# PV PREDICTIONS MADE EASY: FLEXIBILITY THROUGH SIMPLICITY

Marco E. T. GERARDS
University of Twente – the Netherlands
m.e.t.gerards@utwente.nl

Johann L. HURINK
University of Twente – the Netherlands
j.l.hurink@utwente.nl

## ABSTRACT

*Accurate predictions of PV output power play an important role in supporting the energy transition. This article presents an approach that aims at such predictions for PV installations on a household level. It is designed to be implemented easily on home energy management systems with low computational power. The presented prediction algorithm is self-learning, does not need physical parameters of the PV installation and can deal with changing circumstances such as objects (partially) blocking the PV panels. The method is straightforward to implement and a reference implementation in Matlab is given. An evaluation demonstrates that when the inputs used for the approach (irradiance) are correctly predicted, the predicted PV power output is also accurate.*

## INTRODUCTION

Prediction of the power output of PV installations is an important aspect to support the energy transition, especially at the house level. To facilitate upcoming technologies such as local battery storage and demand-side management, Battery Management Systems (BMS) and Home Energy Management Systems (HEMS) are crucial. These systems need to predict the local PV production to be able to make trade-offs that involve future renewable production (see, e.g., [1]). Key to such systems is that they are equipped only with low-power computational hardware. Furthermore, in domestic situations often no details on the available PV installation are available (e.g., orientation). Thus, PV prediction approaches for this domain need to be computationally light and should work well without detailed information on the PV installation.

In literature, often artificial neural networks (ANNs) or other machine learning techniques are used for PV output prediction, using environmental factors and (when available) PV installation parameters as input. However, for other domains it has been shown that performance of methods can be improved when domain knowledge is used. A famous example of such a machine learning based approach is AlphaGo [2], a program developed by Google DeepMind that beat world-class professional players at the board game Go. In the design of AlphaGo, two different ANNs are used for conceptually different parts of the game, in combination with game-specific feature detection.

The approach presented in this paper also relies on splitting different parts of the prediction process, based on domain knowledge. More precisely, the environmental factors are separated from the (unknown) characteristics of the PV installation. Furthermore, we gain additional performance improvement by separating the time-of-the-day influence. The latter is achieved by having different predictors depending on the time of day. The derived prediction algorithm is evaluated using a publically available data set.

## MODELLING

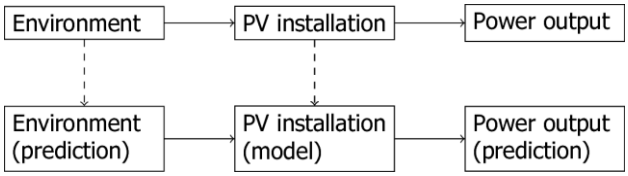


Figure 1: PV power output prediction

In general, two physical components influence the output of a PV installation: the environment and properties of the PV installation (see the top layer in Figure 1, the lower part of this figure is dealt with later). The relevant environment for the PV installation consists of parameters that determine the amount of irradiance, such as the orientation of the sun with respect to the earth (height, angle, solar power, etc.), weather conditions such as cloud coverage, smog, etc. In a first step, we assume that these parts can be well predicted. Later we revisit this assumption since in practical application these predictions are not always available for the exact location of a given PV installation and a correction for the location is needed.

The key part of this article is on the modelling of the PV installation. A PV installation is characterized by parameters such as the size of the installation, location, azimuth/angle of the panels, objects near the panels blocking the solar irradiance (e.g., trees), dirt on the panels, efficiency, etc. Note, that although the objects blocking the irradiance can also be considered as part of the environment, it is more natural to consider them to be part of the PV installation in context of this article.

To, for example, predict the power output of a commercial solar park, it may be possible to create an explicit model of the PV panels. This means that all the aforementioned parameters are collected and integrated in

a detailed model (see, e.g. [3]). For calculating the PV output with such models, a key step is to calculate from the direct and indirect solar irradiance what the irradiance is on the plane of the solar panel, by taking the angle of the sun with respect to the PV panels into account. Clearly, this changes over the year and also depends on the time of the day. This and also some other aspects imply that the resulting models get nonlinear.

However, for small scale installations, creating such explicit models is on the one hand often not a realistic option since the user may not have or provide this information, and parameters (e.g., objects blocking the solar irradiance, dirt, etc.) may change over time. On the other hand, the effort to create the model may be too large. Therefore, this article presents a modelling approach wherein these parameters are treated implicitly and learned by the presented algorithm.

