APPLICATION ARTICLE

Natural Resource Modeling

Bayesian networks for evaluating Climate Change influence in olive crops in Andalusia, Spain.

Rosa F. Ropero[1] | **Rafael Rumí**[2] | **Pedro A. Aguilera**[1]

[1]Dpt. Biology and Geology, University of Almeria, Spain

[2]Dpt. of Mathematics, University of Almeria, Spain

Correspondence
Rosa F. Ropero
Email: rosa.ropero@ual.es

Funding information
Spanish Ministry of Economy and Competitiveness TIN2016-77902-C3-3-P, TIN2013-46638-C3-1-P; Regional Government of Andalusia P12-TIC-2541

Olive crops have become a strategic sector in Andalusia, Spain, providing an element of social cohesion and territorial management, but identified as vulnerable under Climate Change. Their great socio-economic importance makes the mitigation of Climate Change effects an important strategy. The main contribution of this paper is to show the application of Bayesian networks into Climate Change assessment using the evaluation of its impact over olive system in Andalusia. Both classification and regression models were learnt and validated to predict the potential olive grove distribution under an IPCC scenario. A lower error rate was obtained for the regression problem compared to classification. Results predict Climate Change will lead to changes into the territorial distribution of olive crops, with a movement from the river valley to the uplands due to the impact of the predicted increase in temperatures.

Recommendations for Resource Managers

- Bayesian networks are a powerful tool that allow dealing with both discrete and continuous data. However, if continuous data is available in order to retrieve all statistical information from them, discretization process should be avoided and original continuous data used for modeling purpose.
- Olive cropping area follows an altitudinal gradient from the river bed to the mountainous ranges. The minimum temperature limits the establishment of this crop over

certain altitude.
- Under the scenario of Climate Change the altitudinal gradient is lost and the main expanse of olive groves becomes fragmented into smaller patches. Besides, mountainous areas would reach optimal conditions for olive growth due to the increase in temperatures.

KEYWORDS
Agriculture systems, Classification, naïve Bayes, Regression, Tree Augmented naïve Bayes

1 | INTRODUCTION

Spanish agriculture was characterised by diverse and extensive croplands, with a significant proportion of rainfed cereal crops (Sánchez-Martínez et al., 2011). Olive trees (*Olea europaea* L.) are one of the oldest domesticated crops which best adapted to the Spanish climate. Their ability to grow under both dry and drought conditions, meant that olive cropping became an important sector in Spain (Tanasijevic et al., 2014). However, during the 1970s, the fall in olive oil prices made olive cropping much less profitable and a more problematic activity, which drove people to abandon it. However, at the beginning of the nineties, the European laws and grants arising from the Common Agricultural Policy (CAP) encouraged Spanish farmers to maintain and even to increase the surface area and productivity of olive groves. Under these new circumstances, olive cropping became a strategic sector in Spain, accounting for 51% of the European olive crop surface area (Commission, 2011). The autonomous region of Andalusia - particularly Jaén Province - is one of the areas that received most grants. According to the Management Plan for Olive Crops in Andalusia (PDOA, 2015), olives comprise the most symbolic and representative sector in this region with more than 1. 52 million hectares, producing over 75% of Spain's olive oil (Taguas et al., 2015). From the economic point of view, olive exploitation is the agricultural sector that provides most employment, with between 100 and 350 thousand labourers per year (PDOA, 2015). Thus, olive crops comprise an important agriculture sector in Andalusia and configure the so-called *comarcas olivareras* (open landscapes of olive groves), where both social and natural structures are highly conditioned by this crop. It means that olive crops have provided an element of social cohesion, territorial management, employment and wealth generation, while providing society with ecosystem services, such as cultural identity, carbon sequestration and food providing (PDOA, 2015).

The Intergovernmental Panel on Climate Change (IPCC) predicts significant changes in temperature and rainfall patterns in Andalusia, with an increase in the number and force of extreme events (storms and droughts) (Mendez-Jimenez, 2012; Solomon et al., 2007). These changes would potentially have an impact on the distribution of several species, leading to changes in their spatial patterns as optimal conditions relocate. Accordingly, the agriculture sector has been identified as one of the most vulnerable sectors in Andalusia under climatic change. In the case of olive cropping, its huge socio-economic importance makes mitigation of any impacts of climatic change an important strategy. In order to implement any mitigation strategy, climate change impacts firstly need to be clearly known, and this is where statistical modeling becomes crucial.

Climate change modeling approaches are becoming indispensable as a tool for helping management plans and politics to mitigate their effects (Niggol Seo, 2016; Schliep et al., 2015). However, most of the papers related with olive orchards modelling are mainly focused on their effect over soil moisture or soil erosion and, therefore, over olive pro-

ductivity (Viola et al., 2013; García-Ruíz, 2010). In this sense, Viola et al. (2014) develop a crop model divided into two parts, the first for estimating evapotranspiration and assimilation in well-watered conditions whilst the second aimed at reproducing water-stressed conditions. Once the model was calibrated and validated, it was applied to forecast the impact of three scenarios of climate change in a region of Southern Italy. In a similar way, dos Santos et al. (2017) evaluate the impact of climate projections on climatological water balance for olive orchard. Dhiab et al. (2017) use a partial least squares regression method in which olive output is the dependent variable and meteorological and aerobiological information are the independent variables for forecasting the extension of olive crop in Tunisia. Morales et al. (2016) model is more focused on the olive production with a three-dimensional model of canopy photosynthesis and radiation absorption from which simulations about Climate Change were carried out.

Most of these papers are based on deterministic models (Taguas et al., 2017; Militino et al., 2006). Here, we proposed the use of a probabilistic based model. Bayesian networks (BNs) have been defined in recent decades as powerful tools capable of dealing with uncertainty in real-life problems (Quentin Grafton, 2017; Abbal et al., 2016; Phan et al., 2016; Aguilera et al., 2011; Smith et al., 2011). In comparison with other fields, applications in environmental sciences and ecology are still scarce, though some examples of climatic change modelling can be found (Franco et al., 2016; Molina et al., 2013). According to the literature, BNs have been developed to deal with three types of problems: characterisation, classification and regression; but most studies have focused just on characterisation whilst BNs classifiers and BN-based regression, remain scarcely applied (Ropero, 2016).

