Application of Random Forest Regression with Hyper-parameters Tuning to Estimate Reference Evapotranspiration

Satendra Kumar Jain[1]
Research Scholar, Department of Computer Science and Applications, Barkatullah University, Bhopal, Madhya Pradesh, India

Anil Kumar Gupta[2]
Associate Professor, Department of Computer Science and Applications, Barkatullah University, Bhopal, Madhya Pradesh, India

***Abstract*—Estimation of reference evapotranspiration (ETo) is a complex and non-linear problem that is used for the quantification of crop water requirements. In this study, random forest regression based models are developed to predict the ETo of Bhopal city, Madhya Pradesh, India. The meteorological data is collected from IMD, Pune for the periods of the years 2015-16. Based on the correlation among meteorological variables with observed ETo, four different random forest regression models are created. Moreover, the effects of three important hyper-parameters of random forest, such as the number of trees in the forest, depth of the tree, and the number of samples at a leaf node are evaluated to estimate ETo using the proposed models. These hyper-parameters are applied in three different ways to the models such as one hyper-parameter parameter at a time, and combination of hyper-parameters using grid search, and random search approaches. In this study, the result indicates that a random forest regression based model with maximal meteorological input variables exhibits great predictive power in small execution time than minimal input variables. This study also reveals that the model that optimises the hyper-parameters using a grid search approach shows equal predictive power but takes much execution time whereas random search based optimization exhibits the same level of predictive capability in less computation time. Stakeholders can utilize random forest regression models with sufficient meteorological data to estimate crop water requirements, and enhance the food production.**

Keywords*—*Reference evapotranspiration; random forest regression; hyper-parameters; grid search; random search optimization

I. Introduction

Evapotranspiration is a step of the hydrological cycle and has numerous applications such as water management, irrigation scheduling, etc. Evapotranspiration consists of the evaporation and transpiration process. Evaporation removes water from the soils, ponds, and rivers whereas transpiration removes water from the plants. Reference evapotranspiration (ETo) is estimated on smooth grassland which is further used to estimate crop evapotranspiration. The FAO-PM56 is one of the standard empirical methods provided by the Food Agriculture Organization of the United Nations [1]. Such an empirical method suffers from complicated calculations. Weather stations at various places are equipped with power full devices that are constantly observing climatic data. Machine learning based models can be applied to such a huge amount of data to estimate ETo accurately and efficiently. Many authors have applied various machine learning algorithms to estimate ETo.

The ability of M5P, RF, RT, REPT, and KStar and neuro-fuzzy inference systems such as ANFIS, ANFIS-GA, ANFIS-DE, and ANFIS-ICA has been tested to estimate evapotranspiration [2]. Feed-forward artificial neural network with the Levenberg–Marquardt (LM) training algorithm has been investigated to predict evapotranspiration [3]. Genetic programming (GP), support vector machine-firefly algorithm (SVM-FFA), artificial neural network (ANN), and support vector machine–wavelet (SVM–Wavelet) have been analyzed to predict reference evapotranspiration [4]. Extreme learning machine (ELM), back-propagation neural networks optimized by genetic algorithm (GANN), and wavelet neural networks (WNNgra) models have been developed to estimate evapotranspiration [5]. Random forest (RF) and generalized regression neural network (GRNN) models have been applied to estimate daily evapotranspiration [6]. Four tree based ensemble algorithms such as random forest (RF), M5 model tree (M5Tree), gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost)) models have been compared for estimation of evapotranspiration [7] . GRNN, MLP, RBNN, GEP, ANFIS-GP, and ANFIS-SC models have been investigated for modeling evapotranspiration [8]. Genetic (GA) and gene expression programming (GEP) models have been used to estimate reference evapotranspiration [9]. M5P Regression Tree, Bagging, Random Forest (RF), and Support Vector Regression (SVR) have been compared [10]. The performance of kNN k-nearest neighbour, artificial neural network, and Adaptive boosting (AdaBoost) to predict daily evaporation for the potato crop have been investigated [11]. Machine learning algorithms have own hyper-parameters that can be tuned at the training duration. Tuning of hyper-parameters can affect the performance of the algorithm. There are various approaches to tune hyper-parameters. In [12] authors show empirically and theoretically that randomly chosen trails are more efficient than the trails on grid. [13] Performed comparative analysis of various hyper-parameters tuning methods to optimize the accuracy of machine learning algorithms. In [14] hyper-parameters are optimized using weighted random search approaches.

Estimation of ETo plays an important role in water saving and enhancement of food production leading to food security in the world. The selection of machine learning algorithms to estimate ETo is a challenging task because they are not good for all problems. The size and structures of the data affect the performance of the machine learning algorithm. In the current study, the random forest regression algorithm is chosen because of its high performance and handling a complex problem. In this paper, the contribution of works is summarized as follows:

- The reviews of machine learning techniques to estimate reference evapotranspiration and hyper-parameter tuning approaches are done.
- Meteorological data of Bhopal city is collected from IMD Pune. Descriptive analysis is performed on preprocessed data. The correlation coefficient of meteorological data with observed ETo is determined.
- Four different random forest regression based data driven models (based on the correlation among meteorological variables with observed ETo) are developed.
- These hyper-parameters (n_estimators, max_depth, min_samples_leaf) are applied in three different ways (‘one parameter at a time, combinations of parameters using grid and random search approaches) to the four random forest models.
- The performances of twenty models are evaluated and compared with FAO-PM56 using six statistical indicators.

II. Material and Methods

A. Study Site

The proposed random forest regression based data driven models are analyzed in this study using meteorological data from Bhopal city of Madhya Pradesh state, India. Daily meteorological data for the years 2015-16 are obtained from the India Meteorological Data, Pune, which includes input attributes such as minimum temperature (Tmin) in 0C, maximum temperature (Tmax) in 0C, relative humidity (RH) in %, wind speed (u) in m/s and mean solar radiation (Rn) in MJ m-2 day-1. Daily mean sunshine hours of Bhopal city are taken from the Daily Normals of Global & Diffuse Radiation report issued by IMD Pune published in the year 2016. Bhopal city has a subtropical humid climate. It has an average elevation of 500 meters and is located at 23.25 oE latitude and 77.42 oN longitude. Descriptions of training and test datasets are summarized in Table I. The monthly variation of ETo at Bhopal city is observed, where the average minimum ETo is 2.33 mm/day in January 2015 and 2.27 mm/day in December 2016 is noted, similarly average maximum ETo is 7.0 mm/day in May 2015 and 7.47 mm/day in May 2016 is noted. The correlation matrix of observed ETo and the meteorological data of Bhopal city is given in Table II. It can be observed that ETo has a positive correlation with temperature, solar radiation, and wind speed parameters whereas a negative correlation with humidity. Hence it can be said that ETo is an energy driven process and increases as temperature, radiation, and wind speed are increased.

