Journal of Hydrology 540 (2016) 515–526

Contents lists available at ScienceDirect

Journal of Hydrology

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

Regional flood frequency analysis using spatial proximity and basin characteristics: Quantile regression vs. parameter regression technique

Kuk-Hyun Ahn [a,*], Richard Palmer [b]

[a] Northeast Climate Science Center, University of Massachusetts Amherst, MA, United States
[b] Dept. of Civil and Environmental Engineering, University of Massachusetts Amherst, MA, United States

ARTICLE INFO

Article history:
Received 29 October 2015
Received in revised form 21 June 2016
Accepted 22 June 2016
Available online 23 June 2016
This manuscript was handled by G. Syme, Editor-in-Chief

Keywords:
Regional flood frequency analysis (RFFA)
Spatial proximity
Spatial lagged model (SLM)
Spatial error model (SEM)
Quantile regression technique (QRT)
Parameter regression technique (PRT)

ABSTRACT

Despite wide use of regression-based regional flood frequency analysis (RFFA) methods, the majority are based on either ordinary least squares (OLS) or generalized least squares (GLS). This paper proposes 'spatial proximity' based RFFA methods using the spatial lagged model (SLM) and spatial error model (SEM). The proposed methods are represented by two frameworks: the quantile regression technique (QRT) and parameter regression technique (PRT). The QRT develops prediction equations for flooding quantiles in average recurrence intervals (ARIs) of 2, 5, 10, 20, and 100 years whereas the PRT provides prediction of three parameters for the selected distribution. The proposed methods are tested using data incorporating 30 basin characteristics from 237 basins in Northeastern United States. Results show that generalized extreme value (GEV) distribution properly represents flood frequencies in the study gages. Also, basin area, stream network, and precipitation seasonality are found to be the most effective explanatory variables in prediction modeling by the QRT and PRT. 'Spatial proximity' based RFFA methods provide reliable flood quantile estimates compared to simpler methods. Compared to the QRT, the PRT may be recommended due to its accuracy and computational simplicity. The results presented in this paper may serve as one possible guidepost for hydrologists interested in flood analysis at ungaged sites.

1. Introduction

Estimation of the given probabilities or return periods of regional floods is crucial for water resources management (Pandey and Nguyen, 1999). However, streamflow data are not always available at specific sites of interest (ungaged sites) or the record may be too short to provide useful statistics. Moreover, available streamflow data may not accurately represent current conditions of basins due to altered basin characteristics including forest reclamation and urbanization. Regional flood frequency analysis (RFFA) has been proposed to estimate extreme events at ungaged stations in many regions worldwide based on the concept that regional flood flow statistics are closely related to basin and climate characteristics (Chebana and Ouarda, 2008; Gaume et al., 2010; Gupta and Dawdy, 1995; Hussain, 2011; Martel et al., 2011; Micevski and Kuczera, 2009; Nezhad et al., 2010; Noto and La Loggia, 2009; Nyeko-Ogiramoi et al., 2012; Reis et al., 2003).

RFFA research often has adopted a regression-based approach regarding a flood quantile of interest (namely, the quantile regression technique (QRT): Aziz et al., 2015; Gupta et al., 1994; Haddad et al., 2012; Haddad and Rahman, 2012; Thomas and Benson, 1970; Zaman et al., 2012). For instance, Zaman et al. (2012), using the ordinary least squares (OLS) regression, identified a relation between mean annual flood and some basin characteristics, such as area and rainfall intensity for semi-arid and arid regions of Australia. As an alternative approach to the QRT, the parameter regression technique (PRT) has been employed in previous studies (Haddad et al., 2012; Madsen et al., 2002; Malekinezhad et al., 2011; Micevski et al., 2015). This PRT approach estimates the parameters of a selected distribution with explanatory variables: after the parameters are estimated, flood quantiles are derived from the selected distribution and its parameters. For example, Micevski et al. (2015) estimated the parameters of the log-Pearson type III (LP3) using the region of influence (ROI) and Bayesian generalized least squares (GLS). As Haddad et al. (2012) noted, comparison of the PRT with QRT in RFFA is indeed necessary. Few studies comparing the PRT with the QRT (Haddad et al., 2012; Haddad and Rahman, 2012; Taylor et al., 2011) exist; for example, Haddad et al. (2012) compared the QRT and PRT in 53 catchments in Tasmania, Australia and concluded that the QRT provides more accuracy for higher flood percentiles while the PRT offers relatively better performance for lower flood percentiles.

* Corresponding author.
E-mail addresses: ahnkukhyun@gmail.com (K.-H. Ahn), palmer@ecs.umass.edu (R. Palmer).

