Analysing the Determinants of Surface Solar Radiation with Tree-Based Machine Learning Methods: Case of Istanbul

Denizhan Guven (✉ guvende@itu.edu.tr)
Eurasia Institute of Earth Sciences, Istanbul Technical University

Research Article

Keywords: Surface Solar Radiation, Decision Tree, Random Forest, Gradient Boosted Regression Tree, Machine Learning, Forecasting

Posted Date: September 27th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3365597/v1

Additional Declarations: No competing interests reported.

Analysing the Determinants of Surface Solar Radiation with Tree-Based Machine Learning Methods: Case of Istanbul

Denizhan Guven[1]

[1] Eurasia Institute of Earth Sciences, Istanbul Technical University, Istanbul 34469, Turkey

Corresponding Author: Denizhan Guven, ITU Ayazaga Campus, Eurasia Institute of Earth Sciences, 34469, Maslak, Istanbul / Turkey, E-mail: guvende@itu.edu.tr, ORCID: 0000-0003-1605-7661

Abstract

This study estimates both hourly and daily Downward Surface Solar Radiation (SSR) in Istanbul while determining the importance of variables on SSR using tree-based machine learning methods, namely Decision Tree (DT), Random Forest (RF), and Gradient Boosted Regression Tree (GBRT). The hourly and daily data of climatic factors for the period between January 2016 and December 2020 are gathered from the European Centre for Medium-Range Weather Forecasts' (ECMWF) ERA5 reanalysis data sets. In addition to the meteorology data, hourly data of selected aerosols are obtained from the Ministry of Environment, Urbanization and Climate Change. Temperature, cloud coverage, ozone level, precipitation, pressure, and two components of wind speeds, PM_{10}, $PM_{2.5}$, and SO_2 are utilized to train and test the established models. The model performances are determined with the out-of-bag errors by calculating R-squared, MSE, RMSE, and MBE. The GBRT model is found to be the most accurate model with the lowest error rates. Furthermore, this study provides the variable importance in determining the SSR. Although all models provide different values for the variable importance; temperature, ozone level, cloud coverage, and precipitation are found to be the most important variables in estimating daily SSR. For the hourly estimation, the time of day (hour) becomes the most important factor in addition to temperature, ozone level, and cloud coverage. Finally, this study shows that the tree-based machine learning methods used with these variables to estimate hourly and daily SSR results are very accurate when it is not possible to measure the SSR values directly.

Keywords: Surface Solar Radiation, Decision Tree, Random Forest, Gradient Boosted Regression Tree, Machine Learning, Forecasting

Statements and Declarations

The author declares that he/she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Nomenclature

ANN	Artificial Neural Network	NWP	Numerical Weather Prediction
AOD	Aerosol Optical Depth	O_3	Total ozone layer
ARIMA	Autoregressive Integrated Moving Average	PDP	Partial Dependence Plot
BSA	Black Sky Albedo	PM10	Particulate matter with a diameter of 10 microns or less
CART	Classification and Regression Tree	PM2.5	Particulate matter with a diameter of 2.5 microns or less
CC	Cloud Coverage	PREC	Precipitation
CNN	Convolutional Neural Network	PRES	Pressure
COD	Cloud Optical Depth	PSO	Particle Swarm Optimization
DT	Decision Tree	PV	Photovoltaic
ELM	Extreme Learning Machine	RF	Random Forest
GBRT	Gradient Boosted Regression Tree	RMSE	Root Mean Square Error
GCM	Global Circulation Model	RSS	Residual Sum of Squares
IFS	Integrated Forecasting System	RT	Random Tree
IQR	Interquartile Range	RTM	Radiative Transfer Models
LSTM	Long Short Term Memory	SO2	Sulphur dioxide
MAE	Mean Absolute Error	SSR	Surface Solar Radiation
MARS	Multivariate Adaptive Regression Spline	SVR	Support Vector Regression
MBE	Mean Bias Error	Temp	Temperature
MLR	Multiple Linear Regression	u10	Eastward Wind Speed
MSE	Mean Square Error	v10	Northward Wind Speed
NN	Neural Networks	WRF	Weather Research and Forecasting Model
nRMSE	Normalized Root Mean Square Error	ZA	Zenith Angle

1. Introduction

1.1. Research Background

Surface Solar Radiation (SSR), which is the solar radiation reaching the earth's surface, is one of the vital surface energy balance elements. Observing the radiative fluxes using land-based monitoring stations became widespread after the 1950s, mainly metering the downward solar element (Hartmann et al. 2013).

Based on this SSR monitoring, SSR has exhibited notable variations, from declining to raising, which is also denoted as *from global dimming to brightening*. However, it should be noted that the word "global" here refers to "global radiation", rather than denoting the worldwide aspect of the phenomenon. In this context, there are two different SSR trends; i) a decline trend between the 1950s and 1980s (Global Dimming), and ii) an increase trend from the 1980s to the 2000s (Brightening) (Wild et al. 2005; Wild 2009; Ohmura 2009). Rather than external impacts by the sun, the origination of these variations is mostly based on the changes in transparency of the atmosphere resulting from cloud type, aerosols, and radiatively active gases such as water vapour (Wild 2012; Willson and Mordvinov 2003).

SSR, also named as global solar radiation, surface insolation, or downward shortwave irradiance, is one of the main components used to build agricultural crop modelling, analyse watershed and run-off in hydrology, optimize building energy use, estimate solar energy potential (AWG Radiation Budget Application Team 2018). In today's world, the role of solar energy in the transition towards cleaner and greener energy production has been growing expeditiously. The share of solar power in total electricity production has increased from 0.15% in 2010 to 3.7% in 2021 (EMBER 2022). Hence, forecasting and analysing of SSR have become very important, especially for the electricity market.

Currently, there are four primary techniques to obtain SSR, including ground measurements, reanalysis data, Global Circulation Models (GCMs), and satellite observations (Wei et al. 2019). Although most of the machine learning techniques do not provide any information for the physics while forecasting SSR, former studies showed that machine learning methods are one forceful way to estimate SSR using satellite observations in addition to the aforementioned techniques (Fan et al. 2020; Chen et al.

2019). Machine learning techniques provide many advantages: i) determining the most important variables in estimating SSR, ii) providing mechanisms to quantify the uncertainties, iii) utilizing different types of remote sensing data, iv) capturing the non-linear relations between dependent and independent variables, and v) assessing the robustness of the model (Zhou et al. 2017).

1.2. Research significance and case study characteristics

Climate change is one of the most significant phenomena at the present time. It is well-known that greenhouse gases are the main reason of climate change. To abate the greenhouse gases released into the atmosphere, the energy sector has been experiencing a radical change in terms of the entrance of renewable energy sources into the electricity market. However, the amount of electricity generated from renewable energy sources is highly dependent on meteorological parameters such as wind speed and solar irradiance. Since the unpredicted electricity generation from renewable energy sources may cause some problems for both electricity transmission and distribution systems, the estimation of these parameters is also very useful for better planning and integration of the produced energy into the grid (Luiz et al. 2018).

The electricity generated from solar photovoltaic (PV) and solar thermal systems constitutes more than 3.7 percent of global electricity production as of 2021. Besides, PV systems have been the fastest-growing sources for electricity generation for 17 years (EMBER 2022). This means that the prediction of output from PV and solar thermal systems is becoming more important for electricity grid operators, energy companies, and governments (Voyant et al. 2017).

Over the past few years, Türkiye has been actively working towards diversifying its energy sources and boosting its utilization of renewable energy. The country possesses substantial potential for renewable energy, specifically solar, wind, hydro, and geothermal resources. As a result of these endeavours, Türkiye has managed to raise the proportion of renewable energy sources (excluding hydro) in its overall installed capacity from 3.5 percent to 22.96 percent within a decade. In this regard, the installed solar energy capacity of Türkiye as of December 2022 stood at 9425 MW, constituting 9.1% of the total installed power (Ministry of Energy and Natural Resources 2023). Thus, to maximize the benefit of PV systems, analysing the determinants of SSR is crucial.

Determining the most significant variables for the SSR and exhibiting how much these variables influence the SSR may set groundwork for maintaining an efficient and profitable operation for PV systems.

1.3. Adopted literature review

The number of studies on solar radiation forecasting is gradually increasing in the literature. As mentioned in the previous section, there are four major methods to obtain solar radiation. However, this literature review includes studies with machine learning methods.

Artificial Neural Networks (ANN) is one of the most common machine learning models for estimating solar radiation. Within this context, Jiang (2008) utilized the ANN model to determine the monthly mean daily diffuse solar radiation for China using solar radiation data from nine stations. Furthermore, the established ANN model was tested for Zhengzhou station, and the results showed that the model can forecast the actual values with an accuracy of about 94.8 percent. Similarly, Mubiru and Banda (2008) applied the same method to examine SSR in Uganda using sunshine duration, maximum temperature, cloud cover, and location parameters. The prediction accuracy was found to be very successful, with a Mean Bias Error (MBE) of 0.059 MJ/m^2 and a Root Mean Square Error (RMSE) of 0.385 MJ/m^2. Furthermore, Voyant et al. (2011) predicted the daily global solar radiation using ANN for France. They also investigated the contribution of exogenous meteorological data. As a result of this study, the insertion of endogenous and exogenous data into the model provided a 1% decrease in the nRMSE (Normalized Root Mean Square Error) for the power production. More recently, Ryu et al. (2018) used ANN to estimate shortwave radiation, diffuse, and total photosynthetically active radiation globally. The result showed that the inter-annual variability of shortwave radiation at both site and continental levels are captured successfully.

In addition to ANN, there are a few studies using tree-based machine learning models. For instance, Zhou et al. (2017) examined Downward Solar Radiation in the U.S. with the Random Forest method. The spatiotemporal patterns were found to be consistent with the expected trends. Based on the RF model, the black sky albedo (BSA) near infrared band, BSA visible band, and clear day coverage were detected as the most important parameters to estimate downward solar radiation. Using the same method, Hou et al.

(2020) predicted surface downward shortwave radiation for China based on ground-measured data from 86 stations of the Climate Data Center of the Chinese Meteorological Administration. The results suggested that the RF method is adequate to predict surface downward shortwave radiation. For the daily forecasting, the results showed an overall R value of 0.92 and a RMSE value of 35.38 W/m^2. Yang et al. (2018) used the GBRT method to estimate SSR for China. On the daily time scale, the established GBRT model estimated SSR with an R^2 value of 0.82 and a RMSE value of 27.71 W/m^2. Furthermore, on the monthly time scale, it determined SSR with an R^2 value of 0.92 and a RMSE value of 15.4 W/m^2.

There are also some studies that use more than one method. Lima et al. (2016) established a model to predict solar irradiation in the North-eastern Brazil using a combination of the Weather Research and Forecasting Model (WRF) and ANN as post-processing method. The forecasted atmospheric outputs of the WRF model were utilized as forecaster by the established ANN model. This study revealed that the combination of WRF and ANN models provides a remarkable improvement in RMSE and the correlation coefficient. Wei et al. (2019) applied four different machine learning methods (i.e., RF, ANN, GBRT, and MARS (Multivariate Adaptive Regression Spline)) to estimate Downward shortwave radiation for China. The analysis showed that the best-performed method is GBRT. The GBRT model at the daily time scale predicted it with an RMSE of 30.34 W/m^2 and an R^2 of 0.90 under clear sky conditions, while these values were 42.03 W/m^2 and 0.86, respectively, under cloudy sky conditions. In their study, Srivastava et al. (2019) tried to forecast hourly solar radiation in India utilizing four different machine learning methods, namely MARS, Classification and Regression Tree (CART), M5Tree, and RF. It is shown that the RF model performs the best, on the other hand; the CART model is the worst among all four models.

1.4. Research gap and motivation

In Table 1, there is a summary of some examples of solar radiation forecasting studies. As far as is known, there is no specific study to Türkiye using tree-based machine learning models to determine the importance of variables in analysing the downward surface solar radiation. Furthermore, none of these studies present how and how much the selected factors impact the SSR. Hence, this study tries to fill these gaps by applying three different tree-based machine learning methods to the specific case of Istanbul, Türkiye.

