WATER RESOURCES BULLETIN
VOL. 19, NO. 4 AMERICAN WATER RESOURCES ASSOCIATION AUGUST 1983

Received the 1984 AWRA Boggess Award for best paper of 1983 in *Water Resources Bulletin*.

TREND ASSESSMENT OF WATER QUALITY TIME SERIES[1]

A. Ian McLeod, Keith W. Hipel, and Fernando Camacho[2]

ABSTRACT: A general methodology is described for identifying and statistically modeling trends which may be contained in a water quality time series. A range of useful exploratory data analysis tools are suggested for discovering important patterns and statistical characteristics of the data such as trends caused by external interventions. To estimate the entries in an evenly spaced time series when data are available at irregular time intervals, a new procedure based upon seasonal adjustment is described. Intervention analysis is employed at the confirmatory data analysis stage to rigorously model changes in the mean levels of a series which are identified using exploratory data analysis techniques. Furthermore, intervention analysis can be utilized for estimating missing observations when they are not too numerous. The effects of cutting down a forest upon various water quality variables and also the consequences of acid rain upon the alkalinity in a stream provide illustrative applications which demonstrate the effectiveness of the methodology.
(KEY TERMS: confirmatory data analysis; data filling; exploratory data analysis; intervention analysis; seasonal adjustment; statistics; stochastic modeling; water quality.)

INTRODUCTION

The main purpose of this research is to present a comprehensive methodology to identify and, if possible, statistically model any trends which may be present in a water quality time series. These trends, if any, may be due to the presence of known or unknown interventions such as various types of land-use changes. In addition to possibly being affected by external interventions, usually a given water quality variable is measured at irregular time intervals, and often there are large time gaps at which no data are collected. Therefore, a systematic procedure is developed to optimize the amount of meaningful statistical information which can be gleaned from the currently available data.

As explained by Tukey (1977), there are usually two major steps in a statistical study. The first step is called "exploratory data analysis" and the objective of this phase of the work is to uncover important properties of the data by executing simple graphical and numerical studies. Some of the techniques available for this phase include a graph of the data against time, the five-number summary graph which Tukey (1977, ch. 2) calls the box-and-whisker plot, Tukey smoothing (Tukey, 1977, ch. 7) and the autocorrelation function. The purpose of the next step which is referred to as a "confirmatory data anslysis" is to statistically confirm the presence or absence of certain properties in the data. For example, when sufficient measurements have been taken for a water quality variable, exploratory data analysis may indicate that there is a possible trend in the data due to a known external intervention. Following this, intervention analysis (Box and Tiao, 1975) can be utilized as a confirmatory data analysis tool to determine if there has been a significant change in the mean level of the series.

Many exploratory techniques and confirmatory methods require that equally spaced data be available, and as was pointed out before, environmental series are usually measured at uneven time intervals. Accordingly, in the next section a methodology based on seasonal adjustment is devised for estimating the entries of an average monthly time series when daily values are available at irregular time intervals and often there are time gaps spanning many months for which no measurements were taken. In addition to estimating values for a monthly sequence, this procedure can of course be used for estimating averages at other even time intervals such as weekly or quarterly data by having 52 and 4 seasons per year, respectively.

Following the data filling section, specific exploratory and confirmatory data analysis techniques are described. In order to clearly demonstrate the efficacy of the foregoing techniques, practical applications are presented throughout the paper. Possible trends in water quality and river flow series are examined for two locations in Canada. In the Province of Alberta, both exploratory and confirmatory data analysis techniques are employed to ascertain the effects of cutting down a forest upon total organic carbon and turbidity in the Cabin Creek near Seebe. On the Mill River near St. Anthony in Prince Edward Island, exploratory data analysis results suggest that perhaps due to acid rain, alkalinity levels may be decreasing over time. These illustrative applications are in fact part of an extensive environmental study executed by the authors where 50 environmental time series were exhaustively

[1] Paper No. 83035 of the *Water Resources Bulletin*.
[2] Respectively, Assistant Professor, Dept. of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada N6A 5B9; Associate Professor, Dept. of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; and Ph.D. Student, Dept. of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada N6A 5B9.

analyzed (see Acknowledgments). For exact definitions of the various water quality variables, the reader may wish to refer to the book of McNeely, *et al.* (1979).

DATA FILLING USING SEASONAL ADJUSTMENT

Many exploratory data analysis methods are valid for use with either unequally or evenly spaced data. Tukey smoothing, which is explained in the next section, is an example of an exploratory tool where the measurements, or estimates thereof, must be available at equal time intervals before the method should be used. In order to employ any of the stochastic modeling techniques at the confirmatory data analysis stage, it is absolutely necessary that an evenly spaced data sequence be available. Therefore, when data are unevenly spaced procedures are required for creating an evenly spaced sequence which stochastically represents what could have occurred historically. Baracos, *et al.* (1981), explain how intervention analysis can be employed for estimating missing values from an evenly spaced data set when the number of unknown observations are not too large (usually not more than 5 percent of the data set). However, for unevenly spaced daily observations with a large number of missing values a different procedure must be adopted for estimating a sequence of evenly spaced average monthly values. The particular technique presented in this section is related to methods developed for seasonal adjustment models.

In seasonal adjustment modeling, a time series is decomposed into various components, one of which is the seasonal term. Various seasonal adjustment procedures are available and the reader may wish to refer to the statistical literature for a description of these techniques (e.g., Shiskin, *et al.*, 1976; Granger, 1980; Kendall, 1973). Suppose that z_t represents an observation at time t either for the original time series or for some Box-Cox transformation of the given data. The Box-Cox transformation for the observation, Z_t, measured at time t is defined as (Box and Cox, 1964)

$$z_t = \begin{cases} [(Z_t+c)^{\lambda} - 1] & \lambda \neq 0 \\ \ell n\,(Z_t+c) & \lambda = 0 \end{cases} \qquad (1)$$

where c is a constant assigned so that $Z_t+c > 0$ for all t and $z_t = Z_t$ when $\lambda = 1$ and $c = 0$. One reason for invoking a Box-Cox transformation is to cause data that are not normally distributed to approximately follow a normal distribution. For instance, a logarithmic transformation may reduce the skewness and improve the symmetry of the distribution if there are quite a few large values in the series. When the variance of a series depends on the "level" of the series this transformation may rectify the problem. Furthermore, a Box-Cox transformation can often alleviate problems with the properties of the residuals of the stochastic model fitted to the series of equally spaced data (Hipel, *et al.*, 1977a; McLeod, *et al.*, 1977).

