Journal of Advanced Informatics in Water, Soil, and Structure

journal homepage: wss.torbath.ac.ir

Research Article

Daily Solar Radiation Estimation with Gene Expression, Neural Network, and Random Forest Artificial Intelligence Models

Mohammad Porrajab

PhD in Water Engineering, University of Tabriz

ARTICLE INFO

Received date: 07 Sep 2025
Accept date: 24 Nov 2025
Published date: 25 Nov 2025

Keywords:

Artificial intelligence, Gene expression, Neural network, Random Forest, Solar radiation

Abstract

Solar radiation plays a key role in a variety of physical, biological, and chemical processes, including snowmelt, evaporation, plant photosynthesis, and crop production. Solar radiation is also required in biological models for assessing forest fire risk, in hydrological simulation models for natural processes, and in environmental, meteorological, and agricultural research, as well as in atmospheric physics. In the present study, using ten-year data from selected meteorological stations across Iran on a daily scale, artificial intelligence models such as gene expression programming (GEP), neural network (ANN), and random forest (RF) were investigated for solar radiation estimation. By introducing different input combinations to different models, the accuracy of each model was evaluated using scatter index (SI), root mean square error (RMSE), and Nash-Sutcliffe efficiency (NS) criteria. The results show that the neural network model with four inputs, having SI of 0.17, RMSE of 3.25, and NS of 0.71, exhibited higher accuracy compared to other models. Additionally, the random forest model demonstrated satisfactory performance in estimating solar radiation.

Homepage: wss.torbath.ac.ir

*Corresponding Author:
Porrajab, Mohammad
Email: mporrajabali@gmail.com
ORCID: 0009-0009-2051-5785

https://doi.org/10.22048/wss.2026.585524.1030

How to cite this article:

Pourrajab,M . (2025). Daily Solar Radiation Estimation with Gene Expression, Neural Network, and Random Forest Artificial Intelligence Models. *Journal of Advanced Informatics in Water, Soil, and Structure*, *1*(2), 301-311. doi: 10.22048/wss.2026.585524.1030

1- Introduction

Accurate estimation of solar radiation in a region is a crucial issue in hydrological studies, energy engineering, architecture, and agricultural planning. Solar radiation plays a key role in a variety of physical, biological, and chemical processes, including snowmelt and evaporation, plant photosynthesis, and crop production. Solar radiation is also needed in biological models for assessing forest fire risk, in hydrological simulation models for natural processes, and in environmental, meteorological, and agricultural research, as well as in atmospheric physics. However, the main benefit of measuring solar radiation is for simulating and designing solar energy devices. Solar radiation data are crucial for assessing the feasibility of solar energy projects. Estimating solar radiation is the most critical factor in selecting a location and determining the capacity (dimensions) of solar energy projects, as well as selecting the appropriate agricultural crop for a region. However, physical measurement of solar radiation is not possible in all parts of the world due to costs and technological requirements. For this purpose, several methods have been used to predict solar radiation. The two most commonly used approaches are empirical methods and artificial intelligence (AI)-based methods. Both approaches have shown good accuracy in different parts of the world. In order to identify the best method, Navab et al. (2023) conducted a comprehensive review of research articles on solar radiation forecasting to compare different methods. The results of this review showed that AI methods have higher accuracy than empirical methods. Among the empirical models, modified sunshine hour-based models (MSSM) have the highest accuracy, followed by sunshine hour-based models (SSM) and non-sunshine hour-based models (NSM). Although the accuracy of NSM models is slightly lower than that of MSSM and SSM models, these models can provide satisfactory results in conditions where sunshine hour data is not available. Also, the review of the research background confirmed that simple empirical models are also capable of accurate prediction and that increasing the degree of polynomials in empirical models does not necessarily lead to improved results. Among the artificial intelligence methods, artificial neural networks (ANN) and hybrid models have the highest accuracy, followed by support vector machines (SVM) and adaptive neuro-fuzzy inference systems (ANFIS). The increase in efficiency in hybrid models is insignificant, but the complexity of these models requires very advanced programming knowledge. The most important input factors for ANN models are: maximum and minimum temperatures, temperature differences, relative humidity, clarity index, and precipitation. Liu and Scott (2001) concluded that a model that uses only minimum and maximum temperature and rainy day data obtained from 39 stations in Australia obtained the best radiation value. Pauli et al. (2010) used time series processed by a multi-layer neural network called a Prospertron, or MLP, to predict radiation values. The 19-year time series values were obtained from Ajax stations (France). The MLP optimization method provided similar or better results than ARIMA or Bayesian inference, Markov chain or K-nearest neighbor (KNN). In addition, it reduced the error by 6% compared to other common methods such as Markov chain or Bayesian inference. Sohrabi et al. (2023) in a study, conventional multiple linear regression (MLR), multilayer perceptron neural network (MLPNN) and adaptive neuro-fuzzy inference system (ANFIS) methods improved using evolutionary algorithms including particle swarm optimization (PSO), differential evolution (DE), ant colony optimization for continuous domains (ACOR) and genetic algorithm (GA) along with three different ensemble techniques including simple averaging (SAE), weighted averaging (WAE) and neural network ensemble (NNE) were used to model daily solar radiation (DSR) at Belleville station in Illinois, USA. The findings showed that evolutionary algorithms are very effective in improving the performance of the ANFIS model, especially when the number of inputs increases. The results of the complexity analysis showed that although the ANFIS-GA algorithm had the highest accuracy in estimating daily solar radiation in the study area, it was more complex than the ANFIS-PSO and ANFIS-DE algorithms. Also, to deal with missing data of measured radiation, Zarro and Marti (2011) used the Principal Component Analysis (PCA) method from 30 meteorological stations on the Mediterranean coast of Spain. Using multivariate linear regression, it was found that latent variables related to the 4 principal components identified using the PCA method could be obtained with respect to the longitude, latitude and altitude of the station where it is located.

