https://doi.org/10.17221/101/2023-RAE

An effective machine learning model for the estimation of reference evapotranspiration under data-limited conditions

Saravanan Karuppanan[1], Saravanan Ramasamy[2], Balaji Lakshminarayanan[2], Sreemanthrarupini Nariangadu Anuthaman[2]*

[1]*Dhanalakshmi Srinivasan College of Engineering and Technology, Mamallapuram, Chennai, India*
[2]*Centre for Water Resources, Anna University, Chennai, India*

**Corresponding author: slvnsaran@gmail.com*

Citation: Karuppanan S., Ramasamy S., Lakshminarayanan B., Nariangadu Anuthaman S. (2025): An effective machine learning model for the estimation of reference evapotranspiration under data-limited conditions. Res. Agr. Eng., 71: 22–37.

Abstract: Reference crop evapotranspiration (ET_o) is a vital hydrological component influenced by various climate variables that impact the water and energy balances. It plays a crucial role in determining crop water requirements and irrigation scheduling. Despite the availability of numerous approaches for estimation, accurate and reliable ET_o estimation is essential for effective irrigation water management. Therefore, this study aimed to identify the most suitable machine learning model for assessing ET_o using observed daily values of limited input parameters in tropical savannah climate regions. Three machine learning models – a long short-term memory (LSTM) neural network, an artificial neural network (ANN), and support vector regression (SVM) – were developed with four different input combinations, and their performances were compared with those of locally calibrated empirical equations. The models were evaluated using statistical indicators such as the root mean square error (RMSE), coefficient of determination (R^2), and the Nash-Sutcliffe efficiency (NSE). The results showed that the LSTM model, using the combination of temperature and wind speed, provided more reliable predictions with R^2 values greater than 0.75 and RMSEs less than 0.63 mm·day^{-1} across all the considered weather stations. This study concludes that, especially under limited data conditions, the developed deep learning model improves the ET_o estimation more accurately than empirical models for tropical climatic regions.

Keywords: artificial neural networks; empirical equations; long short-term memory neural networks; machine learning; reference crop evapotranspiration, support vector machines

Given the rising demand for water resources due to climate change, population growth, and agricultural activities, it is essential to optimise limited water resources, especially in agricultural production systems. The precise monitoring of all elements in the hydrological cycle and utilising this information for decision support in water resource management can ensure the sustainability of water resources. It is crucial to evaluate various methods for increasing the water usage efficiency to conserve scarce water resources. Understanding crop evapotranspiration (*ET*) is necessary for water management studies, including irrigation scheduling, crop water requirement estimation, and hydrological modelling, as it is a significant component of the hydrologic cycle. Direct measurements of the actual *ET* include water vapour transfer methods (eddy covariance) or water budget measurements (lysimeters), but their availability is limited due to their high complexity and cost. Alternatively, the crop *ET* can be estimated using the reference evapotranspiration (ET_o) and crop coefficient (K_c) (Fan et al. 2018).

https://doi.org/10.17221/101/2023-RAE

The Penman-Monteith (FAO-56 PM) model, recommended by the Food and Agriculture Organization of the United Nations (FAO), incorporates thermodynamic and aerodynamic effects and is the reference model for the accurate estimation of ET_o worldwide (Allen et al. 1998). The main drawback of this model is that it requires several climatic variables (maximum and minimum temperature, relative humidity, solar radiation, and wind speed), which are not readily available at most weather stations in developing countries (Karuppanan et al. 2022). Therefore, there is a need for empirical models that demand fewer climatic variables to predict the ET_o adequately. Various empirical models have been adopted to estimate the ET_o using limited data, such as temperature-based models (Hargreaves and Samani 1985; Ravazzani et al. 2012), radiation-based models (Abtew 1996; Tabari and Talaee 2011), humidity-based models (Schendel 1967), and wind speed-based models (Brockamp and Wenner 1963; Mahringer 1970). However, the estimation accuracy of empirical models varies according to the climatic conditions, and the results may provide substandard estimates, given the non-linear, dynamic, and complex nature of the process.

Advances in data science and artificial intelligence have led to the consideration of deep learning and classical machine learning algorithms for estimating the ET_o. Many studies have addressed the application of machine learning algorithms (Fan et al. 2018; Ferreira et al. 2019; Saggi and Jain 2019; Wu et al. 2019; Ferreira and da Cunha 2020; Raza et al. 2020; Salam and Islam 2020; Bellido-Jiménez et al. 2021; Goyal et al. 2023), such as support vector regression (SVM), artificial neural networks (ANNs), and random forest algorithms, for the estimation of the ET_o. For instance, Ferreira et al. (2019) compared SVM and ANN models with empirical equations under limited data conditions for Brazil and found that ANNs performed better than SVMs. Additionally, Wen et al. (2015) evaluated the applicability of the SVM model to estimate the daily reference evapotranspiration (ET_o) using limited climatic data in the extremely arid region of the Ejina Basin, China. Despite the shortcomings of ANNs, such as slow training, susceptibility to becoming trapped in a local minimum, and the need for substantial training data, researchers have found ANNs to provide better estimations than empirical models.

Deep machine learning models have been effectively utilised in various water-related studies and have shown remarkable results. One commonly used deep learning model, the long short-term memory (LSTM) model, has been applied for predicting streamflow (Ni et al. 2019), soil moisture (Li et al. 2021), and other aspects of the hydrological cycle (Hu et al. 2018). The advantage of the LSTM model is its ability to identify long-term relationships between the input and output in the network. This is achieved by utilising memory cells to store, write, and read data, replacing the traditional hidden layer. The model can be trained for sequence generation by processing actual data sequences one step at a time and predicting the next time step.

In a study by Zhu et al. (2020), three deep learning algorithms (LSTM, deep neural network (DNN), and temporal convolutional networks (TCN)) and two classical machine learning algorithms, SVM and random forest(RF), were compared. The results showed that the deep learning algorithms provided more accurate estimates of the ET_o than did the empirical equations. Zhang et al. (2018) developed an LSTM model to predict water table depth dynamics in subareas of the Hetao Irrigation District, China, showing superior performance compared to other models. Sattari et al. (2021) also utilised LSTM models to estimate the monthly ET_o in the Corum region of Turkey with reasonable accuracy. Furthermore, Sowmya et al. (2020) compared four variants of DNNs in predicting the ET_o in California, USA, and observed convincing performance, with R^2 values ranging between 0.94 and 0.96.

To calculate the ET_o, a comprehensive set of meteorological data, including air temperature, wind speed, sun radiation, and relative humidity, is needed. In some cases, there may be limited ground weather stations collecting continuous and precise data, especially in rural or remote areas. The use of reanalysis data as an alternative data source has its own practical restrictions and limitations, such as observational restrictions and reanalysis reliability based on the location, time frame, and variables under consideration. The LSTM model is a promising tool for handling time series data in irrigation management and agriculture, particularly in ET_o prediction.

This study aimed to evaluate the precision of the LSTM model for the daily ET_o estimation using different combinations of daily climatic data, such as the maximum air temperature, minimum air temperature, wind speed, relative humidity, and sunshine hours, in the tropical savannah climatic re-

https://doi.org/10.17221/101/2023-RAE

gions of South India. The performance of the LSTM model is compared with that of other methods, including ANN, SVM, and empirical equations such as Hargreaves Samani, Schendel, Tabari, and Brockamp and Wenner. The novelty of this work is the accurate estimation of the ET_o by developing machine learning models using minimum input parameters in addition to the empirical models for tropical savannah climatic regions.

MATERIAL AND METHODS

Study area and dataset

In this study, we examined five weather stations located at different elevations (Bangalore, Coimbatore, Annamalai Nagar, Kovilpatti, and Tirupati) under the "tropical savannah" (A_w) climate classification. The location map can be found in Figure 1. We sourced daily climatic data from the Indian Meteorological Department (IMD) Pune, including the maximum and minimum temperatures, maximum and minimum relative humidity, wind speed (measured at 2 m in height), and sunlight hours spanning from 1978 to 2014. We meticulously eliminated the missing and extreme values to minimise any potential errors during measurement. Noisy data encompassed instances such as the minimum temperature exceeding maximum temperature, wind speed surpassing 15 m·s^{-1}, relative humidity surpassing 100% or negative values, as well as negative or implausible sunshine values that exceeded the photoperiod. Table 1 outlines the mean values of the collected meteorological variables and the geographical information of the weather stations.

