Solar panels as radiation sensors

Using photovoltaic power output data to model global solar radiation

Master thesis

Msc Environmental Engineering

Msc student: Daniël Bouman

Supervisors: Dr.ir. A.M. Droste & Dr.ir. R. Taormina

October 2025

Delft University of Technology

Acknowledgements

I would like to express my sincere gratitude to Dr.ir. A.M. Droste and Dr.ir. R. Taormina, for their advice and supervision throughout the course of this project. Furthermore I would like to thank Marijn Leeuwenberg and Gerben Voogt from The Green Village for their assistance in acquiring the necessary data from the TGV data platform. This project would not have been possible without their help.

With sincere appreciation,
Daniël Bouman
Delft, October 2025

Abstract

Accurate measurements of global solar radiation are essential for applications ranging from photovoltaic (PV) yield assessment to grid operation and agricultural management. However, direct measurements with pyranometers are sparse and costly, motivating alternative approaches based on widely available PV power production data. This thesis investigates how accurately global radiation can be estimated using machine learning models trained on PV power output data, supplemented with weather reanalysis data.

Two tree-based models, Random Forest and Gradient Boosting, were applied to PV data from two sites in the Netherlands combined with globally available weather reanalysis data, using on-site pyranometer measurements as the ground truth for training and validation. Both *single-location models*, trained and validated on the same site, and *cross-location models*, trained on one site and applied to another, were evaluated. For comparison, a linear regression baseline model was also tested.

The single-location models achieved strong performance, with $R^2 \approx 0.97$, mean absolute errors (MAE) of 10–15 [W/m^2], and near-zero bias at a 15-minute resolution, substantially outperforming established reanalysis products such as ERA5. Cross-location models retained reasonable accuracy ($R^2 \approx 0.94$, MAE 15–30 [W/m^2]), though with increased bias. Feature importance analysis highlighted the dominant influence of PV power output and the clear-sky index for photovoltaics (K_{PV}), while reanalysis variables contributed little. The top-of-atmosphere radiation proved to be a consistently useful predictor.

These results demonstrate that PV power data can be transformed into accurate, high-frequency radiation estimates, offering a cost-effective complement to pyranometer measurements. While site-specific calibration currently limits full generalisation, the findings point toward the feasibility of scalable models based on transferable features such as K_{PV}. Expanding datasets across more climates and system configurations, and exploring advanced model architectures, are promising next steps towards generalised PV-based radiation models that could be applied without requiring on-site pyranometer measurements for training.

Contents

1 Introduction

Global solar radiation, also referred to as global horizontal irradiance (GHI), is the total shortwave radiation received from the sun on a horizontal surface at ground level (Sandia National Laboratories, 2025a). Many fields of research and application require access to high-temporal-resolution global solar radiation data. These include solar energy forecasting for grid integration, climate modelling of radiative fluxes, crop yield prediction in precision agriculture, and dynamic control in building energy simulations. In each case, accurate and frequent radiation data enable better model calibration, more informed operational decision-making, and more detailed analyses of short-term environmental dynamics such as cloud variability or shading effects.

Despite its importance, the availability of high-resolution global radiation data is extremely limited. Ground-based pyranometers, the primary instruments used for such measurements, are sparsely distributed and often only available at dedicated weather stations. This creates large spatial gaps in the data, especially in regions where environmental or energy monitoring infrastructure is lacking. Even in a well-instrumented country like the Netherlands, there is only one station measuring global radiation for approximately every $1000\,\mathrm{km}^2$ of land area (KNMI, 2024).

Improving access to high-resolution radiation data can significantly enhance the accuracy of solar generation forecasting, enable real-time control of energy systems, and support remote agricultural or environmental monitoring. For example, better irradiance estimates can reduce grid imbalances in areas with high photovoltaic (PV) penetration by improving reserve planning and dispatch accuracy (Kaur, 2015; Goodarzi et al., 2019). In agriculture, solar-powered smart irrigation systems can leverage accurate solar irradiance data to optimise both water use and energy management, improving system efficiency under variable environmental conditions (Mohammed et al., 2023). Similarly, real-time irradiance data supports dynamic control strategies in building energy systems, contributing to energy savings and indoor comfort (Kim et al., 2014).

To address this data gap, this thesis investigates a data-driven method for estimating global solar radiation from PV power output data. While most research has focused on predicting PV output from solar radiation, the reverse—estimating radiation from PV data—has received comparatively little attention. Yet this inverse relationship holds substantial practical value. PV systems are now widely deployed and often collect high-frequency power output data, offering an abundant, underutilised data source for estimating solar radiation.

Recent studies have begun to explore this idea. A proof-of-concept study by Beran (2013) demonstrated the feasibility of estimating solar radiation from monthly PV power data from a residential system, achieving an R^2 of 0.99 against nearby weather station data, though its coarse temporal resolution limits practical use. Later, Engerer et al. (2015) used data from commercial PV systems but focused solely on modelling the diffuse fraction of solar radiation. Nespoli and Medici (2017) later proposed an unsupervised method based on physical PV system models to derive global horizontal irradiance from aggregated PV power signals, showing that their approach outperforms satellite-based products at high temporal resolution. Finally, Carvalho and Corrêa (2019) proposed a model linking PV voltage to radiation using laboratory measurements, although real-world complexities were excluded.

In contrast to the unsupervised, physics-based approach of Nespoli and Medici, this thesis employs supervised machine learning models trained on paired pyranometer and PV power data, supplemented with data from the Open-Meteo Historical Weather API, an open-source reanalysis product that combines modern weather models with observational data to provide global historical weather records.

By transforming widely available PV output data into usable radiation estimates, this approach reduces the need for dense sensor networks, offering a scalable and cost-effective means of improving data coverage and supporting energy and environmental applications in under-instrumented regions. A further contribution is the focus on a 15-minute temporal resolution, whereas most

existing solar radiation models operate at an hourly timescale. Previous studies have shown that modelling solar radiation becomes increasingly challenging at finer timescales, as errors tend to grow due to higher variability in atmospheric conditions (Omoyele et al., 2024; Rajagukguk and Lee, 2023; Li et al., 2023).

The main research question of this thesis is: "How accurately can global solar radiation be estimated using a machine learning model based on photovoltaic power output data?"

To answer this question, this study develops machine learning models that use PV power output and auxiliary weather reanalysis data to estimate global solar radiation. Multiple modelling scenarios are evaluated, including both single-site training and cross-site generalisation. Model performance is assessed using standard error metrics, and feature-importance analysis is conducted to understand model behaviour and input relevance.

To guide the investigation, this study formulates one central research question along with a set of sub-questions. These sub-questions explore the relative performance of different modelling approaches, the extent to which spatial generalisation can be achieved, and the influence and practicality of various input features.

Main research question

> *How accurately can global solar radiation be estimated using a machine learning model based on photovoltaic power output data?*

Sub research questions

1. How accurately can a Random Forest regression model estimate global solar radiation?
2. How accurately can a Gradient Boosting regression model estimate global solar radiation?
3. How accurately can a machine learning regression model trained with normalised photovoltaic power output at one site predict global radiation at a different site?
4. What features are useful for model performance and can be feasibly acquired?

2 Data

This section provides an overview of the data sources and input variables used in the models developed in this study. PV power data and global solar radiation data—measured with a pyranometer—were collected from two locations in the Netherlands: Delft and Wageningen (Figure 1). In addition to these local measurements, reanalysis weather data were retrieved from external sources to supplement local measurements and improve model performance.

The complete input dataset includes a combination of directly measured values, weather reanalysis data, variables derived from solar geometry and time, and features computed from existing inputs, as described in Section 3. A summary of all variables is presented in Table 1. After processing, all data were used at a 15-minute time resolution, and time-normalised to UTC.

A full description of the data preprocessing procedures is provided in Appendix I, along with plots of the PV and global radiation data sets.

Figure 1: The 2 locations in the Netherlands where data were collected.

2.1 Delft

In Delft, both PV power and global solar radiation data were collected at the urban research facility of Delft University of Technology, known as The Green Village. These data can be accessed through their data platform (The Green Village, 2025) [1]. The PV power output was measured at the transformer level, capturing the total alternating current (AC) power generated by a system of 18 south-facing solar panels (model: Bisol BMU-260). The panels have a tilt of 30° and an azimuth of 157°. Global radiation was recorded using a pyranometer installed on-site. All measurements were collected at a temporal resolution of 15 minutes. Combined PV and pyranometer data were available from November 2021 to April 2024, with notable gaps in the PV data occurring around May and December 2022, as well as in March and April 2023. Only time steps for which both PV and pyranometer readings were available were retained, leaving a total of 63,016 timestamps with valid measurements.

For the Delft data, two important preprocessing steps were taken. The PV data from The Green Village returned no data during night time. These data were manually set to 0 [W/m^2] in order to have continuous data. For the combined PV and pyranometer data, three continuous days of data were also removed as the pyranometer dataset had values of 0 [W/m^2] for the entire three

[1] Data can only be observed through this platform. Acquiring the data requires a request to The Green Village.

days, while the PV data showed notable power production. It is assumed that during this time maintenance was being performed on the pyranometer.

Global solar radiation was measured in Delft using an EKO MS-40 pyranometer, a Class C instrument certified under ISO 9060:2018 (EKO Instruments, 2025). Class C pyranometers are intended for general meteorological use and offer moderate accuracy relative to higher-grade instruments. For this study, the measurement uncertainty is conservatively estimated at approximately ± 23 [W/m^2], based on the root-sum-square combination of known error sources such as zero offset, temperature response, and directional sensitivity. A detailed justification for this estimate is provided in Appendix II.

2.2 Wageningen

In Wageningen, PV power data were collected from a rooftop residential installation comprising six west-facing solar panels (model: LG MonoX 2 Black). The panels have a tilt of 45° and an azimuth of 245°. As in Delft, the power output was measured at the transformer, capturing the total AC power generated by the system. Global radiation data were sourced from a pyranometer located at the nearby meteorological observation site 'Veenkampen', situated approximately 3.5 kilometres west-northwest of the city. These data were acquired from the MAQ-observations website (Wageningen University & Research, 2025) and are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). All measurements were recorded at a 15-minute temporal resolution. Combined PV and pyranometer data were available from August 2017 to February 2025.

The PV dataset for Wageningen contains several gaps, most notably in September, October, and December 2017; July, August, September, and November 2018; October 2019; and from August through November 2023. As with the Delft data, only time steps for which both PV and pyranometer readings were available were retained, leaving a total of $208,000$ timestamps with valid measurements.

Global solar radiation was measured at Veenkampen using a Kipp & Zonen CM 11 pyranometer, which is classified as a *secondary standard* instrument under ISO 9060:1990 (Kipp & Zonen B.V., 2008). This designation represents the highest category of radiometric performance for pyranometers, ensuring high accuracy and stability. For this study, the measurement uncertainty is conservatively estimated at approximately ± 13 [W/m^2], based on the root-sum-square combination of key error sources. A full justification of this estimate is provided in Appendix II.

2.3 Reanalysis Weather Data

To supplement the locally collected data, this study incorporates reanalysis weather data from the Open-Meteo Historical Weather API (Open-Meteo, 2024). Open-Meteo is an open-source platform that provides access to reanalysis datasets, which combine numerical weather prediction models with observational data from weather stations and national meteorological services (e.g., the Dutch KNMI).

From Open-Meteo, hourly data were obtained for 14 meteorological variables and linearly interpolated to a 15-minute resolution to match the temporal scale of the PV and pyranometer measurements. Although Open-Meteo provides native 15-minute resolution data for regions such as North America and Europe, this study used the hourly global product to maintain consistency and generalisability across geographic contexts. This approach ensures that the methodology remains applicable to locations where only coarser reanalysis data are available.

2.4 Calculated Variables

Several variables used in this study are not directly observed but are derived from the available PV power and weather reanalysis data. These include the clear-sky PV power estimate (PV_{CLEAR}),

the relative production factor (R_{PV}), and the Clear Sky Index for Photovoltaics (K_{PV}). These variables are briefly described in Table 1 and are discussed in full detail in Sections 3.2 and 3.3.

To capture temporal dependencies in PV performance, the Clear Sky Index was also included in both lagged and lead form. These temporally shifted features are denoted as $K_{PV}^{(n)}$, where n indicates the number of 15-minute intervals by which the original K_{PV} value is shifted. For instance, $K_{PV}^{(-1)}$ corresponds to a 15-minute lag, while $K_{PV}^{(4)}$ represents a one-hour lead. Shifted values were computed for n ranging from -10 to 10, providing a temporal window of ± 2.5 hours. This range was selected based on preliminary experiments, which showed that shifts beyond this interval contributed little additional predictive value.

2.5 Variables from Solar Geometry and Time

Theoretical top-of-atmosphere (TOA) irradiance was calculated using solar geometry equations from Duffie and Beckman (2013), which require only geographic location, date, and time as inputs. This variable represents the solar radiation incident at the top of Earth's atmosphere under clear-sky conditions, and serves as a physically meaningful baseline for incoming solar energy.

Additionally, the binary variable *daylight* indicates whether the sun is above the horizon at a given time step (1 for daylight, 0 for nighttime). It was derived using solar position calculations based on time and location, implemented via the `pvlib` Python library (Anderson et al., 2023).

2.6 Data splitting

To enable model training and evaluation, the collected data were divided into three subsets: training, testing, and validation. The training set was used to fit the models, the testing set to evaluate performance during development and optimise hyperparameters, and the validation set to assess final model performance. A split ratio of 70–20–10 was applied.

Table 1: Overview of all model input variables, grouped by data source.

Variable	Unit	Description
Local Measurements (Delft/Wageningen)		
I_G	W/m^2	Measured global radiation from local pyranometer
PV_{MEAS}	W/m^2	Measured PV power output per unit module area
Open-Meteo Reanalysis API		
T_{2m}	°C	Air temperature at 2 metres height
RH_{2m}	%	Relative humidity at 2 metres height
N	%	Total cloud cover (fraction of sky area)
N_{Low}	%	Low-level cloud cover ($\leq$ 3 km altitude)
N_{Mid}	%	Mid-level cloud cover (3–8 km altitude)
N_{High}	%	High-level cloud cover ($\geq$ 8 km altitude)
P	mm	Total precipitation
P_{Rain}	mm	Rainfall component of precipitation
P_{Snow}	mm	Snowfall component of precipitation
WMO	–	Categorical weather condition code (WMO standard)
p_{msl}	hPa	Air pressure at mean sea level
p_s	hPa	Air pressure at surface level
WS	km/h	Wind speed at 100 m altitude
WD	°	Wind direction at 100 m altitude
Derived from Solar Geometry or Time		
I_{TOA}	W/m^2	Theoretical top-of-atmosphere irradiance
daylight	–	Binary indicator for sun above horizon (0 or 1)
d	–	Ordinal day of the year (1–365)
m	–	Minutes since midnight (0–1439)
Calculated from Data		
PV_{CLEAR}	W/m^2	Estimated PV power output per unit module area (clear-sky conditions)
R_{PV}	–	Relative production factor
K_{PV}	–	Clear Sky Index for photovoltaics
$K_{PV}^{(n)}$	–	K_{PV} shifted by n time steps of 15 minutes

3 Methodology

This section outlines the methodology used to develop, train, and evaluate machine learning models for estimating global solar radiation from photovoltaic (PV) power data. The focus is on assessing both the accuracy and generalisability of the models using data from two test sites in the Netherlands.

To this end, four modelling scenarios are considered—two single-location scenarios and two cross-location scenarios—designed to evaluate performance both within and across sites. A key aspect of the methodology is the use of the Clear Sky Index for Photovoltaics (K_{PV}), which helps normalise PV output data and enables better comparison across systems with different configurations.

Two approaches are implemented to simulate the clear-sky condition PV power needed to compute K_{PV}: a relative production method and the PVWatts model.

The chosen regression models—Random Forest and Gradient Boosting—are described in detail, along with the performance metrics used for evaluation. The models are trained under different generalisation scenarios, and hyperparameter optimisation is carried out separately for single- and cross-location scenarios to best suit their respective objectives. Finally, feature importance is analysed to gain insight into which input variables most strongly influence model predictions.

3.1 Modelling Scenarios and Spatial Generalisation

This study considers four modelling scenarios designed to evaluate both the performance and spatial generalisation capability of machine learning models for estimating global solar radiation from PV power output. Spatial generalisation refers to the ability of a model trained at one location to accurately estimate global radiation at a different site—using only PV power data from that site—without access to location-specific ground truth radiation measurements.

Two categories of modelling scenarios are defined:

- **Single-location models:** The model is both trained and evaluated on data from the same site. These scenarios assess how well a model can learn site-specific patterns and generalise over time. They represent a use case in which a pyranometer is installed temporarily alongside PV modules to collect ground truth radiation data for model calibration. After this initial period, the pyranometer may be removed, and the trained model used for continued radiation estimation based solely on PV output.
- **Cross-location models:** The model is trained on data from one site and evaluated on another. These scenarios are designed to test spatial generalisation. A successful model in this context would allow radiation estimation at new locations using only PV data, without requiring pyranometer measurements at those sites. This approach is particularly useful for scaling solar monitoring to regions where the installation of radiation sensors is impractical.

The four specific scenarios considered in this study are summarised in Table 2. Each scenarios is named according to the training and testing location.

Table 2: Overview of modelling scenarios evaluated in this study.

Scenario Name	Train → Test	Description
`Delft_Single`	Delft → Delft	Trained and tested on data from the Delft site (single-location model)
`Wageningen_Single`	Wageningen → Wageningen	Trained and tested on data from the Wageningen site (single-location model)
`Delft_Wageningen`	Delft → Wageningen	Trained on Delft data, tested on Wageningen data (cross-location model)
`Wageningen_Delft`	Wageningen → Delft	Trained on Wageningen data, tested on Delft data (cross-location model)

These modelling scenarios form the foundation of the methodology, enabling evaluation of model accuracy, robustness, and portability across different geographical locations. In particular, the cross-location results provide insight into the practical viability of using data-driven models to replace physical radiation sensors in new deployments.

