Noname manuscript No.
(will be inserted by the editor)

Learning and Inference methodologies for Hybrid Dynamic Bayesian networks. A case study for a water reservoir system in Andalusia, Spain.

Rosa F. Ropero · Ann E. Nicholson · Pedro A. Aguilera · Rafael Rumí

Received: date / Accepted: date

Abstract Time series analysis requires powerful and robust tools; at the same time the tools must be intuitive for users. Bayesian networks have been widely applied in static problem modelling, but, in some knowledge areas, Dynamic Bayesian networks are hardly known. Such is the case in the environmental sciences, where the application of static Bayesian networks in water resources research is notable, while fewer than five papers have been found in the literature for the dynamic extension. The aim of this paper is to show how Dynamic Bayesian networks can be applied in environmental sciences by means of a case study in water reservoir system management. Two approaches are applied and compared for model learning, and another two for inference. Despite slight differences in terms of model complexity and computational time, both approaches for model learning provide similar results. In the case of in-

Rosa F. Ropero
Informatics and Environmental Laboratory, Dpt. of Biology and Geology,
University of Almería
orcid.org/0000-0003-1756-012X
E-mail: rosa.ropero@ual.es

Ann E. Nicholson
Faculty of Information Technology
Monash University, Melbourne, Australia
orcid.org/0000-0002-2269-9823
E-mail: ann.nicholson@monash.edu

Pedro A. Aguilera
Informatics and Environmental Laboratory, Dpt. of Biology and Geology,
University of Almería
orcid.org/0000-0001-8086-4738
E-mail: aguilera@ual.es

Rafael Rumí
Dpt. of Mathematics,
University of Almería
orcid.org/0000-0001-9189-5468
E-mail: rrumi@ual.es

ference methods, again, there were slight differences in computational time, but the selection of one approach over the other is based on the prediction needed: If the aim is just to go one step forward, both *Window* and *Roll out* approaches are similar, when we need to go more than one step forward; the most appropriate will be *Roll out*.

Keywords Water reservoir · Dynamic Bayesian networks · 1-step approach · 2-step approach · Window approach · Roll out approach

1 Introduction

Nowadays, it is widely recognised that including time as a component of models is an important challenge in the field of data mining, reasoning and decision-support systems (Russel and Norvig 2002; Mihajlovic and Petkovic 2001). In environmental sciences, time series analysis has a wide range of applications, and some models have been successfully applied, such as auto-regresive models (Davidson et al 2016; Parmar and Bhardwaj 2015), hidden Markov models (Lagona et al 2015; Spezia et al 2010), order series method (Arya and Zhang 2015), multi-temporal analysis (Lobo et al 2015), autocorrelation functions (Farah et al 2014), functional depth for outliers (Raña et al 2014), and state space models (Bojarova and Sundberg 2010). However, several temporal models are based on complex mathematical notation that experts from other areas are unfamiliar with, or else they act as black boxes. These featues mean such models are difficult for experts and stakeholders to understand and, in addition, specific literature is usually difficult to find (von Asmuth et al 2012).

Bayesian networks (BNs) are statistical tools whose ability to solve a wide range of tasks, including classification, regression, and performing scenarios of (future) change (Aguilera et al 2011) has been demonstrated. From their definition at the beginning of the 1990s (Jensen and Andersen 1990), they have been developed and applied in various fields, and a consolidated literature is easily found. One of their advantages is that non-specialists and stakeholders can get an intuitive understanding of the model since BNs are configured as a model with a qualitative part, namely a direct acyclic graph representing both variables and the relationships between them; and a quantitative part where the strength of these relationships is represented by probabilities. Moreover, the ability of BNs to deal with large datasets and missing data and information from different sources has also been demonstrated (Fernandes et al 2013). Finally, despite being defined for discrete variables, there are some BN solutions proposed for dealing with continuous or hybrid data (continuous and discrete variables simultaneously). The earliest and most common solution was to discretise the continuous variables and treat them as discrete using the available software. The main disadvantage of this approximation is the loss of information and accuracy (Uusitalo 2007). The next solution proposed was the *Conditional Gaussian* model, but this imposes certain restrictions during the structural learning. Other models, such as the *Mixture of Truncated Exponential* (MTE) (Moral et al 2001), *Mixture of Polynomials* (Shenoy and West

2011), and *Mixture of Truncated Basic Functions* (Langseth et al 2012) were proposed to overcome such structural limitations, but only the MTE model has been applied to environmental problems (Maldonado et al 2016; Ropero et al 2016).

In the environmental field, BNs have been widely demonstrated to be a powerful tool to solve problems under a framework of uncertainty (Kelly et al 2013). In the context of Integrated Water Resource Management, BNs were proposed as an appropriate tool for modeling uncertainty in water research (Castelletti and Soncini-Sessa 2007a,b; Henriksen et al 2007), leading to the application of BNs in some European projects such as the FP5-MERIT (Bromley et al 2005) or NeWater (Henriksen and Barlebo 2008). This has meant that water resource management has become one of the fields in which BNs are most commonly applied (Fienen et al 2016; Phan et al 2016), for both groundwater (Aguilera et al 2013) and surface water (Liu et al 2016) systems.

However, BNs have been mainly used for static problems, where time is not included as a component of the model. Nevertheless, several studies have used BNs to predict change in systems modeled under different scenarios (Keshtkar et al 2013). For example, they have been used for modeling scenarios of Climatic and Global change in different ecosystems and catchments (Dyer et al 2014; Mantyka-Pringle et al 2014; Webster and McLaughlin 2014), changes in management plans for groundwater systems and for species conservation (Shenton et al 2014; Tiller et al 2013), risk assessment in groundwater quality assessment (Aguilera et al 2013) and environmental impacts on species distribution patterns (Meineri et al 2015). Although results from static BNs are robust and appropriate, the conclusions cannot be extrapolated to a particular time, nor time series can be handled (Ropero et al 2017).

The extension of BNs, the so-called Dynamic Bayesian networks (DBNs), has begun to be applied to face this new challenge of including time in environmental models (Hill 2013; Molina et al 2013). Even though they are still being developed, some initial applications of DBNs in the environmental sciences are cited in the literature. In the works of Hill et al (2009) and Hill (2013), hybrid DBNs were applied to the control of streaming climatic data, in an attempt to detect anomalies and errors in the data. Zhang et al (2012) used discrete DBNs to integrate data from different times series into a model to accurately estimate the Leaf Area Index in a region of China. In both cases, the DBN application focused on the pre-processing step, trying to correctly collect the data, or merge different data sets. Molina et al (2013) showed how discrete DBNs were learnt as a Decision Support System to predict the effects of Climate Change scenarios in a groundwater system in Spain over the 2070-2100 horizon. Following the idea of applying DBNs in water resources, Molina et al (2016) aimed to model the temporal behavior of hydrological time series for two different river basins, looking for a proper order in the series. Trifonova et al (2015) presented an application in the field of fisheries ecology, in which a DBN based on Hidden Markov models including latent variables was developed for modeling species dynamics over time and space. Papakosta and Straub (2016) developed a model for predicting the occurrence of wildfires

in a Mediterranean region including environmental and anthropogenic information using a 1x1km grid. Finally, Ropero et al (2017) compared static and dynamic BNs in a regression problem in order to predict the behavior of a water reservoir under changing climatic conditions.

Climate in Spain is characterized by irregular rainfall patterns which provokes periods of severe drought followed by heavy storms, making water scarcity management a challenge. For that reason, dam and reservoir construction has historically been the main solution to water scarcity and irregularity with more than 1200 active reservoirs in Spain. Apart from acting as guarantor for water and agriculture consumption, the current water reservoir system has been designed to control flooding. Thus, accurate prediction of the behavior of reservoirs under unstable climatic conditions is crucial to identify the flood risk and mitigate losses caused by flooding. In the same way, drought periods need to be detected in time in order to be prepared, not only to meet the water demand for human consumption, but also to ensure natural watercourses carry a minimum water flow that allows biodiversity to be safeguarded. Furthermore, the temporal component is even more important since experts need to know when such extreme events will happen.

This paper aims to make Dynamic Bayesian networks more accessible to environmental science experts by comparing different approaches available for learning and performing inference using data from a water resource management case study. However, the objective was not to build a hydrological model, but to extract relevant information from those available in order to reach the methodological objective. Thus, a well known and easy to interpret case study was used so that, no extra uncertainty was added to the model. As a novelty, data were totally continuous, and the *Mixture of Truncated Exponential* models were selected to allow continuous variables to be included in the model.

2 Bayesian networks

2.1 Hybrid Bayesian networks

Bayesian networks (BNs) are defined as a statistical multivariate model for a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, and 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; *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 variables, giving values for the CPTs or even refining the structure previously learnt from data (Aguilera

et al 2011). This structure also means that, with no mathematical calculation involved, the variable(s) that are relevant (or not) for a certain problem can be known (Pearl 1988) so simplifying 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

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

where $pa(x_i)$ is a set of the parent of variable x_i according to the structure of the direct acyclic graph.

BNs were originally developed for discrete variables, but real life problems require both continuous and discrete (hybrid) data to be simultaneously included into modelling processes. It has encouraged the proposal of new models for dealing with hybrid data in BNs. One of these models are the *Mixture of Truncated Exponential* models (MTEs), proposed by Moral et al (2001) and developed in detail in Rumí (2003).

