Zazo, S.; Martín, A.; Molina, J.; Macian-Sorribes, H.; Pulido-Velazquez, M. (2022). Performance assessment of Bayesian Causal Modelling for runoff temporal behaviour through a novel stability framework. Journal of Hydrology. 610:1-12. https://doi.org/10.1016/j.jhydrol.2022.127832

Performance assessment of Bayesian Causal Modelling for runoff temporal behaviour through a novel stability framework

Santiago Zazo[1], Ana-María Martín[2,3], Jose-Luis Molina[2], Hector Macian-Sorribes[4], Manuel Pulido-Velázquez[4]

[1] Universitat Politècnica de València. Department of Hydraulic Engineering and Environment, C/ Camino de Vera s/n, 46022 Valencia, Spain. Corresponding author, e-mail: sanzadel@doctor.upv.es

[2] IGA Research Group. Salamanca University, High Polytechnic School of Engineering Avila. Av. de los Hornos Caleros, 50, 05003 Ávila, Spain

[3] Salamanca University. Department of Statistics, High Polytechnic School of Engineering, Ávila. Avda. de los Hornos Caleros, 50, 05003, Ávila, Spain.

[4] Universitat Politècnica de València. Research Institute of Water and Environmental Engineering (IIAMA). Ciudad Politécnica de la Innovación , C/ Camino de Vera s/n, 46022 Valencia, Spain.

Abstract

A strong innovative tendency is nowadays emerging that largely comprises new hydrological modelling approaches, based on Causal Reasoning through Probabilistic Graphical Modelling (PGM), because its ability to support probabilistic reasoning from data with uncertainty. These novel modelling frameworks are quite diverse and disperse not only in terms of techniques but also regarding its aims. It seems necessary to find a general and robust methodology for assessing its performance. This paper aims to provide a novel general methodology for assessing the performance of PGM based on Bayesian Causality for modelling and analysing the rivers´ runoff behaviour. For it, a structured four-step approach is developed and showed throughout the paper. The proposed methodology begins with the identification of the two main factors that condition the Bayesian Causal (BC) Modelling: the number of synthetic series (data amount) and the number of intervals for probability distributions for training and learning processes. The developed analysis comprises the definition of three levels for the first factor and seven levels for the second one, as well as the design of an innovative stability framework that assesses the level of BC Modelling performance. Furthermore, it has been necessary to create-define two novel indexes, named "Similarity and Stability Indexes" from 21 scenarios arising from the combination of the levels of the both factors. The optimal combination of factors is identified through a bi-objectives recursive approach based on previous indexes. Main results drawn successfully show a high relationship between the level of modelling performance, measured in terms of stability, and the river runoff temporal behaviour, measured in terms of temporal dependence. This research may help water managers and engineers to develop more rigorous and robust hydrological causal modelling implementations.

Key words: Probabilistic Graphical Models, Bayesian Causal Modelling, Stability, Time series analysis.

1. Introduction

Probabilistic Graphical Models (PGMs) since their emergence in the 1980s, within the statistical and artificial intelligence reasoning communities (Javidian et al., 2020), have been increasingly getting more notoriety in scientific community. This is mainly because they are an intuitive formalism to capture the probabilistic (in)dependence information of a domain (Pearl, 1988; Vogel et al., 2018) through the combination of probability-graph theories (Vogel et al., 2014), as well as a powerful framework for reasoning under uncertainty in knowledge-based systems (Cabañas de Paz, 2017; Javidian et al., 2020).

According to Koller and Friedman (2009), a PGM "*use a graph-based representation as the basis for compactly encoding a complex distribution over a high-dimensional space*". Formally, it is defined by a joint probability distribution *P(X)* on the set of variables of the problem *X*, and a graph *G* that represents the (in)dependence relationship amongst variables (Cabañas de Paz, 2017). At its core, they are characterized by a "*qualitative"* part which encodes a set of (conditional) (in)dependence relationships amongst variables, and a second "*quantitative*" one that measures the (in)dependence strength (Cabañas de Paz, 2017).

PGMs have been widely applied to a variety of reasoning tasks related to prediction, diagnosis, classification and risk assessment under original approaches and applied to different scientific disciplines. Engelke and Hitz (2020) introduces a general theory of conditional independence applied to graphical models for extremes. For its part, Javidian et al., (2020) propose a new algorithm (named LCD-AMP) applied to chain graphs, that overcomes the problems of data order dependence. Ramos Fernandez (2018) designs a Naïve Bayes PGM as a medical tool for predicting the severe course of acute bronchitis in infants. Lehikoinen et al., (2019), by Bayesian network (BN) classifiers, assesses the effects of multiple natural and anthropogenic factors on marine ecosystems simultaneously. For the first time, Macian-Sorribes et al., (2020) addresses the spatio-temporal dependence dimensions of inflows in hydrology through BNs. In Vogel et al., (2018) is identified the driving factors of flood loss at residential buildings applying novel PGMs approaches, based on BNs and Markov Blankets. On the other side, the relation among probability distribution and the graph defines the PGM type, existing three classes basically (Diez et al., 2018): (a) undirected graphs (e.g. Markov Networks), (b) directed acyclic graphs (e.g. BNs) and (c) chain graphs which are a unification of the first two (Javidian et al., 2020). Of these, the most widely used is the one based on BNs (Cabañas de Paz, 2017; Javidian et al., 2020), due to its ability to describe the overall behaviour of a system through the propagation of (in)dependence among the variables throughout the whole graphical structure (Vogel et al., 2018). In addition, the fact that BNs are based on the notion of conditional (in)dependence (Steck, 2001) makes them particularly useful for dealing with forecasting and analysing of the general behaviour in complex natural processes (Molina et al., 2019).

Referring specifically to hydrology and water resources, BNs have been satisfactorily applied to issues such as water planning of catchments (Mamitimin et al., 2015; Xue et al., 2017), integrated groundwater management (Carmona et al., 2011; Molina et al., 2013), reservoir management (Malekmohammadi et al., 2009), assessment of environmental risk on large dams (Ahmadi et al., 2015), flood damage (Schroeter et al., 2014) and drought forecasting and management (Shin et al., 2019), among others.

Recently, new hydrological challenges have begun to be addressed through the potential that BNs offer to discover causal structures in observed data (Spirtes, 2010), because the intrinsic (in)dependence relationships within of temporal series are not known in depth (Molina and Zazo, 2018; Molina et al., 2019). This is performed under a novel framework named Bayesian Causal (BC) Modelling, based on Causality, and which is addressed in form of Causal Reasoning (CR), supported by Bayesian modelling (Macian-Sorribes et al., 2020; Zazo, 2017). BC Modelling is a suitable technique for modelling the temporal behaviour of water resources of a basin when: (1) the approach is done from top to down, (2) the analysis is focused on the cause and (3) the objective comprises the prediction of the consequence (Macian-Sorribes et al., 2020: Pearl, 2009), as in the case of Bayesian Causality.

By means of BC Modelling it is highlighted the hidden logical temporal (in)dependence structure, that inherently underlies into hydrological historical records (Macian-Sorribes et al., 2020; Zazo et al., 2020). Under this approach, some noteworthy contributions have emerged in the field of stochastic hydrology. In Molina et al. (2016), a preliminary dynamic analysis of temporal dependence propagation was presented. For its part, Molina and Zazo (2018) discovered and quantified, for the first time, two opposite temporal-fractions within annual runoff time series, one conditioned by time and other no. Molina et al. (2019) developed a dynamic predictive runoff model based on the temporal behaviour trends of previously discovered runoff temporal fractions. Zazo et al. (2019) showed a qualitative approach for assessing temporal dependence through a novel dependence graph. Recently, Macian-Sorribes et al., (2020) analysed and assessed, simultaneous and completely, of spatio-temporal dimension of inflows across two adjacent and parallel river basins. In this sense, BC Modelling framework deliver much more accurate and powerful hydrological simulations and predictions, in line with new trends in stochastic hydrological research based on hybrid approaches between “traditional-novel techniques” (Mohammadi et al., 2006; Nourani et al.,2011; Valipour et al., 2012).

At this point, it is worth highlighting that PGMs are characterised by learning directly from the data when the structure of the network is unknown (Javidian et al., 2020; Malone and Yuan, 2012), and that such learning comprises not only the quantitative dimension of the data but also its intrinsic structure (Koller and Friedman, 2009; Steck, 2001). Therefore, issues such as the number of data or the modelling structure itself are crucial for suitable and reliable model performances (Koller and Friedman, 2009; Vogel et al., 2018). In this sense, there are significant algorithm-based developments for learning model structure from sampled data, which are well known-stablished in the scientific community. For instance, inductive causation (IC) algorithm (Verma and Pearl, 1991,1992), a similar approach independently developed by Peter Spirtes and Clark Glymour, known as PC algorithm (Spirtes et al., 2000; Madsen et al., 2003), based on conditional independence decisions (Le et al., 2016), and the improvement of the latter, named NPC learning algorithm, that introduces the notion of a Necessary Path Condition (Madsen et al., 2003; Steck, 2001). Nevertheless, in respect of the data there are two key driving factors for building a reliable dataset for learning and/or training which have not yet been addressed in depth. These are: (1) “number of synthetic series” and (2) “number of intervals” for discrete probability distributions.

The aim of this research is to provide a methodological framework for achieving an optimal combination between the different considered levels of two main driving factors that condition Learning-Training Processes of BC Modelling. This approach was applied to increase knowledge of the temporal behaviour of water resources in a river basin and have not yet been addressed by any hydrological research based on Bayesian Causality. The first, external one, is “number of synthetic series” that populates the Bayesian causal models (Factor 1; three considered levels in annual synthetic data). The second, internal one, is “number of intervals” for discrete probability distributions of synthetic data (Factor 2; with seven considered levels). This was developed to improve the performance of causal models with respect to their analytical and predictive capacity in the context of temporal behaviour of water resources of a basin. The influence between the levels of both factors was measured through two innovative indices, named "Similarity Index" and "Stability Index". These allowed the identification of the optimal combination between the levels of both factors through a bi-objectives recursive approach, in terms of overall stability of the BC Modelling process. These findings may be extrapolated to other approaches based on this type of PGMs for improving their analytical and forecast capabilities.

