

Estimate Reference Evapotranspiration Using Machine Learning Methods

Marwa Dorai^{1,2}^a, Mehrez Abdellaoui²^b, Bouthaina Douh³^c and Ali Douik²^d

¹University of Sousse, ISITCOM, 4011, Sousse, Tunisia

²University of Sousse, ENISO, NOCCS Research Laboratory, 4054, Sousse, Tunisia

³University of Sousse, ISA-CM, 4042 Sousse, Tunisia

marwa.dorai@isitc.u-sousse.tn, {mehrez.abdellaoui, ali.douik}@eniso.u-sousse.tn, bouthaina_douh@yahoo.fr

Keywords: Reference Evapotranspiration ET_0 , Machine Learning (ML), Internet of Things IoT, Water Stress.

Abstract: Agriculture, a fundamental pillar of human civilisation, not only provides the food we need to survive, but is also a major driver of global economic growth. Yet this critical sector is increasingly threatened by the escalating impacts of climate change, particularly through the exacerbation of water scarcity in key agricultural regions. Changing climate patterns are disrupting rainfall cycles, leading to more frequent droughts and reduced water availability. As the global population grows exponentially and demand rises, farmers require water for irrigation to meet these needs. This growing resource scarcity underscores the urgent need for innovative, sustainable agricultural solutions to adapt to these challenges. To secure the future of water resources and safeguard agricultural productivity, it is crucial to proactively implement cutting-edge technologies such as the Internet of Things (IoT) and Artificial Intelligence (AI). In this context, we present a novel approach for estimating reference evapotranspiration ET_0 with the aim of minimising water waste and improving the efficiency of irrigation water management. The study was carried out in a real-world setting where several sensors were installed to measure various parameters, including temperature, soil moisture and rainfall. The station is connected to a server application from which a dataset was generated after data cleaning and pre-processing. The parameters obtained from the dataset were classified in terms of their correlation with the output value ET_0 . Regression was then performed using various machine learning (ML) tools to predict water stress. The developed algorithms resulted in good performances in terms of coefficient of determination R^2 and loss function RMSE. These performances exceed those of existing methods from the state of the art.

1 INTRODUCTION


The water crisis has recently intensified into one of the most urgent global challenges, particularly in the Mediterranean region, where irrigation is vital for maintaining and enhancing agricultural productivity.


The available water for agriculture in this region is diminishing due to population growth and the increasing frequency of droughts. This mounting pressure on water resources necessitates the development of strategies to enhance water use efficiency and optimise the benefits derived from the available water.


one of the most effective strategies is the control of the irrigation by considering the needs of the crops.


These needs can be estimated by measuring evapotranspiration (ET), is a key player in the water cycle, moving water from soil and plants to sky through evaporation and transpiration.

Understanding and estimating ET accurately is essential for efficient water resource management and irrigation planning. Crop water requirement is a fundamental aspect of irrigation water management. The most effective way to define it is by reference evapotranspiration (ET_0). ET_0 represents water evaporated from the soil and emitted to the atmosphere by plants. Accurate ET_0 calculations are essential for a wide range of research, including irrigation planning, hydrological modeling, crop production forecasting and sustainable water resource management at both local and global scales. Additionally, ET information is used as a basis for a number of international water treaties and agreements, in particular with regard to water allocation policies. Estimation of ET_0 from a

^a <https://orcid.org/0000-0003-2442-3270>

^b <https://orcid.org/0000-0002-2492-5206>

^c <https://orcid.org/0000-0002-3439-2212>

^d <https://orcid.org/0000-0002-2492-5206>

reference area can be achieved either by mathematical modelling or by field trial data from specific sensors such as lysimeters, followed by adjustment of the ET_0 value using empirical crop coefficients. While ET_0 can be precisely measured using field experiments and lysimeters, this approach is often impractical due to the high costs and significant time and energy required. Consequently, significant investment has been made in research and development to create more efficient mathematical models for estimating ET_0 . These models typically utilise basic meteorological parameters, which are readily available locally. Significant efforts are also being made to enhance the capabilities of existing models and to develop new ones. There are various indirect approaches to estimating ET_0 , ranging from simple empirical models to more complex ones. The choice of the most appropriate model depends on several criteria, including data availability, regional characteristics and the degree of precision required.

