J. Hydrol. Hydromech., 71, 2023, 4, 425–438
https://doi.org/10.2478/johh-2023-0032
© Barbora Považanová et al., published by Sciendo. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Licence.

sciendo Journal of Hydrology and Hydromechanics Institute of Hydrology SAS

Using feature engineering and machine learning in FAO reference evapotranspiration estimation

Barbora Považanová[1], Milan Čistý[1*], Zbyněk Bajtek[2]

[1] Faculty of Civil Engineering, Slovak University of Technology in Bratislava, Radlinského 11, 810 05 Bratislava, Slovakia.
[2] Institute of Hydrology, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovak Republic.
[*] Corresponding author. Tel.: +421 904 002 302. E-mail: milan.cisty@stuba.sk

Abstract: The authors of this study investigated the use of machine learning (ML) and feature engineering (FE) techniques to accurately determine FAO reference evapotranspiration (*ETo*) with a minimal number of climate variables being measured. The recommended techniques for areas with insufficient measurements are based solely on daily temperature readings. Various ML methods were tested to evaluate how sophisticated an ML algorithm is for this task necessary. The main emphasis was on feature engineering, which involves converting raw variables into inputs better suited for ML algorithms, resulting in improved results. FE methods for estimating evapotranspiration include approximations of clear-sky solar radiation based on altitude and Julian day, approximate relative humidity and wind velocity, a categorical month variable, and variables interactions. The authors confirmed that the ability of ML in such tasks is not solely dependent on choosing the suitable algorithm but also on this frequently ignored step. The results of computational experiments are presented, accompanied by a comparison of the proposed method against standard *ETo* empiric equations. Machine learning methods, mainly due to the transformation of raw variables using FE, provided better results than traditional empirical methods and sophisticated ML algorithms without FE. In addition, the authors tested the applicability of the developed models in the broader area to evaluate the possibility of their generalizability. The potential of this approach to deliver improved predictions, reduced input requirements, and increased efficiency holds interesting promise for optimizing water management strategies, irrigation planning, and decision-making within the agricultural sector.

Keywords: Reference evapotranspiration; Input data reduction; Machine learning; Feature engineering.

INTRODUCTION

Water loss to the atmosphere occurs through the compound process of evapotranspiration, which involves the evaporation from different surfaces and the transpiration of water through plants. It is a crucial water cycle component in the Earth's climate system. To estimate the actual evapotranspiration in specific crops or landscapes, the potential evapotranspiration from a well-watered defined surface, known as reference evapotranspiration (*ETo*), is used. Computing reference evapotranspiration (*ETo*) is an essential step in the evaluation of the water balance in watersheds (Silva-Júnior et al., 2021), drought threat analysis, estimating crop water requirements (Mehta and Pandey, 2015), irrigation scheduling, and various other tasks in water resources management. The Food and Agriculture Organization (FAO) provides a widely used method for calculating *ETo* based on several meteorological variables, including temperature, humidity, wind speed, and solar radiation (Allen et al., 1998). However, the availability of these variables is often limited in many regions, making it challenging to estimate *ETo* accurately. In addition, evapotranspiration is a process that occurs in natural conditions on spatially inhomogeneous input data (especially in complicated terrain), including the difficulty of capturing the behaviour of vegetation (the most complex part is the calculation of the canopy surface resistance). Hence, various limitations of the mathematical description of many aspects of this physical process exist. The use of machine learning (ML) algorithms, which are suitable for modelling processes that are difficult to understand physically and describe mathematically, is recognized to be a potential solution to this problem, as ML can learn patterns and relationships directly from the data. This work uses ML to solve the regression problem, which aims to achieve the most accurate *ETo* calculation with minimal inputs.

Several studies have investigated using ML algorithms to estimate *ETo* with fewer input variables than the original FAO Penman-Monteith formula demands. For example, Dimitriadou and Nikolakopoulos (2022) explored using multiple linear regression models to predict reference evapotranspiration with limited meteorological data for the Peloponnese region in Greece. They identified the most effective combination of input variables for accurate estimates. Similarly, Mostafa et al. (2023) investigated the efficiency of the random vector functional link and relevance vector machine, improved with quantum-based avian navigation optimizer in modelling *ETo* using only minimum temperature, maximum temperature, and extraterrestrial radiation and obtained a satisfactory performance. Other ML algorithms, e.g., deep learning or tree-based models, have also been used to estimate *ETo* (e.g., Roy et al., 2022; Sattari et al., 2020; Valle Júnior et al., 2021; Wang et al., 2019) but also other parameters in hydrology (Chen et al., 2020; Gomes and Blanco, 2021).

To gain a comprehensive understanding of the existing literature on this topic, we recommend referring to review articles (e.g., Krishnashetty et al., 2021). While several studies on machine learning techniques for estimating reference evapotranspiration have been published recently, some knowledge gaps persist. This study aims to address some of these gaps.

Firstly, the authors consider Occam's Razor principle in testing the effectiveness of both simple ML methods and complex ML algorithms to compute *ETo*. This analysis seeks to determine whether more complicated ML algorithms are indeed necessary for this task.

In addition, we examine several proposed empirical formulas for calculating *ETo* that require fewer measured data. Previous works on this topic have usually not consistently undertaken such comparisons, prompting authors to fill this verification gap. If ML methods are to be helpful and accepted by the professional public, they must be clearly more advantageous concerning the standard empirical equations that have been used for a long time.

This work mainly focuses on the potential benefits of feature engineering in proposed ML models. As far as the authors know, this has not been systematically studied. Its essence is transforming raw data into more sophisticated and meaningful inputs. Sometimes the authors' efforts in ML model applications concentrate only on selecting the most robust ML model. However, data preparation is often the element that will make it possible to achieve better results, as demonstrated in this work. This step enhances the model's ability to capture relevant patterns in data and various mutual relationships in data and, as a result, improves predictive model performance.

Furthermore, authors briefly tested the suitability of the developed ML models for computing *ETo* not only in a climate station, from which were obtained calibration data (and in a small buffer around it), but also for a larger area.

In the remainder of this paper, the authors present a detailed description and verification of proposed method for estimating reference evapotranspiration using machine learning techniques. In the Data and study area section, the paper describes the site and meteorological data used, including the location and duration of the measurements. Next, in the Methods section, the authors briefly overview the ML techniques used in proposed approach and explain the feature engineering techniques employed to improve their performance. The authors then present the results of computational experiments, including a comparison of the proposed method with several existing *ETo* estimation methods (also non-ML) in the Results and Discussion section. The authors interpret their findings in this section and discuss suggestions, limitations, and potential future research. Overall, this article presents various refinements for *ETo* estimation using machine learning techniques, potentially improving water resource management and irrigation scheduling.

Data and study area

The input data in this paper comes from the stations Hurbanovo, Poprad-Tatry and Košice. All stations are in Slovakia, Central Europe. The station Hurbanovo is situated in the southwest of Slovakia, with a longitude of 18.1833°E and a latitude of 47.8666°N. The station Poprad-Tatry is in the northeast of Slovakia, with a longitude of 20.2333 and a latitude of 49.0666. The Košice station is located in the east of Slovakia with a longitude of 21.2166°E and latitude of 48.6666°N. Slovakia is a Central European landlocked country with a moderately continental climate (Fig. 1). Four seasons alternate throughout the territory, with dry, hot summers and wet, cold winters. Altitude is an important factor in weather changes and the amount of precipitation. The Hurbanovo and Košice stations are located in the lowlands, with an altitude of 115 m a.s.l. (Hurbanovo) and 230 m a.s.l. (Košice). The Poprad-Tatry station is in the highest Slovak mountain region, the High Tatras, at an altitude of 694 m a.s.l. The average annual rainfall between 1951 and 2022 was 544 mm at the Hurbanovo station, 604 mm at the Poprad-Tatry station, and 615 mm at the Košice station.



Fig. 1. Study area.

To ensure the reproducibility of the presented research, all data were taken from the ECA&D database (Klein Tank et al., 2002), where they are freely available. Only point data directly from the mentioned stations from this database were used, not data from the interpolated grid form of the ECA&D database, which is available as netCDF files. Data is offered in this database from 1951 to a date close to the present. The modelled variable was the reference evapotranspiration of the Košice station. The input data to models were various combinations of the following climate data: minimum daily temperature (*Tmin*), maximum daily temperature (*Tmax*), daily sunshine hours (*n_sun*), relative air humidity (*RH*), cloud cover (*clouds*), wind speed (*wind*). A total of 25260 days with data were used. There was no need to analyze and impute missing data, as the creators of the ECA&D database provided such activities. The data was divided into training (aka calibration set) and testing sets in the ratio of 70:30. The basic characteristics of these two sets are listed in Table 1. It can be observed that the characteristics of the training and testing data acquire similar values, which is appropriate for machine learning.

