Prediction of evapotranspiration in a Mediterranean region using basic meteorological variables

Daniel Jato-Espino*[1], Susanne M. Charlesworth[2], Sara Perales-Momparler[3], Ignacio Andrés-Doménech[4]

[1] GITECO Research Group, Universidad de Cantabria, Av. de los Castros s/n, 39005 Santander, Spain
[2] Centre for Agroecology, Water and Resilience (CAWR), Coventry University, CV1 5FB, United Kingdom
[3] Green Blue Management, Avda. del Puerto, 180-1B, 46023 Valencia, Spain
[4] Instituto Universitario de Investigación de Ingeniería del Agua y Medio Ambiente (IIAMA), Universitat Politècnica de València, Cno. de Vera s/n, 46022 Valencia, Spain

E-mail addresses: jatod@unican.es (D. Jato-Espino), apx119@coventry.ac.uk (S. M. Charlesworth), sara.perales@greenbluemanagement.com (S. Perales-Momparler), igando@hma.upv.es (I. Andrés-Doménech),

** Corresponding author. Tel.: +34 942 20 39 43; Fax: +34 942 20 17 03.*

Abstract

A critical need for farmers, particularly those in arid and semi-arid areas is to have a reliable, accurate and reasonably accessible means of estimating the evapotranspiration rates of their crops in order to optimize their irrigation requirements. Evapotranspiration is a crucial process due to its influence on the precipitation that is returned to the atmosphere. The calculation of this variable often starts from the estimation of reference evapotranspiration, for which a variety of methods have been developed. However, these methods are very complex either theoretically and/or because of the large amount of parameters on which they are based, which makes the development of a simple and reliable methodology for the prediction of this variable important. This research combined three concepts such as cluster analysis, Multiple Linear Regression (MLR) and Voronoi diagrams to achieve that end. Cluster analysis divided the study area into groups based on its weather characteristics, whose locations were then delimited by drawing the Voronoi regions associated with them. Regression equations were built to predict daily reference evapotranspiration in each cluster using basic climate variables produced in forecasts made by meteorological agencies. Finally, the Voronoi diagrams were used again to regionalize the crop coefficients and calculate evapotranspiration from the values of reference evapotranspiration derived from the regression models. These operations were applied to the Valencian Region (Spain), a Mediterranean area which is partly semi-arid and

for which evapotranspiration is a critical issue. The results demonstrated the usefulness and accuracy of the methodology to predict the water demands of crops and hence enable farmers to plan their irrigation needs.

Keywords

Cluster analysis; Crop coefficient; Evapotranspiration; Multiple linear regression; Reference evapotranspiration; Voronoi diagrams

1. Introduction

Evapotranspiration (ET) is the sum of two processes whereby water is lost from the soil surface (evaporation) and from the crop (transpiration) (Aytek 2009). As such, it is an important factor in the formation of clouds and the occurrence of rainfall and plays a relevant role in several different water-related fields, including aquifer recharge (Healy and Scanlon 2010), ecosystem water balances (Sun et al. 2011), global circulation models (Dolman 1993), hydrology (Sorooshian et al. 1993), irrigation systems (Allen 2000; Bos et al. 2008), land surface modelling (Chen and Dudhia 2001) and water resource management (Biswas 2004). Despite its importance, ET is still one of the most misunderstood variables in the hydrological cycle and its characterization remains limited (Brutsaert 1982; Naoum and Tsanis 2003).

As a global average, ET is responsible for approximately 60% of the precipitation returned to the atmosphere, a figure that increases to up to 90% in arid and semi-arid regions (Brutsaert 2005). Therefore, its measurement is essential in agricultural terms for estimating crop water demand and managing irrigation systems. The calculation of ET is frequently preceded by the determination of reference evapotranspiration (ET_o) (López-Urrea et al. 2006), which is the rate at which available soil water is lost from a specific crop (Jensen et al. 1990) and which can be estimated using climate data (Xing et al. 2008).

There are many methods developed to determine ET_o based on climate data, but the FAO Penman-Monteith equation (Monteith 1981) has been recommended by the Food and Agriculture Organization (Allen et al. 1998) and the American Society of Civil Engineers (Allen et al. 2005) as the standard method for this calculation. This equation can be used worldwide without requiring any local adjustment thanks to its physical foundations, validated by the use of lysimeters (Gocic and Trajkovic 2010). In contrast, the main weak-

ness of the FAO Penman-Monteith (PM) method is the large amount of variables it contains, some of which might not be available in many locations, especially developing countries (Martinez and Thepadia 2010).

Several researchers have pointed to the need for simpler methods to estimate ET_o (George et al. 2002; Sabziparvar et al. 2010; Tabari and Talaee 2011). Since the relationships between ET_o and the climate variables on which it depends are nonlinear (Jackson 1985; Kumar et al. 2002; Parasumaran et al. 2007; Wang et al. 2007; Adamala et al. 2014), Artificial Neural Networks (ANNs), Adaptive Neuro Fuzzy Inference Systems (ANFIS) and Genetic Programming (GP) have been the main methods used during the last decades to model it. Kumar et al. (2002) and Adamala et al. (2014) concluded that ANNs outperformed the PM method for reproducing values of ET_o measured with lysimeters, based on the errors yielded by both approaches. Parasuraman et al. (2007), who went one step further and also included GP in the comparison, demonstrated that both this technique and ANNs performed better than the PM method. Similarly, the results achieved by Wang et al. (2008) and Traore et al. (2010) revealed that ANNs could reach higher accuracy than empirical models such as Hargreaves and Blaney-Criddle in the prediction of ET_o.

Despite the nonlinear nature of ET_o, the linear combination of climate variables has been found to provide a simpler and still reliable and accurate alternative to predict it. Hence, the results obtained by Tabari et al. (2012) indicated that the differences between Multiple Linear Regression (MLR) models and Multiple Nonlinear Regression (MNLR) models were almost negligible, to the extent that MLR outperformed MNLR when the number of predictors used was small. In the same line, the studies carried out by Jain et al. (2008), Mallikarjuna et al. (2013) and Ladlani et al. (2014), who compared the capability of MLR to estimate ET_o with that of nonlinear methods such as ANNs and ANFIS, suggested that the performance of both linear and nonlinear approaches was very similar. The predictive power of the models built by Sanford et al. (2013), which explained around 90% of the proportion of the variance in the ratio of ET over precipitation, also provided evidence of the potential of MLR to estimate this variable.

These previous studies show that although nonlinear methods can be slightly more accurate than MLR, the differences between both approaches might not be significant and the linear combination of climate variables can provide accurate predictions of ET_o. Furthermore, MLR are simpler and easier to understand and interpret than nonlinear techniques, which are frequently used as “black boxes” without having a clear perception of their internal workings. For instance, ANNs, which represent the most widely used nonlinear method to estimate ET_o, require a series of hidden layers to relate inputs and output that are often added arbitrarily to improve the accuracy of the prediction model. This might

lead to overfitting of the model and result in misleadingly high-quality estimates. Besides, ANNs do not directly yield equations to estimate future values of ET_o as MLR do. However, former applications of MLR to predict ET_o did not provide solid evidence of their potential for making new estimates. Moreover, they were limited to the prediction of ET_o and did not include any regionalization methodology to group different locations according to their meteorological characteristics, which together with the fact that they were not built according to data availability in weather forecasts precludes the calculation of ET and therefore the design of aprioristic irrigation strategies.

In this context, the aim of this paper was to build linear equations for the prediction of ET based on weather forecasts, so that users can estimate the water requirements of their crops and determine when and how much to irrigate. This was achieved through a methodology which combined three tools such as cluster analysis, MLR models and Voronoi diagrams to enable the estimation and regionalization of ET using basic meteorological variables. These tools were applied to the Valencian Region in Spain, a Mediterranean area with semi-arid climate zones wherein evapotranspiration is an essential factor in optimizing agricultural production.

