DOI: https://doi.org/10.36253/ijam-2815

Solar Radiation Prediction in Semi-Arid Regions: A Machine Learning Approach and Comprehensive Evaluation in Gadarif, Sudan

Amar Ali Adam Hamad [a], Abdelkarem M. Adam [a,b], Yuan Zheng [c]

[a] College of Resources and Environment, Shanxi Agricultural University, Taiyuan, 030031, China

[b] *College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing, 210098, China*

[c] Renewable Energy Power Generation Engineering Research, MOE; School of Water Resources and Hydropower, Hohai University, 210098 Nanjing, China.

Corresponding author: Abdelkarem M. Adam, abdoadam7878@gmail.com

Solar radiation (H) is a critical factor in Earth's surface processes, influencing climate, ecosystems, agriculture, and energy fluxes. Accurate prediction of daily H is essential for advancing solar power as a sustainable energy source. This study evaluates the effectiveness of machine learning (ML) models-support vector regression (SVR), extreme gradient boosting (XGBoost), boosted regression forest (BRF), and k-nearest neighbors (K-NN)-in predicting daily H in Gadarif, Sudan, a semi-arid region with limited prior research on solar radiation. The models were developed using daily climatic variables, including temperature and a binary precipitation variable (P_t) to account for cloud cover effects. The dataset was split into training (80%) and testing (20%) subsets, with model performance evaluated using key metrics: coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). BRF achieved the best performance with an R^2 of 0.963 and RMSE of 4.38 (MJ m^{-2} d^{-1}) during training. However, model performance decreased during testing, with XGBoost and K-NN showing higher error margins. Including Pt improved the models' ability to account for cloud cover effects, particularly on overcast days. Despite these improvements, challenges

remained in predicting H under extreme climatic conditions, highlighting the need for more advanced approaches. These findings suggest that ML models can be effectively adapted for H prediction in other semi-arid and arid regions. The results underscore the importance of considering precipitation and cloud cover in H predictions, which is crucial for optimizing solar energy systems and enhancing agricultural planning.

Key words: Solar radiation; Machine learning; Renewable energy; Semi-arid climate; Comprehensive evaluation

HIGHLIGHTS

- Machine learning models, including SVR, XGBoost, BRF, and K-NN, were applied to predict daily solar radiation (H).
- BRF outperformed the other models, achieving the highest performance with an R^2 of 0.963 and RMSE of 4.38 (MJ m^{-2} d^{-1}) during training.
- Incorporating a precipitation variable (P_t) improved the models' accuracy by accounting for cloud cover effects.
- Testing showed a performance drop, though BRF maintained strong generalization, needing refinement for extreme conditions.
- The methodology, applied in Gadarif, Sudan, can be adapted for other semi-arid and arid regions for solar energy optimization.

Nomenclature

Parameters		Abbreviation	
C	penalty parameter of the error	ANN	Artificial Neural Networks
H	global solar radiation (MJ m^{-2} day^{-1})	MLP	Multi-layer Perceptron
H_0	extra-terrestrial solar radiation (MJ m^{-2} day^{-1})	SVM	Support Vector Machine
K	kernel function	XGBoost	Extreme Gradient Boosting
I	loss function	ANFIS	adaptive neuro-fuzzy inference system
n	number of observations	RF	Random Forest

N	sunshine duration	AI	Artificial Intelligence
ΔT	diurnal temperature range (°C)	BRF	Boosted Regression Forests
Pt	transformed precipitation	ML	Machine Learning
T_{max}	daily maximum temperature (°C)		
T_{min}	daily minimum temperature (°C)		
X_{min}	minimum observed value in the dataset		
X_{max}	maximum observed value in the dataset		
X_{mean}	mean observed value in the dataset		
C_s	Skewness coefficient		
SD	Stander deviation		
C_k	Kurtosis coefficient		
φ	higher-dimensional feature space		
ω	weights vector		
ε	tube size		
λ	regularization parameter		
γ	minimum loss		
Ω	regularization term		

Constants

a, b, and c, empirical coefficients

1. INTRODUCTION

Solar radiation (H) plays a crucial role in Earth's surface processes, influencing climate systems, hydrology, and ecosystems (Caldwell, M.M., Bornman, J.F., Ballaré, 2007). Its accurate estimation is particularly critical in semi-arid regions where environmental and agricultural systems heavily depend on it. Solar radiation directly impacts photosynthesis, making it a vital variable in crop modeling, where agronomic applications are essential for

optimizing yield predictions (Holzman et al., 2018). Precise H forecasts are essential for improving agricultural planning and water resource management, especially in regions with limited resources.

This study addresses the gap in H prediction for semi-arid regions, focusing on Gadarif, Sudan, by employing advanced machine learning (ML) techniques support vector machines (SVM), extreme gradient boosting (XGBoost), boosted regression forest (BRF), and k-nearest neighbors (K-NN). While traditional studies have focused on temperate climates using statistical models, this research applies ML models to capture complex, non-linear interactions in semi-arid conditions. SVM and XGBoost were selected for their robustness and ability to generalize well across varying datasets, BRF for its ensemble method, which reduces bias and variance, and K-NN for its effectiveness in modeling local relationships. By utilizing a daily temporal scale, this study provides precise short-term H forecasts, enhancing prediction accuracy for agricultural applications in resource-challenged regions like Gadarif.

ML approaches have been increasingly applied to estimate H in various climates. (Wang et al., 2016) conducted a comparative study in China, estimating daily H using models such as multilayer perceptron (MLP), radial basis function (RBF), and generalized regression neural networks (GRNN). The study found that GRNN underperformed compared to MLP and RBF, highlighting the need for more robust models in H prediction. Similarly, (Belmahdi et al., 2020) forecasted daily H one month ahead using ARIMA and ARMA models, with ARIMA demonstrating superior accuracy over a persistence model.

Most previous studies focused on a specific timescale or component of H. For instance, (Belmahdi et al., 2022) introduced a new optimization method to predict hourly H, comparing several models, including feed-forward backpropagation (FFBP), ARIMA, k-NN, and SVM. FFBP and ARIMA models exhibited the highest accuracy, as confirmed by regression plots under clear-sky conditions.

(Fan et al., 2018a) employed SVM and extreme gradient boosting (EGB) models to predict H in humid regions with limited data. They found that SVM outperformed EGB and traditional empirical models in terms of prediction stability. Similarly, (Belaid and Mellit, 2016) explored the use of SVM and artificial neural networks (ANN) for predicting daily and monthly H, concluding that SVM produced better correlations between predicted and observed values at both timescales.

Geographical and meteorological data have also been extensively utilized in H modeling. For example, (Sözen et al., 2008) employed an artificial neural network (ANN) model to estimate H in Turkey, achieving highly accurate predictions. In Algeria, (Mellit et al., 2008) applied both ANN and adaptive neuro-fuzzy inference system (ANFIS) models, also producing reliable results for H estimation. (Chen et al., 2011) found that SVM were dependable model for H predictions across multiple stations, while (Ahmed and Adam, 2013) demonstrated that ANN models outperformed empirical models in predicting H in Qena, Egypt, achieving higher correlations between predicted and observed values.

While these studies have significantly advanced the field of H prediction, they often lack comprehensive evaluations of model performance in semi-arid climates. Furthermore, few studies have incorporated precipitation data to account for cloud cover, a critical factor affecting H in these regions. (He et al., 2020) highlighted the variability of H across different geographic regions; however, the unique climatic conditions of semi-arid areas like Gadarif remain underexplored.

The primary objective of this study is to predict daily H in Gadarif, Sudan, using advanced ML models. This is the first study to apply the Boosted Regression Forest (BRF) model for H prediction in this region. Additionally, the study incorporates precipitation data as a key variable to account for the influence of cloud cover on H, which an aspect that has not been extensively explored.

The novelty of this research lies in its application of BRF, an underutilized yet powerful ensemble method, for H estimation in semi-arid regions. By integrating precipitation as a binary variable, the study enhances the accuracy of solar radiation predictions and agricultural modeling, providing new insights into the interaction between precipitation, cloud cover, and H in Gadarif. This tailored approach fills gaps in existing research and contributes to improving forecasting in resource-constrained environments.

2. MATERIALS AND METHODS

2.1. Study area and data collection

Figure 1 illustrates the study area, Gadarif, located in eastern Sudan, which experiences a hot semi-arid climate (BSh according to the Köppen-Geiger classification). This region faces significant agricultural challenges due to harsh environmental conditions, including high temperatures, erratic rainfall, and limited water resources. These factors contribute to

substantial yield variability and increased vulnerability to drought and heat stress. Moreover, the scarcity of reliable water sources and the fluctuating solar radiation levels emphasize the need for accurate solar radiation predictions, which are essential for effective water management and crop planning.

The study area is primarily agricultural, with sorghum and sesame as the main crops. These crops depend on consistent solar radiation (H) and sufficient water availability, emphasizing the importance of this study for local agricultural management.

Daily meteorological data were collected from 2010 to 2022, covering a 12-year period. The data were obtained from the Sudan Meteorological Authority (SMA) at the Gadarif weather station, a well-established station that records key climatic variables. Equipped with modern weather instrumentation, the station measures H, temperature, and precipitation. This data were supplemented with satellite-derived information from NASA's POWER Data Access Viewer, ensuring the completeness and accuracy of the dataset used in this study. The combined dataset includes daily observations of H, temperature (T_{max}, T_{min}), and precipitation (P_t), which were essential inputs for the ML models.