In recent years, there has been a notable shift in the dominant methods for estimating ET_0 beginning by advances in computer technology and the emergence of numerical techniques such as ML and AI. In (Bidabadi et al., 2024), the authors have demonstrated that machine learning models, including neural networks and neuro-fuzzy systems, can outperform traditional methods such as the Penman-Monteith equation, especially in data-scarce environments. The study demonstrated that ANFIS yielded the best results in estimating ET_0 using minimal input data, including only temperature and wind speed from nearby stations. In (Yassin et al., 2016), the authors assessed the effectiveness of ANNs and gene expression programming (GEP) in estimating ET_0 in arid climate. The studies (Chia et al., 2020) and (Shrestha and Shukla, 2015) have demonstrated that ANNs and support vector machines (SVM) are effective techniques for determining and modelling actual crop ET using climatic data. In (Adnan et al., 2017), the authors used a range of machine learning techniques to develop a model for estimating ET using reduced meteorological parameters.

In this context, the use of artificial intelligence techniques, particularly ML regression models, offers significant potential for improving ET_0 estimation. Unlike traditional approaches that often suffer from the scarcity of meteorological data, ML models can integrate a wide range of parameters, going beyond conventional meteorological data. These models can learn complex relationships between different variables and provide accurate estimations even with limited or incomplete data (Yong et al., 2023).

This research presents a ground-breaking multi-

parameter method for estimating ET_0 , integrating state-of-the-art machine learning techniques and detailed data analysis. This method aims to address the shortcomings of conventional methods and provide more reliable and accurate ET_0 estimates. Thereby, we can contribute to create more efficient water resource management system and better irrigation planning in regions facing water scarcity.

The structure of this paper is as follows: following the introduction, the second section covers the approach and process, detailing the study site, the weather station setup, data collection and processing procedures, and the empirical approaches used. The third section discusses the machine learning algorithms employed. The fourth section focuses on the evaluation metrics applied in the study. The fifth section presents the results and discussion. Finally, the paper concludes with a summary of the key findings and offers suggestions for future research directions.

2 APPROACH AND PROCESS

2.1 Study Site

The study was performed at the Department of Horticultural Systems and Natural Environments Engineering, Higher Agronomic Institute of Chott Mariem, located in central-eastern Tunisia. The institute is located at 35°91' north latitude and 10°55' east longitude, at an altitude of 19 metres above sea level. This region belongs to the semi-arid bioclimatic zone, characterised by mild winters and hot summers. A meteorological station, situated 100 meters from the experimental site, provided climatic data during the study period. The average minimum and maximum temperatures were 14.94°C and 24.16°C, respectively. Relative humidity averaged 69.14 % and wind speed averaged 1.85 m/s. The average annual rainfall in the area is 183.73 millimetres, with an annual evaporation rate of 689.59 millimetres, with a five-month drought period from May to September. The region is characterized by limited and infrequent precipitation, high evaporation rates, and elevated maximum temperatures.

2.2 Weather Station Setup

The weather station is a comprehensive and autonomous device designed to measure various climatic parameters. It is equipped with several specific sensors that monitor temperature, humidity, wind speed and direction, atmospheric pressure and precipitation levels. All these sensors are integrated into a

central unit that collects data in real-time. To facilitate data transmission, the station is equipped with a Wi-Fi module that sends the collected information to the cloud every two hours. When data is required for analysis, it is pre-processed to ensure its accuracy and reliability. This pre-processing involves filtering out anomalies and outliers that may result from sensor errors or environmental disturbances.



Figure 1: Weather Station.

2.3 Data Collection and Processing Procedures

The data collection and processing phase is critical to ensuring the accuracy and usability of the climate data collected by the weather station. This phase begins with the transmission of data from the station to the cloud via a Wi-Fi module. Every two hours, the accumulated data is securely sent to a cloud-based storage platform, *Field Climate* ensuring continuous and reliable data collection.

