EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Hybrid machine learning models: A comprehensive, data-driven evaluation with diverse data partitioning strategies for net radiation estimation

Kristian Lorenz Bajao[1)], Kittisak Phetpan*[1)], Ponlawat Chophuk[2)] and Rattapong Suwalak[1)]

[1)]Department of Engineering, King Mongkut's Institute of Technology Ladkrabang, Prince of Chumphon Campus, Chumphon 86160, Thailand
[2)]Faculty of Informatics, Burapha University, Chonburi 20131, Thailand

Received 7 December 2024
Revised 2 March 2025
Accepted 7 March 2025

Abstract

Surface net radiation (R_n) is crucial for climate modeling and agricultural management but is often not readily available, especially in regions like Thailand. Accurate prediction of R_n is essential for estimating evapotranspiration, which is vital for irrigation planning and agricultural productivity. This study develops a hybrid machine learning framework that incorporates K-Nearest Neighbors (KNN) for missing data imputation, Random Forest-Recursive Feature Elimination (RF-RFE) for feature selection, and machine learning models (Multi-layer Perceptron, K-Nearest Neighbors, and Random Forest) for prediction. The research evaluates various data partitioning methods, including hold-out split, K-fold cross-validation, and growing-window forward-validation (gwFV), alongside hyperparameter tuning using GridSearch to enhance model robustness and prevent overfitting. The primary objectives are to develop and evaluate the hybrid ML models for daily R_n estimation using basic meteorological inputs (temperature, relative humidity, and sunshine duration), assess the impact of different input combinations on prediction accuracy in Sawi, Chumphon, Thailand, and compare data partitioning techniques to determine the optimal model performance. Utilizing FAO56PM-calculated R_n as a reference, this study finds that the Random Forest model, with average temperature and sunshine duration (M2) as inputs evaluated under the gwFV method, achieves the highest stability and high accuracy (R^2 of 0.972, RMSE of 0.457 MJ m^{-2} day^{-1}, and MAPE of 3.50%). The Random Forest demonstrates strong generalization capabilities, making it a reliable choice. Even models using only sunshine duration (M3) perform adequately, offering a solution when data availability is scarce. This study concludes that hybrid machine learning models, combined with careful data partitioning, significantly improve R_n estimation. These advancements provide valuable insights for climate modeling, agricultural management, and irrigation scheduling, particularly in data-scarce regions.

Keywords: Artificial intelligence, Crop water requirement, Smart irrigation, Climate change, Net radiation, Data partitioning

1. Introduction

Solar radiation is the primary energy driver of Earth's climate system and surface dynamics, playing a crucial role in agriculture by influencing plant growth, regulating temperature, and controlling evapotranspiration (ET) [1, 2]. While solar incident radiation also known as global solar radiation (R_s) is directly measured or modeled in many studies, it undergoes complex interactions with the atmosphere and the Earth's surface, leading to surface net radiation (R_n) [3]. R_n plays a pivotal role in various Earth system processes. In agriculture, it is fundamental to estimate reference evapotranspiration (ET_o), a key parameter for determining crop water requirements and optimizing irrigation schedules. [4]. Accurate R_n measurements are also crucial for understanding and predicting weather patterns, assessing water resources, and modeling climate change impacts [5, 6]. R_n represents the balance between incoming and outgoing energy at the Earth's surface, encompassing solar and longwave radiation. However, despite its significance, R_n is not readily available in most locations, especially in Thailand [7]. Standard weather stations primarily measure R_s, while R_n measurements are less common. Its direct or physical measurement requires specialized instruments such as net radiometers, which are costly to acquire and maintain, making them inaccessible for many regions or studies [8]. This limitation hinders the widespread use of standard meteorological data for studying soil-atmosphere interactions.

In response to this challenge, researchers have explored indirect methods to estimate R_n, leveraging various approaches to overcome the limitations of direct measurement. Empirical equations, which rely on commonly available meteorological data, have been widely used to approximate R_n [9-12]. However, these methods often struggle to maintain accuracy across diverse climatic and surface conditions due to their simplified assumptions and limited adaptability [13]. Another promising avenue involves using satellite data, which provides extensive spatial and temporal coverage of radiative and meteorological parameters [14, 15]. Satellite-based approaches can complement ground-based observations by offering insights into surface albedo, cloud cover, and atmospheric properties, which are crucial for R_n estimation [6]. However, such methods may face challenges related to resolution, data availability, and pre-processing complexities [15-17]. The limitations of existing methods underscore the need for more accurate and accessible approaches to estimate R_n.

*Corresponding author.
Email address: kittisak.ph@kmitl.ac.th
doi: 10.14456/easr.2025.21

In recent years, machine learning (ML) has emerged as a powerful tool for estimating environmental variables, offering advantages over traditional empirical equations by effectively capturing complex, nonlinear relationships among variables. Numerous studies have assessed the performance of ML models for predicting R_s, each demonstrating distinct strengths and limitations [18-22]. Artificial neural networks (ANN) have shown great potential among these models. For instance, Puga-Gil et al. [23] examined the performance of various ML models—Random Forest (RF), Support Vector Machine (SVM), and ANN—in predicting global solar irradiation in Rias Baixas, Spain, using meteorological data. Their findings indicated that ANN achieved the highest accuracy during model development and extrapolating to other locations. Chen and Kartini [24] introduced a hybrid k-Nearest Neighbors-ANN (KNN-ANN) model for global solar irradiance (GSI) forecasting, incorporating meteorological data from a PV station and nearby sites. Using KNN pre-processing enhanced ANN performance, enabling more precise short-term GSI predictions.

Benali et al. [25] compared the performance of Smart Persistence, Multi-Layer Perceptron Neural Network (MLP-NN), and RF models in predicting global, beam, and diffuse solar irradiance at Odeillo, France, using a three-year dataset. Despite minimal data pre-processing, RF outperformed the other models, reaffirming its effectiveness as an ensemble learning algorithm for short-term solar irradiance estimation. Vaz et al. [18] compared methods such as Ordinary Least Squares (OLS), Ridge, Lasso, KNN, SVM, Decision Trees, and RF. Among these, RF achieved superior accuracy, demonstrating its robustness in handling nonlinear data patterns while maintaining interpretability. Other researchers, such as Azad et al. [22], have estimated daily R_s in Bangladesh using machine learning models, comparing ensemble models (Bagging-REPT, RF, Bagging-RF) with standalone models (GPR, ANN, SVM). Satellite-derived data from ERA5 reanalysis and NASA POWER project datasets, including meteorological variables, were used as inputs. Results showed that RF outperformed standalone models in RMSE, with Bagging-RF also performing better than conventional models.

Beyond model selection, data pre-processing plays a crucial role in determining the accuracy of ML-based estimations. Key pre-processing steps include selecting relevant features and effectively partitioning the dataset. Feature selection is essential for reducing data dimensionality and improving the performance of proposed frameworks. Numerous methods for feature selection are widely discussed in the literature [26]. A comparative analysis by Ramírez-Rivera and Guerrero-Rodríguez [27] identified Recursive Feature Elimination (RFE) with RF as the most effective method for selecting input variables for R_s prediction, outperforming Pearson correlation, SelectKBest, and Sequential Feature Selection (SFS).

Effective data partitioning is also crucial in ML to ensure model generalization and prevent overfitting. The hold-out method is a commonly used approach, particularly in time series analysis, due to its straightforward implementation. It typically allocates 70–80% of the dataset for training while reserving the remainder for testing, with performance assessed based on test error metrics. Its main drawback is the failure to capture evolutionary trends, leading to unreliable assessments [28].