These data were recorded at daily intervals, which enabling for high -resolution training of the model. However, in scenarios where daily data are not available, the model can be adapted by means of a weekly or monthly average, such as low-ceiling input. In addition, proxy dataset from satellite sources, such as MODIS and CHIRPS precipitation estimate, can serve as a viable alternative to support Modi's estimate and application.

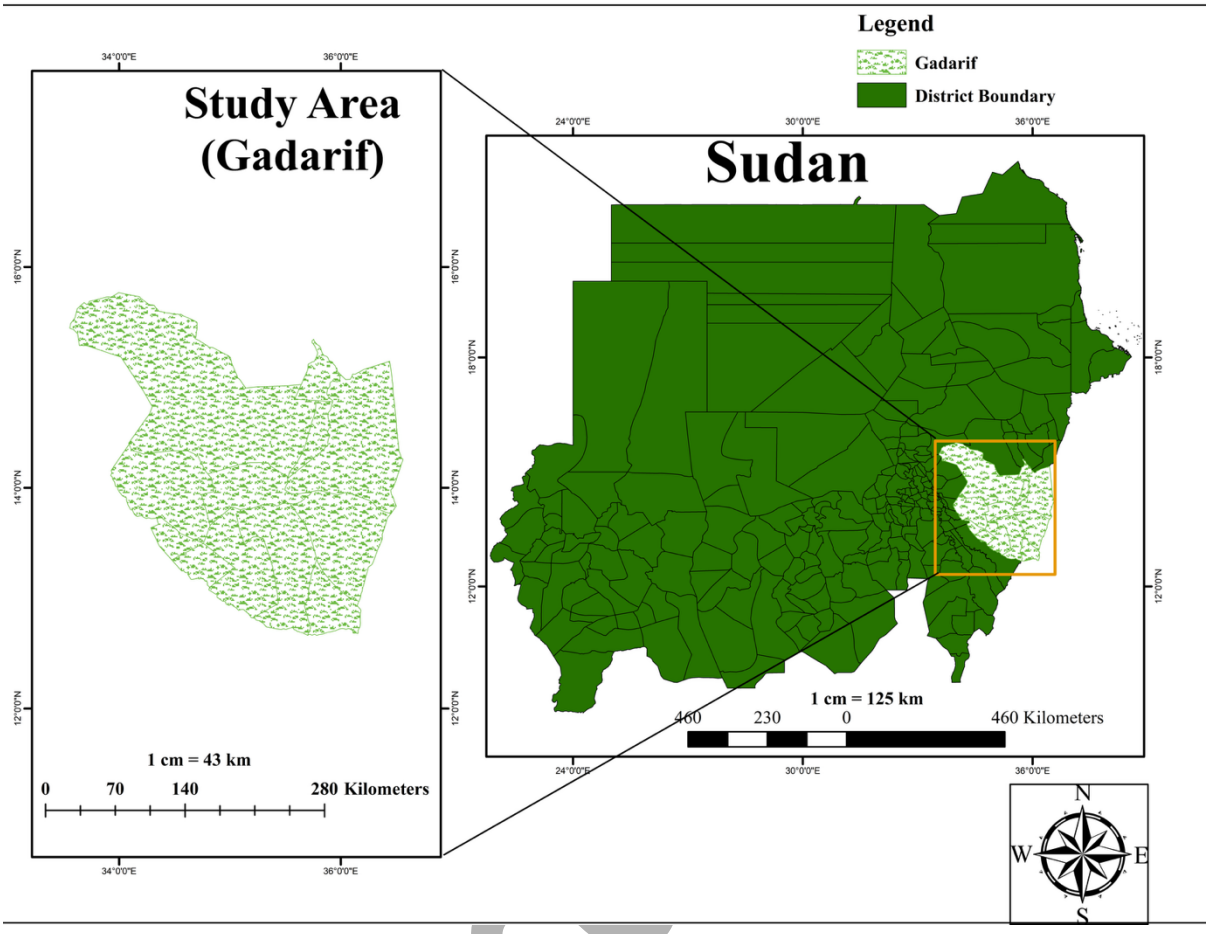


Figure 1. Geographical location of the meteorological station in semi-arid climate region in Sudan.

2.2. Machine learning models

The dataset comprises 4,380 daily records collected over a 12-year period (2010–2022). For the purposes of model development, the data were divided into a training set (80%) and a testing set (20%). The dataset includes daily measurements of H, extraterrestrial radiation (H_0), T_{max}, T_{mean}, T_{min}, and P_t. These variables were used as inputs for the machine learning models to predict H more accurately.

2.2.1. Support vector machines (SVM)

The support vector machine (SVM) model, developed by Vapnik and outlined in (Vapnik, 2006), stands as a widely used supervised AI model for tasks such as data analysis and pattern recognition, particularly in applications involving regression and prediction. The SVM algorithm functions by predicting regression through a series of kernel functions. To ensure methodological clarity, it is important to explain the kernel function in support vector machines (SVM). The kernel function defines the operations and transformations applied to the input

data. By addressing the non-linear characteristics of SVM and the approaches it utilizes to define appropriate decision boundaries, this explanation enhances the understanding of SVM. This understanding, in consequence, empowers them to make well-informed decisions when applying SVM to diverse datasets (Wu, 1999; Tay and Cao, 2001).

The SVM algorithm expresses the approximated function as depicted in the subsequent equation:

$$F(x) = \omega.\varphi(x) + b. \quad (1)$$

In this equation, φ (*x*) denotes the transformation of the input vector x into a higher-dimensional feature space. The parameters ω and b represent the weight vector and a threshold, respectively. These values can be obtained by reducing the regularized risk function, as defined below:

$$R_{SVM}(C) = C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i) + \frac{1}{2} w \|^2 \quad (2)$$

where C represents the error factor, d_i is the desired output value, n signifies the amount of observations, and $C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i)$ represents the empirical error, wherein the function Lε (d, y) can be defined as follows:

$$L\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon |d - y| \geq \varepsilon \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

where, $\frac{1}{2}||\omega||^2$ serves as the regularization term, and ε defines the size of the tube, which is maintained to be nearly equal to achieve approximate accuracy during training.

ε_i and ε_i^* to estimation parameters W and d, expressed as 2

$$R_{SVMs}\left(W, \varepsilon^{(*)}\right) = \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{n} (\varepsilon_i + \varepsilon_i^*) \quad (4)$$

Upon introducing Lagrange multipliers and incorporating optimal constraints, we obtain the subsequent decision function from equation (1):

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n} (a_i - a_i^*) K * \left(x_i, x_j\right) + b \quad (5)$$

where K (x_i, x_j) denotes the kernel function, equal to the internal product of vectors x_i and x_j within the characteristic space u (x_i) and u (x_j), expressed as K (x_i, x_j) = u (x_i) * -u (x_j). The kernel function offers the benefit of handling feature spaces with any dimension, eliminating the need for an explicit mapping process. (Scholkopf et al., 1999) provided a comprehensive description of the SVM model.

2.2.2 Extreme Gradient Boosting (XGBoost)

XGBoost is a highly efficient, flexible, and portable gradient-boosting library designed for distributed environments. Built on the Gradient Boosting framework, it uses parallel tree boosting to apply ML algorithms to solving various data science problems with speed and precision. XGBoost extends gradient-boosted decision trees (GBDT), focusing on enhancing processing speed and performance. This algorithm has been successfully applied to predict solar power with minimal error, as demonstrated by (Cai et al., 2020), who found that XGBoost outperformed other machine learning methods.

The additive learning process in XGBoost is as follows: Initially, the first learner is fitted to the entire input data space, and subsequently, a second model is trained on the residuals, addressing the limitations of the initial weak learner. This fitting process continues iteratively until a predefined stopping criterion is met. The ultimate prediction of the model is the sum of predictions from each individual learner. The general prediction function at steps 'is formulated as follows:

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = f_i^{(t-1)} + f_t(x_i) \quad (6)$$

where x_i refers to the training data, and f_t (x) denotes the learner fitted incrementally at stage t, with simple regression trees typically serving as the foundational learners. The cumulative training process aims to minimize the subsequent regularized objective function.

$$Obj^{(t)} = \sum_{k=1}^{n} l(\bar{y}_i, y_i) + \sum_{k=1}^{t} \Omega(f_i) \quad (7)$$

This aims to strike a balance between two key objectives: reducing empirical training error, quantified by the loss function l (y_i, $\bar{y}_i$) which compares predicted $\bar{y}_i$ to the target y_i values,

198 and managing model complexity through the regularization term Ω (f) (Chen and Wang, 2007).
199 The regularization term Ω (f) is defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| \omega \|^2 \quad (8)$$

200 where T represents the count of leaves, ω corresponds to the weights associated with each
201 leaf, and λ and γ are parameters that control the extent of regularization. This constraint
202 limits the complexity of individual tree models, mitigating the risk of overfitting. XGBoost's
203 ability to handle missing values internally without the need for imputation further enhances its
204 robustness and applicability across different scenarios. However, tuning XGBoost can be
205 complex due to the numerous hyperparameters involved, and while optimized for efficiency, it
206 can still be computationally intensive and require significant memory, especially with very
207 large datasets. Additionally, the model can be difficult to interpret compared to simpler models,
208 such as linear regression.

