Theoretical and Applied Climatology (2024) 155:5279–5296
https://doi.org/10.1007/s00704-024-04953-3

RESEARCH

Estimation of crop evapotranspiration using statistical and machine learning techniques with limited meteorological data: a case study in Udham Singh Nagar, India

Anurag Satpathi[1] · Abhishek Danodia[2] · Ajeet Singh Nain[1] · Makrand Dhyani[3] · Dinesh Kumar Vishwakarma[4] · Ahmed Z. Dewidar[5,6] · Mohamed A. Mattar[5,6,7]

Received: 20 December 2023 / Accepted: 25 March 2024 / Published online: 3 April 2024

Abstract

Accurate forecasting of daily evapotranspiration (ET) is essential for enhancing real-time irrigation scheduling and informed decision-making in water resources allocation. This study investigates the intricate relationships between meteorological variables and evapotranspiration (ET) to enhance the accuracy of ET estimation models. Robust correlations were identified, emphasizing the significance of net radiation (Rn) in predicting ET. The study explores three distinct scenarios, incorporating different combinations of weather variables as input. The first scenario incorporates all weather variables, including date and time, as inputs for model development. The second scenario utilizes only Rn as input to predict ET values. In the third and final scenario, all weather variables, along with date and time, are employed as inputs for comprehensive model development. The multivariate linear regression (MLR) model demonstrated exceptional performance when exclusively using Rn, achieving an impressive R^2 value of 0.99 in both calibration and validation phases. However, limitations were observed when Rn was excluded, highlighting the necessity of a comprehensive set of input data. Penalized regression models, including ridge regression, LASSO, and ELNET, exhibited improved performance with the inclusion of Rn, supporting the importance of this variable in refining ET estimates. Machine learning models displayed remarkable performance, with most achieving R^2 values exceeding 0.95 in scenarios involving extensive input data. The Support Vector Regression (SVR) model faced challenges, indicating potential overfitting in certain scenarios. In scenarios with limited input data, machine learning models exhibited varying performance, with the Random Forest (RF) model emerging as the most robust model with R^2 value of 0.99 and 0.84 during the calibration and validation, respectively.

✉ Dinesh Kumar Vishwakarma
dinesh.vishwakarma4820@gmail.com

✉ Mohamed A. Mattar
mmattar@ksu.edu.sa

Anurag Satpathi
anuragsatpathi50@gmail.com

Abhishek Danodia
abhidanodia@iirs.gov.in

Ajeet Singh Nain
nain_ajeet@hotmail.com

Makrand Dhyani
makdhyani23@gmail.com

Ahmed Z. Dewidar
adewidar@ksu.edu.sa

[1] Department of Agrometeorology, College of Agriculture, G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand 263145, India

[2] Indian Institute of Remote Sensing, Indian Space Research Organisation, Dehradun, Uttarakhand 248001, India

[3] Department of Mathematics, School of Allied Sciences, Graphic Era Hill University, Dehradun, Uttarakhand 248002, India

[4] Department of Irrigation and Drainage Engineering, G.B. Pant, University of Agriculture and Technology, Pantnagar, Uttarakhand 263145, India

[5] Prince Sultan Bin Abdulaziz International Prize for Water Chair, Prince Sultan Institute for Environmental, Water and Desert Research, King Saud University, Riyadh 11451, Saudi Arabia

[6] Department of Agricultural Engineering, College of Food and Agriculture Sciences, King Saud University, Riyadh 11451, Saudi Arabia

[7] Agricultural Research Centre, Agricultural Engineering Research Institute (AEnRI), Giza 12618, Egypt

1 Introduction

Within the domain of hydrological research, the precise determination of crop evapotranspiration (ET) emerges as a crucial variable with far-reaching implications for water resource management, agricultural productivity, and environmental sustainability (Sagar et al. 2022; Raza et al. 2022; Mirzania et al. 2023). Mounting global population, escalating food requirements, and the spectre of climate change exert additional strain on already scarce water resources (Kumar et al. 2022b). Forecasts indicate a looming decline in crop productivity in the foreseeable future (Anapalli et al. 2016), underscoring the urgency of implementing water-conservation strategies and optimized irrigation schedules. Accurate evapotranspiration (ET) estimation provides valuable insights into the water lost through evaporation and transpiration processes, playing a pivotal role in ascertaining crop water requirements (Vishwakarma et al. 2022; Elbeltagi et al. 2023a). Studies suggest that a staggering 90% of agricultural water is lost through crop evapotranspiration in crop systems (Rana and Katerji 2000). Direct measurement of ET demands sophisticated and expensive instrumentation, including lysimeter systems, eddy covariance towers, evaporation pans, Bowen ratio stations, and scintillometer systems (Sagar et al. 2022). Given the cost, complexity, and inconvenience associated with direct measurements (Bachour et al. 2014; Jiang et al. 2016), various empirical equations, such as the Priestley-Taylor, Penman-Montieth, and Food and Agriculture Organization (FAO) crop coefficient methods, have been introduced over time. However, the accuracy of these methods hinges on the precise estimation of crop coefficients (K_c) (Kumar et al. 2021; Elbeltagi et al. 2023b). To address this challenge, the application of machine learning techniques for ET estimation has gained prominence.

The landscape of forecasting models has undergone significant evolution and refinement over time. Initially, the creation of an evapotranspiration (ET) forecast model relied on the simplicity of a stepwise multiple linear regression (SMLR) model, facilitating the identification of optimal predictors from a pool of variables within the model. However, the progression of time witnessed the displacement of these straightforward models by more sophisticated penalized regression techniques, including ridge regression, the least absolute shrinkage selection operator (LASSO), and elastic net (ELNET). In penalized regression models, the inclusion of variables is constrained or reduced to zero through penalization. Advancing beyond these methodologies, the subsequent development of diverse models saw the integration of machine learning (ML) algorithms, drawing inspiration from the intricacies of biological neuron processing (Khaniya et al. 2020; Karunanayake et al. 2020; Ekanayake et al. 2021; Tulla et al. 2024; Heddam et al. 2024).

In recent years, numerous researchers have endeavoured to develop machine learning algorithms for the estimation of crop evapotranspiration (ET) across various crops and regions (Elbeltagi et al. 2022; Mirzania et al. 2023; Vishwakarma et al. 2024). Abyaneh et al. (2011) employed artificial neural networks (ANN) and adaptive neuro-fuzzy inference system (ANFIS) to ascertain the ET requirements of garlic crops. Similarly, Aghajanloo et al. (2013) and Tabari et al. (2013) utilized a suite of approaches, including ANN, ANFIS, neural network-generic algorithm (NNGA), multivariate non-linear regression (MNLR), support vector machine (SVM), k-nearest neighbors (KNN), and AdaBoost, to estimate the evapotranspiration of potato crops. The estimation of K_c and ET values for maize and wheat crops involved the application of various models, such as Generalized Neural Regression (GRNN), fuzzy-genetic (FG), random forest (RF), deep neural network (DNN) and temporal-convolution neural network (CNN) (Feng et al. 2017; Chen et al. 2020; Saggi and Jain 2020). Notably, these studies grapple with the challenge of limited weather data availability for modelling ET values. Thus, there is a pressing need for ET estimation approaches that can effectively operate with a restricted set of weather data.

This research focuses on modelling the crop evapotranspiration process, specifically utilizing sugarcane as the target crop, given its year-round growth across all seasons. Notably, there is a dearth of studies employing machine learning for ET estimation in the Udham Singh Nagar district of Uttarakhand state. Uttarakhand, characterized by 86% hilly terrain and a mere 14% plains, faces topographical constraints, limiting cultivable land to only 14% of the total area. Moreover, 61% of the state is covered by forests (State profile, Government of Uttarakhand). Among the thirteen districts, Haridwar and Udham Singh Nagar stand out as the primary contributors to the plains. Udham Singh Nagar, chosen deliberately for this study, boasts the highest agricultural crop area in Uttarakhand (Directorate of Economics and Statistics). Situated in the Tarai belt at the foothills of the Shivalik range, approximately 80% of the crop area in Udham Singh Nagar is under irrigation (Krishi Vigyan Kendra, Udham Singh Nagar). This emphasis on irrigated land underscores the significance of studying evapotranspiration in this district for improved irrigation and water management practices.

This study introduces novel approaches and contributions to accurately estimate crop evapotranspiration (ET) for sugarcane in the Udham Singh Nagar district. The research encompasses a comprehensive set of models, including one statistical, three penalized regression and four machine learning models. To address limited weather variable availability, the study explores three scenarios: first using Net radiation (Rn) as the sole input, second excluding Rn while including all other weather variables and third incorporating all available weather variables. A thorough comparison of these models across different datasets is conducted to identify the most suitable model for ET estimation under varying data availability scenarios, aiming to enhance prediction accuracy.

2 Site description and data used

2.1 Study area

The research focuses on Udham Singh Nagar district in Uttarakhand, India, situated in the Tarai belt at the foothills of the Shivalik range of the Himalayas within the Kumaon division. The district spans from 28°53′ to 29°23′ N latitude and 78°45′ to 80°08′ E longitude, with an altitude of 214 m (Fig. 1). Known as the food bowl of Uttarakhand State, the district covers a geographical area of 3055 km^2, with approximately 5% of the land under forest. Agriculture serves as the primary occupation for the majority. The district experiences two major cropping seasons, Kharif and Rabi, with prominent crops including rice, wheat, sugarcane, and mustard.

2.2 Data collection

In the formulation of these models, a substantial volume of datasets was indispensable, encompassing both dependent and independent variables. The actual crop ET data based on the Bowen ratio, recorded at one-hour intervals (comprising 12 values per day), was computed from the flux tower situated at the research farm of GBPUAT Pantnagar. Concurrently, weather data, encompassing parameters such as temperature, relative humidity, net solar radiation, wind speed, and surface pressure, was gathered from the same farm utilizing a micro-meteorological flux tower. Of the overall datasets, 80% of the data was utilized for training the models, with the remaining 20% earmarked for model testing. The flowchart detailing the development of various models for ET estimation is delineated in Fig. 2.

The descriptive statistics of the input data for the study area are reported in Table 1.