Defined in Moral et al (2001), MTEs models were designed as an approach to include continuous and discrete variables into BNs with no restriction on the network structure. This approximation proposed dividing the value range of a continuous variable into several intervals, and approximate each using 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, because of its high fitting power, which makes it appropriate to deal with hybrid data.

During the probability inference process, where the posterior distributions of the variables are obtained given some evidence, the intermediate functions are not necessarily density functions. Therefore, a general function called *MTE potential* needs to be defined as follows:

Definition: MTE potential *Let X be a mixed n-dimensional random vector. Let $Z = (Z_1, \ldots, Z_d)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_c)^T$ be the discrete and continuous parts of X, respectively, with $c + d = n$. We say that a function $f : \Omega_X \mapsto \mathbb{R}_[0, \infty)$ is a Mixture of Truncated Exponentials potential (MTE potential) if one of the following conditions holds:*

i. *$Z = \emptyset$ and f can be written as*

$$f(x) = f(y) = a_0 + \sum_{i=1}^{m} a_i \mathrm{e}^{\left\{b_i^T y\right\}} \tag{2}$$

for all $y \in \Omega_{\mathbf{Y}}$, where $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}^c$, $i = 1, \ldots, m$.

ii. *$Z = \emptyset$ and there is a partition $D_1, \ldots, D_k$ of $\Omega_{\mathbf{Y}}$ into hypercubes such that f is defined as*

$$f(\mathbf{x}) = f(y) = f_i(y) \quad \text{if} \quad y \in D_i,$$

where each f_i, $i = 1, \ldots, k$ can be written in the form of Equation (2).

iii. *$Z \neq \emptyset$ and for each fixed value $z \in \Omega_Z$, $f_z(y) = f(z, y)$ can be defined as in* ii.

An MTE potential f is an MTE density *if*

$$\sum_{z \in \Omega_Z} \int_{\Omega_{\mathbf{Y}}} f(z, y) dy = 1.$$

A conditional MTE density can be specified by dividing the domain of the conditioning variables and specifying an MTE density for the conditioned variable for each configuration of splits of the conditioning variables.

Consider two continuous variables Y_1 and Y_2. A possible conditional MTE density for Y_1 given Y_2 is the following:

$$f(y_1 \mid y_2) = \begin{cases} 0.28 + 0.01\mathrm{e}^{1.03y_1} + 0.02\mathrm{e}^{0.01y_1} & \text{if } 0 \leq y_1 < 1,\ 1 \leq y_2 < 3, \\ 0.02 + 0.02\mathrm{e}^{1.01y_1} + 0.12\mathrm{e}^{0.09y_1} & \text{if } 1 \leq y_1 < 3,\ 1 \leq y_2 < 3, \\ 0.49 - 0.12\mathrm{e}^{0.59y_1} - 0.24\mathrm{e}^{-0.08y_1} & \text{if } 0 \leq y_1 < 1,\ 3 \leq y_2 < 4, \\ 0.07 - 0.02\mathrm{e}^{-0.23y_1} + 0.62\mathrm{e}^{-0.23y_1} & \text{if } 1 \leq y_1 < 3,\ 3 \leq y_2 < 4. \end{cases}$$

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 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 (2007).

2.2 Dynamic Bayesian networks

The earliest attempt to deal with time using BNs appeared in Provan (1993), who proposed their use for modeling a generic system in each time step, joining single BNs with links that represent the transition from one time to the next. DBNs have since evolved and nowadays they are defined as *a long-established extension of BNs that can represent the evolution of variables over time* (Nicholson and Flores 2011).

The term dynamic means the system is changing over time, not that the network and the relationships between variables change (Murphy 2002). For simplicity, the proposal of Provan (1993) is followed. Thus, it is assumed that a DBN is a time-invariant model composed by a sequence of identical BNs representing the system at 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). Thus, according to Korb and Nicholson (2011) the components of a DBN are:

– **Time slice**: the state of the system at a particular time t, represented by a static BN identical in each time step.

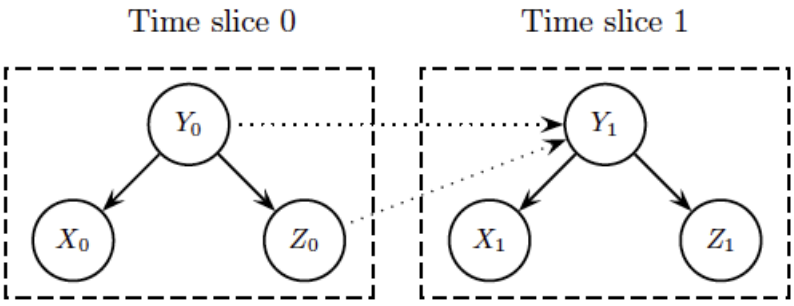


Fig. 1 Example of a Dynamic Bayesian network following the *first-order Markov assumption* comprising three variables X, Z and Y composed of 2 time slices. Solid links represent intra-slice arcs, whilst dotted lines represent inter-slice arcs.

- **Intra-slice arcs**: the relationships between variables in a time-slice (*e.g.* in Figure 1 links between X_0 and Y_0). They remain constant regardless of the particular time.
- **Inter-slice arcs**: also called temporal arcs, they represent the relationships between variables in successive, or not successive, time slices both *(i)* the same variable over time (*e.g.* in Figure 1 links between Y_0 and Y_1) or *(ii)* between different variables over time (*e.g.* in Figure 1 links between Z_0 and Y_1).

These dynamic models could be quite complex due to the existence of sub-models connected through temporal links for each time step. So, 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 assumes 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 both (Figure 1). Only if it is necessary, the DBN can be rolled out and more than two time slices would be represented.

Nowadays two main approaches to learn a DBN are considered: in one (1-step) or two steps (2-step) (Black et al 2014). Following the so-called 1-step approach, the dynamic model, both the structure of the time slice and the inter-slice links, are learnt from the data following the time-invariant property, using specific software, such as Causal Discovery via Minimum Message Length (Korb and Nicholson 2011; O´ Donnell 2000). By contrast, the 2-step approach means that firstly, the structure of a static model (time slice) is learnt using all the information available, and, in the second step, this structure is repeated and connected through (temporal) links. Parameters can be obtained from the data, or elicited from expert knowledge.

Once the models were learnt and validated new information, or *evidence*, can be included into one (or more) variable(s) and used to update the probability distribution of the remaining variables 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 $X_i \notin \mathbf{E}$:

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

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

$$p(x_i, \mathbf{e}) = \sum_{\mathbf{x} \notin \{x_i, \mathbf{e}\}} p_e(x_1, \ldots, x_n), \tag{4}$$

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 $\mathbf{e}$.

Several algorithms have been proposed for both exact inference - *Forward-Backward algorithm* (Baum et al 1970) and *interface algorithm* (Murphy 2002) - and approximate inference - *BK algorithm* (Boyen and Koller 1998) and *FF algorithm* (Murphy and Weiss 2001), in DBN. However, there is no BN software that implements these algorithms in such a way that experts from other fields can easily apply them. For that reason, in this study we propose a framework based on the available algorithms and software to use DBN in a easy way. Since they are represented as a set of identical static BNs connected through (temporal) links, DBNs can be represented and solved as a kind of "static" model divided into different sub-models (model for time 0, model for time 1, and so on), which allows the available algorithms developed for static BNs to be used.

3 DBN for water reservoir system modeling

A case study based on water reservoir modeling in Andalusia, Spain, is presented. For modeling such a complex system, it is necessary to include multivariate approaches including spatial relationships between the reservoirs (which reservoir transfers to what other one), rates of consumption and water for ecological purposes (to safeguard biodiversity in the downstream watercourse). However, this information is often difficult to obtain, so, in this paper, only information about water balance in reservoirs and meteorological variables were used. Besides, the aim of this paper is to study the differences between both learning and inference approaches for DBNs, not the development of an exhaustive hydrological model to explain the behavior of Andalusian water reservoirs in depth.

