Italian Journal of
Agrometeorology
Rivista Italiana di Agrometeorologia

Firenze University Press
www.fupress.com/ijam

Citation: Babakos, K., Papamichail, D., Pisinaras, V., Tziachris, P., Demertzi, K. & Aschonitis, V. (2024). Using a random cross-validation technique to compare typical regression vs. Random Forests for modelling pan evaporation. *Italian Journal of Agrometeorology* (1): 59-72. doi: 10.36253/ijam-2043

Received: February 10, 2023

Accepted: December 7, 2023

Published: August 2, 2024

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

ORCID:
KB: 0000-0003-0771-7816
DP: 0000-0002-3214-1348
VP: 0000-0001-6094-7659
PT: 0000-0002-3454-4132
KD: 0000-0003-4842-1974
VA: 0000-0003-4852-5992

Using a random cross-validation technique to compare typical regression vs. Random Forests for modelling pan evaporation

Konstantinos Babakos[1,2], Dimitris Papamichail[2], Vassilios Pisinaras[1], Panagiotis Tziachris[1], Kleoniki Demertzi[3], Vassilis Aschonitis[1,*]

[1] *Soil and Water Resources Institute, Hellenic Agricultural Organization - Dimitra, Thessaloniki 57001, Greece*
[2] *Department of Hydraulics, Soil science and Agricultural Engineering, Faculty of Agriculture, Forestry and Natural Environment, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece*
[3] *Goulandris Natural History Museum, Greek Biotope/Wetland Centre, Thermi-Thessaloniki 57001, Greece*
*Corresponding author. Email: v.aschonitis@swri.gr

Abstract. Pan evaporation (E_{pan}) of class A pan evaporimeter under local semi-arid conditions was modelled in this study based on meteorological observations as input data using an integrated regression approach that includes three steps: a) first step: appropriate selection of transformations for reducing normality departures of independent variables and ridge regression for selecting variables with low collinearity based on variance inflation factors, b) second step (RCV-REG): regression (REG) of the final model with selected transformed variables of low collinearity implemented using an iterative procedure called "Random Cross-Validation" (RCV) that splits multiple times the data in calibration and validation subsets considering a random selection procedure, c) robustness control of the estimated regression coefficients from RCV-REG by analyzing the sign (+ or -) variation of their iterative solutions using the 95% interval of their Highest Posterior Density Distribution (HPD). The iterative procedure of RCV can also be implemented on machine learning methods (MLs) and for this reason, the ML method of Random Forests (RF) was also applied with RCV (RCV-RF) as an additional case in order to be compared with RCV-REG. Random splitting of data into calibration and validation set (70% and 30%, respectively) was performed 1,000 times in RCV-REG and led to a respective number of solutions of the regression coefficients. The same number of iterations and random splitting for validation was also used in the RCV-RF. The results showed that RCV-REG outperformed RCV-RF at all model performance criteria providing robust regression coefficients associated to independent variables (constant signs of their 95% HPD interval) and better distribution of validation solutions in the iterative 1:1 plots from RCV-RF (RCV-RG: R^2=0.843, RMSE=0.853, MAE=0.642, MAPE=0.081, NSE=0.836, Slope(1:1 plot)=0.998, Intercept(1:1 plot)=0.011, and RCV-RF: R^2=0.835, RMSE=0.904, MAE=0.689, MAPE= 0.088, NSE=0.818, Slope(1:1 plot)=1.120, Intercept(1:1 plot)=-1.011, based on the mean values of 1,000 iterations). The use of RCV approach in various modelling approaches solves the problem of subjective splitting of data into calibration and validation sets, provides a better evaluation of the final modelling approaches and enhances the competitiveness of typical regression models against machine learning models.

Keywords: Class A pan evaporation, random cross validation, regression model, random forest model, machine learning, Highest Posterior Density Distribution of solutions.

1. INTRODUCTION

Evaporation is among the leading components of the hydrologic cycle since it transforms liquid water into gas form, which is diffused into the atmosphere enriching the clouds that regulate precipitation. For this reason, it is always a hot research topic especially during the last years when the analysis of climate change has become a crucial component in developing water resources management plans (Konapala et al., 2020; Althoff et al., 2020). The evaporative water flow rate from large water bodies is significant and the simulation of this loss is prerequisite to understand the contribution of evaporation to hydrologic cycle under varying climatic conditions.

The most common experimental procedure for measuring water evaporation is the pan evaporation (E_{pan}) method. This method is based on measurements of water level fluctuations in evaporimeter tanks (pans), which have specific properties. The most common evaporimeter types are the class A and the Colorado sunken pans (Doorenbos and Pruitt, 1977; Allen et al., 1998). Although, pan evaporation observations are not equivalent to evaporation rates of large water bodies (e.g. lakes), their values are highly correlated and can be useful in understanding the mechanisms that take place between the water surface and the atmosphere, helping to find transition methods between the magnitudes of the two different evaporation types (i.e. pan, lake evaporation) (Finch and Hall, 2001).

Evaporation measurements are extremely useful for researchers and water resources planners, but the installation and the preservation of evaporation pans exhibits a lot of difficulties since their employment cannot be fully automated (e.g. water filling, cleaning of pan etc). For this reason, pan evaporimeter is not a basic instrument of a meteorological station and this led to many efforts for modelling pan evaporation by using meteorological parameters (Finch and Hall, 2001). The first models that were developed are transformations of energy balance equations in combination with terms of water vapor removal (Penman, 1948, 1956; Brutsaert and Lei Yu, 1968). Later, more comprehensive models were developed, which considered more processes involved in the procedure of evaporation, through the evaluation of existing energy balance models (Xu and Singh, 1998; Molina et al., 2006; Valiantzas, 2006). Other models were simply based on regression analysis using measurements from typical meteorological stations and evaporation pans under various climates that differ in the number of required input variables and their form (Irmak and Haman, 2003; Konvoor and Nandagiri, 2007; Almedeij, 2012).

The last years, artificial intelligence has become very popular in many research fields, as well as in hydrology and agrometeorology, and the derived machine learning (ML) algorithms have significantly improved the performance of modelling efforts. Machine learning models do not consist of mathematical equations, which describe physical processes, but they are data driven models, so called black-box models, and their parameterization and performance depends on the attributes of the available data. Since the beginning of 2000, the first Machine Learning implementation approaches have been implemented to calculate daily E_{pan}, mainly using artificial neural network (ANN) methods and, in general, performed better compared to regression models (Bruton et al., 2000; Keskin and Terzi, 2006; Piri et al., 2009; Rahimikhoob, 2009; Shirsath and Singh, 2010; Tabari et al., 2010; Kim et al., 2012; Alsumaiei, 2020). Except for the ANN models, at that time, researchers had developed models using fuzzy logic to estimate E_{pan} (Keskin et al., 2004; Kisi et al., 2005) and other artificial intelligence methods (genetic programming, regression trees), along with ANN, with adequate performance and in some cases with limited available data (Chang et al., 2013; Shiri et al., 2014; Kim et al., 2015). Later, many researchers attempted to calculate E_{pan} by developing models combining machine learning and numerical analysis (hybrid models). The developed hybrid models, in general, further improved the accuracy of the estimations of E_{pan} compared to the machine learning models (Pammar and Deka, 2015; Deo and Samui, 2017; Wang et al., 2017; Ashrafzadeh et al., 2018; Ghorbani et al., 2018; Seifi and Soroush, 2020; Wang et al., 2020). ML models generally show better performance than regression models, but they are highly dependent on the way the data set is split into training and test set and consequently they are prone to fitting the possible noise of the data set (overfitting) (Dietterich, 1995).

