

Article

Not peer-reviewed version

Pv Production Forecast Using Hybrid Models of Time Series with Machine Learning Methods

Thomas Haupt , [Oscar Trull](#) ^{*} , [Mathias Moog](#)

Posted Date: 7 April 2025

doi: 10.20944/preprints202504.0564.v1

Keywords: time series; PV; forecast; machine learning



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

PV Production Forecast Using Hybrid Models of Time Series with Machine Learning Methods

Thomas Haupt ¹, Oscar Trull ^{2,*} and Mathias Moog ¹

¹ Hochschule Ansbach, 91522 Ansbach, Germany

² Department of Applied Statistics, Operational Research and Quality, Universitat Politècnica de València, 46022 Valencia, Spain

* Correspondence: otrull@eio.upv.es

Abstract: Photovoltaic (PV) energy production in Western countries increases yearly. Its production can be carried out in a highly distributed manner, not being necessary to use large concentrations of solar panels. As a result of this situation, electricity production through PV has spread to homes and open-field plans. Production varies substantially depending on the panels' location and weather conditions. However, the integration of PV systems presents a challenge for both grid planning and operation. Furthermore, the predictability of rooftop-installed PV systems can play an essential role in home energy management systems (HEMS) for optimising local self-consumption and integrating small PV systems in the low-voltage grid. In this article, we show a novel methodology used to predict the electrical energy production of a 48 kWp PV system located at the Campus Feuchtwangen, part of Hochschule Ansbach. This methodology involves hybrid time series techniques that include state space models supported by artificial intelligence tools to produce predictions. The results show an accuracy of around 3% on nRMSE for the prediction, depending on the different system orientations.

Keywords: time series; PV; forecast; machine learning

1. Introduction

The installed PV capacity in Europe has risen dramatically from 30 GWp in 2010 to over 250 GWp in 2022 [1]. By the end of 2024, Germany alone has installed 99 GWp. With the goal of achieving climate neutrality, the installed capacity in Germany is expected to reach up to 400 GWp by 2040 [2]. A significant advantage of PV systems is their low levelized cost of electricity (LCOE). Additionally, the nominal efficiency rate is increasing by 0.4 percentage points annually, with current average efficiencies around 21% [3]. Integrating up to 400 GWp of PV generation compared to a load of 80 GW is already a challenge in Germany, especially in the low-voltage grids and strong rural regions.

Apart from participation in the day-ahead market, PV forecasting will play an essential role in the operation of low-voltage grids to implement § 9 EEG. The grid-friendly operation requires local generation and consumption forecasts. However, the accuracy of forecasts for this application is still a relevant but open question. On the other hand, cost-optimised self-consumption in buildings is already playing a significant role [4]. Battery storage systems typically enable this. Additional options for optimised PV self-consumption include battery electric vehicles (BEVs) as well as heat pumps, which are so-called flexible consumers [5]. The optimised control of flexible consumers requires an energy management system (EMS) or, with a focus on single-family households, home energy management systems (HEMS) [6]. The challenge of providing sufficiently precise forecasts for this purpose remains an open research question. In the past, production with traditional generation methods based on thermal sources and very stable prediction gave way to systems that depended on the weather in addition to the functional state of the facilities [7]. The continued increase in generation from renewable sources has led to an increase in the prediction error that has doubled in the last five years and continues to increase [8]. According to the International Energy Agency (IEA), in 2015, there was a production capacity of 196 GW using PV, and a capacity of between 400 and 800 GW was

planned for 2020 [9]. The solution consists of having backup systems generally based on batteries and natural gas power plants, which is expensive [10]. Production stability is also a critical point in prediction. Clear lighting conditions are constant in southern European climates, with little cloud cover. In central Europe, this cloudiness is highly variable, where up to 1/3 of the days of the year can be reached with high cloudiness. This further increases the difficulty of making predictions [11].

In contrast to the systems that are mounted on pitched roofs and point in the direction of the roof, flat roofs offer more degrees of freedom. One of the challenges - also with the system installed on the Campus Feuchtwangen (49°10'47.3" N 10°20'07.9" E) - is that the orientation of the production units not uniform but is distributed along the roof with varying orientations, making it impossible to use global irradiance as a basis for calculation for predicting energy production. For this reason, statistical methods were used for this task. However, given the complexity of the prediction, which must include environmental factors and a high degree of randomness in the observed data, it was decided to include a layer of machine learning that could complement the predictions. The result is a hybrid model that provides predictions at a considerable level.

In this article, we show the electric energy production prediction method that we have at the Campus Feuchtwangen for the solar panels that we have on the Feuchtwangen campus. We present a new model that includes the use of state-space models as the basis for the structural part of the time series, supported by artificial intelligence models that, through the inclusion of meteorological variables, can make accurate predictions of photovoltaic energy production.

It is structured as follows: Section 2 analyses the related work to this article; Section 3 explains the methods used in the article that led to the results shown in Section 4. Sections 5 and 6 discuss the results and raise conclusions.

2. Related Work

The application of PV forecasts is often discussed in the context of HEMS for optimised PV self-consumption. On the one hand, HEMS are simulated in models where perfect forecasts are used for optimisation [6] or focusing on data prediction using rolling horizons within a 24-hour timeframe [12]. A study surveying the practical implementation of 26 HEMS manufacturers revealed that over 50% of these systems currently incorporate PV forecasts [13].

An exact PV forecast will be the prevention of the curtailment of PV systems by §9 EEG. By forecasting generation over the 24 hours ahead, flexible consumers and their storage systems can be controlled to match. However, an intriguing and unresolved issue is the temporal resolution of these forecasts and how they are integrated into the optimisation algorithms. Another critical and increasingly significant aspect is the inclusion of demand forecasts for electricity and heat and generation forecasts.

Reviews of PV forecasts have been abundant recently [14,15]. There is a growing interest in improving predictions, as a considerable increase in installed capacity is expected [9].

