Power Output Reconstruction of Photovoltaic Curtailment

Roman Kohút
Institute of Information Engineering, Automation and Mathematics
Slovak University of Technology in Bratislava
Bratislava, Slovakia
roman.kohut@stuba.sk

Michal Kvasnica
Institute of Information Engineering, Automation and Mathematics
Slovak University of Technology in Bratislava
Bratislava, Slovakia
michal.kvasnica@stuba.sk

***Abstract*—The use of renewable energy sources in the grid's energy mix has recently gained popularity. Especially as solar photovoltaic (PV) generation production has almost zero emissions during its operations, they are preferred over fuel-based electricity production. However, expanding PV generation in grid capacity increases the chance of PV curtailment occurrence. Not to mention that microgrids supported with large-scale PV generation almost certainly create PV curtailment regularly. As the forecast of PV production is one of the electricity grid operation cornerstones, the prediction model should be as accurate as possible. The latest trend is utilizing machine learning (ML) models to predict PV output, thanks to their excellent learning and regression capabilities. However, its performance can be highly influenced by measurements used during the model design. Unfortunately, only some of the research on this topic deals with the PV curtailment problem resulting in underperforming ML models. This paper proposes a novel approach to identify and replace curtailed PV measurements. The methodology includes the physical model as a baseline of truly producable energy, which is then investigated and corrected as a piecewise linear system using Pearson correlation and weather measurements. Through real-life comparative scenarios, the suggested data reconstruction method provides increased accuracy of supervised ML-based solar power prediction.**

***Index Terms*—Photovoltaic curtailment, Solar energy prediction, Physical model, Pearson correlation, Machine learning**

I. Introduction

In recent years the world has changed its leaning towards renewable sources, especially solar energy emerges as one of the leading clean and cheap power production. This change in the energy mix is essential as it establishes a sustainable and ecological electrical system [1]. However, the operating grid can not rely on pure renewable production as its generation is volatile and can not be directly shifted towards the electricity load demand [2]. This burden requires power generation maintainers to predict their production most accurately. Based on this generation's forecasts, microgrid customers can schedule load demand, buy electricity on the short-term market or utilize their model predictive control to optimize the operation of the microgrid [3]. On the other hand, large grid operators can prepare other resources to provide electricity when needed [4]. Hence, the quality of the provided photovoltaic (PV) production predictions is crucial, as many decision-making actions depend on it.

Many approaches exist to designing PV forecasters, with different prediction horizons and steps considering their use. The bases in microgrid control or real-time grid scheduling are day-ahead forecasts with quarter-hourly or hourly steps. The direct solar energy predictions can be categorized as physical, statistical, and machine learning (ML) [5], [6] approaches. The most widely used are ML-based prediction especially deep-learning techniques, thanks to its excellent regression performance through its capability of learning hidden patterns. The bottleneck of this method are data used to train and generate forecasts. To accurately predict future PV production and its fluctuation, it is necessary to correctly identify and predict weather conditions acting on the PV system. In this field, it is standard to use various data preprocessing techniques, as shown in [7], to improve prediction capabilities [8]. However, most of them focus on removing outliers, filtering signals, replacing missing data, and finding the correct feature inputs to the ML models. A few of them solve the problem of the PV curtailment [9], which directly affects solar power measurements see [10], [11]. The critical step to finding the most accurate ML model is to consider all possible influences that can act on the PV system. If the data used contain curtailed PV production model will underperform as it is trained to provide false (lower) PV prediction for specific inputs.

This paper's main contribution and purpose is to improve any ML-based PV power forecasts of the systems where PV curtailment occurs. Proposed data reconstruction algorithm utilizes the physical model of the PV system, the Pearson correlation coefficient, and a wide range of weather measurements. The capability of used data reconstruction method is tested on a real large-scale system with five different scenarios. To demonstrate the benefits of this approach, multiple ML-based models are trained (using the reconstructed and curtailed data) and tested to produce a day-ahead prediction.

II. Preliminaries

A. System and Data Description

In general solar plants consist of multiple assets. The main two are photovoltaic generators that directly produce electricity from sunlight and power converters that transform direct current (DC) to alternating current (AC).

