Artificial Intelligence for Photovoltaic Power Forecasting in Energy Management

Guillermo Blanco[2], Marco Melgarejo[1(✉)], Alejandro Pérez[1], and Camilo Palazuelos[2]

[1] Centro Tecnológico CTC, Parque Científico y Tecnológico de Cantabria, 39011 Santander, Spain
mmelgarejo@centrotecnologicoctc.com

[2] Universidad de Cantabria Av. de los Castros s, 39005 Santander, n, Spain ,

Abstract. Photovoltaic power forecasting plays a key role in the efficient management of renewable energy systems, particularly in installations operating under zero-injection constraints. This paper presents the development of a forecasting model trained on data collected from a pilot photovoltaic installation located in Santander, Spain. The proposed approach combines historical power generation data with meteorological variables to predict the expected generated power and support more efficient energy management decisions. The paper describes the main stages of the workflow, including data preprocessing, feature engineering, model training, and evaluation. The resulting model achieves an R^2 score of 0.92, showing strong predictive performance and demonstrating the potential of machine learning methods for photovoltaic energy forecasting.

Keywords: Machine Learning · Photovoltaic · Artificial Intelligence · Photovoltaic Forecasting · Time Series Prediction · Renewable Energy · XGBoost · Energy Optimization

1 Introduction

The Spanish photovoltaic market is divided into producers and self-consumers. Producers are those who sell their energy to the electrical grid. In contrast, self-consumers do not sell their production; instead, they use it for their own consumption. In Spain, regulatory constraints have been introduced for certain photovoltaic installations. Some self-consumers, typically those with an installed capacity above 100 kW, are required to comply with the zero-injection requirement, ensuring that no power is fed into the grid.

The zero-injection requirement is a regulation that prohibits the injection of surplus photovoltaic generation into the electrical grid. Therefore, self-consumers must adjust their photovoltaic production so that it does not exceed the electrical demand of the installation [8].

Energy flexibility in this context refers to the ability to adapt electricity demand according to external factors such as tariffs or regulatory constraints.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16766, pp. 173–185, 2027.
https://doi.org/10.1007/978-3-032-30539-8_11

Previous studies have shown that predictive control of photovoltaic production can reduce total energy costs by around 15–25%, while also helping to reduce CO_2 emissions. According to the International Energy Agency, in 2025 solar photovoltaic energy made a significant contribution to reducing global emissions, avoiding on the order of 1.5 Gt of CO_2 anually [4].

Energy optimization can be divided into two types according to the time horizon considered. Short-term optimization focuses on hourly energy production in order to minimize costs and increase efficiency over short periods of time. Long-term optimization aims to determine the optimal dimensions of the installation and its components while minimizing installation costs. To estimate the electrical demand of an installation, multiple time-dependent factors must be considered, as shown in Eq. (1):

$$D(t) = B(t) + P(t) + N(t) + A \tag{1}$$

where $B(t)$ represents the basic operating cycle, $P(t)$ the peak demands that introduce maximum-demand events without overlap, $N(t)$ the stochastic noise fluctuations, and A the baseline consumption, ensuring realistic conditions [3].

Spain increased its installed photovoltaic capacity from 11 GW in 2020 to more than 32 GW by the end of 2024 [11]. As a result, photovoltaic power has become one of the country's main electricity generation technologies, even surpassing wind power. At present, photovoltaic energy accounts for 17% of total national electricity generation, making Spain one of the leading countries in solar energy in Europe.

At the international level, photovoltaic energy represents 46% of renewable capacity additions, and the International Energy Agency (IEA) estimates around 3.6 TW of new renewable energy capacity before 2030, of which approximately 80% is expected to come from photovoltaic sources. As 2030 approaches, the photovoltaic industry is expected to strengthen its role as the leading renewable energy technology [4].

Compared with other renewable energy sources, photovoltaic power generally shows a favorable environmental profile. Emissions associated with its manufacturing process are lower than those of other energy sources, such as wind power.