Alternatively, the standard K-Fold cross-validation split enhances model evaluation by dividing the data into k folds, training on k-1 folds, and validating on the remaining folds. Most studies using k-fold cross-validation recommend a minimum test duration of one year for practical evaluation [29, 30]. Hossein Kazemi et al. [29] investigated various k-fold cross-validation methods and the use of temporally distinct hold-out data for predicting reference evapotranspiration (ET_o) with a Gene Expression Programming (GEP) model.

Other studies have incorporated a validation set for hyperparameter tuning before testing. This practice helps prevent overfitting the model to the training data and ensures better generalization performance on unseen data. Ramírez-Rivera and Guerrero-Rodríguez [27] employed a time series cross-validation (tscv), stratified by chronological order, to partition the dataset into training and testing sets (80:20 split). A five k-fold cross-validation was then applied to the training set to create a validation set for model evaluation and tuning. Tejada et al. [31] applied four-fold growing window-forward validation (gwFV) to validate the ET_o model. Schnaubelt's [32] study mentioned that even small changes to the data can lead standard cross-validation techniques to produce highly biased and variable error estimates. In contrast, forward-validation methods provide more reliable estimates of a model's performance on unseen data.

Despite extensive research on R_s estimation using machine learning, R_n remains relatively underexplored, creating a gap in data-driven approaches for its prediction. To address this, we propose a novel hybrid ML framework that integrates KNN imputation for handling missing data, RF-RFE for feature selection, and ML models (MLP, KNN, and RF) to enhance prediction accuracy. A key innovation of this study is the systematic evaluation of multiple data partitioning techniques—hold-out split, K-fold cross-validation, and growing-window forward-validation (gwFV)—where gwFV is incorporated into hyper-parameter tuning via GridSearch to improve model robustness and reduce overfitting. The study aims to develop and assess ML models for estimating daily R_n using minimal meteorological inputs (temperature, relative humidity, and sunshine duration), implement a pre-processing approach that enhances model accuracy and efficiency, examine the impact of different meteorological variable combinations on R_n prediction accuracy in Sawi, Chumphon, Thailand, and compare three data partitioning techniques to evaluate model performance and generalization.

By achieving these objectives, this research offers a cost-effective and data-efficient alternative to traditional methods. The findings contribute to improved irrigation planning, water resource management, and environmental modeling by providing a reliable R_n estimation approach for data-scarce regions, ultimately supporting sustainable agricultural practices and climate adaptation scenarios.

2. Materials and methods

2.1 Database and study location

This study utilized a 14-year daily meteorological dataset (2008–2021) obtained from the Sawi Agro-Meteorological Station in Chumphon, Thailand. The station is situated at 10°20' N latitude and 99°6' E longitude, with an elevation of 13 meters above sea level. Sawi, Chumphon, falls under the tropical monsoon climate category (Am) based on the Köppen–Geiger climate classification [33]. The dataset includes key meteorological variables necessary for the analysis such as crop modeling, irrigation scheduling, and water balance studies.

The average values and associated standard deviations across the entire dataset were as follows: 23.49 ± 1.49°C for minimum temperature (T_{min}), 32.14 ± 2.27°C for maximum temperature (T_{max}), 27.83 ± 1.56°C for average temperature (T_{avg}), 92.18 ± 4.71% for maximum relative humidity (RH_{max}), 58.61 ± 10.27% for minimum relative humidity (RH_{min}), 75.39 ± 6.16% for average relative humidity (RH_{avg}), 5.12 ± 3.22 hours for sunshine duration (SS_h), and 1.26 ± 1.07 m/s for wind speed (W_s).

2.2 Estimation of surface net radiation (R_n)

The FAO-56 Penman-Monteith (PM) technique was applied to compute the daily R_n. This method is recommended to determine the reference crop evapotranspiration (ET_o) utilizing the FAO Penman-Monteith equation. This served as the standard and point of reference for creating and evaluating the models that were being studied. The daily net radiation is determined by subtracting the daily long-wave net radiation from the daily short-wave net radiation:

$$R_n = R_{ns} - R_{nl} \quad \text{(Eq. 1)}$$

where R_{ns} and R_{nl} are the net daily short-wave radiation and net daily long-wave radiation (MJ m^{-2} day^{-1}), respectively.

where R_s is the solar incident radiation or global solar radiation (MJ m^{-2} day^{-1}), albedo or canopy reflection coefficient, which is 0.23 for the hypothetical grass reference crop (dimensionless).

$$R_{ns} = (1 - \text{albedo})R_s \quad \text{(Eq. 2)}$$

$$R_{nl} = \sigma \left[\frac{T_{max}^4 + T_{min}^4}{2}\right] \left(0.34 - 0.14\sqrt{e_a}\right) \left(1.35\frac{R_s}{R_{so}} - 0.35\right) \quad \text{(Eq. 3)}$$

where σ is the Stefan–Boltzmann constant, which is 4.903 x 10^{-9} MJ K^{-4} m^{-2} day^{-1}; T_{max}, T_{min}, and e_a are the maximum temperature (K), minimum temperature (K), and the actual water vapor pressure (kPa), and R_s/R_{so} is relative shortwave radiation (limited to ≤ 1.0). The methodology and theoretical background for this calculation are thoroughly covered in the FAO56 publication [11, 12].

2.3 Random Forest

The Random Forest (RF) method, introduced by Breiman [34], is a supervised ensemble learning algorithm. RF generates predictions by combining outputs from multiple decision trees, each trained on a different subset of the data. For each tree in the RF model, a random portion of the training data is used to construct and train the trees. RF offers variable importance rankings, resists over-fitting and outliers, allows parallelization, and has simple hyperparameter tuning. However, they can be computationally expensive with large datasets and many trees [27].

2.4 Multi-Layer Perceptron (MLP)

Artificial Neural Networks (ANNs) are powerful ML tools that address classification and regression problems. Based on how neurons are connected, ANNs come in two distinct architectures. One prevalent type is the feedforward ANN, exemplified by the Multi-Layer Perceptron (MLP). An MLP comprises an input layer, one or more hidden layers, and an output layer of neurons. Neurons within these layers are interconnected with specific weights, and each neuron computes an output value by applying an activation function to the weighted sum of its inputs. MLP has the advantage of being able to learn complex relationships, whereas single-layer perceptron can only learn linear patterns [22]. Compared to linear statistical techniques, MLP-based ANN is a nonlinear model that is easy to use and understand.

2.5 K-Nearest Neighbors (KNN)

Cover and Hart introduced the K-Nearest Neighbors (KNN) algorithm in 1967 [35], which is utilized in this study for regression analysis. KNN is also utilized for data mining and imputation techniques [36]. As a non-parametric, instance-based learning method, KNN estimates the target variable by identifying and analyzing the nearest data points within the feature space. Unlike conventional regression models, KNN does not rely on a predefined functional form. Instead, it derives predictions by computing the average of the target values associated with the nearest neighbors, using either a simple means or an inverse distance-weight approach.

2.6 Pre-processing and temporal analysis

Forecasting models can be biased by missing data, and outliers in datasets may skew the analysis. To address these issues, the dataset underwent pre-processing. Missing values were imputed using the K-Nearest Neighbors (KNN) method from Scikit-learn, a robust approach to ensuring data completeness [37]. Outliers were identified using Z-score and Interquartile Range (IQR) methods. KNN imputation was subsequently applied to fill these gaps. Additionally, the RH_{avg} and T_{avg} columns were recalculated using the imputed values, ensuring the dataset's integrity and suitability for further analysis.

