Environ. Water Eng., 2022, 8(3), 636-653 DOI: 10.22034/JEWE.2022.312038.1654

Environment and Water Engineering

Homepage: www.jewe.ir

ISSN: 2476-3683

Research Paper

Evaluation of Random Forest-Genetic Algorithm Hybrid Model in Estimating Daily Solar Radiation

Sajjad Hashemi[1], Saeed Samadianfard[2*] and Ali Ashraf Sadraddini[3]

[1]PhD Scholar, Department of Water Engineering, Faculty of Agriculture, Tabriz University, Tabriz, Iran
[2]Assoc. Professor, Department of Water Engineering, Faculty of Agriculture, Tabriz University, Tabriz, Iran
[3]Professor, Department of Water Engineering, Faculty of Agriculture, Tabriz University, Tabriz, Iran

Article information

Received: October 25, 2021
Revised: January 11, 2022
Accepted: January 14, 2022

Keywords:
Ardabil
Intelligent Models
Optimization
Solar Energy

*Corresponding author**:**
s.samadian@tabrizu.ac.ir

Abstract

Solar energy is the most important source of renewable energy, in other words, the main source of energy on Earth. Therefore, estimating the solar radiation parameter with high accuracy is very important. In this regard, in the present study, meteorological data of 3 meteorological stations of Ardabil province, including Meshginshahr, Germi, and Nir for a period of 2 years (2017-2018) on a daily scale were used. Then, the intensity of daily solar radiation in each of the mentioned stations was estimated using random forest and random forest methods-genetic algorithm. The meteorological variables used included minimum, maximum and average temperature, relative humidity, and wind speed, which in eight different combinations were considered as input data in the model calculations. The obtained results were compared with each other using statistical parameters and the best models were selected. By comparing the results, the models of Nir, Meshginshahr, and Germi stations were ranked from highest to lowest modeling accuracy, respectively; So that the GA-RF-V model in Nir station with the root mean square error of 0.346 MJ/m^{2}d and Kling-Gupta efficiency of 0.687 with the least error was introduced as the best model in this study. Also, the results showed that the genetic algorithm has helped to increase the accuracy of all utilized models.

Introduction

The sources of fossil energy are running out and the use of this type of energy has disadvantages such as greenhouse gas emissions, air pollution and global warming. So, there is no doubt that the replacement and use of clean and renewable energy such as solar energy can be the best and most appropriate way for production, growth and economic development of developing countries such as Iran. In addition, solar energy and solar radiation is one of the key factors in the fields of agriculture, hydrology and meteorology. Due to the fact that there are problems in using physical methods and meteorological data (requires complex calculations, high costs, etc.) in

predicting solar radiation, statistical methods and intelligence learning models can be used as a complementary solution that requires much less cost and time. In recent years, researchers have used these methods to model solar radiation. According to previous researches, the importance of using data-driven methods in estimating the intensity of solar radiation is clear. Therefore, in this study, the intensity of solar radiation in three meteorological stations of Ardabil province was estimated using different meteorological data in different combinations as the input of random forest models. Also, using the genetic algorithm, the obtained values from RF models were optimized. Finally, by comparing the results of different scenarios in the three stations of study area, the most accurate model in each station and among all stations was selected and introduced.

Materials and Methods

In the present study the meteorological data of average, minimum, and maximum temperature, relative humidity, and wind speed were utilized to estimate daily solar radiation in three meteorological stations of Ardabil province including Meshgin Shahr, Germi, and Nir from over a period of 2 years (2017-2018) on a daily scale so that the mentioned parameters are used as input data in eight different combinations. These combinations of input parameters consist of 1) mean temperature, 2) minimum temperature and maximum temperature, 3) mean temperature and relative humidity, 4) mean temperature, minimum temperature and maximum temperature, 5) minimum temperature, maximum temperature, and relative humidity, 6) minimum temperature, maximum temperature, relative humidity and wind speed, 7) mean temperature, minimum temperature, maximum temperature, relative humidity and 8) mean temperature, minimum temperature, maximum temperature, relative humidity, and wind speed. The methods used in this research include random forest and random forest optimized by genetic algorithm. Random forest is a set of CART trees and is expressed in four stages: First, the N subset of training samples (D_1, D_2…, D_n) are selected among the sample code set in the training section (D) using the Bootstrap sampling method, and finally, N decision tree will be formed. Then in the N classification tree node index, the m characteristic is selected randomly and according to the principle of minimum node purity, the best characteristic among the M candidate index will be selected. This way the trees will grow. The third step is to repeat the second step. N decision tree is generated. Finally, in the fourth stage, well-grown N decision trees form a random composite forest. The sample on the top floor of the random forest awaits a majority vote. By optimizing the basic parameters in the RF model, the efficiency and accuracy of the model can be improved. Trees have different percentages of accuracy in RF performance, so some trees can make incorrect predictions and reduce model performance. Various strategies are used to increase the accuracy of the model, including the hill climbing strategy and the greedy algorithm, although these strategies have drawbacks such as getting stuck in local optimization and creating a super-optimal subset. Therefore, the GA algorithm is implemented to solve this problem by selecting the best subset of features that can improve the performance of the RF model, and consequently, the RF model, which is optimized by the genetic algorithm (GA-RF), has high accuracy compared to the RF model.

In this study, 70% of the data were used to calibrate the studied models and the remaining 30% were used to validate the models, and then the results obtained from the validation section of each model were compared using statistical indices such as correlation coefficient (CC), root mean square error (RMSE) and Kling-Gupta efficiency (KGE) and the best models were selected. In addition, the Taylor diagram was used to analyze the accuracy of the applied models. Taylor diagram is a graphical solution for evaluating the accuracy of predicted data by simultaneously depicting three statistical parameters: root mean square error, standard deviation, and correlation coefficient.

Results

The results showed that in the random forest method and in Germi station, RF-VI model with CC of 0.77, RMSE of 0.522 MJ/m^2 d and KGE of 0.654, in Meshgin Shahr station, RF-VI model with CC of 0.813, RMSE of 0.417 MJ/m^2 d and KGE of 0.529 and in Nir station RF-V model with CC of 0.778, RMSE of 0.363 MJ/m^2 d and KGE of 0.693 had the best performance. Figure (1) illustrates bar charts of statistical indicators of the best-studied models in each station. Overall, by comparing the results between the three studied stations, the models of Nir station performed more accurately in the RF method than the other two stations, and the results of Meshgin Shahr and Germi stations were in the next rank, respectively. In GA-RF models, Nir,

Meshgin Shahr, and Germi stations were ranked first to third, respectively. It is noteworthy from the obtained results that in the studied stations, the genetic algorithm has improved the performance of all the utilized models so that all GA-RF models have more accurately estimated the intensity of solar radiation by reducing the error.

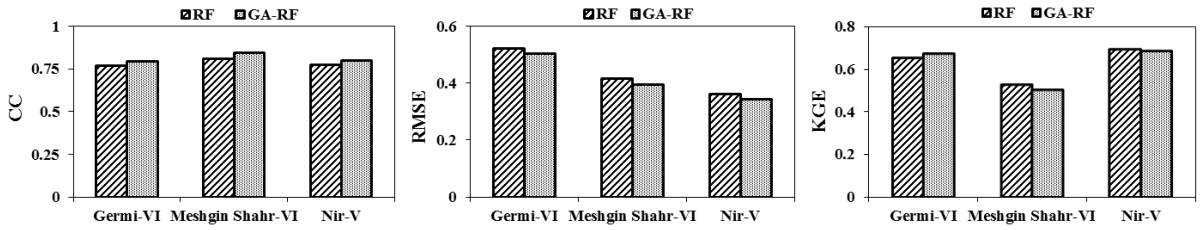


Fig. 1 Bar charts of statistical indicators of the best studied models in each station

Moreover, diagrams of temporal variations of observed and predicted solar radiation values, diagrams of the distribution of observed and predicted solar radiation, and bar graphs of statistical indices of the best studied models in each station showed the better performance of the top models. Similarly, Taylor diagram for the studied models showed that the GA-RF-VI model in Germi and Meshginshahr stations and the GA-RF-V model in Nir station had a smaller radial distance with the observed data and, therefore, indicated higher accuracy in estimating solar radiation. The parameters of minimum and maximum temperature and relative humidity had greatest effects on increasing the accuracy of solar radiation estimation in all three study stations and Taylor diagram showed the superiority of the models with the input of the mentioned parameters.

Conclusion

The overall conclusion showed that: 1) In Germi and Meshgin Shahr stations, the sixth combination models with the input parameters of minimum and maximum temperature, relative humidity and wind speed provided the most desirable results. 2) In Nir station, the fifth combination models with the input parameters of minimum and maximum temperature and relative humidity had the highest accuracy and the lowest error. 3) By comparing the results between mentioned stations, in both RF and GA-RF methods, Nir, Meshgin Shahr and Germi stations were ranked from high to low accuracy, respectively. 4) By examining RF with GA-RF models, it was concluded that the genetic algorithm improved the performance of the models and had a positive effect on all models.

Data Availability

The data can be sent by email by the responsible author upon request.

Conflicts of Interest

The authors of this article declared no conflict of interest regarding the authorship or publication of this article.