An additive seasonal adjustment model can be written as:

$$a_t = C_t + S_t + I_t = C_r + S_m + I_t$$

where t is the Julian day number (i.e., the number of days since January 1, 4713 B.C.), r is the year, m is the month for monthly data, C_t or C_r is the trend factor for modeling relatively long term causes, S_t or S_m is a stable seasonal factor which is assumed not to evolve with time, I_t is the nonseasonal irregular component made up to short-run effects and is not necessarily white noise. The seasonal adjustment algorithm consists of the following steps:

1. Obtain preliminary estimates of C_t, S_t, and I_t. $\tilde{C}_t = \tilde{C}$ is taken to be a constant which is equal to the median of z_t. To get $\tilde{S}_t$ first calculate $\tilde{S}'_m$ as the median of $z_t - \tilde{C}$ for the data in the mth month. Then use $\tilde{S}_m = \tilde{S}'_m - \frac{1}{12}\sum_{m=1}^{12} \tilde{S}'_m$. Estimate the irregular component utilizing

$$\tilde{I}_t = z_t - \tilde{C} - \tilde{S}_m .$$

2. Replace far-out values in the $\tilde{I}_t$ series by the nearest outer fence (see upcoming section on box-and-whisker graphs for definitions of far-out values and outer fences) to form the irregular series $\tilde{I}'_t$. The process of replacing far-out values by outer fences is called "Winsorizing" (Tukey, 1977).

3. Estimate the deseasonalized series given by

$$D_t = \tilde{C} + \tilde{I}'_t .$$

4. Determine the revised trend estimate, $\tilde{\tilde{C}}_t$, where each year in $\tilde{\tilde{C}}_t$ is the mean of D_t for that year. If no data are available for the rth year, the mean of D_t for surrounding years is used.

5. Calculate the revised seasonal component

$$\tilde{\tilde{S}}_m = \tilde{\tilde{S}}'_m - \frac{1}{12}\sum_{m=1}^{12} \tilde{\tilde{S}}'_m$$

where $\tilde{\tilde{S}}'_m$ is the median of $z_t - \tilde{\tilde{C}}_t$.

6. The revised irregular series is estimated using

$$\tilde{\tilde{I}}_t = z_t - \tilde{\tilde{S}}_m - \tilde{\tilde{C}}_r$$

7. Winsorize the revised irregular series, $\tilde{\tilde{I}}_t$, to obtain the Winsorized series, $\tilde{\tilde{I}}'_t$. This is accomplished by replacing the far-out values of $\tilde{\tilde{I}}_t$ by the appropriate outer fences.

8. Obtain an adjusted version (i.e., Winsorized) of the z_t series using

$$z'_t = \tilde{\tilde{C}}_r + \tilde{\tilde{S}}_m + \tilde{\tilde{I}}'_t$$

For a given month for a specified year in which data were originally given, take the median of the z'_t values to get the estimated average monthly value.

9. Adjust the trend for each year by employing $\tilde{\tilde{C}}_r = \tilde{\tilde{C}}_r +$ mean of $\tilde{\tilde{I}}'_t$ for the whole series.

10. To obtain an estimated monthly average value for a given month in which no data were given use

$$\tilde{z}_{r,m} = \tilde{\tilde{C}}_r + \tilde{\tilde{S}}_m$$

where $\tilde{z}_{r,m}$ is the estimated monthly value for the rth year and mth month. The total estimated monthly series is formed by using Steps 8 and 10. Note that if a Box-Cox transformation is taken of the given data, then an inverse Box-Cox transformation must be invoked to obtain the estimated monthly averages for the original untransformed series.

In order to demonstrate how well the seasonal adjustment algorithm works consider the flows in m^3/s of the Cabin Creek near Seebe in Alberta, Canada, from January 1964 till December 1979. A daily flow value has been measured for each day during this time period and for each month in a given year an average monthly value can be readily calculated. Because river flow measurements are often highly skewed, it is advantageous to take natural logarithms of the data. In Figure 1 the natural logarithms of the actual average monthly values are marked with black circles for one particular four-year interval. For exactly the same days on which observations are missing for the turbidity data in the Cabin Creek, the corresponding daily observations are removed from the flow data. Following this, the seasonal adjustment algorithm is employed to estimate the average monthly flows of the logarithmic daily data for the period from 1964 to 1979 and these estimated flows are marked by circles in Figure 1. It should be pointed out that for the turbidity series and hence the estimated flows, only about 8 percent of the data are used in the seasonal adjustment algorithm. In addition, there are many months during which no observations are available. However, as can be seen in Figure 1, the estimated values from the seasonal adjustment algorithm are reasonably close to the actual entries during the four-year period and also the other years not shown in Figure 1.

The seasonal adjustment algorithm allows irregularly spaced observations to be transformed to evenly spaced estimates, so that environmentalists, hydrologists, and other scientists can use all the available statistical tools. In the future, it would be advantageous to design proper sampling programs so that the power of the various statistical methods can be fully realized. As shown in the literature, scientists are cognizant of the importance of sampling for specific types of problems (see for instance Arnold, 1970; Box 1974; Hunter, 1981; and Smeach and Jernigan, 1977). Based upon the properties of the intervention model, Lettenmaier, *et al.* (1978), have suggested specific sampling rules when checking for trends in the data.

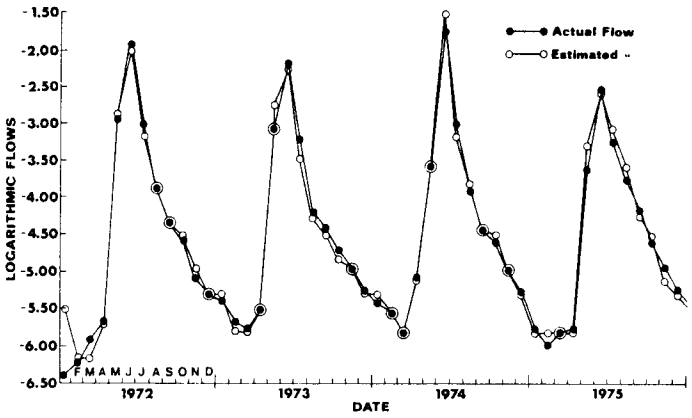

Figure 1. Monthly Logarithmic Flows of the Cabin Creek (m^3/s).

EXPLORATORY DATA ANALYSIS

Various instructive exploratory data analysis tools are available for revealing interesting properties of the data under consideration. Each exploratory technique possesses its own inherent attributes that are useful for uncovering certain data characteristics. Because no single method can clearly portray everything there is to learn about the data, it is advantageous to examine the time series by employing a number of useful investigative graphical and numerical tools. In particular, the techniques utilized in this section include a graph of the data against time, the five-number summary graph which Tukey (1977, ch. 2) calls the box-and-whisker plot, Tukey smoothing (Tukey, 1977, ch. 7), and the autocorrelation function (ACF). In order to employ Tukey smoothing and the ACF, data must be available at equally spaced time intervals and consequently the seasonal adjustment algorithm from the previous section can be used to accomplish this. The plotting of smoothed curves and also the calculation of the ACF at lag one of the average annual time series, constitute valuable methods for detecting possible interventions.

