A study on the prediction of evapotranspiration using freely available meteorological data*

Pedro J. Vaz[1,4][0000−0002−8819−3243], Gabriela Schütz[1,2][0000−0001−5081−3913], Carlos Guerrero[1,3][0000−0001−9907−8235], and Pedro J. S. Cardoso[1,4][0000−0003−4803−7964]

[1] Universidade do Algarve, Portugal
[2] CEOT – Centre for Electronics, Optoelectronics and Telecommunications
[3] MED – Mediterranean Institute for Agriculture, Environment and Development
[4] LARSyS – Laboratory of Robotics and Engineering Systems
{pjmartins-gschutz-cguerre-pcardoso}@ualg.pt

Abstract. Due to climate change, the hydrological drought is assuming a structural character with a tendency to worsen in many countries. The frequency and intensity of droughts is predicted to increase, particularly in the Mediterranean region and in Southern Africa. Since a fraction of the fresh water that is consumed is used to irrigate urban fabric green spaces, which are typically made up of gardens, lanes and roundabouts, it is urgent to implement water waste prevention policies. Evapotranspiration (ETO) is a measurement that can be used to estimate the amount of water being taken up or used by plants, allowing a better management of the watering volumes but, the exact computation of the evapotranspiration volume is not possible without using complex and expensive sensor systems.

In this study, several machine learning models were developed to estimate reference evapotranspiration and solar radiation from a reduced-feature dataset, such has temperature, humidity, and wind. Two main approaches were taken: (i) directly estimate ETO, or (ii) previously estimate solar radiation and then inject it into a function or method that computes ETO. For the later case, two variants were implemented, namely the use of the estimated solar radiation as (ii.1) a feature of the machine learning regressors and (ii.2) the use of FAO-56PM method to compute ETO, which has solar radiation as one of the input parameters. Using experimental data collected from a weather station located in Vale do Lobo, south Portugal, the later approach achieved the best result with a coefficient of determination (R^2) of 0.975 over the test dataset. As a final notice, the reduced-set features were carefully selected so that they are compatible with online freely available weather forecast services.

* This work was supported by the Portuguese Foundation for Science and Technology (FCT), projects UIDB/50009/2020 – LARSyS, UIDB/00631/2020 – CEOT BASE and UIDP/00631/2020 – CEOT PROGRAMÁTICO, UIDB/05183/2020 – MED and by project GSSIC – Green Spaces SMART Irrigation Control, grant number ALG-01-0247-FEDER-047030. Particular thanks to GSSIC project's companies Visualforma - Tecnologias de Informação, S.A. and Itelmatis, Lda.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

Keywords: Evapotranspiration · Machine learning · Public garden · Smart irrigation.

1 Introduction

The hydrological drought is assuming a structural character with a tendency to worsen in regions such as Algarve, Portugal. The problem is not particular to the region occurring, e.g., in most countries of the Mediterranean basin. The climate report, "Climate Change and Land", from August 2019, by the Intergovernmental Panel on Climate Change (IPCC) [18], predicts that, due to climate change, the frequency and intensity of droughts will increase, particularly in the Mediterranean region and in Southern Africa.

A fraction of the fresh water that is consumed by humans is used to irrigate green spaces in the urban fabric, which are typically made up of gardens, lanes and roundabouts, as well as green spaces in hotel and resort chains. The irrigation methodology of these green spaces is commonly done using basic irrigation controllers that are configured according to the experience of those responsible for maintaining the green spaces, without the use of information regarding climate, plants, or soils, as well as data from sensors, nearby weather stations, and from a weather forecast application programming interface (API) that can provide real-time and predicted information. Furthermore, common irrigation controllers have no connectivity, are stand-alone solutions where irrigation schedules are pre-programmed, and only in more complete versions allow irrigation inhibition by means of a rain detection sensor. This is the typical profile that can be found in the overwhelming majority of green space irrigation control systems.

Evapotranspiration (ETO) is a measurement that can be used to estimate the amount of water being taken up or used by plants, allowing a better management of the watering volumes. However, its exact computation is not possible without using complex and expensive sensor systems being many times estimated by formulas or other methods. The use of one over the other depends many times on the available weather parameters.

