energies

MDPI

Article

Machine Learning and Weather Model Combination for PV Production Forecasting

Amedeo Buonanno, Giampaolo Caputo *, Irena Balog, Salvatore Fabozzi, Giovanna Adinolfi, Francesco Pascarella, Gianni Leanza, Giorgio Graditi and Maria Valenti

Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), 00196 Rome, Italy; amedeo.buonanno@enea.it (A.B.); irena.balog@enea.it (I.B.); salvatore.fabozzi@enea.it (S.F.); giovanna.adinolfi@enea.it (G.A.); francesco.pascarella@enea.it (F.P.); gianni.leanza@enea.it (G.L.); giorgio.graditi@enea.it (G.G.); maria.valenti@enea.it (M.V.)
* Correspondence: giampaolo.caputo@enea.it

Abstract: Accurate predictions of photovoltaic generation are essential for effectively managing power system resources, particularly in the face of high variability in solar radiation. This is especially crucial in microgrids and grids, where the proper operation of generation, load, and storage resources is necessary to avoid grid imbalance conditions. Therefore, the availability of reliable prediction models is of utmost importance. Authors address this issue investigating the potential benefits of a machine learning approach in combination with photovoltaic power forecasts generated using weather models. Several machine learning methods have been tested for the combined approach (linear model, Long Short-Term Memory, eXtreme Gradient Boosting, and the Light Gradient Boosting Machine). Among them, the linear models were demonstrated to be the most effective with at least an *RMSE* improvement of 3.7% in photovoltaic production forecasting, with respect to two numerical weather prediction based baseline methods. The conducted analysis shows how machine learning models can be used to refine the prediction of an already established PV generation forecast model and highlights the efficacy of linear models, even in a low-data regime as in the case of recently established plants.

Keywords: machine learning; PV forecasting; microgrids; signal processing

Citation: Buonanno, A.; Caputo, G.; Balog, I.; Fabozzi, S.; Adinolfi, G.; Pascarella, F.; Leanza, G.; Graditi, G.; Valenti, M. Machine Learning and Weather Model Combination for PV Production Forecasting. *Energies* **2024**, *17*, 2203. https://doi.org/10.3390/en17092203

Academic Editor: Younes Mohammadi

Received: 22 March 2024
Revised: 24 April 2024
Accepted: 25 April 2024
Published: 3 May 2024

1. Introduction

Climate change is deeply related to anthropic activities on earth, especially those of the industrial, agriculture, energy production, and transport sectors. Decarbonization is a challenging task in such contexts. It requires novel methods, technologies, strategies, and systems. In the last few decades, considerable efforts have been made by scientists and researchers to facilitate the decarbonization pathway. This is particularly true in the energy sector with numerous implemented initiatives to reduce its Greenhouse Gas (GHG) emissions [1]. These initiatives increasingly also have an economic assessment of the problem [2] underlining how the effects of global warming have negative repercussions on local or global economies, and also how risk mitigation interventions are much less expensive than the consequences of possible extreme events due to ongoing climate change [3]. From a political point of view, both national and international energy transition plans set ambitious decarbonization objectives. Initiatives such as the European Green Deal, Fit for 55%, and the National Ecological Transition Plans emphasize the central role of electrification in different sectors (mobility, heating, etc.), and the substantial contribution of renewable energy sources to electricity generation. In this context, microgrids certainly represent one of the most promising models of transformation of the electricity system. The development of microgrids has emerged as a catalyst for the seamless integration of renewable energy sources into the broader energy ecosystem. Microgrids can function as cohesive entities, capable of islanded operating (grid-off mode) or are able to function in concert with the

main AC grid (grid-on mode). Microgrids represent a solution favoring the integration of renewable energy sources into both Alternating Current (AC) and Direct Current (DC) grids by power electronics converters.

The operation of microgrids depends on characteristics, functioning modes, and the reliability of their internal resources and the power systems they are interfaced with [4]. They can be managed through diverse controller configurations, such as hierarchical, decentralized, or centralized ones. Hierarchically controlled microgrids are of particular significance; each of them is characterized by a Master controller endowed with the capability to oversee and optimize internal resources in alignment with specific energy or economic strategies. Moreover, local Slave controllers are assigned to individual microgrid resources or groups of resources, ranging from renewable power plants and loads to energy storage systems. These controllers are in charge of measures and data transmission, the execution of commands, and alarm activation in the case of critical operation. Data collected and acquired are typically transmitted and processed by Energy Management Systems (EMSs) and specialized platforms, facilitating the planning and optimization of power flows [5].

Within this transformative landscape, Italy has made significant strides, with its photovoltaic (PV) plants boasting an installed capacity of 25 GW as of 2022. Notably, approximately 33% of these installations fall within the 200–1000 kW range [6]. Such sources offer a sustainable and eco-friendly energy solution, but their intermittent and variable production patterns can pose operational challenges that can be effectively addressed through proper modeling [7] and forecasting methodologies [8]. Accurate forecasts play a pivotal role in enabling meticulous planning, strategic commitment, the efficient management of available resources in the renewable energy sector [9], and in ancillary service provision [10]. Errors in energy forecasts can result in significant economic losses for microgrids, including unnecessary energy purchases from the main grid or excessive energy storage. Precise forecasts help us to mitigate these inefficiencies, reduce energy waste, and optimize investments in energy infrastructure, ultimately resulting in substantial economic savings.

For the optimal operation of microgrids, the availability of accurate forecasts is non-negotiable. These forecasts extend to the prediction of generation from renewable sources and from projected microgrid demand.

1.1. Literature Review

While ground-based observations from solar metric stations provide valuable data, they are often burdened by high costs and may not offer continuous, long-term coverage. As an alternative, meteorological satellites provide the means for the indirect determination of solar parameters, including cloudiness, albedo, and solar radiation reaching the earth's surface. Contemporary meteorological forecast models, particularly numerical weather prediction (NWP) ones, have become essential for estimating renewable energy resources, notably solar radiation. Meteorological forecast models, nowadays, are able to predict renewable resources well [11], and their valuable data can be used in further forecasts of solar power plant production (PV [12] and Concentrated Solar Power systems). The NWP models could offer great spatial and temporal coverage against observation data, which are spatially sparse and lack long temporal resolution. Those models can simulate the future and past atmospheric conditions, and they have become very valuable tools in the management of renewable energy plants [13]. Significantly, meteorological forecast models have demonstrated their proficiency in predicting renewable resources and have evolved tools in forecasting and managing renewable energy systems. Recent years have witnessed a surge in the adoption of Machine Learning (ML) techniques predicting PV output [14]. They are deeply analyzed in the literature sector [15]. Several ML models, including Multiple Linear Regression (MLR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), Fully Connected Neural Network (FCNN), Bayesian Neural Network (BNN), Support Vector Regression (SVR), and Regression Tree (RT), have been harnessed and subjected to comparative analyses across diverse studies. For example, in [15], several approaches, grouped into direct (methods that output directly the PV power)

and indirect (methods that require the solar irradiance forecast, the estimation of the plane of array irradiance, and the PV performance model) forecasting techniques, are evaluated.

A comparative analysis of MLR, SVR, RF, GB, and FCNN methodologies is reported in [16] for a day-ahead solar power forecast. It is carried out taking into account historic power production and regional weather prediction related to 152 PV systems. The conducted comparison underscores the fact that GB and RF techniques excel in predicting production for both individual and aggregated PV systems when compared with other considered methods. The BNN method is compared to SVR and RT techniques in [17]. Another study is reported in [18], where the PV production for the University of Manchester plant is forecasted considering different-sized datasets and diverse time horizons.

In [19], the authors employed MLR and Artificial Neural Network (ANN) methods for predicting PV production within a solar microgrid. The paper concluded that MLR offers simplicity and computational efficiency but struggles with capturing complex non-linear relationships, whereas ANN excels in capturing non-linear patterns albeit demanding larger datasets and computational resources with reduced interpretability. Furthermore, the authors in [20] examined enduring techniques for predicting energy consumption, PV, and wind power production.