B. FAO-PM56 Equation

The FAO-56 Penman-Monteith equation is provided by the Food and Agriculture Organization of the United Nation and is considered a standard worldwide accepted method to estimate ETo. It is represented as-

$$ETo = \frac{0.408*(R_n - G) + \gamma(\frac{900}{T+273})u_2(e_s - e_a)}{\nabla + \gamma(1 + 0.34*u_2)} \quad (1)$$

Where

ET_0 = grass reference evapotranspiration in mm day^{-1}

R_n = net radiation in MJ mm^{-2} day^{-1}

G = soil heat flus in MJ mm^{-2} day^{-1}

γ = psychrometric consant in kPa ^{0}c^{-1}

T = mean daily air tempreture in ^0C

u_2 = wind speed at 2m height in m s^{-1}

Δ = slope vapour pressure curve in kPa ^{0}c^{-1}

e_s = saturation vapour pressure in kPa

e_a = actual vapour pourpressure in kPa

$e_s - e_a$ = saturation vapourpressure deficit in kPa

TABLE I. Statistical Description of Meteorological Data

Climatic parameters	Data set	Minimum	Maximum	Mean	Standard Deviation
T_{min}	Training	5.8	32.1	19.63	6.0
	Test	7.9	31.2	20.25	5.75
T_{max}	Training	15.5	46.7	32.17	5.8
	Test	14.2	45.3	32.84	5.74
R_H	Training	12	99	56.85	21.76
	Test	17	98	55.19	23.04
u	Training	0	6.6	1.01	0.65
	Test	0.2	2.9	0.99	0.57
R_n	Training	12.3	26.3	18.83	3.48
	Test	12.3	26.3	19.11	3.65
ET_o	Training	1.71	9.5	4.06	1.57
	Test	1.56	8.1	4.15	1.58

TABLE II. Correlation Coefficient of Meteorological Data with Observed ETo

	T_{min}	T_{max}	R_H	u	R_n	ET_o
T_{min}	1					
T_{max}	0.72	1				
R_H	-0.054	-0.6	1			
u	0.49	0.22	0.09	1		
R_n	0.36	0.70	-0.74	0.14	1	
ET_o	0.71	0.87	-0.60	0.47	0.82	1

ETo is observed by CROPWAT8.0 software in this study, which is a decision support tool and developed by the Land and Water Development division of the Food and Agriculture Organization of the United Nation. Daily minimum temperature (T_{min}), maximum temperature (T_{max}), relative humidity (R_H), bright sunshine hours (l_s), and wind speed (u) are applied as input parameters to CROPWAT8.0 software and it returns daily or monthly solar radiation (R_n) and ETo (mm day^{-1}). The FAO-PM56 is considered superior to other methods if reliable and complete meteorological data are available. Huge amounts of meteorological data are recorded at weather stations. Estimation of ETo from such large data using machine learning based models could be an alternative solution that produces accurate and efficient outcomes.

C. Random Forest Regression

Random forest is a supervised machine learning algorithm that is used for classification as well as regression problems. In this study a random forest machine learning algorithm is used to estimate ETo of Bhopal city, which is considered as a function approximation (regression) problem. It works based on the ensemble learning concept, in which instead of making a single model, multiple models are created on randomly selected data. Therefore the outcome of the random forest regression is made based on estimated results of multiple models [15]. Hence it is considered a highly stable model. It removes the overfitting problem of a decision tree. Multiple trees in the random forest lead to higher accuracy. It works well for large datasets with high dimensions. Various hyper-parameters are provided for the random forest. Tuning of hyper-parameters may improve the performance and predictive capability of random forests. Number of trees in the forest (n_estimators), the longest path between the root and the leaf node (max_depth), the minimum required samples to split a node in the tree (min_samples_split), the maximum number of leaf nodes in the tree (max_leaf_nodes), minimum number of samples at the leaf nodes (min_samples_leaf), and criteria to split the node in the tree (criterion) are considered some important hyper-parameters of random forest. In the present study, the performance of random forest is evaluated by tuning the three hyper-parameters such as n_estimators (10, 20, 30, .., 100), max_depth (2, 3, 4, .., 10), and min_samples_leaf (2, 3, 4, 5). These hyper-parameters are applied in three different ways to the models such as 'one hyper-parameter at a time ', and 'combinations of hyper-parameters' using grid search, and random search approaches. In the case of 'one hyper-parameter at a time', the search space consists of one dimensional hyper parameter values. Grid search and random search approaches are used when multiple hyper-parameters are applied to the model. In this case, the search space consists of a grid of hyper-parameter values, and the model is evaluated at each point in the grid. In the case of random search, the model is evaluated on a randomly opted grid point. Grid search is simple to implement and always finds the best combinations of hyper-parameter. It is a time consuming approach due to the exhaustive search nature. Random search exhibits the same performance in less computation time.

D. Model Development

Model development steps are shown in Fig. 1. Initially, the meteorological and geographical data of Bhopal city is taken into memory. Data preprocessing is a significant step to estimate ETo accurately. It transforms the data in a meaningful way. To obtain the optimum outcomes, missing values are filled in different ways. In the present study, missing values are filled by the mean value of those attributes. Values of all attributes are normalized using the z-score method to make all attributes to the same level of magnitudes so have the same emphasis. Values of ETo are observed using CROPWAT 8.0 software (developed by the Land and Water Development Division of FAO (The Food and Agriculture Organization of the United Nation)) and made as a dependent variable, whereas the remaining attributes (T_{min}, T_{max}, R_n, u, R_H) are designated as independent variables. The whole dataset is partitioned into the training dataset (80%) and the test dataset (20%).

