Journal of Hydrology 538 (2016) 387–398

Contents lists available at ScienceDirect

Journal of Hydrology

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

# Regional Flood Frequency Analysis using Support Vector Regression under historical and future climate

Mesgana Seyoum Gizaw, Thian Yew Gan *

Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta T6G 1H9, Canada

ARTICLE INFO

*Article history:*
Received 13 January 2016
Received in revised form 21 March 2016
Accepted 20 April 2016
Available online 27 April 2016
This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Fi-John Chang, Associate Editor

*Keywords:*
Support Vector Regression
Regional Flood Frequency Analysis
Flood quantiles
Southeastern British Columbia
Southern Ontario

SUMMARY

Regional Flood Frequency Analysis (RFFA) is a statistical method widely used to estimate flood quantiles of catchments with limited streamflow data. In addition, to estimate the flood quantile of ungauged sites, there could be only a limited number of stations with complete dataset are available from hydrologically similar, surrounding catchments. Besides traditional regression based RFFA methods, recent applications of machine learning algorithms such as the artificial neural network (ANN) have shown encouraging results in regional flood quantile estimations. Another novel machine learning technique that is becoming widely applicable in the hydrologic community is the Support Vector Regression (SVR). In this study, an RFFA model based on SVR was developed to estimate regional flood quantiles for two study areas, one with 26 catchments located in southeastern British Columbia (BC) and another with 23 catchments located in southern Ontario (ON), Canada. The SVR-RFFA model for both study sites was developed from 13 sets of physiographic and climatic predictors for the historical period. The *Ef* (Nash Sutcliffe coefficient) and $R^2$ of the SVR-RFFA model was about 0.7 when estimating flood quantiles of 10, 25, 50 and 100 year return periods which indicate satisfactory model performance in both study areas. In addition, the SVR-RFFA model also performed well based on other goodness-of-fit statistics such as *BIAS* (mean bias) and *BIASr* (relative *BIAS*). If the amount of data available for training RFFA models is limited, the SVR-RFFA model was found to perform better than an ANN based RFFA model, and with significantly lower median *CV* (coefficient of variation) of the estimated flood quantiles. The SVR-RFFA model was then used to project changes in flood quantiles over the two study areas under the impact of climate change using the RCP4.5 and RCP8.5 climate projections of five Coupled Model Intercomparison Project (CMIP5) GCMs (Global Climate Models) for the 2041–2100 period. The results suggest that due to a projected increase in the mean annual precipitation, and rainfall of a given return period, the flood quantile is projected to increase by about 7% for the southeastern BC and 29% for southern ON region in the mid- and late 21st century.

## 1. Introduction

Flood frequency analysis has been the classic approach to estimate the magnitude of flood events of various return periods. Assuming floods as stochastic processes, the magnitude and frequency of floods are predicted using certain probability distributions usually characterized by one to three parameters (Chow et al., 1988; Rao and Hamed, 2000) estimated from historical streamflow collected over an extended period. The Bulletin 17B (1982) of the United States Geological Survey (USGS) suggests at least ten years of stream gauging records should be analyzed to warrant statistical analysis as a meaningful basis for estimating future flood events, especially events of high return periods. However, if measured streamflow data is not available, flood frequency analysis can be done using flows simulated by hydrologic model, lumped conceptual models such as the Sacramento model (Burnash et al., 1973) or distributed, physically-based models such as MISBA (Kerkhoven and Gan, 2006) forced with observed climate data, or statistical methods that relate flood quantiles with catchment characteristics.

Regional flood frequency analyses (RFFA) are statistical methods that have been widely used to estimate flood quantiles in catchments where streamflow measurements are either limited or unavailable (Griffis and Stedinger, 2007; Ouarda et al., 2007; Shu and Ouarda, 2007; Haddad and Rahman, 2011; Aziz et al., 2013). In the RFFA based on Quantile Regression Techniques (QRT), a large number of gauged basins are selected from a given geographical region and their flood quantiles, which are estimated from observed streamflow records, are then regressed against

* Corresponding author.
*E-mail address:* tgan@ualberta.ca (T.Y. Gan).

http://dx.doi.org/10.1016/j.jhydrol.2016.04.041
0022-1694/© 2016 Elsevier B.V. All rights reserved.

selected basin characteristics such as catchment area, main channel slope and design rainfall intensity. (Thomas and Benson, 1970; Haddad and Rahman, 2011). The statistical relationship established by RFFA can then be used to estimate flood quantiles of river basins that have limited streamflow records but are located within the same geographic and climatic region of surrounding river basins with sufficient data for regional flood quantile estimation. As empirical methods with regional applications, RFFA are attractive and practical when compared to more physically-based methods that could be computationally intensive with massive input data requirements.

