

# **EFFICIENT FORECASTING OF CROP WATER DEMAND**

Project report submitted in partial fulfillment of the requirement for the degree of  
Bachelor of Technology

in

**Computer Science and Engineering/Information Technology**

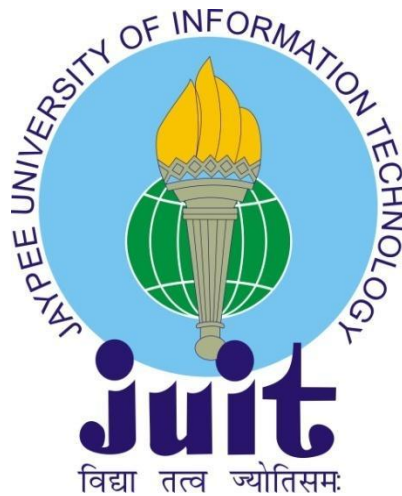
By

SACHIN SHARMA (191283)

Under the supervision of

Dr. Aman Sharma

to




Department of Computer Science & Engineering and Information Technology

**Jaypee University of Information Technology Waknaghat, Solan-173234,  
Himachal Pradesh**

## Candidate's Declaration

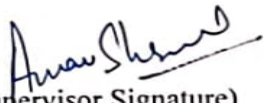
I hereby declare that the work presented in this report entitled “ **Efficient forecasting of crop water demand**” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from July 2022 to May 2023 under the supervision of **Dr.Aman Sharma** (Assistant Professor (SG), Computer Science & Engineering and Information Technology). The matter embodied in the report has not been submitted for the award of any other degree or diploma.



(Student Signature)

Sachin Sharma (191283)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.



(Supervisor Signature)

Dr. Aman Sharma

Assistant Professor (SG)

Computer Science and Engineering

Dated: 5/5/23

**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT**  
**PLAGIARISM VERIFICATION REPORT**

Date: 05/05/23

Type of Document (Tick):  PhD Thesis  M.Tech Dissertation/ Report  B.Tech Project Report  Paper

Name: Sachin Sharma Department: CSE Enrolment No 191283

Contact No. 8629084677 E-mail. sachinpw2017@gmail.com

Name of the Supervisor: Dr. Aman Sharma

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): EFFICIENT FORECASTING OF CROP WATER DEMAND

**UNDERTAKING**

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/ revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

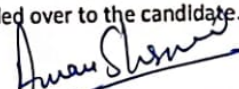
**Complete Thesis/Report Pages Detail:**

- Total No. of Pages = 52
- Total No. of Preliminary pages = 09
- Total No. of pages accommodate bibliography/references = 43

  
(Signature of Student)

**FOR DEPARTMENT USE**

We have checked the thesis/report as per norms and found Similarity Index at 16.....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

  
(Signature of Guide/Supervisor)

  
Signature of HOD

**FOR LRC USE**

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	<ul style="list-style-type: none"> <li>• All Preliminary Pages</li> <li>• Bibliography/Images/Quotes</li> <li>• 14 Words String</li> </ul>	<u>16</u>	Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by  
Name & Signature

Librarian

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at [plagcheck.juit@gmail.com](mailto:plagcheck.juit@gmail.com)

fdgdfgdf

ORIGINALITY REPORT

16%

SIMILARITY INDEX

8%

INTERNET SOURCES

9%

PUBLICATIONS

10%

STUDENT PAPERS

PRIMARY SOURCES

1	<a href="http://www.mdpi.com">www.mdpi.com</a> Internet Source	2%
2	Submitted to University of Greenwich Student Paper	1%
3	Submitted to Aston University Student Paper	1%
4	<a href="http://cropwatch.unl.edu">cropwatch.unl.edu</a> Internet Source	1%
5	<a href="http://www.inass.org">www.inass.org</a> Internet Source	1%
6	<a href="http://resmilitaris.net">resmilitaris.net</a> Internet Source	1%
7	Submitted to University of Northumbria at Newcastle Student Paper	1%
8	<a href="http://link.springer.com">link.springer.com</a> Internet Source	1%
9	<a href="http://www.ijraset.com">www.ijraset.com</a> Internet Source	1%

## Acknowledgement

I would like to express my sincere gratitude and appreciation to all those who have helped me in completing this major project. First and foremost, I would like to thank my project supervisor (Dr. Aman Sharma) for providing me with guidance, support, and valuable feedback throughout the project. Your expertise and knowledge have been invaluable to me.

I would also like to extend my gratitude to Jaypee University of Information Technology, for providing me with the necessary resources and facilities to complete this project. Without your support, this project would not have been possible.

My heartfelt thanks to my family and friends, who have been a constant source of encouragement and motivation throughout my academic journey. Your unwavering support and belief in me have been my greatest strength.

Lastly, I would like to thank all the participants who took part in my study, for sharing their valuable time and insights with me. Your contribution has been instrumental in the successful completion of this project.

Thank you all once again.

  
(Student Signature)

Project Group No. : 34

Student Name: Sachin Sharma

Rollno. : 191283

## Table of Content

Content	Page no.
Certificate	i
Plagiarism Certificate	ii, iii
Acknowledgement	iv
List of figures	vi, vii
List of tables	viii
Abstract	ix
Chapter - 1 Introduction	10
Chapter - 2 Literature survey	15
Chapter - 3 System design and development	29
Chapter - 4 Experiment and results analysis	39
Chapter - 5 Conclusions	43
References	46
Appendices	51

## List of Figures

S.No.	Figure No.	Description	Page no.
1	1.1	Process of evapotranspiration	3
2	3.1	Machine learning pipeline for model building	18
3	3.2	Graph between ET(crop) and other variables.	20,21
4	3.3	Flow Diagram for the proposed methodology	23
5	3.4	Schematic representation of crop coefficients at different growth stages	25
6	3.5	Implementation of XG boosting	27
7	4.1	Graphical representation of residual error	32
8	4.2	The proposed ensemble framework for forecasting of crop water demand	37
9	4.3	Measure of how well our proposed framework predicts the response variable	39
10	5.1	Graphical representation of the comparison of various models with our proposed model	41
11	5.2	Comparison of $R^2$ values of the proposed model with the existing literature	42
12	i	Calculating total null values	47

13	ii	Dataset information	47
14	iii	ET(crop) dependance on max and min temperature	48
15	iv	Proposed model	49
16	v	Gradient boosting regression	49
17	vi	Random forest regressor	50
18	vii	Extreme gradient boosting	50
19	viii	Extra tree regressor	51



## List of Tables

S.No.	Table No.	Title	Page no.
1	1.1	Ault Weather Station Details	5
2	2.1	Comparison of existing approaches for forecasting of crop water demand	14,15
3	3.1	Input variables in our dataset	17,18
4	4.1	Performance of Random Forest against different performance metrics	34
5	4.2	Performance of Extra tree regressor against different performance metrics	35
6	4.3	Performance of XG boosting against different performance metrics	36
7	4.4	Performance of Gradient boosting regressor against different performance metrics	36
8	4.5	Performance of the proposed model against different performance metrics	38
9	4.6	Comparison of performance of various models with the proposed model	38

## Abstract

Due to growing competition for water and scarcity of water, effective water utilization for agricultural uses is essential. Irrigation technology advancements have aided in the development of irrigation planning approaches that minimize water requirements while having little impact on crop quality and yield, hence boosting food security. Crop water requirements must be estimated while designing agricultural and irrigation activities. The purpose of this study was to assess the water needs for wheat crops[10]. The  $ET_0$  and crop coefficient-based crop water requirements, as well as the corresponding irrigation water requirements, can be estimated by analyzing the effective precipitation and crop water requirements. At the CoAgMet station's daily data availability[9], reference evapotranspiration ( $ET_0$ ) was computed using weather information from the maximum and minimum temperatures, average temperature, humidity level, wind speed, solar radiation, and precipitation. Furthermore, the actual evapotranspiration  $ET(\text{crop})$  of the wheat crops produced in the area was calculated by dividing  $ET_0$  by the crop coefficient ( $K_c$ ) value. The irrigation water required for wheat crops was then estimated at points on a grid in various places. This research will help water management systems provide the optimal amount of water required for agriculture while also enhancing irrigation management. The proposed model's results were assessed using a variety of performance metrics, including  $R^2$ , MAE, MSE, RMSE, and MAPE. Using the above-mentioned performance indicators, we compared the suggested model's efficiency against those of the existing models and discovered a significant improvement in performance.

## **Chapter - 1 INTRODUCTION**

### 1.1 Introduction

A crop's evapotranspiration is the quantity of water required by the crop for normal growth. We must calculate evapotranspiration to determine how much water a crop loses in a single day and then calculate how much water we must supply to make up for that loss. Wheat is considered to be one of the world's oldest and most extensively grown food crops, having been cultivated around 10,000 years ago in the Middle East's Fertile Crescent. It was domesticated around the same time as rice but slightly before maize. Wheat, rice, and maize are the three largest staples of the world, accounting for roughly fifty percent of the world's food calories and two-fifths of its protein consumption[3]. Wheat alone contributes just over one fifth of all the world's dietary calories and protein, making it particularly crucial for guaranteeing global food/nutrition security. We will utilize machine learning techniques to forecast the amount of water needed to compensate for water loss due to natural phenomena such as transpiration and evaporation. Evaporation is the process by which soil loses moisture in the form of water vapors, whereas transpiration is the process by which moisture is lost from the surface of crop leaves. In accordance with WaterAid India's 'Beneath the Surface: The State of the World's Water 2019' study, one kilogramme (kg) of wheat requires 1,654 liters of water. Around 217 million hectares farmed annually, it is the most frequently harvested crop in the world.

With sustained global population expansion and a growing demand for wheat-based processed goods in the global South, increasing wheat productivity is critical to ensuring global food security. Population expansion is putting enormous strain on the world's scarce freshwater supplies. The growing need for water coming from the private and industrial sectors is also a

result of population growth. As a result, available water is becoming increasingly scarce, necessitating efficient irrigation water utilization in agriculture. With current water sources practically gone, there is an urgent need to boost water output through effective irrigation. Typically, 15% of given water gets wasted during transportation, another 15% is lost when feeding fields via farm channels, 25% is lost owing to poor water consumption, and only 45% of supplied water is used for farming and agricultural labor. The timing, length, and method of irrigation utilized all influence the effectiveness of irrigation water consumption. Efficient crop irrigation management necessitates information from numerous sources, including soil, crops, and the atmosphere. Evapotranspiration (ET) is the most significant component of the hydrologic cycle, and correct forecasting of ET is critical for a range of applications such as agricultural yield modeling, precise demand for water, irrigation system design, and resource planning and management. Accurate agricultural ET calculations will also be required to improve irrigation system efficiency.

Penman-Monteith developed a formula in 1954 to quantify agricultural water demand based on minimal meteorological circumstances including relative humidity, wind speed, maximum and minimum temperatures, and precipitation[11]. According to Penman, every crop has its crop coefficient, which means we have different crop coefficients. So, in this project, we used climatic data to perform machine learning forecasting of crop water demand. Penman's formula was later applied in various projects involving complex IoT systems, machine learning, and deep learning to forecast the future water requirement of a specific crop.

