Solar Energy 173 (2018) 313–327

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Probabilistic forecasting of day-ahead solar irradiance using quantile gradient boosting

Hadrien Verbois[a,b,c,*], Andrivo Rusydi[a,b,c], Alexandre Thiery[d]

[a] Graduate School for Engineering and Integrative Science, National University of Singapore, Singapore
[b] Solar Energy Research Institute, National University of Singapore, Singapore
[c] Department of Physics, National University of Singapore, Singapore
[d] Department of Probability and Applied Statistics, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Probabilistic prediction
Solar irradiance forecasting
Statistical learning
NWP post-processing

ABSTRACT

Due to the chaotic nature of the underlying physical processes, even state-of-the-art models cannot perfectly forecast the solar irradiance at the surface of the earth. There is, therefore, a growing interest in the research community for forecasting methods that can quantify their own uncertainty.

This paper proposes a novel probabilistic framework for forecasting day-ahead hourly solar irradiance. A principal component analysis (PCA) is used to tightly combine a high-resolution mesoscale numerical weather prediction (NWP) model with a quantile gradient boosting algorithm.

A thorough evaluation of the deterministic and probabilistic properties of the model is conducted for a full year in the tropical island of Singapore. The impact of the sky conditions on its performance is also considered. Furthermore, a rigorous statistical framework is employed to systematically benchmark our model against two state of the art methods, a Lasso model output statistic procedure and an analog ensemble (AnEn). Our model significantly improves the numerical weather prediction model: it achieves a 41% reduction of the MAE and 39% reduction of the RMSE. It is also slightly more accurate than Lasso and has a CRPS 4% lower than that of AnEn.

1. Introduction

As the penetration of solar energy increases, the inherent variability of the solar resource challenges the stability of the power grids and reliable forecasts are becoming ever more important. Despite the significant progress recently made in solar power forecasting, the chaotic nature of atmospheric processes prevents all known forecasting models from achieving a perfect prediction. In particular, under tropical and equatorial climates, uncertainty plays a large role in solar forecasting. It is possible, however, to try and quantify this uncertainty, by delivering not only a point-forecast but a full predictive distribution. Decision makers can then adapt their decisions based on that additional information. Alessandrini et al. (2014) estimated the economic benefits of probabilistic predictions for the day-ahead energy bidding of an Italian wind farm and concluded that probabilistic predictions can give a 23% increase in the annual income in comparison with point-forecasts. Ferruzzi et al. (2016) also showed how considering the uncertainty of forecasts could affect the optimal bidding strategy for a microgrid incorporating a large portion of renewable energy sources.

For day-ahead forecasting, purely statistical forecasts are too short-sighted, and Numerical Weather Prediction (NWP) must be used (Lorenz et al., 2009). Multiple NWP instances can be run simultaneously as an ensemble in an attempt to capture the flow-dependant uncertainty of the atmosphere. The resulting probabilistic forecasts are often under-dispersive (Raftery et al., 2005) and must be post-processed; Ensemble Model Output Statistics (EMOS, Sperati et al., 2016), Bayesian Model Averaging (BMA, Aryaputera et al., 2016) and Quantile Random Forest (QRF, Taillardat et al., 2016) were successfully applied to various NWP ensembles to produce calibrated probabilistic forecasts of solar irradiance or power. Because NWP ensembles require significant computational resources, models that produce a probabilistic forecast from a single NWP have also been developed. Lorenz et al. (2009) combined the European Centre for Medium-Range Weather Forecasts (ECMWF) with a polynomial regression; assuming a normal distribution, they modelled the standard deviation of that distribution as a fourth order polynomial of the clear sky index and the zenith angle. Verzijlbergh et al. (2015) developed a similar procedure with the Global Forecasting System (GFS), assuming a normal distribution with a

* Corresponding author at: Solar Energy Research Institute of Singapore (SERIS), National University of Singapore (NUS), 7 Engineering Drive 1, Singapore 117574, Singapore.

E-mail address: hadrien@u.nus.edu (H. Verbois).

https://doi.org/10.1016/j.solener.2018.07.071
Received 25 March 2018; Received in revised form 5 July 2018; Accepted 23 July 2018
0038-092X/

standard deviation depending on the (predicted) clear sky index only. Analog Ensembles (AnEn) take a different, non-parametric approach to probabilistic forecasting. Partly based on nearest neighbour regression, AnEn were first applied to solar energy by Alessandrini et al. (2015), and compared to quantile regression, another non-parametric model. Subsequent studies further improved AnEn by combining them with Principal Component Analysis (PCA, Davò et al., 2016) or Artificial Neural Network (ANN, Cervone et al., 2017).

Although Almeida et al. (2015) and Nagy et al. (2016) produced probabilistic forecasts using a slightly larger set of NWP outputs (13 and 12 NWP variables, respectively) as input to a quantile random forest, existing probabilistic forecasting research usually restricts the input set to a few NWP variables (Lorenz et al., 2009; van der Meer et al., 2016). Typically this subset will include variables directly related to solar irradiance (cloud fraction, global irradiance, solar position) and/or include traditionally measured meteorological variables (e.g. 10 m wind speed, 2 m temperature, humidity). However, NWP can be configured to output a richer and more informative set of variables. The height of the planetary boundary layer, for example, is a good proxy for the atmosphere stability and could thus be a useful input; variables from the microphysics modules, such as the water vapour concentration, could also be helpful. To this point, but for deterministic forecasts, Verzijlbergh et al. (2015) and Verbois et al. (2018) showed that a stepwise variable regression was able to deliver an improved forecast when given a large set of NWP variables as input.

