Stochastic Hydrology and Hydraulics 8, 301-317 © Springer-Verlag 1994

Entropy principles in the prediction of water quality values at discontinued monitoring stations

A. Kusmulyono and **I. Goulter**
University of Central Queensland, Rockhampton, Queensland 4702, Australia

Abstract: A new methodology for predicting water quality values at discontinued water quality monitoring stations is proposed. The method is based upon the Principle of Maximum Entropy (POME) and provides unbiased predictions of water quality levels at upstream tributaries and on the mainstem of a river given observed changes in the distribution of the same water quality parameter at a downstream location. Changes in the values of water quality parameters which are known a priori to have occurred upstream, but which are not sufficiently large to account for all the observed change in the same water quality parameter at the downstream location are able to be incorporated in the method through the introduction of a new term in the basic entropy expression. Application of the procedure to water quality monitoring on the Mackenzie River in Queensland, Australia indicates the method has considerable potential for prediction of water quality at discontinued stations. The method also has potential for identifying the location of causes of observed changes in water quality at a downstream station.

Key words: Change, discontinued stations, entropy, networks, optimization, prediction, unbiased, water quality.

1 Introduction

Water quality management and the co- and pre-requisite requirements for monitoring have been become one of the most pressing problems for the authorities involved in management of water in river systems. The complexity of the problems associated with effective water quality monitoring are related to a range of factors including (a) the objectives of monitoring, (b) the variables to sample, (c) the locations of stations (d) the frequency of sampling, and (e) how long should a station be operated in relation to the objectives in (a).

All these factors are related to the process of designing a network and/or the monitoring program to be undertaken in relation to a new or existing network. However all designs, be they of the network itself, or of the monitoring program, as well as being concerned with collection of the actual data, must also be effective in gathering those data, and cost efficient in obtaining the information as it relates to those data. [This difference between data and information is summarized nicely in the adage of "data rich but information poor", (Ward et al., 1986)]. These two requirements lead to a need for a means for evaluating the 'performance' of a network. Such an evaluation must include some concept of the 'benefits' of the monitoring in relation to the objectives of that monitoring and the cost, both marginal and average, of obtaining those benefits.

In many cases, network and monitoring program design is controlled by the available budget and the problem becomes one of obtaining the greatest benefit (most information) for that level of budget. Not unexpectedly a number of studies have been conducted over the years to optimize the design of water quality monitoring networks and the monitoring programs to be undertaken with networks.

Loftis and Ward (1980) attempted to identify 'regions' of frequencies of sampling in water quality monitoring. The criterion adopted in their study was based on the width of confidence intervals of water quality variables about the annual sample geometric means. The study identified three general 'regions' within sampling programs, namely: Region 1, which is characterized by high sampling frequencies, where the role of serial correlation is dominant; Region 2, which is characterized by sampling frequencies between approximately 10 and 30 samples per year and where effects of seasonal variation and serial correlation tend to cancel each other out; and Region 3, which is characterized by low sampling frequencies and where seasonal variation plays the dominant role. Loftis and Ward (1980) noted that in Region 2, seasonal variation and serial correlation should either both be considered or both ignored; to consider only seasonal variation will lead to more error than ignoring it.

Palmer and Mackenzie (1985) discussed 'monitoring-effectiveness' and monitoring costs and the use of optimization methods (incorporated into an interactive computer program) to select the aquatic monitoring design that maximizes cost-effectiveness. These authors developed a new approach to cost-effective design of aquatic monitoring networks in which the actual cost minimization issue was addressed by maximizing statistical power for a specified financial budget or, conversely, minimizing cost for a specified statistical power requirement. Both formulations are based on a gradient search algorithm. The results provided by the two models showed that, up to a certain threshold, the potential statistical power available from data (information) is strongly affected by the budget available. Above the threshold very little additional power is gained even with large increases in budget.

Dunnette (1980) noted that ideally a water quality index should be used to determine sampling frequency. In that study the sampling frequencies were actually determined on the basis of observed variability in the Oregon Water Quality Index (OWQI). These sampling frequencies indicated the number of samples required to meet imposed confidence and error limits. It is also mentioned in the paper that the objectives and constraints of the sampling program should be used as the basis for the selection of the time intervals in which to distribute water quality samples.

Harmancioglu (1984) introduced the entropy concept to determine the optimal sampling intervals in water quality monitoring. The entropy principles in that case were applied to determine the information content of stochastic dependent variables in order to identify the optimum sampling intervals with respect to time. This work on the application of entropy principles to design of water quality monitoring networks was subsequently extended to assessment of network efficiency and cost effectiveness (Harmancioglu and Alpaslan, 1992). In that study, the entropy principle was used to quantify information contained within a set of water quality data from a network. Using the quantified information, the efficiency of a network was then analyzed by maximizing the amount of information collected from the network. Cost effectiveness, on the other hand, was evaluated by comparing the costs of monitoring versus the information gained via monitoring. It was shown by Harmancioglu and Alpaslan (1992) that the entropy principle is applicable for network assessment, particularly in cases of rationalization of networks.

All of the studies cited above were directed at identifying procedures or models which were able to identify network designs and/or monitoring programs which give the best information within specified budget limits. This study is directed at another aspect of the budget problem within water quality monitoring, namely, how to predict water quality changes at some upstream locations (tributaries) after the stations at those locations have been discontinued following a period of data collection sufficient to establish the base-line distribution of the water quality at each station. This type of problem may arise in a number of ways. In one situation the budget available for water quality monitoring may be reduced due to external economic factors or the budget may be static (in which case inflation causes a real decrease in funding). In both situations it may be necessary to reduce either the sampling frequency or the number of stations at which water quality sampling is carried out. In the second case of static budgets, or in some situations of slightly rising budgets, it may be necessary to transfer an existing water quality monitoring station from one location to another in response to a more acute need for data at the new location.