A recently published work by Babakos et al. (2020) provided a new approach for assessing the robustness of regression model coefficients but also for comparing the predictive power of regression and machine learning models. The specific approach was a combination of bootstrap and cross-validation techniques that allowed to assess the predictive power during the validation procedure considering the probabilistic range (based on highest posterior density distribution) of regression coefficients, statistical metrics (R^2, RMSE) and the slope and intercept of linear trend line from 1:1 plots between observed vs. predicted values. The analysis was based on pan evaporation measurements for assessing reference crop evapotranspiration. The results of the analysis showed that a Random Forest model (machine learning model) showed slightly better statistical metrics (R^2, RMSE) from a regression model, but it was not balanced showing worse slope and intercept values of the trend line than the regression model.

The aim of this study is to develop a regression model for simulating pan evaporation at local conditions by using only meteorological parameters based on the approach of Babakos et al. (2020) and to compare its strength versus a machine learning approach (Random Forests). The steps presented in this paper could be used as example to investigate and compare the proposed regression method against machine learning methods that use only meteorological data for the simulation of pan evaporation.

2. DATA AND METHODS

2.1. Data

Daily meteorological data of precipitation (P), temperature (T), solar radiation (R_s), relative humidity (RH) and wind speed at 2 m above ground (u_2) covering the warm-dry periods (May to September) of 2008 and 2009 were obtained from the meteorological station located in the Aristotle University Farm (~1 m a.s.l., 40°32’08” N, 22°59’18” E) in Thessaloniki (Greece). The daily values of the meteorological parameters were calculated as mean values of hourly observations of a 24-h period. Moreover, a class A pan evaporimeter made by Monel metal with fetch distance F = 1 m (green upwind fetch - Case A) was used for obtaining daily E_{pan} measurements during the same periods of 2008 and 2009. The climate and evaporation data are representative of the warm-dry season conditions of the Thessaloniki Plain in Greece, where the climate is considered as semi-arid Mediterranean environment (Hastie et al., 2009). The 5-month period of May–September is the period for cultivating summer crops, and it is responsible for more than 70% of the annual reference evapotranspiration in the study area (Aschonitis et al., 2018). The meteorological data were used in this study for building models that estimate daily E_{pan}. The records of rainfall days ($P > 0$) during May–September were excluded from the analysis, leading to a final number of 212 daily records of meteorological and E_{pan} data. The statistical properties and distribution characteristics of the data are given in Table 1 and in Figs. 1,2.

2.2. Methods of analysis

The methodological steps that are going to be followed are provided in the following subsections and in the flowchart presented in Fig. 3.

Table 1. Statistical properties and distribution characteristics of daily measured Class A pan evaporation (E_{pan}), temperature (T), incident solar radiation (R_s), relative humidity (RH), and wind speed at 2 m above ground surface (u_2) after excluding rainfall days.

Parameter	T (°C)	R_s (MJ m^{-2} d^{-1})	RH (%)	u_2 (m s^{-1})	E_{pan} (mm d^{-1})
Minimum	16.70	7.41	31.55	0.85	2.53
Lower quartile	23.79	21.46	47.94	1.20	6.80
Average	25.56	23.87	54.91	1.43	8.14
Median	26.07	24.91	54.99	1.36	8.45
Upper quartile	27.75	27.26	62.10	1.50	9.33
Maximum	31.01	30.04	81.14	3.34	14.85
Range	14.32	22.63	49.58	2.49	12.32
Standard deviation	2.95	4.30	9.65	0.43	2.15
Coeff. of variation	11.56%	18.03%	17.56%	30.32%	26.42%
Stnd. skewness	-3.95	-6.56	0.05	15.26	-0.12
Stnd. kurtosis	0.26	3.66	-1.07	23.09	1.24
Kolmogorov-Smirnov Norm. Test (p-value)*	0.099	<0.05	0.93	<0.05	0.14
Shapiro-Wilk Norm. Test (p-value)*	<0.05	<0.05	0.82	<0.05	<0.05

* p-values <0.05 indicate that data do not follow a normal distribution at 95% confidence level (for both normality tests).

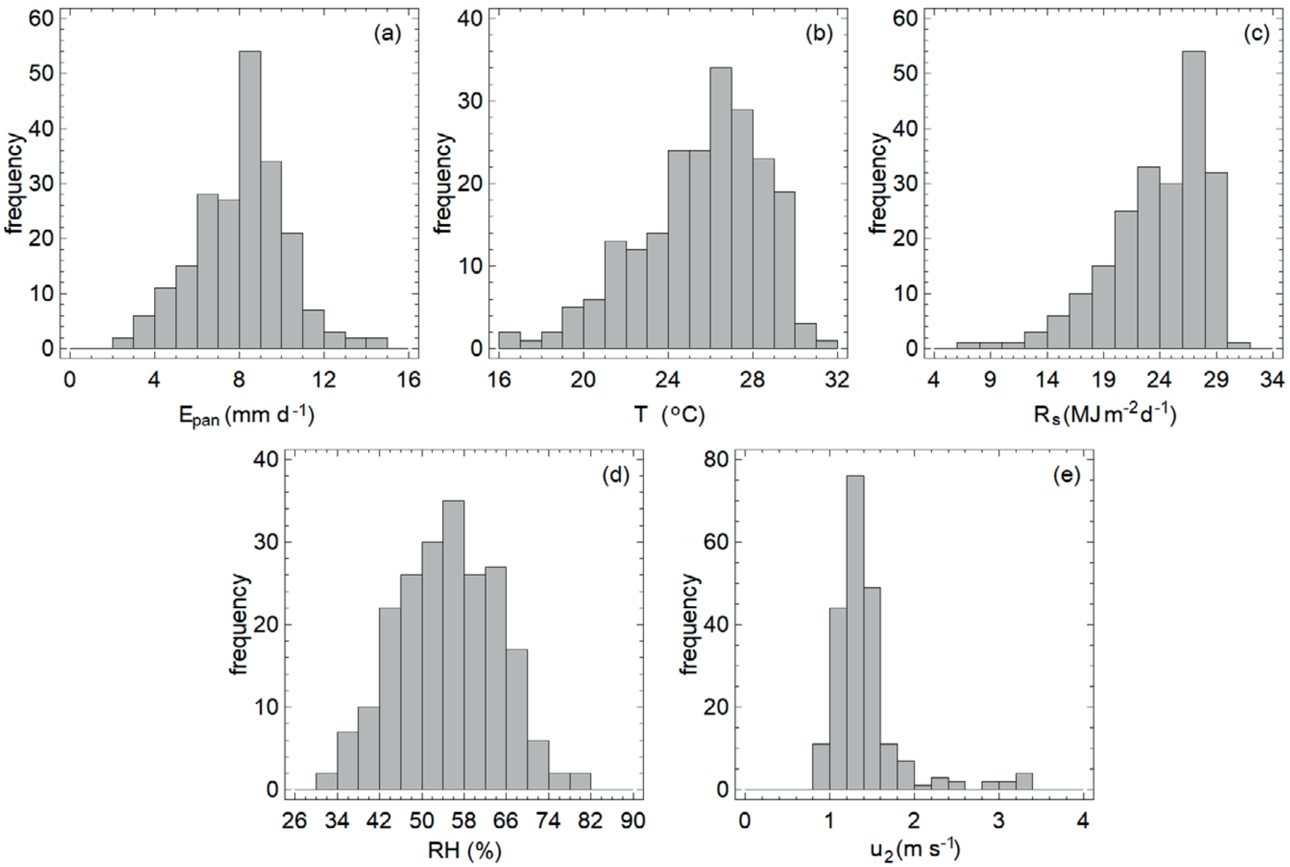

Figure 1. Frequency histograms for daily class A pan evaporation (E_{pan}) data and for daily meteorological parameters of temperature (T), incident solar radiation (R_s), relative humidity (RH) and wind speed at 2 m above ground (u_2).