Production prediction is generally developed from irradiance prediction, especially global horizontal irradiance (GHI). Other variables affect it, such as cloudiness, but the main one is the GHI. The models usually distinguish between a structural part with Irradiance and another stochastic part where time series models and other methodologies are used [16]. Wu et al. [17] analyse the methods applied to predict solar production. A distinction is made between methods based on time series and machine learning. The former is based on statistical models to determine the new predictions, such as exponential smoothing or ARIMA models. The second group is very varied, highlighting neural network, ensemble, and deep learning models. However, hybrid models stand out. Probabilistic models are also emerging as a new alternative. However, time series models in this type of task are usually combined with elements of machine learning [18]

Kavakci et al. [19] use a hybrid prediction method using time series decomposition; it models the trend as a stable element. It uses ML to model the seasonal and stochastic parts. Aziz and Chowdhury [20] apply ARIMA models for production prediction in Bangladesh. This process is more

straightforward, where they use prior information to make univariate predictions. Chodakowska et al. [21] applies SARIMAX models to Jordan and Poland PV stations.

However, there is an incipient growth of academic interest in the development of machine-learning models in this area. Saigustia and Pijarski [22] use XGBoost to determine production in Spain. The results shown are exceptional in terms of fit, although no prediction results are shown. Currently, Long Short-Term Memory (LSTM) Networks and their derivatives are the models that show the most accurate results compared to the rest of the models [23,24]. Scott et al. [25] developed random forest (RF), NN, SVM, and LR models for prediction in different locations, such as Denmark, Holland, America, and Portugal. The proposed models can improve and reduce by half the error measured in terms of RMSE.

One of the main aspects of prediction is the prediction horizon. Predictions can be made for the short or long term. The short-term forecasts (one day ahead) are used for planning, while the former is used for specific operations, with the most common being predictions for one day ahead [26].

3. Materials and Methods

The Campus Feuchtwangen has been operational as a research building since 2018. This facility collects comprehensive data encompassing thermal, electrical, and environmental parameters such as temperature, humidity, and light levels. Additionally, the site is equipped with two independent weather stations that measure various meteorological data, including solar radiation, temperature, humidity, and wind. All this data is logged at one-minute intervals, providing a detailed and continuous record. Additionally, the site is integrated with the Sector Coupling Laboratory.

This laboratory connects flexible loads such as battery storage systems, heat pumps, and charging stations for electric vehicles. Figure 1 pictures this installation. The long-term objective is to optimise the operation of these interconnected systems using forecast data from HEMS.



Figure 1. Solar and weather station installation on the Feuchtwangen campus.

The PV system has a total module capacity of 48 kWp. It consists of 150 modules of the type LG320N1C-G4. The system is divided into two subsystems. System 1 has a module capacity of 20 kWp and uses a 17-kW inverter (SolarEdge 17K). It includes two strings with 31 modules each (9.92 kWp per string), both facing west. System 2 has a module capacity of 28 kWp and a 27.6 kW inverter (SolarEdge 27.6K). It consists of three strings: one with 27 modules facing south (8.64 kWp), one with 32 modules facing east (10.24 kWp), and one with 29 modules also facing east (9.28 kWp). Each module is equipped with a power optimiser (SolarEdge P370) to improve system performance. Overall, the building records approximately 1,000 data points as time series. For this analysis, the collected data includes global solar radiation, temperature, and the AC production profiles of the two PV systems.

Additionally, data from Weather Station 7369 is utilised. The data is tracked by a data logger at a 60-second resolution and stored in an Influx database. A significant portion of the data has been historically recorded for over five years. The long-term goal is to use the PV surplus and forecast data

to optimise the control of flexible loads in a CO₂- and cost-efficient manner. Figure 2 shows the functioning of the HEMS linked to data acquisition and forecasting.

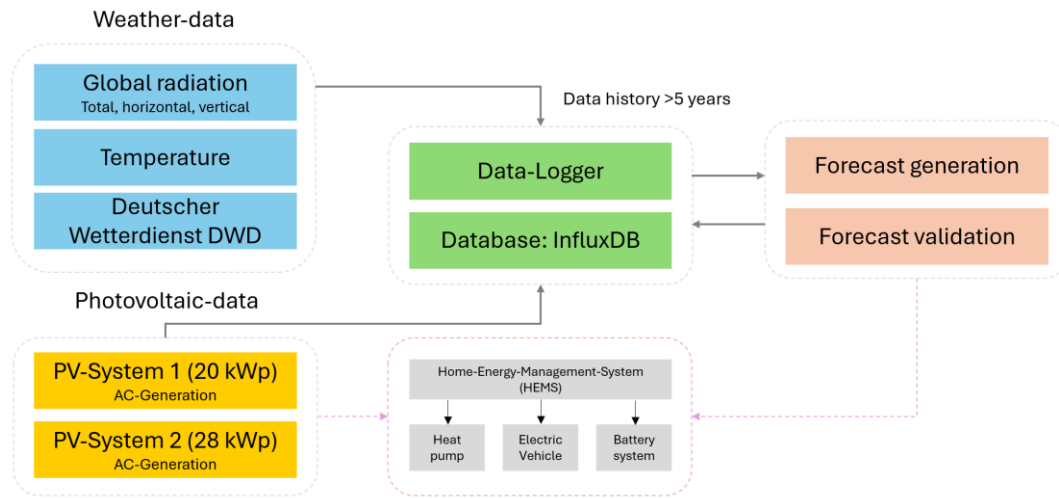


Figure 2. HEMS working schema.

3.1. Machine Learning Methods

3.1.1. Neural Networks

Neural networks are computational models composed of neurons, which are interconnected nodes organised in layers, that produce an output signal processing inputs received from other neurons or input data. The neurons are organised in layers: an input layer, where neurons are connected to the input variables, and an output layer, connected to output variables. All of them are interconnected through the neurons in the hidden layers. This structure can be seen in Figure 3.

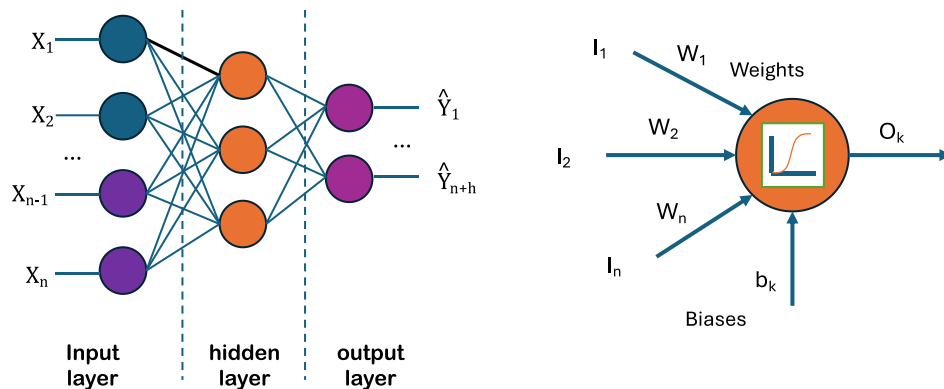


Figure 3. Neural network architecture.

