Ecological Processes

RESEARCH

Open Access

Application of machine learning techniques to simulate the evaporative fraction and its relationship with environmental variables in corn crops

Terenzio Zenone[1*], Luca Vitale[1], Daniela Famulari[2] and Vincenzo Magliulo[1]

Abstract

Background: The evaporative fraction (EF) represents an important biophysical parameter reflecting the distribution of surface available energy. In this study, we investigated the daily and seasonal patterns of EF in a multi-year corn cultivation located in southern Italy and evaluated the performance of five machine learning (ML) classes of algorithms: the linear regression (LR), regression tree (RT), support vector machine (SVM), ensembles of tree (ETs) and Gaussian process regression (GPR) to predict the EF at daily time step. The adopted methodology consisted of three main steps that include: (i) selection of the EF predictors; (ii) comparison of the different classes of ML; (iii) application, cross-validation of the selected ML algorithms and comparison with the observed data.

Results: Our results indicate that SVM and GPR were the best classes of ML at predicting the EF, with a total of four different algorithms: cubic SVM, medium Gaussian SVM, the Matern 5/2 GPR, and the rational quadratic GPR. The comparison between observed and predicted EF in all four algorithms, during the training phase, were within the 95% confidence interval: the R^2 value between observed and predicted EF was 0.76 (RMSE 0.05) for the medium Gaussian SVM, 0.99 (RMSE 0.01) for the rational quadratic GPR, 0.94 (RMSE 0.02) for the Matern 5/2 GPR, and 0.83 (RMSE 0.05) for the cubic SVM algorithms. Similar results were obtained during the testing phase. The results of the cross-validation analysis indicate that the R^2 values obtained between all iterations for each of the four adopted ML algorithms were basically constant, confirming the ability of ML as a tool to predict EF.

Conclusion: ML algorithms represent a valid alternative able to predict the EF especially when remote sensing data are not available, or the sky conditions are not suitable. The application to different geographical areas, or crops, requires further development of the model based on different data sources of soils, climate, and cropping systems.

Keywords: Energy flux, Evapotranspiration, Eddy covariance, Artificial intelligence

Background

Intensive agricultural systems in the Mediterranean areas represent the largest consumers of fresh water and, in some cases, threatening the availability of water resources for other uses (Nguyen et al. 2016; Alexandridis et al. 2009). Projections of future climate change indicate an increasing pressure on water use in the Mediterranean regions as consequence of progressive increase in global temperatures and many countries of this area

*Correspondence: terenzio.zenone@isafom.cnr.it

[1] National Research Council, Institute for Agricultural and Forestry Systems in the Mediterranean, P.Le Enrico Fermi 1, 80055 Portici, Italy
Full list of author information is available at the end of the article

have already experienced water shortages during the last 20 years (Milano et al. 2013).

Several methodologies are currently available to measure the water use in agricultural systems: among them, the direct observation of the evapotranspiration or latent heat flux (LE) using the eddy covariance (EC) technique at ecosystem level, represents the most widely used approach worldwide (Baldocchi 2020). LE is a key component of the energy, carbon, and hydrological cycle: several land surface processes are tightly linked to evapotranspiration through complex feedback mechanisms, such as precipitation (Zveryaev and Allan 2010), surface temperature (Seneviratne et al. 2006), available energy (Gentine et al. 2011) and the amount of soil water available for the plants (de Tomás et al. 2014).

The evaporative fraction (EF), defined as the ratio between LE and the sum of LE and sensible heat flux (*H*), represents an important biophysical parameter reflecting the distribution of surface available energy, and a good tool for interpreting the components of the energy budget (Gentine et al. 2011). In terms of ecophysiological processes, EF is coupled to drought events (Schwalm et al. 2010; Trenberth and Guillemot 1996): when EF approaches unity, most of the available energy is partitioned to LE and water can flow without limitations through the soil–plant–atmosphere continuum; therefore, EF can be viewed as an index of water deficit, ranging from 0, when no water is available, to 1 when there are no water limitations (Schwalm et al. 2010). Several studies investigated the EF of cropland in the last decade. Zhou and Wang (2016) reported a positive correlation between monthly EF, air temperature, NDVI and relative humidity across a network of Ameriflux sites in North America. Yang et al. (2013) analyzed the diurnal patterns of EF, indicating that the self-preservation assumption no longer holds over growing seasons, and diurnal patterns of evapotranspiration are mainly influenced by stomatal regulation. The EF has also been used to predict the evapotranspiration (ET) from remotely sensed instantaneous observations by the application of an improved constant EF Method that include the use of a decoupling factor (Ω) to represents the relative contribution of the radiative and aerodynamic terms to the overall ET (Tang et al. 2017a) or from remotely sensed real time observations with a simplified derivations of a theoretical model (Tang et al. 2017b).

The analysis of daily behavior of EF and its response to biophysical factors have been investigated by Gentine et al. (2011), showing that it is rarely constant and that its temporal power spectrum is wide. EF can either be derived from micrometeorological observations (Schwalm et al. 2010), satellite products such as MODIS (Nutini et al. 2014; Lu et al. 2013a) or from empirical models based on satellite observations and functional relationships (Zhou and Wang 2016). Ground-based observations of EF are normally conducted within experimental observation network such as Fluxnet, ICOS and NEON. Over the past decades the increasing number of geoscientific, atmosphere, and land surface data availability from the research network infrastructures, have co-evolved with development of new machine learning (ML) algorithms (Reichstein et al. 2019).

ML is currently used to simulate a wide range of biophysical and environmental processes, and has been rapidly expanding, covering a wide range of scientific disciplines. ML can be defined as the subset of Artificial Intelligence that provides computer systems the ability to simulate human intelligence (Dash et al. 2021). In recent years ML has been applied in more and more scientific fields including, for example, bioinformatics (Kong et al. 2007), biochemistry (Richardson et al. 2016; Wildenhain et al. 2015), medicine (Kang et al. 2015), meteorology (Kramer et al. 2017; Aybar-Ruiz et al. 2016), economic sciences (Barboza et al. 2017), robotics (Takahashi et al. 2017), aquaculture (López-Cortés et al. 2017), and climatology (Fang et al. 2017). The supervised ML can be classified in two main groups: "classification model" where the target variables are categories, and "regression model" where predictor and target are continuous variables (Liakos et al. 2018). Regression models are among the hottest topics in the development of algorithms able to learn from data and build predictions without being explicitly parameterized for that task. This makes the models able to predict future outcomes after being trained on the basis of past experimental observations where predictor variables are used to train the model with the aim of producing a function that is approximate enough to be able to predict an output (target variables) from new inputs when they are introduced (Sen et al. 2020).

Over the last few years, literature related to the use of ML in agricultural sciences has been growing significantly, and several efforts to predict the energy balance components have been made: Zhao et al. (2019) developed a physics-constrained ML model, able to predict the ET across a series of Fluxnet sites; a similar study was conducted by Tramontana et al. (2016) where several ML methods were used to predict CO_2 and energy exchange (i.e., LE and *H*) across multiple Fluxnet sites. More recently, Pan et al. (2020) used an ensemble of remote sensing, ML and land surface modeling to simulate the ET at global level, while Mosre and Suárez (2021) report the use of ML with in situ remote sensing data to determine the actual ET in arid cold regions. ML algorithms have been also used to predict CO_2 (Guevara-Escobar et al. 2020), latent heat flux (Zhao et al. 2019; Dou and Yang 2018; Yin et al. 2021; Fu et al. 2021; Mosre and Suárez 2021), reference evapotranspiration (Anurag et al. 2021; Borges et al. 2020), and to evaluate terrestrial evapotranspiration at global scale (Pan et al. 2020).

Nevertheless, to our best knowledge, there is still a lack of ML application to predict the daily EF for agricultural crops in Mediterranean regions. The application of new methods to determine EF could represent a valid alternative to conventional remote sensing-based models, especially when the sky conditions are unsuitable, or data are not available. The aim of this study is, therefore, to investigate the daily and seasonal patterns of EF on a multi-year corn cultivation located in southern Italy and evaluate the performance of different ML algorithms to predict the EF at daily time step. In this study, instead of choosing one particular model, we have first trained our dataset to several ML models in parallel, and then we have chosen to develop only the models with the higher performances.

