39. PV-Symposium 2024

Conference paper

https://doi.org/10.52825/pv-symposium.v1i.1063

Published: 05 Aug. 2024

Inclusion of Shading and Soiling With Physical and Data-Driven Algorithms for Solar Power Forecasting

Tim Kappler[1], Anna Sina Starosta[1], Bernhard Schwarz[1], Nina Munzke[1], and Marc Hiller[1]

[1] Karlsruhe Institute of Technology, Germany

Abstract. Shading and soiling are the biggest environmental factors that negatively affect the yield of PV systems. In order to integrate PV systems into the grid as easily and on large scale as possible, it is important that energy generation forecasts are as accurate as possible. The scope of this paper is to present a method to integrate shading and soiling into machine learning based PV forecasts, even if they have already been pre-trained by a large dataset. This paper focuses on shading by buildings, trees, obstacles, while shading by clouds can only be considered to a limited extent by weather forecasts. This study uses a dataset of three years of training data to build a base model. Subsequently, the power loss due to shading and soiling is determined using a digital twin and used to correct the forecast values of the baseline model. Finally, an evaluation of the corrected and original predicted values is performed. This shows that the forecast error can be reduced in the same way as the loss due to shading and soiling using various machine learning methods. The results are compared against a Physics-Informed Neural Network (PINN), which outperformed popular machine learning methods both with and without shading and soiling by 6.6%.

Keywords: Solar Power Forecasting, Shading, Soiling, Machine Learning, Physics-Informed Neural Networks

1. Introduction

The global deployment of photovoltaic (PV) systems and their associated capacity experienced a remarkable increase of 21.8%, from 866 GW in 2021 to 1055 GW in 2022. Anticipated further growth is essential to meet future CO^2 emission targets [1]. However, challenges related to grid stability have come to the forefront due to the variable nature of PV power generation influenced by factors like cloud movements, rain, and irradiation changes [2]. Accurate forecasts are crucial to ensuring stability and availability, with forecasting methods encompassing physical models, statistical approaches, and machine learning methods. Machine learning is favored for its strong generalization capability, adapting to new situations [3]. Despite their advantages, machine learning methods demand substantial data collection over years to achieve high accuracy. Recent trends lean toward employing new neural network architectures for solar power forecasting. Studies have highlighted the effectiveness of Radial basis function (RBF) networks and variational autoencoder networks [4], [5]. While some studies suggest the superiority of Support vector regression (SVR) over neural networks, the choice depends on specific applications [6], [7]. Exogenous data such as irradiation, wind speed, and air temperature are typically used for solar power forecasting [8]. The inclusion of weather forecast data for longer prediction periods has been emphasized in certain studies [9]. The commercial use of solar power forecasting solutions is already ongoing [10]. Challenges arise in predicting unforeseen situations not accounted for in recorded datasets, such as increased shade from

growing trees or newly constructed buildings. Additionally, factors like pollution, dust, and module degradation impact PV system output. Traditional methods struggle to handle shading adequately, leading to increased forecast errors. Existing research on solar power prediction often overlooks explicit consideration of shading effects caused by obstacles. In previous research on the forecast of solar energy, shading effects due to obstacles or soiling are often not explicitly taken into account, as these effects on PV systems are usually not recorded explicitly so that these effects can be allocated. However, these effects are not always avoidable, especially for rooftop systems, and should therefore be taken into account. To the best of the authors' knowledge, there is only one relevant paper which explicitly uses shading and soiling losses in an energy forecast setup from Marca et al. [10]. In this study, the scope of the energy forecast was limited to daily forecasts. In addition, no dynamic changes in shading were considered, but a constant factor was determined for the whole day. The approach utilized Numerical weather prediction (NWP) data for Global horizontal irradiance (GHI) and temperature, incorporating a Photovoltaic (PV) model accounting for shading, optical losses, and system losses [11]. Seasonal fluctuations of shading effects, combined with changing sun positions, pose challenges for solar power forecasting. Few studies extend beyond one year of training data, limiting their ability to address such effects. To address these limitations, a hybrid approach combining data-driven methods and physical models is proposed, explicitly considering shading and soiling effects and quantifying loss effects. This paper focuses on correcting prediction values using a model with extensive training data under shading and soiling conditions to improve generalization capability. In this case, shading and soiling are also taken into account simultaneously for solar power forecasts for the first time. Furthermore, to the best of our knowledge, a Physics-Informed Neural Network is used for the first time for solar power forecasting, which, in addition to physical input variables such as irradiation or temperature, is also based on physical relationships in the form of equations. These relationships were not only implemented but also extended to include shading and soiling.

The quantification of loss effects, such as shading and soiling, enables more robust forecasts, which are essential for energy management strategies in a real environment. This research is unique in combining shading and soiling loss quantification with day-ahead solar power forecasting, recognizing the significant impact of shading and soiling on solar power generation. The validation of the proposed method using shading and soiling setups is presented, followed by a discussion of results and future outlook. Key contributions of this work include the development of solar power forecasts baseline model on a three-year dataset, an enforced shading and soiling setup for solar power forecasting and a comparison of popular machine learning methods under shading and soiling conditions.

