Day-ahead forecasting of regional photovoltaic production using deep learning

Pierre Aillaud
Reuniwatt
Saint-Pierre, France
pierre.aillaud@reuniwatt.com

Jérémie Lequeux
Froglabs
San Fransisco, USA

Johan Mathé
Froglabs
San Fransisco, USA

Laurent Huet
Reuniwatt
Saint-Pierre, France

Caroline Lallemand
Reuniwatt
Saint-Pierre, France

Olivier Liandrat
Reuniwatt
Toulouse, France

Nicolas Sebastien
Reuniwatt
Saint-Pierre, France

Frederik Kurzrock
Reuniwatt
Saint-Pierre, France

Nicolas Schmutz
Reuniwatt
Sainte-Marie, France

Abstract*—Power production based on solar energy is directly related to the state of the atmosphere. As the atmospheric state is undergone, this connection makes the solar energy a non-dispatchable source as opposed to controllable renewable sources such as hydroelectricity. In a context of growing photovoltaic generation, accurate forecast tools at regional scale are then increasingly important to grid operators. Indeed, forecasts allow getting information about future production over the next minutes, hours and days. Forecasting tools offer the possibility of a better grid management strategy specifically for transmission system operator (TSO) that are responsible for balancing renewable power production. High forecast accuracy could also lead to reduced costs for energy trading. In light of this situation, this study focuses on the development and analysis of a regional forecasting tool based on a deep learning approach. The selected model consists in a combination of a convolutional neural network (CNN) with a long short-term memory architecture (LSTM). The CNN layers allow extracting spatial features from Numerical Weather Prediction outputs while the LSTM part supports the temporal relationship. The day ahead regional forecast for Germany is chosen as a case study. The CNN-LSTM is compared to the classical Random Forest model known to be one of the reference techniques for this kind of problematic. Simpler deep learning models are also tested to validate the improvement brought by the CNN-LSTM architecture. All the comparisons are based on the classical root mean squared error (RMSE) metrics. The main result of this study shows that CNN-LSTM model can improve forecast accuracy when compared to state-of-the-art Random Forest. As expected, this improvement is strongly correlated to the amount of historical data which must cover several years according to the sensitivity study realized in this work.

***Index Terms*—forecast, day-ahead, deep learning, Random Forest, CNN, LSTM, NWP**

I. Introduction

The variability of solar resource is a major concern regarding the integration of photovoltaic production systems on the electrical grid. This variability can be divided into two parts. The first one is related to the diurnal cycle and is predictable in a fully deterministic manner. The second part of the variability is linked the state of the atmosphere. The evolution of this state is known to be non-linear and to exhibit a chaotic behavior [1]. In other words, the evolution of the atmospheric system is highly uncertain as a small variation in initial conditions could lead to a strong deviation in the system response.

Injection of electricity coming from the solar primary resource through the use of photovoltaic (PV) panels is, therefore, a major challenge for grid operators that have to ensure a balance between production and consumption at any time of the day [2]. In this context forecasts are required for some applications like electricity trading on the European energy market. It is then a question of providing aggregated forecasts at regional scale including the production of tens, hundreds or even thousands of PV plants. These forecasts are often provided for several look-ahead times spanning intraday and day-ahead horizons.

The brute force approach consists of modelling every PV plants located in the specified region. Such a method can work for small portfolios but is most of the time not applicable. Indeed, it implies a knowledge of the technical details of all PV plants, e.g installed capacity, tilt, orientation, that are generally not or partially available. Some methods are developed to approximate the undefined PV systems with more or less success [3, 4].

Another approach is to use supervised learning using classical methods such as linear regression, polynomial regression or even using ensemble methods such as Random Forest algorithm. In this case, access to the aggregated production history of the entire portfolio is needed. Moreover, explanatory input data, e.g. Global horizontal irradiance (GHI), are also required. The first-order physical data directly affecting production is the GHI. Then other types of meteorological data such as temperature, cloud optical thickness allow refining the production forecasts.