The stationarity of the target variable (R_n) was evaluated using the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [38] and Dickey-Fuller (ADF) [39] tests. For the ADF test, the null hypothesis (H_0) of non-stationarity was rejected (p-value < 0.05), indicating that the data were stationary. Similarly, for the KPSS test, the null hypothesis (H_0) of stationarity was not rejected (p-value > 0.05), further confirming that the R_n data was stationary. Additionally, the presence of seasonality in the data was assessed using a Seasonal-Trend decomposition using LOESS (STL) and the autocorrelation function (ACF). The analysis concluded that the dataset was not strongly seasonal, with an autocorrelation at lag 12 of 0.27 (below the threshold of 0.90). STL decomposition was applied to better understand the trend and residual components, using a periodicity of 365 days for the series.

2.7 Dataset partitioning (splitting) scenarios

Three data partitioning techniques were implemented to ensure robust model evaluation, as shown in Figure 1 and summarized in Table 1. First, a simple hold-out split was applied, with 80% of the data used for training and 20% for testing. This method provides a quick performance estimate but may not fully capture the variability in the dataset. Second, standard K-fold cross-validation with seven folds was utilized, ensuring that each fold used a two-year test set while the remaining data was used for training. This technique helps

reduce bias and variance by allowing all data points to be used for both training and testing at different stages. Lastly, a growing-window forward-validation split was employed, where the training set expands over time while maintaining a fixed two-year test set across seven folds. This approach is particularly useful for time-series data, as it simulates real-world forecasting scenarios by progressively increasing the amount of historical data available for training [32].

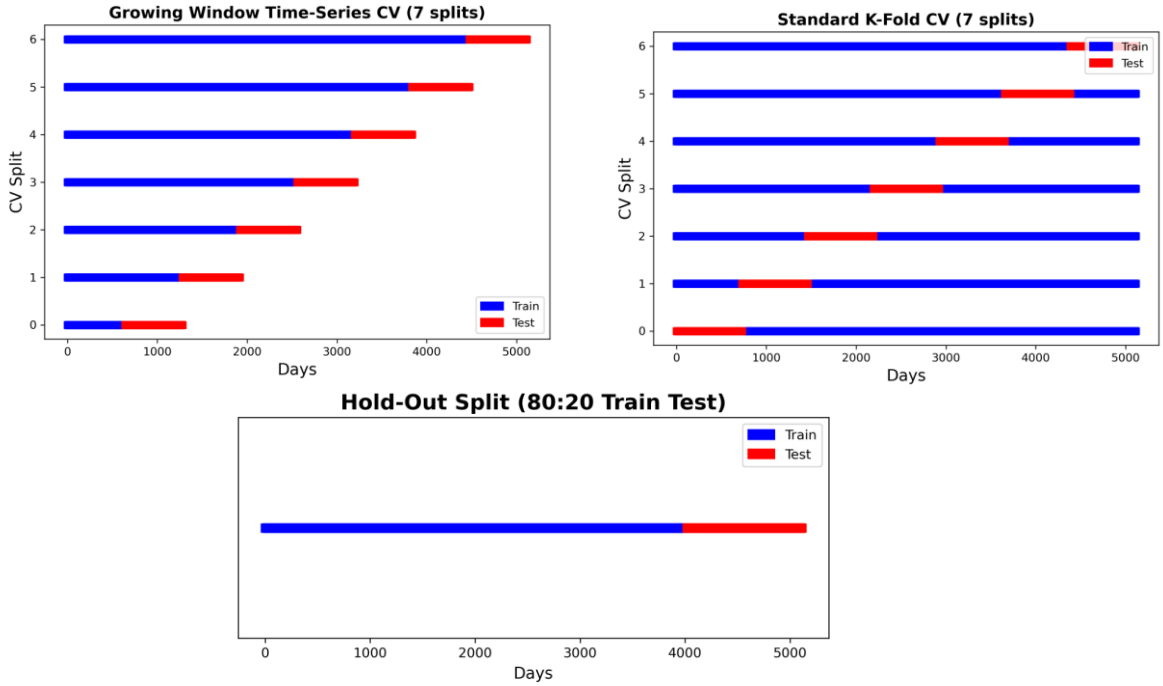

Figure 1 Different data partitioning (Scenario) techniques applied in the study

Table 1 Comparison of data partitioning scenarios

Scenario	Partition technique	Train description	Test description
S1	Hold Out (80/20% split)	The first 10 years were used as training set	The last 4 years were reserved for testing set
S2	Standard K-fold (k=7)	Randomly split into 7 folds	Each fold is used as a test once (2 years)
S3	Growing-window forward-validation (k=7)	Expanding training window every fold	Fixed test window (2 years)

2.8 Feature selection

In this study, feature selection was carried out using Recursive Feature Elimination (RFE), a wrapper-based method combined with an RF (RF-RFE), which progressively eliminated the least important features to retain an optimal subset for enhanced R_n prediction accuracy [2]. RF-RFE was configured to identify the top three features most relevant to R_n estimation. Starting from an initial set, RF-RFE iteratively removed the least important features, producing subsets of five, four, three, two, and one feature(s). This approach allowed the testing of multiple feature combinations to determine the optimal input set for the model. The RF-RFE technique identified the most relevant features for training the machine learning models, as shown in Table 2.

Table 2 Selected features based on RF-RFE feature selection used in ML models

Feature set	Selected features	ML models
M1 (3 features)	T_{max}, T_{avg}, SS_h	MLP1 KNN1 RF1
M2 (2 features)	T_{avg}, SS_h	MLP2 KNN2 RF2
M3 (1 feature)	SS_h	MLP3 KNN3 RF3

2.9 Hyperparameter tuning and model development

MLP, KNN, and RF models' hyper-parameters were fine-tuned through GridSearch (GS), a widely used optimization method for machine learning models [30, 40]. GS with 10-fold time-series cross-validation was used for hyperparameter tuning to optimize the model's performance.

- MLP: Various hidden layer configurations, including (64, 64), (64, 128), and (128, 128), were assessed. Different learning rate strategies (constant, invscaling, and adaptive) were explored, with early stopping applied after 20 iterations to mitigate overfitting. The ReLU activation function was used, and Adam served as the optimizer.
- KNN: Tuned n_neighbors (3, 5, 7, 9) with uniform and distance weighting. Explored Manhattan (p=1) and Euclidean (p=2) distance metrics.

- RF: Tested n_estimators (100, 200, 300, 400), max_depth (10, 20, 30), min_samples_leaf (5, 10, 20), and feature selection methods (sqrt, log2).

The best hyperparameter settings selected based on GS cross-validation for S1-S3 are presented in Table 3-5.

Table 3 Tuned hyperparameters by GS for S1 (Hold-out)

Models	Tuned parameters
MLP1	'hidden_layer_sizes': (128, 128), 'learning_rate': 'adaptive', 'max_iter': 140,
KNN1	'n_neighbors': 9, 'p': 2, 'weights': 'uniform'
RF1	'max_depth': 30, 'max_features': 'auto', 'min_samples_leaf': 10, 'n_estimators': 400
MLP2	'hidden_layer_sizes': (128, 128), 'learning_rate': 'adaptive', 'max_iter': 180,
KNN2	'n_neighbors': 9, 'p': 1, 'weights': 'uniform'
RF2	max_depth': 20, 'max_features': 'auto', 'min_samples_leaf': 10, 'n_estimators': 400
MLP3	'hidden_layer_sizes': (64, 128), 'learning_rate': 'adaptive', 'max_iter': 180,
KNN3	'n_neighbors': 9, 'p': 1, 'weights': 'uniform'
RF3	'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 20, 'n_estimators': 100

Table 4 Tuned hyperparameters by GS for S2 (K-Fold)