The *Field Climate* platform provides a comprehensive solution for managing and analyzing meteorological data collected by various weather stations. It enables real-time monitoring of climatic conditions, data storage, and analysis for applications such as precision agriculture or water stress detection.

The data is then preprocessed to improve its quality using various data-cleaning techniques. For example, missing values, often caused by sensor failures or transmission errors, are handled using imputation methods such as mean, median, or last valid observation carried forward. This ensures that the dataset remains complete and suitable for analysis.

In addition, data integrity is maintained by correcting format errors and ensuring temporal consistency, eliminating inconsistencies such as out-of-sequence timestamps or duplicate entries. After cleaning, the data is formatted and stored in CSV (Comma-Separated Values) files. This standard format facilitates efficient data management and allows for easy access and analysis. The structured storage in CSV files includes feature values such as temperature, humidity, sunshine duration, and solar radiation, as well as the target variable ET_0 , calculated using the Penman formula. Table 1 illustrates all meteorological parameters computed and saved in the dataset.

Table 1: List of the meteorological parameters.

	Meteorological Parameters	Abbreviations
1	Average soil moisture	avg SM
2	Average soil temperature	avg ST
3	Max soil temperature	max ST
4	Min soil temperature	min ST
5	Average air temperature	avg AT
6	Max-Air-Temperature	max-AT
7	Min-Air-Temperature	min-AT
8	Dew Point	DP
9	Min dew point	min DP
10	Solar radiation	SR
11	Vapor pressure deficit	VDP
12	Min Vapor pressure deficit	min VDP
13	HC-Relative-humidity	HC-RH
14	Max-Relative-Humidity	Max-RH
15	Min-Relative-Humidity	Min-RH
16	Precipitation	P
17	U-sonic wind speed	U sws
18	Max wind speed	Max ws
19	Wind gust	w g
20	Delta	D
21	Max delta	Max D
22	Min delta	Min D
23	Sunshine duration	SD
24	Reference evapotranspiration	ET_0

2.3.1 Empirical Methods

A number of empirical methods have been developed for the estimation of reference evapotranspiration ET_0 . These methods employ a variety of climatic data and empirical relationships in order to provide reliable estimates. The most widely recognised of these methods are as follows:

2.3.2 Hargreaves-Samani Method (HS)

The HS method, developed by Hargreaves and Samani, employs temperature data and extraterrestrial radiation to estimate ET_0 . This method's simplicity is one of its main advantages, especially in areas with scarce climatic data (Althoff et al., 2019).

2.3.3 Thornthwaite Method

This method created by C.W. Thornthwaite, it is a popular choice due to its straightforward implementation and minimal data requirements. However, this method is more applicable to humid regions and may require adjustments for arid climates (Thornthwaite, 1948).

2.3.4 Blaney-Criddle Method (BC)

The Blaney-Criddle (BC) method, developed by H. F. Blaney and W. D. Criddle, is a widely used approach in agricultural water management for estimating crop

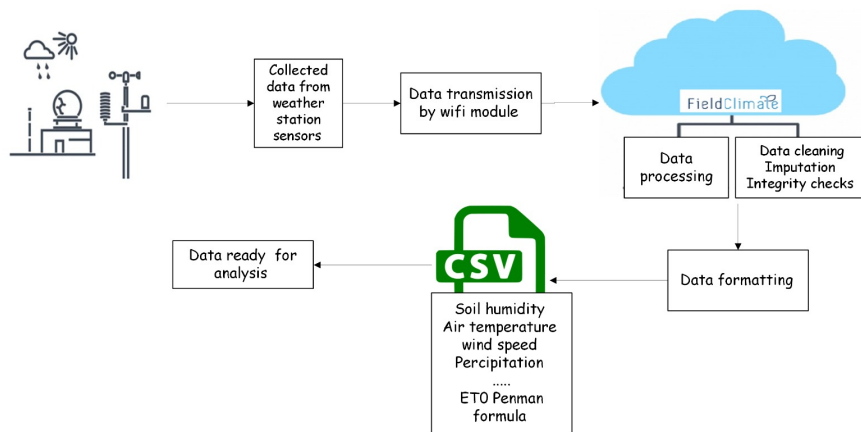


Figure 2: Data collection and processing workflow.

water requirements. It uses mean monthly temperature and the percentage of annual daylight hours. The BC method has undergone several refinements to improve its accuracy under different climatic conditions (Sobrinho et al., 2020).

