Renewable Energy 80 (2015) 770–782

Contents lists available at ScienceDirect

Renewable Energy

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

Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances

Hugo T.C. Pedro, Carlos F.M. Coimbra*

Department of Mechanical and Aerospace Engineering, Jacobs School of Engineering, Center of Excellence in Renewable Energy Integration, Center for Energy Research, University of California San Diego, La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 13 January 2015
Accepted 25 February 2015
Available online 23 March 2015

Keywords:
Smart solar forecasting
Global and direct irradiance
Machine learning
k-nearest-neighbors

ABSTRACT

This work proposes a novel forecast methodology for intra-hour solar irradiance based on optimized pattern recognition from local telemetry and sky imaging. The model, based on the *k*-nearest-neighbors (*k*NN) algorithm, predicts the global (GHI) and direct (DNI) components of irradiance for horizons ranging from 5 min up to 30 min, and the corresponding uncertainty prediction intervals. An optimization algorithm determines the best set of patterns and other free parameters in the model, such as the number of nearest neighbors. Results show that the model achieves significant forecast improvements (between 10% and 25%) over a reference persistence forecast. The results show that large ramps in the irradiance time series are not very well capture by the point forecasts, mostly because those events are underrepresented in the historical dataset. The inclusion of sky images in the pattern recognition results in a small improvement (below 5%) relative to the *k*NN without images, but it helps in the definition of the uncertainty intervals (specially in the case of DNI). The prediction intervals determined with this method show good performance, with high probability coverage ($\approx$90% for GHI and $\approx$85% for DNI) and narrow average normalized width ($\approx$8% for GHI and $\approx$17% for DNI).

1. Introduction

Forecasting of intra-hour solar irradiance and solar power output is important for power grid regulators which seek to reduce power system operation cost by committing appropriate amount of energy resources and reserves. Intra-hour forecasts are also relevant for optimal central plant operations, and is an enabling technology for optimal dispatch of ancillary resources and storage systems [15]. Increased accuracy in irradiance predictions can enable substantial improvement on the automatic generation control tools used to balance generation and load. Moreover, mitigation measures for large drops in solar irradiance, such as demand response, storage and intra-hour scheduling can only be maximized with accurate and reliable intra-our forecast [15].

Short-term fluctuations in solar irradiance are dictated, almost exclusively, by cloud cover. A single passing cloud can bring the power output of a solar farm from full production to minimum and back to full in a mater of minutes or even seconds [27]. Thus, it is not surprising that many recent short-term forecast models incorporate cloud cover information. In general, the methodology for processing ground-based images and including them into the forecast models is based on cloud motion detection and the propagation of the cloud field into the future [8–11,13,21–23,29]. Another approach to model and forecast cloud cover consist of studying the sunshine number, which is a binary number equal to the unity if sun is shining and 0 if it is not. Parametric and semi-parametric approaches based on Markovian dynamics of the sunshine number have shown good results for nowcasting [4,5,26] and are good candidates to improve the intra-hour forecast of solar irradiance, specially when large irradiance fluctuations occur.

Contrary the previous works, no cloud motion detection nor distribution estimation is performed in this work. Here we propose a distribution-free methodology based on the *k*-Nearest-Neighbors (*k*NN) algorithm to predict intra-hour global (GHI) and direct (DNI) irradiances using ground telemetry and sky images. The data used in this work consist of pyranometer measurements of GHI and DNI and sky images captured with a off-the-shelf security camera pointing to the zenith. The main novelty of this study is the development of (i) specific techniques used to extract information from the irradiance data and (ii) of image processing techniques used to quantify sky

* Corresponding author.
E-mail addresses: htcpedro@ucsd.edu (H.T.C. Pedro), ccoimbra@ucsd.edu (C.F.M. Coimbra).

http://dx.doi.org/10.1016/j.renene.2015.02.061
0960-1481/

Nomenclature

B_i	backward cumulative average of irradiance for window $[t - i\delta, t]$
DNI	direct normal irradiance
d, **D**	distance vectors for the kNN forecast
GHI	global horizontal irradiance
I	measured irradiance, GHI or DNI, in Wm^{-2}
$\hat{I}^m$	forecasted irradiance, GHI or DNI, using model m, in Wm^{-2}
L_i	lagged 5 min average of irradiance for window $[t - i\delta, t - (i-1)\delta]$
k	number of nearest neighbors
k_t	clear-sky index, GHI or DNI
kNN	k nearest neighbors
PI	prediction interval
PICP	prediction interval coverage probability
PINAW	prediction interval normalized average width
r, **g**, **b**	red, green and blue channels for the sky images
RMSE	root mean squared error
s	forecast skill
S	set of features used in a given kNN forecast
V_i	irradiance variability for window $[t - i\delta, t]$
t	time, in min
α	weights for the kNN forecast
δ	window size or window increments, in min
Δt	forecast horizon, in min
Δk_t	clear-sky step change
ω	weights for the kNN distances