Models	Tuned parameters
MLP1	'hidden_layer_sizes': (128, 128), 'learning_rate': 'adaptive', 'max_iter': 180,
KNN1	'n_neighbors': 9, 'p': 2, 'weights': 'uniform'
RF1	'max_depth': 30, 'max_features': log2, 'min_samples_leaf': 5, 'n_estimators': 100
MLP2	'hidden_layer_sizes': (128, 128), 'learning_rate': invscaling, 'max_iter': 160,
KNN2	'n_neighbors': 9, 'p': 1, 'weights': 'uniform'
RF2	max_depth': 20, 'max_features': log2, 'min_samples_leaf': 5, 'n_estimators': 100
MLP3	'hidden_layer_sizes': (128, 128), 'learning_rate': 'adaptive', 'max_iter': 200,
KNN3	'n_neighbors': 9, 'p': 1, 'weights': 'uniform'
RF3	'max_depth': 10, 'max_features': log2, 'min_samples_leaf': 20, 'n_estimators': 300

Table 5 Tuned hyperparameters by GS for S3 (gwFV)

Models	Tuned parameters
MLP1	'hidden_layer_sizes': (64, 64), 'learning_rate': constant, 'max_iter': 200,
KNN1	'n_neighbors': 9, 'p': 1, 'weights': 'uniform'
RF1	'max_depth': 20, 'max_features': sqrt, 'min_samples_leaf': 5, 'n_estimators': 200
MLP2	'hidden_layer_sizes': (128, 128), 'learning_rate': constant, 'max_iter': 180,
KNN2	'n_neighbors': 9, 'p': 1, 'weights': 'uniform'
RF2	max_depth': 20, 'max_features': log2, 'min_samples_leaf': 5, 'n_estimators': 100
MLP3	'hidden_layer_sizes': (64, 64), 'learning_rate': constant, 'max_iter': 180,
KNN3	'n_neighbors': 9, 'p': 1, 'weights': 'uniform'
RF3	'max_depth': 10, 'max_features': sqrt, 'min_samples_leaf': 20, 'n_estimators': 300

2.10 Performance metrics

The accuracy of the trained, validated, and predicted net radiation (R_n) data from three ML models (MLP, KNN, RF) were compared to the FAO-56 PM R_n standard using root mean square error (RMSE), the coefficient of determination (R^2), and mean absolute percentage error (MAPE). Higher R^2 values, ideally approaching 1, indicate superior model performance, reflecting a regression line that closely aligns with the data. Conversely, lower RMSE and MAPE values reflect better model accuracy.

3. Results and discussion

3.1 Performance of different models and input features in estimating R_n

This study evaluates the performance of MLP, KNN, and RF models in estimating R_n using three input feature sets: M1 (T_{max}, T_{avg}, SS_h), M2 (T_{avg}, SS_h), and M3 (SS_h only). The models were assessed using three scenarios: Hold-out (S1), K-Fold Cross-Validation (S2), and Growing-Window Forward-Validation (S3). Figure 2 illustrates the model performance comparison for R_n estimation across different input feature sets and scenarios. M1 (MLP1, KNN1, and RF1) consistently performed best. M2 (MLP2, KNN2, and RF2) exhibited similar accuracy, while M3 (MLP3, KNN3, and RF3), which relies solely on sunshine duration, showed the weakest performance across all scenarios.

Table 6 presents key performance metrics, including training and testing R^2, RMSE, and MAPE, while Figure 3 shows the R^2 scatter plot for each ML model across three scenarios. For instance, in S3, the best-performing model, MLP1, achieved the highest testing R^2 (0.974), lowest RMSE (0.443 MJ m^{-2} day^{-1}), and lowest MAPE (3.44%), reinforcing the importance of T_{max}. KNN1 and RF1 followed similar trends, with KNN1 attaining an R^2 of 0.969 and RMSE of 0.479 MJ m^{-2} day^{-1}, while RF1 had an R^2 of 0.966 and RMSE of 0.506 MJ m^{-2} day^{-1}. M2 models did not show significant declines, with MLP2 achieving an R^2 of 0.973 and an RMSE of 0.446 MJ m^{-2} day^{-1}, while KNN2 and RF2 maintained strong performance. In contrast, M3 models demonstrated significant accuracy loss, with MLP3, KNN3, and RF3 yielding much lower R^2 values (0.891–0.904) and higher RMSE (0.860–0.901 MJ m^{-2} day^{-1}). However, their performance remains within an acceptable range for practical applications, highlighting their potential for simplified modeling approaches where data availability is limited.

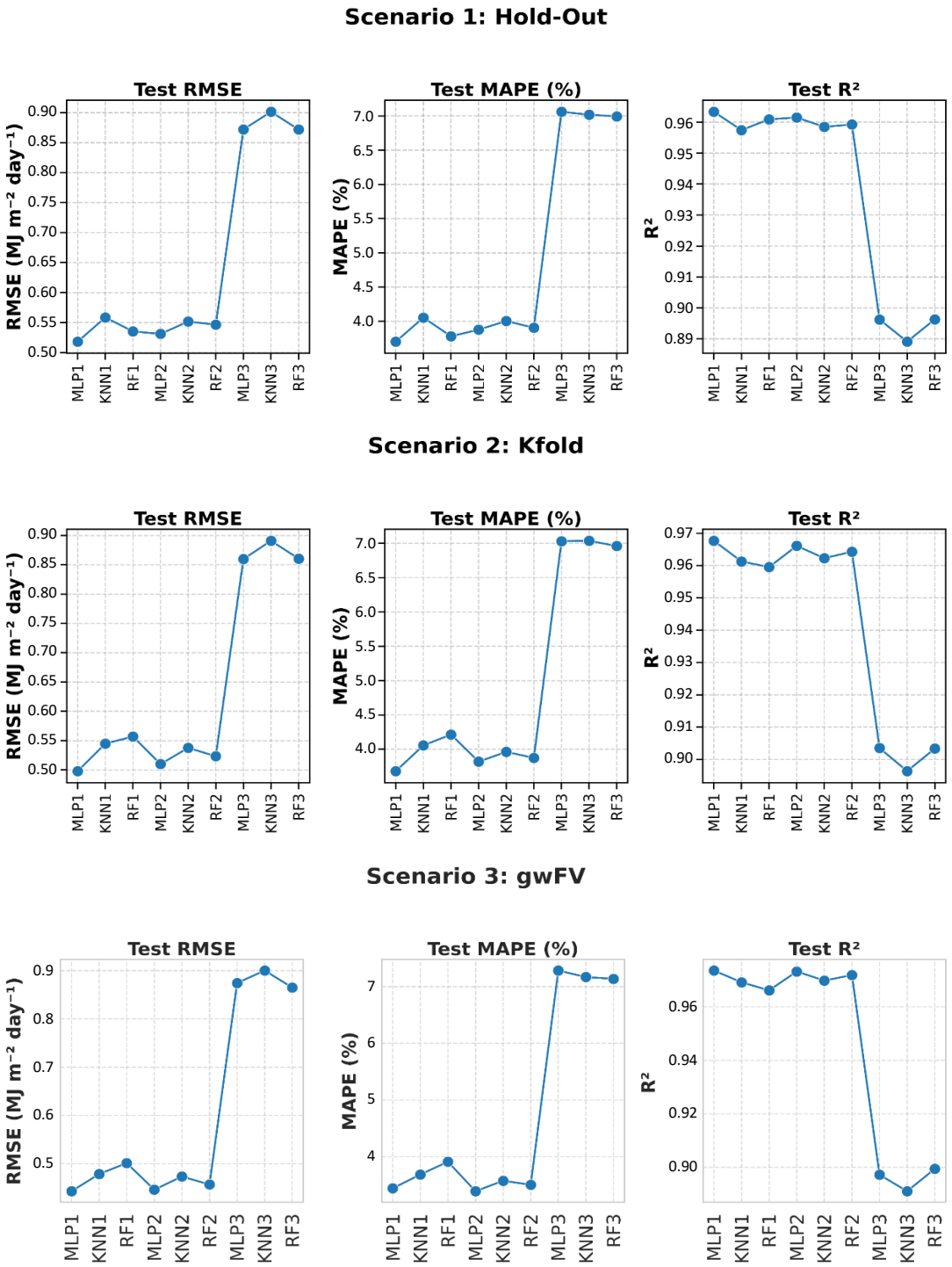

Figure 2 Model performance comparison for R_n estimation of different input feature sets and scenarios