Table 1. Statistical characteristics of training and testing data (*Tmin* – minimal daily temperature, *Tmax* – maximal daily temperature, *n_sun* – sunshine hours, *RH* – relative humidity, *wind* – wind velocity, *clouds* – cloud cover, *ETo* – reference evapotranspiration).

Variable	Train data				Test data			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Tmin (°C)	4.0	7.9	–24.2	21.1	4.5	7.9	–26.9	22.7
Tmax (°C)	13.5	10.1	–13.6	36.4	14.0	10.1	–16.0	37.4
n_sun (*h*)	4.9	4.2	0.0	14.7	5.0	4.4	0.0	15.1
RH (%)	75.9	12.2	34.0	100.0	74.5	13.2	26.0	100.0
clouds (oktas)	4.9	2.5	0.0	8.0	4.8	2.4	0.0	8.0
wind (m/s)	3.7	2.9	0.0	20.3	3.4	2.5	0.0	23.0
ETo (mm)	2.5	1.9	0.0	6.9	2.5	1.9	0.0	7.1

The authors wanted to demonstrate that, for some purposes, other data sources than the primary dataset could be used if it complemented it suitably. Some climatic variables may not be included in the base dataset because they were not measured. In the work, we show how we can also use additional sources of data, e.g., with shorter time periods, which, for example, do not cover the entire period of calculating the target variable *ETo*. To demonstrate such a possibility, see the Methodology section where feature engineering is described; we used the CarpatClim database, which provides data only up to 2010 (Klein Tank et al., 2002; European Commission). Data from this database was used to compute 366 relative humidity and wind speed values for Julian days. Eight countries collaborated in compiling the CarpatClim database, which contains data on 16 climatic variables. It covers the region of the Carpathians mountains and provides data in daily, monthly, and annual time steps for the 1961–2010 period. The grid database is available in a spatial scope of 44°N – 50°N > 17°E –27°E and in a spatial resolution of 0.1°x 0.1° (Szalai et al., 2012). The climate data (relative humidity and wind speed) used are from the grid point near Košice station, with coordinates 48.7°N latitude and 21.2°E longitude.

METHODS

In this chapter, the following methodological issues are addressed. Firstly, selected empirical equations and machine learning methods for computing evapotranspiration are briefly described. Their more detailed description can be found in the referenced literature. Next, the fine-tuning of ML methods is explained, as it is the main issue of their application. Additionally, we present data preparation for ML models to explain and underscore the benefit of feature engineering.

Empirical equations for estimation of *ETo*

Reference evapotranspiration (*ETo*), according to the FAO methodology (FAO56), is a standardized measure that represents the evaporative demand of the atmosphere, quantifying the amount of water that would be transpired by a reference crop under specific meteorological conditions (Allen et al., 1998). This method is often used to evaluate the accuracy of other empirical equations and machine learning approaches in estimating *ETo*. The following equation describes this method used as a reference method also in this paper:

$$ET_o = \frac{o.408 \cdot \Delta \cdot (R_n - G) + \gamma \cdot \frac{900}{T+273} \cdot u_2 \cdot (e_s - e_a)}{\Delta + \gamma \cdot (1 + 0.34u_2)} \quad (1)$$

ETo = reference evapotranspiration (mm day^{-1}), Δ = slope vapour pressure curve (kPa °C^{-1}), R_n = net radiation at the crop surface (MJ m^{-2} day^{-1}), G = soil heat flux density (MJ m^{-2} d^{-1}), γ = psychrometric constant (kPa °C^{-1}), T = mean daily temperature in °C, u_2 = wind speed at 2 m (km h^{-1}), es = saturation vapour pressure (kPa), e_a = mean actual vapour pressure (kPa).

Considering the primary purpose of this article – to develop a procedure for the most accurate determination of reference evapotranspiration with fewer variables – only empirical equations that use fewer variables than the FAO method were investigated. The following methods were selected based on preliminary calculations regarding their accuracy and information from the scientific literature (Yeh, 2017).

The Blaney-Criddle method is a temperature-based method (Blaney et al., 1952). This is one of the four methods presented in FAO Irrigation and Drainage Paper 24 (Doorenbos and Pruitt, 1977). The method is recommended for areas where temperature is the only available climate variable.

$$ETo = p \cdot (0.46 \cdot T+8) \quad (\text{mm day}^{-1}) \quad (2)$$

p = annual daylight percentage of total daytime hours for a given month and latitude, T = the average daily temperature (°C).

Hargreaves-Samani method is a radiation-based method of calculating reference evapotranspiration (Hargraeves and Samani, 1985). The authors of the equation aimed to develop a simple and practical equation that can be used to calculate crop water requirements with a limited number of climatic variables. The method does not require local calibration and uses only maximum and minimum temperatures as input data. The *Ra* value is calculated from the station's latitude; a detailed description is given in Hargreaves and Samani (1985). The equation is as follows:

$$ETo = 0.0023 \cdot Ra \cdot TD^{0.5} \quad (\text{mm day}^{-1}) \quad (3)$$

Ra = extraterrestrial radiation (mm), *TD* = mean maximum and minimum temperature (°C) difference.

McGuinness-Bordne method is a method for calculating potential evapotranspiration (McGuinness and Bordne, 1972). It is also a radiation-based method, using extraterrestrial radiation. The equation is as follows:

$$PE = \frac{Ra}{\lambda\rho} \cdot \frac{Ta+5}{68} \quad (\text{mm day}^{-1}) \quad (4)$$

Ra = extraterrestrial radiation (MJ m^{-2} day^{-1}), depending only on latitude and Julian day, λ = latent heat flux (taken equal to 2.45 MJ kg^{-1}), ρ = density of water (kg m^{-3}), Ta = mean daily air temperature (°C).

Makkink equation is a modification of the original Penman's formula presented by Makkink (1957). It is a radiation-based method. The advantage of this equation is that its behavior is very similar to the basic Penman equation. The calculation includes air temperature and global radiation, which is a measured variable. The necessity of this variable is also a disadvantage of this equation compared to others mentioned in this work.

$$Eto = 0.61 \cdot \left(\frac{\Delta}{\Delta+\gamma} \cdot \frac{Rs}{\lambda}\right) - 0.12 \quad (5)$$

Rs = solar radiation (MJ m^{-2} d^{-1}), Δ = slope of the saturated vapour pressure curve (kPa °C^{-1}), γ = humidity constant (kPa °C^{-1}), λ = latent heat of evaporation (MJ kg^{-1}) (Yeh, 2017).

Machine learning methods

ML methods were chosen from simpler to more complex, as we want to demonstrate whether more complex methods are even necessary for this task. For many tasks, linear regression may be sufficient, so we used it here in its basic and in a regularized form (LASSO) for comparison. Since the "no free lunch" theorem says that we do not know in advance which ML algorithm will be the most suitable for a given task, we chose three principally different ML algorithm types: the random forest representing tree methods, the support vector regression representing kernel methods and the CatBoost representing boosting methods. To further enhance our analysis, we employed the ensemble stacking regression, which is the most complex ML technique used in this study.

The detailed mathematical explanation has been omitted from the algorithm description below, as it can be found in the cited sources.

Multiple linear regression (MLR) solves the linear relationship between independent and dependent variables. This simple, well-known methodology was included here mainly for comparison with other, more powerful methods. The equation for multiple linear regression is:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + \varepsilon \quad (6)$$

Y = dependent variable, X_1, X_2, …, X_p = independent variables, β_0 = intercept term, β_1, β_2, …, β_p = coefficients for each independent variable, and ε is the error term.

Coefficients β_0, β_1, β_2, …, β_p are computed using the least square method, which minimizes the sum of the differences between the actual values of Y and those calculated by the above equation.

LASSO (Least Absolute Shrinkage and Selection Operator) is a regularization method. It was introduced in 1996 by (Friedman et al., 2009). When used in multiple linear regression, it helps to prevent overfitting by shrinking the coefficients of less important variables towards zero. The regularized form of multiple linear regression using LASSO can also be written as Equation (6), but in addition, it is subject to:

$$\sum|\beta_i| \leq \lambda \quad (7)$$

λ = tuning parameter that controls the amount of shrinkage applied to the coefficients. LASSO linear regression usually leads to a simpler model with better generalization ability on unseen data (not used during the model creation) than basic MLR.

Support Vector Regression (SVR) effectively solves nonlinear regression tasks by utilizing the kernel trick technique. The kernel trick helps to simplify complex relationships between inputs and the target variable. By mapping the original data into a higher-dimensional space, SVR transforms the problem into a more manageable linear one. This technique implicitly maps the input data into a higher-dimensional feature space without computing its coordinates. Instead, a kernel function calculates the inner products between all pairs of data points in the feature space. The most commonly used kernel functions are linear, polynomial, radial basis function (RBF), and sigmoid kernels. Once the data is mapped, SVR finds a hyperplane that maximizes the margin between the support vectors and the hyperplane while minimizing the training error. The support vectors are a subset of training samples crucial in defining the decision boundary, as they are the closest to the hyperplane.