Furthermore, in [21], cloud cover, humidity, and temperature impacts on PV generation predictions were evaluated by 175 time series. They were obtained measuring the production of an actual rooftop-mounted PV system installed in Utrecht (Netherlands). Moreover, the authors offered a concise overview of prediction methodologies employed in microgrids, particularly for short-term forecasting.

In the context of a microgrid, PV production forecasting was also discussed in [22], and the Tunicate Swarm Algorithm (TSA)-based Least-Square Support Vector Machine (LSSVM) was applied. It offers the advantages of robustness to noisy data and the ability to capture complex patterns, while suffering from the need for hyperparameter tuning and is computationally demanding. The TSA-based Multilayer Perceptron Neural Network (TSA-MLPNN) provides flexibility in capturing non-linear relationships and uncertainty estimation but demands significant computational resources and may be sensitive to overfitting. On the other hand, the Whales Optimization Algorithm (WOA)-based LSSVM offers advantages in convergence speed and robustness to local optima, but it may require careful hyperparameter tuning and lacks interpretability.

Lastly, in [23], a blended Fuzzy-PSO smart forecasting method is deployed, and its precision is documented and contrasted with Fuzzy and Fuzzy-GA prediction models. The authors underline the advantages of the Fuzzy-PSO smart forecasting method, which include the ability to combine the strengths of both fuzzy logic and Particle Swarm Optimization (PSO), potentially leading to improved accuracy and robustness in prediction. However, this method may require more computational resources due to the optimization process involved in PSO, and its interpretability could be reduced compared to standalone Fuzzy logic models.

Recently, the hybridization of data-driven approaches based on the history of observed production and of physical approaches that employ both weather forecast and a mathematical model of the PV system are emerging as effective solutions [24]. For example, in [25], an ANN employs clear sky solar radiation and weather forecast data to obtain a PV production prediction, while, in [26], an ANN that has inputs and also the output of a five-parameter equivalent model of a PV module (using datasheet data or using an optimization method to identify the parameters) is considered. The studies observe that PV production forecasting is improved by the combined (hybridized) methods.

1.2. Beyond the State of the Art

In this work, we propose a combined approach of an already established NWP-based PV production model with an ML model that leverages the past observations of production to improve the final performance. Differently from works present in the literature, our approach assesses different types of ML models and two baseline models (a physical and a

data-driven one). The ML model does not have direct access to weather forecast variables. The novelty of our approach is hence in the combination of these different elements, where some of them can be assumed to be already available on a considered site, to improve the final forecasting result.

By improving the accuracy of PV production forecasts, our research plays a crucial role in ensuring that solar energy is harnessed to its full potential, furthering the microgrid's support to minimize carbon footprints and advance the cause of the sustainable energy transition.

The analyzed case study refers to the ENEA—Italian National Agency for new technologies, energy and economic sustainable development—microgrid realized at the Research Centre of Portici (Italy). It is a demonstrator of "Multivector Integrated Smart Systems and Intelligent microgrids for accelerating the energy transition" (MISSION) project furnished with renewables (solar and wind), a Combined Heat Power (CHP) plant, storage devices, and a data center critical load.

The objective of achieving carbon-neutral microgrids is actively pursued by optimizing and managing the available resources in alignment with energy and environmental objectives [27]. The proposed models are implemented in the MISSION demonstrator, the functionality of which has strong connections with precise PV output predictions, ensuring appropriate resource management.

A promising aspect of the proposed combined method consists of its significant performance improvement, also in a low-data regime, as seen in the case of recently installed plants. This research is poised to make a substantive contribution to the overarching mission of cultivating sustainable, climate-resilient energy systems and driving the realization of carbon-neutral microgrids. In our previous article [28], we described some preliminary results that were extended with more models for comparison, a more detailed discussion of results, and a more extensive literature review.

The manuscript is organized as follows. Section 2 is dedicated to prediction models and the considered combined approach description. In Section 3, details about the model training process are provided. The conducted tests and obtained results are reported and commented on in Section 4.

2. Materials and Methods

2.1. Weather Research and Forecasting Model

The Weather Research and Forecasting model (WRF) is one member of the NWP model family designed for both atmospheric research and operational forecasting applications [29]. The model is used in different research areas in a wide range of meteorological applications where different time and horizontal resolutions, from tens of meters to thousands of kilometers, could be applied [30].

In this work, the computational spatial resolution of the WRF model is set to 10 km $\times$ 10 km with 151 $\times$ 151 simulation grid points. The computational domain covers the entire region of the Italian peninsula with a center at 41.25° latitude and 13.5° longitude. The model uses 30 sigma atmospheric vertical levels and 4 soil levels. Input boundary conditions are given with four-time-daily runs of the Global Forecast System [31]. The WRF model starts its forecast at 00:00 UTC for the next 48 h where the first hours are considered as a model spin-up. The described WRF model provides different atmospheric outputs. In the assessed application, the authors consider only atmospheric temperature (T_A) and Cloud Cover (CC). The outputs of the described atmospheric forecasting model are used as data inputs for the PV plant production model that will be detailed below.

2.2. BaselineP Model

In PV plant installation sites, where observations of solar irradiance, in particular Global Horizontal Irradiance (GHI) are available, the maximum potential of global irradiance at the horizontal surface can be estimated. The specific maximum potential radiation is estimated when ground horizontal irradiance is considered with no cloud condition.

This information, together with the astronomical sun position and extraterrestrial global horizontal irradiance, provide an input for the "clear-sky" model.

Observations of GHI are expensive, and, with the lack of these data, satellite data and NWP models are used to estimate atmospheric temperature, cloud coverage, relative humidity, and radiation balance outcomes.

First, Cloud Cover (*CC*) is used from a weather model to decrease the intensity of the GHI clear sky value ($GHI_{clear\ sky}$) and to obtain the GHI forecast for a specified location (GHI_{pred}). The mathematical expression of the previous statement is written in Equation (1):

$$GHI_{pred} = GHI_{clear\ sky} \cdot (offset + (1 - offset) \cdot (1 - CC)) \quad (1)$$

where the *offset* is 0.35 and *CC* is 0 for a no clouds condition, and 1 for completely covered cloudy conditions [32]. When GHI_{pred} for the desired location is calculated, the decomposition model of solar radiation can be applied.

Each PV module has its own orientation, azimuth angle (α) with respect to the south direction, and tilt (β) of the panel with respect to the horizontal plane. The decomposition model allows us to compute Global Irradiance on any oriented PV surface (*GI*) as reported in Equation (2). The incident radiation *GI* on the panel surface is composed of three components: the direct radiation I_{dir}, the diffuse radiation I_{diff}, and the radiation reflected from the ground I_{ref}:

$$GI = I_{dir} + I_{diff} + I_{ref} \quad (2)$$

Now, the temperature of PV module (T_M) can be computed as in Equation (3) [33]:

$$T_M = T_A + (NOCT - 20)\frac{GI_{pred}}{G} \quad (3)$$

where T_A stands for ambient temperature [°C] (output from WRF model), *NOCT* is the Nominal Operating Cell temperature [°C] calculated for a wind speed at a PV module height of 1 m/s, an ambient temperature of 20 °C, and an irradiance value of G = 800 W/m^2.

Finally, it is possible to calculate the PV production forecast for a specific plant (PV_{pred}) in the following 24 and 48 h by using the PVWatts model represented in Equation (4):

$$PV_{pred} = \eta\, P_n \frac{GI_{pred}}{GI_0}(1 + K(T_M - T_0))(1 - A) \quad (4)$$

where η is the inverter efficiency, P_n is the nominal power of the PV plant [W], GI_0 is the global solar radiation at standard test condition (=1000 W/m^2), K is the temperature coefficient of PV modules [%°C^{-1}], T_M is the temperature of the PV module, T_0 is the reference cell temperature at the standard test conditions (25 °C), and A represents the system losses [34].