This paper presents part of a framework to estimate ETO supported by the use of machine learning, acquired intelligence, meteorological data from weather stations on the field, as well as meteorological data and forecasts from APIs available on the internet. The framework will include the computation of crop water requirements derived from the ETO prediction methods, providing an optimal irrigation schedule in terms of start(s) and duration(s), in order to optimize water expenditure, energy expenditure, and the well-being of the crop. This allows the development of an intelligent irrigation solution, technologically differentiated from other platforms on the market, using innovative communications technology, hardware and software, aggregating devices such as probes, field controllers, meteorological stations, among others. The development of the full framework will be done in project GSSIC – Green Spaces SMART Irrigation Control which is developing an innovative intelligent irrigation solution, in terms of reducing water consumption, reducing reaction time in solving problems, in-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

creasing efficiency in detecting anomalies, and maintaining the quality of green spaces.

The main contribution of this paper is the study and proposal of a set of methods to estimate ETO and solar radiation using features commonly available in most open weather forecast APIs.

The paper is structured as follows. The next section presents the problem's background and the methodologies used by others to tackle the problem in study. Section 3 explores the dataset and explains the computational setup. The fourth section presents the proposed methods and the associated performance analysis. The final section presents some conclusion and establishes some future work.

2 Problem's Background

Water requirements depend on the reference evapotranspiration (ETO) which is one of the fundamental parameters for irrigation scheduling as well as improving the management and use of water resources. Prediction of reference evapotranspiration for the following days plays a vital role in the design of intelligent irrigation scheduling, as it is proportional to the amount of water that will have to be restored during the irrigation period [5].

Some of the main characteristics that distinguish crop evapotranspiration (ET_c) from ETO are (i) the crop cover density and total leaf area, (ii) the resistance of foliage epidermis and soil surface to the flow of water vapor, (iii) the aerodynamic roughness of the crop canopy, and (iv) the reflectance of the crop and soil surface to short wave radiation [1]. In this context, known the crop coefficient (K_c), the crop evapotranspiration (ET_c) value for a specific time period can be estimated by

$$\mathrm{ET}_c = K_c \mathrm{ETO}. \tag{1}$$

The crop coefficient can be simple or have two components, one representing the basal crop coefficient (K_{cb}) and another representing the soil surface evaporation component (K_e), being computed by

$$K_c = K_s K_{cb} + K_e, \tag{2}$$

where $K_s \in [0, 1]$ is used to introduce a K_c reduction in cases of environmental stresses such as lack of soil water or soil salinity [1].

It is thus clear that to make a prediction of a crop's water requirements (ET_c), it is necessary to accurately estimate the reference evapotranspiration (ETO), which is the evapotranspiration of a reference surface, defined as hypothetical grass with a uniform height of 0.12 m, a fixed surface resistance of 70 sm^{-1}, and an albedo (reflection coefficient) of 0.23 [2].

Historically several deterministic methods have been developed to estimate evapotranspiration using single or limited weather parameters and are generally categorized as: temperature, radiation or combination based. Temperature based methods include Thorntwait [20], Blaney-Criddle [3] and Hargreaves and Samani [7]; Radiation methods include Priestley-Taylor [15] and Makkink [11];

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

Combination methods include Penman [14], modified Penman [4] and FAO56-PM[2].

Shahidian el al. [17] give an overview of several methods and compare their performance under different climate conditions. For most of the methods, the authors concluded that when applied to climates different from those on which they were developed and tested they can yield a poor performance and may require the adjustment of empirical coefficients to accommodate local climate conditions, which is not ideal.

The *Food and Agriculture Organization of the United Nations* (FAO) recommends using the FAO-56 Penman-Monteith (FAO-56PM) formula as a reference method for estimating ETO [2]. To give a deeper idea of the involved parameters, the formula is given by

$$ETO = \frac{0.408\Delta(R_n - G) + \gamma\frac{900}{T+273}u_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)}, \tag{3}$$

where R_n is the net radiation at crop surface $[MJm^{-2}day^{-1}]$, G is the soil heat flux density $[MJm^{-2}day^{-1}]$, T is the air temperature at 2 m height $[^oC]$, u_2 is the wind speed at 2 m height $[ms^{-1}]$, e_s is the saturation vapor pressure $[kPa]$, e_a is the actual vapor pressure $[kPa]$, $e_s - e_a$ is the saturation vapor pressure deficit $[kPa]$, Δ is the slope vapor pressure curve $[kPa^oC^{-1}]$, and γ is the psychrometric constant $[kPa^oC^{-1}]$. Being based on physical principles, the formula has become widely adopted as a standard for ETO computation since it performs well under different climate types. However, to compute ETO using FAO56-PM the following main meteorological parameters are required: temperature, solar radiation, relative humidity, and wind speed.