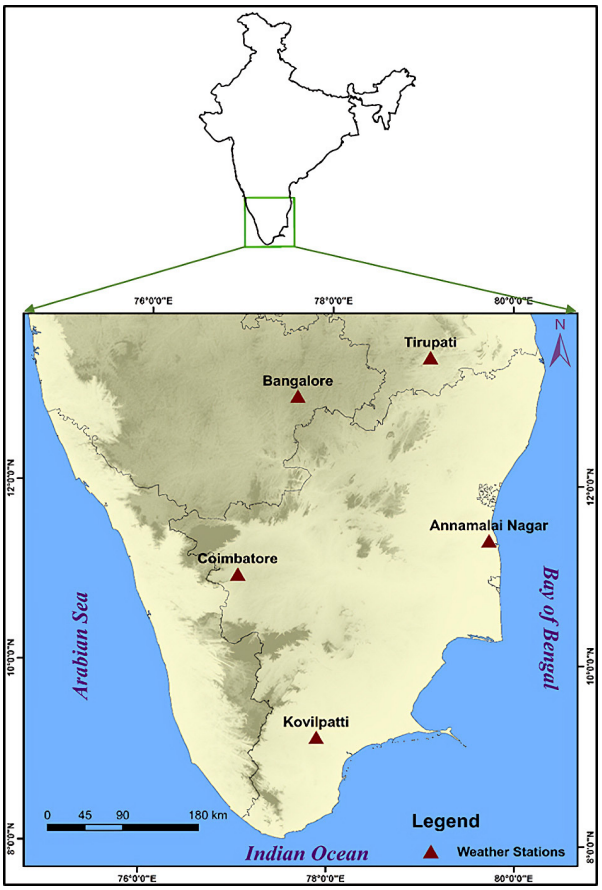

Figure 1. The geographical locations of the selected weather stations

Empirical models

In this study, observed lysimeter data were used for training and testing the machine learning models, as well as for comparison. The performance of the developed models was evaluated against existing empirical models to accurately estimate the ET_o under data-scarce conditions. Vishwakarma et al. (2022) compared 30 empirical methods for humid and subtropical climatic regions and found that radiation-based models exhibit greater accuracy in predicting the ET_o. Niranjan and Nandagiri determined that the calibrated Hargreaves-Samani model provided a precise estimation for the study

Table 1. Mean values of climatic variables and geographical parameters of weather stations

Station	Latitude (N)	Longitude (E)	Altitude (m a.s.l.)	T_{max} (°C)	T_{min} (°C)	RH_{mean} (%)	U_2 (m·s^{-1})	SSH (h)	ET_o (mm·day^{-1})
Bangalore	13°0'	77°37'	920	29.23	18.35	66.82	2.17	6.70	4.41
Coimbatore	11°0'	77°0'	403	31.85	21.48	66.54	2.30	6.56	4.79
Annamalai Nagar	11°24'	79°44'	4	31.82	22.79	75.28	1.84	7.85	4.57
Kovilpatti	9°12'	77°53'	80	35.52	22.08	64.14	1.95	7.12	5.32
Tirupati	13°27'	79°5'	395	33.44	22.65	60.48	2.38	6.41	5.29

T_{max} – maximum temperature; T_{min} – minimum temperature; RH_{mean} – mean relative humidity; U_2 – wind speed (measured at 2 m height); SSH – sunshine hours; ET_o – reference evapotranspiration

https://doi.org/10.17221/101/2023-RAE

area. The Hargreaves-Samani, Schendel, Tabari, and Brackamp and Wenner models were chosen for this study based on their input parameters.

Hargreaves-Samani equation. The Hargreaves-Samani model is one of the simplest empirical models for computing reference evapotranspiration. It is extensively used worldwide due to its superior accuracy and minimal need for climatic data. It is expressed as described in Equations (1–2):

$$ET_o = 0.408 \times 0.0023 R_a \left(T_{\text{mean}} + 17.8\right)\left(T_{\max} - T_{\min}\right)^{0.5} \quad (1)$$

where: ET_o – reference crop evapotranspiration; R_a – extraterrestrial radiation (MJ·m^{-2}·day^{-1}); T_{mean} – mean temperature (°C); $T_{\max}$ – maximum temperature (°C); $T_{\min}$ – minimum temperature (°C).

$$R_a = \frac{24(60)}{\pi} G_{sc} d_r \left[\omega_s \sin(\phi)\sin(\delta) + \cos(\phi)\cos(\delta)\sin(\omega_s)\right] \quad (2)$$

where: G_{sc} – solar constant (0.0820 MJ·m^{-2}·min^{-1}); d_r – inverse relative distance (Earth-Sun); ω_s – sunset hour angle (rad); ϕ – latitude (rad); δ – solar decimation (rad).

Schendel equation. This study uses the Schendel (1967) model, which is a commonly used humidity-based model that requires the mean temperature and relative humidity as inputs. The Schendel (S) model is expressed as shown in Equation (3):

$$ET_o = 16 \times \frac{T_{\text{mean}}}{RH_{\text{mean}}} \quad (3)$$

where: RH_{mean} – relative humidity (%).

Tabari equation. The Tabari and Talee model, which is a radiation-based model, requires only temperature data and solar radiation data as the inputs considered in this study and is expressed according to Equation (4):

$$ET_o = -0.478 + 0.156 R_s - 0.0112 T_{\max} + 0.0733 T_{\min} \quad (4)$$

where: R_s – solar radiation (MJ·m^{-2}·day^{-1}), which was calculated based on the Angstrom formula given in Equation (5):

$$R_s = \left(a_s + b_s \frac{n}{N}\right) R_a \quad (5)$$

where: a_s (intercept) and b_s (slope parameter) – regression coefficients with values of 0.25 and 0.5, respectively; n –actual duration of sunshine (h); N – possible maximum sunshine hours.

Brockamp-Wenner model. The Brockamp and Wenner model uses wind speed-based models and demands only the wind speed and temperature data as input parameters, and it is described according to Equation (6):

$$ET_o = 5.43 u^{0.456} \left(e_s - e_a\right) \quad (6)$$

where: u – wind speed at a height of 2 m (m·s^{-1}); e_s – saturated vapour pressure (kPa); e_a – actual vapour pressure (kPa).

Additionally, all the considered empirical models were regionally calibrated by minimising the sum of square errors by employing the generalised reduced gradient (GRG) non-linear method (Djaman et al. 2019), see Equation (7):

$$\text{SSE} = \sum_{i=1}^{N} \left(Y_{obs,i} - Y_{est,i}\right)^2 \quad (7)$$

where: SSE – sum of square errors; $Y_{obs,i}$ – ET_o calculated using the FAO Penman-Monteith (PM) model; $Y_{est,i}$ – estimated ET_o using the selected equations.

The implemented methodology flow chart is shown in Figure 2.

Artificial neural network

Artificial neural networks (ANNs) are inspired by the complex nervous system of the human brain. ANNs are designed to learn from input datasets to establish non-linear relationships between the inputs and the desired outputs. The fundamental structure of an ANN consists of input, hidden, and output layers interconnected with nodes and activation functions (Dawson and Wilby 1998). The neurons within the network perform numerical computations to calculate the weights and biases. Crucial hyperparameters, such as the number of epochs, number of hidden layers, number of nodes, and type of training algorithm, significantly impact the performance of a neural network.

The network operates in two phases: during the feed-forward phase, input signals propagate in the forward direction, activating functions to produce output values. In backward propagation, errors are used to adjust the layer weights. The error is essen-

https://doi.org/10.17221/101/2023-RAE

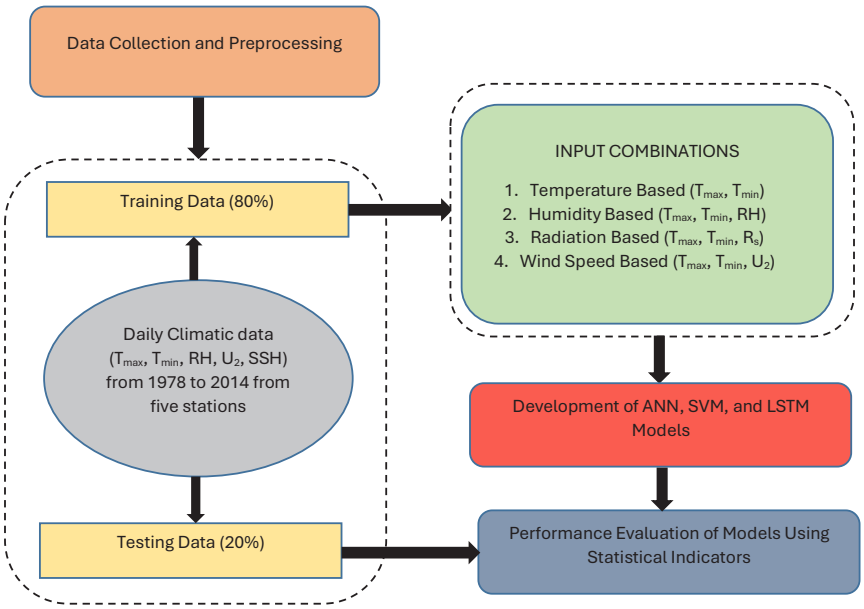

Figure 2. Flow chart of the methodology implemented in this study

T_{max} – maximum temperature; T_{min} – minimum temperature; RH – relative humidity; R_s – solar radiation; U_2 – wind speed (measured at 2 m height); SSH – sunshine hours

tially the difference between the observed values and the model output. The mathematical formulation of the ANN is presented in the following Equation (8):

$$y_k = g_2\left[\sum_{j=1}^{M} W_{kj}\, g_1\left(\sum_{i=1}^{N} W_{ji} x_i + W_{jo}\right) + W_{ko}\right] \quad (8)$$

where: y_k – output at node k; g_1 – activation function for the hidden layers; g_2 – activation function for the output layers; M – number of neurons in the hidden layers; W_{kj} – weight between the hidden node and the output node; N – number of neurons in the input layers; W_{ji} – weight between the input node and the hidden node; x_i – value at node i; W_{jo} – bias of the j^{th} neuron in the hidden layer; W_{ko} – k^{th} neuron in the output layer.

