Taylor & Francis
Taylor & Francis Group

Canadian Water Resources Journal

ISSN: 0701-1784 (Print) 1918-1817 (Online) Journal homepage: https://www.tandfonline.com/loi/tcwr20

EXPLORATORY DATA ANALYSIS

Rory M. Leith , Keith W. Hipel & Herman Goertz

To cite this article: Rory M. Leith , Keith W. Hipel & Herman Goertz (1991) EXPLORATORY DATA ANALYSIS , Canadian Water Resources Journal, 16:1, 81-92, DOI: 10.4296/cwrj1601081

To link to this article: https://doi.org/10.4296/cwrj1601081

EXPLORATORY DATA ANALYSIS

Submitted October, 1989; accepted August, 1990

Rory M. Leith[1], Keith W. Hipel[2] and Herman Goertz[3]

Abstract

Exploratory data analysis techniques are used to detect trends and other statistical characteristics in nine streamflow time series at both the annual and monthly levels. For convenience of interpretation, the output from the analysis is displayed graphically, along with some numerical results from appropriate statistical tests. As well as providing indications and statistical tests of trends, non-normal behaviour and autocorrelated behaviour in flow sequences, exploratory data analysis may be used to place any particular response or collection of responses in context against the range of observed values, thus indicating periods of unusual flow conditions.

Résumé

On utilise ici des techniques d'analyse exploratoire de données pour déceler les tendances et autres caractéristiques statistiques dans neuf séries de débits chronologiques, tant par an que par mois. Pour faciliter l'interprétation, les résultats de l'analyse sont indiqués graphiquement, accompagnés de quelques données numériques provenant de tests statistiques. Tout en fournissant des indications et des tests statistiques sur les tendances, le comportement non normal et le comportement «autocorrélé» dans les séquences de débit, l'analyse permet de placer toute réponse ou tout ensemble de réponses particulières en contexte par rapport à la gamme des valeurs observées, en indiquant les périodes de conditions de débit inhabituelles.

Annual Exploratory Data Analyses

Although exploratory data analyses can be used for revealing a range of statistical properties contained in a data set, this paper emphasizes the detection of trends. To detect long term trends, annual series are examined. Subsequent to discovering trends in annual series, trends within each season of the year are investigated.

To illustrate exploratory data analysis methods for annual series, the results for the Grand River at Galt, Ontario (station number 02GA003 in Table 2) are presented. For this station, the time series of annual mean discharges are displayed in Figure 1. This graph is produced by the program TSCAT. The horizontal axis contains the observation numbers which start in 1913, while the vertical axis gives the magnitudes of the observed values in cubic metres per second (cms). Each of the observations is marked by a cross and successive observations are joined by straight lines. When no cross appears

1. Water Resources Branch, Pacific and Yukon Region, Inland Waters, Environment Canada, Vancouver, British Columbia
2. Department of Systems Design, University of Waterloo, Waterloo, Ontario
3. Water Resources Branch, Ontario Region, Environment Canada, Guelph, Ontario

Table 1: MH Programs Used in this Study

Smoothed Scatter Plot	(TSCAT)
Blurred Smooth Plot	(SMOOTH)
Spectral Analysis	(CUP)
Autocorrelation Function	(USID)
Normal Probability Plot	(NPLOT)
Box and Whisker Plot	(BWPLOT)
Periodic Autocorrelation	(PARID)

above the observation number, this means that the measurement is missing. At the bottom of Figure 1 the number of missing values is provided.

The 95% confidence limit boundaries, which are the horizontal lines at 16 and 54, are derived by assuming that the series of annual flows is normally independently distributed (NID) about the mean level. Since five of the 72 observed values fall outside the 95% confidence limits, there may be some question as to whether the series is NID or white. White sequences are characterized by having no autocorrelation or in the frequency domain a flat power spectrum. A good discussion is provided by Bendat and Piersol, 1971. The normality and independence hypotheses are further explored by other graphs.

When using TSCAT, one way to detect

Table 2: Streamflow Time Series Used In The Exploratory, Data Analysis Study

Water Survey of Canada Station Number	Station Name	Province	Drainage Area km^2	Period of Record
02EA005	North Magnetawan near Burk's Falls	ON	321	1915-1985
02FC001	Saugeen River near Port Elgin	ON	3960	1914-1985
02GA003	Grand River at Galt	ON	3520	1913-1985
02HL001	Moira River near Foxboro	ON	2620	1915-1985
04LJ001	Missinaibi River at Mattice	ON	8940	1920-1983
08HA002	Cowichan River at Lake Cowichan	BC	596	1913-1985
08NE039	Big Sheep Creek near Rossland	BC	347	1929-1985
08NK005	Elk River at Phillips Bridge	BC	4450	1924-1985
08NP001	Flathead River at Flathead	BC	1110	1929-1985

trends is to examine the smooth of the observations. The smooth, labelled RS50 in Figure 1, is obtained using robust locally weighted regression (Cleveland, 1979). In the label, RS stands for robust smooth and the number is a measure of the neighbourhood over which the smoothing takes place. Larger values (neighbourhoods) produce smoother curves. This is demonstrated in Figure 2 which is a plot of Grand River at Galt annual flows, but with a smoothing parameter of 1.00 rather than 0.50 used in Figure 1. In both figures there is an increasing trend with time.