Table 6 Training and test statistics of MLP, KNN, and RF models with different inputs across three scenarios

Model	Scenario 1 (Hold Out)						Scenario 2 (K-fold)						Scenario 3 (gwFV)					
	Training set			Test set			Training set			Test set			Training set			Test set		
	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE
MLP1	0.958	0.539	4.03	0.964	0.517	3.78	0.958	0.538	3.68	0.968	0.498	3.97	0.960	0.528	3.92	0.974	0.443	3.44
KNN1	0.962	0.513	3.79	0.957	0.559	4.05	0.962	0.513	4.06	0.961	0.545	3.75	0.963	0.506	3.71	0.969	0.479	3.68
RF1	0.967	0.481	3.54	0.961	0.535	3.78	0.965	0.490	4.21	0.960	0.557	3.64	0.968	0.475	3.56	0.966	0.506	3.91
MLP2	0.957	0.544	4.02	0.962	0.527	3.74	0.956	0.550	3.82	0.966	0.510	4.08	0.960	0.526	3.86	0.973	0.446	3.39
KNN2	0.962	0.515	3.81	0.962	0.552	4.00	0.962	0.515	3.96	0.962	0.538	3.79	0.964	0.500	3.68	0.970	0.474	3.57
RF2	0.964	0.501	3.72	0.958	0.545	3.91	0.966	0.483	3.88	0.964	0.523	3.54	0.968	0.471	3.46	0.972	0.457	3.50
MLP3	0.902	0.826	6.91	0.896	0.875	7.13	0.900	0.834	7.03	0.904	0.860	6.83	0.894	0.859	7.06	0.897	0.874	7.29
KNN3	0.894	0.860	6.79	0.889	0.901	7.02	0.893	0.860	7.04	0.896	0.891	6.82	0.886	0.889	6.95	0.891	0.891	7.17
RF3	0.904	0.817	6.76	0.896	0.871	7.00	0.903	0.822	6.96	0.903	0.861	6.76	0.897	0.848	6.93	0.899	0.899	7.14

Note: RMSE (MJ m^{-2} day^{-1}) and MAPE (%)

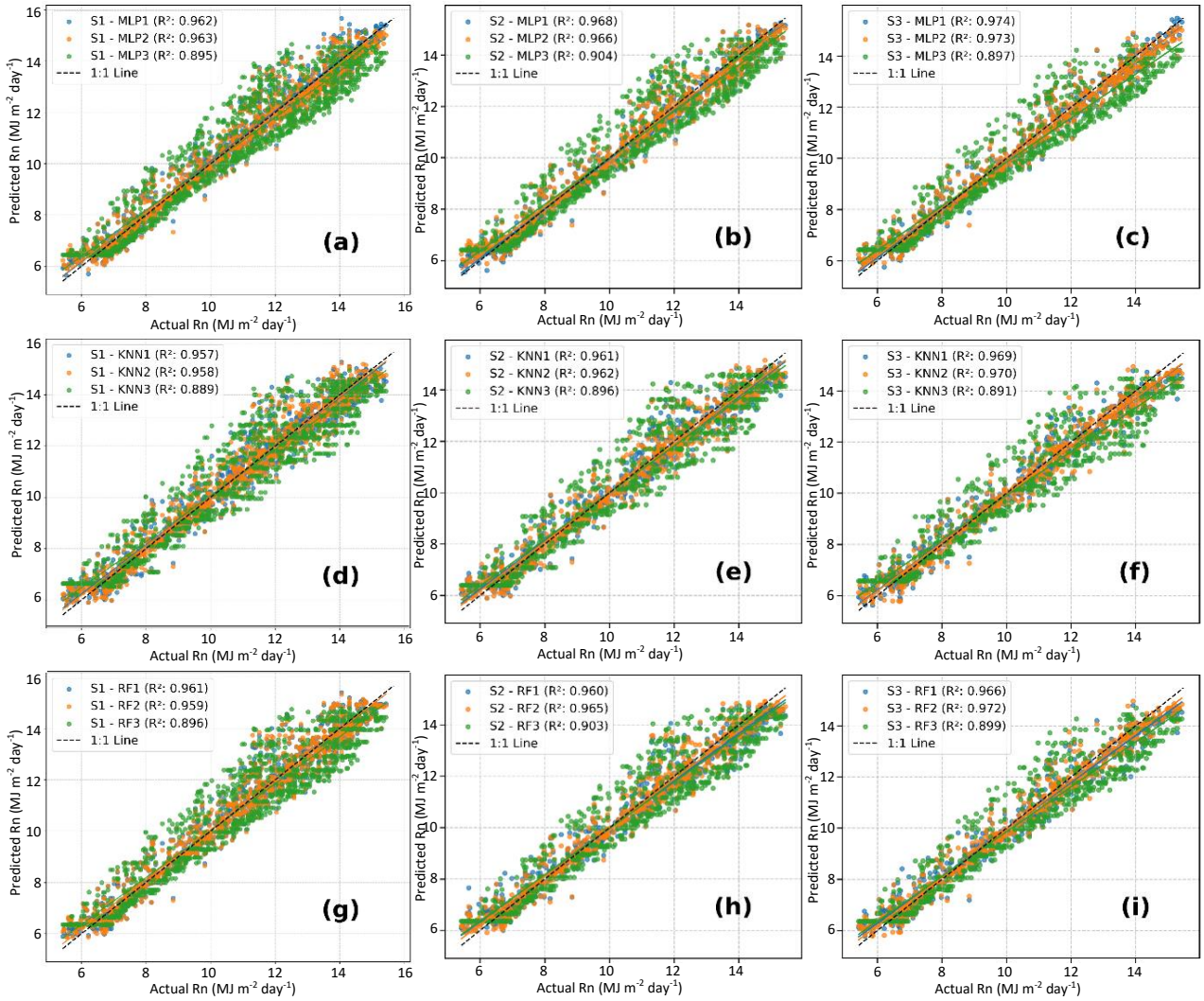

Figure 3 Scatter plots of predicted R_n by three ML models in three scenarios: (a) S1-MLP, (b) S2-MLP, (c) S3-MLP, (d) S1-KNN, (e) S2-KNN, (f) S3-KNN, (g) S1-RF, (h) S2-RF, (i) S3-RF

These findings align with research by Huang et al. [41], emphasizing the dominant role of sunshine duration (SS_h) in R_s estimation. Alizamir et al. [21] found a strong correlation between SS_h and R_s, with values of 0.91 and 0.86 at the Darbandikhan and Dukan stations in Iraq, respectively. These stations are in a hot-summer Mediterranean climate, further underscoring the relevance of SS_h. Similarly, Yu et al. [42] noted that SS_h is the most influential factor in estimating R_n across multiple climate zones, except in the subtropical monsoon zone, where its effect was less significant. Since our study is in a tropical monsoon climate, our results align with their findings in temperate and continental regions, underscoring SS_h's dominant role in R_n estimation. While incorporating T_{max} and T_{avg} improves accuracy, the moderate performance of SS_h-based models suggests its sufficiency in many applications. Future research should explore its impact in varying climatic conditions, especially where humidity and wind play a greater role.

