Data-driven model for river flood forecasting based on a Bayesian network approach

Brahim Boutkhamouine, Hélène Roux, François Pérès

HAL Id: hal-03164840

https://hal.science/hal-03164840

Submitted on 10 Mar 2021

Data-driven model for river flood forecasting based on a Bayesian network approach

Brahim Boutkhamouine[1,2] | **Hélène Roux[1]** | **François Pérés[2]** (iD)

[1]Institut de Mécanique des Fluides de Toulouse (IMFT), Université de Toulouse, CNRS-INPT-UPS, Toulouse, France

[2]Laboratoire Génie de Production (LGP), Université de Toulouse, INP-ENIT, Tarbes, France

Correspondence
François Peres, Institut de Mécanique des Fluides de Toulouse (IMFT), Université de Toulouse, CNRS-INPT-UPS, Toulouse, France.
Email: francois.peres@enit.fr

Abstract

Uncertainty analysis of hydrological models often requires a large number of model runs, which can be time consuming and computationally intensive. In order to reduce the number of runs required for uncertainty prediction, Bayesian networks (BNs) are used to graphically represent conditional probability dependence between the set of variables characterizing a flood event. Bayesian networks (BNs) are relevant due to their capacity to handle uncertainty, combine statistical data and expertise and introduce evidences in real-time flood forecasting. In the present study, a runoff–runoff model is considered. The discharge at a gauging station located is estimated at the outlet of a basin catchment based on discharge measurements at the gauging stations upstream. The BN model shows good performances in estimating the discharges at the basin outlet. Another application of the BN model is to be used as a reverse method. Knowing discharges values at the outlet of the basin, we can propagate back these values through the model to estimate discharges at upstream stations. This turns out to be a practical method to fill the missing data in streamflow records which are critical to the sustainable management of water and the development of hydrological models.

KEYWORDS
flood risk, missing Data, uncertainty prediction

1 | INTRODUCTION

To deal with flood risk, a resistance-based strategy has been often adopted. It consists of building flood control structures, such as levees, channels and dams, in order to control flood threats (Morrison et al., 2017). Although these standard approaches help to reduce the risk of flooding, they do not provide any fully effective guarantee against these feared events. Flood defence infrastructures and regular controls may fail to provide adequate protection against flash or huge flood events, which can result in property damage and even human casualties (Park, 2013, Ogie et al., 2020). Alongside these defence approaches, adaptive methods seek to adjust to changes and deal with flood uncertainty (Pahl-Wostl, 2007, Pant et al., 2018). Societies have to manage and live with flood risk. Real-time forecasts, warning and well-functioning emergency plans help in saving both lives and livelihoods.

Within the framework of the implementation of flood crisis management plans in mountainous sectors, it turns out to be essential more than anywhere to measure the vulnerability of the stakes (human and material) to a rapid rise in water levels observed upstream of the areas to be protected (Fuchs et al., 2019), (Klein, 2019), and (Terzi, 2019). The assessment of this vulnerability, and thus of the solutions to be implemented to reduce it, requires the answer to various questions: How much time do we have to react? Are flood trends predictable? Will we have time to travel to intervene? What is the possible duration of the flood, its probability of occurrence (in the area under observation)?

Anticipation of the phenomenon is therefore crucial. It must be based on information extracted upstream relating to the flows observed on the main river but also on the river's inflows, which also contribute to the river's swelling. Calculation of the risk inherent in

this phenomenon requires, on the one hand, knowledge of the uncertainties linked to the model input data (random uncertainties on rainfall, epistemic uncertainties on flows, measurement uncertainties on flows or water heights...) and, on the other hand, knowledge of how they propagate to the outlet. Analytical models based on hydrological calculations and physical behavioural models are mostly deterministic and have difficulty in establishing the confidence intervals necessary for rational decision-making. This is especially the case for flow records that are established under the assumption of uniform flow when the energy profile is parallel to the water line and the riverbed bottom.

In some cases, however, the energy slope of a channel control may be variable, usually due to a variable downstream boundary condition ("variable turbulence"), or sometimes due to transient flow conditions ("variable flow") (Mansanarez et al., 2017, Ferguson et al., 2019, Bhola et al., 2019).

In this context, the use of frequentist approaches may be an alternative. These methods are independent of the complexity of the system under study, but require databases that are sufficiently representative to provide a trustworthy result. If this condition is met, the use of models based on Bayesian approaches makes sense (Garrote et al., 2007, Yassine, Pérès, Roux, Cassan, & Frysou, 2018, Yassine, Roux, et al., 2018). With the aim of assessing the flow discharge at a given location, they can take into account individual gauge uncertainties in the form of probabilistic distributions. More importantly, they allow the decomposition of uncertainty components related to the model of the gauging curve ("structural error"), parameter estimation ("parametric error") and potentially systematic and non-systematic errors in the water level series ("propagation error") (Boutkhamouine, Roux, & Pérès, 2017, Boutkhamouine, Roux, Pérès, & Vervoort, 2018, Jiang et al., , 2019, Fan, 2019; Leandro, 2019, Dittes, Špačková, & Straub, 2019).

The objective of this paper is to establish such a model. The rest of the paper is divided into 3 parts. Through a brief state of the art of the domain under study confronting the need for operational tools in the field of flood assessment and existing modelling techniques, the scientific issue of the work is outlined. In a second step, the Bayesian network tool adopted for the proposed modelling is described from the point of view of its operating mode and capabilities. The construction principles of the model are then discussed and deployed on a case study described beforehand. A last part is devoted to the analysis of the results obtained, the presentation of possible modes of validation of these results and a discussion on the different alternatives of model using. A final conclusion summarizes the objectives, the approach and the achieved results.

2 | CONTEXT AND ISSUE

The main attractions of hydraulic models, which simulate flood phenomena, are to numerically simulate and predict the spatial and temporal evolution of a river's hydraulic characteristics during a

flood, such as water levels, flow velocities and submergence durations (Molinari, 2019). Hydraulic models are most often deterministic and remain a simplified representation of a physical process to which they are related. However, confidence in the results obtained is limited by the level of uncertainty arising from the structure of the model itself (the natural system is too complex to be modelled using simplified equations) as well as from model parameterization or forcing (spatio-temporal precipitation, initial conditions). Quantifying these uncertainties, in terms of probability distribution and degree of certitude, is of primary importance for risk assessment and decision-making, (Biondi, Versace, & Sirangelo, 2010, Annis, 2020, Tscheikner-Gratl et al., , 2019). Taking into account of these uncertainties would allow the establishment of models for purpose of risk assessment and context-oriented decision-making related to damage prevention and crisis management. Several authors have emphasized the importance of qualifying and quantifying the effect of uncertainties on hydraulic model outputs in order to support the decision-making process (Sieg, 2019). Uncertainty in hydraulic modelling can arise from a variety of sources:

- Topographic data. Topographic and hydrological errors are considered to be the major sources of uncertainty in flood hazard modelling (Casas, 2006);
- Upstream and downstream model boundary conditions, initial conditions (Stephens & Bledsoe, 2020);
- Roughness coefficients (Lim & Anders Brandt, 2019);
- Model calibration–validation data (Ahmadisharaf, 2019);
- Uncertainties related to the structure of the model (Liu & Merwade, 2019)

The level of confidence in the results associated with a decision will be all the better as it is established in knowledge of the probability distributions associated with the different variables involved in the decision to be made. In recent years, uncertainty quantification and reduction have been subject of many scientific breakthroughs. (Matott, Babendreie, & Purucker, 2009, Han & Coulibaly, 2017, Coulibaly & Evora, 2007, Ronalds & Zhang, 2019). In the field of flood estimation, however, most of the approaches rely, directly or indirectly, on Monte Carlo simulations.

The Monte Carlo method is a well-known technique based on the use of random numbers to simulate deterministic systems with stochastic parameters or inputs. It consists of running the model a large number of times using random sampling from the input parameter distributions until sufficient output values have been obtained. If hydrological model is complex or the number of uncertain parameters is high, producing forecasts incorporating uncertainties—that is forecasts in the form of probability distributions—becomes time consuming and computationally expensive.

Yet, the dynamic management of flood-type events requires tools capable of reacting quickly to the input of new information in order to contextualize decision-making with regard to the state of the descriptive variables of the analysed situation. In particular, in small catchment characterized by fast response time, and where flash flood are

often destructive, there is a need for real-time flood forecasts (Ruin, Creutin, Anquetin, & Lutoff, 2008, Roux, 2011; Si, 2019).

In this context of dynamic decision-making, the use of Bayesian networks (BNs) may be a relevant solution (Liu, Pérès, & Tchangani 2016). Since their introduction (Pearl, 1985), Bayesian networks have become indeed a highly popular tool in artificial intelligence to handle uncertainties. In the field of flood prediction, Bayesian networks have many advantages: They allow both quantitative and qualitative factors to be taken into account, which most models do not; they enable causal links between variables to be visualized: risk aggregation is achieved by the very construction of the networks, which avoids the estimation of correlations; they facilitate reduction factors to be detected through inference; and they can be used for diagnostic purposes to search for causes in the presence of a result obtained (upward propagation) or for prognostic purposes to identify the consequences of introducing new information (downward propagation) (Uusitalo, 2007).

