POLITECNICO DI TORINO
Repository ISTITUZIONALE

Simulation-Based Photovoltaic Production Data for Training Data-Driven PV Power Forecasting Models

Original
Simulation-Based Photovoltaic Production Data for Training Data-Driven PV Power Forecasting Models / Gallo, R., Castangia, M., Canfora, A., Macii, A., Aliberti, A., Patti, E.. - (2026), pp. 709-714. (8th Global Power, Energy and Communication Conference (GPECOM) Naples (ITA) 03-05 June 2026) [10.1109/gpecom70462.2026.11578470].

Availability:
This version is available at: 11583/3012653 since: 2026-07-03T10:42:35Z

Publisher:
IEEE

Published
DOI:10.1109/gpecom70462.2026.11578470

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

14 November 2026

Simulation-Based Photovoltaic Production Data for Training Data-Driven PV Power Forecasting Models

Raimondo Gallo*, Marco Castangia*, Andrea Canfora†, Alberto Macii*, Alessandro Aliberti* and Edoardo Patti*
*Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy. Email: name.surname@polito.it
†FiberCop, Rome, Italy. Email: name.surname@fibercop.com

***Abstract*—Accurate day-ahead photovoltaic (PV) power forecasting is essential for grid operation and energy market participation. Nevertheless, PV plants frequently do not have enough historical production data to develop accurate forecasting models. In situations where real measurements are scarce, this work investigates whether physics-based simulated data can support data-driven forecasting models. We propose a methodology that leverages a subset of numerical weather prediction (NWP) variables, provided by the ICON-EU model that predicts day-ahead DC production for a target PV plant. Specifically, a 3D convolutional neural network (3D-CNN) is trained to map spatial-temporal NWP forecasts to day-ahead PV power production. To overcome the issue of data scarcity, we use a physics-based model implemented with `pvlib` to simulate the DC power production for the target PV plant. We compare two experimental settings to assess the use of simulated data as an alternative to scarce real measurements, one using simulated production data and another using the limited available real measurements. Results obtained on a real-world demo site show that the model trained on simulated data achieves a lower root mean squared error (RMSE) than the model trained solely on measured production, highlighting the potential of combining physical modeling and machine learning to improve PV forecasting in data-scarce scenarios such as newly deployed solar plants.**

***Index Terms*—PV power forecasting, Day-ahead forecasting, Synthetic data, Numerical weather prediction, Deep learning**

I. Introduction

Although the amount of carbon dioxide emitted per unit of electricity produced worldwide has declined over the past decade, emissions from electricity generation still account for 40% of the worldwide emissions [1]. In this scenario, photovoltaic (PV) systems continue to expand with their generation expected to overtake wind and nuclear by 2026.

However, renewable energy production, from wind and solar particularly, is highly variable due to weather phenomena hard to predict. This poses challenges to the power grid flexibility, and and accurate power production forecasting becomes crucial. Indeed, grid stability, energy trading, and the correct management of distributed energy resources all depend on the quality of PV forecasts.

In recent years, methods that rely on machine learning and deep learning have shown high effectiveness in PV forecasting tasks. Despite these advances, the performance of data-driven models strongly depends on the availability of historical production measurements. This creates a major challenge for newly installed PV plants, where there may only be a short period of operational data available. In such limited data scenarios, training robust forecasting models becomes challenging due to the limited amount of labeled data and the risk of overfitting.

To overcome this limitation, one approach could be to use synthetic data produced by physics-based simulations. Physical PV models can simulate how much power a particular PV asset can produce based on identified characteristics of both the facility and the environment at its location; thus enabling the generation of a set of training samples without an extensive availability of historical measurements. These simulations could provide a valuable source of training data to data-driven forecasting models.