2. Methodology

2.1. Framework

Evapotranspiration (ET) and reference evapotranspiration (ET_o) can be related through Eq. (1):

$$ET = ET_o \cdot K_c \tag{1}$$

where K_c is the single crop coefficient (dimensionless), which combines the effect of soil evaporation and crop transpiration into a single coefficient and is recommended for irrigation planning, design, management and scheduling (Allen et al. 1998). Since K_c averages evaporation and transpiration, a single crop coefficient is used to determine ET for weekly or longer periods (Allen et al. 1998). Based on findings from several researchers on the temporal scale of K_c for different crops under Mediterranean climate (Ferreira and Carr 2002; Williams et al. 2003; Testi et al. 2004; Amayreh and Al-Abed 2005; Martínez-Cob A. 2008; Villalobos et al. 2009), a monthly period was chosen for the estimation of this coefficient. This is a time horizon that suits the purpose of this research, since it allows the prediction of daily ET for every month.

The FAO PM method is used in Spain for calculating ET_o (Doorenbos and Pruitt 1976). The concept of ET_o was defined by the FAO as the rate of ET from an ideal 12 cm high grass reference crop with a fixed canopy of 70 s·m^{-1} and an albedo of 0.23 (Allen et al. 1998). This reference surface resembles an extensive and well-watered green grass cover of uniform height, actively growing and completely shading the ground (Droogers and Allen 2002). ET_o (mm) can be estimated through Eq. (2), once the aerodynamic and radiation terms derived from the PM equation are combined:

$$ET_o = \frac{0.408 \cdot \Delta \cdot (R_n - G) + \gamma \cdot \frac{900}{T + 273} \cdot U_2 \cdot (e_a - e_d)}{\Delta + \gamma \cdot (1 + 0.34 \cdot U_2)} \quad (2)$$

where R_n is net radiation at the crop surface (MJ·m^{-2}·d^{-1}), G is soil heat flux (MJ·m^{-2}·d^{-1}), T is mean temperature (ºC), U_2 is mean wind speed at 2 m above the ground (m·s^{-1}), $(e_a - e_d)$ is the difference between the actual (e_a) and saturation (e_d) vapor pressure (kPa), Δ is the slope of the vapour pressure curve (kPa·ºC^{-1}) and γ is the psychrometric constant (kPa·ºC^{-1}), computed as shown in Eq. (3) (Brunt 2011):

$$\gamma = 0.00163 \cdot \frac{P}{\lambda} \quad (3)$$

where P is atmospheric pressure (kPa) and λ is latent heat (MJ·kg^{-1}). Eqs. (2) and (3) reveal the complexity of the PM equation and the great amount of parameters required by it, some of which are not provided by meteorological agencies in their weather forecasts. Therefore, there is a justifiable need to develop alternative models to estimate ET using basic meteorological variables.

2.2. Overview

The Valencian Region is divided into three provinces: Alicante, Castellón and Valencia. Table 1 summarizes their main demographic and climate characteristics and indicates the number of valid agrometeorological stations located in each of them. The Spanish Ministry of Agriculture, Food and Environment (MAGRAMA) provides historical daily values of ET_o for these stations calculated using the FAO PM equation (see Eq. (2)).

Table 1. Main characteristics of the provinces forming the Valencian Region

Province	Population	Surface area (km^2)	Valid stations	Average Annual Precipitation (mm)	Average Annual Max Temperature (ºC)	Average Annual Min Temperature (ºC)
Alicante	1,934,127	5,816	16	311.1	23.3	13.2
Castellón	604,344	6,632	10	467.0	22.3	12.7
Valencia	2,578,719	10,763	23	474.9	23.0	13.8

However, conventional weather stations do not record all the information required to complete the equation, which also cannot be used to predict new values of ET, since it is not compatible with the variables that are presented in the daily Spanish Meteorological Agency weather forecasts (AEMET 2016). In accordance with the data included in these forecasting models, predictors that are made available include mean temperature (T_{mean}, ºC), maximum temperature (T_{max}, ºC), minimum temperature (T_{min}, ºC), mean relative humidity (RH_{mean}, %), maximum relative humidity (RH_{max}, %), minimum relative humidity (RH_{min}, %) and mean wind speed (WS_{mean}, $m \cdot s^{-1}$).

The four main steps carried out to develop a methodology capable of predicting ET for a single day in any month using basic meteorological variables are listed below:

- Acquisition of the daily datasets corresponding to the seven predictors for the 49 stations located in the whole region and their subsequent arrangement in months, according to the time horizon of K_c.

- Categorizing the weather stations based on their recorded values in relation to the predictors. Measures of central tendency and variability were used to characterize these stations for clustering.

- Development of regression equations to make predictions of daily ET_o for each month and cluster from the combination of the set of predictors.

- Delimitation of the boundaries associated with both the clusters previously obtained and the values of K_c for each station using Voronoi diagrams.

The fulfilment of these steps enabled daily ET to be determined by multiplying K_c by the regression equation built to estimate ET_o for the month and the cluster corresponding to the coordinates of the study area. The theoretical framework behind the tools on which these last three steps were based is described in the following subsections.

2.3. Cluster analysis

Cluster analysis, a term first introduced by Tryon (1939), is a multivariate data mining technique that uses different algorithms and methods to group objects based on their similarity. As a result, objects within a group are related to one another but unrelated to objects in other groups, so that the distinctness of the clusters increases as the similarity within a group and the difference between groups increase (Tan et al. 2005).

Even though the notion of "cluster" is clear, the definition of the threshold that differentiates two clusters has not been precisely defined. Consequently, many clustering methods have been developed over the years, each of them based on different working principles (Estivill-Castro and Yang 2004). Among them, k-means is one of the most popular algorithms to cluster large datasets in an efficient and simple way (Forgy 1965; MacQueen 1967; Wu et al. 2008).

The k-means algorithm seeks to partition a set of observations n into $k(\leq n)$ clusters by minimising the within-cluster sum of squares ($WCSS$), i.e. the sum of distances of each point in the cluster to its centroid. This algorithm proceeds according to the three following steps (Tan et al. 2005): (1) choose k initial centroids, where k is the number of clusters desired; (2) assign each observation to the closest cluster according to the Euclidean distance between them, i.e. the square root of the sum of their squared differences; and (3) update the centroid of each cluster based on the points assigned to it. The last two steps are repeated until the results converge and there are no point changes in the clusters. In other words, the algorithm stops when the centroids remain the same (Tan et al. 2005).

Two pairs of measures of central tendency and variability were proposed to characterize these variables for each weather station depending on whether they were normally distributed or not: mean ($\bar{x}$) and standard deviation (σ) or median ($\tilde{x}$) and interquartile range (IQR), respectively. The Shapiro-Wilk test (Shapiro and Wilk 1965), which has been found to be more reliable when checking normality than other commonly used tests such as Kolmogorov-Smirnov or Lilliefors (Shapiro et al. 1968), was selected for checking normality.

2.4. Multiple linear regression

Multiple Linear Regression (MLR) aims to model the relationship between two or more predictors (basic meteorological variables) and a predictand (ET_o) by fitting a linear equation to observed data (see Eq. (4)):

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \varepsilon \quad (4)$$

where y is the predictand expressed as a linear combination of a set of K predictors x_k, each of which is multiplied by a coefficient β_k that indicates its relative weight in the equation. The equation also includes a constant β_0 and a random component ε (the residuals) which explain everything that cannot be interpreted from the predictors.