دوره ۸، شماره ۳، صفحات: ۶۳۶-۶۵۳ DOI: 10.22034/JEWE.2022.312038.1654

محیط‌زیست و مهندسی آب

Homepage: www.jewe.ir

ISSN: 2476-3683

مقاله پژوهشی

ارزیابی مدل هیبریدی جنگل تصادفی-الگوریتم ژنتیک در تخمین تابش خورشیدی روزانه

سجاد هاشمی[۱]، سعید صمدیان فرد[۲*] و علی‌اشرف صدرالدینی[۳]

[۱]دانشجوی دکتری، گروه مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران

[۲]دانشیار، گروه مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران

[۳]استاد، گروه مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران

چکیده

انرژی خورشیدی مهم‌ترین منبع انرژی‌های تجدیدپذیر و به عبارتی منبع اصلی انرژی‌های موجود در زمین است. لذا برآورد پارامتر تابش خورشیدی با دقت بالا اهمیت بسیاری دارد. در این راستا، در پژوهش حاضر از داده‌های هواشناسی ۳ ایستگاه هواشناسی استان اردبیل شامل مشگین شهر، گرمی و نیر در بازه زمانی ۲ ساله (۲۰۱۸ -۲۰۱۷) در مقیاس روزانه استفاده شد. سپس با به‌کارگیری روش‌های جنگل تصادفی و جنگل تصادفی-الگوریتم ژنتیک شدت تابش خورشیدی روزانه در هر یک از ایستگاه‌های مذکور برآورد گردید. متغیرهای هواشناسی مورد استفاده شامل حداقل، حداکثر و میانگین دما، رطوبت نسبی و سرعت باد بوده که در هشت ترکیب متفاوت به‌عنوان داده‌های ورودی در محاسبات مدل‌ها در نظر گرفته شد. نتایج به‌دست‌آمده با استفاده از پارامترهای آماری با یکدیگر مقایسه شده و مدل‌های برتر انتخاب شد. با مقایسه کلی نتایج، مدل‌های ایستگاه‌های نیر، مشگین شهر و گرمی به ترتیب از بیشترین به کمترین دقت مدل‌سازی رتبه‌بندی شدند؛ به‌طوری‌که مدل GA-RF-V در ایستگاه نیر با دارا بودن جذر میانگین مربعات خطای MJ/m^2 ۰/۳۴۶ d و راندمان کلینگ-گاپتا ۰/۶۸۷ با کمترین خطا به‌عنوان برترین مدل در این مطالعه معرفی شد. همچنین نتایج به‌دست‌آمده نشان داد که الگوریتم ژنتیک به افزایش دقت همه مدل‌های مورد استفاده کمک شایانی کرده است.

اطلاعات مقاله

تاریخ دریافت: [۱۴۰۰/۰۸/۰۳]

تاریخ بازنگری: [۱۴۰۱/۰۱/۲۱]

تاریخ پذیرش: [۱۴۰۱/۰۱/۳۴]

واژه‌های کلیدی:

اردبیل

انرژی خورشیدی

بهینه‌سازی

مدل‌های هوشمند

***نویسنده مسئول:**

s.samadian@tabrizu.ac.ir

۱- مقدمه

با توجه به این‌که از یک طرف منابع انرژی‌های فسیلی رو به اتمام است و از طرف دیگر استفاده از این نوع انرژی‌ها معایبی همچون انتشار گازهای گلخانه‌ای، آلودگی هوا و گرمایش جهانی دارد، جایگزین کردن و استفاده از انرژی‌های

پاک و تجدیدپذیری مانند انرژی خورشیدی می‌تواند بهترین و مناسب‌ترین راه برای توسعه اقتصادی کشورهای در حال توسعه مانند ایران باشد. علاوه بر موارد فوق، انرژی خورشیدی و پارامتر تابش خورشیدی یکی از عوامل کلیدی در زمینه‌های کشاورزی، هیدرولوژی و هواشناسی است و نقش اساسی در انواع فرایندهای فیزیکی، بیولوژیکی و شیمیایی از جمله ذوب برف، تبخیر و فتوسنتز گیاه و تولید محصول ایفا می‌کند. همچنین تخمین صحیح و دقیق مقدار تابش خورشیدی اهمیت فراوانی در برنامه‌ریزی آبیاری و طراحی سامانه‌ها و شبکه‌های آبیاری دارد (Peter and Steven 1999; Yang et al. 2001; Almorox and Hontoria 2004; Mossad 2004). لذا پیش‌بینی انرژی خورشیدی باعث بهبود مدیریت این انرژی، تخمین و کاهش هزینه‌های نگهداری سامانه‌های بهره‌برداری انرژی خورشیدی می‌شود. بنابراین می‌توان با به‌کارگیری روش‌هایی که شدت تابش خورشیدی را با دقت مناسبی پیش‌بینی می‌کنند کمک شایانی در جهت استفاده از انرژی خورشیدی نمود (Belaid and Mellit 2016; Mousavi et al. 2017; Wu et al. 2019). با توجه به اینکه مشکلاتی در به‌کارگیری روش‌های فیزیکی و داده‌های هواشناسی (نیاز به محاسبات پیچیده، هزینه‌های زیاد و مواردی از این قبیل) در پیش‌بینی تابش خورشیدی وجود دارد، می‌توان از روش‌های آماری و مدل-های یادگیری هوشمند به‌عنوان یک راه‌حل مکمل که به‌مراتب به هزینه و زمان کمتری نیاز دارند، استفاده کرد. در سال‌های گذشته محققان از این روش‌ها برای مدل‌سازی تابش خورشیدی استفاده کرده‌اند از جمله Ibrahim and Khatib (2017) مدل جدیدی را برای پیش‌بینی ساعتی تابش خورشیدی ارائه کردند. این مدل از ترکیب روش جنگل تصادفی با الگوریتم کرم شب‌تاب ایجاد شده که از الگوریتم کرم شب‌تاب برای بهینه‌سازی روش جنگل تصادفی به‌کاربرده شد همچنین از داده‌های هواشناسی ساعتی برای ورودی این مدل استفاده شد. با مقایسه نتایج مدل مذکور با نتایج روش‌های جنگل تصادفی، شبکه عصبی مصنوعی و شبکه عصبی مصنوعی بهینه‌سازی شده با الگوریتم کرم شب‌تاب، نتیجه گرفته شد که مدل معرفی شده در این تحقیق از دقت پیش‌بینی بالاتری نسبت به سایر روش‌ها برخوردار بود. Rao et al. (2018) تأثیر ترکیب‌های مختلفی از پارامترهای هواشناسی را بر تخمین تابش خورشیدی توسط مدل‌های شبکه عصبی مصنوعی[1] (ANN) بررسی کردند. نتایج به‌دست آمده نشان داد که مدلی از ANN با پارامترهای ورودی ساعات آفتابی و تابش فرازمینی بیشترین دقت را در برآورد تابش خورشیدی دارا بود. Benali et al. (2019) مقادیر ساعتی تابش خورشیدی را در منطقه اودیلو فرانسه پیش‌بینی کردند. آن‌ها از فن‌های محاسبات نرم شامل روش‌های شبکه عصبی مصنوعی، پایداری هوشمند و جنگل تصادفی استفاده کرده و بیان کردند که دقت مدل‌های جنگل تصادفی در برآورد تابش خورشیدی به‌طور قابل‌توجهی بالاتر از سایر مدل‌های مورد مطالعه بود. Samadianfard et al. (2019) در مطالعه‌ای از پارامترهای هواشناسی دمای حداقل، دمای حداکثر، رطوبت نسبی ساعات آفتابی، حداکثر ساعات آفتابی، شاخص آسمان صاف، روز از سال و تابش فرازمینی برای ورودی روش‌های رگرسیون بردار پشتیبان[2] (SVR)، مدل درختی[3] (MT)، برنامه‌ریزی بیان ژن[4] (GEP) و سیستم استنتاجی تطبیقی عصبی-فازی[5] (ANFIS) به‌منظور برآورد تابش خورشیدی در ایستگاه سینوپتیک تبریز استفاده کردند. نتایج به‌دست‌آمده نشان داد که مدل-های SVR و MT به ترتیب در رتبه‌های اول و دوم دقیق-ترین مدل‌های بررسی شده قرار گرفتند. همچنین آن‌ها نتایج حاصل از مدل‌های مذکور را با نتایج معادلات تجربی مقایسه کرده و بیان کردند که معادلات تجربی دقت کمتری نسبت به روش‌های هوشمند داشته است. در تحقیق دیگری، Alizamir et al. (2020) کارایی شش روش یادگیری ماشینی مختلف شامل درخت ارتقای گرادیان[6] (GBT)، شبکه عصبی پرسپترون چند لایه[7] (MLPNN)، دو نوع مختلف سیستم استنتاجی تطبیقی عصبی-فازی (ANFIS)، شاخه رگرسیون تطبیقی چند متغیره[8] (MARS) و درخت رگرسیون و طبقه‌بندی[9] (CART) برای پیش‌بینی تابش خورشیدی در دو ایستگاه در دو منطقه متفاوت بررسی کردند. آن‌ها از داده‌های دمای حداقل، دمای حداکثر و رطوبت نسبی برای ورودی مدل‌های توسعه داده شده استفاده و با مقایسه نتایج نشان دادند که مدل GBT مدل-

[1]Artificial Neural Network
[2]Support Vector Regression
[3]Model Tree
[4]Gene Expression Programming
[5]Adaptive Neuro Fuzzy Inference System
[6]Gradient Boosting Tree
[7]Multilayer Perceptron Neural Network
[8]Multivariate Adaptive Regression Spline
[9]Classification and Regression Tree

سازی تابش خورشیدی را بهتر از سایر مدل‌ها انجام داده است. Bayat and Mirlatifi (2009) با استفاده از روش شبکه‌های عصبی مصنوعی شدت تابش کل خورشیدی روزانه را در ایستگاهی که سابقه اندازه‌گیری تابش کل خورشیدی را نداشت و دارای فاکتورهای اقلیمی مشابه با ایستگاهی بود که در آن تابش کل خورشیدی اندازه‌گیری می‌شد، تخمین زدند و نتایج را با شش مدل رگرسیونی مورد مقایسه قرار دادند. از بین تمام مدل‌های استفاده شده جهت تخمین تابش کل خورشیدی روزانه، مدل شبکه عصبی مصنوعی با ورودی ساعات آفتابی حداکثر و تابش فرازمینی روزانه و ساعات آفتابی اندازه‌گیری شده بهترین نتیجه را در سطح معنی‌داری ۱٪ ارائه داد.

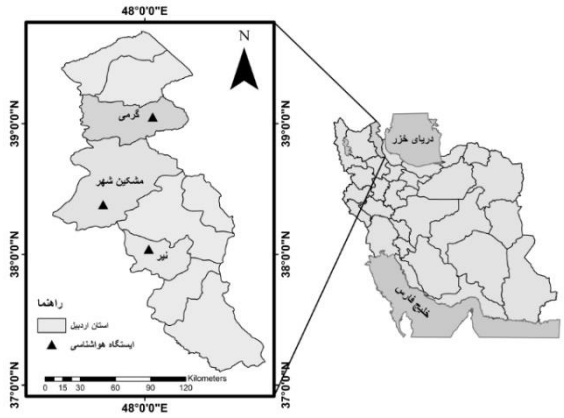
در مطالعه دیگری، Mohammadi et al. (2019) کارایی عملکرد روش تحلیل مؤلفه اصلی[1] (PCA) تئوری آنتروپی[2] (EN) را برای تعیین ورودی مدل‌های شبکه عصبی مصنوعی پرسپترون چندلایه (MLP)، شبکه عصبی مصنوعی تابع پایه شعاعی[3] (RBF)، ماشین بردار پشتیبان[4] (SVM) و برنامه-ریزی ژنتیک[5] (GEP) در برآورد تابش خورشیدی در دو ایستگاه همدید کرمان و مشهد مورد بررسی قرار دادند. با توجه به نتایج به‌دست‌آمده در ایستگاه کرمان، مدل ENT-MLP و در ایستگاه مشهد، مدل PCA-MLP با کمترین خطا، به‌عنوان بهترین مدل معرفی شدند.