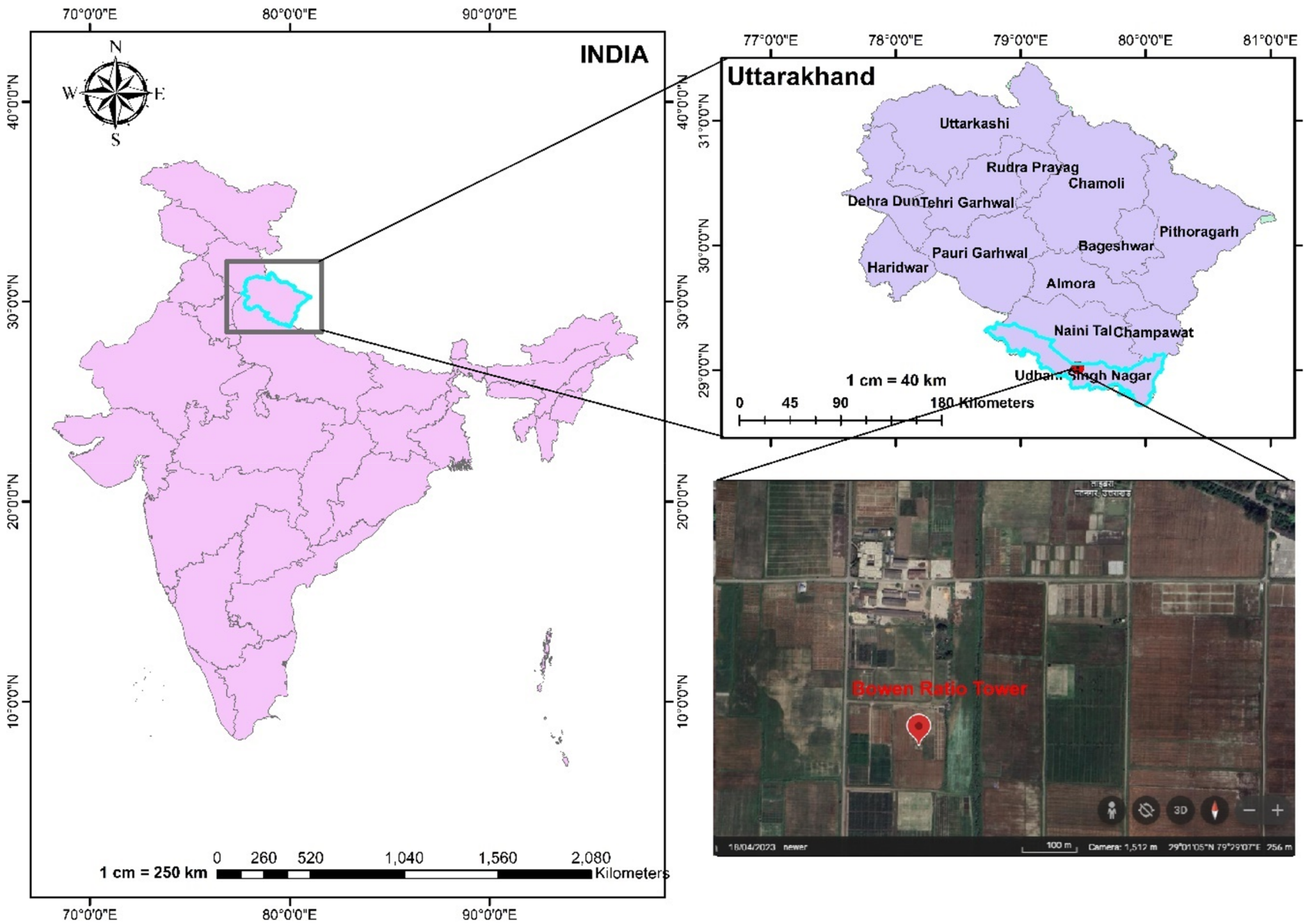


Fig. 1 Location of the Bowen ratio tower over the study area

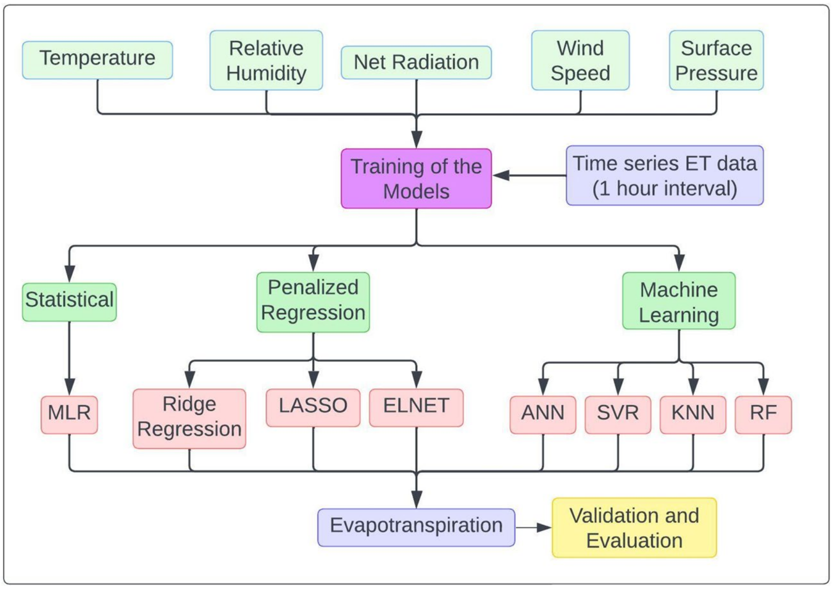


Fig. 2 Flowchart of different model development for ET estimation

Table 1 Data statistics of hourly weather parameters at study stations

Statistics parameters	ET (mm/hour)	Temp (°C)	RH (%)	Rn (W/m^2)	Wind Speed (m/sec)	Pressure (kPa)
Mean	2.54	25.20	82.89	100.23	1.20	102.74
Standard Error	0.05	0.09	0.24	2.14	0.01	0.04
Standard Deviation	5.01	8.76	22.12	200.19	1.36	3.94
Sample Variance	25.09	76.76	489.26	40077.87	1.86	15.51
Kurtosis	0.65	-0.92	-0.02	0.31	2.05	-0.43
Skewness	1.36	-0.23	-1.09	1.25	1.41	-0.86
Range	21.96	41.06	88.61	863.20	10.70	14.01

3 Methodology

3.1 Calculation of ET with Bowen ratio

To employ the Bowen ratio method for calculating evapotranspiration (ET), it is essential to measure temperature and humidity at two different heights (Malek and Bingham 1993; Peacock and Hess 2004; Buttar et al. 2018). In this study, data on temperature and humidity were gathered at 2 m and 4 m above the surface, alongside measurements of net radiation (Rn), wind speed, wind direction, and surface pressure. Net radiation (Rn) was quantified using a Net radiometer positioned on the Bowen ratio tower. The ground heat flux (G) in this investigation was considered to be 10% of the Net radiation (Kato and Yamaguchi 2007; Teixeira et al. 2009). The Bowen ratio, representing the ratio of sensible heat flux to latent heat flux, can be expressed using the formula outlined by Buttar et al. (2018).

$$\mathrm{ET} = \frac{\mathrm{R_n} - \mathrm{G}}{(1 + \beta)} \tag{1}$$

In the context of this study, denoting evapotranspiration as ET, net radiation as Rn, soil heat flux as G, and Bowen ratio as β, the calculation of β within the specified surface layers between two levels can be determined using the formula established by Verma et al. (1978):

$$\beta = \gamma \frac{\Delta \mathrm{T}}{\Delta \mathrm{e}} \tag{2}$$

In the given context, where ΔT and Δe represents the temperature and vapor pressure gradient between the two measured heights, and γ denotes the psychrometric constant, the computation of saturation vapor pressure values was conducted using the formula outlined by Allen et al. (1998):

$$\mathrm{e} = 0.6108 \exp\left(\frac{17.27 + \mathrm{T}}{\mathrm{T} + 237.3}\right) \tag{3}$$

The psychrometric constant, denoted as γ, establishes the connection between the partial pressure of water in the air and the air temperature. The formula for calculating the psychrometric constant (γ) is provided by Allen et al. (1998):

$$\gamma = \frac{\mathrm{C_p P}}{\varepsilon\lambda} \tag{4}$$

In this context, where C_p represents the specific heat at constant pressure, P is the atmospheric pressure, ε denotes the ratio of the molecular weight of water vapor to dry air, and λ stands for the latent heat of vaporization, the values of ET obtained at hourly intervals were employed for training and testing statistical and machine learning models.

3.2 Development of statistical, machine and deep learning models for ET estimation

Over the course of time, there has been a significant evolution in forecasting models. Initially, the creation of an ET forecast model involved employing a stepwise multiple linear regression (SMLR) model, facilitating the identification of optimal predictors from a pool of variables in model development. However, these rudimentary models were gradually supplanted by penalized regression models, including ridge regression, least absolute shrinkage selection operator (LASSO), and elastic net (ELNET). In these penalized regression models, the number of variables is constrained through the imposition of penalties or zero constraints. As the development of various models progressed, a diverse array of machine learning (ML) algorithms, inspired by the processing of biological neurons, made their entrance. An illustrative model in this regard is the artificial neural network (ANN), now extensively utilized across different disciplines. The ANN has proven instrumental in solving a myriad of problems across various fields (Ghiassi et al. 2005; Shukla et al. 2021; Elbeltagi et al. 2022; Saroughi et al. 2023; Mirzania et al. 2023). The neural network model exhibits intelligent learning capabilities during its training process. However, it is imperative to acknowledge certain drawbacks associated with ANN, including its intricate design, potential for offering ambiguous solutions, absence of explicit rules for network structure determination, and reliance solely on numeric information (Azzam et al. 2022). To address these challenges, numerous alternative machine learning models, such as support vector machine (SVM), random forest (RF), and sophisticated deep learning models like convolutional neural network (CNN) and deep neural network (DNN), were developed.