2.3.5 Priestley-Taylor Method (PT)

This methodology, developed by Priestley and Taylor, modifies the Penman equation to estimate ET_0 in environments where radiation is the primary factor influencing evapotranspiration. This offers a more efficient alternative to the original Penman equation, with the introduction of an empirical coefficient, and is particularly useful for humid regions with ample solar radiation (Sobrinho et al., 2020)

2.3.6 Penman–Monteith Method (PM)

Howard Penman developed the Penman Method, which integrates energy balance and aerodynamic principles to calculate ET_0 . This method requires detailed meteorological data, making it data intensive but exceptionally accurate. The PM has been standardised by the FAO and WMO (Sobrinho et al., 2020), (Wright, 1985). Despite its complexity, the Penman method is celebrated for its precision and is widely accepted as the gold standard for ET_0 estimation. As a result, the standardised ET_0 equation (1) is used as the target variable in the modelling phases.

$$ET_0 = \frac{0.408\Delta(R_c - H) + \rho \frac{900}{T_a + 273} V_2 (P_s - P_a)}{\Delta + \rho(1 + 0.34V_2)} \quad (1)$$

Defining the variables as follows:

- ET_0 : Reference ET (mm/day).
- R_c : Crop surface net radiation (MJ/m²/day).
- H : Soil heat flux density (MJ/m²/day).

- T_a : Average daily air temperature at 2 meters above ground level(°C).
- V_2 : Two-meter wind speed (m/s).
- P_s : Saturation pressure of the water vapor (kPa).
- P_a : Actual vapor pressure (kPa).
- $P_s - P_a$: Water vapor deficit (kPa).
- Δ : Temperature coefficient of saturation vapor pressure (kPa/°C).
- ρ : Moisture content coefficient (kPa/°C).

3 MACHINE LEARNING MODELS

This research explores the application of machine learning models, including linear regression, random forest, support vector regression and extreme gradient boosting, to predict ET_0 . These methods are implemented using Python, ensuring robust and consistent results for this research endeavor.

3.1 Linear Regression

Linear Regression is a key method for modeling the relationship between a target variable and one or more predictor variables. The objective is to identify a linear equation that best fits this relationship. This is achieved by representing the target variable as a weighted sum of the predictor variables, together with an intercept. The coefficients are calculated by minimising a loss function, typically the sum of squared errors. By assuming a linear relationship, Linear Regression simplifies the modeling process, making it effective for understanding and predicting patterns in data, especially when the relationship is straightforward.

3.2 Random Forest

RF implemented by (Breiman, 2001), is an ensemble learning approach that consists of several decision tree estimators. Each tree in the forest is constructed from values derived from a randomly sampled subset of the data. The process starts at the root node of each tree and progresses downwards, evaluating all available information at each node. Predictive variables are calculated throughout this process. To prevent over-fitting, a cross-validation technique is employed, which systematically trims the trees to enhance their generalisability.

3.3 Support Vector Regressor

SVR is a machine learning approach that integrates the principles of Support Vector Machine (SVM) and is utilized for non-linear regression (Vapnik, 2013). The objective is to identify a function that accurately approximates the relationship between input variables and target values, while simultaneously minimising both error and model complexity. The process begins with fitting a linear model to the data, followed by applying a nonlinear kernel to capture more intricate patterns. The method focuses on minimising operational risk rather than just prediction error, making it an effective approach for modelling intricate data relationships.