2- Materials and Methods

2-1- Study Area

Meteorological data from the Bushehr, Fasa, Lar, Marvast, and Shahrbabak stations (Figure 1) at a daily scale with a minimum of 10 years were used in the present study. These data include minimum air temperature, maximum air temperature, mean air temperature, extraterrestrial radiation, sunshine hours, and cloudiness degree (Table 1). After initial review and necessary analyses regarding homogeneity and data completion, artificial intelligence models including gene expression programming, neural network, and random forest were developed with specific input combinations. The manner of data utilization by the models was a crucial aspect of this research; various data partitioning methods were examined to select and introduce the best arrangement. Additionally, developing regression-based models considering the key variables governing the solar radiation phenomenon was among the objectives of this study.

Fig. 1. Location of the study stations

Table 1. Information about the study stations

No.	Station	Elevation (m)	Mean Temp. (°C)	ΔT (°C)
1	Bushehr	9	6.30	9.08
2	Fasa	1288	7.44	17.61
3	Lar	792	7.52	16.45
4	Marvast	1546	7.39	17.01
5	Shahrbabak	1834	7.51	16.98

2-2- Gene Expression Programming (GEP) Model

Shiri et al. (2012) performed a sensitivity analysis on the parameters of gene expression programming and introduced the best combination of operators for solar radiation modeling. Therefore, the selection of options at this stage was made based on the results of the aforementioned study using the Gene Xpro software, under the following conditions:

Selection of appropriate fitness function: The fitness function in this method can take various forms (based on absolute error, relative error, etc.). However, based on the studies by Shiri et al. (2012), the root mean square error (RMSE) function is the best fitness function for solar radiation modeling.

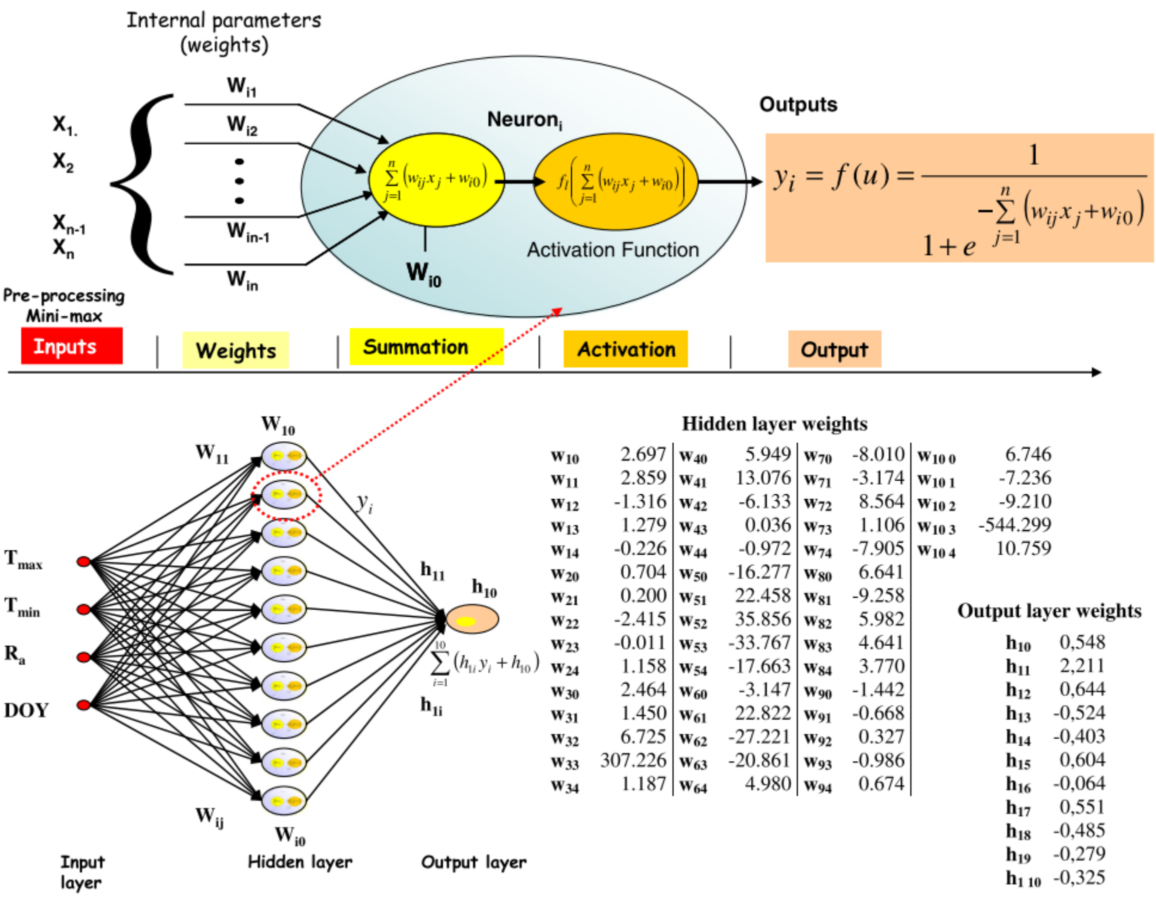
Selection of terminal sets and operator functions for creating chromosomes: In the present study, the terminal set includes meteorological parameters such as air temperature, wind speed, etc. The operator functions used encompass a wide range of different functions, which were selected based on the recommendations of Shiri et al. (2012).

