Journal of Hydrology 535 (2016) 559–580

Contents lists available at ScienceDirect

Journal of Hydrology

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

A complete procedure for multivariate index-flood model application

Ana Isabel Requena [a,*], Fateh Chebana [b], Luis Mediero [a]

[a] Department of Civil Engineering: Hydraulics, Energy and Environment, Technical University of Madrid, C/ Profesor Aranguren s/n, 28040 Madrid, Spain
[b] INRS-ETE, 490 rue de la Couronne, Quebec, QC, Canada

ARTICLE INFO

Article history:
Received 30 June 2015
Received in revised form 27 January 2016
Accepted 4 February 2016
Available online 10 February 2016
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Niko Verhoest, Associate Editor

Keywords:
Multivariate regional hydrological frequency analysis
Index-flood model
Prediction at ungauged sites
Copulas
Bivariate return periods
Ebro catchment

SUMMARY

Multivariate frequency analyses are needed to study floods due to dependence existing among representative variables of the flood hydrograph. Particularly, multivariate analyses are essential when flood-routing processes significantly attenuate flood peaks, such as in dams and flood management in flood-prone areas. Besides, regional analyses improve at-site quantile estimates obtained at gauged sites, especially when short flow series exist, and provide estimates at ungauged sites where flow records are unavailable. However, very few studies deal simultaneously with both multivariate and regional aspects. This study seeks to introduce a complete procedure to conduct a multivariate regional hydrological frequency analysis (HFA), providing guidelines. The methodology joins recent developments achieved in multivariate and regional HFA, such as copulas, multivariate quantiles and the multivariate index-flood model. The proposed multivariate methodology, focused on the bivariate case, is applied to a case study located in Spain by using hydrograph volume and flood peak observed series. As a result, a set of volume-peak events under a bivariate quantile curve can be obtained for a given return period at a target site, providing flexibility to practitioners to check and decide what the design event for a given purpose should be. In addition, the multivariate regional approach can also be used for obtaining the multivariate distribution of the hydrological variables when the aim is to assess the structure failure for a given return period.

1. Introduction

Hydrological frequency analyses (HFAs) can be univariate or multivariate, according to the number of variables involved, and local or regional, depending on the number of sites considered. This leads to four frameworks: (i) univariate local; (ii) univariate regional; (iii) multivariate local; and (iv) multivariate regional (MR-HFAs). Many studies have focused on (i) and (ii). However, a multivariate analysis is required, for instance, when routing processes affect floods, such as in dam design and safety analysis and hydraulic modelling in flood-prone areas. The multivariate approach allows improvement of the analysis of a study phenomenon by considering the relation between variables and by using all the available information (e.g., Volpi and Fiori, 2014). Indeed, the multivariate approach is indispensable for performing suitable flood frequency analysis, as the univariate approach can lead to underestimating or overestimating the risk associated which such events (De Michele et al., 2005). Consequently, more attention has recently been paid to (iii) (e.g., Chebana, 2013; Grimaldi and Serinaldi, 2006; Requena et al., 2013; Shiau et al., 2006; Yue et al., 1999). In addition, though MR-HFAs should be considered when complete flood hydrographs need to be characterised (multivariate) at ungauged sites (regional) or when a reduction in uncertainty at gauged sites is sought (especially when a short data length is available), few studies have dealt with (iv) because of its complexity (e.g., Ben Aissia et al., 2015; Sadri and Burn, 2012).

Univariate regional HFAs are usually based on the index-flood model proposed by Dalrymple (1960). As an extension, and focused on the bivariate case, the multivariate index-flood (MIF) model was presented by Chebana and Ouarda (2009), combining copulas with bivariate quantile curves. In addition, the discordancy and homogeneity tests (Hosking and Wallis, 1997) required for the application of the index-flood model were also extended to the multivariate context (Chebana and Ouarda, 2007; Chebana et al., 2009).

The recently performed MR-HFAs mainly involve either theoretical developments or partial applications. Therefore, this study seeks to introduce a complete multivariate regional procedure based on the MIF model that entails all steps required to undertake an exhaustive analysis, providing guidelines. Analogous to the MIF model, the proposed procedure is focused on the bivariate case,

* Corresponding author. Tel.: +34 913365261.
E-mail addresses: ana.requena@caminos.upm.es (A.I. Requena), fateh.chebana@ete.inrs.ca (F. Chebana), luis.mediero@upm.es (L. Mediero).

http://dx.doi.org/10.1016/j.jhydrol.2016.02.004
0022-1694/© 2016 Elsevier B.V. All rights reserved.

analysing hydrograph volume and flood peak observed series. The contributions of the present study consist of including the MIF model proposed by Chebana and Ouarda (2007, 2009) in a complete framework for its application in practice, and developing the method by using a flexible distribution for simulating more accurate synthetic homogeneous regions regarding the multivariate homogeneity test. In addition, copula goodness-of-fit (GoF) tests are used in a specific way for accomplishing the copula selection process in MR-HFAs, and an uncertainty analysis is performed to show the uncertainty decrease related to the multivariate regional approach in comparison with that of the multivariate local approach.

Moreover, several practical contributions are provided. First, the selection process to identify the multivariate regional distribution (MRD) for applying the MIF model is generalised; given that when the model was presented the focus was to evaluate its performance based on simulated regions, using a known copula and marginal distributions. Second, a partial residual analysis is used in relation to the estimate of the index flood, checking linearity in a multiple regression model. Third, the possibility of using different kinds of multivariate return periods (e.g., Salvadori et al., 2011) for estimating quantile curves is highlighted. Fourth, a recent procedure for event selection developed in a multivariate local HFA (Volpi and Fiori, 2012) is included in the regional procedure. As an illustration, the procedure is applied to a region in the Ebro catchment in Spain. The proposed methodology and required background are shown in Section 2. The application of the methodology, consisting of a description of the region studied and flood data, as well as of the results and discussion provided is presented in Section 3. The conclusions obtained are summarised in Section 4.

