Probabilistic photovoltaic power forecasting model based on deterministic forecasts

L. Alfredo Fernandez-Jimenez*, *Sonia* Terreros-Olarte, *Pedro J.* Zorzano-Santamaria, *Montserrat* Mendoza-Villena, and *Eduardo* Garcia-Garrido

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

Abstract. This paper presents an original probabilistic photovoltaic (PV) power forecasting model for the day-ahead hourly generation in a PV plant. The probabilistic forecasting model is based on 12 deterministic models developed with different techniques. An optimization process, ruled by a genetic algorithm, chooses the forecasts of the deterministic models in order to achieve the probability distribution function (PDF) for the PV generation in each one of the daylight hours of the following day in a parametric approach. The PDFs, which constitute the probabilistic forecasts, are a mixture of normal distributions, each one centred in the forecasts of the selected deterministic models. The genetic algorithm chooses the deterministic forecasts, the variance of the normal distributions and their weights in the mixture. In a case study the proposed model achieves better forecasting results than the obtained with the conditional quantile regression method applied to the same data used to develop the deterministic forecasting models.

1 Introduction

The development of environment-friendly and sustainable power generation systems has led to an increasing penetration of photovoltaic (PV) power in the electricity grids. Currently, the annual growth of global solar PV capacity is greater than all other renewable power technologies combined. Global PV capacity is expected to reach 1TW by the end of 2023 [1].

PV power generation is characterized by a significative variability even in short time periods since it depends on very varying weather variables as solar irradiance, temperature, wind velocity, cloud cover, etc. A secure large-scale integration of solar PV generation in the electric grids requires the availability of PV power generation forecasting models. On one hand, short-term forecasts of the PV power generation are suitable for Transmission System Operators for regulating and dispatching tasks. On the other hand, PV power producers can use short-term forecasts to prepare bids to the electricity markets or to schedule maintenance tasks.

The development of accurate forecasting models for energy sector has been an active scientific field in recent years. Two kind of forecasting models can be identified according to the nature of the forecast: deterministic (or point forecast) and probabilistic models. Deterministic forecasting models (DFMs) offer as forecast the expected future value for the variable of interest while probabilistic forecasting models (PFMs) provide information regarding to the uncertainty in the forecast. DFMs were the first to be developed, while PFMs are more recent. The most mature subdomains in the energy forecasting sector are the deterministic short-term load forecasting and the probabilistic wind power forecasting. Solar power forecasting is identified as a subdomain with great progress in the next years [2].

Most of the published works related to PV power forecasting correspond to DFMs. Two kind of approaches have been used: statistical and machine learning methods. Das *et al.* [3] present a complete summary of the techniques used in PV power forecasting models. The development of PV power PFMs is now in its first stage, following in many cases the trail left by the development of this type of models in the wind energy subdomain.

PFMs can provide as output interval forecasts (prediction intervals) or density forecasts (forecasts of probability distribution function for the desired variable). The prediction intervals can help to the agents of an energy market to trade with low risk, although a density forecast provides a more flexible and complete prediction. A PFM can be fit using several techniques such as distribution-based forecasts, bootstrapped prediction intervals or quantile regression averaging. In this work we have focused on the distribution-based (or parametric) forecast due its easy interpretation and its low computing requirements.

The basis for the distribution-based forecasts is the assumption of the Gausssian noise in the residuals of some deterministic time series models as AR, ARIMA, etc. A PFM can easily be constructed from a deterministic model using this approach. Dudeck [4] presents an example of PFM for electricity price forecast: A Multilayer Perceptron (MLP) neural network constitutes the deterministic model and the probability forecasts are obtained from the deterministic forecasts

* Corresponding author: luisalfredo.fernandez@unirioja.es

and the error distribution on the training data set. Fatemi *et al*. [5] present a PFM for solar irradiance based on the deterministic forecast provided by a linear least mean square error estimator and beta or two-sided power distributions to obtain probabilistic forecasts. Bracale *et al.* [6] use Bayesian inference, Monte Carlo simulation and a modified Gamma distribution to develop a PFM for the hourly power generation in a PV plant.