با توجه به تحقیقات انجام گرفته، اهمیت کاربرد روش‌های داده محور در برآورد شدت تابش خورشیدی مشخص می‌شود. لذا در این پژوهش، با هدف برآورد شدت تابش خورشیدی در سه ایستگاه هواشناسی استان اردبیل، از داده‌های مختلف هواشناسی در ترکیب‌های مختلف به‌عنوان ورودی مدل جنگل تصادفی استفاده شد. همچنین با به‌کارگیری الگوریتم ژنتیک مقادیر به‌دست‌آمده از مدل‌های RF بهینه‌سازی شد. در نهایت با مقایسه نتایج سناریوهای مختلف در سه ایستگاه موردمطالعه دقیق‌ترین مدل در هر ایستگاه و بین همه ایستگاه‌ها انتخاب و معرفی شد.

۲- مواد و روش‌ها

۲-۱- منطقه موردمطالعه

استان اردبیل با وسعتی حدود ۱۷۸۰۰ km^2 حدود ۱/۱٪ از مساحت کل کشور ایران را به خود اختصاص داده است. محدوده مختصات جغرافیایی در این استان در بازه °۳۷/۴۵ تا °۳۹/۴۲ عرض شمالی و °۴۷/۳۰ تا °۴۸/۵۵ طول شرقی قرار داشته و میانگین ارتفاع آن از سطح دریا ۱۴۰۰ m است. ویژگی‌های خاص جغرافیایی و توپوگرافی استان مانند رشته‌کوه‌هایی با ارتفاع بیش از ۴۰۰۰ m و دشت‌های وسیع باعث شده این استان از لحاظ میزان نزولات جوی وضعیت بهتری نسبت به سایر مناطق کشور داشته باشد. در تحقیق حاضر داده‌های تابش خورشیدی (R_s)، میانگین دما (T_m)، حداقل دما (T_{min})، حداکثر دما (T_{max})، رطوبت نسبی (RH) و سرعت باد (WS) سه ایستگاه هواشناسی استان اردبیل شامل مشگین شهر، گرمی و نیر در بازه زمانی ۲ ساله (۲۰۱۷-۲۰۱۸) در مقیاس روزانه به کار گرفته شده به-طوری‌که از ۷۰٪ داده‌ها برای واسنجی مدل‌های موردمطالعه و از ۳۰٪ باقی‌مانده برای صحت‌سنجی مدل‌ها استفاده شده است. در شکل (۱) منطقه مورد مطالعه نشان داده شده است.



شکل ۱- موقعیت جغرافیایی ایستگاه‌های موردمطالعه

Fig. 1 Geographical location of the studied stations

همچنین مختصات جغرافیایی و خلاصه‌ای از آمار داده‌های مشاهداتی ایستگاه‌های موردمطالعه شامل پارامترهای میانگین (X_{mean})، حداقل (X_{min})، حداکثر (X_{max})، انحراف معیار (S_x)، ضریب تغییرات (C_v) و ضریب چولگی (C_{sx}) در جدول (۱) آورده شده است.

[1]Principal Component Analysis
[2]Entropy Theory
[3]Radial Basis Function
[4]Support Vector Machine
[5]Genetic Programming

جدول ۱- مشخصات جغرافیایی و آمار داده‌های مشاهداتی ایستگاه‌های موردمطالعه

Table 1 Geographical characteristics and statics of observational data of the studied stations

Station	Longitude and Latitude	Elevation from sea level (m)	Parameter	X_{mean}	X_{min}	X_{max}	S_x	C_v	C_{sx}
Germi	48° 04´ 39° 01´	850	R_s (MJ/m^{2}d)	2.454	0.378	3.94	1.002	0.408	-0.309
			T_{mean} (°C)	19.325	-4.5	33.5	8.237	0.426	-0.788
			T_{min} (°C)	14.828	-8.8	26.7	7.524	0.507	-0.822
			T_{max} (°C)	24.522	-0.6	40.5	8.727	0.356	-0.808
			RH (%)	64.568	15.875	100	20.602	0.319	-0.214
			WS (m/s)	3.063	0.125	9.5	1.815	0.593	1.091
Meshgin Shahr	47° 40´ 38° 23´	1400	R_s (MJ/m^{2}d)	1.95	0.267	3.06	0.727	0.373	-0.385
			T_{mean} (°C)	16.236	-9.3	27.6	7.501	0.462	-0.953
			T_{min} (°C)	11.582	-12.4	22.4	6.922	0.598	-1.001
			T_{max} (°C)	22.222	-3.4	35.6	8.481	0.382	-0.845
			RH (%)	52.19	17	100	17.92	0.343	0.556
			WS (m/s)	2.192	0.5	9.1	1.102	0.503	2.799
Nir	47° 59´ 38° 02´	1450	R_s (MJ/m^{2}d)	2.276	0	3.453	0.813	0.357	-0.699
			T_{mean} (°C)	14.964	-13.7	27	7.502	0.501	-1.066
			T_{min} (°C)	7.662	-25.5	19.9	6.519	0.851	-1.272
			T_{max} (°C)	23.015	-7.1	38.4	9.088	0.395	-0.864
			RH (%)	58.302	21	97.75	16.32	0.28	0.179
			WS (m/s)	3.941	0.625	11.875	1.884	0.478	1.5

۲-۲- روش‌های مورد استفاده

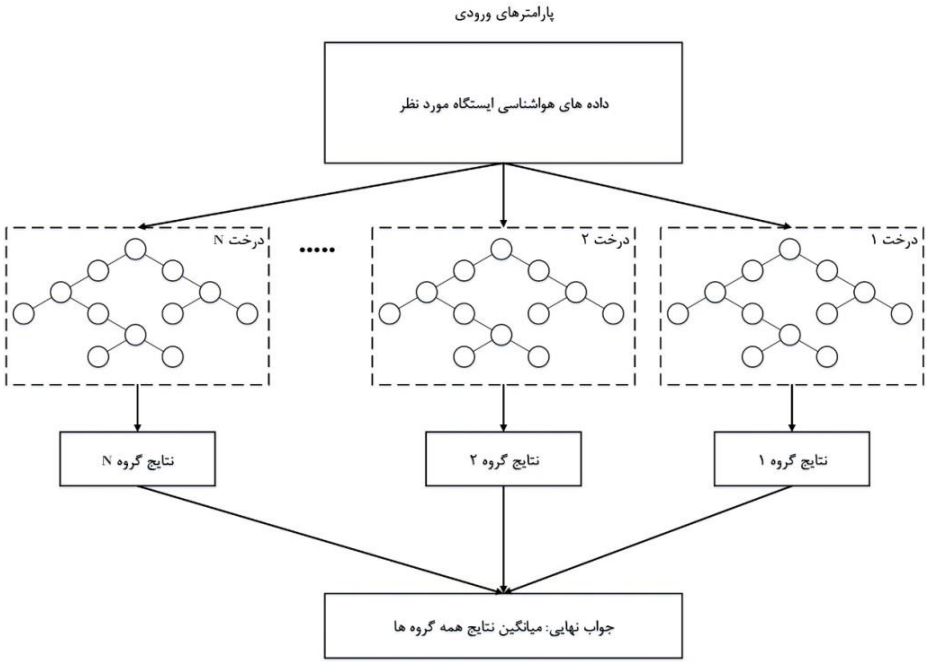
۲-۲-۱- جنگل تصادفی

در الگوریتم جنگل تصادفی برای ایجاد هر درخت، گروه مختلفی از الگوهای موجود با در نظر گرفتن جایگزینی دوباره هر الگوی برگزیده، انتخاب می‌شوند. اندازه این گروه نمونه-برداری شده، برابر تعداد کل الگوهای موجود خواهد بود. این روش در بین روش‌های درختی، روش نسبتاً پیچیده‌ای است که به‌منظور افزایش دقت مدل در آن چندین درخت تصمیم آموزش داده می‌شود. نتیجه حاصل پیش‌بینی گروهی از درختان تصمیم است (Breiman 2001).

در روش یادگیری جنگل تصادفی هر درخت تصمیم با استفاده از یک نمونه تصادفی که از مجموعه داده‌های آموزشی انتخاب شده است، آموزش می‌بینند. انتخاب مجموعه‌ای از متغیرهای پیش‌بینی کننده نیز که برای تقسیم‌بندی گره‌ها استفاده می‌شود، به‌صورت تصادفی انجام می‌گیرد. در روش جنگل تصادفی دو ویژگی m مشخصه و n گره به ترتیب برای تعداد متغیرهای کمکی مورد استفاده در هر زیر مجموعه و تعداد درختان مورد استفاده در جنگل تعیین می‌گردد. تعداد متغیرها می‌تواند از یک تا تعداد کل متغیرهای کمکی را در برگیرد. تعداد درختان توسط کاربر و معمولاً از ۵۰۰ تا ۱۰۰۰ انتخاب می‌گردد (Brungard et al. 2015). به بیان دیگر، در درخت تصمیم با دنبال کردن مجموعه‌ای از سؤالات مرتبط با خصوصیات داده‌ها و نگاه به داده جاری برای اتخاذ تصمیم، طبقه یا دسته آن تعیین می-شود. CART[1] یک الگوریتم درخت باینری در الگوریتم درخت تصمیم است و جنگل تصادفی مجموعه‌ای از درختان CART است و تحت چهار مرحله بیان می‌شود: ابتدا N زیرمجموعه از نمونه‌های آموزش (D1, D2…, Dn) در میان مجموعه کد نمونه‌های موجود در بخش آموزش (D) با استفاده از روش نمونه‌برداری Bootstrap برگزیده می‌شوند و درنهایت N درخت تصمیم تشکیل خواهد شد. سپس در N شاخص گره درخت طبقه‌بندی، m مشخصه به‌طور تصادفی انتخاب می‌شود و مطابق با اصل حداقل خلوص گره، بهترین مشخصه در بین M شاخص کاندید انتخاب خواهد شد. به این ترتیب درختان رشد خواهند کرد. مرحله سوم تکرار گام دوم است. N درخت تصمیم تولید می‌شود. و در نهایت در مرحله چهارم، N درخت تصمیم که به خوبی رشد پیدا

[1]Classification and regression trees

کرده‌اند جنگل تصادفی طبقه‌بندی ترکیبی را ایجاد می‌کنند. نمونه واقع در طبقه نهایی جنگل تصادفی منتظر رأی اکثریت می‌ماند. شکل (۲) نمای کلی روش جنگل تصادفی را نشان می‌دهد.



