EPiC Series in Engineering

Volume 3, 2018, Pages 310–320

HIC 2018. 13th International Conference on Hydroinformatics

Uncertainty analysis of a Temperature-Index Snowmelt Model using Bayesian Networks

Brahim BOUTKHAMOUINE[1,2], Hélène ROUX[1], François PERES[2], R. Willem VERVOORT[3]

[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.
[3] Centre for Carbon, Water and Food, Sydney Institute of Agriculture, The University of Sydney, Sydney, Australia.

Corresponding author: bboutkha@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, we explore in this study the potential of Bayesian Networks (BNs). A BN is created using a simple version of Temperature-Index Snowmelt Model. Next, uncertainty analysis is performed using both the BN method and Monte-Carlo (MC) simulations. The results show that BN method gives similar results to the MC method and can be used for real-time applications.

Keywords: Uncertainty analysis, Deterministic Model (DM), Bayesian Network (BN).

Introduction

Although much work has been done to overcome uncertainty problems in hydrological modelling, several proposed methods are computationally demanding (see a recent review in [**1**], and the references cited therein). The most common method to estimate uncertainties within the hydrological community is to use the classic Monte-Carlo method (MC, [**2**]). It involves running the model successively using random sampling from the distributions of input parameters until sufficient samples of the output distribution have been obtained. Without using smart sampling, such as Latin

Hypercube, it is obvious that the MC simulations requires a large number of samples of the input distribution. If the model is complex or the number of uncertain parameters is high, MC method becomes time consuming and computationally expensive. As an example, for a hydrological model with five uncertain parameters that takes 5 seconds to run, in order to propagate uncertainties tied with these parameters through the model, 10 000 runs will take 50 000 seconds to be done, that is to say more than 13 hours of simulations. In this study we explore the potential of Bayesian Networks (BNs, [**3**]) to allow real-time uncertainty estimation. BNs are mathematical models presented in the form of Directed Acyclic Graphs (DAGs) consisting of nodes and directed arcs relating these nodes. The nodes present the variables and the arcs represent the causal relationships between these variables. They can be used to create "expert systems" including expert knowledge about complicated domains and phenomena. They can also be used as tools for decisions support. In this paper, we will demonstrate uncertainty analysis on a Temperature-Index Snowmelt Model [**4**] using BNs and comparing to MC simulations.

1 Methods: Bayesian Networks

1.1 Definition of Bayesian Networks

A Bayesian Network $\mathrm{B} = \{\mathrm{G}, \mathrm{P}\}$ is a combination of both graph and probability theory. It is defined as: $\mathrm{G} = \{\mathrm{X}, \mathrm{E}\}$, a directed acyclic graph, without circuits, consisting of nodes and arcs relating these nodes. The nodes are associated with a set of random variables, $\mathrm{X} = \{X_1, ..., X_n\}$, explaining the studied phenomenon, e.g. snow melt process, and directed arcs, E, represent the set of causal relationships between these variables. Each node, X_i, in the graph is associated with a conditional probability , $\mathrm{P} = \{\mathrm{P}(X_i/\mathrm{Pa}(X_i)\}$, expressing the effect of the variables, $Pa(X_i)$, that cause X_i in G. P are also called local probabilities and they express the "dependencies strengths" between the nodes (variables). In practice, P are defined in Conditional Probability Tables (CPTs), in the form of tables. The way G is shaped, i.e. a directed acyclic graph, simplifies the computation of both joint and marginal probabilities of the nodes, X, making up the network. The computation of these probabilities comes down to the product or sum of conditional probability terms directly accessible from CPTs. In BNs terms, this operation is called inference:

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

As we can see in equation 1, the computation of $P(X_1, ..., X_n)$comes down to the product of local probabilities terms directly accessible from CPTs associated with the

network. This combination between the theory of graphs and probability is one of the powerful aspects of Bayesian Networks providing an efficient way to preform inference (the process of estimating a posterior probability of a variable of interest giving observations on some other variables in the graph). Algorithms dealing with inference in BNs are explained in details in Pearl (1988) **[5]** and Jensen (1996) **[6]**. The most classical and famous one rests on what it's called as Junction Tree (JT) method (also known as Clique Tree method). It is used to compute marginal probabilities on graphs by creating a tree of cliques, and carrying out a message-passing procedure on this tree. This is done through three stages: Moralisation, Triangulation, and assembling cliques into a junction tree. More details on this method are available on Jensen (1996, p.76) **[6]**.

