GLOBAL SOLAR RADIATION PREDICTION MODEL WITH RANDOM FOREST ALGORITHM

by

Hakan KOR*

Department of Computer Engineering, Faculty of Engineering, Hitit University, Corum, Turkey

Original scientific paper
https://doi.org/10.2298/TSCI200608004K

Global solar radiation estimation is crucial for regional climate assessment and crop growth. Therefore, studies on the prediction of solar radiation are emerging. With the availability of the public data on solar radiation, computerized models have been developed as well. These predictive models play significant role in determining the potentials of regions suitable for renewable energy generation required by engineering and agricultural activities. Herein a computerized model has been presented for estimating global solar radiation. The model utilizes random forest algorithm and reached predictive value of 93.9%.

Key words: *random forest, solar radiation, prediction model*

Introduction

Sufficient but secure energy is one of the main ingredients for welfare and economic development of the society. Nowadays, as its most general definition energy is perceived as the potential for causing changes. Since energy-related activities have significant side effects on the environment, legal enforcements to the energy sector are increasing. Thus, fundamental structural changes in the energy sector, called energy transitions, occur worldwide. When the main source of energy is examined, it can be said that before 1950, the coal ranked the first in energy production and consumption, while oil came to the fore with the discovery of rich oil resources [1]. With the 1973 oil crisis, people understood, that oil was not a reliable sufficient energy source [2]. Population statistics directly influence the world energy balance. The world population, which is currently approaching 8 billion, is estimated to reach 16 billion in 2055 [3]. The desire for faster mass development potentially results in more energy consumption. Nevertheless, fuels with carbon waste cause irreparable damage to the nature. Since energy resources such as coal, oil and natural gas are highly limited and it is known that they will be depleted after a certain period of time, the rapid increase of the population has led people to discover different types of reliable and harmless energy resources.

In order to avoid the dangerous consequences of fossil-based energy sources on nature, it is necessary to turn to natural resources such as sun, wind, geothermal, biomass, seawave and hydrogen, which are also named as green energy sources [4, 5]. Solar energy is considered as a clean energy source, which has not yet been completely exploited [6]. Solar energy is also positively supported, reinforced by the Kyoto protocol and various laws regarding of green energy [7]. This energy can be smoothly converted into electricity, another

* Author's e-mail: hakankor@hitit.edu.tr

more usable form of energy, by using photovoltaic power generation systems to combat global warming. When photovoltaic power generation systems are connected to the electrical grid, predicting global solar radiation becomes significant to stabilize the entire network [8]. There are a wide range of algorithms and methods utilizing to predict the solar radiation. One of the most widespread usage algorithms is perceived as the random forest algorithm [9]. In this study, it is determined that the random forest algorithm has a high estimation rate for solar radiation estimation processes.

Related works

Global solar radiation comprehension is compulsory to predict, study and design the economic viability of solar energy systems. Analyzes were made based on four year global solar radiation data measured on a horizontal surface over the Čačak region, Republic of Serbia. As a result, the annual and monthly optimum tilt angles were determined by converting the available solar radiation data on a horizontal surface into various tilt angles [10].

Premalatha and Valan [11], in their study, used meteorological data collected over the last ten years from five different locations in India to accurately predict solar radiation. It has been developed two artificial neural network (ANN) models with four different algorithms. In addition, the best ANN model was extracted based on the minimum mean absolute error (MAE), root mean square error (RMSE), and maximum linear correlation coefficient, R^2. The developed ANN model evaluates solar energy installations where there are no meteorological data measurement facilities. In the process of estimating monthly average global radiation, low mean absolute percentage error (MAPE) is effective in determining the accuracy and conformity of the model [12], They claim that the global solar radiation should be considered as the most essential parameter in meteorology, renewable energy, and solar energy conversion applications, especially in the dimensioning of independent photovoltaic systems [13]. It is stated that solar radiation is the fundamental source of the Earth's energy, which provides almost 99.97% of the heat energy required for a wide various chemical and physical procedures in the atmosphere, ocean, land and other water partial bodies. It constitutes an essential key role in solar radiation as a renewable energy source.

