Ecological Informatics 81 (2024) 102638

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Predicting the fundamental fluxes of an eddy-covariance station using machine learning methods

David Garcia-Rodriguez [a,*], Pablo Catret Ruber [b], Domingo J. Iglesias Fuente [c], Juan José Martínez Durá [a], Ernesto López Baeza [d,e], Antonio Garcia Celda [a]

[a] University Research Institute on Robotics and Information and Communication Technologies (IRTIC), Universitat de Valencia, Paterna 46980, Spain
[b] Faculty of Mathematical Sciences, Universitat de Valencia, Burjassot 46100, Spain
[c] Valencian Institute of Agricultural Research (IVIA), Moncada, 46113, Spain
[d] Faculty of Physics, Universitat de Valencia, Burjassot 46100, Spain
[e] Albavalor, Science Park Universitat de Valencia, Paterna 46980, Spain

ARTICLE INFO

Keywords:
Eddy covariance fluxes
Machine learning
Deep learning
Agricultural crops
Surface energy balance

ABSTRACT

Monitoring tools are needed to maximise living systems' ability to mitigate emissions and adapt to changing environmental conditions. Therefore, it is important to be able to predict the fundamental fluxes in crops, in this case vineyards, such as sensible heat flux (*H*), latent heat flux (*LE*) and carbon dioxide flux (CO_2), in order to know their capacity to adapt to the environmental effects of climate change. In this study, Linear Regression (LR), Elastic Net (EN) regression, K-Nearest Neighbours (KNN), Gaussian-Process (GP), Decision Tree (TREE) Regression, Random Forest (RF) Regression, XGBoost (XGB) Regression, Support Vector Regression (SVR) and Multi-layer Perceptron (MLP) models have been applied to predict fundamental fluxes of an eddy-covariance station from conventional meteorological parameters. These models reproduced well the estimations of three output parameters from the eddy-covariance station. The performance of each predictive model was evaluated using Root-Mean-Squared Error (RMSE), Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE) and the coefficient of determination (R^2). The findings indicate that for the variable *H*, the GP model outperformed the SVR and all other models, achieving an R^2 value of 0.99. Conversely, the SVR demonstrated superior performance for the variables *LE* and CO_2, with R^2 values of 0.96 for both. In summary, these findings suggest that the three models proposed show a robust performance in the prediction of the studied fluxes, underlining their versatility and adaptability to the various environmental conditions of the vineyard.

1. Introduction

Forests and agroecosystems contribute to emissions reduction through their ability to capture carbon from the atmosphere and store it in various forms (Canadell and Raupach, 2008; Locatelli, 2016). Agricultural ecosystems are particularly promising from this perspective, as they provide a range of services, including food production, carbon sequestration, nutrient recycling, and climate regulation (Andrén and Kätterer, 2001; Iglesias et al., 2013). Additionally, in recent years, there is a growing adoption of technologies to maximise agroecosystem performance while contributing to sustainability through the optimal use of water, soil, and various resources, as well as increased carbon fixation.

In this scenario, there is a need for monitoring tools that maximise living systems' adaptation to the environmental conditions and their emissions mitigation capacity. Traditionally, the eddy covariance technique has been established as a technology that allows continuous and non-destructive measurement of carbon and water fluxes, being widely used with scientific, industrial, or agricultural purposes (Burba, 2013). Technically, it is based on the interpretation of the covariance between a turbulent wind (in the direction normal to the sur face) and the concentrations of a constituent (such as CO_2) as a flux to or from the surface of interest (i.e. Perez-Priego et al., 2018). Taken together, these technologies have contributed to a better understanding of fundamental processes in the biosphere, such as evapotranspiration (Ryu et al., 2012),

* Corresponding author.
E-mail addresses: david.garcia-rodriguez@uv.es (D. Garcia-Rodriguez), carupa@alumni.uv.es (P.C. Ruber), iglesias_dom@gva.es (D.J. Iglesias Fuente), juan.martinez-dura@uv.es (J.J.M. Durá), ernesto.lopez@uv.es (E.L. Baeza), toni.garcia@uv.es (A.G. Celda).

https://doi.org/10.1016/j.ecoinf.2024.102638
Received 25 January 2024; Received in revised form 6 May 2024; Accepted 7 May 2024
Available online 9 May 2024

and they are indeed understood as a main and very promising tool to enhance the capacity of plant systems to combat or adapt to climate change (Guevara-Escobar et al., 2021; Wiesner et al., 2022).

However, although the information provided by these infrastructures is very valuable, they are costly and resource intensive. Also, at the operational level, the mathematical treatment of the data is sometimes complex and requires very specific techniques and methodologies, as for example, gap-filling (Shi et al., 2022; Yao et al., 2021). Recent advances in machine learning have enabled many technical barriers to be overcome in this area, creating many opportunities and challenges that were unapproachable in the past (Huntingford et al., 2019; Tang and Li, 2023). Machine learning (ML) algorithms have advanced exponentially in recent years, leading to breakthroughs, and more recently aiding climate analysis applications (Citakoglu and Coşkun, 2022; Reichstein et al., 2019). Although there has been much discussion and debate in the scientific community about the effectiveness and suitability of machine/deep learning techniques to help improve our understanding of local and global environments, many studies have shown that they have superiority over other methods when upscaling eddy-covariance main fluxes (Shi et al., 2022; Xu et al., 2018; Zhu et al., 2023) to field or regional scales. Among them, ANNs (Artificial Neural Networks), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) algorithms have shown interesting and comparable performance, with RF providing very consistent results with slightly less bias (Kim et al., 2019; Mahabbati et al., 2021). Some attempts to evaluate ML methods for upscaling water fluxes from eddy covariance towers to the regional scale have been proposed (i.e., Campioli et al., 2016; Wang et al., 2016; Xu et al., 2018), but their application to these fields still requires further effort.

Overall, current ML techniques have demonstrated several strengths and limitations, but their future potential seems clear, especially when focused on weather and climate modelling (Bayram and Çıtakoğlu, 2023; Chantry et al., 2021). Although ML allows predictive and probability-based calculations, which are useful tools for evaluating eddy covariance data there is, at the moment, a lack of knowledge about generating flux data from environmental measurements. Previous work of this research group presented a methodology based on neural networks to derive surface net radiation from conventional meteorological parameters (Geraldo-Ferreira et al., 2011). Now, in this paper, data from an eddy covariance station have been combined with meteorological parameters using a ML-based approach to, for the first time (to the authors' knowledge), generate carbon and water fluxes from conventional, easy-to-measure, meteorological measurements.

The main objective of this research work is to apply and compare different ML models to predict sensible heat (*H*), latent heat (*LE*) and carbon dioxide (CO_2) fluxes in a semi-arid vineyard area, from conventional meteorological measurements, to produce results comparable to those of an eddy-covariance station in relation to these fluxes. The remainder of this article is organised as follows. Section 2, on materials and methods, briefly describes the datasets and prediction models. Section 3 shows the results based on the prediction models analysed and those finally proposed. Section 4 presents the discussion of these results. Concluding remarks are drawn in Section 5.

2. Materials and methods

This section is divided into four subsections: description of the dataset, proposed methodology, model comparison, performance evaluation measures and statistical analysis.

2.1. Site study and data

With a total area up to about 100 km^2, the Valencia Anchor Station (VAS) is located in La Plana de Utiel-Requena, approximately 80 km north-west of the city of Valencia (Fig. 1). Established by the University of Valencia in November 2001, the VAS serves as a significant site for the scientific validation of current low and medium spatial resolution remote sensing instruments onboard EO missions from ESA, EUMETSAT, Copernicus, NASA, etc. This area has been characterised from various perspectives, including soil types (Carbó et al., 2021), vegetation cover and land uses (Brown et al., 2019), soil moisture (Colliander et al., 2021), shortwave and longwave radiation (Bodas-Salcedo et al., 2003), etc. It consists primarily of vineyards and fruit trees (approximately 75%) together with other land covers including sclerophyllous vegetation, non-irrigated arable land, and coniferous forests. Within this area, a 1 × 1 km^2 area (Fig. 2) is the specific work zone where most of the field instruments are installed (FAPAR stations, soil moisture probes, eddy covariance station, radiation station, etc.) and a large number and types of mobile measurements are usually carried out.

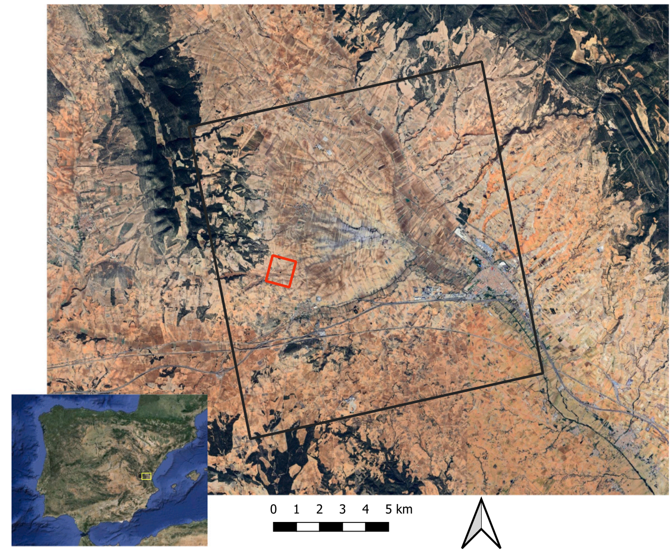

Fig. 1. Satellite image of the Valencia Anchor Station 10 × 10 km^2 validation area (black square) and the 1 × 1 km^2 validation area (red square), which includes most of the field instrumentation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The eddy covariance station (Fig. 2) is located 50 m from the VAS. It contains a CO_2 and H_2O Open-Path Gas Analyzer and 3D sonic anemometer (model EC150), controlled by a CR3000 datalogger. The datalogger performs measurements every 100 ms and provides flow output estimates every 30 min. The flux footprint is approximately of 400 m, taking into consideration the height (4 m) of the tower (Gash, 1986). In addition, the predominant wind direction is NW according to the monthly wind rose provided by the AEMET Spanish Meteorological Service for the period 1991–2020, from one of their stations located about 3.5 km from our site, also over a similar vine area. Furthermore, the VAS includes a complete 2 m high radiation mast containing a Kipp & Zonen albedometer (global radiation and surface reflected radiation), a pair of inverted Kipp & Zonen pyrgeometers (incident longwave atmospheric radiation and ground emitted radiation), and a Garmin GPS-controlled solar tracking system consisting of an Eppley pyrheliometer (direct solar radiation) and a shaded Kipp & Zonen pyranometer with a disk for diffuse radiation measurements.

