# An evaluation of predictor variables for photovoltaic power forecasting

Lennard Visser[1], Tarel AlSkaif[2], and Wilfried van Sark[1]

[1] Copernicus Institute of Sustainable Development, Utrecht University, Princetonlaan 8a, 3584 CB Utrecht, The Netherlands

[2] Information Technology Group, Wageningen University and Research, Hollandseweg 1, 6706 KN Wageningen, The Netherlands

**Abstract.** Accurate forecasts of the electric power generation by solar Photovoltaic (PV) systems are essential to support their vast increasing integration. This study evaluates the interdependence of 14 predictor variables and their importance to machine learning (ML) models that forecast the day-ahead PV power production. To this purpose, we use two feature selection models to rank the predictor variables and accordingly, examine the performance change of two ML forecast models when a growing number of variables is considered. The study is performed using 3 years of data for Utrecht, the Netherlands. The results show the most important variables for PV power forecasting and identifies how many top variables should be considered to achieve an optimal forecast performance accuracy. Additionally, the best forecast model performance is found when only a few predictor variables are considered, including a created variable that estimates the PV power output based on technical system characteristics and physical relations.

**Keywords:** Photovoltaics · Solar power forecasting · Predictor variables · Machine Learning · Weather forecasts.

## 1 Introduction

In recent years, we have experienced a rapid growth of the installed capacity of solar Photovoltaic (PV) systems. Moreover, this growth is foreseen to last during this decade, which will lead to an expected increase from an installed solar PV capacity of 760 GWp in 2020 [8] to $1,800$ GWp in 2025 [11]. Due to the variable nature of power generation from PV systems, forecast models are deemed essential for effective integration in the electricity system, and subsequently, facilitate a high PV penetration level [13].

Artificial intelligence and particularly Machine Learning (ML) have already proven its value to solar forecasting in several studies [3, 15, 17]. Especially those studies that have utilized ML models to post-process numerical weather predictions (NWPs) have found acceptable forecast results, where the obtained accuracy results from the high local climate dependency.

However, the importance and contribution of the various predictor variables, which usually include meteorological variables, to forecast the PV power output are rarely studied. Instead, most studies select few predictor variables, lacking any substantial argumentation concerning their in- or excludance. Besides, the focus of literature related to PV power forecasting has been on the models, and particularly the model performance. An exemption to this concerns a study that assessed the importance of several predictor variables for estimating the PV power output [2]. Similar work related to establishing the importance of the predictor variables for forecasting is unknown to the authors.

In this paper we aim to provide insight into the most important predictor variables for day-ahead solar forecasting. We firstly assess the dependency of the PV power output per predictor variable and examine the interdependence among the predictor variables. Thereafter, we quantify the contribution of the various predictor variables on the model forecast performance.

This paper is organized as follows. The following Section 2 presents the methods. The data collection and assessment of the variable interdependence are discussed in Sections 3. Next, Section 4 presents the results. Conclusions and recommendations are given in Section 5.

## 2 Methods

### 2.1 Feature selection method

In the present study Recursive Feature Elimination (RFE) is used to select and rank the most important predictor variables for PV power forecasting. RFE is a wrapper-based feature selection method that deploys a regression or machine learning model in its core to rank the most important variables. This is accomplished by an iterative process, where the model is fitted to the variables in the training data set after which the least contributing variables are removed. The number of iterations needed depends on the step size (i.e. the number of variables that are removed in one iteration), the desired number of variables and the number of variables present in the data set. Since we aim to rank the variables, we continue this process until we have found the most important variable with a step size of one [6,16]. Furthermore, in this study two different models are used to define the variable ranking, namely a Multivariate Linear Regression (MLR) and Random Forest regression (RF) model. Moreover, in MLR the importance of a variable is per iteration set to be equal to the (regression) coefficient. Since RF is an ensemble regression model [4], the variable importance is based on the Mean Decrease Impurity (MDI). In short, the MDI is established by considering the relative contribution of all the splits for which a specific variable is responsible with respect to the overall model accuracy [6,16].