Support vector machine (SVM)

Vapnik proposed a novel machine learning approach for decision support systems to address classification and regression problems. The main idea underlying SVM is to perform linear regression in this space after non-linearly mapping the data x into a high-dimensional feature space (Boser et al. 1992). In general, SVM is applied to estimate a regression function based on the given set of data points, as described by Equation (9):

$$\{(x_i, y_i)\}_i^n \quad (9)$$

where: x_i – input of the ET_o derived from the algorithm; y_i – desired output values of the ET_o; n – number of training samples.

The inputs are initially non-linearly mapped into a high-dimensional feature space where they are linearly connected with the outputs to perform non-linear regression procedures. To formalise the SVM, the following linear computation function was used, as shown in Equation (10):

$$f(x) = \omega \times \phi(x) + b \quad (10)$$

where: ω – weight vector; ϕ (x) – mapping function that transforms the input vectors into a high-dimensional space; b – constant bias value.

By minimising the regularised risk function, the values of coefficients ω and b can be calculated according to Equations (11–12):

$$R_{reg}(f) = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon\left[f(x_i), y_i\right] + \frac{1}{2}\|\omega\|^2 \quad (11)$$

$$L_\varepsilon\left[f(x) - y\right] = \begin{cases} |f(x) - y| - \varepsilon & \text{for } |f(x) - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

where: C – positive constant that denotes the penalty parameter; ε – tube size; L_ε [f (x_i), y_i] – ε-intensive loss function, which calculates the empirical error; $1/2\|\omega\|^2$ – regularisation term.

Using the Lagrangian and optimum conditions, the following Equation (13) yields a non-linear regression function:

https://doi.org/10.17221/101/2023-RAE

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)k(x_i,x)+b \quad (13)$$

where: a_i, a_i^* – Lagrange multipliers; k (x_i, x) – kernel function.

With the utilisation of the Karush-Kuhn-Tucker conditions, only a restricted number of coefficients will not be zero between a_i and a_i^*. One may call the related data points the support vectors. In the D-dimensional feature space, the kernel function k (x_i, x) describes the inner product, see Equation (14):

$$k(x,y)=\sum_{i=1}^{D}\varphi_j(x)\varphi_j(y) \quad (14)$$

where: D – dimension feature space; ϕ_j – mapping function.

This demonstrates that a dot product in a certain feature space corresponds to any symmetric kernel function k meeting Mercer's condition. Three factors that are effective in the SVM performance are the ε error term, C penalty parameter, and kernel function parameter γ. This study adopts the medium radial basis function (RBF) kernel to perform SVM for predicting evapotranspiration due to its robustness. The RBF kernel function can be expressed by Equation (15):

$$k(x_i,x)=\exp\left(-\frac{1}{2\sigma_2}\|x-x_i\|^2\right) \quad (15)$$

where: σ – variance.

Further details on the SVM can be found in Vapnik (1995).

Long short-term memory

Long short-term memory (LSTM) is a type of recurrent neural network (RNN) proposed by Hochreiter and Schmidhuber (1997). LSTM has a similar architecture to an RNN with input, hidden, and output layers. However, the LSTM replaces the basic unit of a regular RNN with a memory block (Graves et al. 2013). Long-term dependencies can be modelled by LSTM employing a memory unit termed the cell state.

The LSTM memory block consists of three gates – the forget, input, and output gates (Yuan 2018). The designed network for this study consisted of the following parts: input time series data were fed into the 'sequence input layer' of the network followed by an 'LSTM layer', which further connects with the 'fully connected layer' of several hidden neurons followed by a 'regression output layer. The network architecture can be further improved based on the problem dependencies by including an additional LSTM layer and dropout layer to prevent the model from being overfit (Roy 2021) (Figure 3). Thus, the features of LSTM can reduce the effects of the vanishing gradient problem.

The following Equations (16–17) represent the hidden and cell states at any time step t:

$$h_t=o_t\odot\sigma_c(C_t) \quad (16)$$

$$C_t=f_t\odot C_{t-1}+i_t\odot g_t \quad (17)$$

where: h_t – output of the hidden layer at time t; o_t – outcomes of the output gate at timespan t; ⊙ – element-wise multiplication of the vectors (Hadamard product); σ_c – state activation function [generally the hyperbolic tangent function (tanh)]; C_t – cell output state at time t; f_t – outcomes of the forget gate at timespan t; i_t – outcomes of the input gate at timespan t; g_t – outcomes of the candidate state at timespan t.

In LSTM, the first step is to examine whether the information from the cell state is remembered or forgotten. The calculation is expressed by Equation (18):

$$f_t=\sigma\left[W_f\times(h_{t-1},x_t)+b_f\right] \quad (18)$$

where: W_f – weights connecting h_{t-1} and x_t to the forget gate; x_t – input at time t; h_{t-1} – output of the hidden layer at time t–1; b_f – forget gate bias term.

The second step controls which data must be stored in the cell state. The tanh layer creates the new candidate value $\tilde{C}_t$ after a sigmoid layer detects the values that need to be updated. The expression for the calculation is described by Equations (19–20):

$$i_t=\sigma\left[W_i\times(h_{t-1},x_t)+b_i\right] \quad (19)$$

$$\tilde{C}_t=\tanh\left[W_C\times(h_{t-1},x_t)+b_C\right] \quad (20)$$

The next step is updating the previous cell state C_{t-1} by multiplying C_{t-1} by f_t and adding $i_t\times\tilde{C}_t$ to obtain the new value, see Equation (21):

https://doi.org/10.17221/101/2023-RAE

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (21)$$

The cell state is passed via the tanh function and multiplied by the result of the sigmoid gate, and then a sigmoid layer is run to obtain the output according to Equations (22–23):

$$o_t = \sigma\left[W_o \times \left(h_{t-1}, x_t\right) + b_o\right] \quad (22)$$

$$h_t = o_t \times \tanh\left(C_t\right) \quad (23)$$

where: x_t – input at time t; h_{t-1} – output of the hidden layer at time t–1; h_t – output of the hidden layer at time t; C_t – cell output state at time t; C_{t-1} – cell output state at time t–1; $\tilde{C}_t$ – outcome of the input gate at timespan t; f_t – outcomes of the forget gate at timespan t; o_t – outcomes of the output gate at timespan t; W_c – weights connecting h_{t-1} and x_t to the cell state; W_f – weights connecting h_{t-1} and x_t to the forget gate; W_i – weights connecting h_{t-1} and x_t to the input gate; W_o – weights connecting h_{t-1} and x_t to the output gate; b_f – forget gate bias term; b_i – forget gate bias term; b_C – cell gate bias term; b_o – forget gate bias term; tanh – hyperbolic tangent function $e^x - e^{-x}/e^x + e^{-x}$.

In LSTM networks for the ET_o estimation, cell states capture long-term dependencies and patterns. The cell state acts as a conveyor belt for information and is updated and controlled by gates. This allows the LSTM to capture temporal dependencies and non-linear dynamics in meteorological data, improving the accuracy of the ET_o predictions.

Data management scenarios

The selection of relevant input variables is critical for developing machine learning models since the variables provide essential information. This study compared the performances of the LSTM, ANN, SVM, and empirical models with the observed ET_o values. To evaluate all the machine learning and empirical models under limited data conditions, four different combinations were considered for developing the machine learning models in this study, and their respective input parameters are tabulated in Table 2. In addition, the developed machine learning models were compared with the corresponding empirical models according to their input variables.

