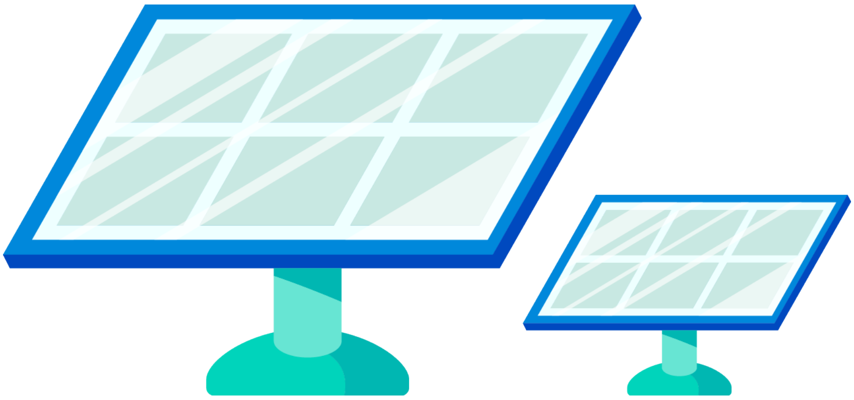
Universiteit Utrecht

Faculty of Geosciences
Copernicus Institute of Sustainable Development

A comparative study of PV simulation and machine learning models on a macrolevel and microlevel



Master's Thesis Energy Science
(GEO4-2510)

A. Zuiker
5550327
Energy Science
System Analysis Track
Utrecht University

Thesis Supervisor:
Dr Atse Louwen
Second Reader:
Prof. Wilfried G. J. H. M. van Sark

Utrecht, May 2019

Abstract

The increasing demand for PV (photovoltaic) modellers brings forward the need for a clear and comprehensive assessment of the most applied PV prediction models. Such an overview should consist of both computer simulation models and machine learning models, as the latter has expanded to the field of PV assessment. Comparative studies have so far been incomplete, limited in accuracy assessment and have never compared simulation modelling with machine learning techniques.

This comparative study determined the modelling accuracy for simulation models PVLib, PVSyst, SAM, PVWatts and Helioscope and for eight different machine learning models. The accuracy is determined by comparing the modelled with the measured DC power output of a commercial PV module, for which the meteorological and performance data is obtained from the Utrecht Photovoltaic Outdoor Test facility. The accuracy is evaluated on a macro- and microlevel, which differentiates between the error of annual electricity yields and the aggregated errors for each data point. The differentiation between the macro- and micro-accuracy provides further insights in a model's optimal application. In addition, the influence of the source of meteorological data, type of solar input irradiance and the resolution of input data on a model's accuracy is determined as well. The sensitivity of the machine learning accuracies to the amount of training data is also determined. Every modelling step is elaborately described to ensure absolute transparency and examination of the model configurations.

It is concluded that four different model combinations under PVLib are the most accurate on both the macro-and microlevel. SAPM is found to most accurately model from global panel-of-array irradiance and the combination of Physical and FSSC the most accurate using global horizontal irradiance. PVSyst and SAM also obtained decent micro-accuracies, but they generally underestimated the electricity yields. The machine learning models proved to accurately predict electricity yields, but generalisation and wider application require more research.

The influence of the source of meteorological data and the type of solar input irradiance influenced a model's micro-accuracy but were not found to consistently influence the macro-accuracy. The time resolution of meteorological data is found to slightly influence both macro- and micro-accuracy. The minimum amount of training data needed for all used data sets that guaranteed decent machine learning accuracies was found to be 8 months.

The results found in this comparative study facilitate in selecting the most suitable PV prediction model for each objective, incorporating both simulation and machine learning options.

Preface

Before you lies the Master's Thesis 'A comparative study of PV simulation and machine learning models on a macrolevel and microlevel'. It was written as a requirement for my Master program Energy Science at the University of Utrecht. The research started in November 2018 and ended in May 2019.

The topic started as an idea of my supervisor Dr Atse Louwen, who used machine learning algorithms to predict annual electricity yields for various prototype PV modules. I was interested to know whether this was more accurate than using computer simulation models. I found out that there was no clear answer to that question, as both techniques were never compared for the application of predicting a PV module's performance. The decision of therefore conducting a comparative study was also aligned with my aspirations to improve my Python programming skills, for which my interest grew after enrolling into several modelling courses.

After finalising my research, I am proud to say that my aspiration is achieved due to hard work and excellent supervision. Atse Louwen always helped me with either my programming or PV modelling questions and guided my research in times that I was unsure about its objectives, for which I like to thank him for.

I hope you enjoy reading my thesis as much as I enjoyed written it.

Aron Zuiker

Utrecht, May 1, 2019

Table of Content

Acronyms and Abbreviations

AC	Alternating Current
AM	Air Mass Coefficient
AOI	Angle-of-incidence
CEC	California Energy Commission
DC	Direct Current
DHI	Direct Horizontal Irradiance
DNI	Direct Normal Irradiance
FSSC	First Solar Spectrum Correction
GHI	Global Horizontal Irradiance
GPOA	Global panel-of-array irradiance
IWEC	International Weather for Energy Calculations
KNMI	Royal Netherlands Meteorological Institute
LID	Light-induced-degrading
MPPT	Maximum Power Point Tracking
NRMSE	Normalised root-mean-error
POA	Plane-of-array (irradiance)
PVPMC	PV Performance Modelling Collaborative
RMSE	Root-mean-square-error
SAM	System Advisor Model
SAPM	Sandia PV Array Performance Model
STC	Standard Test Conditions
UPOT	Utrecht Photovoltaic Outdoor Test facility

Chapter 1 | Introduction

Solar power from photovoltaics (PV) is one of the fastest growing renewable energy sources globally (IEA, 2017b), showing a record 34% growth in 2017 and is expected to triple in power generation in 2023. (IEA, 2018). These figures are a result of the rapidly falling costs of PV modules, increasing efficiency (Kumar Sahu, 2015) and due to its important role in mitigating climate change through its abundant, inexhaustible and clean energy source (Gurupira & Rix, 2017). The rapidly increasing diffusion of PV systems also increases the demand for models that predict PV performance (Kirn & Topic, 2017). The accuracy of modelling PV performance is essential in decision making for assessing the financial viability of PV projects and is required to be as a high as possible to assure economic feasibility and avoid risk (Kirn & Topic, 2017).

Currently there are many different models available for predicting PV performance for both commercial and research purposes. This research distinguishes between two types of prediction models: simulation models and machine learning models. Both are fundamentally different in the way they convert input data into output data. With a simulation model, the model consists of a set of *a priori* rules usually constrained by the laws of physics, i.e. the model is structured and defined before executing the simulation. Machine learning models, on the other hand, set out with an initially undefined model, where both the input and output values are known. Machine learning algorithms train and test a model based on the input and output values. Once the model is trained accordingly and validated for its accuracy it can then be used for different sets of input values (Hackeling, 2017).

Each prediction model has unique features and applications and differs in its degree of complexity and input parameters (Gurupira & Rix, 2017). With the increasing demand for PV performance modellers and the lack of a clear and comprehensive overview, it is a difficult task for new PV modellers to pick the right prediction model that fits their available data and application. Model results can also not easily be compared as they are often running with different sets of software packages, requiring an understanding of different platforms and modelling techniques (J. S. Stein & Klise, 2009). Choosing the right model from the start is essential as modellers are often reluctant to switch later on due to a lock-in effect, primarily due to the time investment that understanding a new model demands.

Comparative studies have been conducted for some simulation models in an attempt to decide which model is the most accurate in simulating PV performance (Dolara et al., 2015; Gurupira & Rix, 2017; B. Marion, 2008). These studies however do not fully cover the most applied or realistic simulation models, nor do they include comparison with machine learning models. PV simulation models such as Helioscope and PVWatts are not covered in comparative studies but are identified by the PV Performance Modelling Collaborative (PVPMC) as widely used simulation models (J. Stein, 2016). Marion (2008) compared three simplistic mathematical models that do not represent realistic weather scenarios and technological performance of PV modules, demonstrating to be unsuitable for accurate simulation of PV performance. Gurupira & Rix (2017) compared the accuracy of the PV simulation models PVSyst, System Advisor Model (SAM) and PVLib but the accuracy was only determined on a macrolevel that did not include more precise error-metrics. The input data was also limited to hourly timesteps, although PVLib and SAM have the ability to work with timesteps of minutes. When striving to the highest model accuracy, taking advantage of the shortest timestep is important to prevent information loss.

Predicting PV performance is not mere limited to simulation models but has expanded to the domain of machine learning. Several studies have been conducted on predicting PV performance based on machine learning algorithms (Kazem et al., 2016; Li et al., 2016). Kazem (2016) demonstrates high accuracies using a machine learning technique called Support Vector Machine (SVM) and proposes similar studies to be conducted that use different machine learning techniques. Lauret et al., (2015) have compared several techniques such as neural networks, Gaussian processes and SVM but these were used for predicting short-term solar irradiance and not for predicting PV performance.

Computer simulation models and machine learning algorithms keep on increasing in complexity and processing power and are being increasingly deployed in predicting PV performance. Although both techniques keep on increasing their presence, they have never both been incorporated in a single comparative study. They remain two separate techniques that so far have not been combined to form a more accurate and powerful method for predicting PV performance. To further develop the field of predicting PV performance it is crucial to start with a comprehensive study that assesses both these techniques and that identifies their strengths and weaknesses. Such comparative study can form the basis for further research on how these techniques can complement each other and to obtain the subsequent synergy.

To provide a clear overview for PV modellers and to stimulate the integration of computer simulation and machine learning, the main aim of this study is to conduct a comprehensive and fair comparison between the most applied PV performance simulation models and various machine learning models. The aim is to assess which of these models is the most accurate and under which conditions and applications it holds. The research question of this comparative study is stated as:

Which of the most applied PV simulation and machine learning models is the most accurate in predicting PV power output?

To understand the conditions and applications under which a certain prediction model is the most accurate, several situations are investigated on their influence on a model's accuracy. These situations are investigated by answering the following four sub-questions:

1) What is the influence of the **source of the meteorological data** on a prediction model's accuracy?

2) What is the influence of the **type of input irradiance** on a prediction model's accuracy?

3) What is the influence of the **input data's resolution** on a prediction model's accuracy?

4) What is the influence of the **time period of input data** on a machine learning model's accuracy?

This study compares the five most applied PV performance simulation models as identified by the PVPMC: PVSyst, SAM, PVLib, Helioscope and PVWatts (J. Stein, 2016). The machine learning techniques used for training the various prediction models are based on the applicable and available methods of scikit-learn. Regression techniques suitable for dealing with large datasets (>10,000 data points) and accessible through scikit-learn are simple and multi linear regression, polynomial regression, k-nearest neighbours regression and decision tree regression (scikit-learn developers, n.d.). For each of these regressors at least one model is trained for predicting PV performance. In addition, three ensemble techniques are applied for improving the machine learning models' accuracy. When speaking of the PV performance this study refers to the direct current (DC) power output of the PV module.

This research identifies two kinds of accuracy: micro-accuracy and macro-accuracy. Micro-accuracy is the degree in which a prediction model can accurately predict the actual power output on a microlevel such as in minutes or hours. Macro-accuracy is the degree in which a model can accurately predict the total annual electricity yield. A model can be inaccurate on the microlevel, but when taken the aggregated results of a whole year it can still be accurate in predicting the annual electricity yield (Tapia & H., 2014). The micro-accuracy of each model is determined by using two indicators of error statistics: root-mean-square-error (RMSE) and normalised-root-mean-square-error (NRMSE). Both indicators present a value of micro-accuracy by indicating the difference between modelled and measured PV performance for every data point (B. Marion, 2008). The macro-accuracy is determined by calculating the error between the measured and modelled electricity yield. In order to determine model accuracy, predicted PV performance is compared with measured performance data of two commercial monocrystalline silicon PV modules.

The two commercial PV modules and corresponding meteorological data have been extensively monitored for 30 months by the Utrecht Photovoltaic Outdoor Testing (UPOT) facility. Meteorological data is also measured by a weather station from the Royal Netherlands Meteorological Institute (KNMI) in De Bilt, which is only 2 km away from the UPOT facility. Meteorological data from both sources is used for predicting PV performance by the prediction models in order to investigate the influence of the data source, regarding sub-question one. The UPOT facility measured the global horizontal irradiance (GHI) and the global panel-of-array (POA) irradiance, which both can be used as input irradiance to predict PV performance. Comparing the influence of these types of input irradiance investigates sub-question two. In addition, the measured UPOT data is resampled into three different timesteps in order to investigate if the input data's resolution influences accuracy, investigating sub-question three. The final and fourth sub-question is answered by varying the time period of the machine learning models' input data and investigating how this influences the accuracy.

Chapter 2 describes the underlying theoretical concepts that are used in modelling PV performance. These concepts explain the physical and chemical phenomena related to PV electricity generation, which form the foundation of the five simulation models. This chapter also provides a brief introduction into machine learning and the various techniques applicable for modelling PV performance. Chapter 3 elaborately describes the modelling steps taken for each simulation and machine learning model and elaborates on the method for determining model accuracy. This chapter also clarifies on the data collected by the two measuring sites, the two commercial PV modules and the various assumptions taken for modelling PV performance. Chapter 4 presents the accuracy results of the prediction models and of several sub-models. These sub-models are internally incorporated by the various simulation models which provide some flexibility in configuring the simulation model. The accuracy of these sub-models is compared in order to select the most accurate internal modelling path of the main simulation model. Chapter 4 ends with a sensitivity analysis of the most accurate simulation model to changing input variables. The discussion section elaborates on the implications and limitations of the model comparison and advocates for further research. The conclusion section ends the report by providing the answer on the main research question and four sub-questions.

Chapter 2 | Theoretical Framework

This chapter consists of three sections. Section 2.1 briefly describes all the theoretical concepts that form the core in modelling PV performance. Section 2.2 introduces the five simulation models that are up for comparison and that are used for modelling PV performance. Section 2.3 provides an introduction into machine learning and describes the various techniques that are used for predicting the PV performance.

2.1 Solar Radiation and Photovoltaic Technology

Every year 3,800,000 EJ of solar radiation is intercepted by earth (Blok & Nieuwlaar, 2017), which is enough energy to power the global energy system for more than 6,500 years (IEA, 2017a). All this solar energy however cannot directly power the energy system and first needs to be converted into useful energy carriers, such as electricity. Converting solar radiation directly into electricity can be accomplished with PV technology, which captures photons that trigger an electromagnetic current in an external circuit (Twidell & Weir, 2015). The entire process of photons travelling down through the atmosphere, touching down on PV modules and being converted in an electric current consists of several steps, each with its own underlying physical dynamics and corresponding energy losses. Modelling PV performance implies accurately modelling all steps of solar energy transport and conversion and requires an understanding of several physical and chemical concepts.

2.1.1 Extra-Terrestrial Irradiance and Weather

Solar radiation reaches the earth's atmosphere with an energy intensity between 1,321 W/m^2 and 1,415 W/m^2 and is called extra-terrestrial irradiance (Blok & Nieuwlaar, 2017; Twidell & Weir, 2015). Passing down through the atmosphere to the surface results in intensity loss as the irradiance is blocked by atmospheric molecules that form a layer around the surface of the earth. Touching upon the surface, the intensity has decreased to an average 1,000 W/m^2 with a clear sky, down to 100 W/m^2 or less on a cloudy day (Blok & Nieuwlaar, 2017). These figures present the substantial effect that the weather has on the solar irradiance available for PV electricity generation.

Besides the loss of radiation, weather also influences the technological performance of PV modules. For modelling PV performance four additional weather parameters are taken into account in this research: the ambient temperature, wind speed, wind direction and relative humidity. All parameters influence the operating temperature of PV systems, and thus its efficiency (Kazem et al., 2014; Touati et al., 2016).

Aside from being affected by weather, the intensity of the solar irradiance reaching the surface is mainly determined by the season and time of day. When the sun is positioned perpendicular to the surface, photons have to travel the shortest distance through the atmosphere, resulting in less diffusion of the irradiance. The shortest distance through the atmosphere is often denoted with an air mass coefficient (AM) of 1, which is defined as "one atmosphere" thickness. The AM increases when the angle between the earth surface and the sun decreases, as the irradiance needs to travel through more atmosphere, resulting in more diffusion (Yella et al., 2011).

The daily cycle of a rising and a setting sun, or day and night, are caused by the rotation of the earth around its axis. There is a tilt in the axis of Earth, called obliquity. Seasons are the effect of a tilted earth orbiting around the sun. This obliquity also has the effect that it is summer on the northern hemisphere, when it is winter on the southern hemisphere, and vice versa. Sun reaches its perceived highest position during summer and is at its lowest during winter, as seen in Figure 1. Hence, PV modules generate the most electricity during the summer (Twidell & Weir, 2015).

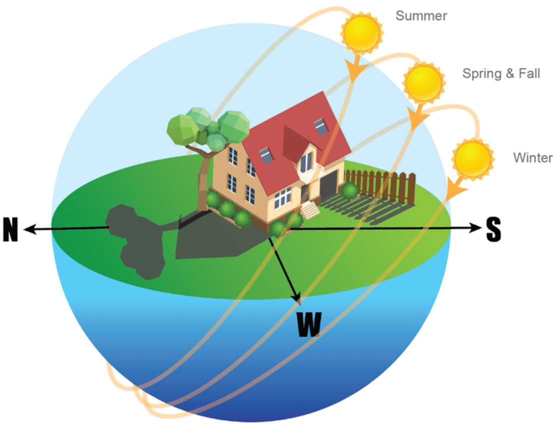


Figure 1. The sun's position in the sky at noon during each season (Water University, n.d.).

The sun's position is indicated using the solar zenith angle and the solar azimuth angle. The solar zenith defines the sun's angular altitude in the sky and the solar azimuth defines the sun's relative position on the horizon. The solar zenith is measured from the vertical, meaning the solar zenith is 0° when the sun is directly overhead and 90° at sunset. The azimuth is commonly defined in the northern hemisphere as equal to 0° in the north cardinal direction and increases clockwise (Reda & Andreas, 2004). The azimuth angle is also used for defining the orientation of PV modules. A PV module with a surface azimuth of 180° means it is faced due south.

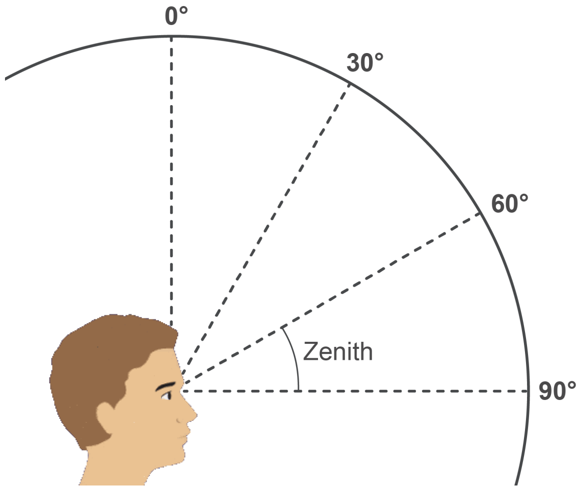


Figure 2. Solar zenith angle (Time and Date, n.d.).

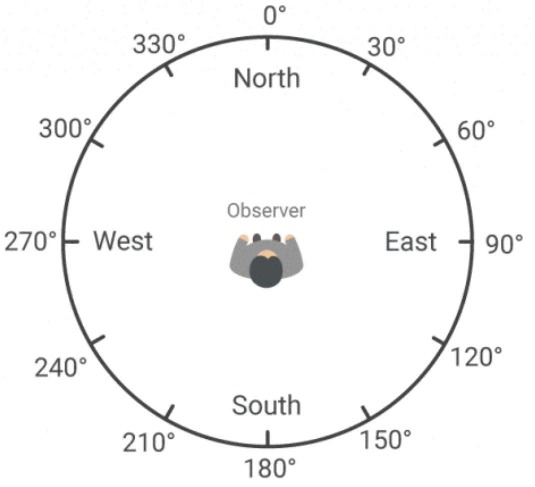


Figure 3. Solar azimuth angle (Time and Date, n.d.).

2.1.2 Incidence Irradiance

According to Twidell and Weir (2015), solar irradiance can be split into direct beam, diffuse radiation and reflected radiation. The first two components are directly related to the weather phenomena discussed in section 2.1.1 and explain why on a cloudy day still 100 W/m^2 of solar irradiance can reach the surface. Direct beam solar radiation is the radiation that is travelling through the atmosphere down to the surface in a straight line without intervention. The direct beam radiation can be completely blocked on a cloudy sky, preventing it from reaching the surface. Fortunately, part of the direct beam radiation is scattered in all directions when colliding with molecules in the atmosphere and can, through the process of diffusion, still reach the surface through a layer of clouds. It has to be noted that even on a cloudless day at least 10% of the solar radiation reaches the surface as diffused radiation (Twidell & Weir, 2015).

Reflected radiation is the radiation that is reflected on the ground and surroundings (everything non-atmospheric) and is dependent on the albedo factor. The albedo factor ranges between 0 and 1. A value of 0 means that the ground is completely non-reflective, and a value of 1 means that all irradiance is completely reflected. Standard values are 0.2 for grassy grounds and 0.6 for snow-covered ground, with the world's average being 0.34 (Luque & Hegedus, 2011).

