https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

Technical Note: A Visual Diagnostic Framework for Identifying Non-Stationarity and Mixed Populations in Flood Series

Paul H. Whitfield[1] and Donald H. Burn[2]

[1]Centre for Hydrology, University of Saskatchewan, Canmore, AB, Canada
0000-0001-6937-9459
[2]Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, ON, Canada
0000-0001-7917-6380

Correspondence to: Paul H. Whitfield (paul.h.whitfield@gmail.com)

Abstract. Practitioners are commonly faced with conducting flood frequency analysis (ffa) with a specific purpose in mind. They are faced with the temptation to use all the available data and assume that the conditions of ffa are met. Flood frequency analysis relies on the assumptions that the flood time series are: [1] stationary, and, [2] independent, widely known as independent and identically distributed (i.i.d.). It is commonly understood that these conditions do not always exist. In many cases, the sample is composed of mixed populations and low outliers often confuse the analyst by biasing the selection of a distribution. Magnitude outliers may come from a different generating mechanism than the main population of peaks. Timing outliers can also indicate alternative generating mechanisms. A diagnostic framework for visual screening of annual maxima and peaks-over-threshold data is described that can better inform the analyst of the nature of the flood series. This integration allows the identification of mixed populations that are often missed in standard routines.

1 Introduction

Flood frequency analysis usually relies on the assumptions that the flood time series are: [1] stationary (meaning their statistical properties don't change over time), and, [2] independent (each flood event is unrelated to others). This is best known as being iid. It is becoming widely understood that in many cases these conditions do not exist (Waylen and Woo 1982; Cohn *et al.* 2013; Burn and Whitfield 2016, 2025; England *et al.* 2018; Ryberg *et al.* 2020; Vidrio-Sahagún *et al.* 2023, 2024) being combinations of low outliers (Cohn *et al.* 2013) mixed processes (Waylen and Woo 1982; Merz and Blöschl 2003; Fischer *et al*. 2016; Barth *et al*. 2017; Tarasova *et al*. 2019), non-stationary magnitudes (Vidrio-Sahagún *et al.* 2023, 2024), or non-stationary generating processes (Burn and Whitfield 2016, 2025). Screening of a flood data plot should identify high outliers that need careful consideration before any analysis (Rahman *et al.* 2014; Whitfield and Burn 2026). Recent research indicates that, in many places in North America, flood generating processes are increasingly rainfall driven in a warming climate (Burn & Whitfield 2025).

In evaluating whether data are suitable for ffa there are three main issues to consider [1] which flood variable, (instantaneous peak, maximum daily average or peaks over threshold (POT)) should be used, [2] completeness (at best, data for many years and

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

complete years, but, with caution, peaks from partial years), and [3] representativeness, which deals with the assumptions of stationarity and independence and a common generating process. Completeness deals with ensuring the annual peaks are actually the annual maxima, ideally the block maximum from a calendar or hydrological year. Stationarity can be addressed using existing screen methods. Before ffa is conducted on a data set, the analyst needs to check for stationarity and independence and determine whether or not all events are likely to arise from the same flood generating process. Multiple flood-generating processes requires special consideration in ffa (Waylen and Woo, 1982; Iacobellis *et al*. 2010; Burn and Whitfield, 2025). This note focuses on approaches for assessing series for stationarity and missing data, and then assessing and screening for multiple generating processes based on examination of outliers; magnitude (high and low), and timing outliers, and clustering of observations based on magnitude and timing.

Whitfield (2012) emphasized the need for improved data screening tools, especially given the growing interest in hydrometric trends driven by climate change concerns. Before hydrometric data can be used effectively, it should be screened to identify significant trends and any associated changepoints (Dierauer *et al*. 2017). Bard *et al*. (2011) developed a comprehensive screening process that included trend and change-point analysis, but it required multiple software tools and was not user-friendly. Dierauer *et al*. (2017) built upon this work by integrating graphical and computational tools into a single, more accessible package. Analyzing hydrometric data is inherently complex due to the numerous factors that need to be considered. Over time, streamflow may be altered by many influencing factors (Merz *et al*. 2012). Streamflow can be affected by both climatic factors (such as natural variability and climate change) and non-climatic factors (such as regulation and land-use changes) (Dierauer *et al*. 2017). Additionally, detecting temporal trends is challenging because the data often contain change points—abrupt shifts in the statistical properties of the time series (Dierauer *et al*. 2017). Despite its importance, time-series screening is often overlooked (Whitfield 2012). One R-package, FlowScreen (Dierauer and Whitfield 2016, 2025), addresses this gap by offering a streamlined approach to data screening. It includes standard statistical tests such as the Mann-Kendall test, supports change-point detection, and methods for assessing completeness. Similarly, CSHShydRology (Anderson *et al.* 2018) contains several functions that can be used to screen for completeness and data quality.

In this work, we use "outliers" in the sense of being important and interesting rather than sampling errors. In the case of floods, the "outlier" is very likely to be the event of interest (Klemeş 1986, 1989). Hu (1987) argues that magnitude outlier tests in a statistical context postulate an assumption that outliers have a unique distribution, which is different from that of the remaining sample observations. This is particularly relevant for ffa if large outliers are not part of the dominant generating process (i.e. Rogue floods - "Rogue" floods are outliers in more than one of magnitude, timing, and density (Whitfield and Burn 2026)).

Guidelines for flood frequency analysis are well developed and available for many countries (Australia, Brazil, China, India, United Kingdom, United States, and others); however, tools suitable for practitioners that can be used for screening and understanding flood data series before such analysis are generally lacking.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

This Technical Note describes several available methods that have been combined to support screening of annual maxima and peaks-over-threshold time series. Here we describe an R function that provides three plots that can be used to inform the analyst about the important aspects of those data.