The power plant is capable of functioning in three distinct modes. The first mode is referred to as the off-the-grid system, which operates independently of any national or local electricity distribution network. The remaining two setups, known as the hybrid and grid-connected, are connected to the primary power grid. Both the hybrid and off-the-grid modes allow the utilization of excess energy through battery storage, although the battery's capacity is typically undersized. In addition, the grid connection for the hybrid and grid-connected setups enables the exportation of excess electricity to the primary grid or importing of electricity during any shortages.

Measurement of this solar power plant generation can be done through inverters or smart electricity meters. The negative of this system from the data perspective is that the measurement collection is limited only to the exact energy production. When the current possible power generation on the solar panels is greater than the load's consumption, naturally, the load will consume only the necessary power. The extra energy is lost (curtailed), as the information about it. If the power plant includes batteries, this behavior is only shifted in time as the capacity of the batteries reaches its maximum and can no longer store energy, so the power production is curtailed. Such behavior occurs mainly in an off-grid system, as others can export excessive electric energy. However, grid contestants must meet agreed deals as it can destabilize the whole grid operation. This concluded in the same situation that PV production is curtailed as the load consumption decrees.

B. Physical Model

The PV system's output relies on multiple factors, categorized as weather conditions and mechanical properties. The more significant factor is the corresponding meteorological state, which behaves as a multi-variable non-linear system, including solar irradiance, temperature, humidity, wind speed, pressure, visibility, and many more. Not all the mentioned parameters are essential to developing a good estimator of accessible power production. The most important is solar irradiance, which directly transforms into electrical energy and temperature, influencing PV panel efficiency.

In general, total solar irradiance G_{t} can be modeled as a sum of three components, surface absorbed irradiance G_{s}, diffused irradiance G_{d}, and ground reflected irradiance G_{g}

$$G_{\mathrm{t}}(k) = G_{\mathrm{s}}(k) + G_{\mathrm{d}}(k) + G_{\mathrm{g}}(k). \tag{1}$$

By direct measurement of $G_{\mathrm{DHI}}(k)$ diffuse horizontal irradiance and $G_{\mathrm{DNI}}(k)$ direct normal irradiance at each time step k is possible to determine each component separately, concerning α_{STA} and α_{SAA} PV surface tiled and azimuth angle respectively. The final output power production can be calculated as energy produced from total solar irradiance affected by current PV panel efficiency

$$P_{\mathrm{t}}(k) = P_{\mathrm{max}} \frac{G_{\mathrm{t}}(k)}{G_{\mathrm{stc}}} N_{\mathrm{p}} \left(1 + \kappa \left(T_{\mathrm{c}}(k) - T_{\mathrm{STC}}\right)\right), \tag{2}$$

where the left part of the brackets represents total irradiation (1) transformed at defined time step k at standard test conditions (STC) $G_{\mathrm{STC}} = 1000\ W/m^2$, $T_{\mathrm{STC}} = 25\ ^{\circ}C$, P_{max} is a maximal (peak) power production for specific PV panel at STC and N_{p} is a number of installed panels. The right side of the brackets defines the effect of PV panel efficiency linked to the cell temperature $T_{\mathrm{c}}(k)$, and κ is the temperature coefficient of P_{max}. As the PV cell temperature is hard to measure in real-time, it can be calculated as follows

$$T_{\mathrm{c}}(k) = T(k) + \frac{G_{\mathrm{d}}(k)}{G_{\mathrm{TC}}} \left(T_{\mathrm{NOCT}} - T_{\mathrm{TC}}\right), \tag{3}$$

where T_{NOCT} is nominal operating cell temperature (NOCT) obtained at test condition (TC) $G_{\mathrm{TC}} = 800\ W/m^2$, $T_{\mathrm{STC}} = 20\ ^{\circ}C$ and $T(k)$ is measured ambient temperature. It follows from the (2) that if the $T_{\mathrm{c}}(k)$ is higher, then T_{STC} efficiency of the PV panel decrease by the coefficient κ.

Defined STC, TC, NOCT, P_{max}, κ can be obtained from PV manufacturer. The number of PV panels, surface tiled, and azimuth angles are unique for each solar power plant installation. This section provides a brief overview of the model used. For more details, see [12].