## PROPOSED METHODOLOGY

In a first step we present a linear model of PV installations that contains all relevant parameters implicitly. Next, we explain how we can train the parameters of the presented model and discuss the advantages of this self-learning approach.

### Linear PV installation model

The aim of a PV installation model is to provide a way to transform the solar irradiance at a given time $t$ to a corresponding output power. However, as mentioned, in general not all parameters of the installation are given and thus predictions need to be used. We choose not to predict parameters of the model, but to directly predict the output of the model, which in our case is the output power. Mathematically, we model the PV installation as a function $p(\boldsymbol{x_{t,d}}, t, d)$, where $\boldsymbol{x_{t,d}}$ is a vector with irradiance values such as the Direct Normal Irradiance (DNI) and Global Horizontal Irradiance (GHI), $t$ is the time within the day and $d$ indicates the day. Although in general this is a nonlinear function, we choose to approximate it by a function that is linear in $\boldsymbol{x_{t,d}}$. To emphasize this, we introduce for each given time $t$ and day $d$ a function $p_{t,d}(\boldsymbol{x})$, which leads to a set of functions indexed by the time and date.

Our model of a PV installation now consists of the set of all such functions $p_{t,d}(\boldsymbol{x})$. In practice, we just determine the functions for a finite set of time intervals (e.g., 1-hour or 15-minute intervals) and a specific day $d$ for which we want to predict the PV power. This means that the model consist of a set of linear functions that give the PV output power for a given vector of irradiance values.

Each of the linear functions is described by the coefficients $c_0$ (referred to as the intercept) and $\boldsymbol{c} = [c_1, \dots, c_K]$, where $K$ is the number of irradiance parameters. To calculate the output power of the PV panel for a given irradiance vector at time $t$ of day $d$, we now may use the linear function:

$$p_{t,d}(\boldsymbol{x}) = c_0 + \boldsymbol{c}^T \boldsymbol{x}. \quad (1)$$

Here, $c_0$ and $\boldsymbol{c}$ form the model of the PV installation, while $p_{t,d}(\boldsymbol{x})$ is the prediction for time $t$ at day $d$. What remains, is to determine these linear functions.

### Training of the model

Instead of deriving the linear functions from an explicit model, we choose to train our model using past irradiance measurements obtained at the location of the PV installation and measurements of the corresponding output power.

Note, that using historical data for training is only possible when the data is in a way representative for the current situation. To make this explicit, the following assumption is needed (and later verified) within this paper: when $\boldsymbol{x}$ and $t$ are fixed, $p_{t,d}(\boldsymbol{x})$ changes only slowly when $d$ changes. This assumption implies that the parameters of the PV installation (i.e., the angle with respect to the sun, dirt on the panels and objects blocking irradiance) are commonly almost the same for consecutive days.

To train the model we require a set of "historical" irradiance values $\boldsymbol{x_{t,d}}$, $d \in \{1, \dots T\}$, and corresponding measurements $y_{t,d}$ for the PV output. These T days of training data should be not too far in the past. The historic data points lead to the following set of $T$ equations for the model:

$$y_{t,d-i} = c_0 + \boldsymbol{c}^T \boldsymbol{x_{t,d-i}}, \qquad \boldsymbol{i} \in \{\boldsymbol{1}, \dots \boldsymbol{T}\}.$$

When $T > K + 1$ this system of equations is over determined and we can calculate a best approximation by using the normal equations, i.e., the model is trained using a multivariate linear regression.

Essential in training the model is a trade-off between sufficient training days without using data that is not representative (e.g., too old).

### Summary of the method

The methodology to predict the power output of a PV installation at a time $t$ at day $d$, takes the following steps:

1) Collect a set of historical $T$ measurements of irradiance and PV output power.
2) Determine $c_0$ and $\boldsymbol{c}$ using multivariate linear regression (normal equations).
3) Obtain the environment predictions $\boldsymbol{x_{t,d}}$.
4) Predict the output power using (1).

## Advantages of the model

The presented model is self-learning since it only depends on historical measurements, instead of detailed physical parameters of the PV installation. In domestic areas where PV panels can be blocked by trees, the foliage density of the trees play a key factor. This property is automatically taken into account in the model, together with other properties such as dirt on the panels.