In this paper, a method to predict the water quality levels at discontinued upstream stations with an approach based on entropy/information theory using measured data at a downstream main channel station is proposed. Alternately, if it can be shown that the water quality values predicted by the method at existing stations are sufficiently accurate, the method also has the capability of identifying opportunities for discontinuing a number of stations when budgetary limitations are causing a 'rationalization' of the network design.

2 Historical use of the entropy concept in water resources

The entropy concept has been introduced to water resources relatively recently. Sonuga (1972) applied the principle to parameter estimation and derivation of frequency distributions. He also applied the concept to derivation of functional rainfall-runoff relationships (Sonuga, 1976). Harmancioglu examined use of the entropy principle in the measurement of the information content of random process (Harmancioglu, 1981); evaluation of information transfer between hydrologic processes (Harmancioglu and Yevjevich, 1987); and assessment of recharge systems for a river basin (Harmancioglu and Baran, 1989). Amorocho and Espildora (1973) utilized the principle to assess the hydrologic model performance while Chiu (1987, 1988, 1989, 1991) and Chiu and Chiou (1986) applied the principle to velocity distributions in open channels. Awumah et al. (1990, 1991) on the other hand developed the principle for use in redundancy measures for water distribution network design.

3 Problem statement

The particular application of entropy examined in this paper embodies the use of the Principle of Maximum Entropy (POME) to develop updated probability distributions of water quality levels in upstream tributaries and the upstream mainstem of a river where monitoring has been discontinued, given 1) an observed change in the distribution of the water quality observed at the downstream location and 2) knowing the previous probability distributions of the water quality levels at the upstream tributary stations. (Note the method is also able to predict water quality levels in the upstream tributaries and upstream mainstem if no changes have been observed downstream. However, under such a scenario of no downstream changes, changes are also unlikely to have occurred at the upstream locations and there is, therefore, relatively little need to predict water quality values at those upstream locations because the distribution of values at the stations can reasonably be expected to be the same as that previously observed).

A potential additional use of the methodology, beyond the simple prediction of water quality values at upstream stations, is in its contribution to the identification of potential locations of causes of changes in downstream water quality values. Once the updated probability distributions at each upstream location have been identified by the method, the likely location(s) (mainstem and/or one or more tributaries) of the cause of the water quality changes observed at the downstream station can be identified. Such an identification of likely location(s) of causes of observed downstream changes might be based upon the statistical likelihood of the difference between the newly predicted distribution of water quality and the 'old' (known) distribution of water quality values.

In the situation where changes affecting water quality are known to have occurred upstream, but where those changes are not sufficiently large to account for all the changes observed downstream, it is possible to modify the method to identify the likely location(s) of the cause of observed changes at the downstream station which are not able to be accounted for solely by the known upstream changes. An important characteristic of the method as it used in this fashion is that, in the absence of actual monitoring at upstream locations, it gives unbiased estimates of the likely locations of upstream changes in the water quality, and indicates, again in an unbiased manner, the likely magnitude of these changes.

It should be noted that the method is applicable for identification of the likely upstream locations of causes of long term changes in water quality at downstream stations rather than for identifying the locations of sources of short term or transient variations in water quality.

The theoretical basis of the procedure and the practical considerations of its application to the prediction of the water quality are described in the following sections.

4 Theoretical background

The basic principle of the procedure is the interpretation of entropy as expressed by Shannon's measure of information (Shannon, 1948). This entropy expression can be interpreted as a measure of uncertainty and can be explained as follows: Let the probabilities of n possible outcomes A_1, A_2, ..., A_n of an experiment be p_1, p_2, ..., p_n respectively. Shannon's formulation for entropy can be written mathematically in these terms as:

$$H = -\sum_{i=1}^{n} p_i \ln p_i \tag{1}$$

where:

$$\sum_{i=1}^{n} p_i = 1 \tag{2}$$

[Equation (2) is needed to ensure development of a complete probability distribution].

The important characteristics of the formulation expressed by Equations (1) and (2) are:

a. H takes on its maximum when all events have the same probability or uncertainty, i.e, $p_i = 1/n$.
b. H takes on its minimum value (equal to 0), when there is a certainty among the events.
c. Any random probabilities will give a value of H between these two extremes.

The character of the entropy function described above also means that the probability distribution with the maximum entropy is the most unbiased distribution consistent with the information specified by the constraints. This observation in turn means that, without any constraints other than Equation 2, the distribution developed by the formulation is the most dispersed, i.e., it is a uniform distribution.

The results obtained from maximizing H [Equation (1)] by assigning values to p_i has been discussed extensively in earlier papers, e.g., Sonuga (1972), and Jaynes (1983) and is known as the Principle of Maximum Entropy (POME). The underlying principle of this assignment of p_i values is that maximizing the value of H in this manner will result in the most unbiased estimate for p_i for any condition defined by the constraints on the values of p_i.

One such formulation involving constraints is as follows:

$$\text{Max } H = -\sum_{i=1}^{n} p_i \ln p_i \tag{3}$$

Subject to:

$$\sum_{i=1}^{n} p_i = 1 \tag{4}$$

$$\sum_{i=1}^{n} p_i x_i = \mu \tag{5}$$

$$\sum_{i=1}^{n} p_i x_i^2 = \mu^2 + \sigma^2 \tag{6}$$

where:
p_i = probability of event x_i
μ = mean of the outcomes of events x_i for all i
σ = standard deviation of the outcomes of events x_i for all i

Solving the above formulation for the unknown values of p_i results in a normal distribution with a mean of μ and a standard deviation of σ. [Note that solving the same formulation but with the constraints defined in Equations 5 and 6 removed, results in the specification of a uniform distribution of x_i.]

Prior information about the distribution of x_i can be incorporated into the above formulation by use of the Kullback-Leibler's Principle of Minimum Discrimination Information (MDI) by modifying Equation (3) to the following form

$$\text{Min } H = \sum_{i=1}^{n} p_i \ln (p_i/q_i) \tag{7}$$

$$\text{or} \quad \text{Max } H = -\sum_{i=1}^{n} p_i \ln (p_i/q_i) \tag{8}$$

where:

q_i = prior knowledge of the probability of event x_i

It should be noted that the 'prior knowledge' of the probability of event x_i can be the probability of the event x_i known from a previous investigation or the probability of x_i estimated to be now occurring as a result of observed changes in the system.