2. Methodology

2.1. Method

This section explains how the machine learning methods are trained and how the dataset required was created. The calculation of losses due to shading (ρ_{Shading}) and soiling (ρ_{Soiling}) is also carried out using a simulation model as well as how the predicted values of the reference model were corrected. A common dataset must be created so that the parameters for the methods can be trained based on the data. For this purpose, the power of PV arrays and the corresponding weather data must be collected, which are to be used as input for the machine learning methods. Numerical simulations from the German Weather Service (DWD) can be used for the weather data, which can be collected as weather forecast data with an hourly resolution of three hours up to 10 days in the future. Power data from the solar park, which is located at 49.1° longitude and 8.4° latitude and can be used as the basis for the forecast. The setup of the solar park and the system structure of the tables are described in more detail in chapter 2.2. Here, data are available with a resolution of one second, which can be mixed down to hourly average values as weather data are only available in hourly resolution. However, these data are subject to disturbances, such as the presence of gaps as missing values

due to the failure of system components or sensor errors. For this reason, the values are filtered so that no negative PV values are allowed and no PV values that are 20% above the maximum system installation capacity. These values are marked as corrupt and excluded from the dataset. As this only accounts for a very small part of the entire dataset, no impact is to be expected during training. The DWD's forecast data, which are used as weather data, have been collected since February 2021, so that a dataset of almost three and a half years can be created until December 2023. Two and a half years are used to train the machine learning algorithms and one year to validate the models on unknown data. Within this test period, the shading and soiling scenarios are also performed. However, since not all-weather features are relevant for the PV forecast, it must first be determined which input data are significant to the actual PV output. The Pearson correlation coefficient and the Mutual Information (MI) index are often used in the literature for this purpose [12]. The pearson coefficient simply calculates the correlation between the respective input variable and the corresponding PV power. The advantage here is that the procedure is easy to interpret. However, as a linear equation, the correlation measure is only able to capture linear relationships. For non-linear correlations, which can also occur in PV forecasting, the Mutual Information Index is often used. For the filtered dataset, only weather data with a correlation coefficient above 0.5 ($|\rho_{\text{Pearson}}| \geq 0.5$) and a Mutual Information Index (MI > 0.6), which is commonly used in the literature, are used [12]. After the dataset was restricted to the meaningful features, various machine learning methods were trained to validate the entire method on different learning algorithms. Classical machine learning methods [13] such as Decision trees (DT), Gaussian process models (GPR), Neural networks (NN) and kernel-based methods such as Support vector regressions (SVR) and Physics informed neural networks (PINN) [14] were used. The unique aspect of PINNs is that they can consider an equation or constraints to bring in additional prior physical knowledge. The PINN uses the Evans equation [15] in this work which is extended by the term $\rho_{\text{Soiling}} \cdot \left(1 - \rho_{\text{Shading}}\right)$ in order to incorporate shading and soiling into the training and testing process. The power according to the extended Evans equation is described by equation 1, which requires the area $A_{\text{PV}} = N_{\text{Series}} \cdot N_{\text{parallel}} \cdot A_{\text{Module}}$ of all modules, the temperature coefficient T_{C} and two system-specific coefficients C_1 and γ, which were determined using the training data.

$$P_{\text{PV}} = C_1 \cdot \rho_{\text{Soiling}}\left(1 - \rho_{\text{Shading}}\right) \cdot A_{\text{PV}} \cdot E_{\text{POA}} \cdot \left(\left(1 - T_{\text{C}}(T_{\text{Cell}} - T_{\text{Ref}}) - C_2 log_{10}\left(\frac{E_{\text{POA}}}{E_{\text{Ref}}}\right)\right)\right) \quad (1)$$

In contrast to simpler PV power models, the Evans equation also takes into account the low-light behavior and temperature characteristics. For this purpose, the Global Horizontal Irradiance (*GHI*) forecast must first be converted into the irradiation values that hit the PV module perpendicular (E_{POA}). Here, an appropriate decomposition and transposition model from Louche et al. [16] is used with the Python library pvlib [17]. For the PINN, there are also more complex models in form of state space equations [18], which offer no additional value for this work due to the limitation of hourly-resolved weather data, as the processes can be seen as stationary. To estimate the temperature of the cells T_{Cell}, the Faiman model [19] is used according to equation (2), which can calculate the cell temperature based on air temperature, irradiation, and wind speed. This empirical model considers the interaction between these factors, accounting for the heat generated by solar radiation and the influence of wind on heat dissipation.

$$T_{\text{Cell}} = T_{\text{Amb}} + \frac{E_{\text{POA}}}{U_1 + U_2 \cdot v_{\text{Wind}}} \quad (2)$$

The Python library SciANN [20] is used to implement the PINN. This provides a wrapper to include the physical knowledge as differential equations. For this purpose, the Evans equation was described by a differential equation whose algebraic solution it corresponds (see equation 3).

$$E_{\mathrm{POA}}{}^2 \cdot \frac{\partial^2}{\partial E_{\mathrm{POA}}{}^2} P_{\mathrm{PV}} + \frac{\partial}{\partial T_{\mathrm{Amb}}} P_{\mathrm{PV}} + C_i \cdot E_{\mathrm{POA}}{}^2 + C_j \cdot E_{\mathrm{POA}} = 0 \quad (3)$$

