Advances in Water Resources 63 (2014) 91–103

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Copula models for frequency analysis what can be learned from a Bayesian perspective?

Eric Parent [a], Anne-Catherine Favre [b,*], Jacques Bernier [a,c], Luc Perreault [c]

[a] Equipe Modélisation, Risques, Statistique, Environnement, UMR 518 INRA-MIA, 16 rue Claude Bernard, 75005 Paris, France
[b] Université Grenoble-Alpes, Ecole Nationale Supérieure de l'Energie, l'Eau et l'Environnement, Institut national polytechnique de Grenoble, Laboratoire d'étude des Transferts en Hydrologie et Environnement, Bâtiment OSUG-B, Domaine universitaire, BP 53, 38041 Grenoble cedex 09, France
[c] Institut de Recherche d'Hydro-Québec, 1800, boulevard Lionel Boulet, Varennes (Québec), J3X 1S1, Canada

ARTICLE INFO

Article history:
Received 7 December 2010
Received in revised form 19 August 2013
Accepted 28 October 2013
Available online 16 November 2013

Keywords:
Frequency analysis
Bayesian inference
Copula
Romaine River
Model selection

ABSTRACT

Large spring floods in the Québec region exhibit correlated peakflow, duration and volume. Consequently, traditional univariate hydrological frequency analyses must be complemented by multivariate probabilistic assessment to provide a meaningful design flood level as requested in hydrological engineering (based on return period evaluation of a single quantity of interest). In this paper we study 47 years of a peak/volume dataset for the Romaine River with a parametric copula model. The margins are modeled with a normal or gamma distribution and the dependence is depicted through a parametric family of copulas (Arch 12 or Arch 14). Parameter joint inference and model selection are performed under the Bayesian paradigm. This approach enlightens specific features of interest for hydrological engineering: (i) cross correlation between margin parameters are stronger than expected , (ii) marginal distributions cannot be forgotten in the model selection process and (iii) special attention must be addressed to model validation as far as extreme values are of concern.

1. Introduction

In several applied statistical areas, like hydrology, the analysis of extreme events is of particular interest. Estimation of quantiles for different characteristics of hydrological events such as spring floods is standard practice in civil engineering. In fact, the choice of an acceptable and cost-effective solution for the design of hydraulic structures and optimal reservoirs operation depend upon such quantities. For instance, critical structures like spillways and dams are generally designed to withstand a flood peak and a flood volume with a given small prespecified exceedance probability.

Generally, quantile estimation is achieved by fitting univariate probability distributions to historical flood peaks and flood volumes. Of course, since these two random variables are correlated, univariate analyses should not be performed independently. In this case, a bivariate analysis is more appropriate. In fact, an univariate analysis can lead either to an underestimation of risk [11], which may introduce tragic consequences, or to an overestimation of an event severity [51], which could lead to unnecessary preemptive spilling or increase in building costs.

Another hydrological application where a multivariate quantile estimation is clearly needed concerns the modeling of the combined risk upstream of the confluence of several rivers (see e.g. [16]) or the summation of several sub-watershed in series. For many applications, the peakflow results indeed from the combination of flows for many intermediate watersheds. In this case, the dependence between flows must be taken into account to predict the joint hydrological behavior correctly.

A problem in the application of the multivariate quantile estimation comes from the concept of *return period* commonly used in univariate analysis of extreme flood events. In multivariate analysis, the return period has no longer a unique definition depending on whether we consider the intersection or the union of two events [51]. In other words, the quantiles corresponding to a given return period cannot map to a unique percentile value. For example, in two dimensions, the quantiles are represented by an iso-curve [45]. Recently Salvadori and De Michele [46] introduced the concept of Kendall's return period in order to solve the problem of uniqueness of the return period.

This paper aims to demonstrate the advantages of addressing quantile estimation of flood peak and flood volume using a Bayesian bivariate model. The dependence structure between the two random variables modeling is here specified by a copula model. The main idea underlining copula construction given in [48] is to separate marginal distributions from the dependence structure, which allows to work with a broader class of parametric multivariate distributions as explained by Renard and Lang [41]. Standard frequentist estimates commonly take advantage of this separation by first fitting independently density functions to each marginal

* Corresponding author.
E-mail address: Anne-Catherine.Favre-Pugin@ense3.grenoble-inp.fr (A.-C. Favre).

0309-1708/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.advwatres.2013.10.013

sample, and then by calibrating the copula function (using for instance the inverse cumulative density functions as in [18]). In a fully Bayesian perspective however, statistical inference, prior elicitation and model selection are treated coherently in a global probabilistic framework at the cost of a deeper involvement in the computation of probability and a fair understanding of its operational interpretation. As a consequence, information from the whole dataset can be jointly transferred to the various components of the Bayesian model (i.e. margins + copula + priors) during the inferential step. Correlations then emerge among Bayesian estimates since they rely on the same pieces of information. Such links may seem surprising for many modelers used to the standard approach, but bring useful information for hydrological engineering:

1. Cross correlation between margin parameters might be stronger than expected.
2. Marginal distributions cannot be forgotten in the model selection process since Bayesian inference considers margins, copula and priors as components which interact together within the model structure.
3. Finally, after data assimilation, the global Bayesian probabilistic framework allows to integrate out the parameter uncertainty and yields predictive distributions for new data. Because of explicit account of parameter partial knowledge, such Bayesian predictive pdf's corresponds to mixtures which can be more dispersed than their frequentist counterparts, for which parameter point estimates are directly plugged into the likelihood function. Therefore the Bayesian predictive approach helps temper overconfident prediction (underdispersed predictive pdf) and therefore can lead to more cautious decisions under uncertainty.

