Predicting Forest Evapotranspiration using Remote Sensing and Machine Learning

Bhawna Yadav[1], Laxmi Kant Sharma[1], Basant Bijarniya[1]

[1]Department of Environmental Science, Central university of Rajasthan, Ajmer, India - bhavna8295@gmail.com ,
(laxmikant_evs,2024phdevs001)@curaj.ac.in

Keywords: Evapotranspiration, Remote sensing, Machine learning, Sun Induced chlorophyll Fluorescence, Artificial Neural Network

Abstract

Evapotranspiration (ET), which constitutes evaporation from soil and water surfaces and transpiration from stomata of plant leaves, is an important indicator for measuring global hydrological and carbon cycle balances. Though it is crucial to monitor ET for water resource management, energy production, and environmental conservation, predicting ET is a complex task and lacks a reliable approach for accurately predicting ET using remote sensing and meteorological data. ML methods, with their ability to handle complex and non-linear relationships to make accurate predictions, can be used to predict ET. In this study, ML algorithms—Random Forest Regression, Support Vector Regressor, Artificial Neural Network, and an ensemble model—are developed to predict forest evapotranspiration. The models are trained with ECMWF ERA5 reanalysis meteorological parameters (max. and min. air temperature, relative humidity, vapor pressure deficit, precipitation, volumetric soil water content, and wind speed), remote sensing data products (MODIS Enhanced Vegetation Index, MODIS Land Surface Temperature, MODIS Fractional Photosynthetically Active Radiation) as independent parameters, and ET data (8-day data) from MODIS as the target variable. All the datasets are interpolated to a 4-day temporal resolution for the years 2016-2018. From the ensemble model, a satisfactory R-squared value of 0.81 and RMSE value of 0.27 mm/day for the prediction were obtained using the parameters chosen from feature analysis. The trained model is used to predict the forest ET map for the Upper Aravali region for the years 2016-2018. Using ML algorithms to estimate ET rates can be useful for proactive resource management, particularly in water-stressed areas.

1 Introduction

Forests play a significant role in controlling the balance of terrestrial water on Earth, which covers about 31% of its land surface. Forests are one of the world’s major biomes and they evapotranspire an adequate amount of water back into the atmosphere. Therefore, forest evapotranspiration has a significant impact on regional and global climate, as well as river flow, which in turn has an impact on water resources, flooding, and sediment transport. (Komatsu et al., 2012). Additionally, carbon fixation is related to forest evapotranspiration and forest biodiversity. Evapotranspiration is a significant part of the global water cycle, which is important to forests. Numerous variables, such as air and land surface temperature, atmospheric humidity, radiation, wind speed, soil moisture, and vegetation characteristics, have an impact on it. Forests typically evapotranspire at higher rates than other types of land cover because of their dense vegetation, which offers a large surface area for evaporation and transpiration. Due to the complexity of the processes involved, measuring evapotranspiration in forests can be difficult. There are several techniques used, including direct measurements made with tools like lysimeters, eddy covariance towers, and sap flow sensors. By analysing data on variables like vegetation indices, land surface temperature, and satellite-based observations, remote sensing techniques, such as those described above, can also provide useful information on evapotranspiration at regional and global scales. (Ha et al., 2015a).

When plants absorb sunlight during photosynthesis, a process known as solar-induced fluorescence occurs. The physiological health and activity of plants can be inferred from this fluorescence emission. As both SIF and T (canopy temperature) heavily rely on APAR (absorbed photosynthetically active radiation). The rationale behind linking canopy T to remotely sensed vegetation indices (VIs) or leaf area index (LAI) in remote sensing-based methods is that canopy T is associated with carbon assimilation through stomatal conductance at the canopy level (gc). However, real photosynthetic activity is not directly determined by VIs or LAI. Instead, they rely on reflectance data, which are common in satellite-based optical temperature approaches. Therefore, it is anticipated that direct proxies for photosynthesis would enhance canopy T prediction, particularly by gc constraint. Recent advances in measuring and interpreting solar-induced chlorophyll fluorescence (SIF) have made it possible to estimate photosynthetic activity from space (Shan et al., 2021).

Chlorophyll fluorescence has been utilised to research the physiology of photosynthetic activities and stomatal conductance at various cellular and subcellular levels, whereas SIF, as an indirect technique, is typically studied at the canopy extent and above. Stomatal conductance can also be estimated using various spaceborne sensors such as GOME-2, GOSAT, and OCO-2 to enhance the regional ET prediction. Several studies have used the global SIF data which has been collected from various space-based equipment’s to predict global and regional ET. This study assesses SIF's ability to monitor temporal changes in stomatal conductance and transpiration for forest and agricultural land by taking into account the overall interlinkage between vegetation, boundary layer, carbon uptake, and water loss. High frequency time series data was used to study the connection between SIF and stomatal conductance. For the ground measurement flux data has been taken from three different sites, the 8-day satellite data was averaged, an empirical regression model is being carried out between stomatal conductance and SIF. This study showed that SIF and stomatal conductance plays a crucial role in estimating plant carbon and water balance as SIF is a proxy for Gross Primary Productivity (GPP) that can be used to measure stomatal conductance under fully amalgamated carbon uptake and transpiration. (Shan et al., 2019).

The direct measurement of evapotranspiration (ET) using flux towers has several limitations, including high costs, spatial constraints, and data gaps. Similarly, remote sensing-based physical models, while valuable, often face challenges such as limited spatial coverage, the need to incorporate multiple parameters, and difficulties in handling complex relationships. Machine learning (ML) techniques offer a promising alternative by automatically capturing intricate non-linear relationships between ET and its controlling factors. ML models can effectively handle complex interactions, learn patterns directly from data, and make accurate large-scale predictions. Unlike traditional approaches, ML models can reproduce complex processes by mapping relationships between input and output variables. Additionally, they are more tolerant to missing or limited data, utilizing techniques such as imputation and feature selection to enhance predictive accuracy. Various meteorological (e.g., wind speed, relative humidity, temperature, and precipitation) and remote sensing-based parameters (e.g., Enhanced Vegetation Index—EVI) can be integrated into ML models for improved ET estimation. Despite the potential of ML in this domain, limited studies have focused on its application for ET prediction, particularly in forest ecosystems. To address this gap, this study evaluates the performance of different ML models—Random Forest Regression, Support Vector Regression (SVR), Artificial Neural Networks (ANN), and an ensemble approach—for predicting forest ET using meteorological and remote sensing datasets.