Another method to detect trends is to use the non-parametric Mann-Kendall test (Mann, 1945; *Hirsch et al.*, 1982). Below the graphs in Figure 1 and 2 are the estimated value of Mann-Kendall rank correla-

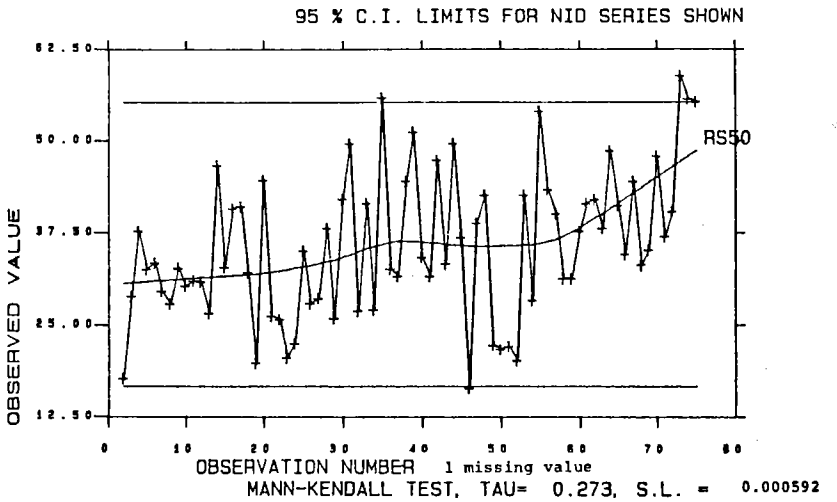

Figure 1: Time Series Plot of Annual Flows for 02GA003 Grand River at Galt for 1913-1987 (Smoothing Parameter 0.50).

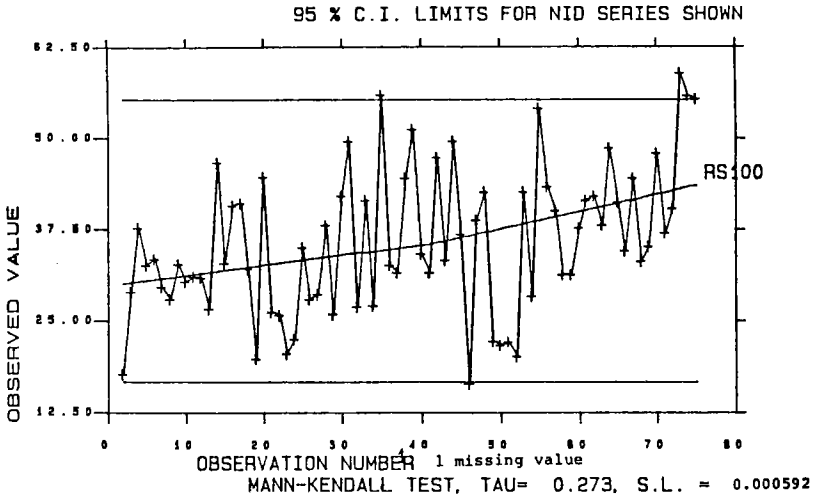

Figure 2: Time Series Plot of Annual Flows for 02GA003 Grand River at Galt for 1913-1987 (Smoothing Parameter 1.00).

tion statistic (tau) along with its significance level (SL). Because the SL 0.00059 is much less than 0.05, one can reject the null hypothesis that the observations are independent and identically distributed and entertain the alternative hypothesis that there is a monotonic trend. Since tau is positive (0.273), the trend is increasing.

TSCAT can also produce another scatter plot which is useful in detecting dependencies among the observations. In Figure 3 annual mean discharges for the Grand River at Galt for year t are plotted against annual mean discharges for year t-1 for the period 1913 to 1987. The robust locally weighted regression smooth labelled RS50 indicates an increasing trend. At the bottom of the plot, the Mann-Kendall rank correlation statistic (tau) along with its significance level are printed. As the SL is (0.013) less than 0.05, the annual mean discharges separated by 1 year do appear to be dependent.

To check if the observations in a series are uncorrelated, the cumulative periodogram (Jenkins and Watts, 1968; Bartlett 1955) can be used. The cumulative periodogram, produced by the program CUP, for the annual Grand River at Galt flows from 1914 to 1987 is displayed in Figure 4. In this figure, the 75%, 90%, 95% and 99% confidence limits are plotted parallel to the line drawn from (0,0) to (0.5,1). The graph for the cumulative periodogram falls well inside the confidence limits, indicating the annual series to be white noise.

