WATER RESOURCES DATA QUALITY ASSESSMENT AND DESCRIPTION OF NATURAL PROCESSES USING ARTIFICIAL INTELLIGENCE TECHNIQUES

by

NICOLAS LAUZON
B.Eng., École Polytechnique de Montréal, 1993
M.A.Sc., École Polytechnique de Montréal, 1995

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
(Department of Civil Engineering)
We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
November 2003
© Nicolas Lauzon, 2003

Abstract

The assessment of the quality of any data is difficult to perform if only because of the subjective nature of this task, where quality may be interpreted differently from one scientific domain to another, viewed differently in various cultures and societies, and considered as a more or less chronic problem depending on the context of application. Data, whether employed directly or as inputs for any data analysis or modeling efforts, are at the base of any decision-making process, and a characterisation of their quality is essential in determining bias in any decision on which they are based. This thesis focuses on the assessment of the quality of data regularly employed in water resources engineering and management, in particular hydrometric data and modeling parameters. New approaches are proposed for the detection of three types of anomalies, outliers, shifts and trends, which are a persistent concern to engineers and managers alike, and have been the focus of much research directed at reducing bias in the estimation of water quantity and quality. Artificial intelligence techniques (AITs) constitute the foundations of these new approaches, which are designed to take advantage of the capacity of AITs to provide representative descriptions of data domains. Based on theoretical experiments of their performance relative to conventional statistical diagnostics, and on applications to real hydrometric data from representative watersheds in Canada, the AIT-based approaches may indeed be used to confirm the results from conventional approaches as well as complement and, in some cases, enhance them. Since the ultimate use of hydrometric data is likely as inputs to hydrologic, hydraulic or water quality models, applications of AITs in the simulation of natural processes are also explored in this work. These applications focus on inflow and algae concentration modeling, and demonstrate that improvements in modeling estimations can be gained from the description of natural processes with AIT. Throughout this thesis, discussions regarding the advantages and disadvantages of these AIT-based approaches are provided along with suggestions for future developments.

Table of Contents

List of Tables

List of Figures

Preface

This thesis has led to the production of the following publications:

- Lauzon, N., and Lence, B. J., 20??. Hybrid fuzzy-mechanistic models for added parameter flexibility. Submission to Environmental Modelling and Software expected in April 2004.
- Lauzon, N., and Lence, B. J., 20??. Detection tests based on artificial intelligence techniques for the identification of outliers, shifts and trends in Canadian hydrometric data. Submission to Canadian Water Resources Journal expected in March 2004.
- Lauzon, N., and Lence, B. J., 20??. New Directions for the Characterization of Anomalies in Data. Submission to Water Resources Research expected in February 2004.
- Lauzon, N., and Lence, B. J., 2003. Detection of anomalies in hydrometric data using artificial intelligence techniques. CSCE 16th Hydrotechnical Conference, Canada Centre for Inland Waters, Burlington, Ontario, Canada, October 22 – 24, on CD-ROM, 10 pages.
- Lauzon, N., Lence, B. J., and Maier, H. R., 2002. Use of Artificial Intelligence Techniques for the Estimation of Cyanobacteria (Blue-Green Alage) Concentrations. The 30th Annual Conference of the Canadian Society for Civil Engineers, Nollet, M.J. and Trépanier, M. (eds.), Montreal, PQ, Canada, June 5 – 8, on CD-ROM, 10 pages.
- Lauzon, N., and Lence, B. J., 2001. Identification of Long Term Patterns in Hydrologic Data Using Fuzzy and Neural Network Techniques. The 2001 International Congress on Modelling and Simulation, The Australian National University, Canberra, Australia, December 10 – 13, F. Ghassemi, D. White, S. Cuddy, and T. Nakanishi (eds.), Volume 4: General Systems, 1925-1930.
- Lauzon, N., and Lence, B. J., 2001. Adaptive Fuzzy Technique for Inflow Modeling. The 3rd International Conference on Intelligent Processing and Manufacturing of Materials, Vancouver, BC, Canada, July 29 – August 3, J.A. Meech, S.S. Veiga, M.M. Veiga, S.R. Leclair and J.F. Maguire (eds.), on CD-ROM, 11 pages.

Acknowledgements

I want to sincerely thank my supervisor, Professor Barbara Lence, for her continuous support through my Ph.D. work. She has given me all the liberty to explore the most unexplored paths of my area of specialty. And her rigor and no-nonsense attitude have made sure I would not get lost in the way. I am very grateful for the contribution of the members of my supervisory committee, namely, Professors Ricardo Foschi, John Meech and Robert Millar (University of British Columbia). Thanks to Professors Guy Dumont and Daniel Moore (University of British Columbia), and Professor William Lennox (University of Waterloo), who accepted to review this thesis prior to its defense. Thanks also to Professor Ian McKendry (University of British Columbia), who chaired the defense committee. Quite deserving of considerations is my current employer, Professor François Anctil (University Laval), for his flexibility prevented me from delaying the completion of my Ph.D. Worthy of mention, for by one way or another making this Ph.D. journey an easier task for me, are Andreas Prucker and Joan liu, as well as Professors Holger Maier (Adelaide University), Micheal Quick (University of British Columbia), and Slobodan Simonovic (University of Western Ontario). Also, I will not forget Professor Donald Burn (University of Waterloo) for his great efficiency on a request I made to him, Professor Jean Rousselle (École Polytechnique de Montréal) for his encouragements, and Danli Wang for her moral support. At last, I wish to acknowledge the sources of funding that helped me financially for the duration of my Ph.D. program, namely: the Natural Science and Engineering Research Council of Canada; Fonds québecois de la recherche sur la nature et les technologies (formely Fonds FCAR, from the province of Québec); Manulife Financial through the Canadian Council of Professional Engineers; and the University of British Columbia through my supervisor, the Department of Civil Engineering, the Faculty of Graduate Studies, the Flood Control Research Fund, Green College and the Earl R. Peterson Memorial Trust.

Chapter 1
Introduction

This work develops and analyzes new approaches based on artificial intelligence techniques (AITs) for evaluating the quality of data sequences employed in the design and management of water resources systems. Factors that affect the quality of data sets are anomalies such as 1) outliers, which are individual data having statistical properties that differ from those of the overall population; 2) shifts, which are sudden changes over time in the statistical properties of the historical records of data; and 3) trends, which are systematic changes over time in the statistical properties. The approaches developed in this thesis, based on AITs, are designed to identify these anomalies if they are present in the data sets under study. These approaches depart from the commonly used traditional approaches, often based on probabilistic and statistical methods. They are shown to perform similarly to the traditional approaches, and as such they confirm the validity of these traditional approaches and provide complementary information. These AIT-based approaches may further complement the traditional approaches in that they may integrate subjective inputs such as experts' judgment in the diagnostic process. Such information may be an important element in the task of evaluating the quality of data sets. The results presented in this work also show that the AIT-based approach may constitute an enhancement to traditional approaches in specific instances, that is, in the diagnostic of shifts and trends for multivariate cases, where more than one sequence of data are tested simultaneously.

For the purpose of the design and management of water resources systems, it is important to be aware of anomalies in data, for they can induce a bias in the estimation of water quantity and quality parameters, and may consequently lead to improper water resource management policies or infrastructure design. The identification of these anomalies has always been a difficult task to undertake, especially in the case of outliers, and consequently only a few consistent identification procedures and guidelines exist for application to water resources, as well as many other domains related to the environment and natural resources. Yet there is a need for such identification procedures, if only because

of the increasing amount of data available in such domains in general. Large scale monitoring networks in water resources, such as those of Environment Canada or the United States Geological Survey, for example, have been in place for a significant period of time now, and space programs have also led to the production of large sources of remote sensing data.

This work also presents some new developments in modeling natural phenomena, again through the help of AITs. Here, these techniques are integrated into well known, standard simulation models so as to better represent natural processes within the model structure in order to improve the accuracy of estimates provided by the models. This approach is applied to both inflow modeling and algae concentration modeling. These applications demonstrate the flexibility of the use of AITs, and show their potential for correctly describing physical mechanisms involved in natural phenomena.

The common ground of the approaches developed in this thesis is the application of AITs to data quality control and modeling. When employed as they are in this work, AITs lead to appreciable technical innovations. An important contribution of this thesis is the development and exploration of the capacity of the AITs for defining knowledge bases. The definition of knowledge bases implies the description of the data domain with respect to features or patterns present in the data. An example of a knowledge base for the evaluation of data quality is a set of patterns, some of which are typical of those of data sequences affected by anomalies, and others of which are typical of those of unaffected data sequences. Other examples of knowledge bases may be the development of definitions that relate 1) surface runoff contributions to antecedent conditions of inflow, precipitation and temperature on a watershed for inflow modeling; and 2) algae growth and mortality rates to water temperature, light intensity and nutrient content for algae concentration modeling. It is not common practice among professionals in water resources to use AITs for the purpose of defining knowledge bases. This work shows that AITs can provide satisfying results when employed for such a purpose. It is also argued that they may offer greater flexibility, and possible improvement when compared with statistical methods that are also used to define knowledge bases.

Chapter 2 provides a literature review on the topics of the evaluation of data quality, the development of simulation models and the definition of knowledge bases, and explains

the need for identifying new approaches, based on AITs, for addressing these topics. Chapter 3 provides the mathematical background related to the AITs employed in this work, that is: fuzzy logic, fuzzy c-means and the Kohonen neural network. This chapter gives a general overview of these techniques, while the specific details related to their use are presented in the subsequent application chapters, that is, Chapters 4 through 7. Chapters 4, 5 and 6 are dedicated to methods of evaluating the quality of data sequences. Chapter 4 presents the tests conducted on one variant of AITs applicable for the detection of shifts and trends in data sequences. This variant is compared with statistical detection tests commonly employed for the identification of shifts and trends, using synthetic data that represent hydrometric data sequences. Chapter 5 focuses on a second variant of AITs, which is applicable to all anomalies, that is, outliers, shifts and trends. Synthetic data representing hydrometric data sequences are employed to evaluate the performance of this variant. Chapter 6 represents the reality test, where the second variant for the detection of outliers, shifts and trends is applied on real hydrometric data for representative river basins in Canada. Chapter 7 is dedicated to the assessment of the developments for the modeling of natural phenomena, based on two applications: inflow modeling and algae concentration modeling. Chapter 8 discusses the overall results of the thesis, the advantages and disadvantages of the approaches presented and future work in this area. Chapter 9 provides the final conclusions and recommendations.

Chapter 2
Literature Review

2.1 Current Practice in Water Resources

The study of water resources systems, or any environmental systems for that matter, often requires the replication through simulation modeling of the mechanisms that link a phenomenon of interest to its generating causes. Figure 2.1 offers the simple, traditional illustration of this process, where the simulation model block includes all the mathematical formulations that explain the behavior of some system under study. The outputs are the response of this model when given a set of inputs. The inputs, often called forcing factors in the literature, represent the constraining elements that limit the behavior of the mechanistic model and include data (i.e, measurements or observations) taken from the field, and predetermined values for the model parameters, as necessary.

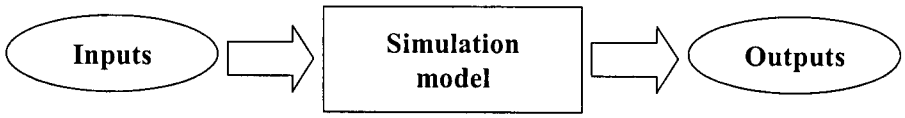


Figure 2.1. Typical model structure.

When the outputs of the model are compared with data from the field, discrepancies generally appear. These differences are due to uncertainties, which Vicens et al. (1975) classify into two categories: natural and informational. Natural uncertainties refer to the difficulty in adequately describing observations from the field, either the inputs for the model or the data required to calibrate and verify the model. Such information may indeed be affected by errors or anomalies, and if these are not detected and addressed appropriately, they can induce a bias in the description of the data (e.g., the mean, standard deviation, probabilistic distribution). If the data serve as model input, the bias may propagate through the simulation model to the output. If the data are compared with the model outputs as in calibration and validation, the bias may taint the calibration and

validation processes. Informational uncertainties are divided into parameter uncertainties and model uncertainties, and refer to the difficulty with current knowledge to properly establish the structure of the model or to select the values of the parameters.

This thesis introduces developments that are designed to help reduce uncertainties, more specifically, in the context of the application cases provided, natural and parameter uncertainties. The goal of this chapter is to provide relevant background knowledge associated with the approaches for reducing data uncertainties. The following section addresses data quality and the control of it through methods that detect errors or anomalies. Adequate control of the quality of data leads to a better description of these data and therefore to a reduction of natural uncertainties. The next section identifies issues related to modeling and methods employed to reduce informational uncertainties. The last section introduces artificial intelligence techniques, justifies these techniques for their use in reducing uncertainties and compares them with conventional methods for doing so.

2.2 Water Resources Data Quality

Few would contest the need for good quality data for the study of water resources systems. Yet, little work has been undertaken that evaluates the quality of data emanating from these systems, compared with the large body of work undertaken in other sectors. Sectors in which analysis of data quality would be very suitable are the manufacturing industry, health sciences, services and social sciences. This suitability is attributed to the fact that such sectors can offer very well controlled systems, where their physical and structural mechanisms are well understood, the processes and experiment protocols employed to observe the systems are clear, consistent and well detailed, and the inputs to and outputs from the systems can be fully accounted for. All of these elements facilitate the study of data quality, for the attention may be focused mainly on finding anomalies in data, and not on trying to understand the mechanisms of the systems. Sometimes these systems can also integrate redundant measurement and control procedures, which allow cross-checking of information to confirm the validity of the datasets. On the other hand, water resources systems, and many other environmental systems in fact, may be considered as being very far from well controlled systems. Some mechanisms of these systems may not be well understood (e.g., in the case of algae growth as a function of the energy and nutrient

factors presented in Chapter 7 of the thesis). The processes and experimental protocol for observing these systems may not be spatially representative and consistent over time. The inputs and outputs may not be fully accounted for (e.g., in the case of underground water transfer from one watershed to another, or of unregulated water diversions). All of these elements complicate the process of assessing water resources data quality, for one cannot be sure whether anomalies in the data set are genuinely anomalous or are the results of misunderstood and therefore unaccounted for mechanisms, of an inadequate observation approach, or of inputs or outputs that cannot be measured. Assessing data quality on such systems is therefore a challenging task, which is in part responsible for the limited literature related to the subject. In the following section, the concept of data quality analysis as applied in the manufacturing industry, health sciences, services and social sciences is presented, and then related to water resources systems. Next, a discussion of three common cases of anomalies affecting the quality of water resources data, outliers, shifts and trends, is presented.

2.2.1 Background

Regarding the assessment of the quality of data in water resources systems, let us consider as suitable two similar definitions of the term quality, one given by the American Society for Quality Control (ASQC, 1983): *"Quality is the totality of features and characteristics of a product or service that bear on its ability to satisfy given needs,"* and the traditional definition as given by Farnum (1994): *"Quality is the conformance to specifications."* Satisfying these definitions implies that the data (the product) must prove adequate with respect to some defined attributes (specifications, features or characteristics), the most common of which as indicated by Wang et al. (1993), Holt and Jones (1998), and Brackstone (1999) as being: accuracy, relevance, completeness, coherence, interpretability, timeliness, and accessibility. With water resources systems, it is often difficult to reach high standards with respect to many of these attributes.

Accuracy is fundamental to quality, although it can never be expected to be achieved absolutely. It is an attribute that somehow stands apart from the other attributes, which are often related to each other, although these other attributes can affect accuracy. Accuracy should at least be reached to a point where the risk of misinterpretation as a result

of the use of the data is minimized (Holt and Jones, 1998). The major constraint of this minimization is that exact measurements are often prohibitively expensive (Brackstone, 1999), especially in light of the fact that the cost due to errors is very hard to quantify (Liepins, 1989). Errors affecting the accuracy of data measured in the field of water resources can be isolated, which are local errors that may occur at regular or irregular intervals (i.e., outliers), or persistent, which are errors that are propagated over some intervals (i.e., shifts or trends). Both isolated and persistent errors can either be random, which means that they are not the result of identifiable structures, or systematic, which implies that they follow some known structure (WMO, 1985). Origins of errors listed by Van Der Shaaf (1984) and the WMO (1985) include factors due to anything from malfunctioning measurement instruments (i.e., quality, technical limitations, condition) to errors in the processing of the data measured (e.g., coding errors, typos). These authors also mention that focusing on the origins of errors in order to identify them in a data set is not a viable option, because often measurement conditions do not allow the possibility to monitor the origins of errors, and some of these origins cannot be easily monitored anyway. Therefore, very often, the data set itself is the only available material and one can only hope to detect errors with respect to distinct features or patterns they exhibit in the data. The methods developed in this thesis are designed to address the attribute of accuracy. Nevertheless, the other attributes associated with data quality are briefly reviewed here.

Relevance is the degree to which the data meet some defined needs. This attribute is as important as accuracy. With respect to this attribute, two topics are of interest concerning environmental systems: scale (temporal and spatial) and domain of study. Scale refers to the appropriate magnitude of the temporal (seconds, minutes, days) and of the spatial (mm, cm, m, km) increments in order to mathematically characterize some given phenomenon. The measurement network must be dense enough and observe frequently enough to provide a good description of the spatial and temporal variability of the phenomenon under study. In the areas of hydrology and hydrodynamics, Klemes (1983), Anderson and Burt (1985), Beven et al. (1988), and Martin et al. (1999) state that one must also consider that some given phenomenon may be the result of many generating causes, each possibly occurring at different scales, which makes the collection of relevant data a complex task. The domain of study refers to the proper delineation of the region where the phenomenon of interest takes

place, and this is a basic issue that is discussed extensively in textbooks such as Freeze and Cherry (1979), Robertson et al. (1988) and Maidment (1993). Relevance is achieved when the measurement network covers the domain adequately. Completeness indicates whether there are enough data in terms of both type and quantity to adequately characterize the natural phenomenon under study. Completeness is strongly related to the attribute of relevance, concerning the similar issues of scale and domain of study. Completeness also includes consideration of the amount of data, that is, whether the database contains all of the behaviors expected for the phenomenon under study or only a few specific situations (Anderson and Burt, 1985; Beven, 2001). Coherence reflects the degree to which data sets can be brought together within a broad framework. To illustrate the importance of this attribute, consider the work of Overton et al. (1993) and Endreny and Jennings (1999), who attempt to merge two databases containing water discharge and water quality parameters, respectively, so as to increase the volume of data available. Indeed, it might be common that two organizations have measurement networks in the same region, observing the same phenomena. One could be very interested in being able to aggregate the observations from the networks of both organizations so as to have a global, larger database. Overton et al. (1993) and Endreny and Jennings (1999) show that merging databases is not straightforward. The measurement networks could have been used for different purposes, and the issue of scale, for example, may call for adjustments in the data before they are merged so as to avoid incoherence.

Interpretability, timeliness and accessibility are attributes that are considered of lesser importance technically, yet they should not be dismissed because they refer to the capacity of people to use the data. Interpretability refers to the capacity to understand the data at hand. This means that the database must be properly documented. Timeliness is related to the delay between the moment the data are measured and the moment they become available. Accessibility indicates whether the data are available at all. It can be argued that the more a database is used the more likely higher data quality can be achieved (Orr, 1998). The larger the user base, the greater the pressure for achieving high quality standards. If the database is used frequently, the users are likely gaining expertise and are therefore more apt to detect changes and anomalies over time. In brief, provided there is a

demand, a good way to maintain and improve data quality is to make the data easily accessible (Orr, 1998).

As mentioned previously, accuracy is the attribute targeted by the methods developed in this thesis, but the other attributes must be kept in mind, for they can possibly affect accuracy. Accuracy is dependent on the capacity to detect errors or anomalies in datasets. In the sections that follow, an overview of methods used to detect anomalies such as outliers (isolated events) and shifts and trends (persistent events) is given.

2.2.2 Outliers

An outlying observation, or an outlier, is a general term that refers to either a contaminant, which is *"any observation that is not a realization from the target* (probability) *distribution,"* or a discordant observation, which is *"any observation that appears surprising or discrepant to the investigator"* (Beckman and Cook, 1983). To summarize, outliers are individual data having statistical properties that appear to differ from those of the overall data population. Beckman and Cook (1983) provide an extensive review on the subject of outliers and classify the methods for addressing outliers as being either of the accommodation type or of the identification type.

Accommodation methods attempt to dampen the effects of outliers through suitable modifications of the methods of analysis describing the data or of the methods estimating the values of model parameters. The strategy is to develop modifications that are not too sensitive to extreme values, which in data sets are often considered as outliers. One simple case, as illustrated by Pearson (2001), is to use the median absolute deviation from the median (MAD) instead of the standard deviation as a way of characterizing the variation of a sample around a reference position (mean or median). The standard deviation is the square of the sum of differences from the mean, and therefore magnifies the effect of extreme values, while the MAD considers the median absolute value of the differences from the median, therefore giving equal weight to both extreme and "normal" values. Similarly, probability-weighted moments, also called linear moments or L-moments, described by Hosking (1989 and 1990), are also less sensitive to outliers than standard moments. L-moments are linear sums of weighted components while standard moments are sums of components elevated to the power of the associated order (i.e., the second order

moments use a power of two, the third order moments use a power of three, etc.). L-moments have often been used in water resources, specifically for frequency analysis of extreme events such as floods (Hosking and Wallis, 1990; Stedinger et al., 1993; GREHYS, 1996). They have also been used specifically for the characterization or identification of outliers (Hosking, 1995; Zafirakou et al., 1998). Accommodation is not the preferred method for addressing outliers. It is not entirely reliable. It is more appropriate to identify outliers before they propagate through the analysis or modeling processes and induce a bias in the final conclusions.

The identification or detection approach, in its simplest form, can often be reduced to a simple visual inspection of the data in which the data points that appear to be far from the range of the overall data population are considered as outliers (Collett and Lewis, 1976; Beckman and Cook, 1983). When the eyes cannot be trusted, statistics have been used.

In the simplest application of statistics for identifying outliers, the points in the data set are considered as independent of each other. The possibly best known, and certainly easiest method for the detection of outliers in this situation is the application of the "3σ edit rule," a statistical test where a point can be considered suspicious if it is located more than three times the standard deviation of the data set from the mean (WMO, 1985; Pearson, 2001). The "3σ edit rule," however, performs poorly under situations with multiple outliers (Pearson, 2001). Many other, more elaborate methods exist for the detection of outliers. In summary, for data sets with independent points, typical developed methods are: 1) graphical techniques, such as those of Bacon-Shone and Fung (1987), based on the measure of distances between data points; 2) statistical tests, such as those presented in Bradu and Hawkins (1982), Jain (1981), Kottegoda (1984), and Rosner (1975); and 3) fitting methods, such as those of Aitkin and Wilson (1980) or Kitagawa (1979), which find the appropriate parameters of assumed distributions, one for the population of normal points and one for the outlier population. Note that the third type of detection methods also constitutes a form of accommodation. These three types of methods have at least one point in common, that is, they require the assumption of specific distributions for the populations of normal points and outliers. These distributions may not always be the Normal distribution, as indicated and demonstrated by Kottegoda (1984). It remains that the need to assume the population distributions is restrictive, for cases with real data are not always in compliance with the

assumed distributions. Many methods, such as all of the aforementioned cited references, can also address situations where several outliers, not only one, are present in the data sets, unlike the "3σ edit rule."

Whatever the methods, the risk of false detection is acknowledged to be present. Masking and swamping effects respectively refer to the cases of outliers falsely detected as normal values and to the case of normal values falsely detected as outliers (Beckman and Cook, 1983). Figure 2.2 illustrates the occurrence of cases of false detection.

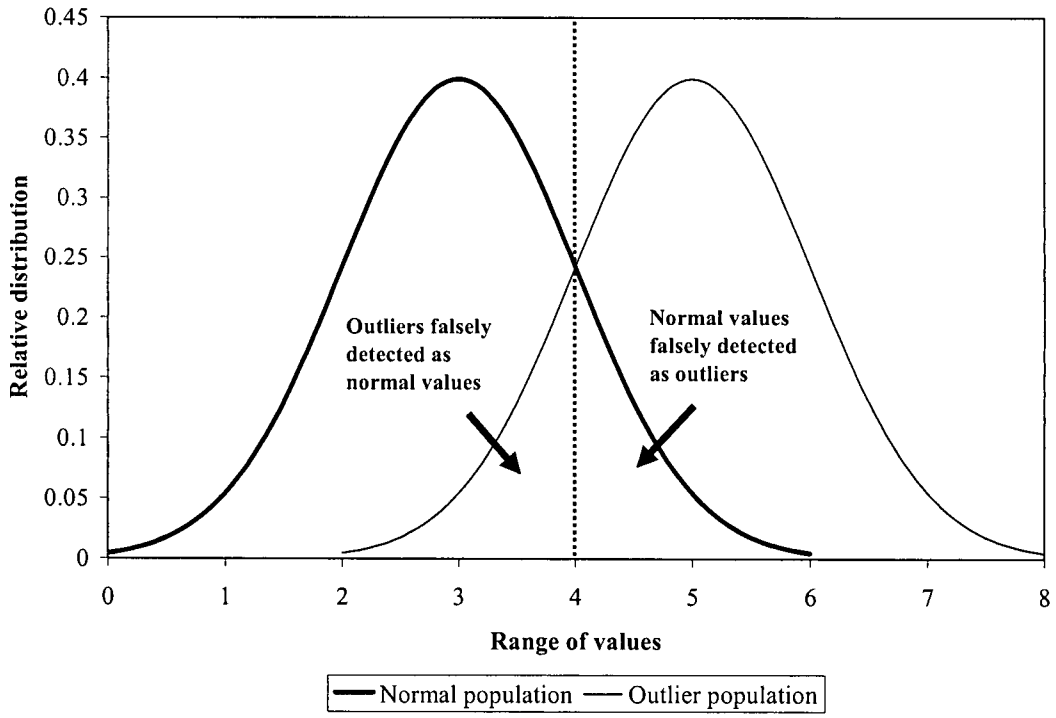


Figure 2.2. Occurrence of false detection for outliers.

If the distributions of the population of normal values and the population of outliers are known, there would be a very high likelihood that both distributions would in part overlap. The purpose of detection methods is to distinguish between normal values and outliers, in the simple and most common cases (i.e., statistical tests), to determine the line (i.e., the dashed line in Figure 2.2) between the two populations. Because the distributions are overlapping, there will be a portion of the outliers that will be judged as normal values (the left tail of the outlier distribution on the left side of the dashed line), and a portion of

the normal values that will be judged as outliers (the right tail of the distribution of normal values on the right side of the dashed line). Figure 2.2 presents the situation where outliers are larger than the normal values, but the conclusions about distribution overlapping and false detection also apply to situations where outliers are smaller than the normal values.

For data sets that are multivariate or time dependent, the graphical, statistical and fitting methods for data sets made of independent points can also be applied. Although such methods do not take the dependence of the data points into account, and do not utilize information regarding dependence to avoid false detection. Linear regression is a common tool for the quantification and modeling of multivariate dependence, and can be used as a basis for the detection of outliers. For instance, Brown (1975), Cook et al. (1982), and Dempster and Gasko-Green (1981) have developed tools to analyze the residuals of linear regression models for the purpose of identifying outliers. Andrews and Pregibon (1978) and Draper and John (1981) have devised statistical measures that evaluate the influence of outliers in the estimation of the parameters of the linear regression models. Box and Tiao (1968) and West (1984) also attempt to evaluate the influence of outliers on model parameters, using a Bayesian analysis. Time series analysis through the use of Box and Jenkins models (Box and Jenkins, 1970), which are a particular form of linear regression models, have also provided opportunities for the detection of outliers in time dependent data. Fox (1972) and Ljung (1993) present tests of detection based on the analysis of the model parameters, while Abraham and Box (1979) analyze the influence of outliers on the model parameters through the use of a Bayesian analysis. Data sets showing spatial dependence, which is a particular case of multivariate dependence, can also take advantage of developments in geostatistics. Although it is not purposely designed for the detection of outliers, the consideration of a "nugget effect" in kriging is a common way to represent micro-variability and possible measurement errors (Kitanidis, 1993). It involves a slight modification (i.e., the addition of a Dirac function) of the semivariogram, which is an expression of the correlation between data points. Also, Bardossy and Kundzewicz (1990) present semivariograms that are specifically designed to be used for the detection of outliers. All of these detection methods applied for multivariate or time dependent cases involve some form of accommodation. Indeed some form of modeling is required, and the value of model parameters, and the structure of the model itself to some extent, can be

affected as a result of the presence of outliers. As previously mentioned, some methods are designed to analyze the effects of outliers on parameter values, yet their efficiency is dependent on the restrictions imposed on the nature of the outliers and of the normal values, that is, they must follow a specific distribution. In fact, with the exception of geostatistical techniques where the restriction is implied from the structure of the semivariogram, all of the techniques require some restrictive assumptions regarding the nature of the outliers and/or of the normal values. Of course, real data are not necessarily in compliance with these assumptions. Again, the risk of false detection is regularly acknowledged.

The methods described in this section for the analysis of outliers quite often constitute fine, rigorous and elaborate mathematical developments. These strengths may also be viewed as a disadvantage at times, for such mathematical developments are not easy to introduce to a wide selection of users. First, because these developments target specific objectives, their inclusion in standard commercial mathematical software (SPSS, Matlab, etc.) is not economically appealing. Second, implementing such methods in an applied environment, even with the help of tools development procedures offered in mathematical software, requires well trained professionals with strong foundations in mathematics, statistics and their domain of application. In general, the transfer of these methods from their theoretical origins to a more applied environment is challenging.

Applications in water resources have permitted some developments for outlier analysis, especially in the analysis of hydrometric observations, which are the type of data employed in this thesis. For example, Kottegoda (1984) employs statistical tests for the detection of outliers in data sequences of annual maximum floods. Hydrometric data are time-dependent products of impulses (i.e., precipitation) and their propagation over time. An easy strategy to identify a given point in the data set as an outlier is to evaluate if the difference between this point and the preceding one is too large to be considered as coming from a likely impulse (i.e., a precipitation event) for this hydrologic system. Such a strategy is employed by Krawjeski and Krawjeski (1989), where the threshold on the difference is determined manually, and by Lauzon (1993), where the threshold is based on the unit hydrograph associated with the hydrometric sequences of inflows. Of course such a strategy can be refined, and the judgment on the extent of the difference between data

points can also be based on a comparison with other data sets, such as those from a neighboring hydrometric or precipitation stations (Rassam et al., 1991). A more elaborate strategy is to build a model to replicate the data set under study. The model output is then assumed to represent the truth, and a data point is considered as an outlier if it differs from the model estimation by too great a margin. Examples include the work of Bennis and Bruneau (1993a and b), who employ Box and Jenkins models with or without the Kalman filter, Krawjeski and Krawjeski (1989), who use a mechanistic model, Perreault et al. (1991) and Nguyen (1993), who develop methods based on a combination of models, and Bérubé et al. (1987) and Rassam et al. (1991), who employ Fourier series. It should be noted that these models must involve some form of outlier accommodation, which means that steps need to be taken to make sure that outliers do not affect the determination of the values of the model parameters. A distinct strategy is that of Krawjeski (1987) and Krawjeski and Krawjeski (1989), for the detection of outliers in radar rainfall and hydrometric data respectively, that make use of influence functions. They use the developments given in Delvin et al. (1975), where the influence functions determine the influence of each data point in the estimation of parameters such as the mean, standard deviation or correlation coefficient. A data point that has a disproportionate influence is suspected to be an outlier. For the detection of outliers in radar rainfall and hydrometric data, the spatial correlation and autocorrelation are the chosen parameters, respectively, and the technique is quite similar to a pattern recognition tool, especially in the case involving radar rainfall data. The disadvantages of this strategy are that the data must follow the standard Normal distribution, and that the influence function does not work well with non-stationary data, that is, when the parameters such as the mean, standard deviation and correlation vary over space or time.

2.2.3 Shifts and Trends

As is generally observed in the literature, these kinds of anomalies involve data sequences that evolve with time. Shifts are sudden changes over time in the statistical properties of the historical records of data, while trends are systematic changes over time in the statistical properties. As with outliers, the methods available for addressing shifts and trends can be divided into two categories: 1) those that attempt to accommodate the

presence of shifts and trends, and 2) those that are meant to detect such anomalies. Accommodation implies the use of models, and there are a large variety of these in water resources, ranging from the simple to the complex. Very often, models are not specifically designed to account for shifts and trends, although they can sometimes accommodate these anomalies. A more general discussion of simulation models is offered in Section 2.3. The few models that directly accommodate shifts and trends are usually Box and Jenkins models, and they have been widely used in water resources, particularly for the modeling of water inflows. The common practice in the development of such models is to first search for trends in the data sequences, and this can be accomplished through the analysis of autocorrelation coefficients, as described in such textbooks as Box and Jenkins (1970), Salas et al. (1980), Pankratz (1983), or Bras and Rodriguez-Iturbe (1985). Seasonal and other periodical phenomena (e.g., El-Nino and sun spots) can be considered as shifts occurring at regular intervals, and Box and Jenkins models offer the flexibility to account for these. The most common modeling strategy is to develop general models that include adequate representation of all periodic elements. Another strategy is to develop a so-called periodic model, which is a series of models, each one meant to be employed at a very specific period in time. Examples of the use of such models can be found in Pagano (1978), Troutman (1979), Thompstone (1983), Thompstone et al. (1985), Vecchia (1985), Fernandez and Salas (1986), and Bartoloni et al. (1988). When shifts occur at irregular intervals, one can resort to intervention analysis, which requires the use of Box and Jenkins models again. As illustrated in the work of Box and Tiao (1975), Hipel (1975), Lettenmaier (1980), Hipel et al. (1981), and McLeod et al. (1983), these models include step functions, which activate a specific part of the model structure when needed so as to account for the presence of shifts in the data sequences. Further developments with these kinds of models have been limited since the beginning of the 90s. Such models are restrictive, as they require that the data follow a specific distribution, usually the Normal distribution, although the Gamma distribution is also used (Fernandez and Salas, 1986). The main criticism of these models, which are linear regression equations, is that they are too simple to adequately represent complex natural phenomena such as river inflows, regardless of the presence of shifts and trends. Building on the concept of intervention analysis are several tests that allow evaluation of the potential location and amplitude of shifts. The Cumulative

Sum of Deviations (CUSUM) and Exponential Weighted Moving Average (EWMA) tests can detect changes in the statistical properties of data, usually the mean, based on the sum of the deviations of the data from their expected values from models. Large sums indicate a shift. Such tests have been applied particularly to the quality control of industrial processes, computer and electrical engineering applications, and in biological and biomedical studies. General description of the CUSUM and EWMA tests can be found in Basseville and Nikiforov (1993) and Farnum (1994). Applications for the detection of shifts with CUSUM can be found in Radharamanan et al. (1994), Mantua et al. (1997), Jarpe and Wessman (2000), Reynolds and Stoumbos (2000), and Khoo and Quah (2002), while EWMA has been applied by Prabhu and Runger (1997), Srivastava and Wu (1997), Cahn and Zhang (2000), and Reynolds and Arnold (2001). Farnum (1994) mentions, however, that the results of these tests can be affected by the model representing the data under investigation. An inappropriate model can lead to a bias in the results of the tests.

As with outliers, the best strategy is to detect shifts and trends before considering building models that can accommodate them. The domain of water resources, particularly in the study of hydrometric data, has provided opportunities to develop and apply detection methods, although at a different time scale than that considered for the purpose of accommodation. The most commonly used statistical tests for the detection of shifts are the Student's and Mann-Whitney tests (Salas, 1993). In the presence of a shift a data sequence is considered as being composed of two subsets of data (i.e., one before and one after the shift), each coming from a different population. Both tests provide a measure of the distance that separates these two subsets. Both tests require that the data points be independent. The Student's test requires that the data set follows a Normal distribution, which is not a necessary requirement for the Mann-Whitney test. The disadvantage of the latter test is in the characteristic of its response, which is assumed to follow a standard Normal distribution, but its real distribution is only an approximation of the Normal distribution (Hirsh et al., 1993). This leads to an imprecision in the decision criteria of the test. Both tests also require that the presumed location of the shift is known. If this information is not known, a practical way to address the problem is simply to test the data at all possible locations of the shifts. A more refined approach is to use Bayesian analysis to probabilistically determine the possible location of the shift, as demonstrated for example

by Lee and Heghinian (1977) and Perreault et al. (1999 and 2000). The analyses performed by these authors require that the data follow a Normal distribution.

For the detection of trends, the most well known test is the Mann-Kendall test, but the Spearman test has also been applied, for instance, by Lettenmaier (1976) on water quality data and by Anderson et al. (1992) on water inflow data. Both tests require that the data points be independent, but do not require that they follow any particular distribution. Like the Mann-Whitney test, the responses of both tests are presumed to follow a specific distribution, the standard Normal distribution for the Mann-Kendall test (Conover, 1980; Salas, 1993) and a numerically established distribution for the Spearman test (Conover, 1980).

The Mann-Kendall test has been frequently used for cases in North America. In water resources, a study of trends involving this test have been performed on water quality data (Lettenmaier, 1976; Hirsh et al., 1982; Hirsh and Slack, 1984), climatology-related data such as temperature and precipitation (Gan, 1992 and 1998; Gan and Kwong, 1992; Burn, 1994; Lettenmaier et al., 1994), and hydrometric data (Lettenmaier et al., 1994; Westmacott and Burn, 1997; Leith and Whitfield, 1998; Yulianti and Burn, 1998; Lins and Slack, 1999; Zhang et al., 2001; Cunderlik and Burn, 2002). Many of these authors also address the case of evaluating the amplitude of the trends. The capacity of the test for avoiding false detection, which is an issue that affects all statistical tests mentioned thus far, is also discussed to various extents in the references cited above.

With regard to the quality of data, the important point to remember is that the accuracy of data sets must be achieved to the extend that the risk of misinterpretation as a result of the use of the data is minimized, even though high accuracy standards can be expensive and their benefits are difficult to quantify. Of course, accuracy is not the only attribute related to data quality. Relevance, completeness, coherence, interpretability, timeliness, and accessibility are attributes that must be kept in mind while analyzing the quality of data. The methods available for the analysis of anomalies such as outliers, shifts and trends require several considerations. All the methods require some restrictive assumptions on the nature of the data analyzed or their results (e.g., following a specific probability distribution). Accommodation methods should be used with care. They may be able to address the anomalies that are assumed to be present and well defined, but might not

be able to address anomalies that are not assumed to be present but are present nonetheless. The best strategy is therefore to design efficient identification methods before resorting to accommodation. Of course, detection methods are not totally immune to the possibility of performing false detection. Finally, for practical purposes, attempts should be made to develop tools that are easy to implement and easy to grasp by potential users.

2.3 Modeling

2.3.1 Overview

As mentioned in Section 2.2, water resources models may be used to accommodate anomalies such as outliers, shifts and trends. The purpose of this section is to discuss issues related to model structure and parameter uncertainties, in order to complement Section 2.2, which addresses the direct source of data uncertainty. Water resources models come in various forms, some of which are quite general while others are designed to address very specific issues. The term water resources model often refers to tools for understanding water responses in continental areas (i.e., surface and groundwater), although atmospheric and oceanographic models can also be considered as water resources models, if only to close the Earth water cycle. As a guide, a classification of water resources models, partly inspired by the classification of Wurbs (1998), is offered in Figure 2.3. All cases consist of both simulation and optimization models, which are tools that define the system under study, and support decision making, respectively. Ultimately, these tools might encompass the whole planetary water cycle, but for now they are only applied to smaller systems for more specific objectives. A hierarchy exists among models. At the highest level in the hierarchy are atmospheric models, for water in the atmosphere, oceanographic models, for water in the ocean, and watershed models, for water on land. These three types of models should normally be inter-related, since processes at the interface between the atmosphere, the oceans and the land affect the water cycle. Watershed models are the direct products of the so-called branch of watershed hydrology, which *"deals with the integration of the hydrologic processes at the watershed scale to determine watershed response"* (Singh and Woolhiser, 2002). They should ideally include all individual hydrologic processes, which are commonly divided into two categories, surface processes and groundwater processes. Watershed modeling, or hydrologic modeling as a whole, is often considered an

information provider for a number of water resources models. Indeed, hydrologic models provide initial conditions for river and lake hydraulic models, water quality response models, biological interaction models (e.g., fish habitat models), operational models for water regulation infrastructures, water distribution hydraulic models, and demand forecasting models.

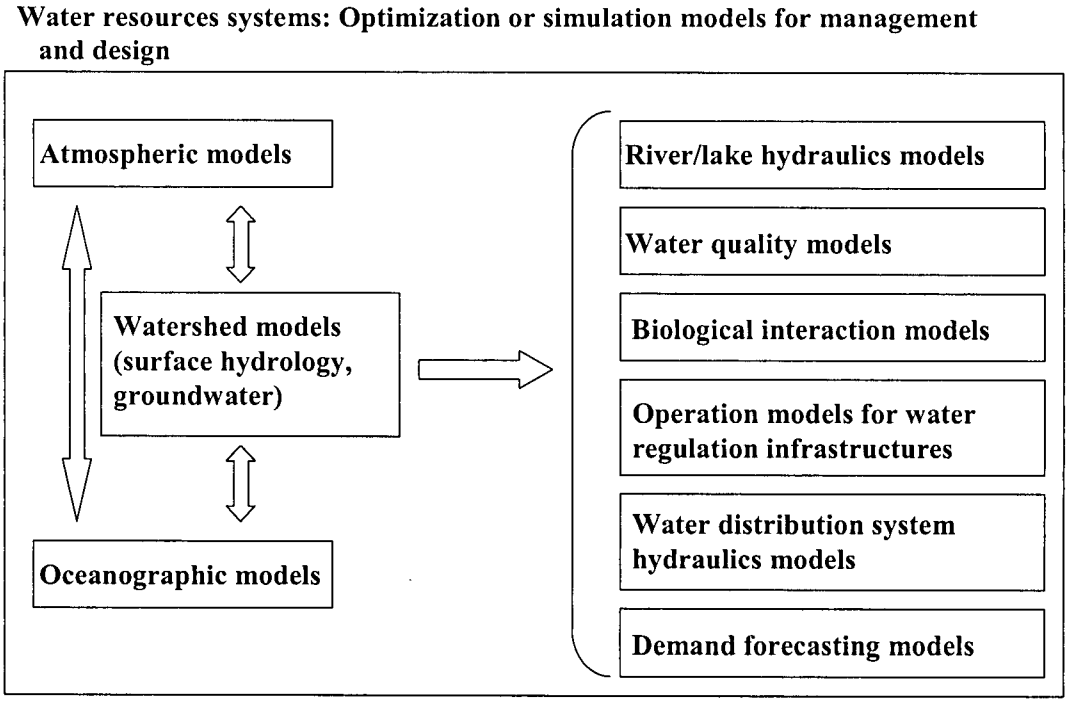


Figure 2.3. Classification of water resources models.

In the sections that follow, the emphasis is placed on watershed models and hydrologic models in general. These models are particularly affected by uncertainties in model structure and parameter estimates (See Sections 2.3.2 and 2.3.3, respectively). A watershed model is also used in an application case of river inflow modeling in Chapter 7. Section 2.3.4 is also dedicated to existing methods for accommodating such uncertainties. To complete the topic of modeling, some details are given on the use of artificial intelligence techniques for modeling purposes.

2.3.2 Model Structure

The structure of a model is the mathematical formulation(s) that explain the behavior of the natural or artificial system under study. A model is said to be physically-based if the mathematical formulations exactly replicate the mechanisms that rule the behavior of the system. At this level, the mathematical formulations are strictly established from the laws of physics (i.e., Newtonian physics in water resources). In contrast, a model is said to be empirical if the mathematical formulations provide only an approximation of the behavior of the system. Here, the mathematical formulations are meant only to provide a rough description of the behavior of the system. Most models fall between empirical and physically-based, depending on the level of accuracy of the mathematical formulations employed to describe the behavior of the system. The number of existing watershed models is large, as exemplified by the long list provided in Singh and Woolhiser (2002). Other lists, and descriptions of models can be found in Singh (1989 and 1995), and Rousselle et al. (1999). WMO (1975, 1986 and 1992) also list a good number of models and evaluate their performance. This large choice of watershed models illustrates the fertile imagination of developers, but also highlights the challenges such developers face with regard to the structure of models.

One of the reasons for the diversity of watershed models is that they are often built to fulfill specific objectives, depending on the interest of the stakeholders. For example, a watershed model employed for agricultural management would need to have strong components that allow a reasonably accurate estimate of soil humidity and evapotranspiration, and would have to be generally more physically-based than a model used for hydro-energetic purposes, where often only a good evaluation of the water inflows at the watershed outlet really matters. Differences may be present for models that fulfill other particular roles also, such as to aid in flood control, the mitigation of environmental impacts, and forest management. Another, rather unfortunate reason for the diversity of models is that the structure of the model to be built is frequently dictated by the availability of data (Singh and Woolhiser, 2002). This reason illustrates the importance of one attribute of data quality: relevance. Uncertainty regarding the validity of the model structure is largely dependent on the validity of the data that explain the behavior of the natural phenomena under study. Statistical tools have been used to determine the potential

magnitude of the effects of input data on outputs (i.e., the relevance). A typical example in watershed modeling involves autocorrelation and cross-correlation analyses as described in Box and Jenkins (1970), Salas et al. (1980), Pankratz (1983), and Bras and Rodriguez-Iturbe (1985). These analyses can then be used to develop time series models. These tools are descriptive in nature, and the importance of such descriptive tools in the future in water resources must be highlighted.

As mentioned in Section 2.2.1, relevance of data is related to the issue of scale, or the appropriate temporal and spatial magnitude employed to characterize a given phenomenon. The natural processes involved in hydrology act at various temporal and spatial scales. Because they are less data intensive and require less computing power, the trend in the past has been to produce hydrologic models that are based on large scales. But as computer power increases, greater efforts have been made to perform studies at smaller scales and to develop more physically-based models. In the 80s and 90s the concerns were in resolving physical issues. This included the determination of the scale that provides an adequate account of the variability of the watershed structure (Beven, 1983 and 1989; Beven et al., 1988; Wood et al., 1991), or the mathematical description of natural processes such as infiltration or evapotranspiration (Abbott et al., 1986a and b; Anderson and Rogers, 1987; Grayson et al., 1992a and b). The developments in these two areas were extensive, but authors such as Anderson and Rogers (1987), Grayson et al. (1992b), Woolhiser (1996), Entekhabi et al. (1999) and Singh and Woolhiser (2002) stress the need for further developments related to physically-based models, and one of their major issues is the need for appropriate (relevant) databases. The data must be representative of the phenomena under study, and measured to reflect an adequate scale. Of course, a measure of relevance can be achieved with descriptive tools such as statistics, as is commonly used on large scale systems, or with the descriptive tools based on artificial intelligence that are proposed in this thesis.

Of particular interest in water resources is the integration of ecological features in the modeling process, such as, among others, the vegetal cover. Entekhabi et al. (1999), and Roberts (2000) indicate the importance of the vegetal cover in the hydrological cycle, as the photosynthesis process can greatly affect the evapotranspiration process and soil humidity. As another example, environmental impact studies quite often attempt to determine the

impacts on the fauna and flora of changes in the characteristics of water resources systems (e.g., water diversions). As indicated by Harte (2002), ecology is based on complex interdependencies, which make the development of predictive tools such as models a difficult task, and this is why descriptive tools such as statistics are often employed instead in this domain. To properly integrate ecological features in the modeling process in water resources, a proper description of these features must be performed and therefore powerful descriptive tools are required.

2.3.3 Parameters and Calibration

Parameters originate from the structure of the models, and therefore the validity of the existence of such parameters is dependent on the validity of the structure of model. Because the validity of the model is largely dependent on the validity of the data, the data quality attribute of relevance therefore affects parameter uncertainty. Data relevance is not the only source of parameter uncertainty, for the calibration process that determines the value of the parameter is also a contributing factor.

Automatic calibration procedures are generally optimization modules that require the following four elements: an objective function, an optimization algorithm, a termination criterion and calibration data. The objective function must be minimized or maximized and must normally express the goodness of fit of the parameters with respect to the data. Probably the most common objective function is the sum of the squared difference between estimated and observed system response values, but Clarke (1973) indicates that such a function might rarely be appropriate. Diskin and Simon (1977) provide several other objective functions. The optimization algorithm is a search procedure that determines the parameter values that optimize the objective function. Reference books such as that of Rao (1979) or Nash and Sofer (1996) provide descriptions of optimization algorithms, many of them quite suitable for cases in water resources. Regardless of the optimization algorithm employed, the main concern is always to find the global optimum. A problem that is ill-defined increases the risk of having the algorithm trapped in a local optimum. The number of local optima also increases as the complexity of the problem to be optimized increases. The termination criterion is a target condition, and the optimization algorithm stops searching when this condition is met. The criteria are imposed on the objective function

(e.g., the search stops when the difference between the current and previous values of the objective function falls below some threshold) or the parameters (e.g., the search stops when the difference between the current and previous values of the parameters falls below some threshold). The importance of the relevance of calibration data has already been mentioned, but completeness is also a data quality attribute to consider. A good model must indeed be able to provide reasonably good estimates regardless of the set of inputs. Therefore, the data employed in the calibration process must be of sufficient quantity to be representative of all the possible behaviors of the system to be modeled (Salas et al., 1980; Salas, 1993; Singh, 1989 and 1995).

2.3.4 Uncertainties and Adaptation

The common practice for accommodating uncertainties in model structure and parameter estimates is to rely on more than one solution or simulation result. For structural uncertainties, this means obtaining a global result from the consideration of several models instead of just one. To obtain the global result, the weighted sum of the results of all the considered models is the strategy employed. The weaknesses of some models in some particular situations can therefore be compensated for by the strengths of other models. Typical examples of this strategy are the studies of Ellis (1988 and 1990) that combine the results of several air contaminant transport models for a better estimation of the concentration of contaminants producing acid rain present at several sensitive geographic locations. The values of the weights are determined through a linear optimization technique. In these studies several objective functions are tested for their performance. In watershed modeling, comparisons such as those of the WMO (1975, 1986 and 1992) have shown that models usually perform similarly overall, but that significant differences can occur locally and therefore the combination of models might prove advantageous. Cavadias and Morin (1986) perform this combination with the models employed in WMO (1986), and show improvements in the performance in 80% of the cases, while the other 20% of cases produce results that are only slightly worse compared with the situations where models are used in isolation. Other noticeable examples in hydrology are the works of Perreault et al. (1991) and Nguyen (1993), who combine models so as to build overall tools that may better detect outliers in hydrometric series.

Parameter uncertainty is also addressed through the process of studying alternatives, also called sensitivity analysis. Once the values of the parameters have been determined, sensitivity analysis involves varying the values of the parameters within a range of the feasible values, and then seeing how those variations affect the results of the model. The sensitivity analysis approach is well documented and is usually employed whenever model calibration is involved (Anderson and Burt, 1985). A robust, well documented reference is the work of Beven and Freer (2001), who present recent developments in sensitivity analysis and integrate these in a general framework in order to address parameter uncertainties. This framework includes the determination of the probability distribution of the parameters for Monte Carlo sampling, the definition of measures of model performance with respect to the values of the parameter, and the determination of the distribution of the model response as a function of the parameters.

Model combination and sensitivity analysis constitute valid ways to address structure and parameter uncertainties, respectively, although they are considered to be "after the fact" approaches. Indeed, these techniques attempt to reduce the effects of the potential deficiencies of the models. The strategy that is presented in this thesis is in the development of tools that attempt to find these deficiencies, so that models can be adapted to limit the effects of these deficiencies and hence to reduce the need for methods dealing with uncertainty.

On the subject of addressing uncertainties, adaptation is another strategy that has been commonly used in water resources, particularly in situations that necessitate real-time forecasting from a model (Anderson and Burt, 1985). Adaptation is established on the assumption that the structure of the model is basically correct, but is relatively static and therefore cannot adapt to the changing conditions of the system. The solution is to use the model, as shown in Figure 2.1, and to add a quality control procedure and feedback mechanism, as shown in Figure 2.4. The quality control procedure receives the outputs from the model and checks the difference between these outputs and the observed values. To close the loop, a feedback mechanism imposes adjustments on the inputs (observed data or parameters), the state variables within the structure of the models, or the outputs themselves, so that the outputs calculated after the adjustments are closer to the observed

values. Several feedback mechanisms are documented in the literature, and they range from manual procedures to fully automated tools.