In this study, we focus on a runoff–runoff model. The aim of the model is to estimate the discharge at a gauging station based on discharge measurements of proxies gauging stations. In other words, we seek to establish the probabilistic relationship between the water flow from an outlet and the discharge from the upstream stations located on the main river or on its tributaries. To do so, we rely on the use of inference techniques supported by Bayesian networks applied to the case of a catchment area located in the Pyrenees chain in the south of France. Thereafter some theoretical elements will be introduced to explain the principles of modelling and calculation inherent to Bayesian networks. The model will then be instantiated to the selected case study.

3 | MODEL BUILDING

3.1 | Bayesian Networks

A Bayesian network is a mathematical model that graphically represents conditional probabilistic dependencies between variables. A Bayesian network B={G,P} is a combination of both graph and probability theories. It is defined as follows: $G = \{X, E\}$, a directed acyclic graph, without circuits, consisting of nodes and arcs linking these nodes (Liu et al., 2016).

The nodes are associated with a set of random variables, $X = \{X_1 \cdots X_n\}$, explaining the studied phenomenon. The directed arcs E, represent the set of causal relationships between the variables. Each node, X_i, in the graph is associated with a conditional probability, $P = \{P(X_i/\text{Pa}(X_i))\}$ expressing the effect of the variables, Pa(X_i) called parents of X_i that cause X_i in G. P are also called local probabilities, and they express the "dependencies strengths" between the nodes.

In practice, P are defined in conditional probability tables (CPT$_s$). The way G is shaped, that is a directed acyclic graph, simplifies the computation of both joint and marginal probabilities of the nodes, X:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i/\text{Pa}(X_i)) \text{ and } P(X_i = x_i) = \sum_{X_{j(j \neq i)}} P(X_i = x_i/X_j) \quad (1)$$

As shown in Equation 1, the computation of $P(X_1 ..., X_n)$ comes down to a product of local probabilities terms directly accessible from CPT$_s$ associated with the network. This operation is called inference. Using the same technique, one can also compute marginal probabilities associated with a node, or a set of nodes, after having introduced some observations on one or various nodes in the graph.

This combination between the theory of graphs and probability is one of the powerful aspects of Bayesian networks providing an efficient way to perform inference. Algorithms dealing with inference in BNs are explained in details in (Lauritzen & Speigelhalter, 1988, Jensen, 1996, Liu, Tchangani, & Pérès, 2016).

The construction of a Bayesian network mainly involves the following steps:

- Structure learning: determine the factor variables (nodes) related to the study object, and then determine the dependent or independent relationships between the nodes so as to construct a directed acyclic network structure.
- Parameter learning: based on a given Bayesian network structure, learn the conditional probability table (CPT) associated to each node.

There are two approaches to specify the structure of the causal graph. The first approach consists in learning the causality directly from data, if they exist. However, resulting causalities may not be consistent with reality due to a lack of representative data. The second approach uses expert's opinion to specify the structure. The BN structure is directly inspired from the hydrographic network of the studied basin. Concerning the conditional probability tables, CPT_s, they are learned from historical discharge data described hereunder.

3.2 | Case description

3.2.1 | Study area

Salat river basin, located in southern France, is an upstream tributary of the Garonne river (Figure 1). It takes it sources from the north slope of Mont Rouch. After a 75km stretch, particularly through the regional natural park of the Pyrenees Ariégeoises, it flows into the Garonne river at Roquefort-sur-Garonne (basin's outlet). Its main tributaries are Garbet and Arac on the right bank and Lez on the left one. Salat river basin has an area of nearly 1570km^2. The slope is important with an altitude which varies from 2,900 m in the upstream zone (Mont Rouch) to 300 m in the downstream part (latitude: 43°09' 52'' N, 00° 58'26''E).

The basin is marked by a strong nival regime due to the melting snow in the spring (Boutkhamouine et al., 2018). It is also subject to an oceanic influence causing heavy rain and storms during April to June. A record snowfall followed by a rapid thaw may cause flash flooding especially when combined with heavy rains. Recently, such case occurred on 17th June 2013 in the region causing victims and extensive property damages amounting to tens of millions of euros.
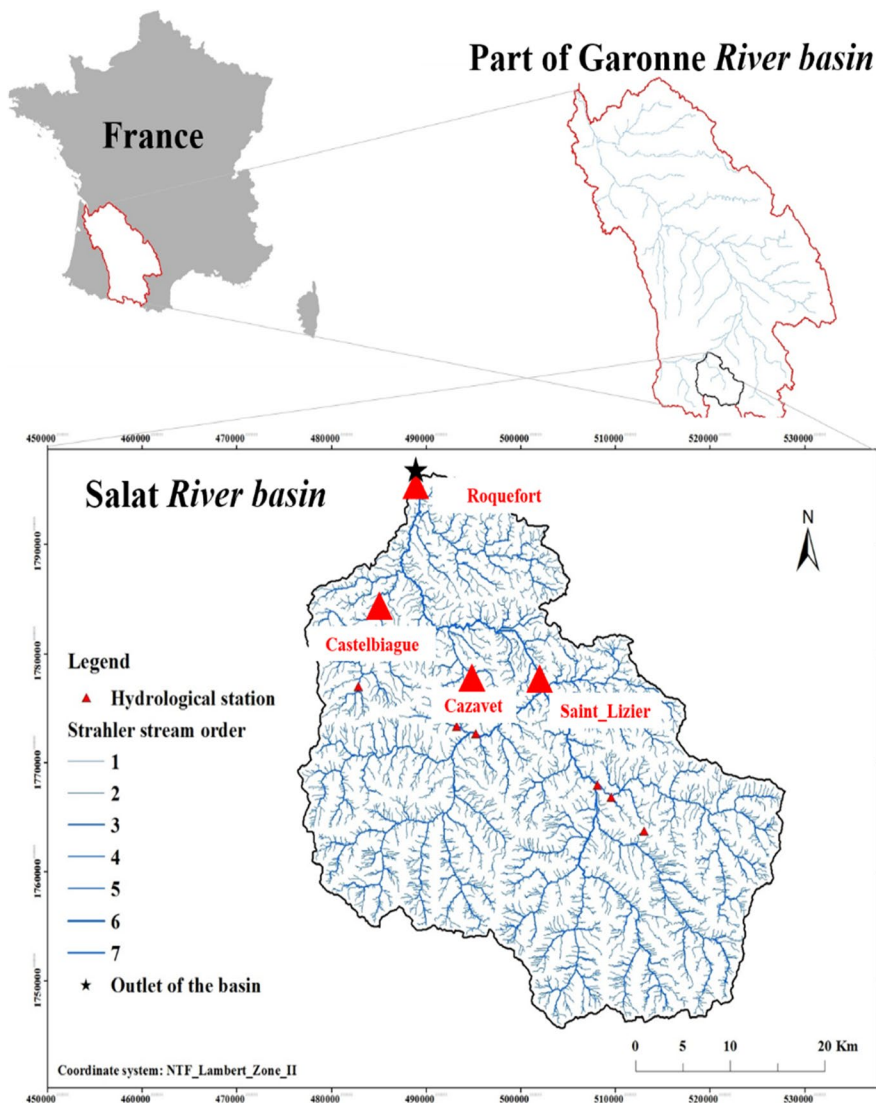
FIGURE 1 Map of the Salat river basin [Colour figure can be viewed at wileyonlinelibrary.com]

3.2.2 | Data

Currently, the Salat basin is controlled by the means of many hydrometric stations measuring the streamflow on different main tributaries of the basin. The historical measurements are available on the French national hydrometric database (HYDRO, http://hydro.eaufrance.fr) provided by SCHAPI (Service Central d'Hydrométéorologie et d'Appui à la Prevision des INondations/Central Service for Hydrometeorology and Flood Prediction Support). In this study, the historical measurements of four stations, Roquefort, Castelbliaque, Cazavet and Saint-Lizier, are used to test the presented Bayesian networks methodology. These stations are located at the downstream of the Salat basin (see Figure 1). The historical measurements downloaded for this study are hourly discharge series from April 2005 to January 2017 in the four stations (Figure 2).

From Figure 2b, one can clearly see a delay between the discharge at the station Roquefort (basin's outlet) and discharges at the upstream stations (Saint-Lizier, Cazavet and Castelbiague). Such time delays correspond to the flow transfer duration between the upstream stations and the outlet of the basin. Flow transfer time, called also lag time, corresponds to the time needed by the flow to move from any location within the river basin to another, particularly, from one upstream hydrological station to another one downstream. This should be estimated and incorporated in the data before training the Bayesian Network CPTs.

