Probabilistic reference model for hourly PV power generation forecasting

L. Alfredo Fernandez-Jimenez*, *Sonia* Terreros-Olarte, *Alberto* Falces, *Pedro M.* Lara-Santillan, *Enrique* Zorzano-Alba, and *Pedro J.* Zorzano-Santamaria

Department of Electrical Engineering, University of La Rioja, 26004 Logroño, Spain

Abstract. This paper presents a new probabilistic forecasting model of the hourly mean power production in a Photovoltaic (PV) plant. It uses the minimal information and it can provide probabilistic forecasts in the form of quantiles for the desired horizon, which ranges from the next hours to any day in the future. The proposed model only needs a time series of hourly mean power production in the PV plant, and it is intended to fill a gap in international literature where hardly any model has been proposed as a reference for comparison or benchmarking purposes with other probabilistic forecasting models. The performance of the proposed forecasting model is tested, in a case study, with the time series of hourly mean power production in a PV plant with 1.9 MW capacity. The results show an improvement with respect to the reference probabilistic PV power forecasting models reported in the literature.

1 Introduction

The global capacity of power plants based on renewable energy sources connected to the electric grids has grown strongly in recent years. The technical, economic and social benefits that these plants provide, together with government policies promoting their integration, have made this growth possible. The expected primary energy supply by 2050 will be covered by a mix of renewable sources, with the solar Photovoltaic (PV) energy as the most important with a 69% [1].

The integration of large PV plants with capacity of tens of MW makes that these facilities must participate as any other power producer in the electricity markets. The participation basically involves the submission of generation bids for the hours or periods covered by a market session. In most of the electricity markets the most important session is the daily or day-ahead session. If the schedule contained in the generation bids is not met, the PV power producer can be penalized with a lower retribution to that fixed in the market session. In order to achieve the maximum economic profit, the PV power producer must submit bids with hourly generation values as close to the actual ones as possible.

PV power generation depends on weather variables such as solar irradiance, temperature, humidity and cloud cover. The variability of these weather variables makes that power production changes at any moment. Since PV power generation is not controllable, a PV power producer needs a short-term forecasting (STF) model to prepare bids to the electricity market. Also, the forecasts provided by the model can be useful for other purposes, such as scheduling maintenance tasks. On the other hand, other agents, as the Transmission System Operators, can need short-term PV power generation forecasting models to foresee values that could be critical in order to maintain the stability of the system.

An important research effort has been carried out in the last two decades with the development of short-term forecasting models for power plants based on renewable resources, mainly wind farms and PV plants. Most of the short-term PV power forecasting models reported in the international literature provide only forecasts of the power production, that is, only the expected values. These models are known as deterministic models. An overview of the techniques used in short-term PV power forecasting models can be found in [2, 3]. Deterministic forecasting models provide a poor output, only the expected value, and since there are not error-free forecasting models, they do not provide information on how to quantify such error (difference between the forecasted value and the future real value).

Probabilistic forecasting models (PFMs) outperform deterministic ones by providing information about the uncertainty associated with the forecast. Moreover, they become an important tool because a probability distribution is needed for risk-based decision making [4]. Short-term PV power PFMs have a young history. One of the first PFMs [5] is derived from a deterministic one using weighted quantile regression conditioned to a clearness index to produce probabilistic forecasts. A prediction interval approach for the global irradiance is described in [6], which can be used to estimate the maximum deviation of the real power from the forecasted values for different weather conditions. The use of ensembles from Numerical Weather Prediction tools is proposed in [7]; an ensemble corresponds to a set of deterministic forecasts of weather variables obtained

* Corresponding author: luisalfredo.fernandez@unirioja.es

with slightly different initial conditions. An ensemble approach using the deterministic forecasts of seven different machine learning algorithms is presented in [8] assuming normal distribution for the forecast errors. In [9] the authors propose a method based on the analog ensemble approach, that is, a set of past cases with "similarity" to the recent ones. A nonparametric approach to obtain the density forecast is presented in [10]. The effect of the aggregation of time series of electricity load and the increasing share of PV power in the net load is evaluated in [11] by means of prediction intervals (PIs) in local electricity distribution grids.