C. Pearson Correlation Coefficient

Purpose of the correlation coefficients is to find and interprets how strong a relationship is between the investigated variables [13]. Pearson correlation coefficient is one of the most popular tools in the field of feature extraction for machine learning inputs with a large number of available variables. This coefficient represents a linear correlation between two continuous variables (x, y), and it is formulated as follows

$$r_{\mathrm{xy}} = \frac{\sum_{i=1}^{M} (x_i - \overline{x}) \sum_{i=1}^{M} (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{M} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{M} (y_i - \overline{y})^2}} \tag{4}$$

where r_{xy} represents correlation coefficient, $\overline{x}$ denotes mean of x and $\overline{y}$ denotes mean of variable y across all samples M. Value form (4) can come from a variance of $r_{\mathrm{xy}} \in [-1, 1]$. The closer the coefficient gets to the 1, the higher the positive correlation is between x and y. Conversely, suppose the coefficient gets closer to the -1. In that case, variables have a higher negative correlation, and if the r_{xy} is close to 0, there is no direct linear correlation between x and y.

III. Problem Statement

As described in Section II-A, power production can be curtailed by many factors despite solar power plant operation modes. The consequence of such a behavior is that the measured information from the smart devices provides incomplete information as other grid parts influence it. This creates unwanted data corruption, which is hard to identify and process before syntheses of the forecasting model. If data are not processed correctly model trained with such data provides underestimated PV production. It affects the utilization of solar power generation as its forecasts are used in model predictive control (MPC), which benefits from knowing the most accurate

predictions. Likewise, various design-making processes are dependent on the best PV forecasts.

This work suggests a novel curtailed data reconstruction approach for supervised machine learning modeling of PV panels production. Utilizing the solar panel physical model and Pearson correlation coefficient to identify and replace curtailed power production with estimated values of a maximal possible generation.

IV. Prediction of Achievable Energy

In general, the designed PV energy forecasting model includes weather forecasts as input producing the one-step prediction in the following form

$$y(k) = f_{\mathrm{P}}(x(k)), \quad k = 0, \ldots, N, \tag{5}$$

where y is prediction of generated power, x represents input features to the predictive model f_{P} for define time step k.

In the field of ML-based solar power prediction, the adopted methodology consists of standard procedures, including:

1) Data collection and pre-processing.
2) Identification and feature extraction.
3) Model selection (type, structure).
4) Model training and validation using weather measurements and historical PV production.
5) Testing procedure using weather forecasts and historical PV production to select best performance model.

The following section extends the pre-processing as it aims to avoid biased predictions and model underperforming. The training data are reconstructed to replace curtailed PV production with estimated achievable production.

A. Data Reconstruction

Assume we have acquired a representative sample of historical weather measurements $W \in \mathbb{R}^{M \times n_{\mathrm{w}}}$, where n_{w} represents the number of unique variables $w \in \mathbb{R}^{M}$ and power production measurements $Y \in \mathbb{R}^{M}$ which includes corrupted information.

The first step in PV data reconstruction is to find the physical model (2) and identify its parameters. Using historical weather measurements and the designed model, we can construct dataset $Y_{\mathrm{P_t}} \in \mathbb{R}^{M}$, which corresponds to the measured one. This data represents a naive guideline of achievable power generation, which is unaware of grid limitations. The curtailment of power production indicates that only interesting information from the modeled dataset $Y_{\mathrm{P_t}}$ is that which is greater than measured information from Y. Using this fact created dataset is modified as follows

$$Y_{\mathrm{P_t}}(m) = y_{\mathrm{p}}, \quad y_{\mathrm{p}} = \begin{cases} Y_{\mathrm{P_t}}(m), & \text{if } Y(m) < Y_{\mathrm{P_t}}(m) \\ Y(m), & \text{otherwise} \end{cases}, \tag{6}$$

where $m = 1, \ldots, M$ represents an index of the single value of the vector. In this way, it is possible to keep original measurements intact and separate potentially corrupted ones.

