Renewable Energy 78 (2015) 203–218

Contents lists available at ScienceDirect

Renewable Energy

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

On the role of lagged exogenous variables and spatio–temporal correlations in improving the accuracy of solar forecasting methods

Athanassios Zagouras, 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 and Center for Energy Research, University of California San Diego, La Jolla, CA 92093, USA

A R T I C L E I N F O

Article history:
Received 30 September 2014
Accepted 30 December 2014
Available online

Keywords:
Solar forecasting
Time-lagged correlation
Gridded data
Exogenous input data
Genetic algorithms

A B S T R A C T

We propose and analyze a spatio–temporal correlation method to improve forecast performance of solar irradiance using gridded satellite-derived global horizontal irradiance (GHI) data. Forecast models are developed for seven locations in California to predict 1-h averaged GHI 1, 2 and 3 h ahead of time. The seven locations were chosen to represent a diverse set of maritime, mediterranean, arid and semi-arid micro-climates. Ground stations from the California Irrigation Management Information System were used to obtain solar irradiance time-series from the points of interest. In this method, firstly, we define areas with the highest correlated time-series between the satellite-derived data and the ground data. Secondly, we select satellite-derived data from these regions as exogenous variables to several forecast models (linear models, Artificial Neural Networks, Support Vector Regression) to predict GHI at the seven locations. The results show that using linear forecasting models and a genetic algorithm to optimize the selection of multiple time-lagged exogenous variables results in significant forecasting improvements over other benchmark models.

1. Introduction

Large-scale utilization of solar energy for power generation requires advanced forecasting technologies to handle the variability of this weather-dependent resource. Several research groups have developed forecast algorithms and methodologies to predict ground irradiance at different forecasting horizons, ranging from minutes out to days ahead into the future [1].

The numerical tools that ingest relevant ground telemetry, remote sensing data and numerical weather prediction results typically include simple deterministic models with complex preprocessing steps, stochastic time series models such as AR, ARMA and ARIMA, machine learning tools such as Artificial Neural Networks, K-Nearest-Neighbors, Supported Vector Regression, and many other post-processing algorithms [1–4]. Most existing operational models use only local data and are limited to observe variations in the input variables in a single point in space, and are “blind” to the motion of weather systems across the earth's surface. A few recent studies employed spatially distributed data to cover small regions in space. Lonij et al. [5] use data from a network of 80 rooftop sensors in Tucson, AZ, Yang et al. [6] use a network of 10 sensors in Singapore and Sözen et al. [7] train a neural network for 12 ground stations spread over Turkey.

In both cases the researchers reported substantially improvements to the forecast skill. The drawback for such approach is that, a network of ground sensors is costly to deploy and maintain due to required regular maintenance. When no such ground truth data is available, is it possible to find some useful approximations to irradiance measurements as Pelland et al. [8] have done by using spatially distributed irradiance forecast from the Canadian Meteorological Centre. Al-Alawi and Al-Hinai in Ref. [9] developed a neural network based forecasting model to predict the global radiation for remote locations in Oman, where ground-based measurements are not available.

In a similar way to these authors, we explore in this work simple forecast models that include irradiance data from locations in the vicinity of the point of interest. Here gridded data comes, not from a numerical weather prediction, but from models that ingest satellite images to estimate the solar irradiance at the ground level. Such models have grown in accuracy in recent years (see for instance [10,11]). Moreover, such data will even be more accurate in the near future with GOES-R, the next generation of geosynchronous environmental satellites by NOAA, that will increase the frequency of full-disk imagery from 30 min to 5 min [12].

* Corresponding author.
E-mail address: ccoimbra@ucsd.edu (C.F.M. Coimbra).

http://dx.doi.org/10.1016/j.renene.2014.12.071
0960-1481/

The objective of this paper is to introduce a forecasting approach that improves the prediction accuracy of ground measured irradiance data throughout the state of California through the use of exogenous satellite-derived variables. Two main aspects are explored by the proposed methodology. First is to investigate how exogenous input data obtained by difference source than a ground-based sensor at the point of interest affect the forecast of the measured observations. It is important to identify the seeking area around the vicinity of a ground location where a model can search for exogenous input data. Secondly, the time latency of both the endogenous and exogenous inputs plays a crucial role in the forecasting performance, as well as the optimization of selecting a number of exogenous variables from different time lags. Given a large dataset of gridded satellite data, we propose a time-lagged correlation analysis between the ground measured times series and the gridded data to define potential seeking regions based on the degree of correlation. We study different types of forecasting models including the persistent model, Artificial Neural Networks (ANN), Support Vector Regression (SVR), linear models using a variety of time-lagged input data and a linear model optimized by Genetic Algorithms (GA). The latter provided the best forecast accuracy for forecast horizons up to 3 h ahead, with respect to any ground location and evaluated on different periodic partitions of an independent testing subset.