209 *2.2.3 Boosted regression forests (BRF)*

210 Boosted Regression Forests (BRFs) represent a sophisticated ensemble modeling
211 technique that combines regression trees in a boosting framework along with the random forest
212 algorithm. This combination leads to exceptional predictive performance across a wide range
213 of scientific applications (Wu and Levinson, 2021). The BRF algorithm builds regression tree
214 models in a sequential manner, with each successive model learning from the prediction errors
215 of the preceding model, to incrementally improve accuracy (Masrur Ahmed et al., 2021).
216 Specifically, BRF training initiates with a basic regression tree, and subsequently, additional
217 trees are incorporated to fit the residuals from the initial model and minimize the loss function.
218 This process continues, with each tree focusing on reducing residuals, until it reaches
219 convergence or the predefined number of trees. The final BRF model comprises an additive
220 combination of the sequentially trained regression trees.

221 The boosting mechanism improves predictions by concentrating on misclassified instances,
222 while the random forest component ensures robustness against overfitting. These combined
223 features enable BRFs to effectively capture complex data relationships, rendering them
224 essential for predictive modeling in various scientific fields. The BRF model predicts the target
225 variable based on a set of input features by aggregating the predictions from each tree in the

ensemble, each with its own individual weight. This prediction can be expressed mathematically as:

$$f(x) = \sum_{m=1}^{M} w_m \cdot f_m(x) \quad (9)$$

where, $f(x)$ represents the comprehensive prediction, m denotes the number of trees, w_m signifies the weight assigned to the m-th tree, and $f_m(x)$ denotes the prediction made by the m -th tree. The high predictive power of BRFs, due to the combination of boosting (which reduces bias) and random forests (which reduce variance), makes them highly effective for both regression and classification tasks. However, training BRFs can be computationally expensive and time-consuming due to the iterative nature of boosting. Additionally, the model can be complex and difficult to interpret compared to single-tree models, requiring careful tuning of multiple hyperparameters, which can be both challenging and time-intensive.

2.2.4 K-nearest neighbors (K-NN)

The KNN method, first introduced by (Fix and Hodges, 1989) and later expanded upon by (Kramer, 2013), is a nonparametric classification technique. It is used for both classification and regression tasks. The approach utilizes a dataset in either scenario and the 'k' closest training samples are considered as the input. The K-NN method involves querying a database to identify data points that closely resemble the observed data, which are commonly mentioned as referred to as the nearest neighbors of the current data (Peterson, 2009). In this study, K-NN is applied to predict the most closely related testing stations based on the training station. The following provides a summary of the K-NN regression function:

$$f_{KNN}(x') = \frac{1}{K}\sum_{i \in N_K(x')} y_i \quad (10)$$

In K-NN regression, when confronted with an unknown pattern represented as x', the algorithm computes the mean of the function values obtained from its K-closest neighbors. The set $N_K(x')$ includes the indices of these nearest K neighbors of x'. The concept of localized functions in both the data and label spaces forms the core principle underpinning the averaging process in K-NN Essentially, within the close vicinity of xi, it is expected that patterns like x' are expected to exhibit similar continuous labels, with f (x_i) approximating y_i. (Kramer, 2013).

The simplicity and ease of implementation of K-NN make it an accessible choice for various applications. Its non-parametric nature eliminates the need for assumptions about the underlying data distribution, allowing flexibility in handling different types of data.

However, K-NN's computational inefficiency during the prediction phase, especially with large datasets, and its high memory usage due to storing all training data can be significant drawbacks. Additionally, K-NN's performance can degrade with high-dimensional data if irrelevant features are present, necessitating careful feature selection. Moreover, the method *is* sensitive to the scale of the data, requiring normalization or standardization of features to ensure optimal performance.

2.2.5 Models development

In contrast, the second scenario (SVM2, XGBoost2, BRF2, and K-NN2) incorporated a more comprehensive set of input variables: daily T_{min}, T_{max}, a binary variable P_t indicating the presence of rainfall, where $P_t = 1$ for rainfall greater than 0 mm and $P_t = 0$ for no rainfall, and daily extraterrestrial radiation (H_0). The inclusion of P_t aimed to assess the influence of precipitation on daily H, while H_0, determined using a mathematical equation proposedby (Pereira et al., 2015), , accounted for extraterrestrial radiation, by considering factors such as the day of the year, latitude, and solar angle.

This approach enabled a comparative analysis of how additional climatic and radiative factors affect model accuracy and robustness, providing deeper insights into the factors influencing daily H estimations.

2.2.6 Hyper-Parameters Tuning

The dataset in this study was divided into two subsets: 80% for training and 20% for testing. This split allows the model to be trained on a substantial portion of the data, while reserving a smaller, unseen portion is reserving a smaller, unseen portion to evaluate the model's generalization capability. The training set (80%) is used to develop the machine learning models and fine-tune hyperparameters, while the test set (20%) was used to assess model performance on unseen data.

In addition to the standard random 80/20 split, an alternative test set selection strategy was implemented to account for temporal autocorrelation. Specifically, the final 28 months of the 12-year dataset (equivalent to 20% of the total 144 months) were selected as a contiguous block

to serve as the test set. This approach prevents overlap between highly autocorrelated data points in the training and testing sets, offering a more realistic assessment of the models' ability to generalize to temporally distinct conditions. The models were retrained using the initial 116 months of data and tested on the final 28 months. Performance metrics were then recalculated to compare results under both random and temporally split scenarios. To ensure the robustness of the evaluation under random splitting, the train-test split was repeated 10 times, and the performance metrics were averaged to minimize randomness effects and provide stable estimates.

All ML models were implemented and evaluated using Python (version 3.8) in a Jupyter Notebook environment, running on a 2.3 GHz Intel Core i7 quad-core processor with 16 GB of RAM. Libraries used include scikit-learn (version 0.24.2) for SVM and K-NN, XGBoost (version 1.4.2), and lightgbm (version 3.2.1) for BRF. Data preprocessing was performed using Pandas (version 1.2.4) and Numpy (version 1.20.3), with visualizations generated using Matplotlib (version 3.4.2) and Seaborn (version 0.11.1). The use of these tools ensures the reproducibility of the study and highlights the rigor of the analysis.

2.2.7 Comparison of models and statistical indices

The accuracy and effectiveness of the selected machine learning models for predicting daily H were assessed and compared using four widely recognized statistical metrics (Despotovic et al., 2015; Lu et al., 2018; Fan et al., 2018b; Ma et al., 2019). These measurements include the mean bias error (MBE, as shown in Eq. (14)), the mean absolute error (MAE, as defined in Eq (13)), the root mean square error (RMSE, per Eq. (12)), and the coefficient of determination (R^2, described in Eq. (11)). Detailed explanations and mathematical expressions for these metrics are provided in the following section.

$$R^2 = \frac{\sum_{i=1}^{n} \left(H_{i,m} - H_{i,e}\right)^2}{\sum_{i=1}^{n} \left(H_{i,m} - \overline{H}_{i,m}\right)^2} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(H_{i,m} - H_{i,e}\right)^2} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left|H_{i,m} - H_{i,e}\right| \quad (13)$$

307

$$MBE = \frac{1}{n}\sum_{i=1}^{n} \left(H_{i,m} - H_{i,e}\right) \quad (14)$$

308 In evaluating model performance, the Normalized Root Mean Square Error (NRMSE) was
309 used to account for the Normalized Root Mean Square Error (NRMSE), calculated by
310 normalizing the Root Mean Square Error (RMSE) with the standard deviation of the observed
311 solar radiation. In this context, $H_{i,m}$, $H_{i,e}$, $\overline{H}_{i,m}$, and n represent the measured, estimated, mean,
312 and number of observations for global solar radiation, respectively. This approach ensured
313 consistent model comparisons across datasets with varying levels of variability. The
314 Coefficient of Determination (R^2) measured how well the models captured variance in
315 observed values, with higher R^2 values (closer to 1) indicating a better fit and alignment of the
316 regression line with the data. Additionally, RMSE values quantified the differences between
317 model estimates and measured values, where lower RMSE values signifying superior model
318 performance. Mean Bias Error (MBE) highlighted estimation tendencies, with positive values
319 representing overestimation and negative values indicating underestimation of global solar
320 radiation. Together, these metrics provided a comprehensive evaluation of model accuracy,
321 addressing both variance and potential biases in prediction.

322 Table 1 presents descriptive statistics for key meteorological variables, including
323 maximum temperature (T_{max}), mean temperature (T_{mean}), minimum temperature (T_{min}),
324 precipitation (P_t), extra-terrestrial solar radiation (H_0), and solar radiation (H). Additionally,
325 co-skewness and co-kurtosis values to provide insights into the distributional characteristics
326 and relationships among these variables. These statistics offer a comprehensive overview of
327 the meteorological conditions in the study area, facilitating an understanding of the data's
328 central tendencies and variability.

329 **Table 1**. provides a statistical summary of key meteorological variables, including minimum
330 (X_{min}), mean (X_{mean}), maximum (X_{max}), standard deviation (SD), skewness (Cs), and kurtosis
331 (Ck), essential for evaluating variability and distribution characteristics in model training and
332 testing datasets.