This paper is organized as follows: after this introduction, the methodological section and data description is addressed. Then, the main results drawn from the research are presented. Finally, in discussion and conclusions section the results are discussed in detail, as well as the general conclusions from the study are outlined.

2. Methodological approach

The framework for assessing the BC Modelling performance, in terms of overall stability process comprises a four-step approach (Fig 1). Step-1 implies the generation of synthetic data from historical records through a parametric Autoregressive Moving Average Model (ARMA) model or any alternative model approach able to generate synthetic time series. Use of observational data would only be possible if enough data to populate the BNs is available. Step-2 designs the BC Modelling by 4 sequential phases (Learning, Preprocessing, Constraints and Structure-Learning Process). In Step-3, probability propagation is addressed by Dependence Mitigation Graphs (DMGs), obtaining also its uncertainty bands through a Bootstrap approach. Finally, Step-4 assesses the BC Modelling performance innovatively. Firstly, a novel Similarity index is defined through an in-depth analysis of the overlap between uncertainty bands of wrap-around averages of DMGs. Secondly, a Stability index is created for assessing global BC Modelling performance of all possible combinations between the levels of both factors. Then, a bi-objectives recursive approach identifies the optimal combination of factors.

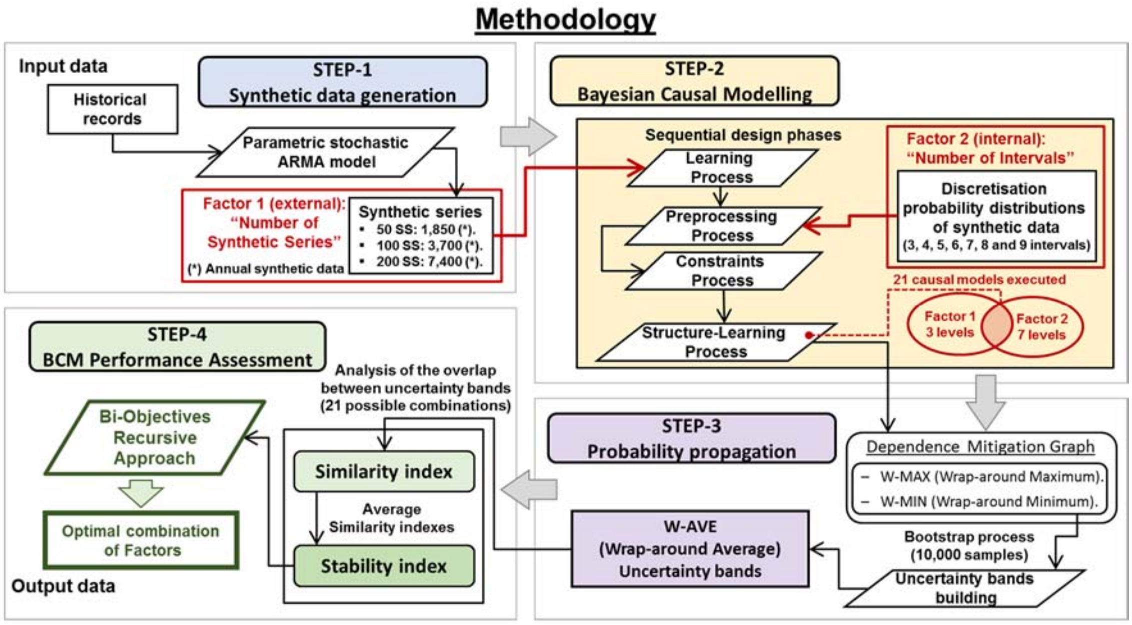

Fig 1. General methodology.

As DMG supposes an innovative way for assessing probability propagation, for understanding of the reader Uncertainty Band generation and their overlap analysis are facilitated by exemplary figures.

On the other side, the application of ARMA models to hydrological analysis/simulation and prediction reveals a wide heterogeneity of data samples. For example, in Srinivas and Srinivasan (2006) is assessed a hybrid model for stochastic simulation of multi season stream flows through 100 synthetic data seasonal-threshold model. Borgomeo et al., (2015) presented a method to generate synthetic monthly streamflow through random sampling from a parametric or a nonparametric distribution, assessed by 100 flow sequences. In Molina et al., (2019) a dynamic predictive model which was populated with 200 synthetic annual runoff series from an ARMA model is developed.

Based on the review of the scientific literature and in the expertise acquired in the development of causal models applied to the analysis and prediction of spatio-temporal behaviour of water resources in a river basin(s) three different numbers of annual synthetic series (Factor 1; 50, 100 y 200) were generated in each case study, all of them with the same length of the historical datasets. In contrast, Factor 2 (number of intervals for probability discretization) and whose influence has not yet been addressed within the framework of hydrological research, was given seven levels (3, 4, 5, 6, 7, 8 and 9 intervals). This led to the design, execution and analysis of the results of 21 causal models for each case study, which ensures the robustness of the findings and the designed methodological framework.

Data description - Case studies

The framework in which this research is conducted involves that the case studies form part of the methodological section. In order to define a robust methodological approach that would allow extrapolating the research findings to the field of prediction based on Bayesian causal models, two headwater case studies were chosen according to the following key aspects: (1) their completely opposite temporal behaviour, temporal (in)dependence, (2) same time period of analysis and (3) lack of spatial correlation between them.

Both headwaters (Porma in the north, and Adaja in the south) belong to Duero River Basin (the largest international river basin between Spain and Portugal). Furthermore, they are defined by gauging stations, located upstream the first regulating reservoir (flows under natural regime). They are: Camposolillo (code 2078) in the first case, and Adaja (code 2046) in the second one (Fig. 2a). The time period covered 37 hydrological years (from October to September) 1974/75 and 2010/11. The historical records (without missing data) were supported by the network of gauging stations of Duero river basin Authority (MITECO, 2021). Table 1 summarizes the main statistical parameters.

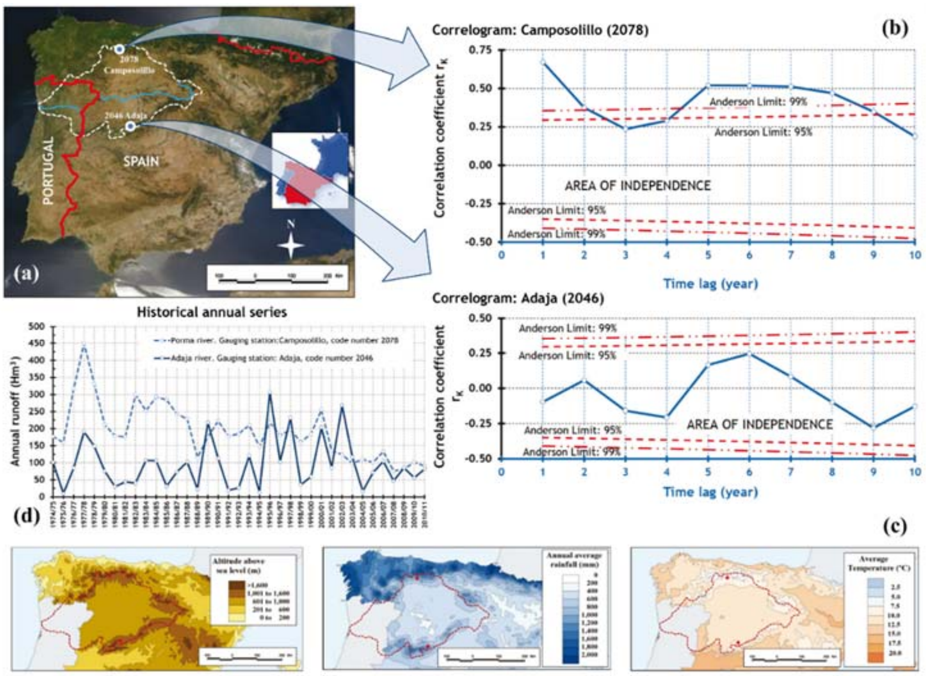

Fig 2. (a) Case studies location. (b) Correlograms and Anderson probability limits for an independent series (95 and 99 percent probability level). (c) Main climatic factors. (d) Historical annual series (1974/75 to 2010/11). Note: Hydrological year in Spain begins in October and ends in September of the following year.

Main statistic parameters	Sub-basins	
	Porma **"Camposolillo" (2078)**	**Adaja** **"Adaja" (2046)**
Annual mean (Hm^3):	191.29	95.90
Standard deviation (Hm^3):	80.92	73.44
Skewness coefficient:	0.98	1.25

Table 1. Historical records. Main statistical parameter.

Porma case study displays a temporal behaviour completely dependent, with its correlation coefficients outside the Anderson limits of the correlogram or very close it. In contrast, Adaja case study, exhibits an opposite behaviour, clearly independent, with all correlation coefficients within the independence area of the correlogram. Fig.2b displays both correlograms.

Both case studies are characterized by a cold Mediterranean climate, highly continental. Porma case study shows an annual average rainfall of 1430 mm. Geologically, there are important formations of limestones, sandstone and alluvial deposits in the fluvial courses. Hydrogeological speaking, it is characterized by low permeability or impermeable materials and low productivity and small extension aquifers (Molina and Zazo, 2018). Adaja case study presents an average annual about 400 mm. Geologically, it shows important areas of arkoses, granitic blocks and alluvial deposits at the bottom of the valley. Regarding hydrogeology, impermeable or very low permeability formations are present (Molina and Zazo, 2018). Fig. 2c displays the main climatic factors and Fig. 2d shows the graphic of both historical annual series.