2 Overview of methods for estimating ET

2.1 Machine Learning

Machine learning (ML) techniques are increasingly being used to estimate evapotranspiration (ET) at regional scales. For example, Yang et al. (2006) employed flux tower measurements from the Ameri Flux network along with three remote sensing variables—Land Surface Temperature (LST), Improved Vegetation Index (IVI), and land cover—along with surface shortwave radiation to estimate eight-day-averaged ET using a Support Vector Machine (SVM) model. Similarly, Lu and Zhuang (2010) utilized Artificial Neural Networks (ANN) to develop a daily ET product by integrating remotely sensed data, meteorological variables, and flux tower observations. The goal of their study was to scale up tower-based ET measurements to a regional level using ML models. In recent research, five widely used ML techniques—Support Vector Machines (SVM), Deep Belief Networks (DBN), Random Forests (RF), and Artificial Neural Networks (ANN)—have been employed for ET estimation. Additionally, ensemble learning methods, which combine multiple ML models to improve predictive performance have been explored for better accuracy. Ruiz-Aĺvarez et al. (2021), Kanan et al. (2023), Piragnolo et al. (2021), Tausif, et al. (2023).

2.1.1 Random Forest: Based on a CART decision tree model, Breiman (2001), Liu, Y., Zhang, S., Zhang, J., Tang, L., & Bai, Y. (2021). created the RF, which includes the algorithms for classification (RFC) and regression (RFR). The basic idea based on statistical theory is to extract K samples and the bootstrap resampling technique is used to repeatedly and randomly create a new set of training samples from the original training samples set N, followed by the production of K decision trees and a random forest based on the bootstrap sample set. Regarding the classification model, the final prediction outcomes of newly collected data are determined by the number of votes received from the classification tree. Similarly, for the regression model, all averages of the decision trees' predictive values are considered as final outcomes.

2.1.2 Support Vector Regressor: Vapnik initially created SVMs for pattern categorisation, but they have now been applied to regression approximation. SVMs convert nonlinear regression to linear regression by mapping the low-dimensional input space to a higher-dimensional feature space with kernel functions that meet Mercer's criterion, V. Vapnik (1991), V. N. Vapnik (1998), Dou, X., & Yang, Y. (2018).

2.1.3 Artificial Neural Networks: A neural network is a computing system designed to mimic the function of the human brain (Haykin, 1998; Antonopoulos & Antonopoulos, 2017). It is widely used for regression tasks because of its ability to approximate complex nonlinear functions. Among the various neural network algorithms, multilayer perceptron’s with backpropagation (MLP-BP) remain the most commonly used (Haykin, 1998). However, MLP-BP faces challenges in selecting the appropriate network topology and optimizing solutions. Its performance depends on several factors, including the number of hidden layers, the number of neurons per layer, activation functions, weight initialization methods, learning rate, momentum, epoch size, complexity penalty functions, and regularization parameters.

3 Materials and Methods

3.1 Site description

One of the oldest hill ranges in the world, the Aravali is symmetrically located in northwest India and spans three states (Haryana, Rajasthan, Gujarat, and Delhi) as well as one union territory. It used to cover an area of almost 75,000 km^2. Situated between the Gangetic plain and the enormous Thar desert, it is an ecotonal, semi-arid woodland environment. Partially and completely encompassing 15 districts, the upper Aravali range is distributed among the states of Delhi, Haryana, and Rajasthan. Its coordinates are 25°30'-29°N and 75°30'-78°E. With an extent of 37593.2 km^2, it underwent orogeny between 2.5 billion years ago. Compared to other Aravali regions, this range has a lower density of forests and is much more urbanised. Nonetheless, there are a few protected areas, such as the Sariska National Park, Aravali Biodiversity Park, Sultanpur National Park (Kumari et al. 2017).

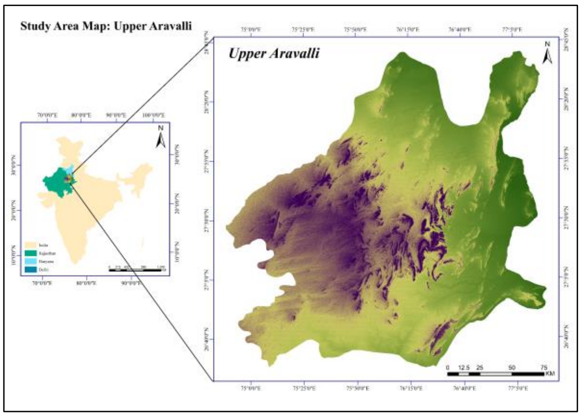


Figure 1. Upper Aravalli Study Area Map

4 Datasets

S. No.	Type	Parameters	Product	Spatial Res. (m)	Temporal Res. (days)
1	Remote Sensing	Evapotranspiration	MOD16A2.061	500	8
2		FPAR	MCD15 FPAR	500	4
3		LST	MOD11 LST	1000	Daily
4		EVI	MOD13 EVI	250	16
5		SIF	OCO-2 (GOSIF)	5000	8
6	Meteorological	VPD	ECMWF ERA5	11132	Daily
7		Soil Moisture	ECMWF ERA5	11132	Daily
8		Relative Humidity	ECMWF ERA5	11132	Daily
9		Precipitation	ECMWF ERA5	11132	Daily
10		Wind Speed	ECMWF ERA5	11132	Daily
11		Air Temperature	ECMWF ERA5	11132	Daily