3.4 Extreme Gradient Boosting Algorithm (XGBoost)

XGBoost, developed by (Chen and Guestrin, 2016), is a ML tool that a machine learning framework leveraging ensemble decision tree gradient boosting for high predictive accuracy. It uses shrinkage (learning rate adjustment) to fine-tune predictions and reduce overfitting. Column subsampling enhances robustness by selecting random feature subsets, reducing correlation. Tree pruning, guided by a gamma threshold, simplifies trees by removing insignificant splits, while L1 and L2 regularization penalties prevent model overcomplexity. XGBoost also handles missing values natively, learns optimal paths without imputation, and employs early stopping to avoid overfitting and save computational effort. These features make it efficient and effective for diverse ML tasks.

4 EVALUATION PERFORMANCE METRICS

The accuracy and performance of the ML models in estimating ET_0 were evaluated using two widely adopted regression metrics: the determination coefficient R^2 and the root-mean-square error (RMSE). The R^2 is employed to assess the correlation and agreement between the actual and predicted daily ET_0 values. The value of R^2 varies between 0 and 1, with $R^2 = 1$ indicating a positive correlation. Whereas RMSE is used to measure the error associated with the estimated models. This metric ranges from 0 to infinity, with lower RMSE values indicating that the model's predictions closely align with the actual values (Zhou et al., 2020) (Zhou et al., 2020). The evaluation metrics are calculated using the following equations.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_{ri} - y_{pi})^2}{\sum_{i=1}^N (y_{ri} - \bar{y}_r)^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{ri} - y_{pi})^2} \quad (3)$$

where :

- N : Total number of samples.
- y_{ri} : Real value of the i -th sample.
- y_{pi} : Predicted value of the i -th sample.
- \bar{y}_r : Mean of the real values

5 RESULTS AND DISCUSSIONS

This study uses a dataset covering the period from September 2022 to May 2024, which originally contained 9946 rows by 61 columns (parameters). After pre-processing, it was reduced to 610 rows and 24 variables. The dataset was then divided into two subsets: 80% for training and 20% for testing. To ensure reliable variable selection for reference evapotranspiration (ET_0) estimation and to avoid data leakage, correlation coefficients were computed only from the training set.

A correlation matrix is a tool for displaying correlations among several variables. The correlations between two variables is represented in each cell of the matrix. The coefficients vary between -1 and 1 and indicated the magnitude and direction of their linear relationship. A perfect correlation can only be detected by a value of 1 or -1, while a value of 0 indicates that none exists. (Agrawal et al., 2022). In this study, a correlation matrix was employed to examine the relationships between various meteorological parameters