There, y_t represents the observed data and \hat{y}_t the predicted values. X_t are variables related to the model. The way the neurons are connected is through weighted links W_i called axioms, with additional biases b_i , and are processed through an activation function f to obtain an output O_i . The activation function introduces non-linearity properties to the network. The function can be either a linear activation, sigmoid, Tanh, or ReLU, among others, depending on the objective of the NN [17]. The output function related to each neuron can be stated as in **Error! Reference source not found..** n means the total number of neurons connected to every neuron.

$$O_k = \sum_{i=1}^n W_i I_i + b_k \quad (1)$$

One common type of neural network is the feedforward neural network, which consists of an input layer, one or more hidden layers, and an output layer. The input layer receives the initial data,

such as images, text, or numerical values. Each neuron in the hidden layers applies a weighted sum of inputs and a nonlinear activation function, transforming the input data into a form suitable for the next layer. Finally, the output layer produces the network's prediction or classification.

The global formulation of the neural network can be expressed as in **Error! Reference source not found..**

$$\hat{y}_{t+1} = f(X_1, X_2, \dots, X_n) \quad (2)$$

Training a neural network involves adjusting the weights of connections between neurons to minimize the difference between the network's predictions and the actual targets. This is typically done using optimisation algorithms like gradient descent, combined with a loss function that quantifies the difference between predicted and actual values, measured using Mean Squared Error (MSE).

3.1.2. Least Squares Boosting Ensemble (LSBoost)

Ensemble methods are becoming more important every day in machine learning. The philosophy of working with these methods is to use apparently simple techniques that would not be expected to obtain great results and combine them so that the precision of their predictions increases considerably. Sets can be made using one of three different ways: bagging, boosting, and stacking. In our case, we have chosen to use the second.

Boosting leverages previous predictor errors to enhance future predictions. By amalgamating multiple weak base learners into a single strong learner, boosting markedly enhances model predictability. The process involves organising weak learners sequentially, allowing each learner to learn from its predecessor, thereby iteratively refining predictive models.

LSBoost algorithm [27] consists of sequentially building a series of regression trees, in which residuals of the previous trees are predicted through a loss function as in **Error! Reference source not found..**

$$L(y_t, F_m) = \frac{(y_t - F_m)^2}{2} \quad (3)$$

Being F_m the regression tree on each stage and the first term $F_0(X) = y$. Each subsequent tree focuses on reducing the errors made by the previous trees, see eq. **Error! Reference source not found.** and thereby gradually improving the overall prediction accuracy.

$$(X, \{\rho_m; \alpha_m\}) = \sum_{m=1}^M \rho_m h(X, \alpha_m) \quad (4)$$

With M the maximum number of trees, and h being a simple parametrized function of the observed data X using the parameters α and ρ . The final prediction is obtained by summing up the predictions from all the trees in the ensemble. An optimisation problem by minimising the errors is applied, and the numerical values of the parameters, as shown in **Error! Reference source not found.**, are obtained.

$$(\rho_m, \alpha_m) = \operatorname{argmin}_{\rho, \alpha} L(y_t, F_m) \quad (5)$$

This direct optimisation of the loss function makes LSBoost well-suited for regression problems, where the goal is to minimise the squared differences between predicted and observed values [28].

3.2. Novel Hybrid TBATS-ML Method

An exhaustive analysis of the time series reveals a deterministic component that traditional methodologies, such as exponential smoothing can predict. On this occasion, we have considered the use of state space models such as TBATS (Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components) [29]. The TBATS models can be expressed in their reduced format according to **Error! Reference source not found..**

$$\varphi_p(L)\eta(L)y_t^{(\omega)} = \theta_q(L)\delta(L)\varepsilon_t, \quad (6)$$

where $\varphi_p(L)$ and $\theta_q(L)$ are polynomials of length p and q , being L the lag operator. $y_t^{(\omega)}$ stands for the Box-Cox transformed y_t with the Box-Cox parameter ω . Finally,

$$\eta(L) = (1-L)(1-\phi L) \prod_{i=1}^T \prod_{j=1}^{k_i} (1 - 2\cos\lambda_i^j L + L^2)$$

$$\delta(L) = [L^2\phi(1-\alpha) + L(\alpha + \phi\beta - \phi - 1) + 1] \prod_{i=1}^T \prod_{j=1}^{k_i} (1 - 2\cos\lambda_i^j L + L^2) +$$

$$(1-L)(1-\phi L) \sum_{i=1}^T \sum_{j=1}^{k_i} \prod_{k=1, k \neq i}^T \prod_{l=1, l \neq j}^{k_l} (1 - 2\cos\lambda_l^{(k)} L + L^2) [(\cos\lambda_j^{(i)} \gamma_{1i} +$$

$$\sin\lambda_j^{(i)} \gamma_{2i}) L^2 - \gamma_{1i} L^3] + (1-L)(1-\phi L) L \prod_{i=1}^T \prod_{j=1}^{k_i} (1 - 2\cos\lambda_i^j L + L^2) \sum_{i=1}^T k_i \gamma_{1i}.$$

The designation TBATS($\omega, \phi, p, q, \{m_1, k_1\}, \{m_2, k_2\}, \dots, \{m_T, k_T\}$) of the model includes also the seasonal period length m_i and the number of harmonics k_i . For more information on how model fitting is performed and how to make forecasts, it is recommended to access the reference [29]. In our case, to achieve the functionality of the model, the forecast library in R has been used [30].

Once the TBATS state space prediction models have been obtained, this information from the deterministic part is used in an integrated way in A.I. models, similar to the process used in [31]. Within a neural network or ensemble regression model, the TBATS-adjusted variable is included in addition to a selected set of climatological variables from the DWD. New synthetic variables are added to this set to help the model in the prediction.

Finally, the model is trained by separating the training set into two subsets, one for fitting and one for validation of the model predictions. It is fitted by minimising the MSE (Mean Squared Error) and validated with the predictions made in the validation subset.