Hydro-Québec's new hydroelectric complex on the Romaine River – which is located in the Eastern part of the province of Québec, Canada- serves as a case study to illustrate the approach and the previous keypoints.

Section 2 provides a description of the hydroelectric complex on the Romaine River and of the data set. Section 3 is devoted to multivariate models in hydrology and parametric copulas. Bayesian inference is developed in Section 4, namely the choice of priors (Section 4.1), the posterior analyses as if the margins were independent (Section 4.2) and the joint estimation of parameters for margins and copulas model (Section 4.3). In Section 5, we consider Bayesian model selection with some implementation issues. Concluding comments are given in Section 7.

2. The Romaine River as a motivating case study

The case study developed throughout this paper concerns the bivariate analysis of peak flow and volume of the Romaine River.

Hydro-Québec plans to build a hydroelectric complex on the Romaine River in the lower North shore of the Saint-Laurent River. The complex will include four important developments located along the first 200 [km] of the river, with an installed capacity of approximately 1,500 megawatts. Project construction will take place from 2009 to 2020. The complex will have an average annual production capacity of 8.0 TWh. Each power plant will include a rockfill dam, a flood spillway, a power plant with two turbine-alternator [3].

Romaine River rises in Lower Northern Québec and flows over a distance of approximately 490 [km] draining into Saint-Laurent Seaway. The localization of the Romaine River is illustrated in Fig. 1.

This basin has numerous lake and rivers and covers an area of $14,470$ [km^2]. Hydro-Québec plans to operate the Romaine River water-resource system with an interesting hydroelectric power generation potential.

Two hydrometric stations have been operated on the Romaine River. The first one (ID 073801) is located a few kilometers upstream the outlet of the river. The spring flood is governed by snowmelt and occurs each year, in average, between the end of March and the end of June. For this station, with a watershed of $13,106$ [km^2], spring peak flood and corresponding volumes are available from 1957 to 2005. For this case study, the volume of each year corresponds to the maximum water discharge based upon a fixed duration of 52 days occurring between March 1st and June 30th. The peak flow is simply the maximum value within the beginning and the end of the corresponding 52 days flood event. This approach is the standard one used for flood frequency analysis at Hydro-Québec. The series consist of $n = 47$ observations, since the data for 1960 are missing. The second station (ID 73802) is located in the upstream part of the river (with a waterbasin of 6675 [km^2]) and was managed for only 10 years (1972–1981). This additional information is used to specify the prior distributions in Section 4.1.

Fig. 2 shows the co-evolution of the two variables of interest (peak flow and volume) during the study period. A careful look suggests that the two variables are related to one another. To confirm this hypothesis, scatterplots of ranks are drawn. The rank plots of Fig. 3 clearly reveal the presence of positive dependence in the pairs (peak, volume).

One might also compute the common Pearson's correlation, r_n, and base a test of independence thereupon. The conclusion is dubious because it is based on the assumption of bivariate normality, which turns out to be inappropriate in this case. A better way to proceed consists of computing Spearman's rank correlation, ρ_n, or Kendall's coefficient of concordance, τ_n, and to base a test of independence on these distribution-free statistics. The values of the three coefficients of dependence and the corresponding P-values for the test of independence are reported in Table 1. As could be expected, the two quantities of interest are highly positively cross-correlated.

Note that Fig. 2 suggests a change of hydrologic regime around 1985. Concomitant breaking points in the peak and volume series of the Romaine River have been observed on most streams in the Northeastern part of the Québec-Labrador Peninsula [39,40,49,23]. Of course, possible concomitant breaking points in the series may also artificially increase the natural correlation. This phenomenon, which might reflect recent possible climatic variations, could be modeled by introducing a temporal structure such as a Markov chain within a mixture model setting [13]. However, in this paper, emphasis is put onto bivariate analysis and we assume that the individual time series are stationary and, since a permutation test for a lag one time correlation gives p-values of respectively 3.9% and 1.1% for the peak and the volume series, we also assume that they exhibit no time auto-correlation (see Section 5.1 in [20]). Accordingly, when considered separately they can be assimilated to random samples from univariate distributions and when considered jointly to independent replicates of a bivariate distribution. Preliminary studies with Generalized Pareto (the appropriate model for values over a threshold) or Generalized Extreme Value (the theoretical limiting behavior of block maxima) did not indicate a better fit than with simpler distributions like normal and gamma pdfs. Peak and volume will then be respectively modeled using gamma or normal marginal distributions. Although not grounded on theoretical models of extremes, such distributions are candidates of common statistical use with the convenience of conjugates for Bayesian analysis. Note that the sample skewness of the volume is 0.0673 showing that this sample can be considered as symmetric. The next section proposes to make recourse to copula models to depict their joint behavior.