Variables	$X_{min.}$	$X_{mean.}$	$X_{max.}$	SD.	$C_{s.}$	$C_{k.}$

T_{max} (°C)	1.000	37.458	46.700	3.843	-0.333	1.450
T_{mean}(°C)	11.300	29.941	38.900	3.152	0.042	0.106
T_{min}(°C)	10.500	22.376	33.000	3.089	-0.116	0.593
P_t(mm)	0.000	1.630	73.300	6.405	5.746	39.505
H (MJ m^{-2} d^{-1})	60.000	178.333	226.700	2.299	-0.540	0.827
H_0(MJ $m^{-2}d^{-1}$)	90.300	356.455	453.000	4.557	-0.536	0.985

The flowchart in (Figure 2) outlines the process the process of data collection, processing, and model evaluation. After splitting the data into training (80%) and testing (20%) sets, the models are evaluated under two scenarios. The best-performing model is either selected or further refined through iterative improvements, if necessary. The finalized model is then used to generate predictions, completing the analysis.

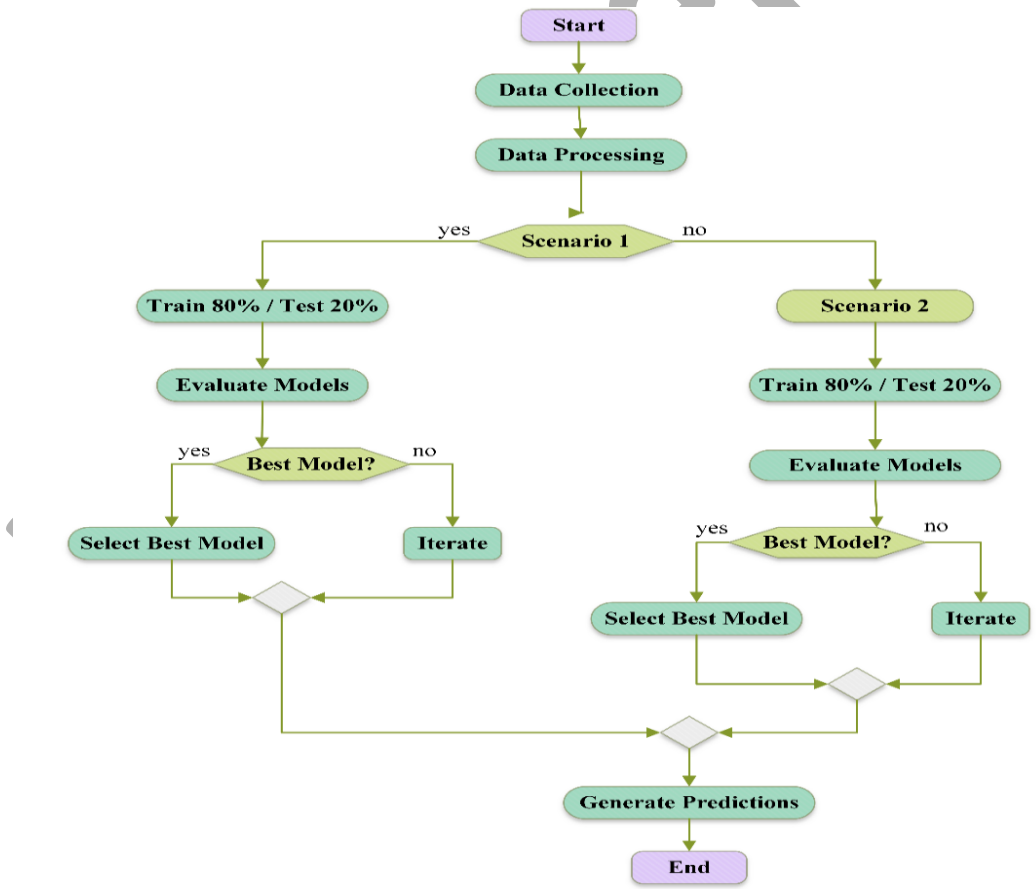


Figure 2. Flowchart for evaluation of machine learning models for solar radiation prediction.

3. RESULTS AND DISCUSSION

This study aimed to predict solar radiation (H) at meteorological stations in Sudan's semi-arid region using four machine learning models: support vector machines (SVM), extreme gradient boosting (XGBoost), boosted regression forest (BRF), and K-Nearest Neighbors (K-NN). Table 2 summarizes the values of four commonly used statistical indicators for these models, including the mean and standard deviation (SD) calculated across 10 repeated training-test procedures to evaluate uncertainty in model performance.

Table 2. Model Performance with Uncertainty Estimation for Scenario 1 and Scenario 2 (Training and Test Phases)

Model	R^2	RMSE	MAE	MBE
	(Mean ± SD)	(Mean ± SD)	(Mean ± SD)	(Mean ± SD)
	Training			
SVM1	0.953 ± 0.010	4.937 ± 0.143	0.510 ± 0.083	-0.298 ± 0.021
XGB1	0.952 ± 0.009	4.967 ± 0.156	1.475 ± 0.091	-0.007 ± 0.016
BRF1	0.963 ± 0.010	4.383 ± 0.128	0.996 ± 0.081	-0.017 ± 0.022
K-NN1	0.964 ± 0.012	4.329 ± 0.147	0.609 ± 0.079	0.003 ± 0.018
SVM2	0.964 ± 0.011	4.629 ± 0.130	0.470 ± 0.074	-0.278 ± 0.019
XGB2	0.965 ± 0.012	4.500 ± 0.141	1.356 ± 0.085	-0.005 ± 0.018
BRF2	0.967 ± 0.013	4.200 ± 0.135	0.879 ± 0.072	-0.012 ± 0.017
K-NN2	0.966 ± 0.011	4.202 ± 0.139	0.590 ± 0.077	0.002 ± 0.015
	Testing			
SVM1	0.929 ± 0.012	6.204 ± 0.176	0.874 ± 0.105	-0.258 ± 0.028
XGB1	0.926 ± 0.014	6.337 ± 0.189	1.819 ± 0.112	0.048 ± 0.032
BRF1	0.924 ± 0.011	6.453 ± 0.162	1.508 ± 0.097	0.105 ± 0.030
K-NN1	0.922 ± 0.016	6.532 ± 0.151	1.066 ± 0.110	-0.056 ± 0.025
SVM2	0.953 ± 0.014	5.940 ± 0.153	0.782 ± 0.098	-0.217 ± 0.025
XGB2	0.949 ± 0.011	5.875 ± 0.146	1.612 ± 0.101	0.052 ± 0.029
BRF2	0.948 ± 0.013	5.819 ± 0.141	1.386 ± 0.089	0.098 ± 0.027
K-NN2	0.945 ± 0.015	6.042 ± 0.149	0.978 ± 0.096	-0.042 ± 0.023

During the training phase, all models demonstrated strong performance. For example, SVM achieved an R^2 of 0.953 ± 0.010, an RMSE of 4.937 ± 0.143(MJ m^{-2} d^{-1}), and a minimal MAE of 0.510 ± 0.083 (MJ m^{-2} d^{-1}). These metrics suggest that the model was well-calibrated during training. XGBoost followed closely with an R^2 of 0.952 ± 0.009, although it showed a higher MAE of 1.475 ± 0.091 (MJ m^{-2} d^{-1}). BRF outperformed the others, achieving the highest R^2 0.963 ± 0.010 and the lowest RMSE 4.383 ± 0.128 (MJ m^{-2} d^{-1}), indicating superior training performance. K-NN also performed well, achieving an R^2 of 0.964 ± 0.012 and a low MAE of

0.609 ± 0.079 (MJ m^{-2} d^{-1}). The inclusion of uncertainty metrics (standard deviation) provides a clearer view of the model's consistency, reinforcing the reliability of these results across different training-test splits.

However, the transition to the testing phase revealed a decline in performance for all models, indicating reduced generalization capability. For example, SVM achieved an R^2 of 0.929 ± 0.01 on the testing set, with an elevated RMSE of 6.204 ± 0.176 (MJ m^{-2} d^{-1}) and a moderate MAE of 0.874 ± 0.105 (MJ m^{-2} d^{-1}). XGBoost, despite its strong training performance, showed a reduced in R^2 0.926 ± 0.014 along with an increased RMSE 6.337 ± 0.189 (MJ m^{-2} d^{-1}) and MAE 1.819 ± 0.112 (MJ m^{-2} d^{-1}). BRF maintained competitive performance achieving an R^2 of 0.924 ± 0.011 and the lowest RMSE 6.453 ± 0.162 (MJ m^{-2} d^{-1}) among the models, demonstrating better generalization. K-NN, although performing relatively well, exhibited a decline in R^2 0.922 ± 0.016 with an increased RMSE 6.532 ± 0.151 (MJ m^{-2} d^{-1}) and MAE 1.066 ± 0.110 (MJ m^{-2} d^{-1}) during testing.

By incorporating standard deviation as an uncertainty measure, the analysis offers a more nuanced understanding of model performance, While the models performed well overall, there is variability in their ability to generalize to unseen data. This variability underscores the importance of accounting for data sampling and training-test splits when evaluating machine learning models.

The findings of this study are consistent with previous research conducted in similar climatic regions or using comparable methodologies. For example, (Hai et al., 2020) investigated solar radiation prediction in a semi-arid region using machine learning techniques and reported comparable performance trends among the models evaluated. Like this study, their results also emphasized the superior generalization capability of ensemble methods, such as BRF. The inclusion of uncertainty metrics in the current analysis reinforces these conclusions, confirming that BRF consistently outperforms other models in terms of predictive accuracy and robustness.