To avoid overfitting, SVR introduces an error limit (ϵ). Errors more minor than or equal to ϵ are considered acceptable, so the algorithm does not focus on small information noise. The aim is to find the optimal line or curve that best fits the data points while keeping the errors within this limit. This avoids overfitting and handling outliers or noisy data points (so the model is not too closely tailored to the training data and may perform well on unseen data as a consequence).

The mathematical formulation of SVR has been documented many times in various research papers, including hydrological studies, due to its strong ability to generalize and make accurate predictions, so we do not include it here (e.g., Mohammadi and Mehdizadeh, 2020; Seifi and Riahi, 2020).

Random forests (RF) are used for classification and regression problems (Breiman, 2001). An RF is an ensemble of regression trees where each tree independently makes predictions. The final prediction of the random forest is obtained by averaging the predictions from all the trees.

A regression tree is a tree-like branching structure that makes decisions based on a series of if-else conditions using the features of the data. The predicted numerical values are stored in the terminal nodes of the tree.

The RF algorithm randomly selects a subset of the available data, known as a bootstrap sample, to train each tree (i.e., suggests the mentioned if-else conditions and values in terminal nodes). This sampling process ensures that each tree sees a slightly different set of data, introducing desirable diversity into the ensemble. The present study utilized the "ranger", a fast and efficient implementation of random forests (Wright and Ziegler, 2017).

CatBoost (Prokhorenkova et al., 2018) is a gradient-boosting algorithm that differs from random forests in its approach to creating an ensemble. While random forests build independent decision trees in parallel, gradient boosting constructs sequential trees that focus on correcting errors from previous models. This involves training an ensemble of weak prediction models sequentially, with subsequent models designed to correct the errors made by the earlier models. By assigning weights to the training instances based on their errors, the subsequent models are trained to minimize those weighted errors. The final prediction is obtained by combining the predictions of all the models in the ensemble. This iterative process results in a model that can handle complex relationships, achieve high accuracy, and become popular in hydrology (Huang et al., 2019). What sets CatBoost apart are its unique features, such as automatic learning of optimal strategies for handling missing values, employing "random permutations" to address overfitting, and effectively handling both numerical and categorical features and from the user's point of view, excellent setting of default values of its parameters.

Stacking regression - H2O autoML is an automated machine learning algorithm included in the H2O framework (Prokhorenkova et al., 2018). H2O AutoML is designed to be simple and produces high-quality models suitable for deployment in an enterprise environment. It automates the training and tuning of machine learning models and automatically searches for the best model(s) and their hyperparameters related to a given dataset. It also supports some level of automatic feature engineering.

H2O AutoML uses stacking, one of the most popular ensemble machine learning techniques, to improve model performance. Multiple models are trained on the same dataset, and the predictions from these models are then used as input to a meta-model that learns how to best combine the predictions from these base models.

H2O AutoML supports training regression, binary classification, and multi-class classification models on a single machine or a cluster of machines. It trains a Random Forest, an Extremely Randomized Forest, Gradient Boosting Machines (GBMs), Deep Neural Nets, and a grid of GLMs (Madni et al., 2023). Finally, a stacked ensemble model is trained.

Feature engineering

When using machine learning models, the input data can either be raw variables like temperature or wind speed or transformed data known in the ML community as "features". Features are derived modifications of raw data that provide a more informative description of the problem, often leading to improved ML model performance. The proposed machine learning pipeline for calculating evapotranspiration emphasizes feature engineering (FE), a crucial step often overlooked.

This work proposes several strategies for FE. In the first, temperature, the essential raw variable for computing *ETo*, is transformed into approximately calculated evapotranspiration.

The Hargraeves-Samani (Eq. (3)) and McGuinness-Bordne equations (Eq. (4)) were used for this purpose.

Next FE procedure utilizes temperature to calculate the saturation vapour pressure using the Magnus-Tetens formula, which also does not require other variables (Eq. (8)):

$$es = 6.1094 \cdot \exp(17.625 \cdot T / (T + 243.04)) \quad (8)$$

es = saturation vapour pressure (kPa), T = air temperature (in °C).

The following equations consider clear-sky solar radiation (R_{so}) as another FE-derived variable, which is also used in calculating the original FAO Penman-Monteith method:

$$Ra = (24 \cdot 60 / \pi) \cdot Gsc \cdot dr \cdot (\cos(\varphi) \cdot \cos(\delta) \cdot \sin(ws) + + ws \cdot \sin(\varphi) \cdot \sin(\delta)) \quad (9)$$

$$R_{so} = (0.75 + 2 \cdot 10^{-5} \cdot z) \cdot Ra \quad (10)$$

R_{so} = clear sky solar radiation (in MJ/m²/day), z = altitude above sea level (in meters), R_a = extraterrestrial radiation on a horizontal surface (in MJ/m²/day), G_{sc} = solar constant (0.0820 MJ/m²/min), d_r = relative inverse distance Earth-Sun, φ is the latitude (in radians), δ = solar declination (in radians), w_s = sunset hour angle (in radians).

All previously mentioned additional variables generated in FE are regarding climatic data based only on temperature measurements.

Additionally, precipitation was used as a supporting variable. It does not directly affect the reference evapotranspiration. Still, it may statistically reflect cloud cover and air humidity, which are not considered in the proposed models because we try to calculate *ETo* with as few measured variables as possible. Although precipitation is another variable, that is, another claim for data availability, which we want to reduce, it is measured quite often and, therefore, usually available, so we used it when experimenting with additional variables in the context of FE.

Similarly, relative humidity and wind speed were indirectly included in the input data. These are not so often available, so we do not implement raw data from ongoing historical measurements of these variables in our models, only statistically processed data representing their usual development during the year. Only the average and smoothed values of these data for a specific day of the year are used, and we consider the same values every year. This can be obtained if at least several years of this data are available (measured). Raw data for generating these features are usually easily obtained everywhere. If unavailable, this raw data may not be from the same period in which evapotranspiration modelling is performed.

Relative humidity and wind speed data were extracted from the CarpatClim database (http://www.carpatclim-eu.org/pages/home/) for all available years in this database 1961–2010. The average values for each Julian day were determined and smoothed by the loess function over the year. This way, 366 values were obtained for both variables and assigned to the corresponding Julian days. In the final data for machine learning computations every year, the same 366 values were used as inputs. The resulting graph in Fig. 2 displays average relative humidity values for 366 Julian days, and as can be seen, these data provide both time series of average values of these variables and capture their seasonal trend as well, which is helpful information for modelling.

A categorical variable for the month was also included in the calculation to account for seasonal variations in estimating evapotranspiration. As a result, ML algorithms create individual models for each of the twelve months, which is beneficial due to the abundance of data available to construct such more specialized models. These categorical variables are not numerical values but represent 12 distinct seasonal categories.

Interactions between variables were incorporated into the input data to enhance the accuracy of the computations. Interactions capture the combined effects of two or more variables, such as humidity and temperature. These interactions were acquired by introducing new variables that are the products of the interacting variables. By including the products of interacting variables as additional features, the synergistic relationship between them could be caught, allowing for a more comprehensive and accurate representation of the evapotranspiration process.

Fine-tuning and evaluation of machine learning models

To enhance the efficiency of machine learning algorithms, it is crucial to fine-tune their various parameters. It means configuring specific parameters that dictate the behaviour of the algorithms used in the models. Next, the procedure used in this work to tune the models is outlined.

In the tuning process, only one parameter was searched for in every model, leaving the other parameters at their default values. A range of possible values for this parameter was predefined, and then the k-fold cross-validation technique was utilized to choose the best value for this adjustable parameter.

K-fold cross-validation involves randomly dividing the training data into k folds (groups). Based on this division, the model is trained k times, each running using a different fold for testing and the remaining k-1 folds for training (calibration). This approach determines the model's performance metrics for each parameter combination on independent data. Using an evaluation criterion, such as RMSE, the model's accuracy was accessed in the testing folds. The optimal parameter was determined by identifying the result with the best performance.

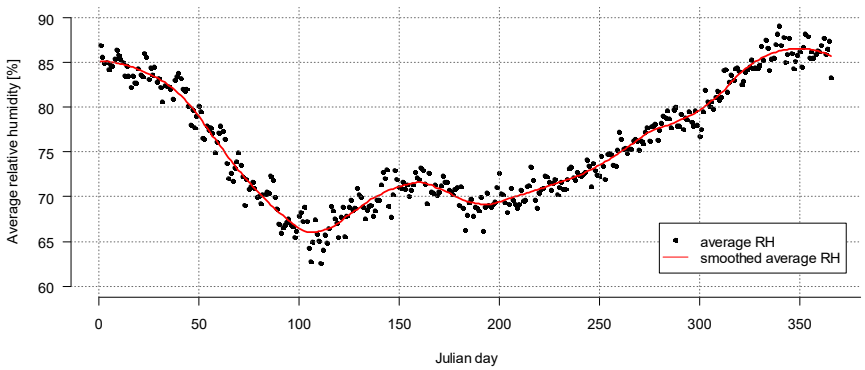


Fig. 2. Average relative humidity for Julian days based on CarpatClim grid point data near Košice.