3.3 Model description

In the present study multiple models were developed based on three distinct sets of meteorological datasets. The first scenario involved utilizing only Rn as an input variable, as it exhibited the highest correlation coefficient when compared to other input variables. In the second scenario, all variables except Rn were employed as input variables as the net radiometer instrument is not available in all the weather observatories. Hence, to find good models which can predict ET values without data of Rn can be very useful. The third scenario incorporated all variables, including Rn, for model development. The specifics of each model are elaborated upon in the subsequent discussion:

3.3.1 Multiple linear regression (MLR)

The Multiple Linear Regression (MLR) stands as a traditional forecasting method, where regression equations are formulated using independent variables. In the context of this study, the MLR approach was compared with other advanced methods. One of the strengths of MLR lies in its capability to assist in the selection of optimal predictor variables from a vast array of candidates, as highlighted in previous works (Singh et al. 2014; Vishwakarma et al. 2018; Das et al. 2018). Notably, a significance level of 0.05 was adopted for p-values during the development of the MLR model.

3.3.2 Ridge regression

Ridge regression introduces a slight bias to predictor variables, mitigating the risk of overfitting in datasets (Li et al. 2010; Pavlou et al. 2016). Its primary objective is to enhance outcomes compared to traditional models by minimizing overfitting. This method affords researchers the capability to estimate coefficients, even in the presence of substantial correlations among predictor variables (Hilt and Seegrist 1977). The ridge regression may exhibit modest performance during training, its overall effectiveness tends to be consistently superior. The loss in ridge regression can be quantified as follows:

$$L_{ridge}(\widehat{\beta}) = \sum_{i=1}^{n}\left(y_i - x_i'\widehat{\beta}\right)^2 + \lambda\sum_{j=1}^{m}\beta_j^2 = \|y - X\widehat{\beta}\|^2 + \lambda\|\widehat{\beta}\|^2 \tag{5}$$

where x and y represent the input and output vectors, respectively. The training dataset comprises n samples, while β denotes the regression coefficient, and λ serves as the penalty parameter.

3.3.3 Least absolute shrinkage selection operator (LASSO)

LASSO, a form of penalized regression model, functions by shrinking coefficients that exhibit correlation toward zero. Positioned as a data-driven model, LASSO is designed to counteract overfitting and promote the generalization of the model. The minimization of the objective function, as articulated by Hilt and Seegrist (1977), is expressed as:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i^{\prime}\hat{\beta})^2 + \lambda \sum_{j=1}^{m} \left|\hat{\beta}_j\right| \tag{6}$$

The LASSO model incorporates a regression coefficient, denoted as β, which is linked to the input parameters. Here, x and y signify the input and output vectors, respectively. The training dataset comprises n samples, and the penalty parameter λ functions as a hyperparameter in the model.

3.3.4 Elastic Net (ELNET)

The penalty of both ridge regression and LASSO gets combined in the ELNET model (Abbas et al. 2020). In LASSO regression, a penalty in the form of the "absolute value of magnitude" is incorporated, while in ridge regression, a penalty is imposed in the form of the "squared magnitude of the coefficient." ELNET integrates both regularization techniques, and the loss can be defined as per the formulation by Zou and Hastie (2005):

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n} \left(y_i - x_i^{\prime}\hat{\beta}\right)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}^2 + \alpha\sum_{j=1}^{m}\left|\hat{\beta}_j\right|\right) \tag{7}$$

where x and y represent the input and output vectors, respectively. The training dataset consists of n samples, and the model parameters include β, the regression coefficient, λ, the penalty parameter, and α, which serves as the mixing parameter between ridge (α = 0) and LASSO (α = 1).

3.3.5 Artificial neural network (ANN)

The Artificial Neural Network (ANN) model, inspired by biological neurons akin to the human brain (Kaur and Sharma 2019; Shukla et al. 2021), is characterized by three layers: the input layer, hidden layer, and output layer. The hidden neurons incorporate an activation function, which transforms the activation level of a unit neuron into an output signal. The pivotal functions are performed within the hidden layer, and the outcomes are subsequently transmitted to the output layer. The determination of the number of nodes in the input layer is contingent upon the count of independent predictors. The expression for the output (hi) of neuron i in the hidden layer is articulated by Wang (2003):

$$h_i = \sigma\left(\sum_{j=1}^{N} V_{ij}x_j + T_i^{hid}\right) \tag{8}$$

here, σ is activation function, N is the number of input neurons, V_{ij} is the weights, x_j is the input to the neurons and T_i^{hid} is the threshold terms of the hidden neurons.

3.3.6 Support vector machine (SVM)

SVM, a machine learning algorithm primarily designed for classification and introduced by Vapnik (1998), extends its utility beyond classification to encompass tasks like time series estimation and regression analysis (Thissen et al. 2003). Its versatility extends to various applications, including hydrology, ecology, climatology, among others (Kushwaha et al. 2021; Kumar et al. 2022a; Singh et al. 2022a, b; Achite et al. 2023). SVM exhibits favourable performance in handling high-dimensional data, as noted by Azzam et al. (2022), although its efficacy diminishes in the presence of noisy and overlapped data. For a comprehensive understanding of the SVM model, interested readers can refer to the details provided by Fan et al. (2018).

3.3.7 K-nearest Neighbour (KNN) regression

The KNN, a non-parametric machine learning algorithm suitable for both classification and regression tasks, operates by assigning the value of an object in regression scenarios as the average of its K nearest neighbours, aligning with its name. KNN boasts advantages in terms of straightforward implementation and rapid real-time response. Nevertheless, its effectiveness diminishes when dealing with high-dimensional datasets and those containing noisy or missing values. For an in-depth understanding of KNN, interested readers can explore the detailed description provided by Kramer (2013).

3.3.8 Random forest (RF)

Random Forest, introduced by Breiman (2001) and investigated by Biau et al. (2008), represents a machine learning approach that integrates multiple decision trees. Each decision tree is generated independently, relying on a random vector sampled from the input data, while maintaining a consistent distribution across all trees (Azzam et al. 2022). The outcome is determined by selecting the most voted estimator among these classifications. Random Forest exhibits robust performance for both small and high-dimensional datasets. However, the model's complexity and training time requirements are notable disadvantages. For a comprehensive understanding of the RF algorithm, the mechanism of the RF model is elucidated in detail in references (Liu et al. 2012; Biau and Scornet 2016).

3.4 Model Evaluation

Assessing the accuracy of a model is a pivotal stage for both developers and users alike. The subsequent section briefly outlines the statistical parameters employed to evaluate the performance of the models:

3.4.1 Coefficient of determination (R^2)

The R^2 metric is employed to assess the linear association between the observed and predicted datasets. Ranging from 0 to 1, a value of 1 indicates a robust linear relationship. Generally, an R^2 value exceeding 0.5 is considered acceptable.

$$R^2 = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)(\widehat{y}_i - \widehat{y}_i)}{\sigma_y \sigma_{\widehat{y}}} \right)^2 \tag{9}$$

In this context, y_i denotes the observed value, $\widehat{y}_i$ is the predicted value for i = 1, 2,….n. $\bar{y}_i$ and $\overline{\widehat{y}}_i$ is the mean of observed and predicted values, respectively. σ_y and $\sigma_{\widehat{y}}$ is the standard deviation of actual and predicted values respectively.

3.4.2 Root mean square error (RMSE)

The Root Mean Square Error (RMSE) serves as a commonly employed metric to quantify the disparity between observed values from the environment and those predicted by the model. Utilizing RMSE allows for the measurement of the error existing between the two datasets. A lower RMSE value indicates superior model performance, while a higher value suggests poorer model performance. The RMSE is calculated using the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n}} \tag{10}$$

Here, y_i is the observed value while $\widehat{y}_i$ is the predicted value and n shows the number of observations. The unit of RMSE is similar to the unit of observed or predicted values.

3.4.3 Normalized root mean square error (nRMSE)

The nRMSE also called as scatter index, is a statistical error indicator which can be calculated as:

$$\text{nRMSE} = \frac{\text{RMSE}}{\overline{\text{y}}_\text{i}} \tag{11}$$

Here, $\overline{y}_i$ is the mean of observed values. It helps to compare the models with different scales. The unit of nRMSE is percentage.

3.4.4 Nash sutcliffe model efficiency coefficient (NSME)

The NSME is a normalized statistic that determines the measure of likelihood or model performance in terms of its accuracy (Nash and Sutcliffe 1970). It can be expressed in terms of equation as:

$$NSME = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{12}$$

Here, y_i is the observed value, $\widehat{y}_i$ is the predicted value and $\overline{y}_i$ is the mean of observed values. The value of NSME ranges from -∞ to 1. A proximity of the NSME value to 1 signifies enhanced model efficiency, with a value of 0 indicating model accuracy comparable to the mean accuracy of the calculated observed data. Conversely, a negative value indicates a deficiency in the model's performance.

3.4.5 Correlation coefficient (CC)

This metric estimates the intensity of the linear association between observed and predicted data values, with the correlation coefficient (CC) spanning from -1 to + 1. A CC value of -1 indicates a robust negative relationship, while a + 1 value signifies a strong positive relationship. The computation of CC is expressed as:

$$CC = \frac{n\sum_{i=1}^{n} y_i\widehat{y}_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n}\widehat{y}_i)}{\sqrt{n(\sum_{i=1}^{n} {y_i}^2) - (\sum_{i=1}^{n} y_i)^2}\sqrt{n(\sum_{i=1}^{n}\widehat{y}_i^2) - (\sum_{i=1}^{n}\widehat{y}_i)^2}} \tag{13}$$

Here, y_i is the observed value, $\widehat{y}_i$ is the predicted value.

3.4.6 Agreement index (d)

Conceived by Willmott (1981), this statistical measure serves as an accuracy assessment tool for models. The range of values for "d" extends from 0 to 1, with a value of 1 indicating a perfect match between observed and predicted values, and 0 signifying no alignment. However, it's important to note that "d" is particularly sensitive to extreme values, attributed to the squared differences. The expression for "d" is articulated as follows:

$$d = 1 - \left(\frac{\sum_{i=1}^{n}((y_i - \widehat{y}_i)^2)}{\sum_{i=1}^{n}(|\widehat{y}_i - \overline{y}_i| + |y_i - \overline{y}_i|)^2} \right) \tag{14}$$

Here, y_i is the observed value, $\hat{y}_i$ is the predicted value and $\overline{y_i}$ is the mean of observed values.