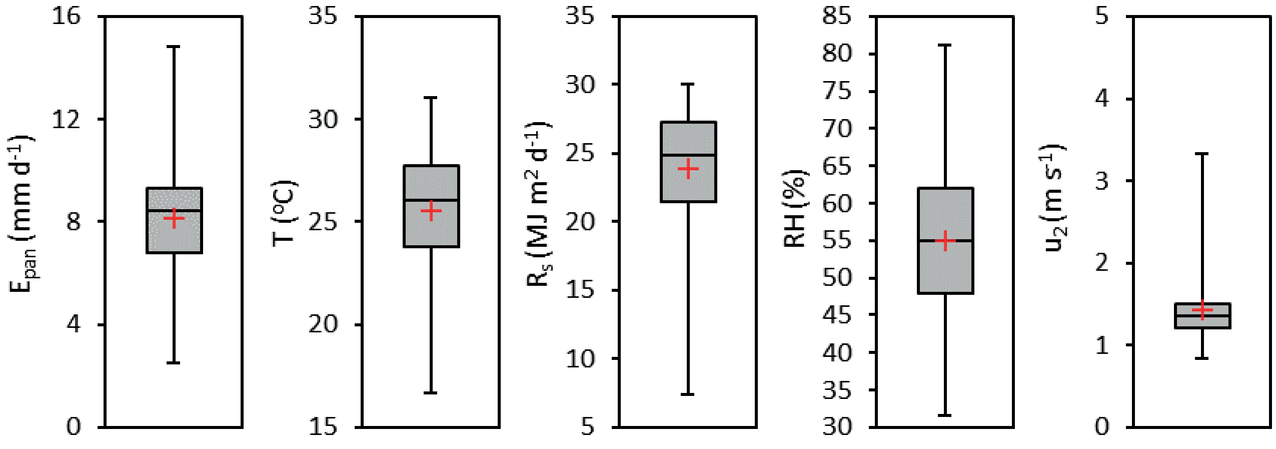

Figure 2. Box-Whisker plots for daily class A pan evaporation (E_{pan}) data and for daily meteorological parameters of temperature (T), incident solar radiation (R_s), relative humidity (RH) and wind speed at 2 m above ground (u_2).

2.2.1. Transformation of variables and Ridge regression

A common problem when building models based on meteorological parameters is the high multicollinearity that may appear among some parameters (e.g. temperature and solar radiation). High multicollinearity among independent variables leads to imprecise estimates of the regression model coefficients using ordinary

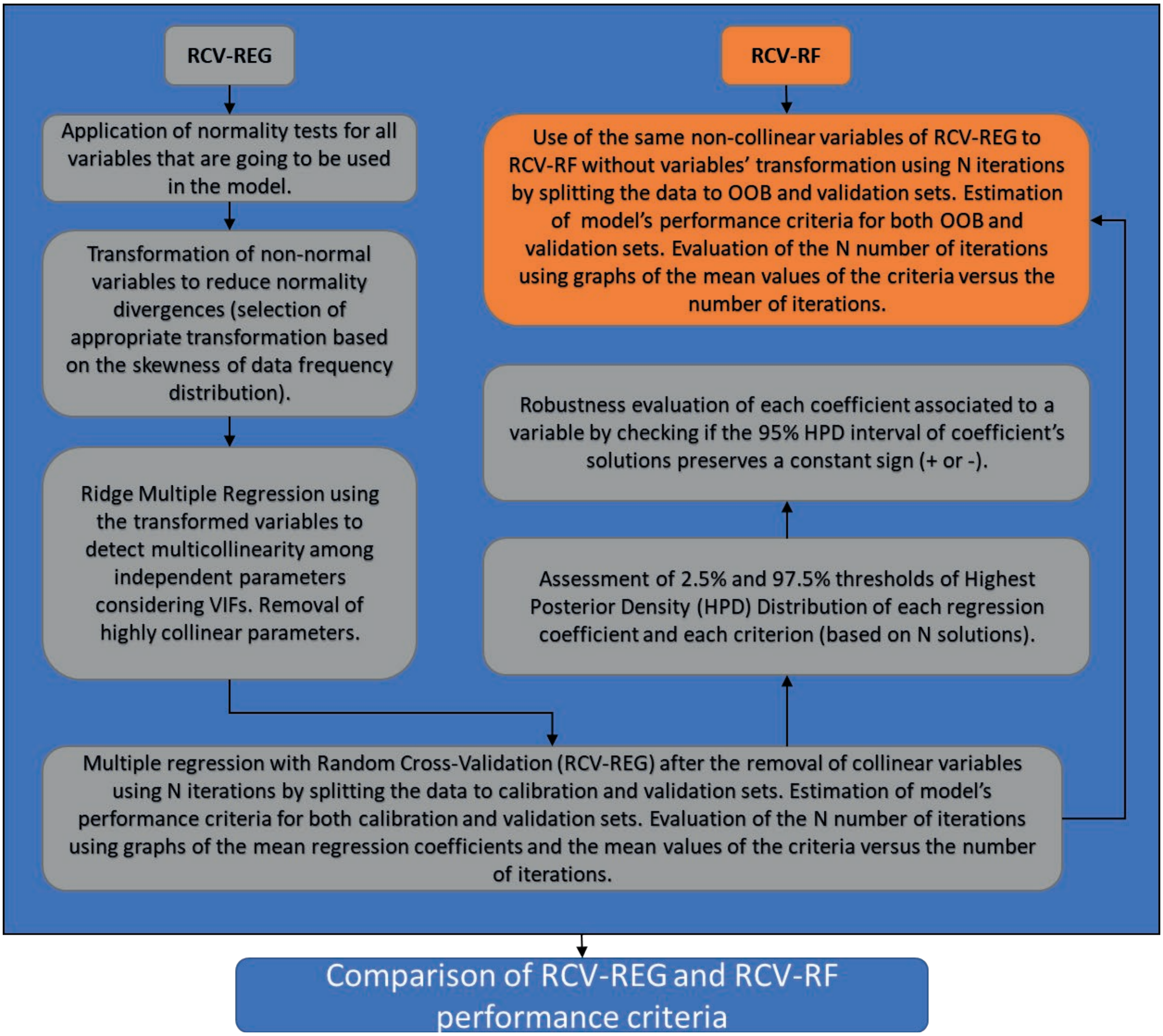
RCV-REG
RCV-RF
Application of normality tests for all variables that are going to be used in the model.
Transformation of non-normal variables to reduce normality divergences (selection of appropriate transformation based on the skewness of data frequency distribution).
Ridge Multiple Regression using the transformed variables to detect multicollinearity among independent parameters considering VIFs. Removal of highly collinear parameters.
Use of the same non-collinear variables of RCV-REG to RCV-RF without variables' transformation using N iterations by splitting the data to OOB and validation sets. Estimation of model's performance criteria for both OOB and validation sets. Evaluation of the N number of iterations using graphs of the mean values of the criteria versus the number of iterations.
Robustness evaluation of each coefficient associated to a variable by checking if the 95% HPD interval of coefficient's solutions preserves a constant sign (+ or -).
Assessment of 2.5% and 97.5% thresholds of Highest Posterior Density (HPD) Distribution of each regression coefficient and each criterion (based on N solutions).
Multiple regression with Random Cross-Validation (RCV-REG) after the removal of collinear variables using N iterations by splitting the data to calibration and validation sets. Estimation of model's performance criteria for both calibration and validation sets. Evaluation of the N number of iterations using graphs of the mean regression coefficients and the mean values of the criteria versus the number of iterations.
Comparison of RCV-REG and RCV-RF performance criteria

Figure 3. Flowchart of the methodological steps.

least squares, whereas the final model tends to overfit the data. A solution to this problem is the use of Ridge regression analysis, which is based on the idea that the variance of the slope estimates can be greatly reduced by introducing some bias into them. Ridge regression analysis also includes the estimation of the variance inflation factor (VIF) of each independent variable, which is a valuable indicator of multicollinearity among them. A large VIF has not been universally defined, but it is commonly considered large when exceeds the threshold value 10; however, some use 4 as threshold value (Kutner et al., 2004; O'brien, 2007; Vatcheva et al., 2016; Helsel et al., 2020).

