# DEVELOPMENT OF THE STREAMFLOW AND FLOOD ANALYSIS SYSTEM USING R (SFASUR-TEC)

MAIKEL MENDEZ[(1)] & LUIS A. CALVO-VALVERDE [(2)]

[(1)] Escuela de Ingeniería en Construcción, Instituto Tecnológico de Costa Rica, Cartago, Costa Rica.
mamendez@itcr.ac.cr
[(2)] Escuela de Ingeniería en Computación, Instituto Tecnológico de Costa Rica, Cartago, Costa Rica.
lcalvo@itcr.ac.cr

## ABSTRACT

The Streamflow and Flood Analysis System Using R (SFASUR-TEC) was developed as part of a research project focusing on hydrological modelling for operational hydropower planning and forecasting in tropical mountainous regions of Costa Rica. Since Costa Rica has an enormous hydropower potential and renewable energy markets keep growing, reliable estimation and assessment of river flows is imperative. SFASUR-TEC was developed to increase the understanding of natural and altered flow regimes linked to relevant physical processes with a particular interest in hydropower applications. Consequently, SFASUR-TEC is a collection of methods to identify and quantify key components of flow regimes and assess their behaviour through time. The system was completely developed using the R programming language based on various criteria, including availability of specialized libraries, open-source implementation and cross-platform compatibility. To evaluate the performance of the system, SFASUR-TEC was applied to the upper Toro River catchment in Costa Rica using a 17-year streamflow data set (1994-2010). The catchment was selected based on its predominantly mountainous and rainy conditions, its relevance in the Costa Rican hydropower generation context and the availability of temporal and spatial information. Results from the Flood Frequency Analysis Block, suggest a 60 $m^3/s$ peak flow for a 20-year return period (FFATr20) along with its 95% confidence intervals with values between 42 and 93 $m^3/s$. Flow Duration Curves (FDCs) show that the dry season expands from February through April, with flow values as low as 1.5 $m^3/s$, essentially sustained by baseflow. Baseflow Separation shows that baseflow is indeed a considerable contribution to total streamflow, with BFI values ranging from 0.12 to 0.40 between the months of May through December. Results from this study are intended to support strategic decision-making to improve the performance of the hydropower facilities currently installed in the upper Toro River catchment and elsewhere.

**Keywords:** Analysis, Flood, R, SFASUR-TEC, Streamflow.

## 1 INTRODUCTION

Costa Rica is among the world's leaders in renewable energy as nearly 93% of its electricity comes from renewable sources. This has been accomplished with considerable investment in developing and expanding renewable energy capacity, particularly hydropower (WEF, 2013). Hydropower is the predominant source of energy which accounts for over 78% of the country’s total generation, while geothermal represents 12%, wind 2.1%, biomass 0.25% and only 7.19% comes from thermal sources (ICE, 2011).

Many of the country’s leading hydropower facilities are located in catchments with tropical mountainous conditions which are characterized by persistent immersion in clouds, an important source of precipitation year-round (IMN, 2008). This precipitation provides streamflow for facilities that run by water stored in reservoirs or those that are run by streamflow directly. Therefore, a reliable assessment of flow regimes in these mountainous catchments is essential to improve inflow predictions and hydropower operation, along with the development of future potentials.

Accordingly, the Streamflow and Flood Analysis System Using R (SFASUR-TEC) was developed to identify and quantify key components of flow regimes and assess their behaviour through time in mountainous tropical catchments. SFASUR-TEC was developed to increase the understanding of natural and altered flow regimes linked to relevant physical processes with a particular interest in hydropower applications.

SFASUR-TEC was applied to the upper Toro River catchment, where Insitituto Costarricense de Electricidad (ICE) operates three hydropower facilities with an installed capacity of 138 MW; approximately 10% of the Costa Rican installed capacity (ICE, 2011). This catchment was selected based on its predominantly mountainous and rainy condition, its relevance in the national hydropower generation context and the availability of temporal and spatial information.

## 2 STUDY AREA AND DATA SOURCES

The upper Toro River catchment (43.15 km2) is located in the province of Alajuela in north-western Costa Rica (Figure 1). The topography is mountainous with elevations ranging from 2593 to 1334 m. The slope is steep with a mean value of 23%. The mean annual rainfall of the area is 4200 mm and the mean annual temperature range is between 17.2 and 32.8 ºC. The land use in the catchment is dominated by forest (62%) and grassland (35%) with minor contributions from other uses; mainly water and urban. The catchment has a highly complex precipitation pattern and its temporal and spatial distribution is influenced by factors such as El Niño southern oscillation (ENOS), geomorphology, rugged terrain and microclimates. Daily and hourly streamflow data for the period were obtained from the ICE 12-6 gauge station. Even when the catchment has been monitored since the early 1980's, hourly records are very sparse. For this reason, only mean daily streamflow data from a 17-year dataset (1994-2010) was used in this study.