The methodology developed is shown in Figure 2 and is explained in detail in this section. First, data were collected and organized in order to fit

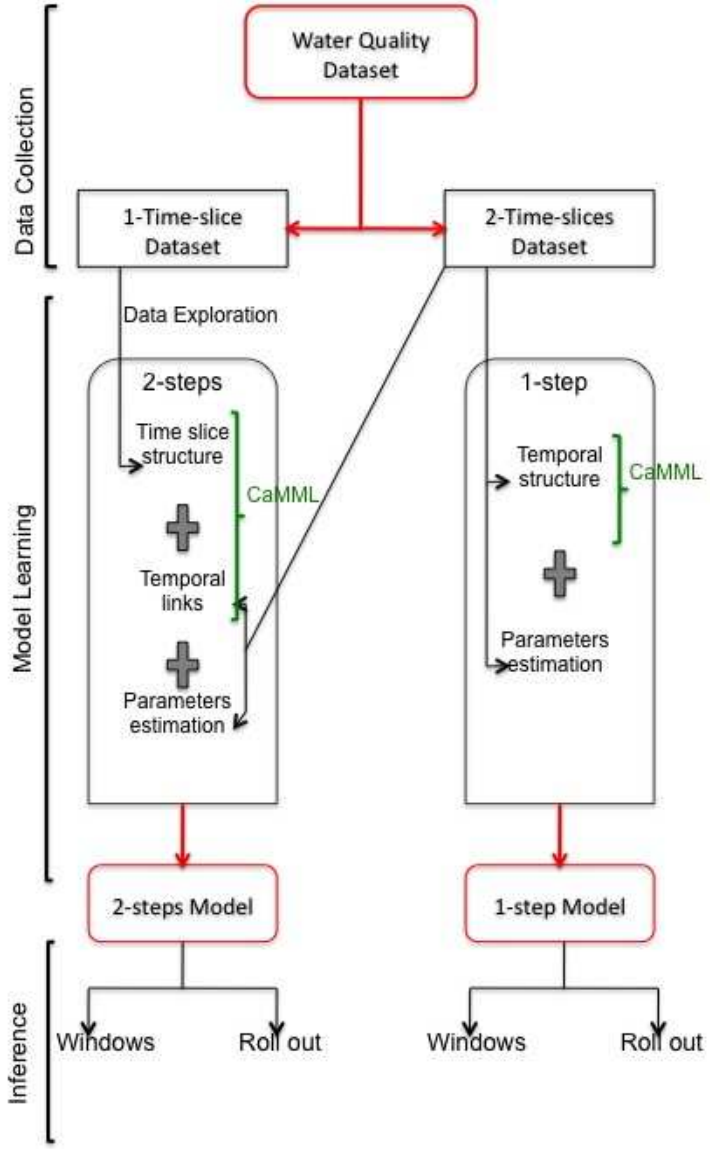


Fig. 2 Flowchart of the methodology applied.

with the software requirements (Section 3.2). Also, an initial data exploration was carried out before any modeling process, using *Omnigram Explorer* to gain some prior information (Section 3.3.1). Once data were collected and analyzed, the modeling process was divided into structure learning, parameter estimation and model validation for both *one-step* and *two-steps* approaches

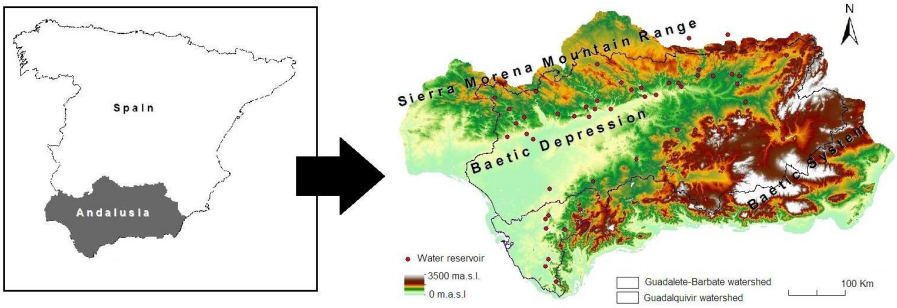


Fig. 3 Location, relief and reservoirs selected from the study area.

(Sections 3.3.2 and 3.3.3). Finally, DBN inference was done using two methods: *Windows* and *Roll-out* approaches (Section 3.4).

3.1 Study area

Andalusia (Figure 3) is located in southern Spain and configures the second largest Autonomous Region of Spain, and the most-densely populated. It covers a surface area of $87,600$ km^2 or 17.3% of the national territory [1]. Its terrain extends over a wide range of altitude, from the *Baetic Depression* to the mountain ranges of the *Sierra Morena* and *Baetic System*, which boast the highest peaks in Spain, over 3000 m.a.s.l. The landscape is quite heterogeneous, with huge differences between the densely populated and rich irrigated croplands areas of the river basin and coastlands, to the sparsely populated forests of the uplands.

Its climate is similarly heterogeneous. Even though Andalusia is included in the Mediterranean climate zone, there are stark differences between the coast and inland. The climate in the southeast is semiarid, with less than 200 mm of annual rainfall, whilst the middle and northern parts are under a continental climate influence, receiving more than 4000 mm rainfall. These patterns are not only spatially, but also temporally irregular cycles of drought and wet periods. This has led to the construction of more than 100 active reservoirs in Andalusia.

Due to Andalusia step relief and climate, it is divided into six catchments. In the current study, only the Guadalquivir and Guadalete-Barbate catchments are considered (Figure 3), encompassing a total of 61 water reservoirs.

[1] Data from the Spanish Statistical Institute

3.2 Data Collection & Organization

Data were collected from the Water Quality Dataset of the Andalusian Regional Environmental Information Network[2] (Andalusian Regional Government) for the 61 reservoirs selected. They consist of seven continuous variables and one discrete, collected monthly from October 1999 to September 2008 (Table 1). *Temperature* in °C (T) and *Rainfall* in m^3/m^2 (R) represent the climatic conditions in the vicinity of the reservoir. *Percentage Evaporation* (E) is the percentage of the water in the reservoir that evaporates. *Water level* (WL) indicates the height of the water column in m.a.s.l., whilst *Percent Fullness* (PF) expresses the percentage of the reservoir capacity that is currently used, from 0 to more than 100% (following a storm event, the reservoir can exceed the dam capacity). Finally, reservoir management is represented by Amount Discharge and Amount Transfer in. *Amount Discharge* (AD measured in m^3) refers to the amount of water that is released to meet ecological, water consumption or regulation purposes. By contrast, *Amount Transfer in* (AT) is the amount of water deliberately added to the reservoir, e.g., pumped in from another reservoir. *Reservoir Use* is the discrete variable that represents the main use of the water reservoir (Hydroelectric, General regulation, Irrigation, Human consumption, Industry, No information, Ecological, Irrigation and other, Irrigation and consumption, Consumption and others).

Table 1 Main statistics of the continuous variables collected.

Variable	Minimum	Mean	Maximum
Rainfall	0.00	0.04	0.43
Temperature	0.0	17.8	30.0
Percentage Evaporation	0.00	0.86	23.40
Amount Discharge	0.00	10.46	1529.39
Amount Transfer in	0.00	10.81	1529.71
Water Level	0.00	355.8	1039.3
Percentage Fullness	0.00	57.20	237.78

With the aim of fitting with the software and modeling requirements, data collected were organized in two different datasets (Figure 2):

- Static dataset. Once the data are collected, values of variables at different times are put together to create a new unique variable, in which time is excluded (*e.g.* in Figure 4(a), the variable *Temperature* is configured by taking the temperature data for October 1999, November 1999 and so on). This static dataset has seven variables and 6588 observations; it was used for the initial data exploration, and the time slice structure learning in the 2-step approach.
- Dynamic dataset. For each reservoir, data are organized into two-time slices, comprising every consecutive pair of months (Figure 4(b)). This

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

Dam	T	R	...
1	$T_{oct1999}$	$R_{oct1999}$	...
2	$T_{oct1999}$	$R_{oct1999}$	...
...	$T_{oct1999}$	$R_{oct1999}$	...
1	$T_{nov1999}$	$R_{nov1999}$	...
2	$T_{nov1999}$	$R_{nov1999}$	...
...	$T_{nov1999}$	$R_{nov1999}$	...
1	$T_{dec1999}$	$R_{dec1999}$	...
2	$T_{dec1999}$	$R_{dec1999}$	...
...	$T_{dec1999}$	$R_{dec1999}$	...

(a) Static dataset

Dam	T_0	R_0	...	T_1	R_1	...
1	$T_{oct1999}$	$R_{oct1999}$	...	$T_{nov1999}$	$R_{nov1999}$	...
1	$T_{nov1999}$	$R_{nov1999}$	...	$T_{dec1999}$	$R_{dec1999}$	...
...	...	...	...	...	...	...
2	$T_{oct1999}$	$R_{oct1999}$	...	$T_{nov1999}$	$R_{nov1999}$	...
2	$T_{nov1999}$	$R_{nov1999}$	...	$T_{dec1999}$	$R_{dec1999}$	...
...	...	...	...	...	...	...

(b) Dynamic dataset

Fig. 4 Example of both static and dynamic datasets for the *Temperature* (T) and *Rainfall* (R) variables.

temporal dataset has fourteen variables (Temperature at time 0, Temperature at time 1, Rainfall at time 0, Rainfall at time 1, and so on) and 6527 observations [3]. This dataset was used for the dynamic structure learning with CaMML during the 1-step approach, and later on for both 1-step and 2-step DBN models parameters estimation using Elvira software. Following this organization, this temporal dataset was also divided into one for learning and model validation, (October 1999 to September 2007) and another for inference (October 2007 to September 2008).

3.3 DBN model learning

3.3.1 Data exploration

According to Aguilera et al (2011), the first step in (D)BN modeling is to state the objective of the model. Here, the focus is on modeling the behavior and evolution of water storage in the reservoir system, represented by the variable *Percentage Fullness*, but also, the relationships with the other variables. No *classification* or *regression* problem (in which the aim is to accurately predict the behavior of one unique discrete or continuous variable, respectively) is faced, but *characterization* (in which the main idea is to learn the structure of the model and study not only the relationships but also their strength). In this case, the model can be obtained directly from the data, through an expert elicitation process, or using both expert and data information. In environmental sciences *characterization* has been widely developed through expert

[3] Note that the difference in the sample size in both dataset is due to the different organization of the data.

elicitation with discrete variables (Aguilera et al 2011). In this paper, continuous information is available and the methodology applied allows us not to discretize them. Thus, the model could be directly obtained from the data, but obtaining first some initial knowledge about the variables and their potential relationships can be extremely useful. This data visualization can assist in understanding the relationships between the variables included, and also greatly simplify communications with stakeholders.

