Simo Ben

Ultra-Short-Term Solar PV Power Forecasting Using Gradient Boosting Regression Trees

School of Technology and Innovation
Master's thesis in Smart Energy Programme

Vaasa 2026

UNIVERSITY OF VAASA
School of Technology and Innovation
Author: Simo Ben
Title of the Thesis: Ultra-Short-Term Solar PV Power Forecasting Using Gradient Boosting Regression
Degree: Master of Science in Technology
Programme: Smart Energy
Supervisor: Petri Välisuo & Timo Mantere
Year: 2026 **Sivumäärä:** 84

ABSTRACT:

As solar PV power generation becomes more significant in the Nordic power systems, operation, flexibility management and local energy management are also becoming more challenging. In high-latitude regions, such as Helsinki, Finland, cloud-induced irradiance fluctuations, large seasonal variations in Solar elevation, shorter daylight in winter and snowy/icy surface conditions can greatly influence PV production, making power forecasting more challenging. These conditions lead to forecast uncertainty at very short time horizons and demand forecasting techniques with the ability to forecast both fast power variations and seasonal operating patterns. The aim of this thesis is to examine the possibilities of ultra short-term (UST) PV power forecasting with a 10–60-minute forecast horizon for the Helsinki Kumpula measurement site data. A supervised machine-learning approach using Gradient Boosting Regression Trees (GBRT) is designed to model the non-linear relationship between PV power output and environmental predictors. The input features are lagged PV power measurements, irradiance-related features, cloud cover, air temperature, wind speed, relative humidity, solar-geometry features, clear-sky normalization, and snow-sensitive predictors. These features are intended to depict the most recent system behavior, evolving meteorological conditions, and seasonal effects in the high latitudes. A time-series-aware validation approach is used to evaluate model performance, and a persistence benchmark is used for comparison. RMSE, NRMSE, MAE, R2 and forecast skills are used to measure forecast accuracy. The outcomes show that the model results in the best accuracy for the smallest forecast time used in the experiment, 10 minutes, with an RMSE of 1.94 kW and an nRMSE of 10.03%. The RMSE and nRMSE values are also slowly increasing as the prediction horizon increases, increasing to 2.41kW and 12.46% respectively after 60 minutes. The GBRT model is superior to the persistence in forecasting skills as seen from the skill scores of 0.22 at 10 minutes and 0.27 at 60 minutes, which are positive values. The results also indicate that recent PV power observations are more significant than the other predictors for short forecast leads, while the meteorological, solar-geometry, and snow-related predictors gain in significance with increasing forecast leads. The central idea of this thesis is a systematic and transparent forecasting framework based on GBRT for the ultra-short-term prediction of PV power in Nordic high latitude conditions. The proposed solution enhances the short-term situational awareness of PV generation, thus paving the way for the seamless integration of solar power into flexible power systems. The results also underpin future hybrid approaches to forecasting, based on machine learning and physical irradiance modelling and sky-camera-based cloud information.

KEYWORDS: Photovoltaic forecasting, Ultra-short-term forecasting, Gradient Boosting, Solar Power, Machine Learning, Clear-sky Index, Snow-aware forecasting.

VAASAN YLIOPISTO
Tekniikan ja innovaatiojohtamisen akateeminen yksikkö
Tekijä: Simo Ben
Tutkielman nimi: **Erittäin lyhyen aikavälin aurinkosähkötehon ennustaminen gradienttivahvistusmalleilla**
Tutkinto: Tekniikan diplomi-insinööri
Oppiaine: Smart Energy
Työn ohjaajat: Petri Välisuo & Timo Mantere
Valmistumisvuosi: 2026 **Sivumäärä:** 84

TIIVISTELMÄ:
Aurinkosähköntuotannon merkityksen kasvaessa pohjoismaisissa sähköjärjestelmissä myös järjestelmän käyttö, joustonhallinta ja paikallinen energianhallinta muuttuvat haastavammiksi. Korkean leveysasteen alueilla, kuten Helsingissä, pilvien aiheuttamat säteilyvaihtelut, suuret vuodenaikaiset muutokset auringon korkeuskulmassa, talven lyhyempi päivänpituus sekä lumiset ja jäiset olosuhteet voivat vaikuttaa merkittävästi aurinkosähkön tuotantoon ja vaikeuttaa tehon ennustamista. Nämä olosuhteet aiheuttavat ennusteen epävarmuutta hyvin lyhyillä ennustehorisonteilla ja edellyttävät ennustusmenetelmiä, jotka kykenevät ennustamaan sekä nopeita tehonvaihteluita että vuodenaikaisia tuotantomalleja. Tämän tutkielman tavoitteena on tarkastella ultra-short-term- eli erittäin lyhyen aikavälin aurinkosähkötehon ennustamista 10–60 minuutin ennustehorisonteilla käyttäen Helsingin Kumpulan mittausaseman aineistoa. Tutkielmassa käytetään ohjattuun koneoppimiseen perustuvaa menetelmää, jossa Gradient Boosting Regression Trees -mallilla (GBRT) mallinnetaan aurinkosähkötehon ja ympäristömuuttujien välistä epälineaarista yhteyttä. Syötemuuttujina käytetään viivästettyjä PV-tehomittauksia, säteilyyn liittyviä muuttujia, pilvisyyttä, ilman lämpötilaa, tuulen nopeutta, suhteellista kosteutta, aurinkogeometriaan liittyviä muuttujia, clear-sky-normalisointia sekä lumelle herkkiä muuttujia. Näiden muuttujien tarkoituksena on kuvata järjestelmän viimeaikaista käyttäytymistä, muuttuvia meteorologisia olosuhteita sekä korkean leveysasteen vuodenaikaisvaikutuksia. Mallin suorituskykyä arvioidaan aikasarjarakenteen huomioivalla validointimenetelmällä, ja vertailukohtana käytetään persistenssimallia. Ennustetarkkuuden arviointiin käytetään RMSE-, nRMSE-, MAE- ja R^2-mittareita sekä ennustetaitoa kuvaavaa skill score -arvoa. Tulokset osoittavat, että malli saavuttaa parhaan tarkkuuden lyhimmällä tarkastellulla ennustehorisontilla, eli 10 minuutissa, jolloin RMSE on 1,94 kW ja nRMSE 10,03 %. RMSE- ja nRMSE-arvot kasvavat ennustehorisontin pidentyessä, ja 60 minuutin horisontilla ne ovat vastaavasti 2,41 kW ja 12,46 %.GBRT-malli suoriutuu persistenssimallia paremmin, mikä näkyy positiivisina skill score -arvoina: 0,22 10 minuutin horisontilla ja 0,27 60 minuutin horisontilla. Tulokset osoittavat myös, että viimeaikaiset PV-tehohavainnot ovat tärkeimpiä lyhyillä ennustehorisonteilla, kun taas meteorologisten, aurinkogeometristen ja lumeen liittyvien muuttujien merkitys kasvaa ennustehorisontin pidentyessä. Tutkielman keskeinen ajatus on systemaattinen ja läpinäkyvä GBRT-pohjainen ennustuskehys aurinkosähkötehon erittäin lyhyen aikavälin ennustamiseen pohjoisissa korkean leveysasteen olosuhteissa. Ehdotettu ratkaisu parantaa aurinkosähkötuotannon lyhyen aikavälin tilannetietoisuutta ja tukee siten aurinkovoiman integrointia joustaviin sähköjärjestelmiin. Tulokset tukevat myös tulevia hybridiennustusmenetelmiä, joissa yhdistetään koneoppimista, fysikaalista säteilymallinnusta ja taivaskamerapohjaista pilvitietoa.

AVAINSANAT: Aurinkosähkön ennustaminen, Ultra-lyhyen aikavälin ennustaminen, Gradient Boosting Koneoppiminen, Selkeän taivaan indeksi, Lumivaikutus, PV-ennustaminen

Contents

1 Introduction 6

1.1 Research problem 7

1.1.1 Research question 7

1.1.2 Scope of the study 8

1.1.3 Outline 8

2 Background and Literature Review 10

2.1 General Research on Solar Power Forecasting 10

2.1.1 Forecasting Inputs and Feature Engineering 11

2.1.2 Machine Learning Techniques 11

2.2 Physical and Images-Based Irradiance Modelling 13

2.2.1 Extraterrestrial Irradiance 13

2.2.2 Air Mass and Atmospheric Effects 14

2.2.3 Sky Camera and Image-Based Forecasting 15

2.2.4 Conclusion and Insights 16

3 Mathematical Theory 18

3.1 PV Forecasting Problem 18

3.1.1 Persistence Baseline 19

3.1.2 Tree Algorithms 20

3.1.3 Regression trees 20

3.1.4 Gradient Boosting Regression Trees 22

3.1.5 Boosting 23

3.1.6 Evaluation Metrics 24

4 Method and Data 25

4.1 Raw Data 26

4.1.1 Photovoltaic Power Data 27

4.1.2 Meteorological Measurements 27

4.1.3 Numerical Weather prediction Data 28

4.1.4 Variability in Data 30
4.2 Data Processing 32
4.2.1 Outlier Handling 34
4.2.2 Clear Sky Normalization 35
4.3 Feature Engineering 36
4.3.1 Feature Selection 39
4.3.2 Cross-Validation 40
4.4 Model Evaluation Framework 42
4.4.1 Root Mean Square Error 42
4.4.2 Mean Absolute Error 43
4.4.3 Normalized RMSE 43
4.4.4 Skill Score 43
4.4.5 Coefficient of Determination 44
4.4.6 Forecast skill 44
4.5 Forecast model 45
4.5.1 Forecasting algorithms 45
4.5.2 Gradient Boosting Regression Trees 46
4.5.3 Hyperparameter Optimization 47
4.6 Benchmark models 48
4.6.1 Persistence Model 48
4.6.2 Moving-Average Baseline 49
5 Results 51
5.1.1 Data Transformation and time series 51
5.1.2 Machine Learning model 54
5.1.3 Hyperparameter 54
5.1.4 RMSE 56
5.1.5 nRMSE 57
5.1.6 Skill scores 57
5.1.7 Run-time 58
6 Discussion 59

6.1 GBRT forecast performance 59
6.1.1 Role of input features 60
6.1.2 Data limitations 62
6.1.3 Error metrics 63
6.1.4 Methodology limitations 64
6.1.5 Physical modelling and future improvements 66
6.1.6 Overall interpretation 67
7 Conclusion and Future work 69
7.1 Conclusion 69
7.2 Future work 70
References 72
Appendices 76
Appendix 1. Additional forecasting results 76
Appendix 2. Python code for data preparation and model implementation 77

Algorithms

Algorithm 1. GBRT-based forecasting 49

Figures

Figure 2.1 Machine-learning framework..... 12
Figure 2.2 Annual Variation in extraterrestrial radiation..... 14
Figure 2.3 Air mass and the solar zenith angle..... 15
Figure 2.4 Sky camera workflow PV forecasting 16

Figure 4.1 Energy output for Helsinki……31

Figure 4.2 Installation PV……32

Figure 4.3 Solar radiation……32

Figure 4.4 Energy output during a sunny day……36

Figure 4.5 Energy output during a cloudy day……36

Figure 4.6 Available samples after forecast horizon filtering……37

Figure 4.7 Illustration of the solar zenith angle……42

Figure 5.1 Normalized PV power output…… 52

Figure 5.2 Actual PV output overtime…… 52

Figure 5.3 Autocorrelation of normalized PV …… 52

Figure 5.4 Normalized PV time series …… 52

Figure 5. 5 Prediction errors across forecast horizons for the Helsinki Kumpula site …… 52

Figure 5. 6 PV output for a selected day at the Helsinki Kumpula site …… 52

Figure 5.7 RMSE for cross validation of the GBRT model for various numbers of trees M and shrinkage values v. This is based on the FMI data set for Helsinki PV. Left, 10-minute forecast length. Right: 60-minute projected range. …… 58

Figure 5.8 RMSE obtained by cross validation for the GBRT model with the shrinkage value v=0.03 and various numbers of trees M. …… 59

Figure 5.9 PV variation and the GBRT forecasting accuracy for the forecast horizon for the Helsinki kumpula site …… 60

Figure 5.10 nRMSE of the GBRT model for the different forecast horizons at the site of Helsinki …… 61

Tables

Table 4.1 Forecasting framework as summarized……29

Table 4.2 PV Site…… 30

Table 4.3 NWP variable obtained Metasomatic……34

Table 4.4 Candidate predictor variables GBRT forecasting framework……44

Table 4.5 Cross-validation…… 45

Table 4.6 Hyperparameters GBRT…… 52

Table 5.1 Forecasting performance of the GBRT…… 57

Table 5.2 Final selected hyperparameters of the GBRT model. 58

Table 5.3 Skill scores of GBRT model 62

Table 5.4 Run-time of the GBRT model 62

Table 6.1 Final GBRT model for various forecasting horizons 65

Abbreviations

AR Autoregressive
ANN Artificial Neural Network
RMSE Root Mean Square Error
CB Cloud Base
R^2 Coefficient of Determination
CSI Clear-Sky Index
CS Clear-Sky
DHI Diffuse Horizontal Irradiance
DNI Direct Normal Irradiance
FMI Finnish Meteorological Institute
GBRT Gradient Boosting Regression Trees
GHI Global Horizontal Irradiance
KNN k-Nearest Neighbors
USTF Ultra-Short-Term Forecasting
MAE Mean Absolute Error
nRMSE Normalized Root Mean Square Error
ML Machine Learning
PV Photovoltaic

Symbols

P(t)	Observed PV power at time t
P̂(t+h)	Forecast PV power at horizon h
h	Forecast horizon (10–60 minutes)
t	Time index
P_pers (t+h)	Persistence forecast
θ_z	Solar zenith angle
L_d	Day length
CSI(t)	Clear-sky index at time t
GHI(t)	Global horizontal irradiance
P_cs (t)	Clear-sky PV power

1 Introduction

As solar photovoltaic (PV) generation becomes more prevalent, the power system faces new challenges in operating, managing local energy, and balancing its energy supply and demand. PV generation is not conventional and cannot be dispatched like normal generation as it depends on the weather. It depends greatly on the position of the sun, cloud cover, air temperature, wind speed and direction. A consequence of this is that PV power production may change rapidly, particularly if irradiance changes rapidly due to movements of the clouds.

PV forecasting plays therefore an important role in lowering uncertainty and the decision-making process. Forecasts can be made for varying forecast horizons ranging from a few minutes in the future up to day-ahead forecasts. The relevance of ultra-short-term forecasts extends to real-time grid operation, energy-storage control, and local power management and is particularly useful for forecast periods of a few minutes to hours. For such short time horizons, PV production can frequently be a good predictor of future production, and for longer time horizons, meteorological variables can play a larger role.

Persistence forecasting is widely used as a point of reference in short-term PV forecasting as it assumes that the PV output or the current irradiance conditions stay constant for a short forecast period. The approach works well at low-risk clear skies and with very short lead-time. The accuracy is generally reduced with longer forecasts and in scenarios involving quickly moving cloud patterns that result in dynamic changes of irradiation.

One of the biggest difficulties of PV power forecasting is in high-latitude areas like Helsinki. PV power output can be influenced by seasonal changes in the sun's elevation, significant variations in the amount of daylight, low elevation angles of the sun in winter, cloud cover, and snow and ice cover. All these conditions cause the nonlinear and highly time-dependent correlation between the production of PV and weather variables. Hence, the recent power observation from PV, meteorological variables, temporal features,

solar geometry effects and winter related factors should all be taken into consideration in forecasting models for Nordic conditions.

Several machine-learning methods can be used for ultra-short-term PV power forecasting, including random forests, artificial neural networks, support vector regression, and gradient-boosting methods. In this thesis, the forecasting task is first considered as a general supervised machine-learning regression problem, where future PV power is predicted using recent PV measurements, meteorological variables, and engineered time-dependent features. The specific modelling method is selected based on the literature review and the characteristics of the available dataset. Gradient Boosting Regression Trees are later chosen because they are suitable for structured tabular data, can model nonlinear interactions between PV output and weather-related predictors, and have shown good performance in previous PV forecasting studies.

1.1 Research problem

The aim of this thesis is to look at the issue of ultra-short term (UST) PV power forecasting in high latitudes. The seasonal variations of PV generation are more pronounced in Helsinki, the sun elevation during the winter period is low, the days during summer are long, the solar irradiance fluctuates significantly over a shorter period of time due to the cloud cover, and the impact of snow. This makes PV forecasting difficult and suggests that a model that can account for recent PV power observations and weather and time-related variables is needed. While simple persistence forecasts may be adequate for short-term forecasts, they are less effective during fast weather changes. A more sophisticated model is required to model the nonlinear relationships between PV power, irradiance, cloud, temperature, snow and temporal factors.