Temperatures range from −15 °C in winter to 45 °C in summer, with an average annual temperature of 14 °C, while average annual precipitation is around 450 mm, with peaks in spring (March and May) and autumn (September and October), with some eventual peaks in ends June and January. However, at the boom of the vine phenological cycle (July and August), there is virtually no rainfall except on very rare occasions. The topography (800 m above sea level) is reasonably flat (variation of ±50 m in 10 × 10 km^2) in the study area (Carbó et al., 2021). The vine phenological cycle extends from April–May to

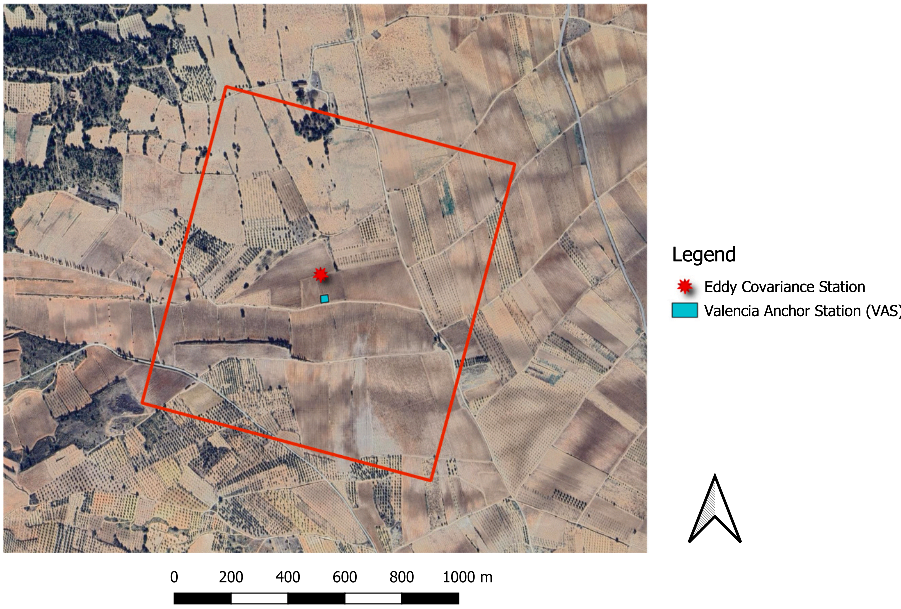

Fig. 2. Satellite image of the 1 × 1 km^2 area (red square above) with the VAS meteorological and eddy-covariance station. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

September–October, with the boom period occurring in July and August, while bare soil conditions persist for the rest of the year (Schwank et al., 2012).

Table 1 shows the different field parameters that have been used to predict *H*, *LE* and CO_2 between 22 June and 4 September 2022 covering most of the vine phenological cycle. The significant features of the dataset have been explored based on multiple parameters such as variable name and its abbreviation, symbol, unit and dataset source. All parameter data have been resampled to 30 min temporal resolution. In addition, antecedent air temperature (*AT*) and relative humidity (*RH*) composites over the previous 12, 24, 36, 48 and 72 h have also been considered.

Table 1
Key Insights from Dataset Features: A Quick Overview.

Variable	Variable abbreviation	Symbol	Unit	Source**
Soil Moisture	SM	θ	m^3/ m^3	Soil Moisture Probe
Pressure	PRES	p	hPa	AEMET*
Precipitation	PREC	P	L/m^2	AEMET*
Air Temperature	AT	T_a	Celsius Degrees	AEMET*
Relative Humidity	RH	RH	Dimensionless (%)	AEMET*
Wind Speed	WS	WS	m/s	AEMET*
Global Radiation	R_S (down)	R_S↓	W/m^2	Radiation Station
Reflected Radiation	R_S (up)	R_S↑	W/m^2	Radiation Station
Atmospheric Radiation	R_L (down)	R_L↓	W/m^2	Radiation Station
Emitted Radiation	R_L (up)	R_L↑	W/m^2	Radiation Station
Sensible Heat Flux	H	H	W/m^2	Eddy Covariance Station
Latent Heat Flux	LE	LE	W/m^2	Eddy Covariance Station
Carbon Dioxide Flux	CO2	CO_2	mg/(m^{2}s)	Eddy Covariance Station

* AEMET: Agencia Estatal de Meteorología.
** The data is semi-hourly

2.2. Descriptive statistical analysis and comparison between variables

The study utilised Pearson's and Spearman's correlation coefficients to analyse the relationships between target and features variables. Pearson's coefficient measured linear correlation, while Spearman's coefficient assessed monotonic relationships, providing valuable insights into the dataset's characteristics.

Table 2 presents a summary of instances before and after the filtering process, along with the percentage of valid data for each variable. The pre-filtered instances represent the total number of data points before any filtering procedures were applied. After filtering, post-filtered instances depict the remaining data points that passed the filtering criteria. The 'Significant data (%)' column indicates the proportion of instances deemed valid after filtering, providing insights into the effectiveness of the data cleaning process. As discussed in more detail in subsection 2.3, data filtering consists of manually removing outliers and smoothing them with a moving average.

Table 3 represents the statistical summaries of our target variable dataset, i.e. *H*, *LE* and CO_2, after filtering. The statistical representation of the different variables includes the count, mean, standard deviation, the minimum, the maximum, the 25th percentiles, 75th percentiles, C_V coefficient of variation, C_{SK} coefficient of skewness and C_K coefficient of kurtosis. Coefficients such as C_{SK} and C_K indicate skewness and kurtosis, respectively, with lower C_V values indicating lower variability relative to the mean. Variables *H* and *LE* exhibit slight right skewness, while CO_2 shows slight left skewness. Variations in kurtosis indicate differences in peakedness or flatness compared to a normal distribution, with all variables displaying relatively stable scatter around the mean.

Fig. 3a shows a graphical representation of the Pearson's correlations for the 13 parameters. On the one hand, it is observed that there is a

Table 2
Summary of Pre- and Post-filtering Instances and Percentage of Valid Data.

Variable	Pre-filtered instances	Post-filtered instances	Significant data (%)
H	3303	3300	99.9
LE	3303	3260	98.7
CO_2	3303	3259	98.7

Table 3
Statistical Summaries for the Target Variable Dataset: Insights and Analysis.

Variable	mean	std	min	25%	75%	max	C_V	C_{SK}	C_K
Filtered H	65.38	98.27	−76.44	−10.96	13.23	141.75	1.49	0.91	−0.36
Filtered LE	42.69	49.02	−15.75	1.61	21.52	77.38	1.15	1.09	0.68
Filtered CO_2	−0.079	0.14	−0.54	−0.18	−0.045	0.038	−1.74	−0.74	−0.18

C_V coefficient of variation, C_{SK} coefficient of skewness, C_K coefficient of kurtosis.

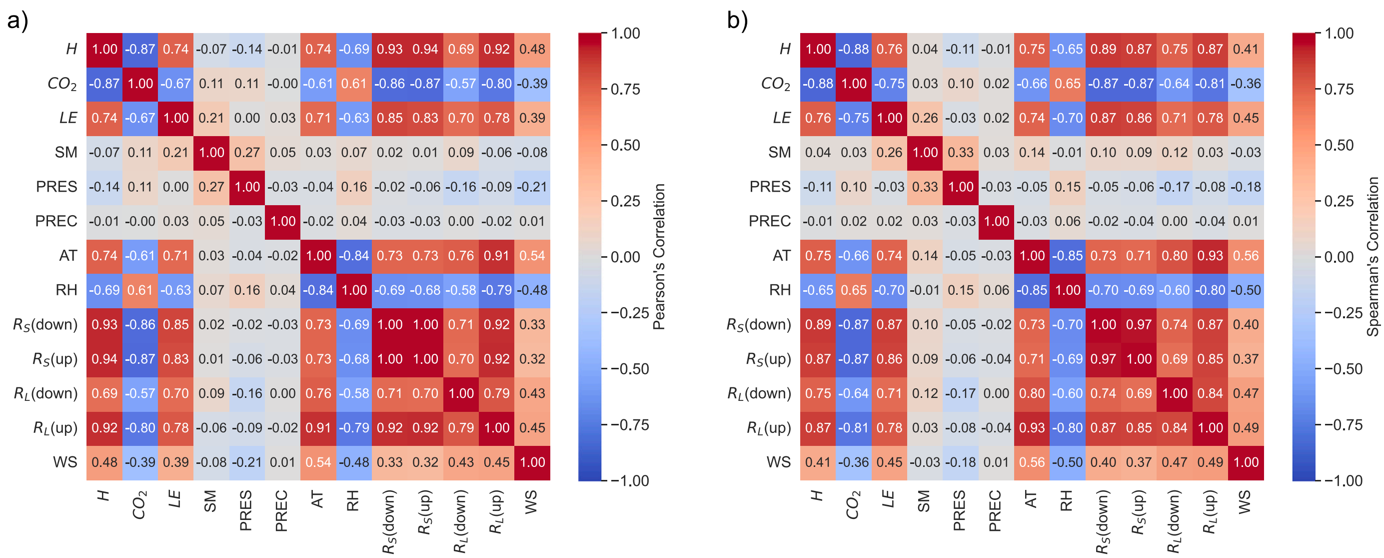

Fig. 3. a) Pearson's and b) Spearman's correlations between the variables.

significant positive correlation between the radiation variables, and with other variables such as *AT*, *H*, *LE* and, to a lesser extent, *WS*. On the other hand, a negative correlation is observed with the rest of the CO_2 and *RH* variables, which is consistent with the natural behaviour of these variables throughout the day. Finally, it is noted that *SM*, *PRES* and *PREC* show hardly any correlation with the rest of the variables. Spearman's correlations have also been calculated, and very similar results have been obtained, showing that there is no clear monotonic non-linear direct relationship between the different variables (Fig. 3b).

2.3. Proposed models

In this study, various regression models including Linear Regression (LR), Elastic Net (EN) regression, K-Nearest Neighbours (KNN), Gaussian-Process (GP), Decision Tree (TREE) Regression, Random Forest (RF) Regression, XGBoost (XGB) Regression, Support Vector Regression (SVR) and Multi-layer Perceptron (MLP) were employed for predicting the target variables using Python 3.8.19 with the following libraries (Pandas 2.0.3, NumPy 1.24.3, Scikit-learn 1.3.0, SciPy 1.10.1, XGBoost 1.7.6, SHAP 0.42.1, Matplotlib 3.7.2 and Seaborn 0.12.2). Before applying these models, a structured approach was adopted, by initially subjecting the dataset to pre-processing and cleaning procedures, dealing with missing values and manually removing outliers. (as indicated in Table 2, with minimal intervention). A 3-window moving average smoothing technique was then applied to mitigate small inherent instantaneous measurement errors in the systems. Subsequently, predictor variables were standardised using the *StandardScaler* implementation from Scikit-learn, adhering to the next formula:

$$z_i = \frac{x_i - \overline{x}}{\sigma_x} \tag{1}$$

where z_i represents the standardised variable, x_i denotes the observation of the variable before standardisation, $\overline{x}$ is the mean of the variable before standardisation, and σ_x is its standard deviation.

The dataset is then divided into training sets, which comprise 80% of the data, and a test set representing the remaining 20%, in order to maximise model training while maintaining a sufficient amount of data for robust model validation (Adjuik and Davis, 2022; Fernández-López et al., 2020). This division is done by random sampling to ensure that both subsets are representative of the overall dataset and minimise the risk of bias. Randomisation helps to preserve the distribution of key features and target variables across both training and test sets, enhancing the generalisability of the model. We leverage the scikit-learn function *GridSearchCV* to meticulously fine-tune the hyperparameters of each model within the confines of the training set, aiming to enhance their predictive performance. A five-fold cross-validation is used and the negative mean square error is set as the target measure. Following this hyperparameter optimisation process, the performance metrics detailed in Section 2.5 are meticulously evaluated using the designated test set. Subsequently, based on the performance metrics obtained, we discern and select the most effective models for each target variable. For an illustrative overview of this model selection process, kindly refer to the visual representation provided in Fig. 4.

2.4. Model comparison

In this section, a concise discussion of the mathematical fundamentals of the predictive models used in the preliminary assessment is presented. LR assumes a linear relationship between variables and estimates coefficients using least squares. EN combines L1 and L2 penalties to deal with multi-collinearity and allows adjustable regularisation (Zou and Hastie, 2005). KNN is non-parametric and estimates the target variable based on the closest observations (Cover and Hart, 1967). GP probabilistically models complex relationships, incorporating prior beliefs and uncertainty (Rasmussen, 2004). TREE regression recursively partitions data based on feature values, providing interpretability (Jena and Dehuri, 2020). RF builds an ensemble of trees, increasing stability and providing insight into the importance of variables (Breiman, 2001).

