[redacted]

ORIGINAL PAPER

[redacted]

Assessment of flood risk in Mediterranean catchments: an approach based on Bayesian networks

M. Julia Flores[1] · Rosa F. Ropero[2] · Rafael Rumí[3]

[redacted]

Abstract

National and international technical reports have demonstrated the increase of extreme event occurrences which becomes more dangerous in coastal areas due to their higher population density. In Spain, flood and storm events are the main reasons for compensation according to the National Insurance Consortium. The aim of this paper is to model the risk of flooding in a Mediterranean catchment in the South of Spain. A hybrid dynamic object-oriented Bayesian network (OOBN) was learnt based on mixture of truncated exponential models, a scenario of rainfall event was included, and the final model was validated. OOBN structure allows the catchment to be divided into five different units and models each of them independently. It transforms a complex problem into a simple and easily interpretable model. Results show that the model is able to accurately watch the evolution of river level, by predicting its increase and the time the river needs to recover normality, which can be defined as the river resilience.

Keywords Flood risk assessment · Dynamic Bayesian networks · Object-oriented Bayesian networks · Mediterranean watershed

1 Introduction

In the last decade, both national and international technical reports have demonstrated the increase of extreme event occurrences (AEMET 2008). In Europe, this has become more evident with the last series of extreme storm like *Klaus* (2009), *Cyntia* (2010) or the storm seasons which devastated UK in 2013–2014. Nowadays, there is an increased concern about the relationship between extreme events, flooding risk and climate change (Paprotny and Morales-Napoles 2017), which becomes more dangerous in coastal areas due to their higher population density (Bolle et al. 2018).

In a national context, the Spanish Meteorological Agency (AEMET) has predicted a noticeable decrease in the annual rainfall values for the twenty-first century. However, rainfall events will be more extreme and torrential (AEMET 2008). In the collective memory, we found the 2017's winter–spring seasons with a series of heavy storms which provoked extensive damage in the Andalusian coastal area. According to the study of the National Insurance Consortium (CCS 2017), in Spain, flood and storm events are the main reasons for compensation. In the temporal series of 1971–2016, a total of 48.7% of files and a 69.9% of total costs have flood damage as a cause, while those provoked by storms (including damage for rain and snow) conform the 45% of files and the 16% of total costs.

With the objective of reducing the risk of damage for both infrastructures and humans well-being, there is a necessity of creating robust tools to predict the behavior of river levels as an initial step to establish the so-called Alert

✉ Rosa F. Ropero
rosa.ropero@ual.es

M. Julia Flores
Julia.Flores@uclm.es

Rafael Rumí
rrumi@ual.es

[1] Computing Systems Department, SIMD I3A, University of Castilla-La Mancha, Campus Univ., Albacete, Spain

[2] Informatics and Environmental Research Group, Department of Biology and Geology, University of Almería, Almería, Spain

[3] Department of Mathematics, University of Almería, Almería, Spain

Systems. These systems configure a management tool able to provide robust information about the risk of flooding with time enough to adopt emergency security measures (Bolle et al. 2018). These constitute a part of the so-called Watershed Management Plan, defined as a set of processes, tools and systems with the sustainable development as a common objective to optimize a balance between socioeconomic benefits and ecological sustainability (Keshtkar et al. 2013). Thus, an integrated flood risk management is crucial to determine the strategies to be followed, mainly in those areas with a high human density, like coastal areas (Jager et al. 2018).

Risk of flooding has been defined as the probability of an event occurrence and the negative consequences that this event can provoke in human health, cultural heritage, economic activity and environment (Commission 2007). One of the main issues to be solved is how different modeling approaches deal with uncertainty (Keshtkar et al. 2013). Besides, large hydrological dataset is often difficult to deal with and needs powerful statistic tools able to manage them (Papacharalampous et al. 2019). For that reason, there is a growing interest in terms like *adaptive management*, *probability* or *probabilistic management*, and in methodologies which allow uncertainty to be properly dealt with.

Defined at the beginning of the nineties (Pearl 1988), Bayesian networks (BNs) are probabilistic graphical models included in the artificial intelligence and data mining family (Koski and Noble 2011). They have been successfully applied in risk and reliability problems (Langseth et al. 2009) in which their versatility and robustness allow BNs to include large datasets with a high degree of complexity dealing with uncertainty, and even, with spatial and temporal data (Marcot and Penman 2019; Yu et al. 2017; Ropero 2016). Besides, thanks to their intuitive structure, BNs have been included in management systems where experts and stakeholders play an important role (Zhu et al. 2018; Landuyt et al. 2013) configuring a decision support systems in several fields (Chan et al. 2012) and forecasting models (Kim et al. 2018; Dlamini 2010). Because of these advantages, BNs have been applied in risk assessment (Wang et al. 2016; Maldonado et al. 2016) and natural hazards modeling (Malekmohammadi and Moghadam 2018), specially in flood risk assessment (Paprotny and Morales-Napoles 2017).

However, modeling real-life problems often implies facing with really complex systems. In order to deal with these situations, a step beyond BNs is the development of the so-called object-oriented Bayesian networks (OOBNs) (Mortera et al. 2013; Langseth and Bangsø 2001). They have the same advantages of BNs but also, the ability to model really complex systems by dividing the main problem into sub-problems (sub-models). They have started to be applied in risk studies (Liu et al. 2016a) and water management field (Gine-Garriga et al. 2018), but neither in environmental risk management, nor flooding risk management.

The aim of this paper is to model the probability of exceeding the river level, so that, risk of flooding becomes highly probable. To achieve this goal, the behavior of a Mediterranean catchment in the South of Spain was modeled based on a dynamic object-oriented Bayesian network. Section 2 explains in detail the theory behind both BNs and OOBNs, and their adaptation to temporal data. Section 3 describes the methodology followed to model flood risk in *Guadalhorce*, a Mediterranean watershed located in the South of Spain. Section 4 shows the results obtained, and finally, Sect. 5 draws the conclusions achieved and identifies future works.

2 Dynamic object-oriented Bayesian networks

Bayesian networks (BNs) (Jensen and Nielsen 2007) are defined as a statistical multivariate model for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, and made up of two components: *i*) the qualitative part, a direct acyclic graph in which each vertex represents one of the variables (when two vertices are linked by an edge, it indicates the existence of statistical dependence between them); *ii*) the quantitative part, a conditional probability distribution for each variable X_i, $i = 1, \ldots, n$, given its parents in the graph $(pa(x_i))$ expressed as 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 role in the model learning process, mostly identifying relationships between variables and giving values for the CPTs or even refining the structure previously learnt from data (Aguilera et al. 2011; Voinov and Bousquet 2010). This structure also means that, with no mathematical calculation involved, the variable that is relevant (or not) for a certain problem can be known (Pearl 1988). That is, it simplifies the joint probability distribution (JPD) of the variables required 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

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

BNs were originally developed for discrete variables, but real-life problems often require continuous, or even both continuous and discrete (hybrid) data to be included into modeling processes. Initially, the *Conditional Gaussian* (CG) model (Lauritzen 1992) was proposed to deal with continuous data, but it requires data to follow a *Multivariate Gaussian distribution* which is not always fitted by real-life environmental data. Besides, in the case of hybrid data, CG imposes a set of topological restrictions in the network that limit the structure of the model (Aguilera et al. 2011). These limitations have encouraged the proposal of new models for dealing with continuous and hybrid data in BNs. One of these models is the *mixture of truncated exponential* models (MTE). Defined in Moral et al. (2001) and developed in detail in Rumí (2003), MTE models were designed as an approach to include continuous and discrete variables into BNs with no restriction on the network structure. This approximation divides the range of a continuous variable into several intervals, and estimates each of them using a linear combination of exponential functions rather than by a constant (Rumí 2003). Moreover, since they are closed under restriction, marginalization and combination MTE models are appropriate for performing inference. This model is able to accurately approximate any theoretical distribution function because of its high fitting power. See Rumí et al. (2006), Rumí and Salmerón (2007), Cobb et al. (2007) for more information about it.

There are many other kinds of probabilistic graphical models, some of which are described in Koller and Friedman (2009) including: variants of BNs, such as influence diagrams (also called Bayesian decision networks), which extend BNs with decision and utility nodes to support decision making; dynamic Bayesian networks (DBNs), which explicitly model changes in the system over time, and object-oriented Bayesian networks (OOBNs), which refer to a particular form of BN whose principles are based on object-oriented philosophy. OOBNs are hierarchical and compositional BNs which support incremental construction and re-usability of sub-models. They provide a very powerful representation to model particular types of problems, where there is repetition of the same kind of elements.