Selection of chromosome architecture: The length of each chromosomal band was set to 8, and the number of genes in each chromosome was set to 3.

Selection of linking function for subprograms: When the subprograms of gene expression programming are of algebraic type (as in the present study), the addition linking function is used to establish connections between subprograms. The final stage of modeling includes the selection of operators.

2-3- Neural Network (ANN) Model

The architecture of neural networks was selected based on the algorithms available in the IPS module of the STATISTICA software. This module replaces the conventional heuristic process in network design with search algorithms that determine the number of input variables (if necessary), the number of hidden neurons, and other important factors of the network structure. The basis of this module is empirical sampling of different network structures, automatically determining the degree of complexity using a wide range of algorithms. In the present study, the number of hidden layer neurons was determined automatically. Accordingly, the multilayer perceptron neural network with hidden neurons ranging from 1 to 15 was evaluated for each input combination, and the best possible structures were selected. The available modeling data were divided into two or more parts based on the data management system used and introduced to the models as training, testing, and validation data. Figure 2 shows the schematic structure of an artificial neural network with four input variables along with the schematic process of the hidden layer.



Hidden layer weights

w_{10}	2.697	w_{40}	5.949	w_{70}	-8.010	$w_{10\,0}$	6.746
w_{11}	2.859	w_{41}	13.076	w_{71}	-3.174	$w_{10\,1}$	-7.236
w_{12}	-1.316	w_{42}	-6.133	w_{72}	8.564	$w_{10\,2}$	-9.210
w_{13}	1.279	w_{43}	0.036	w_{73}	1.106	$w_{10\,3}$	-544.299
w_{14}	-0.226	w_{44}	-0.972	w_{74}	-7.905	$w_{10\,4}$	10.759
w_{20}	0.704	w_{50}	-16.277	w_{80}	6.641		
w_{21}	0.200	w_{51}	22.458	w_{81}	-9.258		
w_{22}	-2.415	w_{52}	35.856	w_{82}	5.982		
w_{23}	-0.011	w_{53}	-33.767	w_{83}	4.641		
w_{24}	1.158	w_{54}	-17.663	w_{84}	3.770		
w_{30}	2.464	w_{60}	-3.147	w_{90}	-1.442		
w_{31}	1.450	w_{61}	22.822	w_{91}	-0.668		
w_{32}	6.725	w_{62}	-27.221	w_{92}	0.327		
w_{33}	307.226	w_{63}	-20.861	w_{93}	-0.986		
w_{34}	1.187	w_{64}	4.980	w_{94}	0.674		

Output layer weights

h_{10}	0,548
h_{11}	2,211
h_{12}	0,644
h_{13}	-0,524
h_{14}	-0,403
h_{15}	0,604
h_{16}	-0,064
h_{17}	0,551
h_{18}	-0,485
h_{19}	-0,279
$h_{1\,10}$	-0,325

Fig. 2. Schematic structure of an artificial neural network including the network training process and middle layer operations for a model with 4 input components

2-4- Random Forest (RF) Algorithm

One of the efficient tools used in problems related to estimating target variables or classifying patterns is the decision tree. A decision tree divides the input space into sets of separate regions and assigns a response value to each region. In the simplest case, this response in regression problems can be determined based on the average of target values associated with patterns placed in each region. Generally, a single decision tree is prone to overfitting and has limited generalization capability. When constructing a decision tree, a small change in the training patterns can cause fundamental changes in the tree structure (Quinlan, 1986). Combining decision trees, known as ensemble methods, prevents this problem. An ensemble consists of multiple learners called base learners. Base learners are typically constructed from training data using a base learning algorithm, which can be a decision tree, artificial neural network, or other learning algorithms. The generalization capability of an ensemble is often stronger than that of individual base learners. Ensemble methods are widely accepted primarily due to their ability to strengthen weak learners (Schapire, 1990).