The RF-RFE method confirms the significance of T_{max}, T_{avg}, and SS_h as input features in estimating R_n. Its effectiveness in this study is evident in its ability to identify the most relevant meteorological factors influencing R_n estimation. The method successfully selected key features, ensuring that only the most relevant variables were retained while reducing redundancy. This aligns with the findings of previous studies [27, 43] that have highlighted the ability of RFE to effectively eliminate highly correlated variables and identify the most influential predictors, ultimately leading to improved model accuracy.

3.2 Performance stability of ML models across different data partitioning scenarios and the role of data pre-processing

Figure 4 illustrates RMSE stability measured as the percentage difference between training and testing RMSE values across various models and scenarios. This section critically analyzes these findings to assess model robustness and consistency. Under the hold-out scenario (S1) for M1 and M2, MLP2 exhibited the highest stability, demonstrating the lowest RMSE percentage change (-3.13%), closely followed by MLP1 (-4.08%). This indicates minimal performance degradation between training and test sets. KNN2, RF2, and KNN1 showed moderate RMSE variations of 7.18%, 8.78%, and 8.97%, respectively, while RF1 exhibited the largest discrepancy at 11.23%. For M3, KNN3 demonstrated the highest stability (4.77%), whereas MLP3 (5.93%) and RF3 (6.61%) showed moderate stability. In the K-fold validation scenario (S2), KNN2 achieved the highest stability, with the lowest RMSE change (4.47%), followed by KNN1 (6.24%). RF2, MLP2, and MLP1 displayed moderate stability levels at 8.28%, -7.27%, and -7.43%, respectively, whereas RF1 exhibited the most considerable instability (13.67%). Within M3, MLP3 demonstrated the highest stability (3.12%), followed closely by KNN3 (3.60%) and RF3 (4.74%). The growing window-forward validation scenario (S3) further highlighted differences in stability trends. For M1 and M2, RF2 exhibited the highest stability (-2.97%), followed by KNN2 (-5.20%) and KNN1 (-5.34%).

Conversely, MLP1 and MLP2 experienced the greatest instability, with RMSE variations of -16.10% and -15.21%, respectively. In M3, KNN3 demonstrated remarkable stability, with a minimal RMSE change (0.22%), followed by MLP3 (1.75%), while RF3 showed a moderate RMSE variation (6.01%).

Overall, the findings suggest that KNN models generally exhibit superior stability across different validation strategies (S1-S3) and model variations (M1-M3), followed by MLP and RF models. However, stability alone does not necessarily imply superior predictive performance. For instance, in S3 under M1, KNN1 achieved the highest stability yet yielded a lower test R^2 (0.969) and higher RMSE (0.479) and MAPE (3.68%) compared to MLP1, which, despite lower stability, attained the highest test R^2 (0.974) alongside the lowest RMSE (0.443) and MAPE (3.44%). A similar trend was observed with MLP2, which closely matched MLP1 in accuracy and stability. These results highlight the inherent trade-off between model stability and accuracy.

Notably, RF2 emerged as an exception, delivering high predictive accuracy without significant stability compromise, reinforcing the importance of balancing both factors in model selection.

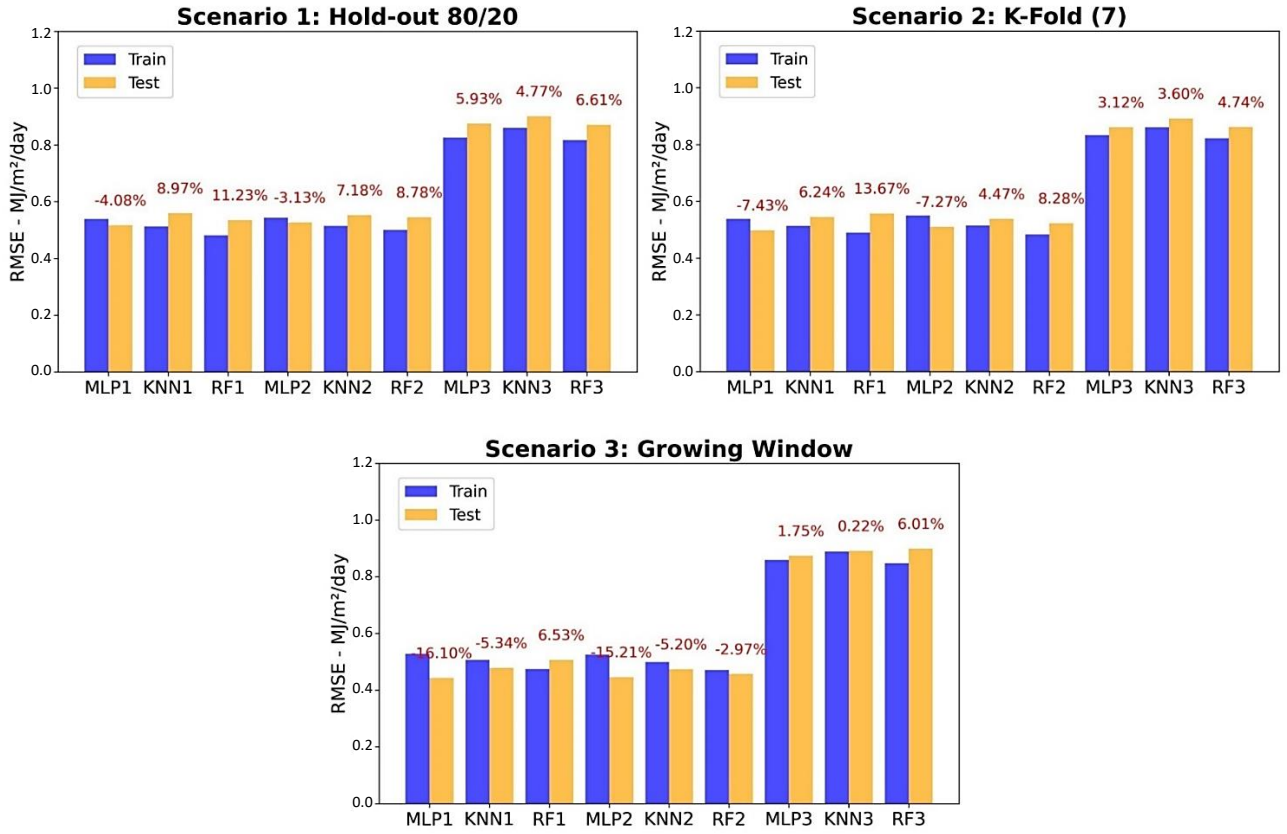

Figure 4 Stability of ML models based on RMSE (Train vs Test) across different scenarios

These findings align with prior research on machine learning applications in environmental modeling. For example, Pagano et al. [36] demonstrated the strong predictive power of both MLP and RF in estimating actual evapotranspiration. Their study found RF to be the best-performing model while also demonstrating that reducing the number of input features did not significantly impact predictive accuracy. This observation aligns with our findings, particularly in M1 and M2, where accuracy and performance remained comparable despite variations in input features, suggesting the feasibility of selectively deploying minimal in-field sensors.