A more sophisticated distribution-based approach using an ensemble of deterministic models is presented in [7]. The authors use seven PV power DFMs based on different machine learning techniques to form an ensemble. Quantiles for the probabilistic forecasts are obtained using three methods (linear method, normal distribution method and normal distribution method with additional features) using the forecasts of the seven DFMs.

The PFM proposed in this paper is based on 12 PV power DFMs. These DFMs were developed using diverse techniques, although all using the same training and testing data sets. Also, as a requirement, the training process of the 12 models (if required) was carried out with a cross-validation scheme with 5 folds in order to minimize overfitting. The proposed PFM provides the probability distribution function (PDF) for each daylight hour in the following day.

The inner structure of each of the deterministic models and their characteristics is out of the scope of this paper, which is focused on the construction of a PFM of the PV power generation from a set of DFMs with different characteristics and forecasts. This could be the requirement of some PV plant managers, who receive deterministic forecasts from different service providers (forecasting services), without knowing anything about the models that generate the forecasts. The proposed PFM obtains probabilistic forecasts, which could be of interest to assess the risk in decision-making issues, outperforming in this aspect to the DFMs, but also improving their own deterministic forecasting results.

The structure of the paper is as follows: section 2 describes briefly the 12 DFMs; section 3 presents the optimization process used to obtain the proposed PFM; section 4 presents the computational results obtained with the proposed model in a case study with data of a real PV plant; finally, section 5 presents the conclusions.

2 Deterministic forecasting models

The proposed model is based on the forecasts of 12 deterministic PV power forecasting models based on different techniques. The forecasts provided by the DFMs can be very different since, depending on the data used in their training or adjustment, they can "specialize" to achieve the lowest forecasting error over different ranges of the input variables. Therefore, the ensemble may contain better predictions than those of any DFM. The models were developed and trained independently, although all of them use the same input or explanatory variables. Table 1 shows the list of the explanatory or input variables used to build the DFMs. Many of the input variables correspond to weather forecasts obtained from a numerical weather prediction model. The weather forecasts are the forecasted hourly values for all the daylight hours of the following day. The output variable for all the models is the hourly power generation in a PV plant for the daylight hour *h* of the following day.

Table 1. Explanatory variables for the DFMs.

Name	Description
temp	Temperature at 2 meters (Kelvin)
swflx	Surface downwelling shortwave flux ($W \cdot m^{-2}$)
mod	Wind module at 10 meters (m/s)
dir	Wind direction at 10 meters (degrees)
rh	Relative humidity at 2 meters (per unit)
cft	Cloud cover at low and medium levels (per unit)
cfl	Cloud cover at low level (per unit)
cfm	Cloud cover at medium level (per unit)
cfh	Cloud cover at high level (per unit)
prec	Accumulated rainfall in the hour ($kg \cdot m^{-2}$)
clear	Clear-sky global horizontal irradiance ($W \cdot m^{-2}$)
aghi	Average global horizontal irradiance ($W \cdot m^{-2}$)
aip	Average irradiance on panel ($W \cdot m^{-2}$)
h1	Cosine of the day fraction for the hour *h*
h2	Sine of the day fraction for the hour *h*

The explanatory variables of Table 1 include forecasts of weather variables (temp, swflx, mod, dir, rh, cft, cfl, cfm, cfh, prec and clear) for the future hour *h* of the following day. Two other input variables correspond to calculated values (aghi and aip). The variable aghi is the average value of swflx (which corresponds to the global horizontal irradiance) and aip is the average value of the irradiance on the PV panel throughout the hour *h*. This last variable (api) is calculated, considering the solar geometry and the type of PV panel (fixed, one-axis tracker or two-axis tracker), as the sum of the direct normal irradiance (DNI) and the total diffuse irradiance on the tilted surface of the PV panel. The DNI is estimated by applying the Erbs model [8] to the forecasted swflx value and the total diffuse irradiance is estimated by means of the King model [9]. The last two variables (h1 and h2) are used to code the hour *h*.