This basic principle of maximizing H subject to a range of constraints on the values of p_i, and having some basic information on the probability distribution to be estimated, i.e., having some prior knowledge of the probabilities, is adopted in the approach to predict water quality levels at upstream locations in a river system described in the following sections.

5 Formulation of the water quality estimation problem

Consider a river basin in which sufficient water quality data have been gathered to define the probability distributions of the water quality on a number of major tributaries and in the upstream and downstream reaches of the main channel of the river. For simplicity of explanation at this time consider the case of a stream with two tributaries described in Figure 1. Define events as the range of the possible values of a water quality parameter at each sampling station on the tributaries and the mainstem of the stream. The range of these 'possible' values can be estimated in a number of ways, for example, as the values lying within four standard deviations either side of the mean values.

Assume that the water quality at the downstream location a is a function of the water quality at the two upstream location b and c, i.e.,

$$x_a = f(x_b, x_c) \tag{9}$$

This assumption implies that there are no inputs of pollutants etc. between locations b and c and location a which significantly effect the value of the water quality parameter in question. The type of function described by f() in Equation (9) depends on the water quality parameter (pollutant) being monitored and the particular physical conditions, i.e., flow, distance between stations for the situation being examined, whether the pollutant is conservative, and the time response of the pollutant if it is in fact non-conservative.

Continuing with the same simple problem now consider the case where the water quality level at the downstream main channel station a has changed considerably. Such a situation indicates that changes in the values of water quality levels are likely occurring either at upstream location b or c or both. The POME [Equations (3)-(6)] is proposed as a means of predicting, without bias, the distribution of water quality levels in the upstream tributaries (and upstream mainstem if appropriate) which are most likely given the change observed at the downstream location.

The modified form of the POME for this case is as follows:

$$\text{Max } H = -\sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij} \ln[p_{ij}/(q_{ij}/m)] \tag{10}$$

Subject to:

$$\sum_{i=1}^{n} p_{ij} = 1/m \qquad (j = 1, 2, ..., m) \tag{11}$$

$$\sum_{i=1}^{n} p_{ij}x_{ij} / \sum_{i=1}^{n} p_{ij} = \mu_j \qquad (j = 1, 2, ..., m) \tag{12}$$

$$\sum_{i=1}^{n} p_{ij}x_{ij}^2 / \sum_{i=1}^{n} p_{ij} = \mu_j^2 + \sigma_j^2 \qquad (j = 1, 2, ..., m) \tag{13}$$

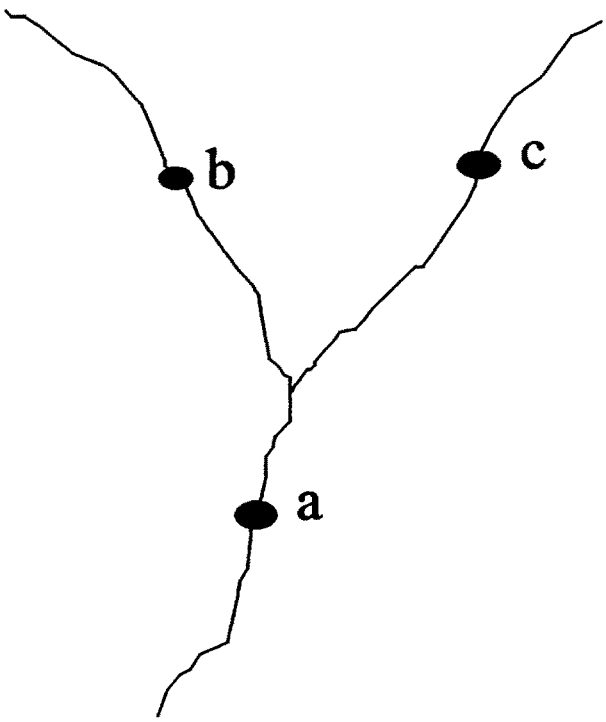

Figure 1. Schematic of explanatory example

$$\mu = f(\mu_1, \mu_2, \ldots, \mu_m) \tag{14}$$

$$0 \leq p_{ij} \leq 1 \qquad \text{for all i, j} \tag{15}$$

$$0 \leq q_{ij} \leq 1 \qquad \text{for all i, j} \tag{16}$$

$$\mu_j \geq 0 \qquad \text{for all j} \tag{17}$$

$$\sigma_j \geq 0 \qquad \text{for all j} \tag{18}$$

where:
x_{ij} = possible water quality level i at station j
q_{ij} = prior probability of event x_{ij}
μ_j = mean of the water quality level at station j
σ_j = standard deviation of water quality level at station j from the prior distribution
μ = observed (changed) mean of the water quality level at the downstream location
p_{ij} = probability of event x_{ij} to be assigned knowing the mean of water quality level downstream μ
m = number of the upstream stations.
n = number of intervals (discrete water quality values) at each station.

This formulation is applied separately for each water quality parameter of interest.

Given an observed change in water quality levels at a downstream location, the above formulation assigns probabilities to each of the possible water quality levels of each of the upstream stations. These new probabilities are then used to develop new, unbiased, estimates of the mean values of the water quality at the upstream stations.

The q_{ij} values in the formulation can be those probabilities which existed prior to the observed downstream change in water quality, or, when changes are known to have occurred upstream, (but where as noted previously these observed upstream changes cannot account the magnitudes of the changes observed downstream), the values of the probabilities which can be associated with (accounted for by) the new known conditions.

6 Demonstration of the use of the technique

Water quality data from the Fitzroy River basin in Central Queensland, Australia are used in this study to demonstrate how the MDI can be used to predict water quality values accurately. The locations of the stations in this basin from which the data used to demonstrate the model are obtained are shown in Figure 2.

