Benchmarking physics-informed machine learning-based short term PV-power forecasting tools

Pombo, Daniel Vázquez; Bacher, Peder; Ziras, Charalampos; Bindner, Henrik W.; Spataru, Sergiu V.; Sørensen, Poul E.

Published in:
Energy Reports

Link to article, DOI:
10.1016/j.egyr.2022.05.006

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Energy Reports 8 (2022) 6512–6520

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

Benchmarking physics-informed machine learning-based short term PV-power forecasting tools

Daniel Vázquez Pombo [a,b,*], Peder Bacher [c], Charalampos Ziras [a], Henrik W. Bindner [a], Sergiu V. Spataru [d], Poul E. Sørensen [e]

[a] Department of Electrical Engineering, Technical University of Denmark (DTU), Frederikborsvej 399, Roskilde, 4000, Denmark
[b] R&D Strategic Development, Vattenfall AB, Evenemangsgatan 13C, Solna, 169 56, Sweden
[c] Department of Computer Science, Technical University of Denmark (DTU), Frederikborsvej 399, Roskilde, 4000, Denmark
[d] Photovoltaic Materials and Systems, Technical University of Denmark (DTU), Frederikborsvej 399, Roskilde, 4000, Denmark
[e] Department of Wind Energy, Technical University of Denmark (DTU), Frederikborsvej 399, Roskilde, 4000, Denmark

ARTICLE INFO

Article history:
Received 19 December 2021
Received in revised form 11 April 2022
Accepted 3 May 2022
Available online xxxx

Keywords:
Machine learning
Physics informed
Solar PV
PV power forecasting

ABSTRACT

Uncertainty is one of the core challenges posed by renewable energy integration in power systems, especially for solar photovoltaic (PV), given its dependence on meteorological phenomena. This has motivated the development of numerous forecasting tools, recently focused on physics informed machine learning (ML). Virtually, every paper claims to provide better accuracy than the previous, yet the replicability of these studies is very low, motivating unfair, or erroneous comparisons. This paper reviews and compares the most relevant ML-methods identified in the literature (Random Forest, Support Vector Regression, Convolutional Neural Networks, Long–Short Term Memory and a Hybrid of the last two) with two statistical methods: persistence and an Semi-Parametric Auto-Regressive model. Furthermore, we propose a methodology to integrate a PV-performance model in ML models to forecast power several hours ahead with 5-min resolution. A basic dataset including power production and meteorological measurements is expanded with physics-informed features that capture the relationship between weather and PV operational state, while keeping strong correlation towards the intrinsic feature. This allows the models to learn about the physical interdependence of different features, potentially yielding a higher accuracy than conventional methods. Then, we also propose a physics-informed feature selection to tighten the search-space of the best performer. A case study of a PV array in Denmark is used for validation using both the original and expanded datasets. Results show how the best ML models consistently used physics-informed features in all cases.

1. Introduction

Due to various economic and environmental concerns, power systems worldwide undergo a transition from traditional fossil-fuel-based technologies towards non-polluting alternatives. However, generation fluctuation caused by the inherent volatility of renewable energy sources (RES) has increased grid operation complexity. System operators (SOs) are increasingly demanding ancillary services from RES, which amplifies the need for accurate short time predictions of the available renewable power. Forecast horizons of a few minutes ahead are critical for control applications and grid support capabilities, such as frequency response or voltage regulation (Ghosh et al., 2017). Longer horizons are used in energy management systems (EMSs) to schedule controllable loads, such as heat pumps or electric vehicles, according to fluctuating market prices (Mohandes et al., 2021). In fact, forecasting has been identified as the most promising technology aiming for an increased deployment of renewables, as it is the key enabler for a variety of other smart grid applications (Wang et al., 2020a). For example, Cordova et al. (2018) showed how marginal improvements in forecasting accuracy of intra-day RES availability significantly increased penetration over the year due to better scheduling, along with reducing system operation cost.

Solar power is widely deployed at all levels of the power system in a very geographically-distributed manner. Most of the available literature targeting forecasting tools for solar power focuses on predicting irradiance, as this is the most critical parameter when estimating solar generation (Wang et al., 2020a,b; Cordova et al., 2018; Ahmed et al., 2020). Generally, irradiance values are then fed to a simple black-box model of a PV unit to cast a power prediction. This approach is not only more prone to errors

* Corresponding author at: Department of Electrical Engineering, Technical University of Denmark (DTU), Frederikborsvej 399, Roskilde, 4000, Denmark.
E-mail address: dvapo@dtu.dk (D.V. Pombo).

https://doi.org/10.1016/j.egyr.2022.05.006

than direct power prediction, but it also hinders transparency, as stated by several authors (Blaga et al., 2019; Ahmad and Chen, 2020). Nevertheless, beyond irradiance, PV power depends on other meteorological factors, such as wind speed or temperature, which play a fundamental role in defining their operating efficiency along with some manufacturing-related parameters.

Motivated by increasing data availability, ML models are increasingly used as forecasting tools. These models, despite their conceptual simplicity, often yield great results. They use historical data from a particular site along with other features to cast a prediction for the desired horizon and resolution. ML models try to find linear and non-linear relationships in the data and use them in the prediction process. Popular ML methods employed in solar forecasting are: random forest (RF), support vector machine (SVM), and artificial neural network (ANN) (Ahmed et al., 2020; Sobri et al., 2018).

Despite the need for high resolution short term power forecast horizons, the vast majority of the available work employing ML focuses on hourly estimations of solar irradiance for day-ahead scenarios which are then used to compute PV power (Wang et al., 2019; Sobri et al., 2018; Voyant et al., 2017; Ahmad and Chen, 2020; Lotfi et al., 2020). This is partly explained by the fact that forecasting with lower resolution (e.g., in the order of one hour) is comparatively easier as the predictions are less affected by noisy readings due to the smoothing effects of averaging. In addition, many published papers evaluate their models in a biased manner, e.g. by predicting during winter periods with lower variability and daylight hours, which overestimates accuracy (Blaga et al., 2019). These gaps are explained by the limited availability of large enough data-sets with high resolution data on the one hand. But also, by a bias towards the non-publication of negative results (Blaga et al., 2019). However, multi-hour horizon forecasting with 5-min resolution is interesting for energy management system of isolated power systems such as microgrids, as this period allows to configure adaptable stochastic dispatchers (Mohandes et al., 2021).