### 2.2 Forecast models

Three different models that forecast the PV power output are considered in this study. For simplicity reasons, the two models selected for feature selection (MLR and RF), are also considered for forecasting. Training data is used to set the parameters and in case of RF also to find the optimal hyper-parameters [4,15]. In addition, in this study we include the Hay-Davies transposition model as a benchmark [7,12]. Moreover, this transposition model directly estimates the PV power output based on the technical properties of the PV system as well as a number of weather variables. These variables include the Global Horizontal Irradiance (GHI), the Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI), Ambient Temperature (AT) and Wind Speed (WS). Since predictions of these variables can be obtained (see Section 3.1), it presents an alternative forecast approach.

Furthermore, the performance of the forecast models is examined by the Mean Absolute Error (MAE). To quantify these results independently of the PV system size, the MAE is expressed in percentages.

## 3 Data description and analysis

### 3.1 Data collection

Power measurements of a rooftop PV system present the target variable and are collected from January 2014 until December 2016. The PV system is located in Utrecht, the Netherlands, with an approximate (GDPR compliant) latitude and longitude of 52.1° and 5.1°. Moreover, the PV system has an installed capacity of 2295 Wp. The system is oriented due south (180°) with a tilt angle of 38°. By means of averaging one-minute power measurements, hourly production values are obtained [15].

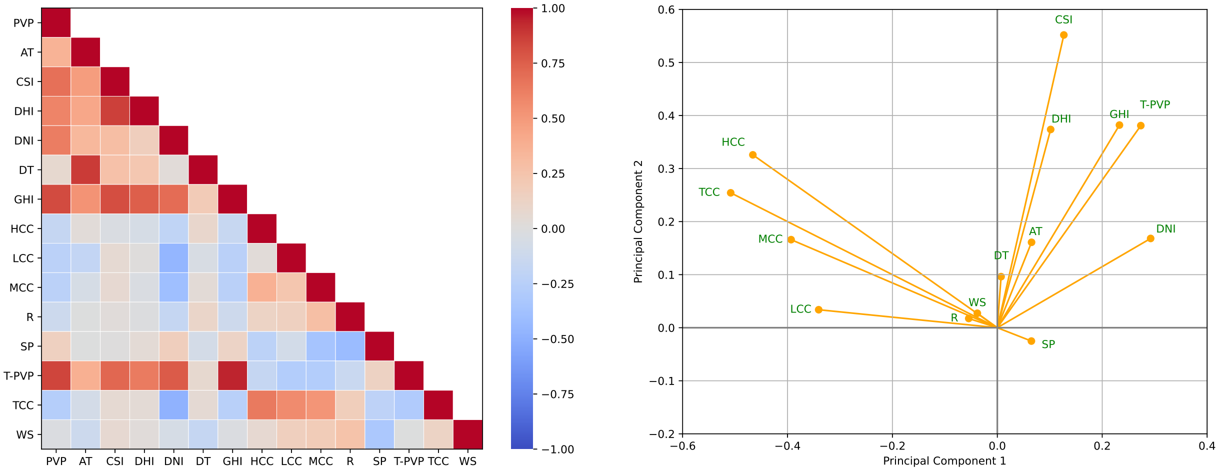
Furthermore, part of the predictor variables concern historical weather predictions. These are collected for the same period. The weather predictions are generated by the High Resolution Forecast Configuration (HRES) of the Integrated Forecast System (IFS) developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) [5]. These variables include: Ambient and Dewpoint Temperature (AT and DT); Global Horizontal Irradiance (GHI); Surface Pressure (SP); Total, Low, Mid and High cloud cover (TCC, LCC, MCC and HCC); and the zonal and meridonal wind vector components. The latter two variables are merely retrieved to estimate the Wind Speed (WS). These variables are collected as it is well-known that they affect the power production of a solar PV system either directly or indirectly [7,10,12].

To comply with the requirements and generate day-ahead PV power forecasts, the retrieved weather predictions are characterized by a 12 hour lead time, a 36 hour time horizon, an hourly temporal resolution and a daily update rate. Besides, a number of additional predictor variables are either retrieved or created. Firstly, the estimated GHI under clear sky conditions according to the Ineichen-Perez model, i.e. Clear Sky Irradiance (CSI) [9]. Secondly, the DIRINT model is used to decompose the predicted GHI into its two components, DNI and DHI [10,12]. Lastly, since we can estimate the PV power output by means of a transposition model (see Section 2.2), we include the expected or transposed PV (T-PVP) power output as an additional predictor variable.

