A comprehensive system for the exploration and analysis of hydrological data

George Tsakalias and Demetris Koutsoyiannis

Department of Water Resources, Faculty of Civil Engineering, National Technical University, Athens, Heroon Polytechneiou 5, GR-157 80, Zografou, Greece (dk@hydro.ntua.gr)

Abstract. A new approach is developed for the automatic (computer based) exploration and analysis of hydrological data, particularly focused on the identification of shifting relationships among hydrological variables. The methodology developed is applicable to many hydrological problems, such as identification of multiple stage-discharge relationships in a river section, data homogeneity analysis, analysis of temporal consistency of hydrological data, detection of outliers, and determination of shifts and trends in hydrological time series. Such problems are examined here as particular applications of the single methodology developed. A general mathematical representation of the data exploration problem is initially proposed, based on set theory considerations. Several statistical tests as well as auxiliary information of physical conditions are systematically combined so as to form an objective function to be optimised. This objective function represents the performance of a solution, (where a solution is a specific partitioning of a data set into subperiods), in a manner that in each subperiod a single relationship among data values holds. It is shown that an exhaustive search of all candidate solutions is intractable. Therefore, a heuristic algorithm is proposed, which emulates the exploratory data analysis of the human expert. This algorithm encodes a number of search strategies in a pattern directed computer program and results in an automatic determination of a satisfactory solution.

Keywords. Hydrological data processing, Homogeneity of hydrological data, Consistency of hydrological data, River stage-discharge relationships, Exploratory data analysis, Statistical tests, Heuristic algorithms, Pattern directed programming, Blackboard systems.

1. Introduction

Hydrological data processing is a high-volume task with great importance. Several teams are usually involved in this processing, such as practitioners, engineering consultants and managers, and in some cases researchers. The typical methodology used includes a wide range of tools, from simple empirical techniques to sophisticated statistical analysis. The task, although routine, includes complicated problems where decision-making is necessary, and the personal knowledge and experience of the analyst or decision-maker affects the results.

The main difficulties in data processing are caused by erroneous or spurious measurements and by shifts or changes, either in the measuring process or in the physical process itself. Our purpose is to develop a general methodology applicable to several phases of the data processing, which will tackle such problems in a systematic and intentionally automatic manner. Such a methodology must be capable of partitioning a hydrological data series into subperiods and detecting outliers, in a way that no spurious measurements neither changes in the physical or measuring process appear within each subperiod. To this aim any relationship or law imposed by the data values must be employed in order to identify the subperiods. In some cases the determination of such a relationship is one of the demands of the problem solution. In other cases the relationship is used only for facilitating the partitioning of the data set and the detection of outliers. The methodology also incorporates any additional information (apart from the data set itself), which is concerned either with the measuring process, the physical process measured, or the environment of this physical process.

To clarify the scope and the objectives of the developed methodology we will start with some examples that will be later used for the detailed application of the method. As a first example we consider the data set consisting of simultaneous measurements of stage and discharge in a typical Mediterranean river (Sakoulevas river in Western Macedonia, Greece) shown in Figure 1(a). In this case we wish to determine the stage-discharge relationship using the data set. Apparently however, there is no clear single relationship between stage and discharge for the entire data set of Figure 1(a), although the river hydraulics impose such a relationship. This is due to changes in the geometry or the roughness of the river bed. In Figure 1(b) we have partitioned the data in three subperiods and excluded some measurements

judged as outliers. (Later we will explain how these outliers were identified). Now there appear three well-established relationships, which can be described by fitting a parametric law (in our example, a power law) to the data of each subperiod. In Figure 1(c) we show the power law for the first subperiod, determined by the least squares method, along with the data values of that subperiod and the outliers, i.e., the data values that depart from the curve significantly. Obviously, the fitting of the law is a very easy task. What is complicated here is the step from Figure 1(a) to Figure 1(b), i.e., the partitioning of the data set into subperiods and the detection of outliers. To this aim we must consider any available information about the river properties as well as statistical analysis of the data.

Our second example is concerned with the traditional double mass analysis for testing the consistency of measured values of rainfall data (see, e.g., Linsley et al., 1975, p. 81). In Figure 2 we show the double mass diagram for two rain gauges in Sterea Hellas, Central Greece (Gregorio and Poros Reganiou). It is very simple to detect in that diagram a change in the slope of the double mass curve, indicating a systematic change in the precipitation measuring regime at one of the two stations. Somehow more difficult is to detect the two outliers indicating spurious measurements during two years. As we will see below, an automatic and objective computer-based procedure is not as easy as it may seem from the geometrical interpretation of Figure 2. As in the previous example, the problem here is to partition the data set in an unknown number of subperiods and detect the outliers such that the cumulative rain depths of the two stations in each subperiod obey a simple law of proportionality. The determination of the law, i.e., the proportionality constant, is also demanded in order to adjust the data values.

Our third example is similar to the second in the sense that it is referred to rainfall measurements at two stations, but, in addition, in this case we are also interested on the temporal consistency of the measurements. More specifically, we consider short-scale rainfall data and we try to detect possible shifts in the time recording of each measurement. Such a shift is often caused by a malfunction of the clock device of the instrument or a fault in the manual filling of the form (the observer may write a day's measurement in a wrong position of the form and then may continue to transfer this error for several subsequent days). The

situation is depicted in Figure 3 for one year's daily measurements of two rain gauges in Sterea Hellas, Central Greece (Poros Reganiou and Anthero). We observe that there exists a subperiod with a left shift and one with a right shift in the time recording, whereas there exist three subperiods with time consistent data. We also observe two cases where the rainfall depths of the two stations are in great disagreement. Again, our problem is to partition the data series into consistent subperiods and to detect any faulty measurements. To this aim we can assume a linear relationship between the two rainfall series. This relationship is not a demand of the problem solution; however it assists the partitioning. The details of the example and the relationship will be discussed later in Section 5.

In all three examples discussed above we have data values which can be described by three coordinates (t, x, y) where t denotes time, x denotes a reference variable (stage in the first example, rainfall depth of a reference station in the other examples) and y denotes a tested variable (discharge in the first example, rainfall depth of the tested station in the other examples). We use the time coordinate for ordering the data set only. The coordinates x and y we use to identify a bivariate relationship

$$y = g(x) \tag{1}$$

between them, which may change from subperiod to subperiod. Our problem is to identify a partition of the data set whose subperiods (blocks) correspond to particular instances of the relationship (1). We must emphasise that, given a specific subperiod with its data set, the identification of (1) is a rather trivial issue. What is nontrivial is the identification of the appropriate (consistent) subperiods. The relationship (1) itself may or may not be required as an output of the problem examined but it is necessary for the identification of subperiods. This notion may be straightforwardly extended for problems involving more than one reference variables, i.e., x_1, x_2, …, in which case the relationship among (x_1, x_2, …, y) is multivariate. It can also be applied in the absence of a specific reference variable, in which case the time t plays the role of the reference variable. The latter case we meet for example in identification of nonstationarities (e.g., linear trends or shifting levels) in hydrological data series.

All the above problems are well known in hydrology and several methods are established to solve them. The solution of each of them involves the use of various tools, which are usually chosen and applied in an organised manner by an experienced scientist or practitioner. Such tools include numerous statistical tests, which assist decision making. An authoritative presentation of such tests is given by Hirsch et al. (1993). Specifically, for the identification of outliers tests based on the departure of an observation from the mean, standardised by the standard deviation (Stedinger et al., 1993, p. 18.45), or based on the influence of an observation (Cook, 1977, also quoted by Hirsch et al., 1993, p. 17.45) are typically used; a number of tools of this type can be found in Barnett and Lewis (1994). For detecting abrupt shifts of some property of the data set the rank-sum test and the two-sample *t* test (Hirsch et al., 1992, pp. 17.21-17.23) may be used. Such tests usually need a prior judgement of the point where the shift takes place, which is often done empirically. However, there are objective methods that can detect the location of the shift by means of a test statistic. We mention here the method by Worsley (1983) for testing linearly related values. In the overall enterprise various graphical displays and other tools known as exploratory data analysis (EDA) are very helpful (Velleman and Hoaglin, 1981; Hoaglin et al., 1983; Hirsch et al., 1993). Apart from statistical tests, other procedures that incorporate knowledge of actual physical conditions of the process examined are used. Recent developments tend to use Decision Support Systems to organise this kind of knowledge; we mention here the works by Gawne and Simonovic (1994) and DeGagne et al., (1996) for the establishment of stage-discharge curves.

In the introductory presentation of the three examples we have tried to outline their similarities, although they are concerned with different problems, which are typically treated by different methods. These similarities allow us to formulate a general methodology to tackle all those problems using a single algorithm. The algorithm we propose can emulate the human reasoning process and result in an automatic determination of a satisfactory solution of problems involving partitioning of data sets into subperiods, identification of shifting laws and detection of outliers. The solution is determined by optimising an appropriate objective function, which incorporates statistical indexes of the data set as well as other information

available for the physical process. The optimisation is performed using certain artificial intelligence techniques. Although the algorithm can incorporate additional information on the physical conditions, it can also operate in a black-box way, i.e., using the information contained in the data set only. This feature of the algorithm is useful in situations where additional information is not available or it is very complicated to extract and use it. In such situations, which in practice are very common, the reliability of results is decreased as no mathematical model can replace perfectly the missing information. However the ability to perform in those situations, which evidently are more difficult to tackle, is a strong point of the proposed algorithm.

The text is organised in six sections including this introductory section. In Section 2 we give the general description and the mathematical representation of the problem and, also, we determine the state space of the problem. In Section 3 we combine several statistical and conceptual tests to build a single measure of the degree of "consistency" of a data set, which can be used thereafter for optimisation. In Section 4 we present the developed heuristic algorithm for the optimisation. In Section 5 we present details on the different applications of the method to the above-described real-world hydrological problems. Finally, Section 6 is devoted to conclusions and discussion.

2. Problem setting

2.1 Basic notation

In this section we give the mathematical background of the method in a somehow abstract way. To increase readability we refer several times to the stage-discharge example. In section 5 we will refer to the other examples, as well. The background and, consequently, the notation used are mainly based on the set theory (see e.g., Moschovakis, 1994; Stoll, 1979; among others).

Let A denote the finite data set to be explored. To each element a of A there corresponds a list of real numbers $(t(a), x(a), y(a))$, where t denotes time and the other coordinates denote the results of hydrological measurements of two quantities at a station (or the same quantity

measured at two stations). Additionally, information about the quality or the physical condition of the measurement may be available for each a, which is not necessarily quantitative. In our stage-discharge example, the data set A consists of all concurrent measurements of the river stage and discharge. In that case t represents the time of each measurement whereas x and y denote the measured stage and discharge, respectively. As already mentioned, the list of coordinates may be expanding by including more than one reference variables, i.e., $x_1, x_2, \ldots$, or contracted if $x \equiv t$. We assume that (a) any two distinct elements of A cannot have the same coordinate t, and (b) the coordinate t defines a linear order in the set A; thus, for any two elements a and b of A we denote $a < b$ if $t(a) < t(b)$. This ordering property is transferred to any subset of A.