Methods

Experimental site

Data collection occurred during the period 2004–2009 on a farm located in southern Italy (40° 31′ 25.5″ N, 14° 57′ 26.8″ E) that is the European southernmost cropland observation candidate site of the ICOS (Integrated Carbon Observation System) European infrastructure. The area is characterized by typical Mediterranean climate: over the last 30 years, the average annual precipitation was 908 mm with an overall mean air temperature of 15.5 °C. Most of the precipitation occurs in October–November while the driest month is July. The site was cultivated with silage corn (*Zea mays*) as main crop; the vegetative season was defined according to its growing cycle (i.e., day of sowing, and day of harvest).

The site was equipped with an EC system composed of a fast response sonic anemometer (R3, Gill Instruments Ltd., Lymington, UK) and open-path infrared CO_2/H_2O gas analyzer (Li-7500, Li-Cor Inc., Lincoln, NE, USA) to measure energy fluxes (Latent Heat Flux, LE and Sensible Heat Flux, H) and the CO_2 net ecosystem exchange (NEE). More detailed information about the EC system and ancillary sensors can be found in Vitale et al. (2007) and Vitale et al. (2016). To achieve a satisfactory upwind fetch, the height of eddy covariance sensors was set to 2 m above the ground while the canopy was shorter than 1 m and later moved following the crop growth up to 3.9 m at harvest time. The experimental field covers an area of 10 ha, whereas the average footprint was 182 m along the prevalent wind direction (NE–SW).

Data streams from both IRGA and sonic anemometer were logged at a frequency of 20 Hz via the Eddymeas software, and the fluxes calculated using the software EddySoft (Kolle and Rebmann 2007). Corrections for flux losses as well as for sensor separation (Horst and Lenschow 2009) and low-pass frequency filters (Moncrieff et al. 2004) were also applied. High-frequency spectral correction was performed according to the model of Eugster and Senn (1995). The flux footprint was computed according to the analytical model of Schuepp et al. (1990), and quality control was applied to half-hourly (30-min) fluxes following Göckede et al. (2004), by assigning a quality flag (0 for good data, 1 for acceptable data, 2 for bad data) to each flux value. The standard WPL terms were considered to correct for density fluctuations (see Webb et al. 1980).

EC data used in this study were part of the Fluxnet 2015 Dataset (https://fluxnet.org/data/fluxnet2015-dataset/). Time series were processed according to the approaches reported in Pastorello et al. (2020). This methodology includes a preliminary processing block, where data quality assurance and quality control (QA/QC) for all the variables investigated are carried out by means of a variable-specific despiking routines. Energy fluxes (H and LE), used in this study were then gap-filled using the MDS method (Reichstein et al. 2005) and the values were adjusted according to a methodology that corrects for un-closure of the energy budget, by assuming a correct Bowen ratio. The corrected fluxes are obtained by multiplying the original, gap-filled LE and H data by an energy balance closure correction factor (EBC_CF), which is calculated on a subset of observations—where all the components needed to calculate the energy balance closure are available—namely: measured net radiation (R_n) and soil heat flux (G), and measured or good-quality gap-filled latent heat and sensible heat fluxes. The correction factor is calculated for each half-hour as (R_n – G) / ($H+LE$), and the time series is filtered removing the values that are outside 1.5 times the interquartile range, then used as a basis to calculate the corrected H and LE fluxes.

Environmental variables were measured at 1 Hz and averaged every half-hour: precipitation was monitored using a rain gauge (Rain Collector II, Davis Instruments, CA, USA) located on the ground, soil temperature and volumetric water content (SWC) at 0.3-m depth were also determined by means of TCAV, 105E thermocouple probes and CS 616 water content reflectometer (Campbell Scientific, Ltd., Shepshed, UK), respectively.

Soil heat flux density was monitored with heat flux plates (HFT3 Campbell Sci. Ltd., Shepshed, UK) at 5 cm below the soil surface. Data were collected at three different locations within the footprint area of the EC tower and with a time step of 30 min. G values coming from the plate were first corrected for the change in heat storage in the soil layer above the plate following the methodology reported in the Instruction for soil-meteorological measurements of the ICOS) protocol (Op de Beeck et al. 2017).

The main components of the solar radiation, i.e., incoming, reflected and net radiation, were monitored using an Eppley pyranometer (Eppley Laboratory Inc., USA) CNR1 net radiometer (Kipp & Zonen, NL), all located at 4 m from the ground, while air temperature and vapor pressure were monitored using the Bowen Ratio System (Campbell Scientific Ltd., Shepshed, UK).

Leaf area index (LAI), monitored for the entire crop cycle by sampling plants at different locations of the field to cover the spatial variability, was determined using a Li-3100 leaf area meter (Li-Cor Biosciences, Lincoln, Nebraska, USA). For the years 2008–2009, LAI data were obtained from high spatial resolution data acquired by SPOT family satellite, as the field data were too sparse. Map elaborations were based on an empirical relationship between LAI and one-view angle measurements of reflectance (rλ) in the red and infrared bands (for more specific information, see Vitale et al. 2016).

Evaporative fraction

The partitioning of incoming energy was evaluated using the EF approach. The instantaneous EF (dimensionless) was calculated from LE (W m^{-2}) and H (W m^{-2}) fluxes, measured by the above-described EC station, as follows:

$$\mathrm{EF}_{\mathrm{daytime}} = \frac{\int_{t1}^{t2}\mathrm{LE}(t)\mathrm{d}t}{\int_{t1}^{t2}[H(t)+\mathrm{LE}(t)]\mathrm{d}t}, \tag{1}$$

where the time difference t_2–t_1 in the present study refers to the time from 08:00 to 18:00 (UTC + 1) (Zenone et al. 2015). An ideal energy balance closure can be achieved when the available energy is equal to the turbulent fluxes, and then, EF can be expressed as:

$$\mathrm{EF} = \frac{\mathrm{LE}}{H+\mathrm{LE}} = \frac{\mathrm{LE}}{\mathrm{Rn}-G}, \tag{2}$$

where R_{n} is the net radiation and G is the soil heat flux. EF is a biophysical parameter related to the partitioning of available energy and therefore to the energy balance closure at the measuring point. EC fluxes of energy typically do not satisfy the energy balance closure, due to different levels of sensor errors, unmeasured storage terms, mismatches in source area and landscape heterogeneity (Foken 2008).

To overcome this issue, EF can be rewritten as

$$\mathrm{EF} = \frac{1}{1+H/\mathrm{LE}}. \tag{3}$$

We could then assume that the errors on LE and H present similar magnitude (Hollinger and Richardson 2005; Richardson et al. 2006; Foken 2008) and are uncorrelated; this allows to mathematically cancel out the errors linked to the lack of energy balance closure (Schwalm et al. 2010).

Methodology flowchart adopted for the ML algorithms application

The ML methodology adopted to predict the EF using the ML algorithms consists of 3 main steps (Fig. 1):

Step 1: Predictors selection. A preliminary analysis was conducted to determine which input variables (Table 1) have the highest statistical significance on the prediction of EF. This involved the use of the neighborhood component analysis (NCA, see Wang and Tan 2017), the minimum redundancy maximum relevance (MRMR, see Jo et al. 2019) algorithms, and a correlation matrix to determine the Pearson coefficient as well as the variance inflation factor (VIF) to check the multicollinearity among the predictors.

Step 2: Selection of the best ML algorithms. In this second step we made a preliminary application of the different ML models to compare their performance, to then identify the best ML algorithm. This included five different classes of ML algorithms: linear regression (LR), regression trees (RT), Gaussian process regression (GPR), support vector machines (SVM), and ensembles of tree (ETs) models. The methods that provided values of $R^2 > 0.65$ during the training process were further developed and investigated: two classes of ML algorithm were selected, the GPR and the SVM.

Step 3: ML application

GPR models

To investigate the relationship between the environmental features selected and the EF variability, a GPR model (Rasmussen and Williams 2006) was adopted. GPR model assume that the output y of a function f with input x can be expressed as

$$y = f(x) + \in, \tag{4}$$

where x is the input vector, f is the function value and y is the observed target value.

We have supposed that the observed y values differ from the function values $f(x)$ by additive noise, and we will further assume that this noise follows an independent, identically distributed Gaussian distribution with zero mean, variance σ_n^2 and $\varepsilon \sim (0, \sigma_n^2)$.