Given the great generality of the WRF model, the BaselineP model can be readily applied to any location.

In this work, we have considered $NOCT = 47$ °C, $P_n = 9000$ W, $K = -0.45\%°C^{-1}$, $A = 14\%$, $\eta = 0.95$ in Equations (3) and (4).

2.3. BaselineD Model

When there is constant monitoring of PV production and ground solar radiation available, an uncomplicated model to forecast PV production could be used, here called BaselineD model. This model represents the relationship between measurements of PV production and GHI observations, as described in Equation (5):

$$PV_{obs} = m \cdot GHI_{obs} + q \quad (5)$$

The parameters m and q are fitted based on the available monitored PV production and GHI values. This mathematical function is then applied on the outputs of Equation (1) (GHI_{pred}) to calculate the forecast of PV production (PV_{pred}). This model could be used

when observations are available. The found linear correlations are valid only for the exact PV plant and cannot be used for different PV locations.

In Figure 1, the relationship between PV plant production (PV_{obs}) and the observation of GHI (GHI_{obs}), and their corresponding correlation coefficients for 3 January 2021, is depicted. The example of a cloudy day is shown to present that an unpredictable radiation condition of PV production also has linear dependence with GHI. In this study, 365 relationships are obtained, one for each day of the year, and the relationship depicted by Equation (5) differs for each day. The correlation of the previous day can be used for the forecast of the next day because the astronomical sun–earth position has small variations. Daily correlations are valid only for the considered PV plant and the same approach could be used where there are available daily measurements of PV production and GHI.

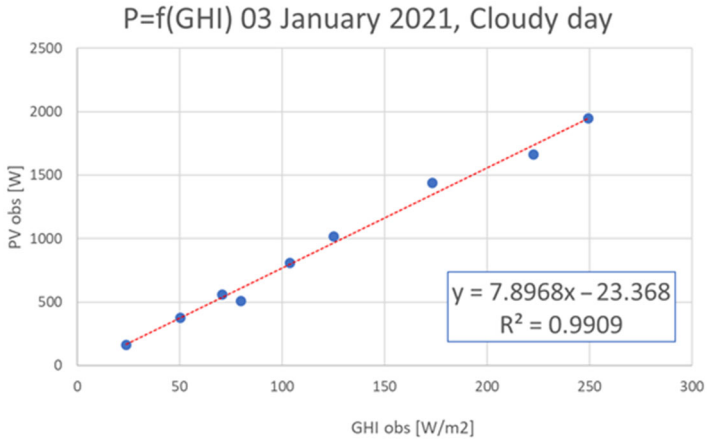

Figure 1. Correlation between the PV production and GHI observations. The blue dots are the observations of GHI and PV production, and red dashed line is the computed regression line.

A conceptual scheme of both BaselineP and BaselineD models is presented in Figure 2.

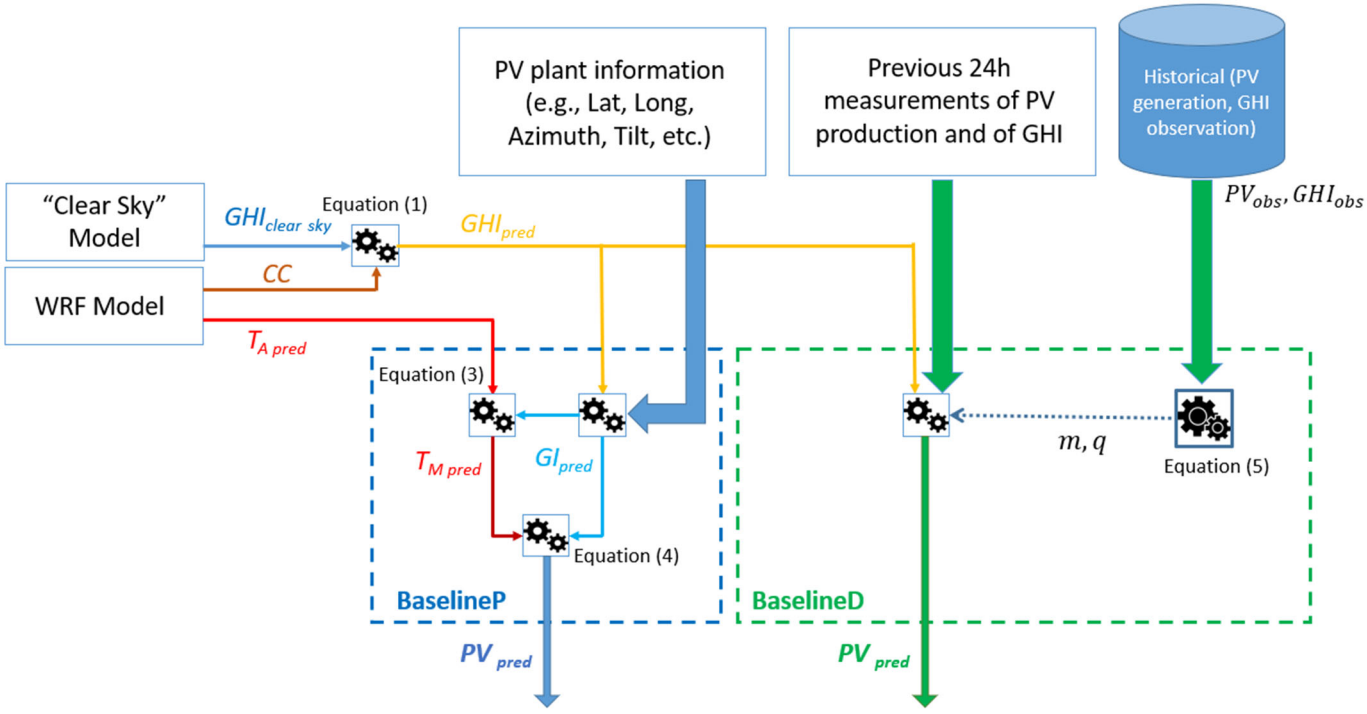

Figure 2. Conceptual scheme of BaselineP (blue) and BaselineD (green).

2.4. Proposed Approach

This section focuses on forecasting a target value (y) for a specific time (t) and for the following $H-1$ time steps where H is the forecasting horizon. There are many possible approaches, but a general scheme involves the following:

- Target values;
- Past covariates: variables influencing the target value, observed in the previous W time steps;
- Future covariates: variables impacting the target value, related to the time t and to the subsequent $H-1$ time steps, and that are known at the prediction time.

This concept is outlined in Equation (6), where the term y_{t-i} denotes the target value in the past, at the time step *t-i*; ${x^p}_{t-j}$ (with $p \in \{1,\ldots,P\}$) is the value of the p-th past covariate at the time step *t-j*; and ${z^f}_{t+k}$ (with $f \in \{1,\ldots,F\}$) indicates the value of the f-th future covariate at the time step $t+k$:

$$y_{t:t+H-1} = f\left(y_{t-1},\ldots,y_{t-W},{x^1}_{t-1},\ldots,{x^1}_{t-W},\ldots,\ldots\ldots{x^P}_{t-W},{z^1}_{t},\ldots,{z^1}_{t+H-1},\ldots,{z^F}_{t},\ldots,{z^F}_{t+H-1}\right) \quad (6)$$

A schematic representation of Equation (6) is presented in Figure 3, where the past target values together with past and future covariates are employed to predict the next H target values.