Despite these achievements, the photovoltaic industry still faces important challenges, one of the main ones being the price of electricity. As this technological sector has expanded rapidly, internal competition has increased significantly, leading to a decline in the selling price of solar energy. This situation has directly affected the profitability of photovoltaic installations [6]. If corrective measures are not taken soon, the situation is expected to worsen by the end of 2026.

The goal of this work is to develop a prediction model that primarily helps installations operating under the zero-injection rule to manage their demand through forecasting. In addition, the proposed model can also be useful for intelligent energy management systems in general. Its purpose is to predict the expected power generated by the photovoltaic installation.

2 Related work

Photovoltaic power forecasting has gained prominence amid Spain's rapid solar expansion. Early models employed statistical data but limited by non-linear weather dependencies.

Machine learning advancements shifted toward tree-based ensembles. *Random forests* and *Extra Trees* improved handling of meteorological inputs like irradiance and temperature, achieving robust short-term predictions on UCI datasets. *Gradient boosting*, particularly *XGBoost*, outperformed these via optimized regularization and feature importance, with R^2 scores exceeding 0.90 in benchmarks [12].

As shown in Table 1, in Spain tools such as Shirenda [15] leverage machine learning on REE data (2015–2020) for nationwide yield estimation, supporting zero-injection compliance under RD 244/2019. Recent studies forecast rooftop PV with tuned *XGBoost* models and emphasize real-time AEMET integration, but they still overlook edge deployment [5].

These contributions enable demand optimization, saving 15-25% costs, but gaps persist in low-resource hardware forecasting and hybrid *skforecast* [13] with *XGBoost* for real pilot data [2], motivating our methodology.

Table 1. Comparison of related forecasting studies

Work	Method	Horizon	Key Metric	Limitation
Chen (2016) [5]	*XGBoost*	Short-term	R^2: 0.92	Offline focus
Zhang (2023) [9]	*LSTM-AdaBoost*	Short	RMSE: 0.043	Synthetic data
Shirenda_PV (2025) [10]	ML on REE	Long-term	N/A (yield est.)	No real-time

This work further investigates the application of machine learning models, especially *XGBoost*, to photovoltaic power forecasting. In particular, it analyzes how these models capture non-linear relationships between meteorological variables and photovoltaic generation, thereby improving forecasting accuracy and supporting more informed operational decisions in solar installations.

3 Methodology

3.1 Data source

The data were collected from a real photovoltaic installation located on the roof of CTC in Santander, Spain, over the course of one year. The power generation and meteorological datasets were merged and temporally synchronized. The same procedure was also applied to another photovoltaic installation in Lorca, Murcia, Spain. However, the present study focuses primarily on the Santander case, while the Lorca installation is used only as a complementary example in the practical applications section.

3.2 Dataset preprocessing

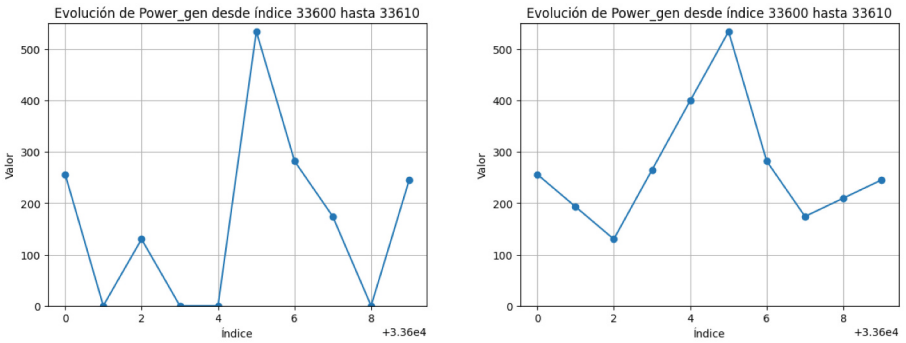
When developing a forecasting model, the first step is to ensure that the dataset used for training is properly prepared and processed. The data must reflect the conditions of the target installation so that the model can learn representative local patterns. This has a direct impact on forecasting accuracy.