In this work, we propose a data-driven methodology for day-ahead PV power forecasting. The approach is based on a 3D convolutional neural network (3D-CNN) that maps spatio-temporal numerical weather prediction (NWP) forecasts to target PV power production. Fig. 1 shows the pipeline of the proposed methodology, including the required data, forecasting modeling and evaluation setup. To address the limited availability of measured data, we leverage a physics-based model implemented with `pvlib` to generate simulated direct current (DC) power production. In this context, we evaluate the approach under two experimental settings, SIM and REAL, which differ in the source of the PV training targets, while sharing the same NWP inputs and real-world test set. Both settings are evaluated on the same real-world test set to ensure a fair comparison. The proposed methodology is evaluated on a real-world demo PV plant located in Mathi, Italy. Results show that the SIM setting leads to improved forecasting performance on measured data compared to the REAL setting, reducing the root mean squared error (RMSE) by more than 250 W. These findings highlight the potential of combining physical modeling and machine learning for PV forecasting in scenarios with limited availability of historical data.

The rest of the paper is organized as follows: Section II reviews related work on PV power forecasting strategies and results. Section III describes the available datasets, PV power simulation models, and the proposed forecasting methodology. Section IV discusses the experimental results, comparing forecasting performance of the models trained exclusively with synthetic and real data respectively. Finally, Section V outlines

This work is partially supported by *SINEGRA - Sistema INtelligente per l'Efficienza e la Gestione delle Reti Avanzate* (MI_DDR_00407), an Italian National project funded by the Ministry of the Environment and Energy Security (MASE) under the Mission Innovation 2.0 initiative.

directions for future research.

II. Related Works

The accurate forecasting of PV power output is essential for power system operation and energy market participation. The approaches to forecast PV power can be categorized into physical and statistical [2]. The physical methods, based on technical specifications of the PV asset, utilize theoretical models of the PV system to estimate the power generation. While statistical models rely on historical data to estimate the PV power. Finally, hybrid approaches combine the theoretical knowledge drawn from physical models with data-driven methods.

In the literature, statistical models are widely used to forecast PV power, as machine learning and deep learning have gained significant attention thanks to their capability at learning complex nonlinear relationships between weather conditions and PV power production. However, based on the forecasting horizon of interest, the choice of the method can change substantially. For instance, satellite and sky images or ground measurements are more suited for short-term forecasts up to 6 hours ahead; while, with NWP models data it is possible to forecast up to days ahead [3].

The joint need for accurate day-ahead power forecasting and the increasing accuracy of statistical methods has driven the development of forecasting methods that leverage NWP forecasts as input predictors. Several studies have demonstrated that combining historical production measurements with NWP data can improve day-ahead forecasting performance [4], [5]. Architectures such as convolutional neural networks, recurrent neural networks, and hybrid models have been successfully applied to solar power prediction tasks [6], [7]. In particular, convolutional architectures can effectively process spatially distributed NWP fields, allowing forecasting models to exploit the spatial structure of atmospheric variables surrounding the PV plant.

Despite these advances, a key challenge remains the limited availability of historical production data for newly installed PV systems. To address this issue, recent works have explored the use of synthetic or simulated data generated through physics-based PV models [8], [9]. By combining plant characteristics with meteorological information, these models can approximate PV production and generate additional training data when real measurements are scarce.

Based on this idea, the present work proposes a day-ahead PV power forecasting methodology that exploits a subset of variables derived from ICON-EU NWP forecasts [10], [11], in addition to PV production data as target variables, while investigating the use of a physics-based PV simulator, developed using `pvlib`, as an alternative source of PV production data. The produced day-ahead forecasts are characterized by a temporal resolution of 15-minutes. As illustrated in Fig. 1, the proposed methodology is evaluated under two experimental settings described in Section III-E. We evaluate the methodology on a real-world demo PV system located in Mathi, northern Italy. In contrast to approaches that rely solely on measured production data, we analyze whether a model trained on simulated data can generalize effectively to real-world observations.