Among the supervised learning techniques, Random Forest based on decision trees, are widely used as it offers good performance with a small number of hyperparameters [5]. This class of method allows reducing the variance of the final model by training several decision trees and using the bagging method. The Random Forest approach is still subject to some limitations. The main limitation comes from the fact that this method does not make it possible to take into account the

978-1-7281-6115-0/20/$31.00 ©2020 IEEE

space-time correlations existing in the input data.

In this context, the work described here aims at qualifying new approaches allowing to take into account these correlations with an aim at improving the predictive performances. In recent years, the so-called Deep Learning models, based on neural networks, have gained in popularity thanks to the good performance obtained on image classification challenges [6, 7]. These approaches, unlike decision trees, take advantage of the Spatio-temporal correlations present in the data. For example, convolutional networks are very efficient in recognizing the spatial characteristics of images. As far as temporal correlations are concerned, recurrent networks are particularly well adapted.

The application of deep learning methods to the problem of regional solar power production forecast is a recent area of research for which few studies have been carried out. There is, therefore, a real need for the qualification of deep learning approaches compared to the classical techniques.

II. Methods and data

This section aims at presenting the data used for the training and evaluation as well as the forecast characteristics in Sec. II-A, the selected deep learning models in Sec. II-B and the reference model in Sec. II-C. Evaluation metric is presented in Sec. II-D followed by the training and testing procedures in Sec. II-E.

A. Data and forecast characteristics

1) Forecasts characteristics: This study focuses on day-ahead forecasts. The day-ahead forecast corresponds to a forecast covering all the horizons from $T+24H$ to $T+48H$ where T is the origin of the forecast. The temporal granularity of the forecast is 15 minutes. These day-ahead forecasts are based on NWP models. The origin of the models, noted T, is 00:00UTC and there is one forecast per day.

2) Reference data: The evaluation is based on nationally aggregated German photovoltaic production data. These public data are available on the ENTSOE-E Transparency Platform operated by the Transmission System Operators [8]. Data from 2014 to 2017 are available for this study. Over this period the total installed PV capacity has evolved from $36GW$ in 2014 to a little more than $45GW$ in 2019. The temporal granularity is 15 minutes.

3) NWP data: Different variables are available and progressively integrated into the modelling process. One of the key variables for forecasting the PV production is GHI that can be retrieved from NWP forecasts. 3 other variables have been identified for the forecasts. These are the ambient temperature [9], the Total Column Water Vapor (TCWV)[10, 11] and the snow depth (SNOD). The GHI variable is extracted from the 00:00 (UTC) run of the ECMWF model . The 3 other variables, i.e. temperature, TCWV and SNOD, are coming from the 00:00 (UTC) run of the GFS model.

4) Clearsky Data: In addition, clearsky irradiance data are also integrated as maps. These maps are built using the ESRA clear-sky model [12].

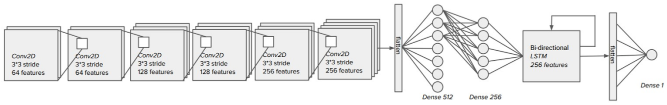

Fig. 1. Graphical representation of the CNN-LSTM model retained for this study.

B. Deep learning model

The problem posed involves Spatio-temporal data, e.g. Spatio-temporal maps from NWP forecasts, production measurements. Thus, the chosen model must be able to manage these Spatio-temporal data. Figure 1 shows the selected architecture. The data flows from the left to the right. The first layers are convolutional layers. They are here to extract the spatial features contained in the input data. The input can be multi-channel with each channel representing one variable described in Sec. II-A. The spatial features are extracted for several consecutive timesteps and then passed to a bi-directional long short term memory (LSTM) network. The LSTM is placed here to extract the temporal features of the data. Finally, a dense layer is place after the LSTM cells to combine the spatio-temporal features and provide a forecast for one time step. The model uses a time window of 12 values representing 3 hours, i.e. for each time step the model uses the last 3 hours of data. To forecast the whole day, the CNN-LSTM has to be run for each time step.

To assess the relevance of this architecture, a simple multi-layer perceptron with a hundred hidden layers has also been implemented. The VGG-16 model [13] model is also used in an attempt to qualify a kind of transfer learning. The VGG-16 is used as a pre-trained CNN based on the ImageNet dataset [6] and is connected to an LSTM that has to be trained on the dataset used in this work.