To correctly identify curtailed power production, it is necessary to investigate modeled data. Whether the mismatching between $Y_{\mathrm{P_t}}$ and Y is caused by effects of the connected grid or exogenous influences of the weather conditions, which are not included in the model (2). Calculating the difference

$$\Delta Y = Y_{\mathrm{P_t}} - Y, \tag{7}$$

and using it in Pearson guided correlation from Section II-C it is possible to determine if the weather measurements w from W significantly impact modelled and measured data mismatch (7). If the premise given is correct, it should be reflected in reconstructed data. Otherwise, we can assume that there was a curtailment in PV-produced energy.

The compensation of weather variables that are not modeled is based on the simple linear model in which coefficients are found as follows

$$\min_{a,b} \quad J = \sum_{m=1}^{M} \Big(\Delta Y(m) - (S(m)a^{\top} + b) \Big)^2, \tag{8a}$$

$$\text{s.t.} \quad S(m)a^{\top} + b \leq \Delta Y(m), \tag{8b}$$

where $a \in \mathbb{R}^{n_{\mathrm{s}}}, b \in \mathbb{R}$ represents coefficient of linear equation and $S \in \mathbb{R}^{M \times n_{\mathrm{s}}}$ includes only dependent variables from W ($n_{\mathrm{s}} \leq n_{\mathrm{w}}$). The weather measurements are selected based on a simple rule. If the investigated weather variable $w^{(i)}$ where $i = 1, \ldots, n_{\mathrm{w}}$, has a greater correlation coefficient (4), then the variables from the model (2) ($G_{\mathrm{DHI}}, G_{\mathrm{DNI}}, T$) it is included in S. After the linear model is found, the final formulation of the resulting preprocess data is provided as

$$Y_{\mathrm{f}}(m) = Y_{\mathrm{P_t}}(m) - (S(m)a^{\top} + b), \quad m = 1, \ldots, M \tag{9}$$

where Y_{f} represents reconstructed data, $Y_{\mathrm{P_t}}$ represents modeled data, and $Sa^{\top} + b$ is a compensation of exogenous weather influences acting on the PV system.

However, as mentioned in Section II-B, the meteorological impact on PV power production is highly non-linear. This means that a simple linear model provides unsatisfying results for a large number of historical samples M containing the different weather conditions (rainy, warm, cold days or changing seasons, etc.). Also, Pearson correlation is a measure of linear dependency between two datasets creating the same problem.

The key step is to investigate a small portion of the time series data $M \Rightarrow \{M_1, \ldots, M_n\}$ at the time. This solution helps to decompose time series data as a piecewise linear process providing the desired results.

B. Implementation Details

The condition on lines (13, 16) can be interpreted as follows. Suppose the correlation of the exogenous weather variable $w^{(i)}$ with ΔY for the defined portion of data is smaller than the modeled variables ($G_{\mathrm{DHI}}, G_{\mathrm{DNI}}, T$). In that case, it is possible to assume that external weather influences did not create the difference ΔY. Otherwise, the linear model (9) will compensate their influence. This condition can be modified by tuning coefficient $p_{\mathrm{r}} \geq 0$, which for non-zero values allows passing less dependant weather variables.

The range of the investigated data Δt should be selected based on the sampling time of the measured data W, Y and weather conditions of the PV panel's location. Chosen time