The process of tuning and evaluating the models was performed in the R language environment, using its packages caret, parallel, registerDoParallel, hydroGOF, glmnet, randomForest, CatBoost, kernlab and H2O (Dorogush et al., 2018; Friedman et al., 2009; H2O.ai, 2022; Karatzoglou et al., 2004; Kuhn et al., 2020; Liaw and Wiener, 2021; R Developement Core Team, 2021; Zambrano-Bigiarini, 2022).

Evaluation of the models was based on their performance with testing data that was not used during the parameter calibration. Three statistical indicators are used for model evaluation, which helps gain insights into different aspects of model performance:

Root Mean Square Error (RMSE): It provides the standard deviation of the model's prediction error, normalized by standard deviation. RMSE quantifies the average magnitude of the residuals between the predicted and actual values. Equation for calculating RMSE is as follows (Zambrano-Bigiarini, 2022):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2} \quad (11)$$

S_i = simulated value, O_i = observed value, N = number of data points.

Percentual Bias (PBIAS): This indicator measures the average deviation, expressed as a percentage, between the model's predictions and the actual values. PBIAS helps assess the systematic bias of the model. PBIAS is calculated by the following equation (Zambrano-Bigiarini, 2022):

$$PBIAS = 100\frac{\sum_{i=1}^{N}(S_i - O_i)}{\sum_{i=1}^{N} O_i} \quad (12)$$

S_i = simulated value, O_i = observed value, N = number of data points.

Coefficient of Determination (R2): R2 represents the proportion of variance in the dependent variable that the independent variables can predict. Ranging from 0 to 1, an R2 value of 1 indicates a perfect fit of the regression predictions to the data. Equation for calculating R2 is as follows (Zambrano-Bigiarini, 2022):

$$R2 = 1 - \frac{SSR}{SST} \quad (13)$$

SSR = sum of squared residuals, SST = total sum of squares.

The mathematical formulation of these indicators is available in the standard statistical literature (e.g., Montgomery and Runger, 2018).

RESULTS AND DISCUSSION

The authors compared the reference evapotranspiration values calculated by the FAO Penman-Monteith method (considered a benchmark method) with the evapotranspiration calculated by the empirical equations of Hargreaves-Samani, Blaney-Criddle, McGuinness-Bordne and Makkink.

Various machine learning techniques, such as multiple linear regression (MLR), regularized linear regression using LASSO, CatBoost, random forest (RF), support vector machine (SVM), and autoML H2O models. These were employed with varying combinations of input climate variables.

ETo was calculated by the same machine learning models and using feature engineering.

In addition, the degree of generalizability of machine learning models is assessed by applying models trained on the Košice station data for calculating evapotranspiration on the Hurbanovo and Poprad stations data.

Calculation of *ETo* by FAO56 and empirical equations

In this work, the FAO Penman-Monteith equation is considered the benchmark method for *ETo* calculation, as it is the most widely used *ETo* calculation method. It was used to compute reference evapotranspiration for the stations Košice, Hurbanovo and Poprad in a daily step in the period 3.1.1951-30.4.2020. Calculations were implemented in the R programming language using the FAO56 package (Ahani and Mousavi Nadoushani, 2021).

The FAO Penman-Monteith equation requires several input variables that may not always be available for a given location. The authors used alternative equations such as the Hargreaves-Samani, Blaney-Criddle, Makkink and McGuinness-Bordne formulas to address this issue, which requires fewer input variables. These equations were computed using the R Evapotranspiration package (Guo et al., 2016). The results were compared with the FAO Penman-Monteith method and are presented in Fig. 3 and Table 2.

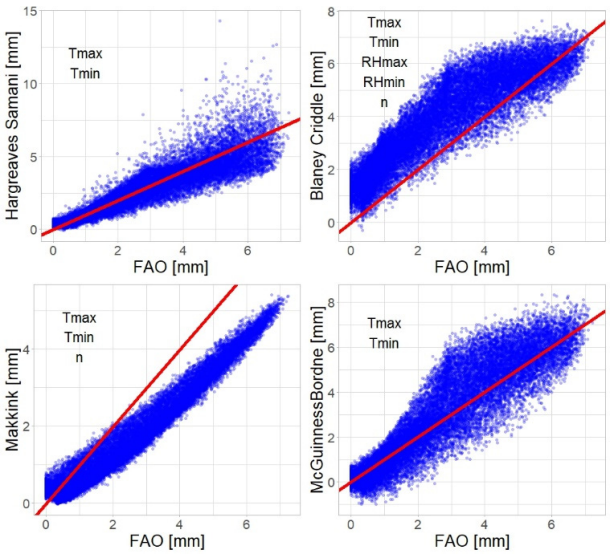


Fig. 3. Comparison of empirical equations with FAO Penman-Monteith method (*Tmin* – minimal daily temperature, *Tmax* – maximal daily temperature, *n* – sunshine hours, *RHmax* – maximum relative humidity, *RHmin* – minimum relative humidity).

Table 2. Statistical characteristics of empirical equations' results (*Tmin* – min. temperature, *Tmax* – max. temperature, *n* – sunshine hours).

Equation	Variables	negat_count	RMSE	PBIAS %	R2
Hargreaves-Samani	*Tmax*, *Tmin*	4	0.73	3.4	0.86
Blaney-Criddle	*Tmax*, *Tmin*	4	1.3	–29.8	0.84
Makkink	*Tmax*, *Tmin*, *n*	12	1.04	51	0.96
McGuinness-Bordne	*Tmax*, *Tmin*	1311	0.86	–6	0.86

The scatter plots, with the input variables displayed, are shown in Fig. 3. Table 2 presents the statistical characteristics of RMSE, PBIAS, and R2 and the number of the calculated negative values. It can be observed that the best results were achieved by the Hargreaves-Samani equation, with a value of RMSE 0.73, PBIAS 3.4 and R2 0.86. Together with the Blaney-Criddle equation, it had the smallest number of negative calculated values, which can be considered incorrect. Regarding the R2 value, the Makkink equation was the best but achieved the highest PBIAS value (51%). It can be concluded that the Hargreaves-Samani equation achieved the best result. This can be attributed to the inclusion of the calculated value of extraterrestrial radiation (*Ra*) based on latitude in this equation.

Calculation of *ETo* by machine learning using raw data

ML models described in the section Methodology were used in this section to calculate the *ETo* of the Košice climatic station. The inputs to these models were various combinations of climate data. A mutual correlation plot between *ETo* and climatic variables is in Fig. 4. The correlation coefficients between the variables are shown in the cells above the diagonal. Variable abbreviations are on the matrix's diagonal with the variable's histograms. The correlation coefficient is for two variables on the diagonal vertically and horizontally from the cell with its value. Also, mini scatter plots between variables are shown below the diagonal.

The input data was divided into training and testing sets in the ratio of 70:30. Only training data was used to train (calibrate) the models, and the trained models were verified using test data. Five combinations of inputs are compared in Table 3 to evaluate how different climatic variables affect the accuracy of the *ETo* computation.

The main (always used) variables were *Tmin* and *Tmax* (model no. 1), usually measured everywhere. They also correlate strongly with *ETo* (Fig. 4). In the remaining models of this table, one of the following variables was added: sunshine hours (*n_sun*), relative humidity (*RH*), wind velocity (*wind*) and cloud cover (*clouds*), to improve the model and find out the effect of this variable on *ETo* computation accuracy. The most significant improvement occurred in model No. 2, where variable sunshine hours was added, which reduced the RMSE value in the best H2O autoML model by 23.8 %, the absolute value of PBIAS changed from 0% to 1%, and R2 improved from 0.77 to 0.87. Sunshine hours correlate most strongly with *ETo* from the variables added to the models in this table (Fig. 4). The most negligible improvement occurred with adding the measured climatic variable cloud cover (model no. 5).

In terms of various ML models performance, it's obvious that nonlinear models (SVM, RF, CatBoost, autoML) consistently outperform linear (MLR, LASSO). For instance, when using minimum and maximum temperatures and sunshine hours, the CatBoost model outperforms linear regression across all statistical indicators. Although linear regression is easier to interpret, regarding precision, it is designed for linear problems, which the calculation of evapotranspiration is not. In this article, the authors do not concentrate on investigating the physical aspect of the evapotranspiration mechanism, for which a better interpretable model can be preferred. Instead, they focused on offering a more straightforward (computed with fewer data) and more accurate alternative calculation. Therefore, using nonlinear models for this purpose is recommended.

After analyzing the information in Tables 2 and 3, it becomes evident that standard empirical equations exhibit more significant systematic errors than machine learning models, as indicated by the larger values of the PBIAS statistic (which quantifies underestimation or overestimation). However, empirical models tend to score better regarding the coefficient of determination R2 and RMSE. This has led the authors to exploit feature engineering to enhance ML models' performance, what will be discussed in the following section.