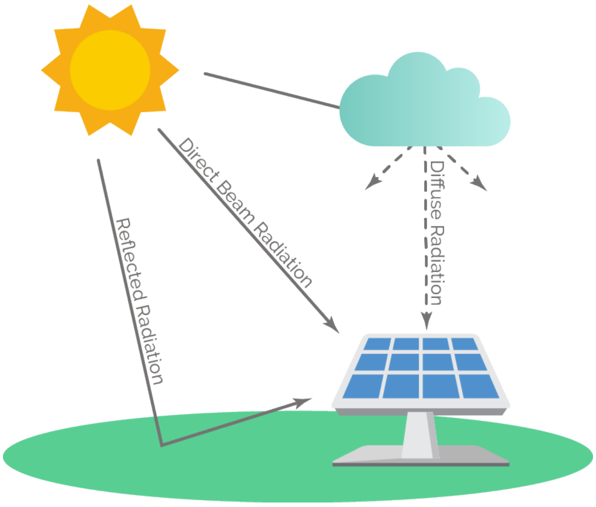


Figure 4. Solar irradiance on a PV module.

The total/global, direct beam and diffuse radiation are denoted by global horizontal irradiance (GHI), direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI) respectively. DHI is usually measured horizontally whereas DNI is measured by the surface normal to the direct beam radiation. DNI and DHI make up GHI according to Formula 1.

$$GHI = DHI + \cos(\theta) \cdot DNI \tag{1}$$

θ is the solar zenith and thus the fraction of DNI increases as the sun is higher in the sky. In practice GHI can only be used for PV performance when split into DHI and DNI. However, often only GHI data is available as it is directly measured by weather sensors. Fortunately several models exist that can convert GHI to DNI and DHI, based on the solar position (Lave et al., 2015).

Because neither GHI, DNI or DHI take into account the orientation and tilt of the PV module, they need to be converted into the POA irradiance before being used for modelling PV performance. For conversion to POA irradiance no exact calculation methods exist, there are only estimation models. The total amount of POA irradiance is called the global POA (GPOA) irradiance, which is the sum of the sky diffuse POA, ground diffuse POA and direct POA irradiance (Lave et al., 2015). After subtracting shading, soiling, reflection and spectral losses from the POA irradiance, that what remains is called the effective POA irradiance and is the actual irradiance that is absorbed by the solar cell and used for electricity generation (IEA-PVPS, 2017).

2.1.3 Shading, Soiling, Reflection and Spectral losses

Shading losses

Shading can have serious consequences for PV modules that are not optimised for shading losses (Martínez-Moreno et al., 2010). PV modules consist of PV cells connected in series. When one or multiple cells are shaded, it affects the output of the whole series of connected and unshaded cells. As a result, the loss in power output can be disproportionate to the shaded area. Depending on which part of the module is shaded, the power output can even decrease to zero. Shading effects can be reduced by using bypass diodes, optimizers or micro-inverters (Harb et al., 2013; Silvestre et al., 2009). For modelling PV performance, shading effects have to be accounted for when solar arrays have an obstructed view.

Soiling losses

Another factor blocking the irradiance is soiling. Soiling is the accumulation of dust and sand particles on PV cells and is generally determined by location and weather (Figgis et al., 2017). PV modules placed on a high rooftop subjected to wind have fewer soiling issues than modules placed on the ground in the desert.

Reflection losses

Reflection losses or angle-of-incidence (AOI) losses are losses of irradiance that is reflected off the PV module. The reflection of the irradiance is mainly determined by the AOI and is measured from the surface vertical. It increases as the sun sets. The larger the angle the more irradiance is reflected (Yusufoglu et al., 2013). Figure 5 presents the relation between the amount of irradiance that is transmitted through the surface glass and foil of a PV module as a function of the AOI. This relationship is important for modelling PV performance as the position of the sun influences the amount of irradiance that reflects off the PV module.

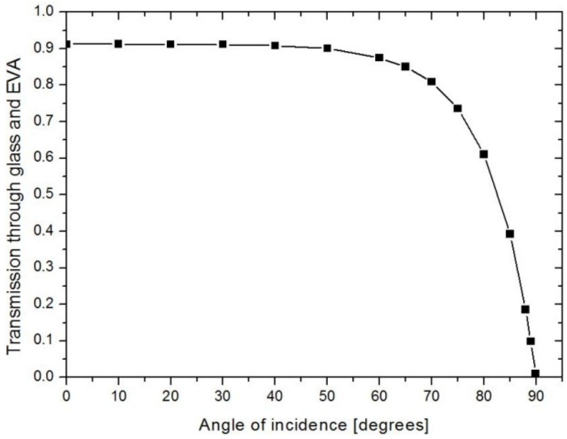


Figure 5. Reflection of irradiance on the module's surface depending on the angle-of-incidence θ (Yusufoglu et al., 2013).

Spectral losses

The solar irradiance consists of a whole spectrum of light with different wavelengths. As the solar irradiance travels through the atmosphere to the ground it loses some of its spectrum through absorption and diffusion of atmospheric particles. The spectral composition that remains after travelling down to the ground depends on the path of the sun (the airmass coefficient AM) and the spectral characteristics of the atmospheric particles. The type of PV technology determines the spectral responsivity of the solar cell, meaning that some wave lengths are better absorbed than other wave lengths (Mavromatakis & Vignola, 2016). Spectral losses thus occur as part of the spectrum of irradiance is lost passing down through the atmosphere and as the PV module is not equally responsive to the entire spectrum of irradiance. Figure 6 illustrates how various PV technologies respond differently to different wave lengths, which highlights the large influence the choice of technology has on reducing spectral losses.

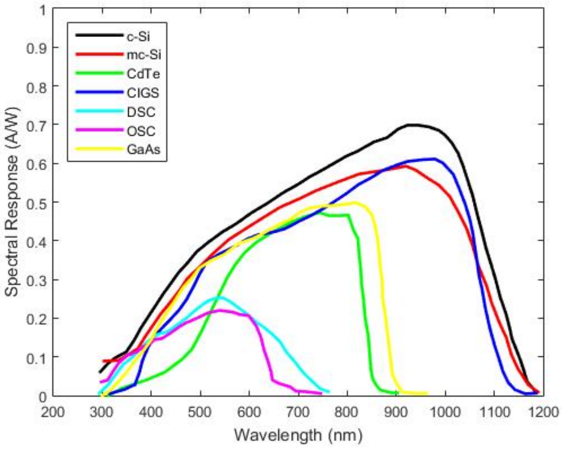


Figure 6. Spectral responsiveness of various solar cells (Sandia, n.d.).

2.1.4 Photovoltaic Cells

A photovoltaic module consists of an array of solar cells, capable of directly converting solar energy into electricity. Solar cells are made up by two flat layers of semiconducting materials that are separated by a tiny gap, called the p-n junction, as presented in Figure 7 (Çengel, A. & Boles, A., 2015). When photons are caught by the semiconductors, they separate positive and negative charge carriers in the absorbing material. Due to a permanently existing electric field at the p-n junction, the charge carriers flow in a direction according to their charge and consequently produce a current in an external circuit (Twidell & Weir, 2015).

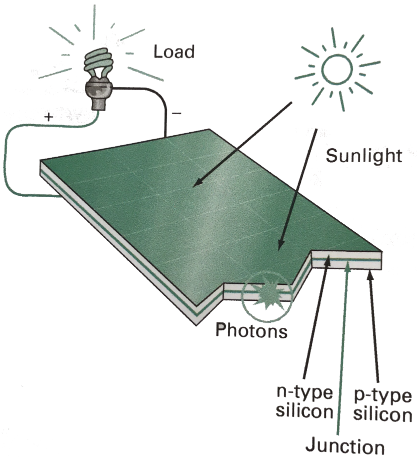


Figure 7. Diagram of solar cells (Twidell & Weir, 2015).

2.1.5 Single Diode Model

The current of solar cells depends on the surface of the solar cell and is often denoted as the current density J, defined by the current I over the cell surface A (Çengel, A. & Boles, A., 2015). PV cells are often simplified with a single diode equivalent circuit model presented in Figure 8.

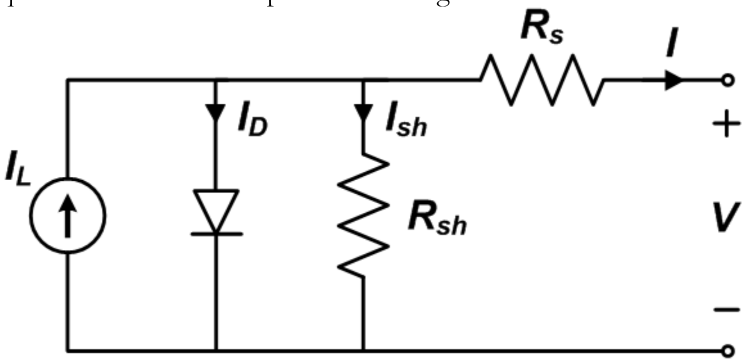


Figure 8. Single diode equivalent circuit model of a solar cell (Luque & Hegedus, 2011).

Here I_L is the light-induced current from the solar cell, I_D the dark diode current (a current occurring even when the solar cell is in the dark), I_{SH} the shunt current, R_{SH} the shunt resistance, R_S the series resistance and V the cell output voltage (Luque & Hegedus, 2011).

The presence of the shunt resistance provides an alternate path for the light-induced current, resulting in an unwanted shunt current which reduces the cell output voltage (Bouzidi et al., 2007). Technological developments have reduced shunt losses to negligible effects, thus shunt resistances are often assumed to be infinite (Twidell & Weir, 2015).

The series resistance is the internal resistance of the solar cell caused by the contact resistance between the metal contacts and the semiconductor, the resistance of the top and rear metal contacts and through the movement of current through the emitter and base of the solar cell (Honsberg & Bowden, n.d.). R_S is higher when several modules are connected in series, due to the extra interconnections. R_S should be minimised to maximise the power output but increases with the temperature. A solar cell therefore decreases in efficiency when the ambient temperature increases (Twidell & Weir, 2015).

The current I can be written according to the single diode equation.

$$I = I_L - I_D - I_{SH} \tag{2}$$

$$I = I_L - I_D - \frac{(V - I \cdot R_S)}{R_{SH}} \tag{3}$$

In practice Formula 3 is hard to analytically solve as the current I is a function of voltage V and current I. Several iterative and analytical approximation methods exist that attempt to solve the equation such as the Lagrange method, Bishop's Algorithm and least-squares numerical techniques (Bishop, 1988). A technique however preferred for its analytical approach (therefore computational preferable for PV modelling) is the Lambert W-function. The Lambert W-function is a more complex single diode equation for which the current I in the right term of the equation is substituted (Jain & Kapoor, 2004). The Lambert W-function is used in this research and is further described in subsection 3.3.2.

The power output of a solar cell is indicated with an IV curve (Figure 9) which plots the generated power as function of the cell voltage and current. The power output is the product of current and voltage and should be maximised. The maximum current that can occur is called the short-circuit current I_{SC} and is the current from a solar cell that occurs when the solar cell is short circuited which happens when the voltage applied is zero. The maximum voltage output is the open-circuit voltage V_{OC} and occurs when the current is zero. It presents the maximum potential of the electric field in the p-n junction (Honsberg & Bowden, n.d.). The maximum potential of the p-n junction decreases with increasing temperature and shifts the IV curve to the left, lowering the power output of the solar cell.

Both the voltage and current are affected by incoming irradiance. V_{OC} increases slightly with increasing irradiance and increases I_{SC} proportionally. Although the current and voltage are the highest at I_{SC} and V_{OC}, the power output at both points equals zero. The maximum power obtainable is denoted by P_{MP} (M_{PP} in Figure 9) and is the product of the maximum power current I_{MP} and maximum power voltage V_{MP} (Twidell & Weir, 2015).

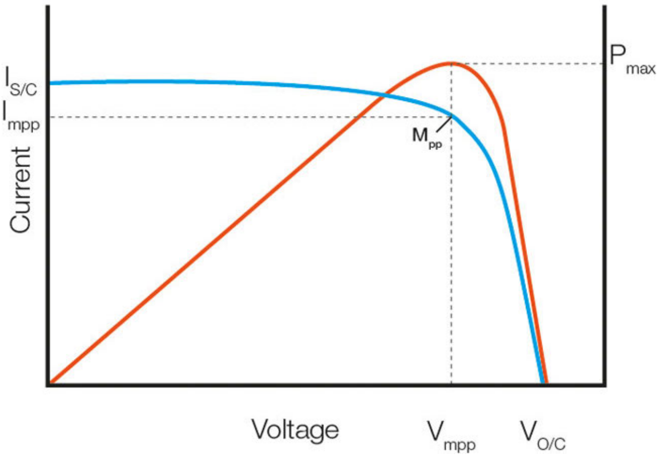


Figure 9. IV curve of a solar cell power output (Seaward, n.d.).

In realistic scenarios the IV curve looks very different from the ideal curve presented in Figure 9. Due to energy losses the IV curve changes, resulting in a decrease in power output. Figure 10 illustrates how the power output decreases as a result of losses.

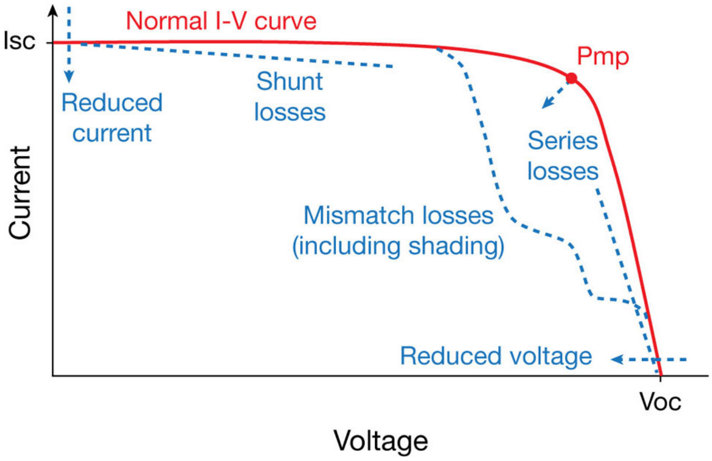


Figure 10. Influence of cell and system losses on IV curve (Hernday, 2011).

Shunt and series losses affect the power by lowering the current and voltage output. In addition, mismatch losses can have a devastating effect on the output of a solar array. Mismatch losses (including shading) occur when PV modules with different IV characteristics are connected in an array. When the modules are connected in series, the module with the lowest current determines the current output of the entire string of modules. Similarly, when modules are connected in parallel, the voltage output is driven by the module with the lowest voltage (Honsberg & Bowden, n.d.).

2.2 Simulation Models

Five simulation models are used for predicting PV performance: PVLib, SAM, PVWatts, PVSyst, and Helioscope. All models are briefly described in this section.

PVLIB is a free and open source library of modelling functions, focused on simulating PV performance and is developed by a group of PV professionals of the PVPMC and Sandia National Laboratories (SNL) (Gurupira & Rix, 2017). PVLib has the most modelling flexibility as it is possible to access and edit the source code directly. This degree of modelling flexibility however requires users to have some knowledge of programming, which might prevent modellers from working with PVLib (Andrews et al., 2014). The PVLib packages are available for Python and MATLAB and are compatible with Windows, macOS and Linux.

SAM is a free software package developed by the National Renewable Energy Laboratory (NREL) that can simulate system performance of a wide variety of renewable energy technologies, such as solar PV (Gurupira & Rix, 2017). SAM is more user friendly than PVLib as it has a graphical user interface (GUI) for inserting input parameters, choosing scenarios and displaying results. SAM is a desktop application that runs on Windows, macOS and Linux.

PVWatts is a free web application developed by NREL for rapid PV production calculations. It is a more simplified version of SAM and is limited in input parameters and scenarios. PVWatts is cross platform and cloud based but can also be run in SAM.

PVSyst is a paid PV performance Windows-only desktop application originally developed by the University of Geneva, but later commercialised. PVSyst can be used for free with full capabilities for a trial period of 30 days. Just like SAM it has a user-friendly GUI that makes it possible to model PV performance for users not experienced with computer modelling (Gurupira & Rix, 2017). PVSyst can only be run on Macs using systems using Boot Camp or through virtualisation software like Parallel Desktops or VMware Fusion.

Helioscope is a paid commercial PV performance model developed by Folsom Labs (Folsom Labs, n.d.). Helioscope has a very user-friendly environment with advanced GUI but offers limited configuration possibilities. Helioscope is a cross platform cloud-based web application that offers a 30-day trial with full capabilities.

2.3 Machine Learning

Machine learning is the study and design of software that utilises information from the past to inform about future probabilities. It encompasses techniques that learn from experiences and observations without being specifically programmed for the specific set of information (Hackeling, 2017). A machine learning algorithm starts with an unknown model but known input and output values. The algorithm then tries to build a mathematical model with the best 'fit' on the available data. The best fit is the trained model for which the bias and variance between the predicted value and the actual given output value is the lowest. Once a model is built it can be used to make predictions with a new set data for which the output data is unknown (Unpingco, 2016). If the model is proven to be accurate enough it can provide valuable information about unknown future probabilities. (Hackeling, 2017).

For predicting PV performance this implies that a machine learning algorithm can accurately predict the power output of a PV module without it having any knowledge on PV technology or on the characteristics of solar irradiance. Instead of using fundamental properties and a set of empirical rules that define the process of going from a certain situation to a certain output, machine learning algorithms merely examine input and output values and ignore any pre-determined model that connects the two (Alpaydin, 2009).

Instead of using the terminology of input and output variables this study defines input variables of machine learning models as independent variables and the output variables as dependent variables. For machine learning it is incorrect to refer to dependent variables as output, as they are also used as input for the algorithm. Independent variables are the pre-determined variables that do not change by the machine learning experiment. Consequently, dependent variables are the values that can change depending on the experiment. Once a machine learning model is built by the algorithm, it converts the independent variables into the dependent variables (Hackeling, 2017).

Today many different machine learning algorithms exist, from simple regression models to advanced mathematical complex algorithms. The aim of this study is not to describe and explain the exact functioning of these algorithms but only to explore the potential of some of these algorithms for predicting PV performance. This section therefore only briefly describes the several machine learning techniques that are used, without giving an in-depth explanation.

2.3.1 Training, Testing and Validating

Before feeding data to a machine learning algorithm the data is usually split into a set of training and testing data. The set of training data, containing both dependent and independent variables, is used by the algorithm to 'train' the model for a best fit. The independent variables from the test data are then given to the trained model to make predictions. The predictions are then compared to the dependent variables of the test data to evaluate the accuracy of the trained model (Hackeling, 2017).

In the case of training a PV model this entails providing part of the measured meteorological data (independent variables) and measured PV performance data (dependent variables) to the machine learning algorithm as training data. The remaining part of meteorological data is then fed to the model as test data, once its trained, to predict its corresponding PV performance. The predicted performance is then compared to the actual measured PV performance.

Providing the model with a set of test data is crucial in assessing if the model is not 'overfitted'. A model is overfitted when it practically memorized its observations by building a too complex model. An overfitted model can accurately predict dependent variables only if it is given the exact same independent variables that were used for training the model (Müller & Guido, 2017). When a new set of unknown independent variables are given to the model it fails to accurately make predictions as it only remembered the outcome of the known processed data. Contrary to overfitting, a model can also be underfitted. An underfitted model fails to sufficiently fit the training data and cannot be generalised with new data. Underfitting is usually the result of a too simple model or an insufficient amount of training data (Hackeling, 2017).

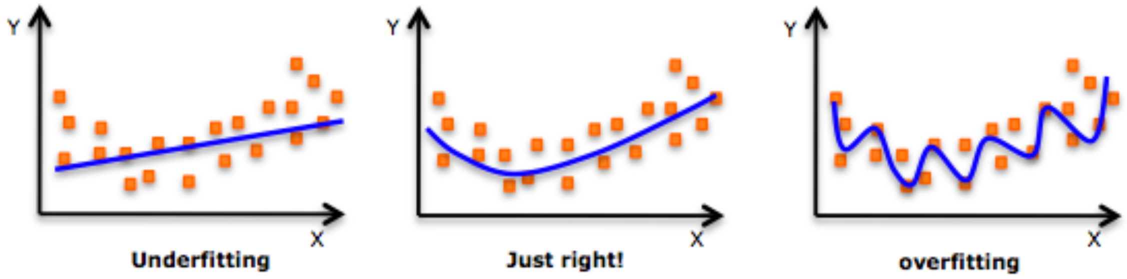


Figure 11. Examples of overfitting and underfitting (Bronshtein, 2017).

Cross Validation

A more advanced method of using split data is by means of cross validation. With cross validation the whole dataset is split into a certain number of folds. The model is then trained using all but one fold and is tested on the remaining fold. The folds are then rotated and used to further train the model until each fold has been used as testing set (Figure 12) (Unpingco, 2016). Cross validation generally makes more efficient use of the available dataset and provides more accurate results than using only a single set of training and testing data (Hackeling, 2017).

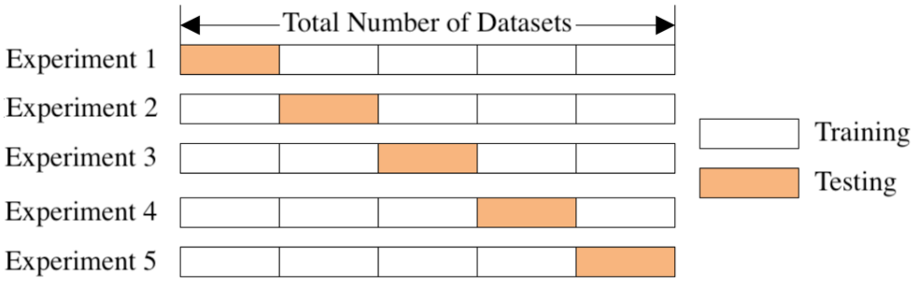


Figure 12. Cross Validation (Pawar, 2018).

