Renewable Energy 72 (2014) 195–202

Contents lists available at ScienceDirect

Renewable Energy

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

An analytical comparison of four approaches to modelling the daily variability of solar irradiance using meteorological records

Jing Huang*, Alberto Troccoli, Peter Coppin

CSIRO Marine and Atmospheric Research, GPO Box 3023, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 10 March 2014
Accepted 8 July 2014
Available online

Keywords:
Solar variability
Statistical downscaling
Global horizontal irradiance
Random forest
Multiple linear regression

ABSTRACT

Temporal solar variability significantly affects the integration of solar power systems into the grid. It is thus essential to predict temporal solar variability, particularly given the increasing popularity of solar power generation globally. In this paper, the daily variability of solar irradiance at four sites across Australia is quantified using observed time series of global horizontal irradiance for 2003–2012. It is shown that the daily variability strongly depends on sky clearness with generally low values under a clear or overcast condition and high values under an intermittent cloudiness condition. Various statistical techniques are adopted to model the daily variability using meteorological variables selected from the ERA-Interim reanalysis as predictors. The nonlinear regression technique (i.e. random forest) is demonstrated to perform the best while the performance of the simple analog method is only slightly worse. Among the four sites, Alice Springs has the lowest daily variability index on average and Rockhampton has the highest daily variability index on average. The modelling results of the four sites produced by random forest have a correlation coefficient of above 0.7 and a median relative error around 40%. While the approach of statistical downscaling from a large spatial domain has been applied for other problems, it is shown in this study that it generally suffices to use only the predictors at a single near point for the problem of solar variability. The relative importance of the involved meteorological variables and the effects of clearness on the modelling of the daily variability are also explored.

1. Introduction

The installed capacity of solar photovoltaic (PV) electricity has increased from about 100 gigawatts (GW) in 2012 to over 138 GW in 2013 worldwide [13]. In Germany, solar PV penetration (the ratio of the electricity generated by solar PV to the total electricity consumption) is about 5.7 percent in 2013 and can be much higher in a short time period [18]. However, solar irradiance received at or near ground is highly variable in nature, which in turn leads to the variability of the power produced by PV panels. Given the trend of the increasing grid penetration of solar power, this has significant impacts on the operation of power systems across a range of time scales. In principle, the earlier and the more accurately system operators and planners know the extent of variability of power production, the more options they will have to accommodate it, and consequently the cheaper it will be to manage the system [14]. While day ahead forecasts for weather information is normally provided by numerical weather prediction (NWP) tools, empirical models are needed to link weather information and the extent of solar variability.

The most common parameter of solar irradiance recorded in local meteorological stations is the global horizontal irradiance (GHI), which is defined as the total (direct plus diffused) solar irradiance projected on a horizontal surface. Using GHI time series, Stein et al. [17] proposed a simple and robust metric (the daily variability index, or DVI) to quantify the mean daily variability of solar irradiance. They also showed that DVI correlates strongly with the ramp rate of GHI, which is essentially important in managing the integration of solar power systems into the grid (see e.g. Refs. [11,14]). In this paper, the 1-min GHI time series measured by the Bureau of Meteorology at four sites across Australia are used and the corresponding DVI series are calculated. Then, statistical models are built to link DVI to atmospheric variables in the ERA-Interim reanalysis so as to understand what the main drivers for DVI are.

Statistical downscaling methods are often used for simulating or predicting local weather conditions from coarse-resolution and large-scale atmospheric model output (see e.g. Refs. [8,9]). The target of these methods is to build an empirical model which links the large-scale atmospheric fields (i.e. predictors) with the local-

* Corresponding author.
E-mail address: jing.duke@gmail.com (J. Huang).

http://dx.doi.org/10.1016/j.renene.2014.07.015

scale quantities of interest (i.e. predictands). In our case, we compare two approaches whereby the predictors are chosen from a single point in the local area in the first instance and from a large spatial domain in the second. In the first case, the statistical model relates a set of predictors and the targeted predictand at the same location. As such, the term "statistical downscaling" is used in a general sense. The methodology has been recently adopted for the problem of wind variability [5,7]. Davy et al. [5] have successfully downscaled reanalysis fields from the US National Centers for Environmental Prediction to model wind variability at Waratah Bay, Victoria, Australia after dimension reduction using empirical orthogonal function (EOF) techniques. The resulting model of wind variability is found to outperform a simple regression method against wind speed as well as models using multiple linear regression (MLR). Following [5], Ellis et al. [7] extended the use of statistical downscaling techniques to the problem of the variability of wind power generation. In the present study, four techniques have been used to build empirical models between DVI and meteorological records: the analog method (see e.g. Ref. [19], the MLR method, a nonlinear regression method called random forest (RF) and persistence, which is used as our benchmark.