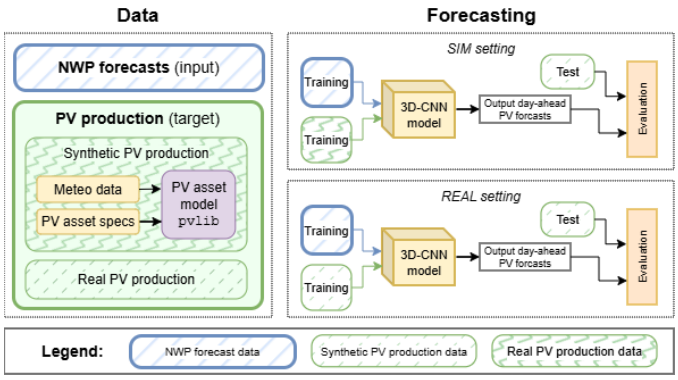


Fig. 1. The pipeline of the proposed methodology and experimental setups.

III. Methodology

In this section, we describe the proposed day-ahead PV power forecasting methodology, detailing the required inputs, processing steps, and model architecture. We also present the physics-based simulation method used to generate the synthetic PV production dataset.

A. NWP forecasts

The NWP models provide forecast data for a vast collection of variables at atmospheric and ground level. Referring to Fig. 1, NWP forecast are the input to the proposed forecasting model. Specifically, we acquired NWP data from the ICON-EU model, the operational regional forecasting system developed by the German Weather Service (DWD) [10], [11]. ICON-EU provides meteorological forecasts over Europe with a resolution of approximately 6.5 km and hourly forecast steps (lead times). For this study, we just considered the forecasts produced by the 00 UTC run and extract the lead times corresponding to the day-ahead forecasting horizon. Specifically, we select lead times from +21 h to +48 h to match the production period of the following day.

The NWP fields (the spatial grids representing the selected meteorological variables) are spatially cropped over the region centered around the location of the real-world PV plant to analyze, corresponding to an area of approximately 90,000 km^2.

We considered a subset of the available variables relevant for solar power generation, specifically surface downward shortwave radiation flux and near-surface temperature. These fields for the selected NWP lead times are stacked as multi-channel spatial tensors and used as input to the proposed forecasting model. Fig. 2 shows a sample image variable of temperature at two meters above the ground from the selected NWP model, including the analyzed cropped region.

We retrieved the historical operational NWP data used in this study from a publicly available repository. As the dataset is maintained by a third-party provider, occasional

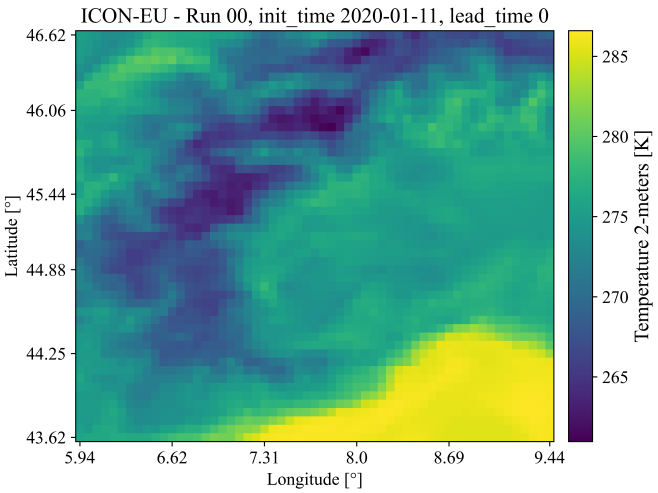


Fig. 2. A sample image variable of temperature at two meters above the ground from ICON-EU for the 00 run and a given init_time and lead_time.

interruptions in data availability may occur due to external service limitations.

B. PV production data

Through the inverter monitoring systems, it is possible to extract operational data from a PV system. These data generally include electrical measurements at both the alternate current (AC) and DC sides, such as voltage and current, as well as additional variables describing the system's operating conditions, for instance inverter temperature, operational status, and cumulative energy production. Due to maintenance or communication issues in the data acquisition system, the PV dataset may contain missing or corrupted measurements.