Once the dataset was assembled, an additional analysis was performed to detect and remove outliers that could negatively affect model accuracy.

The 25th percentile is the value below which 25% of the observations fall, while the 75th percentile is the value below which 75% of the observations fall. The interquartile range (IQR) is defined as the difference between these two percentiles [14]. Outlier removal results in a moderate reduction in the dataset size, which is acceptable, although it also creates gaps in the time series.

Some daytime samples recorded between sunrise and sunset showed zero power generation values. These values were considered physically implausible in the present modelling stage, since zero-injection operation was not included during model training. Instead, the zero-injection constraint is addressed in a subsequent optimization layer, further discussed in Section 4.3, where equipment demand can be scheduled according to the forecasted photovoltaic production in order to avoid surplus injection into the grid.

Consequently, these anomalous zero-valued samples were attributed to possible data logging errors and corrected by linear interpolation. Retaining them would have introduced artificial discontinuities in the generation signal and could have increased the prediction error. Fig. 1 illustrates this preprocessing step: the left panel shows the original signal with anomalous zero-valued samples, whereas the right panel shows the corrected signal after interpolation.



(a) Before interpolation (b) After interpolation

Fig. 1. Effect of linear interpolation on the time series. The left panel shows the original signal with anomalous zero-valued samples, whereas the right panel shows the corrected signal after interpolation.

Finally, some rows contained NaN values in one or more columns, and these rows were removed. After completing the preprocessing stage, the final dataset was obtained, as summarized in Table 2.

Table 2. Dataset descriptive statistics

Statistic	Power_gen(W)	temperature(C^{O})	windSpeed(km/h)	solarRad(W/m^2)	moist(%)	rain(mm)	pressure(hPa)
count	41821.00	41821.00	41821.00	41821.00	41821.00	41821.00	41821.00
mean	86.63	16.32	6.87	101.45	79.29	0.53	30.13
std	156.52	4.22	5.99	162.07	10.71	1.24	0.20
min	0.00	4.44	0.00	0.00	46.00	0.00	29.60
25%	0.00	12.90	3.22	0.00	73.00	0.00	30.00
50%	0.00	16.30	4.83	0.00	82.00	0.00	30.20
75%	94.00	19.80	9.66	153.00	88.00	0.00	30.30
max	780.00	30.60	27.40	631.00	96.00	5.40	30.70

Since the variables in the dataset are measured on different scales, feature scaling was applied. This step must be performed only after splitting the data into training and test subsets; otherwise, information from the training data would leak into the test set. This issue, known as data leakage, can lead to overly optimistic performance estimates.

Scaling maps the variables to a common range, typically between 0 and 1. In this work, this was performed using *MinMaxScaler*.

After preprocessing, the model development and training stage was carried out.

3.3 Model development

Time series are data ordered chronologically. Forecasting consists of predicting future values of a time series by exploiting its past behavior in an autoregressive manner, optionally combined with external variables through tools such as *skforecast* [2]. In this work, exogenous variables such as solar irradiance, temperature, and wind speed were used together with endogenous features derived from previous observations, including mean, median, maximum, and minimum values. Unlike standard regression settings, time plays a central role in time-series forecasting; however, the problem can still be addressed using regression-based models such as *Gradient Boosting* or *Random Forest*.

A forecasting model based on *XGBoost* was trained. The optimal hyperparameters were selected by means of a *GridSearchCV* procedure. After training, feature importance analysis was used to assess the relative contribution of each predictor to the model output, as shown in Fig. 2.

As expected in a tree-based time-series forecasting model, the most influential predictors are the nearest lagged values, namely *Lag_1* and *Lag_2*, which represent the observations recorded 5 and 10 minutes earlier. However, this should not overshadow the relevance of the exogenous variables.

