Improvement the estimation of reference evapotranspiration by combining different types of meteorological data Using machine learning models

Ayoub Ba-ichou[1*], *Abderrahim* Zegoumou [2], *Said* Benhlima[1], *and My Ali* Bekr[1]

1 Lab TSI, Department of Computer Science, Faculty of Sciences, Moulay Ismail University, Meknes, Morocco
2 Tofail University. Faculty of Sciences. Laboratory of Vegetal, Animal and Agro Productions industry, University campus Kenitra BP 133 Morocco

Abstract. Irrigation and the strategic planning thereof play a pivotal role in diverse hydrological inquiries, with reference evapotranspiration (ETo) standing as a paramount variable within this domain. While the equation (FAO-56 PM)is extensively employed for (ETo)estimation, its dependence on numerous weather datas such as solar radiation, temperature, relative humidity, extraterrestrial radiation and wind speed, introduces inherent constraints, the remote computation necessitates a substantial array of sensors, thereby incurring considerable expenses. To surmount this challenge, artificial intelligence methodologies, encompassing various machine learning (ML) models, are harnessed for ETo estimation, requiring only minimal parameters.This investigation scrutinizes the effectiveness of alternative equations (Hargreaves-Samani, Romannenko, Jensen-Haise, ASCE_PM) vis-à-vis (ML) models such as Xgboost,Support Vector Machine (SVM), and Random Forest (RF) in the estimation of ETo across the Meknes region, utilizing diverse permutations of the four measured variables. The study employs an extensive array of hyperparameters in two distinct scenarios: (i) randomization of all data, and (ii) training on one station while validating on another. All methodologies employed in this study yield satisfactory outcomes when juxtaposed against empirical models reliant on minimal meteorological data.

1 Introduction

The estimation of (ETo) is important for irrigation planning and water resource management, especially in countries like Morocco that face drought and water scarcity. The

*Corresponding author:ayoub.baichou@edu.umi.ac.ma

lysimeter is he accurate method of calculating ETo, but is expensive to use. Using intelligent models based on meteorological data is therefore a more suitable and cost-effective approach as it minimizes the variables involved in the equations and can be easily integrated into irrigation management software. This approach allows him to accurately estimate ETo at a lower cost, which is very important in Morocco's vast irrigated areas. The FAO(United Nations Food and Agriculture Organization) recommends employing the FAO-56 PM as the benchmark method for estimating ETo (Allen et al., 1998). Despite its globally acknowledged accuracy (Landeras et al., 2008), This equation requires a large amount of weather data: maximum temperatures, minimum temperatures, solar radiation, relative humidity, and wind speed.This complexity arises from the unavailability of certain parameters in local weather stations. Consequently, efforts have been made to reduce the number of empirical equation parameters for reference evapotranspiration (ETo) in situations where data on relative humidity, solar radiation,or wind speed are unavailable. While these approaches necessitate less data, their performance varies based on the climatic conditions of the respective locations (Maestre-Valero et al., 2013), sometimes yielding unsatisfactory results.

This study commences by showcasing the inadequacy of these models through a comparison with results obtained from FAO-56 Penman-Monteith, revealing performance disparities in the absence of data. Recent endeavors to estimate ETo have explored machine learning models, such as Random Forest (RF) and support vector machines (SVMs). Among these models, SVMs have demonstrated superior performance compared to traditional techniques. In Morocco, for the first time, we leverage artificial intelligence techniques like machine learning (ML) to estimate ETo with minimal parameters. Accurate ETo estimation is crucial for effective irrigation planning and water resource management, especially in countries grappling with drought and water scarcity like Morocco.(López-Urrea et al., 2006)

While lysimeters offer precision in ETo calculation, their expense limits widespread use. Intelligent models based on weather data provide a cost-effective alternative, minimizing variables in equations, integrating seamlessly with irrigation management software, and ensuring accurate ETo estimation. Despite the FAO-56 PM being recommended as a reference method, its extensive meteorological data requirements pose challenges, particularly in local weather stations. Consequently, alternative procedures have been proposed for estimating ETo variables, like solar radiation, relative humidity, and wind speed. However, these approaches exhibit variable performance depending on climatic conditions. To address this, several studies have turned to machine learning models, including RF and SVMs, showcasing the potential of ML, particularly SVMs, in accurate ETo estimation, even in regions with limited data availability. This work marks the first exploration of ML techniques for ETo estimation in Morocco, highlighting their efficacy with a constrained parameter set.