2.1 Step-1. Synthetic data generation

This step is addressed through a parsimonious ARMA (1,1) model for each sub-basin, developed according to Molina et al. (2016). ARMA model plays a crucial role in this research because it provides reliable data to populate the BC Modelling process. In this sense, the purpose of this initial phase is to generate, in form of synthetic series all of them equiprobable to the historical series considered, the amount of information necessary to populate the subsequent Bayesian causal models. This stochastic, parsimonious and unconditioned model confers the highest degree of freedom to Bayesian modelling. An ARMA (p,q) is expressed as (Salas et al., 1980):

$$Y_t = \mu + \sum_{j=1}^{p} \emptyset_j \left(Y_{t-j} - \mu\right) + \epsilon_t - \sum_{j=1}^{q} \theta_j \, \epsilon_{t-j} \qquad (1)$$

where Y_t is the value of the temporal series at a certain time-step t, μ is the mean of the time series; (p,q) are the number of autoregressive (AR) and moving average (MA) average parameters, $\emptyset_j$, θ_j are the coefficient of AR and MA average model respectively and ϵ_t represents the historical residuals.

Before the postulation of the model, a normalization process was applied using three different methods: square root, natural logarithm plus one and natural double logarithm plus one (Salas et al., 1980). These functions are commonly applied to operational hydrology (Ochoa-Rivera et al., 2007). The natural logarithm plus one method was chosen in both cases, as it yielded the lowest skewness coefficient. Afterwards, the synthetic data (equiprobable to historical records) were generated.

The validation of synthetic time series was carried out comparing their relevant statistics to those of historical records (Ochoa-Rivera et al., 2007). In this case, due to use annual time series, this process was focused on mean values (Molina et al., 2021). Given that the levels considered for Factor 1 (50, 100 and 200 synthetic series) and the length of the historical data (37 years) a total of 1850, 3700 and 7400 annual synthetic data were respectively generated in each case study.

2.2 Step-2. Bayesian causal model design

Conceptually BC Modelling, as PGM based on BNs, is defined as a direct acyclic graph (DAG) in which conditional probability among variables is computed (Molina et al., 2021) and propagated omnidirectionally supported by Bayes' Theorem (Zazo et al., 2020). Mathematically, BC Modelling is expressed as (Madsen et al., 2003):

$$P(V) = \prod_{X \in V} P\big(X|pa(X)\big) \ , \tag{2}$$

where $P(V)$ is the joint probability distribution over set of nodes V (random variables, X) present in the graph and which are defined through DAG and $P(X|pa(X))$ is the conditional probability distribution for each variable $X \in V$ that belongs to a set of probability distributions.

BC Modelling application to hydrological time series makes possible to determine/quantify the strength of (in)dependence relationships between variables in a dynamic and step-by-step manner (Macian-Sorribes et al., 2020). This is performed through modifying and analyzing the evolution of the probability distribution over time

(Macian-Sorribes et al., 2020), by maximizing the statistical evidence of highest discrete probability distributions interval in each BN´s variable (Molina et al., 2016). After this maximization over a particular variable (node), the change (its influence) in the expected value of the other ones is assessed and quantified (Molina et al., 2016).

BC Modelling design was developed in 4 phases: (1) learning from synthetic data from ARMA (*1*,*1*) models (Factor 1); (2) preprocessing that implies the discretization of the synthetic data into discrete probability distributions (Factor 2), all of them with the same amplitude; (3) considering constraints initially that the main relationship among variables is natural (consecutive variables linking, time-lag=1); and (4) structure learning, which highlights the real structure of interdependence of runoff series by a constraint-based approach (Steck, 2001; Vogel et al., 2018), which finds a DAG structure from synthetic data through conditional independence statistical tests.

This final stage is the core of BC Modelling because it reveals the hidden, non-trivial and logical interdependence structure (time lag > 1) that underlies into hydrological series. It was done through NPC learning algorithm (Madsen et al., 2003) which allows efficient simplifications to find net-structures (Steck, 2001) under the condition of not including links between conditionally independent nodes (HUGIN, 2021).

BC Modelling design was supported by software HUGIN® Expert (version 7.3) with a significance level of 0.05 for Structure Learning Process. All BC Modelling parameters were fixed except the two factors analyzed.

2.3 Step-3. Probability propagation

This step comprises probability propagation analysis over the time throughout the network. through an innovative Dependence Mitigation Graphs (DMG, Zazo et al., 2019). DMG encompasses two variables: (1) “time” (X-axe; in form of time-lag) and (2) “% relative of change” (Y-axe; propagation of temporal dependence strength over the time). Conceptually, DMG is obtained through the relative percentage of change produced in the child variable, connected with the parent (Molina et al., 2016; Zazo et al., 2019).

Fig. 3 shows two DMGs for both sub-basins. Temporally dependent basins tend to present a dominant positive part (Fig 3a Porma). In contrast, temporally non-dependent basins show a tendency towards symmetric wrap-around results; positive or Wrap-around Maximum (W-MAX), and negative or Wrap-around Minimum (W-MIN), as found in Fig 3b Adaja (Molina et al., 2016).

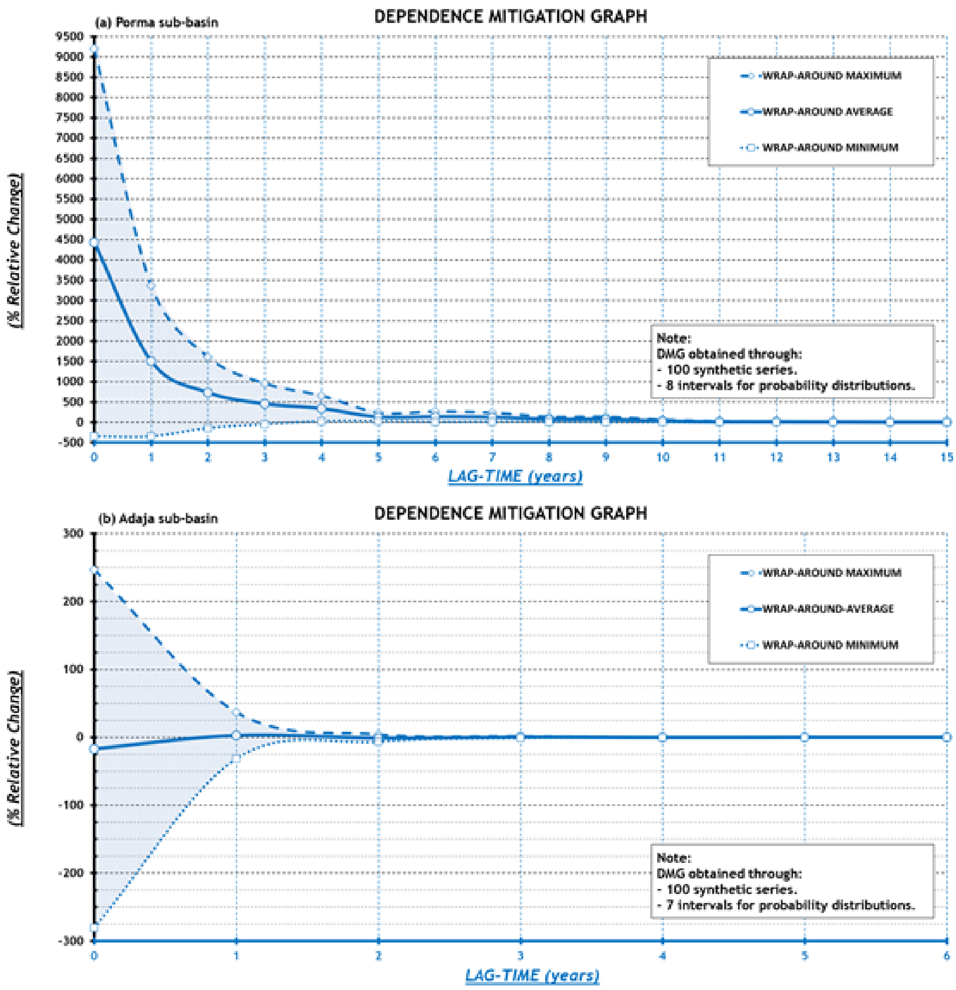

Fig 3. Dependence Mitigation Graphs for Porma (a) and Adaja (b) case studies.

2.3.1 Uncertainty bands generation

In order to study the uncertainty of the wrap-around results before changes in the levels of the two parameters/factors studied, 95 % uncertainty bands for W-MAX and W-MIN series were generated by Bootstrap estimation (Efron and Tibshirani, 1993).

For the band around W-MAX, let $\boldsymbol{x_1, x_2, \ldots, x_n}$ be the values used for getting $\mathbf{\max_i} \boldsymbol{x_i}$ for a particular lag. From these values, 10,000 bootstrap samples $\boldsymbol{x_1^*, x_2^*, \ldots, x_n^*}$ were generated and the maximum value $\mathbf{\max_i} \boldsymbol{x_i^*}$ was calculated for each one. Then, 2.5th and 97.5th percentiles of those maximum values were obtained. Those percentiles were named $(\mathbf{max}\, \boldsymbol{x^*})_{\mathbf{0.025}}$ and $(\mathbf{max}\, \boldsymbol{x^*})_{\mathbf{0.975}}$ respectively. Finally, the 95% confidence interval for this particular time-lag is developed as:

$$\left(2 \max_i x_i - (\max x^*)_{0.975};\ 2 \max_i x_i - (\max x^*)_{0.025}\right) \tag{3}$$

Then 95% uncertainty band for W-MAX series is calculated by joining the upper and lower limits from the confidence intervals. Similarly, it is determined W-MIN 95% confidence intervals. From both bands the 95% uncertainty band for the Wrap-around Average (W-AVE) can be constructed.