The consideration of ready and reliable data is absolutely required for the design, installation, optimization and performance evaluation of solar technologies in any random region. The daily sun data of horizontal and inclined planes are significantly required for daily monitoring of the performance of the applications. These applications include lighting and agricultural processes. Solar radiation data are associated with diverse solar energy conversion devices and appropriate model design to be able to implementation [14].

Chen *et al.* [15], presents several reports on solar radiation measurements. These are appreciated as solar water heating, agricultural studies, wood drying, photovoltaic, which designs some solar energy applications such as thermal load, evaluating potential power levels, solar radiation measurements, atmospheric energy balance studies and meteorological forecasts. This energy could be converted into electricity which is recognized as another beneficial type of energy, using photovoltaic power generation systems to combat global warming [8].

Gumus and Kilic [16], suggested that a brand-new approach for the eastern region for the global solar radiation and sunshine duration of the potentially highest solar energy of Turkey in the basis of data from previous years of forecast. This proposed method estimates key parameters by using time series and an analysis method. This method is perceived as the exponential weighted moving average. This model clearly predicts global solar radiation and

sunlight duration for the next year and is evaluated by statistical parameters, MAPE and coefficient of determination, R^2, to examine the success of the proposed technique. In another study on solar radiation conducted in Turkey, Sanlıurfa, Harran University Mechanical Engineering measurements between the years 2009-2016 by using data obtained from the solar tracking system solar radiation measurement systems of the Department are calculated. The relationship between diffuse radiation rate and the clarity index is utilized to obtain it from eight years of data, three horizontal solar diffuser radiation models have been proposed. Using 15 diffuse radiation models given in the literature and the results obtained from the models, horizontal solar diffuse radiation values were calculated and the obtained data were compared with the measurement data for Sanlıurfa. [17].

It is been asserted that various prediction models have been proposed with solar radiation data on a monthly or annual basis in different countries of the world. A simple model has been developed to estimate the daily global solar radiation contain and to illustrate horizontal plane through nine year meteorological data from the Tabouk region, which is completely located in the geography of Saudi Arabia. Five different meteorological parameters were taken into consideration in creating the model. In addition, statistical comparison of this developed model and other models in the literature was conducted [18]. In the study conducted in China, various diffuse solar radiation models have been developed through 40 years of solar radiation data which were collected from 14 different stations. It has been claimed that the second order polynomial equation provides the healthiest consequences as a result of China being a very common geography in terms of its geographical location. The created models were compared statistically with the measured data [19]. In the research which was conducted in India, solar radiation data from five different cities between 2001 and 2005 were used. It has been concretely announced that angstrom type first-, second-, and third-degree solar radiation models specific to each region have been developed with these data. The monthly total insolation irradiance values of the data measured with the created models were compared statistically [20].

Material and methods

Today, with the population growth and the industrial revolution, dependence on energy has increased. In addition fossil fuels, RES are also frequently used to meet energy needs. Solar energy is the leading RES. In this section, the dataset and the method used are examined in detail. The analysis process of the research was managed with open source python language. In the analysis, the code snippets were absolutely directed on a python-based Jupiter notebook. Libraries such as NumPy, matplotlib, seaborn and pandas are included in Jupiter. The csv file containing 32687 records was operated as the data source. In the next stage, the terms of time and location adjustments were implemented for the *Pacific/Honolulu* region. Grouping operations were performed to present the values of some variables in hourly, daily, weekly or monthly periods. In the next step, the variable rates that could predict the global solar radiation value were determined and subjected to correlation analysis. Correlation analysis is considerably applied to identify the possible relationship between two or more variables. Correlation illustrates the functional form of a linear relationship between two or more variables. Nevertheless, when the value of one of the variables is acknowledged, it provides prediction about the other value [21]. The variables with the best predictive values of the random forest algorithm were determined. The random forest classifier is a meta estimator which fits a set of decision tree classifiers to various subsamples of the dataset and identically utilized the mean to improve prediction accuracy and control overfitting.