Calculation of *ETo* by machine learning with feature engineering

Using features instead of raw data is a crucial step in improving the results of ML models, but it is often overlooked. This work evaluated the influence of various uses of features in ML models.

Table 4 displays models that utilize only the minimum and maximum temperature and one feature derived solely from temperature. These additional data – features - include temperature transformed to approximate evapotranspiration by

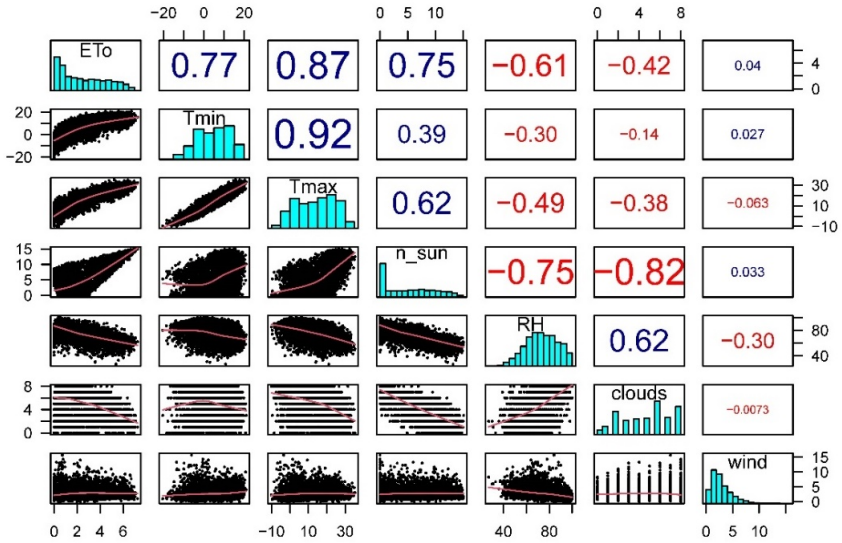
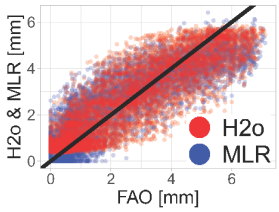
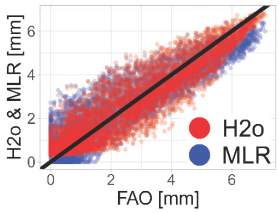
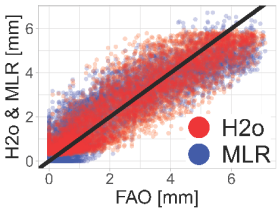
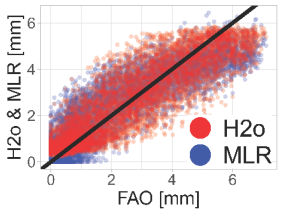
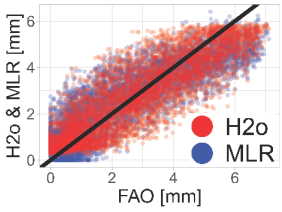


Fig. 4. Correlation matrix of input climatic variables and FAO *ETo* (*Tmin* – minimal daily temperature, *Tmax* – maximal daily temperature, *n_sun* – sunshine hours, *RH* – relative humidity, *wind* – wind velocity, *clouds* – cloud cover).

Table 3. Evaluation of ML models containing a limited amount of raw data (*Tmin* – minimal daily temperature, *Tmax* – maximal daily temperature, *n_sun* – sunshine hours, *RH* – relative humidity, *wind* – wind velocity, *clouds* – cloud cover).

Model No.	Inputs	model	RMSE	PBIAS %	R2	Modelled~Calculated
1	*Tmin*, *Tmax*	MLR	0.963	0.5	0.748	
		LASSO	0.935	–0.4	0.762	
		catboost	0.91	–0.6	0.774	
		ranger	0.915	–0.5	0.772	
		svm	0.915	–2.1	0.774	
		H2O	0.91	0	0.774	
2	*Tmin, Tmax, n_sun*	MLR	0.789	0.8	0.831	
		LASSO	0.751	–0.2	0.846	
		catboost	0.692	–0.4	0.87	
		ranger	0.697	–0.5	0.868	
		svm	0.7	1	0.869	
		H2O	0.693	1	0.869	
3	*Tmin, Tmax, RH*	MLR	0.877	0.9	0.792	
		LASSO	0.84	–0.3	0.808	
		catboost	0.796	–0.3	0.827	
		ranger	0.799	–0.4	0.826	
		svm	0.797	–0.7	0.828	
		H2O	0.796	–0.7	0.827	
4	*Tmin*, *Tmax, wind*	MLR	0.933	0.5	0.764	
		LASSO	0.905	–0.5	0.777	
		catboost	0.863	–0.7	0.797	
		ranger	0.868	–0.7	0.795	
		svm	0.865	–1.5	0.797	
		H2O	0.864	–0.3	0.797	
5	*Tmin, Tmax, clouds*	MLR	0.952	0.5	0.755	
		LASSO	0.903	–0.5	0.778	
		catboost	0.872	–0.6	0.793	
		ranger	0.877	–0.7	0.79	
		svm	0.878	–1	0.792	
		H2O	0.873	–1.2	0.793	

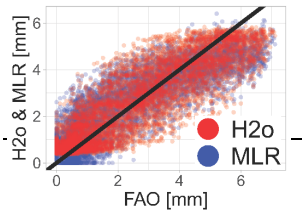
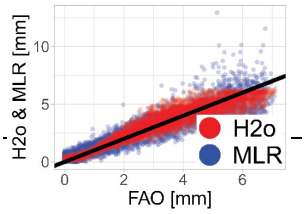
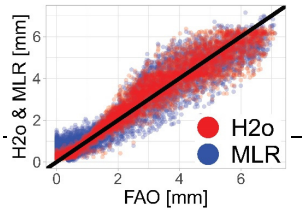
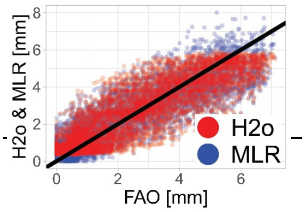
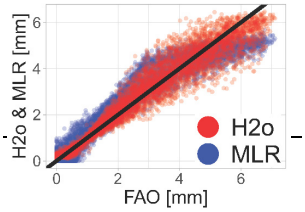
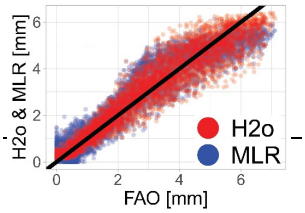
the Hargreaves-Samani (*HS_ETo*) and McGuiness-Bordne (*MGB_PE*) equations, saturation vapour pressure (*es*) and clear-sky solar radiation (*Rso*). In addition, the authors also included among the derived features an easily obtained categorical variable, which identifies months. Deriving features from raw variables is explained in the Methods section. So, in the models presented in this table, the temperature is the only measured climate variable employed, much like in most of the empirical equations in Table 2. Therefore, the models in this table are potential competitors of the standard empirical models evaluated in Table 2 regarding simplicity of inputs and *ETo* calculation accuracy.

It can be observed in Table 4, which presents the main results of this work, that adding features improves the model's results in comparison with Tables 2 and 3, i.e., models utilizing FE are notably better in the computation of *ETo*, than standard equations or ML methods without using FE. The most significant improvement occurred with adding the clear sky solar radiation *Rso* (model No. 5), which is not a measured variable. Adding this variable decreased the RMSE value using, e.g., autoML H2O model in comparison with the best empirical equation (Hargraeves and Samani, 1985) by 36%, the percentual PBIAS value decreased from 3.4 to 0.3, and the coefficient of determination R2 increased from 0.86 to 0.94. Consistency with the results of empirical methods can be seen here, where the best results were provided by the Hargraeves-Samani method, which also uses a similar auxiliary quantity, namely extraterrestrial radiation Ra, but the use of Rs in the context of machine learning methods provides an even more accurate approximation of the reference evapotranspiration. In the methodology section, it is mentioned

that the *Rso* variable is calculated based on the date, latitude, and altitude. This means that by incorporating these inputs, the model's accuracy can be enhanced, and it can be utilized in various altitudes and geographic regions.

Similarly, all investigated ML models which use *Rso* have better results than the best empirical equation or ML models which use only raw data. Furthermore, this statement also holds for other examined features in Table 4 when combined with *Tmin* and *Tmax*. The addition of a basic categorical feature, "*month*," (in model No.6) greatly improved the results. This is because the algorithms were able to internally create a tree of models that focused on calculating *ETo* for each individual month. However, it should be noted that the ranger ML model does not possess this capability. Good results were also achieved by adding the approximate evapotranspiration calculated by empirical equations *HS_ETo* (model no. 2) and *MGB_PE* (model no. 3). The slightest improvement came from adding saturation vapour pressure es.