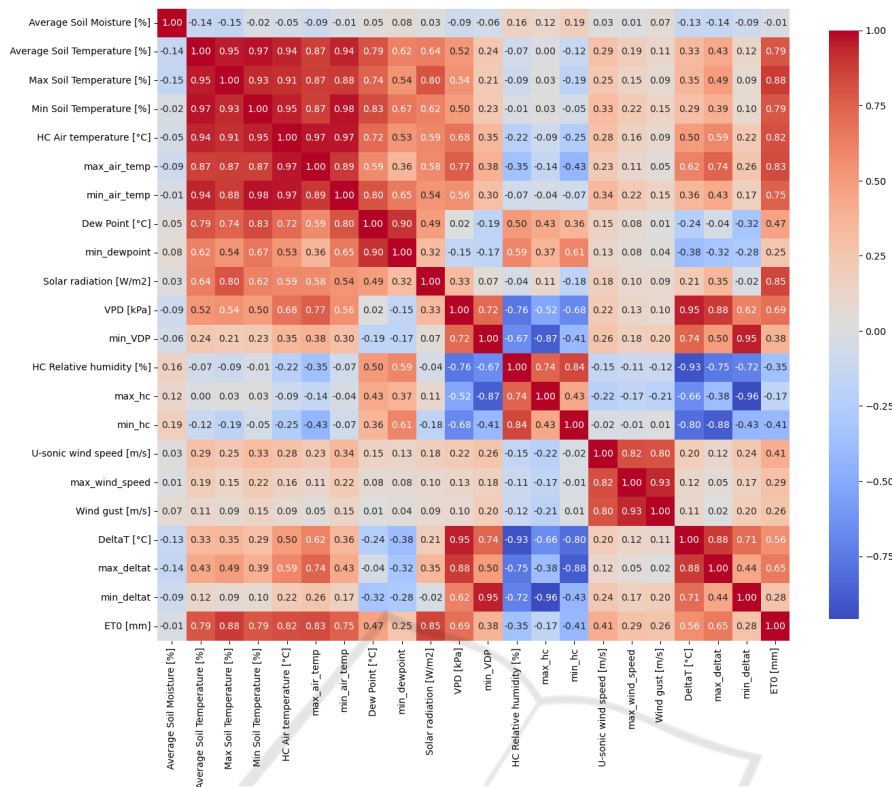


Figure 3: Inter-correlation Heatmap of Various Input Meteorological Parameters.

(inputs) and ET_0 as the output. The correlation coefficients were computed exclusively from the training data to ensure precise variable selection. The results, illustrated in Figure 3 as a heatmap, show that Max_ST has the greatest impact on ET_0 , while Avg_SM has the least significant effect.

Several tests and evaluations have been conducted to achieve the best performance of the proposed method. The idea of the test algorithm is to vary the number of parameters used in the prediction algorithm, considering their ranking from the top 3 to all parameters. The evaluation of the regression scores R^2 when varying the number of parameters involved in the algorithm is shown in the figure below.

After analysing the results obtained, we can conclude that the number of parameters used for regression is crucial to reach the best performances. If we choose 8 parameters or less, R^2 score doesn't exceed 0.94. While, this score reaches the best values when the number of parameters exceeds 13. Each regression method has a unique sensitivity to different sets of input parameters, which affects its performance and accuracy. For example, some methods may achieve optimal results with a larger set of parameters, such as LR with 22 parameters capturing more complex patterns within the data. While oth-

ers, such as XGBoost with only 14 parameters, may perform better with a more streamlined selection, reducing the potential for overfitting and focusing on the most influential variables. In addition to R^2 score, we have computed RMSE for each combination of the input parameters from the top 3 to all parameters. We can conclude that when R^2 increases the RMSE decreases. Figure 5 illustrates the evolution of RMSE values when varying the number of parameters. This variability highlights the importance of tailoring the parameter selection process to the specific characteristics and requirements of each regression method. It also highlights the need for a thorough evaluation and comparison of different models to determine the most effective approach for a given dataset and prediction task. Ultimately, the choice of parameters and regression method is crucial in determining the precision and reliability of the predictive model. To make the right choice, table 2 summarises the best scores obtained for each method and the number of parameters involved. The data in this table allows us to gain valuable insights into the performance and complexity of the four regression models. The SVR model achieves the best R^2 value of 0.9764, with the LR and XGBoost models not far behind, both achieving a value of 0.9759. The maximum R^2 for RF is 0.9622, which