The usefulness of W-AVE is based on the application of the three-fold criterion “statistical-comparability-representativeness”: (1) Statistically, of the three wrap-arounds, W-AVE exhibits the lowest dispersion; (2) in river basins with temporally dependent behavior, W-MAX is clearly predominant with a minimal influence of W-MIN. By contrast, in independent ones, W-MAX and W-MIN tend to compensate each other (Molina et al., 2016; Zazo et al., 2019). Consequently, W-AVE can compare both types of river basins through a unique "wrap-around"; and (3) W-AVE is also able to capture/represent the temporal behavior of water basins. When W-AVE is generally positioned above the DMG X-axis the basin exhibits a dependent behavior; whereas if W-AVE is positioned in both sides of the X-axis the behavior is independent.

Finally, uncertainty band for W-AVE is built by averaging the 95% uncertainty bands for W-MAX and W-MIN series. In other words, the upper (lower) limit of the uncertainty band for the W-AVE series become the average of the upper (lower) limits of the uncertainty bands for the W-MAX and MIN series. In this manner, a single wrap-around (with its uncertainty band) is available. For a better understanding of this process Fig. 4 schematizes the W-AVE generation process and its uncertainty band.

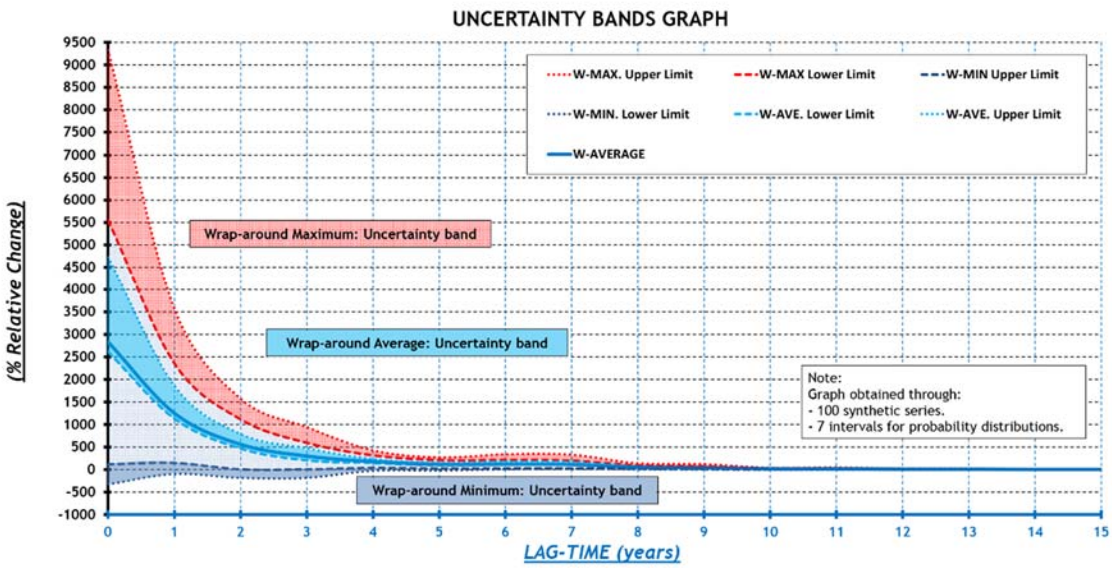

Fig 4. Graph of uncertainty bands. Wrap-around Maximum (W-MAX), Minimum (W-MIN) and Average (W-AVE). Note: Uncertainty bands at 95% confidence level. Example Porma case study.

2.4 Step-4. BC Modelling performance assessment

This step determines how the W-AVE built depends on Factors 1 and 2 of BC Modelling. For this purpose, for each of the 21 scenarios (3x7) considered, W-AVE and 95% uncertainty band are constructed and then the overlapping between all of them is analysed. In this regard, two novel indexes have been defined: (1) Similarity index and (2) Stability index. The first one analyses in depth the overlap between uncertainty bands for the W-AVE, and the second one assesses the overall performance of the process.

Moreover, in the scientific literature there are a variety of indices based on Bayesian approaches, such as the Bayesian Information Criterion (Schwarz, 1978) one of the most widely applied to model selection, those arising from the combination of different indices and causes (Bradley et al., 2015; Kim et al., 2018), or from the combination of different causes to assess a particular problem (Giné-Garriga et al., 2018), among others.

2.4.1 Similarity index

For each pairwise comparison of scenarios a Similarity Index is calculated through a complete Uncertainty Bands Overlapping Analysis (UBOA) as follows:

1. If the two uncertainty bands overlap for a certain time-lag, the value 1 is assigned to that time-lag. Otherwise, a 0 value is assigned. This is mathematically expressed by a Boolean Factor of Overlap (F_o).

2. A weighted sum of the overlaps of all time-lag is obtained through following weights (w_i): 0.5 for lag_0 and lag_1 ($w_0=w_1=0.5$), 0.1 for lag_2 to $lag4$ ($w_2=w_3=w_4=0.1$) and 0.01 for the rest *lags* (w_i=0.01 for $i > 4$). This weight structure is based on the acquired knowledge on DMG (Zazo et al., 2019). DMG allows us to identify two regions of higher and lower mitigation of time dependence. Of these, lag_0 and lag_1 are those that present a higher mitigation gradient, for that $w_0=w_1=0.5$. The limit between these regions has been observed in lag_4 or lag_5, mainly in lag_4, which justifies the assignment of weight 0.1 to $lag_{2\text{-}4}$. From lag_5 the mitigation gradient is gradual, up to a relative percentage of change equal to 0.

3. The weighted sum is corrected through a multiplicative factor (0, 1, 2, 3, 4, 5) that considers the consecutiveness of the overlap of uncertainty bands for the first 5 time-lags, denoted as $F_{(c)}$. If the bands are overlapped within the first 5 lags, multiplier $F_{(c)}$ is 5; 4 if the overlap is within the first 4 lags, and successively. $F_{(c)}=0$ if no overlap in the first 5 lags. This threshold (first 5 time-lags, up to lag_4)

is defined according to the limit between the two regions of dependence mitigation indicated above.

4. The Similarity Index of each pairwise comparison is obtained by the normalization of the previous rate to be enclosed within 0 (no overlap) and 1 (total overlap). This normalization is done by dividing the previous rate by the summation of all time-lag weights and multiplying the result by 5 (maximum F_o value). Fig. 5 schematizes UBOA process for a Similarity index value of 0.71. Similarity Index is formally expressed as:

$$Similarity\ Index = \frac{F_{(c)} \cdot \sum(w_i \cdot F_o)}{5 \cdot \sum w_i}\ , \tag{4}$$

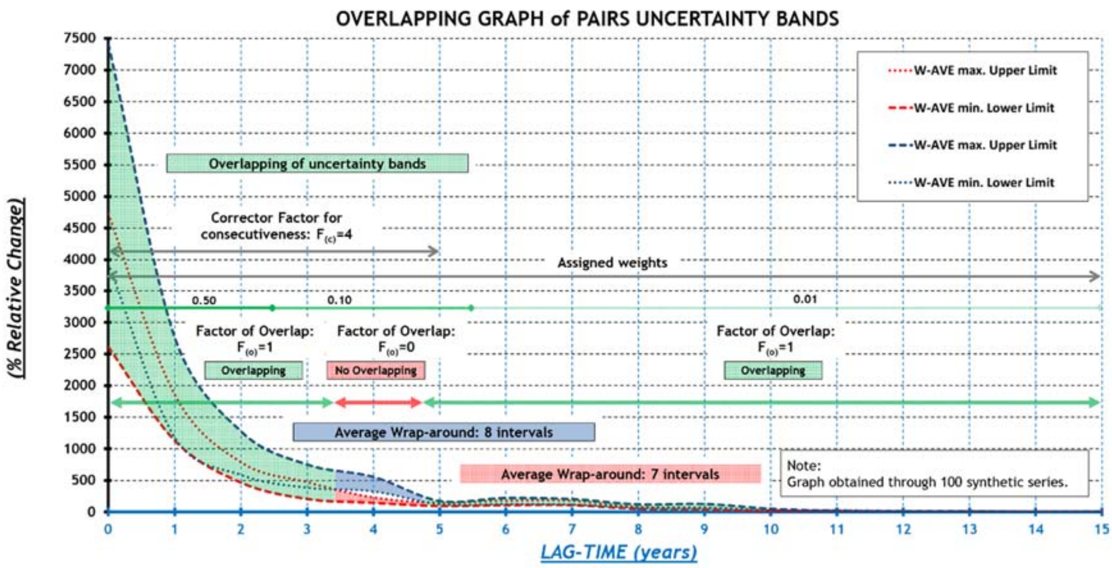

Fig 5. Uncertainty Bands Overlapping Analysis. Example Porma case study of overlapping between uncertainty bands. Note: This analysis process is repeated for each of the 21 possible combinations resulting from Factor levels 1 and 2.

Then, Similarity index values were arranged in matrix form and grouped on two categories and five classes (Table 2). For each case study 10 different matrices were defined: seven 3x3 matrixes for grouping the results of Factor 1 vs Factor 2, and three 7x7 matrixes in the case of Factor 2 vs Factor 1 results.

Similarity categories	Classes	Color code	Description
Similarity Index ≥ 0.50 High evidence of similarity	[1.00 , 0.90]		Fully
	(0.90 , 0.80]		Highly
	(0.80 , 0.65]		Similar
	(0.65 , 0.50]		Slightly
Similarity Index < 0.50 Null or Low evidence of similarity	(0.50 , 0.00]		

Table 2. Similarity Index. Categories and classes.

