Towards An Efficient Regression Model for Solar Energy Prediction

Anuj Prakash
Department Of Computer Science & IT,
Jaypee Institute of Information Technology, Noida, India
anuj1992@gmail.com

Sandeep Kumar Singh
Department Of Computer Science & IT,
Jaypee Institute of Information Technology, Noida, India
sandeepk.singh@jiit.ac.in

***Abstract*— This paper describes a model for forecasting the daily solar energy. The features used in this model include precipitation, flux (long-wave, short wave), air pressure, humidity, cloud cover, temperature, radiation (long-wave and short-wave). These features along with previous data for daily solar energy received for the years 1994-2007 has been used for forecasting. The data for the features comes from a grid of sites in the United States and the data for previous years' daily solar energy comes from 98 sites in Oklahoma, United States. Two algorithms have been used for forecasting—Linear Least Square Regression and Gradient Boosting Regression. Gradient Boosting Regression has shown to be around 2.5% more accurate as compared to Linear Least Square Regression.**

Keywords—Solar Energy, Prediction, Regression model, Computational Intelligence

I. INTRODUCTION

In today's world renewable sources of energy has become a top priority for the governments due to the increasing global concerns about climate change and scarcity of fossil fuels. Today, solar power has become part of our daily lives. Appliances like solar notebooks, solar air-conditioners, solar cars, etc. demonstrate the use of the sustainable power of the sun. As the adverse effects of burning of fossil fuels and the depletion rate of non-renewable energy sources increase, the future of solar energy looks bright. The problem with renewable sources of energy is that they are not easily predictable in advance and vary based on both weather as well as site specific conditions. In the present study, a model for predicting daily solar energy at a site in Oklahoma, United States is described. There are 15 features on which the daily solar energy at that site depends. These features are recorded by real sensors in a 16 X 9 grid spread across United States covering the state of Oklahoma also. On this data, two machine learning algorithms are applied—Linear Least Square Regression and Gradient Boosting Regression. Linear Least Square Regression is applied to compare the accuracy of Gradient Boosting Regression in the domain of solar energy prediction.

II. DATA SET

Daily solar energy data were provided by the Oklahoma Mesonet with the assistance of Dr. Jeffrey Basara. The GEFS Reforecast Version 2 data were developed and provided by Dr. Thomas Hamill.This data was taken from[5].

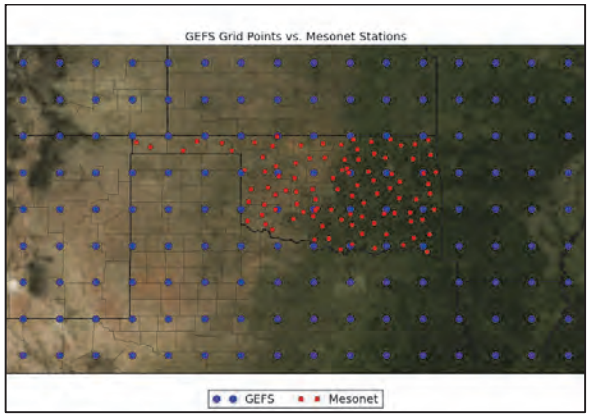

Fig. 1: A map depicting the location of GEFS and Mesonet Sites.

Input features like precipitation, flux (long-wave, short wave), air pressure, humidity, cloud cover, temperature, radiation (long-wave and short-wave), etc. have been recorded at GEFS Sites daily during 1994-2007(14 years). The data for the daily solar energy for the years 1994-2007 (14 years) have been recorded at the Mesonet sites. All the data has been recorded by real sensors at the respective sites. The data for the GEFS features are in netCDF4 files with each file holding the grids for each ensemble member at every time step for a particular variable. Each netCDF file contains the latitude-longitude grid and timesteps values as well as metadata listing the full names of each variable and the associated units. Details about each of the 15 variables are given in Appendix. Also, the data for the total daily incoming solar energy in (J m^{-2}) at 98 Oklahoma Mesonet sites has been recorded by sensors that have been in continuous operation since January 1, 1994. The solar energy was directly measured by a pyranometer at each Mesonet site every 5 minutes and summed from sunrise to 23:55 UTC. A separate file contains the latitude, longitudes, and elevation (meters) of each Mesonet station.