3.3. Accuracy Metrics

Metrics allow us to determine the model's accuracy level in making predictions, all of which are based on prediction error. There are many metrics and no one metric can be considered the most suitable to perform the tasks. However, it is customary to use the MSE to fit prediction models, while others, such as the root mean square error (RMSE), as shown in **Error! Reference source not found.**, and the mean average percentage error (MAPE), are used to compare prediction accuracy and precision. In this case, the MAPE is discarded because it requires the values to be strictly positive, which is not the case.

$$RMSE = \sqrt{\frac{(y_t - \hat{y}_t)^2}{N}} \quad (7)$$

where N is the total number of observed values. The problem with RMSE is that it values the error absolutely, and comparing two stations of different capacities may be ineffective. Following [32], $NRMSE_{installed}$, which is defined in **Error! Reference source not found.**, will be used.

$$NRMSE_{installed}(\%) = \frac{\sqrt{\frac{(y_t - \hat{y}_t)^2}{N}}}{P_{installed, peak}} \cdot 100 \quad (8)$$

4. Case Study: Prediction of Energy Production in PV Systems

The two solar panels located in the Feuchtwangen Campus at the Hochschule Ansbach are named PV-System1 (PV1), with a maximum peak production of 20 kW, and PV-System2 (PV2), with a maximum peak production of 30 kW. The data corresponds to the PV production from 2021 to March 2024, quarter-hourly measured. We reserved data from January 2024 to assess out-of-sample forecasts. The data was down-sampled for hourly measures to be synchronised with climate data obtained from DWD. Figure 4 shows the PV data for the period.

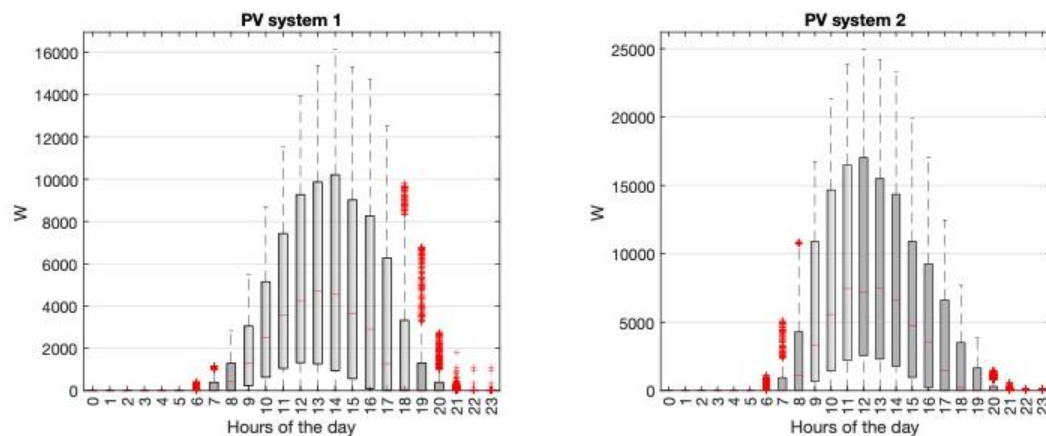


Figure 4. Solar energy production in solar panels PV-System 1 and PV-System 2 located in Campus Feuchtwangen (HS-Ansbach).

The figure shows the different behaviour of both PV installations. The arrangement of the solar panels on the roof of the installation is not the same for both installations. Irradiation and energy production differ slightly between the two systems regardless of weather conditions. If we observe the behaviour of both installations with respect to the irradiance received (see Figure 5), we can see that there is a higher ratio for PV2 compared to PV1 for both Global Horizontal Irradiance (GHI) and Global Vertical Irradiance (GVI), which corroborates the situation described above.

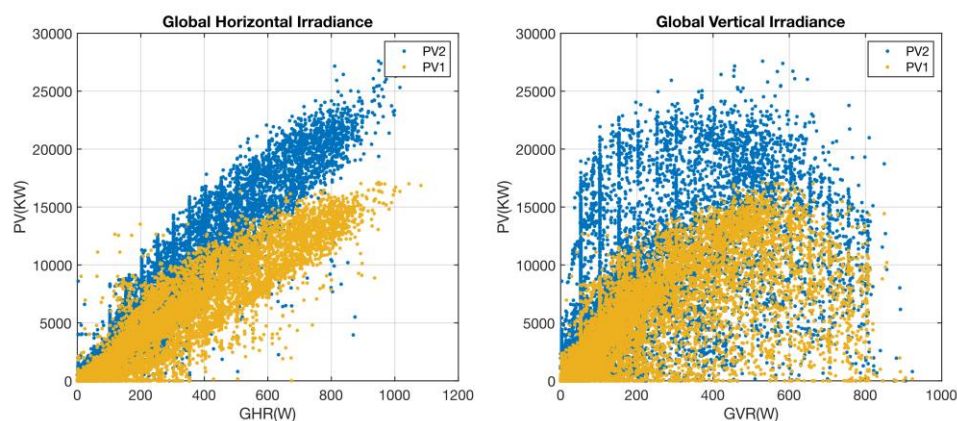


Figure 5. Global radiation relationship with PV production. The left panel represents GHI vs. PV, whilst the right panel represents GVI vs. PV.

Traditional methods of forecasting PV power from clear-sky radiation have limitations due to installation factors [33]. These methods often overlook localised atmospheric variations and installation-specific characteristics, such as panel orientation and shading effects. In addition, they do not account for dynamic weather patterns, resulting in inaccurate predictions. Advanced techniques incorporating site-specific data and machine learning are needed to improve forecast accuracy.

In any case, weather conditions are essential for solar PV production. Therefore, in order to carry out production forecasting, weather conditions must be taken into account. The climatological information used in this article has been obtained from the open database of the *Deutsche Wetterdienst* on its website <https://opendata.dwd.de/> [34].

4.1. Data Cleaning

In any forecasting process, the forecaster must solve existing problems in the time series, either missing data or outliers. On this occasion, it is also necessary to deal with clearly erroneous measurement data, which are produced by errors provided by sensors. These are errors that can go

unnoticed when training the models and introduce noise in the prediction models. Figure 6 reflects this situation, showing the values labelled as Feasible, Anomalous and mismeasurement. It can be seen how the data has been sectorized around the feasible region and the rest of the data to be corrected or eliminated.