Algorithm: Data Reconstruction

Input: W, Y
Output: Y_{f}

Initialization:
Set $\Delta t \leftarrow \mathbb{R} \in (1, M), p_{\mathrm{r}} \leftarrow 0$.
Initial Calculation:
Set $Y_{\mathrm{P_t}} \leftarrow P_{\mathrm{t}}(W)$.
Power Saturation :
for $m = 1 : M$ **do**
  if $(Y_{\mathrm{P_t}}(m) < Y(m))$ **then**
    Set $Y_{\mathrm{P_t}}(m) \leftarrow Y(m)$.
  end if
end for
Set $\Delta Y \leftarrow Y - Y_{\mathrm{P_t}}$.
Main Loop:
for $j = 1 : \Delta t : M - \Delta t$ **do**
  Set $\delta \leftarrow \Delta Y(j : j + \Delta t), \phi \leftarrow Y_{\mathrm{P_t}}(j : j + \Delta t)$ and $\omega \leftarrow W(j : j + \Delta t)$.
  Calculate Correlation for Modeled Variables :
  Set $r_{\mathrm{xy}}^{\mathrm{DHI}} \leftarrow f(G_{\mathrm{DHI}}, \delta)$, $r_{\mathrm{xy}}^{\mathrm{DNI}} \leftarrow f(G_{\mathrm{DNI}}, \delta)$ and $r_{\mathrm{xy}}^{\mathrm{T}} \leftarrow f(T, \delta)$.
  for $i = 1 : m_{\mathrm{w}}$ **do**
    if $\left(\omega^{(i)} \neq \{G_{\mathrm{DHI}}, G_{\mathrm{DNI}}, T\}\right)$ **then**
      Set $r_{\mathrm{xy}}^{\omega} \leftarrow f(\omega^{(i)}, \delta)$.
    end if
    if $\left(\left|r_{\mathrm{xy}}^{\omega}\right| < \left\{\left|r_{\mathrm{xy}}^{\mathrm{DHI}}\right|, \left|r_{\mathrm{xy}}^{\mathrm{DNI}}\right|, \left|r_{\mathrm{xy}}^{\mathrm{T}}\right|\right\} - p_{\mathrm{r}}\right)$ **then**
      Append $S \leftarrow \omega^{(i)}$.
    end if
  end for
  Calculate Reconstructed Data with Compensation :
  Find $a, b \leftarrow \operatorname{argmin} J(\delta, S)$.
  Set $\sigma \leftarrow \phi - \left(Sa^{\top} + b\right)$.
  Append $Y_{\mathrm{f}} \leftarrow \sigma$.
end for

range of investigated data is crucial as it is directly linked to linear models. Using a too large portion of the dataset at once (week, month, year) will lead to inaccurate correlations as the modeled variables are often the most dependent and directly affect power production. Consequences are such that algorithm will never or very rarely compensate possible exogenous effects, and if so, it will be inaccurate. On the contrary, selecting a too small portion of the dataset compared to the sampling time of the measurement will lead to too much compensation of modeled data, and we will end up with an underestimation of power production. It is worth mentioning that the range of the reconstructed data Δt may be different in each iteration. The proposed algorithm does not restrict that. If the data indicates it, it is even a recommended step. Another point is that provided physical model from Section II-B can take any complexity, form, or additional variables. The only difference will be that the S will be a smaller subset of W as the number of the exogenous variables will decrease with the modeled variables increase.

V. Case Study

This work presents a real large-scale PV system on which we test the proposed algorithm and provide its reliability for any case of PV setup. Investigated PV installation is located in Romania and includes 11680 individual solar panels of the same efficiency and peak power P_{max}. The system contains three master slave inverter control, each containing four separate inverters whit individual smart meters. The primary inverter controls the other three, which are connecting and disconnecting based on the power production. The system is directly connected to the main grid without local power consumption, so it does not show frequent power curtailment. This makes it a perfect example as a proposed algorithm can be tested against true power production. Excluding power inverters data one by one from the aggregated dataset, we can simulate PV power curtailment, which is then reconstructed based on the Algorithm 1 shown in the previous section.

TABLE I: Solar panel model specification under the standard test condition: $G_{\mathrm{STC}} = 1000$ W/m^2, $T_{\mathrm{STC}} = 25$ °C and test conditions: $G_{\mathrm{TC}} = 800$ W/m^2, $T_{\mathrm{TC}} = 20$ °C.

α_{STA} [°]	α_{SAA} [°]	P_{max} [W]	κ [%/°C]	T_{NOCT} [°C]
19.5	0	250	−0.44	20

VI. Results and Discussion

The data included in this work represents the total PV productions, weather measurements, and forecasts within 230 days with one hour sampling time $T_{\mathrm{s}} = 1\ h$. Weather measurements and forecasts for defined location are imported from third-party open-source API Tomorrow.io, providing 20 different weather variables, including possible 430h prediction.