Time Series Plots

One of the simplest and more useful exploratory graphical tools is to plot the data against time. Characteristics of the data which may be easily discovered from a perusal of a graph include the detection of extreme values, trends, known and unknown interventions, dependencies between observations, seasonality, need for a data transformation, nonstationarity, and long term cycles (Hipel and McLeod, 1983; Berthouex, *et al.*, 1981).

When considering unequally spaced daily data, the actual time intervals between adjacent observations must be calculated before plotting the observations against time. A convenient technique to employ is to determine the Julian day number for each observation using the formula given by Hewlett-Packard (1977). With this information the gap between adjacent observations can be determined as the difference of the Julian day numbers of the observations. This procedure is employed to obtain the graph in Figure 2 of the

natural logarithms of the turbidity in the Cabin Creek. As shown by the time gaps between observations, there are many days and even months during which no measurements were taken. For instance, from August 2 to November 22, 1975, inclusive, no observations were recorded.

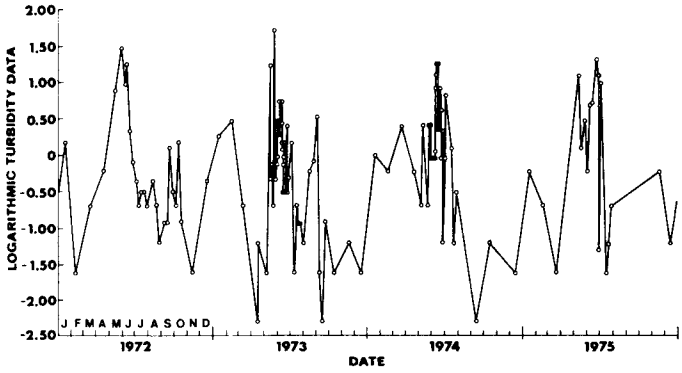

Figure 2. Natural Logarithms of the Turbidity (mg/ℓ) Data for the Cabin Creek.

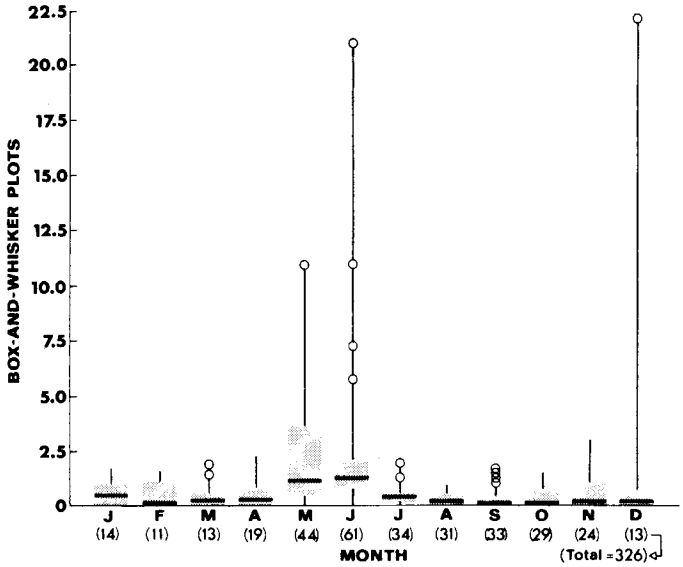

Figure 3. Box-and-Whisker Plots for Turbidity (mg/ℓ) in the Cabin Creek Before July 1, 1974, Where There is No Data Transformation.

Box and Whisker Graphs

The box-and-whisker graph is based upon what is called the five-number summary (Tukey, 1977, ch. 2). For a given data set, the five-number summary consists of the smallest and largest values, the median and the two extreme quartiles, which are called "hinges."

To assist in characterizing extreme values, Tukey (1977) has suggested the following definitions. Let "H spread" be the difference between the two hinges, and a "step" 1.5 times the H-spread. "Inner fences" are one step outside hinges and "outer fences" are two steps outside hinges. Values between an inner fence and its neighboring outer fence are called "outside." Values beyond outer fences are "far-out." When entertaining seasonal data such as monthly or quarterly data, it is instructive to calculate a five-number summary plus outside and far-out values for each season. A convenient manner in which to display this information is to plot "box-and-whisker" diagrams for each season or month. Figure 3 depicts the box-and whisker plots for turbidity in the Cabin Creek before July 1, 1974, when part of the forest was cut down. In this figure the data have not been transformed using a Box-Cox transformation. The upper and lower ends of a rectangle for a given month represent the two hinges and the thick line drawn horizontally within each rectangle is the value of the median. The minimum and maximum values in a particular month are the end points of the lines or "whiskers" attached to the rectangle or "box." The far-out values are indicated by a circle in Figure 3. Below each month is a number which gives the number of data points used to calculate the box-and-whisker graph above the month. The total number of observations across all the months is listed below November and December.

In addition to detecting far-out values, box-and-whisker diagrams can be used to discover the lack of symmetry in the distribution of the data for a given season. If the data are approximately symmetrical with respect to the median, they may follow a symmetric distribution such as the normal distribution. For a given month in a box-and-whisker diagram, symmetric data would cause the median to lie in the middle of the rectangle and the lengths of the upper and lower whiskers would be about the same. Notice in Figure 3 for the turbidity data that the whiskers are almost entirely above the rectangle for all of the months and for six of the months there are a total of 14 far-out values. This lack of symmetry can at least be partially rectified by transforming the given data using the Box-Cox transformation in Equation (1). By comparing Figure 3 to Figure 4 where natural logarithms are taken of the turbidity data, the improvement in symmetry can be clearly seen. Furthermore, the Box-Cox transformation has reduced the number of far-out entries from 14 in Figure 3 to 3 in Figure 4.

Box-and-whisker plots can be employed as an important exploratory tool in intervention studies. If the date of the intervention is known, box-and-whisker diagrams can be constructed for each season for the data before and after the time of intervention. These two graphs can be compared to ascertain for which seasons the intervention has caused noticeable changes. When there is sufficient data, this type of information is crucial for designing a proper intervention model to fit the data at the confirmatory data analysis stage.

The Cabin Creek basin which has an area of 2.12 km^2 was originally forested but from July to October 1974, 40 percent of the forested area was clear-cut. Total organic carbon readings are available from March 17, 1971, to January 10, 1979. Figures 5 and 6 display the box-and-whisker plots of the

natural logarithms of the total organic carbon in mg/ℓ for the Cabin Creek before and after the intervention, respectively, caused by the removal of the trees. As can be observed, there are obvious drops in the medians for almost all the months after the intervention. These and other changes cannot be as easily detected in a plot of the entire series against time.