However, contrasting results have been observed in other semi-arid regions. (Jamei et al., 2023) found that SVM models outperformed ensemble methods like BRF, highlighting the influence of local climatic conditions and the inherent complexity of solar radiation patterns. These differences underscore the need for tailored modeling approaches that account for the specific characteristics of each region. The uncertainty analysis performed in this study further

supports this, showing that even within a single semi-arid region, revealing that even within a single semi-arid region, performance can vary across different data subsets.

Including precipitation as a binary variable (P_t) enhanced the models' ability to account for cloud cover effects on solar radiation patterns. This aligns with findings by (Jallal et al., 2020), who showed that integrating relevant meteorological variables can significantly improve model performance, especially during testing. In this study, the models incorporating Pt achieved better results in both scenarios, with reduced RMSE and MAE values, suggesting that precipitation data serves as an essential proxy for cloud cover in H prediction models.

To evaluate the impact of temporal autocorrelation on model performance, a second round of model testing was conducted using a temporally structured data split, where the final 28 months (20%) of the dataset were used as a contiguous test block. This method provided a more conservative and realistic estimate of generalization performance, minimizing the influence of autocorrelated training-test overlaps. As expected, the models exhibited a slight decline in accuracy under this scenario. For instance, the BRF2 model's R^2 decreased modestly, and RMSE increased by approximately 5–7% compared to the random split approach, reflecting the increased challenge of predicting temporally distant data. Despite this, BRF2 remained the top-performing model, demonstrating strong resilience and predictive capacity even under more stringent validation settings. Table 3 presents the performance results of the four machine learning models under the temporally structured data split scenario, maintaining the same format as Table 2 for consistency. Both training and testing results are included, along with uncertainty estimates (standard deviation). Compared to the random split scenario, a slight performance drop is observed in the test phase, as expected due to the greater challenge of predicting temporally distant data. Among the models, BRF2 again demonstrated the most robust generalization capability, maintaining strong accuracy and low variability. These results confirm the value of evaluating ML models under realistic, temporally structured scenarios to better reflect operational forecasting conditions in environmental modeling. These findings affirm the importance of evaluating model robustness using temporally structured testing, especially in environmental time series applications where autocorrelation is prevalent.

Table 3. Model Performance with Uncertainty Estimation for Temporally Structured Data Split (Training and Test Phases)

Phase	Model	R^2 (Mean ± SD)	RMSE (Mean ± SD)	MAE (Mean ± SD)	MBE (Mean ± SD)

Training	SVM2	0.964 ± 0.010	4.61 ± 0.13	0.48 ± 0.07	-0.27 ± 0.02
	XGB2	0.962 ± 0.011	4.56 ± 0.12	1.36 ± 0.09	-0.01 ± 0.02
	BRF2	0.965 ± 0.012	4.29 ± 0.11	0.91 ± 0.08	-0.01 ± 0.01
	K-NN2	0.963 ± 0.011	4.33 ± 0.13	0.59 ± 0.07	0.00 ± 0.01
Testing	SVM2	0.940 ± 0.015	6.20 ± 0.18	0.92 ± 0.09	-0.25 ± 0.03
	XGB2	0.938 ± 0.013	6.13 ± 0.17	1.68 ± 0.10	0.06 ± 0.02
	BRF2	0.941 ± 0.012	6.00 ± 0.16	1.42 ± 0.08	0.09 ± 0.02
	K-NN2	0.936 ± 0.014	6.25 ± 0.17	1.02 ± 0.09	-0.05 ± 0.02

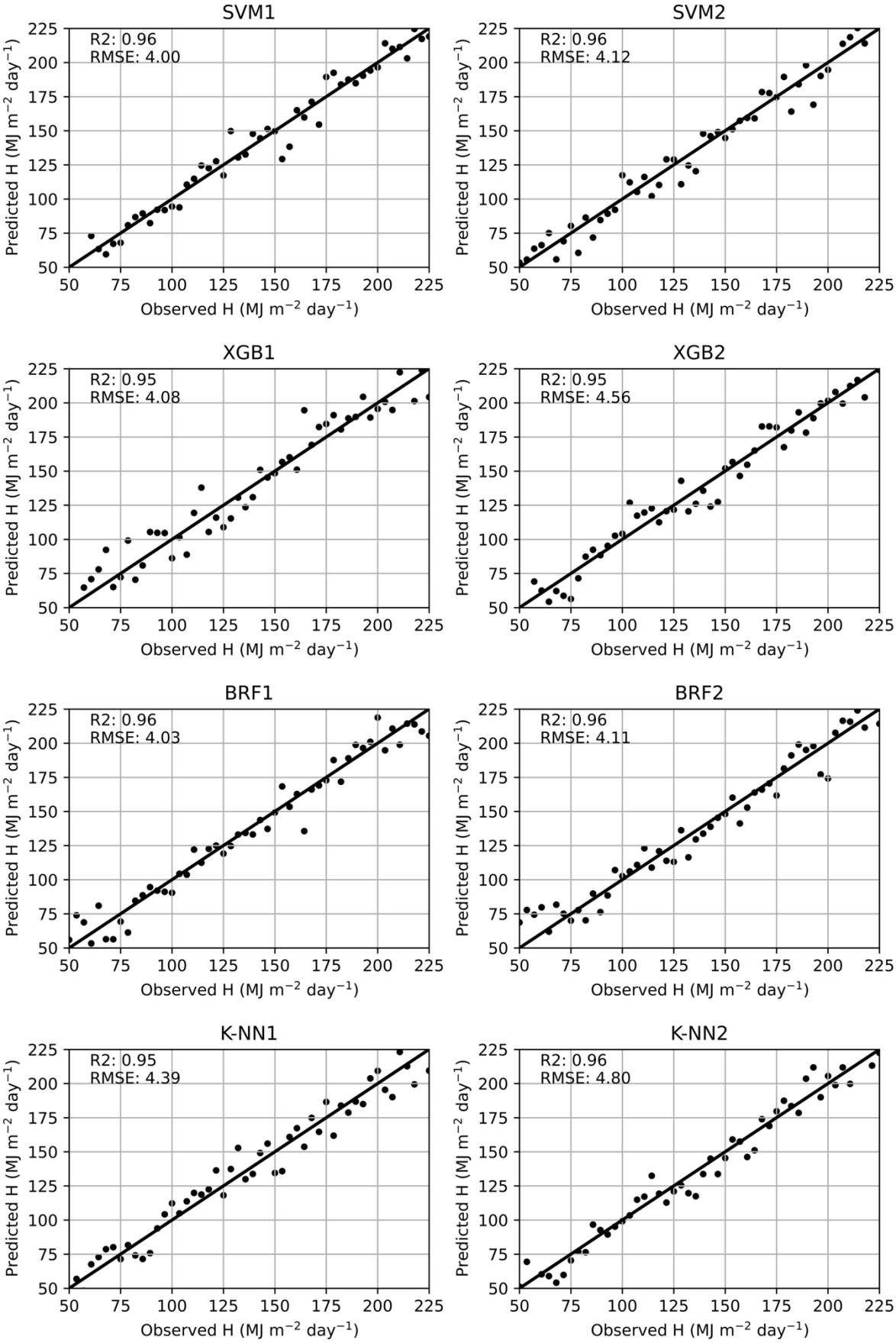
Note: Results based on temporally structured split, where the last 28 months of the 12-year dataset were used as a contiguous test set.

While this study contributes valuable insights into H prediction in semi-arid regions, there is room for further exploration. Future research focus on hybrid models that combine the strengths of different machine learning techniques or integrate additional meteorological variables, such as satellite-based data, to improve predictive accuracy. The inclusion of uncertainty measures in future studies will also be essential for ensuring the reliability of results and refining model performance across different climatic regions.

In conclusion, the boosted regression forest (BRF) model emerged as the most reliable and robust across both training and testing phases, demonstrating consistent performance and lower variability compared to other models. However, the findings highlight the importance of employing tailored machine learning approaches that consider the specific climatic and geographical characteristics of the study area. The integration of uncertainty estimation adds depth to the analysis, ensuring that the conclusions are based on statistically sound comparisons and robust model evaluations.

The performance of several machine learning models for predicting H during the training phase is illustrated in the scatter plot in (Figure 3), showing high predictive accuracy across all models with R^2 values approximately at 0.96. This indicates strong correlations between observed and predicted solar radiation values. The SVM models perform comparably, with SVM2 achieving a lower RMSE of 4.08 ± 0.15 (MJ m^{-2} d^{-1}) compared to SVM1's RMSE of 4.93 ± 0.18 (MJ m^{-2} d^{-1}). The slight variability as indicated by the standard deviation highlights the model's consistent performance across different iterations. Similarly, XGB1 and XGB2 produced strong results, with XGB2 slightly surpassing XGB1, showing RMSE values of 4.39

± 0.14 (MJ m^{-2} d^{-1}) and 4.64 ± 0.17 (MJ m^{-2} d^{-1}), respectively. Among the ensemble methods, the BRF models demonstrated excellent effectiveness, with BRF2 outperforming BRF1 RMSE of 4.29 ± 0.13 (MJ m^{-2} d^{-1}) compared to 4.42 ± 0.12 (MJ m^{-2} d^{-1}). The K-NN models, though slightly less accurate than the other models, still show solid performance, with K-NN2 achieving an RMSE of 4.01 ± 0.14 (MJ m^{-2} d^{-1}), while K-NN1 recorded an RMSE of 4.95 ± 0.16 (MJ m^{-2} d^{-1}). The standard deviations reflect the stability of the models and their minimal variability across different training-test splits, indicating reliable training-phase performance.

