Information theory applied to evaluate the discharge monitoring network of the Magdalena River

Leonardo Alfonso, Liyan He, Arnold Lobbrecht and Roland Price

ABSTRACT

The acquisition of good hydrologic information is an important issue in water management since it is the basis of decisions concerning the allocation of water resources to different users. However, sufficient data are often not available to describe the behaviour of such systems, especially in developing countries, where monitoring networks are inappropriately designed, poorly operated or are inadequate. Therefore, it is of interest to design and evaluate efficient monitoring networks. This paper presents two methodologies to design discharge monitoring networks in rivers using the concepts of Information Theory. The first methodology considers the optimization of Information Theory quantities and the second considers a new method that is based on ranking Information Theory quantities with different possible monitor combinations. The methodologies are tested for the Magdalena River in Colombia, in which the existing monitoring network is also assessed. In addition, the use of monitors at tributaries is explored. It is demonstrated that the ranking method is a promising way of finding the extremes of Pareto fronts generated during multi-objective optimization processes and that better (more informative and less redundant) monitoring network configurations can be found for the Magdalena River.

Key words | hydrodynamic model, information theory, Magdalena River, monitoring networks, total correlation, wetlands

Leonardo Alfonso (corresponding author)
Arnold Lobbrecht
Roland Price
Hydroinformatics Core,
Integrated Water Systems and Governance Department,
UNESCO-IHE,
P.O. Box 3015 2601,
DA Delft,
The Netherlands
E-mail: *l.alfonso@unesco-ihe.org*

Liyan He
DHI China-Inland Water Department, 5th,
No. 303, Xiao Muqiao Road,
Shanghai, 200032,
China

Arnold Lobbrecht
HydroLogic BV,
Stadsring 57, 3811 HN, Q1
Amersfoort,
The Netherlands,
Postbus 2177 3800 CD

INTRODUCTION

The acquisition of hydrologic information is an important issue in water management. Decisions concerning the allocation of water resources should be made with adequate information (Loucks *et al.* 2005). However, sufficient data are often not available to describe the behaviour of such systems. In fact, an evident decline in the amount of river flow data has been noticed in recent years due to reasons that include lack of resources, weak institutional structures and ignorance about how useful the data are (Sene & Farquharson 1998). In particular, in developing countries monitoring networks are often inappropriately designed (gauges are sited at unsuitable locations), poorly operated (time intervals are too long or too short for proper measurement) or inadequate (lack of sufficient gauges or a robust communication system). Under these conditions, water-related decisions may bring negative impacts on the environment, the regional economy and society (WMO 2003).

From a practical perspective, a number of recommendations for discharge gauging location exist (WMO 2008), including considerations such as having a regular and stable river bed, parallel velocities at each point of a cross section and avoiding curved reaches, strong backwater effects, flow bifurcations and aquatic growth. In addition, there exist a number of methods that deal with the design and evaluation of monitoring networks for streamflow data. Some of the most common methods are statistical (see e.g., Moss & Karlinger 1974; Moss & Tasker 1991), entropy-based (e.g., Husain 1989; Yang & Burn 1994), and methods that include direct surveys to assess users needs (e.g., Davar & Brimley 1990). A recent, comprehensive review of available methods for evaluating monitoring

doi: 10.2166/hydro.2012.066

networks is presented by Mishra & Coulibaly (2009). Recent studies have also addressed the issue of model calibration (Sun & Bertrand-Krajewski 2012) and the set up of measuring campaigns (Freni & Mannina 2011).

This paper presents original research to locate flow monitors based on the potential information content that a set of geographical locations can provide, following concepts from Information Theory, which were first introduced in the field of water resources by Amorocho & Espildora (1973). Entropy theories have been used to estimate discharges in a river cross section (see e.g., Chiu & Chen 2003) and to evaluate velocity distributions in pipes (Chiu *et al.* 1993). A review of applications of the theory was made by Singh (1997). For some recent research studies, see e.g., Alfonso *et al.* (2010b), Maruyama *et al.* (2005), Mishra *et al.* (2009) and Ruddell & Kumar (2009).

The paper is organized as follows. First, Information Theory is reviewed briefly, including a numerical example to clarify the concept of Total Correlation. Then, the methodology section is introduced by a brief review on the design and evaluation of monitoring networks, which is followed by a description of a multi-objective optimization method and a rank-based algorithm. Next, the case study of the Magdalena River, Colombia is introduced, followed by the analysis of results obtained from the application of the methodology. The Results section also includes the assessment of the existing monitoring stations for the river and the assessment of a possible monitoring configuration that takes into account the location of the tributaries to the river. The methods are then compared, including a sensitivity analysis of a parameter used in the entropy estimation. Finally, conclusions and recommendations are presented.

Information theory

Shannon (1948) developed Information Theory to measure the information content of random variables. To understand the informational aspect of entropy, suppose there is a set of N events. Uncertainty is when we do not know which of the N events will happen. The degree of uncertainty can be different according to one's knowledge about the events. Once an event occurs and we observe it, our uncertainty decreases, as we receive some information. For this reason information can be regarded as a decrease in uncertainty.

