Exploring the Limits of Machine Learning in the Prediction of Solar Radiation

Giovanni Scabbia, Antonio Sanfilippo, Daniel Perez-Astudillo,
Dunia Bachour, and Christos Fountoukis

Abstract

Predicting solar radiation at diverse time horizons is crucial for optimizing solar energy integration, ensuring grid stability, and regulating energy markets. Two main levels of time granularity are usually recognized as requiring different treatment: *solar nowcasting* for predictions up to 6 h, and *solar forecasting* for predictions beyond 6 h. Solar nowcasting typically relies on machine learning methods, while Numerical Weather Prediction (NWP) models are considered better suited for solar forecasting. The goal of this study was to explore the limits of machine learning in solar forecasting. Our results show that machine learning methods can be profitably used for predicting solar radiation beyond 6 h, with comparable performances to NWP models for day-ahead solar forecasting.

Keywords

Solar radiation forecasting • Autoregressive modeling • Machine learning • Differencing

1 Introduction

In order to meet the future demand for power as well as to diversify the energy mix, Qatar is currently investing into the development of a 700-MWh solar power plant, with plans to exceed 20% in dependency on solar energy by 2030. One of the most important challenges for the near-future energy supply will be the reliable integration of photovoltaic (PV) sources into the present structure of the grid, which can be affected by the sudden variations and uncertainty in solar irradiance caused by meteorological changes. Predictions of solar radiation in both near real time (nowcasting) and longer term (e.g., day ahead) are therefore of primary interest for the network operators to reliably maintain a continuous equilibrium of the power supply/demand balance over the system, and to predict unexpected production deviations that could lead to grid instabilities. Furthermore, predicting the output capacity of PV systems is also crucial for the optimal management of the power storage infrastructure and to lower the overall system complexity and cost.

The choice of a modeling approach for solar forecasting tends to be determined by the prediction horizon (e.g., minutes, hours, days) needed for the application of focus (Inman et al. 2013; Diagne et al. 2013). Stochastic and machine learning models (Mellit and Kalogirou 2008) and satellite cloud motion vectors analyses (Hammer et al. 1999) are typically used for near-real-time and short-term predictions (e.g., from 1 min to 6 h). Due to their lower temporal resolution, Numerical Weather Prediction (NWP) models are better suited for longer-term prediction (e.g., days ahead) (Diagne et al. 2013; Fountoukis et al. 2018).

Several machine learning techniques have been used in solar nowcasting (Diagne et al. 2013; Mellit and Kalogirou 2008; Inman et al. 2013; Pelland et al. 2013; Mellit et al. 2006). Different methods have been developed to enhance the performance of these algorithms. These include: wavelet-based denoising (Lyu et al. 2014), ensemble and multi-modeling strategies (Chaouachi et al. 2010; Mohammed et al. 2017; Sanfilippo et al. 2016a), statistical detrending (Akarslan and Hocaoglu 2016; Sanfilippo et al. 2016b; Sanfilippo 2019), and multivariate modeling prediction (Sfetsos and Coonick 2000).

The objective of this work was to test whether machine learning methods can be extended to longer horizons to provide reliable performance for solar predictions beyond 6 h and day ahead. As a case study, we assessed the performance of univariate autoregressive linear models for solar forecasting in Qatar at 12 step-ahead horizons with different step durations, including 5, 15, 30 and 60 min.

G. Scabbia (✉) · A. Sanfilippo · D. Perez-Astudillo · D. Bachour ·
C. Fountoukis
Qatar Environment and Energy Research Institute, Qatar
Foundation—HBKU, Doha, Qatar
e-mail: gscabbia@hbku.edu.qa

E. Heggy et al. (eds.), *Sustainable Energy-Water-Environment Nexus in Deserts*,
Advances in Science, Technology & Innovation,
https://doi.org/10.1007/978-3-030-76081-6_46

2 Materials and Methods

For this analysis, we used solar radiation data collected by the Qatar Environment and Energy Research Institute (QEERI) in the period spanning from January 2014 to December 2016 over Education City in Doha (25.33° N, 51.43° E) (Perez-Astudillo and Bachour 2014). In particular, we focused on the Global Horizontal Irradiance (GHI) measured from the institute's monitoring station, sampled every second and recorded as 1-min averages in Watt per square meter (W/m^2). We first performed quality checks on the data to ensure that incorrect readings and missing values were discarded and corrected by applying a one-week-long moving average interpolation function.

One of the key conditions of statistical and machine learning forecasting models is for the time series to be stationary (i.e., a signal with time-invariant mean and variance). To normalize the GHI measure, we derived the clearness index (k_t) by computing the ratio of the GHI to the incoming solar radiation on a horizontal surface at the top of the earth's atmosphere. Also, to reduce the air mass and zenith angle dependency, we normalized k_t by the standard clearness global irradiance profile, obtaining the normalized clearness index k_t'. We then derived the k_t' at different time steps by averaging the 1-min-long time series at 5-, 15-, 30- and 60-min steps. We finally filtered out observations with solar zenith angles greater than 80° to discard samples with low accuracy and ensure that only daytime readings for k_t' are selected. Table 1 presents a brief summary of the descriptive statistics of the dataset at the different time steps.