Sometimes a third set of observations is used for additional validation. This can be valuable for cross validation as the third set of observations provides data still unknown to the model, which helps to determine if the model can be generalised to new data (Cielen et al., 2016).

2.3.2 Linear Regression

Linear regression is the most basic machine learning technique and entails training a model that describes a linear relation between a continuous dependent variable and one or more independent variables. The most basic method is simple linear regression, which is a machine learning technique that models the relationship between a single independent variable and single continuous dependent variable (Unpingco, 2016). It can only fit a linear model, meaning it can only accurately predict if the relationship between the dependent and independent variable can be approximated by a linear correlation. The model of a simple linear regression is described by the following equation (Hackeling, 2017).

$$y = a + b \cdot x \tag{4}$$

y is the predicted value (dependent variable) and x the predictor (independent variable). a and b are the intercept and coefficient respectively and are both estimated by the learning algorithm. Multiple linear regression essentially works the same but uses two or more independent variables for training the model. This requires the algorithm to find at least one extra coefficient to build the estimation function. Advanced machine learning models can have even more coefficients and predictors contained by more complex non-linear equations (Hackeling, 2017).

2.3.3 Polynomial Regression

Real life situations can seldom be approximated with just linear relations. A curvilinear relation changes the estimation equation by adding at least a second-order polynomial (Hackeling, 2017).

$$y = \alpha + b_1 \cdot x + b_2 \cdot x^2 \tag{5}$$

Higher order polynomials can be added for fitting the model, but precaution has to be taken as this can lead to extremely complex and overfitted models (see Figure 11). A technique used to restrict the complexity of polynomial regression is called regularisation. Machine learning models with regularisation incorporate an additional parameter (Lambda) in the model equation which penalises complexity. Consequently, a model with the fewest assumptions is favoured over complex models with high-order polynomials. Parameters such as Lambda are set manually before training the model (Unpingco, 2016).

2.3.4 K-Nearest Neighbours Regression

With K-Nearest Neighbours Regression (KNN) the machine learning algorithm uses feature similarity to predict values of new data. New independent variables are evaluated on how close or similar they are to known independent variables and are given a mean value of the corresponding dependent variables. The accuracy of the predicted value depends on the number of neighbours it takes its value from and is determined by the K-value (Müller & Guido, 2017). If only the closest neighbour is used for predicting the new value, the model tends to overfit the data. If the mean of too many neighbours is used, the model is underfitted and not accurate enough. The ideal number of the K-value is presented by an elbow curve, which plots the model error as a function of the K-value. The elbow curve is case specific and is empirically found (Singh, 2018).

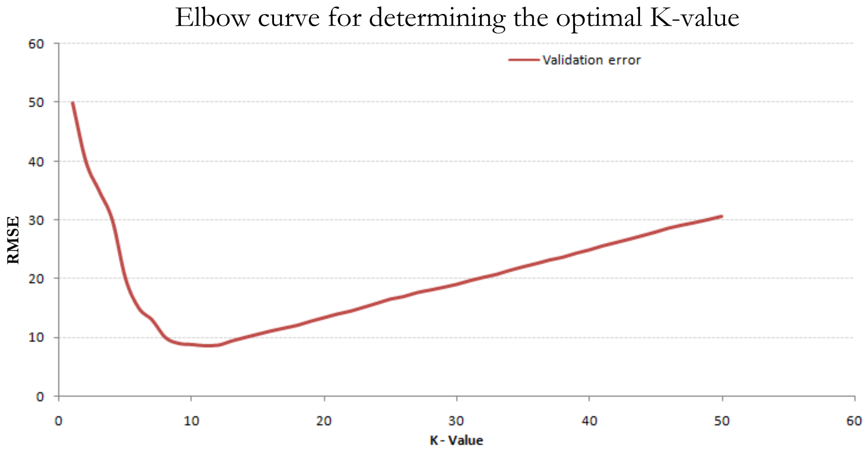


Figure 13. Elbow curve for determining the optimal K-value (Singh, 2018).

2.3.5 Decision Trees

A decision tree machine learning algorithm is a technique that uses a flowchart structure to make categorical decisions that result in predictions. The flowchart is a hierarchical tree diagram built up by categorical statements (nodes) that categorise input data as either true or false, based on the statement of the decision node (Müller & Guido, 2017). The independent input variable works its way down through various decision nodes until it reaches the end node (result). The tree model including all its decision and end nodes is called a decision tree. Figure 14 is an example of a decision tree that uses the weather outlook to predict the number of hours that are played on a day at the tennis court. When a value, in this case the weather outlook, starts at the highest node (the root node) and is evaluated on all conditions of its branch then it ends at an end node that gives the outlook its prediction for the number of hours played.

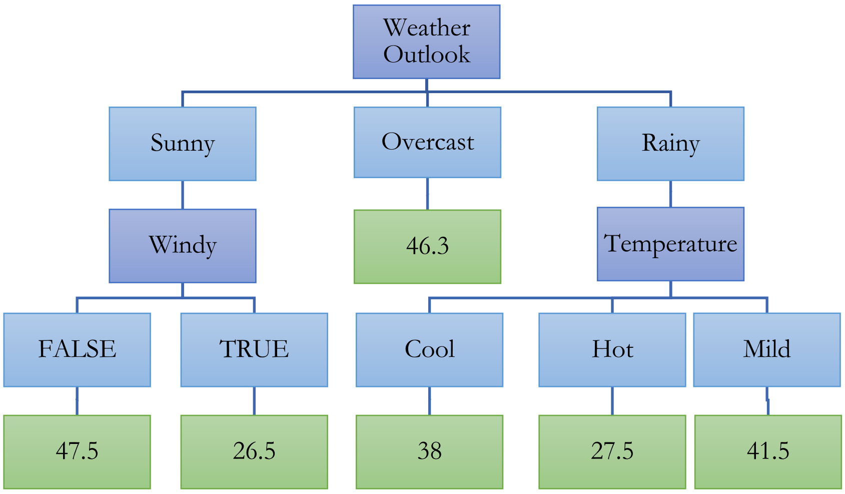


Figure 14. Decision and end nodes making up a decision tree (Das, n.d.).

A decision tree algorithm builds such a tree by grouping data with similar features into separate regions. Every region is then grouped into more sub regions which increases the homogeneity of the data. This process continues until every data point has an individual end note or until the model is called a hold and groups the remaining data points together in an end note (Hackeling, 2017). A decision tree is however not solely limited to making discrete predictions but can also have a continuous output, making it suitable as a regressor. For regression a prediction value is calculated by taking the mean of all the values of the trained data that are within a certain sub region or end node. A decision tree regressor has to be told manually to stop when a certain minimal sample of values in a sub region is left, when the variance of another split node only increases or when an end node is reached (Unpingco, 2016).

2.3.6 Ensemble Methods

An ensemble is a combination of machine learning algorithms that performs better than each of its individual components. Ensemble methods are distinguished between bagging, boosting and stacking.

Bagging

Bagging (bootstrap aggregating) is a method specialised in reducing the variance and bias of a machine learning model. Basically it takes several random samples from the train data and independently trains a model on each of the subsets (Unpingco, 2016). This method is especially useful in improving the accuracy of decision tree models that typically have high variance. Bagging of decision trees consist of creating multiple sub-models or 'trees' for different samples of the training data. When a prediction has to be made for a new data point, each sub-model predicts its own value based on the same independent variables. From the pool of predictions the mean is then taken as the final and main prediction. Increasing the number of trees generally decreases the variance and bias of the model. The marginal return in increased accuracy however decreases after creating more trees and requires more computational processing time. A trade-off is made between processing time and accuracy in choosing the number of trees. The number of trees or models that are trained is defined as the number of estimators and is manually set. (Hackeling, 2017).

Boosting

Boosting is similar to bagging in the sense that it trains multiple models on random subsets of data. The difference is that with bagging the models are trained independently from each other, whereas with boosting the models are trained in a way that dictate each other's training. Values that are incorrectly predicted by one model are more focussed on in other models, by giving more weight to the weakly predicted value. The different models are essentially working together instead of working separately. Boosting can lead to more accurate results than bagging, but is in turn more vulnerable to overfitting (Hackeling, 2017).

Stacking

The last ensemble method entails the stacking of multiple machine learning algorithms. The stacked models are generally divided between two levels of regression, commonly referred to as level 0 and level 1. In level 0 several machine learning models are trained separately with the given set of training data. These models can be based on fundamentally different machine learning techniques such as decision tree, K-Nearest

Neighbours and polynomial regression. The separate models in level 0 are called base-estimators. As soon as all the base-estimators have made their predictions with a new set of data, the predictions are passed on to the next level: level 1. This level is composed of a single estimator called the meta-estimator. The meta-estimator intelligently combines the results of the various base-estimators to increase the accuracy of the final prediction. The meta-estimator uses the best features of each base-estimator and avoids using its inaccurate features. E.g., base-estimator A can be very accurate at dealing with high values whereas base-estimator B is highly accurate with low values. The meta-estimator then assigns more weight to the result of base-estimator A for high values and more weight to base-estimator B for low values. With stacked regression each model's best traits are used, leading to higher accuracies than of each of the individual models (Hackeling, 2017; Unpingco, 2016).

Chapter 3 | Methodology

In this study a total of thirteen prediction models are used to model PV performance: five simulation models and eight machine learning models. The accuracy of all the predictions models is determined by comparing the modelled DC power output of the two commercial PV modules with the actual DC power output measured by the UPOT facility. Accuracies are precisely calculated on a macro- and microlevel. Both macro- and micro-accuracy are calculated using error-metrics, adding to the quantitative nature of this research. The error-metrics are described in section 3.1.

The prediction models use meteorological input data from both the UPOT facility and the KNMI. The KNMI meteorological data is added to investigate the influence of the data source on a model's accuracy. In addition, some weather measurements required for modelling PV performance were not measured by the UPOT facility and are provided by the KNMI. Whereas the KNMI only measures GHI, the UPOT facility also measured the orientation specific global POA irradiance. In order to investigate the influence of these two types of input irradiance, the prediction models estimate the PV performance using both. Some simulation models, however, are not capable of modelling from global POA irradiance. For these models the influence is therefore not investigated. In addition, the timesteps of the UPOT data is resampled to 2, 15 and 60-minute timesteps, in order to investigate the influence of the resolution of input data. The data used for modelling the PV performance is concisely described in section 3.2.

Modelling the PV performance from the collected data is extensively described for each prediction model. For each modelling step it is clearly stated which data is used as input, what the output is and on which assumptions it is based. The modelling steps for the five simulation models are described in section 3.3 and those for the eight machine learning models in section 3.4.

In order to assess the sensitivity of the accuracy results, two sensitivity analyses are conducted. The first sensitivity analysis is conducted by changing the value of several input variables for the most accurate simulation model. The second sensitivity analysis is conducted to assess the influence of changing the time period of input data on the most accurate machine learning model. Both sensitivity analyses are described in section 3.5.

3.1 Determining Accuracy

The accuracy of the prediction models is determined by comparing the modelled DC power output with the measured DC power output on a macro- and microlevel.

3.1.1 Macro-Accuracy

The accuracy on macrolevel is indicated using the yield error of the total modelled DC power output. The macro-accuracy is calculated according to Formula 6.

$$Yield\ Error = \left(\frac{E_x}{E_y} - 1\right) \times 100\% \tag{6}$$

with,

E_x = measured (electricity) yield [kWh]
E_y = modelled (electricity) yield [kWh]

The yield error is the indication for the difference between the modelled and measured electricity yield for a certain period of time. Although this can be any defined period of time, this research defines the yield to be a full year or more to adhere to the macrolevel that is investigated. The macro-accuracy is interesting for modellers only concerned about annual yields, which is often the case.

The yield error in this research is also used for comparing between the modelled and measured global POA irradiance. In this case the term insolation error is used instead of the yield error.

3.1.2 Micro-Accuracy

The accuracy on microlevel is indicated with the root-mean-square-error (RMSE) and the normalized root-mean-square-error (NRMSE). Both indicators measure the error between the modelled and measured PV performance (B. Marion, 2008).

The RMSE expresses the average magnitude of the error between the modelled and measured values and is one of the most commonly used metrics for error analysis. The RMSE squares the errors before they are averaged, giving more weight to larger errors to better identify models with large variance (B. Marion, 2008). The RMSE though falls short in comparing datasets with different scales. The datasets in this research are resampled into three different timesteps, making the RMSE unsuitable to compare them among each other as they contain different scales after resampling. To better compare between different timesteps the NRMSE is used, which normalizes the error, making it more suitable to compare between datasets with different scales. There are several methods for normalizing RMSE, such as dividing by the mean or the range of the measured data, but all methods have the same effect (Tong & Granat, 1999). For this research the RMSE is normalised by dividing it with the standard deviation of the measured data (Formula 9).

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - x_i)^2} \tag{7}$$

$$NRMSE = \frac{RMSE}{s} \tag{8}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{9}$$

with,

y_i	=	i_{th} modelled value
x_i	=	i_{th} measured value
$\bar{x}$	=	mean value of measured data points
n	=	number of measured or modelled data points
s	=	standard deviation

The NRMSE is also used for intermediary modelling steps. Several modelling options that influence the final results are made throughout the modelling process. Due to the exploratory nature of this research and the wide amount of prediction models that are up for comparison, the PV performance is not simulated for all different model configurations. Choices had to be made in order to limit the amount of data forks. A data fork means that a modelling path is split into two or more paths. E.g., different irradiance decomposition models can be selected for decomposing GHI into DNI and DHI in some simulation models. In order to determine which irradiance decomposition model leads to the highest accuracy in the end result, the output of all optional models is used for further modelling. Three modelling choices lead to three data forks, which split the number of modelling paths in at least eight different paths (2^3). Given the fact that the source data is initially resampled into three timesteps it illustrates the importance of limiting the amount of data forks in order to limit computational processing time. Fortunately, UPOT measured the global POA irradiance, which is used as an intermediary value of comparison. After modelling the global POA irradiance, numerous modelling paths can be excluded from further modelling, as only the most accurate path is chosen to continue modelling PV performance. Limiting the amount of data forks is mostly a concern for modelling with PVLib as it offers the widest flexibility in selecting different modelling options.

3.2 Data Collection

The input data used for modelling and evaluating the PV performance of the commercial modules is divided between measured meteorological and performance data, defined technical parameters and various assumptions. This section describes the data sources from which the data is collected and specifies on the exact variables and parameters that are used.

This section is limited to describing external data collection only. More data is used such as meteorological data from Meteonorm and the International Weather for Energy Calculations (IWEC) in Amsterdam but these datasets and others have been internally accessed through several simulation models and are thus seen as part of the corresponding simulation model and its underlying content. It is noted in the modelling steps (section 3.3) when such internal datasets are used.

Meteorological Data

Meteorological data is collected from both the UPOT facility and the KNMI. The UPOT meteorological data consists of GHI, 3 separate measurements of global POA irradiance, air temperature, wind speed, wind direction and relative humidity; all measured with a 5-minute interval at night and a 30-second interval during the day. The measured time period of the weather data is equal to the PV performance data of the two commercial PV modules and starts at 1-1-2015 0:00 and ends at 23-12-2017 0:00.

The KNMI meteorological data is obtained from a weather station in De Bilt, which is only 1,700m away from the UPOT facility and thus represent similar weather. The KNMI dataset consists of GHI, air temperature, wind speed, wind direction, relative humidity, precipitation and air pressure; all measured with a 60-minute interval. The KNMI data is retrieved from 1-1-2015 0:00 until 1-1-2018 0:00. GHI from UPOT and KNMI are further denoted as GHI_{UPOT} and GHI_{KNMI} and global POA irradiance from UPOT as $GPOA_{UPOT.}$ Contrary to the KNMI data, the UPOT data has some data gaps that are presented in Table 1.

Table 1. Missing meteorological UPOT-data.

Year	Start Date	End Date	Data points	Data gaps		
				> 1 hour	> 5 hours	> 24 hours
2015	01-01-2015	31-12-2015	599,308	7	2	0
2016	01-01-2016	23-12-2016	574,670	2	2	0
2017	28-04-2017	22-12-2017	366,395	49	26	0

The year 2015 of the UPOT data has the most data points and is the only year with data from every day (although 2016 only misses one week of winter data). For modelling PV performance for the year 2016 and 2017 KNMI data for the period equivalent to the data gaps of UPOT is removed from the dataset to account for missing data and to make a fair comparison.

PV Performance Data

All measured PV performance data of the two commercial PV modules is obtained from the UPOT facility. Both modules have measured data consisting of global POA irradiance, air temperature, V_{OC}, I_{SC}, P_{MP}, I_{MP}, V_{MP}, R_{SH} and R_S, with measurements taken every 2 minutes during daylight. The DC power output of the PV modules is measured before the inverter, meaning inverter losses are excluded from modelling. Module 1 has the least amount of missing data, although the amount is limited for both modules. The data gaps are presented in Table 2.

Table 2. **Missing PV performance UPOT-data.**

Year	Start Date	End Date	Data points	Data gaps		
				> 1 hour	> 5 hours	> 24 hours
Module 1						
2015	01-01-2015	31-12-2015	115,233	5	0	0
2016	01-01-2016	23-12-2016	114,196	6	0	0
2017	28-04-2017	22-12-2017	59,864	0	0	0
Module 2						
2015	01-01-2015	31-12-2015	97,434	22	0	0
2016	01-01-2016	23-12-2016	96,884	15	0	0
2017	28-04-2017	22-12-2017	54,123	0	0	0

Technical Parameters

The two PV modules measured by the UPOT facility are two identical commercial modules with a rated power of 265 W. Both PV modules are made of monocrystalline silicon solar cells and are mounted with a 37° surface tilt on open racks with fixed axes. The modules are located on the 35 m tall rooftop of the Hans Freudenthal building on the campus of Utrecht University (van Sark et al., 2012). The altitude of the location is 2 m, meaning the altitude of the modules on the roof is 37 m. The modules are orientated south (surface azimuth 180°) with an unobstructed east-south-west horizon. The geographical coordinates of the building are 52°05'15.8"N 5°10'03.0"E. These coordinates are used for modelling the solar position at the location of the PV modules. PV performance is location specific, and thus these coordinates are used to make sure the solar position variables are the same for all models. The simulation models require specific technical parameters of the PV modules as input values. These parameters are obtained from STC measurements and the California Energy Commission (CEC) module database.

Assumptions

Energy losses due to shading are assumed to be zero as the PV modules have an unobstructed view and are placed on a high rooftop. A customary and constant annual albedo factor of 0.2 is assumed (Kotak et al., 2015). Inverter losses are not modelled as the measured DC output is measured before the inverter. Some additional losses are applied with average values retrieved from literature. These losses are soiling losses (1.5%), cabling and connection losses (1.0%), light-induced-degradation (LID) (1.0%) and maximum power point tracking (MPTT) losses (0.1%) (IEA-PVPS, 2017).

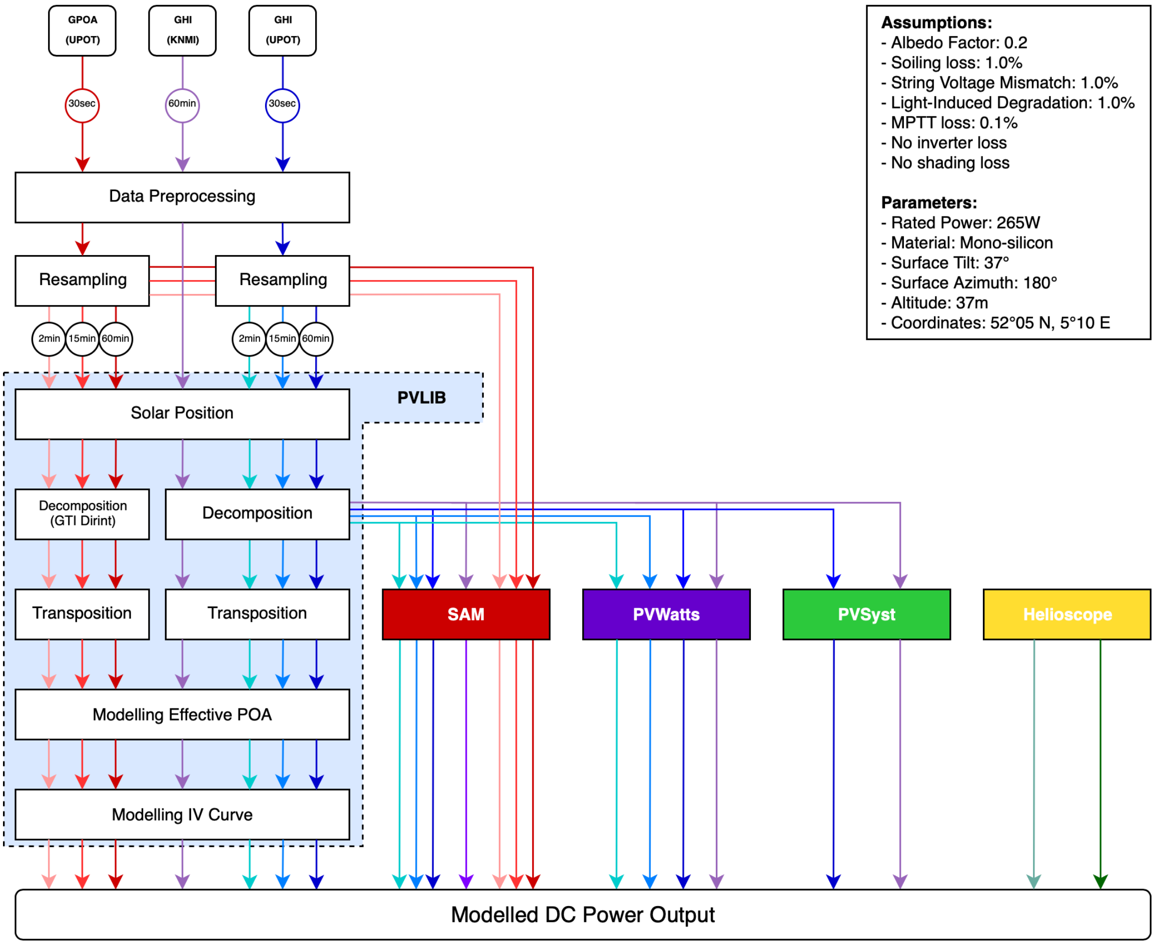
The smallest timestep for which PVSyst and Helioscope can model is a 60-minute timestep. For SAM and PVWatts a 1-min timestep is the limit. PVLib and all machine learning models can run with timesteps of nanoseconds and form no limitations in this regard. As the smallest timestep that all prediction models can run is a 60-minute timestep this timestep is selected as the main timestep used for all the prediction models. To assess the full potential of the other models and to assess the influence of data resolution, they are also run with 2-minute and 15-minute timesteps. The 2-min timestep is equal to the timestep of the measured PV performance data whereas the intermediary 15-minute timestep is chosen to provide additional information on the use of different timesteps. The source data is initially resampled into the three different timesteps.