A standard BN is made up of ordinary nodes, representing random variables. An OOBN class is made up of both nodes and objects, which are instances of other classes. Thus, an object may *encapsulate* multiple sub-networks, giving a composite and hierarchical structure. Objects are connected to other nodes via some of its own ordinary nodes, called its *interface* nodes. The rest of the nodes are not visible to the outside world, thus providing so-called *information hiding*, another key concept. A class can be thought of as a self-contained template for an OOBN object, described by its name, its interface and its hidden part. Finally, interface nodes are divided into input nodes and output nodes. Input nodes are the root nodes within an OOBN class, and when an object (instance) of that class becomes part of another class, each input node may be mapped to a single node (with the same state space) in the encapsulating class. The output nodes are the only nodes that may become parents of nodes in the encapsulating class. Connections to and from an object must be such that the underlying BN is still a directed acyclic graph.

Both BNs and OOBNs can be used to obtain prediction about the change of the system under some (future) scenarios, but the conclusions reached cannot be extrapolated to a particular time, nor time series can be handled. For these reasons, the extension of BNs, the so-called dynamic Bayesian networks (DBNs) (Korb and Nicholson 2011), has begun to be applied to face the new challenge of including time in the model (Molina et al. 2013). The first attempt to deal with time using BNs appeared in Provan (1993), which proposed their use for modeling a generic system in each time step, joining the BNs with links which represent the transition from one time to the next one. They were defined as (Nicholson and Flores 2011): *A long-established extension of BNs that can represent the evolution of variables over time.*

The term dynamic means the system is changing over time, not that the network and the relations between variables change (Murphy 2002). For simplicity, it is assumed that a DBN is a time-invariant model composed by a sequence of identical BNs representing the system in each time step, and a set of temporal links between variables in the different time steps representing a temporal probabilistic dependence between them (Pérez-Ramiréz and Bouwer-Utne 2015).

In order to reduce the potential number of temporal parents in the network, and also the computational cost, the *Markov assumption* is followed (Murphy 2002). That is that *the state of the world at a particular time depends on only a finite history of previous states*. In the simplest case, the current state of the system depends only on the previous state, called a *first-order Markov process*. Given these restrictions, a DBN can be represented with only two consecutive time slices (time 0 and time 1) and the relationship between them. Only if necessary, the DBN can be rolled out and more than two time slices would be represented.

DBNs are still a scarcely used modeling approach in environmental science, but preliminary applications show powerful and promising results in real-life applications (Yung et al. 2016).

Recently, a novel approach called extended OOBN (EOOBN) has been presented (Liu et al. 2016a), where parameters in OOBNs may vary among the different instantiations/objects of the network. The modeling scheme proposed also includes the dynamic dimension. This model has been successfully used for representing risk assessment of flash floods (Liu et al. 2016b), which indicates that this approach seems promising to the current problem. However, this methodology is considerably distinct to our problem, because we deal with continuous values and distributions, while the mentioned work dealt with discrete values and CPTs. They share the capability to change the conditional probabilities, as in our learning process the instances parameterizations are estimated independently too.

Hence, in our paper we use a combination of these two dynamic BNs (DBNs) to model decision making regarding flooding risk (over time) between time-dependent variables, and OOBNs to represent the watershed structure in different objects that correspond to the different river units identified.

3 Flood risk assessment based on Bayesian networks

3.1 Study area

Guadalhorce catchment is located in Málaga province, Andalusia, in the South of Spain (Fig. 1). It is considered one of the most important rivers in Andalusia in terms of length, over 165 km, and flow rate (estimated at 8 m^3 per second according to local government). Andalusian Regional Government determines the so-called areas of potential risk of flooding (APRFs) based on historical data, soil, geomorphologic characteristics and potential impact over human infrastructures and society. For *Guadalhorce* catchment, a total of three APRFs (Fig. 2) have been identified due to its high population settlement and agriculture activity. In the last years, these areas, and Málaga province in general, have suffered events of heavy storm that have provoked critical damage.

This catchment is limited in the North by *Sierra de Archidona* mountainous range, in the East by the *Gibalto, San Jorge, Jobo* and *Camarolos* mountainous ranges, by *Sierra de las Nieves* mountainous range in the West, and Mediterranean Sea in the South. The main tributaries of *Guadalhorce* River are the rivers *Grande, Turón* and *Guadalteba* (Fig. 1). In its incipient watercourse, *Guadalhorce* river is shaped by the orographic conditions of the territory characterized by limestone, clay and gypsum. Then, it runs from west to east through the so-called *Depresión de Antequera*, an area of important agricultural tradition, to come across the vast alluvial plains of a group of municipalities belonging to the Guadalhorce's valley region, which is well known for its irrigated agriculture.

Historically, this area has had a notable population and agricultural activity. But the irregular flow regime of *Guadalhorce* River, characterized by severe droughts and flash floods, has encouraged dam constructions in its middle course. So that, from the beginning of the twentieth century, several hydrological infrastructures have been constructed in order to supply water, regulate water flow,

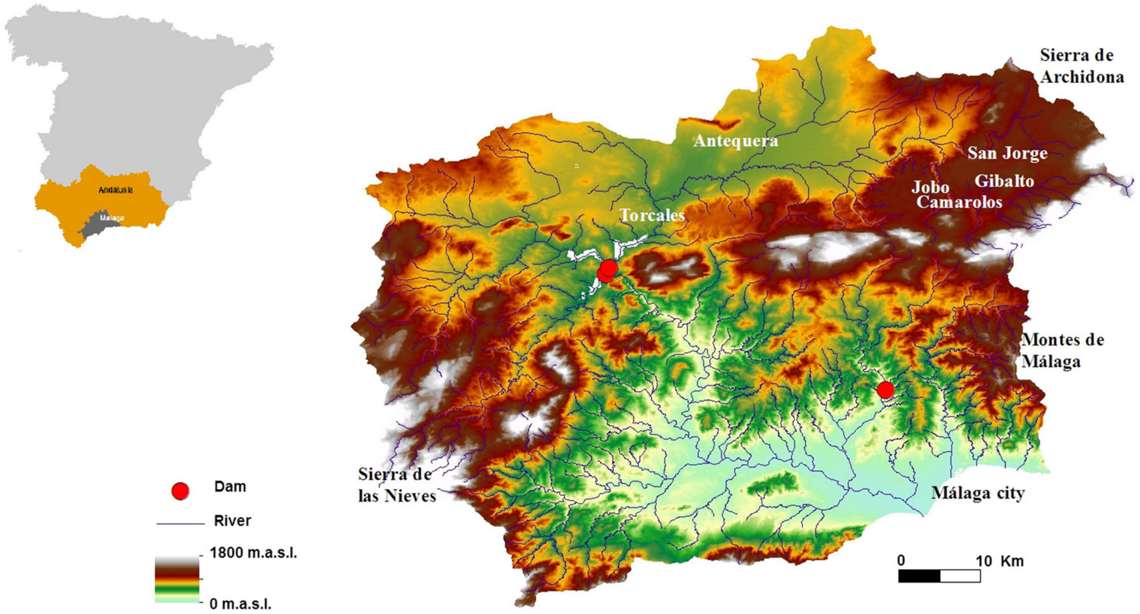


Fig. 1 *Guadalhorce* catchment, its location, relief and hydrographic systems. Dams are marked in red. Only the three dams used in the experiments are shown

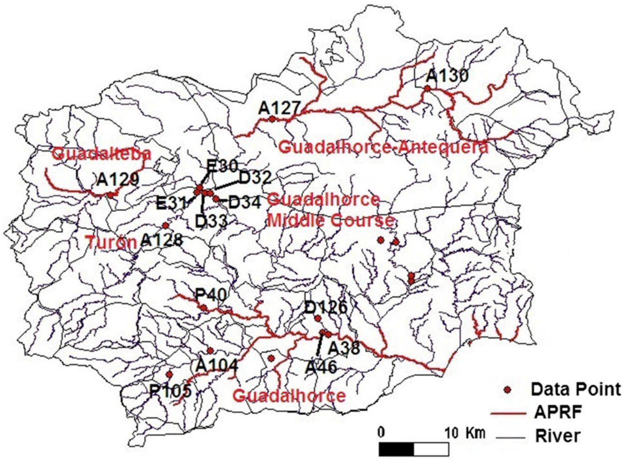


Fig. 2 Points for data collection, sub-catchments and areas of potential risk of flooding (APRF) in the *Guadalhorce* catchment. Those data points with no label were rejected because of the lack of data

provide electricity to the cities and reduce damages provoked by flood and drought. There are a total of five reservoirs in the *Guadalhorce* catchment: *Guadalhorce* reservoir, in the *Turón* River, *Gualdalteba* reservoir, in the *Guadalteba* River, *Guadalhorce* reservoir (connected to the *Guadalteba* reservoir when there is a high level of water in both reservoirs), *Gaitanejo* reservoir and *Tajo de la Encantada* reservoir. Because of the lack of data in the last two reservoirs, they will not be taken into account in this study. Finally, *Guadalhorce* flows into the Mediterranean, close to the capital city of Málaga, in an area that was once a deltaic plain, occupied by marshes that fed on the winter floods, but the reservoirs upstream provoked their disappearance.