Table 4. Evaluation of models with features derived from temperature, date and latitude (*Tmax* – maximal daily temperature, *Tmin* – minimal daily temperature, *HS_ETo* – evapotranspiration by the Hargreaves-Samani equation, *MGB_PE* – evapotranspiration by McGuiness-Bordne equation, *es* – saturation vapour pressure, *Rso* – clear-sky solar radiation).

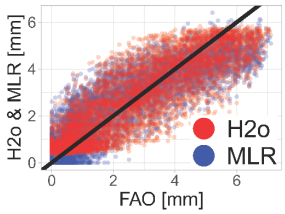
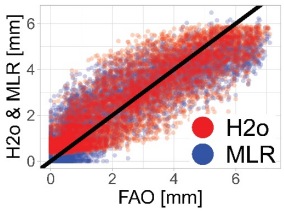
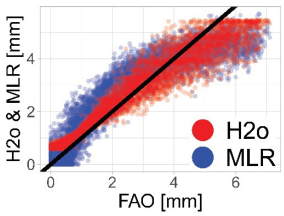
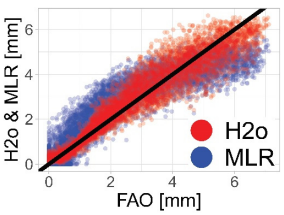
Model no.	Inputs	model	RMSE	PBIAS %	R2	Modelled~Calculated
1	*Tmax*, *Tmin*	MLR	0.963	0.5	0.748	
		LASSO	0.935	–0.4	0.762	
		CatBoost	0.91	–0.6	0.774	
		ranger	0.915	–0.5	0.772	
		svm	0.915	–2.1	0.774	
		H2O	0.91	0	0.774	
2	*Tmax*, *Tmin*, *HS_ETo*	MLR	0.704	–0.5	0.867	
		LASSO	0.636	–0.5	0.892	
		CatBoost	0.485	–0.1	0.937	
		ranger	0.492	–0.1	0.935	
		svm	0.471	–0.6	0.941	
		H2O	0.478	–0.1	0.939	
3	*Tmax*, *Tmin*, *MGB_PE*	MLR	0.604	–0.1	0.902	
		LASSO	0.526	–0.2	0.926	
		CatBoost	0.484	–0.1	0.937	
		ranger	0.484	–0.2	0.937	
		svm	0.478	–0.6	0.939	
		H2O	0.481	1.2	0.938	
4	*Tmax*, *Tmin*, *es*	MLR	0.942	–0.4	0.762	
		LASSO	0.922	–0.7	0.772	
		CatBoost	0.915	–0.6	0.776	
		ranger	0.919	–0.6	0.774	
		svm	0.92	–2.1	0.775	
		H2O	0.915	–1.2	0.776	
5	*Tmax*, *Tmin*, *Rso*	MLR	0.612	0.7	0.9	
		LASSO	0.473	–0.2	0.94	
		CatBoost	0.469	–0.2	0.94	
		ranger	0.475	–0.1	0.94	
		svm	0.471	–0.6	0.94	
		H2O	0.467	0.3	0.94	
6	*Tmax*, *Tmin*, *month*	MLR	0.606	0.3	0.902	
		LASSO	0.522	–0.1	0.927	
		CatBoost	0.522	–0.1	0.927	
		ranger	1.557	–0.4	0.792	
		svm	0.518	–1.3	0.928	
		H2O	0.52	–0.7	0.928	

The next group of investigated features (Table 5) are variables frequently available next to temperatures. Precipitation (*prec*), humidity averages for Julian days (*a_RH*), and average wind speeds for Julian days (*a_wind*) were included in this group. The selection of precipitation as a feature is due to its frequent measurement. Although it does not directly affect the reference evapotranspiration calculation, there is potential in the ML approach to use precipitation as a statistical reflection of air humidity and cloud cover. The other two suggested variables, namely average humidity and wind data, are not intended to be complete historical records for an investigated period but only averages for each day of the year. These values, totalling 366, can be derived from various databases or incomplete measurements in different regions worldwide. The applied procedure to obtain these variables is described in Methods.

Based on Table 5, it can be observed that adding each variable except precipitation (model No. 2) again significantly improved the performance of models. Also is an evident clear advantage of ML models over standard linear regression. Adding the humidity averages of Julian's days contributed to the most significant change. Compared CatBoost with these variables with MLR computed only with *Tmin* and *Tmax*, the RMSE value decreased by 46.1%, the absolute value of PBIAS decreased from 0.5% to –0.1%, and the R2 value increased from 0.748 to 0.928. When comparing CatBoost with the best empirical equation based on temperatures only (Hargraeves-Samani), the RMSE value decreased by 28.9%, the absolute value of PBIAS decreased from 3.4% to –0.1%, and the R2 value increased from 0.860 to 0.928. Please note that the relative humidity and wind speed data are not typical daily measurements. Instead, they are smoothed average values for Julian days, calculated using, e.g., shorter periods of measurements from the given or nearby area where we calculate the reference evapotranspiration. In this way, we can also include climatic variables unavailable for the entire estimated period in the calculation. This study used data from the grid dataset CarpatClim for a shorter period (grid point near Košice).

In the above tables, only one feature was added to *Tmin* and *Tmax*. In Table 6, combinations of features are evaluated to assess if increasing the number of variables helps increase accuracy by their synergistic effect. Model No. 1 (Table 6) used all the temperature, date and longitude derived features and *Tmin* and *Tmax*. The results differ only slightly from model No. 5 in Table 4, in which only the variable *Rso* was used. A more significant improvement occurred when using all features in model No. 2.

Table 5. Evaluation of models with features derived from variables other than temperature (*Tmax* – maximum daily temperature, *Tmin* – minimal daily temperature, *prec* – precipitation, *a_RH* – humidity averages for Julian days, *a_wind* – average wind speeds for Julian days).

Model no.	Inputs	model	RMSE	PBIAS %	R2	Modelled~Calculated
1	*Tmax*, *Tmin*	MLR	0.963	0.5	0.748	
		LASSO	0.935	–0.4	0.762	
		CatBoost	0.91	–0.6	0.774	
		ranger	0.915	–0.5	0.772	
		svm	0.915	–2.1	0.774	
		H2O	0.91	0	0.774	
2	*Tmax*, *Tmin*, *prec*	MLR	0.964	0.5	0.752	
		LASSO	0.933	–0.4	0.767	
		CatBoost	0.902	–0.5	0.782	
		ranger	0.906	–0.6	0.78	
		svm	0.909	–1.9	0.78	
		H2O	0.903	0.8	0.782	
3	*Tmax*, *Tmin*, *a_RH*	MLR	0.812	1.2	0.826	
		LASSO	0.667	–0.4	0.881	
		CatBoost	0.519	–0.1	0.928	
		ranger	0.531	–0.2	0.924	
		svm	0.547	–1.3	0.92	
		H2O	0.609	–0.1	0.934	
4	*Tmax*, *Tmin*, *a_wind*	MLR	0.839	1.5	0.815	
		LASSO	0.741	–0.2	0.853	
		CatBoost	0.532	–0.3	0.924	
		ranger	0.541	–0.2	0.922	
		svm	0.583	–1.3	0.909	
		H2O	0.513	–0.2	0.929	

We can conclude that, generally, better results were achieved with features than with raw climate data. Based on the results of various ML models in the presented tables, it is also possible to suggest the autoML H2O and CatBoost models, which usually achieve the best results. Since the construction of features is simple, the authors suggest their usage as a significant improvement to the more precise computation of the *ETo* with fewer climatic variables than the original FAO method.

Generalizability of the proposed models

When creating machine learning models, they must be trained on data similar to what they will be applied to (as shown in Table 1). Nonetheless, having a model capable of calculating evapotranspiration across various regions without the need for creating new models or obtaining all the necessary training data for each area is advantageous. To clarify this, let's consider an example of rainfall-runoff modelling. A rainfall-runoff ML model can be developed using only rainfall and flows as inputs without requiring other descriptive data like a digital terrain model, land use, soil properties, or hydraulic characteristics of the river bed. Incorporating these factors during ML model training is based only on the fact that they are somewhat mirrored in precipitation and flows. This approach allows the ML model to calculate the unknown, unmeasured flow rates solely with precipitation. Although this approach is simpler and sometimes even more accurate, the downside is that the resulting model is specific to the watershed on which it was trained. Therefore, in this study, the authors have also examined the generalizability of the models created for evapotranspiration, i.e., the extent of applicability of the model trained on one climatic station's data. The authors plan to address a more detailed verification of this in the future with a separate work.

