*J. Earth Syst. Sci.* (2023)132 129
https://doi.org/10.1007/s12040-023-02138-6



# Modelling monthly reference evapotranspiration estimation using machine learning approach in data-scarce North Western Himalaya region (Almora), Uttarakhand

Utkarsh Kumar[1,2]

[1]*ICAR-Vivekananda Parvatiya Krishi Anusandhan Sansthan, Almora 263 601, Uttarakhand, India.*
[2]*Division of Agricultural Engineering, ICAR-Indian Agricultural Research Institute, New Delhi 110 012, Delhi, India.*
*e-mail: utkarsh.kumar@icar.gov.in*

MS received 9 March 2023; revised 18 April 2023; accepted 2 May 2023

Reference evapotranspiration ($ET_0$) is a crucial parameter in hydrology that is used to estimate the amount of water lost through evaporation and transpiration from a standard reference crop under specified climatic conditions. $ET_0$ is calculated based on several meteorological variables such as temperature, humidity, wind speed, and solar radiation. However, in many regions, there may be limited availability of these meteorological data, making it difficult to estimate $ET_0$ using conventional methods. This study aimed at using machine learning (ML) techniques to estimate $ET_0$ with minimal climatic inputs, using the Food and Agriculture Organization (FAO)-56 Penman–Monteith model as the standard reference method. Different ML models, including Long Short-Term Memory (LSTM) neural networks, Gradient Boosting Regressor (GBR), Random Forest (RF), and Support Vector Regression (SVR), were developed with climatic variables as input parameters. To summarize, the study evaluated different machine learning models to estimate $ET_0$ with minimal climatic inputs and compared their performance with the standard Penman–Monteith model for Hawalbagh experimental farm observatory. The results showed that the LSTM model performed better than other ML models, followed by SVR and RF, in estimating $ET_0$ with minimal climate data. The study concluded that LSTM and SVR models are the most robust ML models for estimating $ET_0$ in such scenarios. The study found that even with limited climatic data, such as a two-parameter combination (maximum temperature and relative humidity, relative humidity and sunshine hours) or a three-parameter combination (minimum temperature, relative humidity, and precipitation, minimum temperature, maximum temperature, and precipitation), promising results in $ET_0$ estimation can be achieved. The study's findings are significant for estimating $ET_0$ in data-scarce regions, especially for agricultural water management in semi-arid climates.

**Keywords.** Evapotranspiration; gradient boosting regressor; long short-term memory; Penman–Monteith; random forest; support vector regression.

Published online: 07 August 2023

## 1. Introduction

The accurate estimation of crop water needs is crucial for sustainable management of water resources, particularly in regions affected by climate change such as India. Crop evapotranspiration ($ET_{crop}$) is an important parameter that can help in estimating a field's net irrigation requirement, which varies depending on the crop being grown (Srivastava *et al.* 2017; Kumar *et al.* 2020; Maza *et al.* 2020). Evapotranspiration ($ET_0$) is the process of water movement from soil and vegetation to the atmosphere, which occurs through the combination of plant transpiration and soil and atmospheric evaporation. It is a vital component in agriculture, land management, pollution detection, irrigation planning, hydrological balance, and watershed hydrology (Kumari and Srivastava 2020). Accurate estimation of $ET_0$ is crucial in managing water resources for various purposes, including irrigation, drinking, industrial use, and water reserve management (Kumar *et al.* 2021a, b; 2023a, b). The precise calculation of $ET_0$ can improve irrigation efficiency, water reuse, and seepage control, making it an essential factor in addressing water management problems. The quantity of water needed to irrigate crop fields for the entire period is called crop-water requirement. It can be computed by multiplying crop coefficient with $ET_0$ (Srivastava *et al.* 2018). Accurate irrigation scheduling is impossible with prior knowledge of $ET_0$. There are different methods developed for the calculation of $ET_0$ from meteorological variables, which are temperature-based, radiation-based, and combination-based. The estimation of $ET_0$ through Lysimeter is considered to be the most accurate, but it is labour-intensive, expensive and requires proper maintenance. The performance of different methods varies in different agro-climatic regions. The FAO-56 PM has been recommended worldwide for the calculation of $ET_0$ and used as a reference approach (Tabari *et al.* 2013; Kovoor *et al.* 2018; Valipour *et al.* 2018; Fan *et al.* 2019; Kumar *et al.* 2022). FAO-56 PM method requires large amount of data for $ET_0$ calculation (Shiri *et al.* 2012; Caminha *et al.* 2017; Feng *et al.* 2017; Trigo *et al.* 2018; Kumar *et al.* 2021a). Due to the paucity of weather data in developing countries like India, limited meteorological data are available for the calculation of $ET_0$. In recent years, much attention has been given towards forecasting natural events (Khosravi *et al.* 2018; Yaseen *et al.* 2018; Naganna *et al.* 2019; Xiao *et al.* 2019). Therefore, there is a need for the calculation of evapotranspiration using machine learning algorithm. Several researchers across the globe in different agro-climatic regions have attempted to compute $ET_0$ from different meteorological variables and majority of the method need data that are not readily available, particularly in the hilly region of Uttarakhand (Kumar *et al.* 2023a). In addition to this, some of these methods are applicable to certain specific climatic conditions and they cannot be used under conditions which are different from those they were originally developed. Simple model is basically based on the linear relationship between different meteorological variables. Due to rapid urbanization and industrialization, pollution can affect $ET$ in several ways. For example, air pollution can reduce the amount of sunlight that reaches the land surface, which can reduce photosynthesis and therefore transpiration (Yao 2017). Several studies on assessing the impact of pollution on environmental conditions is conducted in recent years (Ambade *et al.* 2021; Gautam *et al.* 2021, 2023; Nepolian *et al.* 2021; Thapliyal *et al.* 2022). $ET$ is a complex phenomenon and it is highly non-linear in nature. Various researches showed that machine learning algorithms have shown promising results in predicting $ET_0$ (Partal 2009; Cobaner 2013; Falamarzi *et al.* 2014; Adamala *et al.* 2019; Kelley and Pardyjak 2019). Machine learning techniques are efficient in obtaining complicated relationships between input and output variables, which makes them robust tools for $ET_0$ modelling (Abdullah *et al.* 2015; Marti *et al.* 2015; Wang *et al.* 2018; Ferreira *et al.* 2019; Nourani *et al.* 2019; Wu and Fan 2019). In the last decade, various researches have been explored to demonstrate the proficiency of various machine learning model for predicting $ET_0$ (Chauhan and Shrivastava 2008; Zanetti *et al.* 2007; Khoob 2008; Trajkovic 2010; Feng *et al.* 2017). Based on review of the literature, MLP algorithm is the best applied and most efficient neural network to predict $ET_0$. Therefore the complexity of the calculation of evapotranspiration can be solved using machine learning (ML) algorithm with limited meteorological variables. This research evaluates different ML algorithms for the estimation of $ET_0$ in different agro-climatic zones of India. Several studies have emphasized on the estimation of accurate $ET_0$ in different hydrological studies work. ANN and multivariate non-linear regression technique were applied in semi-arid areas of Iran for the estimation of $ET_0$. For

data-limited conditions, ML techniques have been extensively used as an alternative tool for the computation of $ET_0$. Many of the aforementioned studies are done in same agro-climatic zone, particularly in India. Therefore, it is necessary to evaluate different ML techniques in various agro-climatic zones for the calculation of $ET_0$. Numerous studies have been reported in the literature to test the applicability of ML techniques for the estimation of reference $ET_0$. There are various softwares used for calculation, viz., CROPWAT, Daily $ET$ and $ET_0$ calculator. These softwares calculate $ET_0$ based on empirical and semi-empirical equation which requires numerous meteorological and geographical parameter. These softwares do not yield any result if one parameter is missing. In most situations, meteorological variable is scarce and limited to very few numbers of climatic variables. Therefore, present and future estimation of $ET_0$ is confined. Recently there have been many attempts to compute and predict $ET_0$ with higher accuracy and at different time scales. Numerical and statistical techniques have been applied to mimic the random nature of meteorological variables. The difficulties associated with these methods forced scientists and researchers to look for alternative options such as, data-driven techniques and machine learning approach, viz., ANN (Kumar *et al.* 2011). Based on review of literature, it can be summarized that machine learning models are suitable for the computation and prediction of $ET_0$ in data-limited conditions. The study area was selected because of significant wastage of surface and groundwater due to ineffective irrigation planning by local farmers. This results in water scarcity during dry season, highlighting the need for effective irrigation plans. Accurate estimates of $ET$ are useful for this and important for transitioning to more integrated and adaptive water resources management. The current study gives a comparison of four ML-based models to discover the best model for assessing daily $ET_0$ under the state using minimal input variables in the sub-tropical atmospheres. The objectives of the current study are as per the following: (1) to develop different ML models, SVR, LSTM, RF and GBR for modelling $ET_0$ over Hawalbagh experimental farm weather observatory, (2) to assess the performance and stability of these models with different input combinations, and (3) to find an appropriate approach to boost the modelling performance under the limited input factors condition.