Climate in this area is Mediterranean, but the steep relief determines stark differences between areas. The presence of two main mountain ranges, *Sierra de las Nieves* and *Torcales*, acts as barrier reducing the rainfall coming from the Atlantic sea and, also, the coastal influence in local climate. It defines three areas from the climate point of view. The upper catchment, around the municipality of *Antequera*, is the area with the highest variability of temperatures with the coldest values in winter, but also being the warmest in summer, due to its inland location. Secondly, in the middle course, the area around the dams lies between both mountain ranges and presents the highest rainfall values of the catchment. Finally, the lower part has the warmest climate due to the coastal influence. Despite these differences, in terms of rainfall values, autumn and spring seasons are characterized by strong storms which can provoke serious damage in infrastructures, and also, humans well-being, mostly on the upper and middle part due to the mentioned relief.

Table 1 Stations and variables collected in *Guadalhorce* catchment

Station	Type	Variables collected
A130	Hydrological	Level, rainfall
A129		
A128		
A127		
A104		
A38		
D34		
D126	Meteorological	Rainfall
P105		
A46		
P40		
D33		
D32		
E31	Dam	Level
E30		

3.2 Data collection

Data were collected from the Hydrological Information Systems[1] (*Sistema Automático de Información Hidrológica*, SAIH). This system allows data to be obtained in different time periods (monthly, weekly, daily and per hour) for a set of stations in different Mediterranean catchments. These are from three different types, namely dams, meteorological and hydrological stations. Information is mainly related to rainfall patterns and the level of the river in the different points of interest.

For our study, data for *Guadalhorce* catchment were collected per hour from October 2013 to March 2018 (both included). A total of 15 stations—over 20 available—were selected along the riverbed (Fig. 2). The rest were omitted because of their incomplete data series. Table 1 shows a summary of variables collected in each type of station used.

Final dataset contains a total of 33.252 observations. We have followed the division into hydrological years (from October to September), and used just those complete years, from October 2013 to September 2017, for learning and validation processes. Once the model was validated, a scenario of storm event was included into the model with the aim of evaluating its predictive accuracy. For this purpose, the remaining last 6 months (from October 2017 to March 2018) were avoided and just a period of 10 days was used. This period of time includes a significant storm

[1] http://www.redhidrosurmedioambiente.es/saih/presentacion.

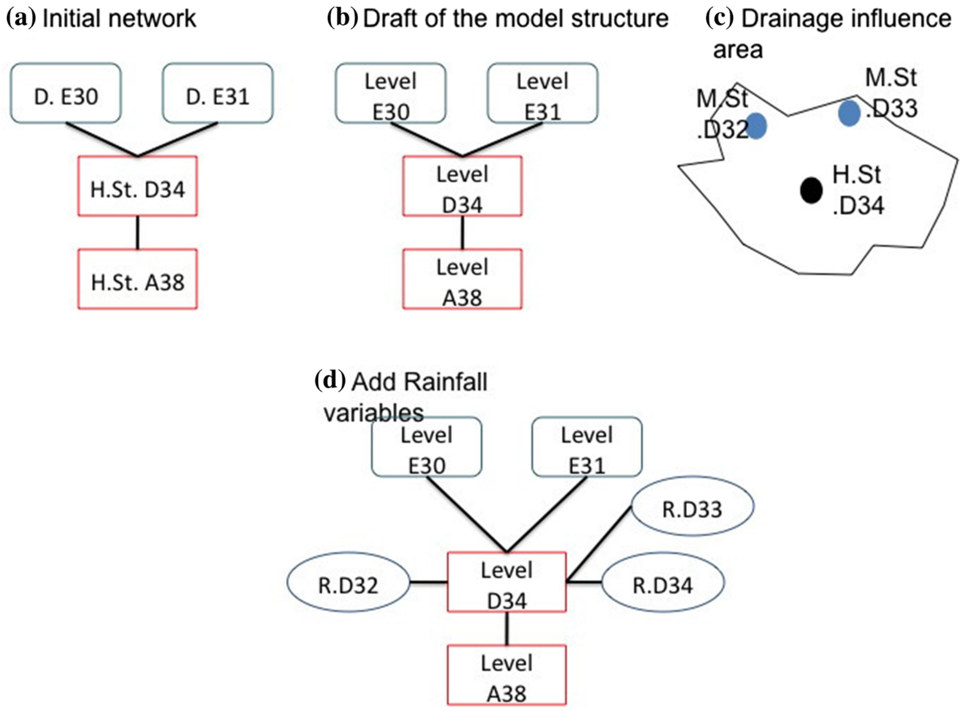


Fig. 3 Procedure to determine the static OOBN structure. Based on the territorial distribution of the watershed, an initial network described the relationships between the stations following the riverbed and connected the variables of level between them. For each hydrological station, its drainage influence area was obtained using the corresponding ArcGIS Toolbox and those meteorological stations lie in were linked to it. So that, rainfall variables were included into the model and connected to the corresponding Level variable. *D* dam, *H.St.* hydrological station, *M.St.* meteorological station, *R.* rainfall

event that takes place from October 16th to 26th, 2017, and provoked flood and damage in human infrastructure.

3.3 Structural model learning

The idea is to model the temporal evolution of the river level. Following a simple approximation, it could be considered that predicted water level at hour t would be equal to the water level at hour $t-1$, which allows to reduce complexity (Papacharalampous et al. 2019). However, the study area lies into a Mediterranean climate, characterized by sudden and heavy storm events, which can produce a quick and dangerous increase of river level. So, water level should not be the same at t than at $t-1$, and a more complex model is needed. In this paper, a model based on OOBN is developed.

There are two main ways of determining the structure of a BN model (Ropero 2016): by automatic structural learning algorithms, and by hand using literature and expert opinion. In this paper, the structure of the model was developed by hand, following the physical networks of relations among the different data points. In this section, an example of model construction is developed, but the final complete model is presented in following sections.

According to Fig. 2, *Guadalhorce* River connects the different points between them and determines the possible structure of the model and the division in different sub-catchments. In this sense, firstly, an initial network of relations between the stations in *Guadalhorce* catchment was defined, and the different areas were separated in order to define the sub-models (Fig. 3). From this initial network, variables of Level for each Hydrological station were linked to each other configuring a draft of the final network structure. By this way, this initial network represents the river flow and connects all the stations between them following a causal relation along the riverbed. Next, rainfall variables were included into the model structure. In this case, there were two possible sources of rainfall information. In the one hand, each hydrological station provides values of rainfall and they were included into the model as new nodes connected to their corresponding Level variable (for example, hydrological station A130 provides two different variables: level at A130 and rainfall at A130). In the other hand, apart from hydrological stations, data were collected from meteorological station and are distributed in the middle and lower part of the catchment. In these cases, the relationships of these new nodes (rainfall at meteorological stations) and the level variables need to be established according to their geographical situation. So that, for each hydrological station, its drainage influence area was calculated using the ArcGIS toolbox,[2] and those meteorological stations belonging to this influence area were connected in the network to the corresponding level variable. For example, the hydrological station *D34* is located in the middle area, and the variable level at *D34* was initially connected upland with both variables of level at *E30* and

[2] For a detailed information of this method see http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/understanding-drainage-systems.htm.

E31 dams, and downland with variable of level at *A38* hydrological station (Fig. 3). Now, rainfall variables can be included in the net, by calculating the drainage influence area of *D34* station and checking that both *D32* and *D33* meteorological stations lie in. So, both Rainfall from *D32*, *D33*, and even *D34* stations, were included in the network and linked with the Level variable from *D34* station. This process was repeated with all stations and the structure for the OOBN was finally completed.