Furthermore, since power is only generated during daytime, nighttime values are discarded from the data set. Besides, in case of any missing values in the data set, the respective timestamp is removed from the data set. Next, the data set is split into a train and test set, whereas the former comprises two years of data (2014-2015) and the latter one (2016). In addition, all variables in the training set have been normalized to obtain only values between 0 and 1. Finally, the test set has been normalized using the data characteristics of the train set.

### 3.2 Data analysis

To provide insights into the predictor and target variables, in this section first Pearson correlation is applied to define the cross-correlation amongst the predictor and target variables. The resulting correlation coefficient between a pair of variables is a measure of their linear dependence. The correlation is calculated over the combined train and test period. The dependency of the target variable (PVP) on the predictor variables are illustrated by the most left row in Fig. 1a. The strongest positive correlations for the PV power production is found for T-PVP and GHI, where correlations values of 0.85 and 0.82 are obtained respectively. This shows the dependency of PV power production on the predicted GHI and T-PVP, indicating that a higher predicted value results in an increasing PV power production. Next, a strong positive correlation for PVP is found with CSI (0.68), DNI (0.63) and DHI (0.60). On the other hand, the strongest negative correlations for PV are found for TCC (-0.27), MCC (-0.23), LCC (-0.23) and HCC (-0.17). This negative relation is simply explained as a higher cloud cover rate will decrease the solar irradiance and therefore result in a lower PV power output. Nevertheless, the reason for the limited extend of the negative correlation can be found in the little effect of the cloud cover rate on the PVP during early and late hours, where the PV power production is also low under a clear sky.



(a) (b)

Fig. 1: Interdependence of the predictor and target variables, presented by the (a) cross correlation (best viewed in color) and (b) biplot of the first and second principal components.

Besides, Fig. 1a presents the interdependence amongst the predictor variables. From the figure a strong positive correlation between GHI and T-PVP can be observed (0.94). This values shows the high dependency of T-PVP on GHI, as explained in Section 2.2. Strong positive correlations are also found among the predicted DT and AT (0.88), and the predicted GHI and CSI (0.81). In this study, the strongest yet relative limited negative correlations are found for LCC, MCC and TCC with DNI (-0.45, -0.40, -0.48). Similarly, a negative correlation is found between T and SP (-0.42).

In this study, a Principal Component Analysis (PCA) is used to further assess the interdependence amongst the predictor variables. In a PCA, the input variables are decomposed and transformed into a new feature subspace. These constructed features, i.e. Principal Components (PCs), are uncorrelated to each other, while holding information from all input variables [1,14]. Fig. 1b depicts a biplot representation of the predictor variables. Moreover, the figure marks the contribution of each predictor variables to the first two principal components, which are orthogonal to each other. Fig. 1b confirms the correlation values obtained and depicted in Fig. 1a. Moreover, it presents a high positive correlation between HCC and TCC, which have a high contribution to the first PC. Similarly, a high positive correlation is found between GHI and T-PVP, as well as DHI and CSI, which all have a high contribution to the second PC. In addition, Fig. 1b shows a negative correlation between SP and WS, which have a small contribution to the first two PCs. A negative correlation is also found for DNI with LCC, MCC and TCC.

## 4 Results and discussions

### 4.1 Predictor variable importance

An overview of the most important variables for day-ahead PV power forecasting is given in Table 1. Both recursive selection algorithms, MLR and RF, have indicated the T-PVP variable as the top predictor variable. In contrast to the others, this variable was constructed based on knowledge on the operation of this PV system. Moreover, the variables describes the expected output of a PV system by its physical dependency on the GHI, DNI, DHI, WS and AT, while also considering the systems' characteristics, e.g. tilt angle and orientation. Consequently, this shows the added value of develop and include predictor variables based on expert knowledge. Furthermore, from Table 1 it is striking that the 4 CC variables are marked of limited importance in both selection algorithms. This is due to the relative low negative correlation found between the CC variables and PV power output (see Section 3.2). Besides, the information presented by the CC variables is partly captured by the GHI, after all an increased cloud cover rate will reduce the GHI.