The goodness-of-fit of a MLR model is often measured through the coefficient of determination (R^2) (Hirsch et al. 1993). The standard R^2 is useful to determine how well the model fits the original data, but has several limitations that compromise its validity to make predictions. It does not capture the influence of the number of predictors in fitting the model, so that the addition of a predictor always results in an increase in R^2. The adjusted R^2 arose as a modified version of the standard R^2 that compares the explanatory power of regression models built with different numbers of predictors. However, although this coefficient improves the reliability of R^2, it still cannot provide accurate predictions of new data, which is the main goal of this research. Another variant of the coefficient of determination, known as predictive R^2, was used to overcome this drawback by making estimates on new observations according to three steps: (1) remove each observation from the dataset, (2) estimate the regression equation without the removed observation and (3) determine how well the model predicts the removed observation. The goodness-of-fit of the models was also tested through the standard error of the regression (S), which represents the average distance from the observed values to the regression line.

Cook's distance was used to show the influence of each observation on the response values and identify erroneous measurements in the predictors (outliers). According to Eq. (5), an observation with a Cook's distance (D_i) larger than three times the mean Cook's distance is considered as an outlier (Stevens 2009):

$$D_i = \frac{\sum_{j=1}^{n} \left(z_j - z_{j(i)}\right)^2}{p \cdot MSE} \quad (5)$$

where z_j is the jth fitted response values, $z_{j(i)}$ is the jth fitted response value where the fit does not include observation i, p is the number of coefficients of the regression model and MSE is the mean squared error.

MLR is based on four assumptions that must be verified to ensure its validity: linearity, independence, homoscedasticity and normality. Violation to these assumptions was diagnosed through the residual plots and the Durbin-Watson statistic (Osbourne and Waters 2002).

2.5. Voronoi diagrams

The concept of Voronoi diagrams (Voronoi 1908), also known as Dirichlet tessellation (Dirichlet 1850) or Thiessen polygons (Thiessen and Alter 1911), consists of dividing a plane containing a series of points following the nearest-neighbor rule, so that each point belongs to the region of the plane closest to it (Aurenhammer 1991), called a Voronoi cell.

Analytically, if $X = \{x_1, x_2, \ldots, x_n\}$ is a set of point sites in the plane, then the Voronoi cell for a point site x_i ($VC(x_i)$) is defined as the set of points y in the plane that are closer to x_i than any other point site (see Eq. (6)):

$$VC(x_i) = \{y \mid d(x_i, y) < d(x_j, y), \forall j \neq i\} \quad (6)$$

where $d(x, y)$ denotes the Euclidean distance between the points x and y. From a graphical point of view, $VC(x_i)$ can also be defined in terms of the intersection of half-planes. The bisector of x and y is equal to the perpendicular line through the centre of the line segment $\overline{xy}$ and separates the plane into two half-planes. Therefore, the Voronoi diagram of X is the tuple of cells $VC(x_i \in X)$. More details about the properties of Voronoi diagrams can be found in Aurenhammer and Klein (2000).

3. Results and discussion

The study period for this research was between 2008 and 2014, since the former was the first year in which all the agrometeorological stations in the Valencian Region (see Table 1) started to work altogether. Figure 1 shows the location of this region in relation to the geography of Spain and the Mediterranean Sea and its division into the provinces of Alicante, Castellón and Valencia.

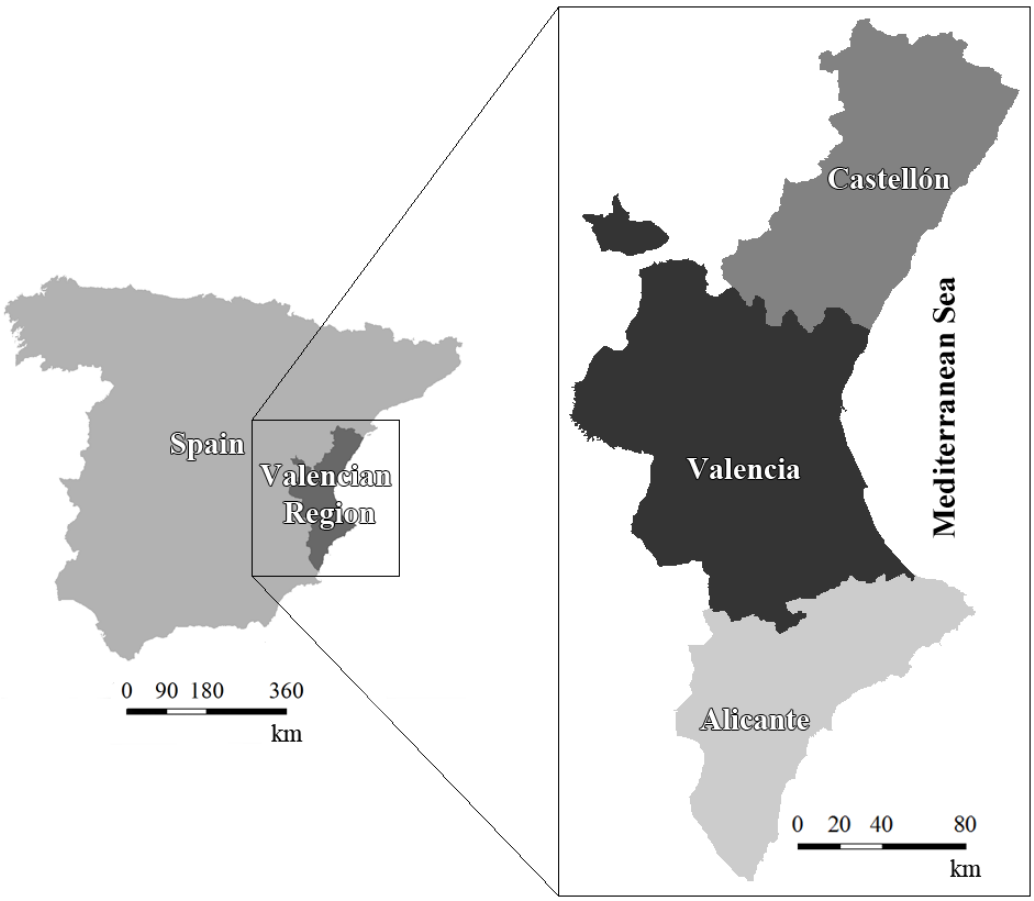


Figure 1. Location and provincial division of the Valencian Region

The first step in the methodology was the regionalization of the Valencian Region according to its weather characteristics, which were provided by the values taken by the basic meteorological variables to be used as predictors for building the regression models in the stations. Normality of this set of possible predictors was checked through the Shapiro-Wilk test, which revealed that the null hypothesis was rejected for all of them (p-values < 0.05). Hence, these variables were characterized for clustering through the median ($\tilde{x}$) and interquartile range (IQR) corresponding to each station. As an exploratory inspection of the variations in ET_o across the Valencian Region, Table 2 lists the monthly values of $\tilde{x}$ and IQR obtained after averaging the stations located in each of the three provinces forming it. The general trend of these data suggested that the highest values of ET_o were recorded in Alicante, which is characterised by having a drier climate than either Castellón or Valencia and therefore, higher temperatures coupled with lower humidity. The Köppen Climate Classification for the Iberian Península (Chazarra et al. 2011) confirmed this inference, since Alicante completely belongs to type B (dry), whereas Castellón and Valencia also have some type C areas (temperate).

Table 2. Average median ($\tilde{x}$) and interquartile range (IQR) of ET_o (mm/month) for each province

Province	Measure	Month											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Alicante	$\tilde{x}$	1.18	1.79	2.74	3.65	4.62	5.45	5.70	5.01	3.67	2.36	1.41	1.05
	IQR	0.70	0.85	1.18	1.41	1.17	0.99	0.75	0.94	1.19	0.86	0.71	0.49
Castellón	$\tilde{x}$	1.07	1.69	2.51	3.42	4.32	5.12	5.31	4.59	3.44	2.17	1.30	0.99
	IQR	0.64	0.85	1.06	1.28	1.32	1.02	0.89	1.10	1.19	0.88	0.63	0.48
Valencia	$\tilde{x}$	1.10	1.75	2.68	3.55	4.49	5.30	5.55	4.86	3.55	2.19	1.30	0.98
	IQR	0.81	1.05	1.32	1.47	1.40	1.13	0.77	0.97	1.30	0.94	0.79	0.60

Many different methods have been developed to optimize the determination of the number of clusters in a dataset, such as the gap statistic, Hartigan's approach or silhouette (Tibshirani et al. 2001). However, since cluster analysis preceded the development of the prediction models, the number of clusters chosen was calculated to maximize the predictive R^2 of subsequent regression equations. The results demonstrated that the optimal number of clusters was 1 in all cases except in May, June, July and August, when it was 2. In other words, the predictive R^2 was maximized for these clusters and then began to decrease its value gradually as the number of clusters increased.