The aim of this paper is to explore the use of BNs in the assessment of the impact of Climate Change on the extent of olive cropping in Andalusia, Spain. This paper is organised as follows: Section 2 defines and explains the concept of Bayesian networks and their use for classification and regression problems, and the inference process. Section 3 describes the methodology used, considering model learning, validation and scenarios of Climatic Change. Section 4 shows the results obtained and the predictions from the Climatic Change scenarios. Finally, Section 5 draws the conclusions obtained.

2 | BAYESIAN NETWORKS

Bayesian networks (BNs) are defined as a statistical multivariate model for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$. They are composed by two components: *i*) the qualitative part, a direct acyclic graph in which each vertex represents one of the variables, linked by an edge which indicates the existence of statistical dependence between them; and *ii*) the quantitative part as the conditional probability distribution for each variable $X_i, i = 1, \ldots, n$, given its parents in the graph ($pa(x_i)$) expressed in Conditional Probability Tables (CPTs) (in the case of discrete variables) or probability functions (for continuous variables).

The qualitative part allows BN models to be easily understood by experts in other fields who are unfamiliar with the model's mathematical context. Thus, experts and stakeholders can play an important part in the model learning process by identifying relationships between the variables, giving values for the CPTs or even refining the structure previously learnt from data (Aguilera et al., 2011).

This structure also allows that, with no mathematical calculation involved, the variable(s) that are relevant (or not) for a certain one can be known (Pearl, 1988), and help to simplify the joint probability distribution (JPD) of the variables necessary to specify the model. Thus, BNs provide a compact representation of the JPD over all the variables, defined as the product of the conditional distributions attached to each node, so that

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid pa(x_i)). \quad (1)$$

where $pa(x_i)$ is the set of parents of variable x_i according to the structure of the directed acyclic graph.

BNs were originally developed for discrete variables, but real life problems require both continuous and discrete (hybrid) data to be simultaneously included in the modelling processes. This necessity has brought about the proposal of new models for dealing with hybrid data in BNs. One of these models is the *Mixture of Truncated Exponential* models (MTEs), proposed by Moral et al. (2001) and developed in detail in Rumí (2003). Similar to discretization methods, through MTE models the range of the variable is divided into a set of intervals, in each of them the distribution is approximated by an exponential function, rather than a constant value like in discretization (for a detail information about MTE see Cobb et al. (2007a); Rumí and Salmerón (2007); Rumí et al. (2006)). This approximation allows both discrete and continuous variables to be included simultaneously in the model with no changes into the methodology followed (Ropero et al., 2014). In environmental science, BNs based on MTE models have been successfully applied for regression (Maldonado et al., 2016a), classification (Maldonado et al., 2016b), characterisation (Ropero et al., 2016), and even dynamic models (Ropero et al., 2017).

Defined in Moral et al. (2001), MTE models divide the value range of a continuous variable into several intervals, and approximate each of them by an exponential function rather than by a constant (Rumí, 2003), since they are closed under restriction, marginalization and combination. It is able to deal with any distribution function, due to of its high fitting power, which makes it appropriate to deal with hybrid data. In the same way as in discretisation, the more intervals used to divide the domain of the continuous variables, the better the MTE model accuracy, but also the more complex. Furthermore, in the case of MTEs, using more exponential terms within each interval substantially improves the fit to the real model, but, again, more complexity is assumed. For more details about learning and inference tasks in MTE models, see Rumí et al. (2006); Rumí and Salmerón (2007) and Cobb et al. (2007b).

2.1 | Classification based on Bayesian networks

A classification problem, in which a discrete class variable C exists, and a set of continuous or discrete explanatory variables (called features) $X_1, \ldots, X_n$, can be expressed as a BN and an individual with observed features $x_1, \ldots, x_n$ will be classified as belonging to class c^* obtained as

$$c^* = \arg\max_{c \in \Omega_C} f(c \mid x_1, \ldots, x_n), \quad (2)$$

where Ω_C denotes the set of possible values of C.

If we consider that $f(c \mid x_1, \ldots, x_n)$ is proportional to $f(c) \times f(x_1, \ldots, x_n \mid c)$, the specification of an n dimensional distribution for $X_1, \ldots, X_n$ given C is required in order to solve the classification problem, which implies a high computational cost, since the number of parameters necessary to specify a joint distribution is exponential to the number of variables. However, using the factorisation determined by the network (Eq. 1), this cost can be broadly reduced.

2.2 | Regression based on Bayesian networks

In the case of regression, the idea is similar to classification problems but the so-called class variable is now continuous. A BN can be used as a regression model for prediction purposes if it contains a continuous response variable Y and a set of discrete and/or continuous feature variables $X_1, \ldots, X_n$. Thus, in order to predict the value for Y from k observed features, with $k \leq n$, the conditional density

$$f(y \mid x_1, \ldots, x_n), \tag{3}$$

is computed, and a numerical prediction for Y is given[1] using the expected value as follows:

$$\hat{y} = g(x_1, \ldots, x_n) = \mathbb{E}[Y \mid x_1, \ldots, x_n] = \int_{\Omega_Y} y f(y \mid x_1, \ldots, x_n) dy, \tag{4}$$

where Ω_Y represents the domain of Y.

Note that $f(y \mid x_1, \ldots, x_n)$ is proportional to $f(y) \times f(x_1, \ldots, x_n \mid y)$, and therefore, solving the regression problem would require a distribution to be specified over the n variables given Y. The associated computational cost can be very high. However, using the factorisation determined by the network, the cost can be again broadly reduced.

2.3 | Constrained structures for classification and regression through BNs

In both classification and regression problems, the aim of the model is to predict, as accurately as possible, the results of the class/response variable (olive crops in our case) rather than properly estimate the parameters of the relationship between all variables, so that the so-called *constrained* or *fixed structure* is developed. The extreme case is the *naïve Bayes* (NB) structure (Duda et al., 2001; Friedman et al., 1997), which consists of a Bayesian network with a single root node and a set of attributes having only the class/response variable as a parent.