Thus, data exploration was performed using the *Omnigram Explorer* (OE) software which allows an initial understanding of how the variables are related, as well as some idea about the system's causal structure (Ropero et al 2015; Taylor et al 2015). This software was designed as a tool for interactive exploration of relationships between variables in an agent-based simulation. It draws on ideas for visualization in the *Attribute Explorer* (Spence and Tweedie 1998), where data is presented as a set of histograms, one per variable [4].

Initially, a data file and model definition are loaded in OE software. The data file contains the joint data sample, in which each variable is represented by a histogram, showing its sample distribution (Figure 5 a)). If a bin is empty (e.g., bin 0 in Rainfall node in Figure 5 a)), a thin horizontal line is drawn at the base. Besides, model definition is included to designate some variables as input or outputs and to use Bayesian network links to represent causal structure or other dependencies. However, these features are for display, while the links included have no meaning. Thus, in this paper, neither links nor model definition was included.

The potential of this tool lies in its interaction modes, where a variable or subset of variables can be selected and their relationship with the remaining variables explored. The selected variables are the focus of attention, which is indicated visually by a red square indicator in the corner of the node (Rainfall variable in Figure 5 a)). When the focus variable changes, all of the other variables are updated to show the corresponding sample values in their distributions. There are four modes of interaction: single node brushing (Figure 5 b), only one variable is the focus), multi node brushing (Figure 5 c), two variables can be the focus), omnibrushing (Figure 5 d), again only one variable can be the focus, but the remaining variables are updated to show each bin what fraction of the data correspond to the focal bins), and sample view (Figure 5 e), with just one focus variable, the difference is the way data are display, in this case, each individual sample is represented as a small colored circle, simultaneously across all variables).

Figure 6 shows the omnibrushing interaction mode for both *Rainfall* and *Temperature* variables. Lower values of *Rainfall* (marked in yellow) are associated with higher *Temperature* values and lower values of *Percentage Fullness*. However, the highest values of *Rainfall* (marked in blue) are not particularly correlated with higher values of *Percentage Fullness*. Besides, there is a negative relationship between *Rainfall* and *Temperature*, while a positive relation-

[4] For more detail information about this software see the link: http://www.tim-taylor.com/omnigram/.

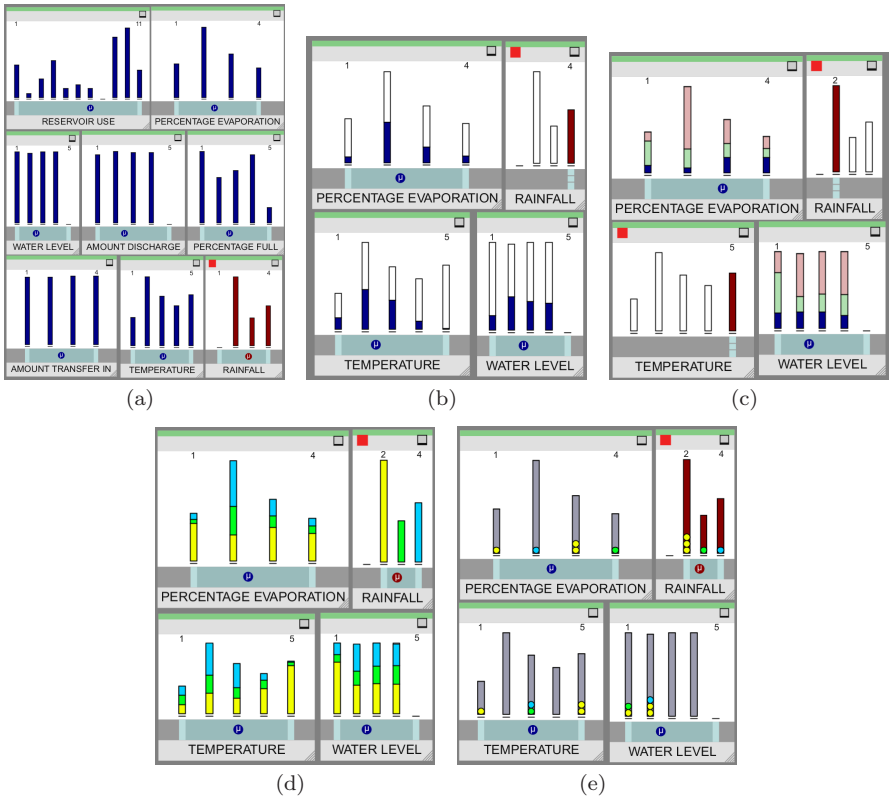
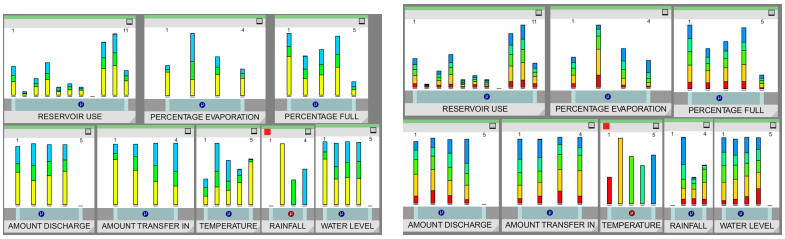


Fig. 5 Screenshot from *Omnigram Explorer* software. Initial histograms for the water reservoir data focussing on the *Rainfall* variable (a) and modes of interaction in OE for a subset of variables: Single node (b), Multi node (c), Omnibrushing (d) and Sample view (e).

ship exists between *Percentage Fullness*, *Water Level* and *Amount Transfer in* (Figure 7). However, the relationships with *Percentage Evaporation* are more ambiguous. When *Rainfall* values are higher, *Percentage Evaporation* tend to be more prevalent in the second bin.

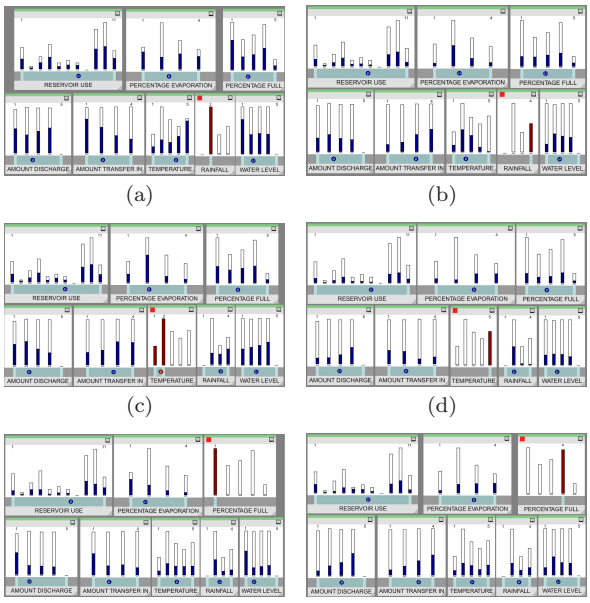
In the case of *Temperature*, moderate values (marked in yellow and green colors) are more prevalent in the rest of the variables than both extremes (bins 1 in red, and 5 in blue colors). If we used single node brushing (Figure 7), when we focus on bins 1 and 2 (corresponding to temperatures lower than 15°C), samples are fairly flat, except for lower *Percentage Evaporation* and slightly higher values of *Rainfall*. If we move now to the highest bin (temperatures above 25°C), it shows low *Rainfall* and higher *Amount Discharge*, presumably to combat drought conditions.

Finally, both *Amount Transfer in* and *Amount Discharge* behave in the same way with respect to *Percentage Fullness* (Figure 7 (e) and (f)) and that the relationship between all three is positive. We computed the Pearson correlation between *Amount Transfer in* and *Amount Discharge* conditioned on



(a) (b)

Fig. 6 Omnibrushing for Rainfall and Temperature variables.



(a) (b) (c) (d) (e) (f)

Fig. 7 Single Node Brushing for *Rainfall*, *Temperature* and *Percentage Fullness*, focus on the lowest and highest values.

water reservoir, which was a very high 0.95. However, they behave in opposite ways in high temperature conditions, so they both will be included into the model.

Through this data exploration, the following information was obtained from the data:

- *Rainfall* and *Temperature* are clearly inversely related.
- whilst *Rainfall*, *Percentage Fullness* and *Water Level* are positively related.
- *Percentage Evaporation* is also related to both *Rainfall* and *Temperature*, but the relationship seems to be more complex.
- *Reservoir use* provides no significant information, so it will not be included into the model. This way, only seven continuous variables were used.

Thus, these relationships need to be included in the model. In both cases, *Rainfall* and *Temperature* act as a posible cause of *Percentage Fullness*, *Percentage Evaporation* and *Water Level*, so they should appear in the network as parents of these. Also, given a fixed *Percentage Fullness*, the variables of *Amount Discharge* and *Amount Transfer in* provide similar information and should be considered to the model to be closely related.