2-5- Input Combinations for Intelligent Models

Considering the components affecting the reference evapotranspiration phenomenon, combinations of input parameters including Ra (extraterrestrial radiation), Tmax (maximum temperature), Tmin (minimum temperature), DOY (day of year), and ICSKY (clear sky index) were considered as input combinations for the artificial intelligence models (Table 2).

Table 2. Input parameters of the intelligent models

Model	Input Parameters
I	Ra
II	Ra, Tmax
III	Ra, Tmax, Tmin
IV	Ra, Tmax, Tmin, DOY
V	Ra, Tmax, Tmin, DOY, ICSKY

2-6- Evaluation Criteria

In this study, three evaluation criteria including scatter index (SI), root mean square error (RMSE), and Nash-Sutcliffe efficiency (NS) were used to assess model accuracy. The equations for each are as follows:

$$SI = \frac{RMSE}{\overline{Rs}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Rs_{model} - Rs_{measured}\right)^2}}{\overline{Rs}} \quad \text{Eq (1)}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Rs_{model} - Rs_{measured}\right)^2} \quad \text{Eq (2)}$$

$$NS = 1 - \frac{\sum_{i=1}^{N}\left(Rs_{model} - Rs_{measured}\right)^2}{\sum_{i=1}^{N}\left(Rs_{measured} - \overline{Rs}\right)^2} \quad \text{Eq (3)}$$

In the above equations, Rs_{model} is the modeled radiation, $Rs_{measured}$ is the observed radiation, $\overline{R}_s$ is the mean observed radiation, and N is the total number of data points.

3- Results and Discussion

3-1- Gene Expression Programming (GEP) Models

In the overall comparison among GEP models, models V and IV showed higher accuracy at the Bushehr station. At the Fasa station, model IV produced the best results. At the Lar station, model III was the best. At the Marvast station, model V yielded better results, and at the Shahrbabak station, model II performed best. Calibration with a single meteorological station at the Bushehr station, in the most accurate model, achieved SI of 0.19, RMSE of 3.64, and NS of 0.64. At the Fasa station, with SI of 0.15, RMSE of 3.01, and NS of 0.74, model IV was the best calibrated model. At the Lar station, with SI of 0.12, RMSE of 2.55, and NS of 0.78, model III was the best. At the Marvast station, with SI of 0.15, RMSE of 2.94, and NS of 0.75, model V was the best calibrated model. At the Shahrbabak station, with SI of 0.17, RMSE of 3.50, and NS of 0.68, model V was the best calibrated model.