Considering the above, ridge regression to fit the E_{pan} data using the T, R_s, RH and u_2 parameters was considered a crucial step to detect if multicollinearity exists among the independent variables. Before ridge regression, the normality tests of Shapiro-Wilk and Kolmogorov-Smirnov (STATGRAPHICS Centurion XV software, StatPointTechnologies Inc.) were used to identiy normality divergences of E_{pan}, T, R_s, RH and u_2 data for $p<0.05$ (Table 1). R_s and u_2 data failed to pass both normality tests, T and E_{pan} succeeded to pass only the test o Kolmogorov-Smirnov while RH succeeded to pass both tests. To reduce normality divergences, different transformations were employed according to the rules of Helsel (2020) that are based in data skewness. T and R_s data were negatively skewed (Fig.1b,c) and for this reason the square transformation (x^2) was used. On the other hand, u_2 data were positively skewed (Fig.1e) and for this reason the logarithmic transformation ln(x+1) was used. E_{pan} was not transformed because different tested transformations did not improve the results of the normality tests. Thus, the final form of the ridge regression model with transformed variables was the following:

$$E_{pan} = a + b \cdot T^2 + \mathrm{c} \cdot R_s{}^2 + d \cdot RH + e \cdot \ln(u_2 + 1) \qquad (1)$$

where E_{pan} is the measured evaporation from the evaporimeter (mm d^{-1}), T is the mean daily air temperature (°C), R_s is the incident solar radiation (MJ m^{-2} d^{-1}), u_2 is the mean daily wind speed at 2 meters height (m s^{-1}), RH is the mean daily relative humidity (%), u_2 is the mean daily wind speed at the height of 2 m (m s^{-1}). The ridge regression is used as a preliminary control procedure to assess multicollinearity before proceeding to the modelling approach of RCV-REG, which is described in the next section. In case of high multicollinearity, the indipedent variables of Eq. 1 are reduced in order to reach an acceptable VIF value of the remaining parameters before their use in RCV-REG.

2.2.2. Regression with Random Cross-Validation (RCV-REG)

The E_{pan} model (Eq. 1) contains non-linear transformations of the independent variables and its predictive power was investigated using a random cross-validation regression (RCV-REG) analysis based on the concept of Babakos et al. (2020). The RCV-REG analysis performs a random splitting of the initial dataset into a calibration set (70% of the records) and a validation set (30% of the records). This random splitting is performed 1000 times (number adusted by the user), leading to a respective number of calibration and validation pairs of datasets. The calibration procedure leads to 1000 estimations of the regression coefficients of Eq. 1. The estimated coefficients of each calibration set were used to validate the model based on the respective validation set. The RCV-REG was built in R software using the "nls.lm" function of the {minpack.lm} package (Guan et al., 2020), which includes the Levenberg-Marquardt nonlinear least-squares algorithm. The range of 1000 solutions of each regression coefficient from calibration and validation procedures was respectively defined by the 95% confidence interval of the highest posterior density (HPD) distribution. The 2.5% and 97.5% thresholds (HPD thresholds) containing the central 95% interval of the HPD distribution were estimated in R software using the "p.interval" function of the {LaplacesDemon} package (Majhi et al., 2020). This function returns unimodal or multimodal HPD intervals depending on the form of the probability distributions. The HPD intervals are extremely valuable since they can provide information about the robustness of regression coefficients, the robustness of the independent variables associated to them and consequently the robustness of the overall model. The following robustness rule was suggested by Babakos et al. (2020) based on the results of the complete RCV-REG procedure: "a model is robust only when the 95% HPD intervals of all its regression coefficients associated to independent variables preserve a constant sign (+ or -)". When a 95% HPD interval of a regression coefficient contains positive and negative values, then it indicates a non-robust coefficient (non clear effect of the indipedent variable) that can significantly affect the robustness of the model.

2.2.3. Random Forests with Random Cross-Validation (RCV-RF)

Random forests (RF) is among the most important machine-learning methods (Breiman, 2001), which is an improvement of the Classification and Regression Trees (CART) method (also called decision trees). RF employs a modification of the bootstrap aggregating technique (bagging), where a large collection of decorrelated, noisy, approximately unbiased trees are constructed and averaged in order to minimize the model variance and instability problems (Hastie et al., 2009). RF is an ensemble method where the aggregation of multiple trees increases the prediction accuracy, with results described by both low bias and low variance (Breiman, 2001; Diaz-Uriarte and De Andres, 2006). Advantages of RF are the ability of modeling high-dimensional nonlinear relationships with few user-defined parameters, relative robustness with resistance to overfitting and estimation of importance of the variables (Dietterich, 1995; Breiman, 2001; Hastie et al., 2009; Diaz-Uriarte and De Andres, 2006; Strobl et al., 2009).

The hyperparameters of the RF model significantly affect model's performance. The hyperparameters that were considered in this study are the number of the regression trees (num.trees), the proportion of train set that was used for building the model (sample.fraction), the number of candidate predictors that were randomly sampled (mtry), and the minimum number of points in the terminal nodes of the regression trees (min.node.size). Different combinations of various values of hyperparameters was built, and by executing the "ranger" package (Wright et al., 2020), their optimal set of values was determined (mtry = 2, num.trees = 1000, sample.fraction = 0.7, and min.node.size = 5). The validation set was set to 30% of the initial dataset in order to be comparable with the RCV-REG. RF also includes an iterative process (Out-Of-Bag - OOB) during training (calibration) by using different sets of the training dataset (the rest 70%), which is used to reduce the variance without changing the bias of the complete ensemble. RF also estimates the predictor variables' importance, which is calculated as the mean-across all trees-decrease (%) of accuracy, expressed by the % change in Mean Squared Error - MSE (%) of the Out-of-Bag (OOB) sample when the variable is not considered by permuting its values randomly and maintaining the others as they were.

The internal random procedures in RF lead every time to different solutions using constant optimal values of hyperparameters (Hastie et al., 2009). For this reason, 1000 RF iterations (RCV-RF) were also made in order to make comparisons with the RCV-REG model (Eq. 1). RF is not restricted by the limitations of a predefined nonlinear form and can be used as a benchmark model for evaluating the predictive accuracy of typical regression models but, also, for assessing the relative importance of the predictor meteorological variables to affect E_{pan}. The main reason for selecting the Random Forests approach is that it does not consider assumptions regarding normality, linearity, homoscedasticity, and collinearity. It also does not demand a high sample-to-predictor ratio, it is very suitable to interaction effects (including nonlinearity) and it is recognized as one of the state-oftheart methods in terms of prediction accuracy (Flach 2012; Geurts et al., 2009; Golino and Gomes, 2016). The RCV-RF analysis was performed using all the same predictor variables of RCV-REG without transforming the variables since RF does not consider assumptions of normality.

2.2.4. Models' performance criteria and evaluation of required number of iterations used in RCV-REG and RCV-RF

The models' performance criteria of R^2, the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the Nash–Sutcliffe efficiency were estimated: a) for each calibration and validation dataset of the RCV-REG analysis for Eq. 1, and b) for the OOB and validation dataset of the RCV-RF analysis; leading to 1000 respective estimations of their values for each model. Moreover, the 1000 estimations of slope and intercept of the trend line in the 1:1 plot of observed vs. predicted E_{pan} models only using the validation sets were also made (for both RCV-REG or RCV-RF) as complementary criteria. The 1000 estimations of the criteria were also analyzed through the computation of HPD intervals.