3.3.2 Model learning

Once data were analyzed and prior information gained, the next step consists of model learning. Both One-step and Two-steps learning approaches were used for model learning. In order to learn in an intuitive way, DBNs following One-step approach, CaMML software is used for structure learning. This software is a machine learning program able to learn causal structure by using a Bayesian metric (Minimum Message Length score) and stochastic search to find the model, or set of models, with the highest posterior probability given the data (for more information see Korb and Nicholson (2011) [5]. Besides, it was updated to incorporate dynamic models learning. However, this model provides us with the causal structure of the model not its parameters, thus, after CaMML, Elvira software was used for parameter estimation.

In the case of one-step approach, as it was previously said, this learning approach means both the time slice structure and temporal links are learnt simultaneously. So, the dynamic dataset was included into the software and carried out the learning process. CaMML runs a search of the optimal dynamic structure of the model according to the data provided using the MML score. By this way, One-step model was obtained.

In contrast, in two-steps first the static structure is obtained and repeated for time 0 and time 1 and, in a second step, temporal links are added. In order to compare the two learning process, CaMML software is applied. This software supports prior information about the structure of the model but just for static model learning. This prior information comprises what variables should be linked (priors), or the partial (or total) order of variables (tiers). The idea of using priors is to assist the discovery process with common sense background knowledge or expert opinion, or, as in this case, with the information that data exploration provides. Inspired by *OE* data exploration, the following tiers and priors were included:

- *Priors:* The following links should exist: from *Rainfall* to *Percentage Fullness*, from *Percentage Evaporation* to *Percentage Fullness*, and from *Water Level* to *Percentage Fullness*.
- *Tiers:* Variables in the model should follow this structure: in a first level *Rainfall* and *Temperature* as parent of *Percentage Evaporation*, *Amount Discharge* and *Amount Transfer in* which are positioned in a second level; and, finally, *Percentage Fullness* and *Water Level*.

[5] For more information visit http://bayesian-intelligence.com/software/

Using the static dataset and this information, a static BN was learnt. The next step consisted of learning the temporal links between each time slice by repeating the static structure and including them in the CaMML, to obtain the temporal links using the dynamic dataset (Figure 2).

Both one-step and two-steps causal structures were learnt, and now parameters of the relationships between variables (both intra and inter-slices) were estimated from the data. As it was mentioned in Section 2.2, for making DBNs learning more feasible for experts in environmental sciences, these models are considered as a kind of complex static model divided into two parts: one for time 0, and the other for time 1. Thus, available algorithms for parameter estimation can be applied. In this case, Elvira software was used with the data from October 1999 to September 2005.

We have followed the approach of Morales et al (2007) to estimate the corresponding conditional distributions based on MTE models. Let X_i and Y be two random variables, and consider the conditional density $f(x_i \mid y)$. The idea is to split the domain of Y by using the equal frequency method with three intervals. Then, the domain of X_i is also split using the properties of the exponential function, which is concave, and increases over its whole domain (see Rumí et al (2006)). Accordingly, the partition consists of a series of intervals whose limits correspond to the points where the empirical density changes between concavity and convexity or decrease and increase. In the case of models with more than one conditioning variable, see Moral et al (2003) for more details.

At this point, a 5-parameter MTE is fitted for each split of the support of X, which means that in each split there will be five parameters to be estimated from data:

$$f(x) = a_0 + a_1 e^{a_2 x} + a_3 e^{a_4 x}, \quad \alpha < x < \beta \tag{5}$$

where α and β define the interval in which the density is estimated.

The reason to use the 5-parameter MTE lies in its ability to fit the most common distributions accurately, while the model complexity and the number of parameters to estimate is low (Cobb et al 2006). The estimation procedure is based on least squares (Romero et al 2006; Rumí et al 2006).

3.3.3 Model validation

According to Aguilera et al (2011), those BNs models for *characterization* purpose are often validated through experts. In our case, the aim is not to develop a hydrological model, so that the validation is more focused on the capability for making appropriate predictions of the reservoir behavior. Thus, for predicting the behavior of one variable, the relationships included into the model network need to be as precise as possible, if not, this variable will not be accurately predicted. In this way; validation was carried out to accurately predict the behavior of the *Percentage Fullness* variable, which is continuous. For that, data from October 2005 to September 2007 was used as evidence to

the model, including values for all variables except *Percentage Fullness* into the model and checking the prediction made for this variable. The accuracy of this prediction was measured using the root mean squared error (rmse) (Witten and Frank 2005) between the actual values of the response variable (*Percentage Fullness*), $y_1, \ldots, y_n$, and those predicted by the model, $\hat{y}_1, \ldots, \hat{y}_n$, following the equation:

$$rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \quad (6)$$

3.4 Inference in DBN

Once models were learnt and validated, a scenario of changes was proposed. The idea is to check if the results obtained from the model are coherent with the reality and if they are easily interpretable by experts. Thus, we predict the future behavior of *Percentage Fullness* when new information comes. Data from October 2007 to September 2008. In each time step, information about *Rainfall*, *Temperature*, *Amount Discharge* and *Amount Transfer in* were included as *evidences* and the Penniless algorithm (Cano et al 2002, 2000) was applied. The value of those variables was fixed and this information was propagated through the network to update the probability distribution of the rest of the variables. From this probability distribution, an estimated value can be obtained like the mean, median, or even mode. Finally, here we used the mean as the estimated value, and the *rmse* of *Percentage Fullness* variable at each time step and their evolution over time was studied.

Two main approaches for inference in DBNs were proposed, the so called *Window* and *Roll out*. The idea behind the *Window* methodology is to keep the DBN as simple as possible by maintaining only two time-slices, following the *first order Markov process*. Figure 8 a) shows the *Window* approach following an example: a DBN with five variables in each time step, and one temporal link. We have information about variables X_1 and X_2, and want to check the temporal behavior of variable X_4.

- Firstly, *evidences* are included into the model (variables X_1 and X_2 at time 0, marked in red color) and propagated. Mean values for the variable X_4 in the next time step are obtained (marked in green color).
- If only two time steps are required, the process is stopped. If not, we need to "move the window" so that, we can see time 1 and 2. In this step, evidences are obtained as the mean value of X_4 variable (values of variables in time 1, marked in blue color), and propagated to the next time step (time 2). There will be as many windows as time steps we need to represent.

Figure 8 b) shows the *Roll out* approach. In this case, the network is replicated for the total number of time slices we need. In this way, new evidences are included simultaneously in all variables rather than in consecutive steps

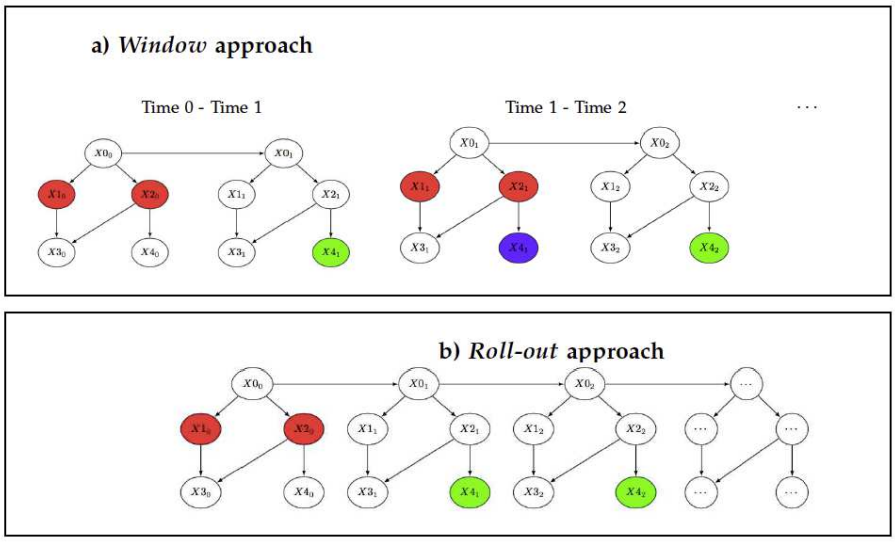


Fig. 8 Outline of the inference in DBN. Red nodes indicate evidenced variables; green nodes indicate the goal variable; blue node indicates a evidenced node obtained from the previous time slice.

as in the previous approach. In our case, the behavior of *Percentage Fullness* variable that we wish to study between October 2007 and September 2008, means the network is *rolled out* to show eleven time slices, whilst the *Window* approach was repeated 10 times.

4 Results

4.1 Comparison between DBN learning processes

Figures 9 and 10 show the 1-step and 2-step DBN models, respectively. Table 2 compares the two models in terms of error rate and structural complexity. Validation was carried out with data not previously used in learning process, and the *rmse* was calculated for *Percentage Fullness* variable at time 0 and time 1 in order to compare them, not for providing a goodness of fit measure. In contrast, complexity was measure, using both the number of links between variables and the time spend in the learning process, the so called computational time [6].

The two approaches provide similar model structures, though in the 1-step process CaMML does not allow prior knowledge to be included. This means that these relationships are evident in the information provided by our data. Besides, both models (1-step and 2-steps) simultaneously validate each other. Specifically, the structure shows a common pattern with a sequence of levels.