1.2 Bayesian Network associated with the snow melt model

1.2.1 Temperature-Index Snowmelt Model

A Temperature-Index Snowmelt Model is a widely-used method to estimate snow melt rates [**4,7**]. It assumes a simple empirical deterministic relationship between air temperature and melt rates. The simplest version of this model can be summarized as: during a rainfall event, the nature of precipitation P reaching the soil surface is closely related to the air temperature (T_a); if $T_a < 0$ then P falls as snow (P_s), otherwise P would be rain. Snow Water Equivalent (SWE) represents the amount of water stored in the snow (vertical depth), it is a function of snow precipitation P_s and melt rate (S_m). The model is generally applied as a time series model with a daily step. For each step, SWE_t is computed:

$$SWE_t = SWE_{t-1} + P_s - S_m \quad with$$
$$S_m = \begin{cases} Min(SWE_{t-1}, M_f * T_a) & if \quad T_a > 0 \\ 0 & else \end{cases} \qquad (2)$$

Where M_f is the melt factor, which is the rate of melt per degree per day. M_f expresses the influence of the meteorological conditions other than air temperature and physiographic characteristics of the basin. This model is purely deterministic with M_f as its unique parameter. From now on, we denote this model as the Deterministic Model (DM). In order to construct a Bayesian network representing this model, two elements are needed (see section 1.1): the structure and the conditional probability tables.

1.2.2 Specifying the structure of BN

The structure of BN is directly inspired by the deterministic model (DM) presented above. The posterior probability of SWE is computed knowing the prior distribution

probabilities of the inputs (P and T_a), of the model parameter (M_f) and anterior conditions SWE_0 (figure 1).

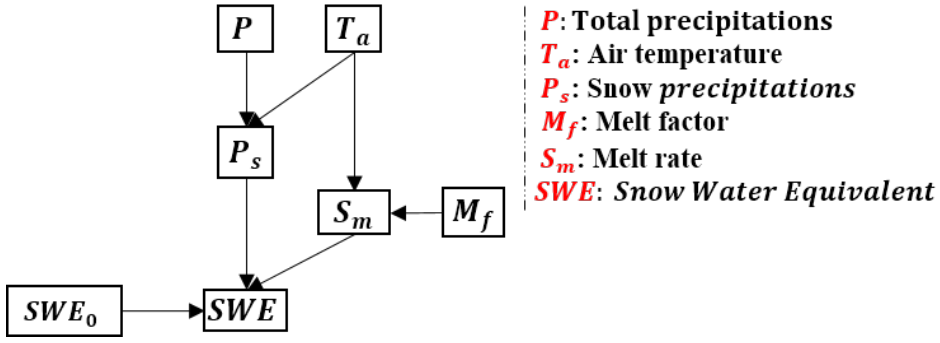


Figure 1: Oriented directed structure of BN inspired from DM.