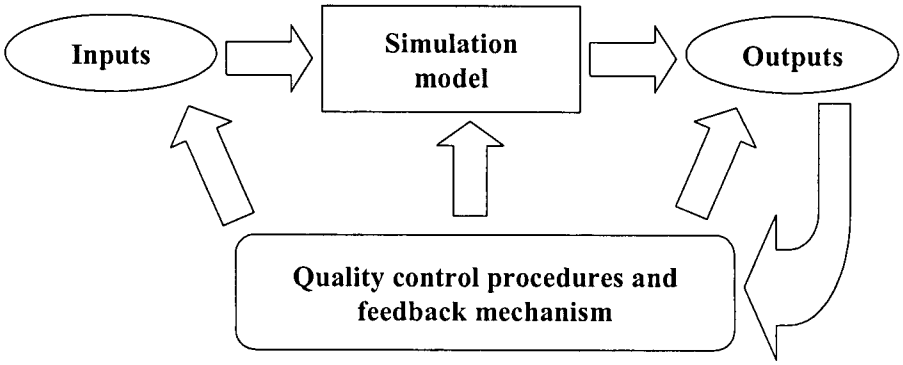


Figure 2.4. Mechanistic model with feedback mechanism.

Most of these feedback mechanisms are model-specific, that is, they have been applied only for the model for which they have been designed, but some generic methods exist. The simplest method adds to the current calculated output the difference between the previously calculated output and the observed values, as demonstrated in the work of Bouchard (1986) and Bouchard and Salesse (1986). A more elaborate method, proposed by Cavadias and Gupta (1978), Lundberg (1982), and Iritz (1988), is to add to the current calculated output an estimated output error as determined by a Box and Jenkins model based on the previous errors. One of the most elaborate, automated and widely used feedback mechanisms is the Kalman filter. A detailed description can be found in Lloyd (1984), but in brief, the Kalman filter performs adjustments based on the assumption that both the elements that must be adjusted and the observed values themselves are subject to errors. It is a refinement of the feedback mechanisms mentioned previously, which, with the exception of that of Iritz (1988), assumes that the observed values are free of errors. The Kalman filter is a recursive procedure, which updates, among other information, the estimation error covariance matrix of the state variables (i.e., the elements to adjust) in order to maintain its adjustment capabilities as the conditions of the system change. Typical examples of the use of the Kalman filter can be found in 1) Georgakakos (1986a and b) for a hydrometeorological model; 2) Lettenmaier and Burges (1976) for a water quality model; 3) McLaughlin (1980) and VanGeer et al. (1991) for groundwater models; and 4) Kitanidis

and Bras (1980a and b), Bergman and Delleur (1985a and b), and Assaf and Quick (1991) for watershed models. The main drawback of this feedback mechanism is that it requires assumed values for the elements of the initial estimation error covariance matrix as well as the covariance matrices for the residuals in the state and measurements equations, these equations being the structural basis of the Kalman filter. This assumption is not readily apparent for any water resources system.

Globally, it can be said that all feedback mechanisms are "after the fact" methods as much as those mentioned previously for addressing model structure and parameter uncertainties. They attempt to reduce the effects of the potential deficiencies, but do not provide any insights that would help to understand these deficiencies and resolve them. It must be noted also that feedback mechanisms, including the Kalman filter, are sensitive to large outliers.

2.3.5 Modeling with Artificial Intelligence Techniques

Neural networks, fuzzy sets and fuzzy logic are artificial intelligence features that have been applied in water resources since the end of the 1980s. Details regarding neural networks can easily be found in general textbooks such as Chen (1996) and Suykens et al. (1996), and more succinct summaries targeted for practitioners and researchers in water resources can be found in Coulibaly et al. (1999) and ASCE (2000a). Neural networks can be very useful in modeling a system whose structure is not well known but is assumed to be non-linear (Chen, 1996; Suykens et al., 1996), and have been often considered as a very advantageous alternative to the frequently employed Box and Jenkins models for water resources applications. Neural networks are simple and very flexible tools, although some care must be given regarding the choice of the inputs and of the network structure in order to ensure an efficient use of this technique (ASCE, 2000a; Maier and Dandy, 2000). ASCE (2000b) provides an extensive list of cases in water resources, with references, where neural networks have been employed. Applications include rainfall-runoff modeling (Achela et al. 1998; Atiya et al. 1999; Coulibaly et al. 2000; Gautam et al. 2000; Imrie et al. 2000; Lauzon et al. 2000; Sajikumar and Thandaveswara 1999; Thirumalaiah and Deo 1998), reservoir management (Neelakantan and Pundarikanthan 2000), surface water quality modeling (Chan-Yan 2000; Lek et al. 1999), groundwater and contaminant

transport modeling (Morshed and Kaluarachchi 1998), and the prediction of the onset of algae blooms (Maier et al. 1998; Maier et al. 2000; Yabunaka et al. 1997). All of the references cited above employ the backpropagation neural network algorithm, which is the most common neural network structure used in water resources. The backpropagation neural network is made of interconnected layers of neurons, including an input layer that receives inputs, an output layer that provides outputs, and one or more layers between the two that process the information.

Details about fuzzy sets and their main derivative, fuzzy logic, can be found in textbooks such as Dubois and Prade (1980), Zimmermann (1991) and Terano et al. (1992), and general examples of their use in natural systems are given in Bardossy and Duckstein (1995). Fuzzy sets and fuzzy logic have been employed in a broader range of circumstance compared with neural networks. For instance, fuzzy sets have been employed as optimization tools by 1) Chang et al. (1997) and Sasikumar and Mujumdar (1998) for water quality management cases; 2) Fontane et al. (1997) for reservoir operation management; and 3) Kindler (1992) for water supply and demand management. They have also been employed as decision making tools in water resources by Yin et al. (1999), Bender and Simonovic (2000), and Despic and Simonovic (2000). For modeling purposes, fuzzy logic has been used for the estimation of 1) a drought index (Pesti et al., 1996); 2) infiltration (Bardossy and Disse, 1993); 3) precipitation (Ozelkan et al., 1996); 4) evapotranspiration (Franks and Beven, 1997); 5) water inflows (See and Openshaw, 1999); 6) reservoir levels (Russell and Campbell, 1996; Shrestha et al., 1996); 7) groundwater (Dou et al., 1995 and 1998); and 8) algae blooms (Maier et al., 2000; Setnes et al., 1997 and 1998).

In the references cited above, AIT have proven to be reasonably easy to implement, and therefore may be attractive for practical purposes. Neural networks are indeed easy to use, but the structure of backpropagation neural network, which is the most commonly employed structure in water resources, is essentially that of an empirical model. Such a structure does not provide an exact replication of the behavior of the system under study and does not easily provide insights that would help give a better understanding of the system for the development of more physically-based models. Fuzzy sets are descriptive tools designed to establish knowledge domains. They may be used to link the various behaviors of a given system to observations or measurements performed on this system.

This description or mapping of the knowledge domain may then be useful for the development of physically-based models. Fuzzy logic is a predictive tool that is designed to provide a global estimate of the behavior of the system given the knowledge domain description established with fuzzy sets. The references cited above on the subject of modeling using fuzzy logic are studies that employ fuzzy sets and fuzzy logic on systems or parts of systems for which the knowledge domains are known or reasonably well known. One possible avenue of interest would be to develop adaptations of fuzzy sets and fuzzy logic so that they can be used on systems or parts of systems for which the knowledge domains are not as well known. Finally, it must be noted that AITs that are used as models are also subject to model structure and parameter uncertainties, as described in Sections 2.3.2 and 2.3.3.

It must be noted that model structure and parameter uncertainties are strongly related to the data. Data relevance is a quality attribute that affects the structure of the model. The common practice is to develop the structure of the model based on the data available, and therefore the data must adequately represent the behavior of the system the model is meant to replicate. Data completeness is a quality attribute that particularly affects the parameters of the model. To guarantee that the values of the parameters fairly represent the system under study, the data set employed to determine these values through calibration must contain all the possible patterns of behavior exhibited by the system. The methods commonly employed to address uncertainties are considered to be "after the fact" methods, for they only focus on reducing the potential deficiencies of the models and do not usually provide insights into understanding these deficiencies. Because model uncertainties can be directly related to data quality, the ideal approach would be to develop descriptive tools that help determine the value of the available data with respect to a specific quality attribute. Relevance and completeness are attributes to keep in mind for models, and data accuracy must also be considered. Indeed, measurement errors in data may introduce a bias that affects the evaluation of data relevance and completeness.

2.4 Description of Data Domains

2.4.1 Context

Throughout this thesis it is assumed that the quality of data sets affects model results. Erroneous data may propagate through a model and therefore induce a bias in the output. Also models are structurally developed and calibrated based on the data that are available. Therefore the validity of a model is ruled by the validity of the data available to represent the behavior of the system the model is meant to replicate. This work is directed toward developing tools that may be used to determine the characteristics of the data before they are used in the various simulation models. Figure 2.5, which is a modification of Figure 2.1, illustrates the process, where the inputs are pre-processed in order to evaluate their characteristics before they are used in a model.

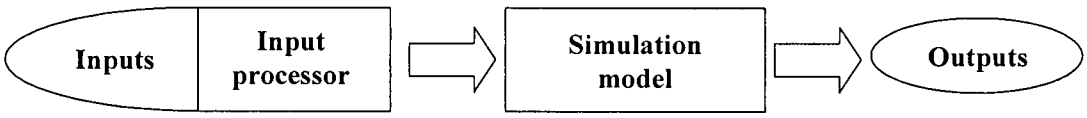


Figure 2.5. Typical model structure with input processor added.

Here, the characteristics of the data are determined by identifying specific patterns within the domain covered by the data sets, for example, patterns of erroneous data versus patterns of correct data, or patterns that are linked to particular points of the data and that correspond to very specific responses of the system under study. This is what is referred to as describing the data domain. A simple illustration is given in Figure 2.6, where the domain of two data series (axes X and Y in Figure 2.6) is subdivided, presumably with respect to the response or behavior of some system.

On one hand, the obvious task of the input processor would be to improve the accuracy attribute of the database by identifying erroneous data so that they are withdrawn or corrected before being fed to the model. On the other hand, the input processor could assist in the building of a model or in the development of modeling strategies. For example, the data might exhibit patterns that correspond to very specific responses of some water resources system, thereby indicating the need to develop more than one model. This implies a specific model for each targeted response, instead of only one model that is expected to

adequately represent all possible responses of the system. Incidentally, with some analysis, the input processor could also help determine the limitations of the models, that is, some system responses may not be adequately estimated with the model(s) available. It does not address data relevance or completeness, but it can offer an indication of where one stands with regard to these two data quality attributes.

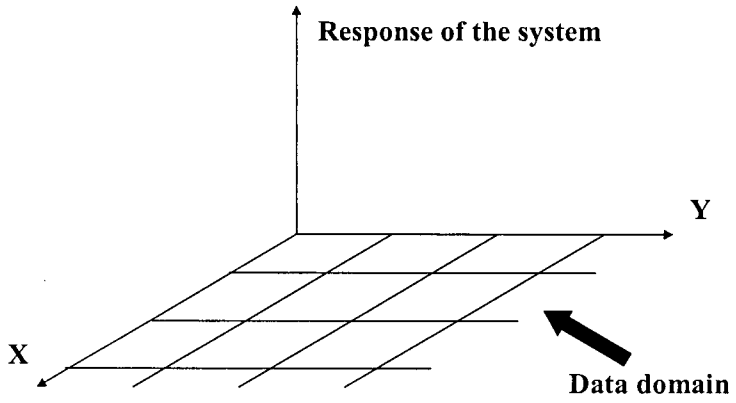


Figure 2.6. Subdivision of the data domain with respect to the response of the system.

The development of tools that describe the data domain for the purpose of finding errors or anomalies in data such as shifts, trends and outliers is a primary contribution of this thesis (Chapters 4, 5 and 6). Some developments are also achieved for the description of the domain of model parameters, which are the other type of inputs fed to the model aside from the data (Chapter 7). AITs are the basis of these developments, and the reasons for choosing these techniques are given in Section 2.4.3 and Chapter 3. A review of a few descriptive tools already employed in water resources is given in Section 2.4.2.

2.4.2 Descriptive Tools in Water Resources

In water resources, probability and statistics have been used extensively to describe data. The methods presented in Section 2.2, addressing the issues of data quality in general and outliers, shifts and trends in particular, have their roots in probability and statistics. In order to complement the review in Section 2.2, a few words are needed about the other techniques that can be useful for the description of data with respect to the patterns they

exhibit: multivariate statistical analysis tools. Analysis of variance is the data description tool used for the development of multiple linear regression models, which are frequently employed in all domains of water resources. The determination of homogeneous hydrologic regions is an activity that has employed description tools other than analysis of variance. For example, Ribeiro et al. (1995) make use of canonical correlations for the determination of homogeneous hydrologic regions, while Birikundavyi et al. (1993) employ correspondence analysis, which is a specific case of principal components analysis. In water resources, Principal Components Analysis has also been used for the precipitation fields (Siew-Yan-Yu et al., 1998), the interpretation of groundwater hydrographs (Winter et al., 2000), and the study of snow parameters from remote sensing data (Derksen et al., 2000), among others. The point in common among these methods, including analysis of variance, is that they assume that there exists an underlying structure between the variables. The goal of this work is to avoid making any inference about the underlying structure between the variables, except in the case of the presence of outliers shifts or trends, and therefore the methods given above are not considered. As mentioned by Dillon and Goldstein (1984), it is the common practice to consider such techniques as sensitive to data anomalies, and therefore not suitable for anomalous data. This statement is however given in the context of when one is not looking for data anomalies. Theoretically, there is no contraindication of the use of multivariate statistical tools purposely for the detection of anomalies such as outliers, shifts and trends. For example, application of Principal Components Analysis or other related factorial analyses for the detection of shifts is performed in Balcerowska et al. (2000) and Kruszewski et al. (2003).

Cluster analysis does not make use of correlations, but rather establishes groups (or patterns) based on the distances between sequences of data. Typical applications in water resources of cluster analysis can be found in Burn (1990) and other subsequent work such as Zrinji and Burn (1994), or Burn (1997), who use it to develop the concept of "*Region of Influence*" for the determination of hydrologic homogeneous regions. Dillon and Goldstein (1984) mention that cluster analysis is sensitive to outliers, although this argument is based on application of the technique to cases where errors in the data are not expected. The work in this thesis shows that AITs can be structured to specifically detect errors and may fulfill the same purpose as cluster analysis. Cluster analysis can also be structured for the

detection of errors. Its disadvantage compared with AITs is that it is computer-intensive in the context of the applications presented in this thesis.

2.4.3 Artificial Intelligence

The references cited in Section 2.3.5 essentially present AITs as modeling tools, although they can also be employed in the description of input domains (i.e., data and parameters). Fuzzy sets are strictly descriptive tools. Related to fuzzy sets is a tool call fuzzy c-means. It is a clustering technique, performing the same function as the statistical cluster analysis while being less computer-intensive for cases involving large data sets. An application of this technique in water resources can be found in Hall and Minns (1999) for the determination of homogeneous hydrologic regions. The backpropagation neural network is not descriptive in nature, but the structure of this network involves discretizing the data domain, as shown in Figure 2.6. Another form of neural network, the Kohonen network, is a strictly descriptive tool. It is a clustering technique, and like fuzzy c-means, it also performs the same function as the statistical cluster analysis while being less computer-intensive for cases involving large data sets. Such a network has rarely been used in water resources so far. Examples of its use can be found in Liong et al. (2000) for the classification of watershed conditions, Hall and Minns (1999) for the determination of hydrologic homogeneous regions, Gotz et al. (1998) for the identification of river pollutant sources, and Bowden et al. (2002) for the study of algae blooms.

Probability and statistics have been the basis of most developments performed for evaluating data quality and model structure and parameter uncertainties. This thesis proposes developments based on AITs as alternatives to probability and statistics in order to address issues such as data quality and uncertainties. Patterns in inputs (data and model parameters) are the subject of this study, and AITs can be structured so as to identify them. Descriptive AITs can be considered as performing at least as well as probability and statistical tools employed for the description of data. Their flexibility and ease of use may be deemed as attractive to researchers and practitioners in water resources. In some instances, the reduced computational burden of AITs compared with that of probability and statistical tools is a definitive asset. In the next chapter, a general description of the artificial intelligence technique employed in this thesis is given.

Chapter 3

General Description of Artificial Intelligence Techniques

3.1 Introduction

AITs are presented here as alternatives to statistical and probabilistic techniques for addressing issues of data quality and model and parameter uncertainties. Statistics and probability may be very acceptable for resolving these issues, and artificial intelligence techniques should be considered as simply an alternative way of looking at such problems. The goal of this thesis is to compare these two general approaches so as to foster new developments in both types of techniques for a better understanding of environmental and hydrologic systems.

AITs have the advantages in that they exhibit applicability, adaptability, implementability, and input flexibility. Applicability here means that they can be applied to a wide range of situations or problems. For any given purpose for which probability and statistics are applicable, the likelihood that tools based on artificial intelligence will be applicable for the same purpose is high. Adaptability refers to the possibility of employing AITs as stand-alone tools, which is demonstrated in Chapters 4, 5 and 6 of this thesis. Here, the Kohonen neural network and fuzzy c-means are applied for the study of data quality. AITs can also be integrated in a larger mathematical scheme, which is illustrated in Chapter 7, where fuzzy logic is incorporated in simulation models for the determination of the values of model parameters with the aim of reducing parameter uncertainties. Implementability indicates that it is relatively easy to build artificial intelligence tools. The mathematics behind basic artificial intelligence techniques can be fairly easy to grasp, and many software products based on artificial intelligence have been developed and made available on the market. Artificial intelligence might also be easier to present to a wide audience. The tasks of practitioners and researchers include the dissemination of their work, in environmental impact studies for example, to a public audience, and this would include the description of the methods employed in work. Artificial intelligence techniques are based on specific features of the brain, that is, the structure (neural networks) of the brain and its ability to describe information (fuzzy sets). The latter feature, in particular, is a

concept that is widely understood by everybody, which might not be the case for statistics. In brief, if one feels more comfortable about the methods one employs for a study, then the results of this study might be better understood if not better accepted. Finally, input flexibility refers to the possibility for employing artificial intelligence with hard data such as measured observations, combined with soft data such as observations or judgment that cannot be quantified exactly. A simple example is the case of describing and predicting floods in a river, which are difficult to measure with stream gauges as they can be damaged or destroyed in floods, but can be qualified with reference to the field information such as marks on the facades of houses or tree trunks or the presence of water at specific locations on the floodplain. Floods can then be characterized as big, medium or small, depending on the height of the marks on the trunks or houses or on the extent of water in the floodplain.

Of particular interest is the capacity of artificial intelligence techniques for use as descriptive tools, as mentioned in Section 2.4. They can subdivide data domains with respect to features or patterns present in the data, as illustrated simply in Figure 2.6. Specific patterns can then be identified (e.g., patterns of anomalous data versus patterns of normal data) and isolated if necessary. Specific modeling solutions can also be developed once the patterns are identified. Thus specific models may be designed for application to specific patterns, a strategy employed by Bowden et al. (2002) for modeling algae concentrations in a river, for example. The downfall of such data description tools is that they require much data in order to assure that all possible patterns are adequately represented.

In this chapter, a general description of the artificial intelligence techniques employed in this work is given, and followed with an overview of the applications that are treated in the thesis. The adaptations of these techniques for the applications in this thesis are developed in the subsequent chapters as appropriate.

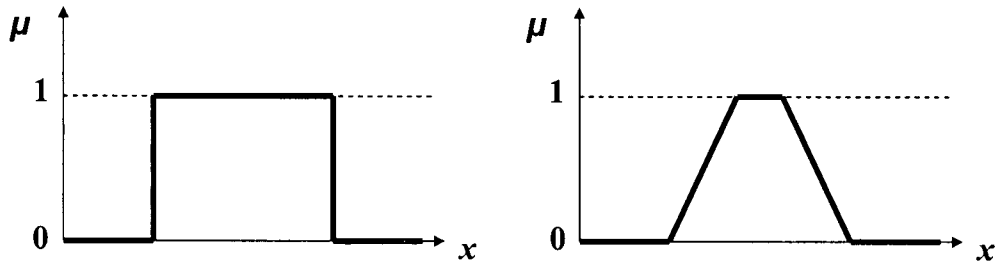
3.2 General Description of the Techniques

3.2.1 Fuzzy Logic

Details of this technique can be found in textbooks by Dubois and Prade (1980), Zimmermann (1991), Terano et al. (1992), and Bardossy and Duckstein (1995). Fuzzy logic is the process that links the input domain to the response of some given system. Given

a set of inputs, a likely response is given based on the characterization of the input domains through fuzzy sets and predefined rules that explain the relationship between inputs and responses. Thus, to build a procedure based on fuzzy logic requires the characterization of the input and response domains through fuzzy sets, and then the derivation of the rules of the system. Once all domains are characterized and the rules are defined, the response of a given set of inputs is established as a function of the degree of fulfillment of the rules that apply, the combination of those rules in order to get a fuzzy response, and a deffuzification procedure that turns the fuzzy response into a crisp value.

Fuzzy sets constitute a departure from classic sets on which all probability rules are based. With classic sets, as shown in Figure 3.1a, an element in the input domain (x) either belongs entirely to the set or it does not belong at all to the set. If it belongs entirely, the membership value (μ) attached to the element equals one. If it does not belong at all to the set, the membership value equals zero.



(a) Classic set **(b) Fuzzy set**

Figure 3.1. Classic set (a) versus fuzzy set (b).

A fuzzy set (Figure 3.1b) accepts that elements do not belong entirely to the set, and defines the membership value for these elements as being between zero and one. This flexibility given to the membership value offers a way to describe how people perceive things. A classic example that illustrates the advantage of such flexibility is the definition of the set of tall people. Let's assume that a person is considered to be tall definitely if he or she measures 1.85 m or more in height. Building a classic set on this premise means that all measurements equal to or greater than 1.85 m are given a membership value of one, while anything smaller is given a membership value of 0, including the measurement of 1.84$\dot{9}$ m.

With a fuzzy set, measurements equal to or greater than 1.85 m can be given a membership value of 1, and the smaller measurement may be given a membership value other than zero. For example, a measurement of 1.80 m can still be considered as fairly tall and be given a membership value of 0.75. In fact, the membership value can decrease linearly from 1.85 m, and reach zero at say 1.75 m. Of course, such fuzziness in membership for a given set or indicator can be observed with natural phenomena. For example, a flood can be considered large, to some extent, or fairly small. A drought can be assumed as very severe to relatively mild, and an algae bloom can be viewed as very large or very small. The input domain (e.g., water inflow measurements) can be entirely characterized by fuzzy sets, which, as shown in Figure 3.2, represent different quality: very small, small, medium, large, and very large.

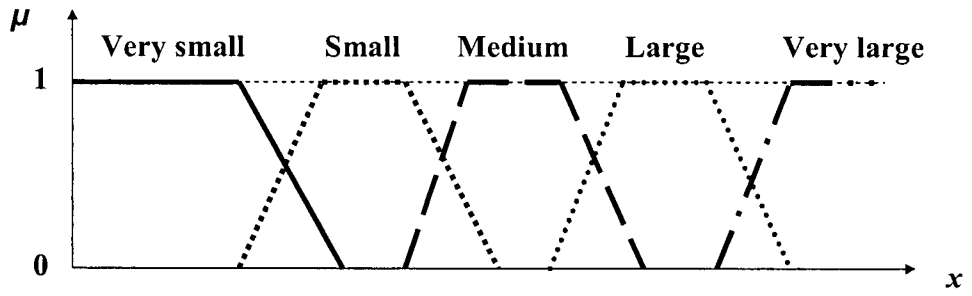


Figure 3.2. Characterization of the input domain with fuzzy sets.

The operators employed with a classic set, such as NOT, AND and OR can also be applied with fuzzy sets. Let's consider three fuzzy sets: A, B, and C, with their respective membership functions $\mu_A(x)$, $\mu_B(x)$ and $\mu_C(x)$ over the input domain x, then for the complement (i.e., NOT A or A'):

$$\mu_{A'}(x) = 1 - \mu_A(x) \tag{3.1}$$

for the intersection or AND operator $(A \cap B)$:

$$\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)) \tag{3.2}$$

and for the union or OR operator $(A \cup B)$:

$$\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)) \tag{3.3}$$

Also, the operator properties are respected with fuzzy sets, For example:

$$(A \cap B) \cap C = A \cap (B \cap C) \quad 3.4$$

and:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad 3.5$$

The only exceptions are the following properties, which are not true for fuzzy sets:

$$A \cap A' = \phi \text{ (void)} \quad 3.6$$

and:

$$A \cup A' = U \text{ (the entire domain or universe)} \quad 3.7$$

The construction of rules, which link the input domain to the response of the system, are based on the operators AND and OR. Let's consider the fuzzy sets A, B, and C applying on variables *a*, *b* and *c*, respectively. Then a rule can be: IF *a* is A AND/OR *b* is B AND/OR c is C, THEN D, where D is the fuzzy set of the response of the system applying to that rule. Bardossy and Duckstein (1995) describes several simple procedures for the construction of rules. Chapter 7 in this thesis provides another construction procedure, based on the optimization of the parameters of a simulation model, and is suitable for the application cases presented in that chapter. For a given set of inputs, more than one rule usually applies. This is the consequence of the fact that fuzzy sets in the input domain are overlapping, as shown in Figure 3.3. If variables *a*, *b*, and *c* had the values indicated by their respective arrow, this means that all the rules built with fuzzy sets A_1 and A_2 for variable *a*, B_2 and B_3 for variable *b*, and C_1 and C_2 for variable *c* would apply. The purpose in fuzzy logic is to determine the rules that apply in a given situation, to give each of these rules a weight often called a degree of fulfillment of the rule, and to combine the response of these rules with respect to their degree of fulfillment. The combination of responses from the rules gives a global response for the system under study, this global response being as a consequence a fuzzy set. A crisp response, that is a single value, can then be determined from the global fuzzy response through the use of a defuzzification procedure, which is the equivalent of getting an estimate such as the mean or the median from the probability distribution of the response of the system.

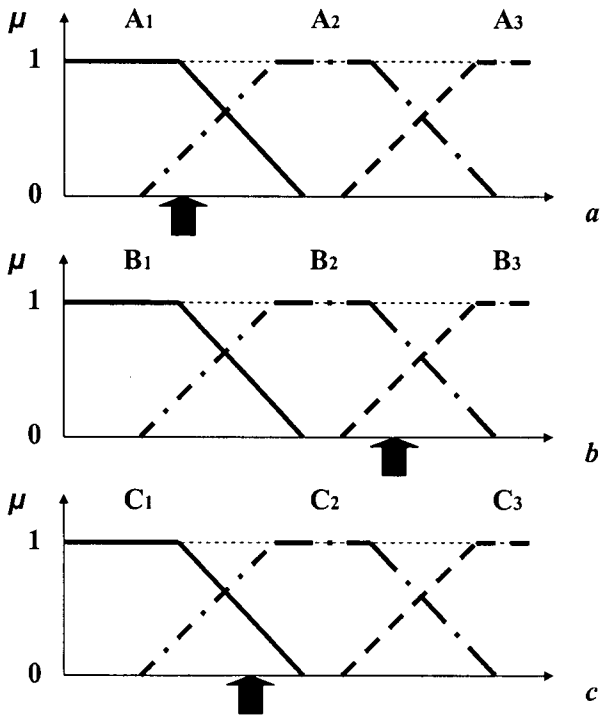


Figure 3.3. Characterization of several variables with fuzzy sets.

The degree of fulfillment of a rule (v) is determined with respect to the operators employed in the rules, either AND or OR. The most common way to determine v is through product inference relationships:

$$v(\mathrm{A}_1 \text{ AND } \mathrm{A}_2) = \mu_{\mathrm{A}_1}(a_1)\mu_{\mathrm{A}_2}(a_2) \qquad 3.8$$

and:

$$v(\mathrm{A}_1 \text{ OR } \mathrm{A}_2) = \mu_{\mathrm{A}_1}(a_1) + \mu_{\mathrm{A}_2}(a_2) - \mu_{\mathrm{A}_1}(a_1)\mu_{\mathrm{A}_2}(a_2) \qquad 3.9$$

or through min and max inference relationships:

$$v(\mathrm{A}_1 \text{ AND } \mathrm{A}_2) = \min\left(\mu_{\mathrm{A}_1}(a_1), \mu_{\mathrm{A}_2}(a_2)\right) \qquad 3.10$$

and;

$$v(\mathrm{A}_1 \text{ OR } \mathrm{A}_2) = \max\left(\mu_{\mathrm{A}_1}(a_1), \mu_{\mathrm{A}_2}(a_2)\right) \qquad 3.11$$

At times, when the structure of the system is not well known, and it is presumed that both AND and OR operator can apply to a given rule, then the degree of fulfillment can be established as the combination of the degrees of fulfillment coming from the use of each operator:

$$\nu = \gamma\nu(\mathrm{OR}) + (1-\gamma)\nu(\mathrm{AND}) \tag{3.12}$$

where γ is the weight given the relative importance of both operators.

The combination of the response of the rules involves determining the membership value (μ_D)of the global fuzzy response, over the response domain d. There are many ways of combining the rules, and they can be divided into two types: min/max combinations, and additive combinations. In this thesis, the additive combination referred to as normed weighted sum combination is employed, which is considered from experience as satisfactory by Bardossy and Duckstein (1995) for application to environmental systems. This combination method determines the membership value of the global fuzzy response as follow:

$$\mu_D(d) = \frac{\sum_{i=1}^{I} \nu_i \beta_i \mu_{D_i}(d)}{\max_u \sum_{i=1}^{I} \nu_i \beta_i \mu_{D_i}(u)} \tag{3.13}$$

where:

$$\frac{1}{\beta_i} = \int_{-\infty}^{+\infty} \mu_{D_i}(d)\mathrm{d}d \tag{3.14}$$

In Equations 3.13 and 3.14, ν_i is the degree of fulfillment of rule i, D_i is the fuzzy response for rule i, and I is the total number of rules that apply for the considered set of inputs. The fraction ensures that the membership value for the global fuzzy response is not greater than 1. The advantage of this combination method is that it can lead to a relatively easy defuzzification process.

In the defuzzification process, the mode ($\max(\mu_D(d))$), mean or median of the global fuzzy response is sought. The value of the mean ($m(D)$) when the normed weighted sum combination is employed is determined as follows:

$$m(D) = \frac{\sum_{i=1}^{I} \nu_i m(D_i)}{\sum_{i=1}^{I} \nu_i} \tag{3.15}$$

3.2.2 Fuzzy C-means

Before describing fuzzy c-means in great detail, clustering techniques in general and the conventional statistical clustering techniques in particular need to be introduced. Cluster techniques are used to reduce the amount of data collected to a form that can be more easily interpreted. The goal of such techniques is to obtain a smaller number of groups or clusters such that data sequences located in each cluster demonstrate a certain degree of similarity with each other. This means that the determined clusters should display small, within-cluster variations, as illustrated simply in Figure 3.4.

A clear description of the conventional statistical clustering technique can be found in Dillon and Goldstein (1984). Let's consider N sequences of data, each sequence having K elements or variables. The clustering analysis starts with building an $N \times N$ matrix, which contains elements that defines the similarity between data sequences. Generally, the measure of similarity considered is the Euclidian distance between data sequences. Data points are then assigned to C clusters according to the similarity measures through clustering techniques, which can be divided into two categories: hierarchical and partitioning. The agglomerative hierarchical techniques assume that each sequence starts with its own cluster, then the two closest sequences are fused together to form one cluster, and the process of fusing clusters, either individual sequences or groups of sequences, continues until all sequences are gathered into one cluster. The divisive hierarchical techniques perform the reverse process. They start with considering all the sequences in one cluster, and then divide the sequences into smaller groups repeatedly until each sequence is in its own cluster. The hierarchical techniques produce a clustering tree that gives the history of agglomeration or division of sequences, and the final number of clusters is determined by where the tree is cut off and by counting the number of branches that have been cut. Partitioning techniques are a refinement of the hierarchical techniques, where the sequences can move from one cluster to another so as to optimize some given criterion, for example, minimizing the sum of the distances between cluster centers and their respective sequence members. Partitioning techniques require that the number of clusters be known or specified.

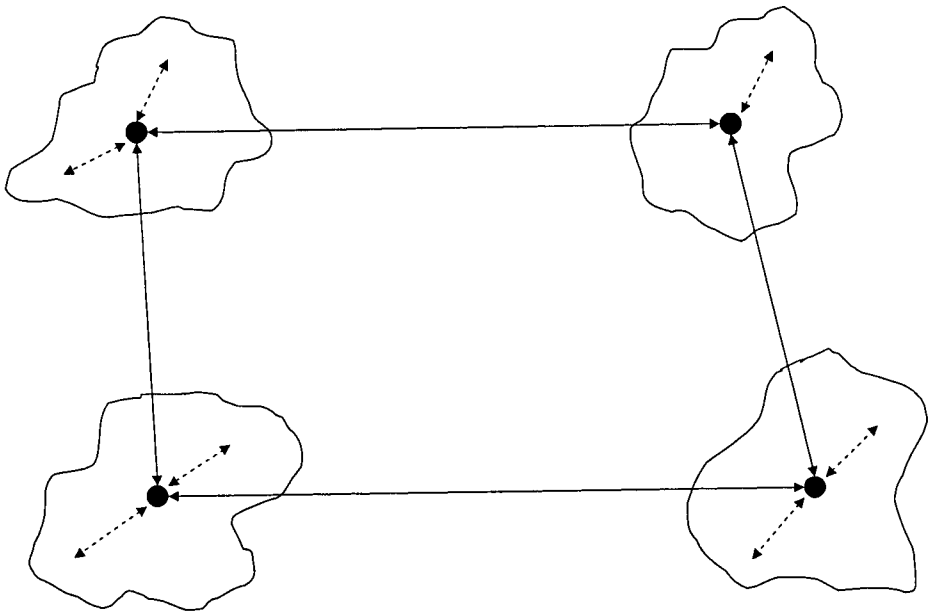


Figure 3.4. Distances between clusters versus distances within cluster.

These conventional clustering techniques can be considered as "hard" because they impose that each of the data sequences belongs entirely to one cluster. Fuzzy c-means is a partitioning clustering technique, but should be considered as a "soft" technique, because it allows the sequences during the optimization process to belong to some degree to more than one cluster, according to a membership value. Fuzzy c-means is simply an extension of the concept of fuzzy sets for the purpose of clustering data, and further details on this technique can be found in Bezdek (1981) or Hall and Minns (1999). Let's consider again N sequences of data, each sequence having K elements or variables. Fuzzy c-means will partition these data points into C clusters, where $2 \leq C \leq K$. The clustering algorithm used in fuzzy c-means is based on minimization of the following objective function:

$$F = \sum_{c=1}^{C} \sum_{n=1}^{N} \mu_{c,n}^{r} d_{c,n}^{2} \qquad 3.16$$

where the μ's denote the membership grade for every data point, the d's denote the Euclidean distance between sequence n and a cluster center c, and r denotes the weighting parameter controlling the amount of fuzziness in the process of classification. The membership values indicate the extent to which the sequences belong to the clusters, and are constrained as follows:

$$\sum_{c=1}^{C} \mu_{c,n} = 1 \text{ for all } n \qquad 3.17$$

and:

$$0 < \sum_{n=1}^{N} \mu_{c,n} < N \text{ for all } c \qquad 3.18$$

The Euclidean distance between a sequence and a cluster center is defined as follows:

$$d_{c,n} = \sqrt{\sum_{k=1}^{K} \left(x_{k,n} - z_{k,c}\right)^2} \qquad 3.19$$

where $x_{k,n}$ is the kth element of sequence n, $z_{k,n}$ is the kth element of cluster c. The weighting parameter, r, is equal to 1 when the clustering is hard, and the clustering becomes softer or fuzzier as r increases from 1. For fuzzy clustering, the value of r used is generally $1.25 \leq r \leq 2$ (Hall and Minns, 1999).

Fuzzy c-means clustering is carried out through an iterative optimization of the objective function (F). It starts with an assigned number of clusters (C) and weighting parameter (r). The cluster centers are initially located in the mass center of all data sequences. With the set of initial cluster centers, every data sequence is assigned an initial and equal membership grade, μ, for each cluster. The μ's are collected in the partition matrix, $U^{(p)}$, which is a $C \times K$ matrix where p denotes the number of iterations. The cluster centers and membership grades for each data sequence are upgraded through an iterative process. This process gradually moves the cluster centers to better sets of cluster centers, and the iteration stops when the difference in the updated U and the previous U is less than a prescribed limit, e. The summary of the process is as follow:

1. Select a number of clusters (C) and a weighting parameter (r), and initialize the partition matrix, $U^{(0)}$,
2. Compute a new set of cluster centers, z:

$$z_{k,c} = \frac{\sum_{n=1}^{N} \mu_{c,n}^{r} x_{k,n}}{\sum_{n=1}^{N} \mu_{c,n}^{r}} \quad \text{for } k = 1, ..., K \text{ and } c = 1, ..., C \qquad 3.20$$

3. Update the elements of the partition matrix as follows:

$$\mu_{c,n}^{(p+1)} = \left(\sum_{j=1}^{C} \left(\frac{d_{c,n}^{(p)}}{d_{j,n}^{(p)}} \right)^{2/(r-1)} \right)^{-1} \qquad 3.21$$

4. Repeat the second and third step, and terminate when $U^{(p+1)}$ does not differ from $U^{(p)}$ by more than e.

Fuzzy c-means is a clustering tool that is relatively easy to employ. Of course, since it involves an optimization procedure, there is always a possibility of being trapped in a local optimum. This places uncertainties on the validity of the calibrated clusters. This problem is addressed in Chapter 5. The advantage of fuzzy c-means clustering, compared with the conventional clustering techniques, is that it may require less computing capabilities. The conventional clustering techniques involve the computation of the distances between sequences, yielding an $N \times N$ matrix. Because the matrix is symmetric and the elements in the diagonal are not used, the number of required elements can be reduced to $(N^2 - N)/2$. This is the largest element load for this technique. All the other elements required by the techniques are considered to be negligible quantities. A case with 10,000 data sequences, which is on the order of magnitude of the number of sequences employed in the applications in Chapter 5, would yield around 50,000,000 elements, therefore requiring a computer memory of about 400 Mb. Many personal computers nowadays would have difficulty to supply such a memory requirement, even assuming the element load is not duplicated for temporary variables for the procedures used by the clustering software. And this is a memory requirement that applies regardless of the desired number of clusters, whether it is 1 or 10,000. On the other hand, the largest requirement of the fuzzy c-means clustering is due to the matrices containing the distances (z) and the membership values (U). Both are $C \times N$ matrices, yielding a total of $2NC$ elements. To achieve equivalent memory load with both the conventional clustering techniques and fuzzy c-means approach, this would means that $(N^2 - N)/2 = 2NC$, or that C be equal to $(N - 1)/4$. Using the same example of 10,000 sequences, this implies that C should be equal

to 2,500 in order for the fuzzy c-means to match the memory requirement of the conventional clustering technique.

3.2.3 Kohonen Neural Network

The Kohonen neural network, which is well described in Kohonen (1990 and 1997), is another clustering technique that can be used as an alternative to the conventional statistical clustering techniques. Neural networks attempt to replicate the structure of the brain, and are built with single units called perceptron as illustrated in Figure 3.5. A perceptron contains a neuron that processes the information received, a set of input synapses that feed the neuron with information from other neurons, and a set of output synapses which ship the product of the neuron processing towards other neurons. The most commonly used neural network structure is the backpropagation network, which is a series of layers of neurons. The raw information is received by the input layer, which basically plays the role of a sensor (e.g., eyes, ears, nose, tongue or skin), and is then transferred to the next layer through the synapses. This movement forward goes on from one layer to the next, until the processed information reaches the output layer, which provides the final answer to any given set of inputs fed to the input layer. Because of this structure, the backpropagation neural network can function as a simulation model, and has often been used as such in water resources application, as an alternative to other empirical or physically-based models.

The Kohonen network, on the other hand, is designed as a tool to analyze inputs. As illustrated in Figure 3.6, the Kohonen network is made of an input layer that receives the data and an output layer composed of several neurons often structured in a two dimensional plane that ultimately sorts the input information in some determined pattern. Given a set of inputs fed to the input layer, only one neuron on the output layer is activated, that is, it returns a value of 1, while all the other neurons on the output layer return a value of 0. Let's look at one neuron on the output layer, as illustrated in Figure 3.7, to see how it can be activated.

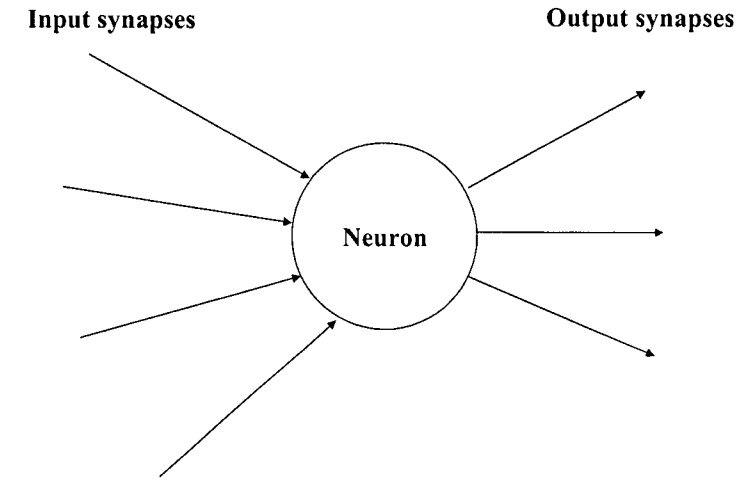


Figure 3.5. Schema of a perceptron.

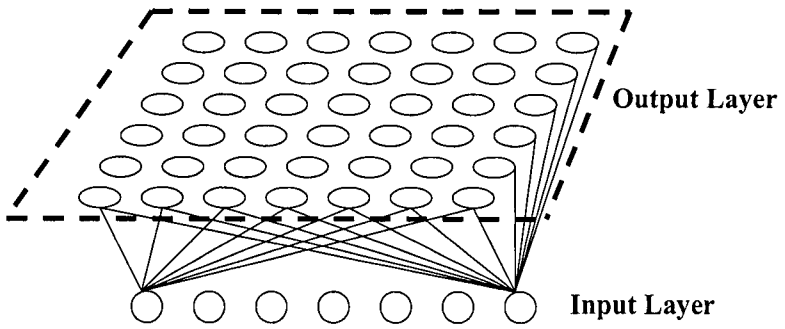


Figure 3.6. Structure of the Kohonen network.

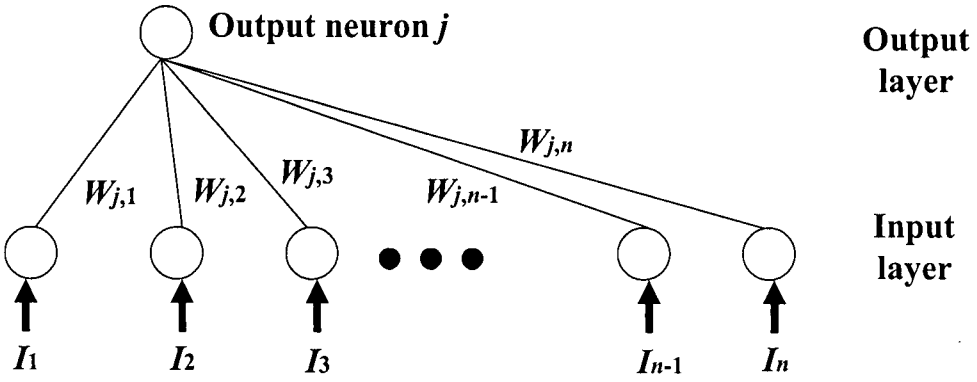


Figure 3.7. Connection between an output neuron and the inputs neurons.

As seen in Figure 3.7, each of the n neurons on the input layer receives a single value I_k, where $k = 1, ..., n$ (e.g., a variable or an element in a data sequence), and transfer the value to the output neuron j through the synapse. Each synapse is associated with a weight $W_{j,k}$, providing a vector of weights W_j for output neuron j. The output neuron calculates the distance D_j between its weight vector and the input vector through the usual relationship:

$$D_j = \sqrt{\sum_{k=1}^{n} \left(I_k - W_{j,k}\right)^2} \qquad 3.22$$

In this context, the weight vector can be considered as the mass center of the output neuron, and the output neuron itself can then be viewed as the center of a cluster. The neuron on the output layer that is activated given an input vector is the one that yields the shortest distance D_j.

The values of the elements in the weight vectors need to be calibrated so that the network covers the whole input domain. This calibration process tends to structure the output layer so that the input pattern can be defined in some meaningful coordinate system (Kohonen, 1990), which is why the Kohonen network is also called a self-organizing map. The most commonly employed types of maps are the grid map (rectangular shape) and the hexagonal map (hexagon shape). The calibration is an iterative process, where one input vector is fed to the network at every iteration. The weight vector of the activated neuron on the output layer is fully updated at every iteration, while the weight vectors of the neighboring output neurons are updated to some extent. A general formula for the updating of the weight vector (W_j), at iteration t, following the feeding of input vector I is:

$$W_j^{(t)} = W_j^{(t-1)} + h_j\left(I - W_j^{(t-1)}\right) \qquad 3.23$$

This formulation simply drives the weight vector to be closer to the input vector, and the weight vector is equal to the input vector if h equals 1. The parameter h determines the amplitude of the updating. For the study of data involving natural phenomena, one appropriate function of h would be (Kohonen, 1991):

$$h_j = h_0 \exp\left(-\left(d_{j,a} / \sigma\right)^2\right) \qquad 3.24$$

In this expression, $d_{j,a}$ is the distance between the activated neuron (a) and another neuron j as determined on the output map (layer). When $j = a$, the exponential equals 1 and the value of h_j is at its maximum value (h_0). The value of h_j decreases as the distance between

activated neuron a and neuron j increases. Parameter h_0 gives the magnitude of the updating. It has a high value at the beginning of the calibration process so as to ensure a rapid spreading of the output neuron over the input domain, and is reduced at every iterative step so that only small adjustments are performed on the weight vector at the end of the calibration process. Parameter σ is a scaling factor on the distance, and indicates the extent of the output map affected by the updating. It is set at a high value at the beginning of the calibration process so that a large neighborhood or number of output neurons are significantly updated, and its value is decreased at every iterative step so that only a small neighborhood or number of output neurons is significantly updated, leading to only small refinements at the end of the calibration process.

There are several ways to express h_j aside from Equation 3.24. Another common way is to employ a step function, which imposes adjustments of the weight vectors of the neurons that are within a specific distance from the activated neuron, and no adjustment on the weight vectors of the neurons that are beyond this specific distance from the activated neuron. The measure of distance can also vary. The one employed in Equation 3.24 is the Euclidian distance, but the Manhattan distance can also be an option for grid maps, and the distance based on the number of links that separate neurons from each other can be employed on hexagonal maps. Existing software automates the calibration processes for the Kohonen network, and usually starts the calibration with the weight vectors of all the output neurons concentrated on the mass center of the data domain. The calibration is performed using a large set of input vectors, which are fed randomly to the network at the rate of one input vector per iteration. A large number of iterations ensures that all input vectors are employed a significant number of times on the average at all times of the calibration process (i.e., from the rapid spreading to the refinements of the map). Visually, with this kind of calibration process, the output layer spreads out over the input domain. A good example of this visualization is presented in Figure 3 in Kohonen (1990, p. 1468), where the results of the calibration process are shown in six 6 different steps. There, the output layer stretches gradually from one calibration step to the next over a two-dimensional input domain (two variables or two elements in the input vector). On a one dimensional input domain, the output layer would simply stretch in the ascending and descending directions of the data. With more than two elements or variables in the input

vectors, the behavior of the calibration is less predictable. With n elements in the input vectors and n dimension in the output layer, the calibration simply sorts the data in ascending or descending order in every dimension. When the dimensions of the output layer are smaller than the dimensions of the input vector, then the calibration process can spread the network over the input domain in an infinite number of ways. The greatest advantage of the Kohonen network is that it can reduce the dimension of a given problem so as to provide an easier grasp of the structure of the data. Indeed, using a two-dimensional output layer map in order to classify an n-dimension input domain ($n > 2$) facilitates the visual interpretation of the patterns or structure of the data. Now, with the prospect that the calibration process can spread the network in so many different ways, there are possibilities that some parts of the input domains or some particular patterns in the data may be overlooked or not adequately represented in the final network. This results in uncertainty about the reliability of the network similar to that about the validity of the calibrated clusters with fuzzy c-means. As mentioned in Section 3.2.2, this problem is addressed in Chapter 5.

Like fuzzy c-means clustering, the Kohonen network approach demands less computer capability than the conventional clustering techniques. In fact, the largest memory requirement comes from the matrix that stores all the weight vectors, a matrix that is required also by the conventional clustering technique and fuzzy c-means approaches. In brief, the Kohonen network has the smallest memory requirement of all other clustering techniques presented. Its disadvantage, as observed by experience in the applications in this thesis, is that achievement of a reasonably good calibration of the weight vectors is computationally intensive. When memory is not an object, the calibration of the Kohonen network takes the longest time compared with the other techniques.

3.3 Applications of AITs

Here, a brief overview and the common points of the application cases investigated in the subsequent chapters are given. The more critical details on how the AITs are structured or adapted to respond to the needs of the various applications are provided in the appropriate chapters as needed. Chapters 4 to 6 address the problem of data accuracy by the development and application of detection methods for the identification of anomalies in the

data. Three type of anomalies are investigated, that is 1) outliers, which are individual data having statistical properties that differ from those of the overall population; 2) shifts, which are sudden changes over time in the statistical properties of the historical records of data; and 3) trends, which are systematic changes over time in the statistical properties. Figure 3.8 illustrates examples of these anomalies.

Chapters 4 and 5, each respectively provides a variant of the use of fuzzy c-means and the Kohonen network for the detection of shifts and trends. These variants are compared to conventional statistical tests of detection for an evaluation of their performance. The work in Chapters 4 and 5 can be viewed as an exercise in determining the reliability of detecting shifts and trends of both the methods proposed here and of the conventional tests that have been used typically. In Chapter 5, the applications of the Kohonen network and fuzzy c-means are also applied for the detection of outliers. In these two chapters synthetic data established to represent hydrometric data observed in Canada are employed. In Chapter 6, real data are used to demonstrate the applicability of these methods.

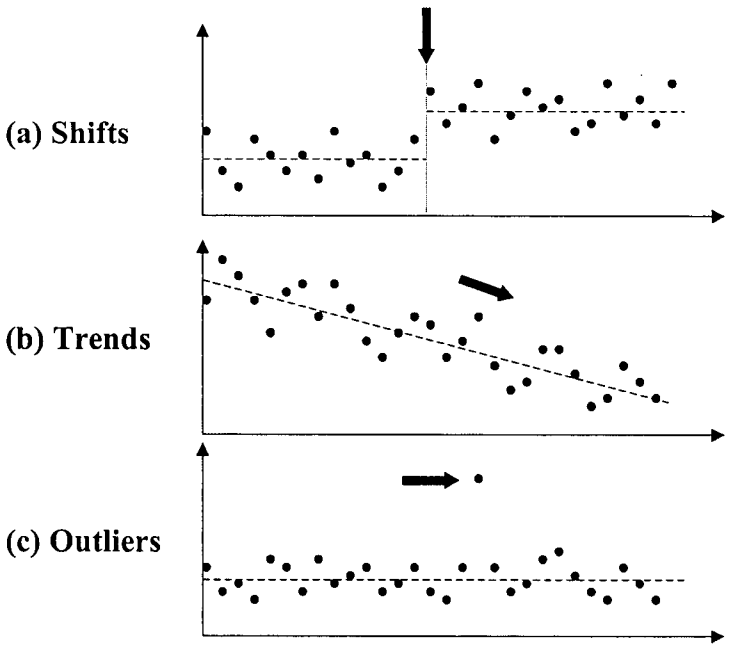


Figure 3.8. Description of anomalies: (a) shifts, (b) trends and (c) outliers.