[6] Using a MacBook Air, 1.6 GHz Intel Core i5. RAM 4GB 1600 MHz DDR3

Climate configures the first level where the three variables, *Rainfall, Temperature* and *Evaporation*, are related to each other, but differences between the two approaches are clearly visible. In the 2-step DBN model, due to the expert knowledge, *Rainfall* and *Temperature* are both parents of *Evaporation*, whilst in the 1-step DBN model it is just the contrary. The priors and tiers included in CaMML have forced the order of these climatic variables, but it does not mean that having no priors implies a poorer model. In 1-step DBN model these variables are related according only to the data information, without any expert knowledge.

Following through the network, both variables of reservoir management are found to be linked to each other (*Amount Transfer in* and *Amount Discharge*) in a similar way. This is an example of how both models validate each other. Despite their differences in the learning approach, these two variables are linked in the same way, even in the temporal structure. It means the relationship is clearly expressed in the data, and both learning process are able to properly represent this information.

Finally, the bottom of the network is represented by *Water Level* and *Percentage Fullness*. Relationships between these last two variables and the rest are different in the two approaches. The 1-step model is simpler and these two variables are related to *Evaporation* and *Rainfall*, but not to each other. In contrast, the 2-step model includes more links between these variables and both climate (*Evaporation*) and reservoir management (*Amount Discharge*) variables, and between them (from *Water Level* to *Percentage Fullness*).

Note that, in the 2-step model, no direct link from *Rainfall* and *Percentage Fullness* is included (even when it was pointed to by the priors information), but there is a relationship between these two variables through the variable *Evaporation*. Priors and tiers included into CaMML forced the establishment of certain relationships with a confidence interval. In our case this interval was set at 0.9. However, if the data does not support this relationship, it will not be included, and an alternative is proposed. In this case, instead of a direct relationship from *Rainfall* to *Percentage Fullness* (that is also not included in 1-step as an intra-slice arc) the alternative is an indirect relationship from *Rainfall* to *Evaporation* and to *Percentage Fullness*.

In terms of model complexity, measured by the number of links, the 1-step approach provides a model with fewer arcs between variables. Instead of including a high number of intra-slice arcs, learning the dynamic structure directly determines the best structure for the temporal problem, in such a way that minimizes the number of links, and the model complexity. In contrast, learning the model in two consecutive steps means more links are included (Table 2).

These differences in term of structure and complexity involve an increase in computational time for the 2-step process (Table 2). Even though the difference was less than a minute, in more complex models (with a greater number of variables and relationships), this difference could imply intractable computational times for the 2-step approach.

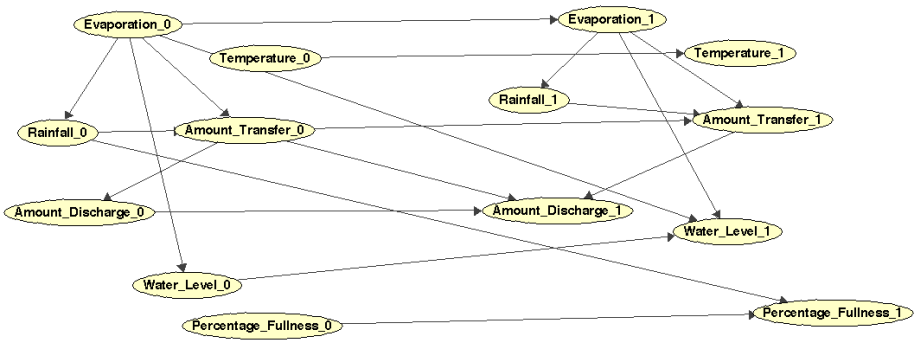


Fig. 9 Dynamic structure of 1-step model.

The two approaches provide similar values of *rmse*. Again, the 1-step approach shows less differences between the *rmse* at time 0 and 1, whilst this difference is greater in the 2-step process.

Table 2 Comparison between the two learning processes in terms of *rmse* for the *PF* variable (both in time 0 and time 1), number of intra and inter-slices links and computational time (seconds).

Model	*1-step*	*2-step*
PF_0 *rmse*	37.35	45.88
PF_1 *rmse*	41.57	39.94
Intra-slices links	5	12
Inter-slices links	10	8
Total links	15	20
Computational Time (sec)	183.3	244.1

4.2 Comparison between DBN inference methodologies

In order to study the applicability of these methodologies in a real case, new data were included into the model and the behavior of *Percentage Fullness* studied. The idea is to show how accurately the prediction could be, and also the interpretability of the results. Figure 11 shows the evolution of *rmse* for the *Percentage Fullness* variable for both *Windows* and *Roll out* approaches.

The *Windows* approach (in green) shows a more stable evolution over time. *rmse* increases until reaching a stable value that persists until the end of the process. In this way, the 1-step model provides a slightly lower error rates than the 2-step model. However, both models present a similar error rate. In contrast, *Roll out* approach (marked in blue) shows greater variability, with

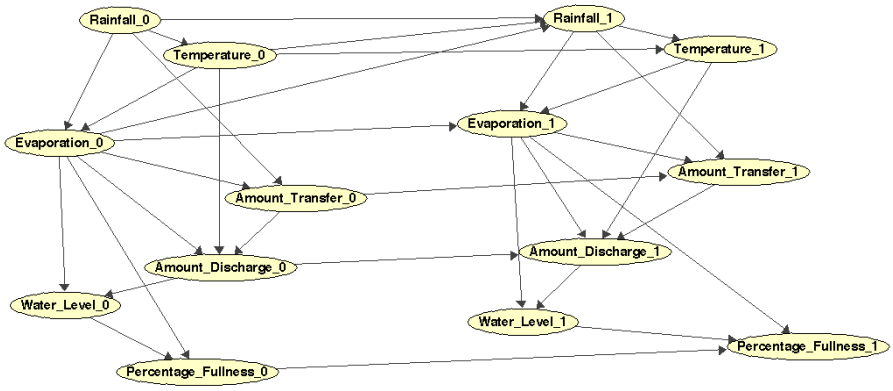


Fig. 10 Dynamic structure of 2-step model.

values depending on the time slice. The 2-step model shows greater variability between one time-slice and the next.

Table 3 compares the computational time for both inference methods, and for the learning phase, for both models. As for the learning process, there is a stark increase in computational time for the 2-step approach. In the case of the inference process, the difference is more evident, taking nearly 4 seconds more for 2-step. Again, even though these differences are small in this case, a more complex 2-step model would lead to a computational time that would make this approach unviable. Model learning is usually carried out once, so time for this task is not so important as for inference. Depending on the case study and the data, it could be necessary to perform several inference process (for example, updating the model every five minutes as new information reaches the model, and interpreting the results). In those cases, an increase in the computational time would lead to an unfeasible process since the management process requires the new information to be propagated and evaluated as soon as possible.

Table 3 Comparison between learning and inference processes in terms of computational time expressed in seconds.

Model	*1-step*	*2-step*
Learning	183.3	244.1
Window	2.5	5.9
Roll out	2.9	6.8

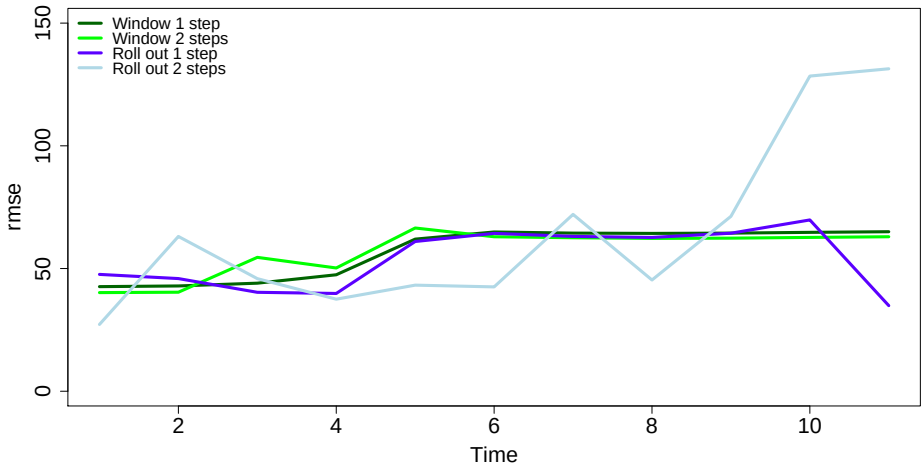


Fig. 11 *Root mean square error* values for both inference approach (*Window* and *Roll out*) in both models (1-step and 2-step).

4.3 Case study results

To showcase the results obtained from a DBN, we selected one reservoir and present the evolution of *Percentage Fullness* over four months, from December to March (Figure 12).

At the beginning of winter season, December, the reservoir is close to its maximum capacity according to all four inference methods. *Roll out* provides results closer to the real values than the *Window* approach. During the winter season, the input from rainfall causes *Percentage Fullness* to increase, as is predicted by all four models, whilst at the beginning of spring (March) reservoir fullness falls according to the decrease in rainfall and increase in evaporation rates.

