energies

MDPI

Article

Advanced PV Performance Modelling Based on Different Levels of Irradiance Data Accuracy

Julián Ascencio-Vásquez [1,*], Jakob Bevc [2], Kristjan Reba [2], Kristijan Brecl [1], Marko Jankovec [1] and Marko Topič [1]

[1] Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia; kristijan.brecl@fe.uni-lj.si (K.B.); marko.jankovec@fe.uni-lj.si (M.J.); marko.topic@fe.uni-lj.si (M.T.)

[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia; jakob.bevc@gmail.com (J.B.); kristjan.reba96@gmail.com (K.R.)

* Correspondence: julian.ascencio@fe.uni-lj.si

Received: 9 March 2020; Accepted: 21 April 2020; Published: 1 May 2020

Abstract: In photovoltaic (PV) systems, energy yield is one of the essential pieces of information to the stakeholders (grid operators, maintenance operators, financial units, etc.). The amount of energy produced by a photovoltaic system in a specific time period depends on the weather conditions, including snow and dust, the actual PV modules' and inverters' efficiency and balance-of-system losses. The energy yield can be estimated by using empirical models with accurate input data. However, most of the PV systems do not include on-site high-class measurement devices for irradiance and other weather conditions. For this reason, the use of reanalysis-based or satellite-based data is currently of significant interest in the PV community and combining the data with decomposition and transposition irradiance models, the actual Plane-of-Array operating conditions can be determined. In this paper, we are proposing an efficient and accurate approach for PV output energy modelling by combining a new data filtering procedure and fast machine learning algorithm Light Gradient Boosting Machine (LightGBM). The applicability of the procedure is presented on three levels of irradiance data accuracy (low, medium, and high) depending on the source or modelling used. A new filtering algorithm is proposed to exclude erroneous data due to system failures or unreal weather conditions (i.e., shading, partial snow coverage, reflections, soiling deposition, etc.). The cleaned data is then used to train three empirical models and three machine learning approaches, where we emphasize the advantages of the LightGBM. The experiments are carried out on a 17 kW roof-top PV system installed in Ljubljana, Slovenia, in a temperate climate zone.

Keywords: PV system; PV performance modelling; data filtering; machine learning; lightGBM

1. Introduction

Projections of the photovoltaic (PV) installed capacity show a strong and fast expansion of the PV deployment in under-developing countries in South America, Africa, and Asia due to the cost-competitiveness that PV systems achieved [1]. This considerable increase in solar electricity could alter the regular operation of electrical grids if energy storage is not considered. Since the solar resource is not constant along the day and can have rapid fluctuations due to moving clouds and other local effects, the forecasting of electricity becomes a non-trivial problem. Therefore, the modelling of the energy output delivered by PV systems is an essential task for grid operators to plan, run, and preserve the stability of the electrical system.

To develop trustworthy PV performance and PV power models, accurate electrical and weather data are needed as input. For a typical PV system, the metadata (mounting configuration, technology, etc.) is known, as well as the time-series of output power since the commissioning date. The weather

data is rarely measured at the same location as the PV modules. Typically, the ground-based local weather data (using pyranometers and temperature sensors) are not available, and the use of global reanalysis-based or satellite-based data is the only option [2–5]. In this regard, the "ERA5" climate reanalysis dataset developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) is reported as a reasonable data source for studies in PV [6–8]. However, the need for improvement in overcast conditions or high latitudes has also been identified [9,10]. Additionally, the decomposition and transposition irradiance models usually need to be applied to translate the horizontal irradiance to the Plane-of-Array (PoA) of the system [11].

The fusion of the electrical data and the weather data can reflect typical data issues such as gaps, mismatches, timeshifts, or outliers. For this reason, a data filtering process is usually applied to reduce the uncertainty and offsets already added by the input data. Once the data is clean, either simple empirical or advanced machine learning approaches can be used to model the output of any PV system [12–15].

Regarding empirical models, by considering the global PoA irradiance (G_{PoA}) and the PV module temperature measured on-site, high accuracy can be reached. However, it is proved that machine learning approaches can provide even better results, although the time setting and processing in some cases can be extensive. The most common models used for PV forecasting are artificial neural networks (ANN) or support vector machines (SVM) [16]. Recent publications highlight a novel algorithm so-called "Light Gradient Boosting Machine" or "LightGBM" [17]. This algorithm has been tested successfully in the finance industry [18–20], the chemistry industry [21,22], and the healthcare sector [23,24]. In the PV industry, the first results were published in [25], highlighting the accuracy and fast speed to estimate the energy output of a PV system. Hereby, we extend and validate the use of several energy yield models for different levels of irradiance data accuracy.

First, different levels of irradiance data accuracy are defined, where irradiance is measured on-site or extracted from the ERA5 climate reanalysis dataset and modelled to the Plane-of-Array (PoA) through decomposition and transposition models. Then, the general energy yield modelling methodology is presented, as well as the filtering algorithm applied. The empirical models and machine learning approaches are described, and finally, the results from the filtering algorithm and each energy yield model are presented and discussed.