All parameters can be easily obtained from weather forecast APIs except for solar radiation. Solar radiation forecasting APIs are, at the moment, not common and present a high-cost penalty. So, apart from the water availability in the topsoil, being the evaporation from a cropped soil mainly determined by the fraction of the solar radiation reaching the soil surface [2], there is the need to (i) develop alternative methods for ETO estimation using limited meteorological parameters, that do not require solar radiation and are compatible with the weather parameters obtained by freely available weather forecast and historical weather data APIs, or (ii) to estimate the solar radiation itself and use it as an approximation on the solar radiation dependent methods. This is also important since in most situations a proper functioning, maintained, and calibrated weather station, with solar radiation measurement capability is not close to the area of interest.

Recently, as an alternative, several authors have used machine and deep learning to estimate ETO. For instance, Granata [6] compared three different evapotranspiration models, which differ in the input variables, using data collected in Central Florida, a humid subtropical climate. For each of this models four variants of machine learning algorithms were applied: M5P Regression Tree, Bagging, Random Forest, and Support Vector Machine (SVM). However, all three models included as input variable the net solar radiation. Wu and Fan [22]

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

evaluated eight machine learning algorithms divided in four classes: neuron based (MLP – Multilayer Perceptron, GRNN – General Regression Neural Network, and ANFIS – Adaptive Network-based Fuzzy Inference System), kernel-based (SVM, KNEA – Kernel-based Non Linear Extension of Arps decline model), tree-based (M5Tree – M5 model tree, Extreme Gradient Boosting – XGBoost), and curve based (MARS – Multivariate Adaptive Regression Spline). The methods were applied to data collected from 14 weather stations in various climatic regions of China and used only temperature or temperature and precipitation as input to the models. Daily ETO estimates were satisfactory, but can be possibly improved by including further weather parameters and using different machine learning algorithms. Ferreira *et al.* [5] used six alternative empirical reduced-set equations, such as Hargreaves and Samani [7], and compared the estimated values with the ones from an Artificial Neural Network (ANN) and SVM model. Data was collected from 203 weather stations and used for daily ETO estimation for the entirety of Brazil. Temperature or temperature and humidity was used as input features. They concluded that in general ANN was the best performing model when including, as input features, data from up to four previous days. Results were good considering that only temperature or temperature and humidity were used as inputs.

In this study, we explore and develop machine learning based ETO prediction models supported on data from a weather station placed in the Algarve region, in south Portugal, as it will be described in Sec. 3.

3 Dataset and experimental setup

Data from February 2019 up to September 2021 was collected from a weather station that uses sensors from Davis Instruments, located in Vale do Lobo, in south Portugal. The following weather parameters were periodically measured throughout the day and stored with a daily resolution: temperature (minimum, maximum, and average), dew point (minimum, maximum, and average), relative humidity (minimum, maximum, and average), solar radiation (maximum and average), wind speed (minimum, maximum, and average), wind direction, atmospheric pressure (minimum, maximum, and average), rain intensity, and precipitation.

The time series was split using a ratio of 75 % for training and 25 % for testing, naturally without shuffling. This results in test data starting from February 4, 2021 onward. Furthermore, train data was divided into 10 folders using time series cross-validation [9] and a grid search was used to tune the hyperparameters for each model that was used. As foreseeable, all presented model evaluation metric values are obtained using the test data that the models never saw while training. In this context, for model statistical evaluation and performance comparison the coefficient of determination (R^2), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used. Considering y_t the actual value and $\hat{y}_t$ the estimated value at instants $t = 1, 2, ..., n$, and $\overline{y}$ the mean value of the

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

actual samples, they are defined as $R^2 = 1 - (\sum_{t=1}^{n}(y_t - \hat{y}_t)^2)/(\sum_{t=1}^{n}(y_t - \overline{y})^2)$, $MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$, and $MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\%$.

The FAO-56PM equation, see Eq. (3), was used to compute the target ETO from the data collected from the weather station. When using solar radiation as target, the average solar radiation from the weather station was used.

During the conduction of this study the following widely known machine learning regression models were used: Ordinary Least Squares (OLS), Ridge regression (Ridge), Lasso regression (Lasso), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Decision Tree (Tree), and Random Forest (Forest) [10,19]. Table 9 (Appendix A) summarizes the sets of hyperparameters used in the grid search procedure, being the final configurations presented in the corresponding sections.

Finally, to carry out the study, Python v3.8.2, Numpy v1.20.3 [8], Pandas v1.3.2 [12,16], Scikit-learn v0.24.2 [13], and PyET v1.0.1 [21] were used. Pandas library was used for data analysis and manipulation, PyET to compute the reference evapotranspiration using the FAO56-PM method, and Sklearn is a widely used python machine learning framework that includes regressors, data preprocessing, and model metrics evaluation tools.