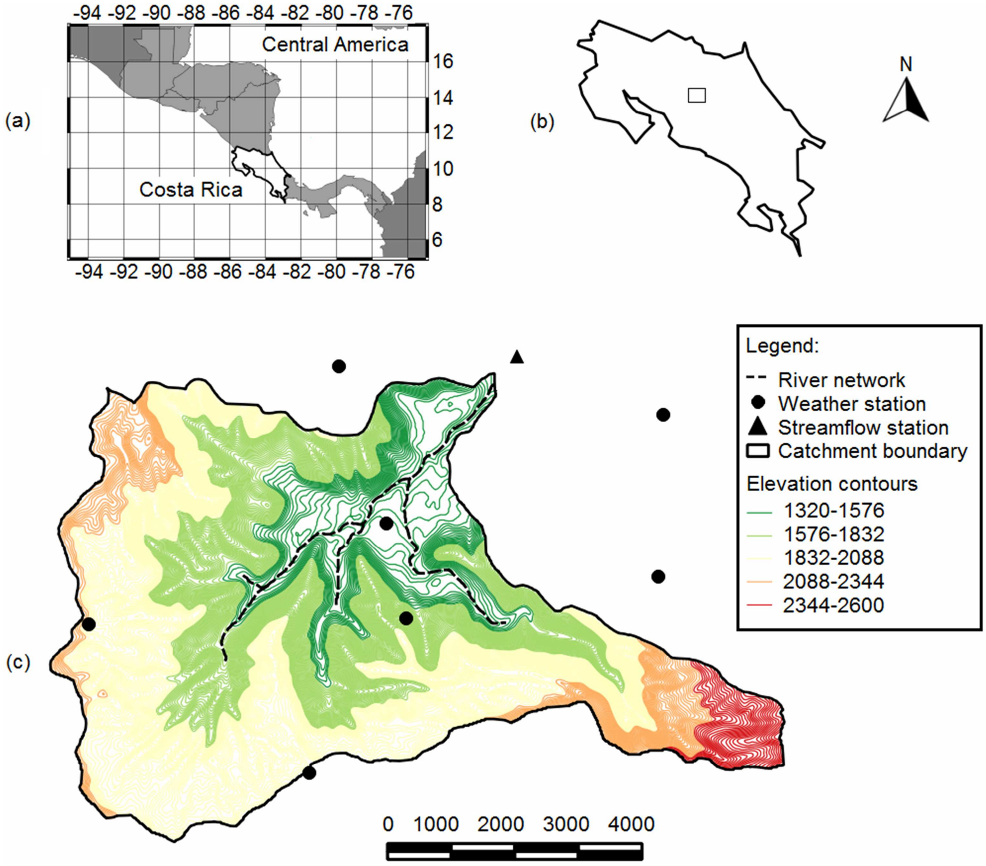


**Figure 1**. (a) Position of Costa Rica in Central America, (b) position of the upper Toro River catchment in Costa Rica; (c) Upper Toro River catchment boundary, river network, elevation contours, rain-gauges and streamflow gauging station.

## 3 SYSTEM DESCRIPTION

As coded using the R programming language (R Core Team, 2016), SFASUR-TEC is divided in various blocks (Figure 2). The authors recommend running SFASUR-TEC under RStudio-IDE (RStudio Team, 2016) as it significantly facilitates interaction with the system. Below follows a brief description of each block. Relevant R packages used in the development of SFASUR-TEC are from now on presented in curly braces {}. For each block, the outcome of the respective analysis is displayed in tabular and graphical form.

### 3.1 Custom Functions - Block

One of the great strengths of R is the user's ability to add custom functions. For instance, this block includes all custom functions along with data filtering routings, descriptive statistics {*pastecs*} and temporal series manipulation {*lubridate*}, {*tidyr*} and {*dplyr*}.

### 3.2 Flood Frequency Analysis Block

This block along with its custom functions, fit a given extreme value distribution to an extreme value series vector (e.g. annual instantaneous maximum value). Log Pearson Type III distribution is selected by default (Maidment, 1993; Chow et.al, 1988). Nonetheless, a list of distributions supported by {*lmomco*} package may also be selected (e.g. Generalized Extreme Value Distribution and Gumbel Distribution). A bootstrapping routing is also applied to randomly sample the extreme value series to estimate confidence interval for each given non-exceedance probability and return period (from 1 to 500 years).