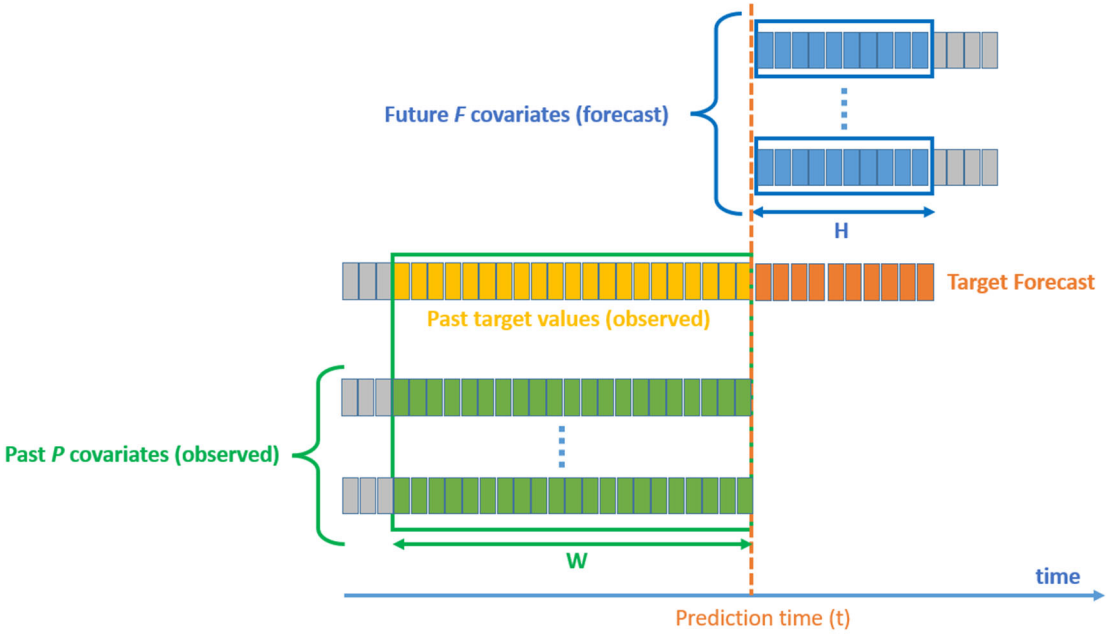

Figure 3. Generic forecasting problem's conceptual layout using future and past covariates.

In a scenario involving the production forecast of PV systems, the past covariates can be represented by various measurements such as solar irradiance and temperature. Conversely, the future covariates can be acquired either through weather forecasts or by utilizing PV production forecasts generated by baseline models, as performed in this work and illustrated in Figure 4.

The weather forecast module generates predictions for various weather variables for the next H hours. Subsequently, the baseline model utilizes these forecasts to provide a corresponding PV production forecast that can be employed by the ML model as future covariates, together with observed production, to enhance and refine the result.

Several ML regression models can be considered. In this work, we have focused on a linear model, Long Short-Term Memory (LSTM), eXtreme Gradient Boosting (XGB), and the Light Gradient Boosting Machine (LGBM), which are briefly described below. Even though these ML models are often employed in forecasting tasks, the originality of our

approach is using their combination with baseline models in order to have a more precise energy prediction model that leverages both historical data and a PV production forecast obtained from BaselineP or BaselineD.

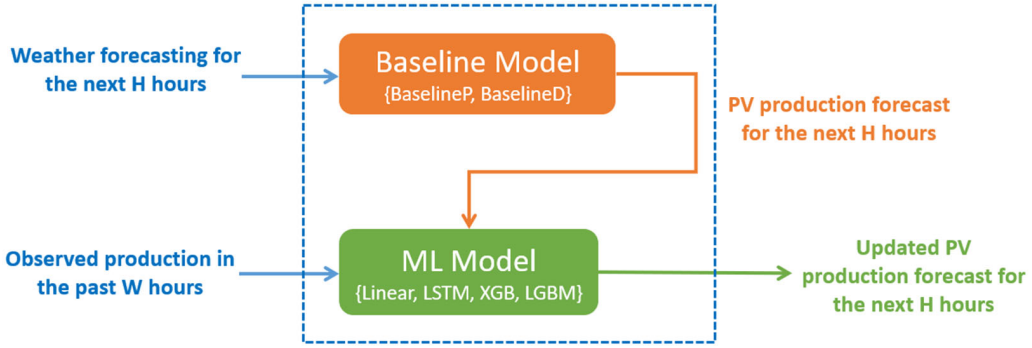

Figure 4. Proposed approach's conceptual scheme.

2.4.1. Linear Model

In the context of linear models [35], a mathematical representation capturing the relationship between the target variable to be predicted at time t (denoted as y_t), the historical target values $\{y_k\}_{k=t-1}^{t-W}$, and the future covariates $\{z_k\}_{k=t}^{t+H-1}$ is reported in Equation (7):

$$y_t = \delta_0 + \left(\sum_{k=t-1}^{t-W} \alpha_k y_k\right) + \left(\sum_{k=t}^{t+H-1} \gamma_k z_k\right) \quad (7)$$

Here, δ_0 represents the intercept, while α_k are the weights associated with the past target values and γ_k are the weights associated with the future covariates.

A crucial prerequisite to the predictions involves obtaining estimations for the future covariates (z_k) and the past target values (y_k). In the present study, the authors adopt a multi-model approach wherein H distinct models are trained. Each model is designed to predict a single value of the target within the interval $[t, t + H - 1]$.

2.4.2. Long Short-Term Memory (LSTM)

Differently by Feed Forward Neural Networks (FFNNs), the Recurrent Neural Networks (RNNs) are specialized in handling sequential data, and the computed output is not only dependent on the input but also on the hidden state of the system that is updated as the sequence is processed [35]. Utilizing the input (x_t) and the state from the preceding step (h_{t-1}), it is possible to update the current state (h_t) and compute the output (o_t) through the following equations:

$$h_t = \sigma_h(W_x \cdot x_t + W_h \cdot h_{t-1} + b_h) \quad (8)$$

$$o_t = \sigma_o(V_h \cdot h_t + b_o) \quad (9)$$

Here, σ_h and σ_o denote activation functions for the state and output, respectively; W_x, W_h, and V_h are the weight matrix for the input–state connection, the recurrent connection between states, and for the state–output connection, respectively; and b_h and b_o serve as the bias vectors for the state and output, respectively. The Long Short-Term Memory (LSTM) [36] is a specific RNN designed to mitigate issues like vanishing and exploding gradients by incorporating a memory element (c_t).

2.4.3. Gradient Boosting Methods

XGB [37] and LGBM [38] are two powerful ML algorithms belonging to the family of Gradient Boosting methods. The former employs a regularizer on the tree complexity to avoid overfitting and the latter makes the tree expand, leaf-wise. They are widely adopted due their speed, scalability, and robustness.

2.5. Datasets

To train and evaluate ML models, we need the PV production forecast obtained from BaselineP and BaselineD, the weather forecast, and the observed PV production. All data cover the year 2021 (from 1 of January to 31 of December).

2.5.1. PV Production

The observed PV production data are obtained by a PV installation above a motorcycle parking lot at the ENEA Research Centre located in Portici (NA), Italy. It consists of 36 glass/mono-crystalline panels, producing a nominal power of 9 kWp with a total covering area of 60 m^2. The installation has a tilt angle of 7° and is 28° south-east oriented (with 0° representing the south). The generated electrical energy is utilized for internal purposes after being delivered to the centre.

2.5.2. The Weather Forecast and the Baselined PV Production Forecast

The weather forecast values have been considered for the entire year of 2021. They represent the GHI and the predicted temperature for each day throughout the year on hourly basis.

These values are employed to predict the PV generation employing the BaselineP and BaselineD models described in Section 2.2 and Section 2.3, respectively.

In Figures 5 and 6, the actual and the predicted PV generation are shown for a sunny day and cloudy day (when the value of the daily clearness index (the ratio between the global irradiance and extraterrestrial irradiance on a horizontal surface) is greater than 0.65, the day is considered sunny otherwise cloudy) for the year 2021, respectively.