The value of the 'new observed mean' at the downstream station from which the new distributions of water quality are to be estimated at the upstream stations may be annual mean values or, in order to reduce the impact of 'one-off' short term trend changes, the mean over a period of years (equivalent to a moving average). (Recall that the method is not for predicting the distribution of values of water quality parameters in the face of short term transient changes in observed water quality values). The change in the probability distribution of water quality at the upstream station occurring as a result of using either annual mean, two year moving average, three year moving average values etc. of the water quality at the downstream stations can be significant depending on the length of record of the data. This issue is discussed in detail in Kusmulyono and Goulter (1994).

In this study the prediction and subsequent comparison of water quality values analysis were performed only for a four year moving average of the mean annual value of the water quality. The data used in the study were collected from 1971 to 1989 and were mainly available on a quarterly basis. The data were divided into two groups:

(a) the data in the first group was assumed to be the data collected in the earlier period of time and constitute the information available to develop the prior probability distribution (q_{ij} values) of the water quality at each station;

(b) the data in the second group constitute the value of the water quality to be estimated at each tributaries knowing only the mean value of the annual mean at the downstream station for the corresponding period of time.

The specific water quality variables considered in the study are conductivity, dissolved ions, dissolved solids and hardness. Importantly these data, which are shown in Table 1, show significant differences between the two groups of data.

Normality tests were conducted on the data using the Saphiro-Wilk W test, and most of the values were found to be normally distributed. Only the Hardness data at station 130106 were found to be not normally distributed. However, the W values in this case is still very close to the W critical specified in the tables. Therefore, it was assumed that normal distribution would still be an appropriate model for this parameter. These normality tests were applied to the data because the model developed in the formulation shown in Equation (10)-(18) is strictly valid only for normal probability distributions. If the data are not, in fact, normally distributed, the model formulation has to be modified, or the data transformed into normal distribution, before the principle can be applied in the form given by Equations (10)-(18).

The procedure to develop the relationship between the water quality value at the upstream and downstream locations in the functional form described by Equation (14) of the constraint set of the entropy formulation is based on the following process. Suppose 10 years of data exist from 1971 to 1980. The data at every station can be grouped into seven '4-yearly' periods (1971-1974, 1972-1975, ..., 1977-1980) and the four year 'moving average' mean of the data for each station calculated. These four year moving average values are then used in a regression analysis between the values at the downstream station and the summation of the values from the upstream tributaries. The values in the upstream tributaries can be weighted by the discharges or the catchment areas of the tributaries in order to recognize the proportional contribution of each tributary to the water quality at the downstream station.

Moving average regression analysis was chosen because the proposed method is intended for prediction of the mean value of the water quality over a period of time rather than for analysis of short term variations. The moving average values are considered to be the best estimate for this relationship as they damp any short term water quality changes which might occurred in the historical record for the upstream and downstream stations. The values of the regression coefficients for the four-year moving average approach used in this paper are shown in Table 2.

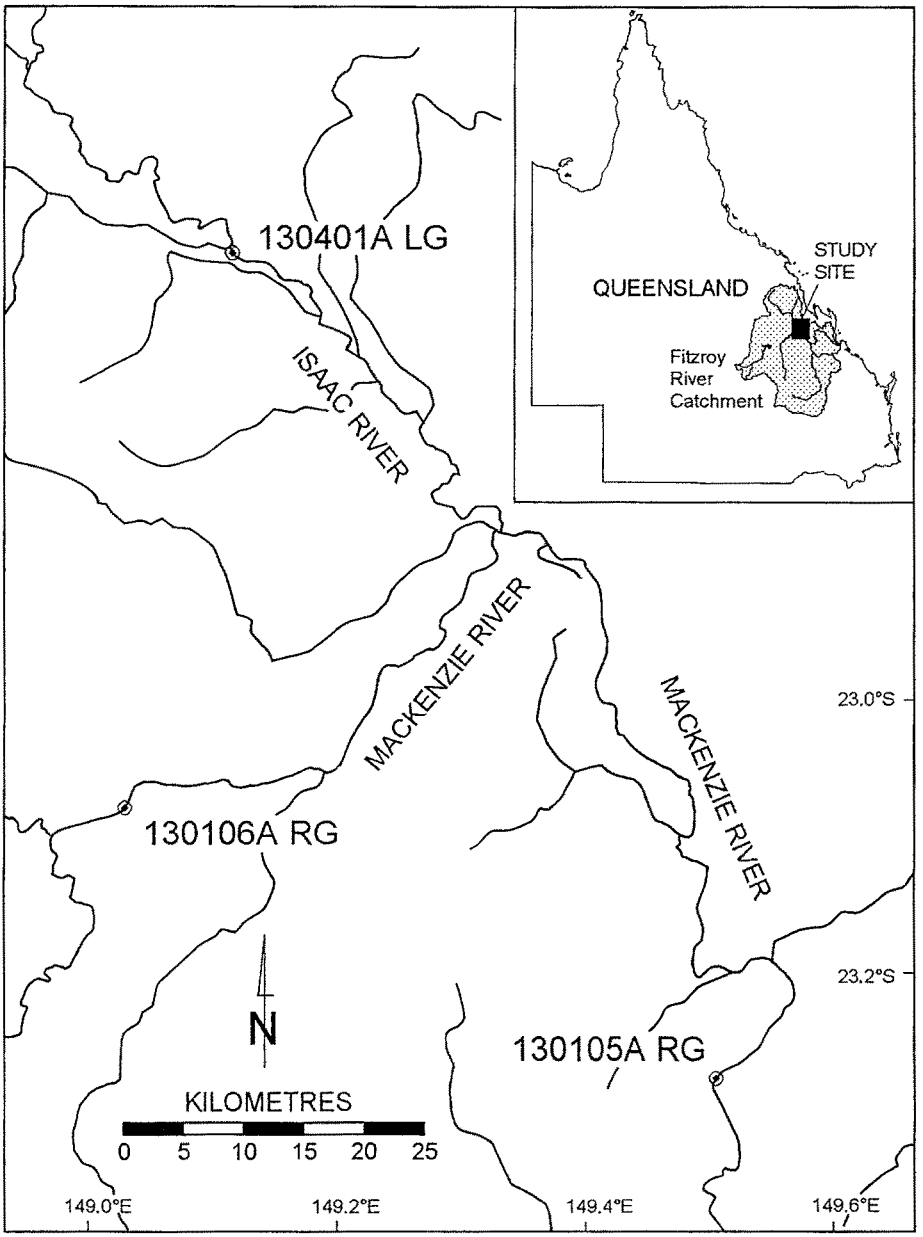

Figure 2. Location of the water quality monitoring stations.