Most of the reported forecasting models contrasts their prediction results with those obtained with at least one reference or baseline model. The most widely used reference model for deterministic forecasting is the so-called persistence model, however there is no accepted common reference model for PFMs.

In this paper we propose a PFM that could be used as a reference model in a similar manner as the persistence model is used to evaluate deterministic forecasting models. The proposed PFM uses minimal information and it can provide forecasts of the hourly PV power generation for any forecasting horizon.

The structure of the paper is as follows: section 2 presents existing work in reference PFMs of hourly PV power generation; section 3 describes the proposed model; section 4 presents the computational results obtained with the proposed model in a case study with a real PV plant; finally, section 5 presents the conclusions.

2 Probabilistic forecasting reference models

A recent work [12], describing guidelines for the presentation and evaluation of newly contributed forecasting models, advices the use of solid benchmarks against which the new proposed models must be tested empirically. Benchmarking is the process of comparing the forecasting results obtained with the new model with those obtained from reference models. The first reference model is the persistence one, which is a deterministic forecasting model.

The persistence model is one of the simplest methods used to forecast future values of a time series. It assumes that the conditions are the same at the current time (moment when the forecasts are produced) and at the future time. It has been used frequently as a reference model for comparison purposes (or benchmarking) in energy forecasting applications (load, wind power, electricity prices, etc.). For solar forecasting applications (irradiance or PV power), with an intraday horizon, the fraction of the power output relative to the clear-sky conditions is the variable that it considered to maintain the same value at the current time and at the future time [13]. For applications with longer forecasting horizon, the persistence model assumes that the PV power production at a specific hour in the future is the same that the last known value at the same hour of the day.

The evaluation of deterministic forecasting models is performed using a set of indicators. The two most common indicators are the well-known Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). A new prediction model, if it has something of predictive value, should improve the values of the indicators achieved by the reference models.

The main characteristics of a probabilistic forecast are accuracy, reliability, and sharpness. Since PFMs are more complex (in term of characteristics) than the deterministic ones the comparison of two PFMs models is difficult and a compromise must be met between the three characteristics. Several skill scores have been proposed in the international literature. The Continuous Ranked Probability Score (CRPS) is the skill score that evaluates jointly the accuracy, the reliability and the sharpness of the probabilistic forecasts; it is the most common skill score used in probabilistic weather forecasting [14].

A simple PFM is the climatological forecasting model. Applied to the PV power forecast, for each future hour it creates an ensemble with generation values from the past (it could be from all the available data, from the same day in the past, from the same month, etc.) corresponding to the same hour. The probabilistic forecasts are produced empirically from the elements in each ensemble (for each hour).

Although for deterministic forecasting the persistence model is considered as the first reference or baseline model, for probabilistic models there is not a reference model widely accepted. Alexandrini et al. [9] propose a PFM as a baseline for the STF of the hourly power production in several PV plants. They use a model called "Persistence Ensemble" (PeEn), which is formed with the 20 most recent measured PV power generation values at the same hour, that is, in the previous 20 days. In other work [15] the same authors use the PeEn, but now with 51 members (51 most recent values at the same hour). The number of elements in the ensembles is vaguely justified.

Zamo *et al*. [7] use as reference model a simple climatological forecasting defined as the quantile sets computed with the measured PV generation in the training data from dates with the same month that the corresponding to the future time. Thus, there are 12 reference models, one for each month of the year.

In the forecast of the solar irradiance Yang [16] proposes as reference model the "complete-history persistence ensemble" (CH-PeEn). This ensemble is composed of all historical clear-sky index measurements that share the same time of day. The model is proposed in order to avoid the subjective choice in the number of members in the ensemble. Since it uses all the historical measurements, it presents similarities with the climatological model, although it corrects the resulting distribution by multiplying it by the clear-sky global horizontal irradiance expectation to obtain the solar irradiance probabilistic forecast.