With $C_i = \frac{2 \cdot C_1 \cdot \rho_{\mathrm{Soiling}} \cdot (1-\rho_{\mathrm{Shading}}) \cdot A_{\mathrm{PV}} \cdot T_{\mathrm{C}}}{U_1 + U_2 \cdot v_{\mathrm{Wind}}}$) and $C_{\mathrm{j}} = C_1 \cdot \rho_{\mathrm{Soiling}} \cdot (1 - \rho_{\mathrm{Shading}}) \cdot A_{\mathrm{PV}} \cdot (C_2 + T_{\mathrm{C}})$
Equation (3) has the form $x_1{}^2 \cdot y_{x_1 x_1} + y_{x_2} + C_{\mathrm{i}} \cdot x_1{}^2 + C_j \cdot x_1 = 0$ and has the extended Evans equation as a solution. It can be used as an initial and boundary condition that no power is generated for no irradiation and the power is known under STC conditions. This means that not only the loss can be minimized due to the data, but also the physical loss due to the governing equations (1) – (3).

Furthermore, it must be shown that the trained models are robust. For this purpose, a 5-fold cross-validation is used to show the robustness. A robust model in machine learning is characterized by its ability to consistently perform well across different datasets and under different conditions. This robustness is important to ensure that the model does not overreact to specific characteristics of a dataset, but still is able to make accurate predictions for new unseen data. *k*-fold cross-validation is a method used to assess the robustness of such a model. Instead of relying on a single random division of the data into training and test sets, it divides the dataset into *k* subsets. *k*-1 subsets are used for training and the remaining subset is used for testing. This procedure is repeated k times, with each subset being assigned the role of test set exactly once. By averaging the error metrics over these *k* iterations, a more reliable model performance is obtained. The use of *k*-Fold cross-validation offers several advantages. It helps to reduce variance in the performance metrics as it is based on multiple training and test splits. It also enables an assessment across the entire dataset, as each data point is used once for validation. This method allows the model to generalize better to different datasets and thus demonstrates its robustness to different conditions and data [21].

The common metrics of time series forecasting are used to evaluate the models. The RMSE measures the deviation between the actual and predicted values in a time series. To calculate the RMSE, the differences between the actual and predicted values are squared, the average of these squared differences is taken and finally the square root is formed. The normalized RMSE (nRMSE) is a variant of the RMSE in which the error is set in relation to the installed capacity (10 kW). This makes it possible to compare the errors across different datasets. Also, the R^2 measure is used to evaluate the performance of the models. R^2 is easy to interpret as it ranges between 0 and 1. The closer the value is to 1, the better the predicted values match the actual values. Both metrics are crucial to evaluate the performance of time series prediction models and to ensure that the predictions are accurate and consistent.

2.2. Experimental Setup

Each PV array of the solar park uses a string inverter which can provide three MPPTs. The 10 kWp installation capacity of one array is made up of two strings with an output of 5 kWp each. The PV arrays examined used Solarwatt Blue modules with a peak output of 250 W according to STC conditions. 20 modules are installed per string and 40 modules in total per array. In order to collect sufficient shading data for the data-driven methods, two shading scenarios have been selected at the solar park from which measurement data has been collected since July 2023 (see Figure 1).

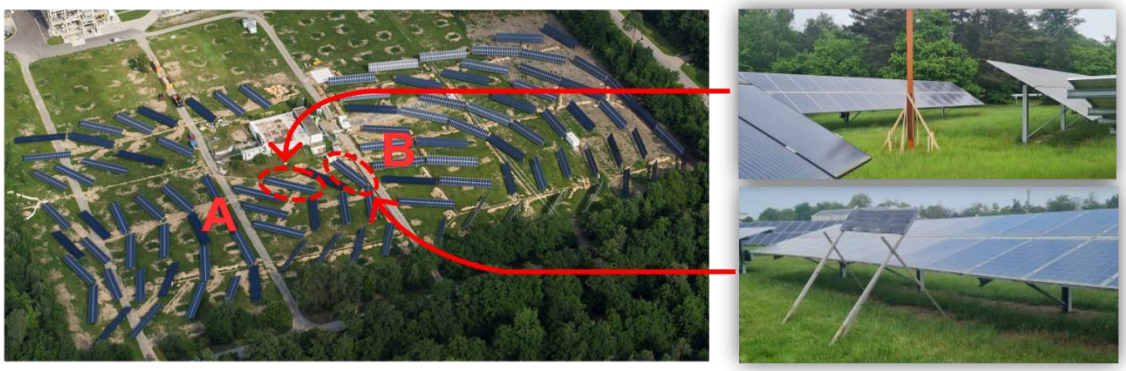

***Figure 1**. PV tables of the solar park on the north campus of the KIT. Selected PV tables for the investigation of shading are circled in red with their respective shading scenarios.*

Shading and soiling scenarios were reproduced at the test arrays (see Table 1), with generated power being recorded and stored in a database. A simulation model is used to quantify the shading and soiling losses.