Table 1. Water quality data at MacKenzie River and Isaac River sub-basin.

a. STATION 130401 (at Isaac River)

CALIBRATION

Year	Conductivity @ 25 C (mS/m)	Dissolved Ions (mg/l)	Dissolved Solids (mg/l)	Hardness (mg/l)
71	270.0	196.1	-	78.0
72	351.3	265.9	-	102.8
73	237.8	173.8	140.8	73.5
74	471.0	301.0	249.5	136.5
75	350.0	211.4	176.0	89.0
76	550.0	359.4	303.0	157.7
77	311.0	204.3	174.2	86.4
78	457.5	304.9	263.0	129.0
79	660.0	411.0	343.0	181.0
80	429.0	269.6	231.8	113.0
Mean	408.8	269.7	235.2	114.7
Standard Deviation	131.0	76.4	68.1	35.9

VERIFICATION

Year	Conductivity @ 25 C (mS/m)	Dissolved Ions (mg/l)	Dissolved Solids (mg/l)	Hardness (mg/l)
81	425.0	200.8	218.5	106.5
82	316.7	200.7	173.3	83.3
83	-	-	-	-
84	332.5	219.3	185.0	83.5
85	283.3	176.7	156.7	72.0
86	276.7	182.6	150.0	75.7
Mean				
(81-85)	339.4	199.4	183.4	86.3
(82-86)	302.3	194.8	166.3	78.6

Table 1 (continued)

b. STATION 130106 (at Mackenzie River)

CALIBRATION

Year	Conductivity @ 25 C (mS/m)	Dissolved Ions (mg/l)	Dissolved Solids (mg/l)	Hardness (mg/l)
71	230.0	199.3	-	91.0
72	204.8	167.1	-	78.6
73	146.7	135.8	100.3	61.7
74	168.6	136.1	104.2	61.7
75	192.2	151.0	117.3	65.7
76	410.0	347.2	257.0	162.0
77	240.6	198.6	151.4	85.2
78	160.0	127.5	100.5	55.2
79	340.0	257.6	199.0	101.0
80	230.3	165.9	135.0	69.3
Mean	232.3	188.6	145.6	83.1
Standard Deviation	83.2	68.2	56.1	31.3

VERIFICATION

Year	Conductivity @ 25 C (mS/m)	Dissolved Ions (mg/l)	Dissolved Solids (mg/l)	Hardness (mg/l)
81	199.3	151.0	114.8	58.8
82	211.3	163.3	122.5	72.5
83	-	-	-	-
84	206.7	161.3	123.3	62.0
85	242.5	166.2	140.0	65.5
86	185.0	150.2	110.0	61.0
Mean				
(81-85)	214.9	160.4	125.2	64.7
(82-86)	211.4	160.3	124.0	65.3

Table 1 (continued)

c. STATION 130105 (at Mackenzie River)

CALIBRATION

Year	Conductivity @ 25 C (mS/m)	Dissolved Ions (mg/l)	Dissolved Solids (mg/l)	Hardness (mg/l)
71	-	-	-	-
72	290.0	199.7	148.5	80.5
73	215.0	159.6	135.8	61.8
74	350.0	233.5	194.0	97.0
75	235.0	158.4	133.7	63.7
76	516.7	336.2	274.0	150.3
77	350.0	248.6	204.0	102.3
78	367.5	258.2	207.5	110.5
79	-	-	-	-
80	411.7	220.0	220.0	107.7
Mean	343.2	231.2	189.7	96.7
Standard Deviation	97.6	58.7	48.3	28.8

VERIFICATION

Year	Conductivity @ 25 C (mS/m)	Dissolved Ions (mg/l)	Dissolved Solids (mg/l)	Hardness (mg/l)
81	248.0	176.7	139.0	71.0
82	260.0	167.0	150.0	63.0
83	-	-	-	-
84	226.3	161.0	130.0	64.0
85	235.0	161.0	140.0	61.5
86	255.0	187.9	146.7	80.7
Mean				
(81-85)	242.3	166.4	139.8	64.9
(82-86)	244.1	169.2	141.7	67.3

Table 2. Values of regression coefficients for various variables.

	CONDUCTIVITY	DISSOLVED IONS	DISSOLVED SOLIDS	HARDNESS
B_0	137.72	100.58	50.978	51.357
B_1	0.5980	0.5914	0.7423	0.5351

The range of water quality values to be considered in the entropy formulation at the upstream station were set at plus and minus four standard deviations either side of the mean of the relevant station. This range of values was discretized into 40 intervals. The discretization of the continuous variables implicit in the base entropy function and necessary for its use in this and other studies involving similar uses of the entropy expression has been identified by Harmancioglu (1992) as being a critical element in the appropriate use of the function. Examination of the impacts of varying the discretization intervals for formulation described in this paper are reported in detail in Kusmulyono and Goulter (1994). The results in Kusmulyono and Goulter (1994) indicate that, while the value of H and the values of probability assigned vary with the size of discretization intervals, the new mean values assigned by the formulation do not change indicating that the model results are effectively independent of the size of the discretization interval.