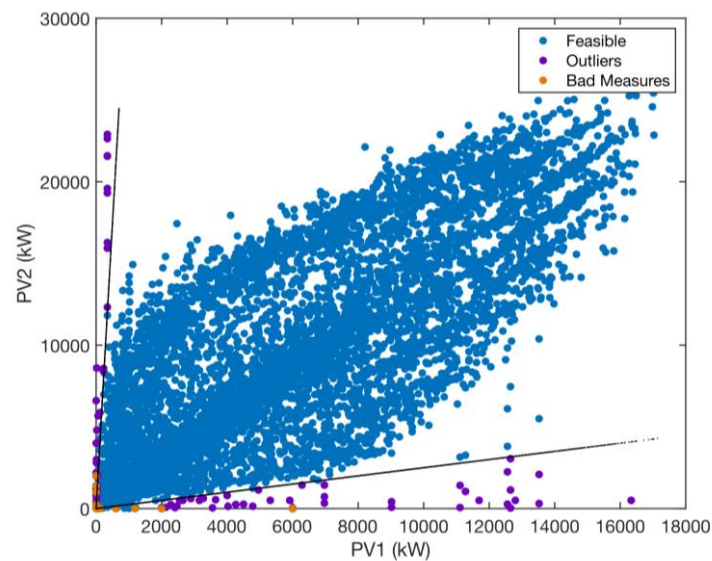


Figure 6. Bad measures and outliers determination in PV1 and PV2.

In the case of climatological variables, there is a similar situation with a large amount of missing data. In this case, we have resorted to generate regression models by means of A.I. using data provided by nearby weather stations that could provide reliable data for those moments of time. Figure 7 shows the stations from which data could be obtained.

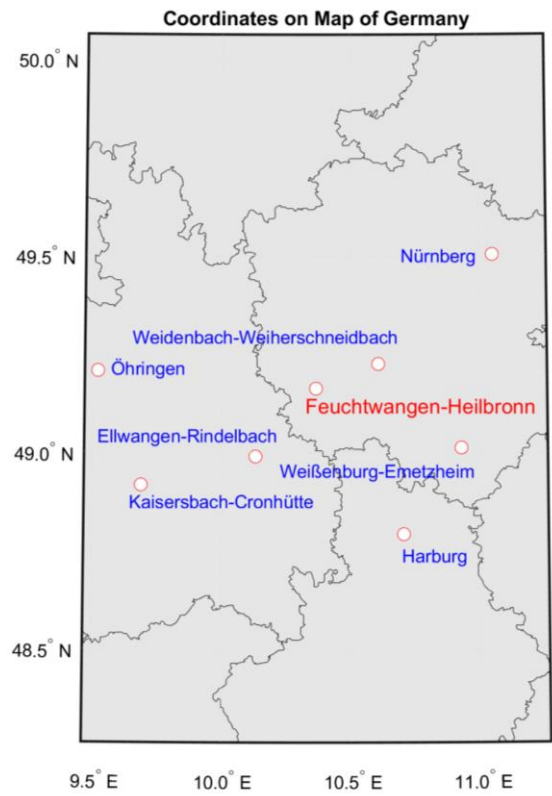


Figure 7. Map of weather stations near Feuchtwangen.

Finally, after completing the series, we proceeded to reconstruct the PV1, PV2, GHI and GVI series. The result of the data cleaning is shown in the attached video. It includes a representation of the variables PV1, PV2, GHI and GVI over time, both in their original and repaired values.

4.2. Training and Forecasting

The data set worked on in this article covers the period from January 1, 2022, to February 2024. These are the data collected by the Feuchtwangen Campus facility. This set has been divided into three subsets: The training set, which includes the first period from January 2022 until September 30, 2023. A second set for in-sample validation, until December 31, 2023. Finally, the out-of-sample test set until March 2024.

The training of the models and their validation has been carried out using the facility's own variables, in addition to the current climatological data provided by the DWD. In addition, synthetic variables that help the model to make predictions are included. These variables are listed in Table 1.

Table 1. Synthetic variables created for the models.

Name	Description	Values
sun	Sun	Binary
PV1/PV2 d24	Power measurements generated in PV1/PV2 delayed 24 hours	Real
Hour/month	Day/month of the year	Integer 1-24/1-12
DayOfYear	Consecutive day within the year	Integer 1-366
WeekOfYear	Calendar week within the year	Integer 1-53
TBATS	Data modeled using TBATS for the structural part of the model	Real

Subsequently, the models have been trained and the variables to be used in the predictions have been selected. The procedure includes a selection by Minimum Redundancy Maximum Relevance (MRMR) Algorithm [35]and F tests [36].

Due to the different characteristics of each installation, the variables used in each case do not coincide. This will be analysed later in the discussion of the results. Nevertheless, the procedure followed to perform the training and predictions with both facilities is the same.

In the case of prediction, two different prediction strategies have been analysed: In the first, current climatological data are used and forward prediction is performed using that observed base; in the second option, predicted climatological data provided by the DWD are used.

5. Results

The models have been analysed according to two strategies. The first one consists of using the current data, with which the models are trained, and from them make predictions of the energy production always based on the past. This makes sense if we take into account the inertia of time and the fact that day-to-day variations are not large. The other strategy is to make forecasts using the current production data, but using the climatological values predicted by the DWD. This delegates the responsibility of the climatology prediction to the DWD, although it depends on it, and it is necessary to know it to be able to make our predictions.

The differences between both strategies are not very dissimilar, however, this second strategy seems to provide better results independently of the installation and the models used, so it has been chosen to continue with it. The results are shown separately for each facility to facilitate reading comprehension.

5.1. PV System 1

The first step in the prediction was the estimation of the TBATS state space model using the “forecast” library of R. The model obtained was TBATS(1, {2,2}, -, {<24,6>, <8760,7>}). As can be seen, two seasonalities have been considered: an intraday and an interannual one.

Starting with all the variables offered by the DWD, an initial screening has been carried out by selecting those variables whose influence on the model is significant, both with the application of the F-Test method and the MRBR method. The use of more variables does not imply an increase in the precision of the model. Once the models have been adjusted, by means of an iterative procedure and aided by the SHAPLEY curves, the number of variables is reduced. Tests using principal component analysis have not been effective, so their use has been discarded.

The strategy selected for this installation has been the use of the predictions provided by DWD. Table 2 shows the selected climatological variables. The table shows the variables and the description of the data obtained from DWD.

Table 2. Main climate variables used in the forecasts.