Among the available literature, it is worth mentioning the case of Liu et al. (2015), which introduces numerical weather prediction (NWP) and historical data into an ANN to predict PV power, and Persson et al. (2017) uses gradient boosted regression trees for six steps ahead PV power prediction. More recently, Feng et al. (2019) used SVM for one hour ahead irradiance forecasting. Then, Jang et al. (2016) claims to predict solar power, but actually obtains irradiance and cloud coverage by feeding satellite images to an SVM. Then, Zhang et al. (2019) presents an ensemble method for day-ahead PV power forecasting using NWP and recorded measurements. Furthermore, (Mohandes et al., 2021) uses deep learning with ANN to predict irradiance. On the other hand, Gigoni et al. (2018) criticizes the lack of studies focused on comparing several methods on the same dataset. More recently, an intrinsic PV power forecasting tool is included in Hafiz et al. (2020) as part of a stochastic optimization problem. Due to the low accuracy obtained with an LSTM on the 1-minute scale, they are forced to use 15-minute intervals with a single step ahead. Then, in Zhen et al. (2020) sky images are used to estimate irradiance over a wide area. The authors compare the performance of several different ANN structures. Similarly, sky images are used in Fu et al. (2021) to capture cloud patterns, where they use them to forecast irradiance with a convolutional auto-encoder. Lastly, in Yan et al. (2021) a frequency-domain decomposition is applied to power measurements and fed them to a CNN. The actual predictions require undoing said decomposition afterwards. They claim that accuracy is improved by avoiding physical dependence. It should be highlighted that all (Liu et al., 2015; Persson et al., 2017; Feng et al., 2019; Jang et al., 2016; Zhang et al., 2019; Mohandes et al., 2021; Gigoni et al., 2018) used hourly resolution, except (Jang et al., 2016), which complain about the lack of off-the-shelf intra-hour solar predictors.

Based on the review presented in this section, we can point to a few unaddressed research gaps regarding ML-based PV power forecasts, which our present work addresses. First, we propose a physics-informed dataset expansion and feature selection process to improve the performance of ML-based PV forecasting techniques. This simple and effective approach reinforces the learning potential of any ML-model for potentially all PV farms as it uses basic datasheet parameters and commonly available meteorological data. Through extensive validations we show how these adaptations can enhance accuracy in a wide variety of ML models, significantly reducing training time by compressing the feature search-space. Second, we address the lack of systematic and detailed comparison between different ML models, by carefully assessing accuracy under varying conditions, in the generally more challenging summer months. We further benchmark the performance of each ML method with both a naive persistence forecast and a more advanced semi-parametric auto-regressive (SPAR) model, which is proven to present high accuracy (Bacher et al., 2013). Thus, we are better able to generalize regarding the performance of different ML approaches and feature selection methods. Then, to the knowledge of the authors there is no work available focusing on the more challenging task of performing high resolution forecasting over a relatively long horizon of 60 steps ahead justified by state of the art energy management systems for isolated grids. Finally, we release the dataset and code in Open Access to facilitate the replicability and transparency of the study, aiming to serve as benchmark for future research in this area (Pombo, 2022; Pombo et al., 2022b). It should be mentioned that, this study does not include NWP or imaging as these were not available (Bracale et al., 2017).

2. Theoretical background

This section introduces the two core aspects of the theoretical background, namely: ML-based time series forecasting, and PV-performance modeling.

2.1. Machine learning-based forecasting models

The available forecasting approaches are usually divided in physical, statistical, and ML-based. Multitude of researchers have analyzed their performance for solar power forecasts in the past few years. Given the growing data availability, results are conclusive in identifying ML methods as the most accurate and promising ones for hourly resolution (Ahmed et al., 2020). However, there are no concrete conclusions for higher resolutions. In general, ML methods rely on the ability of artificial intelligence to learn patterns and complex relationships from historical measurements. The computational needs of these methods are typically high, as they must be iteratively trained; when the model has been trained, a prediction can be cast in a fraction of a second. We have identified five ML methods with good prospects of delivering accurate forecasts from the available scientific literature. In this section, we formally introduce them at a high level. The interested reader is referred to Pombo et al. (2022a), Bacher et al. (2013) for a more thorough presentation.

2.1.1. Random forest

Consisting of groups of simple if-then-else rules, they can be applied in both classification and regression tasks. RF use entropy as a metric to distribute the different values into subsets; defined according to another metric called information gain. Which is used to split the training data into different subsets. Once trained, RF cast predictions as an ensemble, or weighted average of each tree's predictions.

2.1.2. Support vector machine

Similarly to RF, SVM are a group of ML methods based on risk minimization of the training data-set originally developed for classification (Vapnik, 2013). When employed in regression applications such as time-series forecasting, they are known as support vector regressor (SVR). Briefly, SVM translate any input data into a higher dimensional space, subsequently applying linear regression.

2.1.3. Artificial neural networks

ANNs are composed by neurons grouped in layers arranged in purpose-dependent architectures. Information flows within the model according to an activation function that computes and forwards the prediction. Recurrent ANNs are a special subset, which includes long–short term memory (LSTM) models, which are able to capture temporal sequences, thus specializing in time-series forecasting. Convolutional neural network (CNN) are another subset targeting geographically-distributed information, usually employed in image analysis. Lately, hybrid CNN-LSTM structures have been reported as best performers in day-ahead hourly solar forecasting. Theoretically, these models are able to capture simultaneously spatial and temporal relationships. LSTM, CNN and the CNN-LSTM hybrid are included in this research to evaluate the forecasting behavior for short horizons with high sampling rates. However, our parallel research on 1, 2 and 3 day-ahead hourly PV power forecasting showed how, with proper feature selection and tuning, ANN can be outperformed by other methods such as RF and SVR (Pombo et al., 2022a).

Long–Short Term Memory Neural Network: LSTM neural networks are designed to reduce the incidence of the vanishing gradient problem. This well-established phenomenon causes regular ANN to stagger during training due to small convergence gradients in the optimization solution. While in LSTM, cell-vectors are designed to forget certain parts of stored memory when adding new information, thus minimizing the appearance of small gradients. Theoretically, LSTM algorithms should outperform other simpler regression methods such as RF or SVM due to their capacity of capturing temporal dependencies between observations (Mussumeci and Coelho, 2020). In fact, Wang et al. (2019) reviewed deep learning-based forecasting of renewable energy, concluding that LSTM outperforms other more traditional methods for large enough datasets. However, other works have pointed towards a bias dealing with unfair comparisons between methods; for example, using state-of-the-art tuning for some methods, while naive methods for others (Blaga et al., 2019).

Convolutional Neural Networks: CNNs are designed to process spatially distributed data, whose structure consists of stacking different layers. At least one of them performs convolutions which are integral blending operations on two functions expressing the overlap between them, which effectively filters the input data. Different activation functions are used to increase the non-linearity of the network, while over-fitting is avoided by down-sampling the input data using pooling layers.