A graph of a time series may blur patterns in the data which a smoothed plot may reveal (Tukey, 1977). In Figure 5, which is produced by program SMOOTH, the Tukey blurred 3RSR smooth indicates an increasing trend, although this may be biased by the very high last 2 annual flows. The Tukey blurred smooth must be used with evenly spaced data and the smoothing is achieved by repeated medians of 3 (called 3R). The SR in 3RSR refers to repeated splitting. The vertical lines in Figure 5 indicate the median of the absolute values of the roughs for all points,

$$\text{Rough} = \text{Data} - \text{Smooth (i.e. 3RSR)}. \quad (1)$$

The smooth observation is located at the midpoint of the bar.

Another way to check for the presence of correlation in a series is to examine a graph of the sample autocorrelation function (ACF) (Jenkins and Watts, 1968). Figure 6, produced by program USID, shows the correlation coefficients at various lags for the series of annual flows for the Grand

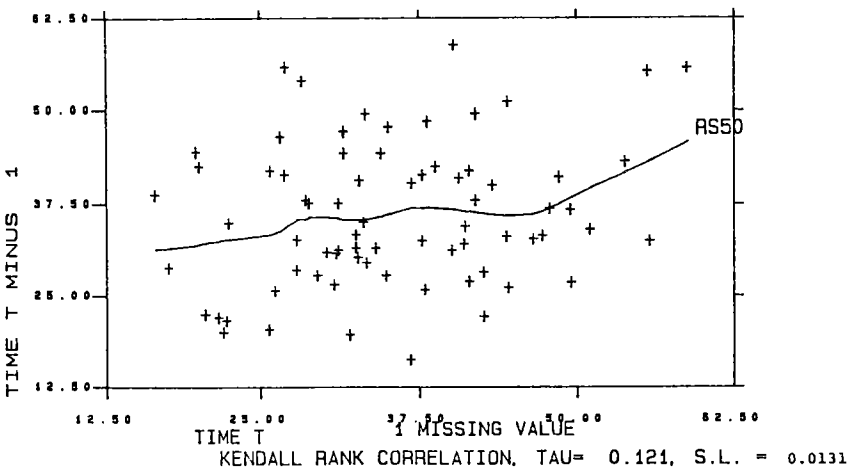

Figure 3: Lag One Scatter Plot for Annual Flows for 02GA003 Grand River at Galt 1913-1987.

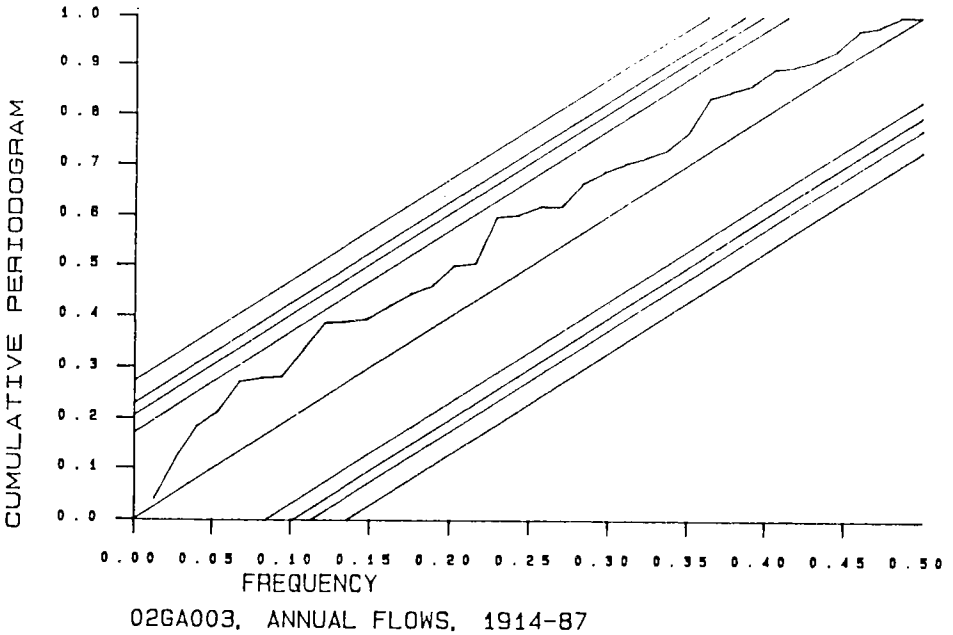

Figure 4: Cumulative Periodogram of Annual Flows for 02Ga003 Grand River at Galt (1913-1987).