C. Reference Model

The reference model is a Random Forest that uses the same input data as the CNN-LSTM. It takes all NWP variables and clearsky maps as input and has one output that is the PV production for the current time step.

D. Evaluation metrics

All models are compared in terms of RMSE defined in Eq. 1,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - o_i\Big)^2}, \quad (1)$$

with y_i the prediction, o_i the observation, i the considered sample and n the number of prediction points. The RMSE is expressed in kW and is normalized by the total installed capacity in Germany in 2017 that is $41000kW$. The normalized RMSE is noted nRMSE.

978-1-7281-6115-0/20/$31.00 ©2020 IEEE

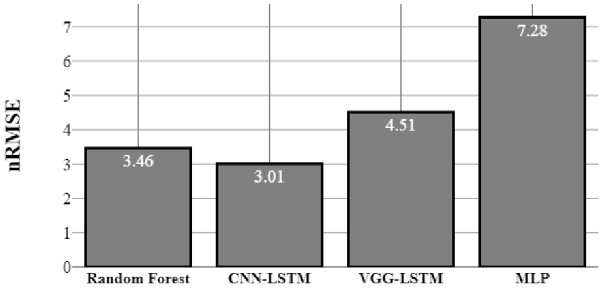

Fig. 2. RMSE normalized by the total installed PV capacity over Germany for the year 2017.

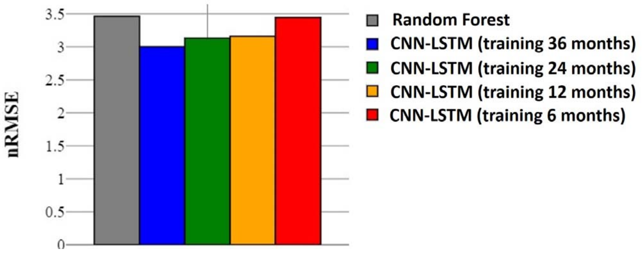

Fig. 3. The sensitivity of the CNN-LSTM network performance according to the volume of data used for training. The nRMSE is computed over the test year, i.e. 2017.

E. Training and testing procedure

The full dataset covers the period from 2014 to 2017. The year 2017 is extracted and kept as the reference test dataset that is never seen by any model before testing. Years 2014, 2015 and 2016 are used as the training/validation dataset to converge for each model to an optimal set of hyperparameters in terms of RMSE. The VGG-16-LSTM is only partially trained as the CNN part is already pre-trained on the ImageNet dataset. Therefore the weights are fixed for the CNN part and optimized through the training procedure for the LSTM and Dense layers using the PV production data of this study.

The Random Forest uses a slightly different procedure for the training as it will require a huge amount of memory to train such algorithm on 3 years of data at 15 minutes timestep. The Random Forest uses a sliding window of 3 months which has been chosen based on an evaluation on the training/validation dataset. The authors are aware that this is not consistent with the training procedure of the CNN-LSTM models but the goal is to compare a less data-intensive approach, i.e. Random Forest using 3 months as a sliding window, to a more data-intensive one, i.e. CNN-LSTM using several years. This kind of qualification will allow to select the best approach regarding the available data.

III. Results

The results obtained with the 4 models described in Secs. II-B and II-C are shown in Fig. 2. MLP model is the worst performer with an nRMSE of 7.28%. Surprisingly, the VGG-LSTM model has a relatively good performance considering the pre-training of the VGG-16 on images not related to weather data. The best deep learning model is the CNN-LSTM fully trained on weather and PV production data. this model obtains an nRMSE of 3.01%. The Random Forest performance is slightly worse than the CNN-LSTM with a nRMSE of 3.46%. For this study, it appears that deep learning models are able to compete with classical ensemble approaches but without offering a significant jump in performance as it has been observed in the past on classification problems [14].