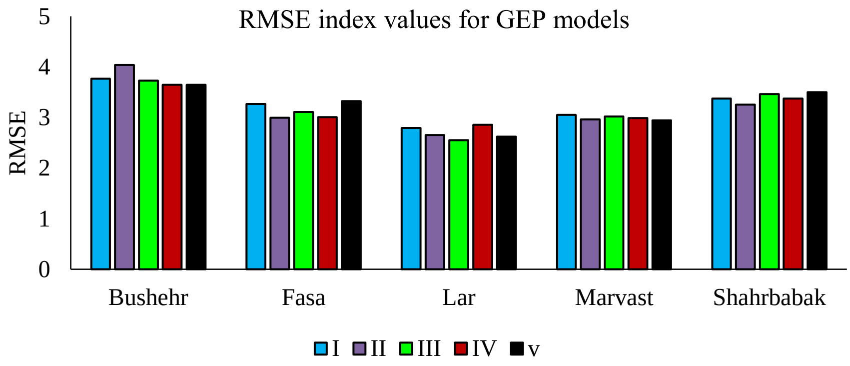


Fig. 3. RMSE values for GEP models

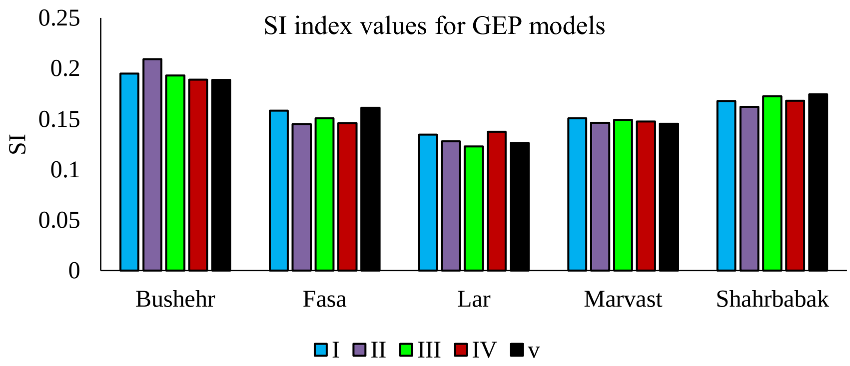


Fig. 4. SI values for GEP models

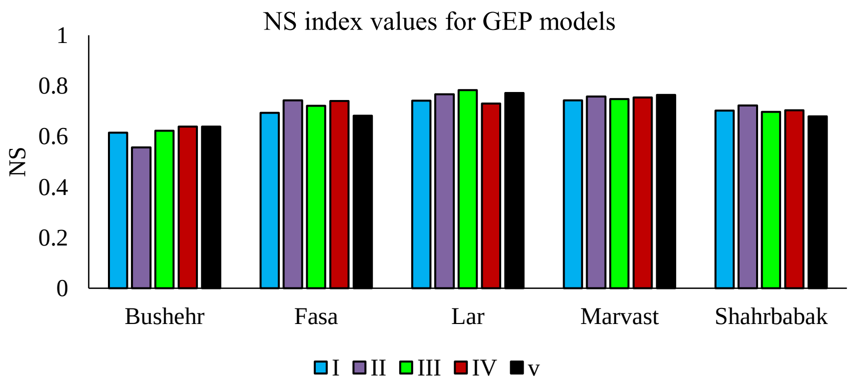


Fig. 5. NS values for GEP models

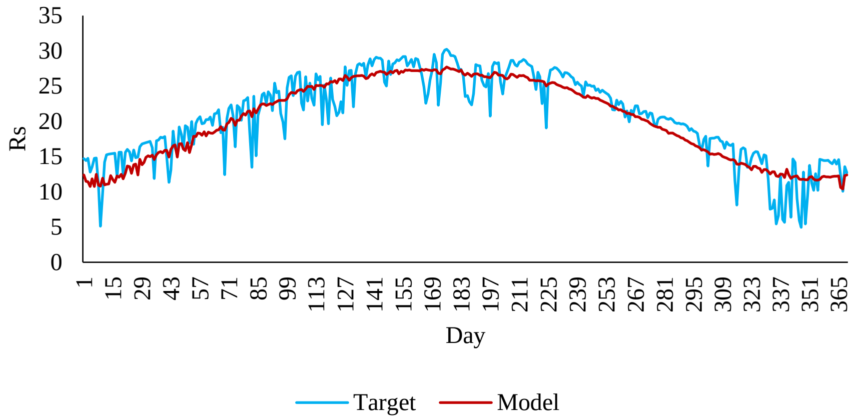


Fig. 6. Comparison of computed solar radiation values with the most accurate GEP model at Marvast station versus measured values as daily average (MJ/m²/day)