The evaluation of the selected number of iterations (i.e. 1000) used in RCV-REG and RCV-RF was performed using individual graphs of the mean regression coefficients (only for RCV-REG) and of the mean criteria values (for both RCV-REG and RCV-RF) versus the number of iterations. Based on these graphs, it was assumed that the required number of iterations is succeeded when the mean value of a regression coefficient or a criterion reaches a stablized plateau.

3. RESULTS

Ridge regression analysis was performed on the transformed variables of Eq. 1 to detect any multicollinearity among them and the VIF results of the variables are shown in Table 2. The VIF values of all the variables were below the threshold value 4, suggesting low multicollinearity degree and low overfitting effects by their combined use. For this reason, indipedent variables were not removed from Eq. 1 and it was used as it is in the RCV-REG approach.

The mean, standard error, minimum, maximum and 2.5% and 97.5% HPD quantiles of the coefficients of the RCV-REG approach for Eq. 1 are given in Table 3. The statistical metrics and the robustness rule of Babakos et al. (2020) based on the 1000 iterations (Table 3) showed that the form of Eq. 1 is robust considering that all the estimated coefficients associated to independent variables have stable sign between 2.5% and 97.5% HPD thresholds. Only the regression coefficient of intercept (coefficient a in Eq. 1) does not follow the rule of robustness, but it is not associated to an independent variable. The regression model's performance is described by the statistical criteria given in Table 4 for both the calibration and validation subsets.

Table 2. VIF values of Ridge regression analysis on transformed data.

	Value of the coefficient	VIF
Constant	-0.51	-
T^2	0.005	1.54
$\ln(u_2+1)$	4.03	1.63
RH	-0.027	1.87
R_S^2	0.006	1.47
R^2 of the regression	0.852	

Table 3. Highest posterior density distributions of 1000 estimations of coefficients of independent variables for the RCV-REG model (Eq. 1).

	a	b	c	d	e
Mean	-0.458	0.005	4.019	-0.028	0.006
St.Error	0.022	0.000	0.012	0.000	7.14E-06
Min	-2.542	0.003	2.611	-0.048	0.005
HPD thres. 2.5%	-1.995	0.004	3.131	-0.040	0.005
median	-0.453	0.005	4.018	-0.028	0.006
HPD thres. 97.5%	0.788	0.006	4.707	-0.017	0.006
Max	1.970	0.006	5.272	-0.008	0.006

The mean, standard error, minimum, maximum and 2.5% and 97.5% HPD thresholds of the importance indicator of the variables from the RCV-RF method are given in Table 5, where the importance of independent variables showed the following ranking R_s>RH>u_2>T.

As regards the implementation of RCV-RF approach, the statistical metrics for both the OOB set and the validation set are given in Table 6.

Considering the mean values of metrics for the validation datasets (1,000 iterations) of RCV-REG and RCV-RF (Table 4 and 6), it is observed that RCV-REG outperformed RCV-RF in all criteria indicating that the proposed regression approach can compete the accuracy of machine learning methods for building evaporation models.

Finally, for the evaluation of the selected number of iterations (i.e. 1000) used in RCV-REG and RCV-RF, the individual graphs of the mean regression coefficients (only for RCV-REG) and of the mean criteria values from the validation procedure (for both RCV-REG and RCV-RF) versus the number of iterations are given in Fig. 4 and 5. Based on these graphs, it is evident that the all the regression coefficients and perfromance criteria reached a stablized plateau even after 500 iterations. Thus, the number of 1000 iterations is considered more than enough for assuming a robust analysis using both approaches.

Table 5. Importance indicators of indipedent variables based on the RCV-RF approach.

Imp. Indicator	T	u_2	RH	R_s
Mean	75.992	84.319	88.730	204.366
St.Error	0.395	0.418	0.412	0.529
Min	36.362	49.824	50.894	152.022
HPD thres. 2.5%	51.118	60.887	63.887	170.893
median	76.208	84.023	87.980	204.772
HPD thres. 97.5%	98.923	111.048	113.947	234.927
Max	111.194	132.797	142.721	264.168

4. DISCUSSION

4.1. Performance of the models

The most possible reason to justify why RCV-REG showed better performance from RCV-RF in the external validation (metrics denoted as pred. in Tables 4 and 6) is probably due to the normality improvement of the data based on proper selection of transformations used in RCV-REG. Transformed variables were not used in

Table 4. Performance criteria for the RCV-REG model.

Criterion		Mean	St.Error	Min	HPD thres. 2.5%	Median	HPD thres. 97.5%	Max
Calibration	R^2	0.854	0.000	0.801	0.827	0.854	0.880	0.893
	RMSE	0.817	0.001	0.715	0.746	0.822	0.887	0.914
	MAE	0.615	0.001	0.544	0.571	0.616	0.669	0.692
	MAPE	0.078	0.000	0.068	0.072	0.078	0.084	0.088
	NSE	0.854	0.000	0.801	0.827	0.854	0.880	0.893
Validation	R^2	0.843	0.001	0.675	0.777	0.845	0.907	0.926
	RMSE	0.853	0.003	0.570	0.703	0.846	1.020	1.073
	MAE	0.642	0.002	0.456	0.526	0.640	0.751	0.850
	MAPE	0.081	0.000	0.057	0.067	0.081	0.096	0.110
	NSE	0.836	0.001	0.660	0.770	0.838	0.903	0.919
	Intercept*	0.011	0.015	-1.493	-0.853	0.005	1.059	1.288
	Slope*	0.998	0.002	0.836	0.875	0.998	1.116	1.197

*Estimated only for the validation set.

Table 6. Performance criteria for the RCV-RF model.

	Criterion	Mean	St.Error	Min	HPD thres. 2.5%	Median	HPD thres. 97.5%	Max
OOB	R2	0.823	0.001	0.761	0.791	0.824	0.852	0.881
	RMSE	0.902	0.001	0.788	0.842	0.903	0.968	1.002
	MAE	0.685	0.001	0.599	0.636	0.684	0.737	0.758
	MAPE	0.088	0.000	0.076	0.081	0.088	0.094	0.097
	NSE	0.821	0.001	0.760	0.790	0.823	0.851	0.880
Validation	R2	0.835	0.001	0.673	0.775	0.838	0.896	0.910
	RMSE	0.904	0.003	0.649	0.744	0.896	1.093	1.267
	MAE	0.689	0.002	0.481	0.564	0.687	0.820	0.918
	MAPE	0.088	0.000	0.061	0.071	0.088	0.107	0.117
	NSE	0.818	0.001	0.663	0.759	0.821	0.881	0.892
	Intercept*	-1.011	0.020	-3.236	-2.299	-1.021	0.120	0.657
	Slope*	1.120	0.002	0.898	0.963	1.121	1.260	1.378

*Estimated only for the validation set.

Random forest method because this approach overcomes problems of normality, linearity, homoscedasticity and collinearity (Flach 2012; Geurts et al., 2009; Golino and Gomes, 2016) since it doesn't use metric distances between data points but applies splits along a tree. Another possible reason is the difference in the degrees of freedom of the two models (associated to the number of coefficients that are free to vary) in combination with the number of records used in this study. Models that have low degrees of freedom (e.g. linear or non linear regression such as RCV-REG) are not usually so flexible to fit the data. Thus, the lower the degrees of freedom of a model the lower is the effect of the noise/bias in the data included in the final model (e.g. noise/bias in the data may come from other sources such as errors associated to the observer). Moreover, when the number of data records to calibrate the model is low, the larger is the effect of the noise/bias included in the final calibration. Based on the above, machine learning models that are generally used to solve problems using Big Data and they have much more degrees of freedom, may not be the proper choice for typical modelling applications, where the number of data is small, because they "absorb" a large portion of the data noise. This may lead to a lower performance of a machine learning model compared to a regression model during external validation. In this study, the number of records were 212 and they are not in the category of Big Data but they are enough for the typical regression analysis. This was an additional reason for including the RCV approach of iterations in both modelling approaches.