Hybrid CNN-LSTM: The hybrid CNN-LSTM concept is expected to effectively capture both spatial and temporal features simultaneously in the same model. Certain authors use LSTM at the input to ensure they capture temporal correlation, while leaving CNN at the core as an output filter, while others reverse the order. To the best of our knowledge, there is no consensus regarding a recommended approach.

2.2. Generic PV-performance model

There are a number of generic PV-performance models available in the scientific literature to estimate the operational state of the array. These range in complexity and precision. In this paper we used an adaptation of the 'power–temperature coefficient PV model' proposed by King et al. (2004). The main advantages of this well-established method is its good accuracy, simplicity and the possibility of building it using basic data-sheet parameters and meteorological measurements, thus making it suitable to be computed for virtually any PV plant. The only required field measurements are irradiance, wind speed and air temperature. The rest of the information is collected from the PV and inverter datasheets. First, the module's temperature (T_m) is estimated as per Eq. (1), where T_a, W_s, a and b represent air temperature, wind speed, and two material-dependent coefficients which can be extracted from King et al. (2004), respectively. Note that the following equations are time dependent, however, the time index is omitted for notational simplicity.

$$T_m = T_a + E_{POA}\ \exp(a + b\ W_s) \tag{1}$$

E_{POA} corresponds to the total solar irradiance incident on a PV panel, which is computed based on the sun's position, array's angle, shadowing and surroundings' reflection. However, in our case, the irradiance sensor is placed in the same orientation and inclination as the array, therefore, we consider that measurement equivalent to E_{POA}. Cell temperature (T_c) is calculated as follows:

$$T_c = T_m + \frac{E_{POA}}{E_{STC}}\Delta T, \tag{2}$$

where E_{STC} stands for the reference irradiance (1000 W/m^2) and ΔT represents the difference between module and cell temperatures. These values are again material-dependent taken from the datasheet. Furthermore, the array's DC output power is obtained with:

$$P_{DC} = N_s N_p \frac{E_{POA}}{E_{STC}} P_{mp,STC}\left[1 + \gamma_{mp}\,(T_c - T_{STC})\right], \tag{3}$$

where $P_{mp,STC}$, T_{STC}, γ_{mp}, N_s and N_p stand for the panel's peak power and temperature measured under standard conditions, the normalized temperature coefficient of peak power, the number of panels connected in series and parallel, respectively. Lastly, the inverter's efficiency is applied in order to obtain the estimated output power of the PV system (P_{AC}) as

$$P_{AC} = \begin{cases} \eta_{max}\ P_{DC}, & \text{if } \eta_{max}\ P_{DC} \le\ P_{ACmax} \\ P_{ACmax}, & \text{if } \eta_{max}\ P_{DC} >\ P_{ACmax} \end{cases} \tag{4}$$

This value can be used not only to identify curtailment periods, but also inaccurate active power measurements and maintenance periods simply by locating substantial deviations (above 20%), and substituting the incorrect value with P_{AC}. This model allows to expand the number of available features in the original dataset. It captures the relationship between the meteorological measurements and the PV production, effectively introducing physical dependence between the features and simplifying the training process.

3. Methodology

In order to assess the benefit of applying a physics-informed approach to improve PV forecasting, we follow the steps outlined in Fig. 1. We begin by pre-processing the raw measurements to obtain a clean dataset, which we expand with additional features during the data-mining stage (Section 3.1). Subsequently, we perform feature selection (Section 3.2); we apply both a Pearson-based and physics informed-based feature selection to validate our hypothesis that the latter improves performance. This selection defines different search-spaces for ML model tuning and optimization (Section 3.3). Lastly, in the evaluation stage (Section 3.4) we compare the results for each ML model and against a number of benchmarks.

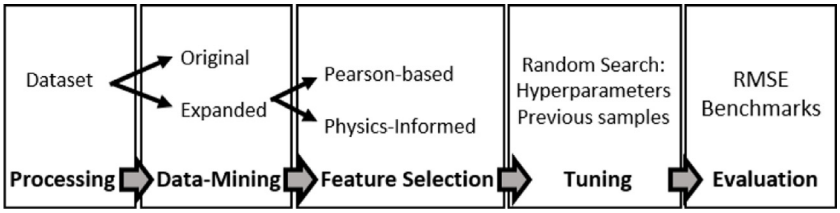

Fig. 1. Methodology overview.

3.1. Data-mining

After performing data cleaning on the available raw measurements (removing outliers, aligning the time-series, etc.), we obtain the *original* dataset. Next, we expand it by include the PV-performance model, described in Section 2.2. This leads to the inclusion of four new features: P_{DC}, P_{AC}, T_m, and T_c. Second, we introduce two features related to the time of the day: an indicator of the 5-min step (M_{day}) one of the hour (H_{day}). The resulting dataset is called *expanded*.

3.2. Feature selection

The objective of this step is to limit the size of the search-space, thus speeding up the identification of the best performer. The common practice in feature selection is to consider large values of Pearson correlation, with the intrinsic features as a marker of how potentially useful a metric will be during learning. Pearson correlation coefficient is a measure of linearity and is given by

$$R_{X,Y} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}. \tag{5}$$

Features with Pearson values closer to the extremes are usually a good starting point during feature selection, since strong linearity is easily captured by ML methods. However, the opposite claim does not hold. Most available research focused on time-series forecasting points to ML's ability of discovering non-linear relationships and learning from them. Despite this, such features are often intentionally not included in the majority of papers (Harrou and Sun, 2020). A simple explanation is the non-existence of a unique non-linear dependence metric, as it can be assessed solely for individual relations, such as exponential, hyperbolic, etc. An additional problem regarding feature selection based on linearity is to establish a threshold allowing to discriminate between 'useful' and 'non-useful' features. To the best of our knowledge, there is no consensus, but based on our survey of relevant literature (Liu et al., 2015; Persson et al., 2017; Feng et al., 2019; Jang et al., 2016; Zhang et al., 2019; Gigoni et al., 2018; Hafiz et al., 2020; Zhen et al., 2020; Fu et al., 2021; Yan et al., 2021), we can point to a value in the vicinity of |0.7|. Therefore, Pearson is a useful tool during the preliminary evaluation of a feature's relevance, however, it is not perfect. In fact, we will show how it fails to identify some features with high physical relevance.

Due to these shortcomings, we apply two different feature selection methods. One is based on Pearson coefficients, and the other on physical relevance. The objective is to test whether there is a clear gain from limiting the search-space based on physical knowledge or if we risk discarding potentially useful feature combinations. We demonstrate in the case study how the latter feature selection is applied.