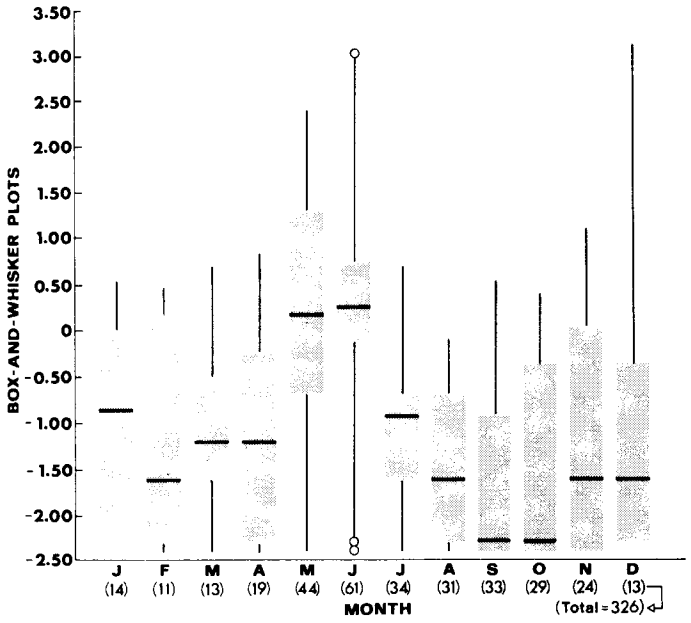

Figure 4. Box-and-Whisker Plots for Turbidity (mg/ℓ) in the Cabin Creek Before July 1, 1974, Where There is a Logarithmic Data Transformation.

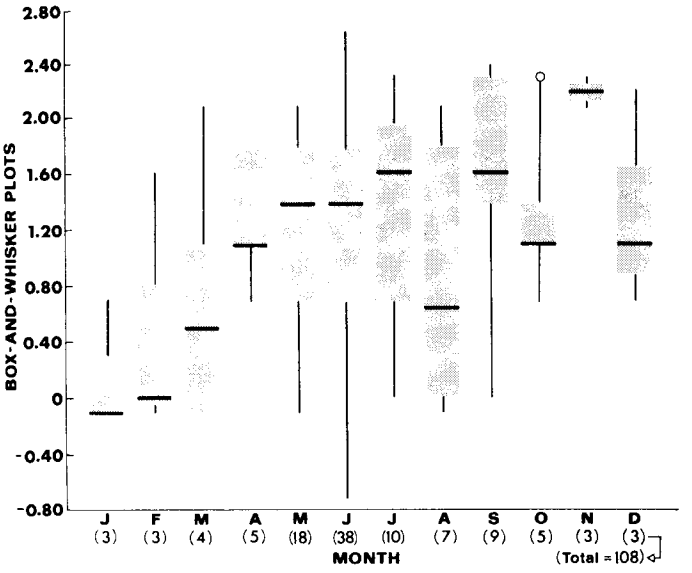

Figure 5. Box-and-Whisker Plots of the Logarithmic Total Organic Carbon (mg/ℓ) in the Cabin Creek Before July 1, 1974.

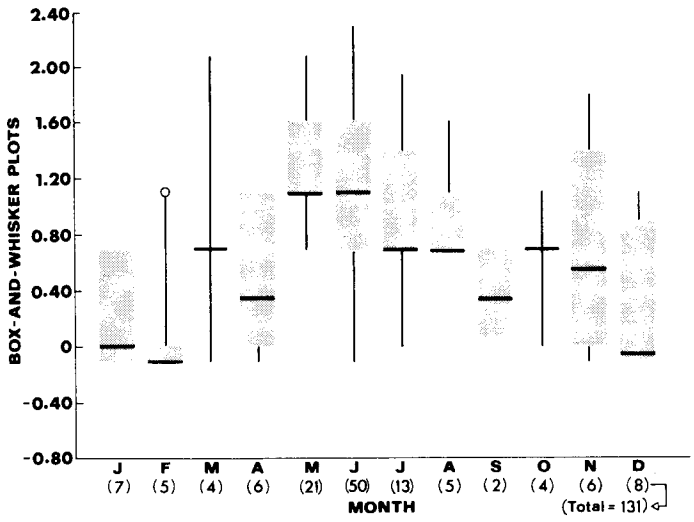

Figure 6. Box-and-Whisker Plots of the Logarithmic Total Organic Carbon (mg/ℓ) in the Cabin Creek After October 31, 1974.

Tukey Smoothing

Sometimes a graph of a given time series "blurs" statistical information in the data which a smoothed plot of the series at equally spaced time intervals may reveal more clearly. Consider, for example, Figure 7 which is a plot of the average annual total organic carbon in mg/ℓ, for the Cabin Creek where the average annual entries are calculated using the estimated monthly values obtained from the seasonal adjustment algorithm developed earlier. From this graph there appears to be a drop in the mean level of the series in the later years compared with the values in the early 1970's. When the "blurred smooth" in Figure 8 is studied the general characteristics of the data are more clearly portrayed. Figure 8 is a blurred smoothed plot of the average annual total organic carbon for the Cabin Creek where the vertical lines reflect the magnitude of the rough, or "blur" of the series and a "smoothed" observation is located at the mid-point of the bar. Notice from Figure 8 that the smoothing characteristics for the data before 1974 are more or less the same but from 1974 onwards there is an obvious decrease in the mean of the series. This property was also suggested by the box-and-whisker plots of the series shown before and after the intervention in Figures 5 and 6, respectively.

Although a smoothed graph does not contain any more information than what is already present in the plot of the raw data, in many instances the smoothed graph portrays the essential features much more clearly. The purpose of a smoothed curve is to reveal the systematic structure and interesting statistical characteristics of the data. Consider, for example, the blurred smoothed graph in Figure 9 for the total alkalinity in mg/ℓ for the Mill River at St. Anthony in Prince Edward Island, Canada. This graph is a blurred smoothed plot of the average annual values which were calculated from the estimated monthly entries obtained from the seasonal adjustment algorithm. In Figure 9, there is an obvious shift downwards in

alkalinity from 1973 to 1977 followed by abrupt decreases in 1978 and 1979. Because the soil in the Mill River basin is sandy, acid rain could quickly drain through the ground without undergoing substantial chemical changes and thereby adversely affect the water quality. Consequently, the decrease in alkalinity in Figure 9 could be mainly due to acid rain which could severely affect the biological life in the river. However, it is still necessary to collect more data and determine when the acid rain intervention came into effect before proper confirmatory data analyses can be executed.