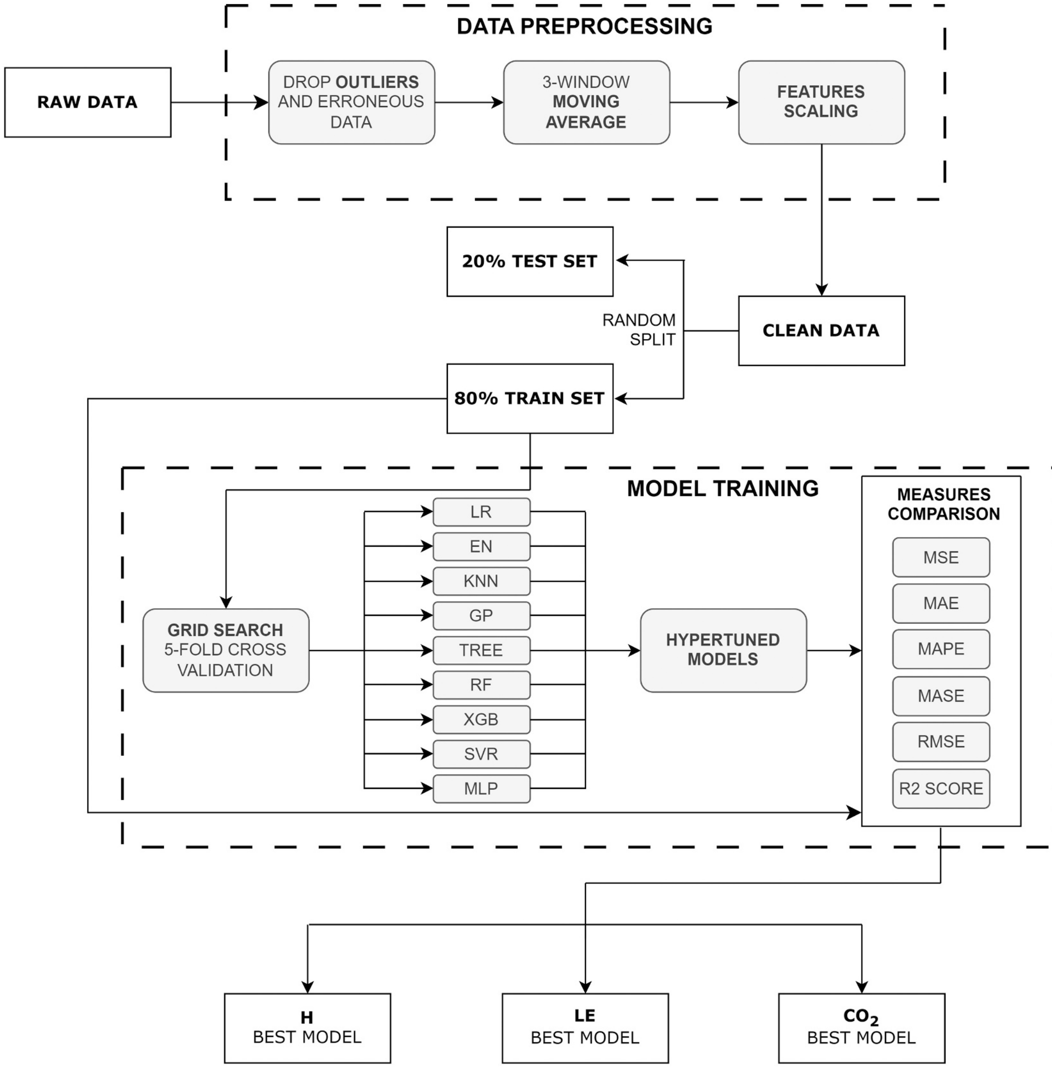

Fig. 4. Flowchart of the proposed models.

XGB optimises predictive modelling through gradient boosting and regularisation (Chen and Guestrin, 2016). SVR uses kernel functions to efficiently handle non-linear data by fitting the best hyperplane within a pre-defined error margin (Hsu et al., 2003). MLP is a feed-forward neural network that learns complex patterns using adjustable weights and activation functions (Kingma and Ba, 2014).

2.5. Performance evaluation measures

Various performance evaluation metrics, i.e., Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scale Error (MASE) and Coefficient of determination (R^2) have been used to measure the performance accuracy, and suitability of prediction models. In this section, the evaluation metrics used are properly recalled with their mathematical rationale (Nagelkerke, 1991).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{5}$$

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i - \bar{y}|} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

where $y \in \mathbb{R}^n$ is the vector of observations, $\hat{y} \in \mathbb{R}^n$ is the vector of predicted values, $\bar{y}$ is the mean of observed values and n is the length of the training set.

3. Results

In this section, the prediction results for H, LE and CO_2, based on the models considered, are discussed. The prediction results on the optimal prediction model are also analysed, together with the significance of the variables, in terms of relative preponderance.

3.1. Best model identification

Table 4 includes details such as the name of the prediction model, its hyperparameters, the parameters selected during the 5-fold cross-validation, and the hyperparameters adjusted according to the selection criteria to determine the most appropriate configuration for the prediction model.

3.2. Performance evaluation of *H*, *LE* and CO_2 prediction models

Table 5 shows the results of the nine regression models based on the six established performance evaluation metrics for the prediction of *H*, *LE* and CO_2. To assess the accuracy of these predictions, R^2 should be high with low MAE, MSE, RMSE, MAPE and MASE. Based on these validation metrics, the performance of each regression model was evaluated to determine the optimal prediction model among them. In this study, the three variables frequently take values close to zero, thus the MAPE result is deceptive and uninformative. Table 5 demonstrates that the MAPE measures are excessively high despite the model having very low error.

In the case of *H*, an R^2 value of 0.99 was achieved for GP and SVR models, with the GP model exhibiting higher performance in terms of MAE, MSE, RMSE and MASE values. Similarly, for *LE*, an R^2 value of 0.96 was obtained for SVR and KNN models, with the SVR model exhibiting higher performance in terms of MAE, MSE, RMSE and MAPE values. Concerning CO_2, the SVR and KNN models yielded the highest R^2 value of 0.96, with the SVR model demonstrating highest overall performance.

In previous statements, GP was considered as the best model for *H* and SVR was considered as the best model for *LE* and CO_2 since it had the least MAE, MSE, RMSE, MAPE and MASE and highest R^2 values. Although the MAE, MSE, RMSE, MAPE and MASE error criteria indicated the correctness of the variables, they do not offer information about the models' distribution. For the comparison of the models, the Taylor diagram (Fig. 5), the violin plot (Fig. 6) and the box error plot (Fig. 7) were used.

The standard deviation and correlation of each model compared to the observed data are shown in each Taylor diagram. From these diagrams, several conclusions can be drawn regarding the performance of

Table 5
Performance evaluation of *H*, *LE* and CO_2 prediction models.

Variable	Model	MAE	MSE	RMSE	MAPE	MASE	R^2
H	XGB	8.58	175.23	13.24	71.82	0.083	0.98
	RF	10.32	248.37	15.76	75.97	0.10	0.97
	LR	20.22	729.70	27.01	220.13	0.20	0.92
	SVR	6.60	136.71	11.69	**35.21**	0.064	**0.99**
	GP	**5.72**	**109.05**	**10.44**	35.23	**0.055**	**0.99**
	EN	19.96	733.36	27.08	212.91	0.19	0.92
	KNN	7.80	178.54	13.36	47.39	0.075	0.98
	MLP	9.91	226.24	15.04	91.42	0.10	0.98
	Tree	13.17	422.45	20.55	110.57	0.13	0.95
LE	XGB	6.58	113.17	10.64	210.65	**0.098**	0.95
	RF	7.42	145.51	12.06	187.45	0.12	0.94
	LR	12.52	341.59	18.48	901.78	0.14	0.85
	SVR	**5.32**	**93.27**	**9.66**	**93.31**	0.10	**0.96**
	GP	5.58	111.61	10.56	163.87	0.23	0.95
	EN	13.63	417.97	20.44	814.42	0.25	0.82
	KNN	5.40	93.91	9.69	99.79	0.19	**0.96**
	MLP	6.99	112.49	10.61	251.32	0.13	0.95
	Tree	10.32	276.78	16.64	408.36	0.10	0.88
CO_2	XGB	0.019	0.00088	0.030	142.74	0.12	0.95
	RF	0.025	0.0013	0.036	188.33	0.16	0.93
	LR	0.047	0.0039	0.062	299.12	0.30	0.79
	SVR	**0.015**	**0.00073**	**0.027**	69.89	**0.096**	**0.96**
	GP	**0.015**	0.00094	0.031	**56.22**	0.098	0.95
	EN	0.092	0.013	0.11	381.45	0.59	0.31
	KNN	0.016	0.00074	**0.027**	68.05	0.10	**0.96**
	MLP	0.030	0.0017	0.041	163.20	0.19	0.91
	Tree	0.034	0.0028	0.053	180.10	0.22	0.84

Table 4
Hyperparameter Grid for Each Model.

Model	Hyperparameter	Parameter Selection	BP* H	BP* LE	BP* CO2
EN	alpha	[0.1, 0.5, 1.0]	0.1	0.1	0.1
	l1_ratio	[0.1, 0.5, 0.9]	0.9	0.9	0.1
KNN	neighbours	[3, 5, 7, 9]	3	3	3
	weights	[uniform, distance]	distance	distance	distance
	p	[1,2]	1	1	1
GP	alpha	[10^{-10}, 10^{-5}, 10^{-2}, 10^{-1}]	0.1	0.1	10^{-5}
TREE	max_depth	[5, 10, 15]	10	10	10
	min_samples_split	[2, 5, 10]	10	5	5
	min_samples_leaf	[1, 2, 4]	4	2	2
	max_features	[None, sqrt, log2]	None	None	None
RF	n_estimators	[100,200,300]	200	100	300
	max_depth	[5, 10, 15]	15	15	15
	min_samples_split	[2, 5, 10]	2	2	2
	min_samples_leaf	[1, 2, 4]	1	1	1
	max_features	[1.0, sqrt, log2]	log2	1	sqrt
XGB	n_estimators	[100, 200, 300]	300	300	300
	learning_rate	[0.01, 0.05, 0.1, 0.2]	0.1	0.2	0.1
	max_depth	[3, 5, 7]	7	5	7
	subsample	[0.8, 1.0]	0.8	0.8	0.8
	colsample_bytree	[0.8, 1.0]	1	1	1
	reg_alpha	[0, 0.1, 0.5, 1]	1	0.5	0.1
	gamma	[0, 0.1, 0.5, 1]	1	1	0
	min_child_weight	[1, 3, 5]	3	1	5
SVR	C	[0.1, 1, 10, 100]	100	100	1
	gamma	[0.001, 0.01, 0.1, 1]	0.5	0.5	0.01
	epsilon	[0.01, 0.1, 0.2, 0.5]	0.1	0.1	0.1
MLP	hidden_layer_sizes	[(50), (100), (50, 50), (100, 50)]	(100, 50)	(100, 50)	(100, 50)
	activation	[relu, tanh, logistic]	relu	relu	relu
	alpha	[0.0001, 0.001, 0.01]	0.001	0.001	0.01
	learning_rate	[constant, invscaling, adaptive]	constant	invscaling	constant
	max_iter	[200, 300, 400]	400	400	300

* BP denotes Best Parameter.