3.2 Clear Sky Index for Photovoltaics

To ensure that a model generalises well across different geographical locations, the selected features should be physically meaningful and not overly dependent on local conditions (Razavi and Gupta, 2015). While the correlation between PV module power output and global radiation is certainly physically meaningful, it is also highly dependent on site-specific factors such as panel orientation and geographical location. For these reasons, rather than relying solely on the direct PV power output as a model feature, the Clear Sky Index for Photovoltaics (K_{PV}) is also employed, following the approach proposed by Engerer and Mills (2014).

The purpose of using K_{PV} is to normalise PV power output with respect to the maximum expected output under clear-sky conditions. This facilitates more meaningful comparisons between sites with different system sizes, orientations, and local atmospheric conditions. By expressing the measured output as a fraction of the clear-sky potential, K_{PV} provides a dimensionless indicator of relative system performance that is less influenced by local configuration and more reflective of transient atmospheric conditions such as cloud cover.

K_{PV} is calculated as:

$$K_{PV} = \frac{PV_{MEAS}}{PV_{CLEAR}} \tag{1}$$

where PV_{MEAS} is the measured PV system power output, and PV_{CLEAR} is the simulated power output of the system under clear-sky conditions. Both quantities are expressed in watts per square metre [W/m^2], making K_{PV} a dimensionless ratio. Throughout this study, PV power production is expressed in watts per square metre of module area to enable comparison between systems of varying sizes. Simulation of PV_{CLEAR} is explained in Section 3.3.

The manner in which K_{PV} is incorporated into the dataset depends on the modelling scenario:

- **Single-location scenarios:** K_{PV} is added as an additional feature alongside raw PV power output. This gives the model access to both absolute and normalised signals, potentially improving site-specific performance.
- **Cross-location scenarios:** The raw PV power output is replaced by K_{PV}, ensuring the input feature is normalised and thus more transferable across different sites. This substitution supports spatial generalisation by reducing dependence on site-specific irradiance and system configuration.

3.3 Clear-sky Power Output Simulation

In this study, two different methods are used to simulate the clear-sky power output (PV_{CLEAR}) required to calculate K_{PV}: a relative production approach and the PVWatts model for PV power estimation (Dobos, 2014).

Estimating clear-sky power output through simulated relative production

The clear-sky power output is first estimated using a relative production approach, as outlined below. The first step involves computing clear-sky irradiance values based on geographical location and time, using the simplified Solis model proposed by Ineichen (2008), which is an analytical approximation of the original Solis model developed by Müller et al. (2004). This model provides clear-sky condition estimates of the three key components of solar irradiance: Direct Normal Irradiance (DNI), Global Horizontal Irradiance (GHI), and Diffuse Horizontal Irradiance (DHI), as illustrated in Figure 2.

- **Direct Normal Irradiance** refers to the solar radiation received per unit area by a surface that is always oriented perpendicular to the incoming solar rays.
- **Global Horizontal Irradiance** is the total solar radiation received per unit area by a horizontal surface. It includes both the direct component (projected onto the horizontal plane) and the diffuse component scattered by the atmosphere.
- **Diffuse Horizontal Irradiance** is the portion of solar radiation received from the sky after scattering by air molecules, aerosols, and clouds, excluding direct sunlight.

These components are related through the following equation:

$$GHI = DNI \cdot \cos(\theta) + DHI \tag{2}$$

where θ is the solar zenith angle, defined as the angle between the sun's rays and the vertical (see Figure 2).

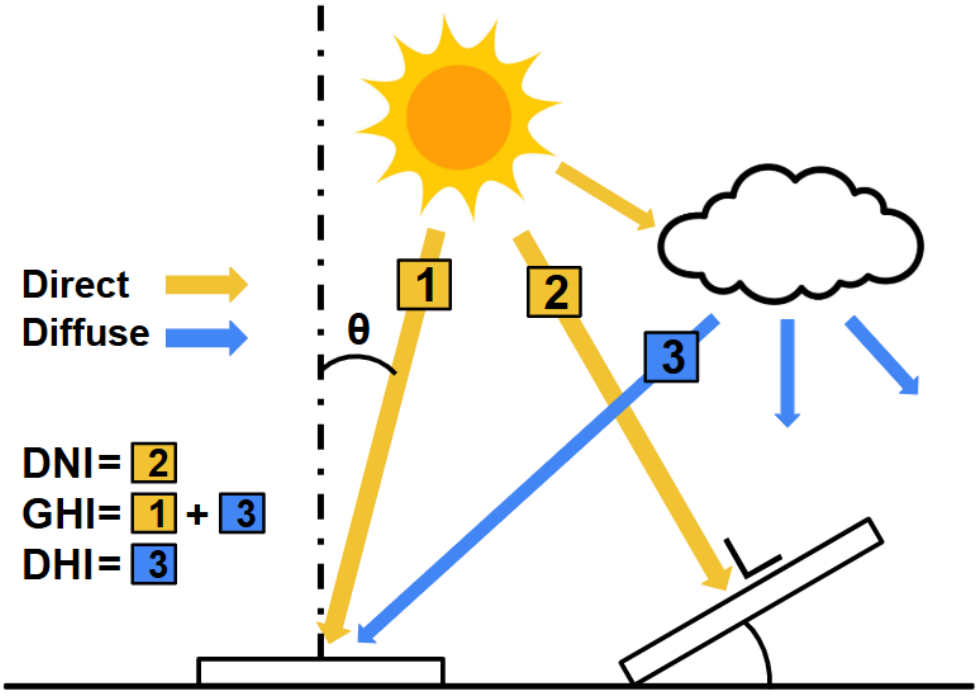


Figure 2: Visual representation DNI, GHI and DHI.

Next, the plane-of-array (POA) irradiance—representing the total solar radiation incident on the tilted surface of the PV module at the test site—is simulated using the isotropic sky diffuse model (Liu and Jordan, 1963). This model assumes the diffuse radiation is uniformly distributed across the sky dome (Sandia National Laboratories, 2025b) and requires the previously calculated DNI, GHI, and DHI values, along with the module's tilt, azimuth, and geographical location.

The POA irradiance is then normalised by the standard test-condition irradiance—typically $1000\,[\mathrm{W/m^2}]$ (Sinovoltaics, 2011)—to obtain the relative production factor R_{PV} (Eq. 3). Then, the clear-sky power output in direct current is estimated by multiplying R_{PV} by the PV module's rated DC power output under standard test conditions (Eq. 4). Finally, the DC power output estimate is multiplied by an inverter loss factor (η_{inv}) to obtain the estimated clear-sky power output for the relative production method, denoted as $PV_{CLEAR,rel}$ (Eq. 5).

The relevant equations are as follows:

$$R_{PV} = \frac{I_{POA}}{I_{STC}} \tag{3}$$

$$PV_{DC,rel} = R_{PV} \cdot PV_{dc0} \tag{4}$$

$$PV_{CLEAR,rel} = \eta_{inv} \cdot PV_{DC,rel} \tag{5}$$

where:

- R_{PV} is the dimensionless relative production factor $[-]$,
- I_{POA} is the simulated Plane of Array irradiance under clear-sky conditions $[\mathrm{W/m^2}]$,
- I_{STC} is the irradiance level for standard test conditions, for both considered PV setups $1000\,[\mathrm{W/m^2}]$,
- $PV_{DC,rel}$ is the estimated clear-sky power output before inverter losses $[\mathrm{W/m^2}]$,
- PV_{dc0} is the rated DC power output of the system under standard test conditions $[\mathrm{W/m^2}]$,
- $PV_{CLEAR,rel}$ is the estimated clear-sky power output for the relative production approach $[\mathrm{W/m^2}]$,
- η_{inv} is the inverter efficiency $[-]$.

Equation 4 assumes a linear relationship between POA irradiance and PV module power output, which is reasonably accurate for most PV module types (Meflah et al., 2024). While reasonable under clear-sky and high-irradiance conditions, this simplification introduces limitations. It neglects temperature effects, which can significantly reduce power output as module temperature increases, and does not account for angle-of-incidence losses, where shallow sunlight angles lead to higher reflection. Additionally, under low-irradiance conditions, non-linear effects such as inverter inefficiencies and shunt resistance become more prominent, potentially reducing accuracy.

To address these limitations, the second approach in this study uses the PVWatts model, which incorporates temperature effects and system-level derate factors.

Estimating clear-sky power output using the PVWatts model

In the second approach, clear-sky power production is estimated using PVWatts, incorporated within a model chain function from the `pvlib` library for Python (Anderson et al., 2023). The resulting estimate, $PV_{CLEAR,pvw}$, serves as an alternative to the relative production method for computing the clear-sky power baseline used in the derivation of the clear-sky index K_{PV}.

PVWatts is a simulation tool developed by the National Renewable Energy Laboratory (NREL) for estimating the energy production of grid-connected photovoltaic systems (Dobos, 2014).

The model chain is set up as follows: First, as with the relative production method, clear-sky irradiance values are computed using the simplified Solis model, yielding DNI, GHI, and DHI. Following this, these irradiance components are then transposed into POA irradiance using the isotropic sky diffuse model, the same as for the relative production approach. Simultaneously, the SAPM model is used to simulate the module temperature, using weather data on air temperature and wind speed as inputs.

Within the model chain spectral losses are neglected, which would otherwise require additional data on atmospheric conditions such as air mass and composition. This is a common simplifying assumption for standard silicon modules under regular conditions, as spectral effects are very small for standard silicon modules (Ishii et al., 2011).

Finally, the PVWatts model is used to estimate the clear-sky power output using the following two equations (Dobos, 2014):

$$PV_{DC,pv} = \frac{I_{POA}}{I_{STC}} \cdot PV_{dc0} \left[1 + \gamma_{pdc} \cdot (T_{cell} - T_{ref})\right] \tag{6}$$

$$PV_{CLEAR,pv} = \eta_{inv} \cdot PV_{DC} \tag{7}$$

where:

- $PV_{DC,pv}$ is the simulated DC power output [W/m^2],
- I_{POA} is the plane-of-array irradiance in [W/m^2],
- I_{STC} is the irradiance level for standard test conditions, typically 1000 [W/m^2],
- PV_{dc0} is the rated DC power output of the system under standard test conditions [W/m^2],
- γ_{pdc} is the power temperature coefficient [°C^{-1}],
- T_{cell} is the PV cell temperature in [°C],
- T_{ref} is the reference cell temperature, set at 25 [°C],
- $PV_{CLEAR,pv}$ is the estimated clear-sky power output for the PVWatts model approach [W/m^2],
- η_{inv} is the inverter efficiency [−],

3.4 Linear Regression

As a baseline model, a linear least squares regression is fitted for the single-location scenarios. Prior to fitting, all features are scaled to have zero mean and unit variance. The model estimates a set of linear coefficients w_1 through w_n, where n is the number of features, along with an intercept term b. These parameters are chosen to minimise the residual sum of squares between the observed target values and the model predictions.

To maintain physical plausibility, any negative predictions produced by the linear regression model are set to zero, as negative global radiation values are physically impossible.

During preliminary testing, certain input features were found to be strongly linearly correlated with others, causing instability in the estimated regression coefficients. To address this, the features PV_{CLEAR}, p_{msl}, and P_{Rain} were excluded from the linear regression inputs. This adjustment was applied exclusively to linear regression models, which are particularly sensitive to multicollinearity (Shrestha, 2020). Furthermore, the models were trained only on datasets with

relative-production-derived K_{PV}, as the goal was not to maximise performance but to provide a simple and interpretable baseline against which tree-based models could be compared.

Linear regression is implemented using the `scikit-learn` Python library (Pedregosa et al., 2011).

3.5 Random Forest Regression

After establishing a baseline using linear least squares regression, a Random Forest regression model is fitted for each modelling scenario.

Random Forest is an ensemble learning method designed for regression tasks and belongs to the family of bagging methods (Figure 3). It constructs multiple decision trees using bootstrap samples of the training data and averages their predictions to produce a final output. Bootstrap aggregation (also known as bagging) involves generating new training datasets by sampling the original data with replacement, introducing variability between the trees. This diversity increases robustness and reduces model variance, leading to improved generalisation performance compared with single decision trees (Breiman, 2001).

A key strength of Random Forest regression is its resistance to noise and overfitting. By averaging the outputs of many decorrelated trees—each trained on a different subset of data—the model provides stable and accurate predictions, even in the presence of non-linear and complex feature interactions. For this reason, Random Forest regression is well suited to this study, where the relationships between PV power output and meteorological variables—such as temperature, irradiance, and cloud cover—are expected to be highly complex and non-linear.

The key hyperparameters of the Random Forest model explored in this study are:

1. `n_estimators`: The number of decision trees in the forest. Increasing this generally improves performance and stability, but incurs greater computational cost.
2. `max_depth`: The maximum depth of each decision tree. Limiting the depth can reduce overfitting and improve generalisation, but may increase bias.
3. `min_samples_leaf`: The minimum number of samples required to be at a leaf node. Increasing this parameter leads to more conservative trees with fewer splits, which can reduce overfitting.
4. `min_samples_split`: The minimum number of samples required to split an internal node. Higher values can reduce variance but may increase bias, potentially leading to underfitting.
5. `max_features`: The fraction of features considered when determining the best split at each node. For example, a value of 0.7 means that 70% of the available features are randomly selected at each split. Smaller values encourage diversity across trees and help mitigate overfitting.
6. `max_samples`: The fraction of the training dataset used to build each tree. A value of 0.5 means each tree is trained on 50% of the original dataset (sampled with replacement). Smaller values can reduce training time but may increase the risk of underfitting.

Random Forest regression is implemented in this study using the `scikit-learn` Python library (Pedregosa et al., 2011).

3.6 Gradient Boosting Regression

Following the application of Random Forest regression, Gradient Boosting regression tree models are fitted for each modelling case to evaluate whether they can achieve better performance.

Gradient Boosting refers to a family of powerful ensemble learning techniques that approximate target functions by iteratively adding models in a sequential manner (Figure 3). Unlike bagging methods, which train multiple models independently on different subsets of data and aggregate

their predictions, boosting trains models sequentially, with each model focusing on correcting the errors of its predecessor (Friedman, 2001). This approach allows the ensemble to gradually improve prediction accuracy by reducing bias and capturing complex patterns in the data.

Gradient Boosting with regression trees begins by predicting the mean of the target variable as the initial estimate. The residuals—i.e., the differences between the observed values and this initial prediction—are then computed. A regression tree is then fitted to these residuals, learning how the model should adjust its predictions. The updated prediction is the sum of the initial estimate and the output of this first tree, scaled by a learning rate. This process is repeated iteratively: at each stage, a new tree is trained on the residuals of the current model, and its predictions are added to the ensemble. The procedure continues until a predefined stopping criterion is met, such as a maximum number of trees or convergence of the model error.

Gradient Boosting regression trees can achieve higher accuracy than Random Forests; however, they also have drawbacks. Training times are slower due to the sequential nature of tree construction, and they are more prone to overfitting than Random Forests (Natekin and Knoll, 2013). They are considered in this study because—like Random Forests—they are well suited to capturing the complex, non-linear relationships between PV power output and the relevant meteorological variables. If overfitting is successfully mitigated, it is hypothesised that this method may yield more accurate results than the Random Forest model.

The key hyperparameters of the Gradient Boosting model explored in this study are:

- `n_estimators`: The number of trees that are added sequentially. A larger number of trees allows the model to capture more complex relations at the cost of a higher risk of overfitting.
- `learning_rate`: A scaling factor applied to the contribution of each tree. It controls the trade-off between bias and variance.

along with `max_depth`, `min_samples_leaf`, `min_samples_split`, and `max_features`, which are explained in subsection 3.5.

Gradient boosting regression trees have been implemented in this study using the `scikit-learn` Python library (Pedregosa et al., 2011).

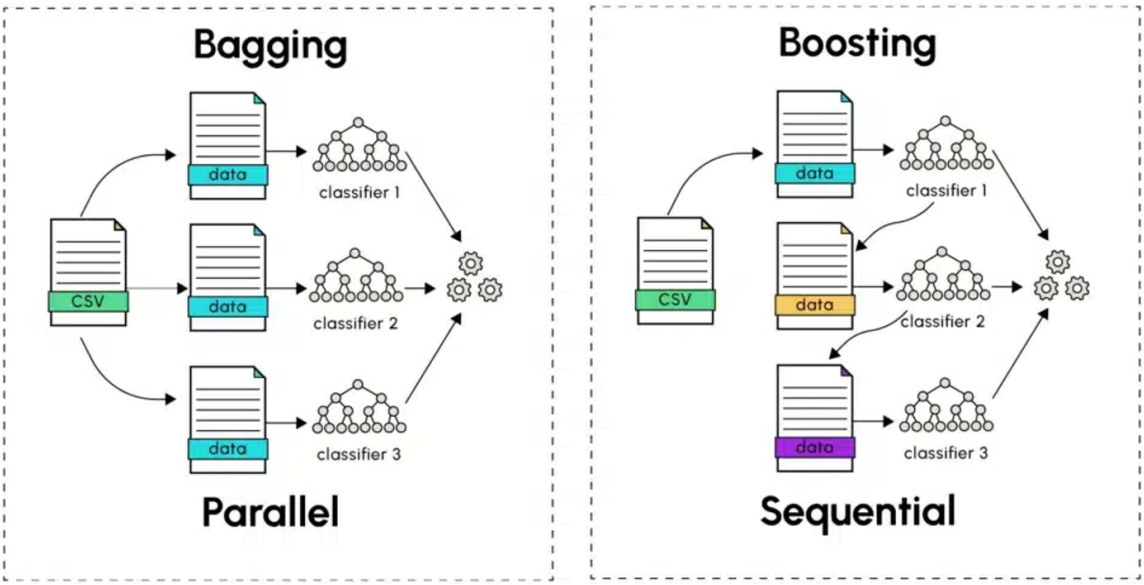


Figure 3: Visual representation of bagging and boosting (Daniel, 2025).