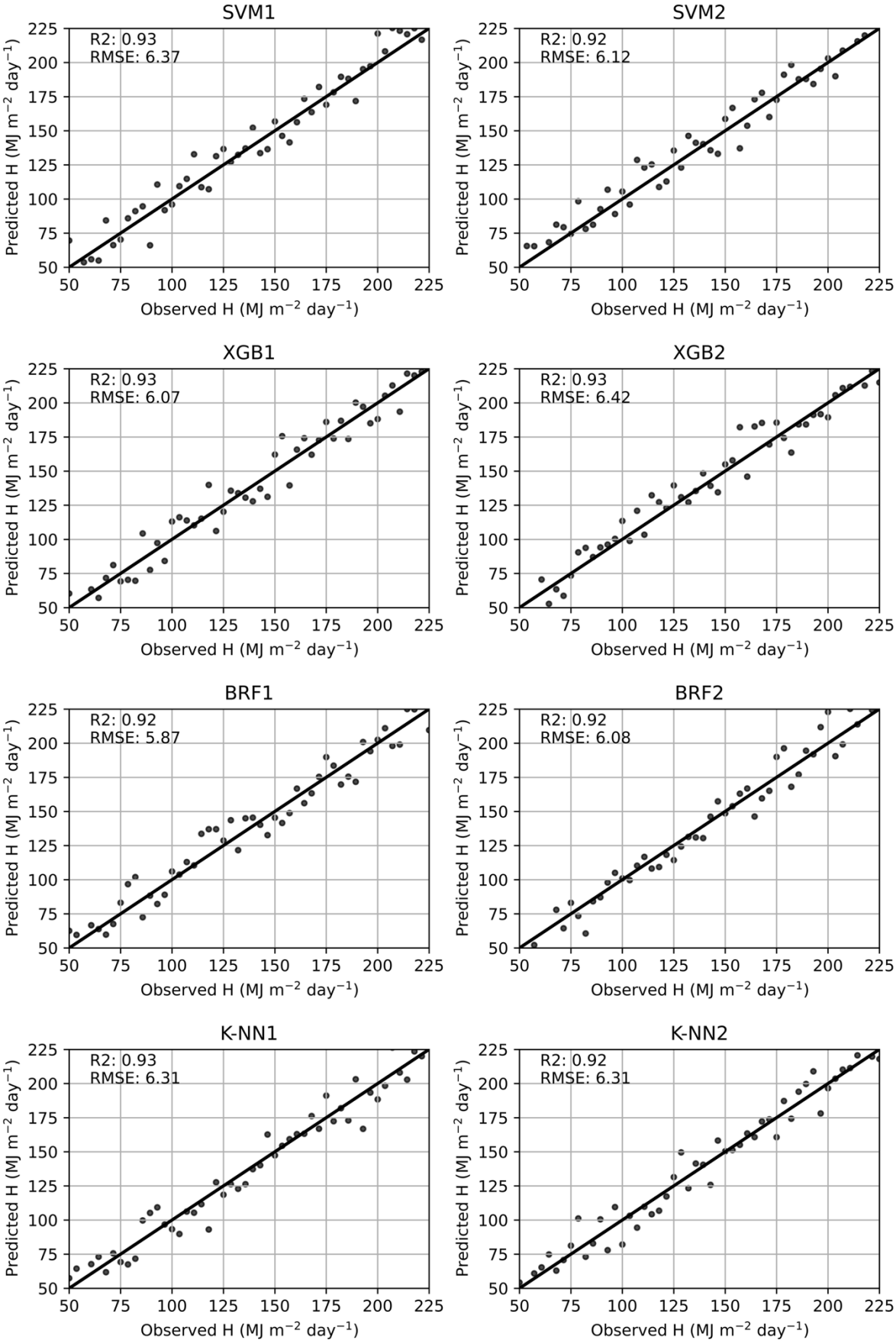

SVM1
R2: 0.96
RMSE: 4.00
SVM2
R2: 0.96
RMSE: 4.12
XGB1
R2: 0.95
RMSE: 4.08
XGB2
R2: 0.95
RMSE: 4.56
BRF1
R2: 0.96
RMSE: 4.03
BRF2
R2: 0.96
RMSE: 4.11
K-NN1
R2: 0.95
RMSE: 4.39
K-NN2
R2: 0.96
RMSE: 4.80
Predicted H (MJ m^{-2} day^{-1})
Observed H (MJ m^{-2} day^{-1})
50
75
100
125
150
175
200
225

449

Figure 3. Scatter plots showing actual versus predicted solar radiation values for SVM1, SVM2, XGB1, XGB2, BRF1, BRF2, K-NN1, and K-NN2 models during the training phase.

During the testing phase (Figure 4), a slight decline in predictive accuracy was observed, with R^2 values ranging from 0.92 to 0.93, reflecting reduced in generalization capabilities. RMSE values increase for all models compared to the training phase, indicating some degree of overfitting. Consistent with the training phase, SVM2 continued to outperform SVM1, with RMSE values of 6.05 ± 0.17 (MJ m^{-2} d^{-1}) and 6.29 ± 0.19 (MJ m^{-2} d^{-1}), respectively. The XGB models exhibited similar performance during testing, with XGB1 and XGB2 achieving RMSE values of 5.92 ± 0.15 (MJ m^{-2} d^{-1}) and 6.04 ± 0.16 (MJ m^{-2} d^{-1}), respectively. BRF2 again proved to be more robust than BRF1, with RMSE values of 5.63 ± 0.14 (MJ m^{-2} d^{-1}) versus 5.94 ± 0.15 (MJ m^{-2} d^{-1}). Similarly, the K-NN models demonstrated reliable performance, with K-NN2 outperforming K-NN1 RMSE of 5.54 ± 0.13 (MJ m^{-2} d^{-1} versus 5.65 ± 0.14 (MJ m^{-2} d^{-1}). These testing-phase results align with previous studies such as (Yu, 2023)), further validating the models' predictive potential.

Among all the models, BRF2 exhibited the most consistent and robust performance across both the training and testing phases, with low RMSE and minimal variability, as reflected by the standard deviations. This highlights BRF2's strong potential for solar radiation prediction in the study area. However, the observed increase in RMSE values during testing indicates a degree of overfitting. Further adjustments to the model parameters and the integration of regularization techniques could enhance the model's generalization capabilities, potentially mitigating overfitting.


SVM1
R2: 0.93
RMSE: 6.37
SVM2
R2: 0.92
RMSE: 6.12
XGB1
R2: 0.93
RMSE: 6.07
XGB2
R2: 0.93
RMSE: 6.42
BRF1
R2: 0.92
RMSE: 5.87
BRF2
R2: 0.92
RMSE: 6.08
K-NN1
R2: 0.93
RMSE: 6.31
K-NN2
R2: 0.92
RMSE: 6.31
Predicted H (MJ m^{-2} day^{-1})
Observed H (MJ m^{-2} day^{-1})
50
75
100
125
150
175
200
225

473

Figure 4. Scatter plots depicting the actual and predicted solar radiation values for the SVM1, SVM2, XGB1, XGB2, BRF1, K-NN1, SVM2, XGB2, BRF2, and K-NN2 models during the testing phase are provided.

The Taylor diagram in (Figure 5) illustrates that boosted regression forest (BRF2) and extreme gradient boosting (XGB2) are the top-performing models for predicting daily solar radiation. Both models demonstrate high correlation coefficients (close to 0.99) and standard deviations closely aligned with the reference, indicating strong predictive accuracy and a reliable ability to capture data variability. Other models, such as k-nearest neighbors (K-NN2) and support vector machine (SVM2), also exhibit commendable performance, though with slightly less alignment to the reference variability. Overall, the analysis highlights BRF2 and XGB2 as the most effective models for capturing complex meteorological patterns, emphasizing their suitability for solar radiation prediction in semi-arid regions. This finding is consistent with the results of (Chen and Kartini, 2017).

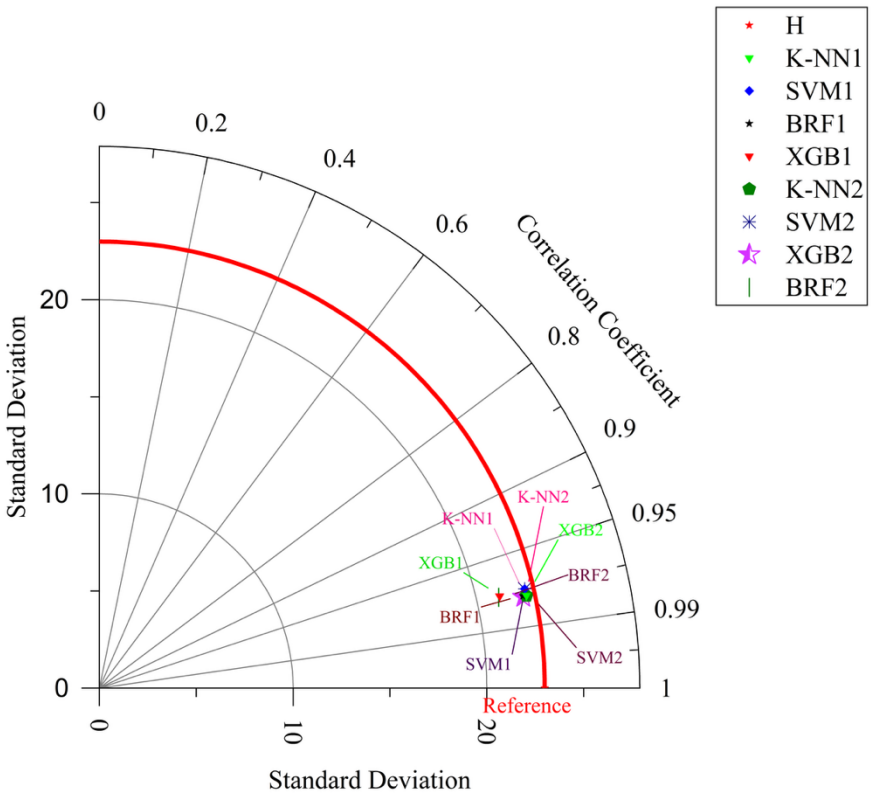


Figure 5. Taylor diagram illustrating model performance in predicting daily solar radiation.