3.3 Simulation Modelling Steps

This section precisely describes all modelling and data pre-processing steps taken for each simulation model. Only a single PV module is modelled as the two measured commercial PV modules have identical characteristics. Aside from Helioscope, meteorological data for all prediction models is pre-processed using Python. Helioscope is limited to only use its own data sources and does not provide the ability to import custom data files. Pre-processing primarily consists of initial resampling into the three timesteps, complementing UPOT meteorological data with KNMI data, dealing with missing and faulty data and constructing readable datasets for each individual simulation model.

Modelling PV performance in this study has two separate starting points. One point is to start modelling from the measured GHI and the other is to start from the measured global POA irradiance. Both starting points are however not possible for all simulation models. PVSyst and PVWatts cannot be modelled using global POA irradiance as initial irradiance and for Helioscope neither one of the two starting points is possible as pointed out earlier. For modelling using KNMI's meteorological data it is also not possible to start from the second point as it only provides GHI.

In addition, SAM, PVWatts and PVSyst cannot model PV performance with only GHI as irradiance input but require either DNI or DHI irradiance in addition to GHI. Neither UPOT or the KNMI has measured DNI or DHI so this needs to be modelled first from the measured GHI. Fortunately, PVLib includes several irradiance decomposition models for splitting GHI into DNI and DHI irradiance. For providing SAM, PVWatts and PVSyst with the necessary input data, GHI is decomposed using PVLib. This means that the first modelling steps of PVLib are also part of SAM, PVWatts and PVSyst's modelling paths. Once DNI and DHI are modelled with PVLib, the data is exported to be used separately for these three models, officially marking the split of the modelling paths of PVLib, SAM, PVWatts and PVSyst. Figure 15 provides a clear flowchart explaining the different sets of input irradiance and how they are used for the simulation models. All steps in the flowchart are explained in further detail in subsections 3.3.1 to 3.3.6.



Assumptions:
- Albedo Factor: 0.2
- Soiling loss: 1.0%
- String Voltage Mismatch: 1.0%
- Light-Induced Degradation: 1.0%
- MPTT loss: 0.1%
- No inverter loss
- No shading loss

Parameters:
- Rated Power: 265W
- Material: Mono-silicon
- Surface Tilt: 37°
- Surface Azimuth: 180°
- Altitude: 37m
- Coordinates: 52°05 N, 5°10 E

Figure 15. Flowchart of simulation models.

3.3.1 Data Pre-Processing

Pre-Processing

1) Two Python libraries are initially imported and used for data pre-processing. These are data analysis library Pandas and the mathematical library Numpy.

2) In order to calculate the accuracy of the prediction models it is firstly required to process the measured PV performance data. The DC power output for both PV modules is measured but some data is missing for both datasets. Both datasets are therefore combined into one complete dataset. The average is taken of both PV modules when both modules have a measured power output. When only one module has data available for a data point then this value is taken.

3) The meteorological data from UPOT and KNMI is imported from csv files. The UPOT dataset consists of 1,540,373 data rows and the KNMI dataset of 26,306 data rows (including a leap year of 2016).

4) All data rows for which GHI is smaller than 0 are removed from the dataset. Negative irradiance is in conflict with the laws of physics and is caused by measuring errors. Negative GHI is removed to prevent a negative modelled DC power output.

5) The UPOT meteorological consists of three separate measurement of global POA irradiance. Although the means are quite similar (245.7, 248.8 and 247.4 W/m^2) the first is the most representative as the measurement was synchronised with the sensor measuring the DC power output of the PV module. Only this measurement of global POA irradiance is therefore used for modelling and not the mean of the three measurements.

Resampling

6) The UPOT data is down-sampled to 2, 15 and 60-minutes data intervals using the nearest resample method. This method prevents data from being smoothed when down-sampled, which would result in lower errors for larger timesteps. Smoothed data makes larger timesteps seem more accurate and thus leads to unfair comparison of data. The nearest resample method takes the value of the data point closest to the newly resampled data point. The right resampling method is crucial in comparing the accuracy of different timesteps. Even with the nearest resample method however, error losses for larger timesteps cannot be completely avoided. Resampling the UPOT data into the three different timesteps represents a data fork, meaning there are now four modelling paths (including the hourly KNMI data).

7) Air pressure and precipitation are not measured by the UPOT facility and is copied from the KNMI dataset. For the timesteps smaller than 60-minutes, air pressure and precipitation values are up-sampled using linear interpolation. Air pressure and precipitation have smaller influence on the PV performance than the earlier resampled solar irradiance, which justifies this method for up-sampling.

3.3.2 PVLib (Python)

Unlike the other simulation models, PVLib is not a stand-alone application but a library containing numerous models relevant for PV modelling. The library is accessible with either Python or MATLAB. Some models from the PVLib library are similar to models used in other simulation models or even originate from them. E.g. PVLib contains a cell temperature model from PVSyst and an IV curve model from PVWatts. Both examples are more elaborately discussed in the modelling steps below.

Solar Position

1) The first step is importing the PVLib modelling library into Python.

2) Before decomposing GHI, it is required to model the solar position. The solar position is modelled by the Ephemeris solar model and consists of the apparent elevation, apparent solar zenith, solar azimuth, elevation, equation of time and solar zenith. Modelling the solar position with this model requires the time index (the exact date and time of each data point), location (latitude, longitude and altitude) and the air pressure. The Ephemeris model came out slightly more accurate for modelling the global POA

irradiance (although the difference is negligible) than Spa$_{numba}$, Spa$_{numpy}$ and Pyephem. The accuracy results of the comparison between these solar sub-models are presented in Table 6 in section 4.1.

3) AOI of the solar irradiance is modelled using the surface tilt of the PV modules (37°), surface azimuth (180°), solar zenith and solar azimuth.

4) The relative airmass is modelled from the solar zenith with the 'Kastenyoung1989' model. Consequently, the absolute airmass AM is calculated using the relative airmass and air pressure. Kastenyoung1989 is the default model from seven possible airmass models. The choice of model however does not matter as they all have the same output.

Irradiance Decomposition

5) The next step is decomposing GHI into DNI and DHI, for which four separate irradiance decomposition models are compared: ERBS, DIRINT, DIRINDEX and DISC. Only ERBS has both DNI and DHI as output whereas the other models only have DNI as output. For these models DHI is then calculated according to Formula 1 of subsection 2.1.2.

   a) The required input variables for ERBS are GHI, time index and solar zenith. In order to prevent ERBS from calculating extreme high values of DNI, the input data for ERBS with AOI greater than 85° is removed. The reason behind these extreme values is that ERBS calculates DNI by dividing by cos(solar zenith). High values of AOI that remain after the previous filter condition (to remove data for which GHI is equal or lower than 0) cause the modelled DNI to skyrocket, which is the case for 979 2-min data points.

   b) The required input variables for DIRINT are GHI, time index, solar zenith and air pressure.

   c) The required input variables for DIRINDEX are GHI, clear sky GHI, clear sky DNI, time index, solar zenith, air pressure and air temperature. The clear sky GHI and clear sky DNI are determined from the extra-terrestrial irradiance with the assumptions that there is no cloud coverage and thus is equal to the maximum amount of GHI and DNI that can reach the ground. Both the clear sky GHI and clear sky DNI are modelled with the Simplified Solis clear sky model, which has a lower error (lower normalized RMSE) than the Ineichen clear sky model. The required input for the clear sky model is the location coordinates and the time index.

   d) The required input variables for DISC are GHI, time index, solar zenith and air pressure.

6) Before exporting the decomposed GHI to be used by the other simulation models several validity checks are carried out. The following conditions are verifying if no physical laws are in conflict with the modelled DNI and DHI.

   The checks are as followed:

   a) Confirming if DNI is always lower than the extra-terrestrial irradiance (DNI extra). DNI extra is modelled using the time index. This value depends on the distance between the sun and the earth and is therefore practically location independent on earth.

   b) Confirming if DHI is always equal to or lower than GHI.

   c) Confirming the absence of negative values of DNI and DHI.

7) The modelled DHI and DNI and measured GHI are exported together with the required meteorological data to model-specific csv files that can be imported by SAM, PVWatts and PVSyst. **From this point on, the modelling path of these models is split from the modelling path of PVLib.**

Irradiance Transposition

8) POA irradiance is now modelled using six separate irradiance transposition models: Perez, Hay & Davies, Klucher, Reindl, King and Isotropic irradiance transposition models, to compare which of these is the most accurate. Table 3 presents the input and output variables that correspond to the different models.

Table 3. In and output of PVLib irradiance transposition models.

Model	Perez	Hay/Davies	Isotropic	Klucher	Reindl	King
Input	- Surface Tilt - Surface Azimuth - GHI - DNI - DHI - DNI$_{extra}$ - Solar Zenith - Solar Azimuth - Relative Airmass	- Surface Tilt - Surface Azimuth - DNI - DHI - DNI$_{extra}$ - Solar Zenith - Solar Azimuth	- Surface Tilt - DHI	- Surface Tilt - Surface Azimuth - GHI - DHI - Solar Zenith - Solar Azimuth	- Surface Tilt - Surface Azimuth - GHI - DNI - DHI - Solar Zenith - Solar Azimuth	- Surface Tilt - GHI - DHI - Solar Zenith
Output	- POA sky diffuse - POA ground diffuse - POA direct	- POA sky diffuse	- POA sky diffuse	- POA sky diffuse	- POA sky diffuse	- POA sky diffuse

Only Perez directly models all three components of POA irradiance. For the other models direct and ground diffuse POA irradiance are modelled using the Beam Component and Get Ground Diffuse models. For the latter the default albedo (0.2) is used. The POA components are all relative to the orientation of the PV module and can therefore be simple summed to calculate the global POA irradiance.

$$POA_{global} = POA_{direct} + POA_{sky\ diffuse} + POA_{ground\ diffuse} \tag{10}$$

$$POA_{diffuse} = POA_{sky\ diffuse} + POA_{ground\ diffuse} \tag{11}$$

9) Now that the global POA irradiance is modelled, the different modelling options are compared with each other using the NRMSE in regard to the measured POA irradiance. The most accurate combination is chosen for further modelling PV performance.

10) This step marks the **second modelling starting point** as from this moment the measured GPOA$_{UPOT}$ irradiance is introduced for modelling. Similar as for the measured GHI, GPOA$_{UPOT}$ needs to be transposed into its separate components. Modelling the different POA component from GPOA$_{UPOT}$ requires the use of a different irradiance transposition model called GTI DIRINT. This model requires the GPOA$_{UPOT}$, AOI, solar zenith, solar azimuth, air pressure, time index, surface tilt, surface azimuth and the albedo factor as input variables. According to the author the model results in large errors for AOI greater than 80° (Bill Marion, 2015). Global POA irradiance with an AOI greater than 80° is thereby not transposed, but set to be equal for diffuse POA, resulting in a direct POA irradiance of zero. The output of GTI DIRINT are GHI, DHI and DNI.

11) Direct POA can now be modelled, using DNI, surface tilt, surface azimuth, solar zenith, and solar azimuth. Subsequently after modelling direct POA, diffuse POA is calculated using Formula 10 and 11.

12) Before modelling the effective POA irradiance a final check is done, verifying if direct POA is never greater than global POA. For these cases direct POA is set to zero and diffuse POA equal to global POA. This condition occurred 460 times for the 2-min dataset, which is equal to 0.13% of the dataset.

Modelling Effective Panel-of-Array Irradiance

13) Until now all modelling steps were independent of the type of PV module used (except for the orientation). Modelling effective POA irradiance includes modelling module specific losses that require various technical module parameters. The technical parameters of the PV modules used in this research are imported from the flash-test results. Some PVLib models however require more parameters than the flash-test provide, so any missing information is complemented with parameters from the CEC database. For the Sandia PV Array Performance Models (SAPM) additional parameters are required that are only found in the Sandia module database. Unfortunately, the PV modules in question are not incorporated in this database, and thus the additional parameters are taken from a similar PV module in the database. For all SAPM models below, the parameters from flash-tests and the CEC database are combined with the parameters from a comparable PV module from the Sandia database, unless otherwise specified. The parameters from the flash-tests and the CEC and Sandia module databases are presented in appendix A.1.

    Effective POA irradiance is calculated from the POA components by applying soiling, reflection, spectrum and shading losses. As shading losses are assumed to be zero, only three types of losses are modelled.

14) The first step in modelling effective POA irradiance is modelling AOI losses (reflection losses). PVLib has three models that can model AOI losses: SAPM, Ashrae and Physical.

    a) SAPM is the only one of these three modules requiring specific module parameters. Besides some module specific parameters, the model requires the modelled AOI.

    b) Ashrae only requires the modelled AOI as input. The default settings of the model (n=1.2526, K=4.0, L = 0.002) are used.

    c) Physical also only requires the modelled AOI as input. The default settings of the model (b=0.05) is used.

15) The second type of losses that influence effective POA irradiance are spectrum losses (spectral mismatch). Two PVLib models are suitable for modelling these losses: SAPM and the First Solar Spectrum Correction (FSSC) model.

    a) Aside from the Sandia module database parameters, the SAPM spectrum losses model only requires the absolute airmass. SAPM does not make any distinction between diffuse POA and direct POA irradiance in applying spectral losses.

    b) The FSSC model uses the CEC module database parameters for modelling spectrum losses. This model requires precipitation and the absolute airmass as input values.

16) Finally, the effective POA irradiance is calculated using Formula 12, which directly applies soling losses (King et al., 2004).

$$POA_{effective} = \left[\frac{f_{spectral}\left(POA_{direct} \cdot f_{aoi} + POA_{diffuse}\right)}{E_0}\right] \cdot SF \tag{12}$$

with,

$f_{spectral}$ = Fraction of spectral loss
f_{aoi} = Fraction of angle-of-incidence loss
SF = Soiling factor (0.99)
E_0 = Reference irradiance (1000 W/m^2)

Modelling IV Curve

17) The first step in modelling IV curves is modelling the cell temperature. For modelling the cell temperature two models are used: SAPM and PVSyst cell temperature models. The SAPM temperature model is used for the effective POA irradiance that is previously modelled using SAPM. The PVSyst temperature model is applied on the other modelled effective POA irradiance.

    a) The SAPM cell temperature model requires the global POA irradiance, wind speed, air temperature and a module setup specification as input. In this case the module setup is defined by 'open rack cell glass back', which corresponds to three required input parameters that are determined for various different module setups. These are the upper limit for module temperatures at low wind speeds and high solar irradiance (a = -3.47), the rate at which the module temperature drops as wind speed increases (b = -0.0594) and the temperature difference between the cell and the module back cover (dT = 3).

    b) The PVSyst cell temperature works with the same required input values but uses different temperature parameters. The default setup specified as 'freestanding' corresponds to a heat loss factor coefficient of 29 and a wind loss factor of 0. The latter assumes the wind does not influence the cell temperature. To include the effect of the wind, the wind loss factor is manually set to -0.0594, which is equal to the one of SAPM.

18) The IV curve is modelled with three different models: SAPM, Single Diode model (CEC), and the PVWatts DC model under PVLib. The SAPM model is used for the previously modelled values with SAPM and both the Single Diode model and PVWatts module (subsection 2.1.5) for the other modelled data.

    a) SAPM calculates five points on the IV curve: V_{oc}, I_{sc}, V_{mp}, I_{mp} and P_{mp}, requiring effective POA irradiance, cell temperature and the specific Sandia module parameters.

    b) The Single Diode model requires I_L, I_D, R_{SH}, R_S, n, N_S, V_{th} and solves the following Lambert W-function, which is a more elaborate version of the single diode equation presented in subsection 2.1.5 (Formula 3).

$$I = I_L - I_D \cdot \left[exp\left(\frac{V + I \cdot R_S}{n \cdot N_S \cdot V_{th}} \right) - 1 \right] - \frac{(V + I \cdot R_S)}{R_{SH}} \qquad (13)$$

with, n = Usual diode ideal factor

N_S = Number of cells in series

V_{th} = Cell thermal voltage under the desired IV curve conditions, which is Boltzmann's constant multiplied with the cell temperature and divided by the electron charge.

All the above parameters first need to be modelled using effective POA irradiance, cell temperature and the CEC module parameters, using the CEC model of Dobos (2012).

    c) The PVWatts DC model only requires effective POA irradiance, cell temperature, rated module power and the temperature coefficient of the module. The latter is obtained from the CEC module database. Contrary to SAPM and the Single Diode model, the PVWatt DC model does not model five point on the IV curve but only the DC power output P_{MP}.

19) At last, the assumed cabling and connection losses (1.0%), LID (1.0%) and MPPT losses (0.1%) are applied on all the different sets of modelled P_{MP}.

20) The modelled performance of all the different PVLib model combinations is exported to csv files, to be imported in the Python evaluation script that is used for determining accuracy and to compare all the different prediction models.

3.3.3 System Advisor Model (SAM)

Before running a simulation in SAM six main model configurations have to be setup. Each of the six different setup steps are described below.

Location and Resource

1) The previously exported SAM specific meteorological and irradiance csv files are imported. In total there are 12 different files: data for the years 2015, 2016 and 2017 each with 2-min, 15-min and 60-min timesteps for UPOT and for the 60-min timesteps for the KNMI data. The location coordinates and time zone are automatically imported from the files.

2) Monthly albedo factor values (0.2) are manually inserted and are assumed to be constant throughout the year.

3) For the sky diffusion model (irradiance transposition model) Perez is selected from three options. Perez is more accurate (lowest NRMSE and insolation error) than the Isotropic and the HDKR model. This statement is validated for the 2015 2-min UPOT dataset, for which the results are presented in Table 9 in section 4.1. The HDKR model is a combination of the irradiance transposition models of Hay/Davies, Klucher and Reindl.

4) SAM can model with either a combination of DNI and GHI or with global POA irradiance. Both options are separately run.

Module

5) The specific PV modules are selected from the CEC module database which, once selected, automatically fills in most of the fields of the PV module parameters.

6) The nominal operating cell temperature (NOCT) method for temperature correction is selected, with rack mounted modules as mounting standoff and an array height of a two-story building or higher.

Inverter

7) As it is not possible to model without selecting an inverter, a custom inverter is chosen for which the efficiency is set to a 100%. This ensures that no inverter losses are taken into account.

System Design

8) The number of modules per string and number of strings in parallel in the subarray are set to 1, as only 1 module is being modelled.

9) The module axis is set to fixed and its orientation parameters (surface tilt and azimuth) are inserted.

Shading and Layout

10) Shading losses are set to zero.

Losses

11) Cabling and connection losses (1.0%), LID (1.0%) and MPPT losses (0.1%) are applied. No further AC, transformer or transmission losses are selected.

12) The simulation is run and the results are exported to csv files that are later imported in the Python evaluation script.

3.3.4 PVWatts (SAM)

Running a simulation with PVWatts in the SAM desktop application consist of only 2 setup steps, which are described below.

Location and Resource

1) The same 12 input files used for SAM are imported. The location coordinates and time zone are automatically imported from the files.

2) No selection can be made in the type of irradiance transposition model nor in the albedo factors. PVWatts can only be run using GHI and DNI and not using global POA irradiance.

System Design

3) The rated power of the module (265 Wp) is inserted, as well as a custom inverter with a 100 % efficiency. Besides the rated power no specific module parameters of the PV modules are required.

4) Orientation information is passed (surface tilt and surface azimuth).

5) Cabling and connection losses (1.0%), LID (1.0%) and MPPT losses (0.1%) are again applied.

6) No additional shading options are applied.

7) The simulation is run and the results are exported to csv files that are later imported in the Python evaluation script.

3.3.5 PVSyst

A PVSyst simulation project consist of a project designation and setting up a system variant. The project designation requires a meteorological file to be selected. Files can either be imported from a custom source or be selected from a list of public weather stations. After importing and selecting the file, the system variant is configured according to the following steps.

Orientation

1. UPOT and KNMI 60-min timestep meteorological data and modelled DNI are imported for the years 2015, 2016 and 2017. PVSyst can only model using hourly timesteps.