The target for the forecasting model is defined as the total DC power generated by the PV system. When multiple PV strings are available, the instantaneous DC power is computed as the sum of the contributions from each string. Eq. 1, describes how DC power can be computed:

$$P_{\mathrm{DC}}(t) = \sum_i P_{\mathrm{PV}i}(t) = \sum_i V_{\mathrm{PV}i}(t) \cdot I_{\mathrm{PV}i}(t) \tag{1}$$

where $V_{\mathrm{PV}i}(t)$ and $I_{\mathrm{PV}i}(t)$ denote the voltage and current measured for the i-th PV string at time t.

We adopt DC power as the prediction target rather than the AC power injected into the grid, allowing the model to capture the actual energy conversion at the module level, independently of inverter efficiency and losses on the grid side, guaranteeing consistency with the physics-based simulated PV power production. The temporal resolution of the PV dataset matches the 15-minute granularity of the day-ahead forecasting task.

C. Synthetic PV production

We generate a synthetic dataset of PV power production using a physics-based model implemented with `pvlib`, allowing the generation of additional training samples to complement the limited real measurements. Table I lists the PV module parameters required by the simulator and their unit. Such information can be acquired from datasheets of the PV asset of interest.

Besides technical PV module information, the simulation also requires environmental inputs, such as global horizontal irradiance (GHI) and its components direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI), ambient temperature, and wind speed. These inputs, together with module-specific parameters, are used to compute the effective irradiance and cell temperature, which determine the simulated DC power at the module level. The outputs of multiple modules and strings are then aggregated to obtain the total DC power of the simulated PV system.

Through the PV asset simulation model, it is possible to generate a synthetic DC power dataset for arbitrary historical periods by feeding the corresponding meteorological measurements. This approach allows the creation of long-term and accurate datasets enriching the amount of training samples for data-driven PV forecasting models, whenever on-site PV measurements are not available.

D. Forecasting Model

To predict day-ahead PV power production, we employ a 3D-CNN deep learning model. This architecture is designed to exploit the spatio-temporal structure of NWP data, which provide meteorological variables over a spatial grid and across multiple forecast lead times. The input to the model consists of a sequence of NWP fields defined over a geographical grid surrounding the PV plant. This results in a four-dimensional input tensor with shape *(C, T, H, W)*, corresponding to NWP fields *(C)*, lead time *(T)*, latitude *(H)*, and longitude *(W)*, respectively.

Referring to Table II, which reports in detail the 3D-CNN's layers and their parameters, the model processes the inputs through a stack of 3D convolutional layers, which jointly learn spatial and temporal patterns from the NWP fields while progressively reducing the spatial resolution of the feature maps. Then, the encoded feature maps are flattened and passed to a set of fully connected layers with nonlinear activation functions and dropout regularization. These layers map the

TABLE I
TECHNICAL PARAMETERS OF THE PV PLANT MODULE REQUIRED BY THE PHYSICS-BASED PV SIMULATOR.

Parameter	Real-world demo PV plant	Unit
Tilt angle	10	°
Orientation azimuth angle	210	°
Cell type	Multicrystalline silicon	-
Number of cells in series	60	-
Nominal DC power	245	W
Open-circuit voltage	37.10	V
Short-circuit current	8.54	A
Maximum power point voltage	30.65	V
Maximum power point current	8.02	A
Power temperature coefficient	-0.46	%/°C

learned spatio-temporal features to the final prediction of PV power production, characterized by a temporal granularity of 15 minutes.

Before training, we standardized the NWP input tensors using z-score normalization. We computed the mean and standard deviation for each selected NWP field in the training set by averaging across the temporal and spatial dimensions, as described in Eq. 2:

$$x' = \frac{x - \mu}{\sigma} \tag{2}$$

where μ and σ denote the mean and standard deviation computed from the training set then used to normalize the validation and test data.