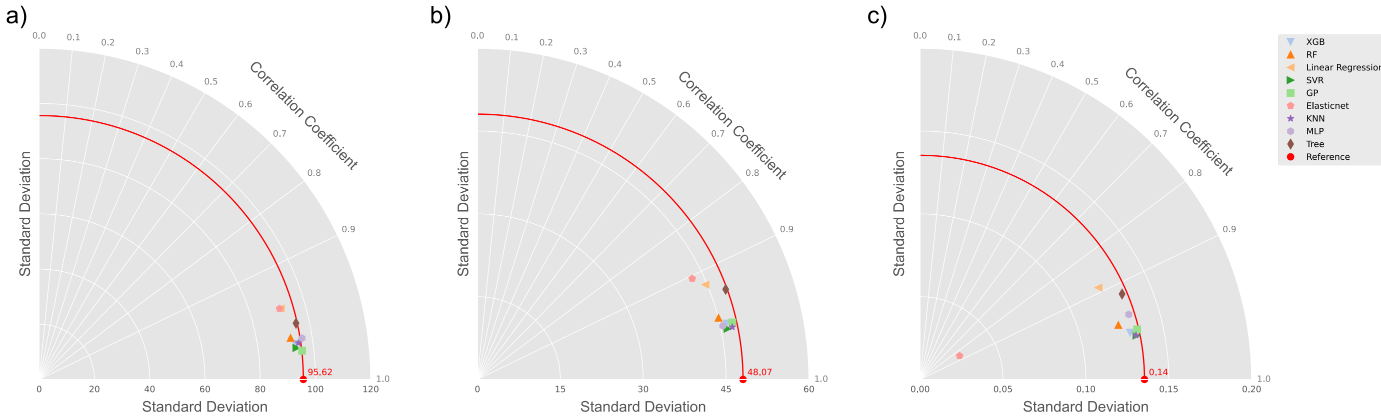

Fig. 5. Taylor diagrams for a) *H*, b) *LE* and c) CO_2.

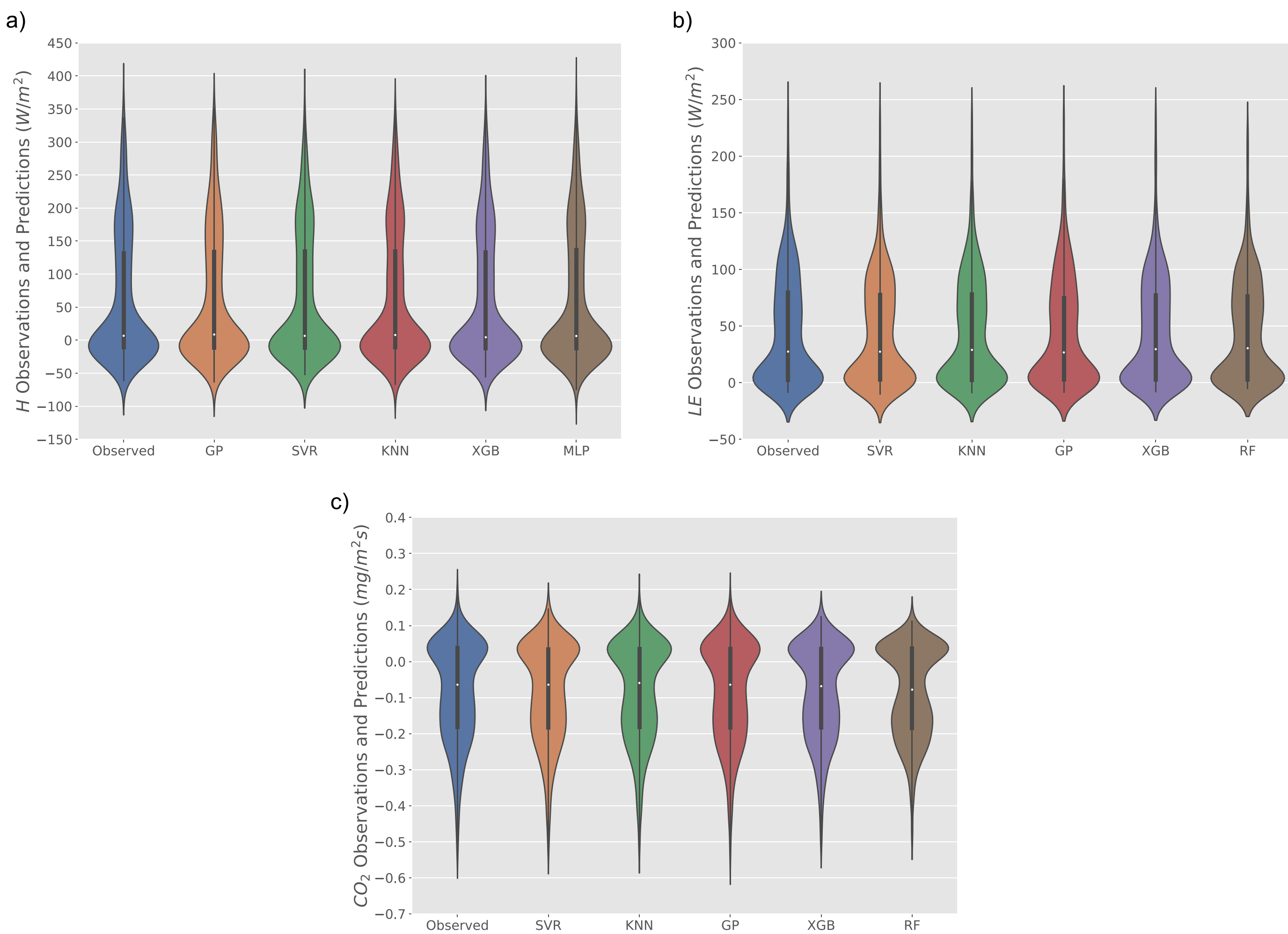

Fig. 6. Violin plots for a) *H*, b) *LE* and c) CO_2.

the models. For the variable *H*, it is generally observed that the models exhibit a relatively low standard deviation almost identical to the observed data, indicating a good fit in terms of the variability of the variable of interest. Moreover, the correlation between the models and the observed data remains relatively consistent across different models, suggesting their strong predictive capability. Regarding the variable *LE*, higher correlations are observed for the GP, KNN, SVR, MLP, XGB, and RF models compared to the Tree, LR, and EN models. Additionally, there is a relatively lower standard deviation for these models, except for Tree, which exhibits the best standard deviation but at the expense of a lower correlation. Finally, for variable CO_2, a significant variation in correlation is observed among the SVR, KNN, XGB, and GP models compared to the others, especially with EN, indicating differences in their ability to reproduce the variability of the variable of interest.

The conformity between estimation data and the observed data was assessed using the violin plot. Statistical comparisons between models were also performed using the violin plot.

Fig. 6 presents a violin plot considering the observed values and the

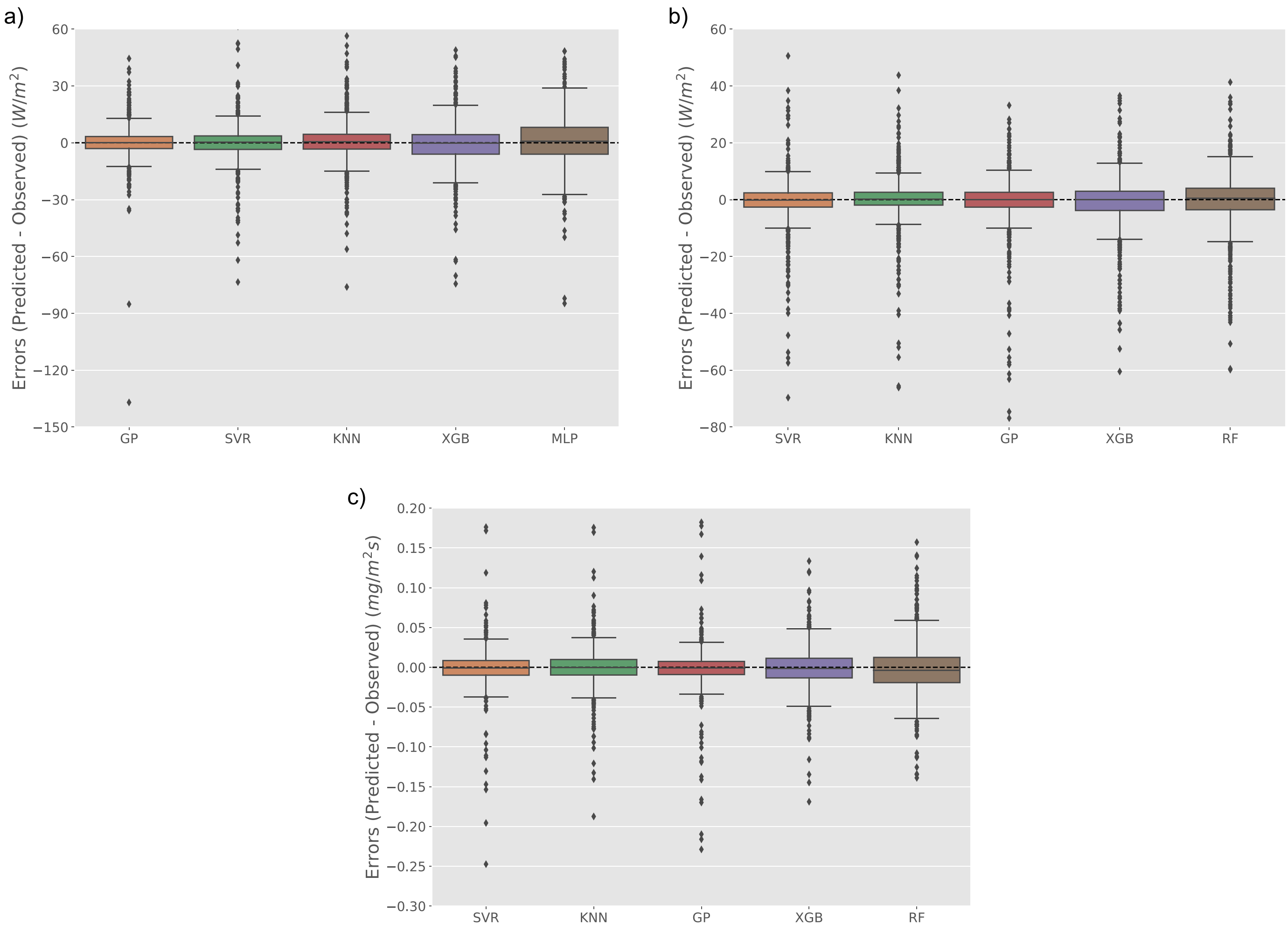

Fig. 7. Error box plot diagrams for a) *H*, b) *LE* and c) *CO2*.

top five models for each target variable. For variable *H* (Fig. 6a), the first five models are remarkable similar to each other; however, subtle differences can be distinguished in the error box plot diagrams presented in Fig. 7a, especially in the interquartile range and outliers, with the GP model appearing the most favourable. Similar patterns are observed with the *LE* and CO_2 variables, where differentiating between models is challenging, as exemplified by the almost identical behaviour of SVR and KNN. In contrast, in the case of MLP, there are small differences, such as a larger interquartile range.

Finally, statistical significance comparisons between the different ML models were made. Firstly, the Kruskal–Wallis test was employed to see whether the median of the observations and the estimates were identical (Kruskal and Wallis, 1952). The test works by ranking the data from all groups together and calculating a test statistic based on the rankings. This statistic is compared with a critical value of the chi-squared distribution to determine if there are significant differences between the groups. In the results provided, if the Kruskal statistic is high and the associated *p*-value is low (usually <0.05), it indicates that at least one of the groups is significantly different from the others. Conversely, if the p-value is high, it suggests that there are no significant differences between the groups, and the null hypothesis of equal population medians cannot be rejected.

In Table 6, the results of the Kruskal-Wallis (KW) test reveal insights for each variable. For variable *H*, all models exhibited no significant differences, as indicated by *p*-values ranging from 0.6179 to 0.9850. Similarly, for variable *LE*, most models showed no significant differences except for the Elastic Net (EN) model, which demonstrated borderline

Table 6
KW test results.