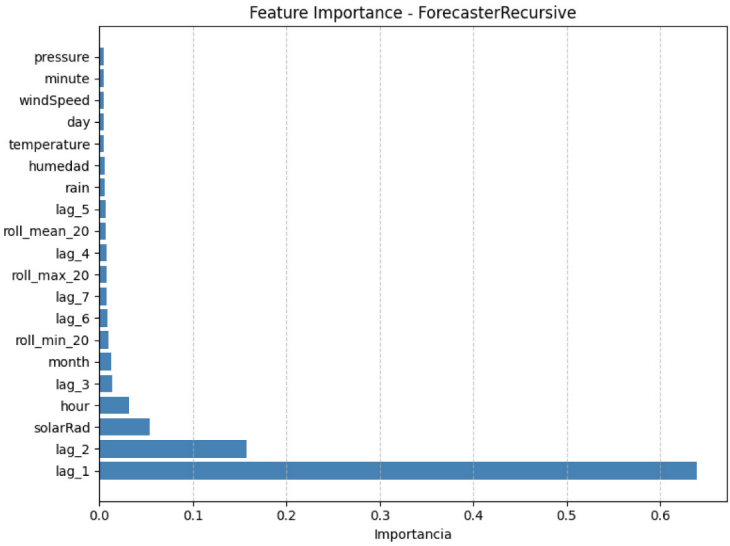


Fig. 2. Feature importance.

In particular, solar irradiance ranks as the third most important feature overall and the most important non-autoregressive predictor, which is consistent with the physical behavior of photovoltaic generation. The hour of the day also appears among the most relevant variables, further supporting the temporal and physical consistency of the model.

In addition, summary statistics derived from previous samples, such as rolling minima and maxima, are also present in the top-ranked features, indicating that local temporal context contributes useful complementary information. Therefore, although the lagged variables dominate the ranking, relying only on *Lag_ 1* and *Lag_ 2* would be an oversimplification, as the remaining predictors help reduce forecasting error and improve the robustness of the model.

It is also essential to verify that the model is not overfitting the training data. Overfitting occurs when the model memorizes the training set instead of learning generalizable predictive patterns, which typically leads to excellent performance on seen data but poor results on unseen data. A common way to assess this behavior is by comparing the evolution of the training and validation errors across boosting iterations. If the validation error increases substantially while the training error continues to decrease, the model can be considered overfitted. Figure 3 shows that the training and validation curves evolve consistently, indicating that the model does not exhibit significant overfitting.

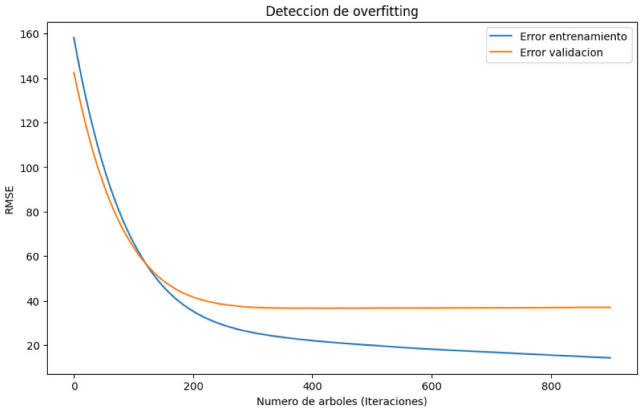


Fig. 3. Training and validation error across iterations.

4 Results

4.1 Prediction metrics

After generating predictions on the test set, the performance metrics reported in Table 3 were obtained.

Table 3. Prediction metrics

MAE	16.52 W
RMSE	40.39 W
R^2	0.92

The results indicate that the proposed model achieves a satisfactory level of predictive accuracy on the test set. This is especially relevant considering that the case study corresponds to a photovoltaic installation located in Cantabria, where frequent meteorological fluctuations and variable cloud cover lead to rapid changes in solar irradiance. Therefore, the forecasting task is more challenging than in scenarios with more stable weather conditions.