Let Q be the subset of A containing the elements that are judged as outliers. Later we will explain how a set of this type can be constructed. Let also D be the complement of Q in A, i.e., $D = A - Q$. We will focus our interest in the so called intervals δ of D, which are subsets of D containing a certain number of consecutive ordered elements, i.e., they have the property that any two consecutive elements in δ are also consecutive in D (but not necessarily consecutive in A). A geometric explanation of this property is depicted in Figure 4. The set of all possible intervals of D, which is a subset of the powerset $\mathcal{P}(D)$, will be denoted by $\mathcal{G}_\delta(D)$. In the latter notation the index δ indicates the type of the set's elements (in this case intervals) and the argument D indicates that these elements are intervals in D.

A set $\Delta = \{\delta_1, \delta_2, \ldots, \delta_n\}$ is called an ordered partition of D if its members are disjoint intervals and their union is D, i.e.,

$$\begin{gathered}(\forall\ i \mid i = 1, \ldots, n)\ [\delta_i \in \mathcal{G}_\delta(D)] \\ (\forall\ i, j \mid i, j = 1, \ldots, n \text{ and } i \neq j)\ [\delta_i \cap \delta_j = \varnothing] \\ \delta_1 \cup \ldots \cup \delta_n = D\end{gathered} \tag{2}$$

The intervals $\delta_1, \delta_2, \ldots, \delta_n$ are called blocks of the partition Δ. The set of all possible partitions Δ of D will be denoted by $\mathcal{G}_\Delta(D)$. We can easily expand the ordering property of A into Δ. Furthermore, we will say that two blocks δ_i and δ_j are adjacent iff $\delta_i \cap \delta_j = \varnothing$ and $\delta_i \cup \delta_j$ is an interval in D.

2.2 The exploration process

We will say that the elements of a block δ are consistent if there exists statistical and/or conceptual evidence that they all obey a single law of the type $y = g(x)$. We will call outliers the elements of A that cannot belong to any block with consistent elements. Thus, an outlier a, which is located in the subperiod covered by the block δ, departs significantly from the law $y = g(x)$; if we attempt to modify δ to include the element a, then the elements of δ will no longer be consistent.

Let us assume that we have no additional information about the data set apart from the values (t, x, y) of its elements. So, when we start the exploration procedure the only known is the data set A. We do not know if there are outliers in this set. Also, we do not know whether the data values are consistent along the whole period covered by the data set or whether they are consistent inside some subperiods only. Thus, the goal of the exploration is to determine a set of outliers $Q \subseteq A$ and a set of intervals $\Delta = \{\delta_i\}$ of $D = A - Q$, each referring to a specific time period, such that the data values in each δ_i are consistent. Following the above notation, a solution of our problem will be a specific couple $S = (\Delta, Q)$. The state space of the problem is the set of all possible S, which will be denoted by $\mathcal{G}_S(A)$. Note that to construct $\mathcal{G}_S(A)$ we need to find (a) all possible sets Q (i.e., all elements of the powerset $\mathcal{P}(A)$), and (b) all partitions Δ of any set $D = A - Q$.

To find the solution S of the problem we need tools to assess the consistency for each S. First, we need a two-valued function that for every $S \in \mathcal{G}_S(A)$ returns 1 if all blocks of Δ have consistent elements, and 0 otherwise. This can be constructed by invoking certain aspects of hypothesis testing theory. However, it is practically certain that this function will return 1 for many different S. Thus, we will also need a procedure to choose the best solution. This can be based on another function, which to each Δ assigns a real number expressing a measure of the goodness of a solution. This can be done by combining measures of the goodness of fit of the relationships $y = g(x)$ for the different blocks. Having this function we can compare two solutions and select the best. Expanding this thinking further, we can pose several criteria that assess the goodness of a solution, and thus introduce a vector function

$$\mathbf{f}: \mathcal{G}_S(A) \rightarrow \mathcal{R}^{\mu} \tag{3}$$

where $\mathcal{R}$ is the set of real numbers and μ is the dimension of the vector, i.e., the number of criteria used. The function **f** will be referred to as the objective function. In that case the "best" solution S is the outcome of a multiobjective optimisation problem with a finite state space. A radical simplification of the problem is to combine the different criteria in a single real function f (i.e., with $\mu = 1$) thus reducing the problem to single objective optimisation. The construction of a specific real objective function by combining several statistical and conceptual criteria is studied in detail in Section 3.

2.3 Size of the state space

As the state space of the problem is finite, our problem enters the family of discrete optimisation problems, which theoretically can be solved by an exhaustive search. This method constructs all possible solutions S, one by one, and assesses the goodness of each solution applying the function **f**. To evaluate the efficiency of this method we need to know the number of elements $|\mathcal{G}_S(A)|$ of the state space. (We will use the symbol |.| throughout this text to denote the number of elements, else known as cardinality, rather than the absolute value; the latter we denote by abs(.)). This can be done as follows.

Let D_N be a specific subset of A with number of elements $|D_N| = N$, and $Q = A - D_N$. By removing the greatest element d_N of D_N we construct a set D_{N-1} with $N - 1$ elements. Let also $\mathcal{G}_\Delta(D_N)$ and $\mathcal{G}_\Delta(D_{N-1})$ be the sets of all possible partitions of D_N and D_{N-1}, respectively. If we know $\mathcal{G}_\Delta(D_{N-1})$ we can construct $\mathcal{G}_\Delta(D_N)$ by modifying each of its elements Δ so as to include d_N. Because the partitions are ordered and d_N is greater than all $d_1, \ldots, d_{N-1}$, there are only two ways to modify Δ: we can either insert d_N to the last block of the partition, or add a new block to the partition, i.e., the block $\{d_N\}$, without changing the other blocks. Consequently,

$$|\mathcal{G}_\Delta(D_N)| = 2\,|\mathcal{G}_\Delta(D_{N-1})| \tag{4}$$

and since $|\mathcal{G}_\Delta(D_1)| = 1$, we conclude that

$$|\mathcal{G}_\Delta(D_N)| = 2^{N-1} \tag{5}$$

Furthermore, if $|A| = M$, there can be $\binom{M}{N}$ different subsets D_N of A, all with N elements. Thus,

$$|\mathcal{G}_S(A)| = 1 + \sum_{N=1}^{M} \binom{M}{N} 2^{N-1} \quad (6)$$

where the term 1 stands for the case that all elements of A are outliers. After algebraic manipulations (6) becomes

$$|\mathcal{G}_S(A)| = \frac{1 + 3^M}{2} \quad (7)$$

We can observe that this number is extremely large even for a relatively low M. Thus, it is practically impossible to use the exhaustive method. For example, for a data set containing 100 values we have $|\mathcal{G}_S(A)| = 2.6 \times 10^{47}$. Assuming that an exhaustive algorithm can construct and assess 10^6 solutions per second (an extremely optimistic estimate of speed) it would take 10^{34} years to complete its execution. Consequently, we need an effective optimisation algorithm that must reduce dramatically the number of searches. Such heuristic algorithms like A* or hill-climbing are widely used in artificial intelligence (see Pearl, 1992 for an authoritative analysis of heuristic algorithms). In this study we propose an algorithm implemented as a pattern-directed computer program, which will be presented in detail in Section 4.

3. Consistency assessment

3.1 The objective function

In our search for the best solution $S = (\Delta, Q)$ of the exploration problem our main criteria are the following.

1. There should be statistical and/or conceptual evidence that the data values of each block δ of Δ are consistent.
2. There should be a measure of the consistency of the data associated with the set Δ, based on statistical criteria, which we wish to be as large as possible.
3. The number of outliers $|Q|$ should be as small as possible, given that we wish to incorporate in our solution as many data values as possible.
4. The number of blocks $|\Delta|$ should be as small as possible, given that we do not wish to spread the data in many subperiods (we rather wish to have all data in one period if possible).

To quantify the criterion 1 we introduce a two-valued function

$$h_e: \mathcal{G}_\delta(D) \to \{0, 1\} \qquad \forall\, D \subseteq A \tag{8}$$

which acts on each block δ_i of Δ and returns zero if there is evidence that the elements are not consistent (for example if there exists an outlier in δ_i) and unity otherwise. The evidence of consistency is provided by the execution of several appropriate statistical or conceptual tests based on the available data set, also considering any available additional information. In each statistical test the null hypothesis defends in some manner the data consistency. If the outcomes of all tests are favourable for data consistency then $h_e(\delta_i) = 1$. If even one outcome is against data consistency then $h_e(\delta_i) = 0$. Such tests are presented in detail in Subsection 3.3. To combine the various $h_e(\delta_i)$ for all i we multiply them thus getting a function applying on Δ, i.e.,

$$H_e: \mathcal{G}_\Delta(D) \to \{0, 1\} \qquad \forall\, D \subseteq A \tag{9}$$

where

$$H_e(\Delta) = \prod_{\delta_i \in \Delta} h_e(\delta_i) \tag{10}$$

Apparently, $H_e(\Delta) = 1$ if the outcomes of all tests at all blocks δ_i of Δ are favourable for data consistency, otherwise $H_e(\Delta) = 0$. The functions h_e and H_e will be referred to as the block consistency evidence and the partition consistency evidence, respectively.

We can quantify the criterion 2 in a similar manner. To each block δ_i of Δ we assign a positive real number $h_m(\delta_i)$, which expresses some measure of consistency of the block's data. The function

$$h_m: \mathcal{G}_\delta(D) \to \mathcal{R}^+ \qquad \forall\, D \subseteq A \tag{11}$$

will be referred to as the block consistency measure. Then we combine the measures $h_m(\delta_i)$ for all i by multiplying them thus getting the partition consistency measure for Δ. This defines the function

$$H_m: \mathcal{G}_\Delta(D) \to \mathcal{R}^+ \qquad \forall\, D \subseteq A \tag{12}$$

where

$$H_{\mathrm{m}}(\Delta) = \prod_{\delta_i \in \Delta} h_{\mathrm{m}}(\delta_i) \tag{13}$$

In order for the multiplication in (13) to be well justified we can define h_{m} and H_{m} in terms of probability of some events, in which case

$$0 \le h_{\mathrm{m}}(\delta),\ H_{\mathrm{m}}(\Delta) \le 1 \tag{14}$$

Specifically, in each block δ_i we consider an event ε_i associated with the consistency of the data of that block so that the events of different blocks be independent. Then we can define

$$h_{\mathrm{m}}(\delta_i) = \Pr(\varepsilon_i) \qquad H_{\mathrm{m}}(\Delta) = \Pr\left(\bigcap_{i=1}^{|\Delta|} \varepsilon_i\right) \tag{15}$$

where Pr(.) denotes probability. Owing to the independence of ε_i, (13) is compatible with (15). The construction of the events ε_i and the determination of the consistency measure functions will be discussed in Subsection 3.5.