3.4.7 Mean biased error (MBE)

The MBE value denotes the average bias in predictions, where a positive MBE signifies overestimation, and a negative MBE indicates underestimation from the datasets. The computation procedure for the MBE value is outlined as follows:

$$MBE = \frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \hat{y}_i) \tag{14}$$

where, y_i and $\hat{y}_i$ is the observed and predicted value respectively.

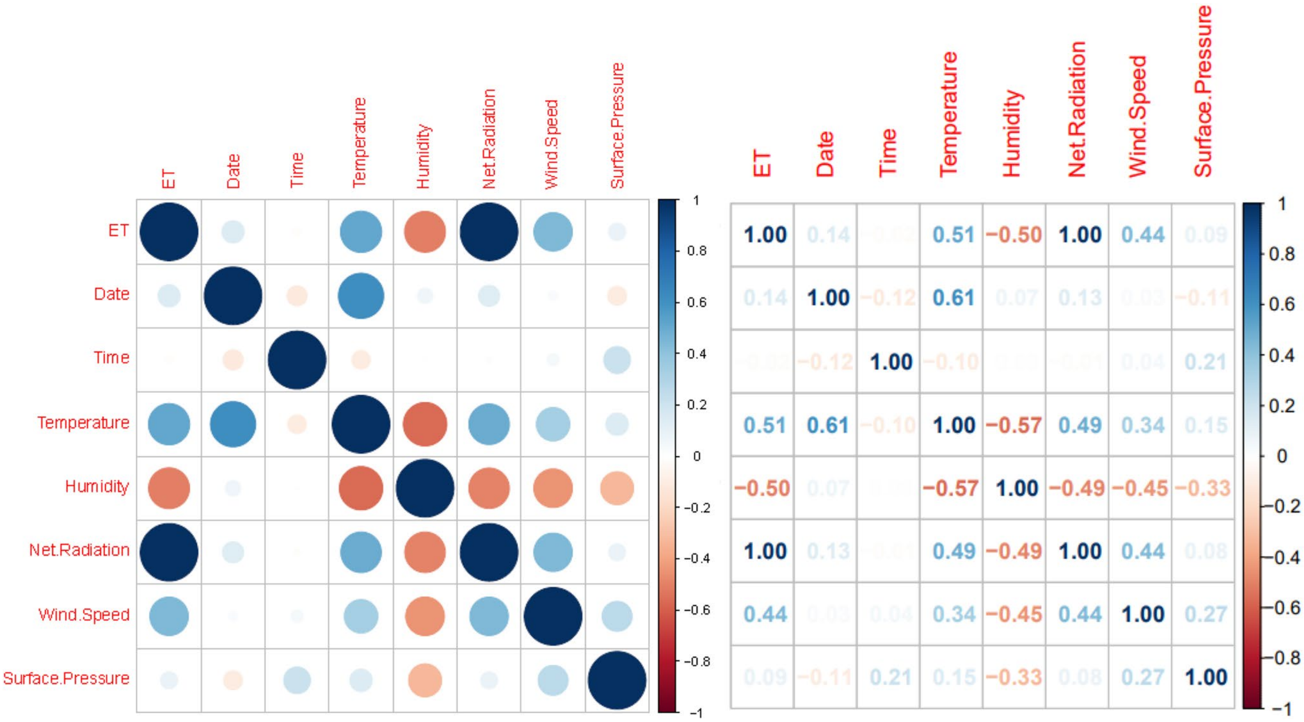
4 Results and discussion

4.1 Correlation study between ET and input weather variables

Figure 3 displays the correlational diagram along with correlation coefficient values depicting the relationships among variables. The analysis reveals robust correlations between ET values and Rn (1.00), Temperature (0.51), Relative Humidity (-0.5), and Wind speed (0.44). Additionally, positive correlations emerge between Date and Temperature (0.61) as well as Net radiation and Temperature (0.49). Conversely, negative correlation coefficients are prominent, particularly between relative humidity and other meteorological variables. This trend is logical, as an increase in net radiation, temperature, and wind speed corresponds to a decrease in the relative humidity in the atmosphere. Additionally, it is noteworthy that high relative humidity in the atmosphere leads to a decrease in the evapotranspiration (ET) value. This can be attributed to the saturated nature of the atmosphere with water, limiting the potential for further evaporation and transpiration. These findings underscore the intricate relationships between meteorological parameters and their impact on ET, contributing valuable insights to the understanding of the dynamic processes governing agricultural water consumption. The examination of these correlation coefficients led to the formulation of three distinct scenarios for model development in this study. The first scenario incorporates all weather variables, including date and time, as inputs for model development. The second scenario utilizes only Rn as input to predict ET values. In the third and final scenario, all weather variables, along with date and time, are employed as inputs for comprehensive model development.

4.2 Evaluation of statistical (MLR) model performance

Table 2 presents the outcomes of the MLR model at both the calibration and validation stages, while Table 3 showcases



	ET	Date	Time	Temperature	Humidity	Net.Radiation	Wind.Speed	Surface.Pressure
ET	1.00	0.14	−0.02	0.51	−0.50	1.00	0.44	0.09
Date	0.14	1.00	−0.12	0.61	0.07	0.13	0.03	−0.11
Time	−0.02	−0.12	1.00	−0.10	0.00	−0.01	0.04	0.21
Temperature	0.51	0.61	−0.10	1.00	−0.57	0.49	0.34	0.15
Humidity	−0.50	0.07	0.00	−0.57	1.00	−0.49	−0.45	−0.33
Net.Radiation	1.00	0.13	−0.01	0.49	−0.49	1.00	0.44	0.08
Wind.Speed	0.44	0.03	0.04	0.34	−0.45	0.44	1.00	0.27
Surface.Pressure	0.09	−0.11	0.21	0.15	−0.33	0.08	0.27	1.00

Fig. 3 Correlation between the ET and different weather variables

Table 2 Performance of MLR model to Forecast ET values

Input/Parameters	Stage	R^2	RMSE	nRMSE	NSME	MBE
Rn	Calibration	0.99	0.43	8.6%	0.99	0
	Validation	0.99	0.25	4.8%	1.00	-0.06
T, RH, WS, SP	Calibration	0.41	3.8	77.1%	0.41	0
	Validation	0.40	4.85	91.4%	0.16	-2.28
Rn, T, RH, WS, SP	Calibration	0.99	0.41	8.2%	0.99	0
	Validation	0.99	0.28	5.2%	1.00	0.06

Table 3 Equations developed in different scenarios of MLR model

Input/Parameters	Equation	Model efficiency (%)
Rn	$ET = 0.044 + 0.0248 \times Rn$	99.25%
T, RH, WS, SP	$ET = 571.99 - 0.013 \times Date + 0.002 \times Time + 0.264 \times T - 0.025 \times RH + 0.972 \times WS - 0.026 \times SP$	40.58%
Rn, T, RH, WS, SP	$ET = -22.67 + 0.0005 \times Date - 0.0008 \times Time + 0.0084 \times T - 0.0025 \times RH + 0.0244 \times Rn - 0.0011 \times WS - 0.0022 \times SP$	99.32%

the corresponding equations derived. The findings underscore the significance of net radiation (Rn) in predicting evapotranspiration (ET). Remarkably, the MLR model demonstrates outstanding performance when exclusively utilizing Rn as an input variable, achieving an impressive R^2 value of 0.99 in both calibration and validation phases. The model's excellence is further evident in other key statistical parameters, with NSME values of 0.99 and 1, as well as nRMSE values of 8.6% and 4.8% during calibration and validation, respectively. This aligns with the findings of Chia et al. (2022), who also achieved highly accurate daily ET estimates by employing Rn as the sole input variable, attaining an R^2 value of 0.96.

In the second scenario, where Rn was excluded as an input variable, there was a notable decline in model performance. The R^2 values dropped to 0.41 and 0.4 during calibration and validation, respectively. Correspondingly, other statistical parameters exhibited diminished performance, with calibration values of NSME at 0.41, RMSE at 3.8, and nRMSE at 77.1%. During validation, these values were NSME = 0.16, RMSE = 4.85, and nRMSE = 91.4%. This emphasizes the limitations of employing a statistical model like MLR for accurate ET estimation when working with a restricted set of input data. In the third scenario, where all parameters were utilized as input variables, the model performances rebounded to excellence. However, a comparison between the results of the first scenario (Only Rn) and the third scenario (All parameters) revealed a marginal improvement during calibration and a slight deterioration during validation. This echoes findings by Sattari et al. (2021), who observed similar trends when using sunshine duration (n) as the sole input variable, generating superior ET estimates compared to multiple meteorological variables (T, WS, RH, n). The scatter plot diagram illustrating all MLR models is depicted in Fig. 4.

4.3 Evaluation of Penalized regression models performance

Ridge regression, LASSO, and ELNET represent penalized regression models, implying the imposition of penalties on input parameters when dealing with numerous input variables. In the first scenario when Only Rn is the input, the number of input variables is merely one, rendering the implementation of penalties impractical. Therefore, in these three penalized regression models, only the second scenario (All parameters except Rn) and the third scenario (All parameters) are applicable. The statistical parameters evaluating the performance of penalized regression models are detailed in Table 4. In the second scenario (All parameters except Rn), the performance of all penalized models was uniformly subpar, with consistent values across statistical metrics. The R^2 value during calibration remained at 0.4 for all models, while during validation, it reached 0.27 for ridge regression and 0.4 for both LASSO and ELNET, indicating unsatisfactory model performance. Furthermore, the MBE values during validation displayed strong negative values, indicative of underestimation by all the penalized models. Consequently, caution is advised against employing penalized regression models for ET estimation when working with a limited set of input variables.