Normalisation of data

Before applying the machine learning models to the data, it is wise to normalise all the input and output values to reduce the convergence problems and disparities related to the different units. In this study, min-max normalisation is implemented

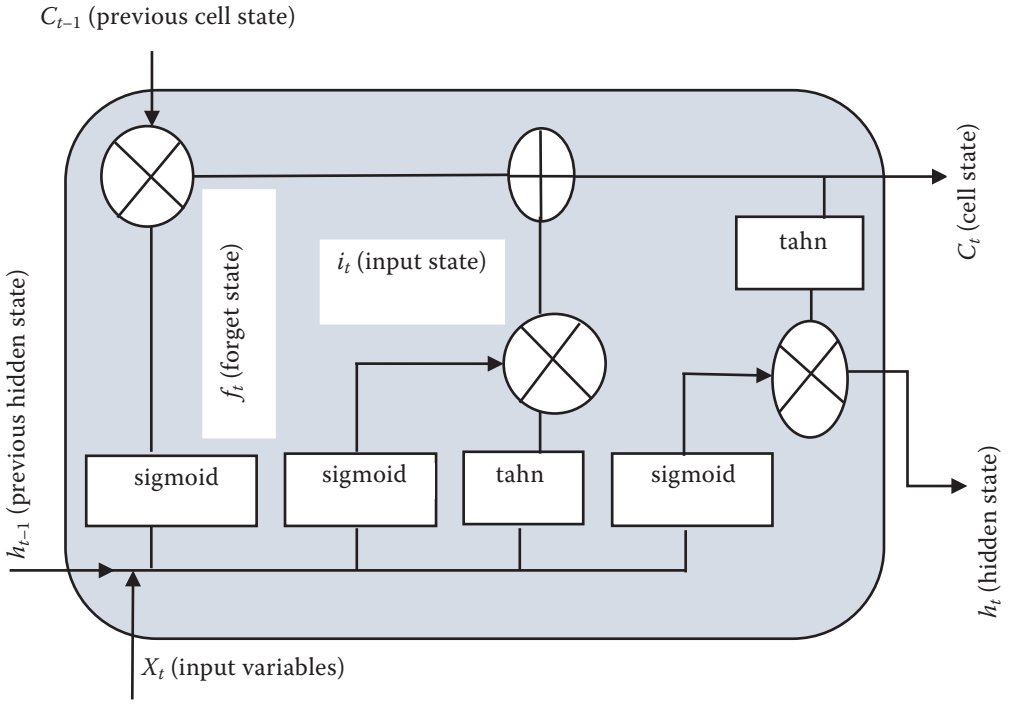

Figure 3. Block diagram of the long short-term memory (LSTM) network model

https://doi.org/10.17221/101/2023-RAE

Table 2. Machine learning models and their respective input parameters

Model	Scenarios	Input parameters
Temperature based models	LSTM1 ANN1 SVM1	maximum temperature (T_{max}), minimum temperature (T_{min}), mean temperature (T_{mean}), extraterrestrial radiation (R_a)
Humidity-based models	LSTM2 ANN2 SVM2	mean temperature (T_{mean}), mean relative humidity (RH_{mean})
Radiation-based models	LSTM3 ANN3 SVM3	maximum temperature (T_{max}), minimum temperature (T_{min}), mean temperature (T_{mean}), solar radiation (R_s)
Wind speed-based models	LSTM4 ANN4 SVM4	maximum temperature (T_{max}), minimum temperature (T_{min}), wind speed (U_2)

to scale the values of the input parameters into a typical range from zero to one before inputting them into the model via the following formula:

$$\text{norm}(X) = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (24)$$

where: norm (X) – normalised input data; X – raw input data; max (X) – maximum input data; min (X) – minimum of the input data.

Local calibration

The performance of empirical ET_o models often varies among stations within the same climate zone due to the differences in the data quality, time periods used, and uncertainties in the coefficient values for the ET_o estimation. To address this, the coefficients of all the empirical models were locally calibrated using the generalised reduced gradient (GRG) non-linear approach to minimise the sum of square errors.

Performance assessment of the calibrated models

The following performance metrics were employed to evaluate the trained models against the measured Lysimeter ET_o values.

Coefficient of determination (R^2), Equation (25):

$$R^2 = \left(\frac{\sum_{i=1}^{N} \left(Y_{est,i} - \overline{Y_{est}}\right)\left(Y_{obs,i} - \overline{Y_{obs}}\right)}{\sqrt{\sum_{i=1}^{N} \left(Y_{est,i} - \overline{Y_{est}}\right)^2 \sum_{i=1}^{N} \left(Y_{obs,i} - \overline{Y_{obs}}\right)^2}} \right)^2 \quad (25)$$

The Nash Sutcliffe Efficiency (NSE), Equation (26):

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{N} \left(Y_{obs,i} - Y_{est,i}\right)^2}{\sum_{i=1}^{N} \left(Y_{obs,i} - \overline{Y_{obs}}\right)^2} \quad (26)$$

Root mean square error (RMSE), Equation (27):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(Y_{obs,i} - Y_{est,i}\right)^2}{N}} \quad (27)$$

where: N – amount of the total data; $Y_{est,i}$ – estimated ET_o using the selected equations; $\overline{Y_{est}}$ – mean of the estimated ET_o values; $Y_{obs,i}$ – ET_o calculated using the FAO PM model; $\overline{Y_{obs}}$ – mean of the observed ET_o values.

To create a perfect model, smaller *RMSE* values and higher R^2 and *NSE* values are preferred.

RESULTS AND DISCUSSION

Local calibration of the empirical models. The calibrated coefficients are shown in Table 3. For instance, the coefficients of the Hargreaves Samani model were calibrated within the ranges of 0.0026 to 0.0099, 13.468–9.147, and 0.336–0.534, compared to the original coefficients of 0.0023, 17.8, and 0.5, respectively. Following calibration, the values for the Bangalore station were aligned with those from a study by Niranjan and Nandagiri (2021). Similarly, the coefficients of the Schendel model were adjusted for various stations. The modified coefficients for all the considered models, along with the original coefficients, were tabulated for the ET_o estimation. After calibration, the R^2 values increased, and the RMSE values decreased across all the stations. Addition-

https://doi.org/10.17221/101/2023-RAE

Table 3. Original and calibrated coefficients of the considered empirical models

Model	Original coefficients	Modified coefficients				
		Annamalai Nagar	Bangalore	Coimbatore	Kovilpatti	Tirupati
Hargreaves Samani	0.0023	0.0055	0.0026	0.0059	0.0099	0.0083
	17.80	–2.88	9.15	–5.31	–13.47	–7.99
	0.50	0.40	0.53	0.34	0.34	0.34
Schendel	16.00	12.46	11.60	10.11	11.35	10.54
Tabari	–0.478	–6.582	–3.360	–5.209	–8.724	–7.371
	0.156	0.138	0.128	0.102	0.122	0.129
	0.0112	–0.2230	–0.1560	–0.1920	–0.2500	–0.2390
	0.0733	0.0559	0.0435	0.0934	0.1250	0.1040
Brockamp and Wenner	5.43	1.88	2.59	2.57	1.52	2.23
	0.456	0.896	0.284	0.263	0.251	0.160

ally, in this study, the performances of the developed machine learning models were compared solely using locally calibrated empirical equations.

Temperature-based models. The locally calibrated Hargreaves Samani model outperformed the original model. This finding contrasts with the findings of Niranjan and Nandagiri (2021), who reported significant improvement in the estimation of the ET_o across all the Bangalore zones using locally calibrated parameters. The performance of the temperature-based models is depicted in Figure 4. The figure shows that the LSTM1 model outperformed the other machine learning and empirical models that were considered. The highest coefficient of determination (0.88) was observed at the Kovilpatti station, while the lowest RMSE value (0.705 mm·day^{-1}) occurred at the Bangalore station. Except for Tirupati, the SVM1 model underperformed compared to the other models at all the stations. In addition to the LSTM1 model, the locally calibrated Hargreaves Samani model performed well at all the locations. Figure 4 shows that the mean R^2 and RMSE values of the LSTM1 model for the temperature-based approach are 0.775 mm·day^{-1} and 0.868 mm·day^{-1}, respectively, while for the other models, they are greater than 0.982 mm·day^{-1}. The NSE values exceeded 0.64 at all the stations for the LSTM1 model, whereas for the other considered models, they were less than 0.45.

Several studies have shown that incorporating temperature data and employing machine learning models can increase the accuracy of the ET_o estimation (Fan et al. 2018; Ferreira et al. 2019; Chia et al. 2020). However, in this study, we found that the accuracy of estimating the ET_o was not improved by the temperature-based ANN and SVM models. On the other hand, the findings of Wen et al. (2015) indicate that SVM-based machine learning models outperform empirical models. Additionally, Ferreira et al. (2019) discovered that, across the entire country of Brazil, temperature-based DNN and SVM models provided more accurate predictions than the Hargreaves-Samani model. Notably, the performance accuracy of machine learning models varies significantly by region (Sun et al. 2020).