2.4.2 Stability index

The Stability Index is computed by averaging the Similarity indexes of each scenario with the others (including the scenario in hand) in the form of a transposed matrix. Consequently, the previous ten matrixes for each case study are summarized in only two: one of dimensions 3x7 in the case of Factor 1, and second one of 7x3 for grouping the results of Factor 2. Both matrixes are expressed as:

$$\boldsymbol{S}_{F-1\,(3x7)} = \begin{pmatrix} \frac{\sum_{m=1}^{3}\left(SI_{L1,F2}\right)_{m,1}}{3}_{L1,F1} & \cdots & \frac{\sum_{m=1}^{3}\left(SI_{L7,F2}\right)_{m,1}}{3}_{L1,F1} \\ \frac{\sum_{m=1}^{3}\left(SI_{L1,F2}\right)_{m,1}}{3}_{L2,F1} & \cdots & \frac{\sum_{m=1}^{3}\left(SI_{L7,F2}\right)_{m,1}}{3}_{L2,F1} \\ \frac{\sum_{m=1}^{3}\left(SI_{L1,F2}\right)_{m,7}}{3}_{L3,F1} & \cdots & \frac{\sum_{m=1}^{3}\left(SI_{L7,F2}\right)_{m,7}}{3}_{L3,F1} \end{pmatrix} \quad (5)$$

$$\boldsymbol{S}_{F-2\,(7x3)} = \begin{pmatrix} \frac{\sum_{m=1}^{7}\left(SI_{L1,F1}\right)_{m,1}}{7} & \frac{\sum_{m=1}^{7}\left(SI_{L2,F1}\right)_{m,1}}{7} & \frac{\sum_{m=1}^{7}\left(SI_{L3,F1}\right)_{m,1}}{7} \\ \vdots & \vdots & \vdots \\ \frac{\sum_{m=1}^{7}\left(SI_{L1,F1}\right)_{m,7}}{7} & \frac{\sum_{m=1}^{7}\left(SI_{L2,F1}\right)_{m,7}}{7} & \frac{\sum_{m=1}^{7}\left(SI_{L3,F1}\right)_{m,7}}{7} \end{pmatrix} \quad (6)$$

where $SI_{Li,F\text{-}1}$ and $SI_{Li,F\text{-}2}$ are the mean values of Similarity Index (*SI*) of each levels of Factors 1 and 2 respectively.

Then, overall performance of Stability indexes was recursively analysed through trend graphs, according to their classification in two categories and five classes (Table 3).

Stability categories	Classes	Description
Stability Index ≥ 0.50 High evidence of stability	[1.00 , 0.90]	Fully
	(0.90 , 0.80]	Highly
	(0.80 , 0.65]	Stable
	(0.65 , 0.50]	Slight stable
Stability Index < 0.50 Null or Low evidence of stability	(0.50 , 0.00]	

Table 3. Stability Index. Categories and classes.

2.4.3 Stability framework

The influence of the factors over the BC Modelling Performance is analyzed through a recursive process that provides the highest possible stability:

1. The process starts with the evaluation of Factor 1, through plotting the performance of the Stability Index (Y-axis) versus Factor 2 (X-axis). In this graph, the Factor 1 level with the highest Stability Indexes is selected.
2. The level value of Factor 2 is determined through plotting the trend of the Stability Index (Y-axis) versus Factor 1 (X-axis). The number of intervals (Factor 2 levels) is chosen so that the Stability Index for the previously determined value of Factor 1 displays the highest value. This is done recursively if there are several values that satisfy the indicated condition.

3. Results

3.1 Generation of synthetic time series

Average values of mean, standard deviation and skewness coefficient were determined from each historical records and synthetic data time series of each case study (Table 4). It is worth to highlight the maintenance of the mean value, according to the average behaviour that annual data offer. Fig. 6 shows the spectrum (wrap-arounds of each set of synthetic data of Factor 1), which will populate the BC Modelling.

Gauging station			Historical time series	50 SS	100 SS	200 SS
	Camposolillo	Annual mean (Hm^3):	191.29	195.88	192.57	189.64
		Standard deviation (Hm^3):	80.92	75.32	74.72	72.55
		Skewness coefficient:	0.96	0.90	0.88	0.85
	Adaja		**Historical time series**	**50 SS**	**100 SS**	**200 SS**
		Annual mean (Hm^3):	95.90	102.16	100.93	101.11
		Standard deviation (Hm^3):	73.44	100.68	97.26	97.77
		Skewness coefficient:	1.25	2.19	2.14	2.15

Table 4. Main statistical parameters: Historical records vs. Synthetic data generation.

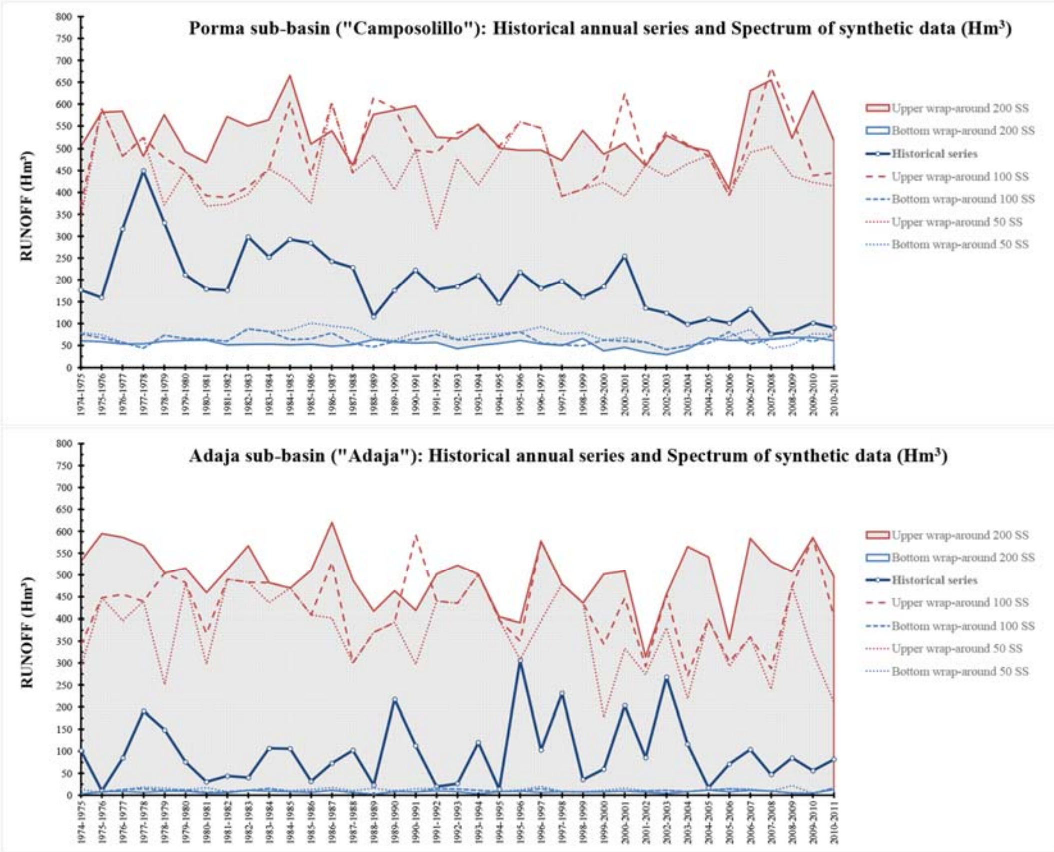

Fig 6. Historical records and Spectrum of synthetic data (3 level) to populate Bayesian Causal Modelling. (Upper) Porma sub-basin. (Bottom) Adaja sub-basin.

3.2 Similarity indexes

Fig. 7 shows Similarity Indexes obtained in matrix form. In general, it is noticeable that the values are relevant for 6 intervals and beyond (Fig. 7 left). For this reason, the analysis of the resulting 7x7 dimension matrixes (Factor 1 vs. Factor 2) focuses on intervals 6 to 9. In contrast, 3x3 matrixes of the Factor 2 vs. Factor 1 are analysed for all of them.

In Porma case study, the highest values are obtained with 50 and 200 SS (Factor 1), with the combination of 200 SS and intervals 6 to 9 (Factor 2) showing the highest indexes [(6-7=0.94), (6-8=0.94), (6-9=0.97), (7-8=0.97), (7-9=0.99) and (8-9=0.74)]. In contrast, in the case of 50 SS the highest values are only obtained with the combination of 7 to 9 intervals [(7-8=0.96), (7-9=0.94) and (8-9=1.00)]. In both cases (50-200 SS), the values are homogeneous, with minimal variability, high evidence of similarity and belonging principally to the "Fully similar" class.

On the other hand, the matrixes (Factor 1 vs. Factor 2) for Adaja case study are less homogeneous, although the highest values are also associated with 50 and 200 SS. Moreover, the 200 SS level shows that for 6 to 9 intervals there is an alternation between values with null or low evidence of similarity [(6-8=0.34), (7-8=0.32), (8-9=0.34)] and values belonging to class slight similarity [(6-7=0.75), (6-9=0.75), (7-9=0.75)]. In the 50 SS level the disparity of values is meaningful, where values of practically null evidence of similarity [(7-8=0.12, 8-9=0.39)] are alternating with values of total similarity [(7-9=0.99)].

In both case studies, the 100 SS level Factor 1 shows the highest variability of indexes, even with the lowest values (the lowest evidence of similarity), Porma: [(6-9=0.07, (7-9=0.08), (8-9=0.19)], Adaja: [(6-8=0.10), (7-8=0.39), (8-9=0.10)]. On the other hand, the analysis of the 3x3 matrixes (Factor 2 vs. Factor 1) highlights clearly that the highest indexes were found for 7 intervals except for the combination of 5 intervals with 100 and 200 SS in the Adaja case study (value of 0.75). In the rest of combinations, the values show null or low evidence of similarity.