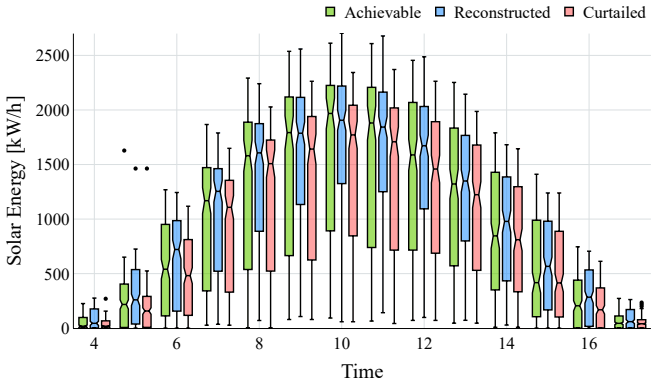
The total PV production is used in five different showcase simulations to provide representative results across a wide range of solar power curtailment. In the presented results, we apply the following scenario scheme:

TABLE II: Utilized data of process grope across all scenarios.

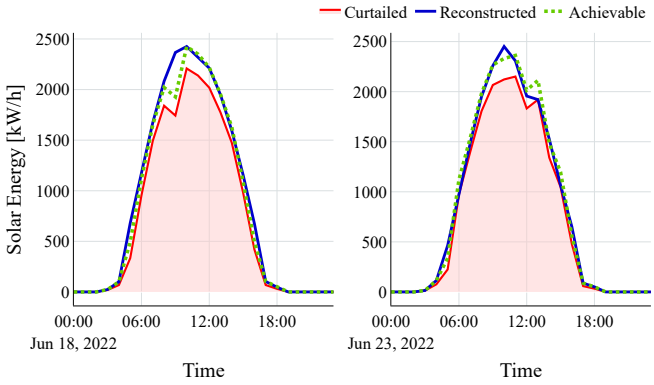
Grope	Case 1	Case 2	Case 3	Case 4	Case 5
1	100%	72.2%	72.2%	53.0%	72.2%
2	84.0%	100%	84.0%	100%	72.5%
3	100%	100%	100%	83.5%	72.1%

From now on, we will refer to the aggregated (the sum of all gropes) information of curtailed production, achievable production (without curtailment), and reconstructed data (curtailed data that have undergone a reconstruction process).

Each case is individually processed, as it contains unique curtailed PV measurements Y, which are then transformed to reconstructed dataset Y_{f} representing maximal possible power generation for a defined time. The range of the investigating data is chosen as $\Delta t = 24$ h. By comparing these datasets Y, Y_{f} with achievable production Y_{R}, we generate results shown in Fig. 1 and Tab. III. The presented Fig. 1a shows a detailed



(a) Statistical results of reconstructed data compare to achievable and curtailed PV production for each hour of the daily active generation.



(b) Reconstructed data for 24 h of the day.

Fig. 1: Visualization of data reconstruction procedure (case 2).

TABLE III: Statistic results for reconstructed data Y_{f} compare to the curtailed Y and achievable Y_{R} PV power generation $[\mathrm{kW/h}]$ across all cases.

Case	P_{miss} [%]	Mean			Variance $\cdot 10^4$		
		Y_{R}	Y	Y_{f}	Y_{R}	Y	Y_{f}
1	5.3	498.7	472.1	504.8	53.5	46.2	49.5
2	9.2	498.7	453.0	499.8	53.5	44.7	49.8
3	14.5	498.7	427.4	483.9	53.5	38.3	45.8
4	21.1	498.7	395.7	474.1	53.5	32.9	44.3
5	27.8	498.7	363.4	461.5	53.5	28.9	43.0

statistical comparison across all measurements (case 2) for each hour of the active PV production (from 17 pm to 4 am is negligible power generation). As we can see, the mean of the reconstructed data is shifted toward the mean of the achievable PV production. As well as, the variance of reconstructed data is stretched compared to curtailed PV measurements, leading to better correlation with achievable power within every hour of the day. As the results suggest, the reconstruction procedure succeeded in mimicking the real measurements (achievable power), as we can see in the example Fig. 1b. The difference between reconstructed and achievable PV production is caused by a one-time change in the weather conditions, which are not captured as significant by Pearson correlation (4), so in conclusion, they are ignored by the reconstruction procedure.