شکل ۲- نمای کلی روش جنگل تصادفی

Fig. 2 Overview of the random forest method

۲-۲-۲- الگوریتم ژنتیک

الگوریتم ژنتیک (GA)[1] یک روش تکاملی با ساختار انتخاب طبیعی است که بر پایه نظریه داروین استوار است که توسط Holland (1992) و Goldberg (1989) توسعه داده شد. الگوریتم‌های ژنتیک اغلب گزینه خوبی برای تکنیک‌های پیش‌بینی بر مبنای رگرسیون هستند. در این روش دو نکته اساسی وجود دارد: اول این‌که این روش خطی است و دوم این‌که به جای این‌که در میان «فضای پارامترها» جستجو شود، پارامترهای مورد استفاده مشخص می‌شوند. با استفاده از الگوریتم‌های ژنتیک یک اَبر فرمول یا یک طرح تنظیم می‌شود. سپس داده‌هایی برای گروهی از متغیرهای مختلف، شاید در حدود ۲۰ متغیر فراهم می‌شود. پس از آن الگوریتم ژنتیک اجرا خواهد شد که بهترین تابع و متغیرها را مورد جستجو قرار می‌دهد. هر فرمولی که از طرح داده شده بالا تبعیت کند؛ فردی از جمعیت فرمول‌های ممکن تلقی می‌شود.

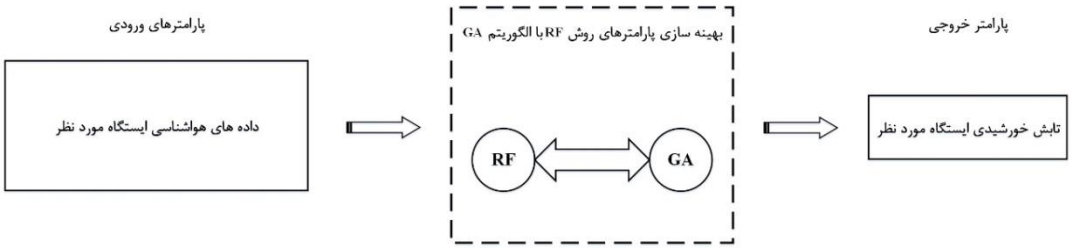
۳-۲-۲- مدل ترکیبی الگوریتم ژنتیک-جنگل تصادفی

با بهینه‌سازی پارامترهای اساسی در مدل RF می‌توان بهره‌وری و دقت مدل را بهبود بخشید. بر اساس مطالعه Adnan et al. (2016)، درختان دقت‌های مختلفی را در عملکرد RF دارا هستند، به‌طوری‌که برخی از درختان می‌توانند پیش‌بینی‌های نادرستی انجام دهند و عملکرد و کارایی مدل را کاهش دهند. از راهبردهای مختلفی برای افزایش دقت مدل استفاده می‌شود که الگوریتم راهبرد تپه نوردی[2] و الگوریتم حریصانه[3] از جمله این فن‌ها هستند، هرچند این استراتژی‌ها دارای نقصان‌هایی مثل گیر کردن در یک بهینه محلی و ایجاد یک زیرمجموعه فوق بهینه هستند. بنابراین، الگوریتم GA برای حل این مشکل با انتخاب بهترین زیر مجموعه از ویژگی‌هایی که قادر به بهبود عملکرد مدل RF است، اجرا شده و در نتیجه، مدل RF که توسط الگوریتم ژنتیک بهینه‌شده است (GA-RF)، در مقایسه با مدل RF از دقت بالایی برخوردار است. شکل (۳) نمودار مدل ترکیبی GA-RF را نشان می‌دهد.

[1]Genetic Algorithm

[2]Hill climbing strategy

[3]Greedy algorithm



شکل ۳- نمای کلی مدل ترکیبی GA-RF

Fig. 3 Overview of the GA-RF hybrid model

۲-۳- معیارهای ارزیابی کارایی مدل‌ها

در این پژوهش برای ارزشیابی دقت مدل‌ها و مقایسه نسبی نتایج مدل‌های برآوردی با مقادیر اندازه‌گیری شده تابش خورشیدی با استفاده از روش‌های هوشمند موردمطالعه، از معیارهای آماری، ضریب همبستگی[1] (CC)، جذر میانگین مربعات خطا[2] (RMSE) و راندمان کلینگ-گاپتا[3] (KGE) استفاده شد که مطابق روابط (۱) تا (۶) محاسبه می‌شوند.

$$CC=\frac{\left(\sum_{i=1}^{n}O_iP_i-\frac{1}{n}\sum_{i=1}^{n}O_i\sum_{i=1}^{n}P_i\right)}{\left(\sum_{i=1}^{n}O_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}O_i\right)^2\right)\left(\sum_{i=1}^{n}P_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}P_i\right)^2\right)} \quad (۱)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i-O_i)^2} \quad (۲)$$

$$KGE=1-\sqrt{(R-1)^2+(\beta-1)^2+(\gamma-1)^2} \quad (۳)$$

$$R=\frac{\left[\sum_{i=1}^{n}(O_i-\bar{O})(P_i-\bar{P})\right]}{\sqrt{\sum_{i=1}^{n}(O_i-\bar{O})^2\sum_{i=1}^{n}(P_i-\bar{P})^2}} \quad (۴)$$

$$\gamma=\frac{CV_P}{CV_O}=\frac{\frac{\sigma_P}{\bar{P}}}{\frac{\sigma_P}{\bar{O}}} \quad (۵)$$

$$\beta=\frac{\bar{P}}{\bar{O}} \quad (۶)$$

که، n تعداد داده‌ها، O_i و P_i به ترتیب مقادیر مشاهداتی و پیش‌بینی‌شده تابش خورشیدی و $\bar{O}$ و $\bar{P}$ به‌ترتیب میانگین مقادیر مشاهداتی و پیش‌بینی‌شده تابش خورشیدی هستند. همچنین σ_o و σ_p به ترتیب انحراف معیار مقادیر مشاهداتی و محاسبه شده با استفاده از روش‌های موردمطالعه (RF و GA-RF) می‌باشد. همچنین، دیاگرام تیلور برای تحلیل دقت روش‌های مورد استفاده در تخمین تابش خورشیدی به کار گرفته شد. دیاگرام تیلور راه‌حلی گرافیکی برای ارزیابی دقت داده‌های پیش‌بینی شده با به تصویر کشیدن همزمان سه پارامتر آماری جذر میانگین مربعات خطا، انحراف معیار و ضریب همبستگی می‌باشد (Taylor 2001).

۳-یافته‌ها و بحث

با استفاده از داده‌های هواشناسی حداقل، حداکثر و میانگین دما، رطوبت نسبی و سرعت باد، هشت ترکیب متفاوت از این داده‌ها بر مبنای ضرایب همبستگی بین پارامتر تابش خورشیدی هر ایستگاه با سایر پارامترهای هواشناسی، به‌عنوان پارامترهای ورودی در محاسبات مدل‌ها در نظر گرفته شد. جدول (۲) ضرایب همبستگی بین پارامتر تابش خورشیدی ایستگاه‌های مورد مطالعه با سایر پارامترهای هواشناسی همان ایستگاه را نشان می‌دهد.

جدول ۲- ضرایب همبستگی بین پارامتر تابش خورشیدی هر ایستگاه با سایر پارامترهای هواشناسی

Table 2 Correlation coefficients between the solar radiation parameter of each station with other meteorological parameters

Station	T_{min}	T_{max}	T_m	RH	WS	R_s
Germi	0.730	0.819	0.805	-0.629	0.411	1
Meshgin Shahr	0.682	0.788	0.759	-0.452	-0.009	1
Nir	0.603	0.780	0.762	-0.357	-0.233	1

با توجه به جدول (۲) در ایستگاه گرمی، پارامتر دمای حداکثر بیش‌ترین همبستگی را با داده‌های تابش خورشیدی دارد. دمای میانگین و دمای حداقل نیز

[1]Correlation Coefficient
[2]Root Mean Square Error
[3]Kling-Gupta efficiency

همبستگی بالایی را با داده‌های تابش خورشیدی نشان دادند. همچنین ضریب همبستگی ۰/۴۱۱ سرعت باد با تابش خورشیدی حاکی از تأثیرگذاری این پارامتر در مقادیر تابش خورشیدی است. در ایستگاه مشگین شهر نیز دمای حداکثر، دمای میانگین و دمای حداقل به ترتیب با ضرایب همبستگی ۰/۷۸۸، ۰/۷۵۹ و ۰/۶۸۲ بیش‌ترین همبستگی را با تابش خورشیدی داشتند. همچنین در ایستگاه نیر همچون دو شهر دیگر به ترتیب سه پارامتر دمای حداکثر، دمای میانگین و دمای حداقل پارامترهایی با بیش‌ترین همبستگی با تابش خورشیدی شناخته شدند. به‌طورکلی نتایج جدول (۲) حاکی از همبستگی بالای داده‌های دما با تابش خورشیدی در منطقه موردمطالعه دارد. در ادامه در جدول (۳) ترکیب‌های مختلفی از پارامترهای هواشناسی به‌عنوان ورودی مدل‌های مورد استفاده آورده شده است.