Humidity-based models. This study compared the performances of humidity-based LSTM2, ANN2, and SVM2 models with that of the locally calibrated Schendel model. The results indicated that the LSTM2 model outperformed the other models, exhibiting a higher R^2 value (0.64) and a lower RMSE value (< 0.92 mm·day^{-1}). In contrast, the locally calibrated Schendel model showed less accurate performance across all the weather stations, with an R^2 value of 0.72 and an RMSE value of 1.137 mm·day^{-1}. On average, the R^2 values of the calibrated Schendel, ANN2, and SVM2 models were 0.058, 0.066, and 0.069 lower, respectively, than that of the LSTM2 model. The NSE values of the LSTM2 model for the Annamalai Nagar, Bangalore, Coimbatore, Kovilpatti, and Tirupati stations were 0.627, 0.287, 0.614, 0.738, and 0.812, respectively (Figure 5).

It should be noted that the humidity-based models, despite incorporating relative humidity as an additional input, exhibited lower performance than the temperature-based models. This phenomenon is attributed to the absence of extraterrestrial radiation as an input in the estimation of ET_o in humidity-based empirical models, whereas temperature-

https://doi.org/10.17221/101/2023-RAE

based equations encompass this factor. However, the inclusion of relative humidity as an input in temperature-based deep learning models and conventional machine learning models significantly enhanced the precision of the ET_o estimation.

According to Zhu et al. (2020), the presence of humidity parameters led to the random forest and SVM models outperforming the LSTM and DNN models. Additionally, the results of Kiafar et al. (2017) corroborate that the estimation accuracy of the Schendel model was inferior to that of the Hargreaves-Samani model. The performance of the proposed humidity-based machine learning models surpassed that of the empirical model by a substantial margin, with the LSTM model demonstrating superior performance compared to the ANN and SVM models among the machine learning models.

Radiation-based models. At all the weather stations studied, the models based on radiation consistently outperformed those based on temperature and humidity. When radiation parameters are incorporated, the machine learning models with temperature variables demonstrate more accurate ET_o forecasting. The findings indicate that the deep learning model LSTM3 outperforms the other models, displaying the highest R^2 and NSE values and the lowest RMSE values. The R^2 value exceeds 0.9, the NSE value surpasses 0.77, and the RMSE value is less than 0.84 mm·day^{-1} across all the weather stations, with the exception of the Coimbatore station (R^2 = 0.82, NSE = 0.53, and RMSE = 0.92 mm·day^{-1}). On aver-

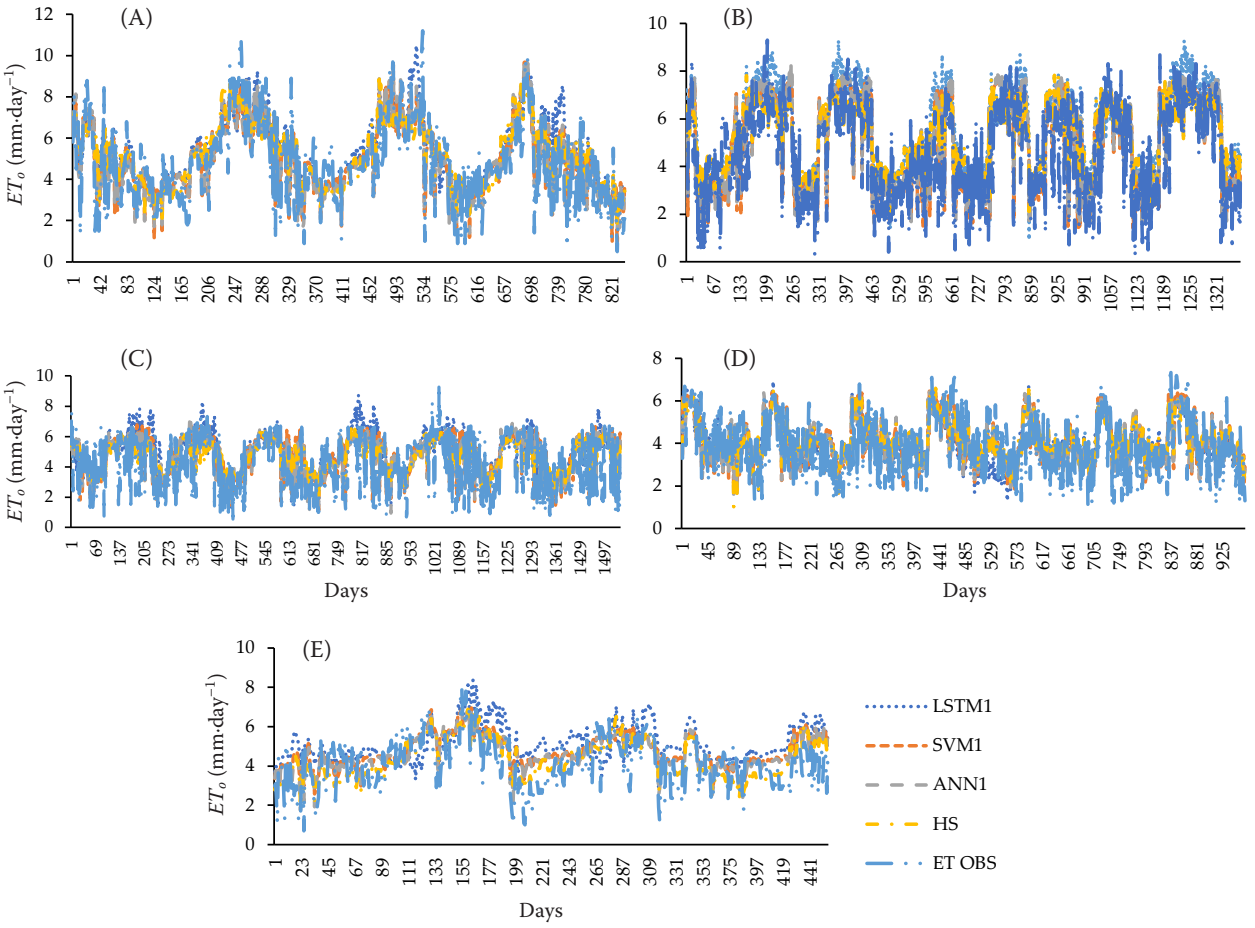

Figure 4. Comparison of the observed reference evapotranspiration (ET_o) and the temperature-based models for the testing period at (A) Tirupati station, (B) Kovilpatti station, (C) Coimbatore station, (D) Bangalore station, and (E) Annamalai Nagar station

LSTM – long short-term memory neural network; SVM – support vector regression; ANN – artificial neural network; HS – Hargreaves-Samani model; ET OBS – observed evapotranspiration