## 2. Study area and data used

To illustrate the idea presented in the paper, we selected ICAR-VPKAS, Hawalbagh experimental farm, observatory (29°36′N; 79°40′E at 1250 m above mean sea level) located in Kumaon division of Uttarakhand, Almora, India. The geographical location of the study area is shown in figure 1. Meteorological data (i.e., minimum and maximum temperature, minimum and maximum relative humidity, sunshine hours, and rainfall) were obtained from the meteorological observatory for the period, January 1985–December 2010, on a monthly basis, as given in table 1. The annual rainfall of the site varies from 693 to 1415 mm with an average rainfall of 1013 mm during the study period. Maximum temperature ($T_{\max}$) ranged from 24.7° to 26.7°C with average $T_{\max}$ equal to 25.93°C and minimum temperature ($T_{\min}$) ranged from 9.08° to 11.5°C with average $T_{\max}$ equal to 10.4°C (figure 2a–d).

## 3. Methodology

This section covers a detailed description of different machine learning models applied in the present study. The complete dataset was divided into two parts. The first part which is 80% of the whole dataset is used for training the selected ML model, while the remaining 20% second part is used to test the developed ML model. This study compared the performance of four machine learning (ML) models for modelling the $ET_0$ using meteorological data collected from the observatory. The ML models used were LSTM, GBR, SVR, and RF regressor. Models were prepared in Python using the Keras library (https://keras.io). The flowchart used in this study is shown in figure 3.

### 3.1 *Optimal input selection of input variable*

The optimal input combination was selected based on seven statistical criteria: MSE, RMSE, $R^2$, adjusted $R^2$, Mallows' Cp, Akaike's AIC, and Amemiya's PC. The combination with the smallest and closest to zero values of MSE, Mallows' Cp, Akaike's AIC, and Amemiya's PC, and the largest values of $R^2$ and adjusted $R^2$ were chosen as the optimal input combination for $ET_0$ modelling. This approach was applied at Hawalbagh experimental farm and five different input combinations were examined. The top five models were selected based on combined scores calculated from above-mentioned statistical

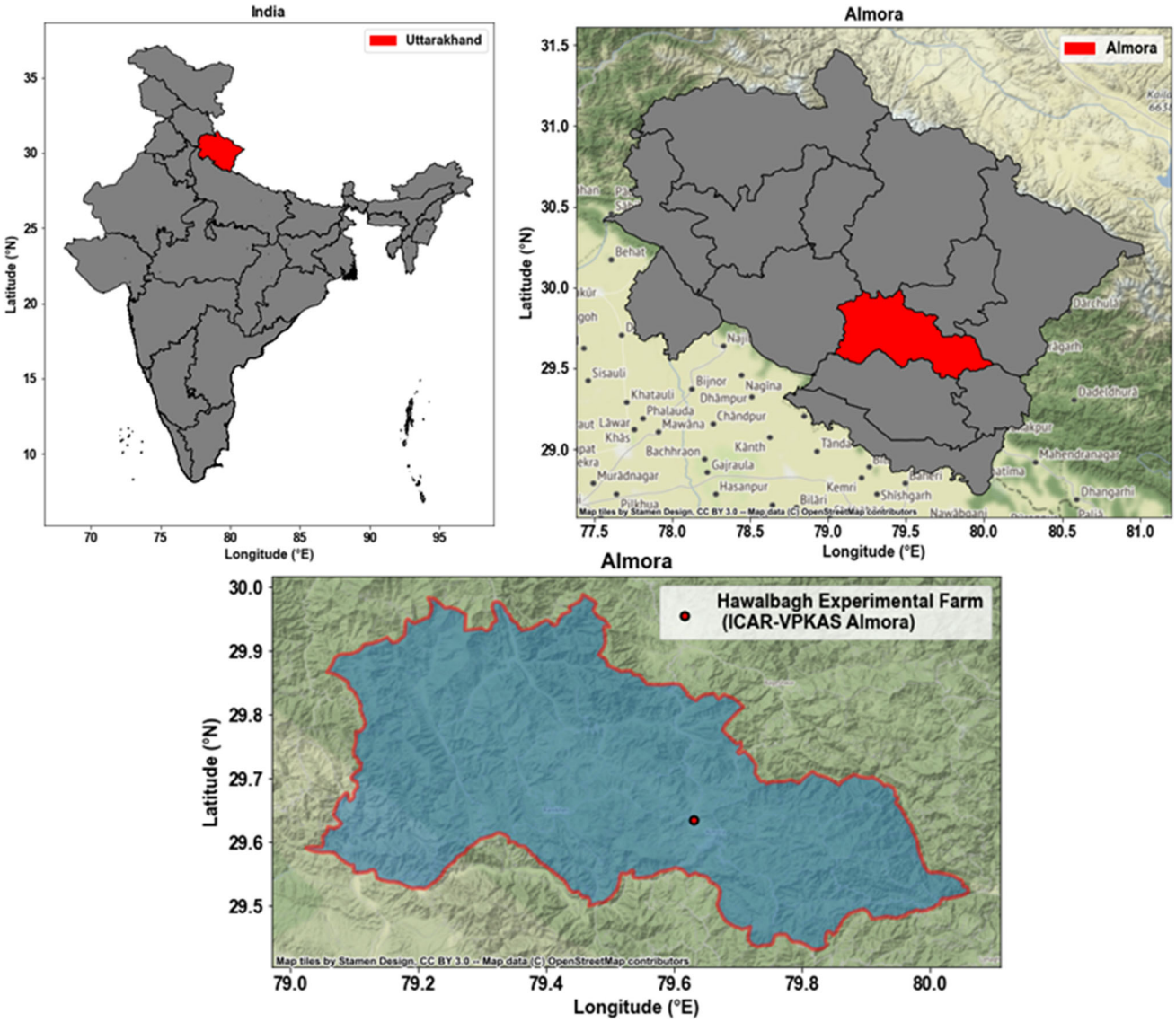


Figure 1. Location of the study area.

Table 1. *Monthly meteorological parameter dynamics for the period of 1985–2010.*

| Meteorological parameter | Mean ± SD | Minimum | Maximum | CV (%) |
|---|---|---|---|---|
| Precipitation (mm) | 84 ± 91 | 0 | 493 | 108.33 |
| Maximum temperature | 25.9 ± 4.53 | 15.5 | 33.9 | 17.49 |
| Minimum temperature | 10.43 ± 7.48 | –1.7 | 21.8 | 71.71 |
| Relative humidity | 67.8 ± 9.41 | 40.25 | 87.3 | 21.8 |
| Sunshine hour | 7.06 ± 1.53 | 2.96 | 10.17 | 0.21 |

parameters. The statistical analysis of these combinations is presented in table 2.

### 3.2 *Model development*

In order to build a reliable predictive model, selecting the appropriate predictors is a crucial step. In the current study, four different machine learning (ML) models were chosen to predict monthly evaporation: the classification and regression tree (CART), the cascade correlation neural network (CCNN), gene expression programming (GEP), and support vector machine (SVM). These models were developed using five different input combinations:

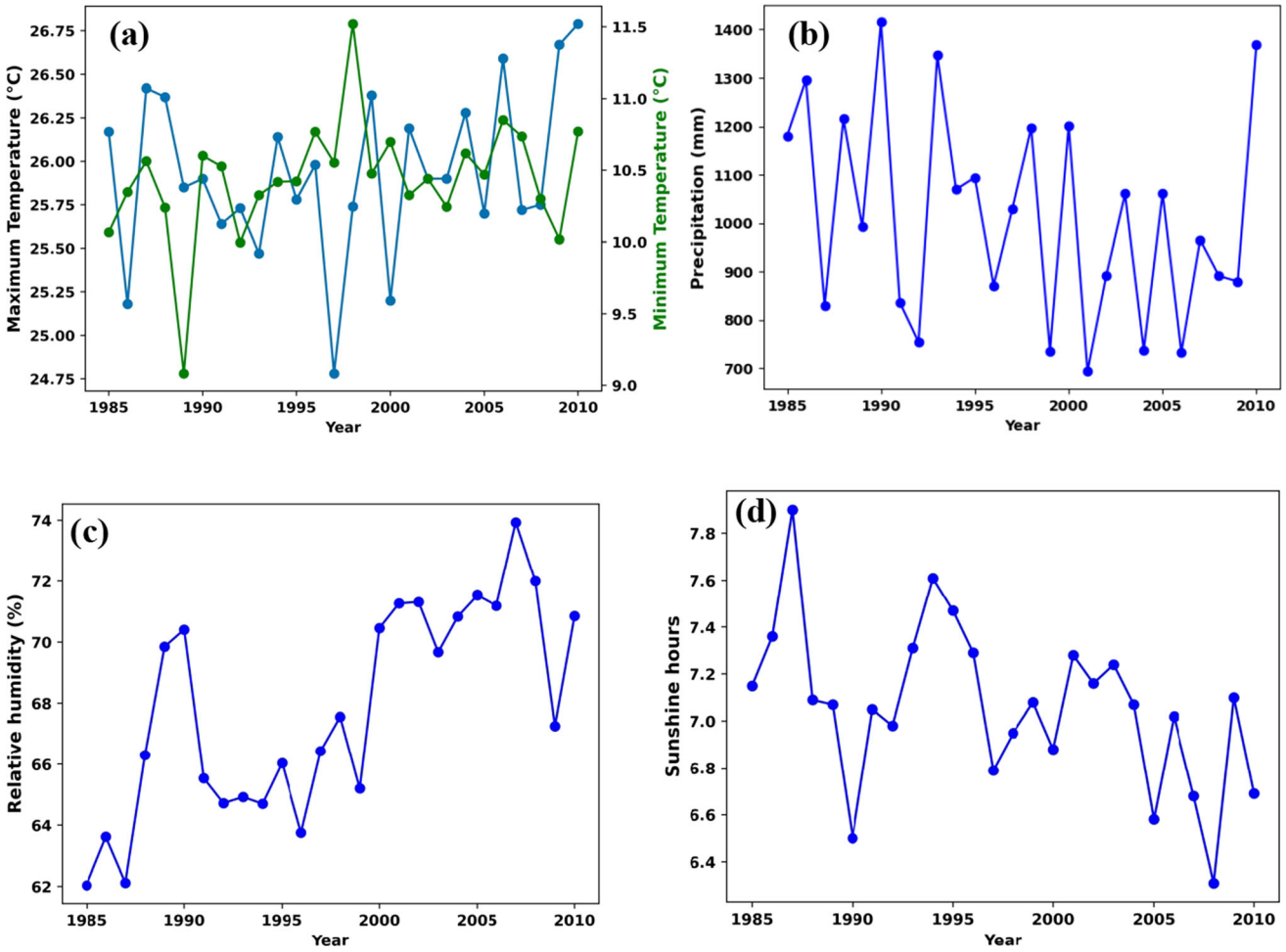


Figure 2. Monthly average climatic parameters for the North Bihar (1985–2011). (**a**) $T_{max}$ and $T_{min}$, **(b)** precipitation; **(c)** relative humidity; and **(d)** hours of sunshine.

$$\text{Model 1 (M1)}: ET_0 = f(T_{max}, \text{RH}_{mean}) \quad (1)$$

$$\text{Model 2 (M2)}: ET_0 = f(T_{min}, \text{RH}_{mean}, \text{RF}) \quad (2)$$

$$\text{Model 3 (M3)}: ET_0 = f(\text{SSh}, \text{RH}_{mean}) \quad (3)$$

$$\text{Model 4 (M4)}: ET_0 = f(T_{min}, \text{RF}, \text{SSh}) \quad (4)$$

$$\text{Model 5 (M5)}: ET_0 = f(T_{min}, T_{max}, \text{RF}) \quad (5)$$

where $ET_0$ is the reference evapotranspiration, WS is the wind speed, RF is the rainfall, RH is the relative humidity, $T_{min}$ is the minimum temperature, $T_{max}$ is the maximum temperature, and SSh is the number of sunshine hours.

### 3.3 *Calculation of reference ET*

There are different equations developed to estimate $ET_0$ across the globe in different agro-climatic zone and extensively evaluated using Lysimeter data. However, the standard and reliable method for computing $ET_0$ is FAO Penman–Monteith (Allen *et al.* 1998). In the present study, we have used FAO Penman–Monteith to calculate $ET_0$. It represents the response of weather parameters to environmental conditions. $K_c$ is different for different crops and represents the crop canopy development and crop management practices through the growing season. Daily meteorological data were collected for five variables, which included maximum air temperature ($T_{max}$ °C), minimum air temperature ($T_{min}$ °C), minimum relative humidity ($RH_{min}$, %), maximum relative humidity ($RH_{max}$, %), wind speed ($U_2$, m $s^{-1}$), and solar radiation ($R_s$, MJ $m^{-2}$ $d^{-1}$). Measurements were taken at a height of 2 m for air temperature and relative humidity, and at a height of 10 m for wind speed above the soil surface. Wind speed data at 2 m ($U_2$) were obtained using the log-wind profile equation from those collected at 10 m.

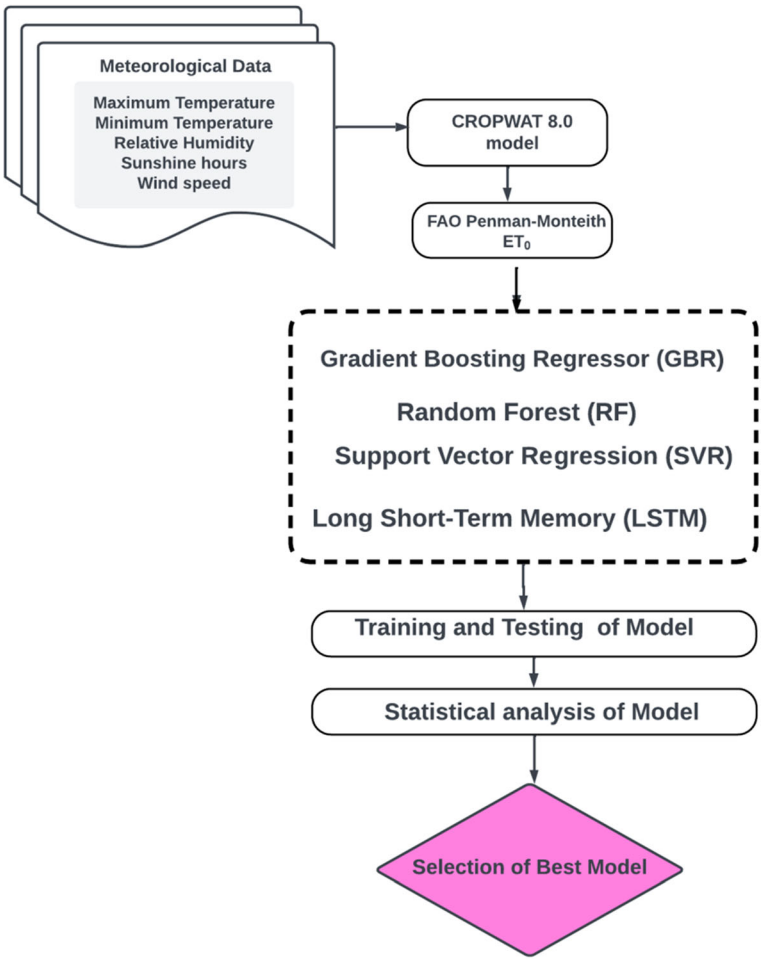


Figure 3. Flow chart used in this study.

Table 2. *The best top five models along with their input combination and statistics values.*

| Input combination | MSE | RMSE | $R^2$ | Adj_$R^2$ | Cp | Amemiya's PC | AIC | Score |
|---|---|---|---|---|---|---|---|---|
| $T_{max}$ $RH_{mean}$ | 107.14 | 10.35 | 0.90 | 0.88 | 120.13 | 90.91 | 107.45 | 0.957 |
| $T_{min}$ $RH_{mean}$ RF | 114.78 | 10.71 | 0.89 | 0.88 | 128.70 | 97.39 | 115.09 | 0.956 |
| $RH_{mean}$ SSh | 107.06 | 10.34 | 0.90 | 0.88 | 120.04 | 90.84 | 107.37 | 0.956 |
| $T_{max}$ RF SSh | 88.22 | 9.39 | 0.93 | 0.92 | 98.91 | 74.85 | 88.52 | 0.955 |
| $T_{min}$ $T_{max}$ RF | 115.43 | 10.74 | 0.90 | 0.88 | 129.42 | 97.94 | 115.73 | 0.951 |

Using the meteorological data available obtained from the weather station. Crop evapotranspiration is calculated as:

$$ET_c = K_c \times ET_0. \tag{6}$$

The FAO-56 PM method was used as a standard method in the present research.

$$ET_0 = \frac{0.408(R_n - G) + \frac{900}{T_a+273}u(e_s - e_a)}{\Delta + \gamma(1 + 0.34u)} \tag{7}$$

where $ET_0$ is the potential reference crop evapotranspiration (mm $\text{day}^{-1}$); $R_n$ is the net radiation (MJ $\text{m}^{-2}$ $\text{day}^{-1}$); $G$ is the soil heat flux (MJ

