Modelling uncertainty of flood quantile estimations at ungauged sites by Bayesian networks

D. Santillán, L. Mediero and L. Garrote

ABSTRACT

Prediction at ungauged sites is essential for water resources planning and management. Ungauged sites have no observations about the magnitude of floods, but some site and basin characteristics are known. Regression models relate physiographic and climatic basin characteristics to flood quantiles, which can be estimated from observed data at gauged sites. However, some of these models assume linear relationships between variables and prediction intervals are estimated by the variance of the residuals in the estimated model. Furthermore, the effect of the uncertainties in the explanatory variables on the dependent variable cannot be assessed. This paper presents a methodology to propagate the uncertainties that arise in the process of predicting flood quantiles at ungauged basins by a regression model. In addition, Bayesian networks (BNs) were explored as a feasible tool for predicting flood quantiles at ungauged sites. Bayesian networks benefit from taking into account uncertainties thanks to their probabilistic nature. They are able to capture non-linear relationships between variables and they give a probability distribution of discharge as a result. The proposed BN model can be applied to supply the estimation uncertainty in national flood discharge mappings. The methodology was applied to a case study in the Tagus basin in Spain.

Key words | Bayesian networks, flood quantiles, prediction at ungauged basins, uncertainty estimation

D. Santillán (corresponding author)
L. Mediero
L. Garrote
Department of Civil Engineering, Hydraulic and Energy Engineering,
Technical University of Madrid,
C/Professor Aranguren s/n,
28040 Madrid,
Spain
E-mail: *david.santillan@upm.es*

ABBREVIATIONS

AMF	Annual maximum flood
ANN	Artificial neural network
BN	Bayesian network
BGLS	Bayesian approach analysis of a GLS technique
CPT	Conditional probability table
DTM	Digital terrain model
GEV	Generalised extreme value
GLS	Generalised least squares
OLS	Ordinary least squares
Pdf	Probability distribution function
PUB	Prediction at ungauged basins
RD	Reliability diagram
SVM	Support vector machine
RPS	Ranked probability score
VIF	Variation inflation factor
WLS	Weighted least squares

NOTATION

α	Scale parameter of the GEV distribution
β	Regression parameters
ε_m	Measurement uncertainties of observed AMF
ε_s	Sampling uncertainty of flood quantile estimations at gauged sites
ζ	Location parameter of the GEV distribution
η	Regression model uncertainty of flood quantile estimations at ungauged sites
Π_X	Parents of variable X
π_x	Values of parents of variable X
ρ	Spearman's rho test
σ^2_{obs}	Variance of ε_m
σ^2_T	Variance of ε_s
σ^2_R	Variance of η
τ	Kendall's tau test
χ^2	Chi-square statistics

doi: 10.2166/hydro.2013.065

A	Basin area
d_j	Observed probability of jth interval
J	Number of intervals of a discretised event in a BN
j	jth interval of a discretised event in a BN
k	Shape parameter of the GEV distribution
H	Conditional entropy
H_{m}	Mean height of basin
h	Number of catchments
$K_n(w)$	Empirical distribution of the nth W variable
$K_{\theta_n}(w)$	Theoretical distribution introduced by a copula of the nth W variable
m	Number of catchment descriptors
P_T	Annual maximum 24-hours rainfall for a given T-year return period
P_j	Probability computed by a BN to jth interval
Q_{ci}	Observed flood at the ith gauged site
Q_T	Flood quantile for the T-year return period
R^2	Coefficient of determination
R^2_{adj}	Adjusted coefficient of determination
$\mathcal{S}_n$	Cramer–von Mises statistic
s	Length of an AMF series
$\mathcal{T}_n$	Kolmogorov–Smirnov statistic
T	Return period
W	Random variable
w	Value of the variable W
X	Random variable
$\boldsymbol{X}_{\mathrm{d}}$	Matrix of physiographic and climatic descriptors of a basin
x	Value of the variable X
Y	Random variable
y	Value of the variable Y
y_T	Vector of log-transformed flood quantiles for a given T-year recurrence interval
$\overline{y_T}$	Mean of y_T
$\widehat{y_T}$	Vector of predicted values of y_T

INTRODUCTION

Flood frequency analyses estimate the frequency of occurrence of a given flood event. They require observed data to conduct a statistical analysis, assuming that flood events are independent and identically distributed (Rao & Hamed 2000). However, observed streamflow series are usually short and at-site analyses present large uncertainties, mainly for high return periods. A regional approach is often used to augment the short records with available data from other sites, assuming they all have a similar frequency distribution. This leads to more accurate quantile estimations than at-site studies (Hosking & Wallis 2005).

Observed data are only available at gauged sites. Nevertheless, prediction in ungauged basins (PUB) is often required. It entails three steps: (i) identification of homogeneous regions; (ii) estimation of regional quantiles at gauged sites for the return period of interest; (iii) use of a regional method to transfer the known information at gauged sites to ungauged basins (Sarhadi & Modarres 2011).

In Spain, a regional flood frequency analysis has been conducted recently (Jiménez-Álvarez *et al*. 2012). Mainland Spain was divided into 36 homogeneous regions following geographical boundaries. Homogeneity was tested by the Wiltshire, and Hosking and Wallis measures (Wiltshire 1986; Hosking & Wallis 2005). Several frequency distributions and different parameter estimation methods were evaluated for estimating the flood frequency curve. The best results were obtained by the generalised extreme value (GEV) distribution estimated by the L-moments method. Estimation of flood quantiles was improved by using historical flood records at sites where this information was available. The regional shape parameter method was selected to estimate the frequency distribution, estimating the scale and location parameters with at-site information. A linear regression model was selected to estimate quantiles at ungauged sites, relating log-transformed flood quantiles to log-transformed catchment descriptors. Several uncorrelated descriptors were used, such as basin area, slope, mean height, basin perimeter, mean annual rainfall and precipitation quantile, among others.

Regression models are often used in PUB studies. Multivariate regression models that relate a hydrological variable to a set of climatic and physiographic characteristics is the most common technique (Kjeldsen & Jones 2009). The parameters of the regression model can be estimated by different methods. Ordinary least squares (OLS) assumes that observations are homoscedastic and independently distributed, which means that sampling uncertainty at gauged sites is not taken into account. Weighted least squares

(WLS) uses the variance as a surrogate of at-site sampling uncertainty (Tasker 1980). The generalised least squares (GLS) technique assumes heteroscedasticity and cross-correlation of residuals, taking into account the correlation between sites (Stedinger & Tasker 1985). In the last decade, Reis *et al*. (2005) developed a Bayesian approach analysis of a GLS technique (BGLS), which improves the estimation of the model error variance. Kjeldsen & Jones (2010) improved the GLS technique including the correlation of model errors with distance in the regression model.

In addition, flood frequency analyses present many sources of uncertainty that are propagated through the estimation process (Merz & Thieken 2005; Alvisi & Franchini 2013). These can be classified into two categories: random uncertainties that represent the natural variability of the system and epistemic uncertainties that represent the incomplete knowledge of the system (Ferson & Ginzburg 1996; Hall 2003; Apel *et al*. 2004). In the case of PUB, epistemic uncertainties can arise from measurement errors in observations, sampling errors, regression model errors, inability of the regression model to simulate the relationship between the hydrological variable and the climatic and physiographic variables, among others.

Regression models provide prediction uncertainty either following a normal distribution with variance equal to the variance of the residuals of the model, in the case of traditional regression models, or following an empirical distribution, in the case of the BGLS technique. The latter gives a constant uncertainty as a mixture of both the regression coefficients and model error empirical distributions, regardless of the values of the dependent variables. In addition, both independent and dependent variables are considered as deterministic and only one realisation is considered to establish the relationship between flood quantiles and basin physiographic characteristics. Furthermore, regression models cannot take into account uncertainties from measurement errors or in the estimation of the climatic and physiographic variables, such as the precipitation. These uncertainties are seldom propagated through the estimation process of flood quantiles at ungauged sites and therefore they are often underestimated.

In this paper, a Monte Carlo experiment is proposed to investigate how uncertainties propagate through the PUB process. As a second step, a data-driven model is built to reproduce the relationships between the variables used in the regression model. The most used data-driven models in hydrology are the artificial neural networks (ANNs), the Bayesian networks (BNs), the M5 model trees and the support vector machines (SVMs) (Solomatine & Dulal 2003; Corzo *et al*. 2009).

An ANN is a parallel-distributed information processing system composed of simple processing units called neurons which work in an interconnected and parallel way. The result is a deterministic non-linear model which has been widely applied in modelling some hydrological processes, such as rainfall–runoff, streamflows, water quality or precipitation estimations (Govindaraju 2000). Dawson *et al*. (2006) applied ANNs for deterministic flood quantile estimation at ungauged sites in the UK. Shu & Burn (2004) used an ANN ensemble size of 20 networks for deterministic estimations of 10-year flood quantiles in the UK. In both cases, inputs to the network were catchment descriptors and the output was the *T*-year flood quantile.