Referance	**Location**	**Method**	**Estimation for**
Tymvios et al. (2005)	Cyprus	ANN, Ångström's linear approach	Global Solar Radiation
Jiang (2008)	China	ANN	Monthly Mean Daily Diffuse Solar Radiation
Mubiru and Banda (2008)	Uganda	ANN	Monthly Average Daily Global Solar Radiation
Hocaoglu et al. (2008)	Türkiye	2-D linear filters and ANN	Daily Solar Radiation
Lam et al. (2008)	China	ANN	Daily Global Solar Radiation
Mellit et al. (2010)	Saudi Arabia	ANN	Hourly Global, Diffuse And Direct Solar Irradiance
Qin et al. (2011)	China	ANN	Monthly-Mean Daily Global Solar Radiation
Voyant et al. (2011)	France	ANN	Daily Global Solar Radiation
Wang et al. (2012)	U.S	ANN	Land Surface Shortwave and Longwave Radiation
Martins et al. (2012)	Brazil	ANN	Downward Solar Radiation
Rahimikhoob et al. (2013)	Iran	ANN	Global Solar Radiation
Lima et al. (2016)	Brazil	ANN, NWP	Surface Solar Irradiance
Tang et al. (2016)	China	ANN	Surface Solar Radiation
Jiang et al. (2016)	U.S	MARS	Surface All-Wave Net Radiation
Zhou et al. (2017)	U.S	RF	Downward Solar Radiation
Deo and Sahin (2017)	Australia	ANN	Global Solar Radiation
Wang et al. (2017)	China	Neuro-Fuzzy and ANN	Daily Global Solar Radiation
Yang et al. (2018)	China	GBRT	Surface Downward Shortwave Radiation
Ryu et al. (2018)	Global	ANN	Shortwave Radiation, Diffuse and Total Photosynthetically Active Radiation
Feng and Li (2018)	China	ANN and M5Tree	Total, direct and diffuse solar radiation
Ghimire et al. (2019)	Australia	SVR, PSO	Global Solar Radiation
Sharafati et al. (2019)	Burkina Faso	RF, RT, RC	Daily Global Solar Radiation
Srivastava et al. (2019)	India	MARS, CART, M5, RF	Daily Solar Radiation
Kisi et al. (2019)	Türkiye	MARS, M5Tree, DENFIS	Monthly average daily global radiation
Wei et al. (2019)	China	RF, GBRT, ANN, MARS	Surface Downward Shortwave Radiation
Hou et al. (2020)	China	RF	Surface Downward Shortwave

			Radiation
Gürel et al. (2020)	Türkiye	ANN, Time series, Empirical	Monthly average daily global radiation
Zeng et al. (2020)	China	RF	Daily Global Solar Radiation
Hai et al. (2020)	Algeria	ELM, ARIMA, MLR	Daily Global Solar Radiation
Chen et al. (2021)	U.S.	RF	Half- hourly Global Solar Radiation
Qin et al. (2021)	Australia	Physics-based method	Surface Downward Shortwave Radiation
Sianturi et al. (2021)	Indonesia	Post-processing methods	Daily and monthly solar radiation
Vakitbilir et al. (2022)	North Cyprus	CNN, LSTM, SVR	Global horizontal irradiation
Singla et al. (2022)	-	LSTM	Global horizontal irradiation
Zhang and Chen (2022)	China	Spatial downscaling and temporal extrapolation	Daily Average Shortwave Solar Radiation
Bhattacharjee and Chowdhury (2022)	US	LSTM	Global horizontal irradiation
Basilio et al. (2023)	Brazil	ANN, Multivariate Adaptive Regression Spline	Daily Global Solar Radiation
Chen et al. (2023)	China	RNN, CNN, LSTM	Surface Solar Radiation

Table 1. Summary of the literature

1.5. Research objectives

In this study, three different tree-based machine learning techniques, namely Decision Tree (DT), Random Forest (RF), and Gradient Boosted Regression Tree (GBRT), are utilized to analyse the SSR. Each tree-based machine learning method has different advantages and disadvantages. While DT provides a very useful graphical display to interpret and explain the output very easily, its prediction accuracy is weaker than other methods, and it is very sensitive to small changes in data. Although RF and GBRT methods require higher computational time to train the model, they can provide variable importance and work with missing values.

Although the Neural Networks (NN) are very popular in prediction problems, tree-based methods do not require a GPU to complete training as NN does. In addition to this, tree-based algorithms may provide higher accuracy than NNs. Deep learning methods are more pertinent for applications including image and speech recognition, and language processing, whereas, tree-based algorithms surpass deep learning methods on tabular-style datasets. Furthermore, where the patterns or features that the model reveals may be more important than the prediction performance of the model, the interpretability power of tree-based models comes to the forefront in applications such as medical and meteorology (Lundberg et al. 2020). Thus, the tree-based models are more convenient for the purpose of this study.

In addition to machine learning methods, Radiative Transfer Models (RTM) also perform well to estimate SSR. However, physical parametrization techniques like RTM usually need various input parameters, such as aerosol optical depth (AOD) and cloud optical depth (COD), cloud height, cloud type, and the height of the cloud top and base. In most locations, it is inconvenient to obtain these parameters accurately and completely. Since the accuracies of these parameters have a great impact on the estimation accuracy, this situation may cause uncertainties for the results (Tymvios et al 2005; Wei et al 2019). Besides, these models are not suitable for common users because of their complexity (Wang et al. 2012). In light of this information, three different tree-based machine learning approaches are employed to fulfil the aim of this study.

Furthermore, certain variables within the dataset exhibit a significant correlation with each other. Given that tree-based ML methods employ decision trees, they possess the ability to effectively address the issue of multicollinearity (Yoon 2021). Hence, in the context of this study, utilizing tree-based approaches would be more advantageous.

Depending on the advantages of tree-based ML methods, the main purpose of this study is three-fold; i) examining the forecast accuracies of tree-based models, ii) investigating the determinants of hourly and daily SSR, and iii) presenting how and how much the selected factors impact the SSR, which is the main contribution of this paper to the literature. This paper is organised as follows: The second section exhibits a review of the methodologies used and explains the data, as well as the estimation results are presented and discussed. Finally, the conclusion is given in the last section.

2. Case Study and Data

To established the above-mentioned methods, the hourly meteorology data covering the January 2016 - December 2020 period were gathered from ERA5. The ERA5 database is created by the efforts of the fifth generation of the European Centre for Medium-Range Weather Forecasts' (ECMWF) atmospheric reanalysis of the global climate (Copernicus Climate Change Service 2021). ERA5 is acquired using data assimilation methods depending on the ECMWF's latest Integrated Forecasting System (IFS). The IFS aggregates model data with all convenient recorded in-situ and space-borne observations. It has been providing hourly data of worldwide atmospheric and surface parameters at the resolution of 0.28125° (31km) since 1979. In addition to the meteorology data, hourly data of PM_{10}, $PM_{2.5}$, and SO_2 are obtained from the Ministry of Environment, Urbanization and Climate Change (2022).

The performance of established models that use reanalysis data can be tested using in-situ observation data. In this study, ground-based global radiation data is obtained from the Meteorological Service. However, the number of data points for the working domain and time period is considerably low to test the established models. Since the number of data points may affect the validation process, ground-based measurement data is not included in this study to test the established ML models.

To obtain daily data, the means of hourly data were calculated for every day. The analysis was examined for Istanbul, Türkiye, and hence the coordinates of the interested area were selected accordingly. Furthermore, Antalya, Türkiye was selected to validate the best model accuracy. The units of data used in the models are given in Table 2.

Data	Unit	Data Source
Surface Solar Radiation (SSR)	W/m^2	ECMWF
Temperature (Temp)	°C	
Cloud Coverage (CC)	%	
Total Ozone column (O_3) *	Dobson unit	
Precipitation (PREC)	m	
Pressure (PRES)	hPa	
Eastward Wind Speed (u10)	m/s	
Northward Wind Speed (v10)	m/s	
Zenith Angle (ZA)	Degree	NOAA
PM_{10} (PM10)	µg/m^3	Ministry of Environment, Urbanization and Climate Change
$PM_{2.5}$ (PM2.5)	µg/m^3	
SO_2 (SO2)	µg/m^3	

Table 2. List of data and their units

* The ozone data used in this study is the total ozone column in the atmosphere. This parameter can also be referred to as vertically integrated ozone.

It is very helpful and useful to use the exploratory data analysis before the analysis. It helps the visualization of descriptive statistics. In this context, the summaries of hourly and daily data for the selected area are shown in Table 3-4, respectively. Hourly Downward Surface Solar Radiation ranges between 0.0 W/m^2 and 949.2 W/m^2, whereas daily average Downward Surface Solar Radiation varies in the range of 8.87 W/m^2 and

341.3 W/m^2. Comparing the interquartile range (IQR) of the aerosols (PM$_{10}$, PM$_{2.5}$, and SO$_2$), it can be said that PM$_{10}$ has a wider range of variability than other particles. Besides, the variability of northward wind speed is slightly bigger than that of eastward wind speed. The Bosphorus (also known as the Strait of Istanbul) could be the main reason for this situation since it lies in the north-south direction.

Variable	Min	q_1	Mean	Median	q_3	Max	St. Dev.	IQR
SSR	0.0	0.0	175.2	7.8	306.6	949.2	251.40	306.6
Temp	270.3	282.9	288.7	288.6	295.0	305.4	6.95	12.14
u10	-10.32	-3.47	-1.38	-1.62	0.66	9.82	2.86	4.13
v10	-14.01	-3.91	-1.52	-2.02	1.01	10.56	3.34	4.92
O_3	256.3	301.5	329.8	324.8	355.0	481.5	35.80	53.48
PREC	0.0	0.0	0.09	0.0	0.02	7.81	0.30	0.02
PRES	979.7	1001.4	1005.7	1005.1	1009.7	1029.7	6.33	8.33
CC	0.0	0.09	0.47	0.41	0.88	1.0	0.38	0.79
PM10	0.24	22.9	40.88	34.18	51.02	68.30	28.21	28.09
ZA	17.6	62.53	89.7	89.65	117.02	162.43	35.63	54.79
PM2.5	0.01	11.59	20.24	17.19	25.00	67.38	13.21	13.41
SO2	0.07	3.18	6.70	5.05	7.98	10.35	7.74	4.8

Table 3. Brief statistical analysis of hourly data

Variable	Min	q_1	Mean	Median	q_3	Max	St. Dev.	IQR
SSR	8.87	90.80	175.19	175.14	258.12	341.43	91.91	167.32
Temp	271.4	283.0	288.7	288.5	295.4	300.6	6.78	12.32
u10	-8.61	-3.28	-1.38	-1.44	0.49	6.34	2.54	3.77
v10	-11.64	-3.78	-1.52	-2.09	0.76	7.48	3.02	4.54
O_3	259.4	302.0	329.8	325.2	355.8	470.3	34.87	53.82
PREC	0.0	0.0	0.09	0.01	0.07	2.86	0.20	0.07
PRES	983.9	1001.4	1005.7	1005.0	1009.4	1027.9	6.14	8.00
CC	0.0	0.19	0.47	0.45	0.74	1.0	0.31	0.55
ZA	17.81	62.28	89.66	89.5	117.07	161.74	35.63	54.79
PM10	5.7	26.63	40.66	36.43	50.45	284.83	20.36	23.82
PM2.5	2.84	12.96	20.08	18.01	24.63	74.61	10.32	11.67
SO2	0.78	3.51	6.67	5.32	8.02	92.4	7.05	4.51

Table 4. Brief statistical analysis of daily data

3. Methodology

3.1. Models

In this section, three tree-based machine learning methods are explained.

3.1.1. Decision Tree

Decision Tree method provides one of the most efficient analyses for both regression and classification cases. DT is a supervised machine learning technique applying a group of hierarchical rules. An illustration with five regions is given in Figure 1. In this illustration, the variables x_1 and x_2 are separated by the set of predetermined points (a_1- a_4).