Training of the model requires solving the normal equations, for which many software libraries are available and only low computational power is required.

Another advantage of the methodology compared to black box machine learning (e.g. ANNs) is that it provides insight and the internal structure may be used to obtain useful results. For example, in general we expect that for the same day and time in two consecutive years the coefficients $c_0$ and $\boldsymbol{c}$ should be (nearly) the same. If we now detect large deviations of these parameters this may indicate something has changed about the PV installation. This way, the coefficients may be used to analyse the "fitness" of the PV installation. This, for example, may lead to automated detection of faults or dirt on panels.

Finally, the model makes it easy to reason about aspects that decrease the quality of the prediction when using real-world data instead of highly accurate measurements.

# DEALING WITH IMPERFECT DATA

In the previous section we assumed that accurate data needed for the training of the model is available for the precise location of the PV installation. However, for most domestic situations, such high quality data, especially for the irradiance cannot be assured. The next section discusses possible problems that may arise in such situations, together with possible solutions.

## Small range of measurement data

When the days within the training horizon just contain days with a clear sky or just days with high cloudiness (i.e., only relatively high or low irradiance values), the resulting linear function is only valid for a small range. It can be used to make predictions for irradiances outside the range for which it was trained. It is known that for such extrapolation, higher uncertainty must be taken into account.

When data for low irradiance values is missing, this may be countered by adding a zero irradiance data point (the $c_0$ coefficient) to the dataset. This value may be based on the $c_0$ values from previous periods where it was reliably determined, since we may assume that the $c_0$ value is stable for a longer time.

When the training data is in a small range, we may address the issue by increasing $T$ until the range of irradiance values is sufficiently large to construct a model that does not rely on extrapolation.

## Location of the weather prediction

In the presented model we assumed that weather measurements and predictions are available for the *exact* location of the PV installation. However, in most situations these measurements are from a weather station that is near to the installation. Consequently, there may be a time difference between certain irradiance levels at the weather station and the PV installation, e.g. because of a cloud that arrives at the weather station some time before arriving at the PV installation.

In general, when using historical data for training, fitness of the data needs to be determined before it is used. In this process, the set of available data points for training needs to be reduced to the trustable data points. Possible methods are:

- If for a data point the difference between the predicted output power and the measured output power needs is too high, i.e. $\left| y_{t,d} - p_{t,d}(\boldsymbol{x}) \right| > e$, for some predetermined threshold $e$, this higher deviation may indicate that the conditions at the location of the irradiance measurement and the PV output measurement deviate significantly, and that this data is not suitable as input for training.
- When the weather predictions/measurements are stable over a long time, the corresponding data can be considered as suitable. For example, a day with no clouds provides clean training data. Also when cloud coverage at a time t is similar to that (measured/predicted for) surrounding time intervals, this can be an indication that weather conditions are stable.

These methods can be used to select only high quality training data. To also improve the quality of the irradiance input data for the predictions, the irradiance of the weather station(s) may be translated to that of the location of the PV installation, for example by taking into account the movement of clouds, rainfall, etc. Since environment predictions are outside the scope of this article, we refer the reader to works focussing on such issues, e.g. [4].

# EVALUATION

In this evaluation we used a data set generated using [5], which is a website that uses a detailed PV model [6] to calculate the PV output power. This data set also contains historical irradiance measurements (DNI, GHI). Since the PV model is calculated this way, there are no imperfections in this data. Therefore, this model is used as the ground truth in our evaluation.