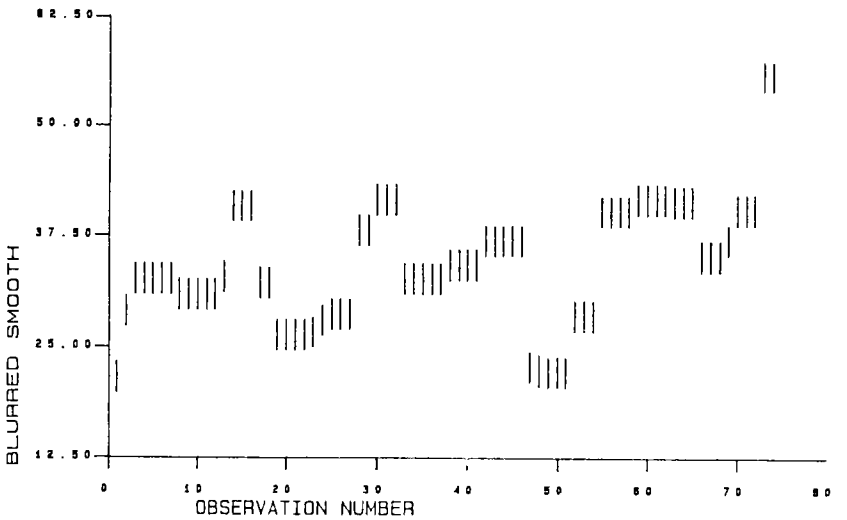

Figure 5: Blurred Smooth Plot of Annual Flows for 02Ga003 Grand River at Galt (1913-1987).

River at Galt. This program requires equally spaced data. The horizontal lines at -0.22 and 0.22 are the 95% confidence limits for the correlation coefficients. They are calculated from Bartlett's approximation (Box and Jenkins, 1976 page 35). Since the values for the sample ACF lie inside the 95% confidence limits, the series can be considered to be uncorrelated.

In addition to dependence, the distribu-

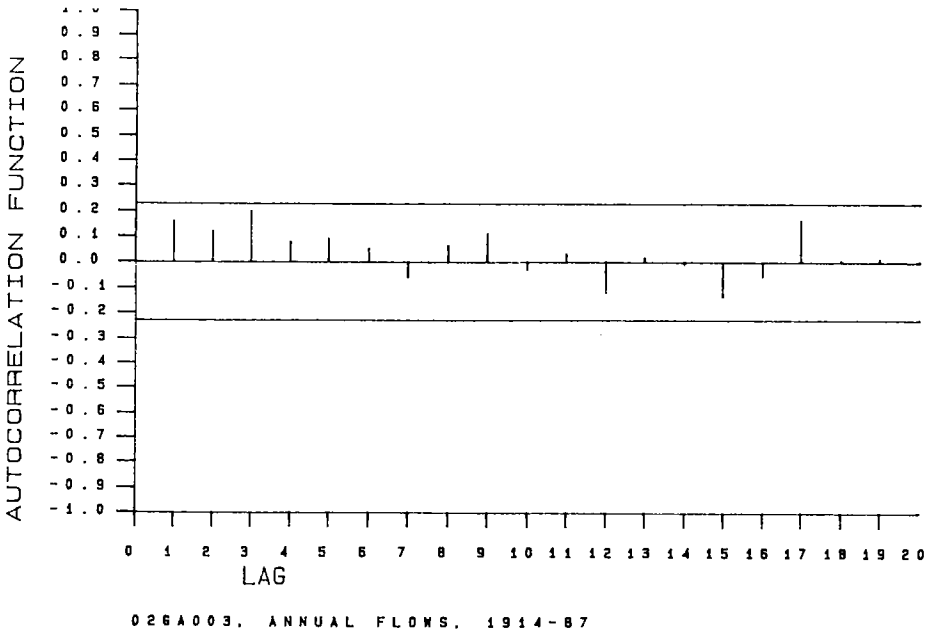

Figure 6: Sample Autocorrelation Function for the Series of Annual Streamflow for 02GA003 Grand River at Galt.

tion of the data may be of interest. The program NPLOT produces the normal probability plot (Looney and Gulledge, 1985) and statistics shown in Figure 7. In this figure the abscissae are the empirical quantities calculated using the annual mean quantiles. The annual mean discharges appear normally distributed as the points plot along the diagonal line in Figure 7 and inside the Kolmogorov-Smirnov 95% confidence limits.

The Box and Whisker graph (Tukey, 1977) which is drawn below the normal probability plot provides another method for examining the distribution of the annual flows. Each cross represents a data point whose value can be determined by referring to the abscissa scale. The vertical line inside the rectangle is the median, while the left and right ends (hinges) of the box represent the halfway points from the median to the appropriate extreme. The extreme whiskers are the horizontal lines protruding from the hinges of the box to the extreme values. Because the box and whisker graph is fairly symmetric, the annual flows appear to follow a symmetric distribution such as a normal distribution.