E. Real-world demo PV site & experimental settings

a) Experimental design: Referring to Fig. 1, we define two experimental settings, namely SIM and REAL, to investigate the effectiveness of simulated data for training the forecasting 3D-CNN model. In both settings, the model receives as input the selected variables from the ICON-EU NWP forecasts, while differing in the source of the PV power targets used during training. In the SIM setting, the model is trained using simulated PV production data and evaluated on real observations. Instead, in the REAL setting, the model is both trained and evaluated using only measured PV production data as target. This comparison enables assessing whether simulated data can be employed as a reliable substitute for real measurements.

b) Real-world demo PV site: The considered real-world PV plant is located in Mathi, in the Piedmont region of northern Italy. It is a 11.27 kW DC system whose specifications are reported in Table I. We retrieved operational data from the inverter monitoring system of the PV plant covering the period from July 2024 to December 2025, with a temporal granularity of 15 minutes. Using the measured currents and voltages of the PV strings, we computed the total generated DC power of the plant using Eq. 1. The resulting DC power time series constitutes the data employed in the REAL experimental setting and for the validation of the day-ahead forecasts for both SIM and REAL settings.

c) Simulated vs measured data: We employed a physics-based model described in Section III-C to generate a synthetic dataset of PV power production for the real-world PV site of Mathi by parameterizing the simulator sourcing from its specification, reported in Table I. Due to the lack of a meteorological station on-site, the environmental inputs for the simulation, including GHI, ambient temperature, and wind speed, are obtained from ARPA Piemonte [12], for the closest meteorological station available located in Caselle Torinese, approximately 11 km away from the PV demo site. The DNI and DHI components are retrieved by decomposing the GHI using Boland model provided by `pvlib`. We generated the synthetic DC power dataset for the period from January 2010 to December 2025 by feeding the ARPA Piedmont meteorological measurements for that period into the PV simulation model. This approach provides a sufficiently long dataset that captures production seasonality, allowing the model to learn from a wide range of operating conditions.

Fig. 3 shows the comparison between simulated and measured PV power for a few sample days, while Table III reports the global errors between the two timeseries for the whole time span of the real dataset. The mathematical formulation and description of the selected metrics is provided in Section IV. The results, evaluated over the full real dataset, indicate that the simulation model closely reproduces with limited errors the observed PV production dynamics. Overall, the model provides a reliable representation of the process, supporting its use for synthetic data generation.

d) Dataset splits: We chronologically split the datasets into training, validation, and test sets. For the simulated dataset, the training set covers the period from January 2020 to June 2024. The validation set includes the second half of 2024 and a few selected months of 2025. The test set

TABLE II
ARCHITECTURE OF THE PROPOSED 3D-CNN FORECASTING MODEL, WHERE *ks* REPRESENTS THE KERNEL SIZE AND *st* IS THE STRIDE OF THE CONVOLUTIONAL LAYERS.

Layer	**Hyperparameters**
Input	tensor shape = (C, T, H, W) (NWP fields, lead times, latitude, longitude)
With	$C = 2$, $T = 28$, $H = 38$, $W = 43$,
Conv3D	channels = $C \rightarrow 128$, ks = $(1, 3, 3)$, ReLU
Conv3D	channels = $128 \rightarrow 128$, ks = $(1, 3, 3)$, st = $(1, 2, 2)$
Conv3D	channels = $128 \rightarrow 128$, ks = $(1, 3, 3)$, st = $(1, 2, 2)$
Conv3D	channels = $128 \rightarrow 128$, ks = $(1, 3, 3)$, st = $(1, 2, 2)$
Conv3D	channels = $128 \rightarrow 128$, ks = $(1, 3, 3)$, st = $(1, 2, 2)$
Temporal Conv3D	ks = $(T, 1, 1)$, ReLU
Flatten	–
Fully Connected	hidden units = ff, ReLU
Dropout	rate = 0.2
Fully Connected	hidden units = ff, ReLU
Dropout	rate = 0.2
Fully Connected	hidden units = ff, ReLU
Output Layer	units = N_{lead}