III. DATA PREPROCESSING AND ANALYSIS

The first step in this work was to extract the data for a variable from its respective netCDF4 file. After, extracting the data for each of the points on the 16 X 9 grid, the feature data is mapped to the GEFS Sites by linear interpolation. This was done so that the features on which the daily solar energy depends on and the amount of daily solar

energy itself are known for the years 1994-2007 for the same sites.

Note: Most of the graphs and calculations are worked out on the basis of data values at the 'ACME' station (Latitude: 34.80833N, Longitude: -98.0233E, Elevation: 397 m) of the Mesonet Sites. Since the values at each of the Mesonet Sites are independent of the values at other sites, calculations for one site shall be used for demonstration purposes.

Fig. 2 – Fig 4 depicts the daily solar energies of some station in different forms.

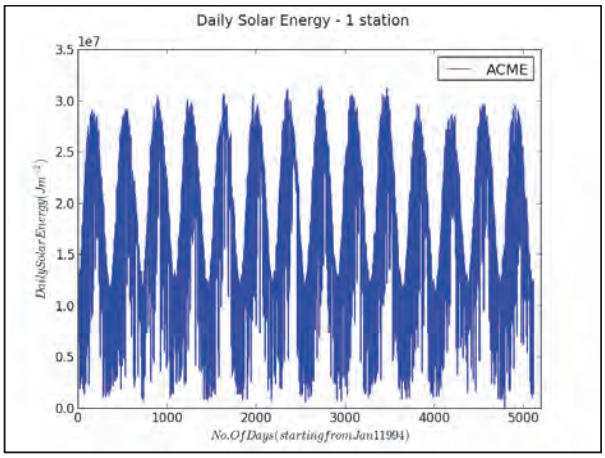

Fig. 2: Daily solar energy at 'ACME' station for years 1997-2007

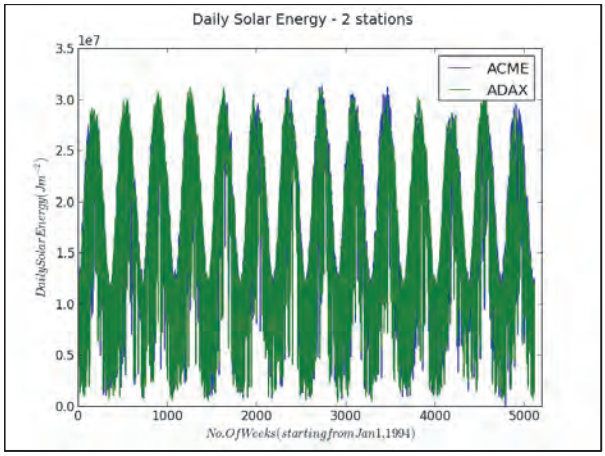

Fig. 3: Daily solar energy at 'ACME' vs. 'ADAX' stations for years 1997-2007

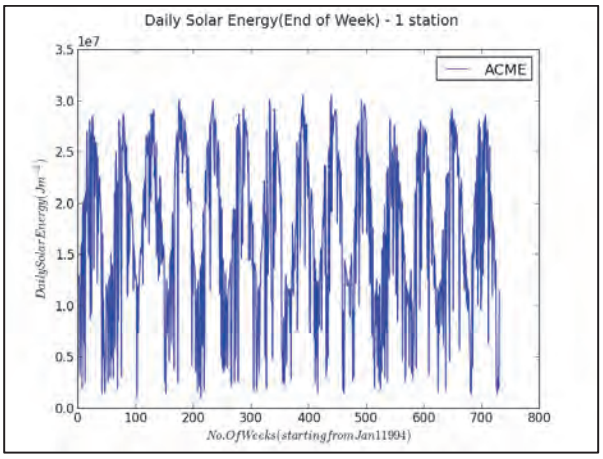

Fig. 4: Daily solar energy at 'ACME' station (every seventh day) for years 1997-2007

After some preprocessing the 15 features were mapped to the Mesonet Sites. Fig. 5-Fig. 9 shows variation of 5 out 15 features [1] for 1100 days starting from January 1, 1994. The data is shown for 'ACME' station.