Table 1. Representation of the datasets used for the study

4.1 MODIS Products

In this study, three MODIS (Moderate Resolution Imaging Spectroradiometer) products are used MOD11A2.006 is the MODIS Land Surface Temperature (LST) product which has 1 km spatial resolution and an 8-day temporal resolution which gives global Terra Land Surface Temperature and Emissivity. It consists of 12 bands of which band 1 LST_Day_1km is being used to carry out the study. MOD13Q1.006 is the MODIS Terra Vegetation Indices (VI) product which has 250 m spatial resolution and a 16-days temporal resolution. It includes 11 bands, featuring two key vegetation indices: NDVI and EVI. This study utilizes band 2 (EVI). MCD15A3H.006 The MODIS Leaf Area Index/FPAR product has a 500 m spatial resolution and a 4-day temporal resolution. It consists of six bands, with band 1 (FPAR) selected for this study. This band covers a wavelength range of 400–700 nm. All MODIS data products are extracted at the point locations chosen randomly in the study area using a Google Earth Engine (GEE) script and exported as CSV files. Additionally, geotiff files for the study area are generated and stored in Google Drive using GEE The training data (CSV) datasets are then interpolated or averaged to match 4-days temporal resolution. If the temporal resolution is greater than 4 days, mathematical average is performed for consecutive 4 days. If the temporal resolution is less than 4 days, the dataset is temporally linearly interpolated for 4 days. For prediction data, the Geotiff files are resampled using the cubic spline to 250m spatial resolution. The geotiff files are generated for every month (average) for the years 2016-2018.

4.1.1 GOSIF: Using a data-driven approach, a global OCO-2 Solar-Induced Fluorescence (SIF) dataset was generated with high spatial (0.05°) and temporal (8-day) resolution for the period 2000–2020. This dataset was created by integrating MODIS remote sensing data, meteorological reanalysis data, and discrete OCO-2 SIF soundings. Pre-processed global GOSIF geotiff files serve as the data source for OCO-2 SIF. These files are downloaded as tar archives and extracted using Python scripts. For the selected locations, pixel values are read and stored as CSV files using the GDAL, NumPy, and pandas libraries. Since the dataset has an 8-day temporal resolution, it is temporally interpolated to a 4-day resolution. For prediction, geotiff files are masked to the study area and resampled to a 250 m spatial resolution using cubic spline interpolation via the GDAL library.

4.1.2. Meteorological datasets: The ERA5-Land dataset, derived from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5-Land Climate Reanalysis, provides daily aggregated climate information, including meteorological variables and land surface parameters. This dataset contains all 50 variables available on the Copernicus Climate Data Store (CDS). For this study, the ERA5-Land Daily Aggregated dataset, accessible via the Google Earth Engine (GEE) data catalogue, is used to extract the required meteorological parameters. For the Upper Aravalli region, meteorological variables—including air temperature, dew point temperature, precipitation, soil moisture, and wind (u & v components)—are extracted for the years 2016–2018 and saved as CSV files using a GEE script. The extracted dataset, containing daily meteorological data (2016–2018) with 1096 rows × 6 columns, is averaged into a four-day resolution, resulting in 274 records. Since ERA5-Land data does not directly provide wind velocity magnitude, relative humidity, or vapor pressure deficit (VPD), these parameters are computed using air temperature and dew point temperature values via a Python script.

Once all necessary parameters are derived, the CSV file is processed further for model training. In addition to the CSV dataset for the study site, geotiff files for all parameters across the Upper Aravalli region are extracted using a GEE script. These files are averaged to a four-day resolution and resampled to a 250 m spatial resolution using cubic spline interpolation.

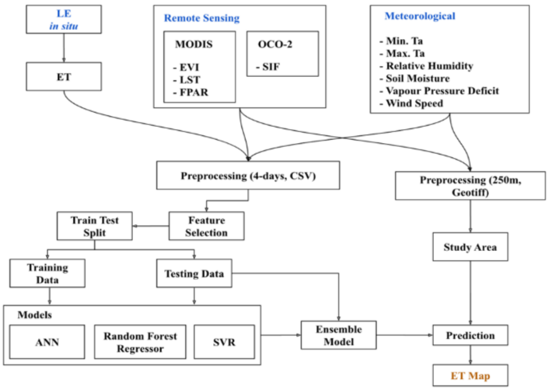


Figure 2. Schematic diagram of methodology workflow

5 Data preprocessing

As outlined in the dataset section, preprocessing of meteorological and remote sensing datasets involves two distinct workflows for training and prediction. For training data, all relevant parameters are extracted as point data for the Upper Aravalli region and stored in a CSV file, where each data point corresponds to the same geographical location. In contrast, for predicting ET across the entire study area, these parameters are extracted as geotiff files for the required time period.

To ensure accurate modelling, it is essential to clean the dataset before proceeding with analysis. Datasets may contain extreme values (outliers) that significantly deviate from the norm and are inconsistent with the rest of the data. Identifying and removing these outliers enhances ML model performance and improves predictive accuracy. Outliers often arise due to measurement **errors** or data processing issues. Statistical techniques can be employed to detect and eliminate such anomalies, ensuring the dataset accurately represents the underlying patterns.

5.1 Feature Selection and Scaling

To improve model efficiency, the number of input parameters can be reduced by removing redundant and non-informative variables through feature selection. A large number of variables can slow down model training and even degrade performance if irrelevant to the target parameter. Different feature selection methods can be applied depending on the variable type. Since all parameters in this study are numerical, Pearson's correlation coefficient and mutual information scores are used to identify the most relevant variables. For the training dataset, selected parameter data points are merged into a common CSV file. These parameters are combined into a Pandas DataFrame and stored as a CSV file, containing 274 records, corresponding to three years of 4-day intervals: (365+365+366)/4 records.