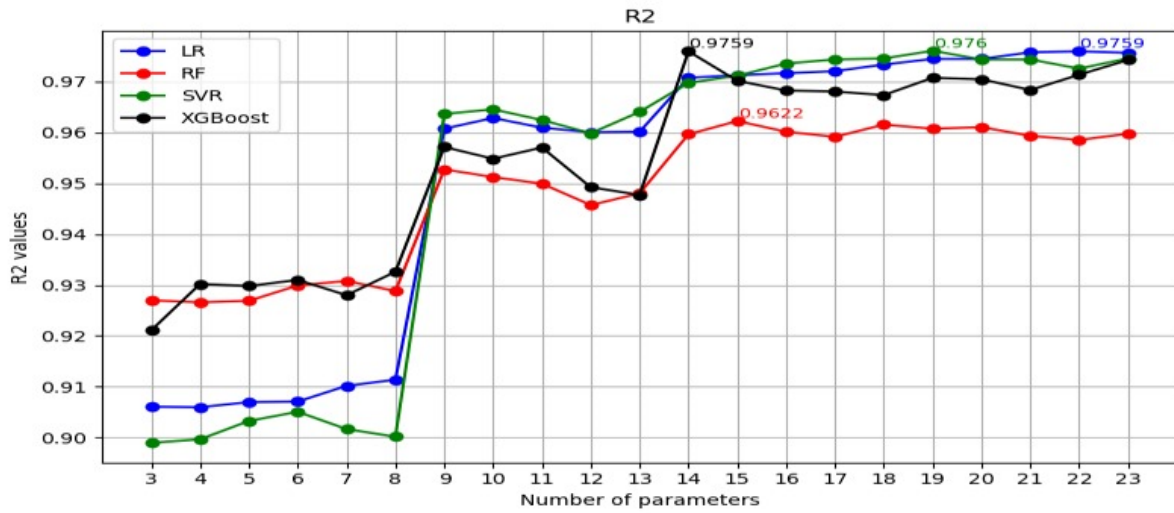


Figure 4: Evolution of R^2 score when varying the number of parameters.

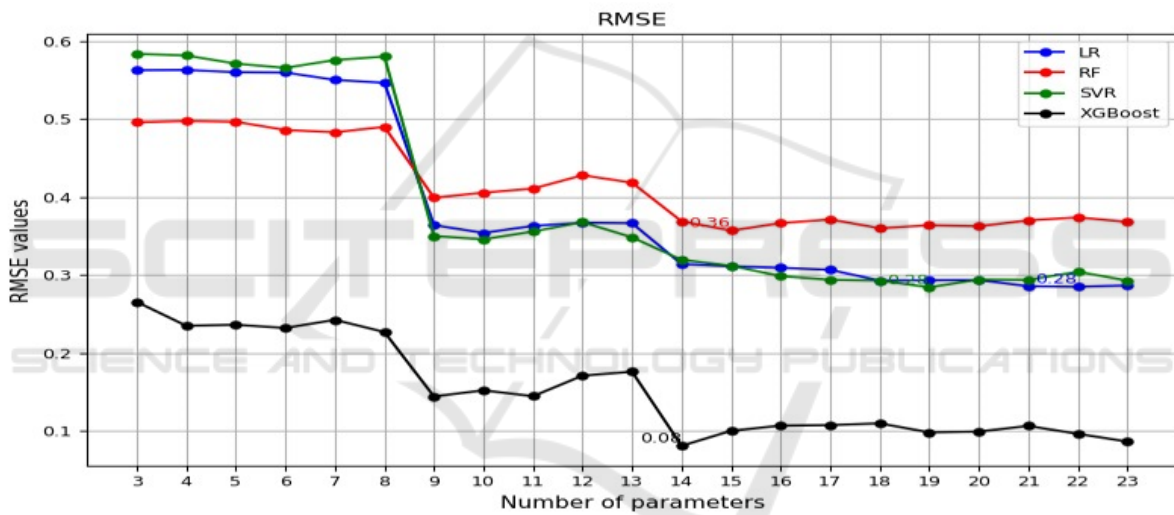


Figure 5: Evolution of RMSE score when varying the number of parameters.

Table 2: Best results of evaluation metrics.

Models	Max R^2	Min R^2	Param's numb
LR	0.9759	0.2849	22
RF	0.9622	0.3568	15
SVR	0.9764	0.2814	19
XGBoost	0.9759	0.0811	14

is still a commendable result. In terms of RMSE, XGBoost has the lowest value of 0.0811, indicating superior prediction accuracy. The next lowest values were achieved by SVR and LR, with scores of 0.2814 and 0.2849, respectively. The highest RMSE was recorded for RF, at 0.3568. In terms of the number of parameters, XGBoost uses the fewest number equal to 14, which demonstrates that it is both high-performing and relatively simple. Additionally, RF is

a relatively simple model with 15 parameters, while SVR and LR are more complex, with 19 and 22 parameters, respectively. Overall, XGBoost is the most effective model due to its lowest RMSE and minimal number of parameters, making it the optimal choice for precise predictions with reduced complexity. SVR shows excellent Max R^2 , but is slightly less effective in terms of RMSE. LR and RF are competitive, but do not outperform the other models on key metrics.