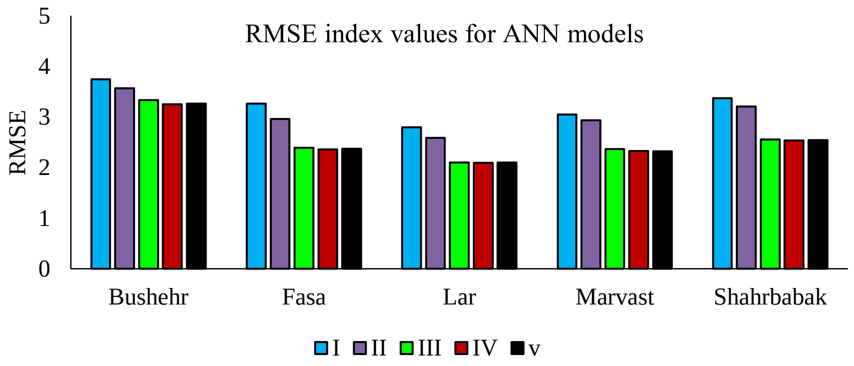


Fig. 7. RMSE values for ANN models

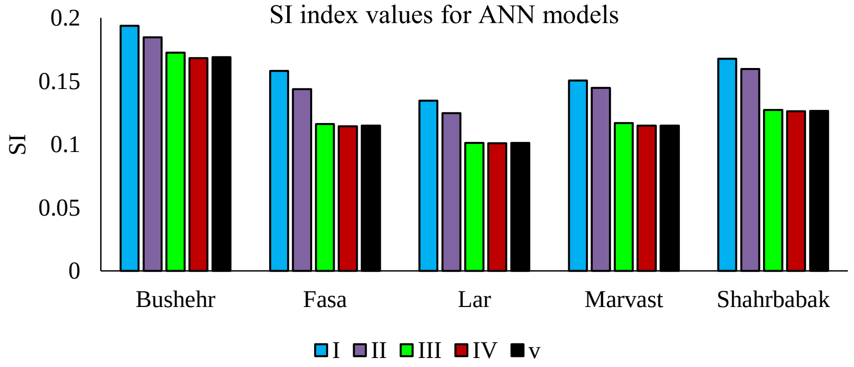


Fig. 8. SI values for ANN models

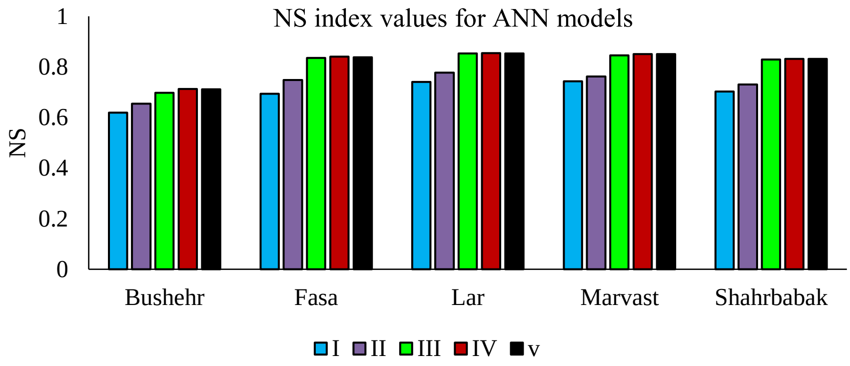


Fig. 9. NS values for ANN models

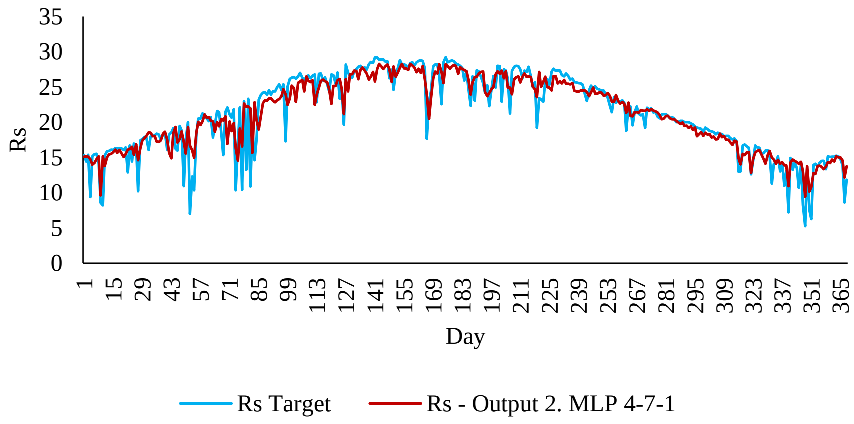


Fig. 10. Comparison of computed solar radiation values with the most accurate ANN model at Lar station versus measured values (MJ/m²/day)

3-2- Neural Network (ANN) Models

In the overall comparison among ANN models, model IV showed higher accuracy at the Bushehr station. At the Fasa station, model IV produced the best results. At the Lar station, model V was the best. At the Marvast station, model V yielded better results, and at the Shahrbabak station, model IV performed best. Calibration with a single meteorological station at the Bushehr station, in the most accurate model, achieved SI of 0.17, RMSE of 3.25, and NS of 0.71. At the Fasa station, with SI of 0.11, RMSE of 2.36, and NS of 0.84, model IV was the best calibrated model. At the Lar station, with SI of 0.10, RMSE of 2.10, and NS of 0.85, model V was the best. At the Marvast station, with SI of 0.11, RMSE of 2.32, and NS of 0.85, model V was the best calibrated model. At the Shahrbabak station, with SI of 0.13, RMSE of 2.54, and NS of 0.83, model IV was the best calibrated model.

3-3- Random Forest (RF) Models

In the overall comparison among RF models, model IV showed higher accuracy at the Bushehr station. At the Fasa station, model III produced the best results. At the Lar station, model III was the best. At the Marvast station, model IV yielded better results, and at the Shahrbabak station, model V performed best. Calibration with a single meteorological station at the Bushehr station, in the most accurate model, achieved SI of 0.17, RMSE of 3.27, and NS of 0.71. At the Fasa station, with SI of 0.12, RMSE of 2.54, and NS of 0.81, model III was the best calibrated model. At the Lar station, with SI of 0.11, RMSE of 2.29, and NS of 0.83, model III was the best. At the Marvast station, with SI of 0.12, RMSE of 2.52, and NS of 0.82, model IV was the best calibrated model. At the Shahrbabak station, with SI of 0.13, RMSE of 2.68, and NS of 0.81, model V was the best calibrated model.