***Table 1**. Used PV Arrays to investigate shading and soiling at the solar park of KIT*

Array	Tilt angle	Orientation
A	15°	30°
B	30° (East)	0° (South)

The validation of the simulation model is explained more briefly in a previous paper [22]. The model provides the power of a PV array that does not experience any shading or soiling and is able to simulate the generated power on a real system. A 1-diode model is used for this purpose [9], which is taken from the data sheet of the corresponding PV modules. The irradiation and module temperature are available as input data. The data are taken from the irradiation and temperature sensors installed on the PV array. The losses simulated due to shading and utilization can be compared with the actual losses by comparing them with reference strings. However, the quantified power losses cannot be used to correct the prediction values as they depend on the fluctuations of the clouds. Therefore, the power losses are converted into ratios that do not change significantly over the different days. The more meaningful ratios for the forecast algorithms were converted by dividing the actual power by the simulated power of the PV array.

Afterwards, the study looks at the extent to which the reduction of the forecast values by the ratios contributed to a reduction of the RMSE. The machine learning methods already mentioned are then trained on the dataset determined and tested under soiling and shading.

Lastly more detailed comparison of the PINN network with regard to the prediction error under shadowing and soiling was also carried out. Here, two-layer neural networks were compared with each other. Both used two layers with 50 and 25 neurons. The MLP referred to a classic two-layer neural network, while the PINN used the same configuration but also used the physical prior knowledge.

3. Results

3.1. Feature Selection

As relevant input characteristics for the PV forecast, the Pearson feature selection and the mutual information index were calculated of the training dataset. The most promising features are shown in Table 2. The features $GHI_{Forecasted}$, $P_{T\text{-}1d}$ and Θ_{Sun} are characterized by particularly

high significance. The number of minutes of sunshine, azimuth and air temperature are moderately significant. The current hour, day of the year and wind speed are only very slightly significant and therefore not included in the dataset. The azimuth angle should be mentioned, which has a low linear correlation with the PV power but a high non-linear correlation with the PV power.

***Table 2**. Pearson feature selection and mutual information index. Features were kept with correlation coefficient above 0.5 or mutual information index above 0.6*

Feature	Description	$\lvert\rho_{Corr}(X,Y)\rvert$	MI(X,Y)	Relevant Feature
t_H	Hour	0.05	0.07	No
DoY	Day of the Year	0.04	0.01	No
Θ_{Sun}	Elevation Angle of the sun	0.72	0.91	Yes
$P_{T\text{-}1d}$	Power value yesterday	0.81	0.91	Yes
T_{Amb}	Air temperature	0.50	0.15	Yes
$GHI_{Forecasted}$	Global Horizontal Irradiance	0.88	0.91	Yes
N_{Sun}	Sunny Minutes per hour	0.79	0.50	Yes
v_{Wind}	Windspeed	0.14	0.05	No
φ_{Sun}	Azimuth	0.04	0.70	Yes

3.2. Correction of the forecast

The results of the test error are listed in Table 3-7 together with the respective parameters of the models. The PINN network has the lowest validation error with an nRMSE of 7.17 % and a coefficient of determination of 0.90. Neural networks and decision tree methods achieved the best results on average, followed by SVR and kernel approximation methods. The GPR models performed the worst with an nRMSE between 7.61 % and 14.1 %.

***Table 3**. Parameter for Decision tree based algorithms*

Name	Method	Min size	Number regressor	Learn-rate	nRMSE / %	R^2
FineTree	Decision Tree	12	1	-	9.35	0.84
BoostedTree	Decision Tree	8	30	0.1	7.67	0.89
BaggedTree	Decision Tree	8	30	-	7.79	0.89

***Table 4**. Parameter for SVR based algorithms*

Name	Method	Kernel	Kernel-size	Box Constant	Epsi-lon	nRMSE / %	R^2
LinearSVM	SVR	Linear	1	Auto	Auto	11.5	0.77
QuadraticSVM	SVR	Quadratic	30	Auto	Auto	8.11	0.89
CubicSVM	SVR	Cubic	30	Auto	Auto	7.74	0.90
GaussianSVM	SVR	Gauß	12	Auto	Auto	7.61	0.90
CoarseGaussi-anSVM	SVR	Gauß	30	Auto	Auto	8.51	0.87

***Table 5**. Parameter for kernel-based algorithms*

Name	Method	Kernel-size	Epsilon	Iteration limit	nRMSE / %	R^2
SVMKernel	SVR	1	Auto	1000	8.34	0.88
LRKernel	LS Kernel	30	Auto	1000	8.19	0.88

***Table 6**. Parameter for neural net-based algorithms*

Name	Method	N_{Layers}	$N_{Neurons}$	Epochs	Acti-vation	nRMSE / %	R^2
Bilayered	Neural Net	2	20	1000	ReLu	7.48	0.90
MediumNet	Neural Net	1	50	1000	ReLu	7.56	0.90
WideNeuralNet-work	Neural Net	1	100	500	ReLu	7.48	0.90
PINN	Neural Net	2	20	500	SciActi vation	7.17	0.90

***Table 7**. Parameter for GPR-based algorithms*

Name	Method	Basisfunc-tion	Kernel-function	Kernel-size	Sigm a	nRMSE / %	R^2
Exponen-tialGPR	GPR	Constant	Exponential	Auto	Auto	7.61	0.90
5/2GPR	GPR	Constant	Matern 5/2	Auto	Auto	9.56	0.84
SquaredGPR	GPR	Constant	Squared Exponential	Auto	Auto	14.1	0.66
RationalGPR	GPR	Constant	Rational Quadratic	Auto	Auto	13.8	0.67