3.7 Model Performance

Model performance is evaluated using four common metrics, alongside an analysis of periodic bias. These metrics are: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE), and the coefficient of determination (R^2). These metrics are computed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

$$\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

Here, y_i denotes the observed value for the i-th sample, $\hat{y}_i$ represents the corresponding predicted value, $\bar{y}$ is the mean of all observed values, and n is the total number of observations. In this study, RMSE, MAE, and MBE are expressed in units of [W/m^2], while R^2 is dimensionless.

Each metric captures a different aspect of model performance. RMSE penalises larger errors more heavily, making it useful when large deviations are particularly undesirable. MAE is less sensitive to outliers and represents the average absolute prediction error. MBE indicates systematic bias, showing whether the model tends to over- or underpredict on average. R^2 expresses the proportion of variance in the observed data explained by the model, serving as a general measure of goodness of fit.

While MBE provides an overall measure of systematic bias, it does not capture temporal variations. In many climate and weather-related applications, it is important to assess whether the model exhibits recurring over- or underestimation patterns on shorter timescales. To address this, MBE is also computed on a monthly basis, following the approach of previous studies (Li et al., 2014; Budiyanto et al., 2018).

3.8 Hyperparameter Optimisation Strategy

To maximise the predictive performance of the Random Forest and Gradient Boosting models, a structured hyperparameter optimisation procedure was implemented. The search was conducted using the Optuna framework for Python (Akiba et al., 2019), which offers an efficient and flexible interface for automated hyperparameter tuning. Optuna supports a range of search strategies; in this work, the Tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al., 2011) was chosen. TPE is a Bayesian optimisation method that models the probability distributions of promising and unpromising hyperparameter configurations based on past evaluations, thereby focusing exploration on regions of the parameter space likely to yield improvements. This makes it considerably more sample-efficient than conventional grid or random search, which is advantageous for computationally demanding models.

The optimisation objective was adapted to the type of experiment:

- **Single-location models:** Hyperparameters were tuned to minimise the RMSE using three-fold cross-validation on the training set. This approach encourages good generalisation within the same site while mitigating overfitting. Final model performance was subsequently assessed on the validation dataset.

- **Cross-location models:** At each optimisation step, the model was trained on the training data from one site and evaluated on the training data from the other. The RMSE on the unseen location served as the optimisation criterion, promoting hyperparameters that transfer well between sites. Because the test set was not involved in this process, final performance could be evaluated on the complete test dataset rather than the validation subset.

All hyperparameters listed in Sections 3.5 and 3.6 were considered for tuning, except for a few fixed values to reduce computational overhead. For Random Forest, only `min_samples_split` was actively tuned among the three regularisation parameters (`max_depth`, `min_samples_leaf`, `min_samples_split`); preliminary experiments showed no significant benefit from optimising all three simultaneously, so `max_depth` and `min_samples_leaf` were retained at their default values (`None` and `1`, respectively).

For Gradient Boosting, the parameter `max_depth` was fixed at 30 to limit training time while maintaining sufficient model complexity to capture non-linear relationships, as boosting builds trees sequentially rather than in parallel.

Due to computational constraints, hyperparameter optimisation was performed only for datasets using K_{PV} derived from the relative production method. The resulting configurations were then also applied to models using PVWatts-derived K_{PV}. Likewise, for cross-location experiments, optimisation was carried out only on the `Delft_Wageningen` scenario, and the same hyperparameters were reused for `Wageningen_Delft`.

3.9 Feature Importance Evaluation

In this study, feature importance is quantified using a normalised impurity-based score, automatically computed by `scikit-learn` for both the Random Forest and Gradient Boosting regression models. These scores are accessible via the `feature_importances_` attribute.

The importance score of a feature is determined by the total reduction in impurity it causes across all trees in the ensemble. In regression trees, impurity is typically measured using the MSE. Whenever a node is split based on a feature, the resulting decrease in MSE is attributed to that feature. These reductions are accumulated over the entire model and then normalised such that the importances sum to one. A feature with an importance score of zero was not used in any split and thus had no influence on the model's structure.

It is important to note that impurity-based feature importance has several limitations. First, it does not directly indicate which features contribute to the model's ability to generalise to unseen data. A feature with a low importance score may still play a critical role in refining predictions—particularly in later stages of the decision process. For instance, the model might initially rely on top-of-atmosphere irradiance for a coarse prediction and then use cloud cover or weather-related features to make finer corrections. Furthermore, when features are strongly correlated, their importance can be diluted as they compete to explain the same variance in the target variable. In such cases, the model may distribute importance across correlated features in a way that under represents their individual contributions.

4 Results

This section presents the performance of the global radiation models. It begins with baseline results from the linear regression approach, followed by evaluation of the Random Forest and Gradient Boosting regressions. Differences in predictive accuracy between Delft and Wageningen are examined to assess the impact of site-specific conditions, and the monthly MBE is analysed to identify shorter-term biases. Finally, the optimised hyperparameters from automated tuning are summarised, offering insight into the configurations that achieved the best performance.

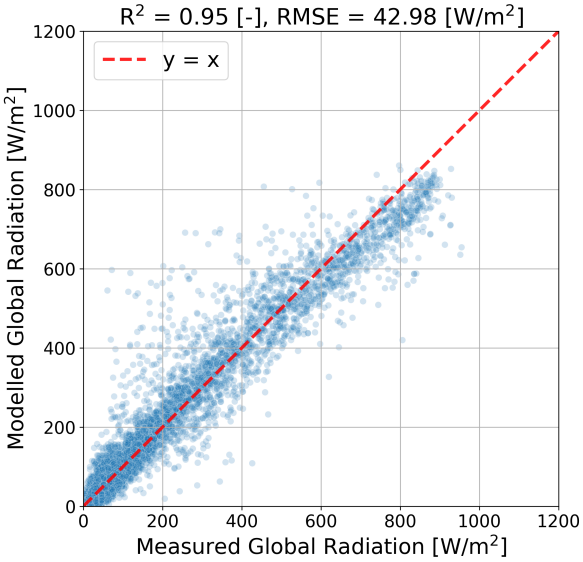
Table 3 provides an overview of the standard performance metrics for all modelling scenarios.

Table 3: Overview of performance metrics for all modelling scenarios and regression methods.

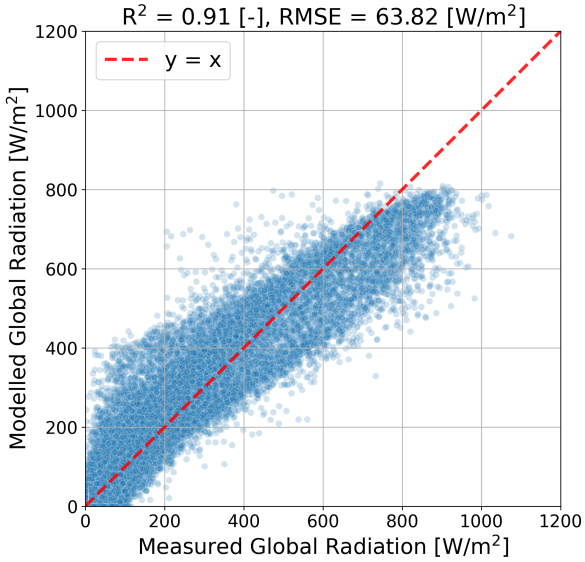
Scenario and Model Type	**RMSE**	**MAE**	**MBE**	**R^2**
	W/m^2	W/m^2	W/m^2	–
Scenario: `Delft_Single`				
Linear Regression	42.98	19.32	1.30	0.95
Random Forest – Relative production	31.80	10.62	-0.52	0.97
Random Forest – PVWatts	31.78	10.60	-0.50	0.97
Gradient Boosting – Relative production	30.73	10.06	-0.50	0.97
Gradient Boosting – PVWatts	30.43	9.96	-0.34	0.97
Scenario: `Wageningen_Single`				
Linear Regression	63.82	36.36	5.67	0.91
Random Forest – Relative production	39.63	15.68	0.30	0.97
Random Forest – PVWatts	39.64	15.69	0.22	0.97
Gradient Boosting – Relative production	38.65	14.97	0.11	0.97
Gradient Boosting – PVWatts	38.54	14.96	0.02	0.97
Scenario: `Delft_Wageningen`				
Linear Regression	109.87	62.58	40.44	0.73
Random Forest – Relative production	60.48	29.18	7.34	0.92
Random Forest – PVWatts	60.88	28.97	9.06	0.92
Gradient Boosting – Relative production	59.94	30.16	4.81	0.92
Gradient Boosting – PVWatts	59.90	30.29	6.59	0.92
Scenario: `Wageningen_Delft`				
Linear Regression	68.44	36.68	2.44	0.87
Random Forest – Relative production	44.25	18.16	-4.15	0.95
Random Forest – PVWatts	47.12	19.16	-7.01	0.94
Gradient Boosting – Relative production	44.50	19.44	-2.07	0.95
Gradient Boosting – PVWatts	46.03	20.54	-4.17	0.94

4.1 Linear Regression Results

This section presents results for the baseline linear regression model across the four modelling scenarios. Predicted global radiation values are plotted against measured pyranometer data from the test sets. The results for the single-location scenarios are shown in Figure 4, while the cross-location scenarios are shown in Figure 5.

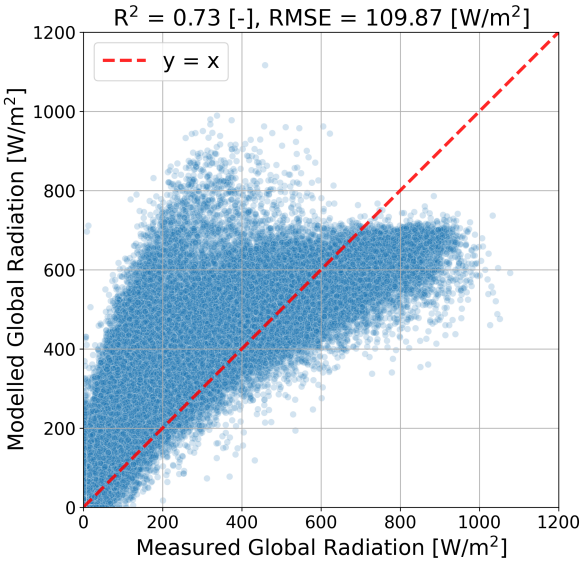


(a) Scenario: `Delft_Single`

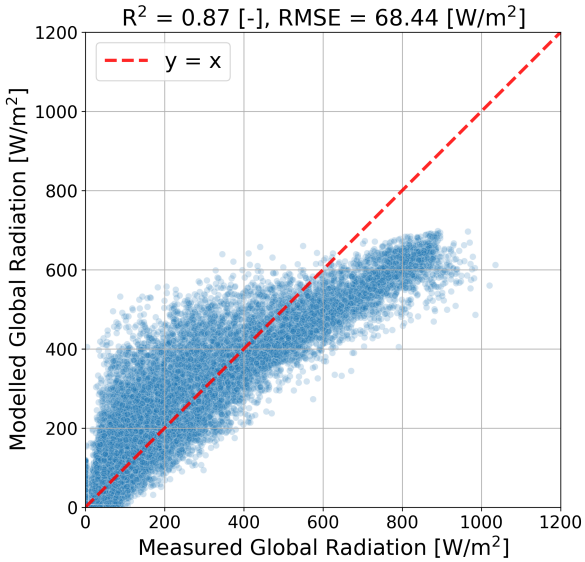


(b) Scenario: `Wageningen_Single`

Figure 4: Linear regression model results for both single-location scenarios.



(a) Scenario: `Delft_Wageningen`



(b) Scenario: `Wageningen_Delft`

Figure 5: Linear regression model results for both cross-location scenarios.

The linear model performs reasonably well for the `Delft_Single` scenario, although it exhibits a clear tendency to underestimate global radiation at higher values. This can be attributed to the model's assumption of a linear relationship between input features derived from PV power output, and global radiation, whereas the actual relationship becomes increasingly non-linear at higher irradiance levels, mostly due to temperature effects (Buni et al., 2018).

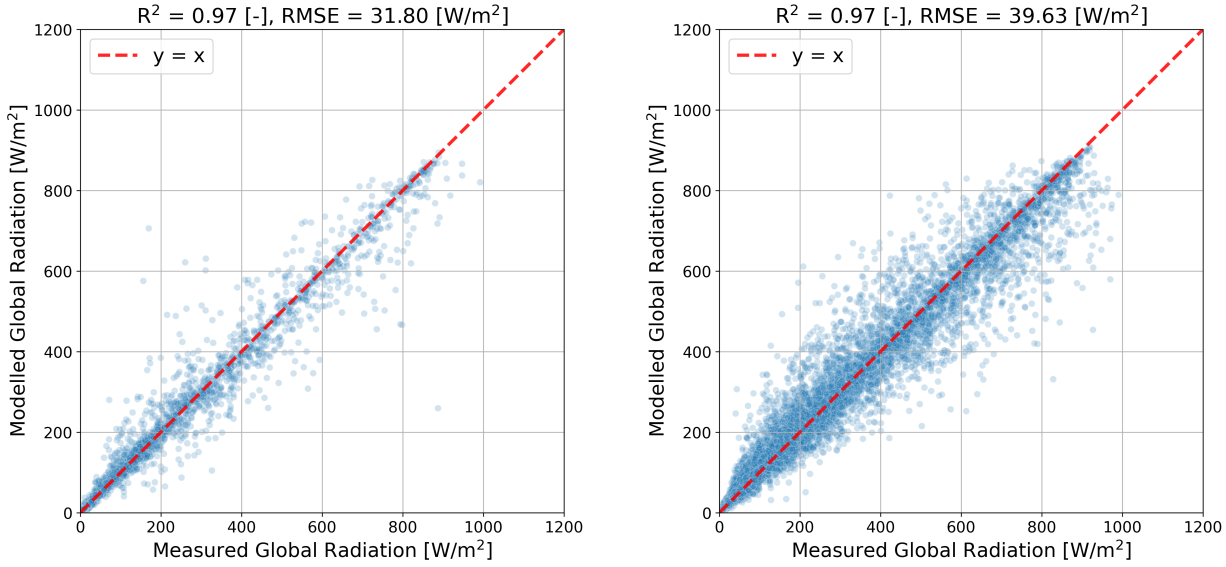
In contrast, the model performs significantly worse for the `Wageningen_Single` scenario, with the RMSE going up from 43 to 64 [W/m^2]. This trend is also observed across other models and is discussed in more detail in Section 4.4. Additionally, the model shows a systematic overestimation of global radiation in this case, with a MBE of 5.66 [W/m^2]. Similar to the Delft scenario, it also underestimates values at the higher end of the irradiance spectrum.

The cross-location scenarios show a clear decline in model performance, which is consistent with expectations. This reduction can be attributed to site-specific factors such as shading and panel

orientation, both of which strongly influence the relationship between the input features and global radiation. As a result, a model trained on one site is inherently limited when applied to another with distinct characteristics. Even so, both cross-location linear regression models retain some predictive skill, with R^2 values above 0.7, indicating that they are able to reproduce the overall trend. However, the very high RMSE of 110 [W/m^2] observed in the `Delft_Wageningen` scenario highlights the inability of linear regression to capture the more complex, non-linear relationships required for robust generalisation across sites. This suggests that while linear models can provide a baseline, they are not well-suited for developing generalisable models of global radiation.

4.2 Random Forest Regression Results

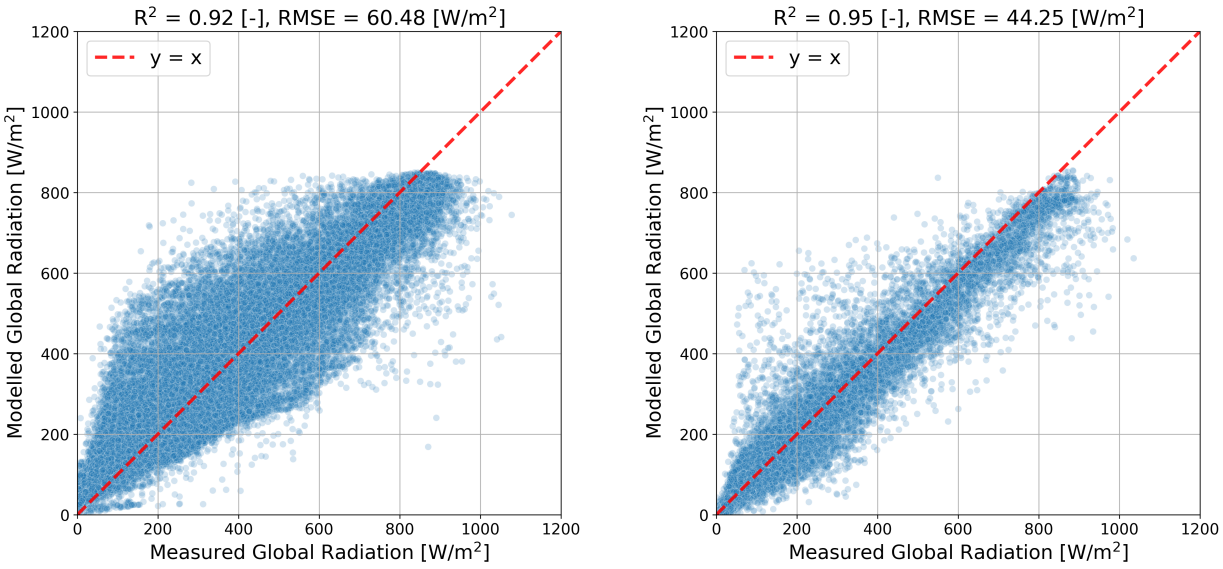
The performance of the Random Forest regression models across all modelling scenarios is presented here. Results obtained using the Clear Sky Index for Photovoltaics (K_{PV}), calculated via the relative production approach, are shown in Figure 6 for the single-location scenarios and in Figure 7 for the cross-location scenarios. Similarly, results based on K_{PV} values computed using the PVWatts model are shown in Figures 8 and 9.