2. Data accuracy and Models

2.1. Definition of Data Accuracy Levels

A PV performance model can be designed using the measured electrical power and the weather at specific locations if operational data exist. The output power is typically measured with high accuracy (<1%), as well as the ambient temperature. However, the global PoA irradiance (G_{PoA}) accuracy varies depending on different cases (see Figure 1) and is defined as follows:

- Low accuracy: the solar global horizontal irradiance (*GHI*) is extracted from a satellite-based or reanalysis-based dataset without post-processing (uncertainty "μ_{01}") and estimated at the PoA using decomposition (uncertainty "μ_1") and transposition (uncertainty "μ_2") models.
- Medium accuracy: the *GHI* is measured using a pyranometer (uncertainty "μ_{02}") and G_{PoA} estimated using decomposition and transposition models (uncertainties "μ_1" and "μ_2").
- High accuracy: the G_{PoA} is measured using a pyranometer in the plane of array (uncertainty "μ_{02}").

As reference values, μ_{01} in terms of average normalized mean bias error (nMBE) ranges from 3.47% to 5.14% [4], μ_{02} can be close to 1.5% [26]. Through modelling, the μ_1 and μ_2 could add errors below 5% (nMBE with Erbs model) and below 1.5% (nMBE with Hay/Davies model), respectively [11].

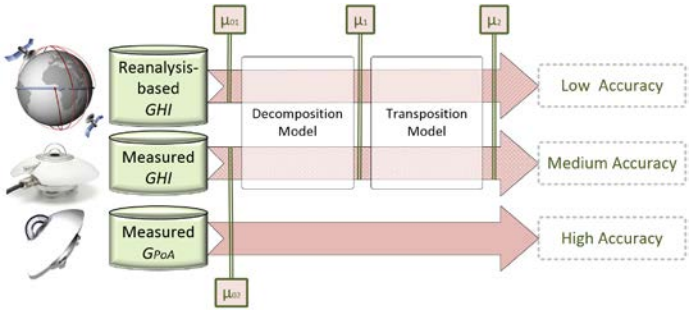


Figure 1. Flowchart for the definition of different levels of irradiance data accuracy at Plane-of-Array (PoA) from measured and modelled data sources.

2.2. Definition of Empirical Models

Empirical models are defined by mathematical equations combining the input variables (i.e., weather variables) and numerical coefficients. The empirical modelling is a straightforward and fast method to predict the PV energy yield. Usually, least-square approximations are performed to extract the coefficients from the recorded operational data. Hereby, we define three empirical models based on the structure of well-performing models in the PV community.

- Empirical #1—Modified PVGIS model: An empirical model combining logarithmic regressions of normalized irradiance and PV module temperature with six empirical coefficients has been reported to provide excellent results over large geographical regions [27,28] (e.g., used in the PVGIS online tool [29]). In our simulations, we are using the measured ambient temperature (T_{amb}) and measured/modelled plane-of-array irradiance without normalization as an input. Equation (1) shows the mathematical expression.

$$P(G,T) = 1 + k_1 \ln(G) + C_2 \ln(G)^2 + T\left(k_3 + k_4 \ln(G) + k_5 \ln(G)^2\right) + k_6 T^2 \quad (1)$$

- Empirical #2—SRCL2014 model: developed by S. Ransome et al. [30], combines first and second-order regressions with logarithmical functions and four empirical coefficients to estimate the output power from the irradiance. The mathematical expression is presented in Equation (2).

$$P(G) = G(k_1 \ln(G) + k_2)\left(1 - (1 - k_3)G^2\right)k_4 \quad (2)$$

- Empirical #3—Polynomial model: a polynomial function of the irradiance can achieve a simple mathematical approximation of the output power. We are using a 4th order polynomial function.

$$P(G) = G\left(k_1 + k_2 G + k_3 G^2 + k_4 G^3 + k_5 G^4\right) \quad (3)$$

2.3. Machine Learning Approaches

Advanced mathematical models, so-called machine learning (ML) approaches, have shown the improvement of accuracy in many different fields of science, including PV. In this work, we compare three approaches and validate them using the hold-out method.

- Artificial neural networks (ANN): are machine learning models inspired by biological neural networks. They consist of mathematical units called neurons and connections between them called weights. For ANN to learn a task, weights have to be optimized. This is usually done through gradient-based optimization techniques [31].
- Support vector machine (SVM): a supervised learning model that can be used for regression as well as classification tasks. SVM separates the data linearly, and by using a kernel trick, it transforms

the data into higher dimensional feature space where a linear separation with a hyperplane is performed [32].

- Gradient boosting machines and LightGBM: gradient boosting decision tree (GBDT) [33,34] is a widely used machine learning algorithm, which achieves state-of-the-art results in many tasks and offers interpretability. The GBDT is an ensemble model in which predictors are trained sequentially. In each iteration, a weak prediction model, such as one level tree, fits the residual errors of the previous model. The main computational cost of such algorithm originates from the learning of decision trees, where the bottleneck is to find optimal split points with the highest information gain. LightGBM is a novel GBDT model that involves two novel techniques to deal with the problems of finding the optimal split point: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). The GOSS is applied to reduce the number of data instances, and EFB is used to reduce the feature space. Applying the LightGBM, the time processing can be reduced considerably in comparison to the ANN and SVM approaches.

2.4. Characteristics of the Photovoltaic (PV) System Used

Models are trained and tested using the operational data recorded on a 17 kW PV system installed on the rooftop of the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. This installation site is located in the temperate climate with medium irradiation regarding the Köppen-Geiger-Photovoltaic (KGPV) climate classification [5] and the climate stress based on temperature, humidity and irradiance equals to −0.355%/a, which is lower than the median based on the degradation map presented in [7].