Notice that we use OOBN framework for modeling purposes, but this does not imply that every instance of river unit is identical. On the one hand, the parameters for the probability distributions may differ as they are learnt from the corresponding data. Furthermore, when modeling the particular unit we also consider the territory structure, so that we consider the individual features such as rain trajectory and riverbed.

However, the aim of this model is to evaluate the risk of flooding by means of observing the level of the river along time. To accurately determine the level of the river in a specific time, information about the temporal evolution of this river is necessary. That is why this OOBN model needs to be transformed into a temporal model.

To transform a static (OO)BN model into a dynamic one, the most common solution consists in duplicating the static network structure, and connecting nodes by temporal links [detailed information about this methodology can be found in Ropero et al. (2018)]. It means all relations included in the network will be replicated and connected, as it is shown in Fig. 4, option a).

In our model, it would imply repeating both rainfall and river-level variables. However, as a modeling decision we did not include temporal relation in rainfall variables, for the sake of simplicity. Given the climate conditions of this particular geographical area, the value of rainfall at time $i + 1$ is not strongly dependent on the rainfall at i. Rain can happen abruptly at any time, and it is, in fact a rare occurrence. So, this is the circumstance we would like to predict, when sudden rains could arrive. To empirically prove this *weak dependence*, we have computed statistical information. Per every rainfall variable, we got the correlation values at time i with respect to the same variable at time $i + 1$. When all the data points are used, these correlations values are mostly below 0.50, having only five stations out of 13 with a value between 0.5 and 0.64. When removing those data points where the Rainfall value is 0.0 both for moments i and $i + 1$, which are not informative for our purposes, the correlation is effectively reduced for all the stations. In this second case, 9 of 13 stations present correlation lower than 0.40, and the most correlated does not even reach 0.55.

By contrast, to evaluate the river level and the risk of flooding, the behavior of the river in the previous time is important. We have empirically corroborated that the dependence between river level at time i and $i + 1$ is strong enough to be included into the model (mean correlation close to 0.9). Here, we could ask how much information would be relevant, 1 h, 2 h, or more? Data available present a time step of 1 h, so that lower range is not feasible. The other idea would be to increase the time step, 2 h, 3 h or more. The goal of the model is to predict the behavior of

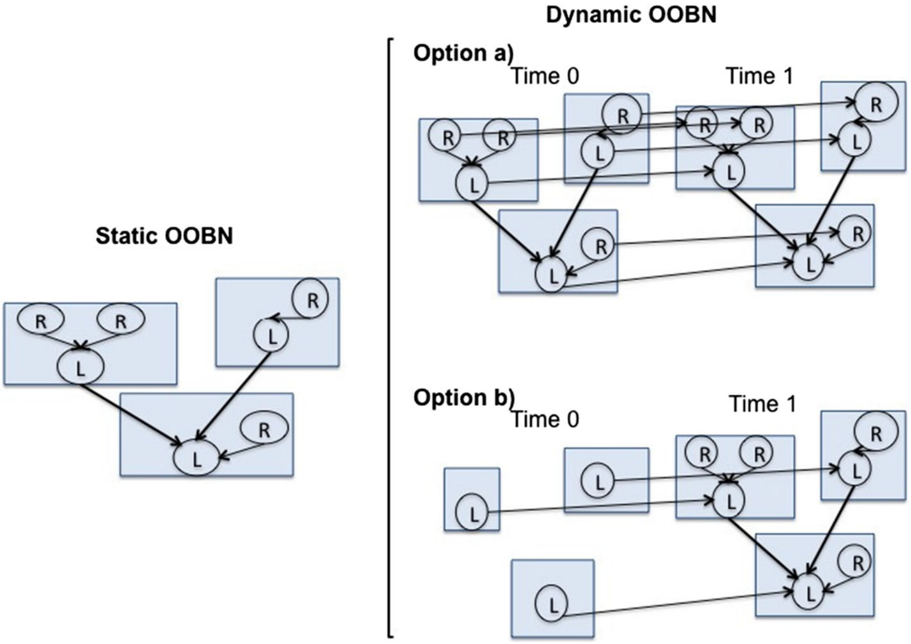


Fig. 4 Example of the methodology followed to transform a static OOBN into a dynamic OOBN. *R.* rainfall, *L.* level

the river during a storm event, which is usually short in time but intensive in Mediterranean areas. It means that, considering a large time step probably implies losing information for the river behavior. For that reason, 1 h time step is considered. Thus, only level variables will be definitely duplicated as it is shown in Fig. 4, option b).

3.4 Parameter estimation and model validation

As in the model structure learning, there are two main approaches to estimate the parameters in a BN: by experts or literature. These approaches are mainly used in the case of discrete variables, and through the use of automatic learning algorithms that estimate the parameters from the provided data.

In this paper, since all variables are continuous, and there is a large dataset available (from October 2013 to September 2017, 33.252 observations), automatic parameter estimation was carried out using Elvira software (Elvira-Consortium 2002). A 5-parameter MTE (*mixture of truncated exponential* model, explained in Sect. 2) distribution was fitted for every probability distribution due its ability to fit the most common distributions accurately, while both model's complexity and the number of parameters to be estimated remain low. This function fits a function of the form $p_i(x) = k + ae^{bx} + ce^{dx}$, where k, a, b, c and $d \in R$, for every piece in which the domain of the variable is divided, according to changes in increasing/decreasing and concavity/convexity. This way every function $p_i(x)$ fits an easy-shaped function, and so it is able to accurately approximate a wide variety of functions. In the case of conditional distributions (for non-root variables in the graph), the domain of the conditioning variables is also split according to equal frequency criteria. So, in fact, the 5-parameter MTE function is piecewise function where, in each piece i a function $p_i(x)$ if fitted to the data. See Moral et al. (2001) and Rumí et al. (2006) for more information.

In order to validate the complete model, k-fold cross-validation (Stone 1974) process was applied. It is a widely used technique in artificial intelligence in which the aim is to check how predictive a model is when confronted with data that have not been previously used for learning. It is based on the holdout method in which the data set is separated into two complementary sets, one for learning (D_l) and another for testing (D_t). In this way, we can estimate the error of a model built from D_l according to set D_t. To check the accuracy of the model, the *root mean square error* (RMSE) (Witten and Frank 2005) was calculated for level variables:

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

where $y_1, \ldots, y_n$ are the actual values of the level variable, and $\hat{y}_1, \ldots, \hat{y}_n$ are the values predicted by the model.[3]

To reduce variability, the dataset is initially divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as D_t and the other $k-1$ subsets are put together to form D_l. For the case study presented in this paper, and due to the temporal nature of the data, a forward k-fold cross-validation was performed. It means the first D_t corresponds to the first hydrological year (October 2013 to September 2014), and the first D_l the second hydrological year (October 2014 to September 2015). Later on, the second D_t corresponds to the first and second hydrological years (October 2013 to September 2015), and the second D_l, is the third hydrological year (October 2015 to September 2016) and so on. Then, the average error across all k trials, 3 in our case, is computed.

3.5 Scenario of rainfall event

Once the model is obtained 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 $\mathbf{E}$, 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 $X_i \not\in \mathbf{E}$:

$$p(x_i|\mathbf{e}) = \frac{p(x_i, \mathbf{e})}{p(\mathbf{e})} \propto p(x_i, \mathbf{e}), \tag{3}$$

since $p(\mathbf{e})$ is constant for all $X_i \not\in \mathbf{E}$. So, this process can be carried out computing and normalizing the marginal probabilities $p(x_i, \mathbf{e})$, in the following way:

$$p(x_i, \mathbf{e}) = \int_{\mathbf{x}} \not\in \{x_i, \mathbf{e}\} p_e(x_1, \ldots, x_n)\mathrm{d}x, \tag{4}$$

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 $\mathbf{E}$ by their values $\mathbf{e}$.

Distribution $p(x_i, \mathbf{e})$ would be the result of the inference process, i.e., the output predictive probability distribution. This is one of the advantages of modeling using Bayesian networks; the output is a complete distribution of values for the goal variable (Fig. 5), instead of just a single-point prediction. However, since for validation purposes we need to compare to the real values, a point prediction $\hat{y}$ is computed:

[3] Computation of values $\hat{y}_i$ is explained in Sect. 3.5.