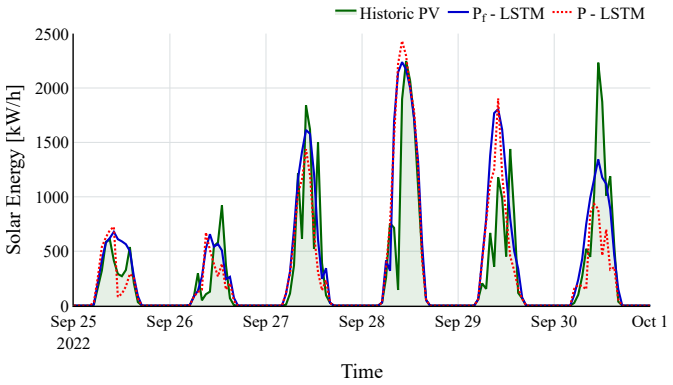


Fig. 2: Day-ahead PV forecast, using LSTM trained with reconstructed P_{f} and curtailed P PV data compared to achievable historical PV production (case 2).

In Tab. III, we can see statistic results for the whole dataset across all cases. Naturally, increases in the missing total produced power P_{miss} decrease the overall accuracy of the proposed algorithm. Looking at the mean and variance of the reconstructed data, we can observe a slower decline compared to curtailed PV production, which decreases at a much greater rate. Showing that the reconstruction procedure provides beneficial results as it scales with higher power curtailment.

To show the advantages of reconstructed data in PV power prediction, we design pair of the long-short term memory (LSTM see [14]) for every case from Tab. II. Each scenario is then represented by LSTM trained using curtailed PV measurement, and LSTM trained with reconstructed data to predict PV power P, P_{f} notated subsequently. We divide our datasets into three groups. First, the training dataset includes measurements within 138 days, and it is used to estimate the trainable parameters of the LSTMs. The second group represents validation data (23 days) for the model validation. Finally, the last group includes 69 days of measurements on which we perform accuracy analysis presented in Tab. IV and Fig. 2. The structure of the LSTM networks is selected as follows. Weather feature inputs are chosen using provided Algorithm 1 by including all modeled variables and those which were used in model compensation (9). In summary, inputs include (direct normal irradiance, diffuse horizontal irradiance, outside temperature, cloud base, dew point, humidity, precipitation intensity, wind speed, and wind direction). Ideal hyperparasites of the ML model are found using Bayesian optimization, the number of LSTM units is 543, the learning rate is selected as $3.715 \cdot 10^{-4}$. The 2900 epochs of minimizing the sum of squared errors were sufficient in order to find optimal weights and basis of the LSTM units with ADAM optimizer. The inputs to the LSTM are not changing, and outputs are not so much different in provided scenarios, so we fix these parameters for all ten recurrent neural networks.

In Fig. 2, we can see a day-ahead forecast (prediction for 24 h from the midnight of the previous day) comparison of

TABLE IV: Overall accuracy results for day-ahead forecasting of PV power generation [MW/h].

Case	P_{tot}	Maximal		MSE		Surplus		Deficit	
		ΔP	ΔP_{f}	P	P_{f}	P	P_{f}	P	P_{f}
1	372.5	4.4	2.0	220	86	10.6	19.1	42.6	18.7
2	372.5	4.4	2.1	238	91	9.0	17.3	47.1	21.9
3	372.5	4.6	2.6	329	135	6.3	12.1	60.8	33.5
4	372.5	4.9	3.2	612	201	4.3	11.4	92.3	44.3
5	372.5	5.1	3.4	716	243	3.1	10.6	101.1	49.7

constructed models with historical values for case 2. As shown, LSTM trained with reconstructed data is more reliable on days with higher power generation. For the days with lower production, both LSTMs provide similar results as the training data are not often curtailed in such small production. More detailed results can be found in Tab. IV. Where P_{tot} represents the total power produced without curtailment from our testing dataset, which is the same across all scenarios. From the mean squared error (MSE), we can declare that LSTM trained with reconstructed data provide superior results. The maximal absolute error between aggregated true daily production and the day-ahead forecast is shown in the third column. As we can see, LSTM trained with curtailed power production has almost twice larger maximal error $2\Delta P \approx \Delta P_{\text{f}}$. We chose the sum of daily aggregated surplus and deficit as the last indicator across all daily samples. As it already been pointed out, LSTM trained with reconstructed data naturally provides a larger surplus (overestimated PV prediction) competed to LSTM trained with achievable PV measurements, which on the contrary, provides a more significant deficit (underestimated PV prediction). However, as we can see, the sum of the deficit and surplus for both of the LSTM models compared to P_{tot} are proportionally different on a large scale showing the benefits of the proposed data reconstruction.