The solar irradiance at the PoA and also at horizontal position are measured using pyranometers Kipp and Zonnen CMP-21 and CMP-6, respectively, as shown in Figure 2a. The PV system comprises 74 modules Bisol BMU233 and one additional reference module, as presented in Figure 2b,c. The data evaluated ranges from May 2014 to December 2019.

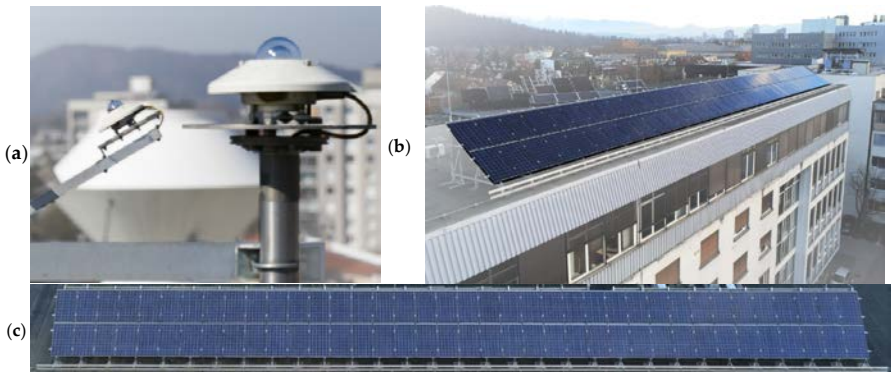


Figure 2. Images of the 17 kW Photovoltaic (PV) system installed on the rooftop of the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. (**a**) Pyranometers installed at the Plane-of-Array and horizontal. (**b**) PV system from a perspective showing the building where installed. (**c**) The PV system from a perpendicular angle.

3. Methodology of PV Energy Yield Modelling

The flowchart in Figure 3 presents the energy yield modelling steps. The raw datasets with different levels of data accuracy are divided into "training set" (e.g., 70% of the data) and "test set" (e.g., 30% of the data). Before training the models, and since recorded data might contain outliers, gaps of missing data, and corrupted or incoherent values, the filtering is applied. The filtered training dataset is then used to create the PV energy yield model, either with empirical or machine learning approaches. The models predict the energy yield on the Test set. Finally, for the validation step, the unrealistic measured points together with their modelled points are filtered.

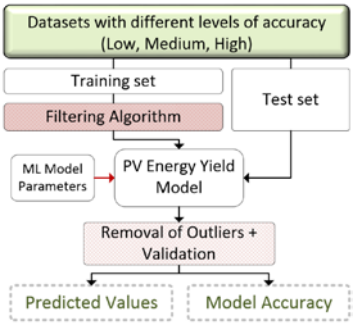


Figure 3. Flowchart for the data processing, including filtering algorithm stage, training, testing, and validation of empirical models and machine learning approaches.

3.1. Data Filtering Process

The automatic filtering algorithm is presented in Table 1, which is inspired by the idea proposed in [35]. The power output is correlated to the plane-of-array irradiance, where a non-linear relationship is expected between both variables. The data filtering starts by clustering the power data depending on the irradiance. The Gaussian distribution is calculated per each cluster, and a lower and upper "percentile curves" are created by connecting the values of each group (i.e., the 5th and 95th percentiles). A third middle curve is identified by defining the 50th percentile values of each cluster and approximating them by a polynomial function.

In the second step, we applied an additional smoothing step of the lower and upper percentile curves. For both curves, the Gaussian distribution of the distances to the 50th percentile curve is calculated, and the selected percentile limit is used to define the side curves (25th for the lower and 75th for the upper curve in our case). All points outside the lower and upper percentile curves are removed. This approach turned out to be very effective to remove the unrealistic data values.

3.2. PV Energy Yield Modelling Procedure

In addition to the filtering procedure, the evaluation of the long-term measured PV energy yield is recommended to identify probable degradation of the PV system. We use a statistical model, so-called Holt-Winters (HW) seasonal exponential smoothing [36], applied to the monthly Performance Ratio (*PR*), defined as the ratio of the normalized total energy yield and incident solar irradiance.

The filtered dataset is split into the training set (e.g., 70% of the data) and the test set (e.g., 30% of the data). A standard normalization of each value is applied by subtracting the average of the training set and dividing it by its standard deviation. The feature sets, or input variables, for each ML approach, used defined as (1) Irradiance, (2) Irradiance and ambient temperature, and (3) Irradiance, ambient temperature, and the sun position angles (azimuth and zenith).

The numerical coefficients of each empirical model are fitted by applying the least square linear regression method. The machine learning approaches are optimized by manual tuning of the hyper parameters using python libraries (LightGBM [17] and scikit-learn [37]).

3.3. Uncertainty Indicators

The overall accuracy of the PV energy yield models can be evaluated using two indicators: the standard error and normalized root-mean-square-error (nRMSE) (see Equations (4)–(6)). For ML approaches, the processing time is also measured.

We use the correlation of the temperature and irradiation with the standard error to identify the easiest and hardest periods of prediction in terms of climate conditions.

$$\text{Error } [\%] = \frac{\left| y_{model} - y_{real} \right|}{y_{real}} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ireal} - y_{imodel})} \quad (5)$$

$$\mathrm{nRMSE} = \frac{\mathrm{RMSE}}{P_{\mathrm{nominal}}} \quad (6)$$

Table 1. Algorithm for data filtering (Clustering).