The M5 model trees are inspired by the modular modelling. The problem to model is divided into simple sub-tasks whose solutions are combined to solve the initial problem. Sub-tasks are modelled by linear regression equations. M5 models have been applied to flood forecasting (Solomatine & Xue 2004) and rainfall–runoff modelling (Solomatine & Dulal 2003), among other applications.

SVMs were initially developed to solve classification problems. However, their applications have been extended to regression problems. SVM employs the structural risk minimisation principle. The goal of support vector regression is to map the input data in a higher dimensional feature space, in which data may be approximately linear, and then find in that space a function that can best approximate the output with a given error tolerance. SVMs have been applied for flood forecasting (Yu *et al*. 2006) and rainfall–runoff modelling (Elshorbagy *et al*. 2010), among other applications.

The BN model was selected and was trained on the results of the experiment. A BN is a directed acyclic graph that infers the joint probability distribution of several related variables from observations through nodes, which represent random variables, and links, which represent causal dependencies between them (Charniak 1991). BNs have a

probabilistic nature which allows them to be a suitable expert system tool for encoding uncertainty (Heckerman 1997). BNs are useful for cases involving a high level of uncertainty or where the relationships between variables are non-linear and complex (Chen & Pollino 2012), among other situations. Moreover, BNs can be used for diagnostic or explanatory purposes, i.e. given the catchment descriptors, the flood quantile is estimated by the network or given the flood quantile and several catchment descriptors, the remaining descriptors are predicted.

BNs have been applied widely and successfully to many scientific fields such as medicine, informatics, industry, biology or meteorology (Cano *et al*. 2004; Cinicioglu *et al*. 2012). However, applications in the field of hydrology are more recent, being used under two approaches: supporting decision systems and forecasting flood events. Molina *et al*. (2005) developed a tool based on BNs that provides support in making decisions about hydraulic actions during floods. BNs were a useful tool thanks to their ability to cope with imperfect and incomplete information. Ames *et al*. (2005) modelled a watershed management decision system by means of BNs. Farmani *et al*. (2009, 2012) proposed a methodology powered by Bayesian belief networks and evolutionary multi-objective optimisation as a decision-making tool which took into account uncertainties. The tool was successfully applied to a real groundwater management case. Mediero *et al*. (2007) presented a methodology to support the decision of the best reservoir operation strategy during flood events based on BNs. Garrote *et al*. (2007) presented a flash flood forecasting model based on a BN model trained on the results of a deterministic rainfall–runoff model, estimating the probability of occurrence of a flood event in a much shorter time than a rainfall–runoff model.

This paper presents a methodology to propagate the uncertainties that arise in the process of predicting flood quantiles at ungauged basins which traditionally is performed by deterministic techniques. A procedure for propagating uncertainties using traditional deterministic models and Monte Carlo simulations is proposed. First, uncertainties in the estimation of the regional quantiles at gauged sites are taken into account by a Monte Carlo based procedure. Then, an ensemble of regression equations is fitted to propagate the uncertainties through the regression model. Finally, a BN model is used to capture the uncertainty of the whole process, taking advantage of its probabilistic nature. The result is a probability distribution of the flood quantile at any ungauged site in the region. The ability of the model to reproduce the uncertainties in the whole process is assessed. This methodology allows the user to (a) propagate uncertainties from flood observations to the estimated flood quantiles, (b) quantify the influence of each source of uncertainty and (c) find out which part of the process needs to be improved for obtaining more reliable estimations. In addition, this methodology can be applied to develop flood discharge mappings that supply a more accurate description of the uncertainties in flood quantile predictions. In recent years, some national discharge maps have been developed, such as the national flood discharge mapping in Austria (Merz *et al*. 2008) and the map of maximum flows in Spain (Jiménez-Álvarez *et al*. 2012). These maps give simplified information about the prediction uncertainty. BNs can overcome this weakness by supplying more accurate estimation of the uncertainties in short computation times.

This paper is organised as follows. The next section introduces the methodology which is followed by a section which presents the case study where the methodology is applied. Results are then described before conclusions are presented in the final section.

METHODOLOGY

The methodology shows how uncertainties in the PUB process are propagated by means of the following steps (Figure 1). First, annual maximum flood (AMF) series recorded at gauged basins are disturbed by measurement uncertainties via a Monte Carlo simulation. Several samples of AMF series are generated and a flood frequency curve is fitted to each sample, obtaining an ensemble of flood frequency curves. Regional flood quantiles are computed for a given return period. Secondly, the ensemble of quantiles is disturbed by sampling uncertainties which are also introduced by a Monte Carlo simulation. Thirdly, an ensemble of regression models is fitted. Finally, a BN is trained on the ensemble of regression models. As BNs need to be trained with a large enough data set, a stochastic generator

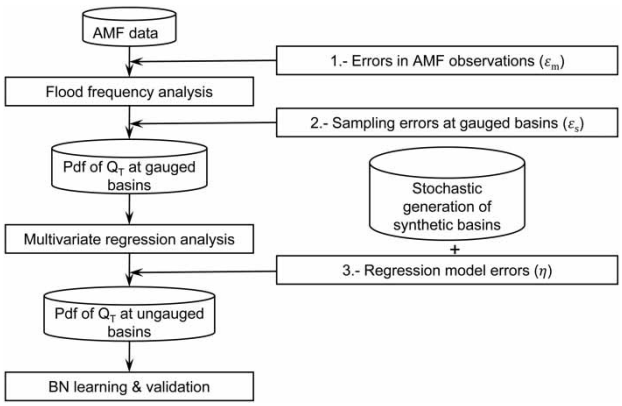

Figure 1 | Flow chart of the methodology (Pdf – probability distribution function).

of synthetic basins is used to extend the data set of real basins. A copula model is used to maintain the statistical properties of the observed catchment descriptors.

Uncertainty analysis

The most important uncertainties when estimating flood quantiles at ungauged sites are summarised in Table 1. Three main sources of uncertainty are accounted for: (a) measurement uncertainties of observed annual maximum floods (ε_m); (b) sampling uncertainties of flood quantile estimations at gauged sites (ε_s); and (c) regression model uncertainties of flood quantile estimations at ungauged sites (η).

Table 1 | Sources of uncertainty in the PUB process

Analysis	Random uncertainty	Epistemic uncertainty
Flood frequency estimation	Flood variability	Measurement errors in the observations Assumptions: stationarity, independency and non-seasonality of AMF series Selection of a distribution function Parameter estimation method Sampling uncertainty: length of time series Regional technique
Regression model	Natural variability of climatic descriptors	Inability of the regression model to simulate the relationship between the dependent and independent variables Method to estimate the regression coefficients Regression model errors

For the sake of simplicity uncertainties from assumptions of stationarity, independency and non-seasonality of AMF series are not taken into account in this study. The regional study carried out by Jiménez-Álvarez *et al.* (2012) provided the most suitable distribution function, parameter estimation method and method to estimate the coefficients of the regression model for each homogeneous region. Additionally, the most appropriate regional technique was selected. As a result, uncertainties derived from these sources were neglected.

Finally, the uncertainties from the regression model were not considered, as the scope of this paper is focused on proving the ability of a BN model to represent the probability distributions of the uncertainty in flood predictions. Therefore, the simple OLS regression model used currently in Spain was selected.

Flood frequency analysis

Streamflow discharges are estimated indirectly from observations of water levels by a rating curve. Errors can arise from measurements of the water level, transformations of the water level into discharge by a rating curve and changes in the rating curve over time. The total uncertainty from this process (ε_m) is assumed to follow a Gaussian distribution with zero mean and variance equal to σ^2_{obs}. σ^2_{obs} can be parameterised as a function of the observed flow (q_{obs}) and a constant (ε_{obs}), implying larger uncertainties for higher flows:

$$\sigma^2_{obs} = (\varepsilon_{obs} q_{obs})^2 \quad (1)$$

ε_{obs} usually takes a value of 0.1 for high discharges (Clark *et al.* 2008).

In Spain, flood quantiles at a given gauged site are estimated by a three-parameter GEV distribution (Equation (2)) using the L-moments method with a regional shape parameter (Jiménez-Álvarez *et al.* 2012). The flood quantile for a given T-year return period (Q_T) can be obtained by Equation (3),

$$F(x) = \exp\left\{-\left(1 - k\frac{x-\zeta}{\alpha}\right)^{1/k}\right\} \quad (2)$$

$$Q_T = \zeta + \frac{\alpha}{k}\left\{1 - \left[-\ln\left(1 - \frac{1}{T}\right)\right]^k\right\} \quad (3)$$

where ζ is the location parameter, α is the scale parameter and k is the shape parameter.