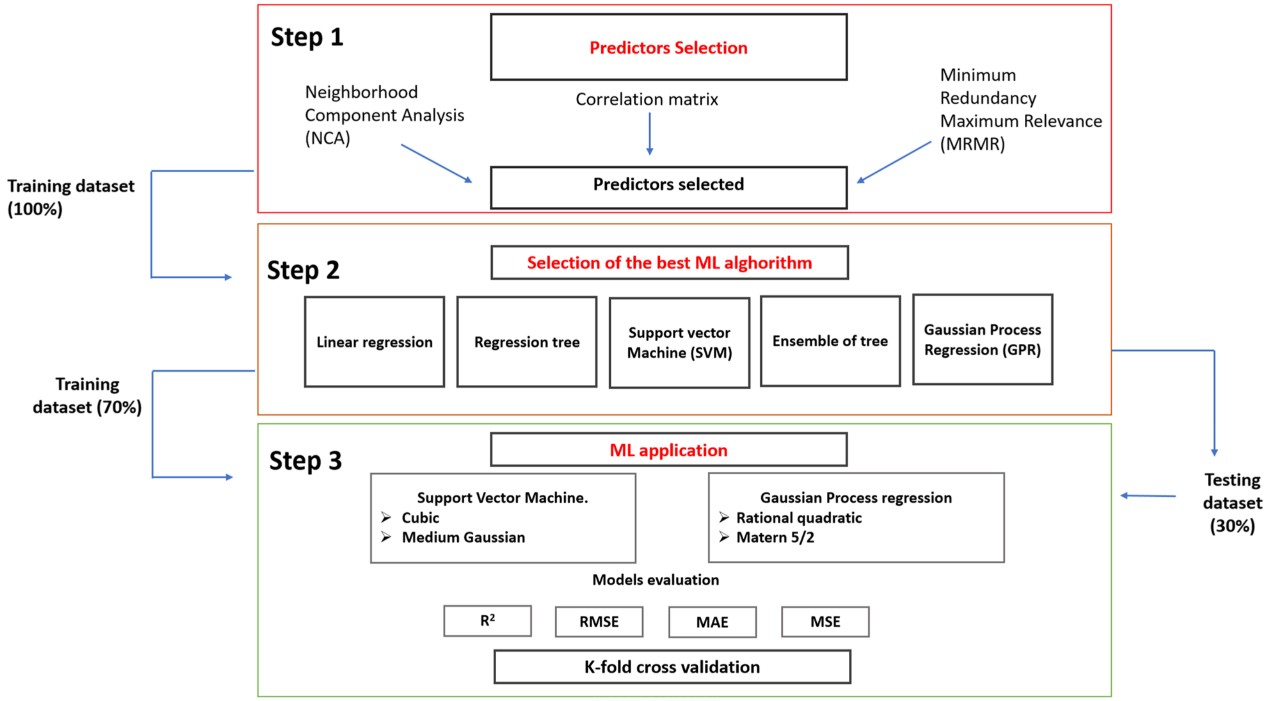

Fig. 1 Flowchart of the methodology adopted to select the predictor variables and select the ML algorithms

Table 1 Abbreviation and description of the predictor variables used in the machine learning algorithm of the study

Abbreviated variable	Description	Variable category	Unit	Sampling frequencies
LAI	Leaf area index	Biophysical	Number (#)	Bi-monthly
NEE*	CO_2 net ecosystem exchange	Biophysical	μmol m^{-2} s^{-1}	30 min
SWC*	Soil water content	Environmental/management	%	30 min
T_a*	Air temperature	Meteorology	°C	30 min
R_n	Net radiation	Meteorology	W m^{-2}	30 min
*G**	Soil heat fluxes	Meteorology	W m^{-2}	30 min
VPD*	Vapor pressure deficit	Meteorology	Pascal	30 min

*Predictor variable selected after Step 1

GPR model assume also that not only the error term $\in$, but also f is considered as a random variable (Ballabio et al. 2019). The GPR $f(x)$ is distributed as a Gaussian process:

$$f(x) \sim gp(\mu(x), k(x, x^*)), \tag{5}$$

where $f(x)$ is defined by its mean $\mu(x)$ and covariance $k(x,x*)$.

The covariance function k, is known as the kernel function of the GPR models and analyzes the dependence of the function values between different values of x. The kernel function of a GPR model can be considered equivalent to Kriging (Stein 1999). While Kriging is, in general, performed on a geographical space, the GPR is applied arbitrarily to a number of different covariates. The choice of the appropriate kernel function is based on the structure, or peculiar patterns of the data investigated (Ballabio et al. 2019).

In this study, two kernel functions, the Matern 5/2 (Eqs. 6, 7) and the rational quadratic (Eq. 8) were used:

$$k(x_i, x_j) = \sigma_f^2\left(1 + \frac{\sqrt{5r}}{\sigma_l} + \frac{5r^2}{3\sigma_l^2}\right)\exp\left(-\frac{\sqrt{5r}}{\sigma_l}\right), \tag{6}$$

where

$$r = \sqrt{(x_i - x_j)^T (x_i - x_j)}, \tag{7}$$

is the Euclidean distance between x_i and x_j and σ_l is the characteristic length scale.

$$k(x_i, x_j|\theta) = \sigma_f^2 \left(1 + \frac{r^2}{2\alpha\sigma_l^2}\right)^{-\alpha}, \tag{8}$$

where σ_l is the characteristic length scale, α is a positive scale-mixture parameter, and

$$r = \sqrt{(x_i - x_j)^T (x_i - x_j)}. \tag{9}$$

is the Euclidean distance between x_i and x_j.

It is possible to use a separate length scale σ_m for each predictor m ($m = 1, 2, \ldots, d$).

The built-in kernel (covariance) functions with a separate length scale for each predictor implement automatic relevance determination. The unconstrained parametrization θ in this case is

$$\theta_m = \log \sigma_m \quad \text{for } m = 1, 2 \ldots g, \tag{10}$$

$$\theta_{d+1} = log\sigma_f. \tag{11}$$

SVM regression

SVM analysis is a popular machine learning tool for classification and regression, first identified by Vapnik (1999). SVM regression is considered a non-parametric technique because, as GPR, it relies on kernel functions. SVM employs a model able to build a decision surface by mapping the input and target variables into a high-dimensional (or infinite-dimensional) feature space. Next, a linear regression is executed in the high-dimensional feature space. This mapping operation is required because in many cases, the relation between a multidimensional input (i.e., predictor variables) and the output (i.e., target variables) is unknown and very likely to be nonlinear. SVM regression aims at finding a linear hyperplane, which fits the multidimensional input vectors to output values (Wauters and Vanhoucke 2014). Two SVM kernel functions, the cubic Gaussian (Eq. 12) and polynomial (Eq. 13) were used:

$$G(x_j, x_k) = \exp(-\|x_j - x_k\|^2), \tag{12}$$

$$G(x_j, x_k) = (1 + x_j' x_k)^q. \tag{13}$$

where q is in the set {2,3,...} and G is a Gram matrix of an n-by-n matrix that contains elements $g_{i,j} = G(x_i, x_j)$. Each element $g_{i,j}$ is equal to the inner product of the predictors as transformed by ϕ. However, we do not need to know ϕ, because we can use the kernel function to generate Gram matrix directly. Using this method, nonlinear SVM finds the optimal function $f(x)$ in the transformed predictor space.

Quality assessment

For the evaluation of the results of the ML models, several criteria were used: the determination coefficient (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean squared error (MSE). R^2 is a key output in regression analysis, while RMSE is a measure of the average squared difference between the predicted and actual outputs of a model, the MAE measures the average error between them. In contrast to R^2, lower values of RMSE and MAE indicate a better performance of an algorithm.

Cross-validation analysis

In the current analysis, the evaluation of ML model performance was based on EF data split into 5-year training set (2004 to 2008) and 1-year testing set (2009). However, this analysis remained only partially reliable since the predictive performance obtained for only one test set can be different to that obtained for another test set (Rodríguez et al. 2010). In order to tackle this potential issue, we apply the K-fold cross-validation to the original dataset; all the data collected were split into ten randomly partitioned sets ($K = 10$ folds) of almost equal size. In the first iteration, the first fold ($K = 1$) was used to test ML models and the remaining folds ($K = 2$ to 10) were used for the training. In the next iteration, the second fold ($K = 2$) served for the testing stage, while the rest ($K = 1$ and $K = 3$ to 10) were used for the training. The process was repeated until the final iteration was reached.

Results

Based on the growing seasons full datasets, average diurnal cycles of LE + H, H and EF were calculated for each month, as shown in Fig. 2. Heat fluxes show the typical Gaussian shape, whereas EF features a characteristic concave U profile, with higher values in the early morning (e.g., 0.7–0.8 for July and August) and late afternoon hours while lower values (e.g., 0.4–0.5) were observed in late spring (May), early summer (June) and early fall (September). Considering only data spanning across midday (i.e., from 11:00 to 13:00), the EF behavior appears to be more constant demonstrating a little variance of the daytime values, with the average values ranging between 0.4 and 0.6, during the entire growth cycle.

