THE IMPACT OF AVAILABLE DATA HISTORY ON THE PERFORMANCE OF PHOTOVOLTAIC GENERATION FORECASTING MODELS

Arthur BOSSAVY
MINES ParisTech - France
arthur.bossavy@mines-paristech.fr

Andrea MICHIORRI
MINES ParisTech - France
andrea.michiorri@mines-paristech.fr

Robin GIRARD
MINES ParisTech - France
robin.girard@mines-paristech.fr

Georges KARINIOTAKIS
MINES ParisTech - France
georges.kariniotakis@mines-paristech.fr

ABSTRACT

The continuous growth of solar power capacity raises challenges to distribution system operators regarding power quality and security of supply. Network management systems must be enhanced with short-term forecasting functionalities able to predict the solar plants production in the next hours or days. The provision of individual forecasts for each solar plant on the network is often required. To that purpose, historical measurements are needed for tuning the forecasting models. The situation is challenging for new plants for which long history of measurements is not yet available. In that case, models able to provide accurate production forecasts based on few historical production data, are required. In this paper, we investigate the performance of state-of-the-art short-term PV forecasting models as a function of the historical data available for tuning. We compare the results with those obtained by a reference model whose utilization does not require more than one day of past production data. Our analysis relies on production data from a 200 kWc solar plant located in the south-east of France. It shows that satisfactory performances can be expected from state-of-the-art models, when calibrated with no more than one or two weeks of training data.

1. INTRODUCTION

Today, solar power, and namely photovoltaic power (PV) plants capacity is undergoing a fast growth. The development of network management systems facilitating its penetration in the distribution network while ensuring power quality and supply may need individual forecasts for each plant connected to the grid. Recent research works have undertaken the development of dedicated short-term (a few hours to a few days ahead) PV forecasting models.

Generally, solar power production forecast algorithms are based on a combination of up to date meteorological forecasts and historical production data. In the literature, photovoltaic output forecasts are based on: i) meteorological forecasts coupled with a physical model, ii) time series analysis or iii) a combination of the two approaches. Research is particularly active on the second approach thanks to its possibility of using only parameters relative to the PV system as input. Examples of this approach can be found in [1], [2] and [3]. The approaches proposed in [1] and [2] are based on neural networks, while [3] proposes the use of an ARMA method coupled with a Kalman filter. An example of the third approach can be found in [4], where meteorological solar irradiation forecasts are used as an alternative to local solar irradiation measurements, in combination with a NARX neural network.

Various power system management tasks would benefit from accurate production forecasts available as soon as a considered plant is commissioned. At that early stage, only few production data is available for tuning forecasting models. Forecasts derived from advanced models may be of poor accuracy since these models generally rely on statistical methods and require, *a priori*, a large amount of data to be calibrated. On the other hand, the cost of storing this data and the computational burden of using it into forecast algorithms may also decrease the value of forecasts derived from those models. For those models to be handled appropriately, it is necessary to know what performances can be expected depending on the amount of available data.

In this paper, we analyze the performance of two state-of-the-art short-term PV forecasting models as a function of the size of the training period dedicated to their calibration. We consider both a linear and a non-linear statistical model. Historical production data as well as solar irradiation Numerical Weather Predictions (NWP) are used as input. We estimate what amount of data is needed to outperform a reference model which does not require any calibration, or more than one day of past production data. We investigate what performances increase can be expected through advanced models, while more data becomes available, then giving some clue about the considered models value. This information can then be used to optimize the design of a solar power forecasting tool, able to maintain high performances, particularly in the case of new installations with few available data.

The paper is organized as follows: in section 2, we describe the methodology of our study, i.e. the considered reference and advanced models, details about the latter's calibration, and data used. In section 3, we give results from an analysis realized with production data from a 200 kWc solar plant. Finally, conclusions and perspectives are given in section 4.

2. METHODOLOGY

In this section, we describe the forecasting models, training procedure, evaluation criteria, and the data used in our study.

State-of-the-art and reference models

In order to forecast the PV production p_{t+h} at a given time origin t for horizon h, state-of-the-art forecasting models generally combine past production data m_t, with meteorological predictions $NWP_{t+h|t}$. Such a model can be written:

$$\hat{p}_{t+h|t} = \hat{f}_h(m_t, NWP_{t+h|t}) \tag{1}$$

where $\hat{f}_h$ is the estimation of a modelling function f_h. This function can be linear as in [5], where production data has been first normalized using a so-called *clear sky model* to obtain more stationary power time series. Nevertheless, to our knowledge, the usefulness of such normalization has not been yet demonstrated. Since we get similar performance results while not considering such normalization, we will consider in this study a usual linear model:

$$f_h = c_h^1 + c_h^2 m_t + c_h^3 NWP_{t+h|t} \tag{2}$$

To capture complex non-linearities, so-called *black-box* models (e.g. Neural-Networks) are often used to forecast the PV production as in [1] and [2]. We consider such a type of model and evaluate the performances of the Random Forest (RF) algorithm [6]. The latter assumes a local constant function f_h on a plane-parallel subdivision of the inputs space. Different estimations (i.e. subdivisions) derived from bootstrap replica of data are aggregated in the last place. The Random Forest algorithm has been reported to be effective in dealing with complex non-linearities in some weather related processes [7], and has been recently used for the particular case of short-term wind power forecasting [8].