Random forest algorithms

Random forest algorithm is one of the most popular machine learning algorithms [22]. The random forest algorithm is referred as a combination of random decision trees [23]. This algorithm produces a prediction value by averaging the estimates of randomly generated decision trees [24]. The mathematical equation which is executed to produce the random forest algorithm predictive value is given [25]:

$$\hat{Y} = \frac{1}{N}\sum_{n=1}^{N} T_n(X) \tag{1}$$

where the average values of Y comes from n, N, and $T_n(x)$. The X input parameters refer to the number of decision trees in the N random forest. The equation describes the average of decision trees T_n, $n = 1,2,...,N$ for the input X with the aim of obtaining a robust prediction.

Honolulu region and amount of energy daily produced

Acknowledging the radiation data of a certain region is a very important parameter in the design of solar powered systems. In the figure, from a general perspective, the amount of energy produced per square meter on a monthly basis according to different solar panel locations in the Honolulu region can be seen.

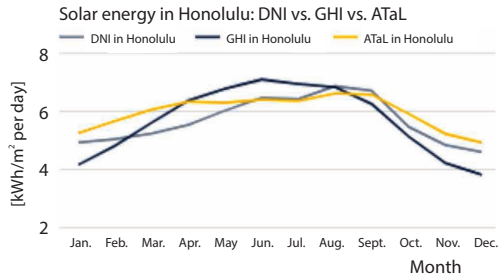

Figure 1. Energy amounts produced from solar panels in Honolulu region [27]

Let's briefly explain the terms in fig. 1.

Direct normal irradiance (DNI): The total amount of solar radiation received per unit area by a surface that is always perpendicular to the sun rays that come in a straight line from the direction of the sun at its current position in the sky.

Global horizontal irradiance (GHI): The total amount of solar radiation received per unit area by a surface that is always positioned in a horizontal manner.

Average tilt at latitude (ATaL): The total amount of solar radiation received per unit area by a surface that is tilted toward the equator at an angle equal to the current latitude. The ATaL will often produce the optimum energy output [26].

The research data includes 32687 lines and 11 different variables which belong to the Pacific/Honolulu region, which were registered between September-December 2016.

The section containing a very small part of the research data is shown in fig. 2. In the research data, the variables of time, radiation value, temperature, pressure, humidity, wind direction, speed, sunrise time and sunset time are included.

	UNIXTime	Data	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	TimeSunRise	TimeSunSet
7416	1472724008	9/1/2016 12:00:00 AM	00:00:08	2.58	51	30.43	103	77.27	11.25	06:07:00	18:38:00
7415	1472724310	9/1/2016 12:00:00 AM	00:05:10	2.83	51	30.43	103	153.44	9.00	06:07:00	18:38:00
7414	1472725206	9/1/2016 12:00:00 AM	00:20:06	2.16	51	30.43	103	142.04	7.87	06:07:00	18:38:00
7413	1472725505	9/1/2016 12:00:00 AM	00:25:05	2.21	51	30.43	103	144.12	18.00	06:07:00	18:38:00
7412	1472725809	9/1/2016 12:00:00 AM	00:30:09	2.25	51	30.43	103	67.42	11.25	06:07:00	18:38:00

Figure 2. Part of the data set

Model evaluation

The success rate of the model was compared to R^2, RMSE values. In the implementation of the model, 11 variables were used as inputs for estimating the radiation value, and after processing with five different algorithms, the random forest algorithm giving the best prediction rate was selected. Bagging meta-estimator, forests of randomized trees, AdaBoost, gradient tree boosting and histogram-based gradient algorithms were not used because of their low predictive value.

Examining of hourly and monthly values

In this part of the research, the radiation, temperature, pressure and humidity values which were obtained on an hourly and monthly basis were examined.

Hourly and monthly solar radiation values of Honolulu region is seen in fig. 3. According to these values, when the solar radiation value showed a logarithmic increase between 07-13 hours on an hourly basis, it showed a logarithmic decrease between 13-18 hours. When monthly values are examined, it is seen that the average radiation values in September, Octobe, and November were the same, but decreased in December.