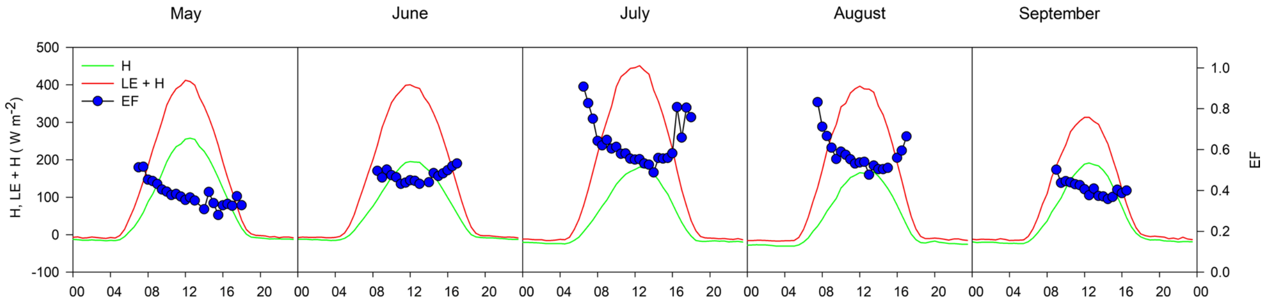

Fig. 2 Monthly diurnal cycles of available energy (LE + *H*), *H*, and daytime-EF, measured over the growing seasons of the corn crop. EF is calculated every half-hour from the fluxes, according to Eq. 1

Step 1: predictors selection

The correlation matrix reported in Table 2 indicated, as expected, a significant correlation between several of the predictors as indicated by the *p* values < 0.05. The correlation between the predictors and EF were all significant and the highest correlation was found between EF and *G* ($R = 0.40$, $\sigma = 4 \times 10^{-10}$). The VIF indicates an average value among all the predictors of 1.858 and a collinearity tolerance of 0.580. The results from the other two feature selection methods, i.e., NCA and MRMR are illustrated in Fig. 3: both methodologies identified LAI and R_n as predictors with the lowest weight (close to zero). The feature weights obtained from the NCA analysis suggested that (classified from highest to lowest values) air temperature (T_a), NEE, SWC, *G* and VPD were the most influencing factors on EF predictions (Fig. 3B). The feature weights obtained from the MRMR partially confirmed the NCA results, indicating NEE, T_a, and *G* as the main predictors, while VPD and SWC score was close to zero (Fig. 3A). In view of these findings, NEE, SWC, T_a, VPD, and *G* were selected for the ML predictive analysis on EF prediction: NEE represents the net CO_2 exchange between the ecosystem and the atmosphere, measured with the EC technique, to emphasize the role of the gas exchange, the volumetric SWC representing the amount of water potentially available for the vegetation, the T_a and VPD represents the atmospheric physical conditions, and *G* is component of the energy balance of the ecosystem investigated.

Selection of the best ML algorithm

The results from the different ML algorithms proposed for the training phase are shown in Table 3, in terms of the different evaluation criteria adopted (see "Quality assessment" section). The linear regression models among the four methods (see Table 3) show an average R^2 of 0.44 (±0.06) and an average RMSE of 0.1 (±0.005), while for the regression tree group R^2 was 0.41 (±0.05) and RMSE 0.19 (±0.005). Similar results were found for the SVM model, with an average R^2 of 0.44 (±0.11) and RMSE of 0.01 (±0.002), while slightly better performance was found for the ensemble of tree models: R^2 0.54 (±0.15) and RMSE 0.009 (±0.002). The best performances were found for two algorithms: cubic Gaussian SVM (Eq. 9) and polynomial (Eq. 10) SVM classes; and the Matern 5/2 (Eq. 6) and rational quadratic (Eq. 7) GPR classes. The results of the training-and-testing phase instead are reported in Table 4.

Table 2 Pearson correlation coefficient (*R*) of all predictor variables selected for the application of the ML algorithm

	LAI	R_n	*G*	VPD	T_a	NEE	SWC	EF
LAI	1	0.51*	− 0.64*	0.12	0.27*	0.29*	0.22*	− 0.35*
R_n	0.51*	1	− 0.35*	− 0.11	0.01	0.11	0.32*	− 0.29*
G	− 0.64*	− 0.35	1	0.02	0.05	− 0.55*	− 0.08	0.40*
VPD	0.12	− 0.11*	0.02	1	0.57*	− 0.15*	− 0.08	− 0.26*
T_a	0.27*	0.01	0.05	0.57*	1	− 0.02	− 0.05	− 0.39*
NEE	0.29*	0.11	− 0.55*	− 0.15*	− 0.02	1	0.11	− 0.31*
SWC	0.22*	0.32	− 0.08	− 0.08	− 0.05	0.11	1	− 0.27*
EF	− 0.35*	− 0.29*	0.40*	− 0.26*	− 0.39*	− 0.31*	− 0.27*	1

EF is the dependent variable

*Significant correlation (*p* values < 0.05)

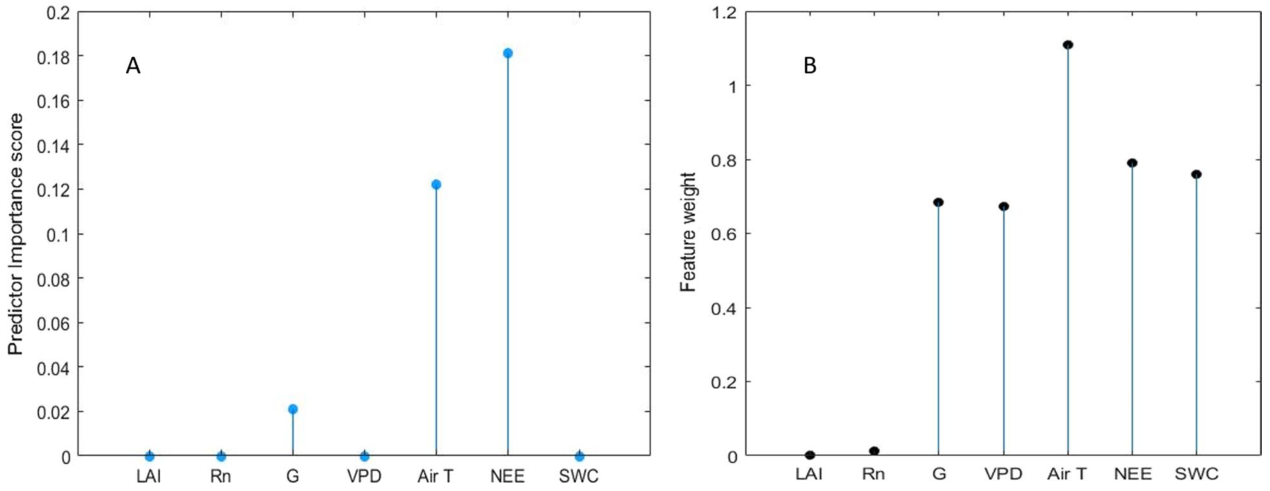

Fig. 3 Results of the feature selection analysis by NCA and MRMR algorithms

Table 3 Results of the evaluation criteria applied on the different ML algorithms adopted for the training phase

Prediction model		**RMSE**	**R^2**	**MSE**	**MAE**
		Training phase			
Linear regression	Linear	0.109	0.40	0.012	0.086
	Interaction linear	0.101	0.48	0.010	0.076
	Robust linear	0.110	0.39	0.012	0.086
	Stepwise linear regression	0.098	0.52	0.009	0.074
Regression tree	Fine Tree	0.103	0.48	0.010	0.077
	Medium Tree	0.111	0.39	0.012	0.084
	Coarse Tree	0.113	0.37	0.012	0.088
Support vector machine (SVM)	Linear SVM	0.111	0.38	0.012	0.085
	Quadratic SVM	0.088	0.61	0.007	0.065
	Fine Gaussian SVM	0.113	0.35	0.012	0.086
	Coarse Gaussian SVM	0.104	0.45	0.011	0.081
Ensemble of Trees	Boosted Trees	0.085	0.64	0.007	0.063
	Bagged Trees	0.107	0.43	0.011	0.086

ML application

The results from the application of the selected 4 ML on the training phase are shown in Fig. 4, where the time series of daily EF values are plotted; the values of the EF, for the all years considered in the study, ranged from 0.07 to 0.74: the comparison between observed and predicted EF shows that for all 4 models the predicted values were within the 95% confidence interval: with an average R^2 among the ML algorithms of 0.88 (±0.1). The R^2 and RMSE obtained for each ML model are reported in Fig. 4.