Figure 2 shows the quantified power losses after shading and soiling using the structures described in section 2.2. The measured power loss is calculated so that one string remains unshaded and unsoiled while the other is shaded by the structure. The same procedure applies to the soiling measurement. This power loss is due to the fact that the simulated power is compared with the actual power based on the irradiation and temperature data. This resulted in an RMSE of 0.03 kW for soiling and 0.12 kW for shading for the simulation model over the observation timeframe of one month. Both errors are negligibly small in relation to the 10 kW system (0.3% nRMSE and 1.2% nRMSE). The two power losses from measurement and simulation are therefore in good agreement. When the soiling occurred on 8th September, a reduction in the soiling ratio from 0.97 to around 0.91 can be seen in Figure 3. This is due to the soiling, as the soiling loss increases according to Figure 2. This means that 6 % of the output on this day has been lost due to the soiling.

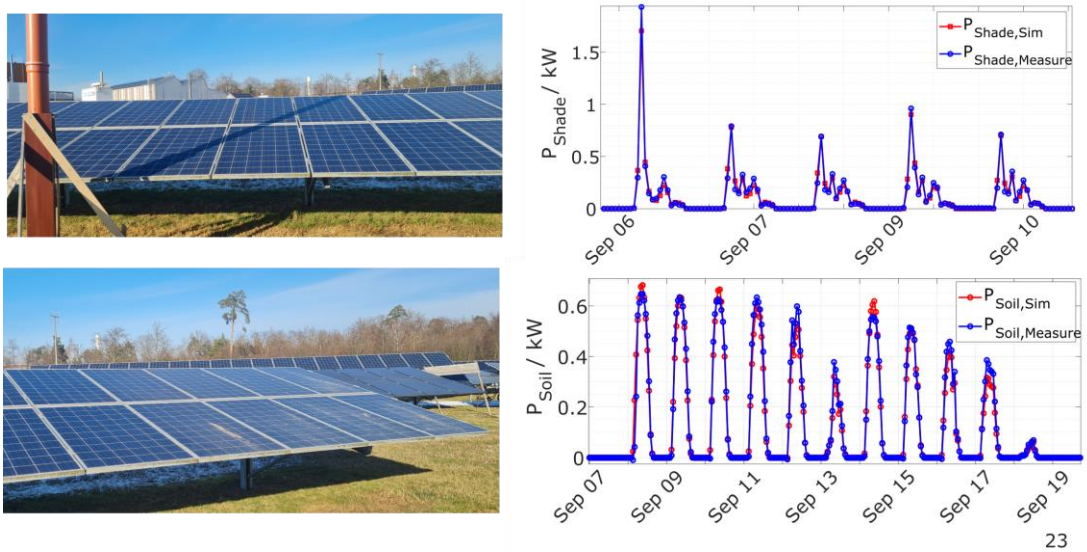

***Figure 2**. Estimated power loss by simulation of the PV array model for shading and soiling compared to the measured power loss compared to the unshaded and unsoiled PV string (Measure)*

It can also be seen in Figure 2 and Figure 3 that there was a clean-up on September 13 due to the onset of rain. This was followed by a further soiling of the same PV string, which was also removed after three days. Figure 2 shows that the shading can be clearly recognized, especially in the morning. This is due to the low position of the sun, which leads to wide shadows being cast. Figure 3 shows the shading ratio, which leads to a reduction of the shading ratio to up to 0.2 in the corresponding morning hours. This means that 80% of the power is lost due to the shadow in relation to an unshaded table at these times.

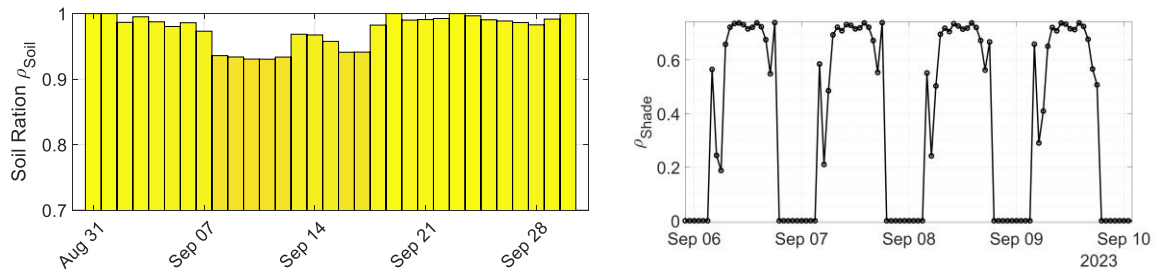

***Figure 3**. Soiling and Shading ratio for correcting the existing forecast values*

The results for the improvement in the forecast under the two scenarios for shading and pollution are shown in Figure 4. The forecast error under pollution (left) and shading (right) is shown in red throughout the day. The forecast error was evaluated over the entire pollution period from 31st August to 30th September and the forecast error was calculated for each hour. Using the calculated shading and soiling conditions, a reduction in the forecast error can be recognized at the corresponding times as shown in Figure 4. A reduction can be seen for soiling over the entire day, while it can only be seen with the shading structure at the corresponding shading time points. These shading points are the times at which the shading ratio was reduced. As the quotient is calculated daily for soiling, the value here is a constant with a value of 0.95, as in addition to the actual reduction in the quotient, there are also reductions due to rain. In the case of shading, a change in the shading ratio can be seen throughout the day. In contrast, the deviation from the value of 1 between 5 and 8 a.m. (UTC) is particularly noticeable as this is when the shading is most significant. During the night hours the ratio is zero throughout the day.