The uncertainty associated with the estimation of Q_T or sampling uncertainty (ε_s) is usually expressed as a confidence interval, whose mean is the computed value of the quantile and its lower and upper bounds are given by a Gaussian distribution with zero mean and variance σ_T^2. The asymptotic variance of the three-parameter GEV quantile for the case of a regional shape parameter is given by Equation (4) (Lu & Stedinger 1992):

$$\sigma_T^2 = \alpha^2 \frac{b_1 + b_2 z + b_3 z^2}{s} \quad (4)$$

where:

$$z = \begin{cases} \left\{1 - [-\ln(1 - 1/T)]^k\right\}, & k \neq 0 \\ -\ln[-\ln(1 - 1/T)], & k = 0 \end{cases}$$
$$b_1 = s\mathrm{Var}(\zeta)/\alpha^2 \quad (5)$$
$$b_2 = 2s\mathrm{Cov}(\zeta, \alpha)/\alpha^2$$
$$b_3 = s\mathrm{Var}(\alpha)/\alpha^2$$

where s is the length of the AMF series, Cov is the covariance and Var is the variance.

Uncertainties are propagated through the flood frequency analysis at each gauged site as follows. First, an ensemble of randomised samples of each record in AMF series (Q_{ci}) is generated disturbing the observation by ε_m (Figure 2(a)). Then, a flood frequency curve is fitted to each randomised AMF series, obtaining an ensemble of flood frequency curves for each gauged site (Figure 2(b)). Afterwards, each Q_T from the ensemble of flood frequency curves is randomised by ε_s. Finally, the probability distribution of Q_T at each gauged site is obtained (Figure 2(c)).

The uncertainty propagation of ε_m and ε_s is shown in Figure 2(c). The effect of both errors on flood quantile estimation is an increase of the variance of its probability distribution.

Multivariate regression analysis

Flood quantiles at ungauged basins located within a hydrological homogeneous region are usually estimated by a multivariate regression analysis, which relates flood quantiles to physiographic and climatic features of the basins (Equation (6)):

$$y_T = \boldsymbol{X}_d \beta + \eta \quad (6)$$

where y_T is a vector of log-transformed flood quantiles for a given T, $\boldsymbol{X}_d$ is a matrix of physiographic and climatic descriptors with a first column of unity, β is a vector with the regression model coefficients and η is the error of the regression model.

The regression coefficients (β) are estimated by the OLS method (Equation (7)), as it was used in the regional analysis carried out in Spain recently (Jiménez-Álvarez *et al*. 2012). In

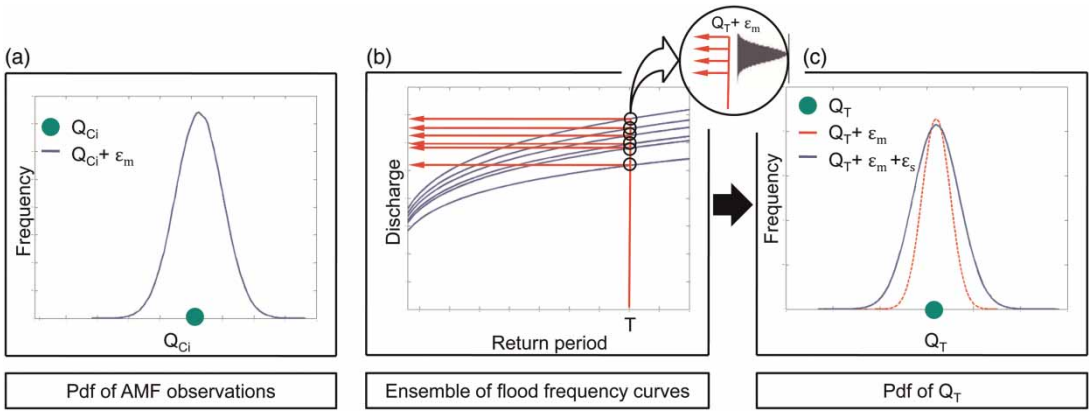

Figure 2 | Uncertainty propagation through the flood frequency analysis.

this case, the standard error of prediction of the regression model (σ_R) is given by the variance of the residuals (Equation (8)).

$$\beta = \left(\boldsymbol{X}_d^T \boldsymbol{X}_d\right)^{-1} \boldsymbol{X}_d^T y_T \tag{7}$$

$$\sigma_R = \sqrt{\frac{\left(y_T - \widehat{y_T}\right)^2}{h - m - 1}} \tag{8}$$

where $\widehat{y_T}$ is a vector of predicted values of y_T, h is the length of vector y_T and m is the number of catchment descriptors used in the regression model.

Goodness of fit of the multivariate regression model is measured by the coefficient of determination (R^2) and the adjusted coefficient of determination (R^2_{adj}), given by Equations (9) and (10), respectively. Multicollinearity is quantified by the variance inflation factor (VIF) (Equation (11)). A value larger than five indicates that multicollinearity exits.

$$R^2 = 1 - \frac{y_T - \widehat{y_T}}{y_T - \overline{y_T}} \tag{9}$$

$$R^2_{adj} = 1 - (1 - R_2)\frac{h - 1}{h - m - 1} \tag{10}$$

$$VIF = \frac{1}{1 - R_k^2} \tag{11}$$

where $\overline{y_T}$ is the mean of y_T and R_k^2 is the coefficient of determination between the k^{th} catchment descriptor and the remaining m^{-1} catchment descriptors used in the regression model.

Generally, a regression model is fitted to a set of deterministic values of y_T (Figure 3(a)). However, in the case of the present paper the values of y_T are probabilistic, as a result of the uncertainty analysis carried out above (Figure 3(b)). In order to propagate uncertainties through the regression model, an ensemble of regression models were fitted to an ensemble of vectors y_T (Figure 3(c)). Quantiles at ungauged sites are estimated by the ensemble of regression equations, randomising the results by the regression modelling error (η), which is assumed to follow a Gaussian distribution with zero mean and variance σ_R^2 (Equation (8)).

If quantiles are computed by the traditional model regression, only uncertainties from the regression model are considered. However, the proposed methodology can deal with more sources of uncertainty and, consequently, the variance of the probability distribution of quantiles increases (Figure 3(d)).

Flood quantile estimation by Bayesian networks

Once the propagation of the uncertainty in the PUB process was developed, flood quantiles could be estimated at any ungauged site in the homogeneous region. The result is not only a deterministic value, but a probability distribution that takes into account the uncertainties in the process. In order to avoid replicating the process every time we need to estimate a quantile at an ungauged site, a BN was used to learn the variability of all the process in terms of conditional probabilities.

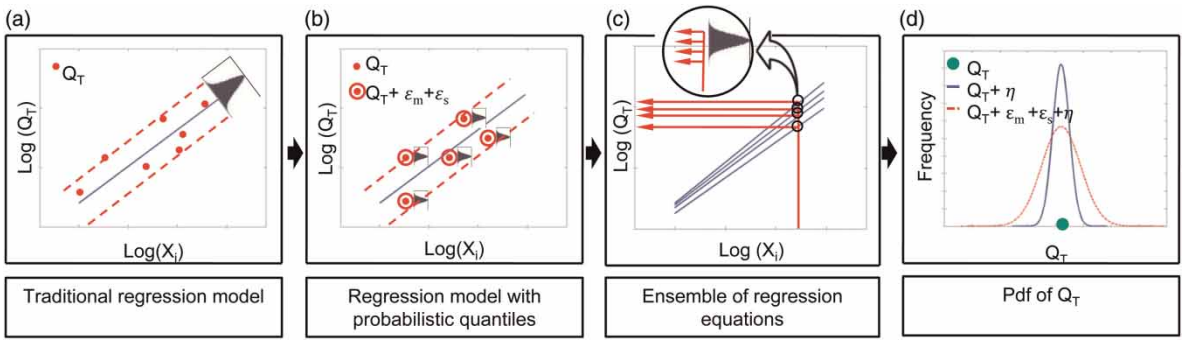

Figure 3 | Uncertainty propagation through the regression model.

BNs are directed acyclic graphs composed of nodes and links. Nodes represent random variables and links represent causal dependencies among them. The strength of the relationships between variables is represented by conditional probability distributions attached to each variable, which can be inferred from observations or synthetic data. A BN estimates the conditional probabilities of a set of nodes given the observations or evidence at the rest of nodes, using the Bayes' theorem. Evidence can be propagated downward or upward, so that the BNs can be used for diagnostic or explanatory purposes. More details can be found in Ames *et al.* (2005).