جدول ۳- ترکیب‌های متفاوت از پارامترهای هواشناسی مورداستفاده

Table 3 Different combinations of utilized meteorological parameters

Model	Input parameters
I	T_m
II	T_{min}, T_{max}
III	T_m, RH
IV	T_m, T_{min}, T_{max}
V	T_{min}, T_{max}, RH
VI	T_{min}, T_{max}, RH, WS
VII	T_m, T_{min}, T_{max}, RH
VIII	T_m, T_{min}, T_{max}, RH, WS

با در نظر گرفتن اینکه داده‌های دما بیش‌ترین همبستگی را با تابش خورشیدی داشتند سعی شده است بیشتر از داده‌های دما در ترکیب پارامترهای ورودی استفاده شود به‌طوری‌که با توجه به جدول (۳)، هشت ترکیب پارامترهای ورودی شامل دمای میانگین، دمای حداقل و دمای حداکثر، دمای میانگین و رطوبت نسبی، دمای میانگین، دمای حداقل و دمای حداکثر، دمای حداقل، دمای حداکثر و رطوبت نسبی، دمای حداقل، دمای حداکثر، رطوبت نسبی و سرعت باد، دمای میانگین، دمای حداقل، دمای حداکثر، رطوبت نسبی و دمای میانگین، دمای حداقل، دمای حداکثر، رطوبت نسبی و سرعت باد می‌باشند.

در جدول (۴) خلاصه‌ای از آمار داده‌های هواشناسی مورد استفاده شامل پارامترهای میانگین (X_{mean})، حداقل (X_{min})، حداکثر (X_{max})، انحراف معیار (S_x)، ضریب تغییرات (C_v) و ضریب چولگی (C_{sx}) به تفکیک برای هر دو بخش واسنجی و صحت‌سنجی ارائه شده است.

با توجه به جدول (۴) در مقایسه داده‌های تابش خورشیدی، ایستگاه گرمی به‌طور میانگین مقادیر بزرگ‌تری را در مقایسه با دیگر ایستگاه‌ها دارد. در مقایسه داده‌های دما نیز ایستگاه گرمی بالاترین مقدار دما را ثبت کرده است و در حداقل میزان دما نیز ایستگاه نیر رتبه نخست را دارد. مقادیر انحراف معیار داده‌های تابش خورشیدی در هر سه ایستگاه موردمطالعه و در دو بخش واسنجی و صحت‌سنجی پایین بوده و نشان‌دهنده پراکندگی اندک این داده‌ها می‌باشد. ضریب چولگی بیانگر میزان عدم تقارن توزیع احتمال داده‌ها حول میانگین آن‌هاست که در بخش واسنجی داده‌های تابش خورشیدی در ایستگاه‌های گرمی و مشگین شهر و در ایستگاه نیر داده‌های رطوبت نسبی مقادیر چولگی نزدیک به صفر و بیش‌ترین تقارن را نشان دادند. همچنین در بخش صحت‌سنجی داده‌های دمای میانگین در دو شهر گرمی و مشگین شهر و رطوبت نسبی در شهر نیر دارای مقادیر چولگی نزدیک به صفر بودند.

برای مقایسه دقیق‌تر و نتیجه‌گیری مطلوب در جدول (۵) مقادیر پارامترهای آماری CC، RMSE و KGE برای مدل‌های مختلف RF و GA-RF در ایستگاه‌های مورد مطالعه آورده شده است.

جدول ۴- آمار داده‌های هواشناسی در ایستگاه‌های موردمطالعه در هر دو بخش واسنجی و صحت‌سنجی

Table 4 Statistics of meteorological data at studied stations in both calibration and validation sections

Station	Stage	Parameter	X_{mean}	X_{min}	X_{max}	S_x	C_v	C_{sx}
Germi	Calibration	R_s	2.267	0.378	3.94	1.021	0.45	-0.09
		T_{mean}	17.308	-4.5	31.1	8.778	0.507	-0.42
		T_{min}	12.92	-8.8	26.7	8.011	0.62	-0.433
		T_{max}	22.435	-0.6	36.3	9.318	0.415	-0.453
		RH	64.846	20.75	100	20.663	0.319	-0.1
		WS	2.784	0.125	9.5	1.903	0.683	1.409
	Validation	R_s	2.896	0.772	3.933	0.801	0.277	-0.692
		T_{mean}	24.087	15.4	33.5	3.752	0.156	0.012
		T_{min}	19.333	12.2	26.4	3.172	0.164	0.17
		T_{max}	29.451	19.5	40.5	4.09	0.139	-0.044
		RH	63.914	15.875	95.875	20.568	0.322	-0.491
		WS	3.721	0.875	8.125	1.392	0.374	0.821
Meshgin Shahr	Calibration	R_s	1.843	0.352	3.06	0.741	0.402	-0.103
		T_{mean}	14.618	-9.3	27	8.149	0.557	-0.603
		T_{min}	10.057	-12.4	22.4	7.561	0.752	-0.62
		T_{max}	20.421	-3.4	34.8	9.137	0.447	-0.516
		RH	51.301	17	100	19.2	0.4	0.686
		WS	2.318	0.5	9.1	1.254	0.541	2.462
	Validation	R_s	2.201	0.267	2.967	0.629	0.286	-1.095
		T_{mean}	20.055	10.3	27.6	3.434	0.171	-0.032
		T_{min}	15.183	9.2	21.8	2.795	0.184	0.466
		T_{max}	26.475	12.2	35.6	4.388	0.166	-0.58
		RH	54.289	18	95	14.447	0.3	0.125
		WS	1.893	0.6	3.1	0.497	0.262	-0.219
Nir	Calibration	R_s	2.128	0	3.405	0.854	0.401	-0.426
		T_{mean}	13.318	-13.7	26.5	8.105	0.609	-0.759
		T_{min}	6.114	-25.5	16.9	6.949	1.136	-1.035
		T_{max}	21.082	-7.1	37.8	9.725	0.461	-0.587
		RH	57.03	21	94.625	16.312	0.286	0.331
		WS	4.254	0.625	11.875	2.112	0.497	1.143
	Validation	R_s	2.627	0.399	3.453	0.577	0.219	-1.353
		T_{mean}	18.851	8.6	27	3.581	0.19	0.176
		T_{min}	11.316	5.1	19.9	3.166	0.28	0.428
		T_{max}	27.581	14.2	38.4	5.004	0.181	-0.263
		RH	61.305	21.857	97.75	16.039	0.262	-0.172
		WS	3.202	1.857	5.375	0.795	0.248	0.289

با توجه به جدول (۵) در روش جنگل تصادفی و در ایستگاه گرمی، مدل RF-VI با ضریب همبستگی ۰/۷۷، جذر میانگین مربعات خطای ۰/۵۲۲ MJ/m^2 d و راندمان کلینگ-گاپتا ۰/۶۵۴ به‌عنوان دقیق‌ترین مدل در این روش انتخاب شد. مدل RF-VIII با پنج پارامتر ورودی شامل میانگین، حداقل و حداکثر دما، رطوبت نسبی و سرعت باد و مدل RF-IV با سه پارامتر ورودی شامل رطوبت نسبی، حداقل و حداکثر دما به ترتیب در جایگاه‌های دوم و سوم مدل‌های این روش قرار گرفتند. در این روش سناریو اول (مدل RF-I) کم‌دقت‌ترین عملکرد را داشت. با این وجود، در روش GA-RF تمامی مدل‌ها با کاهش خطا و افزایش دقت همراه بودند، به‌طوری‌که سه مدل برتر یعنی مدل‌های GA-RF-VI، GA-RF-VIII و GA-RF-IV به ترتیب با افزایش

ضریب همبستگی و کاهش جذر میانگین مربعات خطا همراه بودند.

جدول ۵- پارامترهای آماری مدل‌های مختلف RF و GA-RF در ایستگاه‌های موردمطالعه

Table 5 Statistical parameters of different RF and GA-RF models in the studied stations