3.3. Training, tuning & performance metrics

Once features have been selected, the dataset is divided in training, validation and testing sets as 70%, 20% and 10% of the samples, respectively. Then, we define a search space covering different feature combinations, number of previous samples (Pre), and relevant hyperparameters for each of the investigated models. In this way, we are able to perform an unbiased grid search to find the best possible performer. Mean absolute error (MAE) was selected as the loss function during training, as it has the advantage of applying uniform significance to all samples.

$$MAE_h = \frac{\sum_i^n \left|P_{i,h} - \hat{P}_{i,h}\right|}{n}, \quad \forall h \in \mathcal{H}, \tag{6}$$

On the other hand, root mean square error (RMSE) was selected as the accuracy assessment metric during testing, in order to penalize large deviations. $RMSE_h$ represents the RMSE for all tested samples $i \in \{1, \ldots, n\}$ and for prediction step h.

$$RMSE_h = \sqrt{\frac{\sum_i^n \left(P_{i,h} - \hat{P}_{i,h}\right)^2}{n}}, \quad \forall h \in \mathcal{H}, \tag{7}$$

where $P_{i,h}$ stands for the ith observation for prediction step h, and $\hat{P}_{i,h}$ stands for the corresponding. We rank the different models by using the average RMSE over the whole horizon ($\overline{RMSE}$), as in Ahmad and Chen (2020); but also the average MAPE ($\overline{MAPE}$). In addition, the residual sum of squares (RSS) and the coefficient of determination (R^2) are used to further gain insight into the performance of the models, rank the methods, and act as tie-breaker.

$$RSS = \sum_i^n \left|P_{i,h} - \hat{P}_{i,h}\right|^2, \quad \forall h \in \mathcal{H}, \tag{8}$$

$$R^2 = \frac{RSS}{\sum_i^n \left|P_{i,h} - \overline{P}_{i,h}\right|}, \quad \forall h \in \mathcal{H}, \tag{9}$$

3.4. Evaluation

The evaluation consists of comparing the trained ML models between them and against three benchmarks: intrinsic, persistence and SPAR. The objective is to find the best performers on each dataset. However, in a practical implementation it is important to pay attention to the computational needs of a particular model's training process. While there might be a small accuracy improvement, an increase in training time or allocated memory might tilt the balance in a different direction.

3.4.1. Intrinsic

This simply consists on the most basic ML model that can be built by including only the intrinsic feature, i.e., past measurements of PV power.

3.4.2. Persistence

Persistence is the simplest forecasting method. It assumes that a given system state in time t will remain unchanged until $t+H$, where H represents the forecasting horizon. It can be formally expressed as.

$$\hat{P}_{(t+h|t)} = P_t, \ \forall\, h \ \in \ \mathcal{H}, \tag{10}$$

where P_t is the measured active power at time t, and $\hat{P}_t$ the predicted value.

3.4.3. Semi-parametric auto-regressive model

SPAR techniques have been widely used in statistical modeling (Hastie et al., 2009; Bacher et al., 2013). SPAR is an auto-regressive model whose parameters change as a smooth function of the time of day, using basis splines to obtain these functions. The model includes a diurnal curve intercept and an auto-regressive term as:

$$\hat{P}_{(t+h|t)} = f_{\text{bs}}(t_{\text{day}}) + f_{\text{bs}}(t_{\text{day}})P_t, \ \forall\, h \ \in \ \mathcal{H}, \tag{11}$$

where $f_{\text{bs}}()$ is a spline function adapting to the time of the day or t_{day}. The model fits the data in two stages, as described in Bacher et al. (2013): first the basis splines are calculated with spline functions set to 10 degrees of freedom, resulting in 20 separate time-series, then regression is performed. This process is applied separately for all horizons with a recursive least squares scheme, such that the coefficients adapt to data over time. This model does not rely on other data than the latest power observation, but could be extended to include other variables and functional relationships to improve performance.

4. Case study

Our case study is based on a small-scale PV system situated in Denmark. We considered a high resolution of 5-minutes and a horizon of 60 steps ahead, equal to 5 h. We chose this particular resolution and horizon due to their relevance in stochastic energy management (Hafiz et al., 2020).

4.1. Data description

Measurements correspond to a 10 kW PV array in SYSLAB, a laboratory focused on distributed energy resources, belonging to the Danish Technical University (DTU). In addition, the recordings of a meteorological mast are included. The mast is located 10 meters away from the PV, thus presenting nearly identical conditions. Measurements are transmitted via TCP/IP to a central data storage with second resolution, which are then averaged to build the 5-minute dataset. We selected a 15 month period from June 1st 2019 to August 31st of 2020 (131,616 5-minute samples). The recordings are: active power on the inverter AC side, global horizontal irradiance (GHI), plane of array irradiance (E_{POA}), air temperature, humidity, wind speed and direction. The latest calibration estimated the measurement error as $\pm 1\%$ for all metrics, with the exception of humidity ($\pm 0.6\%$) and irradiance ($\pm 3\%$). The array is tilted 60° and partially shadowed by vegetation, a wind turbine and a building. Due to research activity in SYSLAB, the PV arrays are seldom curtailed. Thus, the recorded active power is not always equivalent to the available power; this issue is addressed during data pre-processing.

Table 1
SYSLAB PV system specifications.

Parameter	Value
a	−3.56
b	−0.075
ΔT	3
$P_{\text{mp,STC}}$	200 W
γ_{mp}	−0.00478
N_{s}	18
N_{p}	2
P_{ACmax}	10 kW
PV manufacturer	Schüco
Inverter	SMA SunnyTripower 10000TL

4.2. Dataset definition

The original dataset includes the measured active the PV power (P_{Solar}), T_{a}, humidity, W_{s}, wind direction (W_{dir}), GHI and E_{POA}. The expanded includes the original dataset plus P_{DC}, P_{AC}, T_{m}, T_{c}, M_{day} and H_{day}. Note that these are computed using the data presented in Table 1. The heading of Table 2 presents the different features, which are identified by an ordinal number. Following that logic, intrinsic includes only feature 0, original from 0 to 6 and expanded all of them.