In particular, the R^2 value shows that the model explains a large proportion of the variance in the target variable, while the MAE and RMSE values indicate that the prediction error remains moderate in relation to the typical range of photovoltaic power generation observed in the dataset. Overall, these results suggest that the model is robust and suitable for short-term photovoltaic forecasting in a real operating scenario with significant weather variability.

4.2 Graphical representation

A useful way to assess the model performance is to compare the predicted values with the observed ones. Fig. 4 presents this comparison over the full test set, where the predictions are shown in red and the actual values in blue.

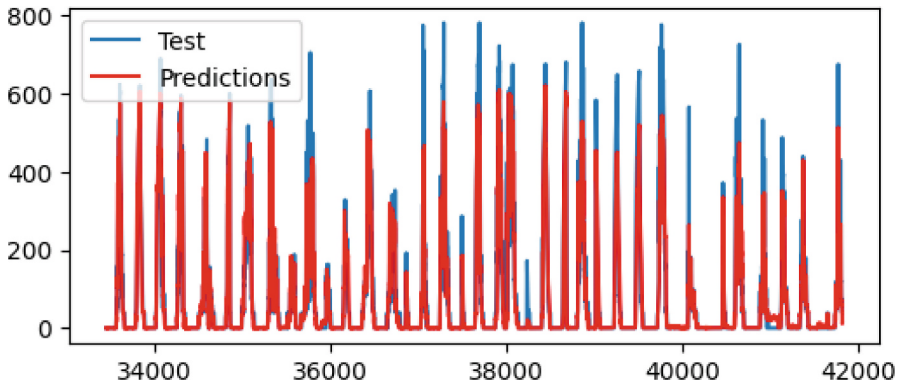


Fig. 4. Prediction through the whole dataset.

A more detailed evaluation can be performed over shorter forecasting horizons, such as the next 24 or 48 hours. Fig. 5 shows the 48-hour forecasting example, again using red for predicted values and blue for observed values. The similarity between both curves indicates that the model is able to capture the overall trend of the power generated by the installation.

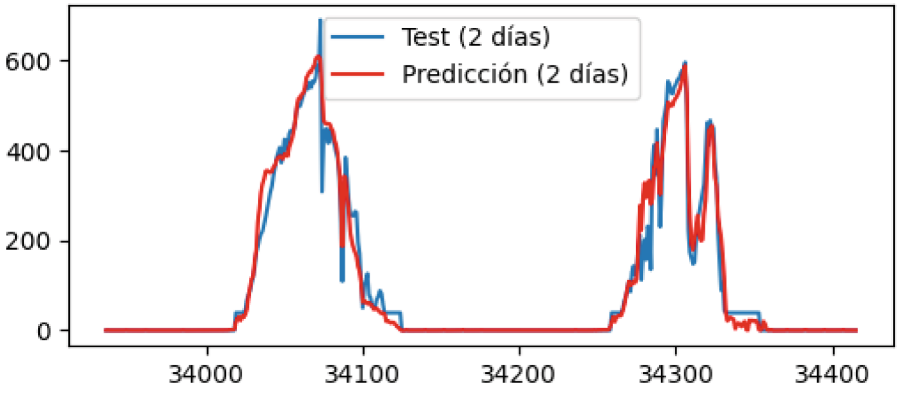


Fig. 5. Prediction of the next 48 hours.

Another way to evaluate model performance is to compare predicted and observed values in a scatter plot rather than along the time axis. Fig. 6 presents

this comparison, where the closer the points are to the dashed red diagonal, the better the agreement between predictions and observations. The predicted values are shown on the Y-axis and the observed values on the X-axis.