The joint probability distribution of a set of N quantitative random variables, assuming independence, is given by:

$$P(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} P(x_i|\pi_{x_i}) \tag{12}$$

where x_i is the quantitative value of the i^{th} variable X_i, $P(x_1, x_2, \ldots x_N)$ is the joint probability, π_{x_i} are the quantitative values of the set of causes or parents of variable X_i and $P(x|y)$ is the conditional probability of x given y is known (Castillo *et al.* 1997).

Training of BNs involves two steps: structure learning, which entails choosing a network given by a set of variables and links between them, from a number of possible networks, and parameter learning or fitting, which obtains the parameters of the chosen network from observed data (Buntine 1996).

The sample length used in the training process has a significant effect on the results given by the BN. Small training data sets can lead to incomplete trained networks, while large data sets can lead to an overfitting. Consequently, the result of the training process can be quantified by means of the conditional entropy (H), which assesses the joint probability distribution of each combination of values and the probability distribution computed by the network after the training process (Equation (13)) (Cooper & Herskovits 1992):

$$H_{X|\pi_x} = -\sum_{\Pi_X=\pi_x} P(\Pi_X = \pi_x) \sum_{X=x} P(X = x|\Pi_X = \pi_x) \times \ln P(X = x|\Pi_X = \pi_x) \tag{13}$$

where X is any variable of the network, Π_X is the set of the parents of the variable X and π_x is a set of given values of the parents of X.

H quantifies the remaining entropy of a random variable and is a measure of its uncertainty (Ihara 1993). Consequently, its value can be neither high nor low. The evolution of H over the length of the training data set was analysed, with the optimal size being reached when H approaches a constant value.

Once a BN is trained, the validation process is conducted to assess its quality. There are several validation measures. In this study the reliability diagram (RD) and the ranked probability score (RPS) were selected.

Reliability describes how often an observation has occurred given a particular prediction (Wilks 2000). RDs are used to evaluate the prediction reliability, showing the forecast probability against the observed relative frequency. A perfect reliability is represented by the main diagonal, where the events predicted with a given probability, p, are observed with that same probability, p.

The RPS is used to assess the overall prediction performance of the probability forecast (Franz & Sorooshian 2002). RPS is obtained from Equation (14),

$$\text{RPS} = \frac{1}{J-1} \sum_{m=1}^{J} \left(\sum_{j=1}^{m} P_j - \sum_{j=1}^{m} d_j \right)^2 \tag{14}$$

where J is the number of bins in which the variable has been discretised, P_j is the probability given by the BN for the bin j and d_j is the observed frequency for the bin j. Values of RPS can lie within the interval [0, 1]. A zero value is a perfect prediction and a one value is the worst prediction. Therefore, a prediction will be better if its value is close to zero.

Stochastic generator of synthetic basins

Since real catchment descriptors are obtained from a digital terrain model (DTM), the number of real ungauged basins in a region is limited by the cell size of the DTM. In certain cases, a BN could need to be trained with a larger data set than that obtained from the DTM. Consequently, a stochastic generator of synthetic basins was developed. Dependencies between physiographic and climatic variables in the real

region were maintained. A copula model was used, as copulas have the ability to generate synthetic data keeping the marginal and joint statistical properties of the observed variables, such as the basin area, mean height or mean slope.

A copula between a pair of variables (X, Y) is a multivariate distribution that reproduces the dependence relationships between the two variables, regardless of their marginal distributions (Sklar 1973). Copulas are classified into families, with the elliptical and the Archimedean being the most used. In the last two decades, copulas have spread in many fields, such as finance, biomedical studies and engineering (Yan 2007). They have been recently used in hydrology (Salvadori & De Michele 2007; Mediero *et al.* 2010; Requena *et al.* 2013).

The copula family that best represents the dependence properties of the observed data was selected. Among the several families of copulas that have been proposed in the literature, the Archimedean have received much attention in hydrological analyses for their large variety and ease of construction (Nelsen 2006). Therefore, this copula family was used in this study. Archimedean copulas are divided into several subfamilies, with the Clayton, Frank and Gumbel being the most used. The selection process is composed of several steps. First, independence tests are conducted to determine whether there is a significant association between the variables. The Spearman's rho (ρ) and the Kendall's tau (τ) tests were used.

Secondly, the most suitable copula is selected. The choice of a particular subfamily is based on goodness of fit tests, such as the Cramer–von Mises ($\mathcal{S}_n$) (Equation (15)), or the Kolmogorov–Smirnov ($\mathcal{T}_n$) (Equation (16)),

$$\mathcal{S}_n = \int_0^1 |\mathbb{K}_n(w)|^2 K_{\theta_n}(w)\, \mathrm{d}w \qquad \mathbb{K}_n(w) = \sqrt{n}\{K_n(w) - K_{\theta_n}(w)\} \tag{15}$$

$$\mathcal{T}_n = \max_{0 \le w \le 1} |\mathbb{K}_n(w)| \tag{16}$$

where $K_n(w)$ is the empirical distribution of the variables $(W_1, \ldots, W_n)$, $K_{\theta_n}(w)$ is the theoretical distribution introduced by the copula and the variables $(W_1, \ldots, W_n)$ are a transformation of the pairs (X_i, Y_i).

Lower values of the goodness of fit tests lead to better fittings of the copula. Consequently, the copula with the lowest value of the statistics will be selected. Afterwards, a large number of synthetic basins can be randomised with the chosen copula model.

CASE STUDY

The homogeneous region 32 was selected from the regional flood frequency analysis conducted in mainland Spain (Jiménez-Álvarez *et al.* 2012). This region is located in the Tagus basin in the central part of Spain (Figure 4(a)). In the above-mentioned study, 26 gauged sites were used in

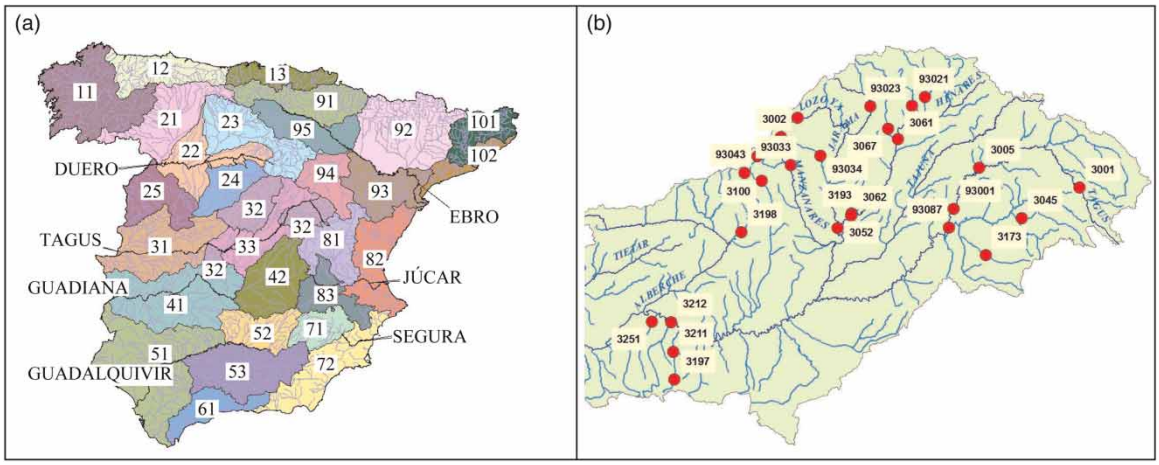

Figure 4 | (a) Homogeneous regions considered in Jiménez-Álvarez *et al.* (2012). (b) Red points show the gauging sites located in region 32. Please refer to the online version of this paper to see this figure in colour: http://www.iwaponline.com/jh/toc.htm.

this region which are represented by red points in Figure 4(b). The GEV distribution was selected with a given regional shape parameter equal to −0.1273. A multivariate regression analysis was also carried out, using the following catchment descriptors: the basin area in km^2 (A), the annual maximum 24-hour rainfall for a given T recurrence interval in mm (P_T) and the average height of the basin above mean sea level in m (H_m). Regression parameters (β) were estimated by the OLS method (Equation (7)). Therefore, only uncertainties associated with flood quantile estimation via the regression model (η) were assessed in that study, assuming that errors follow a normal distribution with zero mean and variance given by σ_T^2. The results of R^2, R^2_{adj} and VIF statistics are listed in Table 2.

A set of eight gauged basins was selected to show how uncertainties propagate through the PUB process. Catchment descriptors of these basins are shown in Table 3 and their location is depicted in Figure 4(b). A set of 11,000 descriptors of real basins were obtained in region 32 from a digital terrain model with a cell size of 500 m, in order to calibrate the copula model for generating synthetic basins.