As reference model we consider a model that forecasts the production based on the previous day of production data, without any modelling function involved. We consider a mix of persistence and diurnal persistence as in [5]:

$$\hat{p}_{t+h|t} = m_t = \begin{cases} p_t, \text{ if } h \leq 2 \text{ h and } h = 24 \text{ h} \\ p_{t+h \bmod 24 - 24}, \text{ otherwise} \end{cases} \tag{3}$$

We consider the same past production data m_t to forecast the production with either the reference or the advanced models. Thus, the latter differ from the reference model by the use of additional meteorological (solar irradiation) predictions as inputs, all through a modelling function statistically calibrated.

Training procedure & evaluation

To appropriately capture the slow seasonal variations in the PV production process, we have to estimate the parameters of the modelling function f_h adaptively. Then, the calibration of the advanced models is realized through a sliding window whose size represents the training period size, and varies from 5 to 120 days in our study. The step with which the considered window moves along the time series, namely the updating step, is determined by the updating rate of meteorological predictions (i.e. daily at 12 h UTC). Such a training procedure has already been proposed for the forecasting of other non-stationary time series, such as wind speed time series [9]. In some cases, e.g. with linear models, the update of the model parameters estimation can be determined through recursive formulae, thus increasing the computing efficiency (see [5] Appendix B, or [10] for more details).

To evaluate the performances of the different models, we consider two criteria adapted to the evaluation of forecasts of the mean production level. We consider the mean forecast error, namely the bias *BIAS*, and the root mean square error *RMSE*. If $e_{t+h|t}$ denotes the forecast error at instant t for horizon h, then these two criteria are defined by:

$$RMSE_h = \sqrt{\frac{1}{N}\sum_{t=1}^{N} e_{t+h|t}^2} \quad BIAS_h = \frac{1}{N}\sum_{t=1}^{N} e_{t+h|t} \tag{4}$$

where N denotes the test set sample size.

Data

In our study, we consider historical PV production data from a 200 kWc plant located in the south-east of France. The power output time series covers a 14 months period from 06/11 to 08/12, and consists of 10-minute averages of power production records.

We consider NWP delivered by the European Centre for Medium-Range Weather Forecasts (ECMWF), at the closest grid point from the solar plant. The variables consist of surface global solar irradiation forecasts. Initially available at a 3-hour time resolution, they have been linearly interpolated to an hourly resolution, and cover forecast horizons from 1 h to 36 h ahead. Since no information is available about the orientation or the tilt of the plant, it is not possible to recalibrate a priori those forecasts from a horizontal to a well-oriented plan. Some clue about such orientation and tilt may be determined from the performances analysis (e.g. in the potential presence of a seasonal bias in the forecasts), and a recalibration procedure could be further envisaged.

3. RESULTS

In this section, we give the results of our analysis on the forecast performances evolution with the training period size for the proposed case study.

Considering a 5 days training period may seem highly insufficient to allow a reliable tuning, and satisfactory accuracy of the considered advanced forecasting models. Nevertheless, our study shows that the accuracy increases with the training period size (see Figure 1). Near optimal performance is reached with a training period size of about 20 days. The non-linear RF based model shows an overall better performance than the linear model for lower training period sizes.

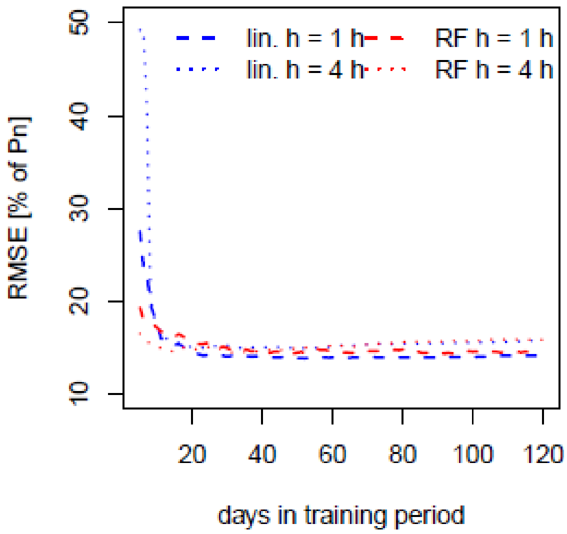

Figure 1: Performances measured through the RMSE criterion of the advanced forecasting models as a function of the training period size. Results have been normalized by the solar plant nominal power P_n. They are given here for the linear (lin.) and non-linear (RF) models and forecast horizons of 1 h and 4 h.

