Journal of Hydrology 408 (2011) 67–77

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

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

Spatially smooth regional estimation of the flood frequency curve (with uncertainty)

F. Laio [a,*], D. Ganora [a], P. Claps [a], G. Galeati [b]

[a] Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, corso Duca degli Abruzzi 24, Torino, Italy
[b] ENEL S.p.A., Unità di Idrologia, via Torino 14, Mestre (VE), Italy

ARTICLE INFO

Article history:
Received 25 June 2010
Received in revised form 14 June 2011
Accepted 17 July 2011
Available online 27 July 2011

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Alin Andrei Carsteanu, Associate Editor

Keywords:
Regional flood frequency analysis
L-moments
Uncertainty
Ungauged basins
Short records

SUMMARY

Identification of the flood frequency curve in ungauged basins is usually performed by means of regional models based on the grouping of data recorded at various gauging stations. The present work aims at implementing a regional procedure that overcomes some of the limitations of the standard approaches and adds a clearer representation of the uncertainty components of the estimation.

The information in the sample records is summarized in a set of sample *L*-moments, that become the variables to be regionalized. To transfer the information to ungauged basins we adopt a regional model for each of the *L*-moments, based on a comprehensive multiple regression approach. The independent variables of the regression are selected among a large number of geomorpholoclimatic catchment descriptors. Each model is calibrated on the entire dataset of stations using non-standard least-squares techniques accounting for the sample variability of *L*-moments, without resorting to any grouping procedure to create sub-regions. In this way, *L*-moments are allowed to vary smoothly from site to site in the descriptor space, following the variation of the descriptors selected in the regression models. This approach overcomes the subjectivity affecting the techniques for the definition and verification of the homogeneous regions. In addition, the method provides accurate confidence bands for the frequency curves estimated in ungauged basins.

The procedure has been applied to a vast region in North-Western Italy (about 30,000 km^2). Cross-validation techniques are used to assess the efficiency of this approach in reconstructing the flood frequency curves, demonstrating the feasibility and the robustness of the approach.

1. Introduction

The evaluation of the frequency of flood events in ungauged catchments is usually approached by building suitable statistical relationships (models) between flood statistics and basins characteristics, calibrated on a set of records of annual maxima. These models are used to transfer the information available at the gauged sites to the target basin, where only morphoclimatic catchment's characteristics are available. This type of procedure is called a *regional model*, because it identifies a subset of basins, called *region*, that is used as a pooling set where the information to be transferred to ungauged site resides. In standard regional models, the basins, which are assumed to belong to a homogeneous region, donate their (common) statistical properties of the flood frequency curve to the ungauged basins that are assumed to fall in the same region.

Various methods to achieve this goal have been proposed in the literature (see for example the review by Cunnane (1988) and Grimaldi et al. (2011)), differing to each other mainly on the basis of the distribution used to describe the at-site data (see e.g. Hosking and Wallis, 1997 for a bouquet of distributions), and on the pooling criterion used for the delineation of regions. Several techniques have been proposed for region delineation. Among others, we can mention: cluster analysis and proximity pooling (Burn, 1990), hierarchical approaches (Fiorentino et al., 1987; Gabriele and Arnell, 1991), neural network classifiers (Hall and Minns, 1999) and mixed approaches (Merz and Bloschl, 2005). For any of these techniques the check for statistical homogeneity within the regions is an important issue (Viglione et al., 2007; Castellarin et al., 2008).

However, most of the standard statistical tools for the estimation of the flood frequency curve in ungauged basins present limitations. In particular, (i) the subdivision of the domain of interest in homogeneous regions, and (ii) the choice of an a priori probability distribution to describe the sample data, can be considered as limiting factors, due to the difficulties of managing estimations where abrupt changes occur across regions, or distributions demonstrate not to keep their properties inside and across regions.

Regarding the point (i), different approaches exist to create homogeneous regions. For instance, regions can be created by splitting in separated areas the geographical space or a multidimensional space of the physiographic basin's characteristics

* Corresponding author. Tel.: +39 011 0905617.
E-mail addresses: francesco.laio@polito.it (F. Laio), daniele.ganora@polito.it (D. Ganora), pierluigi.claps@polito.it (P. Claps), giorgo.galeati@enel.com (G. Galeati).

0022-1694/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2011.07.022

(e.g. Ouarda et al., 2001, Fig. 1). The regions can be defined by means of fixed boundaries (e.g. cluster analysis procedures) or by means of a pooling technique that does not define fixed regions, as in the region-of-influence (ROI) approach (Burn, 1990). The ROI approach is more flexible than the fixed-regions approach because it creates site-dependent regions. However, the estimates are not smooth (both in geographic or physiographic spaces) due to possible discontinuities at the border between one ROI and another.

The main limitation of the approaches that use a subdivision in separate regions is the difficulty to assess a reliable and stable configuration of the regions (e.g. which catchments to include or not in a particular region). In fact, since there is no prior information about the regions configuration, any algorithm used for regions delineation induces some errors. Then, the regions must be tested for their statistical homogeneity, although the related tests can be rather weak in the estimation of statistical heterogeneity (Viglione et al., 2007). A few papers have tried to overcome this problem proposing methods based on the interpolation of the hydrological variable in the descriptors space (Chokmani and Ouarda, 2004; Chebana and Ouarda, 2008), or based on the so-called top-kriging (Skoien et al., 2006). The first technique presents problems in the definition of the descriptors used for the interpolation, while the top-kriging is heavily dependent on the availability of large datasets that would support a reliable construction of a "objective" variogram. The idea not to resort to a grouping procedure to form the regions has been also developed by Stedinger and Tasker (1985), and recently improved by Griffis and Stedinger (2007), where the advantages of using this approach are underlined. Using no regions there is no longer the need for an homogeneity test: the statistical characteristics of the floods can vary from site to site and the model will try to reproduce this variability.