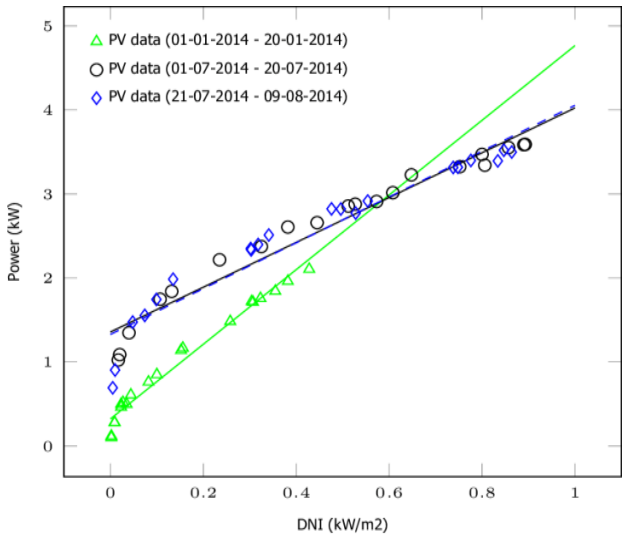


**Figure 2: correlation between irradiance and power output (13:00, 2014)**

## Verification of assumptions

The proposed methodology relies on several assumptions. One of the assumptions is that the relation between irradiance and PV power output is linear and does not change much between consecutive days. To be able to visualize this relationship, we first consider only the DNI values of the irradiance. Figure 2 shows the correlation between the DNI and PV output power for three different time periods. From this it is clear that for the relevant DNI values (i.e. values not very close to 0) the relation between DNI and the power output can be approximated quite well by a linear function. Furthermore, the two different time periods in summer have a similar relationship (the blue dashed and black graphs overlap), indicating that using recent days to train the model is promising approach. However, considering the data for January (in green), we may state that this data cannot be used to train a model for summer.

To indicate this effect on a long term, Figure 3 shows the resulting predicted coefficients for 345 days from the data set, whereby the prediction was based on the DNI and GHI ($K = 2$) of the $T = 20$ previous days. The intercept ($c_0$) is commonly a small negative number and fairly stable. This supports our conclusion that the intercept from previous models can be used for training a new model when low irradiance values are missing in the training data. Also the coefficient for the DNI is very stable, but slightly lower in summer months.