4 Models for ETO estimation: tuning and feature selection

This section divides in the following way. Firstly, in Sec. 4.1, a baseline method for ETO estimation using the referred ML algorithms and having as input features all weather parameters that are provided by the weather station is presented. Then, and while still using the measured solar radiation, a first attempt is made at reducing the number of input features that are used, while maintaining similar model performance metrics. In Sec. 4.2 ML ETO estimation models that do not use solar radiation as a feature are explored. In general, solar radiation either as a measurement or as a forecast is not available, hence the need to develop models that do not use it as an input feature. Finally, since solar radiation seems to be one of the main ETO drivers, in Sec. 4.3 ML solar radiation estimation models that use a reduced feature set are explored. Then, two different approaches are taken: (i) inject the solar radiation estimation into another ML model to predict ETO or (ii) use FAO56-PM formula to compute ETO having as input the estimated solar radiation.

4.1 An ETO baseline using ML methods and measured solar radiation

To establish a baseline, the ML regression models were trained using all features available in the data collected from Vale do Lobo weather station, including the measured solar radiation. Furthermore, using the set of parameters described in Tab. 9, the conducted grid search established the parameters values outlined

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

in Tab. 10 (Appendix A) as the best configurations. Table 1 summarizes the attained metrics results. It can be seen that the best performing methods are OLS, Ridge, and Random Forest regressors with the best R^2 equal to 0.981, which corresponds to a MAE of 0.18 mm and a MAPE of 5.51 %.

In a second phase, with the objective of reducing the feature set served as input to the algorithms (recall that, besides algorithms constraints, this reduction is important since many weather stations and weather APIs do not provide data for all relevant weather parameters), analysis of the Lasso coefficients and of the Random Forest feature importance was conducted, resulting in a new model with a reduced-set of features. As it can be seen on Tab. 2, it was observed that when using maximum and minimum temperature, average humidity, average wind, and average solar radiation as features, the models had similar performance to the previous results, and some even improved their metrics values. In this case, random forest gives the best R^2 score being closely followed by OLS. This is an important result since, except for solar radiation, these features are easily obtained through weather forecast APIs.

4.2 ETO estimation using ML methods with limited set of features (excluding solar radiation)

In a first attempt to use ML algorithms to directly estimate ETO without using solar radiation as a feature, and using a tuning strategy similar to the one described in Sec. 4.1 (to simplify our explanation and avoid an exhaustive description, due to space constraints, only final settings and conclusions are summarized), it was found that $Month \in \{1, 2, \ldots, 12\}$ was an important feature. I.e., when comparing with the feature-set used to obtain the results in Tab. 2,

Table 1. Comparison of several regression methods for ETO estimation using all available features, including measured solar radiation, namely: $Month$, Day, $TempMax$, $TempAvg$, $TempMin$, $HumidityMax$, $HumididtyAvg$, $HumidityMin$, $DewpointMax$, $DewpointAvg$, $DewpointMin$, $PressureMax$, $PressureAvg$, $PressureMin$, $WindMax$, $WindAvg$, $WindGust$, $RainIntensity$, $Precipitation$, $SolarRadiationAVG$, and $SolarRadiationMax$.

	OLS	Ridge	Lasso	kNN	SVM	Tree	Forest
R^2	**0.981**	0.979	0.967	0.910	0.967	0.938	0.972
MAE (mm)	**0.18**	0.19	0.22	0.40	0.21	0.34	0.20
MAPE (%)	**5.51**	5.60	6.41	10.67	6.60	9.18	5.54

Table 2. Comparison of several regression methods for ETO estimation using limited features, but including measured solar radiation, namely: $TempMax$, $TempMin$, $HumididtyAvg$, $WindAvg$, and $SolarRadiationAVG$.

	OLS	Ridge	Lasso	kNN	SVM	Tree	Forest
R^2	0.969	0.962	0.967	0.934	0.967	0.933	**0.971**
MAE (mm)	**0.21**	0.23	0.22	0.31	0.22	0.32	**0.21**
MAPE (%)	6.60	6.52	6.70	7.70	6.72	8.91	**5.89**