At the top of Figure 7 are a series of statistics which allow hypothesis testing. F1 and F2 are the coefficients of skewness (D'Aqostino, 1970) and kurtosis respectively. Assuming that the data are normal one can calculate the significance level (SL) for each of these statistics. The SLs of 0.347 and 0.469 for F1 and F2, respectively, are greater than 0.05 which indicates no significant departure from normality.

The statistics labelled as W at the top of Figure 7 refers to the Shapiro-Wilk's statistic (Shapiro and Wilk, 1965). The estimated value of 0.975 has an SL of 0.405, which is much larger than 0.05. This test indicates the annual Grand River flows are normally distributed.

Blom's correlation (Looney and Gulledge, 1985) coefficient provides another test of normality. In Figure 7 this statistic has a value of 0.995 and a significance level greater than 0.10 which is obtained from the tabulated empirical percentage points

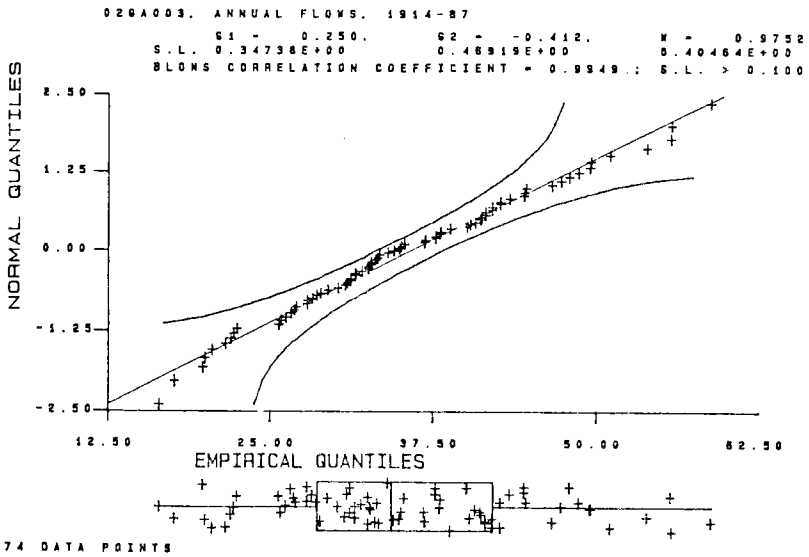

Figure 7: Normal Probability Plot for the Sequence of Annual Flows for 02GA003 Grand River at Galt.

printed in Looney and Gulledge's (1985) paper. Blom's correlation coefficient does confirm the data are normal.

By scanning the results produced by the various plots and test statistics one can summarize the main statistical properties of the data. In particular, the annual Grand River streamflow series has an upward trend (Figures 1 and 2), is white noise (Figures 4 and 6) and is normally distributed (Figure 7). The main statistical findings for all the annual series are listed in Table 2.

Seasonal Exploratory Data Analysis

The seasonal (i.e. monthly) flow series for the stations listed in Table 2 can be subjected to a range of exploratory data analyses including: time series plots, monthly box-and-whisker graphs and periodic autocorrelation function and partial autocorrelation function plots.

These tests and plots produced a large mass of material, so selected results for the 08NP001 Flathead River at Flathead and the 02GA003 Grand River at Galt are presented.

Figure 8 shows a time series plot of September (season 9) flows for the Flathead River. There are two flows outside the 95% confidence intervals; the larger is for 1951. The robust locally weighted regression smooth labelled RS91 clearly shows an upward trend. This trend is confirmed by the Mann-Kendall test statistics, significance level 0.0028 and Mann-Kendall tau is 0.273 indicating a significant increasing trend. The only other months for which the trends are significant at the 5% significance level are July (tau = 0.162, SL = 0.077) and August (tau = 0.181, SL = 0.48). Since the tau values are positive for these months the trends are increasing.

The Fisher's combination method (Fisher, 1970) can be used to ascertain if there is a significant trend across a set of months such as July, August and September or across an entire year. When the results for m seasons are combined, Fisher shows that:

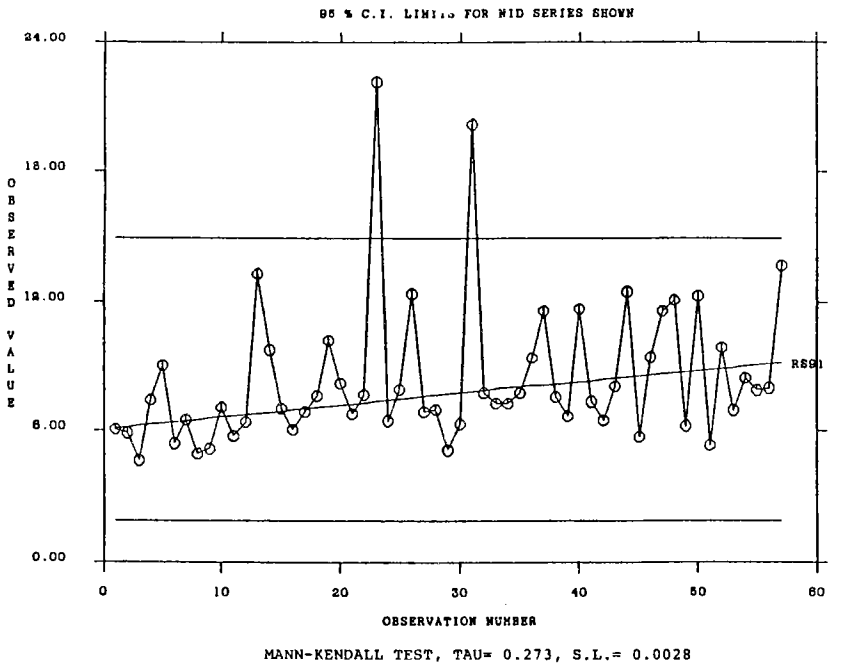

Figure 8: Time Series Plot of September Flows for 08NP001 Flathead River at Flathead.

$$-2\sum_{i=1}^{m} \ln SLi \qquad (2)$$

where ln is log to base e

approximately follows a Chi-square distribution with 2 m degrees of freedom.

For the Flathead River series for 12 months the Fisher's combination is x^2 = 42.51 on 24 degrees of freedom which has a significance level of 0.0113. Consequently, across all months there is an increasing trend which is significant.

Although for the high flow months of May and June there is no significant trend.

For the Flathead River at Flathead, the Mann-Kendall test indicates no significant trend in the annual series. So the results for the annual series and the Fisher's combination of all months are contradictory. When contradictions such as this occur, the output for both the seasonal and annual series should be carefully compared.

The annual flows for the Grand River at Galt show increasing trend. This behaviour is also detected by the using seasonal Mann-Kendall tests. There are significant increasing trends for all months except March, April, and May. Figure 10 shows the increasing trend for September flows. Fisher's combination across all months indicates a significant upward trend.

The results of seasonal Mann-Kendall trend tests are summarized in Table 4. Further investigation is suggested by the contradictory results for the records from stations 02HL001Moira River near Foxboro and 08HA002 Cowichan River at Lake Cowichan.

Box and Whisker plots allow season to season examination of the behaviour of different streams. Figures 10, 11, and 12 depict the Box and Whisker plots of monthly flows for three streamflow stations in Ontario. Inspection of these three plots shows some interesting attributes of these series.

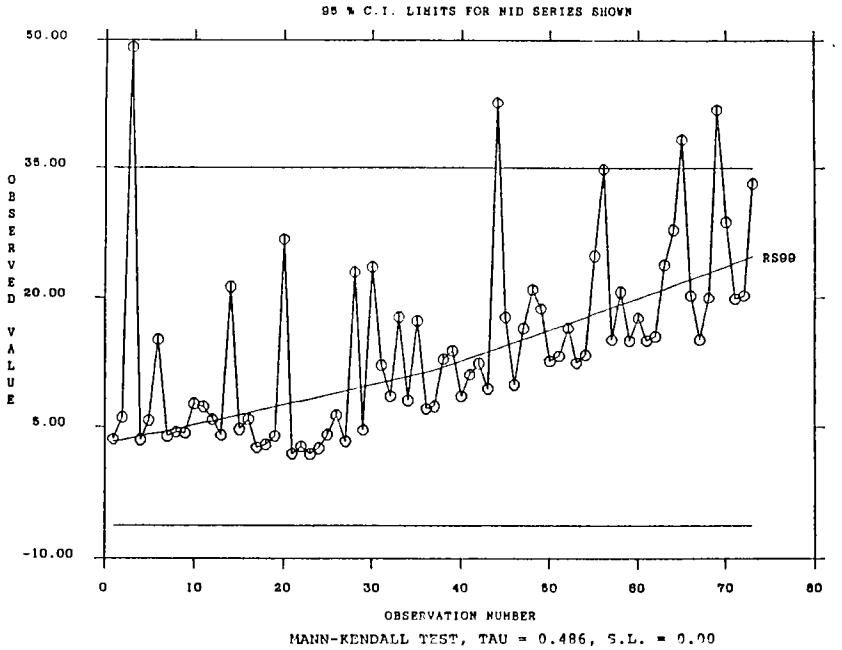

Figure 9: Time Series Plot of September Flows for 02GA003 Grand River at Galt.