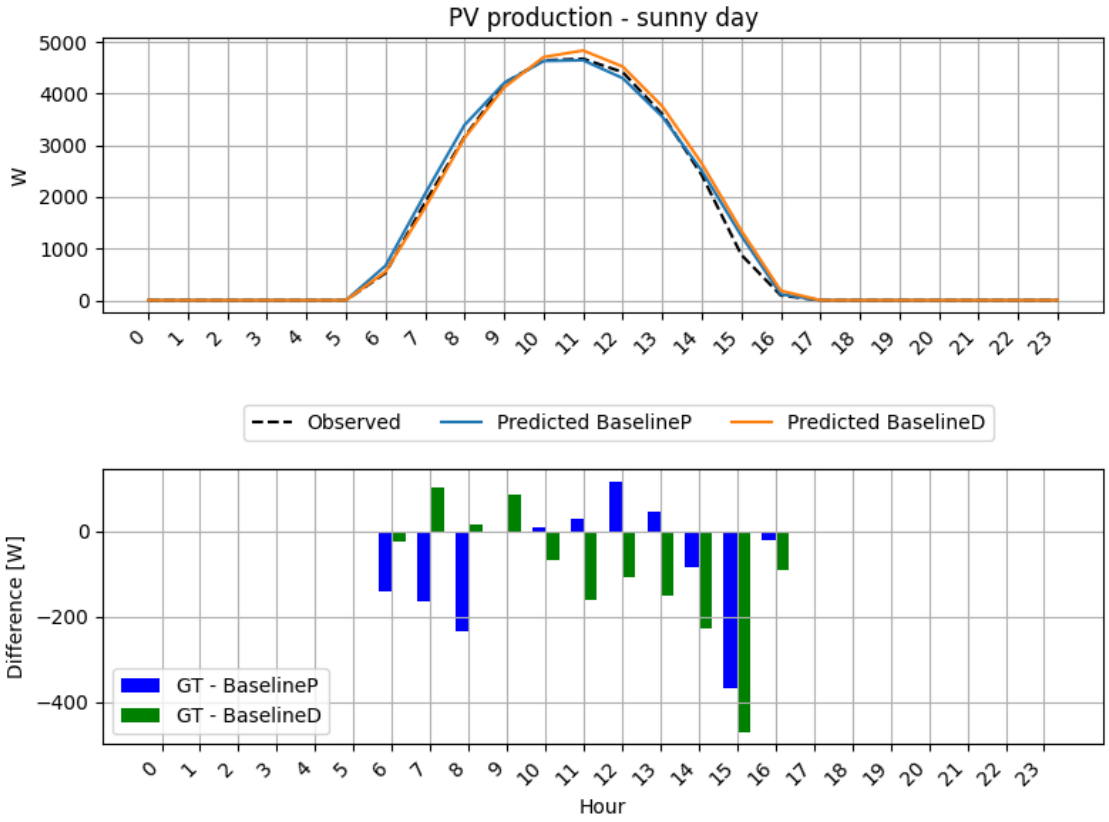

Figure 5. (**Top**) the actual and predicted PV production, using BaselineP and BaselineD, for a sunny day in the test set; (**bottom**) difference between the ground truth (observed) and the prediction results obtained using BaselineP and BaselineD.

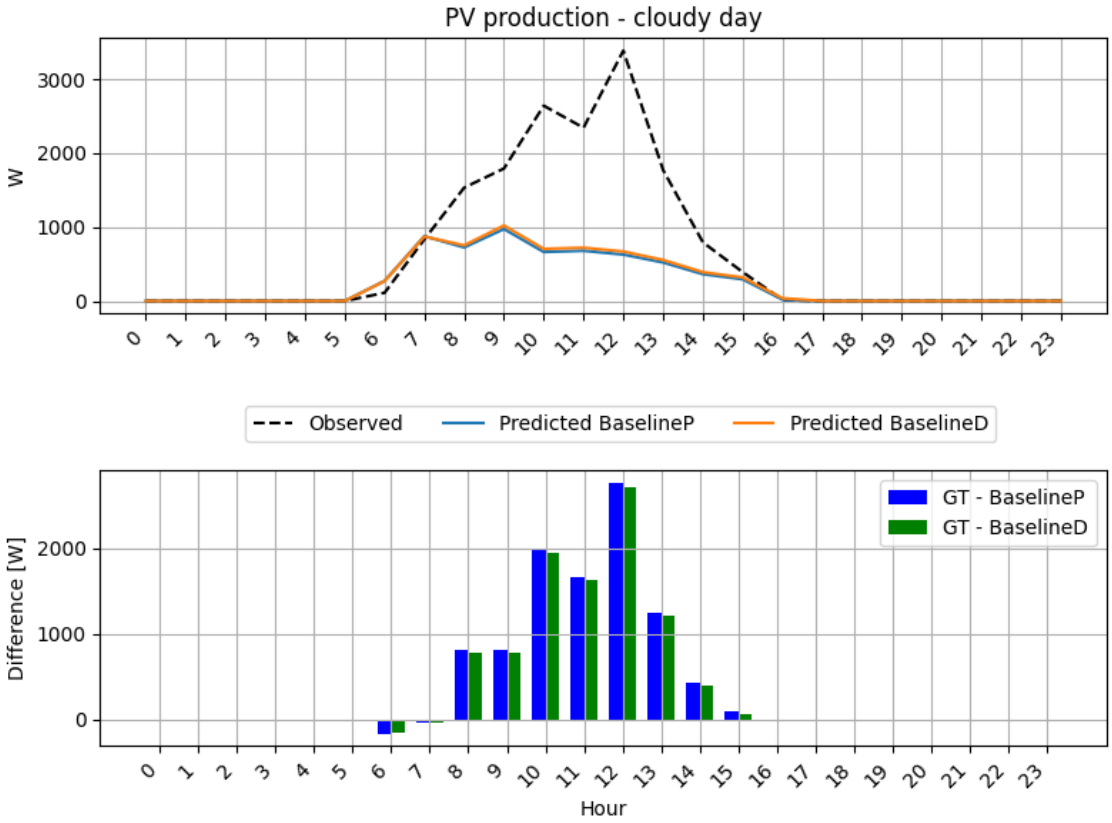

Figure 6. (**Top**) the actual and predicted PV production, using BaselineP and BaselineD, for a cloudy day in the test set; (**bottom**) difference between the ground truth (observed) and the prediction results obtained using BaselineP and BaselineD.

2.6. Metrics

The Root-Mean-Square-Error (*RMSE*) and the Coefficient of Variation (*CV*) are the metrics employed to evaluate the performance of the implemented models. The *RMSE* quantifies how closely a model's predictions $\hat{y}(t)$ align with the actual target values $y(t)$, often referred to as the ground truth. In this case, N represents the number of values taken into account. Differently, *CV* offers insights into the dispersion of errors relative to the average observed value $\overline{y}$. The *RMSE*, expressed in Watts, is computed using Equation (10), while the *CV* is dimensionless and is computed using Equation (11):

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y(t) - \hat{y}(t))^2} \tag{10}$$

$$CV = \frac{RMSE}{\overline{y}} * 100 \tag{11}$$

3. Model Training

The considered dataset is divided into three sets: the training set, validation set, and test set. Namely, the training set goes from 1 January 2021 0:00 to 30 September 2021 23:00, the validation set is internal to the training set and goes from 1 August 2021 0:00 to 30 September 2021 23:00, and the test set goes from 1 October 2021 0:00 to 31 December 2021 23:00.

The authors utilized the pre-known information of the sunrise and sunset time (obtained from Sunrise Sunset API (https://sunrise-sunset.org/api accessed on 26 February 2024)) to refine the predictions.

