**1. Study Area and Data Collection**

The southwestern Indian state of Karnataka has a wide range of climatic conditions because to its varied physiographic and geographic characteristics. This study focuses on three important sites: Hiriyur (Babbur Farm), Mysore (Nagenahalli), and Bangalore (GKVK). These locations are classified under three different agroclimatic zones, which are distinguished by semi-arid conditions and considerable rainfall variability: the Eastern Dry Region, Southern Dry Zone, and Central Dry Zone. These areas are essential for growing plantation crops (like coconut and Arecanut), pulses (like Pigeonpea, cowpea, and groundnut), and cereals (like paddy, ragi, and maize), which emphasizes the importance of precise agroclimatic assessment.

Agrometeorological observatories within or close to the respective locations provided the daily meteorological parameters, which included maximum (Tmax) and minimum (Tmin) temperature (°C), relative humidity (RH I and RH II %), wind speed (m/s), solar radiation in MJm⁻²day⁻¹, bright sunshine hours in hours, and evaporation in millimeters. The data was gathered for the years 1996–2024 for Bangalore and Mysore and 2011–2021 for Hiriyur.

Additionally, NASA-POWER reanalysis datasets were used to fill up the gaps in the data, especially for wind speed. For continuity and correctness, missing values were filled after checks for temporal and spatial consistency. To access the NASA-POWER data, use this link: This link: <https://power.larc.nasa.gov/data-access-viewer/>

**Table 1 Summary of input settings used to implement statistical, penalized and machine learning models**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **A1** | **A2** | **A3** | **A4** | **A5** | **A6** | **A7** | **A8** | **A9** |
| **Tmax** | \* | \* | \* | \* | \* | \* | \* |  |  |
| **Tmin** | \* | \* | \* | \* | \* | \* | \* |  |  |
| **Tmean** |  |  |  |  |  |  |  | \* | \* |
| **RHmean** |  | \* | \* | \* | \* |  | \* | \* | \* |
| **SR** | \* |  |  | \* |  |  |  |  | \* |
| **SSH** |  |  |  |  | \* | \* | \* | \* |  |
| **WS** |  |  | \* | \* | \* |  |  | \* | \* |

A1 = Temperature and solar radiation; A2 = Temperature and relative humidity; A3 = Temperature, relative humidity, and wind speed; A4 = Temperature, relative humidity, solar radiation, and wind speed; A5 = Temperature, relative humidity, sunshine hours, and wind speed; A6 = Temperature and sunshine hours; A7 = Temperature, relative humidity, and sunshine hours; A8 = Temperature, relative humidity, sunshine hours, and wind speed; A9 = Temperature, relative humidity, solar radiation, and wind speed; **(Tmax = Maximum temperature; Tmin = Minimum temperature; RH = Relative humidity (%); WS = Wind speed (m s⁻¹); SR = Solar radiation (MJ m⁻² day⁻¹); SSH = Sun shine hours)**

**FAO56 Penman–Monteith (FAO 56 PM) Method**

This method was adopted by the FAO as the standard method of estimating *ETo* as it gives more consistent ETo estimates, and it has been shown to perform better than other *ETo* methods [8,9]. According to [8,9], the FAO 56 PM method is summarized by the following equation:

where *ETo* is the reference evapotranspiration (mm d−1), *Rn* is the daily net solar radiation (MJ m−2 d−1), *G* is the soil heat flux (MJ m−2 d−1), *T* is the average daily air temperature at a height of 2 m (◦C), *U2* is the daily mean of the wind speed at a height of 2 m (m s−1), *es* is the saturation vapor pressure (kPa), *ea* is the actual vapor pressure (kPa), ∆ is the slope of the saturation vapor pressure versus the air temperature curve (kPa ◦C−1), and *γ* is the psychrometric constant (kPa ◦C−1). All parameters were calculated using equations provided by [9]. The soil heat flux (*G*) was assumed to be zero over the calculation time step period (24 h) [8].

Extraterrestrial radiation (*Ra*) and theoretical sunshine (*N*), which are astronomical data, can easily be estimated for a certain day and location, according to [9], as follows:

where *Ra* is extraterrestrial radiation (MJ/m2 d), *Gsc* is the solar constant (0.0820 MJ/m2 min), *dr* is the inverse relative distance between the Earth and the Sun, *ωs* is the sunset hour angle (rad), *φ* is latitude (rad), *δ* is solar declination (rad), and *N* is theoretical sunshine (h).

Random Forest for regression (RFr)

Random forest for regression (RFr) utilizes numerous individual models, known as decision trees, which are ultimately aggregated into one, aiming to minimize both the variance and bias of the base learner, which is the decision tree, to the greatest extent achievable by the system. This approach, referred to as ensemble learning [36–38], harnesses the combined knowledge of multiple models to enhance the overall performance of the learning system. That is, a random forest comprises a collection of regression trees (decision trees), utilizing their information collectively.

In the training process, the observations within the fitting dataset are employed to create numerous regression trees, each having distinct training parameters, thereby contributing uniquely to the prediction process. The ultimate observation prediction results from the amalgamation of all individual predictions, thereby leveraging the diverse internal characteristics of each tree to improve generalization. It represents a form of decision structure learning, centered on a predictive model, with the aim of accurately estimating the dependent variable based on the observed values of independent variables.

Each individual regression tree comprises a connected flowchart. In this structure, there is a solitary starting node from which two branches initially extend and lead to ‘child’ nodes stemming from their parent nodes. Each node has a specific satisfaction condition (impurity criterium), and if this objective is not met, the process advances to a new node and its corresponding children. The ensemble method used for ETo model construction was bootstrapped aggregation (bagging) [36,39–43]. This approach entails training multiple independent models on random subsets (bootstraps) of the fitting data. The algorithm selects a random subset of the available features while ensuring there is no correlation among the decision tree estimators. These estimators showed, as expected, high variance, since they perfectly capture the pattern of the particular sample data. Ultimately, when the predictions from these individual models (regression trees) were combined through averaging, the variance in aggregation was significantly reduced.

For the development of a precise and reliable RFr model, it is essential to appropriately adjust its learning hyperparameters. The most critical factors in this process are the quantity of decision trees and their maximum depth within the modeling system. Additionally, careful consideration should be given to parameters such as the minimum number of samples necessary to split a node, the minimum number of samples required for a leaf node, and the number of features to be considered when searching for the optimal split. To identify the most suitable combination of hyperparameters for the RFr model, we utilized a trial-and-error approach, aiming to minimize the mean square error. This involved iterating through various settings for the first two hyperparameters, while keeping the last three at their default values, until we achieved the desired target error. We examined a range of values for the number of decision trees and their maximum depth, spanning from 50 to 500 per unit for the former and 5 to 12 per unit for the latter. The default values for the minimum number of samples required to split a node, the minimum number of samples needed for a leaf node, and the number of features considered for the optimal split were 2, 1, and 1, respectively

RFisanensemblemachinealgorithm,whichiseD cientinbothclassiBcationandregression.Itisgen erallyusedforpredictiveanalysis.Inpredictingthe Bnaloutput,combinationofalldecisiontreesiscon sideredratherthandependingonindividualdecision trees.ThefoundationofRFisbasedonsupervised