Journal of Agrometeorology

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 26 (1) : 63 - 68 (March - 2024)
https://doi.org/10.54386/jam.v26i1.2462
https://journal.agrimetassociation.org/index.php/jam

Research Paper

Machine learning methods for estimating reference evapotranspiration

AMIT BIJLWAN[1], SHWETA POKHARIYAL[1], RAJEEV RANJAN[1*], R.K SINGH[1] and ANKITA JHA[2]

[1]Department of Agrometeorology, G.B Pant University of Agriculture and Technology Pantnagar 263145, Uttarakhand, India
[2]ICAR-Indian Institute of Water Management, Bhubaneswar, Odisha
**Corresponding author- rajeevranjanagri@gmail.com*

ABSTRACT

Precise estimation of evapotranspiration is crucial for optimizing crop water uses particularly in the context of agriculture and horticultural production. In this study, various machine learning techniques was used to determine reference evapotranspiration by leveraging historical weather data. The models tested include artificial neural networks (ANN), Lasso, Ridge, Random Forest, LGBM regressor, and Gradient boosting regressor. LGBM regressor emerged as the top-performing model, exhibiting exceptional accuracy with a testing R-squared of 1.0. ANN also demonstrated notable performance, achieving a testing R-squared of 0.99. Moreover, the Random Forest and Gradient boosting regressor models showcased strong predictive capabilities, with R^2 values of 0.99 and 0.98, respectively. These models offer valuable alternatives for estimating evapotranspiration, providing robustness and adaptability to diverse environmental datasets.

***Keywords**:* Reference Evapotranspiration, LGBM regressor, ANN, Random Forest, Gradient boosting regressor

The scope of agricultural water resources still has limited availability, necessitating the implementation of effective irrigation management that prioritize water conservation (Anapalli *et al*., 2016; Adak *et al.,* 2015). Evapotranspiration (ET) is an important component of the regional water budgeting as a function of interactions between the soil, vegetation and atmosphere (Allen *et al*., 1998; Liu *et al*., 2013). Although direct measuring techniques like eddy covariance, Bowen ratio, or lysimeters are available, soil heterogeneity and dynamic energy transfer make them difficult for calculating spatially resolved reference crop evapotranspiration (ETc) at large scales (Jiang *et al.,* 2016), making it expensive, time-consuming, and inconvenient to use these methods. Climatic factors such as solar radiation, air and soil temperature, atmospheric humidity, wind speed etc. have a considerable impact on evaporation while crop characteristics and agricultural practices have a significant impact on transpiration (Rana and Katerji, 2000). In order to overcome these constraints, machine learning algorithms have emerged as viable alternative for estimating reference evapotranspiration (ETo) and ETc. Machine learning models captures intricate patterns and correlations in data by leveraging large datasets and strong computing methods, resulting in more accurate ETo predictions. These models are trained using historical meteorological data, satellite images, and ground-based evapotranspiration observations. Regression analysis is a popular machine learning technique for estimating ET and a variety of regression methods including decision tree, random forest, support vector machine and linear regression are employed.

Tree-based machine learning algorithms have been shown as effective in calculating groundwater levels, solar radiation, soil moisture and evaporation (Hassan *et al.,* 2017; Bhattacharya, *et al.,* 2018). These methods not only excel at detecting patterns and trends, but they are also computationally efficient, especially for reasonably big datasets. Tree-based methods, when compared to other machine learning techniques, provide a simple yet robust solution for solving complicated environmental modelling issues. Artificial neural networks (ANNs) are another method for estimating evapotranspiration adept at capturing non-linear relationships and complex patterns, making them suitable for modeling the intricate nature of evapotranspiration. Machine Learning models find applications in various fields, including yield prediction (Gupta *et al*., 2022; Saravanan and Bhagavathiappan, 2022). Setiya *et al.,* (2022) used five distinct approaches—SMLR, LASSO, ELNET, Ridge regression, and ANN to explore the correlation between yield

***Article info - DOI**:* *https://doi.org/10.54386/jam.v26i1.2462*
Received:1 December 2023; Accepted: 21 February 2024; Published online : 1 March 2024

and weather parameters and Patel *et al.,* (2023) integrated remote sensing variables with machine learning techniques to predict crop yield. In the present study the various models viz. Artificial neural networks (ANN), Lasso, Ridge, Random forest, LGBM regressor, and Gradient boosting regressor and been used to estimate the reference evapotranspiration at Pantnagar, Uttarakhand.