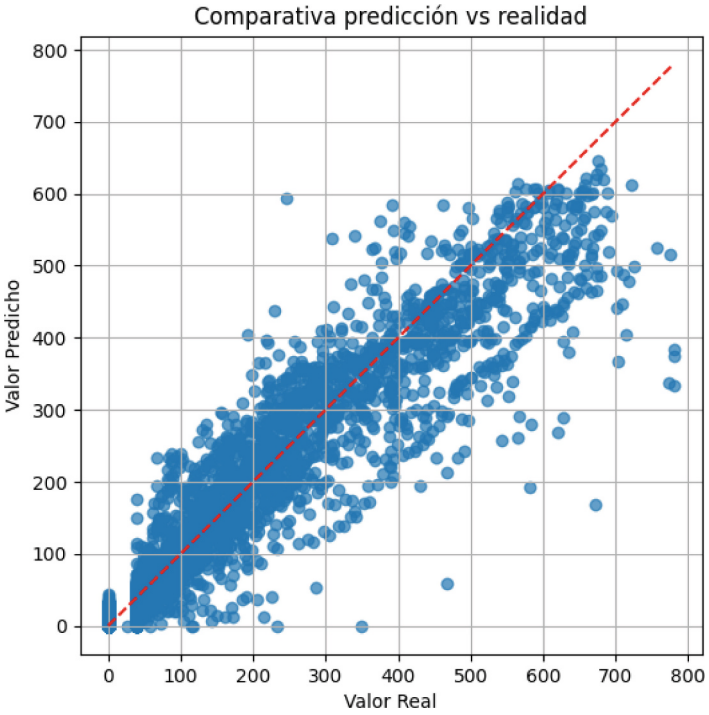


Fig. 6. Predicted vs. real values for the full test set.

4.3 Practical applications

Although the main case study of this paper is the Santander installation, the Lorca installation is included as a complementary example to illustrate the practical value of the proposed forecasting approach for industrial and commercial energy management.

From an operational perspective, the main usefulness of the forecasting model lies not only in point predictions, but also in deriving aggregated indicators of photovoltaic availability over decision-relevant time windows. In practice, daily forecasts can be summarized as average production levels for morning and afternoon shifts, as shown in Fig. 7. These indicators provide a simple representation of the photovoltaic energy expected to be available in each shift, together with the overall daily average.

Such information can be directly used to support load-balancing strategies. In industrial environments, several loads may be flexible, including pumping systems, compressed-air equipment, thermal processes, charging systems, or non-critical production tasks. If higher photovoltaic generation is forecast for a given shift, these loads can be shifted or intensified during that period to increase self-consumption. In this way, the forecast becomes a decision-support tool for adapting demand to the expected renewable generation profile.

This application is especially relevant under zero-injection constraints. In a strict zero-feed-in regime, all photovoltaic energy must be consumed locally at each time step; otherwise, export-limiting systems enforce curtailment or disconnection of the photovoltaic plant. Therefore, relying only on anti-export hardware may ensure compliance, but it can also waste usable renewable energy and increase implementation costs. By contrast, forecast-based demand adaptation allows a larger share of the locally generated energy to be used before curtailment becomes necessary. Both strategies can therefore be combined: export-limiting hardware can remain as a safety layer, while predictive energy management acts as the main mechanism to maximize photovoltaic utilization.

Once these forecasted averages are available, different automatic optimization strategies can be applied to schedule flexible loads. A first option is a heuristic or metaheuristic scheduler that assigns shiftable loads to the periods with the highest predicted photovoltaic availability. Previous works [1,7] have explored particle swarm optimization and other multi-objective approaches to improve self-consumption and reduce grid dependence. Heuristic multi-agent coordination schemes have also been proposed for scheduling shiftable devices, offering high-quality solutions with lower computational cost. In addition, robust multi-objective formulations can explicitly account for photovoltaic uncertainty while optimizing economic and environmental criteria. All these approaches are compatible with the present application, where the goal is to synchronize flexible demand with forecasted photovoltaic production.

Forecast uncertainty must also be considered, since it increases as the prediction horizon moves further away from the present time. For this reason, indicators such as those in Fig. 7 can be complemented with error bars or confidence intervals, allowing the operator or the energy management system to distinguish between more and less reliable forecast horizons. This enables a hierarchical planning strategy: day-ahead forecasts can define an initial schedule, while intraday updates progressively refine it as new meteorological information and more accurate short-term predictions become available. In other words, the load-balancing plan can be continuously corrected in a rolling-horizon manner as the target operating period approaches.