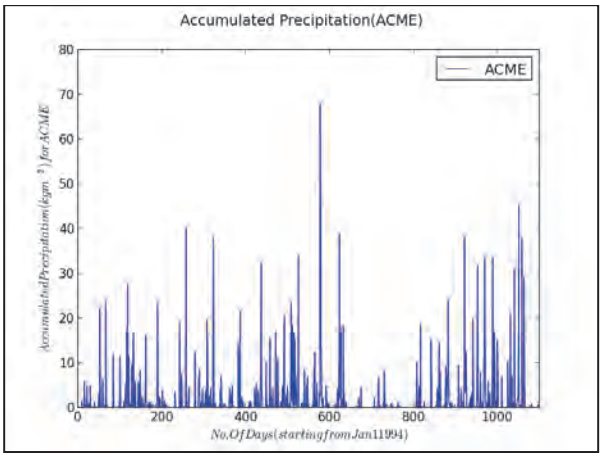

Fig. 5: Accumulated Precipitation of 'ACME' for 1100 days starting January 1, 1994.

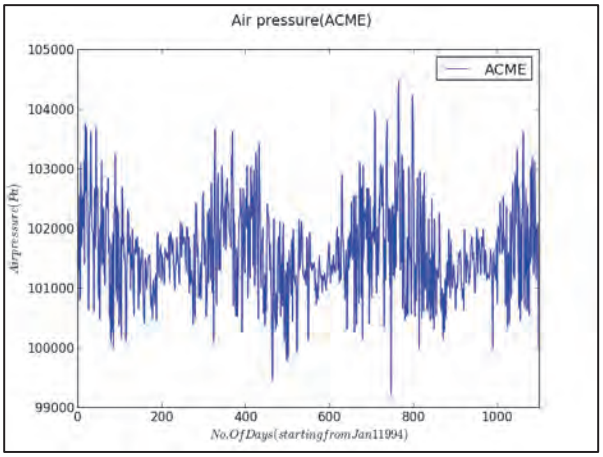

Fig. 6: Air Pressure of 'ACME' for 1100 days starting January 1, 1994.

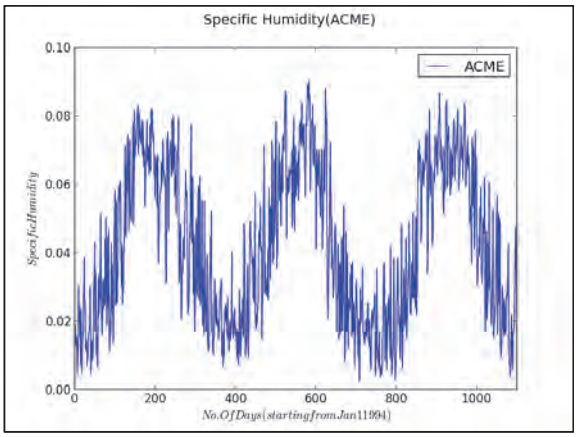

Fig. 7: Specific Humidity of 'ACME' for 1100 days starting January 1, 1994.

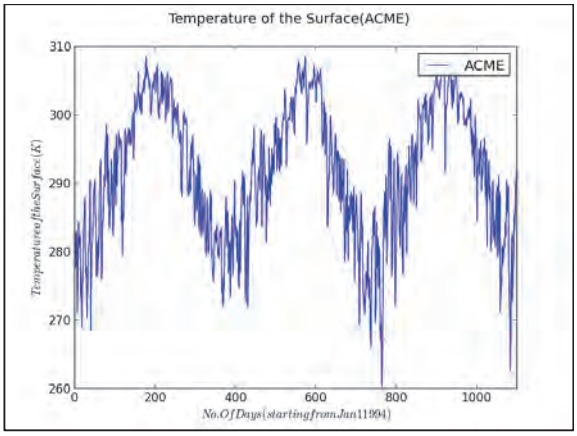

Fig. 8: Temperature at the Surface for 'ACME' for 1100 days starting January 1, 1994.