Furthermore, we can combine the consistency evidence functions and the consistency measure functions to get two uniting functions, which will be referred to as the consistency assessment functions (for block and partition, respectively):

$$h(\delta) = h_{\mathrm{e}}(\delta)\, h_{\mathrm{m}}(\delta), \qquad H(\Delta) = H_{\mathrm{e}}(\Delta)\, H_{\mathrm{m}}(\Delta) \tag{16}$$

We must note that the definition of functions H_{e} and H_{m} is somehow arbitrary and subjective. However, this may not be a critical issue, as a wide range of trial functions that have some justified general properties and an appropriate set of parameters (which is more important than the function type) gave the same results (i.e., the same final solution S) in several experimental applications of the method.

The quantification of the remaining two objectives is very simple and is given by the following equation, which combines all four objectives in a four-dimensional objective function:

$$\mathbf{f}(\Delta, Q) = \begin{bmatrix} H_{\mathrm{e}}(\Delta) \\ H_{\mathrm{m}}(\Delta) \\ |A - Q| \\ |A| - |\Delta| \end{bmatrix} \tag{17}$$

We wish to maximise all the four components of this function, noting that the first component is two-valued and the second is highly nonlinear. We observe that the second and fourth components of the function usually are not competing. As a result of the definition of H_m in (13) as product of h_m terms that normally are less than unity, the smaller is the number of blocks $|\Delta|$ the bigger are the values of both $|A| - |\Delta|$ and $H_m(\Delta)$. Thus, we can ignore the fourth component. Furthermore we can combine the remaining three components in a single real function by introducing priorities to each one, thus reducing the problem to single objective optimisation. To this aim, we give first priority to the consistency evidence function, second priority to the number of outliers and third priority to the consistency measure. Observing that the first component is two-valued (0 or 1), the third is integer, and the second is a real number between 0 and 1, we formulate our final single objective function as

$$f(\Delta, Q) = H_e(\Delta) \left[|A - Q| + H_m(\Delta) \right] \tag{18}$$

which is compatible with the priorities that we have set. (For example, among two solutions with $H_e(\Delta) \neq 0$, the solution with the fewer outliers always results in greater value of $f(\Delta, Q)$ regardless of the value of $H_m(\Delta)$).

3.2 Construction of the block consistency evidence function

As we have stated earlier, the block consistency evidence function is a two-valued function that combines the results of several statistical and conceptual tests applied to the data of each block. Thus, it can be written as

$$h_e(\delta) = h_e^1(\delta)\, h_e^2(\delta)\, h_e^3(\delta) \cdots \tag{19}$$

where each of the components $h_e^l(\delta)$ is one if the outcome of the relevant test defends the data consistency and zero otherwise. In the next subsections 3.3 and 3.4 we describe some representative tests that we have explored and found to have good performance in the model applications in real-world hydrological problems. To increase readability the reader is prompted to have in mind the stage-discharge example, where the consistency of a block's data is expressed in terms of the goodness of fit of the relationship

$$y = g(x) = c\, x^d \tag{20}$$

to the data values, where y denotes discharge, x denotes stage, and c and d are parameters statistically estimated from the data values of the specific block. The equivalent log-log linear relationship, which has the advantage of making residuals homoscedastic, may also be considered.

3.3 Statistical components of the consistency evidence function

We emphasise that no single statistical test can describe entirely what we mean with the term consistency in this context. For example, if we merely use a test based on the determination coefficient of (20), this would not capture an outlier in the block examined, as that outlier may not affect seriously the determination coefficient. Thus, we need a combination of statistical tests. Below we give a collection of the most important tests and the resulting functions $h_e^l(\delta)$. The collection may be expanded to include other tests suitable for specific problems. Also, not all tests of the given collection are necessary for any problem; the developed system has the ability to enable or disable each of the tests.

Determination coefficient. Let e_j be the departure of the data element d_j, from the assumed law of interval δ, i.e., $e_j = y_j - \hat{y}_j$ where $\hat{y}_j = g(x_j)$. The determination coefficient is

$$r^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2} \tag{21}$$

where σ_e^2 and σ_y^2 are the variances of e and y respectively, estimated from all data of interval δ. Let ρ^2 be the corresponding population parameter and ρ_0^2 be a reference value greater than zero (e.g., $\rho_0^2 = 0.9$). We test the null hypothesis H_0: $\rho^2 = \rho_0^2$ against the alternative hypothesis H_1: $\rho^2 < \rho_0^2$ at a significance level α_1 (e.g., $\alpha_1 = 0.05$). If the null hypothesis is rejected then we set the corresponding consistency evidence function equal to zero, otherwise we set it equal to unity. (Notice the difference from the typical correlation coefficient test where the null hypothesis is $\rho = 0$; also notice that the null hypothesis, as set above, is favourable for the data consistency, whereas in the typical correlation coefficient test is not.) We use the Fisher transformation for r, i.e., $\zeta(r) = (1/2) \ln [(1 + r) / (1 - r)]$ and assume that R^2, the random variable whose estimate is r^2, is the squared inverse Fisher transformation of an approximately normally distributed variable, whose mean is $\zeta(\rho_0)$ and standard deviation is

$$\sigma_\zeta^2 \approx \frac{1}{|\delta| - 3} \tag{22}$$

where $|\delta|$ is the number of elements of δ. With these assumptions the resulting consistency evidence function is

$$h_{\mathrm{e}}^1(\delta) = \begin{cases} 0 & \alpha(r^2) < \alpha_1 \\ 1 & \text{otherwise} \end{cases} \tag{23}$$

where $\alpha(r^2)$ denotes the attained significance level for the observed r^2; it can be shown (D. Koutsoyiannis, unpublished document, 1994) that this is given by

$$\alpha(r^2) = F_{\mathrm{G}}\left(\frac{\zeta(r) - \zeta(\rho_0)}{\sigma_\zeta}\right) - F_{\mathrm{G}}\left(\frac{-\zeta(r) - \zeta(\rho_0)}{\sigma_\zeta}\right) \tag{24}$$

In the last equation r and ρ_0 are the positive square roots of r^2 and ρ_0^2, respectively, and F_{G} denotes the standardised normal distribution function.

In this test we have two parameters that must be selected by the user: the reference value ρ_0 and the significance level α_1. We note that the alternative hypothesis H_1 has no meaning if the reference value ρ_0 equals zero. To avoid numerical effects for very low values of ρ_0 we suggest a minimum value of $\rho_0 = 1.3\ \alpha_1$ (which was found by numerical investigation). For example for $\alpha_1 = 0.05$ the minimum value is $\rho_0 = 0.065$. This implies no practical limitation as for most problems the usual values of ρ_0 are greater than 0.9 (see examples of Section 5).

Standardised residual. The second component accounts for extraordinary departures of the data values from the law $y = g(x)$ and thus assists for the identification of outliers. It depends on the maximum absolute residual e_j standardised by the standard error σ_e and is defined by

$$h_{\mathrm{e}}^2(\delta) = \begin{cases} 0 & (\exists\, j \mid j = 1, \ldots, |\delta|)\left[\dfrac{\mathrm{abs}(y_j - \hat{y}_j)}{\sigma_e} > b_2\right] \\ 1 & \text{otherwise} \end{cases} \tag{25}$$

where abs(.) denotes absolute value (we have used this notation to avoid conflict with the symbol for set cardinality) and b_2 is an acceptable threshold corresponding to a significance level α_2. For example, if the residual is normally distributed, we can adopt $b_2 = 2.58$ to characterise as outliers the extreme 1% of the data values. Alternatively we can determine b_2 as a function of $|\delta|$ from the equation by Stedinger et al. (1993, p. 18.45), based on tabulated

data by Interagency Advisory Committee on Water Data (1982). It is clear from (25) that the consistency evidence function equals zero even if only one outlier is contained in δ.

An alternative way to standardise residuals is to divide by the measured value y_j instead of the standard deviation of residuals. This agrees with ISO Standard 1100 for determination of stage-discharge relationships (International Organization for Standardization, 1973), which sets a maximum difference ±20% of the measured discharge from the adopted curve.

Standard deviation of residuals. Sometimes, it is desirable to keep the standard deviation of residuals below a certain level σ_0, a practice known in the quality control theory as control of the standard deviation (Papoulis, 1990, pp. 342-347). This leads to the one-sided test H_0: $\sigma_e = \sigma_0$ against H_1: $\sigma_e > \sigma_0$. From the theory of statistics we obtain the following component of the consistency evidence function

$$h_e^3(\delta) = \begin{cases} 0 & \sigma_e > b_3 \\ 1 & \text{otherwise} \end{cases} \tag{26}$$

where

$$b_3 = \sigma_0 \sqrt{\frac{\chi^2_{1-\alpha_3}(|\delta| - 1)}{|\delta| - 1}} \tag{27}$$

in which α_3 is the adopted significance level for this test, and the numerator under the root denotes the $(1 - \alpha_3)$-quantile of the χ^2 distribution for $(|\delta| - 1)$ degrees of freedom.

Runs. The fourth component accounts for the existence of unusual patterns within δ. An unusual pattern is a sub-interval δ_p of δ, also termed run, that consists of a considerable number of neighbouring points whose residuals (i.e., departures from the law $y = g(x)$) are all positive or all negative. The existence of such a pattern within δ may imply another separate law for the sub-interval δ_p. Such a pattern is captured by

Error! (28)

where b_4 is a critical value. Using the theory of runs (Hald, 1965, pp. 342-353) it can be shown (D. Koutsoyiannis, unpublished document, 1994) that

$$b_4 \approx \beta \ln |\delta| + \gamma \tag{29}$$

where

$$\beta = \frac{11}{\ln(-2^{13}\ln(1-\alpha_4))}\ , \qquad \gamma = 1 \qquad \text{if } |\delta| < -2^{13}\ln(1-\alpha_4)$$
$$\beta = \frac{1}{\ln 2}\ , \qquad \gamma = \frac{-\ln(-2\ln(1-\alpha_4))}{\ln 2} \qquad \text{otherwise} \tag{30}$$

whereas α_4 is the confidence level of the relevant statistical test. The critical length obtained by the above equations for $\alpha = 0.05$ and $|\delta| = 10$, 30 and 100 is 5, 7 and 9 respectively. Note for comparison that in a similar situation concerning the construction of stage-discharge curves the ISO Standard 1100 (International Organization for Standardization, 1973) suggests a critical number of consecutive points above or below the curve equal to 7, regardless of the total number of points defining the curve. Apparently however, the critical length given by (29) as a function of $|\delta|$ is more reasonable.

Marginal outliers. Essentially, all the previous components assess the goodness of fit of the law $y = g(x)$ to the data. This component accounts for marginal outliers, i.e., based on merely each one of the variables x or y.

$$h_e^5(\delta) = \begin{cases} 0 & (\exists j \,|\, j = 1, \ldots, |\delta|)\left[\dfrac{\text{abs}(x_j - \overline{x})}{\sigma_x} > b_5 \text{ or } \dfrac{\text{abs}(y_j - \overline{y})}{\sigma_y} > b_5\right] \\ 1 & \text{otherwise} \end{cases} \tag{31}$$

where b_5 is a critical value corresponding to a significance level α_5. In the above equation $\overline{x}$ and σ_x are the mean and standard deviation of x, respectively, whereas $\overline{y}$ and σ_y are the mean and standard deviation of y. We note that this test, besides its marginal action on each variable separately, is also important for the test of the goodness of fit of the law $y = g(x)$. To understand this consider the case that δ contains some data values that form an irregular concentrated group of points (x, y) plus a point (x_f, y_f) which is a far outlier with respect to both x and y. It is very probable that the far outlier can lead to the faulty conclusion that there exists some law $y = g(x)$, despite of the fact that most of the points are spread irregularly. This case is surely captured by the component h_e^5 while it may not be traced by the previous components as the departures for the hypothetical law may be low.