The remainder of the paper is organized in four parts: in Section §2, the case study will be presented. Section §3 is devoted to a description of the time-lagged correlation analysis and the developed models. Section §4 analyses the experimental results. The conclusion is reported in Section §5.

2. Case study

2.1. CIMIS ground data

Ground-based hourly irradiance data used in this study were obtained from 7 weather stations operated by California Irrigation Management Information System (CIMIS). A network of 120 automated weather stations throughout the entire state of California constitute the CIMIS program managed by the Office of Water Use Efficiency (OWUE), California Department of Water Resources (DWR). Aiming to assist in the efficient management of irrigation water resources, DWR and the University of California, Davis established the CIMIS program in 1982. The CIMIS weather stations are located at key agricultural and municipal sites to either measure various meteorological data on an hourly basis or calculate them from measure values. The weather stations are equipped to measure data such as solar radiation, air and soil temperature, relative humidity, precipitation, and wind speed and wind direction. In the present study, we selected seven weather stations located at distinct sites in California, from the north to the south as well as at coastal and inland sites (Fig. 1). The wide range of stations was based on the estimated solar variability that each location experiences and their geographic proximity to regions where antipodal climatic conditions occur. For instance, San Diego is exposed to the marine layer of the Pacific ocean which substantially affects solar variability, whereas Merced, located in the Central Valley of California and has hot, dry, cloudless summers and notably low solar variability.

Table 2 summarizes geographical information for the sites used. Solar radiation data were obtained for a 45-month period from January 2009 to September 2012. The total solar radiation that reaches each weather station is measured by a LI200S Li-Cor photovoltaic pyranometer. The 60 readings per minute are then averaged to hourly data and published on the CMIS portal. The error accuracy of the Li-Cor pyranometer is within ±5% under natural sunlight conditions.

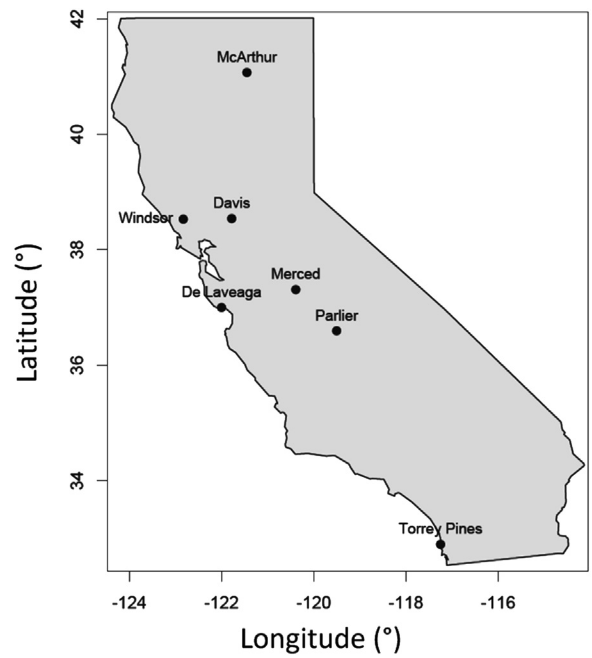

Fig. 1. Map of California in gray with selected CIMIS weather stations marked in black and labeled.

The ground data were split into training, validation and testing sets, respectively. The training set was used for the time-lagged correlation calculations and the estimation of the model parameters, the validation set was used for the parameter optimization and the testing was applied for evaluating the generalization capacity over independent data. The three data sets were created with the following algorithm: the 1st week of every month was assigned to the validation set, the 2nd to testing and the remaining data were used for training. This resulted in a typical data partition with 50% for training and 25% for validation and testing, respectively. This splitting algorithm ensures that all the subsets share similar seasonal trends.

2.2. Satellite-derived data

In addition to the CIMIS data, we use gridded data values of GHI for the state of California provided by the *SolarAnywhere®* [13]. The Enhanced Resolution dataset, for the same period as the CIMIS dataset, consists of satellite-derived GHI data on an approximately 0.01° resolution grid both in latitude and longitude and a 30 min time sampling rate. *SolarAnywhere®* global and direct irradiance values are derived from the SUNY semi-empirical satellite-to-irradiance model, which relies on imagery collected from the GOES-W and GOES-E satellites in conjunction with other meteorological data sources [14,15]. A number of studies validates the accuracy of SUNY model at several individual locations with a typical error of the derived global and direct irradiance estimates correspond to 5%

Table 1
List of possible exogenous variables to include into the forecast models.

$k_1(t)$	$k_2(t)$	$\cdots$	$k_{N_{exo}}(t)$
$k_1(t-\Delta t)$	$k_2(t-\Delta t)$	$\cdots$	$k_{N_{exo}}(t-\Delta t)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$k_1(t-(N_{lag}-1)\Delta t)$	$k_2(t-(N_{lag}-1)\Delta t)$	$\cdots$	$k_{N_{exo}}(t-(N_{lag}-1)\Delta t)$