Data accuracy is the major issue of this thesis, and the goal is to develop methods that allow a description of the input (i.e., data in this case) domain in such a way that anomalous patterns in the data, generated by the presence of outliers, shifts or trends, can be identified. A secondary effort is also dedicated to the issue of another source of uncertainty in simulation modeling, that is, the model parameters, which are another type of input fed to the model. In Chapter 7, fuzzy logic is employed to map the value of some of the parameters of models with respect to relevant indicators of the system under study. From this mapping, specific values of the parameters can then be determined with respect to the system conditions in order to obtain improved model estimates compared with those from the model version where the parameter values remain constant regardless of the condition of the system. In Chapter 7, this parameter mapping is employed on a watershed model to estimate water inflow and on an algae growth model to estimate algae concentrations in a river.

Chapter 4

Classification Procedures for Detecting Anomalies

4.1 Introduction

This chapter proposes one approach for detecting data anomalies, more precisely shifts and trends. The evaluation of the approach is achieved with an experiment that employs synthetic data designed to replicate hydrometric observations. Such data are regularly considered in studies for the detection of shifts and trends. Because this approach cannot be applied for the detection of outliers, this type of anomaly is not considered here. A second approach, which is presented in Chapter 5, addresses all anomalies of interest in this thesis.

Hydrometric measurement sequences can be subject to shifts and trends due to the effects of anthropogenic and natural changes on water inflow regimes in watersheds over time. In the planning and management of water resources systems, it is important to be aware of such patterns in order to estimate water availability as accurately as possible. Recently, several extensive studies, such as those of Anderson et al. (1992), Yulianti and Burn (1998), and Zhang et al. (2001) in Canada and Lettenmaier et al. (1994) and Lins and Slack (1999) in the United States, have been undertaken with the goal of determining whether trends, and in some cases shifts, are present in historical records of streamflow measurements. These studies provide a wealth of information since many stations over large territories are assessed. However, the authors of these studies imply that the validity of the results depends greatly on the accuracy of the statistical tests used for the detection of shifts and trends. The statistical tests also impose restrictions, for example, that data be independent, or that data or results of the tests follow a specific distribution, often the Normal distribution. Independence of the data is necessary, but the imposition of a distribution may lead to a bias in the results if the distribution is not adequately representative of reality. The application of AITs for pattern recognition does not require any assumption regarding the distribution of the data or the test results.

This chapter first describes the conventional statistical detection tests in some detail, and develops approaches for using AITs to detect shifts and trends. In the following

section, a description of the experimental design and database employed for the evaluation and the comparison of the performance of all the methods involved, both conventional and artificial intelligence based, is provided. This database is comprised of univariate cases, for which the conventional detection tests are commonly applied, as well as multivariate cases. Examples of multivariate cases are sets of streamflow data from several gauging stations or time sequences of different variates such as streamflows and stream temperatures. A review and discussion of the experimental results constitute the last section of this chapter.

4.2 Conventional Detection Tests

4.2.1 Shifts

The Student's and the Mann-Whitney tests can be used for detecting shifts. Given a sequence of individuals x_t, $t = 1, \ldots, N$, that is divided into two continuous sub-sequences of size n_1 and n_2 ($n_1+n_2 = N$), both tests assume a null hypothesis stating that both sub-sequences come from the same population. The Student's test determines the absolute value of the standardized difference of the means of the sub-sequences (T), as indicated in Equation 4.1.

$$T = \frac{\left|\bar{x}_2 - \bar{x}_1\right|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad 4.1$$

with:

$$S = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{N - 1}} \qquad 4.2$$

where $\bar{x}_1$ and s_1 are the mean and standard deviation of sub-sequence 1, prior to the shift, and $\bar{x}_2$ and s_2 are the mean and standard deviation of sub-sequence 2, posterior to the shift. The null hypothesis is T is greater than $T_{1-\alpha/2,\nu}$, which is the $1-\alpha/2$ quantile of the Student's distribution, with $\nu = N-2$ degrees of freedom and α as the significance level of the test. For the Mann-Whitney test, the entire sequence is sorted in ascending order, and the mean of the position in the sorted sequence of the individuals of the first sub-sequence is calculated. Equation 4.3 shows the process of calculating this mean, which is also standardized. The absolute value of the absolute mean gives the value of u. Formally:

$$u = \frac{\left(\sum_{t=1}^{n_1} R(x_t)\right) - 0.5n_1(n_1 + n_2 + 1)}{(n_1 n_2 (n_1 + n_2 + 1)/12)^{0.5}} \quad 4.3$$

where $R(x_t)$ is the rank in the sorted sequence of element x_t. The null hypothesis is rejected if u is greater than $u_{1-\alpha/2}$, which is the $1-\alpha/2$ quantile of the standard Normal distribution, with α as the significance level of the test. For both tests, the location of the shift and consequently the size of n_1 and n_2 are assumed to be known (Salas, 1993). A practical way to circumvent the problem of an unknown shift location is to apply the test at all potential shift locations, and to assume that the shift actually occurs at the location of the largest test value above the value of $T_{1-\alpha/2,\nu}$ and $u_{1-\alpha/2}$, respectively, for the Student's and Mann-Whitney tests.

4.2.2 Trends

The Mann-Kendall and the Spearman tests can be used for detecting trends. Given a sequence of individuals x_t, $t = 1, \ldots, N$, both tests assume a null hypothesis stating that there is no trend in the sequence. The Mann-Kendall test considers the sum (Z) of the gradients between each individual x_t, $t = 1, \ldots, N$-1, and all the subsequent individuals $x_{t'}$, $t' = t+1, \ldots, N$, in the sequence, as shown in Equation 4.4.

$$Z = \sum_{t=1}^{N-1} \sum_{t'=t+1}^{N} z_{t,t'} \quad 4.4$$

where:

$$z_{t,t'} = \begin{cases} 1 & \text{if } x_{t'} > x_t \\ 0 & \text{if } x_{t'} = x_t \\ -1 & \text{if } x_{t'} < x_t \end{cases} \quad 4.5$$

The sum of the sign of the gradients is then standardized as indicated in Equation 4.6 in order to give u.

$$u = \frac{Z + m}{\sqrt{V(Z)}} \quad 4.6$$

with:

$$V(Z) = \frac{1}{18} N(N-1)(2N+5) \quad 4.7$$

where $m = 1$ if $Z < 0$, and $m = -1$ if $Z > 0$. Variable $V(Z)$ is the estimated variance for the sum of gradients Z, and must be slightly altered as shown in Salas (1993) when there are duplicates in the sequence (i.e., individuals in the sequence with the same value). The null hypothesis is rejected if the absolute standardized value of u is greater than $u_{1-\alpha/2}$, which is the $1-\alpha/2$ quantile of the standard Normal distribution, with α as the significance level of the test. For the Spearman test, the sum of the difference between the actual position of each individual in the sequence, and its position in the sequence when it is sorted in ascending order, is performed, then standardized to yield quantity ρ, as given in Equation 4.8.

$$\rho = 1 - \frac{6\sum_{i=1}^{N}\left(R(x_i) - i\right)^2}{N\left(N^2 - 1\right)} \qquad 4.8$$

where $R(x_i)$ is the rank in the sorted sequence of element x_i. The null hypothesis is rejected if the absolute standardized value is greater than $\rho_{1-\alpha/2}$, which is the $1-\alpha/2$ quantile of the probability distribution related to the Spearman test, with α as the significance level of the test. The distribution can be found in Conover (1980). The results of the Mann-Kendall test may be affected if the sequence exhibits a significant auto-correlation. One can account for this by performing the test on the pre-whitened sequence. Assuming that r is the lag-1 auto-correlation, the pre-whitened sequence is $x_2 - rx_1, \ldots, x_n - rx_{n-1}$. It is suggested that one pre-whiten the sequence if $r > 0.1$ (Zhang et al., 2001).

4.2.3 Multivariate Cases

With the exception of the Mann-Kendall test, the aforementioned tests for shifts and trends, as described by Conover (1980) and Salas (1993), can only assess one data sequence at a time. It is not possible to use these tests to assess several sequences simultaneously to obtain an overall diagnosis of the presence of a shift or trend in such cases. Hirsh et al. (1982) develop a variation of the Mann-Kendall test for the detection of trends that provides an overall diagnosis when applied to several data sequences. This variation is an extension of the traditional univariate Mann-Kendall test. Hirsh and Slack (1984) extend the Mann-Kendall test further to account for serial dependence between the data sequences.

4.3 Detection Tests Based on Artificial Intelligence Techniques

4.3.1 Shifts

The purpose of clustering techniques such as the Kohonen network is to classify the individuals of a data sequence with respect to some specified features. Consider a univariate data sequence made of individuals coming from two distinct populations, as would be the case if there were a shift. If this data sequence were used to calibrate a Kohonen network, it would be expected that the individuals from the first population would activate a particular region of neurons on the output map, while the individuals from the other population would activate other neurons. On the Kohonen network map, as shown in Figure 4.1, the centroids (i.e., the bold dots) of the regions affected by each population can be calculated the same way centroids can be determined on a topographic map, and the distance between centroids thus becomes an indicator of the magnitude of the shift.

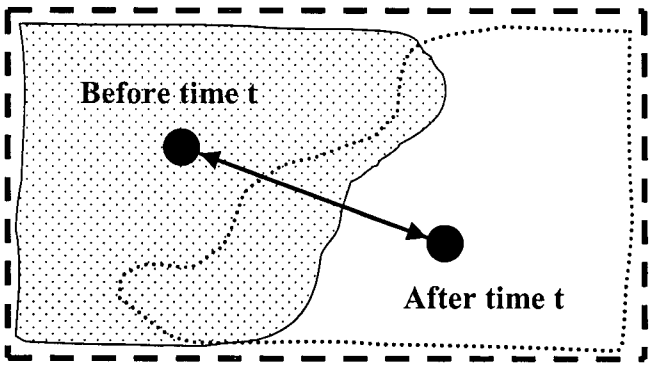


Figure 4.1. Detection of a shift with the Kohonen network.

For this application, distances obtained from the network are divided by the maximum distance that can possibly be achieved by the network. This standardization, which limits the distance values between 0 and 1, allows comparison between distances from different networks, whatever their respective dimensions. Similarly, if the clusters in fuzzy c-means can be ordered to form a map, then centroids and standardized distances can also be determined. Employed as such, the Kohonen network and fuzzy c-means perform exactly the same function as the Student's and Mann-Whitney tests which also provide a measure

of the distance that differentiates two populations. For all techniques, conventional and artificial intelligence tests, if the location of the shift in the data sequence is not known, the strategy is to verify all potential locations, and the location that produces the largest differentiating distance is assumed to be the most likely location of the shift. Also, if there is no shift at all in the data set, then all techniques should produce distances equal to zero at all potential locations.

4.3.2 Trends

With trends, it is assumed that features in the data sample constantly vary over time. This is the equivalent of having shifts at all possible locations on the data sample. Under this circumstance, all techniques employed for the detection of shifts would provide distances that are greater than zero for all potential locations. The Kohonen network, and fuzzy c-means once the clusters are ordered, would also give non-zero distances at all locations. This means, for example, that distinct regions of the Kohonen map (or the map of the clusters with fuzzy c-means) would be activated by the data before and after each location. Therefore a measure of the presence of a trend would be the mean of these distances. A large mean distance would indicate the likely presence of a trend while a small mean distance would indicate a lesser or no trend. Another way to use the Kohonen network and fuzzy c-means to detect trends is by verifying how the individuals of the data sample are grouped in each neuron and cluster. Indeed, if either technique were calibrated for data that fall on a smooth curve, then the chosen weights would assign each neuron or cluster to only a specific part of that curve. Similarly, with a data set that represents the record of a trend over time, each neuron or cluster should be activated by the data coming from a specific period of time. By extension, specific regions of the Kohonen map (or the map of the clusters with fuzzy c-means) would be activated with data coming from a specific period of time, as illustrated in Figure 4.2. The evaluation of the mean distance or the data grouping represents a measure of cohesion between data. The Mann-Kendall and Spearman tests evaluate how each individual in the data set ranks with respect to the other individuals, and therefore also provide a measure of cohesion between data.

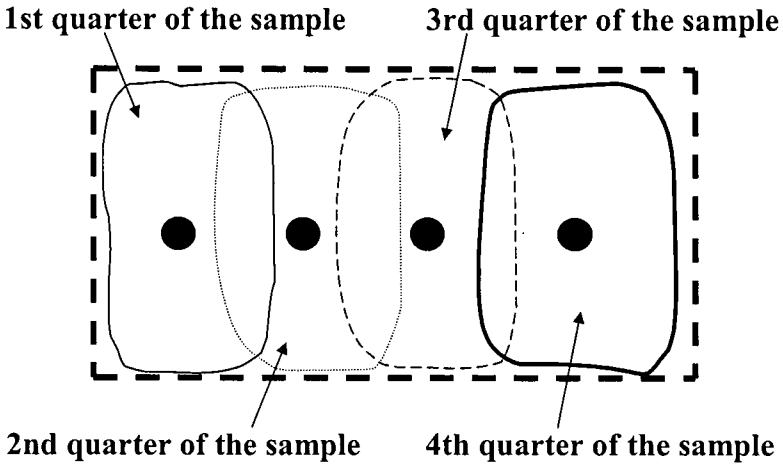


Figure 4.2. Detection of a trend with the Kohonen network.

4.3.3 Advantages and Disadvantages

The approaches for using AITs as detection tests are straightforward. For the univariate case, one data sequence is employed for the calibration of the Kohonen network or the clusters of fuzzy c-means, feeding only one individual of the sequence at a time. This calibration provides a one-dimension map. After calibration, the individuals of the same sequence are fed again to the Kohonen network or the clusters of fuzzy c-means, to see which neurons or clusters are activated and to evaluate the distance or cohesion, as explained in Section 4.3.1 and 4.3.2. The mapping properties of the Kohonen network make this technique a very suitable tool for the estimation of distance or cohesion between data elements. Fuzzy c-means is less suitable because it requires that the clusters be sorted before they are used as a map for estimating both distance and cohesion. The preliminary analysis of the results show that the fuzzy c-means approach used as indicated in this chapter is slightly inferior to the Kohonen network. The emphasis in this chapter is therefore placed on the Kohonen network and the comparison of its performance with that of the conventional statistical detection tests. In Chapter 5, a different detection strategy is proposed, where the use of both the Kohonen network and fuzzy c-means is suitable.

A disadvantage of the use of AITs for the detection of shifts and trends is that there is no decision criterion, or threshold, such as those under the conventional statistical tests, to indicate, with some level of confidence, whether or not there is a shift or a trend. Here an approach is proposed for determining the threshold for the occurrence of a shift or trend,

and it is based on minimizing the number of false detections by the AIT-based detection tests. This approach is applied with the AITs so as to provide them with a decision criterion, but it is also applied on the conventional statistical tests to compare the threshold based on a confidence level with that based on the minimization of false detection. The advantage of the Kohonen network is that it easily accommodates multivariate cases, that is, more than one data sequence or variate, unlike most statistical tests described in Section 4.2. For multivariate cases, one would increase the number of neurons in the input layer by one additional neuron per variate, and the number of weights per neuron in the output layer by one additional weight per variate.

In order to calibrate the artificial intelligence techniques, there cannot be more parameters (i.e, weights of the output neurons for the Kohonen network and elements of the cluster centers for fuzzy c-means) than there are input vectors from the data sequences. Here, the number of parameters (i.e., the product of the number of variates and the number of output neurons or clusters) is restricted so that it is approximately 20% of the number of input vectors (i.e., the number of individuals in the data sequences). For example, consider a sequence of 30-individual data with two variates. For this multivariate case, the Kohonen network may be employed, and could accommodate six total weights (i.e., 30 individuals $\times$ 0.20 = 6). Therefore, for the two variates, three output neurons would be possible (i.e., six total weights $\div$ two variates).

4.4 Experimental Design and Database

Synthetic streamflow data sequences, generated by Monte-Carlo simulations, are used here to evaluate the performance of the tests for shifts and trends. The necessary random generators used in these Monte-Carlo simulations are those provided in MATLAB. This Monte-Carlo approach is inspired by that of Hirsh et al. (1982), although the data type and method of analysis are different and several conventional tests as well as the Kohonen network and fuzzy c-means tests are evaluated. The means and coefficients of variation of the data are specified to reflect those of natural streamflow records found in Canada. These statistical properties could represent annual or seasonal streamflow peak and mean sequences.

The performance of each test for detecting a shift or trend is evaluated for cases that represent variations in (1) the length of the data sequences, (2) the coefficient of variation of the individuals in the sequence, and (3) the amplitude of the shifts or trends imposed on the data sequences. Synthetic streamflow data sequences comprised of 30, 40 or 50-individual data sequences are used. A batch is defined as a grouping of five sets of sequences (each set being comprised of 10,000 sequences). Each set in a batch represents data for a specific variate, and experiments are conducted for the univariate case, as well as for cases involving two and five variates. The mean and the coefficient of variation of the data sequences vary uniformly between 1 and 20,000 and between 0.05 and 0.5, respectively. Each point in a sequence is created randomly, following a Normal distribution.

In order to determine the success in detecting shifts or trends, half of the data sequences in each batch are corrupted each with one shift or one trend, depending on the test conducted, and the other half are uncorrupted. If a sequence in a set is corrupted, then all the corresponding sequences in the other sets (i.e., for the other variates) of the batch are also corrupted. For tests of shifts, the shift amplitude is chosen randomly, following a Uniform distribution, and can be as much as ±25% of the mean of the sequence prior to the shift. The location of the shift is determined randomly, following a Uniform distribution, and can be anywhere in the sequence except within the first and last five individuals. The location of the shift, when there is one, is the same for all corresponding sequences in the sets (i.e., for the other variates) of a batch. For tests of trends, the amplitude of the increase or decrease of the mean is chosen randomly, following a Uniform distribution, and can be as much as ±0.5% of the initial mean per time step. The direction of the trend, when there is one, is the same, either downward or upward, in all corresponding sequences in the sets of a batch. For detection of both shifts and trends, two types of batches are examined, each having the same sequence length (i.e., either 30, 40 or 50 individuals). In one batch, corruption within corresponding sequences is of the same amplitude, while in the other batch the amplitude of the corruption varies within corresponding sequences. Thus the performance of the detection tests is investigated for twelve batches of data, that is, two batches each for data sequences of 30-, 40- and 50-individuals, and this for both tests for detection of shifts and trends.

The cases examined are referenced based on identification codes presented in Table 4.1. Each detection test is identified by two letters, which are followed by a number and a series of letters that represents the number of variates examined and variations in the data. For example, MK5BW represents the Mann-Kendall test, multivariate case where five corresponding sequences are tested simultaneously, in which trends of different amplitude exist within corresponding data sequences, and data sequences are whitened. As a memory aid for distinguishing the conventional statistical test from the tests using the Kohonen network, remember that the identification code for the latter tests always starts with a K. Similarly, the identification code for fuzzy c-means always starts with an F.

Table 4.1. Cases of detection tests evaluated.

Initials	Description
(a) Shifts	
MW	Mann-Whitney test
ST	Student's test
FS	Fuzzy c-means test for shifts
KS	Kohonen network test for shifts
(b) Trends	
SP	Spearman test
MK	Mann-Kendall test
FT	Fuzzy c-means test for trends
KT	Kohonen network test for trends
(c) Further details following the first two letters identifying the test	
1	Univariate case, one sequence tested at a time
2	Multivariate case, two sequences tested simultaneously
5	Multivariate case, five sequences tested simultaneously
A	Cases of shifts or trends of same amplitude within corresponding sequences (multivariate cases only)
B	Cases of shifts or trends of different amplitude within corresponding sequences (multivariate cases only)
W	Withened sequences (apply to Mann-Kendall test only)

4.5 Results

4.5.1 Corrupted Versus Uncorrupted Data

The value of a detection test is determined by its capacity to differentiate between corrupted and uncorrupted data, in other words by its capacity to avoid false detection. False detection occurs when corrupted sequences are falsely identified as uncorrupted and when uncorrupted sequences are falsely identified as corrupted. All detection tests for shifts

give a measure of distance while all detection tests for trends give a measure of cohesion. Larger distances or cohesion, indicate larger likelihood of shifts or trends, respectively. Of course, the greater the amplitude of a shift or trend, the greater the likelihood of a high measure of distance or cohesion, although the variance of the data sequence may also affect the measure of the distance or cohesion. Consequently, the efficiency of detection tests is analyzed here with respect to the ratio of the amplitude of the corruption to the coefficient of variation of the data sequences (Amp/CV). Figure 4.3 provides a typical representation of the relationship between the average maximum distances and the Amp/CV ratio. The data in Figure 4.3 show the results for the MW1 case for shifts for all sequence sizes tested, although similar results are obtained for all other tests, conventional and AIT-based, applied to univariate cases, for both the detection of shifts and trends.

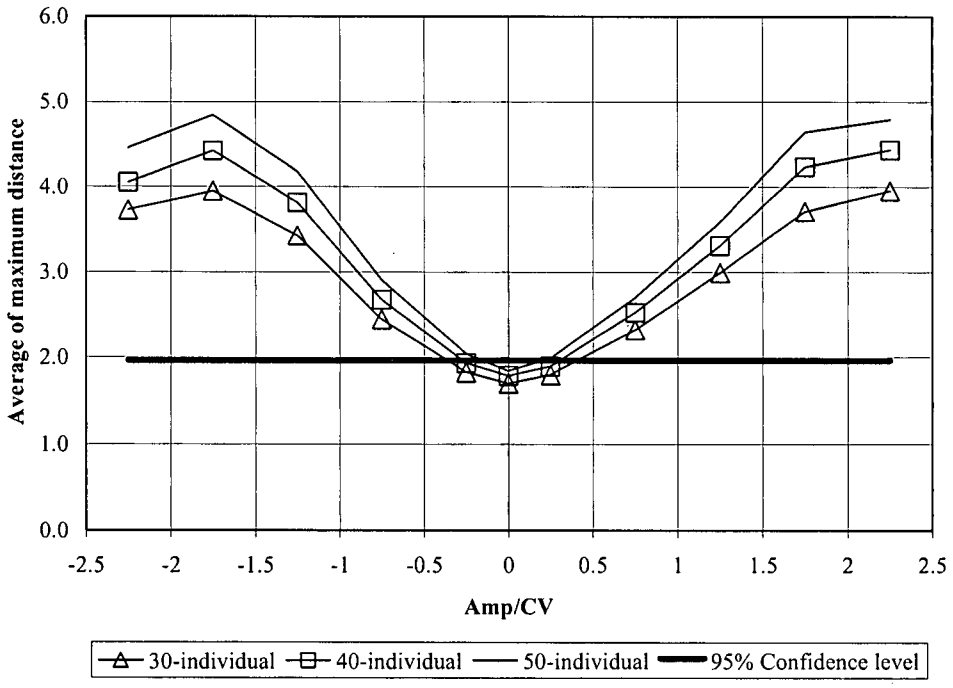


Figure 4.3. Mann-Whitney test for the detection of shifts for univariate cases.

Figure 4.3 shows that the average maximum distances are high for both largely positive and largely negative Amp/CV ratios. Indeed, providing the amplitude of the shift or trend is significantly large compared with the coefficient of variation, one can easily see the corruption in a chronologically plotted data sequence. As the Amp/CV ratios decrease,

the average maximum distances become smaller and reach a minimum when Amp/CV = 0. In data sequences affected by shifts or trends of small amplitude compared with the coefficient of variation it becomes more difficult to visually identify the corruption. Importantly, Figure 4.3 shows that the average distances for data sequences with small Amp/CV ratios are quite close to the average distance for uncorrupted data sequences (i.e., Amp/CV = 0). In practice, a threshold distance value, based on a confidence level (e.g., 95%, as shown in Figure 4.3) in the case of the conventional statistical tests, is normally used to distinguish between corrupted and uncorrupted sequences. A false detection occurs when the distance value of an uncorrupted data sequence is higher than the threshold value, or when the distance value of a corrupted sequence is smaller than the threshold value. False detection for corrupted sequences is highly possible when the Amp/CV ratio is small. In an ideal situation, the distance values for uncorrupted sequences should be as close as possible to zero, while the distance values for corrupted sequences should be as high as possible, thus allowing for a clear threshold to be determined that minimizes, if not prevents, false detection.

4.5.2 Setting the Threshold Values

For conventional statistical tests, there is no indication that setting a confidence level at 90, 95 or 99% would provide optimal results in terms of reducing false detection. This confidence level is only a statistical concept that has little meaning relative to the data being analyzed. Here, for the Kohonen network and the conventional tests, the optimal threshold is set at a value that minimizes false detection expressed by the sum of 1) the ratio of uncorrupted sequences falsely detected as corrupted over the total number of sequences, plus 2) the ratio of corrupted sequences falsely detected as uncorrupted over the total number of sequences. In the minimization process, equal weights are given to both ratios, for it is just as unacceptable to falsely detect corrupted sequences as it is to falsely detect uncorrupted ones. This approach is necessary for the detection tests employing the Kohonen network, as there is no other way to determine a threshold value for these tests. It is also applied for the conventional statistical tests, as shown in Table 4.2 for shifts and Table 4.3 for trends. For the conventional tests, the most commonly used confidence level of 95% is also shown, for comparison sake. In Tables 4.2 and 4.3, column 1 is the detection

test employed, column 2 is the threshold value obtained from the minimization procedure, column 3 is the ratio (U, in %) of falsely-detected uncorrupted sequences, over the total number of sequences, column 4 is the ratio (C, in %) of falsely-detected corrupted sequences over the total number of sequences, column 5 is the sum of columns 3 and 4, and columns 6 to 9 are the same as columns 2 to 5, with the threshold values based on the 95% confidence level. In Table 4.2, the Us can be added to the Cs directly, because there is an equal proportion of uncorrupted and corrupted sequences in the database.

Table 4.2 indicates that the conventional statistical tests perform slightly better than the Kohonen network, the best test being the Student's test (ST1). Also, as the sequence size increases, the Mann-Whitney test (MW1) performs better when the threshold is chosen based on the minimization of false detection than when the 95% confidence level is used. Compared with the thresholds based on the 95% confidence level, the thresholds based on the minimization process result in a lower ratio of uncorrupted sequences falsely detected as corrupted and in a higher ratio of corrupted sequences falsely detected as uncorrupted. Overall, the tests for the detection of shifts perform a false detection about 30% of the time.

Table 4.2. Thresholds for univariate cases with shifts.

Case	Optimal				95% confidence level			
	Threshold	False detection ratio (%)			Threshold	False detection ratio (%)		
		U	C	U+C		U	C	U+C
(a) 30-individual sequences								
MW1	2.26	7.6	23.8	31.3	1.96	14.7	17.8	32.5
ST1	2.43	7.6	23.3	30.9	2.37	8.5	22.4	30.9
KS1	0.31	8.6	23.2	31.8	NA	NA	NA	NA
(b) 40-individual sequences								
MW1	2.44	6.0	23.7	29.6	1.96	17.1	14.9	31.9
ST1	2.53	7.0	22.2	29.2	2.33	10.2	19.4	29.6
KS1	0.29	7.7	24.1	31.8	NA	NA	NA	NA
(c) 50-individual sequences								
MW1	2.43	7.0	21.1	28.1	1.96	18.5	12.8	31.3
ST1	2.55	7.0	20.5	27.6	2.31	11.2	17.0	28.2
KS1	0.27	8.0	22.3	30.3	NA	NA	NA	NA

Note: NA = Not Applicable.

Table 4.3 indicates that the Kohonen network using a threshold based on the minimization process performs slightly better than the conventional statistical tests with thresholds based on the 95% confidence level. When the thresholds based on the

minimization process are employed, the conventional statistical tests perform as well as the Kohonen network. Also, it appears that whitening the sequences prior to using the Mann-Kendall test is detrimental, as this approach results in the worst performance of all detection tests for trends. Compared with the thresholds based on the 95% confidence level, the thresholds based on the minimization process result in a higher ratio of uncorrupted sequences falsely detected as corrupted and in a lower ratio of corrupted sequences falsely detected as uncorrupted. Overall, the tests employed for the detection of trends perform a false detection around 40, 35 and 30% of the time, for 30, 40 and 50-individual sequences, respectively. The improvement of the performance as the sequence size increases in the cases of trends is due to the nature of the corruption or trend in that it becomes more obvious with time. When there is a trend, the average of the sequence is increased or decreased at every time step by some percentage of the initial average (0.5% per time step being the maximum), and this systematic modification is easier to detect in the long term (large sequence size) than in the short term (small sequence size).

Table 4.3. Thresholds for univariate cases with trends.

Case	Optimal				95% confidence level			
	Threshold	False detection ratio (%)			Threshold	False detection ratio (%)		
		U	C	U+C		U	C	U+C
(a) 30-individual sequences								
MK1	1.25	10.0	32.0	42.0	1.96	2.4	41.9	44.2
MK1W	1.03	13.4	29.5	42.9	1.96	1.6	44.8	46.4
SP1	0.23	10.6	31.3	41.9	0.36	2.5	41.6	44.1
KT1	0.16	12.8	28.7	41.6	NA	NA	NA	NA
(b) 40-individual sequences								
MK1	1.34	8.6	27.6	36.2	1.96	2.4	36.5	39.0
MK1W	1.19	10.4	26.7	37.2	1.96	1.7	39.8	41.5
SP1	0.20	10.6	25.5	36.1	0.31	2.6	36.1	38.7
KT1	0.14	8.4	28.3	36.7	NA	NA	NA	NA
(c) 50-individual sequences								
MK1	1.42	7.5	23.4	31.0	1.96	2.4	30.7	33.1
MK1W	1.31	8.6	23.3	31.8	1.96	1.9	33.5	35.4
SP1	0.19	8.8	22.2	30.9	0.28	2.5	30.5	33.0
KT1	0.13	8.4	22.8	31.2	NA	NA	NA	NA

Note: NA = Not Applicable.

4.5.3 Finding the Location of Shifts

In the case of shifts, the challenge is not only in determining whether or not there is a shift, but also in finding the location of the shift when there is one. For the univariate

cases, Table 4.4 presents the success rate of all tests in identifying the location of the shifts. Here, the results consider only the corrupted sequences, and success of identification occurs when the test properly detects the presence of a shift (i.e., the differentiating distance value above the detection threshold) and identifies its location exactly or within ±1, 2, 3, 4, or 5 time steps. Table 4.4 indicates that all tests exactly identify the location of the shift on corrupted sequences approximately 20% of the time, and within ±5 time steps slightly more than 40% of the time. Conventional statistical tests perform slightly better than the Kohonen network. The performance slightly decreases as the sequence size increases, because it becomes more difficult to properly identify the location of the shift as the number of possible locations increases. Of course, the success rate varies with respect to the Amp/CV ratio, that is, a sequence with a highly positive or negative Amp/CV ratio (i.e., a greater distance value on the average) is more likely to provide a clearer indication of the location of the shift than a sequence with a smaller Amp/CV ratio. When the Amp/CV is greater than | ± 1.75 |, the success rate of finding the exact location of the shift is 80% or above for all detection tests, and is 100% when identifying the location of the shift within ±5 time steps. With small Amp/CV values, less than | ±0.25 |, the success rate of finding the exact location of the shift is approximately 5% for all tests, and is slightly more than 20% when identifying the location of the shift within ±5 time steps.

Table 4.4. Success rate in identifying the location of the shift with univariate cases.

Case	Success rate (%) in identifying the location of the shift at plus minus					
	0 time step	1 time step	2 time steps	3 time steps	4 time steps	5 time steps
(a) 30-individual sequences						
MW1	19	31	37	41	44	46
ST1	20	31	37	41	44	46
KS1	17	27	32	36	39	41
(b) 40-individual sequences						
MW1	19	30	36	39	42	43
ST1	21	31	38	41	44	46
KS1	16	24	29	32	35	36
(c) 50-individual sequences						
MW1	19	31	37	41	43	46
ST1	20	31	37	41	43	46
KS1	15	24	29	32	34	36

4.5.4 Multivariate Cases

Typically, an analysis of shifts and trends with a multivariate sequence is expected to provide a clearer response than that from analyzing univariate cases separately. As indicated previously, the Kohonen network can detect both shifts and trends for multivariate cases, and the Mann-Kendall test can detect trends for multivariate cases. In the case of the Mann-Kendall test, the approach developed by Hirsh and Slack (1984), which takes the serial dependence between sequences into account, is evaluated here. The results presented show that detection tests for multivariate cases indeed perform better than those for univariate cases. The improvement in the performance is illustrated by values of distance (shifts) or cohesion (trends) between corrupted and uncorrupted sequences that are more distinctive than those under the univariate case. Similarly to Figure 4.3, Figure 4.4 presents the average maximum distances with respect to the Amp/CV ratio obtained from the Kohonen network, testing five sequences simultaneously, for the detection of shifts.

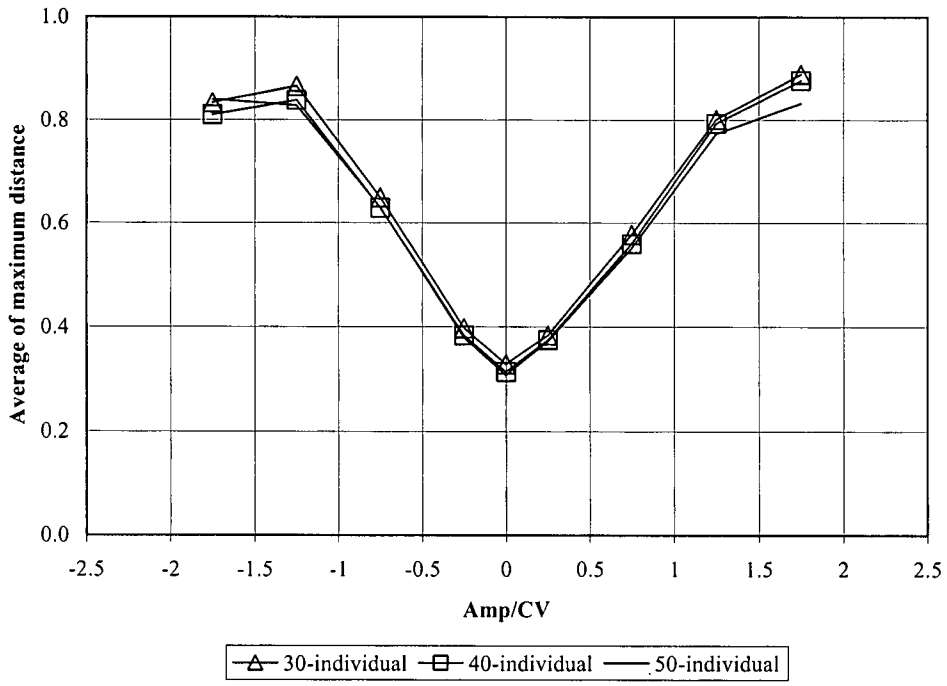


Figure 4.4. Kohonen network for the detection of shifts for multivariate cases.

Again, particular attention should be given to the region with small Amp/CV ratios. The distances in Figures 4.3 and 4.4 are on different scales. However, the ratio of the average

distance for corrupted sequences (i.e., at an Amp/CV = ±0.25) and the average distance for uncorrupted sequences (i.e., at an Amp/CV = 0) is proportionally larger for the data shown in Figure 4.4 than that for the data shown in Figure 4.3. This ratio equals 1.07, 1.07 and 1.09 for univariate 30, 40 and 50-individual sequence cases, respectively, while it is equal to 1.19, 1.21 and 1.22 for multivariate 30, 40 and 50-individiual sequence cases. Similar conclusions can be drawn for corrupted sequences with larger Amp/CV ratios versus uncorrupted sequences (Amp/CV = 0). This implies that there is more potential in the multivariate case to establish a threshold value that differentiates between corrupted and uncorrupted sequences, and this results in a reduced occurrence of false detection compared with that of the univariate case.

Tables 4.5 and 4.6 show the percentage of falsely detected sequences for multivariate cases for shifts and trends, respectively. The data in Tables 4.5 and 4.6 confirm the reduced occurrence of false detection in the multivariate cases compared with that of univariate cases presented in Tables 4.2 and 4.3.

Table 4.5. Thresholds for multivariate cases with shifts.

Case	Optimal			
	Threshold	False detection ratio (%)		
		U	C	U+C
(a) 30-individual sequences				
KS2A	0.42	6.9	23.6	30.5
KS2B	0.41	9.0	21.6	30.6
KS5A	0.52	5.1	20.1	25.3
KS5B	0.50	6.6	14.4	21.0
(b) 40-individual sequences				
KS2A	0.38	5.9	24.3	30.3
KS2B	0.37	6.2	23.3	29.5
KS5A	0.49	6.0	18.5	24.6
KS5B	0.49	5.8	13.6	19.4
(c) 50-individual sequences				
KS2A	0.32	8.5	21.0	29.5
KS2B	0.32	8.7	21.2	29.9
KS5A	0.47	6.7	17.3	24.0
KS5B	0.45	7.4	12.0	19.4

The results related to shifts in Tables 4.5 and 4.2 show that there is no reduction of false detection when two sequences are tested simultaneously (KS2A and KS2B), but the

reduction is definitely noticeable when five sequences are tested simultaneously (KS5A and KS5B). The KS5B case (i.e., where the B signifies the case with shifts of different amplitude within corresponding sequences) provides better results than the KS5A case (i.e., where the A signifies the case of shifts of same amplitude within corresponding sequences). This can be explained by considering the relative sizes of amplitudes of shifts. For all A cases a sequence with a small, difficult to detect, amplitude is combined with other sequences with the same small, equally difficult to detect, amplitude, making the overall detection of the shift on all sequences difficult. For all B cases a sequence with a small, difficult to detect, amplitude may be combined with a sequence with a large, potentially easier to detect, amplitude, making the overall detection of the shift on these sequences easy. Hence the overall performance of tests for the B cases is greater than that of the A cases.

For multivariate tests of trends, eight possible cases can be generated from the Mann-Kendall test (MK2A, MK2AW, MK2B, MK2BW, MK5A, MK5AW, MK5B, MK5BW), and two cases can be generated from the Kohonen network (KT2A and KT2B), but for the sake of conciseness Table 4.6 only presents a sampling of all possible cases, which are considered representative. The Kohonen network cannot be used for cases of five sequences tested simultaneously (i.e., KT5A and KT5B) due to the restriction imposed on the number of output neurons with respect to the sequence size.

Table 4.6. Thresholds for multivariate cases with trends.

Case	Optimal				95% confidence level			
	Threshold	False detection ratio (%)			Threshold	False detection ratio (%)		
		U	C	U+C		U	C	U+C
(a) 30-individual sequences								
MK2B	1.23	12.0	25.8	37.8	1.96	3.9	37.6	41.6
MK5B	1.30	10.9	17.6	28.5	1.96	4.0	30.7	34.7
MK5BW	1.18	11.6	19.8	31.4	1.96	3.7	35.9	39.5
KT2B	0.21	10.2	31.3	41.5	NA	NA	NA	NA
(b) 40-individual sequences								
MK2B	1.48	8.2	22.3	30.5	1.96	3.7	30.0	33.6
MK5B	1.49	7.9	10.5	18.4	1.96	3.9	17.6	21.5
MK5BW	1.31	9.3	11.2	20.5	1.96	3.4	23.3	26.7
KT2B	0.19	6.4	31.2	37.6	NA	NA	NA	NA
(c) 50-individual sequences								
MK2B	1.58	6.6	16.2	22.8	1.96	3.3	21.4	24.8
MK5B	1.79	4.9	6.0	10.9	1.96	3.8	7.6	11.4
MK5BW	1.58	6.3	6.5	12.9	1.96	3.5	11.7	15.2
KT2B	0.16	7.1	25.6	32.6	NA	NA	NA	NA

Note: NA = Not Applicable.

The results related to trends in Tables 4.6 and 4.3 show that a noticeable reduction of false detection occurs when two sequences are tested simultaneously with the Mann-Kendall test (MK2B), and that the reduction is even more pronounced when five sequences are tested simultaneously (MK5B and MK5BW). The Kohonen network for the detection of trends yields no noticeable improvement in the case of two sequences tested simultaneously. Note that whitening the sequences before the test (i.e., the W cases) results in greater false detection. As was observed for the univariate case, false detection decreases as the sequence size increases. As was observed for the tests for the detection of shifts, for detection of trends, the B cases yield better results than the A cases. For all cases in detection of trends, the threshold values based on the 95% confidence level always lead to more false detection than the cases in which the threshold is based on the minimization of false detection.

In the case of shifts, the approach for finding the location of the shift is more successful with the multivariate cases than it is with univariate cases. Table 4.7 shows the success rate of finding the location exactly and within ±5 time steps. In comparison with the results given in Table 4.4, there is no real improvement when two sequences are tested simultaneously (KS2A and KS2B), but the improvement is apparent when five sequences are tested simultaneously (KS5A and KS5B).

4.6 Discussion and Conclusion

The results presented are in agreement with those of Hirsh et al. (1982) with respect to the behavior of the detection tests, and while AIT-based tests may be thought of as confirming the conventional tests, one must use all of these tests with caution. Indeed, the rate of false detection for univariate cases of at best around 30% for shifts and trends shown in this work is certainly high, even from a hydrologic perspective where errors due to uncertainties related to data adequacy and model structures can be significant. The results worsen as the amplitude of the shift or the trend decreases, yet it is desirable that a technique be able to reliably detect shifts and trends of low amplitude, for even a small change of the mean in the streamflow regime can make the difference between a profitable and non-profitable water resources project.

Table 4.7. Success rate in identifying the location of the shift with multivariate cases.

Case	Success rate (%) in identifying the location of the shift at plus minus					
	0 time step	1 time step	2 time steps	3 time steps	4 time steps	5 time steps
(a) 30-individual sequences						
KS2A	23	32	36	40	42	44
KS2B	22	33	38	42	44	46
KS5A	35	45	49	52	53	54
KS5B	39	52	57	61	63	64
(b) 40-individual sequences						
KS2A	21	29	33	36	38	40
KS2B	20	29	33	37	39	41
KS5A	36	45	50	52	54	55
KS5B	40	51	57	60	62	64
(c) 50-individual sequences						
KS2A	20	28	32	35	37	39
KS2B	19	27	32	35	38	39
KS5A	36	46	50	52	55	56
KS5B	39	51	57	61	63	65

The performance of the detection tests is evaluated using synthetic data, which are designed to represent the mean and coefficient of variation of data originating from hydrometric measurements in Canada. No attempt has been made to add more parameters in the characterization of data, such as through auto-correlation to account for possible dependence among data points or the skewness coefficient so as to depart from the Normal distribution. It would require a large amount of data sequences to account for the variability of all these parameters, and this would greatly increase the computational burden for the evaluation of the detection tests.

The results show that whitening the data, which removes the dependence among data points in a sequence, prior to using the Mann-Kendall test for trends leads to a reduction of the performance of the test. It must be admitted here that the data sequences employed in this application are the cause of the reduction of the performance of the test. Because the points in the sequences are designed to be independent of each other, there is thus no point in performing any whitening.

The evaluation of the performance is accomplished with respect to the Amp/CV ratio, although it is not a ratio that one can readily use in practice, for one normally does not know the amplitude of the shift or trend in advance. This ratio is nevertheless used

throughout this study because it is a better indicator of the difficulty in identifying shifts and trends than either the amplitude of the anomaly or the coefficient of variation of the sequence alone. One needs only to randomly generate corrupted data sequences to notice that even a shift or a trend of large amplitude may not be easy to identify visually if the coefficient of variation is also large. Additional analyses of the performance with respect to the amplitude alone and the coefficient of variation alone have been undertaken, and confirm commonsense, that is, that the performance of the detection tests decreases as the coefficient of variation of the sequence increases and as the amplitude of the anomaly decreases.

This application also demonstrates that the threshold values, which identify the separation between corrupted and uncorrupted data, may not always be optimal if based on a specific confidence level. It is important that these thresholds be set so as to minimize the occurrence of false detection, rather than to satisfy a statistical concept that has little meaning for the data under study. This statement challenges the common practice of using thresholds based on some confidence level, although this is not a new challenge. Many researchers and practitioners often employ several thresholds at various confidence levels when applying detection tests so as to evaluate a range of potential results.

Detection tests for more than one data sequence at once are also evaluated for their performance in identifying shifts and trends, and results shows that these tests can provide more reliable diagnoses compared with those applied to only one sequence at a time. When the Mann-Kendall test for detection of trends is applied to a case of five sequences tested simultaneously, the percentage of overall false detection reduces to around 15% with 50-individual sequences, the best case being as low as 11% (MK5B, see Table 4.6). This false detection rate could be considered as a reasonable error by hydrologic standards, and is definitely better than the best performance obtained for the univariate cases (31%, MK1, SP1 and KT1, Table 4.3). Whenever possible, it may be preferable to employ these detection tests for multivariate cases, for example, when analyzing hydrometric data from stations that are close to each other geographically. With a large number of hydrometric stations, performing a regionalisation analysis prior to the use of the detection tests may be beneficial, assuming the regionalisation technique is reliable and does not generate errors that counter the benefit of multivariate detection tests.

The Kohonen network and fuzzy c-means are relatively new techniques that can help identify patterns in data, and have been structured in this application so as to replicate the behavior of the conventional statistical tests. The Kohonen network, especially, provides similar performance to that of conventional detection tests. The AIT-based tests may be used to confirm the results of conventional detection tests. They also constitute an enhancement relative to the conventional detection tests for multivariate cases. AIT-based tests require more computational time than the conventional tests, but this is viewed as reasonable, given the large number of sequences assessed in this chapter (i.e., tens of thousands). In a practical context, where only a few sequences might be tested, the added computational burden of the AIT-based tests would be negligible. In the case of the Kohonen network, its advantage is its capacity to adapt for multivariate cases, which only one of the conventional tests for trends presented here, and none for the case of shifts, can do. The Kohonen network and fuzzy c-means can be structured in other meaningful ways for the identification of shifts and trends, and one approach for doing so is developed and demonstrated in Chapter 5

Chapter 5
Mapping Procedures for Detecting Anomalies

The approaches based on the Kohonen neural network and fuzzy c-means developed in Chapter 4 are constrained by the size of the database. The database employed to calibrate the weights of the output neurons or the clusters is comprised of individual sequences (i.e., one for the univariate cases and two to five for multivariate cases), the inputs being data points (i.e., one for the univariate cases and two to five for multivariate cases). The sequences employed in Chapter 4 are chosen to represent annual hydrologic events, leading to relatively few individuals per sequence, the usual amount in Canada being between 30 to 50 annual events. As a result, the number of output neurons on the Kohonen map or the number of clusters in the fuzzy c-means approach must be small, limiting the task of these methods to roughly sorting the data in the sequence in ascending or descending order. As such, the Kohonen neural network and fuzzy c-means approach are not used to their full potential.

In this chapter, the proposed approach is to use entire sequences as inputs to the AITs instead of only data points within the sequences. This implies that the AIT tools are calibrated on large sets of sequences instead of single sequences as previously undertaken. The objective of the Kohonen network and fuzzy c-means is to sort sequences with respect to the patterns present in them, that is, the patterns associated with the absence in sequences of anomalies such as outliers, shifts and trends, and patterns associated with the presence in sequences of anomalies of more or less large magnitude. Of course, the database is still a constraining factor, but since it may be comprised of a very large number of sequences, the number of output neurons or clusters may then be large, allowing for a relatively exhaustive discrimination of possible patterns present in the sequences. This approach can take a greater advantage of the potential of both the Kohonen neural network and fuzzy c-means, and can also address the case of outliers, which is not achievable with the approach employed in Chapter 4. In Section 5.1, the protocol of the experiment and the uncertainties related to the calibration process are presented. These two issues are common for the application to shifts and trends as well as to outliers. The case of outliers requires a

treatment that differs from that of shifts and trends, and thus tests for detecting shifts and trends are addressed Section 5.2, while those for outliers are addressed in Section 5.3. Section 5.4 provides the conclusions to this chapter.

5.1 Common Elements of the Applications

5.1.1 Protocol of Experiment

The goal here is to develop Kohonen maps or fuzzy c-means cluster sets that could be used to differentiate between sequences corrupted with anomalies from those sequences that are uncorrupted. In addition, in the case of anomalous sequences, such maps or cluster sets should provide an estimate for the amplitude of the corruption and for the location of the corruption when this applies (i.e., shifts and outliers). This requires that the behavior of the maps or cluster sets be known, and therefore the calibration and validation process should be conducted using data sequences for which the characteristics are known. Synthetic data are thus employed to create the databases for both the calibration and validation steps so as to permit the appropriate interpretation of the maps and cluster sets. Of course, the databases must be as representative as possible of the real data on which the maps and cluster sets are to be ultimately employed. Fortunately, shifts and trends are relatively easy to replicate synthetically. Outliers constitute a more complex case, yet can be reduced to a manageable number of characteristic types.

The calibration of a Kohonen map or fuzzy c-means cluster set make use of one set of data sequences for univariate cases of outliers, shifts or trends, and either two or five sets of sequences for multivariate cases of shifts or trends. The number of sequences in the set depends on the number of weights that must be calibrated. Similar to the rule established for the calibration process in the application in Chapter 4, the number of weights to calibrate must be only a fraction (e.g. 20%) of the number of available data sequences. For example, if one builds a Kohonen map with an output layer of 10×10 neurons, with 30-individual sequences, this implies that the calibration of 10×10×30 = 3000 weights is necessary. If it is imposed that the number of weights be only 20% of the number of data sequences, then 15000 sequences are required for the calibration process (i.e., 3000÷0.20). The fraction of 20% is employed here for all univariate cases involving outliers, shifts and trends, as it is in Chapter 4. With multivariate cases for shifts and trends, the fraction stands

at 20% for cases involving 2 sets of sequences, but is set at 50% for cases involving 5 sets of sequences so as to reduce the size of the database to a manageable proportion considering the computer capabilities available. For multivariate cases, as in Chapter 4, if one sequence in a set is corrupted, then the corresponding sequences in the other sets are similarly corrupted. When an anomaly is present, the sequences employed in the calibration are designed to exhibit no variation aside from that caused by the anomalies. Initial tests have shown that such structure for the sequences leads to the calibration of Kohonen maps and fuzzy c-means clusters that have higher performance. The main criterion in the design of the calibration sequences is that the differences between these sequences and those from the validation sets, or other real data, are minimized.

The validation sequences are meant to be as close as possible to reality. In the case of shifts and trends, the sequences employed are generated in the same way as those used in Chapter 4. In the case of outliers, real data sequences, from hydrometric stations known to be of good quality and assumed to be free of outliers are used, and deliberately corrupted when necessary. As is the case with the calibration sets, the Kohonen maps or fuzzy c-means clusters are fed with one set of data sequences for univariate cases of outliers, shifts or trends, and either two or five sets of sequences for multivariate cases of shifts or trends. The detection performance of the Kohonen maps or fuzzy c-means clusters is determined with respect to the ratio of amplitude of the corruption over the coefficient of variation of the sequences. The capacity of the maps or cluster sets to determine the amplitude of the corruption with accuracy is also established based on the validation sets. The reliability of these AIT in finding the location of shifts or outliers in sequences is also analyzed based on the validation sets.

5.1.2 Addressing Uncertainties in Calibration

The calibration process for both the Kohonen network and fuzzy c-means may lead to instances where some patterns present in the data are not adequately represented. In the case of the Kohonen network, the calibration sequences are fed one at a time, randomly, and the order by which the sequences are presented to the network may affect the direction in which the map unfolds. A simple example is that of two identical maps, with one inverted relative to the other one. In more complex situations, the map might be unfolded in

a way that favors some patterns to the detriment of others. In the case of fuzzy c-means, because the calibration involves an optimization procedure, there is the possibility that the final solution is trapped at a local optimum, therefore reducing the validity of the cluster set. To circumvent these calibration problems, more than one Kohonen map or cluster sets is relied upon. Similar to the application of Monte-Carlo simulations for addressing uncertainties, several maps or cluster sets are calibrated, and the final result of a detection diagnostic is based on the aggregation of the results coming from all of the available maps or cluster sets, as shown in Figure 5.1. Here, the maps and cluster sets are meant to provide estimations of the ratio of the amplitude of the corruption over the coefficient of variation of the sample or the location of the shift or the outliers. The aggregation procedure employed simply consists of averaging the estimates from all of the maps or cluster sets available. Because the calibration procedures for both the Kohonen network and fuzzy c-means are rather lengthy, a total of 10 maps or cluster sets per case are aggregated.