Definitions:
class_size: Size of each class
P_{Gi}: Power output values per cluster
L_{xth}: Lower percentile used as threshold
U_{xth}: Upper percentile used as threshold
$P_{Gi\text{-}Lxth}$: L_{xth} percentile of P_{Gi}
$P_{Gi\text{-}50th}$: 50th percentile of P_{Gi}
$P_{Gi\text{-}Uxth}$: U_{xth} percentile of P_{Gi}
$P_{Gi\text{-}50th}\{G_i\}$: 50th percentile of output power per class G_i
Functions:
$\mathrm{polyfit}(x) = a{\cdot}x + b{\cdot}x^2 + c{\cdot}x^3 + d{\cdot}x^4 + e$
$f_{Gaussian}$: Gaussian distribution
• Clustering by Irradiance
for G_i range from class_size to 1300 in steps of class_size
$G_i = \{G_i\_\mathrm{class\_size}, G_i\}$
$P_{Gi} = P_{output}\{G_i\}$
• Calculate thresholds per class (G_i)
$P_{Gi\text{-}Lxth} = f_{Gaussian}(L_{xth})$
$P_{Gi\text{-}50th} = f_{Gaussian}(50th)$
$P_{Gi\text{-}Uxth} = f_{Gaussian}(U_{xth})$
• Filtering the 50th percentile curve
$P_{shift}\{G_i\} = \lvert P_{Gi\text{-}50th}\{G_i\} - P_{Gi\text{-}50th}\{G_{i\text{-}2}\}\rvert^2 + \lvert P_{Gi\text{-}50th}\{G_i\} - P_{Gi\text{-}50th}\{G_{i-1}\}\rvert +$ $\lvert P_{Gi\text{-}50th}\{G_i\} - P_{Gi\text{-}50th}\{G_{i+1}\}\rvert + \lvert P_{Gi\text{-}50th}\{G_i\} - P_{Gi\text{-}50th}\{G_{i+2}\}\rvert^2$
$P_{Gi\text{-}50th\text{-}Filter}\{G_i\} = P_{Gi\text{-}50th}\{G_i\} > P_{shift}\{90th\ percentile\}$
• Polynomial Fitting the 50th percentile curve
$P_{fit\text{-}50th}\{G_i\} = \mathrm{polyfit}\{P_{Gi\text{-}50th\text{-}Filter}\}$
• Filtering the L_{xth} and U_{xth} percentile curves
$\mathrm{Diff}_{Lxth}\{G_i\} = P_{fit\text{-}50th}\{G_i\} - P_{Gi\text{-}Uxth}$
$\mathrm{Diff}_{Uxth}\{G_i\} = P_{fit\text{-}50th}\{G_i\} - P_{Gi\text{-}Uxth}$
$P_{Gi\text{-}Lxth\text{-}Filter}\{G_i\} = P_{Gi\text{-}Lxth} < \mathrm{Diff}_{Lxth}\{75th\ percentile\}$
$P_{Gi\text{-}Uxth\text{-}Filter}\{G_i\} = P_{Gi\text{-}Uxth} > \mathrm{Diff}_{Uxth}\{25th\ percentile\}$
• Fitting the L_{xth} and U_{xth} percentile curves
$P_{fit\text{-}Lxth}\{G_i\} = \mathrm{polyfit}\{P_{Gi\text{-}Lxth\text{-}Filter}\}$
$P_{fit\text{-}Uxth}\{G_i\} = \mathrm{polyfit}\{P_{Gi\text{-}Uxth\text{-}Filter}\}$

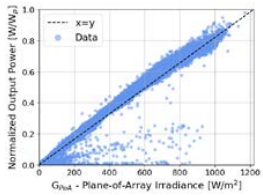
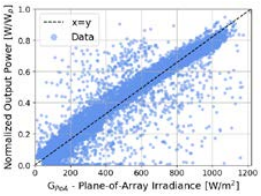
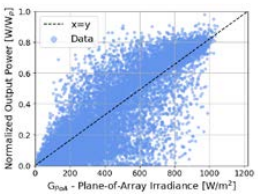
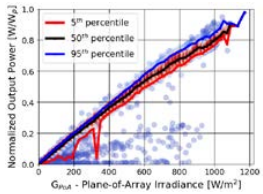
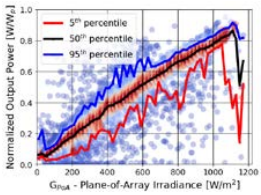
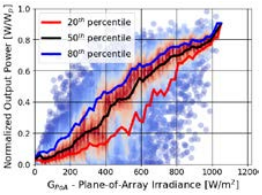
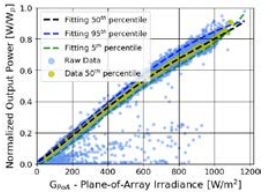
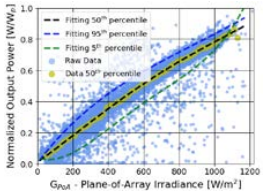
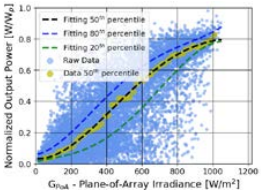
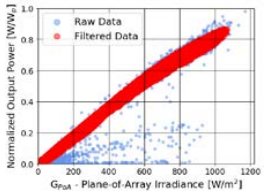
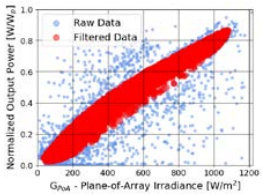
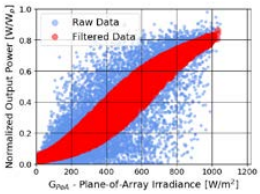
4. Results

The differences in accuracy are determined by the type of irradiance used. For the case of "high accuracy" dataset, the PoA irradiance is measured. The "medium accuracy" dataset constitutes the horizontal irradiance measured on-site and decomposition modelling by using the Erbs model and the Hay-Davies model to model transposition to estimate the PoA irradiance. The "low accuracy" dataset uses the *GHI* extracted from the ERA5 climate reanalysis dataset provided by the ECMWF and modelled as the previous case.