1.1.1 Research question

This thesis aims to build and evaluate a Gradient Boosting Regression Trees model for ultra-short-term photovoltaic power forecasting at the Helsinki Kumpula site. Measured

PV power output, meteorological variables, irradiance-related predictors, and engineered features are used in the model to predict PV power over short forecast horizons.

The research question is :

How accurately can a machine-learning model forecast photovoltaic power 10–60 minutes ahead at the Helsinki Kumpula site under high-latitude conditions ?

In this thesis, GBRT is selected as the implemented machine-learning model based on the literature review and its suitability for the available tabular PV and meteorological dataset.

1.1.2 Scope of the study

This study is limited to ultra-short-term photovoltaic power forecasting at the Helsinki Kumpula site. The analysis uses measured PV power output, local meteorological and irradiance observations, and selected numerical weather prediction variables. The forecasting horizons are 10, 30, and 60 minutes ahead.

The study evaluates a deterministic Gradient Boosting Regression Trees model that produces point forecasts of future PV power. Model performance is assessed using RMSE, nRMSE, MAE, R^2 and forecast skill against benchmark models. The results are therefore site-specific and should not be generalized directly to other PV installations without further validation.

Physical irradiance modelling, sky-camera forecasting and probabilistic forecasting are outside the main scope of the thesis. These approaches are discussed only as possible directions for future research.

1.1.3 Outline

Thesis is organised as follows.

Chapter 2 provides theoretical background and available literature about PV power forecasting, ultra-short term forecasting, machine learning algorithms and PV forecasting challenges at high latitudes.

Chapter 3 describes the methodology behind the forecasting model and metrics.

Chapter 4 Details the data collection, pre-processing, feature engineering and model training.

Chapter 5 Discusses the forecasting results (RMSE, nRMSE, skill scores, hyperparameter tuning).

Chapter 6 Reflects on the results with respect to the research question through model performance, predictor importance, limitations and suggestions for improvement.

Chapter 7 Draws conclusions and proposes future work.

2 Background and Literature Review

This chapter provides an overview of all the research that has been carried out on PV power forecasting, especially in the fields of short-term and ultra-short-term forecasting. The growing deployment of renewable energy and the need for accurate operational planning in the power system has led to solar forecasting gaining popularity as a topic of interest. There are several types of forecasting methods that are used, such as persistence-based forecast, statistical forecast, physical forecast, machine-learning forecast, and hybrid forecast. Each method has a different outlook on performance depending on the forecast period, availability of input data, site characteristics, and application.

2.1 General Research on Solar Power Forecasting

Solar photovoltaic (PV) power forecasting has become increasingly significant with the growing integration of PV generation into modern power systems. Weather conditions have a vital influence on PV production, and forecasting is a key requirement for balancing with the grid, scheduling PV generation, controlling energy storage and taking operations decisions. Forecasting time horizons may span intra-hour to day-ahead applications, and the most appropriate modelling technique is determined by the application, the data available and the time of the forecast. Intra-hour and ultra-short-term forecasts play a crucial role in energy-storage management, power smoothing, smart-grid operation and electricity dispatch (Al-Dahidi et al., 2024, pp. 2–3).

The literature review in this area reveals that, due to weather variability, data quality issues, forecast-horizon dependence and site-specific environmental conditions, solar PV forecasting is still in the state of being technically challenging. Physical, statistical, machine-learning and hybrid forecasting techniques have been studied. But none of the forecasting methods work best in every situation. Rather, model performance is a function of the interplay of input data, temporal resolution, forecast lead time and the environment of the forecasting site (Al-Dahidi et al., 2024, pp. 2–5).

2.1.1 Forecasting Inputs and Feature Engineering

The accuracy of the PV forecasts relies heavily on the quality, selection and availability of the input variables. Prediction performance is affected by weather conditions, forecast horizon, geographical location and availability of data. The direct inputs (global solar radiation, historical PV power) are of particular significance, as they are closely linked to PV generation. Indirect inputs like ambient temperature, wind speed and relative humidity can also increase the accuracy of the forecast as these influence the temperature of the cells, atmospheric conditions and system efficiency (Al-Dahidi et al., 2024, pp. 8–12).

Further, feature engineering and preprocessing play significant roles in the PV forecasting task. The results of previous studies indicate that normalization, dimensionality reduction, feature extraction, filtering, clustering and weather classification can be beneficial for prediction performance. In many cases, the raw weather data is not enough, and feature engineering can yield more meaningful features while mitigating the impact of noise, redundancy, and missing data (Al-Dahidi et al., 2024, pp. 8–12).

For short-horizon PV forecasting, it becomes essential to consider the historical sequence data. There has been recent research on forecasting with GBRT that demonstrates that models based on trees can perform better if they are fed with past observations and additional features derived from the sequence. For GBRT methods, the authors Zheng et al. contend that the historical dependence should be explicitly represented, as the temporal dynamics can only be represented in short time span by the timestamp variables. This will facilitate the use of lagged PV observations and sequence-based features in GBRT forecasting models (Zheng et al., 2024, pp. 1–3).

2.1.2 Machine Learning Techniques

Machine-learning methods gain importance in PV forecasting due to the ability of these methods to flexibly model nonlinear relationships between PV output and environmental variables, which is not accomplished by many conventional methods. The solar

forecasting community has been actively leveraging machine learning and deep learning techniques in recent years, with review studies indicating that these techniques are widely adopted and demonstrate high forecasting accuracy for various horizons. Furthermore, ensemble and hybrid methods frequently yield high forecasting precision for solar forecasting at different horizons, based on recent review studies. The literature also shows that the choice of predictors and predictor preprocessing are related to model performance, particularly when the variability is significant from a weather perspective (Al-Dahidi et al., 2024, pp. 3–5).

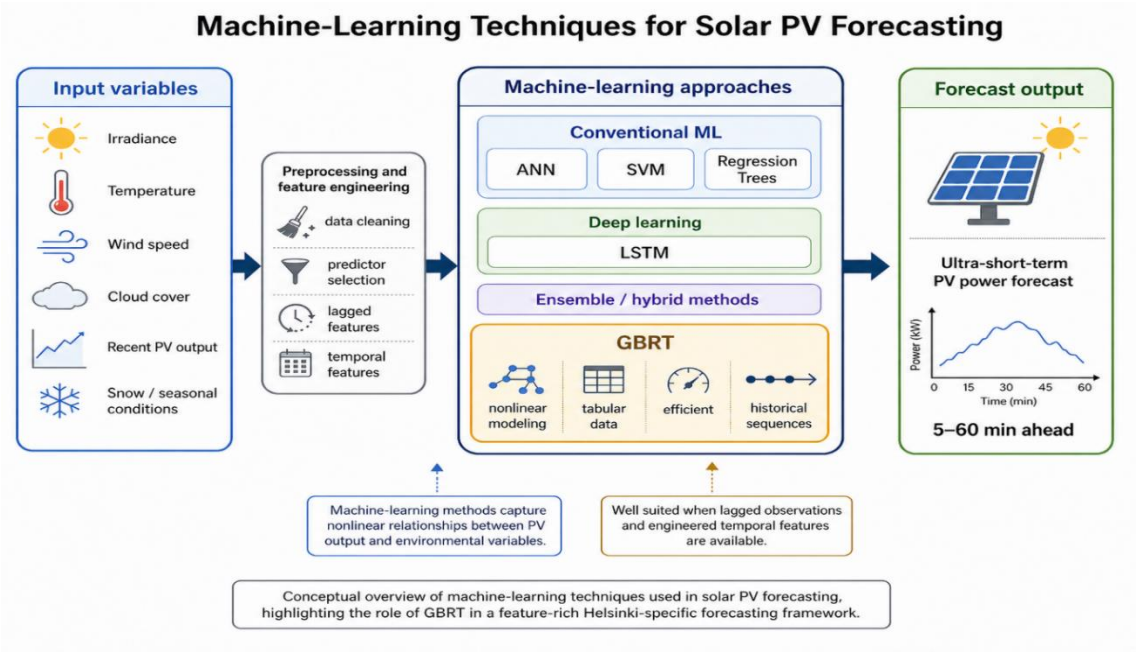


Figure 2.1 Machine-learning framework for ultra-short-term PV forecasting Al-Dahidi et al. (2024)

For the present thesis, Gradient Boosting Regression Trees are particularly relevant. Zheng et al. describe GBRT methods such as XGBoost as accurate and efficient tools for regression problems on tabular data, even when compared with more complex deep learning solutions. Their results show that enriching GBRT with features representing historical sequences improves forecasting accuracy across different horizons and datasets while maintaining favorable computational efficiency. This is methodologically important because it supports the use of GBRT as a practical forecasting framework when

lagged information and engineered temporal features are available (Zheng et al., 2024, pp. 1–2).

In the context of solar PV forecasting, machine-learning approaches are especially attractive because photovoltaic power depends on nonlinear interactions among irradiance, temperature, wind, cloud conditions, and recent production history. Recent review literature also notes that, for cold and snowy Nordic conditions, approaches such as ensemble methods, weather clustering, optimization algorithms, and snow-aware modelling strategies are valuable for improving prediction performance. This strengthens the case for using a feature-rich GBRT framework in a Helsinki-specific forecasting study (Al-Dahidi et al., 2024, pp. 3–5; Zheng et al., 2024, pp. 1–3).

2.2 Physical and Images-Based Irradiance Modelling

2.2.1 Extraterrestrial Irradiance

Extraterrestrial irradiance is the amount of solar radiation received at the top of the Earth's atmosphere without being filtered by the atmosphere, scattered by the atmosphere, or cloud attenuation. It is the theoretical maximum of solar radiation that a surface can receive when it is beyond the atmosphere and is typically given in watts per square metre (W/m^2). The extraterrestrial irradiance is largely dependent upon the solar constant and the distance between the earth and the sun, which changes from year to year due to the fact that the earth's orbit is elliptical.

The radiation from above the atmosphere follows a regular annual cycle as illustrated in Figure 2.2. The value will be the greatest during the closest approach of the Earth to the Sun, and the least during the greatest distance. This seasonal variation is used to provide an important reference for modelling solar energy, which sets the highest limit of solar radiation available before considering any variation caused by the atmosphere.

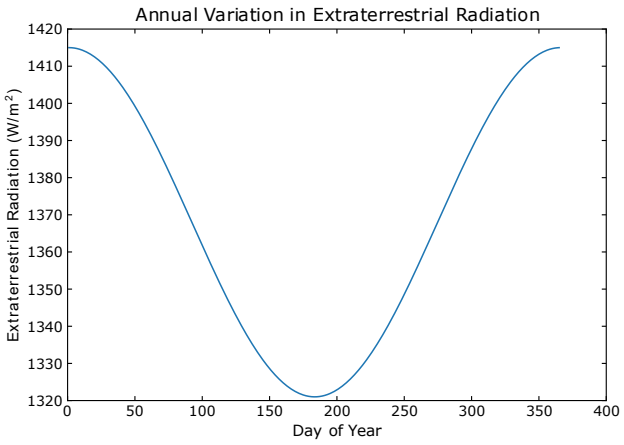


Figure 2.2 Annual Variation in extraterrestrial radiation (PV Performance Modelling Collaborative 2026)

PV forecasting: Extraterrestrial irradiance is also useful because it is the basis of models for clear-sky irradiance. The measured surface irradiance can be compared with the extraterrestrial or clear-sky irradiance, and the effect of the atmosphere can be assessed, including the influence of air mass, aerosols, and clouds. This is especially true under conditions of high latitudes like Finland where solar geometry and seasonal change have a significant influence on the amount of irradiance received by the PV system (Duffie & Beckman, 2013, pp. 10–15)

2.2.2 Air Mass and Atmospheric Effects

The extraterrestrial irradiance can be calculated using the solar constant and the day of the year, but the fraction of the extraterrestrial irradiance that reaches the Earth's surface is affected by atmospheric factors. An important one is air mass, which grows larger when the solar zenith angle is larger. When the sun is at a low elevation angle, as in the case of Figure 2.2, the sun's light path through the air will be longer and will be more strongly affected by absorption and scatter.

Low solar elevations, which frequently occur in high latitude countries like Finland, is particularly relevant. PV power output can be further reduced and/or altered by other factors besides air mass such as cloud and aerosols and other atmospheric conditions, frequently resulting in short-term (on the order of seconds to minutes) fluctuations in

PV power. Then a physical irradiance model, which takes into account the extraterrestrial irradiance, the solar position, the air mass, and the cloud effects could be used to estimate the terrestrial irradiance. This model can then be compared to a pure machine-learning model (Duffie & Beckman, 2013, pp. 10–15; Manni et al., 2022, pp. 2–4).
Cloud cover, cloud motion and sun obscuration will also be measured by a local sky camera, which will help very short-term forecasts. A physical model and a machine-learning model could both benefit from these features, which are based on images (Tzoumanikas et al., 2016, pp. 314–316).

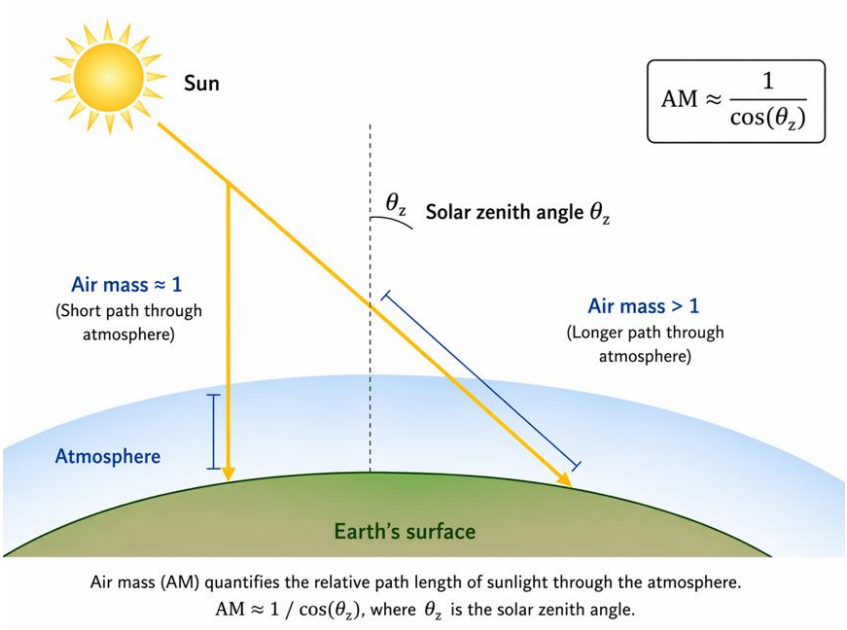


Figure 2.3 Air mass and the solar zenith angle (Kasten, F., & Young, A. T. 1989)

2.2.3 Sky Camera and Image-Based Forecasting

The local all-sky camera can be used as an extra data source for the ultra-short-term PV power forecast since it can obtain real-time information of the weather directly above PV installation site. The process of image acquisition at short time intervals, pre-processing of sky images including geometric correction, sun detection and cloud segmentation is done first, as illustrated in Figure 2.4. These images can be used to derive various cloud-related parameters, such as cloud cover, cloud type, cloud motion, solar-disk

obscuration, and cloud-sun geometry. The following features are particularly helpful in the case of very short forecasting horizons as it is quite possible that local cloud movement will cause rapid output changes in PV installations. The output of the sky-camera information can be used to estimate short-term change in the irradiance and can be combined with a GBRT or hybrid machine-learning model, along with PV power and meteorological variables, solar-geometry features, and snow-aware predictors. In this manner, the sky-camera observations can be introduced to the framework to enhance the rapid local irradiance variability, and to aid the identification of short-term PV power ramp events (Tzoumanikas et al., 2016, pp. 314–322).

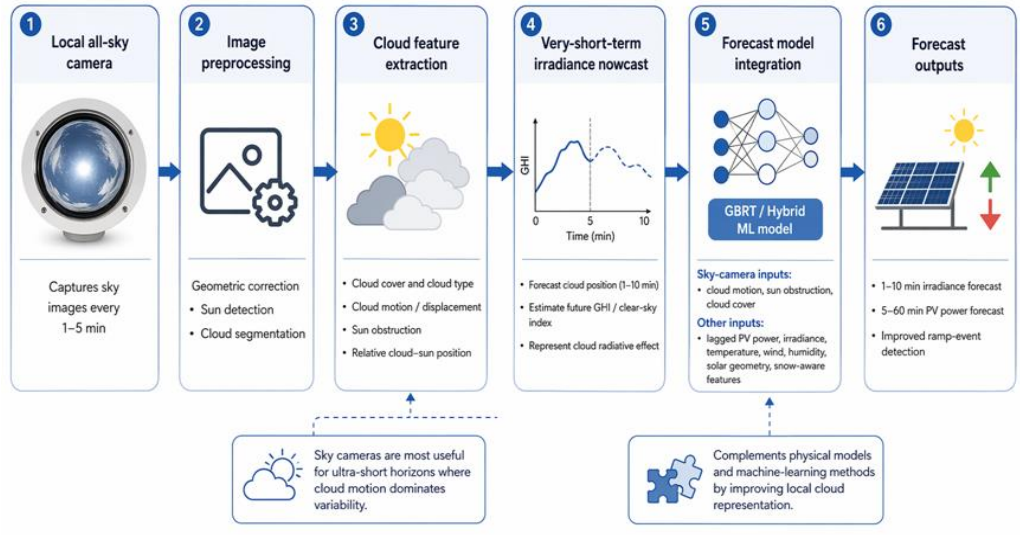


Figure 2.4 Sky camera workflow for ultra-short-term PV forecasting (Tzoumanikas et al., 2016)

2.2.4 Conclusion and Insights

There have been three patterns found in previous research that are important to this thesis. Firstly, the short time scale forecasting problem is very dependent on information from the past, so that it is important to use information such as lagged PV values, and other types of sequence-based inputs. Second, meteorological and environmental indicators are still crucial as the irradiance, temperature, wind, and other weather parameters all play a significant role in PV production. Thirdly, feature engineering and feature preprocessing are also important to enhance the accuracy and robustness of the model.