https://doi.org/10.17221/101/2023-RAE

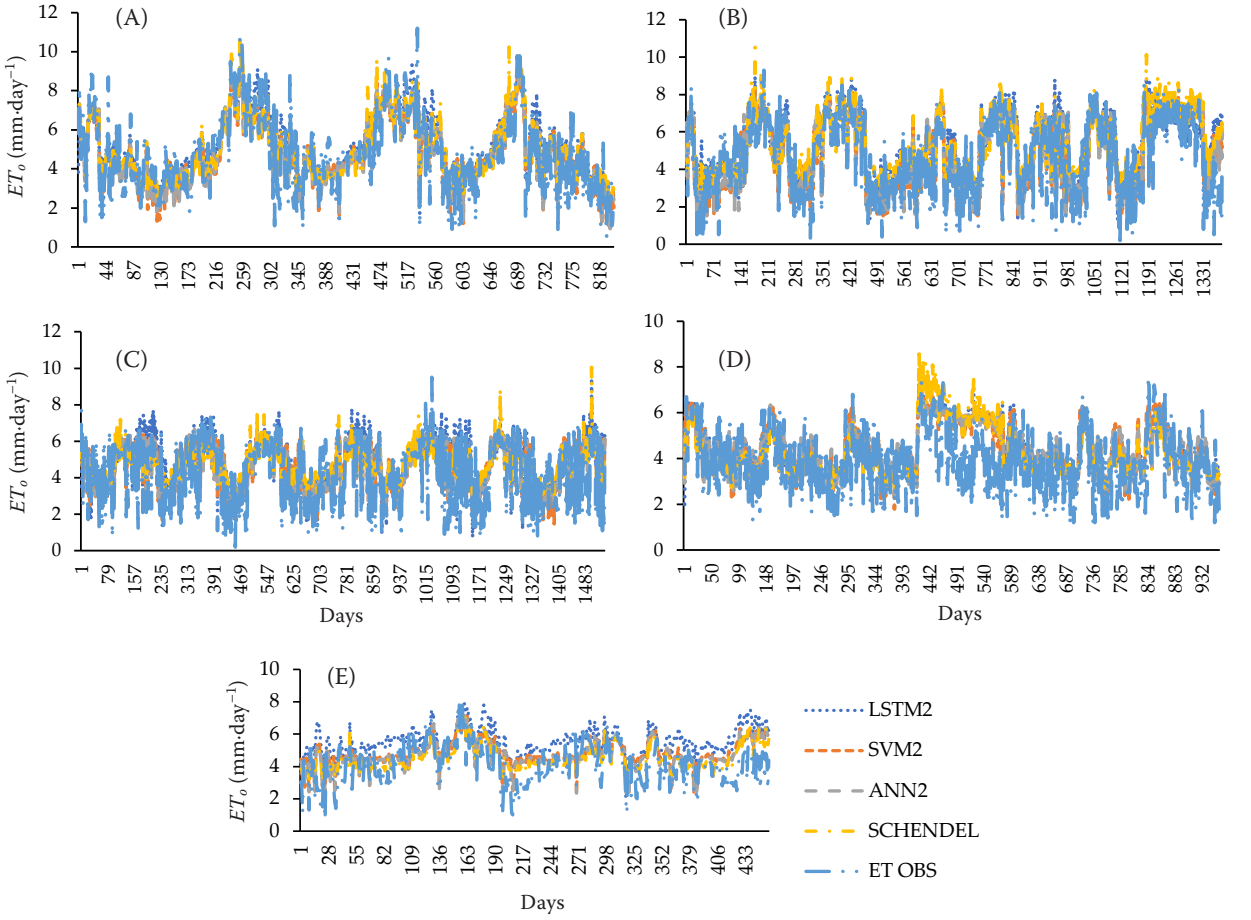

Figure 5. Comparison of the observed reference evapotranspiration (ET_o) and humidity-based models for the testing period at (A) Tirupati station, (B) Kovilpatti station, (C) Coimbatore station, (D) Bangalore station, and (E) Annamalai Nagar station

LSTM – long short-term memory neural network; SVM – support vector regression; ANN – artificial neural network; HS – Hargreaves-Samani model; ET OBS – observed evapotranspiration

age, the RMSEs of the radiation-based models were 0.2 mm·day^{-1} and 0.15 mm·day^{-1} lower than those of the temperature-based and humidity-based models, respectively (Figure 6).

According to experts, temperature and radiation parameters contribute to 80% of the variation in the reference evapotranspiration (Irmak et al. 2003; Samani 2003; Tabari and Talaee 2011). These results align with those of Sharma et al. (2022), who found that the input combination of temperature and solar radiation produced the best performance at the Ludhiana and Amritsar stations in India. Sharma et al. considered six different input combinations to develop two hybrid deep learning networks. These conclusions also correspond with the findings of Kaya et al. (2021), who determined that the combination of temperature and solar radiation improved the performance using a multilayer perceptron. Moreover, they confirmed that solar radiation is the most influential parameter in estimating the ET_o using soft computing approaches.

Wind speed-based models. Based on Figure 7, it can be concluded that among all the combinations of input parameters, the machine learning models based on wind speed outperformed the others across all the stations under consideration. Specifically, the deep learning-based LSTM4 model demonstrated superior performance, yielding R^2 > 0.88, RMSE < 0.6 mm·day^{-1}, and NSE > 0.8. In contrast, the empirical model exhibited the lowest performance, with an RMSE exceeding 0.925 mm·day^{-1} relative to the machine learning algorithms. The combination of the wind speed and temperature consistently delivered favourable results across all the stations,

https://doi.org/10.17221/101/2023-RAE

unlike the other input combinations. Although the coefficient of determination of the empirical model surpassed that of the other empirical models, it is essential to locally calibrate the wind speed models (Sharafi and Mohammadi Ghaleni 2021).

Maroufpoor et al. (2020) suggested that the combination of temperature and wind speed yields more accurate results across the various climates in Iran's 31 provinces. Moreover, Muhammad et al. (2019) observed that models based on wind speed were more dependable than those based on radiation. In recent years, there has been a growing interest in using machine learning models such as ANNs, RFs, long short-term memory (LSTM) models, and support vector machines (SVMs) to calculate the ET_o with a minimal number of meteorological input factors. These machine learning models outperform traditional empirical equations. The superior performance of the LSTM model can likely be attributed to its internal structure, which enables it to retain and utilise past information.

Vapour removal is heavily dependent on wind and air turbulence, which cause large volumes of air to flow across evaporating surfaces. As water evaporates, the air above the surface becomes saturated with water vapour. Without the continuous exchange of this air with drier air, the rates of water vapour removal and evapotranspiration will decrease (Allen et al. 1998). For instance, in hot and dry weather conditions, the need for evapotranspiration is sig-

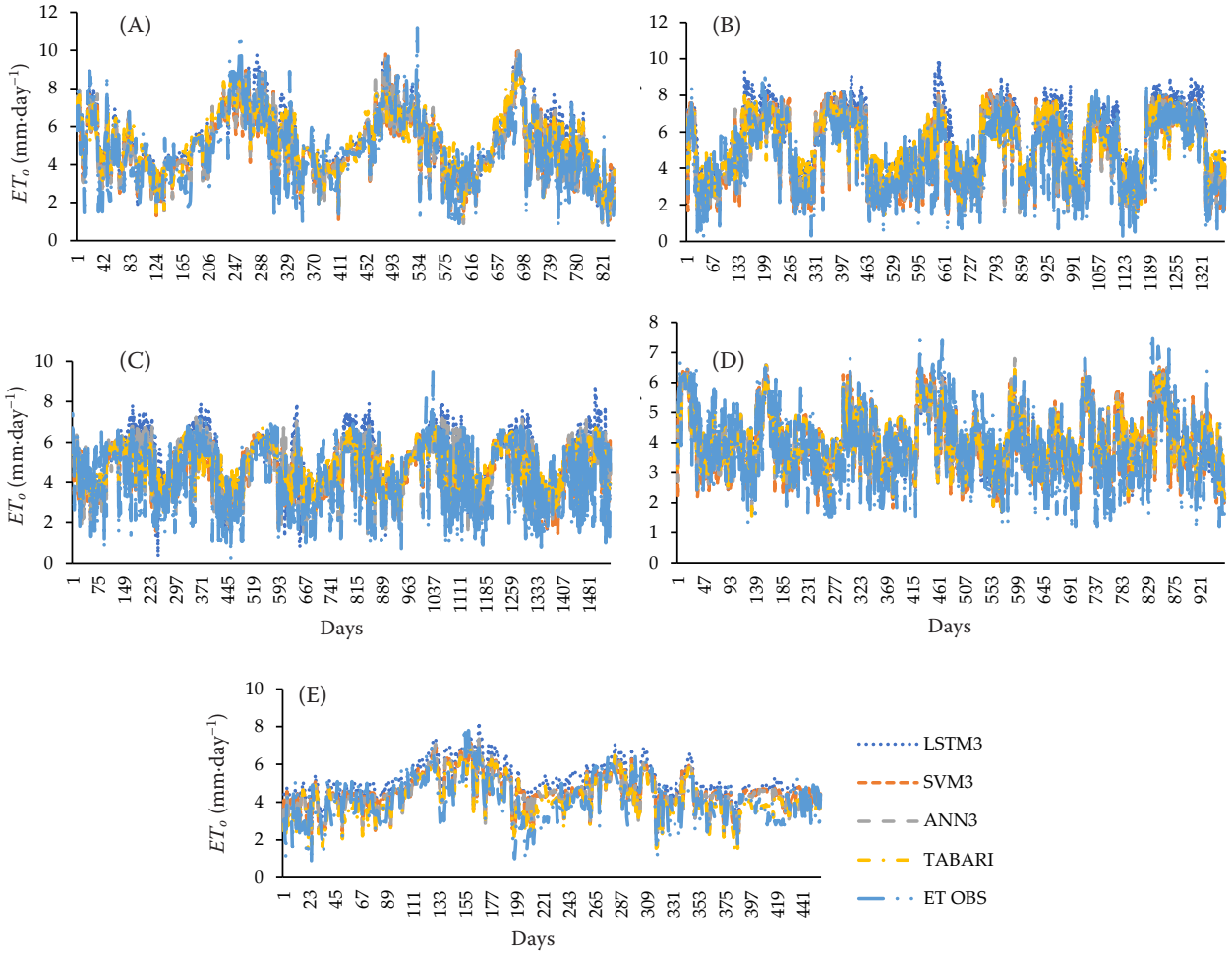

Figure 6. Comparison of the observed reference evapotranspiration (ET_o) and radiation-based models for the testing period at (A) Tirupati station, (B) Kovilpatti station, (C) Coimbatore station, (D) Bangalore station, and (E) Annamalai Nagar station