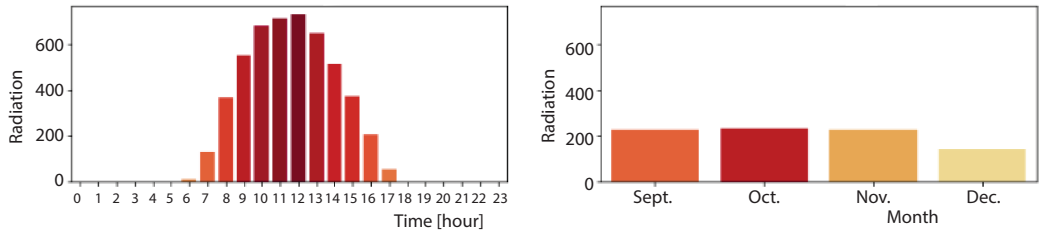

Figure 3. Hourly and monthly solar radiation values

Hourly and monthly temperature values of Honolulu region are seen in fig. 4. According to these values, it was observed that the temperature value decreased between 01-06 hours, increased between 07-13 hours and decreased again between 14-24 hours on an hourly basis. When the monthly values are examined, it has been determined that there is a decrease in the temperature value between September and November.

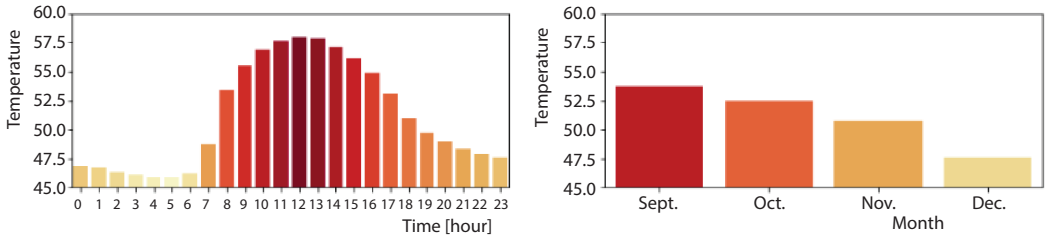

Figure 4. Hourly and monthly temperature values

Hourly and monthly pressure values of Honolulu region is seen in fig. 5. According to these values, it was observed that there was a decrease in the temperature value between 01-04 hours on an hourly basis, an increase between 04-10 hours, a decrease between 11-16 hours, an increase between 17-22 hours and a decrease between 23-24 hours. When the monthly pressure values are analyzed, it is seen that the pressure value increased between September and November, while a sharp decrease was observed in December.

Hourly and monthly humidity values of Honolulu region is seen in fig. 6. According to these values, although there is not much difference between 01-08 hours on an hourly basis, a slight decrease is observed between 3-6 hours. While humidity values increased between 8-16 hours, it decreased between 17-24 hours. When the monthly values are analyzed, it is seen that the months of September and October were almost the same, but there was a sharp decrease in November and a sharp increase in December.