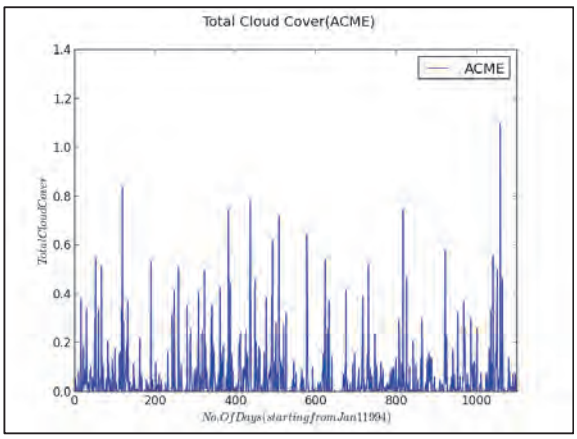

Fig.9: Total Cloud Cover for 'ACME' for 1100 days starting January 1, 1994.

IV. Prediction Models

Both the observational and forecasted are presented in a time-series format. The data for the features and daily solar energy is obtained for 14 years (1997- 2004 or 5113 days). This data is divided in training, cross-validation and test ing set. Train set has 3097 days whereas cross validation and test set each have 1023 days of data. As mentioned earlier two models are used for prediction—Linear Least Square Regression and Gradient Boosting Regression.

A. Linear Least Square Regression

Linear least squares regression is a commonly-used technique to predict the relationship between a dependent or response variable, e.g., daily solar energy, and a set of independent features on which the daily solar energy may depend. The regression minimizes the sum of the squared differences between the observed solar energy and the solar energy predicted by a linear approximation of the forecast weather models [2, 4]. Each feature is assigned a weight which indicates the contribution or importance of that feature in predicting the daily solar energy. The Root Mean Squared Error (RMSE) when Linear Least Square Regression is applied to values on the cross- validation set is found to be 5.23 * 10^6 J m^{-2}. Fig. 10 shows the observed and forecasted daily solar energy for the 1023 days in the cross-validation set when Linear Least Square Regression is applied to it.

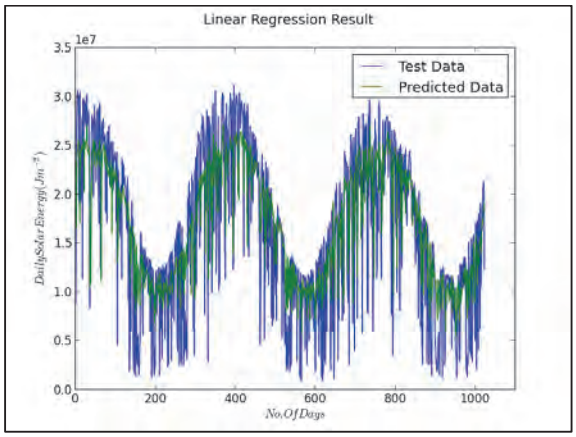

Fig.10: Observed and forecasted daily solar energy when linear least square regression is applied to the cross- validation set.

B. Gradient Boosting Regression

Gradient Boosting is a regression technique, wherein a prediction model is produced by an collection of models such as decision trees. The model is built stage-by-stage and is then generalized by optimizing a chosen loss function [6]. The loss function used in this case in 'Least Squares'. The Scikit-Learn package of Python [3] is used for implementation of Gradient Boosted Regression. Initially, the default values for number of boosting stages (n_estimators = 100) and learning rate (lr = 1.0) was used. This gave an RMSE 5.17 * 10^6 J m^{-2} when applied on the cross validation set. Fig. 11 shows the observed and forecasted daily solar energy for the 1023 days in the cross-validation set when Gradient Boosted Regression is applied to it.

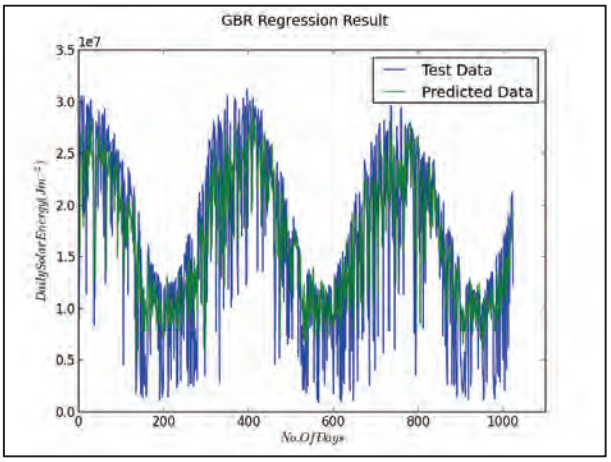

Fig.11: Observed and forecasted daily solar energy when Gradient Boosted Regression is applied to the cross- validation set.