2 Methods

The data used here are based on daily mean flow observations downloaded from Water Survey of Canada using Environment Canada Data Explorer (ECDE). The annual maximum for each year was obtained using functions from CSHShydRology; complete and partial years were used for demonstration. Alternately, one might use instantaneous peak values, or peaks-over-threshold. The methods described here can be used for visual screening of either annual maxima series or peaks-over-threshold series. In these examples, the amax or POT series were prepared using functions from the package CSHShydRology.

Screening of time series needs to consider whether the data are representative and whether they conform with the assumptions of ffa. Addressing these visually considers completeness, data quality, stationarity, mixed populations, and outliers and Rogue events.

2.1 Completeness and Quality

Flood frequency analysis depends on long series of representative data. Assessing this depends on considering whether annual maxima are based on entire years or if partial years are suitable.

Many sources of streamflow data provide flags or other systems to communicate about the provenance of individual data points. Of particular interest when considering annual or over threshold peaks are cases where: [1] the value is an estimate where the observed stage during the event exceeds the highest observed stage discharge in the rating curve, [2] the value is based on a partial day of record as could occur when a gauge is destroyed by the event (Pomeroy *et al*. 2016), and, [3] where there is a backwater from downstream or from ice. While these flags are qualitative, they can provide important guidance to the assessment of the data.

2.2 Trend and Randomness

An essential data screening is to plot the data as a time series and assess it for missing data, randomness, and stationarity. A basic question is the support for the sampled time series of peaks (Figure 1). Annual maxima based on less than an entire year might be suitable. The time-series screening plots differentiate between annual maxima based on 365(366) days (solid symbols), and those for partial years (open circles). Rather than presume that the maximum value is properly captured there may be a need to investigate further. It is common to avoid this by removing values where only part years were available; in that case, missing

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

years are counted and pink vertical lines indicate missing years. For POT data, this reflects years without observations greater than the threshold and missing years.

Wald-Wolfowitz is a non-parametric test used to assess randomness against a trend (Wald & Wolfowicz 1943; Vivekanandan 2025). This test considers runs of observations above and below the median. The plot provides the time series with points above the median in blue, below the median in red, and the median in gray. Transitions between above and below are marked in vertical gray dashed lines. Mann-Kendall is a non-parametric test commonly used to test for a trend in a time series. A loess line is added to the plot and the colour and width reflect the Mann-Kendall significance (grey thin line – NS $p > 0.1$, dark red dashed line $p <= 0.1$, red solid line $p <= 0.05$). Text is added to provide the actual p values of both tests.

2.2.1 Magnitude

Flood magnitude is commonly the focus of flood studies (Vogel *et al*. 2001, 2019). As with any data, data should be tested for outliers using different methods (Barnett 1978). Since they may come from a distribution, it is recommended that low outliers in annual maxima should be censored to avoid influencing the fitting of the model of interest (Lamontange *et al.* 2013, 2016; Cohn *et al.* 2013). Cohn *et al.* (2013) identified low outliers that could influence implementation of log-Pearson Type III flood frequency calculations using the Grubbs' test. In this screening, the Multiple Grubbs-Beck Test (MGBT) developed by Cohn *et al.* (2013). Analysts should carefully consider whether retaining low outliers would result in the selection of an inappropriate distribution rather than one that fits the higher events.

Testing for high outliers is commonly done using the Grubbs' test (Grubbs 1950, 1969; Wilkinson 2017). The original Grubbs' test detected outliers normally distributed univariate data. Available versions of the tests detect only one outlier at a time. Testing for more than a single outlier depends on repeated censoring the series and repeating of the test (Wilkinson 2017). The Grubbs' test is valid where the series is randomly sampled from a normal distribution, and the location and scale estimates are unbiased (Wilkinson 2017). We implemented a function that would do the sequential Grubbs' test for up to 15 high outliers; the high values are not censored, but those values that would be considered large outliers are identified. Flood data are seldom normally distributed, so the function allows the Grubbs' test to be applied to log transformed (default), untransformed, and Box-Cox transformed values.

High outliers can be generated by a different process than typical annual maxima. The classic approach is the Grubbs test, and this is implemented so that all high outliers are identified. Cohn *et al.* (2013) provide the method used to identify low outliers.

2.2.2 Timing

When considering flood timing, day of year needs to be treated as a circular variable to accommodate seasonality (Whitfield 2018). Several circular methods are available for this, the Rayleigh test is the most common method for circular data. The Rayleigh test asks "is this observation at a different time/angle from the others?" Another aspect is the concentration of points

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

around the circle, known as κ (kappa); high values of κ indicate that the data is concentrated in one portion of the circle while a κ of 0.0 indicates they are evenly spread ($1/\kappa$, is analogous to σ^2, the variance).

A method of classifying timing outlier using estimators based on the minimum distance for the unknown parameters of a parametric density on the unit sphere was developed by Sau & Rodrigues (2018). These estimators are consistent and asymptotically normally distributed and allows detection of potential atypical values in flood event timing. Timing outliers are identified using the technique described in Sau and Rodrigues (2018). This method was implemented based on code shared by Daniela Rodrigues (personal communication).

2.2.3 Density using Fuzzy Clustering

Fuzzy clustering addresses magnitude and timing simultaneously and has the added benefit of determining how many clusters are needed to describe the data and also outliers. Whitfield (2018) explored methods that could be used to cluster floods based on timing. Fuzzy clustering coupled to outlier detection is a recent development. Cebeci (2019) and Cebeci *et al.* (2017, 2022) proposed a methodology using possibilistic algorithms that allows each data point to belong to multiple clusters with different membership measures, distinct from each point being assigned to a single cluster as in k-means. The advantage of this fuzzy clustering is that timing and magnitude outliers can be determined based on a measure of membership of a point to every cluster (Cebeci *et al*. 2017).