The list of DFMs of the hourly PV power generation is the following:

1. Bayesian additive regression trees (Bart) [10]: A sum of trees model with prior regularization and Bayesian backfitting.
2. Bayesian Lasso regression averaged model (Blasso) [11]: A linear L1 regularization model.
3. Bayesian ridge regression model (Bridge): A linear L2 regularization model.
4. Bayesian regularised feed-forward neural network (Brnn) [12]: A neural network with one hidden layer and regularization with a Bayesian approach.
5. Cubist model (Cubist) [13]: A rule-based model with an associated multivariate linear model to each rule.
6. Extreme learning machine (Elm) [14]: A neural network with a single hidden layer trained with the Elm algorithm.
7. Elastic net model (Enet) [15]: A linear model with regularization and variable selection.

8. Generalized boosted regression model (Gbm) [16]: Ensemble of decision trees with a stage-wise model by optimizing the loss function.
9. Least angle regression model (Lars) [17]: Linear regression model with feature selection.
10. M5rules model (M5rules) [18]: non-parametric model with propositional regression rules extracted from model trees.
11. Projection pursuit regression model (Ppr) [19]: projection of input data into a low dimensional subspace.
12. Quantile regression neural network (Qrnn) [20]: non-linear quantile regression with a single hidden layer neural network.

The models are adjusted or trained using 5-folds cross-validation. The value of tuning parameters, if needed, is selected according to the lowest average root mean square error (RMSE) with the 5 folds used as testing sets.

3 Proposed probabilistic model

Our parametric approach is based on a mixture of normal distributions as the PDF of the PV power generation. Normal distribution is defined by two parameters, the mean and the standard deviation. The values of these two parameters are necessary for each member (distribution) in the mix. Figure 1 plots the PDF of a mix of four normal distributions; each distribution has its mean and standard deviation values. In order to form the PDF of the mix, another parameter is needed for each member, a weighting factor. The sum of the weighting factors of all members in the mix must be equal to 1. Note that the PDF of the mixture in the figure isn't Gaussian shaped.

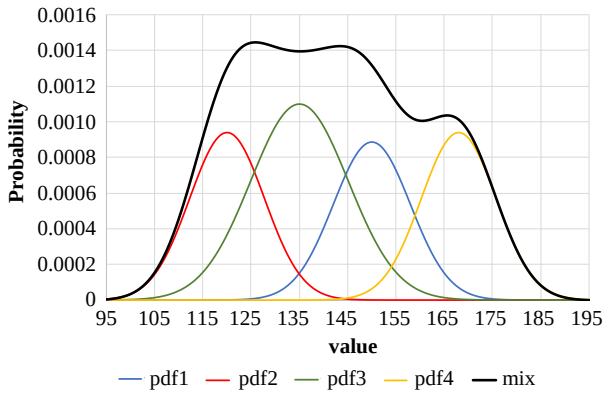

Fig. 1. PDF of a mix of four normal distributions.

The basic idea of the proposed model is to obtain the PDF of the hourly PV power generation in a PV plant as a mixture of normal distributions from the forecasts of the 12 abovementioned DFMs. But we don't necessarily require the forecasts of the 12 models, since some of them may be highly correlated and their inclusion in the mixture can only worsen the global outcome. In addition, the mean, standard deviation and weighting factor values must be selected for each of the members of the mixture. For the mean value of each member we take the forecast of the corresponding DFM. Even so, the values of a broad set of parameters need to be determined.