Figure 2 illustrates the Voronoi regions obtained for each of these months from the pair of values $(\tilde{x}, IQR)$ calculated from each station. These were the warmer months of the year and those in which the combination of weather effects resulted in the highest and most varying values of ET (see Table 2), justifying the need to partition the whole workspace into two zones. The clustering patterns were consistent with that premise, since they separated the coastal and interior areas of the region, which were the zones wherein such variability became more accentuated.

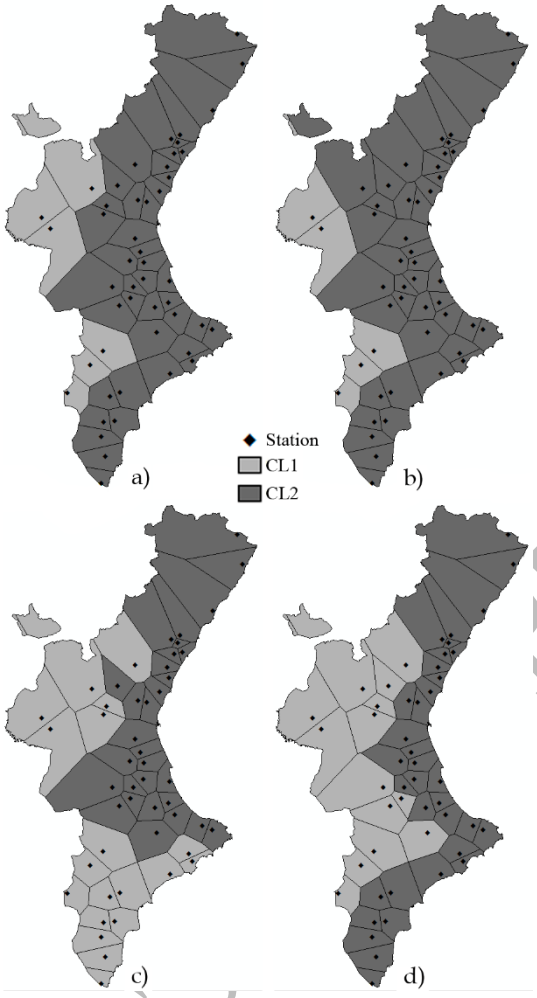


Figure 2. Clusters obtained for a) May b) June c) July d) August

From there, multiple linear regression models were built to estimate daily ET_o for each month and cluster by adapting Eq. (4) to the specifics of this research: $y = ET_o$ (mm/day); $x_1 = T_{mean}$ (ºC); $x_2 = T_{max}$ (ºC); $x_3 = T_{min}$ (ºC); $x_4 = RH_{mean}$ (%); $x_5 = RH_{max}$ (%); $x_6 = RH_{min}$ (%); $x_7 = WS_{mean}$ (m·s^{-1}). A 95% confidence interval (p-value < 0.05) was set to choose predictors stepwise, whilst Cook's distances were calculated using Eq. (5) to detect and remove influential points. Table 3 shows the regression coefficients and goodness-of-fit measures obtained for the number of days (N) corresponding to each month and cluster (CL) between 2008 and 2014.

Table 3. Summary of the regression models to predict ET_o (mm/day) for each month and cluster

Month	CL	N	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	S	Pred. R^2
1	1	8309	1.131	-	0.029	0.008	-	-0.007	-0.007	0.464	0.070	0.967
2	1	7473	1.126	-	0.079	-0.006	-	-0.009	-0.009	0.463	0.102	0.970
3	1	8440	1.021	-	0.124	-0.010	-	-0.009	-0.014	0.514	0.174	0.957
4	1	8151	0.881	0.307	-	-0.132	-	-0.007	-0.021	0.553	0.208	0.953
5	1	1027	1.638	-	0.174	-0.035	-0.016	-	-0.020	0.545	0.195	0.962
	2	7070	1.261	0.266	-	-0.114	-0.008	-	-0.016	0.722	0.211	0.915
6	1	967	1.410	0.246	-	-0.069	-	0.002	-0.035	0.616	0.202	0.952
	2	6817	1.969	0.243	-	-0.102	-	-0.007	-0.014	0.673	0.177	0.896
7	1	3092	2.320	0.041	0.077	-	-0.009	-	-0.013	0.830	0.156	0.936
	2	4864	2.751	0.191	-	-0.079	-	-0.007	-0.019	0.692	0.161	0.861
8	1	2467	-0.014	0.038	0.126	-	-0.012	-	-0.009	0.976	0.240	0.926
	2	6175	-0.735	0.353	-	-0.144	-	-0.009	-0.015	0.823	0.237	0.830
9	1	8372	-1.189	0.346	-	-0.149	-	-0.005	-0.018	0.853	0.191	0.945
10	1	8471	-0.452	0.222	-	-0.092	-	0.002	-0.020	0.628	0.168	0.921
11	1	8518	0.568	0.016	0.052	-	-	-0.006	-0.009	0.538	0.082	0.961
12	1	8260	0.755	-	0.028	0.009	-	-0.005	-0.006	0.497	0.043	0.980

The results were 16 regression equations consisting of 5 predictors in each case. Variations in the coefficients associated with the predictors (see Table 3) demonstrated the need to build monthly regression models for the prediction of ET_o, because weather attributes vary over the year (e.g. increased temperature in summer). Although the predictors included in each model varied in some cases depending on the month and cluster, all regression models consisted of two temperature-related variables ('T_{mean} AND T_{min}' OR 'T_{mean} AND T_{max}' OR 'T_{min} AND T_{max}'), two humidity-related variables ('RH_{mean} AND RH_{min}' OR 'RH_{mean} AND RH_{max}' OR 'RH_{min} AND RH_{max}') and WS_{mean}. The most influential predictors were found to be those related to temperature with an average contribution around 50% to estimate ET_o, except for the colder months, in which the combination of relative humidity and wind speed explained up to 80% of the variations in the predictand. The physical relationships between the mean predictors (x_1, x_4 and x_7), which are the most representative ones for each type of variable (temperature, humidity and wind), and the predictand were logical in all cases. The pores of plants in which water is released open if they are surrounded by warmer air, i.e. there is an increase in transpiration (Crawford et al. 2012). In contrast, relative humidity is inversely proportional to evapotranspiration, since the evaporation of water into the air is hindered as this becomes more saturated (Thut 1938). As for wind speed, moving air facilitates the process of evapotranspiration, since it is less saturated than stagnant air and can absorb water vapor more easily (Moore et al. 2003).

The reliability of the regression models for making predictions was guaranteed by the high and low values of predictive R^2 and S reached, respectively. The values of predictive R^2 indicated that these regression models can make estimates for new values of daily ET_o with an accuracy of at least 83% through a linear combination of basic variables related to temperature, humidity and wind. The ratio between S and the average monthly values of $\tilde{x}$ and IQR (see Table 3) was at most 7% and 25%, respectively, which demonstrates that the errors in the regression models were very small in relation to the typical values and spread of ET_o. The relationships between the climate variables used as predictors and ET_o were nonlinear in general. Figure 3 illustrates this circumstance for April, in which the predictand varied nonlinearly in relation to all predictors except RH_{min}, whose relationship to ET_o could be assumed to be linear. Therefore, these results confirmed that the linear combination of climate variables can provide accurate predictions of ET_o, even though their individual correlations are mostly nonlinear.