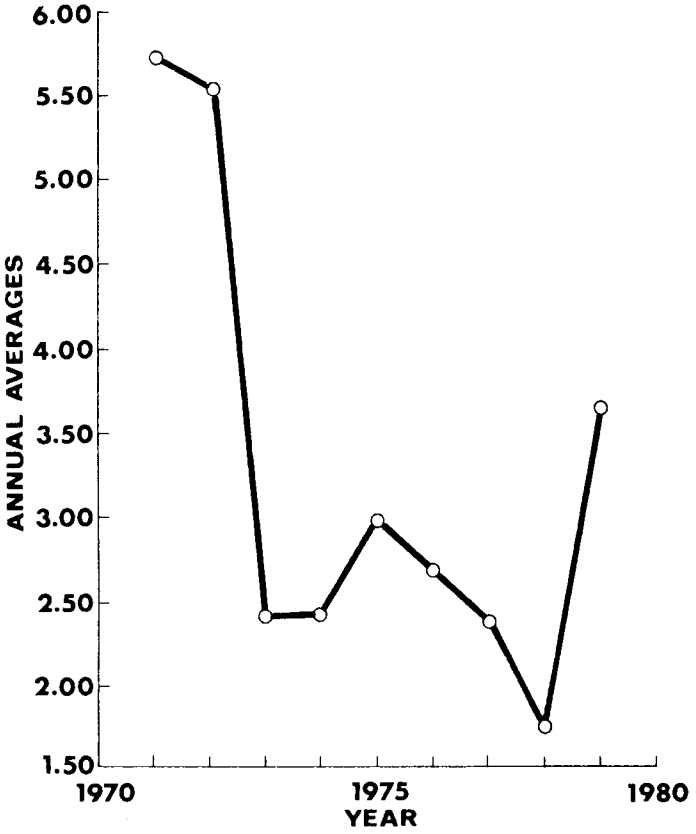

Figure 7. Estimated Annual Values of the Total Organic Carbon (mg/ℓ) in the Cabin Creek.

To construct a smoothed curve consider qualitatively subdividing a given time series as

Data = Smooth + Rough

By filtering out the rough or noise portion of the data, the smoothed curve can be examined for important statistical features. The filter which maps the given series into a smoothed curve is referred to as a smoother.

The nonlinear smoothers developed by Tukey (1977, ch. 7) and also discussed by McNeil (1977), are very flexible when used in practical applications and are capable of detecting all of the items discussed for a plot of the series except, possibly, for occasional outliers. Mallows (1980) explains the desirable properties that any smoother should possess and also presents some theoretical mathematical results for Tukey smoothers. Some of the more important attributes that a smoother should have include the ability to be responsive to abrupt changes in level, marginal distribution, and covariance structure.

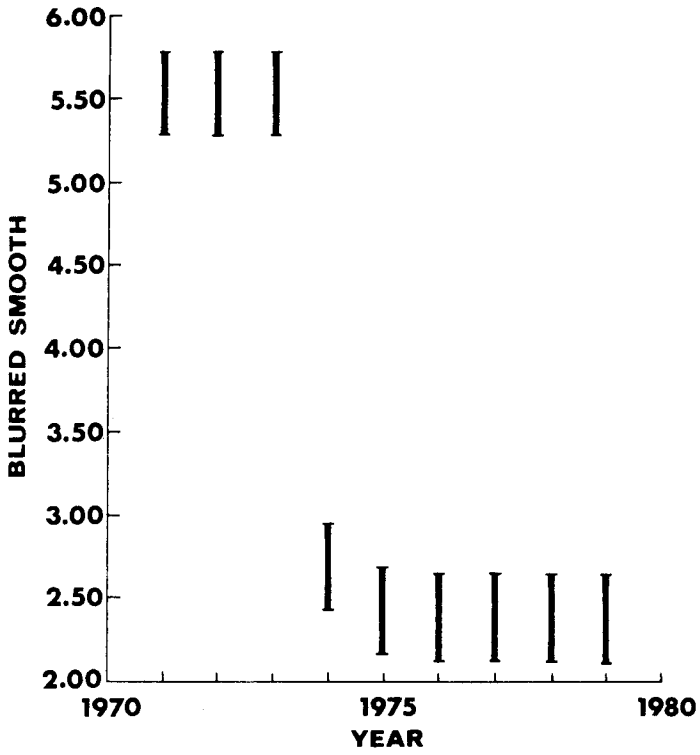

Figure 8. Blurred Smooth of the Estimated Average Annual Total Organic Carbon (mg/ℓ) in the Cabin Creek.

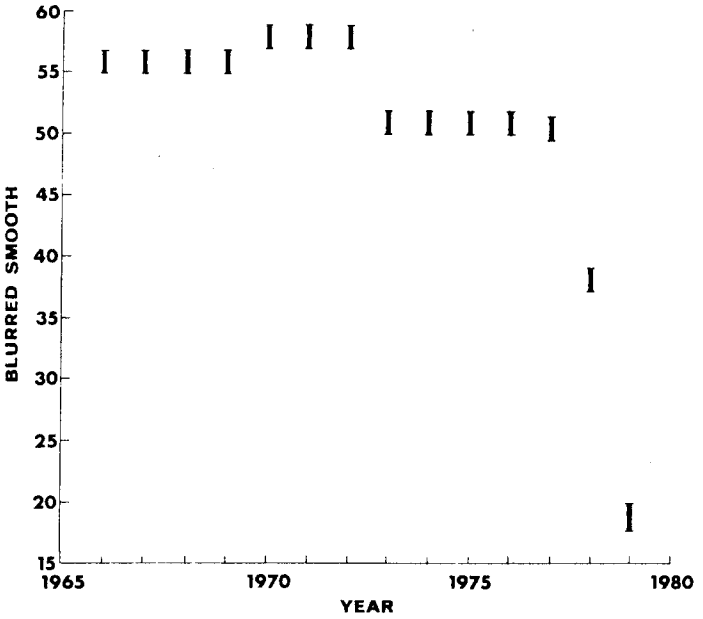

Figure 9. Blurred Smooth of the Estimated Average Annual Total Alkalinity (mg/ℓ) in the Mill River.

Figures 8 and 9 are examples of what Tukey (1977, ch. 7) calls a blurred "3RSR" smooth. Although Tukey defines various types of nonlinear smoothers, the blurred 3RSR version is particularly well designed for use in exploratory data analysis with environmental time series. For a detailed explanation of how to calculate a blurred 3RSR smooth the reader can refer to Tukey (1977, ch. 7). A computer program for a 3RSR smooth is given by McNeil (1977, ch. 6).

Autocorrelation Function

The ACF at lag k for a given time series reflects the linear dependence between values which are separated by k time lags. The estimate for the ACF at lag k for an evenly spaced series, z_t, of length N, can be calculated using (Jenkins and Watts, 1968)

$$r_k = \frac{\sum_{t=1}^{N-k} (z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{N} (z_t - \bar{z})^2}, \quad k > 0 \qquad (2)$$

where $\bar{z}$ is the estimated mean of the z_t series. The value of r_k can range from −1 to +1 where r_0 always has a value of unity. Because the ACF is symmetrical about lag zero, it is only plotted for positive lags. When the theoretical ACF is zero and therefore the series is white noise, r_k is asymptotically normally independently distributed with a mean of zero and variance of 1/N. Using simulation experiments, Cox (1966) demonstrated that when r_1 is calculated for a sequence of uncorrelated samples the sampling distribution of r_1 is very stable under changes of distribution and the asymptotic normal form of the sampling distribution is a reasonable approximation even in samples as small as ten.