All the above approaches require, at the initial stage, an hypothesis on the at-site frequency distribution chosen to describe the data CDF (cumulative distribution function) and to estimate flood quantiles. In fact, these methods basically perform more or less refined interpolation techniques on the flood quantiles estimated at site. This bring us back to point (ii) above, which is related to the choice of an a priori CDF to represent the data. However, different probability distributions can fit equally well the data for low return periods, while they may produce diverging estimates when extrapolated to high return periods (an example will be given in the following Fig. 6). This effect becomes even more evident in the case of short records, which are particularly important in data-scarce regions.

In this paper, we followed the idea of transferring hydrological information assuming no regions nor pooling groups, and we use the *L*-moments and their dimensionless ratios as statistical variables to be transferred to the ungauged sites. In particular, we select the sample *L*-moment of order one (the mean), the coefficient of *L*-variation (L_{CV}) and the *L*-skewness (L_{CA}) of the record. Regionalizing these three *L*-moments allows one to reconstruct the whole flood frequency curve, at least if three-parameter curves are selected. The choice of the mean, L_{CV} and L_{CA} as hydrological signatures in a regional framework can be interpreted in an index-flood framework (Dalrymple, 1960) considering the mean as the scale factor and the *L*-moments ratios as the descriptors of the dimensionless growth curve. A similar approach has been applied by Vogel et al. (1999) to the annual streamflow, who regionalized the first two moments instead of the *L*-moments.

The use of the mean, L_{CV} and L_{CA} instead of a quantile or distribution-parameter is also helpful, for both calibration and prediction purposes, when catchments with short sample records are used in the analysis. In fact, during the model calibration phase, sample *L*-moments are computed even if their sample variability is high (but known or quantifiable), without resorting to often inefficient estimates of the at-site parent distribution. This avoids information loss due to the elimination of short records. On the other hand, if one is interested in the local quantile prediction at a gauged site with a short record, it is still possible to compute, for instance, the index-flood (Q_{ind}) and the L_{CV} directly on the sample record, leaving to the regional procedure the estimation of L_{CA}. From this point of view, this approach extends the original index-flood method, in which Q_{ind} is often estimated locally, based on even few at-site measurements, while the growth curve is derived by a regional model.

The relationships built to transfer the information to the ungauged sites are based on multiple regressions and are discussed in Section 2.2. The choice of the probability distribution used for the final quantile estimation is based on a model averaging approach and is reported in Section 3. The proposed methodology is applied to an area of about 30.000 km^2 located in North-Western Italy, including 70 gauging stations. The application is presented in Section 4 and final remarks are reported in the conclusions section.

2. Model definition

2.1. At-site estimates: systematic and non-systematic information

The first step of the procedure is to check the available data and use them to compute suitable statistical indicators at the gauged sites. Among the possible types of data which can be used in the statistical analysis (e.g. Stedinger et al., 1993), common procedures implicitly assume a record of *n* systematic measures. Sometimes, however, systematic records of data can be integrated with additional data, derived from measurements of significant occasional events. This can be particularly useful when the original systematic record is short. When a number of occasional additional measurements is available, one can merge them with the systematic ones to improve the robustness of the final estimates (e.g. Bayliss and Reed, 2001).

To calculate the probability weighted moments (PWMs) of the extended record, we use a method suggested by Wang (1990): the merged sample of total length n_{all} is arranged in increasing order

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n_{all}-l+1)} \leqslant x_{(n_{all}-l+2)} \leqslant \ldots \leqslant x_{(n_{all})} \qquad (1)$$

where the subscript in round brackets indicates the sorted position; the *l* largest events, exceeding a threshold x_0, are considered as a censored sample, whose elements can be either systematic or occasional data. Then, the "equivalent" length *m* is associate to the complete record, equal to the number of years between the first and the last measurement of both the systematic and the occasional record, considered together. When working with censored samples, the theoretical formula for the PWM of order *r* of a random variable *X* with distribution function $F(x) = P(X \leqslant x)$, as $\beta_r = \int_0^1 x(F)F^r dF$, must be split in two components (Wang, 1990),

$$\beta_r = \int_0^{F_0} x(F)F^r\,dF + \int_{F_0}^1 x(F)F^r\,dF = \beta''_r + \beta'_r \qquad (2)$$

where $F_0 = F(x_0)$ is the non-exceedance probability relative to the censoring threshold x_0. The unbiased estimator of β''_r is then (Wang, 1990):

$$b''_r = \frac{1}{n}\sum_{i=1}^{n_{all}} \frac{(i-1)(i-2)\ldots(i-r)}{(n-1)(n-2)\ldots(n-r)} x''_{(i)} \qquad (3)$$

where $x''_{(i)}$ is deducted from the sorted sample as

$$x''_{(i)} = \begin{cases} x_{(i)} & \text{if } x_{(i)} < x_0, \\ 0 & \text{if } x_{(i)} \geqslant x_0. \end{cases}$$

On the other hand, the estimator of β'_r is (Wang, 1990)