Its name comes from the naive assumption that feature variables $X_1, \ldots, X_n$ are assumed to be independent given the class/response variable. This strong conditional independence assumption is somehow compensated by the reduction of the number of parameters to be estimated from data, since in this case, it holds that

$$f(c \mid x_1, \ldots, x_n) \propto f(c) \prod_{i=1}^{n} f(x_i \mid c), \tag{5}$$

which means that, instead of one n-dimensional conditional distribution, n one-dimensional conditional distributions are estimated. Despite this extreme independence assumption, the results are amazing in many cases, and it is for this reason that it has become the most widely used Bayesian classifier in the literature.

A step beyond is to allow each feature to have one more parent beside the response/class variable, so configuring a *Tree Augmented naïve Bayes*. For each dataset there are several possible TAN structures, so the way to choose between them is to learn a maximum weight spanning tree structure with feature variables using the mutual information with respect to the response/class variable as the weight of each edge (Chow and Liu, 1968), defined as

[1] Note that in the BN framework, a prediction of Y can be obtained even when some of the variables are not observed.

$$I(X_i, X_j \mid C) = \sum_{X_i, X_j, C} p(x_i, x_j \mid c) \log \frac{p(x_i, x_j \mid c)}{p(x_i \mid c) p(x_j \mid c)} \tag{6}$$

These extra relationships are not based on an environmental interpretation but on the amount of information they share with the response/class variable. Finally, relationships from the response/class to each feature are included. In general, the increased complexity, in both the structure and the number of parameters results in richer and more accurate models (Friedman et al., 1997).

2.4 | Inference in Bayesian networks

Once the model is learnt and validated, BNs allow new information, or *evidence*, to be included into the model, through the so-called *inference process* or *probabilistic propagation*. If we denote the set of *evidenced* variables as **E**, and its value as e, then the inference process consists of calculating the posterior distribution $p(x_i|\mathbf{e})$, for each variable of interest $X_i \notin \mathbf{E}$:

$$p(x_i|\mathbf{e}) = \frac{p(x_i, \mathbf{e})}{p(\mathbf{e})} \propto p(x_i, \mathbf{e}), \tag{7}$$

since $p(\mathbf{e})$ is constant for all $X_i \notin \mathbf{E}$. So, this process can be carried out computing and normalising the marginal probabilities $p(x_i, \mathbf{e})$, in the following way:

$$p(x_i, \mathbf{e}) = \sum_{\mathbf{x} \notin \{x_i, \mathbf{e}\}} p_e(x_1, \ldots, x_n), \tag{8}$$

where $p_e(x_1, \ldots, x_n)$ is the probability function obtained from replacing in $p(x_1, \ldots, x_n)$ the evidenced variables **E** by their values **e**.

3 | METHODOLOGY

The cover area by olive crops in Andalusia was predicted through BNs according to a set of land uses and climatic variables. In the literature, it is common to change continuous into discrete variable in order to deal with a classification problem rather than a regression since most available software are not able to deal with continuous or hybrid variables (Aguilera et al., 2011). In order to compare both approximations, once regression model was learnt with the original continuous data, the olive crops variable was discretised and a classification model built. Both techniques were compared in terms of error rate. Lastly, a scenario of climatic change with two horizons, 2040 and 2100, was included and the evolution of olive orchard was evaluated.

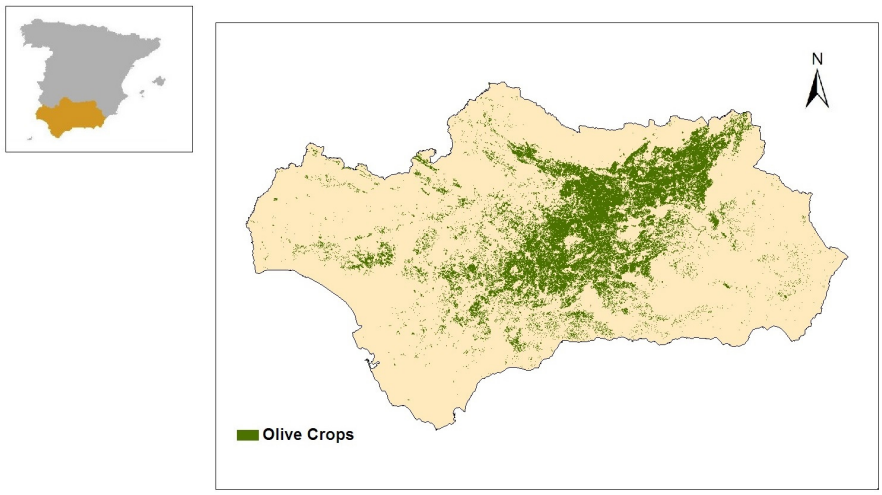


FIGURE 1 Location of Andalusia in southern Spain and the extent of olive cropping.

3.1 | Study area

Andalusia is located in southern Spain (Figure 1) forming the second largest autonomous region (covering more than 87.000km^2) and the most densely populated according to data from the National Statistics Institute [2].

From the climatic point of view, Andalusia presents a Mediterranean climate, characterised by mild annual temperatures and an irregular rainfall pattern, with periods of drought followed by strong storm events. This irregularity is also territorial, with stark differences between inland and coastal Andalusia. Whilst coastal areas, mainly in the southeast, are characterised by semi-arid conditions, inland Andalusia is more humid.

Inland Andalusia enclosed the so-called Baetic depression, with the *Guadalquivir* river basin area. Two extensive mountainous ranges - the *Sierra Morena* mountain range and the Baetic system- separate this depression from the coastal and north area of Andalusia. This river valley plain presents optimum conditions for agricultural settlement, mainly rainfed herbaceous and woody crops (olive is the most common crop). In this area, more than 50% of territory is used for agriculture. In addition, its population has close links to agriculture, through direct employment (in the primary economic sector), and through indirect employment (manufacturing industries).

3.2 | Data collection & pre-processing

Data were collected from the Andalusian Regional Environmental Information Network [3], from the Regional Goverment of Andalusia. Data from different thematic maps (climate and land uses) were incorporated into the geographic information system ArcGIS v.9.3. For all datasets used, the coordinate system is based on the European Terrestrial Reference System 1989 (ETRS89). A 5x5 km grid cell was used for obtaining the information for all variables, giving a total of 3630 observations.

Climatic information was obtained from raster maps based on the 1971-2000 time series. For each cell, values of

[2] http://www.ine.es

[3] http://www.juntadeandalucia.es/medioambiente/site/rediam

TABLE 1 Minimum, maximum and mean values for all variables collected. PET, Potential Evapotranspiration.