The commonly used RFFA include the rational method, index flood method and QRT (Aziz et al., 2013). In QRT, regional flood prediction equations are developed by regressing flood quantiles (predictands) estimated from a large number of gauged catchments in a given geographic region based on the catchments' physiographic and climatic variables (predictors) (Thomas and Benson, 1970; Haddad and Rahman, 2011). Early RFFA studies that utilized QRT methods based on ordinary least square (OLS) regression related flood quantiles with hydrologic characteristics of catchments (Thomas and Benson, 1970). However, many studies have shown that QRT based on generalized least square (GLS) regressions is more efficient than OLS. GLS based regressions generally provide more precise estimates because they account for differences between the variance of streamflow from various sites that mainly arise from differences in record length and cross correlation among concurrent streamflows (Stedinger and Tasker, 1985; Griffis and Stedinger, 2007; Haddad and Rahman, 2011). On the other hand, there have been RFFA studies based on artificial neural networks (ANN). ANNs are machine learning algorithms which are information processing systems that partly function like the human brain (Shu and Ouarda, 2007). With certain built-in search algorithms, and only input and output data, ANNs are capable of finding optimal nonlinear relationships of basin-scale hydrologic processes without requiring detailed physical information or data related to these processes (Nor et al., 2007). Recent RFFA studies based on ANNs have been shown to be better than regression models in modeling complex relationships between flood quantiles and climatic/physiographic properties of a catchment. For example, Aziz et al. (2013) developed an ANN-based RFFA model for estimating flood quantiles in eastern Australia which gave a better performance than RFFA models based on QRT. Similarly, from testing different ANN based RFFA models in Québec, Canada, Shu and Ouarda (2007) showed that ANN based models are easier to apply and tend to be more accurate than regression based RFFA models.

Another machine learning approach that can also potentially be used for RFFA is Support Vector Regression techniques developed from a kernel-based classification algorithm called the Support Vector Machines (SVM). Over the years SVM has been extended as a regression tool referred to as Support Vector Regression Machine (SVR) (Drucker et al., 1997; Smola and Schölkopf, 1998). Even as relatively new techniques, SVM and SVR have been applied in various hydrologic studies, and in particular for streamflow prediction. Liong and Sivapragasam (2002) used SVM to predict the flood stage of the Brahmaputra, Ganges and Meghna Rivers which merge at the city of Dhaka, Bangladesh and they concluded that SVM is at least as good if not better than ANNs and has better generalization ability when the training dataset available is limited. Dibike et al. (2001) used SVM in the classification of remotely sensed data and rainfall-runoff modeling and found the generalization and performance of SVM to be better than other classification methods and traditional conceptual rainfall runoff models. Yu et al. (2006) used SVR to develop a real-time flood stage forecasting model that could effectively forecast flood stages up to six hours of lead time. Wu et al. (2008) also used SVR for river stage prediction and concluded that their model could predict river stages more accurately than other machine learning algorithms such as ANN. Samui (2011) showed that a Least Square Support Vector Machine (LS-SVM) model could predict evaporation loss from reservoirs more accurately than and an ANN model. Zakaria and Shabri (2012) found that SVM to be better than multiple linear regressions (MLR) in predicting the streamflow of ungauged sites. SVM has also been used in predicting groundwater levels in coastal aquifers where it showed to have more superior generalization ability than ANN models (Yoon et al., 2011).

As mentioned in the previous paragraph, the comparatively better generalization ability of SVR makes it an attractive alternative approach to perform RFFA with limited number of gauged surrounding river basins for estimating flood quantiles of an ungauged basin. In this study, our objective is to investigate the performance of a machine learning technique (SVR) in a RFFA and to compare its performance with ANN based RFFA models for two groups of river basins located in British Columbia (BC) and Ontario (ON) of western and eastern Canada, respectively. We will also extend the application of SVR-RFFA model to predict changes in projected flood quantiles over the two study areas under the impact of climate change. With this introduction, a brief description of SVR is given in Section 2, data and methodology in Section 3, discussion and results in Section 4, and summary and conclusions in Section 5.

## 2. Support Vector Regression

The Support Vector (SV) algorithm is a class of nonlinear search algorithm based on a statistical learning theory developed by Vapnik and Chervonenkis (Vapnik and Chervonenkis (1964), Smola and Schölkopf (1998)). Over the years, the SV algorithm has been successively developed as a classification tool such as the Support Vector Machines (SVM), and later combined with a regression technique to become the Support Vector Regression (SVR) (Drucker et al., 1997; Smola and Schölkopf, 1998). For a given $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset X \times \Re$, the SVR technique aims to find a function $f(x)$ that has an $\varepsilon$ deviation from the observed targets $y_i$ for all training data $x_i$. $f(x)$ can be written for linear SVR as

$$f(x) = \langle \omega, x \rangle + b \text{ with } \omega \in X, b \in \Re \tag{1}$$

where $\langle \omega, x \rangle$ represents the dot product. In order to get a suitable fitting function $f$ one will search for a small $\omega$(weighting factor) and a constant $C$ which will optimize an objective function given as

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\underset{for\ i=1\ to\ l}{\text{Subject to}} \quad \begin{cases} y_i - \langle \omega, x_i \rangle - b \leqslant \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i \leqslant \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geqslant 0 \end{cases} \end{aligned} \tag{2}$$

where slack variables $\xi_i, \xi_i^*$ are introduced so that the function $f$ that approximates all pairs of $(x_i, y_i)$ are given rooms for errors that are beyond the targeted deviation $\varepsilon$. The constant $C > 0$ determines the amount of slack that can be tolerated beyond the deviation target $\varepsilon$ to achieve an optimal search for the fitting function $f$. The solution of (2) can be found by introducing a Lagrangian function with a dual set of variables given as

$$\begin{aligned} L = {} & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ & - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - y_i + \langle \omega, x_i \rangle + b) \\ & - \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i + y_i - \langle \omega, x_i \rangle - b) \end{aligned} \tag{3}$$