The results of the testing phase are shown instead in Fig. 5: the average R^2 between observed and predicted daily EF of 0.66 (±0.06). The results of the cross-validation analysis (see "Cross-validation analysis" section) are reported in Table 5: the R^2 obtained among all the iterations for each of the 4 adopted ML algorithms were fundamentally constant with a very narrow range between the min and max R^2 obtained.

The results confirm the superiority of the GPR rational quadratic model, as indicated by the highest R^2 and lowest RMSE values, compared to the other ML models.

Discussion

The different components of the land surface energy balance follow the diurnal trend of the incoming radiation forcing with different amplitude and phase characteristics (Gentine et al. 2011). The EF is considered as a

Table 4 Results of the evaluation criteria applied on the 4 ML algorithms selected during the training-and-testing phase

		R^2	RMSE	MAE	MSE
Training phase					
Machine learning model					
Support vector machine	Polynomial SVM	0.76	0.0530	0.0388	0.001
	Cubic Gaussian SVM	0.83	0.0540	0.0472	0.002
Gaussian process regression	Matern 5/2	0.94	0.0293	0.0236	0.000
	Rational quadratic	0.99	0.0134	0.0106	0.000
Testing phase					
Machine learning model					
Support vector machine	Cubic Gaussian SVM	0.66	0.0806	0.0678	0.004
	Polynomial SVM	0.70	0.0741	0.0662	0.004
Gaussian process regression	Matern 5/2	0.82	0.0593	0.0458	0.002
	Rational quadratic	0.72	0.0554	0.0418	0.001

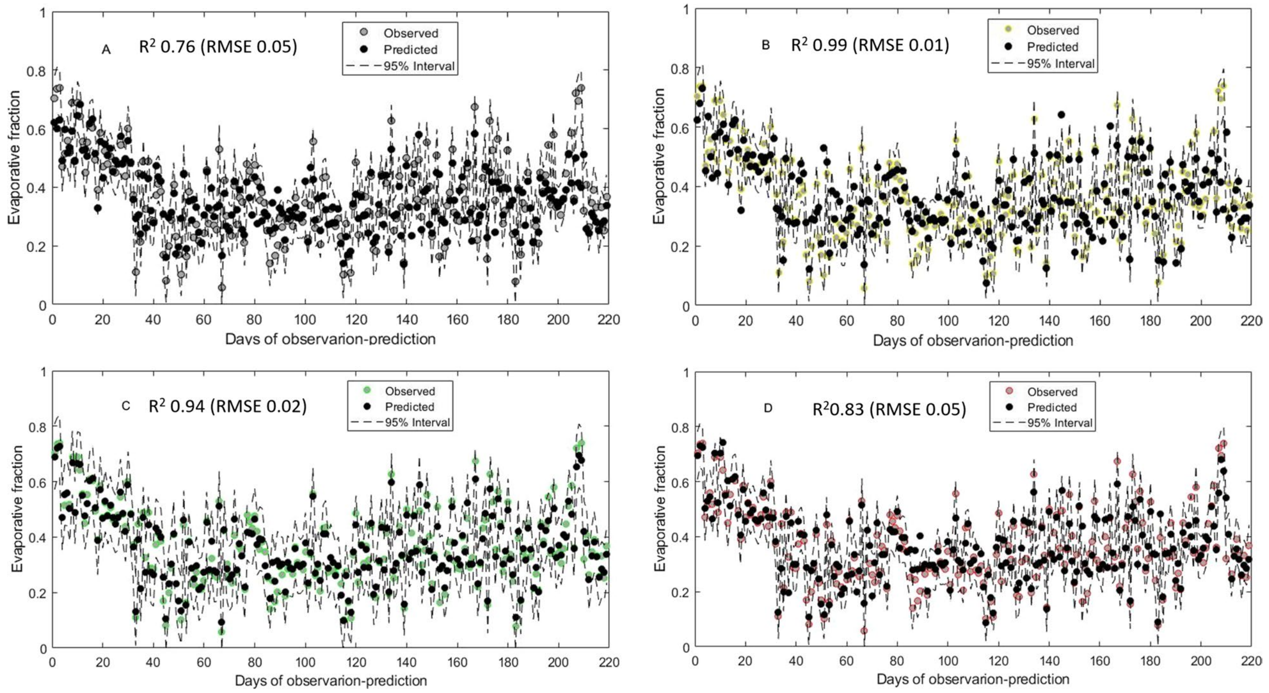

Fig. 4 Comparison between observed and predicted daily EF, during the training phase, of the four different approaches: SVM polynomial model (**A**), GRP rational quadratic (**B**), GRP Matern 5/2 (**C**), and SVM cubic Gaussian (**D**). Colored circles represent the observed values, black circles the predicted values, and the dashed lines represent the 95% confidence interval

diagnostic of the surface energy balance and is rarely constant during daytime (Gentine et al. 2011). Daytime self-preservation of the EF is mainly due to the high level of solar radiation around midday, while it is sensitive to the presence of warm and dry air above the atmospheric boundary layer (Gentine et al. 2011), typical of drought periods and therefore has been used as water deficit index (Hu et al. 2019; Schwalm et al. 2010) or to determine the daily evapotranspiration (Liu et al. 2020). Water availability increases EF values (Lhomme and Elguero 1999; Schwalm et al. 2010), and under concurrent conditions of a cloudless sky, it contributes to daily trends of EF featuring a typical concave shape. On the other hand, when water is the limiting factor, an enhancement of EF in the late afternoon is often found from field studies (Gentine et al. 2007).

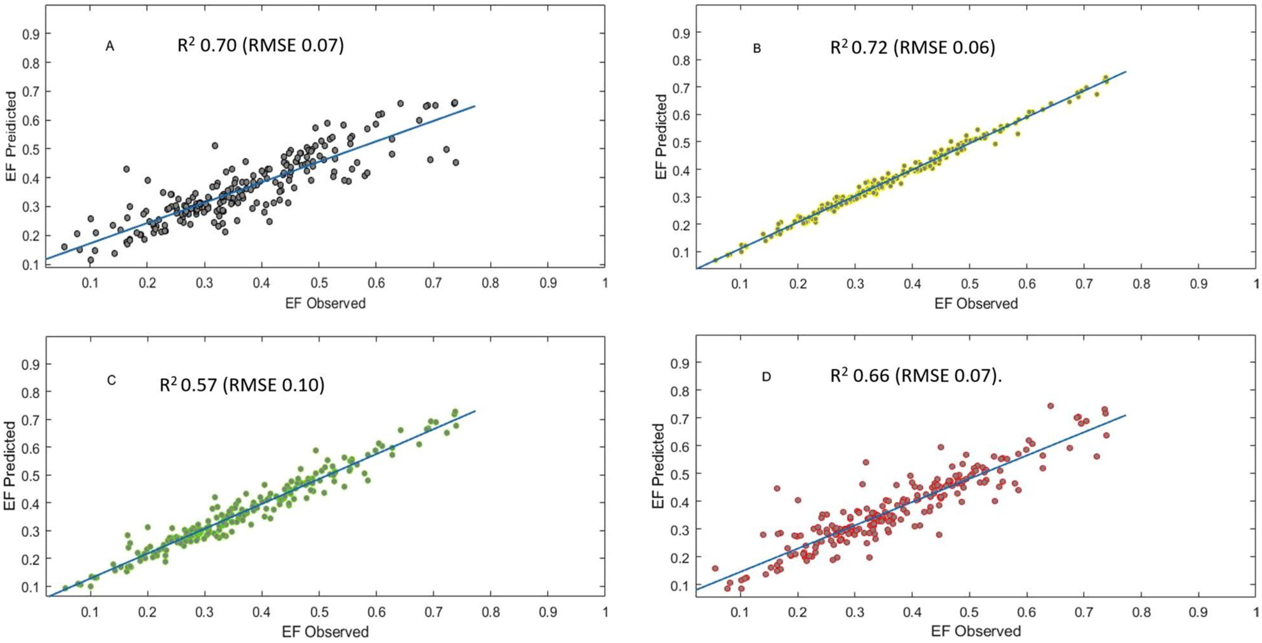

Fig. 5 Scatter plot between observed and predicted EF, during the testing phase of the four different approaches: the SVM polynomial model (**A**), GRP rational quadratic (**B**), GRP Matern 5/2 (**C**), and SVM cubic Gaussian (**D**)