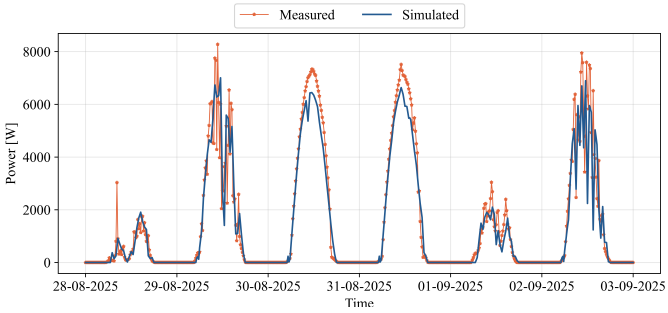


Fig. 3. Comparison between simulated and measured PV power for a sample period of time.

TABLE III
THE PERFORMANCE METRICS COMPUTED BETWEEN THE SIMULATED AND MEASURED PV POWER GENERATION FROM JULY 2024 TO DECEMBER 2025, THE ENTIRE TEMPORAL COVERAGE OF THE MEASURED DATASET.

RMSE [W]	**nMAE**	**MBE [W]**	**R^2**
643.55	0.21	-119.55	0.92

consists of four representative months of 2025 (April, July, September, and December), chosen to cover different seasons. As shown in Fig. 1, the test set is always defined using real PV production data and is shared across both SIM and REAL experimental settings. This ensures a fair comparison during the performance evaluation of the SIM and REAL setting. For the real dataset, the training and validation sets are selected within the available period between June 2024 and late 2025. Due to the limited availability of measurements, non-consecutive months are used for training, while a smaller subset is reserved for validation.

The temporal splits are designed to cover different seasons of the year, ensuring that training, validation, and test sets include a representative range of meteorological conditions influencing PV production. In this way, we make sure that both models are evaluated under identical real-world conditions, isolating the effect of the training data source.

We consider a training or test sample to be valid only if both the NWP input variables and the corresponding PV power target are available. If either the NWP inputs or the target value is missing, the sample is excluded from both training and evaluation. This filtering step inevitably reduces the number of usable samples. Table IV reports the number of valid samples available for each split in the SIM and REAL experimental settings. Test samples are identical for both settings.

The model parameters were optimized using the training set, while the validation set was used for hyperparameter tuning and early stopping. Specifically, we specified a number of epochs equal to 1000 with a patience of 200, and a batch size of 256. The experiments were conducted on an NVIDIA A5000 GPU.

IV. Results

In this section, we present the evaluation of the forecasting model, comparing its performance for the SIM and REAL experimental settings. We first introduce the chosen performance metrics, and then discuss the results across different test periods, highlighting differences in prediction errors and the ability to capture PV production variability.

A. Evaluation Metrics

We evaluated the performance of the day-ahead forecasting models in the SIM and REAL experimental settings in terms of the root mean squared error (RMSE), the normalized mean absolute error (nMAE), the mean bias error (MBE), and the coefficient of determination (R^2).

The RMSE measures the square root of the average squared difference between predicted and observed values, giving higher weight to large errors. The nMAE represents the mean absolute error normalized by the installed capacity of the PV system, providing a scale-independent indicator that facilitates comparison across different plants. The MBE quantifies the average bias of the forecasts and indicates whether the model tends to systematically overestimate or underestimate PV production. Finally, the R^2 measures the proportion of variance in the observed data explained by the model predictions. Eq. 3, 4, 5, 6 describe the mathematical formulation of the selected performance metrics:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(y_{test,n} - y_{pred,n})^2}{N}} \tag{3}$$

$$nMAE = \frac{\sum_{i=1}^{N}|y_{test,n} - y_{pred,n}|}{\overline{y}_{test}} \tag{4}$$

$$MBE = \frac{\sum_{n=1}^{N} y_{test,n} - y_{pred,n}}{N} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_{test,n} - y_{pred,n})^2}{\sum_{n=1}^{N}(y_{test,n} - \overline{y}_{test})^2} \tag{6}$$

where $y_{pred,n}$ and $y_{test,n}$ are the predicted and observed value for sample n respectively, $\overline{y}_{test}$ is the mean value of the observed values, and N is the total number of predictions.