For this study, we used univariate linear learning techniques such as the autoregressive (AR) model, which outputs a linear weighted combination of past observations of the input variable. We referred to AR (p) as the p-order AR model, which considers p lagged values as input of the regression. We selected the best lag length accordingly to the case-specific evaluation of the Akaike (AIC), Bayesian (BIC) and Hannan–Quinn (HQIC) Information Criteria (Burnham and Anderson 2004; Hannan and Quinn 1979). The autoregressive coefficients are computed using the unconditional maximum likelihood estimation method over a training subset of the data. The model performance measure was instead evaluated over a test subset of the available data using standard metrics such as the relative Root Mean Square Error (rRMSE) normalized respect to the mean value of the measured data. The implementation was carried out in Python using the *sklearn* and *statsmodels* libraries.

To obtain a reliable and generalized evaluation of the model performance, we employ a tenfold rolling-window cross-validation procedure. Cross-validation allows obtaining a generalized result independent of the explicit statistical distribution of the dataset, and of the time frame selected in the specific fold. Furthermore, the rolling-window approach ensures that the temporal autocorrelation of the series is preserved, and avoids the so-called look-ahead bias, which would occur if we were to use future data during the training process. Each model was trained on a fold subset (80% of the fold size), and its predictive power was tested on the remaining portion of the fold that was not used for training. The model forecasts the next 12 forward horizons which were compared with the true values from the test set. Finally, the persistence model was used as baseline to compare the performance of each different model considered in this study. This naive predictor assumes that future values of k_t' are equal to the variable observed at the previous time step.

3 Results

Our results, summarized in Fig. 1, show that the linear autoregressive models are able to predict the variation of k_t' up to 12 steps-ahead with better accuracy than the baseline persistence model, for all horizons (5/15/30/60 min). Most importantly, forecasts with relatively longer time step that generate day-ahead predictions (i.e., 60 min) show an error rate that is comparable to the performance of day-ahead forecasting with NWP models. For example, the day-ahead GHI forecast was presented in Fountoukis et al. (2018) for the same venue reports rRMSE values for the months of May, August and January (13.1, 12.1 and 21.5%) that are very close to the rRMSE of day-ahead k_t' predictions, i.e., the 12th step-ahead at 60 min-step (18.2, 8.5, 19.6%).

We also found that the difference in performance between the persistence and autoregressive models narrows as the time-step duration shortens. This is an expected result, as signals that are closer in time tend to be more similar. Also expected is the better performance of the AR model during the summer months (Apr–Sept) which present lower weather variability than the local winter months (Oct–Mar). Finally, the somewhat peculiar results in the month of March (10% lower mean k_t' and double to triple higher standard deviation) that affect the overall performance of the model are explained by the severe dust storms that hit Qatar in that period (Fig. 2).

Table 1 Summary of descriptive statistics

Variable: k_t'	5 min	15 min	30 min	60 min
Obs.	137,835	45,929	23,018	11,458
Mean	0.678	0.678	0.678	0.678
Std. Dev.	0.135	0.129	0.126	0.123

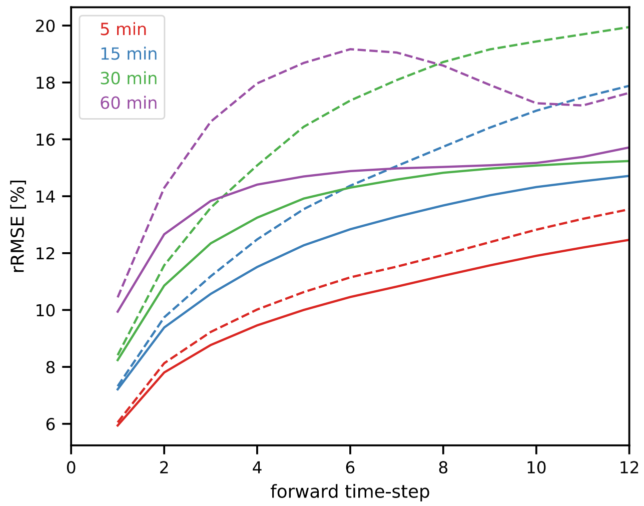

Fig. 1 Prediction rRMSE by horizon for the persistence model (dotted line) and the AR model (full line) for the different time-step durations (night values removed)