Table 5 R^2 of the *K*-fold cross-validation analysis

***K*-fold iteration**	**ML algorithms**			
	Cubic Gaussian SVM	**Polynomial SVM**	**Matern 5/2**	**Rational quadratic**
1st iteration	0.84 (0.052)	0.76 (0.057)	0.94 (0.03)	0.96 (0.02)
2nd iteration	0.82 (0.056)	0.75 (0.05)	0.91 (0.03)	0.90 (0.03)
3rd iteration	0.82 (0.053)	0.76 (0.05)	0.94 (0.03)	0.97 (0.02)
4th iteration	0.75 (0.067)	0.75 (0.05)	0.93 (0.03)	0.97 (0.02)
5th iteration	0.82 (0.056)	0.75 (0.05)	0.91 (0.03)	0.97 (0.02)
6th iteration	0.81 (0.056)	0.74 (0.05)	0.94 (0.03)	0.89 (0.04)
7th iteration	0.82 (0.054)	0.76 (0.05)	0.90 (0.03)	0.97 (0.02)
8th iteration	0.83 (0.053)	0.75 (0.05)	0.91 (0.03)	0.96 (0.02)
9th iteration	0.80 (0.057)	0.76 (0.05)	0.93 (0.03)	0.96 (0.02)
10th iteration	0.82 (0.052)	0.76 (0.05)	0.92 (0.03)	0.93 (0.03)
R^2 mean	0.81	0.75	0.92	0.94
RMSE mean	0.05	0.05	0.03	0.02

RMSE values are in parentheses

It is generally accepted (Zhou and Wang 2016; Liu et al. 2019; Liu et al. 2020, Peng et al. 2013) to refer operationally to $EF_{daytime}$, i.e., the EF determined during daylight hours: considering that LE and H are near zero during nighttime, the estimated EF values can fluctuate greatly and can therefore be of little practical use. Moreover, during the transition periods from daytime to nighttime, the hourly EF is likely to be very unstable, and most of the times it fluctuates abruptly. For this reason, only $EF_{daytime}$ values were considered and simulated in this study.

When not available from direct observation, the EF is commonly simulated using a wide range of remote-sensing-based modeling schemes that have been proposed over past few decades, and goes from simple empirical formulas to complex land surface process models (Norman et al. 1995; Bastiaanssen et al. 1998; Su 2002; Nishida et al. 2003; Allen et al. 2007; Anderson et al. 2007; Mu et al. 2007; Bateni et al. 2013; Xu et al. 2014). Moreover, several studies have documented the advantages of the surface temperature-vegetation index method, and its applicability in the estimation of EF has been tested in several different regions across the world. (Rahimzadeh-Bajgiran et al. 2012; Yang et al. 2015; Liu et al. 2017; Hu et al. 2018; Carlson and Petropoulos 2019).

Our results demonstrate a potential for using ML to predict the EF. When tested against 6 years of EF measurements, the selected ML models were able to explain up to 99% (on the best case, and 81% on average) of the observed variability of the daily EF values.

When compared to ground observations of EC measurements, the remote-sensing-based EF model showed results comparable with the ML algorithms proposed in our study. Peng and Loew (2014) reported, for cropland sites, an R^2 ranging from 0.84 to 0.79 between observed and MODIS-TOA EF. Lu et al. (2013b) developed a method for estimating daily EF derived by day–night differences in surface temperature, air temperature, and net radiation showing a good agreement with measurements from an EC system corrected by the residual energy method with an R^2 of 0.857. Zhu et al. (2020) proposed a revised version of the surface temperature-vegetation index to retrieve simultaneously soil moisture and EF, showing an R^2 around 0.70 between observed and estimated EF. Remote-sensing-based EF models can therefore be considered the most widely used methodology to derive EF, especially as they allow the estimate at different spatio-temporal scales, including areas not covered by experimental network sites (Zhu et al. 2020); however, they require complex parameterization schemes and clear sky conditions which are not always achievable. The comparable performance of the ML algorithms proposed in our study with conventional remote-sensing-based EF models shows the potential of ML algorithms as a valid alternative to the conventional remote-sensing models.

The ability of ML algorithms to predict any target variable depends on the functional relationships between predictor variables and the target variable itself, as learned from the data rather than depending on an underlying process-level understanding (Breiman 2001). The ability of ML algorithms to predict the target variable (in our case, EF), is therefore correlated with the input variables available (in our case NEE, SWC, T_a, VPD and G). This introduces a potential limit in predicting the effects of novel conditions, which instead does not affect conventional models, with their ability to reproduce complex biophysical interactions. Nevertheless, ML algorithms might prove very useful when remote sensing data are not available. Indeed, in the ML models used in this study, key predictor variables are few (see Table 1): two weather variables (T_a and VPD), one component of the energy balance (G), SWC, and CO_2 NEE. All these factors are well known drivers of EF, and conventionally measured in experimental sites within research infrastructure networks, such as ICOS, Fluxnet, NEON and LTER.

Previous studies by Williams et al. (2015) and Puma et al. (2013) focused on the influence of vegetation on EF showing that the state of the crop canopy (e.g., LAI) exercises a stronger control on EF than SWC, and that LAI variability led to seasonal differences in LE and H, and thus EF. Bagley et al. (2017) report that differences in LAI between winter wheat and grassland/pasture led to differences in LE and H, and therefore on the magnitude of the observed EF. In our study, the application of both the neighborhood component analysis and the minimum redundancy maximum relevance indicate a negligible contribution of LAI as predictor, and therefore it was not included among the predictor variables. Moreover, including the LAI among the predictors resulted only in a limited improvement on the ML models: in particular the R^2 increased from 0.72 to 0.77 for the GPR Matern 5/2; from 0.75 to 0.79 for the GPR Rational quadratic; from 0.68 to 0.71 for the polynomial SVM; for the cubic Gaussian SVM instead, we observed a reduction of the R^2 from 0.72 to 0.67 (data not shown). Another important variable, able to predict up to 40% of the EF variability is the normalized difference vegetation index (NDVI), as reported by Zhou and Wang (2016) in a study over a series of agricultural sites within the Ameriflux network. In this study, we have intentionally avoided using NDVI, or other predictors that cannot be easily measured at site level from conventional meteorological observations, in the attempt to produce a simple model to predict the EF from field observations.

While our challenge was to test the ML models ability to reliably reproduce the EF at a specific site, it would be naïve to believe that the proposed ML algorithms from this study might be successfully applied to different soils, climates, or crop production systems considering that they were trained only on a dataset from a single corn crop site. Further research in this direction should focus on analyses including multiple sites across different terrestrial ecosystems, in order to extend the applicability of ML algorithms to effectively predict the EF.

Conclusions

From this study, daily EF predictions have been reliably derived for a corn crop in a Mediterranean region using ML algorithms. The application to other geographical areas and crops requires further improvements, applying model training based on diverse data sources from different soils, climate, cropping systems and agronomic managements. Our results show that support vector machine and Gaussian process regression algorithms are able, with limited input measurement data, to explain a substantial portion of the EF variability. The performance of the tested ML algorithms has proven to be comparable to the conventional remote sensing-based models and can be used when the sky conditions are not suitable for remote sensing observations. In addition, ML algorithms facilitate the interpretation of interactions between the predictors and the EF. Our results also suggest that in principle the integration of ML algorithms with remote sensing-based models could be an opportunity to improve the predictability of EF.

Acknowledgements
The authors thank Dr Paul di Tomasi for the technical support of the EC instrumentation, and Dr Andrea Esposito for the informatic support of the data acquisition. We thank also Mr Gaetano and Lazzaro Jemma owners of the "Jemma farm" who hosted our experimental EC installation

Author contributions
TZ: analyzed the data and wrote the manuscript; DF: reviewed the manuscript and provided the data; LV: reviewed the manuscript; VM: reviewed the manuscript. All authors read and approved the final manuscript.

Funding
The study was conducted under the aegis of the European Integrated Carbon Observation System (ICOS).

Availability of data and materials
Data are available upon reasonable request.

Declarations

Ethics approval and consent to participate
The manuscript does not involve human participants, human data or human tissue.

Consent for publication
The manuscript does not contain any individual person's data.

Competing interests
There are no financial and non-financial competing interests.