(a) Scenario: `Delft_Single`

(b) Scenario: `Wageningen_Single`

Figure 6: Random Forest regression model using K_{PV} derived from relative production. Results for both single-location scenarios.



(a) Scenario: `Delft_Wageningen`

(b) Scenario: `Wageningen_Delft`

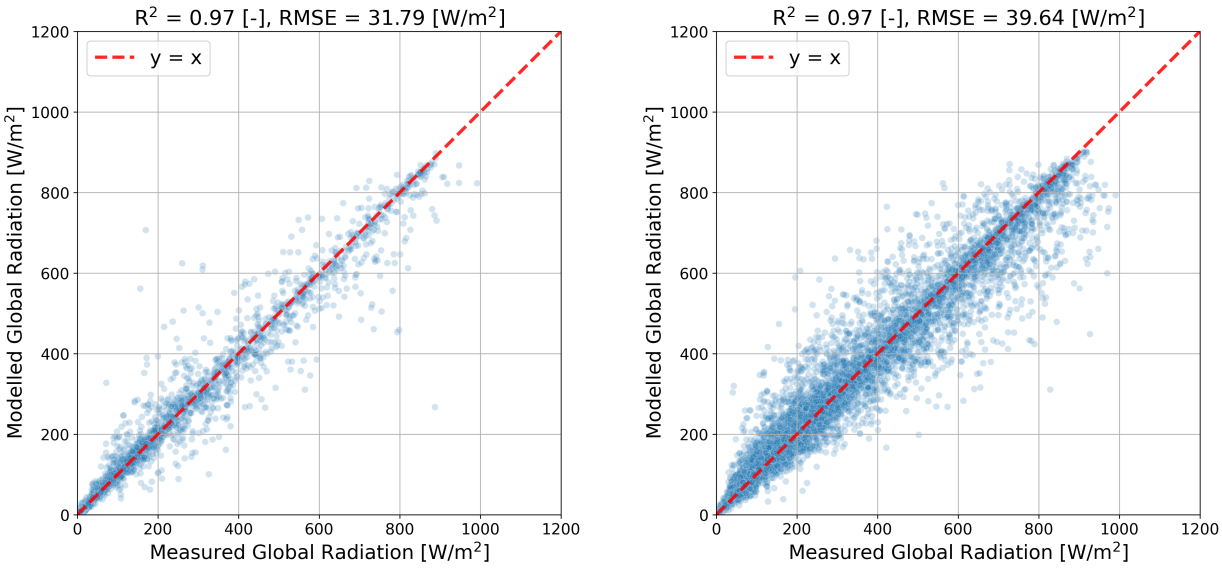
Figure 7: Random Forest regression model using K_{PV} derived from relative production. Results for both cross-location scenarios.

From Figure 6, it is evident that the Random Forest models using K_{PV} derived from relative production achieve substantially higher performance than the baseline linear regression models for the single-location scenarios. Unlike linear regression, which assumes a strictly linear relationship between PV power and radiation, Random Forests can capture complex, non-linear dependencies and interactions among input features. This flexibility allows the model to reproduce both the general trend and the more subtle variations in the data. For both locations, the models are capable of reproducing global radiation values with high accuracy.

When considering the estimated pyranometer measurement uncertainties of 23 [W/m^2] for Delft and 13 [W/m^2] for Wageningen, the `Delft_Single` model demonstrates excellent agreement with measured data, yielding an RMSE of 31.80 [W/m^2] and a MAE of 10.62 [W/m^2]. However, consistent with the trend observed in the linear regression results, performance for the `Wageningen_Single` scenario is notably lower. This performance gap is likely due to a stronger correlation between PV power output and global radiation at the Delft site compared to the Wageningen site. Further discussion of this difference is provided in Section 4.4.

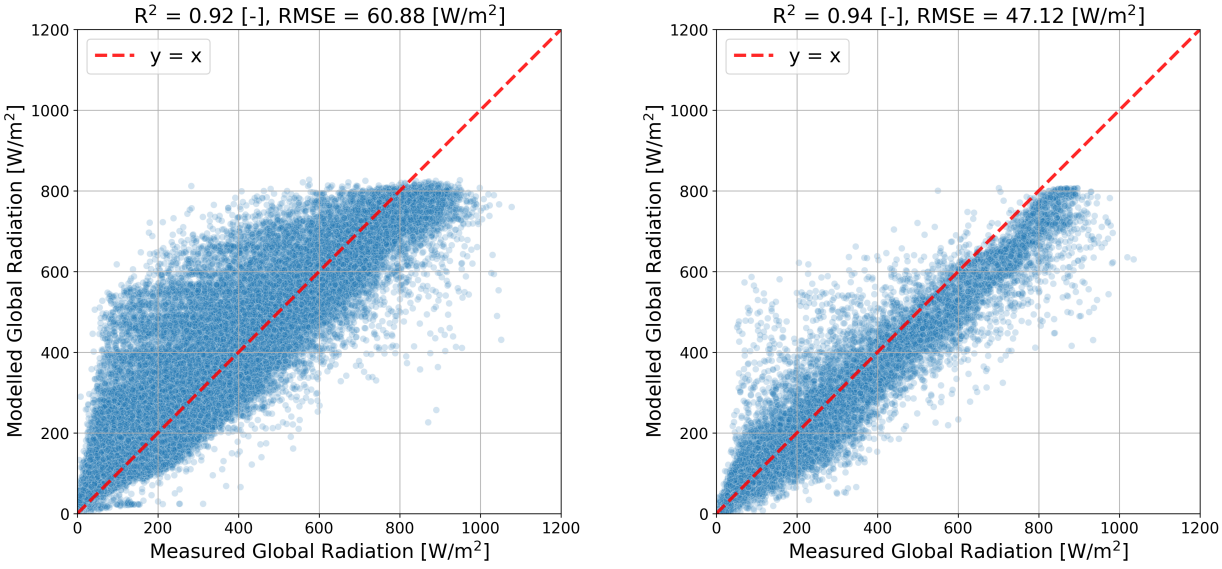
For the cross-location scenarios, performance has also improved substantially compared to the linear baseline model. This improvement highlights the advantage of Random Forests in dealing with shifts in feature–target relationships between sites. While linear models are constrained by a single global coefficient for each feature, Random Forests can partition the feature space in a more complex manner, allowing in this case for better adaptation to site-specific conditions.

It can again be noted that the model trained on Wageningen data and evaluated on Delft data outperforms the reverse case by a considerable margin. This asymmetry is once more attributed to the stronger correlation between PV power output and global radiation in the Delft dataset.



(a) Scenario: `Delft_Single` *(b) Scenario:* `Wageningen_Single`

Figure 8: Random Forest regression model using K_{PV} derived from PVWatts. Results for both single-location scenarios.



(a) Scenario: `Delft_Wageningen` *(b) Scenario:* `Wageningen_Delft`

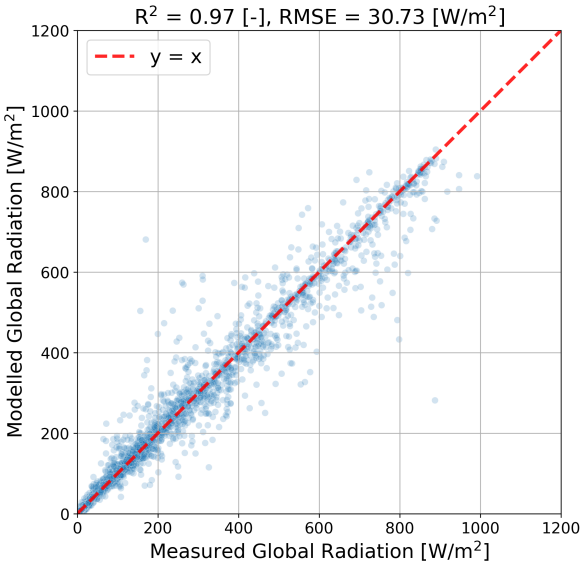
Figure 9: Random Forest regression model using K_{PV} derived from PVWatts. Results for both cross-location scenarios.

When comparing the relative production approach with the PVWatts-based K_{PV} (Figures 8 and 9), no significant differences are observed. As shown in Table 3, the performance metrics for both approaches are nearly identical, indicating that the choice of K_{PV} calculation method has little impact on Random Forest performance in this study.

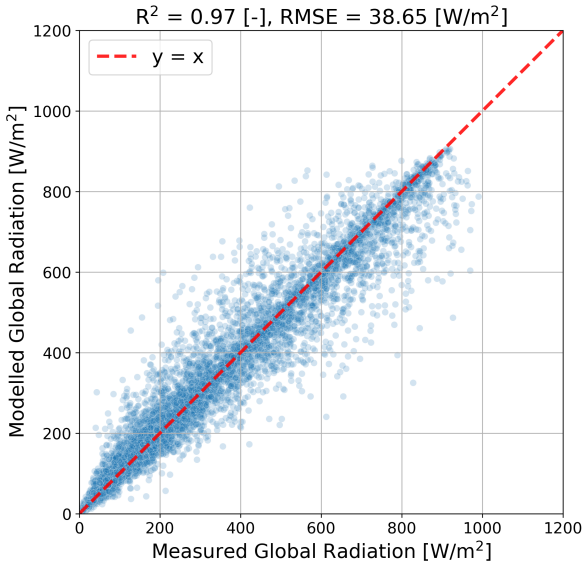
4.3 Gradient Boosting Regression Results

This section presents the results for the Gradient Boosting regression models across all modelling scenarios. Figures 10 and 11 display the results for the single- and cross-location scenarios, respectively, when using K_{PV} values computed via the relative production approach. Similarly,

Figures 12 and 13 present the results obtained with K_{PV} values calculated using the PVWatts model.

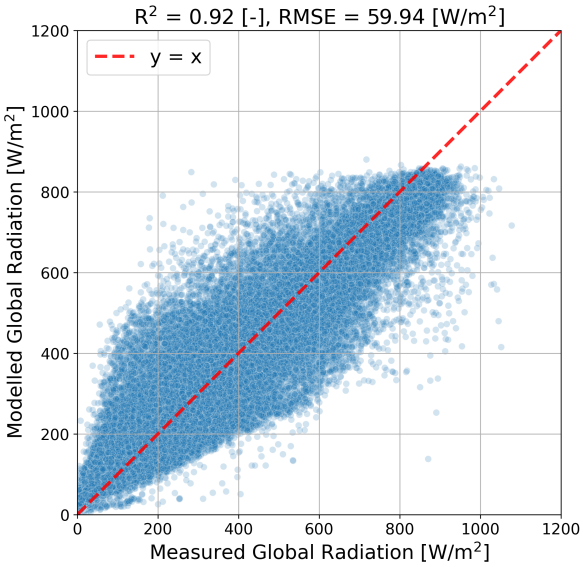


(a) Scenario: `Delft_Single`

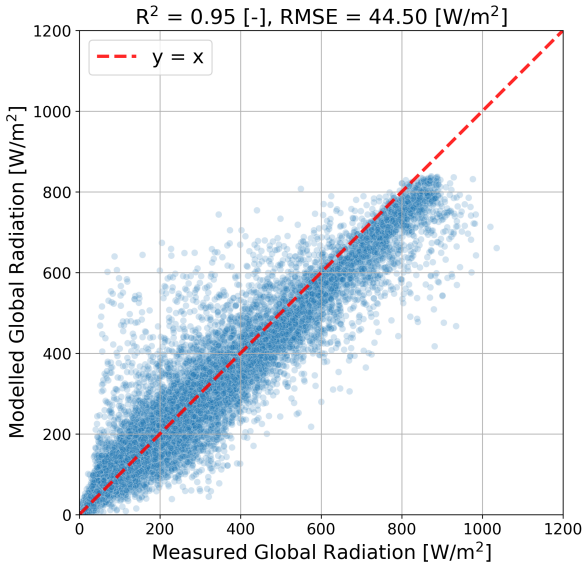


(b) Scenario: `Wageningen_Single`

Figure 10: Gradient Boosting regression model using K_{PV} derived from relative production. Results for both single-location scenarios.



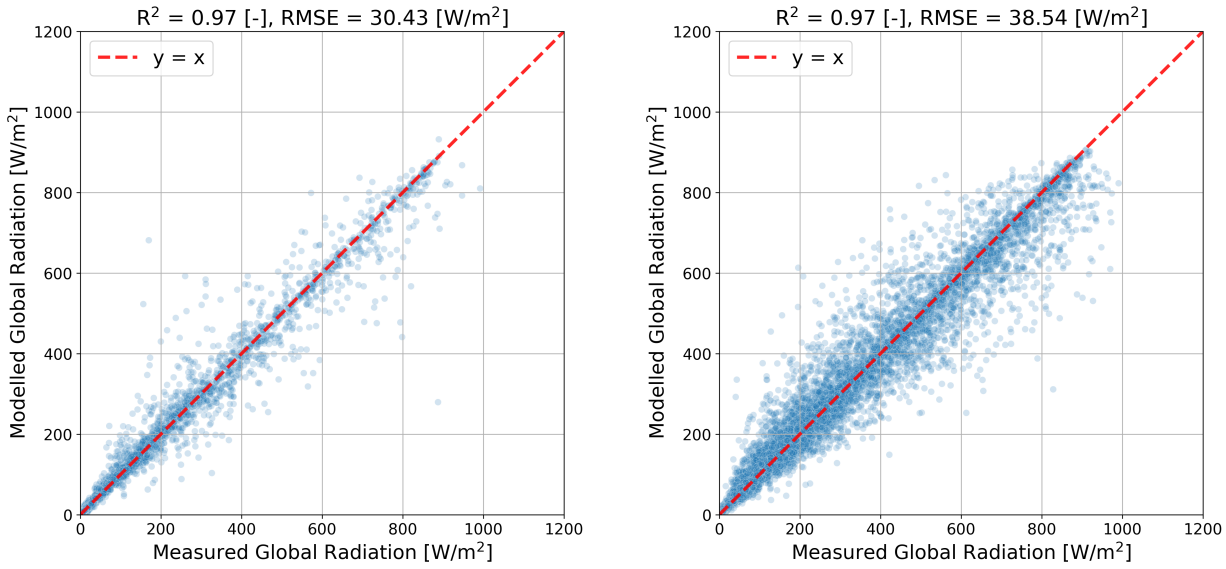
(a) Scenario: `Delft_Wageningen`



(b) Scenario: `Wageningen_Delft`

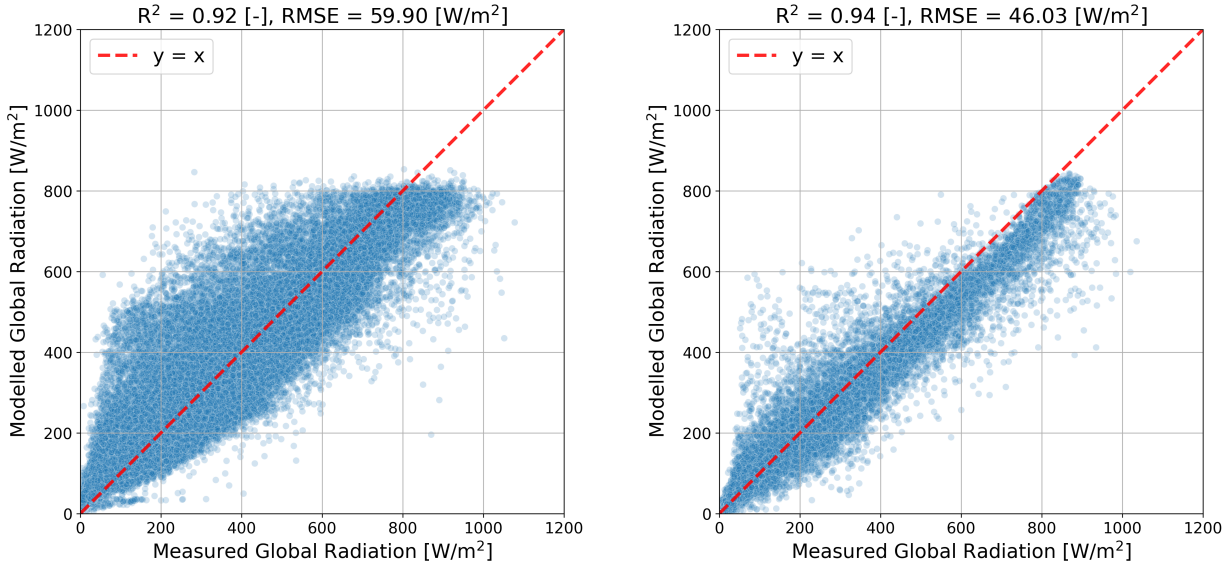
Figure 11: Gradient Boosting regression model using K_{PV} derived from relative production. Results for both cross-location scenarios.

When K_{PV} is derived from the relative production approach, the Gradient Boosting models perform very similarly to the Random Forest models based on the same K_{PV}. As shown in Table 3, differences across most performance metrics are minimal. The most notable distinction observed here is a lower MBE in the cross-location scenarios, indicating reduced systematic bias compared to the Random Forest models. This suggests that Gradient Boosting is better able to fine-tune predictions under differing site conditions, even if overall error metrics remain similar.