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

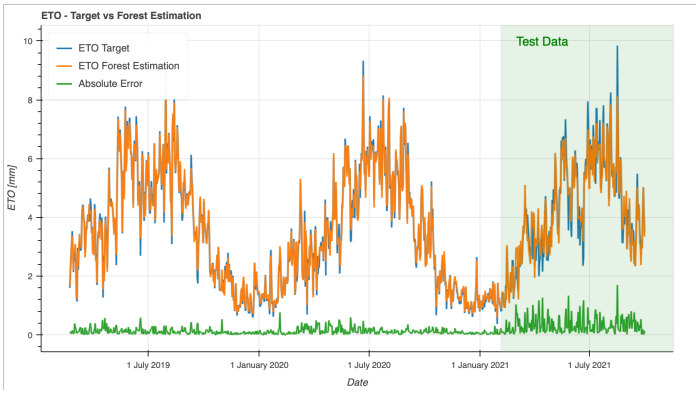

Fig. 1. Target ETO vs Random Forest estimation where solar radiation, actual or estimated, was not used as feature.

it can be seen that the used features are similar except for adding $Month$ and dropping the average solar radiation. Table 3 shows the results obtained with this set of features, and it can be clearly seen that Random Forest is the best performing model with an R^2 of 0.936, a MAE of 0.32 mm, and a MAPE of 9.11 %. Figure 1 sketches the ETO target, the ETO estimated using the Random Forest model, and the absolute error. The plot includes the predictions for the full dataset but, the shadowed region corresponds to the test data, the one used to compute the metrics, being visible the increase of the absolute error for those dates.

Maintaining the hyperparameters tuning strategy, attempts were made to improve the models' performance by doing some feature engineering, namely with new features constructed by: (i) computing the inverse of the features values (justified by the fact that some features appear in the denominator of reference FAO56-PM equation, Eq. (3)), (ii) polynomial features, and (iii) adding time lags. However, the success was minor and not noticeable to be presented here but, the idea was not abandoned as will be seen in the next sections.

4.3 ETO estimation using approximated solar radiation values

In order to try to improve the limitation and results obtained in the previous sections, a different approach was tried. The idea was to use a reduced-set of

Table 3. Comparison of several regression methods for ETO estimation using limited features, namely: $Month$, $TempMax$, $TempMin$, $HumididtyAvg$, $WindAvg$.

	OLS	Ridge	Lasso	kNN	SVM	Tree	Forest
R^2	0.856	0.855	0.859	0.893	0.855	0.814	**0.936**
MAE (mm)	0.53	0.53	0.53	0.43	0.53	0.56	**0.32**
MAPE (%)	15.15	15.20	14.78	11.93	15.16	16.24	**9.11**

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

features to previously estimate solar radiation and then either inject that solar radiation prediction into another ML regressor (with the same reduced-set features) or use FAO56-PM formula, Eq. (3), to simply approximate the ETO values. Both approaches are presented next.

Estimating solar radiation using ML methods In this section the solar radiation measured in the weather station was used as the target, i.e., the value to be estimated. Following the same tuning procedures as before (namely, the analysis of Lasso coefficients and Random Forest feature importance), the conclusion was that the best configuration for solar radiation estimation was attained for the Random Forest method with the following features: month, day, maximum and minimum temperature, average humidity, average wind speed, and average dew point. More precisely, the results presented in Tab. 4 show that the Random Forest model is the one with more satisfactory performance, with an R^2 of 0.814, a MAE of 21.31 W/m^2, and a MAPE of 11.29 %.

Again, further attempts were made to improve the models' performance by doing feature engineering such as polynomial features, inverse of features, and adding time lags. Of these, only polynomial features were helpful in improving models' performance. After individually analyzing the features' relevance for the models, it was found that by adding the following reduced-set polynomial feature $Month^2 \times Day$, the performance metrics were improved for all models, except Ridge and Lasso. The justification for such is not obvious and, as such, will not be discussed here. Detailed in Tab. 5, Random Forest is still the best performing model, now with an R^2 of 0.822, a MAE of 20.63 W/m^2, and a MAPE of 10.99 %. Figure 2 plots the target solar radiation, the approximated solar radiation obtained with the Random Forest method, and absolute error curves (shadowed is the test set). This improved solar radiation estimation will be used next to predict the ETO values.

Table 4. Comparison of several regression methods for average solar radiation estimation using limited features, namely: $Month$, Day, $TempMax$, $TempMin$, $HumididtyAvg$, $WindAvg$, $DewpointAvg$.

	OLS	Ridge	Lasso	kNN	SVM	Tree	Forest
R^2	0.532	0.505	0.382	0.580	0.312	0.594	**0.814**
MAE (W/m^2)	39.05	40.48	45.59	36.30	48.09	30.67	**21.31**
MAPE (%)	19.61	20.55	22.32	19.33	23.26	16.34	**11.29**

Table 5. Comparison of several regression methods for average solar radiation estimation using polynomial features, namely: $Month$, Day, $TempMax$, $TempMin$, $HumididtyAvg$, $WindAvg$, $DewpointAvg$, $Month^2 \times Day$.