The model presented in this paper differs from those using the PeEN or the CH-PeEn, since it selects the elements for the ensembles using an exploration window which width is chosen according to the results obtained with the data of previous years: the width of the window which offers the best value of the selected skill score

with the data of previous years, is used to provide the forecasts in the recent one. Moreover, the model proposed can be used for any horizon, offering probabilistic forecasts, in the form of quantiles, for any day in the future.

The purpose of this model is to be used as a reference to check the performance of other forecasting models. It uses the minimal information, just only the time series of hourly power production in the PV plant. Therefore, it could be considered as the probabilistic equivalent to deterministic persistence model.

3 Proposed probabilistic reference model

The base of the proposed PFM for the hourly power generation in a PV plant is the following: suppose we want to forecast the PV power generation for the hour *h* of day *d* of the year. In that moment the sun will have a position in the sky very similar to the one it had, at the same hour, in the days previous to that day *d*. Moreover, the position of the sun in that moment will be very similar to the corresponding to the hour *h* of days close to the day *d* but in previous years. Since the position of the sun is almost similar, very similar values of power to those of the past can be expected for the hour *h* of the day *d* in the current year or in a future year. So, the idea is to collect these similar power values from the past to form an ensemble from which calculate the probabilistic forecast. Obviously, the forecasts will depend on the number of members in the ensemble, that is, on the definition of days close to day *d*.

The starting point for the development of the proposed model is a time series with historical hourly generation data of the PV plant. This series will constitute the training data set (there is not a training process, but we use that name because the structure of the ensemble is selected with that data set). The proposed PFM for the hourly power production in a PV plant provides, as probabilistic forecasts, a quantile set (quantiles 0.05 to 0.95 in 0.05 steps) computed with the measured PV generation values in the training data from days in other years close to the day corresponding to the future time. The selection of the close dates is carried out by means of a window with the proper width centred on the same day *d* of previous years in the training data. The performance of different window widths is evaluated with the training data, and the width of the window that offers the best probabilistic error is selected to provide the probabilistic forecast for the future time.

Suppose we have a time series of power generation values in a PV plant composed of the hourly data in *n* years. In order to produce the probabilistic forecasts for all the daylight hours of tomorrow (day *d* of the year), the procedure would be as follows:

1. The day *d* of the first year is selected from the training data set. For each of the daylight hours, the PV power generation values for the day *d* and days within a window centred on them (days *d*-*wy* to *d*+*wy*) from the rest of the years are chosen. Thus, for each hour of day *d* of the first year there is a set (ensemble) of values with power generations in similar hours in the training data set. The process is repeated for all the days *d* of the other years, so it is obtained an ensemble of power generation values for each of the daylight hours of all the days *d* of the different years in the training data set. Each ensemble has $(2 \cdot wy+1) \cdot (n-1)$ elements.
2. With the data of each ensemble (for each of the daylight hours of the *n* days selected in stage 1) the cumulative distribution function is calculated empirically obtaining afterwards 19 quantiles, from 0.05 to 0.95 in 0.05 steps.
3. Stages 1 and 2 are repeated with different window widths (*wy* values). The *wy* value that achieves the lowest CRPS with the *n* days (days *d*) belonging to the training data set is selected.
4. A similar process to the described in stage 1 is carried out again with the days *d* of the previous *n* years but considering now windows that only select previous days, that is, days from *d*-1 to *d*-*wr*, where *wr* is the width of this second window. Now, the number of elements in each ensemble is *wr* because only values from the same year are selected.
5. The cumulative distribution function is calculated with the new ensembles obtained in stage 4, and the 19 quantiles obtained.
6. Stages 4 and 5 are repeated with different values of *wr*, and it is selected the value of *wr* which achieves the lowest CRPS with the daylight hours of the *n* days belonging to the training data set.