The following four cases were used in demonstrating the methodology. Firstly, the variances of the data at every station were maintained at the values associated with previously determined probability distributions. Secondly, the variance of the water levels at the upstream locations were bounded only by the requirement that their sum equal a function of the observed variance of the water quality values at the downstream station. In this second case, the variance of the values at the downstream station was assumed to be constant at the level of the prior distribution at the downstream station. However, in reality any appropriate variance, e.g., a new variance observed at the downstream station can be specified. Note that, in this second case, the MDI model determines the variances of the water quality at each upstream station. In the last two cases, the 'prior' probabilities q_{ij} are modified to demonstrate how known changes in the environment around the upstream tributaries can be accommodated. This incorporation of known changes may be undertaken with the variances at the upstream stations constrained to historical or known values (equivalent to Case I) or with the variances at the upstream stations unconstrained other than their sum must relate to the observed variance at the downstream station (equivalent to Case II)

The formulation was solved using the nonlinear optimization package program GRG2 (Lasdon and Warren, 1986) . Table 3 summaries the values achieved for the four variations of the basic model.

7 Discussion

In Table 3, the new mean water quality values for each variable predicted by the model are shown in column (4) as the assigned values.

CASE I [Table 3 (a)]. The results in this table show how well the values predicted by the basic entropy function match the values derived from observed data. This closeness occurs even though there are significant differences between the values used to derive the q_{ij} values in the entropy function and the values which the entropy function is trying to predict. It can be seen that the greatest percentage error occurs in the predicted value of Hardness at station 130106 ($\simeq$15%). The high error percentage for Hardness in this case is believed to be due in part to the normal approximation being applied to data that are not actually normally distributed.

CASE II [Table 3(b)]. Recall that in this case the formulation has been modified such that the variances of the new probability distributions at the upstream stations were also assigned by the method. The only requirement on the variances is that their sum should be equal to the sum of the variances from the observed data. The result shows that the effect of permitting the change on the variance is not significant to the assignment of the new mean values. (The new mean values assigned by the method are not significantly different from CASE I).

Table 3(a). Case I. Fixed variance

Station 130401

Variables	Prior Mean Value (Observed)	New Mean Value (Observed)	New Mean Value (Assigned)	Entropy (H)	% Error of Predicted Value	
(1)	(2)	(3)	(4)	(5)	(6) = $\frac{(4)-(3)}{(3)}$ x 100%	average
Conductivity	408.8	(i) 339.4	318.6	-0.131	-6.12	6.00
		(ii) 302.3	320.1	-0.127	5.89	
Dissolved Ions	269.7	(i) 199.4	215.9	-0.151	-8.30	10.10
		(ii) 194.8	218.0	-0.139	-11.90	
Dissolved Solids	235.2	(i) 183.4	171.0	-0.256	-6.75	5.30
		(ii) 166.3	172.8	-0.241	3.94	
Hardness	114.7	(i) 86.3	92.3	-0.118	6.91	13.20
		(ii) 78.6	93.9	-0.102	13.47	

Station 130106

Variables	Prior Mean Value (Observed)	New Mean Value (Observed)	New Mean Value (Assigned)	Entropy (H)	% Error of Predicted Value	
(1)	(2)	(3)	(4)	(5)	(6) = $\frac{(4)-(3)}{(3)}$ x 100%	average
Conductivity	232.2	(i) 214.9	213.4	-0.131	-0.72	0.91
		(ii) 211.4	213.7	-0.127	1.11	
Dissolved Ions	188.6	(i) 160.4	166.4	-0.151	3.71	4.02
		(ii) 160.3	167.2	-0.139	4.34	
Dissolved Solids	145.6	(i) 125.2	123.5	-0.256	-1.33	0.76
		(ii) 124.0	124.2	-0.241	0.16	
Hardness	83.1	(i) 64.7	74.2	-0.118	14.67	14.73
		(ii) 65.3	74.9	-0.102	14.79	

Table 3(b). Case II. Non fixed variance

Station 130401

Variables	Prior Mean Value (Observed)	New Mean Value (Observed)	New Mean Value (Assigned)	Entropy (H)	% Error of Predicted Value	
(1)	(2)	(3)	(4)	(5)	(6) = $\frac{(4)-(3)}{(3)}$ x 100%	average
Conductivity	408.8	(i) 339.4	318.6	-0.131	-6.12	6.00
		(ii) 302.3	320.1	-0.127	5.89	
Dissolved Ions	269.7	(i) 199.4	215.9	-0.151	-8.30	10.10
		(ii) 194.8	218.0	-0.139	-11.90	
Dissolved Solids	235.2	(i) 183.4	171.0	-0.256	-6.75	5.30
		(ii) 166.3	172.8	-0.241	3.94	
Hardness	114.7	(i) 86.3	92.2	-0.118	6.84	13.15
		(ii) 78.6	93.9	-0.102	13.47	

Table 3(b) (continued)

Station 130106

Variables	Prior Mean Value (Observed)	New Mean Value (Observed)	New Mean Value (Assigned)	Entropy (H)	% Error of Predicted Value	
(1)	(2)	(3)	(4)	(5)	(6) = $\frac{(4)-(3)}{(3)}$ x 100%	average
Conductivity	232.2	(i) 214.9	213.4	-0.131	-0.72	0.91
		(ii) 211.4	213.7	-0.127	1.11	
Dissolved Ions	188.6	(i) 160.4	166.4	-0.151	3.71	4.02
		(ii) 160.3	167.2	-0.139	4.34	
Dissolved Solids	145.6	(i) 125.2	123.5	-0.256	-1.33	0.72
		(ii) 124.0	124.2	-0.241	0.08	
Hardness	83.1	(i) 64.7	74.2	-0.118	14.67	14.73
		(ii) 65.3	74.9	-0.102	14.79	