	OLS	Ridge	Lasso	kNN	SVM	Tree	Forest
R^2	0.553	0.313	0.375	0.590	0.400	0.605	**0.822**
MAE (W/m^2)	38.04	49.08	46.53	32.48	43.87	30.36	**20.63**
MAPE (%)	18.99	24.74	23.4	16.69	22.21	16.33	**10.99**

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

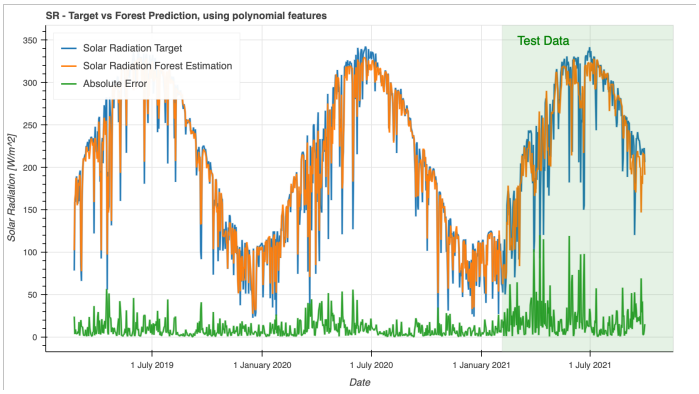

Fig. 2. Target solar radiation vs Random Forest estimation with polynomial features.

ETO estimation using ML and the approximated solar radiation The predicted values from the best performing solar radiation estimation model in the previous section, which was the Random Forest model with polynomial restricted features (see Tab. 5), were injected as a feature together with maximum temperature, average humidity and average wind into the early studied methods to estimate the ETO, being the obtained results summarized in Tab. 6. It can be seen that the Random Forest is the best performing model, with an R^2 of 0.951, a MAE of 0.26 mm and a MAPE of 7.44 %. This result is close to the before presented ETO baseline that used ML and limited features, but included the measured solar radiation (see Tab. 2). As a reference, in that case, the MAPE was equal to 5.89 %. As before, some feature engineering was tested but brought no further improvement. Figure 3 (top) plots the target ETO, the estimated ETO, and corresponding error curves for the train and test (shadowed) dataset.

ETO estimation using FAO56-PM equation and the approximated solar radiation To finalize our study, a hybrid approach was tested. In this case, the predicted solar radiation is used with the FAO56-PM equation to estimate target ETO, being the results shown on Tab. 7. With an R^2 of 0.975, MAE of 0.18 mm and MAPE of 5.51 % over the unseen test data, this result is better than any of the previously obtained ones, even better than the ML reduced-set

Table 6. Comparison of several regression methods for ETO estimation using limited features and previously estimated solar radiation, namely: $TempMax$, $HumididtyAvg$, $WindAvg$, SolarRadAVG_prediction_forest

	OLS	Ridge	Lasso	kNN	SVM	Tree	Forest
R^2	0.944	0.944	0.893	0.934	0.937	0.921	**0.951**
MAE (mm)	0.30	0.30	0.38	0.32	0.31	0.37	**0.26**
MAPE (%)	9.36	8.99	9.64	8.58	9.60	10.54	**7.44**

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

Table 7. Result obtained when computing ETO using FAO56-PM equation and using as solar radiation the previously calculated prediction from best performing Random Forest model.