3.4 Conceptual components

Apart from statistical tools the data analysts use practical rules and additional information about the data set. This additional information is employed in our comprehensive system in three ways. First, we use it in the formulation of the consistency evidence function, in order to form some conceptual tests formally similar with statistical tests. Such tests are described below in this subsection. Second, we utilise it in the search strategy for the best solution, as it will be discussed in Section 4. Third, the experienced analyser may alter the final solution obtained automatically by the system, as it is not obligatory for the user to adopt that solution; rather the interactive form of the system allows the user to modify the solution $S = (\Delta, Q)$ and ask the system to best fit a specific law to the modified solution, or even propose his or her own fit.

Number of elements of blocks. This component, $h_e^6(\delta)$, guarantees a minimum number of elements of δ and it is defined by

$$h_e^6(\delta) = \begin{cases} 0 & |\delta| < b_6 \\ 1 & \text{otherwise} \end{cases} \tag{32}$$

where b_6 is a critical value. ISO Standard 1100 for determination of stage-discharge relationships (International Organization for Standardization, 1973) sets a lower limit $b_6 = 6$ for the number of measurements that define a valid stage-discharge curve.

Known outliers. Additional information about the measurements may lead us to characterise as outlier a given point, whose physical conditions depart from normal. For example, stage-discharge measurements taken during periods when the river is iced, do not obey to any particular law and thus are characterised as outliers. Thus, before we apply any statistical procedure we may form a set Q_o containing such points, which must be excluded from the blocks of any solution. The component of the consistency evidence function that prevents these points to be included in a block is

$$h_e^7(\delta) = \begin{cases} 0 & \delta \bigcap Q_o \neq \varnothing \\ 1 & \text{otherwise} \end{cases} \tag{33}$$

Known breakpoints. In many situations it is known that the physical conditions of the measured processes have changed at a certain time. For example, an abrupt change in the

cross section geometry of the river channel, caused by a major flood, leads to a permanent shift of the stage-discharge relationship. Thus, if t^* is the (known) time that such an event has been occurred, then we should not have data points before and after t^* belonging to the same block. This is captured by the component $h_e^8(\delta)$ of the consistency evidence function, which can be formally written as

$$h_e^7(\delta) = \begin{cases} 0 & (\exists\, a, b \in \delta)\ [t(a) < t^* \le t(b)] \\ 1 & \text{otherwise} \end{cases} \tag{34}$$

The previous two components account for additional information that induces particular decisions about the final solution. To take such actions the analyser must be certain about the validity of this information. However, there are cases of "suspect" measurements or "likely" breakpoints, where the analyser cannot be certain in taking actions such as excluding data points from the final solution or setting a breakpoint at a certain time. For example, a stage-discharge measurement taken during a flood may not obey the law of other measurements referring to more steady flow conditions. In such cases the use of components (33) and (34) is not appropriate, because we cannot exclude such a data point from the final solution without testing it first. However, we can take advantage of this kind "fuzzy" information in the search strategy of the final solution, as it is explained in Section 4.

3.5 Construction of the block consistency measure function

As we have stated in Subsection 3.1 the block consistency measure is defined as the probability of an event ε_i associated with the data values of interval δ_i. This event must be related with the consistency of the data in a manner that its probability be close to 1 for an apparently consistent data set and close to zero for an evidently inconsistent data set. Such an event can be defined in terms of the test statistics discussed in Subsection 3.3. Since the various statistics examined are apparently jointly dependent, it is very difficult to describe our event by more than one statistics. Thus, we must choose one of them, the more representative for the specific problem studied. Without loss of generality we will choose the determination coefficient R^2. This was found empirically to be the most appropriate for the studied

examples, which can be theoretically justified by its feature to be an overall measure of the goodness of fit. An event with the desired properties is

$$\varepsilon = \{R^2 < r^2\} \tag{35}$$

The resulting block consistency measure is then

$$h_{\mathrm{m}}(\delta) = \Pr(R^2 < r^2) = \alpha(r^2) \tag{36}$$

where $\alpha(r^2)$ is the attained significance level of the test, given by (24). Alternatively, we can define the consistency measure by means of the conditional probability $\Pr(R^2 < r^2 \mid R^2 < \rho_{\mathrm{c}}^2)$ where ρ_{c}^2 is the critical value of R^2 for the relevant test described in Subsection 3.3 at the significance level α_1. In this case the consistency measure takes the form

$$h_{\mathrm{m}}(\delta) = \begin{cases} 0 & \alpha(r^2) \le \alpha_1 \\ \dfrac{\alpha(r^2) - \alpha_1}{1 - \alpha_1} & \alpha(r^2) > \alpha_1 \end{cases} \tag{37}$$

which is consistent with $h_{\mathrm{e}}^1(\delta)$ defined in (23), as both $h_{\mathrm{m}}(\delta)$ and $h_{\mathrm{e}}^1(\delta)$ are equal to zero in the critical region of the test. We clarify that $\alpha(r^2)$ that is used in the above equations has already been calculated during the evaluation of $h_{\mathrm{e}}^1(\delta)$ (i.e., the application of the determination coefficient test), and thus the evaluation of $h_{\mathrm{m}}(\delta)$ needs no additional statistical calculation.

4. The proposed heuristic algorithm

4.1 Human reasoning strategies and state space reduction

The actions of a human expert trying to explore the consistency of a given data set can be represented by a process of successive transformations from a problem state $S = (\Delta, Q)$ to a new state $S' = (\Delta', Q')$, i.e.,

$$(\Delta', Q') \leftarrow (\Delta, Q) \tag{38}$$

This step-by-step search is obviously not exhaustive as not all possible problem states are evaluated. The expert always selects the most promising new state S', thus heuristically guiding the search towards the "best" solution. There are some decisions associated with the selection of the direction of the search, i.e., the selection of the most promising new problem state. These decisions are supported by the use of the statistical tools described in previous

section together with the expert's experience and intuition. The junction of scientific knowledge, experience and intuition can be viewed as heuristic rules or search strategies that the expert employs in order to guide the search towards the most promising direction.

We will now assume that a data set A is given to a human expert and that he or she tries to separate it into consistent intervals and to isolate the outliers of the set, i.e., he or she searches for a solution (Δ, Q) of the problem. We will trace in the following paragraphs some of the most preferable search strategies, used to select and test the most promising problem states. Such strategies are depicted in the schematic representation of Figure 5, in the form of a simplified example indicating the consecutive steps of the search process. In the initial step L^1 the data set is totally unexplored, while in the final step L^8 the final solution has been found. The strategies include:

1. Opening of *windows* in the data set, in order to locate consistent intervals (Figure 5, steps L^2 and L^6).
2. *Expansion* of a consistent interval integrating elements adjacent to this interval that do not belong to any other consistent interval but rather lie in between the already found consistent intervals (Figure 5, steps L^3, L^5 and L^7).
3. *Merging* of adjacent consistent intervals when their union is judged consistent (Figure 5, step L^6).
4. *Shifting* of the boundaries of adjacent consistent intervals that cannot be merged, by exchanging elements between them (Figure 5, step L^7).
5. *Isolation* of some elements (outliers) belonging to an interval that cannot be judged as consistent (Figure 5, steps L^4 and L^6).
6. *Re-integration* of some elements that were left out as outliers but as the problem state evolves they might be integrated in a recently found consistent interval (Figure 5, step L^8).
7. *Replacement* of isolated elements by other elements of a consistent interval (Figure 5, step L^8).
8. *Reduction* of the problem to distinct sub-problems with lower complexity. This results in arbitrary partitioning of the data set and processing of each interval separately.
9. *Reconnection* of sub-problems. This is the reverse of the reduction procedure.

4.2 Representation scheme

The heuristic algorithm we propose in this study emulates the human reasoning process and thus we will use a similar representation. The various search strategies used by the human expert are represented in the algorithm as numerous transformations. The algorithm, in order to reach a solution state, transforms the current state of the problem to a new state, as described by (38), which is equivalently written as

$$(\Phi', \Pi', Q') \leftarrow (\Phi, \Pi, Q) \tag{39}$$

where Δ is now separated in two disjoint sets Φ and Π, defined by

$$\Phi \cup \Pi = \Delta, \pi \in \Pi \Rightarrow h(\pi) > 0 \tag{40}$$

This means that the set Π holds all members of Δ already found to be consistent, while the set Φ holds the rest members of Δ. These members are either tested and judged inconsistent, or not yet tested.

The initial state of the algorithm is

$$S_I = (\Phi_I, \Pi_I, Q_I) = (\{A\}, \varnothing, \varnothing) \tag{41}$$

Each transformation consists of a *conditions* part and an *action* part. The conditions part of a transformation represents a search of the current state for a pattern on which it can apply its action. The action part represents the modification of the current state that will be performed if the conditions part succeed in the search of the pattern of interest.

The algorithm will terminate when it is no more able to transform the current state. The algorithm can be schematically represented by the following seven steps:

(1) Form the initial state of the problem and assert it as the current state.

(2) If all transformations are marked as *used* then go to step (7).

(3) Select a transformation not marked as *used*.

(4) Search if the conditions of the transformation hold in the context of the current state.

(5) If the conditions do not hold, mark the transformation as *used* and go to step (2). Else continue.

(6) Instantiate the actions of the transformation thus modifying the current state. Remove the *used* mark from all transformations and go to step 2.

(7) Exit with the current state as a solution of the problem.

In the following subsections we describe the specific transformations used by the algorithm to reach the goal state. For each transformation we describe the conditions and the actions part and give some comments.

4.3 Opening of windows

Opening of windows is one of the most popular procedures in signal analysis and pattern classification (Duda and Hart, 1973, pp. 88-95). As an isolated procedure it is not so effective, but when combined with other types of actions it results in a powerful tool. The procedure searches for a block φ of Φ such that at least one subinterval π of φ is consistent. If it succeeds, it appends the interval π to the set Π and modifies appropriately the set Φ without changing the set Q. The conditions that must hold in order for the transformation to instantiate its action are

$$(\exists\, \varphi \in \Phi)\, (\exists\, \pi \subseteq \varphi,\, \pi \in \mathcal{G}_\delta(D))\, [h(\pi) > 0] \tag{42}$$

The "exists" ($\exists$) quantifier may be interpreted as the indicator of an appropriate search. The relation $\pi \in \mathcal{G}_\delta(D)$ indicates that π is an interval in D, where D is the union of all members of both Π and Φ. The result of the transformation is

$$\Phi' = (\Phi - \{\varphi\}) \cup \{\varphi_1, \varphi_2\}, \quad \Pi' = \Pi \cup \{\pi\}, \quad Q' = Q \tag{43}$$

where φ_1 and φ_2 are subintervals of φ containing those members that are smaller and greater, respectively, than the members of π.