The ACF furnishes a method for interpreting trends in the data. If, for example, there is a large positive correlation at lag one, this means that in the plot of a series a sequence of high values will often be grouped together and low values will often follow other low values. In other words, when r_1 and other sample ACF's are significantly different from zero, this indicates the presence of stochastic trends in the data. If, for instance, the significance level is less than 0.05 this means that r_1 is significantly different from zero at the 5 percent significance level. The value of r_1 for the annual total organic carbon series in Figure 7 is 0.371 with a significance level of 0.137. Consequently, r_1 is significantly different from zero. When there is an intervention which causes a significant change in the mean level of a series such as the change shown in Figures 7 and 8 for the total organic carbon, this introduces a trend in the data due to the observations fluctuating about different mean levels at specified sections in the series. This enforced trend should cause a rather large value for r_1 for the entire series which is the case for the total organic carbon series. Likewise, an overall trend in the data can cause r_1 to be large. Experience and theory suggest that the trend test based on r_1 is often more powerful than the usual nonparametric tests such as the runs tests and those tests described in Kendall (1973, ch. 2).

CONFIRMATORY DATA ANALYSIS

The General Intervention Model

When analyzing time series, various types of stochastic models constitute useful confirmatory data analysis tools. An assumption underlying all of the stochastic models which can be employed in practical applications is that the data sets to which they are fitted consist of observations separated by equal time intervals. Although it would be desirable to possess stochastic models which can readily handle data consisting of any kind of unevenly spaced observations, currently no such practical models exist and, indeed, it may turn out to be mathematically intractable to develop these types of stochastic models. In practice, if the measurements are not evenly spaced, appropriate techniques must be utilized to produce a series of equally spaced data that is estimated from the given information. Of course, practitioners are advised to design future sampling programs so that evenly spaced data are collected at suitable time intervals. In this way, the inherent assets of the confirmatory data analysis tools can be fully exploited.

Specific time series models (Box and Jenkins, 1970) have been employed to model water quality data (see, for instance, Fuller and Tsokos, 1971). Intervention analysis constitutes a powerful confirmatory data analysis tool which is extremely useful in environmental impact assessment for analyzing the effects of natural and man induced interventions on the environment (Box and Tiao, 1975). The method was originally suggested for use in hydrology by Hipel, *et al.*, in 1975 and has been successfully applied to a variety of hydrologic and environmental problems. Intervention analysis has been used in hydrology to determine statistically the effects of dam construction on annual (Hipel, *et al.*, 1975) and monthly (Hipel, *et al.*, 1977b) downstream river flows. Hipel, *et al.* (1977c, 1978) used the technique to ascertain the stochastic effects of a forest fire on monthly river flows, and D'Astous and Hipel employed intervention analysis to model the effectiveness of water pollution abatement measures. Baracos, *et al.* (1981), used intervention analysis to determine how changing the type of measuring gauge affects snow measurements. Based upon the theory of the intervention model, Lettenmaier, *et al.* (1978), explained how to design data collection procedures. Intervention analysis has also been employed to estimate missing data points in a time series (Baracos, *et al.*, 1981; D'Astous and Hipel, 1979; Lettenmaier, 1980). Within this paper, intervention analysis is used for the first time in conjunction with the seasonal adjustment algorithm and exploratory data analysis tools, to study complex water quality problems.

A mathematical description of the general intervention model is given by Baracos, *et al.* (1981). In addition to modeling the statistical effects of external interventions, the general

intervention model can be used to estimate missing observations. Furthermore, when covariate series are available it is possible to include them in the general intervention model. For instance, river flows and another water quality variable may be used as inputs in a water quality intervention model.

No matter what type of stochastic model is being fitted to a given data set, it is recommended to follow the identification, estimation, and diagnostic stages of model development (Box and Jenkins, 1970; Hipel and McLeod, 1983). Baracos, *et al.* (1981), Hipel, *et al.* (1977b), and Box and Tiao (1975) suggest specific methods for constructing intervention models. Within this paper, the method of McLeod (1977) is employed for obtaining maximum likelihood estimates (MLE's) of the model parameters although other recommended maximum likelihood procedures include those of Ansley (1979) and Ljung and Box (1979). When parameters are estimated for a number of models, a convenient method for choosing the most appropriate model is to select the model that has the minimum value of the Akaike Information Criterion (AIC) (Akaike, 1974). Within the hydrological literature, the method of employing the AIC in conjunction with the three stages of model development has been clearly explained (Hipel and McLeod, 1983; Hipel, 1981) and the efficacy of the AIC has been confirmed by a wide range of stochastic modeling applications (e.g., McLeod, *et al.* 1977; McLeod and Hipel, 1978a; Hipel and McLeod, 1983; Hipel, 1981). If a suitable range of models is considered, it has been found in practice that the model possessing the minimum AIC value also satisfies diagnostic tests of the model residuals such as those proposed by McLeod (1978) and Hipel, *et al.* (1977a).

Applications

In 1961 the Marmot Creek experimental basin was established on the eastern slopes of the Rocky Mountains in Alberta, Canada (Jeffrey, 1965; Golding, 1980). The objective of the study was to determine the hydrology of the area so guidelines that are consistent with the importance of the eastern slopes as a water supply area for Alberta and Saskatchewan could be formulated for harvesting trees. Both the Middle Fork and Cabin Creeks are located within the Marmot basin in the Province of Alberta. From July to October 1974 an intervention took place in the Cabin Creek basin when 40 percent of the forested area was clear-cut. Because the trees in the forested Middle Fork basin were not cut down, and the basin is located close to the Cabin Creek basin, the appropriate series from the Middle Fork Creek can be used as covariate series for intervention models developed for the Cabin Creek data sets. In this way the intervention components in the intervention models will more accurately measure the effects of the intervention in the Cabin Creek series.

Intervention models were developed for 12 water quality variables on the Cabin Creek although representative results are only shown in this section for the total organic carbon intervention model. For each water quality intervention model, the covariate series are the same water quality series for the Middle Fork basin and also the monthly flows of the Cabin Creek. The general structure of the model which is employed with the series is

$$z_t - \bar{z} = \sum_{i=1}^{12} \omega_{0i}\xi_{ti} + \omega_{013}(x_t - \bar{x}) + \omega_{014}(y_t - \bar{y}) + N_t \quad (3)$$

where z_t is the average monthly water quality series for the Cabin Creek that was estimated using the seasonal adjustment algorithm; $\bar{z}$ is the mean of the z_t series; ξ_{ti} is the intervention series for a given month where it is given a value of one for the month it represents from the intervention onwards and a value of zero elsewhere; ω_{0i} is the transfer function parameter for the ξ_{ti} series and the MLE for ω_{0i} can be used to ascertain the effects of the intervention for the month being studied; x_t is the estimated monthly logarithmic series for the Cabin Creek where the seasonal adjustment algorithm is used to estimate the monthly flows from daily flows that occur at the same time as the water quality observations; $\bar{x}$ is the mean of the x_t series; ω_{013} is the transfer function for the Cabin Creek flow series; y_t is the same estimated monthly water quality series as z_t but for the Middle Fork Creek, the seasonal adjustment algorithm is used to estimate y_t; $\bar{y}$ is the mean of the y_t series; ω_{014} is the transfer function parameter for the covariate Middle Fork water quality series; and N_t is the noise term which can be modeled by an autoregressive integrated moving average (ARIMA) model (Box and Jenkins, 1970) and it contains a white noise series denoted by a_t.