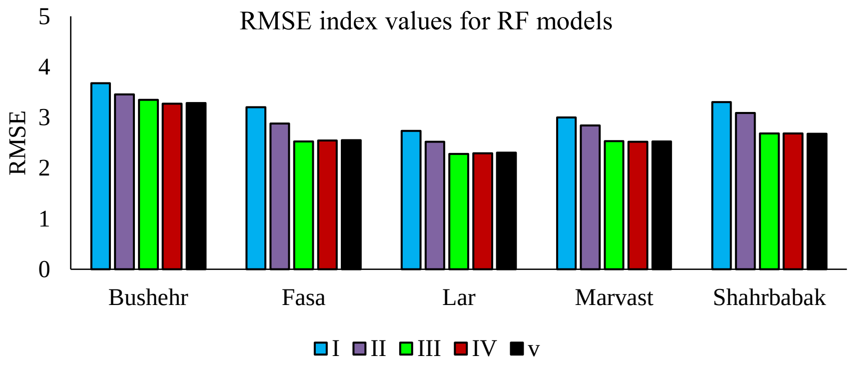


Fig. 11. RMSE values for RF models

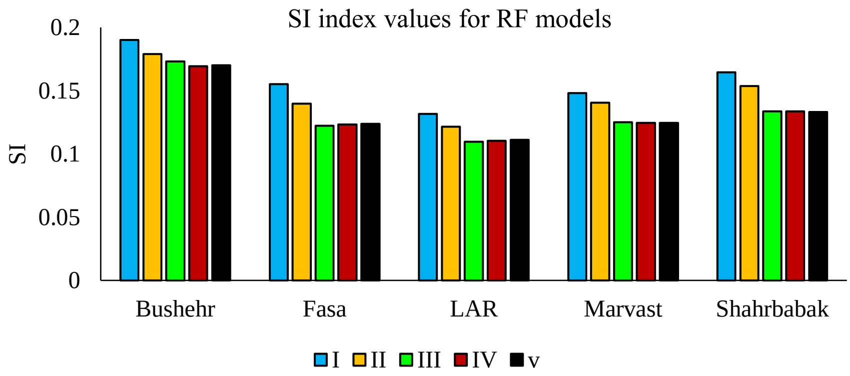


Fig. 12. SI values for RF models

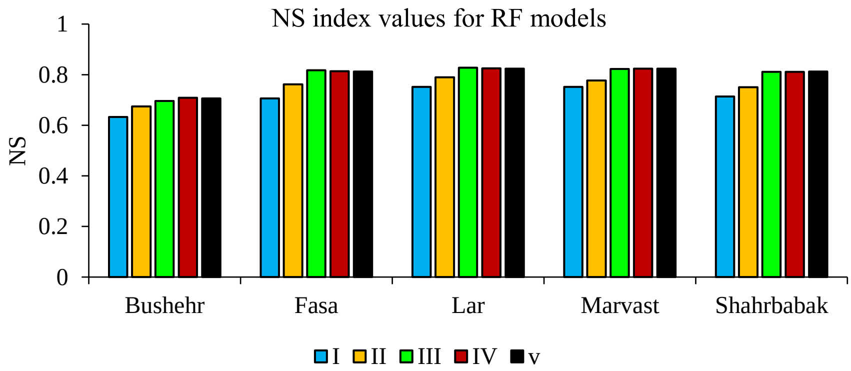


Fig. 13. NS values for RF models

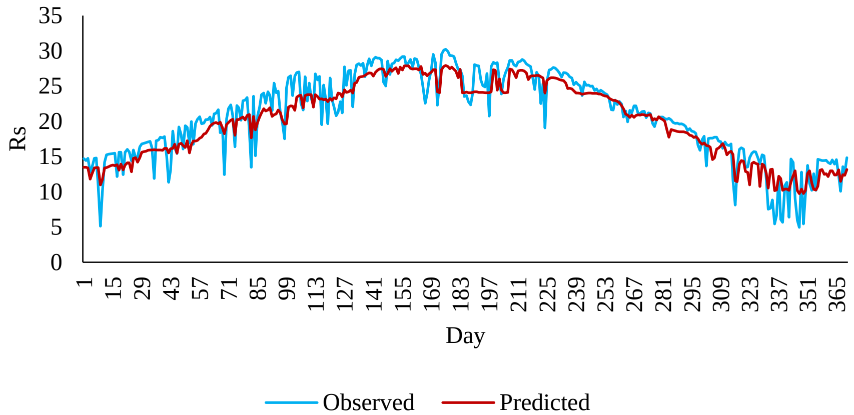


Fig. 14. Comparison of computed solar radiation values with the most accurate RF model at Marvast station versus measured values (MJ/m²/day)

Table 3. RMSE, SI, and NS values for GEP, ANN, and RF models at different stations with different input combinations