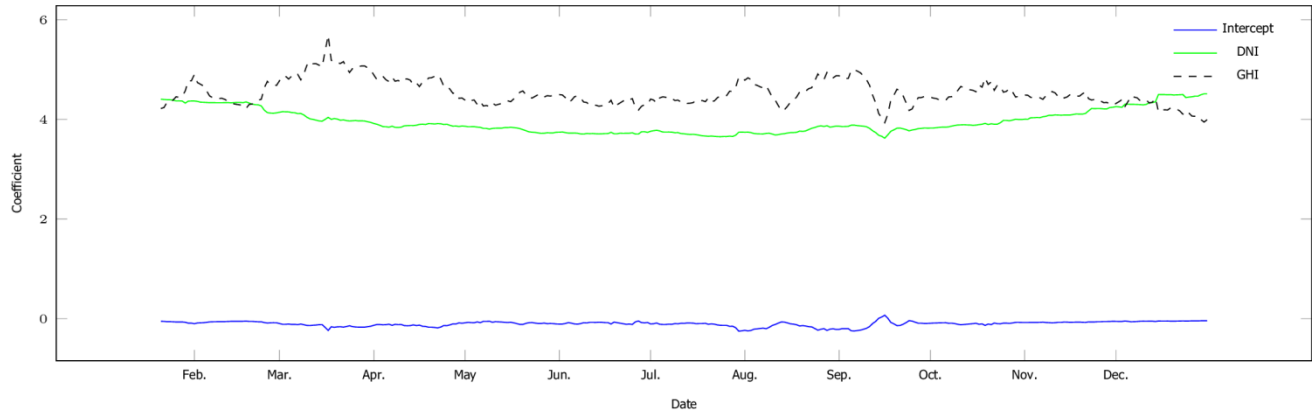


**Figure 3: coefficients for 14:00**

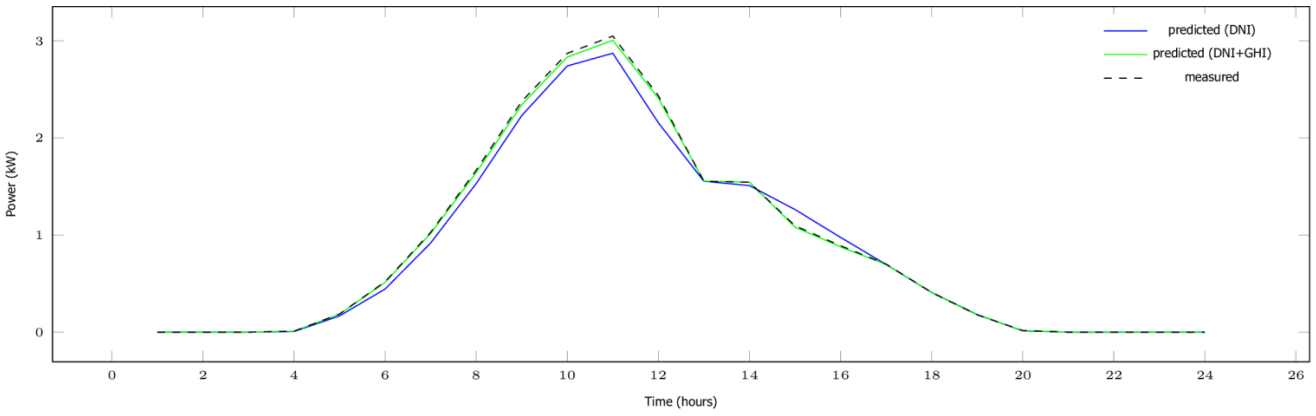


**Figure 4: PV power output prediction for 21-07-2014 based on 20 day training data**

Next, we investigate the predictive power of the model. For this, we train two different models ($K = 1$ and $K = 2$) with $T = 20$ consecutive training days in July, and use them to calculate ("predict") the PV output the subsequent day based on irradiance measurements, as shown in Figure 4. When using only the DNI as irradiance input ($K = 1$), the measured output power and the predicted output power slightly deviate, especially at low irradiance levels as we have seen before. However, when we also use the GHI ($K = 2$), the predictions and measurements closely match.

## MATLAB IMPLEMENTATION

For the input data we assume that we have the column vectors `dni` and `ghi` of size `T` with historical irradiance measurements from the last `T` days taken at the same time as needs to be predicted, together with a corresponding vector `p` with historical output power values. To train the model using Matlab, we calculate the coefficients for the intercept, DNI and GHI values (K=2) with:

```
coeff = [ones(size(p)), dni, ghi] \ p;
```

To remove low quality training data, we may remove rows from the matrix `[ones(size(p)),dni, ghi]` before applying the multivariate regression (the backslash operator).

We can multiply the three dimensional vector `coeff` with a row vector with predictions of the DNI and GHI (`pdni` and `pghi`) to predict the output power of the PV installation:

```
predp = [1, pdni, pghi] * coeff;
```

Note, that this example can be easily changed to include more (or fewer) sources of irradiance data.

## DISCUSSION AND CONCLUSIONS

This article discusses a simple and easy to adapt prediction algorithm for the power output of a PV installation. It uses (predicted) irradiance data as input, and for each time of the day a different linear predictor is determined. These linear predictors are created by using multivariate linear regression on a set of historic training data. We have shown that using the same times for recent days gives reliable results. Also, we provided methods to determine what part of the training data is unreliable, such that we can deal with measurement errors or unstable weather situations.

The results show that the modelling of the PV panel is very accurate. This high accuracy made it possible to use a simpler version of our model already successfully in several real-world projects (e.g., [7], [8]).

## REFERENCES

[1] M. E. T. Gerards, H. A. Toersche, G. Hoogsteen, T. v. d. Klauw, J. L. Hurink and G. J. M. Smit, "Demand side management using profile steering," in *2015 IEEE PowerTech*, Eindhoven, 2015.

[2] David Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature,* vol. 529, no. 7587, pp. 484-489, 2016.

[3] Z. Bouzid and N. Ghellai, "Estimation of Solar Irradiation on Inclined Surface and Design Method for an Autonomous Photovoltaic System. Application to Algeria.," in *Renewable and Sustainable Energy Conference (IRSEC)*, 2015.

[4] S. Al-Alawi and H. Al-Hinai, "An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation," *Renewable Energy,* vol. 4, no. 1, pp. 199-204, 1998.

[5] S. Pfenninger and I. Staffell, "renewables.ninja," [Online]. Available: https://www.renewables.ninja/. [Accessed 05 09 2018].

[6] S. Pfenninger and I. Staffell, "Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data," *Energy,* pp. 1251-1265, 2016.

[7] M. E. T. Gerards, J. L. Hurink and R. Hübner, "Demand side management in a field test: lessons learned," *CIRED-Open Access Proceedings Journal,* vol. 2017, no. 1, pp. 1678-1681, 2017.

[8] S. Nykamp, T. Rott, K. Keller and T. Knop, "Forecast the grid oriented battery operation to enable a multi-use approach and discussion of the regulatory framework," *CIRED-Open Access Proceedings Journal,* vol. 2017, no. 1, pp. 2760-2763, 2017.