MATERIAL AND METHODS

The daily records for various meteorological variables for the period 1996 to 2017 were obtained from the Agrometeorological observatory, Department of Agrometeorology, Govind Ballabh Pant University of Agriculture & Technology, Pantnagar located at 29^0N latitude and 79.30^0E longitude. The meteorological variables include evaporation, sunshine hours (which were converted into radiation using the Angstrom's equation), total rainfall, wind direction, wind speed, vapor pressure, relative humidity, dry bulb temperature, air temperature etc. The dataset provides comprehensive information about the climatic conditions and environmental factors that influence wheat growth in the specified location over the 9-year period comprising 860 days in total. To facilitate the development and testing of the model, the entire dataset underwent partitioning into training and testing sets. Specifically, the data was allocated into 80% for training and 20% for testing purposes. Additionally, radiation derived from sunshine hours provides insights into the availability of solar energy, which is vital for photosynthesis and crop development. There were missing values (null values) in the dataset comprising Evaporation and climatic variables. To address this issue, the dataset was interpolated using various interpolation methods to fill up the missing values. The following methods were employed for interpolation: linear, nearest, zero, slinear, quadratic, and cubic.

Interpolation methods

The interpolation process aimed to provide a completer and more accurate dataset for further analysis and modelling, ensuring that the missing numbers have no adverse effect on the results and insights obtained from the data. The dataset can be efficiently used for investigating the relationship between reference evapotranspiration and meteorological variables by utilizing multiple interpolation approaches after selecting the best-performing method, promoting valuable agricultural research and climate studies.

The process involved estimating the missing values in the dataset using each interpolation method. The linear interpolation method computes new data points within the range of existing data points by connecting them with straight lines. The nearest interpolation method uses the value of the nearest data point to fill the missing value. The zero method sets all the missing values to zero. The slinear, quadratic, and cubic methods use piecewise linear, quadratic, and cubic polynomials, respectively, to estimate missing values. After applying each interpolation method, the accuracy of the interpolated data was evaluated by calculating the root mean square error between the interpolated as well as the original dataset with the missing values filled using linear interpolation. Finally, the interpolation method with the lowest RMSE was selected as the best method for filling the missing values in the dataset.

Estimating reference evapotranspiration (ETo)

The Penman-Monteith equation was used for calculating ETo (Allen *et al*., 1998)

$$ETo = \frac{0.408 * \Delta(R_n - G) + \gamma * \frac{900}{T+273} * u2 * (e_s - e_a)}{\Delta + \gamma * (1 + 0.34 * u2)}$$

Where:

ETo is the reference evaporation (mm/day).

Δ is the slope of vapour pressure curve (kPa/^{0}C).

Rn is the net radiation at the crop surface (MJ/m^2/day).

G is the soil heat flux density (MJ/m^2/day).

γ is the psychrometric constant (kPa/^{0}C).

T is the mean daily air temperature at 2m height (^{0}C).

u2 is the wind speed at 2 meters height (m/s).

es is the saturation vapour pressure (kPa).

Ea is the actual vapour pressure (kPa).

While determining ETo, different weather parameter such as air temperature, humidity, wind speed and sunshine hours are taken into account.

Artificial neural network (ANN)

The Keras library was used to develop the ANN. The network architecture consists of four dense layers: the first two hidden layers have 50 neurons each, the third layer has 100 neurons and the output layer contains a single neuron for linear regression. Rectified linear unit (ReLU) activation function is employed in all hidden layers, however the output layer employs a linear activation function. The model is compiled using the Adam optimizer, a widely used mean squared error and the stochastic optimization algorithm loss function to quantify the difference between actual and predicted values. The training process iterates over the dataset for 100 epochs and mini batch gradient descent is performed having a batch size of 50. After that, the model is trained using the provided training data. Additionally, ANN model evaluates the model's performance on the training and test data by making predictions for both datasets.