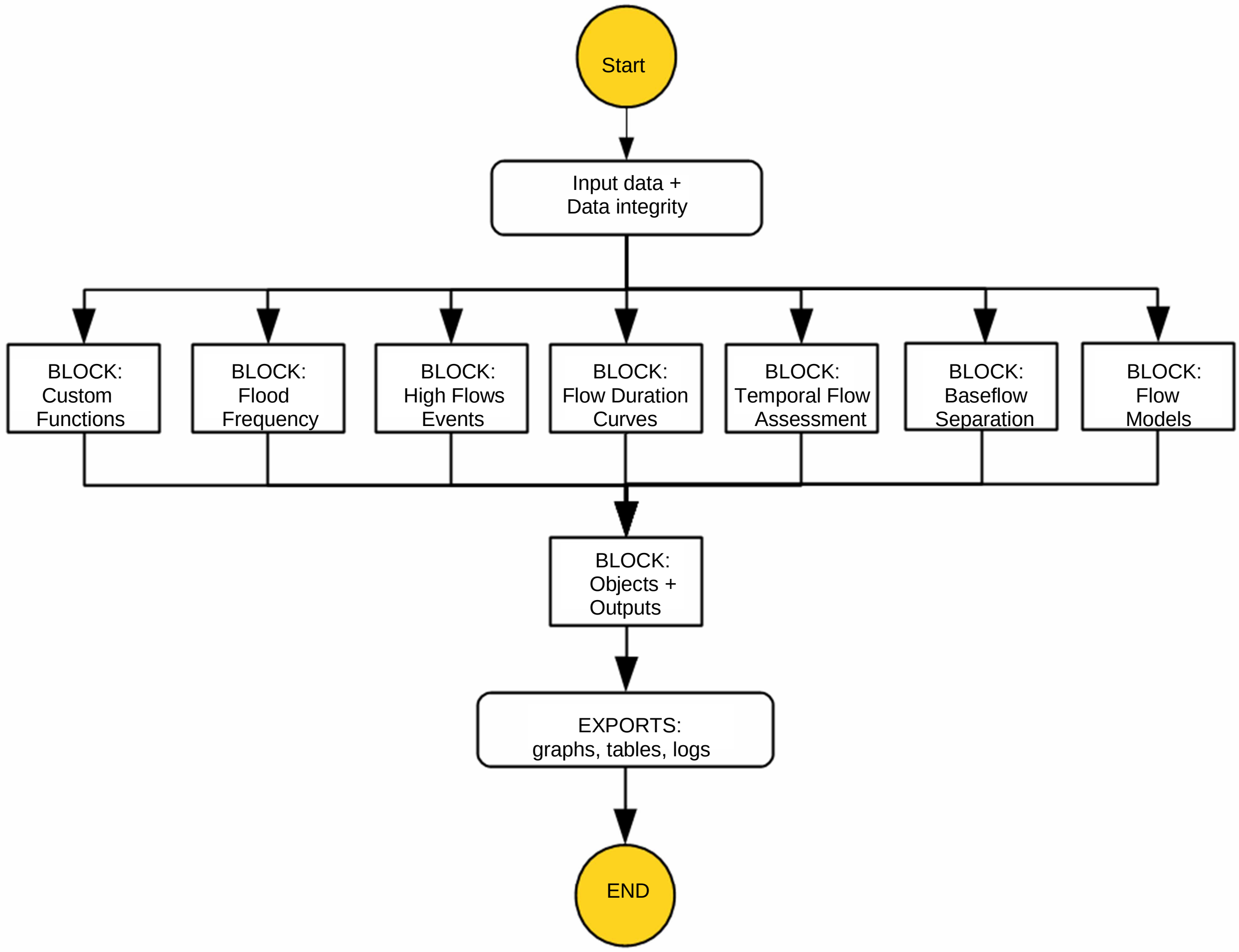


**Figure 2**. SFASUR-TEC model structure.

3.3 Highflow Events - Block

Based on the results from the Flood Frequency Analysis Block, highflow events with defined threshold recurrence intervals (1.5 and 10 years by default) are identified and quantified and a new subset data frame is created. Subsequently, the extreme value series vector is numerically and graphically compared with this subset. The main objective of this block is to quantify the amount and magnitude of exceedance probability highflow events included in the entire time series. Another functionality included in this block is a hydrograph explorer, which graphically displays streamflow data records in a variety of temporal aggregations (annual and monthly). The main objective is to gain a better appreciation of the entire sequence of events involved in the time series. Descriptive statistics (mainly the mean and median) are also included.

3.4 Flow Duration Curves - Block

Flow Duration Curves (FDCs) are computed using specialized functions from the {*hydroTSM*} library which follows the methods described by Vogel and Fennessey (1994). In this method, individual FDCs are calculated for each year and month. Subsequently, a single median value is calculated for each exceedance point using the collection of values at the same exceedance taken from the multiple annual FDCs. These new median values are then plotted to create a single new FDC. The same approach is followed by Metcalfe et.al (2013). Two different temporal aggregations are supplied; one continues which includes all computed values and another one discrete, which includes user defined values for comparative purposes. A daily hydrograph discretized by year and showing the mean and median values, is also prepared by this block.

3.5 Temporal Flow Assessment - Block

The main objective of this block is to graphically show the streamflow variability using monthly and yearly box-plots and violin-plots supplied by {*ggplot2*} library. Graphs are plotted in both linear and logarithmic scales. In the case of the box-plots, the upper whisker extends from the hinge to the highest value that is within 1.5 * IQR of the hinge, where IQR is the inter-quartile range, or distance between the first and third quartiles. The lower whisker extends from the hinge to the lowest value within 1.5 * IQR of the hinge. Data beyond the end of the whiskers are outliers and plotted as points according to the Tukey criteria. These box-plots make it possible to visualize the year-to-year variation of median annual flows around the longer-term median flow for the recorded period. It is also possible to identify patterns of wet and dry years from these box-plots.

### 3.6 Baseflow Separation - Block

Baseflow Separation is accomplished by means of the {*EcoHydRology*} library, which reads a streamflow dataset and produces a baseflow-quickflow dataset using 1, 2 or 3 passes (Nathan and McMahon, 1990). Results are show graphically using normalized and absolute box-plots and hydrographs. The Baseflow Index (BFI), which is the ratio between baseflow and total flow, is also calculated. Therefore a BFI of 0.5 indicates that 50% of total streamflow can be attributed to baseflow for the respective time period, either monthly or yearly.

### 3.7 Flow Assessment Models - Block

Taking advantages of the highly developed machine learning capabilities available in R, specialized functions from the {*stats*} library are used to compute linear models (lm) at yearly temporal resolution of the mean, median and maximum values of both total-streamflow and baseflow. The idea behind this block is to identify increasing or decreasing trends through time. A detailed summary of the model metrics, which includes coefficients and residuals, are presented both graphically and tabulated.

### 3.8 Exports - Block

Relevant outputs from all block are exported in csv, txt and png formats, which allows user to further manipulate data outside the R ecosystem (e.g. MS-EXCEL, STATA).

## 4 RESULTS AND ANALYSIS

Results from the Flood Frequency Analysis Block, suggest a 60 $m^3/s$ peak flow for a 20-year return period (FFATr20) along with its 95% confidence intervals with values between 42 and 93 $m^3/s$ (Figure 3). Projections beyond a 20-year period show considerable uncertainty as observed record is based on a 17-year period. Highflow Events on the other hand (Figure 4), show a total of 25 events exciding a 1.5-year return period (FFATr1.5 = 28.3 $m^3/s$) and only 2 events exciding a 10-year return period (FFATr10 = 50.4 $m^3/s$). This means that the FFATr1.5 is exceeded at least once every year and the FFATr10 is exceed twice in 17 years. This proves the applicability of the R methods to estimate the magnitude of low and high-frequency extreme floods that can be further used in the decision-making process regarding operation of the hydropower facilities in the Upper Toro river catchment. These projections could also be used in hydrodynamic simulations of sluices, derivations and other hydraulic structures along the river bank.