4.3. Pre-processing

The measurement system is asynchronous, therefore, the first step is to ensure the alignment of the different recordings. This was done by generating all possible timestamps for the studied period and match them with the available data, leaving the rest with NAN values (less than 0.01% of the dataset). Then, P_{AC} was used to fill any missing power measurement. Subsequently, outliers were identified as values outside a reasonable bandwidth. Thereafter, the dataset was split into different sets in the aforementioned proportions for training, validation and testing. Then, values were scaled from 0 to 1 to avoid cross-contamination. We followed Alanazi's procedure Alanazi and Khodaei (2016) to stationarise the data by using clear sky irradiance to remove the offset for GHI, and E_{POA}, effectively detrending it. Then, night time periods can be removed by knowing the sunset and surise time of each day, and Al-Sadah's polynomical model can be used to detrend and normalize. Other metrics such as humidity, W_{s}, or W_{dir} presented negative p-values in the Augmented Dickey-Fuller test, hence allowing to reject the H1 hypothesis and confirming their stationarity.

4.4. Search space definition

The correlation between the different features are presented in Table 2. From the original dataset, E_{POA} and GHI have a very strong correlation with measured active PV power (P_{Solar}). Air humidity presents a relatively strong anti-correlation, but is below the 0.7 threshold. However, we decided to include humidity as part of the search space, as it directly affects heat transmission given the thermodynamic dependence of PV efficiency upon temperature.

T_{a} and especially W_{s} present very low correlation. However, after including them in (1) and (2), the new features T_{m} and T_{c} present high correlation and are included as potential features in the expanded dataset. This highlights the importance of capturing non-linear relationships as reformulated features in the data-mining stage. ML models should be able to discover these relationships, given both a large enough dataset and sufficient number of iterations. However, by reformulating these metrics, the learning task is effectively eased. P_{DC} and P_{AC} are also potential features because they show perfect linear correlation. Based

Table 2
Feature Correlation. Boldness indicates original dataset.

	0-T_a [C]	**1-Humidity [%]**	**2-W_s [m/s]**	**3-W_{dir} [deg]**	**4-GHI [kW/m^2]**	**5-E_{POA}**	7-P_{DC}	8-P_{AC}	9-T_m	10-T_c	11-M_{day} [0-287]	12-H_{day} [0-23]
6-P_{Solar}	0.38	−0.57	0.14	0.2	0.93	1	1	1	0.76	0.79	−0.04	−0.04

Table 3
Features included in each dataset search-space.

	0-T_a [C]	1-Humidity [%]	W_s [m/s]	3-W_{dir} [deg]	4-GHI [kW/m^2]	5-E_{POA}	6-P_{Solar}	7-P_{DC}	8-P_{AC}	9-T_m	10-T_c	11-M_{day} [0-287]	12-H_{day} [0-23]
Original	X	X	X	X	X	X	X						
Exp-Pe		X			X	X	X	X	X	X	X		
Exp-Py		X				X	X		X		X		X

on the Pearson correlation we thus choose humidity, GHI, E_{POA}, P_{DC}, P_{AC}, T_m and T_c as potential features. We refer to this feature set as *Expanded-Pearson*.

Next, we present a physics-aware approach in defining a feature set, which is named *Expanded-Physics*, and in the results section we show how this improves the performance of all ML models. It is worth noting how both M_{day} and H_{day} present no linear relationship, as their correlation is practically zero, which would cause to discard them. However, we are intuitively aware that time of day directly affects PV production as it captures the day/night cycle, despite the near zero linear correlation. We choose H_{day} over M_{day} due to the reduced number of values it can take, which effectively smooths out oscillations and noise; this approach has proved useful in a previous study for wind power (Pombo et al., 2021). As a next step, we can reduce the number of features based on their physical relevance by discarding those containing redundant information. GHI is discarded in favor of E_{POA}, as we know both metrics contain the same information (irradiance), and the latter is more representative of the system's behavior as it captures the solar energy hitting the PV array. Similarly, we drop P_{DC} in favor of P_{AC}, as it also captures power conversion efficiency. Additionally, even if T_a and W_s show high correlation, they are discarded since they are used to compute T_c. Then, T_m and T_c represent temperature of module and cell. The first is then discarded as cell temperature is a more precise marker of the PV's dynamical state.

Table 3 summarizes which features are included on each search-space.

4.5. Model tuning

The number of previous samples and hyperparameters were tuned in two runs: the first one been quite broad, and the second with reduced possibilities based on the conclusions from the first. By performing a two-fold grid search we tighten the search-space, thus improving the chances of finding good performers among the trained models. A practical implementation of a forecasting system should choose a single datasets and focus on tuning hyperparameter configurations and number of previous samples, as their purpose is to minimize the search space in order to speed up training and reduce computational costs. In this way, the possible number of previous samples to be used was set from 0 to 120 in steps of 10 during the first run, while in the second only from 0 to 50 in steps of 5. It should be noted how, the shape of the previous samples is only modified in those models requiring tensors, that is ANN.

Hyperparameters depend on the ML method. RF could take any number of trees from 100 to 5000 in steps of 100. However, this was further tightened to 400 to 3000 in a second run of the search space. SVM could use rbf, polynomial (from 1 up to 5 degrees) and sigmoid kernels, with the C parameter as an integer from 1 to 7, epsilon from 0.1 to 1 in 0.1 steps. However, we limited the second run to the rbf kernel only. Regarding ANN in general, we found that batch sizes of 16 and 32 were the most convenient along with epocs from 800 to 1000. Particularly for CNN, it could take from 16 to 48 filters in steps of 16, kernel and pooling sizes as integers from 1 to 7. For LSTM, in the first run we allowed structures from 1 to 3 layers with 1 to 15 neurons per layer in steps of 5. And up to two dense additional layers with 1 to 5 neurons. However, in the second we allowed only either 2 or 3 non-dense layers ranging from 1 to 20 neurons. Lastly, the hybrid CNN-LSTM had the same options as CNN and LSTM.

Table 4
Model performance in terms of $\overline{RMSE}$ [%].

Model	Intrinsic	Original	Exp-Pearson	Exp-Physics
RF	12.10	11.83	11.80	11.77
SVR	13.74	13.71	13.59	13.04
CNN	15.32	15.27	14.64	14.03
LSTM	14.95	14.89	14.82	14.55
Hybrid	15.75	15.72	15.14	14.99

5. Results & discussions

This section evaluates the different models in terms of RMSE over the testing set (July and August), omitting night periods. Due to the large number of different models, feature combinations and hyperparameters, we have used an HPC cluster with Dual Intel® Xeon® Processor E5-2650 v4 (2 *x* 12 core, 2.20 GHz) with nodes of 256GB/512 GB memory, to train and test the models (DTU Computing Center, 2021). We only present the most relevant results for each model, along with remarks regarding the conducted sensitivity analysis.

5.1. Influence of feature sets on ML models

Fig. 2 shows the RMSE performance of persistence and SPAR against each ML method for every feature selection approach, for all 60 prediction time steps. Note that the brute force approach, i.e., finding the best model by considering all features through an exhaustive search, always returns the best feature combination as a subset of those included in the Expanded-Physics set.