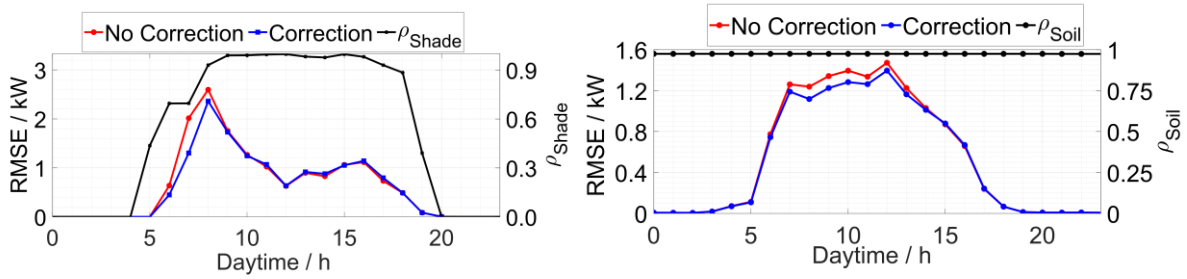

***Figure 4**. Improvement of the RMSE by correcting the predicted values using the shading and soiling ratio*

It can be seen in both graphs that the forecast error is reduced at the times when the effects occur depending on the value of the corresponding loss ratio. The parameters of the neural network used are listed in Table 8.

***Table 8**. Used parameters of the neural network*

N_{Layers}	$N_{Neurons}$	Optimizer	Learn rate	Epochs	Batch size
2	15	Adam	1e-3	300	24·7

Validation by use of shading and soiling can be seen in Figure 5. The reduction in the forecast error at 8 o'clock as a result of the correction using the shading ratio can be seen here. This can also be seen from the curve through the loss ratio (black). The remaining reduction in the forecast error is due to the soiling ratio. Here the overall RMSE could be reduced by 16%.

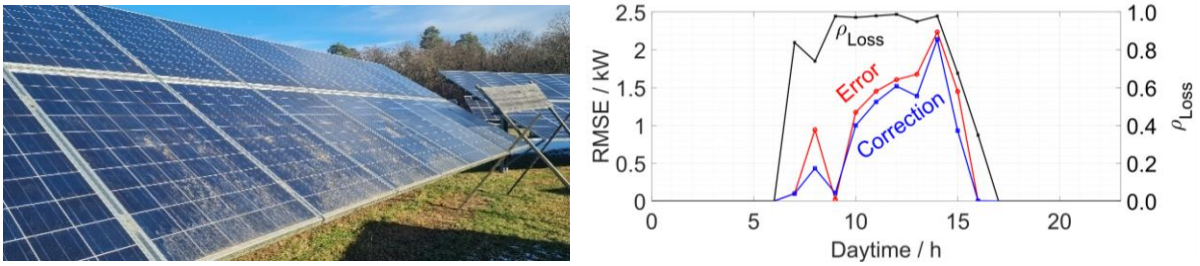

***Figure 5**. Comparison of calculated output due to shading and soiling on a sunny winter day. The comparison is made with an unshaded and unsoiled string of the same PV-array*

It was shown what effect the individual effects (soiling and shading) have on the reduction of the RMSE. Once only the soiling ratio was used for correction and another time only the shading ratio. Then, the entire ratio was used and it can be seen in Figure 6 that there is no negative influence of the two ratios, as the corresponding curves can be added together. It is also evident that the machine learning methods are not able to correct the error completely because they cannot obtain enough information about the shading.

A list of all parameters of the investigated models is provided in Table 3-7. It should be emphasised here that the improvement under soiling, was similarly good for all methods. This is in contrast to shading, where model and configuration-dependent models improve the prediction error to a greater and sometimes to a lesser extent.