After feature selection, the training dataset is split into training (80%) and testing (20%) subsets using the scikit-learn library. To ensure consistency across all variables, feature scaling (data normalization) is applied to the independent variables. Without normalization, the ML algorithm may become biased, assigning lower importance to smaller values and overemphasizing larger ones. There are multiple feature scaling techniques available such as, 1) Min-Max Scaling, 2) Absolute maximum scaling, 3) Normalisation, 4) Robust Scaling. For this study, Min-Max Scaling is used to normalize the data values, ensuring they fall within a standardized range

5.2 Model training and testing

Once feature scaling is applied, the ML models are ready for training and evaluation. The Random Forest Regression (RF) and Support Vector Regression (SVR) models are implemented using the scikit-learn library, while the Artificial Neural Network (ANN) model is built using TensorFlow. Since ML model performance is highly dependent on hyperparameter tuning, Grid Search Cross-Validation is employed to identify the optimal parameters for the RF model. This process evaluates multiple parameter combinations, validating the model against the dataset and selecting the best-performing configuration. Once tuning is complete, a new RF model is trained with the optimized parameters and used for prediction. To account for variability in model performance due to different train-test splits, cross-validation is performed to compute the average RMSE (Root Mean Square Error) and R² values for both RF and SVR models. The ANN model used in this study consists of four layers: one input layer, two hidden layers, and one output layer. For validation, the ANN model uses Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as performance metrics. Similarly, for all three models (RF, SVR, and ANN), RMSE and R² values are calculated to assess model accuracy.

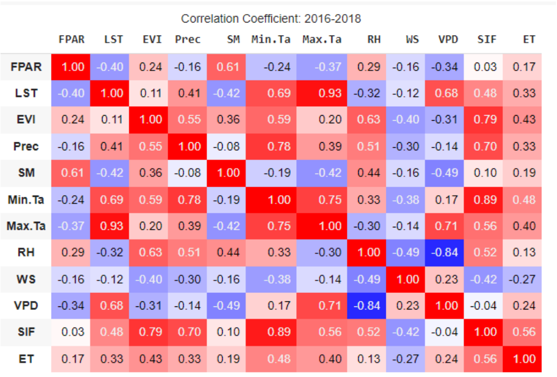
In machine learning (ML), ensemble techniques are employed to combine multiple models, resulting in improved predictive performance. There are several methods to construct an ensemble model. Bootstrap Aggregation (Bagging) involves training multiple models on different subsets of the training dataset to reduce variance and improve stability, with Random Forest being a widely used bagging-based model. Stacked Generalization (Stacking) enhances prediction accuracy by combining outputs from multiple base models and using a meta-model to refine predictions. Boosting is a sequential technique where models are added iteratively, correcting errors from previous models to improve overall performance, producing a weighted average of predictions for enhanced accuracy. In this study, the ensemble model uses the different models that are trained with the same dataset and averages the prediction to get the final ensemble prediction.

6 Result and Discussion

6.1 Feature Selection - Correlation coefficient & Mutual Information

The relationship between two variables in a dataset is measured statistically using feature correlation coefficients (FCCs). They show the linear relationship between the features' strengths and directions. There are several ways to calculate correlation coefficients, but Pearson's correlation coefficient is by far the most popular. The "r" symbol stands for Pearson's correlation coefficient, which calculates the linear correlation between two continuous variables. It accepts numbers in the range of -1 and 1, with -1 denoting a perfect negative correlation, 1 denoting a perfect positive correlation, and 0 denoting no correlation.

Below mentioned are the feature correlation coefficients for the given study period, based on these features it will decide the kind of relation each feature shares with each other. The features which share equal importance with each other from them one feature will be eliminated. However, only linear relationships are being measured by Pearson's correlation.



Correlation Coefficient: 2016-2018

	FPAR	LST	EVI	Prec	SM	Min.Ta	Max.Ta	RH	WS	VPD	SIF	ET
FPAR	1.00	-0.40	0.24	-0.16	0.61	-0.24	-0.37	0.29	-0.16	-0.34	0.03	0.17
LST	-0.40	1.00	0.11	0.41	-0.42	0.69	0.93	-0.32	-0.12	0.68	0.48	0.33
EVI	0.24	0.11	1.00	0.55	0.36	0.59	0.20	0.63	-0.40	-0.31	0.79	0.43
Prec	-0.16	0.41	0.55	1.00	-0.08	0.78	0.39	0.51	-0.30	-0.14	0.70	0.33
SM	0.61	-0.42	0.36	-0.08	1.00	-0.19	-0.42	0.44	-0.16	-0.49	0.10	0.19
Min.Ta	-0.24	0.69	0.59	0.78	-0.19	1.00	0.75	0.33	-0.38	0.17	0.89	0.48
Max.Ta	-0.37	0.93	0.20	0.39	-0.42	0.75	1.00	-0.30	-0.14	0.71	0.56	0.40
RH	0.29	-0.32	0.63	0.51	0.44	0.33	-0.30	1.00	-0.49	-0.84	0.52	0.13
WS	-0.16	-0.12	-0.40	-0.30	-0.16	-0.38	-0.14	-0.49	1.00	0.23	-0.42	-0.27
VPD	-0.34	0.68	-0.31	-0.14	-0.49	0.17	0.71	-0.84	0.23	1.00	-0.04	0.24
SIF	0.03	0.48	0.79	0.70	0.10	0.89	0.56	0.52	-0.42	-0.04	1.00	0.56
ET	0.17	0.33	0.43	0.33	0.19	0.48	0.40	0.13	-0.27	0.24	0.56	1.00

Figure 3. Correlation coefficient matrix of years 2016-2018

SIF has the highest correlation with ET which is 0.56 followed by Min Ta, EVI, Max Ta (0.48, 0.43, 0.40) respectively. From the above observations we can conclude that there is a hidden covariation between SIF and ET and based on that we can derive a relationship.