RESULTS AND DISCUSSION

The proposed methodology was applied to the case study. The highest return period considered in Jiménez-Álvarez *et al.* (2012), 500 years, was selected to assess the methodology, since sampling uncertainties increase as the return period does. Table 3 shows flood quantiles for the 500-year return period (Q_{500}) estimated by the deterministic regression model at the eight gauging sites with the regression parameters of Table 2.

Flood frequency analysis

An ensemble of 100,000 members of AMF series was generated at each gauging station in the 32 region. AMF observed series were disturbed by ε_m using a Monte Carlo simulation. A three-parameter GEV distribution with a regional shape parameter was fitted to each member of the ensemble of disturbed AMF series. Consequently, an ensemble of 100,000 flood frequency curves was generated at each gauging station. Afterwards, an ensemble of Q_{500} was computed at the gauging stations, which was disturbed by ε_s through the asymptotic variance associated with each flood frequency curve (σ_T). Probability distributions of flood quantiles disturbed by ε_m and ε_s are shown in black dotted lines in Figure 6. For the selected sites, the deterministic estimations of Q_{500} by the flood frequency analysis can be compared to the probability distributions of Q_{500}, computed by the Monte Carlo experiment, which considered the uncertainties in the process. The influence of each single uncertainty is measured by the standard deviation of the quantiles disturbed by each source. The variance of the quantile distribution disturbed by ϵ_m and ε_s is listed in Table 4. Sampling uncertainties of flood quantile estimations are higher than the measurements uncertainties in all selected sites.

Multivariate regression analysis

Once an ensemble of 100,000 Q_{500} was estimated at each gauged site, a regression model was fitted to each ensemble member. The catchment descriptors used in the previous study were maintained, in order to compare the improvement

Table 2 | Summary statistics of the multivariate regression analysis carried out in Jiménez-Álvarez *et al.* (2012)

	Regression parameters							VIF		
T	Intercept (β_0)	ln [A] (β_1)	ln [P_T] (β_2)	ln [H] (β_3)	σ_R^2	R^2	R^2_{adj}	ln [A]	ln [P_T]	ln [H]
2	− 4.344	0.679	0.843	0.936	0.0167	0.887	0.872	2.038	3.169	1.833
5	− 2.982	0.657	0.857	0.553	0.0130	0.923	0.912	1.935	3.112	1.893
10	− 2.397	0.642	0.775	0.445	0.0127	0.932	0.923	1.885	3.105	1.932
25	− 1.796	0.623	0.647	0.369	0.0130	0.933	0.924	1.823	3.075	1.971
100	− 0.830	0.601	0.458	0.229	0.0139	0.938	0.930	1.735	2.999	2.002
500	0.303	0.575	0.193	0.103	0.0158	0.931	0.922	1.715	3.010	2.033

Table 3 | Catchment descriptors and Q_{500} from the deterministic regression model for the selected gauged sites

Code	A (km^2)	P_{500} (mm)	H_m (m)	Q_{500} (m^3/s)
3002	44	168	1,753	85
93041	9	187	1,658	25
3100	236	165	1,153	299
3197	62	121	963	179
3211	299	115	777	381
3251	224	124	646	208
93021	282	132	1,101	259
3193	258	106	784	215

of considering the uncertainties that arise in the process. 100,000 regression equations and their σ_R were obtained. Consequently, probabilistic regression model coefficients (β) were obtained, instead of deterministic ones. Figure 5 depicts the a posteriori probabilistic distribution of these coefficients. As can be appreciated, the effect of the uncertainties over the coefficients consists of an increase of its variance. At a given basin with known physiographic features, the ensemble of regression models provides an ensemble of deterministic flood quantiles, which were disturbed by η. Finally, a probability distribution at the given basin was obtained.

The probability distribution of Q_{500} at the eight selected gauged sites listed in Table 2 is depicted in Figure 6. Deterministic estimations via the flood frequency analysis without considering uncertainties are represented by green points. Probability distributions of quantiles estimated by the deterministic multivariate regression analysis affected by η are depicted in solid blue lines while probability distributions of quantiles computed by the ensemble of regression models affected by ε_m, ε_s and η errors are plotted in solid red lines. The standard deviation of the disturbed quantiles by η errors is listed in Table 4. As single uncertainties cannot be added, they partially compensate each other when combined.

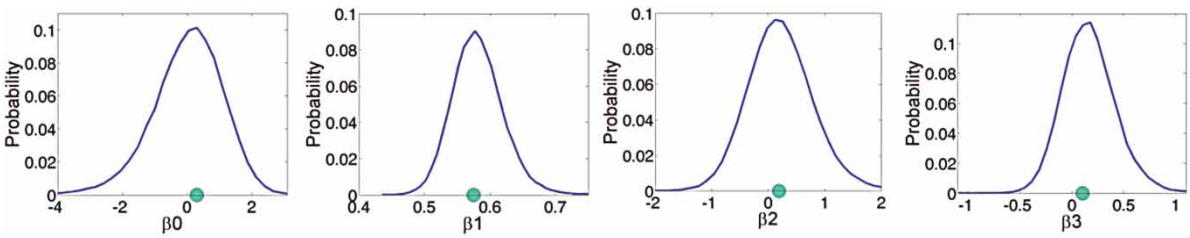

Figure 5 | Probability distributions of the regression coefficients (β). Solid lines represent the results of the ensemble of regression models and points the traditional regression model.

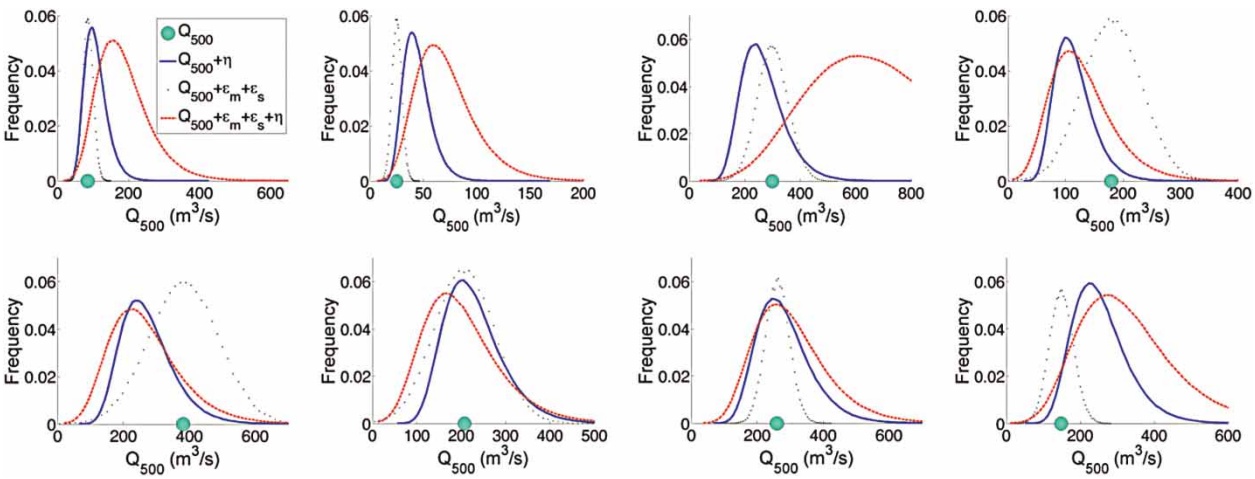

Figure 6 | Flood quantile estimation by deterministic flood frequency analysis (Q_{500}), by the regression model with uncertainties ($Q_{500}+\eta$) and by the regression model with the three sources of uncertainty considering the study ($Q_{500}+\varepsilon_m+\varepsilon_s+\eta$) at eight gauged sites. From left to right and top to bottom: 3002, 93041, 3100, 3197, 3211, 3251, 93021 and 3193 site. Please refer to the online version of this paper to see this figure in colour: http://www.iwaponline.com/jh/toc.htm.

Table 4 | Standard deviation of the disturbed quantiles for the selected gauged sites

Code	σ_{ε_m}	σ_{ε_s}	σ_η	$\sigma_{\varepsilon_m}+\sigma_{\varepsilon_s}+\sigma_\eta$
3002	3.180	14.173	31.214	44.349
93041	0.971	4.488	12.862	18.092
3100	10.843	54.560	77.418	115.557
3197	7.353	50.915	33.280	45.414
3211	14.088	104.777	79.151	104.864
3251	9.905	67.774	66.600	89.471
93021	6.325	35.635	81.843	108.963
3193	10.360	74.909	71.437	98.660

Estimations via the ensemble of regression models have a greater variance than the others, as the ensemble of regressions takes into account more sources of uncertainties. The main sources of uncertainty are caused by ε_s and η errors. In some sites the largest error comes from the regression model, while in others the largest uncertainty is derived from the flood frequency analysis. Uncertainties increase as the basin area does in most cases. However, there is no clear relationship between uncertainties and the remaining catchment descriptors.

