Journal of Hydrology 527 (2015) 704–717

Contents lists available at ScienceDirect

Journal of Hydrology

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

Handling uncertainty in bivariate quantile estimation – An application to flood hazard analysis in the Mekong Delta

Nguyen Viet Dung [a,c,*], B. Merz [a,d], A. Bárdossy [b], H. Apel [a]

[a] GFZ German Research Centre for Geosciences, Section 5.4 Hydrology, Potsdam, Germany
[b] Institute of Hydraulic Engineering, University of Stuttgart, Stuttgart 70569, Germany
[c] Southern Institute of Water Resources Research SIWRR, Ho Chi Minh City, Viet Nam
[d] Institute of Earth and Environmental Science, University Potsdam, Germany

ARTICLE INFO

Article history:
Received 7 March 2015
Received in revised form 12 May 2015
Accepted 15 May 2015
Available online 22 May 2015
This manuscript was handled by Corrado Corradini, Editor-in-Chief

Keywords:
Bivariate flood quantile
Copulas
Uncertainty estimation
Bootstrapping
Mekong Delta

SUMMARY

The hydrological load causing flood hazard is in many instances not only determined by peak discharge, but is a multidimensional problem. While the methodology for multivariate frequency analysis is well established, the estimation of the associated uncertainty is rarely studied. In this paper, a method is developed to quantify the different sources of uncertainty for a bivariate flood frequency analysis. The method is exemplarily developed for the Mekong Delta (MD), one of the largest and most densely populated river deltas worldwide. Floods in the MD are the basis for the livelihoods of the local population, but they are also the major hazard. This hazard has, however, not been studied within the frame of a probabilistic flood hazard analysis. The nature of the floods in the MD suggests a bivariate approach, because the societal flood severity is determined by both peak discharge and flood volume. The uncertainty caused by selection of statistical models and parameter estimation procedures are analyzed by applying different models and methods. For the quantification of the sampling uncertainty two bootstrapping methods were applied. The developed bootstrapping-based uncertainty estimation method shows that large uncertainties are associated with the estimation of bivariate flood quantiles. This uncertainty is much larger than the model selection and fitting uncertainty. Given the rather long data series of 88 years, it is concluded that bivariate flood frequency analysis is expected to carry significant uncertainty and that the quantification and reduction of uncertainty merit greater attention. But despite this uncertainty the proposed approach has certainly major advantages compared to a univariate approach, because (a) it reflects the two essential aspects of floods in this region, (b) the uncertainties are inherent for every bivariate frequency analysis in hydrology due to the general limited length of observations and can hardly be avoided, and (c) a framework for the quantification of the uncertainties is given, which can be used and interpreted in the hazard assessment. In addition it is shown by a parametric bootstrapping experiment how longer observation time series can reduce the sampling uncertainty. Based on this finding it is concluded that bivariate frequency analyses in hydrology would greatly benefit from discharge time series augmented by proxy or historical data, or by causal hydrologic expansion of time series.

1. Introduction

Flood risk is often not only determined by peak discharge, but is a multidimensional problem. Examples are the joint occurrence of flood discharge at river confluences (Wang et al., 2009; Bender et al., 2013), the superposition of river flooding and storm surges at coasts (Kew et al., 2013), or the important role of flood duration, besides peak discharge, for dike failure (Vorogushyn et al., 2010) and for flood losses (Merz et al., 2013). While the methodology for multivariate frequency analysis is well established, the estimation of the associated uncertainty is rarely studied. In this paper, a method is developed to quantify the different sources of uncertainty for the case of bivariate flood frequency analysis.

This method is exemplarily developed for the Mekong Delta (MD), one of the largest and most densely populated river deltas worldwide. Floods occur annually and are the basis of the livelihoods of several million people in Cambodia and Vietnam. The MD is known as rice bowl of South East Asia and has one of the world's most productive fisheries. This high productivity is a

* Corresponding author at: GFZ German Research Centre for Geosciences, Section 5.4 Hydrology, Telegrafenberg, Building C4, room 2.04, 14473 Potsdam, Germany. Tel.: +49 331 288 28784; fax: +49 331 288 1570.
E-mail address: dung@gfz-potsdam.de (N.V. Dung).

http://dx.doi.org/10.1016/j.jhydrol.2015.05.033
0022-1694/

consequence of annual flooding providing large amounts of sediments and nutrients (Manh et al., 2015). Ecosystems and agriculture are well adapted to flooding. However, extreme floods (e.g. flood years 1961, 1971, 2000, 2001, 2002 and 2011) can cause huge damage (Hoa et al., 2007; MRC, 2009) and pose a serious threat to millions of people. Given the enormous relevance of flooding, it is surprising that, to the authors' knowledge, a probabilistic flood hazard assessment for the Mekong Delta does not exist, neither in scientific journals nor in reports from local authorities. The current flood risk management simply uses the most disastrous flood in the recent history, the flood season in 2000, as design event for flood protection.

Besides the annual maximum flood discharge (Q), the flood volume (V) is an equally important factor governing the inundation in the MD. One illustrative example is the devastating flood in 2000. The peak discharge was not extraordinarily high, but the flood volume was the largest recorded in 88 years of observation, causing extended inundation, prolonged water logging and huge damages. Q and V are stochastically correlated which necessitates a bivariate frequency model. Classical bivariate approaches are usually applied when the margins of the two random variables (or their transformations) follow the same family of distributions (e.g. Kelly and Krzysztofowicz, 1997; Yue, 2001; Yue et al., 1999). For example, Adamson et al. (1999) applied the bivariate Gumbel distribution to the gauge Vientiane in Lao PDR in the Mekong River Basin using 79 years of observed discharges.