1.2.3 Estimating CPTs of BN

Once the model structure is fixed, model parameters have to be estimated. This phase consists of estimating all the CPTs involved in the causal relationships of the network. This is done by using algorithms in which BN is trained based on a dataset of cases [**8**]. In our case, the dataset of 100 000 cases is created using the DM model. The DM is forced with stochastic precipitation and temperature with random values of M_f. The results are stored in a text file text used as our database for estimating CPTs. Next, we used Netica [**9**], a Bayesian network toolkit, and specifically its counting algorithm to learn the network CPTs from the created data. 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 (for details, see [**10**]). Because of the requirements imposed by many BN software packages as it is the case of Netica, which do not support continuous variables, the discretisation of continuous nodes is often necessary in BN construction. To address this issue, the users can either rely on the discretisation solutions offered by these packages, or by finding their own ones. The discretization policy followed here is simple. It consists of scanning the historic data, presented below (section 1.2.5), in order to identify the minimum and the maximum values of all observed variables, and then divide the continuous nodes into a number of intervals. Ideally, it is only a few intervals should be associated with each node in order to reduce the dimension of BN and optimize its time inference computation. Two versions of BN model were constructed using two arbitrary discretization (table 1).

Nodes	P & P_s	T_a	S_m	M_f	SWE & SWE
Range Variation	*[0, 135] mm*	*[-16, 24] C°*	*[0, 250] mm*	*[0, 10] mm/d/C°*	*[0, 2470] mm*
Discretisation **D20**	20	20	50	20	**20**
Discretisation **D100**	20	20	50	20	**100**

Table 1: Range variations of BN nodes and their two discretization.

1.2.4 Used software

In this study, the two BN versions were developed flowing the following the discretization fixed in table 1. the two versions of BN are developed using the software package of Netica version API Java (Netica-J, [**9**]) to. Netica-J is a complete program for working with Bayesian Networks. It has many practical utilities such as the ability to develop, train, change and store networks and determine better resolutions. Netica performs inference which solves the network by finding the marginal posterior probability for each node using the Junction Tree algorithm. One advantage of Netica is the comprehensive, flexible and user friendly graphical user interface included in the package.

1.2.5 Case Site study: MtHood field site

The data from the Mount Hood field site **[11]** has been chosen to demonstrate the BNs. Mount Hood field site, NRCS site number 651, is located in northern Oregon, USA, at approximately 1637 meters above sea level (latitude: 45.32, and longitude: -121.72). The site is operated by National Resources Conservation Service (NRCS) and provides observation data of snow water equivalent, snow depth, precipitation, temperature and other climatic variables in hourly, daily, monthly and yearly increments. The daily observed precipitation, air temperature and snow water equivalent (SWE) between 01/10/2005 to 30/09/2013 were downloaded and used to perform uncertainty analysis for both the DM and BN Model.

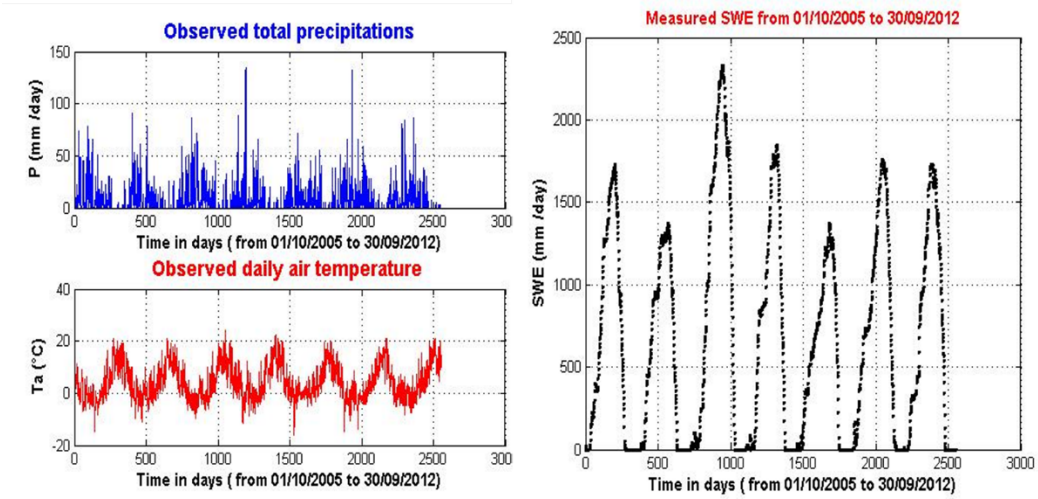


Figure 2: **(a)** data of daily observed precipitation, temperature, and **(b)** snow water equivalent (SWE) at Mt Hood field site.