In some gauging stations, the probability distributions when considering η are not centred with the value of quantiles estimated by the flood frequency analysis. This is caused by the regression model that gives an estimation of the mean hydrological behaviour in the region (Figure 3(a)).

Stochastic generation of synthetic basins

As the 11,000 real triplets of catchment descriptors that represent real ungauged basins in the selected region were inadequate to train the BN model, a larger set of synthetic basins was generated by means of two copulas fitted to the real basins. The first copula models the dependency between H_m and A. The second one between H_m and P_{500}. The variable H_m has the same values in both sets.

As a first step, independency tests between the variables were conducted, which are based on the ρ and τ statistics. Their results are shown in Table 5. The high values of the *p*-value indicate that the null hypothesis may be rejected. Therefore both copula models can be used to simulate the observed data.

Table 5 | Results of the Spearman's rho and the Kendall's tau statistics for a confidence level of 5%

	ρ	*p*-value	τ	*p*-value
(H_m, A)	0.10	10.57	0.06	10.26
(H_m, P_{500})	0.70	73.11	0.50	77.98

The Clayton, Frank and Gumbel families of Archimedean copulas were fitted to the observed data. The choice of the best model is based on the $\mathcal{S}_n$ and $\mathcal{T}_n$ statistics, whose values are listed in Table 6. The most suitable copula model for both cases is the Clayton copula. Lower values of the statistics indicate a better fit of the copula to the observed data.

An ensemble of 100,000 synthetic basins was randomised from both fitted copulas, giving combinations of the three catchment descriptors as a result. Cumulative probability distributions of the synthetic data set were compared to those of the real data (Figure 7). It can be seen that cumulative distributions of synthetic basins are quite similar to those of observed basins for the three catchment descriptors.

Flood quantiles at the synthetic basins were estimated by the ensemble of regression models. For a given synthetic basin, 10 regression equations were chosen randomly from the ensemble of regression models and 10 plausible flood quantiles were estimated at the basin. The result was disturbed by uncertainties from the regression model. A sample of 10,000,000 Q_{500} at 100,000 basins was generated at the end of the process, in order to have a large enough data set to be used in the training process of a BN.

Prediction in ungauged basins by Bayesian networks

Flood quantiles were estimated by the BN depicted in Figure 8. Inputs are the variables P_T, A and H_m. The output is the variable Q_T. Dependencies between A, H_m

Table 6 | Results of $\mathcal{S}_n$ and $\mathcal{T}_n$ statistics applied to (H_m, A) and (H_m, P_{500}) set

		Clayton	Frank	Gumbel
(H_m, A)	$\mathcal{S}_n$	0.52	1.56	2.32
	$\mathcal{T}_n$	1.96	2.29	2.75
(H_m, P_{500})	$\mathcal{S}_n$	6.76	9.36	19.57
	$\mathcal{T}_n$	4.80	5.92	7.57

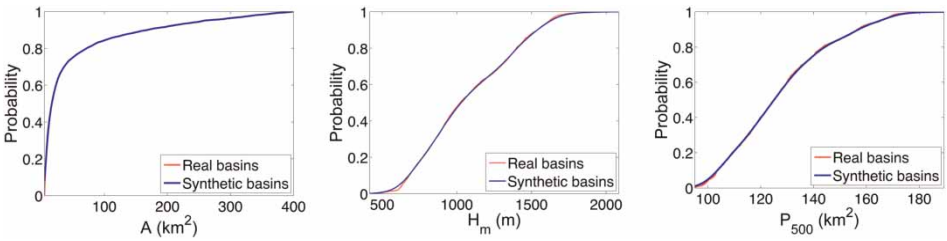

Figure 7 | Comparison of cumulative probability distributions of synthetic data and real basins.

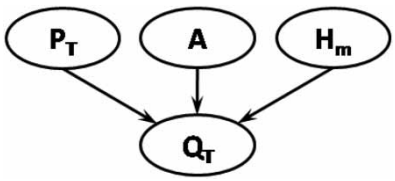

Figure 8 | BN structure.

and P_T were neglected as the ability of the network to reproduce the uncertainties were made comparing BN to the multivariate regional regression analysis performed in Jiménez-Álvarez *et al.* (2012), which does not take into account relationships between inputs.

Variables in a BN can be either continuous, providing probability distributions by a given distribution function, or discrete, providing probability distributions by conditional probability tables (CPT). As the uncertainty propagation in the PUB process is assumed not to follow a Gaussian distribution, a discrete BN was chosen. Consequently, variables were discretised into intervals. The number of intervals was determined by a trial and error process. The most accurate predictions were provided by binning P_T, A and Q_T into 25 equally spaced containers and H_m into 20. Since the topology of the network employed in this study has no hidden nodes and the training data set will be large enough thanks to the use of synthetic data, its parameters will be estimated by the maximum likelihood method, which is the simplest and quickest technique. The BN was trained and validated with the ensemble generated in the previous section, being split into two groups. The first group, named training set, had a length of 9,000,000 samples. The second group, named validation set, had a length of 1,000,000 samples.

The size of the training data set has a significant effect over the BN response. Therefore, the minimum size of the training data set was selected by means of the H measure (Figure 9). It can be seen that the asymptotic value of 0.27 is reached when the data set has a length of 2,000,000 samples. Consequently, the BN must be trained on a minimum data set of 2,000,000 samples. However, better results could be achieved with a larger data length. Three training data sets of 2,000,000; 5,000,000 and 9,000,000 members were used to study the BN behaviour in terms of the validation measures.

The trained BN was validated by means of both RD and RPS using the validation data set. The RD compares the observed frequency of an event with its prediction probability computed by the BN. The more similar both variables are, the more reliable the network is. Figure 10 shows that as the size of the training set increases, the prediction probability is more similar to the observed frequency, except for the highest probabilities where a deterioration occurs for lengths of 2,000,000 and 5,000,000. The BNs trained with 9,000,000 and 5,000,000 samples improve slightly the overall result compared to the one trained with 2,000,000 samples.

The RPS measures the prediction performance of the BN. Its results are shown in Table 7. As the training set size increases, the RPS value is closer to zero, meaning that the trained BN has a better prediction performance. It

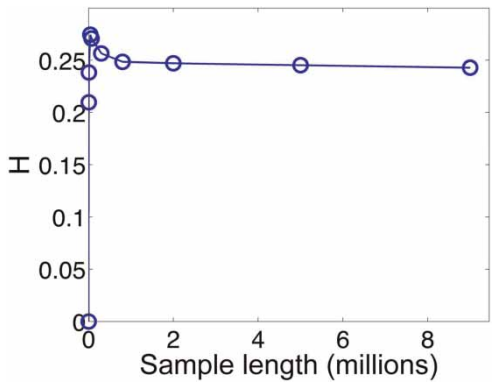

Figure 9 | Sensitivity analysis of the sample length used in the training process.

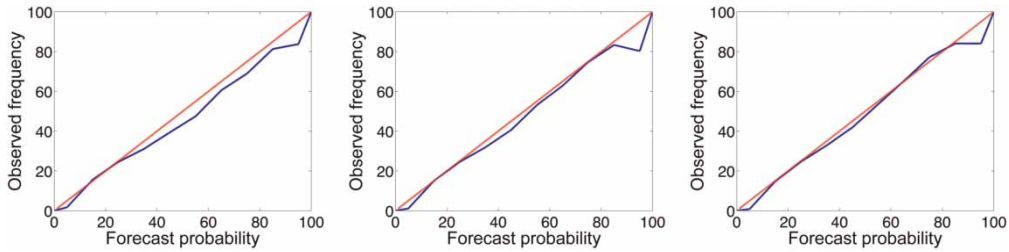

Figure 10 | Reliability diagram results for different training sample lengths. From left to right: 2,000,000; 5,000,000 and 9,000,000 samples.

Table 7 | Results of the RPS

Number of samples	2,000,000	5,000,000	9,000,000
RPS	0.1150	0.0550	0.0402

can be seen that the BN trained with 5,000,000 improves the results of the BN trained with 2,000,000 samples, but the BN trained on 9,000,000 slightly improves the results with 5,000,000 samples. Accordingly, the BN trained with 5,000,000 was selected, as it showed a better performance than the BN trained with 2,000,000 members. On the other hand, the BN trained with 9,000,000 members does not lead to a significant improvement.