This can also be seen with an analysis of the lowest training period size, indicated hereafter as LBR, for which the advanced models outperform the reference model. Such a characteristic size is shown for both advanced models considered here in Figure 2**.** In this figure, the LBR is represented as a function of the forecast horizon. It is noted that the forecast horizon coincides to the hour of the day: *h* = 1 h to 6 h (i.e. 1 pm to 6 pm), and *h* = 18 h to 30 h (i.e. 6 am to 6 pm, the next day). One can notice that the non-linear RF model outperforms the reference model even with a tuning based on only 5 days of data, except for *h* = 1 h. Because of some persistence in the PV production process, forecasting at noon for 1 pm, with just the actual production level, allows satisfying performances. Getting higher accuracy level from the considered advanced models requires a calibration with a longer training period. Forecasting the production at the very first hours of daytime may also be challenging, and require longer training periods. Actually in our study, we did not observe any performance improvement with respect to the reference model for *h* = 18 h (i.e. 6 am the next day). Nevertheless, accurate forecasts of the midday production, when the production and forecast errors level may have a significant impact on the network management are considered more relevant for industrial use.

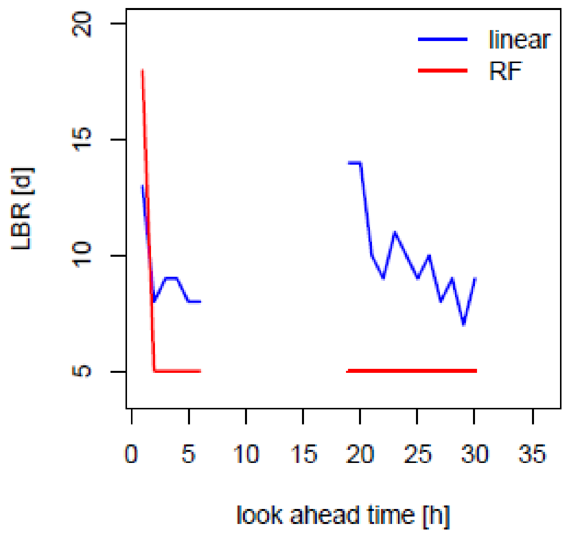

Figure 2: Lowest training period size (LBR) allowing to outperform the reference model, as a function of the forecast horizon. The considered horizons range from 1 h to 6 h and 18 h to 30 h ahead, and coincide with daytime hours.

When considering training period size higher than the LBR, the rate and maximum level to which the performances of advanced models improve, also depend on the forecast horizon. In Figure 3 is shown the performance improvement with respect to the reference model, for both the advanced models and different forecast horizons, depending on the training period size. The results associated to the first forecast day are given on the left panel, while those associated to further horizons are given in the right panel. For each forecast horizon, the performance improvement is displayed from the LBR to an optimal training period size. The latter is here defined as the lowest training period size allowing performances close to the optimal ones (with a tolerated relative decrease in performances not exceeding 1%).

The non-linear model first outperforms the reference model for very little training period size (except for *h* = 1 h), and thus with a significant improvement. The linear model requires more data to be calibrated and after a training period of more than about 10 days, it outperforms the reference model and shows a sharp performance increase until it reaches near optimal performances. The optimal performances of the two advanced models are rather similar. The linear model may perform better in forecasting the production 1 h ahead.

The results associated to the forecasts' bias showed a bias which tends to decrease with the training period size. Then, the associated decrease in performances is due to a prevailing increase in the forecast errors variability. The results also showed a higher bias for the linear (parametric) than for the non-linear (non-parametric) model. We did not observe the presence of a seasonal