According to Shannon & Weaver (1949), the entropy of a discrete random variable X, which has discrete values x_1, $x_2, \ldots, x_n$ with probabilities $p(x_1)$, $p(x_2), \ldots, p(x_n)$, where n is the number of elementary events, is given by:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\ \log p(x_i) \tag{1}$$

The amount of information content between two random variables is of interest. For stochastically independent random variables X_1 and X_2, the total entropy is $H(X_1) + H(X_2)$. However, if X_1 and X_2 are stochastically dependent, part of the information is shared by them and the total entropy is not the simple sum of individual entropies. Instead, the Joint Entropy is defined by:

$$H(X_1, X_2) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_{1i}, x_{2j})\ \log p(x_{1i}, x_{2j}) \tag{2}$$

where $p(x_{1i}, x_{2j})$ is the joint distribution between variables X_1 and X_2.

However, natural processes are typically influenced by a significant number of variables, and a way to understand these processes is to look at the relationships between the variables. The assessment of the dependencies among a set of variables can be carried out by means of the concept of Total Correlation (McGill 1954; Watanabe 1960), which gives the amount of information shared by all variables at the same time, taking into account all the dependencies between their partial combinations:

$$C(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i) - H(X_1, X_2, \ldots, X_N) \tag{3}$$

The sum of the entropies of all the variables is always no less than their Joint Entropy, so the total correlation can only be non-negative. If no dependency exists between them, then the total correlation is zero.

The main difficulty with Equation (3) is that it requires the estimation of the Joint Entropy of multiple variables $H(X_1, X_2, \ldots, X_N)$, which in turn needs a proper estimation of the joint distribution $p(x_1, x_2, \ldots x_N)$, a difficult task for a big value of N. This drawback, however, can be resolved using the grouping property of mutual information

(Kraskov *et al.* 2003), in which new variables are built up by agglomerating pairs of variables in such a way that the entropy of each new variable is equivalent to the Joint Entropy of the original pair. This agglomeration can be achieved by putting the corresponding records of the two original variables together, forming a new set of symbols. The Total Correlation is calculated by summing up all those pair-wise joint entropies obtained for each built variable.

In order to clarify the procedure, a numerical example to calculate the Total Correlation of the time series X, Y and Z is presented in Table 1a. In order to estimate the probabilities, the records are transformed into symbols that will be used in a frequency analysis by the quantization method discussed in the next section. For the scope of this example, the floor function will be used to convert each value in X, Y and Z into a corresponding integer Xt, Yt, Zt (Table 1b). The probability of occurrence of a particular symbol i in each series Xt, Yt, Zt is presented in Table 1c.

From Equation (1), $H(Xt) = 1.357$, $H(Yt) = 2.171$, $H(Zt) = 2.522$ bits. Next, a new variable A is produced by agglomerating the variables Xt and Yt in such a way that each unique pair of symbols (xt, yt) produces a unique symbol xy. This can be achieved by the expression $a = 10x + y$. Every record a_i in A will have an associated probability of occurrence within the agglomerated time series, (Table 2a). From Equation (1), $H(A) = 2.922$ bits, which is exactly the same value that would be obtained with Equation (2). In other words, $H(A) = H(X,Y)$. Now consider agglomerating time series A and Z, producing a new variable B, using the expression $b = 10a + z$. The variable and the associated probabilities are presented in Table 2b.

From Equation (1), $H(B) = 3.322$ bits, which is equivalent to the Joint Entropy $H(A, Z) = H(X, Y, Z)$. Finally, from Equation (3),

$$\begin{aligned} C(X, Y, Z) &= H(X) + H(Y) + H(Z) - H(X, Y, Z) \\ &= 1.357 + 2.171 + 2.522 - 3.322 \\ &= 2.728 \text{ bits} \end{aligned}$$

This means that 2.728 bits of information are shared among the three variables, under all possible degrees of interaction (X and Y, X and Z, Y and Z and X, Y and Z). Although the length of the time series in the example is insufficient to adequately calculate probabilities, they are used to provide clarity on the method.

Quantization

Although there exist a number of nonparametric methods to estimate mutual information (see e.g., Moon *et al.* 1995), in this paper the expressions (1) to (3) are estimated using a histogram-based frequency analysis. For this purpose, the concept of quantization, a procedure of constraining a continuous set of values to a discrete set is used. From the mechanical standpoint, the quantization rounds a value x to its nearest lowest integer multiple of a namely x_q to give:

$$x_q = a\left\lfloor \frac{2x + a}{2a} \right\rfloor \tag{4}$$

Table 1 | Simplified example of Total Correlation calculation

a. Variables			b. Transformed variables			c. Probabilities		
X	***Y***	***Z***	***Xt***	***Yt***	***Zt***	***P(Xt$_i$)***	***P(Yt$_i$)***	***P(Zt$_i$)***
3.24	2.20	5.45	3	2	5	0.1	0.3	0.2
4.25	2.08	4.18	4	2	4	0.1		0.2
5.30	1.15	2.37	5	1	2	0.7	0.1	0.2
5.35	4.81	2.02	5	4	2		0.3	
5.45	5.40	3.01	5	5	3		0.2	0.1
5.70	4.36	4.75	5	4	4			
6.55	4.60	6.85	6	4	6	0.1		0.2
5.42	5.21	6.04	5	5	6			
5.40	3.13	5.32	5	3	5		0.1	
5.25	2.91	7.99	5	2	7			0.1