(a) Scenario: `Delft_Single` *(b) Scenario:* `Wageningen_Single`

Figure 12: Gradient Boosting regression model using K_{PV} derived from PVWatts. Results for both single-location scenarios.



(a) Scenario: `Delft_Wageningen` *(b) Scenario:* `Wageningen_Delft`

Figure 13: Gradient Boosting regression model using K_{PV} derived from PVWatts. Results for both cross-location scenarios.

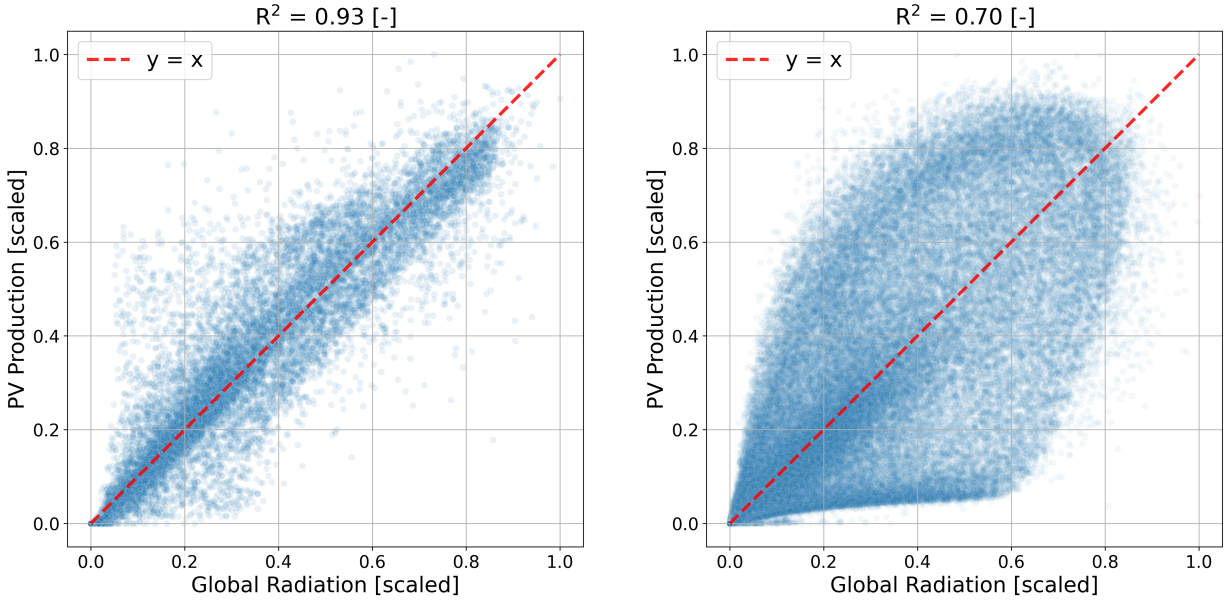
For the single-location scenario, Gradient Boosting models using K_{PV} from PVWatts show marginally better performance than those using K_{PV} from the relative production approach. In contrast, for cross-location scenarios, their performance is slightly lower. As with the Random Forest models, the choice of K_{PV} calculation method has only a minor effect on Gradient Boosting model performance in this study.

4.4 Difference in Results Between Delft and Wageningen

An analysis of the single-location results reveals that the `Wageningen_Single` scenario consistently underperforms compared to `Delft_Single` across all model types. A similar trend is observed in the cross-location experiments: models trained on Wageningen data and evaluated on Delft data

outperform the reverse configuration. The most plausible explanation lies in the difference in PV panel orientation between the two sites, which alters the relationship between global radiation and PV power output.

Figure 14 plots global radiation against PV power (both normalised to unit scale) for each site. It is evident that Delft exhibits a stronger and more linear correlation between the two variables compared to Wageningen.



(a) Location Delft

(b) Location Wageningen

Figure 14: Global radiation plotted against photovoltaic power data (unit scale) for both locations.

To illustrate that the weaker correlation at Wageningen is primarily driven by panel orientation, Figure 15 shows data for clear summer days at the Wageningen site. The dataset was filtered to include only summer dates with total cloud cover (N) below 0.3 (fractional area), ensuring that the relationship between global radiation and PV power could be observed with minimal cloud interference. The colour bar indicates the time of day for each measurement. The plot reveals that PV power increases only marginally during the morning, despite steadily rising global radiation, until shortly before noon—when the sun passes over a rooftop notch that shades the panels. As a result, peak power production occurs later in the afternoon, when the sun–panel angle becomes more favourable, rather than at solar noon when global radiation is highest.

The data points that lie closer to the $y = x$ reference line in Figure 14b correspond mainly to cloudier days. On such days, PV power generation is driven more by diffuse than by direct radiation. Because diffuse radiation is less dependent on the angle between the sun and the PV panels, the relationship between global radiation and PV output becomes more linear.

Together, these two Figures show how panel orientation can noticeably change the relation between global radiation and PV power production.

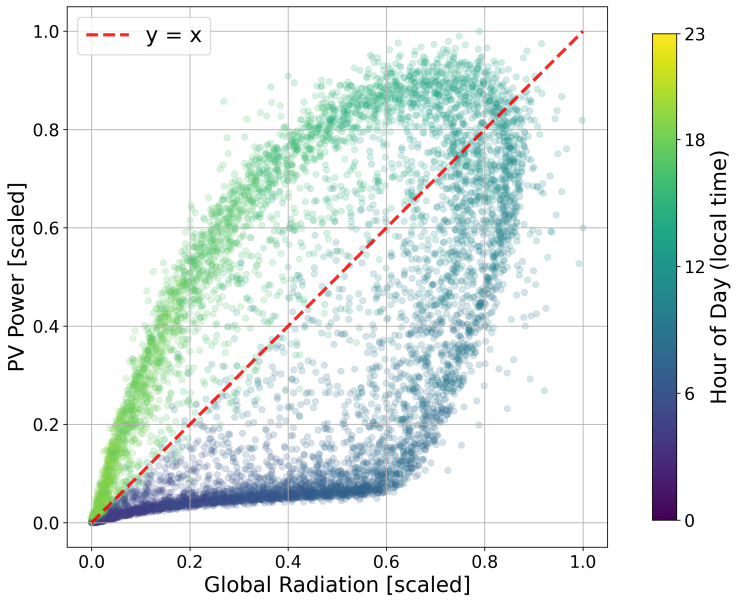


Figure 15: Global radiation versus photovoltaic power (unit scale) for clear summer days in Wageningen. Colours indicate the time of day of measurement.

4.5 Monthly Mean Bias Error

As outlined in the Methodology, the MBE was also evaluated on a monthly basis to assess potential seasonal trends in model performance. To ensure statistical reliability, only months with at least 200 data points in the validation set were considered.

The results for the single-location scenarios are presented in Figure 16, while the cross-location results are shown in Figure 17. Both figures use box plots to summarise the distribution of monthly MBEs, with whiskers extending to 1.5 times the interquartile range.

The distributions indicate that the two machine learning models yield highly similar results, and that the choice of K_{PV} calculation method—whether based on relative production or PVWatts—has only a minor influence on the observed monthly biases of the single-location scenarios. For the cross-location scenario `Wageningen_Delft` however it can be seen that using the relative production based K_{PV} gives a notably tighter distribution, closer to the zero line.

These figures highlight another important point: the mean bias error computed over the entire dataset can mask substantial month-to-month variability. For instance, the `Wageningen_Single` scenario with the Random Forest relative production model has a very low MBE of 0.30 [W/m^2] over the whole data range, but some individual months can have biases as high as 13 [W/m^2].

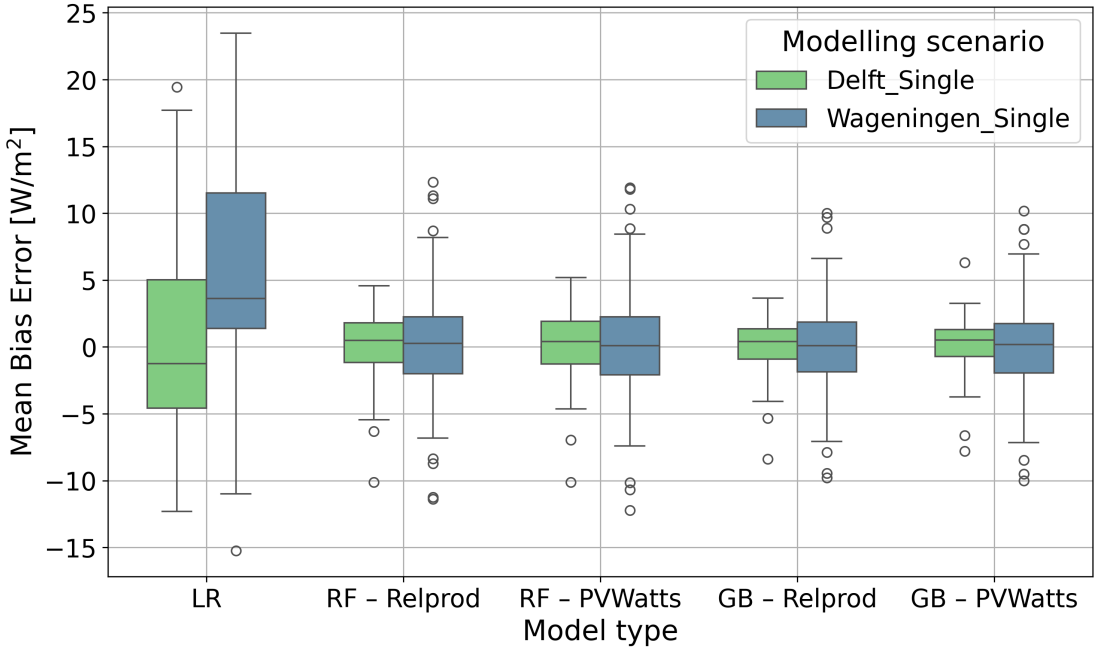


Figure 16: Distribution of monthly MBEs for the single-location scenarios. Whiskers extend to 1.5 *times the interquartile range.*

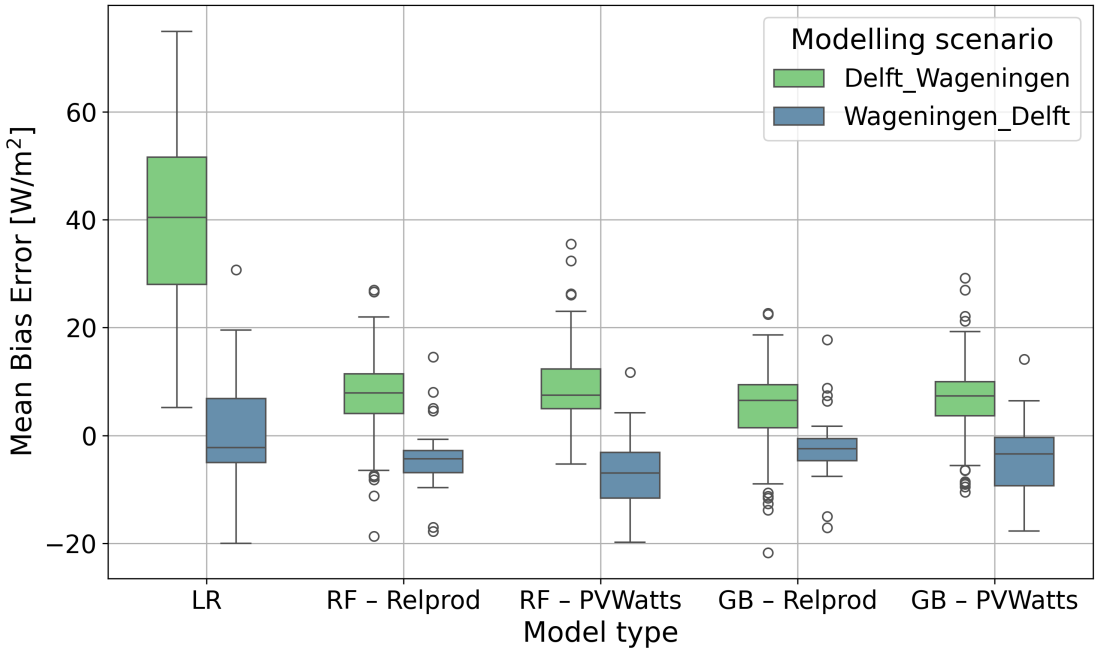


Figure 17: Distribution of monthly MBEs for the cross-location scenarios. Whiskers extend to 1.5 *times the interquartile range.*

4.6 Optimised Hyperparameters

Table 4 summarises the hyperparameter configurations obtained through optimisation with the Optuna framework for Python. As outlined in the Methodology (Section 3.8), optimisation was performed exclusively on the dataset where K_{PV} was derived using the relative production approach. For the cross-location experiments, optimisation was carried out only for the `Delft_Wageningen` scenario.

Consequently, the optimised hyperparameters identified in these settings were reused for the cor-

responding models based on K_{PV} derived from PVWatts, as well as for the `Wageningen_Delft` scenario. This approach ensured methodological consistency while limiting computational overhead.

Table 4: Overview of model configurations for each modelling scenario.

Scenario and Model Type	**n_estimators**	**min_samples_split**	**max_features**	**learning_rate**	**max_depth**
SCENARIO: `Delft_Single`					
Random Forest – Relative production	1500	2	0.65	–	**None**
Random Forest – PVWatts	1500	2	0.65	–	**None**
Gradient Boosting – Relative production	400	83	0.30	0.0265	30
Gradient Boosting – PVWatts	400	83	0.30	0.0265	30
SCENARIO: `Wageningen_Single`					
Random Forest – Relative production	900	3	0.65	-	**None**
Random Forest – PVWatts	900	3	0.65	-	**None**
Gradient Boosting – Relative production	400	95	0.30	0.0201	30
Gradient Boosting – PVWatts	400	95	0.30	0.0201	30
SCENARIO: `Delft_Wageningen`					
Random Forest – Relative production	700	236	0.50	–	**None**
Random Forest – PVWatts	700	236	0.50	–	**None**
Gradient Boosting – Relative production	400	2	0.40	0.0096	30
Gradient Boosting – PVWatts	400	2	0.40	0.0096	30
SCENARIO: `Wageningen_Delft`					
Random Forest – Relative production	700	236	0.50	–	**None**
Random Forest – PVWatts	700	236	0.50	–	**None**
Gradient Boosting – Relative production	400	2	0.40	0.0096	30
Gradient Boosting – PVWatts	400	2	0.40	0.0096	30

4.7 Feature Importance Results

Feature importance scores for the Random Forest and Gradient Boosting models using K_{PV} derived from the relative production approach are shown in Figure 18 for the single-location scenarios and in Figure 19 for the cross-location scenarios. Feature importance scores for the models using PVWatts were not significantly different and as such are not shown here. They can be found in Appendix III.

For the single-location models, feature importance is dominated by PV power output, Top of Atmosphere radiation (TOA), and K_{PV}. As shown in Figure 18, the `Delft_Single` models rely more heavily on PV power output and less on TOA, whereas the `Wageningen_Single` models show the opposite trend. This is consistent with the stronger correlation between global radiation and PV power output observed in Delft, as discussed in Section 4.4. The reanalysis weather data features exhibit low or negligible importance scores, indicating that they contribute minimally to

the model's predictions.

In the cross-location models, a similar pattern emerges. For the `Delft_Wageningen` scenario, the models rely more on K_{PV} than on TOA, while the opposite is observed for `Wageningen_Delft`. Additionally, the importance of the time-shifted K_{PV} increases compared to the single-location models, indicating that this feature provides useful information for improving spatial generalisation.

It is important to note that PV power output itself is excluded from the cross-location models due to its strong dependence on site-specific conditions (see Section 3.2). Similarly, the relative output R_{PV} was removed after early testing showed poor generalisation across locations.

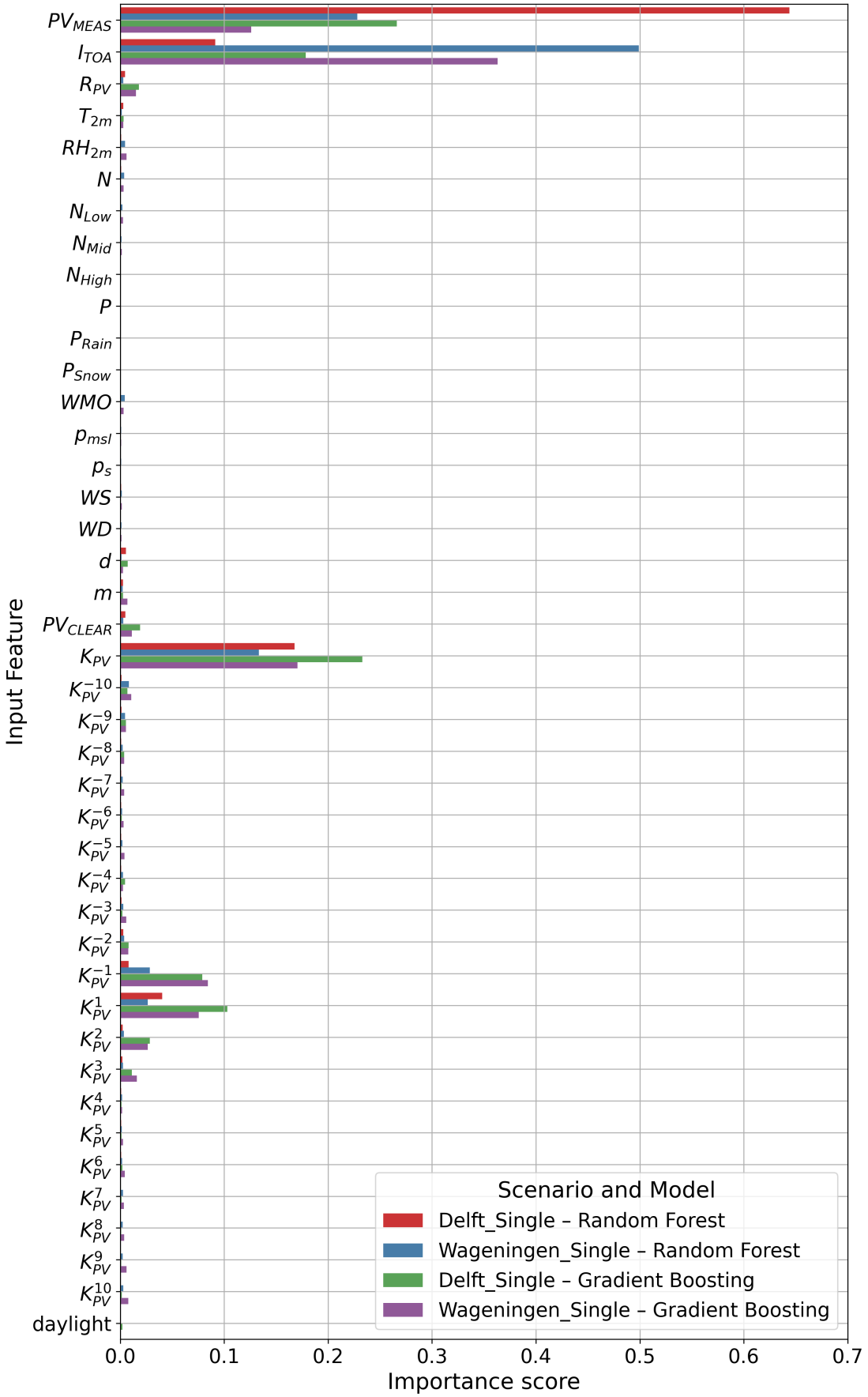


Figure 18: Feature importance scores for the single-location models trained with K_{PV} derived from the relative production method.