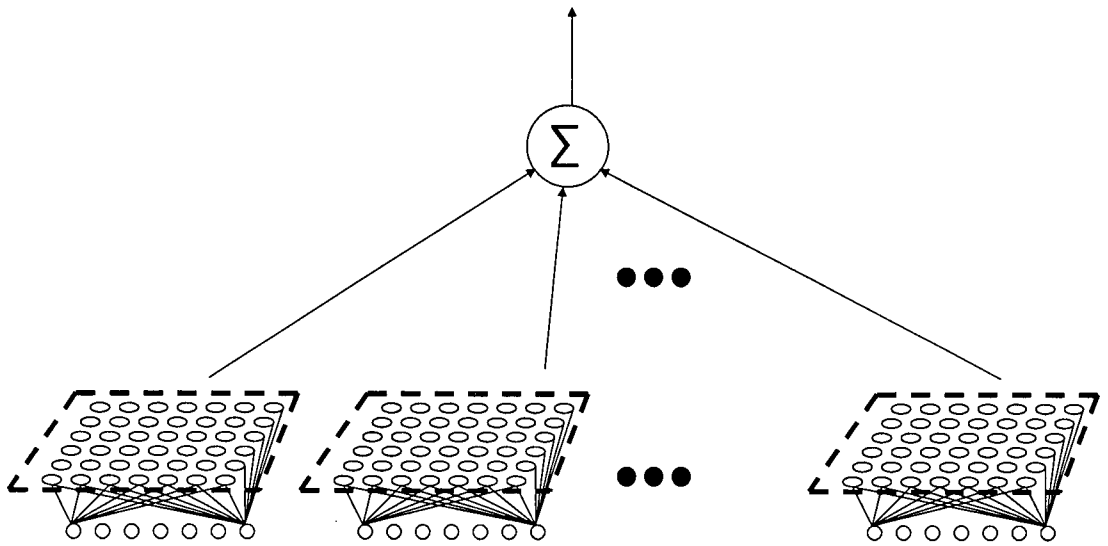


Figure 5.1. Aggregation of results from Kohonen network maps.

5.2 Application on Shifts and Trends

As a reminder, the goal of this work is to develop tools for the analysis of hydrometric data, more particularly annual indicators such as the annual extremes (min or max) or annual averages. As stated in Chapter 4, these annual indicators have been the

subject of numerous studies in North America, and the need for the kind of analyses produced in these studies can still be regularly present in practice. It must be noted that, as is the case of conventional statistical tests employed in these studies, the sets of Kohonen maps and fuzzy c-means clusters can be used for data other than those from hydrometric observations. The only requirement is the respect of the length limitation of the input vector for both the Kohonen maps and the fuzzy c-means clusters.

In the next section, a detailed account of the databases employed for the calibration and validation of the maps and cluster sets is provided. The results and a discussion of these, respectively, are presented in the subsequent two sections.

5.2.1 Databases

The calibration sets of data are kept structurally simple to ease the determination of Kohonen maps and fuzzy c-means cluster sets. The Figures 5.2a and 5.2b are typical corrupted sequences of data generated for the calibration sets for shifts and trends, respectively. Whether the data are real or synthetic, a linear normalization is always performed on the data so that a sequence of points ranges between 0 and 1. This normalization ensures that Kohonen maps and cluster sets can be used regardless of the original scale of the data points. In the case of a shift, the sequence of points follow a straight line at some level, and the evolution is broken at some location of the sequence, that is, at the location of the shift. Following that, the points are still on a straight line, but at a different level (see Figure 5.2a). In the case of a trend, the sequence of points follows a straight line with either a positive or negative slope (see Figure 5.2b). The magnitude of the step between lines for shifts and the magnitude of the slope of the straight line for trends determine the amplitude of the corruption. A large step or a steep slope is representative of a shift or a trend of high amplitude, respectively. A small step or a mild slope is representative of a shift or a trend of small amplitude, respectively. In both cases, shifts and trends, when there is no corruption present, the sequence of points follow a straight line at level 0.5. This results in a difference of levels or a slope equal to zero.

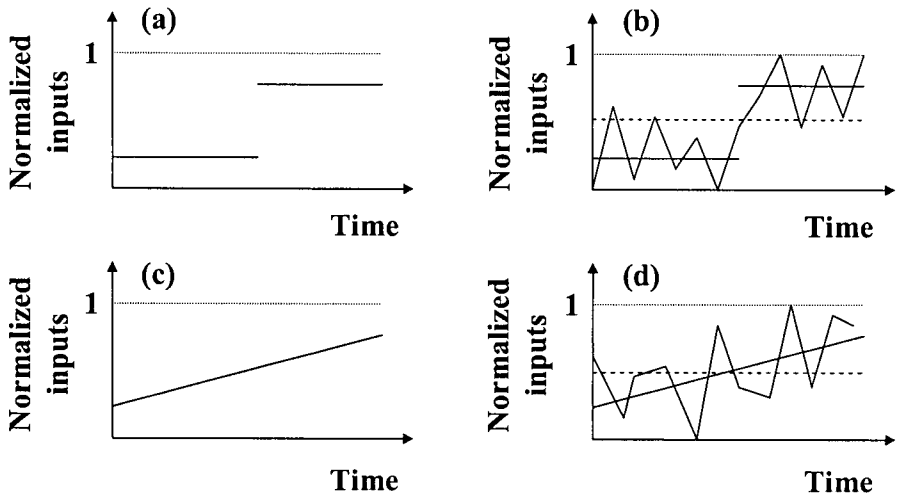


Figure 5.2. Calibration and validation sequences for shifts and trends.

Preliminary results show that such simple calibration data sets, where no variability is present aside from that caused by the corruption, lead to more reliable Kohonen maps and fuzzy c-means cluster sets, compared with those from calibration sets where variability other than that caused by the corruption is also present. The validation data sets are made of synthetic sequences that are meant to be as close to real data as possible, and this implies that variability other than that caused by corruption must be present, as exhibited by the continually broken lines in Figures 5.2c and 5.2d. These two figures also demonstrate the validity of the calibration sets. Consider the case of a real sequence of data affected with a shift, which after normalization is compared with a continuous straight line at level 0.5 (representing an uncorrupted sequence in the calibration sets) and with a straight line broken at some location (representing a corrupted sequence in the calibration sets). This is the case presented in Figure 5.2c. It can be observed from this graph that the sum of the squared differences between the real case and the broken straight line case would be smaller than the sum of the squared difference between the real case and the continuous straight line. An output neuron on the Kohonen map or a cluster in fuzzy c-means is considered as activated if, among all output neurons or clusters, it yields the smallest sum of squared differences between the inputs and the weights associated with this neuron or cluster. Therefore, a neuron or a cluster that characterizes the broken straight line would be activated rather than the neuron or cluster characterizing the continuous straight line. This

argument can also be made for the case of trends, using Figure 5.2d. This explains why the calibration data sets can lead to the production of Kohonen maps or fuzzy c-means cluster sets capable of detecting shifts or trends in real data sequences or synthetic data sequences designed to represent reality.

The structure of the databases employed here is similar that presented in Chapter 4, and again generated with MATLAB. The performance of the maps and cluster sets is evaluated for cases that represent variations in (1) the length of the data sequences, (2) the coefficient of variation of the individuals in the sequences, and (3) the amplitude of the shifts or trends imposed on the data sequences. The calibration database for univariate cases is made of six batches of ten sets of data sequences. Three batches are for cases of shifts, that is, one for the 30, 40 and 50-individual sequences, respectively. Similarly, three batches are used in the case of trends, one for the 30, 40 and 50-individual sequences, respectively. Each set of sequences is employed to calibrate one Kohonen map and one set of fuzzy c-means clusters. Therefore, a group of ten maps and ten cluster sets is calibrated for each batch. The size of the sets vary with respect to the length of the data sequences so as to be in agreement with the rule regulating the data requirement for the calibration of the weights of the output neurons and clusters. Kohonen maps of 10×10 output neurons as well as fuzzy c-means cluster sets of 100 clusters are considered, leading to the calibration of 3,000, 4,000 and 5,000 weights for each map and cluster set, for cases of 30-, 40- and 50-individual sequences, respectively. The number of weights must be 20% of the number of data sequences, and this implies that the number of sequences per set for cases of 30, 40 and 50-individual sequences must be 15,000, 20,000 and 25,000, respectively. The calibration database for multivariate cases is also made of six batches of sets, three for shifts and three for trends, representing cases of 30-, 40- and 50-individual sequences. There is a total of ten groups of five sets per batch. Each group of sets is used to calibrate one Kohonen map and one set of fuzzy c-means clusters. Again, a group of ten maps and ten cluster sets is calibrated for each batch. Two sets of sequences per group are employed for cases of two variates, while the entire five sets of sequences are used for cases of five variates. For cases of 30, 40 and 50-individual sequences, the number of weights to calibrate either a map or a cluster set is 6,000, 8,000 and 10,000, respectively, for the cases of two variates. The number of weights is 15,000, 20,000 and 25,000, respectively, for the

cases of five variates. A number of sequences per set of 30,000, 40,000 and 50,000 is considered, respectively for cases of 30-, 40- and 50- individual sequences. In all calibration sets, for univariate and multivariate cases, 50% of the sequences are corrupted (i.e., magnitude of levels for shifts and slope for trends that are different from zero).

The sequences in the validation database have characteristics that are similar to those of the sequences used in Chapter 4. The mean and the coefficient of variation of the data sequences vary uniformly between 1 and 20,000 and between 0.05 and 0.5, respectively. Each point in a sequence is created randomly, following a Normal distribution. For multivariate cases, if a sequence is corrupted, then the corresponding sequences in the associated sets of a group are also corrupted. For cases of shifts, the amplitude is chosen randomly, following a Uniform distribution, and can be as much as ±25% of the mean of the sequence prior to the shift. The location of the shift is determined randomly, following a Uniform distribution, and can be anywhere in the sequence except within the first and last five individuals. The location of the shift, when there is one, is the same for all corresponding sequences in the associated sets (i.e., for the other variates) of a group. For tests of trends, the amplitude of the increase or decrease of the mean is chosen randomly, following a Uniform distribution, and can be as much as ±0.5% of the initial mean per time step. The direction of the trend, when there is one, is the same, having either a positive or negative slope, in all corresponding sequences in the associated sets of a group. The amplitude of the corruption, for cases of both shifts and trends, varies within corresponding sequences. For univariate cases, a total of 6 sets of sequences are produced, three for shifts and three for trends, for the 30, 40 and 50-individual sequences, respectively. For multivariate cases, six groups of five sets, covering shifts and trends, and their variability in the length of the sequences, are produced. Each set is used to validate the corresponding group of ten maps and ten cluster sets obtained from the calibration process. The number of sequences per set is 50,000 for all cases.

5.2.2 Results

Corrupted Versus Uncorrupted Data

When the calibration data sets are employed to validate the Kohonen maps or the fuzzy c-means cluster sets the division between uncorrupted and corrupted sequences is

very clear. All the uncorrupted sequences activate one specific output neuron or cluster, while the corrupted sequences activate the other output neurons or clusters. A few corrupted sequences activate the output neuron or cluster dedicated to the uncorrupted sequence, the worst case being 0.2% of the corrupted sequences falsely detected as uncorrupted. These few corrupted sequences represent cases of corruption (i.e., shifts or trends) of very small amplitude and simply indicate that the resolution of the maps and the cluster sets is not fine enough to properly identify these corrupted sequences. Larger maps or cluster sets than those employed here would have allowed a finer discrimination of the possible patterns present in the sequences, making it possible to improve the identification of sequences with corruption of small amplitude. The issue of resolution must be noted, especially in the cases of shifts, where the purpose is not only to detect the shift, but also to locate it. Compared with cases of trends, the number of patterns to examine in the cases of shifts increases by a power of two.

When the validation data sets are employed to check the performance of the Kohonen maps or the cluster sets, the division between uncorrupted and corrupted sequences is not clearly defined. A typical result for a Kohonen map is presented in Figure 5.3. This map has been calibrated for the detection of shifts in 30-individual sequences. Figure 5.3a and 5.3b provide the same map, illustrated in the form of an array of values, with each element representing one output neuron. The number in each element of the arrays is the number of sequences, uncorrupted in Figure 5.3a, and corrupted in Figure 5.3b, activating the neuron associated with that element. The data in Figure 5.3a show that the uncorrupted sequences activate neurons in a particular region of the array (i.e., the map). It is assumed that the neurons that are not activated in Figure 5.3a, shaded with the number in bold in the figure, are only meant to be activated by corrupted sequences, and indeed, these neurons are shown to be activated in Figure 5.3b by some corrupted sequences. The non-shaded neuron with the number in bold and in italics in Figure 5.3 is the one that is activated by the uncorrupted sequences in the calibration sets.

0	0	85	0	0	0	0	0	0	0
0	0	150	48	0	0	0	0	0	0
0	0	8	1325	198	47	7	0	0	0
0	0	12	774	1732	491	113	7	0	0
0	0	10	286	2028	1233	472	21	0	0
0	0	17	521	940	827	677	125	0	0
0	0	11	319	865	306	783	539	221	39
0	0	3	175	733	472	224	398	1322	1516
0	0	3	85	673	291	162	6	83	478
0	0	0	103	1571	1323	3	*6*	2	77

(a)

0	8	60	18	0	2	1	0	2	0
0	27	277	210	93	67	75	65	13	0
0	18	157	677	710	493	383	196	35	2
0	21	164	777	1122	926	684	245	31	1
0	24	246	645	928	809	748	298	10	0
0	75	379	767	507	387	552	363	92	13
0	56	303	701	573	115	386	424	347	165
2	62	273	607	605	238	97	197	641	947
0	16	174	408	594	160	71	2	30	180
0	1	47	340	1253	596	3	*2*	2	33

(b)

Figure 5.3. Kohonen map with (a) uncorrupted sequence, and (b) corrupted sequences.

The data in Figure 5.3 show that a few corrupted sequences can be detected as such, because they activate neurons in the shaded areas. The bulk of the corrupted sequences, however, activate neurons that are also activated by uncorrupted sequences. Table 5.1 presents the success rate of the Kohonen maps and the fuzzy c-means cluster sets for identifying corrupted sequences based on the differentiation between the shaded and non-shaded neurons. The numbers in Table 5.1 are the average success rates from the group of ten maps or ten cluster sets, and there is very little variability from one map or cluster set to another. The success rate for all cases never exceeds 10%. Combined with a 100% success rate for identifying uncorrupted sequences, this translates to a false detection rate of both corrupted and uncorrupted sequences of 46% in the best case (i.e., fuzzy c-means applied to 50-individual sequences corrupted with trends) and of close to 50% in the worst case (i.e., Kohonen network applied to 30-individual sequences corrupted with trends).

Table 5.1. Success rate in identifying corrupted sequences.

Corruption	Sequence size	Success rate (%)	
		Kohonen	Fuzzy c-means
Shifts	30	3.86	3.14
	40	5.13	4.33
	50	6.05	4.42
Trends	30	0.54	0.57
	40	2.95	3.06
	50	7.87	8.32

Differentiating uncorrupted from corrupted sequences based on a strict separation of regions on the Kohonen maps or of cluster sets is not attractive. As shown in the data in Figure 5.3, for example, the neuron in the third column from the left, second row from the bottom is activated by only 3 uncorrupted sequences and by as much as 174 corrupted sequences. Considering this neuron as a detector of corrupted sequences would have a very small cost in terms of false detection of uncorrupted sequences for and the benefit would be great in terms of adequately detecting corrupted sequences (i.e., a presumed ratio of cost to benefit of 3 to 174). The calibration data sets are designed to develop maps or cluster sets that can provide estimates of the amplitude of the corruptions, either shifts or trends, and the detection approach can be based on this feature. This implies the determination of threshold values applied on the estimates of amplitude of corruption to differentiate between presumed uncorrupted and corrupted sequences.

Setting the Threshold Values

Because the properties of the sequences in the validation sets are all known, it is easy to determine the average properties of the sequences activating some given neuron or cluster, and this for all neurons and clusters. The criterion chosen for the separation between uncorrupted and corrupted sequences is the ratio of the amplitude of the corruption over the coefficient of variation of the sequences (i.e., Amp/CV, for shifts or trends), as described and employed in Chapter 4. Thus all neurons on a map or all clusters in a set are characterized based on the average ratio of amplitude to CV of their associated sequences, and a typical result is that of Figure 5.4. This figure shows the evolution of the Amp/CV ratio over a Kohonen map established for the detection of shifts in 30-individual sequences. The axes in Figure 5.4 represent the output neuron number, and the scale on the right

indicates the range of the Amp/CV ratio. In Figure 5.4, the upper left corner of the map is activated by sequences corrupted by a positive shift, leading to a sudden increase in the mean of the sequence after the location of the corruption (i.e., positive values of the Amp/CV ratio, from 0 to 3), while the lower right corner of the map is activated by a negative shift (i.e., negative values of the Amp/CV ratio, from 0 to -3). Of course, uncorrupted sequences have an Amp/CV ratio equal to zero, and should normally activate neurons in the middle part of the map, on the right side in Figure 5.4.

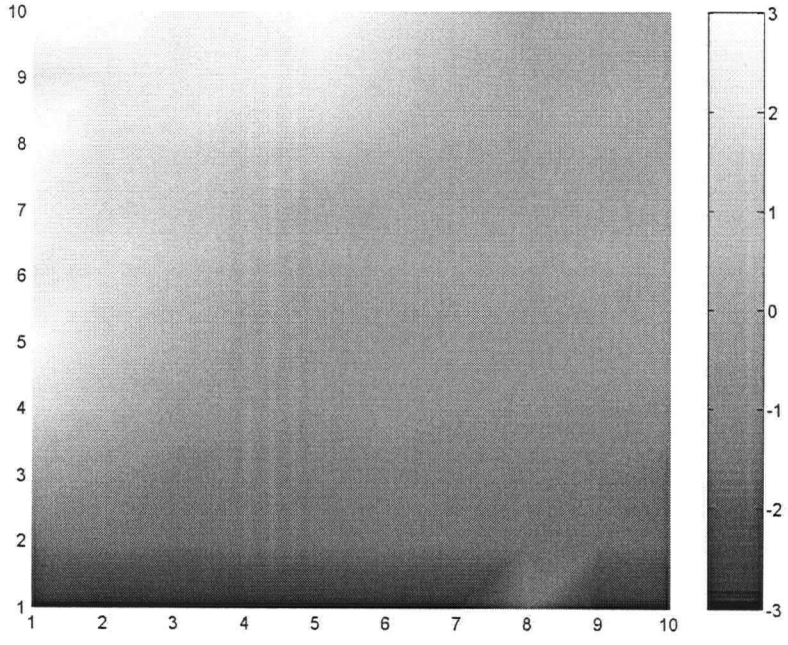


Figure 5.4. Ratio Amp/CV for shift with 30-individual sequences.

The aggregation of the results from a group of maps or clusters for each type of corruption and length of the sequences is undertaken in this work. A group of ten maps and one of ten fuzzy c-means cluster sets have been calibrated in the attempt to reduce the uncertainties from the calibration processes. The aggregation consists of evaluating the average of the estimates of the Amp/CV ratio from all the maps or cluster sets in a given group. The accuracy of the maps and cluster sets in determining the Amp/CV ratio, expressed as the root of the mean squared errors between the actual and estimated ratio (RMSE), is given in Tables 5.2 and 5.3, for cases of shifts and trends, respectively. The RMSE is established for each map or cluster set, and the values in the table give the

average of the RMSE (i.e., average of maps or cluster sets alone in the tables). The RMSE is also determined from the results of the aggregation of a group of maps or cluster sets (i.e., aggregation of maps or cluster sets in the tables). In Tables 5.2 and 5.3, the improvement in the RMSE provided by the aggregation of maps or cluster sets compared with the corresponding average of maps or cluster sets considered alone is also presented. Table 5.2 indicates that the estimates of the Amp/CV ratio differ from the actual ratio by 0.4 on average, this for a ratio for which the range in absolute value goes from 0 to 3 for the validation data sets employed. The results in Table 5.2 show that an improvement in the estimates of the ratio of around 6% is provided by the aggregation of maps or cluster sets compared with the strategy of considering only one map or cluster set. Regardless of the performance obtained, the AIT-based approach proposed in this chapter is beneficial from the point of view of being able to determine an estimate of the amplitude of a shift in sequences. And from a practical point of view, the Kohonen maps and fuzzy c-means cluster sets are very easy to employ, once calibrated. The amplitude of a shift can be inferred from the results of conventional statistical detection tests, although such tools are not intended for that purpose. Other than the methods presented here, only Bayesian analysis as developed by Lee and Heghinian (1977) and Perreault et al. (1999 and 2000) could be applied for the estimation of the amplitude of a shift.

Table 5.2. Estimated versus actual ratio Amp/CV for univariate cases of shifts.

Technique	Sequence size	RMSE for the		Improvement (%) of aggregation relative to non- aggregation
		Average of maps or cluster sets alone	Aggregation of maps or cluster sets	
Knet	30	0.402	0.382	5.05
	40	0.385	0.363	5.80
	50	0.381	0.355	6.91
Fuzzy	30	0.425	0.401	5.77
	40	0.412	0.385	6.35
	50	0.408	0.381	6.67

For the determination of the amplitude of a trend, a commonly used estimator is that proposed by Sen (1968), expressed as:

$$A = median\left(\frac{x_j - x_i}{j - i}\right) \text{ for all } j < i \qquad 5.1$$

where A is the amplitude of the trend, and the xs are individuals in the data sequence. Given the nature of the validation data sets employed, the Amp/CV ratio is obtained when the estimator in Equation 5.1 is divided by the standard deviation of the sequence. In Table 5.3, the performance of this estimator, shown in the third column, is compared with that of the maps and cluster sets. For this Median estimator, the difference between the estimated and actual ratio ranges from 0.016 in the worst case to 0.014 in the best case on the average. For the case of trends, the absolute values of the Amp/CV ratio range from 0 to 0.1 for the validation data sets employed. The worst improvement obtained with the maps and cluster sets is 0.013 while the best improvement is at 0.008. There is very little improvement from the aggregation of maps or cluster sets relative to the strategy of considering only one map or cluster set. The improvement is, however, noticeable with the aggregation of maps or cluster sets relative to the Median estimator. Improvement increases as the length of the sequences increases, which is normal since it is easier to detect trends on long sequences than on short ones.

Table 5.3. Estimated versus actual ratio Amp/CV for univariate cases of trends.

Technique	Sequences Size	RMSE for the			Improvement (%) of aggregation relative to	
		Median	Average of maps or cluster sets alone	Aggregation of maps or cluster sets	the Median	non-aggregation
Knet	30	0.0161	0.0127	0.0127	20.68	0.10
	40	0.0151	0.00982	0.00981	35.22	0.13
	50	0.0143	0.00767	0.00766	46.48	0.15
Fuzzy	30	0.0161	0.0127	0.0127	20.67	0.06
	40	0.0151	0.00982	0.00981	35.21	0.08
	50	0.0143	0.00767	0.00766	46.47	0.10

The results in Tables 5.2 and 5.3 are an indicator of the capacity of the maps and cluster sets to differentiate between uncorrupted and corrupted sequences based on the Amp/CV ratio. The smaller the RMSEs, the smaller the likelihood of false detection becomes. The detection strategy employed here is identical to the one employed in Chapter 4, and it is the determination of threshold values within the range of the Amp/CV ratio that minimize the risk of false detection. Both uncorrupted sequences falsely detected as corrupted and corrupted sequences falsely detected as uncorrupted are given equal weight in the minimization procedure, for it is as detrimental to perform either one of these

misdiagnosis. The results of this minimization procedure applied to the maps and cluster sets are given in Tables 5.4 and 5.5, for cases of shifts and trends, respectively. For comparison-sake, the Mann-Whitney and Student's tests for shifts and the Mann-Kendall and Spearman tests for trends are applied to the validation data sets. In the case of these tests, the minimization procedure for false detection is also applied. In Tables 5.4 and 5.5, column 1 gives the detection techniques employed, column 2 provides the threshold values, column 3 (U) lists the ratio of the number of uncorrupted sequences falsely detected as corrupted over the total number of sequences, column 4 (C) lists to the ratio of the number of corrupted sequences falsely detected as uncorrupted over the total number of sequences, and column 5 (U+C) gives the sum of columns 3 and 4. In Tables 5.4 and 5.5, the Us can be added to the Cs directly, because there is an equal proportion of uncorrupted and corrupted sequences in the database. The results shown in Tables 5.4 and 5.5 for the conventional statistical detection tests are similar to those presented in Tables 4.2 and 4.3, indicating the similarities of the data sets employed here and in Chapter 4. The results shown in Tables 5.4 and 5.5 for the Kohonen maps and the cluster sets come from the aggregation of the maps and cluster sets.

Table 5.4. False detection ratio for univariate cases with shifts.

Case	Optimal			
	Threshold	False detection ratio (%)		
		U	C	U+C
(a) 30-individual sequences				
MW	2.18	9.1	22.1	31.3
ST	2.38	8.3	22.5	30.8
Knet	0.268	9.8	22.1	31.9
Fuzzy	0.316	8.2	25.0	33.2
(b) 40-individual sequences				
MW	2.32	7.9	21.3	29.2
ST	2.46	8.0	20.9	28.8
Knet	0.302	6.7	23.5	30.2
Fuzzy	0.302	7.8	24.6	32.4
(c) 50-individual sequences				
MW	2.46	6.5	21.3	27.9
ST	2.60	6.4	21.0	27.5
Knet	0.290	7.2	22.0	29.3
Fuzzy	0.295	7.3	25.0	32.2

For the cases of shifts, as in Chapter 4, it appears again that the conventional statistical detection tests for shifts perform slightly better than the AITs (see Table 5.4), and the Kohonen maps also perform better than the fuzzy c-means cluster sets. The false detection ratio is essentially 30% regardless of the sequence length. For the cases of trends (see Table 5.5), as in Chapter 4, the AITs perform slightly better than the conventional statistical detection tests for trends. The false detection ratio is about 40% for 30-individual sequences, and drops to 30% for 50-individual sequences. In all cases, shifts and trends, the ratio of uncorrupted sequences falsely detected as corrupted is always smaller than the ratio of corrupted sequences falsely detected as uncorrupted.

Table 5.5. False detection ratio for univariate cases with trends.

Case	Optimal			
	Threshold	False detection ratio (%)		
		U	C	U+C
(a) 30-individual sequences				
MK	1.18	11.3	30.4	41.8
SP	0.21	12.8	28.8	41.6
Knet	0.005	15.0	26.4	41.5
Fuzzy	0.006	11.9	29.6	41.5
(b) 40-individual sequences				
MK	1.39	7.9	28.5	36.4
SP	0.22	8.6	27.8	36.4
Knet	0.006	10.7	25.4	36.2
Fuzzy	0.007	9.2	27.0	36.2
(c) 50-individual sequences				
MK	1.42	7.7	23.2	30.9
SP	0.21	7.7	23.1	30.8
Knet	0.006	7.9	22.7	30.6
Fuzzy	0.006	7.4	23.3	30.7

Finding the Location of the Shift

As with the Amp/CV ratio, the location of the shift can be characterized on the maps or the cluster sets, leading to the production of maps similar to that illustrated in Figure 5.4. The accuracy of the maps and cluster sets for determining the location of the shift can be evaluated with the RMSE comparing the estimated with actual location of the shift. The RMSE obtained with the maps and cluster sets considered alone as well as aggregated are presented in Table 5.6. Also shown are those obtained with the Mann-

Whitney and Student's tests for comparison-sake. The RMSE increases with all conventional and AIT-based detection tests as the length of the sequences increases. This is explained by the fact that, as the length of the sequences increases, the number of potential locations of the shift increases, and accordingly the risk of missing the exact location of the shift also increases. The AITs provide more accurate estimates of the location of the shift than the conventional statistical detection tests. The conventional tests yield RMSEs from 6 for the 30-individual sequences case to 12 for 50-individual sequences case, while the RMSE ranges from 5 to 10 in the case of the AITs. The last three columns of Table 5.6 show the improvement obtained with the aggregation of maps relative to the conventional tests and the maps and cluster sets considered alone. The improvement due to the aggregation ranges from 15 to 22% when compared with the conventional tests. There is no significant improvement due to the aggregation compared with the maps and cluster sets considered alone.

Table 5.6. RMSE for the location of the shift with univariate cases.

Technique	Sequence size	RMSE				Improvement (%) of aggregation relative to		
		MW	ST	Average of maps or cluster sets	Aggrega-tion	MW	ST	non-aggregation
Knet	30	6.42	6.41	5.00	4.94	22.97	22.88	1.11
	40	9.39	9.34	7.66	7.59	19.22	18.79	1.00
	50	11.96	11.86	10.18	10.05	15.97	15.23	1.28
Fuzzy	30	6.42	6.41	5.03	4.97	22.50	22.40	1.07
	40	9.39	9.34	7.70	7.61	18.93	18.50	1.14
	50	11.96	11.86	10.26	10.14	15.23	14.49	1.15

The results in Table 5.6 do not include the use of threshold values. The RMSE are those obtained from corrupted sequences without consideration as to whether the sequences would be detected or not as corrupted based on the threshold values. The data in Table 5.7 accounts for the use of the threshold values. Columns 2 through 7 list the success rates for identifying the location of the shift exactly and up to as much as plus or minus five time steps from the corruption, in sequences detected as corrupted.

Table 5.7. Success rate in identifying the location of the shift with univariate cases.

Case	Success rate (%) in identifying the location of the shift at plus minus					
	0 time step	1 time step	2 time steps	3 time steps	4 time steps	5 time steps
(a) 30-individual sequences						
MW	20	32	38	41	44	46
ST	21	32	38	42	44	46
Knet	4	14	24	32	38	43
Fuzzy	5	14	23	29	35	39
(b) 40-individual sequences						
MW	19	31	37	41	43	45
ST	20	32	38	42	44	46
Knet	4	11	18	25	30	34
Fuzzy	3	11	19	24	28	32
(c) 50-individual sequences						
MW	20	31	36	40	43	45
ST	21	32	37	41	44	46
Knet	2	8	16	21	26	30
Fuzzy	2	10	16	21	25	28

The success rates of the Kohonen maps and cluster sets in identifying exactly the location of the shift is close to zero, due to the rather coarse resolution of the maps and cluster sets. Essentially, in the determination of the values of the weights, the calibration procedures employed for the Kohonen maps and fuzzy c-means cluster set only lead to a blurred rendition of the location of the shifts. These calibration procedures accomplish a weighted sum of the input vectors fed to the output neurons or cluster sets, and the strongly weighted vectors for a given neuron or cluster may include several different locations, thus causing the blurring. The success rates for AITs increase as less precision is allowed with respect to the identification of the location of the shifts. The AIT-based tests almost have the same success rates as those of the conventional statistical detection tests in identifying the location of the shifts at plus or minus five time steps, with 30-individual sequences. The Kohonen maps and the cluster sets provide inferior success rates to those of the conventional tests, although the success rates of the AITs eventually reach and, in the case of the Kohonen maps, surpass those of the conventional tests when the location of the shift may be determined at more than plus or minus five time steps.

The conclusions that may be made based on an analysis of the results of both Tables 5.6 and 5.7 are that the AITs provide more accurate estimation of the location of the shift,

but that the use of threshold values to separate between uncorrupted and corrupted sequences is detrimental. Many corrupted sequences for which the AITs yield reasonably accurate estimates of the location of the shift are not considered in the calculation of the success rate presented in Tables 5.7 because they are considered as uncorrupted based on the threshold values. Yet some benefit may be obtained from the use of AITs based on the results in Table 5.6. One approach might be to use the conventional statistical detection tests for shifts to differentiate between uncorrupted and corrupted sequences, and then to use the Kohonen maps or clusters to estimate the location of the shift on the sequences detected as corrupted. The results of Tables 5.6 and 5.7 also highlight the finding that all of the detection methods, conventional and AITs alike, do not differentiate between uncorrupted and uncorrupted sequences in exactly the same manner. For example, there may be a significant number of sequences that are detected as corrupted by one method but not by the others. However, there are a large number of sequences that are similarly diagnosed (i.e., assumed uncorrupted or corrupted) by all methods, and this observation is demonstrated in the application of these methods to actual hydrometric data in Chapter 6.

Multivariate Cases

The multivariate cases require a much larger number of weights to calibrate for the Kohonen maps and fuzzy c-means clusters than the univariate cases, and this requirement ultimately affects the detection performance. The multivariate cases imply a larger number of patterns to differentiate than the univariate case, and this accentuates the issue of resolution. Also, the characterization of the maps and cluster sets based on the Amp/CV ratio is not as clear as with univariate cases. Unlike the univariate cases, where the characterization is based on the Amp/CV ratio of each of the sequences, the average of the Amp/CV ratio of corresponding sequences is the factor used in characterizing the output neurons or clusters for the multivariate case. All of these issues must be considered in the analysis of the results. Only the Kohonen network is tested with multivariate cases, as this AIT consistently provides better results than the fuzzy c-means for the univariate cases.

Tables 5.8 and 5.9, for shifts and trends, respectively, provide the RMSE between estimated and actual average values of the Amp/CV ratio, and therefore can be compared with Tables 5.2 and 5.3. In Tables 5.8 and 5.9, the terms Knet2 and Knet5 respectively

refer to the cases where two and five variates are tested at once. For shifts (Tables 5.8), the multivariate 30-individual cases show improvement in the estimation of the Amp/CV ratio compared with the univariate 30-individual cases. Deterioration of the quality of the estimates of the Amp/CV ratio is observed with multivariate 40- and 50-individual cases compared with the corresponding univariate cases. It must also be noted that the aggregation of the maps provides better results than the maps considered alone, and the improvement is higher with multivariate cases than it is with univariate cases (e.g., 5 to 6% improvement from non-aggregation to aggregation in the univariate cases of 30-individual sequences versus 10% improvement for multivariate cases of 30-individual sequences).

Table 5.8. Estimated versus actual ratio Amp/CV for multivariate cases of shifts.

Technique	Sequence size	RMSE		Improvement (%) from average of maps alone
		Average of maps alone	Aggregation of maps	
Knet2	30	0.331	0.298	9.98
	40	0.479	0.431	10.15
	50	0.486	0.440	9.34
Knet5	30	0.253	0.210	17.14
	40	0.496	0.460	7.27
	50	0.478	0.430	10.00

For trends (Table 5.9), all multivariate cases constitute an improvement from their corresponding univariate cases in the estimation of the Amp/CV ratio. The amplitude of the improvement is the highest for cases with short sequences (i.e., 30-individual sequences), and consistently reduces for cases with increasing lengths of sequences. There is a benefit to using the aggregation of the maps rather than the maps alone, a benefit that is more noticeable with multivariate cases than it is with univariate cases.

Table 5.9. Estimated versus actual ratio Amp/CV for multivariate cases of trends.

Technique	Sequence size	RMSE		Improvement (%) from average of maps alone
		Average of maps alone	Aggregation of maps	
Knet2	30	0.0106	0.0104	2.03
	40	0.0104	0.0101	2.29
	50	0.0087	0.0084	3.78
Knet5	30	0.0087	0.0080	8.30
	40	0.0089	0.0082	7.39
	50	0.0073	0.0066	10.06

Table 5.10 lists the false detection ratios obtained with all multivariate cases with shifts, from 30- to 50-individual sequences, with the Kohonen maps structured to test 2 or 5 sequences at once. The improvement in the estimation of the Amp/CV ratio with Kohonen maps for multivariate 30-individual cases clearly translates in a reduction of the false detection ratio, as shown by the results in Table 5.10. The false detection ratio for multivariate 30-individual cases is significantly lower than any other false detection ratio associated with 30-individual sequences corrupted with shifts, including the multivariate cases analyzed in Chapter 4 (see Tables 4.5 and 5.4). The results shown in Table 5.10 for multivariate 40- and 50-individual cases are more ambivalent. The multivariate cases with 40-individual sequences do not yield any improvement from the univariate cases. The cases with 50-individual sequences involving two variates tested at once with the Kohonen map are not any better than the univariate cases (see Table 5.4). The case where five variates are tested at once shows a small improvement compared with univariate cases (see Table 5.4), but is not as good as the multivariate case with five sequences tested at once applied in Chapter 4 (see Table 4.5).

Table 5.11 provides the same information as Table 5.10, but for cases with trends. Aside from the Kohonen maps, the Mann Kendall test can also be applied to cases with more than one variate at once, and therefore its performance on the calibration sets is also given in Table 5.11. In Table 5.11, MK2 represents the Mann Kendall test applied to two sequences of variables at once, while MK5 is for the Mann Kendall test applied to five sequences at once. The improvement in the estimation of the Amp/CV ratio with Kohonen maps for multivariate cases with trends for all sequence lengths reflects an improvement of the use of the maps compared with any univariate cases, whether conventional tests or AITs are employed (see Table 5.5). The Kohonen maps applied to multiple 30-individual sequences perform better than the Mann Kendall test applied on the same multiple sequences. However, the Mann Kendall test performs better than the Kohonen maps for the cases with 40 and 50-individual sequences.

Table 5.10. False detection ratio for multivariate cases with shifts.

Case	Optimal			
	Threshold	False detection ratio (%)		
		U	C	U+C
(a) 30-individual sequences				
Knet2	0.295	7.1	15.7	22.8
Knet5	0.290	4.0	6.8	10.8
(b) 40-individual sequences				
Knet2	0.161	11.0	19.3	30. 3
Knet5	0.089	9.1	22.5	31.6
(c) 50-individual sequences				
Knet2	0.202	8.5	22.8	31.4
Knet5	0.150	10.1	16.2	26.3

Table 5.11. False detection ratio for multivariate cases with trends.

Case	Optimal			
	Threshold	False detection ratio (%)		
		U	C	U+C
(a) 30-individual sequences				
MK2	1.20	11.8	25.4	37.2
MK5	1.27	10.9	18.0	28.9
Knet2	0.006	11.6	24.8	36.4
Knet5	0.008	8.4	17.5	25.9
(b) 40-individual sequences				
MK2	1.40	7.9	22.1	30.0
MK5	1.48	7.7	10.7	18.4
Knet2	0.007	13.7	20.9	34.6
Knet5	0.008	9.5	15.5	25.0
(c) 50-individual sequences				
MK2	1.65	6.9	16.2	23.1
MK5	1.72	4.5	6.6	11.1
Knet2	0.008	8.0	19.7	27.7
Knet5	0.007	6.3	11.4	17.8

The capacity of the Kohonen maps to estimate the location of the shifts in the multivariate cases is expressed in terms of the RMSE in Table 5.12. As in Table 5.6, the results in Table 5.12 are established based on corrupted sequences only, whether or not the sequences are detected as corrupted based on the threshold values. The results in Tables 5.12 are in agreement with those in Tables 5.8 and 5.10, that is, there is a benefit to using the Kohonen maps in a multivariate setting for the 30-individual cases, but no benefit is observed for the 40 and 50-individual cases. The RMSEs for 30-individual sequence cases

are smaller in Table 5.12 (i.e., multivariate cases) than they are in Table 5.6 (i.e., univariate cases), and the opposite is true for the 40 and 50-individual sequence cases. The benefit of the aggregation of maps compared with the use of maps alone in multivariate cases is also greater with the cases of 30-individual sequences than it is with the cases of 40 and 50-individual sequences.

Table 5.12. RMSE for the location of the shift with mutlivariate cases.

Technique	Sequence Size	RMSE		Improvement (%) of aggregation relative to non-aggregation
		Average of maps alone	Aggregation of maps	
Knet2	30	4.72	4.55	3.61
	40	8.22	8.16	0.74
	50	11.15	11.08	0.65
Knet5	30	4.19	3.80	9.38
	40	8.32	8.29	0.31
	50	11.24	11.21	0.30

Table 5.13 shows the success rates in identifying the location of the shift in multivariate cases when the use of the threshold values is taken into account, as it is in Table 5.7 with univariate cases. Due to the resolution issue mentioned in this section, the results of the approach for identifying the location of the shift exactly or quite closely with Kohonen maps are poorer than those for the conventional tests. As more flexibility is allowed in the identification of the location of the shift (e.g., the location of a shift at plus or minus four or five time steps), the Kohonen maps emerge as a better option than other methods (see Table 5.7) for the cases of the 30-individual sequences. For the cases of the 40 and 50-individual sequences, the use of the Kohonen maps in a multivariate setting is definitely a poor choice.

The use of the Kohonen maps in a multivariate setting shows promise, as attested by the results obtained for cases involving 30-individual sequences. The ambivalent results obtained with cases for the 40- and 50-individual sequences indicates that one must be prudent in the determination of the structure of the Kohonen network. As the length of the sequences increases, and as the number of variates tested at once increases, the number of patterns to be represented in the map also increases, particularly in the cases of shifts where both the amplitude of the shift and its location must be assessed. In this application, the number of output neurons has been kept constant at 100 for all cases due to computing

capability constraints. This number should normally be a function of the length of the sequences and also the number of sequences tested at once so as to ensure a constant resolution from one case to another. The number of iterations in the calibration process should also reflect the sequence size and whether univariate or multivariate cases may be assessed.

Table 5.13. Success rate in identifying the location of the shift with multivariate cases.

Case	Success rate (%) in identifying the location of the shift at plus minus					
	0 time step	1 time step	2 time steps	3 time steps	4 time steps	5 time steps
(a) 30-individual sequences						
Knet2	6	18	32	42	50	56
Knet5	7	26	45	59	69	75
(b) 40-individual sequences						
Knet2	2	5	10	16	21	25
Knet5	1	4	9	13	17	21
(c) 50-individual sequences						
Knet2	0	4	7	11	14	17
Knet5	1	4	8	12	15	19

5.2.3 Discussion

The results in the previous section give case by case indications of which conventional statistical tests and AITs provide the best opportunity for adequately detecting shifts or trends. For univariate cases of shifts, conventional tests appear better suited to diagnose the presence or absence of a shift, although AITs as employed in this chapter seem better suited to determine the location of the shift when there is one, regardless of whether this shift is properly diagnosed as such. With univariate cases of trends, all methods essentially provide the same detection performance, although at the limit a very slight edge can be granted to AITs. The overall conclusion is that all methods, whether for univariate cases of shifts or trends, yield rather equivalent detection performance and confirm each other in their validity. If all methods are sound and valid, then one might be led to concluding that there is a limit to the detection capacity when only one sequence or variate is available for testing. Depending on the properties of the sequences (i.e., variance, skewness, distribution, etc.), there are instances that cannot be properly detected because the resolution of the tests is not sufficiently fine. This highlights the importance of

concurrent sequences or supporting information to aid in the detection process. Multivariate methods of detection demonstrate that considering more than one sequence or variate leads to gains in detection performance, and therefore should be used to complement and enhance univariate methods when one is certain that sequences can be grouped together. The application here shows that AITs can be suitable to assess multivariate cases, and can provide enhanced results relative to conventional statistical tests.

5.3 Application on Outliers

In an attempt to address all types of anomalies discussed in this work (i.e., outliers, shifts and trends), preliminary detection tests based on Kohonen maps and fuzzy c-means cluster sets are developed here for the identification of outliers. In hydrology, unlike shifts and trends, which are usually evaluated on annual sequences, the presence of outliers is often assessed for short time-step records, such as daily sequences of hydrometric observations. The presence of outliers is an important issue among managers of water resources systems whether for the purpose of hydro-energetic production or flood control. Outliers can induce a bias in the calibration of inflow prediction models employed for short to long-term water resources management. On an operational basis, the real-time detection of outliers in observations is a significant issue, because a wrongful real-time decision based on erroneous data can have effects that could linger for a long period of time. Because of the nature of outliers, the procedure for the development of Kohonen maps and fuzzy c-means cluster is slightly different from that in Section 5.2 for shifts and trends. The procedure has similar steps to those in Section 5.2, which are 1) a calibration phase to build the maps and cluster sets and 2) a validation phase for the evaluation of the performance of the detection of outliers. Conventional outlier detection methods, based on gradients between points in data records, as described in Krajewsky and Krajewsky (1989) are also considered here in the validation phase for comparison-sake. Outliers involve the consideration of a much larger number of possible patterns, and this affects the construction of synthetic data sequences for the calibration and validation processes.

The experimental databases employed for the calibration and validation of the maps and cluster sets are presented in the following section. Next, the results of the experiment and a discussion of these results are presented.

5.3.1 Databases

Outliers are aberrances that disrupt the expected pattern of the data. If one considers a set of points following a straight horizontal line, and it is expected according to the underlying processes being investigated that these points follow a straight line, then a point in the set would be considered as an outlier if it is significantly above or below this straight line. This outlier in itself creates a pattern that differs from the one defined by the set of points when no outlier is present. The Kohonen maps and fuzzy c-means cluster sets constructed here are designed to differentiate between patterns where outliers are present and patterns where no outliers are present, given a continuous subset of an historical record of data. A subset of hydrometric data points can in itself be the source of many patterns. In a given subset of data points with one outlier present, several patterns may exist depending on the location of the outlier, and its amplitude and direction (i.e., positive or negative). The number of patterns increases more if more than one outlier is present in the subset of data points.

The calibration data sets employed here attempt to represent the diversity of patterns for cases where no outlier or, only one outlier, is present in the subset. Figure 5.5 illustrates a few examples of possible subsets of points in the calibration data sets. As in cases of shifts and trends, the range of values varies between zero and one for the purpose of reducing the scale of the problem. Real data would need to be standardized before being fed to the maps and cluster sets established with the calibration data sets employed here. The uncorrupted subsets are designed to represent expected patterns found in hydrometric observations, including continuous ascensions (e.g., the rise of the hydrograph, see Figure 5a), a continuous descent (e.g., the recession of the hydrograph), the various possibilities of ascension followed by a recession, and the various possibilities of recession followed by an ascension (e.g., see Figure 5b). Simple power functions, with power coefficient between 0.5 and 1.5 are employed to generate the data points in any given subset. The corrupted subsets must reflect the uncorrupted patterns, with the addition of an outlier. For corrupted subset, patterns may vary with respect to the amplitude of the outlier, its location (e.g., see Figures 5c and d), and its direction, either positive or negative (e.g., see Figures 5e and f). The amplitude of the outlier is determined by how far it is from the other data points in the

subset. The amplitude is determined randomly and can vary uniformly between 0 and 1. The number of possible patterns also varies with respect to the length of the subset. The longer the subset, the greater the number of patterns.

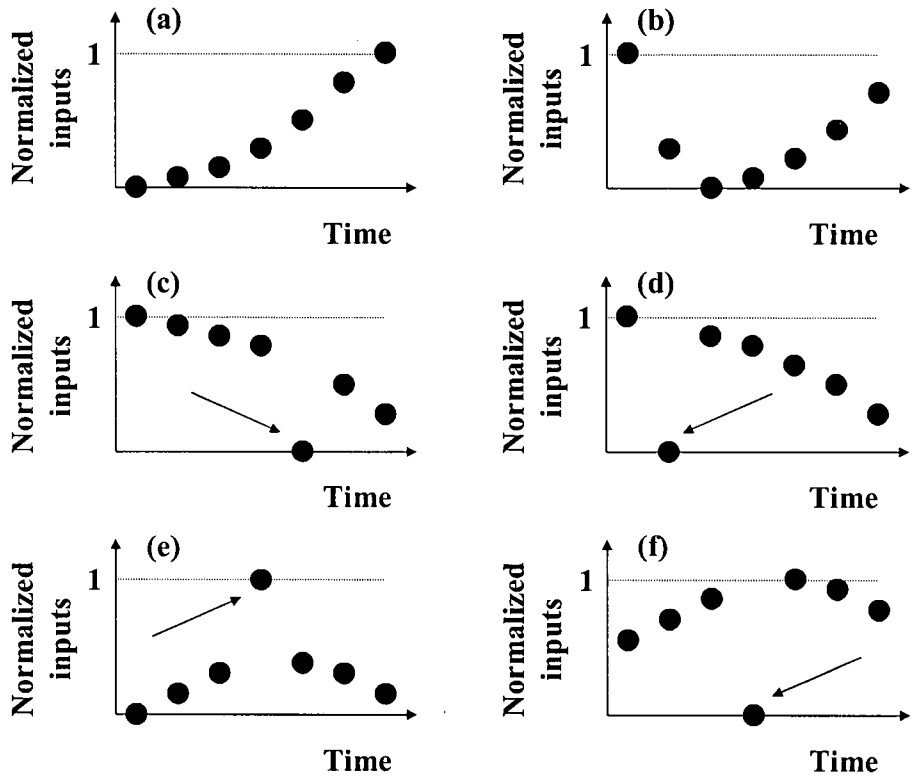


Figure 5.5. Examples of calibration sequences for outliers.

Maps and cluster sets are calibrated for subsets containing 5, 7 or 10 data points (e.g., 5 to 10 daily observations of inflows or water levels). Larger maps and cluster sets than those employed for shifts and trends are considered here for outliers because of the large number of possible patterns involved. A total of 225 neurons (15×15) and clusters per map and set, respectively, are considered for cases with 5- and 7-point subsets, while 324 neurons (18×18) and clusters per map and set are used for cases with 10-point subsets. This implies the calibration of 1,125, 1,575 and 3,240 weights per map and cluster set, respectively for the 5, 7 and 10-point subsets. As with shifts and trends, ten maps and cluster sets are calibrated for each case (i.e., 5, 7 and 10-point subsets) to attempt to reduce the uncertainties inherent to the calibration procedures. Therefore, 10 calibration sets of

subsets are created per case, the sets containing 11,000, 16,000 and 35,000 subsets, respectively, for the 5, 7 and 10-point subset cases. The length of the sets is such that the ratio of the number of weights to calibrate versus the number of subsets equals around 10%, as opposed to the 20% employed for shifts and trends. For all calibration data sets, an equal number of uncorrupted and corrupted subsets are created.

Unlike shifts and trends, it is not easy to build synthetic validation data sets for outliers that are presumed to be as close as possible to reality because of all the possible situations that arise with hydrometric data at a short time scale. To circumvent these difficulties, real data sets, presumed free of outliers, are used. They are then corrupted with outliers for which the properties are known. The chosen data are those coming from inflow observations at hydrometric stations on the Mistassibi, Harricana, and San Juan Rivers (i.e., Environment Canada station numbers 02RD002, 04NA001 and 08HA010, respectively, data taken from the HYDAT CD-ROM, version 96-1.04). The first two rivers are in the Province of Quebec, while the last one is in British Columbia on Vancouver Island. The size of the watersheds at the stations are 9,320, 3,680 and 580 km^2, respectively, and the distinct location of the stations ensures some diversity in terms of hydrologic behavior so as to demonstrate the versatility of the maps and cluster sets. A continuous 10-year record is taken from each of these 3 stations, and then these original records are duplicated 100 times each, allowing the construction of around one million subsets for all cases (i.e., 5, 7 and 10-point subsets). The production of a large number of subsets with MATLAB, similar to a Monte Carlo simulation, attempts to ensure that all possible patterns are represented. Before the subsets are built, the duplicated records are corrupted with outliers. A corruption is a value added to or subtracted from the original value of a point in the data sequence. Of the one million data points generated from the duplication, only 5% of them are corrupted in an attempt to have only one outlier per subset, when there is one present. The large size of the database, one million subsets, ensures nevertheless that a large number of outliers are available for investigation, that is, around 20,000. The outliers (i.e., the added or subtracted values) are designed to follow a Normal distribution of mean 0, and of standard deviation equal to that of the corresponding original record of data, and are added to the values of the data points.

5.3.2 Results

For comparison-sake, two simple outlier detection methods proposed in Krajewsky and Krajewsky (1989) are tested here, along with the Kohonen maps and fuzzy c-means cluster sets. The first method is based on the gradient between two data points, that is:

$$G_t = q_t - q_{t-1} \qquad 5.2$$

where G_t is the gradient at time t, and q_t and q_{t-1} are the data points (e.g., daily inflows) at time t and t-1. It is assumed that q_t is affected by an outlier if G_t is higher or lower than a given threshold value. The second method is a product of gradients (GP_t), that is:

$$GP_t = G_t G_{t+1} \qquad 5.3$$

where G_t and G_{t+1} are calculated with equation 5.2. It is assumed that q_t is affected by an outlier if GP_t is higher or lower than a given threshold value.