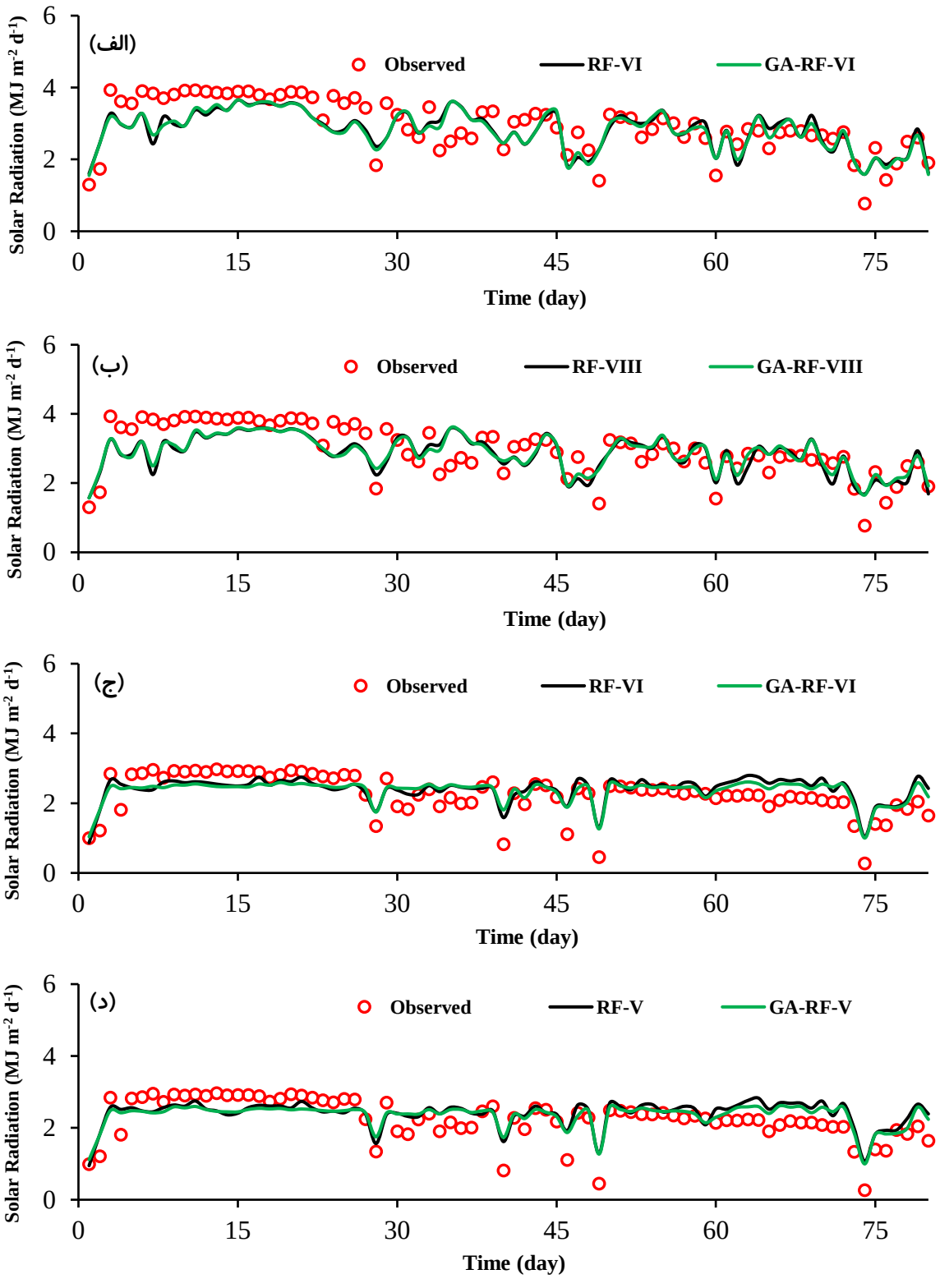
Station	Scenario	RF			GA-RF		
		CC	RMSE	KGE	CC	RMSE	KGE
Germi	I	0.463	0.732	0.343	0.698	0.589	0.426
	II	0.732	0.543	0.596	0.772	0.523	0.544
	III	0.667	0.616	0.614	0.705	0.58	0.586
	IV	0.757	0.527	0.571	0.775	0.511	0.592
	V	0.758	0.534	0.656	0.772	0.52	0.651
	VI	0.77	0.522	0.654	0.794	0.505	0.673
	VII	0.752	0.536	0.629	0.776	0.512	0.632
	VIII	0.76	0.527	0.639	0.786	0.507	0.614
Meshgin Shahr	I	0.418	0.59	0.276	0.631	0.505	0.328
	II	0.754	0.459	0.515	0.809	0.418	0.513
	III	0.695	0.487	0.396	0.755	0.454	0.425
	IV	0.765	0.451	0.486	0.804	0.434	0.454
	V	0.805	0.423	0.526	0.846	0.396	0.507
	VI	0.813	0.417	0.529	0.847	0.396	0.505
	VII	0.808	0.422	0.518	0.827	0.41	0.496
	VIII	0.811	0.43	0.497	0.839	0.413	0.468
Nir	I	0.36	0.569	0.292	0.593	0.466	0.34
	II	0.699	0.414	0.614	0.769	0.371	0.631
	III	0.698	0.412	0.607	0.747	0.385	0.579
	IV	0.672	0.427	0.485	0.752	0.389	0.533
	V	0.778	0.363	0.693	0.801	0.346	0.687
	VI	0.774	0.373	0.708	0.785	0.36	0.666
	VII	0.749	0.383	0.584	0.787	0.363	0.593
	VIII	0.747	0.387	0.581	0.776	0.372	0.577

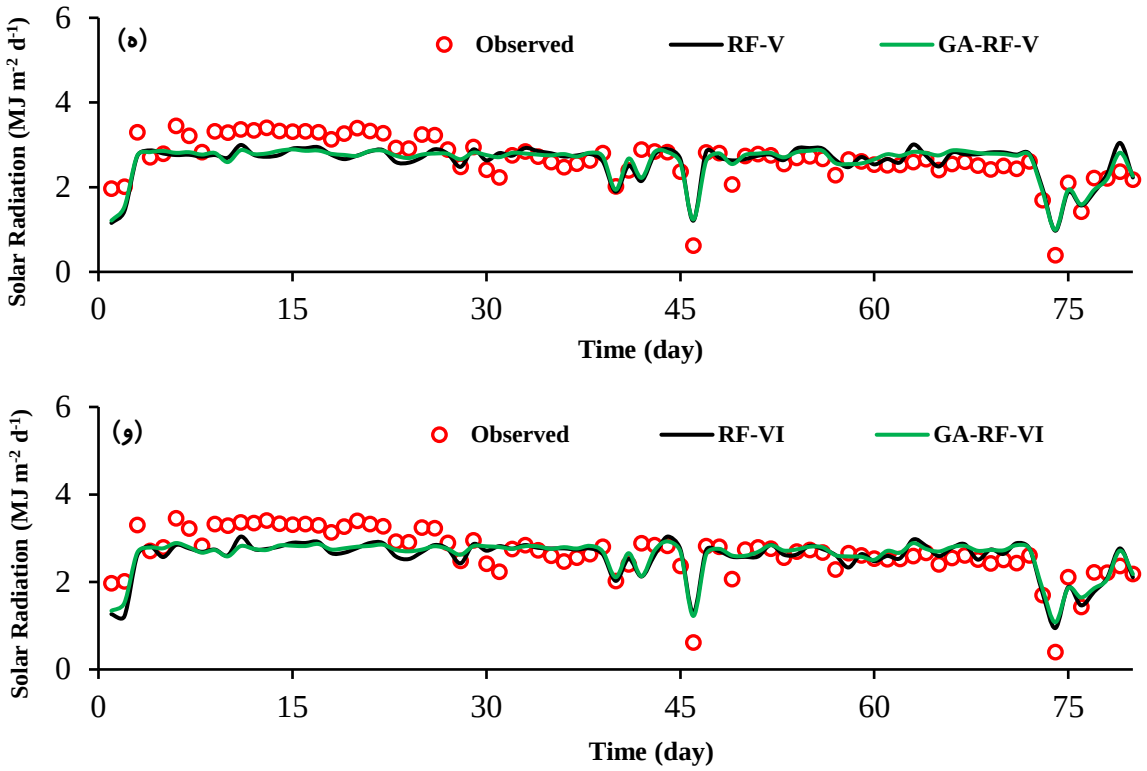
با توجه به جدول (۵) در ایستگاه مشگین شهر و در روش جنگل تصادفی مدل RF-VI با دارا بودن ضریب همبستگی ۰/۸۱۳، جذر میانگین مربعات خطای ۰/۴۱۷ MJ/m^2 d و راندمان کلینگ-گاپتا ۰/۵۲۹ بهترین مدل شناخته شد. در رتبه‌های بعدی، مدل‌های RF-V و RF-VII با ضریب همبستگی بالا و خطای کمتر مناسب‌ترین عملکرد را در بین مدل‌های جنگل تصادفی داشتند. مدل‌های RF-I و RF-III نیز کم‌دقت‌ترین مدل‌های این روش بودند. در این ایستگاه نیز، الگوریتم ژنتیک باعث افزایش کارایی مدل‌های جنگل تصادفی شد به‌طوری‌که تمام مدل‌های GA-RF باعث افزایش تطابق داده‌های محاسباتی با مشاهدات شدند. در این روش سه مدل GA-RF-V، GA-RF-VI و GA-RF-VII با داشتن ضریب همبستگی بالا و افزایش دقت مدل‌های جنگل تصادفی، به‌عنوان سه مدل برتر شناخته شدند.

نتایج به‌دست‌آمده در ایستگاه نیر (جدول ۵) نیز نشان داد که در بین مدل‌های RF، مدل RF-V دارای ضریب همبستگی ۰/۷۷۸، جذر میانگین مربعات خطای ۰/۳۶۳ MJ/m^2 d و راندمان کلینگ-گاپتا ۰/۶۹۳ بوده و با کم‌ترین خطا، مطلوب‌ترین نتایج را در بین مدل‌های RF ارائه کرد. مدل‌های RF-VI و RF-VII نیز هر دو با چهار پارامتر ورودی جزو مدل‌های برتر روش جنگل تصادفی بودند. در روش الگوریتم ژنتیک-جنگل تصادفی نیز همه مدل‌ها عملکرد قابل‌قبول و تأثیر مثبتی در کاهش خطای مدل‌های RF داشتند. به‌طوری‌که مدل‌های GA-RF-V، GA-RF-VI و GA-RF-VII با افزایش ضریب همبستگی و کاهش جذر میانگین مربعات خطا عملکرد مناسب خود را نشان دادند.

به‌طورکلی با مقایسه نتایج بین سه ایستگاه موردمطالعه، مدل‌های ایستگاه نیر در روش RF عملکرد دقیق‌تری نسبت

به دو ایستگاه دیگر داشتند و نتایج ایستگاه‌های مشگین شهر و گرمی نیز به ترتیب در رتبه‌های بعدی قرار گرفتند. در مدل‌های GA-RF نیز همانند روش RF ایستگاه‌های نیر، مشگین شهر و گرمی به ترتیب رتبه‌های اول تا سوم را به خود اختصاص دادند. نکته قابل‌توجه از نتایج به‌دست‌آمده این است که در ایستگاه‌های موردمطالعه، الگوریتم ژنتیک باعث بهبود عملکرد همه مدل‌های مورداستفاده شده است، به‌طوری‌که همه مدل‌های GA-RF با کاهش خطا شدت تابش خورشیدی را با دقت بیشتری برآورد کرده‌اند. شکل (۴) نمودار تغییرات زمانی مقادیر تابش خورشیدی مشاهداتی و پیش‌بینی‌شده با استفاده از برترین مدل‌های RF، GA-RF را در ایستگاه‌های موردمطالعه نشان می‌دهد.