In Section 2, the GHI measurement at four sites and the ERA-Interim reanalysis are introduced. Also, the background of the downscaling techniques is given and the metrics of performance evaluation used in this paper are defined. In Section 3, the DVI and daily clearness index (DCI) of solar irradiance are defined and their statistical properties are discussed. The meteorological fields from the ERA-Interim reanalysis are selected and used as predictors to construct downscaling models using four downscaling techniques in Section 4. Concluding remarks are given in Section 5.

2. Data and methodology

2.1. Data preparation

The Bureau of Meteorology (BOM) has been undertaking high frequency observations of various solar parameters across Australia. The observation data is freely available through its website (http://reg.bom.gov.au/climate/reg/oneminsolar/index.shtml). This study uses the GHI 1-min time series from 2003 to 2012 at four observation sites in Australia, which are Alice Springs, Darwin, Rockhampton and Wagga Wagga, respectively. The GHI is measured at 1 Hz using CM-11 pyranometers manufactured by Kipp and Zonen, and then is averaged over the preceding 1 min to produce the 1-min time series. The locations, elevations and the rates of missing data of the four observation sites are given in Table 1. The four sites are selected to represent a desert climate (Alice Springs), a tropical climate (Darwin), a humid subtropical climate (Rockhampton) and a temperate climate (Wagga Wagga), respectively.

The ERA-Interim reanalysis is a global atmospheric reanalysis product of the European Centre for Medium-Range Weather Forecasts (ECMWF) covering the period from 1979 to near present. Its resolution is 6-hourly in time, viz. 4AM, 10AM, 4PM and 10PM in Australian Eastern Standard Time (AEST) and $0.75^\circ \times 0.75^\circ$ in space. Although the ERA-Interim reanalysis is one of the best reanalysis products available in the world, it should be realised that it is not completely equivalent to observation whereby errors may arise from the process of data assimilation or running numerical models when global observations are unavailable for some variables.

Table 1
Detailed information about the four observation sites used in the paper. RMD stands for the rate of missing data.

Location	Latitude	Longitude	Elevation [m]	RMD [%]
Alice Springs	−23.7951	133.8890	546	1.7
Darwin	−12.4239	130.8925	30.4	4.8
Rockhampton	−23.3753	150.4775	10.4	2.9
Wagga Wagga	−35.1583	147.4573	212	0.76

2.2. Statistical models

Random forest (RF) is a nonlinear ensemble learning method recently developed by Refs. [2,3], which constructs multiple-predictor models. RF can be used for both regression and classification problems. For regression its algorithm contains three major steps as follows [12]. First, n bootstrap samples are drawn with each sample including approximately 64% of the original training data. Then, an unpruned regression tree is grown for each of the bootstrap samples. However, rather than using the best split among all p predictors, only m of the p predictors are randomly sampled and the best split is chosen from among these m variables. Finally, the prediction is formed by averaging the output of the n trees. In the present study, the default values of $n = 500$ and $m = p/3$ are used while a change of m and n within a reasonable range does not affect the resulting performance of RF significantly in this study. In addition to constructing multiple-predictor models, RF also produces scores measuring the relative importance of each predictor on the predictand. This score is estimated by calculating the mean decrease in accuracy due to permuting the associated predictor while leaving the others unchanged [12].

RF has been used to study complex problems where nonlinearity often plays an important role (e.g. Refs. [1,5]). One prominent feature of random forest is its robustness against overfitting [2,16]. Eccel et al. [6] downscaled the output of two NWP models to predict the minimum spring temperature in an alpine region using RF and artificial neural networks (ANN), and found that RF performs similarly as ANN, if not superior. The *random Forest* library [12] in the statistical software *R* [15] is used to perform the RF analysis in this study.

In addition to RF, three other techniques are also used including MLR, persistence (i.e. use the DVI of yesterday to predict the DVI of today) and the analog method. As its name suggests, the analog method identifies similar patterns of the predictors at large scale in their historical records, and then the simultaneous local predictand is associated with the large scale pattern. In practice, the similarity is normally quantified by the Euclidean distance in the multi-dimensional hyperspace of the predictors and the prediction is calculated as the mean of the predictand in the analogs identified [9]. Zorita and Storch [19] used the analog method as a statistical downscaling technique and compared it with other more complicated methods, such as canonical correlation analysis and neural networks and found that the analog method performs in general as well as the more complicated methods.

2.3. Metrics of performance evaluation

Three metrics are used in the following results to quantitatively evaluate the performance of a model: Pearson's correlation coefficient r, the mean absolute error (MAE) and the median relative error (MeRE). While one metric may not be able to evaluate and compare the performance of multiple models accurately, it is expected that the models can be evaluated comprehensively by studying the three metrics together. Assuming the observation values are x_i and the corresponding modelled values are y_i, the three metrics are defined as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{1}$$