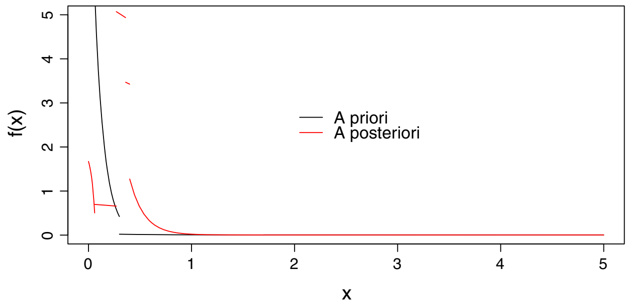


Fig. 5 Example of probability distributions both a priori and a posteriori for A130 variable. The corresponding forecasts are obtained computing the expected value for the a-posteriori (red-line) density

$$\hat{y} = \int x\,p(x_i, \mathbf{e})\mathrm{d}x_i$$

Inference process was applied for validation purpose using real observations from October 16th to 26th, 2017, when a series of rainfall events were detected (please note that this dataset was not previously used for the learning and validation processes). Data for rainfall variables were included as *evidence* and propagated to obtain the predicted posterior probability distribution of level variables.

Besides, in order to compare with the naive assumption mentioned in Sect. 3.3, inference is done, but considering water level at $t-1$ is equal to its initial value in the dataset (it means, the water level at t_0), and RMSE is also computed. Since it is just a simple comparison, only one variable is shown.

4 Result and discussions

Figure 6 shows the model obtained. The OOBN was divided into five different units or BNs corresponding to the physical structure of the watershed. In the upper area, the *Guadalhorce-Antequera* river unit corresponds to the beginning of river course, and two data points are located (A130 and A127) with their corresponding rainfall and level information. Besides, in this upper area but located in the west, the *Guadalteba* and *Turón* river units are found, both with just one station. These three units converge into the *Guadalhorce middle river course* unit, where both E30 an E31 dams, and D34 hydrological station are located. Finally, the *Guadalhorce* river unit represents the lower part of the watershed with two hydrological and four meteorological stations.

Temporal behavior is represented by a set of level variables (*i.e., Level A129_0*) marked in white and linked to the present model with dashed lines. These variables represent the state of the river level in the previous hour.

Table 2 shows the RMSE values for all level variables, and their range. This metric gives us information about the difference between the predicted and real value. Taking into account the range of the variables, the RMSE for level variables is relatively low in all variables except dams. In our model, just information about rainfall and river level was considered; however, the level achieved in a dam depends not only on the incoming water (through rainfall and runoff coming from its drainage area), but also on the water consumption (both human and agriculture), and even on management decisions (pumped water from other dams).[4] For these reasons, the difference between the value predicted by the model and the real value is higher. This is an initial approximation of flooding risk modeling in Mediterranean watershed through dynamic OOBN that, as far as we know, has not been previously deal with. So the aim of this paper is to provide a simple and easy to understand model to show a real application and the potentiality of modeling this problem using BNs. This decision is also related with the fact that data obtained for learning and validation purposes include only drought years. The official website used for data collecting can provide data from 2013, information prior this date was not available, so just the last drought period was used. It means that dams were most of the time under their values of maximum capacity, so, during a storm event, they can be used as flood container and avoid flooding in the upper area has a deep impact over the middle and lower area of the catchment.

Once the model was obtained and validated, a scenario of Rainfall event was included into the model. In this scenario, we have considered dams were under their maximum capacity, so that they can admit the increase of river level and avoid this increase to reach the middle and lower river course. Besides, considering that the prediction of dam levels needs further information (which is not available) for the analysis of the scenario results, dam's level was not taken into account. First, Fig. 7 shows the temporal series of rainfall variables used as evidences in the scenario proposed. During 18th and 19th October, there was a storm event that provoked a heavy storm in all units of our model with the higher values in the lower part, *Guadalhorce* unit, mainly in the area of the A38 data point which is the closer point to the city of *Málaga* and lies in an area considered to be at high risk of flooding. According to the methodology described in Sect. 3.5, new information was included into the model for just rainfall variables, so that, water-level variables at time $t-1$ are computed by the model and it is not fixed. However, in order to compare with the naive assumption mentioned previously, water level at $t-1$ is considered constant. Taking variable A127

[4] This information was not available for this study.

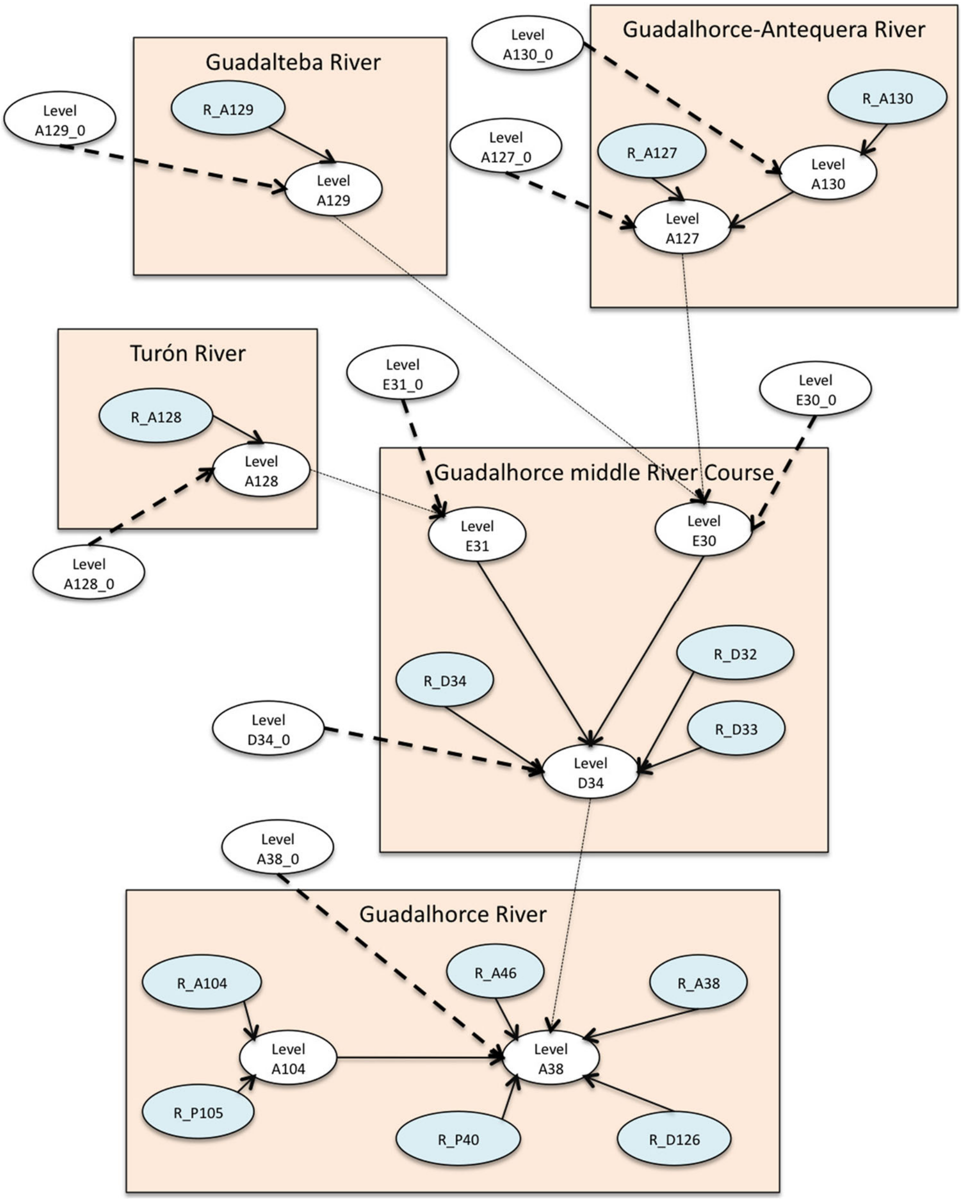


Fig. 6 Structure of the final dynamic OOBN obtained. The final notation "_0" indicates the variable in the previous time. Dotted lines represent the temporal links, while solid lines represented the links between variables in the same time step. *R* rainfall

as an example, the computed RMSE for this naive assumption reaches 0.29, against 0.018 obtained when these variables are not constant (and its values are computed by the model). Results obtained are able to predict the river flood (Fig. 8), while in the case of naive approximation, is always constant, so risk of flooding could not be studied.

Results of the scenario proposed in the upper area are shown in Figs. 9 and 10 in which four data points were evaluated, those included into *Guadalhorce-Antequera* units (A127 and A130), and *Turón* and *Guadalteba* units (A128 and A129). The model provides the results as a set of posterior probability distributions. For making the interpretation of results easier, the means of the distributions are computed and compared with the real observed values. So that, values predicted by the model (red) were represented against real values (black) for all data points, with the RMSE obtained. Besides, the degree of river-level change is shown in Figs. 9 and 10b.