In the preceding scenario where Rn was integrated as an input variable, there was a remarkable enhancement in the performance of the penalized models. The R^2 and NSME values surpassed 0.98 for both calibration and validation across all penalized models, signifying an outstanding

Fig. 4 Scatter plot between observed ET and predicted ET of MLR model

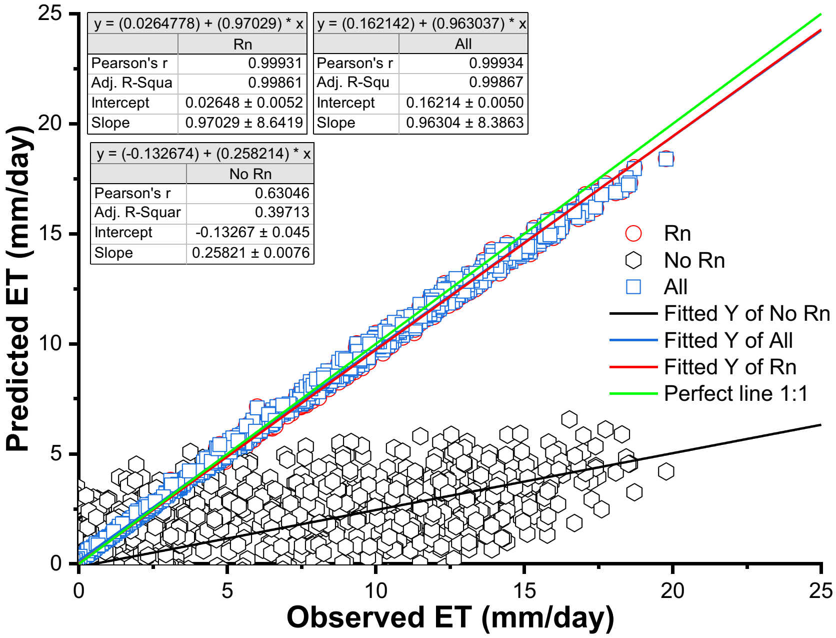


Table 4 Performance of Penalized regression models to estimate ET values

Input/Parameters	Stage	R^2	RMSE	nRMSE	NSME	MBE
Ridge regression						
T, RH, WS, SP	Calibration	0.4	3.81	77.2%	0.4	0
	Validation	0.27	4.54	85.6%	0.27	-1.43
Rn, T, RH, WS, SP	Calibration	0.98	0.65	13.2%	0.98	0
	Validation	0.98	0.78	14.7%	0.98	-0.15
LASSO						
T, RH, WS, SP	Calibration	0.41	3.8	77.1%	0.41	0
	Validation	0.4	4.84	91.4%	0.6	-2.27
Rn, T, RH, WS, SP	Calibration	0.99	0.41	8.3%	0.99	0
	Validation	0.99	0.78	5.4%	1.00	-0.02
ELNET						
T, RH, WS, SP	Calibration	0.41	3.8	77.1%	0.41	0
	Validation	0.4	4.84	91.3%	0.17	-2.26
Rn, T, RH, WS, SP	Calibration	0.99	0.41	8.3%	0.99	-0.001
	Validation	0.99	0.29	5.6%	1.00	0.001

level of model performance. However, the nRMSE values indicated good performance (< 20%) for ridge regression, while LASSO and ELNET exhibited excellent performance (< 10%). During the validation stages, the MBE values were -0.15, -0.02, and -0.001 for ridge, LASSO, and ELNET, respectively, suggesting a minor underestimation. These findings align with the research of Zhou et al. (2020), which supports for the incorporation of Rn in combination with other weather variables to achieve superior results in arid and semi-arid regions. The scatter plot diagram illustrating all penalized regression models is presented in Fig. 5.

4.4 Evaluation of machine learning models performance

In the first and third scenarios, all machine learning models exhibited exceptional performance, boasting R^2 values exceeding 0.95, except for the Support Vector Regression (SVR) model, which demonstrated subpar performance ($R^2 = 0.53$) during the validation stage in the third scenario, possibly resulted due to data overfitting (Table 5). The additional statistical metrics of Artificial Neural Network (ANN), K-Nearest Neighbours (KNN), and Random Forest (RF) models, including nRMSE (< 10%) and NSME

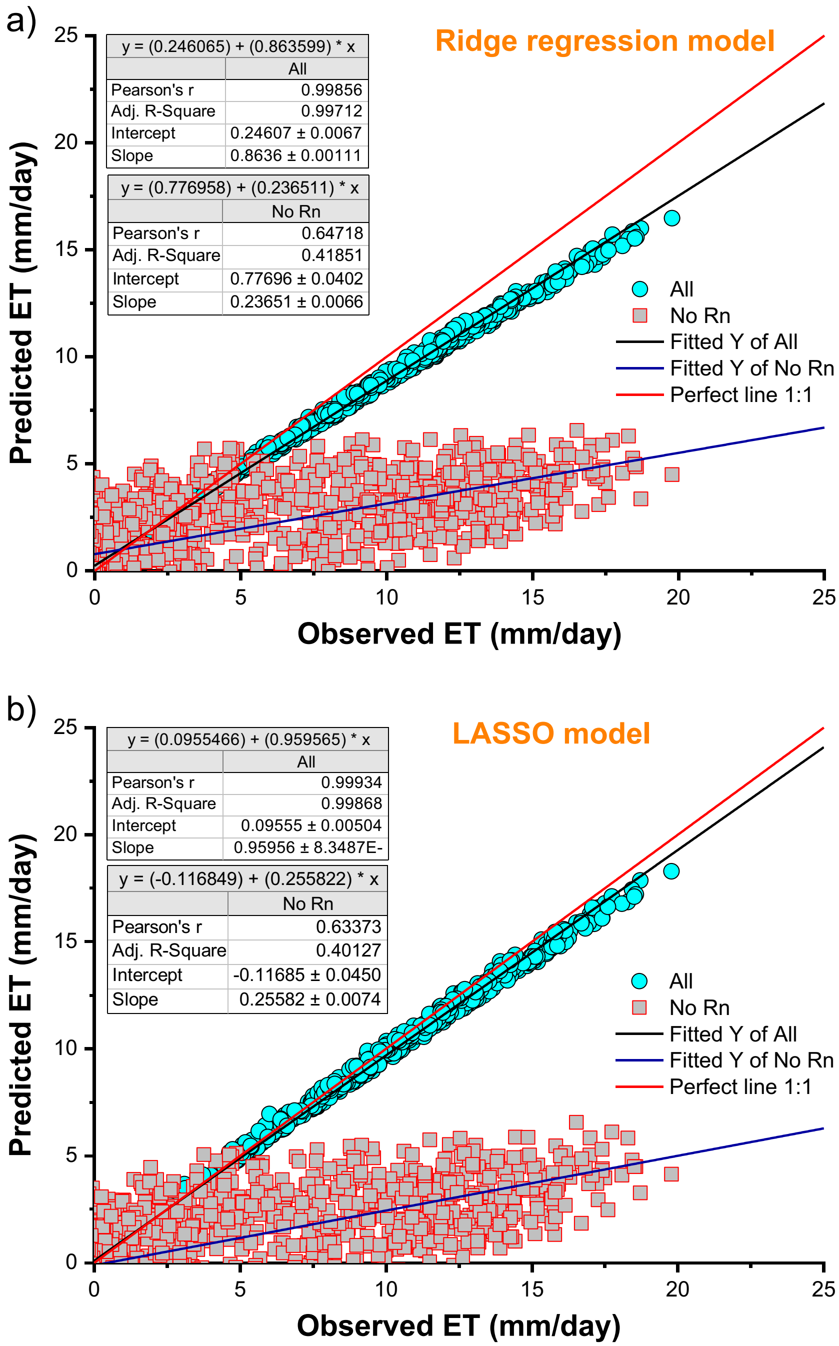


Fig. 5 Scatter plot between observed ET and predicted ET of penalized (**a**) Ridge regression (**b**) LASSO and (**c**) ELNET regression models

(> 0.98), consistently indicated outstanding model performance. With the inclusion of Rn as an input variable, any machine learning model, except SVR, proved suitable for accurate ET estimation across the study region. Comparable findings were reported by Üneş et al. (2020) when Rn was employed as the sole input variable for Support Vector Machines (SVM), resulting in more accurate ET estimates.

In the second scenario, where input data for weather variables was limited, diverse outcomes were observed across different models. The Artificial Neural Network (ANN) model yielded moderate results, with R^2 values of 0.58 and 0.61 during calibration and validation, respectively. However, the nRMSE values of 64.7% and 62.3% during calibration and validation stages indicated suboptimal model performance. Correspondingly, NSME values (0.58 and 0.61) suggested a fair level of model performance (Khaniya et al. 2020; Karunanayake et al. 2020; Ekanayake et al. 2021; Heddam et al. 2024). The Support Vector Regression (SVR) model displayed poor performance during the validation stage, registering an R^2 value of 0. K-Nearest Neighbours (KNN) exhibited commendable performance during calibration, with an R^2 value of 0.72, nRMSE of 10.2%, and NSME

Fig. 5 (continued)