Even though results are coherent with the real situation, there is a difference with respect to real values marked in red. The explanation is that the models learnt in this study were proposed as a simplified example of a complex real-life problem (reservoir systems management) which demand that additional information be taken into account (rates of consumption, which reservoir provides the amount transfer in, to which reservoir the water is discharged). In this regard, our models considered the output of reservoirs comes mainly from evaporation losses and transfers to other reservoirs, without taking into account the water consumption use.

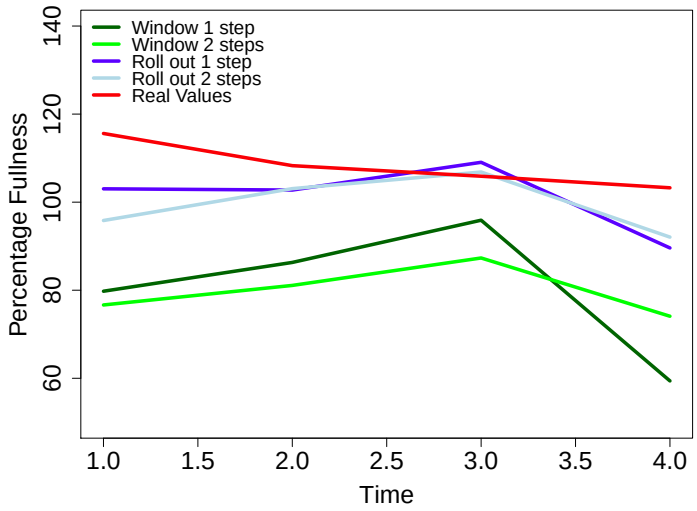


Fig. 12 Example of the evolution of variable *Percentage Fullness* over winter months

5 Discussion and Conclusions

Analysis of environmental temporal series data requires powerful tools capable to dealing with the inherent complexity of (socio-)natural systems over time. Several methods have been successfully proposed and applied, but their mathematical notation and, often, the fact that models act as a black box, mean they are not easily understood by non-mathematical experts. The role of expert and stakeholders in environmental modeling is crucial and their inclusion in the learning and validation processes necessary. This truth has encouraged new methodologies to be proposed that balance model accuracy and robustness with an intuitive component.

In this sense, the ability of BNs has been widely demonstrated in terms of the ability to deal with (static) environmental data whether including stakeholders or learning the mode solely from data. Besides, through the so-called *inference* process, scenarios of change can be included, and changes in the behavior of the system easily studied. Recently, the extension of BNs, in the form of Dynamic Bayesian networks, began to be applied to tackle temporal environmental problems.

A comparison between of the two learning methodologies demonstrated minor differences. In terms of structure, both models present the same three levels: climate, reservoir management and Fullness (measured using *Percentage Fullness* and *Water Level*). However, significant difference appears in terms of the complexity of the network, measured as the number of links. Whereas the 1-step approach learns the structure in a single step and the links between variables are fewer, the 2-step approach in contrast, repeats a static structure

and the number of intra-slices tends to be higher. Thus, computational time is higher for the 2-step approach.

The advantage of the 2-step over the 1-step approach is that no specific algorithm for dynamic model learning is required, which means that it provides an easier way to learn dynamic models for experts in environmental sciences who are familiar with static BNs. In general, depending on the goal of the model, either learning method can be applied and results obtained would be similar (in terms of error rate). However, in applications where the number of variables is high, the 1-step approach is recommended in order to reduce complexity and keep computational cost viable.

Inference in DBNs can be also handled with two approaches: *Window* and *Roll out*. Results in terms of *rmse* show a slight differences between the two approaches. In the case of *Window*, since the temporal structure is reduced to only two time-slices, complexity is not increased, and computational time is kept low in comparison with the *Roll out* approach. Furthermore, by moving the window, results are less influenced by the time-slice. However, in each step, the evidences are included as the mean value of the variable in the previous step, which could imply some information losses. In contrast, *Roll out* simultaneously represents all the time-slices involved in the inference process. Thus, it allows all the evidence propagation to be seen just in one step, rather than having to check the behavior of the system in several windows. Moreover, information from the previous time step is collected from the complete probability distribution, not from just the mean value.

However, this difference only applies when prediction is needed over more than one time step. If, for example, information about October is available, and only the behavior of water reservoir in November is required, *Window* and *Roll out* processes are equivalent. But, for prediction over all winter months (December, January, February and March), since *Roll out* shows the complete sequence of time slices, it is the most appropriate process, as we can seen from Figure 12 .

It is common that initially, only partial information is available concerning the problem to be modeled. In such cases, the model is an initial approximation used to learn about the problem. Through this probabilistic methodology, even when more variables and information are encouraged to be included, the initial model obtained provides stakeholders and experts with useful feedback, which help to improve the process. For the case study, the application of the DBN allows water reservoir behavior to be evaluated in the face of changing conditions. Even though further efforts are needed in modeling water reservoir management in Andalusia, an initial understanding of the problem and the variable relationships have been gained through the use of OE software and DBNs. Besides, under the scenario of change proposed, the evolution of percentage fullness of each reservoir can be computed, and studied in detail.

The following steps would consist of modeling a water reservoir system in depth using DBN, including more information about consumption rates, spatial information (relationships between different reservoirs) and water demand. In addition, taking advantage of the DBN versatility, this model can be

developed as an alarm system, which identifies the situations when reservoirs exceed a critical fullness threshold (to avoid both drought and flooding risk).

The literature contains numerous applications of static BNs for modeling natural systems under scenarios of future change. In these studies, the conclusions obtained are appropriate and robust, but temporal behavior is not properly studied. The systems modeled are based on a static picture, so that, when the new information is included, only some of the components of our picture are modified but the temporal relationships between the component of the systems are not taken into account. However, if a Dynamic BN is learnt, the system is modeled including relationships between components during each time period, but temporal interactions are also included in the dataset. Thus, performing scenarios of change using static models gives us information about the structure of relationships between components of the system, and how these components are related (*i.e.* if one variable changes, how it affects the others). But, if the objective is to studty a temporal process, and temporal data are available, the best solution is to model using a Dynamic BN.

A novelty of the current study is the use of MTEs models in DBNs. Their inclusion was firstly proved in Ropero et al (2017), where both continuous and discrete variables were simultaneously included into the models. In the current study, only continuous data were used, and MTEs was applied in the same way as in the study by Ropero et al (2017). It means MTEs models allow different type of data to be included with no modification into the algorithm or the learning process.

Nowadays, algorithms for both DBNs learning and inference are still under development. For a successful application in environmental sciences further effort is needed to encourage ecologists to apply them.

Acknowledgements This work has been supported by the Spanish Ministry of Economy and Competitiveness through projects TIN2016-77902-C3-3-P and TIN2013-46638-C3-1-P. Rosa F. Ropero is supported by a post-doctorate *Contrato Puente* funded by the University of Almería. The research presented in this paper was performed while R.F. Ropero stayed at Monash University; this research stay was supported by the Spanish Ministry of Education, Culture and Sport (grant EST14/00352).

References

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

Aguilera PA, Fernández A, Ropero RF, Molina L (2013) Groundwater quality assessment using data clustering based on hybrid Bayesian networks. Stochastic Environmental Research & Risk Assessment 27(2):435–447

Arya FK, Zhang L (2015) Time series analysis of water quality parameters at Stillaguamish river using order series method. Stochastic Environmental Research & Risk Assessment 29:227–239

von Asmuth JR, Maas K, Knotters M, Bierkens MFP, Bakker M, Olsthoorn T, Cirkel DG, Lenunk I, Schaars F, von Asmuth DC (2012) Software for hydrogeologic time series analysis, interfacing data with physical insight. Environmental Modelling & Software 38:178–190