Krajewsky and Krajewsky (1989) do not commit to the determination of specific threshold values to distinguish between instances of corruption or non-corruption for their tests, denoted in this work as Test 1 for Equation 5.2 (i.e., the gradient) and Test 2 for Equation 5.3 (i.e., the gradient product), because of the unknown distributions of the responses of these tests. This is also problematic for the cases of the detection tests based on fuzzy c-means and the Kohonen network. As a result, the determination of threshold values for all tests are determined based on the minimization of false detection, as performed with shift and trends in Chapters 4 and 5. Distinct threshold values are determined for each river to see if there is any variability.

From figures similar to that of Figures 5.3 and 5.4, assessment of the capacity of maps or cluster sets to distinguish between uncorrupted and corrupted subsets can be performed. Average characteristics are calculated for all output neurons in a map or clusters in a set based on the properties of the subsets that activate these neurons and clusters. Here the chosen characteristics are the location of the outlier in the subset and the ratio of the amplitude of the outlier to the standard deviation of the data sets (Amp/SD). To facilitate comparisons, the values of Test 1 and Test 2 are respectively divided by the standard deviation (SD) and the variance (SD^2) of the data sets to correspond to the Amp/SD ratio employed with the maps and cluster sets.

Table 5.14 gives the optimal false detection ratios obtained with all tests, and in which T is the threshold, U is the number of uncorrupted data points falsely detected as

corrupted over the total number of uncorrupted data points, C is the number of corrupted data points falsely detected as uncorrupted over the total number of corrupted data points, Test 1 and Test 2 are already defined, and Knet and Fuzzy are the detection tests based on the Kohonen maps and fuzzy c-means applied to the 5, 7 and 10-individual subsets.

Table 5.14. False detection ratio for cases with outliers.

Test	Harricana river			Mistassibi river			San Juan river		
	T	U	C	T	U	C	T	U	C
Test 1	0.181	7.71	14.14	0.170	14.70	13.37	0.163	30.52	11.57
Test 2	0.176	0.68	32.20	0.223	2.30	36.14	0.514	9.42	49.09
Knet 5	0.269	10.39	18.65	0.035	8.81	17.14	0.132	14.52	33.15
Fuzzy 5	0.175	10.19	17.88	0.061	6.48	16.99	0.144	12.24	34.66
Knet 7	0.327	9.52	20.04	0.046	6.16	23.19	0.433	6.34	41.85
Fuzzy 7	0.266	11.17	19.77	0.066	5.21	21.59	0.224	18.12	37.37
Knet 10	0.262	9.17	29.22	0.074	4.78	35.03	0.445	10.95	48.92
Fuzzy 10	0.216	8.34	52.27	0.091	4.17	55.34	0.467	7.78	63.39

The results in Tables 5.14 must be interpreted in light of Table 5.15, which gives some basic characteristics of the river watersheds.

Table 5.15. Characteristics of the watersheds.

River	Watershed area (km^2)	Daily inflow (m^3/s)		Coefficient of variation	Average over area ($l/(s \times km^2)$)
		Average	Standard Deviation		
Harricana	3680	57.74	42.80	0.74	15.69
Mistassibi	9320	197.89	197.44	1.00	21.23
San Juan	580	51.20	80.46	1.57	88.28

The threshold value that minimizes the false detection ratio differs from one river to another for any given test, and this indicates that this parameter is dependent on the characteristics of the data under investigation. It is expected that the highest threshold values be obtained for the San Juan River considering the relatively high variability of daily inflow, that is, a high coefficient of variation or standard deviation compared with the mean flow. In this context, even uncorrupted instances can lead to high test values, and the determination of high threshold values reflects the need to consider as uncorrupted these high values to minimize the risk of falsely detecting uncorrupted data points as

uncorrupted. To conclude that threshold values are proportional to the coefficient of variation of the data is not appropriate though, based on the results for the Harricana River. The threshold values are higher for this river than those of the Mistassibi River except Test 1, and higher than a few of those for the San Juan River, in spite of the fact that the coefficient of variation for the Harricana River is lower than that of the other two rivers. Note that the lowest false detection ratios on uncorrupted data points are encountered in the case of the Mistassibi River, even though its threshold values are the lowest of all rivers. A visual inspection of the Harricana River data shows a very smooth annual cycle with very little noise. Many lakes are found on the watershed of this river, for these are a common feature of the region where the river is located (i.e., around James Bay and Hudson Bay), and these seem to ensure an important routing of precipitation events. A study with more data sets would be recommended to provide a reliable diagnostic on the presumed unusual behavior encountered in the Harricana River, but this present application does highlight the difficulty of finding acceptable threshold values for detection tests for outliers. The optimization process employed here for the determination of the threshold is the minimization of the sum of false detection ratios, that is, the minimization of U+C (see Tables 5.14). However, the variability of this sum around the minimum is not overly large for most tests. Another alternative to the minimization approach could be to find a threshold such that U = C. This leads to an equal uncertainty regarding the false detection of uncorrupted or corrupted data points with little increase in the sum U+C (+7 in the worst case). It does not change the aforementioned conclusion that the threshold values of the Harricana River are high compared with threshold values for the other rivers. In any case, it is most likely preferable to use the minimization of the sum of U+C for most tests because this leads to smaller ratios of uncorrupted points falsely detected as corrupted. With regard to the issue of outliers, it is assumed that the number of uncorrupted data points in a data set largely exceeds the number of corrupted points, and this implies that it is preferable to be in a situation where the overall false detection ratio, that is, the total number of falsely detected points whether uncorrupted or corrupted over the total number of points, is minimized. If a particular test must be chosen as best, Test 1 should be selected for the Harricana and San Juan Rivers while the Fuzzy 5 test should be considered for the Mistassibi River. This conclusion regarding Test 1 remains as such even if the threshold

values on the Harricana and Mistassibi Rivers are increased for this test so that the U ratio is reduced - to the detriment of the C ratio - and are more in line with those obtained with the best tests based on AITs.

On one hand, the results presented in Tables 5.14 for AITs represent the best solutions provided for such techniques. This involves the aggregation of maps or cluster sets as performed for shifts and trends in Chapter 5. Also, a given outlier obviously appears in more than one subset, that is, 5, 7 or 10 times depending on the size of the subsets, for a subset is simply a snapshot of the whole data set within a small time frame moving chronologically in regular time steps. The Amp/SD ratio of a targeted data point, whether corrupted or not, is the average of the Amp/SD ratio obtained with the a group of 10 maps or ten cluster sets applied on either the 5, 7 or 10-individual subsets containing this particular data point. Therefore, for any given data point, 50, 70 or 100 values of the Amp/SD ratio are averaged for the 5, 7 or 10- individual subsets, respectively. On the other hand, the results presented in Tables 5.14 for AITs represent the most stringent solutions provided by such techniques. With AIT tests, there is also the necessity to determine the location of the outlier, its position changing from one subset to another. Table 5.16 gives the success rate for identifying the location of outliers by the variants of tests based on AITs, and this for each river.

Table 5.16. Success rate in identifying the location of the outliers.

River	Case	Success rate (%) in identifying the location of the outlier at plus minus (time steps)									
		0	1	2	3	4	5	6	7	8	9
Harricana	Knet 5	81.35	81.41	81.47	81.51	81.53					
	Fuzzy 5	82.12	82.20	82.26	82.30	82.32					
Mistassibi	Knet 5	82.86	85.21	85.33	85.40	85.41					
	Fuzzy 5	83.01	84.14	84.23	84.28	84.29					
San Juan	Knet 5	66.85	67.61	68.45	68.93	69.07					
	Fuzzy 5	65.34	65.79	66.40	66.72	66.83					
Harricana	Knet 7	79.96	80.40	80.45	80.50	80.53	80.57	80.58			
	Fuzzy 7	80.23	80.82	80.89	80.94	80.98	81.02	81.04			
Mistassibi	Knet 7	76.81	81.39	81.85	81.92	81.97	82.00	82.01			
	Fuzzy 7	78.41	81.86	82.06	82.13	82.17	82.21	82.21			
San Juan	Knet 7	58.15	58.40	58.70	59.02	59.30	59.49	59.54			
	Fuzzy 7	62.63	64.00	64.84	65.54	66.10	66.45	66.50			
Harricana	Knet 10	70.78	78.13	78.62	78.71	78.77	78.81	78.84	78.86	78.86	78.87
	Fuzzy 10	47.73	72.52	75.47	76.32	76.75	76.83	76.86	76.88	76.89	76.89
Mistassibi	Knet 10	64.97	74.15	76.43	76.92	77.03	77.08	77.12	77.15	77.15	77.15
	Fuzzy 10	44.66	67.86	72.00	73.32	73.71	73.80	73.83	73.85	73.86	73.86
San Juan	Knet 10	51.08	54.80	55.36	55.82	56.31	56.74	57.08	57.27	57.34	57.35
	Fuzzy 10	36.61	50.84	52.27	52.98	53.50	53.80	53.98	54.06	54.10	54.13

The results in Table 5.16 consider the outliers that have been correctly detected as corrupted data points based on the threshold values on the Amp/SD ratio, and the location of the outliers is evaluated from the 50, 70 or 100 responses provided by the maps or cluster sets applied to the 5, 7, or 10-individual subsets, respectively. In order to be fair to Test 1 and Test 2, the false detection ratio on corrupted data points (i.e., C in Table 5.14) is established based on the requirement that the location of the outlier be identified exactly. The false detection ratios for corrupted points in Table 5.14 (see Columns 4, 7 and 10) for AIT detection tests are equal to 100 minus the values in Column 3 of Tables 5.16 (i.e., the success rate in exactly identifying the location of the outliers that have been correctly diagnosed as corrupted points based on the threshold values of the Amp/SD ratio). This stringent requirement does not affect the Knet 5, Knet 7, Fuzzy 5 and Fuzzy 7 cases, for these cases provide the exact location of the outliers in almost all instances that are correctly identified as corrupted. For example, the success rate of Knet 5 for the Harricana River of both adequately diagnosing corruption and the location of corrupted points is 81.35% (see the value in Row 3, Column 3 of Table 5.16). The success rate for this case of adequately diagnosing corruption regardless of location is only slightly higher, that is, 81.53% (see the last value in Row 3 of Table 5.16). For Knet 7 for the Harricana River, the numbers are respectively 79.96% and 80.58%, which represent a slightly lower success rate because of the larger number of possible locations of outliers to choose from in Knet 7 than in Knet 5. For Knet 10, and particularly with Fuzzy 10, the difference between identifying the exact location and having no regard for location is much greater for all rivers and this is responsible for the poor results in Table 5.14 for these two detection tests.

The success of all tests is related to their capacity for determining the amplitude of outliers to some extent. The Kohenen maps and fuzzy c-means cluster sets are established so as to estimate this amplitude directly. Test 1 and Test 2 are not designed for this purpose, but can accomplish the quantification of the amplitude to some extent. Table 5.17 presents the RMSEs between observed and calculated Amp/SD ratios with Test 1 and Test 2. The RMSEs in Table 5.16 represent the best attempt at estimating Amp/SD, where a value of 0 for Amp/SD is automatically given when the tests diagnose the data point as uncorrupted. Another approach, which provides worse results than those in Table 5.16, is to neglect the diagnostic of the tests. Here, the Amp/SD ratio is only equal to Equation 5.2 divided by SD

for Test 1 and to Equation 5.3 divided by SD^2 for Test 2. The results in Table 5.17 may be compared with those of Table 5.18, which provides the RMSEs obtained with the Kohonen maps and fuzzy c-means cluster sets.

Table 5.17. Ratio Amp/SD for cases of outliers with conventional tests.

Case	RMSE for		
	Harricana	Mistassibi	San Juan
Test 1	0.398	0.416	0.825
Test 2	0.463	0.469	1.666

The most interesting point to observe from Tables 5.17 and 5.18 is the relatively small variability from one river to another in the RMSEs obtained from the AITs compared with those of Test 1 and Test 2. In spite of the difference in the statistical properties of the data from each river, a relative stability in the estimation of the amplitude of outliers can be achieved with AITs, but cannot be obtained with the conventional tests employed here, Test 1 and Test 2. Note that the failure comes only from the estimation of Amp/SD with the San Juan River, which is the most difficult case due to its high variability of the daily inflows. When this river is neglected, and only the Harricana and Mistassibi Rivers are considered, then it can be said that Test 1 and Test 2 provide relatively stable estimates of the amplitude of outliers, although they are not as good as those provided by the AIT-based tests. The estimation of the amplitude of outliers might be of interest for those who want to quantify the effect of such anomalies on decision-making processes. Indeed, if there is a fear of removing potential outliers from data sequences based on detection tests because of the risk of false detection that would lead to the removal of valuable, uncorrupted data, then the development and use of outlier detection tests becomes irrelevant. Of course, if outliers are not removed, then they might induce a bias in the estimation of water quantity and this must be assessed. From the determination of the average amplitude of outliers potentially present in hydrometric data sequences, an evaluation of the amplitude of the bias induced by these outliers in the estimation of water quantity could possibly be achieved. Such evaluation of bias could be translated into economic consequences. A better estimate of the amplitude of outliers may lead to a better understanding of the bias induced by these outliers in the assessment of water quantity and economic effects.

Table 5.18. Ratio Amp/SD for cases of outliers with AIT.

River	Case	RMSE for		Improvement from		
		Average of maps	Aggregation of maps	Average of maps	Test 1	Test 2
Harricana	Knet 5	0.333	0.315	5.62	20.86	31.98
	Fuzzy 5	0.304	0.295	2.96	25.74	36.18
Mistassibi	Knet 5	0.311	0.303	2.63	27.16	35.29
	Fuzzy 5	0.305	0.301	1.27	27.72	35.79
San Juan	Knet 5	0.375	0.361	3.58	56.23	78.31
	Fuzzy 5	0.371	0.364	1.90	55.89	78.15
Harricana	Knet 7	0.367	0.348	5.28	12.56	24.85
	Fuzzy 7	0.350	0.340	2.83	14.48	26.50
Mistassibi	Knet 7	0.345	0.336	2.61	19.17	28.19
	Fuzzy 7	0.345	0.340	1.40	18.27	27.40
San Juan	Knet 7	0.435	0.417	4.03	49.47	74.96
	Fuzzy 7	0.437	0.428	2.21	48.17	74.32
Harricana	Knet 10	0.399	0.383	4.00	3.73	17.26
	Fuzzy 10	0.396	0.386	2.34	2.80	16.46
Mistassibi	Knet 10	0.392	0.383	2.11	7.90	18.18
	Fuzzy 10	0.393	0.388	1.32	6.77	17.18
San Juan	Knet 10	0.505	0.486	3.79	41.07	70.80
	Fuzzy 10	0.499	0.488	2.17	40.81	70.68

5.3.3 Discussion

The results presented on outliers in this section must be considered carefully, for the detection tests are not compared using exactly the same validation data sets. Because all the tests can only detect single isolated outliers, cases in the validation database that would have involved detection of multiple outliers have been removed prior to calculating false detection ratios or estimating amplitudes. For Test 1 and Test 2, this involves removing cases where two or more outliers appear immediately after one another, and this leads to trimming the validation database by around 2%. For tests based on AITs, this amounts to removing all subsets that contain more than one outlier, yielding a reduction of the validation database by slightly more than 3%. It is deemed that these reductions of data have little impact, considering that the validation data base is made up of more than one million test data points. One must remember that the results focus only on the detection of single, isolated outliers.

The choice of the size of the subsets has been driven by the variability in the time response of watersheds. The fundamental basis of pattern recognition techniques such as those employed here is the ability to make decisions with respect to overall picture of the

data. In this application, the picture is a subset of data taken within a given time frame (i.e., 5, 7 or 10 time steps), and it is expected that subsets affected by outliers would present different features than those of subsets that are uncorrupted. Long subsets could possibly be employed on data representing slow hydrologic regimes, with long rises and recessions, therefore yielding rather smooth monotonic evolutions unless "spikes" like outliers are present. The results presented here show that 5 and 7-individual subsets are the maximum length admissible, at least for the watersheds investigated. The cases with the 10-individual subsets yield the worst performance of all tests analyzed, and they are also more susceptible to failure due to multiple outliers within subsets. This does not mean however that shorter subsets would be preferable. Test 1 can be considered as a pattern recognition technique working on 2-individual subsets, and indeed provides the best performance except for the Mistassibi River. Test 2 can be viewed as a pattern recognition technique working on 3-individual subsets, and produces performance that is worse than AIT-based detection tests. The watersheds employed here with the tests have been chosen because they are deemed to represent most of the hydrologic regimes present in Canada. Of course, if time was not a constraint, it would have been preferable to evaluate the test on as many watersheds as possible.

5.4 Conclusion

The results in Chapters 4 and 5 justify the application of AIT-based tests because at the least they can help confirm the results obtained with conventional detection tests for shifts, trends and outliers. The AIT-based tests may in fact complement conventional detection tests, as all conventional and AIT-based tests, even though they may perform similarly in terms of false detection, may behave differently. For a given situation, some tests could provide a diagnostic of detection, while the others may conclude to no detection. This difference in the behavior of the tests is shown in Chapter 6, with applications on real hydrometric data. AIT-based tests may constitute an enhancement of conventional detection tests, as shown in some multivariate cases, and this should be considered a rationale for further developments.

The conventional tests applied here are very simple, especially those for outliers, which make them very attractive for practitioners. Nevertheless, simple tests such as the

Mann-Withney test for shifts and Mann-Kendall test for trends are accepted widely as standards, and it is normal to consider them as references for comparison, in spite of the possibility of more sophisticated tools, depending on the anomaly under investigation. AIT-based tests can be considered as being more sophisticated on a technical standpoint than the conventional tests employed here, yet do not consistently provide significant improvement in performance. As stated previously, for univariate cases, AIT-based tests only confirm and to some degree complement the results obtained by conventional tests. It would be of interest to evaluate a greater selection of techniques (see Chapter 2) for the detection of anomalies under an experiment similar to that presented in this work, in Chapters 4 and 5. Among other issues, the variability in the false detection performance should be given particular attention, for the performance of the approaches presented here is remarkably and consistently similar for each anomaly. It may be useful to investigate if this observation is true for a greater array of methods.

With regard to AITs applied for the detection of anomalies, one must note that their greatest advantage is their capacity to provide a good description of data domains. This description implies a long calibration process, which is a disadvantage of AITs-based tests. The processing time, for the detection of anomalies after calibration, is also slightly longer for AITs than for conventional techniques, but remains negligible if one treats only a few sequences, as is normally the case in practice. The AIT-based approaches developed in this work might not provide significant improvement from other conventional methods, but these approaches might represent only a very limited sample of the possible strategies for addressing the issues of data anomalies with AITs. The AITs employed in this work are actually based on very simple structures, which make them very flexible and able to be molded in numerous ways depending on the inputs provided. There is no rule regulating this modeling process, the only limitation being the imagination of the user. As a result, much trial and error may be necessary before finding a preferred approach, and the approaches in this thesis may very likely represent only a few possibilities.

Patience and perseverance is necessary in the development of AIT-based approaches for the detection of data anomalies. It must be noted that only simple cases have been evaluated here, where only one shift or trend per sequence, or one outlier per subset are evaluated. More complicated with multiple shifts, trends and outliers, or with

combinations of anomalies (i.e., shifts with trends or several outliers of a different nature) should be investigated, for they constitute more realistic cases. The choice of the random generators (i.e., Normal and Uniform) for the creation of synthetic data sequences and anomalies should also be revised to better reflect the reality of natural processes. It must be noted though that the validity of tests becomes more questionable as the cases are more complex, and one must therefore rely on additional information (i.e., multivariate cases) or more heuristic approaches (e.g., expert judgment) to provide a decision regarding the detection of anomalies. AITs might in fact prove very useful for the assessment of these complex cases. It is shown in this work that AITs may constitute an improvement beyond that of conventional tests for multivariate cases, and they can also integrate soft data such as expert judgment in the description of data domains. The use of soft data is not demonstrated in this work, because the author's experience would not be extensive enough to provide a strong expertise base. A panel of experts should be called upon for the development of this expertise base. Ultimately, an expert system could be built based on AITs, for which the detection of anomalies would be accomplished with respect to several requirements for the data or thresholds for the tests.

Before developing further approaches based on AITs, it is necessary to demonstrate and evaluate the tests developed here using real hydrometric data in order to strengthen their validity. This is accomplished in Chapter 6.

Chapter 6
Practical Applications of Mapping Procedures

The purpose of this chapter is to demonstrate the applicability of the methods presented in Chapter 5 for real cases. The databases employed here are much smaller than those of the experiments of Chapters 4 and 5 and therefore permit a more specific analysis of the applicability of the methods assessed in this work. A more in-depth discussion of the validity and potential concerns regarding the application of AITs for the assessment of anomalies in data is offered, based on the results obtained here and in the previous chapters. Notably, the results highlight the value of AIT-based tests as complements to conventional detection tests. Two general applications, first for shifts and trends, and second for outliers, are presented, each with a description of the database used and a discussion of the results obtained. Finally, broad conclusions based on these two applications are offered.

6.1 Application to Shifts and Trends

6.1.1 Description of the Application Case

The context of application here is similar to that presented in Anderson et al. (1992), Yulianti and Burn (1998), Zhang et al. (2001), or Cunderlik and Burn (2002), where a large number of hydrometric stations over a large territory are tested so as to determine regional patterns. The hydrometric stations chosen here come from the pool of stations employed by these authors, who have selected them because they measure inflows on unregulated rivers and have relatively long data records of assumed high quality. Consequently, comparisons between the results of these authors and those of this application can be made to some extent. In particular, this analysis is designed to be similar to that of Zhang et al. (2001), considering similar data sets for tests, that is, annual mean inflows and annual maximum daily inflows, and performing detection tests on the same periods as much as possible, that is, from 1947 to 1996 (50 years), 1957 to 1996 (40 years), and 1967 to 1996 (30 years). Table 6.1 lists the hydrometric stations employed here, the data for these stations being taken from Environment Canada HYDAT CD-ROM, version 96-1.04. Most of the stations are located in the southern regions of Canada, with a few stations, all with relatively short

records, in the north around the Hudson Bay, in the western provinces and the territories (Northwest and Nunavut).

Table 6.1. Hydrometric stations employed for detection tests.

Province	Station	Name	Period of availability (years) for					
			Annual mean			Annual daily max		
			30	40	50	30	40	50
NB	01AK001	Shogomoc Stream near Trans Canada	Y	Y	Y	Y	Y	Y
NB	01AQ001	Lepreau River at Lepreau	Y	Y	Y	Y	Y	Y
NB	01BE001	Upsalquitch River at Upsalquitch	Y	Y	Y	Y	Y	Y
NS	01DG003	Beaverbank River near Kinsac	Y	Y	Y	Y	Y	Y
NS	01EC001	Roseway River at Lower Ohio	Y	Y	Y	Y	Y	Y
NS	01EF001	Lehave River at St. Margatets Bay	Y	Y	Y	Y	Y	Y
NS	01EH001	East River at St. Margatets Bay	Y	Y	Y	Y	Y	Y
NS	01EO001	St. Marys River at Stillwater	Y	Y	Y	Y	Y	Y
ON	02AA001	Pigeon River at Middle Falls	Y	Y	Y	Y	Y	Y
ON	02EA005	North Magnetawan River near Burk's Fall	Y	Y	Y	Y	Y	Y
ON	02EC002	Black River near Washago	Y	Y	Y	Y	Y	Y
QC	02NF003	Matawin & Saint-Michel-Des-Saints	Y	Y	Y	Y	Y	Y
QC	02PJ007	Beaurivage (Riviere) & Saint-Etienne	Y	Y	Y	Y	Y	Y
NL	02YL001	Upper Humber River near Reidville	Y	Y	Y	Y	Y	Y
QC	03MB002	Rivere a la Baleine pres de l'embouchure -1	Y			Y		
ON	04LJ001	Missinaibi River at Mattice	Y	Y	Y	Y	Y	Y
QC	04NA001	Harricana (Riviere) & Amos	Y	Y	Y	Y	Y	Y
AB	05AA004	Pincher Creek at Brodin's Farm	Y			Y		
AB	05AD003	Waterton River near Waterton Park	Y	Y		Y	Y	
AB	05AF010	Manyberries Creek at Brodin's Farm				Y	Y	Y
AB	05BB001	Bow River at Banff	Y	Y	Y	Y	Y	Y
AB	05BJ004	Elbow River at Bragg Creek				Y	Y	Y
MN	05LJ005	Ochre River at Ochre River	Y	Y		Y	Y	
ON	05PB014	Turtle River near Mine Center	Y	Y	Y	Y	Y	Y
ON	05OA002	English River at Umfreville	Y	Y	Y	Y	Y	Y
MB	06GD001	Seal River below Great Island	Y			Y		
AB	07BE001	Athabasca River at Athabasca	Y	Y		Y	Y	Y
BC	07FB001	Pine River at East Pine	Y			Y		
NT	07OB001	Hay River near Hay River	Y			Y		
NT	07RD001	Lockhart River at outlet of Artillery Lake				Y		
BC	08CB001	Stikine River above Grand Canyon	Y			Y		
BC	08CD001	Tuya River near Telegraph Creek	Y			Y		
BC	08FB007	Bella Coola River above Burnt Bridge Creek				Y		
BC	08JB002	Stellako River at Glenannan	Y	Y		Y	Y	
BC	08JE001	Stuart River near Fort St-James	Y	Y		Y	Y	Y
BC	08KH006	Quesnel River near Quesnel	Y	Y		Y	Y	Y
BC	08LA001	Clearwater River near Clearwater Station	Y	Y		Y	Y	
BC	08LD001	Adams River near Squilax	Y	Y		Y	Y	
BC	08MA002	Chilko River at outlet of Chilko Lake				Y	Y	Y
BC	08NL007	Similkameen River at Princeton	Y	Y	Y	Y	Y	Y
NT	10EB001	South Nahanni River above Virginia Falls				Y		
NU	10RC001	Black River above Hermann River	Y			Y		
AB	11AA026	Sage Creek at Q Ranch near Wildhorse	Y	Y	Y	Y	Y	Y

Note: Y stands for yes.

For annual mean inflows, the availability of a station is determined based on the length of the overall records and the number of missing data within the periods of interest.

If more than 20% of data are missing in a given year, then this year is deemed unavailable for the station. The periods are not fixed in time, and therefore the data for some stations do not exactly fall in the time periods from 1947 to 1996 for the 50-year period, 1957 to 1996 for the 40-year period, and 1967 to 1996 for the 30-year period, although they are close to these ranges. If 30 years of data cannot be gathered from 1962 to 1996, then the station is considered as unavailable for the 30-year period. Similarly, 40 years of data from 1952 to 1996 for the 40-year period and 50 years of data from 1942 to 1996 for the 50-year period must be gathered for a station to be considered usable. For annual maximum daily inflows, the availability of a station is based on the length of the overall records and missing data within the flood periods. For a given year, if there are missing data for a large part of the periods where floods occur in the region where the station is located, then this year is considered as unusable. The number of usable years of data for the 30, 40 and 50-year periods is established as for the annual mean inflows. The last six columns of Table 6.1 describe which stations are available (i.e., Y) for each period length (i.e., 30, 40 and 50 years) and for each data set (i.e., annual mean inflows and annual maximum daily inflows). For the annual mean inflows, 37, 29 and 21 stations are usable for the 30, 40 and 50-year periods, respectively, while 43, 32 and 27 stations are usable for the 30, 40 and 50-year periods, respectively, for the annual maximum daily inflows.

6.1.2 Results and Discussion

All detection tests for univariate cases involved in Chapter 5 are applied to the data here, namely the Mann-Whitney and Student's tests for shifts, the Mann-Kendall and Spearman tests for trends, the tests based on Kohonen maps for shifts and trends, and the tests based on fuzzy c-means cluster sets for shifts and trends. The threshold values that distinguish between a diagnostic of corruption or non-corruption are those given in Tables 5.4 and 5.5. Because conventional and AIT-based detection tests produce equivalent performance, the approach applied here is to not rely on only one test, but on all of them, that is, four tests each for shifts and trends. If a sequence is diagnosed as uncorrupted or corrupted by all available tests, it may be safe to assume that these unanimous diagnostics confirm one another as opposed to the case where the tests provide conflicting diagnostics. Tables 6.2 and 6.3 demonstrate this assumption for the detection tests for shifts and trends,

respectively, based on the validation data sets employed in Chapter 5 for univariate cases. Tables 6.2 and 6.3 also provide the results with respect to 30, 40 and 50-individual sequences, and indicate the number of sequences that activate 0, 1, 2, 3 or 4 tests. A test is said to be activated if it diagnoses a given sequence as corrupted. Also, the sequences are of course divided between uncorrupted and corrupted cases (Columns U and C in Tables 6.2 and 6.3). For example, Table 6.2 indicates that, for the case of 30-individual sequences, 12,715 sequences lead to the activation of all four available tests, and of these 12,715 sequences, 85% are corrupted while the remaining 15% are uncorrupted. Therefore, based on the validation data sets in Chapter 5, there is an 85% chance that a given 30-individual sequence is actually corrupted if all four tests are activated, and this is an improvement from the consideration of only one test. Obviously, the results are more ambiguous when only half the tests are activated.

Table 6.2. Conjoint results from detection tests for shifts.

Number of tests activated	30-individual sequences			40-individual sequences			50-individual sequences		
	Number of sequences	Percentage of		Number of sequences	Percentage of		Number of Sequences	Percentage of	
		U	C		U	C		U	C
0	26,083	67	33	26,468	68	32	26,444	70	30
1	4,113	58	42	4,199	60	40	4,761	60	40
2	4,209	49	51	4,470	44	56	4,225	45	55
3	2,880	38	62	2,575	36	64	2,897	31	69
4	12,715	15	85	12,288	12	88	11,673	10	90

Note: U stand for uncorrupted sequences and C stands for corrupted sequences.

Table 6.3. Conjoint results from detection tests for trends.

Number of tests activated	30-individual sequences			40-individual sequences			50-individual sequences		
	Number of sequences	Percentage of		Number of sequences	Percentage of		Number of sequences	Percentage of	
		U	C		U	C		U	C
0	28653	57	43	31223	61	39	31349	66	34
1	2509	52	48	2091	51	49	757	54	46
2	2978	50	50	2273	48	52	2543	49	51
3	2086	48	52	1455	45	55	571	45	55
4	13774	34	66	12958	24	76	14780	19	81

Note: U stand for uncorrupted sequences and C stands for corrupted sequences.

As a whole, the chance of detecting a shift when all four tests are activated is fairly high regardless of the sequence length, ranging from 85 to 90% (see Table 6.2). On the other hand, the chance of detecting an uncorrupted sequence when no tests are activated ranges only from 67 to 70%, and this is a good indication of why all tests falsely detect corrupted sequences as uncorrupted at a higher rate than they falsely detect uncorrupted

sequences as corrupted. In the best case, where sequences are considered as corrupted when two or more tests are activated and as uncorrupted when no test or only one test is activated, the false detection ratios (uncorrupted and corrupted) are essentially equal to those presented in Table 5.4. Similarly, the combination of the results from tests does not yield significant improvements in the estimation of the Amp/CV ratio or the location of the shift. Strictly from a performance standpoint, the combination of tests does not lead to improvement from the consideration of only one test. From a practical standpoint, to favor one specific test would not be adequate, because all tests are equally reliable or unreliable. The combination of tests, as applied here on the hydrometric data listed in Table 6.1, is designed to ensure some relatively unambiguous diagnoses (i.e., zero or four activated tests) for at least a few of the sequences.

For cases of trends, similar conclusions to those stated for cases of shifts may be drawn. In the best case of combinations, if sequences are considered as corrupted when two or more tests are activated and as uncorrupted when no test or only one test is activated, then the false detection ratios (uncorrupted and corrupted) are essentially equal to those presented in Table 5.5. Also, the combination of the results from the tests would not yield significant improvements in the estimation of the Amp/CV ratio of the trends. The benefit of the combination of the tests is from the perspective that relatively unambiguous diagnoses (i.e. zero or four activated tests) can be obtained for at least a few of the sequences, although this reduction of ambiguity is less prevalent for trends than it is for shifts, as can be seen by the comparison of Table 6.2 with Table 6.3.

Table 6.4 details the results obtained from detection tests for shifts when applied to the hydrometric data from the stations listed in Table 6.1. Table 6.4 distinguishes between the variables under evaluation, that is annual mean inflows and annual maximum daily inflows, and among the length of the data sequences. As in Tables 6.2 and 6.3, the sequences are grouped as a function of the number of tests they activate. In Table 6.4, the numbers in parentheses are the number of sequences available for tests for a given variable and sequence length. A large portion of the data sequences of annual mean inflows do not activate any of the detection tests available or activate only one test, which indicates the presence of very few shifts in this group of hydrometric stations. The proportion of sequences of annual mean inflows that activated two or more tests is quite a bit larger for

30-year sequences than it is for 40- and 50-year sequences. This proportion must be considered with caution, however, for the high proportion with 30-year sequences is in large part due to hydrometric stations that are not available for tests on 40- and 50-year sequences. As a result, there is no possibility to determine if the shifts in the 30-year sequences are noticeable in the 40- and 50-year sequences. The estimation of the location of the shifts from all tests seems to indicate that, if a shift is present in a sequence, then it occurs in the last 30 years (i.e., between 1967 and 1996), the most likely decade being the 70s. The proportion of sequences of annual maximum daily inflows that activate two or more tests is more likely to be sustained from one sequence length to another. First, the number of possibly shifted sequences of annual maximum daily inflows is larger than the number of possibly shifted sequences of annual mean inflows. Second, this sustained proportion from one sequence length to another of possibly shifted sequences is also due to the fact that a large portion of the stations that are possibly affected by shifts are available for tests at all sequence lengths. This main change comes from the stations in the Atlantic Region, all having long data records, with a large portion of them having possibly shifted annual maximum daily inflows and mostly unaffected annual mean inflows. Note the large number of ambiguous sequences of annual maximum inflows (i.e. one to three tests activated). This is explained by the relatively large coefficients of variation of annual maximum daily inflow sequences compared with those of annual mean inflow sequences. All detection tests are more likely to fail to provide adequate diagnostics as the coefficient of variation of the data sequences increases.

Table 6.4. Detection tests for shifts with hydrometric data.

Number of tests activated	Sequences of annual mean inflows			Sequences of annual maximum daily inflows		
	30-year sequences (37)	40-year sequences (29)	50-year sequences (21)	30-year sequences (43)	40-year sequences (32)	50-year sequences (27)
0	16	16	12	16	9	8
1	7	5	4	5	4	5
2	4	4	2	7	10	4
3	1	1	1	6	3	0
4	9	3	2	9	6	10

Based on the estimation of the Amp/CV ratio for the AIT-based detection tests and on the test values obtained for conventional detection test, it is deemed that a reduction over time in the annual mean inflows occurs for the 30- and 40-year sequences for most

stations. The exceptions to this conclusion are some stations in the north (Northern BC and, in and around the territories), and two stations in the Great Lakes / Saint Laurence Region. With 50-year sequences of annual mean inflows, a slight increase over time is noticed consistently for stations in the Great Lakes / Saint Laurence Region and also for some stations in the Atlantic Region. Slight to potentially significant decreases are observed for all other stations. For 30-year sequences of annual maximum inflows, reduction over time of inflows is consistent in the southern regions (i.e., the Atlantic, Great Lakes / Saint Laurence, and Southern BC Regions) while an increase is noticed for the northern regions (Northern BC, and in and around the territories). For 40- and 50-year sequences, the reduction persists for the Atlantic Region and Southern BC and might be significant for both regions, while a mix of slight reductions and increases are observed for the stations in the Great Lakes / Saint Laurence Region and in Ontario and Quebec around the Hudson Bay. As with the annual mean inflows, if a shift is present in the sequence, it is likely that it has occurred recently, most specifically in the 70s. Overall, for both annual mean and maximum daily inflows, the most affected region is that of Southern BC with consistent, possibly significant reductions of inflows.

Table 6.5 provides information of the same nature as that of Table 6.4, but this time with detection tests for trends. For both annual mean and maximum daily inflows, the number of unambiguous cases is large, that is, most sequences either activate no tests or activate all four available detection tests for trends. However, as seen in the results in Table 6.3, the possibility of false detection remains high even with presumably unambiguous diagnostics. Many cases of sequences affected by a trend could possibly lead to inactivation of all tests, and many cases where no trend is present could potentially produce the activation of all tests. The results in Table 6.5 are deemed valid, nevertheless, for they mostly confirm those of Table 6.4. As mentioned in Chapter 4, detection tests for shifts and trends perform essentially the same function. Detection tests for shifts can be employed for the identification of trends, for a trend can be considered to be the result of shifts that occur at regular intervals. Detection tests for trends can be employed for the identification of shifts, for a shift involves a change in the characteristics of the data as much as a trend does. If the interest is in the detection of a reduction or increase in the values of inflows regardless of the sources (i.e., shifts or trends), then detection tests for shifts and trends can

be used together. Table 6.6 shows the results of employing all tests for shifts and trends, which means that a sequence has the possibility to activate up to 8 tests.

Table 6.5. Detection tests for trends with hydrometric data.

Number of Tests Activated	Sequences of annual mean inflows			Sequences of annual maximum daily inflows		
	30-year sequences (37)	40-year sequences (29)	50-year sequences (21)	30-year sequences (43)	40-year sequences (32)	50-year sequences (27)
0	22	21	17	20	19	13
1	2	2	0	3	1	0
2	3	1	1	5	3	4
3	0	0	1	3	1	0
4	10	5	2	12	8	10

The results from Tables 6.5 and 6.6 confirm the conclusions based on the data in Table 6.4. For annual mean inflows, consistent reductions are observed for all southern regions of Canada in the short to medium term (30- and 40-year), while increases are obtained for the northern regions in the short term (30-year). Slight reductions persist in the southern regions on the long term (50-year), except in the Great Lakes / Saint Laurence and in the Atlantic Region to some extent, where slight increases are noticed. As suggested by Zhang et al. (2001), interdecadal variability might be the cause of the specific patterns per region of ups and downs over time in the annual mean inflows. In the short term (30-year), reduction over time of annual maximum daily inflows is consistent in the southern regions (i.e., the Atlantic, Great Lakes / Saint Laurence, and Southern BC Regions) while an increase is noticed for the northern regions (Northern BC, and in and around the territories). The reduction persists for the Atlantic Region and Southern BC in the medium to long term (40- and 50-year), while a mix of slight reduction and increase is observed for the stations in the Great Lakes / Saint Laurence Region and in Ontario and Quebec around the Hudson Bay.

The results presented here are in large measure in agreement with those of Zhang et al. (2001), who perform extensive cross-country analyses of trends on inflows. Both applications lead to the conclusion that Southern BC is the most affected region of the country with respect to changes in annual mean and maximum daily inflows. From short to long-terms periods (i.e., 30 to 50 years), decreases are generally observed in the southern regions of the country. Small regional differences can be identified between the two applications such as, for example, the assessment of the annual mean inflows in the

Atlantic Region in the long term. Admittedly the application here is based on a much more limited number of hydrometric stations than in Zhang et al. (2001), and this reduces the validity of more specific, region to region analyses.

Table 6.6. Detection tests for shifts and trends with hydrometric data.

Number of Tests activated	Sequences of annual mean inflows			Sequences of annual maximum daily inflows		
	30-year sequences (37)	40-year sequences (29)	50-year sequences (21)	30-year sequences (43)	40-year sequences (32)	50-year sequences (27)
0	15	14	12	13	8	8
1	4	6	4	4	4	5
2	3	2	1	6	8	0
3	1	1	0	0	0	0
4	5	1	1	2	1	1
5	1	0	0	3	3	0
6	0	2	0	6	1	6
7	0	1	2	4	2	0
8	8	2	1	5	5	7

The results here do not lead to the conclusion that any of the detection tests distinguishes itself from the others, but rather that they may complement each other. With real data, there is no way to know for sure if the sequences are affected by anomalies, and consequently there is no possibility to accurately determine deficiencies in the tests. In such a circumstance one must rely entirely on the analytical results. In large part, the tests have been consistent in their behavior, that is, either they have all been activated or none have been activated for the majority of the data sequences presented to them. This application confirms that all tests are performing essentially equally and this validates the results obtained in Chapter 5 for synthetic data.

6.2 Application to Outliers

6.2.1 Description of the Application Case

The purpose of this application is more in the evaluation of the behavior of all tests than in the validation of the performance of those presented in Chapter 5. Even though some of the tests seem to exhibit similar performance, they do not necessarily detect the same kind of outliers, and this is demonstrated in this section. The gradient (Equation 5.2, referred to as Test 1) and the product of gradient (Equation 5.3, referred to as Test 2) tests, as well as the detection tests based on the Kohonen maps and fuzzy c-means cluster sets for 5-point subsets, respectively referred to as Knet 5 and Fuzzy 5, are employed here for the

detection of outliers on six hydrometric stations. The Kohonen maps and fuzzy c-means cluster sets for 7- and 10-point subsets are not considered, as they do not perform as well as Test 1, Test 2, Knet 5 and Fuzzy 5. Table 6.7 details the chosen stations, five of which record inflows, while the last one records water levels. Three of these stations are actually those considered in Chapter 5 for the construction of the validation database for the evaluation of the performance of the detection tests for outliers, although different 10-year periods are chosen. The choice of a station recording water level is justified to determine the versatility of the detection tests and their capacity to perform for different kinds of measurements. The hydrometric stations observing inflows are chosen so as to have diversity in the size and productivity of the watersheds, as shown in Table 6.7, and to examine the tests under different hydrologic regimes, as exhibited in Figure 6.1.

Table 6.7. Characteristics of the hydrometric stations.

Name	Province	Identification in the text	Type	Period (in year, inclusive)	Area (km^2)	Mean (m^3/s or m)	Standard deviation (m^3/s or m)
Coquihalla River near Hope	BC	C	Inflow	1970/79	741	30.87	33.21
Harricana River at Amos	QC	H	Inflow	1970/79	3,680	59.20	48.32
Metabetchouane River	QC	ME	Inflow	1980/89	2,280	46.70	55.52
Mistassibi River	QC	MI	Inflow	1970/79	9,320	204.83	221.85
San Juan River near Port Renfrew	BC	S	Inflow	1963/72	580	51.23	80.04
Saint-Jean Lake at Saint-Gédéon	QC	G	Level	1981/90	73,000	100.44	1.11

Figure 6.1 illustrates typical records for all the stations, namely (a) Coquihalla, (b) Harricana, (c) Metabechouane, (d) Mistassibi, (e) San Juan and (f) Saint-Gédéon stations. In terms of hydrologic regime, the Harricana station is in a northern region (i.e., Northen Quebec, close to James Bay) where inflows peak during the spring snowmelt, and then almost constantly recede until the next spring. The San Juan station is located on Vancouver Island, BC, where precipitation falls mostly as rain, generating a hydrograph that evolves with respect to the amplitude of precipitation inputs. The Coquihalla, Metabetchouane and Mistassibi stations are located in regions that receive regular water inputs, a large part as snow, and typical hydrographs for these watersheds usually contain two peaks, one in spring due to the snowmelt, and one in fall due to abundant precipitation and high soil moisture.

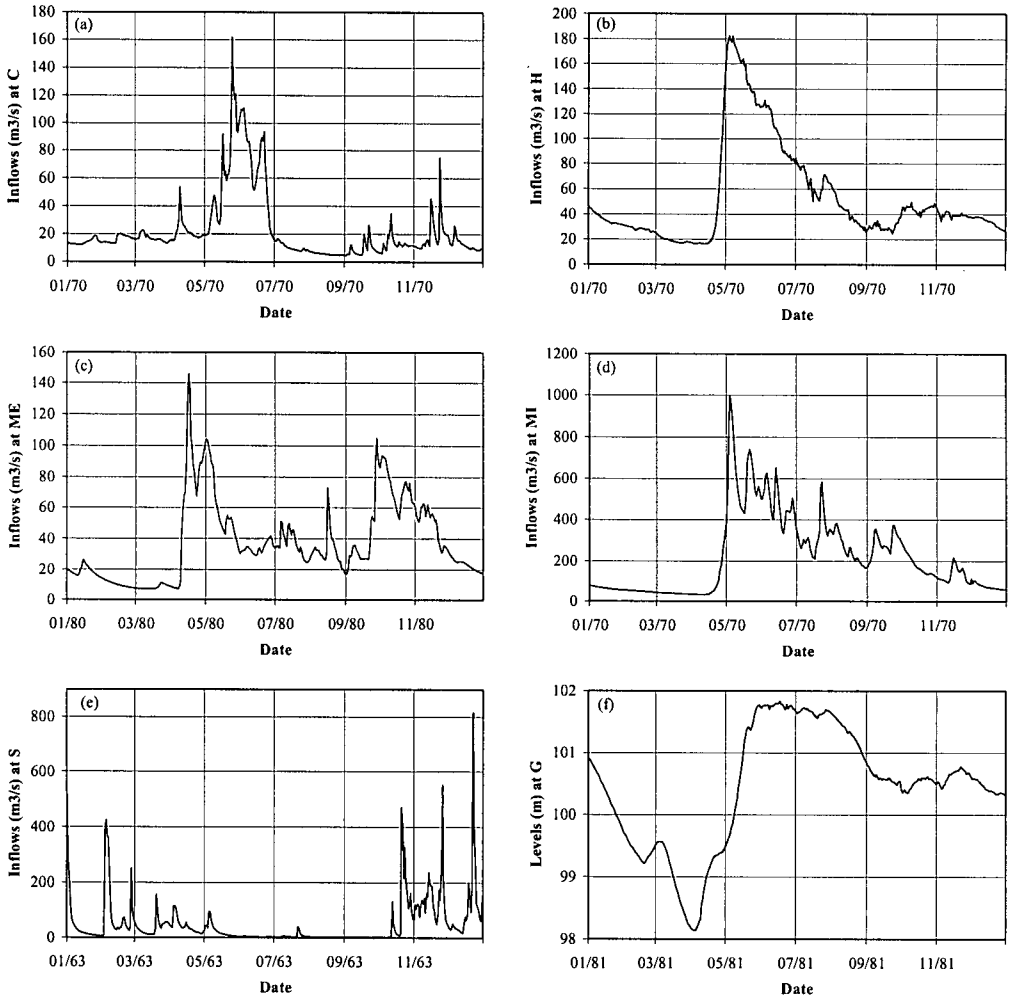


Figure 6.1. Inflows or water levels at all stations.

From Figure 6.1, it may be concluded that the evolution of the hydrographs come from natural causes. Suspicions might be raised in the case of the inflow at the Coquihalla station on May 25 (Figure 6.1a) and on the inflow at the San Juan station on November 26 (Figure 6.1e), for they seem to look like spikes that could be difficult to explain considering the possible response time of the respective watersheds. A naked eye examination of the 10-year period of all of the stations indicates that all the data sequences are of good quality, with only a few possible suspected cases of outliers. The data are taken from the Environment Canada HYDAT database, and the quality indicator provided by this source indicates that no data points have been revised within the chosen time period for all stations. The data on which the detection tests are applied are considered as essentially free

of outliers, and this application constitutes a way to determine the kind of data that will induce the tests to misdiagnose uncorrupted points as corrupted.

6.2.2 Results and Discussion

The first issue in the application of the tests is to determine the thresholds beyond which test values are assumed to be indicative of the presence of outliers. For the Harricana, Mistassibi and San Juan stations, the threshold values employed are those determined for these stations in Chapter 5 (Table 5.14). The threshold values assigned to the Mistassibi station are also those considered for the Coquihalla and Metabetchouane stations, because of similarities in the hydrologic regime of the watersheds and in the coefficient of variation of the data. Based on the similarity of the hydrologic regimes, the threshold values for the Mistassibi station are also assigned to the Saint-Gédéon station. The coefficient of variation for the data at the Saint-Gédéon station is small, which means that detection tests might not be activated (i.e., diagnosing a corruption or outlier) under any circumstance, if the threshold values are large. The threshold values for the Mistassibi station are among the smallest available for all tests.It must be said that the Knet 5 and Fuzzy 5 tests are employed more liberally for the application case here, that is, the detection diagnostic is based only on the threshold value, regardless of the location of the outliers. The neglect of the location of the outliers for Knet 5 and Fuzzy 5 is deemed preferable in order to avoid any conflicting decision should there be subsets of data points with multiple outliers.

The potential benefits of combining the results of all tests is examined, as accomplished for shifts and trends in Section 6.1. For outliers, the combination of test results provides a poorer performance than that of tests used alone when the validation database for outliers in Chapter 5 is employed. The failure to obtain better performance is due to the heterogeneity of the behavior of the tests. On one hand, detection tests for shifts and trends yields relatively uniform diagnostics, and their combination simply helps the decision process in the case of only a few cases causing divergence in the diagnostic. On the other hand, detection tests for outliers provide divergent diagnostics in many cases, and therefore very few clear decisions can be made based on the combination of the tests. As a consequence, the tests are evaluated for their performance individually for each of the

stations, which remains a reasonable, case-by-case task considering the small number of stations and the relatively small size of the time periods.

Table 6.8 indicates the number of times each test is activated for each station, and the numbers must be viewed in light of the fact that, considering the length of the data sequence, each test has been applied a total of 3,650 times per station. Because the data are assumed to be essentially free of outliers, it can be presumed that a test falsely diagnoses a corruption each time it is activated. For the Mistassibi and San Juan stations, the data in Table 6.8 exhibit false detection ratios on uncorrupted points for all tests that are considerably similar to those presented in Chapter 5 (see Table 5.14). This could lead to the conclusion that there is uniformity in the behavior of the tests within stations regardless of the time period, although this does not hold in the case of the Harricana station, for which false detection ratios are quite smaller in Table 6.8 than those in Table 5.14. The false detection ratios for the Mistassibi and Metabetchouane stations are relatively similar, and this can possibly be explained by the similarities in terms of the hydrologic regime of the watershed and the coefficient of variation of the data. The ratios are higher for the Metabetchouane station than for the Mistassibi station because factors that differentiate these two stations might not allow direct transfer of the threshold values for the Mistassibi station to the Metabetchouane station. The direct transfer of threshold values from one station to another is even more questionable in the case of the Coquihalla station, for which the threshold values for the Mistassibi station lead to very frequent activation of the tests. At the Saint-Gédéon station, the conventional tests (Test 1 and Test 2) are activated much less often than the AIT-based tests (Knet 5 and Fuzzy 5), and this indicates the difference in the behavior from one tests to another, as illustrated in the figures that follow.

Figures 6.2 to 6.4 present the hydrographs of selected stations and years, where the black dots represent the points in the data that are diagnosed as corrupted by the tests. There are four graphs per figure, that is, one for each test: Test 1, Test 2, Knet 5 and Fuzzy 5. Figure 6.2 illustrates a yearly hydrograph observed at the Mistassibi station, and it can be seen that each detection test possesses different behavior in the diagnostic of outliers. Test 1 and Test 2 (see Figure 6.2a and b) check the difference between data points, and the point under investigation is then considered as an outlier if this difference exceeds a given threshold.

Table 6.8. Activation of detection tests for outliers.