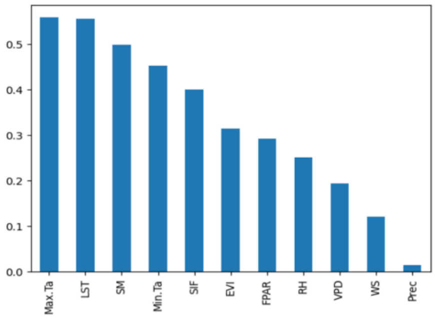


Figure 4. Mutual information score of parameters

Mutual information describes the relationship in terms of uncertainty. It measures the statistical dependency or association between two variables such as feature and the target variable and shows their relevance. It quantifies the amount of information that one variable contains about the other. Non-linear relationship between two random variables is being provided by this. Hence, the high mutual information indicates large reduction in uncertainty, whereas low value indicates small reduction and when the mutual information is zero it means two variables are independent. Max air temperature and LST show highest mutual information, followed by soil moisture, Min air temperature, SIF, EVI, FPAR while precipitation shows lowest mutual information. Based on a high mutual information score only the important features are being selected.

6.2 Relationship between selected features and ET

The above stated are the important features which are being selected based on the mutual information provided are plotted as scatterplots. Water evaporates more quickly from surfaces such as soil, water, and plant leaves at higher temperatures. The transition of water molecules from a liquid to a vapour state is accelerated by increased thermal energy, which also increases the kinetic energy of the water molecules.

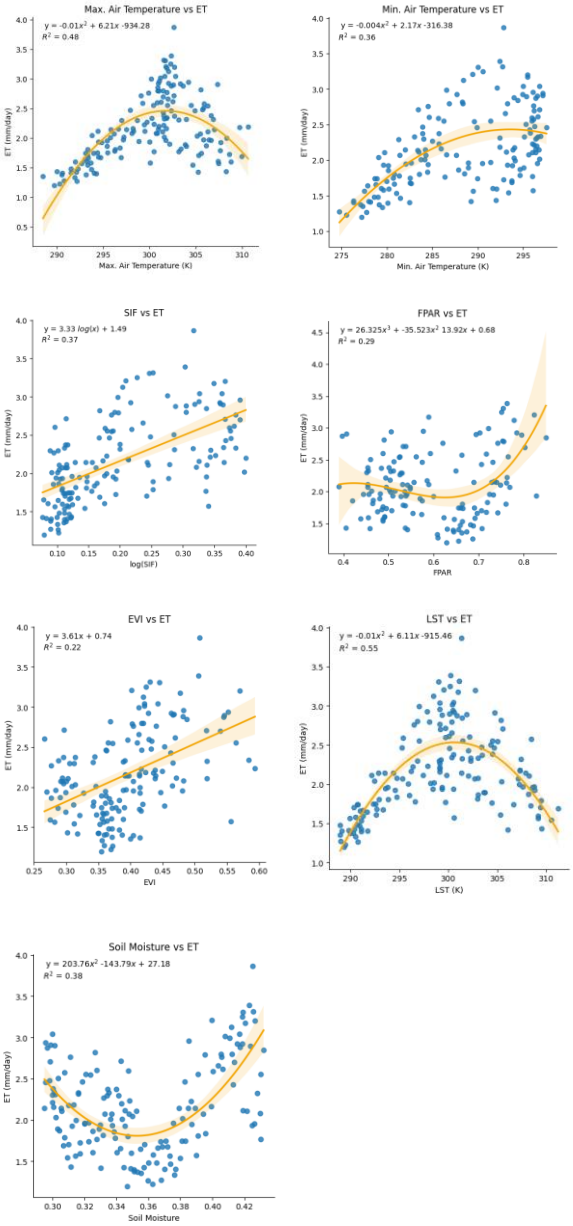


Figure 5. Relationship between important features selected based on mutual information score vs ET.

6.2.1 Temperature Response to Transpiration:

Transpiration, the process of water loss from plants through stomata on leaves, is strongly influenced by temperature. As temperatures rise, stomata open, leading to the release of water vapor, which typically causes transpiration rates to increase. However, this relationship is non-linear.

There is a positive correlation between evapotranspiration and maximum air temperature, although the connection is not purely linear. The degree of evapotranspiration is also affected by various other factors, such as humidity, wind speed, and the availability of soil moisture. Additionally, the form and strength of this relationship can vary depending on the specific geography, plant type, and regional environmental conditions.

The relationship between maximum air temperature (independent variable) and evapotranspiration (ET) (dependent variable) was plotted, with $R^2 = 0.41$. The plot shows that ET increases with temperature up to a certain point, after which it begins to decrease. This behaviour occurs because, at higher temperatures, the stomata begin to close, reducing water loss and thus lowering ET. The temperature in this analysis varied from 285 K to 315 K.
Similarly, when minimum air temperature is plotted against ET, the relationship follows a similar pattern. Initially, ET increases and reaches a peak of about 2.5 mm/day. However, once a certain temperature threshold is reached, ET starts to decrease, with the relationship showing an R^2 value of 0.3. This suggests that while min air temperature has an effect on ET, it is less pronounced than that of max air temperature

6.1.2 SIF and ET: Solar Induced Fluorescence (SIF) and evapotranspiration (ET) share a positive relationship because SIF is an indicator of the light emitted by plants during photosynthesis, and transpiration is closely linked to photosynthetic activity. When plants photosynthesize and transpire, water vapor is released, contributing to evapotranspiration. As SIF values increase, it indicates higher photosynthetic activity, which typically leads to higher transpiration rates, thereby increasing the rate of evapotranspiration.
In this study, SIF shows an R^2 value of 0.37 with ET. To improve the relationship and reduce skewness, a logarithmic transformation is applied to SIF for feature transformation. The plot reveals that ET starts at 1.6 mm/day and increases steadily, reaching up to 2.8 mm/day as SIF rises, reflecting the connection between increased photosynthetic activity and higher evapotranspiration.