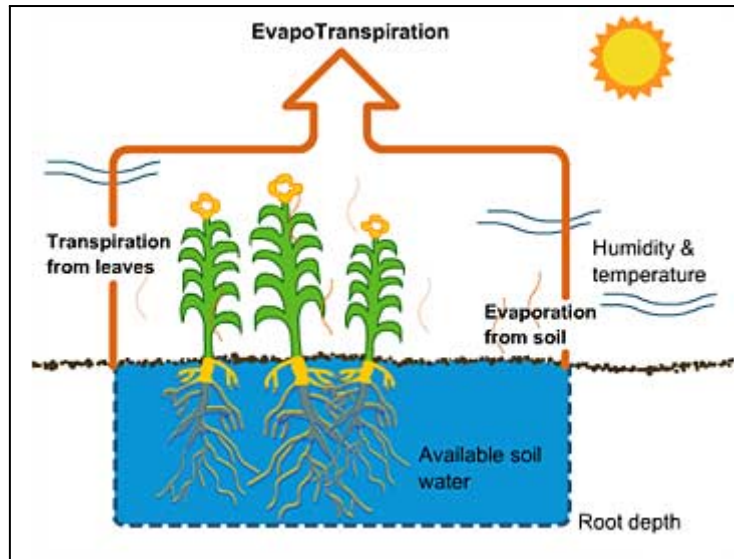


Fig 1.1 Process of evapotranspiration

When employing water balance techniques for irrigation management, however, the effective values of inflows and outflows are always assumed to be known. For instance, if the total amount of rainfall for a certain site has been determined, the effective amount that's going to infiltrate the soil must be computed. If the irrigation system is inefficient and irrigation is not supplied directly at the area where soil moisture measurements are made, the fraction of the overall irrigation amount which contributes to the soil's moisture replenishment must be known. A portion of the applied water can flow off on any particular day prior to infiltrating the soil profile, thanks to quick drainage and surface sealing. Similarly, some water may be lost as deep percolation.

Water demand estimation can be divided into two approaches: conceptual theory and system theory (Pulido-Calvo et al. 2009[10], Zhou 2002[5], Alvisi 2007[3]). Conceptual models forecast irrigation water requirements based on a variety of factors including soil moisture, infiltration, and evaporation. These criteria are used by irrigation managers for estimating the water needed for irrigation for the whole season. Nevertheless, estimated water demand at the

start of the watering season may not accurately reflect true water consumption for a variety of reasons. The system theory technique is the second approach to calculating water demand. In this method, the model undergoes training on available data before being utilized to forecast future water demand. System-theoretic approaches outperform conceptual approaches in terms of efficiency and accuracy (Pulido-Calvo 2008)[9]. Furthermore, it is based solely on publicly available data.

Evapotranspiration (commonly abbreviated ET) is the quantity of water lost through plants and the ground as a result of meteorological conditions. ET is calculated as the total of both transpiration and evaporation. Evaporation is the transfer of water from the earth's surface to the atmosphere. Transpiration is the loss of water from the plant's leaves and body to the air. The amount of evapotranspiration is affected by temperature, humidity, wind, amount of sunlight, and plant type. ET is measured in inches of water. ET refers to the amount of water lost to the atmosphere; this water is replaced through precipitation, rain, and irrigation. As a result, the places with most demand for water for crops are those that are hot, dry, windy, and sunny. When it is chilly, moist, and cloudy, with no or a slight breeze, the lowest readings are found. It is obvious from the information provided above that a single crop cultivated in different temperature zones will have variable water requirements. A given kind of maize produced in a cold region, for example, will use a smaller amount of water daily than an identical variety of maize planted in a warmer climate. It is so useful to select a standard crop or reference crop to figure out how much water that crop requires per day in various regions. The benchmark or reference crop was chosen to be grass.

The efficiency of irrigation water consumption is determined by the time, length, and technique of irrigation employed. To effectively manage crop irrigation needs, information from numerous sources, such as soil, crop, and

atmosphere is required. ETc estimation is also one of the key elements employed in irrigation planning, design, and operation (Rowshon et al., 2013)[17]. Jensen et al. (1990)[13] gives extensive assessments of commonly employed methods for predicting plant water requirements and determining transpiration. In our project we have collected the data CoAgmET, stands for Colorado's Mesonet, it is an organization of Colorado State University for collecting daily data of different regions in Colorado[9]. We have worked on the dataset of the winter wheat crop and this data is collected by Ault weather station.

Table 1.1: Ault Weather Station Details

Station ID	ALT01
Location	1 mi SE Ault.
Latitude	40.5690
Longitude	-104.7200
Elevation	4910 ft.

## 1.2 Problem Statement

People utilize incorrect irrigation methods due to a lack of information about water management and forecasts. This wastes readily available fresh water. Farmers are also not getting a good yield. In this age of overcrowding, we must use scarce freshwater resources responsibly. To optimize water management and irrigation systems, one should understand agricultural water demand, or how much water a crop requires during its growth phase.

There are various types of irrigation systems available today, but one must determine which irrigation system to use to get a good crop yield by efficiently using fresh groundwater and rainwater, such as drip systems, center-pivot irrigation, sub-irrigation, spray irrigation, and so on. To select the appropriate irrigation system, we must first understand the local climate. After we have this climate-related dataset, we use machine learning algorithms to effectively forecast the water requirements of a specific crop (ET crop). Understanding of (ET crop) will offer us the greatest notion of which irrigation system we should use in our location to produce the highest yield of crops while utilizing the least amount of fresh water. The paucity of pure water supplies around the world has created a demand for their best use.

### 1.3 Objectives

Our project's main goal is to anticipate crop water demand (ET crop) based on minimal provided meteorological variables such as the highest and lowest temperatures, humidity, solar radiation, wind speed, gust speed, and precipitation of a certain region. Crop water need refers to the efficient water supply required for the highest possible yield during the development phase. Following data visualization and standardization, several machine learning algorithms are implemented on the collected dataset, and the best four models developed using machine learning out of all implemented models are ensemble to forecast the result and provide the best prediction of ET crop. The more precise our forecasting, the more effectively and methodically planning for agriculture can be carried out. Water needs can be utilized to determine which irrigation method an individual should use for higher crop yield while using fresh water more efficiently and properly. The proposed



method employs stacking-based ensemble learning to increase classifier diversity. The suggested framework's performance is compared to current literature using the coefficient of determination (R<sup>2</sup> score), mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE).

#### 1.4 Methodology

In this section, we will go over the project methodology step by step. I noticed relationships between other factors after deleting outliers and thoroughly analyzing the data. The data was then divided into two parts. 70% of the data is training data, and 30% is testing data. I incorporated all of the models listed above in the background and preliminary part, and after examining all predefined models, we discovered that these models helped enhance overall accuracy. Given the literature study on agricultural water demand prediction, the suggested ensemble framework naturally incorporated the majority of authors who achieved substantial accuracy in their machine learning models.

The experimental location (40.5690°N, -104.7200°E) is in Ault, Colorado, USA, at the ALT01 weather station[9]. First, we measured how multiple parameters affect crop water requirements. We only evaluated relative humidity, precipitation, solar radiation, maximum temperature, minimum temperature, and wind speed. Crop factors such as crop density, kind, crop height, and leaf breadth have also been excluded. We collected winter wheat data from March to December for three consecutive years, from 2020 to 2022.

We began with the first phase in machine learning, which was data collecting and processing. We gathered information on the winter wheat crop and meteorological conditions. Three years of data were not available at the exact

same time. So we gathered information for three years in a row and compiled it. Table 3.1 displays a list of all the variable names and types of input variables, together with their measuring units, that are taken into account throughout the prediction process. According to this data, ET crop is our expected target variable. We processed our dataset after obtaining the necessary data. We calculate the total number of fields with null values for this purpose. After counting the null values we filled the null values with the average value of that column.

Following preprocessing, the following step we carried out was feature selection, which involved determining the right correlation between various variables. We used heat map visualizations to identify the relationship among different variables and visualization techniques to comprehend how some of the features are purposefully affecting our target variable, ET crop, but others are all independent and no short-term immediate impact of that variable can be observed. These visualization results showed us how some of the variables ought to be removed before running regression models on our dataset.

Crops' ideal growing conditions depend heavily on climate when it comes to their irrigation needs; that is, how much water they require to thrive: In hot, sunny environments, crops will need more frequent watering compared to cooler, cloudier settings. Crops like rice or sugarcane need more water compared to beans or wheat. Crops that are fully developed require more water compared to recently planted crops. Crops can receive the required water amount through a combination of irrigation and rainfall. Rainwater is usually supplemented or complemented by irrigation water. All the water that crops require must be supplied by irrigation only in desert or dry areas or during times of drought.

The irrigation water needed is the difference between the crop water requirement and the amount of rainfall that the crops can use (effective rainfall). The computation of irrigation water requirements serves as the foundation for developing an irrigation plan (often by an agronomist) and constructing an irrigation scheme, such as channel dimensions (typically by engineers). The following phase was model selection. We used the decision tree regressor, random forest regressor, extra tree regressor, gradient boosting regressor, KNN (k - closest neighbor), and linear regression models. We chose the best four models and then combined them to build a proposed model for more accurately predicting crop water requirements.

The supervised learning technique includes the well-known Random Forest machine learning algorithm. It applies to both ML classification and regression issues. The concept of ensemble learning is utilized, where multiple classifiers are combined to address complex problems and enhance model accuracy. By using multiple decision trees taken from diverse parts within an inputted dataset, rather than just one tree alone, a random forest classifier can improve overall prediction accuracy—hence its name. The prediction of final output by Random Forest depends on the majority vote of predictions from each tree.[15].

AutoML tools are used to train Extra Trees, which is a supervised machine learning method based on decision trees and also known as Extremely Randomized Trees. It may be faster than Random Forest, but it is comparable. The Extra-Trees technique generates fewer decision trees than the Random Forests algorithm, however, each tree uses random samples without any permutation[15]. A special sample data set is produced for each tree. A specific number of features are randomly chosen from the entire set of

features for each tree. Extra Trees' most significant and distinctive aspect is the random selection of feature split values. Instead of computing the locally best value for splitting the data using Gini or entropy, the algorithm randomly selects a split value. Because of this, the tree is diverse and uncorrelated. Gradient Boosting is a technique that builds models in a sequential manner while attempting to minimize the error rate of the prior model.

Regression or classification can both be performed using decision trees. The way it operates is by creating a tree-like division of the data into smaller parts. When a set of features' output value is anticipated, the portion that contains the set of features corresponds is taken into consideration. After assessing the effectiveness of the model using the ensemble technique, various evaluators are used to gauge its performance. To assess our models, we have employed MSE, MAE,  $R^2$ . With our suggested ensemble model, we have contrasted various models[15].

## **Chapter - 2 LITERATURE REVIEW**

Water managers have a huge obligation to use the water effectively because agricultural irrigation uses a lot of it. The fields lose a significant amount of water through evaporation and transpiration. It is challenging to distinguish between the two processes since transpiration and evaporation frequently occur simultaneously. This section covered a variety of earlier research projects and how they tackled the issue.