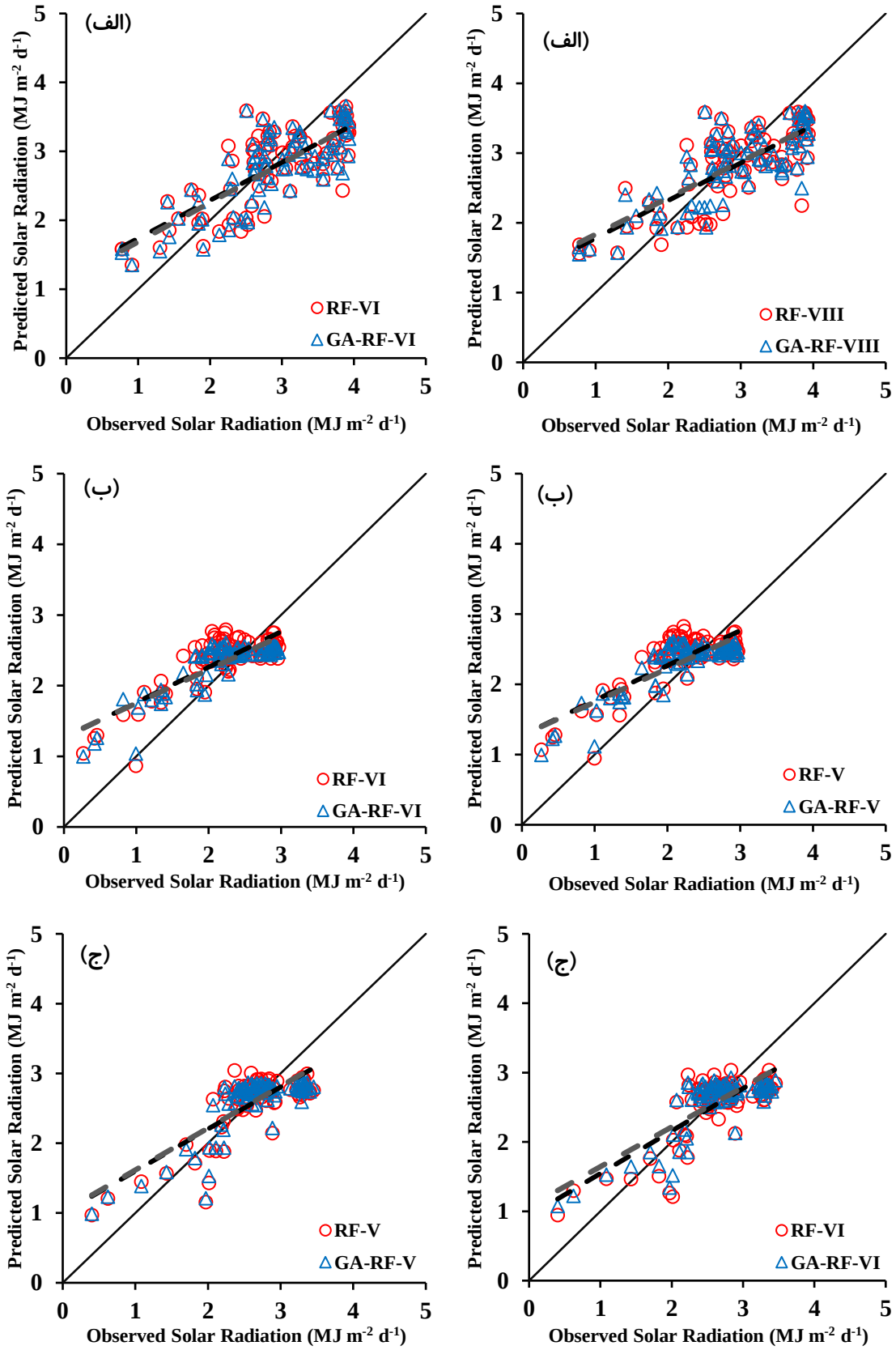




شکل ۴- نمودار تغییرات زمانی شدت تابش خورشیدی مشاهداتی و پیش‌بینی‌شده با استفاده از برترین مدل‌های موردمطالعه: الف) ایستگاه گرمی، مدل‌های RF-VI و GA-RF-VI، ب) ایستگاه گرمی، مدل‌های RF-VIII و GA-RF-VIII، ج) ایستگاه مشگین شهر، مدل‌های RF-VI و GA-RF-VI، د) ایستگاه مشگین شهر، مدل‌های RF-V و GA-RF-V ، ه) ، ایستگاه نیر، مدل‌های RF-V و GA-RF-V و) ایستگاه نیر، مدل‌های RF-VI و GA-RF-VI

Fig. 4 Graph of temporal variations of observed and predicted solar radiation values using the best studied models: a) Germi station, RF-VI and GA-RF-VI models, b) Germi station, RF-VIII and GA-RF-VIII models, c) Meshgin Shahr station, RF-VI and GA-RF-VI models, d) d) Meshgin Shahr station, RF-V and GA-RF-V models, e) Nir station, RF-V and GA-RF-V models, f) Nir station, RF-VI and GA-RF-VI models

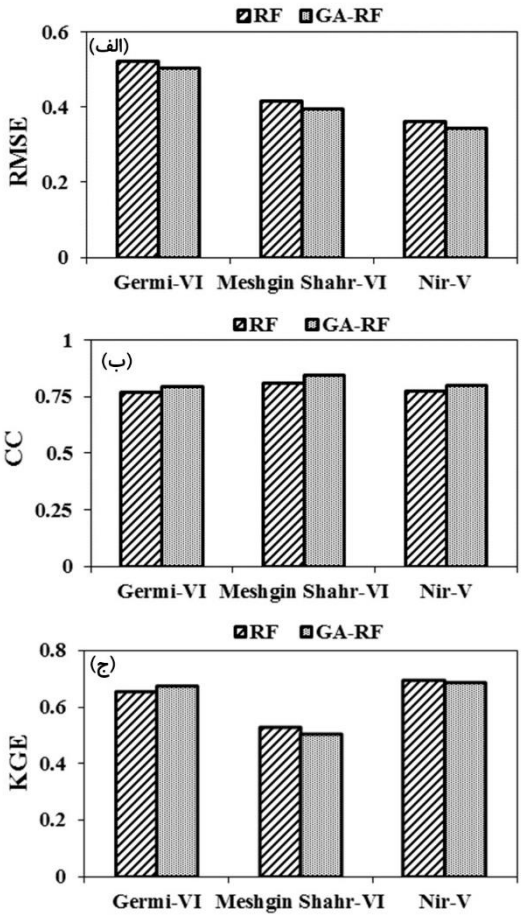
با توجه به نمودارهای تغییرات زمانی مقادیر تابش خورشیدی مشاهداتی و پیش‌بینی‌شده با استفاده از روش‌های مورد مطالعه و با به‌کارگیری پارامترهای هواشناسی به‌عنوان داده‌های ورودی، روند اشاره شده در مورد مدل‌های برتر نیز قابل استنباط است. به عبارت دیگر ، از شکل (۴) نتیجه گرفته می‌شود که در ایستگاه گرمی، مدل‌های RF-VI، RF-VIII، GA-RF-VI و GA-RF-VIII بهترین عملکرد را داشتند به‌طوری‌که پارامترهای ورودی برترین مدل‌ها (سناریو ششم) شامل دمای حداقل، دمای حداکثر، رطوبت نسبی و سرعت باد می‌باشند. در ایستگاه مشگین شهر نیز همانند گرمی مدل‌های سناریو ششم (RF-VI و GA-RF-VI)، بیشترین تطابق را با داده‌های مشاهداتی در مقایسه با سایر مدل‌ها داشتند. همچنین در این ایستگاه مدل‌های RF-V و GA-RF-V از نظر دقت عملکرد رتبه دوم را به خود اختصاص دادند. در ایستگاه نیر نیز مدل‌های RF-V، RF-VI، GA-RF-V و GA-RF-VI مطلوب‌ترین نتایج را ارائه دادند. در این ایستگاه ترکیب سه پارامتر دمای حداقل، دمای حداکثر و رطوبت نسبی دقیق‌ترین مدل‌ها را به ثبت رساند. لذا به‌طورکلی دمای حداقل، دمای حداکثر و رطوبت نسبی به‌عنوان سه پارامتر کلیدی در تخمین شدت تابش خورشیدی در همه ایستگاه‌های مطالعه شده شناخته شدند. همچنین پارامتر سرعت باد نیز به افزایش دقت مدل‌ها در ایستگاه‌های گرمی و مشگین شهر کمک کرد. در ادامه برای نتیجه‌گیری بهتر نمودارهای پراکنش تابش خورشیدی مشاهداتی و پیش‌بینی‌شده با به‌کارگیری همه روش‌های موردمطالعه و با استفاده از پارامترهای هواشناسی در شکل (۵) آورده شده است.



شکل ۵- نمودارهای پراکنش تابش خورشیدی مشاهداتی و پیش‌بینی‌شده با استفاده از برترین مدل‌های موردمطالعه در ایستگاه‌های مختلف: الف) گرمی، ب) مشگین‌شهر و ج) نیر

Fig. 5 Observed and predicted solar radiation distribution diagrams using the best studied models at different stations: a) Germi, b) Meshgin Shahr, and c) Nir

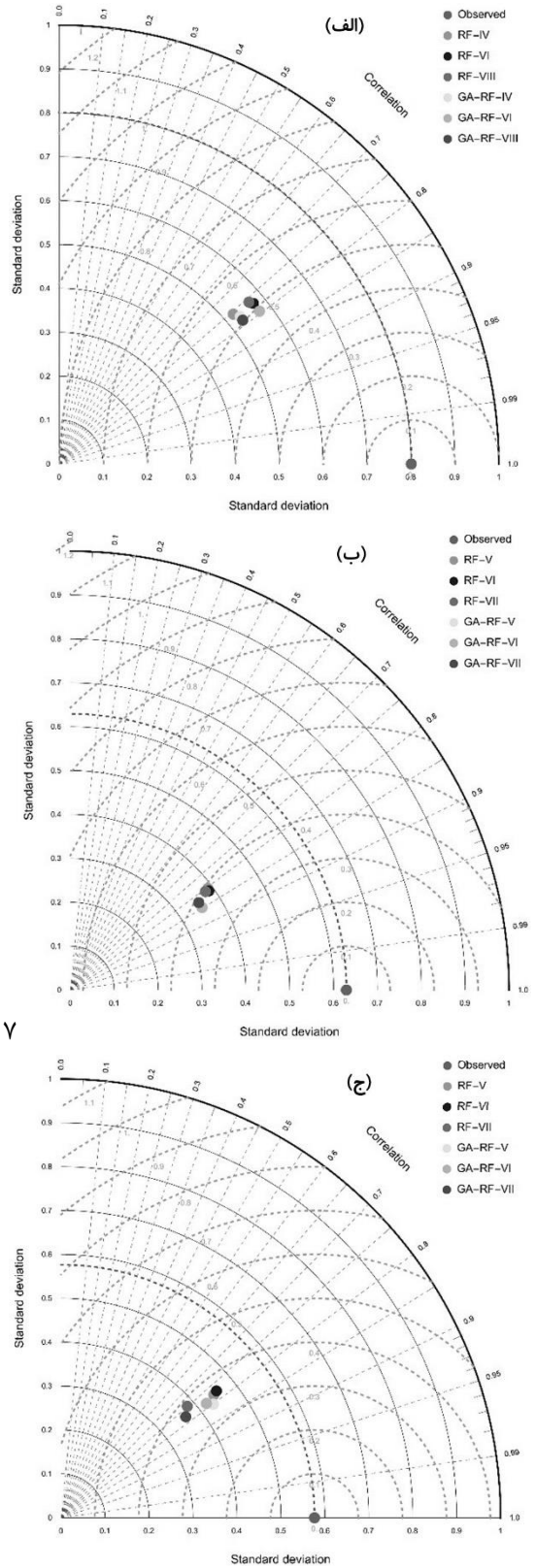
در شکل (۵) مشاهده می‌شود که مدل‌های برتر مذکور در هر ایستگاه، کم‌ترین پراکنش را حول محور نیم‌ساز داشته و دارای بیش‌ترین دقت و کم‌ترین خطا بوده‌اند. در شکل (۶) نیز نمودار میله‌ای شاخص‌های آماری برترین مدل‌های مورد مطالعه در هر ایستگاه ارائه شده است.