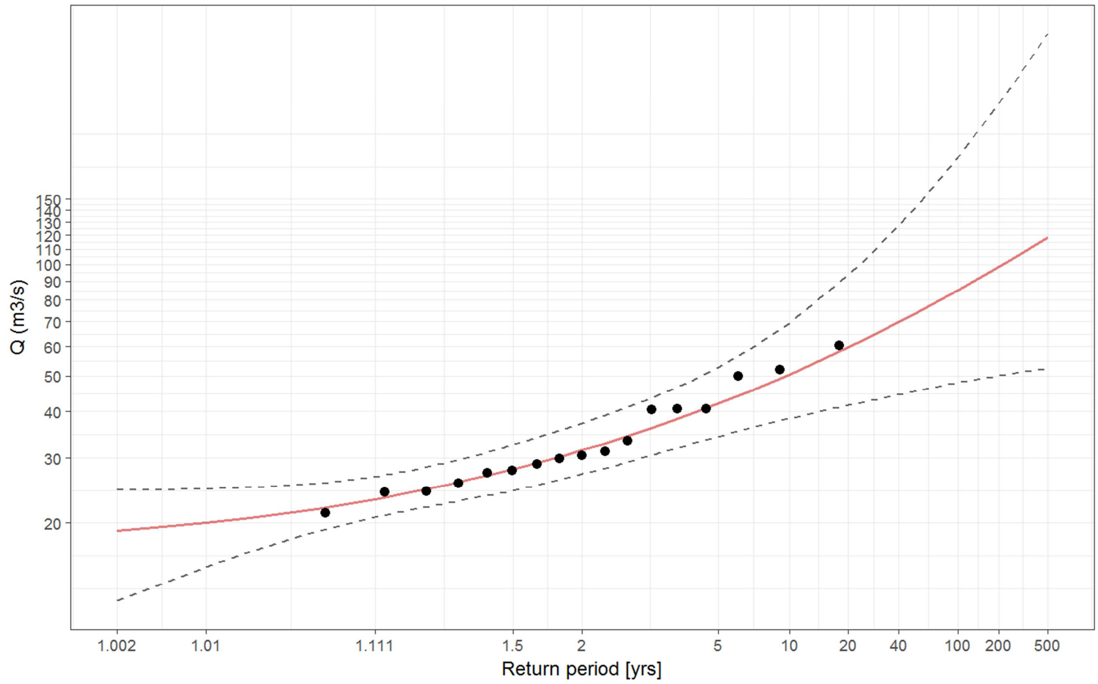


**Figure 3**. Upper Toro river catchment Flood Frequency Analysis (1994-2010). Log Pearson Type 3. Dotted points represent observed values, continuous red line represents the fitted model and grey dash lines represent CI at 95%.

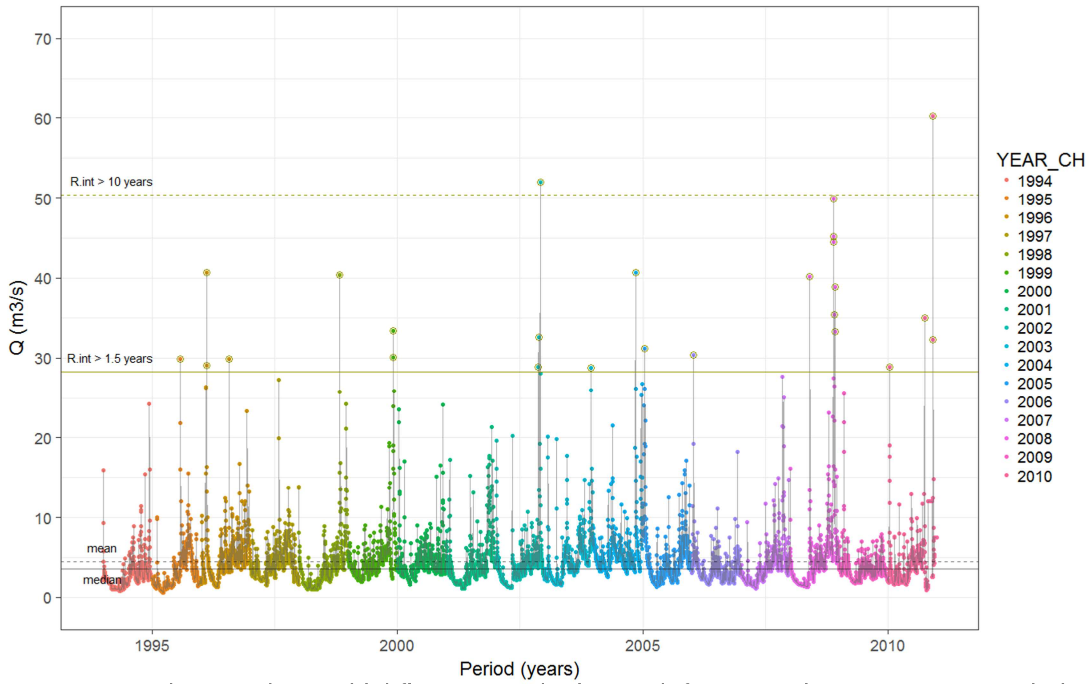


**Figure 4**. Upper Toro river catchment highflow events hydrograph for 1.5 and 10-year return periods.

Concerning Flow Duration Curves (FDCs), the plot shows the percentage of time that flow in the stream is likely to equal or exceed a value of interest, in this case, 80% of the time (Figure 5). It can be seen that the dry season expands from February through April, with flow values as low as 1.5 $m^3$/s, essentially sustained by baseflow. From June until December nonetheless, flow recovers to values above 2.5 $m^3$/s. November on the other hand is the most productive month, with sustained values close to 4.0 $m^3$/s. This behaviour is somehow supported by the daily hydrograph discretized by year (Figure 6), where the daily mean and median trends show lower values from February through April and higher values from October through December. The shape of FDCs in their upper (above 80% of the time) and lower (below 20% of the time) regions is particularly significant in evaluating the stream and catchment characteristics (). In this case, the months of March and May are unable to sustain flow above a minimum threshold of 1.0 $m^3$/s. Oppositely, the months of November and December are most likely to cause flooding due to heavy precipitation, with values high above 50 $m^3$/s.