	FAO56-PM + SR_pred
R^2	0.975
MAE (mm)	0.18
MAPE (%)	5.51

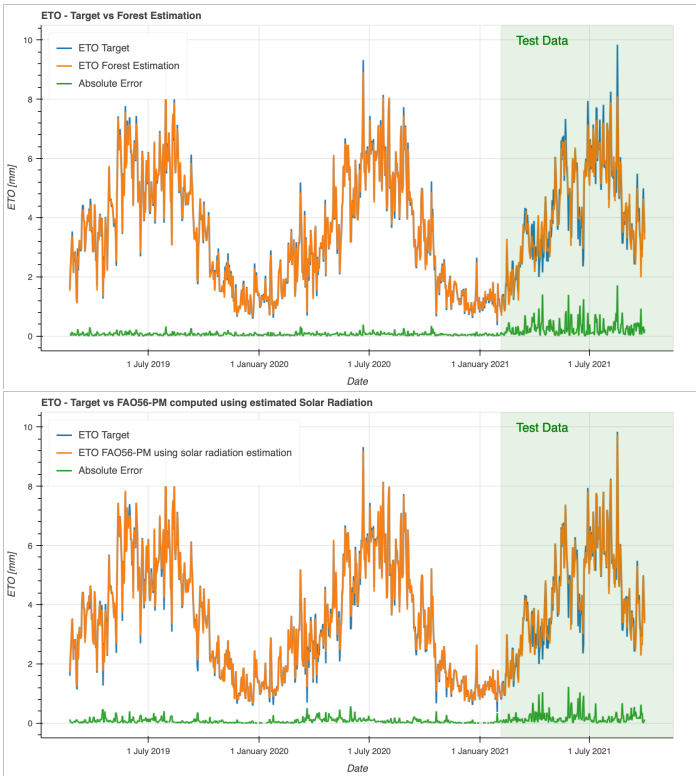

Fig. 3. Target ETO vs Random Forest estimation (top) and FAO56-PM equation (bottom) using estimated solar radiation.

baseline (see Sec. 4.1) that used the weather station measured solar radiation as a feature. Figure 3 (bottom) plots ETO target, estimated ETO, and error curves, being evident the improvement in the error when compared with the top plot.

In short, Tab. 8 presents an overview of the best ETO estimators that where previously presented. Comparing the first two columns it can be concluded that when the measured solar radiation is available, the use of a reduced-set has low impact on model performance. Further, the use of previously estimated solar radiation (last two columns) improves results when solar radiation measurement

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

Table 8. Overview of the best ETO estimators for each method that was presented.

	Measured solar radiation		No solar radiation	Estimated solar radiation	
	Table 1	Table 2	Table 3	Table 6	Table 7
R^2	**0.981**	0.971	0.936	0.951	0.975
MAE (mm)	**0.18**	0.21	0.32	0.26	**0.18**
MAPE (%)	**5.51**	5.89	9.11	7.44	**5.51**

is not available. The hybrid method (last column) gives similar results to those of the ML baseline when using all the weather parameters provided by the weather station, and gives better performance than the reduced-set ML baseline that used the actual measured solar radiation.

5 Conclusion and future work

In this study, several ML models and a hybrid approach for the ETO estimation were tested with different degrees of success. Since solar radiation is the main ETO driver, as stated by several authors and also concluded by us, models were also developed for estimating solar radiation using features usually available in the common weather forecast APIs. This allowed both the injection of the previously estimated solar radiation in ML regressors to estimate ETO, but also the possibility to use the hybrid approach where solar radiation is previously estimated and then FAO56-PM algorithm is used to finally compute ETO. The latter yielded the best results, with an R^2 of 0.975, a MAE of 0.18 mm and an MAPE of 5.51 %, which when compared with other authors works, is a good result considering the limited weather parameter features that were used.

Future work will include the use of other prediction methods (such as, recurrent neural network models) and a more extensive dataset, by using the existing weather station infrastructure that is installed in the Algarve region, in south Portugal. The objective will be to develop local and pooled models of ETO predictors for the Algarve region. Also, since all limited feature models here presented are compatible with freely available weather forecast APIs a study needs to be made to assess the impact of using such APIs as input data to the ML models here developed.

References

1. Allen, R.: Crop coefficients. Encyclopaedia of Water Science (2003)
2. Allen, R.G., Pereira, L.S., Raes, D., Smith, M., et al.: Crop evapotranspiration-guidelines for computing crop water requirements-FAO irrigation and drainage paper 56. FAO, Rome **300**(9), D05109 (1998)
3. Blaney, H.F., Criddle, W.D.: Determining consumptive use and irrigation water requirements. No. 1275, US Department of Agriculture (1962)
4. Doorenbos, J.: Guidelines for predicting crop water requirements. FAO irrigation and drainage paper **24**, 1–179 (1977)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