Author details
[1]National Research Council, Institute for Agricultural and Forestry Systems in the Mediterranean, P.Le Enrico Fermi 1, 80055 Portici, Italy. [2]National Research Council CNR, Institute for Bioeconomy, Via P. Gobetti, 101, 40129 Bologna, Italy.

Received: 3 May 2022 Accepted: 23 August 2022
Published online: 05 September 2022

References

Alexandridis T, Cherif I, Chemin Y, Silleos G, Stavrinos E, Zalidis G (2009) Integrated methodology for estimating water use in mediterranean agricultural areas. Remote Sens 1:445–465

Allen RG, Tasumi M, Trezza R (2007) Satellite-based energy balance for mapping evapotranspiration with internalized calibration (METRIC)—Model. J Irrig Drain Eng 133:380–394

Anderson MC, Norman JM, Mecikalski JR, Otkin JA, Kustas WP (2007) A climatological study of evapotranspiration and moisture stress across the continental United States based on the thermal remote sensing: 1. Model formulation. J Geophys Res-Atmos 112:D10117

Anurag M, Yazid T, Nadhir AA, Shamsuddin ShH, Harkanwaljot S, Sekhon RKP, Priya Rai KP, Padam S, Ahmed E, Saad S (2021) Daily pan-evaporation estimation in different agro-climatic zones using novel hybrid support vector regression optimized by Salp swarm algorithm in conjunction with gamma test. Eng Appl Comput Fluid Mech 15(1):1075–1094

Aybar-Ruiz A, Jiménez-Fernández S, Cornejo-Bueno L, Casanova-Mateo C, Sanz-Justo J, Salvador-González P, Salcedo-Sanz S (2016) A novel grouping genetic algorithm - extreme learning machine approach for global solar radiation prediction from numerical weather models inputs. Sol Energy 132:129–142

Bagley JE, Kueppers LM, Billesbach DP, Williams IN, Biraud SC, Torn MS (2017) The influence of land cover on surface energy partitioning and evaporative fraction regimes in the US Southern Great Plains. J Geophys Res-Atmos 122(11):5793–5807

Baldocchi DD (2020) How eddy covariance flux measurements have contributed to our understanding of global change biology. Glob Change Biol 26(1):242–260

Ballabio C, Lugato E, Fernández-Ugalde O, Orgiazzi A, Jones A, Borrelli P, Panagos P (2019) Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. Geoderma 355:113912

Barboza F, Kimura H, Altman E (2017) Machine learning models and bankruptcy prediction. Expert Syst Appl 83:405–417

Bastiaanssen WGM, Menenti M, Feddes RA, Holtslag AAM (1998) A remote sensing surface energy balance algorithm for land (SEBAL)—1. Formulation. J Hydrol 213:198–212

Bateni SM, Entekhabi D, Jeng DS (2013) Variational assimilation of land surface temperature and the estimation of surface energy balance components. J Hydrol 481:143–156

Borges L, Fernando F, da Cunha F (2020) New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agric Water Manag 234:106113

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Carlson TN, Petropoulos GP (2019) A new method for estimating of evapotranspiration and surface soil moisture from optical and thermal infrared measurements: the simplified triangle. Int J Remote Sens 40(20):7716–7729

Dash SS, Nayak SK, Mishra D (2021) A review on machine learning algorithms. Intell Cloud Comput 495–507.

de Tomás A, Nieto H, Guzinski R, Salas J, Sandholt I, Berliner P (2014) Validation and scale dependencies of the triangle method for the evaporative fraction estimation over heterogeneous areas. Rem Sens Environ 152:493–511

Dou X, Yang Y (2018) Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems. Comput Electron Agricult 148:95–106

Eugster W, Senn WA (1995) Cospectral correction model for measurement of turbulent NO_2 flux. Bound-Layer Meteorol 74(4):321–340

Fang K, Shen C, Kifer D, Yang X (2017) Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. Geophys Res Lett 44:11030–11039

Foken T (2008) The energy balance closure problem—an overview. Ecol Appl 18:1351–1367

Fu T, Li X, Jia R, Feng L (2021) A novel integrated method based on a machine learning model for estimating evapotranspiration in dryland. J Hydrol 603:126881

Gentine P, Entekhabi D, Chehbouni A, Boulet G, Duchemin B (2007) Analysis of evaporative fraction diurnal behaviour. Agric For Meteorol 143(1–2):13–29

Gentine P, Entekhabi D, Polcher J (2011) The diurnal behavior of evaporative fraction in the soil–vegetation–atmospheric boundary layer continuum. J Hydrometeorol 12(6):1530–1546

Göckede M, Rebmann C, Foken T (2004) A combination of quality assessment tools for eddy covariance measurements with footprint modelling for the characterization of complex sites. Agric For Meteorol 127:175–188

Guevara-Escobar A, González-Sosa E, Cervantes-Jiménez M, Suzán-Azpiri H, Queijeiro-Bolaños ME, Carrillo Ángeles I, Cambrón-Sandoval VH (2020) Eddy covariance carbon flux in a scrub in the Mexican highland.Biogeosci Discuss 2020:1-16

Hollinger DY, Richardson AD (2005) Uncertainty in eddy covariance measurements and its application to physiological models. Tree Physiol 25:873–885

Horst TW, Lenschow DH (2009) Attenuation of scalar fluxes measured with spatially displaced sensors. Bound-Layer Meteorol 130:275–300

Hu X, Shi L, Lin L, Zhang B, Zha Y (2018) Optical-based and thermal-based surface conductance and actual evapotranspiration estimation, an evaluation study in the North China Plain. Agric For Meteorol 263:449–464

Hu X, Shi L, Lin L, Zha Y (2019) Nonlinear boundaries of land surface temperature–vegetation index space to estimate water deficit index and evaporation fraction. Agric For Meteorol 279:107736

Jo I, Lee S, Oh S (2019) Improved measures of redundancy and relevance for mRMR feature selection. Computers 8(2):42

Kang J, Schwartz R, Flickinger J, Beriwal S (2015) Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. Int J Radiat Oncol Biol Phys 93:1127–1135

Kolle O, Rebmann C (2007) Eddysoft-documentation of a software package to acquire and process eddy covariance data. Tech Rep Max Planck Inst Biogeochem 10:88

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G (2007) Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35:345–349

Lhomme JP, Elguero E (1999) Examination of evaporative fraction diurnal behaviour using a soil-vegetation model coupled with a mixed-layer model. Hydrol Earth Syst Sci 3:259–270

Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D (2018) Machine learning in agriculture: A review. Sensors 18(8):2674

Liu L, Liao J, Chen X, Zhou G, Su Y, Xiang Z, Wang Z, Liu X, Li Y, Wu J, Xiong X, Shao H (2017) The microwave temperature vegetation drought index (MTVDI) based on AMSR-E brightness temperatures for long-term drought assessment across China (2003–2010). Remote Sens Environ 199:302–320

Liu Q, Wang T, Han Q, Sun S, Liu CQ, Chen X (2019) Diagnosing environmental controls on actual evapotranspiration and evaporative fraction in a water-limited region from northwest China. J Hydrol 578:124045

Liu X, Xu J, Zhou X, Wang W, Yang S (2020) Evaporative fraction and its application in estimating daily evapotranspiration of water-saving irrigated rice field. J Hydrol 584:124317

López-Cortés XA, Nachtigall FM, Olate VR, Araya M, Oyanedel S, Diaz V, Jakob E, Ríos-Momberg M, Santos LS (2017) Fast detection of pathogens in salmon farming industry. Aquaculture 470:17–24

Lu J, Li ZL, Tang R, Tang BH, Wu H, Yang F, Zhou G (2013a) Evaluating the SEBS-estimated evaporative fraction from MODIS data for a complex underlying surface. Hydrol Process 27(22):3139–3149

Lu J, Tang R, Tang H, Li ZL (2013b) Derivation of daily evaporative fraction based on temporal variations in surface temperature, air temperature, and net radiation. Remote Sens 5(10):5369–5396

Mosre J, Suárez F (2021) Actual evapotranspiration estimates in arid cold regions using machine learning algorithms with in situ and remote sensing data. Water 13(6):870

Milano M, Ruelland D, Fernandez S, Dezetter A, Fabre J, Servat E, Fritsch JM, Ardoin-Bardin S, Thivet G (2013) Current state of Mediterranean water resources and future trends under climatic and anthropogenic changes. Hydrol Sci J 58:498–518

Moncrieff JB, Clement R, Finnigan J, Meyers T (2004) Averaging, detrending and filtering of eddy covariance time series. In: Lee X, Massman WJ, Law BE (eds) Handbook of micrometeorology: a guide for surface flux measurement and analysis. Kluwer Academic Publisher, Dordrecht, pp 7–32