The integrated qualitative and quantitative approaches to comprehend socio-cultural behavior and environmental elements influencing agricultural water demand, quality, and crop sustainability were described by Mohamed Ali Abunnour et al. in 2016[2].

The process by which water evaporates from the ground's surface by evaporation as well as from the crop through transpiration together is known as evapotranspiration. For bumper crop production in typical conditions, the crops need a specific amount of water to make up for water losses through evapotranspiration.

Several authors have put forth different methods for forecasting crop water demands. In this case, a succinct analysis of a few significant contributions to the body of literature is provided.

In their 2015 study, Shafika Sultan Abdullah et al.[4] used data from a region in northern Iraq called Nineveh Governorate, with a total size of 37,323 km<sup>2</sup> and a height of 222.6 m above sea level. Weather data from the main meteorological station in Mosul (Global Station Code 608), containing daily averages of maximum air temperature, minimum air temperature, relative humidity, radiation hours, and wind speed from 1980 to 2005, was analyzed

using the quick and reliable extreme machine learning algorithm. As an evaluation standard, the  $R^2$  and RMSE coefficients of determination were applied.

In 2017, Siraj Sheko et al.[6] reviewed the Water Use Efficiency, Irrigation Frequency, and Crop Water Requirement Estimates for the Production of Cabbage. The Journal of Geoscience and Environment Protection discussed the need to plan irrigation in order to efficiently utilize limited water resources for crop production. He also gave a concise summary of how machine learning can be used to predict plant water needs and how we can utilize it for better water management.

In their article titled "Water demand forecasting using machine learning methods: A case study of Beijing-Tianjin-Hebei Region of China," Qing Shuang and Rui Xhao (2021)[7] used a variety of models, including statistical models for linear regression, ridge regression, and bayesian ridge regression. SVM, decision trees, and ML techniques were also employed by him. As part of his ensemble approaches, he also suggested using random forests and adaboost, as well as 10-fold cross validation for the data used for training to avoid overfitting situations. Extreme water shortages exist in the Beijing-Tianjin-Hebei region of China. Making more effective use of available water resources can be accomplished with the use of water demand forecasting.

P. Mohan et al. (2018)[8] from REVA University used a variety of data mining techniques, including ANN, K-means, SVM, and others. Agriculture system models employ robust estimate approaches such as SCG and (BFGS) QuasiNewton, both of which are based on neural network algorithms. Predict the soil moisture content using this algorithm to control farm irrigation. Data

mining and deep neural network methodologies were used. He also suggested using a different layered model to forecast crops. Accuracy, precision, recall, sensitivity, and specificity are some of the evaluation measures that are used. The suggested approach is contrasted with the following approaches already in use: SOM-DNN, SOM-KNN, weighted-SOM-KNN, Random Forests - Multiple Linear Regressions, and SOM Learning Vector Quantization (LVQ). This machine-learning technique was examined by J.H. Jeong, J.P. Resop, N.D. Mueller, D.H. Fleisher, K. Yun, E.E. Butler, D.J. Timlin, K.M. Shim, J.S. Gerber, V.R. Reddy, and S.H. Kim.

Studies that have already been done have predicted water demand using one or more models. For the first time, this work provides a comprehensive comparative analysis of multiple machine learning and statistical models incorporated in IPS and EPS. The IPS uses training and test data from the same time period, while the EPS constructs models using historical training data and applies them to predict future water demand.

In their experimentation, Sidhu et al. (2020)[1] utilized seven distinct machine learning algorithms, including SVM, decision tree, random forest regressor, extra tree regressor, gradient boosting regressor, adaboost regressor, and neural network, to obtain results. The most accurate model, he found, was the adaboost regressor. Utilizing ten folds, we took it a step further in reducing overfitting in our model, following his three-fold strategy that impacted our study. He utilized  $R^2$ , accuracy, and mean square error as a means of evaluating performance.

Direct measuring methods are ineffective for determining evapotranspiration across vast irrigated regions. They are mostly utilized for research by skilled professionals. Evapotranspiration is widely assessed using an array of methods

that need climatic parameter observations. Ravneet Kaur Sidhu evaluated the effectiveness of several proposed models using MSE, MAE,  $R^2$ , and explained variance scores. The use of n-fold cross validation in Beijing-Tianjin (2021) influenced our research effort, allowing us to remove model overfitting. Our study advances crop water demand forecasting because we assembled the most effective model for making more precise predictions.

Table 2.1 Existing approaches for crop water demand forecasting

S. No.	Author	Approach	Dataset Used	Performance Parameters used
1	Ravneet Kaur Sidhu et al. (2019)[1]	SVM, Decision tree, Random Forest	CSSRI	Accuracy, Recall
2	Qing Shuang et al. (2021)[3]	Linear Regression, SVM, Decision tree, Random forest, Adaboost	Annual water resources report, China statistics yearbook	$R^2$ , MAE
3	Abdullah et al. (2015)[4]	Extreme learning Machines, ANN	Weather station in Mosul (609)	$R^2$ , MSPE
4	Abunnour et al. (2016)[2]	K-NN, Decision tree, Random Forest		Accuracy, ROC-AUC
5	Ravinder Kumar et al. (2019)[5]	LSTM neural network	CSSRI	TOPSIS
6	P. Mohan et al.(2018)[6]	ANN, KNN, K-means	Raitamitra-Karnataka State Department of Agriculture (KSDA)	Accuracy, recall, precision, sensitivity



7	Dean C. J. Rice et al. (2017)[8]	Linear Regression, Quotient Method, Feed-Forward Neural Network	Local greenhouse operation	RMSE, NRMSE
8	Pulido-Calvo et al. (2008)[10]	Feed-Forward CNN, Genetic algorithm	-	Accuracy
9	MA Khan et al. (2011)[11]	Decision Tree	CICL water delivery statements, Weather stations in New South Wales	Accuracy
10	Sara Sadri et al. (2022)[13]	Random Forest algorithm	SMAP Level-3, SMAP Level-4, MODIS, MSWEP V280	R <sup>2</sup> , RMSE, KGE indicators
11	RG Perea et al. (2015)[14]	ANN	Agroclimatic station in BMD, Andalusia (Southern Spain)	R <sup>2</sup> , SEP

## Chapter - 3 SYSTEM DESIGN AND DEVELOPMENT

### 3.1 Model Development

The development, design, and analysis of machine learning algorithms are discussed in this section. With the help of CoAgMet, our data was prepared. A network of agricultural weather stations is operated by CoAgMET throughout the state of Colorado (USA)[9]. All the variables used in our experiment to predict the (ET crop) are shown in Table 3.1. One can utilize the data obtained from the weather station to compute the ET crop values and simulate the water consumption of different crops.

The proposed framework for efficient forecasting of crop water demand is depicted in Figure 4.2. After analyzing the data thoroughly and discovering the relationship among the parameters, we removed all the outliers. After partitioning the data, we have 70% of instances for training data and 30% for testing data. The overall accuracy was increased with the help of the baseline predefined machine learning models that we included. To evaluate their models, the majority of other authors who have worked on a similar project have taken into account R<sup>2</sup>, MSE, and MAE. A significant increase in accuracy can be achieved using various bagging and boosting algorithms. Apart from LightGBM and XGBoost, other algorithms such as gradient boosting, decision trees, and random forests were also considered when selecting models for a stacked ensemble-based regressor.

#### 3.1.1 Experimental Analysis

The data was retrieved from CoAgMET which operates weather stations for agricultural purposes around the state of Colorado[9]. The data from these

stations are used to calculate and predict the Evapotranspiration (ET) values to model the water use for various crops. The data that we have selected for our experiment is taken from Ault station as shown in table 1.1. Table 3.1 shows the variables in our dataset contains a total 12 number of input variables which includes  $ET_r$ , Hourly evapotranspiration,  $ET_0$ , evapotranspiration of crop i.e.  $ET(\text{crop})$ . In this project we are using these given parameters to predict the value of evapotranspiration of winter wheat  $ET(\text{crop})$  which depicts the amount of water needed for the optimal growth of the crop. All these provided variables play an important role to determine the  $ET(\text{crop})$ . So,  $ET_{\text{crop}}$  is our target variable which we will be predicting with the help of other given parameters.

Table 3.1 Input variables in our dataset

Attribute	Description
Avg temp	Average of all the daily temperatures recorded during the observation period
Max temp	Maximum temperature limit for optimal growth
Min temp	Minimum temperature limit for optimal growth
$RH_{\text{max}}$	Maximum relative humidity on a particular day
$RH_{\text{min}}$	Minimum relative humidity on a particular day
Solar radiation	Different solar radiation amounts
Precipitation	Amount of water deposited in the atmosphere
Wind run	Wind speed
Gust	Measured brief increase in wind speed
$ET_r$	Relative evapotranspiration rate

$ET_0$	Reference evapotranspiration rate
Hourly ET	Hourly evapotranspiration rate
$ET_{crop}$	Amount of water required for optimal growth

We have followed each step of the machine learning algorithm for implementing all our baseline machine learning algorithms as shown in Figure 3.1. Figure represents the complete pipeline of machine learning. We have also implemented our proposed stacked model using these steps as it is represented in the given Figure 3.3.

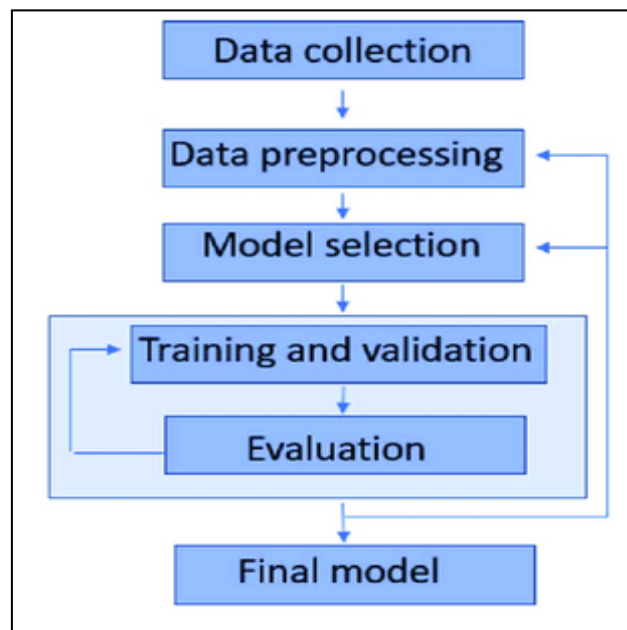


Fig. 3.1 Machine learning pipeline for model building

Our proposed methodology, which is divided into three different phases, is described by the algorithm below.

Phase 1: Preprocessing phase

Phase 2: Training phase

Phase 3: Testing phase

Data mining is the responsibility of the preprocessing phase. To discover various insights from the data and implement necessary modifications, data visualization is performed. At this stage, data distortions and class imbalances have been fixed and duplicate rows have been removed as an example. During the training phase, the framework is built by finding the best accuracy-based models and learning all the weights and parameters required for making accurate predictions. Our proposed clustering framework is analyzed by evaluating the parameters and weights obtained on unseen data during the testing stage, which is the last stage.