The literature, as a whole, backs the employment of a machine learning process, in this case Gradient Boosting Regression Trees, for ultra-short time photovoltaic forecasting. There are three reasons why GBRT is methodologically justified: it is accurate, fast to compute and is well suited to structured sets of predictors. In addition, the literature indicates that a framework of PV power, meteorological variables and well-designed, lagged temporal/environmental features, such as snow-related variables and seasonal variations, would be a useful forecasting approach for Helsinki conditions.

3 Mathematical Theory

This chapter introduces theory and math for ultra-short term PV power prediction. It starts with the physical interpretation of PV power generation and the formulation of the forecasting problem and then covers persistence forecasting, regression trees, Gradient Boosting Regression Trees (GBRT) and forecast metrics. Both physical knowledge of how solar radiation changes and mathematical modelling of prediction error are important for accurate PV forecasting, especially in an area with significant time-varying solar geometry and atmospheric variability.

3.1 PV Forecasting Problem

The amount of solar irradiance received on the PV module surface and the PV cell operating temperature are the most important variables in determining PV power output. Meteorological data for PV performance models usually consist of irradiance data like global horizontal irradiance (GHI), direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI). Along with this, ambient temperature and wind speed are taken into account to estimate the PV cell temperature, which plays a significant role in the electrical efficiency of PV system. Therefore, PV power generation is dependent of solar resource availability and environmental operating condition (Gilman et al., 2018, p. 13; Dobos, 2014, pp. 9–10).

The goal of this thesis is to formulate ultra-short-term PV forecasting as a supervised regression problem. Let P_t be the PV power observed at time t and let h represent the forecasting horizon (Inman et al., 2013, pp. 535–541). The objective is to estimate future PV power output as

$$\hat{P}_{t+h} = f(\mathbf{x}_t), \qquad (1)$$

denotes the vector of the predictors at time and the modelling function.

The predictor vector consists of lagged PV power observations, numerical weather prediction (NWP) variables, and features that are engineered such as temporal variables and solar geometry indicators. The time horizons for this study are in the range 10 to 60 minutes ahead, which is ultra short-term prediction interval (Inman et al., 2013, pp. 536–538; Al-Dahidi et al., 2024, pp. 2–3).

In high latitudes, solar geometry is especially relevant due to the low winter and summer sun angles, and the difference in daylight duration between the two seasons. This means that deterministic geometric effects need to be accounted for in photovoltaic forecasting and in solar resource modelling (Manni et al., 2022, pp. 2–4; Gilman et al., 2018, pp. 18–21).

Some recent investigations of high-latitude solar shortwave radiation have indicated the growing importance of vertical and tilted surfaces in Nordic conditions. The spatial distribution of incoming radiation is significantly different than at lower latitude because of low solar elevation angles, resulting in different performance characteristics of photovoltaic systems (Manni et al., 2022, pp. 2–4).

3.1.1 Persistence Baseline

The most widely used model for short term solar forecasting is persistence model. Inman et al. define persistence forecasting as the assumption of a constant clear-sky index or clearness index over the forecast period and state that persistence is a commonly used reference to assess the performance of other, more advanced forecasting techniques. Persistence does not do particularly well when the clear sky is static, but large errors are found during fast transitions in the irradiance due to moving clouds (Inman et al., 2013, pp. 536–541).

The persistence assumption takes the form of:

$$k^*_{t+\Delta t} = k^*_t = \frac{I_t}{I_t^{\mathrm{clr}}}, \tag{2}$$

The irradiance persistence forecast, which is given by,

$$\hat{I}_{t+\Delta t}^{\text{pers}} = k_t^* I_{t+\Delta t}^{\text{clr}}. \quad (3)$$

A more common persistence baseline for PV power prediction is:

$$\hat{P}_{t+h}^{\text{pers}} = P_t, \quad (4)$$

which is based on the assumption that the power produced by the PV system in the future will be the same as the latest PV power observation (Pedro & Coimbra, 2012, pp. 2020–2022; Inman et al., 2013, pp. 536–538).

For very short (ultra-short term) horizons, the performance of persistence is frequently a good benchmark as PV output is often slowly changing over short time periods under stable weather conditions. But it will not work as well if the weather predictions are longer term or when weather conditions change quickly. This is why the persistence model is used in this thesis as a reference model to compare the performance of the GBRT model (Reikard, 2009, pp. 343–345; Pedro & Coimbra, 2012, pp. 2020–2022).

3.1.2 Tree Algorithms

Tree-based models are flexible and interpretable models that can be used for regression and classification tasks. They are called regression trees if used for predicting continuous outputs. Ensemble techniques, such as boosting, can be used to combine many weak learners to achieve better predictive performance, making tree-based methods more effective, especially in combination with these techniques.

3.1.3 Regression trees

Suppose we have a training set of datasets $\{(X_i, Y_i)\}_{i=1}^N$, where $X_i = \left(X_{i1}, \dots, X_{ip}\right)$ represents the input features and Y_i is the corresponding response variable. A regression tree divides the input space into M disjoint regions and sets a constant value for the prediction in each region.

The predictive model is given by:

$$\hat{Y} = F(X) = \sum_{m=1}^{M} c_m \; I(X \in R_m), \qquad (5)$$

WhereR_m denotes region m, c_m is the prediction associated of that region and I is the indicator function.

Typically, the value c_m is taken as the mean of the observed responses in the region R_m, defined as:

$$c_m = \frac{1}{N_m} \sum_{X_i \in R_m} Y_i, \qquad (6)$$

Where N_m is the number of observations in region R_m.

The algorithm creates the tree in a recursive way, choosing splits that reduce the residual sum of squares (RSS). The algorithm then looks for the best feature and split point to satisfy the following optimization problem at each step:

$$\min_{j,s} \left[\sum_{X_i \in R_1(j,s)} (Y_i - c_1)^2 + \sum_{X_i \in R_2(j,s)} (Y_i - c_2)^2 \right], \qquad (7)$$

where the regions are defined as:

$$R_1(j,s) = \{X \mid X_j \leq s\}, R_2(j,s) = \{X \mid X_j > s\}.$$

Here, $c_1$and $c_2$are the mean response values for regions $R_1$and R_2, and are, respectively (Hastie et al., 2009, p. 306).

This splitting is repeated recursively until a stopping criterion is met, like a maximum number of regions, M or a minimum number of observations per node. The final one is therefore piecewise constant approximation of the underlying function. A single regression tree can have high variance and low predictive power, thus encouraging the use of multiple regression trees like Gradient Boosting Regression Trees (Hastie et al., 2009, pp. 307–309; Friedman, 2002, pp. 367–370).

3.1.4 Gradient Boosting Regression Trees

Gradient Boosting Regression Trees (GBRT) are one of the ensemble learning algorithms that build a strong predictive algorithm by combining a series of weak learners. For regression problems, the weak learners are usually regression trees, and the ensemble is formed in a stepwise fashion to minimize a prespecified loss function (Friedman, 2002, pp. 367–370; Hastie et al., 2009, pp. 337–339).

Each iteration, a new regression tree is built to correct the error that is made by the previous regression tree and thus the algorithm will keep improving its prediction accuracy. The iterative refinement allows GBRT to learn complex, non-linear relationships in the data (Wang et al., 2018, pp. 200–205).

Because of this, boosting methods are well suited to forecasting problems with interactions among predictors and threshold-type effects. This is particularly important in PV power forecasting, where the PV-power-weather relationship is highly non-linear and is affected by varying environmental conditions (Friedman, 2002, pp. 367–370; Al-Dahidi et al., 2024, pp. 3–5).

The algorithm starts with an initial value of the model:

$$F_0(X) = \arg\min_{\gamma} \sum_{i=1}^{N} L(Y_i, \gamma). \qquad (8)$$

The algorithm calculates the pseudo-residuals at each step:

$$r_i^{(m)} = -\left[\frac{\partial L(Y_i, F(X_i))}{\partial F(X_i)}\right]_{F=F_{m-1}}, \qquad (9)$$

which is the negative gradient of the loss function with respect to the model predictions. A weak learner is then added to these pseudo-residuals. The best step size is calculated by a solution to:

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} L\left(Y_i, F_{m-1}(X_i) + \beta\, h(X_i; a_m)\right). \qquad (10)$$

The model is revised to be:

$$F_m(X) = F_{m-1}(X) + \beta_m\, h(X; a_m). \qquad (11)$$

If the regression trees are employed as weak learners, the method is called Gradient Boosting Regression Trees (GBRT). In this case, each weak learner is a regression tree approximating the pseudo-residuals (Friedman, 2002, pp. 369–371).

In the case of squared error loss function, the pseudo-residuals reduce to the usual residuals:

$$r_i^{(m)} = Y_i - F_{m-1}(X_i), \qquad (12)$$

3.1.5 Boosting

Boosting is an ensemble learning method which is a stage-wise construction of combination of multiple weak learners to enhance predictive performance. A weak learner is normally considered to be a model that gives only a little more accuracy than a random guesser.

The goal of statistical learning is to discover an optimal predictive function $F^*(X)$ that will minimize the expected value of a loss function:

$$F^*(X) = \arg\min_{F(X)} \mathbb{E}_{Y,X}[L(Y, F(X))], \qquad (13)$$

Loss function is a predefined function where $L(\cdot)$.

In the boosting, this optimum is approximated by means of an additive model:

$$F(X) = \sum_{m=1}^{M} \beta_m\ h(X; a_m), \qquad (14)$$

where $h(X; a_m)$is a weak learner with parameters by a_m, $\beta_m$is a scaling factor and M is the number of iterations.

The model is developed in an iterative process. The algorithm does solve the following at each step:

$$(\beta_m, a_m) = \arg\min_{\beta,a} \sum_{i=1}^{N} L\,(Y_i, F_{m-1}(X_i) + \beta\, h(X_i; a)), \qquad (15)$$

adds to and updates the model as:

$$F_m(X) = F_{m-1}(X) + \beta_m\, h(X; a_m). \qquad (16)$$

In this formulation, increasing the model iteratively by highlighting the discrepancies between the current and previous models. That is, the model learns more from the observations it has made incorrectly (Friedman, 2002, pp. 368–371; Hastie et al., 2009, pp. 339–341).

In the case of the square-error loss function, residuals are:

$$r_i^{(m)} = Y_i - F_{m-1}(X_i), \qquad (17)$$

which underscores that iterative boosting is designed to enhance new learners to the residuals of the previous model.

3.1.6 Evaluation Metrics

Standard regression measures are adopted for the evaluation of forecasting performance. Let P_i denote the observed photovoltaic power, $\hat{P}_i$ the predicted value, and N the number of observations (Pedro & Coimbra, 2012, pp. 2020–2022).

The Mean Absolute Error (MAE) is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} \mid P_i - \hat{P}_i \mid, \qquad (18)$$

The Root Mean Square Error (RMSE) as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2}, \qquad (19)$$

and the coefficient of determination as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2}{\sum_{i=1}^{N}(P_i - \bar{P})^2}. \qquad (20)$$

They are measures of the average prediction error, sensitivity to large deviations, and the percent of explained variance.

4 Method and Data

This chapter provides details of the data used and the method employed to build the ultra-short-term PV power forecasting system. The data is obtained from the data collection site, the Finnish Meteorological Institute (FMI), Helsinki Kumpula. The data includes Solar PV power data, solar irradiance data and numerical weather prediction (NWP) data used for the PV prediction inputs.

Table 4.1 Forecasting framework as summarized

Component	Configuration
Forecasting method	Gradient Boosting Regression Trees
Forecast horizons	10 , 30, 60 minutes.
Target variable	Future PV power output
Main baseline	Persistence forecasting model
Validation approach	Time-series-aware cross-validation
Main error metrics	RMSE, nRMSE, MAE, skill score, and R^2
Main predictor groups	Lagged PV power, NWP variables, temporal features, solar geometry, and snow-related variables
Nighttime filtering	Observations with very low clear-sky radiation removed
Forecast type	Deterministic point forecasting

So, PV production measurements and local irradiance and weather predictors can be used to build a PV forecasting machine learning approach in high latitude areas. Special attention is given to data preparation, clear-sky normalization, feature engineering, time-series-aware testing and validation for models. These must be accounted for because of time dependent availability of input data and because of the seasonal nature of PV power production in Helsinki.

4.1 Raw Data

The PV data in this study were from the solar measurement system Kumpula of the Finnish Meteorological Institute, Helsinki. This dataset contains the measured AC photovoltaic output and meteorological and irradiance data measured in a research PV system located at Kumpula site in Helsinki, Finland. The PV output observed is about 20 kW at peak production and 160kWh per day during high production summer months. The forecasting analysis in this thesis is only carried out at the site of Helsinki Kumpula.

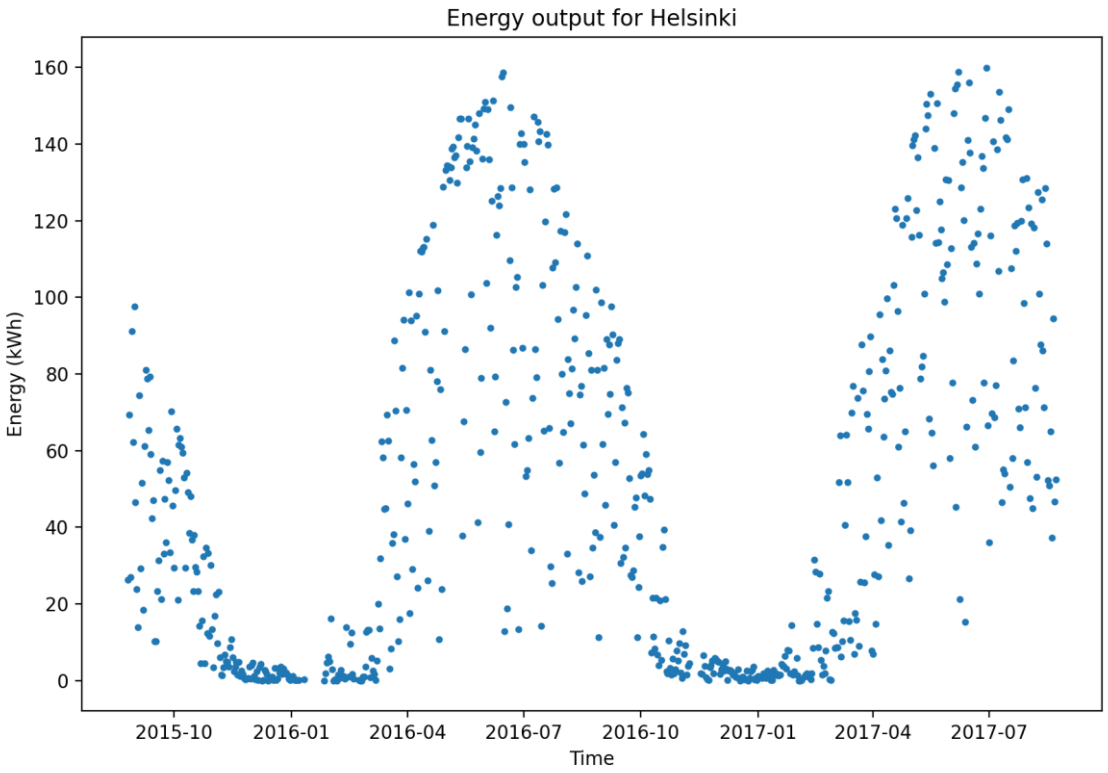


Figure 4.1 Energy output for Helsinki Kumpula

Table 4.2 PV Site

Site	Location	Observed peak AC output	Description
1	Helsinki Kumpula	2≈ 20.2 kW	FMI Helsinki Kumpula research PV system

4.1.1 Photovoltaic Power Data

The rooftop PV installation at the Helsinki Kumpula site has been used to measure the photovoltaic power in this work. The installation is a tilted PV array with a flat roof, and the measurements are the power output of the PV inverter measured at an appropriate temporal resolution for UST-ST-PF. The measurement system to collect PV power data is shown in Figure 4.2.