2 Materials & methods

2.1 Datasets

2.1.1 Datasets& data combinations

The data used in this study were gathered from two stations in Morocco: ENSAM-Meknes (latit: 33°85' N, longi: −5°57' W, elevation: 557m), ENAM-Meknes (latit: 33.843199, longi: -5.473981, elevation: 631.9m). We procured a time series encompassing climate observations spanning from 2018 to 2023. These observations include maximum air temperature (Tmax), minimum air temperature (Tmin), mean relative humidity (RH), mean solar radiation (SR),extraterrestrial radiation (ER), and mean wind speed ($u2$).

The gathered data were partitioned into two primary scenarios, denoted as A and B. Scenario A encompasses nine distinct inputs, covering all possible combinations of climate data, employing cross-validation across all data. In scenario (B), the combinations from Scenario A are maintained, but with a training phase involving one station and validation utilizing the other.

2.1.2 Data cleaning

Concerning data cleansing, we removed lines featuring anomalies such as instances where the minimum temperature exceeded the maximum temperature. Additionally, we excluded data points showing relative humidity beyond the acceptable range of 0–100%, cases where the (RH)min surpassed the (RH)max, negative ($u2$), negative (SR), or instances where (SR) exceeded (ER). Moreover, any missing ET0 values were imputed through a direct calculation using the (FAO56-PM) equation.

2.2 Calculation of $ET0$

2.2.1 ThePenman-Monteith FAO-56 PM

The (PM) equation, specifically the FAO-56 PM utilized for estimating ET0, is a physically grounded approach that seamlessly incorporates both physiological and aerodynamic parameters (XU and AL 2006). Widely acknowledged as the most reliable and globally accepted method, the FAO-56 PM method is designed for estimating ET0 across diverse climatic conditions. The primary expression for computing ET0 within the PM method is commonly presented as outlined by (Allen and Al (1998)):

$$\boldsymbol{ET_0} = \frac{\mathbf{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u_2 (e_s - e_a)}}{\mathbf{\Delta + \gamma(1 + 0.34 u_2)}} \qquad (1)$$

2.2.2 Romanenkoequation.

Romanenko (1961) formulated an evaporation equation by establishing a relationship utilizing mean temperature and relative humidity (Rh)(Xu & Singh, 2001a):

$$\boldsymbol{ET = 0.0018(25 + T_a)^2 (100 - R_a)} \quad (2)$$

$$\boldsymbol{R_a = \frac{e^0(T_d)}{e^0(T_a)}} \qquad (2)$$

2.2.3 Jensen and Haiseequation.

Assessed over a span of 35 years and 3000 observations, the general equation was derived by Xu and Singh(Xu & Singh, 2001b):

$$\boldsymbol{\lambda ET = C_t(T - T_x)R_s} \qquad (3)$$

Subsequent modifications were made, resulting in the following expression:

$$\boldsymbol{ET = (0.0252T_{mean} + 0.078).R_s} \quad (4)$$

2.2.4 Hargreaves and Samaniequation.

This expression is expressed as:

$$\boldsymbol{ET_0 = 0.0023(T_{mean} + 17.8)(T_{max} - T_{min})^{0.5}R_a} \quad (5)$$

2.3 Metrics

Model loss assessment utilized (MSE), the mean squared error, a commonly embraced metric for errors. The MSE is determined by averaging the squared variances between the real and estimated value, as illustrated below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(ET_{0,e}\, i - ET_0 i)2$$

In assessing the effectiveness and efficiency of our model, we integrated additional commonly used metrics, including mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination R^2. These metrics are extensively applied in various studies to evaluate the (ML) models.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|ET_{0,e}i\ - ET_0 i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ET_{0,e}i - \ ET_0 i)2}$$

$$R^2 = \frac{\sum_{i=1}^{n}(ET_0{}^{i} - \overline{ET_0{}^{i}})(ET_{0,e}i - \overline{ET_{0,e}i})}{\sqrt{\sum_{i=1}^{n}(ET_0{}^{i} - \overline{ET_0{}^{i}})\sum_{i=1}^{n}(ET_{0,e}i - \overline{ET_{0,e}i})}}$$

In this context, n signifies the count of samples, $ET0$i represents the actual value of the ET0 calculated by equation (1) at the i-th instance, and $ET_{0,e}i$ denotes the predicted value. Additionally, $\overline{ET_0{}^{i}}$and$\overline{ET_{0,e}i}$are respectively to the mean of the real value and to the mean of the model value at the i-th instant.