6.1.3 FPAR vs ET: FPAR (Fraction of Photosynthetically Active Radiation) and evapotranspiration (ET) are closely related, as FPAR plays a key role in regulating ET. FPAR measures the fraction of solar radiation (in the 400-700 nm range) absorbed by plants, and higher FPAR values typically indicate denser, healthier vegetation canopies. This leads to an increase in leaf area and transpiration rates, resulting in higher ET. Additionally, FPAR is an important parameter in the energy balance equation; when FPAR is high, more energy is absorbed by the vegetation, which contributes to a higher potential for evapotranspiration.

Moreover, FPAR plays a key role in the energy balance equation: higher FPAR suggests that more solar energy is absorbed by the vegetation, which in turn increases the potential for evapotranspiration. The relationship between FPAR and ET in this study shows an R^2 value of 0.26. Using a third-order polynomial equation, it was observed that ET increases initially, then begins to decrease, possibly due to vegetation stress. After reaching a certain threshold, ET starts to increase again as favourable conditions return. The ET values range from 2.2 mm/day to 3.4 mm/day, reflecting this dynamic interaction between FPAR and evapotranspiration.

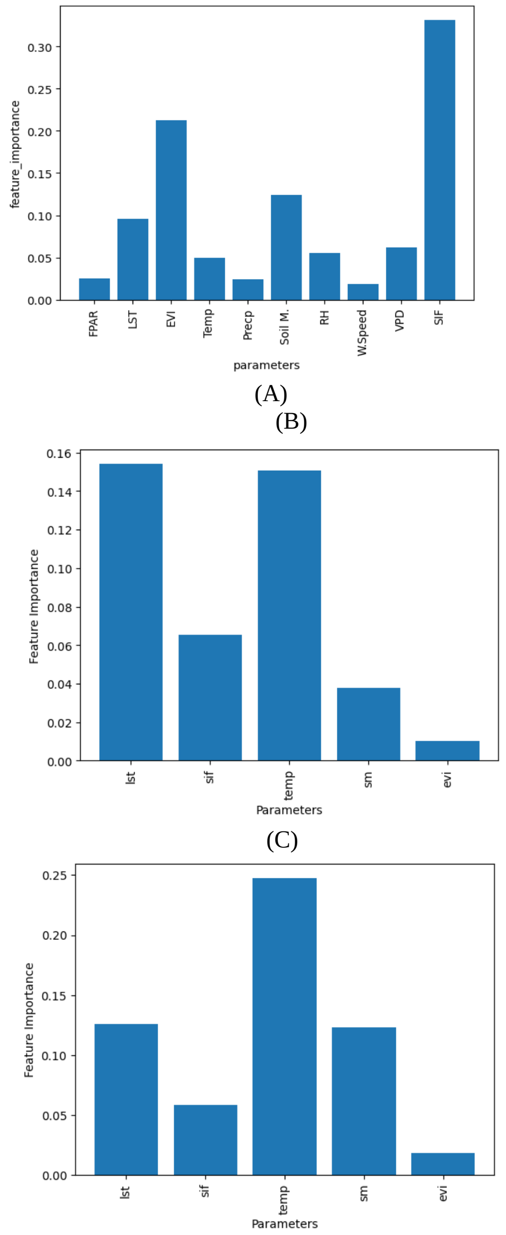
6.1.4 EVI vs ET: The Enhanced Vegetation Index (EVI) is an indicator that measures the greenness and density of vegetation cover, serving as a proxy for plant health and vitality. Since plant density and health are closely linked to transpiration rates, denser and healthier plants are capable of evaporating more water. This increased transpiration contributes to higher evapotranspiration. As a result, EVI and ET share a positive correlation—higher EVI values generally signify healthier vegetation and higher transpiration rates, leading to increased ET.
In this study, the relationship between EVI and ET shows an R^2 value of 0.22 and follows a linear trend. ET starts at 1.5 mm/day and rises to 2.9 mm/day as EVI values increase, ranging from 0.25 to 0.60. This further confirms the link between vegetation health, transpiration, and evapotranspiration.

6.1.5 LST vs ET: Land surface temperature and evapotranspiration are positively related to each other as LST influences the rate of evapotranspiration. Higher LST values generally indicate warmer surface temperatures. Warmer surfaces tend to have increased evaporation rates, as the heat accelerates the conversion of liquid water into vapour. Transpiration refers to the process by which plants release moisture through their leaves. Higher LST can stimulate plant stomatal opening, allowing more water vapour to escape through transpiration. LST having $\boldsymbol{R^2}$ value of 0.48 and following the second order equation, it starts to increase from 285 and increases up to a certain threshold of 300 K, while the ET value goes from 1.2 mm/day to 2.5 mm/day.

6.1.6 Soil moisture vs ET: Soil moisture refers to the amount of water present in the soil, and it plays a critical role in determining the amount of water available for plant uptake and transpiration. When soil moisture levels are high, plants have sufficient water to actively transpire, leading to an increase in evapotranspiration. Therefore, a positive relationship exists between soil moisture and evapotranspiration—higher soil moisture values generally indicate more water available for plant transpiration, resulting in greater ET. In this study, soil moisture has an R^2 value of 0.36. Initially, ET starts at 2.5 mm/day, but it decreases with the onset of the summer season as soil moisture declines due to rising temperatures. As moisture levels decrease, ET also decreases. However, with the arrival of rainfall, soil moisture increases, and ET begins to rise again, peaking at 3.2 mm/day. This demonstrates the dynamic interplay between soil moisture, temperature, and evapotranspiration.

6.3 Feature Importance

Feature Importance quantifies the importance of each parameter in the model. It provides the insights into which features are most influential in predicting the target variable. Across different models, Max.Ta and SIF show higher feature importance. EVI is the least used parameter by the models for the prediction. LST is consistently being used by all the models moderately.



(A)

(B)

(C)

Figure 6. Feature importance of parameters based on different ML models (A) Random Forest, (B) Support Vector Regressor, (C) ANN.