4.2. Limitations of RCV-REG approach

The RCV-REG iterative procedure in combination with the preliminary analyses of the Ridge regression and the tranformation of variables can be considered a complete methodology that takes into account all the nessecary elements for building a robust model. On the other hand, the final form of the model is based on the experience of the user, which may not be adequate to achieve the maximum potential of the methodology. Moreover, RCV-REG approach is limited to provide a graphical representation of the results. For example, it is typical in modelling approaches to provide 1:1, quantile:quantile (Q:Q) plots of observed vs. predicted data, 2D plots of the respective joint probability distribution etc. The problem of the RCV-REG is that there are 1,000 iterations that neither can be plotted seperately (due to the large number) nor to merge the results of all iterations in one graphic type. The second case is feasible for 1:1 plot but leads to clouds of points that come from different 30% of the initial data.

4.3. Reasons for excluding records of cold season and rainfall days

The reason for not including pan evaporation measurements of the cold season and of the rainy days was because these measurements have a lot of bias. During the six-month cold season in this location, E_{pan} is low and generally falls in the range 0-2 mm/day not only due to the lower temperature but also due to high relative humidity with a lot of foggy days and condensed moisture (dew) in the leaves of the surrounding vegetation

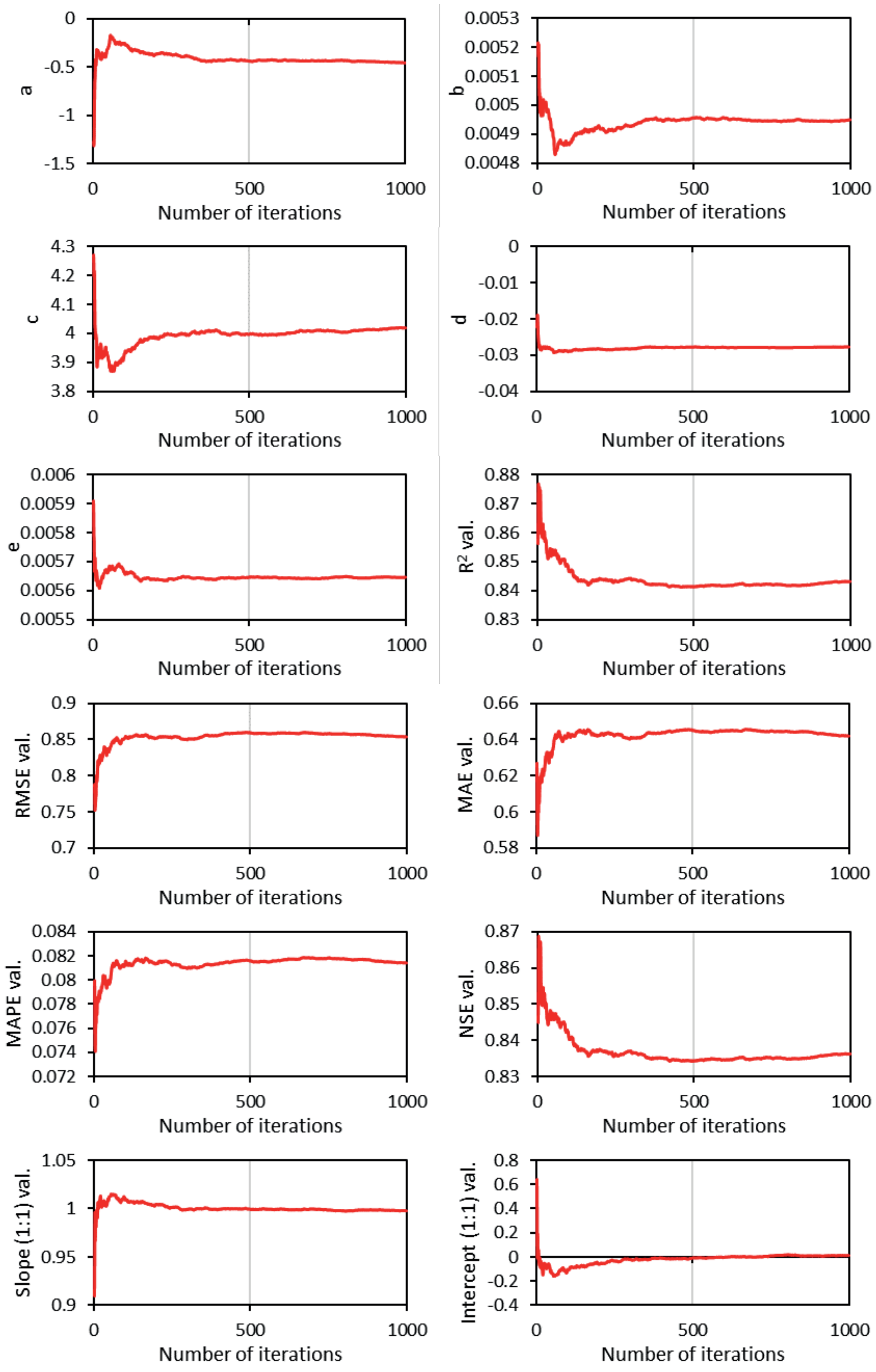

Figure 4. Variation of mean values of regression coefficients of Eq. 1 and performance criteria from the validation procedure versus the number of iterations for RCV-REG.