5. Ferreira, L.B., da Cunha, F.F., de Oliveira, R.A., Filho, E.I.F.: Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM: A new approach. Journal of Hydrology **572**, 556–570 (May 2019). https://doi.org/10.1016/j.jhydrol.2019.03.028
6. Granata, F.: Evapotranspiration evaluation models based on machine learning algorithms. Agricultural Water Management **217**, 303–315 (2019)
7. Hargreaves, G.H., Samani, Z.A.: Estimating potential evapotranspiration. Journal of the Irrigation and Drainage Division **108**(3), 225–230 (1982)
8. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (Sep 2020). https://doi.org/10.1038/s41586-020-2649-2
9. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts, Australia, 3rd edn. (May 2021)
10. Kishore Ayyadevara, V.: Pro machine learning algorithms. APRESS, New York, NY, 1 edn. (Jul 2018)
11. Makkink, G.F.: Testing the Penman formula by means of lysimeters. Journal of the Institution of Water Engineers **11**, 277–288 (1957)
12. McKinney, W.: Data Structures for Statistical Computing in Python. In: Stéfan van der Walt, Jarrod Millman (eds.) Proceedings of the 9th Python in Science Conference. pp. 56 – 61 (2010). https://doi.org/10.25080/Majora-92bf1922-00a
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
14. Penman, H.L.: Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **193**(1032), 120–145 (1948)
15. Priestley, C.H.B., Taylor, R.J.: On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly weather rev. **100**(2), 81–92 (1972)
16. Reback, J., Jbrockmendel, McKinney, W., Van Den Bossche, J., et al.: pandas-dev/pandas: Pandas 1.4.1 (2022). https://doi.org/10.5281/ZENODO.3509134
17. Shahidian, S., Serralheiro, R., Serrano, J., Teixeira, J., Haie, N., Santos, F.: Hargreaves and other reduced-set methods for calculating evapotranspiration. IntechOpen (2012)
18. Shukla, P., Skea, J., Calvo Buendia, E., Masson-Delmotte, V., Pörtner, H., Roberts, D., Zhai, P., Slade, R., Connors, S., Van Diemen, R., et al.: IPCC, 2019: Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. Intergovernmental Panel on Climate Change (2019)
19. Skiena, S.S.: The Data Science Design Manual. Springer International Publishing (2017). https://doi.org/10.1007/978-3-319-55444-0
20. Thornthwaite, C.W.: An approach toward a rational classification of climate. Geographical review **38**(1), 55–94 (1948)
21. Vremec, M., Collenteur, R., X.Z.: PyEt – open source python package for calculating reference and potential evaporation (v1.0.1). Zenodo (2021)
22. Wu, L., Fan, J.: Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration. PloS one **14**(5), e0217520 (2019)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37

A Machine Learning algorithms parameters

Table 9. Sets of hyperparameters used in the grid search procedure (according to the available parameters in the Scikit-learn suite [13])

Model	Hyperparameters	Range explored
OLS	fit_intercept	{False; True}
	normalize	{False; True
Ridge	alpha	$\{10^{-4}, 10^{-3}, \ldots, 10^{2}\}$
	fit_intercept	{False; True}
	normalize	{False; True}
Lasso	alpha	$\{10^{-4}, 10^{-3}, \ldots, 10^{2}\}$
	fit_intercept	{False; True}
	normalize	{False; True}
KNN	n_neighbours	{1, 2, . . . , 7}
	weights	{uniform; distance}
	leaf_size	{1, 3, 5, 10, 20, 30, 40}
SVM	C	{0.01, 0.1, 0.5, 1.0, 10.0, 100}
	max_iter	10000
	fit_intercept	{False; True}
DT	splitter	{best; random}
	criterion	{mse; friedman_mse; mae; poisson}
Forest	n_estimators	{10, 100, 250, 500, 750, 1000}
	min_samples_leaf	{1, 2, 3, 5, 10}
	max_depth	{3, 5, 10}
	criterion	{mse}
	max_features	{None; sqrt; log2}

Table 10. Tunned parameters

Model	Hyperparameter	Table 1	Table 2	Table 3	Table 4	Table 5	Table 6
OLS	fit_intercept	True	False	False	True	True	False
	normalize	True	False	False	False	True	False
Ridge	alpha	100	0.1	100	100	1	0.1
	fit_intercept	True	True	False	True	True	True
	normalize	False	True	False	False	True	True
Lasso	alpha	0.1	0.1	0.1	20	100	0.01
	fit_intercept	True	False	False	False	False	True
	normalize	False	False	False	False	False	True
KNN	n_neighbours	4	2	2	4	2	2
	weights	distance	distance	distance	distance	distance	distance
	leaf_size	1	1	1	1	1	1
SVM	C	0,01	0,01	0,01	0,1	0,1	0.01
	max_iter	10000	10000	10000	10000	10000	10000
	fit_intercept	True	False	False	False	True	False
DT	splitter	Random	best	Random	best	best	Random
	criterion	friedman_mse	mae	friedman_mse	friedman_mse	friedman_mse	friedman_mse
Forest	n_estimators	100	1000	500	100	100	1000
	min_samples_leaf	1	1	1	1	1	1
	max_depth	10	10	10	10	10	10
	criterion	mse	mse	mse	mse	mse	mse
	max_features	None	None	None	None	None	None

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08760-8_37