Superscripts

clr	clear-sky model
g, *d*	GHI, DNI

conditions from the sky images and (iii) of integrating the information into the forecast models. The images are not used to monitor could dynamics, instead, they are used to provide features, such as the entropy of the Red, Green and Blue channels, to determine the nearest neighbors from the historical dataset.

The kNN method is one of the simplest methods among the machine learning algorithms. In contrast to the statistical methods that attempt to find models from the available data, the kNN uses the training dataset as the model. Despite the fact that the kNN model was originally developed for pattern classification it can easily be applied to regression problems for time series [34], such as the forecast of solar radiation. Although there are very few articles in literature that apply kNN to the forecast of solar irradiation (see for instance [25,27]), kNN has been extensively applied as a forecasting technique to problems such as: electricity load and electricity price [18,19], daily river water temperature [31], water inflow [1] and weather forecast [2]. In the field of meteorology and weather forecasting the kNN method is also known as the analog method [30,33,35].

For the purpose of forecasting time series the kNN model consists of looking into the time series history and identifying the time stamps in the past that resemble the current conditions most closely - the nearest neighbors. Once these are determined, the forecasting is computed from the time series values subsequent to the matches. In essence, the kNN model resembles a lookup table for which previous features are used as indicators of sequential behavior.

Another goal for this work is to propose a method to estimate the PIs for the forecast. A PI is an estimate of an interval in which a future measured value falls, given a lower and an upper bound [7]. PIs provide information about the forecast uncertainty and are very important for operational planning. Usually a probability is associated with the PI to quantify the confidence level of the interval. In the field of wind power forecasting, there are many significant works dedicated to calculating PIs [3,17,28,32]. The same does not happen in the field of solar power forecast where most literature focuses on point forecast, while the PIs are rarely provided. A few exceptions are the works by Refs. [6,12,16,20,24]. In this work we propose a simple algorithm to calculate the PI bounds using the nearest neighbors. The quality of the PIs is measured in terms of their probability coverage and average normalized width.

2. Data

The solar irradiance measurements (GHI and DNI) were obtained at Folsom, CA, 38.64° N and 121.14° W, at a sampling of 1 measurement per minute. The dataset contains one year of data (from December 2012 to December 2013). The irradiance data was averaged into 5-min windows and divided into three disjointed datasets. The first dataset, denoted as training or historical dataset, is used to create the kNN database of features from where the nearest neighbors are selected. The second dataset, denoted as optimization dataset, is used in the optimization algorithm to determine the several free parameters (explained below) in the forecasting model. The third dataset, the independent testing set, is used to assess the performance of forecasting model. The three datasets were constructed by grouping disjoint subsets for each month in the whole dataset, so that the training, optimization and testing sets all contain similar data.

Fig. 1 shows the all the data for the three datasets. In this figure the scatter plots show k_t for GHI versus the normalized time ω ($\omega = -1$ sunrise and $\omega = 1$ sunset). The whiskers allow to visualize the data variability as a function of the normalized time. The histograms show the relative frequency of k_t for the entire datasets, in logarithmic scale. These plots show that the three datasets have similar data distributions. The size of each dataset is also shown in the figure.

Sky images were captured at the rate of one image per minute with a security camera (Vivotek, model FE8171V) pointing to the zenith in the same period. The advantages of this camera include high-resolution, easy installation, low cost, and absence of moving parts. On the other hand, given that direct sunlight is not blocked, the circumsolar region is affected by glare caused by forward Mie scattering and also from light scattered from the dome. The camera provides 24-bit compressed jpgs, with 8 bits per color channel (Red, Green and Blue), at 1563 by 1538 pixels. After removing the pixels that do not correspond to the sky dome and the saturated pixels (yellow and orange shaded areas in Fig. 1d, respectively) only ≈50% of the pixels are usable. The images were synchronized with the irradiance data and also divided in three datasets. Finally, in this work we consider only time stamps for which the sun elevation angle is above 5°.

3. Methodology

3.1. Persistence

The persistence model is used in this work as a baseline model. In this model we assume that the clear-sky index for the forecasted variable (GHI or DNI) persists in to the future. The clear-sky index is defined as