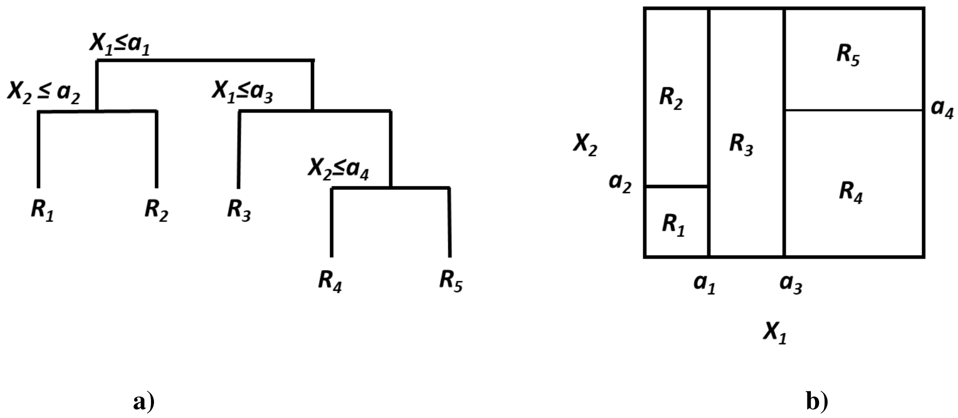

Fig. 1 a) The partition of a two-dimensional example **b)** The output of recursive binary splitting example

A decision tree structure comprises of a root node, a set of internal nodes, a group of terminal node known also as leaf, and a set of branches which attaches the nodes (See Figure 2). The main concept of DT method is to divide a complicated decision into diverse more elementary decisions, which conduct to an explanation that is simpler to interpret. Throughout the dividing process, DT algorithm separates the predictor space into *J* district and non-overlapping regions such as R_1, R_2, …, R_J. The model targets to find rectangles R_1, R_2, …, R_J that minimizes the Residual Sum of Squares (RSS) which is given as

$$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i-\hat{y}_{R_j})^2 \tag{1}$$

where $\hat{y}_{R_j}$ stands for the mean dependent variable for the training data within the j^{th} rectangle. The calculation of RSS is given in Eq. 2 (Gareth et al. 2013).

$$RSS=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{2}$$

On the other hand, it is not practicable to regard all potential partition of the feature space into *J* rectangles. Hence, a top-down approach, called *recursive binary splitting*, is applied. To apply it, the predictor X_j and the threshold point a that separates the predictor spaces into two non-overlapping regions $\{X|X_j \le a\}$ and $\{X|X_j > a\}$ must be selected, while minimizing the RSS given in Eq. 2 (Gareth et al. 2013).

$$\sum_{i:x_i\in R_1(j,a)=1}^{J}(y_i-\hat{y}_{R_1})^2+\sum_{i:x_i\in R_2(j,a)=1}^{J}(y_i-\hat{y}_{R_2})^2 \tag{3}$$

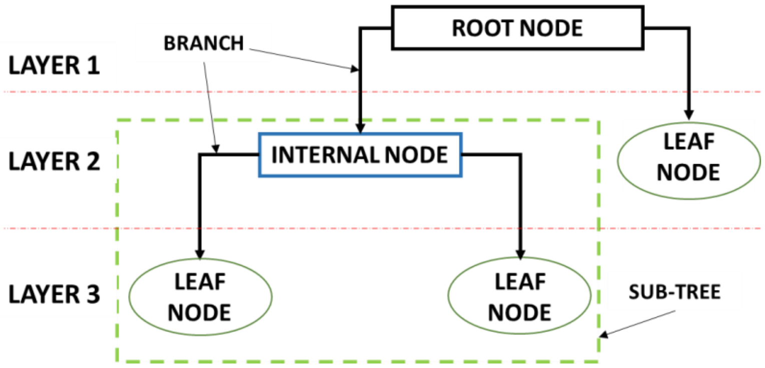

Fig. 2 Structure of Decision Tree

3.1.2. Random Forest

Random Forest (RF) algorithm also depends on the same concept with DT. After a number of trees (the forest) is composed, RF model uses a vote to unify the trees' forecasts. The structure of a RF model is shown in Figure 3. In Figure 3, the red circles indicate the prediction of each tree. After determining the prediction, the outputs of all trees are aggregated based on majority voting (for classification) or averaging (for regression).

RF possesses a built-in feature selection algorithm and, hence it is capable of managing many input variables without having to remove some variables to reduce dimensionality. Variable importance scores can be evaluated by calculating the growth in prediction error in the case of the values of a variable are exchanged across the out-of-bag observations, so called permutation testing. This score is determined for each individual tree, averaged across the whole ensemble and divided by the standard deviation (Shaikhina et al. 2019).

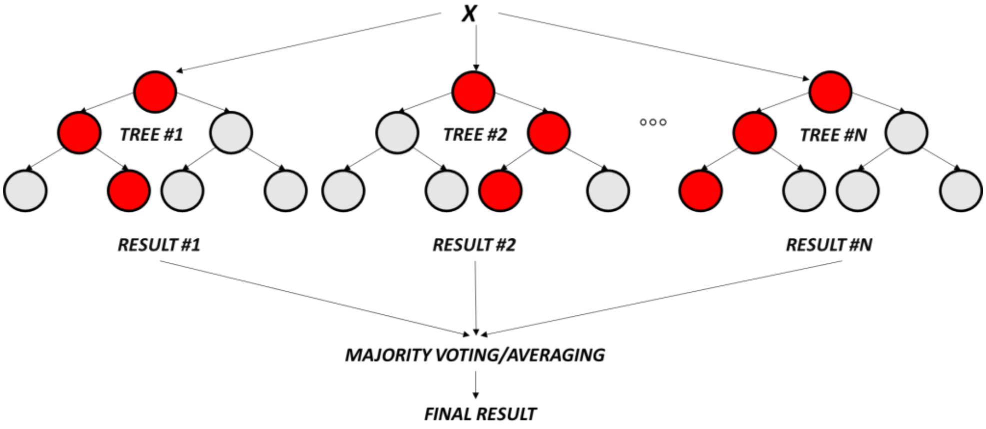

Fig. 3 Structure of a Random Forest model

The equation of average prediction of trees is (Breiman 2001):

$$F(x) = \frac{1}{J}\sum_{j=1}^{J} c_{j_{full}} + \sum_{k=1}^{K}(\frac{1}{J}\sum_{j=1}^{J} contribution_j(x,k)) \tag{4}$$

where c_{full} is the mean of the whole dataset, K is the total number of features, and J represents the number of trees in the forest.

3.1.3. Gradient boosted regression tree

Gradient Boosted Regression Tree algorithm originates an assemblage of shallow trees in a row with each tree learning and developing on the preceding tree, whereas RF algorithm establishes an ensemble of deep independent trees. Input of GBRT consists of Data $\{(x_i, y_i)\}_{i=1}^{n}$ and a differentiable Loss Function $L(y_i, F(x))$. The first step is to build a tree with only one leaf which is the mean value of all predictions (Eq. 4) (Yang et al. 2018):

$$F(x)_0 = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{5}$$

Second step is to estimate the pseudo-residuals for every sample

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for i=1,…,n} \tag{6}$$

where M is the maximum number of tree. After estimating the pseudo-residuals, a regression tree to the r_{im} is fitted and terminal regions R_{jm} for j=1,…,J_m is established. The output value minimizing the RSS is calculated for each leaf.

$$\gamma_{jm} = argmin \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma) \tag{7}$$

As the last step, new predictions are made by renewing the predictions based on the learning rate ($lr \in (0,1)$) (Yang et al. 2018).

$$F_m(x) = F_{m-1}(x) + lr \times \sum_{j=1}^{J_m} \gamma_m I(x \in R_{jm}) \tag{8}$$

There are three significant parameters to build up a GBRT model, namely the number of trees (K), the learning rate or shrinkage (R), and the maximum depth of interactions (D). The optimum performance of the model may be obtained by choosing the best compound of these parameters. (Zhang and Haghani 2015). While the number of trees is tantamount to the number of iterations for minimizing the future risks associated with prediction, the learning rate specifies the influence of each tree on the result. The maximum depth of variable interactions indicates how complex the tree is. Reducing the learning rate and utilizing a larger number of trees may improve the generalizability capability of the established model and refrain from over-fitting. In this study, a grid search method is utilized to calibrate the parameters. To achieve the best combination, three learning rates (0.0005, 0.001, 0.01), and three tree complexities (1, 3, 5) are tested with the number of trees from 1 to 10000 for the daily model. For the hourly model, due to the high number of data points, four learning rates (0.0005, 0.001, 0.01, 0.1) and three tree complexities (1, 3, 5) are evaluated with the number of trees from 1 to 40000 to find the best grid structure. 10-fold cross-validation is applied to provide performance robustness for both models.

In addition to accurate prediction capability, GBRT models can also present the variable's importance in two different ways, namely relative influence (impurity-based) and permutation-based variable importance. Impurity-based importance may dedicate higher importance to variables that are not predictive on unseen data, in cases where the established model is overfitting. On the contrary, permutation-based importance does not have this kind of issue as it is computed on unseen data.

The idea behind assessing the relative importance of a predictor is that the more frequently it is used to split a tree, the more significant it becomes. In this particular scenario, when considering a single decision tree "*T*", the equation below is used to measure the relative importance of the predictor x_i (Breiman et al. 1984).

$$I_i^2(T) = \sum_j^{J-1} \tau_j^2 I(v(j) = i) \tag{9}$$

where j stands for the j^{th} splitting node and J is the number of nodes in tree *T*. x_i is the splitting predictor related to the node *j*, whereas τ^2_j symbolizes the corresponding improvement in the model performance following the splitting at the j^{th} node.

The final predictor importance is calculated by averaging the predictor importance of all individual trees unitedly following the estimating the predictor importance of all trees in an ensemble tree model involving *K* trees (Breiman et al. 1984):

$$I_i^2(K) = \frac{1}{K} \sum_{k=1}^{K} I_i^2(T_k) \tag{10}$$

In addition to the relative importance, the final permutation-based importance of predictors can be calculated as (Breiman et al. 1984):

$$I_i = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \tag{11}$$

where s is the reference score of the model m on data D; j is the feature (column of D); *K* is the number of trees and $s_{k,j}$ is the score of the model m on corrupted data $\dot{D}_{k,j}$. The structure of GBRT model is provided in Figure 4.

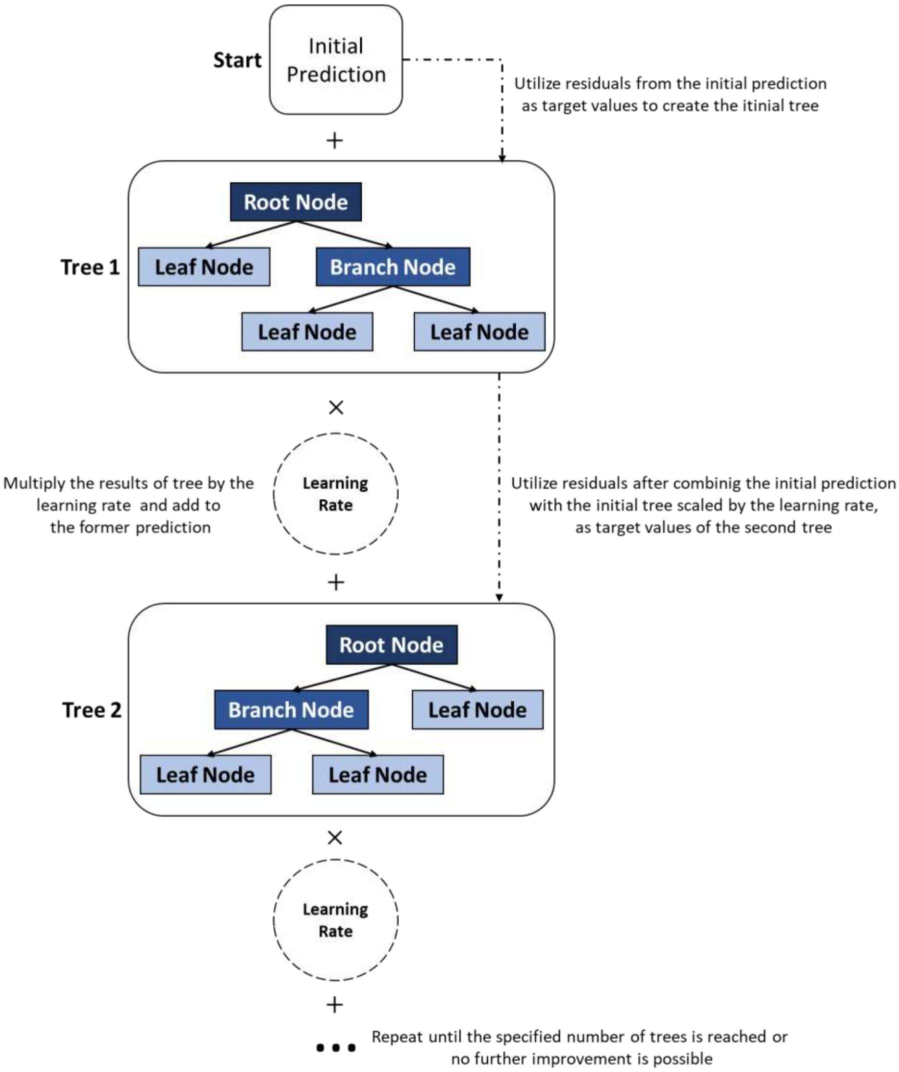
Start
Initial Prediction
Utilize residuals from the initial prediction as target values to create the itinial tree
+
Tree 1
Root Node
Leaf Node
Branch Node
Leaf Node
Leaf Node
×
Multiply the results of tree by the learning rate and add to the former prediction
Learning Rate
Utilize residuals after combinig the initial prediction with the initial tree scaled by the learning rate, as target values of the second tree
+
Tree 2
Root Node
Branch Node
Leaf Node
Leaf Node
Leaf Node
×
Learning Rate
+
• • •
Repeat until the specified number of trees is reached or no further improvement is possible

Fig. 4 Structure of GBRT model

The GBRT model can also provide the partial influence of factors on SSR. The partial dependence plot (PDP) presents the marginal impact that one or two factors have on the prediction output of a machine learning model (Friedman 2001). A PDP may exhibit whether the relationship between the dependent variable and a factor is linear, monotonic, or more complex. Most of the machine learning techniques are black-box algorithm, that do not provide an equation. However, PDPs are very useful tools to present the relations between predictors and dependent variables, as aforementioned.

