Energy Conversion and Management 119 (2016) 121–129

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Assessing the potential of random forest method for estimating solar radiation using air pollution index

Huaiwei Sun [a,b,*], Dongwei Gui [c], Baowei Yan [a], Yi Liu [a], Weihong Liao [d], Yan Zhu [b], Chengwei Lu [a], Na Zhao [a]

[a] School of Hydropower and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[b] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China
[c] Cele National Station of Observation & Research for Desert-Grassland Ecosystem in Xinjiang, Chinese Academy of Sciences, Urumqi 830000, China
[d] State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China

ARTICLE INFO

Article history:
Received 24 November 2015
Received in revised form 12 April 2016
Accepted 14 April 2016
Available online 18 April 2016

Keywords:
Random forest
Solar radiation
Variable importance
Air pollution index

ABSTRACT

Simulations of solar radiation have become increasingly common in recent years because of the rapid global development and deployment of solar energy technologies. The effect of air pollution on solar radiation is well known. However, few studies have attempting to evaluate the potential of the air pollution index in estimating solar radiation. In this study, meteorological data, solar radiation, and air pollution index data from three sites having different air pollution index conditions are used to develop random forest models. We propose different random forest models with and without considering air pollution index data, and then compare their respective performance with that of empirical methodologies. In addition, a variable importance approach based on random forest is applied in order to assess input variables. The results show that the performance of random forest models with air pollution index data is better than that of the empirical methodologies, generating 9.1–17.0% lower values of root-mean-square error in a fitted period and 2.0–17.4% lower values of root-mean-square error in a predicted period. Both the comparative results of different random forest models and variance importance indicate that applying air pollution index data is improves estimation of solar radiation. Also, although the air pollution index values varied largely from season to season, the random forest models appear more robust performances in different seasons than different models. The findings can act as a guide in selecting used variables to estimate daily solar radiation and improve the accuracy of solar radiation estimation.

1. Introduction

Solar energy technology plays an essential role in providing a sustainable energy future. Nearly half of the global solar photovoltaic (PV) is produced in China, which is the largest solar PV manufacturer in the world. In addition, a Chinese long-term plan estimates a target install capacity for solar PVs for the year 2020 to be 20 GW. For a given location, information on solar radiation is essential to the design and evaluation of solar energy technologies. However, much of this information is not readily available because of a limited number of meteorological stations in the world. Several studies have devised methods to estimate solar radiation [1].

A major challenge in estimating solar radiation concerns its variability. Several studies have observed alternating rates of rising and declining solar radiation at land surfaces [2]. Spatiotemporal changes in the amount of solar irradiance have been investigated by means of data collected over many years [3]. As numerous studies reveal, urban development [4] and aerosols linked to air pollution affect variations in solar radiation [5]. At least one study [6] has shown that human activities such as burning of fuels and mining cause pollutants (tiny particles of elemental carbon, for example) to be released into the atmosphere. Aerosols are known to act as both direct and indirect radiative forcings in reducing radiation [7]. Direct radiative forcings include methods of scattering and absorbing solar radiation. Indirect radiative forcings include material means of attenuating surface solar radiation by acting as cloud condensation nuclei to increasing cloud reflectivity and lifespan [8]. Confidential evidence exists that increased air pollution leads to declines in solar radiation and thus the number of hours of sunshine in China [9]. In addition, at least one study has deduced that

* Corresponding author at: School of Hydropower and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China.
E-mail address: huaiweisun@whu.edu.cn (H. Sun).

http://dx.doi.org/10.1016/j.enconman.2016.04.051
0196-8904/

Table 1
Information about the general geography, climates, and API data records of the three sites.

Stations	Latitude	Longitude	Daily solar radiation (MJ/m^2)			Average daily API in the different seasons				
			Mean	Max.	Min.	All	Spring	Summer	Autumn	Winter
Haikou	20°01′48″	110°21′00″	14.09	30.52	0.02	36.4	35.8	27.1	39.6	43.1
Changchun	43°54′00″	125°13′12″	13.28	31.50	0.01	72.4	75.9	58.6	68.6	87.0
Urumqi	43°46′48″	87°39′00″	14.04	34.62	0	103.1	86.6	60.4	89.6	177.4