A more flexible and more general approach is based on copulas. They enable the use of different probability distribution functions for the different variables. Copulas are also able to integrate different types of marginal distributions with time-varying parameters (Joe, 1997). While copulas have frequently been applied in "high-risk" financial and accrual sectors (Cherubini et al., 2004; Embrechts et al., 2003; Patton, 2012), their use in water resources is quite recent, for example for analyzing rainfall data (Singh et al., 2005), flood behavior (Klein et al., 2010; Salvadori and De Michele, 2004), drought analysis (Lee et al., 2013), geostatistical groundwater quality models (Bárdossy, 2006), or spatial interpolation of rainfall (Bárdossy and Pegram, 2013).

In this paper, a copula-based bivariate model for flood frequency analysis of peak flow and volume is developed for Kratie, the gauge which is commonly used as the upstream gauge of the MD. A comprehensive uncertainty analysis is performed by investigating parameter estimation method uncertainty, model selection uncertainty and sampling uncertainty. A method is developed to quantify the sampling uncertainty associated with the bivariate frequency analysis. This uncertainty source is usually ignored in the context of multivariate frequency analysis due to its difficult estimation and interpretation. To the knowledge of the authors there have been just a few studies discussing this important topic (e.g. Serinaldi, 2013; Serinaldi and Kilsby, 2015). In this study it is shown how this particular source of uncertainty can be quantified, and the implications of this uncertainty estimation on bivariate flood frequency analysis are discussed This provides the basis for a suitable statistical framework used for a probabilistic flood hazard assessment for the Mekong Delta, as well as recommendations for its use in practical applications and further improvements. The transferability of the presented approach is also discussed.

2. Methodology

2.1. Copula theory

For the sake of simplicity, only a short theory on copulas is presented below. More details can be found in Joe (1997), Nelsen (2006), Salvadori and De Michele (2004) and Cherubini et al. (2004). By definition, a copula is a multivariate distribution function with uniform margins on the interval [0, 1]. The theory of copulas is based on Sklar's theorem (Sklar, 1959). In the bivariate context which is the focus of this study, it can be written in the form:

$$F_{X,Y}(x,y) = C\{F_X(x), F_Y(y)\} \quad (x,y) \in \mathbb{R}^2, \tag{1}$$

where $F_X(\cdot)$ and $F_Y(\cdot)$ are the marginal cumulative density functions (cdf) of random variables X and Y, respectively, while $F_{X,Y}(\cdot,\cdot)$ is the bivariate joint cdf of X and Y. $C\{\cdot,\cdot\}$ denotes the bivariate copula function which is a mapping from $[0,1]^2$ to $[0,1]$.

A copula sets up a link between the joint distribution and its marginal distribution functions. The bivariate model for two random variables can be uniquely constructed based on chosen marginal distributions and the copula representing the dependence between variables independently. Eq. (1) can be rewritten for the parametric form as:

$$F_{X,Y}(x,y,\theta) = C\{F_X(x,\theta_x), F_Y(y,\theta_y), \theta_C\} \quad (x,y) \in \mathbb{R}^2$$

where $\theta_X, \theta_Y, \theta_C, \theta = (\theta_X, \theta_Y, \theta_C)$ are parameters of the marginal distributions, the copula and the bivariate distribution, respectively.

In order to build up the copula-based bivariate statistical model for frequency analysis for a particular case study comprising two studied random variables X, Y, two fundamental and steps need to be taken: parameter estimation and goodness-of-fit testing. The former is to estimate parameter θ assuming that the fitting distribution belongs to a known distribution class. The selection of a candidate distribution class (marginal, copula, bivariate) for the fitting practically depends on studied variables. The latter is to test the validity of that assumption. Within the context of this study, we do not aim to present a general procedure to accomplish these two complementary tasks. We would rather present the tasks performed in case study in the Mekong Delta. Readers interested in theoretical details are referred to the papers by Genest and Favre (2007) and Genest et al. (2009).

2.2. Quantiles in hydrology

The quantile notion is important in hydrology (Chebana and Ouarda, 2011). It represents a locus of points (quantile set) of considered variables corresponding to a given hazard level p, $p \in (0,1)$. For each p the univariate quantile is well defined by a single real value being simply derived from the inverse of the cumulative distribution function of the studied random variable. However, the multivariate quantile provides an infinite combination of studied variables and the estimation of them depends on the multivariate design of case studies. In this study we are focused on the estimation of bivariate flood quantiles for the Mekong Delta, where the design and quantile estimation is based on the definition of flood events of interest presented in Section 3.4.

2.3. Uncertainty estimation of bivariate quantiles

Any estimation comes along with uncertainty. However, the issue of uncertainty estimation is often overlooked in literature possibly because of its technical difficulty even for univariate analyses. This issue is extensively discussed in the recent paper by Serinaldi (2014). Serinaldi (2014) proposed three algorithms for uncertainty estimation of joint quantiles. In this study we independently developed an alternative bootstrapping based algorithm for this purpose (Algorithm A), and additionally implemented a modified version of ALGO-C of Serinaldi (2014) (Algorithm B), both for