Table 3(c). Case III. Fixed variance and shifted mean

Station 130401

Variables	Prior Mean Value (Observed)	New Mean Value (Observed)	New Mean Value (Assigned)	Entropy (H)	% Error of Predicted Value	
(1)	(2)	(3)	(4)	(5)	(6) = $\frac{(4)-(3)}{(3)}$ x 100%	average
Conductivity	408.8	(i) 339.4	323.3	-0.118	-4.74	6.08
		(ii) 302.3	324.7	-0.114	7.41	
Dissolved Ions	269.7	(i) 199.4	220.2	-0.128	10.43	12.27
		(ii) 194.8	222.3	-0.117	14.12	
Dissolved Solids	235.2	(i) 183.4	175.4	-0.222	-4.36	5.46
		(ii) 166.3	177.2	-0.208	6.55	
Hardness	114.7	(i) 86.3	96.6	-0.077	11.94	18.44
		(ii) 78.6	98.2	-0.064	24.94	

Station 130106

Variables	Prior Mean Value (Observed)	New Mean Value (Observed)	New Mean Value (Assigned)	Entropy (H)	% Error of Predicted Value	
(1)	(2)	(3)	(4)	(5)	(6) = $\frac{(4)-(3)}{(3)}$ x 100%	average
Conductivity	232.2	(i) 214.9	204.4	-0.118	-4.89	4.00
		(ii) 211.4	204.8	-0.114	-3.12	
Dissolved Ions	188.6	(i) 160.4	158.1	-0.128	-1.43	1.12
		(ii) 160.3	159.0	-0.117	-0.81	
Dissolved Solids	145.6	(i) 125.2	115.0	-0.222	-8.15	7.42
		(ii) 124.0	115.7	-0.208	-6.69	
Hardness	83.1	(i) 64.7	65.9	-0.077	1.85	1.92
		(ii) 65.3	66.6	-0.064	1.99	

Table 3(d). Case IV. Non fixed variance and shifted mean

Station 130401

Variables	Prior Mean Value (Observed)	New Mean Value (Observed)	New Mean Value (Assigned)	Entropy (H)	% Error of Predicted Value	
(1)	(2)	(3)	(4)	(5)	(6) = $\frac{(4)-(3)}{(3)}$ x 100%	average
Conductivity	408.8	(i) 339.4	323.3	-0.118	-4.74	6.08
		(ii) 302.3	324.7	-0.114	7.41	
Dissolved Ions	269.7	(i) 199.4	220.2	-0.128	10.43	12.27
		(ii) 194.8	222.3	-0.117	14.12	
Dissolved Solids	235.2	(i) 183.4	175.4	-0.222	-4.36	5.46
		(ii) 166.3	177.2	-0.208	6.55	
Hardness	114.7	(i) 86.3	96.6	-0.077	11.94	18.44
		(ii) 78.6	98.2	-0.064	24.94	

Station 130106

Variables	Prior Mean Value (Observed)	New Mean Value (Observed)	New Mean Value (Assigned)	Entropy (H)	% Error of Predicted Value	
(1)	(2)	(3)	(4)	(5)	(6) = $\frac{(4)-(3)}{(3)}$ x 100%	average
Conductivity	232.2	(i) 214.9	204.4	-0.118	-4.89	4.00
		(ii) 211.4	204.8	-0.114	-3.12	
Dissolved Ions	188.6	(i) 160.4	158.1	-0.128	-1.43	1.12
		(ii) 160.3	159.0	-0.117	-0.81	
Dissolved Solids	145.6	(i) 125.2	115.0	-0.222	-8.15	7.42
		(ii) 124.0	115.7	-0.208	-6.69	
Hardness	83.1	(i) 64.7	65.9	-0.077	1.85	1.92
		(ii) 65.3	66.6	-0.064	1.99	

CASE III [Table 3(c)]. The mean values of the prior distribution at station 130106 were decreased by an amount which is estimated from observed changes in the environment of the basin in which that station is located. In this case, the value from every variable was reduced by 10 units. Note that this selection of a value of 10 units was made without reference to the differences between the data in the period (1971-1980) used to define the distributions and the data from the period (1981-1989) used to validate the model, other than knowing the values of the water quality parameter had decreased at the downstream stations and therefore had likely decreased at one or both of the upstream stations. The change in the prior distribution, i.e., reduction in the mean value by a certain number of units, in this case has resulted in the new mean values assigned for some variables being much closer to the observed values and others being further away. Two conclusions may be drawn from these results. In the first instance, no change may have occurred at station 130106 and the change in the prior distribution was inappropriate. The other conclusion might be that the basic model (i.e., no change in the historical distribution) is, in fact, the more robust of the two formulations and should be used in preference to the other, thereby eliminating the need to assess potential changes. However, more work is needed to examine this issue further.

CASE IV [Table 3(d)] is a combination of CASE II and III, in that both the prior values at station 130106 were reduced and changes in variances were permitted. The results do not differ significantly from CASE III, which suggests that the changes of the prior q_{ij} values are the primary mechanism for a change (improvement in the results).

It is apparent, in CASES III and IV in particular, that the method is both accurate in its predictions (assignments of probability distributions) and flexible in a sense that it allows users to make a reasonable assumption on the condition of the sub-basin and to use that assumption in the input in the model through specification of the 'prior' distribution. When good approximations of the prior distributions are available, the model gives, non-unexpectedly, better predictions of the values relative to the observed value. The problem is then how to actually estimate the mean value of the

'prior' distribution, if it is known in advance that such a change in the environment has occurred in the sub basin (i.e., development of a new housing area or a new industrial area). More work is also needed in this area.

The results gained from the method are theoretically (from the fundamental principles behind POME) the most unbiased probability distribution, consistent to the information given in the constraints. Inappropriate or erroneous predictions can occur therefore when incorrect information is included in, or some information is omitted from, the constraints. Therefore, it is critical in the application of this technique to convey the information expressed in the constraints carefully.

In this study it was also assumed that there is no dependency among the data from the upstream stations. This assumption may be to be unrealistic in some, or even in a majority of, cases. An important step in further development of the procedure would be to include and observe the effect of any sum dependency in the model.