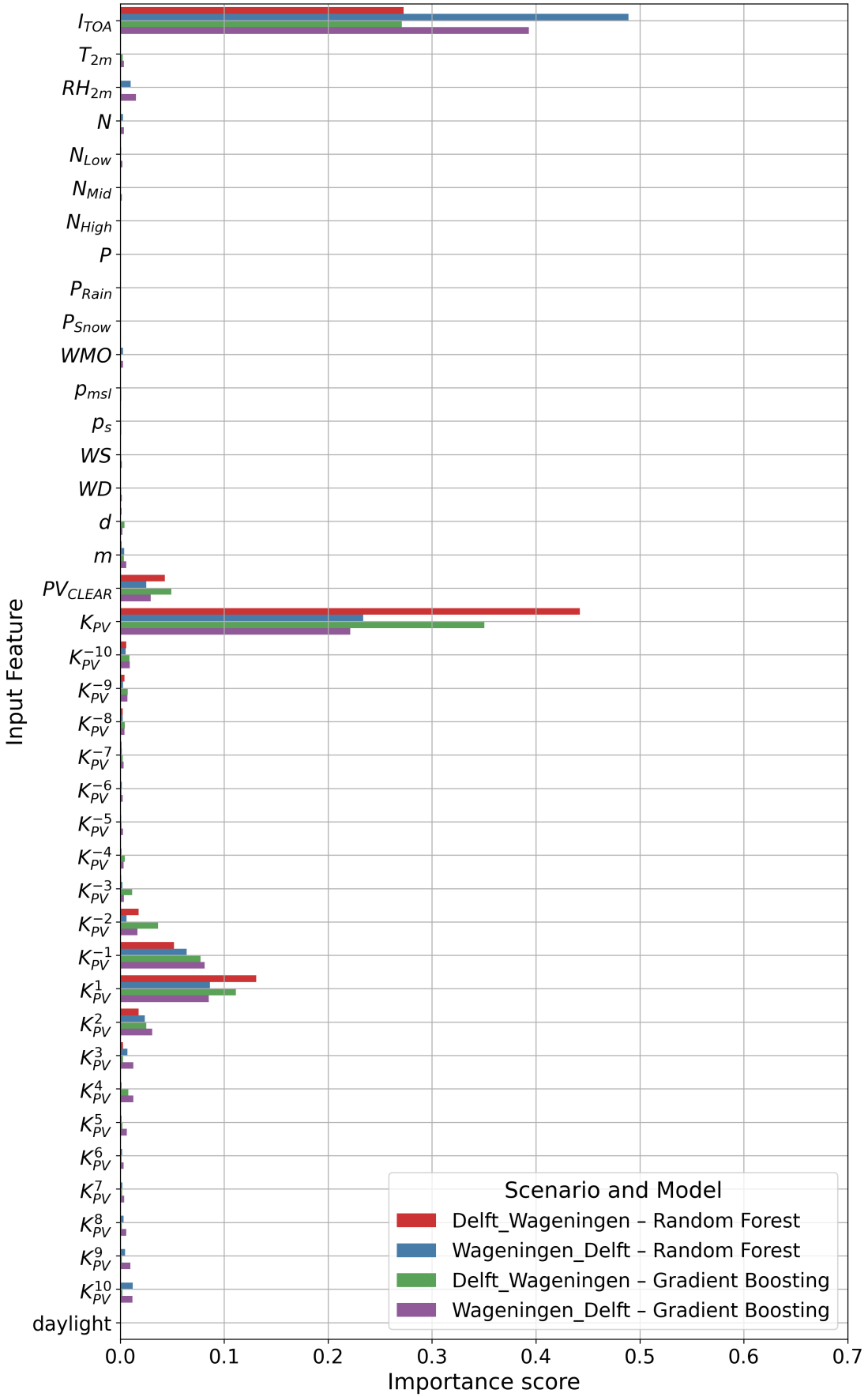


Figure 19: Feature importance scores for the cross-location models trained with K_{PV} derived from the relative production method.

4.8 Summary of Results

Overall, the results demonstrate that global solar radiation can be estimated with high accuracy from PV power output data. Linear regression provides a reasonable baseline for single-location scenarios, but its performance drops sharply when applied across locations, highlighting its limited ability to generalise. In contrast, tree-based machine learning models (Random Forest and Gradient Boosting) achieve substantially higher accuracy, with $R^2 \approx 0.97$, MAEs of 10–15 [W/m^2], and very low bias (|MBE| < 0.6 [W/m^2]) at a 15-minute timescale. Even in the more challenging cross-location scenarios, these models maintain reasonable performance ($R^2 \approx 0.92$, MAE 15–30 [W/m^2]). Performance is consistently stronger for Delft than Wageningen, reflecting site-specific differences in the correlation between PV output and radiation. Between the two machine learning methods, Gradient Boosting tends to reduce systematic bias compared to Random Forest, particularly in cross-location settings. Feature importance analysis further highlights PV power output and K_{PV} (including its time-shifted form) as the most influential predictors. Finally, differences between the two approaches for calculating K_{PV} (relative production vs. PVWatts) were found to have only a minor impact on model performance.

5 Discussion

This study demonstrates that global solar radiation can be estimated with high accuracy and low bias using machine learning models trained on photovoltaic power output data. Both Random Forest and Gradient Boosting achieved strong performance, with R^2 values of 0.97 and MAEs of 10–15 [W/m^2], while still retaining reasonable accuracy when applied across locations. Model performance varied by site, with consistently better results for Delft than for Wageningen, likely reflecting a stronger correlation between PV power output and radiation in the Delft dataset.

Feature importance analysis highlighted the central role of PV power output and K_{PV}, with time-shifted K_{PV} values further improving spatial generalisation between sites. By contrast, weather reanalysis variables contributed little, suggesting that under the conditions examined here, PV-based features alone are sufficient to capture most of the radiation variability. Nonetheless, reanalysis data may still provide added value if other variables are considered or if the models are applied in climates with stronger atmospheric variability.

These results confirm that tree-based machine learning models are well suited to capturing the complex, non-linear relationship between PV power output and solar radiation, which explains their strong predictive skill. The weak performance of linear regression supports this point: while it reproduced the overall trend, its errors were much larger, its bias higher, and its R^2 markedly lower. This performance gap highlights the need for flexible non-linear models capable of reflecting the complex PV response to changing atmospheric conditions. The strong role of K_{PV} in cross-location models reinforces the findings of Engerer and Mills (2014), namely that normalising PV output relative to clear-sky performance improves data comparability across sites.

When benchmarked against established alternatives, the advantages of this approach become clear. For instance, Urraca et al. (2018) reported an average MAE of 23.13 [W/m^2] and an average MBE of 4.54 [W/m^2] for the ERA5 reanalysis product when estimating daily averages of global radiation. In contrast, the single-location models developed here achieved substantially lower MAEs on a much finer 15-minute timescale, while also maintaining a near-zero bias ($|\text{MBE}| < 0.6$ [W/m^2]). This demonstrates a key strength of supervised PV-based models: they provide accurate, high-frequency estimates while avoiding systematic errors that can easily propagate into downstream applications such as PV yield assessments, grid operation strategies, or agricultural planning.

To put these errors into perspective, the MAEs of 10–15 [W/m^2] are smaller than typical magnitudes of important urban climatology fluxes such as net all-wave radiation (100–200 [W/m^2]) and sensible heat flux (50–150 [W/m^2]) Oke (2002). This highlights the potential of PV-based radiation modelling to generate data that is not only accurate for PV yield applications but also valuable for urban climate and energy exchange studies.

As expected, performance declined in cross-location applications, but error levels remained comparable to those of ERA5 despite operating at a much higher temporal resolution. Similarly, the RMSE values of 30–40 [W/m^2] observed in the single-location models are similar to the findings of Nespoli and Medici (2017), who derived GHI from aggregated PV signals using an unsupervised, physics-based approach. While their method avoids the need for ground-truth pyranometer data, it relies more heavily on model assumptions and was validated at only two sites in Switzerland.

The comparison between K_{PV} derived via the relative production method and the PVWatts model revealed only small performance differences. The most notable case was `Wageningen_Delft`, where the relative production method yielded lower biases that were also more consistent over time, as shown in the monthly MBE analysis. This suggests that a simple approach to simulating clear-sky power output is sufficient to obtain robust results, at least in the context of this study.

The monthly MBE analysis also revealed that the mean bias error, when computed over the entire dataset, can obscure considerable month-to-month variability, as periods of over- and underestimation may cancel each other out in the long-term average.

An additional question concerns how well the models can be applied when no data from the target site are available at all. Appendix IV explored this "true transferability" scenario by applying single-location hyperparameters directly to another site without any tuning on its data. The resulting performance was very close to that of models optimised specifically for cross-location use, and in one case even slightly better. This suggests that much of the skill in cross-site prediction stems from the general structure of the models rather than site-specific hyperparameter choices, supporting their potential as accessible tools in data-sparse settings.

Another practical aspect is the availability of input variables. Because the main experiments included reanalysis features and a time-lead version of K_{PV}, they cannot be deployed in real time without modification. As shown in Appendix V, retraining the Random Forest models with only instantaneous features (removing reanalysis data and the lead K_{PV}) led to only a marginal loss in accuracy. This indicates that the proposed approach can be used not only for retrospective analyses but also for near real-time monitoring of global radiation.

Most optimised hyperparameters were unremarkable, with one exception: in all Gradient Boosting cross-location models, the parameter `min_samples_split` was set to two by the optimisation process. Higher values were expected for cross-location scenarios (as observed in Random Forest models), since low values can encourage overfitting to site-specific training data. In this case, however, overfitting was likely prevented by the maximum tree depth of 30 and the relatively low learning rate.

These findings suggest that in locations lacking direct radiation measurements, but with access to PV power data, global radiation can be modelled with high accuracy and low bias at a fine temporal resolution using relatively simple tree-based machine learning models. Such models could provide a cost-effective means of deriving reliable radiation estimates, particularly for applications requiring high-frequency data. Nevertheless, their practical use still requires temporary deployment of a pyranometer for site-specific calibration.

This reliance on local calibration is a key limitation of the approach. Without pyranometer data for training, the high accuracy achieved in the single-location models cannot be replicated. This motivated the exploration of cross-location models, which showed some generalisation capability but still underperformed compared to locally trained models. Other limitations include the narrow scope of test sites and configurations: both sites were located in the Netherlands, and only two panel orientations were considered. Even so, the sites exhibited substantially different PV–radiation relationships, and the models performed well in both cases, supporting their broader applicability. A further limitation concerns measurement uncertainty: the pyranometer used in Delft had a combined mean error of approximately ± 23 [W/m^2] (Appendix II), which constrains the maximum achievable model accuracy. This could be addressed by using pyranometers with higher precision.

Although this study demonstrates the strong potential of PV-based machine learning models, further research is needed to enhance their robustness, generalisability, and practical applicability. Several directions are apparent. First, the findings should be validated on a broader range of sites and PV configurations to confirm their broader applicability. With access to more diverse datasets, a logical next step would be to develop multi-site models that incorporate system metadata such as azimuth and tilt as features. Such models could eliminate the need for site-specific calibration, enabling radiation estimation wherever PV power data are available. A pragmatic intermediate step might be to develop a model applicable across the Netherlands before expanding to larger regions. Further promising directions include exploring more advanced architectures, such as deep neural networks or hybrid models, and extending the framework towards short-term solar radiation forecasting.

6 Conclusions

This thesis set out to answer the research question: *"How accurately can global solar radiation be estimated using a machine learning model based on photovoltaic power output data?"* The overarching objective was to assess whether photovoltaic (PV) power data can serve as a reliable proxy for global radiation where direct measurements are unavailable.

The study explored two machine learning model types—Random Forest and Gradient Boosting—across two types of modelling scenarios: *single-location* scenarios, where models were trained and tested on the same site, and *cross-location* scenarios, where a model was trained on one site and validated on another. The former illustrate the potential when local ground-truth measurements are available for calibration, while the latter evaluate the feasibility of developing more generalised models that operate without site-specific pyranometer data.

The results show that tree-based machine learning models, trained on PV power output and supplementary reanalysis data, can estimate global radiation with high accuracy and very low bias. Both Random Forest and Gradient Boosting achieved strong predictive performance, with R^2 values of about 0.97, MAEs of 10–15 [W/m^2], and mean bias errors below 0.6 [W/m^2] on a 15-minute timescale, for the single-location scenarios. Compared to the widely used reanalysis product ERA5, these models achieved lower errors at a much finer temporal resolution. Moreover, the models retained reasonable accuracy when applied across sites, with the cross-location models having R^2 values of around 0.94, MAEs of 15–30 [W/m^2], and absolute mean bias errors ranging 2–9 [W/m^2].

Feature importance analysis highlighted the dominant influence of PV power output and the clear-sky index for photovoltaics (K_{PV})—a normalised ratio of actual to expected clear-sky generation—with time-shifted K_{PV} values further improving spatial generalisation. In contrast, weather reanalysis features contributed little to the models' predictive skill. By comparison, the theoretical top-of-atmosphere radiation emerged as a consistently valuable predictor.

This study contributes to the fields of solar radiation modelling and machine learning for environmental data in several ways. First, it provides empirical evidence that PV power data can be leveraged to estimate global solar radiation at high temporal resolution with minimal bias, offering a scalable and cost-effective alternative to direct pyranometer measurements. Second, it demonstrates that machine learning models can generalise across sites if normalised, transferable features such as K_{PV} are used. Together, these findings highlight the potential of PV-based radiation modelling for applications in solar resource assessment, grid operation planning, and agricultural management.

Despite these promising results, several limitations must be acknowledged. Single-location models, while highly accurate, still require local pyranometer data for training, meaning they cannot entirely replace direct measurements. Cross-location models showed promise for wider applicability without on-site calibration, although their accuracy was lower. Furthermore, the study was restricted to two sites in the Netherlands, which limits generalisability to other climates and system configurations. Finally, the pyranometer used in Delft introduced a measurement uncertainty of approximately ± 23 [W/m^2], which constrained the maximum achievable model accuracy.

Future research should focus on extending the dataset to include a wider range of PV systems with diverse orientations and configurations across more locations and climates. Such data would likely enable the development of generalised models capable of accurately estimating radiation without local pyranometer calibration. This would allow solar radiation to be estimated at any site where PV panels are installed and their power output recorded. Incorporating system metadata (e.g., azimuth, tilt) as model features could further enhance robustness. In addition, exploring more advanced model architectures such as deep neural networks or hybrid approaches may yield further performance improvements. Finally, this modelling framework could be expanded towards short-term solar radiation forecasting, broadening its practical utility.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Anderson, K. S., Hansen, C. W., Holmgren, W. F., Jensen, A. R., Mikofski, M. A., and Driesse, A. (2023). pvlib python: 2023 project update. *Journal of Open Source Software*, 8(92):5994.

Beran, M. (2013). Can photovoltaic solar panels be calibrated to monitor solar radiation? *Weather*, 68(6):157–161.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Budiyanto, M. A., Nawara, R., et al. (2018). Comparative study of the monthly global solar radiation estimation data in jakarta. In *IOP Conference Series: Earth and Environmental Science*, volume 105, page 012111. IOP Publishing.

Buni, M. J., Al-Walie, A. A., and Al-Asadi, K. (2018). Effect of solar radiation on photovoltaic cell. *International Research Journal of Advanced Engineering and Science*, 3(3):47–51.

Carvalho, I. F. and Corrêa, M. (2019). Techniques of Solar Irradiance Estimation from Datasheet Information of Photovoltaic Panels. *IEEE Brazilian Power Electronics Conference and Southern Power Electronics Conference*.

Daniel (2025). Bagging vs Boosting: What is the Difference? `https://datascientest.com/en/bagging-vs-boosting`.

Dobos, A. P. (2014). PVWatts Version 5 Manual. Technical Report NREL/TP-6A20-62641, National Renewable Energy Laboratory (NREL).