The implementation of the models was carried out utilizing Python, and some libraries such as Darts [39] version 0.27.1, Numpy version 1.24.3, pandas version 1.5.3, and Scipy version 1.10.1.

The experiments were conducted on a Personal Computer (PC) equipped with an Intel Core i7-9700 CPU running at 3.00 GHz with eight cores, 16 GB of RAM, and a NVIDIA GeForce GTX 1050Ti GPU. The operating system used was Windows 10 Pro.

4. Discussion

Figures 7 and 8 depict the forecasting for two representative days in the test set, a sunny and a cloudy day, respectively, together with the difference between the ground truth and the prediction.

Table 1 presents the outcomes of the PV generation forecasting, utilizing only the two baseline models, BaselineP and BaselineD, and the combined models LinearD, LinearP, LSTMD, LSTMP, XGBD, XGBP, LGBMD, and LGBMP, where the final P and D indicate which is the employed baseline model, BaselineP or BaselineD, respectively. All models (except baselines) employ a 48 h window for past target values, with a forecasting horizon of 24 h.

For each metric, considering the days in the test set, the authors calculated the average and standard deviation (in parentheses). The Wilcoxon signed rank test was used to compare the metrics obtained from the combined approach and the related baseline. The null hypothesis under consideration was that the paired samples obtained by the baseline and the model that utilizes it originate from the same distribution. In Tables 1 and 2, * represents the rejection of the null hypothesis considering a p-value < 0.05, associated with a statistically significant difference between the baseline and related combined model. Table 2 presents the average and standard deviation of *RMSE* computed over the days of the test set for all considered models, differentiating between sunny and cloudy days, as defined previously.

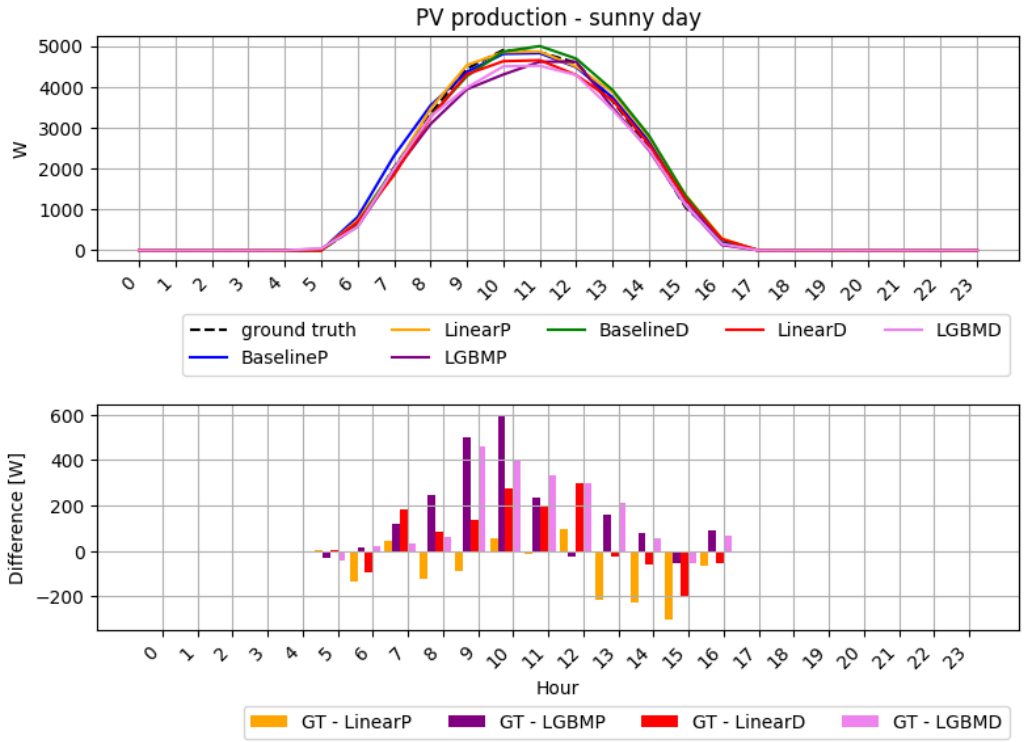

Figure 7. (**Top**) ground truth compared to the forecasting results of PV production obtained using only the baselines (BaselineP and BaselineD) and four combined models (LinearP, LinearD, LGBMP, LGBMD) for a cloudy day; (**bottom**) prediction error.

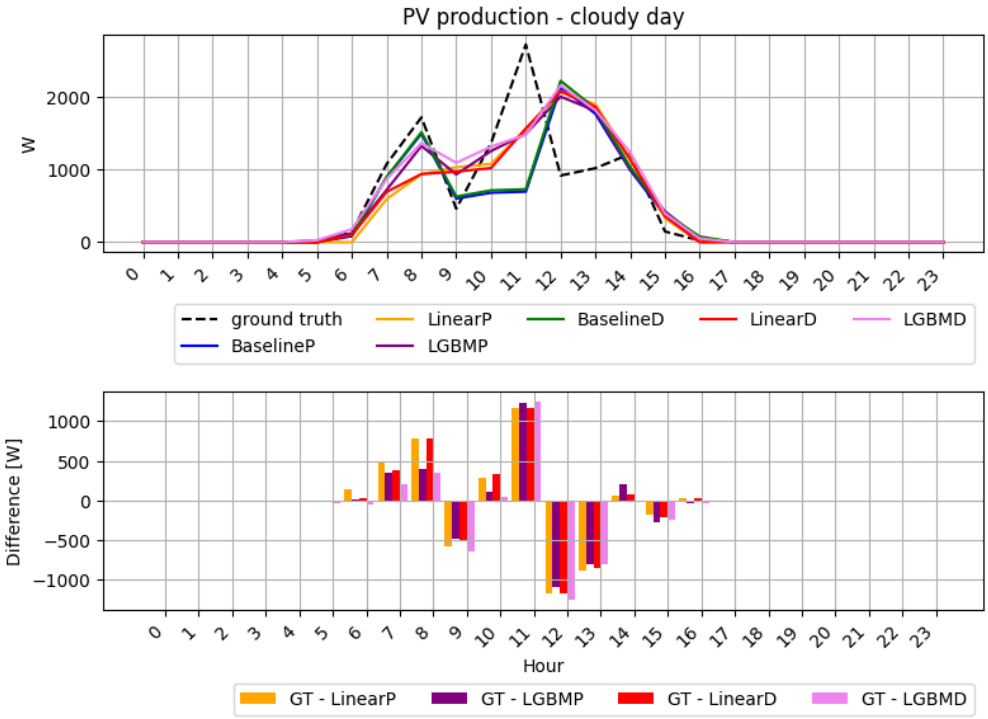

Figure 8. (**Top**) ground truth compared to the forecasting results of PV production obtained using only the baselines (BaselineP and BaselineD) and four combined models (LinearP, LinearD, LGBMP, LGBMD) for a cloudy day; (**bottom**) prediction error.

Table 1. PV generation forecasting results. For the days in the test set, the average (standard deviation) of *RMSE* and *CV* is calculated.

Type	Model	*RMSE*	*CV*
D	BaselineD	436.87 (253.95)	100.19 (91.62)
	LinearD	**420.72 (221.13)** *	**99.30 (95.83)** *
	LSTMD	435.18 (245.29)	100.15 (95.84)
	XGBD	502.75 (259.69) *	115.01 (106.85) *
	LGBMD	445.17 (245.45)	99.33 (85.65)
P	BaselineP	470.73 (248.73)	107.50 (96.66)
	LinearP	**438.47 (231.66)** *	103.26 (96.53) *
	LSTMP	454.46 (260.33) *	102.53 (91.77) *
	XGBP	487.92 (255.42)	112.51 (106.10)
	LGBMP	443.08 (250.66) *	**99.91 (90.17)** *

Table 2. PV generation forecasting results divided for sunny and cloudy days. For the days in the test set, the average (standard deviation) of *RMSE* is calculated.