3.2.3 | Estimation of lag time between the upstream stations and Roquefort using observed data

Estimating the time of flow transfer between two hydrological stations is a high complex task. In fact, it depends on many parameters such as the river slope, the river bottom roughness, the length of the flow path and the flow depth. (Green & Nelson, 2002, McCuen, 2009). Hereunder, we estimate the flow time transfer between each upstream station and Roquefort using a simple statistical method applied to the observed data of Figure 2. This method is similar to the one described in (Seyam & Othman, 2014). Given a flood specific discharge, q, at the station Roquefort, the idea is to look through the

FIGURE 2 Historical discharges measured in the upstream hydrological stations of Salat. (a) Plot of hourly streamflow in different stations from April 2005 to January 2017. (b) Random zoom on a particular period, from 10th June to 14th June 2008 [Colour figure can be viewed at wileyonlinelibrary.com]
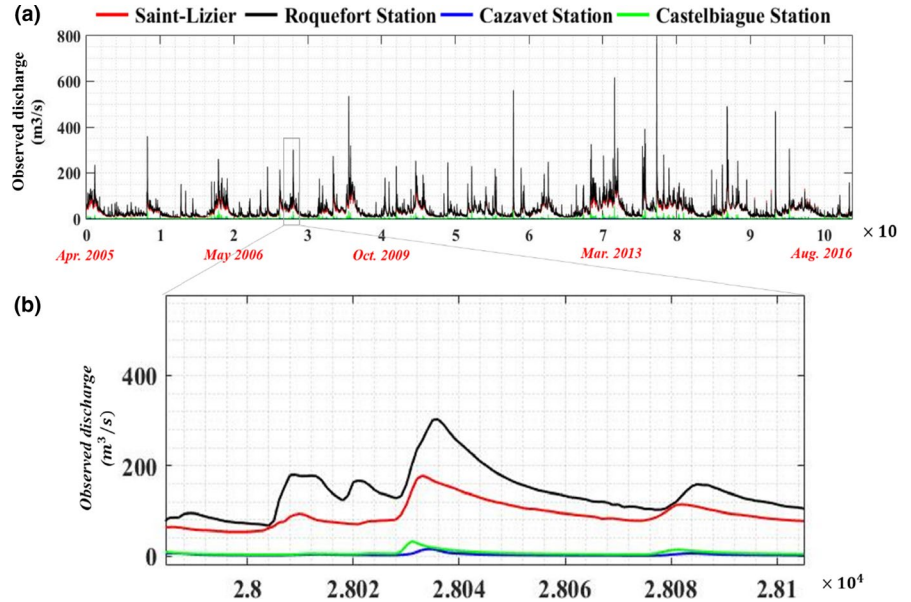

data and extract all Roquefort discharge peaks above this threshold. Then, for each peak identify all discharge peaks of upstream stations causing this peak. In this study, we have chosen for threshold a flood specific discharge $q = 0.1 m^3/s/Km^2$ which corresponds approximatively to a flood discharge $Q = 0.1 * 1570 \sim 150 m^3/s$ at Roquefort station. Figure 3 shows the estimated cumulative distribution functions of lag times for each upstream station and Roquefort. T_{10}, T_{50} and T_{90} denote, respectively, the 10th, 50th and 90th percentiles of the distribution.

3.2.4 | Data pre-processing

As mentioned previously, the parameters of BN model are learned from historical discharge data. Nevertheless, before starting any learning process, the data have to be prepared. The performances of

the learning process depend on the quality of data used for learning but also on the format as well as on the features included such as time delay here. A twofold procedure has been used to prepare the learning data (Figure 4).

1. Formatting the data: All hourly discharge data for the 4 stations are stored in a text file, column by column (each column corresponds to a station). Each line within the text file denotes a case (discharge value at different stations). In order to take in consideration, the time transfer between the upstream stations (Castelbiague, Cazavet and Saint-Lizier) and Roquefort, we have chosen to displace the columns associated with upstreaming stations, compared to Roquefort one, using the corresponding T50 of calculated time transfer (Figure 3), which are approximately 3 hr for Castelbiague, 2 hr for Cazavet and finally 3 hr for Saint-Lizier.
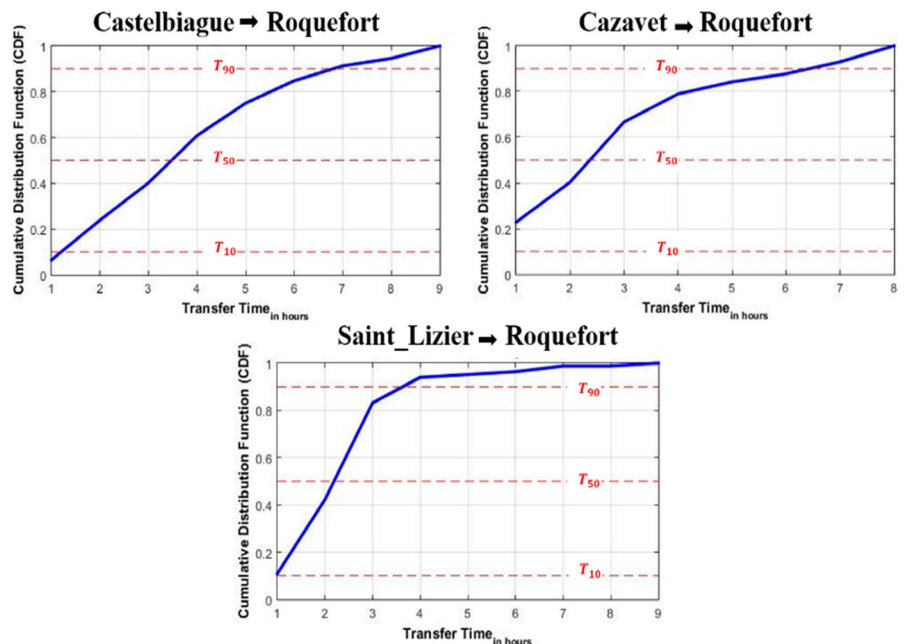
FIGURE 3 Flow time transfer calculated for each upstream station from the data presented in Figure 2 using a specific discharge $q = 0.1 m^3/s/km^2$ at Roquefort station [Colour figure can be viewed at wileyonlinelibrary.com]

Raw data

Cases /stations	Saint-Lizier	Cazavet	Castelbiague	Roquefort
01/04/2005 00:00	52,10	0,55	2,05	65,80
01/04/2005 01:00	51,80	0,54	2,04	65,70
01/04/2005 02:00	51,1	0,52	2,03	63,40
01/04/2005 03:00	51,80	0,51	2,02	62,80
.	.	.	.	.
.	.	.	.	.
01/01/2017 21:00	32,10	1,50	1,51	29,30
01/01/2017 22:00	32,50	1,54	1,51	35,40
01/01/2017 23:00	31,50	1,55	1,51	35,80

Insert time delay, $T_{50} \approx 3\,h$, $T_{50} \approx 2\,h$, $T_{50} \approx 3\,h$

Treated data incorporating time delay between upstream stations and Roquefort

Cases /stations	Saint-Lizier	Cazavet	Castelbiague	Roquefort
01/04/2005 03:00	52,10	0,52	2,05	62,80
01/04/2005 04:00	51,80	0,51	2,04	62,80
01/04/2005 05:00	51,1	0,50	2,03	62,80
01/04/2005 06:00	51,80	0,49	2,02	60,10
.	.	.	.	.
.	.	.	.	.
01/01/2017 21:00	32,10	1,26	1,54	29,30
01/01/2017 22:00	32,30	1,40	1,52	35,40
01/01/2017 23:00	32,30	1,50	1,51	35,80

2. Sampling: At the end of first step, 103,772 cases were stored. The first 93,772 cases, from April 2005 to 6th December 2015, are reserved for learning the BN CPTs, and the remaining 10,000 cases, from 7th December 2015 to January 2017, are dedicated to the model validation. It should be noted that the choice of learning and validation period is completely arbitrary and the two periods are totally independent.

3.3 | Model structure and parameter learning

3.3.1 | Model structure

As previously said, the BN structure used in this study is created from the hydrographic network of Salat basin displayed in Figure 1. Given the geographical position of the studied hydrological stations, the discharge at Roquefort station can be predicted using the flows at three upstream stations, namely Castelbiague, Cazavet and Saint-Lizier (Figure 5). These three3 stations are thus considered parents of Roquefort. It should be noted that this assumption neglects other possible water intakes like groundwater contributions and potential runoffs on the immediate surfaces between the three upstream stations and basin's outlet.

3.3.2 | Discretization and parameter learning

The model was built using the software package Netica Java (Norsys, 2015). It is shown in Figure 6. The associated development code, under Eclipse IDE, is showed in Appendix A.