The filtering process is applied to the training set, and the results are presented step-by-step in Table 2. For high and medium accuracy, the input parameters for $L_{xth}$and U_{xth} are equal to 5th and 95th percentile, respectively, while for low accuracy, they are set to 20th and 80th percentile. Those values are selected manually in accordance with the level of data accuracy. The long-term evaluation of the

PV energy yield is presented in Figure 4. The monthly *PR* is calculated from the high accuracy dataset. The linear regression by using the Holt-Winters (HW) seasonal exponential smoothing method and the linear regression calculated to identify degradation of the PV system close to −0.27%/a. The average *PR* of the training set and test set are 0.882%/a and 0.872%/a, respectively. Thus, not large deviations in the system operation can be found between "training" and "test" datasets.

Table 2. Filtering algorithm step-by-step on the datasets for each level of data accuracy.

High Accuracy Data	Medium Accuracy Data	Low Accuracy Data
(1) Measured Power Output (2) On-site measured G_{PoA}	(1) Measured Power Output (2) On-site measured *GHI* (3) Estimated G_{PoA} from *GHI*	(1) Measured Power Output (2) Reanalysis *GHI* (3) Estimated G_{PoA} from *GHI*
Parameters: L_{xth} = 5th, U_{xth}= 95th	Parameters: L_{xth} = 5th, U_{xth}= 95th	Parameters: L_{xth} = 20th, U_{xth} = 80th
		
		
		
		

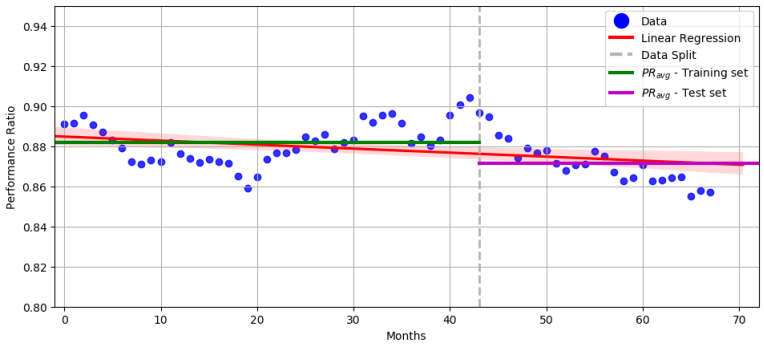


Figure 4. Linear regression of monthly Performance Ratio calculated using the Holt-Winters (HW) seasonal exponential smoothing for the 17-kW PV system in the period May 2014–Dec 2019. The average *PR* for training set and test set are also presented.

The accuracy of each model, in view of the climate conditions, can be easily observed by clustering the standard errors for irradiances (see Figure 5a) and for temperatures (see Figure 5b). Both plots illustrate the mode per cluster using the high accuracy dataset. The coloured dots represent the Gaussian distribution of the LightGBM model per each cluster.

Machine learning approaches perform well also under low irradiance conditions (below 300 W/m^2). A systematically linear error is observed as a function of the temperature (see Figure 5b), which could be improved by adding new features such as wind speed or measured back-side module operating temperature.

In the case of the empirical models, the best model is the #1 (modified PVGIS model), which uses irradiance and temperature as data input. Empirical models #2 and #3 show their limitations when using only irradiance as an input variable, clearly observed in Figure 5b, where the best accuracy is achieved at around 25 °C, which is also defined as the temperature at standard test conditions.

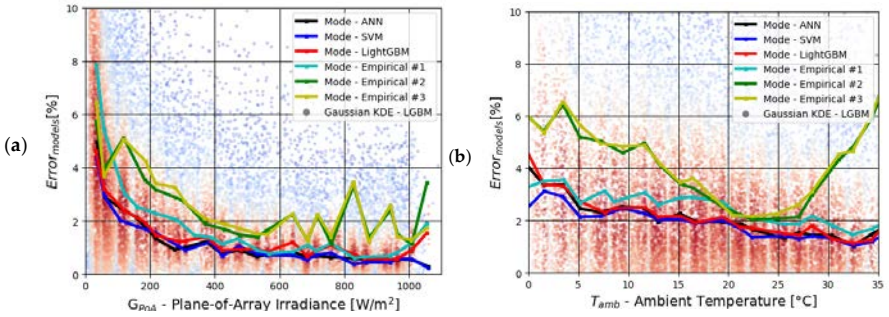


Figure 5. Mode of the standard error per cluster of (**a**) PoA irradiance and (**b**) ambient temperature for each model. Coloured dots illustrate as an example of the Gaussian distribution of the Light Gradient Boosting Machine (LightGBM) model per cluster.