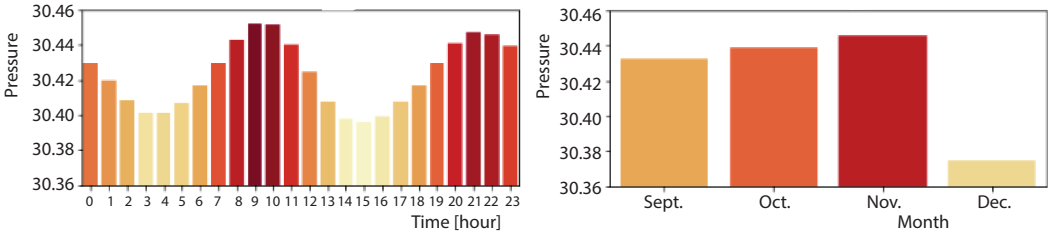

Figure 5. Hourly and monthly pressure values

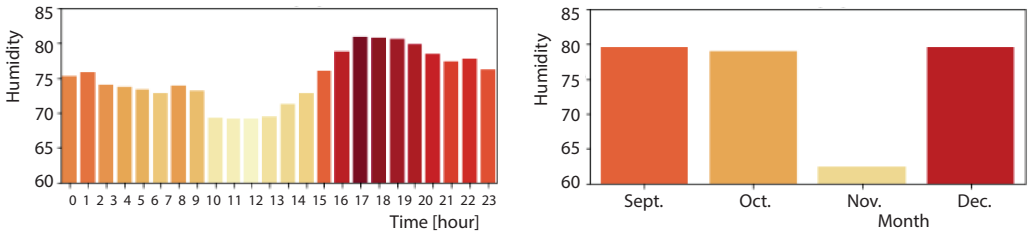

Figure 6. Hourly and monthly humidity values

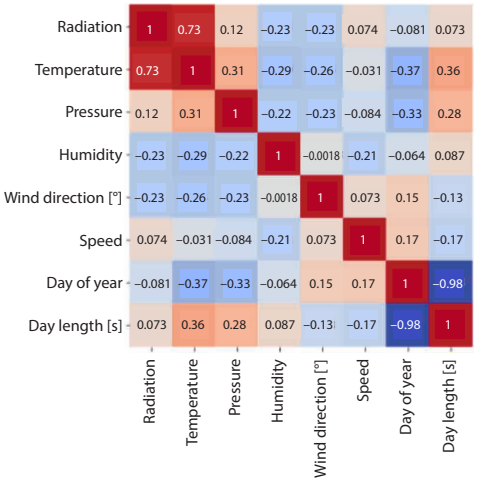

Figure 7. Correlation of variables

	Features	R^2 Score
0	Temperature, Pressure, Humidity, WindDirection...	0.932824
1	Temperature, Humidity, WindDirection(Degrees),...	0.931121
2	Temperature, Humidity, DayOfYear, TimeOfDay(s)	0.933630
3	Temperature, DayOfYear, TimeOfDay(s)	0.932989
4	Temperature, TimeOfDay(s)	0.800729

Figure 8. Features and R^2 scores

Result

While 10% of the randomly selected data set was reserved for testing without being included in the training, 80% of the remaining was reserved for training and 20% for validation. After that, hyper parameter setting was done.

The correlation relationship is shown in fig. 7. Accordingly, there is a positive relationship between temperature and radiation.

In fig. 8, effect of properties on R^2 prediction scores are seen. It is seen that temperature, humidity, day of the year, and time of day values have a positive effect on prediction success, while it is determined that they have the lowest estimation rate when two variables, namely temperature and time of day. In addition, explained variance, MSE and R^2 values are given below.

explained variance = 0.9390530586138153

MSE = 6288.783329309797

R^2 = 0.9389912461683354

Coefficient of determination

The stability coefficient can be used to determine the linear relationship between calculated and measured values [27]:

$$R^2 \frac{\sum_{i=1}^{n}(X - X_m)(Y - Y_m)}{\sqrt{\left[\sum_{i=1}^{n}(X - X_m)^2\right]\left[\sum_{i=1}^{n}(Y - Y_m)^2\right]}} \tag{2}$$

The coefficient R^2 in the previous equation explains the model's statistical ability and determines predictions for future outcomes. Namely, suppose the model's independent variable is x and dependent variable y, then the coefficient R^2 governs the variation in y when x changes. It is among the key factors in correlation analysis and is used when one wants to predict future outcomes, even testing models having related information. The coefficient R^2 lies between 0 and 1. It is indeed the square of the correlation coefficient. Like as it is true for the correlation coefficient, the closer the value of R^2 to 1, the better the prediction and strength of the model.

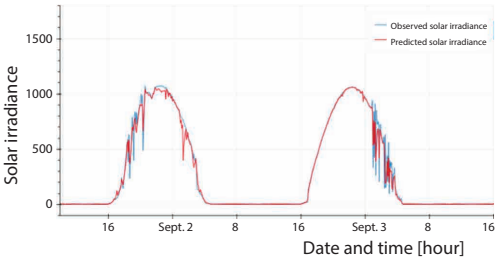

Figure 9. Observed and predicted values

Figure 9 shows the difference between the observed and predicted values on the test data. As a result, observed and predicted values are compatible with each other.

Conclusion and discussion

Today, as a result of the extensive use of fossil-based resources such as coal, oil, and natural gas, it is known that serious damage is inflicted on nature and causes global warming. Therefore, more focus should be placed on the wind, solar, solar-hydrogen, hydroelectric, and geothermal energy sources known as green energy.