6.4 Model Evaluation

Model evaluation is based on comparing measured ET (mm/day) at location points with predicted ET values. The accuracy is assessed using R² and RMSE. R² indicates how well the predictor variables explain the variance in the response variable, with higher values reflecting better model accuracy. RMSE measures the difference between observed and predicted ET values, with lower values indicating better model performance. These metrics help determine how accurately the models predict evapotranspiration.

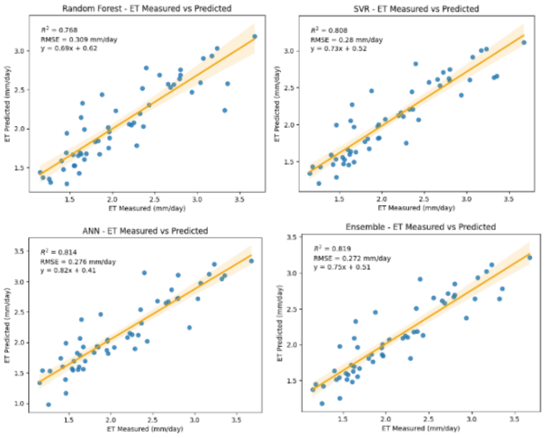


Figure 7. Model evaluation based on predicted and measure value of ET.

The R² value ranges from -1 to +1, with higher values indicating better model performance. RMSE, on the other hand, measures the difference between actual and predicted values, with lower values signifying a better fit to the dataset. The Random Forest Regression model shows an R² value of 0.768 and an RMSE of 0.309 mm/day, demonstrating good performance. The ET values predicted by this model range from 1.45 mm/day to 3.00 mm/day, following a linear trend. The Support Vector Regression (SVR) model outperforms RF, with an R² value of 0.808 and an RMSE of 0.28 mm/day, indicating a better fit and more accurate predictions.

The ANN model achieves an R² value of 0.814 and an RMSE of 0.276 mm/day, delivering the best performance among the individual models. An Ensemble model, which aggregates all the individual models, was also developed to improve accuracy. While the Ensemble model did show a slight improvement with an R² of 0.819 and an RMSE of 0.272 mm/day, the enhancement in accuracy was not substantial. Thus, while the Ensemble model performs better than individual models, the ANN model remains the best-performing model overall. Therefore, ANN is considered the optimal choice for this study.

6.3.1 Influence of SIF in ET prediction: SIF (Solar-Induced Fluorescence) is the measurement of light emitted by plants during photosynthesis in the form of fluorescence. This emission is influenced by various factors, including the photosynthetic activity of pigments like chlorophyll, the health of the vegetation, light conditions, and environmental stressors. SIF serves as an indicator of plant physiological processes, reflecting both the plant's ability to photosynthesize and its overall state.

Ensemble Model Performance	R^2	RMSE (mm/day)
Without SIF With SIF	0.789 0.819	0.294 0.272

Table 2. Ensemble model performance with and without SIF

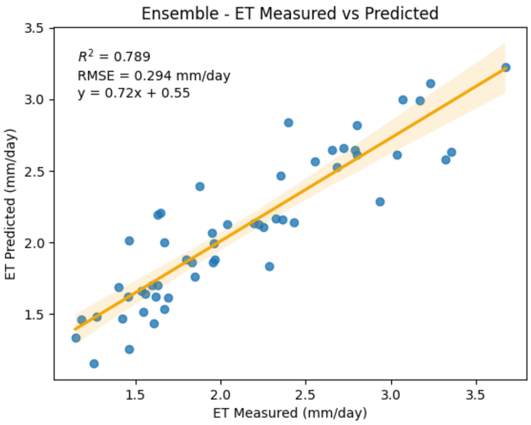


Figure 8. Ensemble model evaluation of ET predicted and measured after including SIF

The inclusion of SIF as an independent parameter slightly improves model performance, showing higher feature importance during prediction. However, SIF is highly correlated with other variables such as Min.Ta, Max.Ta, and EVI, which the models predominantly rely on in the absence of SIF. Consequently, the absence of SIF does not significantly impact the model results. Without SIF, the model yields an R^2 value of 0.789 and an RMSE of 0.294 mm/day, while with SIF included, the R^2 value improves to 0.819, and the RMSE decreases to 0.272 mm/day. This suggests that although SIF has some influence, it does not drastically affect model accuracy.
4o mini

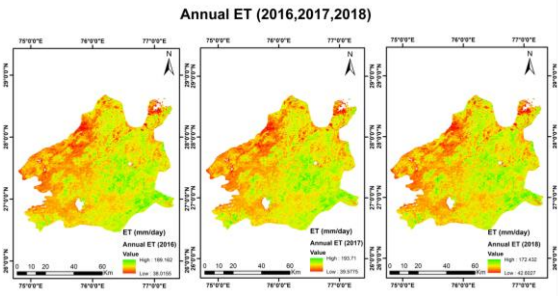


Figure 9. Annual ET maps of years 2016-2018.

The annual ET predicted for three years from 2016 – 2018 is found to be maximum in 2017. It ranges from 39.58 mm/day to 193.71 mm/day. As there is more foliage and more greenness the rate of evapotranspiration itself will be higher. While the lowest ET is seen in 2018 and in 2016 it is comparatively high.

Model	R^2	RMSE
RF	0.768	0.309
SVR	0.808	0.280
ANN	0.814	0.276
Ensemble	0.819	0.272

Table 3. Summary of the results

7 Conclusion

In this study, a comprehensive analysis of evapotranspiration (ET) was conducted using multiple machine learning models, including Random Forest Regression (RF), Support Vector Regression (SVR), and Artificial Neural Networks (ANN), with the aim of predicting ET dynamics in the Upper Aravali region. Several meteorological and remote sensing parameters, such as temperature, soil moisture, SIF, and vegetation indices, were utilized to enhance model prediction accuracy. Among the models tested, ANN emerged as the best performer, achieving the highest R^2 value of 0.814 and the lowest RMSE of 0.276 mm/day. The ensemble model, though slightly improving the results, did not show substantial improvement over the ANN model, confirming that ANN provided the most accurate predictions.

