Solar Energy 150 (2017) 383–393

Contents lists available at ScienceDirect

Solar Energy

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

Evaluation of statistical learning configurations for gridded solar irradiance forecasting

David John Gagne II [a,b,*], Amy McGovern [c], Sue Ellen Haupt [a], John K. Williams [d]

[a] National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307, United States[1]
[b] School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd., Suite 5900, Norman, OK 73072, United States
[c] School of Computer Science, University of Oklahoma, 110 W. Boyd St., Devon Energy Hall, Rm. 150, Norman, OK 73019, United States
[d] The Weather Company, an IBM Business, 400 Minuteman Rd, Andover, MA 01810, United States

ARTICLE INFO

Article history:
Received 21 November 2016
Received in revised form 28 March 2017
Accepted 9 April 2017

Keywords:
Statistical learning
Solar irradiance
Forecasting

ABSTRACT

Gridded forecasts of solar irradiance are increasingly needed to integrate power into the electric grid from distributed solar installations and newer large-scale installations that don't have long records of observed irradiance. We evaluate different combinations of statistical learning models and aggregations of weather data from observed sites to identify which combination produces the lowest forecast errors at independent sites. The evaluation reveals how statistical learning model choice, closeness of fit to training data, training data aggregation, and interpolation method affect forecasts of clearness index at Oklahoma Mesonet sites not included in the training data. It shows that the choices of statistical learning model, interpolation scheme, and loss function have the biggest impacts on performance. Errors tend to be lower at testing sites with sunnier weather and those that are closer to training sites. All of the statistical learning methods and the NWP model output produce reliable predictions but underestimate the frequency of cloudiness compared to observations.

1. Introduction

Solar-based electricity generation and its share of the power supply have been growing rapidly over the past decade (Shaker et al., 2016). As solar power achieves higher penetration and becomes more critical to the electric infrastructure, accurate forecasts of solar irradiance and solar power are needed in order to maintain a balanced electric load under varying weather conditions (Renné, 2014). Current state-of-the-art solar and wind energy forecast systems combine Numerical Weather Prediction (NWP) model output with statistical learning models trained on a historical archive to produce solar irradiance or power output forecasts with minimal bias (Orwig et al., 2015). This approach is very effective for sites that have been operating for a long period of time, but with new utility-scale solar plants coming online more frequently and distributed solar installations increasingly impacting the measured load, accurate solar predictions are needed for larger areas where observing sites either have very short records or are not available at all (Tuohy et al., 2015).

Generating accurate irradiance predictions at sites without observations can be accomplished by fusing static and dynamic data sources together within a statistical learning framework. The amount of solar irradiance at the surface is primarily driven by the position of the sun in the sky as well as the amount and type of aerosols and clouds scattering the sunlight. Obstructions by terrain, buildings, and trees can also impact solar irradiance at lower sun angles. Solar position can be directly calculated given a location and time, and information about terrain and land cover type is available from high resolution gridded datasets. Cloud cover and aerosol information can be extracted from NWP model output, but operational NWP models generally do not represent either very well and may be subject to other systematic biases (Diagne et al., 2013). Statistical learning models can determine the effects of cloud cover from other NWP model conditions associated with observed cloudiness. They can also incorporate information from data sources unavailable to an NWP model, including climatological information and statistics concerning spatial and temporal variability (McCandless et al., 2015, 2016a,b).

Current operational statistical gridded forecasting systems use linear bias correction methods to calibrate raw NWP model output to either observations or analyses, which are a gridded fusion of

[1] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

* Corresponding author at: National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307, United States.

E-mail addresses: dgagne@ucar.edu (D.J. Gagne II), amcgovern@ou.edu (A. McGovern), haupt@ucar.edu (S.E. Haupt), John.Williams@weather.com (J.K. Williams).

http://dx.doi.org/10.1016/j.solener.2017.04.031

observations and background NWP model output, and then interpolate those corrections to a fine grid. The National Weather Service Gridded Model Output Statistics (MOS) system (Glahn et al., 2009) performs linear regression corrections at each observation site and then uses the Cressman (1959) successive correction method and an elevation correction to interpolate the site-based MOS forecasts to a grid. The Australian Bureau of Meteorology, which has to account for a sparse observation network across most of the country, performs a bias correction of model output on a coarse grid and then builds a weighted consensus that is statistically downscaled to a fine grid (Engel and Ebert, 2012).

The purpose of this paper is to evaluate different statistical learning models and configurations for gridded solar irradiance forecasting. The primary hypothesis is that ensemble decision tree methods produce more accurate gridded solar irradiance forecasts than linear regression and raw NWP model output. In the preprocessing stage, the set of input variables, NWP model configuration, and division of training data are investigated. Multiple types of statistical learning models, as well as different configurations of those models, are evaluated to determine which parameter choices impact performance. Finally, different methods for applying the calibrated statistical learning models to unknown sites are evaluated based on their forecast errors and the realism of their forecast distributions. This article is adapted from Chapter 5 in Gagne (2016).