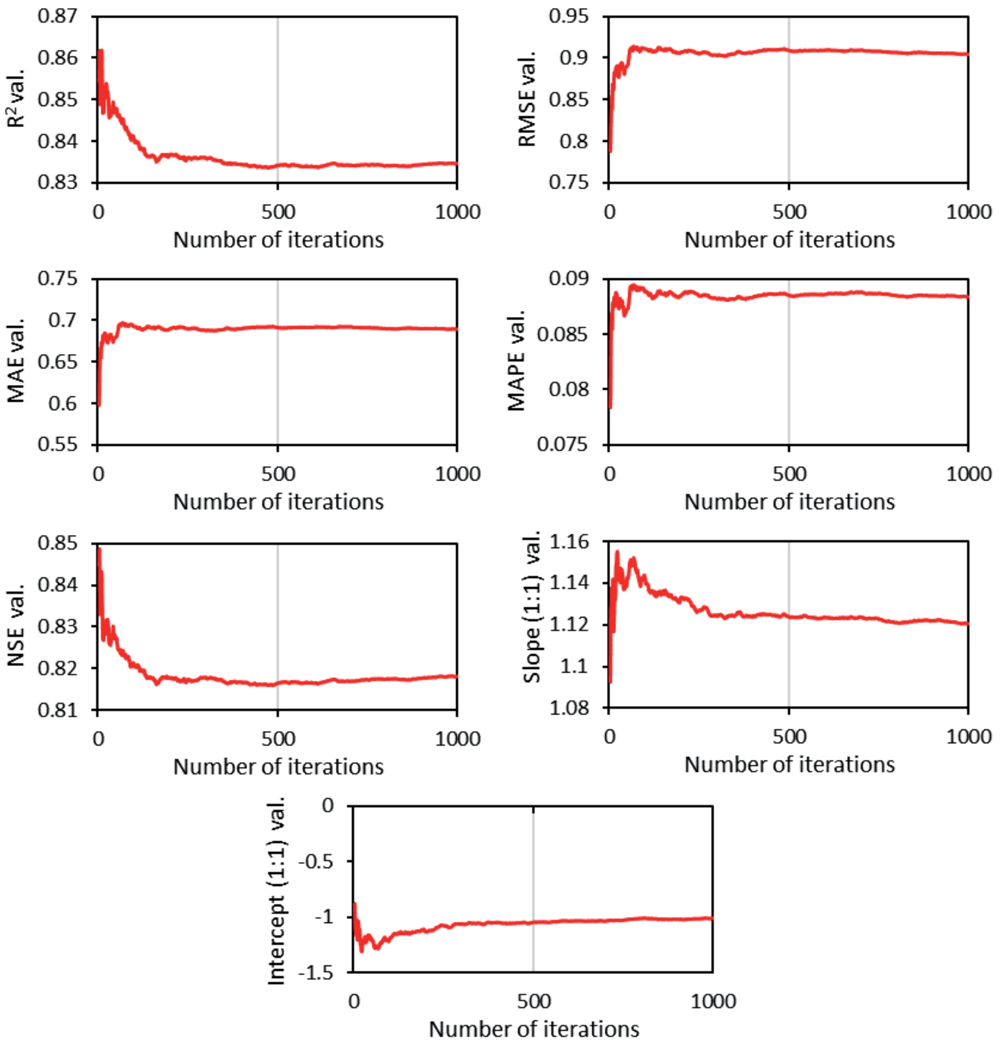

Figure 5. Variation of mean values of performance criteria from the validation procedure versus the number of iterations for RCV-RF.

during the morning. During these months, even during April or October, it was observed the occurrence of negative E_{pan} measurements due to condensed dew input in the pan. The rainfall days during the warm season were also excluded because the temperature of rainfall is quite different from the water temperature in the pan (even 10°C) and their mixing changes the evaporative energy demand. The records of rainfall days could only be used in the case of deterministic modelling approaches based on energy budget where the water temperature is used as input parameter.

Another reason for not including the data from the six-month cold season is that the seasonal variation of temperature and solar radiation between summer and winter leads the two variables to be collinear and thus the one should be removed from the modelling approach. Using only the data of the warm season, the two variables present lower collinearity that allows their combined use in the models.

The general concept and methodological steps presented in this study are valid for all areas that have distinct warm and cold seasons. If cold season does not exist (e.g. tropical environments), data from all seasons can be included. Rainfall days can also be included but it is expected to lead to a reduction in the predictive accuracy of the final model. The larger the proportion of the rainfall days in the final record, the larger is expected to be the reduction of the predictive accuracy. For the inclusion of rainfall days, it is proposed the inclusion of rainfall variable in Eq. 1 or the use of a categorical variable for splitting the records in rainfall and no-rainfall days. The second case is considered as categorical regression.

5. CONCLUSIONS

The implementation of the RCV-REG method, which includes regression with transformed variables of low collinearity and analysis of the 95% HPD of regression coefficients, was found to be an extremely powerful approach for E_{pan} analysis that can compete machine learning methods and can provide a complete evaluation of the regression coefficients robustness. As it was shown from the results of this study for modelling E_{pan} using meteorological variables, the specific method succeeded to outperform RCV-RF in all performance criteria. Moreover, RCV-REG gave a better aspect and evaluation of the robust effect of the indipedent variables (through the HPD analysis on the regression coefficients associated to the indipedent variables) in comparison to RCV-RF that is able to provide only a relative ranking of indipedent variables' importance.

Moreover, the use of the RCV data splitting approach in various modelling approaches solves the problem of subjective splitting of data into calibration and validation sets, provides a better evaluation of the final modelling approaches and enhances the competitiveness of typical regression models against machine learning models. Despite the fact that machine learning methods are more advanced in comparison to typical regression methods, mostly by handling better Big Data, they still face the problem of transferability from the developer to the user for various reasons, such as non-availability of the calibrated code or its form (since it is a black box) and lack of users's expertise to handle such models. On the other hand, the resulting models through typical regression approaches do not require advanced skills and can be used in other studies either by adopting the entire calibrated model or by adopting the general form of the model.

Future studies should focus on (a) the investigation of E_{pan} models with the inclusion of records of rainfall days and (b) the investigation of new graphical methods for representing different elements of the results of the RCV-REG method.

REFERENCES

Allen R.G., Pereira L.S., Raes D., Smith M., 1998. Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements. Irrigation and Drainage Paper 56, Food and Agriculture Organization of the United Nations: Rome.

Almedeij J., 2012. Modeling Pan Evaporation for Kuwait by Multiple Linear Regression. The Scientific World Journal, Ar. ID 574742: 1-9. https://doi.org/10.1100/2012/574742

Alsumaiei A.A., 2020. Utility of artificial neural networks in modeling pan evaporation in hyper-arid climates. Water (Switzerland), 1508: 1-12. https://doi.org/10.3390/w12051508

Althoff D., Rodrigues L.N., da Silva D.D., 2020. Impacts of climate change on the evaporation and availability of water in small reservoirs in the Brazilian savannah. Climatic Change, 159: 215–232. https://doi.org/10.1007/s10584-020-02656-y

Aschonitis V., Diamantopoulou M., Papamichail D., 2018. Modeling plant density and ponding water effects on flooded rice evapotranspiration and crop coefficients: critical discussion about the concepts used in current methods. Theoretical and Applied Climatology, 132: 1165-1186. https://doi.org/10.1007/s00704-017-2164-z

Ashrafzadeh A., Malik A., Jothiprakash V., Ghorbani M.A., Biazar S.M., 2018. Estimation of daily pan evaporation using neural networks and meta-heuristic approaches. ISH Journal of Hydraulic Engineering, 26(4): 421-429. https://doi.org/10.1080/09715010.2018.1498754

Babakos K., Papamichail D., Tziachris P., Pisinaras V., Demertzi K., Aschonitis V., 2020. Assessing the robustness of pan evaporation models for estimating reference crop evapotranspiration during recalibration at local conditions. Hydrology, 7(3): 62. https://doi.org/10.3390/hydrologfy7030062

Breiman L., 2001. Random forests. Machine Learning, 45: 5-32. https://doi.org/10.1023/A:1010933404324

Bruton J.M., McClendon R.W., Hoogenboom G.,. 2000. Estimating daily pan evaporation with artificial neural networks. Transactions of the American Society of Agricultural and Biological Engineers, 43(2), 491-496. https://doi.org/10.13031/2013.2730

Brutsaert B., Lei Yu S., 1968. Mass Transfer Aspects of Pan Evaporation. Journal of Applied Meteorology and Climatology, 7: 563–566. DOI:10.1175/1520-0450(1968)007<0563:MTAOPE>2.0.CO,2

Chang F.J., Sun W., Chung C.H., 2013. Dynamic factor analysis and artificial neural network for estimating pan evaporation at multiple stations in northern Taiwan. Hydrological Sciences Journal, 58(4): 813-825. https://doi.org/10.1080/02626667.2013.775447

Deo R.C., Samui P., 2017. Forecasting evaporative loss by least-square support-vector regression and evaluation with genetic programming, gaussian process, and minimax probability machine regression: Case study of Brisbane city. Journal of Hydrologic Engineering, 22(6), art. no. 05017003. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001506

Díaz-Uriarte R., De Andres S.A., 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7: 1-13. https://doi.org/10.1186/1471-2105-7-3

Dietterich T.G., 1995. Overfitting and undercomputing in machine learning. ACM Computing Surveys, 27(3): 326-327. https://doi.org/10.1145/212094.212114

Doorenbos J., Pruitt W.O., 1977. Guidelines for Predicting Crop Water Requirements. Irrigation and Drainage Paper No. 24, 2nd ed., Food and Agriculture Organization of the United Nations: Rome.

Finch J.W., Hall R.L., 2001. Estimation of open water evaporation–a review of methods. R&D Technical Report W6–043/TR. Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol.

Flach, P. 2012. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge: Cambridge University Press.