The transformation will process every member φ of Φ until it finds one having at least one consistent subinterval π. The search for the existence of such a subinterval begins with a trial window with number of elements $|\pi| = c$ with $c = |\varphi|$. If the algorithm does not find a window π with consistent members, it reduces the number of elements of the search window to $c = c / c_r$ where $1 < c_r < c$. The search is completed when $c < c_{min}$. The parameters c_r and c_{min} are defined by the user. If the parameter c_r is very large then we get a fast but risky search. If the

parameter c_r is near to 1 we get a slow scanning of the set φ. An interesting special case is $\varphi_1 = \varphi_2 = \varnothing$ in which $\pi = \varphi$.

4.4 Expansion of a consistent interval

When a consistent interval (a member π of the set Π) is located in the current state, a natural action would be to expand this consistent interval to its unexplored neighbourhood (a member φ of the set Φ). This is accomplished by cutting a piece φ_1 from the adjacent φ set and merging it with π. The conditions that must hold in the current state are

$$(\exists\, \pi \in \Pi)\ (\exists\, \varphi \in \Phi\ )\ (\exists\, \varphi_1 \subseteq \varphi\ )\ [\pi \cup \varphi_1 \in \mathcal{G}_\delta(D) \text{ and } h(\pi \cup \varphi_1) > 0] \tag{44}$$

where the interval π is adjacent to φ as well as to the subinterval φ_1 of φ. The result of the transformation is

$$\Pi' = (\Pi - \{\pi\}) \cup \{\pi \cup \varphi_1\}, \qquad \Phi' = (\Phi\ - \{\varphi\}) \cup \{\varphi - \varphi_1\}, \quad Q' = Q \tag{45}$$

Many alternative sets φ_1 are tested during this expansion process. Firstly, the algorithm tries to integrate the whole set φ in the consistent interval π. Thus, the number of elements c of the first set φ_1 is equal to that of φ, i.e., $|\varphi_1| = |\varphi|$. If the union of π and φ_1 is not consistent then the number of elements of φ_1 is reduced by one and the new union of π and φ_1 is tested. This reduction is done repeatedly until $c = 1$.

4.5 Merging of adjacent consistent intervals

When two consistent intervals are adjacent, a natural action is to merge them, if possible. The relevant transformation modifies the set Π without affecting the sets Φ and Q. The necessary conditions are

$$(\exists\ \{\pi_1, \pi_2\} \subseteq \Pi)\ [\pi_1 \cup \pi_2 \in \mathcal{G}_\delta(D) \text{ and } h(\pi_1 \cup \pi_2\ ) > h(\pi_1)\ h(\pi_2)] \tag{46}$$

where the intervals π_1 and π_2 are adjacent. The result is

$$\Phi' = \Phi\ , \quad \Pi' = (\Pi - \{\pi_1, \pi_2\}) \cup \{\pi_1 \cup \pi_2\}, \quad Q' = Q \tag{47}$$

The algorithm keeps track of all unsuccessful attempts to merge consistent intervals by inserting appropriate marks in a special space of the computer memory. Thus, the algorithm prevents the application of the transformation under the same conditions (application to the

same consistent intervals). Marks of this type are inserted for other transformations as well, but it is not our purpose to present all details of the particular implementation of the proposed algorithm.

4.6 Shifting of consistent intervals boundaries

In case of a smooth transition from an instance of a law relating the data values to a different instance of the law, the boundary between any two instances may be not well defined. Very often, there is no abrupt change in the data of the two law instances. The data values in the boundary of two consecutive consistent intervals lay in a "gray" area and may belong to either law instance. Thus, we need a transformation that will force the two consistent intervals to exchange elements until a better partitioning is found. Its conditions are

$$(\exists\ \{\pi_1, \pi_2\} \subseteq \Pi)\ (\exists\ \{\pi'_1, \pi'_2\} \subseteq \mathcal{G}_\delta(D))\ [\pi'_1 \cup \pi'_2 = \pi_1 \cup \pi_2 \text{ and } h(\pi'_1)\ h(\pi'_2) > h(\pi_1)\ h(\pi_2) \quad (48)$$

where π_1 and π_2 are adjacent intervals in D, π'_1 and π'_2 are also adjacent intervals in D, and $\pi'_1 \neq \pi_1$. The modification of the current state is

$$\Phi' = \Phi, \quad \Pi' = (\Pi - \{\pi_1, \pi_2\}) \cup \{\pi'_1, \pi'_2\}, \quad Q' = Q \qquad (49)$$

The number of elements c of π'_1 determines the time position where the new intervals are separated. The transformation starts the search with $c = |\pi_1| \pm 1$. If the new partitioning does not improve the solution, the number of elements c becomes $c = |\pi_1| \pm 2$. This scanning goes on until $c = |\pi_1| \pm c_m$ where c_m is defined by the user (e.g., $c_m = 3$).

4.7 Isolation of elements

The detailed mathematical representation of this transformation is somewhat complicated (G. Tsakalias, unpublished document, 1994) and, therefore, we will give a verbal description of the transformation. When an unexplored set φ contains outliers, it cannot be used by any of the former transformations. For example, the transformation that expands a consistent interval cannot expand a set π toward an adjacent set φ if the first element of φ is an outlier. This outlier is an obstacle that must be removed. The transformation takes as input a set φ, isolates a subset z of this set, and returns to the algorithm a "clean" set $\varphi' = \varphi - z$. Thus the expansion transformation (or any other transformation) will now act on the new set φ'. If the isolation of

the outliers is successful (the set z contains the obstacles) then the members of the set φ' will be utilised and the transformation will not act again on the set φ. Otherwise, a different set z will be isolated and the process will stop with the isolation of the entire set φ.

The isolation of some elements does not necessarily mean that these elements will remain isolated in the final solution. They must be considered as temporarily "disabled" because, as the solutions evolve, other transformations may succeed in utilising them again. An exception to this rule is the case where there exists additional, physically-based, information about the measurements, that may lead us to characterise as outlier a given point, whose physical conditions depart from normal. If such outliers are a priori known positively, they can be entirely excluded from the data set. However, if the additional information is no more than a "suspicion" about the correctness of some data points, then these data point are introduced in the solution procedure by assigning them a priority (against other points) for being isolated by the algorithm.

4.8 Re-integration of elements

This transformation is opposite to the previous one, as it attempts to re-integrate temporarily isolated data. It modifies Π and Q without affecting Φ. The necessary conditions are

$$(\exists\, Q_R \subseteq Q)\ (\exists\, \pi \in \Pi\ )\ [\pi \cup Q_R \in \mathcal{G}_\delta(D \cup Q_R) \text{ and } h(\pi \cup Q_R) > 0] \tag{50}$$

The first relation in square brackets states that the new set $\pi \cup Q_R$ must be an interval in $D \cup Q_R$. The above conditions hold in case that the algorithm in a previous stage had isolated some elements (which are contained in the set Q_R) and it is now coming to revise this decision trying to reconnect these elements. The transformation of the current state is

$$\Phi' = \Phi,\quad \Pi' = (\Pi - \{\pi\})\ \cup \{\pi \cup Q_R\},\quad Q' = Q - Q_R, \tag{51}$$

The set Q_R is chosen so that $|Q_R| \le c_R$, where c_R is an integer constant defined by the user and by default set to $c_R = 1$.

4.9 Replacement of elements

This transformation is similar to the previous one, but here the re-integrated elements are not appended to a specific interval but rather they replace some other elements of this interval. The conditions that must hold in order for the transformation to apply are

Error! (52)

The first relation in square brackets states that the new set $(\pi - \pi_E) \cup Q_E$ must be an interval in $(D - \pi_E) \cup Q_E$. These conditions describe a search for the re-integration of some previously isolated elements (members of the set Q_E) that replace an equal number of connected elements (members of the set π_E) inside a consistent interval (π) of the current state. The result of the transformation is

$$\Phi' = \Phi, \quad \Pi' = (\Pi - \{\pi\}) \cup \{(\pi - \pi_E) \cup Q_E\}, \quad Q' = (Q - Q_E) \cup \pi_E \tag{53}$$

The set Q_E is chosen so that $|Q_E| \le c_E$, where c_E is an integer constant defined by the user, which is by default set to $c_E = 1$.

4.10 Problem reduction and reconnection

Very often, a human expert does not consider simultaneously all data values of a large set (e.g., with one thousand or more measurements), but rather splits the data values in "pages" and considers each page separately. As the search evolves, he or she tries to reconnect these pages. To emulate this human strategy we have implemented a transformation that partitions a large interval φ into two subintervals φ_1 and $\varphi - \varphi_1$. The conditions that must hold are obvious:

$$(\exists\, \varphi \in \Phi)\, (\exists\, \varphi_1 \subseteq \varphi)\, [\{\varphi_1, \varphi - \varphi_1\} \subseteq \mathcal{G}_\delta(D)] \tag{54}$$

The relation in square brackets means both φ_1 and $\varphi - \varphi_1$ must be intervals in D. The result is

$$\Phi' = (\Phi - \{\varphi\}) \cup \{\varphi_1, \varphi - \varphi_1\}, \quad \Pi' = \Pi, \quad Q' = Q \tag{55}$$

The number of elements of the set φ_1 is fixed to some integer c_f defined by the user. As the transformation can be applied many times, it is possible that the initial set φ will be finally separated into many intervals. However, this does not mean that these intervals will remain

separated. In fact this is quite unlikely because the separation is done in arbitrary locations. Although this transformation does not affect the quality of the final solution, it affects the speed of the method in case of large data sets.

Another case of problem reduction is met when there exists auxiliary, physically-based information (such as that discussed in the end of subsection 3.4), about specific permanent shifts in the measured hydrological process or its environment of, or even in the measuring process itself. In such a case the problem reduction is permanent and no reconnection is needed. This, apparently, simplifies the problem as its dimensionality is permanently reduced.

4.11 Additional remarks

In addition to the above-described transformations there are some other with complementary action. For example, there is a transformation that collects the "garbage" produced by other transformations.

After the completion of the algorithm, the set Π contains the blocks found to be consistent, while the set Φ contains inconsistent blocks. The elements of Φ are appended to the set of outliers Q, so that the final solution (Δ_F, Q_F) is

$$\Delta_F = \Pi, \qquad Q_F = \bigcup_{\varphi_i \in \Phi} \varphi_i \cup Q \tag{56}$$

There is no guarantee that this solution is the optimum. Due to the extremely complicated determination of the consistency assessment function there can be no theoretical proof that the algorithm reaches the absolutely optimal solution. However, in the experiments we have done we never found manually a solution with higher value of the objective function than that obtained by the algorithm.

We must notice that the algorithm, as described above, does not evaluate directly the objective function $f(\Delta, Q)$ of equation (18) in each step (i.e., in each execution of a transformation). In fact, it evaluates the block consistency assessment function $h(\delta)$ for each interval δ it considers. This simplifies the computations. All transformations applied, increase successively the value of the function $h(\delta)$ of the interval δ that they modify. Consequently, they increase always the value of the objective function $f(\Delta, Q)$. This feature of the algorithm

prevents circular actions evolving a number of transformations. To this rule there is the exception of transformations that do not modify the set Π, which do not alter the value of objective function. For example, the transformation that isolates elements does not affect the value of this function. However, it employs an internal explicit mechanism for preventing circular actions.