In Table 3, the nRMSE of predicted values for empirical and machine learning approaches are presented for different levels of data accuracy. We compared three cases with regard to the filtering procedure: (1) the training and testing is carried out with raw datasets (no filtering stages); (2) the models are trained with filtered data and the testing is carried out on the raw data; and (3), the models are trained with filtered data and the removal of outliers applied to the predicted values.

The nRMSE on the testing stage decreases considerably if the input data accuracy is higher. In low accuracy dataset, the training stage is similar among empirical and ML approaches. The integration of a filtering stage helps to obtain better predictions using ML approaches for Medium and high data accuracy. The ML approaches show a performance close to 1% nRMSE when validating the predicted values without outliers.

The processing time of LightGBM on a standard PC (processor i5-4200U and 8GB of RAM) is much lower compared to the ANN and SVM, and it does not vary significantly under different data accuracies. It is evident that the LightGBM approach offers high modelling accuracy with the processing time of an empirical approach.

The performance of the machine learning approaches using a high accuracy data set as an input will also depend on the selected features. In Table 4, the nRMSE of the models are presented by using only the G_{PoA} as input, the G_{PoA}, and the T_{amb} together, and a third case, including the sun position (SP) defined by the sun azimuth and sun zenith.

Table 3. Uncertainty indicators of ML approaches for different levels of data accuracy, including the normalized root-mean-square-error (nRMSE) of the testing set and the processing time. Features considered: G_{PoA}, T_{amb}, and sun position angles.

Model	Low Accuracy (%, s)				Medium Accuracy (%, s)				High Accuracy (%, s)			
	(1)	(2)	(3)	Speed	(1)	(2)	(3)	Speed	(1)	(2)	(3)	Speed
LightGBM	11.64	12.19	4.58	0.031	5.93	5.89	2.47	0.047	1.98	1.45	0.99	0.047
ANN	11.55	12.08	4.52	75.69	6.11	6.03	2.45	85.43	2.27	1.46	1.01	133.09
SVM	12.05	12.18	4.59	4.33	5.92	5.98	2.49	15.47	1.44	1.45	1.01	38.74
Empirical #1	12.07	12.26	5.05	0.031	6.41	6.32	2.99	0.031	2.33	1.60	1.18	0.031
Empirical #2	11.91	12.18	4.98	0.078	6.40	6.52	3.41	0.047	2.03	2.13	1.46	0.063
Empirical #3	11.91	12.23	4.86	0.125	6.36	6.57	3.40	0.203	2.02	2.12	1.45	0.188

(1) Raw Training and Test sets. No filtering algorithm applied; (2) Filtered Training set and Raw Unfiltered Test set; (3) Filtered Training set and removal of outliers on the Test set.

Table 4. nRMSE of each machine learning approach using different input features for the high accuracy dataset.

Input Features	G_{PoA}	G_{PoA} + T_{amb}	G_{PoA} + T_{amb} +SP
LightGBM	1.49 %	1.10 %	0.99 %
ANN	1.44 %	1.16 %	1.01 %
SVM	1.52%	1.19%	1.01%

5. Discussions

The modelling of the PV energy yield includes several sources of uncertainties. In this manuscript, we addressed the accuracy of the solar irradiance as input and the modelled PV energy yield using empirical and machine learning approaches. Additionally, we observe that the *GHI* obtained from the ERA5 reanalysis dataset underestimates the real solar resource, and site-specific adaption techniques should be considered before using this source.

In terms of the overall methodology, the proposed filtering algorithm, as well as the LightGBM machine learning approach, gives a robust solution for PV energy yield modelling using high accurate input data. Additionally, the automatization of the procedure can be achieved thanks to the fast time processing of the models used. For the same reason, the tuning of LightGBM can be quickly optimized.

The placing of the filtering stage gives the flexibility to address different applications. For example, in the case of forecasting or gap filling, the real output power is unknown; thus, the filtering algorithm cannot be used in the predicted values. In cases of failure detection, the real measured power needs to be compared to the predicted values; thus, a filtering algorithm has to be applied to the post-processed data to remove the outliers.

Regarding the input variables or features of ML approaches, by only considering irradiance as input, the methods achieve reasonable estimations. Including related operational variables such as the ambient temperature and sun position angles, the accuracies are even better. Additionally, the wind speed could improve the model by considering the cooling effect of modules.

Future research can be address to compare the accuracy of machine learning approaches for systems in different climate zones, where extreme weather conditions, such as snow loads, dust deposition, or strong wind gusts, are impacting the PV modules.

6. Conclusions

The energy yield of PV systems is one of the key indicators for technical and financial stakeholders since it is directly related to the return-of-investment of the project. For this reason, the modelling is an important task to be carried out. This variable depends mostly on the weather conditions, such as temperature and irradiance. If weather measurement devices are installed on-site, the data recorded can be considered as highly accurate. However, in many cases, limited or no data are measured at

the site location. For this reason, external sources of weather data and irradiance models, such as decomposition and transposition models, can help to fill the gaps of missing data, but in general, losing accuracy.

In this article, we compared three empirical models, and three machine learning approaches used to model the energy yield of PV systems together with the new data filtering procedure. The filtering procedure is based on the predefined middle and side data percentile ranges where a new efficient smoothing step of side limits is used. Data filtering is an essential step to be applied in both sets of data (training and testing) due to a large number of unrealistic data/outliers contained in PV-related time-series. The outliers are a result of wrong measurements or unnatural conditions (i.e., a sudden temporal change in the irradiance due to clouds or reflections which is not expressed in the PV module temperature change). Additionally, when using a transposition model, unfiltered data will also be largely influenced by the transposition model error.