Table 3: Statistical Results for the Nine Annual Riverflow Series

Station Number	Trend	White Noise	Normally Distributed
02EA005	No There is a large outlier in 1928	Yes	Yes
02FC 001	Slight increasing trend	Yes	Yes
02GA003	Increasing trend	Yes	Yes
02HL001	No	Yes	Yes
04LJ001	Robust locally weighted regression and Tukey smooth suggest that data collected from 1920-30 differ statistically from the other data and should be carefully checked		
08HA002	No	Yes	Yes
08NE039	No	Yes	Yes
08Nk005	No	Yes	Yes
08NP001	No	Yes	Yes

The Box and Whisker plot for 02AD008 Nipigon River at Pine Portage shows an even distribution across the months, with the median level and variability similar for all months (Figure 10). As would be suspected from this pattern of monthly flows, the flows of this river at this site are highly regulated. This station gauges a river with relatively large flows and which is located in the Lake Superior drainage basin of Northern Ontario.

The Box Cox Parameter (Box and Cox, 1964), indicated as 1.0 in Figure 10, indicates that no transformation of data has taken place. The units of flow on the vertical axis are cubic metres per second.

The outside values and the far out values at the top of Figure 10 are explained by fences. H-spread is the difference between values of the hinges, and a step is a 1.5 times the H-spread. The inner fences are one step beyond the hinges and outer fences are two steps beyond the hinges. Data between the inner fence and outer fence are outside values. Data outside the outer fence are far out.

Figure 11 shows the monthly Box and Whisker (B-W) plots for station 02BC004 White River below White Lake. This basin is located in the Lake Superior drainage basin but flows are considerably less than for the Nipigon River at Pine Portage and the flow regime is closer to natural. The B-W plot for this station exhibits large flows in May and June and lower flows in other months.

Another use of the B-W plot is to assess the flows of any given month or year against the historical flow data. This can be achieved by plotting a symbol for the particular monthly flow values onto the B-W plot. The box-whisker diagram elements summarize the historical flow data for each month and the symbol for the particular month is plotted onto the diagram at the appropriate flow level. This immediately indicates in which quartile the flow value falls or if the flow value is excessively large or small. This is shown in Figure 10 for the Nipigon River at Pine Portage. The particular monthly flow values for 1987 are shown with small circles on the B-W plot.

The extremely low monthly flows for 1987 (usually the lowest on record) (May through December) are immediately demonstrated.

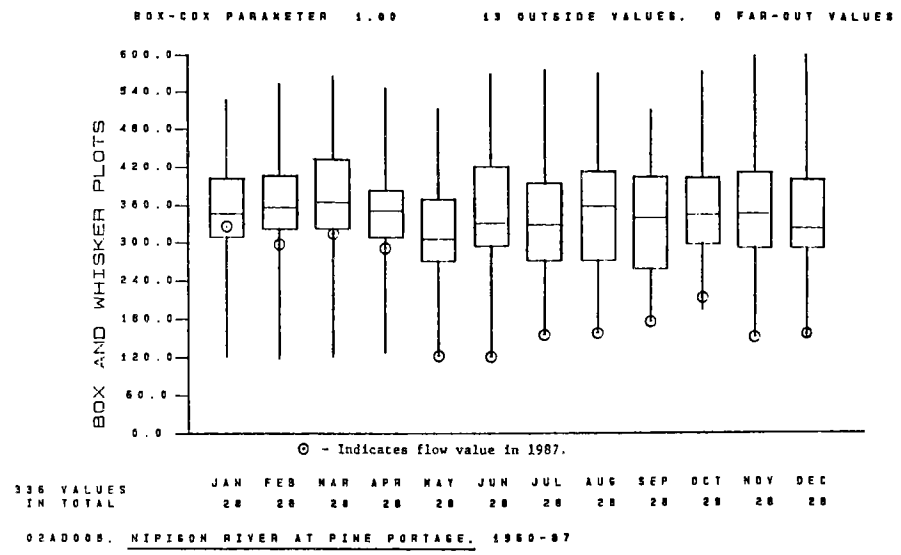

Figure 10: Monthly Box-Whisker Plots for 02AD002 Nipigon River at Pine Portage.