Persistence was easily outperformed by all models, and its performance deteriorates in a quasi-linear manner, reaching 30% RMSE by the end of the horizon. Therefore, we have scaled the plots to focus on the rest of the models. In contrast, SPAR is far superior, as it is able to even beat some of the ML models.

Table 4 presents the percentage $\overline{RMSE}$ of each ML model for the three different feature sets and the intrinsic feature. Note that persistence and SPAR present an error of 21.19 and 13.72 respectively.

Sub Fig. 2(a) presents the results for RF, where we can see how the Expanded-Physics model is able to beat all three benchmarks and feature selection approaches. Sub Fig. 2(b) presents

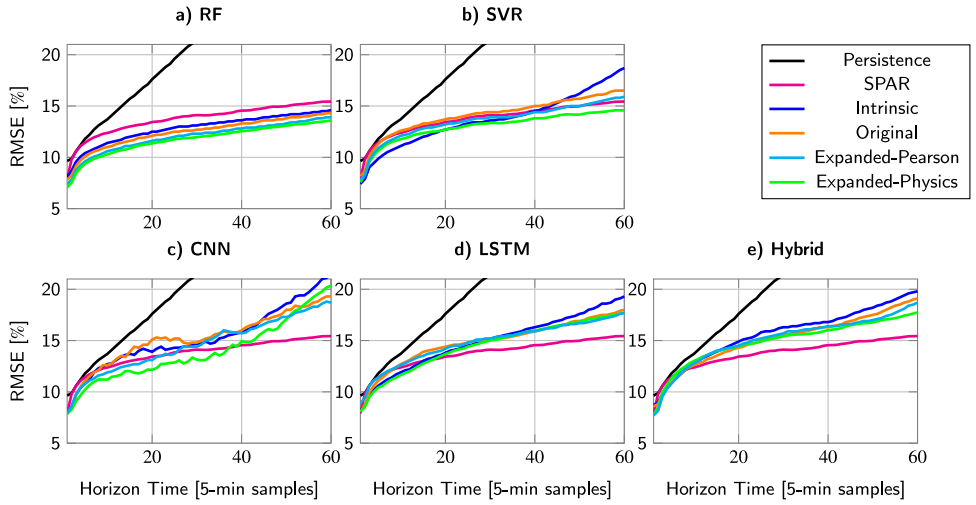

Fig. 2. RMSE Comparison of the different feature selection methods for each ML model.

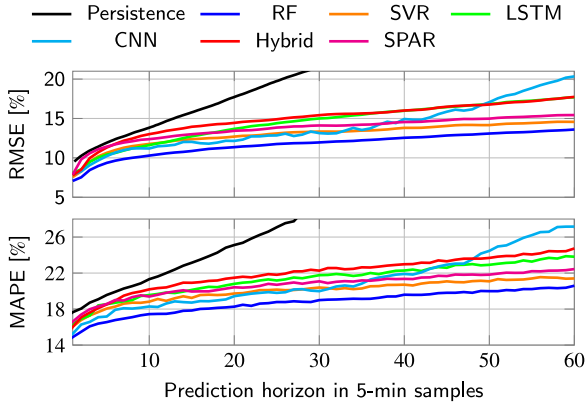

Fig. 3. Comparison of the best performers.

the results for SVR. Intrinsic and Ex-Pe models present the same structure using *rbf* kernel. The first fails to beat SPAR, while the latter manages to do so by including T_m and P_{AC}, presenting an $\overline{RMSE}$ of 13.59%. On the other hand, Original and Expanded-Physics present an $\overline{RMSE}$ of 13.71%, and, 13.04% respectively. The remaining three models under-perform when compared to SPAR.

By using various ML models, we can generalize some of the results from this study. For instance, including exogenous features helps reducing prediction error, while feature combination has a higher effect on the overall performance than the actual hyperparameters, which only improve performance over a relatively narrow bandwidth. However, this is crucially dependent upon the ML method in the first place. Also, there are only a few different feature combinations shared among the best performers, which belong to the search space of the physics-informed dataset (Expanded-Physics). Summarizing, we see how by expanding the dataset with the PV-performance model and H_{day} we outperform both Intrinsic models and those trained with the Original dataset. Furthermore, by avoiding the inclusion of redundant features in the selection we contribute to reducing the search-space, thus decreasing the time and effort invested in finding the best possible feature combination for a given model. Not only there is a difference in performance, but also on the number of models capable to outperform persistence and SPAR, as including redundant features tended to harm accuracy, especially in ANNs. In addition, models trained on the Original and Expanded-Pearson datasets, tended to be more complex, taking longer to train and

Table 5
Best performers ranking and summary.

Method	Pre	Features	$\overline{RMSE}$ [%]	$\overline{MAPE}$ [%]	R^2
RF	30	5-6-8-10-12	11.77	18.65	0.94
SVR	0	6-10-12	13.04	19.92	0.9
SPAR	1	6	13.72	20.63	0.86
CNN	30	1-5-6-10	14.03	20.85	0.83
LSTM	10	5-6-7-10-12	14.55	21.22	0.79
Hybrid	0	5-6-12	14.99	21.83	0.75
Persistence	1	6	21.19	32.07	0.62

requiring higher memory. Generalizations should be made cautiously in ML, given the broad spectrum of training possibilities of the same model over a single dataset. However, Fig. 2 shows a clear tendency, consistent in all different models. Hence, we feel confident by stating that our dataset expansion approach combined with a physics-informed feature selection contributes to simpler to train and more accurate predictors.

Fig. 3 compares only the best performers of each ML method for each 5-min step over the 5 h horizon, in terms of $\overline{RMSE}$ and $\overline{MAPE}$. These metrics are also included in Table 5 which presents a ranked summary based on model performance, showing the main features number of previous samples, along with R^2. Note that the features are identified with the ordinals of Table 2.

In general, we can mention a few factors impeding the comparison of our work against other published papers, such as different size and placement of PV arrays, data-set length, monitored error metrics, time resolution and horizon. In fact, we only managed to find one study suitable for a relatively fair comparison. There, data from a 200 kW array is used to forecast one step ahead with an ANN obtaining % $\overline{RMSE}$ ranging from 6.6 to 11% (Sheng et al., 2017). Despite this result being smaller than the presented models, the difference in the horizon should be taken into account, as predicting only one step ahead is a substantially easier task than predicting 60, as in this paper. Fig. 4 presents 3 consecutive predictions for a random day in August; that is with both horizon and refreshing rate of 60 samples. Its purpose is to illustrate the behavior of the different methods and showcase their ability to capture the different trends. We can see how the performance deteriorates fast after roughly 20 samples for all ANN while RF, SVR and SPAR hold high accuracy over most of the horizon.