The value of kappa is a measure of concentration of the data about the circle that can be used to inform clustering. High values indicate that the data are concentrated in one direction. When multiple directions are common, as in mixed distributions, the values of κ are low. By default, when κ is less than 5, fuzzy clustering is used to determine how many clusters are needed to describe the data (from 2 to 5). Clustering always infers there are at least two groups; using the κ threshold lets a single population be a valid result (Whitfield 2018).

3 Visualization

Three visualization approaches to flood screening are presented and these can be applied to both annual maxima and POT series. In plots where POT data were used, they are so labelled. For each visualization method, two examples are shown, one for annual maxima and the second for POT. The supplementary material provides these three plots for four different sites and for both annual maxima and POT. The supplementary material covers a much broader range of cases than can be covered here.

The first visualization provides a time series plot of the data and information about quality, trends, and changepoints that might affect subsequent analysis. In these plots, solid points indicate complete years of data; open points indicate annual maxima from years with some part of the year's observations being missing. Years without observations are marked with vertical pink lines, and

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

the number of missing years, or years with no POT events is noted in the upper left in red text. Points that are flagged using a "SYM" code from Water Survey of Canada have a colored square surrounding that point ["A" partial day (green), "B" backwater (cyan), "D" dry (orange), and "E" estimated (red)].

Following from Wald-Wolfowitz, points above the median are blue, below the median red, and the median value(s) in grey. Dashed vertical lines mark Wald-Wolfowitz shifts between above and below the median and the significance of the Walf-Wolfowitz test against trend is given in the lower left. A loess line shows the tendency of the series, and it is coloured based on the Mann-Kendall trend test reported in the lower right. If the MK is not significant a thin gray line is shown, if p <0.10 a dark red line is shown, and if the Mann-Kendall p value is <0.05 a thicker red line is shown. The results of the Pettitt test for a changepoint are shown in the upper right. The changepoint from the Pettitt test is shown as a vertical line using the same colours for p values as Mann-Kendall. When the time series plot is amax the pink lines indicate years without data; when the series is POT they indicate years without observations above the threshold and would include missing years and any years where there were no peaks greater than the threshold.

The second visualization plots the flood series, either annual maxima or peaks over threshold (POT) against the plotting position of a relevant distribution and information about seasonality, low, high, and timing outliers that might affect subsequent analysis. For annual maximum series, Gumbel plotting position is used. For POT series, the plotting position used is Pareto. These plots show the confidence intervals and the estimate based on the distribution and the magnitude of peaks on a return period axis. This basic element is common in plots used in flood frequency analysis; plot also shows results from tests for low, high, and timing outliers using symbols that indicate an observation that is a high, low, and/or timing outlier. In addition, the symbols are coloured indicating months of the year thus indicating seasonal distributions of floods and a histogram of these is provided in a legend displaying the colours and the distribution of annual peaks across months. A second legend includes the counts of high, normal, low, and timing outliers.

The third visualization of the flood series is a polar plot with information on the temporal distribution, mixed populations and outliers that might affect subsequent analysis. The polar plot shows the seasonal distribution of floods as a histogram around the outer rim, and their relative magnitude on the radial axis (i.e. the outer rim has the magnitude of the flood-of-record. High and low outliers are marked with symbols (triangle and square) on the main plot and timing outliers are shown as red squares on the outer rim. The flood of record is highlighted. Also included are two kernel density estimates (bandwidths of 15 (red) and 50 (green) as a visual indication of the seasonality. Details about the flood of record are listed with the counts of the outliers and are shown below the plot.

Kappa κ is a measure of the concentration of data around the mean and high values indicate concentration and low spread. On the polar plots, when the value of kappa is less than the κ threshold (5), a fuzzy clustering is conducted and the points for the annual maxima are coloured by the cluster number (from 2 to 5 clusters are possible), and fuzzy clustering outlier cases are coloured red.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

Highly concentrated cases, i.e. $\kappa > 5$ are considered to have only one cluster. The information in the polar plot can be used to consider the nature and timing of mixed distributions and the potential impact of outliers.

Burn and Whitfield (2016, 2025) and Whitfield and Burn (2026) describe the background of these methods and other alternatives relating to timing and density outliers in more detail.

4 Selected examples

One amax series and one peaks-over-threshold series are used to demonstrate this screening approach. Two flood data sets are shown in the Figures, one annual maximum series and one peaks-over-threshold that covers the natural scope of flood frequency screening. The supplementary material provides examples of annual maxima and POT series for four rivers. The reader is encouraged to review the supplementary material as those examples cover all aspects of the screening which is not possible with a limited number of figures.

In the first example, Adams River at Squilax, BC, (Figure 1a) shows there are no years without observations (missing years), the Wald-Wolfowitz and Mann-Kendall tests indicate no trend, and the Pettit test indicates no significant changepoint. The return period plot (Figure 2a) indicates that all the annual peaks occur in May to July and there are no magnitude outliers. All flood events occur in May/June/July and there are timing outliers (9) but these are not of high magnitude. The polar plot for this site shows that the timing outliers are those events at the margins of this main period (Figure 3a). The annual maxima are highly concentrated in one time period as indicated by a high value for κ (>4) and there is likely only one population of events. This suggests that this annual maximum series should be suitable for standard flood frequency analysis.