2. The first part of a PVSyst simulation is defining the orientation of the solar array. The surface tilt and surface azimuth are inserted in the demanding fields. It is important to note that an array faced south has a surface azimuth of 0° in PVSyst, contrary to the other simulation models. Setting the surface azimuth to 180° leads to very low PV performance after simulating.

System

3. The number of modules is set to 1 by setting the number of modules in series and the number of strings to 1.

4. The PV modules available in the PVSyst module database are not identical to the UPOT PV modules. The most similar modules in the database are selected instead. The main difference is that these modules are polycrystalline instead of monocrystalline, meaning its performance is expected to be slightly lower.

5. A simulation in PVSyst cannot be completed without selecting an inverter. Unfortunately, a custom inverter cannot be configured with an efficiency of 100%, so an arbitrary inverter (Micro-0.25-I-OUTD-US-240) is selected from the database. The type of inverter does however not influence the required results as the output file includes the modelled DC power output before the inverter.

Detailed Losses

6. For defining thermal parameters, the mounting setup is defined as a free mounted module with air circulation. The corresponding field thermal loss factor and NOCT equivalent factor are left at its default values.

7. Ohmic losses in the DC circuit of the solar array are set to zero, as the circuit only consists of a single module. Internal circuit losses of the PV module are accounted for in the next step.

8. Cabling and connection losses (1.0%), LID (1.0%) and MPPT losses (0.1%) are applied.

9. Soiling losses are set to 1.0%.

10. Spectral correction (spectral losses) settings are left at its default configuration.

11. The (default) Ashrae model is selected for modelling AOI losses.

12. Losses due to ageing, unavailability, auxiliary equipment, and shading are set to zero.

13. Optional settings for advanced shading, self-consumption, energy storage, energy management are left unchecked.

14. The simulation is run and the results are exported to csv files that are later imported in the Python evaluation script.

3.3.6 Helioscope

Helioscope does not offer the ability to import meteorological files but relies on meteorological data from Meteonorm or IWEC data from various weather stations. Meteonorm uses satellite meteorological data (TMY) with a 10km x 10km grid scale resolution. The closest available IWEC weather station is located in Amsterdam. This makes it initially difficult to compare its results with the other simulation models as the input data is not equal for Helioscope. In addition, the meteorological data used from Meteonorm or IWEC is not given for a specific year but is given as an average over several years. Fortunately, the meteorological data used by Helioscope can be exported after running the model. The file, which consists of 60-min timesteps, is then imported and undergoes the same PVLib modelling steps (subsection 3.3.1) as the UPOT and KNMI data. This ensures the models can be fairly compared on its working principles without being distorted by different sets of input variables.

Designs

1. The first part of using Helioscope is choosing a geographical location from satellite imagery, which results in a top view of the area intended for the desired PV modelling project. For this research it means navigating to the Hans Freudenthal building and selecting its roof as a field segment.

2. After selecting the roof as field segment the type of PV module is selected from a list, which fortunately includes the UPOT PV modules. The racking is defined as fixed tilt racking. The surface tilt, surface azimuth and height of the building are inserted in the demanding fields. Additional setup options such as spacing, keep-out areas and inverter options are ignored. For this research only a single module is modelled, thus not requiring the advanced design setup for large solar arrays.

Conditions

3. After designing the solar array, the type of meteorological data is selected from two available options. The first one is to use Meteonorm meteorological data and the second to use IWEC data from the weather stations in Amsterdam. To compare, the model is run twice, each with a different data source.

4. A soiling loss of 1% is applied.

5. The Sandia cell temperature model is selected.

6. Mismatch settings are left unchanged and are kept at default settings.

7. Perez is selected as irradiance transposition model and the source of solar angle is selected to be equal to the data source (Meteonorm or IWEC data) and not with the coordinates of the Hans Freudenthal building.

Shading

8. No shading profiles are applied.

9. The simulation is run and the results are exported to csv files that are later imported in the Python evaluation script.

3.4 Machine Learning Models

All machine learning techniques used for predicting the PV performance are accessible through Python's machine learning library scikit-learn. Through scikit-learn these techniques are easily imported without the need of understanding the exact underlying mathematical algorithms. This section does not further elaborate on the methodology of the different machine learning techniques than section 2.3. They are simply used as 'black box' functions made possible by scikit-learn. The scikit-learn functions are called estimators and are further addressed as such. In total eight machine learning techniques are used for training prediction models.

For all different estimators GHI_{UPOT}, GHI_{KNMI} and $GPOA_{UPOT}$ are used as main independent input variables. All estimators train models for three different timesteps for GHI_{UPOT} and $GPOA_{UPOT}$ and for one 60-min timestep for GHI_{KNMI}. In total each estimator is thus used seven times to train a machine learning model. Additional independent variables are the air temperature, AOI, wind speed, air pressure, precipitation and the relative humidity. Only the simple linear regression estimator does not take these additional variables as it can only take a single variable. The measured DC power output of the UPOT PV modules is used as the dependent output variable. Both input and output data are used by the estimator for training and testing in order to build a model that best fits the available data. The data of 2015 is used for training and testing as this year contains the least amount of missing data points. The trained models are then used for predicting the DC power output using GHI_{UPOT}, GHI_{KNMI} and $GPOA_{UPOT}$ for the year 2016. The results are then compared with the measured DC power output of 2016 for validation, using the RMSE and NRMSE.

To compare fairly between the different estimators and to reproduce the results, the data is split into a training and testing set using a pseudo-random number of 42. This guarantees the split data is always equal and that the results remain the same when re-training a model. For all the estimators the data is split into five sets for which each set is used as both training and testing data with cross validation. The machine learning techniques and corresponding estimators used for training are presented in Table 4.

Table 4. Machine learning techniques and corresponding estimators.

Regression Technique	Estimator	Data Split	Input
Simple Linear Regression	Linear Regressor	Cross Validation	- GHI_{UPOT} - GHI_{KNMI} - $\text{GPOA}_{\text{UPOT}}$
Multiple Linear Regression	Linear Regressor		**UPOT & KNMI:** - GHI - Air Temperature - AOI - Wind Speed - Air Pressure - Precipitation - Relative Humidity **UPOT:** - GPOA
Polynomial Regression	Ridge Regressor		
K-Nearest Neighbour Regression	KNeighbours Regressor		
Decision Tree Regression	Decision Tree Regressor		
Ensemble - Bagging	Extra Forest Regressor		
Ensemble - Boosting	XGB Regressor		
Ensemble - Stacking	**Base Estimators:** - Extra Forest Regressor - Random Forest Regressor - XGB Regressor - Ada Boost Regressor **Meta-Estimator:** - KNeighbours Regressor		

Figure 16 presents a flowchart that explains how the different solar irradiance from UPOT and the KNMI is used as input for all eight machine learning models. Data pre-processing is the same as describes in section 3.3.1.

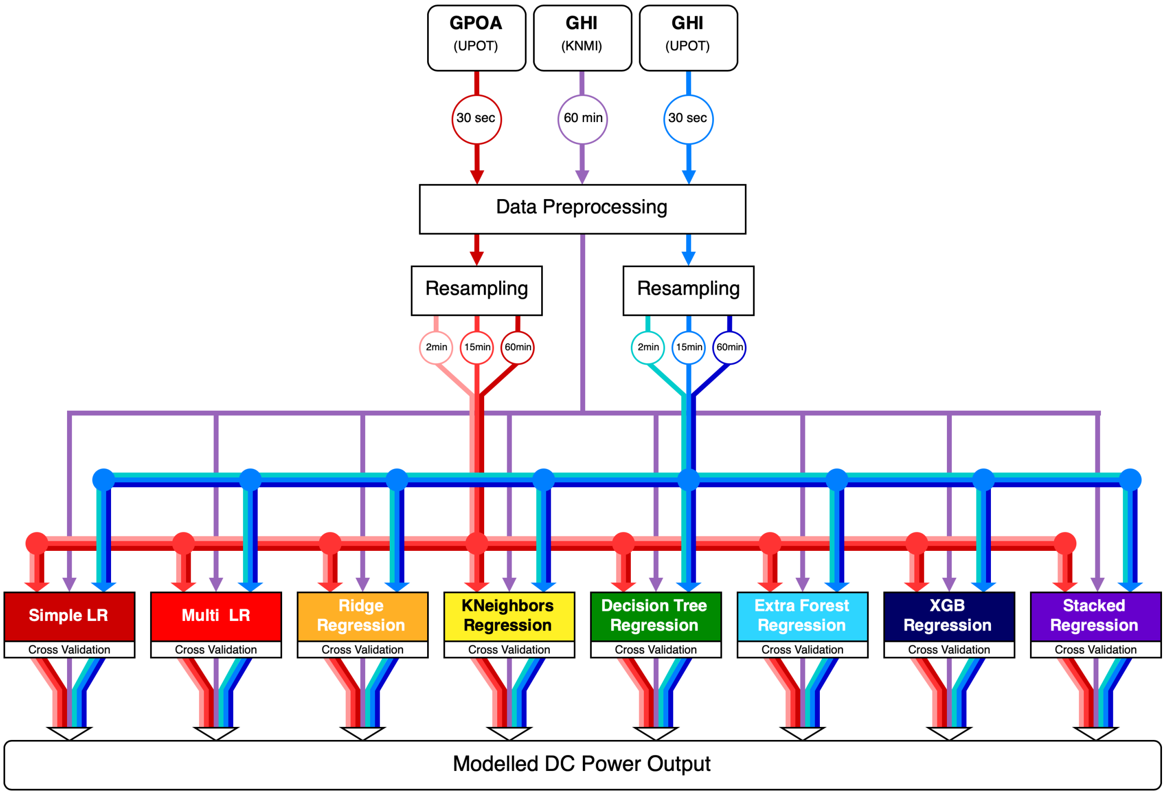


Figure 16. Flowchart of all eight machine learning regressors.

1. Simple Linear Regression
The first technique for predicting PV performance is simple linear regression using the linear regression estimator. Although the estimator can take multiple independent input variables, it is only given GHI_{UPOT}, GHI_{KNMI} and in turn $GPOA_{UPOT}$ to predict the DC power output.

2. Multiple Linear Regression
For multiple linear regression the same linear regression estimator is used, but now it is given multiple independent variables and thus the additional independent input variables are used.

3. Ridge Regression
The Ridge regression estimator is used as a polynomial regressor. The Ridge regressor parameter alpha is set to 2 as this value is found to be the most accurate. Alpha penalizes complex trained models preventing the estimator from overfitting.

4. K-Nearest Neighbours Regression
The KNeighbours regression estimator is used and the number of neighbours is set to 10. This number is retrieved from an elbow curve based on several model runs with the 60-min GHI_{UPOT} as primary irradiance input. The regressor for the elbow curve was only run for the 60-min timestep to limit processing time. The optimal K-value found in the elbow curve is used in training the KNeighbours regression model and is found in Figure A.1 in appendix A.2.

5. Decision Tree Regression
The basic decision tree regressor is used. The default settings are left untouched (no additional arguments are passed to the estimator).

6. Extra Forest Regression
The extra forest regressor is an estimator based on decision tree regression. The extra forest regression has a built-in bagging ensemble function, which is manually set to perform 100 estimations. This number is selected as a higher number of estimations does not lead to major improvements in accuracy and to limit computational processing time. The bagging module is expected to increase the accuracy substantially compared to a basis decision tree regression.

7. XGB Regression
As a boosting ensemble estimator, the XGB Regressor is used and the number of estimators is set to 100, for the same reasons as for the extra forest regression.

8. Stacked Regression
Four base-estimators and one meta-estimator are used for stacked regression. The base-estimators are a multi linear regressor, Ridge regressor, extra forest regressor and a K-Neighbours regressor. The meta-estimator is an XGB regressor. These estimators are chosen as they are fundamentally different in order to benefit from their different strengths. The number of estimators is again set to 100.

3.5 Sensitivity Analysis

Two sensitivity analyses are conducted: one for the most accurate simulation model and one for the most accurate machine learning model.

Simulation models
The first sensitivity analysis is conducted for the most accurate simulation model in order to assess the robustness of the confirmation as most accurate simulation model. Simulation models are generally optimised for a specific range of variables. When certain values outside the optimised range are used, the model can have an unrealistic output. Even small changes in input variables can lead to noteworthy changes in the output of a model. The model's results are tested on its sensitivity by changing the following input variables and parameters: albedo factor, air temperature, air pressure, relative humidity, wind speed and precipitation. Their values are modified based on scenarios of extreme low and high values. The scenarios and corresponding input variables are presented in Table 5.

Table 5. **Minimum and maximum scenarios for various UPOT input variables.**

Input variable	Minimum	Maximum
Albedo	Albedo factor 0 (no reflection).	Albedo factor is 1 (max. reflection).
Air Temperature	Average air temperature is 1°C lower.	Average air temperature is 1°C higher.
Air Pressure	Constant air pressure of 96,920 Pa (lowest measured)	Constant air pressure of 104,110 Pa (highest measured)
Wind Speed	No wind	Wind speed is twice as high
Relative Humidity	Constant relative humidity of 24.5% (lowest measured)	Constant relative humidity of 100% (highest measured)
Precipitation	No precipitation	Twice as much precipitation (when it falls)

Machine Learning models

The second sensitivity analysis is conducted on the most accurate machine learning model in order to assess the model's sensitivity to a smaller set of training data. Machine learning is only possible if there is a set of empirical data to learn from. This means that predictions cannot be made for specific PV modules, unless there is already some measured performance data. When choosing for machine learning methods, it is essential to know the amount of data needed to train an accurate model. A sensitivity analysis is therefore conducted that assesses both macro- and micro-accuracy for a changing size of training and testing data, in order to find the minimum amount of measured data required to train an accurate prediction model.

The analysis is done by varying the amount of input data used for training the model and plotting the resulting NRMSE and yield error as a function of the amount of input data. The NRMSE is plotted for GHI_{UPOT} and $GPOA_{UPOT}$ both for a 2-min and 60-min timesteps. The NRMSE is plotted for these timesteps to assess if the 2-min timestep more quickly leads to an increase in accuracy as it contains 30 times more data than the 60-min dataset. The time periods range from only 1 month of data to a full year of 12 months. This is done four times, each with a different starting month in order to assess the influence of the season in which the data is collected.

Chapter 4 | Results

This chapter is divided into four sections. The first section presents the macro- and micro-accuracies of the different sub-models. The second and third section of this chapter provide the macro- and micro-accuracy results for the different simulation and machine learning models. The final section is a sensitivity analysis of the most accurate simulation and machine learning model.

4.1 Sub-model Accuracies

Although only the most accurate sub-models are used for simulating the final results, the intermediary results presented in this section still present valuable information. The comparison results of sub-models are presented for modelling the solar position, irradiance decomposition and irradiance transposition. The highest accuracies in the tables are accentuated in bold green.

Solar Position Sub-Models

All four PVLib solar position sub-models are compared on their macro- and micro-accuracy. The solar position is modelled using location coordinates and the time index and is consequently used for irradiance decomposition and irradiance transposition. The effect of using a different solar position model on the transposed irradiance is essentially what is presented in Table 6. The results in the table include the total amount of modelled global POA irradiance, which is the aggregate of the years **2015, 2016 and 2017** (corrected for missing values).

The differences between the sub-models are very limited and negligible for PV modelling. The Ephemeris sub-model only has a 0.00002 lower NRMSE for the 2-min timesteps. In terms of macro-accuracy all solar position models have practically the same insolation error. This is sensible as the solar position is not a modelled approximation as it can be mathematically calculated.

Table 6. PVLib solar position sub-model accuracy for modelling global POA irradiance (2015-2017)

Solar Model	SPA numba	SPA numpy	Pyephem	Ephemeris	UPOT$_{\text{source}}$
Global POA irradiance for 2015, 2016 & 2017 [kWh/m^2]					
2-min $_{\text{res.}}$	3,164.32	3,164.32	3,164.30	3,164.23	**3,157.96**
15-min $_{\text{res.}}$	3,147.02	3,147.02	3,147.00	3,146.92	
60-min $_{\text{res.}}$	3,150.49	3,150.48	3,150.46	3,150.31	
Insolation Error (macro-accuracy)					
2-min $_{\text{res.}}$	0.20%	0.20%	0.20%	0.20%	**0.00%**
15-min $_{\text{res.}}$	-0.35%	-0.35%	-0.35%	-0.35%	
60-min $_{\text{res.}}$	-0.24%	-0.24%	-0.24%	-0.24%	
Normalised RMSE (micro-accuracy)					
2-min $_{\text{res.}}$	0.047821	0.047821	0.047820	**0.047818**	**0.00**
15-min $_{\text{res.}}$	0.045441	0.045441	0.045440	**0.045438**	
60-min $_{\text{res.}}$	0.043211	0.043211	**0.043209**	0.043218	

Irradiance Decomposition Sub-Models

A comparison is made between four irradiance decomposition sub-models, the results are presented in Table 7. For all the different timesteps Dirint is the most accurate irradiance decomposition sub-model on a micro-level for modelling global POA irradiance and has the lowest insolation error for the 15- and 60-min timestep data. Although Dirindex is considered an improved version of Dirint (Chain et al., 2002), it overall scores lower than Dirint. Dirindex only has a lower insolation error over the three-year time period for the 2-min timestep data.

Table 7. PVLib irradiance decomposition sub-model accuracy for modelling global POA irradiance (2015-2017)

Model	Erbs	Dirint	Dirindex	Disc	UPOT$_{source}$
Global POA irradiance for 2015, 2016 & 2017 [kWh/m^2]					
2-min res.	3,172	3,177	3,156	3,202	**3,158**
15-min res.	3,169	3,162	3,138	3,198	
60-min res.	3,180	3,164	3,145	3,209	
Insolation Error (macro-accuracy)					
2-min res.	0.45%	0.60%	**-0.08%**	1.40%	**0.00%**
15-min res.	0.36%	**0.11%**	-0.64%	1.28%	
60-min res.	0.70%	**0.20%**	-0.42%	1.61%	
Normalised RMSE (micro-accuracy)					
2-min res.	0.059	**0.048**	0.049	0.052	**0.00**
15-min res.	0.056	**0.044**	0.047	0.048	
60-min res.	0.051	**0.042**	0.045	0.044	

Irradiance Transposition Sub-Models

Six PVLib and three SAM irradiance transposition sub-models are compared on their accuracy. The results are presented in Tables 8 and 9. The most accurate PVLib irradiance transposition sub-model is Perez. Perez has the lowest NRMSE for all timesteps and the lowest insolation error for 15- and 60-min timesteps. Only the King and Klucher models have a slightly lower insolation error for 2-min timestep data.

Table 8. PVLib irradiance transposition model accuracy for modelling global POA irradiance (2015-2017)

Model	Perez	Hay/Davies	Isotropic	Klucher	Reindl	King	UPOT$_{source}$
Global POA irradiance for 2015, 2016 & 2017 [kWh/m^2]							
UPOT 2-min	3,177	3,104	3,004	3,143	3,115	3,163	**3,158**
UPOT 15-min	3,162	3,087	2,980	3,132	3,099	3,139	
UPOT 60-min	3,164	3,089	2,979	3,138	3,102	3,138	
Insolation Error (macro-accuracy)							
UPOT 2-min	0.60%	-1.72%	-4.88%	-0.47%	-1.36%	**0.15%**	**0.00%**
UPOT 15-min	**0.11%**	-2.26%	-5.63%	-0.81%	-1.87%	-0.61%	
UPOT 60-min	**0.20%**	-2.19%	-5.68%	-0.64%	-1.78%	-0.63%	
Normalised RMSE (micro-accuracy)							
UPOT 2-min	**0.048**	0.050	0.065	0.049	0.050	0.060	**0.00**
UPOT 15-min	**0.044**	0.046	0.064	0.047	0.046	0.057	
UPOT 60-min	**0.042**	0.043	0.062	0.045	0.043	0.055	

The same is true for the irradiance transposition sub-models of SAM. Table 9 illustrates that Perez has the lowest NRMSE and the lowest insolation error of the three irradiance transposition models. Even the combination of Hay/Davies, Klucher and Reindl (HDKR) in SAM is still not more accurate than using Perez.

Table 9. SAM irradiance transposition model accuracy for modelling global POA irradiance (2015)

Model	Perez	HDKR	Isotropic	$\text{UPOT}_{\text{source}}$
Global POA irradiance for 2015 [kWh/m^2]				
UPOT 2-min	1,174	1,171	1,129	**1,211**
Insolation Error (macro-accuracy)				
UPOT 2-min	**-3.02%**	-3.26%	-6.76%	**0.00%**
Normalised RMSE (micro-accuracy)				
UPOT 2-min	**0.110**	0.117	0.119	**0.00**

4.2 Simulation Model Accuracy

The accuracy results for all five simulation models are presented in Table 10 and Table 11. The highest accuracies for both GHI and global POA irradiance are accentuated in bold green. Table 10 presents the PVLib accuracies for four different combinations of sub-models. As explained in subsection 3.3.2. different sub-models are used for modelling AOI and spectral losses, cell temperature and the IV curve. The results of both tables include the electricity yield for the years 2015, 2016 and 2017 combined. Dividing by three does not present a yearly average as the dataset is incomplete.

Table 10. Final accuracies for modelling the DC power output for different combinations of PVLib sub-models, based on the combined data of 2015, 2016 and 2017.