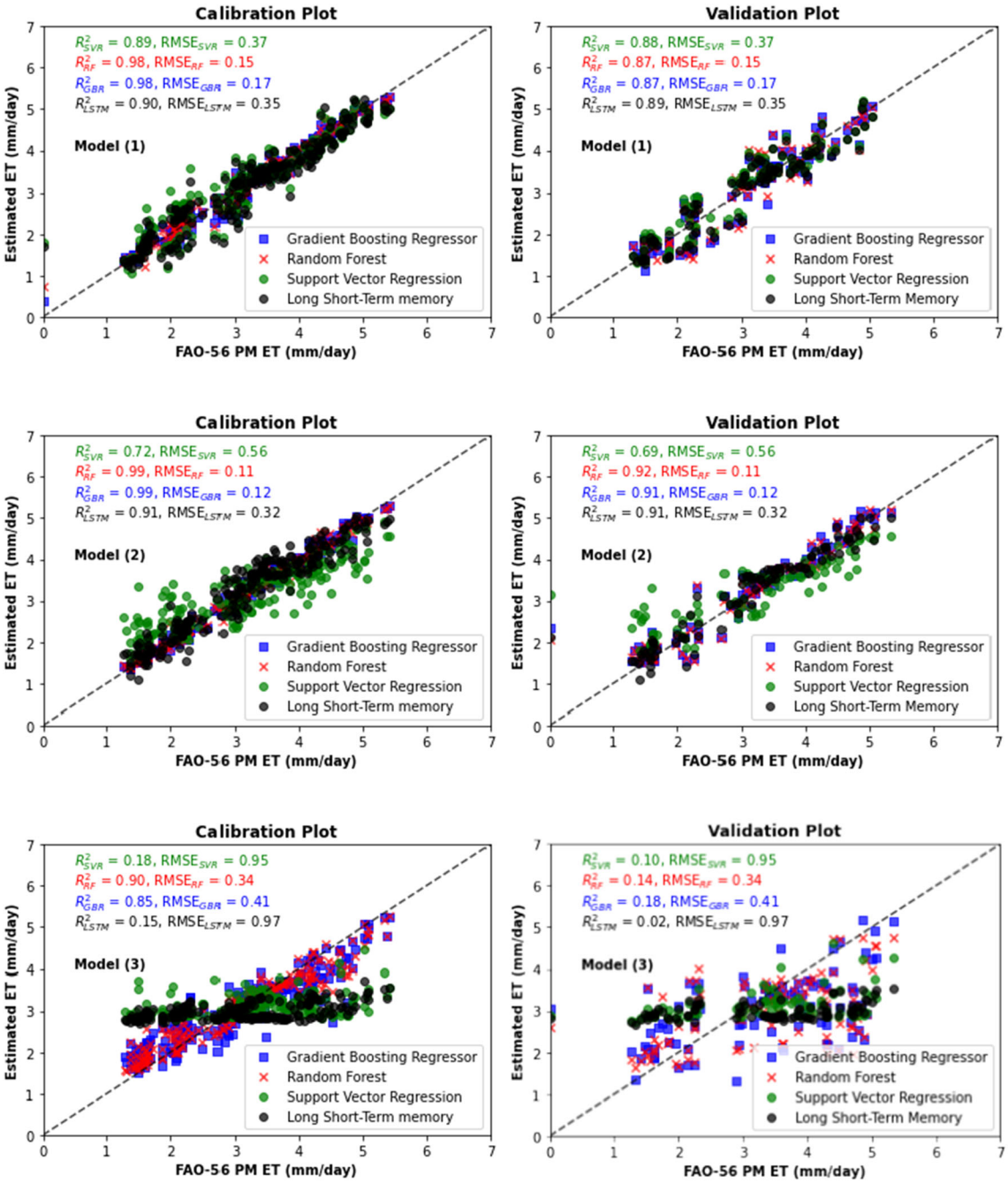


Figure 4. Scatter plots of the FAO-56 *ET* referred as $ET_0$ and those estimated by the ML models. **(a)** SVR, **(b)** RF, **(c)** GBR, and **(d)** LSTM with all parameters of input for during the testing period at Hawalbagh experimental farm.

$m^{-2}\,day^{-1}$), which is assumed as null for daily periods; $T_a$ is the average daily temperature (°C); $u$ is the wind speed at a height of 2 m (m $sec^{-1}$); $e_s$ is the saturation vapour pressure deficit (kPa); $e_s$ is the actual vapour pressure deficit (kPa); $e_s - e_a$ is the vapour pressure deficit (kPa); $\Delta$ is the slope of the saturation vapour pressure–temperature curve (kPa $°C^{-1}$); and $\gamma$ is the psychometric constant (kPa $°C^{-1}$).

### 3.4 *Machine learning models*

#### 3.4.1 *Gradient boosting regressor*

The gradient boosting regressor (GBR) is an ensemble machine learning model that uses a collection of sequentially arranged tree models to make predictions. It is designed to improve the accuracy of weak prediction models, typically decision trees, by iteratively building on the errors

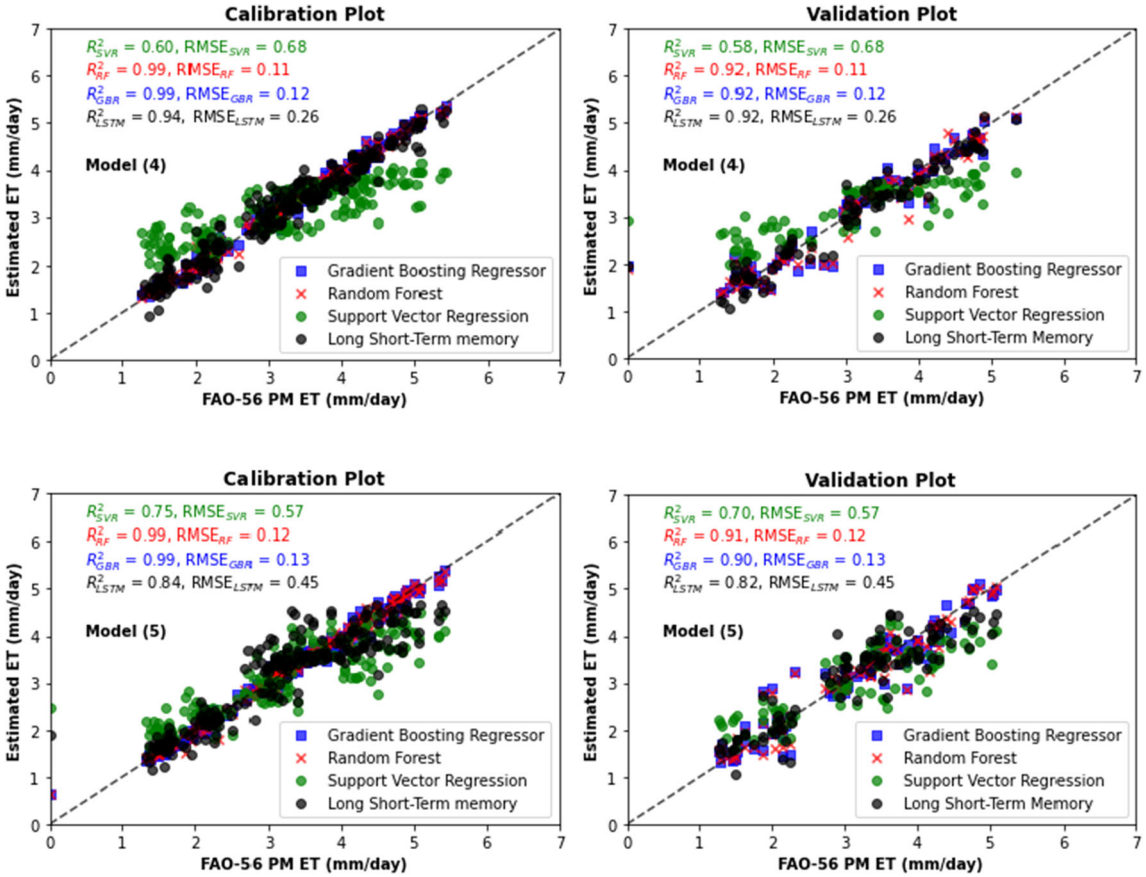


Figure 4. (*Continued.*)

of previous models in the ensemble. This technique is known as 'boosting'.

As the GBR algorithm iteratively builds each model, it focuses on improving the predictions of the previous model by learning from the errors it made. The result is a more robust and accurate model that is capable of making better predictions on the given dataset. Overall, the GBR model is a powerful tool for machine learning tasks that involve predicting numerical values. A GBR with $M$ number of trees can be stated as:

$$f_M(x_j) = \sum_{m}^{M} \gamma_m h_m(x_j) \tag{8}$$

where $h_m$ is a weak learner that performs poorly individually and $\gamma_m$ is a scaling factor adding the contribution of a tree to the model.