Therefore, the practical contribution of the proposed forecasting model goes beyond estimating photovoltaic generation. Its main added value lies in enabling predictive load scheduling and energy optimization strategies that increase self-consumption, reduce unnecessary curtailment, and mitigate the operational drawbacks associated with zero-injection requirements.

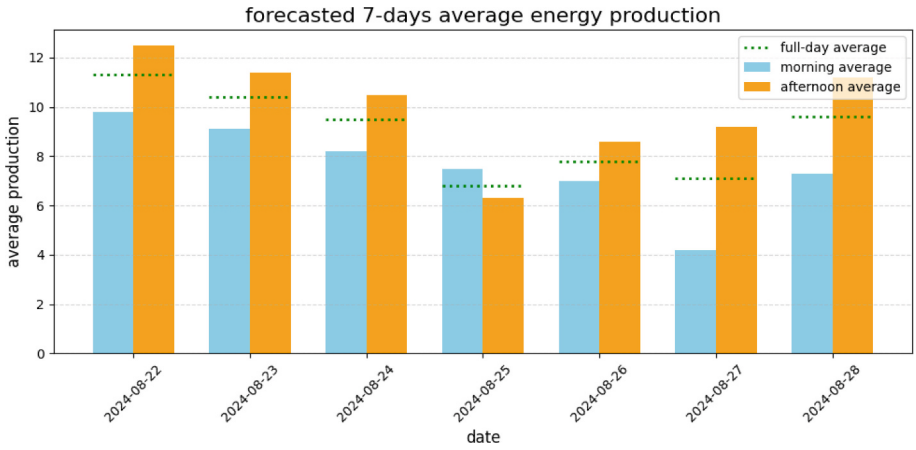


Fig. 7. Average production profile for the Lorca installation, included as a complementary application example.

5 Conclusions and future steps

The development of a forecasting model is only meaningful when it serves a practical purpose. As stated earlier, the main objective of this work was to support the electrical management of an installation operating under the zero-injection rule. Thanks to the proposed model, it is possible to estimate the amount of energy that will be generated and adjust the electrical demand accordingly.

Compared to the state of the art, the proposed methodology demonstrates competitive performance while addressing relevant gaps in the literature. The model achieved an R^2 of approximately 0.92, which is in line with previously reported short-term forecasting results. More importantly, whereas several recent studies rely on synthetic data or do not consider real-time edge deployment, the proposed approach combines skforecast and XGBoost on real pilot data. This strengthens the practical relevance of the methodology and supports its potential for low-resource hardware forecasting.

As future work, the model could be extended by incorporating external meteorological forecasts, such as those provided by AEMET, as exogenous variables during prediction, rather than relying only on historical local measurements from the on-site weather station. Although these forecasts are usually generated on spatial grids, they can be adapted to the exact installation location through statistical downscaling or simpler geographical interpolation methods.

In addition, the same approach could provide access to a broader historical meteorological record, enabling the construction of a more representative training dataset. This would improve the robustness of the model and further support its deployment in practical photovoltaic energy management applications.

Data Availibility. The data supporting the results of this study are available from the corresponding author upon request.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Funding. The presented work is part of the project titled *Posicionamiento Estratégico de Especialización Inteligente y Sostenible para las Factorías del FUTuro de CANtabria (FUTCAN)*, which is being carried out at the Centro Tecnológico CTC. This entity has received, for the FUTCAN project, co-financed assistance from the European Regional Development Fund (ERDF) through the 2021–2027 Cantabria Operational Program, under the funding line *Grants for research projects with high industrial potential of excellence technology agents for industrial competitiveness (TCNIC).*