Hereafter, we will simulate the snow water equivalent (SWE) at this station giving the observed precipitation and air temperature presented in figure 2.a, and then compared it to the observed SWE showed in figure 2.b. The simulations are done using both Deterministic Model (DM) and Bayesian Network (BN) model.

2 Results and discussion

2.1 Calibration of the model DM

In order to estimate the optimum value of the parameter melt day factor M_f, we have calibrated the model DM using one year of historic data of precipitation, air temperature and observed SWE presented in Figure 2 (from October 2012 to September 2013). Several values of M_f were tested and the performance of the model DM was estimated using the Nash- Sutcliffe (NS) efficiency criteria [**12**]. It is defined as one minus the sum of the absolute squared differences between the predicted and observed values normalized by the variance of the observed values during the period under investigation. NS ranges between 1.0 (perfect fit) and $-\infty$. An NS lower than zero indicates that the mean value of the observed values would better predictor than ones given by the model. Several runs of the model DM have been carried out with several values of M_f. For each run, NS is calculated (figure 3). According to this criteria, $M_f = 1.4\ \mathrm{mm/d/C°}$ is found to be the optimum value of DM with a value of NS=0.7634.

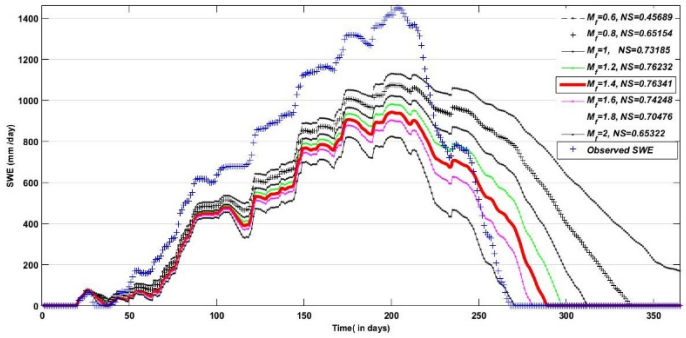


Figure 3: Calibration process of the DM model. An optimum value of $M_f = 1.4\ \mathrm{mm/d/C°}$ is found.

Later on, after we have set the discretisation D100 for BN model and learned CPTs using the methodology described in section **1.2.3**, we compute for each time step the posterior probability of SWE (figure 4). This probability, in the form of a state vector, corresponds to all possible values of SWE giving the inputs (precipitations and air temperature), and the optimum value of Mf as evidences. In the Bayesian network terms, this operation is called inference. For each time step, a state vector of SWE is computed (figure 4), in which we can access to the most probable value, mean value, standard deviation etc. The performance of the BN model can be estimated by comparing the most probable values of SWE of each time step to the observed SWE using NS criteria. In our case we found an NS=0.6769. This value is slightly lower than the one calculated for DM (NS=0.7634). This is due to the discretisation (D100), associated with the BN.

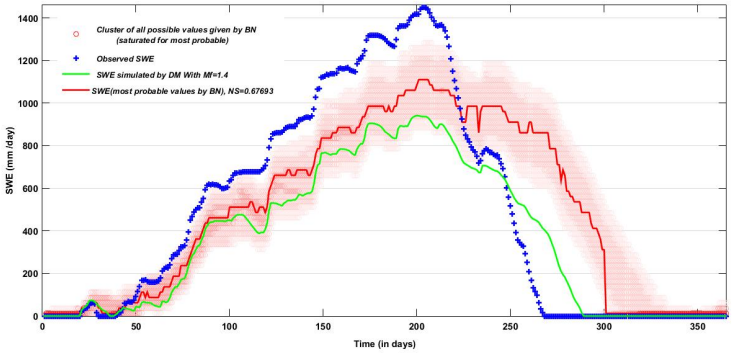


Figure 4: Posterior probabilities of SWE calculated by BN model with $M_f = 1.4\ \mathrm{mm/d/C°}$. for each time step, the saturated color corresponds to the most probable values.