Factor 1 (Synthetic series) vs. Factor 2 (Intervals)

50 Synthetic series

	Porma sub-basin								Adaja sub-basin						
	3	4	5	6	7	8	9		3	4	5	6	7	8	9
3	1.00	0.30	0.07	0.08	0.00	0.00	0.00	3	1.00	0.09	0.08	0.09	0.09	0.08	0.69
4	0.30	1.00	0.94	0.33	0.00	0.00	0.00	4	0.09	1.00	0.18	0.37	0.54	0.11	0.97
5	0.07	0.94	1.00	0.97	0.08	0.09	0.08	5	0.08	0.18	1.00	0.04	0.21	0.09	0.48
6	0.08	0.33	0.97	1.00	0.35	0.10	0.33	6	0.09	0.37	0.04	1.00	0.55	0.99	0.39
7	0.00	0.00	0.08	0.35	1.00	0.96	0.99	7	0.09	0.54	0.21	0.55	1.00	0.12	0.99
8	0.00	0.00	0.09	0.10	0.96	1.00	1.00	8	0.08	0.11	0.09	0.99	0.12	1.00	0.39
9	0.00	0.00	0.08	0.33	0.99	1.00	1.00	9	0.69	0.97	0.48	0.39	0.99	0.39	1.00
µ	0.21	0.37	0.46	0.45	0.48	0.45	0.49	µ	0.30	0.47	0.30	0.49	0.50	0.40	0.70

100 Synthetic series

	3	4	5	6	7	8	9		3	4	5	6	7	8	9
3	1.00	0.17	0.00	0.00	0.00	0.00	0.00	3	1.00	0.46	0.08	0.08	0.08	0.08	0.07
4	0.17	1.00	0.08	0.00	0.01	0.00	0.00	4	0.46	1.00	0.09	0.09	0.09	0.09	0.09
5	0.00	0.08	1.00	0.34	0.92	0.31	0.27	5	0.08	0.09	1.00	0.34	0.75	0.12	0.51
6	0.00	0.00	0.34	1.00	0.72	0.92	0.07	6	0.08	0.09	0.34	1.00	0.75	0.10	0.51
7	0.00	0.01	0.92	0.72	1.00	0.71	0.08	7	0.08	0.09	0.75	0.75	1.00	0.39	0.51
8	0.00	0.00	0.31	0.92	0.71	1.00	0.19	8	0.08	0.09	0.12	0.10	0.39	1.00	0.10
9	0.00	0.00	0.27	0.07	0.08	0.19	1.00	9	0.07	0.09	0.51	0.51	0.51	0.10	1.00
µ	0.17	0.18	0.42	0.44	0.49	0.45	0.23	µ	0.26	0.27	0.41	0.41	0.51	0.27	0.40

200 Synthetic series

	3	4	5	6	7	8	9		3	4	5	6	7	8	9
3	1.00	0.27	0.00	0.00	0.00	0.07	0.00	3	1.00	0.37	0.11	0.37	0.49	0.30	0.71
4	0.27	1.00	0.30	0.07	0.08	0.07	0.07	4	0.37	1.00	0.53	0.49	0.49	0.30	0.70
5	0.00	0.30	1.00	0.30	0.32	0.89	0.31	5	0.11	0.53	1.00	1.00	0.52	0.34	0.75
6	0.00	0.07	0.30	1.00	0.94	0.96	0.94	6	0.37	0.49	1.00	1.00	0.75	0.34	0.75
7	0.00	0.08	0.32	0.94	1.00	0.97	0.99	7	0.49	0.49	0.52	0.75	1.00	0.32	0.75
8	0.07	0.07	0.89	0.96	0.97	1.00	0.74	8	0.30	0.30	0.34	0.34	0.32	1.00	0.34
9	0.00	0.07	0.31	0.94	0.99	0.74	1.00	9	0.71	0.70	0.75	0.75	0.75	0.34	1.00
µ	0.19	0.27	0.45	0.60	0.61	0.67	0.58	µ	0.48	0.55	0.61	0.67	0.62	0.42	0.71

Factor 2 (Intervals) vs. Factor 1 (Synthetic series)

	Porma sub-basin				Adaja sub-basin		
3 intervals	50 SS	100 SS	200 SS		50 SS	100 SS	200 SS
50 SS	1.00	0.08	0.00	50 SS	1.00	0.21	0.00
100 SS	0.08	1.00	0.10	100 SS	0.21	1.00	0.02
200 SS	0.00	0.10	1.00	200 SS	0.00	0.02	1.00
µ	0.36	0.39	0.37	µ	0.40	0.41	0.34
4 intervals	50 SS	100 SS	200 SS		50 SS	100 SS	200 SS
50 SS	1.00	0.00	0.00	50 SS	1.00	0.09	0.29
100 SS	0.00	1.00	0.10	100 SS	0.09	1.00	0.00
200 SS	0.00	0.10	1.00	200 SS	0.29	0.00	1.00
µ	0.33	0.37	0.37	µ	0.46	0.36	0.43
5 intervals	50 SS	100 SS	200 SS		50 SS	100 SS	200 SS
50 SS	1.00	0.08	0.00	50 SS	1.00	0.19	0.19
100 SS	0.08	1.00	0.30	100 SS	0.19	1.00	0.75
200 SS	0.00	0.30	1.00	200 SS	0.19	0.75	1.00
µ	0.36	0.46	0.43	µ	0.46	0.65	0.65
6 intervals	50 SS	100 SS	200 SS		50 SS	100 SS	200 SS
50 SS	1.00	0.00	0.00	50 SS	1.00	0.23	0.23
100 SS	0.00	1.00	0.50	100 SS	0.23	1.00	0.37
200 SS	0.00	0.50	1.00	200 SS	0.23	0.37	1.00
µ	0.33	0.50	0.50	µ	0.49	0.53	0.53
7 intervals	50 SS	100 SS	200 SS		50 SS	100 SS	200 SS
50 SS	1.00	0.95	0.04	50 SS	1.00	0.12	0.55
100 SS	0.95	1.00	0.49	100 SS	0.12	1.00	0.75
200 SS	0.04	0.49	1.00	200 SS	0.55	0.75	1.00
µ	0.66	0.81	0.51	µ	0.56	0.62	0.77
8 intervals	50 SS	100 SS	200 SS		50 SS	100 SS	200 SS
50 SS	1.00	0.03	0.21	50 SS	1.00	0.97	0.10
100 SS	0.03	1.00	0.33	100 SS	0.97	1.00	0.00
200 SS	0.21	0.33	1.00	200 SS	0.10	0.00	1.00
µ	0.41	0.45	0.51	µ	0.69	0.66	0.37
9 intervals	50 SS	100 SS	200 SS		50 SS	100 SS	200 SS
50 SS	1.00	0.29	0.03	50 SS	1.00	0.99	0.99
100 SS	0.29	1.00	0.31	100 SS	0.99	1.00	0.51
200 SS	0.03	0.31	1.00	200 SS	0.99	0.51	1.00
µ	0.44	0.53	0.45	µ	0.99	0.83	0.83

Fig 7. Similarity Indexes. (Left) Matrixes (7x7): Factor 1 vs. Factor 2. (Right) Matrixes (3x3): Factor 2 vs. Factor 1. Please note it additionally includes the mean value of each scenario denoted by µ, which justifies the posterior Stability Index.

Finally, it is evident that the highest similarity is achieved with 200 SS, being its combinations with 7 to 9 intervals the ones that “a priori” offer the greatest similarity. The assessment of the overall performance of both factors through the Stability Indexes, and by means bi-objectives recursive process, will allow defining the optimal combination of both factors.

3.3 BC Modelling performance assessment

3.3.1 Stability indexes

Table 5 summarizes the Stability indexes of Factor 1 vs. Factor 2. In contrast, Table 6 shows the Stability indexes resulting according to the levels of Factor 2 for fixed levels of Factor 1. It is observed that the highest values of the Stability Index are mainly associated with the combination of 200 SS (Factor 1) with 7 to 9 intervals (Factor 2). In particular, for Porma case study: [200 SS vs. (7 intervals = 0.61), (8 intervals = 0.67)], and for Adaja case study: [200 SS vs. (6 intervals = 0.67), (7 intervals = 0.62), (9 intervals = 0.71)] (see Table 5).

On the other side, it is worth noting that in Porma case study no stability index is above 0.5 (lower limit of the stability threshold) for the levels 50 and 100 SS of Factor 1, and almost the same result is observed in Adaja save two combinations: [(50 SS, 9 intervals) = 0.70] and [(100 SS, 7 intervals) = 0.51]. Furthermore, in Adaja stability indices for all combinations of Factor 1 with 8 intervals are lower than the preceding and following ones (0.40, 0.27, 0.42).

		Gauging stations					
		Porma (2078)			**Adaja (2046)**		
		Synthetic series			**Synthetic series**		
		50	**100**	**200**	**50**	**100**	**200**
Intervals	3	0.21	0.17	0.19	0.30	0.26	0.48
	4	0.37	0.19	0.27	0.47	0.27	0.55
	5	0.46	0.42	0.45	0.30	0.41	0.61
	6	0.45	0.44	0.60	0.49	0.41	**0.67**
	7	0.48	0.49	**0.61**	0.50	0.51	**0.62**
	8	0.45	0.45	**0.67**	0.40	0.27	0.42
	9	0.49	0.23	0.58	**0.70**	0.40	**0.71**

Table 5. Stability Indexes of levels of Factor 2 fixing the levels of Factor 1. Note: The highest values are indicated in bold type. Please see Figure 7 where the Stability Index values are justified.

Additionally, in both case studies the Stability Indexes for 200 SS and 7 intervals exhibit a high and almost equal evidence of stability (0.61 and 0.62 respectively), being the combination that offers the greatest stability of results with the minimum spread of values.