Sub-Model	**PVLib**								**UPOT** source
Reflection Model	**SAPM**		**Physical**		**Ashrae**		**Ashrae**		
Spectral Mismatch Model			**FSSC**		**FSSC**		**PVWatts**		
Temperature Model			**PVSyst**		**PVSyst**		**PVSyst**		
IV Curve Model			**Single Diode**		**Single Diode**		**PVWatts**		
Combination nr.	**I**		**II**		**III**		**IV**		
Source	GHI	GPOA	GHI	GPOA	GHI	GPOA	GHI	GPOA	
Electricity Yield 2015, 2016 & 2017 [kWh/kWp]									
UPOT 2-min	2,845	2,903	2,901	2,956	2,919	2,957	2,821	2,859	**2,894**
UPOT 15-min	2,838	2,901	2,894	2,954	2,911	2,955	2,814	2,857	
UPOT 60-min	2,845	2,909	2,901	2,962	2,917	2,963	2,820	2,864	
KNMI $_{60-min}$	2,893	-	2,942	-	2,957	-	2,957	-	
Yield Error (macro-accuracy)									
UPOT 2-min	-1.69%	**0.33%**	**0.24%**	2.17%	0.86%	2.20%	-2.50%	-1.21%	**0**
UPOT 15-min	-1.93%	**0.24%**	**0.01%**	2.09%	0.59%	2.12%	-2.77%	-1.28%	
UPOT 60-min	-1.69%	**0.53%**	**0.26%**	2.37%	0.81%	2.40%	-2.56%	-1.01%	
KNMI $_{60-min}$	**-0.03%**	-	1.69%	-	2.20%	-	2.20%	-	
Normalised RMSE (micro-accuracy)									
UPOT 2-min	0.076	0.061	0.076	**0.060**	**0.074**	**0.060**	0.078	0.066	**0**
UPOT 15-min	0.071	0.055	0.071	**0.054**	**0.069**	**0.054**	0.074	0.060	
UPOT 60-min	0.073	0.056	0.073	**0.055**	**0.071**	**0.055**	0.076	0.061	
KNMI $_{60-min}$	**0.181**	-	0.182	-	0.182	-	0.182	-	
RMSE (micro-accuracy)									
UPOT 2-min	17.79	16.92	17.68	16.86	**17.52**	16.85	17.63	**16.82**	**0**
UPOT 15-min	16.20	15.04	16.06	14.94	**15.88**	**14.93**	16.17	15.05	
UPOT 60-min	16.10	14.88	15.96	14.78	**15.76**	**14.77**	16.11	14.93	
KNMI $_{60-min}$	26.842	-	**26.835**	-	26.836	-	26.850	-	

Table 10 shows that combinations I, II and III have about the same NRMSE for GHI and global POA irradiance. The RMSE's of combination III are however slightly more accurate than those of the other combinations. The combinations show more differences when looking at the yield error. Combination II is overall most accurate in predicting the total yield from GHI$_{UPOT}$, which has a yield error of 0.26% or less for each timestep. Combination I, however, has the lowest yield error for GHI$_{KNMI}$ and GPOA$_{UPOT}$. Overall every combination has a yield error under 3%. When comparing between different timesteps for UPOT the NRMSE is the lowest for the 15- and 60-min timesteps.

There is no absolute winner among the four different combinations. It seems that combinations II & III are the most accurate for GHI$_{\text{UPOT}}$ and combination I (SAPM) for GPOA$_{\text{UPOT}}$ and GHI$_{\text{KNMI}}$.

Table 11. Final accuracies for modelling the DC power output for SAM, PVWatts, PVSyst and Helioscope, based on the combined data of 2015, 2016 and 2017.

Model	SAM		PVWatts	PVSyst	Helioscope		UPOT source
Source	GHI	GPOA	GHI	GHI	Meteonorm	IWEC	
Electricity Yield 2015, 2016 & 2017 [kWh/kWp]							
UPOT 2-min	2,730	2,715	2,748		2,858	2,837	**2,894**
UPOT 15-min	2,719	2,714	2,734				
UPOT 60-min	2,634	2,520	2,574	2,732			
KNMI $_{\text{60-min}}$	2,614	-	2,468	2,912			
Yield Error (macro-accuracy)							
UPOT 2-min	-5.66%	-6.17%	-5.05%	-	**-1.25%**	**-1.97%**	**0**
UPOT 15-min	-6.03%	-6.20%	-5.52%	-			
UPOT 60-min	-8.97%	-12.93%	-11.04%	-5.60%			
KNMI $_{\text{60-min}}$	-9.67%	-	-14.71%	**0.62%**			
Normalised RMSE (micro-accuracy)							
UPOT 2-min	**0.097**	**0.084**	0.117	-	0.388	0.368	**0**
UPOT 15-min	**0.093**	**0.079**	0.111	-			
UPOT 60-min	0.131	**0.117**	0.170	**0.083**			
KNMI $_{\text{60-min}}$	0.199	-	0.220	**0.190**			
RMSE (micro-accuracy)							
UPOT 2-min	**18.94**	**18.49**	20.88	-	49.81	48.53	**0**
UPOT 15-min	**17.52**	**16.86**	19.33	-			
UPOT 60-min	21.25	**23.51**	25.87	**16.24**			
KNMI $_{\text{60-min}}$	28.54	-	30.41	**27.41**			

According to table 11 all four simulation models have a lower micro-accuracy than all four sub-model combinations of PVLib. The model with the lowest micro-accuracy in Table 10 for GHI$_{\text{UPOT}}$ and GHI$_{\text{KNMI}}$ is PVSyst. The only model from Table 11 capable of modelling from GPOA$_{\text{UPOT}}$ is SAM, it however scores lower than all PVLib combinations both in macro- and micro-accuracy.

Contrary to the PVLib model combinations, all simulation models from Table 11 have negative yield errors (except for GHI$_{\text{KNMI}}$ from PVSyst). It seems that all these models underestimate the performance of the UPOT PV modules and increases with the timesteps. For PVSyst this could be explained as the modelled PV module is a polycrystalline and not a monocrystalline solar cell. For SAM, PVWatts and PVSyst these errors are of significant magnitude in the range of -5.02% to -14.71%. Remarkably Helioscope has a higher macro-accuracy, even though it uses different meteorological data. It is still capable of accurately modelling the electricity yield over the years 2015, 2016 and 2017, with yield errors of only -1.25% and -1.97% using Meteonorm and IWEC respectively. In this regard, Helioscope is more accurate in predicting the electricity yield over several years than SAM, PVWatts and PVSyst.

The macro-accuracy of Helioscope however looks different when modelling a single year. Table 12 presents the accuracy results when using the IWEC meteorological data of a single year from Helioscope as input for PVLib. The PVLib results are the average of all four sub-model combinations of Table 10 and are compared with the measured DC power output of 2015.

Table 12. Comparing Helioscope and PVLib with normalised meteorological data.

Simulation Model	Helioscope	PVLib
Data Source	IWEC	
Electricity Yield 2015	1,075 kWh/kWp	1,066 kWh/kWp
Yield Error (macro-accuracy)	-6.97%	-7.75%
Normalised RMSE (micro-accuracy)	0.377	0.19
RMSE (micro-accuracy)	50.44	27.01

According to Table 12, the NRMSE and RMSE of Helioscope is almost double that of PVLib and has a yield error equal to those of the simulation models in Table 11. These numbers clearly demonstrate that Helioscope remains a relatively inaccurate model on the micro-level when accounting for different meteorological input data and that its macro-accuracy is only high for modelling an average year.

Although there is no clear winner of the most accurate PVLib sub-model combination, all its combinations have higher macro- and micro-accuracies than the other stimulations models. **PVLib** is therefore considered the most accurate simulation model based on both macro- and micro-accuracy results.

Microlevel Prediction

To understand what is happening with the accuracy of the simulation models on a microlevel, six graphs (Figure 17 a through f) are presented below, to illustrate the difference between the predictions made for a sunny day with a clear sky and for a sunny day with passing clouds. All graphs are based on GHI_{UPOT} in order to plot the data of the various simulation models in the same graph. The PVLib curve is constructed using the modelled DC power output of combination III (Table 10), as this sub-model combination is the most accurate using GHI.

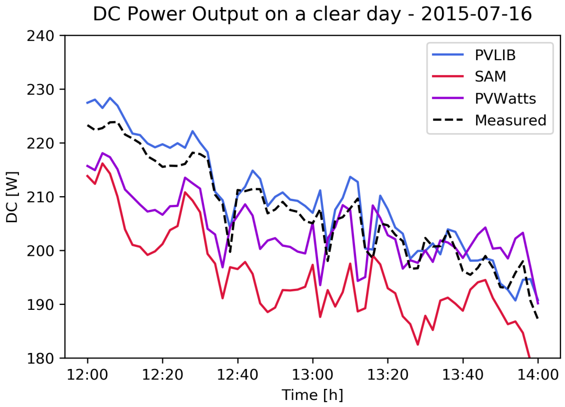


Figure 17a. Modelled vs. measured DC power output on a summer day with clear sky (2-min timestep).

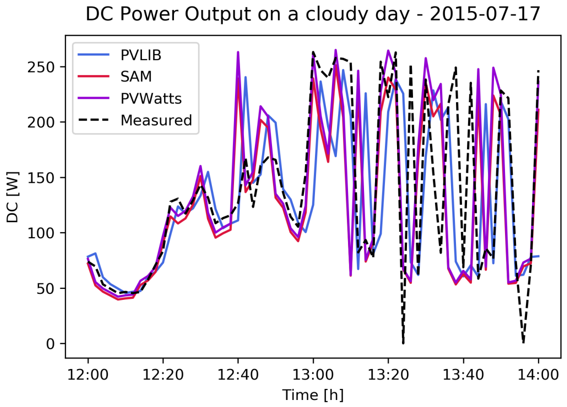


Figure 17b. Modelled vs. measured DC power output on a cloudy summer day (2-min timestep).

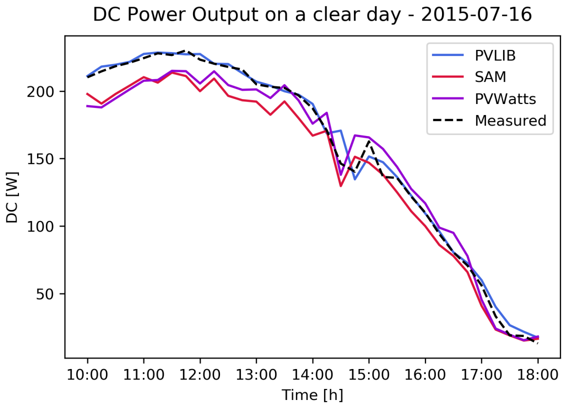


Figure 17c. Modelled vs. measured DC power output on a summer day with clear sky (15-min timestep).

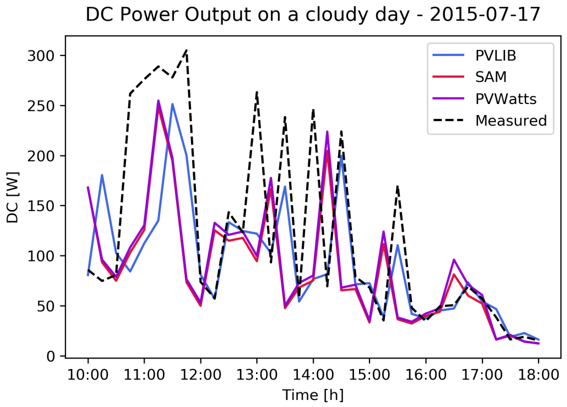


Figure 17d. Modelled vs. measured DC power output on a cloudy summer day (15-min timestep).

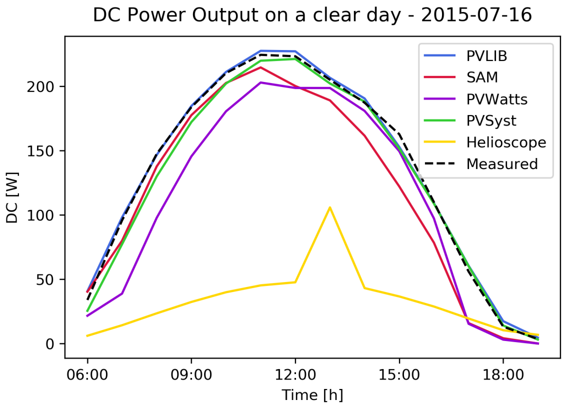


Figure 17e. Modelled vs. measured DC power output on a summer day with clear sky (60-min timestep).

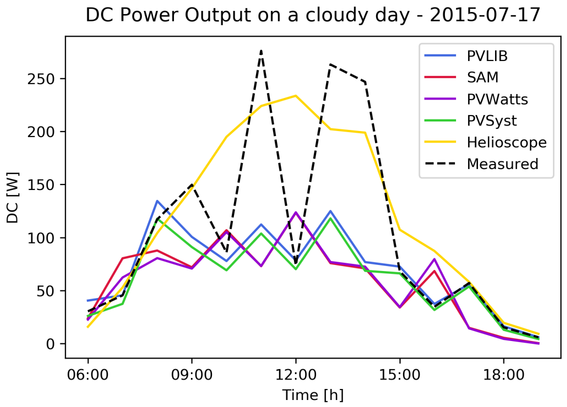


Figure 17f. Modelled vs. measured DC power output on a cloudy summer day (60-min timestep).

Figure 17a through 17f illustrate PVLib to be most aligned with the actual measured DC power output, with PVSyst following closely (for the 60-min timestep). When looking at the figures it appears that SAM and PVWatts generally have the same trend and respond similarly to incoming solar irradiance. Both model curves are regularly below the measured DC curve, meaning they generally predict a DC power output lower than the actual power output. These observations are aligned with the negative yield errors of SAM and PVWatts in Table 11. Helioscope is clearly the least aligned, which is a logical observation based on the difference in meteorological input data. For all six plots the RMSE and NRMSE are presented in Table 13. Important to note is that these metrics are calculated for all timesteps with the same time period, as in Figure 17 e & f, in order to make a fair comparison.

Table 13. Micro-accuracy comparison for clear sky and cloudy day between different timesteps.

Weather	Clear Sky		Cloudy Sky	
	RMSE	**NRMSE**	**RMSE**	**NRMSE**
UPOT 2-min	5.43	0.0417	51.35	0.3193
UPOT 15-min	4.54	0.0340	57.00	0.3422
UPOT 60-min	4.79	0.0415	72.12	0.4202

Table 13 shows that the micro-accuracy for modelling a day with relatively stable solar irradiance is much higher than for a day where passing clouds disturb solar irradiance frequently. In addition, it seems that 60-min timestep data has the hardest time dealing with these fluctuations, whereas the smallest 2-min timestep is more responsive due to its smaller timestep. The clear sky results for the different timesteps are comparable to the results found for the simulation models in section 4.2, for which the 15-min timestep has the highest micro-accuracy. It seems that the 15-minute timestep in on average best capable of dealing with both insolation scenarios, which could explain why this model has the highest micro-accuracy for all simulation models.

4.3 Machine Learning Model Accuracy

Tables 14a and 14b present the accuracy results of the eight machine learning models for predicting the DC power output of 2016. All models are trained using meteorological and PV performance data of 2015. The highest accuracies for both GHI and global POA irradiance are once again accentuated in bold green.

Table 14a. Accuracy machine learning models for prediction PV performance for the year 2016.

Machine Learning Model	**Simple LR**		**Multi LR**		**Ridge Regression**		**KNeighbours Regression**		**UPOT** **source**
Ensemble	-		-		-		-		
Source	GHI	GPOA	GHI	GPOA	GHI	GPOA	GHI	GPOA	
Electricity Yield 2016 [kWh/kWp]									
UPOT 2-min	1,132	1,134	1,109	1,128	1,109	1,128	1,121	1,132	**1,119**
UPOT 15-min	1,155	1,160	1,133	1,152	1,133	1,152	1,158	1,157	**1,142**
UPOT 60-min	1,155	1,156	1,137	1,151	1,137	1,151	1,149	1,143	**1,144**
KNMI $_{60\text{-min}}$	1,146	1,181	1,113	1,101	1,113	1,101	1,127	1,121	**1,142**
Yield Error (macro-accuracy)									
UPOT 2-min	1.12%	1.33%	-0.91%	0.77%	-0.91%	0.77%	**0.18%**	1.10%	**0**
UPOT 15-min	1.17%	1.57%	-0.76%	**0.91%**	-0.76%	**0.91%**	1.44%	1.32%	
UPOT 60-min	0.91%	1.06%	-0.60%	0.58%	-0.60%	0.58%	0.39%	**-0.08%**	
KNMI $_{60\text{-min}}$	**0.17%**	-	-2.70%	-	-2.69%	-	-1.47%	-	
Normalised RMSE (micro-accuracy)									
UPOT 2-min	0.259	0.125	0.258	0.124	0.258	0.124	0.245	0.135	**0**
UPOT 15-min	0.236	0.109	0.239	0.109	0.239	0.109	0.224	0.115	
UPOT 60-min	0.220	0.103	0.229	0.105	0.229	0.105	0.220	0.114	
KNMI $_{60\text{-min}}$	0.387	-	0.375	-	0.375	-	0.401	-	
RMSE (micro-accuracy)									
UPOT 2-min	33.4	24.2	31.7	23.9	31.7	23.9	33.5	25.2	**0**
UPOT 15-min	30.2	20.5	28.5	20.2	28.5	20.2	29.3	20.6	
UPOT 60-min	29.0	19.8	27.7	19.6	27.7	19.6	28.7	20.2	
KNMI $_{60\text{-min}}$	42.9	-	41.3	-	41.3	-	44.5	-	

Table 14b. **Accuracy machine learning models for prediction PV performance for the year 2016.**

Machine Learning Model	Decision Tree Regression		Extra Forest Regressor		XGB Regressor		Stacked Regressors *		UPOT source
Ensemble	-		Bagging		Boosting		Stacking		
Source	GHI	GPOA	GHI	GPOA	GHI	GPOA	GHI	GPOA	
Electricity Yield 2016 [kWh/kWp]									
UPOT 2-min	1,086	1,109	1,094	1,121	1,129	1,134	1,094	1,121	**1,192**
UPOT 15-min	1,160	1,153	1,149	1,154	1,151	1,153	1,147	1,154	**1,142**
UPOT 60-min	1,145	1,160	1,147	1,153	1,146	1,150	1,143	1,153	**1,144**
KNMI 60-min	1,137	-	1,134	-	1,111	-	1,132	-	**1,142**
Yield Error (macro-accuracy)									
UPOT 2-min	-2.92%	-0.93%	-2.30%	**0.21%**	0.91%	1.32%	-2.23%	0.22%	**0**
UPOT 15-min	1.58%	0.98%	**0.63%**	1.13%	0.84%	0.96%	0.65%	1.10%	
UPOT 60-min	**0.06%**	1.36%	0.29%	0.76%	0.19%	0.52%	0.07%	0.81%	
KNMI 60-min	-0.58%	-	-0.87%	-	-2.89%	-	-0.95%	-	
Normalised RMSE (micro-accuracy)									
UPOT 2-min	0.282	0.195	0.199	0.125	**0.198**	**0.107**	0.199	0.125	**0**
UPOT 15-min	0.215	0.142	0.166	0.103	0.179	**0.095**	**0.166**	0.103	
UPOT 60-min	0.214	0.145	0.158	0.102	0.170	**0.097**	**0.158**	0.102	
KNMI 60-min	0.456	-	0.357	-	0.360	-	**0.356**	-	
RMSE (micro-accuracy)									
UPOT 2-min	43.9	38.4	28.6	24.2	**27.6**	**23.2**	28.8	24.2	**0**
UPOT 15-min	34.0	28.7	**24.3**	20.5	**24.3**	**19.5**	**24.3**	20.5	
UPOT 60-min	35.0	30.6	24.0	20.0	**23.8**	**19.2**	24.0	19.9	
KNMI 60-min	55.4	-	41.5	-	**40.8**	-	41.3	-	

Based on RMSE, the machine learning model with the highest micro-accuracy for GHI_{UPOT}, GHI_{KNMI} and $GPOA_{UPOT}$ for all timesteps is the XGB regression model. For both GHI_{UPOT} and GHI_{KNMI} the stacked regressors result in the highest micro-accuracy. The NRMSE of the multiple regressor and extra forest regressor are only slightly lower for GHI_{KNMI} and for GHI_{UPOT} for the 15-min and 60-min timestep data.

Yield error results are overall similar to those of PVLib where all machine learning models have a yield error lower than 3%. There is no single machine learning model with the highest macro-accuracy as this greatly depends on the timestep, data source and irradiance input. It appears to be arbitrary which model has the lowest yield error, as the highest accuracy various over six different machine learning models. Similar to the simulation models, all machine learning models have the highest micro-accuracies for the 15- and 60-min timesteps and by far the lowest micro-accuracy for GHI_{KNMI}.

The simple linear regression model is overall more accurate than the multiple linear regression model. Adding more variables thus does not necessary lead to a higher accuracy. The multi linear and polynomials regression models score almost identical and only differ in output starting from four decimals (not shown in Table14a). The results show that all three ensemble methods are substantially more accurate than the other five methods on a microlevel, especially when modelling from GHI.

Although the macro-accuracy results do not differ that much, the **XGB regressor** is considered to be the most accurate machine learning model primarily based on its micro-accuracy results.

4.4 Sensitivity Analysis

4.4.1 Variable and Parameter Sensitivity

The most accurate simulation model identified is PVLib. For this model six different input variables are changed according to the minimum and maximum scenarios defined in section 3.5. The sensitivity on the yield error and NRMSE is only assessed for GHI$_{UPOT}$ 2-min timestep data for the year 2015. Because the performance is modelled from GHI, PVLib sub-model combination I (see Table 10) is used for this sensitivity analysis. The sensitivity results of the albedo factor and air temperature are plotted as a function of different monthly time periods presented in Figure 18 and 19. The sensitivity results of the four remaining variables are presented in Table 14. The variables in Table 14 all have minimal impact on the results and are therefore not as elaborately illustrated as the albedo factor and the air temperature.