On the other hand, Puga-Gil et al. [23] reported that MLP consistently outperformed RF and SVM in predicting monthly global solar irradiation. This aligns with our observations that MLP models, despite variations in stability, can generalize effectively across different datasets when properly trained. Conversely, Yamaç and Todorovic [44] found that KNN performed best when meteorological data were limited, whereas MLP achieved the highest accuracy when all variables were available. This aligns with the observed trade-off in our study, where KNN demonstrated superior stability in several cases but was often outperformed by MLP in terms of test accuracy. Santos et al. [45] further support the effectiveness of MLP and RF for environmental modeling, particularly in data-limited regions. This reinforces our findings that RF and MLP can achieve strong performance even in challenging environmental conditions when appropriately trained and optimized.

Data partitioning significantly impacts model evaluation. Our findings align with the study of Hossein Kazemi et al. [29], which demonstrates that K-Fold (S2) is superior to hold-out (S1). Furthermore, gwFV (S3) emerged as the most effective strategy, minimizing look-ahead bias and providing a more realistic out-of-sample error estimate, especially in dynamic datasets [32]. Hyperparameter tuning, a crucial yet subjective process, can significantly influence results, highlighting the importance of careful consideration and consistent practices [10].

Advanced denoising techniques like wavelet transforms could enhance machine learning model performance by effectively isolating and removing noise while preserving key features. Shiri [46] demonstrated that integrating wavelet transforms with RF significantly improved evapotranspiration estimation accuracy. This suggests potential benefits for net R_n estimation. Future work will explore integrating wavelet-based denoising by decomposing meteorological input data (e.g., T_{max}, T_{avg}, SS_h), denoising components, and reconstructing signals before model input. This approach aims to mitigate noise effects and improve model robustness, especially when data quality is suboptimal.

To further contextualize our findings, we compare them with recent studies in the field of solar radiation and environmental modeling. For instance, Landeras et al. [47] found that neural network-based approaches, particularly MLP, consistently outperformed other methods regarding accuracy and robustness. This aligns with our observations, where MLP models demonstrated superior performance in estimating R_n, especially when incorporating key input features such as T_{max}, T_{avg}, and SS_h. The strong performance of MLP in both studies underscores its versatility and effectiveness in solar radiation modeling, even across different geographic and climatic contexts. Additionally, Ikram et al. [48] emphasized the importance of optimizing machine learning models through advanced algorithms, which resonates with our findings on the impact of hyperparameter tuning and feature selection. For example, using GridSearchCV for hyperparameter optimization and RFE-RF for feature selection significantly enhanced model accuracy and stability. Their work reinforces that optimization strategies are critical for maximizing model performance in environmental applications.

In a recent study, Alizamir et al. [49] proposed a hybrid deep learning model combining Long Short-Term Memory (LSTM) networks with wavelet transforms to improve daily solar radiation (R_s) prediction. Their approach leverages LSTM's ability to capture temporal dependencies and wavelet transforms' denoising capabilities, particularly for handling non-stationary and noisy environmental data, resulting in superior predictive performance. In contrast, our study focuses on stationary data, as confirmed in the methodology, allowing us to optimize traditional machine learning models (MLP, KNN, and RF) for R_n estimation without the need for advanced techniques to address non-stationarity. However, their findings suggest that future research could explore hybrid approaches, such as integrating wavelet transforms or deep learning architectures, to further enhance model robustness, particularly in datasets where non-stationarity is a concern.

4. Conclusion

This study successfully developed and evaluated a novel hybrid machine-learning framework for estimating daily net radiation (R_n) using minimal meteorological inputs. The integration of KNN imputation, RF-RFE feature selection, and machine learning models (MLP, KNN, and RF) effectively enhanced prediction accuracy. A key contribution was the rigorous evaluation of data partitioning techniques, including the hold-out split, K-fold cross-validation, and growing-window forward-validation (gwFV), with time-series cross-validation incorporated into hyperparameter tuning for improved robustness.

Our findings highlight the importance of including average temperature (T_{avg}) for accurate R_n estimation, as models incorporating this variable (M1 and M2) generally outperformed those using only sunshine duration (M3). While M3 (SS_h only) exhibited slightly lower accuracy, its performance remained acceptable for practical applications, demonstrating the potential for simplified modeling approaches in data-limited situations. RF-RFE effectively identified relevant meteorological features, optimizing model efficiency. While K-Nearest Neighbors (KNN) demonstrated superior stability in some validation scenarios, RF and MLP emerged as highly reliable models due to their robustness and ability to handle complex, nonlinear relationships in the data.

This research offers a cost-effective and data-efficient approach for Rn estimation, particularly valuable in data-scarce regions, and contributes to improved water resource management and sustainable agricultural practices.

Future research should address the study's limitations, particularly its limited geographic scope, by testing the framework across diverse climates and broader datasets. Specifically, evaluating the model's performance in regions where factors such as humidity and wind speed have a more direct impact on R_n estimation (e.g., coastal areas, and humid tropical climates) is essential. This will help assess the framework's robustness and generalizability under varying environmental conditions.

5. Acknowledgements

The authors would like to thank the Sawi Agro-Meteorological (Sawi Agromet) Station in Chumphon, Thailand, Meteorological Department, for sharing their meteorological dataset.

6. References

[1] Gürel AE, Ağbulut Ü, Bakır H, Ergün A, Yıldız G. A state of art review on estimation of solar radiation with various models. Heliyon. 2023;9(2):e13167.
[2] Hissou H, Benkirane S, Guezzaz A, Azrour M, Beni-Hssane A. A novel machine learning approach for solar radiation estimation. Sustainability. 2023;15(13):10609.
[3] Carmona F, Rivas R, Kruse E. Estimating daily net radiation in the FAO Penman–Monteith method. Theor Appl Climatol. 2017;129(1):89-95.
[4] Dimitriadou S, Nikolakopoulos KG. Artificial neural networks for the prediction of the reference evapotranspiration of the Peloponnese Peninsula, Greece. Water. 2022;14(13):2027.
[5] Wu B, Liu S, Zhu W, Yan N, Xing Q, Tan S. An improved approach for estimating daily net radiation over the Heihe River Basin. Sensors. 2017;17(1):86.
[6] Liu X, Zhang J, Yan H, Yang H. Estimation of the surface net radiation under clear-sky conditions in areas with complex terrain: a case study in Haihe River Basin. Front Ecol Evol. 2022;10:935250.
[7] Tohsing K, Phoemwong C, Uearsri C, Saiplang P. An estimation of net radiation from global solar radiation in the main regions of Thailand. J Phys: Conf Ser. 2023;2431(1):012021.
[8] Carmona F, Rivas R, Caselles V. Development of a general model to estimate the instantaneous, daily, and daytime net radiation with satellite data on clear-sky days. Remote Sens Environ. 2015;171:1-13.
[9] Flumignan DL, Rezende MKA, Comunello É, Fietz CR. Empirical methods for estimating reference surface net radiation from solar radiation. Engenharia Agrícola. 2018;38(1):32-7.
[10] Jiang B, Zhang Y, Liang S, Wohlfahrt G, Arain A, Cescatti A, et al. Empirical estimation of daytime net radiation from shortwave radiation and ancillary information. Agric For Meteorol. 2015;211-212:23-36.
[11] Allen RG, Walter IA, Elliott RL, Howell TA, Itenfisu D, Jensen ME, et al. The ASCE standardized reference evapotranspiration equation. USA: ASCE; 2005.
[12] Allen R, Pereira L, Raes D, Smith M. FAO Irrigation and drainage paper No. 56. Rome: Food and Agriculture Organization of the United Nations; 1998.