Regarding the Stability Indexes of Factor 1 levels after setting the levels of Factor 2 (Table 6), the highest stability indexes are achieved, in general, from 7 intervals in all levels of Factor 1. In this sense, the highest values are observed in Adaja case study for 9 intervals (0.99, 0.83 and 0.83) and in Porma case study for 100 SS and 7 intervals (0.81). However, a detailed analysis reveals that the level of the 7 intervals of Factor 2 is the only one for which the resulting values are all higher than 0.50 in both cases. In particular, Adaja case study shows increasing values, from 0.56 to 0.77.

Gauging stations			Number of intervals						
			3	**4**	**5**	**6**	**7**	**8**	**9**
	Porma	50 SS	0.36	0.33	0.36	0.33	**0.66**	0.41	0.44
		100 SS	0.39	0.37	0.46	0.50	**0.81**	0.45	0.53
		200 SS	0.37	0.37	0.43	0.50	0.51	0.51	0.45
			Number of intervals						
			3	**4**	**5**	**6**	**7**	**8**	**9**
	Adaja	50 SS	0.40	0.46	0.46	0.49	0.56	**0.69**	**0.99**
		100 SS	0.41	0.36	**0.65**	0.53	**0.62**	**0.66**	**0.83**
		200 SS	0.34	0.43	**0.65**	0.53	**0.77**	0.37	**0.83**

Table 6. Stability Indexes of levels of Factor 1 fixing the levels of Factor 2. Note: The highest values are indicated in bold type.

3.3.2 Bi-objective recursive process. Optimal combination of levels

The influence of both factors over the general BC Modelling Performance is analysed through a bi-objective recursive approach inspired by a One-Way ANOVA.

As indicated previously, the highest stability index for Factor 1 level is determined firstly, and then the stability trend for the levels of Factor 2 is observed for the set level. In this manner, both factors are controlled simultaneously and recursively. Fig. 8 displays the overall stability framework of the process.

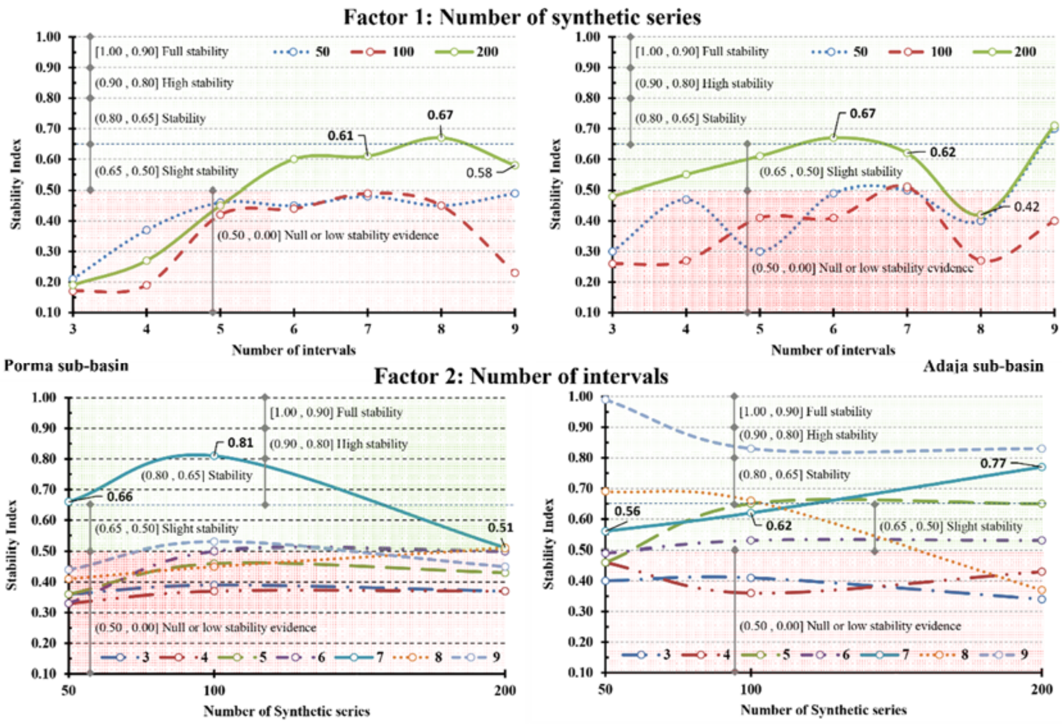

Fig 8. Overall stability framework. (Upper) Factor 1 (3 levels) vs. Factor 2. (Bottom) Factor 2 (7 levels) vs. Factor 1. Please note that the category of null or low evidence of stability, (0.50 , 0.00], is coloured in light red. In contrast, the category of high evidence of stability, [1.00 , 0.50], is coloured in light green. For Stability index thresholds please see Table 3.

The choice of the level of Factor 1 (upper Fig. 8) is clear. In both case studies the highest overall Stability indexes and trends are only achieved for the level 200 SS of Factor 1. In the case of Porma sub-basin, the values display an increasing trend from 0.19 (minimum value; 3 intervals) to 0.60 (6 intervals). From this point a stability is observed, with a maximum of 0.67 (8 intervals) and a gradual decrease to 0.58 (9 intervals). In the case of the other two levels of Factor 1 (50 and 100 SS), the described behavioural pattern is generally maintained, but always inside the category of null or low evidence of stability. On the other hand, in the Adaja sub-basin case, the highest overall stability trend is found for the levels of 6 and 7 intervals of Factor 2 (0.67 maximum and 0.62 respectively). The other two levels for Factor 1 exhibit an irregular stability behaviour.

Once set the level of Factor 1 (200 SS), the best level of Factor 2 is determined the graphs at the bottom of Figure 8. In Porma case study, all levels of Factor 2 display a pattern of lack of relevant stability in the results, excluding the level of 7 intervals. The overall stability indexes of this level are higher than 0.50 in all cases. In contrast, in Adaja case study is not so evident with two meaningful levels. Whilst the 9-interval level shows the highest stability values, with a clearly stable trend at the value 0.83 for 100 and 200 SS (see Table 6), the 7-interval level shows a clearly increasing trend of stability up to

0.77 (200 SS). Considering the minimal difference between both values of the stability index (0.83 vs. 0.77, roughly 7%), both values confirm a high evidence of stability and, for consistency and homogeneity of results with the previous case, the best level of intervals is achieved for 7 ones.

Consequently, the optimal combination of levels of both factors is obtained with 200 SS and 7 intervals.

4. Discussion and Conclusions

This research proposes an innovative methodological framework for assessing and improving the performance of BC Modelling, applied to the analysis and the prediction of runoff temporal behaviour of a river basin, based on Causality. For that, a robust stability framework to assess performance Learning-Training Processes of BC Modelling is set up employing an innovative bi-objective recursive approach. This is conditioned by two main factors: the number of synthetic series that populate the causal models (Factor 1), and the number of intervals defined for discrete probability distributions of synthetic data (Factor 2). These factors have not yet been addressed by any hydrological research based on Bayesian Causality.

This innovative stability framework has led to the development of two novel indicators. The first one, named "Similarity Index", enables an in-depth analysis of the overlap between wrap-around of DMGs (in this case focused on W-AVE uncertainty bands). The second one, called "Stability Index", assesses the overall BC Modelling performance process. In addition, the number of causal models executed/analyzed together with the strong mathematical foundations applied highlights the robustness and reliability of the findings.

Moreover, this research has revealed, for the first time, the influence of the number of intervals of probability distributions of synthetic data (Factor 2) on the BC Modelling process, and by extension to Bayesian PGMs, by means of Similarity and Stability indexes. This factor had not yet been analysed in a hydrological causal reasoning context.

This research, focused on two exemplary temporal case studies through bi-objective recursive developed approach, highlights that the most stable combination of factors is achieved through the combination of the 200 SS level of Factor 1 and 7-interval level of Factor 2.

As the improvement of BC Modelling performance has been conducted on the decision variables of the process itself, in this case annual normalizable runoff time series, the optimal combination of levels achieved may also be effective for other temporal scales such as seasonal and/or monthly, normalizable ones as well. Furthermore, the fact that the case studies show opposite, and exemplary temporal behaviors indicates that the conclusions of this research can be extrapolated to other river basins.

On the other hand, the optimal combination of factor levels will allow a better understanding of statistical evidence propagation between the model variables. This will undoubtedly make possible to predict future events and analyze the temporal and spatio-temporal behaviour of water resources in a river basin in a more reliably way.

Furthermore, the optimization of the BC Modelling will enable progressing in future work on the DMG by incorporating uncertainty bands and developing new indicators of temporal and even spatio-temporal behaviour, these latter addressed by means of multivariate analysis.

Additionally, the framework developed might be applied to both flow classification and prediction. It provides an in-depth knowledge of the historical flow patterns, as an alternative to standard statistical analyses, and maybe automatized (subject to data requirements) to classify river reaches according to how the conditioned and the unconditioned fraction behave and interact. In addition, BC Modelling might be used to predict streamflows in an alternative way compared to traditional hydrometeorological forecasting methods (e.g. Ensemble Streamflow Prediction, use of stochastic models, use of meteorological forecasts to force a hydrological model). In fact, it is one of the most promising applications of the methodology proposed.

Finally, the methodological framework may serve as a basis for further advances in the field of classification applied to Bayesian PGMs.

Credit authorship contribution statement

Santiago Zazo: Conceptualization, Methodology, Formal analysis, Visualization and leadership of Investigation. Ana María Martín: Methodology, Software, Formal analysis, Investigation. Jose-Luis Molina: Conceptualization, Methodology, Supervision, Validation. Hector Macian-Sorribes and Manuel Pulido-Velázquez: Supervision, Validation. The Discussion and Conclusions sections were addressed by all authors, and all authors wrote the paper.

Acknowledgments

The authors would like to thank Salamanca University for the funding given to this study through the "Call for Concept Testing and Results Protection of the University of Salamanca. TCUE PLAN 2018-2020".

References

Ahmadi, A., Moridi, A., Han, D. (2015). "Uncertainty Assessment in Environmental Risk through Bayesian Networks". *Journal of Environmental Informatics*. 25, 46-59.