Baum L, Peterie T, Souled G, Weiss N (1970) A maximization technique ocurring tin the statistical analusis of probabilistics functions of Markov chains. The Annals of Mathematical Statistics 40(1):164–171
Black A, Korb KB, Nicholson AE (2014) Intrinsic Learning of Dynamic Bayesian Networks. In: Pham N, Park S (eds) PRICAI 2014, LNAI 8862, pp 256–269
Bojarova J, Sundberg R (2010) Non-gaussian state space models in decomposition of ice core time series in long and short time-scales. Environmetrics 21:562–587
Boyen X, Koller D (1998) Tractable inference for complex stochastic processes. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp 33–42
Bromley J, Jackson NA, Clymer OJ, Giacomello AM, Jensen FV (2005) The use of Hugin® to develop Bayesian networks as aid to integrated water resource planning. Environmental Modelling & Software 20:231–242
Cano A, Moral S, Salmerón A (2000) Penniless propagation in join trees. International Journal of Intelligent Systems 15:1027–1059
Cano A, Moral S, Salmerón A (2002) Lazy evaluation in Penniless propagation over join trees. Networks 39:175–185
Castelletti A, Soncini-Sessa R (2007a) Bayesian networks and participatory modelling in water resource management. Environmental Modelling & Software 22:1075–1088
Castelletti A, Soncini-Sessa R (2007b) Coupling real-time control and socio-economic issues in participatory river basin planning. Environmental Modelling & Software 22:1114–1128
Cobb BR, Shenoy PP, Rumí R (2006) Approximating probability density functions with mixtures of truncated exponentials. Statistics and Computing 16:293–308
Cobb BR, Rumí R, Salmerón A (2007) Advances in probabilistic graphical models, Springer, chap Bayesian networks models with discrete and continuous variables, pp 81–102. Studies in Fuzziness and Soft Computing
Davidson JE, Stephenson DB, Turasie AA (2016) Time series modeling of paleoclimate data. Environmetrics 27:55–65
Dyer F, ElSawah S, Croke B, Griffiths R, Harrison E, Lucena-Moya P, Jakeman AJ (2014) The effects of climate change on ecologically-relevant flow regime and water quality attributes. Stochastic Environmental Research & Risk Assessment 28:67–82
Farah W, Nakhlé MM, Abboud M, Annesi-Maesano I, Zaarour R, Saliba N, Germanos G, Gerard J (2014) Time series analysis of air pollutants in Beirut, Lebanon. Environmental Monitoring Assessment 186:8203–8213
Fernandes JA, Lozano JA, Inza I, Irigoien X, Pérez A, Rodríguez JD (2013) Supervised pre-procesing approaches in multiple class variables clasification for fish recruitment forecasting. Environmental Modelling & Software 40:245–254
Fienen MN, Nolan BT, Feinstein DT (2016) Evaluating the sources of water to wells: Three techniques for metamodeling of a groundwater flow model. Environmental Modelling & Software 77:95–107
Henriksen HJ, Barlebo HC (2008) Reflections on the use of Bayesian belief networks for adaptive management. Journal of Environmental Management 88:1025–1036
Henriksen HJ, Rasmussen P, Brandt G, von Bülow D, Jensen FV (2007) Public participation modelling using Bayesian networks in management of groundwater contamination. Environmental Modelling & Software 22:1101–1113
Hill D, Minsker BS, Amir E (2009) Real-time Bayesian anomaly detection in streaming environmental data. Water Resource Research 45:1–16
Hill DJ (2013) Automated Bayesian quality control of streaming rain gauge data. Environmental Modelling & Software 40:289–301
Jensen F, Andersen S (1990) Approximations in Bayesian belief universes for knowledge-based systems. In: Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence, pp 162–169
Kelly R, Jakeman AJ, Barreteau O, Borsuk M, ElSawah S, Hamilton S, Henriksen HJ, Kuikka S, Maier H, Rizzoli E, Delden H, Voinov A (2013) Selecting among five common approaches for integrated environmental assessment and management. Environmental Modelling & Software 47:159–181
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). Ecological Modelling 268:48–54

Korb KB, Nicholson AE (2011) Bayesian Artificial Intelligence. CRC Press
Lagona F, Picone M, Maruotti A (2015) A hidden mark model for the analysis of cylindrical time series. Environmetrics 26:534–544
Langseth H, Nielsen TD, Rumí R, Salmerón A (2012) Mixtures of Truncated Basis Functions. International Journal of Approximate Reasoning 53(2):212–227
Liu R, Chen Y, Wu J, Gao L, Barret D, Xu T, Li L, Huang C, Yu J (2016) Assessing spatial likelihood of flooding hazard using naïve bayes and GIS: a case study in Bowen Basin, Australia. Stochastic Environmental Research & Risk Assessment 30:1575–1590
Lobo FL, Costa MP, Novo EM (2015) Time-series analysis of Landsat-MSS/TM/OLI images over Amazonian waters impacted by gold mining activities. Remote Sensing of Environment 157:170–184
Maldonado A, Aguilera P, Salmerón A (2016) Continuous Bayesian networks for probabilistic environmental risk mapping. Stochastic Environmental Research & Risk Assessment 30(5):1441–1455, DOI 10.1007/s00477-015-1133-2
Mantyka-Pringle CS, Martin TG, Moffatt DB, Linke S, Rhodes JR (2014) Understanding and predicting the combined effects of climate change and land-use change on freshwater macroinvertebrates and fish. Journal of Applied Ecology 51:572–581
Meineri E, Dahlberg CJ, Hylander K (2015) Using Gaussian Bayesian Network to disentangle direct and indirect associations between landscape physiography, environmental variables and species distribution. Ecological Modelling 313:127–136
Mihajlovic V, Petkovic M (2001) Dynamic Bayesian Networks: A State of the Art. Tech. rep., Electrical Engineering, Mathematics and Computer Science (EEMCS)
Molina J, Zazo S, Rodríguez-Gonzálvez P, González-Aguilera D (2016) Innovative analysis of runoff temporal behavior through bayesian networks. Water 8:1–21
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. Journal of Hydrology 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, Springer, vol 2143, pp 156–167
Moral S, Rumí R, Salmerón A (2003) Approximating conditional MTE distributions by means of mixed trees. In: ECSQARU'03. Lecture Notes in Artificial Intelligence, Springer, vol 2711, pp 173–183
Morales M, Rodríguez C, Salmerón A (2007) Selective naïve Bayes for regression using mixtures of truncated exponentials. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 15:697–716
Murphy K, Weiss Y (2001) The factores frontier algorithm for approximate inference in DBNs. In: Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, pp 378–385
Murphy KP (2002) Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis, University of California, Berkeley
Nicholson A, Flores J (2011) Combining state and transition models with dynamic Bayesian networks. Ecological Modelling 222:555–566
O´ Donnell R (2000) Flexible Causal Discovery with MML. PhD thesis, Faculty of Information Technology (Clayton). Monash University, Australia, 3800
Papakosta P, Straub D (2016) Probabilistic prediction of daily dire occurrence in the mediterranean with readily available spatio-temporal data. iForest Biogeosciences and Forestry 10:32–40
Parmar KS, Bhardwaj R (2015) Statistical, time series, and fractal analysis of full stretch of river Yamuna (India) for water quality management. Environmenta Science Pollutant Resource 22:397–414
Pearl J (1988) Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference. San Mateo, California
Pérez-Ramiréz PA, Bouwer-Utne I (2015) Use of dynamic Bayesian networks for life extension assessment of ageing systems. Reliability Engineering and Systems Safety 133:119–136
Phan T, Smart JC, Capon S, Hadwen W, Sahin O (2016) Applications of Bayesian belief networks in water resource management: A systematic review. Environmental Modelling

& Software 85:98–111
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
Raña P, Aneiros G, Vilar JM (2014) Detection of outliers in functional time series. Environmetrics 26:178–191
Romero V, Rumí R, Salmerón A (2006) Learning hybrid Bayesian networks using mixtures of truncated exponentials. International Journal of Approximate Reasoning 42:54–68
Ropero RF, Nicholson A, Korb K (2015) Using a new tool to visualize environmental data for bayesian network modelling. In: CAEPIA'15, Albacete, Spain
Ropero RF, Rumí R, Aguilera P (2016) Modelling uncertainty in social-natural interactions. Environmental Modelling & Software 75:362–372
Ropero RF, Flores MJ, Rumí R, Aguilera PA (2017) Applications of hybrid dynamic bayesian networks to water reservoir management. Environmetrics 28:1–11
Rumí R (2003) Modelos de redes bayesianas con variables discretas y continuas. PhD thesis, Universidad de Almería
Rumí R, Salmerón A (2007) Approximate probability propagation with mixtures of truncated exponentials. International Journal of Approximate Reasoning 45:191–210
Rumí R, Salmerón A, Moral S (2006) Estimating mixtures of truncated exponentials in hybrid Bayesian networks. Test 15:397–421
Russel S, Norvig P (2002) Artificial Intelligence: A Modern Approach, Pearson, chap Probabilistic reasoning over time, pp 542–583
Shenoy PP, West JC (2011) Inference in hybrid Bayesian networks using mixtures of polynomials. International Journal of Approximate Reasoning 52(5):641–657
Shenton W, Hart BT, Chan TU (2014) A Bayesian network approach to support environmental flow restoration decisions in the Yarra river, Australia. Stochastic Environmental Research & Risk Assessment 28:58–65
Spence R, Tweedie L (1998) The attribute explorer: information synthesis via exploration. Interacting with Computers 11:137–146
Spezia L, Futter MN, Brewer MJ (2010) Periodic multivariate normal hidden markov models for the analysis of water quality time series. Environmetrics 22:304–317
Taylor T, Dorin A, Korb K (2015) Omnigram Explorer: A Simple Tool for the Initial Exploration of Complex Systems, ECAL 2015
Tiller R, Gentry R, Richards R (2013) Stakeholder driven future scenarios as an element of interdisciplinary management tools; the case of future offshore aquaculture development and the potential effects on fishermen in Santa Barbara, California. Ocean & Coastal Management 73:127–135
Trifonova N, Kenny A, Maxwell D, Duplisea D, Fernandes J, Tucker A (2015) Spatio-temporal bayesian network models with latent variables for revealing trophic dynamics and functional networks in fisheries ecology. Ecological Informatics 30:142–158
Uusitalo L (2007) Advantages and challenges of Bayesian networks in environmental modelling. Ecological Modelling 203:312–318
Webster KL, McLaughlin JW (2014) Application of a bayesian belief network for assessing the vulnerability of permafrost to thaw and implications for greenhouse gas production and climate feedback. Environmental Science & Policy 38:28–44
Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann
Zhang Y, Qu Y, Wan J, Liang S, Liu Y (2012) Estimating leaf area index from MODIS and surface meteorological data using a dynamic bayesian network. Remote Sensing of Environment 127:30–43