In Equation (3) the seasonally adjusted monthly flows are employed as a covariate series. The reason for using the seasonally adjusted series rather than the known monthly river flows is that this may help to eliminate any problems due to seasonal adjustment that are contained in the z_t series. It should be kept in mind that by considering the flows as a covariate series, the stochastic or statistical relationship between the flow, x_t, and the water quality series, z_t, is formally modeled through the transfer function parameter, ω_{013} in the overall intervention model in Equation (3).

When constructing the water quality intervention models in Equation (3) the identification, estimation, and diagnostic check stages of model development were adhered to. Although the transfer functions for all the water quality series are the same as those in Equation (3), it should be pointed out that quite a few different types of transfer functions were actually tested. For instance, because not too many observations for each month are available after the intervention, a step intervention along with a ω_{0i} parameter is included in the first term for each month on the right hand side in Equation (3). If more data were available the possibility of including a parameter in the denominator of each transfer function would have been feasible. Hipel, *et al.* (1977c, 1978) show how a term in the denominator can model the attenuating effects of a forest fire upon river flows as the forest slowly recovers over the years. Finally, a specific seasonal ARIMA model had to be identified separately for modeling N_t in Equation (3) for each water quality intervention model.

To calculate the percentage change in the mean level of the z_t series for a given month due to the intervention, the following formula is employed when the series is transformed using natural logarithms (see Hipel, *et al.*, 1977b, for a derivation).

$$\text{Percent Change} = (e^{\hat{\omega}_{0i}} - 1)\,100 \tag{4}$$

where $\hat{\omega}_{0i}$ is the MLE of the intervention parameter for the ith month. To calculate the 95 percent confidence limits simply add and subtract 1.96 times the standard error to $\hat{\omega}_{0i}$ and then substitute these two values into Equation (3). If the z_t series is not transformed by a Box-Cox transformation, the percentage change in the mean for the ith month is given by

$$\text{Percent Change} = \frac{\hat{\omega}_{0i}}{\bar{z}_{bi}} \times 100 \tag{5}$$

where $\bar{z}_{bi}$ is the monthly mean for the ith month before the intervention.

Total Organic Carbon Application. As was explained earlier, exploratory data analyses clearly detect the effects of the forest clearing upon the total organic carbon series for the Cabin Creek. For example, when the box-and-whisker graphs for before and after the intervention are compared in Figures 5 and 6, respectively, the decrease in the median level after the intervention can be easily seen for almost all the months. Likewise, the average annual plot in Figure 7 and the blurred smooth in Figure 8 clearly detect the drop in the mean level of total organic carbon in later years.

The foregoing exploratory facts are rigorously confirmed in a statistical sense by fitting the intervention model in Equation (3) to the total organic carbon series (mg/ℓ) which is available from the start of 1971 to the end of 1978. Natural logarithms are used for the two total organic carbon series given by z_t and y_t for the Cabin and Middle Fork Creeks, respectively. The seasonal ARIMA models identified for the noise term, N_t, contains one nonseasonal autoregressive parameter and one seasonal autoregressive parameter. The parameter ω_{013} which relates the Cabin Creek flows to the total organic carbon in the Cabin Creek has a MLE of 0.081 with a standard error of 0.095. Since the MLE of ω_{013} is about the same size as its standard error, it may be worthwhile to include the flows as a covariate series in the intervention model. The MLE for ω_{014} is 0.620 with a standard error of 0.082 and, consequently, it is very informative to incorporate the covariate total organic series from Middle Fork Creek into the model. In Table 1, the MLE's and standard errors are presented for the 12 intervention parameters contained in the first component on the right hand side of Equation (3). Also included in Table 1 is the percentage change in mean level for each month along with the 95 percent confidence limits which are calculated using Equation (4). For all the months where zero is not included in the 95 percent confidence limits, the percentage change in the mean level is confirmed to be significantly different from zero. Accordingly, from Table 1 it can be seen that there is a significant drop in the mean level of total organic carbon in the Cabin Creek during the months of June, July, and August.

CONCLUSIONS

As demonstrated by the practical applications, a comprehensive procedure is now available for identifying and modeling trends in environmental time series caused by known or unknown interventions. Box-and-whiskers graphs, for example (see Figures 3 to 6), are useful at the exploratory data analysis stage for discovering basic statistical characteristics of the data such as the types of trends caused within each season due to an intervention. Even though environmental data are often measured at irregular time intervals and large segments of the measurements may be missing, the new seasonal adjustment algorithm can often be employed to obtain reasonable estimates for equally spaced data. In this way, the exploratory data analysis tools such as Tukey Smoothing and the ACF which depend upon evenly spaced data, can be used by practitioners. Furthermore, at the confirmatory data analysis stage, intervention analysis can be utilized to ascertain if there is a significant change in the mean level of a time series due to trends caused by known interventions. When available, covariate series can be incorporated into the intervention model to make it more precise and if there are not too many missing data points they can be estimated by introducing appropriate components into the intervention model.

ACKNOWLEDGMENTS

The authors appreciate the support of the Water Quality Branch of the Inland Waters Directorate at Environment Canada for funding an extensive water quality study during June 1981. Particular individuals who should be thanked include William D. Gummer (Chief of the Water Quality Branch for the Western and Northern Region), Roger McNeely (now head of the Carbon 14 Laboratory of the Geological Survey of Canada), Howard O. B. Block (District Resource Officer of the Water Quality Branch for the Western and Northern Region), and Simon Whitlow (Head of the Data Systems Section of the Water Quality Branch). In the study, a total of 50 time series comprised of 18 different environmental variables were exhaustively examined using exploratory and confirmatory data analysis tools. The methods employed in this study along with some representative results are presented within this paper. The authors are grateful for being able to use the computational facilities at the Statistical Laboratory within the Department of Statistical and Actuarial Sciences, The University of Western Ontario, during the execution of the project.

LITERATURE CITED

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control 19:716-723.

Ansley, C. F., 1979. An Algorithm for the Exact Likelihood of a Mixed Autoregressive-Moving Average Process. Biometrica 66:59-65.

Arnold, J. C., 1970. A Markovian Sampling Policy Applied to Water Quality Monitoring of Streams. Biometrics 26:739-747.

Berthouex, P. M., W. G. Hunter, and L. Pallesen, 1981. Wastewater Treatment: A Review of Statistical Applications. Environmentrics

TABLE 1. Intervention Parameter Estimates for the Total Organic Carbon Model for the Cabin Creek.