At the end of stage 6, the width of two windows used to select the members of the ensembles for each daylight hour have been chosen: the first one used to select values from previous years (window with total width of $2 \cdot wy+1$) and the second to select values from the current year (window with width *wr*). Now, both windows are applied jointly in order to select the members in the ensembles used to empirically obtain the quantiles for the probabilistic forecast for each hour of the future day (day corresponding to the current day plus the forecasting horizon). The total number of members in each ensemble is $(2 \cdot wy+1) \cdot n+wr-ho$, where *ho* represents the forecasting horizon in days (the most recent *ho*-1 days are not considered since they belong to the future at the time the prediction is carried out).

4 Case study

The proposed probabilistic STF PV power generation model has been applied to a PV plant composed of two-axis solar trackers. The capacity of the PV plant is 1.9 MW and it is located in the north of Spain. A time series with the hourly production for 2.5 years were available (from 01/10/2008 to 31/03/2011). This was the longest hourly PV power generation time series we had at our disposal. No outlier or missing value was found in such time series.

In order to evaluate the performance of the proposed model, the available data were divided into two sets. The first one with the data of the two first years (from 01/10/2008 to 30/09/2010) was used as the training data

set. The second with the data of the last six months (from 01/10/2010 to 31/03/2011) as the testing data set.

The proposed PFM aims to forecast the hourly mean power production in the PV plant. In the context of this paper, the forecasts are carried out at any moment between the sundown of the day *d*-1 and the beginning of day *d*, and they correspond to the hourly mean power generation in the PV plant for all the hours of the future day. The probabilistic forecasts are formed by 19 quantiles for each daylight hour, from quantile 0.05 to quantile 0.95 in 0.05 steps. The quantile 0.5 is used as the deterministic forecast, for comparison purposes with the persistence model.

The methodology with 6 stages presented in the previous section was applied to all the days in the testing data set with forecasting horizons covering from 1 to 7 days. Also, a forecasting model with undetermined horizon was developed following only stages 1 to 3 (that is, using only data in the ensembles corresponding to previous years). In the selection of the windows only data from the training data set were used.

First, we will present the deterministic forecasting results where the forecasts provided by the proposed model correspond to the quantile 0.5 for each hour. Table 1 shows the results obtained with the deterministic persistence model with the data of the testing set. In this case, the forecast of the hourly power generation at hour *h* of day *d* is the same that the actual power in the previous day (or the last known day for longer forecasting horizons), and the RMSE and MAE are expressed in the row "Day *d*" (note that the forecast process is carried out in the last hours of day *d*-1). The results shown in the row "Any day in future" correspond to the RMSE and MAE obtained using as forecast for each hour in the testing data set the mean value of PV power generation at the same hour in all the days in the training data set.

Table 1. RMSE and MAE for the deterministic persistence model in the testing data set.

Horizon	RMSE (kW)	MAE (kW)
Day *d*	615.52	414.75
Day *d* +1	651.68	449.75
Day *d* +2	693.34	495.79
Day *d* +3	729.31	527.80
Day *d* +4	742.10	542.95
Day *d* +5	714.83	521.29
Day *d* +6	712.65	515.67
Any day in future	738.56	642.60

Table 2 shows the results obtained using the proposed PFM as a deterministic model, that is, using the value forecasted for the quantile 0.5 as the expected value of the PV power generation. The RMSE and MAE are lower than the obtained with the persistence model with improvements between 13.7% and 28.1% depending on the forecasting horizon. Notice that the proposed model achieves better deterministic results with any horizon in the future (forecasting horizon longer than seven days) than the persistence model with a forecasting horizon of only one day.

Table 2. RMSE and MAE for the proposed model in the testing data set.