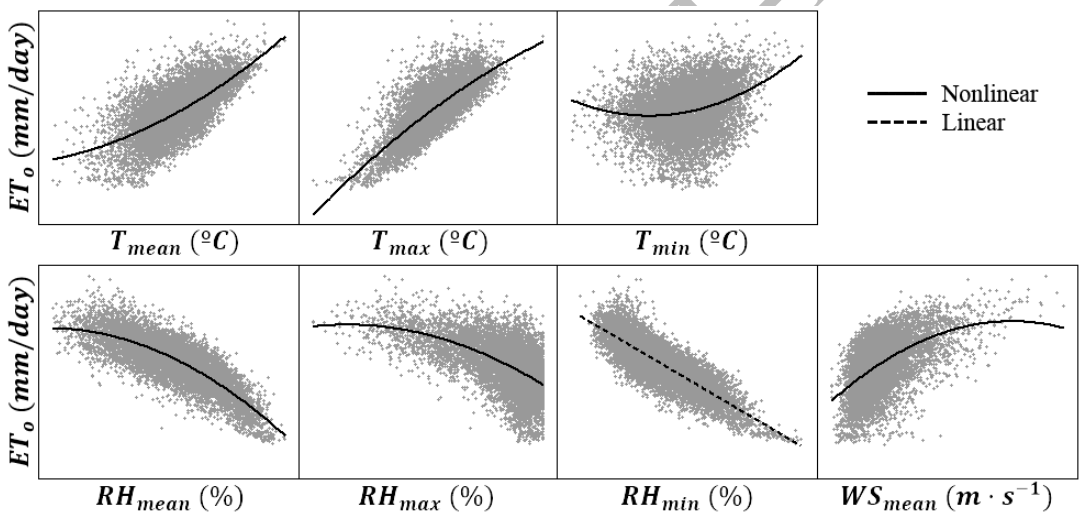


Figure 3. Relationships between the predictors and the predictand (ET_o) in the regression model for April

Figure 4 shows the histograms and scatterplots of standardized residuals against fitted values for two months representing different weather conditions (April (1 cluster) and June (2 clusters)), which provide graphical diagnose verifying whether the assumptions of normality, linearity and homoscedasticity were violated or not. The symmetrical bell-shape of the histograms, which fitted their corresponding theoretical normal curves with high accuracy, suggested that the normality assumption was valid. Moreover, the absence of curvilinear distributions and marked trends (e.g. increasing dispersion as the fitted values increase) in the scatterplots confirmed both the linearity and homoscedasticity of the residuals. Finally, the Durbin-Watson statistics were between 1.5 and 2.5 (Durbin and Watson 1950; Durbin and Watson 1951) in all three cases (1.740 for April and 1.596

(CL1) and 1.591 (CL2) for June), which implied that there was no time trends nor serial correlations in the residuals and their independence could be assumed too. Furthermore, the values of Variance Inflation Factor (VIF) obtained for the predictors, which were always below 10 (Belsley et al. 1980), ensured that they were not highly correlated to each other and multicollinearity was not an issue.

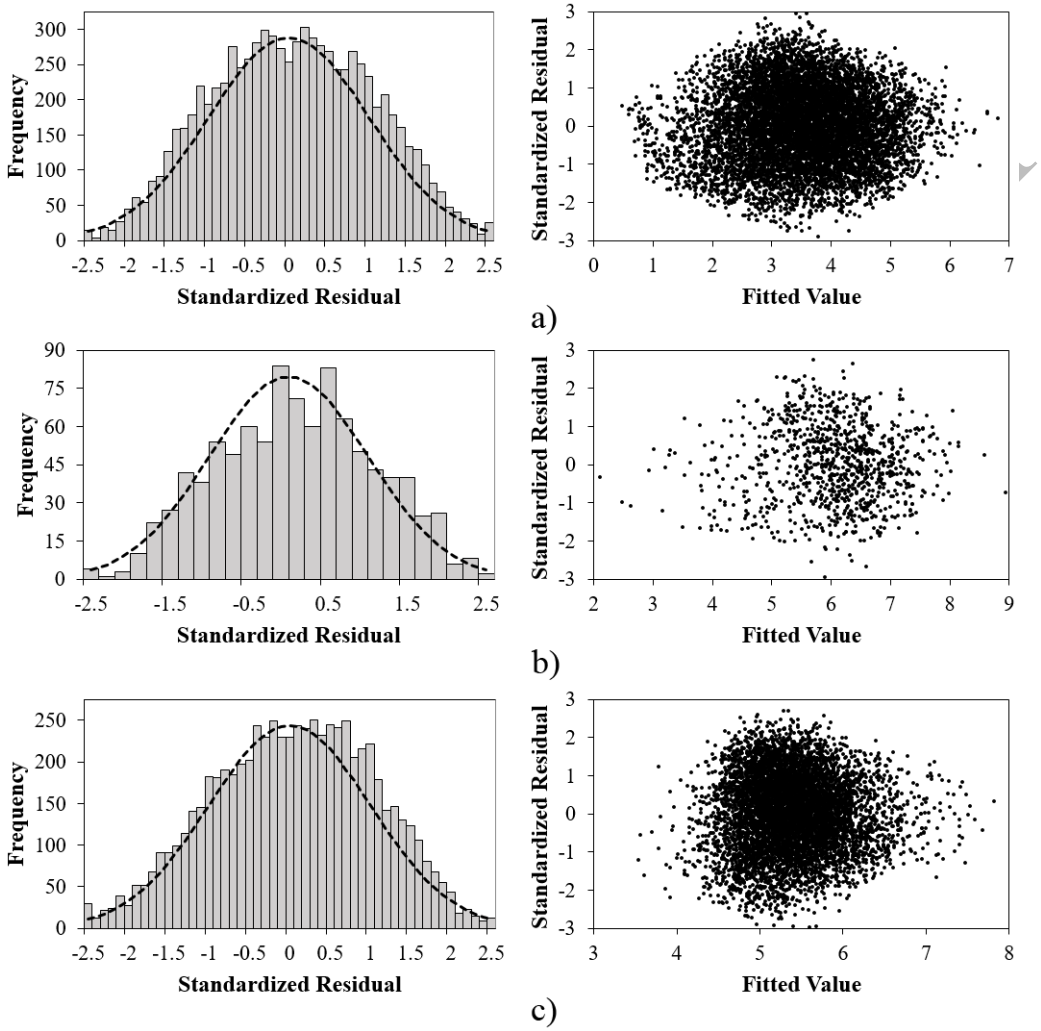
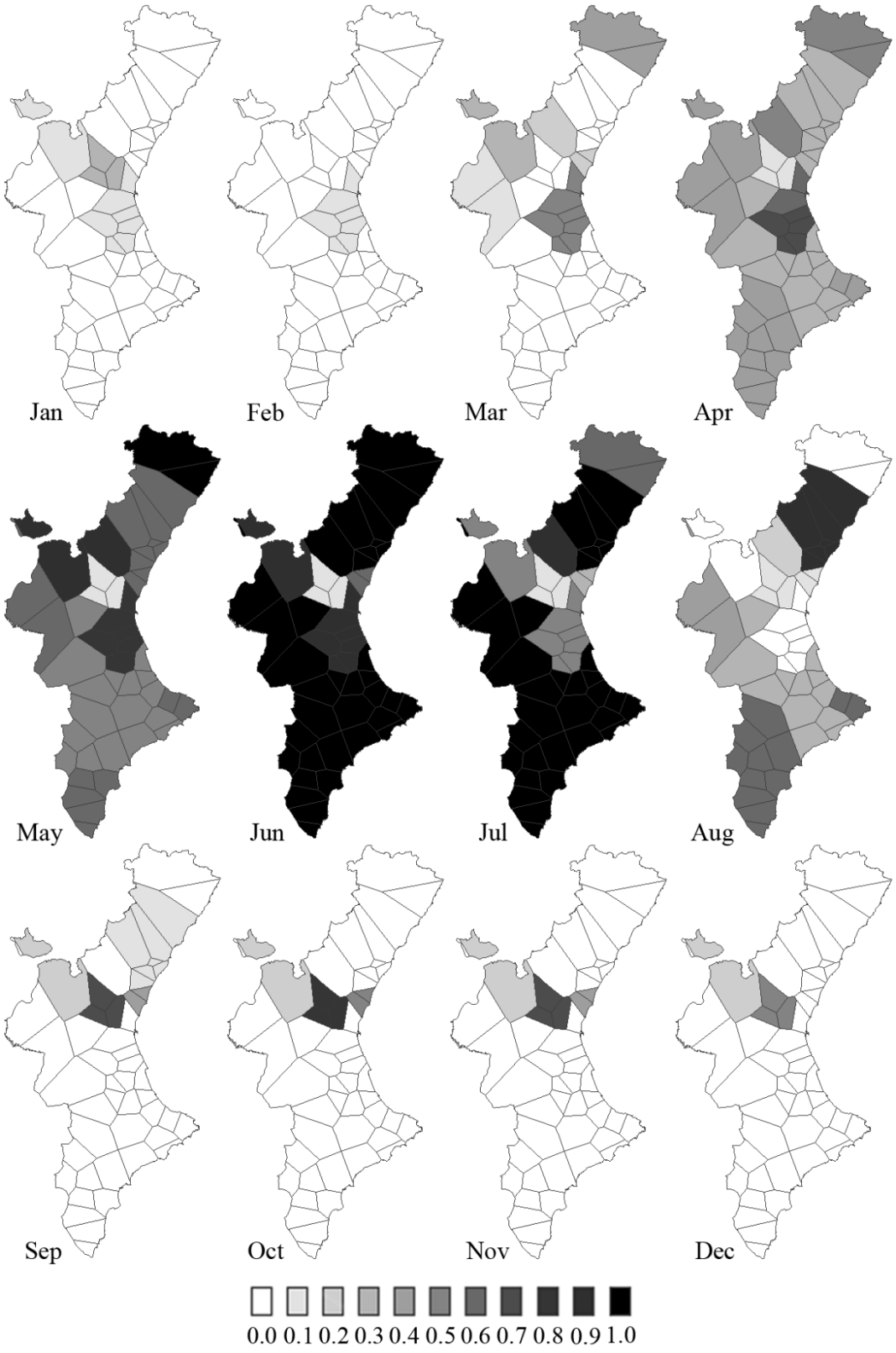