شکل ۶- نمودار میله‌ای شاخص‌های آماری برترین مدل‌های مورد مطالعه در هر ایستگاه: الف) جذر میانگین مربعات خطا، ب) ضریب همبستگی، و ج) راندمان کلینگ-گاپتا

Fig. 6 Bar charts of statistical indicators of the best studied models in each station: a) RMSE, b) CC, and c) KGE

با توجه به شکل (۶) ضریب همبستگی سناریوهای برتر به عدد ۱ نزدیک‌تر است که نشان از عملکرد مناسب مدل‌ها دارد. همچنین الگوریتم ژنتیک باعث افزایش این ضریب در هر سه ایستگاه مطالعاتی شده است. مقادیر کوچک‌تر RMSE نیز نشان‌دهنده‌ی برتر بودن آن مدل نسبت به سایر مدل‌هاست و در مدل‌های برتر مقدار این شاخص کمتر از سایر مدل‌ها بوده و تابش خورشیدی با خطای کمتر برآورد شده است. این شاخص نیز بیانگر تأثیرگذاری الگوریتم ژنتیک در کاهش خطای مدل‌ها و بهینه‌سازی نتایج می‌باشد.



شکل ۷- دیاگرام تیلور تابش خورشیدی مشاهداتی و پیش‌بینی‌شده با استفاده از سه سناریو برتر هر ایستگاه: الف) گرمی، ب) مشگین شهر، و ج) نیر

Fig. 7 Taylor diagram of observed and predicted solar radiation using the top three scenarios of each station: a) Germi, b) Meshgin Shahr, and c) Nir

در شاخص KGE نیز عدد ۱ تطابق کامل بین مقادیر مشاهدات و محاسبات را بیان می‌کند و هر چه مقدار این شاخص کمتر باشد نشان از نامناسب بودن مدل دارد که مدل‌های برتر مقادیر KGE بالاتری را به ثبت رسانده‌اند. همچنین، دیاگرام تیلور به منظور بررسی و تحلیل مقادیر انحراف معیار، ضریب همبستگی و جذر میانگین مربعات خطا بین داده‌های مشاهداتی و پیش‌بینی شده توسط سه سناریو برتر روش‌های RF و GA-RF در هر ایستگاه رسم گردید لازم به ذکر است که در دیاگرام تیلور، فاصله شعاعی از نقطه مشاهداتی، نشان دهنده مقدار جذر میانگین مربعات خطای روش‌های مورد مطالعه می‌باشد. همانطوری که از شکل (۷) مشاهده می‌گردد، مدل GA-RF-VI در ایستگاه‌های گرمی و مشگین شهر و مدل GA-RF-V در ایستگاه نیر، فاصله شعاعی کمتری با داده‌های مشاهداتی داشته و بنابراین، دقت بالاتری را در تخمین تابش خورشیدی داشته‌اند. همانطور که اشاره شد پارامترهای دمای حداقل، دمای حداکثر و رطوبت نسبی بیشترین تأثیرگذاری را در تخمین دقیق‌تر تابش خورشیدی در هر سه ایستگاه مطالعاتی داشته‌اند و دیاگرام تیلور نیز برتری مدل‌ها با ورودی پارامترهای مذکور را نشان می‌دهد. در مطالعه مشابهی Ibrahim and Khatib (2017) از روش جنگل تصادفی و ترکیب این روش با الگوریتم کرم شب تاب برای برآورد شدت تابش خورشیدی استفاده کردند و گزارش کردند که روش ترکیبی در مقایسه با سایر روش‌های مورد مقایسه دارای خطای کمتر و دقت بالا بوده و نتایج مطلوبی را ارائه داده است. لذا با توجه به موفقیت‌آمیز بودن روش جنگل تصادفی و مدل‌های ترکیبی این روش در این پژوهش نیز دقت این نوع مدل‌ها در برآورد تابش خورشیدی روزانه مورد آزمایش قرار گرفت و نتایج مناسبی به دست آمد.

۴- نتیجه‌گیری

تابش خورشیدی یکی از پارامترهای کلیدی در بسیاری از زمینه‌های کشاورزی، هیدرولوژی و هواشناسی است و تعیین دقیق مقدار این پارامتر در هر یک از زمینه‌های مذکور اهمیت فراوانی دارد. بر این اساس و در تحقیق حاضر توانایی و دقت روش‌های یادگیری هوشمند جنگل تصادفی و جنگل تصادفی بهینه‌سازی شده با الگوریتم ژنتیک جهت برآورد مقادیر تابش خورشیدی در سه ایستگاه استان اردبیل شامل گرمی، مشگین شهر و نیر مورد سنجش و ارزیابی قرار گرفت. نتایج به‌دست‌آمده با استفاده از پارامترهای آماری با یکدیگر مقایسه شده و مدل‌های برتر انتخاب گردید. در نهایت با استفاده از نتایج نشان داده شد که:

۱- در ایستگاه گرمی و مشگین شهر، مدل‌های ترکیب ششم با پارامترهای ورودی کمینه و بیشینه دما، رطوبت نسبی و سرعت باد مطلوب‌ترین نتایج را ارائه دادند.

۲- در ایستگاه نیر مدل‌های ترکیب پنجم با پارامترهای ورودی کمینه و بیشینه دما و رطوبت نسبی دارای بیشترین دقت و کمترین خطا بودند.

۳- در مقایسه نتایج بین ایستگاه‌ها نیز در هر دو روش RF و GA-RF ایستگاه‌های نیر، مشگین شهر و گرمی به ترتیب از دقت بیشتر به کمتر رتبه‌بندی شدند.

۴- با بررسی مدل‌های RF با GA-RF استنباط گردید که الگوریتم ژنتیک باعث بهبود عملکرد مدل‌ها شده و تأثیر مثبتی بر همه مدل‌ها داشته است.

دسترسی به داده‌ها

داده‌ها حسب درخواست، از طرف نویسنده مسئول از طریق ایمیل قابل ارسال می‌باشد.

تضاد منافع نویسندگان

نویسندگان این مقاله اعلام می‌دارند که هیچ تضاد منافعی در رابطه با نویسندگی و یا انتشار این مقاله ندارند

References

Adnan, M. N. and Islam, M. Z. (2016). Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm. Knowl. Based Syst., 110, 86-97.

Alizamir, M., Kim, S., Kisi, O. and Zounemat-Kermani, M. (2020). A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions. Energy, 197, 117239.

Almorox, J. and Hontoria, C. (2004). Global solar radiation estimation using sunshine duration in Spain. Energy Convers. Manag., 45, 1529-1535.

Bayat, K. and Mirlatifi, S. M. (2009). Estimation of total daily solar radiation using regression

models and artificial neural networks. J. Agri. Sci. Nat. Resour., 16(3), 270-280 [In Persian].

Belaid, S. and Mellit, A. (2016). Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate. Energy Convers. Manag., 118, 105-118.

Benali, L., Notton, G., Fouilloy, A., Voyant, C. and Dizene, R. (2019). Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. Renew. Energy, 132, 871-884.

Breiman, L. (2001). Random forests. Mach. Learn., 45(1), 5-32.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A. and Edwards, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. Geoderm., 239-240(1), 68-83.

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley Longman Publishing Co., Inc.

Holland, J. H. (1992). Genetic algorithms. Sci. Am., 267, 66-72.

Ibrahim, I. A. and Khatib, T. (2017). A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers. Manag., 138, 413-425.

Mossad, E. (2004). Simple new methods to estimate global solar radiation based on meteorological data in Egypt. Atmos. Res., 69, 217-239.

Peter, E. T. and Steven W. R. (1999). An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity and precipitation. Agric. For. Meteorol., 93, 211-228.

Mohammadi, B., Aghashariatmadari, Z. and Moazenzadeh R. (2019). Determination of input variables to estimate solar radiation using entropy theory and principal component analysis. Iran J. Soil Water Res, 50(3), 625-639 [In Persian].

Mousavi, S. M., Mostafavi, E. S. and Jiao, P. (2017). Next generation prediction model for daily solar radiation on horizontal surface using a hybrid neural network and simulated annealing method. Manag., 153, 671-682.

Rao, D. V. S., Premalatha, M. and Naveen, C. (2018). Analysis of different combinations of meteorological parameters in predicting the horizontal global solar radiation with ANN approach: A case study. Renew. Sust. Energ. Rev., 91, 248-258.

Samadianfard, S., Majnooni-Heris, A., Qasem, S. N., Kisi, O., Shamshirband, S. and Chau, K. W. (2019). Daily global solar radiation modeling using datadriven techniques and empirical equations in a semi-arid climate. Eng. Appl. Comput. Fluid Mech., 13(1), 142-157.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atm., 106, 7183-7192.

Wu, L., Huang, G., Fan, J., Zhang, F., Wang, X. and Zeng, W. (2019). Potential of kernel-based nonlinear extension of Arps decline model and gradient boosting with categorical features support for predicting daily global solar radiation in humid regions. Energy Convers. Manag., 183, 280-295.

Yang, K., Huang, G. W. and Tamai, N. (2001). A hybrid model for estimating global solar radiation. Sol. Energy, 70(1), 13-22.

How to cite this paper:
Hashemi, S., Samadianfard, S. and Sadraddini, A. A. (2022). Evaluation of random forest-genetic algorithm hybrid model in estimating daily solar radiation. Environ. Water Eng., 8(3), 636–653. DOI: 10.22034/JEWE.2022.312038.1654