2.2 Results of uncertainty analysis using DM and BN model

Figure 3 shows that DM model gives acceptable results for M_f ranging from 1 to 1.8 mm/d/C° (NS>0.70), with a pic of performance at $M_f = 1.4\,\mathrm{mm/d/C°}$. Hereafter, we suppose that M_f has a Gaussian distribution around this optimum value with a standard deviation, equals to 0.4 mm/d/C°, expressing the average magnitude of uncertainty associated with this parameter (figure 5).

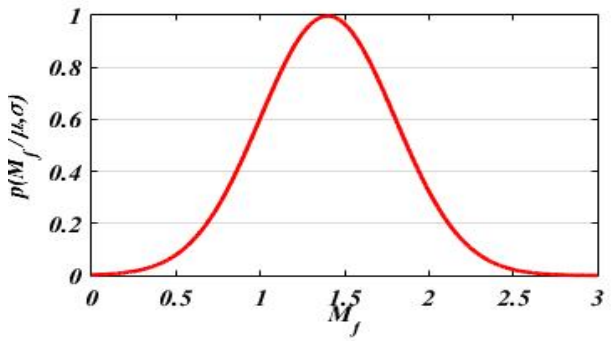


Figure 5: Uncertainty distribution function of M_f, $p(\mu,\sigma) = N(1.4,0.4)$ *mm/d/C°*.

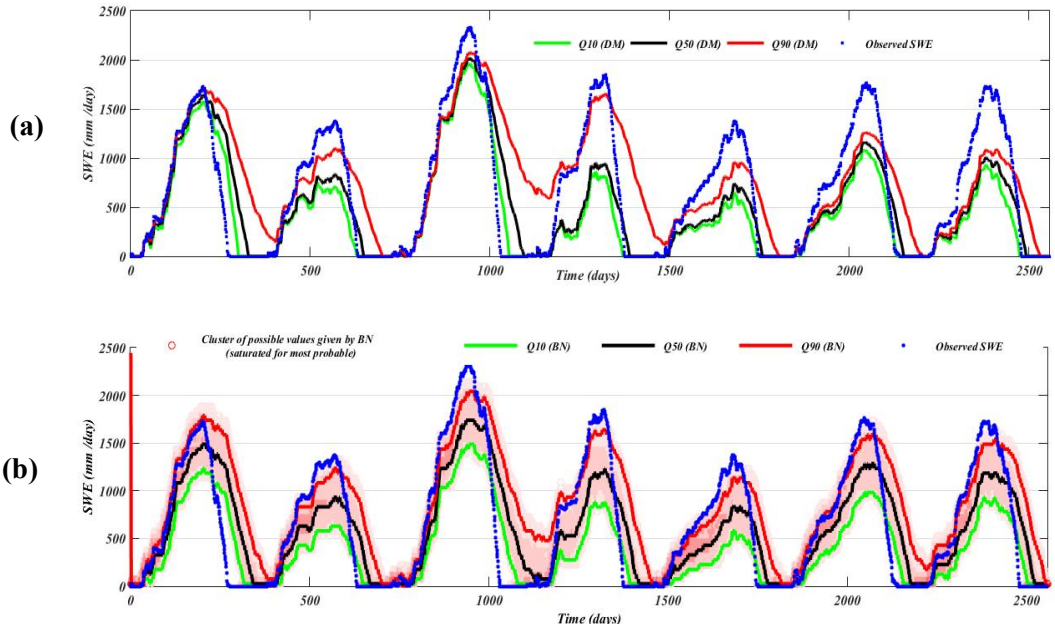
Generally speaking, taking into consideration the uncertainties from different sources (inputs, model parameters, model structure etc.), will certainly improve the reliability of the outcomes (forecasts) giving by a hydrological model. It also helps the decision maker well interpret the results. The assessment of uncertainty in a hydrological model output requires the propagation of different sources of uncertainty through the modelling system. Monte-Carlo (MC) procedure is the most useful method used to achieve this purpose within hydrology community. This by forcing the model with perturbations in model parameters through random sampling from prescribed probability distributions using a MC simulation framework [**13, 14**].