Mu Q, Heinsch F, Zhao M, Running S (2007) Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. Remote Sens Environ 111:519–536

Nguyen TPL, Mula L, Cortignani R, Seddaiu G, Dono G, Virdis SG, Roggero PP (2016) Perceptions of present and future climate change impacts on water availability for agricultural systems in the western Mediterranean region. Water 8(11):523

Nishida K, Nemani RR, Running SW, Glassy JM (2003) An operational remote sensing algorithm of land surface evaporation. J Geophys Res-Atmos 108(D9):4270

Norman JM, Kustas WP, Humes KS (1995) A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature. Agric For Meteorol 77:263–293

Nutini F, Boschetti M, Candiani G, Bocchi S, Brivio PA (2014) Evaporative fraction as an indicator of moisture condition and water stress status in semi-arid rangeland ecosystems. Remote Sens 6(7):6300–6323

Op de Beeck M, Sabbatini S, Papale D (2017) ICOS ecosystem instructions for soil meteorological measurements (TS, SWC, G) (Version 20180615). ICOS Ecosystem Thematic Centre. https://doi.org/10.18160/1a28-gex6

Pan S, Pan N, Tian H, Friedlingstein P, Sitch S, Shi H, Running SW (2020) Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling. Hydrol Earth Syst Sci 24(3):1485–1509

Pastorello G, Trotta C, Canfora E et al (2020) The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci Data 7:225. https://doi.org/10.1038/s41597-020-0534-3

Peng J, Loew A (2014) Evaluation of daytime evaporative fraction from MODIS TOA radiances using FLUXNET observations. Remote Sens 6(7):5959–5975

Peng J, Borsche M, Liu Y, Loew A (2013) How representative are instantaneous evaporative fraction measurements for daytime fluxes? Hydrol Earth Syst Sci 17:3913–3919

Puma MJ, Koster RD, Cook BI (2013) Phenological versus meteorological controls on land-atmosphere water and carbon fluxes. J Geophys Res-Biogeosci 118:14–29. https://doi.org/10.1029/2012JG002088

Rahimzadeh-Bajgiran P, Omasa K, Shimizu Y (2012) Comparative evaluation of the Vegetation Dryness Index (DVI), the Temperature Vegetation Dryness Index (TVDI) and the improved TVDI (iTVDI) for water stress detection in semi-arid regions of Iran. ISPRS J Photogramm Remote Sens 68:1–12

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566(7743):195–204

Reichstein M, Falge E, Baldocchi D, Papale D, Aubinet M, Berbigier P, Valentini R (2005) On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Glob change Biol 11(9):1424–1439

Richardson AD, Hollinger DY, Burba GG et al (2006) A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. Agric For Meteorol 136:1–18

Richardson A, Signor BM, Lidbury BA, Badrick T (2016) Clinical chemistry in higher dimensions: machine-learning and enhanced prediction from routine clinical chemistry data. Clin Biochem 49:1213–1220

Rodríguez JD, Pérez A, Lozano JA (2010) Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 32:569–575. https://doi.org/10.1109/TPAMI.2009.187

Schuepp PH, Leclerc MY, MacPherson JI, Desjardins RL (1990) Footprint prediction of scalar fluxes from analytical solutions of the diffusion equation. Bound-Layer Meteorol 50(1):355–373

Schwalm CR, Williams CA, Schaefer K, Arneth A, Bonal D, Buchmann N, Reichstein M (2010) Assimilation exceeds respiration sensitivity to drought: a FLUXNET synthesis. Glob Change Biol 16(2):657–670

Sen PC, Hajra M, Ghosh M (2020) Supervised classification algorithms in machine learning: a survey and review. In: Mandal J, Bhattacharya D (eds) Emerging technology in modelling and graphics. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore

Seneviratne SI, Luthi D, Litschi M, Schar C (2006) Land–atmosphere coupling and climate change in Europe. Nature 443(7108):205–209

Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer Science & Business Media.

Su Z (2002) The surface energy balance system (SEBS) for estimation of turbulent heat fluxes. Hydrol Earth Syst Sc 6:85–99

Takahashi K, Kim K, Ogata T, Sugano S (2017) Tool-body assimilation model considering grasping motion through deep learning. Rob Auton Syst 91:115–127

Tang R, Li ZL (2017a) An improved constant evaporative fraction method for estimating daily evapotranspiration from remotely sensed instantaneous observations. Geophys Res Lett 44:2319–2326. https://doi.org/10.1002/2017GL072621

Tang R, Li Z-L (2017b) Estimating daily evapotranspiration from remotely sensed instantaneous observations with simplified derivations of a theoretical model. J Geophys Res-Atmos 122:10177–10190. https://doi.org/10.1002/2017JD027094

Tramontana G, Jung M, Schwalm CR, Ichii K, Camps-Valls G, Ráduly B, Papale D (2016) Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. Biogeosciences 13(14):4291–4313

Trenberth KE, Guillemot CJ (1996) Physical processes involved in the 1988 drought and 1993 floods in North America. J Clim 9:1288–1298

Vapnik V (1999) The nature of statistical learning theory. Springer Science & Business Media, Berlin

Vitale L, Di Tommasi P, Arena C, Fierro A, De Santo AV, Magliulo V (2007) Effects of water stress on gas exchange of field grown *Zea mays* L. in Southern Italy: an analysis at canopy and leaf level. Acta Physiol Plant 29(4):317–326

Vitale L, Di Tommasi P, D'Urso G, Magliulo V (2016) The response of ecosystem carbon fluxes to LAI and environmental drivers in a maize crop grown in two contrasting seasons. Int J Biometeorol 60(3):411–420

Wang D, Tan X (2017) Bayesian neighborhood component analysis. IEEE Trans Neural Netw Learn Syst 29(7):3140–3151

Wauters M, Vanhoucke M (2014) Support vector machine regression for project control forecasting. Autom Constr 47:92–106

Webb EK, Pearman G, Leuning R (1980) Correction of flux measurements for density effects due to heat and water vapour transfer. Q J R Meteorol Soc 106:85–100

Wildenhain J, Spitzer M, Dolma S, Jarvik N, White R, Roy M, Griffiths E, Bellows DS, Wright GD, Tyers M (2015) Prediction of synergism from chemical–genetic interactions by machine learning. Cell Syst 1:383–395

Williams IN, Torn MS (2015) Vegetation controls on surface heat flux partitioning, and land–atmosphere coupling. Geophys Res Lett 42(21):9416–9424

Xu T, Bateni SM, Liang S, Entekhabi D, Mao K (2014) Estimation of surface turbulent heat fluxes via variational assimilation of sequences of land surface temperatures from geostationary operational environmental satellites. J Geophys Res-Atmos 119(18):10780–10798

Yang D, He W, Chen HE, Lei HM (2013) Analysis of the diurnal pattern of evaporative fraction and its controlling factors over croplands in the Northern China. J Integr Agric 12(8):1316–1329

Yang Y, Long D, Guan H, Liang W, Simmons C, Batelaan O (2015) Comparison of three dual-source remote sensing evapotranspiration models during the MUSOEXE-12 campaign: revisit of model physics. Water Resour Res 51:3145–3165

Yin L, Tao F, Chen Y, Liu F, Hu J (2021) Improving terrestrial evapotranspiration estimation across China during 2000–2018 with machine learning methods. J Hydrol 600:126538

Zenone T, Fischer M, Arriga N, Broeckx LS, Verlinden MS, Vanbeveren S, Zona D, Ceulemans R (2015) Biophysical drivers of the carbon dioxide, water vapor, and energy exchanges of a short-rotation poplar coppice. Agric For Meteorol 209:22–35

Zhao WL, Gentine P, Reichstein M, Zhang Y, Zhou S, Wen Y, Qiu GY (2019) Physics-constrained machine learning of evapotranspiration. Geophys Res Lett 46(24):14496–14507

Zhou C, Wang K (2016) Biological and environmental controls on evaporative fractions at AmeriFlux sites. J Appl Meteorol Climatol 55(1):145–161

Zhu W, Jia S, Lall U, Cheng Y, Gentine P (2020) An observation-driven optimization method for continuous estimation of evaporative fraction over large heterogeneous areas. Remote Sens Environ 247:111887

Zveryaev II, Allan RP (2010) Summertime precipitation variability over Europe and its links to atmospheric dynamics and evaporation. J Geophys Res: Atmos 115(D12).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.