Pseudo code : Phase 1

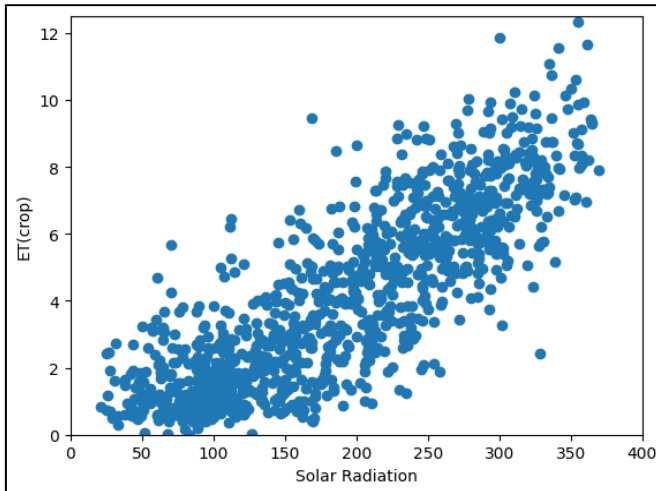
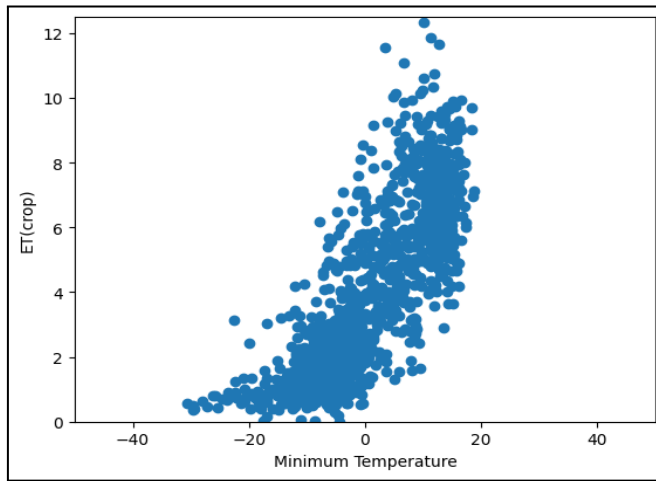
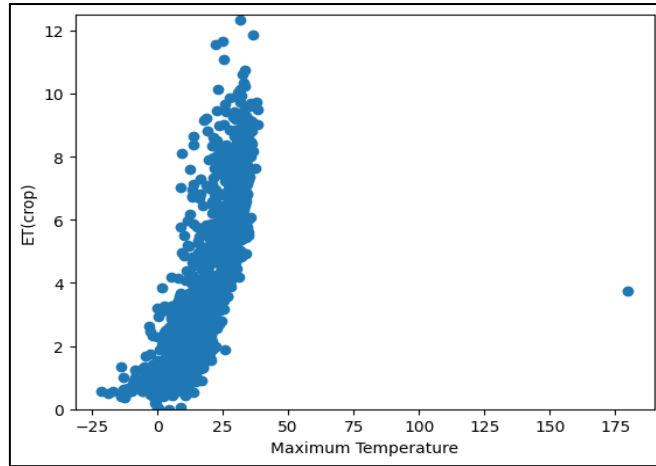
The chosen dataset is sourced from CoAgMET, which runs a network of weather stations for agricultural weather in Colorado as Step 1.

In order to decrease bias in predictions, the dataset undergoes processing by eliminating duplicate entries, as Step 2.

Using the mean value of the specific variable, fill all empty fields in the fields with null values in Step 3.

The data that has been pre-processed is split into training and testing datasets as step 4.

Figure 3.2 shows that in the preprocessing phase we have done some visualization to understand the relation between the various parameters and correlation between them. Further these correlations will help us in predicting the Evapotranspiration more accurately.



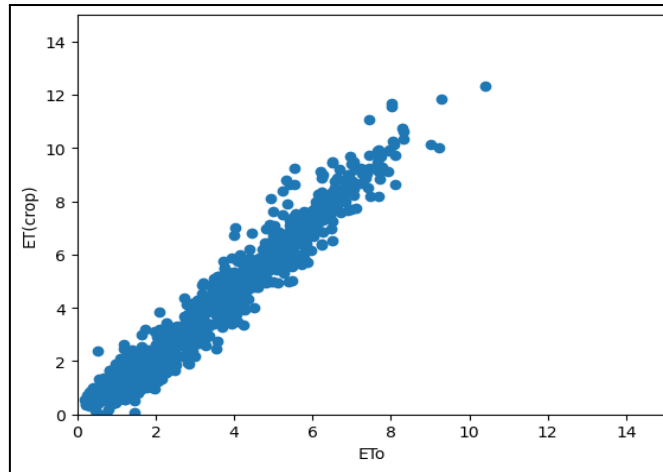


Fig. 3.2 Graph between ET(crop) and other variables.

**Training Phase** is responsible for training the dataset and judging the algorithm. We have chosen 70% of data for the testing purpose. Different models were trained on this data. The performance matrix can be of users choice, in this project we have considered coefficient of determination as our major performance matrix based on which we have made our model go through.

ML algorithms with best performance were taken into consideration which then were stacked together to create a better model for better prediction. As a result, the model learns from the data to perform a set of tasks. Over time, with training, the model gets better at predicting.

Pseudo Code : Phase 2

The first step involves training various machine learning algorithms. For stacking purposes, the four most accurate models are chosen based on the coefficient of determination ( $R^2$ ).

To perform stacking, the four most accurate models are stacked together in step 2.

**Testing Phase:** During the machine learning process's testing phase, the stacked model parameters are tested on a 30% testing dataset. The accuracy is first tested and then a regression report is generated. In the further sections of the report, we discuss that our proposed algorithm has a  $R^2(\%)$  of 96.55%. The baseline machine learning models were outperformed.

Pseudo Code : Phase 3

The weights and parameters are trained on an unseen dataset  $Y$ , and the obtained results are based on the value of  $R^2$ .

In addition to the coefficient of determination, various performance metrics are utilized to compare our outcomes in step 2.

The depth or amount of water needed to compensate for water loss through evapotranspiration is known as crop water demand (crop ET). The amount of water required for optimal growth varies among different crops, in other words. Crop water needs always pertain to a crop that is cultivated under ideal circumstances, meaning a uniform crop that is actively growing, completely covering the soil, free from diseases, and under favorable soil conditions (which include fertility and water). In the given environment, the crop achieves its maximum production potential.



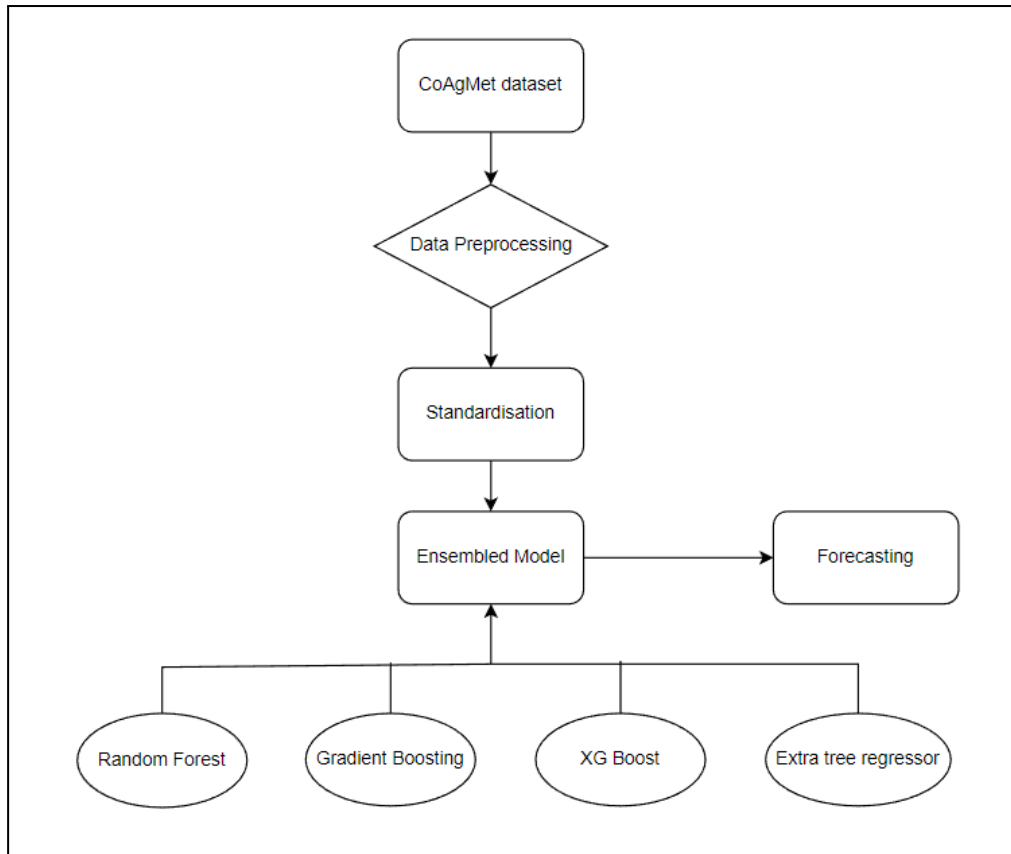


Fig. 3.3 Flow Diagram for the proposed methodology

The most popular strategy for estimating the crop is the Penman-Monteith method. Determining the evapotranspiration reference (ET<sub>ref</sub>) is necessary before determining the rate of evapotranspiration or plant water demand rate for a specific crop. Then, we use the crop coefficient product (K<sub>c</sub>) to calculate the amount of evapotranspiration in order to determine what real crop water demand is.

$$ET_c = ET_{ref} * K_c$$

where,

ET<sub>c</sub> - Evapotranspiration rate ( crop water demand), unit : (inches/day)

ET<sub>ref</sub> is alfalfa - reference evapotranspiration rate, unit : (inches/day)

K<sub>c</sub> is crop coefficient

The essential meteorological variables that are measured by a weather station are used to produce  $ET_{ref}$ , or lucerne reference evapotranspiration (inches/day).

Typically, the crop coefficient, known as  $K_c$ , is established through experimentation for a specific crop. The  $K_c$  values reflect the combined impact of alterations in crop characteristics such as leaf area, plant height, crop development rate, irrigation method, crop planting date, degree of canopy cover, canopy resistance, soil and climate conditions, and management practices. For each agronomic crop, there exists a unique set of crop coefficients that can be used to predict varying rates of water consumption during different growth stages. Figure 3.4 shows a  $K_c$  curve example plotted against days or weeks after planting.

In general, there are four key growth stages for crops: early, crop development, midseason, and late season. Each of these phases has a different duration, which is influenced by the temperature, latitude, height, planting date, crop type, and cultural practices. The best way to determine crop growth stage is by local field observations, thus modify the theoretical  $K_c$  values accordingly.