Variable	Model	Kruskal Statistic	Kruskal *P*-value	H_0*
	XGB	0.04698	0.8284	Accept
H	RF	0.00035	0.9850	Accept
	LR	0.2488	0.6179	Accept
	SVR	0.01078	0.9173	Accept
	GP	0.001464	0.9695	Accept
	EN	0.2333	0.6291	Accept
	KNN	0.09426	0.7588	Accept
	MLP	0.00208	0.9636	Accept
	Tree	0.00378	0.9509	Accept
	XGB	0.2698	0.6034	Accept
LE	RF	0.6252	0.4291	Accept
	LR	0.2563	0.6127	Accept
	SVR	0.1851	0.6671	Accept
	GP	0.1547	0.6941	Accept
	EN	3.7681	0.0522	Accept
	KNN	0.0638	0.8005	Accept
	MLP	1.4098	0.2351	Accept
	Tree	0.2040	0.6515	Accept
	XGB	0.0054	0.9413	Accept
CO2	RF	0.1102	0.7399	Accept
	LR	0.7735	0.3791	Accept
	SVR	0.0129	0.9096	Accept
	GP	0.1513	0.6973	Accept
	EN	0.7408	0.3894	Accept
	KNN	0.0315	0.8592	Accept
	MLP	0.2797	0.5969	Accept
	Tree	0.0001	0.9918	Accept

* H_0: Predicted median and measurements median are equal.

significance with a p-value of 0.0522. This suggests a potential difference in performance compared to other models for variable LE. Regarding the variable CO_2, again, no significant differences were observed across the models, with p-values ranging from 0.3791 to 0.9918, well above the threshold for significance ($p > 0.05$). Thus, there are no substantial deviations in model performance across the tested models for this variable.

3.3. H, LE and CO_2 prediction model and relative significance of the characteristic field variables

3.3.1. Sensible heat flux (H)

For the experimental analysis, *H* has been taken as one of the regression targets. As previously indicated, the data from 22/06/2022 to 04/09/2022, of the dataset from Table 1, have been used for training and testing purposes with a ratio of 80% and 20%, respectively.

The regression hyperparameter tuning has been performed for all the nine prediction models (XGB, RF, SVR, GP, EN, KNN, MLP, Tree and LR), and based on that, the best regression parameter has been used for training the model.

As illustrated in Fig. 8, the x-axis denotes the sequence of observed values, and the y-axis represents the differences between the predicted and observed values. As can be seen, there is a constant trend in the distribution of differences between predicted and observed values as the observed values increase on the X-axis. Particularly, the values on the Y-axis remain consistently close to 0, both above and below the reference line, indicating relative stability in the discrepancies between predictions and actual values across the range of observed values.

To evaluate the importance of predictor variables, we employed the Shapley Additive Explanation (SHAP) method, which integrated six explanatory models using game theory (Shapley, 1953). By calculating the weighted average marginal contribution (i.e., absolute Shapley values) of each modelling feature across all combinations, SHAP assigns importance value to each predictor variable in the ML model (Garcia and Aznarte, 2020). It is worth noting that a positive SHAP value indicates an increase in *H* relative to a baseline value, while a negative value implies a decrease in predicted *H* (Lundberg et al., 2020). In Fig. 9, the results showed that R_S (down) and R_S (up) are the most influential, followed by other significant variables previously established, such as

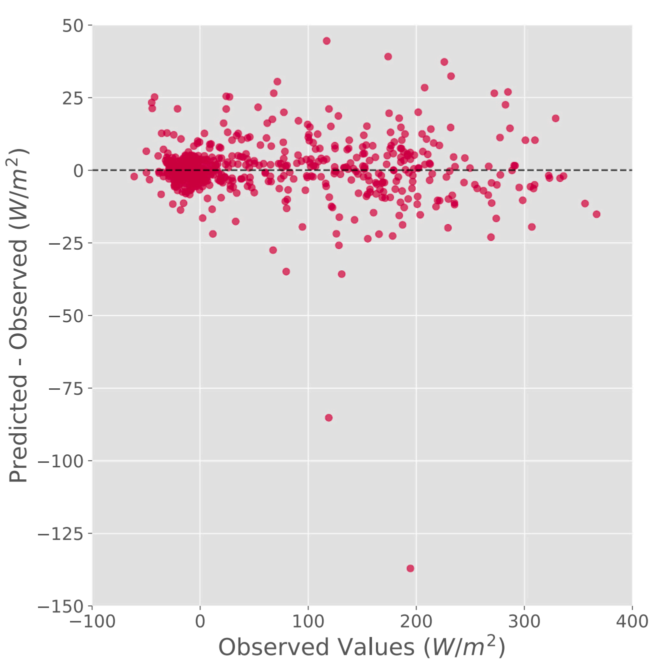

Fig. 8. Scatterplot between observed and predicted-observed values for *H* using GP.

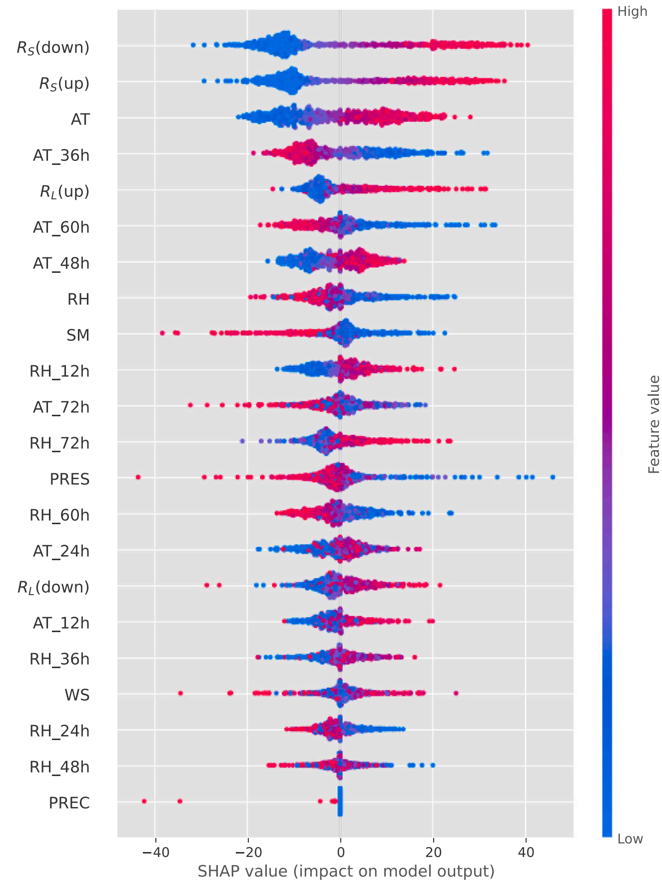

Fig. 9. SHAP summary plot of the GP model for *H*.

AT_36h and *AT*, which is logical since sensible heat, convective flux between the surface and the atmosphere, depends on those radiative components of the surface radiation balance and followed by the antecedent atmospheric temperature. As expected, this convective flux between the surface and the atmosphere depends on the temperature difference between their respective temperatures. As for R_L *(up)*, it is observed that high values exert a positive impact on the model, similar to R_S *(down)* and R_S *(up)*. However, for R_L *(down)*, it is not considered significant, as shown by the absence of a substantial impact. Being summer, the surface-emitted longwave flux seems to be more significant than the atmospheric one due to the high surface temperatures reached during the day.

3.3.2. Latent heat flux (LE)

Consistent with subsection 3.3.1, five out of nine models (XGB, SVR, GP, KNN and MLP) also demonstrate a high degree of accuracy in replicating or predicting ($R^2 \geq 0.95$). In terms of overall performance (high R^2 and low MAE, MSE, RMSE, MAPE and MASE), the SVR model outperforms all other regression models (see Table 5).

In Fig. 10, a consistent overall trend around the reference line is observed, with some high outliers when observed values are low. However, as observed values increase, there is a slight tendency to underestimate, indicated by numerous negative outliers. Ultimately, the model demonstrates reasonable accuracy for low values, yet tends to underestimate slightly as these values increase.

In this scenario, as far as the latent heat flux is concerned, R_S *(down)* and *AT* remain the most critical variables, followed by R_S*(up)* and the longwave variables R_L *(down)* and R_L *(up)*, as can be seen in Fig. 11. Latent heat flux is generally referred to as evapotranspiration, which describes the total evaporation from land surface and transpiration by plants. In particular, the two longwave variables are of substantial

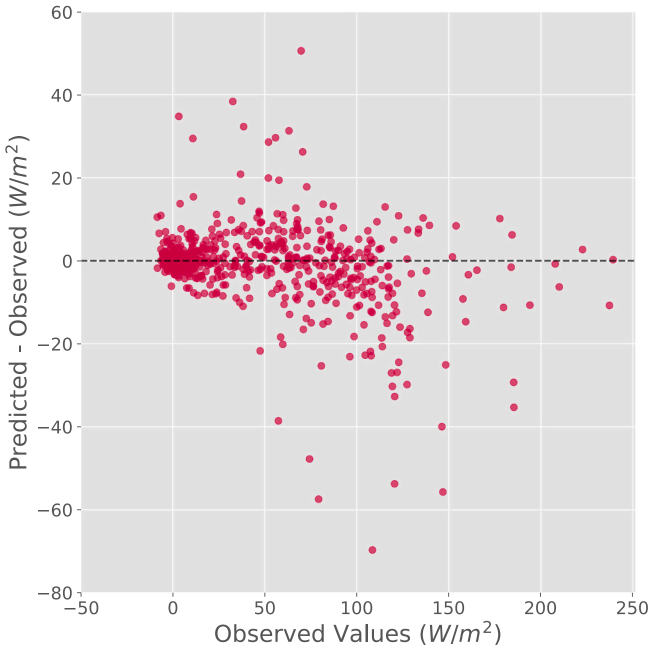

Fig. 10. Scatterplot between observed and predicted-observed values for *LE* using SVR.

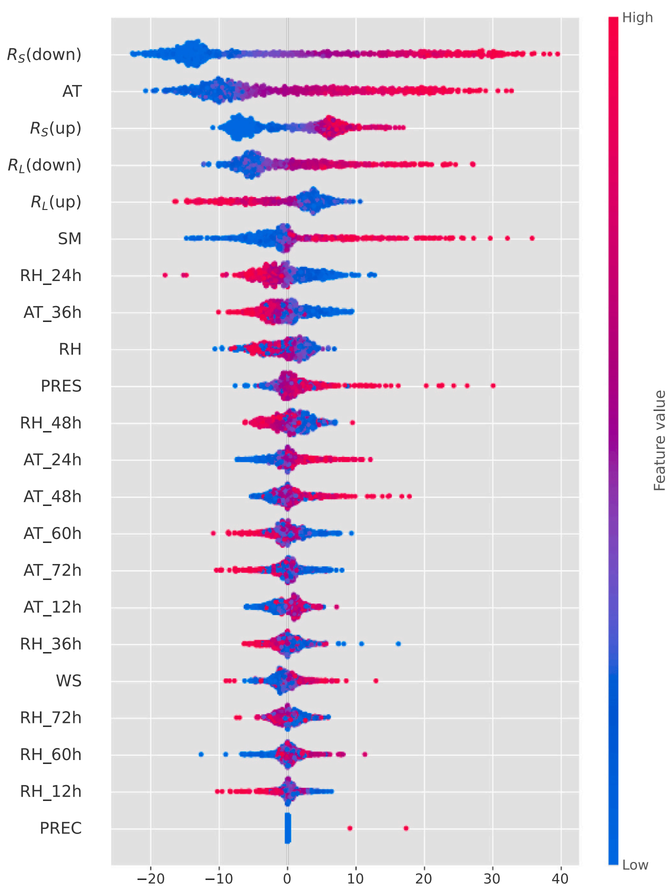

Fig. 11. Summary SHAP plot of the SVR model for *LE*.

importance, in this case including atmospheric properties such as relative humidity, wind speed and soil moisture that most significantly affect R_L *(down)*. In addition, high R_S *(down)*, R_S *(up)*, R_L *(down)* and *AT* values may induce a larger variation in *LE* and both of their SHAP values have long right tails. The SHAP values of *RH* exhibited a relatively balanced pattern at both high and low values.