Four random forest regression based models such as RFR-Model1, RFR-Model2, RFR-Model3, and RFR-Model4 are created. Different combinations of meteorological input parameters (made based on high correlation coefficient with observed ETo values) are applied to these models. In the RFR-Model1, T_{min}, and T_{max} are applied. In the RFR-Model2, T_{min}, T_{max}, and R_n are applied. In the RFR-Model3, T_{min}, T_{max}, R_n, and u are applied. And finally in the RFR-Model4, T_{min}, T_{max}, R_n, u, and R_H are applied. In addition to the input combinations of meteorological parameters, three important hyper-parameters are tuned in each model. These hyper-parameters are tuned and applied to the proposed four models in three different ways: 'one hyper-parameter at a time', and combinations of hyper-parameters are using grid search, and random search optimization approach. Taking into consideration four different models and the applicability of three hyper-parameters to the models produces twenty combinations. Therefore in this study, the performances of twenty models are evaluated. Six different statistical indicators are used in this study to evaluate the performance of the models such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), Pearson correlation coefficient (r), r^2 (coefficient of determination), and Nash-Sutcliffe(NS). These models are implemented in Python with the help of Pandas, Numpy, Sklearn and Matplotlib libraries.

E. Performance Evaluation Indices

Predictive skills of RFR-MODEL1, RFR-MODEL2, RFR-MODEL3, and RFR-MODEL4are evaluated using the following parameters:

Mean absolute error (MAE).

$$MAE = \frac{\sum_{i=1}^{n} \left| E_{pi} - E_{oi} \right|}{n} \quad (2)$$

where

E_{pi} = predicted evapotranspiration.

E_{oi} = observed evapotranspiration.

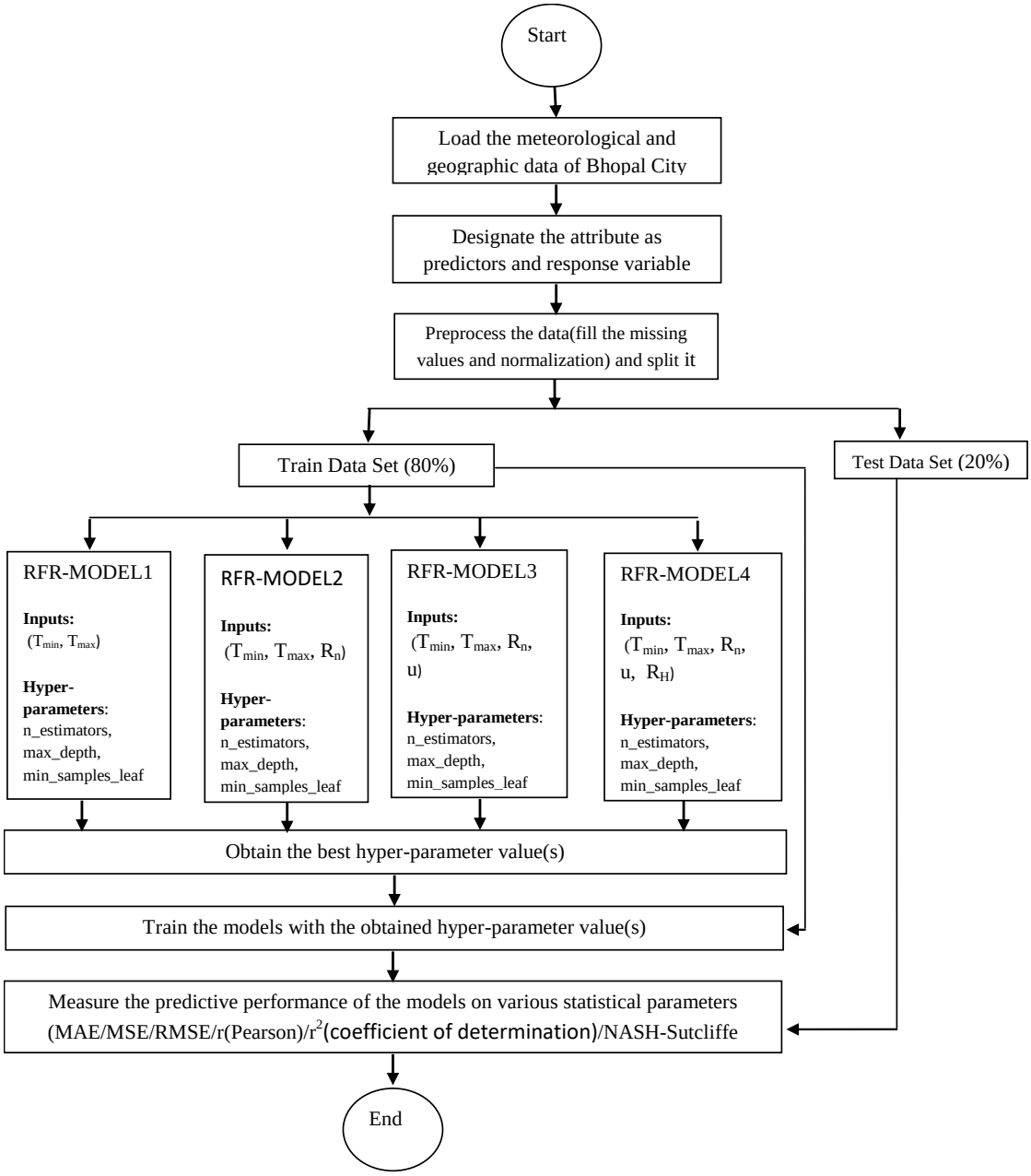

Fig. 1. Flow Chart of Random Forest Regression based Models.