LSTM – long short-term memory neural network; SVM – support vector regression; ANN – artificial neural network; HS – Hargreaves-Samani model; ET OBS – observed evapotranspiration

https://doi.org/10.17221/101/2023-RAE

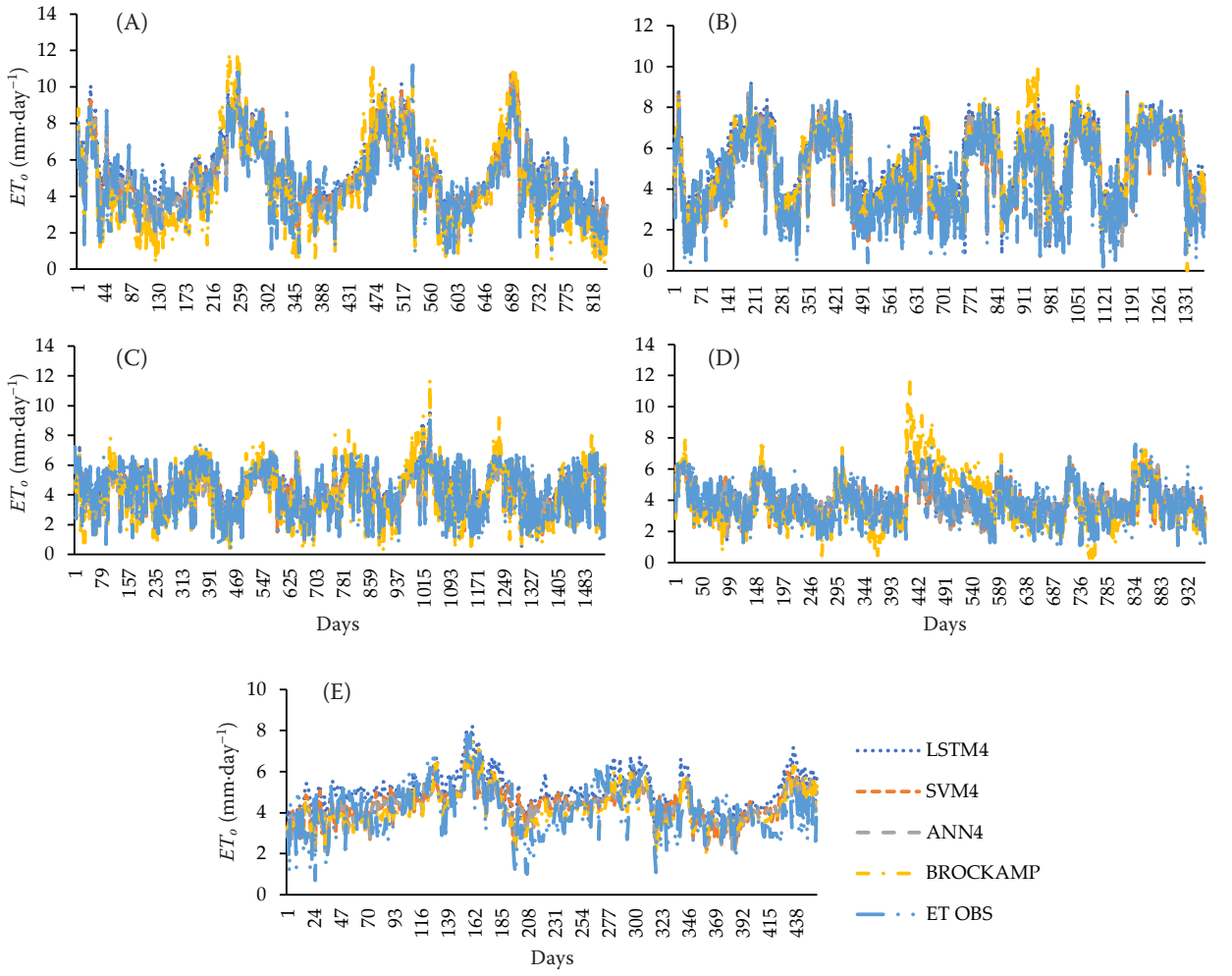

Figure 7. Comparison of the observed reference evapotranspiration (ET_o) and the wind speed-based models for the testing period at (A) Tirupati station, (B) Kovilpatti station, (C) Coimbatore station, (D) Bangalore station, and (E) Annamalai Nagar station

LSTM – long short-term memory neural network; SVM – support vector regression; ANN – artificial neural network; HS – Hargreaves-Samani model; ET OBS – observed evapotranspiration

nificant due to the low moisture content in the air and the abundance of energy available in the form of direct solar radiation and latent heat.

The findings of this study contradict those of previous research (Gonzalez del Cerro et al. 2021; Niranjan and Nandagiri 2021), which concluded that wind speed has the least influence on the estimation of the ET_o. A comparison of the performances of the developed machine learning models and empirical models is presented in Table S1 in the electronic supplementary material (ESM). Additionally, the performance of the developed LSTM models was assessed using box and whisker plots, as illustrated in Figure 8.

CONCLUSION

This study focuses on the application of machine learning models, including LSTM, ANN, and SVM, to estimate the ET_o using a limited set of input parameters specific to tropical savannah climatic regions.

In this study, the performances of the LSTM, ANN, and SVM models were compared with traditional empirical models, including Hargreaves-Samani, Schendel, Tabari, and Brockamp and Wenner. Notably, the performance of the developed machine learning models was superior to that of the traditional empirical formulas across all the considered locations. Specifically, the LSTM model exhibited

https://doi.org/10.17221/101/2023-RAE

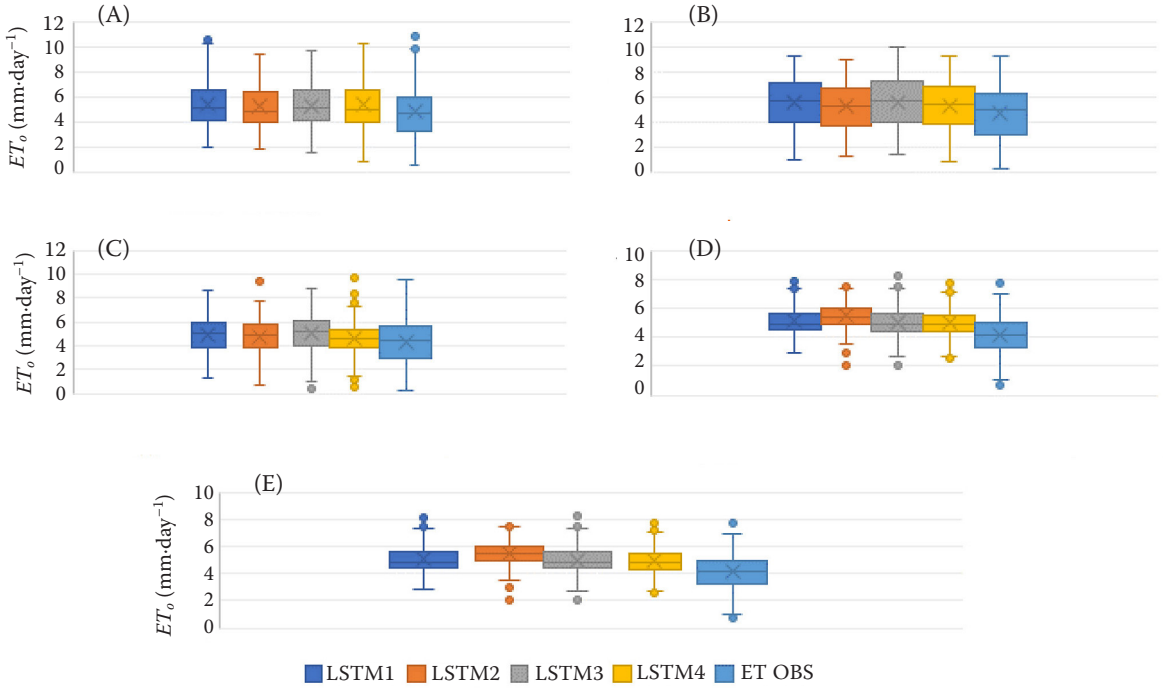

Figure 8. Box and whisker plot for the LSTM (long short-term memory) models in all the weather stations, namely (A) Tirupati station, (B) Kovilpatti station, (C) Coimbatore station, (D) Bangalore station, and (E) Annamalai Nagar station

LSTM – long short-term memory neural network; SVM – support vector regression; ANN – artificial neural network; HS – Hargreaves-Samani model; ET OBS – observed evapotranspiration

the highest performance when utilising a combination of maximum temperature, minimum temperature, and wind speed, with R^2 values exceeding 0.75 and RMSEs below 0.63 mm·day^{-1} at all the stations. Furthermore, the combination of temperature and solar radiation (LSTM3) performed well in certain locations, while in others, the temperature and wind speed combination (LSTM4) yielded better results. Conversely, the performance of the temperature-based input combination was relatively lower, with RMSE values exceeding 0.9 mm·day^{-1}.

The study also highlighted the potential of the developed machine learning models in providing accurate ET_o estimations, particularly under data-limited conditions. The findings indicated that the superior performance of the LSTM model could be attributed to its ability to effectively retain or forget information, surpassing the capabilities of SVMs and ANNs. Ultimately, the implementation of the developed machine learning models could provide valuable support to researchers, water management agencies, and irrigators by enabling accurate ET_o estimations in tropical savannah climatic regions, even when meteorological data were limited.