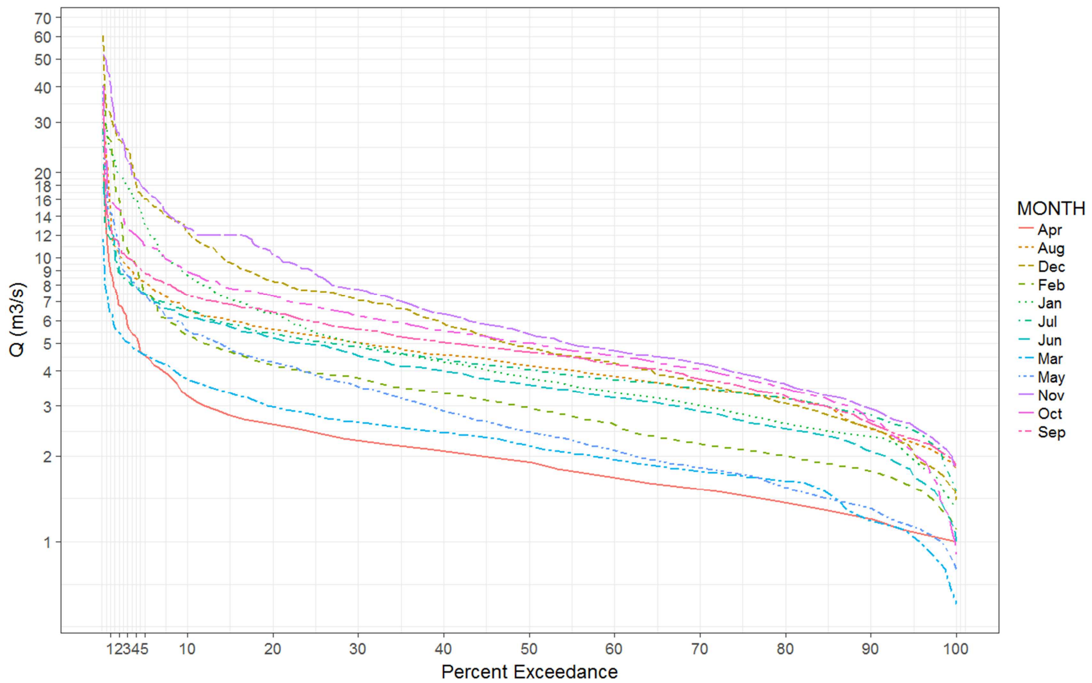


**Figure 5**. Upper Toro River Catchment monthly Flow Duration Curves. Continuous Values (1994-2010).

Regarding Temporal Flow Assessment, daily streamflow variability box-plots discretized by month (Figure 7) confirm what was presented by the FDCs; dry season expands from February through April. Box-plots also show extreme outlier values, either high or low. IQR for each month are clearly presented. Once more, extreme high values (beyond normality) are concentrated from November through January, and extreme low values are concentrated during the month of March. Yearly box-plots on the other hand (Figure 8), show that 1994 and 1995 were particularly dry when compared to the remaining years. This was most likely to be a consequence of El Niño southern oscillation (ENOS). The year 2004 was however, particularly wet as compared to the rest of the period.

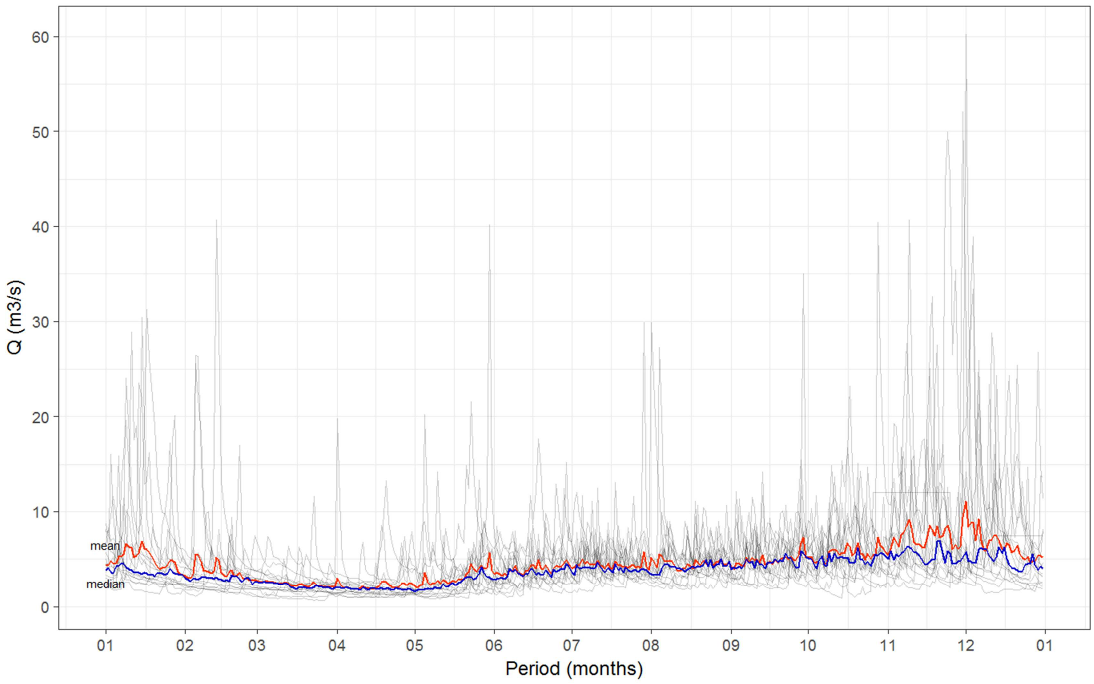


**Figure 6**. Upper Toro River Catchment daily hydrograph discretized by year. Red continuous line represents mean whereas the blue line represents the median.