6 CONCLUSIONS

This research sought to assess the effectiveness of various machine learning models in estimating ET_0 using the FAO Penman method. The models tested included

linear regression, random forest, support vector regression, and XGBoost. These tests were conducted by varying the number of meteorological parameters, ranging from the three most correlated to ET₀ to the complete set of parameters. Our findings demonstrate that the efficacy of the models is clearly influenced by the algorithm employed and the number of parameters incorporated into the predictions. In general, more sophisticated algorithms such as SVR and XGBoost demonstrated superior performances, although each model exhibited particular strengths depending on the evaluation metrics used. In conclusion, the study emphasises the significance of algorithm selection and parameter inclusion for enhancing the precision of ET₀ estimations. The XGBoost model demonstrated particular effectiveness in terms of RMSE, indicating its capacity to provide highly accurate estimations with relatively few parameters. The choice of algorithm for ET₀ estimation is significantly influenced by the available parameters and data. For applications requiring high precision, models like SVR and XGBoost are recommended. However, future studies could focus on hyperparameter optimisation and the use of ensemble techniques to potentially further enhance estimation performance.

ACKNOWLEDGEMENTS

We are extremely grateful to the Department of Horticultural Systems and Natural Environments Engineering at the Higher Agronomic Institute of Chott Mariem for their generous support. It is a true honor to have been entrusted with access to your confidential information, and we deeply appreciate your trust in our work.

REFERENCES

- Adnan, M., Latif, M. A., Nazir, M., et al. (2017). Estimating evapotranspiration using machine learning techniques. *International journal of advanced computer science and applications*, 8(9):108–113.
- Agrawal, Y., Kumar, M., Ananthakrishnan, S., and Kumarpuram, G. (2022). Evapotranspiration modeling using different tree based ensembled machine learning algorithm. *Water Resources Management*, 36(3):1025–1042.
- Althoff, D., Santos, R. A. d., Bazame, H. C., Cunha, F. F. d., and Filgueiras, R. (2019). Improvement of hargreaves–samani reference evapotranspiration estimates with local calibration. *Water*, 11(11):2272.
- Bidabadi, M., Babazadeh, H., Shiri, J., and Saremi, A. (2024). Estimation reference crop evapotranspiration (et₀) using artificial intelligence model in an arid climate with external data. *Applied Water Science*, 14(1):3.
- Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- Chia, M. Y., Huang, Y. F., and Koo, C. H. (2020). Support vector machine enhanced empirical reference evapotranspiration estimation with limited meteorological parameters. *Computers and Electronics in Agriculture*, 175:105577.
- Shrestha, N. and Shukla, S. (2015). Support vector machine based modeling of evapotranspiration using hydroclimatic variables in a sub-tropical environment. *Agricultural and forest meteorology*, 200:172–184.
- Sobrinho, O. P. L., Júnior, W. L. C., dos Santos, L. N. S., da Silva, G. S., Pereira, Á. I. S., and Tavares, G. G. (2020). Empirical methods for reference evapotranspiration estimation. *Scientia Agraria Paranaensis*, pages 203–210.
- Thorntwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical review*, 38(1):55–94.
- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.
- Wright, J. L. (1985). Evapotranspiration and irrigation water requirements.
- Yassin, M. A., Alazba, A., and Mattar, M. A. (2016). Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. *Agricultural Water Management*, 163:110–124.
- Yong, S. L. S., Ng, J. L., Huang, Y. F., and Ang, C. K. (2023). Estimation of reference crop evapotranspiration with three different machine learning models and limited meteorological variables. *Agronomy*, 13(4):1048.
- Zhou, Z., Zhao, L., Lin, A., Qin, W., Lu, Y., Li, J., Zhong, Y., and He, L. (2020). Exploring the potential of deep factorization machine and various gradient boosting models in modeling daily reference evapotranspiration in china. *Arabian Journal of Geosciences*, 13:1–20.