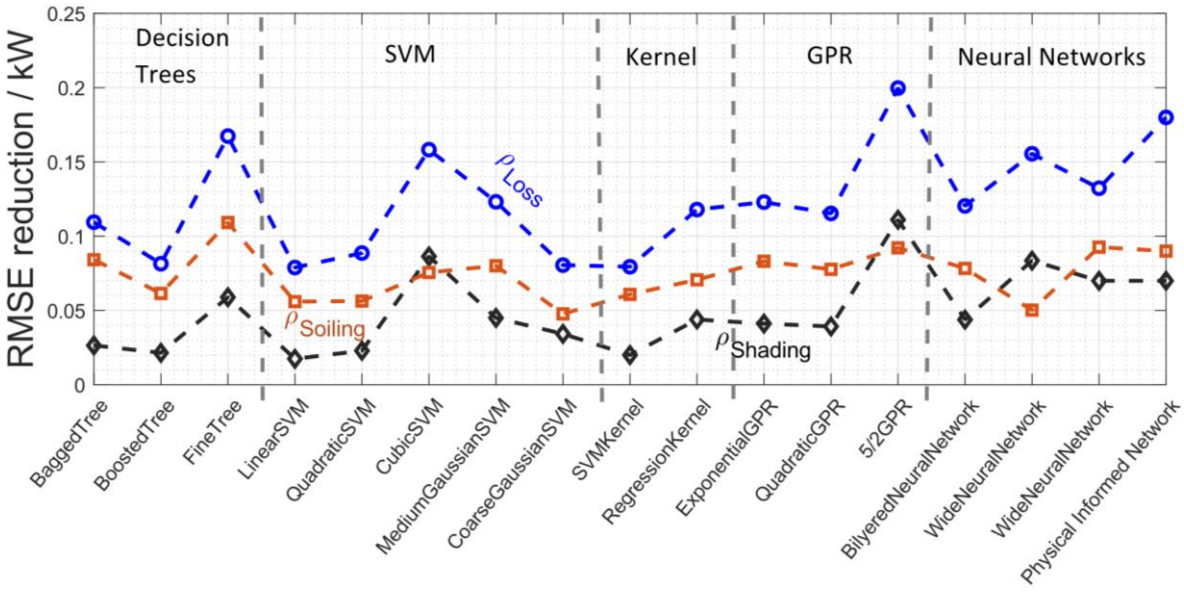

***Figure 6**. Comparison of the error reduction of different machine learning methods with different configurations*

3.3. Prediction error of the PINN network

The parameters for the Evans equation and the Faiman model are determined to be $C_1 = 4.2 \cdot 10^{-4}$,$C_2 = 0.9$ and $U_1 = 9.8$ with the training dataset. Since the forecast values are

used and the wind speed data show no correlation with the PV data, the U_2 Parameter is omitted from the Faiman equation and only the influence of irradiation is considered. It can be seen in Figure 7 that the PINN performs best here in terms of the RMSE and also provides better forecast values at times of low irradiation, which can be attributed to the Evans equation. The PINN was able to show comparable results to the reference model in the noon hours when the soiling is strongest, based on a one-month timeframe. In summary, the PINN was able to outperform the previous state-of-the-art solutions in this study by 6.6 %. The forecast RMSE over the entire period was 0.775 kW for the MLP, 0.633 kW for the corrected MLP and 0.591 kW for the PINN.

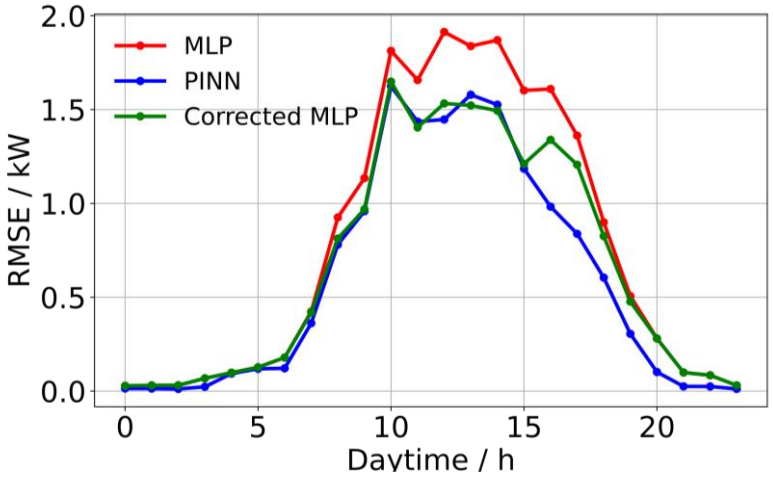

Figure 7. *Comparison of MLP, corrected MLP and PINN*

4. Conclusion and discussion

The method was tested for the first time using several yield-reducing effects simultaneously (shading and soiling). The model, which had already been extensively validated on shading scenarios, was further validated and successfully quantified soiling to improve the forecasting error. Using these data, various machine learning methods were trained and compared. A k-fold procedure was used to ensure robustness.

The dataset was split in such a way that a whole year could be used to validate the models in order to be able to include seasonal effects. As the model is based only on weather data and power data, it can also be used for other PV systems. Important here are the high-resolution irradiation and temperature data, which can accurately record the POA irradiation using appropriate sensors and thus enable the simulation of the PV power. The method was able to reduce the forecast error in day-ahead forecasts under shaded and soiled conditions for all machine learning methods examined. It should be emphasized that the physically informed networks were able to achieve a similar improvement in both soiling and shading as was achieved when correcting the forecasts. In contrast to the other models, no subsequent correction of the models was necessary here, as the loss ratios are directly included in the model. The PINN outperformed other machine learning models and configurations by 6.6 % in a shading setup, also under shaded and soiled conditions. Since the correction lowers the forecast output, there may be some negative influences on the prediction, for example if the forecast model generally underestimates the output. In this case, a retrospective correction would not lead to an increase in the forecast error, even under shading and soiling. In addition, the simulation model is also dependent on an accurate irradiation measurement. Shading of the irradiation sensor would result in the shading not being recognized and therefore not being corrected. As an outlook, further combination measurements of shading and soiling will be collected in order to validate the method even better. An extension to include other degrading effects is also planned.

Data availability statement

The authors do not have permission to share data.