3.3.3. Carbon dioxide flux (CO_2)

Also consistent with subsection 3.3.1, four out of the nine models (XGB, SVR, GP and KNN) demonstrate a high degree of accuracy in replicating or predicting CO_2 as well ($R^2 \geq 0.95$). In terms of overall performance (high R^2 and low MAE, MSE, RMSE, MAPE and MASE), the SVR model outperforms all other regression models (see Table 5).

A consistent trend around the reference line is observed in Fig. 12, as observed values decrease, with some minor outliers when observed values are near 0 but of little significance.

Like the previous analyses, in the case of CO_2 flux, when studying the importance and significance of the variables in the optimal models, it is observed that those with the greatest weight are the global radiation flux, that is, R_S *(down)*, followed by reflected radiation R_S *(up)*, as can be seen in Fig. 13. In addition, the *AT_60h* also shows a relative significance compared to the other variables. As expected, R_S *(down)* and R_S *(up)*, together with air temperature are pivotal drivers of the CO_2 flux. Furthermore, low values of R_S *(down)* and R_S *(up)* positively influence the model, in contrast to *AT_36h*, where the opposite is true (Fig. 13).

Finally, radiative fluxes shown in Table 1 can be combined to provide the surface radiation balance, surface net radiation (R_n), as indicated in eq. 8

$$R_n = R_S\downarrow - R_S\uparrow + R_L\downarrow - R_L\uparrow \tag{8}$$

Soil heat flux, H_C, can then be determined by difference based on the surface energy balance, as indicated in Eq. 9, valid in absence of advective fluxes

$$H_c = R_n - H - LE \tag{9}$$

Fig. 14 shows now the evolution of these surface energy balance fluxes, as an example, during the week of 18 to 24 July 2022, in the significant central stage of the phenological cycle of the vine.

Evapotranspiration (*ET*) is estimated from the latent heat flux (*LE*) component of the eddy covariance measurements. *LE* represents the energy consumed or released during the phase change of water from liquid to vapor during evaporation or transpiration processes. It can be

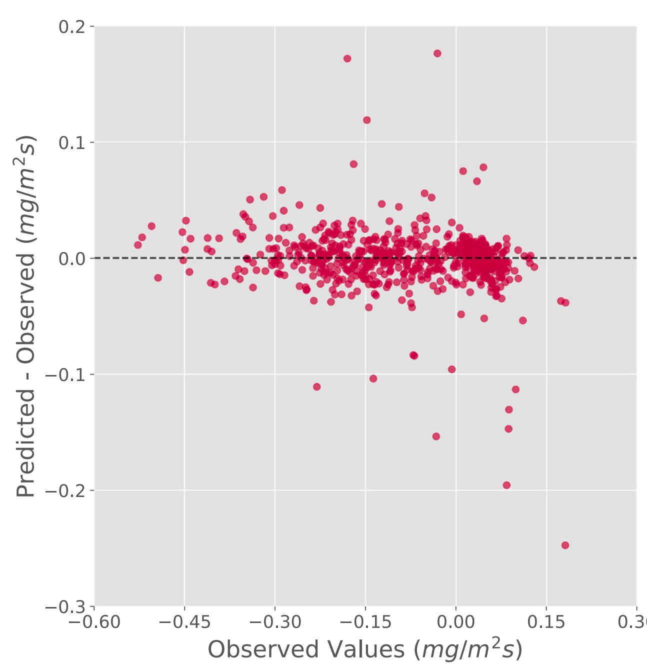

Fig. 12. Scatterplot between observed and predicted-observed values for CO_2 using SVR.

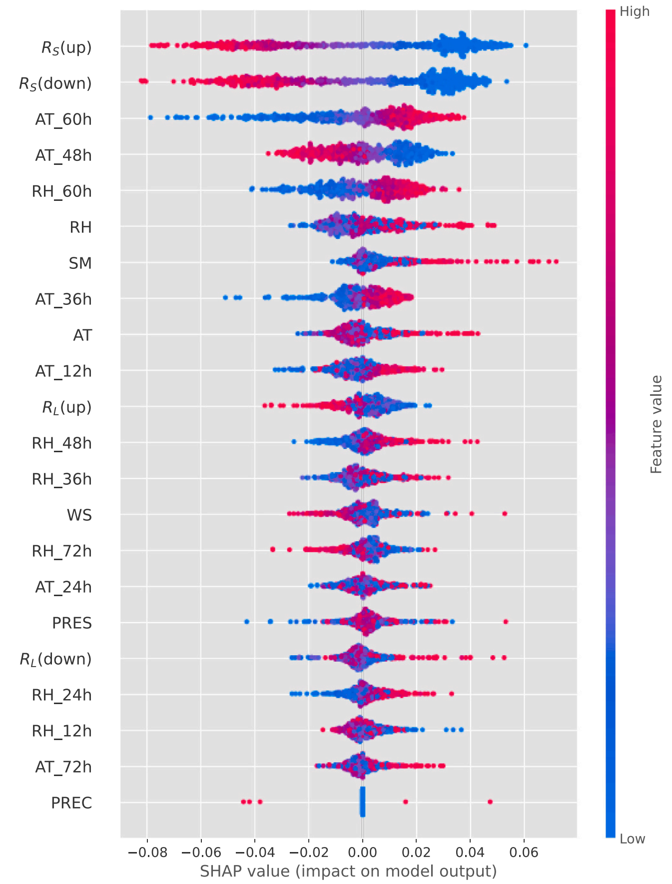

Fig. 13. Summary SHAP plot of the SVR model for CO_2.

calculated using the eddy covariance equation:

$$LE = -\rho \bullet \lambda \bullet \overline{w'q'} \tag{10}$$

where LE is the latent heat flux (W/m^2), ρ is the air density (kg/m^3), λ is the latent heat of vaporization (J/kg) and $\overline{w'q'}$ is the covariance of vertical wind speed and water vapor concentration (m/s · kg/kg).

By knowing the latent heat of vaporization (λ) (Harrison, 1963), and the latent heat flux measured by the eddy covariance system, the rate of evapotranspiration can be calculated according to Allen et al. (1998):

$$ET = \frac{LE}{\lambda} = \frac{LE}{2.501 - 2.361 \cdot 10^{-3} T_a} \tag{11}$$

By way of example, and for the same representative period, the CO_2 flux and evapotranspiration from the eddy-covariance station measurements can be plotted together (Fig. 15).

Fig. 15 illustrates the evolution of CO_2 and water fluxes over a typical week in July in our study area. As expected, both fluxes show an opposite trend, consistent with the theoretical biological behaviour of vines. Interestingly, the combination of these two fluxes, CO_2 and ET, will be used in future to also obtain water use efficiency in vines. The solid lines show the trend of the data from the eddy covariance station (CO_2, direct measurement; ET, transformation according to Eq. (11)), while the dashed lines indicate the output of the model. As shown in the graph, the degree of concordance is very high and it is noted that the daily trend is regular throughout the week. On the one hand, the maximum ET peaks are reached in the central hours of the day and the minimum during the night (Egipto et al., 2023), coinciding with the particular stomatal dynamics of the vegetation at that time of the year, a period characterised by moderate or elevated daytime temperatures and a medium or high degree of water deficit. On the other hand, net CO_2 exchange shows the inverse pattern, also corresponding to this time of the year in semiarid agroecosystems. Similar results have been obtained in other studies and in different species (i.e. Falge et al., 2002).

Our research also suggests that ET (estimated from latent heat flux eddy-covariance measurements, as indicated earlier) can be reproduced from meteorological measurements using ML methods. It is worth noting that, complementary, other authors have been able to derive ET from their conventional meteorological station measurements, also using ML or deep learning methods (Zouzou and Citakoglu, 2023).

4. Discussion

This research involves the application of ML models to estimate fluxes of carbon dioxide (CO_2), sensible heat (H), and latent heat (LE) over a vineyard. By utilising meteorological parameters as inputs, the study aimed to assess the feasibility of deriving these fluxes from conventional weather data. The findings of this study may serve as a precursor for more extensive investigations in other crop types, providing a promising approach for estimating ecosystem fluxes across a range of agricultural contexts.

Statistical summaries offer valuable insights into the distribution and characteristics of the variables examined in this study. By analysing the mean, standard deviation, minimum, maximum, quartiles, coefficient of variation (C_V), coefficient of skewness (C_{SK}), and coefficient of kurtosis (C_K), an overall understanding of the variability, skewness, and shape of the data distribution for each variable has been obtained. The coefficient of skewness (C_{SK}) allowed to assess the degree of asymmetry in the distribution of the filtered variables. Both filtered variables H and LE exhibited slight right skewness, with C_{SK} coefficients of 0.91 and 1.09, respectively, indicating a tendency towards higher values. Conversely, the filtered variable CO_2 showed a slight left skewness ($C_{SK} = -0.74$), suggesting an asymmetry towards lower (negative) values. These observations provide valuable insights into the distributional

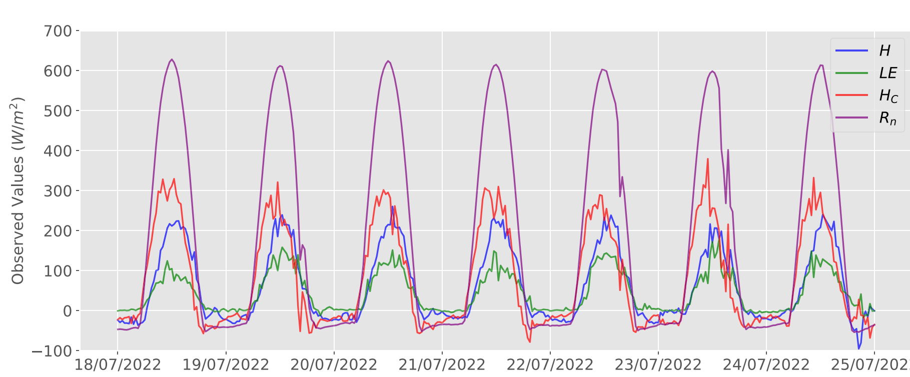

Fig. 14. Evolution of the surface energy balance fluxes along the week of 18 to 24 July 2022.

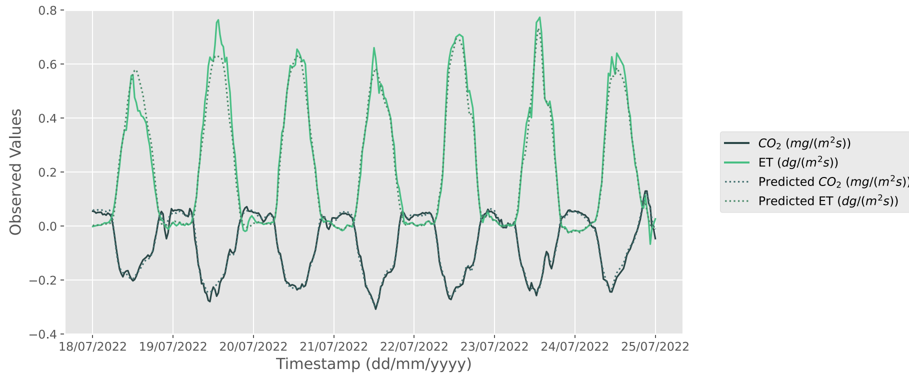

Fig. 15. Evolution of the CO_2 and ET fluxes along the week of 18 to 24 July 2022. *ET* is expressed in dg/(m^2·s) in order to make it comparable and coherent with the CO_2 flux.

characteristics of the variables and highlight the presence of subtle asymmetries in the data. Furthermore, the coefficient of kurtosis (C_K) provided information about the degree of peakedness or flatness of the distribution compared to a normal distribution. The filtered variable *H* showed a slightly flattened distribution ($C_K = -0.36$), while the filtered variable *LE* showed a slightly sharper distribution around the mean ($C_K = 0.68$). In contrast, the filtered variable CO_2 demonstrated a distribution closer in terms of flatness to normal ($C_K = -0.18$). These findings indicate variations in the shape of the distributions and underscore the importance of considering the kurtosis coefficient in the analysis of the data.