GBR model is implemented using the gradient boosting regressor (GBR) method provided by Scikit-learn, a widely used machine learning library. The method is based on the algorithm developed by Jerome H Friedman in the late 1990s, which involves iteratively improving the performance of weak learners by adding them to an ensemble model. The resulting model is a more accurate and robust predictor than any of the individual weak learners used to build it. Overall, the GBR model is an effective approach to solving regression problems in machine learning.

### 3.4.2 *Support Vector Regression*

Vapnik (1995) introduced support vector machine algorithm, which is one of the most popular method among machine learning techniques (Wu *et al.* 2008). SVM is used by several researchers across the globe for the estimation of *ET* (Shrestha and Shukla 2015; Dou and Yang 2018; Tang *et al.* 2019). The most important property of SVM is inclusion of kernel, which makes it powerful to deal with nonlinear properties of the system. SVM is a supervised machine learning method used for both regression and classification analysis in agriculture. It plots data into a high-dimensional feature space

Table 3. *Performance of random forest (RF), support vector regressor (SVR), gradient boosting regressor (GBR), and long shortterm memory (LSTM).*

| | | Calibration | | | | Validation | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | ML algorithm | MSE | $R^2$ | NSE | RMSE | MSE | $R^2$ | NSE | RMSE |
| M1 | RF | 0.02 | 0.98 | 0.98 | 0.15 | 0.14 | 0.87 | 0.87 | 0.38 |
| | SVR | 0.14 | 0.89 | 0.89 | 0.37 | 0.13 | 0.88 | 0.88 | 0.36 |
| | GBR | 0.03 | 0.98 | 0.98 | 0.17 | 0.13 | 0.87 | 0.87 | 0.36 |
| | LSTM | 0.12 | 0.90 | 0.90 | 0.35 | 0.12 | 0.89 | 0.89 | 0.34 |
| M2 | RF | 0.01 | 0.99 | 0.99 | 0.11 | 0.11 | 0.92 | 0.92 | 0.34 |
| | SVR | 0.32 | 0.72 | 0.72 | 0.56 | 0.43 | 0.69 | 0.69 | 0.66 |
| | GBR | 0.01 | 0.99 | 0.99 | 0.12 | 0.13 | 0.91 | 0.91 | 0.36 |
| | LSTM | 0.10 | 0.91 | 0.91 | 0.32 | 0.13 | 0.91 | 0.91 | 0.36 |
| M3 | RF | 0.11 | 0.90 | 0.90 | 0.34 | 1.19 | 0.14 | 0.14 | 1.09 |
| | SVR | 0.91 | 0.18 | 0.18 | 0.95 | 1.24 | 0.10 | 0.10 | 1.11 |
| | GBR | 0.17 | 0.85 | 0.85 | 0.41 | 1.14 | 0.18 | 0.18 | 1.07 |
| | LSTM | 0.94 | 0.15 | 0.15 | 0.97 | 1.36 | 0.02 | 0.02 | 1.16 |
| M4 | RF | 0.01 | 0.99 | 0.99 | 0.11 | 0.10 | 0.92 | 0.92 | 0.32 |
| | SVR | 0.46 | 0.60 | 0.60 | 0.68 | 0.56 | 0.58 | 0.58 | 0.75 |
| | GBR | 0.01 | 0.99 | 0.99 | 0.12 | 0.11 | 0.92 | 0.92 | 0.33 |
| | LSTM | 0.07 | 0.94 | 0.94 | 0.26 | 0.11 | 0.92 | 0.92 | 0.33 |
| M5 | RF | 0.01 | 0.99 | 0.99 | 0.12 | 0.09 | 0.91 | 0.91 | 0.31 |
| | SVR | 0.32 | 0.75 | 0.75 | 0.57 | 0.33 | 0.70 | 0.70 | 0.57 |
| | GBR | 0.02 | 0.99 | 0.99 | 0.13 | 0.11 | 0.90 | 0.90 | 0.33 |
| | LSTM | 0.20 | 0.84 | 0.84 | 0.45 | 0.20 | 0.82 | 0.82 | 0.44 |

using kernels and can classify nonlinearly separable datasets. The accuracy of the SVR model depends on appropriate selection of kernels and parameters, and RBF has been shown to have favourable performance in forecasting. The present study used the LIBSVM library to develop SVR models with the RBF kernel. The fundamental requirement of a kernel is that it must satisfy Mercer's theorem. The optimization algorithm has a global optimization function that makes it different from other machine learning techniques such as ANN, which consider the local maxima. Kernels are the foundation of SVM. The typical assembly of the SVM model is shown in figure 4. The SVM (Support Vector Machine) technique applies the SRM (structural risk minimization) principle to minimize the risk of overfitting and find the optimal decision boundary between different classes.

### 3.4.3 *Random forest*

RF is an ensemble machine algorithm, which is efficient in both classification and regression. It is generally used for predictive analysis. In predicting the final output, combination of all decision trees is considered rather than depending on individual decision trees. The foundation of RF is based on supervised machine learning approaches, which are constantly used in machine learning and popularly used in hydrology. Observed and predicted values are used to calculate the sum of square error (SSE). This process occurs repeatedly until the whole set of data is covered. Mathematically RF can be expressed as:

$$u(x) = u_0(x) + u_1(x) + u_2(x) + u_3(x) + \cdots \quad (9)$$

where the resulting function $u$ is the addition of individual base model $u_i$, where each individual base repressor is the individual decision tree. The fundamental principle behind RF model is to integrate several decision trees in estimating the final result rather than individual decision trees.

Random forests or random decision forests are a machine learning technique used for classification, regression, and other tasks. They work by creating multiple decision trees during training and then using them to make predictions by taking the mode of the classes for classification or the mean prediction for regression. Random forests help to reduce the overfitting of decision trees to their training set (Smith *et al.* 2013; Misra and Li 2020). Extra trees (extremely randomized trees) is an ensemble learning method that builds on the Random Forest algorithm (Breiman 2001). Each decision tree in the extra trees forest is constructed

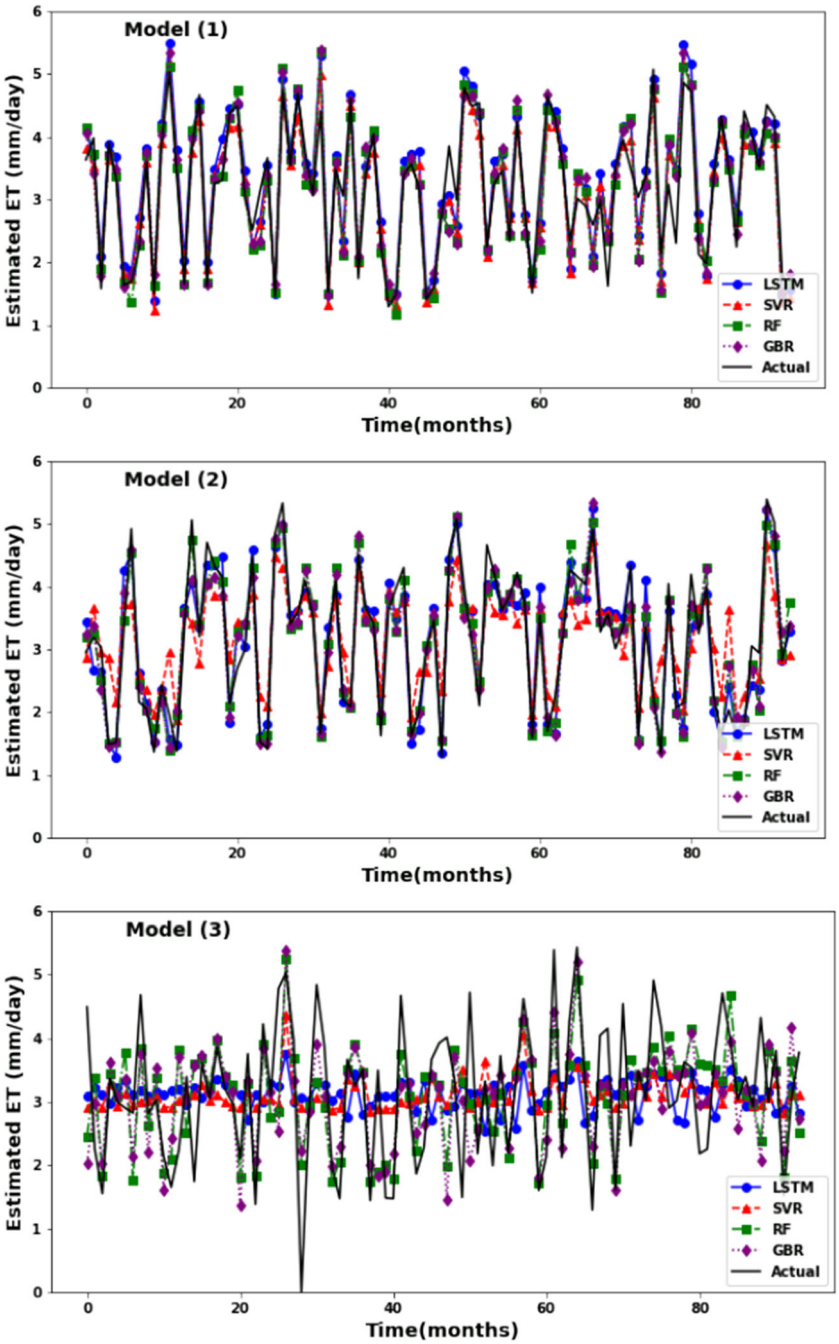


Figure 5. Comparison of observed and estimated $ET$ referred as $ET_0$ by different models (GBR, LSTM, RF, and SVR) with varying parameters of input for the validation period at Hawalbagh experimental farm.