Geurts P., Irrthum A., Wehenkel L., 2009. Supervised Learning with Decision Tree-based Methods in Computational and Systems Biology. Molecular Biosystems, 5(12), 1593-1605. https://doi.org/10.1039/b907946g

Ghorbani M.A., Deo R.C., Yaseen Z.M., Kashani M.H., Mohammadi B., 2018. Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. Theoretical and Applied Climatology, 133: 1119-1131. https://doi.org/10.1007/s00704-017-2244-0

Golino H.F., Gomes C.M.A., 2016. Random forest as an imputation method for education and psychology research: its impact on item fit and difficulty of the Rasch model. International Journal of Research and Method in Education, 39(4): 401-421. https://doi.org/10.1080/1743727X.2016.1168798

Guan Y., Mohammadi B., Pham Q.B., Adarsh S., Balkhair K.S., Rahman K.U., Linh N.T.T., Tri D.Q., 2020. A novel approach for predicting daily pan evaporation in the coastal regions of Iran using support vector regression coupled with krill herd algorithm model. Theoretical and Applied Climatology, 142: 349-367. https://doi.org/10.1007/s00704-020-03283-4

Hastie T., Tibshirani R., Friedman J., 2009. The elements of statistical learning. 2nd ed. Springer, New York.

Helsel D.R, Hirsch R.M, Ryberg K.R, Archfield S.A, Gilroy E.J., 2020. Statistical Methods in Water Resources. In Book 4, Hydrologic Analysis and Interpretation, U.S. Geological Survey, U.S., 4–A3, pp. 460.

Irmak S., Haman D., 2003. Evaluation of Five Methods for Estimating Class A Pan Evaporation in a Humid Climate. Horttechnology, 13: 500-508. https://doi.org/10.21273/HORTTECH.13.3.0500

Keskin M.E., Terzi Ö., Taylan D., 2004. Fuzzy logic model approaches to daily pan evaporation estimation in western Turkey / Estimation de l'évaporation journalière du bac dans l'Ouest de la Turquie par des modèles à base de logique floue. Hydrological Sciences Journal, 49: 1001-1010. https://doi.org/10.1623/hysj.49.6.1001.55718

Keskin M.E., Terzi Ö., 2006. Artificial neural network models of daily pan evaporation. Journal of Hydrologic Engineering, 11: 65-70. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:1(65)

Kim S., Shiri J., Kisi O., 2012. Pan Evaporation Modeling Using Neural Computing Approach for Different Climatic Zones. Water Resources Management, 26: 3231-3249. https://doi.org/10.1007/s11269-012-0069-2

Kim S., Shiri J., Singh V.P., Kisi O., Landeras G., 2015. Predicting daily pan evaporation by soft computing models with limited climatic data. Hydrological Sciences Journal, 60(6): 1120-1136. https://doi.org/10.1080/02626667.2014.945937

Kisi O., Keskyn E.M., Terzy Ö., Taylan D., 2005. Discussion of "Fuzzy logic model approaches to daily pan evaporation estimation in western Turkey". Hydrological Sciences Journal, 50(4): 727-730. https://doi.org/10.1623/hysj.2005.50.4.727

Konapala G., Mishra A. K., Wada Y., Mann M. E., 2020. Climate change will affect global water availability through compounding changes in seasonal precipitation and evaporation. Nature Communications, 11, art. no. 3044. https://doi.org/10.1038/s41467-020-16757-w

Kovoor G.M., Nandagiri L., 2007. Developing regression models for predicting pan evaporation from climatic data - A comparison of multiple least-squares, principal components, and partial least-squares approaches. Journal of Irrigation and Drainage Engineering, 133: 444-454. https://doi.org/10.1061/(ASCE)0733-9437(2007)133:5(444)

Kutner M., Nachtsheim C., Neter J., 2004. Applied Linear Regression Models, 4rd ed., McGraw Hill Irwin, pp. 495.

Majhi B., Naidu D., Mishra A.P., Satapathy S.C., 2020. Improved prediction of daily pan evaporation using Deep-LSTM model. Neural Computing and Applications, 32: 7823-7838. https://doi.org/10.1007/s00521-019-04127-7

Molina J.M., Martínez V., González-Real M.M., Baille A., 2006. A simulation model for predicting hourly pan evaporation from meteorological data. Journal of Hydrology, 318: 250–261. https://doi.org/10.1016/j.jhydrol.2005.06.016

O'brien R.M., 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity, 41: 673-690. https://doi.org/10.1007/s11135-006-9018-6

Pammar L., Deka P.C., 2015. Forecasting daily pan evaporation using hybrid model of wavelet transform and support vector machines. International Journal of Hydrology Science and Technology, 5: 274-294.

Penman H.L., 1948. Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 193: 120–145. https://doi.org/10.1098/rspa.1948.0037

Penman H.L., 1956. Evaporation: an introductory survey. Netherlands Journal of Agricultural Science, 4: 9–29. https://doi.org/10.4236/jss.2016.43010

Piri J., Amin S., Moghaddamnia A., Keshavarz A., Han D., Remesan R., 2009. Daily pan evaporation modeling in a hot and dry climate. Journal of Hydrologic Engineering, 14: 803-811. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000056

Rahimikhoob A., 2009. Estimating daily pan evaporation using artificial neural network in a semi-arid environment. Theoretical and Applied Climatology, 98: 101-105. https://doi.org/10.1007/s00704-008-0096-3

Shirsath P.B., Singh A.K., 2010. A Comparative Study of Daily Pan Evaporation Estimation Using ANN, Regression and Climate Based Models. Water Resources Management, 24: 1571-1581. https://doi.org/10.1007/s11269-009-9514-2

Seifi A., Soroush F., 2020. Pan evaporation estimation and derivation of explicit optimized equations by novel hybrid meta-heuristic ANN based methods in different climates of Iran. Computers and Electronics in Agriculture, 173, art. no. 105418. https://doi.org/10.1016/j.compag.2020.105418

Shiri J., Marti P., Singh V.P., 2014. Evaluation of gene expression programming approaches for estimating daily evaporation through spatial and temporal data scanning. Hydrological Processes, 28(3): 1215-1225. https://doi.org/10.1002/hyp.9669

Strobl C., Malley J., Tutz G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods, 14(4): 323-348. https://doi.org/10.1037/a0016973

Tabari H., Marofi S., Sabziparvar A.-A., 2010. Estimation of daily pan evaporation using artificial neural network and multivariate non-linear regression. Irrigation Science, 28: 399-406. https://doi.org/10.1007/s00271-009-0201-0

Valiantzas J.D., 2006. Simplified versions for the Penman evaporation equation using routine weather data. Journal of Hydrology, 331: 690–702. https://doi.org/10.1016/j.jhydrol.2006.06.012

Vatcheva K.P, Lee M, McCormick J.B, Rahbar M.H., 2016. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiology, 6(2): 227-246. https://doi.org/10.4172/2161-1165.1000227

Wang L., Niu Z., Kisi O., Li C., Yu D., 2017. Pan evaporation modeling using four different heuristic approaches. Computers and Electronics in Agriculture, 140: 203-213. https://doi.org/10.1016/j.compag.2017.05.036

Wang H., Yan H., Zeng W., Lei G., Ao C., Zha Y., 2020. A novel nonlinear Arps decline model with salp swarm algorithm for predicting pan evaporation in the arid and semi-arid regions of China. Journal of Hydrology, 582, art. no. 124545. https://doi.org/10.1016/j.jhydrol.2020.124545

Wright M.N., Wager S., Probst P., Package 'ranger': 2020. A Fast Implementation of Random Forests,. Available online: https://github.com/imbs-hl/ranger (accessed on 1/5/2020)

Xu C.-Y., Singh V.P., 1998. Dependence of evaporation on meteorological variables at different timescales and intercomparison of estimation methods. Hydrological Processes, 12: 429–442. https://doi.org/10.1002/(sici)1099-1085(19980315)12:3<429::aid-hyp581>3.0.co,2-a