VII. Conclusion

This paper proposes a novel approach to photovoltaic (PV) power production measurement reconstruction for systems where PV generation occasionally exceeds load consumption. The methodology is based on constructing the physical model of a real PV system providing the baseline of maximal production. Investigating the curtailed PV measurements as a piecewise linear system, we were able to utilize the Pearson correlation coefficient and linear model of weather measurements. Both tools are used to compensate the difference between the modeled and curtailed PV power production. Using this approach, we reconstruct historical PV measurements and preserve the behavior of a non-linear system. The provided algorithm was tested on a real large-scale PV system, including 11680 individual panels. To investigate the performance of our scheme, we have simulated various levels of curtailment by artificially excluding information from a subset of PV inverters. In all scenarios algorithm successfully reconstruct corrupted (curtailed PV) measurements with some degree of accuracy, which is linked to the size of power curtailment.

The direct benefit of the proposed approach is improved machine learning (ML) based photovoltaic power production forecasts. Our study involves generating and comparing day-ahead predictions using ML models for a defined system, both with and without data reconstruction. From the investigation of different scenarios, we conclude that LSTM trained with reconstructed data provide superior results. Not only maximal error between the historical values and its predictions is almost twice lower. We also provide a good trade-off between the prediction deficit and surplus by possible tuning of the data reconstruction algorithm.

Acknowledgment

This research is funded by the European Commission under the grant no. 101079342 (Fostering Opportunities Towards Slovak Excellence in Advanced Control for Smart Industries). The authors gratefully acknowledge the contribution of the Scientific Grant Agency of the Slovak Republic under the grant 1/0490/23 and the Slovak Research and Development Agency under the project APVV-20-0261.

References

[1] S. Sen and S. Ganguly, "Opportunities, barriers and issues with renewable energy development – a discussion," *Renewable and Sustainable Energy Reviews*, vol. 69, pp. 1170–1181, 2017.

[2] M. Anvari, G. Lohmann, M. Wächter, P. Milan, E. Lorenz, D. Heinemann, M. R. R. Tabar, and J. Peinke, "Short term fluctuations of wind and solar power systems," *New Journal of Physics*, vol. 18, no. 6, p. 063027, 2016.

[3] G. Gust, T. Brandt, S. Mashayekh, M. Heleno, N. DeForest, M. Stadler, and D. Neumann, "Strategies for microgrid operation under real-world conditions," *European Journal of Operational Research*, vol. 292, no. 1, pp. 339–352, 2021.

[4] N. Sunar and J. R. Birge, "Strategic commitment to a production schedule with uncertain supply and demand: Renewable energy in day-ahead electricity markets," *Management Science*, vol. 65, no. 2, pp. 714–734, 2019.

[5] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar energy*, vol. 136, pp. 78–111, 2016.

[6] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, pp. 569–582, 2017.

[7] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC Trends in Analytical Chemistry*, vol. 132, p. 116045, 2020.

[8] M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Forecasting of pv power generation using weather input data-preprocessing techniques," *Energy Procedia*, vol. 126, pp. 651–658, 2017.

[9] E. O'Shaughnessy, J. R. Cruce, and K. Xu, "Too much of a good thing? global trends in the curtailment of solar pv," *Solar Energy*, vol. 208, pp. 1068–1077, 2020.

[10] M. Z. Liu, A. T. Procopiou, K. Petrou, L. F. Ochoa, T. Langstaff, J. Harding, and J. Theunissen, "On the fairness of pv curtailment schemes in residential distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4502–4512, 2020.

[11] N. Stringer, N. Haghdadi, A. Bruce, and I. MacGill, "Fair consumer outcomes in the balance: Data driven analysis of distributed pv curtailment," *Renewable Energy*, vol. 173, pp. 972–986, 2021.

[12] S. A. Kalogirou, *Solar energy engineering: processes and systems*. Academic press, 2013.

[13] H. Xu and Y. Deng, "Dependent evidence combination based on shearman coefficient and pearson coefficient," *Ieee Access*, vol. 6, pp. 11 634–11 640, 2017.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.