On the other hand, notable differences in the importance of the AT, DT, GHI, DNI and R variables can be observed from Table 1. These differences are explained by the model type, i.e. the MLR model relies on a linear relation between the predictor and target variables, while RF is a nonlinear model. A good example of this concerns GHI, which is due to its high positive correlation with T-PVP and the linear nature of the MLR model, has a limited importance as scored by the MLR model.

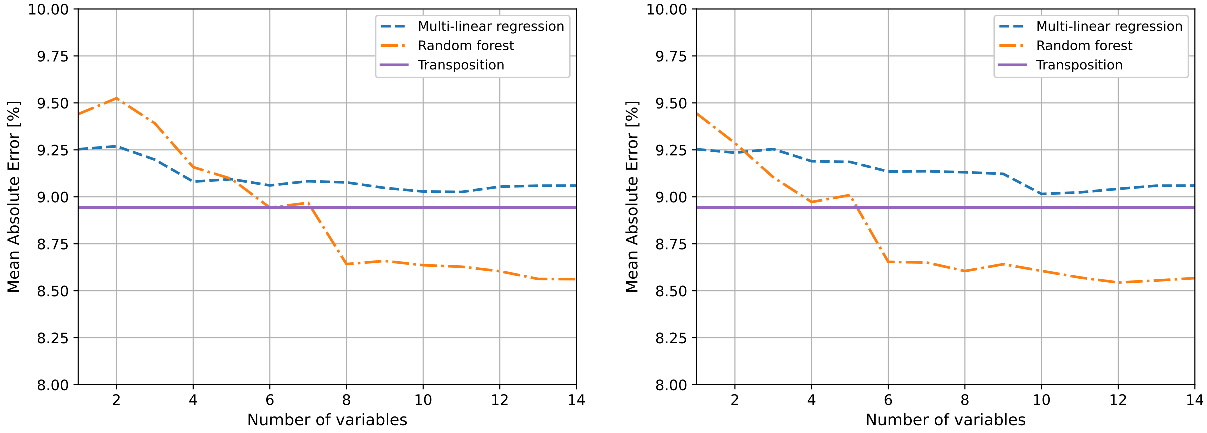
Table 1: Top predictor variables per selection method, where rank 1 presents the most important variable.

| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MLR selection | T-PVP | AT | DT | CSI | GHI | SP | DNI | TCC | DHI | WS | R | MCC | LCC | HCC |
| RF selection | T-PVP | GHI | SP | CSI | AT | TCC | R | WS | DHI | DT | LCC | MCC | DNI | HCC |

### 4.2 Forecast performance

Fig. 2 presents the forecast performance in terms of the MAE for a varying number of included variables. Moreover, the variables in Fig. 2a are based on the MLR selection model, whereas the variables in Fig. 2b come from the RF model. In general, both figures show a trend of an improving forecast, i.e. reduced MAE, with an increasing number of variables. However, some remarkable exceptions occur where the performance accuracy deteriorates with an added variable. In addition, after including 4 and 6 variables for MLR and RF, respectively, and considering their respective selection models the performances are barely found to improve with additional variables. Consequently, an optimal model performance is found by selecting the top 4 and 6 variables in case of MLR and RF. However, although the difference is limited, in the end the best forecast performances are found for both models when the RF selection model is considered. Moreover, this optimum forecast performance requires the top 10 variables in case of the MLR forecast model and 12 for the RF forecast model.

Furthermore, the poor performance of the MLR model compared to the transposition model is outstanding, especially when we keep in mind that the MLR model is fed with this exact information through the T-PVP variable. From Fig. 2, we find that in this study the best forecast performance is obtained for the RF model. Moreover, at least 6 and 8 variables are needed to outperform the Transposition model in case of considering the RF and MLR selection models, respectively. Most remarkably, in both cases the TCC variable turns out to be the missing variable in order to outperform the transposition model. Considering this significant performance improvement because of TCC, this raises the question if earlier adoption of this feature would have improved the model performance at an earlier stage (i.e., for less features).



(a) (b)

Fig. 2: Forecasting model performance according to $n$ top ranked predictor variables, (a) MLR selection and (b) RF selection. The results for the Transposition model are independent of the number of variables selected (see Section 2.2).