Through feature selection and analysis, key variables such as Max.Ta, SIF, and soil moisture demonstrated significant importance in influencing ET rates. Interestingly, while the inclusion of SIF resulted in slight performance improvements, its correlation with other variables such as temperature and vegetation indices indicated that SIF's role in the model was somewhat redundant, without dramatically enhancing accuracy. Nevertheless, the study underscores the importance of combining multiple variables to effectively capture the complexity of ET processes, highlighting the value of remote sensing data and meteorological parameters in regional-scale evapotranspiration predictions. The findings of this research provide valuable insights into the use of machine learning for environmental monitoring, specifically in regions impacted by land-use changes and climate variability. Future work can focus on integrating multiple data fusion approaches to improve model accuracy. Leveraging cloud computing and cloud-based engines for near real-time monitoring and prediction will enable efficient data processing and timely decision-making, particularly for applications like water management, agriculture, and disaster response. This approach offers scalable solutions for real-time decision support across diverse regions.

.

References

Antonopoulos, V. Z., & Antonopoulos, A. V. 2017. Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. *Computers and Electronics in Agriculture*, *132*, 86-96.

Bai, Y., Zhang, S., Bhattarai, N., Mallick, K., Liu, Q., Tang, L., ... & Zhang, J. 2021. On the use of machine learning based ensemble approaches to improve evapotranspiration estimates from croplands across a wide environmental gradient. *Agricultural and Forest Meteorology*, *298*, 108308.

Dou, X., & Yang, Y. 2018. Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems. *Computers and Electronics in Agriculture*, *148*, 95-106.

Ha, W., Kolb, T. E., Springer, A. E., Dore, S., O’Donnell, F. C., Martinez Morales, R., Masek Lopez, S., & Koch, G. W. 2015. Evapotranspiration comparisons between eddy covariance measurements and meteorological and remote-sensing-based models in disturbed ponderosa pine forests. *Ecohydrology*, *8*(7), 1335–1350. https://doi.org/10.1002/ECO.1586

Haykin, S. 1998. *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Kanan, A.H., Pirotti, F., Rahman, M.M., 2023. Mapping inundation from sea level rise and its interaction with land cover in the Sundarbans mangrove forest. Climatic Change 176, 104. https://doi.org/10.1007/s10584-023-03574-5

Komatsu, H., & Kume, T. 2020. Modeling of evapotranspiration changes with forest management practices: A genealogical review. *Journal of Hydrology,585*. https://doi.org/10.1016/J.JHYDROL.2020.124835

Kumari, R., Kant, K., & Garg, M. 2017. Natural radioactivity in rock samples of Aravali hills in India. *International Journal of Radiation Research, 15*(4), 391-398.

Liu, Y., Zhang, S., Zhang, J., Tang, L., & Bai, Y. 2021. Assessment and comparison of six machine learning models in estimating evapotranspiration over croplands using remote sensing and meteorological factors. *Remote Sensing*, *13*(19), 3838.

Lu, X., & Zhuang, Q. 2010. Evaluating evapotranspiration and water-use efficiency of terrestrial ecosystems in the conterminous United States using MODIS and AmeriFlux data. *Remote Sensing of Environment*, *114*(9), 1924-1939.

Lu, X., Liu, Z., An, S., Miralles, D. G., Maes, W., Liu, Y., & Tang, J. 2018. Potential of solar-induced chlorophyll fluorescence to estimate transpiration in a temperate forest. *Agricultural and Forest Meteorology, 252,* 75–87. https://doi.org/10.1016/J.AGRFORMET.2018.01.017

Piragnolo, M., ..., Grigolato, S., 2021. Responding to Large-Scale Forest Damage in an Alpine Environment with Remote Sensing, Machine Learning, and Web-GIS. Remote Sensing 13, 1541. https://doi.org/10.3390/rs13081541

Ruiz-AÍvarez, M., Gomariz-Castillo, F., & Alonso-Sarría, F. 2021. Evapotranspiration response to climate change in semi-arid areas: Using random forest as multi-model ensemble method. *Water*, *13*(2), 222.

Shan, N., Zhang, Y., Chen, J. M., Ju, W., Migliavacca, M., Peñuelas, J., ... & Goulas, Y. 2021. A model for estimating transpiration from remotely sensed solar-induced chlorophyll fluorescence. Remote Sensing of Environment, 252, 112134.

Shan, N., Ju, W., Migliavacca, M., Martini, D., Guanter, L., Chen, J., ... & Zhang, Y. 2019. Modeling canopy conductance and transpiration from solar-induced chlorophyll fluorescence. *Agricultural and Forest Meteorology*, *268*, 189-201.

Shi, H., Zhang, Y., Luo, G., Hellwich, O., Zhang, W., Xie, M., ... & Van de Voorde, T. 2024. Machine learning-based investigation of forest evapotranspiration, net ecosystem productivity, water use efficiency and their climate controls at meteorological station level. *Journal of Hydrology*, *641*, 131811.

Tausif, M., Dilshad, S., Umer, Q., Iqbal, M. W., Latif, Z., Lee, C., & Bashir, R. N. 2023. Ensemble learning-based estimation of reference evapotranspiration (ETo). *Internet of Things*, *24*, 100973.

Vapnik, V. 1998. Statistical learning theory. *John Wiley & Sons google schola*, *2*, 831-842.

Vapnik, V., & Chervonenkis, A. 1991. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, *1*(3), 283-305.

Wu, M., Feng, Q., Wen, X., Deo, R. C., Yin, Z., Yang, L., & Sheng, D. 2020. Random forest predictive model development with uncertainty analysis capability for the estimation of evapotranspiration in an arid oasis region. *Hydrology Research*, *51*(4), 648-665.

Yang, F., White, M. A., Michaelis, A. R., Ichii, K., Hashimoto, H., Votava, P., ... & Nemani, R. R. 2006. Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(11), 3452-3461.