Duffie, J. A. and Beckman, W. A. (2013). *Solar engineering of thermal processes*. John Wiley & Sons.

EKO Instruments (2025). Ms-40 pyranometer – specifications. `https://eko-instruments.com/us/product/ms-40-pyranometer/?cn-reloaded=1#specifications`. Accessed: 2025-07-23.

Engerer, N. and Mills, F. (2014). Kpv: A clear-sky index for photovoltaics. *Solar Energy*, 105:679–693.

Engerer, N., Xu, Y., et al. (2015). A simple model for estimating the diffuse fraction of solar irradiance from photovoltaic array power output. In *21st International Congress on Modelling and Simulation (MODSIM2015). Gold Coast, Australia*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Goodarzi, S., Perera, H. N., and Bunn, D. (2019). The impact of renewable energy forecast errors on imbalance volumes and electricity spot prices. *Energy Policy*, 134:110827.

Ineichen, P. (2008). A broadband simplified version of the solis clear sky model. *Solar Energy*, 82(8):758–762.

Ishii, T., Otani, K., and Takashima, T. (2011). Effects of solar spectrum and module temperature on outdoor performance of photovoltaic modules in round-robin measurements in japan. *Progress in Photovoltaics: Research and Applications*, 19(2):141–148.

Joint Committee for Guides in Metrology (2008). Evaluation of measurement data—guide to the expression of uncertainty in measurement. JCGM 100:2008. Accessed: 2025-07-07.

Kaur, A. (2015). *Forecasting for power grids with high solar penetration*. University of California, San Diego.

Kim, D., Witmer, L., Brownson, J. R., and Braun, J. E. (2014). Impact of solar irradiance data on mpc performance of multizone buildings. In *Proceedings of the 5th International High Performance Buildings Conference*. Purdue University.

Kipp & Zonen B.V. (2008). Cm 11 pyranometer / cm 14 albedometer manual. `https://www.kippzonen.com/Download/48/CM-11-Pyranometer-CM-14-Albedometer-Manual?ShowInfo=true`. Accessed: 2025-07-07.

KNMI (2024). Automatische weerstations. `https://www.knmi.nl/kennis-en-datacentrum/uitleg/automatische-weerstations`. Accessed: 2025-07-22.

Li, H., Cao, F., Wang, X., and Ma, W. (2014). A temperature-based model for estimating monthly average daily global solar radiation in china. *The Scientific World Journal*, 2014(1):128754.

Li, R., Wang, D., Wang, W., and Nemani, R. (2023). A geonex-based high-spatiotemporal-resolution product of land surface downward shortwave radiation and photosynthetically active radiation. *Earth System Science Data*, 15(3):1419–1436.

Liu, B. Y. and Jordan, R. C. (1963). The long-term average performance of flat-plate solar-energy collectors: With design data for the us, its outlying possessions and canada. *Solar energy*, 7(2):53–74.

Meflah, A., Chekired, F., Drir, N., and Canale, L. (2024). Accurate method for solar power generation estimation for different pv (photovoltaic panels) technologies. *Resources*, 13(12):166.

Mohammed, M., Hamdoun, H., and Sagheer, A. (2023). Toward sustainable farming: implementing artificial intelligence to predict optimum water and energy requirements for sensor-based micro irrigation systems powered by solar pv. *Agronomy*, 13(4):1081.

Müller, R. W., Dagestad, K.-F., Ineichen, P., Schroedter-Homscheidt, M., Cros, S., Dumortier, D., Kuhlemann, R., Olseth, J., Piernavieja, G., Reise, C., et al. (2004). Rethinking satellite-based solar irradiance modelling: The solis clear-sky module. *Remote sensing of Environment*, 91(2):160–174.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21.

Nespoli, L. and Medici, V. (2017). An unsupervised method for estimating the global horizontal irradiance from photovoltaic power measurements. *Solar Energy*, 158:701–710.

Oke, T. R. (2002). *Boundary layer climates*. Routledge.

Omoyele, O., Hoffmann, M., Koivisto, M., Larraneta, M., Weinand, J. M., Linßen, J., and Stolten, D. (2024). Increasing the resolution of solar and wind time series for energy system modeling: A review. *Renewable and Sustainable Energy Reviews*, 189:113792.

Open-Meteo (2024). Open-meteo historical weather api. `https://open-meteo.com/`. Accessed: 2025-05-14.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Rajagukguk, R. A. and Lee, H. (2023). Enhancing the performance of solar radiation decomposition models using deep learning. *Journal of the Korean Solar Energy Society*, 43(3):73–86.

Razavi, S. and Gupta, H. V. (2015). What do we mean by sensitivity analysis? the need for comprehensive characterization of "global" sensitivity in earth and environmental systems models. *Water Resources Research*, 51(5):3070–3092.

Sandia National Laboratories (2025a). Global horizontal irradiance. `https://pvpmc.sandia.gov/modeling-guide/1-weather-design-inputs/irradiance-insolation/global-horizontal-irradiance`. [Online; accessed 1. Oct. 2025].

Sandia National Laboratories (2025b). Isotropic Sky Diffuse Model. `https://pvpmc.sandia.gov/modeling-guide/1-weather-design-inputs/plane-of-array-poa-irradiance/calculating-poa-irradiance/poa-sky-diffuse/isotropic-sky-diffuse-model`. [Online; accessed 5. Jun. 2025].

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American journal of applied mathematics and statistics*, 8(2):39–42.

Sinovoltaics (2011). Standard test conditions (stc): definition and problems. `https://sinovoltaics.com/learning-center/quality/standard-test-conditions-stc-definition-and-problems/`. Accessed: 2025-07-21.

The Green Village (2025). The Green Village data platform. `https://www.thegreenvillage.org/dataplatform`. Accessed: 2025-05-04.

Urraca, R., Huld, T., Gracia-Amillo, A., Martinez-de Pison, F. J., Kaspar, F., and Sanz-Garcia, A. (2018). Evaluation of global horizontal irradiance estimates from era5 and cosmo-rea6 reanalyses using ground and satellite-based data. *Solar Energy*, 164:339–354.

Wageningen University & Research (2025). Veenkampen observations. `https://maq-observations.nl/data-downloads`. Accessed: 16 May 2025.

I Data Pre Processing

Here the preprocessing of the data for both sites is provided. The PV production and global radiation data can be seen plotted over time in Figures 20, 21, 22 and 23.

I.1 Delft Data

For the Delft data the following preprocessing steps were taken. The PV power data was set to zero during nighttime, to replace missing values during the night. The data was received in Dutch local time and was then transformed to UTC.

The pyranometer data from the Green Village is measured on a 1-minute resolution. This data was then averaged over 15 minutes in order to match the time resolution of the other data. Then all measurements above a theoretical limit of 1362 [W/m^2] were discarded. Furthermore, three continuous days of data were also removed as the pyranometer dataset had values of 0 [W/m^2] for the entire three days, whilst the PV data showed notable power production. It is assumed that during this time maintenance was being performed on the pyranometer.

PV and pyranometer data were then combined to one DataFrame and all time stamps where either measurement had no data were dropped. After this there were a total of 63,016 timestamps with valid measurements for the Delft location.

I.2 Wageningen Data

For the Wageningen data the following preprocessing steps were taken. The PV power data had no missing data at night as opposed to the Delft data so the data only had to be transformed to UTC.

The pyranometer data from the Veenkampen measurement site was directly acquired at the desired 15-minute resolution and had no noticeable outliers.

Again PV and pyranometer data were combined to one DataFrame and all time stamps where either measurement had no data were dropped. After this there were a total of 208,000 timestamps with valid measurements for the Wageningen location.

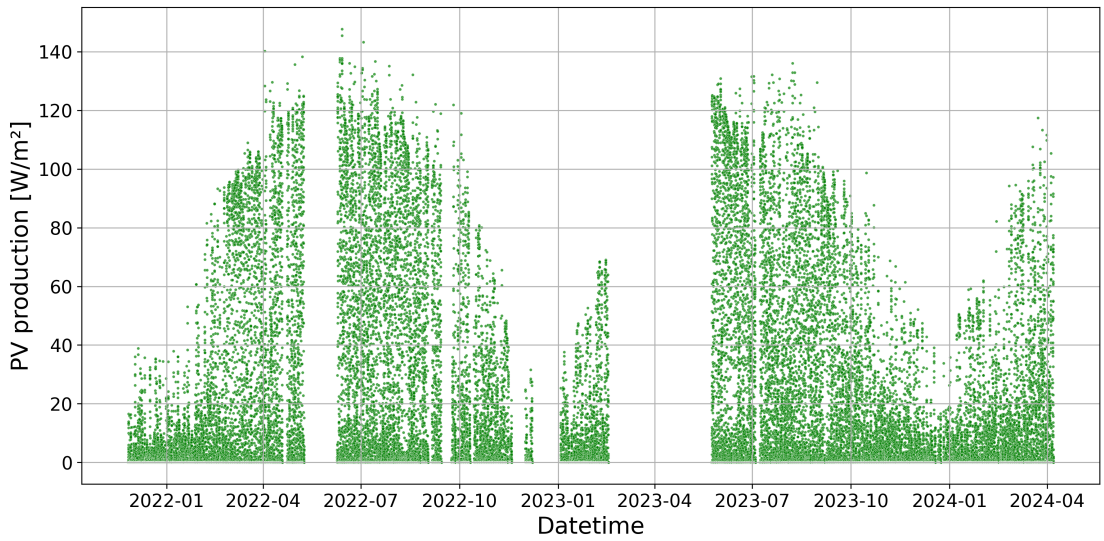


Figure 20: PV production data for the Delft location.

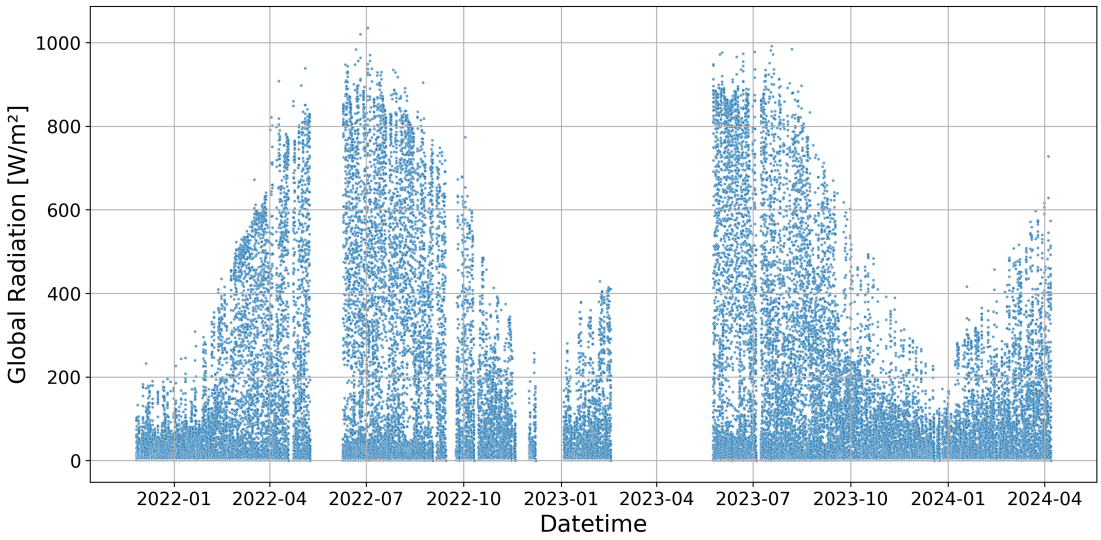


Figure 21: Global radiation data for the Delft location.

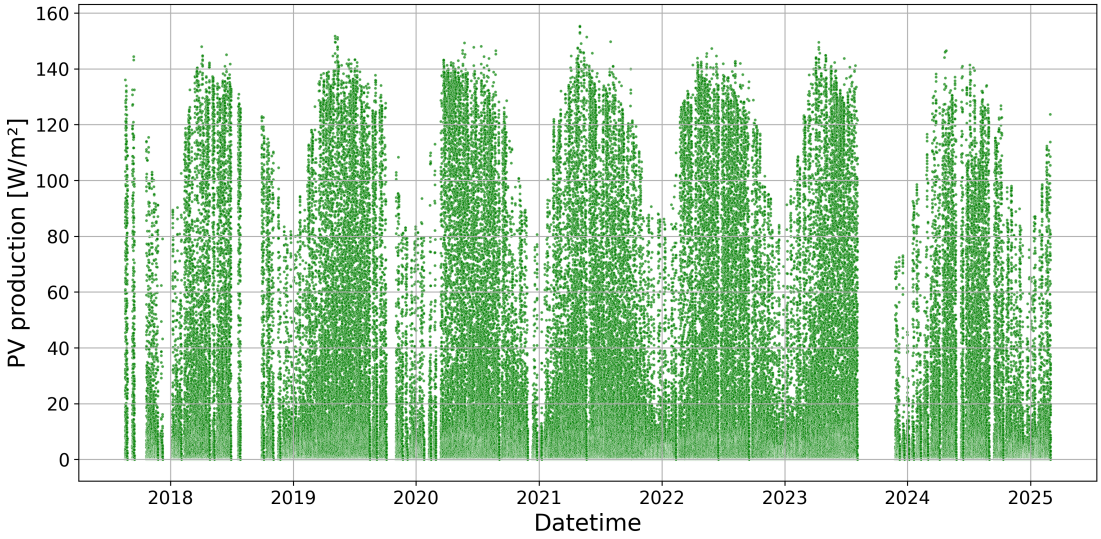


Figure 22: PV production data for the Wageningen location.

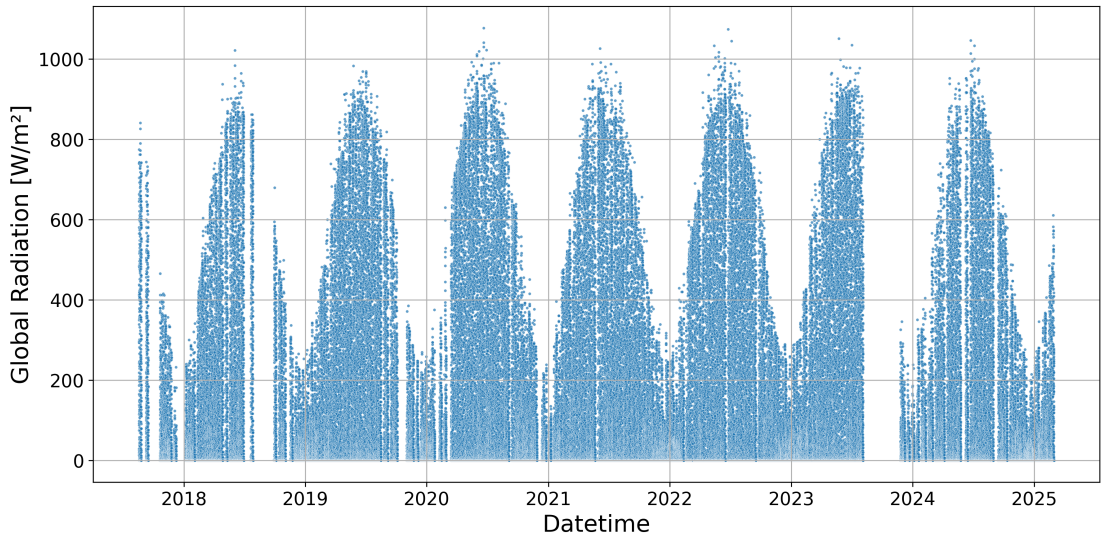


Figure 23: Global radiation data for the Wageningen location.

II Pyranometer details

II.1 Pyranometer Delft

Global solar radiation was measured using an EKO MS-40 pyranometer, classified as a Class C instrument under ISO 9060:2018. This classification denotes suitability for general meteorological applications, offering moderate accuracy relative to higher-grade instruments. The MS-40 features a response time of less than 18 seconds and a non-linearity error of ±1% across the range of 100 to 1000 [W/m^2]. Directional response errors (cosine errors) can reach up to ±20 [W/m^2] at high solar zenith angles near 80°, potentially leading to slight underestimation of irradiance during early morning and late afternoon hours. The sensor exhibits a temperature response of ±3%, and its combined zero offset can be as high as ±17 [W/m^2], which may cause small non-zero readings under low-light or nighttime conditions. The sensor has a non-stability of ±1.5% per year (EKO Instruments, 2025).

Representative Error Calculation

Based on manufacturer specifications and typical field conditions, the expected measurement uncertainty of the MS-40 pyranometer can be estimated by aggregating several individual sources of error. The most relevant specifications are:

- **Zero offset (total)**: up to ±17 [W/m^2]
- **Non-linearity**: ±1% over the range 100–1000 [W/m^2]
- **Temperature response**: ±3%
- **Directional (cosine) response**: up to ±20 [W/m^2] at high zenith angles
- **Non-stability (drift)**: ±1.5% per year

Assuming moderate environmental conditions and regular maintenance (e.g., recalibration every two years), a conservative estimate of the typical mean error can be calculated. Since these error sources are largely independent, they are combined using the root-sum-square (RSS) method (Joint Committee for Guides in Metrology, 2008):

$$\text{Mean error} \approx \sqrt{17^2 + (0.01 \cdot 500)^2 + (0.015 \cdot 500)^2 + 10^2 + (0.015 \cdot 500)^2}$$

Here, 500 W/m^2 is chosen as a representative mid-range irradiance level. Half of the temperature and directional response errors are included, reflecting typical diurnal conditions. The long-term drift is averaged over a two-year recalibration interval. Substituting the values yields:

$$\text{Mean error} \approx \sqrt{17^2 + 5^2 + 7.5^2 + 10^2 + 7.5^2} = \sqrt{526.5} \approx 22.9 \text{ W/m}^2$$

Accordingly, under standard operational conditions, the expected mean measurement error of the MS-40 pyranometer is estimated at approximately $\pm 23\,[\text{W/m}^2]$. This value is considered conservative, as it includes the full non-linearity error and significant portions of the temperature and cosine response uncertainties.