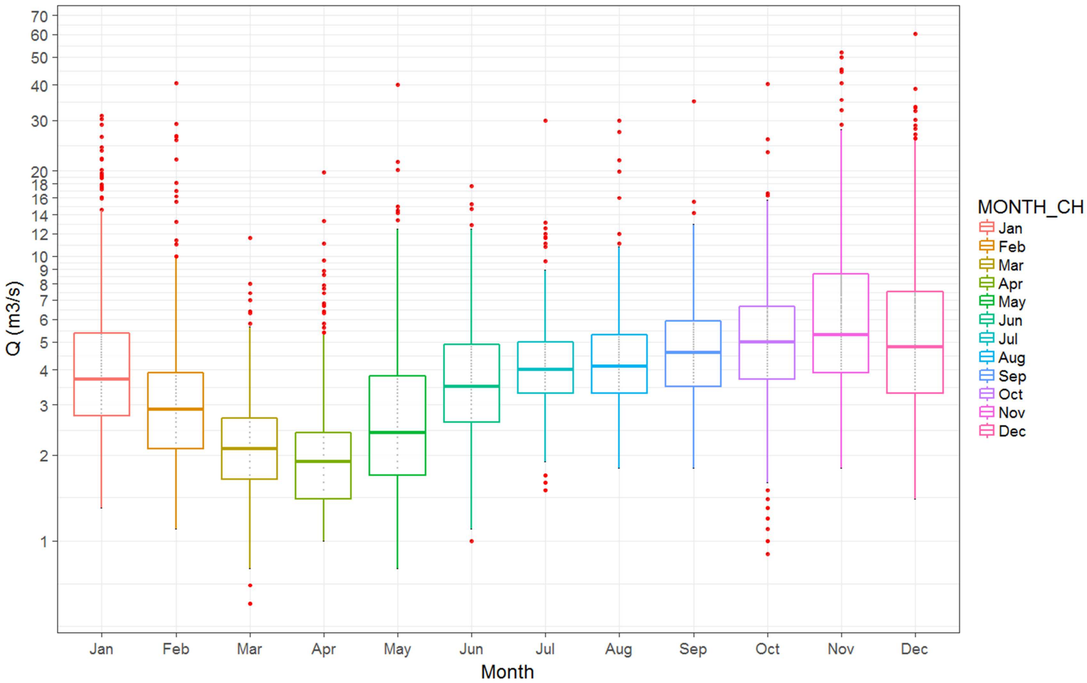


**Figure 7**. Upper Toro River Catchment. Daily box-plots discretized by month (1994-2010).

In relation to Baseflow Separation, the daily Baseflow Index (BFI) box-plots discretized by month (Figure 9), show how much of the total flow can is attributed to baseflow during each month. It can be seen that between the months of May through December, 50% of total streamflow (box-plots interquartile range) present BFI values ranging from 0.12 to 0.40, a considerable contribution to total flow. Special attention however must be

paid to box-plots corresponding to the months of January through April, as many observations appear as outliers (beyond normality) but are in fact, indicators of how relevant was baseflow during those observations. A daily hydrograph in logarithmic scale, along with the median and mean values for the whole period also shows the relative contribution of baseflow to total flow (Figure 10). In this case, baseflow is below the median and mean of the total flow, except during the wet season, where the relative contribution of baseflow considerably increases.
On the subject of Flow Assessment Models, an increasing trend in magnitude with time can be observed for both streamflow maximum-annual (Figure 11) and baseflow maximum-annual values (Figure 12) according to their respective linear models (LM). Even when adjusted R2 values in both cases are considerably low and the confidence intervals broad, the trends themselves are undeniable.  This  increasing  trend  could  suggest changes in the catchment land use or a direct consequence of Climate Change.