Variable	Min	Max	Mean
PET	0	962	832.0
Average Temperature	0	19.03	15.94
Average Rainfall	0	1753	567
Herbaceous Crops	0	100	17.87
Woody Crops	0	100	4.1
Forest	0	100	48.3
Water Surface	0	100	2.78
Bared Areas	0	100	3.17
Olive Crops	0	100	16.37

Potential Evapotranspiration (PET, expressed in mm), Annual Average Temperature (°C) and Annual Average Rainfall (mm) were collected.

Land use information was obtained from the SIOSE 2011 [4]. Initially, a total of 138 different land uses were represented on the thematic map. This number was reduced into 10 by merging them using similarity criteria (*i.e.*, all different types of herbaceous crops were merged into one unique variable: *Herbaceous Crops*). Besides, those land uses occupying less than 1% of the total surface were removed, as well those variables for which more than 75% of data were equal to 0. Finally, a total of six land use variables were preserved and expressed as the percentage of the cell occupied by this land use.

All variables collected were continuous, and Table 1 summarises their main statistics. The surface area of olive groves was expressed in the continuous variable *Olive crops*, which is the variable of interest in our models. In order to compare against regression, *Olive crops* was discretised into five intervals for the classification model: 0-Absence; 1-Less than 25%; 2- Between 25-50%; 3- Between 50-75%; 4- More than 75% land cover.

3.3 | Models Learning

The objective of this paper is to predict, as accurately as possible, the potential cover of olive crops in Andalusia, which is originally a continuous variable. It means a regression problem is faced. Model learning was addressed using Elvira software (Elvira-Consortium, 2002), based on MTE models, and two different structures were tested, a NB and a TAN [5]. A 5-parameter MTE distribution was fitted for every probability distribution due its ability ti fit the most common distributions accurately while both model complexity and the number of parameters to be estimated remains low.

In order to compare the performance of this continuous model against classification methods, a BN classifier was learnt using the discretised version of the *Olive crops* variable, whilst all features remain continuous. Again, both NB and TAN structures were used, so that the comparison is more reliable since the model structure remains constrained.

[4] http://www.siose.es

[5] The software and datasets are available as supplementary material

Model parameters were again estimated using Elvira software (Elvira-Consortium, 2002) and based on 5-parameter MTE models.

So, a total of four BNs models were obtained: two of these treat the *Olive crops* variable as continuous - NB for regression (NBr) and TAN for regression (TANr) - while the other two treat the *Olive crops* variable as discrete - NB classifier (NBc) and TAN classifier (TANc).

3.4 | Models Validation

Models obtained were tested through *k*-fold cross validation (Stone, 1974). This is a commonly used technique in Artificial Intelligence for model validation that is based on the holdout method to check how predictive a model is when confronted with data that have not been previously used for learning. Data is separated into two complementary sets, one for learning (D_l) and one for testing (D_t), and the root mean square error (rmse) of a model built from D_l according to D_t was estimated using the following equation:

$$rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{9}$$

To reduce variability, data is initially divided into *k* subsets, and the method is repeated *k* times. Each time, one of the *k* subsets is used as D_t and the other *k*-1 subsets are combined to form D_l. Finally, the average error of *k* trials is computed. In this paper, *k* was set to 10.

Equation 9 was developed for continuous variables. So, in order to compare regression model with the classification, the rmse equation needs to be re-computed for the discrete version as:

$$rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - ca(\hat{c}_i))^2}\,, \tag{10}$$

where $ca(\hat{c}_i)$ is the class average for the predicted category after propagating the records in the discrete case, and y_i is the actual continuous value for the response variable. Note that, once the data are discretised, the original continuous values are still necessary to compute this version of the rmse. So that, a 10-fold cross validation was computed for both regression and classification models.

3.5 | Scenarios of Climatic Change

In this paper, the aim is to predict how the land use of olive crops might change as a consequence of Climatic Change through an inference process.

The IPCC considers two main scenarios of Climatic Change for Andalusia: A2 and B2 (Mendez-Jimenez, 2012). The A2 scenario describes a heterogeneous world, where self-reliance and preservation of local identity are key. Population increases continuously and economic development is based on national decisions (regionally oriented), whilst per capita economic growth and technological change are fragmented and slow (Gasca, 2014; Solomon et al., 2007). By contrast, the B2 scenario describes a situation in which economic development is not important and the environmental and socio-economic problems are solved at local level. This scenario implies a slow population increase (Gasca,

2014; Solomon et al., 2007). In our study, we focus on the A2 scenario, since we consider it closer to the current trend of socio-ecological change.

Taking the information provided by the IPCC, both national and regional governments have developed climate change scenarios for their territory for a set of variables. In Andalusia, the Environmental Information Network (REDIAM) provides information as a shapefile, with prediction for a set of climatic and land use variables according to the Assessment of the International Panel on Climate Change (Stocker et al., 2013). By this means, the information for the evidences was collected from the REDIAM for only climatic variables, which means just three variables were used as *evidences* (*Rainfall*, *Temperature* and *PET*).

One advantage of BNs is that it is not necessary to include information for all feature variables in order to be able to make the prediction (Ropero et al., 2014). Rather, only new information is included as evidences in those variables for which we have knowledge about their change. In our case, evidences were included in those variables we have information about, *i.e.* the climatic variables and propagated to predict the most probable state of the variable *Olive crops*.