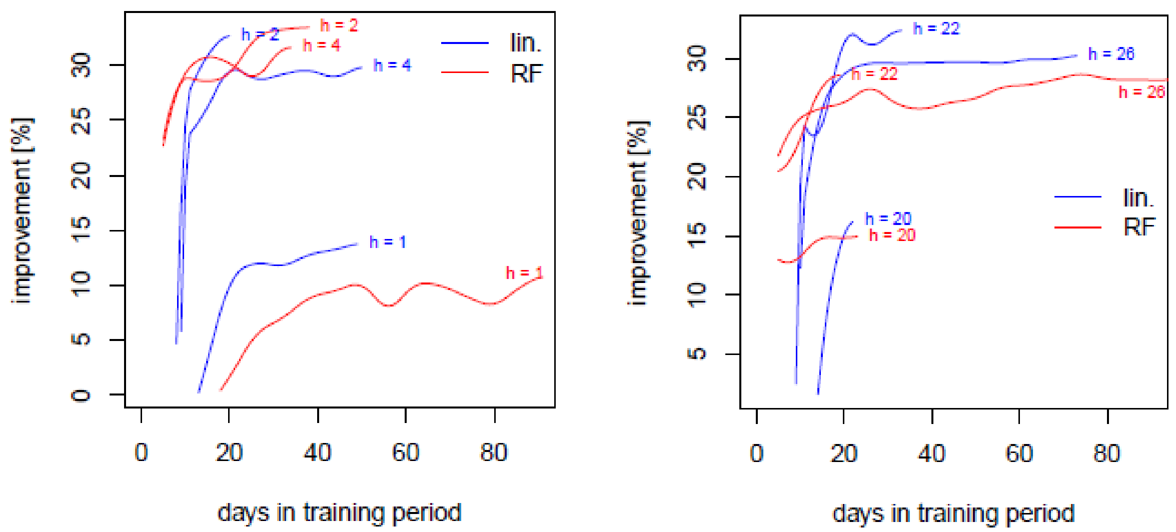

Figure 3: Performances improvement of the advanced models with respect to the reference model, for different forecast horizons *h* (in hours), as a function of the training period size. For each horizon, such improvement is displayed from the LBR to some optimal training size (see the text for more details).

bias. Then, the necessity of recalibrating solar irradiation forecasts from a horizontal to a well-oriented plan will need further investigations.

4. CONCLUSIONS

In this paper, we presented results of an analysis of the performances evolution of short-term state-of-the-art PV forecasting models with the training period size. From a 200 kWc solar plant production data, we showed that a calibration based on a few days only (about 5 to 15 days) was necessary for those models to outperform a reference model whose utilization requires no more than one day of past production data. Further analysis on different case studies should be carried out to confirm these results. The influence of using additional input should also be investigated, as it probably would require longer training period for a model to be calibrated. Finally, the work could be repeated considering probabilistic forecasts of the PV production.

AKNOWLEDGEMENTS

The authors wish to thank the personnel of ERDF Méditerranée for their support. This research is funded within the framework of the NiceGrid and Grid4EU projects. The European Centre for Medium-Range Weather Forecasts (ECMWF) is acknowledged for the provision of weather forecasts.

REFERENCES

[1] L. Ciabattoni, G. Ippoliti, S. Longhi, M. Cavalletti, M. Rocchetti, 2012, "Solar irradiation forecasting using RBF networks for PV systems with storage", *Industrial Technology (ICIT), 2012 IEEE International Conference on*, pp. 699-704.

[2] W. Fei, M. Zengqiang, S. Shi, Z. Chengcheng, 2011, "A practical model for single-step power prediction of grid-connected PV plant using artificial neural network", *Innovative Smart Grid Technologies Asia (ISGT)*, 2011 IEEE PES, pp. 1-4.

[3] M. Hassanzadeh, M. Etezadi-Amoli, M. S. Fadali, 2010, "Practical approach for sub-hourly and hourly prediction of PV power output", *North American Power Symposium (NAPS)*, pp. 1-5.

[4] T. Cai, S. Duan, C. Chen, 2010, "Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement", *Power Electronics for Distributed Generation Systems (PEDG)*, 2010 2nd IEEE International Symposium on, pp. 773-777.

[5] P. Bacher, H. Madsen, H.Aa. Nielsen, 2009, “Online short-term solar power forecasting”, *Solar Energy* vol 83, pp. 1772-1783.

[6] L. Breiman, 2001, “Random forests”, *Machine Learning* vol 45(1), pp. 5–32.

[7] R.J. Davy, M.J. Woods, C.J. Russel, P.A. Copin, 2010, “Statistical downscaling of wind variability from meteorological fields”, *Boundary-Layer Meteorology* vol 135, pp. 161-175.

[8] A. Bossavy, R. Girard, G. Kariniotakis, 2012, “Forecasting ramps of wind power production with numerical weather prediction ensembles”, *Wind Energy* (early view), DOI: 10.1002/we.526.

[9] T. Gneiting, K. Larson, K. Westrick, M.G. Genton, E. Aldrich, 2006, "Calibrated probabilistic forecasting at the stateline wind energy center - The regime-switching space-time method", *J. Am. Stat. Ass.* vol 101, pp. 968-979.

[10] H. Madsen, 2008, *Time Series Analysis*, Chapman & Hall, New York, U.S.A.