During the crop germination and establishment phase at the start of the growing season, the majority of evapotranspiration takes place through soil surface evaporation. Evaporation from the soil surface decreases as the crop canopy develops and covers the soil surface, leading to an increase in the transpiration component of evapotranspiration. During the initial  $K_c$  stage, when the plants are small, both the water consumption rate and  $K_c$  value are low. However, as the plant develops, the crop ET rate increases, as shown in Figure 3.4.

The maximum ET rate for agronomic plants occurs when the crop is fully developed (mid-season  $K_c$ ). As the plant completes its growth phase and attains physical maturity ( $K_c$  end of season), the rate of ET starts to fall once more.

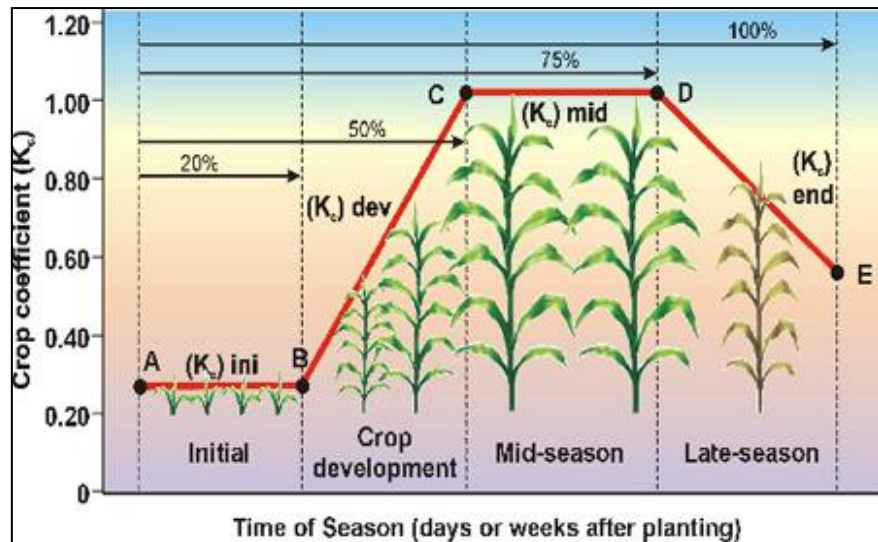


Fig. 3.4 Schematic representation of crop coefficients at different growth stages

In this project various machine learning models are used in the proposed framework. Some extra existing models are also trained and used to measure the performance difference w.r.t. the proposed framework.

#### 1. Random Forest Regressor[15]

The supervised learning approach includes the well-known machine learning method random forest. It can be used to resolve classification and regression ML issues. Its foundation is the concept of ensemble learning, that combines different classifiers to handle complex problems and improve model functionality.

According to what its name implies, "a random forest is a classifier that takes a set of decision trees over different subsets of a given dataset and averages them to improve the prediction accuracy of that dataset." Instead of depending on decision trees, Random Forest utilizes forecasts from each tree to estimate the final outcome based on the forecasts' majority vote. The more trees in the forest, the more accurate and less prone to overfitting problems the model is.

## 2. Extra Tree Regressor[15]

A decision tree-based ensemble supervised machine learning method called Extra Trees, commonly referred to as Extremely Randomized Trees, is trained using AutoML tools. It is comparable to Random Forest but perhaps faster.

The Random Forests algorithm produces more decision trees than the Extra-Trees technique, but there is no permutation and the samples in each tree are random. For each tree, this produces a special sample data set. Additionally, a predetermined number of features from the whole collection of features are randomly selected for each tree. The most important and distinctive feature of Extra Trees is its random choice of feature split values. Instead of determining the locally optimal value for separating the data using Gini or entropy, the algorithm chooses a split value at random. The result is a diversified and uncorrelated tree.

## 3. Gradient Boosting[15]

This algorithm's fundamental principle is to build models in a sequential manner, each model attempting to lessen the error of the one before it. To accomplish this, a new model is constructed using the residuals or errors from the prior model.

Use the gradient boosting regressor if the target columns are continuous; if the

classification is challenging, use the gradient boosting classifier. The "loss function" represents the only distinction between the two. By using gradient descent and weak learners, we want to reduce this loss function. Since it is based on a loss function, classification and regression problems each have a unique loss function, such as mean squared error (MSE).

#### 4. Extreme Gradient Boosting[15]

Gradient Boosted decision trees are implemented using XGBoost technology. In this method, decision trees are created in a sequential manner. Weights matter a lot with XGBoost. Before being put into the decision tree that predicts results, every variable that is independent is assigned a weight. Following that, the variables are fed into the following decision tree with enhanced weights for parameters that the tree misjudged. Then, a robust and precise model is created by combining these various classifiers/predictors. It can be applied to the resolution of issues with regression, classification, ranking, and personalized prediction. Figure 3.5 depicts the application of XG boosting.

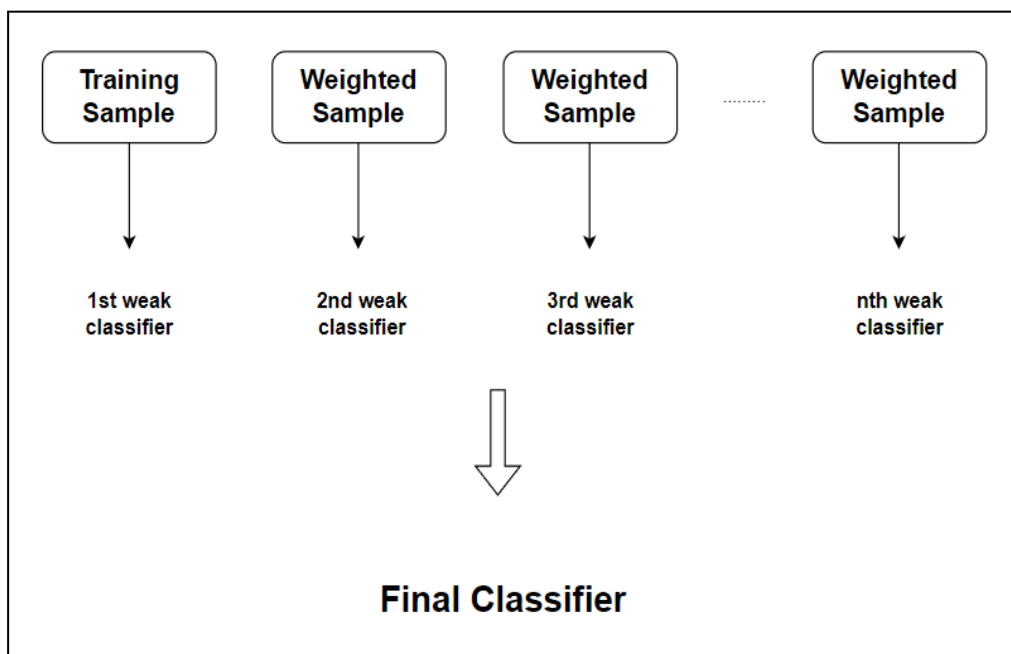


Fig. 3.5 Implementation of XG Boosting

## 5. Decision Tree Regressor[15]

Decision tree regression trains a model in the form of tree structure to forecast data in the future and generate useful continuous output by observing the properties of an item. Continuous output denotes the absence of discrete output, i.e., output that is not only represented by a discrete, well-known collection of numbers or values.

A weather forecasting model that forecasts whether it will rain or not on a specific day is an example of a discrete output.

A profit forecasting system that estimates the likely profit that may be made from selling of a product is an example of a continuous output.

## 6. AdaBoost Regressor[15]

The first truly successful boosting algorithm created for binary classification was called AdaBoost. Multiple "weak classifiers" are combined into a single "strong classifier" using the boosting technique known as "AdaBoost," which stands for "Adaptive Boosting." Yoav Freund and Robert Schapire created it. For their work, they were also awarded the 2003 Gödel Prize.

## 7. K Neighbours Regressor[15]

By averaging each observation in the same neighborhood, the nonparametric KNN regression approach approximately measures the relationship among independent variables and the continuous result. The analyst must decide on the neighborhood's size, or cross-validation can be used to determine the size that reduces mean-squared error.

## 8. Ensemble Learning[16]

Ensemble learning is learned by running the Basic Learner multiple times.

Final votes are cast on hypotheses and final weights are placed on the "metamodel". Different kinds of ensemble techniques include bagging and boosting.

## Chapter - 4 EXPERIMENTS & RESULT ANALYSIS

In this section, I've compared many machine learning algorithms and described the outcomes of our suggested ensemble model's performance analysis when compared to earlier machine learning models. To enhance the accuracy of agricultural water demand forecasting (ET crop), I combined many machine learning algorithms into one. I've developed a number of well-known baseline models, but for stacking, the best models were those with the highest performance. The prediction becomes more diverse when multiple models are combined.

In order to increase the effectiveness of forecasting crop water demand, I first contrasted the performance of the various available models and then chose the models with the highest performance. I then stacked these models in our proposed ensemble model. We have measured the performance of our model against other current models using R<sup>2</sup>, MSE, MAE, RMSE, and MAPE in order to compare the effectiveness of other machine learning models.

### **Coefficient of determination (R<sup>2</sup>)[1]**

A statistical indicator of how well a regression model fits the data is the coefficient of determination. Its value, which ranges from 0 to 1, indicates how well a model predicts a result. The model predicts the data better the closer R<sup>2</sup>'s value is to 1. A negative coefficient of determination (R<sup>2</sup>) indicates that our model does not adequately account for the selected data. Formula for calculating the value of coefficient of determination is given as :

$$R^2 = SSR / SST = 1 - SSE / SST$$



where,

SST - sum of square total and

SSE - sum of square error.

### **Mean square error (MSE)[1]**

The MSE can be used to assess how much a line of regression resembles a collection of points. The "errors"—the distances among each point and the regression line—are squared to achieve this. Figure 4.1's graphic representation of the MSE error emphasizes the mismatch among the best-fit line or regression line to the actual values. Mean squared error is calculated as the average of the squared errors obtained from the function-related data. While a lower MSE suggests the opposite, a larger MSE indicates a broader range of the data points about the mean. Its formula comes from:

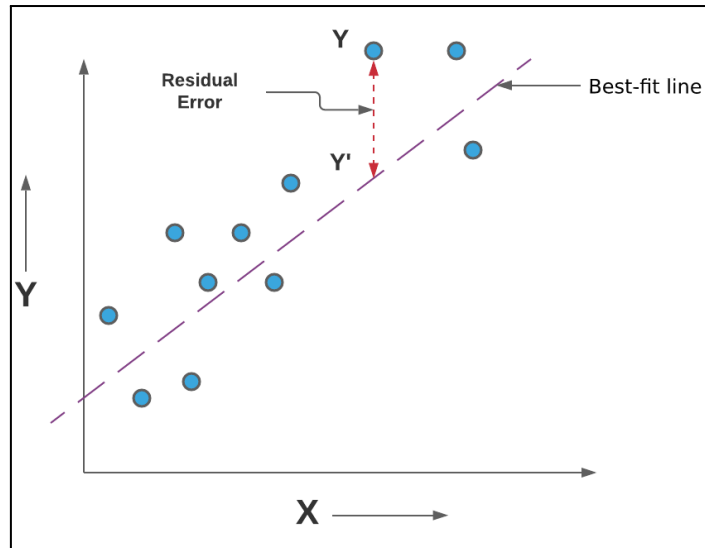
$$\text{MSE} = (1/n) * \Sigma(y - y')^2$$

Where,

n - Total observations

y - actual value of target variable

y' - predicted value of predicted variable



### Mean absolute error (MAE)[1]

To determine how effectively our model is functioning, regression models also employ mean absolute error as an evaluation metric. The mean absolute error of the algorithm in regard to the test set is calculated as the average of the absolute values of each forecasting error for each instance in the test set. Each prediction mistake is represented as the discrepancy between the instance's true value and its expected value. The formula for mean absolute error is given by :

$$MAE = \frac{\sum_{i=0}^n abs(y_i - \lambda(x_i))}{n}$$

where,

$y_i$  - true target value for  $x_i$ ,

$x_i$  - is the predicted value for  $x_i$

$n$  - total test instances.