Ridge regression

Ridge regression model was developed using pre-processing steps and implements with hyperparameter tuning (Table 1), and evaluation of model performance using different metrics. Hyperparameter tuning was done through iterates a list of alpha values for Ridge regression regularization. For each alpha value, a Ridge regressor is trained training data and evaluated on the validation data. The best model with the lowermost mean squared error (MSE) on the validation data is selected. The cross-validation results for Ridge regression outperform the non-cross-validation results, indicating that the model's performance is more reliable and generalizes better to unseen data. It provides a more robust estimate of the model's performance, leading to higher metric scores, such as lower MSE, improved R-squared and enhanced overall model stability.

Table 1: Regularization parameter for Model Development

Parameter	ANN	Ridge regression	Lasso regression	Random forest	Light GBM	Gradient boosting regressor
Cross Validation (Kfold=5)	✓	✓	✓	✓	✓	✓
Randomized SearchCV	✓	✓	✓	✓	✓	✓
L1 Regularization			✓		✓	
L2 Regularization		✓				

Table 2: Evaluation matrix of ANN, Lasso, Ridge, Random Forest, LGBM regressor and Gradient boosting regressor

Model	ANN	Lasso	Ridge	Random Forest	LGBM regressor	Gradient boosting regressor
Training R^2	0.998	0.991	0.991	0.995	1.000	1.000
Training R^2_adj	0.998	0.991	0.991	0.995	1.000	1.000
Training RMSE	0.048	0.098	0.098	0.068	0.008	0.004
Training MSE	0.002	0.010	0.010	0.005	0.006	$1.86E^{-05}$
Training MBE	0.003	$5.51e^{-16}$	$-2.0e^{-16}$	-0.0011	$-4.56e^{-09}$	$-4.82e^{-17}$
Testing R^2	0.997	0.990	0.990	0.983	0.990	0.991
Testing R^2_adj	0.997	0.989	0.989	0.982	0.097	0.990
Testing RMSE	0.060	0.104	0.104	0.133	0.074	0.097
Testing MSE	0.004	0.011	0.011	0.018	0.000	0.009
Testing MBE	0.048	-0.002	-0.002	-0.006	0.004	0.003

Lasso regression

Lasso regression model was developed using hyperparameter tuning (Table 1). It fits the Lasso regressor to different training subsets and predicts target values on the respective validation subsets. The cross-validated predictions are obtained, allowing a more robust assessment of the model's generalization performance. It defines a hyperparameter grid with alpha values to be explored. It performs an exhaustive search, training a Lasso regressor for each combination of hyperparameters. It chooses the best model based on the model with the lowest MSE and obtains the optimal alpha found during the search.

Random forest

Comprehensive approach was used to develop Random forest regressor, including hyperparameter tuning (Table 1), and model evaluation. The final model is then constructed using Random forest regressor. It defines a set of hyperparameters and their potential values to explore. These hyperparameters include the total number of estimators (trees) in a forest, the number of features to consider at each division, the maximum depth of the trees. The minimum number of samples required for splitting a node, minimum number of samples required at each leaf node, whether bootstrap samples are used for training. Based on the MSE, it selects the best possible model and returns best hyperparameters found during the search.

LightGBM (LGBM) model

LightGBM is a gradient boosting framework that performs classification and regression tasks using a tree-based learning approach. LightGBM uses a histogram-based approach to bin continuous features, which helps to reduce memory usage and speed up training. LGBM regressor model is created with default hyperparameters. It searches for the best combination of hyperparameters from the specified distributions. The hyperparameters to tune learning rate, number of estimators (trees), highest depth of trees, minimum number of samples necessary to create a leaf node, fraction of features to evaluate with each split, fraction of the input data used for training each of the trees, plus L1 regularization term. The algorithm performs 50 iterations to explore different hyperparameter combinations. The best LGBM regressor model with optimal hyperparameters is obtained. The best model is then used to make predictions on the test set.