Mean square error (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(E_{oi} - E_{pi})^2 \quad (3)$$

Root mean square error (RMSE) -A small value of RMSE denotes the model fits the datasets strongly.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_{oi} - E_{pi})^2} \quad (4)$$

Pearson correlation coefficient shows the strength of the relationship between observed and predicted ETo.

$$Pearson\ correlation(r) = \frac{\sum_{i=1}^{n}(E_{pi} - \overline{E_{pi}})(E_{oi} - \overline{E_{oi}})}{\sqrt{\sum_{i=1}^{n}(E_{pi} - \overline{E_{pi}})^2 \sum_{i=1}^{n}(E_{oi} - \overline{E_{oi}})^2}} \quad (5)$$

r^2 (coefficient of determination) . A larger value of r^2 indicates the model fits the datasets strongly.

$$Coefficient\ of\ determination(r^2) = r * r \quad (6)$$

Nash-Sutcliffe efficiency (NS) is used to assess the predictive skills of ANN models.

$$Nash-Sutcliffe efficiency(NS) = 1 - \frac{\sum_{i=1}^{n}(E_{pi} - E_{oi})^2}{\sum_{i=1}^{n}(E_{pi} - \overline{E_o})^2} \quad (7)$$

III. Result and Discussion

As stated earlier in the model development section, taking into consideration four different random forest regression based models and the applicability of three hyper-parameters to the models in different ways produces twenty combinations. Therefore in this study, the performances of twenty models are evaluated. The execution time span of each model is calculated from the beginning of the training period to the end of the testing period.

A. Performance of the RFR-Model1

In this model, only two meteorological inputs T_{min}, and T_{max} are applied. The performance of this model is demonstrated in Table III, where it exhibits mae of 0.48, mse of 0.39, rmse of 0.62, r of 0.92, r^2 of 0.85, and Nash-Sutcliffe of 0.85 when the n_estimators hyper-parameter is tuned. Table IV exhibits the performance with mae of 0.45, mse of 0.34, rmse of 0.59, r of 0.93, r^2 of 0.87, and Nash-Sutcliffe of 0.86 when the max_depth hyper-parameter is tuned. Table V exhibits the performance with mae of 0.45, mse of 0.35, rmse of 0.59, r of 0.93, r^2 of 0.87, and Nash-Sutcliffe of 0.86 when the min_samples_leaf hyper-parameter is tuned. Table VI exhibits the performance with mae of 0.46, mse of 0.36, rmse of 0.6, r of 0.93, r^2 of 0.86, and Nash-Sutcliffe of 0.86 when the combination of three hyper-parameters (n_estimators, max_depth, min_samples_leaf) are tuned using a grid search approach. Similarly, Table VII exhibits the performance with mae of 0.46, mse of 0.35, rmse of 0.59, r of 0.93, r^2 of 0.87, and Nash-Sutcliffe of 0.86 when the same combination of hyper-parameters are tuned using a random search approach. It can be observed that RFR-Model1 shows almost the same predictive capability in all scenarios. The Computation time of this model is represented in Table VIII. It takes 10.5 seconds when the n_estimators hyper-parameter is tuned, 29.38 seconds when the max_depth hyper-parameter is tuned, 70 seconds when the min_samples_leaf hyper-parameter is tuned, 301.33 seconds when a grid search approach is applied, and 11.62 seconds when a random search approach is applied respectively in order to estimate ETo. Regression analysis of the RFR-Model1 is shown in Fig. 2 for all scenarios.

B. Performance of the RFR-Model2

In this model, only three meteorological inputs T_{min}, $T_{max,}$ and R_n are applied. The performance of this model is demonstrated in Table III, where it exhibits mae of 0..29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.93, and Nash-Sutcliffe of 0.93 when the n_estimators hyper-parameter is tuned. Table IV exhibits the performance with mae of 0.29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.94, and Nash-Sutcliffe of 0.93 when the max_depth hyper-parameter is tuned. Table V exhibits the performance with mae of 0.29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.94, and Nash-Sutcliffe of 0.93 when the min_samples_leaf hyper-parameter is tuned. Table VI exhibits the performance with mae of 0.29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.94, and Nash-Sutcliffe of 0.93 when the combination of three hyper-parameters (n_estimators, max_depth, min_samples_leaf) are tuned using a grid search approach. Similarly, Table VII exhibits the performance with mae of 0.29, mse of 0.17, rmse of 0.41, r of 0.97, r^2 of 0.94, and Nash-Sutcliffe of 0.93 when the same combination of hyper-parameters are tuned using a random search approach. It can be observed that RFR-Model2 shows the same predictive capability in all scenarios but higher than RFR-Model1. The Computation time of this model is represented in Table VIII. It takes 11.62 seconds when the n_estimators hyper-parameter is tuned, 31.9 seconds when the max_depth hyper-parameter is tuned, 69 seconds when the min_samples_leaf hyper-parameter is tuned, 321.39 seconds when a grid search approach is applied, and 9.72 seconds when a random search approach is applied respectively in order to estimate ETo. Regression analysis of the RFR-Model2 is shown in Fig. 3 for all scenarios.

TABLE III. Model Performance when 'N_estimators' Hyper Parameter is Tuned

Performance Indices	RFR-Model 1	RFR-Model 2	RFR-Model 3	RFR-Model 4
MAE	0.48	0.29	0.17	0.15
MSE	0.39	0.17	0.05	0.05
RMSE	0.62	0.41	0.23	0.22
Pearson(r)	0.92	0.97	0.99	0.99
r^2	0.85	0.93	0.98	0.98
Nash-Sutcliff e	0.85	0.93	0.98	0.98

TABLE IV. Model Performance When 'Max_depth' Hyper Parameter is Tuned

Performance Indices	RFR-Model1	RFR-Model2	RFR-Model3	RFR-Model4
MAE	0.45	0.29	0.17	0.15
MSE	0.34	0.17	0.05	0.05
RMSE	0.59	0.41	0.23	0.22
Pearson(r)	0.93	0.97	0.99	0.99
r^2	0.87	0.94	0.98	0.98
Nash-Sutcliffe	0.86	0.93	0.98	0.98

TABLE V. Model Performance when 'Max_samples_leaf' Hyper Parameter is Tuned

Performance Indices	RFR-Model1	RFR-Model2	RFR-Model3	RFR-Model4
MAE	0.45	0.29	0.18	0.16
MSE	0.35	0.17	0.06	0.06
RMSE	0.59	0.41	0.25	0.23
Pearson(r)	0.93	0.97	0.99	0.99
r^2	0.87	0.94	0.98	0.98
Nash-Sutcliffe	0.86	0.93	0.97	0.98