Station	**Inputs**	**RMSE**	**SI**	**NS**	**RMSE**	**SI**	**NS**	**RMSE**	**SI**	**NS**
		GEP			**ANN**			**RF**		
Bushehr	Ra	3.76	0.19	0.61	3.74	0.19	0.62	3.67	0.19	0.63
	Ra, Tmax	4.04	0.21	0.56	3.57	0.18	0.65	3.46	0.18	0.68
	Ra, Tmax, Tmin	3.73	0.19	0.62	3.34	0.17	0.70	3.35	0.17	0.70
	Ra, Tmax, Tmin, DOY	3.65	0.19	0.64	3.25	0.17	0.71	3.27	0.17	0.71
	Ra, Tmax, Tmin, DOY, ICSKY	3.64	0.19	0.64	3.26	0.17	0.71	3.29	0.17	0.71
Fasa	Ra	3.26	0.16	0.69	3.26	0.16	0.69	3.20	0.16	0.71
	Ra, Tmax	2.99	0.15	0.74	2.96	0.14	0.75	2.88	0.14	0.76
	Ra, Tmax, Tmin	3.11	0.15	0.72	2.39	0.12	0.84	2.52	0.12	0.82
	Ra, Tmax, Tmin, DOY	3.01	0.15	0.74	2.36	0.11	0.84	2.54	0.12	0.81
	Ra, Tmax, Tmin, DOY, ICSKY	3.32	0.16	0.68	2.37	0.11	0.84	2.55	0.12	0.81
Lar	Ra	2.79	0.13	0.74	2.79	0.13	0.74	2.73	0.13	0.75
	Ra, Tmax	2.65	0.13	0.77	2.59	0.12	0.78	2.52	0.12	0.79
	Ra, Tmax, Tmin	2.55	0.12	0.78	2.10	0.10	0.85	2.28	0.11	0.83
	Ra, Tmax, Tmin, DOY	2.85	0.14	0.73	2.10	0.10	0.85	2.29	0.11	0.83
	Ra, Tmax, Tmin, DOY, ICSKY	2.62	0.13	0.77	2.10	0.10	0.85	2.31	0.11	0.82
Marvast	Ra	3.05	0.15	0.74	3.05	0.15	0.74	3.00	0.15	0.75
	Ra, Tmax	2.96	0.15	0.76	2.93	0.14	0.76	2.84	0.14	0.78
	Ra, Tmax, Tmin	3.02	0.15	0.75	2.37	0.12	0.84	2.53	0.12	0.82
	Ra, Tmax, Tmin, DOY	2.98	0.15	0.75	2.33	0.11	0.85	2.52	0.12	0.82
	Ra, Tmax, Tmin, DOY, ICSKY	2.94	0.15	0.76	2.32	0.11	0.85	2.53	0.12	0.82
Shahrbabak	Ra	3.37	0.17	0.70	3.37	0.17	0.70	3.31	0.16	0.71
	Ra, Tmax	3.25	0.16	0.72	3.21	0.16	0.73	3.09	0.15	0.75
	Ra, Tmax, Tmin	3.46	0.17	0.70	2.56	0.13	0.83	2.69	0.13	0.81
	Ra, Tmax, Tmin, DOY	3.37	0.17	0.70	2.53	0.13	0.83	2.69	0.13	0.81
	Ra, Tmax, Tmin, DOY, ICSKY	3.50	0.17	0.68	2.54	0.13	0.83	2.68	0.13	0.81

4- Conclusions

In the present study, five solar radiation models were compared for five meteorological stations across Iran with a twelve-year statistical period. The models used in this research include two empirical solar radiation models and three intelligent models obtained using gene expression programming, neural network, and random forest methods. The overall conclusions of this study can be summarized as follows: AI-based solar radiation estimation models exhibit high accuracy compared to empirical methods. These results hold true for both the initial accuracy of the models and the calibrated models. Calibration has further improved the accuracy of most models. The intelligent models, especially the neural network and random forest, generally showed higher accuracy than other models. The data management method used has enabled more accurate estimation of solar radiation, which is particularly important for stations facing data limitations. In the study area, temperature and energy components have greater importance than other factors affecting solar radiation. It is also recommended that the accuracy of other artificial intelligence and empirical models for estimating solar radiation in the study area be evaluated. The usability of data from arid regions for estimating solar radiation in semi-arid regions be investigated, and a study similar to the present study be conducted for other regions, especially humid regions of the country.

Acknowledgments

The author would like to acknowledge the Iran Meteorological Organization for providing the meteorological data used in this research.

Funding Sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The author declares that there is no conflict of interest to disclose regarding the submission of this paper. The author has no financial or personal relationships that could potentially bias the content or findings of this research.

Authors Contribution Statement

The sole author conceived the study, performed the data analysis, developed the models, and wrote the manuscript.

References

Liu, D.L., & Scott, B.J. (2001). Estimation of solar radiation in Australia from rainfall and temperature observations. *Agricultural and Forest Meteorology, 106*(1), 41-59.

Nawab, F. and Sopian, K. 2023. Solar irradiation prediction using empirical and artificial intelligence methods: A comparative review. *Heliyon*, 9(2), pp.2405-2410.

Paoli, C., Voyant, C., Muselli, M., & Nivet, M.L. (2010). Forecasting of preprocessed daily solar radiation time series using neural networks. *Solar Energy, 84*(12), 2146-2160.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106.

Reikard, G. (2009). Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy, 83*(3), 342-349.

Sohrabi, F. Golabi, M.R. 2023. Daily solar radiation estimation in Belleville station, Illinois, using ensemble artificial intelligence approaches. *Engineering Applications of Artificial Intelligence,* 120(3), pp.342-349.

Schapire, R.E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197-227.

Shiri, J., Kisi, O., Yoon, H., Kim, N.K., & Hosseinzadeh, M. (2012). Sensitivity analysis of gene expression programming parameters in solar radiation modeling. *Computers and Electronics in Agriculture, 89*, 159-165.

Zarzo, M., & Marti, P. (2011). Modeling the variability of solar radiation data among weather stations by means of principal components analysis. *Applied Energy, 88*(8), 2775-2784.