	Number of times each test is activated (i.e., diagnosis of corruption)					
Test	C	H	ME	MI	S	G
Test 1	842	134	371	373	978	43
Test 2	249	9	91	48	333	0
Knet 5	1,131	61	434	317	507	349
Fuzzy 5	826	102	263	204	483	211

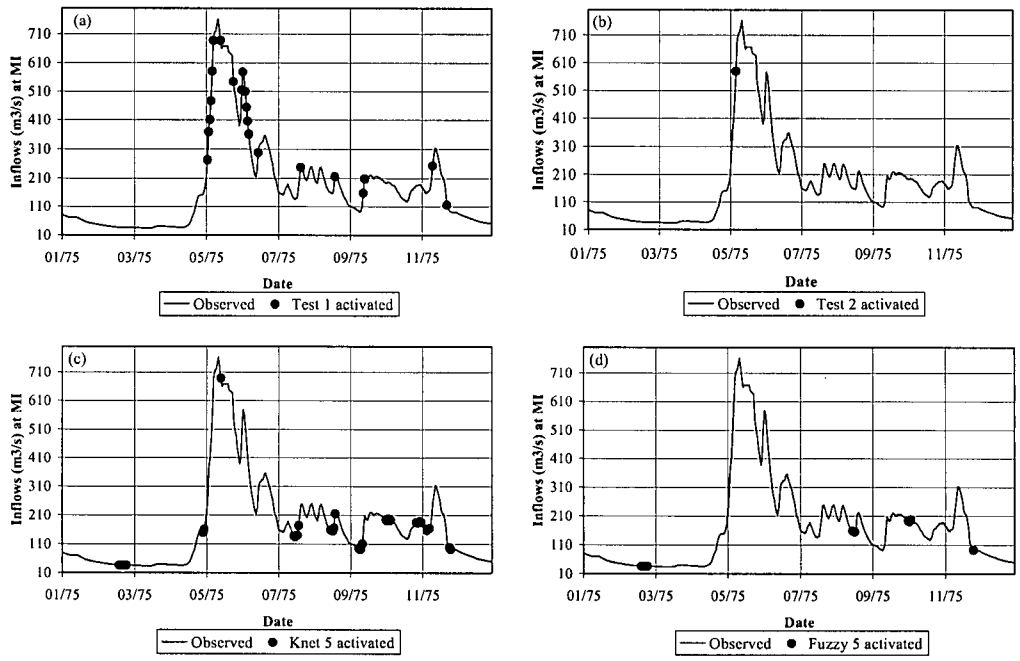


Figure 6.2. Activation of tests on observations at MI.

Physically, this means that a difference is too large to be explained by the response of the watershed to a precipitation impulse (case of a rise) or by the capacity of the watershed to drain itself (case of a recession). This explains why Test1 and Test 2 are activated during the swift rises and recessions present in the hydrograph, because only such events can exceed the threshold values. Test 2 is less sensitive than Test 1, because Test 2 is based on the differences of inflows (or water levels) prior to and posterior to the point under investigation, while only the difference prior to the point is considered in Test 1. Knet 5 and Fuzzy 5 (see Figures 6.2c and d) are activated under different situations from those activating Test 1 and Test 2. Knet 5 and Fuzzy 5 are built to identify patterns, where

an outlier would disrupt the continuous progression of a given process. Assume a situation whereby a process would lead to discrete observations that remain constant over time (i.e., yielding a straight line). In this case, even an observation that only slightly departs from the straight line could possibly be considered as an outlier. This explains why Knet 5 and Fuzzy 5 can be activated by points that show little variation from the surrounding points. Of course, there can be situations when conventional and AIT-based tests converge to the same diagnostic, and this is the situation in Figure 6.2, where, at the end of November, Test 1, Knet 5 and Fuzzy 5 are activated by a swift recession.

Convergence in the diagnostics of tests is more noticeable in Figure 6.3, which illustrates a yearly hydrograph observed at the Coquihalla station.

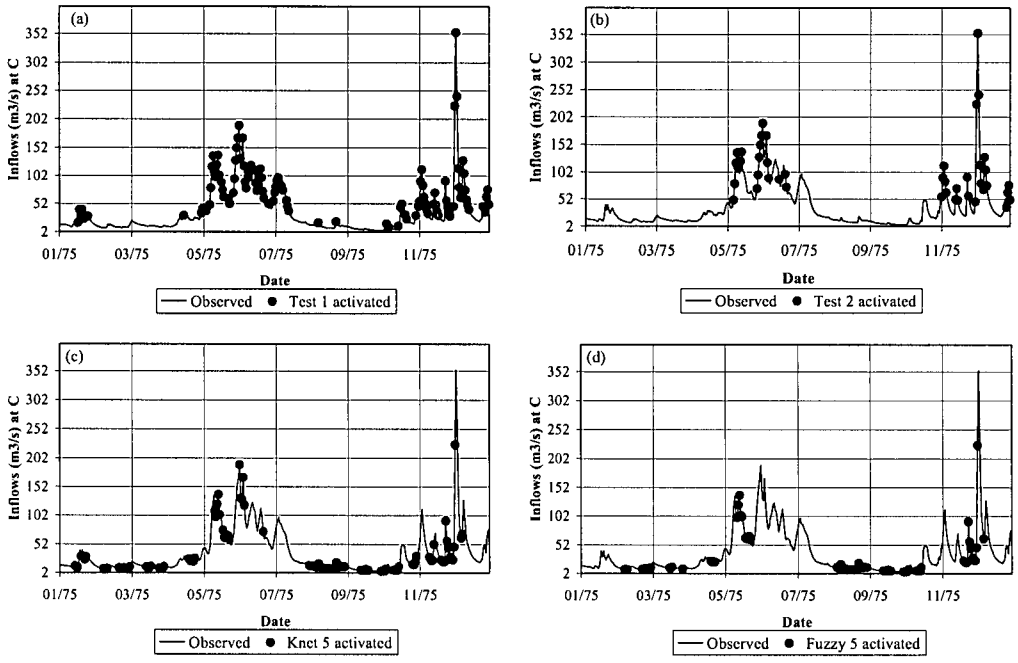


Figure 6.3. Activation of tests on observations at C.

Figure 6.3 shows that the AIT-based tests can also be activated by situations that normally activate conventional tests. All tests are very sensitive to small watersheds, for they are activated frequently on the data of the Coquihalla station, as seen in Figure 6.1, and also on the data of the San Juan station. These small watersheds have likely fast response times and frequent large precipitation impulses, creating swift rises and recessions

that are misleading for the detection tests, as illustrated by the number of activations during summer and fall in Figure 6.3. These swift rises and recessions particularly affect Test 1 and Test 2, since they base their decision about corruption and non-corruption directly on differences between data points. AIT-based tests are also affected by these swift rises and recessions although to a lesser extent. The converse observations are made for the Harricana station, where inflows have a slow response time and few large precipitation inputs exist, producing smoother hydrographs than those on the Coquihalla and San Juan watersheds. As a result, all tests are activated much less often for the Harricana station, as indicated in Table 6.8.

All tests are not activated often for the Saint-Gédéon station as well. This station measures the water level of the Saint-Jean Lake, which has a rather large retention time and therefore leads to very small daily variation in the level of the Lake. Test 2 is not even activated for the whole 10 years of data. Test 1 is activated only during spring, when the variation in the lake level is the highest. Knet 5 and Fuzzy 5 favor the fall season, when the observations of water levels are somewhat noisy.

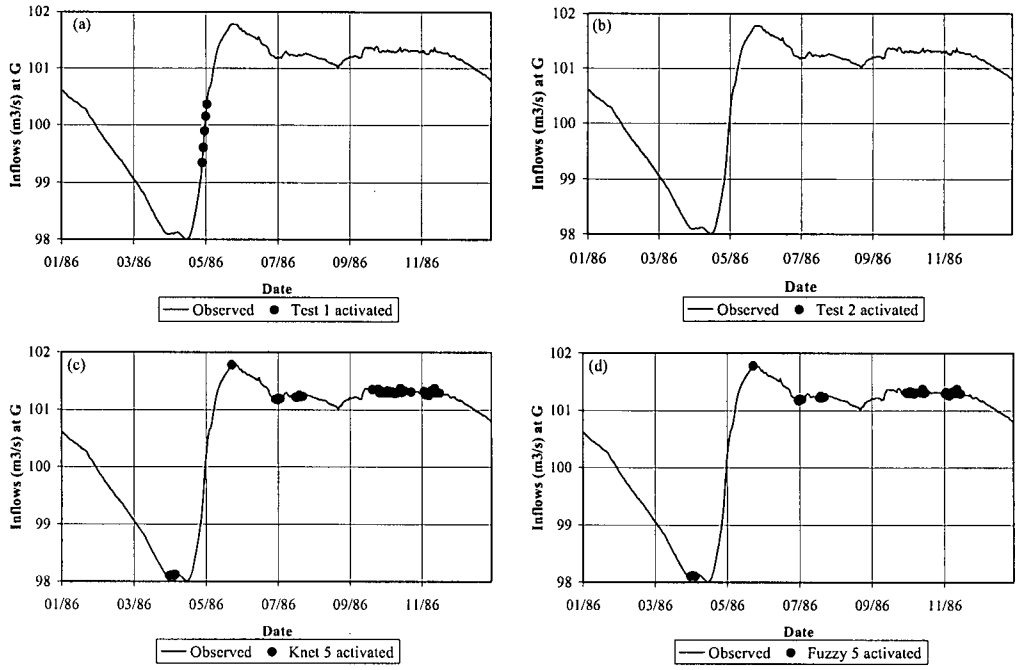


Figure 6.4. Activation of tests on observations at G.

The results presented in this section show the differences in the behavior of the tests, and effort must be made to take advantage of these differences. Considering more than one approach results in a diversity of solutions, where the weaknesses of some can be compensated with the forces of others. Similarly to what is shown in Section 6.1 for the application to shifts and trends, effort should be made to build a framework by which several tests are employed and used to complement each other in the decision-making process. All tests presented here are assumed to be equally valid and could all be used in this framework. AIT-based tests can be sensitive to small noises, but safeguards can be added to them to prevent false detection in such situations. Such an application to real data helps to understand the strengths and weaknesses of all tests, and these can be taken into account in the development of a decision framework. The issue of finding an efficient method to determine threshold values for all tests should also be of interest. It is obvious from the results here that direct transfer of the threshold values from one station to another does not constitute an entirely satisfying solution (e.g., transfer from Mistassibi to Coquihalla). This method of finding thresholds should better take into account the specific characteristics of the data sequences under investigation, as well as associated information (e.g., hydrologic regime, size of the watershed). It must be said that a threshold value for a given station need not be constant under all circumstances. It can possibly vary with respect to time or specific hydrologic conditions.

6.3 Conclusion

This chapter shows the viability of applying AIT-based tests to real data. The largest disadvantage of the AIT-based tests presented in this work is the computational burden they exert. They take more time to implement than the conventional tests, yet they do not yield significant advantages in terms of performance for the detection of anomalies. However, AIT-based tests should not be discarded solely based on this argument, because they do represent a distinct alternative to conventional tests. AIT-based tests for the detection of outliers exhibit a definitely different behavior from the conventional tests, as attested by the results shown in this chapter, and the same can be said to some extent of the AIT-based tests for the detection of shifts and trends. AIT-based tests detect anomalies that are not detected by the conventional tests and vice versa. All these tests, AIT-based and

conventional, should be considered as a set of tools, which might be combined to enhance performance in detecting anomalies.

Even though this issue has not been fully explored in this chapter, recall that the AIT-based tests seem to be attractive tools for the estimation of the amplitude of anomalies, as indicated in Chapter 5 (i.e., Amp/CV for shifts and trends, and Amp/SD for outliers). One may not always be interested in the strict detection of anomalies. If one does not want to risk making a false detection, then one can simply assume that anomalies are possibly present based on the estimation of factors such as the Amp/CV or Amp/SD ratio, and quantify the impacts on the calculation of water quantity or quality for a given water resources system with respect to these ratios.

Chapter 7
AIT Approaches for Model Parameters

In addition to data inaccuracies or anomalies that can greatly affect the results of a simulation model, uncertainties that originate from the structure or the parameters of the model cannot be neglected. This chapter is therefore dedicated to the issue of parameter uncertainties. This is in line with the material discussed in Section 2.1, which states that both data and parameters are inputs to models. They are therefore treated in this work through the same line of reasoning, that is, with focus on the description of input (i.e., either data or parameters) domains. Chapters 4 to 7 focus on the issue of data and parameter uncertainties as described in Section 2.1. A third source of uncertainties, related to the model structure, is not addressed here. It is outside the focus of this work, that is, the assessment of inputs. However, the results presented in this chapter may imply links between parameter and model structure uncertainties.

The applications presented here involve two very different problems and systems. Nevertheless, they have two points in common: the solution employed to reduce the magnitude of parameter uncertainties, and the calibration process used to calibrate the model parameters. These two points are detailed in Section 7.1, while the applications to water inflow modeling and algae concentration modeling and their respective specificities are given in Sections 7.2 and 7.3, respectively. Section 7.4 gives the general conclusions that can be drawn from the results of both applications.

7.1 Common Elements of the Applications

7.1.1 Description of the Parameter Domain

This chapter proposes a method that allows for a more flexible determination of the values of some of the parameters of a simulation model so as to help reduce the magnitude of uncertainty associated with model parameters. The method, which makes use of fuzzy logic, is based on the same line of reasoning employed for dealing with outliers, shifts and trends, that is, the description of the input domain, this time parameters instead of observed data. The description of the parameter domain is established with respect to indicators that

are representative of the conditions of the systems. Quite often parameters in simulation models are given constant values while it can be regularly assumed that they vary with respect to the conditions of the systems. An example is that of runoff coefficients used in lumped conceptual watershed models, where it is acknowledged that the values of such parameters would vary with respect to the level of humidity in the soil. When soil humidity is not observed or cannot be reliably estimated, indicators of the level of humidity in the soil at any given time, such as previous precipitation and inflow records, can be employed to determine the values for the runoff coefficients. The greater freedom given to the determination of the values of the parameters can therefore allow an improved replication of the response of the system under study.

The modeling structure that combines the description of the parameter domain with a given simulation model is given in Figure 7.1.

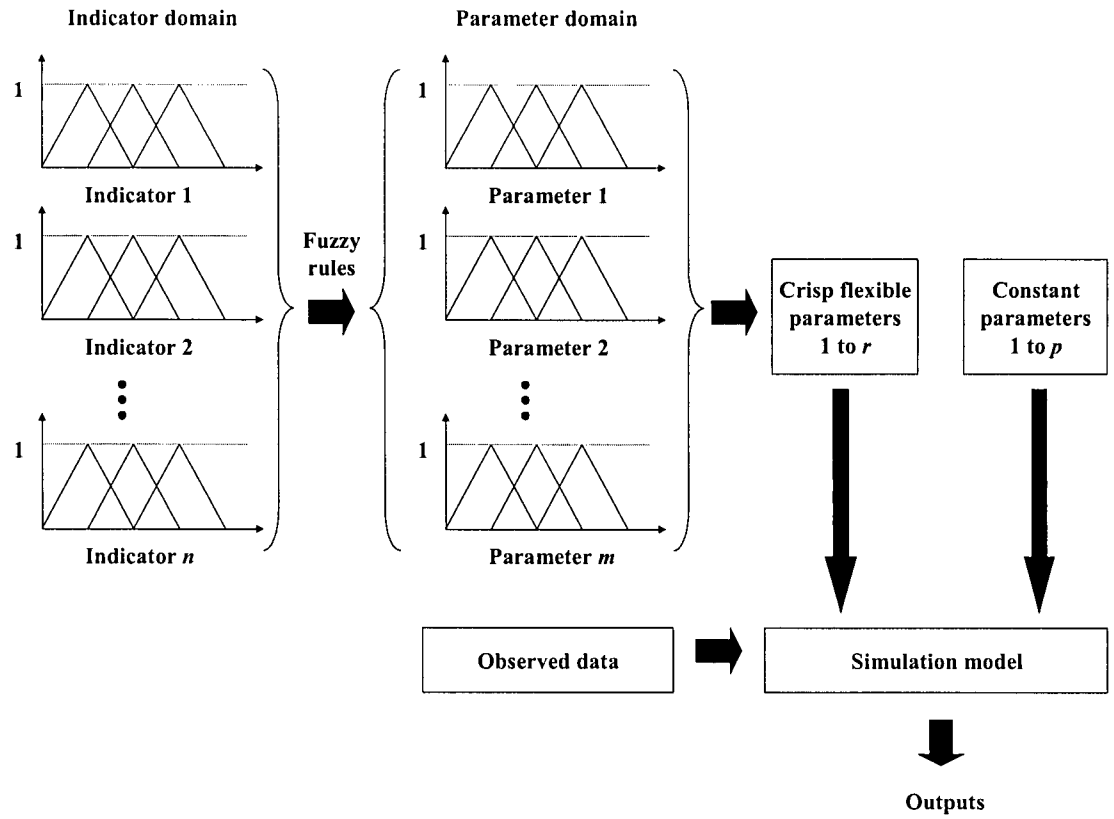


Figure 7.1. Hybrid fuzzy logic and simulation model.

Here fuzzy sets and fuzzy logic are employed in the description of the parameter domain for they allow a structured inference of the values of the parameters as a function of the values of the indicators, such as temperature or nutrient for algae concentrations. Also, they can easily be integrated in the structure of the simulation model, which facilitates the calibration process. In Figure 7.1, the domain of the indicators is described using fuzzy sets. Then rules that relate the parameter domain to the indicator domain can be built. For example, if temperature and the nutrient load are high, the algae growth rate, which is a parameter in the algae concentration model presented in this chapter, will have a given value, most likely a high one. From that point, given a set of indicator values, the calculation of the degree of fulfillment for each rule, the combination of the responses of the rules and the defuzzification process as detailed in Section 3.2 are employed to provide crisp parameter values (i.e., the crisp flexible parameters 1 to r in Figure 7.1). The product inference with the AND operator is taken for the determination of the degree of fulfillment of each rule. The normed weighted sum is employed for the combination of the results of the rules, and then the mean defuzzification is applied to yield crisp parameter values. The relationship that encompasses the aforementioned steps is that of Equation 3.15, repeated here as Equation 7.1:

$$m(p_j) = \frac{\sum_{i=1}^{I} v_i m(p_{j,i})}{\sum_{i=1}^{I} v_i} \qquad 7.1$$

The variable $m(p_j)$ is the mean of parameter p_j, for $j = 1, \ldots, r$, $m(p_{j,i})$ is the mean of parameter p_j associated with rule i, and v_i is the degree of fulfillment of rule i. In the context of the applications in Sections 7.2 and 7.3, the degree of fulfillment of all rules can be known given a set of indicator values. The only unknowns of Equation 7.1 are the $m(p_{j,i})$s, and they are determined through the same calibration procedure employed on the model for constant parameters 1 to p in Figure 7.1. In this work, genetic algorithms are used to obtain the values of the constant parameters and the $m(p_{j,i})$s.

7.1.2 Optimization with Genetic Algorithms

Genetic algorithms are considered a competitive alternative to more conventional optimization procedures such as linear or non-linear programming, particularly when the optimization problem has few constraints. Unlike linear and non-linear programming procedures, genetic algorithms easily accept discontinuities present in the formulation of the problem. They are also less dependent on initial conditions as they consider more than one set of initial conditions simultaneously, which reduces the risk of ending the optimization process at a local optimum. Genetic algorithms are described in details in Goldberg (1989), and only the basics applied to the applications of this chapter are given here. The term genetic algorithm comes from the analogy between this method and biology and, more particularly, genetics. First, the decision variables (i.e., the flexible and constant parameters of the simulation model) for a given optimization problem are coded so as to be represented altogether by a string of characters, very much like genetic material. Quite often a string is made of zeros and ones so as to give a binary code. Any string has its own fitness or robustness value, which is the value of the objective function of the optimization problem obtained with the values of the parameters represented in the string. In the applications of this chapter, the objective function is the inverse of the sum of the square difference between observed and calculated outputs. Hence, a set of parameter values that does not lead to a good fit on the observed output yields a weak fitness value, a set leading to a good fit yields a strong fitness value, and a perfect fit results in a fitness value equal to infinity. With genetic algorithms, a population of strings is initially created and, through a series of operations that are equivalent to the evolution phenomena in nature, the strings are updated so as to produce a population whose average fitness becomes better as the algorithm proceeds from one iteration or generation to another. The most common operations are reproduction, crossover and mutation, and are repeated one after another at any generation (iteration).

By the process of reproduction, an entire string at a given generation can be allowed to be present in the next generation. If string ABCDE is allowed to reproduce once at a given generation, then one string ABCDE will be present in the population in the next generation. This is a random procedure which gives a preference to the best fitted strings. The probability that a particular string i be chosen is f_i/f_{tot}, where f_i is the fitness value of

string i, and f_{tot} is the sum of the fitness values of all the strings of the population. To renew a population of n strings for the next generation, the reproduction process must be conducted n times. The reproduction process is normally one with replacement, which means that the respective probability of the strings to be picked remains constant from one draft to another until a full population is built. Of course, by this process, the strongest strings are likely to be picked more often and therefore will likely increase the total fitness of the population. The reproduction process allows the population to become stronger at every generation, but does not perform any modification in the strings. This means that, if the optimum point is not represented among the strings in the population, there is no possibility to converge to the optimal solution. The crossover process allows modifications on the string or, more specifically in biological terms, allows the exchange of genetic material. Let's consider the following two strings: ABCDE and abcde. By the crossover process, first, the location of a breaking point in the strings is chosen, second, all the elements of the strings that are on the right of the breaking point are traded between the strings. For example, if the breaking point is between the second and third elements in the two strings above, then, after the trade, the resulting strings are: ABcde and abCDE. Thus, at every generation, the strings in the population are grouped by pair. Usually, for each pair, there is a given probability that the crossover occurs. If crossover does occur for a given pair, a breaking point location is chosen randomly and the trade is performed. The mutation process allows a change in the value of only one element in a string. For example, consider a particular string ABCDE and its resultant after mutation ABCDX. Obviously the mutation occurred on the last element of the original string. Biologically, mutations occur frequently and ensure the genetic diversity of a species, which, by this fact, become likely more resistant to diseases or other catastrophes. For optimization purposes, the mutation process must be considered as a safety guard against local optima. For some reasons, the general evolution of a population may be directed towards a local optimum, but, by the mutation process, new strong strings directed towards the global optima could be created and, through reproduction and crossover, could change the trend of the population favorably. Note that the reverse situation, that is to move from a global optimum to a local one because of mutation, can also be possible. The mutation process, like the reproduction and crossover processes, is also random. A mutation occurs upon a given probability,

which is usually low so as to ensure that the creation of the population does not become a pure random process, independent of the information contained in the previous generation. Genetic algorithms use operations that rely heavily on random procedures, yet Goldberg (1989) demonstrates that improvement in the overall fitness increases from one generation to another as long as the probability of mutation remains very small.

To conclude, recall the procedure to follow with any optimization problem:

1. Code the decision variables (i.e., parameters) and the objective function to be compatible with the use of genetic algorithms,
2. Initially, create a random population of strings, which would be the first generation
3. At every generation, use the reproduction, crossover and mutation phenomena to create a new generation,
4. Always keep in memory the best string at every iteration, but keep updating the population until a defined number of generations is reached or some other stopping criteria are satisfied.

7.2 Application to Inflow Modeling

The watershed model employed in this application is based on the TANK model, described among others by WMO (1992), Singh (1995), and Rousselle et al. (1999). It is a relatively typical model, comprised of a series of conceptual reservoirs, each of which with outlets that describe the fractions of water contributing to the inflow, being transferred from one reservoir to another or being lost through evapotranspiration or groundwater transfer. Originally, the parameters of this model are constant regardless of the watershed conditions, and this weakness is corrected with some of the parameters, through the use of fuzzy sets for the description of indicator and parameter domains, and of fuzzy logic for the determination of the parameter values that are the most appropriate for the conditions of the watershed. In a first step, the model is described, along with the integration of the fuzzy logic inference engine within the structure of the model. The database employed to test the model is then given, followed by the protocol of the experiment. The results and the ensuing conclusions are at last presented in detail.

7.2.1 Description of the Model

Given the initial conditions of the watershed, a water input from a rainfall event or snowmelt would be routed through the surface and the ground to ultimately reach the outlet of the watershed. The process of vertical transfer of water from the surface to the ground can be defined by a vertical series of reservoirs, where the top reservoir receives the water input, releases some quantity for the outlet of the watershed, directs another amount out of the system to represent losses, and releases the remaining water to the reservoir below. The reservoir below accomplishes the same function as the first one for an underlying layer of soil, and so on. This idea of reservoirs in series is the structural basis of many of the existing deterministic conceptual inflow models, and it implies that the inflow at the outlet is the sum of sub-inflows coming from each of the reservoirs. Figure 7.2 shows a typical set of reservoirs in series based on the TANK model, and which is used in this work. The physical meaning of each of the water transfers is provided in Figure 7.2, and the losses are generally considered as due to the effect of evapotranspiration or groundwater transfer that never reaches the outlet of the watershed. The water inflow going out through a reservoir sluice (O_t) at time t is usually defined by a power function:

$$O_t = aH_t^b \tag{7.2}$$

where H_t is the level of water in the reservoir at time t, and a and b are coefficients. To account for the travel time to reach the outlet of the watershed, most of the sub-inflows are spread over time, which is in agreement with the well established concepts of hydrographs and hydrograph separation as a function of the types of sub-inflows. The concepts of hydrographs and hydrograph separation are illustrated in Figure 7.3. The contributors to the hydrograph, as indicated in Figure 7.3, are the sub-inflows from the reservoirs shown in Figure 7.2, that is, the surface runoff, the interflow and the baseflow from groundwater. A discrete mathematical model of daily inflows based on such contributions can be of the form:

$$Q_t = \frac{10^6}{86400} A\left(I_{s,t} + I_{w,t} + I_{g,t}\right) + e_t \tag{7.3}$$

where Q_t is the inflow at day t (m^3/s), A is the area of the watershed (km^2), $I_{s,t}$, $I_{w,t}$ and $I_{g,t}$ are respectively the total daily surface, interflow and groundwater contribution (m), e_t is a

residual (m^3/s), and the numerical constant accounts for the conversion of units of area (km^2 to m^2) and time (day to s).

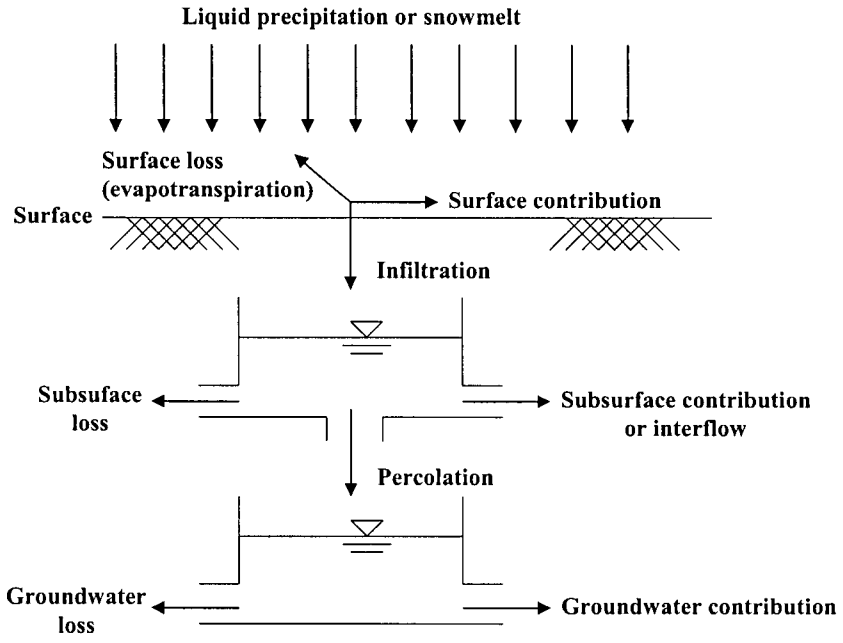


Figure 7.2. Conceptual reservoirs of a hydrologic inflow model.

In the model employed here, the total surface runoff and interflow contributions are the sum of contributions from the previous days until day t, as defined by their respective hydrographs or impulse-response functions, which means that:

$$I_{s,t} = \sum_{\tau} i_{s,t-\tau} \qquad 7.4$$

and:

$$I_{w,t} = \sum_{\tau} i_{w,t-\tau} \qquad 7.5$$

where $i_{s,t}$ and $i_{w,t}$ are the surface runoff and interflow contribution associated with day t, respectively, and τ is the time lag. In the model, the total groundwater contribution at day t is simply represented by the inflow from the sluice of the groundwater reservoir at day t, as defined by Equation 7.2.

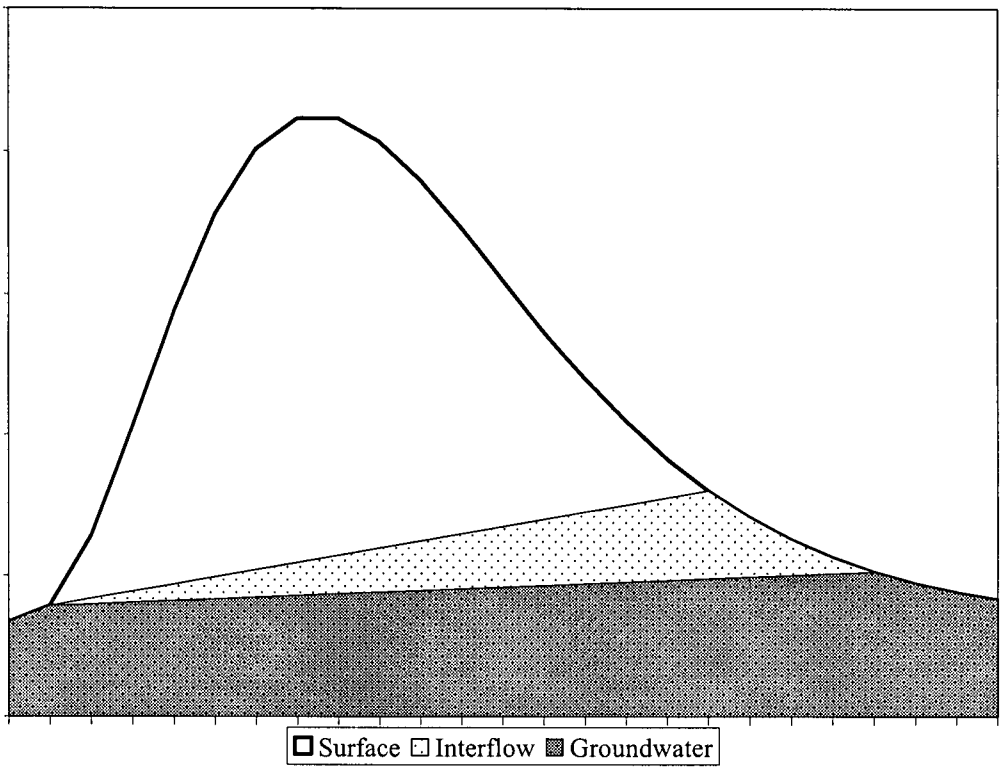


Figure 7.3. Separation of sources of inflow on an idealized hydrograph.

As indicated in Equations 7.4 and 7.5, the surface runoff and interflow contributions can be represented following their respective hydrograph also referred to as the impulse-response function. The equation of the hydrograph can be written as:

$$i_{t+\tau} = R_t C_t f_{t+\tau} \qquad 7.6$$

where R_t is equal to the quantity H^b in Equation 7.2, C_t is a production factor that is equal to a in Equation 7.2, and f_t is a function that represents the hydrograph and therefore spreads the contribution over time. This function can be determined manually using inflow measurements from the watershed in conjunction with precipitation data. However, this procedure is inaccurate because there is no easy way to distinguish between surface inflow, interflow and groundwater flow just by looking at the total hydrograph. One alternative is to approximate the shape of the hydrograph, or f_t using a triangle (Rodriguez-Iturbe and Valdes, 1979). The time to peak (t_p) and the duration (t_d) of the hydrograph are enough to describe the shape of the triangle. This approximation is used in this work, as well as another, which is to consider the hydrograph as a Gamma function:

$$f_t = \frac{\lambda}{\Gamma(\alpha)}(\lambda t)^{\alpha-1} e^{-\lambda t} \qquad 7.7$$

In Equation 7.7, α and λ are functions of the time to peak (t_p) and the duration (t_d) of the hydrograph. The relationships between α, λ, t_p and t_d are defined as:

$$t_p = \frac{\alpha - 1}{\lambda} \qquad 7.8$$

and:

$$\int_0^d f(t)dt = K \qquad 7.9$$

In equation 7.9, K is the proportion of the area of the Gamma function considered as part of the hydrograph. The whole area of the Gamma distribution is obtained only when t equals infinity, which is of no practical interest. Only a part of the area can be accounted for, but it can nevertheless be a very large part (e.g., K may equal as much as 97.5%). The unaccounted for part of the hydrograph is often considered negligible relative to the potential accuracy of the model. The Gamma function has been chosen for its similarity to hydrographs that are often observed in nature. Figure 7.4 illustrates some examples obtained with the Gamma functions for different values of t_p and t_d.

Obviously, inflows are triggered by water input received at the surface of the watershed. When there is no snow cover, the water input is simply equal to the liquid precipitation. When a snow cover is present, a snowmelt model must be used to determine the quantity of water available for runoff. One of the models employed here is described by the following equation:

$$R_t = 0.01 \times \max(S_{t-1} - S_t, 0) \qquad 7.10$$

and:

$$\begin{cases} S_t = S_{t-1} - \left(d + \dfrac{p_{liq,t}}{80}\right)(t_{mean,t} - t_b) + 0.10\,p_{sol,t} & \text{if } t_{mean,t} > t_b \\ S_t = S_{t-1} + 0.10\,p_{sol,t} & \text{if } t_{mean,t} \le t_b \end{cases} \qquad 7.11$$

where S_t is the water equivalent of the snow pack at day t (cm), d is the degree-day factor ($cm \cdot °C^{-1} \cdot day^{-1}$), and t_b is the base temperature (equal to 5°C for this application). Also, p_{liq}, p_{sol} and t_{mean} are the daily liquid precipitation, solid precipitation and mean temperature, respectively. The constant factor 0.10 before $p_{sol,t}$ is the assumed value of snow density, in order to obtain the water equivalent from the solid precipitation. Equations 7.10 and 7.11

constitute a simple model of snowmelt which combines the degree-day method as described in Gray and Male (1981) and Rango and Martinec (1995), and the energy contribution of the rain following the relation given in Linsley et al. (1982). This snowmelt model is calibrated by itself so as to provide the data needed for the inflow model developed in this work.

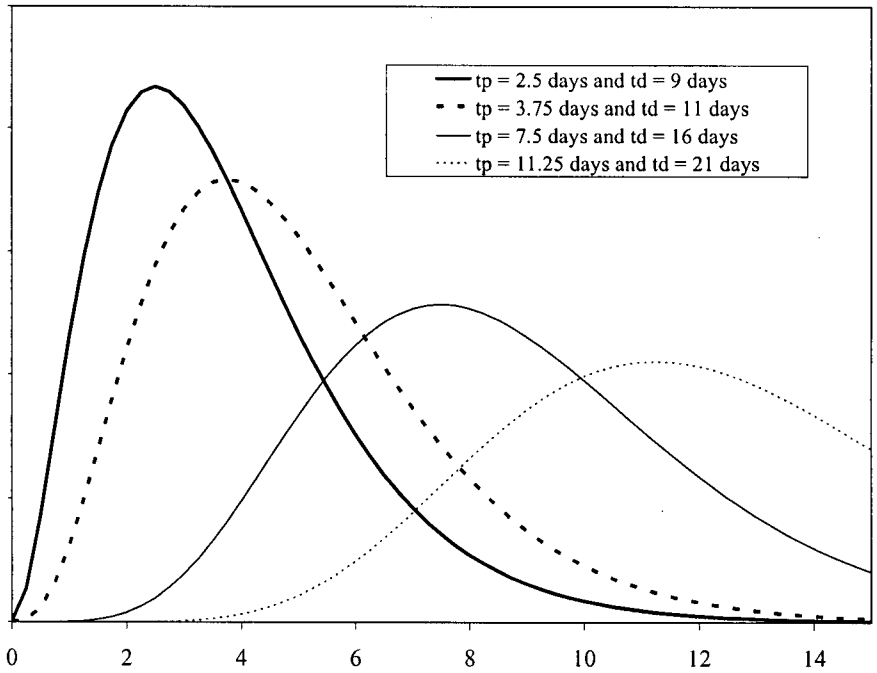


Figure 7.4. Examples of hydrograph as obtained with the Gamma function.

An equivalent, but slightly more complex model, described in Bouchard (1986) is also tested in this application. In the model, Equation 7.10 remains, but Equation 7.11 is replaced by Equations 7.12 through 7.15:

$$\begin{cases} S_t = S_{t-1} - C_1 \cdot SUME_t \cdot E_t^{1.4} + C_2 SUMP_t \cdot p_{liq,t} + 0.10 p_{sol,t} & \text{if } t_{mean,t} > 0 \\ S_t = S_{t-1} + 0.10 p_{sol,t} & \text{if } t_{mean,t} \leq 0 \end{cases} \qquad 7.12$$

with:

$$E_t = \frac{2t_{\max,t} - t_{\min,t}}{3} - t_b \qquad 7.13$$

$$\begin{cases} SUME_t = \dfrac{\sum_t E_t}{MAXSUME} & \text{if } \sum_t E_t < MAXSUME \\ 1 & \text{if } \sum_t E_t \geq MAXSUME \end{cases} \quad 7.14$$

and:

$$\begin{cases} SUMP_t = \dfrac{\sum_t p_{liq,t}}{MAXSUMP} & \text{if } \sum_t p_{liq,t} < MAXSUMP \\ 1 & \text{if } \sum_t p_{liq,t} \geq MAXSUMP \end{cases} \quad 7.15$$

where $t_{max,t}$ and $t_{min,t}$ are the daily maximal and minimal temperature, respectively, and the elements C_1, C_2, t_b, *MAXSUME* and *MAXSUMP* are parameters that must be calibrated. The distinct features of this snowmelt model are the weighting functions, $SUME_t$ and $SUMP_t$, which reduce the amount of snowmelt at the beginning of the spring season to account for the delay in the snowmelt process due to the energy mass contained in the snow cover. In Equations 7.14 and 7.15, the summations of E_t and $p_{liq,t}$ start at the beginning of the snowmelt season, which is the end of March for the watersheds studied in this application. As with the first snowmelt model, the parameters of the second model are calibrated by themselves so as to provide the data needed for the inflow model developed here.

Considering two different hydrograph shapes and two different snowmelt models offers the possibility to construct four different modeling scenarios. This number of scenarios is expanded to account for the variants of the parameter domain tested in the application, that is:

1. The standard variant (SV), where there is no description of parameter domain, and therefore where all parameters remain constant,
2. The flexible hydrograph variant, one indicator (H1), where the description of the domain of the parameters related to the hydrograph functions is performed using the 7-day averaged inflow as the indicator,
3. The flexible hydrograph variant, three indicators (H3), where the description of the domain of the parameters related to the hydrograph functions is performed using the 7-day averaged inflow, precipitation and temperature as indicators,

4. The flexible soil variant, one indicator (S1), where the description of the domain of the parameters related to the outlet of the conceptual reservoirs is performed using the 7-day averaged inflow as indicator.

Hence, considering this latter subdivision, a total of sixteen model scenarios can be constructed.

With variants H1 and H3, the flexible parameters are the *a*s for the surface and interflow contribution rates ($a_{s,t}$ and $a_{w,t}$), often called runoff coefficients, and the time to peak and duration for the surface runoff and interflow hydrographs ($t_{p,s,t}$, $t_{p,w,t}$, $t_{d,s,t}$ and $t_{d,s,t}$). As a consequence, the constant parameters are 1) the *a*s for all the loss components (surface, subsurface and groundwater); 2) the *a*s respectively related to the groundwater contribution to the watershed outlet, and percolation (from the intermediate to the groundwater reservoir); and 3) all the *b*s (i.e., 7 *b*s in total, with the one for the surface contribution assumed to be equal to 1). One last constant parameter is the one that sets the level of the groundwater reservoir on March 31 of each year, the level of the subsurface reservoir being always assumed to be equal to 0 on March 31. Note that the water that infiltrates is the remains of the precipitation once the surface contribution and loss are accounted for. All the flexible parameters in variants H1 and H3 are related to the hydrograph, and the flexibility allows the implementation of the statement advocated by Rodriguez-Iturbe and Valdes (1979), which stipulates that the properties of the hydrograph vary with respect to the conditions of the watershed. For example, it can be assumed that, during large rainfall events, the surface runoff time to peak ($t_{p,s}$) is shortened and the surface runoff production coefficient $a_{s,t}$ increases. Basically, a greater part of the area of the watershed becomes water saturated, thus a greater part of the watershed contributes to the surface runoff, which explains the increased value of $a_{s,t}$. At the same time, velocities increase on the surface due to the higher water depth on the surface during large rainfall events, and this explains the reduction of the time to peak. The reverse conclusions that result from small rainfall events are also true, and similar kinds of reasoning can be used to explain interflows. Such kinds of inferences constitute one way practitioners often use to adjust conceptual inflow models, and in this application, fuzzy logic performs these inferences based on the description of the parameter domain with respect to the indicators. The 7-day averaged inflow for H1 and H3, plus the 7-day averaged precipitation and 7-day

averaged temperature for H3 are chosen as indicators because they can indirectly provide some insights about soil moisture, which is one important condition of the watershed that may not be well accounted for by the structure of the model in the standard variant (SV).

With variant S1, the flexible parameters are the *a*s for the surface, interflows and groundwater contributions to the inflows, the *a*s for the surface, subsurface and groundwater loss components, and the *a* for the percolation. This variant has been added as a consequence of a preliminary analysis of the results of variants H1 and H3, where it is noticed that the parameters that benefit most from added flexibility are the contribution rates ($a_{s,t}$ and $a_{w,t}$). In variant S1, all the flexible parameters are contribution rates (i.e., the *a*s), which are related to the movement of water within the ground. The rationale for this variant is that these flexible parameters can interact, for it is indeed normal to assume an action-reaction chain occurring among these parameters. All the flexible parameters should be sensitive to soil moisture conditions on the watershed, and therefore the 7-day averaged inflow can be employed as an indicator.

The details of the procedure to perform the description of the indicator domain, and incidentally of the parameter domain through inferences with fuzzy logic are provided in Section 7.1.1, but below is a summary of this procedure as applied to this case of inflow modeling

1. The domain of each indicator is described by Gaussian membership functions so as to avoid any discontinuity in the inference process,
2. For each rule, algebraic products link the indicators together,
3. The combination of rules is achieved using the normed weighted sum procedure,
4. The mean deffuzification process is employed to obtain a crisp value for the flexible parameters.

A crisp value for a given parameter is given by Equation 7.1, presented again in Equation 7.16.

$$m(p_j) = \frac{\sum_{i=1}^{I} v_i m(p_{j,i})}{\sum_{i=1}^{I} v_i} \qquad 7.16$$

where variable $m(p_j)$ is the mean of parameter p_j, for $j = 1, ..., r$, $m(p_{j,i})$ is the mean of parameter p_j sets for rule i, and v_i is the degree of fulfillment of rule i. In the context of the

applications, the values for the $m(p_{j,i})$ are not known and must be calibrated the same way as the constant parameters in the model.

Genetic algorithms are employed for the calibration of the parameters, based on the minimization of the sum of the difference between observed and calculated inflows, or rather the maximization of the inverse of this sum. In Equation 7.3, the difference between observed and calculated inflow is represented by the residual term e_t. When this term e_t is isolated in Equation 7.3, the objective function (i.e., fitness function) to be used in this application is:

$$\max\left(\sum e_t\right)^{-1} = \max\left(\sum\left\{Q_t - \frac{10^6}{86400}\left(I_{s,t} + I_{w,t} + I_{g,t}\right)\right\}\right)^{-1} \qquad 7.17$$

7.2.2 Description of the Application Case

The data collected on the Ashuapmushuan River, Mistassibi River and Chute du Diable watersheds are employed in this application. These three watersheds are located in the Saguenay-Lac-Saint-Jean hydrographic system in the Province of Quebec, Canada, as indicated in Figure 7.5. These watersheds are respectively 15,300, 9,320 and 9,700 km^2, which is large enough to allow inflow modeling using daily time steps. On these watersheds, the annual peak flood usually results from the snowmelt during spring, but isolated precipitation events can generate significant flow peaks during summer and fall. These watersheds have significantly different shapes and this results in different hydrological responses. Daily sequences of inflows, temperatures and liquid and solid precipitations are used in this application, along with the estimation of the water equivalent of the snowpack for all three watersheds. All sequences range from 1963 to 1994. The data from 1963 to 1984 are used for the calibration while the data from 1985 to 1994 are employed for the validation of the adaptive model.

With three watersheds, two hydrograph shapes, two snowmelt models, and four variants in terms of the description of parameter domains (SV, H1, H3, and S1), a total of 48 modeling scenarios are analyzed. The calibration process involves many parameters for all scenarios, and therefore can be lengthy. The strategy is to calibrate the parameters of the model in variant SV, for all three watersheds. In this variant, all 18 parameters are constant. In variants H1, H3 and S1, respectively 12, 12 and 11 parameters remain constant and are

given the values determined during the calibration for variant SV. Therefore only the flexible parameters are calibrated in variants H1, H3 and S1. The flexible parameters are represented by the $m(p_{j,i})$s, and the number of these latter parameters varies from one variant to another. With variant H1, the indicator, the 7-day averaged inflow, is described using 3 fuzzy sets, leading to the formation of three rules per parameter, and consequently to the attribution of three $m(p_{j,i})$s per parameter. Thus, with variant H1, a total of 18 $m(p_{j,i})$s must be calibrated to describe all 6 flexible parameters. With variant H3, the three indicators are each described with 2 fuzzy sets, leading to the formation of 8 rules (2×2×2) per parameter and the calibration of 48 $m(p_{j,i})$s to describe all 6 flexible parameters. With variant S1, the indicator, the 7-day averaged inflow, is described with three fuzzy sets, leading to three rules per parameter and the calibration of 21 $m(p_{j,i})$s to describe all 7 flexible parameters.

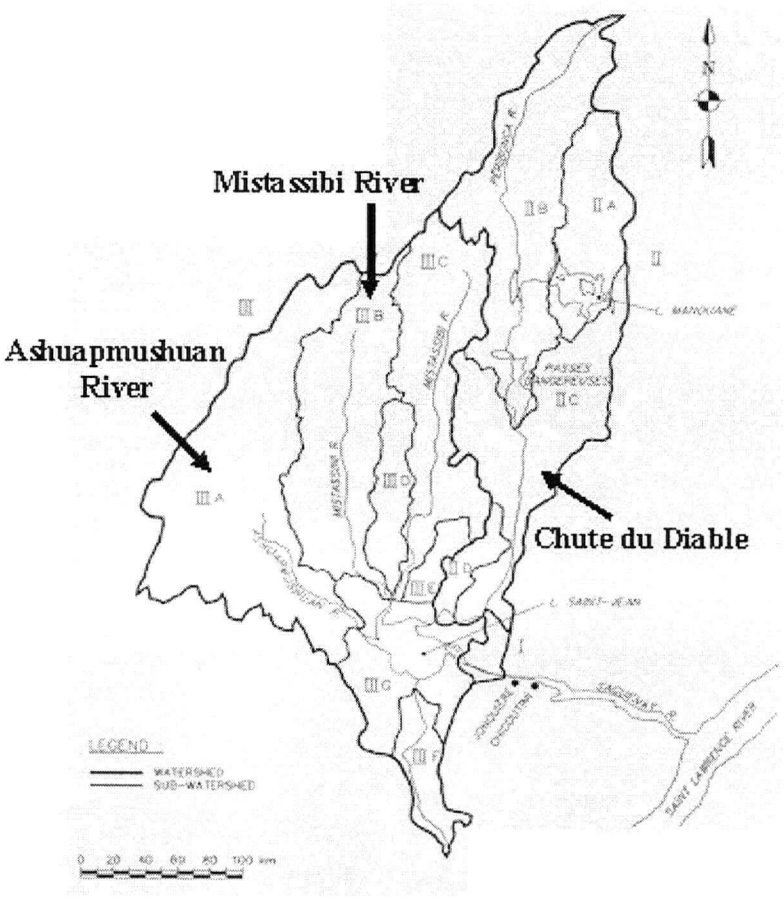


Figure 7.5. Saguenay-Lac-Saint-Jean hydrographic system.

For the evaluation of the results, the differences in the annual means and standard deviations between the observed and calculated inflow sequences are analyzed, and so are the correlation coefficient (CCR_j) and the Nash ($Nash_j$) coefficient. These criteria are determined annually, and their equations are as follows:

$$CCR_j = \frac{\sum_{t=1}^{N}(Q_{obs,t} - \overline{Q}_{obs})(Q_{cal,t} - \overline{Q}_{cal})}{\sqrt{\sum_{t=1}^{N}(Q_{obs,t} - \overline{Q}_{obs})^2 \sum_{t=1}^{N}(Q_{cal,t} - \overline{Q}_{cal})^2}} \qquad 7.18$$

and:

$$Nash_j = 1 - \frac{\sum_{t=1}^{N}(Q_{obs,t} - Q_{cal,t})^2}{\sum_{t=1}^{N}(Q_{obs,t} - \overline{Q}_{obs})^2} \qquad 7.19$$

where $Q_{obs,t}$ are the observed inflows, $Q_{cal,t}$ are the calculated inflows, $\overline{Q}_{obs}$ is the mean of the observed inflows, $\overline{Q}_{cal}$ is the mean of the calculated inflows, and N is the size of the sample (i.e., 365). Index j refers to the year. These are both well-established criteria to evaluate the performance of inflow models (Bouchard, 1986). They both equal one when the calculated values perfectly match the observed values. Also, when the Nash criterion is smaller than zero, this implies that the mean of the observed inflows would be a better estimate of the observed inflows than the inflows calculated by the model. One last element to consider is the adequacy of the description of the parameter domain, and this can be undertaken through the analysis of the graphs showing the parameters versus the indicators.

7.2.3 Results and Discussion

The advantage of the flexibility of the parameters is at first measured by the gain in performance under the modeling variants that make use of this flexibility (H1, H3 and S1) compared with the modeling variant that keeps all parameters constant (SV). Tables 7.1 and 7.2 present a detailed view of the results obtained with variant SV. In Table 7.1, the average absolute difference (in %) between observed and calculated annual means and standard deviations for the inflows are given. For example, modeling variant SV applied on the Mistassibi watershed, using the calibration data set, the Gamma-shaped hydrograph and

the snowmelt model number 1 provides estimates of inflows that lead to annual means that differ by 7.20% on the average from the annual means obtained from the observed inflows. Similarly, for this case, the calculated inflows lead to annual standard deviations that differ by 22.54% on the average from the annual standard deviations obtained from the observed inflows. In Table 7.2, the *CCR* and *Nash* criteria values, as estimated using Equations 7.18 and 7.19, respectively, are provided.

Table 7.1. Means and standard deviations for variant SV.

Watershed	Data set	Hydrograph			
		Gamma		Triangle	
		Mean	Standard deviation	Mean	Standard deviation
(a) Snowmelt model 1 (Equations 7.10 and 7.11)					
Mistassibi	Calibration	7.20	22.54	9.02	9.74
	Validation	5.12	17.98	6.18	11.50
Chute du Diable	Calibration	11.76	13.33	15.12	12.62
	Validation	13.31	8.25	19.04	8.04
Ashuapmushuan	Calibration	11.50	14.50	13.48	14.10
	Validation	10.18	7.54	11.83	8.23
(b) Snowmelt model 2 (Equations 7.10 and 7.12 to 7.15)					
Mistassibi	Calibration	7.03	11.53	9.39	10.16
	Validation	3.88	11.18	5.79	9.39
Chute du Diable	Calibration	12.07	12.29	12.70	12.04
	Validation	11.87	11.96	12.75	11.82
Ashuapmushuan	Calibration	7.80	15.17	6.77	15.27
	Validation	8.43	17.56	7.29	10.62

Note: The values in the table are the average absolute differences (in %) between observed and calculated annual means and standard deviations for the inflows.

Table 7.2. Performance criteria for variant SV.