4 | RESULTS AND DISCUSSION

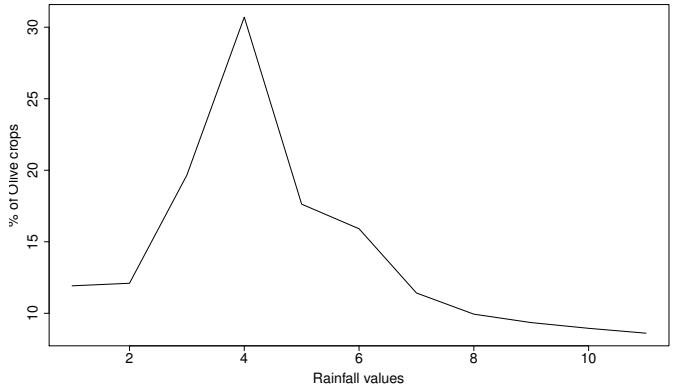
4.1 | Models comparison

Table 2 shows a comparison between BNs and traditional methodologies in terms of their ability to include both continuous and discrete variables simultaneously, dealing with regression and classification problems, and provide interpretable results and a quality output. BNs are able to deal with these five items, and besides, provide additional advantages. Firstly, it is possible to establish specific types of relationships between each feature and the response variable (*i.e.*, linear relationships in the case of Multi Linear Regression), and also, no relationships among the features are often possible. By contrast, BN models allow relationships between features to be included with the aim of improving the response variable prediction. In our paper, only constrained structures have been built, but general networks can be used in which the relationships between features do not respond only to optimise the parameters of the response variable, but also to include environmental knowledge. Also, other types of relationships can be modelled in a BN not only linear relationships. Figure 2(a) shows the type of relationship identified with the regression model with a BN with a TAN structure. These relationships were identified by including several values for each feature and obtaining the value of the probability distribution. In order to summarise the evolution of the probability function when each feature is increasing, the mean of the distribution was calculated and represented. In this case, BNs identify not only direct (Figure 2(a) *PET* variable represents with a +) and inverse (Figure 2(a) *Herbaceous Crops* variable represents with a -) relationships, but also those variables that present a complex relationship with *Olive crops* (Figure 2(a) *Average Rainfall* variable epresents with a +-). Figure 2 b) shows the relationship between *Average Rainfall* and *Olive Crops*. At the beginning of the range, an increase in *Average Rainfall* leads to an increase in *Olive Crops*. Then, at the fourth point of the *Rainfall* variable distribution which is equivalent to the value 526 mm, the relationship changes and further increase in *Average Rainfall* provokes a decrease in *Olive Crops*. From an environmental point of view, olives flourish in not so humid areas; thus, while an initial increase in rainfall would improve conditions for growth, above a certain rainfall threshold, the conditions become sub-optimal for olive crops again.

Another important advantage of BNs is that not all feature variables need to be evidenced in order to predict the response variable. In the case of traditional regression methodologies, if new information about a subset of features needs to be included into the model, the remaining unknown features have to be set with a value (*i.e.*, the mean or the initial value) which implies non-real situations are modeled and some noise is included. In the case of BNs, a scenario

Variable	Relation BNs
PET	+
Average Temperature	+-
Average Rainfall	+-
Herbaceous Crops	-
Woody Crops	-
Forest	-
Water Surface	+-
Bared Areas	+-

(a) Type of relationship



(b) Relationship Rainfall - Olive crops

FIGURE 2 Type of relationships identified by BNs based regression with a TAN structure (a). + means a direct relationship; - means an inverse relationship; +- means a complex relationship. An example of relationships between a feature (*Average Rainfall* variable) and the *Olive Crops* variable (b) in which total of 12 points (minimum, maximum and 10 equidistant points) were included into the feature and the mean of the posterior probability distribution for *Olive Crops* at each point, was calculated.

TABLE 2 Comparisons between Bayesian networks and traditional methodologies. Y expresses that this method is able to deal appropriately with this item, whilst X expresses this method is not. Quality output refers to the information contained in the output *i.e.*, if it returns only a value or a distribution of values, which is more informative

Method	Discrete/Continuous	Regression	Classification	Interpretability	Quality output
Bayesian networks	Y	Y	Y	Y	Y
Linear regression	X	Y	X	Y	Y
Regression trees	Y	Y	X	Y	X
Classification trees	Y	X	Y	Y	X
Logistic regression	X	X	Y	X	Y
Neural network	X	X	Y	X	X

TABLE 3 Root mean square error for the 4 BNs models: NB classifier (NBc), TAN classifier (TANc), NB for regression (NBr) and TAN for regression (TANr); and the accuracy of BNs classifiers.

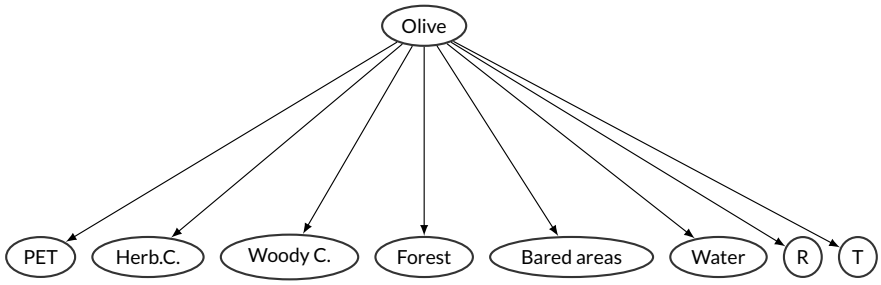
Model	rmse	Accuracy
NBr	21.67	-
TANr	19.64	-
NBc	35.80	0.507
TANc	31.06	0.549

of change can be included to take into account a subset of feature variables, and keeping the rest as non-evidenced variables, rather than including an estimate (Ropero et al., 2014). In our case, information about just three features (over 8) were included as *evidences*, and results were properly obtained. This is why we have decided not to compare against other traditional methodologies.

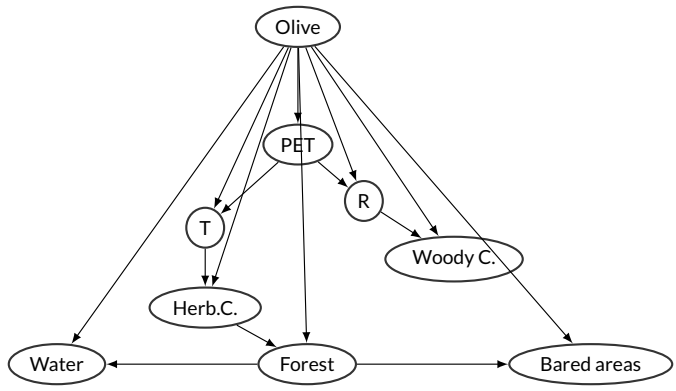
Table 3 shows the rmse values for the BN models, which are mainly used in this paper as a way to compare different models in an appropriate way, rather than to measure the goodness of the models. Firstly, regression models provide smaller errors than classification. Data discretisation is the most commonly used solution when dealing with continuous data. Even though when several approaches have been proposed for dealing with hybrid or continuous data, the literature shows that a high percentage of models continue discretising the data (Ropero, 2016; Aguilera et al., 2011). According to our results, discretising only the class variable causes rmse to increase from 19.64 to 31.06 in the case of the TAN structure, and from 21.67 to 35.80 for the NB case. Thus, it is demonstrated that dealing with the original data provides more accurate results. Besides, In Table 3 the accuracy of both BN classifiers are also shown, and results are below 0.6, which implies less than 60% of results predicted by the model agree with the real data (used for validation purpose).