The storm during the 18th and 19th October (observations from 50 to 100) provoked an increase in the river level in all stations, which come back to the initial levels in A127 and A129, but not in A130 where level after the storm was higher. Predictions made by the model show the same tendency of level increase and, also, decrease. In the case of A130, the results predict the increase in the river level after the storm period. The percentage of level increase was calculated for all stations and shown in Fig. 9b). Again, model results are in concordance with the real observations.

Table 2 Variable ranges and RMSE values from the forward 3-cross validation

Variable	RMSE	Variable range
Level A38	0.059	0.0–1.77
Level A104	15.81	0.0–36.0
Level A127	0.408	0.0–8.1
Level A128	0.493	0.0–13.1
Level A129	0.216	0.0–3.0
Level A130	0.132	0.0–5.0
Level D34	0.008	0.0–2.0
Level E30	11.26	347.0–362.0
Level E31	37.97	327.0–355.0

Level variables are expressed in meters

In the upper area, this storm event provoked more than 10 mm per hour in the *Guadalhorce-Antequera* unit (Fig. 9), where A130 station suffered an increase in its river level from 0.1 m to close to 1 m (more than 2% of change each hour). This increase is even more evident in the A127, located down in the riverbed, in which the level increases from 0.2 m to more than 1.5 m (more than 4% of change each hour). At the end of the storm, river does not recover back to its initial level in A130 station, but in A127. By contrast, in *Guadalteba* unit, represented by A129 station, rainfall hardly reaches 7 mm per hour and does not provoke an important river-level change (0.1–0.15 m and less than 0.5% of change). These two units lie on areas of potential risk of flooding (Fig. 2) since their steep relief and soil characteristics encourage a rapid runoff which feeds the river level. Besides, results show that the river presents a high resilience since it is able to recover its initial level after the rainfall event.

By contrast, *Turón* (Fig. 10, A128 station) is not considered as an area of flooding risk, and it is explained through the results achieved. Figure 7 shows that rainfall values around A128 station reached more than 5 mm per hour; however, river level in this station is hardly affected (around 1% of change).

Peculiar results are found in the case of D34 station (*Guadalhorce middle course* unit). Figure 11 shows that no change was predicted in the river level for this station. This unit divides the watershed into the upper and lower area. Dams located in this unit do not just provide water supply for population and agriculture, but they are also designed to control the river course and flooding. The four hydrological years used for learning and validation purposes correspond to drought years in which the dams level hardly achieved the higher values. Thus, in this period, when a rainfall event or storm affected the upper areas, the impact on the lower area was minimized by the dams which functioned as

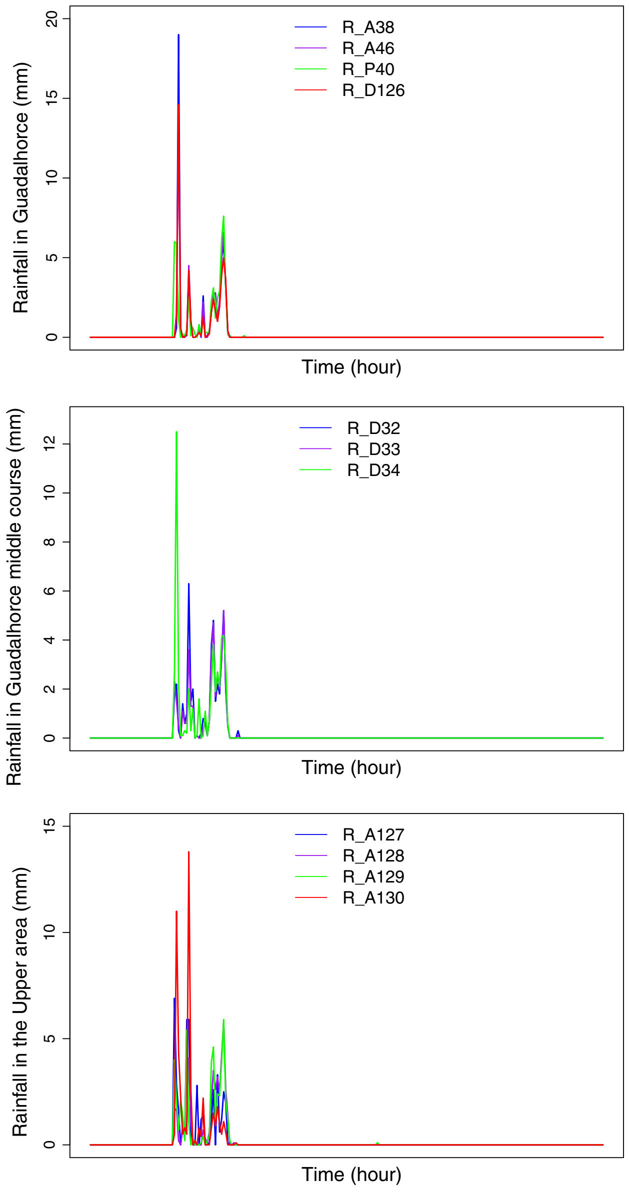


Fig. 7 Rainfall variables for the scenario of rainfall event, from October 16th to 26th, 2017

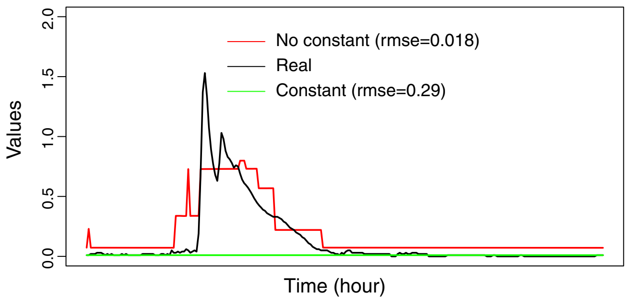


Fig. 8 Comparisons between prediction made considering water-level values constants (marked in green), against computing them by the model (marked in red)

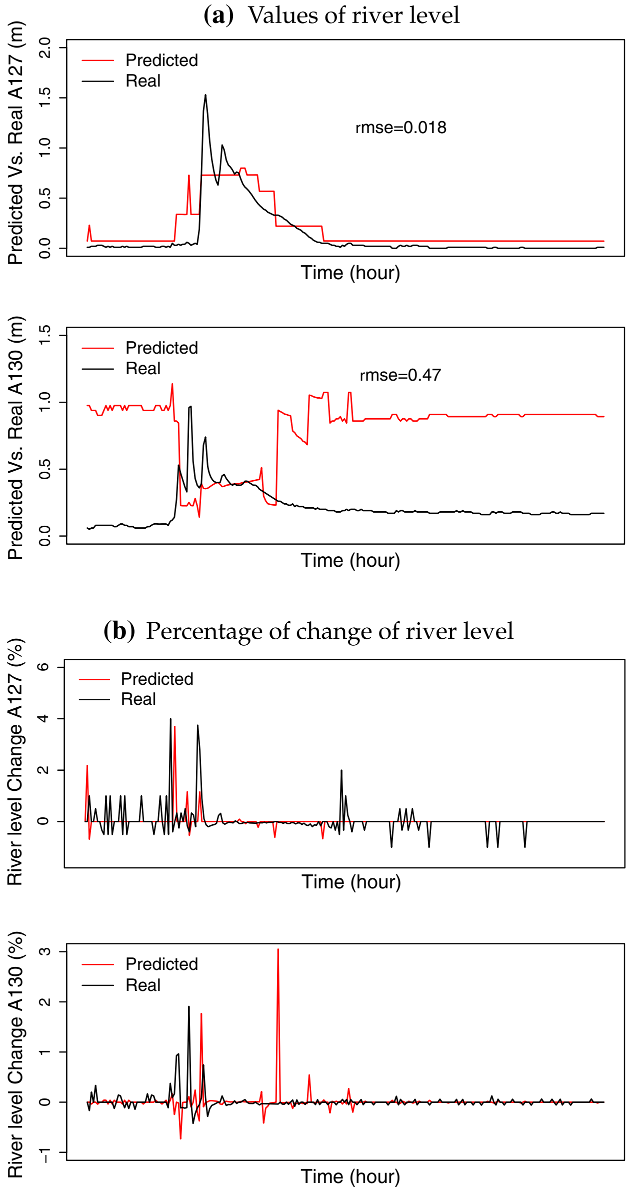


Fig. 9 Predicted versus real values and the RMSE obtained (**a**), and evolution of the river-level change (**b**) in the river points of the *Guadalhorce-Antequera* unit, for the scenario proposed, from October 16th to 26th, 2017