In this study, Uncertainty analysis using both MC procedure on DM model and BN model are carried out. Our aim is to propagate the uncertainty tied with M_f (Figure 5) and see how it affects the outcome SWE. The simulations are done using the data presented in section 1.2.5. The simulation period was set to 1 October 2005, until 30 September 2012, with 1 day as time step length.

The DM model was run 20 000 times. For each time, a single random value of M_f was chosen from the distribution and the SWE result is recorded. Using the corresponding MC simulations, we can calculate, for each time step, some statistical properties such as moments (mean and variance), quantiles (Q10, Q50, Q90) of the simulated SWE.

The BN model is also used to propagate the uncertainty distribution function of M_f through the model. Both versions were tested, corresponding to the discretisation D20, D100 (Table 1). For each version, we simulate the same data, as we did with DM model, by entering for each time step the inputs (P, T_a) and the Gaussian distribution function of $M_f, N(1.4,0.4)$, as evidences. This is done in one run thanks to the

Bayesian Networks inference algorithms. Figure 6 shows different results obtained with DM model and the BN model discretized at D20 and D100. First, one can see that DM model and BN versions, exhibit similar time evolution patterns. This is totally expected since the BN model is a graphical implementation of deterministic equations making up DM. The widths of the prediction uncertainty obtained with DM and BN (version D20) are however different (figure 6.a and 6.b)). We can see a big portion of observed SWE (in blue color) are inside this large interval. When we refine the discretisation of BN (version D100, figure 6.c), the calculated uncertainty interval becomes tighter, but does not include most of the observed SWE. In fact, the more we refine the discretisation of BN the more we converge towards the mean of DM simulations: a coefficient correlation R=0.97 is calculated between BN (version D100) and DM simulations in this case. This proves the deterministic equations (DM model) used to train CPTs of BN are not able to reproduce the process of snow melt at this studied area. The DM, as simple as it is, is clearly not able to reproduce the complexity of snow melt process at this studied site. We can also evaluate the obtained results by considering a global index for uncertainty prediction. It is defined as the percentage of observed SWE points located inside the uncertainty interval Q90-Q10. The DM model and BN model with D20 discretisation give the best performances with 93.74% and 95.42% respectively. This is because the corresponding uncertainty intervals are quite wide. For finer discretization of BN, we calculate lower values of this index; 55.06% for BN (version D100).



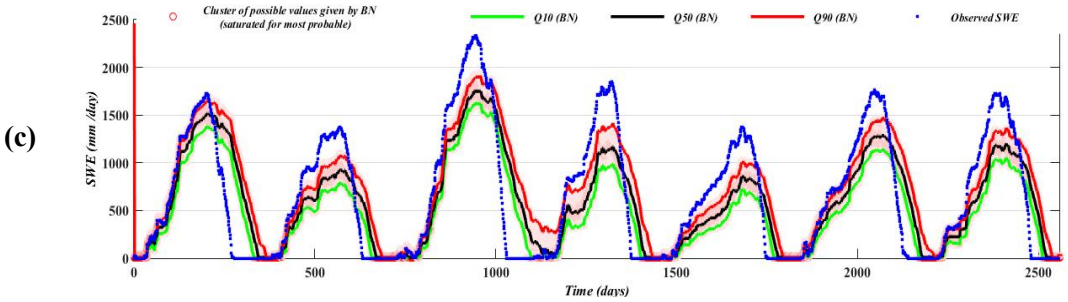


Figure 6: SWE uncertainty simulations resulting from the propagation of a Gaussian uncertainty tied with the parameter $M_f, N(1.4,0.4)$. The prediction uncertainty is bounded by the calculated Q90 and Q10. (a) the propagation is done by forcing the DM model with 20 000 values of M_f chosen randomly form $N(1.4,0.4)$. (b) propagation of uncertainty using BN model with discretization D20. (c) propagation of uncertainty using BN model with discretization D100 (table 1).