II.2 Pyranometer Wageningen

Global solar radiation was measured at a meteorological data collection site called 'Veenkampen', approximately 3.5 kilometres west-northwest of Wageningen, using a Kipp & Zonen CM11 pyranometer, classified as a *Secondary Standard* instrument under ISO 9060:2018. This is the highest classification defined by the standard, indicating that the sensor meets strict requirements on accuracy and stability, necessary for for high-precision scientific and meteorological applications. The CM11 features a response time of less than 15 seconds and a non-linearity of less than $\pm 0.6\%$ over the range up to 1000 [W/m^2]. Directional (cosine) response errors are below ± 10 [W/m^2] for direct beam irradiance at high zenith angles, contributing to improved accuracy during early morning and late afternoon measurements. The sensor's temperature dependence remains under $\pm 1\%$ between -10°C and 40°C, and the zero offset due to thermal radiation is specified to be below ± 7 [W/m^2]. With a stated non-stability of less than $\pm 0.5\%$ per year, the CM11 is well-suited for long-term monitoring applications with minimal calibration drift (Kipp & Zonen B.V., 2008).

Representative Error Calculation

To estimate the expected measurement uncertainty of the Kipp & Zonen CM11 pyranometer, key error sources from the manufacturer specifications are aggregated under typical operational conditions. The most relevant parameters include:

- **Zero offset (thermal response)**: up to $\pm 7\,[\text{W/m}^2]$
- **Non-linearity**: $< \pm 0.6\%$ for irradiance $< 1000\,[\text{W/m}^2]$
- **Temperature dependence**: $< \pm 1\%$ in the range -10°C to 40°C
- **Directional (cosine) error**: up to $\pm 10\,[\text{W/m}^2]$ at high zenith angles
- **Non-stability (drift)**: $< \pm 0.5\%$ per year

Assuming moderate ambient temperatures, a zenith angle distribution skewed toward mid-day (i.e., smaller cosine errors), and a two-year calibration interval, a representative mean error can be estimated using the root-sum-square (RSS) method (Joint Committee for Guides in Metrology, 2008). The following expression is used, again for a typical irradiance of 500 W/m^2:

$$\text{Mean error} \approx \sqrt{7^2 + (0.006 \cdot 500)^2 + (0.005 \cdot 500)^2 + 10^2 + (0.005 \cdot 500)^2}$$

This includes the full zero offset, full non-linearity, and average drift error over two years, while including half of the temperature and directional response contributions. Substituting values yields:

$$\text{Mean error} \approx \sqrt{7^2 + 3^2 + 2.5^2 + 10^2 + 2.5^2} = \sqrt{170.5} \approx 13.1 \text{ W/m}^2$$

Under standard operating conditions, the expected mean measurement error of the CM11 pyranometer is therefore estimated at approximately $\pm 13\,[\mathrm{W/m^2}]$. This reflects a higher accuracy compared to Class C sensors, aligning with the CM11's first-class classification.

III Additional Results and Figures

Figures 24 and 25 show the feature importance scores for the models using K_{PV} derived from the PVWatts model.

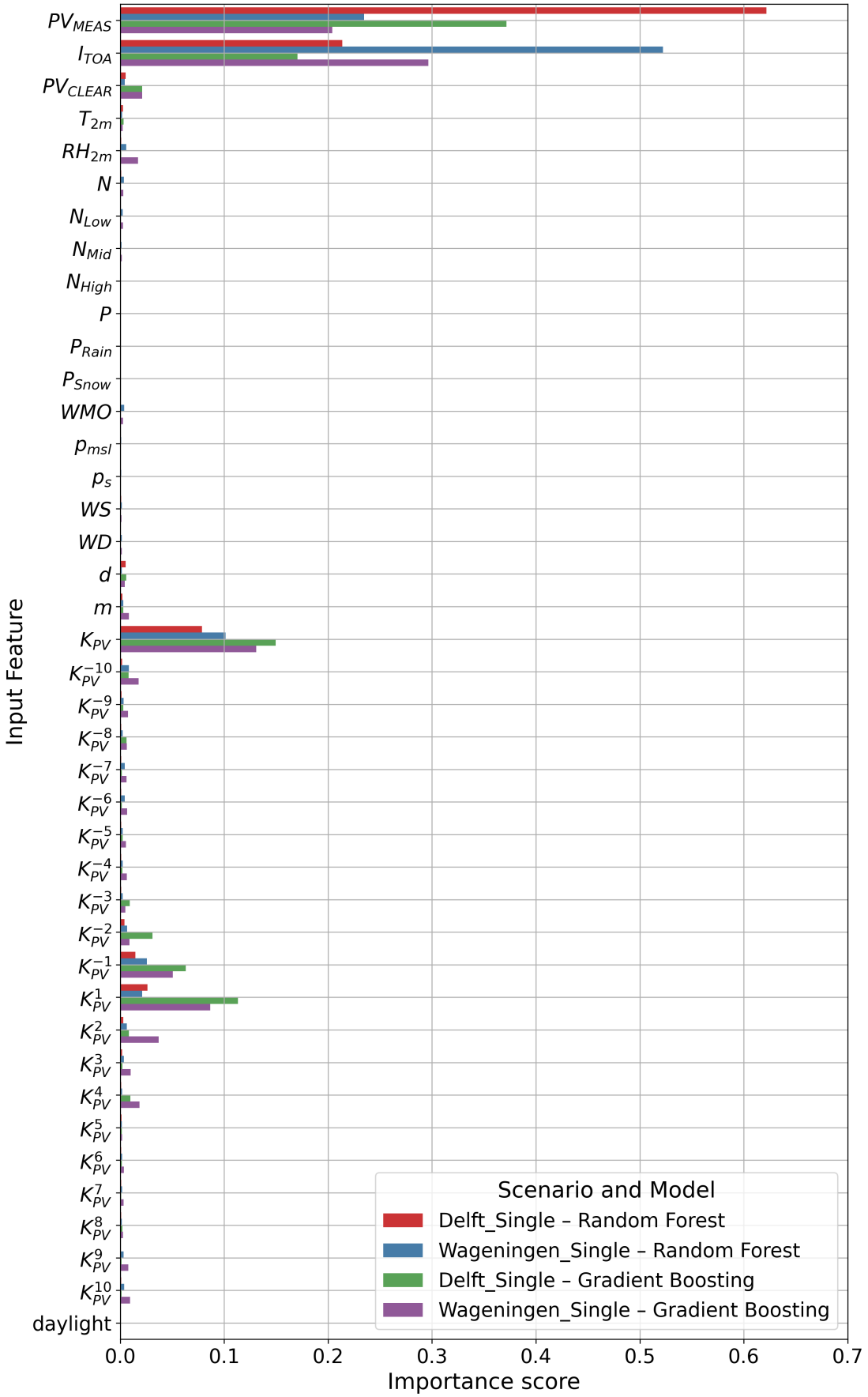


Figure 24: Feature importance scores for the single-location models, using PVWatts.

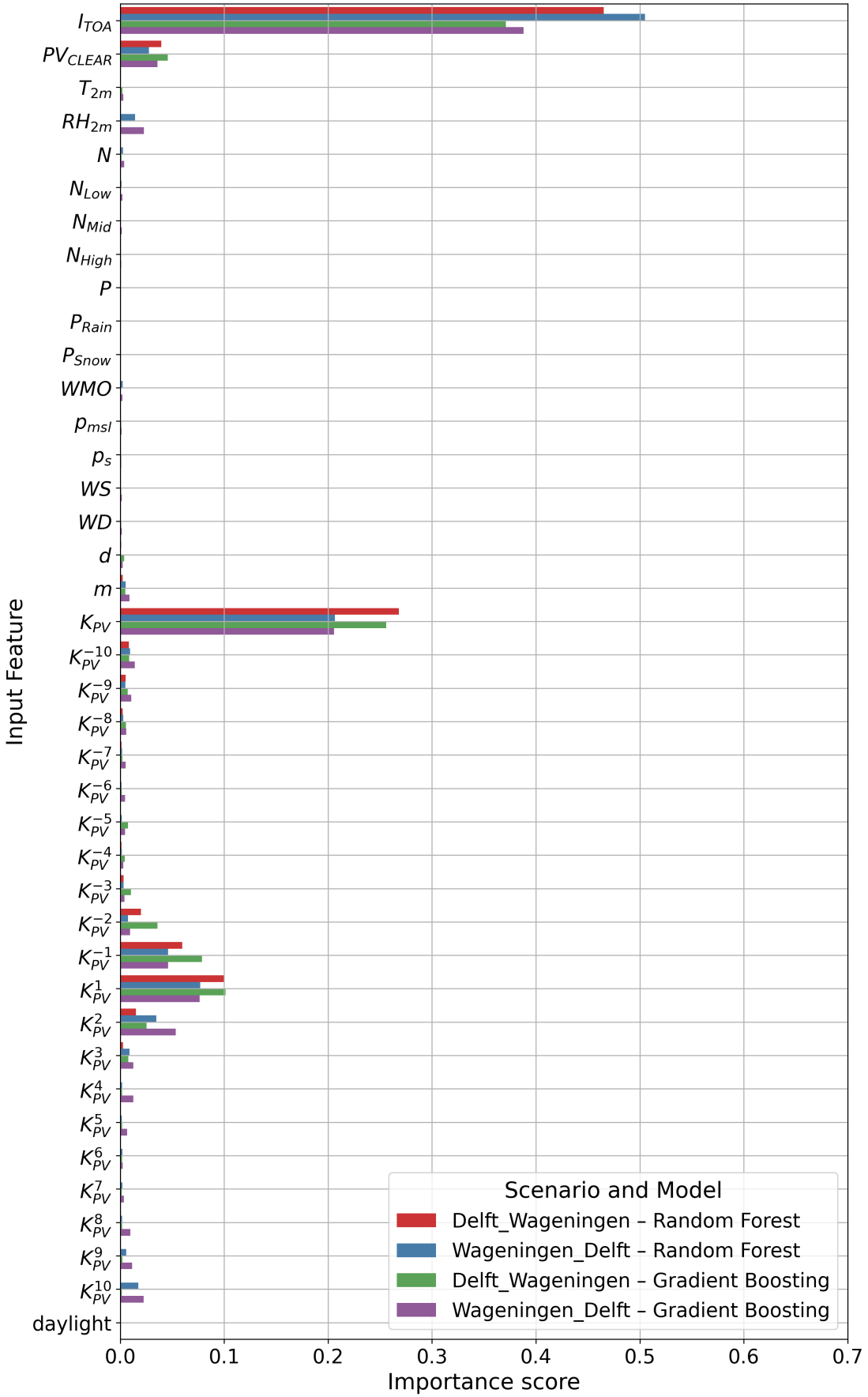


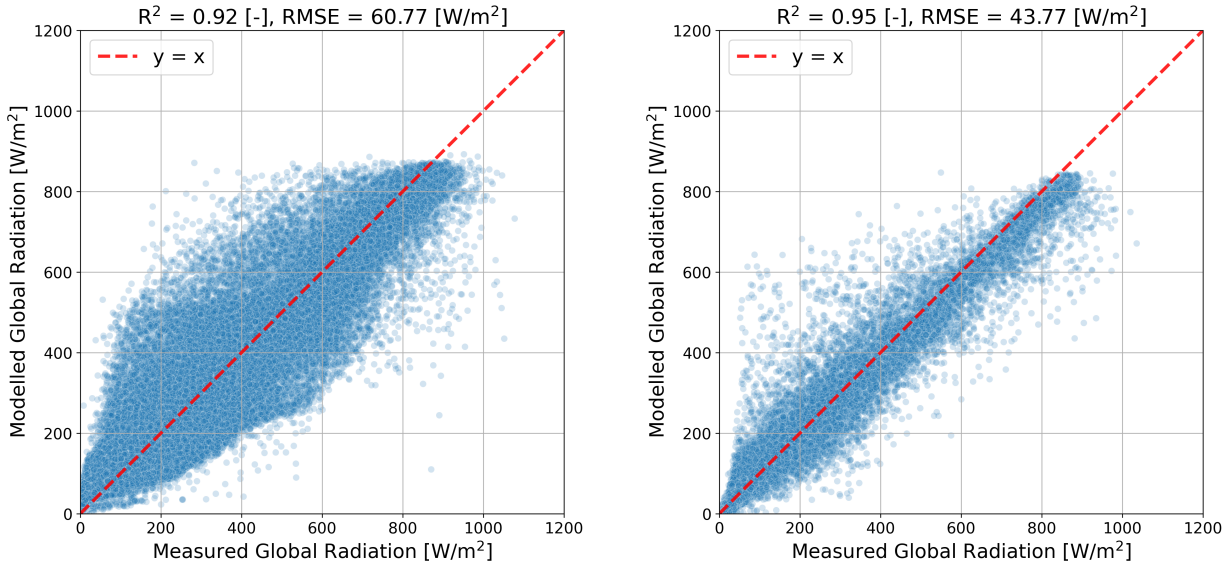
Figure 25: Feature importance scores for the cross-location models, using PVWatts.

IV Transferability

In the main results, the cross-location models were optimised using data from the target site. While this approach helps achieve good performance, it does not fully reflect a situation in which no global radiation data are available from the target location. The use of target-site data during optimisation could therefore overestimate how well the models generalise in a true "zero-calibration" scenario.

To provide a more rigorous assessment of transferability, the Random Forest models based on the relative-production K_{PV} were retrained using only the hyperparameters obtained from the single-location optimisation procedure. These models were then applied directly to the test data of the opposite site, without involving any measurements from that site in either training or hyperparameter selection. This setup simulates a deployment case in which no radiation measurements are available for calibration at the target location.

The resulting performance is shown in Figure 26.



(a) Scenario: `Delft_Wageningen`

(b) Scenario: `Wageningen_Delft`

Figure 26: Random Forest regression models using K_{PV} derived from relative production, with hyperparameters taken from the single-location optimisation. Results are shown for both cross-location scenarios.

The results indicate that model accuracy changes very little compared with models whose hyperparameters were tuned specifically for cross-location performance. Remarkably, the `Wageningen_Delft` scenario even achieves a slightly lower RMSE when using hyperparameters from the single-location optimisation. This suggests that the Random Forest approach is relatively robust to hyperparameter selection and that good transferability can be achieved without access to radiation data from the target site.

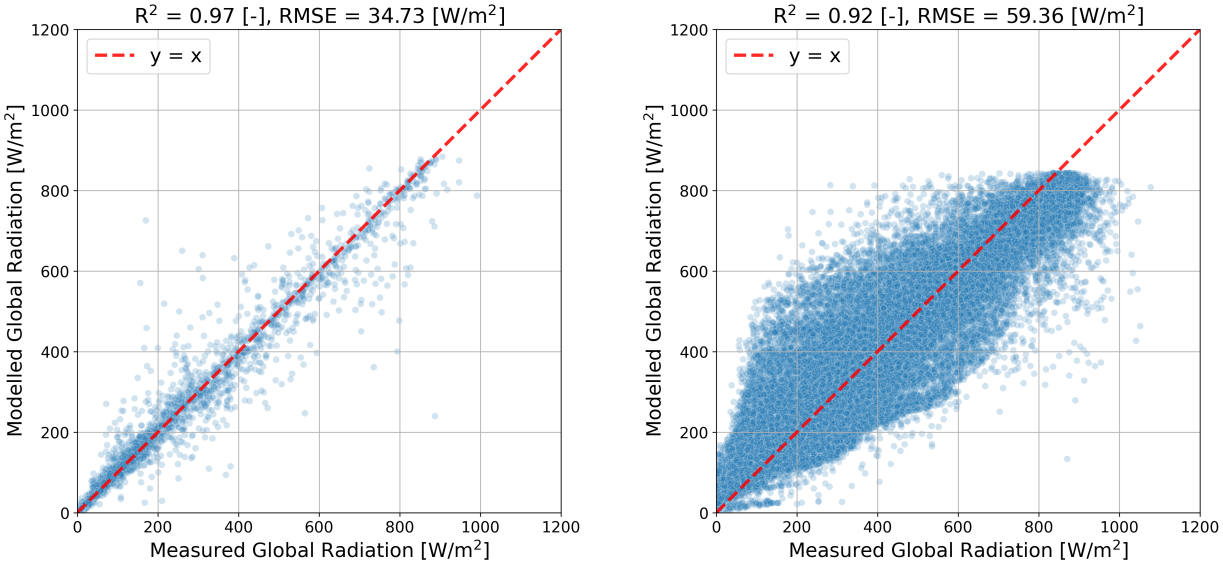
V Real-Time Modelling

The models presented in the main results make use of input features from reanalysis products, as well as the lead (time-shifted) form of K_{PV}. While effective for retrospective analysis, this setup prevents direct application in real-time scenarios, since reanalysis data are typically released with delays ranging from several weeks to months.

To assess the feasibility of real-time radiation estimation, the Random Forest models based on K_{PV} derived from the relative production method were re-evaluated using only features that are

available instantaneously. Specifically, reanalysis variables and the time-lead version of K_{PV} were excluded, while all hyperparameters were kept identical to those used in the main experiments. This procedure was applied to the `Delft_Single` and `Delft_Wageningen` scenarios.

The resulting performance is shown in Figure 27.



(a) Scenario: `Delft_Single`

(b) Scenario: `Delft_Wageningen`

Figure 27: Random Forest regression models using K_{PV} derived from relative production, with only input features available in real time.

The results show that removing reanalysis variables and the lead form of K_{PV} has only a minor effect on model accuracy. This suggests that the proposed approach can be applied effectively for real-time estimation of global solar radiation, without substantial loss of performance.