from the original training sample. However, at each test node, each tree is provided with a random sample of features from the feature set, and each decision tree must select the best feature to split the data based on some mathematical criteria, typically the Gini Index. This process of using a random subset of features at each test node results in the creation of multiple de-correlated decision trees, which can improve the accuracy and reduce overfitting.

### 3.4.4 *Long short-term memory*

Long short-term memory (LSTM) is a type of recurrent neural network (RNN) that is designed to address the vanishing gradient problem and the inability of traditional RNNs to handle long-term dependencies in sequential data.

LSTM networks are comprised of memory cells that can selectively remember or forget information based on inputs from different gates. These gates are

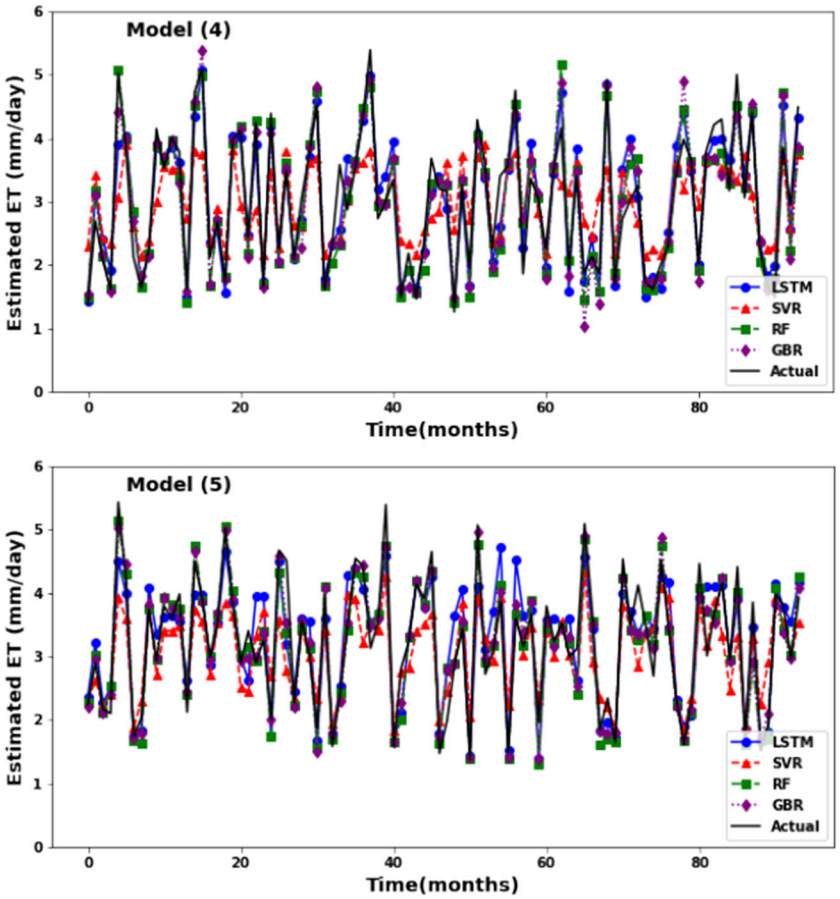


Figure 5. (*Continued.*)

made up of different layers, including the input gate, forget gate, output gate, and cell state layer. The interactions between these layers allow the LSTM network to selectively retain or discard information from previous time steps, making it more effective for modelling long-term dependencies. The structure of an LSTM model is composed of an input gate, an output gate and a forget gate. The gate's value in LSTM is calculated using the previous cell value $C_{t-1}$, previously hidden values $h_{t-1}$ and input $x_t$.

$$i_t = F(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + bias_i) \quad (10)$$

$$O_t = F(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}C_{t-1} + bias_o) \quad (11)$$

$$f_t = F\big(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + bias_f\big). \quad (12)$$

And the cell value is calculated using

$$C_t = f_t C_{t-1} + i_t F(W_{xc}x_t + W_{hc}h_{t-1} + bias_c), \quad (13)$$

$$h_t = O_t \tan^{-1}(C_t). \quad (14)$$

### 3.5 *Model assessment*

To examine the accuracy of $ET_0$ predicted by different algorithms with observed values, different statistical indicators coefficient of determination ($R^2$), root mean squared error (RMSE), mean square error (MSE) and Nash–Sutcliffe Efficiency (NSE) were used. The mathematical expressions are as follows:

$$R^2 = 1 - \frac{\sum (ET_{\text{Pre}} - ET_{\text{Obs}})^2}{\sum (ET_{\text{Obs}} - ET_{\text{Mean}})^2} \quad (15)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (ET_{\text{Pre}} - ET_{\text{Obs}})^2}{n}} \quad (16)$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (ET_{\text{Pre}} - ET_{\text{Obs}})^2 \quad (17)$$

$$\text{NSE} = 1 - \left[\frac{\sum_{i=1}^{n} (ET_{\text{Obs}} - ET_{\text{Pre}})^2}{\sum_{i=1}^{n} (ET_{\text{Pre}} - ET_{\text{Mean}})^2}\right] \quad (18)$$

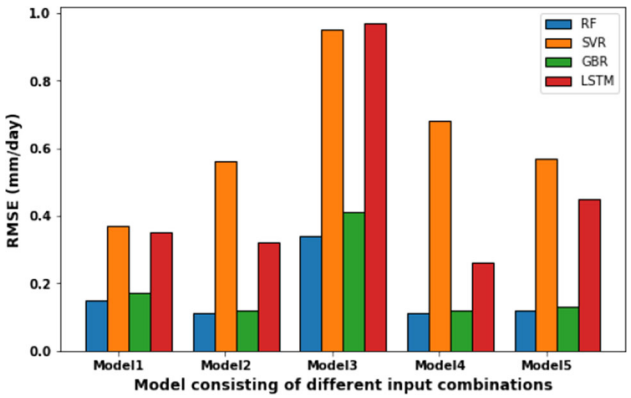


Figure 6. Comparison of RMSE values for Hawalbagh experimental farm for different input combinations.

where $ET_{\text{Pre}}$ (mm/day) is predicted $ET_0$ using different machine learning algorithm at instant $i$, $ET_{\text{Obs}}$ (mm/day) is observed $ET_0$ at instant $i$, $ET_{\text{Mean}}$ (mm/day) is the average $ET_0$ at instant $i$, and $n$ is the number of total observations.

## 4. Results and discussion

The study analysed the performance of LSTM, SVR, GBR, and RF models in estimating daily $ET_0$ using four pre-processing data methods: Pearson correlation, PCA, Lasso model, and random forest. The climate variables considered were daily maximum temperature, minimum temperature, relative humidity, and solar radiation. The optimal input combinations and each model's performance in terms of $R^2$, RMSE, NSE, and MSE were presented for Hawalbagh experimental farm observatory. The Grid Search CV python library module was used to optimize SVR parameters. The performance of different machine learning algorithms in terms of statistical parameters MSE, RMSE, MSE, and $R^2$ were summarized in table 3.