## 5 Conclusion and outlook

In this study, we have assessed the interdependence of 14 predictor variables and their value when utilized for PV power forecasting. The latter is reached by ranking the importance of the respective variables by two feature selection models and evaluating their contribution to the performance of two forecast models. Firstly, this study finds the RF forecast model to outperform the MLR and Transposition, if provided with an adequate selection of predictor variables. Furthermore, the results show the value of including expert knowledge within the operation of ML models to forecast the PV power production, as the top rated variable is the constructed expected power output based on the Transposition model.

In future work we will extend this study by assessing the importance of predictor variables for other locations located in different climate regions. In addition, we will create and consider an increasing number of predictor variables. Finally, by examining the forecast model performance for different data subsets that in- and exclude the created variables i.e. T-PVP, we will quantify the value of expert knowledge in PV power forecasting.

## Acknowledgements

This work is part of the eNErgy intrAneTs (NEAT) project, which is funded by the Dutch Research Council NWO in the framework of the Energy Systems Integration & Big Data programme.

## References

1. AlSkaif, T., Dev, S., Visser, L., Hossari, M., van Sark, W.: On the interdependence and importance of meteorological variables for photovoltaic output power estimation. In: 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC). pp. 2117–2120. IEEE (2019)
2. AlSkaif, T., Dev, S., Visser, L., Hossari, M., van Sark, W.: A systematic analysis of meteorological variables for pv output power estimation. Renewable Energy **153**, 12–22 (2020)
3. Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F., Antonanzas-Torres, F.: Review of photovoltaic power forecasting. Solar Energy **136**, 78–111 (2016)
4. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
5. ECMWF: European Centre for Medium-range Weather Forecasts, ECMWF (2020), https://www.ecmwf.int/en/forecasts/datasets/archive-datasets
6. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature extraction: foundations and applications, vol. 207. Springer (2008)
7. Hay, J.E.: Calculating solar radiation for inclined surfaces: Practical approaches. Renewable energy **3**(4-5), 373–380 (1993)
8. IEA: Snapshot of global pv markets 2021. Tech. rep., International Energy Agency (2021), https://iea-pvps.org/snapshot-reports/snapshot-2021/
9. Ineichen, P., Perez, R.: A new airmass independent formulation for the linke turbidity coefficient. Solar Energy **73**(3), 151–157 (2002)
10. Ineichen, P., Perez, R., Seal, R., Maxwell, E., Zalenka, A.: Dynamic global-to-direct irradiance conversion models. Ashrae Transactions **98**(1), 354–369 (1992)
11. IRENA: Renewable capacity statistics 2020. Tech. rep., International Renewable Energy Agency, Abu Dhabi (2021), https://www.irena.org/publications/2021/March/Renewable-Capacity-Statistics-2021
12. Lave, M., Hayes, W., Pohl, A., Hansen, C.W.: Evaluation of global horizontal irradiance to plane-of-array irradiance models at locations across the united states. IEEE journal of Photovoltaics **5**(2), 597–606 (2015)
13. Lorenz, E., Scheidsteger, T., Hurka, J., Heinemann, D., Kurz, C.: Regional pv power prediction for improved grid integration. Progress in Photovoltaics: Research and Applications **19**(7), 757–771 (2011)
14. Raschka, S., Mirjalili, V.: Python machine learning: Machine learning and deep learning with python. Scikit-Learn, and TensorFlow. Second edition ed (2017)
15. Visser, L., AlSkaif, T., van Sark, W.: Benchmark analysis of day-ahead solar power forecasting techniques using weather predictions. In: 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC). pp. 2111–2116. IEEE (2019)
16. Visser, L., AlSkaif, T., Van Sark, W.: The importance of predictor variables and feature selection in day-ahead electricity price forecasting. In: 2020 International Conference on Smart Energy Systems and Technologies (SEST). pp. 1–6. IEEE (2020)
17. Voyant, C., Notton, G., Kalogirou, S., Nivet, M.L., Paoli, C., Motte, F., Fouilloy, A.: Machine learning methods for solar radiation forecasting: A review. Renewable Energy **105**, 569–582 (2017)