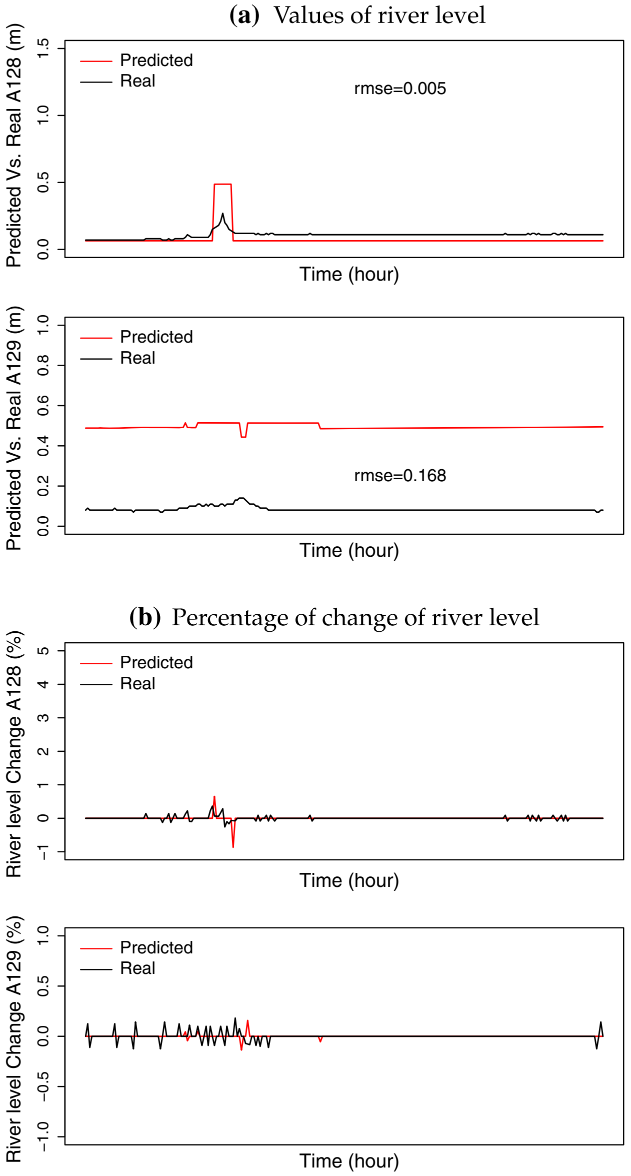
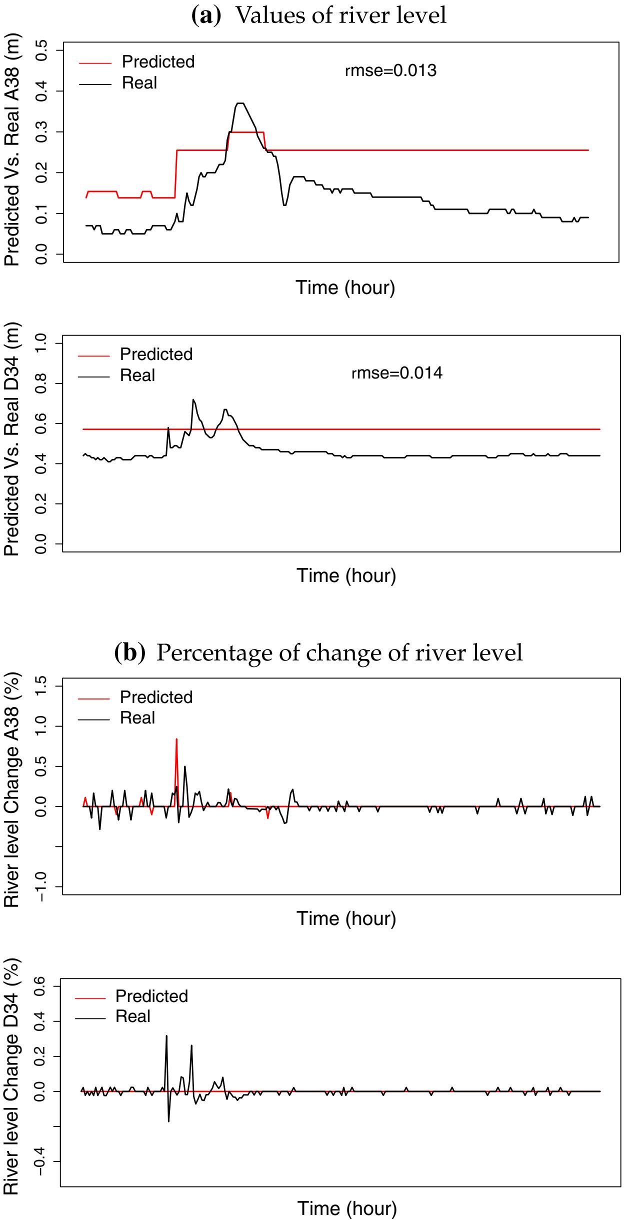


Fig. 10 Predicted versus real values and the RMSE obtained (**a**), and evolution of the river-level change (**b**) in the river points of the *Turón* and *Guadalteba* units for the scenario proposed, from October 16th to 26th, 2017

containment for the flood. In this case, since D34 is located just below the dams, real values show changes lower than 0.4% (from 0.4 to 0.7 m) which were probably due to the minimum water flow provided by the dams to maintain the natural river regime downstream during drought periods (Guadalquivir Plan 2007). This behavior responds to water management decision not to rainfall values, so that model predicts no changes into D34 river level as a consequence of the storm.

Finally, Fig. 11 shows the results for the *Guadalhorce* (A38 station) unit. In this area, rainfall reached the higher values in this storm (Fig. 7), but the impact on the river is hardly noticeable (slight percentage of changes, and an increase from 0.1 m to less than 0.4 m). In this area, the river is close to the sea and the soft relief and wider riverbed makes this river area more stable and resilient, since in spite of the high values of rainfall, the river hardly appreciates the impact. However, due to the proximity of important human infrastructures and population (it is the closer data point to *Málaga* city, with more than 570.000 inhabitants), it is considered as an area of potential risk of flooding (Fig. 2). During drought periods, since the dams are under their maximum capacity, they contain the flood avoiding storms in the upper areas can have an important impact on lower part of the watershed.


(a) Values of river level

(b) Percentage of change of river level

Fig. 11 Predicted versus real values and the RMSE obtained (**a**), and evolution of the river-level change (**b**) in the river points of the *Guadalhorce middle course* and *Guadalhorce* units for the scenario proposed, from October 16th to 26th, 2017

5 Conclusions

Flooding risk management is a crucial aspect of the so-called Watershed Management Plan and Alert Systems. In coastal areas, due to their higher density of human infrastructures, the vigilance and control of river levels becomes an essential task which helps to reduce economic and society's vulnerabilities. The risk of flooding has been defined as the probability of an event occurrence and the negative consequences that this event can provoke, where the main issue to be solved is related with the uncertainty inherent to this natural process.

In this paper, a dynamic object-oriented Bayesian network was applied to determine the probability of exceeding the river level in order to predict the risk of flooding in *Guadalhorce* catchment, in the South of Spain. The use of OOBNs has allowed the complexity of this catchment to be divided into five different units according to the physical structure of the territory. Each of them was modeled independently but connected to each other, in such a way that, if a disturbance takes place in one unit, the focus of our attention can be set on just the same unit, in a related one, or even in all the system. In the proposed scenario of rainfall event, new information about rainfall data was included in all units, but the results were evaluated per unit.

These results show that the model obtained presents low error rates for river-level variables. This implies that, under the scenario proposed, the predicted values do not present a deep difference, neither high errors, when confronted against the real values. During an event of rainfall or storm, the model is able to predict the increase in the river level, and also the time the river needs to recover the normality, which means the resilience of the river.

Through this modeling approach, uncertainty is managed using probability theory; a well-founded formalism and there are a wide range of algorithms and procedures, already developed and validated in the literature, which can be applied to parameter estimation and inference. In our case, the use of dynamic OOBNs contributes to uncertainty analysis in several ways: (1) in the representation of the complexity and (in)dependences of the variables of the model which, by its nature, is intuitive in this type of model, and (2) by the different ways in which results can be displayed. All these advantages make up a broad range of tools to aid the decision making by experts regarding the uncertainty in the framework of adaptive catchment management.