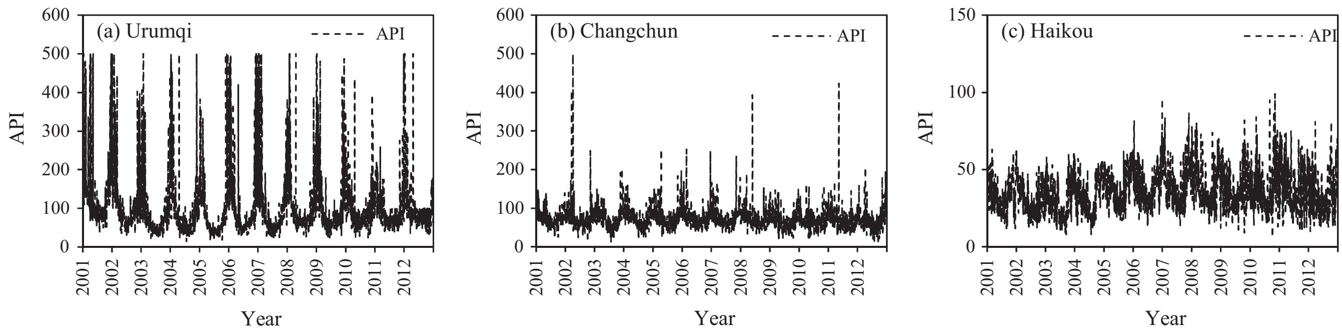

Fig. 1. The daily API in three different sites.

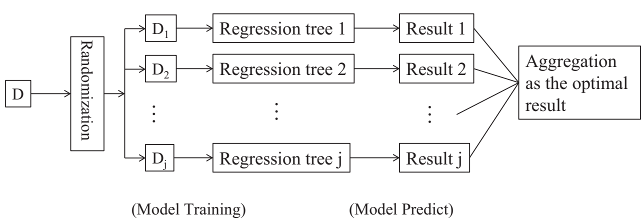

Fig. 2. Random forest regression principle.

air pollution can affect solar radiation estimation that utilizes meteorological variables. Specifically, Zhao et al. [10] have developed and validated several expression models for predicting daily global radiation by introducing API data as a new variable. A more thorough understanding of the API will enable us to better estimate solar radiation.

Previous studies have revealed that sunshine duration is the most common variable used to estimate solar radiation. In addition to the well-known Angstrom–Prescott (A–P) model, different models have been introduced to improve estimation performance [11]. Modifications to the A–P model are mainly for three reasons [12]. First, the A–P coefficients are related to geographic elements or other meteorological variables [13]. Second, the basic structure of models have been changed from linear to nonlinear forms [14], such as raising relative sunshine to second or third order [15]. Third, easily measured meteorological variables have been added to the original A–P formula. Hargreaves and Samani [16] proposed a method to estimate solar radiation by maximum and minimum temperature. Relative humidity data were also identified to be relevant with solar radiation and employed to estimate solar radiation by Elagib et al. [17]. Also, De Jong and Stewart [18] related global solar radiation to air temperatures, precipitation, and cloudiness in the wheat-growing area of western Canada. As reported by Persaud et al. [19] in Botswana and Niger, the coefficients of the A–P models are affected by air pollution, revealing that the API has major potential in accurately estimating solar radiation. However, as more variables have been introduced into the prediction model, the high-dimensional aspect of solar radiation estimation has created an urgent need to measure variable importance and avoid overfitting. An increasing number of machine learning techniques have been studied in the area of energy. The random forest (RF) technique [20], which includes an ensemble of decision trees and incorporates feature selection and interactions naturally in the learning process, presents a good option for estimating solar radiation. By using the RF method, one can obtain predictions and identify predictors that are associated with the response via RF's inbuilt variable importance measures, and also it provides the possibility of taking the ordering information into account and may yield improved predictions. As far as we know, the potential of air pollution index has been identified in estimating solar radiation [10], but the importance of air pollution index has not been studied.

The aim of this study is to apply a RF approach for estimation of solar radiation in three different sites with different API conditions

Table 2
Input variables for estimating solar radiation proposed in literatures.

Models	Types	Input variables	Equations	Source
Sunshine duration based	Linear	S, S_0, S/S_0	$Rs = Ra(aS/S_0 + b)$	[26,27]
	Cubic	S, S_0, S/S_0, S/S_0^2, S/S_0^3	$Rs = Ra(aS/S_0 + bS/S_0^2 + cS/S_0^3 + d)$	[28]
	Exponential	S, S_0, S/S_0, $exp(S/S_0)$	$Rs = Ra(aS/S_0 + b\,exp(S/S_0) + c)$	[11,29]
Temperature based	Power	$Tmax$, $Tmin$, ΔT, $\Delta T^{0.5}$	$Rs = Ra(a\Delta T^{0.5} + b)$	[30,16]
	Exponential	$Tmax$, $Tmin$, ΔT, $exp(\Delta T)$	$Rs = Ra(a - a\,exp(b\Delta T))$	[31]
API based	Linear	API/100	$Rs = Ra(a + bS/S_0 + c\mathrm{API}/100 + dS/S_0(\mathrm{API}/100))$	[10]
	Exponential	exp(API/100)	$Rs = Ra(a + bS/S_0 + c\,exp(\mathrm{API}/100) + dS/S_0\,exp(\mathrm{API}/100))$	[10]
	Logarithmic	log(API/100)	$Rs = Ra(a + bS/S_0 + c\log(\mathrm{API}/100) + dS/S_0\log(\mathrm{API}/100))$	[10]