Energy Conversion and Management 164 (2018) 102–111

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China

Junliang Fan[a], Xiukang Wang[b], Lifeng Wu[c,*], Hanmi Zhou[d], Fucang Zhang[a,e], Xiang Yu[f], Xianghui Lu[c], Youzhen Xiang[a]

[a] *Institute of Water-saving Agriculture in Arid Areas of China, Northwest A&F University, Yangling 712100, China*
[b] *College of Life Sciences, Yan'an University, Yan'an 716000, China*
[c] *School of Hydraulic and Ecological Engineering, Nanchang Institute of Technology, Nanchang 330099, China*
[d] *College of Agricultural Engineering, Henan University of Science and Technology, Luoyang 471003, China*
[e] *Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas of Ministry of Education, Northwest A&F University, Yangling 712100, China*
[f] *Provincial Key Laboratory for Water Information Cooperative Sensing and Intelligent Processing, Nanchang Institute of Technology, Nanchang 330099, China*

ARTICLE INFO

Keywords:
Global solar radiation
Support Vector Machine
Extreme Gradient Boosting
Temperature
Precipitation

ABSTRACT

The knowledge of global solar radiation (H) is a prerequisite for the use of renewable solar energy, but H measurements are always not available due to high costs and technical complexities. The present study proposes two machine learning algorithms, i.e. Support Vector Machine (SVM) and a novel simple tree-based ensemble method named Extreme Gradient Boosting (XGBoost), for accurate prediction of daily H using limited meteorological data. Daily H, maximum and minimum air temperatures (T_{max} and T_{min}), transformed precipitation (P_t, 1 for rainfall > 0 and 0 for rainfall = 0) and extra-terrestrial solar radiation (H_0) during 1966–2000 and 2001–2015 from three radiation stations in humid subtropical China were used to train and test the models, respectively. Two combinations of input parameters, i.e. (i) only T_{max}, T_{min} and R_a, and (ii) complete data were considered for simulations. The proposed machine learning models were also compared with four well-known empirical models to evaluate their performances. The results suggest that the SVM and XGBoost models outperformed the selected empirical models. The performance of the machine learning models was improved by 5.9–12.2% for training phase and by 8.0–11.5% for testing phase in terms of RMSE when information of precipitation was further included. Compared with the SVM model, the XGBoost model generally showed better performance for training phase, and slightly weaker but comparable performance for testing phase in terms of accuracy. However, the XGBoost model was more stable with average increase of 6.3% in RMSE, compared to 10.5% for the SVM algorithm. Also, the XGBoost model (3.02 s and 0.05 s for training and testing phase, respectively) showed much higher computation speed than the SVM model (27.48 s and 4.13 s for training and testing phase, respectively). By jointly considering the prediction accuracy, model stability and computational efficiency, the XGBoost model is highly recommended to estimate daily H using commonly available temperature and precipitation data with excellent performance in humid subtropical climates.

1. Introduction

Accurate estimation of global solar radiation (H) is of great importance for the design and optimization of solar energy systems [61,57,54,34]. However, unlike other meteorological data (e.g. temperature and precipitation), measurements of global solar radiation are always not available for many worldwide locations owing to the high costs and technical complexities [33]. Therefore, various approaches have been proposed to predict H where lack of global solar radiation data, e.g. empirical models [12,35,1], artificial intelligence-based models [15,17,52] and satellite-based methods [36,67,8], etc. Among the above methods, empirical and intelligence-based models are most commonly used due to their model simplicity and high prediction accuracy, respectively [61,30,33,24].

Over the past few decades, many efforts have been made to predict H from different types of empirical models, e.g. sunshine-based models [4,9,10], cloudiness-based models [27,7,39], temperature-based models [44,68,30], day number-based models [41,37,53] and hybrid

* Corresponding author.
E-mail address: china.sw@163.com (L. Wu).

https://doi.org/10.1016/j.enconman.2018.02.087
Received 4 January 2018; Received in revised form 25 February 2018; Accepted 26 February 2018

Nomenclature

Variables

C	penalty parameter of the error
H	global solar radiation (MJ m^{-2} day^{-1})
H_0	extra-terrestrial solar radiation (MJ m^{-2} day^{-1})
K	kernel function
l	loss function
n	number of observations
ΔT	diurnal temperature range (°C)
P	precipitation (mm)
P_t	transformed precipitation
T_{max}	daily maximum temperature (°C)
T_{min}	daily minimum temperature (°C)
φ	higher-dimensional feature space
ω	weights vector
ε	tube size
λ	regularization parameter
γ	minimum loss
Ω	regularization term