5.2. Discussion on model training and performance

Next, findings related to the sensitivity analysis of the different model configuration are presented. Regarding hyperparameter tuning, RF below 400 trees and above 3000 systematically under-performed for the same feature combination, which is explained by under and over-fitting the training data. A similar behavior was encountered in SVR with C values above 7 or using any kernel function other than rbf. LSTM and Hybrid deeper than 4 layers, or including 2 sequential LSTM and a dense one also under-performed, irrespective of the number of neurons. Despite the general recommendation of using 2-D convolutional layers in CNN and hybrid, we found that 1-D convolutions returned more consistent results. In general, 600 to 1100 epocs with a batch size of 16 or 32 presented a satisfactory performance for all ANN.

Regarding feature selection, intrinsic models had good accuracy in general. Features T_a, W_s, and W_{dir} either harmed accuracy or did not present noticeable improvements. GHI and E_{POA} should not be included together, as they basically hold the same information, while E_{POA} seems to consistently outperform GHI. Similarly, T_c is preferred over T_m and P_{AC} over P_{DC}. Lastly, T_c is the only exogenous feature present in all the best performers, while H_{day} appears in all but one. Intuitively, we understand the relationship between the time of day and solar production; yet, a marker such as Pearson correlation does not reflect it. In fact, H_{day} presents the lowest correlation in Table 2, proving once again how such a naive approach to feature selection should be taken cautiously. In general, E_{POA}, P_{AC}, T_c, and H_{day} were consistently part of the best performers in different combinations for all the different ML-models.

Regarding training time, RF required from 1 to 3 h, while SVR took from 30 to 50. Among the ANN, CNN were the fastest, ranging from a couple of hours to 24, while LSTM and Hybrid took similar time, from 12 to 35. Note that the time range is affected by several factors, like number of previous samples and exogenous features. Regarding memory needs all models required less than 2 GB, except for RF, which reached 90 GB. While nor time neither memory were practical constraints during this research, it should be taken into account when designing a practical implementation. Lastly, less than a second is required to cast a prediction with a trained model.

Our results show that persistence represents a weak benchmark model for 5-min resolution, and especially a 5 h horizon. Hence, non-naive forecasting methods such as SPAR must be used when assessing the accuracy of any given model. Further, expanding a dataset with a PV performance model and the time of the day allows to introduce physics-informed features in the ML models. These capture the highly non-linear relationship between the thermodynamical state and the meteorology. Given the extensive testing procedure followed in the study case, we feel confident when stating that these are generally useful for any ML model. Finally, feature selection based in the physical relationships between the different metrics rather than Pearson correlation allows to further reduce the search-space of the best performer. For example, features such as GHI and E_{POA} hold the same information, thus being redundant. The same happens with P_{AC} and P_{DC}, T_m and T_c. Combining the expansion of the dataset with the reduction of the search-space makes the proposed method a simple yet effective way to approach the development of ML-based forecasting tools for PV power. The best performer is RF, however, for a practical implementation, training times and memory allocation capabilities must be taken in consideration.

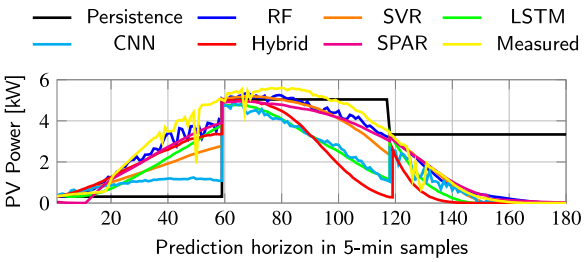

Fig. 4. One day forecast comparison.

6. Conclusions

Forecasting represents the key enabler in reducing RES uncertainty. Particularly, ML-based models have recently gained relevance in scientific literature due to increasing data availability and computational power. In this paper, we propose a physics-informed dataset expansion and feature selection process that improves the accuracy of high resolution PV power forecasters that are based on ML models. We achieve this by integrating a PV-performance model along with other physically relevant features within a ML method. This allows the model to learn the physical relationships among the different features, yielding better accuracy than conventional methods. The added features capture the relationships of a PV's physical domain, but sometimes appear as redundant, and present weak linear correlation towards the intrinsic feature. The simplicity of the proposed approach facilitates its integration in virtually any PV farm as it uses datasheet parameters and basic meteorological data. Further, we showed how the Pearson criterion fails to discriminate between redundant features and discards other physically relevant ones. Hence, we proposed an alternative criterion that, despite still using Pearson, considers physical relevance. Through extensive validations in the most challenging period we show how these adaptations enhance accuracy in a wide variety of ML methods, and significantly reduce training time by compressing the feature search-space. In this way, we are able to offer the forecasting community performance benchmark for both the ML and statistical methods discussed in this paper (RF, SVR, CNN, LSTM, CNN-LSTM, SPAR and Persistence). Both the employed dataset and code are made available to the interested reader in order to facilitate the replication and comparison of this study.

We showcased the usefulness of this methodology with a case study, using a PV array installed in Denmark. We used a two-fold random search approach in both the original and expanded datasets in order to find the best performer. Results showed a clear pattern in prediction accuracy common for all considered models. Intrinsic models presented the lowest accuracy, followed by those using the exogenous variables from the original dataset, followed by those in the extended dataset including only all the highly correlated metrics (Expanded-Pearson). Lastly, the set including only the physically-relevant features ranked best (Expanded-Physics). Therefore, we conclude that by expanding the dataset with physically-informed metrics, ML methods are able to cast more accurate predictions. Hence, by reducing the search-space selecting only non-redundant, physically-informed features, we effectively tighten the search-space easing the training and tuning stage. RF proved to be the best performer, although its memory needs during training might tilt the balance towards choosing other methods requiring less resources. However, this ranking could change if other features were available, such as sky imaging or NWP.

CRediT authorship contribution statement

Daniel Vázquez Pombo: Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data curation, Writing – original draft, Visualization. **Peder Bacher:** Conceptualization, Methodology, Software, Validation, Formal analysis, Resources. **Charalampos Ziras:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Supervision. **Henrik W. Bindner:** Supervision. **Sergiu V. Spataru:** Methodology, Resources, Data curation, Supervision. **Poul E. Sørensen:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

In order to increase transparency and the replicability of the work we share the employed dataset in a repository at DTU DATA Pombo (2022). Furthermore, we also share several Python scripts based on Open Access libraries with two purposes. First, to facility the access to the dataset and, second, provide an easy to use platform applying the physics informed methodology discussed in this work. A full description can be found in Pombo et al. (2022b).