The building of Bayesian network requires to follow the following steps:
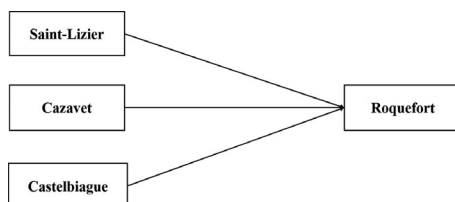
FIGURE 5 Structure of the Bayesian network inspired from the hydrographic network presented in Figure 1 [Colour figure can be viewed at wileyonlinelibrary.com]

1. Create an empty BN.
2. Add nodes to the network.
3. Set the causal structure of the network: set the arcs between the nodes.
4. Discretize the nodes.
5. Learn BN CPT s from data using a parameter learning algorithm.
6. Save the BN into a file.

Even if the variables (discharges) in the model are continuous, they have to be discretized. Indeed, Netica as most of the software-based Bayesian network does not support the introduction of continuous variables. The discretization principles are simple. It consists of scanning the historical hourly discharge data, presented above, in order to identify all minimum and maximum values of all observed discharges at each hydrological station (node), and divide the continuous nodes into a number of intervals. Ideally, only a few intervals should be associated with each node. This makes it possible on the one hand to reduce the size of the BN in order to optimize its temporal inference calculation but also, on the other hand, to lower the prediction uncertainty of the model. Here, we seeked to estimate the water discharge at "Roquefort" using the discharges at the hydrological upstream stations (Figure 5). Hourly hydrograph for each station can be split into two parts: Lower values of discharge, which are very numerous, require shorter intervals to gain enough precision, and for the remaining ranges, which corresponds to the flood discharges, the discretization intervals are further apart.

After determining the topological structure of Bayesian network and the discretization associated with each node, the parameter learning of the network can be performed using the data of hourly discharge for the four stations described above. The Counting–Learning algorithm available on Netica was used to learn parameters, prior and conditional probabilities, of all the nodes in the model. The counting algorithm uses a traditional one-pass method to determine the probabilities, which essentially amounts to counting the number of times a node takes on a certain value given each configuration of the parents (Russell & Norvig, 2010). The learning process was performed using the training dataset containing 93,772 cases (see data pre-processing section above). Once the parameter values were learned, the model could be compiled and was ready for use (Figure 6).
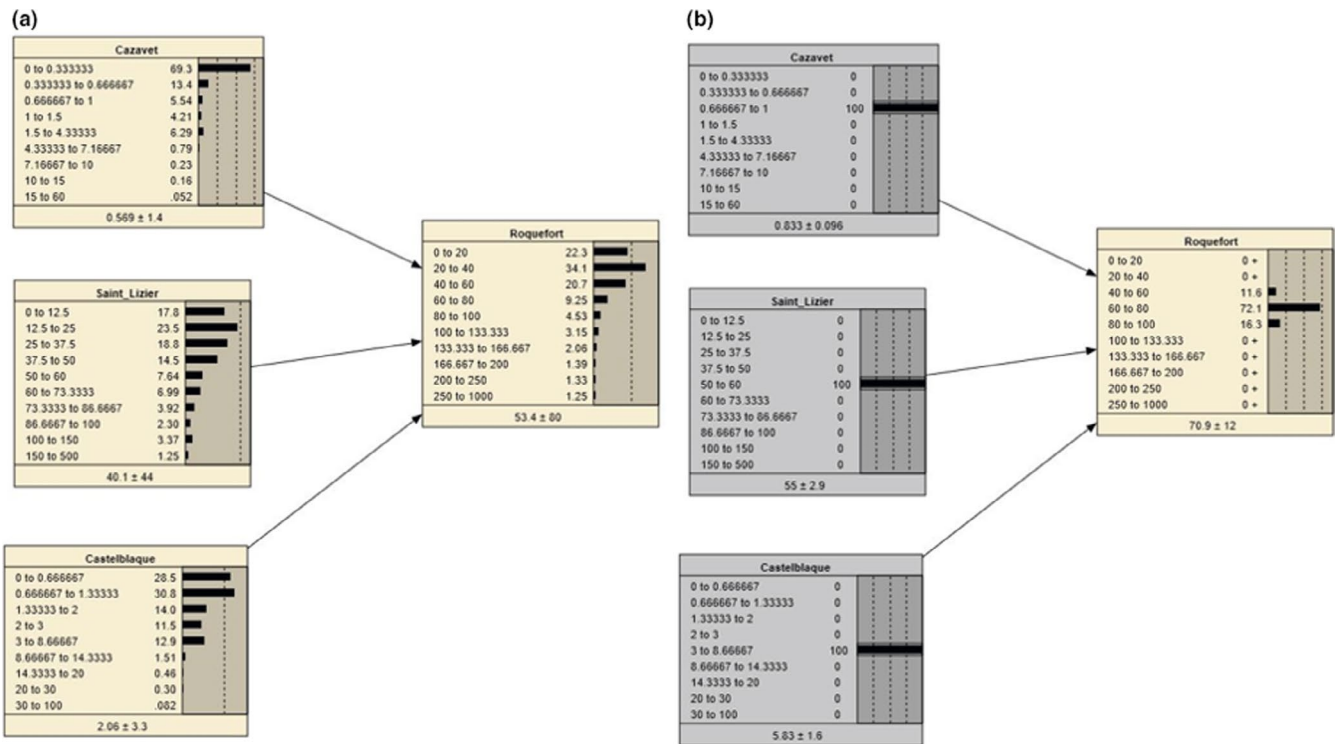
FIGURE 6 A Bayesian network modelling the discharge at the Roquefort as function of the information coming from the upstream stations. (a) Compiled model based on a priori knowledge. (b) Model state when the model is inferring the observation flows of the 3 upstream stations. Grey boxes indicate that observations were entered in corresponding nodes; the model uses the available new information to update the state of unobserved node (here Roquefort node) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 7 Sensitivity of Roquefort station to findings at upstream nodes using variance reduction method implemented in Netica [Colour figure can be viewed at wileyonlinelibrary.com]
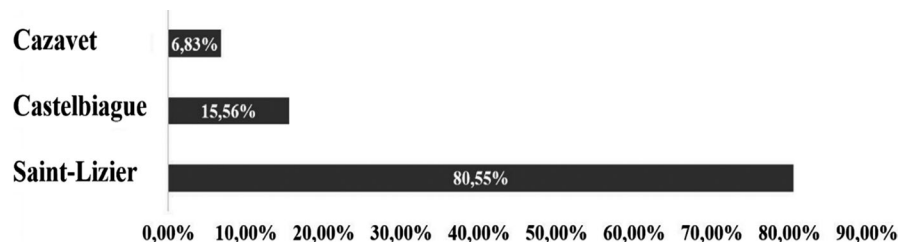
4 | RESULTS AND DISCUSSION

4.1 | Sensitivity analysis

After compiling the model, a sensitivity analysis was performed using the variance reduction method (Marcot, 2012; Marcot, 2012). The variance reduction estimates the impact of a change in the state of a node on the state of a target one. Its ranges from 0% to 100%, and higher value indicates a higher influence. Here, the target node is Roquefort (i.e. the discharge at the Roquefort station). Figure 7 summarizes sensitivity analysis results of Roquefort node using the variance reduction method.

As expected, Saint-Lizier has a higher influence on the Roquefort station. This is obviously logical given the hydrographic network presented in Figure 1. Saint-Lizier station is located at the main Salat's river supplying the outlet (Roquefort). The effects of other stations are relatively low given their locations on secondary tributaries.

4.2 | Model validation

As we explained previously, we have estimated the BN CPTs using 90% of data. The remaining 10%—10,000 cases—of the discharge observations at the 4 stations, from 07/12/2015 to 31/01/2017, were withheld to test the model and evaluate its prediction accuracy. In terms of platform, the Netica-J, the Java version of Netica (Norsys, 2015; Norsys, 2015) was used to evaluate the prediction accuracy of a model. The steps followed to test the model are listed in the frame below. The associated development code, under Eclipse IDE, is given in Appendix B.

1. Read a BN version from the model database.
2. Get the nodes of the BN.
3. Test the BN performances: perform sensitivity analysis.
4. Use the model to perform inference.

    For each time step:

TABLE 1 Confusion matrix comparing BN-estimated and observed discharge values at Roquefort

n = 10,000 cases	Predicted discharge intervals (m3/s)										
Observed discharge intervals		[0, 20]	[20, 40]	[40, 60]	[60, 80]	[80, 100]	[100, 133.3]	[133.3, 166.6]	[166.6, 200]	[200, 250]	[250, 1000]	Σ cases
	[0, 20]	3,414	395	0	0	0	0	0	0	0	0	3,809
	[20, 40]	300	2,074	314	2	0	0	0	0	0	0	2,690
	[40, 60]	1	153	2,481	85	0	0	0	0	0	0	2,720
	[60, 80]	0	0	172	430	6	2	0	0	0	0	610
	[80, 100]	0	0	0	36	44	7	2	0	0	0	89
	[100, 133.3]	0	0	0	2	0	26	6	0	0	0	44
	[133.3, 166.6]	0	0	0	1	0	0	23	0	0	0	24
	[166.6, 200]	0	0	0	0	0	4	4	0	1	0	9
	[200, 250]	0	0	0	0	0	0	0	8	8	0	8
	[250, 1000]	0	0	0	0	0	0	0	1	0	6	7

I Read values of known nodes from the validation data file and set them as observations in the model (evidences).