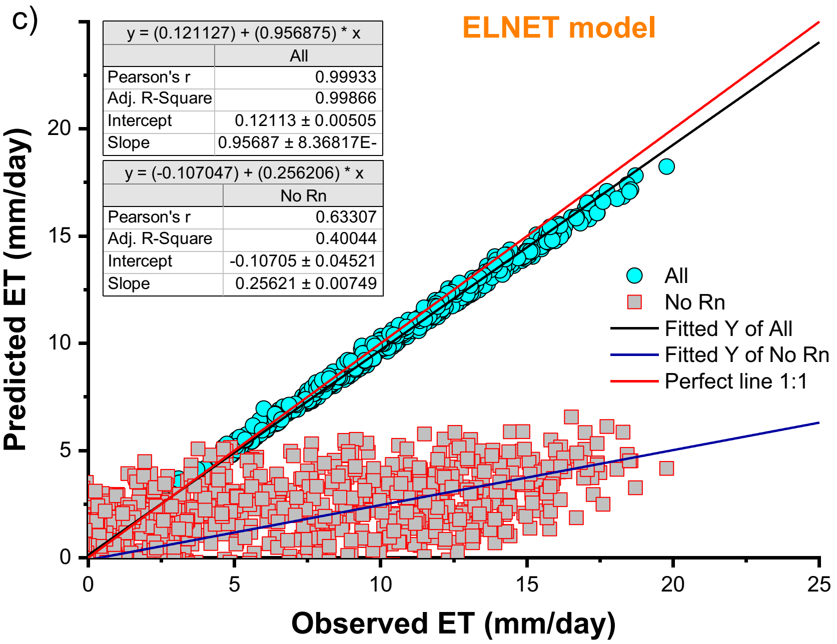


Table 5 Performance of machine learning models to estimate ET values

Input/Parameters	Stage	R^2	RMSE	nRMSE	NSME	MBE
ANN						
Rn	Calibration	0.99	0.37	7.4%	0.99	0.011
	Validation	0.99	0.22	4.2%	1.00	-0.045
T, RH, WS, SP	Calibration	0.58	3.19	64.7%	0.58	-0.018
	Validation	0.61	3.3	62.3%	0.61	0.039
Rn, T, RH, WS, SP	Calibration	0.99	0.05	0.9%	1.00	-0.001
	Validation	0.99	0.04	0.8%	1.00	0.001
SVR						
Rn	Calibration	0.99	0.39	7.9%	0.99	-0.095
	Validation	0.99	0.27	5.1%	1.00	-0.164
T, RH, WS, SP	Calibration	0.67	2.83	57.5%	0.67	-0.27
	Validation	0.00	5.44	102.7%	-0.05	0.13
Rn, T, RH, WS, SP	Calibration	0.99	0.23	4.7%	1.00	0.006
	Validation	0.53	4.31	81.3%	0.69	1.59
KNN						
Rn	Calibration	0.99	0.35	7%	1.00	-0.001
	Validation	0.99	0.27	5.2%	1.00	-0.058
T, RH, WS, SP	Calibration	0.72	0.5	10.2%	0.99	-0.001
	Validation	0.35	4.48	84.5%	0.29	-1.36
Rn, T, RH, WS, SP	Calibration	0.99	0.32	6.4%	1.00	-0.019
	Validation	0.95	1.43	26.9%	0.98	-0.46
Random Forest						
Rn	Calibration	0.99	0.23	4.6%	1.00	0
	Validation	0.99	0.33	6.2%	1.00	-0.053
T, RH, WS, SP	Calibration	0.99	0.5	10.2%	0.99	-0.001
	Validation	0.84	2.56	48.3%	0.77	-0.66
Rn, T, RH, WS, SP	Calibration	0.99	0.09	1.9%	1.00	-0.001
	Validation	0.99	0.8	15.1%	0.98	-0.224

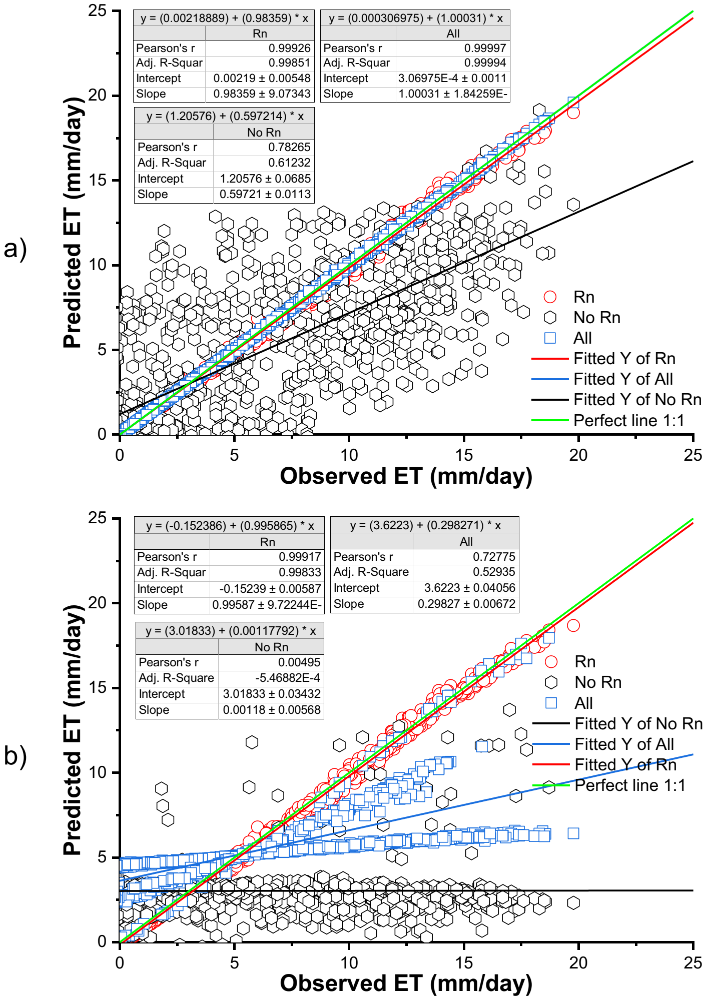


Fig. 6 Scatter plot between observed ET and predicted ET of machine learning: (**a**) ANN, (**b**) SVR, (**c**) KNNand (**d**) RF model

of 0.99. However, its performance diminished significantly during validation, with R^2 and nRMSE values of 0.35 and 84.5%, respectively. Among all the machine learning models with limited input weather variables, the Random Forest (RF) model demonstrated the best performance. Its R^2, nRMSE, and NSME values during calibration were 0.99, 10.2%, and 0.99, and during validation were 0.84, 48.3%, and 0.77, respectively. Similar results were reported by Shiri et al. (2014). The scatter plot diagram depicting all machine learning models is illustrated in Fig. 6.

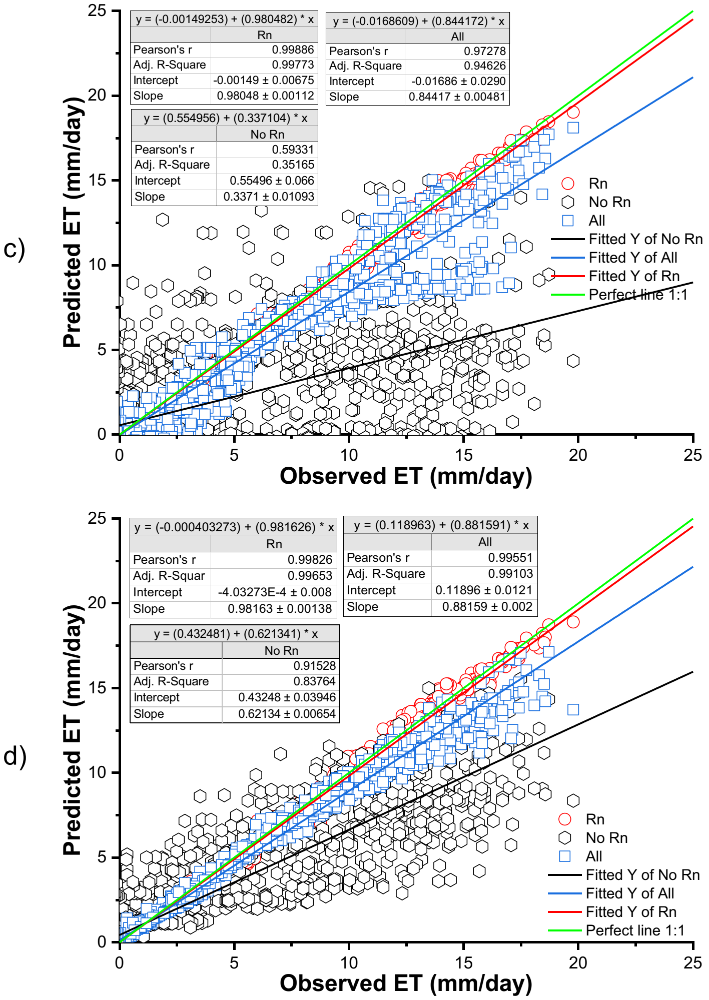
Figure 7(a-b) show the Taylor diagram for observed and predicted ET at Pantnagar station. Taylor diagram represents the correlation coefficient (r), root mean square deviation (RMSD), and standard deviation (SD) (Markuna et al. 2023; Vishwakarma et al. 2023). In penalized regression, it is clear from Fig. 7a that the LASSO and ELNET exhibited excellent performance model has the highest correlation values while the lowest RMSD and SD values for both all input and without Rn parameter. While Fig. 7b show that ANN model exhibited excellent performance model has the highest correlation values while the lowest RMSD and SD values for both all input and without Rn parameter and Random Forest model show better performance in only Rn input.

Predicting future ET (evapotranspiration) under different climatic scenarios involves considering the potential impacts of climate change on key meteorological variables. Given the correlations identified in the analysis, the following considerations can be made regarding ET predictions for future climatic scenarios:

Temperature changes If future climate scenarios involve temperature increases, it is likely to influence ET positively, as there is a positive correlation between temperature and ET. Elevated temperatures generally enhance the rate of evaporation and transpiration.

Net radiation Changes in net radiation may also play a role in influencing ET. The identified positive correlation between Net radiation and Temperature suggests that alterations in net radiation could impact ET in tandem with temperature changes.