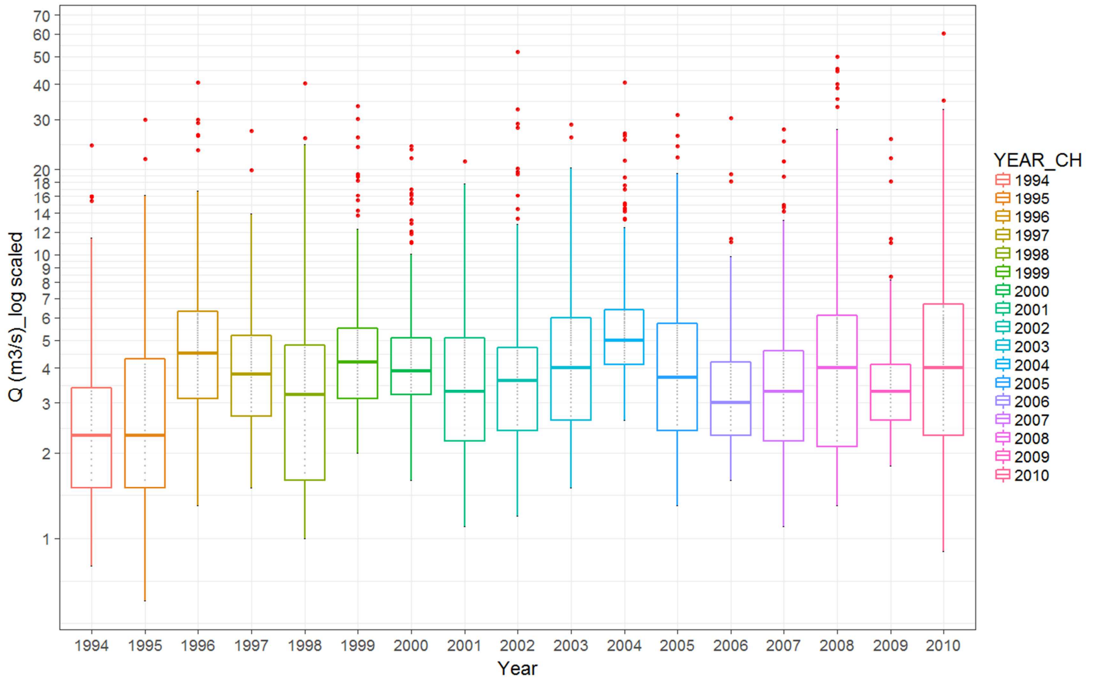


**Figure 8**. Upper Toro River Catchment. Daily box-plots discretized by year (1994-2010).

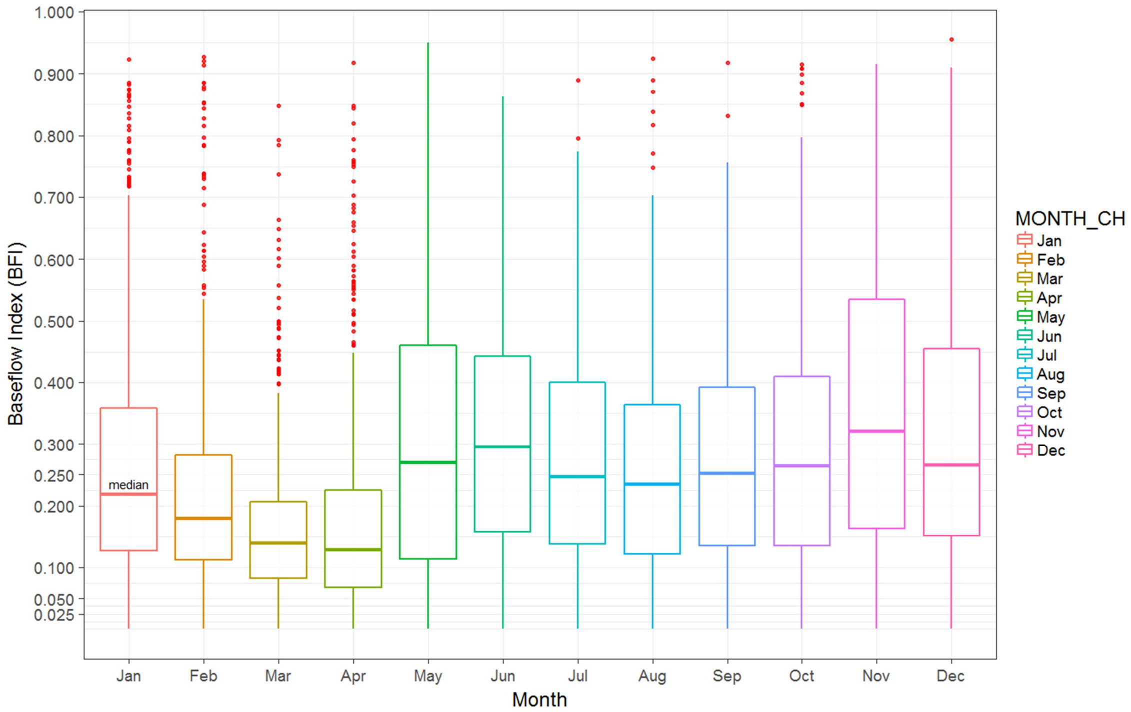


**Figure 9**. Upper Toro River Catchment daily Baseflow Index (BFI) box-plots discretized by month.

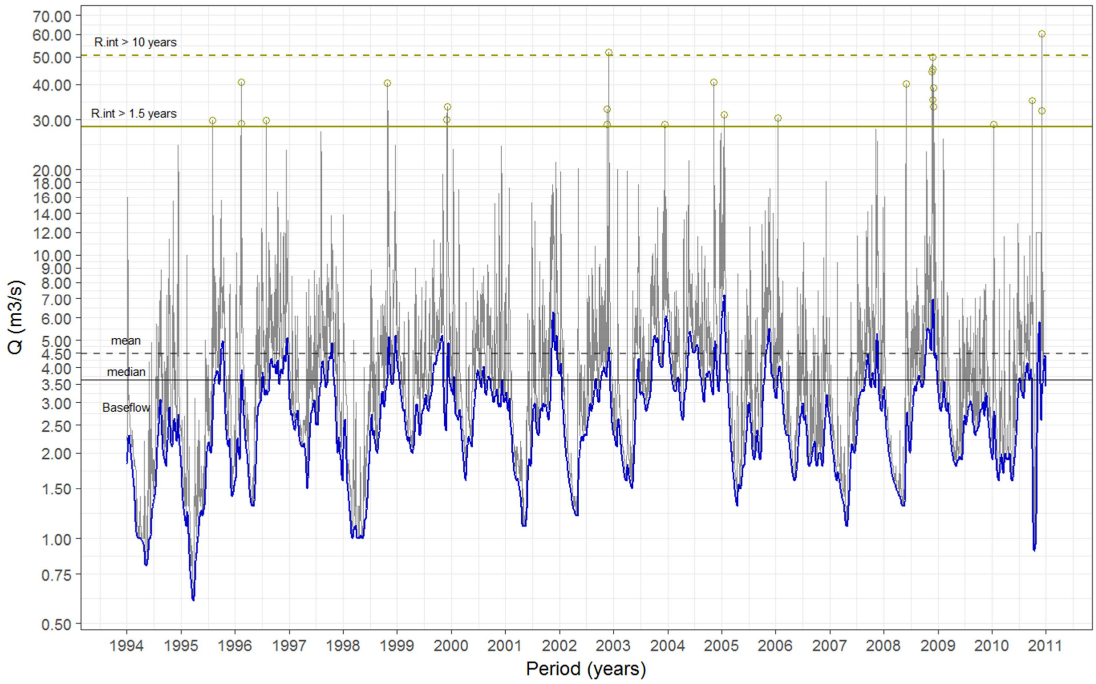


**Figure 10**. Upper Toro River Catchment daily hydrograph in logarithmic scale. Blue continuous line represents the baseflow contribution whereas the grey line represents total flow.