Besides deterministic equations, the BN CPTs can be estimated also using expert's opinion and experimental data. Depending on the discretisation, training CPTs using deterministic modes could be costly and time consuming both in the learning and inference stage since the consequent Bayesian Network is complex with a lot of nodes and links. Another issue related to this method is that the performances of BNs depend on the credibility of the deterministic equation used to train CPTs. The best way to estimate CPTs is using experimental data. This way, the BNs can capture the real response of environmental system studied. Unfortunately, experimental data are not always available or presenting missing values. Learning algorithms tempting to handle this problem often delete the observations with missing values or fill in the missing values. Such procedures may however lead to biased results. One other option is to use expert's knowledge to estimate some CPTs in the network. but its limited to the qualitative nodes since it's very hard to estimate manually tables with big sizes associated with quantitative ones.

Conclusion

Uncertainty analysis on hydrological models could be done using BNs. The advantage of BNs is that we don't need to run the model several times in real-time as the computations of the CPTs can be done a priori. Such advantages could be used to perform uncertainty predictions within operational hydrological model such as flood forecasting systems. BNs make possible to integrate both deterministic, as we did in this article, expert's knowledge, estimating the CPTs using expert's knowledge, into one framework, which can improve its performances. However, one major challenge facing the BN methodology is the need for discretisation of distributions of continuous

variables. Environmental variables and parameters often have continuous values, but the BN methodology, or especially the available software's developed to construct BNs, are very limited in its abilities to deal with such variables. Hence, these values are often discretised, which can lead
to loss of information. A common way to transform continuous values into discrete values is by to divide the continuous distribution into intervals as we did in section 1.2.3.

References

[1] Matot T, L. S., J. E.Babendreie Ret S. T. Purucker, (2009). Evaluating uncertainty in integrated environmental models: A review of concepts and tools. Water Resources Research, 45: W06421, 2009.
[2] Jeremy E. Oakley and Anthony O'Hagan (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. J. R. Statist. Soc. B (2004) 66, Part 3, pp. 751–769.
[3] Jensen F., (1996). Introduction to Bayesian Networks. Springer-Verlag.
Lauritzen S., Speigelhalter D.J., (1988). Local computations with probabilities on graphical structures and their application to expert systems. Royal statistical Society B, 50, 157–224.
[4] Rango A., Martinec J. (1995). Revisiting the degree-day method for snowmelt computations. Water Resour Bull 31: 657±669.
[5] Pearl, J., Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 1988.
[6] Jensen F., (1996). Introduction to Bayesian Networks. Springer Verlag.
[7] Rango A., Martinec J. (1995). Revisiting the degree-day method for snowmelt computations. Water Resour Bull 31: 657±669.
[8] Ramoni, M., Sebastiani, P.: Robust learning with missing data. Mach. Learn. 45(2), 147–170 (2001).
[9] https://www.norsys.com/netica-j.html
[10] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Los Altos, CA, Morgan-Kaufman
(2009)
[11] https://www.nrcs.usda.gov/wps/portal/nrcs/detail/or/home/?cid=nrcs142p2_046290
[12] Nash, J. E. and Sutcliffe, J. V. : River flow forecasting through conceptual models, Part I - A discussion of principles, J. Hydrol., 10, 282–290, 1970.
[13] Beven, K., and A. Binley (1992). The future of distributed models: Model calibration and uncertainty predition, Hydrol. Processes, 6(3), 279-298, doi: 10.1002/hps.3360060305.
[14] Helton, J. C., and F. J. Davis (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliab. Eng. Syst. Saf., 81(1), 23– 69, doi:10.1016/S0951-8320(03)00058-9.
(2009)