## Root mean square error (RMSE)[1]

RMSE is the term used to describe the residuals' standard deviation (mistakes in the predictions). The residuals and RMSE both measure how far apart the residuals are from the regression line, which is the separation between the data points. To put it another way, it offers details on how closely the data are grouped about the line of greatest fit. The RMSE is commonly used to validate experimental findings in the fields of forecasting, regression analysis, and weather research. The formula for RMSE is give as:

$$\text{RMSE} = \sqrt{(f - o)^2}$$

where,

f = predicted value for given input,

o = original value

In this project, I employed k-fold cross validation to get around our training model's overfitting. In our project, we used k = 10 for several machine learning models. Training and testing dataset was split into a 7:3 ratio.

### 1. Random Forest Regressor[15]

The supervised learning approach includes the well-known ML algorithm random forest. It can be used to resolve classification and regression issues. Its foundation is the concept of ensemble learning, which combines different classifiers to handle complex problems and improve model functionality. According to its name, "a random forest is a classifier that takes a set of decision trees over different subsets of a given dataset and averages them to improve the prediction accuracy of that dataset." Instead of depending on decision trees, Random Forest employs forecasts from each tree to estimate

the final outcome based on the forecasts' majority vote. The more trees in the forest, the more accurate and less prone to overfitting problems the model is. Table 4.1 shows the different performance metrics that are tested against our random forest regressor model.

Table 4.1 Performance of Random Forest regressor against different performance metrics

S. No.	Performance metrics	Performance measure value
1	MAE	0.3439
2	MSE	0.2243
3	RMSE	0.4681
4	R <sup>2</sup> percentage	96.84
5	MAPE	0.2251

## 2. Extra Tree Regressor[15]

Extra Trees is a decision tree-based ensemble supervised machine learning technique that is trained using AutoML tools. It is similar to Random Forest, but maybe quicker. The Random Forests algorithm produces more decision trees than the Extra-Trees technique, but there is no permutation and the samples in each tree are random. This results in a distinctive sample data set for every tree. Additionally, a predetermined number of features from the whole collection of features are randomly selected for each tree. The most important and distinctive feature of Extra Trees is the arbitrary choice of feature split values. Instead of determining the locally optimal value for separating the data using Gini or entropy, the algorithm chooses a split value at

random. Because of this, the tree is diverse and uncorrelated. Table 4.2 shows the different performance metrics that are tested against our extra tree regressor model.

Table 4.2 Performance of Extra tree regressor against different performance metrics

S. No.	Performance metrics	Performance measure value
1	MAE	0.3498
2	MSE	0.2291
3	RMSE	0.4718
4	R <sup>2</sup> percentage	96.78
5	MAPE	0.1520

### 3. Extreme Gradient Boosting[15]

Gradient Boosted decision trees are embodied in XGBoost. In this method, decision trees are produced in order. In XGBoost, weights matter a lot. Prior to entering the decision tree that predicts outcomes, every variable that is independent is given a weight. The following tree then receives the variables with improved weights for the variables that the first tree misjudged. To create a strong and reliable model, these many classifiers and predictors are then combined. Regression, classification, ranking, and custom prediction are just a few of the issues it can be used to solve. Table 4.3 shows the different performance metrics that are tested against our XG boosting model.

Table 4.3 Performance of XG boosting against different performance metrics

S. No.	Performance metrics	Performance measure value
1	MAE	0.3747
2	MSE	0.2707
3	RMSE	0.5128
4	R <sup>2</sup> percentage	96.19
5	MAPE	0.2415

#### 4. Gradient boosting regression[15]

By integrating weak learners or weak predictive models, the Gradient Boosting technique creates an ensemble model. The various performance indicators that are compared to our gradient boosting regression model are shown in Table 4.4. Models can be trained using the gradient boosting approach for both classification and regression issues. The approach for fitting a model that forecasts a continuous value uses gradient boosting.

Table 4.4 Performance of Gradient boosting regressor against different performance metrics

S. No.	Performance metrics	Performance measure value
1	MAE	0.3516
2	MSE	0.2308
3	RMSE	0.4728
4	R <sup>2</sup> percentage	96.77
5	MAPE	0.2330

Proposed model[16]

In stacking, a sort of ensemble learning, predictions made by base learners are based on those of the meta-learner and are achieved by the union of several algorithms. The suggested approach selects the best models from each algorithm and combines them to provide even greater accuracy. The data is divided between 70% training data and 30% testing data, and it is acquired from the CoAgMet weather station[9].

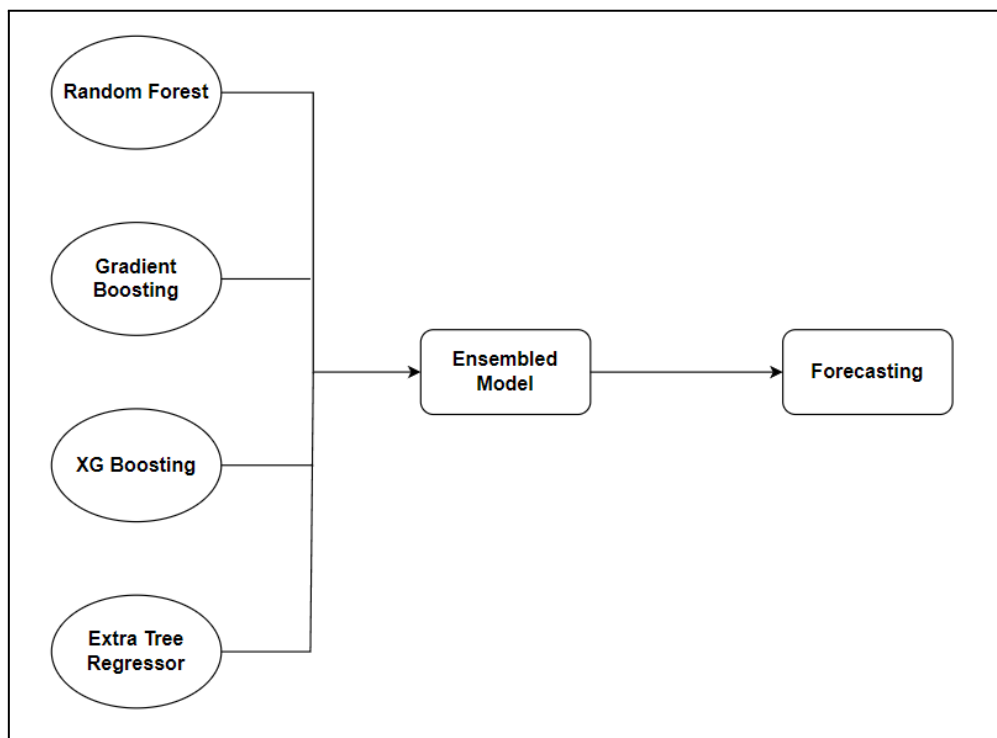


Fig. 4.2 The proposed ensemble framework for forecasting of crop water demand

Figure 4.2 shows the working of the proposed ensemble framework for the forecasting of crop water demand. Models with best performance were selected for the proposed stack based ensemble learning.

Table 4.5 Different performance metrics that are tested against our proposed model.

S. No.	Performance metrics	Performance measure value
1	MAE	0.3421
2	MSE	0.2218
3	RMSE	0.4639
4	R <sup>2</sup> percentage	96.89
5	MAPE	0.2261

Table 4.6 Comparison of the performance of various models with the proposed model

S. No.	Model	R <sup>2</sup>
1	Extra Trees Regressor	0.9648
2	Random Forest Regressor	0.9608
3	Extreme Gradient Boosting	0.9606
4	Gradient Boosting Regressor	0.9624
5	Decision Tree Regressor	0.9412
6	K Neighbours Regressor	0.9082
7	Proposed Model	0.9655



Table 4.6 shows the comparison of various pre-existing models with our proposed model. It can be clearly seen that our proposed model has shown slight increase in the overall performance.

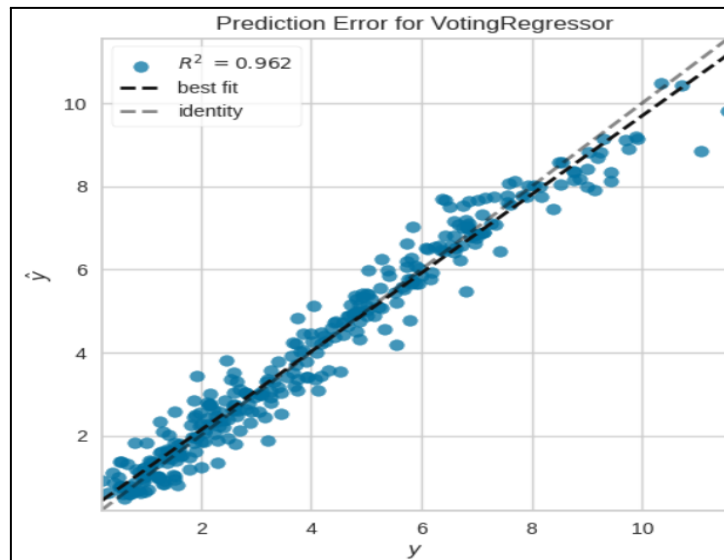


Fig. 4.3 Measure of how well our proposed framework predicts the response variable.

## Chapter - 5 CONCLUSION

### 5.1 Conclusions

This study offers a computational framework based on machine learning for accurately forecasting crop water needs. The suggested framework was successful in predicting plants' needs for water for the best possible growth. Additionally, ensemble learning promotes diversity among foundational learners, improving prediction accuracy. Utilize the coefficient of determination ( $R^2$ ), mean square error (MSE), mean absolute error (MAE), RMSE (Root mean square error), and mean absolute percentage error (MAPE) to assess the effectiveness of each strategy. A single reference data set is used for both training and testing all algorithms. When developing tools for forecasting crop water demand, the majority of current strategies neglect the importance of taking into consideration the diversity of data and skewed data.