### 4.1 *ML models performance with various input combinations*

The study evaluated the performance of LSTM, SVR, GBR, and RF models for estimating daily $ET_0$ using different pre-processing methods to select the best set of input variables. The four pre-processing methods evaluated were Pearson correlation, principal component analysis (PCA), Lasso model, and random forest. The optimal input combinations and the performance of each model in terms of $R^2$, RMSE, NSE, and MSE were listed in table 3 for Hawalbagh experimental farm observatory. Table 3 showed the summary of the LSTM, SVR, GBR, and RF model performances for Hawalbagh experimental farm observatory. Taking everything into account, the RF and GBR models are the most robust among the four ML models regardless of under which input combination, trailed by LSTM and SVR models, which could generally accomplish agreeable accuracy. Figure 4 showed the comparisons between observed $ET_0$ and model-estimated values in the form of scatter plots of the FAO-56 $ET_0$ and those estimated by the ML models with all input parameters during the testing periods. All scatter plots of different ML models showed various distributions. The study compared the performance of LSTM, SVR, GBR, and RF models in estimating daily $ET_0$ using different input combinations and pre-processing methods. The RF and GBR models showed closer agreement with observed $ET_0$ and the RF model performed marginally better than the GBR model, with an estimated $R^2$ value of 0.99 for all model input combinations. The RF model achieved the best performance among the evaluated models with all input combinations at the Hawalbagh experimental station, followed by GBR. The RF model achieved excellent performance, with an RMSE of 0.11 mm/day, MSE of 0.01 mm/day, and $R^2$ of 0.99, followed by GBR with an RMSE of 0.17 mm/day, MSE of 0.03 mm/day, and $R^2$ of 0.98 at the Hawalbagh experimental farm observatory (figure 4). Figure 5 shows the comparisons between observed $ET_0$ and model-estimated values in the form of a line plot. The model results were also compared with $ET_0$ calculated using the Penman–Monteith equation, and the results were found to be very similar to the FAO-56 $ET_0$ based comparison, indicating that the model results were not overstated. In summary, the RF model is the most robust among the four ML models, followed by the GBR and SVR models. The comparison of RMSE values for Hawalbagh experimental farm for different input combinations is shown in figure 6. The RF and GBR models achieved acceptable accuracy across different input combinations of meteorological variables, which makes them more cost-effective and practical for development and application. These models can simulate $ET_0$ where meteorological information is limited.

### 4.2 *Comparison with similar studies*

The paper investigated the performance of four ML models in estimating daily reference

evapotranspiration, including LSTM, GBR, RF, and SVR, and found that RF performed the best, with GBR, SVR, and LSTM also producing accurate results. This is consistent with previous studies that have found ANN, LSTM, and SVR to be effective for $ET_0$ estimation. It has also shown that this study resembles the assessment of reference evapotranspiration found by Raza *et al.* (2020) and Heramb *et al.* (2023), where SVR was found to be used as an alternative $ET_0$ estimation model to the subsistence of conventional methods. Based on table 3, the RF and GBR models performed well in terms of accuracy and computational demand, with the RF model being particularly efficient and achieving low test RMSE values regardless of input combination. The study also found that the two input combinations of maximum temperature and mean relative humidity worked comparatively well with the RF model, followed by the GBR model. The RF and GBR model performed well in estimating $ET_0$ with limited inputs, and therefore, it can be suggested for $ET_0$ estimation in situations with limited meteorological parameter availability.

## 5. Conclusions

This study aimed to compare the performance of four machine learning models, namely LSTM, SVR, GBR, and RF, in estimating $ET_0$ using four different input combinations at two different stations.

The study concluded that using all input variables provided the best performance, but also found that using a combination of three variables (temperature, wind speed, and relative humidity) or two variables (temperature and relative humidity, temperature and wind speed) can provide nearly identical results to using all variables. Therefore, the study suggests that using all variables should be the first priority, followed by three and two variable combinations.

The study found that the RF and GBR model had the best performance among the four tested models, regardless of station or input combination. The LSTM and SVR models were able to achieve good performance using only temperature, relative humidity, and wind speed data, making them more practical and useful for application. The input combination of temperature and wind speed followed by temperature and relative humidity also showed good results, while other models did not perform as well as the LSTM and SVR models with the mentioned input combinations.

The study concluded that using a combination of three or two meteorological parameters, such as temperature, wind speed, and relative humidity, can still provide accurate estimation of reference $ET_0$, even if not all parameter information is available. The LSTM and SVR models were found to be highly relevant for modelling $ET_0$ at spatiotemporal scales, even when meteorological parameter information is limited or fragmented, and are strongly recommended for such cases.

The findings of this study can be applied to model $ET_{\text{crop}}$ for various crops grown in the region, including rice, wheat, and pulses. Additionally, this study provides a useful reference for researchers to develop similar models for other study areas. Since machine learning techniques are data-driven techniques, performance of these ML techniques would deteriorate in case of limited meteorological data. Further, the application of different optimizers in conjunction with machine learning and deep learning models with cross-station data in different agroclimatic zone can be employed in $ET_0$ modelling.

## Acknowledgements

Author is thankful to all the field and laboratory staff for their help in this study and very much thankful to the ICAR-VPKAS, Almora as well as ICAR, New Delhi for providing financial support during the course of investigation.

## Author statement

UK did the modelling exercise, prepared the manuscript, result interpretation, and manuscript revision during peer-review process.

## References

Abdullah S S, Malek M A, Abdullah N S, Kisi O and Yap K S 2015 Extreme learning machines: A new approach for prediction of reference evapotranspiration; *J. Hydrol.* **527** 184–195.

Adamala S, Raghuwanshi N S, Mishra A and Singh R 2019 Generalized wavelet neural networks for evapotranspiration modeling in India; *ISH J. Hydraul. Eng.* **25**(**2**) 119–131.

Allen R, Pereira L, Rae D and Smith M 1998 Crop evapotranspiration-guidelines for computing crop water requirements-FAO irrigation and drainage paper 56. Food and Agriculture Organization, United Nations, Rome.

Ambade B, Sankar T K, Kumar A, Gautam A S and Gautam S 2021 COVID-19 lockdowns reduce the Black carbon and

polycyclic aromatic hydrocarbons of the Asian atmosphere: Source apportionment and health hazard evaluation; *Environ. Dev. Sustain.* **23** 12,252–12,271, https://doi.org/10.1007/s10668-020-01167-1.

Breiman L 2001 Random forests; *Machine Learning* **45** 5–32, https://doi.org/10.1023/A:1010933404324.

Caminha H D, da Silva T C, da Rocha A R and Lima S C R V 2017 Estimating reference evapotranspiration using data mining prediction models and feature selection; *ICEIS* **1** 272–279.

Chauhan S and Shrivastava R K 2008 Performance evaluation of reference evapotranspiration estimation using climate based methods and artificial neural networks; *Water Resour. Manag.* **23** 825–837.

Cobaner M 2013 Reference evapotranspiration based on Class A pan evaporation via wavelet regression technique; *Irrig. Sci.* **31** 119–134.

Dou X and Yang Y 2018 Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems; *Comput. Electron Agric.* **148** 95–106.

Falamarzi Y, Palizdan N, Huang Y F and Lee T S 2014 Estimating evapotranspiration from temperature and wind speed data using artificial and wavelet neural networks (WNNs); *Agric. Water Manag.* **140** 26–36.

Fan J, Ma X, Wu L, Zhang F, Yu X and Zeng W 2019 Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data; *Agric. Water Manag.* **225** 105758.

Feng Y, Cui N, Gong D, Zhang Q and Zhao L 2017 Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling; *Agric. Water Manag.* **193** 163–173.

Ferreira L B, da Cunha F F, de Oliveira R A and Fernandes Filho E I 2019 Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach; *J. Hydrol.* **572** 556–570.

Gautam A S, Kumar S, Gautam S, Anand A, Kumar R, Joshi A, Bauddh K and Singh K 2021 Pandemic induced lockdown as a boon to the environment: Trends in air pollution concentration across India; *Asia-Pacific J. Atmos. Sci.* **57** 741–756, https://doi.org/10.1007/s13143-021-00232-7.

Gautam A S, Singh K, Sharma M, Sneha G, Joshi A and Kumar S 2023 Classification of different sky conditions based on solar radiation extinction and the variability of aerosol optical depth, angstrom exponent, fine particles over Tehri Garhwal, Uttarakhand, India; *MAPAN* **38** 21–36, https://doi.org/10.1007/s12647-022-00533-w.

Heramb P, Ramana Rao K V, Subeesh A and Srivastava A 2023 Predictive modelling of reference evapotranspiration using machine learning models coupled with Grey Wolf Optimizer; *Water* **15**(**5**) 856, https://doi.org/10.3390/w15050856.

Kelley J and Pardyjak E R 2019 Using neural networks to estimate site-specific crop evapotranspiration with low-cost sensors; *Agronomy* **9**(**2**) 108.

Khoob A R 2008 Artificial neural network estimation of reference evapotranspiration from pan evaporation in a semi-arid environment; *Irrig. Sci.* **27**(**1**) 35–39.

Khosravi K, Mao L, Kisi O, Yaseen Z M and Shahid S 2018 Quantifying hourly suspended sediment load using data mining models: Case study of a glacierized Andean catchment in Chile; *J. Hydrol.* **567** 165–179.