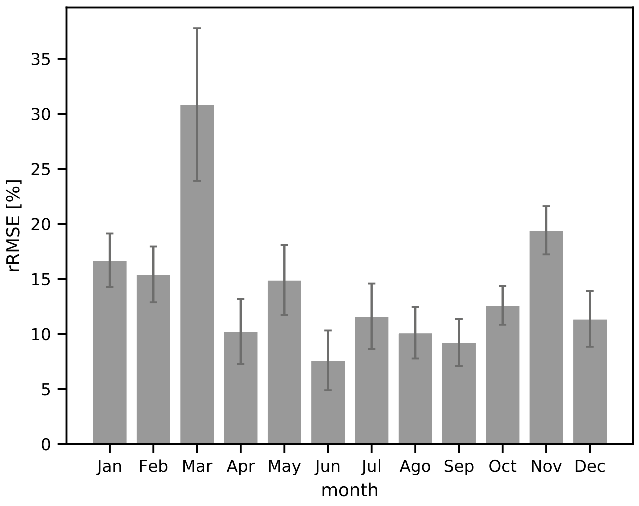

Fig. 2 Average prediction rRMSE by month and across horizons of the AR model (bars represent standard deviation of the mean)

4 Conclusion

The overall results of this analysis indicate that univariate autoregressive models provide reliable performance for the forecasting solar irradiance in Qatar both in near real time and day ahead, with rRMSE values in the 5–16% interval. Of particular interest is the ability of the autoregressive day-ahead predictions to yield comparable performance with NWP models.

Future work will be directed at improving the accuracy in forecasting solar radiation by including the use of exogenous variables (e.g., temperature, humidity, pressure, wind speed and direction) in addition to solar radiation measurements. Furthermore, we will also test the prediction performance of multi-modeling approaches where specific forecasting configurations are determined dynamically for each choice of time series input.

References

E. Akarslan, F.O. Hocaoglu, A novel adaptive approach for hourly solar radiation forecasting. Renew. Energy **87**, 628–633 (2016)

K.P. Burnham, D.R. Anderson, Multimodel inference: understanding AIC and BIC in model selection. Sociol. Methods Res. **33**, 261–304 (2004)

A. Chaouachi, R.M. Kamel, K. Nagasaka, Neural network ensemble-based solar power generation short-term forecasting. J. Adv. Comput. Intell. Intell. Inf. **14**(1), 69–75 (2010)

M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renew. Sustain. Energy Rev. **27**, 65–76 (2013)

C. Fountoukis, L. Martín-Pomares, D. Perez-Astudillo, D. Bachour, I. Gladich, Simulating global horizontal irradiance in the Arabian Peninsula: sensitivity to explicit treatment of aerosols. Sol. Energy **15**, 347–355 (2018)

A. Hammer, D. Heinemann, E. Lorenz, B. Lückehe, Short-term forecasting of solar radiation: a statistical approach using satellite data. Sol. Energy **67**(1–3), 139–150 (1999)

E.J. Hannan, B.G. Quinn, The determination of the order of an autoregression. J. Roy. Stat. Soc.: Ser. B (Methodol.) **41**, 190–195 (1979)

R.H. Inman, H.T.C. Pedro, C.F.M. Coimbra, Solar forecasting methods for renewable energy integration. Prog. Energy Combust. Sci. **39** (6), 535–576 (2013)

L. Lyu, M. Kantardzic, E. Arabmakki, Solar irradiance forecasting by using wavelet based denoising, in *2014 IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES)* (IEEE, 2014), pp. 110–116

A. Mellit, S.A. Kalogirou, Artificial intelligence techniques for photovoltaic applications: a review. Prog. Energy Combust. Sci. **34**, 574–632 (2008)

A. Mellit, M. Benghanem, S.A. Kalogirou, An adaptive wavelet-network model for forecasting daily total solar-radiation. Appl. Energy **83**(7), 705–722 (2006)

A.A. Mohammed, W. Yaqub, Z. Aung, Probabilistic forecasting of solar power: an ensemble learning approach, in *International Conference on Intelligent Decision Technologies* (Springer, Cham, 2017), pp. 449–458

S. Pelland, J. Remund, J. Kleissl, T. Oozeki, K. De Brabandere, Photovoltaic and solar forecasting: state of the art. IEA PVPS Task 14, Subtask 3.1. Report IEA-PVPS T14–01: October 2013

D. Perez-Astudillo, D. Bachour, DNI, GHI and DHI ground measurements in Doha, Qatar. Energy Procedia **49**, 2398–2404 (2014)

A. Sanfilippo, Solar nowcasting, in *Solar Resources Mapping* (Springer, Cham, 2019), pp. 353–367

A. Sanfilippo, L. Martin-Pomares, N. Mohandes, D. Perez-Astudillo, D. Bachour, An adaptive multi-modeling approach to solar nowcasting. Sol. Energy **125**, 77–85 (2016a)

A. Sanfilippo, L. Pomares, D. Perez-Astudillo, N. Mohandes, D. Bachour, Optimal selection of training datasets for solar nowcasting models, in *Proceedings to the 32nd European Photovoltaic Solar Energy Conference and Exhibition*, pp. 1482–1484 (2016b)

A. Sfetsos, A.H. Coonick, Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Sol. Energy **68**(2), 169–178 (2000)