Performance improvement obtained with the CNN-LSTM is achieved with the use of a relatively large volume of data. Figure 3 shows how the performance of the CNN-LSTM evolves when the amount of data is decreased. The classical behavior of deep learning architecture is observed, i.e. accuracy is improved when the amount of data is increased. With 6 months of data used for training, the performance is similar to the Random Forest model based on 3 months of sliding window.

IV. Conclusions

This study carried out the qualification of the deep learning approaches applied to the problem of forecasting aggregate PV production at the national level. Several deep learning architectures have been evaluated and compared to the Random Forest model chosen as reference. A basic Multi-layers perceptron (MLP) architecture was implemented as a deep learning model of relatively basic complexity. In parallel a more complex CNN-LSTM architecture has also been set up. The results on the day-ahead forecast showed that the CNN-LSTM achieves a better performance in terms of the metric considered here, i.e. RMSE. This performance improvement is directly linked to the ability of the CNN-LSTM to ingest and take advantage of a large volume of data. Indeed, while the performance is good when using 36 months for training, a sensitivity study showed that the performance of the CNN-LSTM is similar to the performance of the Random Forest when only 6 months are used for the training. According to this study, the choice of the model for the regional forecast of aggregated PV power will depend on the amount of data available.

Acknowledgment

The authors acknowledge the regional council of La Réunion and the European Regional Development Fund (ERDF) for the grant provided for this work.

References

[1] Roberto Buizza. Chaos and Weather Prediction. *ECMWF*, 2002.

[2] D. Remon, A. M. Cantarellas, J. M. Mauricio, and P. Rodriguez. Power system stability analysis under increasing penetration of photovoltaic power plants with

synchronous power controllers. *IET Renewable Power Generation*, 11(6):733–741, 2017.

[3] Y M Saint-drenan, G H Good, and M Braun. A probabilistic approach to the estimation of regional photovoltaic power production. *Solar Energy*, 147:257–276, 2017. ISSN 0038-092X. doi: 10.1016/j.solener.2017.03.007.

[4] Yves-marie Saint-drenan, Stephan Vogt, Sven Killinger, Jamie M Bright, Rafael Fritz, Roland Potthast, Mines Paristech, O I E Centre Observation, and Sophia Antipolis. Bayesian parameterisation of a regional photovoltaic model Application to forecasting. *Solar Energy*, 188(June):760–774, 2019. ISSN 0038-092X. doi: 10.1016/j.solener.2019.06.053.

[5] M. W. Ahmad, Y. Mourshed, and Y. Rezgui. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy*, 164:465–474, 2018. ISSN 03605442. doi: 10.1016/j.energy.2018.08.207.

[6] J. Deng, W. Dong, R. Socher, L-J. Li, K. Li, and L. Fei-fei. ImageNet : A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[7] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539.

[8] Lion Hirth, Jonathan Mühlenpfordt, and Marisa Bulkeley. The ENTSO-E Transparency Platform A review of Europe ’ s most ambitious electricity data platform. *Applied Energy*, 225(December 2017):1054–1067, 2018. ISSN 0306-2619. doi: 10.1016/j.apenergy.2018.04.048.

[9] S. Dubey, J. N. Sarvaiya, and B. Seshadri. Temperature Dependent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the World - A Review. *Energy Procedia*, 33:311–321, 2013. doi: 10.1016/j.egypro.2013.05.072.

[10] Q. Fu, G. Lesins, J. Higgins, T. Charlock, P. Chylek, and J. Michalsky. Broadband water vapor absorption of solar radiation tested using ARM data. *Geophysical Research Letters*, 25(8):1169–1172, 1998.

[11] C. Hill and R. L. Jones. Absorption of solar radiation by water vapor in clear and cloudy skie: Implications for anomalous absorption. *Journal of Geophysical Research*, 105:9421–9428, 2000.

[12] C. Rigollier, O. Bauer, and L. Wald. On the clear sky model of the ESRA - European Solar Radiation Atlas with respect to the Heliosat method To cite this version : HAL Id : hal-00361373. *Solar Energy*, 68:33–48, 2000.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[14] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber. Multi-Column Deep Neural Network for Traffic Sign Classification. *Neural Networks*, 32:333–338, 2012.