II Compile the model with these observations to get the posterior probabilities of unknown nodes.

III Save the predicted posterior probabilities into a result file to be compared with actual observation for each time step.

5. Close the BN model.

During the testing process, the values of Roquefort node were considered as unknown. The evaluation was conducted by introducing the cases into the validation data file one by one (i.e. line by line). For each case, Netica-J read in the values of nodes, excepted the value for Roquefort. Next, the model was compiled and a probability distribution of Roquefort node was generated by inferring these values in the case study. The prediction at Roquefort node was then compared with the observation value supplied in the data file. Several different measures of prediction accuracy were tested.

One way to estimate the model accuracy prediction is to use the confusion matrix, which compares predicted to observation outcomes (Table 1). For example, in this table, for the all 610 cases in which the observations discharge at Roquefort station are included in the interval $[60, 80]$m^3/s, the model predicted well 430 cases (i.e. predicated values included in $[60, 80]$m^3/s), 172 cases in $[40, 60]$m^3/s, 6 cases in $[80, 100]$ and 2 cases in $[100, 133.3]$m^3/s.

The overall model accuracy rate can be estimated as the number of cases where the model correctly predicted the actual discharge value (observation discharges) over all tested cases (here 10,000 cases). From Table 1, an accuracy rate of 85.10% was calculated. Instead of using the most likely state as a prediction for each case (confusion matrix), one can use the actual belief levels of each state to assess the model's prediction accuracy. Some standard scoring rules evaluating the model accuracy like logarithmic loss, spherical payoff or quadratic loss do not merely take the most likely state as a prediction but rather considers the actual posterior probabilities (belief levels) of each state to determine how well the predictions agree with the observations in the validation dataset. These scores are well explained in (Morgan 1990, Stehman, 1997). Spherical payoff, perhaps the most useful index, varies between 0 and 1, (1 indicating the best possible model performance). Logarithmic loss score ranges between 0 and ∞, (smaller values suggest better performance). Regarding the quadratic loss, it varies from 0 to 2, (0 indicating the best performance). In the case under study, the calculated values of spherical payoff, logarithmic loss and quadratic loss were 0.8796, 0.3736 and 0.2154, respectively. These values can be interpreted as an indication of the model's strong predictive capability.

To find out more about the prediction capacity of the model to reproduce different discharge regimes at Roquefort station, the uncertainty interval was assessed, using quantiles Q_{20} and Q_{80} (i.e. Roquefort node states associated to probability of non-exceedance 0.2 and 0.8, respectively) and the predicted mean discharge, Q_{mean},

from the predicted probability distribution of each tested case. Q_{mean}, (also called expected discharge value), is calculated as follows:

$$Q_{mean} = \sum_{1}^{n} p_i \cdot q_i \qquad (2)$$

where q_i designs a possible state of the node "Roquefort," p_i stands for the corresponding predicted probability (i.e. the posterior probability of Roquefort node associated with the state q_i after inferring the observations of upstream stations through the model), and n denotes the number of states of Roquefort node. In Figure 8, the predicted uncertainty interval, $Q_{80} - Q_{20}$ are plotted, calculated from the 10 000 tested cases, together with predicted Q_{mean} and observed discharges at Roquefort station.

Figures 8 and 9 highlight the consistent performance of the model. BN-estimated discharges properly follow the observations both in low and high flows. Over the entire testing period—10,000 cases—69,97% of observed discharge points are counted inside the predicted uncertainty interval $Q_{80} - Q_{20}$.

A comparison between predicted Q_{mean} and observed discharges is done using the classic Nash–Sutcliffe (NS) efficiency criteria (Nash & Sutcliffe, 1970). NS ranges between 1.0 (perfect fit) and −∞. An NS lower than zero indicates that the mean value of the observed values would be better predictor than the model. Here, the entire validation period was considered and gave rise to a NS = 0.8037, which proves once again the model good performances.

However, it is worth pointing out that some lower performances of the model can be observed when estimating high discharges (Figures 8b and 9). The model tends actually to overestimate the flow exceeding approximatively 250 m^3/s. Such high uncertainty can be the result of low number of cases observed within the dataset and used to train the model CPTs. Indeed, the learning dataset contains only 1,25% of cases where discharge at Roquefort exceeds 250 m^3/s (see Figure 7).

Despite its high predictive performance, as shown in this case study, and the above-mentioned advantages, the accuracy of BNs can be severely impaired by the discretization mode associated with each node in terms of number of classes and range of their respective intervals (Nojavan, Qian, & Stow, 2017). This vulnerability explains why the model fails sometimes to properly reproduce the observed discharges for low and medium values (Figure 8.c).

## 4.3 | Filling missing discharge data using BN model

The reconstruction of missing streamflow data has always been a challenging task in hydrology. Several approaches dealing with gaps in discharge measurement records are reported in literature. Among others, we find regression methods (Tencaliec, Favre, Prieur, & Mathevet, 2015) and artificial neural networks (Coulibaly & Evora, 2007)(Coulibaly & Evora, 2007). Bayesian networks approach may join these existing methods with the further advantage of being
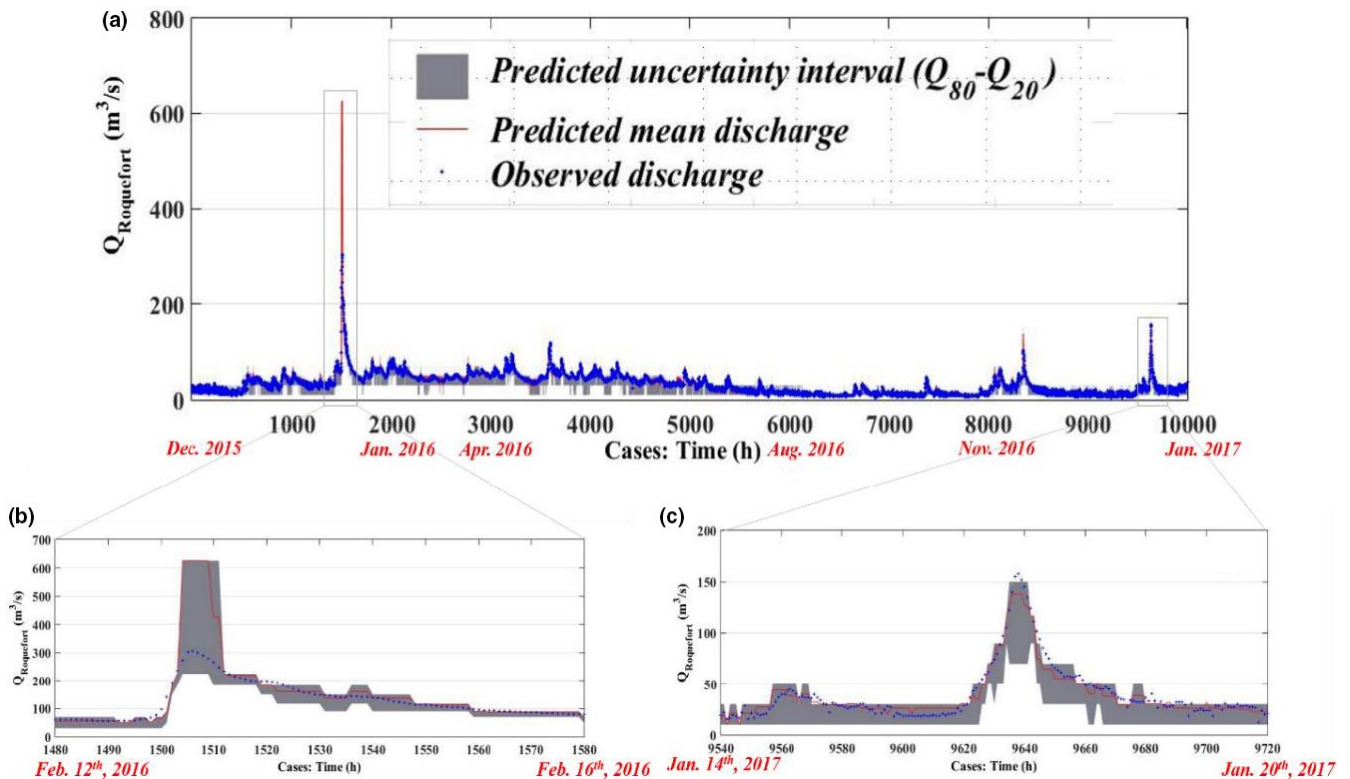
FIGURE 8 (a) Observed (blue dots), Q_{mean} (red lines) predicted mean hydrographs for the validation dataset at the Roquefort station and predicted uncertainty interval estimated ($Q_{80} - Q_{20}$, black range). (b) Zoom on the cases 1,480 to 1,580. (c) Zoom on the cases 9,540 to 9,720 [Colour figure can be viewed at wileyonlinelibrary.com]
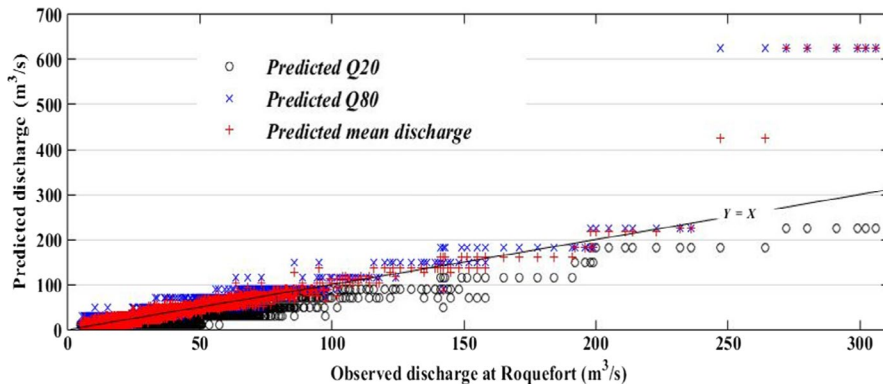
FIGURE 9 Observed discharges at Roquefort compared to BN-estimated discharges during the validation period. Correlation coefficients $R_{20} = 0.87$, $R_{80} = 0.90$ and $R_{mean} = 0.93$ have been calculated, respectively, between observed discharges/Q_{20}, observed discharges/Q_{80} and observed discharges/Q_{mean} [Colour figure can be viewed at wileyonlinelibrary.com]
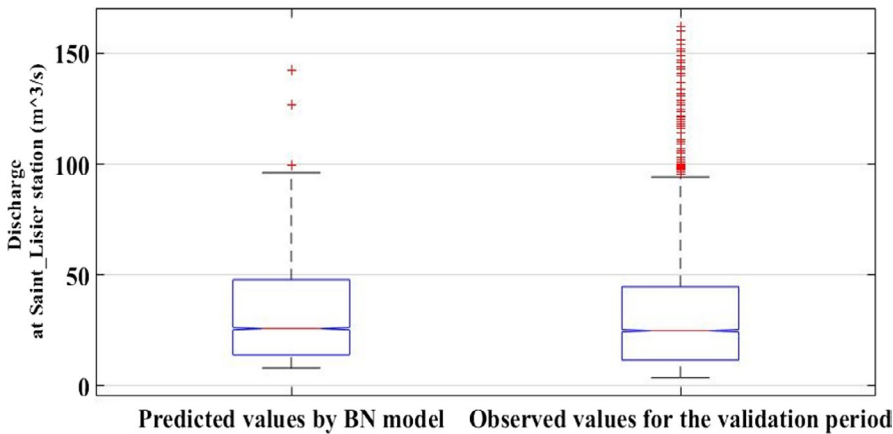
FIGURE 10 Observed discharges at Saint-Lizier compared to BN-reconstructed discharges data during the validation period. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate, respectively, the 25th and 75th percentiles. The outliers are plotted individually using the red "+" symbol [Colour figure can be viewed at wileyonlinelibrary.com]