2.4 Models

2.4.1 Hyperparameter tuning

Hyperparameter tuning is a critical step in optimizing the performance of machine learning models, including Support Vector Machine (SVM), Random Forest (RF), and XGBoost. Understanding the nuances of this process is essential for researchers and practitioners alike. In essence, hyperparameter tuning involves selecting the optimal values for parameters that govern the learning process, such as regularization strength, kernel type, and tree depth. For SVM, key hyperparameters like C (regularization parameter) and kernel choice significantly influence model performance. Similarly, in RF, parameters like the number of trees (n_estimators) and maximum depth of trees (max_depth) play pivotal roles. XGBoost, a gradient boosting algorithm, relies on hyperparameters such as learning rate, tree depth, and minimum child weight for effective tuning. Various techniques exist for hyperparameter optimization, including grid search, random search, and Bayesian optimization, each with its advantages and drawbacks. Practical recommendations suggest a nuanced approach, considering both computational resources and desired performance metrics when selecting the most appropriate tuning strategy for each model. This discussion integrates insights from seminal works in machine learning literature, providing a comprehensive understanding of hyperparameter tuning processes for SVM, RF, and XGBoost models.

2.4.2 Support Vector Regression.

SVR is a regression algorithm that extends the principles of (SVM) to the context of regression analysis. Unlike traditional regression methods, SVR operates by finding a hyperplane that best fits the data while minimizing the prediction error. SVR relies on mathematical principles derived from convex optimization and the theory of support vectors. The algorithm aims to find a hyperplane that maximizes the margin between data points, with a focus on minimizing the errors within that margin. This section will delve into the mathematical formulations of SVR, exploring how it transforms the regression problem into a dual optimization task.

Within this study, the kernels 'linear', 'poly', and 'rbf' were specifically utilized. Furthermore, diverse variations of C, epsilon, and gamma were explored to identify optimal parameters for each: ('linear', 'poly', 'rbf'), (1.1, 5.4, 170, 1001), (0.1, 0.0003, 0.007, 0.0109, 0.019), (0.7001, 0.008, 0.001).

2.4.3 RF :Random Forest Regressor.

Ensemble learning has gained prominence in machine learning for its ability to enhance predictive performance and robustness. This paper shifts its focus to the (RFR), a powerful ensemble learning algorithm designed for regression tasks.

Optimizing the performance of a Random Forest Regressor involves tuning hyperparameters such as the number of trees(n_estimators), tree depth(max_depth), and feature selection criteria. This section provides insights into the impact of hyperparameter choices on the model's accuracy and generalization, guiding practitioners in fine-tuning their models for these specific applications.Concerning the training process, RF typically demands less hyperparameter tuning. In this particular investigation, the primary hyperparameters subject to tuning include the (n_estimators), the (max_features), and the

tree depth (max_depth). The ensuing values were explored for the respective hyperparameters: (100, 200, 400, and 500 trees), (all features), and (4, 5, 6, 7, 8).
.

2.4.4 XGBoost:Extreme gradient boosting.

Boosting algorithms iteratively improve model performance by combining the strengths of multiple weak learners. XGBoost, an extension of the Gradient Boosting framework, has gained popularity for its efficiency and predictive accuracy in regression tasks. This section introduces the fundamental concepts of boosting and outlines the motivation behind exploring XGBoost for regression problems.
XGBoost, a recent proposition by Chen and Guestrin (2016), has garnered substantial attention within the machine learning domain. Similar to RF, this algorithm builds on decision trees but distinguishes itself in the construction of the tree ensemble.
During hyperparameter optimization, adjustments were made to the (n_estimators) and the (max_depth). The values experimented with for the number of trees were 500, 400, 300, and 200, while the (max_depth) underwent testing with (3, 5, 6, 7, 8, 9, and 10).

3 Results

In this section, the results obtained from the five equations are presented, followed by the results obtained from machine learning models.

3.1 The application of the five equations

After the application of the five equations, the following table presents the obtained results.

Table 1. Scores of each of the five equations

Equation	Fao56	Asce-pm	Romanenko	Hargreves-samani	Jensen and haise
Score	1	0.96	0.40	-1.33	0.75

3.2 Machine learning models results (Daily dataset)

In this section, the three models (SVR, RF, XGBoost) of machine learning were used to predict the ET0 value in different scenarios, each model is trained with each possible combination of variables.
A first scenario. Cross-validation of data sets in Meknes to prove whether the model fits the training data sets and whether it is valid locally and regionally.