Figure 4. Histograms and scatterplots of standardized residuals against fitted values for a) April b) June - Cluster 1 c) June - Cluster 2

The final step was the regionalization of the Valencian Region according to the crop coefficients (K_c) in each station, in order to obtain a value for ET from ET_o using Eq. (1). Due to space constraints, this last process was limited to only one crop type: midseason potato. This specific crop was selected because it proved to be variable in terms of both location and time. The daily values of K_c provided by the MAGRAMA through its Agroclimatic Information System for Irrigation (SiAR 2016), which were constant for each

month during the years of study, reaffirmed the convenience of choosing a monthly period for the estimation of this coefficient. Therefore, the Voronoi regions were drawn as shown in Figure 5 according to the values of K_c for each station and month of the year. The procedure would be the same for any other crop, with the only difference that the Voronoi regions should be particularized to the monthly values of K_c associated with the specifics of the crop under study.

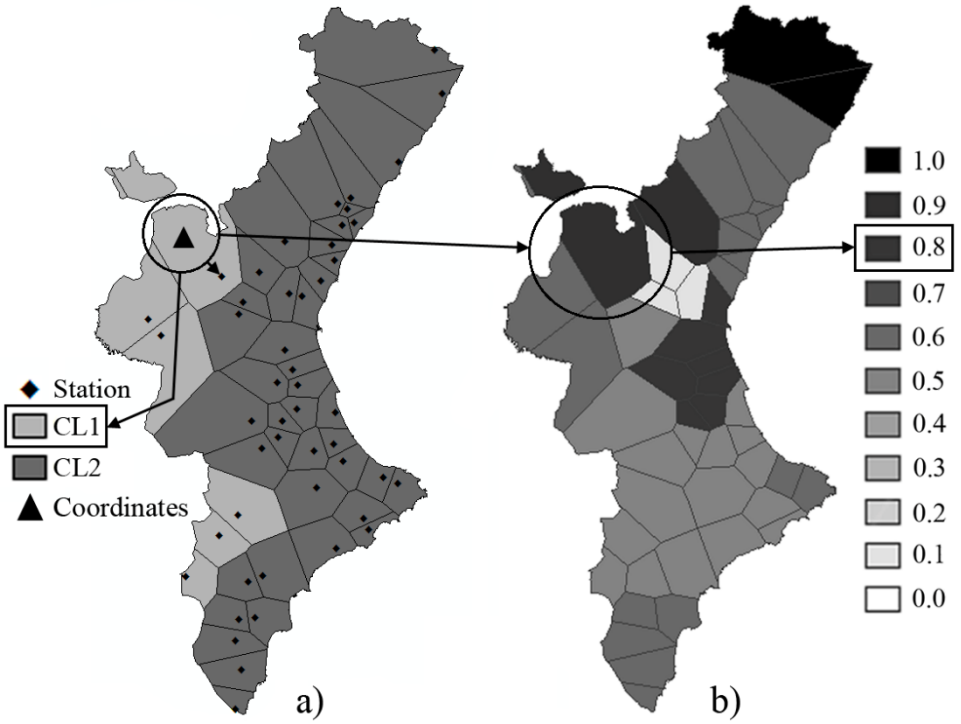


444

445 **Figure 5.** Monthly crop coefficients (K_c) in the Valencian Region for midseason potato

446

Knowing the coordinates for where irrigation was planned, the multiplication of crop coefficients in this area (see Figure 5) by the regression equations summarized in Table 3 enabled an estimation to be made of the water demands of this crop for a single day in any month using basic meteorological variables available from official weather forecasts. For instance, Figure 6 particularizes the procedure for the case of a farmer who planted midseason potatoes in April in the geographic coordinates (39º55'57'' N, 1º04'10'' W) and would like to estimate ET in a day in May. To illustrate the example, the historical average values for May recorded in the closest station to the specified coordinates were taken as the climate variables to be acquired from daily weather forecasts. According to the clusters identified in Figure 6a) and Figure 6b), these coordinates corresponded to CL1 and a Voronoi region with a value of K_c equal to 0.8. The application of the regression equation in Table 3 for these predictors, month and cluster yielded a value of ET_o of 4.66 mm/day. The multiplication of ET_o by K_c as formulated in Eq. (1) resulted in a final value of ET equal to 3.77 mm/day.



Month	T_{max} (ºC)	T_{min} (ºC)	RH_{mean} (%)	RH_{min} (%)	WS_{mean} ($m \cdot s^{-1}$)
May	23.84	11.23	61.62	32.45	1.66

c)

$$ET_o = 1.638 + 0.174 \cdot T_{max} - 0.035 \cdot T_{min} - 0.016 \cdot RH_{mean} - 0.020 \cdot RH_{min} + 0.545 \cdot WS_{mean} = 4.66\ mm/day$$

d)

$$ET = ET_o \cdot K_c = 3.73\ mm/day$$

e)

Figure 6. Estimation of ET in May for midseason potato in the coordinates (39º55'57'' N, 1º04'10'' W) a) Cluster b) Monthly crop coefficient (K_c) c) Historical average values for the predictors in the closest station to the coordinates d) Calculation of ET_o (mm/day) e) Determination of ET (mm/day)

4. Conclusions

This paper presents a methodology for the prediction of daily evapotranspiration based on the combination of cluster analysis, multiple linear regression models and Voronoi diagrams. The first was used to partition the study area according to its weather characteristics, so that regression equations to estimate daily reference evapotranspiration could be built for the resultant clusters using basic meteorological variables. Voronoi diagrams enabled regionalization of the workspace in terms of both clusters and crop coefficients

associated with it, whose multiplication by reference evapotranspiration yielded the value for real evapotranspiration which was being sought.