To carry out that selection task we propose an optimization process ruled by a genetic algorithm (GA). A chromosome including 37 genes with real-value coding is used. The meaning of each gene is shown in Table 2. The first gene is used to code the DFMs that are used in the mixture. This gene ranges 1 to 4095 and its transformation to binary number gives the set of selected DFMs. For example, a value of 39 for this gene means that the DFMs 7, 10, 11 and 12 are the selected to form the mixture (the binary transformation is 000000100111, what indicates the selected DFMs). The standard deviation for the normal distribution corresponding to each DFM is coded with two genes. That value is calculated as a linear function of the mean, as expressed in equation (1), where σ_i represents the standard deviation for the forecasts of DFM i, μ_i represents the mean of the gaussian distribution (that is, the forecast provided by the DFM i), and $c_{0,i}$ and $c_{1,i}$ are the coefficients coded into two genes. Finally, another gene is needed for each DFM to code the weighting factor.

$$\sigma_i = c_{0,i} + c_{1,i}\,\mu_i \tag{1}$$

In Table 2, genes 2 to 13 represent the proportional part of the standard deviation to the mean for each DFM ($c_{1,i}$ values). Genes 14 to 25 represent a fixed term for the standard deviation for each DFM ($c_{0,i}$ values). The last genes, from 26 to 37, represent the weighting factors for each DFM, although to apply the proposed normal mixture distribution, the weights of the selected DFM need to be normalized so that they sum 1.

Table 2. Structure of the chromosome.

Gene	Range	Meaning
1	1 to 4095	Selected DFMs
2-13	0.02 to 0.25	Proportional coefficient for standard deviation, c_1
14-25	0 to 50	Fixed coefficient for standard deviation, c_0
26-37	0.02 to 2.1	Prior weighting factor

Accuracy, reliability, and sharpness are the main characteristics of a probabilistic forecast. The selection of the best PFM is the search for a compromise between the three characteristics. The Continuous Ranked Probability Score (CRPS) is the skill score used to evaluate jointly the accuracy, the reliability and the sharpness of the probabilistic forecasts [21]. A more limited skill score is the root-mean-square deviation (RMSD) from complete histogram flatness [22]. A probabilistic forecast is indistinguishable from future true values of the predicted variable if these values fall with an equal probability in segments or "bins". For example, in the segment of PV power limited by the values of quantiles 0.6 and 0.65, the probability for the true PV power to fall in that segment is 0.05. In essence, the RMSD is an indicator of the reliability of the

probabilistic prediction. A lower RMSD value corresponds to a more reliable prediction. The RMSD is expressed by equation (2), 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 segment. The fitness function for the GA optimization could be the negative value of the CRPS with the training data set (forecasts of the 12 DFMs corresponding to the training data set), the negative value of the RMSD, or a combination of both of them.

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

4 Case study

The proposed methodology was applied to a real case: a PV plant composed of two-axis trackers with a rated capacity near 2 MW. The data available for the plant corresponded to an hourly PV power generation time series for 30 months. The data show high intra-hour variability in the PV power generation. Figure 2 plots the percentage of hours with power output variations greater than the 10% of the rated capacity of the PV plant. Only the five central hours of the day have been considered. The vertical axis of Figure 2 represents the percentage of cases (hours) which power generation value minus the value in the previous hour (difference in absolute value) is greater than 10%, 20%, etc., of the rated capacity. At least 21% of hours present variability over 10% of rated capacity for all the months. The winter months are the months with the highest variability.

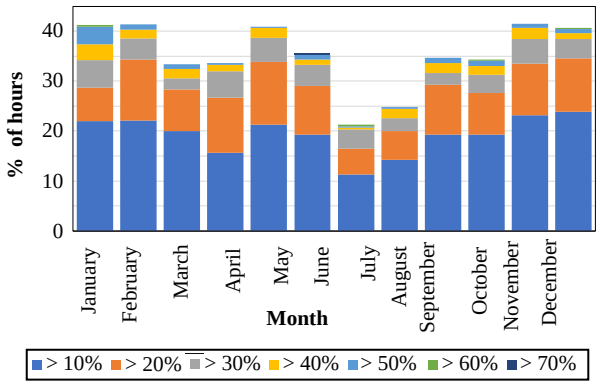

Fig. 2. Percentage of hours with power output variations over 10% of rated capacity.