8 Summary and conclusion

A new method for the prediction of water quality values at discontinued monitoring stations, given observed changes at a downstream station is proposed. The method is based upon entropy theory and, more specifically, on the Principle of Maximum Entropy and uses observed changes in the distribution of water quality values at a downstream station to assign distributions to the water quality values at stations on upstream tributaries or on the upstream mainstem of the stream.

The distributions of water quality provided by the methodology for the upstream points are an unbiased prediction of the new conditions at those locations. As such the method has the ability to make unbiased predictions of the upstream locations of causes of observed changes of water quality value at the downstream station. In the case where changes in the distribution of water quality at upstream station(s) are known to have occurred but where those changes are not sufficiently large to account for all the observed changes at the downstream station, the method can be modified to incorporate the prior information about the system in the assignment of the new probabilities of water quality values.

An evaluation of the method was undertaken by applying the procedure to the prediction of water quality values in the Mackenzie River, Queensland, Australia. In this evaluation one set of the available data was used to calibrate the procedure by developing the prior distributions of water quality at each station. The other set was used to compare the prediction of the entropy model with the observed values. It was found that, in spite of the data used for comparison being significantly different from those in the calibration step, i.e., both the downstream and upstream water quality values were quite different in the two sections of data, the entropy model provided predicted water quality values at the upstream station that were remarkably close to the observed values. These results indicate the potential of the method for predicting water quality values at discontinued stations and for identifying the locations of causes of unaccounted changes in the distribution of water quality values observed at downstream stations.

References

Amorocho, J.; Espildora, B. 1973: Entropy in the assessment of uncertainty of hydrologic systems and models, Water Resour. Res. 9(6), 1515-1522

Awumah, K.; Goulter, I; Bhatt, S. 1990: "Assessment of Reliability in Water Distribution Networks Using Entropy Based Measures", Stochastic Hydrology and Hydraulics 4, 309-320

Awumah, K.; Goulter, I.; Bhatt, S. 1991: "Entropy Based Redundancy Measures in Water Distribution Network Design", Journal of Hydraulic Engineering, ASCE 117(10), 595-614

Chiu, C.-L. 1987: "Entropy and Probability Concepts in Hydraulics" Journal of Hydraulics Engineering, ASCE 113(5), 583-600

Chiu, C.-L. 1988: "Entropy and 2-D Velocity Distribution in Open Channels", Journal of Hydraulics Engineering, ASCE 114(7), 738-756

Chiu, C.-L. 1989: "Velocity Distribution in Open Channels Flow" Journal of Hydraulics Engineering, ASCE 115(5), 576-594

Chiu, C.-L. 1991: "Application of Entropy Concept in Open Channel Flow Study", Journal of Hydraulics Engineering, ASCE 117(5), 615-628

Chiu, C.-L.; Chiou, J.D. 1986: "Structures of 3-D Flow in Rectangular Open Channels" Journal of Hydraulics Engineering, ASCE 112(11), 1050-1068

Dunnette, D.A. 1980: "Sampling Frequency Optimisation Using a Water Quality Index", Journal of Water Pollution Control Federation 52(11), 2807-2811

Harmancioglu, N.B. 1981: "Measuring the Information Content of Hydrological Processes by the Entropy Concept", Centennial of Ataturk's Birth, Journal of Civil Engineering, Faculty of Ege Univ., 13-38

Harmancioglu, N.B. 1984: "Entropy Concept as used in Determination of Optimum Sampling Intervals", Proceedings of Hydrosoft '84, International Conference on Hydraulic Engineering Software, Portoroz, Yugoslavia, 6-99

Harmancioglu, N.B.; Alpaslan, N. 1992: "Water Quality Monitoring Network Design: A Problem of Multi-Objective Decision Making", Water Resources Bulletin, A.W.R.A. 28(1), 179-192

Harmancioglu, N.B.; Baran, T. 1989: "Effects of Recharge System on Hydrologic Information Transfer Along Rivers", IAHS, Proc. of the Third Scientific Assembly of the International Association of Hydrologic Science - New Direction for Surface Water Modelling, IAHS Publ. 181, 223-233

Harmancioglu, N.B.; Yevjevich, V. 1987: "Transfer of Hydrologic information Among River Points", Journal of Hydrology, 91, 103-118

Jaynes, E.T. 1983: "Papers on Probability Statistics and Statistical Physics", D.Reidel Publishing Company, Dordrecht, Holland, 434 p

Kapur, J.N. 1989: "Maximum Entropy Models in Science and Engineering", Wiley Eastern Ltd., New Delhi, India, 635 p

Kusmulyono, A.; Goulter, I. 1994: "Computational Aspects in Use of Entropy Theory in Predicting Water Quality Levels at Discontinued Stations", Stochastic Hydrology and Hydraulics (This issue)

Lasdon, L.S.; Warren, A.D. 1986: "GRG2 User's Guide", Department of General Business Administration, The University of Texas at Austin, Austin, Texas, 60 p

Loftis, J.C.; Ward, R.C. 1980: "Water Quality Monitoring-Some Practical Sampling Frequency Consideration" Environmental Management 4(6), 521-526

Palmer, R.N.; MacKenzie, M.C. 1985: "Optimization of Water Quality Monitoring Networks", Journal of Water Resources Planning and Management 111(4), 478-493

Shannon, C.E. 1948: "A Mathematical Theory of Communication", Bell System Technical Journal 27(3), 379-423, 623-659

Sonuga, J.O. 1972: "Principle of Maximum Entropy in Hydrologic Frequency Analysis", Journal of Hydrology 17, 177-191

Sonuga, J.O. 1976: "Entropy Principle Applied to the Rainfall-Runoff Process", Journal of Hydrology 30, 81-94

Ward, R.C.; Loftis, J.C.; McBride, G.B. 1986: "The 'Data-rich but Information-poor' Syndrome in Water Quality Monitoring", Environmental Management 10(3), 291-297