[13] Gupta S, Singh AK, Mishra S, Vishnuram P, Dharavat N, Rajamanickam N, et al. Estimation of solar radiation with consideration of terrestrial losses at a selected location—a review. Sustainability. 2023;15(13):9962.
[14] Tang W, Yang K, Qin J, Li X, Niu X. A 16-year dataset (2000–2015) of high-resolution (3 h, 10 km) global surface solar radiation. Earth Syst Sci Data. 2019;11(4):1905-15.
[15] Li S, Jiang B, Liang S, Peng J, Liang H, Han J, et al. Evaluation of nine machine learning methods for estimating daily land surface radiation budget from MODIS satellite data. Int J Digit Earth. 2022;15(1):1784-816.
[16] Arshad MJ, Ali S, Khan SN, Arshad A, Liu J, Mumtaz F, et al. Multispectral assessment of net radiations using comprehensive multi-satellite data. Water. 2024;16(23):3378.
[17] Ramírez-Cuesta JM, Vanella D, Consoli S, Motisi A, Minacapilli M. A satellite stand-alone procedure for deriving net radiation by using SEVIRI and MODIS products. Int J Appl Earth Obs Geoinf. 2018;73:786-99.
[18] Vaz PJ, Schütz G, Guerrero C, Cardoso PJS. Hybrid neural network based models for evapotranspiration prediction over limited weather parameters. IEEE Access. 2023;11:963-76.
[19] Sohrabi Geshnigani F, Golabi MR, Mirabbasi R, Tahroudi MN. Daily solar radiation estimation in Belleville station, Illinois, using ensemble artificial intelligence approaches. Eng Appl Artif Intell. 2023;120:105839.
[20] Belmahdi B, Louzazni M, Marzband M, El Bouardi A. Global solar radiation forecasting based on hybrid model with combinations of meteorological parameters: Morocco case study. Forecasting. 2023;5(1):172-95.
[21] Alizamir M, Othman Ahmed K, Shiri J, Fakheri Fard A, Kim S, Heddam S, et al. A new insight for daily solar radiation prediction by meteorological data using an advanced artificial intelligence algorithm: deep extreme learning machine integrated with variational mode decomposition technique. Sustainability. 2023;15(14):11275.
[22] Azad MAK, Mallick J, Islam ARMT, Ayen K, Hasanuzzaman M. Estimation of solar radiation in data-scarce subtropical region using ensemble learning models based on a novel CART-based feature selection. Theor Appl Climatol. 2024;155(1):349-69.
[23] Puga-Gil D, Astray G, Barreiro E, Gálvez JF, Mejuto JC. Global solar irradiation modelling and prediction using machine learning models for their potential use in renewable energy applications. Mathematics. 2022;10(24):4746.
[24] Chen CR, Kartini UT. k-Nearest neighbor neural network models for very short-term global solar irradiance forecasting based on meteorological data. Energies. 2017;10(2):186.
[25] Benali L, Notton G, Fouilloy A, Voyant C, Dizene R. Solar radiation forecasting using artificial neural network and random forest methods: application to normal beam, horizontal diffuse and global components. Renew Energy. 2019;132:871-84.
[26] Dhal P, Azad C. A comprehensive survey on feature selection in the various fields of machine learning. Appl Intell. 2022;52(4):4543-81.
[27] Ramírez-Rivera FA, Guerrero-Rodríguez NF. Ensemble learning algorithms for solar radiation prediction in Santo Domingo: measurements and evaluation. Sustainability. 2024;16(18):8015.
[28] Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf Sci. 2012;191:192-213.
[29] Hossein Kazemi M, Shiri J, Marti P, Majnooni-Heris A. Assessing temporal data partitioning scenarios for estimating reference evapotranspiration with machine learning techniques in arid regions. J Hydrol. 2020;590:125252.
[30] Elzain HE, Abdalla OA, Abdallah M, Al-Maktoumi A, Eltayeb M, Abba SI. Innovative approach for predicting daily reference evapotranspiration using improved shallow and deep learning models in a coastal region: a comparative study. J Environ Manage. 2024;354:120246.
[31] Tejada AT Jr, Ella VB, Lampayan RM, Reaño CE. Modeling reference crop evapotranspiration using Support Vector Machine (SVM) and Extreme Learning Machine (ELM) in region IV-A, Philippines. Water. 2022;14(5):754.
[32] Schnaubelt M. A comparison of machine learning model validation schemes for non-stationary time series data. FAU Discussion Papers in Economics, No. 11/2019. Nürnberg: Friedrich-Alexander-Universität Erlangen-Nürnberg; 2019.
[33] Phumkokrux N. A study of Köppen-Geiger climate classification change in Thailand from 1987–2017. In: Monprapussorn S, Lin Z, Sitthi A, Wetchayont P, editors. Geoinformatics for Sustainable Development in Asian Cities; 2018 Jul 19-20; Bangkok, Thailand. Cham: Springer; 2020. p. 109-17.
[34] Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
[35] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21-7.
[36] Pagano A, Amato F, Ippolito M, De Caro D, Croce D, Motisi A, et al. Machine learning models to predict daily actual evapotranspiration of citrus orchards under regulated deficit irrigation. Ecol Inform. 2023;76:102133.
[37] Juna A, Umer M, Sadiq S, Karamti H, Eshmawi AA, Mohamed A, et al. Water quality prediction using KNN imputer and multilayer perceptron. Water. 2022;14(17):2592.
[38] Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root?. J Econom. 1992;54(1-3):159-78.
[39] Dickey DA, Fuller WA. Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc. 1979;74(366):427-31.
[40] Youssef MA, Peters RT, El-Shirbeny M, Abd-ElGawad AM, Rashad YM, Hafez M, et al. Enhancing irrigation water management based on ETo prediction using machine learning to mitigate climate change. Cogent Food Agric. 2024;10(1):2348697.
[41] Huang L, Kang J, Wan M, Fang L, Zhang C, Zeng Z. Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. Front Earth Sci. 2021;9:596860.
[42] Yu H, Jiang S, Chen M, Wang M, Shi R, Li S, et al. Machine learning models for daily net radiation prediction across different climatic zones of China. Sci Rep. 2024;14(1):20454.
[43] Hissou H, Benkirane S, Guezzaz A, Abderrahim B. Accurate solar radiation forecasting using an effective time series with feature selection [Internet]. Research Square [Preprint]. 2023 [cited 2025 Feb 10]. Available from: https://www.researchsquare.com/article/rs-2421924/v1.
[44] Yamaç SS, Todorovic M. Estimation of daily potato crop evapotranspiration using three different machine learning algorithms and four scenarios of available meteorological data. Agric Water Manag. 2020;228:105875.
[45] Santos PABd, Schwerz F, Carvalho LG, Baptista VBS. Machine learning and conventional method for reference evapotranspiration estimation using limited climatic data scenarios [Internet]. Research Square [Preprint]. 2022 [cited 2025 Feb 10]. Available from: https://www.researchsquare.com/article/rs-2002124/v1.

[46] Shiri J. Improving the performance of the mass transfer-based reference evapotranspiration estimation approaches through a coupled wavelet-random forest methodology. J Hydrol. 2018;561:737-50.
[47] Landeras G, López JJ, Kisi O, Shiri J. Comparison of gene expression programming with neuro-fuzzy and neural network computing techniques in estimating daily incoming solar radiation in the Basque Country (Northern Spain). Energy Convers Manag. 2012;62:1-13.
[48] Ikram RMA, Dai HL, Ewees AA, Shiri J, Kisi O, Zounemat-Kermani M. Application of improved version of multi verse optimizer algorithm for modeling solar radiation. Energy Rep. 2022;8:12063-80.
[49] Alizamir M, Shiri J, Fard AF, Kim S, Gorgij AD, Heddam S, et al. Improving the accuracy of daily solar radiation prediction by climatic data using an efficient hybrid deep learning model: Long Short-Term Memory (LSTM) network coupled with wavelet transform. Eng Appl Artif Intell. 2023;123:106199.