simple to use, not requiring a great amount of data and suitable for prediction in real time, as they are easily updated and there are no assumptions to check for.

As mentioned earlier, BNs offer the capacity to propagate information in any direction if a new information is introduced. Consequently, this interesting feature of the tool makes it possible to reconstruct by rearward propagation any missing information associated with an upstream node if the flow values are known downstream. By using the model described in this paper, one can therefore estimate the discharge data at one of the upstream stations, for example Saint-Lizier, by inferring, through the model, observed data on other stations. By using the same validation dataset, and the methodology described in section 3.2, "Saint-Lizier" node was treated as unknown during the simulation period (i.e. the values of the node are not observed).

The observations of other proxy stations (i.e. Roquefort, Cazavet and Castelbiague) can then be used, through the model to retrieve the discharge streamflow at Saint-Lizier gauge station. The results of simulation are plotted in Figure 10, simulated quantiles, Q_{25} and Q_{75}, and median streamflow discharges are plotted against observed discharges at Saint-Lizier during the validation period.

Figure 10 shows that the BN model successfully replicates the data, for low and medium discharge values, more specifically for those below 50 m³/s. One can also note that, for so many higher values of discharge at Saint-Lizier (observed red "+" values) above 100

m³/s, the BN model detects only 3 values (predicted red "+" values). This proves poor performances of the BN model for higher flow at this station. As mentioned earlier, this can be explained by the learning process which is not based on a sufficiently comprehensive dataset. Actually, the conditional probability table of Roquefort station is not well trained as the model did not learn enough experiences corresponding to high streamflow at Saint-Lizier station (only 4.62% of cases related to streamflow at Saint-Lizier exceed 100 m³/s).

4.4 | Caveats

We used a complete dataset, with no discharge missing data in four stations, to train the model CPTs. However, incomplete data or missing data are a common situation in the flood forecasting. Bayesian network can deal with this by using adequate learning algorithms such as Expectation–Maximization algorithm (Lauritzen 1995) or Gradient Decent algorithm (Russell, Binder, Koller, & Kanazawa, 1995, Castillo, Gutiérrez, & Hadi, 1996). All these algorithms are implemented in most of Bayesian network packages and tools including Netica.

It should also be kept in mind that a pre-treatment of learning data is needed. The time transfer between a node and its parents has to be incorporated, as a delay, in the learning data file. In addition, if the model is used for a different region or the data used for CPTs training are different, node features such as range, minimum

and maximum values have to be adjusted to new conditions before starting the learning process.

A major challenge in the BN approach is the need for discretization of distributions of continuous variables for most of the available softwares dealing with Bayesian Networks like Netica. Continuous values need to be discretized, which can lead to loss of information. In the future, we plan to use continuous Bayesian networks, more precisely Linear-Gaussian models (Cowell, Dawid, Lauritzen, & Spiegelhalter, 1999, Lauritzen, 1995), to perform river flood prediction. A Linear-Gaussian model is a BN where all the nodes are Gaussian, and each node's mean is linear in the values of its parents. They have fewer parameters then the discrete BNs and provide an efficient inference specially for complex graphs.

5 | CONCLUSION

In this paper, we have been describing the application of a Bayesian network to predict flood risk uncertainty interval on hydrometric station based on its statistical relationships to proxy stations. The potential of Bayesian networks approach was also tested to generate missing data of hourly discharges record. The procedure turned out to be simple, easy to implement, requiring a low modelling effort and very fast to evaluate uncertainty intervals for risk assessment and decision-making. The model showed good performances estimating the river flow at the outlet of a river basin based on the river flows upstream. The model also performed relatively well estimating the missing discharge data, especially for low flow values, at one station through the basin by channelling back the river flow data downstream and in the stations nearby. On this basis, it appears that Bayesian networks are a suitable tool for estimating the flood risk uncertainty and reconstructing discharge measurement records, an essential and less-explored domain for the application of probabilistic direct graphical models.

ORCID
François Pérés https://orcid.org/0000-0001-7821-6651

REFERENCES

Ahmadisharaf, E., Camacho, R. A., Zhang, H. X., Hantush, M. M., & Mohamoud, Y. M. (2019). Calibration and validation of watershed models and advances in uncertainty analysis in TMDL studies. *Journal of Hydrologic Engineering*, *24*(7), 3119001. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001794

Annis, A., Nardi, F., Volpi, E., & Fiori, A. (2020). Quantifying the relative impact of hydrological and hydraulic modelling parameterizations on uncertainty of inundation maps. *Hydrological Sciences Journal*, , *65*(4), 507–523. https://doi.org/10.1080/02626667.2019.1709640

Bhola, P. K., Leandro, J., & Disse, M. (2019). Reducing uncertainties in flood inundation outputs of a two-dimensional hydrodynamic model

by constraining roughness. *Natural Hazards and Earth System Sciences*, *19*(7), 1445–1457. https://doi.org/10.5194/nhess-19-1445-2019

Biondi, D., Versace, P., & Sirangelo, B. (2010). Uncertainty assessment through a precipitation dependent hydrologic uncertainty processor: An application to a small catchment in southern Italy. *Journal of Hydrology*, *386*(1–4), 38–54. https://doi.org/10.1016/j.jhydrol.2010.03.004

Boutkhamouine, B., Roux, H., & Pérès, F. (2017). A Bayesian Network approach for flash flood risk assessment. *EGU General Assembly Conference*, *19*, 10447.