Finally, Figures 3(a) and 3(b) show the NB and TAN structures obtained. In terms of complexity, measured as number of links, NB is simpler (8 links, between *Olive crops* variable and the features) than TAN (15 links, between

Olive crops variable and each feature, but also among the features). Besides, the computational time was calculated using a MacBook Air, 1.6 GHz Intel Core i5. RAM 4GB 1600 MHz DDR3, and again, NB is faster (11min 42sec) than TAN (4h 21min). Despite this results, TAN provide a smaller error in both classification and regression models (Table 3). This difference is smaller in the case of regression. Including relationships between features, despite the increase in model complexity and computational cost, means that the model is better able to predict the response variable.



(a) NB



(b) TAN

FIGURE 3 Regression models for olive crops based on both NB and TAN structure. T., Temperature; R., Rainfall; PET, Potential Evapotranspiration; Herb. herbaceous.; C. Crops.

4.2 | Evolution of olive crops in Andalusia under Climate Change

A Bayesian network model for regression based on a TAN structure was selected, and scenarios of climatic change for 2040 and 2100 horizons included. Figure 4 shows the results *a priori*, before any scenario of change, and the predictions for 2040 and 2100 horizons.

According to the Technical Report on Andalusian Climatic Change Adaptation in the Agriculture Sector (Mendez-Jimenez, 2012), optimal conditions for olive crops are characterised by a warm mean annual temperature (16-22 °C) and 650 mm of mean annual rainfall. However, the main limiting factors are the maximum summer temperatures (above 35-40 °C photosynthesis is affected) and sudden falls in the minimum temperature during certain moment

of the cycle (*i.e.*, around 0 °C during the flowering period can provoke irreparable damage). In terms of rainfall, even though olives are quite drought-resistant, when annual rainfall is less than 200 mm, production is drastically reduced. Also, rapid and sudden changes in temperature or rainfall can have important consequences on olive production if they take place during the flowering or sprouting period.

Results *a priori* (Figure 4(a)) show the (current) distribution of olive crops in Andalusia. A gradient of color shows the extent of olive crops per grid cell. The main area of olive groves, corresponds to the *Guadalquivir* river plain. This is Andalusian's largest river catchment, and it has a strong agricultural character. Through history, a warm climate and rich soils have encouraged agricultural settlement here. Further West, olive cropping decreases and around the river mouth it covers less than 30 % of land surface. In this area rainfed crops could be replaced by crops which necessities fit better with the new environmental conditions. By contrast, in southeastern Andalusia, the scarcity of olive groves is explained by two factors: the semi-arid conditions, and the steep relief. This area is characterised by an abrupt mountain relief with peaks rising to more than 2000 m.a.s.l.; the climate is semi-arid, most markedly so in the so-called *Desert of Tabernas*. However, over recent decades, in this desert area, significant olive cropping has been establish thanks to the use of groundwater irrigation systems and stable climatic conditions. Finally, the areas marked in red in Figure 4(a), correspond to the mountainous landscapes in Andalusia. In these areas, the minimum temperature limits the establishment of this mediterranean crop.

Both the 2040 and 2100 horizons were evaluated under the A2 scenario of Climatic Change. Predictions were made using a BN regression model based on a TAN structure. They predict significant changes in olive crop distribution in Andalusia.

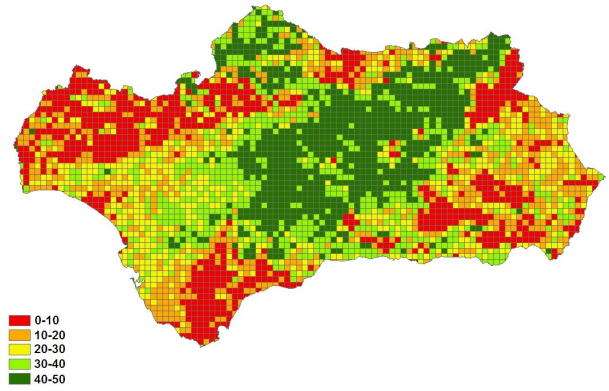
In comparison with *a priori*, under the 2040 horizon (Figure 4(b)) the main area of olive crops is now fragmented into several patches Besides, the *Guadalquivir* plain becomes sub-optimal to support its *a priori* cover of olive crops, and predict less than 30% cover. This change is driven by the increase in annual average temperature in this area provoked by more intense heatwaves and a higher maximum temperatures in summer. This 2040 scenario includes warmer summers rising to more than 40 degrees. This would provoke a reduction in photosynthesis and losses in olive production. In the same way, mountainous areas would suffer increased temperatures, but in this case the situation is possitive for *Olive crops*: fewer frosts and a higher minimum temperature means these areas would reach optimal conditions for olive growth. Comparing *a priori* and 2040 maps, the extension of olive crops increase from less than 10% to between 10 and 30%.

This relocation from the river plain to the uplands is even more emphatic under the 2100 horizon (Figure 4(c)). Under the 2100 scenario, only a few cells are predicted to have olive groves covering less than 10% and these are mainly located in the south-east. However, compared to the 2040 horizon, the continued increase in temperature means several areas change from 40-50% cover of *Olive crops* to 30-40% of olive extension.