The second example, peaks-over-threshold data for the St John River at Fort Kent (Figure 1b), indicates one year without observations, which in this case is the result of the threshold used being larger than the smallest annual maximum. There is no indication of trends with either Wald-Wolfowitz or Mann-Kendall, nor a changepoint. The return period plot (Figure 2b) indicates that for this POT data there are many low outliers (97), one high outlier, and many timing outliers of low magnitude (42). The large number of low outliers indicates that the threshold used is too low. The histogram in Figure 2b shows that most flows occur in April-May but also that events do occur in the summer and fall. This suggests that there may be mixed populations of events. The polar plot (Figure 3b), shows that while the peak events are concentrated in May, there are some POT events from July to January. Fuzzy clustering indicates two clusters, one in May and another in October-November with some cluster outliers in July-August and December. Low outliers occur in both clusters. The number of low outliers suggests that flood frequency analysis should consider censoring the low outliers (Robson and Reed 1999; Cohn *et al.* 2013; England *et al.* 2018) and the analyst be aware that there are likely two populations of peaks that may be from alternate generating processes (see Burn and Whitfield 2025, Whitfield and Burn 2026).

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

Example plots for annual maxima and POT series from four rivers are provided in the supplementary material, that illustrate and reinforce the results shown in Figures 1-3.

A checklist for ffa includes:

1. Check the time series plots;
    a. Identify and assess years with missing observations, or years with less than an entire year of observations.
    b. Assess the potential impacts of quality assurance flags.
    c. Determine if there are apparent trends in the magnitude that might compromise flood frequency analysis (Vormoor *et al*. 2015; Vidrio-Sahagún *et al*. 2024). Trends may be absent in the annual maximum series, but be induced in POT series by shifting distributions of peaks (Burn and Whitfield 2025).
    d. Determine if there is a significant changepoint; a changepoint may indicate a methodological change, or a structural change related to changes in the seasonal distribution of peaks.
2. Examine the return period plot;
    a. Determine if there are low outliers. Low outliers may be influential in selecting an incorrect distribution of flood events (Cohn *et al.* 2013).
    b. Determine if there are high outliers. High outliers may indicate that the statistical distribution is fat tailed, but if there are many high outliers there may be a mixed population (Vogel *et al*. 2019; Whitfield and Pomeroy 2016).
    c. Examine timing outliers in relation to high and low outliers. Timing outliers, particularly of high magnitude events, may indicate separate rare “Rogue” events (Whitfield and Burn 2026).
3. Consider polar plot;
    a. Consider the kernel density lines in relation to the main population of events that may be a single cluster or multiple clusters indicating single or multiple generating processes.
    b. Assess where high and low outliers occur seasonally. High outliers in different seasons may indicate multiple processes that demand more complicated flood frequency analysis (Whitfield and Burn 2026).

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

c. Consider timing outliers which may be indicative of "Rogue" events (Whitfield and Burn 2026) or shifting processes (Burn and Whitfield 2025). Rogue events are rare but demand consideration from a regional perspective (Whitfield and Burn 2026).

d. Assess the fuzzy clustering which may be supportive of considering mixed population ffa (Barth *et al.* 2017, Yu *et al*. 2022, Burn and Whitfield 2025) or of trends in flood generating processes (Burn and Whitfield 2025).

Screening before flood frequency analysis should take all of the above into consideration; knowledge of mixed and/or changing flood generating processes may indicate that a combined distribution approach may be required to reflect different flood generating mechanisms.

In the examples presented here the sample size is large. The methods incorporated have limitations and sensitivity to small samples. The methods used are suitable for natural rivers and streams and applying them to regulated or urbanized basins should be done with caution. Selection of a threshold for POT analysis is an area of active research and the screening presented will address only some of the selection issues, such as low outliers. Our intention was to provide a screening that was simple to implement to support the decisions that need to be made when undertaking flood frequency analysis.

While the statistical tests are not new, the *synthesis* of these tests into a graphical screening framework provides a critical diagnostic that will help prevent the misuse of ffa. This framework addresses two important problems:

Practitioner Bias - the tendency to subconsciously (or consciously) force the data to fit a preconceived model or a desired outcome. A practitioner may be tempted to use every year of record available, ignoring the fact that early data might come from a different climatic regime or a less reliable gauge. A practitioner might ignore visual evidence that the data doesn't follow a standard distribution. When requiring a higher 100-year flood value there may be a bias regarding high outliers. Conversely, dismissing them when the desire is to minimize costs.

Outlier Confusion - which occurs when the analyst cannot distinguish between statistical noise and physical reality. In arid or semi-arid regions, "dry years" can produce very small annual maxima. These "zeros" or near-zeros can severely bias the slope of a frequency curve, making the 100-year estimate look much smaller than it actually is. Is a flood that is larger than anything else in the record a measurement error, or is it a "Rogue" event? A flood that occurs in October in a basin where floods *always* happen in June (snowmelt). Without screening, an analyst might treat a massive rain-on-snow event (like the 2013 Calgary flood, Pomeroy *et al*. 2016) and a standard spring melt as part of the same "population." a clear violation of the i.i.d. assumption.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

5 Summary

A visualization based "diagnostic framework" is provided that integrates methods that identify missing and problematic data and mixed populations that are sometimes missed by standard automated routines. Screening data is a necessary step in any analysis (Whitfield 2012). This is particularly true for flood frequency analysis (Kidson and Richards 2005), where the underlying assumptions are often violated:

- stationarity of the flood series
- independence of individual events
- homogenous population of floods indicating a common generating mechanism

The screening presented supports assessment of data completeness and quality. The screening can identify magnitude and timing outliers that need careful consideration before any analysis (Rahman *et al.* 2014) and also provides a way to determine if there are mixed processes (Waylen & Woo 1982). Our recent research indicates that, in many places in North America, flood generating processes are becoming more mixed in a warming climate (Burn & Whitfield 2025). By putting all tests together, an analyst can't ignore the conflict between a *good fit* and a *non-stationary trend* and it forces the analyst to look at the *timing* and *type* of the flood by separating "outliers" into their proper physical populations. This complements (Merz *et al*. 2015) who called for more scientific rigour in dealing with flood time series.