For a training data involving N samples and p predictors, the partial dependence function of the jth predictor is calculated by (Friedman 2001)

$$\emptyset_j(x) = \frac{1}{N}\sum_{k=1}^{N} f(x_{1,k}, \ldots, x_{j-1,k}, x, x_{j+1,k}, \ldots, x_{p,k}) \tag{12}$$

where f(.) represents the approximation function of the tree ensemble model.

3.2. Performance metrics

As mentioned in the Introduction section, the main purpose of this study is three-fold. The first purpose is to examine the forecast accuracies of tree-based models. The accuracy can be computed using a number of methods. In this study, the accuracy of models is evaluated with four different error measurements, including Mean Bias Error (MBE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and R-squared (R^2).

MBE is the arithmetic mean of the errors between actual value and predicted value. It is estimated as

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{13}$$

where y_i is the predicted value and $\hat{y}_i$the actual value. This method does take into account the direction of errors. Values close to 0 are considered optimal, negative values indicate underestimation, and positive values indicate overestimation. Additionally, all individual differences have the same weight, hence, it does not penalise the larger differences. Since RMSE is a quadratic scoring method which is the square root of the average of squared differences between predicted value and actual value, it penalises large error. This denotes that RMSE is more functional in case large errors are particularly undesired. The equation of RMSE is given as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

MSE measures the magnitude of the average of the squares of the errors, and its formulation is given in Eq. 15.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

Finally, R-squared (R^2), also known as the coefficient of determination, is a statistical measure used to evaluate the goodness-of-fit of a model. It indicates the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. R-squared is a value between 0 and 1. A higher R-squared value indicates a better fit of the model to the data. The equation of R^2 is given as:

$$R^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares} = 1 - \frac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2} \tag{16}$$

where $\bar{y}$ is the mean of the observed data.

Based on these error calculation methods, the model with the lowest out-of-bag RMSE will be utilized for forecasting. However, the data is divided into two parts, namely, the train dataset and the test dataset. Train and test datasets for hourly data consist of values of 35040 (01.01.2016-31.12.2019) and 8760 (01.01.2020-31.12.2020) hours, respectively. On the other hand, train and test datasets for daily estimation comprise 1460 (01.01.2016-31.12.2019) and 365 (01.01.2020-31.12.2020) days, respectively. The established models are trained with the train dataset, and their errors are calculated based on the test dataset (out-of-bag error). In this regard, the flowchart of the proposed models utilized in this paper is shown in Figure 5.

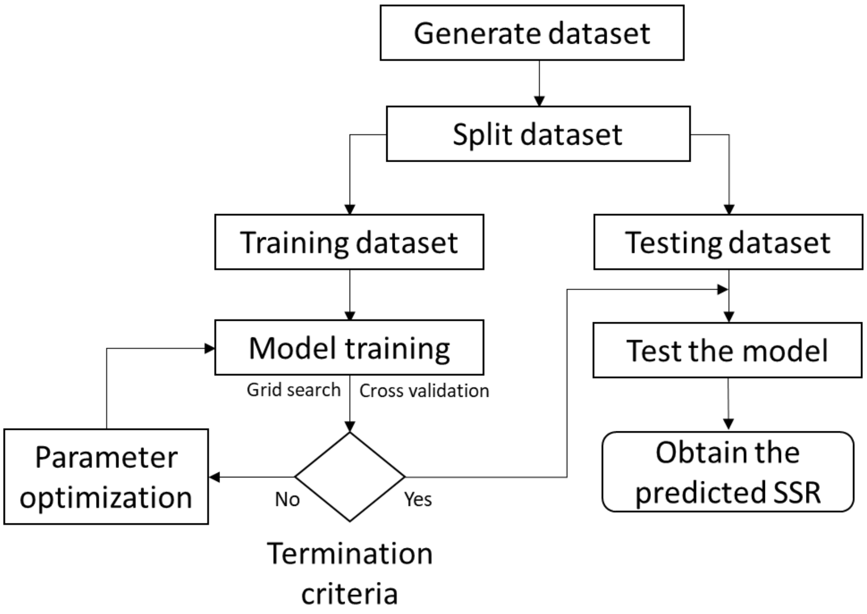

Fig. 5 Flowchart of the tree-based machine learning methods

4. Applications Results and Discussion

4.1. Application Results

The out-of-bag error comparison of established models are given in Table 5. As seen from Table 5, the method with lowest out-of-bag error rate is found to be Gradient Boosted Regression Tree for both daily and hourly estimations.

Method	**MBE**		**MSE**		**RMSE**		**R^2**	
	Hourly	**Daily**	**Hourly**	**Daily**	**Hourly**	**Daily**	**Hourly**	**Daily**
Decision Tree	8.23	6.00	5912.31	2471.60	76.89	49.72	0.91	0.71
Random Forest	7.70	5.99	3913.92	1715.35	62.56	41.08	0.94	0.80
Gradient Boosted Reg. Tree	**7.47**	**5.78**	**3671.50**	**1599.79**	**60.59**	**39.99**	**0.95**	**0.81**

Table 5. Out-of-bag error comparison of estimations

Considering the different cloud coverage percentages, two examples of the diurnal variations of the values obtained from established tree-based models compared to the real data of ERA5 are shown in Figure 6. As seen from Figure 6, all ML methods perform very well in cases of low/partially cloud coverage. Among the three ML methods, DT has the lowest ability to predict the peak points of SSR. For the case of high cloud coverage, the performance of ML methods (especially DT) decreases compared to the low/partially cloud coverage case. On the other hand, the GBRT model still performs considerably well under high variability.

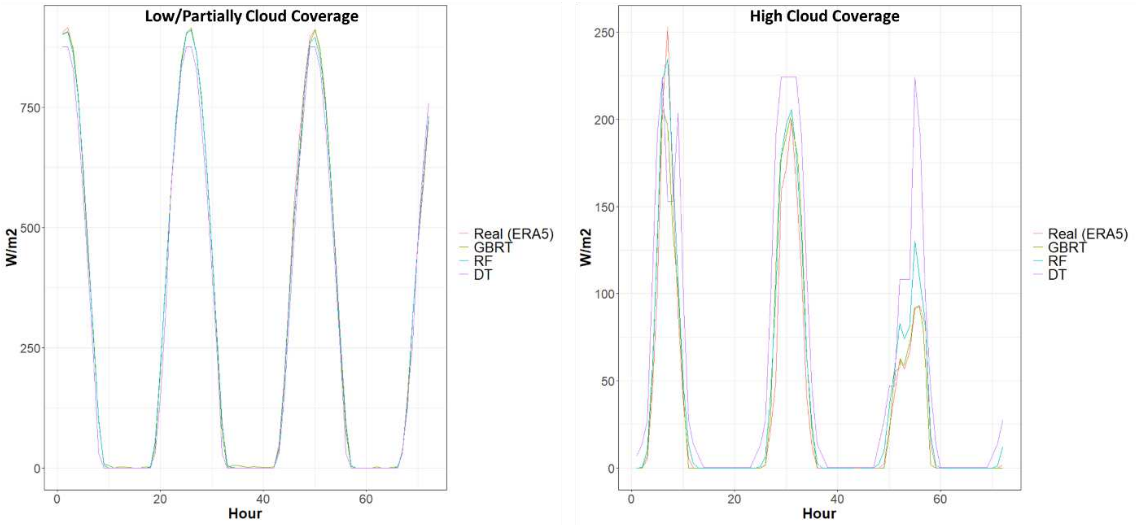

Fig. 6 Hourly model output comparison with real data

In Figure 7, the performance examples of the established ML models are given for different seasons. Similar to the low/partially cloud coverage case, all ML models show outstanding performance for spring and summer. While the output of the GBRT model is the best among all tree-based ML methods, the DT is the weakest method to reflect the variability of the SSR. Considering the fall and winter seasons, it is obvious that the performances of all methods are lower compared to their performances in the spring and summer. On the other hand, it can be revealed that DT overestimates the SSR, and the GBRT model is capable of catching the variability of the SSR. As a result, the GBRT model performs quite better than the RF and DT models.

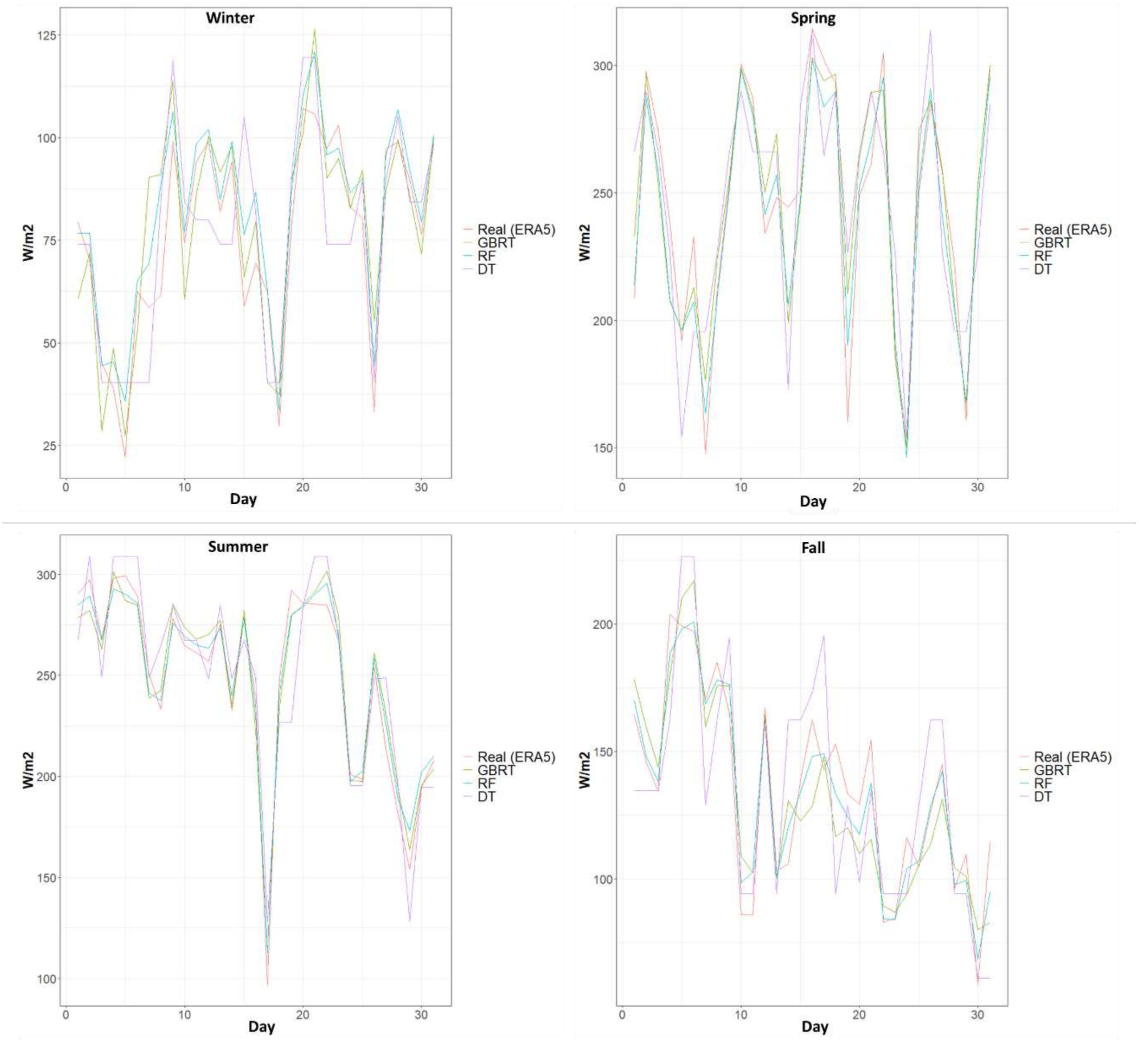

Fig. 7 Daily model output comparison with real data

4.1.1. Decision Tree Model

The second aim of this study is to investigate the determinants of Downward Surface Solar Radiation. As mentioned in the Methodology section, the Decision Tree method is very useful for explaining the model output in graphical form. To run the DT model, the “rpart” package of R software is utilized. The DT model for daily data is given in Figure 8a. In the case of a temperature smaller than 293K, precipitation larger than 0.1m, a solar zenith angle larger than 105°, and cloud coverage larger than 80%, the daily average SSR is found to be at the lowest level with a value of 40 W/m^2. On the contrary, the daily average SSR is at the highest point with a value of 298 W/m^2 when the temperature is higher than 293K, the ozone level is higher than 316 Dobson, and total precipitation is less than 0.01 mm. The other way around, SSR is at its highest point with a value of 790 W/m^2 when cloud coverage is less than 30% and the hour is between 10:00 and 13:00

(See Fig. 8b). Moreover, when the decision tree nodes are examined, the time of day (hour) is detected as the most important factor in determining the downward surface solar radiation.