Type	Model	Sunny Days	Cloudy Days
D	BaselineD	**223.57 (159.13)**	485.62 (246.31)
	LinearD	267.35 (170.66) *	**455.77 (216.37)** *
	LSTMD	260.65 (176.97) *	475.07 (228.16)
	XGBD	358.11 (188.82) *	535.81 (262.36) *
	LGBMD	287.85 (184.33) *	481.12 (243.511)
P	BaselineP	281.16 (159.12)	514.06 (245.22)
	LinearP	**254.81 (149.02)** *	480.44 (226.68) *
	LSTMP	257.04 (195.29)	499.59 (252.20)
	XGBP	335.36 (193.64) *	522.80 (255.04)
	LGBMP	290.70 (213.31)	**477.92 (245.50)** *

From Table 1, we can observe that BaselineD already produces promising results, having considered aspects not modeled by BaselineP and having discovered the observation of real data. Among the evaluated models, the linear ones outperform the other ones in terms of *RMSE* and consistently outperform the baselines, with a statistically significant improvement of 6.9% in the P case and of 3.7% in D case. Moreover, in the D case, compared to baseline, the linear model returns to a lower *RMSE* in 58.1% of all test set days (and in 67.1% of cloudy days), and in 69.8% of all test set days (and in 68.6% of cloudy days) in the P case.

In the D case, when specifically addressing only sunny days within the test set (Table 2), the baseline is optimal, but, considering only the cloudy days in the test set (83% of the total), the most effective solution becomes the linear model.

In P case, instead, linear model is the best model for sunny days and is the second-best approach after LGBM for cloudy days.

In Figure 9 the difference between the ground truth and the prediction is estimated on an hourly basis (from 5:00 to 17:00) and only for baselines and linear models. For the P case, the median error of the linear model is nearer to zero than that of the baseline. For the D case, instead, the median error of the baseline is lower, but the errors are usually more dispersed, presenting a higher range than the linear model.

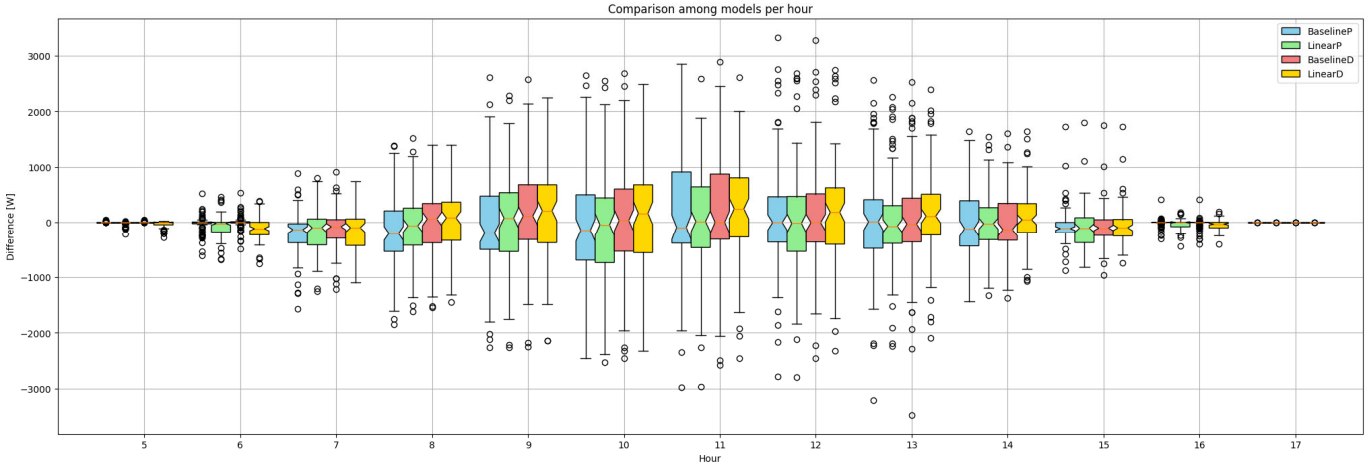

Figure 9. Box-and-whisker plot of the difference between ground truth and the prediction for the BaselineP, BaselineD, LinearP, and LinearD models. The white circles are the outliers.

Although the usage of one year of data is a limitation of this study, it allows authors to demonstrate the efficacy of the proposed methodology and of linear models even in a low-data regime, which is frequently the case in scenarios within a newly established plant. Integrating additional years into the dataset has the potential to help the ML models and to enable the exploration of more complex models that may have benefits from an extensive dataset. Moreover, the authors would like to point out that the proposed approach is general, and it is suitable to be applied in emerging photovoltaic systems where a weather forecast module and historical energy production dataset are available.

5. Conclusions

The forecasting of PV power holds essential significance in accurately strategizing and managing resources for grids and microgrids. It plays a pivotal role in aligning supply and demand, thereby aiding in preventing grid imbalances.

In this study, the authors explored the application of an approach leveraging information derived from both historically observed PV power generation and PV power forecasts generated by numerical weather prediction models. The conducted analysis highlights how machine learning models can be utilized to enhance the prediction of an already established PV generation forecast model. Several machine learning methods, including

the linear model, Long Short-Term Memory, eXtreme Gradient Boosting, and the Light Gradient Boosting Machine, were tested for the combined approach. However, the linear models proved to be the most effective, showing at least a 3.7% improvement in *RMSE* in PV production forecasting compared to two numerical weather prediction-based baseline methods. Among the employed machine learning models, the linear models demonstrated their validity, surpassing baseline performances and showcasing their effectiveness with only one year of data.

Author Contributions: Conceptualization, A.B. and G.C.; methodology, A.B., G.C., I.B. and G.A.; software, A.B., G.C. and I.B.; data curation, G.C., I.B., F.P. and G.L.; writing—original draft preparation, A.B., G.C., I.B., S.F. and G.A.; writing—review and editing, A.B., G.C., I.B., S.F., G.A., G.G. and M.V.; visualization, A.B., G.C. and I.B.; supervision, M.V. and G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Fund for the Italian Electrical System under the contract agreement "Accordo di Programma Mission Innovation 2021–2024—Project MISSION (POA Smart Grid)" between ENEA and the Ministry of the Environment and Energetic Safety (MASE).

Data Availability Statement: The dataset related to PV generation is available on request from the authors.