Numerous studies have been conducted since the first studies to calculate the global solar radio, which began in 1924. Wu *et al.* [28] applied different models to calculate daily global solar radiation in their studies, and the best value was calculated as 93% . Yao *et al.* [29] achieved 87.6% correct calculation success in their study with support vector machines. Zhan *et al.* [30] compared 12 different models for global solar radiation prediction and found the highest estimated rate of 91%. In her study in Spain-Madrid, Almorox [31] reached a rate of 92% at a high confidence level. According to the results of the literature review, many studies have been done on global solar radiation calculations and estimation and continue to be done with different models. In this study, 93.9% correct estimation value was reached. Based on the research data, it has been determined that the random forest algorithm is more effective. In subsequent studies, it is aimed to compare data sets containing the same variable values from different geographical regions.

References

[1] Kallioglu, M. A., Niğde İli İçin Yatay Düzenleme Gelen Günlük Tüm. Yayılı ve Direkt Güneş Işınımını Hesaplama Modeli Geliştirilmesi (Improving a Model for Calculating Daily Global, Diffuse and Direct Solar Radiation on Horizontal Surfaces for Niğde – in Turkish), Ph. D. thesis, Nigde University, Nigde, Turkey, 2014

[2] Gurbuz, A., Enerji Piyasası İçerisinde Yenilenebilir (Temiz) Enerji Kaynaklarının Yeri ve Önemi", *Proceedings*, (The Place and Importance of Renewable (Clean) Energy Resources in the Energy Market – in Turkish), 5th Inter. Advanced Technologies Symp., Karabuk, Turkey, 2009
[3] Rifkin, J., Howard, T., *Entropi Dünyaya Yeni Bir Bakış* (A New Look at the Entropic World – in Turkish), Yayınevi, Istanbul, Turkey, 1997
[4] Ultanır, M. O., Solar Energy is on the Verge of the Century, *Bilim ve Teknik Dergisi*, *340* (1996), pp. 50-55
[5] Varinca, K. B., Varank, G., Rüzgar Kaynaklı Enerji Üretim Sistemlerinde Çevresel Etkilerin Değerlendirilmesi ve Çözüm Önerileri (Assessment of Environmental Impacts and Solution Suggestions in Wind Based Power Generation Systems – in Turkish), *Proceedings*, New and Renewable Energy Resources/ Energy Management Symposium, Kayseri, Turkey, pp. 367-376, 2005
[6] Sudirman, R., *et al*., Comparison of Methods Used for Forecasting Solar Radiation, *Proceedings*, IEEE Green Technologies Conference, Tuksa, Okla., USA, 2012, pp. 1-3
[7] Martin, L., *et al*., Prediction of Global Solar Irradiance Based on Time Series Analysis: Application Solar Thermal Power Plants Energy Production Planning, *Solar Energy*, *84* (2010), 10, pp. 1772-1781
[8] Kamadinata, J. O., *et al*., Global Solar Radiation Prediction Methodology using Artificial Neural Networks for Photovoltaic Power Generation Systems, *Proceedings*, SMARTGREENS, Porto, Portugal, 2017, pp. 15-22
[9] Ozdemir, S., Random Forest Yöntemi kullanılarak potansiyel dağılım modellemesi ve haritalaması: Yukarıgökdere Yöresi örneği (Potential Distribution Modeling and Mapping Using Random Forest Method: the Case of Yukarigökdere Region – in Turkish), *Turkish Journal of Forestry*, *19* (2018), 1, pp. 51-56
[10] Dragićević, S., Vučković, N. M., Evaluation of Distributional Solar Radiation Parameters of Čačak Using Long-Term Measured Global Solar Radiation Data, *Thermal Science*, *11* (2007), 4, pp. 125-134
[11] Premalatha, N., Valan, A. A., Prediction of Solar Radiation for Solar Systems by Using ANN Models with Different Back Propagation Algorithms, *Journal of Applied Research and Technology*, *14* (2016), 3, pp. 206-214