The applicability of all six PV energy yield approaches is demonstrated on three levels of accuracy of input irradiation data: measured on-site, estimated from satellite and reanalysis models. The decomposition and transposition models are used to obtain the Plane-of-Array irradiance when it is not measured.

The comparison of all approaches showed that LightGBM is the most performing algorithm in terms of accuracy and processing time. Given the three levels of data accuracy, the use of low accuracy data results in a nRMSE below 5%. At the medium data accuracy it gets close to 2%, and at high data accuracy, the nRMSE can reach below 1%.

Author Contributions: J.A.-V., J.B. and K.R. worked on the data processing, modelling, and analyses. K.B., M.J. and M.T. discussed the results. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the European Union's H2020 programme SOLAR-TRAIN under grant agreement No 721452 and research programme P2-0197 funded by Slovenia Research Agency.

Conflicts of Interest: The authors declare no conflict of interest.

List of Symbols and Abbreviations

T_{amb}	Ambient Temperature (°C)
ANN	Artificial Neural Networks
ECMWF	European Centre for Medium-Range Weather Forecasts
EFB	Exclusive Feature Bundling
GBDT	Gradient Boosting Decision Tree
GHI	Global horizontal irradiance (W/m^2)
GOSS	Gradient-based One-Side Sampling
G_{PoA}	Global PoA irradiance (W/m^2)
HW	Holt-Winters
KGPV	Köppen-Geiger-Photovoltaic
LightGBM	Light Gradient Boosting Machine
nRMSE	normalized Root-Mean-Square-Error (%)
PoA	Plane-of-Array
PR	Performance Ratio
PV	Photovoltaic
nMBE	normalized Mean Bias Error (%)
SVM	Support Vector Machines

References

1. ETIP PV: The European Technology and Innovation Platform for Photovoltaics Photovoltaic Solar Energy: Big and Beyond. Sustainable Energy to Reach the 1.5 Degrees Climate Target. Available online: https://etip-pv.eu/news/other-news/photovoltaic-solar-energy-big-and-beyond-etip-pv-publishes-vision-for-future-energy-supply/ (accessed on 31 January 2020).
2. Urraca, R.; Gracia-Amillo, A.M.; Huld, T.; Martinez-de-Pison, F.J.; Trentmann, J.; Lindfors, A.V.; Riihelä, A.; Sanz-Garcia, A. Quality control of global solar radiation data with satellite-based products. *Sol. Energy* **2017**, *158*, 49–62. [CrossRef]
3. Palmer, D.; Koubli, E.; Cole, I.; Betts, T.; Gottschalg, R. Satellite or ground-based measurements for production of site specific hourly irradiance data: Which is most accurate and where? *Sol. Energy* **2018**, *165*, 240–255. [CrossRef]
4. Urraca, R.; Huld, T.; Gracia-Amillo, A.; Martinez-de-Pison, F.J.; Kaspar, F.; Sanz-Garcia, A. Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. *Sol. Energy* **2018**, *164*, 339–354. [CrossRef]
5. Ascencio-Vásquez, J.; Brecl, K.; Topič, M. Methodology of Köppen-Geiger-Photovoltaic climate classification and implications to worldwide mapping of PV system performance. *Sol. Energy* **2019**, *191*, 672–685. [CrossRef]
6. Copernicus Climate Change Service (C3S) ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. In *Copernicus Climate Change Service Climate Data Store (CDS)*; Copernicus Climate Change Service (C3S), 2017.
7. Ascencio-Vásquez, J.; Kaaya, I.; Brecl, K.; Weiss, K.A.; Topič, M. Global Climate Data Processing and Mapping of Degradation Mechanisms and Degradation Rates of PV Modules. *Energies* **2019**, *12*, 4749. [CrossRef]
8. Camargo, L.R.; Schmidt, J. Simulation of Long-Term Time Series of Solar Photovoltaic Power: Is the ERA5-Land Reanalysis the Next Big Step? Available online: https://arxiv.org/abs/2003.04131 (accessed on 6 March 2020).
9. Babar, B.; Graversen, R.; Boström, T. Solar radiation estimation at high latitudes: Assessment of the CMSAF databases, ASR and ERA5. *Sol. Energy* **2019**, *182*, 397–411. [CrossRef]
10. Jiang, H.; Yang, Y.; Bai, Y.; Wang, H. Evaluation of the Total, Direct, and Diffuse Solar Radiations From the ERA5 Reanalysis Data in China. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 47–51. [CrossRef]
11. Lave, M.; Hayes, W.; Pohl, A.; Hansen, C.W. Evaluation of Global Horizontal Irradiance to Plane-of-Array Irradiance Models at Locations Across the United States. *IEEE J. Photovolt.* **2015**, *5*, 597–606. [CrossRef]
12. Mosavi, A.; Salimi, M.; Faizollahzadeh Ardabili, S.; Rabczuk, T.; Shamshirband, S.; Varkonyi-Koczy, A. State of the Art of Machine Learning Models in Energy Systems, a Systematic Review. *Energies* **2019**, *12*, 1301. [CrossRef]
13. Kirn, B.; Brecl, K.; Topič, M. A new PV module performance model based on separation of diffuse and direct light. *Sol. Energy* **2015**, *113*, 212–220. [CrossRef]
14. Livera, A.; Theristis, M.; Makrides, G.; Sutterlueti, J.; Ransome, S.; Georghiou, G.E. Performance Analysis of Mechanistic and Machine Learning models for Photovoltaic energy yield prediction. In Proceedings of the 36th European Photovoltaic Solar Energy Conference and Exhibition, Marseille, France, 9–13 September 2019; pp. 1272–1277.
15. Fernández, Á.; Gala, Y.; Dorronsoro, J.R. Machine Learning Prediction of Large Area Photovoltaic Energy Production. In *Data Analytics for Renewable Energy Integration*; Woon, W.L., Aung, Z., Madnick, S., Eds.; Springer International Publishing: Cham, Switzerland, 2014; Volume 8817, pp. 38–53. ISBN 978-3-319-13289-1.
16. Mellit, A.; Massi Pavan, A.; Ogliari, E.; Leva, S.; Lughi, V. Advanced Methods for Photovoltaic Output Power Forecasting: A Review. *Appl. Sci.* **2020**, *10*, 487. [CrossRef]
17. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIP 2017), Long Beach, CA, USA, 4–9 December 2017.
18. Daoud, E.A. Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. *Int. J. Comput. Inf. Eng.* **2019**, *13*. [CrossRef]
19. Minastireanu, E.-A.; Mesnita, G. Light GBM Machine Learning Algorithm to Online Click Fraud Detection. *J. Inf. Assur. Cybersecur.* **2019**, *2019*, 15. [CrossRef]