Despite the relationships between climate variables and reference evapotranspiration are generally nonlinear, the results proved that the linear combination of the former can provide accurate estimates of the latter. The models obtained using multiple linear regression analysis met the four hypotheses related to their residuals and reached high predictive coefficients of determination, which ensured their reliability and capability to make new estimates from daily weather forecasts. As for cluster analysis and Voronoi diagrams, their combination was found to be a simple and effective method for local application of the predictive regression equations and regionalization of crop coefficients, which enabled determining real evapotranspiration without any need to take into account complex physical considerations.

This methodology is proposed as a tool to be used by farmers for irrigation planning and scheduling based on the estimation of water demands of their crops. The daily value of evapotranspiration corresponding to a given date, coordinates and crop can be determined through the cluster, regression equation and crop coefficient associated with the day and region under study, since they are based on primary weather variables that are available from the daily forecasts made by meteorological agencies. Although the validity of these results is not compromised by the size of the study area, further research should consider the application of this methodology to larger locations, in order to delimit different climate zones and develop regional prediction equations at larger scales.

Acknowledgments

This paper was possible thanks to the research project RHIVU (Ref. BIA2012-32463), financed by the Spanish Ministry of Economy and Competitiveness with funds from the State General Budget (PGE) and the European Regional Development Fund (ERDF). The authors also wish to express their gratitude to the Spanish Ministry of Agriculture, Food and Environment (MAGRAMA) for providing the data necessary to develop this study.

References

Adamala, S., Raghuwanshi, N. S., Mishra, A., and Tiwari, M. K. (2014). "Evapotranspiration modeling using second-order neural networks." J.Hydrol.Eng., 19(6), 1131-1140.

AEMET. (2016). "Agencia estatal de meteorología." www.aemet.es (August, 2016).

512 Allen, R. G. (2000). "Using the FAO-56 dual crop coefficient method over an irrigated
513 region as part of an evapotranspiration intercomparison study." J.Hydrol., 229(1-2), 27-
514 41.

515 Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). "Crop evapotranspiration -
516 Guidelines for computing crop water requirements." FAO Irrigation and Drainage Pa-
517 per, 56(9), 1-174.

518 Allen, R. G., Walter, I. A., Elliott, R., Howell, T., Itenfisu, D., and Jensen, M. (2005).
519 "The ASCE standardized reference evapotranspiration equation." Rep. No. Final Re-
520 port, Environmental and Water Resources Institute (EWRI) of the American Society of
521 Civil Engineers (ASCE), Reston, Virginia (U.S.).

522 Amayreh, J., and Al-Abed, N. (2005). "Developing crop coefficients for field-grown to-
523 mato (Lycopersicon esculentum Mill.) under drip irrigation with black plastic mulch."
524 Agric.Water Manage., 73(3), 247-254.

525 Aurenhammer, F. (1991). "Voronoi Diagrams - A Survey of a Fundamental Geometric
526 Data Structure." ACM Computing Surveys, 23(3), 345-405.

527 Aurenhammer, F., and Klein, R. (2000). "Voronoi Diagrams." Handbook of Computa-
528 tional Geometry, V. J. Sack, and G. Urrutia, eds., Elsevier, Amsterdam (Netherlands),
529 201-290.

530 Aytek, A. (2009). "Co-active neurofuzzy inference system for evapotranspiration mod-
531 eling." Soft Comput., 13(7), 691-700.

532 Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). Regression Diagnostics: Identifying
533 Influential Data and Sources of Collinearity. John Wiley & Sons, New Jersey (U.S.).

534 Biswas, A. K. (2004). "Integrated water resources management: A reassessment." Water
535 Int., 29(2), 248-256.

536 Bos, M. G., Kselik, R. A. L., Allen, R. G., and Molden, D. (2008). Water requirements
537 for irrigation and the environment. Springer, Dordrecht (The Netherlands).

538 Brunt, D. (2011). Physical and dynamical meteorology. Cambridge University Press,
539 Cambridge (U.K.).

540 Brutsaert, W. (2005). Hydrology: An Introduction. Cambridge University Press, Cam-
541 bridge (U.K.).

Brutsaert, W. H. (1982). Evaporation into the Atmosphere. D. Reidel, Dordrecht (The Netherlands).

Chazarra, A., Mestre Barceló, A., Pires, V., Cunha, S., Mendes, M., and Neto, J. (2011). "Atlas Climático Ibérico. Agencia Estatal de Meteorología and Instituto de Meteorología de Portugal." Rep. No. Catálogo General de publicaciones oficiales, Agencia Estatal de Meteorología (AEMET) and Instituto de Meteorologia de Portugal, Madrid (Spain) and Lisboa (Portugal).

Chen, F., and Dudhia, J. (2001). "Coupling and advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity." Mon.Weather Rev., 129(4), 569-585.

Crawford, A. J., McLachlan, D. H., Hetherington, A. M., and Franklin, K. A. (2012). "High temperature exposure increases plant cooling capacity." Curr.Biol., 22(10), R396-R397.

Dirichlet, G. L. (1850). "Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen." J. Reine Angew. Math., 40 209-227.

Dolman, A. J. (1993). "A multiple-source land surface energy balance model for use in general circulation models." Agric.for.Meterol., 65(1-2), 21-45.

Doorenbos, J., and Pruitt, W. O. (1976). "Guidelines for predicting crop water requirements." Rep. No. FAO Irrigation and Drainage Paper 24, Food and Agriculture Organization of the United Nations, Rome (Italy).

Droogers, P., and Allen, R. G. (2002). "Estimating reference evapotranspiration under inaccurate data conditions." Irrig.Drain.Syst., 16(1), 33-45.

Durbin, J., and Watson, G. S. (1951). "Testing for serial correlation in least squares regression. II." Biometrika, 38(1-2), 159-177.

Durbin, J., and Watson, G. S. (1950). "Testing for serial correlation in least squares regression. I." Biometrika, 37(3-4), 409-428.

Estivill-Castro, V., and Yang, J. (2004). "Fast and robust general purpose clustering algorithms." Data Min.Knowl.Discov., 8(2), 127-150.

570 Ferreira, T. C., and Carr, M. K. V. (2002). "Responses of potatoes (Solanum tuberosum
571 L.) to irrigation and nitrogen in a hot, dry climate I. Water use." Field Crops Res., 78(1),
572 51-64.

573 Forgy, E. W. (1965). "Cluster analysis of multivariate data: efficiency versus interpreta-
574 bility of classifications." Biometrics, 21 768-769.

575 George, B. A., Reddy, B. R. S., Raghuwanshi, N. S., and Wallender, W. W. (2002).
576 "Decision support system for estimating reference evapotranspiration." J.Ir-
577 rig.Drain.Eng., 128(1), 1-10.

578 Gocic, M., and Trajkovic, S. (2010). "Software for estimating reference evapotranspira-
579 tion using limited weather data." Comput.Electron.Agric., 71(2), 158-162.

580 Healy, R. W., and Scanlon, B. R. (2010). Estimating groundwater recharge. Cambridge
581 University Press, Cambridge (U.K.).

582 Hirsch, R. M., Helsel, D. R., Cohn, T. A., and Gilroy, E. J. (1993). "Statistical Analysis
583 of Hydrologic Data." Handbook of Hydrology, D. R. Maidment, ed., McGraw-Hill,
584 New York (U.S.), 1-55.

585 Jackson, R. D. (1985). "Evaluating Evapotranspiration at Local and Regional Scales."
586 Proc IEEE, 73(6), 1086-1096.