The authors suppose that their *ETo* ML model, trained on Košice data, will not be applicable globally, but they tested its applicability tentatively within Slovakia. This is accomplished in Table 7, evaluating one of the simplest but precise model evaluated earlier in this work, with inputs *Tmin*, *Tmax*, and clear sky solar radiation (*Rso*) as inputs. *Rso* is not a measured variable, it is estimated based on date, latitude and altitude (Eqs. 9, 10), so the model can indirectly consider (different) geographic locations and altitudes. The authors suppose that the similarity of the variables affecting evapotranspiration, but not used in the model, is essential for its applicability in different areas. Fig. 5 illustrates this similarity in testing areas and training locality - Košice, and Table 7 evaluates the model's performance in Hurbanovo and Poprad climatic stations, having quite different climatic conditions within Slovakia.

In Table 7, the models trained on the data of the Košice station were used to calculate the reference evapotranspiration in the quite different Hurbanovo and Poprad stations. The computed data were compared with the evapotranspiration values calculated by the FAO Penman-Monteith method. All stations have similar climatic characteristics (Fig. 5), but given the Slovak conditions, they are sufficiently different that one can generalize from the modelled results obtained in them. Some difference exists because the Poprad station is in mountainous terrain, while the Košice station is in the lowlands. The Hurbanovo station has the highest average temperatures among the available stations in Slovakia. Another reason for utilizing these stations is the contrasting altitudes between Košice (230 m. a.s.l.) and Poprad (694 m. a.s.l.). This variation in altitude also contributes to some differences in climate observed between the stations, which is advantageous for the testing process. In Table 7, the original models, trained on the data of the Košice station, are also shown. The generalizability of the models seems to be at a very high level, as the resulting values of the goodness of fit tests are near the results of the models for Košice. However, more detailed testing is still necessary. The H2O model achieved the best statistical characteristics in all cases. Therefore, it can be concluded that its degree of generalizability is the highest among the selected ML models.

This research investigated the potential of utilizing machine learning (ML) techniques to compute Reference Evapotranspiration (*ETo*) using the FAO methodology. The FAO Penman-Monteith equation is broadly accepted as a reliable, physically-based approach for *ETo* calculation. However, integrating ML into *ETo* calculation poses specific challenges and considerations that require discussion.

Table 6. Evaluation of models with more features included in their inputs (*Tmax* – maximal daily temperature, *Tmin* – minimal daily temperature, *HS_ETo* – evapotranspiration by the Hargreaves-Samani equation, *MGB_PE* – evapotranspiration by McGuiness-Bordne equation, *es* – saturation vapour pressure, *Rso* – clear-sky solar radiation, *prec* – precipitation, *RH* – relative humidity, *wind* – wind velocity).

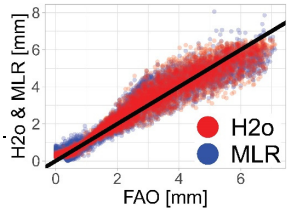
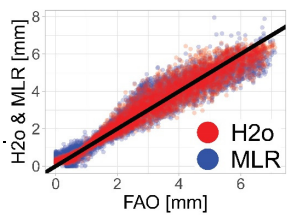
Model no.	Inputs	model	RMSE	PBIAS %	R2	Modelled~Calculated
1	*Tmax, Tmin, HS_ETo, MGB_PE, es, Rso, month*	MLR	0.536	–0.1	0.923	
		LASSO	0.467	–0.1	0.942	
		CatBoost	0.471	–0.1	0.941	
		ranger	1.055	–0.3	0.901	
		svm	0.468	–0.8	0.941	
		H2O	0.468	1.3	0.942	
2	*Tmax, Tmin, HS_ETo, MGB_PE, es, Rso, month, prec, RH, wind*	MLR	0.514	0	0.929	
		LASSO	0.444	–0.1	0.947	
		CatBoost	0.412	–0.3	0.955	
		ranger	0.876	–0.2	0.925	
		svm	0.426	–0.3	0.951	
		H2O	0.407	–0.1	0.956	

Table 7. Testing model trained with Košice data on two different climatic stations (*Rso* – clear-sky solar radiation, *Tmin* – minimal daily temperature, *Tmax* – maximum daily temperature).

No. of inputs	Inputs	Data	model	RMSE	PBIAS %	R2
3	*ETo, Rso, Tmin, Tmax*	Train: Kosice Test: Košice	MLR	0.616	0.7	0.897
			LASSO	0.477	–0.2	0.938
			catboost	0.472	–0.2	0.939
			ranger	0.478	–0.1	0.938
			svm	0.474	–0.7	0.939
			H2O	0.47	–0.3	0.94
3	*Rso, Tmin, Tmax*	Train: Košice Test: Hurba-novo	MLR	0.636	0.2	0.904
			LASSO	0.456	0.4	0.949
			catboost	0.454	0.2	0.949
			ranger	0.462	0.3	0.948
			svm	0.443	0.1	0.952
			H2O	0.452	0.3	0.95
3	*Rso, Tmin, Tmax*	Train: Košice Test: Poprad	MLR	0.58	4.2	0.904
			LASSO	0.45	0.8	0.941
			catboost	0.455	0.3	0.939
			ranger	0.463	–0.6	0.937
			svm	0.459	–0.5	0.938
			H2O	0.452	0.1	0.94

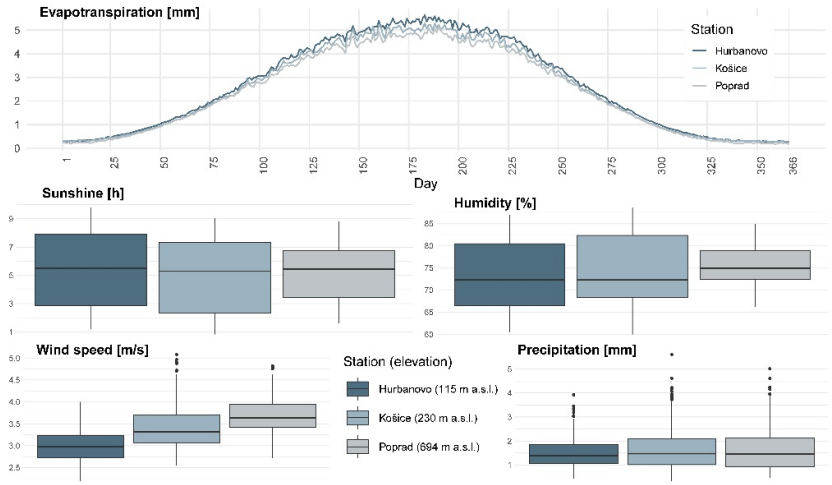


Fig. 5. Climatic characteristics of stations Košice, Hurbanovo and Poprad.

One of the main concerns when employing ML for *ETo* estimation is the quality and availability of input data. Unfortunately, meteorological data can be limited, inconsistent, or contain errors in many regions. To address this, efforts should be made to enhance data quality through regular maintenance of meteorological stations, data validation, and the development of robust data imputation techniques. Improving data quality can reduce the impact of issues with available data on ML model performance. In this study, the data used have already been verified by their provider to the ECA&D database, which is The Slovak Hydrometeorological Institute, and by the database creators themselves, so the data can be considered reliable from this perspective. However, the authors are aware of the existence of this problem and want to address it in further works.

ML models, particularly more complex ones, can capture intricate relationships within data. However, the complexity can pose a challenge when interpreting and validating the models and their predictions. The FAO methodology is known for its transparency and adherence to physical principles. Therefore, retaining and incorporating domain expertise into model development is crucial. Collaboration between domain experts and data scientists can facilitate the incorporation of physical principles and ensure that ML-based *ETo* estimates remain grounded in established scientific knowledge. The authors of this paper are from the domain of water management, and in our opinion, this situation will be helped by bringing learning into the arsenal of experts from other fields. This is also the contemporary trend in ML domain and was demonstrated in this study by using the H2O autoML model, which automates many details of machine learning model development.

CONCLUSIONS

In this study, the authors explored the application of machine learning (ML) techniques and feature engineering (FE) in the calculation of reference evapotranspiration (*ETo*) according to the FAO Penman-Monteith method by using minimal measured

climate variables. In essence, the article attempts to use only temperature from continuously measured climate variables to reduce the requirements for input data for calculating evapotranspiration the rest of the inputs are easily obtainable constructed variables - features. To make this work beneficial, surpassing the accuracy of traditional empirical equations such as Hargreaves-Samani or Blaney-Criddle was required, which was confirmed. Various ML algorithms were used, including LASSO regularized linear regression, SVM, random forest, CatBoost, and stacked model autoML from the H2O framework. CatBoost and autoML were proven to be the most promising models, providing superior results, especially when feature engineering was incorporated into the ML modelling. To the best of the authors' knowledge, feature engineering was not systematically evaluated in previous works on this theme, and this work confirms that it significantly improves the predictive performance of ML models.

Although the models developed in this study based on data from one climatic station from a given region demonstrated promising results in surrounding geographical locations, further research is necessary to verify their generalizability and robustness across a broader range of environmental conditions. However, the authors hypothesize that the models will offer good results in a relatively wide region (e.g., Slovakia, from where the case study is presented), and training the model based on data from several stations is still possible to achieve a generally applicable model for a wider area.