Name	Description	Values
V_VV	visibility	m
TT, TT_TU	air temperature	°C
RF	relative humidity	%
FF	Wind speed	°C
V_TEXXX	Soil temperature in XXX cm depth	°C

The SHAPLEY graph shown in Figure 8 shows the importance of each of the variables included in the model. In this case, soil temperature is observed to be influential in the model.

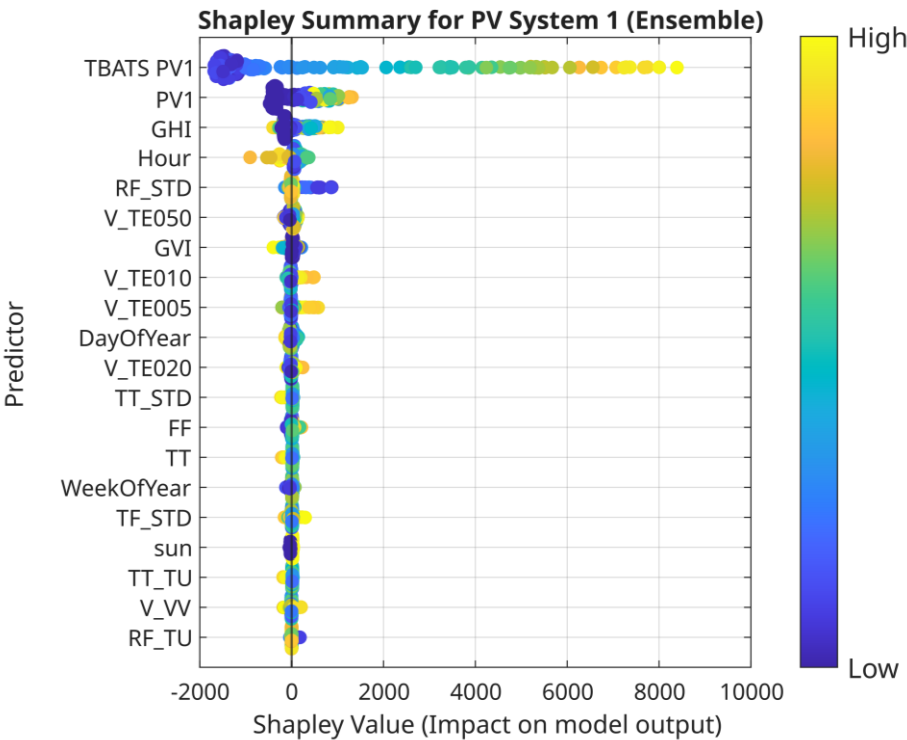


Figure 8. SHAPLEY diagram of the top 20 variables included in the prediction model for PV1. The colour indicates the value of the characteristic, with cool colours being those of lower value and warm colours those of higher value.

The model has been fitted and 24-hour forward predictions have been made on both the validation set and the out-of-sample test set. Table 3 shows a comparison of the models tested and the results obtained in their predictions.

Table 3. Comparison of tested models.

Model	Validation RMSE	Forecasting RMSE	Forecasting NRMSE
Ensemble (LSBoost, min.leaf size:1, L.R. 0.042)	622	445	2.25%
Neural Network (Shallow, 1 Layer Size: 24)	699	621	3.11%
Tree (Non surrogate)	688	626	3.13%

5.2. PV System 2

Following the same procedure as the previous case, the TBATS model obtained for the structural part of the model is: TBATS(1, {2,2}, -, {<24,6>, <8760,5>}).

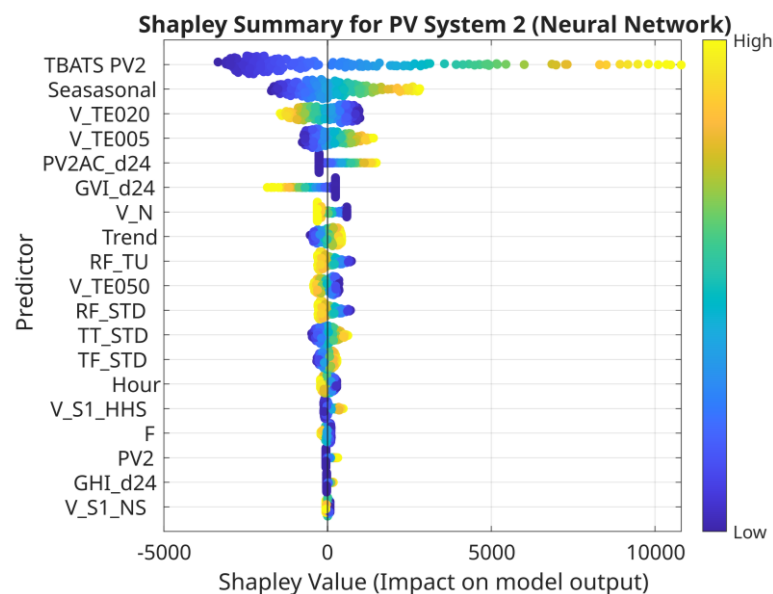
The climatological variables included in the models for PV-System 2 are shown in Table 4. As can be seen, the variables do not coincide with those in Table 2, corroborating in part the different characteristics of both installations. There is a lower presence of variables related to temperature and a higher presence of variables related to cloudiness.

Table 4. Climate variables used in the forecasts.

Name	Description	Values
V_N	visibility	1-8
TT_TU,TT_STD	air temperature	°C
RF_XXX	relative humidity	%
V_TEXXX	Soil temperature in XXX cm depth	°C
V_S1_Ns	Height of clouds 1 st layer	m

Finally, Table 5 shows the values of the metrics obtained using each of the models. On this occasion, the neural network model outperforms the ensemble model in both validation and out-of-sample test predictions.

Figure 9 shows the SHAPLEY diagram in swarm format for the model chosen for the PV 2 installation. The climatological components become more important and the seasonality is more pronounced.

**Figure 9.** SHAPLEY diagram of the top 20 variables included in the prediction model for PV2.

The prediction results are shown in Table 5. Unlike the previous case, the model with neural networks has provided the best results. The NRMSE values are around 3.5% while for the previous case they were around 3%. This loss of precision in the predictions is mainly due to the fact that the

structural component has less significance and that the climatological variables generally have greater variability in the prediction.

Table 5. Comparison of tested models.