The results of Gradient Boosting Regression are only slightly better than Linear Least Square Regression. Therefore, some changes in Gradient Boosting Regression were made to give better results. Two main parameters—number of boosting stages and learning rate are analyzed to get better results. Fig. 11- Fig 13 show variations in RMSE with respect to number of boosting stages and learning rate.

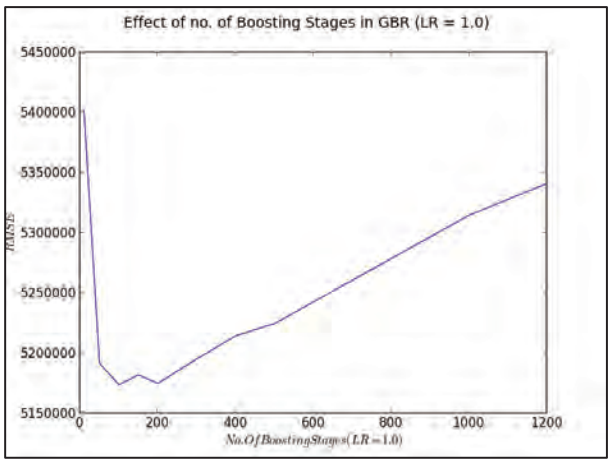

Fig.12: Effect of number of boosting stages on RMSE keeping learning rate fixed at 1.0

Here, it is clear that if learning rate is fixed at 1.0 then RMSE is minimum for 100 boosting stages.

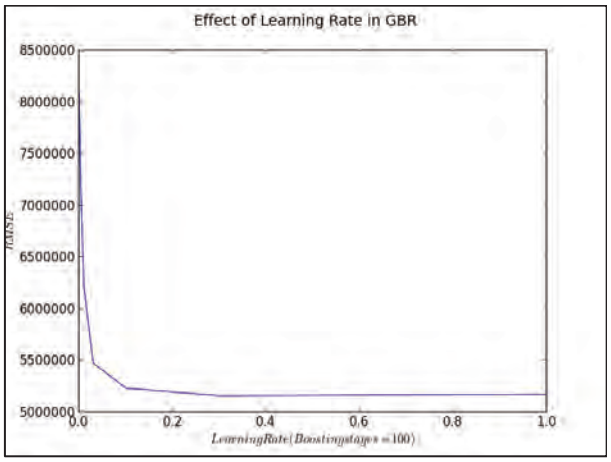

Fig.13: Effect of learning rate on RMSE keeping number of boosting stages fixed at 100

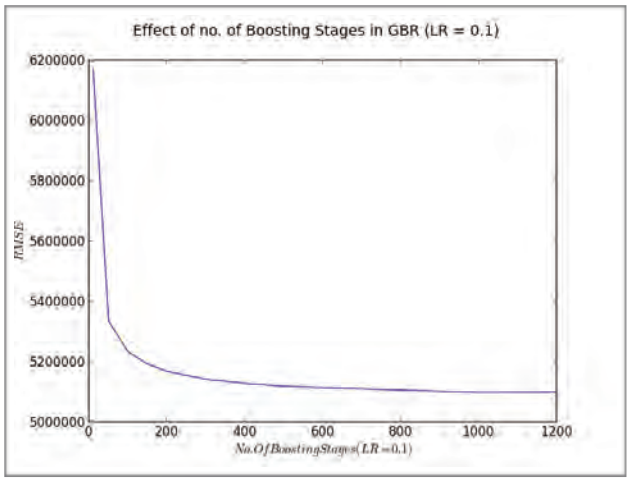

Fig.14: Effect of number of boosting stages on RMSE keeping learning rate fixed at 0.1

After, varying the parameters—number of boosting stages and learning rate, it is concluded that Gradient Boosted Regression will give best results for boosting stages set to 1000 and learning rate set to 0.1. These parameters gave a RMSE 5.09 * 10^6 J m^{-2} when applied on the cross validation set. Fig 14 shows results of the modified Gradient Boosted Regression on the cross- validation set. This result is 2.6 % more accurate than Linear Least Square Regression.