TABLE VI. MODEL PERFORMANCE WHEN (N_ESTIMATORS, MAX_DEPTH, MAX_SAMPLES_LEAF) HYPER PARAMETERS ARE TUNED USING GRID SEARCH

Performance Indices	RFR-Model1	RFR-Model2	RFR-Model3	RFR-Model4
MAE	0.46	0.29	0.18	0.17
MSE	0.36	0.17	0.06	0.06
RMSE	0.6	0.41	0.25	0.24
Pearson(r)	0.93	0.97	0.99	0.99
r^2	0.86	0.94	0.98	0.98
Nash-Sutcliffe	0.86	0.93	0.97	0.98

TABLE VII. MODEL PERFORMANCE WHEN N_ESTIMATORS, MAX_DEPTH, MAX_SAMPLES_LEAF) HYPER PARAMETERS ARE TUNED USING RANDOM SEARCH

Performance Indices	RFR-Model1	RFR-Model2	RFR-Model3	RFR-Model4
MAE	0.46	0.29	0.19	0.16
MSE	0.35	0.17	0.07	0.06
RMSE	0.59	0.41	0.26	0.24
Pearson(r)	0.93	0.97	0.99	0.99
r^2	0.87	0.94	0.98	0.98
Nash-Sutcliffe	0.86	0.93	0.97	0.98

TABLE VIII. EXECUTION TIME (SECONDS) TAKEN BY EACH MODEL

	One hyper-parameter at time			grid search	random search
Models	number of trees	depth of trees	samples at leaf node		
RFR-Model1	10.5	29.38	70	301.33	11.62
RFR-Model2	11.62	31.9	69	321.39	9.72
RFR-Model3	11.87	33.24	71.2	316.31	9.48
RFR-Model4	12.44	36.8	75.57	329.98	11.12

C. Performance of the RFR-Model3

In this model, four meteorological inputs T_{min}, $T_{max,}$ R_n, and u are applied. The performance of this model is demonstrated in Table III, where it exhibits mae of 0.17, mse of 0.05, rmse of 0.23, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.98 when the n_estimators hyper-parameter is tuned. Table IV exhibits the performance with mae of 0.17, mse of 0.05, rmse of 0.23, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.98 when the max_depth hyper-parameter is tuned. Table V exhibits the performance with mae of 0.18, mse of 0.06, rmse of 0.25, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.97 when the min_samples_leaf hyper-parameter is tuned. Table VI exhibits the performance with mae of 0.18, mse of 0.06, rmse of 0.25, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.97 when the combination of three hyper-parameters (n_estimators, max_depth, min_samples_leaf) are tuned using a grid search approach. Similarly, Table VII exhibits the performance with mae of 0.19, mse of 0.07, rmse of 0.26, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.97 when the same combination of hyper-parameters are tuned using a random search approach. It can be observed that RFR-Model3 shows almost the same predictive capability with minor variations in all scenarios but higher than RFR-Model1 and RFR-Model2. The Computation time of this model is represented in Table VIII. It takes 11.87 seconds when the n_estimators hyper-parameter is tuned, 33.24 seconds when the max_depth hyper-parameter is tuned, 71.2 seconds when the min_samples_leaf hyper-parameter is tuned, 316.31 seconds when a grid search approach is applied, and 9.48 seconds when a random search approach is applied respectively in order to estimate ETo. Regression analysis of the RFR-Model3 is shown in Fig. 4 for all scenarios.

D. Performance of the RFR-Model4

In this model, five meteorological inputs T_{min}, T_{max}, R_n, u, and R_H are applied. The performance of this model is demonstrated in Table III, where it exhibits mae of 0.15, mse of 0.05, rmse of 0.22, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.98 when the n_estimators hyper-parameter is tuned. Table IV exhibits the performance with mae of 0.15, mse of 0.05, rmse of 0.22, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.98 when the max_depth hyper-parameter is tuned. Table V exhibits the performance with mae of 0.16, mse of 0.06, rmse of 0.23, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.98 when the min_samples_leaf hyper-parameter is tuned. Table VI exhibits the performance with mae of 0.17, mse of 0.06, rmse of 0.24, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.98 when the combination of three hyper-parameters (n_estimators, max_depth, min_samples_leaf) are tuned using a grid search approach. Similarly, Table VII exhibits the performance with mae of 0.16, mse of 0.06, rmse of 0.24, r of 0.99, r^2 of 0.98, and Nash-Sutcliffe of 0.98 when the same combination of hyper-parameters are tuned using a random search approach. It can be observed that RFR-Model4 shows almost the same predictive capability in all scenarios but higher than RFR-Model1, RFR-Model2 and RFR-Model3. The Computation time of this model is represented in Table VIII. It takes 12.44 seconds when the n_estimators hyper-parameter is tuned, 36.8 seconds when the max_depth hyper-parameter is tuned, 75.57 seconds when the min_samples_leaf hyper-parameter is tuned, 329.98 seconds when a grid search approach is applied, and 11.12 seconds when a random search approach is applied respectively in order to estimate ETo. Regression analysis of the RFR-Model4 is shown in Fig. 5 for all scenarios.

It can be observed that RFR-Model1 demonstrates poor predictive performance. The performance of the models is improving gradually when the maximal meteorological input variables are taken into consideration. Grid search based optimization demonstrates the same level of performance but takes much execution time and will not be feasible when size of search spaces increases whereas random search based optimization exhibits better performance than grid search. Computation time is shown in Fig. 6.