4.12 Program implementation

The algorithm presented in this study is implemented using the Prolog language and following the principles of pattern directed programming. In a pattern directed program the procedures are autonomously triggered when patterns of data are found in the memory of the computer. This is the main difference from conventional programs whose procedures are called by other procedures with a standard order. More specifically, the architecture of the implemented program is quite similar to the architecture of the systems called blackboard systems (Hayes-Roth and Hewett, 1988). The transformations are represented as knowledge sources and the current problem state is represented as a blackboard. The sets used by the algorithm are represented as blackboard objects recorded on the blackboard. Each time a knowledge source finds a pattern (i.e., a combination of objects on which it can apply) it instantiates its action, thus transforming the blackboard (i.e., the current state of the problem). A number of techniques that improve the efficiency of the program are employed in order to avoid unnecessary searching, circular transformations and to increase the program speed.

The system is implemented on RISC workstations under the name PINAX and is connected to the HYDROSCOPE relational database (the Greek national data bank for hydrological and meteorological information) (Papakostas et al., 1994), through the OPSIS time series processing and visualising system (Tsakalias and Koutsoyiannis, 1994).

5. Hydrological applications

In this section we will reconsider the examples already discussed in the Introduction, in order to present the details of the application of the proposed method.

5.1 Stage-discharge relationships

Our first application is concerned with the construction of stage-discharge curves by using simultaneous measurements of water level (or stage) and discharge of a river. This is a very common hydrologic problem as the stage-discharge curves along with the detailed stage records are used to extract the discharge records almost in any river. Although there have been developed standards for the establishment of stage-discharge curves (International Organization for Standardization, 1973) the problem remains difficult and tedious. Its solution has a high degree of subjectivity as it depends on the expert's knowledge, experience and intuition. Recent developments tend to use Decision Support Systems to facilitate the establishment of stage-discharge curves (Gawne and Simonovic, 1994; DeGagne et al., 1996). Such systems are physically based as they consider the hydraulic characteristics of the river, such as energy slope, friction, channel aggradation and degradation, etc.

In Sections 1 and 3.2 we have already discussed the formulation of the problem from our point of view. According to this, rather black box, approach the only data needed is the set of concurrent measurements of stage and discharge. Apparently, the knowledge of the physical conditions of the river channel, and their changes in time, is quite helpful. However, in the example presented here, the Sakoulevas river, no such information is available. In Greece, almost all rivers exhibit significant instability in stage-discharge relationship (owing to frequent changes of river bed conditions) and thus, the construction of multiple stage-discharge curves is obligatory. The problem of constructing the stage-discharge curves becomes more complicated because of the large percentage of erroneous data, caused by non-representative measurements or, more rarely, by incorrect values.

The data set (the set A), depicted in Figure 1 consists of 78 measurements of Sakoulevas river for an eight-year period. The relatively small number of measurements facilitates the visualisation of the results (although we have successfully tested cases with much larger data sets).

We recall from Subsection 3.2 that the relationship among stage (x) and discharge (y) data is a power law $y = c\,x^d$ with $c > 0$. The parameters of the various components of the block consistency assessment function, as defined in Subsection 3.3, are shown in Table 1.

According to (7), the number of elements of the state space of the problem is 8.2×10^{36}. The algorithm evaluated 181 states and obtained in three seconds the results shown in Figure 1b. We observe that the system separated the data in three periods. The outliers (members of the set Q) are not shown in this figure. In Figure 1c we present one of the three blocks with the connected elements (the members of the consistent subset) along with the outliers of that period and the corresponding power law. As explained in section 3.1, due to the form of the objective function (which gives high priority to the number of outliers), the final number of outliers is the minimum possible. This means that the outliers shown in Figure 1c (even if seemingly their departure from the power law is not extremely large) could not be included to the block of consistent elements (in that case the consistency evidence function H_e would become zero).

5.2 Double mass analysis

In this example, the members of the input data set A contains 22 measurements of annual rainfall depths at two adjacent rain gauges located at Gregorio and Poros Reganiou in Sterea Hellas, Central Greece. The target of the problem is to test whether the data set is homogeneous or not, and, in the latter case, to partition it into homogeneous subperiods. A straight line passing from the origin, i.e.,

$$y = g(x) = c\,x \tag{57}$$

should fit the double mass plot of the two series if both are homogeneous during the entire period of the data set. Otherwise, more than one instances of the same law must be identified, each corresponding to a subperiod with homogeneous data. The variables x and y for each block $\delta = \{d_r, r = 1, \ldots, k\}$ are defined by

$$y_j = \sum_{r=1}^{j} h_2(d_r), \qquad x_j = \sum_{r=1}^{j} h_1(d_r) \tag{58}$$

where $h_1(d_r)$ and $h_2(d_r)$ denote contemporaneous rainfall depths in the two neighboring stations. Note that in this example, to determine the problem variables we need first to choose a partitioning into blocks.

The parameters of the specific tests of the block consistency function used are shown in Table 1. According to (7), the number of elements of the state space of the problem is 1.6×10^{10}. The algorithm evaluated 21 states and obtained in less than one second the results shown in Figure 2. We observe that the system isolated two elements (at years 1963 and 1961) and separated the other elements into two homogeneous periods (1983 - 1974 and 1973 - 1962). The isolation of the two elements is quite reasonable. The annual rainfall depth at Gregorio is generally greater than that at Poros Reganiou with a ratio of about 1.4 (the standard deviation is 0.2), whereas in 1963 this ratio becomes double this value and in 1961 becomes half this normal value. These are definitely extraordinary values suspect for measuring errors.

The separation into two periods is also justified. Geometrically this is indicated in Figure 2 by the significant departures of the second period's data from the straight line defined by the first period's data. To verify this further, we performed *t* tests at both data series for the significance of the difference between the means of the annual rainfall depth (not the cumulative values) of the two periods. For the Poros Reganiou data set no significant difference between the two periods was found. For the Gregorio station, when all data values are considered (including the two isolated elements), the difference between the two means (1983-1984 against 1973-1963) is significant at a significance level $\alpha = 0.05$. When the two isolated elements are removed, the difference between the means of the two periods becomes significant at $\alpha = 0.01$.

Interestingly, the same results were obtained by the system using the 264 monthly rainfall depths instead of the 22 annual values.

5.3 Analysis of temporal consistency

In the third example, already mentioned in the Introduction, our goal of consistency exploration is to find the start and the end of periods with faulty time entries of rainfall measurements, giving us the opportunity to check and correct the data record.

Again, in this problem we must partition the data set of the tested rain-gauge, into consistent blocks, in a manner that all rainfall measurements y of each block $\delta = \{d_r, r = 1, \ldots,$

$k\}$ have the same shift τ from the correct time, whereas measurements of different blocks have different lags. To this aim, we use as reference the concurrent measurements x of another rain-gauge, located near the tested rain-gauge. We assume that the measurements of the reference station have correct time entries, and, therefore, any measurement $y(t)$ is in fact concurrent with $x(t - \tau)$. Given the small distance between the locations of the two stations, we expect that their measurements, after the appropriate time shift, should be linearly correlated. Therefore, the law to be established is

$$y(t) = a + b\, x(t - \tau) \tag{59}$$

where all parameters a, b and τ have to be determined. The most difficult to determine is the values of the lag τ for the different blocks, which, besides, are the demands of the problem solution (parameters a and b are not required items of the solution). The problem is simplified if we consider τ as an integer multiple of the temporal resolution of measurements ε, i.e., $\tau = i\, \varepsilon$, where i is an integer ranging in a prespecified interval $[-k, k]$. All integer values of i in this interval must be tested for each block δ, and finally the value that maximises the function $h(\delta)$ is selected.

For the application of the method we have used the daily rainfall series of one year (1982) for stations Poros Reganiou and Anthero, with a distance of 6 km between each other. The parameters used for the components of the consistency assessment function are shown in Table 1. The data values of both stations were consistent in time, so we altered them to check if the algorithm can find the alterations. Specifically, we altered the February's data of Anthero, shifting the time field of each measurement by one day to the left. Also, we altered the October's data of the same station, shifting the time field of each measurement by one day to the right. In addition, we modified one measurement of rainfall depth at the same station, replacing the existing value with an extremely high one. In Figure 3 we give the altered hyetograph of Anthero and the original one at Poros Reganiou. The consistent periods identified by the system are separated by the vertical lines and the isolated observations are marked in this figure (the system found one more outlier). We observe that the system detected our "cheat", finding exactly the points where the "malfunction" appeared and disappeared.

6. Conclusions and discussion

The ascending hydrological information overload requires new computer-based methods for tackling common hydrological problems, formerly accomplished by manual methods. Some of these problems, such as the identification of shifting relationships between bivariate data sets and the exploration and testing of consistency or homogeneity of data sets, are not at all easy tasks when we attempt to solve them automatically with a computer program. They need a rigorous mathematical formulation and an intelligent and quick algorithm to find a solution. In this study we attempted to establish a generalised framework for solving such problems.

Thus, the first achievement of this work is the general mathematical formulation of a seemingly diverse collection of problems concerning hydrologic data processing. All examined problems are formulated as a single problem whose solution is the partitioning of the data set into consistent ordered blocks and the extraction from the data set of a subset holding the outliers. This formulation leads to the determination of the state space of the problem, which appears to be a huge set of possible solutions. The candidate solutions must be evaluated by optimisation of an appropriate objective function.

A wide range of statistical tests is investigated and systematically combined so as to form an objective function representing a measure of the consistency of a data set partitioned into blocks. This function is initially multidimensional and thus the optimisation is multiobjective. By simplifying the objective function, the problem has been reduced to a single-objective optimisation. The specific function proposed integrates the outcomes of several statistical and conceptual tests, and it can be easily expanded or modified, to include more components. Also, it includes several parameters that can be chosen to fit the needs of each specific problem. The parameter sets given in the examples studied may be used as a guide for similar situations.

Due to the magnitude of the state space of the problem, an exhaustive search of all candidate solutions is impossible, even for the smallest problems. Therefore, a heuristic algorithm is proposed, which makes the solution of the problem feasible by emulating the exploratory analysis of the human expert. This algorithm encodes a number of search strategies in a pattern directed computer program.

Three different applications of the method to real world hydrological problems are included to test the system's performance. They are concerned with typical hydrological problems such as the identification of stage-discharge curves, the double mass analysis and the analysis of temporal consistency of hydrological data. In all cases the performance of the system was very satisfactory. Moreover, in all our experiments the system provided improved solutions in comparison with those obtained by human experts, as in all cases the system utilised more data and its solutions had a higher consistency measure than any expert's solution. Conclusively, the system can assist the human expert by undertaking many of his or her tedious tasks.