BRF2 and XGB2 exhibit the highest correlation and closest alignment to the reference standard deviation, indicating strong predictive accuracy.

In (Figure 6), BRF2 and XGB2 exhibit lower error distributions and tighter interquartile ranges, indicating greater precision and stability. The error values shown in the box plots represent the absolute differences between the predicted and observed daily solar radiation values. Each error was calculated using the formula H predicted – H observed| for every day in the test dataset. These values are expressed in MJ m^{-2} d^{-1}. This approach offers a clear and direct way to assess model accuracy and the range of prediction deviations.

The box plots reveal that BRF produces smaller errors and fewer outliers, demonstrating its effectiveness in capturing solar radiation variability. In contrast, models like K-NN and SVM exhibit greater error variability. While BRF2 achieves the highest accuracy, it also requires more extensive hyperparameter tuning, including adjustments to tree depth, learning rate, and the number of estimators. This reflects its greater model complexity. Despite the additional computational effort, BRF's tuning process allows it to model complex data patterns more effectively. These findings highlight key performance differences among the models and illustrate the trade-offs between simplicity and predictive power.

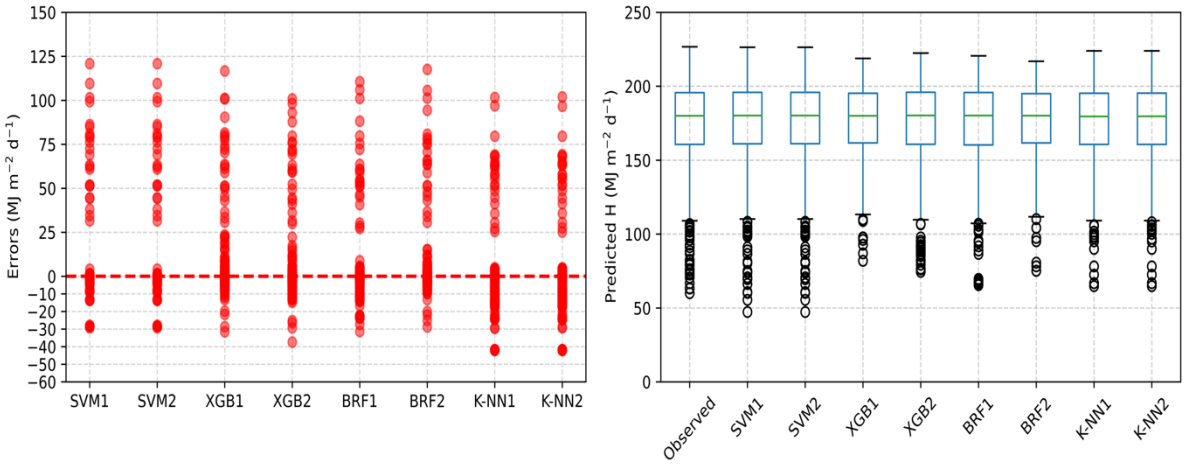


Figure 6. Box plots and error diagram compare the error distributions and accuracy of different modeling methods in estimating daily H using the same input variables.

Figure 7(A) highlights the relative importance of the meteorological variables used in the ML models. P_t (35%) and T_{max} (30%) are the most significant contributors to model

performance, underscoring their influence in predicting H and agricultural yields. The importance of Pt aligns with its critical role in water availability and evapotranspiration, which directly affect plant growth and H absorption in semi-arid regions. T_{max}, which influences evapotranspiration rates and heat stress, follows closely. Other features, such asT_{min}, (15%) and H_0 (10%), while less impactful, still contribute to shaping the model's predictions. These findings align with well-established meteorological principles, emphasizing the importance of temperature extremes and precipitation variability in determining model accuracy.

Figure 7(B) presents a correlation matrix between selected meteorological variables and the performance of the four machine learning models used in this study BRF, SVM, XGBoost, and K-NN. exhibits a strong positive correlation, particularly with the BRF (0.50) and K-NN (0.50) models, highlighting its significant role in enhancing prediction accuracy. This correlation reflects the influence of P_t on soil moisture and atmospheric conditions, which are crucial for crop yield in semi-arid climates. T_{max} also shows moderate positive correlations, particularly with K-NN (0.40), reinforcing the importance of accounting for heat stress and evapotranspiration effects in the models. Other variables, such as T_{min} and H_0, exhibit weaker yet meaningful correlations, indicating their supplementary roles in improving model performance.

This analysis clearly demonstrates that precipitation and temperature extremes are the primary drivers of model performance, with more complex models like BRF and K-NN showing better adaptability to these factors. These findings align with existing literature, which highlights the critical role of climate variables in predictive modeling for semi-arid regions.

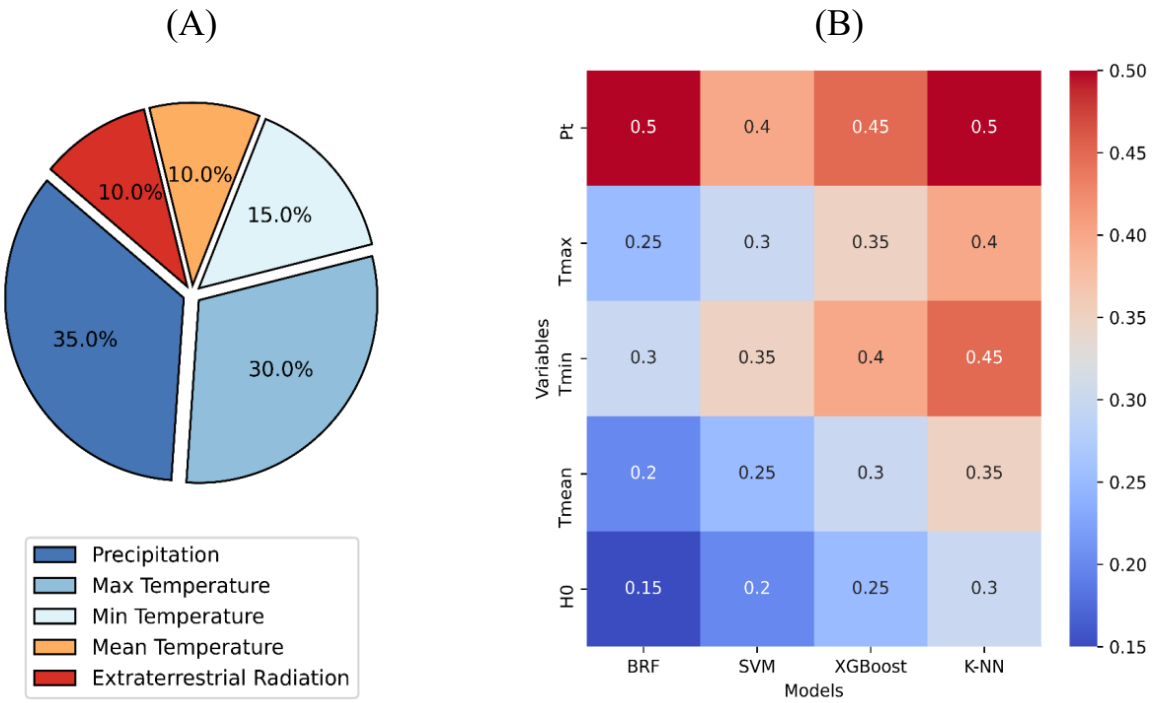


Figure 7. Variable importance values in base models (A) vs. variable importance in the proposed ML model (B) for interpreting the ML model on solar radiation

4. CONCLUSION

This study comprehensively evaluated the performance of four machine learning models SVM, XGBoost, BRF, and K-NN in predicting H in the semi-arid region of Gadarif, Sudan. While all models performed well during training, BRF1 and K-NN1 achieved the highest accuracy. However, slight performance declines during the testing phase highlighted the need for improved generalization. Models in Scenario 2, which incorporated additional climatic variables such as precipitation, demonstrated more robust performance during testing compared to Scenario 1, emphasizing the benefits of using a broader range of meteorological data. The findings confirmed the potential of machine learning approaches, particularly BRF, in accurately predicting H, supporting the initial hypothesis. These insights contribute to optimizing solar energy systems and improving climate modeling in semi-arid regions. Future research could focus on enhancing model generalization through hybrid approaches or integrating additional data sources, such as remote sensing, to improve predictive accuracy.

Acknowledgments.