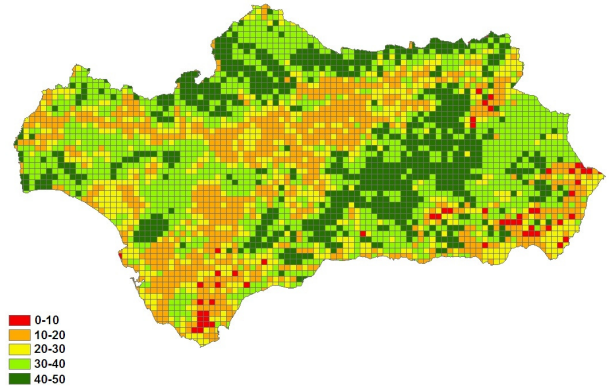
These results are not deterministic since they are based on probabilistic approach. The value of each cell is represented as a probability function rather than an absolute value of extension. From this probability distribution a set of statistics and measurements can be obtained. In this paper, for obtaining the maps of Figure 4, the mean of the probability distribution was obtained. So, the uncertainty enclosed in the results depends on the reliability of the evidences included into the model and the robustness of the model with respect to the rest of variables in the horizon studied.

5 | CONCLUSIONS

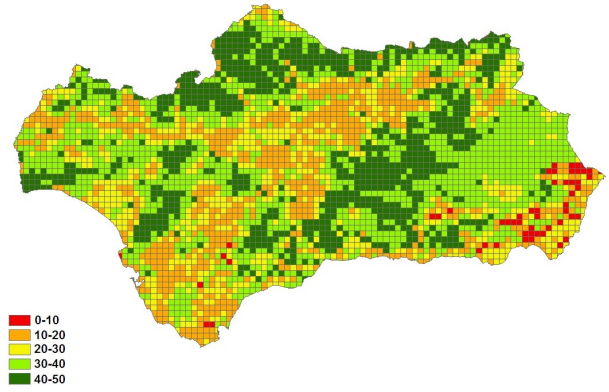
Olive cropping has become one of the most important agricultural activities in Andalusia. Its role in the socio-economy means this crop is the basis of a complex *agroecosystem*. Olive production is the main economic activity in some parts



(a) A priori



(b) 2040



(c) 2100

FIGURE 4 Percentage surface area occupied by olive crops in Andalusia according to results from regression based on BNs model with a TAN structure

of Andalusia and potential changes in its distribution would affect its annual production and impact the socioeconomic structure of these areas.

According to our results, the *a priori* situation confirms a significant area of olive production near the *Guadalquivir* river plain and a limit of olive groves marked by altitude. In the mountains, the lower temperatures means that olive crops are not currently planted, and there is a gradual transition from upland to lowland planting pattern. However, this altitudinal structure is lost under the A2 scenario. The increase of temperature and decrease of rainfall means that olive planting would be relocated from the valley plain to mountainous areas. This territorial change would have an impact on the local socio-economy. Nowadays, a significant proportion of the population depend, directly or indirectly, on olive production. In these areas of olive crops, the reduction of their extension and productivity would imply a decrease in employment rates and richness, which encourage the movement of the population. Besides, economic sector should be diversified in order to adapt to this change and not to depend only on one agriculture activity. The impact of Climatic Change on the distribution of this crop in turn implies a significant potential impact on socioeconomic structure.

Bayesian networks have been extensively applied into several scientific fields to evaluate the impact of new conditions on the system being modelled. In environmental sciences, applications of BNs to study Global Environmental Change and Climatic Change impacts are still scarce despite their advantages over classical methodologies. In the present application, data available were totally continuous and no discretization method was applied in order to fit with the software requirements, but for comparing a totally continuous against a hybrid model. Thus, regression and classification models were obtained and results shows regression models give a smaller error than classification ones.Besides, the methodology applied was not modified in order to deal with both continuous and discrete/continuous or hybrid data, which is an advantage over other traditional methods. Another advantage of BNs is that relationships between each feature and the goal variable can be studied in detail. In this paper, the nature of the relationships between each feature and the goal variable *Olive crops* were studied and both positive and negative relations, but also, more complex ones, were found. Finally, in order to perform an accurate and real scenario of change, BNs allow new information to be included in just those variables for which well-known information is available, climate variables in our case. This means that the remaining variables remain as non-evidenced and no estimates are made of their values.

This paper presents an initial study on the impact of Climatic Change on olive cropping, which included only climatic and land use variables. Due to their great importance in the socioeconomic systems of Andalusia, these models should be extended to include socioeconomic information in order to provide greater focus in that field.

ACKNOWLEDGEMENTS

This study was supported by the Spanish Ministry of Economy and Competitiveness through projects TIN2016-77902-C3-3-P and TIN2013-46638-C3-1-P, and by the Regional Government of Andalusia through project P12-TIC-2541. Rosa F. Ropero is supported by a post-doc *Contrato Puente* funded by the University of Almería.

REFERENCES

Abbal, P., Sablayrolles, J., Matzner-Lober, E., Boursiquot, J., Baudrit, C. and Carbonneau, A. (2016) A decision support system for vine growers based on a bayesian network. *Journal of Agricultural, Biological and Environmental Statistics*, **21**, 131–151.

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R. and Salmerón, A. (2011) Bayesian networks in environmental modelling. *Environmental Modelling & Software*, **26**, 1376–1388.

Chow, C. K. and Liu, C. N. (1968) Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, **14**, 462–467.

Cobb, B., Rumí, R. and Salmerón, A. (2007a) *Bayesian networks models with discrete and continuous variables*, chap. Studies in Fuzziness and Soft Computing, 81–102. Advances in probabilistic graphical models.

Cobb, B. R., Rumí, R. and Salmerón, A. (2007b) *Advances in probabilistic graphical models*, chap. Bayesian networks models with discrete and continuous variables, 81–102. Studies in Fuzziness and Soft Computing. Springer.

Commission, E. (2011) Agriculture and fishery statistics, main results 2009-10. European Union-Eurostat. *Tech. rep.*, European Commission.

Dhiab, A., Mimoun, M., Oteros, J., Garcia-Mozo, H., Dominguez-Vilches, E., Galan, C., Abichou, M. and Msallem, M. (2017) Modeling olive-crop forecasting in Tunisia. *Theor. Appl.Climatol*, **128**, 541–549.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001) *Pattern classification*. Wiley Interscience.

Elvira-Consortium (2002) Elvira: An Environment for Creating and Using Probabilistic Graphical Models. In *Proceedings of the First European Workshop on Probabilistic Graphical Models*, 222–230. URL: `http://leo.ugr.es/elvira`.

Franco, C., Appleby Hepburn, L., Smith, D., Nimrod, S. and Tucker, A. (2016) A bayesian belief network to assess rate of changes in coral reef ecosystems. *Environmental Modelling & Software,*, **80**, 132–142.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian network classifiers. *Machine Learning*, **29**, 131–163.