Borgomeo, E., Farmer, C. L., Hall, J. W. (2015). Numerical rivers: A synthetic streamflow generator for water resources vulnerability assessments. *Water Resources Research*. 51 (7), 5382–5405.

Bradley, A. A., Habib, M., Schwartz, S. S. (2015). "Climate index weighting of ensemble streamflow forecasts using a simple Bayesian approach". *Water Resources Research*, 51(9), 7382-7400.

Cabañas de Paz, R. (2017). New Methods and Data Structures for Evaluating Influence Diagrams. Ph.D. Thesis, University of Granada, Granada, Spain, 334 pp.

Carmona, G., Varela-Ortega, C., Bromley, J. (2011). "The Use of Participatory Object-Oriented Bayesian Networks and Agro-Economic Models for Groundwater Management in Spain". *Water Resour.Manage*. 25, 1509-1524.

Diez, F.J., Luque, M., Bermejo, I. (2018). "Decision analysis networks". *International Journal of Approximate Reasoning*. 96, 1-17.

Efron, B., Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Monographs on statistics and applied probability. Chapman and Hall/CRC.

Engelke, S., Hitz, A.S., (2020). "Graphical models for extremes". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 82, 871-932.

Giné-Garriga, R., Requejo, D., Molina, J.-L., Pérez-Foguet, A. (2018). "A novel planning approach for the water, sanitation and hygiene (WaSH) sector: The use of object-oriented bayesian networks". *Environmental modelling & software*, 103, 1-15.

HUGIN. (2021). http://data.biotracer.hugin.com/htmlhelp/index.html (06/04/2021).

Javidian, M. A., Valtorta, M., Jamshidi, P. (2020). "AMP chain graphs: Minimal separators and structure learning algorithms". *Journal of Artificial Intelligence Research*, 69, 419-470.

Kim, S., Parhi, P., Jun, H., Lee, J. (2018). "Evaluation of drought severity with a Bayesian network analysis of multiple drought indices". *Journal of Water Resources Planning and Management*, 144(1), 05017016.

Koller, D., Friedman, N., 2009. *Probabilistic graphical models: principles and techniques*. MIT press. Cambridge, Massachusetts, USA.

Le, T. D., Hoang, T., Li, J., Liu, L., Liu, H., Hu, S. (2016). "A fast PC algorithm for high dimensional causal discovery with multi-core PCs". *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5), 1483-1495.

Lehikoinen, A., Olsson, J., Bergström, L., Bergström, U., Bryhn, A., Fredriksson, R., Uusitaloc, L. (2019). "Evaluating complex relationships between ecological indicators and environmental factors in the Baltic Sea: A machine learning approach". *Ecol.Ind*. 101, 117-125.

Macian-Sorribes, H., Molina, J.L., Zazo, S., Pulido-Velázquez, M. (2020). "Analysis of spatio-temporal dependence of inflow time series through Bayesian causal modelling". *Journal of Hydrology*. 125722.

Madsen, A.L., Lang, M., Kjaerulff, U.B., Jensen, F. (2003). "The Hugin Tool for learning Bayesian networks". (Contributed Paper). Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Proceeding. 2711, 594-605.

Malekmohammadi, B., Kerachian, R., Zahraie, B. (2009). "Developing monthly operating rules for a cascade system of reservoirs: Application of Bayesian Networks". *Environmental Modelling & Software*. 24, 1420-1432.

Malone, B., Yuan, C. (2012). "A Bounded Error, Anytime, Parallel Algorithm for Exact Bayesian Network Structure Learning" (Contributed Paper). Proceedings of the Sixth European Workshop on Probabilistic Graphical Models, Granada, Spain, 235-242.

Mamitimin, Y., Feike, T., Doluschitz, R. (2015). "Bayesian Network Modeling to Improve Water Pricing Practices in Northwest China". *Water*. 7, 5617-5637.

MITECO. (2021). Ministerio Para la Transición Ecológica. Gobierno de España. https://sig.mapama.gob.es/redes-seguimiento/ (10/03/2021).

Mohammadi, K., Eslami, H.R., Kahawita, R. (2006). "Parameter Estimation of an ARMA Model for River Flow Forecasting using Goal Programming". *J. Hydrol.*, 331, 293–299.

Molina, J.L., Patino-Alonso, C., Zazo, S. (2021). "Hybrid Causal Multivariate Linear Modelling (H_CMLM) method for the analysis of temporal rivers runoff". *Journal of Hydrology*. 126501.

Molina, J.L., Zazo, S., Martín-Casado, A.M. (2019). "Causal Reasoning: Towards Dynamic Predictive Models for Runoff Temporal Behavior of High Dependence Rivers". *Water*. 11, 877.

Molina, J.L., Zazo, S. (2018). “Assessment of Temporally Conditioned Runoff Fractions in Unregulated Rivers”. *J.Hydrol.Eng*. 23, 04018015.

Molina, J.L., Zazo, S., Rodriguez-Gonzalvez, P., Gonzalez-Aguilera, D. (2016). “Innovative Analysis of Runoff Temporal Behavior through Bayesian Networks”. *Water*. 8, 484.

Molina, J.L., Pulido-Velazquez, D., Garcia-Arostegui, J.L., Pulido-Velazquez, M. (2013). “Dynamic Bayesian Networks as a Decision Support tool for assessing Climate Change impacts on highly stressed groundwater systems”. *Journal of Hydrology*. 479, 113-129.

Nourani, V., Kisi, O., Komasi, M. (2011). “Two hybrid Artificial Intelligence approaches for modeling rainfall-runoff process”. *Journal of Hydrology*. 402, 41-59.

Ochoa-Rivera, J.C, Andreu, J., García-Bartual, R. (2007). “Influence of inflows modeling on management simulation of water resources system”. *J.Water Resour.Plann.Manage*. 133, 106-116.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Francisco, USA.

Pearl, J. (2009). *Causality: models, reasoning, and inference*. 2nd edn Cambridge University Press. New York City, NY, USA.

Ramos Fernández, J. M. (2018). Implementación de los modelos gráficos probabilísticos bayesianos en la ayuda al manejo clínico de la bronquiolitis aguda del lactante. Ph.D. Thesis, University of Malaga, 137 pp.

Salas, J. D., Delleur, J. W., Yevjevich, V. M., and Lane, W. L. (1980). *Applied Modeling of Hydrologic Time Series*. Water Resources Publications, Littleton, Colorado, USA.

Schroeter, K., Kreibich, H., Vogel, K., Riggelsen, C., Scherbaum, F., Merz, B. (2014). “How useful are complex flood damage models?” *Water Resour.Res*. 50, 3378-3395.

Shin, J.Y., Kwon, H., Lee, J.H., Kim, T. (2019). “Probabilistic long-term hydrological drought forecast using Bayesian networks and drought propagation”. *Meteorol.Appl*. 27(1), e1827.

Spirtes, P., Glymour, C., Scheines, R. (2000). Causation, Prediction, and Search. Adaptive Computation and Machine Learning. (2nd edition). MIT Press, Cambridge, Massachusetts, USA.

Spirtes, P. (2010). “Introduction to Causal Inference”. *Journal of Machine Learning Research*. 11, 1643-1662.

Srinivas, V. V., Srinivasan, K. (2006). Hybrid matched-block bootstrap for stochastic simulation of multiseason streamflows. *Journal of Hydrology*. 329 (1–2), 1–15.

Steck, H. (2001). Constraint-based structural learning in Bayesian networks using finite data sets. Ph.D. Thesis, Technische Universität München, Germany, 171 pp.

Schwarz, G. (1978). "Estimating the dimension of a model". *The annals of statistics*, 461-464.

Valipour, M., Banihabib, M.E., Behbahani, S.M.R. (2012). "Parameters Estimate of Autoregressive Moving Average and Autoregressive Integrated Moving Average Models and Compare their Ability for Inflow Forecasting". *J. Math. Stat.*, 8, 330–338.

Verma, T., Pearl, J. (1991). "Equivalence and synthesis of causal models". In Proceedings of the Sixth UAI Conference, UAI '90, pages 255–270, 1991.

Verma, T., Pearl, J. (1992). "An algorithm for deciding if a set of observed independencies has a causal explanation". In Uncertainty in artificial intelligence (pp. 323-330). Morgan Kaufmann.

Vogel, K., Riggelsen, C., Korup, O., Scherbaum, F. (2014). "Bayesian network learning for natural hazard analyses". *Natural Hazards and Earth System Sciences*. 14, 2605-2626.

Vogel, K., Weise, L., Schroeter, K., Thieken, A.H. (2018). "Identifying Driving Factors in Flood-Damaging Processes Using Graphical Models". *Water Resour.Res*. 54, 8864-8889.

Xue, J., Gui, D., Lei, J., Zeng, F., Mao, D., Zhang, Z. (2017). "Model development of a participatory Bayesian network for coupling ecosystem services into integrated water resources management". *Journal of Hydrology*. 554, 50-65.

Zazo, S., Molina, J.L., Ruiz-Ortiz, V., Vélez-Nicolás, M., García-López, S. (2020). "Modeling River Runoff Temporal Behavior through a Hybrid Causal–Hydrological (HCH) Method". *Water*. 12, 3137.

Zazo S., Macian-Sorribes H., Sena-Fael C.M., Martín-Casado, A.M., Molina J.L. and Pulido-Velazquez, M. (2019). "Qualitative approach for assessing runoff temporal dependence through geometrical symmetry". (Contributed Paper). Proceedings Internacional Congress on Engineering. Engineering for Evolution (ICEUBI2019). 27-29 November 2019, Covilhã, Portugal.

Zazo, S. (2017). Analysis of the Hydrodynamic Fluvial Behaviour through Causal Reasoning and Artificial Vision. Ph.D. Thesis, University of Salamanca, Ávila, Spain, 159 pp.