I developed a number of baseline models, however for stacking reasons, the most accurate models were selected since ensemble modeling increases the variety of the forecasts.  $R^2$  is the metric I initially used to evaluate models. Table 4.6 displays the coefficients for all the models that were used. To provide even better predictions, the top four performing models - extreme gradient boosting, random forest regressor, gradient boosting regressor, and extra tree regressor are stacked. In addition, I evaluated the algorithm's performance using a number of metrics, such as  $R^2$ , MAE, MSE, RMSE, and MAPE. The comparison of several machine learning algorithms with the suggested approach is depicted visually in Figure 5.1. We used these models of regression to achieve a 96.55%  $R^2$  percentage. For comparisons with other algorithms, we also calculated MSE, MAE, and RMSE, as given in table 4.5. We added diversity and multiplicity to our model with the aid of an ensemble

model. Additionally, stacked-based models add assortment, which means there is a chance that another model used in the ensemble will correctly determine the same feature if an individual model makes an inaccurate prediction about it. The evapotranspiration (ET) of crops could be predicted with the use of the proposed framework. The ensemble learning increases the variety of the base learners, which enhances prediction accuracy. Performance indicators such as the R2, MSE, MAE, RMSE, and MAPE are used to evaluate the effectiveness of each method. The bulk of other approaches do not employ stacking techniques to ensemble numerous models to forecast agricultural water needs while constructing prediction tools.

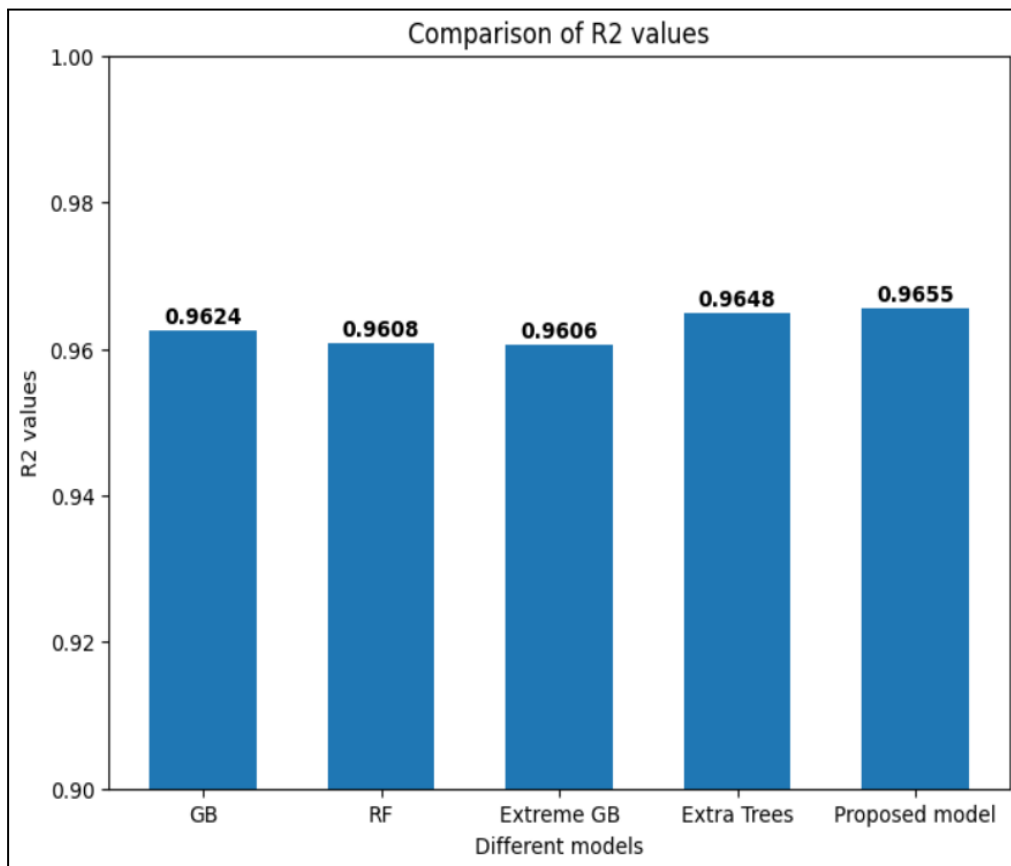


Fig. 5.1 Comparison of various existing models with our proposed model

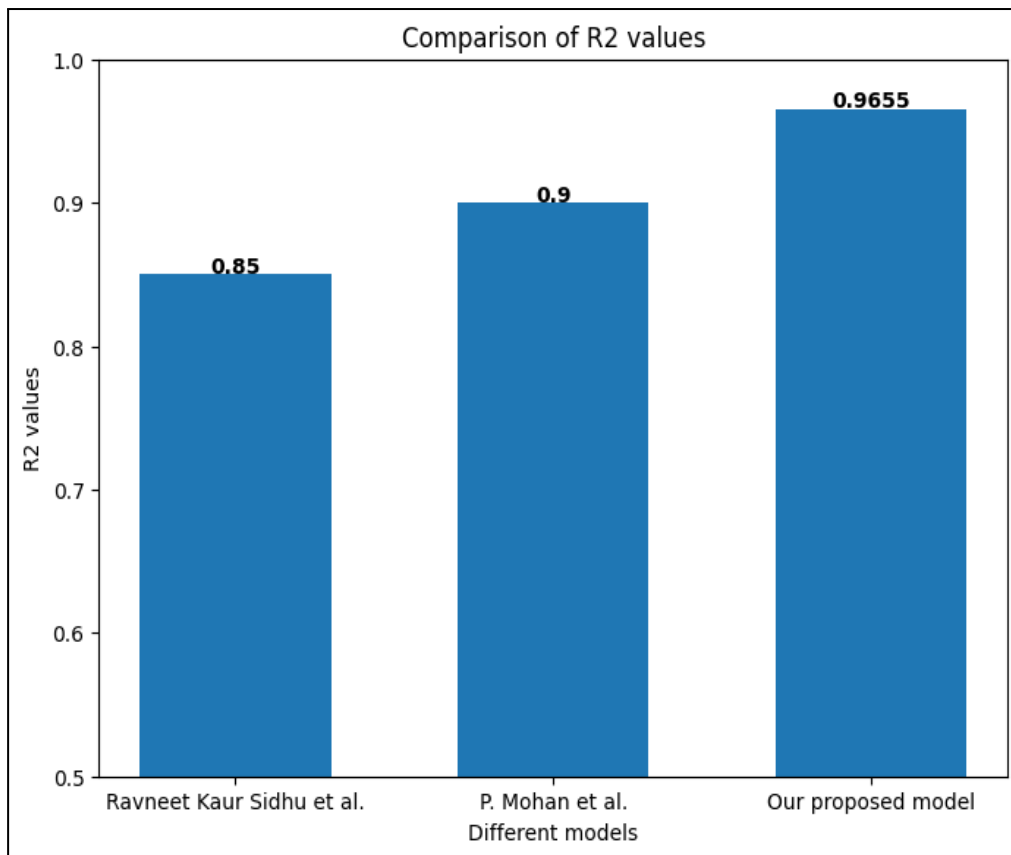


Fig. 5.2 Comparison of  $R^2$  values of the proposed model with the existing literature

## 5.2 Contributions

I've utilized 10-fold cross validation in this framework to get around the issue of an overfitting training dataset. The other authors I've cited have all employed a maximum of three folds.

The suggested methodology uses stacking-based ensemble learning to increase the classifier's diversity. It improves the precision of crop water demand predictions (ET crop).

The proposed framework outperformed other authors who worked on similar projects for predicting the crop water demand when performance was contrasted with previous research on the foundation of R2 score, MAE, MSE, and RMSE.

## References

- [1]Sidhu, R.K., Kumar, R. and Rana, P.S., 2020. Machine learning based crop water demand forecasting using minimum climatological data. *Multimedia Tools and Applications*, 79(19), pp.13109-13124.
- [2]Abdullah, S. S., Malek, M. A., Abdullah, N. S., Kisi, O., & Yap, K. S. (2015). Extreme learning machines: a new approach for prediction of reference evapotranspiration. *Journal of Hydrology*, 527, 184-195.
- [3]Awika J (2011) Major cereal grains production and use around the world. In: Awika J, Piironen V, Bean S (eds) *Advances in cereal science: implications to food processing and health promotion*. American Chemical Society, Atlantic City, pp 1–13
- [4]Dixon J (2007) The economics of wheat: research challenges from field to fork. In: Buck H, Nisi J, Salomon N (eds) *Wheat production in stressed environments*. Springer, Dordrecht, pp 9–22
- [5]Shiferaw B, Smale M, Braun H, Duveiller E, Reynolds MP, Muricho G (2013) Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. *Food Sci* 5:291–317.
- [6]Beshir, S., 2017. Review on estimation of crop water requirement, irrigation frequency and water use efficiency of cabbage production. *Journal of Geoscience and Environment Protection*, 5(07), p.59.

[7]Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, 133(1), 59-68.

[8]Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7, 21.

[9]Coagmet homepage (2016) CoAgMET. Available at: <http://www.coagmet.colostate.edu/>

[10]Shuang, Qing, and Rui Ting Zhao. "Water demand prediction using machine learning methods: A case study of the Beijing–Tianjin–Hebei region in China." *Water* 13.3 (2021): 310.

[11]Karandish, F., & Šimůnek, J. (2016). A comparison of numerical and machine-learning modeling of soil water content with limited input data. *Journal of Hydrology*, 543, 892-909.

[12]Veenadhari, S., Misra, B. and Singh, C.D., 2014, January. Machine learning approach for forecasting crop yield based on climatic parameters. In 2014 International Conference on Computer Communication and Informatics (pp. 1-5). IEEE.

[13] Emami, M., Ahmadi, A., Daccache, A., Nazif, S., Mousavi, S.F. and Karami, H., 2022. County-Level Irrigation Water Demand Estimation Using Machine Learning: Case Study of California. *Water*, 14(12), p.1937.

[14] Mohan, P. and Patil, K.K., 2018. Deep learning based weighted SOM to forecast weather and crop prediction for agriculture application. *Int. J. Intell. Eng. Syst*, 11, pp.167-176.

[15]“Types of Regression Techniques in ML,” GeeksforGeeks, Jan. 16, 2019. <https://www.geeksforgeeks.org/types-of-regression-techniques/> .

[16]J. Brownlee, “A Gentle Introduction to Ensemble Learning Algorithms - MachineLearningMastery.com,” MachineLearningMastery.com, Apr. 18, 2021. <https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/>.

[17] Maina, M. M., Amin, M. S., Rowshon, M. K., Aimrun, W., Samsuzana, A. A., & Yazid, M. A. (2014). Effects of crop evapotranspiration estimation techniques and weather parameters on rice crop water requirement. *Australian Journal of Crop Science*, 8(4), 495–501.