References

Ahmad, T., Chen, H., 2020. A review on machine learning forecasting growth trends and their real-time applications in different energy systems. Sustainable Cities Soc. 54, 102010.

Ahmed, R., et al., 2020. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renew. Sustain. Energy Rev. 124, 109792.

Alanazi, M., Khodaei, A., 2016. Day-ahead solar forecasting using time series stationarization and feed-forward neural network. In: 2016 North American Power Symposium (NAPS). IEEE, pp. 1–6.

Bacher, P., et al., 2013. Short-term heat load forecasting for single family houses. Energy Build. 65, 101–112.

Blaga, R., et al., 2019. A current perspective on the accuracy of incoming solar energy forecasting. Prog. Energy Combust. Sci. 70, 119–144.

Bracale, A., et al., 2017. A probabilistic competitive ensemble method for short-term photovoltaic power forecasting. IEEE Trans. Sustain. Energy 8 (2), 551–560. http://dx.doi.org/10.1109/TSTE.2016.2610523.

Cordova, S., Rudnick, H., Lorca, A., Martinez, V., 2018. An efficient forecasting-optimization scheme for the intraday unit commitment process under significant wind and solar power. IEEE Trans. Sustain. Energy 9 (4), 1899–1909. http://dx.doi.org/10.1109/TSTE.2018.2818979.

DTU Computing Center, 2021. DTU Computing Center Resources. Technical University of Denmark, http://dx.doi.org/10.48714/DTU.HPC.0001.

Feng, C., et al., 2019. Unsupervised clustering-based short-term solar forecasting. IEEE Trans. Sustain. Energy 10 (4), 2174–2185. http://dx.doi.org/10.1109/TSTE.2018.2881531.

Fu, Y., et al., 2021. Sky image prediction model based on convolutional auto-encoder for minutely solar PV power forecasting. IEEE Trans. Ind. Appl. 57 (4), 3272–3281.

Ghosh, S., et al., 2017. Distribution voltage regulation through active power curtailment with PV inverters and solar generation forecasts. IEEE Trans. Sustain. Energy 8 (1), 13–22. http://dx.doi.org/10.1109/TSTE.2016.2577559.

Gigoni, L., et al., 2018. Day-ahead hourly forecasting of power generation from photovoltaic plants. IEEE Trans. Sustain. Energy 9 (2), 831–842. http://dx.doi.org/10.1109/TSTE.2017.2762435.

Hafiz, F., Awal, M., de Queiroz, A.R., Husain, I., 2020. Real-time stochastic optimization of energy storage management using deep learning-based forecasts for residential PV applications. IEEE Trans. Ind. Appl. 56 (3), 2216–2226.

Harrou, F., Sun, Y., 2020. Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems.

Hastie, T., et al., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Jang, H.S., et al., 2016. Solar power prediction based on satellite images and support vector machine. IEEE Trans. Sustain. Energy 7 (3), 1255–1263. http://dx.doi.org/10.1109/TSTE.2016.2535466.

King, D.L., Kratochvil, J.A., Boyson, W.E., 2004. Photovoltaic Array Performance Model. United States. Department of Energy.

Liu, J., et al., 2015. An improved photovoltaic power forecasting model with the assistance of aerosol index data. IEEE Trans. Sustain. Energy 6 (2), 434–442. http://dx.doi.org/10.1109/TSTE.2014.2381224.

Lotfi, M., Javadi, M., Osório, G.J., Monteiro, C., Catalão, J.P., 2020. A novel ensemble algorithm for solar power forecasting based on kernel density estimation. Energies 13 (1), 216.

Mohandes, B., et al., 2021. Optimal design and operation of a renewable power plant with dispatching capability: A renewable energy management system. IEEE Trans. Sustain. Energy 1. http://dx.doi.org/10.1109/TSTE.2021.3058252.

Mussumeci, E., Coelho, F.C., 2020. Large-scale multivariate forecasting models for Dengue-LSTM versus random forest regression. Spatial Spatio-Temporal Epidemiol. 35, 100372.

Persson, C., Bacher, P., Shiga, T., Madsen, H., 2017. Multi-site solar power forecasting using gradient boosted regression trees. Sol. Energy 150, 423–436.

Pombo, D.V., 2022. The SOLETE dataset. http://dx.doi.org/10.11583/DTU.17040767, URL https://data.dtu.dk/articles/dataset/The_SOLETE_dataset/17040767.

Pombo, D.V., Bindner, H.W., Spataru, S.V., Sørensen, P.E., Bacher, P., 2022a. Increasing the accuracy of hourly multi-output solar power forecast with physics-informed machine learning. Sensors 22 (3), 749.

Pombo, D.V., Gehrke, O., Bindner, H.W., 2022b. SOLETE, a 15-month long holistic dataset including: Meteorology, co-located wind and solar PV power from. Data Brief 42, 108046.

Pombo, D.V., et al., 2021. Data-driven wind power forecast for very short horizons. In: 2021 4th International Conference on Smart Energy Systems and Technologies (SEST). IEEE, pp. 1–6.

Sheng, H., Xiao, J., Cheng, Y., Ni, Q., Wang, S., 2017. Short-term solar power forecasting based on weighted Gaussian process regression. IEEE Trans. Ind. Electron. 65 (1), 300–308.

Sobri, S., Koohi-Kamali, S., Rahim, N.A., 2018. Solar photovoltaic generation forecasting methods: A review. Energy Convers. Manage. 156, 459–497.

Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer science & business media.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., Fouilloy, A., 2017. Machine learning methods for solar radiation forecasting: A review. Renew. Energy 105, 569–582.

Wang, H., et al., 2019. A review of deep learning for renewable energy forecasting. Energy Convers. Manage. 198, 111799.

Wang, F., et al., 2020a. A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework. Energy Convers. Manage. 212, 112766.

Wang, H., et al., 2020b. Taxonomy research of artificial intelligence for deterministic solar power forecasting. Energy Convers. Manage. 214, 112909.

Yan, J., et al., 2021. Frequency-domain decomposition and deep learning based solar PV power ultra-short-term forecasting model. IEEE Trans. Ind. Appl..

Zhang, X., et al., 2019. A solar time based analog ensemble method for regional solar power forecasting. IEEE Trans. Sustain. Energy 10 (1), 268–279. http://dx.doi.org/10.1109/TSTE.2018.2832634.

Zhen, Z., et al., 2020. Deep learning based surface irradiance mapping model for solar PV power forecasting using sky image. IEEE Trans. Ind. Appl. 56 (4), 3385–3396.