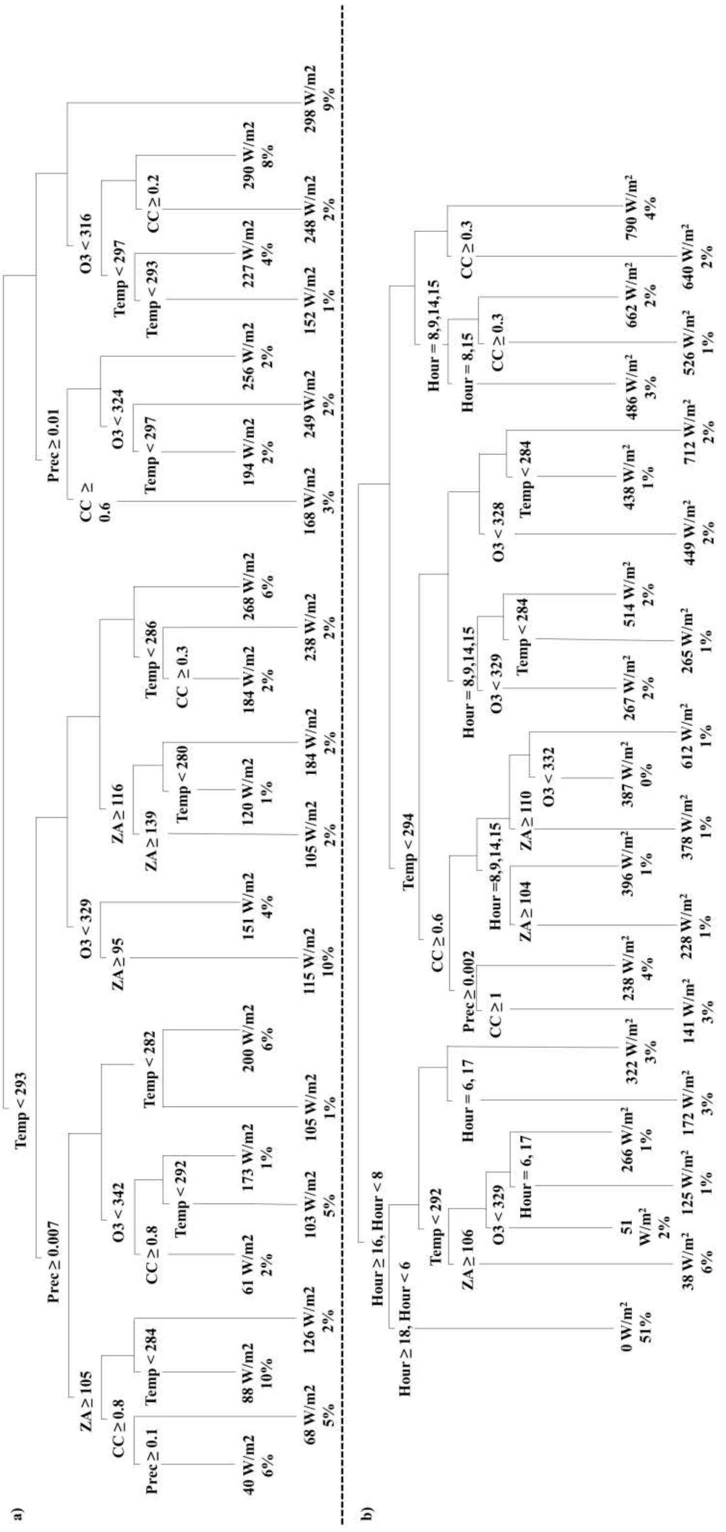

Fig. 8 Decision Tree model for a) daily b) hourly estimation

4.1.2. Random Forest Model

To execute the daily and hourly Random Forest models, the “randomForest” package of R software is used. As mentioned in the Methodology section, the argument K (*mtry*) implies how many predictors are to be taken into consideration for each split of the tree. In addition to the *mtry* argument, the number of grown trees (*ntree*) is another argument that affects the performance of the model. Figure 9 shows the RMSE comparison of different configurations of *ntree* and *mtry* arguments. The configurations with the lowest RMSE for both daily and hourly models are chosen for determining the variance importance. The model with mtry= 7 and ntree= 1750 provides the lowest error rate for the daily model, whereas the model with mtry= 8 and ntree= 750 performs the best for the hourly model.

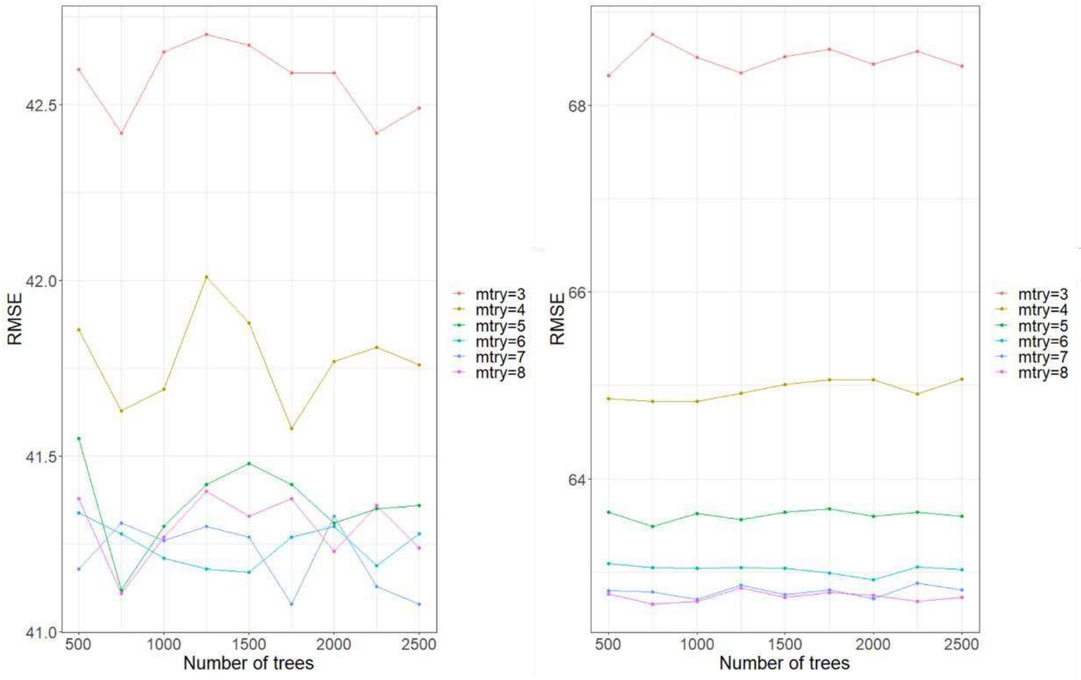

Fig. 9 RMSE comparison of different models for daily and hourly estimations

A graphical portrait of the relative variable importance of the daily and hourly data is shown in Figure 10. Temperature is found to be the most important factor to estimate the daily SSR followed by zenith angle and cloud coverage. Temperature accounts for 29.2% of the overall variable importance. Contrariwise, aerosols (SO_2, PM_{10}, $PM_{2.5}$) and eastward wind speed (u10) have minimal impact on determining the SSR, with values of 1.1, 1.0, 0.9, and 0.9 percent, respectively. When it comes to the hourly model, the time of day (hour) becomes the most important factor, with a value of 68.1 percent. Similar to the daily model, temperature is detected as the most important atmospheric variable for the hourly model, while eastward wind speed (u10) is determined as the least important one.

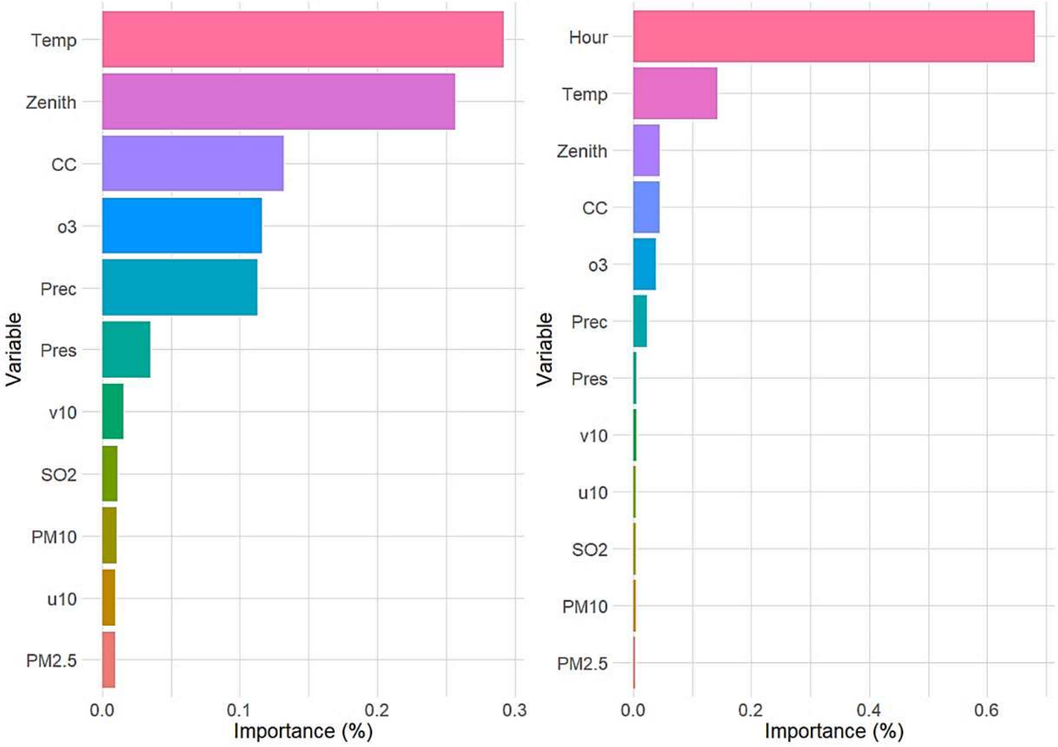

Fig. 10 Relative importance of a) daily and b) hourly RF model

4.1.3. Gradient Boosted Regression Tree Model

As mentioned in the Methodology section, the GBRT model should be tuned to achieve the desired error rates. In this context, the best hypergrid structures used in the Tuned-GBRT resulted from minimizing RMSE are given both for the daily and hourly models in Table 6 and 7, respectively. The best performance for daily model is attained with the learning rate parameter of 0.0005, the subsample parameter of 0.25, the interaction depth parameter of 5, and the number of trees parameter of 32241. Moreover, for the hourly model, the lowest error is achieved when the learning rate is 0.01, the interaction depth parameter is 5, the subsample parameter is 0.85, and the number of trees parameter is 39998.

Shrinkage	Interaction depth	n.minobsinnode	bag.fraction	optimal_trees	min_RMSE
0.0005	5	5	0.25	32241	30.621
0.0005	5	7	0.25	39987	30.624
0.0005	5	3	0.5	40000	30.679
0.001	5	7	0.25	16437	30.768
0.01	5	5	0.25	1591	30.801

Table 6. Best hyper grid structures for daily model

Shrinkage	Interaction depth	n.minobsinnode	bag.fraction	optimal_trees	min_RMSE
0.01	5	3	0.85	39998	40.280
0.01	5	7	0.85	39998	40.351
0.01	5	3	0.85	39964	40.366
0.01	5	5	0.85	40000	40.390
0.01	5	7	0.85	39996	40.415

Table 7. Best hyper grid structures for hourly model

When the main determinants of daily SSR resulting from the RF model are compared with those of the GBRT model (See Fig. 11a), temperature is obviously more important in the GBRT model. At the same time, ozone level, zenith angle, precipitation, and cloud coverage are found to be the most important factors in determining the SSR for the Istanbul region, whereas PM_{10} and SO_2 are detected as the least important ones.

The permutation-based importance provides better performance with the unseen data, and hence, it gives more precise results. Although temperature is determined to be the most important factor in determining the SSR by both importance calculation methods, the importance of ozone level is higher than that of cloud coverage and zenith angle in the permutation-based importance method.

Similar to the RF model, the most significant variable in determining the hourly SSR is the time of day (hour) for both permutation-based and relative importance methods (See Figure 11b). Following the hour variable, temperature is detected as the second-most important factor. Notwithstanding that the importance ranks of aerosols are the same in both variable importance methods, ozone level is observed to be more important than cloud coverage and zenith angle in the permutation-based method.