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

References

1. Kumi, E.N.; Mahama, M. Greenhouse gas (GHG) emissions reduction in the electricity sector: Implications of increasing renewable energy penetration in Ghana's electricity generation mix. *Sci. Afr.* **2023**, *21*, e01843. [CrossRef]
2. Tol, R.S.J. A meta-analysis of the total economic impact of climate change. *Energy Policy* **2024**, *185*, 113922. [CrossRef]
3. Rezai, A.; Taylor, L.; Foley, D. Economic Growth, Income Distribution, and Climate Change. *Ecol. Econ.* **2018**, *146*, 164–172. [CrossRef]
4. Adinolfi, G.; Ciavarella, R.; Graditi, G.; Ricca, A.; Valenti, M. A Planning Tool for Reliability Assessment of Overhead Distribution Lines in Hybrid AC/DC Grids. *Sustainability* **2021**, *13*, 6099. [CrossRef]
5. Vinothine, S.; Arachchige, L.N.W.; Rajapakse, A.D.; Kaluthanthrige, R. Microgrid Energy Management and Methods for Managing Forecast Uncertainties. *Energies* **2022**, *15*, 8525. [CrossRef]
6. Gestore dei Servizi Energetici. Rapporto Statistico Solare Fotovoltaico 2022. Available online: https://www.gse.it/documenti_site/Documenti%20GSE/Rapporti%20statistici/GSE%20-%20Solare%20Fotovoltaico%20-%20Rapporto%20Statistico%202022.pdf (accessed on 26 February 2024).
7. Buonanno, A.; Caliano, M.; Di Somma, M.; Graditi, G.; Valenti, M. A Comprehensive Tool for Scenario Generation of Solar Irradiance Profiles. *Energies* **2022**, *15*, 8830. [CrossRef]
8. Yang, D.; Wang, W.; Gueymard, C.A.; Hong, T.; Kleissl, J.; Huang, J.; Perez, M.J.; Perez, R.; Bright, J.M.; Xia, X.; et al. A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. *Renew. Sustain. Energy Rev.* **2022**, *161*, 112348. [CrossRef]
9. Graditi, G.; Buonanno, A.; Caliano, M.; Di Somma, M.; Valenti, M. *Machine Learning Applications for Renewable-Based Energy Systems*; EAI/Springer Innovations in Communication and Computing; Springer: Cham, Switzerland, 2023; Volume Part F665, pp. 177–198. [CrossRef]
10. Bekhit, R.; Bianco, G.; Delfino, F.; Ferro, G.; Noce, C.; Orrù, L.; Parodi, L.; Robba, M.; Rossi, M.; Valtorta, G. A platform for demand response and intentional islanding in distribution grids: The LIVING GRID demonstration project. *Results Control Optim.* **2023**, *12*, 100294. [CrossRef]
11. Climate Models | NOAA Climate.gov. Available online: https://www.climate.gov/maps-data/climate-data-primer/predicting-climate/climate-models (accessed on 26 February 2024).
12. Fuoco, D.; Mendicino, G.; Senatore, A.; Balog, I.; Caputo, G.; Spinelli, F.; Lepore, M.; Franconiero, D.; Mautone, P.; Oliviero, M. Modelli Previsionali di Producibilità: Ambiti Applicativi. Rapporto Tecnico di Ricerca Industriale D5.3a. Available online: http://www.comesto.eu/wp-content/uploads/2020/11/D5.3a_Modelli-previsionali-di-producibilit%C3%A0_ambiti-applicativi.pdf (accessed on 26 February 2024).
13. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Third Edition—IEA-PVPS. Available online: https://iea-pvps.org/key-topics/best-practices-handbook-for-the-collection-and-use-of-solar-resource-data-for-solar-energy-applications-third-edition/ (accessed on 26 February 2024).
14. Ledmaoui, Y.; El Maghraoui, A.; El Aroussi, M.; Saadane, R.; Chebak, A.; Chehri, A. Forecasting solar energy production: A comparative study of machine learning algorithms. *Energy Rep.* **2023**, *10*, 1004–1012. [CrossRef]
15. Gupta, P.; Singh, R. PV power forecasting based on data-driven models: A review. *Int. J. Sustain. Eng.* **2021**, *14*, 1733–1755. [CrossRef]

16. Visser, L.; AlSkaif, T.; van Sark, W. Benchmark analysis of day-ahead solar power forecasting techniques using weather predictions. In Proceedings of the 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC), Chicago, IL, USA, 16–21 June 2019; pp. 2111–2116.
17. Theocharides, S.; Theristis, M.; Makrides, G.; Kynigos, M.; Spanias, C.; Georghiou, G.E. Comparative Analysis of Machine Learning Models for Day-Ahead Photovoltaic Power Production Forecasting. *Energies* **2021**, *14*, 1081. [CrossRef]
18. Scott, C.; Ahsan, M.; Albarbar, A. Machine learning for forecasting a photovoltaic (PV) generation system. *Energy* **2023**, *278*, 127807. [CrossRef]
19. Kallio, S.; Siroux, M. Photovoltaic power prediction for solar micro-grid optimal control. *Energy Rep.* **2023**, *9*, 594–601. [CrossRef]
20. Dutta, S.; Li, Y.; Venkataraman, A.; Costa, L.M.; Jiang, T.; Plana, R.; Tordjman, P.; Choo, F.H.; Foo, C.F.; Puttgen, H.B. Load and Renewable Energy Forecasting for a Microgrid using Persistence Technique. *Energy Procedia* **2017**, *143*, 617–622. [CrossRef]
21. Gaboitaolelwe, J.; Zungeru, A.M.; Yahya, A.; Lebekwe, C.K.; Vinod, D.N.; Salau, A.O. Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison. *IEEE Access* **2023**, *11*, 40820–40845. [CrossRef]
22. Tayab, U.B.; Yang, F.; Metwally, A.S.M.; Lu, J. Solar photovoltaic power forecasting for microgrid energy management system using an ensemble forecasting strategy. *Energy Sources Part A Recover. Util. Environ. Eff.* **2022**, *44*, 10045–10070. [CrossRef]
23. Teferra, D.M.; Ngoo, L.M.; Nyakoe, G.N. Fuzzy-based prediction of solar PV and wind power generation for microgrid modeling using particle swarm optimization. *Heliyon* **2023**, *9*, e12802. [CrossRef] [PubMed]
24. Mayer, M.J. Benefits of physical and machine learning hybridization for photovoltaic power forecasting. *Renew. Sustain. Energy Rev.* **2022**, *168*, 112772. [CrossRef]
25. Ogliari, E.; Dolara, A.; Manzolini, G.; Leva, S. Physical and hybrid methods comparison for the day ahead PV output power forecast. *Renew. Energy* **2017**, *113*, 11–21. [CrossRef]
26. Niccolai, A.; Dolara, A.; Ogliari, E. Hybrid PV Power Forecasting Methods: A Comparison of Different Approaches. *Energies* **2021**, *14*, 451. [CrossRef]
27. Fabozzi, S.; Graditi, G.; Valenti, M. Techno-economic design of a smart multienergy microgrid. In Proceedings of the 2022 AEIT International Annual Conference (AEIT), Rome, Italy, 3–5 October 2022.
28. Buonanno, A.; Caputo, G.; Balog, I.; Adinolfi, G.; Pascarella, F.; Leanza, G.; Fabozzi, S.; Graditi, G.; Valenti, M. Combined Machine Learning and weather models for photovoltaic production forecasting in microgrid systems. In Proceedings of the 2023 International Conference on Clean Electrical Power (ICCEP), Santa Margherita Ligure, Italy, 27–29 June 2017; pp. 216–222.
29. WRF Model Users Site. Available online: https://www2.mmm.ucar.edu/wrf/users/ (accessed on 26 February 2024).
30. WRF Community. Weather Research and Forecasting (WRF) Model, UCAR/NCAR. 2000. Available online: https://www2.mmm.ucar.edu/wrf/users/ (accessed on 26 February 2024).
31. Global Forecast System (GFS) | National Centers for Environmental Information (NCEI). Available online: https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast (accessed on 26 February 2024).
32. Larson, D.P.; Nonnenmacher, L.; Coimbra, C.F. Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest. *Renew. Energy* **2016**, *91*, 11–20. [CrossRef]
33. CEI 82-25: 2008 Guide for Design and Installation of Photovoltaic. Available online: https://www.intertekinform.com/en-au/standards/cei-82-25-2008-319110_saig_cei_cei_735215/ (accessed on 26 February 2024).
34. Dobos, A.P. PVWatts Version 5 Manual. 2014. Available online: www.nrel.gov/publications (accessed on 26 February 2024).
35. Murphy, K.P. *Probabilistic Machine Learning: An Introduction*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2022.
36. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
37. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
38. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 3149–3157. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf (accessed on 26 February 2024).
39. Herzen, J.; Lässig, F.; Piazzetta, S.G.; Neuer, T.; Tafti, L.; Raille, G.; Van Pottelbergh, T.; Pasieka, M.; Skrodzki, A.; Huguenin, N.; et al. Darts: User-Friendly Modern Machine Learning for Time Series. *J. Mach. Learn. Res.* **2022**, *23*, 1–6. Available online: http://jmlr.org/papers/v23/21-1177.html (accessed on 26 February 2024).