Acknowledgments. The research leading to this paper was performed within the framework of the project Development of a National Data Bank for Hydrological and Meteorological Information (HYDROSCOPE), funded by the European Union in the framework of the STRIDE HELLAS research program (1992-94), and approved by the Greek General Secretariat of Research and Technology (GSRT). Additional funding was provided by the following Greek Organisations: the Ministry of Industry, Energy and Technology, the Ministry of Environment, Regional Planning and Public Works, the Ministry of Agriculture, the Ministry of Education, the Water Supply and Sewage Corporation of Athens, the National Meteorological Service, the Public Power Corporation, and GSRT. Computer sources were provided by the HYDROSCOPE project of National Technical University of Athens. The authors wish to thank V. Zoukos for his suggestions in the mathematical development of the method, I. Nalbantis for his comments, A. Manetas for his help in programming, and N. Mamassis for his help in the applications. The comments of the reviewers are gratefully appreciated.

References

Barnett, V. and T. Lewis, 1994, *Outliers in Statistical data*, 3rd edition, John Wiley and Sons, New York.

Cook, R. D., 1977, *Detection of Influential Observations in Linear Regression*, Technometrics, 19, 15-18.

Dodson, R. D., 1993, Advances in hydrologic computation, in *Handbook of hydrology*, D. R. Maidment (ed.), McGraw-Hill.

DeGagne, M. P. J., G. G. Douglas, H. R. Hudson and S. P. Simonovic, 1996, A decision support system for the analysis and use of stage-discharge rating curves, *J. of Hydrol.*, 184, 225-241.

Duda, R. O., and P. E. Hart, 1973, *Pattern classification and scene analysis*, John Wiley and Sons, New York.

Gawne, K. D., and S. P. Simonovic, 1994, A computer-based system for modelling the stage-discharge relationships in steady state conditions, *Hydrol. Sci. J.*, 39(5), 487-506.

Hald, A., 1965, *Statistical theory with engineering applications*, John Wiley and Sons, New York, 6th ed..

Hayes-Roth, B. and M. Hewett, 1988, BB1: An Implementation of the Blackboard Control Architecture, in *Blackboard Systems,* Engelmore and Morgan (eds.), Addison-Wesley, 297-313.

Hirsch, R. M., D. R. Helsel, T. A. Cohn, and E. J. Gilroy, 1993, Statistical analysis of hydrologic data, in *Handbook of hydrology*, D. R. Maidment (ed.), McGraw-Hill.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey, 1983, *Understanding robust and exploratory data analysis*, John Wiley and Sons, New York.

Interagency Advisory Committee on Water Data, 1982, *Guidelines for determining flood flow frequency*, Bulletin 17B, Hydrology Subcommittee, U. S. Department of the Interior, U. S. Geological Survey, Office of Water Data Coordination, Reston, Va..

International Organization for Standardization, 1973, ISO Standard 1100, Liquid flow measurements in open channels - Establishment and operation of a gauging station & Determination of the stage-discharge relation, ISO, Geneva, Switzerland.

Linsley, R. K., M. A. Kohler and J. L. H. Paulus, 1975, *Hydrology for Engineers*, 2nd edition, McGraw-Hill, Tokyo.

Moschovakis, Y. N., 1994, *Notes on set theory*, Undergraduate text in Mathematics (UTM), Springer-Verlag, NY (Greek edition: Nefeli, Athens, 1993).

Papakostas, N., I. Nalbantis and D. Koutsoyiannis, 1994, Modern computer technologies in hydrologic data management, *Proc. 2nd European Conference on Advances in Water Resources Technology and Management*, Lisbon, 14-18 June 1994, Balkema, Rotterdam, 285-293.

Papoulis, A., 1990, *Probability and Statistics*, Prentice-Hall, New Jersey.

Pearl, J., 1992, *Heuristics: Intelligent Search Strategies for Computer Problem Solving*, Addison-Wesley.

Stedinger, J. R., R. V. Vogel, and E., Foufoula-Georgiou, 1993, Frequency analysis of extreme events, in *Handbook of hydrology*, D. R. Maidment (ed.), McGraw-Hill.

Stoll, R. S., 1979, *Set theory and logic*, Dover Publications, NY.

Tsakalias, G., and D. Koutsoyiannis, 1994, OPSIS: An intelligent tool for hydrologic data processing and visualization, *Proc. 2nd European Conference on Advances in Water Resources Technology and Management*, Lisbon, 14-18 June 1994, Balkema, Rotterdam, 45-50.

Velleman, P. F., and D. C. Hoaglin, 1981, *Applications, Basics, and Computing of Exploratory Data Analysis,* Duxbury, Boston.

Worsley, K. J., 1983, Testing for a two-phase multiple regression, *Technometrics* 25(1), 35-41.

Table 1. Parameters of the block consistency assessment function for the application of the method in the three case studies presented in the paper.

Component	Parameters for the case study of		
	stage-discharge	double mass	temporal consistency
Determination coefficient*	$\rho_0^2 = 0.9$, $\alpha_1 = 0.05$	$\rho_0^2 = 0.9$, $\alpha_1 = 0.05$	$\rho_0^2 = 0.9$, $\alpha_1 = 0.05$
Standardised residual	$\alpha_2 = 0.025$	$\alpha_2 = 0.05$	$\alpha_2 = 0.05$
Standard deviation of residuals	$\sigma_0 = 0.35$, $\alpha_3 = 0.05$	$\sigma_0 = 5$, $\alpha_3 = 0.05$	$\sigma_0 = 1.2$, $\alpha_3 = 0.05$
Runs	$\alpha_4 = 0.05$	$\alpha_4 = 0.01$	$\alpha_4 = 0.01$
Marginal outliers	Disabled	$\alpha_5 = 0.01$	Disabled
Number of elements of blocks	$b_6 = 10$	$b_6 = 8$	$b_6 = 8$

* For both consistency assessment and consistency measure functions.

List of Figures

Figure 1. Construction of stage-discharge curves for the Sakoulevas river: (a) The whole initial data set. (b) The consistent blocks separated by the system, without the outliers; triangles correspond to period 1 (02/1964-05/1966), circles correspond to period 2 (06/1966-05/1969) and rectangles to period 3 (06/1969-01/1972). (c) The data of period 1 (triangles) along with the outliers (symbols) belonging to this period, and the corresponding least squares power curve.

Figure 2. Double mass plot of annual rainfall at stations Poros Reganiou and Gregorio. Data are plotted in reverse order (the point close to the origin corresponds to the last year). Circles correspond the first period (1983-1974), rectangles to the second period (1973-1962) while the symbols correspond to the outliers (1963 and 1961). The straight line is the least-square line of the first period.

Figure 3. One year (1982) hyetographs of two rain-gauges as an example of the temporal consistency analysis. Dotted line represents the measurements at the reference station (Poros Reganiou), while continuous line represents the measurements of the tested (and altered) station (Anthero). For clarity of the diagram we have removed prolonged intermediate dry periods, although the system considered the whole data sequence.

Figure 4. Geometric explanation of the basic definitions and notation. (a) Set-theory representation: The data set *A* to be explored is represented by the thick straight line and its elements, ordered by time, are represented by the circles. Some of the elements of *A* form the set of outliers *Q*. The other elements are partitioned into three blocks, forming the ordered partition $\Delta = \{\delta 1, \delta 2, \delta 3\}$.

Figure 5. Schematic representation of eight ($L^1 \ldots L^8$) consecutive steps of the search for the best partition of a data set with 24 points. In the initial step (L^1) all 24 data points (open circles) are unexplored whereas in the final step (L^8) 21 data points (filled circles) have been partitioned into three blocks (gray frames) and other three points (gray squares) are marked as outliers. Labels indicate the transformations instantiated.

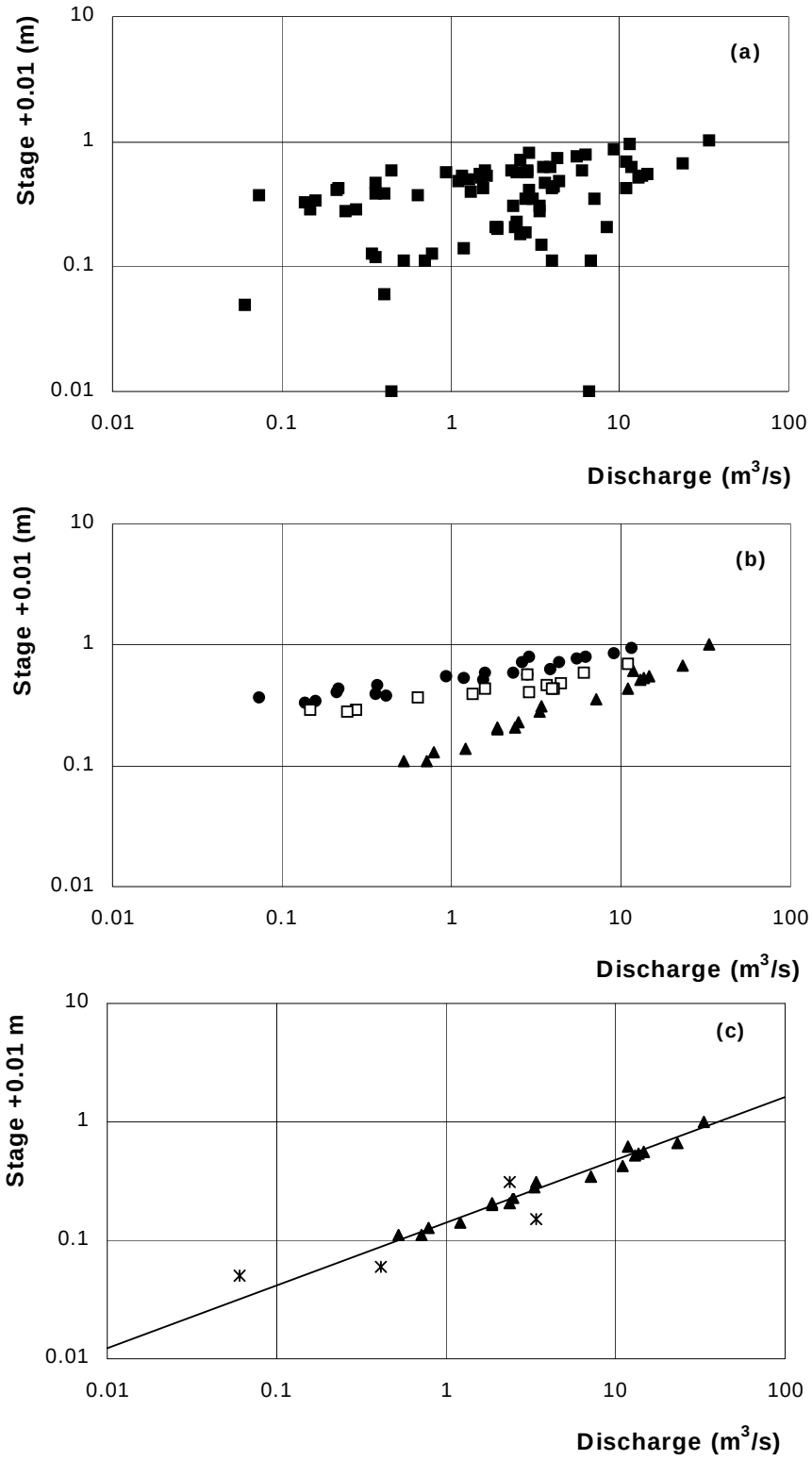

Figure 1. Construction of stage-discharge curves for the Sakoulevas river: (a) The whole initial data set. (b) The consistent blocks separated by the system, without the outliers; triangles correspond to period 1 (02/1964-05/1966), circles correspond to period 2 (06/1966-05/1969) and rectangles to period 3 (06/1969-01/1972). (c) The data of period 1 (triangles) along with the outliers (symbols ✳) belonging to this period, and the corresponding least squares power curve.