Code and data availability

The function described here, `ch_ffa_screen`, is available in the R-package `CSHShydRology`. Supporting functions (`SR_htest, ch_high_Grubbs_test, mretlev_uvplot` [derived from package `POT`] are contained within that package. The function ch_ffa_screen returns a list containing the information about the dataset, and the tests conducted, and a dataframe containing the data and outlier indexes. There are options to allow the user to choose the type of data transformation, and kappa thresholds. In addition to the plots, the function returns the details of the incorporated tests.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

Author contributions

PHW and DHB designed the visualization framework. PHW developed the model code with input from DHB. PHW prepared the manuscript with contributions from DHB.

Competing interests

The authors declare no competing interests.

Acknowledgements

Many colleagues have provided useful discussions that have informed this work. These include Malcolm Clark, Laurent de Rham, Alain Pietroniro, Kevin Shook, Cuauhtémoc Vidrio-Sahagún, and Stilian Stoev.

Financial support

This research was partially supported by funding provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada First Research Excellent Fund's Global Water Futures program, the Canada Research Chairs program and Environment and Climate Change Canada.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

References

Anderson, E. P., Chlumsky, R., McCaffrey, D., Trubilowicz, J. W., Shook, K. R., and Whitfield, P. H.: R-functions for Canadian hydrologists: a Canada-wide collaboration, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 44, 108-112, 10.1080/07011784.2018.1492884, 2018.

Bard, A., Renard, B., and Lang, M.: The AdaptAlp Dataset: Description, guidance, and analyses, Cemagraf, Lyon, France, 15, 2011.

Barnett, V.: The study of outliers: purpose and model, Journal of the Royal Statistical Society: Series C (Applied Statistics), 27, 242-250, 10.2307/2347159, 1978.

Barth, N. A., Villarini, G., Nayak, M. A., and White, K.: Mixed populations and annual flood frequency estimates in the western United States: The role of atmospheric rivers, Water Resources Research, 53, 257-269, 10.1002/2016WR019064, 2017.

Burn, D. H. and Whitfield, P. H.: Changes in floods and flood regimes in Canada, Canadian Water Resources Journal, 41, 139-150, 10.1080/07011784.2015.1026844, 2016.

Burn, D. H. and Whitfield, P. H.: Shifting cold regions streamflow regimes in North America affect flood frequency analysis, Hydrological Sciences Journal, 70, 51-70, 10.1080/02626667.2024.2422531, 2025.

Cebeci, Z.: Comparison of internal validity indices for fuzzy clustering, Journal of Agricultural Informatics/Agrárinformatika, 10, 1-14, 10.17700/jai.2019.10.2.537, 2019.

Cebeci, Z., Kavlak, A. T., and Yildiz, F.: Validation of fuzzy and possibilistic clustering results, 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 16-17 Sept. 2017, 10.1109/IDAP.2017.8090183, 2017.

Cebeci, Z., Cebeci, C., Tahtali, Y., and Bayyurt, L.: Two novel outlier detection approaches based on unsupervised possibilistic and fuzzy clustering, PeerJ Computer Science, 8, e1060, 10.7717/peerj-cs.1060, 2022.

Cohn, T. A., England, J. F., Berenbrock, C. E., Mason, R. R., Stedinger, J. R., and Lamontagne, J. R.: A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series, Water Resources Research, 49, 5047-5058, 10.1002/wrcr.20392, 2013.

Dierauer, J. and Whitfield, P.: FlowScreen: Daily streamflow trend and change point screening (1.2) [code], 2016.

Dierauer, J. and Whitfield, P. H.: A new version of the R-package FlowScreen, Water News, 44, 24-39, 2025.

Dierauer, J., Whitfield, P. H., and Allen, D. M.: Assessing the Suitability of Hydrometric Data for Trend Analysis: The 'FlowScreen' package for R, Canadian Water Resources Journal, 42, 269-275, 10.1080/07011784.2017.1290553, 2017.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

England Jr, J. F., Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas Jr, W. O., Veilleux, A. G., Kiang, J. E., and Mason, J. R. R.: Guidelines for determining flood flow frequency — Bulletin 17C, Reston, VA, Report 4-B5, 168, 10.3133/tm4B5, 2018.

Fischer, S., Schumann, A., and Schulte, M.: Characterisation of seasonal flood types according to timescales in mixed probability distributions, Journal of Hydrology, 539, 38-56, 10.1016/j.jhydrol.2016.05.005, 2016.

Grubbs, F. E.: Sample criteria for testing outlying observations, Annals of Mathematical Statistics, 21, 27-54, http://www.jstor.org/stable/2236553., 1950.

Grubbs, F. E.: Procedures for detecting outlying observations in samples, Technometrics, 11, 1-21, 10.1080/00401706.1969.10490657, 1969.

Hu, S.: Problems with outlier test methods in flood frequency analysis, Journal of Hydrology, 96, 375-383, 10.1016/0022-1694(87)90167-3, 1987.

Iacobellis, V., Fiorentino, M., Gioia, A., and Manfreda, S.: Best fit and selection of theoretical flood frequency distributions based on different runoff generation mechanisms, Water, 2, 239-256, 2010.

Kidson, R. and Richards, K. S.: Flood frequency analysis: assumptions and alternatives, Progress in Physical Geography, 29, 392-410, 10.1191/0309133305pp454ra, 2005.

Klemeš, V.: Dilettantism in hydrology: Transition or destiny? Water Resources Research, 22, 1778-1788, 10.1029/WR022i09Sp0177S, 1986.