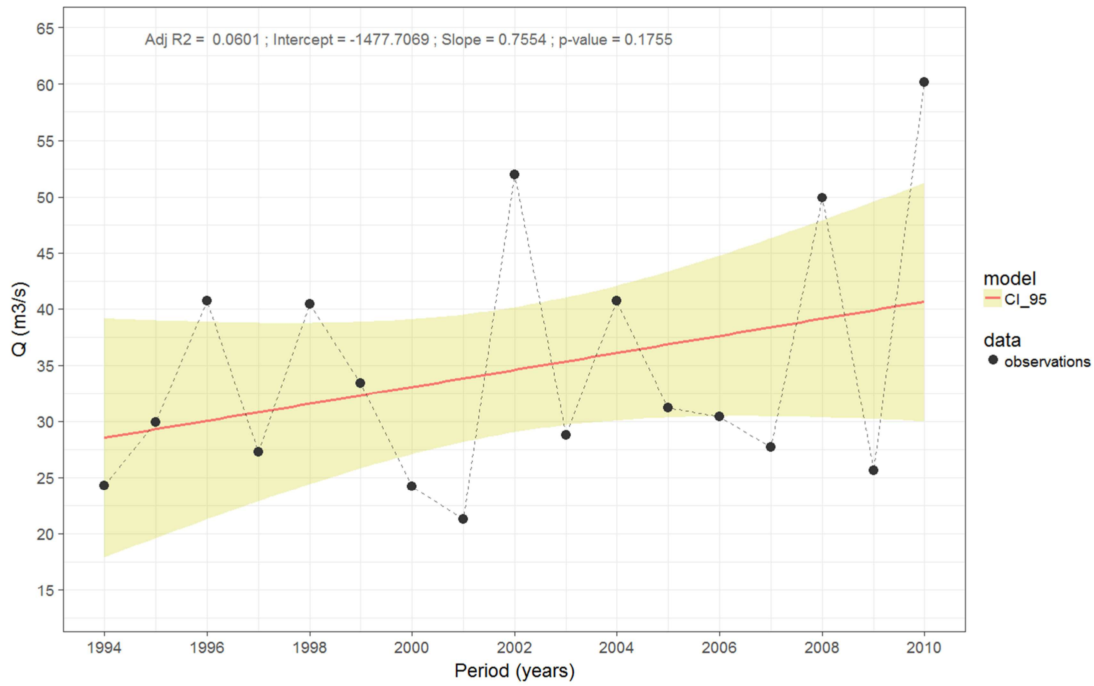


**Figure 11**. Upper Toro River Catchment. Streamflow Annual-Max Linear Model-LM (1994-2010).

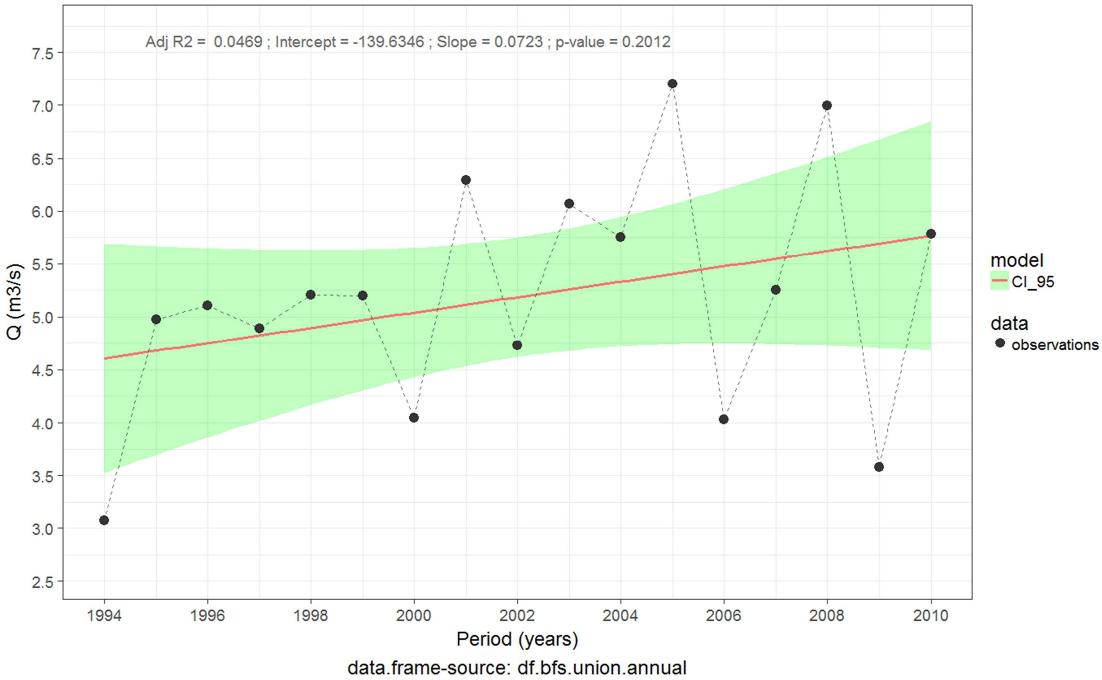


**Figure 12**. Upper Toro River Catchment. Baseflow Annual-Max Linear Model-LM (1994-2010).

## 5 CONCLUSIONS

The Streamflow and Flood Analysis System Using R (SFASUR-TEC), developed using the R programming language, was applied to for the upper Toro River catchment, Costa Rica. The following conclusions can be drawn:

a) SFASUR-TEC represents a flexible and convenient system to identify and quantify key components of flow regimes and assess their behaviour through time.
b) The use of the R programming language allows for a flexible and robust framework that facilitates the incorporation of the data analysis and high-level visualization capabilities available in R into hydrological modelling.
c) Results from the Flood Frequency Analysis Block, suggest a 60 $m^3/s$ peak flow for a 20-year return period (FFATr20) along with its 95% confidence intervals with values between 42 and 93 $m^3/s$.
d) Flow Duration Curves (FDCs) show that the dry season expands from February through April, with flow values as low as 1.5 $m^3/s$, essentially sustained by baseflow.
e) Monthly Temporal Flow Assessment also supports the above conclusion, whereas Yearly Assessment indicates that 1994 and 1995 were the driest years in period considered.
f) Baseflow Separation shows that baseflow is indeed a considerable contribution to total streamflow, with BFI values ranging from 0.12 to 0.40 between the months of May through December.
g) The Flow Assessment Models exhibit an increasing trend in magnitude with time can be observed for both streamflow maximum-annual and baseflow maximum-annual values.

Future development of SFASUR-TEC will focus on:

a) Further validation over additional catchments aiming to verify the efficiency of the code and identify potential improvements.
b) Incorporation of advanced predicting and forecasting time-series analysis techniques available in R including Autoregressive Integrated Moving Averages Models (e.g. ARIMA and ARMA), Seasonal Decomposition and Neural Network Autoregression.
c) Development of an R shiny web user interface for public access.

**ACKNOWLEDGMENTS**

This research was supported by Vicerrectoría de Investigación & Extensión, Instituto Tecnológico de Costa Rica (TEC). The authors are grateful to Centro de Servicio de Estudios Básicos de Ingeniería & Construcción of Instituto Costarricense de Electricidad (ICE) for providing observed data for the Upper Toro River catchment.

**REFERENCES**

Chow, V.T., Maidment D.R. and Mays, L.W. (1988). *Applied Hydrology*. New York: McGraw-Hill.

ICE. (2011). *Centros de Producción Peñas Blancas, Cariblanco, Toro 1 y Toro 2*. Instituto Costarricense de Electricidad (ICE), San José, Costa Rica.

IMN. (2008). *Clima, variabilidad y cambio climático en Costa Rica*. Instituto Meteorológico Nacional (IMN), San José, Costa Rica.

Maidment, D.R. (1993). *Handbook of Hydrology*. New York: McGraw-Hill.

Metcalfe, R.A., Mackereth, R.W., Grantham, B., Jones, N., Pyrce, R.S., Haxton,T., Luce, J.J. and Stainton, R. (2013). *Aquatic Ecosystem Assessments for Rivers (AEAR)*. Science and Research Branch, Ministry of Natural Resources and Forestry, Peterborough, Ontario.

Nathan, R.J. and McMahon, T.A. (1990). Evaluation of automated techniques for base flow and recession analysis. *Water Resources Research*, 26(7), 1465-1473.

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

RStudio Team (2016). *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA URL, 2016. URL http://www.rstudio.com/.

Vogel, R. and Fennessey, N.M. (1994). Flow duration curves I: A new interpretation and confidence intervals. *ASCE, Journal of Water Resources Planning and Management*, 120(4), 485-504.

WEF. (2014). *The Global Energy Architecture Performance Index Report 2014*. The World Economic Forum. REF 031213.