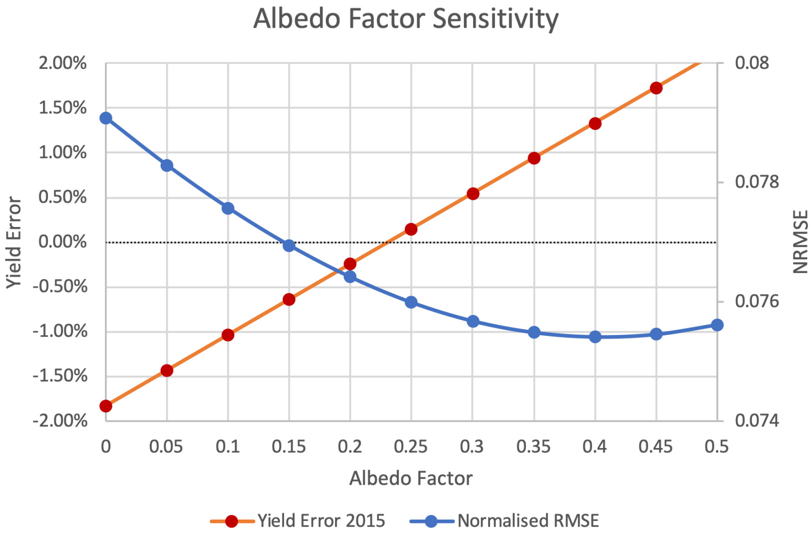


Figure 18. Sensitivity of yield error and NRMSE to a changing albedo factor based on GHI$_{UPOT}$ 2-min timestep data of 2015

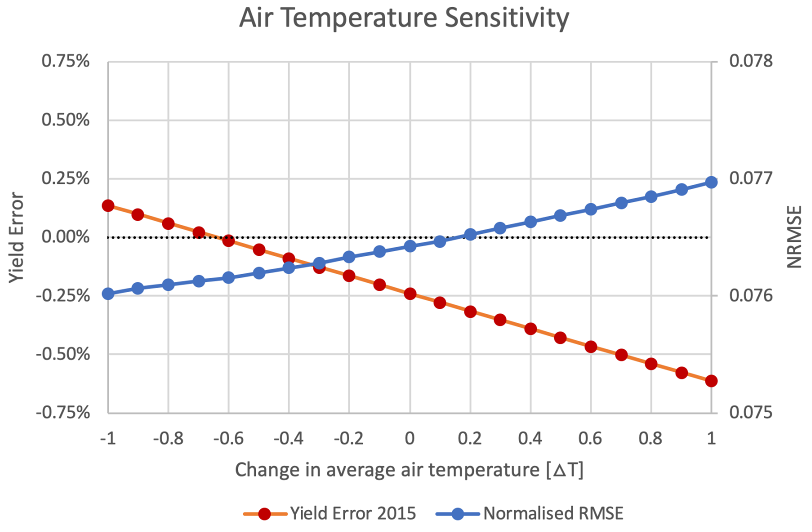


Figure 19. Sensitivity of yield error and NRMSE to changing average air temperature based on GHI$_{UPOT}$ 2-min timestep data of 2015

Figure 18 illustrates that the yield error is positively linearly correlated with the albedo factor and intersect the y-axis for an albedo factor of approximately 0.23. These results are logical, as the amount of incoming reflective irradiance increases with an increasing albedo factor. The NRMSE on the other hand shows a curvilinear relationship with a change in the albedo factor, reaching its lowest point for an albedo factor of approximately 0.4. These results verify the observation that a model has different optimum points for its lowest macro-accuracy and for its lowest micro-accuracy.

Figure 19 presents a similar result for which the air temperature has a negative linear correlation with the yield error and the NRMSE has a curvilinear relation with the change of average air temperature. The yield decreases for a rising air temperature as the efficiency is negatively affected. The yield error intersects the y-axis for a change in average air temperature of approximately -0.6. The NRMSE does not change considerably for a difference in average air temperature.

Table 15. Sensitivity of yield error and NRMSE to changing input variables based on GHI$_{UPOT}$ 2-min timestep data of 2015.

Variable	Minimum		Reference	Maximum
Air Pressure	Yield Error:	-0.55%	-0.28%	0.17%
	NRMSE:	0.0771	0.0765	0.0763
Wind Speed	Error:	0.24%	-0.28%	-0.31%
	NRMSE:	0.0764	0.0765	0.0766
Relative Humidity	Yield Error:	-0.28%	-0.28%	-0.28%
	NRMSE:	0.0765	0.0765	0.0765
Precipitation	Yield Error:	-0.28%	-0.28%	-0.28%
	NRMSE:	0.0765	0.0765	0.0765

According to Table 15, different input values for air pressure, wind speed, relative humidity and precipitation have a minimal effect on both macro- and micro-accuracies. Only the air pressure scenarios result in a small difference.

If all the above scenarios are applied in the most unfavourable direction than this leads to an NRMSE of 0.0817 and a yield error of -2.89%, for which PVLib is still considered to be the most accurate simulation model. The confirmation of PVLib as most accurate simulation model thus holds including the insights from the sensitivity analysis.

4.4.2 Time Period Sensitivity

A sensitivity analysis is conducted on the most accurate machine learning model, identified to be the XGB regressor. This analysis investigates the effects of a changing time period of data on the NRMSE and yield error for 2-min and 60-min timestep data and for both GHI$_{\text{UPOT}}$ and GPOA$_{\text{UPOT}}$ input irradiance. The XGB regressor has trained a model to predict the PV performance of 2016 using 2015 for each setup. The results are presented below in Figures 20a-f with the legend indicating the different starting months.

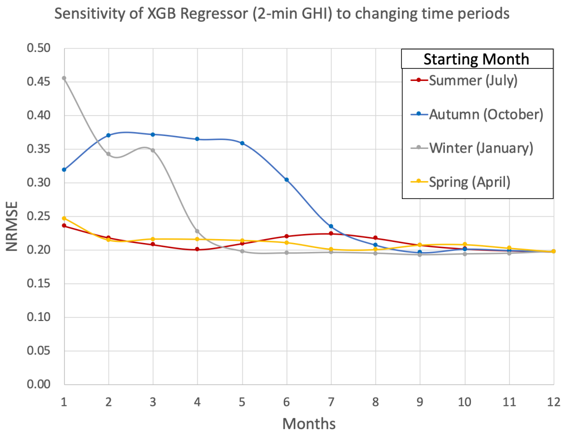


Figure 20a. NRMSE of predicted PV performance for 2016 for different time period using an XGB regressor and 2-min GHI$_{\text{UPOT}}$.

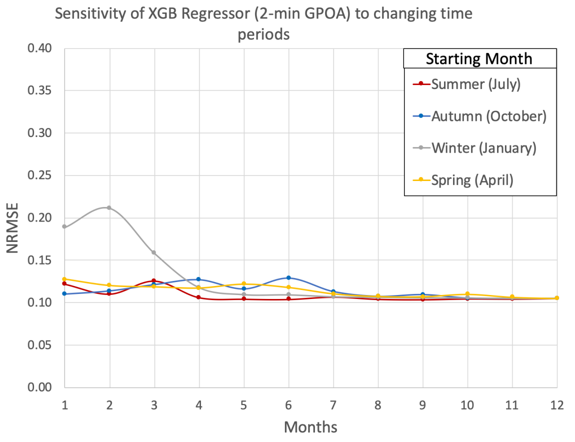


Figure 20b. NRMSE of predicted PV performance for 2016 for different time period using an XGB regressor and 2-min GPOA$_{\text{UPOT}}$.

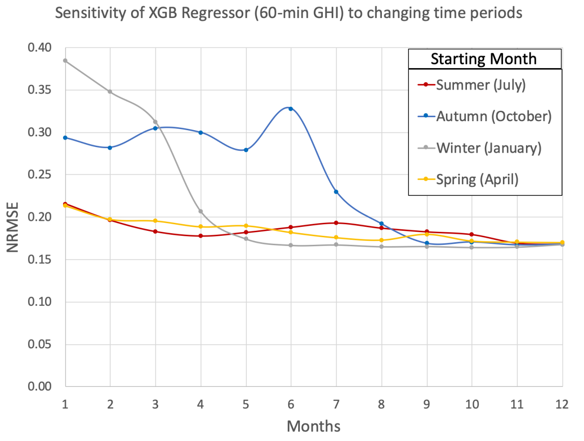


Figure 20c. NRMSE of predicted PV performance for 2016 for different time period using an XGB regressor and 60-min GHI$_{\text{UPOT}}$.

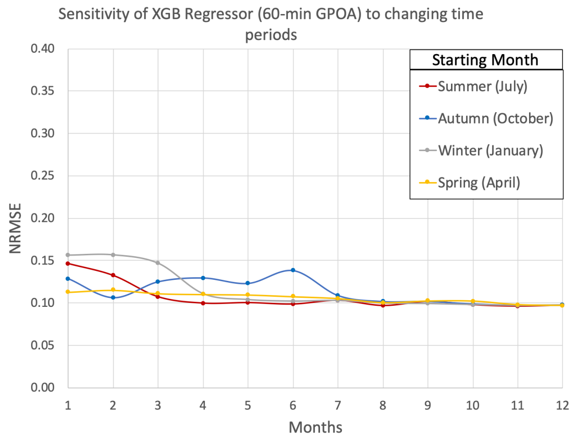


Figure 20d. NRMSE of predicted PV performance for 2016 for different time period using an XGB regressor and 60-min GPOA$_{\text{UPOT}}$.

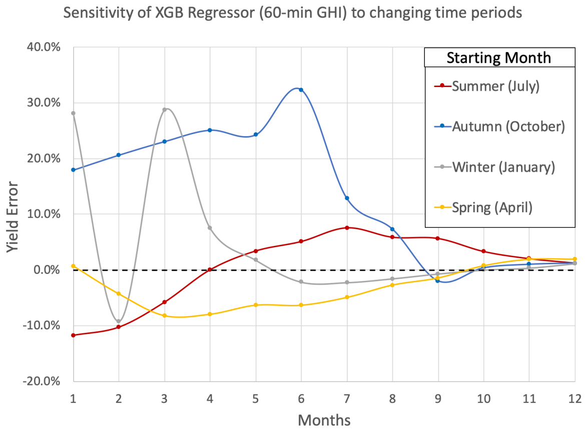


Figure 20e. Yield error of predicted PV performance for 2016 for different time period using an XGB regressor and 60-min GHI$_{\text{UPOT}}$.

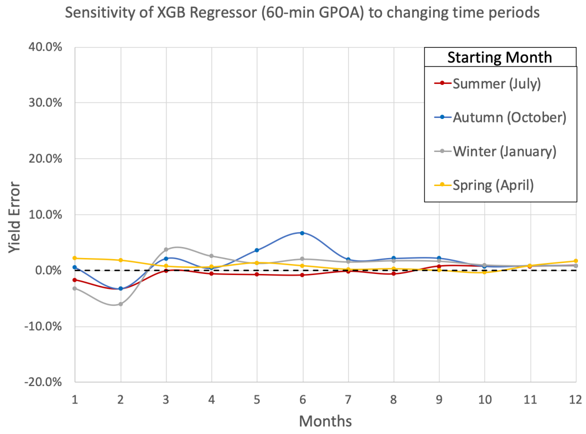


Figure 20f. Yield error of predicted PV performance for 2016 for different time period using an XGB regressor and 60-min GPOA$_{\text{UPOT}}$.

Figures 20a and 20c illustrate that starting with data from either the summer or the spring leads to the highest initial accuracy modelling from GHI. Autumn and winter months initially result in relatively low accuracies and only reach accuracies comparable to summer and spring after 5 and 8 months respectively. Figure 20b and 20d support this statement for global POA irradiance for the winter months, although the NRMSE is initially much lower for all months.

According to Figures 20a and 20b using the 2-min timestep data does not result in noteworthy faster saturation of NRMSE. All four curves' marginal increase in accuracy diminish after the same numbers of months for both 2-min and 60-min timestep data. It appears that it is more important for a dataset to span several different months than having more data from a limited number of months.

Figure 20e shows that the yield error for GHI is highly unstable for the first measuring months. The winter curve is the first to reach a stable yield error and the summer curve the latest. For global POA irradiance in Figure 20f the initial yield errors are considerably lower and initially acceptably low and stable for the spring curve.

Having a dataset spanning a minimum of 8 months guarantees a relatively high micro-accuracy using GHI, after which the marginal decrease in NRMSE is greatly reduced for all months. A stable yield error is however only obtained after 9 months for the winter, spring and autumn curves and only after 11 months for the summer curve. For GPOA this state of saturation is reached after only 7 months for both the NRMSE and yield error. For the summer and spring curve the yield error is already saturated after 3 months.

Using a spring month as initial measuring month overall seems to result in the lowest initial NRMSE and yield error and has overall the best accuracy for the first 4 months. After 4 months the winter curve has a lower and more stable NRMSE. Acceptable accuracy results (both NRMSE and yield error) can thus already be obtained after 1 months for global POA irradiance when choosing April as starting month. The same is true for GHI but in the months after that, there is still some noteworthy increase in accuracy.

4.5 Summary

Tables 16 and 17 compare the simulation models with the machine learning models for both macro- and micro-accuracies. Only the 3 most accurate prediction models are included in these tables.

Table 16. Top 3 prediction models with the highest macro-accuracies.

#	GHI_{UPOT}	$\text{GPOA}_{\text{UPOT}}$	GHI_{KNMI}
1	PVLib	PVLib	PVLib
2	All Machine Learning Models	All Machine Learning Models	All Machine Learning Models
3	Helioscope	Helioscope	Helioscope

Table 17. Top 3 prediction models with the highest micro-accuracies.

#	GHI_{UPOT}	$\text{GPOA}_{\text{UPOT}}$	GHI_{KNMI}
1	PVLib	PVLib	PVLib
2	PVSyst	SAM	PVSyst
3	SAM	XGB Regressor	SAM

Overall it can be stated that modelling with PVLib is the most accurate in predicting PV performance both on macro- and microlevel. All four PVLib sub-model combinations were able to most accurately predict the DC power output for both sources of data, both types of solar irradiance and for all three timesteps.

The second most accurate prediction model on a macrolevel is the use of machine learning models. There is no consistent machine learning model with the highest macro-accuracy, but all score substantially better than SAM, PVWatts, PVSyst and Helioscope. The prediction model with the third highest macro-accuracy is Helioscope. It is important to note that this is only the case for predicting the yield of an average year or over several years (>2). Helioscope is not able to model the PV performance for specific years, nor from specific custom irradiance.

The second most accurate prediction model on a microlevel when modelling from GHI is PVSyst, followed third by SAM. SAM is only placed second for modelling from GPOA, but that is because it is the only simulation model after PVLib that is able to do so. Straightforwardly this means that the most accurate machine learning model (XGB regressor) ends up at number 3.

Discussion

This study was conducted to provide a clear and comprehensive assessment of the performance of various simulation and machine learning models, used for prediction PV performance, and to stimulate further integration of both techniques in the field of PV assessment. The first part of this discussion section examines what the implications of the results mean for these two objectives and how they advance the scientific domain of PV modelling. These implications are however based on a whole construct of assumptions and concessions and therefore have their limitations. The second section subsequently discusses these limitations and to what extend they influenced the results. The final section advocates for additional research to further advance the integration of simulation and machine learning into PV modelling.

Recommendations

The results of this comparative study are meant to inform on the performance of the available prediction models and to form a clear overview that can help in selecting the most suitable prediction model. This section further elaborates what the accuracy results mean for the practical application of the investigated prediction models.

To examine the practical application of the results, a deliberate distinction was made between macro-accuracy and micro-accuracy. Some users are more concerned if a model is accurate in predicting electricity yields instead of accurately predicting an hourly or daily DC output curve. Although PVLib would be the most suitable for both these aspirations, it might not be the most favourable due to its complexity and required programming knowledge of either Python or MATLAB. In addition, the most accurate sub-models of PVLib such as SAPM, Ashrae and Physical require specific parameters only found in either the CEC or Sandia module databases. This limits the modelling application for PV modules that are not incorporated in these databases, such as prototype or state-of-the-art PV modules. For these applications it is necessary to use less accurate sub-models or to select parameters from a different PV module in these databases.

For users looking for a more user-friendly approach to accurately and swiftly predict annual yields it is recommended to use Helioscope, taken that no yield of specific years should be modelled. In the scenario where micro-accuracy is favoured over macro-accuracy, PVSyst and SAM are more recommended. Although PVSyst is more extensive and accurate than SAM the latter is still the better option for users that do not want to purchase the PVSyst software. In addition, running PVWatts from the SAM environment is not recommended as it is less accurate and setting up the extra configurations needed to run SAM does not require a lot of extra effort. Another drawback of PVLib is that it requires the user to collect and provide its own meteorological data. SAM, PVWatts, PVSyst and Helioscope all offer the possibility to easily import meteorological data from weather stations and satellite data all over the globe, making it more suitable for modelling for various different locations. SAM, PVWatts and PVSyst however substantially underestimated the electricity yield, so caution has to be taken applying system losses when using these simulation models.

Machine learning models are recommended for users that have a limited amount of measured performance data but that need to predict the annual electricity yield or the performance for different years. This has implications for PV module developers that conduct measurements for testing new prototypes but do not have the time or resources for full year measurements. For those developers that initially know the measuring time of a PV project, the optimal time period found in this study can be very valuable. If the performance needs to be extrapolated to the entire year using machine learning, then it turns out that if only between 1 and 4 months of data can be accumulated, the best month to start measuring would be April. When the measured irradiance is the global POA irradiance then measuring only the month of April already leads to decent accuracies. This phenomenon is likely to be explained by the average temperature of the months of spring. Measuring only in wintertime makes the machine learning algorithm overestimate the annual electricity yield as the module efficiency is highest for low temperatures. The opposite is true for only measuring summer months where the algorithm underestimates the annual electricity yield. The optimum measuring period, however, is not simply the average year temperature but is likely to be the period for which the average air temperature is equal to the average temperature at which the aggerated electricity of the whole year was generated. The average annual air temperature is not representative for PV electricity generation as most solar irradiance is collected in the summer and thus explains why measuring in spring is better than measuring during the end of summer/beginning of autumn.

The application of machine learning is in principle limited to predictions made for the same PV module as for which the machine learning model was trained. For a more generalized application it is recommended to train and validate a model with several different monocrystalline silicon cells. The model can then be used to predict the performance of another type of monocrystalline silicon PV module, taken that it is corrected for the rated power. In the last section a suggestion for further research is made, intended to develop generalized PV prediction models in order to expand the application of machine learning.

The above recommendations are however in dispute with the results from Gurupira & Rix (2017). According to their comparative study of PVSyst, SAM and PVLib, PVSyst had the highest accuracy and PVLib the lowest. Their results were however only based on the yield-error, which values nonetheless still contradict the results. Their exclusion of micro-accuracy error-metrics led to a limited comparison of these three models. This study built on that by differentiating between two levels of accuracy and to incorporate additional simulation and machine learning models. The incorporation of machine learning models also complemented the research conducted by Kazem (2016), who demonstrated similar accuracies obtained by a SVM machine learning model.

Limitations

The objective of this research was to include the most comprehensive number of PV prediction models. The design of this comparative study was therefore of exploratory nature to incorporate as many simulation and machine learning models as possible. This meant making concessions between including as much prediction models as possible and thoroughly analysing and configuring each one of them. The consequences of these concessions are concisely discussed in the next three paragraphs.

A limitation of this research was its limited effort in model optimisation. Both simulation and machine learning models offered the possibility in additional optimisation by changing parameters or modelling options. Many default options and parameters were left untouched, which could have impacted the accuracy of the models. Especially complex and flexible models such as PVSyst and PVLib offered advanced optimisations that have not been fully utilised in this study. Detailed model optimisation was outside the scope of this research and requires advanced knowledge of each prediction model. In addition, a sensitivity analysis was only conducted for the most accurate simulation and machine learning model. Other models might behave differently when varying input variables as they use different modelling parameters. These insights could have only been revealed if multiple sensitivity analyses were conducted.

Some model configurations were known to be incorrect but had to be used due to data and model limitations. For PVSyst it was only possible to model the polysilicon version of the UPOT PV module, although the PV module in question is made from monocrystalline silicon solar cells. For the SAPM sub-models technical parameters had to be taken from a similar PV module available in the Sandia module database. Such incorrect configurations are likely to have affected the results, although it remains unknown to what extent. An additional sensitivity analysis could have given insights on the influence of these configurations.

This study was mainly based on the output of the different simulation models, which was the modelled DC power output. The output files of all simulation models however included many different variables of intermediary steps. This means that a certain model can be the most accurate based on the DC power output, but that it is not necessarily the most accurate for all its intermediary steps. E.g. it could be that SAM's cell temperature model is more accurate than the cell temperature model of PVSyst, even though PVSyst was more accurate in modelling DC power output. In addition, this research only compared the modelled DC power output but not the modelled AC power output. Each simulation model has a different method for modelling AC from DC, so including this step could have led to a different result. DC/AC modelling is however not that complex and was therefore not expected to have a large effect on the results.

Further Research

In response to the limitations, it is valuable to further optimise the simulation models and make intermediary comparisons after every modelling step. Conducting a similar study with an expert for each simulation model ensures each model is run in its most optimal configuration. In addition, all simulation models provide the option to export datafiles of variables for intermediary modelling steps. Determining accuracy for intermediary modelling steps reveals what the most accurate overall model path is, which is one that could even overlap different simulation models. Such research leads to great insights when striving to develop the most accurate simulation model.