20. Machado, M.R.; Karray, S.; de Sousa, I.T. LightGBM: An Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry. In Proceedings of the 2019 14th International Conference on Computer Science & Education (ICCSE), Toronto, ON, Canada, 19–21 August 2019; pp. 1111–1116.
21. Song, Y.; Jiao, X.; Qiao, Y.; Liu, X.; Qiang, Y.; Liu, Z. Prediction of Double-High Biochemical Indicators Based on LightGBM and XGBoost. In Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science—AICS 2019, Wuhan, Hubei, China, 12–13 July 2019; pp. 189–193.
22. Zhang, J.; Mucs, D.; Norinder, U.; Svensson, F. LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 4150–4158. [CrossRef]
23. Zeng, H.; Yang, C.; Zhang, H.; Wu, Z.; Zhang, J.; Dai, G.; Babiloni, F.; Kong, W. A LightGBM-Based EEG Analysis Method for Driver Mental States Classification. *Comput. Intell. Neurosci.* **2019**, *2019*, 1–11. [CrossRef] [PubMed]
24. Wang, D.; Zhang, Y.; Zhao, Y. LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients. In Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics—ICCBB 2017, Newark, NJ, USA, 18–20 October 2017; pp. 7–11.
25. Reba, K.; Bevc, J.; Ascencio-Vásquez, J.; Jankovec, M.; Topič, M. Photovoltaic Energy Production Forecasting using LightGBM. In Proceedings of the 55rd International Conference on Microelectronics, Devices and Materials, Bled, Slovenia, 22–27 September 2019.
26. Mariottini, F.; Belluardo, G.; Bliss, M.; Isherwood, P.J.M.; Cole, I.R.; Betts, T.R. Assessment and improvement of thermoelectric pyranometer measurements. In Proceedings of the 36th European Photovoltaic Solar Energy Conference and Exhibition, Marseille, France, 9–13 September 2019.
27. Huld, T.; Amillo, A. Estimating PV Module Performance over Large Geographical Regions: The Role of Irradiance, Air Temperature, Wind Speed and Solar Spectrum. *Energies* **2015**, *8*, 5159–5181. [CrossRef]
28. Huld, T.; Gottschalg, R.; Beyer, H.G.; Topič, M. Mapping the performance of PV modules, effects of module type and data averaging. *Sol. Energy* **2010**, *84*, 324–338. [CrossRef]
29. European Commission, Joint Research Centre Photovoltaic Geographical Information System (PVGIS), Online Tool. Available online: https://ec.europa.eu/jrc/en/pvgis (accessed on 3 December 2019).
30. Ransome, S.; Sutterlueti, J. How to Choose the Best Empirical Model for Optimum Energy Yield Predictions. In Proceedings of the 2017 IEEE 44th Photovoltaic Specialist Conference (PVSC), Washington, DC, USA, 25–30 June 2017; pp. 652–657.
31. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press, Inc.: Oxford, UK, 1995; ISBN 0-19-853864-2.
32. Drucker, H.; Burges, C.J.C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Denver, CO, USA, 1–6 December 1997; pp. 155–161.
33. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2000**, *29*, 1189–1232. [CrossRef]
34. Anghel, A.; Papandreou, N.; Parnell, T.; De Palma, A.; Pozidis, H. Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms. *arXiv* **2019**, arXiv:1809.04559.
35. Tsafarakis, O.; Sinapis, K.; van Sark, W. PV System Performance Evaluation by Clustering Production Data to Normal and Non-Normal Operation. *Energies* **2018**, *11*, 977. [CrossRef]
36. Theristis, M.; Stein, J.S. PV Degradation Modeling, PV Performance Modeling Collaborative, Sandia National Laboratories, SAND2019-15366 W. Available online: https://pvpmc.sandia.gov/pv-research/pv-lifetime-project/pv-degradation-modeling/ (accessed on 12 February 2020).
37. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.