The selected BN was used to compute flood quantiles at the eight selected gauged sites and at eight synthetic basins generated by the copula model (Tables 3 and 8). The flood quantile probability distributions computed by the ensemble of regression models and disturbed by the three sources of uncertainty considered (ε_m, ε_s and η) and the BN outputs were tested to come from the same population by means of the Chi-square test at the 5% significance level. Both estimations are plotted in Figures 11 and 12 and results of the Chi-square statistic χ^2 and the p-value of the test are provided in Table 9. Values of the test indicated that flood quantiles estimated with the ensemble and the BN models are samples from the same population, with both probability distributions being quite similar. However, the BN model computed the probability distributions in a much shorter computation time than the ensemble.

Table 8 | Catchment descriptors of eight synthetic basins

Code	A (km^2)	P_{500} (mm)	H_m (m)
S1	41	135	1,351
S2	104	140	795
S3	151	113	813
S4	230	105	705
S5	15	174	1,754
S6	94	101	728
S7	314	132	1,402
S8	368	97	569

CONCLUSIONS

Prediction of flood quantiles at ungauged sites is usually carried out by a multivariate regression analysis, which relates quantiles to physiographic and climatic catchment descriptors. However, some of the regression models assume a linear relationship between deterministic variables.

In this paper, uncertainty propagation through the PUB process was studied. Uncertainties from flood measurement errors, sampling uncertainties from flood frequency estimations and uncertainties from regression model estimations were introduced in the analysis. The effect of each uncertainty source was quantified. Sampling and modelling errors were found to be the main sources of uncertainty. A large ensemble of regression models was obtained to account for these uncertainties. In addition, the ability of Bayesian networks to deal with uncertainties when flood quantiles are estimated at ungauged basins was assessed. However, Bayesian networks need a large data set in the training process. For this purpose, a set of synthetic basins were generated by means of copula models, which maintain the statistical properties of observed data.

A Bayesian network model was applied to a case study in the Tagus basin in Spain. The topology of the Bayesian network was composed of the annual maximum 24-hour rainfall for a given return period, the area of the basin and the mean height of the basin, as input variables. The

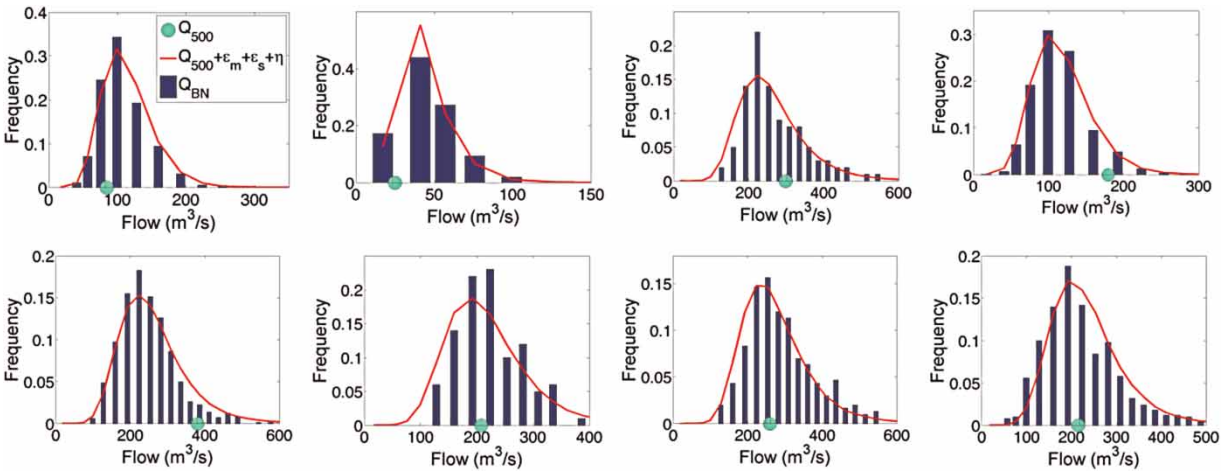

Figure 11 | Flood quantile estimations at eight gauged sites by deterministic flood frequency analysis (Q_{500}), by propagating the uncertainty in the PUB process ($Q_{500}+\varepsilon_m+\varepsilon_s+\eta$) and by the BN ($Q_{BN}$). From left to right and top to bottom: 3002, 93041, 3100, 3197, 3211, 3251, 93021 and 3193 site.

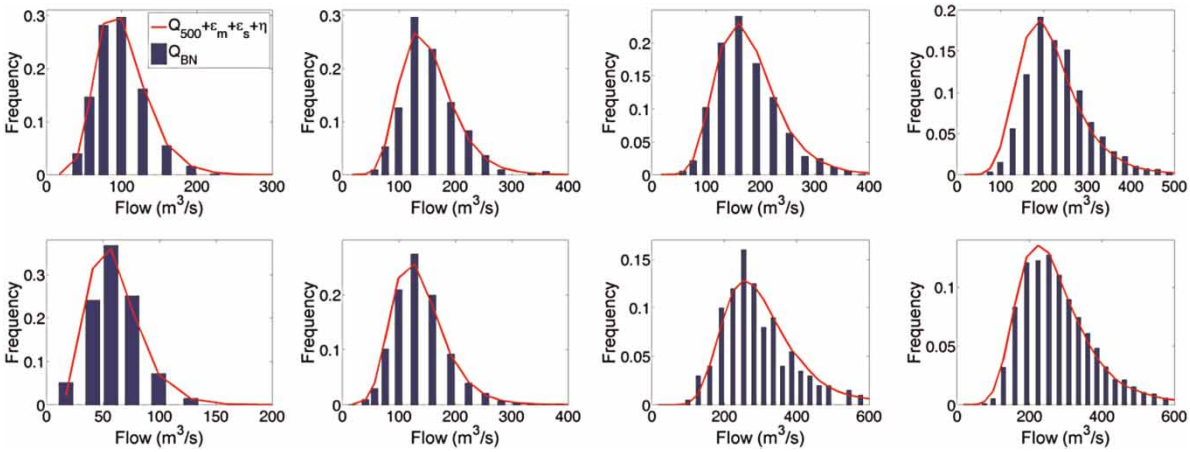

Figure 12 | Flood quantile estimation at eight synthetic basins by propagating the uncertainty in the PUB process ($Q_{500}+\varepsilon_m+\varepsilon_s+\eta$) and by the BN ($Q_{BN}$). From left to right and top to bottom: S1 to S8 basin.

Table 9 | χ^2 statistic and *p*-values of the Chi-square test

Gauged sites			Synthetic basins		
Code	χ^2	***p*-value**	**Code**	χ^2	***p*-value**
3002	0.0160	0.8995	S1	0.0005	0.9822
93041	0.0201	0.8872	S2	0.0016	0.9682
3100	0.0006	0.9813	S3	0.0041	0.9491
3197	0.0001	0.9918	S4	0.0417	0.8382
3211	0.0147	0.9036	S5	0.0055	0.9407
3251	0.0164	0.8982	S6	0.0035	0.9528
93021	0.0125	0.9111	S7	0.0216	0.8832
3193	0.0471	0.8281	S8	0.0054	0.9417

output is the flood quantile probability distribution for a given return period.

The Bayesian network was trained on the ensemble of regression models obtained as a result of the uncertainty propagation process and applied to the large set of synthetic basins generated for this purpose. Afterward, the Bayesian network was validated by the reliability diagram and the rank probability score. The analysis showed that a network trained on 5,000,000 samples carried out good predictions, with high reliability and great prediction performance. Also, the comparison of the Bayesian network output to the result of the uncertainty propagation experiment showed

that this tool can deal with the uncertainties considered in this process. In addition, the Bayesian network model computes the probability distributions in a much shorter computation time than the ensemble of regression models.

This methodology can be useful for improving national flood discharge mappings, where simplified information about the prediction uncertainty is usually provided. More information would require replicating the Monte Carlo procedure presented in this paper each time a user asks for information at a given point. Therefore, a tool based on this procedure would be hampered by an excessive computing time. The proposed methodology based on BNs can supply the probability distribution of the uncertainty in flood predictions in a very short time.

Finally, Bayesian networks can deal with more sources of uncertainties, such as uncertainties in the input variables. The uncertainty in the estimation of the annual maximum 24-hour rainfall could be included if available. In addition, more sophisticated regression models could be introduced to improve the representation of regression model errors if they were available in the future.

In conclusion, Bayesian networks are an efficient tool for flood quantile estimation at ungauged basins. They can reproduce uncertainties from several sources and can propagate them efficiently.

ACKNOWLEDGEMENTS

The authors acknowledge support from the MODEX project (CGL2011-22868) 'Physically-based modelling for characterisation of extreme hydrologic response under a probabilistic approach. Application to dam safety analysis and optimisation of reservoir operation during floods', funded by the Spanish Ministry of Economy and Competitiveness, and from COST action FLOODFREQ (ES0901) 'European Procedures for Flood Frequency Estimation'.