Fig. 6 (continued)



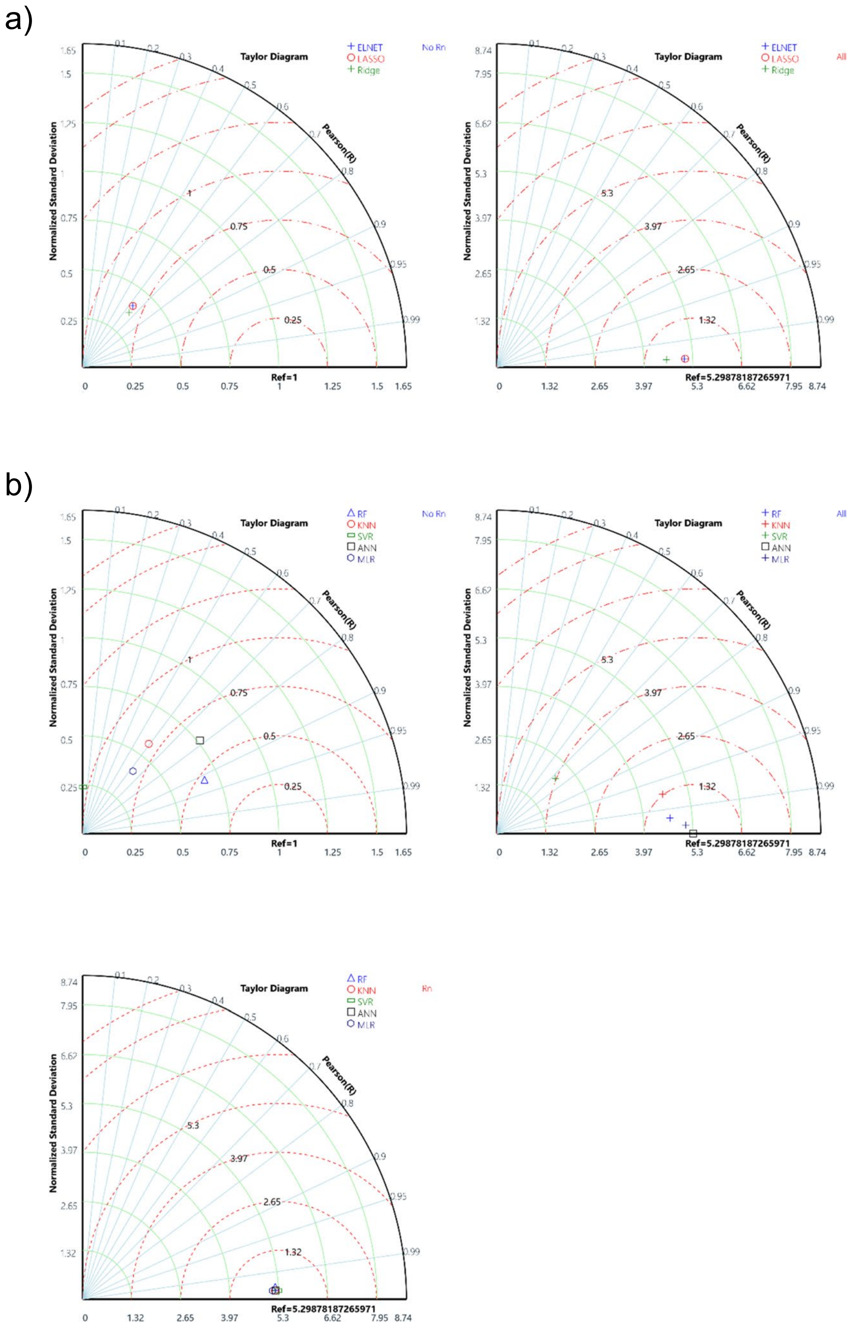
Relative humidity Future scenarios with decreased relative humidity might further contribute to increased ET, given the negative correlation observed. Lower humidity implies a drier atmosphere, potentially facilitating higher rates of evaporation.

Wind speed If future climates bring changes in wind speed, this could impact ET as well. The positive correlation with wind speed indicates that higher wind speeds might enhance ET.

5 Conclusion

Our comprehensive investigation into the relationships between various meteorological variables and evapotranspiration (ET) revealed robust correlations, highlighting the pivotal role of net radiation (Rn) in ET prediction. The multivariate linear regression (MLR) model excelled when solely utilizing Rn as an input, showcasing exceptional performance supported by key statistical parameters. However, limitations emerged when excluding Rn, emphasizing the need for a comprehensive set of input data. The incorporation of Rn enhanced the performance of penalized and machine learning models, demonstrating the importance of this variable. Penalized regression models, including ridge regression, LASSO, and ELNET, demonstrated improved performance with the inclusion of Rn, aligning with the findings of previous studies. In scenarios with limited input data, machine learning models showed varying performance, with the Random Forest (RF) model emerging as the most robust. However, caution is warranted when employing certain models, such as the Support Vector Regression (SVR), in limited-input scenarios due to potential performance issues. In summary, our study contributes valuable insights into the complex dynamics of ET estimation, emphasizing the importance of considering specific meteorological variables and the suitability of different modelling approaches based on data availability. The findings offer practical guidance for researchers and practitioners

Fig. 7 Taylor Diagram of (**a**) penalized regression and (**b**) MLR and machine learning models



working in regions with varying data constraints, facilitating more accurate and reliable ET predictions.

Acknowledgements The authors extend their appreciation to the Deanship of Scientific Research, King Saud University for funding through the Vice Deanship of Scientific Research Chairs; Research Chair of Prince Sultan Bin Abdulaziz International Prize for Water. Authors also thank the Department of Agrometeorology, G.B. Pant University of Agriculture and Technology for providing the required facilities to conduct the study.

Author contribution Anurag Satpathi and Abhishek Danodia conducted the formal analysis, investigation and wrote the manuscript. Ajeet Singh Nain edited and supervised the work. Makrand Dhyani, Dinesh Kumar Vishwakarma, Ahmed Z. Dewidar, and Mohamed A. Mattar prepared the necessary figures and revised the manuscript. All authors have reviewed the results and approved the final version of the manuscript.

Funding statement This research was funded by the Deanship of Scientific Research, King Saud University through the Vice Deanship of Scientific Research Chairs; Research Chair of Prince Sultan Bin Abdulaziz International Prize for Water.

Data availability The datasets analysed during the current study are available from the corresponding author on reasonable request.

Code availability The code are available from the corresponding author on reasonable request.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Abbas F, Afzaal H, Farooque AA, Tang S (2020) Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. Agronomy 10:1046. https://doi.org/10.3390/agronomy10071046

Abyaneh HZ, Nia AM, Varkeshi MB et al (2011) Performance Evaluation of ANN and ANFIS Models for Estimating Garlic Crop Evapotranspiration. J Irrig Drain Eng 137:280–286. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000298

Achite M, Elshaboury N, Jehanzaib M et al (2023) Performance of Machine Learning Techniques for Meteorological Drought Forecasting in the Wadi Mina Basin, Algeria. Water 15:765. https://doi.org/10.3390/w15040765

Aghajanloo M-B, Sabziparvar A-A, Hosseinzadeh Talaee P (2013) Artificial neural network–genetic algorithm for estimation of crop evapotranspiration in a semi-arid region of Iran. Neural Comput Appl 23:1387–1393. https://doi.org/10.1007/s00521-012-1087-y

Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. FAO - Food and Agriculture Organization of the United Nations Rome 300(9):D05109

Anapalli SS, Ahuja LR, Gowda PH et al (2016) Simulation of crop evapotranspiration and crop coefficients with data in weighing lysimeters. Agric Water Manag 177:274–283. https://doi.org/10.1016/j.agwat.2016.08.009

Azzam A, Zhang W, Akhtar F et al (2022) Estimation of green and blue water evapotranspiration using machine learning algorithms with limited meteorological data: A case study in Amu Darya River Basin, Central Asia. Comput Electron Agric 202:107403. https://doi.org/10.1016/j.compag.2022.107403

Bachour R, Walker WR, Ticlavilca AM et al (2014) Estimation of Spatially Distributed Evapotranspiration Using Remote Sensing and a Relevance Vector Machine. J Irrig Drain Eng 140:4014029. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000754

Biau G, Devroye L, Lugosi G (2008) Consistency of random forests and other averaging classifiers. J Mach Learn Res 9(9):2015–2033

Biau G, Scornet E (2016) A random forest guided tour. TEST 25:197–227. https://doi.org/10.1007/s11749-016-0481-7

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Buttar NA, Yongguang H, Shabbir A et al (2018) Estimation of evapotranspiration using Bowen ratio method. IFAC-PapersOnLine 51:807–810. https://doi.org/10.1016/j.ifacol.2018.08.096

Chen Z, Sun S, Wang Y et al (2020) Temporal convolution-network-based models for modeling maize evapotranspiration under mulched drip irrigation. Comput Electron Agric 169:105206. https://doi.org/10.1016/j.compag.2019.105206

Chia MY, Huang YF, Koo CH et al (2022) Long-term forecasting of monthly mean reference evapotranspiration using deep neural network: A comparison of training strategies and approaches. Appl Soft Comput 126:109221. https://doi.org/10.1016/j.asoc.2022.109221

Das B, Nair B, Reddy VK, Venkatesh P (2018) Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. Int J Biometeorol 62:1809–1822. https://doi.org/10.1007/s00484-018-1583-6

de Teixeira AHC, Bastiaanssen WGM, Ahmad MD, Bos MG (2009) Reviewing SEBAL input parameters for assessing evapotranspiration and water productivity for the Low-Middle São Francisco River basin, Brazil. Agric For Meteorol 149:462–476. https://doi.org/10.1016/j.agrformet.2008.09.016

Ekanayake P, Wickramasinghe L, Jayasinghe JMJW, Rathnayake U (2021) Regression-Based Prediction of Power Generation at Samanalawewa Hydropower Plant in Sri Lanka Using Machine Learning. Math Probl Eng 2021:1–12. https://doi.org/10.1155/2021/4913824

Elbeltagi A, Kushwaha NL, Rajput J et al (2022) Modelling daily reference evapotranspiration based on stacking hybridization of ANN with meta-heuristic algorithms under diverse agro-climatic conditions. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-022-02196-0

Elbeltagi A, Al-Mukhtar M, Kushwaha NL et al (2023a) Forecasting monthly pan evaporation using hybrid additive regression and data-driven models in a semi-arid environment. Appl Water Sci 13:42. https://doi.org/10.1007/s13201-022-01846-6

Elbeltagi A, Seifi A, Ehteram M et al (2023b) GLUE analysis of meteorological-based crop coefficient predictions to derive the explicit equation. Neural Comput Appl. https://doi.org/10.1007/s00521-023-08466-4

Fan J, Yue W, Wu L et al (2018) Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. Agric for Meteorol 263:225–241. https://doi.org/10.1016/j.agrformet.2018.08.019

Feng Y, Gong D, Mei X, Cui N (2017) Estimation of maize evapotranspiration using extreme learning machine and generalized regression neural network on the China Loess Plateau. Hydrol Res 48:1156–1168. https://doi.org/10.2166/nh.2016.099

Ghiassi M, Saidane H, Zimbra DK (2005) A dynamic artificial neural network model for forecasting time series events. Int J Forecast 21:341–362. https://doi.org/10.1016/j.ijforecast.2004.10.008