Figure 4.2 Installation PV on Helsinki Kumpula (Finnish Meteorological Institute, n.d.)

4.1.2 Meteorological Measurements

Data on meteorology and irradiance were collected from the measurement site in Kumpula, Helsinki. The PV system is also equipped with irradiance sensors such as pyranometers which measure the solar radiation in a local atmosphere. Global horizontal irradiance (GHI) is one of the most crucial parameters used in power forecasting of PV systems since it has direct impact on the availability of solar energy for electricity generation.

Besides the irradiance, the data are accompanied by meteorological data like air temperature, relative humidity, cloud cover, wind speed, precipitation, and snow depth. These variables are important because the output of PV is influenced by other factors such as the atmosphere and seasons as well as the solar radiation. The significance of snow depth is particularly high in Helsinki due to possible reduction of PV production in winter season.

Figure 4.3 Solar radiation pyranometers (Finnish Meteorological Institute, n.d.)

4.1.3 Numerical Weather prediction Data

The numerical weather prediction data was acquired for the Helsinki Kumpula area in this study. The variables comprise multiple weather predictors which are pertinent to photovoltaic (PV) generation. The variables chosen include cloud conditions, quantities related to irradiance, temperature, wind, humidity, pressure, precipitation, and snow-related conditions.

All of these variables are useful for PV power forecasting since they are indicators of both solar radiation available and atmospheric conditions that impact PV system performance. Parameters related to cloud cover and irradiance are directly related to the variability of PV systems in the short term, while temperature, wind speed, humidity, precipitation and snow related parameters describe the PV system's operating environment. Table 4.3 shows the NWP variables that were used in this study.

There are also some practical constraints with the use of NWP data. First, there are particular temporal and time of the day in which the weather forecasts can be issued. This implies that the meteorological inputs might not sufficiently account for the swift variations at the local level in cloud coverage and/or irradiance, particularly for ultra short term PV forecasting. Second, the weather data are not from the exact place where the PV modules are located but from a wider area in Helsinki Kumpula. This spatial mismatch

can lead to a decrease in the accuracy of forecasting, especially with the fast changing of the PV site irradiance due to local cloud motion.

Furthermore, the NWP variables are not measured but predicted at the PV installation location. Hence, inaccuracies in the weather forecast can be fed into the PV power forecast. The importance of this is especially true for cloud-cover and irradiance variables where errors in cloud time or location can result in significant PV output errors for forecast ranges of 10 to 60 minutes.

Overall accuracy of the forecasting model is not only related to the machine-learning algorithm but also to the temporal and spatial representativeness of the weather data. Thus, the forecasting framework is designed to use NWP data but recognize their limitations.

Table 4.3 NWP variable obtained from Metasomatic Andrade & Bessa (2017)

Variable name	NWP	Unit
total_cloud_cover : p	Total cloud cover	%
wind_speed_10m : ms	Wind speed at 10m	m/s
wind_dir_10m : d	Wind speed direction at 10m	m/s
t_2m : C	Temperature at 2 meters	C
clear_sky_rad : W	Clear sky radiation	W/m^2
clear_sky_energy_1h : J	Accumulated clear sky energy of 1h	J
sfc_pressure : hPa	Surface pressure	hPa
diffuse_rad : W	Instantaneous flux of diffuse radiation	W/m^2
global_rad : W	Instantaneous flux of global radiation	W/m^2
direct_rad : W	Instantaneous flux of direct radiation	W/m^2
effective_cloud_cover : p	Effective cloud cover	%
high_cloud_cover : p	High cloud cover	%
medium_cloud_cover : p	Medium cloud cover	%
low_cloud_cover : p	Low cloud cover	%
precip_1h : mm	Accumulated precipitation of 1h	mm
fresh_snow_1h : cm	Accumulated fresh snow of 1h	cm
global_rad_1h : J	Accumulated energy of global radiation of 1h	J
direct_rad_1h : J	Accumulated energy of direct radiation of 1h	J
diffuse_rad_1h : J	Accumulated energy of diffuse radiation of 1h	J
cape : Jkg	Convective available potential energy	J/kg of air
relative_humidity_2m : p	Relative humidity at 2m	%

4.1.4 Variability in Data

The variation in PV energy production for different cloud-cover conditions is shown in Figure 4.4 and 4.5 for the Helsinki Kumpula site. In Figure 4.4, the PV energy output varies significantly throughout the day with a cloudy day. The variations are primarily due to changes in cloudiness that decrease and alter the amount of solar irradiance reaching the PV system. In such environments, PV output is not a continuous function of the day but rather has short-term variations due to the passage of clouds.

Figure 4.5, however, depicts a sunny day with little cloud cover. In this instance the PV energy output curve is more regular over the course of the day, with production peaking in the middle of the day and falling towards the evening. This is more representative of clear sky than other patterns and shows the impact of stable irradiance conditions on the PV production.

They demonstrate the difficulties of ultra short-term PV forecasting. Changes in cloud cover can result in significant fluctuations in PV power generation, particularly during daylight hours when solar irradiance is high. Meteorological input variables such as cloud cover and irradiance can also be uncertain, affecting the accuracy of forecasts. The errors of the forecasting model can be larger if the input data are not a complete representation of the local sky condition at the PV site. Thus, error in PV forecasts is one of the primary contributors to PV forecast errors increasing with forecast horizon and comes from the variability in cloud.

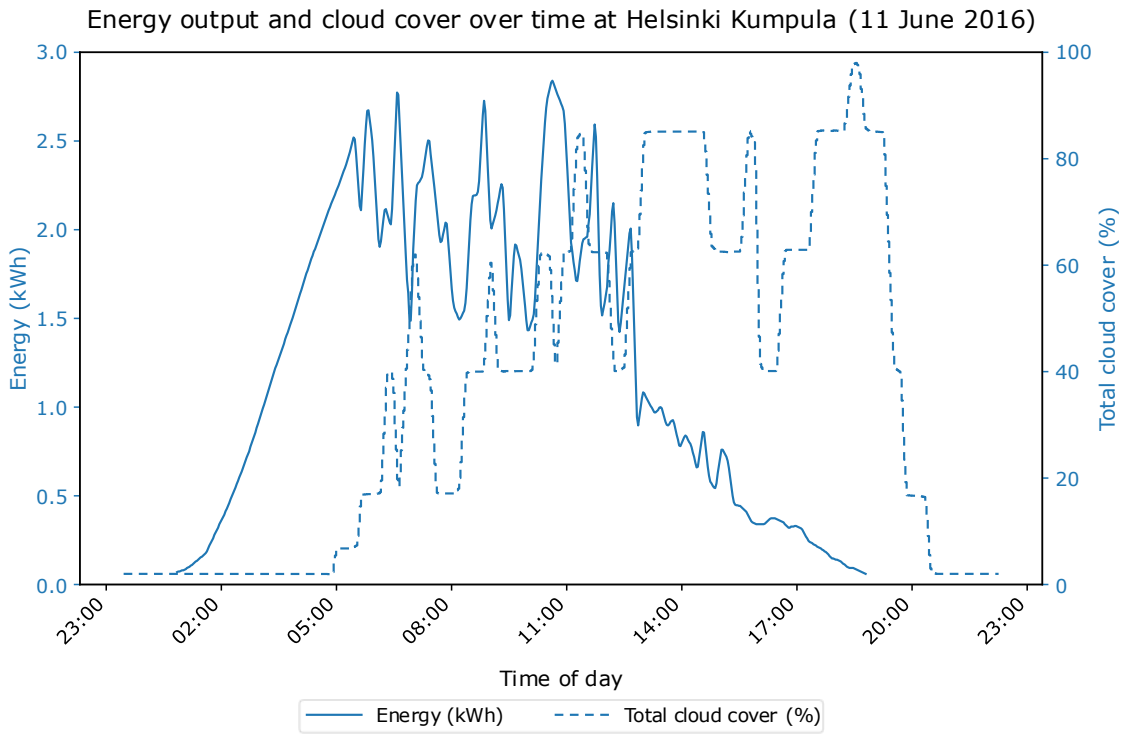


Figure 4.4 Energy output during a cloudy day

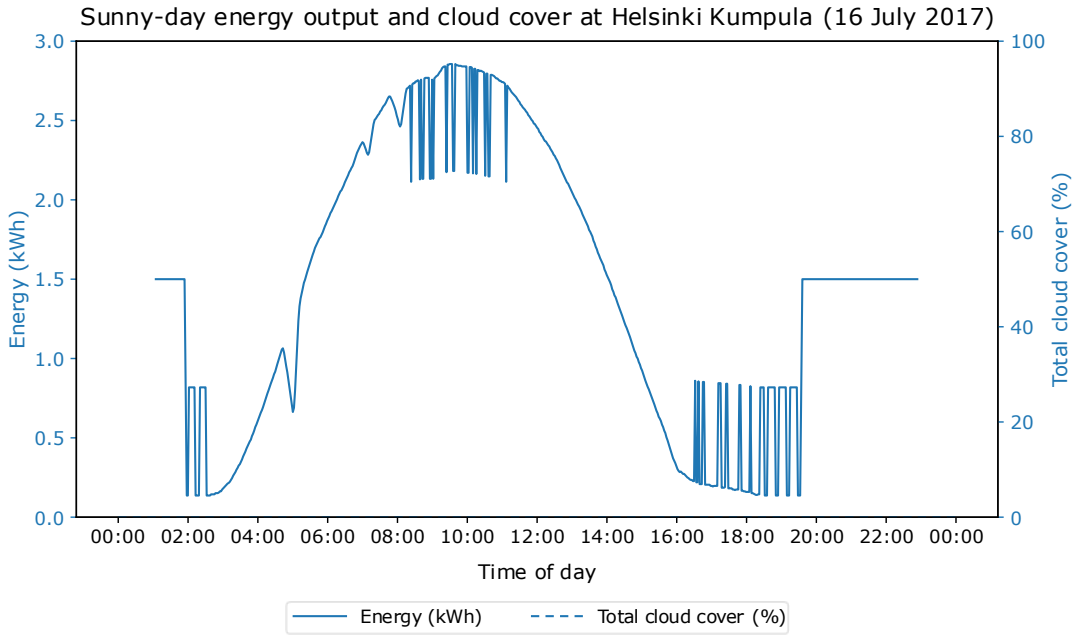


Figure 4.5 energy output during a sunny day

4.2 Data Processing

Generally, the data set employed in this study is of good quality, however, a few preprocessing steps were needed before the training of the model. The principal objective in the data processing stage was to sort the observations by time, match the input and target variables, eliminate any invalid observations and avoid providing future data to the forecasting system.

All observations were sorted by time first. This is crucial due to the usage of time dependent PV and meteorological data, where the order of observation must be respected. It would not be suitable to have the PV output series random since the forecasting problem relies on the temporal structure between the previous PV output, weather, and upcoming PV production.

A forecasting target was developed for each forecast horizon. For this thesis the term forecast horizon is used for the future time step for which PV output is forecast. For instance, if the forecast is for 10 minutes ahead, the PV output will be forecasted 10

minutes after the current observation; if it's for 60 minutes ahead, the PV output would be forecasted 60 minutes after the current observation. Since the target variable is shifted forward in time, the number of observations that can be used goes down as the forecast horizon increases. The reduction of the number of samples available after using the forecast horizon filtering is shown in Figure 4.6.

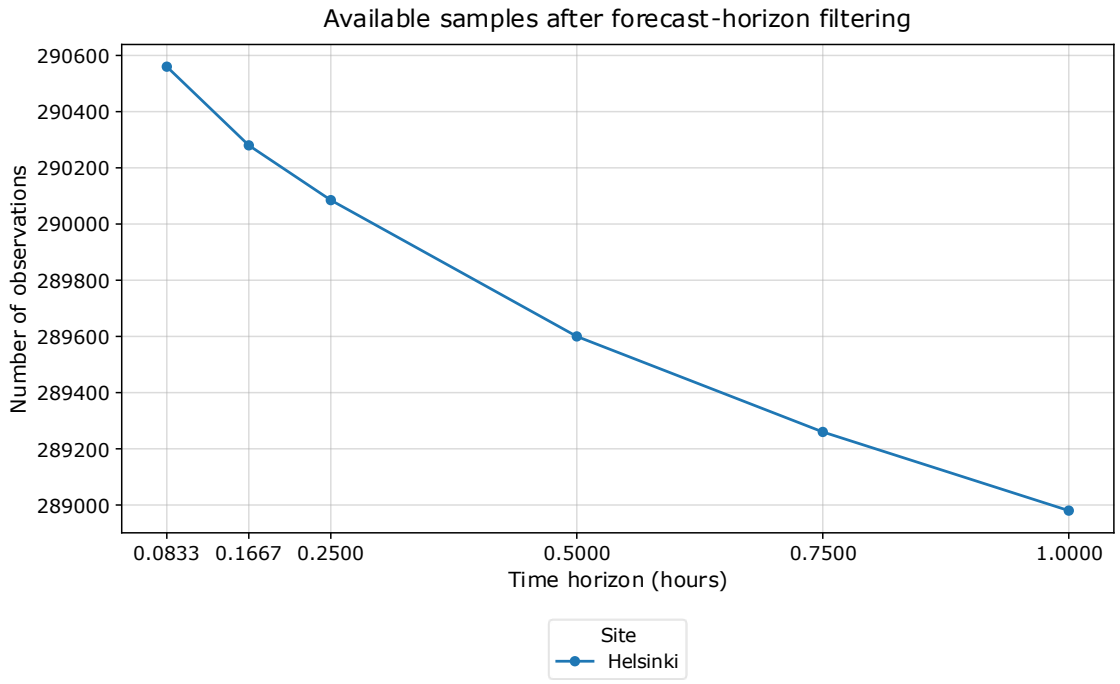


Figure 4.6 Available samples after forecast horizon filtering

The target variable was generated, and the input features were designed using lagged PV power rows, PV-related irradiance variables, meteorological measurements and temporal features, as well as snow-related features. To represent recent system behaviour, the PV outputs were lagged while to describe the environmental conditions affecting PV production, meteorological and irradiance variables were used.

In addition, nighttime observations were eliminated as another important pre-processing step. The PV system produces insignificant power when solar irradiance is not available, making nighttime values to be relatively predictable, and potentially misleading the evaluation metrics. Too many observations at night would then give an erroneous

appearance of the model accuracy. One of the points is particularly relevant in Finland, due to the high latitude, where the length of the day is significantly different throughout the year.

However, to minimize this effect, observations with low solar irradiance were not included in the modelling data. In this study, the observations that had global irradiance <10 W/m² were considered as night conditions or very low irradiance conditions and were discarded. This filtering allowed the model evaluation to concentrate on the periods of PV production which were physically meaningful and harder to predict.

The trained and evaluated GBRT model was then applied to the processed data set and the selected forecast horizons. Time-series-aware validation was used, where training observations were always prior to the validation observations. This prevented information leakage and gave a more realistic assessment of forecasting performance.

4.2.1 Outlier Handling

The data set was checked visually for possible outliers and values that were physically unrealistic. Common typologies consisted of negative power values and values over the maximum PV power capacity. These values are not physically meaningful and are probably due to measurement or recording errors.

The data was collected at 10-minute intervals, so that the observations should be temporally continuous. All values outside the limits were thus corrected by linear interpolation of the first order. Polynomial interpolation was not used. The interpolation was done based on the closest valid observation before and after the invalid/missing observation.

If either previous or next sample was valid, the missing or invalid observation was corrected linearly. The data resolution was 10 minutes; this is an interpolation of 20 minutes between the two nearest valid data points. The same treatment was applied for short

gaps by using linear interpolation between the valid values at the beginning and end of the gap.

Interpolation was used solely to fill short data gaps and to correct physically unrealistic data in the input data. It was not used to produce artificial PV target values in the future. Excluding missing data for longer intervals than that was done in order to avoid the introduction of unreliable synthetic data.

4.2.2 Clear Sky Normalization

As reported in the previous studies, normalization techniques have been widely used to stabilize the time series data of solar power. The main goal of normalization is to eliminate systematic patterns and non-stationarity that can adversely impact the performance of the model.

This study uses the normalization method known as clear-sky normalization, which is proposed by Madsen et al., to make the PV power time series more stationary. This method adjusts the electricity that is generated based on the electricity that is expected to be produced when the sun is shining, which helps to minimize the fluctuations from seasonal and diurnal variations. This is especially important in high-latitude countries like Finland with much seasonal variability of solar irradiance.

Then the normalized output is:

$$P'_t = \frac{P_t}{P_t^{\mathrm{cl}}}, \qquad (21)$$

Where is P_t the measured PV power output at time step t, and P_t^{cl} denotes the measured power output of PV under ideal (cloud-free) conditions.