Gradient boosting regressor (GBR)

A comprehensive process of GBR model development, was done using hyperparameter tuning (Table 1). Model defines a hyperparameter grid with different learning rates, numbers of estimators, random state values, and maximum depths to be explored. The GBR model is created using Gradient boosting regressor as the regressor.

RESULT AND DISCUSSION

Machine learning approaches provide accurate and reliable predictions of reference evapotranspiration. Comparison of various machine learning models' performance metrics on predicting ETo has been depicted in Table 2. It compares various regression models based on their accuracy, R-squared values, and prediction errors (root mean square error (RMSE), MSE, mean biased error (MBE)) on both the training and testing datasets. The best-performing model would have higher R-squared values, lower RMSE and MSE, and an MBE close to zero, indicating an appropriate balance between fitting the training data and generalizing to previously unknown data.

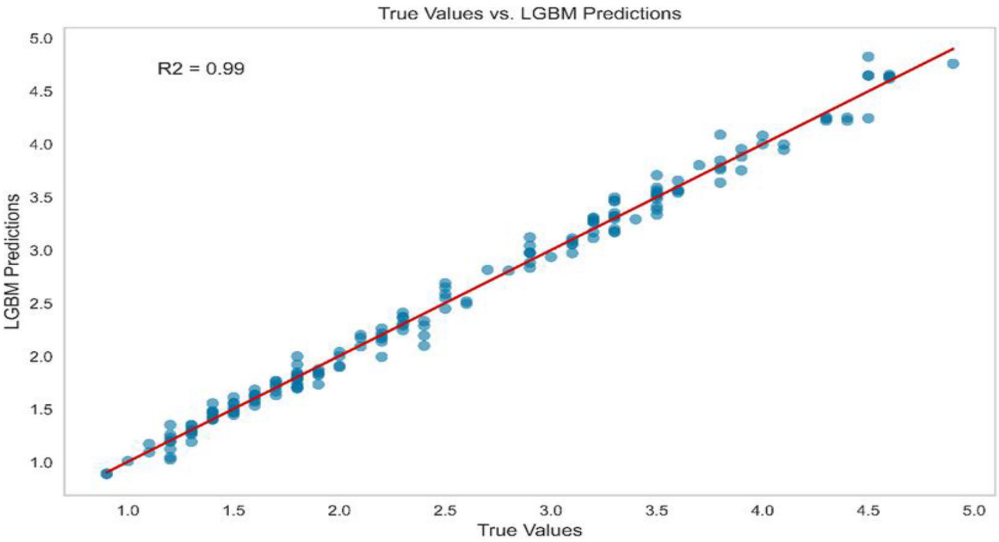

Fig 1: Reference ET calculated by Penman monteith equation and LGBM regressor

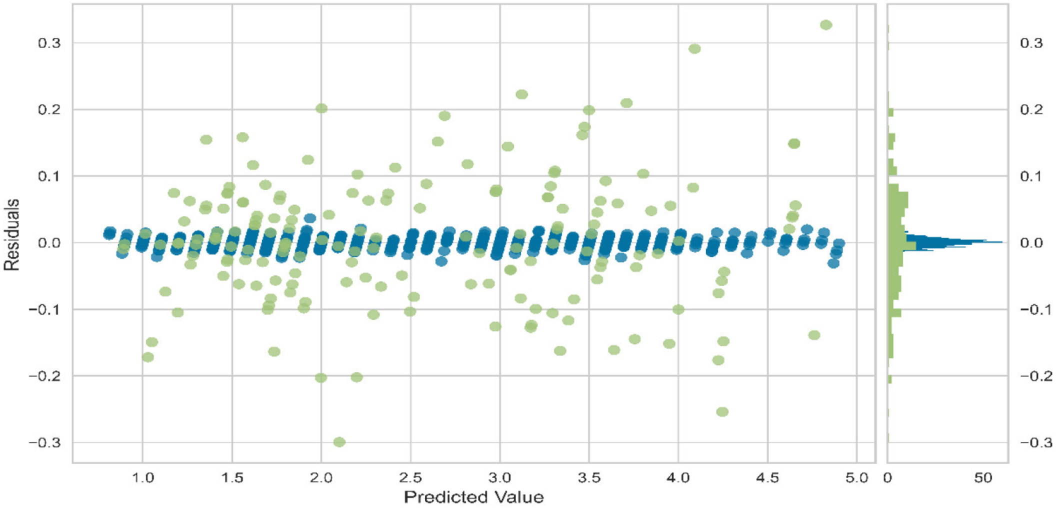

Fig 2: Residual plot and distribution of training and testing of LGBM regressor

Both the training and testing sets demonstrate the exceptional performance of LGBM regressor. Based on the training data, it achieves a perfect R-squared value of 1.00, indicating that it is perfectly suited to the data. The high coefficient of determination (R^2 = 0.99) indicating that it captures 99% of the variance in the testing data (Fig. 1). The low RMSE and MSE values indicate that its predictions are very close to the actual values in the testing set. LGBM regressor has an MBE of approximately 0 on both the training and testing datasets, implying that the model's predictions are, on average, very close to the true values. This also suggests that the model does not have a significant bias in its predictions.

The GBR and ANN also performs exceptionally well in terms of training accuracy, with very high R-squared values (1.00 and 0.998), indicating an excellent fit to the training data (Fig. 2). The model also demonstrates good generalization capabilities, as evident from the relatively high R-squared value on the testing data. The RMSE and MSE values on both training and testing datasets are quite low, indicating accurate predictions. The MBE is also close to zero, showing minimal prediction bias. The Random Forest Regressor also exhibits strong performance with high R-squared values on both training and testing data, indicating a good fit and generalization. The RMSE and MSE on both training and testing datasets are relatively low, suggesting accurate predictions (Fig. 3). Overall, LGBM regressor, GBR and ANN are strong performing models, with the LGBM regressor slightly outperforming them in terms of R-squared, RMSE, and MSE. Also, the ANN's strong performance, characterized by high accuracy, low bias, and excellent generalization capabilities, positions it as a compelling and reliable model for the given task