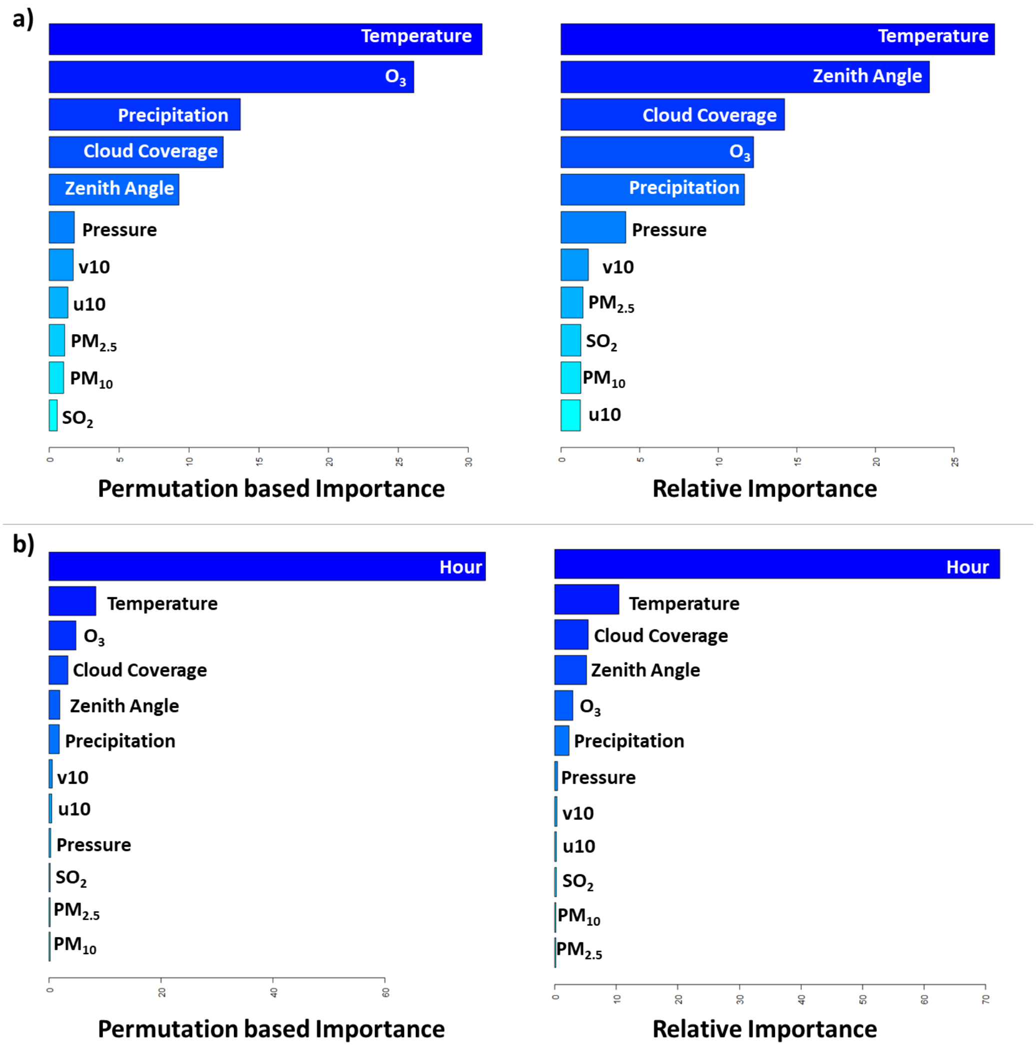

Fig. 11 Importance plots of a) daily and b) hourly GBRT models

In furtherance, the partial dependence plots of factors are given in Figure 12-15. It can be seen that there is a positive relationship between hourly and daily downward surface solar radiation and temperature. On the other hand, there is a negative relationship between precipitation, cloud coverage, and downward surface solar radiation, as expected. After the temperature exceeds 292K, daily downward surface solar radiation starts to increase sharply. The positive impact of ozone levels on hourly SSR continues until the level of 400 Dobson, at which point it starts to have a negative effect. Although the influence of zenith angle on SSR is almost stable between 0 and 90 degrees, SSR sharply decreases after the solar zenith angle exceeds 90 degrees. The impact of westward winds is almost negligible, whereas, eastward winds have a positive impact on SSR. On the contrary, when the northward wind speed increases, hourly SSR decreases almost linearly.

Although the impacts of the aerosols are relatively smaller than the meteorological parameters, it can be said that when $PM_{2.5}$ increases until the level of 55 µg/cm^3, it has a dimmering effect on the calculated SSR. Furthermore, PM10 has a positive relationship with SSR until it reaches 50 µg/cm^3. After this level, the impact of PM_{10} is almost neglectable.

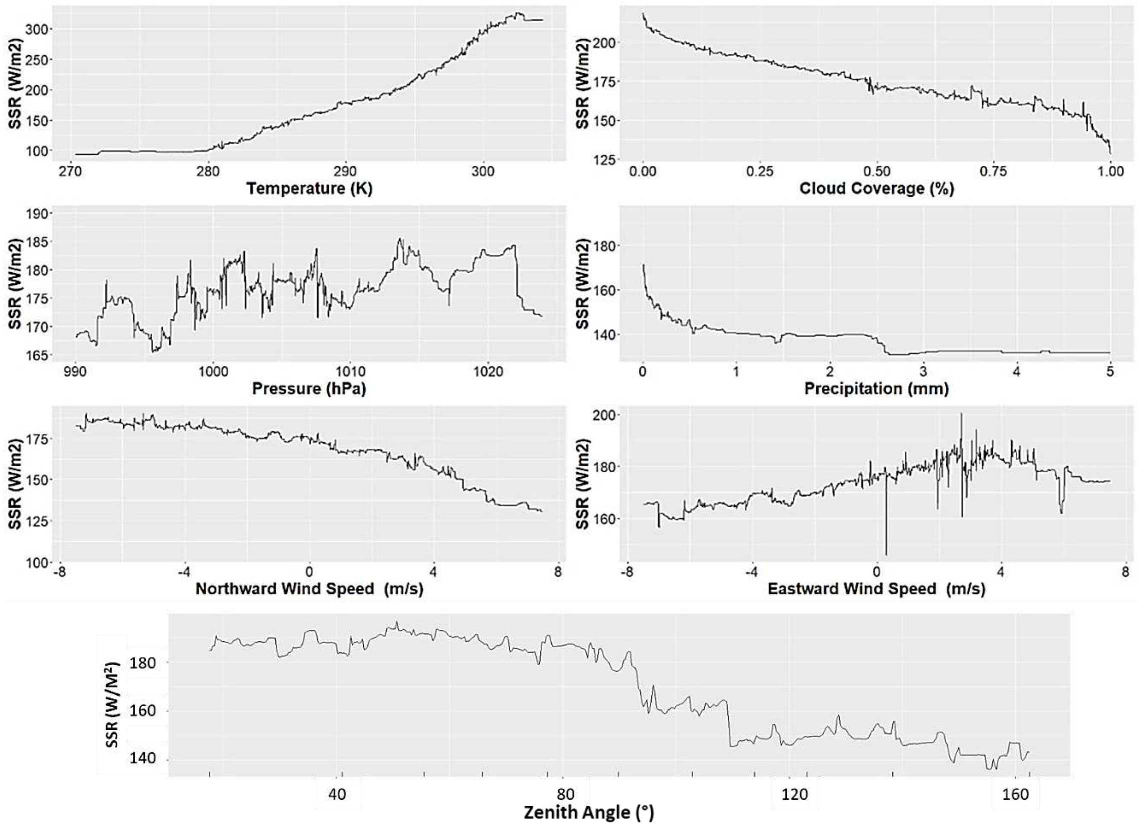

Fig. 12 Partial dependence plots of meteorological variables (hourly)

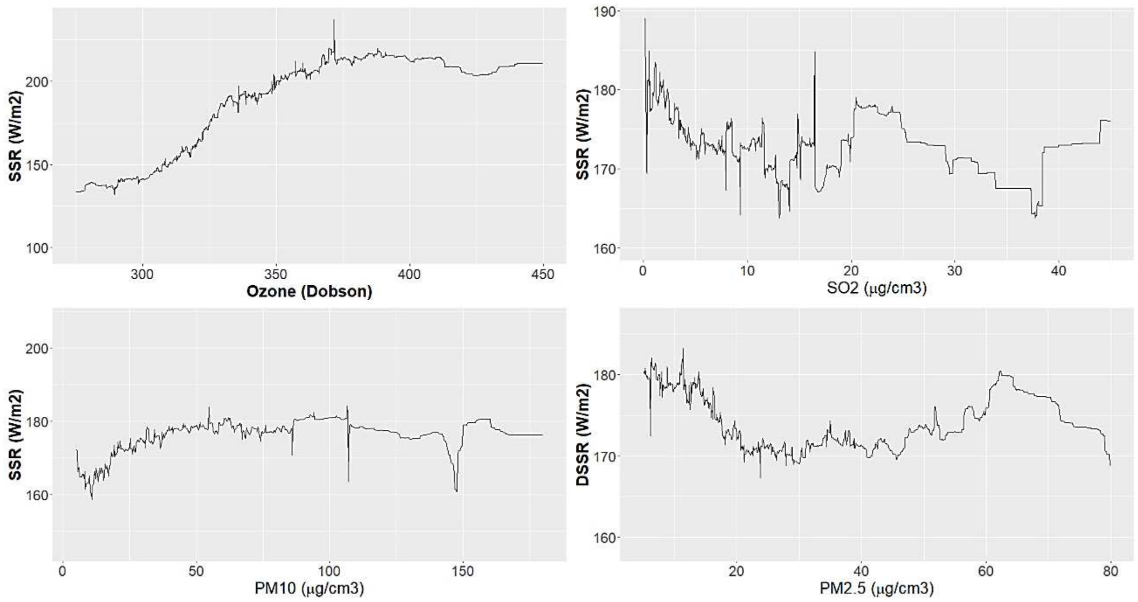

Fig. 13 Partial dependence plots of aerosols (hourly)

Similar to the hourly PDP, there is a negative relationship between northward wind speed and the daily SSR. However, there is some local peaks for the eastward wind speed (See Fig. 14). The negative impacts of $PM_{2.5}$ and SO_2 on daily SSR continue until they reach levels of 35 and 12.5 μg/cm^3, respectively, at which point they start to have a positive effect. Surprisingly, PM_{10} has an additive effect on the daily SSR calculation, while it is practically ineffectual on determining hourly SSR.

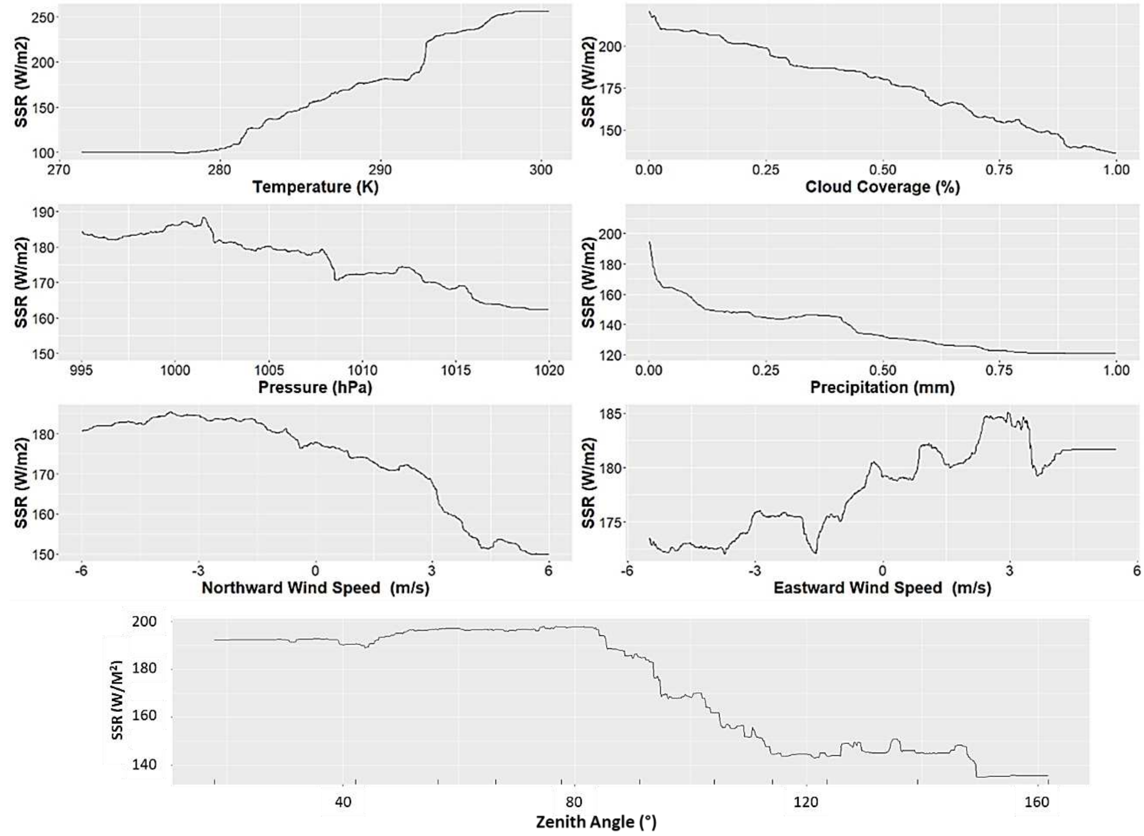

Fig. 14 Partial dependence plots of meteorological variables (daily)