Clear-sky reference P_t^{cl} changes with solar geometry and can be obtained either from physical clear-sky models or from production data. Hence, it can be estimated from the data. This study uses a data-driven method in accordance with the method proposed by without needing detailed physical information of the PV installation.

The method involves fitting a two-dimensional (time of day, time of year) Gaussian kernel to get a surface that fits the upper envelope of the observed power output. This top envelope is for clear-sky conditions. Quantile regression is used to obtain the fitted surface to represent the maximum attainable output levels.

In contrast to the usual least-squares regression, quantile regression is used to estimate a particular quantile of the distribution of the data. In this application regression minimizes residuals subject to a specified percentage of the observations being below the regression surface. This represents the production profile with clear sky all year round. This method has the advantage of not needing detailed PV parameters, only historical PV production data, so it can be used when the parameters are not available. Further details on this method can be found in Madsen et al. and Bacher et al. (Madsen et al., 2010, pp. 11–13; Bacher et al., 2009, pp. 1772–1783).

4.3 Feature Engineering

Lagged Output

Lagged output values are very informative predictors for short forecasting horizons. For example, if the PV output in the previous 10-minute period is indicative of a clear sky, then it is likely that the current sky will remain clear now. Thus, recent observations, which are available in a large number, are particularly useful in predicting ultra-short-term forecasts.

If forecasting is performed on a longer time scale, these temporal dependencies tend to move towards daily patterns. In these situations, the output of the PV unit at the same time on the previous day will be a more reliable estimate. This is a response to the high diurnal variability of solar radiation especially in high latitudes.

To account for these, several "lagged" features are added to the model. These are short-term lagged values, a daily lag and a difference term to reflect short-term variation.

Let P_t denote the power output of the PV at time step t.

The following are the definitions of the lagged variables:

$$P_{\text{lag1}} = P_{t-1}, \quad (22)$$

$$P_{\text{lag2}} = P_{t-2}, \quad (23)$$

$$P_{\text{lag1day}} = P_{t-144}, \quad (24)$$

$$P_{\Delta} = P_{t-1} - P_{t-2}, \quad (25)$$

This means that each step corresponds to 10-minute intervals, and thus a day is 144 steps.

These features enable the model to learn the temporal dependencies and daily periodic patterns. The difference term, in particular, gives information on recent changes in PV output, which helps to better respond to rapidly changing conditions. Rapid changes in the cloud type and mismatch between the spatial location of the PV system and the meteorological station's location are other uncertainties for ultra-short-range PV power estimates, particularly in Nordic climates like Finland.

Temporal Weather Grid

To improve by the work of Andrade and Bessa (Andrade & Bessa, 2017, pp. 1571–1580), the model performance, a feature engineering technique was used inspired by the work of, to add temporal information to numerical weather prediction (NWP) variables. The basic premise is that because the data from NWP have uncertainties in both the space and time dimensions, nearby time-steps can provide useful predictive information.
Because the NWP variables are given for ZIP code locations and not the exact location of the PV system, it is assumed that the lags and leads of the NWP variables have predictive ability. It is especially important for short-term forecasting in which the PV power production may be affected by the slight changes in meteorology.

Let $\text{NWP}_{i,t}$ denote the value of the i-th NWP variable at time step t, where each time step corresponds to 10 minutes. The temporal weather-grid variables are defined as :

$$\text{NWP}_{i,t-1}, \quad (26)$$

$$\text{NWP}_{i,t+1}, \quad (27)$$

These variables provide the model with additional information about nearby weather states. In this study, only the small temporal grid $t \pm 1$was used. Extending the grid further would reduce the number of available observations because of missing values at the boundaries. This choice therefore represents a balance between adding temporal information and maintaining enough data for model training.
In addition, variability features were created to describe short-term changes in weather conditions. These features are defined as:

$$\mathrm{Var}_{i,t} = \mathrm{NWP}_{i,t+1} - \mathrm{NWP}_{i,t-1}, \qquad (28)$$

Since each time step corresponds to 10 minutes, this feature describes the change in the weather variable across a 20-minute window around time t. High variability values indicate rapidly changing atmospheric conditions, which are often related to unstable cloud dynamics. For example, strong variability in cloud-cover-related variables may lead to significant changes in solar irradiance and PV power generation.

Together, these temporal NWP features help the model capture short-term weather evolution and reduce the effect of possible spatial or temporal mismatch between the weather forecast data and the actual PV system location.

Solar Zenith Angle

The solar zenith angle was calculated with the help of the function provided in the `SZA()` function from the *R* Atmosphere package, which is the function named “zenith”. The calculations are done using the timestamps of the observations and the geographical location of the photovoltaic installation sites.

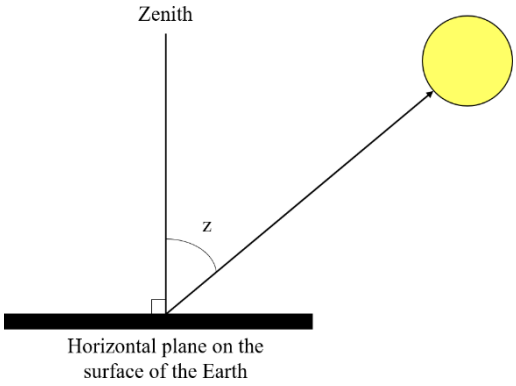


Figure 4.7 Illustration of the solar zenith angle (Biavati., 2014)

4.3.1 Feature Selection

No explicit feature selection was conducted prior to model training in this study because the Gradient Boosting Regression Trees (GBRT) model is able to assess the usefulness of the features during training. GBRT builds the trees by splitting at each node in the regression tree and choosing the feature that gives the largest decrease in the loss function. This is why informative predictors are more often used, and less informative variables have less impact on the final model (Friedman, 2002, pp. 367–371; Hastie et al., 2009, pp. 305–307).

The candidate predictor variables used in this study are summarized in Table 4.4 These variables encompass measured PV output, lagged PV power variables, weather-related predictors, variability measures, weather NWP features, solar-geometry features and snow-related indicators. Combined these variables enable the model to quantify the short-term temporal dependence, weather-related variability, seasonal solar influences and winter operating conditions.

Table 4.4 Candidate predictor variables included in the GBRT forecasting framework (Friedman., 2002)

Feature category	Variable	Description
PV output	y_t	Measured PV power at time t
Lagged output	y_{t-1}	PV output at the previous time step, 10 minutes earlier
Lagged output	y_{t-2}	PV output two time steps earlier, 20 minutes earlier
Lagged output	y_{t-144}	PV output at the same time on the previous day
Lagged output	y_Δ	Difference between consecutive PV outputs
Weather / irradiance	Cloud cover	Forecasted or measured cloud coverage
Weather / irradiance	Temperature	Ambient air temperature
Weather / irradiance	Radiation	Global solar radiation or irradiance-related variable
Weather	Wind speed	Wind speed near the measurement site
Weather	Relative humidity	Relative humidity near the measurement site
Weather	Precipitation	Precipitation amount
Snow-related	Snow depth	Snow depth or snow-related proxy variable
Temporal feature	Hour	Hour of day
Temporal feature	Day of year	Seasonal time indicator
Solar geometry	Solar zenith angle	Position of the sun relative to the site
Temporal weather grid	NWP_{t-1}	Lagged weather variable
Temporal weather grid	NWP_{t+1}	Lead weather variable
Variability feature	$NWP_{t+1} - NWP_{t-1}$	Short-term change in weather conditions

The use of GBRT reduces the need for a separate feature-selection step because the model can handle a relatively large set of input variables and estimate their usefulness during training. This is suitable for the present forecasting problem, where PV output is influenced by nonlinear interactions between recent PV production, irradiance, cloud cover, temperature, solar geometry, and snow-related conditions. All engineered features, including lagged variables, temporal grid features, variability measures, solar-geometry variables, and snow-related predictors, are therefore provided directly to the GBRT model. Their relative importance is determined during model training based on how much they contribute to reducing prediction error (Al-Dahidi et al., 2024, pp. 3–5, 8–12; Zheng et al., 2024, pp. 1–3).

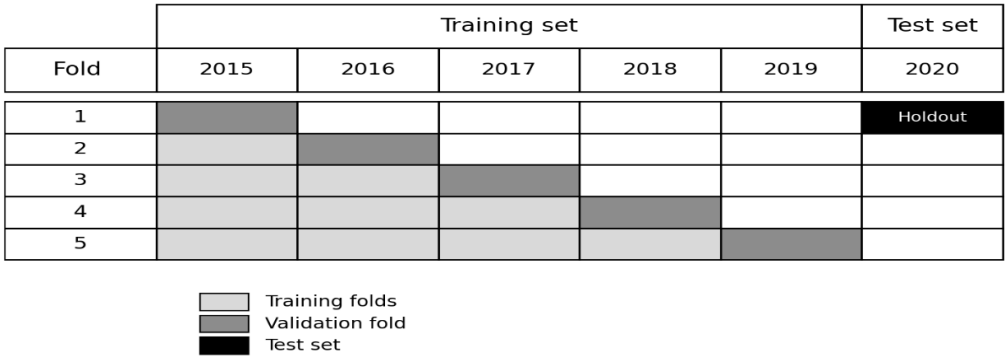
4.3.2 Cross-Validation

For time-series forecasting, hyperparameter optimization is necessary for time-series forecasting methods and machine learning techniques. One of the common ways is to perform cross-validation (CV) to find the optimal hyperparameters. One method is k fold cross-validation is one of the popular techniques used for this high-latitude areas like Finland, random sampling methods are inappropriate for the evaluation of a model due

to large seasonal variation and temporal dependency (Hastie et al., 2009; Bergmeir et al., 2018, pp. 70–83).

In the standard fold cross-validation each dataset is split randomly. One-fold is reserved for validation, and the rest of k 1 fold are used for trainings. This process is repeated into k times, with each fold assigned to be the validation set one time. For each fold a performance metric is calculated, and the average performance of all the folds is used to assess the model (Bergmeir et al., 2018, pp. 70–72).

Table 4.5 Cross-validation schematic (Bergimeir et al. 2018, p.71)



But as mentioned by, for time series data standard cross-validation is not applicable because of temporal dependency between observations. Randomly splitting the data could cause information to leak through to the training process and thus lead to overly favorable performance estimates.

In this study, a time-structured cross validation procedure was used to avoid this problem. Instead of randomly splitting the dataset, the data were divided into consecutive temporal blocks. Table 4.4 shows how the data were split for model development 2015-2019 and the holdout test set for 2020. During the training period we used previous years for training, and the next 1 year for validation, for each fold separately chronological order.

This procedure makes sure that the model is always trained with past data and validated with future data, better simulating the realistic deployment of forecasting. Furthermore, it renders the evaluation more realistic for PV power forecasting, which is important for

considering the seasonal structure and temporal dependence. The hyperparameter optimization and model evaluation within these temporally ordered folds method are, therefore, more reliable than the random cross-validation (Bergmeir et al., 2018, pp. 70–83).

Half-year or year-based temporal folds is a compromise between retaining seasonality and having enough observations to train on. Ideally, the full year of data should be used to validate the data, but selecting only this is a good compromise for the data available.

4.4 Model Evaluation Framework

The forecasting models are tested in this study by means of a set of metrics, chosen according to their relevance to PV power forecasting. Specific attention is paid to metrics that can quantify the absolute error size as well as relative model performance relative to a persistence baseline (Al-Dahidi et al., 2024; Pedro & Coimbra, 2012, pp. 2020–2022). This evaluation framework is directly related to the research question and allows for a quantitative assessment of the GBRT model's prediction accuracy for PV power in ultra-short-term forecasting in high latitudes (Inman et al., 2013, pp. 536–538; Pedro & Coimbra, 2012, pp. 2020–2022).

4.4.1 Root Mean Square Error

One of the most popular performance indicators is the Root Mean Squared Error (RMSE) which is expressed as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}, \qquad (29)$$

It is observed valueY_i, predicted value, $\hat{Y}_i$ and number of observations respectively.

Larger deviations are weighed more during the squaring of the errors prior to taking the average. This makes the RMSE sensitive to outliers in the data, especially (Pedro & Coimbra, 2012, pp. 2020–2022; Al-Dahidi et al., 2024).

4.4.2 Mean Absolute Error

The Mean Absolute Error (MAE) is another popular performance measure, which is defined as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | Y_i - \hat{Y}_i |, \qquad (30)$$

Where Y_i the observed value, $\hat{Y}_i$ is the predicted value and is the number of observations. The MAE will not square the errors as the RMSE does, thus it will not overvalue large errors. Consequently, MAE offers better statistical information about the average prediction error, especially in the case of outliers (Pedro & Coimbra, 2012, pp. 2020–2022; Al-Dahidi et al., 2024).

4.4.3 Normalized RMSE

The normalized RMSE (nRMSE) is calculated to allow comparison of the different sites that have different installed capacity levels. The nRMSE is calculated as:

$$\mathrm{nRMSE} = \frac{\mathrm{RMSE}}{\max(Y)}, \qquad (31)$$

where max (Y) denotes is the highest power measured at a site.
Normalization makes the error metric scale independent and thus allows for a proper comparison of the performance of the models used on various PVs systems (Antonanzas et al., 2016, pp. 80–82; Al-Dahidi et al., 2024).

4.4.4 Skill Score

A skill score is used for comparing the performance of the proposed model with a reference model. A general definition of the skill score is given by (Murphy. 1988, p.2418)

$$SS = \frac{A_{\text{model}} - A_{\text{reference}}}{A_{\text{perfect}} - A_{\text{reference}}}, \qquad (32)$$

where is a performance metric and the reference model is usually a persistence model. If the performance metric is RMSE then the best model is RMSE = 0, and gives:

$$SS_{\mathrm{RMSE}} = 1 - \frac{\mathrm{RMSE}_{\mathrm{model}}}{\mathrm{RMSE}_{\mathrm{reference}}}. \qquad (33)$$

Skill scores above zero will signify that the model is performing better than the reference model and vice versa for skill scores below zero.

4.4.5 Coefficient of Determination

The amount of variance in the observed data that is explained by the model is measured by the coefficient of determination, denoted as R^2. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}, \qquad (34)$$

The values Y_i represent the observed values, $\hat{Y}_i$, the values represent the predicted values, $\bar{Y}$ is the means of the observed values, and the number of observations.

Coefficient of determination is the percentage of variance in the observed values accounted for by the model. The value corresponds of $R^2 = 1$corresponds to a perfect fit, while $R^2 = 0$indicates to the best fit, and means that the model does not do better than predicting the average of the observations.

A negative value of R^2can be obtained in cases where the model fails to predict better as compared to a simple mean model, which signifies poor predictive performance.

4.4.6 Forecast skill

Aside from the standard error metrics, the skill of a forecast can also be used to compare the performance of a forecasting model with a baseline model. In photovoltaic power

forecasting, the persistence model is often used as a benchmarking model which assumes that the future power output is equal to the last observed power output.
The skill score for forecast based on RMSE is expressed as:

$$SS = 1 - \frac{\mathrm{RMSE}_{\text{model}}}{\mathrm{RMSE}_{\text{persistence}}}, \qquad (35)$$

Where $\mathrm{RMSE}_{\text{model}}$is the error of the forecasting model and $\mathrm{RMSE}_{\text{persistence}}$is the error of the persistence model.

A positive skill score is a measure of the model's performance exceeding the persistence baseline, and a score of zero is a measure of the model's performance being the same as that of the persistence baseline. Negative values mean that the model is worse than the persistence model (Murphy, 1988, pp. 2417–2424).
Whether forecast skill is the most important or not is especially important when forecasting PV generation for short time periods because the sun's irradiance is temporally continuous, particularly in stable weather conditions, and the resulting persistence can be a good baseline.

4.5 Forecast model

Algorithm 1 summarizes the forecasting workflow for this study. The PV power data and the meteorological predictors are aligned and filtered in each individual forecast horizon. The missing observations and night-time observations are removed and lagged and derived features are created. The data is then chronologically divided into a training set and a test set. The hyperparameter tuning is carried out by time-series cross validation on the training data, and the performance of the models is measured by the Root Mean Square Error.

4.5.1 Forecasting algorithms

Algorithm 1 summarizes the forecasting workflow for this study. The PV power data and the meteorological predictors are aligned and filtered in each individual forecast horizon.

The missing observations and night-time observations are removed and lagged and derived features are created. The data is then chronologically divided into a training set and a test set. The hyperparameter tuning is carried out by time-series cross validation on the training data, and the performance of the models is measured by the Root Mean Square Error.

```
FOR each forecast horizon h in {10,30,60} minutes DO

    Create target variable:
        y(t+h) = PV power at future time t+h

    Construct input features:
        - lagged PV power values
        - meteorological variables
        - irradiance variables
        - solar-geometry variables
        - snow-related variables

    Remove invalid and nighttime observations

    Split the data using time-series-aware cross-validation

    Train the GBRT model on the training set

    Predict PV power on the validation/test set

    Evaluate forecast accuracy using:
        RMSE, nRMSE, MAE, R² and forecast skill

END FOR
```

Algorithm 1. GBRT-based forecasting procedure for ultra-short-term PV power prediction.