Variable selection represents a pivotal aspect in ML models (Cateni and Colla, 2016). Pearson's correlation played a crucial role in elucidating the relationships between the variables. For instance, R_s*(up)* exhibits a high correlation with R_s *(down)* via the albedo, which remains virtually constant (except for low sun angles). Nevertheless, this did not hinder the model from learning satisfactorily. However, it is worth noting that certain variables, such as *PREC*, could have been potentially excluded from the outset, given their limited significance in our study case. This opens up the possibility of attempting to construct a model with fewer input variables.

The comparison among the nine classifiers (i.e., LR, EN, KNN, GP, TREE, RF, XGB, SVR, and MLP) showed that models performed globally well with reliable accuracy (i.e., $R^2 > 0.79$ except for EN for CO_2) for the three variables with slight differences between them. The fine-tuning of hyperparameters contributed to these differences, although a balance was found between them.

Although all ML models work sufficiently well, as they learn from a specific dataset, a paradigm shift in new data could affect the performance of the models, potentially hindering their performance. For example, if vegetation is removed, or drastically increased, it could greatly affect the performance of the models. In addition, on the one hand, the absence of precipitation during the observation period may limit the generalisability of the results obtained to other meteorological conditions. Additionally, the methods employed are localised and may exhibit varying performance in broader contexts. On the other hand, if the data behave similarly to the cases found in the training set, the models will perform correctly. As for the different regression algorithms, each has advantages and disadvantages. One limitation of the linear regression and Elasticnet models is that they may perform poorly in cases with nonlinear relationships. For their part, SVR and GP are slow to train and scale greatly with increasing amounts of data. Decision trees tend to perform worse than other models. KNN can be sensitive to noisy data and outliers, and since it needs to store the data in memory, it can be computationally expensive with large amounts of data. And XGBoost and Random Forest can be prone to overfitting depending on the selection of their hyperparameters.

The outcomes of the Kruskal-Wallis (KW) test provide valuable insights into the comparative performance of various models across different variables. Notably, for the variable *H*, none of the models displayed significant differences, with all Kruskal *P*-values indicating acceptance of the null hypothesis ($p > 0.05$). This suggests a consistent performance across the tested models in capturing the dynamics of this variable. In contrast, for the *LE* variable, the Elastic Net (EN) model showed borderline significance ($p = 0.0522$), indicating a potential divergence in performance compared to other models, although not conclusively significant. While further investigation may be warranted to elucidate the specific factors contributing to this observed trend, it suggests the need for careful consideration in model selection to ensure optimal performance across different variables. In contrast, the KW test did not reveal any significant differences among the models for variable CO_2. With all Kruskal P-values exceeding the threshold for significance ($p > 0.05$), it appears that the performance of the models remained consistent across the tested variations for this variable. This uniformity in performance highlights the robustness of the models in capturing the dynamics of the CO_2 variable, reflecting their effectiveness in addressing the complexities associated with carbon dioxide flux dynamics.

The investigation focused on three key variables: sensible heat flux (*H*), latent heat flux (*LE*), and carbon dioxide flux (CO_2), with each variable subjected to rigorous analysis and model evaluation. The subsequent examination of variable importance using the Shapley Additive Explanation (SHAP) method revealed the pivotal influence of radiative components, particularly R_S *(down)* and R_S *(up)*, on the predictive accuracy of the model. These findings underscore the significance of radiative variables in characterising sensible heat flux dynamics, highlighting their crucial role in surface-atmosphere interactions during the summer period. Similarly, the SHAP analysis for *LE* identified R_S *(down)* and *AT* as the primary drivers of *LE* dynamics, emphasizing their crucial role in elucidating evapotranspiration processes. Furthermore, the influence of atmospheric properties, including relative humidity (*RH*), wind speed (*SW*), and soil moisture (*SM*), on *LE* was elucidated, highlighting the multifaceted nature of factors contributing to latent heat flux variations. It is well known that plant physiology is highly dependent on environmental conditions (i.e. Driesen et al., 2020). Thus, temperature is a fundamental variable in the regulation of chemical reactions in living beings. In the case of plants this relationship is especially important in reactions involving the exchange of water and CO_2 through the stomata, there existing a very marked correlation between temperature, stomatal dynamics and subsequent metabolism (Greer and Weedon, 2012). In this sense, the pronounced correlation between this environmental variable and CO_2 and water fluxes is not surprising, as the results here reflect. In fact, the data confirm that temperature-related parameters (*AT_60h*, *AT_48h*) play a significant role in the resulting model (see Fig. 13).

The response of CO_2 and water fluxes to plant water status has been widely discussed in the literature because of their crucial role on stomatal dynamics (Buckley, 2019). Certain authors even suggest that it constitutes a key factor in the acclimation of the stomatal response to CO_2 (i.e. Talbott et al., 2003), with variations to air humidity well established (El-Sharkawy et al., 1985). In this model, RH also plays a major role, possibly because it has a limiting role when plants are grown in semiarid, rainfed conditions -and even more so during the summer season-, as is the case in the present study.

Finally, it is worth noting that $R_L(up)$ has been widely documented for its relationship with fluxes in ecosystems (Heitman et al., 2010), directly associated with soil temperature (Thakur et al., 2022). Its role in the physiology of CO_2 exchange is connected to its influence on soil respiration – dependent on soil temperature and moisture - that occurs in the upper soil layers (Meyer et al., 2018). In particular, sensible heat balance have been reported to be closely associated to soil-water evaporation processes (Heitman et al., 2008).

In summary, using meteorological variables and different mathematical models of machine learning we have successfully estimated, for the first time, CO_2 and water fluxes in a semi-arid model plantation. These results can be very significant to estimate -without the need of expensive and complex infrastructures-, these balances, so important from an application point of view, and especially from an environmental and agronomic perspective.

5. Conclusions

In practice, advanced agronomic work that helps in decision making regarding better adaptation to climate change requires the evaluation of field quantities of a higher order than the fundamental variables, such as surface energy balance fluxes, CO_2 flux and carbon assimilation, etc. Herein, this work tries to demonstrate that it is possible to improve knowledge on the carbon footprint in agriculture, efficiency in water use and also strategic issues such as prediction of crop yields in risk situations like droughts, heat waves, etc. This research is particularly significant in the present climate change scenario. In this scenario, strategies involving few economic and material resources are highly desirable. Our research provides ML methods that allow these fluxes to be estimated from more conventional meteorological variables.

Nine ML models were evaluated (i.e., LR, EN, KNN, GP, TREE, RF, XGB, SVR, and MLP) to predict H, LE and CO_2 fluxes over a vineyard. The results revealed that the from original ten features (θ, p, P, Ta, RH, RS↓, RS↑, RL↑, RL↓ and WS), the radiative components provided all the critical information. However, some variables such as precipitation (P) might be disposable due to the characteristics of the study area. The data show that, according to low MAE, MSE, RMSE, MASE and high R^2 for H, the study indicates an R^2 value of 0.99, designating the GP model as the best performing model. In the case of the LE study, an R^2 value of 0.96 was obtained and the SVR model proved to be the best performing. Finally, for CO_2 flux with the SVR model identified as the optimal performance model, machine learning–based methods have been found to give satisfactory results in the prediction of H, LE and CO_2.

Finally, the SHAP method revealed the most important variables for the prediction of H, LE and CO_2. For H, R_S *(down)* and R_S *(up)* are the most influential variables. For LE, the variables R_S (*down*) and AT remain the most critical ones and for CO_2, R_S *(down)*, followed by reflected radiation R_S *(up)* and AT_60h remain the most critical variables. Present findings are expected to provide important information to farmers in the area. Moreover, the main results of our study will serve as a valuable tool for classifying other similar agricultural areas.

CRediT authorship contribution statement

David Garcia-Rodriguez: Writing – original draft, Visualization, Validation, Supervision, Resources, Investigation, Conceptualization. **Pablo Catret Ruber:** Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Domingo J. Iglesias Fuente:** Writing – review & editing, Visualization, Validation, Investigation. **Juan José Martínez Durá:** Project administration, Funding acquisition. **Ernesto López Baeza:** Writing – review & editing, Visualization, Validation, Investigation, Data curation, Conceptualization. **Antonio Garcia Celda:** Writing – review & editing, Validation, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This publication is part of the R + D + i project "PID2020-120438RB-I00" funded by MCIN/AEI/10.13039/501100011033.

References

Adjuik, T.A., Davis, S.C., 2022. Machine learning approach to simulate soil CO2 fluxes under cropping systems. Agronomy 12 (1), 197. https://doi.org/10.3390/agronomy12010197.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements. Food and Agriculture Organization, Rome, Italy (FAO Irrigation and Drainage Paper, No. 56. 300 p.).

Andrén, O., Kätterer, T., 2001. Basic principles for soil carbon sequestration and calculating dynamic country-level balances including future scenarios. Assess. Methods Soil Carbon 495–511.

Bayram, S., Çıtakoğlu, H., 2023. Modeling monthly reference evapotranspiration process in Turkey: application of machine learning methods. Environ. Monit. Assess. 195 (1), 67. https://doi.org/10.1007/s10661-022-10662-z.

Bodas-Salcedo, A., López-Baeza, E., Martínez, F., Mateu, J., Montes, F., 2003. Spatiotemporal modeling and prediction of solar radiation. J. Geophys. Res. Atmos. 108 (D24) https://doi.org/10.1029/2003JD003903.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brown, L.A., Dash, J., Lidón, A.L., Lopez-Baeza, E., Dransfeld, S., 2019. Synergetic exploitation of the Sentinel-2 missions for validating the Sentinel-3 ocean and land color instrument terrestrial chlorophyll index over a vineyard dominated Mediterranean environment. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (7), 2244–2251. July. https://doi.org/10.1109/JSTARS.2019.2899998.

Buckley, T.N., 2019. How do stomata respond to water status? New Phytol. 224 (1), 21–36. https://doi.org/10.1111/nph.15899.

Burba, G., 2013. Eddy Covariance Method for Scientific, Industrial, Agricultural and Regulatory Applications: A Field Book on Measuring Ecosystem Gas Exchange and Areal Emission Rates. https://doi.org/10.13140/RG.2.1.4247.8561.

Campioli, M., Malhi, Y., Vicca, S., Luyssaert, S., Papale, D., Peñuelas, J., Reichstein, M., Migliavacca, M., Arain, M.A., Janssens, I.A., 2016. Evaluating the convergence between eddy-covariance and biometric meth- ods for assessing carbon budgets of forests. Nat. Commun. 7 https://doi.org/10.1038/ncomms13717.

Canadell, J., Raupach, M., 2008. Managing forests for climate change mitigation. Science (New York, N.Y.) 320, 1456–1457. https://doi.org/10.1126/science.1155458.

Carbó, E., Juan, P., Añó, C., Chaudhuri, S., Diaz-Avalos, C., López-Baeza, E., 2021. Modeling influence of soil properties in different gradients of soil moisture: the case of the Valencia Anchor Station validation site, Spain. Remote Sens. 13 (24), 5155. https://doi.org/10.3390/rs13245155.

Cateni, S., Colla, V., 2016. Variable selection for efficient design of machine learning-based models: Efficient approaches for industrial applications. In: Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17. Springer International Publishing, pp. 352–366. https://doi.org/10.1007/978-3-319-44188-7_27.

Chantry, M., Christensen, H., Dueben, P., Palmer, T., 2021. Opportunities and challenges for machine learning in weather and climate modelling: hard, medium, and soft AI. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 379, 20200083. https://doi.org/10.1098/rsta.2020.0083.