B. Forecasting results

Table V reports the forecasting performance obtained under the two experimental settings in terms of the selected metrics. The SIM setting achieves comparable and often superior performance compared to REAL. Considering the overall results, the SIM model achieves an RMSE of 932.62 W, considerably lower than the 1270.13 W obtained by the REAL model. A similar trend is observed for the nMAE, which decreases from 0.41 in the REAL setting to 0.33 in the SIM setting. Furthermore, R^2 improves from 0.71 to 0.85, indicating that the SIM model better captures the variability of the observed PV production.

The most significant differences emerge during high-irradiance conditions. For instance, in July 2025, the SIM model achieves an RMSE of 1067.82 W compared to 1929.8 W for the REAL model, while also improving the R^2

TABLE IV
NUMBER OF VALID SAMPLES FOR EACH SET IN THE SIMULATED AND REAL DATASETS.

Exp. setting	Training	Validation	Test
SIM	1394	339	75
REAL	278	85	75

TABLE V
PERFORMANCE METRICS COMPARISON ON THE DAY-HEAD FORECASTS FOR THE SIM AND REAL EXPERIMENTAL SETTINGS, ON THE SHARED EVALUATION TEST SET.

Period	Exp. setting	RMSE [W]	nMAE	MBE [W]	R^2
Apr 2025	SIM	926.04	0.31	-267.82	0.87
	REAL	**872.93**	**0.26**	**48.52**	**0.89**
Jul 2025	SIM	**1067.82**	**0.30**	**1.80**	**0.84**
	REAL	1929.8	0.53	-874.32	0.47
Sep 2025	SIM	**1083.60**	**0.42**	**-118.06**	**0.73**
	REAL	1137.53	0.43	-6.17	0.70
Dec 2025	SIM	500.73	0.52	**-78.96**	0.75
	REAL	**498.06**	**0.44**	-86.01	0.75
Total	SIM	**932.62**	**0.33**	**-127.0**	**0.85**
	REAL	1270.13	0.41	-257.83	0.71

from 0.47 to 0.84. Similarly, in September 2025, the SIM setting slightly outperforms the REAL setting, with an RMSE of 1083.60 W against 1137.53 W and a higher R^2 (0.73 vs 0.70). Instead, during lower-irradiance periods, the performance of the two settings becomes more comparable. For example, in December 2025, both models achieve similar RMSE values (500.73 W for SIM and 498.06 W for REAL), although the REAL model exhibits a slightly lower nMAE. In April 2025, the REAL setting shows marginally better performance, with an RMSE of 872.93 W compared to 926.04 W for SIM, suggesting that real data may still provide an advantage under certain conditions.

Generally, the SIM setting exhibits a more stable behavior across different seasons, especially in periods characterized by higher variability in PV production. This is supported by the observed lower MBE bias, which remains closer to zero compared to those of the REAL setting. We can associate these improvements to the larger temporal coverage and variability of the simulated dataset, which includes a wider range of operating conditions.

Fig. 4 illustrates the comparison between measured PV production and day-ahead predictions generated by the models trained under the SIM and REAL settings for a few sample days in the test set. The figure highlights how the model trained with synthetic data captures the overall variability of the actual PV production.

The results indicate that training on simulated data can effectively enhance the generalization capability of the forecasting model when evaluated on real-world observations, especially in scenarios characterized by limited availability of historical measurements.