Boutkhamouine, B., Roux, H., Pérès, F., & Vervoort, W. (2018). Uncertainty Analysis of a Temperature-Index Snowmelt Model Using Bayesian Networks. *EPiC Series in Engineering*, *3*, 310–320.

Casas, A., Benito, G., Thorndycraft, V. R., & Rico, M. (2006). The topographic data source of digital terrain models as a key element in the accuracy of hydraulic flood modelling. *Earth Surface Processes and Landforms*, *31*(4), 444–456. https://doi.org/10.1002/esp.1278

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (1996). A new method for efficient symbolic propagation in discrete Bayesian Networks. *Networks*, *28*(1), 31–43. https://doi.org/10.1002/(SICI)1097-0037(199608)28:1<31:AID-NET5>3.0.CO;2-E

Coulibaly, P., & Evora, N. (2007). Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology*, *341*(1–2), 27–41. https://doi.org/10.1016/j.jhydrol.2007.04.020

Cowell, R., Dawid, A., Lauritzen, S., & Spiegelhalter, D. (1999). *Probabilistic Networks and Expert Systems*. New York, NY: Springer-Verlag.

Dittes, B., Špačková, O., & Straub, D. (2019). Managing uncertainty in design flood magnitude: Flexible protection strategies versus safety factors. *Journal of Flood Risk Management*, *12*(2), e12455. https://doi.org/10.1111/jfr3.12455

Fan, Y., Huang, K., Huang, G., Li, Y., & Wang, F. (2019). *An Uncertainty Partition Approach for Inferring Interactive Hydrologic Risks*. Hydrology and Earth System Sciences Discussions, 1–58.

Ferguson, R. I., Hardy, R. J., & Hodge, R. A. (2019). Flow resistance and hydraulic geometry in bedrock rivers with multiple roughness length scales. *Earth Surf. Process. Landforms*, *44*, 2437–2449. https://doi.org/10.1002/esp.4673

Fuchs, S., Heiser, M., Schlögl, M., Zischg, A., Papathoma-Köhle, M., & Keiler, M. (2019). Short communication: A model to predict flood loss in mountain areas, Environmental Modelling & Software, Volume 117. *ISSN*, *176–180*, 1364–8152. https://doi.org/10.1016/j.envsoft.2019.03.026

Garrote, L., Molina, M., & Mediero, L. (2007). Probabilistic forecasts using Bayesian Networks calibrated with deterministic rainfall-runoff models. *NATO Science Series*, *78*, 173–183.

Green, J., & Nelson, E. (2002). Calculation of time of concentration for hydrologic design and analysis using geographic information system vector objects. *Journal of Hydroinformatics*, *4*(2), 75–81. https://doi.org/10.2166/hydro.2002.0009

Han, S., & Coulibaly, P. (2017). Bayesian flood forecasting methods: A review. *Journal of Hydrology*, *551*, 340–351. https://doi.org/10.1016/j.jhydrol.2017.06.004

Jensen, F. (1996). *Introduction to Bayesian Networks*. Secaucus, NJ: Springer-Verlag.

Jiang, X., Liang, Z., Qian, M., Zhang, X., Chen, Y., Li, B., & Fu, X. (2019). Method for probabilistic flood forecasting considering rainfall and model parameter uncertainties. *Journal of Hydrologic Engineering*, *24*(12), 4019056.

Klein, J. A., Tucker, C. M., Steger, C. E., Nolin, A., Reid, R., Hopping, K. A., … Yager, K. (2019). An integrated community and ecosystem-based approach to disaster risk reduction in mountain systems. *Environmental Science & Policy*, *143–152*, 1462–9011. https://doi.org/10.1016/j.envsci.2018.12.034

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, *Issue*, *19*(2), 191–201. https://doi.org/10.1016/0167-9473(93)E0056-A

Lauritzen, S., & Speigelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *50*(2), 157–194. https://doi.org/10.1111/j.2517-6161.1988.tb01721.x

Leandro, J., Gander, A., Beg, M., Bhola, P., Konnerth, I., Willems, W., ... Disse, M. (2019). Forecasting upper and lower uncertainty bands of river flood discharges with high predictive skill. *Journal of Hydrology*, *576*, 749–763. https://doi.org/10.1016/j.jhydrol.2019.06.052

Lim, N. J., & Anders Brandt, S. (2019). Flood map boundary sensitivity due to combined effects of DEM resolution and roughness in relation to model performance. *Geomatics, Natural Hazards and Risk*, *10*(1), 1613–1647. https://doi.org/10.1080/19475705.2019.1604573

Liu, Q., Pérès, F., & Tchangani, A. (2016). Object Oriented Bayesian Network for complex system risk assessment. *IFAC-PapersOnLine*, *49*(28), 31–36.

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

Liu, Z., & Merwade, V. (2019). Separation and prioritization of uncertainty sources in a raster based flood inundation model using hierarchical Bayesian model averaging. *Journal of Hydrology*, *578*, 124100. https://doi.org/10.1016/j.jhydrol.2019.124100

Mansanarez, V., Le Boursicaud, R., Le Coz, J., Renard, B., Lang, M., Horner, I., ... Pobanz, K. (2017). BaRatin-SFD, analyse bayésienne des courbes de tarage à double échelle et de leurs incertitudes. *La Houille Blanche - Revue internationale de l'eau, EDP Sciences*, *2017*(5), 22–28.

Marcot, B. (2012). Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modelling*, *230*, 50–62. https://doi.org/10.1016/j.ecolmodel.2012.01.013

Matott, L., Babendreie, J. E., & Purucker, S. T. (2009). Evaluating uncertainty in integrated environmental models: A review of concepts and tools. *Water Resources Research*, *45*(6), https://doi.org/10.1029/2008WR007301

McCuen, R. (2009). Uncertainty analyses of watershed time parameters. *Journal of Hydrologic Engineering*, *14*(5), 490–498. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000011

Molinari, D., De Bruijn, K. M., Castillo-Rodríguez, J. T., Aronica, G. T., & Bouwer, L. M. (2019). Validation of flood risk models: Current practice and possible improvements. *International Journal of Disaster Risk Reduction*, *33*, 441–448. https://doi.org/10.1016/j.ijdrr.2018.10.022

Morgan, M. G. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. New York, NY: Cambridge Press.

Morrison, A., Westbrook, C., & Noble, B. (2017). A review of the flood risk management governance and resilience literature. *Journal of Flood Risk Management*, *11*(3), 291–304. https://doi.org/10.1111/jfr3.12315

Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models, Part I - A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nojavan, A. F., Qian, S. S., & Stow, C. A. (2017). Comparative analysis of discretization methods in Bayesian networks. *Environmental Modelling & Software*, *87*, 64–71.

Norsys Software Corp (2015). *Norsys Software Corp*. Retrieved from http://www.norsys.com

Ogie, R. I., Adam, C., & Perez, P. (2020). A review of structural approach to flood management in coastal megacities of developing nations: current research and future directions. *Journal of Environmental Planning and Management*, *63*(2), 127–147. https://doi.org/10.1080/09640568.2018.1547693

Pahl-Wostl, C. (2007). Transitions towards adaptive management of water facing climate and global change. *Water Resources Management*, *21*(1), 49–62. https://doi.org/10.1007/s11269-006-9040-4

Pant, R., Thacker, S., Hall, J., Alderson, D., & Barr, S. (2018). Critical infrastructure impact assessment due to flood exposure. *Journal of Flood Risk Management*, *11*, 22–33.

Park, J., Seager, T. P., Rao, P. S. C., Convertino, M., & Linkov, I. (2013). Integrating risk and resilience approaches to catastrophe management in engineering systems. *Risk Analysis*, *3*, 33. https://doi.org/10.1111/j.1539-6924.2012.01885.x

Pearl, J. (1985). *Bayesian networks: A model of self-activated memory for evidential reasoning*. Proc. of Cognitive Science Society, (CSS-7).

Ronalds, R., & Zhang, H. (2019). Assessing the impact of urban development and on-site Stormwater detention on regional hydrology using Monte Carlo simulated rainfall. *Water Resources Management*, *33*(7), 2517–2536. https://doi.org/10.1007/s11269-019-02275-y

Roux, H., Labat, D., Garambois, P.-A., Maubourguet, M.-M., Chorda, J., & Dartus, D. (2011). A physically-based parsimonious hydrological model for flash floods in Mediterranean catchments. *Natural Hazards and Earth System Science*, *11*(9), 2567–2582. https://doi.org/10.5194/nhess-11-2567-2011

Ruin, I., Creutin, J.-D., Anquetin, S., & Lutoff, C. (2008). Human exposure to flash floods – Relation between flood parameters and human vulnerability during a storm of September 2002 in Southern France. *Journal of Hydrology*, *361*(1-2), 199–213. https://doi.org/10.1016/j.jhydrol.2008.07.044

Russell, S., Binder, J., Koller, D., & Kanazawa, K. (1995). Local learning in probabilistic networks with hiden variables. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, *2*, 1146–1152.