## Appendices

```
df.isna().sum()
Avg Temp      0
Max Temp      0
Min Temp      0
RH Max        0
RH Min        0
Precip        0
Gust Speed    0
Solar Rad     0
ASCE ETr      0
PK ET        0
ASCE Hourly ET 0
ASCE ETo      0
dtype: int64
```

Fig. (i) Calculating total null values

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1090 entries, 0 to 1089
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  ---                -
0   Avg Temp              1090 non-null   float64
1   Max Temp              1090 non-null   float64
2   Min Temp              1090 non-null   float64
3   RH Max                1090 non-null   float64
4   RH Min                1090 non-null   float64
5   Precip                1090 non-null   float64
6   Gust Speed            1090 non-null   float64
7   Solar Rad             1090 non-null   float64
8   ASCE ETr              1090 non-null   float64
9   PK ET                 1090 non-null   float64
10  ASCE Hourly ET        1090 non-null   float64
11  ASCE ETo               1090 non-null   float64
dtypes: float64(12)
memory usage: 102.3 KB
```

Fig. (ii) Dataset information

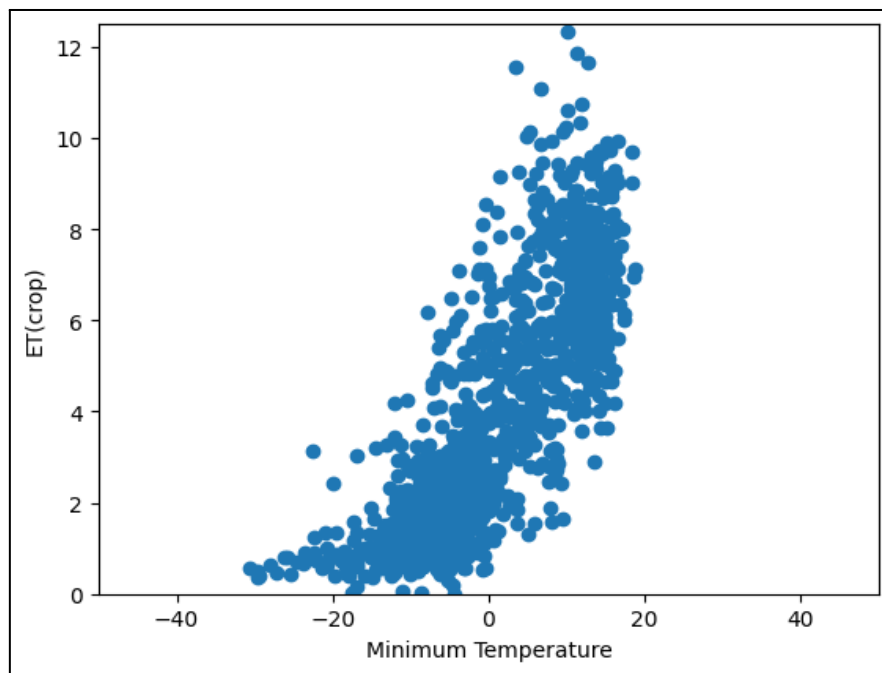
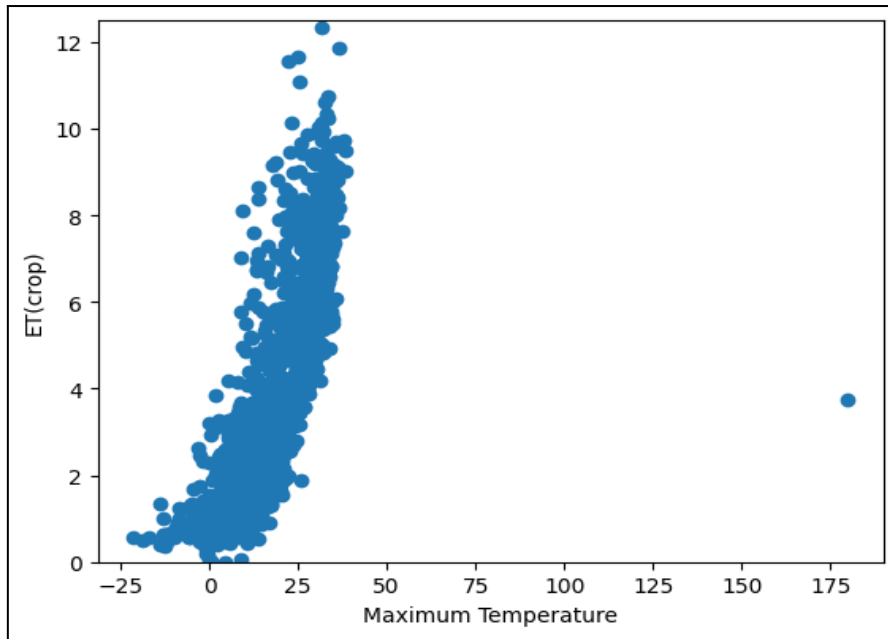


Fig. (iii) ET(crop) vs max. and min. temperature

	MAE	MSE	RMSE	R2	RMSLE	MAPE
<b>Fold</b>						
0	0.3641	0.2570	0.5069	0.9626	0.1149	0.1533
1	0.3261	0.1989	0.4460	0.9681	0.1468	0.4262
2	0.2522	0.1218	0.3491	0.9819	0.0899	0.0952
3	0.3015	0.1539	0.3923	0.9784	0.1055	0.1344
4	0.2743	0.1520	0.3899	0.9779	0.1000	0.1217
5	0.3263	0.1783	0.4222	0.9724	0.1100	0.3239
6	0.3995	0.2783	0.5275	0.9670	0.1465	0.1617
7	0.3157	0.1821	0.4267	0.9692	0.0981	0.1198
8	0.4370	0.3828	0.6187	0.9505	0.1700	0.5848
9	0.4239	0.3132	0.5596	0.9609	0.1265	0.1404
<b>Mean</b>	<b>0.3421</b>	<b>0.2218</b>	<b>0.4639</b>	<b>0.9689</b>	<b>0.1208</b>	<b>0.2261</b>
<b>Std</b>	<b>0.0592</b>	<b>0.0788</b>	<b>0.0814</b>	<b>0.0090</b>	<b>0.0246</b>	<b>0.1558</b>

Fig. (iv) Proposed model

	MAE	MSE	RMSE	R2	RMSLE	MAPE
<b>Fold</b>						
0	0.3597	0.2490	0.4990	0.9638	0.1101	0.1491
1	0.3448	0.2098	0.4580	0.9663	0.1497	0.4390
2	0.2672	0.1270	0.3564	0.9812	0.0905	0.0958
3	0.3100	0.1554	0.3942	0.9782	0.1058	0.1394
4	0.2971	0.1591	0.3989	0.9769	0.1026	0.1304
5	0.3373	0.1881	0.4337	0.9709	0.1102	0.3377
6	0.3925	0.2775	0.5268	0.9671	0.1459	0.1572
7	0.3102	0.1844	0.4294	0.9688	0.1010	0.1264
8	0.4400	0.3849	0.6204	0.9502	0.1737	0.6020
9	0.4569	0.3732	0.6109	0.9534	0.1386	0.1531
<b>Mean</b>	<b>0.3516</b>	<b>0.2308</b>	<b>0.4728</b>	<b>0.9677</b>	<b>0.1228</b>	<b>0.2330</b>
<b>Std</b>	<b>0.0587</b>	<b>0.0852</b>	<b>0.0856</b>	<b>0.0096</b>	<b>0.0257</b>	<b>0.1607</b>

Fig. (v) Gradient boosting regression

	MAE	MSE	RMSE	R2	RMSLE	MAPE
<b>Fold</b>						
0	0.3630	0.2829	0.5319	0.9589	0.1205	0.1591
1	0.3242	0.1935	0.4399	0.9689	0.1489	0.4307
2	0.2547	0.1246	0.3530	0.9815	0.0932	0.0979
3	0.3054	0.1677	0.4095	0.9765	0.1050	0.1327
4	0.2921	0.1655	0.4068	0.9760	0.1049	0.1289
5	0.3285	0.1842	0.4292	0.9714	0.1118	0.3147
6	0.3943	0.2790	0.5282	0.9669	0.1418	0.1606
7	0.3409	0.1958	0.4425	0.9669	0.1053	0.1292
8	0.4269	0.3378	0.5812	0.9563	0.1660	0.5621
9	0.4091	0.3119	0.5584	0.9611	0.1253	0.1353
<b>Mean</b>	<b>0.3439</b>	<b>0.2243</b>	<b>0.4681</b>	<b>0.9684</b>	<b>0.1223</b>	<b>0.2251</b>
<b>Std</b>	<b>0.0518</b>	<b>0.0685</b>	<b>0.0721</b>	<b>0.0077</b>	<b>0.0221</b>	<b>0.1495</b>

Fig. (vi) Random forest regressor

	MAE	MSE	RMSE	R2	RMSLE	MAPE
<b>Fold</b>						
0	0.4112	0.3392	0.5824	0.9507	0.1294	0.1684
1	0.3766	0.2683	0.5180	0.9569	0.1523	0.4334
2	0.3122	0.1797	0.4239	0.9733	0.1044	0.1165
3	0.3246	0.1795	0.4237	0.9748	0.1213	0.1463
4	0.2822	0.1626	0.4032	0.9764	0.1027	0.1323
5	0.3421	0.1965	0.4433	0.9696	0.1138	0.3531
6	0.4310	0.3307	0.5750	0.9608	0.1596	0.1736
7	0.3341	0.2287	0.4782	0.9613	0.1070	0.1258
8	0.4762	0.4583	0.6770	0.9407	0.1782	0.6194
9	0.4568	0.3637	0.6031	0.9546	0.1315	0.1463
<b>Mean</b>	<b>0.3747</b>	<b>0.2707</b>	<b>0.5128</b>	<b>0.9619</b>	<b>0.1300</b>	<b>0.2415</b>
<b>Std</b>	<b>0.0626</b>	<b>0.0937</b>	<b>0.0882</b>	<b>0.0111</b>	<b>0.0244</b>	<b>0.1616</b>

Fig. (vii) Extreme gradient boosting

	MAE	MSE	RMSE	R2	RMSLE	MAPE
<b>Fold</b>						
0	0.3492	0.2298	0.4793	0.9666	0.1146	0.1498
1	0.3250	0.1922	0.4384	0.9691	0.1485	0.4336
2	0.2563	0.1293	0.3595	0.9808	0.0917	0.0931
3	0.2999	0.1701	0.4124	0.9761	0.1042	0.1266
4	0.2922	0.1514	0.3891	0.9780	0.1028	0.1305
5	0.3486	0.2108	0.4592	0.9673	0.1176	0.3000
6	0.4150	0.2958	0.5439	0.9649	0.1485	0.1653
7	0.3353	0.1975	0.4444	0.9666	0.0986	0.1165
8	0.4532	0.4129	0.6426	0.9466	0.1704	0.5680
9	0.4229	0.3017	0.5493	0.9623	0.1255	0.1404
<b>Mean</b>	0.3498	0.2291	0.4718	0.9678	0.1223	0.2224
<b>Std</b>	0.0597	0.0811	0.0809	0.0092	0.0245	0.1520

Fig. (viii) Extra tree regressor