V. Conclusions

This work investigated the use of simulated PV production data for training deep learning models for day-ahead PV power forecasting. A `pvlib` physics-based simulation framework was used to generate synthetic PV power time series based on meteorological inputs and plant characteristics. The approach was evaluated under two experimental settings, SIM and REAL, with models tested on a shared real-world dataset. Results on a real-world demo PV site show that the SIM setting achieves competitive and often superior performance compared to the REAL setting, with lower errors and higher R^2, particularly under high-irradiance and highly variable conditions. This indicates that simulated data can effectively support model training and improve generalization when real measurements are limited. Overall, simulation-based datasets represent a valuable alternative in data-scarce scenarios, enabling the development of robust forecasting models. Future work will extend the analysis to multiple plants and locations, and investigate hybrid strategies combining simulated and real data through transfer learning techniques.

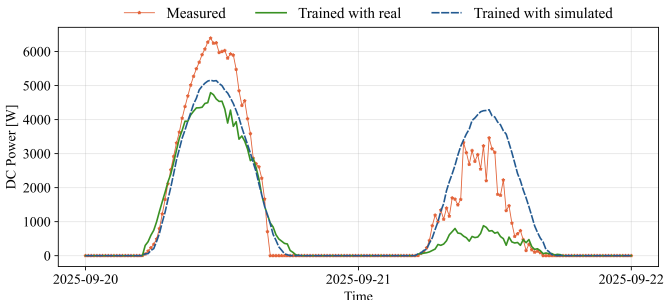


Fig. 4. Day-ahead predictions from the synthetic-data and real-data trained models compared to the target measured power production.

References

[1] IEA, "Electricity 2026," IEA, Paris, Tech. Rep., 2026.

[2] M. J. Mayer and G. Gróf, "Extensive comparison of physical models for photovoltaic power forecasting," *Applied Energy*, vol. 283, p. 116239, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261920316330

[3] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, pp. 569–582, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148116311648

[4] H. Dai, Y. Zhang, and F. Wang, "A day-ahead pv power forecasting method based on irradiance correction and weather mode reliability decision," *Energies*, vol. 18, no. 11, 2025. [Online]. Available: https://www.mdpi.com/1996-1073/18/11/2809

[5] M. Liu, Z. Lai, Y. Fang, and Q. Ling, "Day-ahead photovoltaic power forecasting based on corrected numeric weather prediction and domain generalization," *Energy and Buildings*, vol. 329, p. 115212, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778824013288

[6] V. Suresh, P. Janik, J. Rezmer, and Z. Leonowicz, "Forecasting solar pv output using convolutional neural networks with a sliding window algorithm," *Energies*, vol. 13, no. 3, 2020. [Online]. Available: https://www.mdpi.com/1996-1073/13/3/723

[7] S.-C. Lim, J.-H. Huh, S.-H. Hong, C.-Y. Park, and J.-C. Kim, "Solar power forecasting using cnn-lstm hybrid model," *Energies*, vol. 15, no. 21, 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/21/8233

[8] U. Yahaya, D. Chenvidhya, Y. Sangpongsanont, B. Muenpinij, and T. Chenvidhya, "Simulation of photovoltaic power output using era5 reanalysis dataset validated with high-resolution observational measurements," *Japanese Journal of Applied Physics*, vol. 64, no. 5, p. 05SP22, may 2025. [Online]. Available: https://doi.org/10.35848/1347-4065/add169

[9] L. de Oliveira Santos, T. AlSkaif, G. C. Barroso, and P. C. M. de Carvalho, "Photovoltaic power estimation and forecast models integrating physics and machine learning: A review on hybrid techniques," *Solar Energy*, vol. 284, p. 113044, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X24007394

[10] G. Zängl, D. Reinert, P. Rípodas, and M. Baldauf, "The icon (icosahedral non-hydrostatic) modelling framework of dwd and mpi-m: Description of the non-hydrostatic dynamical core," *Quarterly Journal of the Royal Meteorological Society*, vol. 141, no. 687, pp. 563–579, 2015. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2378

[11] Deutscher Wetterdienst, "Icon-eu numerical weather prediction model," 2024, available at: https://opendata.dwd.de/.

[12] ARPA Piemonte, "Agenzia regionale per la protezione ambientale," 2025, available at: https://www.arpa.piemonte.it/.