Chen, T., Guestrin, C., 2016. Xgboost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. https://doi.org/10.1145/2939672.2939785.

Citakoglu, H., Coşkun, Ö., 2022. Comparison of hybrid machine learning methods for the prediction of short-term meteorological droughts of Sakarya Meteorological Station in Turkey. Environ. Sci. Pollut. Res. 29 (50), 75487–75511. https://doi.org/10.1007/s11356-022-21083-3.

Colliander, A., Reichle, R.H., Crow, W.T., Cosh, M.H., Chen, F., Chan, S., Yueh, S.H., 2021. Validation of soil moisture data products from the NASA SMAP mission. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 364–392. https://doi.org/10.1109/JSTARS.2021.3124743.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27. January. https://doi.org/10.1109/TIT.1967.1053964.

Driesen, E., Van den Ende, W., De Proft, M., Saeys, W., 2020. Influence of Environmental Factors Light, co2, Temperature, and Relative Humidity on Stomatal Opening and Development: A Review. https://doi.org/10.3390/agronomy10121975.

Egipto, R., Aquino, A., Costa, J.M., Andújar, J.M., 2023. Predicting crop evapotranspiration under non-standard conditions using machine learning algorithms, a case study for vitis vinifera l. cv tempranillo. Agronomy 13, 2463. https://doi.org/10.3390/agronomy13102463.

El-Sharkawy, M.A., Cock, J.H., Del Pilar Hernandez, A., 1985. Stomatal response to air humidity and its relation to stomatal density in a wide range of warm climate species. Photosynth. Res. 7, 137–149. https://doi.org/10.1007/BF00037004.

Falge, E., Baldocchi, D., Tenhunen, J., Aubinet, M., Bakwin, P., Berbigier, P., Bernhofer, C., Burba, G., Clement, R., Davis, K.J., Elbers, J.A., Goldstein, A.H., Grelle, A., Granier, A., Guomundsson, J., Hollinger, D., Kowalski, A.S., Katul, G., Law, B.E., Malhi, Y., Meyers, T., Monson, R.K., Munger, J.W., Oechel, W., Paw, K.T., Pilegaard, K., Rannik, U., Rebmann, C., Suyker, A., Valentini, R., Wilson, K., Wofsy, S., 2002. Seasonality of ecosystem respiration and gross primary production as derived from FLUXNET measurements. Agric. Forest Meteorol. 113, 53–74. https://doi.org/10.1016/S0168-1923(02)00102-8.

Fernández-López, A., Marín-Sánchez, D., García-Mateos, G., Ruiz-Canales, A., Ferrández-Villena-García, M., Molina-Martínez, J.M., 2020. A machine learning method to estimate reference evapotranspiration using soil moisture sensors. Appl. Sci. 10 (6), 1912. https://doi.org/10.3390/app10061912.

Garcia, M.V., Aznarte, J.L., 2020. Shapley additive explanations for NO2 forecasting. Eco. Inform. 56, 101039 https://doi.org/10.1016/j.ecoinf.2019.101039.

Gash, J.H.C., 1986. A note on estimating the effect of a limited fetch on micrometeorological evaporation measurements. Bound.-Layer Meteorol. 35 (4), 409–413. https://doi.org/10.1007/bf00118567.

Geraldo-Ferreira, A., Soria-Olivas, E., Gómez-Sanchis, J., Serrano-López, A.J., Velázquez-Blazquez, A., López-Baeza, E., 2011. Modelling net radiation at surface using "in situ" netpyrradiometer measurements with artificial neural networks. Expert Syst. Appl. https://doi.org/10.1016/j.eswa.2011.04.231.

Greer, D.H., Weedon, M.M., 2012. Modelling photosynthetic responses to temperature of grapevine (vitis vinifera cv. semillon) leaves on vines grown in a hot climate. Plant Cell Environ. 35, 1050–1064. https://doi.org/10.1111/j.1365-3040.2011.02471.x.

Guevara-Escobar, A., González-Sosa, E., Cervantes-Jiménez, M., Suzán- Azpiri, H., Queijeiro-Bolaños, M.E., Carrillo-Ángeles, I., Cambrón- Sandoval, V.H., 2021. Machine learning estimates of eddy covariance carbon flux in a scrub in the mexican highland. Biogeosciences 18, 367–392. https://doi.org/10.5194/bg-18-367-2021.

Harrison, L.P., 1963. Fundamentals concepts and definitions relating to humidity. In: Wexler, A. (Ed.), Humidity and Moisture, vol. 3. Reinhold Publishing Co., N.Y.

Heitman, J.L., Horton, R., Sauer, T.J., DeSutter, T.M., 2008. Sensible heat observations reveal soil-water evaporation dynamics. J. Hydrometeorol. 9 (1), 165–171. https://doi.org/10.1175/2007JHM963.1.

Heitman, J.L., Horton, R., Sauer, T.J., Ren, T.S., Xiao, X., 2010. Latent heat in soil heat flux measurements. Agric. For. Meteorol. 150, 1147–1153. https://doi.org/10.1016/j.agrformet.2010.04.017.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A Practical Guide to Support Vector Classification.

Huntingford, C., Jeffers, E.S., Bonsall, M.B., Christensen, H.M., Lees, T., Yang, H., 2019. Machine learning and artificial intelligence to aid climate change research and preparedness. Environ. Res. Lett. 14, 124007 https://doi.org/10.1088/1748-9326/ab4e55.

Iglesias, D.J., Quinones, A., Font, A., Martínez-Alcántara, B., Forner-Giner, M.A., Legaz, F., Primo-Millo, E., 2013. Carbon balance of citrus plantations in eastern Spain. Agric. Ecosyst. Environ. 171, 103–111. https://doi.org/10.1016/j.agee.2013.03.011.

Jena, M., Dehuri, S., 2020. Decision tree for classification and regression: a state-of-the art review. Informatica 44 (4), 405–420. https://doi.org/10.31449/inf.v44i4.3023.

Kim, Y., Johnson, M.S., Knox, S.H., Black, T.A., Dalmagro, H.J., Kang, M., Kim, J., Baldocchi, D., 2019. Gap-filling approaches for eddy covariance methane fluxes: a comparison of three machine learning algorithms and a traditional method with principal component analysis. Glob. Chang. Biol. 26, 1499–1518. https://doi.org/10.1111/gcb.14845.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization arXiv preprint arXiv:1412.6980.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47 (260), 583–621. https://doi.org/10.2307/2280779.

Locatelli, B., 2016. Ecosystem services and climate change. In: Routledge Handbook of Ecosystem Services. Routledge, pp. 481–490.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable ai for trees. Nat. Mach. Intell. 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Mahabbati, A., Beringer, J., Leopold, M., McHugh, I., Cleverly, J., Isaac, P., Izady, A., 2021. A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers. Geosci. Instrum. Methods Data Syst. 10, 123–140. https://doi.org/10.5194/gi-10-123-2021.

Meyer, N., Welp, G., Amelung, W., 2018. The temperature sensitivity (q10) of soil respiration: controlling factors and spatial prediction at regional scale based on environmental soil classes. Glob. Biogeochem. Cycles 32, 306–323. https://doi.org/10.1002/2017gb005644.

Nagelkerke, N.J.D., 1991. A note on a general definition of the coefficient of determination. Biometrika 78, 691–692. https://doi.org/10.1093/biomet/78.3.691.

Perez-Priego, O., Katul, G., Reichstein, M., El-Madany, T.S., Ahrens, B., Carrara, A., Scanlon, T.M., Migliavacca, M., 2018. Partitioning eddy covariance water flux components using physiological and micrometeo- rological approaches. J. Geophys. Res. Biogeosci. 123, 3353–3370. https://doi.org/10.1029/2018JG004637.

Rasmussen, C.E., 2004. Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (Eds.), Advanced Lectures on Machine Learning. ML 2003, Lecture Notes in Computer Science, vol. 3176. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_4.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Ryu, Y., Baldocchi, D.D., Black, T.A., Detto, M., Law, B.E., Leuning, R., Miyata, A., Reichstein, M., Vargas, R., Ammann, C., Beringer, J., Flanagan, L.B., Gu, L., Hutley, L.B., Kim, J., McCaughey, H., Moors, E.J., Rambal, S., Vesala, T., 2012. On the temporal upscaling of evapotranspiration from instantaneous remote sensing measurements to 8-day mean daily-sums. Agric. For. Meteorol. 152, 212–222. https://doi.org/10.1016/j.agrformet.2011.09.010.

Schwank, M., Wigneron, J.P., Lopez-Baeza, E., Volksch, I., Matzler, C., Kerr, Y.H., 2012. L-band radiative properties of vine vegetation at the MELBEX III SMOS cal/val site. IEEE Trans. Geosci. Remote Sens. 50 (5), 1587–1601. https://doi.org/10.1109/TGRS.2012.2184126.

Shapley, L.S., 1953. A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games II. Princeton University Press, Princeton, pp. 307–317.

Shi, H., Luo, G., Hellwich, O., Xie, M., Zhang, C., Zhang, Y., Wang, Y., Yuan, X., Ma, X., Zhang, W., Kurban, A., Maeyer, P.D., de Voorde, T.V., 2022. Variability and uncertainty in flux-site scale net ecosystem exchange simulations based on machine learning and remote sensing: A systematic evaluation. https://doi.org/10.5194/bg-2022-46.

Talbott, L.D., Rahveh, E., Zeiger, E., 2003. Relative humidity is a key factor in the acclimation of the stomatal response to CO2. J. Exp. Bot. 54, 2141–2147. https://doi.org/10.1093/jxb/erg215.

Tang, Y., Li, H., 2023. Comparing the performance of machine learning methods in predicting soil seed bank persistence. Eco. Inform. 77 (January), 102188 https://doi.org/10.1016/j.ecoinf.2023.102188.

Thakur, G., Schymanski, S.J., Mallick, K., Trebs, I., Sulis, M., 2022. Downwelling longwave radiation and sensible heat flux observations are critical for surface temperature and emissivity estimation from flux tower data. Sci. Rep. 12, 8592. https://doi.org/10.1038/s41598-022-12304-3.

Wang, H., Jia, G., Zhang, A., Miao, C., 2016. Assessment of spatial representativeness of eddy covariance flux data from flux tower to regional grid. Remote Sens. 8, 742. https://doi.org/10.3390/rs8090742.

Wiesner, S., Desai, A.R., Duff, A.J., Metzger, S., Stoy, P.C., 2022. Quanti- fying the natural climate solution potential of agricultural systems by combining eddy covariance and remote sensing. J. Geophys. Res. Biogeosci. 127 https://doi.org/10.1029/2022jg006895.

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., Song, L., 2018. Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. J. Geophys. Res. Atmos. 123, 8674–8690. https://doi.org/10.1029/2018jd028447.

Yao, J., Gao, Z., Huang, J., Liu, H., Wang, G., 2021. Technical note: Uncer- tainties in eddy covariance co2 fluxes in a semiarid sagebrush ecosystem caused by gap-filling approaches. https://doi.org/10.5194/acp-2021-631.

Zhu, S., Olde, L., Lewis, K., Quaife, T., Cardenas, L., Loick, N., Xu, J., Hill, T., 2023. Eddy covariance fluxes over managed ecosystems extrapolated to field scales at fine spatial resolutions. Agric. For. Meteorol. 342, 109675. https://doi.org/10.1016/j.agrformet.2023.109675.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat Methodol.) 67 (2), 301–320. http://www.jstor.org/stable/3647580.

Zouzou, Y., Citakoglu, H., 2023. General and regional cross-station assessment of machine learning models for estimating reference evapotranspiration. Acta Geophys. 71 (2), 927–947. https://doi.org/10.1007/s11600-022-00939-9.