Table 2. Results of the first scenario

Number of training variables	Best model	Score	Combination
Four	Random forest	0.91	Temperature,humidity,solarradiation,wind speed
Three	Support vector regressor	0.92	Temperature, solar radiation, wind speed
Two	Random forest	0.89	Temperature,solar radiation
one	Random forest	0.87	Solar radiation

Second scenario. The cross-validation, in this secondary, is characterized by mixed data.

Table 3. Results of the second scenario

Numbers variable d'entrainement	Best model	Score	Combination
Four	XGBoost	0.98	Temperature,humidity,solarradiation,wind speed
Three	Support vector regressor	0.98	Temperature, solar radiation, wind speed
Two	Random forest	0.96	Temperature,solar radiation
One	Random forest	0.87	Solar radiation

4 Discussion

The results obtained from Machine Learning models by comparing them with the five alternative equations are more accurate and useful for practical applications.

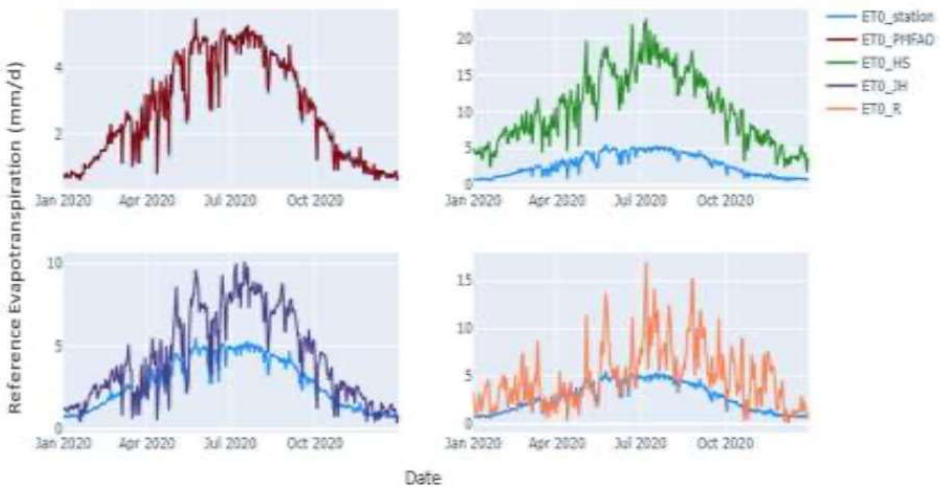

Fig. 1. Comparison between machine learning models and alternatives equations

With regard to climate change, four aspects to be taken into consideration: rise of temperature, wind speed, lowing of humidity and modification of radiation regime.
All models are trained with the combination of four variables: in the first scenario, the support vector regressor model has good accuracy with the combination of the three variables (solar radiation, temperature, wind speed).

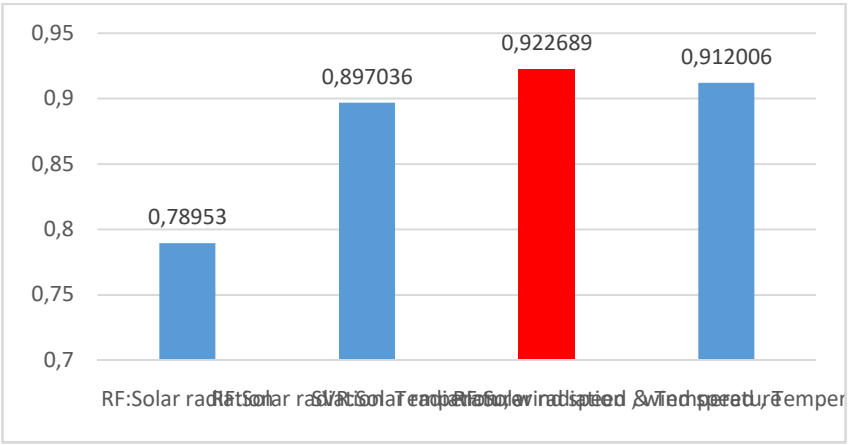

Fig. 2. Comparison of the best result of models trained with all possible combinations in the first scenario

The support vector regressor combined with the same three variables gives the best result in the second scenario.