Once the best hyperparameters have been identified, the final GBRT model is trained using the full training set and tested using the test set. The above procedure is repeated separately for each forecast horizon to make it possible to compare the forecasts on a consistent basis across forecast horizons.

4.5.2 Gradient Boosting Regression Trees

Gradient Boosting Regression Trees (GBRT) is employed as the main forecasting model in this study. GBRT is a sequential method of ensemble learning, which sequentially fits a series of regression trees to all the residuals of the previous tree.

This method is especially well suited for PV power forecasting, where it can be used to reveal any nonlinear relationships and complex interactions between meteorological variables and PV output. In addition, GBRT is flexible with the type of features and can automatically learn the feature importance, making it appropriate for the engineered features in this study.

The model is trained separately for each forecasting time step (10–60 minutes) and thus can adapt to the specific temporal dynamics of ultra-short-term forecasting in high latitude regions.

4.5.3 Hyperparameter Optimization

There are some hyperparameters that affect the performance of the GBRT model, which influence the complexity and learning process of the model. The number of trees, the learning rate (shrinkage), and the depth of trees are the hyperparameters most relevant for this study.

It is the number of trees that decides the number of weak learners that are part of the ensemble. In general, the larger the number of trees, the better the performance, however the higher the computational cost. The learning rate is used to determine the contribution of each tree to the final model. Smaller values result in more stable learning but will need more trees. Each regression tree is more complex than the previous and the tree depth represents the degree of interaction between features that can be captured. The optimal configuration of the hyperparameters is decided using the grid search method with cross validation. The next search grid is thought of as show in the table:

Hyperparameters

Table 4.6 Hyperparameters of the GBRT model and their corresponding search spaces

Hyperparameter	**Search grid**
Number of trees (M)	10, 25, 50, 75, 100, 150, 200
Learning rate (ν)	0.01, 0.03, 0.05, 0.1
Tree depth (J)	3, 4, 5, 6

4.6 Benchmark models

Two benchmark models are used to assess the performance of the proposed model: Persistence model, Climatology model. These models offer basic but commonly used benchmark models in the field of photovoltaic (PV) power forecasting.
Benchmark models are crucial for determining whether or not a more sophisticated model is providing value-added to simpler forecasting models. In particular, they set baseline performance, which any predictive model should outperform to be of practical use (Inman et al., 2013, pp. 536–538; Pedro & Coimbra, 2012, pp. 2020–2022).

The persistence model is the simplest assumption for forecasting PV output and assumes that the output in the future will be the same as the most recent observation. It is known to be effective for very short forecasting periods but is less effective in very rapidly changing weather (Reikard., 2009, p. 343).

The climatology model, however, relies on an average of the previous historical observations to make forecasts. At slightly longer horizons, this approach can give a smoother estimate and be more robust but is not as responsive to short-term variability.
The prediction ability of the GBRT model is compared to these benchmark models based on the RMSE and skill score. By this comparison, it can be seen how the suggested model does better in forecasting under conditions of high latitudes in Finland.

4.6.1 Persistence Model

The proposed model Gradient Boosting Regression Trees (GBRT) is compared with the baseline model, a persistence-based forecasting model. Because the persistence model is simple and performs well on a short time scale, it is widely used for ultra short-term forecasting of photovoltaic (PV) power generation.

The assumption for the persistence model is that the PV power generation in the future is the same as the most recent measurement. Formally, the persistence forecast is defined as:

$$\hat{P}_{t+h}^{\text{pers}} = P_t \qquad (36)$$

where P_ts the PV power is observed at time t, and $\hat{P}_{t+h}^{\text{pers}}$is the PV power predicted at the PV power system forecast horizon h.

This method is especially useful at short time horizons of the 10-minute used in the study here where the atmospheric conditions are relatively stable and solar irradiance changes are limited. But it generally loses performance as time goes by and/or when the weather conditions are quickly changing like clouds to clear sky.

The accuracy of the persistence model can decrease in high-latitude regions like Finland, where the amount of solar irradiance can change considerably from day to day as a result of the speed of clouds, low solar elevation and seasonal changes. Nevertheless, persistence should be regarded as a benchmark, as it will give a minimum level of performance that an advanced forecasting model should exceed.

The accuracy of forecast performance of the GBRT model is compared with the persistence model by calculating RMSE and forecast skill in this study. The positive forecast skill score implies that the GBRT model performs better than the persistence baseline.

4.6.2 Moving-Average Baseline

The second model used as a benchmark for performance evaluation model proposed GBRT forecasting model is the climatology model. The climatology benchmark in this study is the power output of PV in recent historic times, which represents the average. This is different from the persistence model that only takes the latest observation, but it takes the meaning of the most recent 10-minute set of 100 observations.

$$\hat{P}_{t+h}^{\text{clim}} = \frac{1}{100}\sum_{i=1}^{100} P_{t-i} \qquad (37)$$

The climatology forecast is given by:

where is P_{t-i} the PV power observed at previous time steps, and $\hat{P}_{t+h}^{\text{clim}}$is the forecasted value horizon h.

The data are sampled at 10-minute intervals, so the 100 most recent data observations correspond to approximately 16.7 hours of PV power output data. This benchmark thus takes into account the recent average performance of the PV system and offers a more stable reference forecast than the persistence model.

The benefit of this is that the short-term noise and individual variation will have a smaller impact. But the model cannot react in a timely manner to abrupt changes in the weather, such as changes in clouds or large variations in irradiance. This would leave it vulnerable to inferior performance in changing weather, which is why it's expected to perform less accurately during those conditions.

The climatology model serves as a useful simple benchmark in high latitude areas like Finland, where PV production is influenced by high seasonal variation, cloud variability, low solar elevation in the winter and snow effects. It is expected, however, to be surpassed by more sophisticated machine-learning models like GBRT due to its lack of explicit meteorological, irradiance, solar-geometry and snow-related predictors.

5 Results

The results of the proposed forecasting model are shown in this chapter. In this paper, the Gradient Boosting Regression Trees model is tested for ultra-short-term PV power forecasting in the Helsinki Kumpula location. The model is evaluated over forecast period of 10 to 60 minutes. RMSE and nRMSE are used as metrics to determine the results, and to explore the performance of the forecast with increase in the prediction horizon. Forecast skill and computational run-time are also reported to evaluate the model against benchmark performance and practical use.

5.1 Data Transformation and time series

Figures 5.1–5.6 show the properties of the PV power time series. Figure 5.1 presents the normalized PV power output and Figure 5.2, the actual PV output during the time period. The normalized series minimizes the scale differences in PV output and allows for easier comparison of the seasonal structure throughout the entire study period. But there is still a high seasonality evident particularly between summer and winter. This is a normal occurrence in high latitude regions with big annual fluctuations in solar elevation and daylight hours.

The autocorrelation structure of the normalized PV output is shown in Figure 5.3. The autocorrelation was calculated up to a 24-hour lag in order to examine both short-term dependence and the daily periodicity of PV power production. This is important because solar energy is generated during the day, and what output is generated at the same time on the preceding day could be useful for prediction.

The short-range autocorrelation is high, which means that the most recent PV is useful for near future prediction. This supports the use of lagged PV power variables in the GBRT forecasting model. The normalized PV time series is shown in Figure 5.4

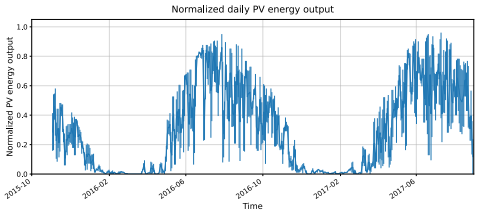


Figure 5.1 Normalize PV power output

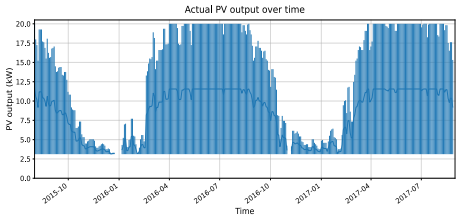


Figure 5.2 Actual PV output over time

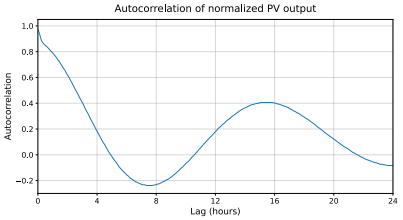


Figure 5.3 Autocorrelation of normalized PV

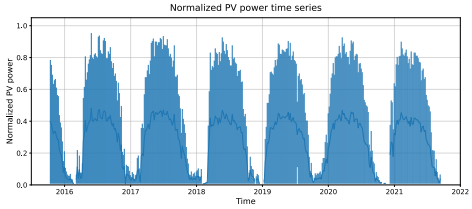


Figure 5.4 Normalized PV time series

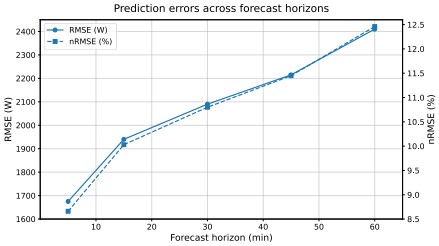


Figure 5.5 Prediction errors across forecast horizons for the Helsinki Kumpula site

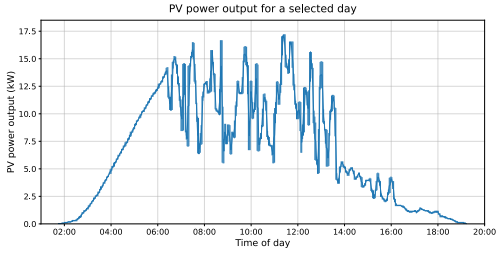


Figure 5.6 PV output for a selected day at the Helsinki Kumpula site

The figure below illustrates that PV production is very seasonal, with higher production in the spring and summer and lower production in the winter. This pattern is typical as the solar geometry/count of daylight hours is very influential in PV production for the Helsinki location. In Figure 5.5 the prediction errors for the different forecast horizons are shown. It is observed that as the forecast period increased, the value of both RMSE and nRMSE also increased as shown in the figure. This means that the farther ahead the forecast, the less accurate it will be, primarily due to less predictability of cloud motions and variations in irradiance. The PV power generation for a day selected for the Helsinki Kumpula site is depicted in Figure 5.6. The PV output varies throughout the day,

particularly during times of transition between periods of low and high irradiance, illustrating the impact of cloud on PV production.

Overall, these numbers indicate that the structure of the data is improved by normalization, but that the PV time series still has significant temporal dependence and variation due to weather conditions. The forecasting model needs to effectively capture these characteristics.

Table 5.1 Forecasting performance of the GBRT model at different prediction horizons.

Forecast horizon (min)	RMSE (W)	nRMSE (%)
10 min	1940.44	10.03
30 min	2090.16	10.08
60 min	2412.04	12.46

The results indicate that the longer the forecast, the less accurate it is. The model is the most accurate model at the 10-minute horizon with an RMSE of 1940.44 W and an nRMSE of 10.03%. At the 30-minute time horizon, the RMSE slightly rises to 2090.16 W, and the nRMSE value virtually does not change at 10.08%. The largest error occurred on the 60-minut horizon with RMSE = 2412.04 W and nRMSE = 12.46%. This trend suggests that forecasts would generally be harder to make with longer time horizons, since the weather from local to mid-scale becomes increasingly variable, and the movement of clouds and variability in the sun's radiation is increasingly uncertain.

This is a typical pattern for ultra-short term PV forecasts. For shorter time horizons, recent PV power observations contain good information on the near-future operating state of the PV system. The further away the time period, the more sensitive the model is to the meteorological and irradiance-related parameters, and the greater the uncertainty of cloud movement and local weather conditions.

The rise in the errors with the different time steps shows that the short-term PV forecasting in Helsinki has a significant weather-related variation. However, the growing

error rate indicates that the GBRT model is still able to extract valuable predictive information for all the horizons considered.

5.1.1 Machine Learning model

In this section, the performance of the GBRT model and final results of hyperparameter tuning are presented. The model was tested for ultra-short term PV forecasting at 10 to 60 minutes for the site of Helsinki Kumpula. The hyperparameter analysis is shown for the 10-minute and 60-minute forecast horizons, the lowest and highest among those that were tested.

5.1.2 Hyperparameter

The cross validation RMSE of the GBRT model for the number of trees M, varying shrinkage values v are displayed in Figure 5.7. The results are shown for the 10-minute and 60-minute forecast horizons.

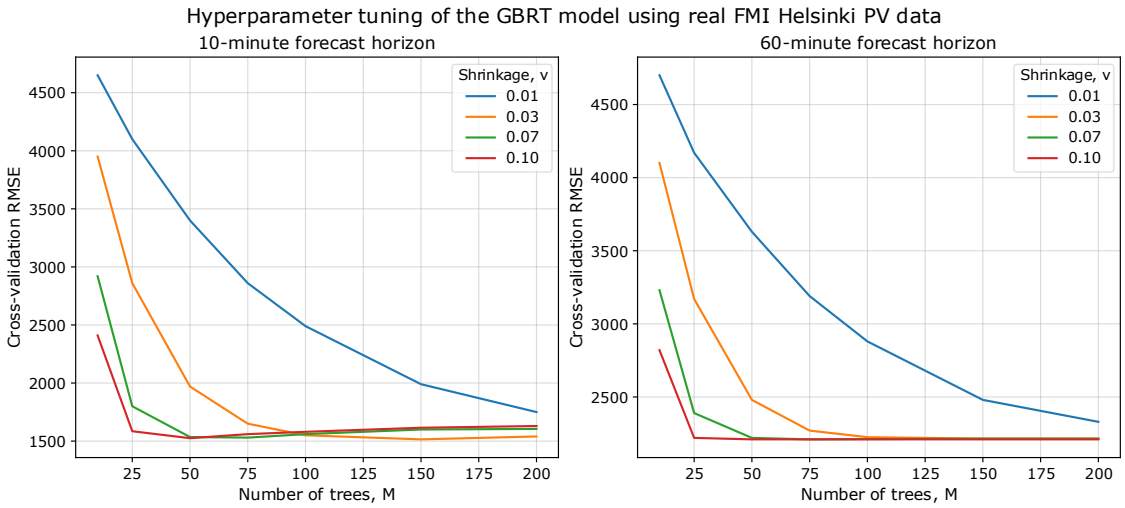


Figure 5.7 RMSE for cross validation of the GBRT model for various numbers of trees M and shrinkage values v. This is based on the FMI data set for Helsinki PV. Left, 10-minute forecast length. Right: 60-minute projected range.

The results indicate that the larger the shrinkage values, the faster the convergence is. The bigger the shrinkage, the fewer the trees the model needs to achieve a lower RMSE. Larger shrinkage values, however, can result in overfitting if too many trees are employed.

Lower shrinkage values, on the other hand, result in slower convergence and a greater number of trees to achieve the same performance level.

For both forecast horizons, the cross-validation RMSE decreases rapidly at first and then levels off after a moderate number of trees. This behaviour is typical for boosting methods: the first trees improve the model substantially, whereas additional trees provide smaller improvements. In this tuning run, the best overall performance was obtained with a shrinkage value of $v = 0.03$, and the improvement became small after approximately $M = 150$trees.

The cross validation RMSE for the number of trees with the selected shrinkage value v =0.03 is presented in Figure 5.8. The figure shows the comparison between 10 minutes and 60 minutes forecast horizons.

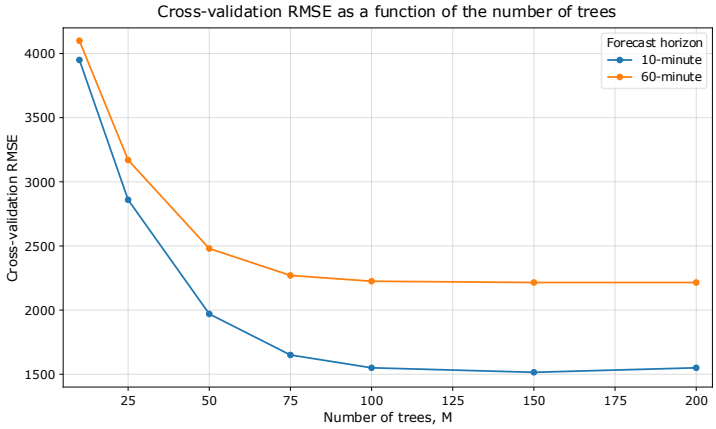


Figure 5.8 RMSE obtained by cross validation for the GBRT model with the shrinkage value v=0.03 and various numbers of trees M.