Watershed	Data set	Hydrograph			
		Gamma		Triangle	
		CCR	*Nash*	*CCR*	*Nash*
(a) Snowmelt model 1 (Equations 7.10 and 7.11)					
Mistassibi	Calibration	0.896	0.773	0.941	0.848
	Validation	0.881	0.743	0.923	0.787
Chute du Diable	Calibration	0.898	0.755	0.905	0.755
	Validation	0.911	0.780	0.921	0.776
Ashuapmushuan	Calibration	0.862	0.669	0.874	0.681
	Validation	0.882	0.725	0.903	0.756
(b) Snowmelt model 2 (Equations 7.10 and 7.12 to 7.15)					
Mistassibi	Calibration	0.914	0.807	0.916	0.802
	Validation	0.901	0.786	0.902	0.785
Chute du Diable	Calibration	0.896	0.748	0.902	0.762
	Validation	0.904	0.764	0.910	0.776
Ashuapmushuan	Calibration	0.890	0.714	0.903	0.767
	Validation	0.885	0.651	0.905	0.744

With respect to the annual means, it can be seen that modeling variant SV is typically different in terms of the observed annual means by 10 to 15%, with extremes that range from 3.88% to 19.04%. There is no indication in the results that the model systematically underestimates or overestimates inflows. An assumption can then be made that the water input (i.e., precipitation or snowmelt) data employed in the model do not systematically bias upward or downward the actual quantity of water received by the watershed and eventually translated into inflows. The timing of the precipitation can be a factor of course, but otherwise a greater attention can be paid to the model structure and parameters to try to explain the difference between calculated and observed means. With respect to the annual standard deviations, it can be said that the modeling variant is typically different from the observed annual standard deviations by 10 to 15%, with extremes that range from 8.04% to 22.54% different. The results indicate that there is a systematic bias produced by the model, that is, the annual standard deviations from calculated inflows are consistently lower than those from the observed inflows. In brief, the calculated inflows are flatter than the observed inflows. With respect the *CCR* and *Nash* criteria, Table 7.2 shows that the modeling variant SV performs similarly well on the Mistassibi and Chute du Diable watersheds. The modeling variant SV provides a poorer performance on the Ashuapmushuan watershed, in spite of the fact that the differences between observed and calculated annual means and standard deviations (Table 7.1) are neither markedly worse nor better than those for the Mistassibi and Chute du Diable watersheds. This is indicative of the presence of a lag between observed and calculated inflows on the Ahuapmushuan watershed, as has been confirmed by looking at the graphs of observed versus calculated inflows. Indeed, the lags on the Ashuapmushuan watershed seem more noticeable than those on the Mistassibi and Chute du Diable watersheds. The lag can be attributed to a bad timing between observed precipitations versus inflows, or a weakness in the model in representing the behavior of the watershed. The latter reason can possibly be partially overcome by the description of the domain of some parameters.

Another element to notice in Tables 7.1 and 7.2 is that the results obtained with the validation data sets are better than those obtained with the calibration data sets. Usually the contrary happens, and this is what justifies the use of validation data sets. Obviously, the calibration data sets present hydrologic situations that are more difficult to account for by

the model, and this is observed not only with the standard modeling variant (SV) but also with all other variants (H1, H3 and S1). As a consequence, the results that are presented in the following tables are determined after having merged the calibration and validation data sets for each watershed so as to obtain a more severe analysis of the modeling variants. It is also interesting to notice in Tables 7.1 and 7.2 that there is no clear winner in terms of performance with respect to the type of snowmelt models or hydrograph shapes used. A slight edge can be given to the snowmelt model number 2 over the snowmelt model number 1, as well as to the triangular-shaped hydrograph over the Gamma-shaped hydrograph. Nevertheless, because of the relative equivalence of these components (snowmelt models and hydrograph shapes), the results are merged for the sake of simplicity.

Hence, in Tables 7.3 and 7.4, the distinction is only made on the watersheds and the modeling variant. The results for modeling variant SV are presented again, and those for modeling variants H1, H3 and S1 are given for comparison with SV. Table 7.3 is similar to Table 7.1, and details the average absolute differences (in %) between observed and calculated annual means and standard deviations for the inflows. Table 7.4 is similar to Table 7.2, and lists the *CCR* and *Nash* criteria values as calculated by Equations 7.18 and 7.19. In Tables 7.3 and 7.4, the values in parenthesis represent the improvement (in %) produced by variants H1, H3 or S1 when compared with the variant of reference (SV), for each watershed. A negative value in the parenthesis means a degradation of the result compared with variant SV.

Table 7.3. Means and standard deviations for variants H1, H3 and S1.

Watershed	Mean				Standard deviation			
	SV	H1	H3	S1	SV	H1	H3	S1
Mistassibi	7.25	6.26	6.42	5.94	13.18	12.55	11.81	12.59
		(13.66)	(11.45)	(18.01)		(4.78)	(10.42)	(4.52)
Chute du Diable	13.33	11.67	11.70	9.21	11.77	12.94	11.89	12.66
		(12.45)	(12.24)	(30.86)		(-9.95)	(-0.96)	(-7.54)
Ashuapmushuan	9.74	8.87	8.86	5.67	13.58	13.29	13.22	13.47
		(8.94)	(9.05)	(41.82)		(2.11)	(2.64)	(0.84)

Note: the values in parenthesis represent the improvement (in %) produced by variants H1, H3 or S1 when compared with the variant of reference (SV). A negative value represents a degradation of the result.

Table 7.4. Performance criteria for variants H1, H3, and S1.

Watershed	*CCR*				*Nash*			
	SV	H1	H3	S1	SV	H1	H3	S1
Mistassibi	0.912	0.925	0.926	0.933	0.797	0.814	0.827	0.833
		(1.47)	(1.48)	(2.33)		(2.13)	(3.69)	(4.45)
Chute du Diable	0.904	0.910	0.909	0.915	0.761	0.778	0.781	0.792
		(0.69)	(0.55)	(1.29)		(2.20)	(2.70)	(4.05)
Ashuapmushuan	0.886	0.899	0.898	0.914	0.711	0.737	0.748	0.778
		(1.47)	(1.41)	(3.19)		(3.64)	(5.13)	(9.32)

Note: the values in parenthesis represent the improvement (in %) produced by variants H1, H3 or S1 when compared with the variant of reference (SV). A negative value represents a degradation of the result.

The results in Tables 7.3 and 7.4 demonstrate that variants in which the description of parameter domains is applied yield better performance than the standard variant where all parameters are constant. The improvement is particularly noticeable for the differences of the annual means, where the largest improvement compared with variant SV is observed, the most impressive case of all being variant S1 with an improvement on the order of 20 to 40%. In terms of the performance criteria (*CCR* and *Nash*), the improvements are smaller, because there is little room for increases in the values of those criteria. At the start, with variant SV, the performance criteria values are already fairly high. Mixed results are observed for the case of the differences between calculated and observed annual standard deviations, where improvement is present for the Mistassibi and Ashuapmushuan watersheds and degradation of the performance affects the Chute du Diable watershed. Because it affects only one watershed, a systematic error of the model structure can be discarded and therefore the most likely reason for this degradation is assumed to be the calibration procedure. Overall, from a performance standpoint, the exercise of proceeding to a description of parameter domains is sound. The flexibility of some of the parameters provide the structure of the model with some flexibility to more adequately replicate the behavior of the system under study, and as a result, to produce estimates that are in a better agreement with the observed outputs.

On the subject of parameter flexibility, one must evaluate whether the description of the parameter domain makes sense, and this can be undertaken through the analysis of the graphs of the parameter value versus the indicator value. Each variant leads to a large number of modeling scenarios, and therefore the analysis presented here focuses only on typical behavior. Such is the case of the parameter values presented in Figure 7.6, which applies for the Mistassibi watershed, variant H1 while making use of the triangular-shaped

hydrograph and the snowmelt model number 1. In the figure, t_{p_s}, t_{d_s}, t_{p_w}, t_{d_w}, a_s and a_w are respectively the surface hydrograph time to peak and duration, the interflow hydrograph time to peak and duration, and the surface and interflow contribution rates. Figure 7.6 presents the parameters in both their constant values (term *_cst* added) and their values if allowed to be flexible.

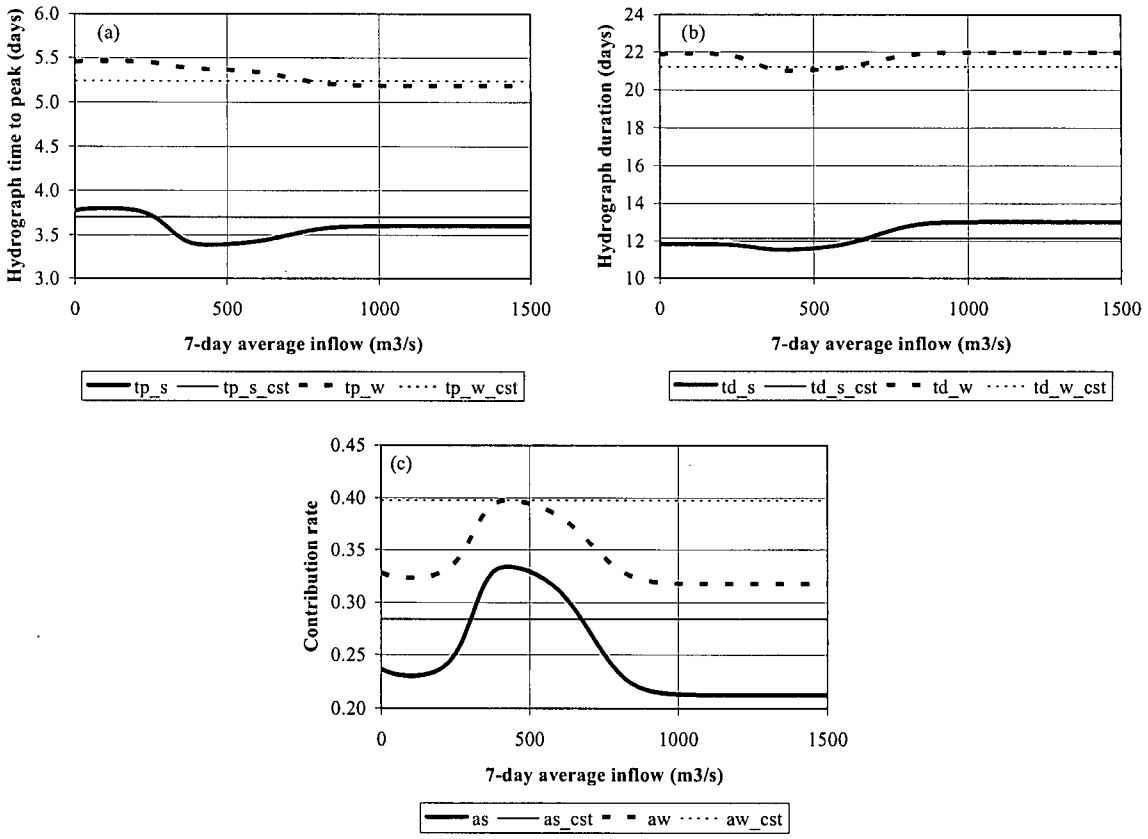


Figure 7.6. Flexible parameters for variant H1.

As the inflow indicator increases, soil humidity is presumed to increase as well, and the expected behavior from the parameter would be that the hydrograph times to peak become shorter, the hydrograph durations become longer, and the contribution rates, which are runoff coefficients, increase. The evolution of the parameter values with respect to the 7-day average inflow, as exhibited in Figure 7.6, roughly follow the expected behavior. There is no instance of scenarios that is in total agreement with the expected behavior. It must be noted that the hydrograph times to peak and durations are relatively insensitive

parameters, that is, they do not vary much with respect to the indicator. At the limit, their variability can almost be considered as background noise. The contribution rates are much more sensitive, and as a result, their disagreement with the expected behavior can be considered as more critical. The disagreement affects the region of the high average inflow, and three reasons may explain this issue.

The first reason is driven by an error in the structure of the model, that is, the contribution rate does not represent a runoff coefficient, but actually a runoff coefficient weighted by a factor related to flood routing within the streams in the watershed. The runoff coefficient is affected by the movement of water within the ground, and as the ground becomes more and more saturated with water due to large precipitations, here represented by the amount of inflow, more water is then available for transfer in the streams of the watershed. This is why the expected evolution of the runoff coefficient as precipitations become more abundant is that of a monotonically increasing curve. Now, as the inflow increases, more water could go through the floodplain, where the roughness is higher than on the streambed, and this would reduce the speed by which water is evacuated and create retention within the watershed, thus leading to a reduction of the evacuation rate of water at the watershed outlet. It is a basic flood routing process that spreads the evacuation of water over time, and this is what the contribution rates in Figure 7.6c might reflect. The solution would be to structure the model so as to more clearly separate the process of moving water within the ground from the process of moving water once it reaches the streams of the watershed.

The second reason would be the reduced number of data in the high inflow range. Obviously, high inflows are extreme cases and appear in smaller number in the historical records. In the case of the Mistassibi watershed, there is a very large number of inflow observations between 0 and 500 m^3/s, facilitating an adequate description of the parameter domain in that range. The number of observation is already much more reduced in the range from 500 to 1000 m^3/s, and only a few cases are available in the range from 1000 to 1500 m^3/s. On account of the rarity of data available for the calibration process, legitimate doubts can be raised as to the validity of the description of the parameter domains for high inflows if the available data do not represent all possible conditions of the watershed. The third reason would be due to the calibration process, for which the only objective is to

minimize the difference between observed and calculated inflows, and is not designed to specifically retain solutions that would be deemed physically adequate.

The first reason, related to the structure of the model, is the most likely rationale to explain the evolution of the parameter values with respect to inflows for variant H1. With variant H3, the calibration process can be a more important factor to explain the evolution of the parameter values with respect to the indicators. With variant H3, the description of the parameter domain is rarely in agreement with the behavior expected from the parameters with respect to the indicators, and there does not seem to be any consistent trend in the description of the parameter domain. A greater number of parameters (the $m(p_{j,i})$s) must be calibrated with variant H3 (48) compared with variant H1 (18), and this even if the description of the parameter domain is coarser with variant H3 (2 fuzzy sets per indicator) than with variant H1 (3 fuzzy sets for the indicator). This means that a much larger set of parameter values must be searched by the calibration procedure for variant H3 compared with variant H1. Obviously, there are many possible sets of parameter values that can provide satisfying answers, for variant H3 constitutes an improvement in performance from variant SV. However, the likelihood of finding a set of parameter values that performs well and leads to a physically meaningful description of the parameter domain most probably declines as the number of parameters to calibrate increases.

Variant S1 leads to similar conclusions as those of variant H1. Variant S1 was added following a preliminary analysis of variant H1, for which it was noticed that the hydrograph time to peak and duration are not particularly sensitive parameters. The contribution rates are sensitive, and it is logical to allow all contribution rates within the model to be flexible so as to remove the possible constraints on action-reaction chains within these parameters, thus the creation of variant S1. Figure 7.7 presents a typical case for variant S1 of the evolution of the parameter values with respect to the indicator, the 7-day average inflow. This is taken from the modeling scenario applied to the Mistassibi watershed, and includes the use of the triangle-shape hydrograph and snowmelt model number 1. In Figure 7.7, a_s, a_w, a_g, a_{ls}, a_{lw}, a_{lg}, and a_p are respectively the surface, interflow and groundwater contribution rates to the inflows, the surface, interflow and groundwater contribution rate to losses of water, and the percolation rate between the subsurface and

groundwater reservoir. The figure presents the parameters in both their constant values (term *_cst* added) and their values if allowed to be flexible.

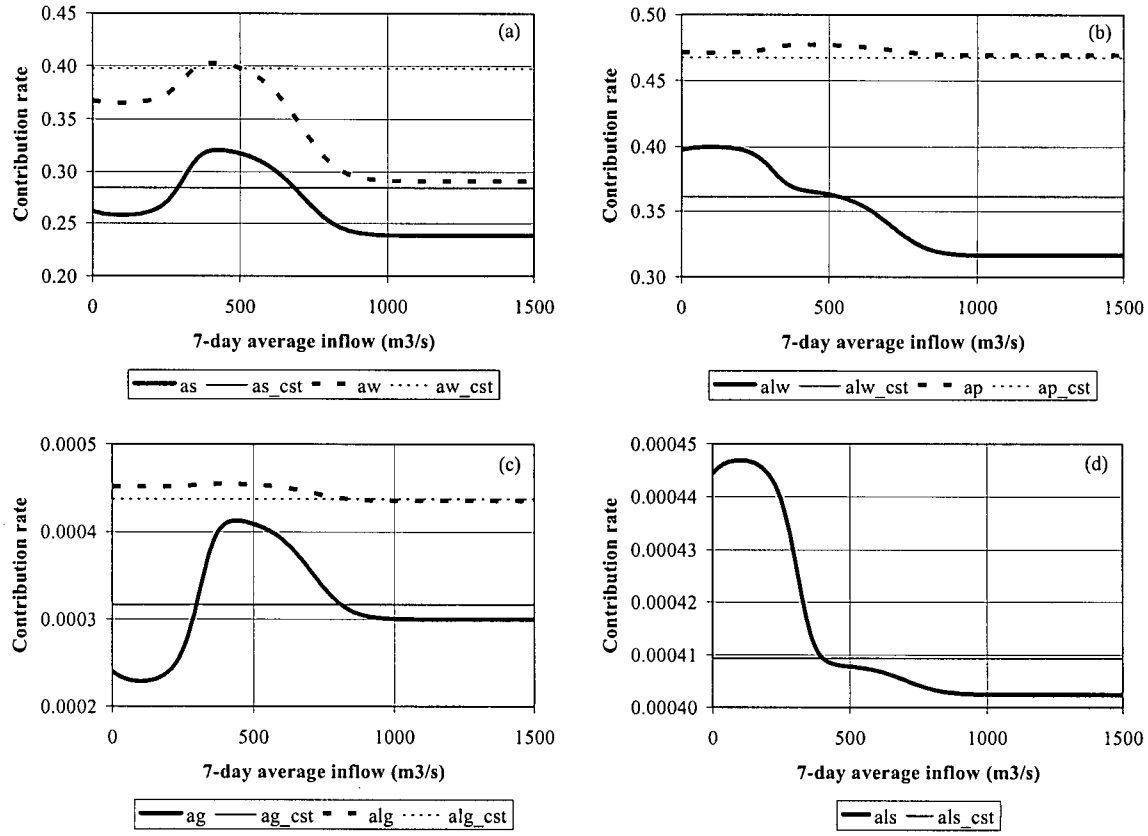


Figure 7.7. Flexible parameters for variant S1.

The conclusions drawn about surface and interflow contribution rates to the inflows in variant H1 can apply to the surface, interflow and groundwater contribution rates to the inflows in variant S1. The behavior of these parameters in the high inflow range is not in agreement with what would naturally be expected, and the disagreement can most likely be explained by the fact that these contribution rates are actually runoff coefficients weighted by flood routing factors. The percolation rate, as a rule, seems to follow the same behavior as that of the contribution rates to inflows, but can also be considered as rather insensitive to inflows. Indeed, there is generally little variation with respect to inflows for this parameter for all scenarios in variant S1. The case of the behavior of the contribution rates

to losses is more difficult to explain. Soil moisture, as represented by the 7-day average inflow, might not be a factor that is as important for these contribution rates to losses as it is for the contribution rates to inflows. Other meteorological factors such as air humidity and temperature can possibly play a definitive role in the description of the domain of these rates to losses. If it can be assumed that soil moisture and air moisture are positively correlated, this could mean that if the 7-day average inflow is high, then soil moisture is high and air moisture is possibly high as well. If air has high water content or is even water-saturated, then the transfer of water from the soil to the air through direct evaporation or transpiration from the vegetal cover, when water is collected in subsurface and groundwater by the root network, can be reduced and even stopped. This may explain why the rates of losses decrease as the inflows increases

In conclusion, the process of describing parameter domains provides a greater structural flexibility to the inflow model, which can then more closely replicate the observed inflows. Modeling performance is improved, and the uncertainties as to the adequacy of the values of the parameter are reduced. This application involves natural processes that are relatively well known, and therefore it has been possible to evaluate whether this process of describing parameter domains could lead to a description that would physically make sense. The results presented here demonstrate that the process could provide acceptable descriptions. One must not forget to carefully analyze whether the results from this process are influenced by the structure of the model employed, by the distribution of data points used as indicators (e.g., few data in a particular range of values), or by the calibration burden (i.e., many parameters to calibrate).

7.3 Application to Algae Modeling

Mechanistic models constitute a long-standing option for modeling algae concentrations, with numerous developments in the 1970s and 1980s that include landmark advances such as those of Di Toro et al. (1971) and Whitehead and Hornberger (1984). Chapter 6 in US EPA (1985) presents an exhaustive literature review of mechanistic algae modeling developments up to the middle of the 1980s, and subsequent progress has been accomplished, as attested by the work of Lung and Larson (1995), Cloot and Pieterse (1999), and Rutherford et al. (2000), for example. In these models, the evolution of algae

concentration is explained through budget equations based on rates, such as algae growth and mortality rates, the settling rate and transport rates from one river section to another, and the loss rate due to predators. Structure and parameter uncertainties are very prevalent in such models. All possible situations of algae growth and decay may not be properly presented in the models. In addition, the data available may not be entirely adequate (i.e., relevant) to represent all natural processes involved, thereby possibly further limiting the application of the models. It is with the uncertainties of the mechanistic approach in mind that researchers and practitioners have recently attempted to implement an alternative modeling approach based on AITs, that is, neural networks (Yabunaka et al., 1997; Maier et al., 1998 and 2000) and fuzzy logic (Setnes et al., 1997 and 1998). Employed as they are by these authors, AITs are only replacements to mechanistic models. In spite of their flexibility, they constitute only black boxes that do not help understand the natural processes involved in algae growth and decay, and this is not in agreement with the objective of this thesis.

Yet, the structure of the usual mechanistic models for the estimation of algae concentration allows very easy implementation of an inference engine based on fuzzy logic for the description of parameter domains, to the point where such descriptions are actually the natural extensions of these models. Similar to that of Section 7.2, this section starts with the description of the mechanistic model, and follows with details of the integration of the fuzzy logic inference engine within the structure of the model. The database employed to test the model is then given, followed by the protocol of the experiment. The results and the ensuing conclusions are at last presented in detail.

7.3.1 Description of the Model

In this investigation of the combined mechanistic/fuzzy logic approach, the simple mechanistic model extensively described in US EPA (1985) is selected. When the algae mass is expressed in terms of cell number per volume in this model, the budget equation is a differential equation relating the rate over time of the algae cell concentration (A) to the growth (G), mortality (M), settling (Se) and predatory (Pr) rates. Equation 7.20 represents this budget equation, while Equation 7.21 details the growth rate (G).

$$\frac{dA}{dt} = (G - M - Se - Pr)A \qquad 7.20$$

$$G = G_{opt} f(T) f(L,N,C,P,S) \qquad 7.21$$

In Equation 7.21, G_{opt} is the optimal growth rate, and $f(.)$ are weighting factors for temperature (T), light intensity (L) and common nutrients, that is, nitrogen (N), carbon (C), phosphorus (P) and silica (S). The multiplicative and the minimum formulations, Equation 7.22 and 7.23, respectively, are the most commonly used ones for $f(L,N,C,P,S)$:

$$f(L,N,C,P,S) = f(L)f(N)f(C)f(P)f(S) \qquad 7.22$$

$$f(L,N,C,P,S) = \min\{f(L), f(N), f(C), f(P), f(S)\} \qquad 7.23$$

Within the natural boundaries of the relevant indicators (i.e., temperature, light and nutrients considered separately), all factors $f(.)$ can vary between zero and one, leaving G_{opt} in Equation 7.21 as the sole parameter describing the volume of growth. Under ideal conditions, that is, when all factors equal one, then the growth rate is optimal.

The other rates in Equation 7.20, i.e., the mortality, settling and predatory rates, are not described in the literature in as much detail as the growth rate. A maximum mortality rate (M_{max}) weighted by a temperature factor is the most common formulation for the mortality rate. Often, this rate is also combined with the predatory and settling rates (Pr and Se) in order to form a global mortality rate. The settling rate is a function of the hydraulic conditions in the water body and the physical properties of the algae, and a typical formulation is the ratio of a settling velocity over water depth, yielding the equivalent of the Froude number. As for the predatory rate, when it is not considered constant, a common formulation is simply a maximum predatory rate weighted by a temperature factor. Of course, predators are living species as much as algae are, and as such, their concentrations in the water body over time could be described by a budget equation similar to Equation 7.20, given an adequate database. Speaking of database, application of algae concentration models often can be constrained by the availability of relevant data. The application presented here is no exception, and this is another reason why the concept of describing parameter domains is interesting, since they may compensate for inadequate data.

As mentioned above, all factors $f(.)$ can vary from zero to one within the natural boundaries of the data. Figure 7.8 provides examples of such factors for temperature, light and a typical nutrient, respectively. These functions are essentially equivalent to fuzzy sets.

As a reminder, the purpose of a fuzzy set is to define a variable or an indicator (e.g., temperature, light intensity, or nutrient concentration) in its entire domain according to a degree of membership expressed by a value varying from zero (low membership) to one (high membership).

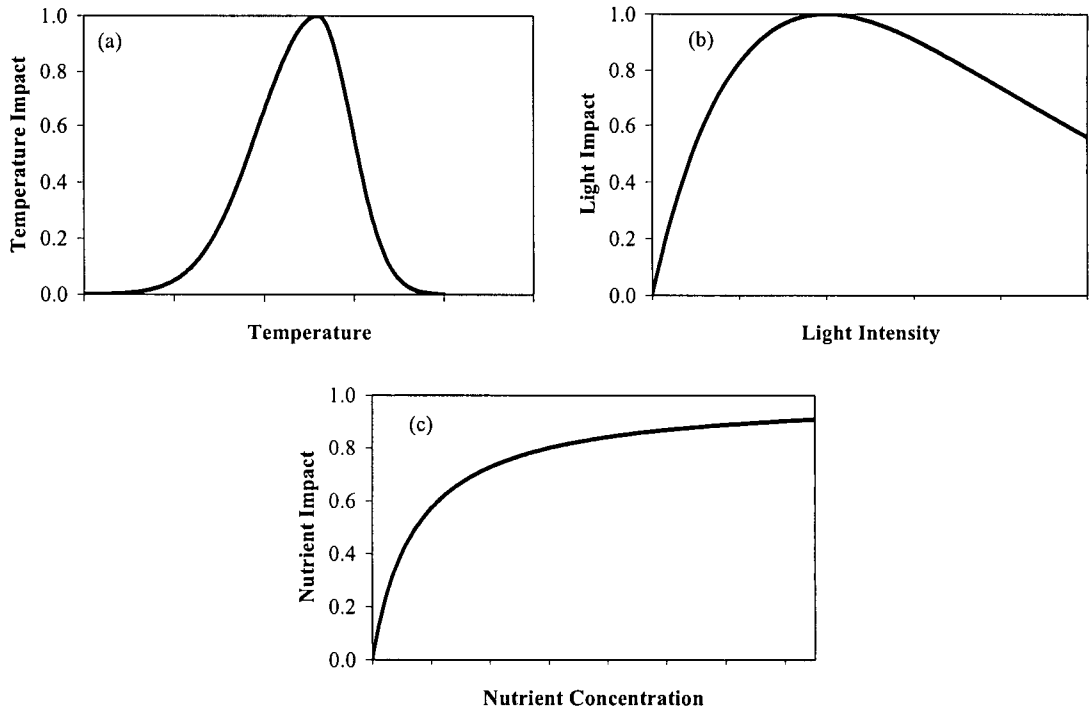


Figure 7.8. Typical weighting factors for (a) temperature, (b) light, and (c) nutrients.

This membership is related to some given attribute, say, in the case of Figure 7.8, the impact of temperature on algae growth, the impact of the amount of light energy on photosynthesis, and the impact of the nutrient concentration on algae growth. Furthermore, Equations 7.22 and 7.23 represent rule inferences commonly used in fuzzy logic, namely the product inference and min-max inference (Bardossy and Duckstein, 1995). For purely mechanistic models the equation for the growth rate (Equations 7.21 and 7.22 or 7.23) can be considered as a fuzzy logic formulation in which there is one and only one rule about the impacts of the variables on this rate. In this application, the concept of rule inference based on Equations 7.21 and 7.22 is extended in order to accommodate more than one rule about the impacts of the indicators on the growth rate. This requires a change in the definition of

the indicators (temperature, light intensity, and nutrients), where their respective domain is described by more than one fuzzy set. Rules can then be constructed and combined. Following the defuzzification process, the growth rate can be expressed as:

$$G = G_{opt} \frac{\sum_{i=1}^{I} v_i m(p_i)}{\sum_{i=1}^{I} v_i} \qquad 7.24$$

where $m(p_i)$ is the factor $f(.)$ that applies to rule i, and v_i is the degree of fulfillment of rule i.

To consider more than one rule adds more flexibility to the model. Indeed, each rule is meant to describe as accurately as possible only a specific subset of situations of algae growth (the combination of all rules covering the entire set of situations), while only the equivalent of one rule covers the entire set of situations in the purely mechanistic formulation. It implies that, as the data or indicator domain is more finely subdivided by a greater number of fuzzy sets, the higher the number of rules is, and therefore the higher the accuracy of the model may be. However, for any given application, the size of the database employed limits the number of rules that can be implemented. There cannot be more parameters (G_is and others) in the mechanistic/fuzzy logic model than there are data points. The use of fuzzy logic can also help compensate for the lack of adequate data. For example, in mechanistic models, solar radiance values are essential in the determination of light intensity effects on algae cells. In many cases, solar radiance is unavailable, but a combination of temperature, daily sunshine (in unit of time) and turbidity, all available variables in the case study presented here, can be used instead to explain light intensity. With fuzzy logic, no mechanism need be described since only intuitive inferences are involved. It is reasonable to infer that if temperature and daily sunshine are high while turbidity is low, then the impact of light intensity is likely to be high. Inversely, if temperature and daily sunshine are low while turbidity is high, then the impact of light intensity is likely to be low. The indicators available for the application may also be employed to help describe the settling rate domain. The rules for the impact of settling may involve the use of the inflow and the water elevation of the water body, which are adequate indicators to infer the settling velocity and water depth normally employed in the calculation of the settling rate. Combination of the rules and deffuzification in the case of the settling rate can lead to a formulation equivalent to Equation 7.24.

Several modeling formulations can be constructed, and five types are explored in this application, with data from a site on the River Murray in Australia. The first type involves a purely mechanistic formulation, that is, Equation 7.20 in its integrated form used with Equation 7.22. In Equation 7.20, the settling, mortality and predatory rates (*Se*, *M*, and *Pr*) are combined to form only one global mortality rate. Such a combination is often accomplished in practice, as is the case in the application of this work, due to the lack of data to allow calculation of separate settling and predatory rates (US EPA, 1985). Two options are examined for the global mortality rate, one which considers the global mortality rate constant over time and the other where the global mortality rate is weighted by a temperature factor $f(T)$. The temperature factor $f(T)$, for both the growth rate and the global mortality rate when necessary, is assumed to follow a skewed Normal distribution, as described by Lehman et al. (1975). As for the light factor $f(L)$ for the growth rate, the formulation proposed by Walker (1975), which assumes photoinhibition, is employed. In the River Murray case, solar radiance is not available, and is therefore replaced in Walker's formulation by the daily number of hours of sunshine. Finally, the classic saturation-type relationship, considering a half saturation constant, is used for nutrient factors ($f(P)$, $f(N)$) for the growth rate. Several model scenarios are tested, each one with their particular combination of variables (temperature plus light only, temperature plus phosphorus only, temperature plus light plus phosphorus, etc.).

The last four types of model formulation involve the description of the parameter domain. For these types, the domain of each indicator employed is divided into Gaussian-shaped fuzzy sets, then Equation 7.20 in its integrated form is used with Equation 7.24. Again several scenarios are possible, depending on the combination of variables employed, and whether the global mortality rate is constant over time or weighted by a temperature factor. The second model formulation type includes all scenarios where the growth rate is based only on indicators related to energy input available for algae (temperature, hour of sunshine and turbidity). For the third type, the growth rate is based on indicators related to energy input and nutrients (nitrogen and phosphorus). For the fourth type, indicators related to energy input and nutrients are used again, but this time a distinct settling rate is considered, calculated using Equation 7.24, with rules based on inflow and water level data available for this application. The fifth type includes scenarios where the inflow is

considered with energy input and nutrients for the calculation of the growth rate. These latter scenarios are particular cases, which attempt to describe a stratification process that has been observed at a location downstream of the site investigated. This stratification process has been recently analyzed by Baker et al. (2000), and it has been assumed that this process could affect algae growth. Stratification may be present on the site of our case study, and inflow values combined with temperature values for the calculation of the growth can be indicative of the presence of stratification in the water body.

In total, sixty model scenarios are tested. The calibration of the parameters of the models is accomplished with genetic algorithms, structured so as to minimize the difference between observed and estimated values of algae concentrations. In the section discussing results (Section 7.3.3), all these scenarios are assessed globally.

7.3.2 Description of the Application Case

Located in southeastern Australia, the River Murray constitutes the major surface water resource of the state of South Australia. Figure 7.9 illustrates this river system and gives a more detailed view of the region of interest for this application. Water is pumped from this river to several major cities of this state. As a result of many activities in the watershed, this river has been subject to many outbreaks of toxic cyanobacterial (blue-green algae) blooms. At Morgan, more particularly, these outbreaks are a significant water supply operational problem, as a water treatment plant is located in this town and several cities get their water from this location. Adequate prediction of algae blooms in this location is therefore important for the optimal operation of the water treatment plant (Maier et al., 2000).

The data used for this application are from Morgan or in its vicinity, and include weekly values of blue-green algae concentration (*Anabaena* spp., in cells/ml), water temperature (°C), turbidity (NTU), hours of sunshine for the day (hr), total kjedahl nitrogen concentration (mg/l), total phosphorus concentration (mg/l), water level (m) of the river at Morgan and river inflow (Ml/day) at the border between South Australia and Victoria (about a hundred kilometers upstream of Morgan). The period of the data used extends from June 1984 to November 1996, and because it is a rather short segment, the whole set has been used for the purpose of calibration. For calibration, other considerations must be

taken into account. Figure 7.10 shows the record of algae concentration and indicates how sensitive the models have to be in order to adequately replicate the evolution over time of the algae mass. Indeed, the concentration can stay unchanged for a significant period of time, then it increases to substantial values very rapidly and finally decreases to nearly zero just as rapidly. To better accommodate the models faced with such variability, daily input sequences from the linear interpolation of the weekly data set are provided to the models. Also the initial conditions, that is, the algae concentrations calculated by the models, are regularly updated with observed concentrations. It is a procedure that is frequently used in practice with iterative models such as those applied here, and helps reduce error propagation.

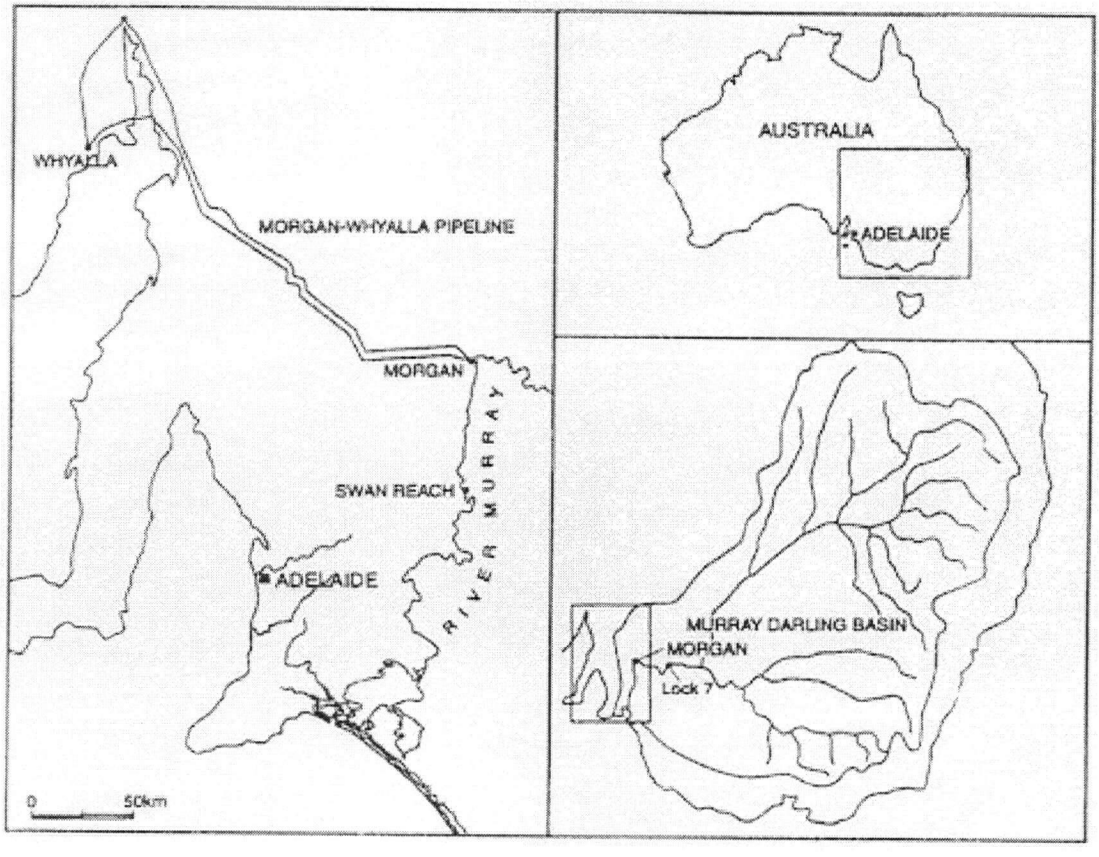


Figure 7.9. The River Murray in South Australia and Morgan.

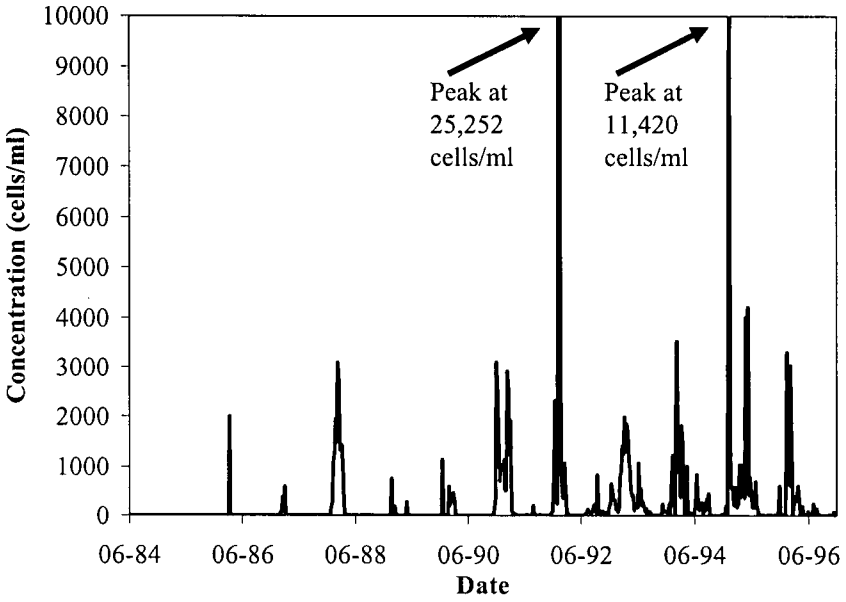


Figure 7.10. Blue-green algae (*Anabaena* spp.) cell concentrations at Morgan.

7.3.3 Results and Discussion

Table 7.5 shows the mean and standard deviation for the observed values of algae concentration and the mean, standard deviation and root mean squared errors (*Rmse*) obtained on the average for each type of model formulation. The *Rmse* is a common criterion of algae model performance. It is the root of the average squared differences between observed and calculated values. It is clear from Table 7.5 that the models, whatever their type, perform somewhat poorly. In the best case (type 2: mechanistic/fuzzy logic based on energy inputs only) the estimates of algae concentration are almost 50% lower on the average (see the means) and can account for only about 50% of the standard deviation. For this particular case, these results are not surprising. Other model developments have been undertaken for this river (Maier et al., 1998 and 2000) and a similar performance has been obtained, although their models have been subject to different calibration and modeling conditions. It can be said nevertheless that the means and standard deviations of the estimated algae concentrations are of the same order of magnitude of the mean and standard deviation of the observed values. Also the models do follow more or less the evolution of algae concentration over time, that is, a peak is often estimated by the model when there is one observed and a period of consistently low concentration is often estimated when such a period is present in the observed values. On the subject of the

estimation of peaks by the models, it must be noted that in almost all instances there is a time delay between the observed and estimated peaks. Figure 7.11 illustrates a typical example of delays between peaks, and it is these time lags that in part explain the high *Rmse* values presented in Table 7.5. The adequate prediction of algae concentration peaks for this river is a difficulty that has already been observed with other model developments (Maier et al., 1998 and 2000).

Table 7.5. Indication of performance for each type of models.

Type	Form of model	Constant global mortality rate			Weighted global mortality rate		
		Mean (Cells/ml)	St.dev. (Cells/ml)	*Rmse* (Cells/ml)	Mean (Cells/ml)	St. dev. (Cells/ml)	*Rmse* (Cells/ml)
-----	Observed data	285	1240	-----	285	1240	-----
1	Pure mechanistic	80	379	1,160	115	499	1,170
2	Energy inputs only	146	606	1,158	139	580	1,146
3	Energy plus nutrient	129	556	1,155	130	490	1,160
4	Distinct settling rate	142	594	1,151	113	413	1,211
5	Stratification cases	123	529	1,151	128	510	1,181

*Note: Models of type 2 to 5 all involve the description of parameter domains (mechanistic/ fuzzy logic models).

Another explanation for the poor performance of the models is the variability of the observed algae concentrations. Algae in significant quantities are present only in periods of bloom, and the concentration is considered equal to zero for more than half of the time. One bloom event in 1992, which peaks at 25,252 cells/ml, as indicated in Figure 7.10, is particularly significant. No models have been able to estimate this peak well, and therefore this biases the evaluation of the performance. Indeed, if this event is not considered in the calculation of the means, standard deviations and *Rmse*s for the models, the results in Table 7.5 change significantly. The length of this event represents about 2.5% of the length of the whole data set, but when it is ignored, the *Rmse* for all models change from an average of 1,164 cell/ml to 664 cell/ml, which means a diminution of 43%. The standard deviations of the estimated values also get closer to the observed standard deviation by 9% when this event of 1992 is not included. There is only a slight improvement (1%) with respect to the means of the estimated values of concentration.

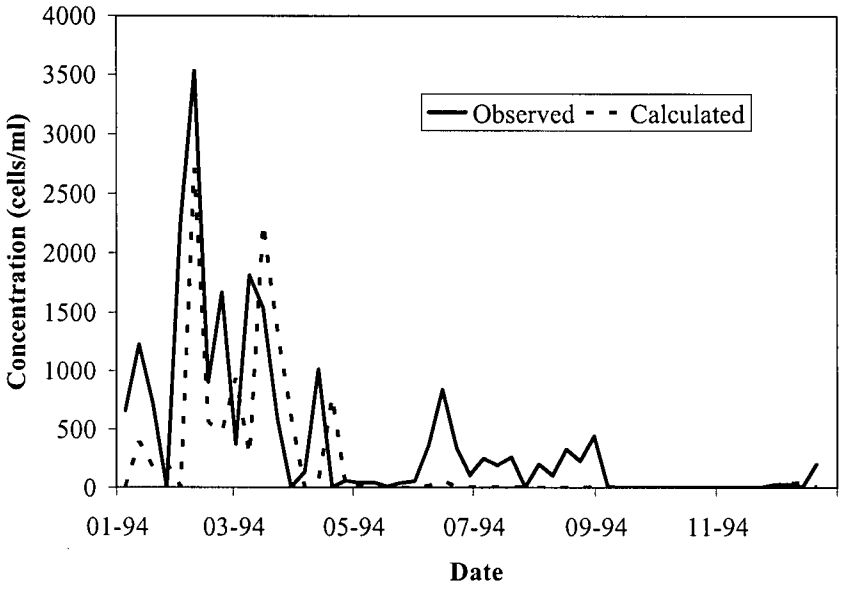


Figure 7.11. Lags between observed and calculated algae concentrations.

Note from Table 7.5 that the models involving the description of parameter domains (mechanistic/fuzzy logic models) provide a superior performance when compared with the purely mechanistic models. The means and standard deviations obtained using the mechanistic/fuzzy logic models are for most cases closer to the mean and standard deviation of the observed values than those from the purely mechanistic models. The *Rmse*s are quite similar for all of the models, but this is due to the fact that all models are afflicted with the problem of estimating the peaks after the observed peaks. Therefore the flexibility gained with the mechanistic/fuzzy logic models is an advantage compared with purely mechanistic models. The mechanistic/fuzzy logic models are an attempt to reduce the structural limitations of the purely mechanistic model while keeping the physical meaning of this mechanistic model. Of course, a claim can be made that this mechanistic part may be flawed and may constitute a limit to good performance. However there also seems to be a limit to the gain in flexibility, as can be concluded from the comparison of the different types of mechanistic/fuzzy logic models tested. Indeed, between type 2 and type 5 models, there is an increase in the number of variables employed, which means that flexibility increases in going from type 2 to type 5 models. However, this addition of flexibility has not translated into improvement of the performance. In fact, the models of types 3 to 5 have produced a poorer performance than the models of type 2. An increase in flexibility implies

the use of a greater number of fuzzy rules and a greater number of parameters must be calibrated as a result. As the number of parameters that need to be to calibrated increases, it becomes more difficult for the optimization procedure to find the optimal values for the parameters. In this application, the same number of iterations has been imposed on the optimization procedure in order to find an optimal set of parameters, regardless of the number of parameters that need to be calibrated. A fairer treatment would have been to impose a number of iterations proportional to the number of parameters in order to allow a more exhaustive search when the number of parameters is large. This limitation imposed on the optimization procedure is a technical point that can explain why model performance has reached a limit as flexibility increase.

Data limitations also influence the quality of estimates, and it is assumed that the data used in this application suffer from some flaws. First, doubts can be raised as to the accuracy of the data sequences, and in particular as to the accuracy of the algae concentration measurements. The work of Jones et al. (2000) describes the protocol for monitoring cyanobacteria used in the measurements of algae concentrations employed in this application, and it also indicates that largely erroneous measurements can be present in the data. It is therefore suspected that many outliers are present throughout the sequence of algae concentrations, the most suspicious of all potential outliers being the high peak of 25,252 cell/ml in 1992. If this particular peak event were withdrawn from the data sequences for the calibration of the models, the variability of the algae concentration would be reduced and the calibration process might lead to models that respond better to the other observed algae blooms. The second possible flaw is that the sequences of data used as inputs in the models (i.e., temperature, hours of sunshine, phosphorus, etc.) may not be representative (i.e., relevant) of the algae concentration sequence. In an analysis prior to the development of these models, it was observed that the statistical properties of the inputs when there are algae blooms are not really different from those of the inputs when there are no algae present at all. The means, standard deviations and probability distributions of the inputs with algae present and those with no algae present are only slightly different. This implies that the same vector of inputs for a given day can be indicative of the presence of algae growth as well as indicative of a total absence of algae. This complicates the

calibration process for a model since no clear difference between algae or no algae situations can be made.

There are two causes that can potentially explain why the inputs are not representative (relevant). The first cause is that the original data set available for this application is discretized on a weekly time step, and this is very likely too large of a time step to characterize algae growth. Indeed, algae concentrations on this river can vary rapidly, and so can inputs such as phosphorus, turbidity or hours of sunshine, for example. Daily measurements of the inputs and algae concentrations would provide a more suitable time step so as to capture the variability over time of all of these indicators. Unfortunately, only a few indicators are consistently measured on a daily basis, while the other indicators are measured on a weekly basis, or a longer time step. The second cause is that the data set is restricted only to a particular section of the river, and it is too small a domain of study. Obviously, algae growth occurs at other locations upstream on the river, and this production of algae can then be in part transported to our point of study. Therefore, the algae concentrations measured at Morgan, the site of study, includes algae produced on site and very likely algae produced upstream. Other indicators measured on site, such as temperature or phosphorus, cannot necessarily be indicative of the condition of algae growth upstream. To solve this problem, it would therefore be required to model algae growth at several locations in the river.

Due to the overall adverse modeling conditions (i.e., data relevance and completeness, plus model structure), a large variability is observed in the description of the parameter domain with respect to the various results. Consequently, very much like variant H3 in the inflow modeling application, it is not possible to provide a graph of the parameter value versus indicator value that would be typical, representative, or even satisfying. At times, the description of the parameter domain would make sense physically, that is, it would adequately represent the photoinhibition phenomenon or describe the algae growth potential adequately (i.e., high growth at mid range temperature and low growth at low and high extreme temperature). At other times, the disagreement between the description and the expected physical reality is noticeable. In spite of the results, the concept of describing parameter domains should be considered as sound and valid. From a performance

standpoint, this concept leads to better modeling estimates compared with the modeling situation where the parameters are kept constant.

7.4 Conclusions

With respect to the two applications presented in this chapter, the following conclusions can be drawn:

1. Improvement in the performance is observed for cases where the flexible parameters are allowed, compared with cases where the parameters are kept constant, due to the added modeling flexibility the description of parameter domains provides to the model.
2. With favorable applications, the description of parameter domains can be physically sound, that is, it can be in agreement with processes observed in nature.
3. Beware of the model structure, for it dictates the nature of the parameters, and a biased model structure can yield inappropriate parameters, no matter if the description of their domains is performed or not.
4. Data adequacy and completeness must always be kept in mind, for they can influence the description of the parameter domains.
5. The calibration procedure must be adapted so as to account for the added burden the description of parameter domains imposes.

Watershed inflow production is reasonably well known physically, which allowed the evaluation of the behavior of the concept investigated in this chapter. In fact, the benefit of this concept is also apparent when applied to the estimation of algae concentration, even though the phenomena of algae growth and decay might not be as well known physically as is inflow production. The challenge would be to work on cases where the natural processes are not well known. This is where the concept presented in this chapter can provide the greatest benefit, by possibly allowing a better understanding of the natural processes involved, as long as the remarks above (i.e., points 1 to 5) are kept in mind.

Chapter 8
Discussion

On the matter of data quality, this thesis has mainly focused, from Chapters 4 to 6, on a specific attribute, accuracy, and on specific problems of accuracy: shifts, trends and outliers. Data adequacy and completeness are data quality attributes that are also highlighted in Chapter 7. In any case, this thesis is not driven solely on resolving particular issues of data quality. To increase awareness related to the major issue of data quality is certainly more important, and this work must be considered as one way of achieving this endeavor. This work was initially motivated by the observation that the issue of data quality is often overlooked in practice, in water resources. On one hand, data are precious commodities that are not necessarily easy to obtain from water resources systems. Routinely, data exist that are not entirely suitable for the intended purpose of the users, yet they may represent the only available information, and consequently pressure to employ them may be high even when their quality is in doubt. On the other hand, the task of building methods that can assess data quality attributes is not easy. The most difficult problem is not really in developing the tools, but rather in finding reference data on which to test these tools so as to determine their performance capabilities. The reference data can be real observations for which the true properties can never be known exactly. They might be affected by some undetected bias that affects data quality attributes. Also, the reference data can be synthetic data, which have been used throughout this work, and for which the properties are exactly known. The performance of the tools for evaluating data quality attributes can be determined based on the knowledge of the properties of the synthetic data. Unfortunately, the synthetic data might not be entirely representative of the real data they are meant to replicate, and this can result in a bias in the determination of the performance of the tools for evaluating data quality attributes. In spite of these difficulties, work dedicated to the evaluation of data quality attributes must continue, and this applies to water resources data as well as to any environmental or natural resource data. It is a question of prevention and cost savings. Data are the foundations of decision-making processes that may have large economic impacts, and are employed in mathematical tools

(e.g. models) that can be very expensive to develop and run. From a practical standpoint, it is important that reliable tools assessing data quality attributes be available. It may not be necessary to exactly detect errors, such as is attempted here for shifts, trends and outliers. At the limit, it may be enough to acknowledge the presence of flaws in data, and to evaluate their impacts so that proper evaluation can be made of the validity of employing specific mathematical tools or of the exact value of a final result of a decision-making process performed based on the data.

Detection tests for shifts, trends and outliers based on AITs are developed as an alternative to conventional tests often employed in practice. AITs are definitely suitable technically for the construction of such tests. However, one must look beyond the technical aspects of the AITs, and focus on the basic foundations of these techniques as they relate to data quality control. Data quality might impose the achievement of specific quantifiable objectives, yet quality is not a fixed and definitive concept. The term quality may have a different meaning from one culture to another or from one scientific domain to another. Even in a specific culture or scientific domain, the necessity for quality control may differ from one application to another. It is with the varying perception of data quality from one context to another in mind that AITs have been chosen as tools in this work. Here, AITs fulfill a very specific task, that is, the detection of anomalies. However, their foundations, which rely on the description of data domains and on inference for determining system behavior, as well as their structural flexibility, are deemed very suitable for addressing concepts as vague as data quality.