Model	Validation RMSE	Forecasting RMSE	Forecasting NRMSE
Ensemble (LSBoost, min.leaf size:1, L.R. 0.042)	1358	987	3.52%
Neural Network (3 Layers of Size [4,1,9], Std.)	1226	917	3.28%
Tree (Non Surrogate)	1494	1097	3.92%

5.3. Discussion

The use of TBATS in the model allows for more accurate predictions, as shown by the nMAPEs of PV1 and PV2. In the PV-System 1 installation this variable is the main variable and by far more important than the rest. In PV-System 2, the influence of TBATS is minor, although it is the main predictor. It was an expected result due to the fact that the panels of both systems have different solar alignments. It can be observed how in the second installation, the model is more affected by the climatological variables, especially those related to ground and air temperature.

This is mainly due to the fact that the arrangement of the plates and the number of plates makes one of the installations more affected by the influence of climatology, while the PV2 installation is more affected by climatology.

In order to establish whether the predictions are within an acceptable level, they have been compared with the literature. Thus, for example, in [17] several papers with nRMSE around 9.5% [37], 5.29% [38] or 1.33% [39] are mentioned. The problem is that there is no single criterion for their calculation. Thus, in [38], the calculation for 24-hour-ahead forecast accuracy is normalised over the range of the prediction, while in [37], it is calculated as we have done, using the installed capacity. In any case, the values obtained indicate that the predictions are accurate.

6. Conclusions

In this paper, machine learning methods have been used to predict the PV generation of the Campus Feuchtwangen with a 48 kWp system. The layout of the strings and the orientation of the panels do not allow the simplified use of traditional prediction methods, which are generally based on the use of the orientation angle and the location of the installation.

Therefore, past observed data and the use of climatological data obtained from the DWD have been used to predict both energy production and irradiance, presenting a hybrid prediction model based on time series and with the help of artificial intelligence. The new method is based on the introduction of TBATS state space models that, through the use of AI with the help of climatological variables, allow predictions to be made.

The results obtained indicate that the prediction accuracy is around 3% in nMAPE, which confirms that the incorporation of TBATS into the model improves its performance. Depending on the setup, different methodologies were required. Thus, for the PV1 installation, ensemble models with LSBoost have provided the best results, with nMAPE of 2.25%, while for PV2 systems neural networks have been used with an accuracy of 3.28%. The fact that system 1 is only west-facing and system 2 has an east, south and west aspect must be taken into account.

However, the results obtained are only framed for the period considered and for one building (Campus Feuchtwangen). In future works we will try to go deeper into the physical elements of the facility to increase the accuracy of the predictions.

Author Contributions Conceptualization, O.T. and M.M.; methodology, O.T. and T.H.; investigation, O.T. and T.H.; writing—original draft preparation, O.T. and T.H.; writing—review and editing, M.M.;

Funding: This research received no external funding.

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

References

1. Statista Installed PV Capacity in Europe Available online: <https://de.statista.com/statistik/daten/studie/303323/umfrage/installierte-pv-leistung-ind-europa/> (accessed on 16 May 2024).
2. Bundesrat Gesetz Zu Sofortmaßnahmen Für Einen Beschleunigten Ausbau Der Erneuerbaren Energien Und Weiteren Maßnahmen Im Stromsektor [Law on Immediate Measures for an Accelerated Expansion of Renewable Energies and Further Measures in the Electricity Sector]. *Drucksache* 315/22 2022.
3. Dr. Simon Philipps *Photovoltaics Report*; 2023;
4. Zhan, S.; Dong, B.; Chong, A. Improving Energy Flexibility and PV Self-Consumption for a Tropical Net Zero Energy Office Building. *Energy Build* **2023**, *278*, 112606, doi:<https://doi.org/10.1016/j.enbuild.2022.112606>.
5. Meyers, C. *Teaching Students to Think Critically*; Jossey-Bass: San Francisco, CA, 1986; ISBN 1-55542-011-7.
6. Mascherbauer, P.; Kranzl, L.; Yu, S.; Haupt, T. Investigating the Impact of Smart Energy Management System on the Residential Electricity Consumption in Austria. *Energy* **2022**, *249*, 123665, doi:<https://doi.org/10.1016/j.energy.2022.123665>.
7. Ciupăgeanu, D.-A.; Lăzăroiu, G.; Barelli, L. Wind Energy Integration: Variability Analysis and Power System Impact Assessment. *Energy* **2019**, *185*, 1183–1196, doi:<https://doi.org/10.1016/j.energy.2019.07.136>.
8. Kazmi, H.; Tao, Z. How Good Are TSO Load and Renewable Generation Forecasts: Learning Curves, Challenges, and the Road Ahead. *Appl Energy* **2022**, *323*, doi:10.1016/j.apenergy.2022.119565.
9. Sobri, S.; Koochi-Kamali, S.; Rahim, N.A. Solar Photovoltaic Generation Forecasting Methods: A Review. *Energy Convers Manag* **2018**, *156*, 459–497.
10. Iheanetu, K.J. Solar Photovoltaic Power Forecasting: A Review. *Sustainability (Switzerland)* **2022**, *14*.
11. Mierzwiak, M.; Kroszczyński, K.; Araszkievicz, A. On Solar Radiation Prediction for the East–Central European Region. *Energies (Basel)* **2022**, *15*, doi:10.3390/en15093153.
12. Lemos-Vinasco, J.; Schledorn, A.; Pourmousavi, S.A.; Guericke, D. *Economic Evaluation of Stochastic Home Energy Management Systems in a Realistic Rolling Horizon Setting*; 2022;
13. Thomas Haupt; Kevin Settler; Johannes Jungwirth; Haresh Vaidya Home-Energy-Management-Systeme (HEMS): Ein Marktüberblick Für Deutschland [Home Energy Management Systems (HEMS): A Market Overview for Germany]. In Proceedings of the PV-Symposium / BIPV-Forum; Kloster Banz, Bad Staffelstein, February 7 2024.
14. Ahmed, M.; Seraj, R.; Islam, S.M. The K-Means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics (Basel)* **2020**, *9*.
15. Antonanzas, J.; Osorio, N.; Escobar, R.; Urraca, R.; Martinez-de-Pison, F.J.; Antonanzas-Torres, F. Review of Photovoltaic Power Forecasting. *Solar Energy* **2016**, *136*, 78–111.
16. Lorenz, E.; Hurka, J.; Heinemann, D.; Beyer, H.G. Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems. *IEEE J Sel Top Appl Earth Obs Remote Sens* **2009**, *2*, 2–10, doi:10.1109/JSTARS.2009.2020300.
17. Wu, Y.K.; Huang, C.L.; Phan, Q.T.; Li, Y.Y. Completed Review of Various Solar Power Forecasting Techniques Considering Different Viewpoints. *Energies (Basel)* **2022**, *15*.
18. Tarmanini, C.; Sarma, N.; Gezegin, C.; Ozgonenel, O. Short Term Load Forecasting Based on ARIMA and ANN Approaches. *Energy Reports* **2023**, *9*, 550–557, doi:<https://doi.org/10.1016/j.egy.2023.01.060>.
19. Kavakci, G.; Cicekdag, B.; Ertekin, S. Time Series Prediction of Solar Power Generation Using Trend Decomposition. *Energy Technology* **2024**, *12*, 2300914, doi:10.1002/ente.202300914.
20. Aziz, S.; Chowdhury, S.A. Early Experience of the Generation Pattern of Grid Connected Solar PV System in Bangladesh: A SARIMA Analysis. In Proceedings of the 2021 6th International Conference on Development in Renewable Energy Technology (ICDRET); 2021; pp. 1–4.
21. Chodakowska, E.; Nazarko, J.; Nazarko, Ł.; Rabayah, H.S.; Abendeh, R.M.; Alawneh, R. ARIMA Models in Solar Radiation Forecasting in Different Geographic Locations. *Energies (Basel)* **2023**, *16*, doi:10.3390/en16135029.
22. Saigustia, C.; Pijarski, P. Time Series Analysis and Forecasting of Solar Generation in Spain Using EXtreme Gradient Boosting: A Machine Learning Approach. *Energies (Basel)* **2023**, *16*, doi:10.3390/en16227618.