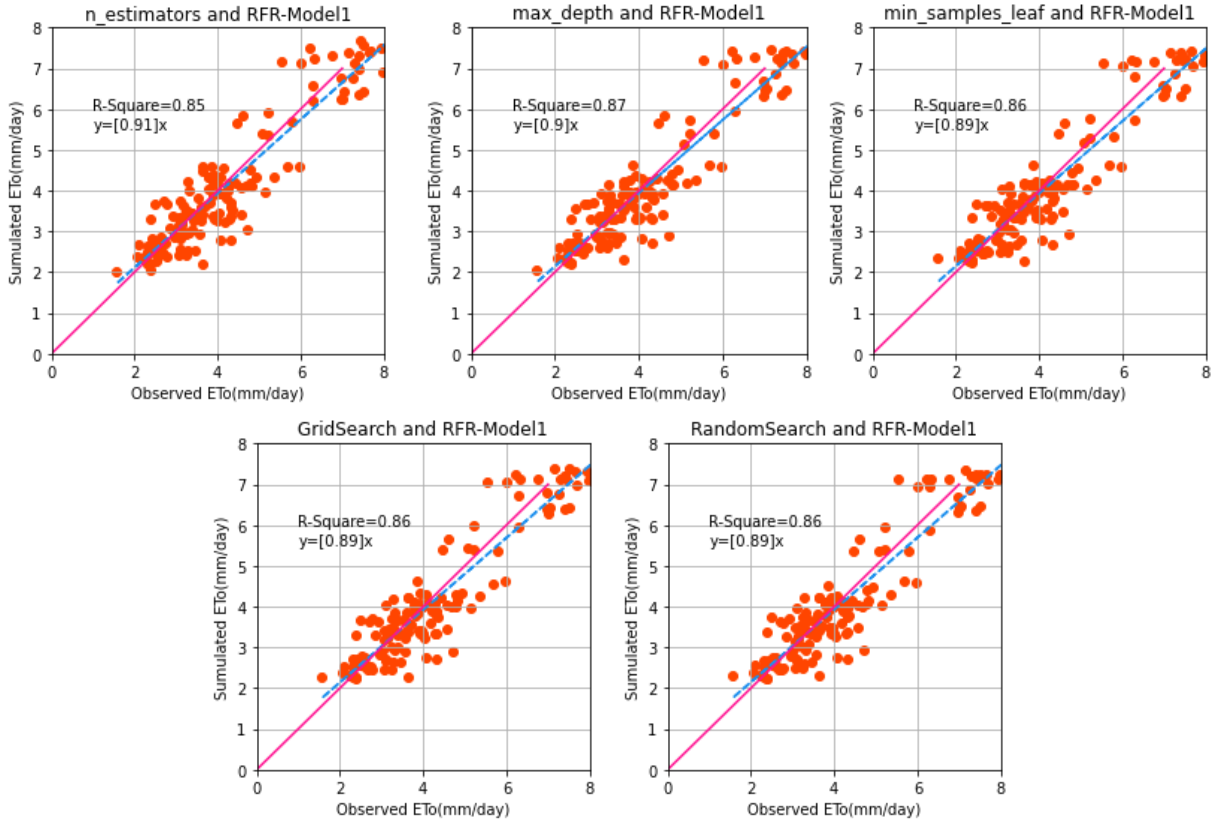

Fig. 2. Regression Analysis of the RFR-Model1 in all Scenarios.

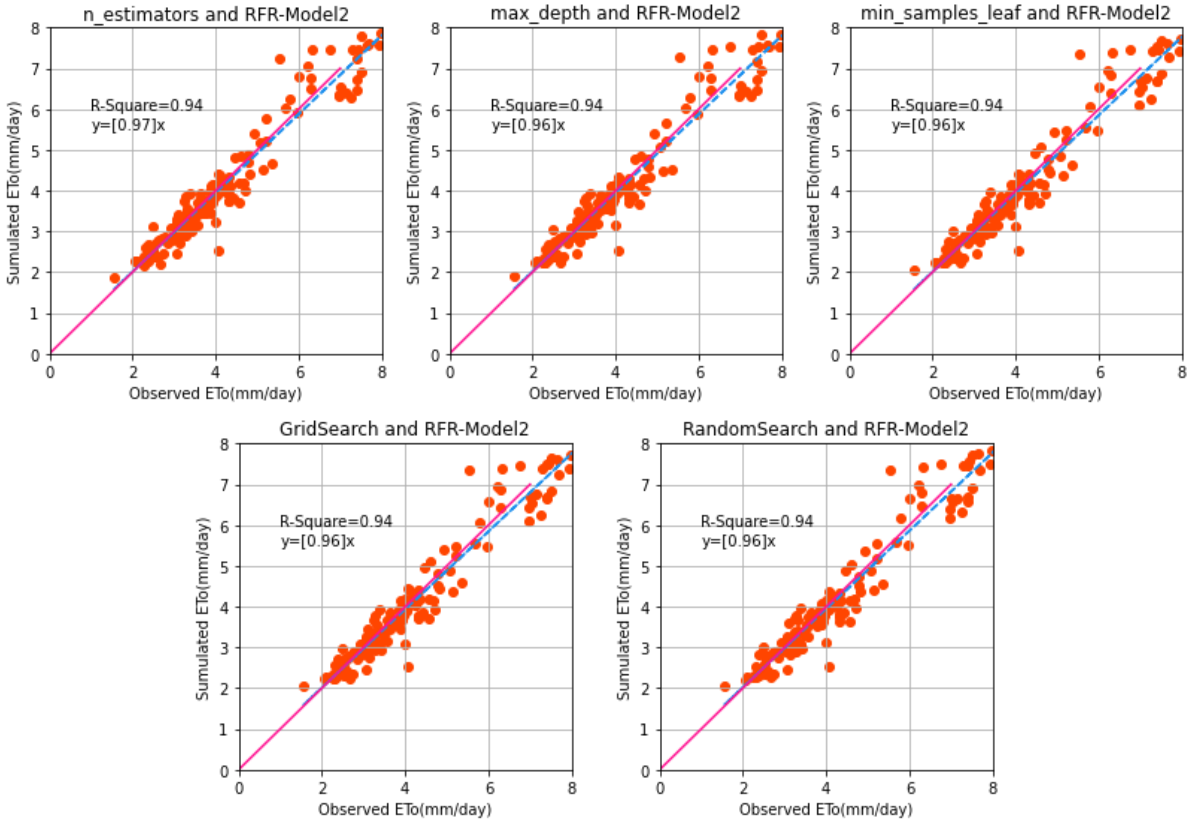

Fig. 3. Regression Analysis of the RFR-Model2 in all Scenarios.