Constants

a, b, c, d, e, f, g	empirical coefficients

Abbreviations

ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Networks
DT	Decision Tree
Extra-Trees	Extremely Randomized Trees
FFA	Firefly Algorithm
GB	Gradient Boosting
GRNN	Generalized Regression Neural Networks
MAE	mean absolute error (MJ m^{-2} day^{-1})
MBE	mean bias error (MJ m^{-2} day^{-1})
ML	machine learning
MLP	Multi-layer Perceptron
R^2	coefficient of determination
RF	Random Forest
RMSE	root mean square error (MJ m^{-2} day^{-1})
SVM	Support Vector Machine
SVR	Support Vector Regression
WT	Wavelet Transform
XGBoost	Extreme Gradient Boosting

models which estimate H by introducing more meteorological variables [4,12,45]. Although the sunshine-based and cloudiness-based empirical models generally provide better H estimates than those based on air temperatures [70,20], they are not readily accessible at many locations around the world. Therefore, various temperature-based empirical models have been established to predict daily H from daily maximum/minimum temperature for being the most available weather parameters at any stations [30]. Hargreaves and Samani [28] proposed the first temperature-based model for estimating H by assuming that the difference between maximum and minimum temperatures was largely related to the ratio of global and extraterrestrial solar radiation. Following this work, various modified forms of Hargreaves and Samani model have been developed to improve H estimates from the diurnal temperature range, e.g. power models [13,29,3,11], logarithmic models [18] and polynomial models [56,30,23]. The accuracy of single temperature-based models were found to be further improved by incorporating other readily accessible weather data such as relative humidity and precipitation [42,19,5,54].

Although empirical models have been widely employed for H estimation, they are difficult to handle complex and nonlinear relationships between independent and dependent variables in noisy environments, particularly in humid regions where H is greatly affected by heavy clouds during rainy days [38]. Thus, many machine learning techniques have been used to predict H, e.g. Artificial Neural Networks (ANN) [69,43], Support Vector Machines (SVM) [17,52], Generalized Regression Neural Networks (GRNN) [65,25] and Adaptive Neuro Fuzzy Inference System (ANFIS) [60,50,71], etc. Among these artificial intelligence models, the ANN algorithms are the most frequently used models [64]. However, the SVM model has been recently proposed as a promising alternative for H estimation as a result of higher prediction accuracy and computational efficiency compared with the ANN model [55,46,64]. Wu and Liu [66] applied the SVM model to estimate monthly mean daily H from ambient temperatures at 24 stations across China. It was found that their newly developed SVM model outperformed the traditional empirical models. Chen et al. [16] compared seven SVM models and five sunshine-based empirical models for the estimation of daily H using different inputs of sunshine duration at three stations in China. All the SVM models gave much better performance than the studied empirical models. Ramli et al. [55] investigated the performance of SVM and ANN models for H estimation on the tilted surface at two sites in Saudi Arabia. The SVM model had significantly higher accuracy, robustness and computational speed in predicting H compared to the ANN model. Quej et al. [52] evaluated the performance of three artificial intelligence models (i.e. SVM, ANN and ANFIS) to estimate the daily H based on measured meteorological variables for a warm sub-humid environment in México. They concluded that the SVM model outperformed the other two machine learning models. The SVM model has also been hybridized with other algorithms, e.g. Firefly Algorithm (FFA) [49] and Wavelet Transform (WT) [47,48,21] to optimize the calibration process and improve the prediction accuracy.

Most of the well-established artificial intelligence models, however, are complex and require high computational costs during training phase. Rule-based Decision Tree (DT) and tree-based ensemble methods, e.g., Gradient Boosting (GB) and Random Forest (RF) and Extremely Randomized Trees (Extra-Trees) have recently begun to attract people's attention, because they are simple but still powerful and robust predictive algorithms [31,26,51]. Hassan et al. [31] explored the potential of four tree-based methods (i.e. Bagging, GB, RF and DT) for modeling H and compared their performance with Multi-layer Perceptron (MLP) and Support Vector Regression (SVR). The results indicate that the proposed tree-based models could provide accurate H estimation despite of being relatively simple. Recently a novel simple tree-based ensemble method named XGBoost has been developed by Chen et al. [17], which is an improved version of gradient boosting with higher computation efficiency and better capability to deal with over-fitting problems [17]. Despite being widely used in many other fields [58,6,59], the model is rarely applied in global solar radiation studies. To the best of the authors' knowledge, only Urraca et al. [62] tested its applicability for H estimation in central Spain, and compared its performance with four different approaches, i.e. Antonanzas model, on-ground Ordinary Kriging, a satellite-based dataset of CM-SAF-SARAH and a reanalysis dataset of ERA-Interim. The results show that errors obtained with the XGBoost model were slightly higher than the satellite-based SARAH and the Ordinary Kriging interpolation, but lower than the other two methods.