The AIT-based detection tests proposed in this work do provide acceptable performance in diagnosing anomalies that are equivalent to those of conventional tests. However, the AIT-based tests require much more effort to implement than the conventional tests employed in this work, and this constitutes a disadvantage from a practical standpoint. The AIT-based tests should not be rejected as a result of this conclusion. The applications presented in this work demonstrate that each of the detection tests has its own behavior, properly diagnosing the presence of an anomaly while, the others do not, and vice-versa. These distinct behaviors are particularly prevalent with detection tests for outliers, and they are also observed for detection tests for shifts and trends, although to a lesser extent. AIT-based tests are valuable because they behave differently from the conventional tests.

Attempts have been made to take advantage of these distinct behaviors in Chapter 6 for the application to shifts and trends, and it is recommended that further efforts be made to develop procedures that combine the results from several tests to improve detection performance. In fact, these procedures might be based on AITs, where the reliability of the tests may be inferred based on characteristics of the data under investigation and the known behavior of the tests.

It must be noted as well that the AIT-based tests have provided some positive results. First, they lead to improvement in detection performance for multivariate cases of shifts and trends, where several data sequences are tested simultaneously, compared with univariate cases, where data sequences are tested individually. The results highlight the advantage of having several sources of information, where, for example, one strongly shifted sequence can help in the detection of the sequences of its group for which the shift is less pronounced. As long as one is certain that sequences can be grouped together, then using tests that can be applied to multivariate cases is a more reliable option than the tests for univariate cases. Second, AIT-based tests appear to provide better estimates of the characteristics of anomalies (i.e., the location of shifts and the amplitude of shifts, trends and outliers, based on Amp/CV and Amp/SD ratios) than those derived from the conventional tests. If one does not remove anomalies in data sequences for fear of false detection, then one can use these estimates of location and amplitude to quantify the impacts of these possible anomalies when the data are employed in models or in decision-making processes. It is expected that the accuracy of these estimates would improve if the resolution of the AIT-based tests is finer, that is, if the Kohonen maps have more output neurons and the fuzzy c-means cluster sets have more clusters. On the whole, it would be interesting to analyze the effects of increasing or decreasing the number of output neurons or clusters in the detection performance of the AIT-based tests. Possibly, false detection might reduce as the number of output neurons or clusters increases. Third, compared with other statistical techniques such as clustering, AIT-based techniques like the Kohonen network and fuzzy c-means would require less computer processing time and memory for the calibration procedure.

Many avenues for further research can be derived from the work accomplished in this thesis. First AITs should be compared with other detection techniques to have a more

general demonstration of the utility of all techniques, AIT-based and conventional. For example, the domain of industrial processes, biology, medicine, and computer and electrical engineering profess a large confidence toward in CUSUM and EWMA tests for the detection of shifts. Such techniques should be investigated further, along with AIT-based tests, for comparison sake and to see if the joint study of all these techniques may help in the development of more robust methods of detection. Second, only cases of single outliers, shifts and trends have been investigated in this work, and these do not represent very realistic instances of what occurs in nature. Cases involving multiple outliers, shifts and trends should be studied with AIT-based tests to strengthen the validity of these techniques. For example, CUSUM and EWMA tests can assess cases of multiple anomalies, and AIT-based tests should be developed as well to address these cases as well. It is actually possible to develop such AIT-based tests, for it is only a question of calibrating them to be able to consider patterns related to cases of multiple anomalies. Such developments have not been accomplished in this work, mainly because the restriction in computer capabilities has limited the calibration of Kohonen maps and fuzzy c-means cluster sets to within a reasonable amount of time. However, as computer capabilities increase, the calibration of such maps and cluster sets will become a more reasonable enterprise. Third, only cases of shift and trend detection in the mean of data sequences have been undertaken in this work. Effort should be made to develop AIT-based tests for the detection of shifts and trends in other statistical properties, like the variance, for example. Such developments would increase the versatility of AIT-based tests. Fourth, this work has focused on the identification of errors, yet the determination of the sources or causes of these errors is also an interesting topic. Determining sources or causes is however a rather elusive endeavor. Often, measurement stations are automatic, without human surveillance, and consequently there are very few mechanisms to help identify sources and causes of errors. Nevertheless, with a good panel of experts, it may be possible to define a knowledge base that details possible circumstances, whether natural or technical, by which errors in data can be generated. And AITs would be suitable to manipulate this knowledge base so as to infer possible sources and causes of errors. Fifth, and finally, it must be kept in mind the interest of developing decision support systems to provide decisive diagnostic on the

detection of outliers, shifts and trends based on the results of several techniques, whether conventional or AIT-based.

This work has not only focused on data quality control methods applied to observations, but has also directed attention on other types of inputs, that is, model parameters. Positive results are obtained in this respect. Here, the descriptive power of AITs is employed in order to improve the performance of inflow and algae growth models. The characterization of the domains of parameters is accomplished, and fuzzy logic is employed as an inference engine to determine the values of parameters that are deemed to be the most suitable given the conditions of the system under study. In this work, the conditions considered are soil moisture in watersheds for the production of inflows, and energy and nutrient availability for the growth and decay of algae concentrations in a river. This process of describing parameter domains and of using the inference engine has provided some flexibility in models that would have otherwise been constrained by constant parameters values regardless of the conditions of the systems. The idea of conceiving hybrid mechanistic-AIT models is based on the fact that several physical mechanisms that regulate some given processes (e.g., inflows or algae concentration) are well known and it is therefore with relative confidence that they can be mathematically formulated in mechanistic models. Other mechanisms may not be as well known and consequently can lead to a mathematical formulation of a mechanistic model that does not take into account all the possible behaviors manifested by these mechanisms. The role of AITs as applied in this work is to offer a characterization of these lesser known mechanisms through the description of parameter domains and to integrate these mechanisms through the inference engine into mechanistic models.

Of course, the lesser known mechanisms considered in this work are actually known with some degree of certainty, and could have been mathematically formulated in a mechanistic model with some relative confidence. The goal has been to test the validity of hybrid mechanistic-AIT models, and the results obtained here are encouraging. However, it must be noted that some balance must be achieved between the resolution of the description of parameter domains and the efficiency of the optimization procedure employed to calibrate the parameters of the models. The finer the description of parameter domains, the larger the number of parameters needed to calibrate, and the more difficult it becomes for

the optimization procedure to find the global optimum. It may also be harder to find optimal solutions that physically make sense as the number of parameters to calibrate increases. These mechanistic-AIT models should be considered as development tools, where lesser known mechanisms are investigated through the descriptive and inference powers of AITs. From the obtained results, accurate and comprehensive mathematical formulations can be constructed for these lesser known mechanisms. Testing these hybrid models on the largest number of situations, where the involved mechanisms are relatively well known, is needed to further consolidate the validity of this approach.

The descriptive power of AITs, whether for characterizing patterns in data or defining the domains of parameters, is a key factor in the development of the techniques undertaken in this thesis. A possible application that emanates directly from the work accomplished in this thesis is the characterization of a complete database prior to using it for a model. Depending on the patterns present in the data, a database can be subdivided, and specific models can be built for each of the subdivisions. Such procedures would be based on the assumption that each model, which would be applied to particular conditions, would provide better results than a single global model calibrated with the whole database. In water resources, the descriptive power of AITs should be kept in mind if the trends of giving greater focus to ecological impacts from the management of water resources systems continues. These trends refer for example to concerns about water quality parameters, for which knowledge has been acquired and methods have been developed to characterize these parameters. The domain of water quality analysis offers large possibilities for developments, for it often involves working on smaller systems where complex biological features are largely present. In brief, the need for knowledge acquisition through the description of mechanisms and modeling development might still be relatively large for water quality systems. These trends also refer more and more to interactions between water resources and fauna and flora, so that water resources specialists should no longer talk only about water and energy budgets, but also about biological budgets. It is in this respect that efficient descriptive tools are needed. Much remains to be discovered about interactions between physical elements such as water resources and biological elements such as animals and plants, not to mention interactions between living species. And the tasks are not only to

provide descriptions of interactions, but also to quantify the impacts of these interactions, and this might be accomplished with the inference engines provided by AITs.

Chapter 9
Conclusion and Recommendations

The bulk of this work has been dedicated to the development and evaluation of an alternative approach to conventional methods, based on artificial techniques (AIT), for the detection in hydrometric data of anomalies such as shifts, trends and outliers. The results show that the AIT-based detection tests yield performances similar to those of conventional detection tests. As such, AIT-based tests can be used to confirm the results obtained by conventional tests, and also to complement them, because this work shows that tests behave differently from one another. Further work, which includes more complex cases of anomalies and integrates soft data such as experts' judgments, should be accomplished to further verify the case for AITs. The AIT-based tests already show some promise. They may constitute an improvement beyond conventional tests for multivariate cases, and for the estimation of the characteristics of anomalies, such as the location of shifts, or the amplitude of anomalies based on Amp/CV or Amp/SD ratios.

As a whole, it is the descriptive power of AITs that is important to remember, as well as their inference power for modeling scenarios. In water resources, AITs are much more known for their predictive ability. The descriptive power of AITs are of course employed on the issue of data quality for the detection of anomalies, but has also been used here for the characterization of parameter domains on two modeling instances, inflow modeling and algae concentration modeling. Used as descriptive and inference tools, one particular AIT, fuzzy logic, has led to the construction of hybrid mechanistic-AIT models providing improved estimates of inflows and algae concentrations compared with traditional mechanistic models. The mechanistic part is employed to characterize the well known physical mechanisms of the process under investigation (i.e., inflows or algae concentration), while the AIT part describes presumed lesser known mechanisms. This idea of hybrid models should be tested with other modeling scenarios to further verify this approach, and eventually may be used to characterize physical mechanism that are not very well known.

In conclusion, the descriptive power of AITs should be explored further in water resources so as to gain expertise with these approaches and favor developments that would improve the ability to describe processes. Compared with other descriptive statistical techniques, such as clustering, AITs have the advantage of requiring less processing time and memory. If the need to better understand the ecological impacts of water resources management really becomes more important in the future, then powerful descriptive tools would be very useful. In terms of future research and developments related to AITs and data quality, below are a few suggestions that are derived from this thesis.

1. Compare AIT-based tests with other detection techniques such as CUSUM and EWMA.
2. Develop AIT-based detection procedures for cases of multiple outliers, shifts and trends.
3. Develop AIT-based procedures for the detection of shifts and trends in other statistical properties than the mean, like the variance.
4. Develop AIT-based procedure for the identification of sources and causes of errors.
5. Develop decision support systems to provide more decisive diagnostic in the detection of outliers, shifts and trends based on the results of several tests, whether conventional or AIT-based.

Bibliography

Abbott, M.B., Bathurst, J.C., Cunje, J.A., O'Connell, P.E., and Rasmussen, J., 1986a. An Introduction to the European Hydrologic System - Systeme Hydrologique Europeen, SHE 1: History and Philosophy of a Physically-Based, Distributed Modeling System. Journal of Hydrology, 87:45-59.

Abbott, M.B., Bathurst, J.C., Cunje, J.A., O'Connell, P.E., and Rasmussen, J., 1986b. An Introduction to the European Hydrologic System - Systeme Hydrologique Europeen, SHE 2: Structure of a Physically-Based, Distributed Modeling System. Journal of Hydrology, 87:61-77.

Abraham, B., and Box, G.E.P., 1979. Bayesian Analysis of Some Outlier Problems in Time Series. Biometrika, 66(2): 229-236.

Achela, D., Fernando, K., and Jayawardena, A.W., 1998. Runoff Forecasting Using RBF Networks with OLS Algorithm. ASCE Journal of Hydrologic Engineering, 3(3): 203-209.

Aitkin, M., and Wilson, G.T., 1980. Mixture Models, Outliers and the EM Algorithm. Technometrics, 22(3): 325-331.

Anderson, M.G. and Burt, T.P., 1985. Hydrological Forecasting. John wiley & Sons, New York, NY, USA.

Anderson M.G., and Rogers C.C.M., 1987. Catchment Scale Distributed Hydrological Models – A Discussion of Research Directions. Progress in Physical Geography,.11(1): 28-51.

Anderson, J.E., Shiau, S.Y., and Harvey, D., 1992. Preliminary Investigation of Trend/Patterns in Surface Water Characteristics and Climate Variations. National Hydrology research Institute Symposion No. 8, Saskatoon, SK, Canada, April 8-9, 189-201.

Andrews, D. F., and Pregibon, D., 1978. Finding the Outliers that Matter. Journal of the Royal Statistical Society (series B), 40(1): 85-93.

ASCE, 2000a. Artificial Neural Networks in Hydrology - I: Preliminary Concepts. ASCE Journal of Hydrologic Engineering, 5(2): 115-123.

ASCE, 2000b. Artificial Neural Networks in Hydrology - II: Hydrologic Applications. ASCE Journal of Hydrologic Engineering, 5(2): 124-137.

ASQC, 1983. Glossary and Tables for Statistical Quality Control. American Society for Quality Control, Milwaukee, WI, USA.

Assaf, H., and Quick, M.C., 1991. Updating Hydrological Forecasts. Canadian Journal of Civil Engineering, 18(4): 663-674.

Atiya, A.F., El-Shoura, S.M., Shaheen, S.I., and El-Sherif, M.S., 1999. Comparison Between Neural-Network Forecasting Techniques - Case Study: River Flow Forecasting. IEEE Transactions on Neural Networks, 10(2): 402-409.

Bacon-Shone, J., and Fung, W.K., 1987. A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data. Journal of the Royal Statistical Society (Series C), 36(2): 153-162.

Baker, P.D., Brookes, J.D., Burch, M.D., Maier, H.R. and Ganf, G.G., 2000. Advection, Growth and Nutrient Status of Phytoplankton Populations in the Lower River Murray, South Australia. Regulated Rivers: Research and Management, 16: 327-344.

Balcerowska, G., Siuda, R., and Engelhard, H., 2000. Application of PCA and FA in Electron Spectroscopy. II. The minimum energy shift of factors that can be resolved in a set of noisy spectra. Surface and Interface Analysis, 29(8): 492-499.

Bardossy, A., and Disse, M., 1993. Fuzzy Rule-Based Models for Infiltration. Water Resources Research, 29(2): 373-382.

Bardossy, A., and Duckstein, L., 1995. Fuzzy Ruled-Based Modeling with Applications to Geophysical, Biological and Engineering Systems, CRC Press, Boca Raton, Florida, USA.

Bardossy, A., and Kundzewicz, Z.W., 1990. Geostatistical Methods for Detection of Outliers in Groundwater Quality Spatial Fields. Journal of Hydrology, 115: 343-359.

Bartoloni, P., Salas, J.D., and Obeysekera, J.T.B., 1988. Multivariate Periodic ARMA (1,1) Processes. Water Resources Research, 24(8): 1237-1246.

Basseville, M., and Nikiforov, I.V., 1993. Detection of Abrupt Changes: Theory and Application. Prentice Hall, Upper Saddle River, New Jersey, USA.

Beckman, R.J., and Cook, R.D., 1983. Outlier.....s. Technometrics, 25(2): 119-149.

Bender, M.J., and Simonovic, S.P., 2000. A Fuzzy Compromise Approach to Water Resources Systems Planning Under Uncertainty. Fuzzy Sets and Systems, 115: 35-44.

Bennis, S., and Bruneau, P., 1993a. Comparaison de méthodes d'estimation des débits journaliers. Canadian Journal of Civil Engineering, 20(3): 480-489.

Bennis, S., and Bruneau, P., 1993b. Amélioration de méthodes d'estimation des débits journaliers. Canadian Journal of Civil Engineering, 20(3): 490-499.

Bergman, M.J., and Delleur, J.W., 1985a. Kalman Filter Estimation and Prediction of Daily Stream Flows: I. Review, Algorithm, and Simulation Experiments. Water Resources Bulletin, 21(3): 815-825.

Bergman, M.J., and Delleur, J.W., 1985b. Kalman Filter Estimation and Prediction of Daily Stream Flows: II. Application to the Potomac River. Water Resources Bulletin, 21(3): 827-832.

Bérubé, R., Charbonneau, R., and Dolbec, M., 1987. Validations des données hydrométriques historiques de la rivière des Outaouais. Technical Report, Hydro-Québec, Montreal, PQ, Canada.

Beven, K.J., 1983. Surface Water Hydrology - Runoff Generation and Basin Structure. Reviews of Geophysics and Space Physics, 21(3): 721-730.

Beven, K.J., 1989. Changing Ideas in Hydrology - The Case of Physically-Based Models. Journal of Hydrology, 105:157-172.

Beven, K.J., 2001. Rainfall-Runoff Modeling: The Primer. John Wiley & Sons, New York, NY, USA.

Beven, K.J., and Freer, J., 2001. Equifinality, Data Assimilation, and Uncertainty Estimation in Mechanistic Modelling of Complex Environmental Systems Using the GLUE Methodology. Journal of Hydrology, 249: 11-29.

Beven, K.J., Wood, E.F., Sivapalan, M., 1988. On Hydrological Heterogeneity - Catchment Morphology and Catchment Response. Journal of Hydrology, 100: 352-375.

Bezdek, J.C., 1981. Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York, NY, USA.

Birikundavyi, S., Rousselle, J., and Nguyen, V.T.V., 1993. Determination des regions homogenes pour le Quebec et l'Ontario: Une approche par l'analyse des correspondances et la classification ascendante hierarchique. Final report, NSERC grant STR 0118482, Departement de genie civil, Ecole Polytechnique de Montreal, Montreal, PQ, Canada.

Bouchard, S., 1986. Amelioration d'un modele hydrologique deterministe et son application a la prevision des ruissellements du bassin du Lac-Saint-Jean. Master's thesis, Universite du Quebec a Chicoutimi, Chicoutimi, PQ, Canada

Bouchard, S. and Salesse, L., 1986. Amelioration et structuration du systeme de prevision hydrologique a court terme PREVIS. Report RH 86-01, Groupe Ressources Hydriques, EEQ, SECAL, Jonquiere, PQ, Canada.

Bowden, G.J., Maier, H.R., and Dandy, G.C., 2002. Optimal division of data for neural network models in water resources applications. Water Resources Research, 38(2):2.1-2.11.

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

Box, G.E.P., and Tiao, G.C., 1968. A Bayesian Approach to Some Outlier Problems. Biometrika, 55(1): 119-129.

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

Brackstone, G., 1999. Managing Data Quality in a Statistical Agency. Statistics Canada, Survey Methodology, Catalogue no. 12-001-XPB, Vol. 25, No.2.

Bradu, D., and Hawkins, D.M., 1982. Location of Multiple Outliers in Two-Way Tables, Using Tetrads. Technometrics, 24(2): 103-108.

Bras, R.L., and Rodriguez-Iturbe, I., 1985. Random Functions and Hydrology. Addison-Wesley Publishing Company, Reading, MA, USA.

Brown, B.M., 1975. A Short-Cut Test for Outliers Using Residuals. Biometrika, 62(3): 623-629.

Burn, D.H., 1990. Evaluation of Regional Flood Frequency Analysis with a Region of Influence Approach. Water Resources Research, 26(10): 2257-2265.

Burn, D.H., 1994. Hydrologic Effects of Climatic Change in West-Central Canada. Journal of Hydrology, 160: 53-70.

Burn, D.H., 1997. Catchment Similarity for Regional Flood Frequency Analysis Using Seasonality Measures. Journal of Hydrology, 202: 212-230.

Cavadias, G.S., and Gupta, S.K., 1978. Stochastic Analysis of the Residuals of a Conceptual Model. International Symposium on Risk and Reliability in Water Resources, University of Waterloo, Waterloo, ON, Canada, June 26-28, E.A. McBean, K.W. Hipel and T.E. Unny, Water Resources Publications, Highlands Ranch, Colorado, 536-555.

Cavadias, G., and Morin, G., 1986. The Combination of Simulated Discharges of Hydrological Models. Nordic Hydrology, 17(1): 21-32.

Chan, L.K., and Zhang, J., 2000. Some Issues in the Design of EWMA Charts. Communications in Statistics Part B: Simulation and Computation, 29(1): 207-217.

Chan-Yan, D.A., 2000. Reservoir Turbidity Modelling Using Artificial Neural Networks and the Estimation of Performance Indicators. Master's Thesis, The University of British Columbia, Vancouver, BC, Canada.

Chang, N.B., Chen, H.W., Shaw, D.G., and Yang, C.H., 1997. Water Pollution Control in River Basin by Interactive Fuzzy Interval Multiobjective Programming. ASCE Jounral of Environmental Engineering, 123(12): 1208-1216.

Chen, C.H., 1996. Fuzzy Logic and Neural Network Handbook, McGraw-Hill, New York, NY, USA.

Clarke, R.T., 1973. A Review of Some Mathematical Models Used in Hydrology, with Observations on their Calibration and Use. Journal of Hydrology, 19: 1-20.

Cloot, A.H.J. and Pieterse, A.J.H., 1999. Modelling Phytoplankton in the Vaal River (South Africa). Water Science and Technology, 40: 119-124.

Collett, D., and Lewis, T., 1976. The Subjective Nature of Outlier Rejection Procedures. Journal of the Royal Statistical Society, Series C, 25(3): 228-237.

Conover, W.J., 1980. Practical Nonparametric Statistics, 2nd Edition. John Wiley & Sons, New York, NY, USA.

Cook, R.D., Holschuh, N., and Weisberg, S., 1982. A Note on an Alternative Outlier Model. Journal of the Royal Statistical Society (Series B), 44(3): 370-376.

Coulibaly, P., Anctil, F., and Bobee, B., 1999. Prevision hydrologique par reseaux de neurones artificiels: etat de l'art. Canadian Journal of Civil Engineering, 26(3): 293-304.

Coulibaly, P., Anctil, F., and Bobee, B., 2000. Daily Reservoir Inflow Forecasting Using Artificial Neural Networks with Stopped Training Approach. Journal of Hydrology, 230(3): 244-257.

Cunderlik, J.M., and Burn, D.H., 2002. Local and Regional Trends in Monthly Maximum Flows in Southern British Columbia. Canadian Water Resources Journal, 27(2): 191-212.

Delvin, S.J., Gnanadesikan, R., and Kettenring, J.R., 1975. Robust Estimation and Outlier Detection with Correlation Coefficients. Biometrika, 62(3): 531- 545.

Dempster, A.P., and Gasko-Green, M., 1981. New Tools for Residual Analysis. The Annal of Statistics, 9(5): 945-959.

Derksen, C., LeDrew, E., Walker, A., and Goodison, B., 2000. Influence of Sensors Overpass Time on Passive Microwave-Derived Snow Cover Parameters. Remote Snesing of Environement, 71(3): 297-308.

Despic, O., and Simonovic, S.P., 2000. Aggregation Operators for Soft Decision Making in Water Resources. Fuzzy Sets and Systems, 115: 11-33.

Dillon, W.R., and Goldstein, M., 1984. Multivariate Analysis: Methods and Applications. John Wiley & Sons, New York, NY, USA.

Di Toro, D.M., O'Connor, D.J. and Thomann, R.V., 1971. A Dynamic Model of the Phytoplankton Population in the Sacramento – San Joaquin Delta, Advance in Chemistry Series. American Chemical Society, 106: 131-180.

Diskin, M.H., and Simon, E., 1997. A Procedure for the Selection of Objective Functions for Hydrologic Simulation Models. Journal of Hydrology, 34: 129-149.

Dou, C., Woldt, W., and Bogardi, I., 1999. Fuzzy Rule-Based Approach to Describe Solute Transport in the Unsaturated Zone. Journal of Hydrology, 220(1): 74-85.

Dou, C., Woldt, W., Bogardi, I., and Dahab, M., 1995. Steady State Groundwater Flow Simulation with Imprecise Parameters. Water Resources Research, 31(11): 2709-2719.

Draper, N.R., and John, J.A., 1981. Influential Observations and Outliers in Regression. Technometrics, 23(1): 21-26.

Dubois, D., and Prade, H., 1980. Fuzzy Sets and Systems: Theory and Applications. Academic Press, New York, NY, USA.

Ellis, J.H., 1988. Acid Rain Control Stategies - Options Exist Despite Scientific Uncertainties. Environmental Science and Technology, 22(11): 1248-1255.

Ellis, J.H., 1990. Integrating Multiple Long-Range Transport Models into Optimization Methodologies for Acid Rain Policy Analysis. European Journal of Operational Research, 46(3): 313-321.

Endreny, T.E., and Jennings, G.D., 1999. A Decision Support System for Water Quality Data Augmentation: A Case Study. Journal of the American Water Resources Association, 35(2): 363-377.

Entekhabi, D., Asrar, G.R., Betts, A.K., Beven, K.J., Bras, R.L., Duffy, C.J., Dunne, T., Koster, R.D., Lettenmaier, D.P., McLaughlin, D.B., Shuttleworth, W.J., VanGenuchten, M.T., Wei, M.Y., Wood, E.F., 1999. An Agenda for Land Surface Hydrology Research and a Call for the Second International Hydrological Decade. Bulletin of the American Meteorological Society, 80(10): 2043-2058.

EPA, 1985. Rates, Constants, and Kinetics Formulations in Surface Water Quality Modeling, 2nd ed. Report EPA/600/3-85/040, Environmental Research Laboratory, Office of Research and Developpment, U.S. Environmental Protection Agency, Athens, GA, USA.

Farnum, N.R., 1994. Modern Statistical Quality Control and Improvement. Duxbury Press, Belmont, CA, USA.

Fernandez, B., and Salas, J.D., 1986. Periodic Gamma Autoregressive Processes for Operational Hydrology. Water Resources Research, 22(10): 1385-1396.

Fontane,, D.G., Gates, T.K., and Moncada, E., 1997. Planning Reservoir Operations with Imprecise Objectives. ASCE Journal of Water Resources Planning and Management, 123(3): 154-162.

Fox, A.J., 1972. Outliers in Time Series. Journal of the Royal Statistical Society (Series B), 34(3): 350-363.

Franks, S.W., and Beven, K.J., 1997. Estimation of Evapotranspiration at the Landscape Scale: A Fuzzy Disaggregation Approach. Water Resources Research, 33(12): 2929-2938.

Freeze, R.A., and Cherry, J.A., 1979. Groundwater. Prentice Hall, Englewood Cliffs, NJ, USA.

Gan, T.Y., 1992. Finding Trends in Air Temperature and Precipitation for Canada and North-EAsthern United States. National Hydrology research Institute Symposion No. 8, Saskatoon, SK, Canada, April 8-9, 57-78.

Gan, T.Y., 1998. Hydroclimatic Trends and Possible Climatic Warming in the Canadian Prairies. Water Resources Research, 34(11): 3009-3015.

Gan, T.Y., and Kwong, Y.T.J., 1992. Identification of Warming Trends in Northern Alberta and Southern Northwest Territories by the Non-Parametric Kenndall's Test. National Hydrology research Institute Symposion No. 8, Saskatoon, SK, Canada, April 8-9, 43-56.

Gautam, M.R., Watanabe, K., and Saegusa, H., 2000. Runoff Analysis in Humid Forest Catchment with Artificial Neural Network. Journal of Hydrology, 235(1): 117-136.

Georgakakos, K.P., 1986a. A Generalized Stochastic Hydrometeorological Model for Flood and Flash-Flood Foercasting - 1. Formulation. Water Resources Research, 22(13):2083-2095.

Georgakakos, K.P., 1986b. A Generalized Stochastic Hydrometeorological Model for Flood and Flash-Flood Foercasting - 2. Case Studies. Water Resources Research, 22(13):2096-2106.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, Reading, MA, USA.

Gotz, R., Steiner, B., Sievers, S., Friesel, P., Roch, K., Schworer, R., and Haag, F. (1998). "Dioxin, Dioxin-Like PCBS and Organotin Compounds in the River Elbe and the Hamburg Harbour: Identification of Sources." Water Science and Technology, 37(6-7), 207-215.

Gray, D.M., and Male, D.H., 1981. Handbook of Snow: Principles, Processes, Management and Use. Pergamon Press Canada, Toronto, ON, Canada.

Grayson, R.B., Moore, I.D., and McMahon, T.A., 1992a. Physically Based Hydrologic Modeling: 1. A Terrain-Based Model for Investigative Purposes. Water Resources Research, 28(10): 2639-2658.

Grayson, R.B., Moore, I.D., and McMahon, T.A., 1992b. Physically Based Hydrologic Modeling: 2. Is the Concept Realistic. Water Resources Research, 28(10): 2659-2666.

GREHYS, 1996. Presentation and Review of Some Methods for Regional Flood Frequency Analysis. Journal of Hydrology, 186(1): 63-84.

Hall, M. J., and Minns, A. W. (1999). "The Classification of Hydrologically Homogeneous Regions." Hydrological Sciences Journal, 44(5), 693-704.

Harte, J., 2002. Toward a Synthesis of the Newtonian and Darwinian Worldviews. Physics Today, 55(10): 29-34.

Hipel, K.W., 1975. Contemporary Box-Jenkins Modelling in Hydrology. Ph.D. thesis, Department of Civil Engineering, University of Waterloo, Waterloo, ON, Canada.

Hipel, K.W., McLeod, A.I., and Noakes, D.J., 1981. Fitting Dynamic Models to Hydrological Time Series. International Conference on Time Series Methods in Hydrosciences, Canada Center for Inland Water, Burlington, Canada, October 6-8, 110-129.

Hirsh, R.M., and Slack, J.R., 1984. A Nonparametric Trend test for Seasonal Data with Serial Dependence. Water Resources Research, 20(6): 727-732.

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

Hirsh, R.M., Helsel, D.R., Cohn, T.A., and Gilroy, E.J., 1993. Statistical Analysis of Hydrologic Data. In Handbook of Hydrology, D.R. Maidment (ed.), McGraw-Hill, New York, NY, USA, Chapter 17.

Holt, T., and Jones, T., 1998. Quality Work and Conflicting Quality Objectives. 84th Director Generals of National Statistical Institutes (DGINS) Conference, Stockholm, Sweden, May 28-29.

Hosking, J.R.M., 1989. The Theory of probability Weighted Moments. Report RC 12210, IBM Research Division, T.J. Watson Research Center, Yorktown Heigths, NY, USA.

Hosking, J.R.M., 1990. L-Moments: Analysis and Estimation of Distributions Using Linear Combination of Order Statistics. Journal of the Royal Statistical Society (Series B), 52(1): 105-124.

Hosking, J.R.M., 1995. The Use of L-Moments in the Analysis of Censored Data. In Recent Advances in Life-Testing and Reliability, N. Balakrishnan (ed.), CRC Press, Boca Raton, FL, USA, Chapter 29.

Hosking, J.R.M., and Wallis, J.R., 1990. Regional Flood Frequency Analysis Using L-Moments. Report RC 15658, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY, USA.

Iritz, L., 1988. Application of an Adaptive Forecast Algorithm to the River Vasterdalalven. Nordic Hydrology, 19(5): 293-302.

Imrie, C.E., Durucan, S., and Korre, A., 2000. River Flow Prediction Using Artificial Neural Networks: Generalisation Beyond the Calibration Range. Journal of Hydrology, 233(1): 138-153.

Jain, R.B., 1981. Percentage Points of Many-Outlier Detection Procedures. Technometrics, 23(1): 71-75.

Jarpe, E., and Wessman, P., 2000. Some Power Aspects of Methods for Detecting Different Shifts in the Mean. Communications in Statistics. Part B: Simulation and Computation, 29(2): 633-646.

Jones, G., Baker, P.D. and Burch, M.D., 2000. National Protocol for the Monitoring of Cyanobacteria in Surface Waters. Final report, Australian Center for Water Quality, Adelaide, SA, Australia.

Khoo, M.B.C., and Quah, S.H., 2002. Computing the Percentage Points of the Run-Length Distributions of Multivariate CUSUM Control Charts. Quality Engineering, 15(2): 299-310.

Kindler, J., 1992. Rationalizing Water Requirements with Aid of Fuzzy Allocation Model. ASCE Journal of Water Resources Planning and Management, 118(3): 308-323.

Kitagawa, G., 1979. On the Use of the AIC for the Detection of Outliers. Technometrics, 21(2): 193-199.

Kitanidis, P.K., 1993. Geostatistics. In Handbook of Hydrology, D.R. Maidment (ed.), McGraw-Hill, New York, NY, USA, Chapter 20.

Kitanidis, P.K., and Bras, R.L., 1980a. Real-Time Forecasting with a Conceptual Hydrologic Model - 1. Analysis of Uncertainty. Water Resources Research, 16(6): 1025-1033.

Kitanidis, P.K., and Bras, R.L., 1980b. Real-Time Forecasting with a Conceptual Hydrologic Model - 1. Applications and Results. Water Resources Research, 16(6): 1034-1044.

Klemes, V., 1983. Conceptualization and Scale in Hydrology. Journal of Hydrology, 65: 1-23.

Kohonen, T., 1990. The Self-Organizing Map. Proceedings of the IEEE, 79(9): 1464-1480.

Kohonen, T., 1997. Self-Organizing Maps, Second Edition. Springer-Verlag, Berlin, Germany.

Kottegoda, N.T., 1984. Investigation of Outliers in Annual Maximum Flow Series. Journal of Hydrology, 72: 105-137.

Krawjeski, W.F., 1987. Radar Rainfall Data Quality Control by the Influence Function Method. Water Resources Research, 23(5): 837-844.

Krawjeski, W.F., and Krawjeski, K.L., 1989. Real-Time Quality Control of Streamflow Data - A Simulation Study. Water Resources Bulletin, 25(2): 391-399.

Kruszewski, S., Siuda, R., Ziomkowska, B., and Cyrankiewicz, M., 2003. PCA and FA Analysis of Steady-State Fluorescence Spectra of Camptothecin. Proceedings of SPIE - The International Society for Optical Engineering, 5064: 84-90.

Lauzon, N., 1993. Méthodes de validation et de prévisions à court terme des apports naturels. Master's thesis, Département de génie civil, École Polytechnique de Montréal, Montreal, PQ, Canada.

Lauzon, N., Rousselle, J., Birikundavyi, S., and Trung, H.T., 2000. Real-Time Daily Flow Forecasting Using Black-Box Models, Diffusion Processes, and Neural Networks. Canadian Journal of Civil Engineering, 27(4): 671-682.

Lee, A.F.S., and Heghinian, S.M., 1977. A Shift of the Mean Level of Independent Normal Random Variables - A Bayesian Approach. Technometrics, 19(4): 503-506.

Lehman, J.T., Botkin, D.B. and Likens, G.E., 1975. The Assumption and Rationales of a Computer Model of Phytoplankton Population Dynamics. Limnology and Oceanography, 20: 343-364.

Leith, R.M.M., and Whitfield, P.H., 1998. Evidence of Climate Change Effects on the Hydrology of Streams in South-Central BC. Canadian Water Resources Journal, 23(3): 219-230.

Lek, S., Guiresse, M., and Giraudel, J.L., 1999. Predicting Stream Nitrogen Concentration from Watershed Features Using Neural Networks. Water Research, 33(16): 3469-3478.

Lettenmaier, D.P., 1976. Detection of Trends in Water Quality Data from Records with Dependent Observations. Water Resources Research, 12(5): 1037-1046.

Lettenmaier, D.P., 1980. Intervention Analysis with Missing Data. Water Resources Research, 16(1): 159-171.

Lettenmaier, D.P., and Burges, S.J., 1976. Use of State Estimation Techniques in Water Resource System Modeling. Water Resources Bulletin, 12(1): 83-98.

Lettenmaier, D.P., Wood, E.F., and Wallis, J.R., 1994. Hydro-Climatological Trends in the Continental United States, 1948-1988. Journal of Climate, 7: 586-607.

Liepins, G.E., 1989. Sound Data Are a Sound Investment. Quality Progress, 22(9): 61-64.

Lins, H.F., and Slack, J.R., 1999. Streamflow Trends in the United States. Geophysical Research Letters, 26(2): 227-230.

Linsley, R.K., Kohler, M. A., and Paulhus, J., 1982. Hydrology for Engineers, 3rd ed. McGraw-Hill, New York, NY, USA.

Liong, S. Y., Lim, W. H., Kojiri, T., and Hori, T. (2000). "Advance Flood Forecasting for Flood Stricken Bangladesh with a Fuzzy Reasoning Method." Hydrological Processes, 14(3), 431-448.

Ljung, G.M., 1993. On Outlier Detection in Time Series. Journal of the Royal Statistical Society (Series B), 55(2): 559-567.

Lloyd, E., 1984. Handbook of Applicable Mathematics, Volume VI: Statistics, Part B. John Wiley & Sons, New York, NY, USA.

Lundberg, A., 1982. Combination of a Conceptual Model and an Autoregressive Error Model for Improving Short Time Forecasting. Nordic Hydrology, 13(4): 233-246.

Lung, W.S. and Larson, C.E., 1995. Water Quality Modeling of Upper Mississippi River and Lake Pepin. Journal of Environmental Engineering, ASCE, 121: 691-699.

Maidment. D.R. (ed.), 1993. Handbook of Hydrology. McGraw-Hill, New York, NY, USA.

Maier, H.R., and Dandy, G.C., 2000. Neural Networks for the Prediction and Forecasting of Water Resources Variables: A Review of Modelling Issues and Applications. Environmental Modelling and Software, 15(1): 101-124.

Maier, H.R., Dandy, G.C., and Burch, M.D., 1998. Use of Artificial Neural Networks for Modelling Cyanobacteria Anabaena spp. in the River Murray, South Australia. Ecological Modeling, 105: 257-272.

Maier, H.R., Sayed, T., and Lence, B.J., 2000. Forecasting Cyanobacterial Concentrations Using B-Spline Networks. ASCE Journal of Computing in Civil Engineering, 14(3): 183-189.

Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J.M., and Francis, R.C., 1997. A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. Bulletin of the American Meteorological Society, 78(6): 1069–1079.

Martin, J.L., McCutcheon, S.C., Schottman, R. W., 1999. Hydrodynamics and transport for water quality modeling. Lewis Publishers, Boca Raton, FD, USA.

McLaughlin, D.B., 1980. Application of Kalman Filtering to Groundwater Basin Modeling and Prediction. In Real-Time Forecasting / Control of Water Resources, E.F. Wood (ed.), Pergamon Press, Oxford, UK, 109-123.

McLeod, A.I., Hipel, K.W., and Comancho, F., 1983. Trend Assessment of Water Quality Time Series. Water Resources Bulletin, 19(4): 537-547.

Nash, S.G., and Sofer, A., 1996. Linear and Nonlinear Programming. McGraw-Hill, New York, NY, USA.

Nguyen, V.T.V., 1993. Validation des données des apport naturel journaliers. Research Report No. WRM93/1, Water Resources Management Series, McGill University, Montreal, PQ, Canada.

Morshed, J., and Kaluarachchi, J.J., 1998. Application of Artificial Neural Network and Genetic Algorithm in Flow and Transport Simulations. Advances in Water Resources, 22(2): 145-158.

Neelakantan, T.R., and Pundarikanthan, N.V., 2000. Neural Network-Based Simulation-Optimization Model for Reservoir Operation. ASCE Journal of Water Resources Planning and Management, 126(2): 57-64.

Orr, K., 1998. Data Quality and Systems Theory. Communications of the ACM, 41(2): 66-71.

Overton, J.McC., Young, T.C., and, Overton, W.S., 1993. Using 'Found" Data to Augment a Probability Sample: Procedure and Case Study. Environmental Monitoring and Assessment, 26: 65-83.

Ozelkan, E.C., Ni, F., and Duckstein, L., 1996. Relationship between Monthly Atmospheric Circulation Patterns and Precipitation: Fuzzy Logic and Regression Approaches. Water Resources Research, 32(7): 2097-2103.

Pagano, M., 1978. On Periodic and Multiple Autoregressions. The Annals of Statistics, 6(6): 1310-1317.

Pankratz, A., 1983. Forecasting with Univariate Box-Jenkins Models: Concepts and Cases. John Wiley & Sons, New York, NY, USA.

Pearson, R.K., 2001. Exploring Process Data. Journal of Process Control, 11(2): 179-194.

Perreault, L., Roy, R., Bobée, B., and Mathier, L., 1991. Validation et estimation des apports journaliers. Research Report 271, INRS-Eau, Université du Québec, Sainte-Foy, PQ, Canada.

Perreault, L., Haché, M., Slivitzky, M., and Bobée, B., 1999. Detection of Changes in Precipitation and Runoff over Eastern Canada and U.S. Using a Bayesian Approach. Stochastic Environmental Research and Risk Assessment, 13: 201-216.

Perreault, L., Parent, E., Bernier, J., Bobée, B., and Slivitzky, M., 2000. Stochastic Environmental Research and Risk Assessment, 14: 243-261.

Pesti, G., Shrestha, B.P., Duckstein, L., and Bogardi, I., 1996. A Fuzzy Rule-Based Approach to Drought Assessment. Water Resources Research, 32(6): 1741-1747.

Prabhu, S.S., and Runger, G.C., 1997. Designing a Multivariate EWMA Control Chart. Journal of Quality Technology, 29(1): 8-15.

Radharamanan, R., Galelli, A., Alex, D.T., and Perez, L.L., 1994. Sensitivity Analysis on the CUSUM Method. International Journal of Production Economics, 33(1-3): 89-95

Rao, S.S., 1979. Optimization: Theory and Applications. John Wiley & Sons, New York, NY, USA.

Rango, A., and Martinec, J., 1995. Revisiting the Degree-Day Method for Snowmelt Computations. Water Resources Bulletin, AWRA, 31(4): 657-669.

Rassam, J.C., Bérubé, R., Bisson, J.L., Carpentier, A., Hoang, V.D., Jaworski, L., Nix, G.A., Pilon, P. Roy, R., Tremblay, D., and Villar, J., 1991. Gestion du risque du système hydrique de la rivière des Outaouais. Final report, sous-comité Gestion du risque, Commission de planification de la régularisation de la rivière des Outaouais, Montreal, PQ, Canada.

Reynolds M.R., and Arnold, J.C., 2001. EWMA Control Charts with Variable Sample Sizes and Variable Sampling Intervals. IIE Transactions (Institute of Industrial Engineers), 33(6): 511-530.

Reynolds, M.R., and Stoumbos, Z.G., 2000. General Approach to Modeling CUSUM Charts for a Proportion. IIE Transactions (Institute of Industrial Engineers), 32(6): 515-535.

Ribeiro-Correa, J., Cavadias, G.S., Clement, B., and Rousselle, J., 1995. Identification of Hydrological Neighborhoods Using Canonical Correlation analysis. Journal of Hydrology, 173(1-4):71-89.

Roberts, J., 2000. Influence of physical and physiological characteristics of vegetation on their hydrological response. Hydrological Processes, 14(16-17): 2885-2901.

Robertson, J.A., Cassidy, J.J., and Chaudhry, M.H., 1988. Hydraulic Engineering. Houghton Mifflin Company, Boston, MA, USA.

Rodriguez-Iturbe, I., and Valdes, J.B., 1979. The geomorphological structure of hydrologic response. Water Resources Research, 15(6), 1409-1420.

Rosner, B., 1975. On the Detection of Many Outliers. Technometrics, 17(2): 221-227.

Rousselle, J., Debs, A., Lauzon, N., and Birikundavyi, S., 1999. Modèles hydrologiques de prévision des apports – Revue de littérature. Project CDT-P2350, Centre de développement technologique, École Polytechnique de Montréal, Montréal, PQ, Canada.

Russell, S.O., and Campbell, P.F., 1996. Reservoir Operating Rules with Fuzzy Programming. ASCE Journal of Water Resources Planning and Management, 122(3): 165-170.

Rutherford, J. C., Scarsbrook, M. R. and Broekhuizen, N., 2000. Grazer Control of Stream Algae: Modeling Temperature and Flood Effects. Journal of Environmental Engineering, ASCE, 126: 331-339.

Sajikumar, N., and Thandaveswara, B.S., 1999. Non-Linear Rainfall-Runoff Model Using an Artificial Neural Network. Journal of Hydrology, 216(1-2): 32-55.

Salas, J.D., 1993. Analysis and Modeling of Hydrologic Time Series. In Handbook of Hydrology, D.R. Maidment (ed.), McGraw-Hill, New York, NY, USA, Chapter 19.

Salas, J.D., Delleur, J.W., Yevjevich, V., and Lane, W.L., 1980. Applied Modeling of Hydrologic Time Series. Water Resources Publication, Littleton, CO, USA.

Sasikumar, K., and Mujumdar, P.P., 1998. Fuzzy Optimization Model for Water Quality Management of a River System. ASCE Journal of Water Resources Planning and Management, 124(2): 79-88.

See, L., and Openshaw, S., 1999. Applying Soft Computing Approaches to River Level Forecasting. Hydrological Sciences Journal, 44(5): 763-778.

Sen, P. K., 1968. Estimates of the Regression Coefficient Based on Kendall's Tau. Journal of the American Statistical Association, 63: 1379-1389.

Setnes, M., Babuska, R., Verbruggen, H.B., Sanchez, M.D. and van den Boogaard, H.F.P., 1997. Fuzzy Modeling and Similarity Analysis Applied to Ecological Data. 6th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'97), Barcelona, Spain, 1:415-420.

Setnes, M., Babuska, R., and Verbruggen, H.B., 1998. Transparent Fuzzy Modelling. International Journal of Human Computer Studies, 49(2): 159-179.

Shrestha, B.P., Duckstein, L., and Stakhiv, E.Z., 1996. Fuzzy Rule-Based Modeling of Reservoir Operation. ASCE Journal of Water Resources Planning and Management, 122(4): 262-269.

Siew-Yan-Yu, T.O., Rousselle, J., Jacques, G., and Nguyen, V.T.V., 1998. Regionalisation du regime des precipitations dans la region des Bois-Francs et de l'Estrie par l'analyse en composantes principales. Canadian Journal of Civil Engineering, 25(6): 1050-1058.

Singh, V.P., 1989. Hydrologic Systems - Volume II: Watershed Modeling. Prentice Hall, Englewood Cliffs, NJ, USA.

Singh, V.P., 1995. Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, CO., USA.

Singh, V.P., and Woolhiser, D.A., 2002. Mathematical Modeling of Watershed Hydrology. ASCE Journal of Hydrologic Engineering, 7(4): 270-292.

Srivastava, M.S., and Wu, Y., 1997. Evaluation of Optimum Weights and Average Run Lengths in EWMA Control Schemes. Communications in Statistics. Part A: Theory and Methods, 26(5): 1253-1267.

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou, E., 1993. Frequency Analysis of Extreme Events. In Handbook of Hydrology, D.R. Maidment (ed.), McGraw-Hill, New York, NY, USA, Chapter 18.

Suykens, J.A.K., Vandewalle, J.P.L., and DeMoor, B. L. R., 1996. Artificial Neural Networks for Modelling and Control of Non-Linear Systems, Kluwer Academic Publishers, Boston, MA, USA.

Terano, T., Asai, K., and Sugeno, M., 1992. Fuzzy Systems Theory and Its Applications. Academic Press, New York, NY, USA.

Thirumalaiah, K., and Deo, M.C., 1998. River Stage Forecasting Using Artificial Neural Networks. ASCE Journal of Hydrologic Engineering, 3(1): 26-32.

Thompstone, R.M., 1983. Topics in Hydrological Time Series Modeling. Ph.D. thesis, Department of Civil Engineering, University of Waterloo, Waterloo, ON, Canada.

Thompstone, R.M., Hipel, K.W., and McLeod, A.I., 1985. Forecasting Quater-Monthly Riverflow. Water Resources Bulletin, 21(3): 731-741.

Troutman, B.M., 1979. Some Results in Periodic Autoregression. Biometrika, 66(2): 219-228.

VanDerShaaf, S., 1984. Errors in Level Recorder Data: Prevention and Detection. Journal of Hydrology, 73:373-382.

VanGeer, F.C., TeStroet, C.B.M., and Yangxiao, Z., 1991. Using Kalman Filtering to Improve and Quantify the Uncertainty of Numerical Groundwater Simulations - 1. The Role of System Noise and Its Calibration. Water Resources Research, 27(8): 1987-1994.

Vecchia, A.V., 1985. Periodic Autoregressive-Moving Average (PARMA) Modeling with Applications to Water Resources. Water Resources Bulletin, 21(5): 721-730.

Vicens, G.J., Rodriguez-Iturbe, I., and Schaake, J.C., 1975. A Bayesian Framework for the Use of Regional Information in Hydrology. Water Resources Research, 11(3): 405-414.

Walker, W.W., 1975. Description of the Charles River Basin Model, Final Report on the Storrow Lagoon Demonstration Plant. Submitted to the Commonwealth of Massachusetts, Metropolitan District Commission, By Process Research, Inc., Cambridge, MA, USA, Chapter 6.

Wang, R.Y., Kon, H.B., and Madnick, S.E., 1993. Data Quality Requirements Analysis and Modeling. Ninth International Conference on Data Engineering, Vienna, Austria, April 19-23.

West, M., 1984. Outlier Models and Prior Distribution in Bayesian Linear regression. Journal of the Royal Statistical Society (Series B), 46(3): 431-439.

Westmaccott, J.R., and Burn, D.H., 1997. Climate Change Effects on the Hydrologic Regime within the Churchill-Nelson River Basin. Journal of Hydrology, 202: 263-279.

Whitehead, P.G., and Hornberger, G.M., 1984. Modelling Algal Behaviour in the River Thames. Water Research, 18: 945-953.

Winter, T.C., Mallory, S.E., Allen, T.R., and Rosenberry, D.O., 2000. Use of Principal Component Analysis for Interpreting Ground Water Hydrographs. Ground-Water, 38(2): 243-246.

WMO, 1975. Intercomparison of Conceptual Models Used in Operational Hydrological Forecasting. Operational Hydrology Report No. 7, WMO Report No. 429, Worl Meteorological Organization, Geneva, Switzerland.

WMO, 1985. Guidelines for Computerized Data Processing in Operational Hydrology and Land and Water Management. WMO Report no. 634, World Meteorological Organization, Joint FAO/WMO publication, Geneva, Switzerland.

WMO, 1986. Intercomparison of Models of Snowmelt Runoff. Operational Hydrology Report No. 23, WMO Report No. 646, Worl Meteorological Organization, Geneva, Switzerland.

WMO, 1992., Simulated Real-Time Intercomparison of Hydrological Models. Operational Hydrology Report No. 38, WMO Report No. 779, Worl Meteorological Organization, Geneva, Switzerland.

Wood, E.F., Sivapalan, M., and Beven, K., 1990. Similarity and Scale in Catchment Storm Response. Reviews of Geophysics, 28(1): 1-18.

Woolhiser, D.A., 1996. Search for Physically Based Runoff Model - A Hydrologic El Dorado? ASCE Journal of Hydraulic Engineering, 122(3): 122-129.

Wurbs, R.A., 1998. Dissemination of Generalized Water Resources in the United States. Water International, 23(3): 190-198.

Yabunaka, K.I., Hosomi, M., and Murakami, A., 1997. Novel Application of a Back-Propagation Artificial Neural Network Model Formulated to Predict Algal Bloom. Water Science and Technology, 36(5): 89-97.

Yin, Y.Y., Huang, G.H., and Hipel, K.W., 1999. Fuzzy Relation Analysis for Multicritria Water Resources Management. Journal of Water resources Planning and Management, 125(1): 41-47.

Yulianti, J.S., and Burn, D.H., 1998. Investigating Links Between Climatic Warming and Low Streamflow in the Prairies Region of Canada. Canadian Water Resources Journal, 23(1): 45-60.

Zafirakou, A., Vogel, R.M., Craig, S.M., and Habermeier, J., 1998. L-Moment Diagrams for Censored Observations. Water Resources Research, 34(5): 1241-1249.

Zhang, X., Harvey, K.D., Hogg, W.D., Yuzyk, T.R., 2001. Trends in Canadian Streamflow. Water Resources Research, 37(4): 987-998.

Zimmermann, H.J., 1991. Fuzzy Set Theory and Its Applications, Second, Revised Edition. Kluwer Academic Publishers, Boston, MA, USA.

Zrinji, Z., Burn, D.H., 1994. Flood Frequency Analysis for Ungauged Sites Using a Region of influence Approach. Journal of Hydrology, 153: 1-21.