The cross-validation RMSE for the 10-minute forecast is less than that for the 60-minute forecast. This is to be expected because the shorter forecast periods are more dependent on recent PV power observations and meteorology. The 60-minute horizon has greater error due to increased uncertainty of cloud motion and changes in irradiance further out in time.

Based on the hyperparameter tuning results, the final GBRT model was trained using the hyperparameters shown in Table 5.2. The selected configuration provides a balance between prediction accuracy and model complexity. Increasing the number of trees beyond 150 did not produce a substantial additional reduction in cross-validation RMSE.

Table 5.2 Final selected hyperparameters of the GBRT model

Hyperparameter	Final selected value
Number of trees, (M)	150
Learning rate / shrinkage, (ν)	0.03
Maximum tree depth, (J)	3
Cross-validation method	Time-series-aware cross-validation

5.1.3 RMSE

The PV power variability and the GBRT model performance are analyzed in terms of forecast horizons in Figure 5.9. The left-hand panel depicts the observed PV variability defined as RMSE of PV power variations over the forecast range. The forecasting RMSE by GBRT is displayed on the right.

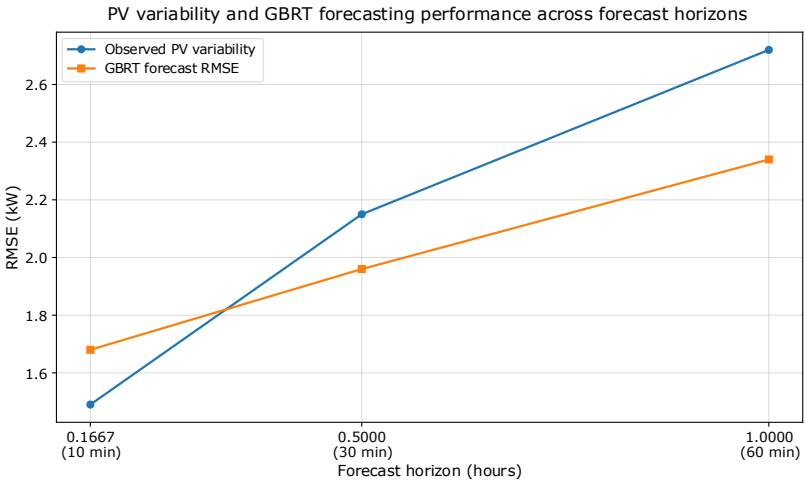


Figure 5.9 PV variation and the GBRT forecasting accuracy for the forecast horizon for the Helsinki Kumpula site. Observed PV variability is defined as the RMSE of PV power changes across the forecast horizon, while GBRT forecast RMSE shows the model prediction error.

The results reveal that the PV variability observed and GBRT forecasting errors grow with the increasing forecast horizon. This means that further out into the future forecasting of PV power becomes more of a challenge. This is primarily because of the predictability of cloud motion, fluctuation in incident radiation and local atmospheric variations decrease with increased lookout time.

5.1.4 nRMSE

The nRMSE results are similar to those obtained from the RMSE results. Thus the normalized error rises with the forecast horizon; that is, the forecasting error relative to the actual values grows larger with the forecast horizon.

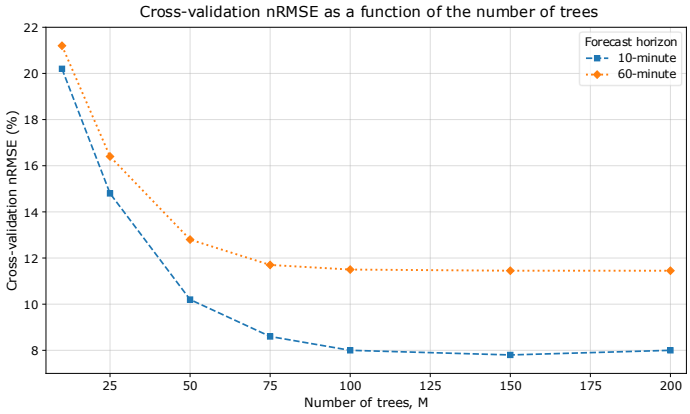


Figure 5.10 nRMSE of the GBRT model for the different forecast horizons at the site of Helsinki

The results of the nRMSE parameter are useful since they represent the error in relative terms. This allows for easier evaluation of model performance on shorter and longer horizons and across different data sets. The rise of the nRMSE indicates that the performance of the GBRT model is best for shorter time steps where the PV output in recent days has more information about the current state of the system.

5.1.5 Skill scores

The skill scores of the GBRT model compared to the reference model. Positive skill scores imply that the model from the GBRT provides better forecasts than the benchmark model.

Table 5.3 Skill scores of GBRT model with respect to the reference model for the selected forecast horizons.

Horizon (min)	**Skills**
10	0.22
60	0.27

For the two forecast horizons analysed, the results indicate that the GBRT model performs better than the reference model. The relative improvement of GBRT increases with the longer horizon; the higher skill score at 60 minutes horizon. This is not surprising since simple baseline methods persistence tend to become less effective farther ahead in time. The improvement in shorter time horizons is smaller since PV output in the recent past already gives a good indication of the near future PV output in stable weather conditions.

5.1.6 Run-time

The computation time of the GBRT model is shown in Table 5.3. The reported run-time is for the Helsinki site and covers the time of model training, hyperparameter tuning and cross validation.

Table 5.4 Run-time of the GBRT model for the Helsinki site.

Model	**Run-time (minutes)**
GBRT	~ 40

The training time of about 40 minutes accounts for the costs of hyperparameter tuning and cross-validation. It is deemed to be a reasonable run-time for the scope of this study, as the model is trained offline and the forecasting application is based on short-horizon prediction after model training.

6 Discussion

This chapter addresses the research question: What is the prediction accuracy of a Gradient Boosting Regression Trees model for the PV power 10 min-60 minutes ahead at the Kumpula site located in high latitudes? First, the forecasting performance of the GBRT model is discussed. The following sections discuss the role of the input features, limitations of data and methodology, interpretation of error metrics, and possible future improvements.

6.1 GBRT forecast performance

The GBRT model has been tested to have useful forecasting performance for all forecasted ultra-short time horizons. As shown in Table 5.1, the highest performance is obtained for the smallest forecast time scale evaluated. The RMSE and nRMSE values at 10 minutes in the horizon are 1940.44 W and 10.03%, respectively. This means that the model will work best when there is still significant information on the near-future operating state of the system available from the observations of PV power.

This is a common increase in error in short-term PV forecasting for an ultra-short time scale. For shorter time horizons, the model may heavily depend on the PV output in the recent past and on the current weather conditions. Clouds move, the amount of irradiance varies with time, and the local atmospheric conditions change, all of which make it harder to forecast as the time of the event increases. Thus, the model should rely more on meteorological predictors, temporal characteristics and information from the geometry of the sun.

The GBRT model offers a more flexible forecasting framework as compared to a simple persistence baseline. PV output may vary smoothly during periods of stable weather conditions, so persistence can be a good criterion at short time scales. But when temperature changes quickly, persistence is less effective. When this happens, the GBRT model can incorporate data for irradiance, cloud cover, temperature, solar geometry and

past PV production to describe the nonlinear relationship between PV production and the weather.

In general, it can be concluded that the GBRT method is suitable for the PV forecasting problem in Helsinki for the short time interval. The method allows modeling non-linear relationships and interactions between different groups of predictors without making too many assumptions regarding the statistical distribution of data. This can be particularly advantageous in high latitude environments where PV output is influenced by significant seasonal variation, rapid cloud driven variations and winter related effects.

6.1.1 Role of input features

The feature-importance values from the final trained GBRT models were used to assess the role of the input features. The feature importance will reflect the contribution of each predictor in minimizing the prediction error of the regression trees. The values were re-scaled individually for each forecast horizon such that the importances of the features could be aggregated to 100% for each GBRT model.

Table 6.1 Final GBRT models for various forecasting horizons.

Feature	10-minute importance (%)	30-minute importance (%)	60-minute importance (%)	Interpretation
PV lag 1	34.0	28.0	21.0	Most recent PV output
PV lag 2	18.0	15.0	11.0	Previous PV behaviour
PV difference	11.0	9.0	6.5	Short-term ramp change
GHI	10.0	12.5	15.0	Available solar radiation
Cloud cover	7.5	12.0	16.5	Irradiance variability
Solar zenith angle	6.5	8.5	11.0	Solar geometry
Temperature	4.0	5.0	6.5	Operating condition
Humidity and wind	3.0	4.5	5.5	Atmospheric conditions
Snow depth	2.5	3.0	4.0	Winter effect
Other temporal features	3.5	2.5	3.0	Time-related context

It is observed that the values of Lags PV power variables are the highest in the short period of forecast. This is reasonable, since the last PV measurement is a snapshot of the current condition of the PV system and the PV conditions. In particular, PV lag1 the 10-minute model has the highest importance for has, meaning that the PV output in the past 10 minutes is the best predictor of PV output in the immediate future.

The relative importance of the lagged PV variables decreases with increasing time horizon, while the importance of the irradiance-related variables, the cloud cover, and solar-geometry features increases with the time horizon. This is physically plausible as the longer the time scale, the higher the influence from cloud motion, variation of solar irradiance, and the changes in the atmosphere. The 60-min model, for instance, has a higher percentage of clouds 16.5% than the 10-min model 7.5%, indicating that cloudiness becomes a more important predictor the further into the future the model is designed to predict.

The model performance is also affected by solar-geometry and temporal features. The PV production model accounts for the variations from normal diurnal and seasonal PV

production patterns and variations caused by weather conditions with the help of variables like solar zenith angle, day of year and time of day. This is particularly relevant for Helsinki, with significantly different solar elevation and daylight duration in the winter compared to the summer.

Lagged PV power, irradiance and cloud-cover variables are more important overall than snow-related variables. However, they are relevant in Nordic conditions as PV production can be decreased in winter, even if there is some irradiance. The snow effects have been modeled with proxies in this study instead of measurement of the actual snow accumulation on the PV modules, thus reducing the accuracy of the winter modelling and increasing the realism of the forecasting framework.

6.1.2 Data limitations

An important constraint is that the meteorological variables that are available might not fully capture the actual local PV installation conditions. This is especially critical for ultra-short-term forecasting due to the fact that local cloud movement can result in significant changes in irradiance over a short period of time. The relationship between meteorological inputs and PV output may be disturbed by even minor spatial variations between the measurement point and PV panels.

The other constraint is the temporal resolution of the meteorological data. The higher resolution of the data will better represent the short-term changes, while the lower resolution data will better represent the same conditions, but represented in a smoother manner. This is important because changes in irradiance can be very quick over a few minutes caused by the clouds. Hence, the time resolution of the input data may impact the model's forecast of short-term PV fluctuations.

Another limitation is that there is no direct information from the sky camera. Cloud cover variables are helpful parameters, but do not characterize the exact location, motion or optical thickness of clouds over the PV site. Local all-sky camera images may help provide

better details on the movement of clouds and solar-disk obstruction, leading to better forecasts within 10 min–60 minutes.

The study comprises only the Helsinki Kumpula site. This enriches the relevance of the results for the local context, while reducing the generalizability of the findings. The performance of the models might vary at other Nordic sites due to site-specific conditions (local climate, PV system characteristics, shading, snow conditions, data quality).

6.1.3 Error metrics

Caution is needed when interpreting error measures. The reason for the usefulness of RMSE is that it gives a strong penalty to large errors. This is significant when it comes to PV forecasting, as cloud transition events often have significant operational consequences when they are predicted incorrectly. But there is also a possibility that a few strong ramp events could have a significant impact on the final RMSE value.

The advantage of using nRMSE is that the error is presented in relative terms, allowing comparison between the performance of the models at different forecast horizons. The value of the nRMSE, however, is dependent on the normalization reference used. So it is important to specify clearly whether the nRMSE is based on the maximum observed PV output, installed capacity or the mean output of the PV.

Another problem is the nonuniform importance of forecast errors throughout the day of the year. PV output is generally low in the morning and evening, which results in lower absolute errors. Forecast errors could be more practically relevant for energy management around midday when production is higher. Hence, average error measures may not necessarily reflect the usefulness of the forecast in practice.

The elimination of nighttime observations is of course methodologically suitable because PV output is either zero or very small at night. To include nighttime would make the model look more accurate than it actually is. Even if the nighttime observations are

excluded, seasonal differences are still significant. The longer daytime during summer and the shorter it is during the winter result in lower production during the winter and higher useful data during the summer.

Future studies could then report error measures for each season or weather regime or production level. For instance, independent estimation with the models during the winter, summer, peak production hours, clear sky and cloudy sky would give more detailed information about the performance of the models.

6.1.4 Methodology limitations

The key methodological advantage of this study is the application of a structured forecasting process which combines lagged PV power, meteorological predictors, solar-geometry features, clear-sky normalization and snow-related variables. This makes the model more applicable to the Helsinki environment than a univariate time-series model.

But it does have difficulties when using a single model for an entire year. PV operating conditions vary significantly in Helsinki and between winter and summer. During the winter months, the sun rises and sets at low angles, and the days are short, and sometimes production is limited due to snow cover. During summer PV production is significantly greater, and duration of daylight is long. One model should be effective in all these seasons, which could mean poor performance at certain times of the year.

A possible improvement: divisionalization of the model for different seasons. A winter model could emphasize low elevation of the sun, snow effects, and short daylight hours while a summer model could emphasize high production levels and cloud driven variations. This could lead to better performance as each model would be required to model a smaller operating range.

It would be better to use models specific to each weather regime. Four sets of models could be trained for 4 different weather conditions: clear-sky, partly cloudy, overcast,

and snowy. This may enable the forecasting system to be more flexible with respect to various modes of PV variability.

This study is also constrained by deterministic forecasting. The GBRT model does not explicitly estimate forecast uncertainty, but it does provide point forecasts. In the real application of energy management, uncertainty information can be beneficial as operators could have to know not only the predicted PV output, but also the range of future PV output. It would therefore be very helpful to have some sort of probabilistic forecast.

While modelling, future enhancements are incorporated which are highlighted as follows:

The results show that GBRT could be a proper method for ultra-short term PV forecasting in Helsinki. The model is flexible, is able to deal with nonlinear relationships and can integrate various kinds of predictors. Still, the method could be enhanced by incorporating machine learning with physical irradiance modelling.

Physical predictors might involve the extraterrestrial irradiance, air mass, solar zenith angle, and atmospheric attenuation, and could be a part of a hybrid model. These variables might be able to give a physically meaningful estimation of the irradiance that can be expected, and the GBRT model could be trained to learn deviations due to cloud, local weather conditions, and snow cover.

It would be a great addition to get local sky-camera measurements incorporated. All-sky images can provide real-time data on the location, motion and blocking of cloud cover from the Sun's disk. This could be very beneficial for shorter forecast lead times, for example when forecasting in the next 10 to 60 minutes, where the impact of local cloud motion on PV output is significant.

Additional research may also examine other aspects of weather data, such as increased spatial resolution, seasonal forecasting, weather-regime classification, and probabilistic forecasting. The extensions may increase the precision and usefulness of the PV forecasting systems in Nordic high latitude regions.

6.1.5 Physical modelling and future improvements

The results indicate that GBRT can be used as a proper forecasting tool for the ultra-short-term PV forecasting in Helsinki. The model is flexible, it is possible to deal with the nonlinear relationship, and it is possible to include various types of predictors. The method may however be further enhanced by incorporating machine learning with physical irradiance modelling.

Physical predictors could be considered such as extraterrestrial irradiance, air mass, solar zenith angle, and atmospheric attenuation, in a hybrid model. These variables might give a physically sound estimate of expected irradiance, and the GBRT model might learn deviation due to the clouds, local weather condition and snow cover.

The other significant improvement would be the incorporation of the local sky-camera measurements. All-sky images can be used to get current data on cloud location, cloud motion, and solar-disk obstruction. This may be particularly helpful for forecasting on time scales of 10-60 minutes, where local cloud dynamics have a significant impact on PV generation.

A range of future issues could also be explored, such as the use of higher resolution spatial weather information, season forecasting models, weather-regime classification and probabilistic forecasting. The extensions may enhance the accuracy and usefulness of PV forecasting systems in Nordic high-latitude regions.

6.1.6 Overall interpretation

Overall, the results indicate that the GBRT model is applicable to predicting PV power at the Helsinki Kumpula site at an ultra-short time scale. The model is most effective when the forecast time is the smallest, because the most recent power observations from the PV can provide valuable information on the near-future power production. Forecasting at longer ranging times is more difficult due to more uncertain cloud motion, variation of the irradiance, and local atmospheric changes. But the meteorological variables and the features of solar geometry and engineered predictors still convey meaningful information to the model.

The results also demonstrate that historical PV production data is not enough for accurate PV forecasting at high latitudes. Production history and variables related to irradiance, meteorological predictors, solar geometry, and winter-related variables are all key in enhancing forecasting accuracy. This is particularly true during the seasonal variations in solar elevation, length of daylight hours, cloudiness, snow cover, in Helsinki.