Certain issues remain to be addressed in the future, which could improve the application of this tool in flood risk management. Firstly, in this paper a simple and easy model was obtained in order to show the applicability of dynamic OOBNs in this field, but information about management decisions, natural characteristics of the riverbed and consumption rates could be included to make the model richer. Besides, in Mediterranean areas there are successive cycles of drought and humid years, but data collected correspond only to a drought period. It means our model is still no able to predict the behavior of the river level during a humid period. For future works, dataset used for model learning needs to be extended to include both periods.

Acknowledgements This study was supported by the Spanish Ministry of Economy and Competitiveness through Projects TIN2016-77902-C3-1-P and TIN2016-77902-C3-3-P, and by the Regional Government of Andalusia through project P12-TIC-2541.

References

AEMET (2008) Generación de escenarios regionalizados de cambio climático en España. Informe técnico. Technical report, Ministerio de Economía, Industria y Competitividad

Aguilera PA, Fernández A, Fernández R, Rumí R, Salmerón A (2011) Bayesian networks in environmental modelling. Environ Model Softw 26:1376–1388

Bolle A, das Neves L, Smets S, Mollaert J, Buitrago S (2018) An impact-oriented early warning and Bayesian-based decision support system for flood risks in Zeebrugge harbour. Coast Eng 134:191–202

CCS (2017) Estadística de Riesgos Extraordinarios. Serie 1971–2016. Technical report, Consorcio de Compensación de Seguros

Chan TU, Hart BT, Kennard MJ, Pusey BJ, Shenton W, Douglas MM, Valentine E, Patel S (2012) Bayesian network models for environmental flow decision making in the Daly river, Northern territory, Australia. River Res Appl 28:283–301

Cobb BR, Rumí R, Salmerón A (2007) Bayesian networks models with discrete and continuous variables. In: Lucas P, Gámez JA, Salmerón A (eds) Advances in probabilistic graphical models. Studies in fuzziness and soft computing. Springer, Berlin, pp 81–102

Commission E (2007) European Commission, 2007. Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks

Dlamini WM (2010) A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. Environ Model Softw 25:199–208

Elvira-Consortium (2002) Elvira: an environment for creating and using probabilistic graphical models. In: Proceedings of the first European workshop on probabilistic graphical models, pp 222–230. http://www.ia.uned.es/investig/proyectos/elvira/

Gine-Garriga R, Requejo D, Molina J, Perez-Foguet A (2018) A novel planning approach for the water, sanitation and hygiene (wash) sector: the use of object-oriented Bayesian networks. Environ Model Softw 103:1–15

Guadalquivir Plan (2007) Plan especial de actuación en situaciones de alerta y eventual sequía de la cuenca hidrográfica del Guadalquivir. Technical report, Ministerio de Medio Ambiente

Jager W, Christie E, Hanea A, den Heijer C, Spencer T (2018) A Bayesian network approach for coastal risk analysis and decision making. Coast Eng 134:48–61

Jensen FV, Nielsen TD (2007) Bayesian networks and decision graphs. Springer, Berlin

Keshtkar AR, Slajegheh A, Sadoddin A, Allan MG (2013) Application of Bayesian networks for sustainability assessment in catchment modeling and management (case study: the Hablehrood river catchment). Ecol Model 268:48–54

Kim K, Lee S, Jin Y (2018) Forecasting quarterly inflow to reservoirs combining a copula-based Bayesian network method with drought forecasting. Water 10:233

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge

Korb KB, Nicholson AE (2011) Bayesian artificial intelligence. CRC Press, Boca Raton

Koski T, Noble J (2011) Bayesian networks: an introduction. Wiley, New York

Landuyt D, Broekx S, D'hondt R, Engelen G, Aertsens J, Geothals P (2013) A review of Bayesian belief networks in ecosystem service modelling. Environ Model Softw. https://doi.org/10.1016/j.envsoft.2013.03.011

Langseth H, Bangsø O (2001) Parameter learning in object-oriented Bayesian networks. Ann Math Artif Intell 32(1):221–243

Langseth H, Nielsen TD, Rumí R, Salmerón A (2009) Inference in hybrid Bayesian networks. Reliab Eng Syst Saf 94:1499–1509

Lauritzen SL (1992) Propagation of probabilities, means and variances in mixed graphical association models. J Am Stat Assoc 87:1098–1108

Liu Q, Peres F, Tchangani T (2016a) Object-oriented Bayesian network for complex system risk assessment. IFAC 49:31–36

Liu Q, Tchangani A, Pérès F (2016b) Modelling complex large scale systems using object oriented Bayesian networks (OOBN). IFAC-PapersOnLine 49(12):127–132

Maldonado A, Aguilera P, Salmerón A (2016) Continuous Bayesian networks for probabilistic environmental risk mapping. Stoch Environ Res Risk Assess 30(5):1441–1455. https://doi.org/10.1007/s00477-015-1133-2

Malekmohammadi B, Moghadam N (2018) Application of Bayesian networks in a hierarchical structure for environmental risk assessment: a case study of the Gabric Dam, Iran. Environ Monit Assess 190:1–17

Marcot BG, Penman T (2019) Advances in Bayesian network modelling: integration of modelling technologies. Environ Model Softw 111:386–393

Molina JL, Pulido-Veláquez D, García-Aróstegui J, Pulido-Velázquez M (2013) Dynamic Bayesian network as a decision support tool for assessing climate change impacts on highly stressed groundwater systems. J Hydrol 479:113–129

Moral S, Rumí R, Salmerón A (2001) Mixtures of truncated exponentials in hybrid Bayesian networks. In: ECSQARU'01. Lecture notes in artificial intelligence, vol 2143. Springer, Berlin, pp 156–167

Mortera J, Vicard P, Vergari C (2013) Object-oriented Bayesian networks for a decision support system for antitrust enforcement. Ann Appl Stat 7:714–738

Murphy KP (2002) Dynamic Bayesian networks: representation, inference and learning. Ph.D. thesis, University of California, Berkeley

Nicholson A, Flores J (2011) Combining state and transition models with dynamic Bayesian networks. Ecol Model 222:555–566

Papacharalampous G, Tyralis H, Koutsoyiannis D (2019) Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. Stoch Environ Res Risk Assess 33(2):481–514

Paprotny D, Morales-Napoles O (2017) Estimating extreme river discharges in Europe through a Bayesian networks. Hydrol Earth Syst Sci 21:2615–2636

Pearl J (1988) Probabilistic reasoning in intelligent systems: network of plausible inference. Morgan Kaufmann, San Mateo

Pérez-Ramiréz PA, Bouwer-Utne I (2015) Use of dynamic Bayesian networks for life extension assessment of ageing systems. Reliab Eng Syst Saf 133:119–136

Provan GM (1993) Tradeoffs in constructing and evaluating temporal influence diagrams. In: Proceedings of the 9th conference of the uncertainty in artificial intelligence, pp 40–47

Ropero RF (2016) Hybrid Bayesian networks: a statistical tool in ecology and environmental sciences. Ph.D. thesis, Department of Biology and Geology, University of Almería

Ropero RF, Nicholson A, Rumí R, Aguilera P (2018) Learning and inference methodologies for hybrid dynamic Bayesian networks: a case study for a water reservoir system in Andalusia, Spain. Stoch Environ Res Risk Assess 32(11):3117–3135. https://doi.org/10.1007/s00477-018-1566-5

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

Rumí R, Salmerón A (2007) Approximate probability propagation with mixtures of truncated exponentials. Int J Approx Reason 45:191–210

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

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B (Methodol) 36(2):111–147

Voinov A, Bousquet F (2010) Modelling with stakeholders. Environ Model Softw 24:1268–1281

Wang X, Zhu J, Ma F, Li C, Cai Y, Yang Z (2016) Bayesian network-based risk assessment for hazmat transportation on the Middle Route of the South-to-North Water Transfer Project in China. Stoch Environ Res Risk Assess 30:841–857

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Mateo

Yu J, Xu L, Xie X, Hou D, Huang P, Zhang G, Zhang H (2017) Contamination event detection method using multi-stations temporal–spatial information based on Bayesian network in water distribution systems. Water 9:894

Yung EC, Wilkinson L, Nicholson A, Quintana-Ascencio P, Fauth J, Hall D, Ponzio K, Rumpff L (2016) Modelling spatial and temporal changes with GIS and spatial and dynamic Bayesian networks. Environ Model Softw 82:108–120

Zhu X, Zhang G, Yuan K, Ling H, Xu H (2018) Evaluation of agricultural water pricing in an irrigation district based on a Bayesian network. Water 10:768

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