2. Methodology and background

The proposed complete practical procedure for performing a MR-HFA based on the MIF model consists of four main steps (see the overview in Fig. 1): (i) screening of the data to find incorrect records through detecting multivariate outliers and discordant sites; (ii) delineation of homogeneous regions to cluster sites based on the similarity of catchment descriptors; (iii) selection of the multivariate regional (copula-based) distribution composed of the marginal distributions that represent the variables separately and the copula that characterises the dependence among them; and (iv) estimate of multivariate quantiles for given return period values and selection of design events for a target site.

Screening of the data is a preliminary step, while the others are related to the main stages of a regional HFA. The procedure is presented for the bivariate case with variables X and Y, which correspond to the flood hydrograph volume (V) and annual maximum flood peak (Q) in this study. The bivariate context is considered because it is less complex and more didactic than cases that entail higher dimensions. However, the procedure could be adapted to higher dimensions by solving certain theoretical and practical difficulties, such as those related to the use of multivariate copulas for modelling complex structures and computational issues (Chebana and Ouarda, 2009). The data set is denoted as (x_{ij}, y_{ij}) with $i = 1 : N$ and $j = 1 : n_i$, where N is the number of sites and n_i is the data length of site i. As usually considered in both local and regional analyses, flood events are assumed to be independent and identically distributed. For the sake of simplicity, no cross-correlation is considered between one studied site and another.

The index-flood model assumes that the frequency distributions at all sites belonging to a homogeneous region are identical, except for a scale factor (the so-called index flood) (e.g., Bocchiola et al., 2003; Grover et al., 2002). Both as a reminder and as a manner to facilitate the understanding of the connection between univariate and bivariate approaches, the index-flood models are formulated below for a given non-exceedance probability $p \in (0, 1)$ at a target site i:

$$\text{Univariate}: \quad \widehat{Q}^{i}_{X}(p) = \hat{\mu}^{i}_{X}\hat{q}_{X}(p), \tag{1}$$

$$\text{Bivariate}: \quad \widehat{Q}^{i}_{X,Y}(p) = \begin{bmatrix} \hat{\mu}^{i}_{X} \\ \hat{\mu}^{i}_{Y} \end{bmatrix} \hat{q}_{X,Y}(p), \tag{2}$$

where $\widehat{Q}^{i}_{X}(p)$ is the univariate quantile of X at the site i for a given probability p, $\hat{q}_{X}(p)$ is the value of the (dimensionless) regional growth curve of X for p, $\hat{\mu}^{i}_{X}$ is the index flood of X estimated at the target site, and $\widehat{Q}^{i}_{X,Y}(p)$ is the bivariate quantile curve of X and Y at the site i for a given (simultaneous non-exceedance) probability p, consisting of points formed by X–Y components that are obtained by multiplying componentwise the (dimensionless) bivariate regional quantile curve for p, $\hat{q}_{X,Y}(p)$, by the estimated index floods $\hat{\mu}^{i}_{X}$ and $\hat{\mu}^{i}_{Y}$. It should be noted that in the univariate approach the quantile is a single value, whereas in the bivariate approach it is a curve.

A set of packages of the free software R (R Development Core Team, 2012) has been used as a basis for performing this study, such as copula (Hofert et al., 2012), CDVine (Brechmann and Schepsmeier, 2013), lmom (Hosking, 2014a), lmomRFA (Hosking, 2014b) and nsRFA (Viglione, 2013).

2.1. Screening of the data

Screening of the data is an essential step to find incorrect records, sites with an anomalous hydrological response and infrequent events that could affect the results of a hydrological frequency analysis. Traditional techniques used in univariate HFA are extended to the multivariate case. Specifically, this step is composed of two tests (see Fig. 2): (a) at-site multivariate outlier detection which identifies unusual observations or gross errors that could have a significant impact on the fitting and, consequently, on the quantile estimates mainly for high return periods; and (b) a multivariate discordancy test, which detects discordant sites in a region due to an unusual hydrological response at the catchment scale or gross errors.

2.1.1. At-site multivariate outlier detection

The detection of multivariate outliers is performed at each site with the aim of removing incorrect measurements and gross errors from multivariate observed series. Outliers are unusual or inconsistent extreme events in relation to the whole data series which may be associated with rare observations or human error. Hence, outliers should be first identified and then checked individually to decide whether to either remove them in case of data errors or keep them if they are infrequent but real events. In the multivariate context, a datum is considered as an outlier if its outlyingness value exceeds a given threshold (Dang and Serfling, 2010). In the present study, spatial and two outlyingness functions based on depth functions (Mahalanobis and Tukey functions) are used for this purpose (Chebana and Ouarda, 2011a). These outlyingness functions measure how 'far' a given point is regarding the entire sample by setting a threshold based on the ratio of false outliers (δ):

$$\delta = \frac{\alpha_n}{\varepsilon_n}, \tag{3}$$

where α_n is the proportion of non-outliers misidentified as outliers and ε_n is the real theoretical proportion of outliers. The value of α_n should be less than ε_n (often equal to 0.15). Common values of δ are 0.10 and 0.15 (Chebana and Ouarda, 2011a; Dang and Serfling, 2010).