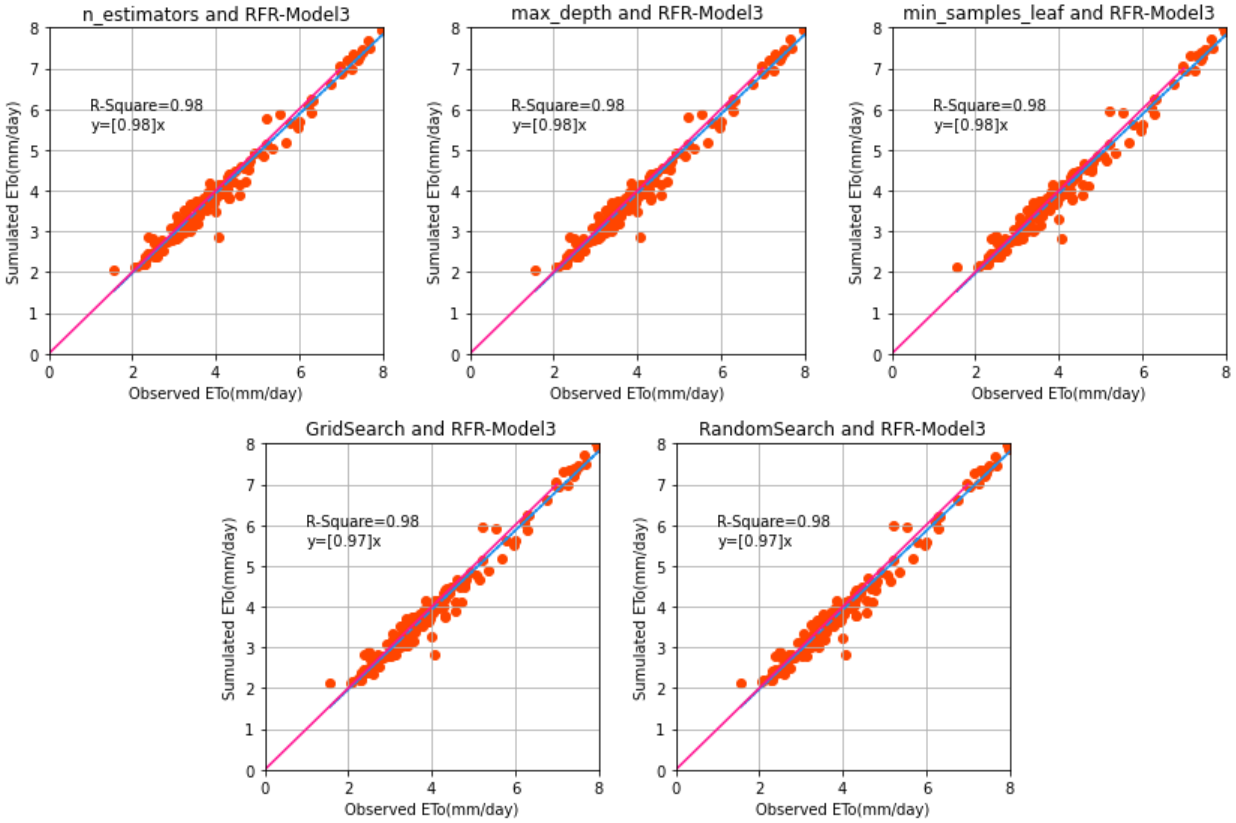

Fig. 4. Regression Analysis of the RFR-Model3 in all Scenarios.

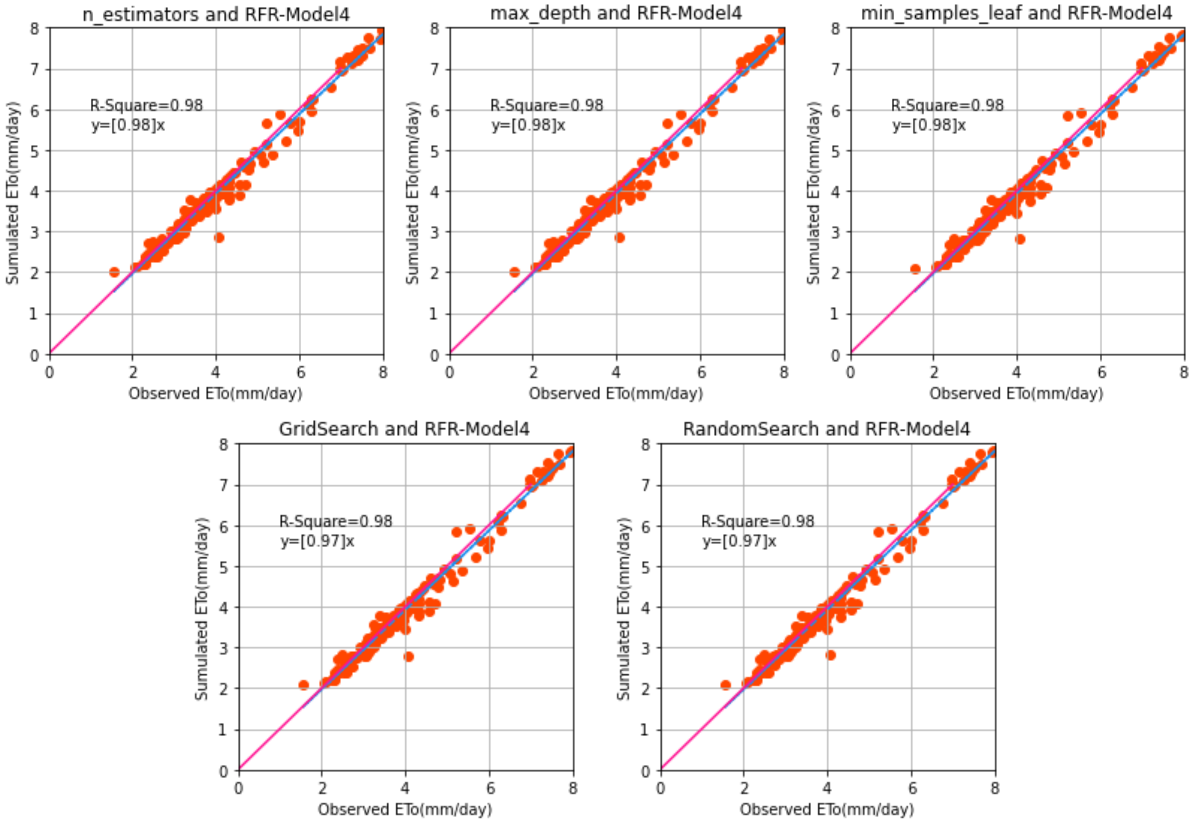

Fig. 5. Regression Analysis of the RFR-Model4 in all Scenarios.