Author contributions

Tim Kappler: Conceptualization, Methodology, Software, Investigation, Writing-Original Draft, Visualization, Resources. **Anna Sina Starosta**: Writing - Review & Editing. **Nina Munzke**: Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Bernhard Schwarz**: Writing - Review & Editing, Supervision, Project administration. **Marc Hiller**: Project administration, Funding acquisition.

Competing interests

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.

Funding

This work was supported by the Solarpark 2.0 project funded (funding code 03EE1135A) by the Federal Ministry for Economic Affairs and Climate Action (BMWK)

Acknowledgement

This work contributes to the research performed at KIT Battery Technology Center. The results were generated within the ”Solarpark 2.0” project. The authors thank the project management organization Jülich (PTJ) and the BMWK.

References

[1] I. R. E. Agency. [Online]. Available: https://www.irena.org/Energy-Transition/Technology/Solar-energy. [Accessed 01 02 2024].

[2] A. S. Saidi, "Impact of grid-tied photovoltaic systems on voltage stability of tunisian," *Ain Shams Engineering Journal,* March 2022.

[3] g. Y. H. C. Haoyin Ye, "State-Of-The-Art Solar Energy Forecasting Approaches: Critical Potentials and Challenges," *Frontiers in Energy Research,* 15 March 2022.

[4] J. a. Q. W. Zeng, "Short-term solar power prediction using an RBF neural network," *2011 IEEE Power and Energy Society General Meeting,* pp. 1-8, 2011.

[5] A. Dairi, F. Harrou, Y. Sun and S. Khadraoui, "Short-Term Forecasting of Photovoltaic Solar Power Production Using Variational Auto-Encoder Driven Deep Learning Approach," *Applied Science,* 10 2020.

[6] L. B. M. M. a. B. C. A. Fentis, "Short-term solar power forecasting using Support Vector Regression and feed-forward NN," in *International New Circuits and Systems Conference*, Strasbourg, 2017.

[7] A. a. K. K. a. J. P. a. M. N. a. H. M. Starosta, "A Comparative Analysis of Forecasting Methods for Photovoltaic Power and Energy Generation with and without Exogenous Inputs," in *38th European Photovoltaic Solar Energy Conference and Exhibition*, 2021.

[8] F. A. a. P. P.-H. a. E. F. F. a. L. Hontoria, "A methodology based on dynamic artificial neural network for short-term forecasting of the power output of a PV generator," *Energy Conversion and Management,* pp. 389-398, 09 2014.

[9] H. M. H. A. N. Peder Bacher, "Online short-term solar power forecasting," *Solar Energy,* 10 2009.

[10] M. C.-C. E. M. E. C.-M. A. G. F. M.-H. D. Masa-Bote, "Improving photovoltaics grid integration through short time forecasting and self-consumption," *Applied Energy,* no. 123, pp. 103-113, 2014.

[11] N. O. R. E. R. U. F. M.-d.-P. F. A.-T. J. Antonanzas, "Review of photovoltaic power forecasting," *Solar Energy,* pp. 78-111, 2016.

[12] "Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing," *Applied Energy,* p. 115023, 15 June 2020.

[13] D. M. a. M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," *Renewable and Sustainable Energy Reviews,* p. 112364, 2022.

[14] M. R. a. P. P. a. G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics,* pp. 686-707.

[15] D. L. EVANS, "Simplified Method for predicting Photovoltaic Array Output," *Solar Energy,* pp. 555-560, 1981.

[16] G. N. P. P. G. S. A. Louche, "Correlations for direct normal and global horizontal irradiation on a French Mediterranean site," *Solar Energy,* pp. 261-266, 1991.

[17] C. W. H. a. M. A. M. William F. Holmgren, "pvlib python: a python package for modeling solar energy systems," *Journal of Open Source Software,* 03 August 2018.

[18] E. I. a. A. G. a. C. I. R. a. B. T. R. a. P. B. C. Batzelis, "A State-Space Dynamic Model for Photovoltaic Systems With Full Ancillary Services Support," *IEEE Transactions on Sustainable Energy,* pp. 1399-1409, 2019.

[19] D. Faiman, "Assessing the outdoor operating temperature of photovoltaic modules," *Progress in Photovoltaics,* 21 Feb 2008.

[20] E. a. J. R. Haghighat, "SciANN: A Keras/TensorFlow wrapper for scientific computations and physics-informed deep learning using artificial neural networks," *Computer Methods in Applied Mechanics and Engineering,* p. 113552.

[21] P. T. L. L. H. Refaeilzadeh, "Cross-Validation," in *Encyclopedia of Database Systems*, Boston, MA, Springer US, 2009, pp. 532--538.

[22] T. a. S. A. S. a. M. N. a. S. B. a. H. M. Kappler, "Detection of Shading for Solar Power Forecasting Using Machine Learning Techniques," *40th European Photovoltaic Solar Energy Conference and Exhibition,* 18 09 2023.

[23] S. U. a. K. J. M. Jazayeri, "A simple MATLAB/Simulink simulation for PV modules based on one diode model," in *High Capacity Optical Networks and Emerging/Enabling Technologies*, Cyprus, 2013.

[24] C. K. L. R. G. a. S. S. Jonathan Lehmann, "Benchmark of eight commercial solutions for deterministic intra-day solar forecast," *EPJ Photovoltaics,* 1 July 2022.