Month	Parameter	MLE	Standard Error	Percentage Change of Total Organic Carbon in mg/ℓ	95 Percent Confidence Limits
January	ω_{01}	0.002	0.231	0.17	−36.33, 57.59
February	ω_{02}	0.085	0.231	8.92	−30.71, 71.22
March	ω_{03}	−0.169	0.227	−15.52	−45.86, 31.82
April	ω_{04}	−0.216	0.224	−19.42	−48.11, 25.11
May	ω_{05}	−0.053	0.228	−5.12	−39.25, 48.20
June	ω_{06}	−0.716	0.227	−51.12	−68.69, −23.68
July	ω_{07}	−0.524	0.256	−40.81	−64.18, −2.20
August	ω_{08}	−0.566	0.260	−43.21	−65.88, −5.49
September	ω_{09}	0.029	0.260	2.65	−38.34, 70.90
October	ω_{010}	−0.019	0.258	−1.91	−40.86, 62.69
November	ω_{011}	0.048	0.266	4.90	−37.75, 76.79
December	ω_{012}	−0.344	0.270	−29.13	−58.26, 20.32

81: Selected Papers. Selections from USEPA-SIAM-SIMS Conference, Alexandria, Virginia, pp. 77-99.

Baracos, P. C., K. W. Hipel, and A. I. McLeod, 1981. Modelling Hydrologic Time Series from the Arctic. Water Resources Bulletin 17: 414-422.

Box, G. E. P., 1974. Statistics and the Environment. J. Wash. Acad. Sci. 64:52-59.

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

Box, G. E. P. and G. M. Jenkins, 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, California.

Box, G. E. P. and G. C. Tiao, 1975. Intervention Analysis With Applications to Economic and Environmental Problems. Journal of the American Statistical Association 70(349):70-79.

Cox, D. R., 1966. The Null Distribution of the First Serial Correlation Coefficient. Biometrika 53:923-926.

D'Astous, F. and K. W. Hipel, 1979. Analyzing Environmental Time Series. Journal of the Environmental Engineering Division, American Society of Civil Engineers 105:979-992.

Fuller, C. F., Jr. and C. P. Tsokos, 1971. Time Series Analysis of Water Pollution. Biometrics 27:1017-1034.

Golding, D. L., 1980. Calibration Methods for Detecting Changes in Streamflow Quantity and Regime. *In:* The Influence of Man on the Hydrological Regime with Special Reference to Representative and Experimental Basins. Proceedings of the Helsinki Symposium, IAHS-AISH 130:3-7.

Granger, C. W. J., 1980. Forecasting in Business and Economics. Academic Press, New York, New York.

Hewlett-Packard, 1977. HP-29C Applications Book.

Hipel, K. W., 1981. Geophysical Model Discrimination Using the Akaike Information Criterion. IEEE Transactions on Automatic Control 26:358-378.

Hipel, K. W., W. C. Lennox, T. E. Unny, and A. I. McLeod, 1975. Intervention Analysis in Water Resources. Water Resources Research 11: 855-861.

Hipel, K. W., D. P. Lettenmaier, and A. I. McLeod, 1978. Assessment of Environmental Impacts, Part One: Intervention Analysis. Environmental Management 2:529-535.

Hipel, K. W. and A. I. McLeod, 1983. Time Series Modelling for Water Resources and Environmental Engineers. Elsevier, Amsterdam (in press).

Hipel, K. W., A. I. McLeod, and W. C. Lennox, 1977a. Advances in Box-Jenkins Modelling, 1, Model Construction. Water Resources Research 13:567-575.

Hipel, K. W., A. I. McLeod, and E. A. McBean, 1977b. Stochastic Modelling of the Effects of Reservoir Operation. Journal of Hydrology 32:97-113.

Hipel, K. W., A. I. McLeod, T. E. Unny, and W. C. Lennox, 1977c. Intervention Analysis to Test for Changes in the Mean Level of a Stochastic Process. *In:* Stochastic Processes in Water Resources Engineering, L. Gottschalk, G. Lindh, and L. Mare (Editors). Water Resources Publications, Fort Collins, Colorado, 1:93-113.

Hunter, J. S., 1981. Environmetrics: Mathematics and Statistics in Service of the Environment. Environmetrics 81: Selected Papers. Selections from USEPA-SIAM-SIMS Conference, Alexandria, Virginia, pp. 3-10.

Jeffrey, W. W., 1965. Experimental Watersheds in the Rocky Mountains, Alberta, Canada. *In:* Symposium on Budapest. Proceedings of the Symposium on Representative and Experimental Areas, IAHS 66:502-521.

Jenkins, G. M. and D. G. Watts, 1968. Spectral Analysis and Its Applications. Holden-Day, San Francisco, California.

Kendall, M. G., 1973. Time-Series. Hafner Press, New York, New York.

Lettenmaier, D. P., 1980. Intervention Analysis with Missing Data. Water Resources Research 16:159-171.

Lettenmaier, D. P., K. W. Hipel, and A. I. McLeod, 1978. Assessment of Environmental Impacts, Part Two: Data Collection. Environmental Management 2:537-554.

Ljung, G. M. and G. E. P. Box, 1979. The Likelihood Function of Stationary Autoregressive-Moving Average Models. Biometrika 66:265-270.

Mallows, C. L., 1980. Resistant Smoothing. *In:* Time Series. Proceedings of the International Conference held at Nottingham University, March 1979, O. D. Anderson (Editor). North-Holland, pp. 147-155.

McLeod, A. I., 1977. Improved Box-Jenkins Estimators. Biometrika 64:531-534.

McLeod, A. I., 1978. On the Distribution of Residual Autocorrelations in Box-Jenkins Models. Journal of the Royal Statistical Society, Series B, 40:296-302.

McLeod, A. I. and K. W. Hipel, 1978a. Preservation of the Rescaled Adjusted Range, 1, A Reassessment of the Hurst Phenomenon. Water Resources Research 14:491-508.

McLeod, A. I., K. W. Hipel, and W. C. Lennox, 1977. Advances in Box-Jenkins Modelling, 2, Applications. Water Resources Research 13: 577-586.

McNeely, R. N., V. P. Neimanis, and L. Dwyer, 1979. Water Quality Source-book – A Guide to Water Quality Parameters. Environment Canada, Inland Waters Directorate, Water Quality Branch, Ottawa, Canada.

McNeil, D. R., 1977. Interactive Data Analysis, Wiley, New York, New York.

Shiskin, J., A. H. Young, and J. C. Musgrave, 1976. The X-11 Variant of the Census Method II Seasonal Adjustment Program. Report No. BEA-76-01, Bureau of Economic Analysis, U.S. Department of Commerce, Washington, D.C.

Smeach, S. C. and R. W. Jernigan, 1977. Further Aspects of a Markovian Sampling Policy for Water Quality Monitoring. Biometrics 33: 41-46.

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