Klemeš, V.: The improbable probabilities of extremes floods and droughts, in: Hydrology of Disasters, edited by: Starosolszky, O., and Melder, O. M., James and James, London 43-51, 10.4324/9781315073583, 1989.

Lamontagne, J. R., Stedinger, J. R., Yu, X., Whealton, C. A., and Xu, Z.: Robust flood frequency analysis: Performance of EMA with multiple Grubbs-Beck outlier tests, Water Resources Research, 52, 3068-3084, 10.1002/2015WR018093, 2016.

Merz, B. and Blöschl, G.: A process typology of regional floods, Water Resources Research, 39, 10.1029/2002WR001952, 2003.

Merz, B., Vorogushyn, S., Uhlemann, S., Delgado, J., and Hundecha, Y.: More efforts and scientific rigour are needed to attribute trends in flood time series, Hydrology and Earth System Sciences 16, 1379-1387, 10.5194/hess-16-1379-2012, 2012.

Pomeroy, J. W., Stewart, R. E., and Whitfield, P. H.: The 2013 flood event in the Bow and Oldman River basins; causes, assessment, and damages Canadian Water Resources Journal, 41, 105-117, 10.1080/07011784.2015.1089190, 2016b.

Rahman, A. S., Haddad, K., and Rahman, A.: Impacts of outliers in flood frequency analysis: A case study for Eastern Australia, Journal of Hydrology and Environment Research, 2, 17-13, 2014.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

Robson, A. J. and Reed, D.: Statistical procedures for flood frequency estimation. Volume 3 of the Flood Estimation Handbook, 1999.

Ryberg, K. R., Kolars, K. A., Kiang, J. E., and Carr, M. L.: Flood-frequency estimation for very low annual exceedance probabilities using historical, paleoflood, and regional information with consideration of nonstationarity, US Geological Survey2328-0328, 2020.

Sau, M. F. and Rodriguez, D.: Minimum distance method for directional data and outlier detection, Advances in Data Analysis and Classification, 12, 587-603, 10.1007/s11634-017-0287-9, 2018.

Tarasova, L., Merz, R., Kiss, A., Basso, S., Blöschl, G., Merz, B., Viglione, A., Plötner, S., Guse, B., and Schumann, A.: Causative classification of river flood events, Wiley Interdisciplinary Reviews: Water, e1353, 10.1002/wat2.1353, 2019.

Vidrio-Sahagún, C. T., He, J., and Pietroniro, A.: Nonstationary hydrological frequency analysis using the Metastatistical extreme value distribution, Advances in Water Resources, 176, 104460, 2023.

Vidrio-Sahagún, C. T., Ruschkowski, J., He, J., and Pietroniro, A.: A practice-oriented framework for stationary and nonstationary flood frequency analysis, Environmental Modelling & Software, 173, 105940, 2024.

Vivekanandan, N.: Effect of data length on estimation of peak flood discharge using L-moments of five probability distributions, Natural Hazards, 1-20, 2025.

Vogel, R. M., Castellarin, A., Matalas, N. C., England, J. F., and Zafirakou, A.: Hydrologic record events, in: Statistical analysis of hydrologic variables: Methods and applications, edited by: Teegavarapu, R. S., Salas, J. D., and Stedinger, J. R., 491-536, 2019.

Vogel, R. M., Zafirakou-Koulouris, A., and Matalas, N. C.: Frequency of record-breaking floods in the United States, Water Resources Research, 37, 1723-1731, 10.1029/2001WR900019, 2001.

Vormoor, K., Lawrence, D., Heistermann, M., and Bronstert, A.: Climate change impacts on the seasonality and generation processes of floods in catchments with mixed snowmelt/rainfall regimes: projections and uncertainties, Hydrology and Earth System Sciences, 19, 913-931, 10.5194/hess-19-913-2015, 2015.

Wald, A. and Wolfowitz, J.: An exact test for randomness in the non-parametric case based on serial correlation, The Annals of Mathematical Statistics, 14, 378-388, https://www.jstor.org/stable/2235925, 1943.

Waylen, P. and Woo, M.-K.: Prediction of annual floods generated by mixed processes, Water Resources Research, 18, 1283-1286, 10.1029/WR018i004p01283, 1982.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

Whitfield, P. H.: Why the provenance of data matters: Assessing “Fitness for Purpose” for environmental data, Canadian Water Resources Journal, 37, 23-36, 2012.

Whitfield, P. H.: Clustering of seasonal events: A simulation study using circular methods, Communications in Statistics - Simulation and Computation, 47, 3008-3030, 10.1080/03610918.2017.1367805, 2018.

Whitfield, P. H. and Burn, D. H.: Rogue and Extreme Floods in North America, Journal of Hydrology, 2026.

Whitfield, P. H. and Pomeroy, J. W.: Changes to flood peaks of a mountain river: implications for analysis of the 2013 flood in the upper Bow River, Canada, Hydrological Processes, 30, 4657-4673, 10.1002/hyp.10957, 2016.

Wilkinson, L.: Visualizing big data outliers through distributed aggregation, IEEE transactions on visualization and computer graphics, 24, 256-266, 10.1109/TVCG.2017.2744685, 2017.