From the power system point of view, the accurate forecasting of PV for the UST is crucial for the following purposes: power system operation, local energy management, power system battery control, and power system balancing. For instance, a PV forecast of 10–60 minutes duration can be used to determine whether the battery should be charged or discharged or if there will be a need for reserve power or if PV production is expected to improve or worsen with changing cloud conditions. This will enhance the capacity of the power system to respond to significant PV power fluctuations.

The power grid may be affected by prediction errors practically. PV production being overestimated could cause the system to expect more solar power than is available, resulting in power imbalance, reserve activation, or incorrect battery-discharge decision. Under-estimating PV production can result in underutilization of available solar energy, inappropriate battery charging schedule or unnecessarily operating local flexibility resources. Large errors in the forecast of PV generation on short time scales can also create

voltage-control issues in distribution systems, particularly when PV generation fluctuates rapidly with change in cloud type and movement.

Hence, even small gains in short-term PV forecasting can be applied. A more precise 10–60-minute forecast would help in planning the operation, would lower risk of imbalance, would facilitate storage management and would enhance the reliability of PV integration in flexible power systems. The GBRT based model hence offers a statistical improvement not only but also serves as a practical tool to manage the fluctuations of solar power in Nordic grid conditions.

7 Conclusion and Future work

7.1 Conclusion

This thesis proposed an ultra-short term (UST) photovoltaic power forecast model for the site Kumpula in Helsinki using Gradient Boosting Regression Trees (GBRT). The study was dedicated to 10 to 60 minutes forecast of PV power output for the Nordic high-latitude environment. The forecasting framework applied lagged PV power observations, meteorological and irradiance data, temporal features, solar-geometry data, clear-sky normalization, and snow-related predictors.

The results indicate that the GBRT model has desirable forecasting performance at all considered ultra-short time forecast horizons. The best performance was obtained at the shortest forecast horizon used, at which point the PV power measurements have information about the near future PV operating state. The error of the forecast was also growing with the length of the forecast period. This is to be expected from the further into the future, the more difficult it is to foresee the movement of clouds, the changes of the irradiance, and the uncertainty of local weather conditions.

The results also reveal that the prediction of PV generation in the short term (in the order of minutes) in Helsinki cannot solely be based on past knowledge of PV power generation. PV production is particularly relevant at short time scales, and meteorological variables, PV irradiance data, PV geometry and PV snow prediction are also important to capture the environment conditions influencing PV production. This is especially true in high latitudes where PV production is highly dependent on seasonality, solar elevation (low in winter, high in summer), and snow effects.

The most important part of this thesis is the development of a structured and repeatable GBRT-based forecasting workflow for the site of Helsinki Kumpula. The proposed solution solves many problems in Nordic PV forecast, such as seasonal variation, quick changes in irradiance due to clouds, or production during the winter due to snow and

low sun elevation. In general, the results suggest that GBRT is appropriate for the ultra-short-term forecasting of PV in Helsinki.

7.2 Future work

There are several directions for ongoing research. The first is the possibility of including sky-camera observations in future studies, which would be of local origin. Sky-camera images can be used to give information on the position of the clouds, their movement and the presence of clouds on the sun's disk. These would be particularly helpful for short forecast periods when local cloud change has a significant influence on PV production.

Third, future studies might consider the use of hybrid forecasting models that integrate machine learning with physical irradiance modelling. Other physical parameters like extraterrestrial irradiance, air mass, solar zenith angle and atmospheric attenuation may be used to provide further information on expected clear-sky conditions. The physically based predictors could be used to train GBRT or other machine learning techniques to enhance the accuracy of the prediction.

Thirdly, there could be more detailed spatial weather data included. The model would benefit from spatial information for meteorological or cloud-related data around the PV site, to reflect the approach of weather systems to this site. This may help forecasts, particularly for the 10-to-60-minute time periods.

Fourth, future research may focus on models for specific seasons or weather regimes. A separate model for winter and summertime, or even for clear-sky, cloudy and snowy conditions might provide better results, when the operating range of a single model is very broad.

Finally, probabilistic forecasting can be considered. This study is based on deterministic point forecasts, but there is a need for uncertainty estimates in operational energy

management. Prediction intervals or probabilistic forecasts would offer not only information about the expected PV output, but also about the range of possible outcomes.

References

Ahmad, T., Chen, H., & Guo, Y. (2018). Discuss different modelling techniques used to forecast PV electricity generation. Renewable and Sustainable Energy Reviews, 82, 254–264. https://doi.org/10.1016/j.rser.2017.09.009

Al-Dahidi, S., Madhiarasan, M., Al-Ghussain, L., Abubaker, A. M., Ahmad, A. D., Alrbai, M., Aghaei, M., Alahmer, H., Alahmer, A., Baraldi, P., & Zio, E. (2024). A holistic review and novel solar PV power production forecasting framework based on data inputs. Energies, 17(16), 4145. https://doi.org/10.3390/en17164145

Andrade, J. R., & Bessa, R. J. (2017). Enhancing the forecasting of renewable energy by using a grid of numerical weather predictions. IEEE Transactions on Sustainable Energy, 8(4), 1571–1580. https://doi.org/10.1109/TSTE.2017.2694340

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martínez-de-Pisón, F. J., Antonanzas-Torres, F. (2016). A survey of PV power forecasting. Solar Energy, 136, 78–111. https://doi.org/10.1016/j.solener.2016.06.069

Bacher, P., Madsen, H., & Nielsen, H. A. (2009). Short-time solar power forecasting using the Internet. Solar Energy, 83(10), 1772–1783. https://doi.org/10.1016/j.solener.2009.05.016

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). Cross validation and its applications for time series prediction with AR models: a note. Computational Statistics & Data Analysis, 120, 70–83. https://doi.org/10.1016/j.csda.2017.11.003

Dobos, A. P. (2014). PVWatts version 5 manual. The National Renewable Energy Laboratory. https://doi.org/10.2172/1158421

Duffie, J. A., & Beckman, W. A. (2013). Fourth Edition of Solar engineering of thermal processes. John Wiley & Sons. https://doi.org/10.1002/9781118671603

Finnish Meteorological Institute. (n.d.). Weather observations. The information was accessed on 25 January 2026 from: https://en.ilmatieteenlaitos.fi/weather-observations

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gilman, P., Dobos, A., DiOrio, N., Freeman, J., Janzou, S., & Ryberg, D. (2018). SAM PV model technical reference update. The National Renewable Energy Laboratory. NREL/TP-6A20-67399.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition). Springer. https://doi.org/10.1007/978-0-387-84858-7

Inman, R. H., Pedro, H. T. C., & Coimbra, C. F. M. (2013). Techniques for forecasting solar energy for integration with renewable energy. Progress in Energy and Combustion Science, 39(6), 535–576. https://doi.org/10.1016/j.pecs.2013.06.002

Kasten, F., & Young, A. T. (1989). Evolution of optical air mass tables and approximation formula. Applied Optics, 28(22), 4735–4738. https://doi.org/10.1364/AO.28.004735

Manni, M., Nocente, A., Kong, G., Skeie, K., Fan, H., & Lobaccaro, G. (2022). A model chain consisting of solar irradiation models, LiDAR scanner and high detailed 3D building model for solar energy digitalization at high latitudes. Frontiers in Energy Research, 10, 1082092. https://doi.org/10.3389/fenrg.2022.1082092

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review, 116(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Pedro, H. T. C., & Coimbra, C. F. M. (2012). Evaluation of the forecast models of the solar energy generation without using exogenous variables. Solar Energy, 86(7), 2017–2028. https://doi.org/10.1016/j.solener.2012.04.004

PV Performance Modeling Collaborative. (n.d.). Air mass. Sandia National Laboratories. Retrieved on April 24, 2026 from https://pvpmc.sandia.gov/modeling-guide/1-weather-design-inputs/irradiance-insolation/air-mass/.

PV Performance Modeling Collaborative. (n.d.). Extraterrestrial radiation. Sandia National Laboratories. Modeling Guide, 1 – Design Inputs – Weather, Retrieved April 24, 2026, from https://pvpmc.sandia.gov/modeling-guide/1-weather-design-inputs/irradiance-insolation/extraterrestrial-radiation/.

Reikard, G. (2009). Comparing time series predictions of solar radiation at high resolutions. Solar Energy, 83(3), 342–349. https://doi.org/10.1016/j.solener.2008.08.007

Tzoumanikas, P., Nikitidou, E., Bais, A. F., & Kazantzidis, A. (2016). Surface solar irradiance (SSI) effect caused by clouds obtained from all-sky images. Renewable Energy, 95, 314–322. https://doi.org/10.1016/j.renene.2016.04.026

Zheng, X., Bagloee, S. A., & Sarvi, M. (2024). Memory-fused XGBoost for time-series forecasting: RecVAE-GBRT. P. Francois, "In 2024 International Joint Conference on Neural Networks (IJCNN)", in proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN) (2024) (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN60899.2024.10650508

Appendices

Appendix 1. Additional forecasting results

This appendix contains additional forecasting information used to support the analysis in Chapter 5. The purpose of the appendix is to give more details, but not too much in the results chapter.

Table 5.1 shows the forecast performance of the Gradient Boosting Regression Trees model for the selected ultra-short-term forecast horizons.

Forecast horizon (min)	RMSE (W)	nRMSE (%)
10 min	1940.44	10.03
60 min	2412.04	12.46

From the results, it is observed that the GBRT model outperforms in the shortest forecast period evaluated. The longer the forecast time horizon the greater the forecast error will be, increasing from 1940.44 W at 10 minutes ahead to 2412.04 W at 60 minutes ahead. This means that the longer-range forecasts are more challenging as the movement of clouds, variability in irradiance, and local weather conditions further away from the forecast are less predictable.

Appendix 2. Python code for data preparation and model implementation

Main implementation steps of the GBRT forecasting workflow are given in this appendix. The code shows data preprocessing, feature engineering, model training, hyperparameter tuning and evaluation.

```
import pandas as pd
import numpy as np

from sklearn.ensemble import GradientBoostingRegressor
from sklearn.model_selection import TimeSeriesSplit
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score


# Load meteorological data from the Helsinki Kumpula station.
weather = pd.read_csv(
    "Helsinki Kumpula_ 25.8.2015 - 22.8.2017_10Minute.csv"
)

# Create timestamp from separate date and time columns.
weather["timestamp"] = pd.to_datetime(
    weather["Year"].astype(str) + "-" +
    weather["Month"].astype(str).str.zfill(2) + "-" +
    weather["Day"].astype(str).str.zfill(2) + " " +
    weather["Time [Local time]"]
)

# Convert cloud-cover text into a numerical value.
# Example: "Mostly cloudy (7/8)" becomes 7.
weather["cloud_cover"] = (
    weather["Cloud cover [1/8]"]
    .astype(str)
    .str.extract(r"\((\d)/8\)")
    .astype(float)
)

# Treat values where cloudiness cannot be determined as missing.
weather.loc[weather["cloud_cover"] == 9, "cloud_cover"] = np.nan

# Rename meteorological variables for easier modelling.
weather = weather.rename(columns={
    "Air temperature mean [°C]": "air_temperature",
    "Snow depth mean [cm]": "snow_depth",
    "Wind speed mean [m/s]": "wind_speed",
```

```
    "Precipitation amount mean [mm]": "precipitation",
    "Relative humidity mean [%]": "relative_humidity"
})

# Convert numerical columns.
numeric_columns = [
    "air_temperature",
    "snow_depth",
    "wind_speed",
    "precipitation",
    "relative_humidity"
]

for column in numeric_columns:
    weather[column] = pd.to_numeric(weather[column], er-
rors="coerce")

# In the FMI data, negative snow-depth values are treated as
no measurable snow.
weather["snow_depth"] = weather["snow_depth"].clip(lower=0)

# Create a snow indicator.
weather["snow_indicator"] = (weather["snow_depth"] >
0).astype(int)

# Keep only variables used in the forecasting model.
weather = weather[
    [
        "timestamp",
        "air_temperature",
        "cloud_cover",
        "snow_depth",
        "snow_indicator",
        "wind_speed",
        "precipitation",
        "relative_humidity"
    ]
]

# Model training and evaluation

forecast_horizons = {
    10: 1,
    30: 3,
    60: 6
}

results = []

for horizon_minutes, horizon_steps in forecast_hori-
zons.items():

    data_h = data.copy()
```

```
# Create target variable.
    data_h["target"]    =    data_h["pv_power"].shift(-hori-
zon_steps)
# Load PV power data from the Helsinki Kumpula PV installa-
tion.
# The PV file should contain at least timestamp and pv_power
columns.
pv      =      pd.read_csv("helsinki_kumpula_pv_power.csv",
parse_dates=["timestamp"])

# Sort both datasets chronologically.
weather = weather.sort_values("timestamp")
pv = pv.sort_values("timestamp")

# Merge weather data with PV power observations.
# The nearest previous weather observation is matched to each
PV timestamp.
data = pd.merge_asof(
    pv,
    weather,
    on="timestamp",
    direction="backward"
)

# Create time-related variables.
data["hour"] = data["timestamp"].dt.hour
data["day_of_year"] = data["timestamp"].dt.dayofyear
data["month"] = data["timestamp"].dt.month

# Create lagged PV power variables.
data["pv_lag_1"] = data["pv_power"].shift(1)
data["pv_lag_2"] = data["pv_power"].shift(2)
data["pv_lag_1_day"] = data["pv_power"].shift(144)

# The value 144 assumes 10-minute PV data:
# 6 observations per hour × 24 hours = 144 observations per
day.

data["pv_delta"] = data["pv_lag_1"] - data["pv_lag_2"]

# Create simple weather-variability features.
data["cloud_cover_lag_1"] = data["cloud_cover"].shift(1)
data["cloud_cover_lead_1"] = data["cloud_cover"].shift(-1)
data["cloud_cover_variability"] = (
    data["cloud_cover_lead_1"] - data["cloud_cover_lag_1"]
)

data["temperature_lag_1"] = data["air_temperature"].shift(1)
data["temperature_lead_1"] = data["air_temperature"].shift(-
1)
data["temperature_variability"] = (
    data["temperature_lead_1"] - data["temperature_lag_1"]
```

```
)

# Define the forecast horizon.
# For 10-minute PV data, a 60-minute-ahead forecast equals 6
steps.
forecast_horizon_steps = 6

data["target"]   =   data["pv_power"].shift(-forecast_hori-
zon_steps)

# Define predictor variables.
feature_columns = [
    "pv_lag_1",
    "pv_lag_2",
    "pv_lag_1_day",
    "pv_delta",
    "air_temperature",
    "cloud_cover",
    "snow_depth",
    "snow_indicator",
    "wind_speed",
    "precipitation",
    "relative_humidity",
    "hour",
    "day_of_year",
    "month",
    "cloud_cover_variability",
    "temperature_variability"
]

# Remove missing values caused by lagging, merging, and tar-
get shifting.
model_data = data.dropna(subset=feature_columns + ["target"])

X = model_data[feature_columns]
y = model_data["target"]

# Time-series-aware cross-validation.
tscv = TimeSeriesSplit(n_splits=5)

best_rmse = float("inf")
best_n_estimators = None
best_model = None

for n_estimators in [50, 100, 150, 200]:
    rmses = []

    for train_idx, val_idx in tscv.split(X):
        X_train, X_val = X.iloc[train_idx], X.iloc[val_idx]
        y_train, y_val = y.iloc[train_idx], y.iloc[val_idx]

        model = GradientBoostingRegressor(
            n_estimators=n_estimators,
```

```
            learning_rate=0.03,
            max_depth=3,
            random_state=42
        )

        model.fit(X_train, y_train)
        y_pred = model.predict(X_val)

        rmse = np.sqrt(mean_squared_error(y_val, y_pred))
        rmses.append(rmse)

    avg_rmse = np.mean(rmses)

    if avg_rmse < best_rmse:
        best_rmse = avg_rmse
        best_n_estimators = n_estimators

# Train final model using the selected number of trees.
best_model = GradientBoostingRegressor(
    n_estimators=best_n_estimators,
    learning_rate=0.03,
    max_depth=3,
    random_state=42
)

best_model.fit(X, y)
y_pred = best_model.predict(X)

# Persistence baseline.
persistence_prediction = model_data["pv_power"]

# Evaluation metrics.
rmse = np.sqrt(mean_squared_error(y, y_pred))
mae = mean_absolute_error(y, y_pred)
r2 = r2_score(y, y_pred)

rmse_persistence = np.sqrt(mean_squared_error(y, persis-
tence_prediction))
skill_score = 1 - (rmse / rmse_persistence)

print("Best number of estimators:", best_n_estimators)
print("RMSE:", rmse)
print("MAE:", mae)
print("R²:", r2)
print("Skill score:", skill_score)
results = pd.DataFrame(results)

print(results)
```