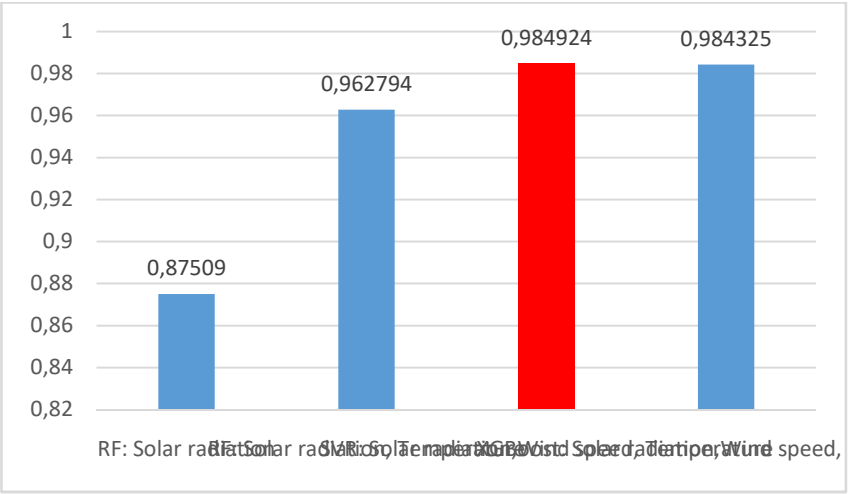

Fig. 3.Comparison of the best result of models trained with all possible combinations in the second scenario

In the realm of predicting evapotranspiration (ET) using machine learning models with minimal parameters, significant challenges arise concerning the generalizability of results across diverse regions and climatic conditions. Despite the appeal of minimal-parameter models for their computational efficiency and simplicity, their applicability beyond the training domain remains uncertain. The inherent complexities of ET estimation, influenced by multifaceted interactions between meteorological variables, land surface characteristics, and vegetation dynamics, amplify these challenges. Transferring models trained on data from one region or climatic zone to another presents formidable obstacles. Issues such as data representativeness across regions, the transferability of model features, and the robustness of algorithms to diverse climatic conditions underscore the limitations and uncertainties in generalizability. To address these concerns, strategies must be devised to enhance the generalizability of machine learning-based ET prediction models. These include incorporating domain knowledge into model development, leveraging transfer learning approaches, and conducting rigorous validation across diverse environmental settings. By embracing these strategies, researchers and practitioners can mitigate the limitations and uncertainties associated with the generalizability of minimal-parameter machine learning models for predicting evapotranspiration, fostering more reliable and robust predictions across various regions and climatic conditions.

5 Conclusion

A precise estimation for evapotranspiration is the preliminary condition for the planification and management of irrigation systems, hydrological modeling, the performance of culture, etc. Despite the existence of numerous methods to calculate ET0, it is difficult to obtain precise measures given to the complex nature of the process itself.
To reduce the number of variables, we used different algorithms of machine learning to estimate the value of evapotranspiration using different scenarios and a combination of the four variables (wind speed, temperature, humidity, and solar radiation). The obtained results from these algorithms are close to the values given by the meteorological stations

using fewer variables in the training. This demonstrates the potential of machine learning in the resolution of such problems.

References

1. Allen, R.G., Pereira, L.S., Raes, D., Smith.: Crop evapotranspiration guidelines for compu-ting crop water requirements. FAO Irrigation and Drainage, Paper No. 56, Food and Agri-culture Organization of the United Nations, Rome. (1998).
2. Saeid, M., Javad, B., K, K.: Using MARS, SVM, GEP and empirical equations for estima-tion of monthly mean reference evapotranspiration. Computers and Electronics in Agricul-ture 139, 103–114 (2017).
3. Junliang, F., Wenjun, Y., Lifeng, W., Fucang, Z., Huanjie, C., Xiukang, W., Xianghui, L., Youzhen, X.: Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. Agricultural and Forest Meteorology 263, 225–241 (2018).
4. Sevim, S.Y., Mladen, T.: Estimation of daily potato crop evapotranspiration using three dif-ferent machine learning algorithms and four scenarios of available meteorological data. Ag-ricultural Water Management 228, 105875 (2020).
5. "Grid Search: Selecting Good Support Vector Machine Parameters" par Chris Thornton, et al., publié dans le journal "Proceedings of the Ninth Australian Conference on Neural Networks" (1998).
6. 1Vu, M.T., Jardani, A., Massei, N., Fournier, M.: Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network. Journal of Hy-drology 597, 125776 (2021).
7. Mohamed, A.Y., Alazba, A.A., Mohamed, A.M.: Artificial neural networks versus gene ex-pression programming for estimating reference evapotranspiration in arid climate. Agricul-tural Water Management 163, 110–124 (2016).
8. "A Practical Guide to Support Vector Classification" par Chih-Chung Chang et Chih-Jen Lin, publié dans le journal "Technical Report, Department of Computer Science and Information Engineering, National Taiwan University" (2001).