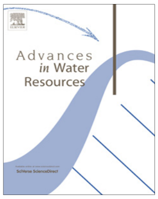
Advances in Water Resources 68 (2014) 1–12

Contents lists available at ScienceDirect

Advances in Water Resources

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



Use of Gene Expression Programming in regionalization of flow duration curve

Muhammad Z. Hashmi [a,*], Asaad Y. Shamseldin [b]

[a] *Water Resources and Glaciology Section, Global Change Impact Studies Centre (GCISC), Pakistan*
[b] *Department of Civil and Environmental Engineering, The University of Auckland, New Zealand*

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form 30 January 2014
Accepted 23 February 2014
Available online 7 March 2014

Keywords:
Hydrology
Artificial intelligence
Catchment
Non-linear regression

ABSTRACT

In this paper, a recently introduced artificial intelligence technique known as Gene Expression Programming (GEP) has been employed to perform symbolic regression for developing a parametric scheme of flow duration curve (FDC) regionalization, to relate selected FDC characteristics to catchment characteristics. Stream flow records of selected catchments located in the Auckland Region of New Zealand were used. FDCs of the selected catchments were normalised by dividing the ordinates by their median value. Input for the symbolic regression analysis using GEP was (a) selected characteristics of normalised FDCs; and (b) 26 catchment characteristics related to climate, morphology, soil properties and land cover properties obtained using the observed data and GIS analysis. Our study showed that application of this artificial intelligence technique expedites the selection of a set of the most relevant independent variables out of a large set, because these are automatically selected through the GEP process. Values of the FDC characteristics obtained from the developed relationships have high correlations with the observed values.

1. Introduction

Knowledge of the flow regime of streams is required by catchment managers for water allocation and for ensuring sufficient low flow for aquatic life and for recreation. The flow regime of a stream is influenced by many factors related to its catchment such as climate, storage and morphology [1–3]. In addition, land use has an indirect but large effect on streamflow regime, through its effect on evapotranspiration and infiltration. If flow gauging has been carried out in a stream for a sufficiently long period (at least several years), its flow regime is determined directly from the streamflow record. More indirect methods have to be used to determine the streamflow regimes of un-gauged catchments. Regionalization is one such indirect method.

1.1. Flow duration curves

The two popular approaches for regionalization are deterministic rainfall–runoff models and the flow duration curve (FDC) approach [4]. Whereas, the former has some limitations related to the determination of model parameters and data requirement, and it is furthermore regarded as “complex and information consuming”, see for example [5], the latter may be preferred particularly in regions with scarce rainfall data.

FDCs show the percentage of time discharges of various sizes (from the lowest to the highest recorded in the catchment) are exceeded [6]. An FDC encapsulates much information about the flow regime of a stream, such as median flow, time distribution of low flow and high flow, and variability of flow. Some of its water resources applications include water resources planning, water supply, reservoir operation, reservoir sedimentation, and hydropower planning and generation [7]. The FDC approach avoids some of the problems of the rainfall–runoff modelling approach, such as model parameter determination and high data requirement, thus offering a simpler method of estimating a flow regime. Regionalization based on the FDC approach has wide application for estimating flow regimes of ungauged streams [4].

Flow duration curves can be normalised by dividing all discharge values by the annual mean discharge (MF), or by the annual median discharge (Q50), allowing FDCs for different catchments to be compared. The slope of the normalised FDC is a measure of streamflow variability [8]. The flow duration curve for a catchment with little variability will have a flat slope in the middle part (indicating that the mean or median flow occurs most of the time) and the flow duration curve for a stream with highly variable flow will have a steep slope. Flow variability depends on the hydro-geological characteristics of a catchment. A useful measure of low-flow is the discharge that is exceeded 95% of the time (Q95). It has been found

---

* Corresponding author. Tel.: +92 333 5575535.
E-mail address: mhas074@aucklanduni.ac.nz (M.Z. Hashmi).

http://dx.doi.org/10.1016/j.advwatres.2014.02.009
0309-1708/© 2014 Elsevier Ltd. All rights reserved.

that there is a strong statistical relationship between Q95/MF (dividing by MF reduces the influence of climate) and suitable measures of the hydro-geology of a catchment (such as metrics indicating soil and bedrock storages and flow rates), see for example [3]. Q95/MF is, in effect, a measure of catchment permeability – its ability to release stored water during periods of low rainfall [3].

There are a number of FDC regionalization techniques available which can be broadly categorised as [7]:

(i) Statistical techniques.
(ii) Parametric techniques.
(iii) Graphical techniques.

Statistical techniques (STs) involve the fitting of a probability distribution to streamflows. Castellarin et al. [7] have summarised some of the early attempts towards the development of statistical techniques for regionalization of FDCs, including (*a*) the regional hydrological model of [9], in which a two parameter log-normal distribution is fitted to the daily stream flows; (*b*) application of a five parameter mixed log-normal distribution and basin clustering by [10]; (*c*) the Claps and Fiorentino [11] approach, which is based on fitting a two-parameter log-normal distribution to each member of the series of annual FDCs (a FDC for each year of streamflow record); (*d*) the statistically based model of Singh et al. [12] developed for the Himalayan region; and (*e*) the model of Croker et al. [13] for ephemeral streams, using the theory of total probability.