REFERENCES

Abtew W. (1996): Evapotranspiration measurements and modeling for three wetland systems in south Florida. Water Resources Bulletin, 32: 465–473.

Allen R.G., Pereira L.S., Raes D., Smith M. (1998): Crop Evapotranspiration – Guidelines for Computing Crop Water Requirements – FAO Irrigation and Drainage Paper 56. Rome, FAO: 50.

Bellido-Jiménez J.A., Estévez J., García-Marín A.P. (2021): New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semiarid region of Spain. Agricultural Water Management, 245: 106558.

Boser E., Guyon I.M., Vapnik N. (1992): Training Algorithm Margin for Optimal Classifiers. In: Haussler D. (ed): Proceedings of the 5th Annual Workshop on Computational Learning Theory, July 27–29, 1992, Pittsburgh, USA: 144–152.

https://doi.org/10.17221/101/2023-RAE

Brockamp B., Wenner H. (1963): Verdunstungsmessungen auf den Steiner See bei Münster. Dt Gewässerkundl Mitt, 7: 149–154. (in German)

Chia M.Y., Huang Y.F., Koo C.H. (2020): Support vector machine enhanced empirical reference evapotranspiration estimation with limited meteorological parameters. Computers and Electronics in Agriculture, 175: 105577.

Dawson C.W., Wilby R. (1998): An artificial neural network approach to rainfall-runoff modeling. Hydrological Sciences Journal, 43: 47–66.

Djaman K., ONeill M., Diop L., Bodian A., Allen S., Koudahe K., Lombard K. (2019): Evaluation of the Penman-Monteith and other 34 reference evapotranspiration equations under limited data in a semiarid dry climate. Theoretical and Applied Climatology, 137: 729–743.

Fan J., Yue W., Wu L., Zhang F., Cai H., Wang X., Lu X., Xiang Y. (2018a): Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. Agricultural and Forest Meteorology, 263: 225–241.

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

Ferreira L.B., da Cunha F.F. (2020): New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agricultural Water Management, 234: 106113.

Gonzalez del Cerro R.T., Subathra M.S.P., Manoj Kumar N., Verrastro S., George S.T. (2021): Modeling the daily reference evapotranspiration in semi-arid region of South India: A case study comparing ANFIS and empirical models. Information Processing in Agriculture, 8: 173–184.

Goyal P., Kumar S., Sharda R. (2023): A review of the Artificial Intelligence (AI) based techniques for estimating reference evapotranspiration: Current trends and future perspectives. Computers and Electronics in Agriculture, 209: 107836.

Graves A., Mohamed A., Hinton G. (2013): Speech recognition with deep recurrent neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 26–31, 2013, Vancouver, Canada: 6645–6649.

Hargreaves G.H., Samani Z.A. (1985): Reference Crop Evapotranspiration from Temperature. Applied Engineering in Agriculture, 1: 96–99.

Hochreiter S., Schmidhuber J. (1997): Long short-term memory. Neural Computation, 9: 1735–1780.

Hu C., Wu Q., Li H., Jian S., Li N., Lou Z. (2018): Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. Water, 10: 1–16.

Irmak S., Irmak A., Allen R.G., Jones J.W. (2003): Solar and net radiation-based equations to estimate reference evapotranspiration in humid climates. Journal of Irrigation and Drainage Engineering, 129: 336–347.

Karuppanan S., Ramasamy S., Narayanan B.L., Anuthaman S.N. (2022): An effective alternate empirical model to estimate reference evapotranspiration for tropical climatic region. Journal of Environmental Protection and Ecology, 23: 2737–2746.

Kaya Y.Z., Zelenakova M., Üneş F., Demirci M., Hlavata H., Mesaros P. (2021): Estimation of daily evapotranspiration in Košice City (Slovakia) using several soft computing techniques. Theoretical and Applied Climatology, 144: 287–298.

Kiafar H., Babazadeh H., Marti P., Kisi O., Landeras G., Karimi S., et al. (2017): Evaluating the generalizability of GEP models for estimating reference evapotranspiration in distant humid and arid locations. Theoretical and Applied Climatology, 130: 377–389.

Li Q., Wang Z., Shangguan W., Li L., Yao Y., Yu F. (2021): Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning. Journal of Hydrology, 600: 126698.

Mahringer W. (1970): Verdunstungsstudien am Neusiedler See. Archives for Meteorology Geophysics and Bioclimatology Series A Meteorology and Atmospheric Physics, 18: 1–20. (in German)

Maroufpoor S., Bozorg-Haddad O., Maroufpoor E. (2020): Reference evapotranspiration estimating based on optimal input combination and hybrid artificial intelligent model: Hybridization of artificial neural network with grey wolf optimizer algorithm. Journal of Hydrology, 588: 125060.

Muhammad M.K.I., Nashwan M.S., Shahid S., bin Ismail T., Song Y.H., Chung E.S. (2019): Evaluation of empirical reference evapotranspiration models using compromise programming: A case study of Peninsular Malaysia. Sustainability, 11: 4267.

Ni L., Wang D., Singh V.P., Wu J., Wang Y., Tao Y., Zhang J. (2019): Streamflow and rainfall forecasting by two long short-term memory-based models. Journal of Hydrology, 583: 124296.

Niranjan S., Nandagiri L. (2021): Effect of local calibration on the performance of the Hargreaves reference crop evapotranspiration equation. Journal of Water and Climate Change, 12: 2654–2673.

Ravazzani G., Corbari C., Morella S., Gianol P., Mancini M. (2012): Modified Hargreaves-Samani equation for the assessment of reference evapotranspiration in alpine river

https://doi.org/10.17221/101/2023-RAE

basins. Journal of Irrigation and Drainage Engineering, 138: 592–599.

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

Roy D.K. (2021): Long short-term memory networks to predict one-step ahead reference evapotranspiration in a subtropical climatic zone. Environmental Processes, 8: 911–941.

Saggi M.K., Jain S. (2019): Reference evapotranspiration estimation and modeling of the Punjab Northern India using deep learning. Computers and Electronics in Agriculture, 156: 387–398.

Salam R., Islam A.R.M.T. (2020): Potential of RT, bagging and RS ensemble learning algorithms for reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh. Journal of Hydrology, 590: 125241.

Samani Z. (2003): Estimating solar radiation and evapotranspiration using minimum climatological data. Journal of Irrigation and Drainage Engineering, 126: 265–267.

Sattari M.T., Apaydin H., Band S.S., Mosavi A., Prasad R. (2021): Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. Hydrology and Earth System Sciences, 25: 603–618.

Schendel U. (1967): Vegetationswasserverbrauch und-wasserbedarf. [Habilitation.] Kiel, Kiel University: 137. (in German)

Sharafi S., Mohammadi Ghaleni M. (2021): Calibration of empirical equations for estimating reference evapotranspiration in different climates of Iran. Theoretical and Applied Climatology, 145: 925–939.

Sharma G., Singh A., Jain S. (2022): A hybrid deep neural network approach to estimate reference evapotranspiration using limited climate data. Neural Computing and Applications, 34: 4013–4032.

Sowmya M.R., Santosh Kumar M.B., Ambat S.K. (2020): Comparison of deep neural networks for reference evapotranspiration prediction using minimal meteorological data. In: Proceedings of the 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), July 2–4, 2020, Cochin, India: 27–33.

Tabari H., Talaee P.H. (2011): Local calibration of the Hargreaves and Priestley-Taylor equations for estimating reference evapotranspiration in arid and cold climates of Iran based on the Penman-Monteith model. Journal of Hydrologic Engineering, 16: 837–845.

Vapnik V. (1995): The Nature of Statistical Learning Theory. New York, Springer: 314.

Vishwakarma D.K., Pandey K., Kaur A., Kushwaha N.L., Kumar R., Ali R., Elbeltagi A., Kuriqi A. (2022): Methods to estimate evapotranspiration in humid and subtropical climate conditions. Agricultural Water Management, 261: 107378.

Wen X., Si J., He Z., Wu J., Shao H., Yu H. (2015): Support-vector-machine-based models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions. Water Resources Management, 29: 3195–3209.

Wu L., Peng Y., Fan J., Wang Y. (2019): Machine learning models for the estimation of monthly mean daily reference evapotranspiration based on cross-station and synthetic data. Hydrology Research, 50: 1730–1750.

Yuan X. (2018): Monthly runoff forecasting based on LSTM-ALO model. Stochastic Environmental Research and Risk Assessment, 32: 2199–2212.

Zhang J., Zhu Y., Zhang X., Ye M., Yang J. (2018): Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. Journal of Hydrology, 561: 918–929.

Received: October 16, 2023
Accepted: November 5, 2024
Published online: March 12, 2025