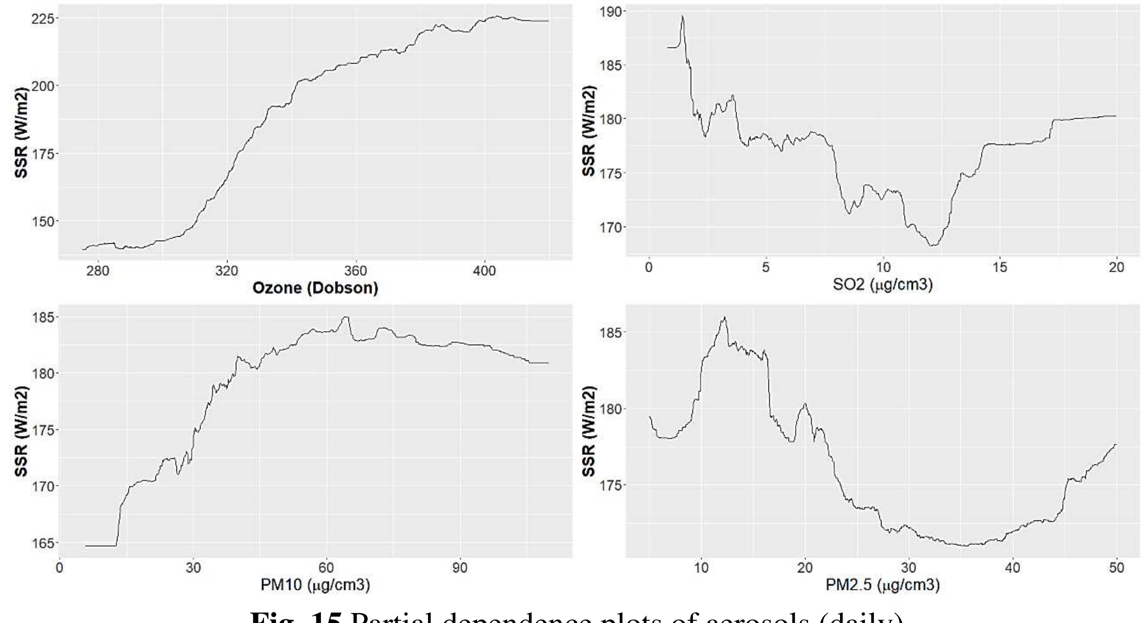

Fig. 15 Partial dependence plots of aerosols (daily)

Another output of this analysis is the interaction plots which are very useful method to show relationships between two predictors and a dependent variable. In the interaction plots, the predictors are examined two at a time. An x–y grid is set up based on the possible combinations of predictors in the range of both variables. The rest of the predictors are held at their means. Model predictions (SSR, in this case) are presented on the z axis. The interaction plots of the most important variables are given in Figure 16 and 17 for the hourly and daily models, respectively.

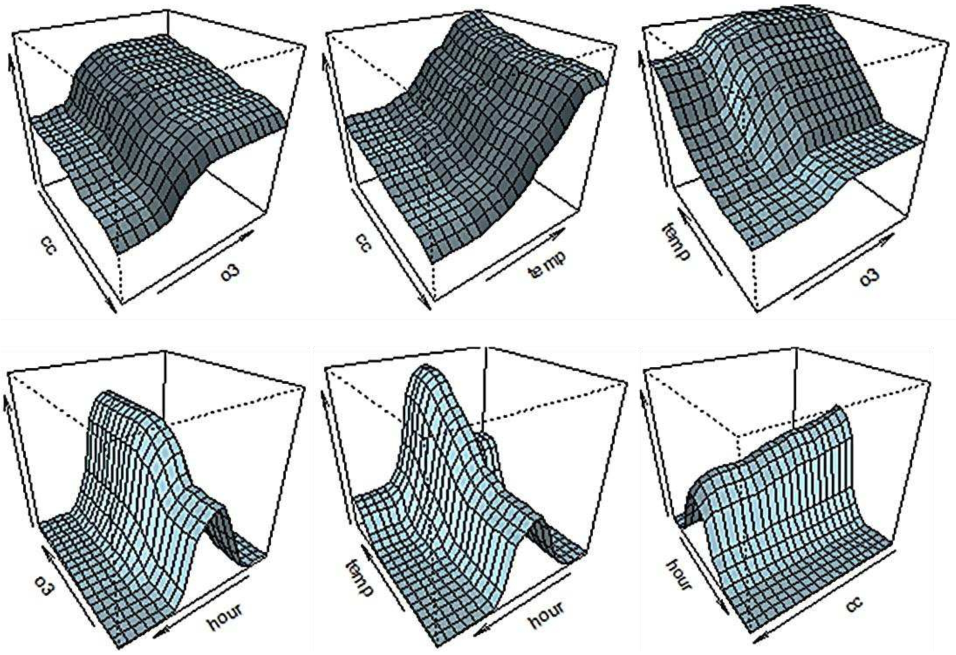

Fig. 16 Interaction plots of most important variables (hourly)

From Fig. 16, it can be revealed that hourly SSR reaches its maximum value when the cloud coverage is at its minimum and the ozone level is at its maximum. Furthermore, SSR shows a rapid increase when the ozone level exceeds a certain level. Different from cloud coverage-ozone level combination, which has an S-shaped graphic, the combination of cloud coverage and temperature shows a more linear profile. The hourly SSR regarding to temperature and ozone level reaches its maximum level when both temperature and ozone level are at their highest points. The combination of increasing temperature and ozone level creates an S-shaped increase in SSR levels, similar to the combination of cloud coverage and ozone level.

When the interaction plot of daily average SSR is examined, it can be said that the combination of cloud coverage and temperature almost has a linear relationship with daily average SSR. Similar to the hourly plot, SSR reaches its maximum value when the cloud coverage is at its minimum and the temperature is at its maximum. However, daily average SSR has a sharper increasing trend regarding the temperature compared to hourly SSR. Daily and hourly plots of cloud coverage and ozone level show the same S-shaped profile. On the other hand, the impact of precipitation with the combination of temperature, cloud coverage, and ozone level on daily SSR can be seen better in Figure 17. When it is rainy, SSR decreases suddenly, and then it follows a linear trend.

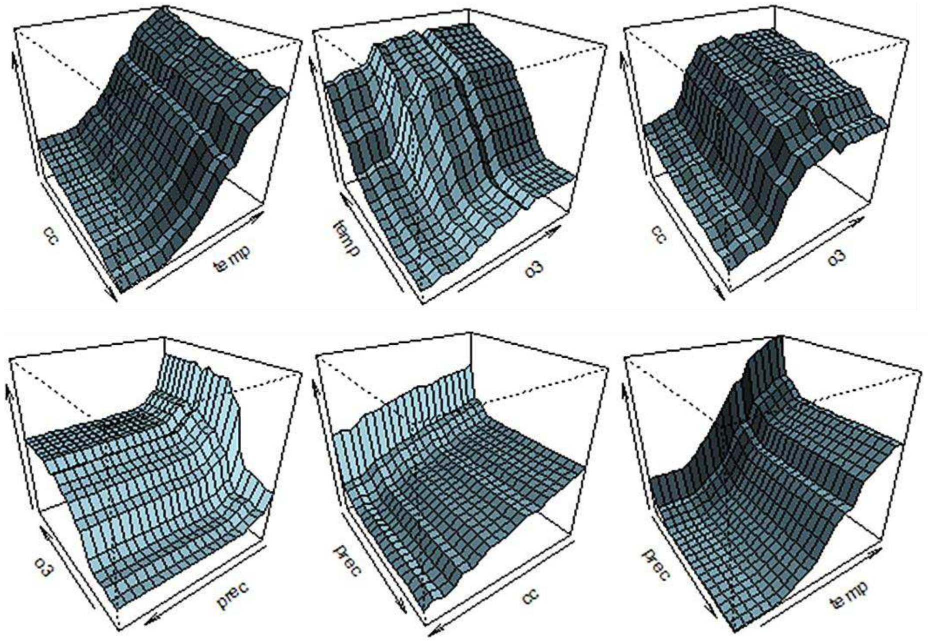

Fig. 17 Interaction plots of most important variables (daily)

To validate the capability of established model, it is applied to another location which has a relatively similar climate to Istanbul. Thus, the daily and hourly models, which are established with data gathered from Istanbul, are utilized for Antalya, Türkiye. While comparing two different datasets, it is needed to use normalized performance indicators. Thus, nMBE (%) and nRMSE (%) are used to compare model performances for both locations. These performance criteria can be calculated by dividing Eq. 13 and Eq. 14 by the mean of the SSR of each location. The performances of both daily and hourly models are given in Table 8. These results indicate that both daily and hourly GBRT models perform well with the datasets gathered from different locations.

Model	Location	nMBE (%)	nRMSE (%)
Daily	Istanbul	3.3%	22.8%
	Antalya	3.9%	23.8%
Hourly	Istanbul	4.3%	34.6%
	Antalya	4.1%	34.2%

Table 8. Performances of daily and hourly models for Istanbul and Antalya

4.2. Discussion

As a result of all these models, it can be revealed that the most important atmospheric determinant in calculating both daily and hourly SSR for Istanbul is temperature. Based on the RF model, temperature has a more than 29 percent impact on the daily SSR, whereas aerosols and eastward wind speed have a minimal impact on determining the daily SSR. The importance of temperature is detected as more than 30 percent in the GBRT model; and aerosols and eastward wind speed are found to be the least important factors, similar to the RF model.

When it comes to the hourly model, in both the GBRT and RF models, the time of day (hour) comes to the forefront, constituting more than 68 percent and 70 percent of the total importance in determining the hourly SSR, respectively. Moreover, temperature and ozone level are the most important atmospheric factors, similar as in the daily model. Although the impacts of the zenith angle and precipitation on hourly SSR are limited, the importance of aerosols, wind speed components, and pressure are almost nugatory.

In light of this information, to forecast daily and hourly SSR levels of Istanbul, the tuned-GBRT model, whose optimized hypergrid structures are given in Table 6 and 7, is the best tree-based machine learning method with lower error rates. The RMSE score of this study is less than 40 W/m^2 for daily resolution, implying the established algorithm had a relatively trustworthy estimation performance (Tang et al. 2016).

Based on the results of this study, it can be said that tree-based machine learning techniques are powerful tools to determine and analyse downward surface solar radiation. With the transition towards cleaner and greener energy production, the importance of renewable energy technologies, such as solar systems, has been increasing gradually. However, the output power of solar systems is closely related to the shortwave solar radiation at the surface. To benefit from photovoltaic systems (PV) at the maximum level, analysing the determinants of SSR is crucial. Since SSR is extremely variable because of the meteorological patterns and particles in the air, power production from PV systems is not stable, dissimilar to dispatchable energy sources like thermal power plants. Hence, determining the most important variables for the SSR and presenting how and how much these variables impact the SSR, which are the main purposes of this study, can provide a basis to maintain an efficient and profitable operation for PV systems.

5. Conclusion

This paper aims to estimate Surface Solar Radiation in Istanbul using tree-based machine learning methods, including Decision Tree, Random Forest and Gradient Boosted Regression Tree, while determining the importance of variables on SSR. In this context, the accuracies of models are evaluated with out-of-bag errors by calculating MSE, RMSE, MBE, and R-squared. Based on the error rates, the Gradient Boosted Regression Tree model provides the best performance with the lowest RMSE for both daily and hourly models. Considering the MBE, the biases of the best models in W/m^2 are calculated as 5.78 W/m^2 for daily and 7.47 W/m^2 for hourly data. Furthermore, to validate the established models, the best models are run for another location. The results indicate that both daily and hourly GBRT models perform well with the datasets gathered from different locations.

In addition to model accuracies, tree-based machine learning methods are powerful tools to present the variable importance. Although all models provide different values for the variable importance; temperature, ozone level, cloud coverage, and zenith angle are found to be the most important variables in estimating daily SSR. For the hourly estimation, the time of day (hour) becomes the most important factor in addition to temperature, ozone level, and cloud coverage. On the other hand, the impacts of the wind speed components and aerosols are almost neglectable. Finally, this study shows that the tree-based machine learning methods used with these variables to estimate hourly and daily SSR results are very accurate when it is not possible to measure the SSR values directly.

Author Declarations

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The author declares that he/she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors Contributions

Material preparation, data collection, analysis, the first draft and final manuscript were performed by DG.

Availability of data and materials

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability

The codes generated during the current study are available from the corresponding author on reasonable request.

Ethical Approval

This material is the author's own original work, which has not been previously published elsewhere. The paper is not currently being considered for publication elsewhere.