More recently, Castellarin et al. [14] presented the regionalization of FDCs through an index flow model (IFM) proposed by Castellarin et al. [15]. As reported by the authors, the IFM is able to derive period-of-record as well as yearly FDCs. The results of this study were validated by applying a jack-knife re-sampling methodology to quantify the uncertainty. The authors reported their results to be comparable or better than other regionalization approaches in practice.

A probabilistic framework, for the design and performance evaluation of small hydro plants and quantification of the associated uncertainty, was presented by [16], based on the Claps and Fiorentino [11] approach. Annual daily flows are assumed to be following a three parameter log-normal distribution (unlike the traditional approach which uses two parameter log-normal).

Iacobellis [17] proposed the model "EtaBeta" for producing FDCs with a specified return period. The EtaBeta model is based on the two-parameter complementary beta distribution, and is bounded by the annual minimum and maximum discharges. The model has the advantage that it produces distributions of the annual minimum daily flow and total annual streamflow.

A regional model has been developed by Ganora et al. [18] for estimating dimensionless non-parametric FDCs for ungauged basins. Their model works on a distance-based approach for quantifying the dissimilarity among the FDCs and basin characteristics (geographic, geomorphologic and climatic) of different basins.

Botter et al. [19] extended the seasonal model (introduced in a break through work of Botter et al. [20–22] for the development of a relationship between the probability density functions of stream flow and parameters related to eco-hydrology, climate and stream transport) to annual time scales. They propose a gamma distribution for representing annual stream flows. The presented model proved to be efficiently reproducing the observed FDC at the study sites. In a very recent work by Botter et al. [23], this model has been now applied to more than 110 seasonal flow regimes in different catchments spread throughout the US and Alps spanning a broad range of climate and morphologic conditions.

Parametric techniques (PTs) employ the development of relationships between the FDC of a gauged site and different catchment descriptors [7]. These relationships are then used for the FDC estimation at ungauged sites.

Examples of the parametric techniques used for FDC regionalization have been reported in a previous review by Castellarin et al. [7]. These include the exponential equation proposed by Quimpo et al. [24] containing two parameters (catchment area and a regression coefficient); the third order polynomial equation of Mimikkou and Kaemaki [25] with four parameters; the study of Franchini and Suppo [26] involving a three parameter analytical equation; and uncertainty assessment of regional FDCs by Yu et al. [27].

Holmes et al. [28] present a 'region of influence' approach for regionalization of FDCs for United Kingdom streams, which builds upon a previous work by Gustard et al. [29] and Young et al. [3]. In their method, any ungauged catchment for which an FDC is required becomes the centre of a 'region of influence' specific to that catchment, and from which a regional FDC curve is derived.

There have been several developments since the review by Castellarin et al. [7]. Patel [30] proposed a simplified solution for low flow estimation in ungauged catchments by the application of exponential and linear models. In the quest for simple methods for FDC estimation, Rojanamon et al. [31] fitted the FDCs of 21 stream gauging stations in the Salawin river basin by logarithmic, quadratic, cubic, power and exponential distributions. The performance of logarithmic and exponential distributions in terms of flow prediction was found to be better than any other distribution used in their study. On the basis of the findings of their study, they proposed the use of simple FDC estimation methods involving very few catchment parameters for exploring the potential of small hydropower sites. Mohamoud [32] has recently proposed a new methodology for regionalization of FDCs with many features not available in other existing methodologies of this nature. It involved the following steps: construction of normalised FDCs; a step-wise regression process for the identification of the most important catchment descriptors from a large set consisting of climatological, morphological and hydrological parameters; development of regional models; prediction at ungauged site using the regional models; and performance evaluation of the regional models. As reported by Mohamoud [32], a unique feature of this approach is that it treats stream flows in two parts namely a magnitude component (the FDC) and a sequence component (time series). The study results have shown that this approach is more flexible than other methods and can have wide application in many water resources related problems. A more

Table 1
Comparison of GEP technique to GP and GAs [37].

Step	Genetic programming	Genetic Algorithms	Gene Expression Programming
1.	In GP, population individuals (chromosomes) are non-linear, varying in length as well as shape (also known as 'parse trees')	In GAs, population individuals (chromosomes) are linear and of fixed length	In GEP, population individuals (chromosomes) are linear entities of fixed length that are converted to non-linear entities of varying sizes and length (expression trees or computer programs) at a later stage
2.	Uses a single entity working as genome (gene) and phenome (body) at the same time	Similar to GP	GEP has totally separated genome and phenome
3.	Sometimes, invalid expressions can be obtained	Similar to GP	Always produces valid expressions
4.	GP is not yet established beyond the replicator threshold	Similar to GP	GEP is well established beyond the replicator threshold