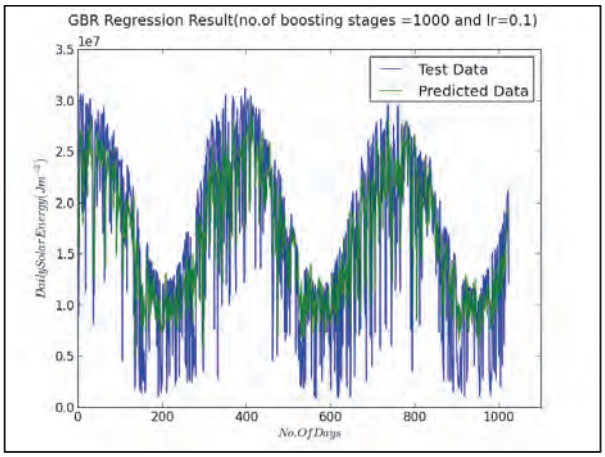

Fig.15: Observed and forecasted daily solar energy when Gradient Boosted Regression (number of boosting stages = 1000 and learning rate = 0.1) is applied to the cross- validation set.

V. CONCLUSION

Gradient Boosted Regression technique applied to the domain of solar energy prediction has proved to be more accurate as compared to Linear Least Square Regression. The findings of this paper are:

- Linear Least Square Regression gives a RMSE of 5.23 * 10^6 J m^{-2} when applied to the cross-validation set.
- Gradient Boosted Regression (number of boosting stages = 100 and learning rate = 1.0) gives a RMSE of 5.17 * 10^6 J m^{-2} when applied to the cross-validation set.
- Gradient Boosted Regression (number of boosting stages = 1000 and learning rate = 0.1) gives a RMSE of 5.09 * 10^6 J m^{-2} when applied to the cross-validation set.
- This result is 2.6 % more accurate than Linear Least Square Regression.

Thus, we can conclude that by adjusting learning rate and boosting stage parameters, Gradient Boosting Regression gives better results than Linear Least Square Regression for the solar energy domain.

REFERENCES

1. Mellit, H. Eleuch, M. Benghanem, C. Elaoun, A. Massi Pavan, "An adaptive model for predicting of global, direct and diffuse hourly solar irradiance", Energy Conversion and Management 51 (2010) 771–782, 2009.
2. N. Sharma, J. Gummeson, D. Irwin, and P. Shenoy, "Predicting Solar Generation from Weather Forecasts Using Machine Learning," University of Massachusetts Amherst
3. Scikit-learn: Machine Learning in Python
4. N. Sharma, J. Gummeson, D. Irwin, and P. Shenoy, "Leveraging Weather Forecasts in Energy Harvesting Systems," University of Massachusetts Amherst, Tech. Rep., September 2011.
5. Kaggle: Go from Big Data to Big Analytics, http://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/data
6. J. Friedman. Greedy function approximation : A gradient boosting machine. Technical report, Stanford University, 1999.

APPENDIX

This list contains all the 15 features on which daily solar energy may depend along with their units. This data was taken from [5].

Variable	Description	Units
apcp_sfc	3-Hour accumulated precipitation at the surface	kg m^{-2}
dlwrf_sfc	Downward long-wave radiative flux average at the surface	W m^{-2}
dswrf_sfc	Downward short-wave radiative flux average at the surface	W m^{-2}
pres_msl	Air pressure at mean sea level	Pa
pwat_eatm	Precipitable Water over the entire depth of the atmosphere	kg m^{-2}
spfh_2m	Specific Humidity at 2 m above ground	kg kg^{-1}
tcdc_eatm	Total cloud cover over the entire depth of the atmosphere	%
tcolc_eatm	Total column-integrated condensate over the entire atmos.	kg m^{-2}
tmax_2m	Maximum Temperature over the past 3 hours at 2 m above the ground	K
tmin_2m	Mininmum Temperature over the past 3 hours at 2 m above the ground	K
tmp_2m	Current temperature at 2 m above the ground	K
tmp_sfc	Temperature of the surface	K
ulwrf_sfc	Upward long-wave radiation at the surface	W m^{-2}
ulwrf_tatm	Upward long-wave radiation at the top of the atmosphere	W m^{-2}
uswrf_sfc	Upward short-wave radiation at the surface	W m^{-2}