Heddam S, Vishwakarma DK, Abed SA et al (2024) Hybrid river stage forecasting based on machine learning with empirical mode decomposition. Appl Water Sci 14:46. https://doi.org/10.1007/s13201-024-02103-8

Hilt DE, Seegrist DW (1977) Ridge, a computer program for calculating ridge regression estimates. USDA Forest Service Research Note NE-236, United States, Department of Agriculture, Forest Service, Northeastern Forest Experiment Station

Jiang X, Kang S, Tong L, Li F (2016) Modification of evapotranspiration model based on effective resistance to estimate evapotranspiration of maize for seed production in an arid region of northwest China. J Hydrol 538:194–207. https://doi.org/10.1016/j.jhydrol.2016.04.002

Karunanayake C, Gunathilake MB, Rathnayake U (2020) Inflow Forecast of Iranamadu Reservoir, Sri Lanka, under Projected Climate Scenarios Using Artificial Neural Networks. Appl Comput Intell Soft Comput 2020:1–11. https://doi.org/10.1155/2020/8821627

Kato S, Yamaguchi Y (2007) Estimation of storage heat flux in an urban area using ASTER data. Remote Sens Environ 110:1–17. https://doi.org/10.1016/j.rse.2007.02.011

Kaur R, Sharma S (2019) An ANN Based Approach for Software Fault Prediction Using Object Oriented Metrics. In: Luhach AK, Singh D, Hsiung P-A, et al. (eds) Advanced Informatics for Computing

Research, ICAICR 2018. Communications in Computer and Information Science, vol 955. Springer Singapore, pp 341–354

Khaniya B, Karunanayake C, Gunathilake MB, Rathnayake U (2020) Projection of Future Hydropower Generation in Samanalawewa Power Plant, Sri Lanka. Math Probl Eng 2020:1–11. https://doi.org/10.1155/2020/8862067

Kramer O (2013) Dimensionality Reduction with Unsupervised Nearest Neighbors. Springer Berlin Heidelberg, Berlin

Kumar R, Lone MA, Bhat OA (2021) Determination of water requirement and crop coefficients for green gram in temperate region using lysimeter water balance. Int J Hydrol Sci Technol 12:1. https://doi.org/10.1504/ijhst.2021.10038778

Kumar A, Singh VK, Saran B et al (2022a) Development of Novel Hybrid Models for Prediction of Drought- and Stress-Tolerance Indices in Teosinte Introgressed Maize Lines Using Artificial Intelligence Techniques. Sustainability 14:2287. https://doi.org/10.3390/su14042287

Kumar R, Manzoor S, Vishwakarma DK et al (2022b) Assessment of Climate Change Impact on Snowmelt Runoff in Himalayan Region. Sustainability 14:1–23. https://doi.org/10.3390/su14031150

Kushwaha NL, Rajput J, Elbeltagi A et al (2021) Data Intelligence Model and Meta-Heuristic Algorithms-Based Pan Evaporation Modelling in Two Different Agro-Climatic Zones: A Case Study from Northern India. Atmosphere (basel) 12:1654. https://doi.org/10.3390/atmos12121654

Li Y-F, Xie M, Goh T-N (2010) Adaptive ridge regression system for software cost estimating on multi-collinear datasets. J Syst Softw 83:2332–2343. https://doi.org/10.1016/j.jss.2010.07.032

Liu Y, Wang Y, Zhang J (2012) New Machine Learning Algorithm: Random Forest. In: Liu B, Ma M, Chang J (eds) Information Computing and Applications. ICICA 2012. Lecture Notes in Computer Science, vol 7473. Springer Berlin Heidelberg, Berlin, pp 246–252

Malek E, Bingham GE (1993) Comparison of the Bowen ratio-energy balance and the water balance methods for the measurement of evapotranspiration. J Hydrol 146:209–220. https://doi.org/10.1016/0022-1694(93)90276-F

Markuna S, Kumar P, Ali R et al (2023) Application of Innovative Machine Learning Techniques for Long-Term Rainfall Prediction. Pure Appl Geophys 180:335–363. https://doi.org/10.1007/s00024-022-03189-4

Mirzania E, Vishwakarma DK, Bui Q-AT et al (2023) A novel hybrid AIG-SVR model for estimating daily reference evapotranspiration. Arab J Geosci 16:301. https://doi.org/10.1007/s12517-023-11387-0

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I — A discussion of principles. J Hydrol 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Pavlou M, Ambler G, Seaman S et al (2016) Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Stat Med 35:1159–1177. https://doi.org/10.1002/sim.6782

Peacock CE, Hess TM (2004) Estimating evapotranspiration from a reed bed using the Bowen ratio energy balance method. Hydrol Process 18:247–260. https://doi.org/10.1002/hyp.1373

Rana G, Katerji N (2000) Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: a review. Eur J Agron 13:125–153. https://doi.org/10.1016/S1161-0301(00)00070-8

Raza A, Al-Ansari N, Hu Y et al (2022) Misconceptions of Reference and Potential Evapotranspiration: A PRISMA-Guided Comprehensive Review. Hydrology 9:153. https://doi.org/10.3390/hydrology9090153

Sagar A, Hasan M, Singh DK et al (2022) Development of Smart Weighing Lysimeter for Measuring Evapotranspiration and Developing Crop Coefficient for Greenhouse Chrysanthemum. Sensors 22:6239. https://doi.org/10.3390/s22166239

Saggi MK, Jain S (2020) Application of fuzzy-genetic and regularization random forest (FG-RRF): Estimation of crop evapotranspiration (ET) for maize and wheat crops. Agric Water Manag 229:105907. https://doi.org/10.1016/j.agwat.2019.105907

Saroughi M, Mirzania E, Vishwakarma DK et al (2023) A Novel Hybrid Algorithms for Groundwater Level Prediction. Iran J Sci Technol Trans Civ Eng. https://doi.org/10.1007/s40996-023-01068-z

Sattari MT, Apaydin H, Band SS et al (2021) Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. Hydrol Earth Syst Sci 25:603–618. https://doi.org/10.5194/hess-25-603-2021

Shiri J, Nazemi AH, Sadraddini AA et al (2014) Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Comput Electron Agric 108:230–241. https://doi.org/10.1016/j.compag.2014.08.007

Shukla R, Kumar P, Vishwakarma DK et al (2021) Modeling of stage-discharge using back propagation ANN-, ANFIS-, and WANN-based computing techniques. Theor Appl Climatol. https://doi.org/10.1007/s00704-021-03863-y

Singh RS, Patel C, Yadav MK, Singh KK (2014) Yield forecasting of rice and wheat crops for eastern Uttar Pradesh. J Agrometeorol 16:199–202. https://doi.org/10.54386/jam.v16i2.1521

Singh AK, Kumar P, Ali R et al (2022a) An Integrated Statistical-Machine Learning Approach for Runoff Prediction. Sustainability 14:8209. https://doi.org/10.3390/su14138209

Singh VK, Panda KC, Sagar A et al (2022b) Novel Genetic Algorithm (GA) based hybrid machine learning-pedotransfer Function (ML-PTF) for prediction of spatial pattern of saturated hydraulic conductivity. Eng Appl Comput Fluid Mech 16:1082–1099. https://doi.org/10.1080/19942060.2022.2071994

Tabari H, Martinez C, Ezani A, Hosseinzadeh Talaee P (2013) Applicability of support vector machines and adaptive neurofuzzy inference system for modeling potato crop evapotranspiration. Irrig Sci 31:575–588. https://doi.org/10.1007/s00271-012-0332-6

Thissen U, van Brakel R, de Weijer A et al (2003) Using support vector machines for time series prediction. Chemom Intell Lab Syst 69:35–49. https://doi.org/10.1016/S0169-7439(03)00111-4

Tulla PS, Kumar P, Vishwakarma DK et al (2024) Daily suspended sediment yield estimation using soft-computing algorithms for hilly watersheds in a data-scarce situation: a case study of Bino watershed, Uttarakhand. Theor Appl Climatol. https://doi.org/10.1007/s00704-024-04862-5

Üneş F, Kaya YZ, Mamak M (2020) Daily reference evapotranspiration prediction based on climatic conditions applying different data mining techniques and empirical equations. Theor Appl Climatol 141:763–773. https://doi.org/10.1007/s00704-020-03225-0

Vapnik V (1998) Statistical learning theory. John Wiley & Sons, Inc., Oxford

Verma SB, Rosenberg NJ, Blad BL (1978) Turbulent Exchange Coefficients for Sensible Heat and Water Vapor under Advective Conditions. J Appl Meteorol 17:330–338

Vishwakarma DK, Kumar R, Pandey K et al (2018) Modeling of Rainfall and Ground Water Fluctuation of Gonda District Uttar Pradesh, India. Int J Curr Microbiol Appl Sci 7:2613–2618. https://doi.org/10.20546/ijcmas.2018.705.302

Vishwakarma DK, Pandey K, Kaur A et al (2022) Methods to estimate evapotranspiration in humid and subtropical climate conditions. Agric Water Manag 261:107378. https://doi.org/10.1016/j.agwat.2021.107378

Vishwakarma DK, Kuriqi A, Abed SA et al (2023) Forecasting of stage-discharge in a non-perennial river using machine learning with gamma test. Heliyon 9:e16290. https://doi.org/10.1016/j.heliyon.2023.e16290

Vishwakarma DK, Kumar P, Yadav KK et al (2024) Evaluation of CatBoost Method for Predicting Weekly Pan Evaporation in Subtropical and Sub-Humid Regions. Pure Appl Geophys. https://doi.org/10.1007/s00024-023-03426-4

Wang S-C (2003) Artificial Neural Network. In: Wang S-C (ed) Interdisciplinary Computing in Java Programming. Springer US, Boston, pp 81–100

Willmott CJ (1981) On the validation of models. Phys Geogr 2:184–194. https://doi.org/10.1080/02723646.1981.10642213

Zhou Z, Zhao L, Lin A et al (2020) Exploring the potential of deep factorization machine and various gradient boosting models in modeling daily reference evapotranspiration in China. Arab J Geosci 13:1287. https://doi.org/10.1007/s12517-020-06293-8

Zou H, Hastie T (2005) Regularization and Variable Selection Via the Elastic Net. J R Stat Soc Ser B Stat Methodol 67:301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.