Yu, G., Wright, D. B., and Davenport, F. V.: Diverse physical processes drive upper-tail flood quantiles in the US mountain west, Geophysical Research Letters, 49, e2022GL098855, 10.1029/2022GL098855, 2022.

a) b)

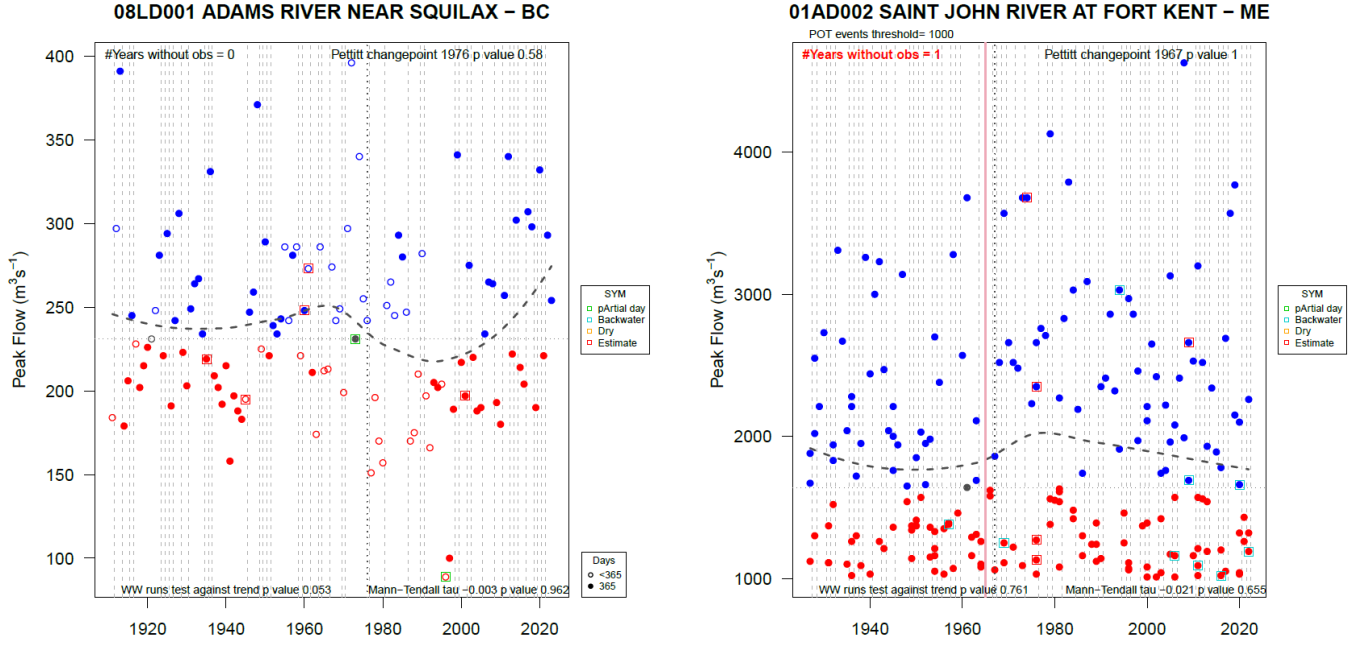


Figure 1. Annotated time series plots (a) annual maximum series from 08LD001 Adams River near Squilax, BC, and (b) peaks-over-threshold series for 01AD002 St John River at Fort Kent, ME. Amax series with partial years have open symbols. Years

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

without observations are shown as pink vertical lines. Coloured squares mark points that are flagged by Water Survey of Canada as partial days, backwater, dry, or estimated. Points above/below the median are shown in blue/red following Wald-Wolfowitz test against trend. The loess line indicates tendency, and shows significance of Mann-Kendall trend test (NS – dotted gray, $p < 0.10$ - dashed dark red, or $p < 0.05$ – solid red line). A vertical line shows the changepoint for the Pettit test following the same scheme. Text indicates the number of years without observations if > 0 (upper left), Pettit test (upper right), Wald-Wolfowitz test (lower left), and Mann-Kendall (lower right).

a) b)