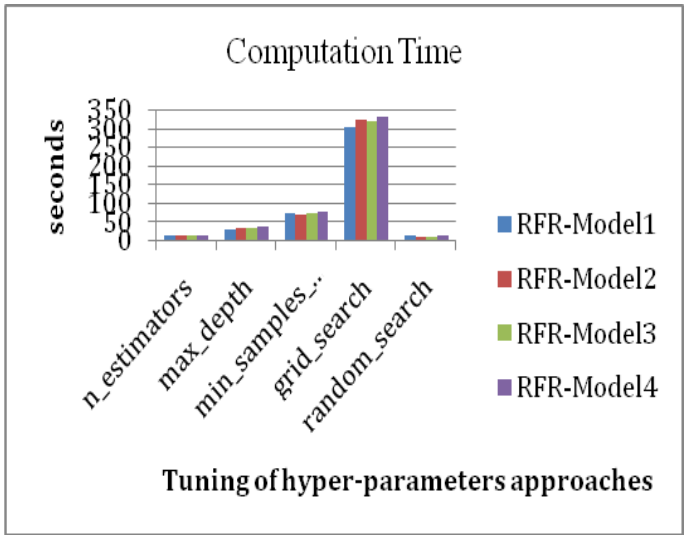

Fig. 6. Computation Time of each Model.

IV. Conclusion

Estimation of ETo has numerous applications. Irrigation scheduling is one of them. In this study, random forest regression based four different models are developed to estimate ETo. Different combinations of meteorological input variables (made based on high correlation coefficient with observed ETo values) are applied to these models. Moreover, the effects of three important hyper-parameters of random forest regression, such as the number of trees in the forest, depth of the trees, and the number of samples at a leaf nod are evaluated to estimate ETo using the proposed models. These hyper-parameters are optimized and applied in three different ways to the models such as one parameter at a time, and combinations of hyper parameters using grid search, and random search. This study reveals that the models with less meteorological input variables demonstrate poor performance than models with maximal input variables (r is of 0.99, r^2 is of 0.98 and Nash-Sutcliffe is of 0.98 in the case of RFR-Model4). Models based on grid search based optimization exhibit the same predictive power but take much computation time. The findings of this study are that random forest regression based models with sufficient meteorological data demonstrate better performance and are useful to the stakeholders such as farmers, engineers for irrigation scheduling and water management. In the future, more hyper-parameter optimization techniques will be applied to estimate accurate ETo for various places in India. This estimated ETo will be used to calculate crop water requirements of Wheat and Maize crops

References

[1] R. G. Allen, L. S. Pereira, D. Raes, and M. Smith, "Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56," 1998.

[2] K. Khosravi et al., "Meteorological data mining and hybrid data-intelligence models for reference evaporation simulation: A case study in Iraq," Comput. Electron. Agric., vol. 167, Dec. 2019, doi: 10.1016/j.compag.2019.105041.

[3] O. Kisi, "Evapotranspiration modelling from climatic data using a neural computing technique," Hydrol. Process., vol. 21, no. 14, pp. 1925–1934, Jul. 2007, doi: 10.1002/hyp.6403.

[4] M. Gocić et al., "Soft computing approaches for forecasting reference evapotranspiration," Comput. Electron. Agric., vol. 113, pp. 164–173, Apr. 2015, doi: 10.1016/j.compag.2015.02.010.

[5] Y. Feng, N. Cui, L. Zhao, X. Hu, and D. Gong, "Comparison of ELM, GANN, WNN and empirical models for estimating reference evapotranspiration in humid region of Southwest China," J. Hydrol., vol. 536, pp. 376–383, May 2016, doi: 10.1016/j.jhydrol.2016.02.053.

[6] Y. Feng, N. Cui, D. Gong, Q. Zhang, and L. Zhao, "Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling," Agric. Water Manag., vol. 193, pp. 163–173, Nov. 2017, doi: 10.1016/j.agwat.2017.08.003.

[7] J. Fan et al., "Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China," Agric. For. Meteorol., vol. 263, pp. 225–241, Dec. 2018, doi: 10.1016/j.agrformet.2018.08.019.

[8] H. Sanikhani, O. Kisi, E. Maroufpoor, and Z. M. Yaseen, "Temperature-based modeling of reference evapotranspiration using several artificial intelligence models: application of different modeling scenarios," Theor. Appl. Climatol., vol. 135, no. 1–2, pp. 449–462, Jan. 2019, doi: 10.1007/s00704-018-2390-z.

[9] M. Valipour, M. A. G. Sefidkouhi, M. Raeini-Sarjaz, and S. M. Guzman, "A hybrid data-driven machine learning technique for evapotranspiration modeling in various climates," Atmosphere (Basel)., vol. 10, no. 6, Jun. 2019, doi: 10.3390/atmos10060311.

[10] F. Granata, "Evapotranspiration evaluation models based on machine learning algorithms—A comparative study," Agric. Water Manag., vol. 217, pp. 303–315, May 2019, doi: 10.1016/j.agwat.2019.03.015.

[11] S. S. Yamaç and M. Todorovic, "Estimation of daily potato crop evapotranspiration using three different machine learning algorithms and four scenarios of available meteorological data," Agric. Water Manag., vol. 228, Feb. 2020, doi: 10.1016/j.agwat.2019.105875.

[12] J. Bergstra, J. B. Ca, and Y. B. Ca, "Random Search for Hyper-Parameter Optimization Yoshua Bengio," 2012. [Online]. Available: http://scikit-learn.sourceforge.net.

[13] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," Informatics, vol. 8, no. 4, Dec. 2021, doi: 10.3390/informatics8040079.

[14] R. Andonie and A. C. Florea, "Weighted random search for CNN hyperparameter optimization," Int. J. Comput. Commun. Control, vol. 15, no. 2, pp. 1–11, 2020, doi: 10.15837/IJCCC.2020.2.3868.

[15] L. Breiman, "Random Forests," 2001.