REFERENCES

Alvisi, S. & Franchini, M. 2013 A grey-based method for evaluating the effects of rating curve uncertainty on frequency analysis of annual maxima. *Journal of Hydroinformatics* **15** (1), 194–210.

Ames, D., Neilson, B., Stevens, D. & Lall, U. 2005 Using Bayesian networks to model watershed management decisions: an East Canyon Creek case study. *Journal of Hydroinformatics* **7**, 267–282.

Apel, H., Thieken, A. H., Merz, B. & Blöschl, G. 2004 Flood risk assessment and associated uncertainty. *Natural Hazards and Earth System Sciences* **4**, 295–308.

Buntine, W. 1996 A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering* **8** (2), 195–210.

Cano, R., Sordo, C. & Gutiérrez, J. M. 2004 Applications of Bayesian networks in meteorology. *Studies in Fuzziness and Soft Computing* **146**, 309–328.

Castillo, E., Gutiérrez, J. M. & Hadi, A. S. 1997 *Expert Systems and Probabilistic Network Models*. Springer Verlag, New York, NY, USA, 251 pp.

Charniak, E. 1991 Bayesian networks without tears. *AI Magazine* **12** (4), 50.

Chen, S. H. & Pollino, C. A. 2012 Good practice in Bayesian network modelling. *Environmental Modelling & Software* **37**, 134–145.

Cinicioglu, E. N., Önsel, S. & Ülengin, F. 2012 Competitiveness analysis of automotive industry in Turkey using Bayesian networks. *Expert Systems with Applications* **39** (12), 10923–10932.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J. & Uddstrom, M. J. 2008 Hydrological data assimilation with the ensemble Kalman filter: use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources* **31** (10), 1309–1324.

Cooper, G. F. & Herskovits, E. 1992 A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9** (4), 309–347.

Corzo, G. A., Solomatine, D. P., Wit, M., Werner, M., Uhlenbrook, S. & Price, R. K. 2009 Combining semi-distributed process-based and data-driven models in flow simulation: a case study of the Meuse river basin. *Hydrology and Earth System Sciences* **13** (9), 1619–1634.

Dawson, C. W., Abrahart, R. J., Shamseldin, A. Y. & Wilby, R. L. 2006 Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology* **319** (1), 391–409.

Elshorbagy, A., Corzo, G., Srinivasulu, S. & Solomatine, D. P. 2010 Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 2: application. *Hydrology and Earth System Sciences* **14**, 1943–1961.

Farmani, R., Henriksen, H. J. & Savic, D. 2009 An evolutionary Bayesian belief network methodology for optimum management of groundwater contamination. *Environmental Modelling & Software* **24** (3), 303–310.

Farmani, R., Henriksen, H. J., Savic, D. & Butler, D. 2012 An evolutionary Bayesian belief network methodology for participatory decision making under uncertainty: an application to groundwater management. *Integrated Environmental Assessment and Management* **8** (3), 456–461.

Ferson, S. & Ginzburg, L. R. 1996 Different methods are needed to propagate ignorance and variability. *Reliability Engineering & System Safety* **54** (2), 133–144.

Franz, K. J. & Sorooshian, S. 2002 Verification of national weather service probabilistic hydrologic forecasts. Report of the Department of Hydrology and Water Resources of The University of Arizona, Tucson, AZ, EE.UU, pp. 3–8.

Garrote, L., Molina, M. & Mediero, L. 2007 Probabilistic forecasts using Bayesian networks calibrated with deterministic rainfall–runoff models. In: *Extreme Hydrological Events: New Concepts for Security* (O. F. Vasiliev, P. H. A. J. M. van Gelder, E. J. Plate & M. V. Bolgov, eds). Springer, Dordrecht, The Netherlands, pp. 173–183.

Govindaraju, R. S. 2000 Artificial neural networks in hydrology. II: hydrology applications. *Journal of Hydrologic Engineering* **5** (2), 124–137.

Hall, J. 2003 Handling uncertainty in the hydroinformatic process. *Journal of Hydroinformatics* **5**, 215–232.

Heckerman, D. 1997 Bayesian network for data mining. *Data Mining and Knowledge Discovery* **1** (1), 79–119.

Hosking, J. R. M. & Wallis, J. R. 2005 *Regional Frequency Analysis: An Approach Based on L-moments*. Cambridge University Press, Cambridge, UK, pp. 1–11.

Ihara, S. 1993 *Information Theory for Continuous Systems*. World Scientific Pub Co Inc, Singapore, vol. 2, pp. 1–56.

Jiménez-Álvarez, A., García-Montañés, C., Mediero, L., Inicio, L. & Garrote, J. 2012 El mapa de caudales máximos de las cuencas intercomunitarias. *Revista De Obras Públicas* **159** (3533), 7–32.

Kjeldsen, T. R. & Jones, D. A. 2009 An exploratory analysis of error components in hydrological regression modeling. *Water Resources Research* **45**, W02407.

Kjeldsen, T. R. & Jones, D. A. 2010 Predicting the index flood in ungauged UK catchments: on the link between data-transfer and spatial model error structure. *Journal of Hydrology* **387** (1–2), 1–9.

Lu, L. H. & Stedinger, J. R. 1992 Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test. *Journal of Hydrology* **138** (1–2), 223–245.

Mediero, L., Garrote, L. & Martín-Carrasco, F. 2007 A probabilistic model to support reservoir operation decisions during flash floods. *Hydrological Sciences Journal* **52** (3), 523–537.

Mediero, L., Jimenez-Álvarez, A. & Garrote, L. 2010 Design flood hydrographs from the relationship between flood peak and volume. *Hydrology and Earth System Sciences* **14** (12), 2495–2505.

Merz, B. & Thieken, A. H. 2005 Separating natural and epistemic uncertainty in flood frequency analysis. *Journal of Hydrology* **309** (1–4), 114–132.

Merz, B., Blöschl, G. & Humer, G. 2008 National flood discharge mapping in Austria. *Natural Hazards* **46**, 53–72.

Molina, M., Fuentetaja, R. & Garrote, L. 2005 Hydrologic models for emergency decision support using Bayesian networks. In: *Lecture Notes in Computer Sciences. Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (L. Godo, ed.). Springer-Verlag, Barcelona, Spain, vol. 3571, pp. 88–99.

Nelsen, R. B. 2006 *An Introduction to Copulas*. Springer Verlag, New York, NY, USA, pp. 109–155.

Rao, A. R. & Hamed, K. H. 2000 *Flood Frequency Analysis*. CRC Press LLC, Boca Raton, FL, USA, pp. 1–16.

Reis Jr, D. S., Stedinger, J. R. & Martins, E. S. 2005 Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. *Water Resources Research* **41**, W10419.

Requena, A. I., Mediero, L. & Garrote, L. 2013 Bivariate return period based on copulas for hydrologic dam design: comparison of theoretical and empirical approach. *Hydrology and Earth System Sciences Discussion* **10**, 557–596.

Salvadori, G. & De Michele, C. 2007 On the use of copulas in hydrology: theory and practice. *Journal of Hydrologic Engineering* **12** (4), 369–380.

Sarhadi, A. & Modarres, R. 2011 Flood seasonality-based regionalization methods: a data-based comparison. *Hydrological Processes* **25** (23), 3613–3624.

Shu, C. & Burn, D. H. 2004 Artificial neural networks ensembles and their application in pooled flood frequency analysis. *Water Resources Research* **40**, W09301.

Sklar, A. 1973 Random variables, joint distribution functions and copulas. *Kybernetika* **9** (6), 449–460.

Solomatine, D. P. & Dulal, K. N. 2003 Model trees as an alternative to neural networks in rainfall–runoff modelling. *Hydrological Sciences Journal* **48** (3), 399–411.

Solomatine, D. P. & Xue, Y. 2004 M5 model trees and neural networks: applications to flood forecasting in the Upper Reach of the Huai River in China. *Journal of Hydrologic Engineering* **9** (6), 491–501.

Stedinger, J. R. & Tasker, G. D. 1985 Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared. *Water Resources Research* **21** (9), 1421–1432.

Tasker, G. D. 1980 Hydrologic regression with weighted least squares. *Water Resources Research* **16** (6), 1107–1113.

Wilks, D. S. 2000 Diagnostic verification of the climate prediction center long-lead outlooks, 1995–98. *Journal of Climate* **13** (13), 2389–2403.

Wiltshire, S. E. 1986 Identification of homogeneous regions for flood frequency analysis. *Journal of Hydrology* **84**, 287–302.

Yan, J. 2007 Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software* **21** (4), 1–21.

Yu, P., Chen, S. & Chang, I. 2006 Support vector regression for real-time flood stage forecasting. *Journal of Hydrology* **138** (3), 704–716.

First received 22 May 2013; accepted in revised form 2 October 2013. Available online 30 November 2013