587 Jain, S. K., Nayak, P. C., and Sudheer, K. P. (2008). "Models for estimating evapotran-
588 spiration using Artificial Neural Networks, and their physical interpretation." Hy-
589 drol.Processes, 22(13), 2225-2234.

590 Jensen, M. E., Burman, R. D., and Allen, R. G. (1990). Evapotranspiration and irriga-
591 tion water requirements. ASCE, New York (U.S.).

592 Kumar, M., Raghuwanshi, N. S., Singh, R., Wallender, W. W., and Pruitt, W. O.
593 (2002). "Estimating evapotranspiration using Artificial Neural Network." J.Ir-
594 rig.Drain.Eng., 128(4), 224-233.

595 Ladlani, I., Houichi, L., Djemili, L., Heddam, S., and Belouz, K. (2014). "Estimation of
596 Daily Reference Evapotranspiration (ET0) in the North of Algeria Using Adaptive
597 Neuro-Fuzzy Inference System (ANFIS) and Multiple Linear Regression (MLR) Mod-
598 els: A Comparative Study." Arab.J.Sci.Eng., 39(8), 5959-5969.

López-Urrea, R., Martín de Santa Olalla, F., Fabeiro, C., and Moratalla, A. (2006). "Testing evapotranspiration equations using lysimeter observations in a semiarid climate." Agric.Water Manage., 85(1-2), 15-26.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, University of California Press, Oakland, California (U.S.), 281-297.

Mallikarjuna, P., Jyothy, S. A., and Sekhar Reddy, K. C. (2013). "Daily Reference Evapotranspiration Estimation using Linear Regression and ANN Models." J. Inst. Eng. India Ser. A, 93(4) 215-221.

Martinez, C. J., and Thepadia, M. (2010). "Estimating reference evapotranspiration with minimum data in florida." J.Irrig.Drain.Eng., 136(7), 494-501.

Martínez-Cob A. (2008). "Use of thermal units to estimate corn crop coefficients under semiarid climatic conditions." Irrig.Sci., 26(4), 335-345.

Monteith, J. L. (1981). "Evaporation and surface temperature." Quarterly Journal, Royal Meterological Society, 107(451), 1-27.

Moore, R., Clark, W. D., and Vodopich, D. S. (2003). Botany. McGraw-Hill, New York (U.S.).

Naoum, S., and Tsanis, I. K. (2003). "Hydroinformatics in evapotranspiration estimation." Environ.Model.Softw., 18(3), 261-271.

Osbourne, J. W., and Waters, E. (2002). "Four Assumptions of Multiple Regression That Researchers Should Always Test." Practical Assessment, Research & Evaluation, 8(2) 1-7.

Parasuraman, K., Elshorbagy, A., and Carey, S. K. (2007). "Modelling the dynamics of the evapotranspiration process using genetic programming." Hydrol.Sci.J., 52(3), 563-578.

Sabziparvar, A. -., Tabari, H., Aeini, A., and Ghafouri, M. (2010). "Evaluation of class a pan coefficient models for estimation of reference crop evapotranspiration in cold semi-arid and warm arid climates." Water Resour.Manage., 24(5), 909-920.

627 Sanford, W. E., and Selnick, D. L. (2013). "Estimation of evapotranspiration across the
628 conterminous United States using a regression with climate and land-cover data." J. Am.
629 Water Resour. Assoc., 49(1) 217-230.

630 Shapiro, S. S., and Wilk, M. B. (1965). "An analysis of variance test for normality." Bi-
631 ometrika, 52(3-4) 591-611.

632 Shapiro, S. S., Wilk, M. B., and Chen, H. J. (1968). "A Comparative Study of Various
633 Tests for Normality." J. Am. Stat. Assoc., 63(324), 1343-1372.

634 SiAR. (2016). "Necesidades netas." eportal.magrama.gob.es/websiar/NecesidadesHidri-
635 cas.aspx (August, 2016).

636 Sorooshian, S., Duan, Q., and Gupta, V. K. (1993). "Calibration of rainfall-runoff mod-
637 els: Application of global optimization to the Sacramento Soil Moisture Accounting
638 Model." Water Resour.Res., 29(4), 1185-1194.

639 Stevens, J. (2009). Applied Multivariate Statistics for the Social Sciences. Taylor &
640 Francis, Mahwah, New Jersey (U.S.).

641 Sun, G., Alstad, K., Chen, J., Chen, S., Ford, C. R., Lin, G., Liu, C., Lu, N., Mcnulty, S.
642 G., Miao, H., Noormets, A., Vose, J. M., Wilske, B., Zeppel, M., Zhang, Y., and Zhang,
643 Z. (2011). "A general predictive model for estimating monthly ecosystem evapotranspi-
644 ration." Ecohydrology, 4(2), 245-255.

645 Tabari, H., Kisi, O., Ezani, A., and Hosseinzadeh Talaee, P. (2012). "SVM, ANFIS, re-
646 gression and climate based models for reference evapotranspiration modeling using lim-
647 ited climatic data in a semi-arid highland environment." J.Hydrol., 444-445 78-89.

648 Tabari, H., and Talaee, P. H. (2011). "Local Calibration of the Hargreaves and Priest-
649 ley-Taylor Equations for Estimating Reference Evapotranspiration in Arid and Cold
650 Climates of Iran Based on the Penman-Monteith Model." J.Hydrol.Eng., 16(10), 837-
651 845.

652 Tan, P. -., Steinbach, M., and Kumar, V. (2005). "Cluster Analysis: Basic Concepts and
653 Algorithms." Introduction to Data Mining, Addison-Wesley, Boston, Massachusetts
654 (U.S.), 487-568.

655 Testi, L., Villalobos, F. J., and Orgaz, F. (2004). "Evapotranspiration of a young irri-
656 gated olive orchard in southern Spain." Agric.for.Meterol., 121(1-2), 1-18.

Thiessen, A. J., and Alter, J. C. (1911). "Precipitation Averages for Larger Areas." Mon. Weather Rev., 39(7), 1082-1084.

Thut, H. F. (1938). "Relative Humidity Variations Affecting Transpiration." Am.J.Bot., 25(8), 589-595.

Tibshirani, R., Walther, G., and Hastie, T. (2001). "Estimating the number of clusters in a data set via the gap statistic." J.R.Stat.Soc.Ser.B Stat.Methodol., 63(2), 411-423.

Traore, S., Wang, Y. -., and Kerh, T. (2010). "Artificial neural network for modeling reference evapotranspiration complex process in Sudano-Sahelian zone." Agric.Water Manage., 97(5), 707-714.

Tryon, R. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers Malloy, Ann Arbor, Michigan (U.S.).

Villalobos, F. J., Testi, L., and Moreno-Perez, M. F. (2009). "Evaporation and canopy conductance of citrus orchards." Agric.Water Manage., 96(4), 565-573.

Voronoi, G. M. (1908). "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième Mémoire: Recherches sur les parallélloèdres primitifs." J. Reine Angew. Math., 134 198-287.

Wang, Y. -., Traore, S., and Kerh, T. (2008). "Neural network approach for estimating reference evapotranspiration from limited climatic data in Burkina Faso." WSEAS Trans.Comput., 7(6), 704-713.

Williams, L. E., Phene, C. J., Grimes, D. W., and Trout, T. J. (2003). "Water use of mature Thompson Seedless grapevines in California." Irrig.Sci., 22(1), 11-18.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. -., Steinbach, M., Hand, D. J., and Steinberg, D. (2008). "Top 10 algorithms in data mining." Knowl.Inf.Systems.Syst., 14(1), 1-37.

Xing, Z., Chow, L., Meng, F. -., Rees, H. W., Monteith, J., and Lionel, S. (2008). "Testing reference evapotranspiration estimation methods using evaporation pan and modeling in maritime region of Canada." J.Irrig.Drain.Eng., 134(4), 417-424.