Russell, S., & Norvig, P. (2010). *Artificial intelligence a modern approach*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Seyam, M., & Othman, F. (2014). The influence of accurate lag time estimation on the performance of stream flow data-driven based models. *Water Resources Management*, *28*(9), 2583–2597. https://doi.org/10.1007/s11269-014-0628-9

Si, W., Gupta, H. V., Bao, W., Jiang, P., & Wang, W. (2019). Improved dynamic system response curve method for real-time flood forecast updating. *Water Resources Research*, *55*(9), 7493–7519. https://doi.org/10.1029/2019WR025520

Sieg, T., Schinko, T., Vogel, K., Mechler, R., Merz, B., & Kreibich, H. (2019). Integrated assessment of short-term direct and indirect economic flood impacts including uncertainty quantification. *PLoS One*, *14*, 4. https://doi.org/10.1371/journal.pone.0212932

Stehman, S. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89. https://doi.org/10.1016/S0034-4257(97)00083-7

Stephens, T. A., & Bledsoe, B. P. (2020). Probabilistic mapping of flood hazards: Depicting uncertainty in streamflow, land use, and geomorphic adjustment. *Anthropocene*, *29*, 100231. https://doi.org/10.1016/j.ancene.2019.100231

Tencaliec, P., Favre, A.-C., Prieur, C., & Mathevet, T. (2015). Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research*, *51*(12), 9447–9463. https://doi.org/10.1002/2015WR017399

Terzi, S., Torresan, S., Schneiderbauer, S., Critto, A., Zebisch, M., & Marcomini, A. (2019). Multi-risk assessment in mountain regions: A review of modelling approaches for climate change adaptation. *Journal of Environmental Management*, *232*, 759–771. https://doi.org/10.1016/j.jenvman.2018.11.100

Tscheikner-Gratl, F., Bellos, V., Schellart, A., Moreno-Rodenas, A., Muthusamy, M., Langeveld, J., ... Tait, S. (2019). Recent insights on uncertainties present in integrated catchment water quality modelling. *Water Research*, *150*, 368–379. https://doi.org/10.1016/j.watres.2018.11.079

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, *203*, 312–318. https://doi.org/10.1016/j.ecolmodel.2006.11.033

Yassine, R., Pérès, F., Roux, H., Cassan, L., & Frysou, O. (2018). A Bayesian network approach to study hydromorphological modifications over

space and time in the framework of a sustainable river restoration project: the "Lac des Gaves" case study. *IFAC-PapersOnLine*, *51*(5), 7–12. https://doi.org/10.1016/j.ifacol.2018.06.191

Yassine, R., Roux, H., Recking, A., Peres, F., Cassan, L., & Frysou, O. (2018). Space-time scale dependences phenomena in the framework of river restoration projects-the Lac des Gaves case study. *EGU General Assembly Conference Abstracts*, *20*, 7292.

APPENDIX A

The following code is used to create the BN for simulating the discharge at "Roquefort" station using Netica Java-API. This version of Netica APIs offers a clean and an object-oriented way to develop and work with BN.

```
/Import Netica java libraries needed for BN development import norsys.netica.*;/Start the public class for the model creation. public class Model_development { /public static void main(String[] args)throws NeticaException {/ Set the work file path String PathBN="D:/These/.../BN/";/ Create the environment for the network Environ env = new Environ("NULL");/ Create an empty BN Net BN = new Net(env);/Node creation Node Castelblaque = new Node("Castelblaque",0, BN);Node Saint_Lizier = new Node("Saint_Lizier",0, BN);Node Cazavet = new Node("Cazavet",0, BN);Node Roquefort = new Node("Roquefort",0, BN);/Set manually the BN structure/ Set parents of "Roquefort node" Roquefort.addLink(Castelblaque);Roquefort.addLink(Saint_Lizier);Roquefort.addLink(Cazavet);/Node Discretization/ The class of discretization function DiscrtFunction Dist = new DiscrtFunction();/ Castelbiague node Castelblaque.setLevels(Dist.ConcatArray(Dist.Values(0.0, 2.0, 4), Dist.Values(3.0, 20.0, 4),Dist.Values(30.0, 100.0, 2)));/ Saint_lizier node Saint_Lizier.setLevels(Dist.ConcatArray(Dist.Values(0.0, 50.0, 5), Dist.Values(50.01, 200.0, 3),Dist.Values(200.01, 500.0, 3)));/ Cazavet node Cazavet.setLevels(Dist.ConcatArray(Dist.Values(0.0, 1.0, 4), Dist.Values(1.5, 10.0, 4),Dist.Values(15.0, 60.0, 2)));/ Roquefort node Roquefort.setLevels(Dist.ConcatArray(Dist.Values(0.0, 80.0, 5), Dist.Values(100.0, 200.0, 4),Dist.Values(250.0, 1,000.0, 2)));/ Learning CPTs from data double d = 1;NodeList nodes = BN.getNodes();Streamer caseFile = new Streamer (PathBN+ "LerData.txt");Caseset cases = new Caseset();cases.addCases(caseFile, 1.0, null);/ Specify the learning method Learner learner = new Learner(Learner.COUNTING_LEARNING);learner.setMaxTolerance(1e-4);learner.learnCPTs(nodes, cases, d);learner.finalize();cases.finalize();/Compile the model BN.compile();/ Save the model into a model database. Streamer modelstream = new Streamer(PathBN+ "BN_SALAT4Stations.dne");BN.write(modelstream);BN.finalize;/ Close BN Env.finalize;/ Close working environment.}}
```

APPENDIX B

After creating and learned BN, it can be used to calculate the posterior probabilities of a target node, here Roquefort, using observations on others nodes.

```
/ Netica java libraries import norsys.netica.*;/ Model validation public class public class Model_validation {public static void main(String[] args)throws NeticaException, IOException {/ Create new environment Environ env = new Environ("NULL");/ Model work file path String PathBN="D:/These/.../BN/";/ Read the BN from the model database Net BN = new Net(new Streamer (PathBN+"BN_SALAT4Stations.dne"));/ Read the nodes of the BN Node Castelbiague = BN.getNode("Castelebiaque");Node Saint_Lizier = BN.getNode("Saint_Lizier");Node Cazavet = BN.getNode("Cazavet");Node Roquefort = BN.getNode("Roquefort");/ BN compile BN.compile();/Sensitivity analysis/ Target node Node queryNode = BN.getNode("Roquefort ");/ Varying nodes NodeList varyingNodes = Roquefort.getParents();/ To measure mutual information using variance reduction method Sensitivity realVarSens = new Sensitivity( queryNode,varyingNodes ,Sensitivity.VARIANCE_OF_REAL_SENSV);/ for example the mutual information between « Roquefort» and one of its parents can be estimated double realVar = realVarSens.getVarianceOfReal(Castelbiague);System.out.println("The variance of real between Castelbiague and Roquefort is "+ realVar);realVarSens.finalize();/ Use the model to perform inference/Read the validation data for observed nodes for the validation period Data data = new Data();double [] Qcavt = data.ReadData(PathBN+" Cazavet.txt");double [] Qctlque = data.ReadData(PathBN+"Castelblaque.txt");double [] QSaintL = data.ReadData(PathBN+"Saint_Lizier.txt");int n = Qcavt.length;/ data length// Create a file writer to save results of "Roquefort" posterior probabilities FileWriter fw = new FileWriter(PathBN+"RoquefortFindings");PrintWriter pw = new PrintWriter(fw);/Get and save the possible values (discretization) of Roquefort double [] RoqLevels = Roquefort.getLevels();for (int j = 0;j < Roquefort.getNumStates();j++) {pw.print((RoqLevels[j]+RoqLevels[j + 1])/2 +";");}pw.print("\n");/ Inferring observations through the model to estimate the posterior probabilities of "Roquefort node" for each time step. float [] probs = new float [Roquefort.getNumStates()];System.out.println("Roquefort number of states: "+Roquefort.getNumStates());double prob; int j,m; Double maxprob;for (int i = 0;i < n;i++) {/ Clear previous observations of upstream stations. Cazavet.finding().clear();Castelblaque.finding().clear();Saint_Lizier.finding().clear();/Set observations of observed nodes Cazavet.finding().enterReal(Qcavt[i]);Castelblaque.finding().enterReal(Qctlque[i]);Saint_Lizier.finding().enterReal(QSaintL[i]);/ Interfere the observations through the model BN.compile();/ Show time step System.out.println("pas:"+i);/ Get posterior probabilities of "Roquefort node for each time step j = 0; m = 0;maxprob = 0.00;do {prob = Roquefort.getBelief("#"+String.valueOf(j));pw.print(prob +";"); j++;} while ( j < Roquefort.getNumStates());probs=(float []) Roquefort.getBeliefs();pw.print("\n");}pw.close();BN.finalize();}}
```