2. Methods

2.1. Observations

Observed solar irradiance data come from the Oklahoma Mesonet (McPherson et al., 2007). The Mesonet reports the 5-min-averaged global horizontal irradiance (GHI) every 5 min using Li-Cor LI-200 silicon photodiode-type pyranometers. The instruments are regularly calibrated and are monitored by both humans and automated algorithms for quality assurance. Extraterrestrial solar radiation and solar position angles are calculated using a Python implementation of the National Renewable Energy Laboratory (NREL) Solar Position Algorithm (SPA) (Reda and Andreas, 2003) within the PVLIB Python library (Holmgren et al., 2015). Solar zenith (θ_s), elevation, and azimuth angles are calculated every 5 min and are used to estimate the idealized clear-sky irradiance at the top of the atmosphere I_{toa}. The clearness index K_t is calculated from the Mesonet solar irradiance I_s as

$$K_t = \frac{I_s}{I_{toa}\cos\theta_s}. \tag{1}$$

The 5-min irradiance and clearness index values are then averaged over the previous hour to determine the hourly-averaged values. The hourly-averaged K_t is then used as truth for the statistical learning model experiments. Times without sun or with data outages are removed from the dataset.

2.2. GRAFS

The gridded statistical learning forecasts are generated using the research version of the Gridded Atmospheric Forecast System (GRAFS; Gagne et al., 2015), a community platform for testing different statistical learning systems developed at the National Center for Atmospheric Research. The statistical learning model experiments are performed with the NOAA National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) model. The GFS is a global spectral model run operationally by NCEP four times a day out to 16 days. The raw GFS model output with approximately 28 km grid spacing is interpolated onto an approximately 4 km grid that uses uniform latitude and longitude values over the contiguous United States. The 3-hourly output of the GFS is interpolated to hourly values by interpolating the clearness index value and then converting back to solar irradiance to account for the changes in solar position. Incoming hourly averaged clearness index and total cloud cover percentage are extracted from 349 GFS model runs initialized at 0000 UTC for the period from 5 June 2015 through 31 May 2016. Forecast hours 14–24 are used for the analysis. All of the input variables to the statistical learning models are listed in Table 1.

2.3. Gridded forecast evaluation procedure

Two procedures are evaluated for training statistical learning models at locations with irradiance observations and applying them to unobserved locations. For "Single Site" models, separate statistical learning models are trained using data from each training site. Then, predictions are made at each of these sites, and finally the predictions are interpolated to the testing sites using the Cressman (1959) successive correction interpolation method. For each interpolation point f_i, a distance-weighted average of the predictions at the stations with distances d_j within a radius of influence R is computed such that

$$f_i = \frac{\sum_j^J w_j f_{sj}}{\sum_{j=1}^J w_j}; \quad \begin{array}{ll} w_j = \frac{R^2 - d_j^2}{R^2 + d_j^2} & d_j < R \\ w_j = 0 & d_j \geqslant R. \end{array} \tag{2}$$

The test sites were initialized with the mean of the predictions at the training sites, and then four passes were performed with the Cressman filter with a decrease in radius for each pass to capture local effects. The 90th, 75th, 50th, and 25th percentiles of distances among Mesonet sites were used, which corresponds to 4.34, 3.28, 2.23, and 1.42 degrees latitude and longitude, respectively. The Cressman interpolation method was chosen because the NWS gridded MOS system (Glahn et al., 2009) also uses it for interpolation from training sites to grid points.

In the "Multi Site" approach, the data from all training sites are aggregated and are used to train a single statistical learning model. This model is then applied at all testing sites using the NWP model and clear sky model output at that location. This approach requires training a single statistical model and can thus utilize a larger training set than the Single Site method. Applying the same model to all grid points also eliminates discontinuities that may be found in approaches that use separate statistical learning models for different regions. However, this approach is less able to correct for local biases and conditions.

Generating calibrated gridded solar irradiance forecasts requires determining the best estimate of irradiance at unobserved locations. In order to simulate this condition and still score the different procedures, the set of 120 Oklahoma Mesonet sites are randomly split into training set sites and testing set sites. In addition, testing days are withheld from the training data to prevent temporal contamination. Every third day during the training period is used as a testing day so that both training and testing sets are sampled from the same seasons.

Forecasts from each statistical learning model are evaluated based on their accuracy, systematic bias, and sharpness (Murphy, 1993). Forecast accuracy is assessed using the mean absolute error because it is less sensitive to outlier errors than the root mean squared error. Systematic bias is evaluated through the mean error and determines if the models tend to over or underforecast clearness index. Sharpness, or the range of forecast values compared to the range of observations, is assessed by examining the distribution of forecasts and comparing them with the distribution of observations.