This paper proposes a novel approach to probabilistic post-processing of NWP that utilises state-of-the-art machine learning techniques to fully leverage the multidimensional nature of NWP output. We implement a probabilistic model based on principal component analysis (PCA) and quantile gradient boosting (QGB) and evaluate the benefit of using a large set of NWP output variables as input to our proposed model. While gradient boosting has been used to deliver deterministic forecasts (Gagne et al., 2017; Huang and Perry, 2016), the performance of its probabilistic counterpart–QGB–remains to be investigated. An AnEn (Alessandrini et al., 2015) and a Lasso regression are also implemented as benchmarks.

This paper is organised as follows. The observational and meteorological data are described in Section 2, the validation metrics used are presented in Section 3, the forecasting models are introduced in Section 4, results are presented and analysed in Section 5, and a discussion and conclusion are given in Section 6.

2. Meteorological and observational data

2.1. Irradiance measurements

Singapore is a 700 km^2 city-state located 1° north of the equator, at a longitude of 104°. Under the Koeppen climate classification system, Singapore has a tropical rainforest climate, characterized by abundant rainfall throughout the year. The Solar Energy Research Institute of Singapore (SERIS) operates a network of 25 meteorological stations in Singapore. These stations measure the global horizontal irradiance (GHI) with a resolution of 1 min using IMT-Solar Si-02-PT-100 silicon sensors, calibrated every two years at A∗Star National Metrology Center in Singapore, and with an accuracy of 5%.

Fig. 1 illustrates the characteristics of the average irradiance over Singapore for year 2016. The clear sky index k_t, calculated as follows, is used as a proxy for the sky condition:

$$k_t = \frac{I_t}{I_{t,clear\ sky}}, \tag{1}$$

where I_t is the global horizontal irradiance and $I_{t,clear\ sky}$ is the corresponding estimate for the clear sky irradiance, estimated by a parametric model fitted to Singapore by Yang et al. (2014). The distribution of hourly k_t is shown in Fig. 1a; the intra-hour standard deviation of k_t within each bin–calculated from minute values of k_t–is also given (blue line). The k_t distribution is unimodal, with a mode between 0.8 and 0.85, and, expectedly, the variability is higher for cloudy and partially cloudy skies ($0.3 < kt < 0.8$) than for clear or overcast skies. Fig. 1b illustrates the monthly k_t mean and the monthly k_t variations throughout 2016. We see that although Singapore does not have winters and summers, the monthly sky conditions vary significantly across the year.

The 25, 50 and 75% percentiles of hourly k_t are shown in dashed red lines in Fig. 1a; in this article, they are used as thresholds to classify the data according to sky condition. Hours with $k_t \leq 0.45$ are considered overcast, $0.45 < k_t \leq 0.68$ cloudy, $0.68 < k_t \leq 0.82$ partially cloudy, and $0.82 < k_t$ clear sky. It should be noted that, by construction, these four groups contain the same number of points.

In this work, we are concerned with the average hourly GHI over Singapore, which is obtained by averaging minute values for all 25 stations. Zenith angles above 85° are discarded. Three years of data are used: from January 2014 to December 2016.

2.2. Numerical weather prediction

2.2.1. Model description

The Weather Research and Forecasting model (WRF) (Skamarock et al., 2008) is a mesoscale NWP primarily developed by the National Centre for Atmospheric Research (NCAR). It is open source and a large community contributes to its improvement. In this work, we follow recent work by Verbois et al. (2018) and use a high-resolution implementation of WRF solar (Jimenez et al., 2016), decomposed in four nested domains with horizontal spatial resolutions of 27, 9, 3 and 1 km, and 35 vertical levels. We use the updated Kain-Fritsch cumulus scheme (Kain and Kain, 2004) (in the two largest domains only), the 2nd Mellor-Yamada-Nakanishi-Niino planetary boundary layer (PBL) scheme (Nakanishi and Niino, 2006), the unified Noah land-surface model (Ek et al., 2003) and the RRTMG longwave scheme (Iacono et al., 2008). A monthly climatological map from the Goddard Chemistry Aerosol Radiation and Transport (GOCART) model (Ginoux et al., 2001) provides initial aerosols concentrations for the aerosol aware Thomson microphysics and the RRTMG shortwave schemes (Thompson and Eidhammer, 2014; Ruiz-Arias et al., 2014). The initial and boundary conditions of the WRF model are provided by the 0.5° release of the Global Forecast System (GFS). For more detail about this WRF implementation and its characteristics, we refer readers to Verbois et al. (2018).

The model is run for every day from January 2014 to December 2016. To simulate operational day-ahead forecasting, it is initiated at 8 pm SGT with a horizon of 24 h. The model thus has a 12 h spin-up during the night and the horizon of the forecasts of interest (during the day) ranges from 12 to 22 h.

2.2.2. Model output

In this work, we utilize variables from several WRF parametrization schemes: the two Radiative schemes, the cumulus physics scheme, the microphysics scheme, the surface schemes and the planetary boundary layer (PBL) scheme. Variables from the dynamical core related to pressure, temperature, humidity and wind are also logged, as well as aerosol-related variables. Global, direct and diffuse irradiances are calculated and logged every minute by WRF; they are subsequently averaged to get hourly values. Other WRF output variables are available every hour.

As we are interested in Singapore average irradiance, every WRF output variable is averaged horizontally over the island. Furthermore, we follow Verbois et al. (2018) and apply a Principal Component Analysis (PCA) to each WRF 3D variable to further decrease the dimensionality of the output. The dimension of each WRF 3D variable is thereby reduced from 35 vertical levels to 1–5 components. The average clear sky irradiance over Singapore is also used as regressor.

The list of WRF variables used is given in Table A.1, together with