Kovoor G M and Nandagiri L 2018 *Sensitivity analysis of FAO-56 Penman–Monteith reference evapotranspiration estimates using Monte Carlo simulations; Hydrologic modeling, Water Science and Technology Library*; Vol. 81, Springer, Singapore, pp. 73–84.

Kumar M, Raghuwanshi N S and Singh R 2011 Artificial neural networks approach in evapotranspiration modeling: A review; *Irrig. Sci.* **29** 11–25.

Kumar U, Sahoo B, Chatterjee C and Raghuwanshi N S 2020 Evaluation of simplified surface energy balance index (S-SEBI) method for estimating actual evapotranspiration in Kangsabati Reservoir command using Landsat 8 Imagery; *J. Indian Soc. Remote Sens.* **48** 1421–1432.

Kumar U, Srivastava A, Kumari N, Sahoo B, Chatterjee C and Raghuwanshi N S 2021a Evaluation of spatio-temporal evapotranspiration using satellite-based approach and Lysimeter in the agriculture dominated catchment; *J. Indian Soc. Remote Sens.* **49** 1939–1950.

Kumar U, Rashmi, Chatterjee C and Raghuwanshi N S 2021b Comparative evaluation of simplified surface energy balance index-based actual ET against Lysimeter data in a tropical river basin; *Sustainability* **13** 13786.

Kumar U, Panday S C, Kumar J, Parihar M, Meena V S, Bisht J K and Kant L 2022 Use of a decision support system to establish the best model for estimating reference evapotranspiration in sub-temperate climate: Almora, Uttarakhand; *Agr. Eng. Int.: CIGR J.* **24**(**1**) 41–50.

Kumar U, Singh D K, Panday S C, Bisht J K and Kant L 2023a Spatio-temporal trend and change detection of rainfall for Kosi River basin, Uttarakhand using long-term (115 years) gridded data; *Arab. J. Geosci.* **16** 173.

Kumar U, Rashmi, Srivastava A, Kumari N, Chatterjee C and Raghuwanshi N S 2023b Evaluation of standardized MODIS-Terra satellite-derived evapotranspiration using genetic algorithm for better field applicability in a tropical river basin; *J. Indian Soc. Remote Sens.* **51** 1001–1012.

Kumari N and Srivastava A 2020 An approach for estimation of evapotranspiration by standardizing parsimonious method; *Agric. Res.* **9** 301–309, https://doi.org/10.1007/s40003-019-00441-7.

Ling Yao 2017 Causative impact of air pollution on evapotranspiration in the North China Plain; *Environ. Res.* **158** 436–442.

Marti P, Nazemi A H, Sadraddini A A, Kisi O, Landeras G and Fakheri F A 2015 Local vs. external training of neuro-fuzzy and neural networks models for estimating reference evapotranspiration assessed through k-fold testing; *Hydrol. Res.* 4**6** 72.

Maza M, Srivastava A, Bisht D S, Raghuwanshi N S, Bandyopadhyay A, Chatterjee C and Bhadra A 2020 Simulating hydrological response of a monsoon dominated reservoir catchment and command with heterogeneous cropping pattern using VIC model; *J. Earth Syst. Sci.* **129**(**1**) 1–16, https://doi.org/10.1007/s12040-020-01468-z.

Misra S and Li H 2020 Chapter 9: Noninvasive fracture characterization based on the classification of SonicWave travel times; In: *Machine learning for subsurface characterization* (eds) Misra S, Li H and He J, Gulf Professional Publishing: Houston, TX, USA, pp. 243–287, ISBN 978-0-12-817736-5.

Naganna S R, Deka P C, Ghorbani M A, Biazar S M, Al-Ansari N and Yaseen Z M 2019 Dew point temperature estimation: Application of artificial intelligence model integrated with nature-inspired optimization algorithms; *Water* **11**(**4**) 742.

Nepolian J V, Siingh D, Singh R P, Gautam A S and Sneha G 2021 Analysis of positive and negative atmospheric air ions during new particle formation (NPF) events over urban city of India; *Aerosol. Sci. Eng.* **5** 460–477, https://doi.org/10.1007/s41810-021-00115-4.

Nourani V, Elkiran G and Abdullahi J 2019 Multi-station artificial intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements; *J. Hydrol.* **577** 123958.

Pangam H, Rao K V R, Subeesh A and Srivastava A 2023 Predictive modelling of reference evapotranspiration using machine learning models coupled with grey wolf optimizer; *Water* **15**(**5**) 856, https://doi.org/10.3390/w15050856.

Partal T 2009 Modeling evapotranspiration using discrete wavelet transform and neural networks; *Hydrol. Process.* **23**(**25**) 3545–3555.

Raza A, Shoaib M, Faiz M A, Baig F, Khan M M, Ullah M K and Zubair M 2020 Comparative assessment of reference evapotranspiration estimation using conventional method and machine learning algorithms in four climatic regions; *Pure Appl. Geophys.* **177** 4479–4508.

Shiri J, Kisi O, Landeras G, López J J, Nazemi A H and Stuyt L C P M 2012 Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain); *J. Hydrol.* **414–415** 302–316.

Shrestha N K and Shukla S 2015 Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment; *Agric. Forest Meteorol.* **200** 172–184, https://doi.org/10.1016/j.agrformet.2014.09.025.

Smith P F, Ganesh S and Liu P A 2013 Comparison of random forest regression and multiple linear regression for prediction in neuroscience; *J. Neurosci. Methods* **220** 85–91.

Srivastava A, Sahoo B, Raghuwanshi N S and Singh R 2017 Evaluation of variable-infiltration capacity model and MODIS-terra satellite-derived grid-scale evapotranspiration estimates in a River Basin with Tropical Monsoon-Type climatology; *J. Irrig. Drain. Eng.* **143**(**8**) 04017028, https://doi.org/10.1061/(ASCE)IR.1943-4774.0001199.

Srivastava A, Sahoo B, Raghuwanshi N S and Chatterjee C 2018 Modelling the dynamics of evapotranspiration using variable infiltration capacity model and regionally calibrated Hargreaves approach; *Irrig. Sci.* **36** 289–300, https://doi.org/10.1007/s00271-018-0583-y.

Tabari H and Hosseinzadeh Talaee P 2013 Multilayer perceptron for reference evapotranspiration estimation in a semiarid region; *Neural Comput. Appl.* **23** 341–348.

Tang W, Li Z and Cassar N 2019 Machine learning estimates of global marine nitrogen fixation; *J. Geophys. Res.: Biogeosci.* **124** 717–730, https://doi.org/10.1029/2018JG004828.

Thapliyal J, Bhattacharyya M, Prakash S, Patni B, Gautam S and Gautam A S 2022 Addressing the relevance of COVID-19 pandemic in nature and human socio-economic fate; *Stoch. Environ. Res. Risk Assess.* **36** 3239–3253, https://doi.org/10.1007/s00477-022-02191-5.

Trajkovic S 2010 Testing hourly reference evapotranspiration approaches using lysimeter measurements in a semiarid climate; *Hydrol. Res.* **41**(**1**) 38–49.

Trigo I F, de Bruin H, Beyrich F, Bosveld F C, Gavilán P, Groh J and López-Urrea R 2018 Validation of reference evapotranspiration from Meteosat Second Generation (MSG) observations; *Agric. Forest Meteorol.* **259** 271–285.

Valipour M and Sefidkouhi M A G 2018 Temporal analysis of reference evapotranspiration to detect variation factors; *Int. J. Global Warming* **14**(**3**) 385–401.

Vapnik V N 1995 *The Nature of Statistical Learning Theory*; New York: Springer.

Wang J, Ma Y, Zhang L, Gao R X and Wu D 2018 Deep learning for smart manufacturing: Methods and applications; *J. Manuf. Syst.* **48** 144–156.

Wu L and Fan J 2019 Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration; *PLoS ONE* **14** e0217520.

Wu X, Kumar V, Ross Quinlan J *et al.* 2008 Top 10 algorithms in data mining; *Knowl. Inf. Syst.* **14** 1–37, https://doi.org/10.1007/s10115-007-0114-2.

Xiao Q, Li C, Tang Y, Li L and Li L 2019 A knowledge-driven method of adaptively optimizing process parameters for energy efficient turning; *Energy* **166** 142–156.

Yaseen Z M, Allawi M F, Yousif A A, Jaafar O, Hamzah F M and El-Shafie A 2018 Non-tuned machine learning approach for hydrological time series forecasting; *Neural Comput. Appl.* **30**(**5**) 1479–1491.

Zanetti S S, Sousa E F, Oliveira V P, Almeida F T and Bernardo S 2007 Estimating evapotranspiration using artificial neural network and minimum climatological data; *J. Irrig. Drain. Eng.* **133** 83–89.

Corresponding editor: Parthasarathi Mukhopadhyay