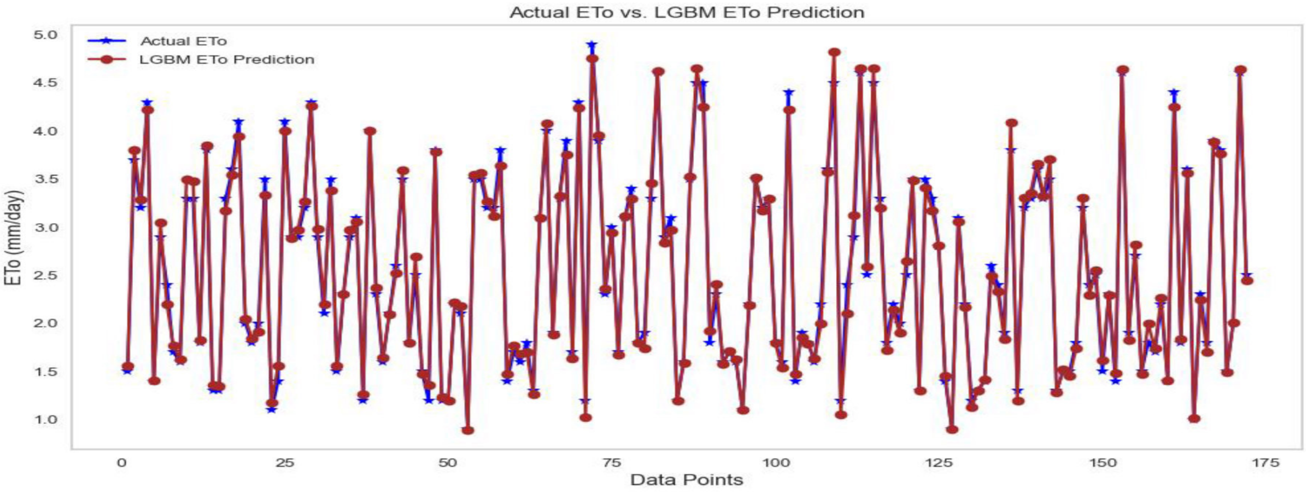

Fig 3: Penman Monteith evapotranspiration (ETo) vs. LGBM prediction

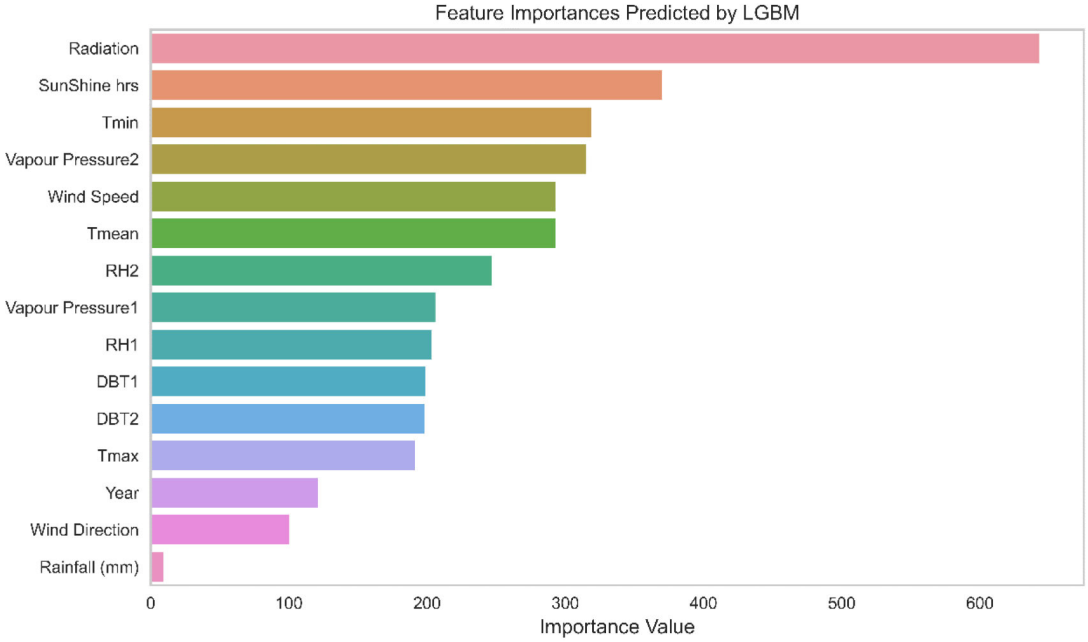

Fig 4: Feature importance as predicted by LGBM

Feature importance values that are non-negative indicate the positive contribution of each feature to the predictive performance of a model (Fig.4). Feature importance values are non-negative and represent the weightage of each feature towards the evaporation loss. Higher positive values generally indicate that the feature is more important in predicting the evaporation loss, and lower values imply lower importance. Feature importance suggests that radiation, sunshine hours, vapour pressure, Tmin and wind speed are the most important feature. More radiation, and higher duration of sunshine hours, lead to higher temperatures and it leads to further increased in evaporation rates. It suggests that higher wind speeds have a significant impact on increasing evaporation loss.

The results show that machine learning algorithms can outperform traditional approaches for estimating ET. The results indicated that LGBM regressor provides the highest testing R-squared value among the tested models highlighting its potential for accurate wheat evapotranspiration estimation. Similar results have been reported by Jatav et al. (2023) and Gao *et al.* (2020).

CONCLUSION

The study has demonstrated that machine learning technoques, including the LGBM regressor, GBR and ANN, have a significant potential for estimating ETo with high precision. LGBM regressor emerged as the top-performing model in this study, achieving a testing R^2 value of 1.0, indicating a near-perfect fit to the data. These models not only fit the training data well but also demonstrated strong generalization capabilities, as evidenced by low RMSE and MSE values on both training and testing datasets. Additionally, the low MBE values indicate minimal prediction bias. The feature importance analysis revealed that higher levels of solar radiation and wind velocity contributed to increase evaporation rates.

ACKNOWLEDGEMENT

The authors are grateful to the Department of Agrometeorology, G.B. Pant University of Agriculture and Technology, Pantnagar and India Meteorological Department for providing laboratory facilities and meteorological data in carrying out this research.