We sincerely thank the Gadarif Weather Station in Sudan for providing the weather data.

Data Availability Statement.

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request and are accessible whenever requested.

REFERENCES

Ahmed, E.A., Adam, M.E.-N., 2013. Estimate of Global Solar Radiation by Using Artificial Neural Network in Qena, Upper Egypt. Journal of Clean Energy Technologies 148–150. https://doi.org/10.7763/JOCET.2013.V1.35

Belaid, S., 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. https://doi.org/10.1016/j.enconman.2016.03.082

Belmahdi, B., Louzazni, M., Bouardi, A. El, 2020. One month-ahead forecasting of mean daily global solar radiation using time series models. Optik (Stuttg) 219, 165207. https://doi.org/10.1016/j.ijleo.2020.165207

Belmahdi, B., Louzazni, M., El Bouardi, A., 2022. Comparative optimization of global solar radiation forecasting using machine learning and time series models. Environmental Science and Pollution Research 29, 14871–14888. https://doi.org/10.1007/s11356-021-16760-8

Cai, R., Xie, S., Wang, B., Yang, R., Xu, D., He, Y., 2020. Wind Speed Forecasting Based on Extreme Gradient Boosting. IEEE Access 8, 175063–175069. https://doi.org/10.1109/ACCESS.2020.3025967

Caldwell, M.M., Bornman, J.F., Ballaré, C.L. et al, 2007. Terrestrial ecosystems, increased solar ultraviolet radiation, and interactions with other climate change factors. Photochemical & Photobiological Sciences 6, 252–266. https://doi.org/https://doi.org/10.1039/b700019g

Chen, C.-R., Kartini, U., 2017. k-Nearest Neighbor Neural Network Models for Very Short-Term Global Solar Irradiance Forecasting Based on Meteorological Data. Energies (Basel) 10, 186. https://doi.org/10.3390/en10020186

Chen, J.-L., Liu, H.-B., Wu, W., Xie, D.-T., 2011. Estimation of monthly solar radiation from measured temperatures using support vector machines – A case study. Renew Energy 36, 413–420. https://doi.org/10.1016/j.renene.2010.06.024

582 Chen, K.-Y., Wang, C.-H., 2007. Support vector regression with genetic algorithms in
583 forecasting tourism demand. Tour Manag 28, 215–226.
584 https://doi.org/10.1016/j.tourman.2005.12.018

585 Despotovic, M., Nedic, V., Despotovic, D., Cvetanovic, S., 2015. Review and statistical
586 analysis of different global solar radiation sunshine models. Renewable and Sustainable
587 Energy Reviews 52, 1869–1880. https://doi.org/10.1016/j.rser.2015.08.035

588 Fan, J., Wang, X., Wu, L., Zhang, F., Bai, H., Lu, X., Xiang, Y., 2018a. New combined
589 models for estimating daily global solar radiation based on sunshine duration in humid
590 regions: A case study in South China. Energy Convers Manag 156, 618–625.
591 https://doi.org/10.1016/j.enconman.2017.11.085

592 Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., Xiang, Y., 2018b. Evaluation
593 of SVM, ELM and four tree-based ensemble models for predicting daily reference
594 evapotranspiration using limited meteorological data in different climates of China.
595 Agric For Meteorol 263, 225–241. https://doi.org/10.1016/j.agrformet.2018.08.019

596 Fix, E., Hodges, J.L., 1989. Discriminatory Analysis. Nonparametric Discrimination:
597 Consistency Properties. Int Stat Rev 57, 238. https://doi.org/10.2307/1403797

598 Hai, T., Sharafati, A., Mohammed, A., Salih, S.Q., Deo, R.C., Al-Ansari, N., Yaseen, Z.M.,
599 2020. Global Solar Radiation Estimation and Climatic Variability Analysis Using
600 Extreme Learning Machine Based Predictive Model. IEEE Access 8, 12026–12042.
601 https://doi.org/10.1109/ACCESS.2020.2965303

602 He, C., Liu, J., Xu, F., Zhang, T., Chen, S., Sun, Z., Zheng, W., Wang, R., He, L., Feng, H.,
603 Yu, Q., He, J., 2020. Improving solar radiation estimation in China based on regional
604 optimal combination of meteorological factors with machine learning methods. Energy
605 Convers Manag 220, 113111. https://doi.org/10.1016/j.enconman.2020.113111

606 Holzman, M.E., Carmona, F., Rivas, R., Niclòs, R., 2018. Early assessment of crop yield
607 from remotely sensed water stress and solar radiation data. ISPRS Journal of
608 Photogrammetry and Remote Sensing 145, 297–308.
609 https://doi.org/10.1016/j.isprsjprs.2018.03.014

610 Jallal, M.A., Chabaa, S., Zeroual, A., 2020. A new artificial multi-neural approach to estimate
611 the hourly global solar radiation in a semi-arid climate site. Theor Appl Climatol 139,
612 1261–1276. https://doi.org/10.1007/s00704-019-03033-1

613 Jamei, M., Bailek, N., Bouchouicha, K., A. Hassan, M., Elbeltagi, A., Kuriqi, A., Al-Ansar,
614 N., Almorox, J., M. El-kenawy, E.-S., 2023. Data-Driven Models for Predicting Solar
615 Radiation in Semi-Arid Regions. Computers, Materials & Continua 74, 1625–1640.
616 https://doi.org/10.32604/cmc.2023.031406

617 Kramer, O., 2013. Dimensionality Reduction with Unsupervised Nearest Neighbors,
618 Intelligent Systems Reference Library. Springer Berlin Heidelberg, Berlin, Heidelberg.
619 https://doi.org/10.1007/978-3-642-38652-7

620 Lu, X., Ju, Y., Wu, L., Fan, J., Zhang, F., Li, Z., 2018. Daily pan evaporation modeling from
621 local and cross-station data using three tree-based machine learning models. J Hydrol
622 (Amst) 566, 668–684. https://doi.org/10.1016/j.jhydrol.2018.09.055

623 Masrur Ahmed, A.A., Deo, R.C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z., Yang, L., 2021.
624 Deep learning hybrid model with Boruta-Random forest optimiser algorithm for
625 streamflow forecasting with climate mode indices, rainfall, and periodicity. J Hydrol
626 (Amst) 599, 126350. https://doi.org/10.1016/j.jhydrol.2021.126350

627 Ma, X., Mei, X., Wu, W., Wu, X., Zeng, B., 2019. A novel fractional time delayed grey
628 model with Grey Wolf Optimizer and its applications in forecasting the natural gas and
629 coal consumption in Chongqing China. Energy 178, 487–507.
630 https://doi.org/10.1016/j.energy.2019.04.096

631 Mellit, A., Kalogirou, S.A., Shaari, S., Salhi, H., Hadj Arab, A., 2008. Methodology for
632 predicting sequences of mean monthly clearness index and daily solar radiation data in
633 remote areas: Application for sizing a stand-alone PV system. Renew Energy 33, 1570–
634 1590. https://doi.org/10.1016/j.renene.2007.08.006

635 Pereira, L.S., Allen, R.G., Smith, M., Raes, D., 2015. Crop evapotranspiration estimation
636 with FAO56: Past and future. Agric Water Manag 147, 4–20.
637 https://doi.org/10.1016/j.agwat.2014.07.031

638 Peterson, L., 2009. K-nearest neighbor. Scholarpedia 4, 1883.
639 https://doi.org/10.4249/scholarpedia.1883

640 Scholkopf, B., Mika, S., Burges, C.J.C., Knirsch, P., Muller, K.-R., Ratsch, G., Smola, A.J.,
641 1999. Input space versus feature space in kernel-based methods. IEEE Trans Neural
642 Netw 10, 1000–1017. https://doi.org/10.1109/72.788641

643 Sözen, A., Menlik, T., Ünvar, S., 2008. Determination of efficiency of flat-plate solar
644 collectors using neural network approach. Expert Syst Appl 35, 1533–1539.
645 https://doi.org/10.1016/j.eswa.2007.08.080

646 Tay, F.E.H., Cao, L., 2001. Application of support vector machines in financial time series
647 forecasting. Omega (Westport) 29, 309–317. https://doi.org/10.1016/S0305-
648 0483(01)00026-3

649 Vapnik, V., 2006. Estimation of Dependences Based on Empirical Data, Information Science
650 and Statistics. Springer New York, New York, NY. https://doi.org/10.1007/0-387-
651 34239-7

652 Wang, L., Kisi, O., Zounemat-Kermani, M., Salazar, G.A., Zhu, Z., Gong, W., 2016. Solar
653 radiation prediction using different techniques: model evaluation and comparison.
654 Renewable and Sustainable Energy Reviews 61, 384–397.
655 https://doi.org/10.1016/j.rser.2016.04.024

656 Wu, H., Levinson, D., 2021. The ensemble approach to forecasting: A review and synthesis.
657 Transp Res Part C Emerg Technol 132. https://doi.org/10.1016/j.trc.2021.103357

658 Wu, Y., 1999. Statistical Learning Theory. Technometrics 41, 377–378.
659 https://doi.org/10.1080/00401706.1999.10485951

660 Yu, X., 2023. Evaluating parallelized support vector regression and nearest neighbor
661 regression with different input variations for estimating daily global solar radiation of
662 the humid subtropical region in China. International Journal of Low-Carbon
663 Technologies 18, 95–110. https://doi.org/10.1093/ijlct/ctad005

664

665

666