http://dx.doi.org/10.1016/j.jhydrol.2016.06.047
0022-1694/

The OLS technique enjoyed wide use as an efficient estimator before Stedinger and Tasker (1985, 1986) demonstrated the superiority of the GLS model that can account for inter-site correlation. Since those seminal papers, a majority of RFFA research has incorporated the GLS approach based on so-called 'acceptably homogenous regions' (Eng et al., 2005; Griffis and Stedinger, 2007; Gruber and Stedinger, 2008; Haddad et al., 2012; Haddad and Rahman, 2012; Madsen et al., 1997; Micevski and Kuczera, 2009). However, regardless of technique used to estimate the relationship between explanatory variables and floods (the OLS or GLS), observing underground characteristics presents a major obstacle in identifying relevant basin characteristics (Oudin et al., 2008). To circumvent this limitation, using the concept of 'spatial proximity' can offer an attractive alternative to RFFA. Using information from adjacent basins may be quite beneficial since many hydrologic and physical variables including flooding quantiles are spatially dependent (Oudin et al., 2008). Therefore, a 'spatial proximity' based method such as the spatial lagged model (SLM) is valuable from the viewpoint of RFFA. To be specific, the SLM and spatial error model (SEM), the most representative methods utilizing the concept of 'spatial proximity', has the advantage of sharing the same explanatory variables and information about a nearby region. Therefore, some regionalization papers adopt the SLM and SEM methods to estimate daily or monthly streamflow (Kokkonen et al., 2003; Oudin et al., 2008; Parajka et al., 2007; Steinschneider et al., 2015). For example, Steinschneider et al. (2015) used spatial error regression in the parameters of the *abcd* model at 22 gages in Southeast United States and demonstrated that spatial error regression yields reliable hydrologic simulations. Despite the potential benefits of the 'spatial proximity' approach, few studies have incorporated the 'spatial proximity' approach in RFFA. Merz and Blöschl (2005) proposed the spatial proximity method based on kriging and found spatial proximity to be a better predictor of regional flood frequencies than are catchment attributes. Kjeldsen et al. (2014) and Kjeldsen and Jones (2007, 2010) employed a spatial proximity based regression to estimate the index flood defined as the median annual maximum discharge for a number of gages in the United Kingdom. Even though several studies exist, investigation with the PRT considering 'spatial proximity' continues to receive little attention. Therefore, the objectives of this paper are twofold: (1) propose a 'spatial proximity' methodology in the PRT and QRT (2) compare the PRT and QRT based on 'spatial proximity' methodology. To verify the possibility of the 'spatial proximity' concept, two basic predictions, naïve prediction and OLS, are also provided in this study. The 'spatial proximity' concept is represented by the SLM and SEM methods. Applicability of the suggested methods at ungaged sites is evaluated by a jack-knifing 'leave-one-out' procedure. This jack-knifing 'leave-one-out' procedure estimates how reliably each method represents an ungaged site (Merz and Blöschl, 2005).

The remainder of the paper is structured as follows: Section 2 briefly describes the study area and data. Section 3 presents the methodologies. Results and conclusions are presented in Sections 4 and 5, respectively.

2. Study area and data

To accomplish the objectives of this study, a sufficient number of streamflow gages and their available long term records are required as the study area. The Northeast U.S. spanning twenty-three states is selected for this study (Fig. 1). Missouri, Kentucky, Illinois in Northeast U.S. are reported as the most flood-prone areas in the U.S. (Changnon et al., 2001). The northeast U.S. has an area of 2,149,777 km^2 representing approximately 27.6% of the continental U.S (NALCC, 2002). The climate of these regions can be summarized as humid continental climate and humid subtropical climate (Peel et al., 2007).

Daily streamflow data are initially downloaded at 3713 gages throughout the Northeast U.S. from the USGS stream gage network (http://wateratch.usgs.gov). During pre-processing, many stations are eliminated due to two criteria: (1) streamflow data must be available from January 1, 1980 to December 31, 2014 (35 years) with less than ten days missing data. If data are missing, they are simply replaced by the arithmetic mean of two data from adjacent dates; (2) gages experience minimal water withdrawals; gages classified as "reference" in the Geospatial Attributes of Gages for Evaluating Streamflow version II (GAGES II; Falcone et al., 2010) are used since all basin characteristics considered in this study are adopted by the GAGES II data. Here "reference" represents hydrologic conditions which are least disturbed by human influences. In summary, out of the initial 3713 gages, 237 are selected in this study (Fig. 1).

Twenty-one gages among the selected gages exhibit significant changes for annual maximum streamflow with the Mann-Kendall test (Kendall, 1955; Mann, 1945) at a 95% confidence interval. However, this portion of gages showing significant changes represents less than 10% of the total gages and thus stationary approach, the fundamental assumption for the RFFA, may be acceptable for the study area.

The GAGES II data offers a number of watershed characteristics compiled from national data sources, including environmental features (e.g. historical precipitation, geology, soils, topography) and anthropogenic influences (e.g. land use, road density, presence of dams) (Falcone et al., 2010). Among them, a total of thirty basin characteristics (Table 1) known to have a strong theoretical relationship to streamflow in past studies (denoted in Table 1) are initially selected; to define the components of land use, the National Land Cover Database (NLCD) categories are reclassified into four simple land uses; forest, agriculture, water and developed area. These reclassified land types have been preferred in previous studies (Ahn and Merwade, 2015; Price et al., 2011). In addition, the 'Precipitation Intensity' is generated by using two variables (annual precipitation and annual number of days of measurable precipitation) provided in GAGES II. All except the above five variables are utilized as they have already been provided in GAGES II. The selected basin characteristics are classified into 9 types – Basin topography, Basin morphometry, Channel network, Hydrologic property, Land use, Soil, Basin aspect, Slope, and Precipitation.

3. Methodology

The methodology applied consists identifying regional flood quantiles at all the gages used in this study, defining their relationships with basin characteristics, and predicting flood quantiles at ungaged sites. After the optimal distribution and its parameters are determined for a gage, flood quantiles for the average recurrence intervals (ARIs) of 2, 5, 10, 20, and 100 years are estimated. Then, regression-based models are established to investigate relationships with basin characteristics. At this stage, three regressions including the ordinary least square model (OLS), spatial lagged model (SLM), and spatial error model (SEM) are implemented. Finally, the effectiveness of the defined models is verified and compared by using a jack-knifing 'leave-one-out' procedure. The details of each methodology are provided in the following sub-sections.

3.1. Regional flood frequency analysis

For regional flood frequency analysis, four three-parameter distributions – Generalized Logistic (GLO), Generalized Extreme Value (GEV), Generalized Pareto (GPA), and Pearson type III (PE3) – are