***Funding**:* The research work was conducted without external funding; however, it benefitted from the support of a university fellowship and resources.

***Conflict of Interests*:** The authors declare that there is no conflict of interest in this article.

Data availability: Data will be available through corresponding author on reasonable ground.

Author's contribution: **A. Bijlwan**: Methodology, Writing and editing; **S. Pokhriyal**: Methodology, Visualization; **R. Ranjan**: Investigation, Supervision; **A. Jha**: Writing-review and editing; **R.K Singh**: Supervision

Disclaimer: The contents, opinions, and views expressed in the research communication published in the Journal of Agrometeorology are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

Publisher's Note: The periodical remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

Adak, T., Kumar, K. and Singh, V. K. (2015). Assessment of variations in reference evapotranspiration, yield and water use efficiency of mango under different fertigation regimes. *J. Soil Water Conserv.,* 14(3): 232-240.

Allen, R. G., Pereira, L. S., Raes, D. and Smith, M. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. FAO, Rome, 300(9): D05109.

Anapalli, S. S., Ahuja, L. R., Gowda, P. H., Ma, L., Marek, G., Evett, S. R. and Howell, T. A. (2016). Simulation of crop evapotranspiration and crop coefficients with data in weighing lysimeters. *Agric. Water Manag*, 177:274-283. https://doi.org/10.1016/j.agwat.2016.08.009

Bhattacharya, P., Maity, P. P., Ray, M. and Krishnan, P. (2018). Comparison of artificial neural network and multi-linear regression for prediction of field capacity soil moisture content. *J. Agril Phys,* 18(2): 173-180.

Gao, W., Li, Z. and Liu, X. (2020). A machine learning approach for estimating crop evapotranspiration based on historical weather data. *Comput Electron Agric*, 177:105701.

Gupta, S., Vashisth, A., Krishnan, P., Lama, A., Prasad, S., and K. S. Aravind. (2022). Multistage wheat yield prediction using hybrid machine learning techniques. *J. Agrometeorol.*, *24*(4): 373–379. https://doi.org/10.54386/jam.v24i4.1835

Hassan, M. A., Khalil, A., Kaseb, S. and Kassem, M. A. (2017). Exploring the potential of tree-based ensemble methods in solar radiation modeling. *Appl. Energy*, 203: 897-916. https://doi.org/10.1016/j.apenergy.2017.06.104

Jatav, M. S., Sarangia, A., Singha, D. K., Sahoo, R. N. and Varghese, C. (2023). Advanced machine learning-based kharif maize evapotranspiration estimation in semi-arid climate. *Water Sci. Techn.,* https://doii.org/10.2166/wst.2023.253.

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

Liu, S.M., Xu ZW, Zhu ZL, Jia, Z.Z. and Zhu, M.J. (2013). Measurements of evapotranspiration from eddy-covariance systems and large aperture scintillometers in the Hai River Basin, China. *J. Hydrol*., 487:24-38. https://doi.org/10.1016/j.jhydrol.2013.02.025

Patel, N. R., Pokhariyal, S., and Singh, R. P. (2023). Advancements in remote sensing-based crop yield modelling in India. *J. Agrometeorol*, 25(3): 343-351. https://doi.org/10.54386/jam.v25i3.2316

Rana, G. and Katerji, N. (2000). Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: a review. *Eur J. Agron*, 13(2-3): 125-153.https://doi.org/10.1016/S1161-0301(00)00070-8

Saravanan, Krithikha Sanju and Velammal Bhagavathiappan. (2022). A comprehensive approach on predicting the crop yield using hybrid machine learning algorithms. *J. Agrometeorol*, *24*(2): 179–185. https://doi.org/10.54386/jam.v24i2.1561

Setiya, P., Satpathi, A., Nain, A. S., and Das, B. (2022). Comparison of weather-based wheat yield forecasting models for different districts of Uttarakhand using statistical and machine learning techniques. *J. Agrometeorol.*, 24(3): 255-261. https://doi.org/10.54386/jam.v24i3.1571