Weather forecasts for all the hours in the time series were obtained from the Meteogalicia server (a regional weather forecast service). This server provided hourly forecasts of the 10 first variables (from “temp” to “prec”) shown in Table 1 in a grid format. The forecasts for the location of the PV plant were obtained by a bilinear interpolation of the forecasts for the nearest points (locations) in the grid. The forecasts for the “clear” variable were obtained using the clear sky Ineichen/Perez model [23]. The values for the last variables of Table 1 were calculated as explained in section 2. The data were divided into two sets. The first set with the data of 27 months (from October 2008 to December 2010) was the training data set. The data of the last three months were used as the testing data set.

The 12 DFMs were trained using the functions of R package caret [24]. Several combinations of parameters were tested in order to optimize the DFMs (all except models 2 and 3). For each DFM were chosen the parameter values that provided the lowest mean RMSE with the 5 folds. Once the optimal values for the parameters of the models were determined, they were trained with the complete training data set. Table 3 shows the forecasting results of the 12 DFMs with the training data set and with the testing data set. Some of the models present a similar performance (Mod2, Mod3, Mod7 and Mod9) with quite similar results for both data sets. One of the models presents an overfitting behaviour, with RMSE for training data set significative lower than those for testing data set (Mod5).

Table 3. Deterministic forecasting models results.

Model	Technique	RMSE train (kW)	RMSE test (kW)
Mod1	Bart	303.61	345.20
Mod2	Blasso averaged	367.84	379.27
Mod3	Bridge	367.85	379.44
Mod4	Brnn	343.77	352.29
Mod5	Cubist	218.49	360.61
Mod6	Elm	342.07	356.55
Mod7	Enet	367.74	378.95
Mod8	Gbm	302.41	350.13
Mod9	Lars	367.75	378.97
Mod10	M5rules	338.39	369.21
Mod11	Ppr	348.49	365.63
Mod12	Qrnn	347.67	348.99

The proposed methodology with the optimization with the GA was applied to the forecasts of the 12 DFMs. A total of 100 generations with 100 individuals per generation was carried out with the forecasts for the training data set of the 12 DFMs. Elitism was applied to the best four individuals, the crossover and mutation rates were fixed in 0.8 and 0.1, respectively.

Several tests were carried out only with the data of the training data set (forecasts of the 12 DFMs). The first test selected the negative value of the CRPS as fitness function achieving a value of 138.04 kW, but with a RMSD value of 304.35 cases. In the second test the negative value of the RMSD with the training data set was used as the fitness function; the CRPS value of the optimized model was 149.23 kW and the RMSD value was 180.12 cases. The conclusion reached with these and other tests was that both skill scores (CRPS and RMSD) had a reverse behaviour: when one got better, the other got worse. As a compromise between the two skill scores, bearing in mind that they have different measurement units, the fitness function was fixed as the negative value of the sum of the CRPS and the RMSD referred to their best values, that is, the CRPS divided by 138.04 kW and the RMSD divided by 180.12 cases.

Once the model was optimized using the new fitness function (negative value of the sum of the two skill scores referred to their best values), the model was applied to the testing data set (forecasts of the 12 DFMs for the testing data set). The DFMs selected by the GA were Mod1, Mod5 and Mod12. The probabilistic forecasting results with the testing data set are shown in Table 4. The RMSE value corresponds to the error with the mean value of the mix distribution for each hour in the testing data set. As it is shown, this RMSE value is lower than the achieved for the best of the DFMs.

Table 4. Forecasting results of Probabilistic Models.

Model	CRPS (kW)	RMSD	RMSE (kW)
Proposed model	182.83	40.11	339.40
Empirical distribution	217.85	98.92	347.27
Quantile regression	214.30	11.91	394.28

In order to assess the probabilistic forecasting results of the proposed model, two other PFMs were developed. The first one was build using the empirical distribution provided by the 12 forecasts of the DFMs: for each hour in the testing data set, the forecasts of the 12 DFMs were used as a sample of an empirical distribution. The second one was a conditional quantile regression model adjusted with the same data used to train/adjust the 12 DFMs, that is, the training data set with the variables shown in Table 1. For these new probabilistic models, quantiles from 0.05 to 0.95 in 0.05 steps were obtained and they were used to calculate the probabilistic skill scores in Table 4. The quantile 0.5 was used as the deterministic forecast for these two models. As it is shown in the table, the proposed model achieves better probabilistic forecasts (lower CRPS) and point forecasts (lower RMSE) than the obtained with the two other models. Only the value of the RMSD is worse than that obtained with the conditional quantile regression model.