García-Ruíz, J. (2010) The effects of land uses on soil erosion in Spain: A review. *Catena*, **8**, 1–11.

Gasca, A. M. (2014) *Guía de escenarios regionalizados de cambio climático sobre España a partir de los resultados del IPCC-AR4*. AEMET, Ministerio de Agricultura, Alimentación y Medio Ambiente.

Maldonado, A., Aguilera, P. and Salmerón, A. (2016a) Continuous Bayesian networks for probabilistic environmental risk mapping. *Stochastic Environmental Research & Risk Assessment*, **30(5)**, 1441–1455.

— (2016b) Modeling zero-inflated explantory variables in hybrid bayesian network classifiers for species occurrence prediction. *Environmental Modelling & Software*, **82**, 31–43.

Mendez-Jimenez, M. (2012) Estudio básico de adaptación al cambio climático. *Tech. rep.*, Regional Goverment of Andalusia.

Militino, A., Ugarte, M., Goigoa, T. and Gonzalez-Audicana, M. (2006) Using small area models to estimate the total area occupied by olive trees. *Journal of Agricultural, Biological and Environmental Statistics*, **11**, 450–461.

Molina, J. L., Pulido-Veláquez, D., García-Aróstegui, J. and Pulido-Velázquez, M. (2013) Dynamic Bayesian Network as a Decision Support tool for assessing Climate Change impacts on highly stressed groundwater systems. *Journal of Hydrology*, **479**, 113–129.

Moral, S., Rumí, R. and Salmerón, A. (2001) Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. In *ECSQARU'01. Lecture Notes in Artificial Intelligence*, vol. 2143, 156–167. Springer.

Morales, A., Leffelaar, P., Testi, L., Orgaz, F. and Villalobos, F. (2016) A dynamic model of potential growth of olive (olea europaea l.) orchards. *European Journal of Agronomy*, **74**, 93–102.

Niggol Seo, S. (2016) The micro-behavioral framework for estimating total damage of global warming on natural resource enterprises with full adaptations. *Journal of Agrcultural, Biological and Environmental Statistics*, **21**, 328–347.

PDOA (2015) Plan director del olivar andaluz. *Tech. rep.*, Regional Goverment of Andalusia.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*. San Mateo, California.

Phan, T., Smart, J. C., Capon, S., Hadwen, W. and Sahin, O. (2016) Applications of Bayesian belief networks in water resource management: A systematic review. *Environmental Modelling & Software*, **85**, 98–111.

Quentin Grafton, R. (2017) Risks, resilience and natural resource management: Lessons from selected findings. *Natural Resource Modeling*, **30**, 91–111.

Ropero, R. F. (2016) *Hybrid Bayesian Networks: A statistical tool in Ecology and Environmental Sciences.* Ph.D. thesis, Dpt. Biology and Geology. University of Almería.

Ropero, R. F., Aguilera, P. A., Fernández, A. and Rumí, R. (2014) Regression using hybrid Bayesian networks: Modelling landscape-socioeconomy relationships. *Environmental Modelling & Software*, **57**, 127–137.

Ropero, R. F., Flores, M. J., Rumí, R. and Aguilera, P. A. (2017) Applications of hybrid dynamic bayesian networks to water reservoir management. *Environmetrics*, **28**, 1–11.

Ropero, R. F., Rumí, R. and Aguilera, P. (2016) Modelling uncertainty in social-natural interactions. *Environmental Modelling & Software*, **75**, 362–372.

Rumí, R. (2003) *Modelos de redes bayesianas con variables discretas y continuas.* Ph.D. thesis, Universidad de Almería.

Rumí, R. and Salmerón, A. (2007) Approximate probability propagation with mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, **45**, 191–210.

Rumí, R., Salmerón, A. and Moral, S. (2006) Estimating mixtures of truncated exponentials in hybrid Bayesian networks. *Test*, **15**, 397–421.

Sánchez-Martínez, J., Gallego-Simón, V. J. and Araque-Jiménez, E. (2011) The andalusian olive grove and its recent changes. *Estudios Geográficos*, **270**, 203–229.

dos Santos, D. F., Martins, F. and Torres, R. (2017) Impacts of climate projections on water balance and implications on olive crop in minas gerais. *Revista Brasileira de Engenharia Agricola e Ambiental*, **21**, 77–82.

Schliep, E., Gelfand, A. and Clark, J. (2015) Stochastic modeling for velocity of climate change. *Journal of Agricultural, Biological and Environmental Statistics*, **20**, 323–342.

Smith, R. I., Dick, J. M. and Scott, E. M. (2011) The role of statistics in the analysis of ecosystem services. *Environmetrics*, **22**, 608–617.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M. and Miller, H. (2007) *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

Stocker, T., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V. and Midgley, P. (2013) *Climate Change 2013. The Physical Science Basis. Working Group I Contribution to the fifth Assessment Report of the Intergovermental Panel on Climate Change.* WMO, UNEP.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36 (2)**, 111–147.

Taguas, E., Vanderlinden, K., Pedrera-Parrilla, A., Giraldez, J. and Gomez, J. (2017) Spatial and temporal variability of spontaneous grass cover and its influence on sediment losses in an extensive olive orchad catchment. *Catena*, **157**, 58–66.

Taguas, E. V., Gomez, J., Denisi, P. and Mateos, L. (2015) Modelling the rainfall-runoff relationships in a large olive orchard catchment in southern Spain. *Water Resource Management*, **29**, 2361–2375.

Tanasijevic, L., Todorovic, M., Pereira, L., Pizzigalli, C. and Lionello, P. (2014) Impacts of climate change on olive crop evaporation and irrigation requirements in the mediterranean region. *Agricultural Water Management*, **144**, 54–68.

Viola, F., Caracciolo, D., Pumo, D. and Noto, L. (2013) Olive yield and future climate forcings. In *FOUR DECADES OF PROGRESS IN MONITORING AND MODELING OF PROCESSES IN THE SOIL-PLANT-ATMOSPHERE SYSTEM: APPLICATIONS AND CHALLENGES*.

— (2014) Future climate forcings and olive yield in a mediterranean orchard. *Water*, **6**, 1562–1580.