Horizon	RMSE (kW)	MAE (kW)
Day *d*	530.95	403.36
Day *d* +1	533.04	405.05
Day *d* +2	535.19	407.61
Day *d* +3	536.75	407.51
Day *d* +4	533.42	404.34
Day *d* +5	533.29	404.21
Day *d* +6	535.05	405.46
Any day in future	571.72	424.94

The average widths of the windows for the testing data set were 28.75 days for *wy* and 26.48 days for *wr*. Table 3 shows the probabilistic forecasting indicators achieved with the proposed forecasting model and other three models proposed as reference in other works.

Table 3. Probabilistic results for the day-ahead.

Model	CRPS (kW)	RMSD	RMSE (kW)
PeEn 20	263.26	18.26	559.80
PeEn 51	271.75	28.08	550.97
Climatological	413.33	30.80	738.56
Proposed model	255.75	13.42	530.95

As shown in Table 3, the proposed model outperforms the probabilistic results obtained with the PFMs used as reference models in other works. The "PeEn 20" (20 members in the ensemble) corresponds to the reference model proposed in [9], the "PeEn 51" corresponds to the reference model proposed in [15] and the "climatological" corresponds to a PFM computed using all the PV power values in the training data set. The error indicators expressed in Table 3 are the CRPS (in order to assess the reliability and sharpness of the probabilistic forecasts, where a lower value indicates a better performance), the RMSD (in order to evaluate the spread of the forecast) and the RMSE (for the deterministic performance). The proposed model achieves better results for all the indicators than the obtained with the other three PFMs used as reference.

The RMSD values expressed in Table 3 correspond to the root-mean-square deviation from complete histogram flatness [17]. This indicator is used to asses the flatness of a rank histogram. The rank histogram is a common tool for meteorologists, who use it to quantify the reliability of a probabilistic forecast, that is, if the probabilistic forecast represents the true distribution of the predicted variable. The underlying assumption is that a probabilistic forecast is indistinguishable from the verifying observations (future true values) if these observations fall with an equal probability in segments or "bins" of the predicted variable. A rank histogram is drawn dividing the probabilistic forecast into N+1 segments of equal probability. In a good rank histogram true observation will fall with an equal probability in each of the N+1 segments, what corresponds to a flat diagram. The RMSD tries to evaluate the deviation from the flatness (in the rank histogram) and it expressed by

equation (1), where N+1 is the number of segments with equal probability, M is the total number of observations, and s_k is the number of observations in each particular segment. For our case study, we define 20 segments with 0.05 probability (the segments are limited by the 19 calculated quantiles), and the total number of observations is 2050 (all the daylight hours in the testing data set). A lower value of RMSD indicates a better spread of the probabilistic forecasts.

$$RMSD = \sqrt{\frac{1}{N+1}\sum_{k=1}^{N+1}\left(s_k - \frac{M}{N+1}\right)^2} \quad (1)$$

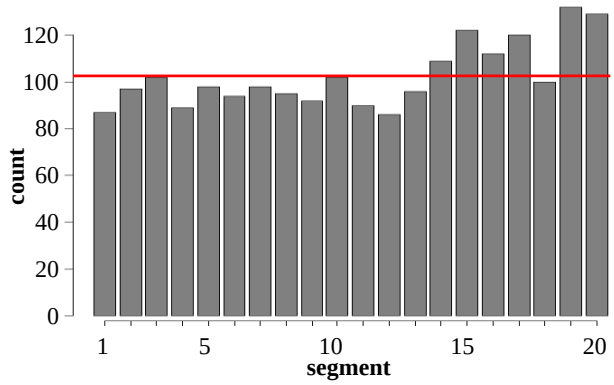

Fig. 1. Rank histogram for the testing data set.

Figure 1 plots the rank histogram corresponding to the probabilistic forecasts of all the daylight hours in the testing data set for a forecasting horizon of 1 day. The rank histogram is almost flat denoting a good spread in the forecasts (with a RMSD value of 13.42 cases). The red line denotes the ideal count of cases for each segment (5% of the 2050 hours in the testing data set).