[12] Yadav, A. K., Chandel, S. S., Solar Radiation Prediction Using Artificial Neural Network Techniques: A Review, *Renewable and Sustainable Energy Reviews*, *33* (2014), May, pp. 772-781
[13] Kalogirou, S. A., *Solar Energy Engineering: Processes and Systems*, Academic Press, New York, USA
[14] Khalil, S. A., Shaffie, A. M., A Comparative Study of Total, Direct and Diffuse Solar Irradiance by Using Different Models on Horizontal and Inclined Surfaces for Cairo, Egypt, *Renewable and Sustainable Energy Reviews*, *27* (2013), Nov., pp. 853-863
[15] Chen, C., *et al*., Smart Energy Management System for Optimal Microgrid Economic Operation, *IET Renewable Power Generation*, *5* (2011), 3, pp. 258-267
[16] Gumus, B., Kilic, H., Time Dependent Prediction of Monthly Global Solar Radiation and Sunshine Duration Using Exponentially Weighted Moving Average in Southeastern of Turkey, *Thermal Science*, *22* (2018), 2, pp. 943-951
[17] Beyazıt, N. I., *et al*., Modelling of the Hourly Horizontal Solar Diffuse Radiation in Sanliurfa, Turkey, *Thermal Science*, *24* (2019), 2, pp. 939-950
[18] Maghrabi, A. H., Parameterization of a Simple Model to Estimate Monthly Global Solar Radiation Based on Meteorological Variables, and Evaluation of Existing Solar Radiation Models for Tabouk, Saudi Arabia, *Energy Conversion and Management*, *50* (2009), 11, pp. 2754-2760
[19] Che, H. Z., *et al*., Analysis of 40 Years of Solar Radiation Data from China, 1961-2000, *Geophysical Research Letters*, *32* (2005), 6, pp. 1-5
[20] Katiyar, A. K., Pandey, C. K., Simple Correlation for Estimating the Global Solar Radiation on Horizontal Surfaces in India, *Energy*, *35* (2010), 12, pp. 5043-5048
[21] Draper, N. R., Smith, H., *Applied Regression Analysis*, John Wiley & Sons, New York, USA, 1998
[22] Ren, S., *et al*., Global Refinement of Random Forest, *Proceedings*, IEEE Conference on Computer Vision and Pattern Recognition, Boston, Mass., USA, 2015, pp. 723-730
[23] Biau, G., Scornet, E., A Random Forest Guided Tour, *Test*, *25* (2016), 2, pp. 197-227
[24] Criminisi, A., *et al*., Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning, *Foundations and Trends® in Computer Graphics and Vision*, *7* (2012), 2-3, pp. 81-227
[25] Ahmad, M. W., *et al.*, Trees vs Neurons: Comparison between Random Forest and ANN for High-Resolution Prediction of Building Energy Consumption, *Energy and Buildings*, *147* (2017), July, pp. 77-89
[26] ***, Solar Energy in Honolulu, https://www.solarenergylocal.com/states/hawaii/honolulu/
[27] Ulgen, K., Hepbasli, A., Solar Radiation Models – Part 2: Comparison and Developing New Models, *Energy Sources*, *26* (2004), 5, pp. 521-530

[28] Wu, G., *et al*., Methods and Strategy for Modelling Daily Global Solar Radiation with Measured Meteorological Data – A Case Study in Nanchang Station, China, *Energy Conversion and Management*, *48* (2007), 9, pp. 2447-2452
[29] Yao, W., *et al.*, A Support Vector Machine Approach to Estimate Global Solar Radiation with the Influence of Fog and Haze, *Renewable Energy*, *128* (2018), Part A, pp. 155-162
[30] Zhang, Q., *et al.*, Comparative Analysis of Global Solar Radiation Models in Different Regions of China, *Advances in Meteorology*, *2018* (2018), ID 3894831
[31] Almorox, J., Estimating Global Solar Radiation from Common Meteorological Data in Aranjuez, Spain, *Turkish Journal of Physics*, *35* (2011), 1, pp. 53-64

Paper submitted: June 8, 2020
Paper revised: October 19, 2020
Paper accepted: October 26, 2020