5 Conclusions

In this work, we propose a methodology for generating a PFM for the hourly power production in a PV plant using point forecasts from a set of DFMs developed with different techniques. The PFM is based on a parametric approach by assuming that the PDF of the PV power production is a mix of normal distributions and that the forecasting errors on the training and testing data sets have the same distributions. The proposed PFM model is optimized by means of a GA which selects the DFMs that compose the mix, as well as the variance and weight that each normal distribution have in the final mixture. The results, in the form of a PDF for each hour, constitute the probabilistic forecasts. The mean value of the mix is used as the point forecast of the PFM. The results obtained with a testing data set, not used to build the PFM and the DFMs, show a better point forecasts than the provided for any of the DFMs and includes complete probabilistic information to assess the uncertainty associated to the forecasts.

Further research works are underway to improve the forecasting results obtained with the proposed model. Other mixture of distributions, such as truncated normal distributions, t distributions, etc., should be tested and also a dynamic structure which could select, for each day, the DFMs that constitute the mix according to their accuracy in the previous days.

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. International Energy Agency, *Renewables 2018: Analysis and Forecasts to 2023* (IEA, Paris, 2018)
2. T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, R.J. Hyndman, Int. J. Forecast. **32**, 896 (2016)
3. 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)
4. G. Dudeck, Int. J. Forecast. **32**, 1057 (2016)
5. S.A. Fatemi, A. Kuh, M. Fripp, Renew. Energy **129**, 666, (2018)
6. A. Bracale, P. Caramia, G. Carpinelli, A.R. Di Fazio, Energies **6**, 733 (2013)
7. A.A. Mohammed, Z. Aung, Energies **9**, 1017 (2016)
8. D.G. Erbs, S.A. Klein, J.A. Duffie, Sol. Energy **28**, 293 (1982)
9. M. Lave, W. Hayes, A. Pohl, C. Hansen, IEEE J. Photovolt. **5**, 597 (2015)
10. H.A. Chipman, E.I. George, R.E. McCulloch, Ann. Appl. Stat. **4**, 266 (2010)
11. C. Hans, Biometrika **96**, 835 (2009)
12. D.J.C. MacKay, Neural Comput. **4**, 415 (1992)
13. J.R. Quinlan, *Proc. of the Tenth International Conf. on Machine Learning*, 236 (1993)
14. G. Bin Huang, Q.Y. Zhu, C.K. Siew, Neurocomputing **70**, 489 (2006)
15. H. Zou, T. Hastie, J. R. Stat. Soc. Ser. B-Stat. Methodol. **67**, 301 (2005)
16. J.H. Friedman, Comput. Stat. Data Anal. **38**, 367 (2002)
17. B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Ann. Stat. **32**, 407 (2004)
18. G. Holmes, M. Hall, E. Frank, Lect. Notes Comput. Sc. **1747**, 1 (1999)
19. J.H. Friedman, W. Stuetzle, J. Am. Stat. Assoc. **76**, 817 (1981)
20. A.J. Cannon, Comput. Geosci. **37**, 1277 (2011)
21. H. Verbois, A. Rusydi, A. Thyery, Sol Energy **173**, 13 (2018)
22. N.V. Balashov, A.M. Thompson, G.S. Young, J. Appl Meteorol Climatol **56**, 297 (2017)
23. M. Reno, C. Hansen J. Stein, Sandia National Laboratories, SAND2012-2389 (2012)
24. M. Kuhn, J. Stat. Softw. **28**, 1 (2008)