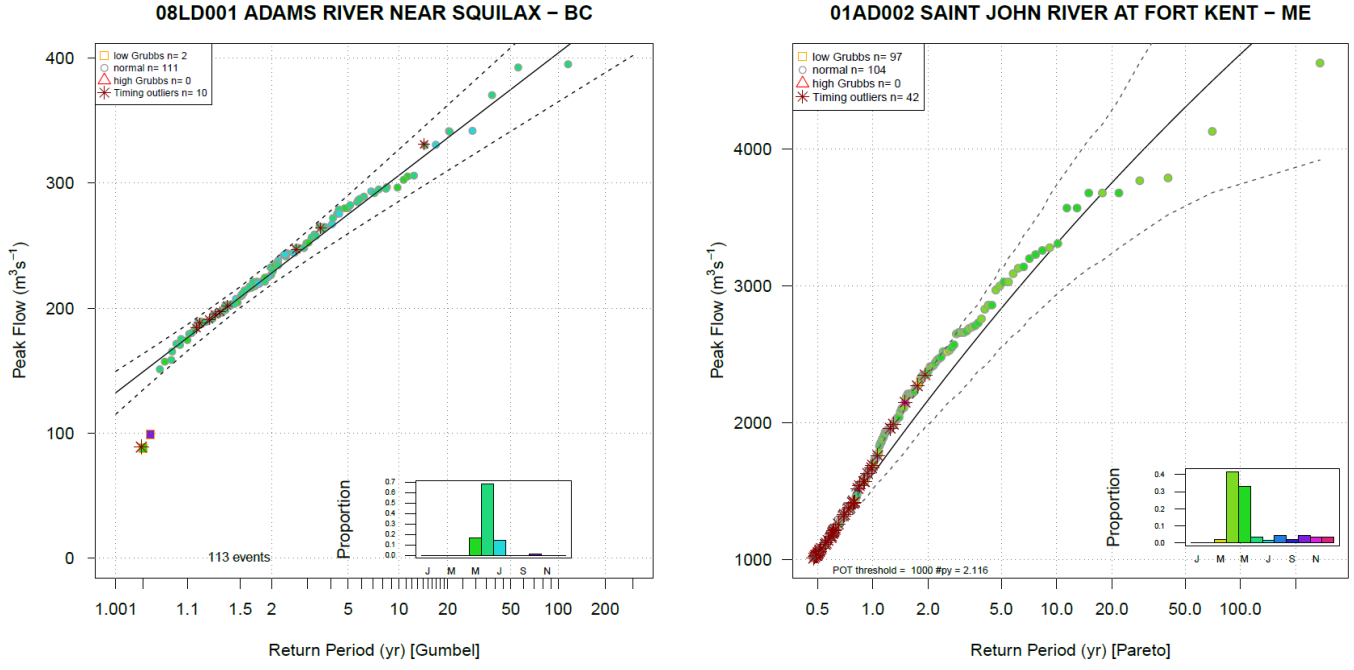
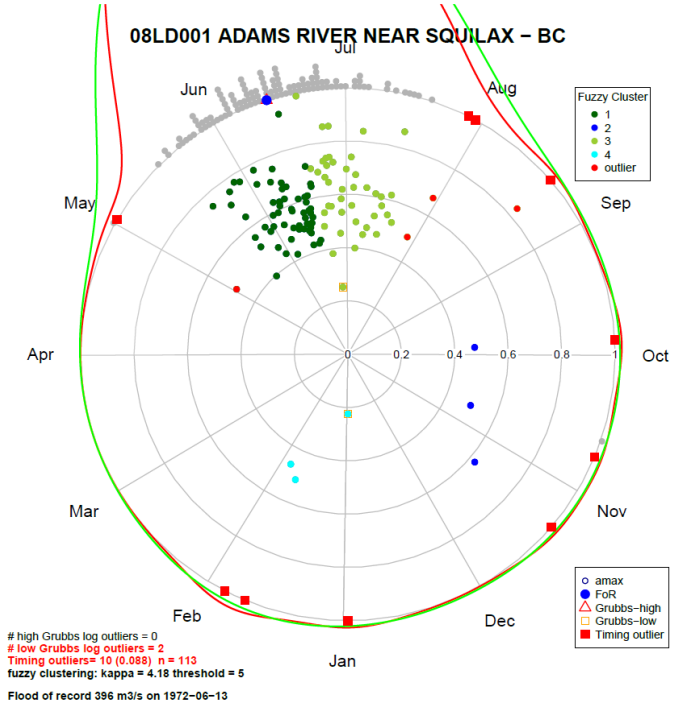


Figure 2. Return period plots showing the magnitude of annual maximum flood peaks (a) annual maximum series from 08LD001 Adams River near Squilax, BC, and (b) peaks-over-threshold series for 01AD002 St John River at Fort Kent, ME. The annual maxima series is shown using Gumbel, and the peaks-over-threshold using a Pareto. Commonly used plots in flood frequency analysis, here include tests for high, low, and timing outliers. Symbols indicate whether an observation is a high (triangle), low (square), or timing (asterisk) outliers. The colored symbols indicated the months of the year thus indicating seasonal distributions of floods and a histogram of these is placed in a corner of the plot to show the distribution of peaks across months. The legend includes the counts of normal observations and high, low, and timing outliers. In this case, the number of low outliers in (b) suggests that the POT threshold was set too high.

https://doi.org/10.5194/egusphere-2026-2067
Preprint. Discussion started: 22 April 2026

a)



b)

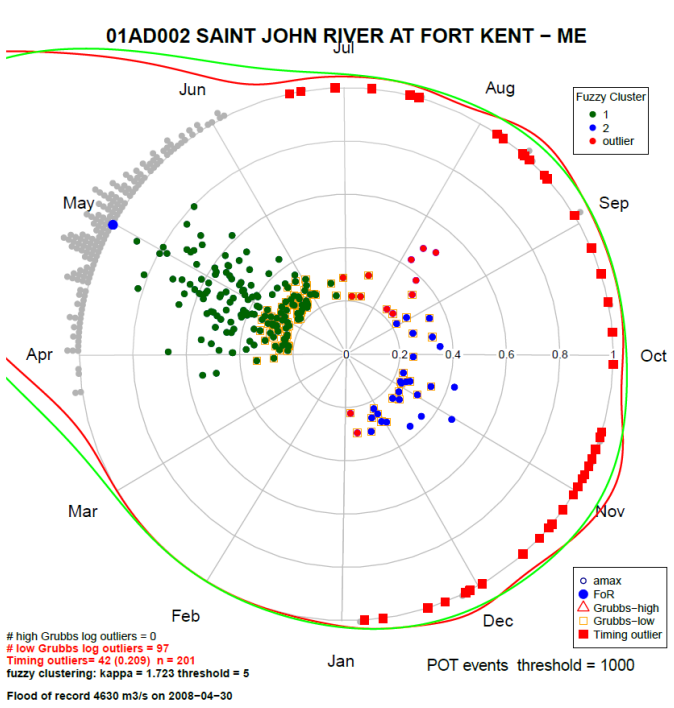


Figure 3. Flood series as a polar plot for (a) annual maximum series from 08LD001 Adams River near Squilax, BC, and (b) peaks-over-threshold series for 01AD002 St John River at Fort Kent, ME. The plots show the seasonal distribution of floods and their relative magnitude on the radial axis. High (red triangle) and low (orange square) outliers are marked with surrounding symbols on the main plot. Timing outliers are shown (red solid squares) on the unit circle. The flood of record (FoR) is highlighted as a blue dot, and a histogram of individual events is on the outside of the circle. Two kernel density estimates (bandwidths of 15 and 50, red and green respectively) as a visual indication of the seasonality. Fuzzy clustering is performed when κ is less than 5, and individual maxima are coloured based on cluster membership shown in the legend in the upper right and may include cluster outliers. Details about the flood of record with the counts of the outliers are shown below the plot.