To further expand the application of machine learning in the domain of predicting PV performance, a study must be conducted with access to a large database of measured PV performance for a wide variety of PV modules. A large model can then be trained that takes the specific technical module parameters as independent input variables. If the model is trained with data from a sufficient number of different PV modules, then perhaps the model can be used to accurately predict the performance for new modules that do not have measured performance data. Such a model could be expanded to include location dependencies, different PV technologies and orientation. If organisations such as the NREL, Sandia National Laboratories and the California Energy Commission could combine and publicise their data then this would create a large enough data source for such an experiment. Developing a universal model for predicting PV performance would certainly be a breakthrough that substantially advances the field of PV modelling.

Overall, all four sub-model combinations of PVLib were found to be the most accurate for predicting PV performance. Although the differences are small, SAPM is the most accurate modelling from global POA irradiance and the combination of Physical and FSSC for modelling from GHI. More user-friendly models such as PVSyst and SAM however also lead to decent accuracies, but caution has to be taken for applying system losses as these models tend to underestimate electricity yields. In addition, machine learning models prove to be accurate in predicting electricity yields, but generalisation and wider application require more research with more extensive data.

Conclusion

This comparative study was set up to compare the accuracies of the most applied PV simulation models and of various machine learning techniques in predicting the DC power output of a PV module. The main research question was therefore stated as followed:

Which of the most applied PV simulation and machine learning models is the most accurate in predicting PV power output?

The prediction models' accuracies were further investigated according to four sub-questions that assessed the influence of the source of meteorological data, the type of input irradiance, the resolution of the input data and the time period of training data for the machine learning models. The prediction models for which the accuracy was determined were simulation models PVLib, SAM, PVWatts, PVSyst and Helioscope and machine learning models that were based on single linear regression, multi linear regression, polynomial regression, K-Nearest neighbours regression, decision tree regression and on three ensemble methods called bagging, boosting and stacking. The accuracy was determined on a macro- and microlevel.

The prediction model with the highest macro- and micro-accuracy in predicting PV power output turned out to be the four sub-model combinations of **PVLib**. The PVLib combinations were able to most accurately predict the DC power output for both sources of data, both types of solar irradiance and for all three timesteps. The sensitivity analysis further showed that even over a range of possible input values the PVLib sub-model combinations presented the most accurate results. In particular, sub-model SAPM was the most accurate in modelling from global POA irradiance, with an NRMSE of 0.0056 and a lowest yield error of 0.24%, and modelling from KNMI with an NRMSE of 0.181 and a yield error of -0.03%. The single diode combination with sub-models Physical and FSSC was the most accurate in modelling from GHI, with an NRMSE of 0.071 and a lowest yield error of 0.01%.

The influence of the source of meteorological data is concluded to substantially influence the micro-accuracy of both simulation and machine learning models. The NRMSE for all prediction models was on average twice as high for the KNMI data than for the UPOT data. Measuring the GHI from the source of the PV modules is thus essential in obtaining the highest micro-accuracy, even when the measurements are taken from a weather station just 2 km away. This statement is aligned with the Helioscope micro-accuracy from using the IWEC as a data source. The IWEC weather station in Amsterdam is located 40 km away from the UPOT facility and resulted in a NRMSE that was on average even three times higher than for the UPOT data. The macro-accuracy, on the other hand, was not considered to be higher for either one of the data sources and seems to be more influenced by the type of prediction model rather than by the source. The source however still needs to be representable for the weather at the location of the PV module.

The solar input irradiance is found to influence the micro-accuracy for all prediction models. The NRMSE of the simulation models was on average 0.02 higher using global POA irradiance and for the machine learning models even 0.09. Especially using machine learning models for predicting PV performance it is thus beneficial to train using measured global POA irradiance as the NRMSE was almost twice as high. The macro-accuracy, however, did not seem to be influenced by the type of solar input irradiance.

The resolution or timestep of the meteorological data seems to slightly influence both the macro- and micro-accuracy. For all prediction models the highest micro-accuracy was obtained for the 15-minute timestep data, followed second by the 60-minute timestep. It seemed that the 15-minute timestep had the optimal balance, as it was capable of accurately modelling both for days with constant solar irradiance as for days with highly fluctuating insolation. It was however not clear to what extend the smoothing of down-sampled data influenced these results. The macro-accuracy for the machine learning models seemed on average to slightly increase with an increasing timestep. Remarkably, the opposite was found to be true for the simulation models, although this was not consistent with all prediction models. These macro-accuracy observations were however only determined for the use of GHI. The timestep did not seem to influence the macro-accuracy using the global POA irradiance.

Both the micro- and macro-accuracy is found to be substantially influenced by the time period of training data for a machine learning model. The influence however decreases over time as the marginal increase in accuracy was found to saturate after 8 months for all four time period scenarios. The time period had the smallest influence and highest initial accuracy for the month April as starting month. Using measured global POA irradiance and PV performance data of only the month of April as training data led to a NRMSE of 0.113 and yield error of 2.20%. The minimum required amount of training data for acceptable accuracies is found not to be influenced by the timestep of training data. Although the 2-min dataset contained 30 times more data than the 60-min dataset, it did not require a smaller time period of training data. The machine learning accuracy is primarily determined by the number of months used as training data and the seasonal position of the measured months, for which spring months are the best to begin with.

For correctly setting up a PV modelling project it is crucial to initially review the available data, time and professional resources and to decide whether the objective is better achieved by aiming for high macro- or micro-accuracy. The results found in this comparative study facilitate in selecting the most suitable PV prediction model for each objective, incorporating both simulation and machine learning options.

Acknowledgements

I want to thank my supervisor Dr Atse Louwen for guiding me in conducting this research and for providing extensive feedback in writing the report. Many thanks to the UPOT facility for measuring and providing the needed data to conduct this comparative study. In addition, I like to thank Marian Habets and Jeroen Teeuwisse for helping me in the didactic revision of the report.

References

Alpaydin, E. (2009). *Introduction to Machine* (2nd Editio). London: The MIT Press. https://doi.org/10.5170/CERN-2016-002.1

Andrews, R. W., Stein, J. S., Hansen, C., & Riley, D. (2014). Introduction to the open source PV LIB for python Photovoltaic system modelling package. *2014 IEEE 40th Photovoltaic Specialist Conference, PVSC 2014*, 170–174. https://doi.org/10.1109/PVSC.2014.6925501

Bishop, J. W. (1988). Computer simulation of the effects of electrical mismatches in photovoltaic cell interconnection circuits. *Solar Cells*, *25*(1), 73–89. https://doi.org/10.1016/0379-6787(88)90059-2

Blok, K., & Nieuwlaar, E. (2017). *Introduction to Energy Analysis* (2nd Editio). New York: Routledge.

Bouzidi, K., Chegaar, M., & Bouhemadou, A. (2007). Solar cells parameters evaluation considering the series and shunt resistance. *Solar Energy Materials and Solar Cells*, *91*(18), 1647–1651. https://doi.org/10.1016/j.solmat.2007.05.019

Bronshtein, A. (2017). Train/Test Split and Cross Validation in Python. Retrieved April 11, 2019, from https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

Çengel, A., Y., & Boles, A., M. (2015). *Thermodynamics: An Engineering Approach* (8th Editio). New York: McGraw-Hill Education.

Chain, C., George, R. A. Y., Vignola, F., Perez, R., Ineichen, P., Moore, K., & Kmiecik, M. (2002). A new operational model for satellite-derived irradiances: description and validation. *Solar Energy*, *73*(5), 307–317. https://doi.org/10.1016/S0038-092X(02)00122-6

Cielen, D., Meysman, A. D. B., & Ali, M. (2016). *Introducing Data Science* (First Edit). New York: Manning Publications Co.

Das, A. (n.d.). Python | Decision Tree Regression using sklearn. Retrieved April 11, 2019, from https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

Dobos, A. (2012). An improved coefficient calculator for the California energy commission 6 parameter photovoltaic module model. *Journal of Solar Energy Engineering*, *134*(2).

Dolara, A., Leva, S., & Manzolini, G. (2015). Comparison of different physical models for PV power output prediction. *Solar Energy*, *119*(June), 83–99. https://doi.org/10.1016/j.solener.2015.06.017

Figgis, B., Ennaoui, A., Ahzi, S., & Rémond, Y. (2017). Review of PV soiling particle mechanics in desert environments. *Renewable and Sustainable Energy Reviews*, *76*(January), 872–881. https://doi.org/10.1016/j.rser.2017.03.100

Folsom Labs. (n.d.). HelioScope. Retrieved November 26, 2018, from https://www.helioscope.com

Gurupira, T., & Rix, A. J. (2017). Pv Simulation Software Comparisons : Pvsyst , Nrel Sam and Pvlib, (February). Retrieved from https://www.researchgate.net/publication/313249367_PV_SIMULATION_SOFTWARE_COMPARISONS_PVSYST_NREL_SAM_AND_PVLIB

Hackeling, G. (2017). *Mastering Machine Learning with scikit-learn Second Edition* (Second Edi). Birmingham: Packt Publishing Ltd.

Harb, S., Kedia, M., Zhang, H., & Balog, R. S. (2013). Microinverter and string inverter grid-connected photovoltaic system - A comprehensive study. *Conference Record of the IEEE Photovoltaic Specialists Conference*, 2885–2890. https://doi.org/10.1109/PVSC.2013.6745072

Hernday, P. (2011). Field Applications for I-V Curve Tracers. Retrieved December 6, 2018, from https://solarprofessional.com/articles/design-installation/field-applications-for-i-v-curve-tracers/page/0/3#.XAkTCC3WAWo

Honsberg, C., & Bowden, S. (n.d.). Photovoltaic Education. Retrieved November 23, 2018, from https://www.pveducation.org/pvcdrom/solar-cell-operation/series-resistance

IEA-PVPS. (2017). *Technical Assumptions Used in PV Financial Models. Review of Current Practices and Recommendations.*

IEA. (2017a). *Key world energy statistics*. Paris: International Energy Agency.

IEA. (2017b). *Renewables 2017*. Paris.

IEA. (2018). Solar PV. Retrieved November 16, 2018, from https://www.iea.org/tcep/power/renewables/solar/

Jain, A., & Kapoor, A. (2004). Exact analytical solutions of the parameters of real solar cells using Lambert W-function. *Solar Energy Materials and Solar Cells*, *81*(2), 269–277. https://doi.org/10.1016/j.solmat.2003.11.018

Kazem, H. A., Chaichan, M. T., Al-shezawi, I. M., Al-saidi, H. S., Al-rubkhi, H. S., Al-sinani, K., & Al-waeli, A. H. A. (2014). Effect of humidity on the PV performance in Oman. *Asian Transactions on Engineering*, *2*(4), 29–32.

Kazem, H. A., Yousif, J., & Chaichan, M. (2016). Modelling of Daily Solar Energy System Prediction using Support Vector Machine for Oman, *11*(20), 10166–10172.

King, D. L., Boyson, W. E., & Kratochvill, J. A. (2004). SANDIA REPORT Photovoltaic Array Performance Model, (December). Retrieved from http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

Kirn, B., & Topic, M. (2017). Diffuse and direct light solar spectra modeling in PV module performance rating. *Solar Energy*, *150*, 310–316. https://doi.org/10.1016/j.solener.2017.04.047

Kotak, Y., Gul, M. S., Muneer, T., & Ivanova, S. M. (2015). Impact of Ground Albedo on the Performance of PV Systems and its economic analysis. *7th International Conference on Solar Radiation and Daylight*, (April), 1–16.

Kumar Sahu, B. (2015). A study on global solar PV energy developments and policies with special focus on the top ten solar PV power producing countries. *Renewable and Sustainable Energy Reviews*, *43*, 621–634. https://doi.org/10.1016/j.rser.2014.11.058

Lauret, P., Voyant, C., Soubdhan, T., David, M., & Poggi, P. (2015). A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. *Solar Energy*, *112*, 446–457. https://doi.org/10.1016/j.solener.2014.12.014

Lave, M., Hayes, W., Pohl, A., & Hansen, C. W. (2015). Evaluation of Global Horizontal Irradiance to Plane of Array Irradiance Models at Locations across the United States. *IEEE Journal of Photovoltaics*, *5*(2), 597–606. https://doi.org/10.1109/JPHOTOV.2015.2392938.c

Li, J., Ward, J. K., Tong, J., Collins, L., & Platt, G. (2016). Machine learning for solar irradiance forecasting of photovoltaic system. *Renewable Energy*, *90*, 542–553. https://doi.org/10.1016/j.renene.2015.12.069

Luque, A., & Hegedus, S. (2011). *Handbook of Photovoltaic Science and Engineering* (2nd Editio). Wiley.

Marion, B. (2008). Comparison of predictive models for photovoltaic module performance. *Conference Record of the IEEE Photovoltaic Specialists Conference*, (May). https://doi.org/10.1109/PVSC.2008.4922586

Marion, B. (2015). A model for deriving the direct normal and diffuse horizontal irradiance from the global tilted irradiance. *Solar Energy*, *122*, 1037–1046. https://doi.org/10.1016/j.solener.2015.10.024

Martínez-Moreno, F., Muñoz, J., & Lorenzo, E. (2010). Experimental model to estimate shading losses on PV arrays. *Solar Energy Materials and Solar Cells*, *94*(12), 2298–2303. https://doi.org/10.1016/j.solmat.2010.07.029

Mavromatakis, F., & Vignola, F. (2016). Spectral Performance of PV Modules of Different Technologies. https://doi.org/10.22618/tp.ei.20163.389015

Müller, A. C., & Guido, S. (2017). *Introduction to machine learning with Python* (First Edit). Sebastopol: O'Reilly Media, Inc. Retrieved from https://github.com/justmarkham/scikit-learn-videos%0Ahttp://oreilly.com/catalog/errata.csp?isbn=9781449369415 for

Pawar, D. (2018). Improving Performance of Convolutional Neural Network.

Reda, I., & Andreas, A. (2004). Solar position algorithm for solar radiation applications. *Solar Energy*, *76*(5), 577–589. https://doi.org/10.1016/j.solener.2003.12.003

Sandia. (n.d.). Spectral Responce. Retrieved April 4, 2019, from https://pvpmc.sandia.gov/modeling-steps/2-dc-module-iv/effective-irradiance/spectral-response/

Scikit-Learn Developers. (n.d.). Scikit-learn algorithm cheat-sheet. Retrieved February 4, 2019, from https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Seaward. (n.d.). What is solar PV I-V curve tracing? Retrieved November 23, 2018, from http://www.seaward-groupusa.com/userfiles/curve-tracing.php

Silvestre, S., Boronat, A., & Chouder, A. (2009). Study of bypass diodes configuration on PV modules. *Applied Energy*, *86*(9), 1632–1640. https://doi.org/10.1016/j.apenergy.2009.01.020

Singh, A. (2018). A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code).

Stein, J. (2016). 6th PV Performance Modelling Workshop [PowerPoint Presentation]. In *PhD*. Fraunhofer ISE, Freiburg: Sandia National Laboratories (SNL).

Stein, J. S., & Klise, G. T. (2009). Models used to assess the performance of photovoltaic systems., (January). https://doi.org/10.2172/974415

Tapia, M. H., & H., R. (2014). Evaluation of Performance Models against Actual Performance of Grid Connected PV Systems, 1–36. Retrieved from http://oops.uni-oldenburg.de/2433/7/Thesis_TapiaM.pdf%0Ahttps://d-nb.info/1077657072/34

Time and Date. (n.d.). Altitude & Azimuth: The Horizontal Coordinate System. Retrieved April 3, 2019, from https://www.timeanddate.com/astronomy/horizontal-coordinate-system.html

Tong, K., & Granat, M. (1999). A practical gait analysis system using gyroscopes. *Medical Engineering & Physics*, *21*, 87–94. https://doi.org/10.1016/j.jsams.2009.01.005

Touati, F., Al-Hitmi, M. A., Chowdhury, N. A., Hamad, J. A., & San Pedro Gonzales, A. J. R. (2016). Investigation of solar PV performance under Doha weather using a customized measurement and monitoring system. *Renewable Energy*, *89*, 564–577. https://doi.org/10.1016/j.renene.2015.12.046

Twidell, J., & Weir, T. (2015). *Renewable Energy Resources* (Third Edit). New York: Routledge.

Unpingco, J. (2016). *Python for Probability, Statistics, and Machine Learning*. Springer International Publishing AG Switzerland. https://doi.org/10.1007/978-3-319-30717-6

van Sark, W., Louwen, A., de Waal, A., Elsinga, B., & Schropp, R. (2012). UPOT: THE UTRECHT PHOTOVOLTAIC OUTDOOR TEST FACILITY Wilfried. *27th European Photovoltaic Solar Energy Conference and Exhibition*.

Water University. (n.d.). Sun vs Shade. Retrieved November 22, 2018, from https://wateruniversity.tamu.edu/plants/sun-vs-shade/

Yella, A., Lee, H. W., Tsao, H. N., Yi, C., Chandiran, A. K., Nazeeruddin, M. K., & Grätzel, M. (2011). Porphyrin-Sensitized Solar Cells with Cobalt (II/III)–Based Redox Electrolyte Exceed 12 Percent Efficiency. *Science*, *334*((6056)), 629–634. https://doi.org/10.1007/s11133-011-9215-z

Yusufoglu, U. A., Pletzer, T. M., Min, B., Van Mo̍lken, J., Litzenburger, B., Pingel, S., … Kurz, H. (2013). A simulation study on the annual energy yield gain of solar modules by reduction of mismatch losses through sorting of solar cells. *28th European Photovoltaic Solar Energy Conference and Exhibition*, 3203–3206.

Appendix

A.1 PV Module Parameters

Table A.1. Flash-Test Results	
Module	Commercial (UPOT) PV Module
Technology	c-Si n-type
Area	1.6335
Rated Power	265
rImp	8.55
rVMP	31
rIsc	8.93
rVoc	39
alpha Isc	0.0004
beta Voc	-0.0033
gamma Pmp	-0.0042

Table A.2. CEC Module Database	
Name	Commercial (UPOT) PV Module
BIPV	N
Date	02/01/2012
T_NOCT	45
A_c	1.634
Technology	Mono-c-Si
N_s	60
I_sc_ref	9.35
V_oc_ref	38.28
I_mp_ref	8.73
V_mp_ref	30.38
alpha_sc	0.004114
beta_oc	-0.11484
a_ref	1.4502
I_L_ref	9.37
I_o_ref	3.15E-11
R_s	0.41
R_sh_ref	194.16
Adjust	-0.03344
gamma_r	-0.377
Version	NRELv1
PTC	244.4

Table A.3. Sandia Module Database	
Name	Most comparable PV Module
Vintage	2009 (E)
Area	1,643
Material	c-Si
Cells_in_Series	60
Parallel_Strings	1
Isco	8.52
Voco	37.5
Impo	7.93
Vmpo	30.78
Aisc	0.00029
Aimp	-000022
C0	1.003
C1	-0.003
Bvoco	-0.126
Mbvoc	0
Bvmpo	-0.135
Mbvmp	0
N	1.323
C2	0.001
C3	-8.711
A0	0.9315
A1	0.05975
A2	-0.01067
A3	0.0008
A4	-2.24E-05
B0	1
B1	-0.002438
B2	0.00031
B3	-1.246E-05
B4	2.11E-07
B5	-1.36E-09
DTC	3
FD	1
A	-3.47
B	-0.0594
C4	0.992
C5	0.008
IXO	8.35
IXXO	5.61
C6	1.128
C7	-0.128
Notes	Source: Sandia National Laboratories Updated 9/25/2012 Module Database

A.2. Optimal K-Value

Figure A.1 presents the elbow curve for predicting the DC power output for the year 2016 from a KNeighbours regression model trained on 2015 data. The optimal K-value was found to be close to 10, for which the curve has the lowest RMSE. The optimal K-value found in Figure A.1. was used in training the KNeighbours regression model.

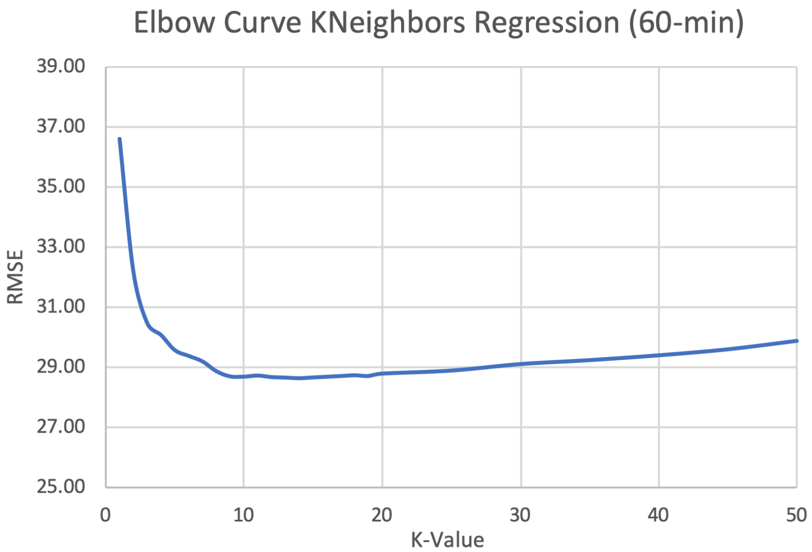


Figure A.1. Elbow curve for predicting DC power output for various K-values. Based on 60-min UPOT GHI as primary input irradiance (2016).