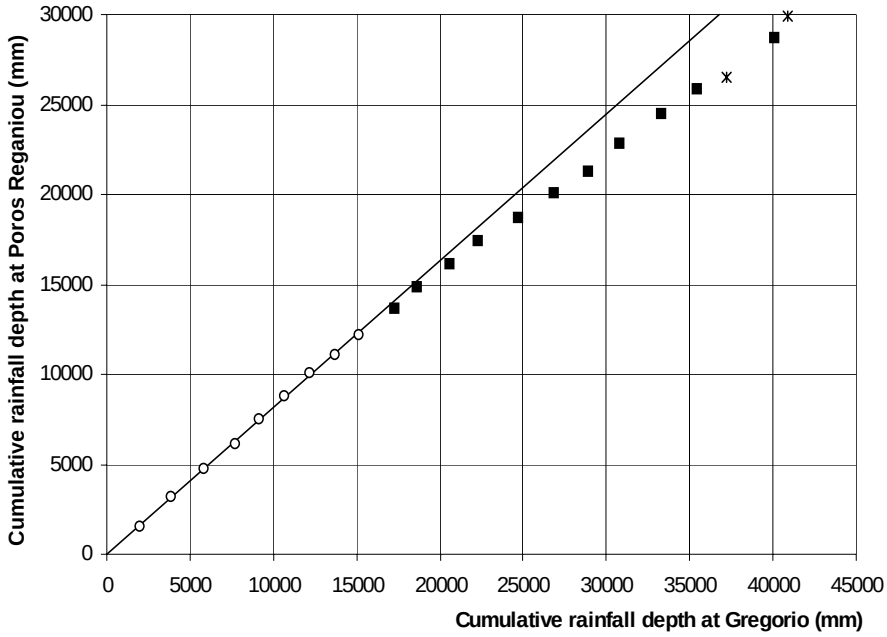

Figure 2. Double mass plot of annual rainfall at stations Poros Reganiou and Gregorio. Data are plotted in reverse order (the point close to the origin corresponds to the last year). Circles correspond the first period (1983-1974), rectangles to the second period (1973-1962) while the symbols ⧓ correspond to the outliers (1963 and 1961). The straight line is the least-square line of the first period.

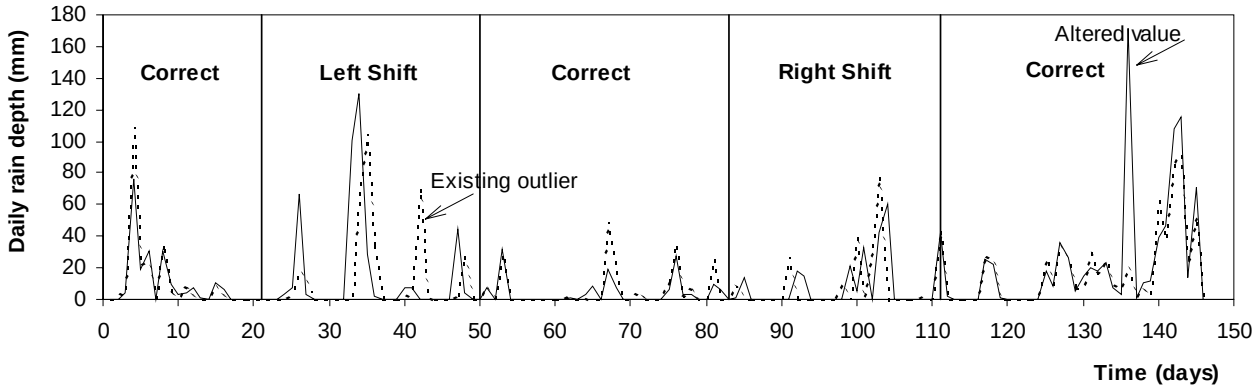

Figure 3. One year (1982) hyetographs of two rain-gauges as an example of the temporal consistency analysis. Dotted line represents the measurements at the reference station (Poros Reganiou), while continuous line represents the measurements of the tested (and altered) station (Anthero). For clarity of the diagram we have removed prolonged intermediate dry periods, although the system considered the whole data sequence.

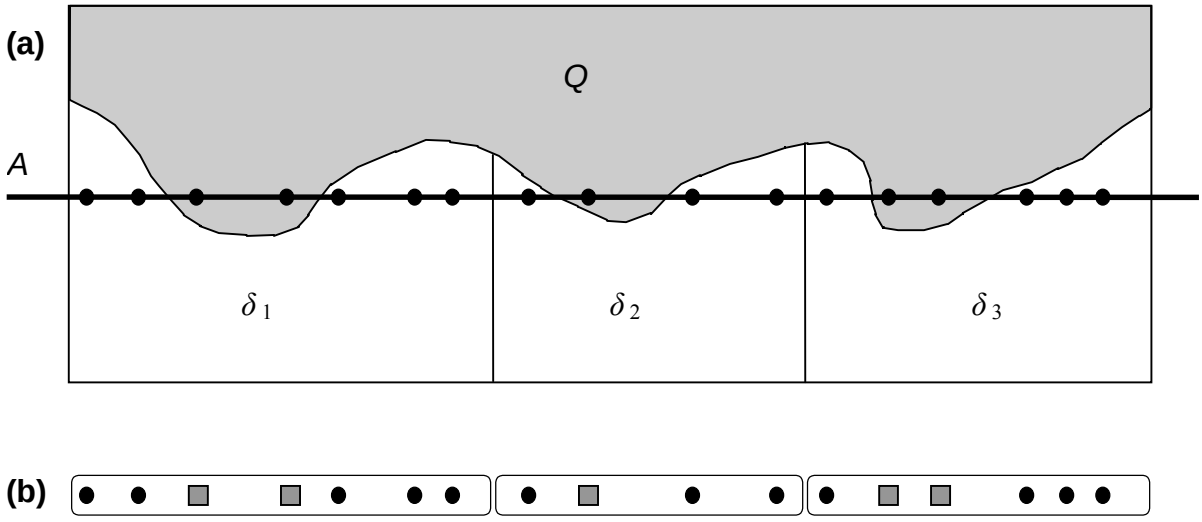

Figure 4. Geometric explanation of the basic definitions and notation. (a) Set-theory representation: The data set A to be explored is represented by the thick straight line and its elements, ordered by time, are represented by the circles. Some of the elements of A form the set of outliers Q. The other elements are partitioned into three blocks, forming the ordered partition $\Delta = \{\delta_1, \delta_2, \delta_3\}$. (b) Simplified schematic representation of the same.

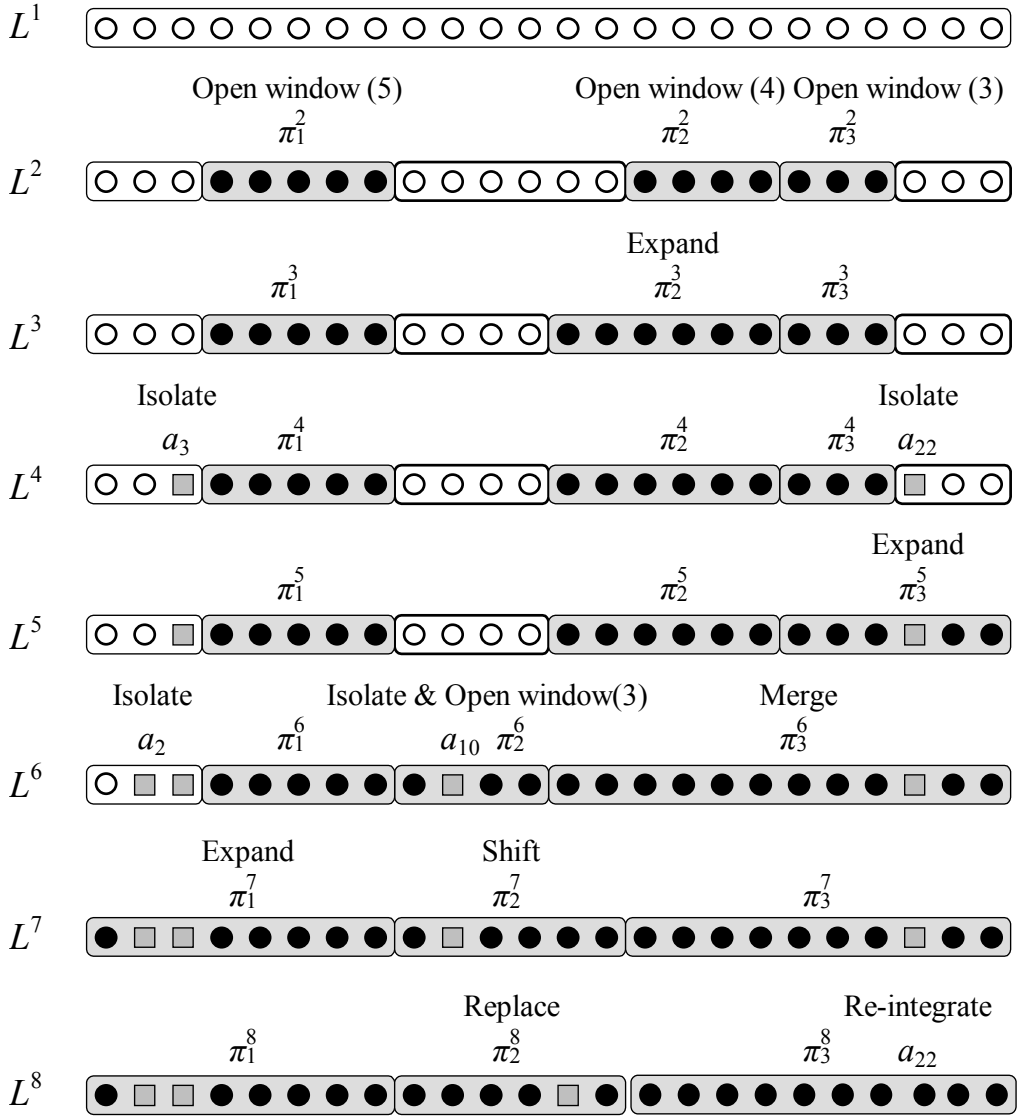

Figure 5. Schematic representation of eight (L^1 … L^8) consecutive steps of the search for the best partition of a data set with 24 points. In the initial step (L^1) all 24 data points (open circles) are unexplored whereas in the final step (L^8) 21 data points (filled circles) have been partitioned into three blocks (gray frames) and other three points (gray squares) are marked as outliers. Labels indicate the transformations instantiated.