Consent to Participate

Not applicable

Consent to Publish

Not applicable

References

AWG Radiation Budget Application Team (2018) GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document for Downward Shortwave Radiation (Surface), and Reflected Shortwave Radiation (TOA), NOAA NESDIS Center for Satellite Applications and Research. *NOAA NESDIS CENTER for SATELLITE APPLICATIONS and RESEARCH.*

Basílio SDCA., Putti FF, Cunha AC, Goliatt L (2023) An evolutionary-assisted machine learning model for global solar radiation prediction in Minas Gerais region, southeastern Brazil. Earth Science Informatics: 1-19. https://doi.org/10.1007/s12145-023-00990-0

Bhattacharjee AD, Chowdhury AR (2022) Short-Term Solar Irradiance Fore-casting Using Long Short Term Memory Variants. In Proceedings of International Con-ference on Data Science and Applications (pp. 227-243). Springer, Singapore.

Breiman L (2001) Random forests. Machine learning 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees (1st ed.). CRC Press. https://doi.org/10.1201/9781315139470

Chen F, Zhou Z, Lin A, Niu J, Qin W, Yang Z (2019) Evaluation of direct horizontal irradiance in China using a physically-based model and machine learning methods. Energies 12(1): 150. https://doi.org/10.3390/en12010150

Chen J, Zhu W, Yu Q (2021) Estimating half-hourly solar radiation over the Continental United States using GOES-16 data with iterative random forest. Renewable Energy 178: 916-929. https://doi.org/10.1016/j.renene.2021.06.129

Chen Y, Bai M, Zhang Y, Liu J, Yu D (2023) Error revision during morning period for deep learning and multi-variable historical data-based day-ahead solar irradiance forecast: towards a more accurate daytime forecast. Earth Science Informatics: 1-23. https://doi.org/10.1007/s12145-023-01026-3

Copernicus Climate Change Service (C3S) (2021) ERA5: fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS).

Deo RC, Şahin M (2017) Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. Renewable and Sustainable Energy Reviews 72: 828-848. https://doi.org/10.1016/j.rser.2017.01.114

EMBER (2022) Global Electricity Review 2022. https://ember-climate.org/app/uploads/2022/03/Report-GER22.pdf

Fan J, Wu L, Ma X, Zhou H, Zhang F (2020) Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. Renewable Energy 145: 2034-2045. https://doi.org/10.1016/j.renene.2019.07.104

Feng Y, Li Y (2018) Estimated spatiotemporal variability of total, direct and diffuse solar radiation across China during 1958–2016. International Journal of Climatology 38(12): 4395-4404. https://doi.org/10.1002/joc.5676

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Annals of statistics, 1189-1232.

Gareth J, Daniela W, Trevor H, Robert T (2013) An introduction to statistical learning: with applications in R. Spinger.

Ghimire S, Deo RC, Raj N, Mi J (2019) Wavelet-based 3-phase hybrid SVR model trained with satellite-derived predictors, particle swarm optimization and maximum overlap discrete wavelet transform for solar radiation prediction. Renewable and Sustainable Energy Reviews, 113: 109247. https://doi.org/10.1016/j.rser.2019.109247

Gürel AE, Ağbulut Ü, Biçen Y (2020) Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation. Journal of Cleaner Production 277: 122353. https://doi.org/10.1016/j.jclepro.2020.122353

Hai T Sharafati A, Mohammed A, Salih SQ, Deo RC, Al-Ansari N, Yaseen ZM (2020) Global solar radiation estimation and climatic variability analysis using extreme learning machine based predictive model. IEEE Access 8: 2026-12042. Doi:10.1109/ACCESS.2020.2965303

Hartmann DL, Tank AMK, Rusticucci M, Alexander LV, Brönnimann S, Charabi YAR, ..., Soden BJ (2013) Observations: atmosphere and surface. In Climate change 2013 the physical science basis: Working group I contribution to the fifth assessment report of the intergovernmental panel on climate change (pp. 159-254). Cambridge University Press.

Hocaoğlu FO, Gerek ÖN, Kurban M (2008) Hourly solar radiation forecasting using optimal coefficient 2-D linear filters and feed-forward neural networks. Solar energy 82(8): 714-726. https://doi.org/10.1016/j.solener.2008.02.003

Hou N, Zhang X, Zhang W, Wei Y, Jia K, Yao Y, ..., Cheng J (2020) Estimation of Surface Downward Shortwave Radiation over China from Himawari-8 AHI Data Based on Random Forest. Remote Sensing 12(1): 181. https://doi.org/10.3390/rs12010181

Jiang Y (2008) Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. Energy policy 36(10): 3833-3837. https://doi.org/10.1016/j.enpol.2008.06.030

Jiang B, Liang S, Ma H, Zhang X, Xiao Z, Zhao X, ..., Jia A (2016) GLASS daytime all-wave net radiation product: Algorithm development and preliminary validation. Remote Sensing 8(3): 222. https://doi.org/10.3390/rs8030222

Kisi O, Heddam S, Yaseen ZM (2019) The implementation of univariable scheme-based air temperature for solar radiation prediction: New development of dynamic evolving neural-fuzzy inference system model. Applied Energy 241: 184-195. https://doi.org/10.1016/j.apenergy.2019.03.089

Lam JC, Wan KK, Yang L (2008) Solar radiation modelling using ANNs for different climates in China. Energy Conversion and Management 49(5): 1080-1090. https://doi.org/10.1016/j.enconman.2007.09.021

Lima FJ, Martins FR, Pereira EB, Lorenz E, Heinemann D (2016) Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks. Renewable Energy 87: 807-818. https://doi.org/10.1016/j.renene.2015.11.005

Luiz EW, Martins FR, Gonçalves AR, Pereira EB (2018). Analysis of intra-day solar irradiance variability in different Brazilian climate zones. Solar Energy 167: 210-219. https://doi.org/10.1016/j.solener.2018.04.005

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, ... Lee SI (2020) From local explanations to global understanding with explainable AI for trees. Nature machine intelligence 2(1): 56-67. https://doi.org/10.1038/s42256-019-0138-9

Martins FR, Pereira EB, Guarnieri RA (2012) Solar radiation forecast using artificial neural networks. International Journal of Energy Science 2(6).

Mellit A, Eleuch H, Benghanem M, Elaoun C, Pavan AM (2010) An adaptive model for predicting of global, direct and diffuse hourly solar irradiance. Energy Conversion and Management 51(4): 771-782. https://doi.org/10.1016/j.enconman.2009.10.034

Ministry of Energy and Natural Resources (2023) Renewable Energy. https://enerji.gov.tr/eigm-resources-en

Ministry of Environment, Urbanization and Climate Change (2022) Air Quality Databank. https://sim.csb.gov.tr/

Mubiru J, Banda EJKB (2008) Estimation of monthly average daily global solar irradiation using artificial neural networks. Solar Energy 82(2): 181-187. https://doi.org/10.1016/j.solener.2007.06.003

Ohmura A (2009) Observed decadal variations in surface solar radiation and their causes. Journal of Geophysical Research: Atmospheres 114:(D10). https://doi.org/10.1029/2008JD011290

Rahimikhoob A, Behbahani SMR, Banihabib ME (2013) Comparative study of statistical and artificial neural network's methodologies for deriving global solar radiation from NOAA satellite images. International Journal of Climatology 33(2): 480-486. https://doi.org/10.1002/joc.3441

Ryu Y, Jiang C, Kobayashi H, Detto M (2018) MODIS-derived global land products of shortwave radiation and diffuse and total photosynthetically active radiation at 5 km resolution from 2000. Remote Sensing of Environment 204: 812-825. https://doi.org/10.1016/j.rse.2017.09.021

Qin J, Chen Z, Yang K, Liang S, Tang W (2011) Estimation of monthly-mean daily global solar radiation based on MODIS and TRMM products. Applied energy 88(7): 2480-2489. https://doi.org/10.1016/j.apenergy.2011.01.018

Qin Y, Huang J, McVicar TR, West S, Khan M, Steven AD (2021) Estimating surface solar irradiance from geostationary Himawari-8 over Australia: A physics-based method with calibration. Solar Energy 220: 119-129. https://doi.org/10.1016/j.solener.2021.03.029

Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N (2019) Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomedical Signal Processing and Control 52: 456-462. https://doi.org/10.1016/j.bspc.2017.01.012

Sharafati A, Khosravi K, Khosravinia P, Ahmed K, Salman SA, Yaseen ZM, Shahid S (2019) The potential of novel data mining models for global solar radiation prediction.

International Journal of Environmental Science and Technology 16(11): 7147-7164. https://doi.org/10.1007/s13762-019-02344-0

Sianturi Y, Sopaheluwakan A, Sartika KA (2021, November) Evaluation of ECMWF model to predict daily and monthly solar radiation over Indonesia region. In IOP Conference Series: Earth and Environmental Science (Vol. 893, No. 1, p. 012074). IOP Publishing.

Singla P, Duhan M, Saroha S (2022) Solar Irradiation Forecasting by Long-Short Term Memory Using Different Training Algorithms. In Renewable Energy Optimization, Planning and Control (pp. 81-89). Springer, Singapore.

Srivastava R, Tiwari AN, Giri VK (2019) Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India. Heliyon 5(10): e02692. https://doi.org/10.1016/j.heliyon.2019.e02692

Tang W, Qin J, Yang K, Liu S, Lu N, Niu X (2016) Retrieving high-resolution surface solar radiation with cloud parameters derived by combining MODIS and MTSAT data. Atmospheric Chemistry and Physics 16(4): 2543-2557. https://doi.org/10.5194/acp-16-2543-2016

Tymvios FS, Jacovides CP, Michaelides SC, Scouteli C (2005) Comparative study of Ångström's and artificial neural networks' methodologies in estimating global solar radiation. Solar energy 78(6): 752-762. https://doi.org/10.1016/j.solener.2004.09.007

Vakitbilir N, Hilal A, Direkoğlu C (2022) Hybrid deep learning models for multivariate forecasting of global horizontal irradiation. Neural Computing and Applications: 1-22.

Voyant C, Muselli M, Paoli C, Nivet ML (2011) Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. Energy 36(1): 348-359. https://doi.org/10.1016/j.energy.2010.10.032

Voyant C, Notton G, Kalogirou S, Nivet ML, Paoli C, Motte F, Fouilloy A (2017) Machine learning methods for solar radiation forecasting: A review. Renewable energy 105: 569-582. https://doi.org/10.1016/j.renene.2016.12.095

Wang L, Kisi O, Zounemat-Kermani M, Zhu Z, Gong W, Niu Z, ..., Liu Z (2017) Prediction of solar radiation in China using different adaptive neuro-fuzzy methods and M5 model tree. International Journal of Climatology 37(3): 1141-1155. https://doi.org/10.1002/joc.4762

Wang T, Yan G, Chen L (2012) Consistent retrieval methods to estimate land surface shortwave and longwave radiative flux components under clear-sky conditions. Remote Sensing of Environment 124: 61-71. https://doi.org/10.1016/j.rse.2012.04.026

Wei Y, Zhang X, Hou N, Zhang W, Jia K, Yao Y (2019) Estimation of surface downward shortwave radiation over China from AVHRR data based on four machine learning methods. Solar Energy 177: 32-46. https://doi.org/10.1016/j.solener.2018.11.008

Wild M (2009) Global dimming and brightening: A review. Journal of Geophysical Research: Atmospheres 114:(D10).

Wild M (2012) Enlightening global dimming and brightening. Bulletin of the American Meteorological Society 93(1): 27-37.

Wild M, Gilgen H, Roesch A, Ohmura A, Long CN, Dutton EG, ..., Tsvetkov A (2005) From dimming to brightening: Decadal changes in solar radiation at Earth's surface. Science 308(5723): 847-850. DOI: 10.1126/science.1103215

Willson RC, Mordvinov AV (2003) Secular total solar irradiance trend during solar cycles 21–23. Geophysical Research Letters 30(5). https://doi.org/10.1029/2002GL016038

Yang L, Zhang X, Liang S, Yao Y, Jia K, Jia A (2018) Estimating surface downward shortwave radiation over China based on the gradient boosting decision tree method. Remote Sensing 10(2): 185. https://doi.org/10.3390/rs10020185

Yoon J (2021) Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. Computational Economics 57(1): 247-265. doi: https://doi.org/10.1007/s10614-020-10054-w.

Zeng Z, Wang Z, Gui K, Yan X, Gao M, Luo M, ..., Yang Y (2020) Daily Global Solar Radiation in China Estimated from High-Density Meteorological Observations: A Random Forest Model Framework. Earth and Space Science 7(2): e2019EA001058. https://doi.org/10.1029/2019EA001058

Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies 58: 308-324. https://doi.org/10.1016/j.trc.2015.02.019

Zhang Y, Chen L (2022) Estimation of Daily Average Shortwave Solar Radiation under Clear-Sky Conditions by the Spatial Downscaling and Temporal Extrapolation of Satellite Products in Mountainous Areas. Remote Sensing 14(11): 2710. https://doi.org/10.3390/rs14112710

Zhou Q, Flores A, Glenn NF, Walters R, Han B (2017) A machine learning approach to estimation of downward solar radiation from satellite-derived data products: An application over a semi-arid ecosystem in the US. PLoS One 12(8): e0180239. https://doi.org/10.1371/journal.pone.0180239