In conclusion, this study establishes the value of using feature engineering in machine learning for accurate reference evapotranspiration estimation. The potential for improved predictions, reduced input requirements, and increased efficiency holds compelling promise for optimizing water management strategies, crop planning, and decision-making in the agricultural sector.

Acknowledgements. This work was supported by the Slovak Research and Development Agency under Contract No. APVV-19-0383 "Natural and technical measures oriented to water retention in sub-mountain watersheds of Slovakia" and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/0067/23.and Grant No. 2/0015/23 “Comprehensive analysis of the quantity and quality of water regime development in streams and their mutual dependence in selected Slovak basins”

REFERENCES

Ahani, A., Mousavi Nadoushani, S.S., 2021. FAO56: Evapotranspiration Based on FAO Penman-Monteith Equation: R package version 0.1.0 [WWW Document]. URL https://CRAN.R-project.org/package=FAO56 (accessed 6.7.2023).

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration – Guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper 56. FAO, Rome 300, D05109.

Blaney, H.F., Criddle, W.D., 1950. Determining water requirements in irrigated areas from climatological and irrigation data. United States Department of Agriculture, Soil Conservation Service, Washington 25, DC, 48 p.

Breiman, L., 2001. Random forests. Machine Learning, 45, 5–32.

Dimitriadou, S., Nikolakopoulos, K.G., 2022. Multiple linear regression models with limited data for the prediction of reference evapotranspiration of the Peloponnese, Greece. Hydrology, 9, 7, 124. https://doi.org/10.3390/hydrology9070124

Doorenbos, J., Pruitt, W.O., 1977. Crop water requirements. FAO Irrigation and Drainage Paper 24. Land and Water Development Division, FAO, Rome, 144 p.

Dorogush, A.V., Ershov, V., Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. ArXiv preprint arXiv. https://doi.org/https://doi.org/10.48550/arXiv.1810.11363

European Commission, Joint Research Centre. Agri4Cast dataset [WWW Document]. URL https://agri4cast.jrc.ec.europa.eu/dataportal/ (accessed 6.7.2023).

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., 2009. Glmnet: Lasso and elastic-net regularized generalized linear models: R package version 1, 24 p.

Gomes, E.P., Blanco, C.J.C., 2021. Daily rainfall estimates considering seasonality from a MODWT-ANN hybrid model. Journal of Hydrology and Hydromechanics, 69, 13–28. https://doi.org/10.2478/johh-2020-0043

Guo, D., Westra, S., Maier, H.R., 2016. An R package for modelling actual, potential and reference evapotranspiration. Environmental Modelling & Software, 78, 216–224. https://doi.org/10.1016/j.envsoft.2015.12.019

Hargreaves, G.H., Samani, Z.A., 1985. Reference crop evapotranspiration from temperature. Applied Engineering in Agriculture 1, 96–99. https://doi.org/10.13031/2013.26773

Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zeng, W., Zhou, H., 2019. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. Journal of Hydrology, 574, 1029–1041. https://doi.org/10.1016/j.jhydrol.2019.04.085

H2O.ai, 2022. H2O Documentation: H2O for R Users [WWW Document]. URL https://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html#h2o-package-for-r (accessed 6.7.2023).

Chen, X., Parajka, J., Széles, B., Strauss, P., Blöschl, G., 2020. Controls on event runoff coefficients and recession coefficients for different runoff generation mechanisms identified by three regression methods. Journal of Hydrology and Hydromechanics, 68, 155–169. https://doi.org/10.2478/johh-2020-0008

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. Kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software, 11, 9, 1–20. https://doi.org/10.18637/jss.v011.i09

Klein Tank, A.M.G., Wijngaard, J.B., Können, G.P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A.F.V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L.V., Petrovic, P., 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. International Journal of Climatology, 22, 1441–1453. https://doi.org/10.1002/joc.773

Krishnashetty, P.H., Balasangameshwara, J., Sreeman, S., Desai, S., Kantharaju, A.B., 2021. Cognitive computing models for estimation of reference evapotranspiration: A review. Cognitive Systems Research, 70, 109–116. https://doi.org/10.1016/j.cogsys.2021.07.012

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., 2020. Caret: Classification and Regression Training. R package version 6.0-86.

Liaw, A., Wiener, M., 2015. RandomForest: Breiman and Cutler’s random forests for classification and regression. R package version 4, 14.

Madni, H.A., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., Hamdi, M., Ashraf, I., 2023. Water-quality prediction based on H_2O AutoML and explainable AI techniques.

Water, 15, 3, 475. https://doi.org/10.3390/w15030475
Makkink, G.F., 1957. Testing the Penman formula by means of lysimeters. Journal of the Institution of Water Engineers, 11, 277–288.
McGuinness, J.L., Bordne, E.F., 1972. A comparison of lysimeter-derived potential evapotranspiration with computed values. US Department of Agriculture.
Mehta, R., Pandey, V., 2015. Reference evapotranspiration (ETo) and crop water requirement (ETc) of wheat and maize in Gujarat. Journal of Agrometeorology, 17, 107–113. https://doi.org/10.54386/jam.v17i1.984
Mohammadi, B., Mehdizadeh, S., 2020. Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm. Agricultural Water Management, 237, 106145. https://doi.org/10.1016/j.agwat.2020.106145
Montgomery, D.C., Runger, G.C., 2018. Applied Statistics and Probability for Engineers. 7th Ed. John Wiley.
Mostafa, R.R., Kisi, O., Adnan, R.M., Sadeghifar, T., Kuriqi, A., 2023. Modeling potential evapotranspiration by improved machine learning methods using limited climatic data. Water, 15, 3, 486. https://doi.org/10.3390/w15030486
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features: unbiased boosting with categorical features. Advances in Neural Information Processing Systems 31.
R Developement Core Team, 2009. A language and environment for statistical computing. http://www. R-project. org.
Roy, D.K., Sarkar, T.K., Kamar, S.S.A., Goswami, T., Muktadir, M.A., Al-Ghobari, H.M., Alataway, A., Dewidar, A.Z., El-Shafei, A.A., Mattar, M.A., 2022. Daily prediction and multi-step forward forecasting of reference evapotranspiration using LSTM and Bi-LSTM models. Agronomy, 12, 3, 594. https://doi.org/10.3390/agronomy12030594
Sattari, M.T., Apaydin, H., Shamshirband, S., 2020. Performance evaluation of deep learning-based Gated Recurrent Units (GRUs) and tree-based models for estimating ETo by using limited meteorological variables. Mathematics, 8, 972. https://doi.org/10.3390/math8060972
Seifi, A., Riahi, H., 2020. Estimating daily reference evapotranspiration using hybrid gamma test-least square support vector machine, gamma test-ANN, and gamma test-ANFIS models in an arid area of Iran. Journal of Water and Climate Change, 11, 217–240. https://doi.org/10.2166/wcc.2018.003
Silva-Júnior, R.O., Souza-Filho, P.W.M., Salomão, G.N., Tavares, A.L., Santos, J.F., Santos, D.C., Dias, L.C., Silva, M.S., Melo, A.M.Q., Souza-Costa, C.E.A., Rocha, E.J.P., 2021. Response of water balance components to changes in soil use and vegetation cover over three decades in the Eastern Amazon. Frontiers in Water, 3, 749507. https://doi.org/10.3389/frwa.2021.749507
Szalai, S., Nejedlik, P., Štastny, P., Mikulová, K., Szentimrey, T., Bihari, Z., Lakatos, M., 2012. Climate of the Carpathian Region, a project for a high resolution harmonized gridded database. Forum Carpaticum 2012.
Valle Júnior, L.C.G. do, Vourlitis, G.L., Curado, L.F.A., Palácios, R. da S., Nogueira, J. de S., Lobo, F. de A., Islam, A.R.M.T., Rodrigues, T.R., 2021. Evaluation of FAO-56 procedures for estimating reference evapotranspiration using missing climatic data for a Brazilian tropical savanna. Water, 13, 1763. https://doi.org/10.3390/w13131763
Wang, S., Lian, J., Peng, Y., Hu, B., Chen, H., 2019. Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China. Agricultural Water Management, 221, 220–230. https://doi.org/10.1016/j.agwat.2019.03.027
Wright, M.N., Ziegler, A., 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77, 1, 1–17. https://doi.org/10.18637/jss.v077.i01
Yeh, H.-F., 2017. Comparison of evapotranspiration methods under limited data. In: Bucur, D. (Ed.): Current Perspective to Predict Actual Evapotranspiration. Intech Open. https://doi.org/10.5772/intechopen.68495
Zambrano-Bigiarini, M., 2022. HydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series 1–77. https://doi.org/10.5281/zenodo.839854
CarpatClim, Deliverable D1.6 [WWW Document]. URL http://www.carpatclim-eu.org/docs/deliverables/D1_6.pdf (accessed 6.7.2023).

Received 1 July 2023
Accepted 4 October 2023