References

1. Abdalla, M.A.A., Min, W., Bing, W., Ishag, A.M., Saleh, B.: Double-layer home energy management strategy for increasing pv self-consumption and cost reduction through appliances scheduling, ev, and storage. Energy Rep. **10**, 3494–3518 (2023). https://doi.org/10.1016/j.egyr.2023.10.019
2. Amat Rodrigo, J., Escobar Ortiz, J.: skforecast (Sep 2023). https://doi.org/10.5281/ZENODO.8382788, https://doi.org/10.5281/ZENODO.8382788, accessed on: May 8, 2026
3. Castán-Lascorz, M.A., Mselle, B.D., Zabaleta, L.L., León, M., Arroyo, J.: Energy flexibility optimization in industry: A hybrid approach with synthetic data evaluation. In: 2025 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC). pp. 1–10. Valencia, Spain (Jun 2025). https://doi.org/10.1109/ICE/ITMC65658.2025.11106527
4. IEA: Global energy review 2026: Co2 emissions. Tech. rep., International Energy Agency, Paris (2026), https://www.iea.org/reports/global-energy-review-2026/co2-emissions, accessed on: May 8, 2026. Licence: CC BY 4.0
5. IJPEDS: Article download 24057/15126 (2026), https://ijpeds.iaescore.com/index.php/IJPEDS/article/download/24057/15126, accessed on: March 6, 2026
6. International Renewable Energy Agency: Renewable power generation costs in 2023. Tech. rep., International Renewable Energy Agency, Abu Dhabi (Sep 2024), https://www.irena.org/Publications/2024/Sep/Renewable-Power-Generation-Costs-in-2023, accessed on: May 6, 2026
7. Luz, G.P., Brito, M., Sousa, J., Vieira, S.: Coordinating shiftable loads for collective photovoltaic self-consumption: A multi-agent approach. Energy **229**, 120573 (2021). https://doi.org/10.1016/j.energy.2021.120573, https://www.sciencedirect.com/science/article/pii/S0360544221008227
8. Ministerio para la Transición Ecológica: Real decreto 244/2019, de 5 de abril, por el que se regulan las condiciones administrativas, técnicas y económicas del autoconsumo de energía eléctrica. BOE-A-2019-5089, pp. 35674–35719 (2019), https://www.boe.es/eli/es/rd/2019/04/05/244, accessed on: February 23, 2026

9. PMC: Article pmc11002275 (2026), https://pmc.ncbi.nlm.nih.gov/articles/PMC11002275/, accessed on: March 6, 2026
10. PV Magazine: Nueva base de datos abierta para estimar la producción fotovoltaica en españa (2025), https://www.pv-magazine.es/2025/02/03/nueva-base-de-datos-abierta-para-estimar-la-produccion-fotovoltaica-en-espana/, accessed on: March 6, 2026
11. Red Eléctrica de España: Informe de energías renovables: potencia instalada, solar fotovoltaica (2025), https://www.sistemaelectrico-ree.es/es/informe-de-energias-renovables/sol/potencia-instalada/solar-fotovoltaica-solpotencia, information prepared using provisional data up to January 2026. Accessed on: May 6, 2026
12. ScienceDirect: Article s2352484725001490 (2026), https://www.sciencedirect.com/science/article/pii/S2352484725001490, accessed on: March 6, 2026
13. skforecast: skforecast documentation (v0.20.1) (2026), https://skforecast.org/0.20.1/index.html, accessed on: March 7, 2026
14. Statistics How To: Interquartile range (iqr): What it is and how to find it (2025), https://www.statisticshowto.com/probability-and-statistics/interquartile-range/, accessed on: December 1, 2025
15. Sánchez-Hernández, G., Jiménez-Garrote, A., López-Cuesta, M., Galván, I.M., Aler, R., Pozo-Vázquez, D.: A novel method for modeling renewable power production using era5: Spanish solar pv energy. Renewable Energy **240**, 122120 (2025). https://doi.org/10.1016/j.renene.2024.122120, https://doi.org/10.1016/j.renene.2024.122120