Figure 2 plots the probabilistic forecasts for the daylight hours of five days in the testing data set with a forecasting horizon of 7 days. The days represented in the figure are from 12/12/2010 to 16/12/2010 and the forecast is carried out with data corresponding to, at least, 7 days before. In the figure are represented the quantiles 0.05, 0.25, 0.5, 0.75 and 0.95, and the real value.

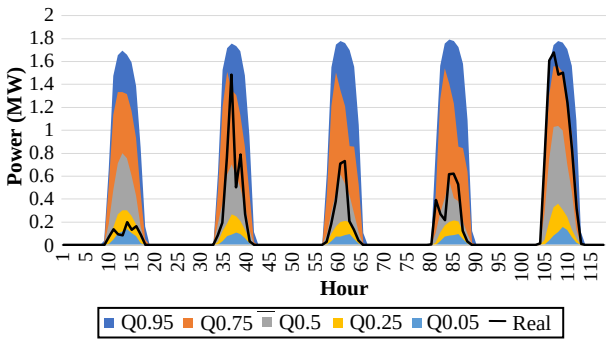

Fig. 2. Probabilistic forecast for five days in the testing data set with a forecasting horizon of 7 days.

5 Conclusions

A new PFM for PV power generation has been proposed in order to serve as reference model for benchmarking purposes. The model uses only past values of PV power generation and achieves better results than other reference PFM reported in the literature.

Further research works are underway to validate the proposed model with data from other PV plants and to enhance its forecasting performance including elements as ageing coefficients that could better adapt the power production from the past to the present.

Acknowledgments. The authors would like to thank the "Ministerio de Economia, Industria y Competitividad" of the Spanish Government for supporting this research under the project ENE2016-78509-C3-3-P and the ERDF funds of the European Union.

References

1. M. Ram *et al. Global Energy System based on 100% Renewable Energy* (LUT University and Energy Watch Group, 2019)
2. U.K. Das, K.S. Tey, M. Seyedmahmoudian, S. Mekhilef, M.Y.I. Idris, W. Van Deventer, B. Horan, A. Stojcevski, Renew Sust Energ Rev **81**, 912 (2018)
3. M.N. Akhter, S. Mekhilef, H. Mokhlis, N.M. Shah, IET Renew Power Gener **13**, 1009 (2019)
4. S.A. Fatemi, A. Kuh, M. Fripp, Renew Energy **129**, 666 (2018)
5. P. Bacher, H. Madsen, H.A. Nielsen, Sol Energy **83**, 1772 (2009)
6. E. Lorenz, J. Hurka, D. Heinemann, H. Beyer, IEEE J Special Top Earth Observ Remote Sens **2**, 2 (2009)
7. M. Zamo, O. Mestre, P. Arbogast, O. Pannekoucke, Sol Energy **105**, 804 (2014)
8. A.A. Mohammed, W. Yaqub, Z. Aung, *Intelligent Decision Technologies* (Springer, Cham, 2015)
9. S. Alessandrini, L. Delle Monache, S. Sperati, G. Cervone, Appl Energy **157**, 95 (2015)
10. F. Golestaneh, P. Pinson, H.B. Gooi, IEEE Trans Power Syst **31**, 3850 (2016)
11. D.W. van der Meer, J. Munkhammar, J. Widén, Sol Energy **171**, 397 (2018)
12. C. Croonenbroeck, G. Stadtmann, Renew Sust Energ Rev **108**, 312 (2019)
13. H.T.C. Pedro, C.F.M. Coimbra, Sol Energy **86**, 2017 (2012)
14. H. Verbois, A. Rusydi, A. Thyery, Sol Energy **173**, 313 (2018)
15. S. Sperati, S. Alessandrini, L. Delle Monache, Sol Energy **133**, 437 (2016)
16. D. Yang, Sol Energy **184**, 410 (2019)
17. N.V. Balashov, A.M. Thompson, G.S. Young, J. Appl Meteorol Climatol **56**, 297 (2017)