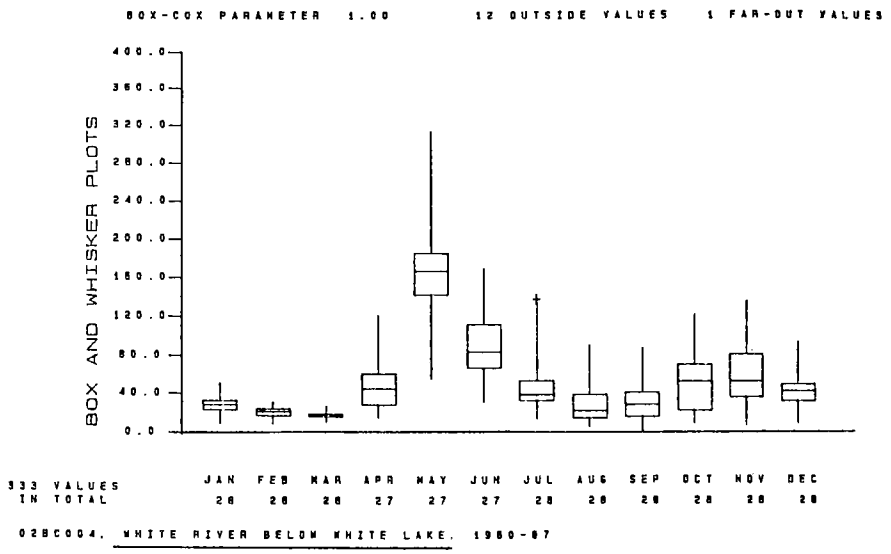

Figure 11: Monthly Box-Whisker Plots for 02BC004 White River below White Lake.

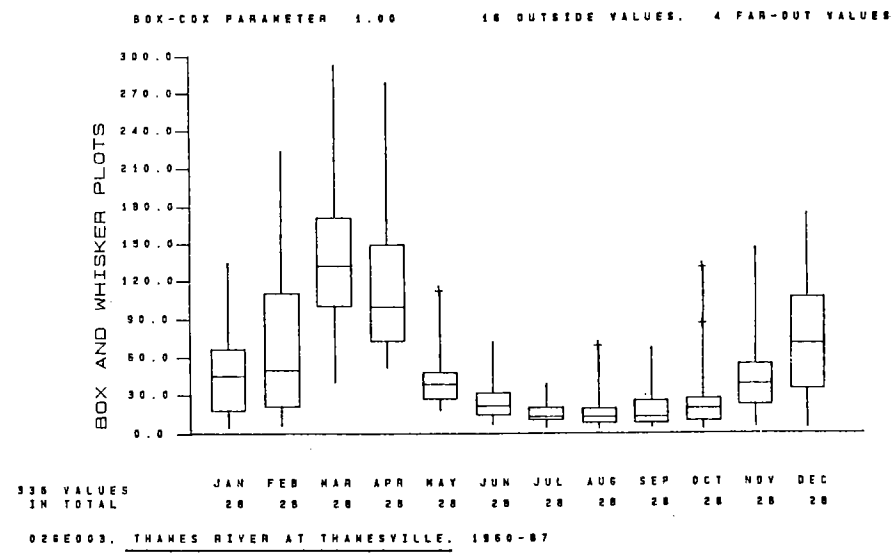

Figure 12: Monthly Box-Whisker Plots for 02GK003 Thames River at Thamesville.

Conclusions

The MH time series package has demonstrated value in providing a comprehensive range of exploratory data analysis tools for analyzing hydrologic time series. The analysis found significant trends in the flows, both monthly and annually for the Grand River at Galt and pointed to data from the Flathead River at Flathead which

requires further investigation. The package also demonstrated ability to display and measure normality and correlation structure. This information is of use in planning gauging strategy and in providing more complete data products to data users.

The Box and Whisker plots allow assessment of sequences of monthly flows and clearly demonstrate differences between streams and indicate clearly an extreme event such as the occurrence of a series of low flows. These plots make behaviour of streams readily visible and should be made a part of monthly hydrology reports.

References

Bartlett, M.S. 1946. On the Theoretical Specification of Sampling Properties of Autocorrelated Time Series. *Journal of the Royal Statistical Society*, Series B, 8: 27-41.

Bendat, J.S. and A.G. Piersol. 1971. *Random Data: Analysis and Measurement Procedures*. John Wiley & Sons, Ltd.

Box, G.E.P. and D.R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society*, Series B, 26: 211-252.

Cleveland, W.S. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74 (368): 829-836.

D'Agostino, R.B. 1970. Transformation to Normality of the Null Distribution of g. *Biometrika*, 57: 679-681.

Fisher, R.A. 1970. *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, Scotland.

Hirsch, R.M., Slack, J.R. and R.A. Smith. 1982. Techniques for Trend Assessment for Monthly Water Quality Data. *Water Resources Research*. 18 (1): 107-121.

Jenkins, G.M. and D.G. Watts. 1968. *Spectral Analysis and its Application*. Holden-Day San Francisco.

Loney, S.W. and T.R. Gulledge Jr. Use of the Correlation Coefficient with Normal Probability Plots. *The American Statictician*.

Mann, H.B. 1945. Nonparametric Tests Against Trend. *Econometrics*. 13: 245-259.

Shapiro, S.S. and M.B. Wilk. 1965. An Analysis of Variance Test for Normality. *Biometrica*, 52: 591-661.

Tukey, J. 1977. *Exploratory Data Analysis*. Addison-Wesley. Reading, Massachusetts.

Table 1 and 3 are not directly referenced in the text; information provided in the tables is used throughout the manuscript.