23. Rajagukguk, R.A.; Ramadhan, R.A.A.; Lee, H.J. A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power. *Energies (Basel)* **2020**, *13*, doi:10.3390/en13246623.
24. Wang, L.; Mao, M.; Xie, J.; Liao, Z.; Zhang, H.; Li, H. Accurate Solar PV Power Prediction Interval Method Based on Frequency-Domain Decomposition and LSTM Model. *Energy* **2023**, *262*, 125592, doi:https://doi.org/10.1016/j.energy.2022.125592.
25. Scott, C.; Ahsan, M.; Albarbar, A. Machine Learning for Forecasting a Photovoltaic (PV) Generation System. *Energy* **2023**, *278*, 127807, doi:https://doi.org/10.1016/j.energy.2023.127807.
26. Singla, P.; Duhan, M.; Saroha, S. A Comprehensive Review and Analysis of Solar Forecasting Techniques. *Frontiers in Energy* **2022**, *16*, 187–223.
27. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat* **2001**, 1189–1232.
28. Alajmi, M.S.; Almeshal, A.M. Least Squares Boosting Ensemble and Quantum-Behaved Particle Swarm Optimization for Predicting the Surface Roughness in Face Milling Process of Aluminum Material. *Applied Sciences* **2021**, *11*, doi:10.3390/app11052126.
29. De Livera, A.M.; Hyndman, R.J.; Snyder, R.D. Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing. *J Am Stat Assoc* **2011**, *106*, 1513–1527.
30. Hyndman, R.; Athanasopoulos, G.; Bergmeir, C.; Caceres, G.; Chhay, L.; O'Hara-Wild, M.; Petropoulos, F.; Razbash, S.; Wang, E.; Yasmeeen, F. Forecast : Forecasting Functions for Time Series and Linear Models. R package version 2022.
31. Dai, G.; Luo, S.; Chen, H.; Ji, Y. Efficient Method for Photovoltaic Power Generation Forecasting Based on State Space Modeling and BiTCN. *Sensors* **2024**, *24*, 6590, doi:10.3390/s24206590.
32. Nguyen, T.N.; Müsgens, F. What Drives the Accuracy of PV Output Forecasts? *Appl Energy* **2022**, *323*, 119603, doi:https://doi.org/10.1016/j.apenergy.2022.119603.
33. Maitanova, N.; Telle, J.S.; Hanke, B.; Grottkke, M.; Schmidt, T.; Von Maydell, K.; Agert, C. A Machine Learning Approach to Low-Cost Photovoltaic Power Prediction Based on Publicly Available Weather Reports. *Energies (Basel)* **2020**, *13*, doi:10.3390/en13030735.
34. Deutscher Wetterdienst Service Open Data DWD Available online: <https://www.dwd.de/EN/ourservices/opendata/opendata.html> (accessed on 17 May 2024).
35. Peng, H.; Ding, C. Minimum Redundancy and Maximum Relevance Feature Selection and Recent Advances in Cancer Classification. In Proceedings of the Proceedings of the 2005 SIAM International Conference on Data Mining (SDM) Proceedings of the 2005 SIAM International Conference on Data Mining (SDM); Kargupta, H., Srivastava, J., Kamath, C., Goodman, A., Eds.; Society for Industrial and Applied Mathematics: Newport Beach, April 23 2005; Vol. 52, pp. 52–59.
36. Frank, A. Uci Machine Learning Repository. Irvine, ca: University of California, School of Information and Computer Science. <http://archive.ics.uci.edu/ml> **2010**.
37. Wen, Y.; AlHakeem, D.; Mandal, P.; Chakraborty, S.; Wu, Y.-K.; Senjyu, T.; Paudyal, S.; Tseng, T.-L. Performance Evaluation of Probabilistic Methods Based on Bootstrap and Quantile Regression to Quantify PV Power Point Forecast Uncertainty. *IEEE Trans Neural Netw Learn Syst* **2020**, *31*, 1134–1144, doi:10.1109/TNNLS.2019.2918795.
38. Ray, B.; Shah, R.; Islam, M.R.; Islam, S. A New Data Driven Long-Term Solar Yield Analysis Model of Photovoltaic Power Plants. *IEEE Access* **2020**, *8*, 136223–136233, doi:10.1109/ACCESS.2020.3011982.
39. Massaoudi, M.; Chihi, I.; Sidhom, L.; Trabelsi, M.; Refaat, S.S.; Abu-Rub, H.; Oueslati, F.S. An Effective Hybrid NARX-LSTM Model for Point and Interval PV Power Forecasting. *IEEE Access* **2021**, *9*, 36571–36588, doi:10.1109/ACCESS.2021.3062776.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.