Table 2 | Example of agglomeration of variables for Total Correlation calculation

A	P(Ai)	B	P(Bi)
32	0.1	325	0.1
42	0.1	424	0.1
51	0.1	512	0.1
54	0.2	542	0.1
55	0.2	553	0.1
54		544	0.1
64	0.1	646	0.1
55		556	0.1
53	0.1	535	0.1
52	0.1	527	0.1

It is well known that entropy-related quantities are sensitive to the size of the bins used in the frequency analysis and this is also the case with the parameter a. For this reason, a sensitivity analysis is included below for the case study. The selection of the parameter a is explained when introducing the case study.

METHODOLOGY

Evaluation of monitoring networks

From the information point of view, an optimal monitoring network should satisfy at least two basic objectives: maximum information content at each gauging site and minimum dependency between monitoring points. The first objective can be achieved by maximizing the Joint Entropy of the selected gauges (see Equation (2) for the case of N variables), which represents the amount of information that is potentially contained in the network. A number of authors have used the concept of transinformation (see e.g., Krstanovic & Singh 1992; Mogheir & Singh 2002; Mogheir *et al.* 2006) or normalized versions (Yang & Burn 1994; Alfonso *et al.* 2010b) to evaluate the second objective. In this paper the concept of Total Correlation, Equation (3), is used as a measure of dependency.

Although the ideal monitoring network would be composed of a set of gauges that provide the maximum information content and that are able to capture independent information, Total Correlation and Joint Entropy are conflicting objectives: when the Joint Entropy of a set of variables increases, the Total Correlation decreases and vice versa. For this reason, the best location of N monitors would be such that they simultaneously fulfil both objectives. The mathematical formulation of the optimization problem is:

$$\begin{aligned} &\min\{C(X_1, X_2, \ldots, X_N)\} \\ &\max\{H(X_1, X_2, \ldots, X_N)\} \end{aligned} \tag{5}$$

where the decision variables $X_1, X_2, \ldots, X_N$ are the geographical locations of N gauges. A simplification of the problem can be graphically seen as selecting from Figure 1 the best three circles such that their contained area is the biggest and at the same time that their overlapped area is the smallest.

The optimization problem posed in Equation (5) is solved using Multi-objective Optimization with Genetic Algorithms (MOGA). A rank-based greedy algorithm that optimizes both objectives independently is also provided in order to compare results. Both approaches are described below.

Multi-objective optimization approach

One way of solving the problem is to pose it as a multi-objective optimization problem (MOOP), whose output consists of sets of quasi-optimal, non-dominated solutions that define a Pareto front. This front describes what can be achieved in terms of decisions, showing how an

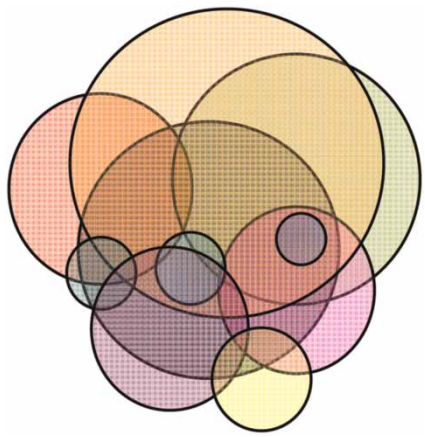

Figure 1 | Graphical, simplified version of the gauge-selection problem. Adapted from Alfonso *et al.* (2010c).

improvement in one criterion leads to a deterioration in other criteria. The MOOP consists of looking for such a set of decision variables that simultaneously satisfy constraints and optimize the values of the objective functions. In general, these objective functions conflict with each other, and therefore a solution that fully satisfies all the objectives at the same time may not exist. In this context, to optimize means to find a compromise among the objectives.

Mathematically, a vector of decision variables X^* is optimal if there is no other vector X that improves an objective without degrading another objective. This means that X is not better than X^* or, in other words, X^* is a non-dominated solution. Mathematically, X is Pareto-optimal if either $f_i([X]) \geq f_i([X^*])$, or for at least one $i = \{1, 2, \ldots, k\}$ $f_i([X^*]) < f_i([X])$, where k is the number of total objectives, X and X^* are vectors of decision variables, and f are objective functions.

In order to provide an example of the concept, consider the simplified problem schematized in Figure 1. The final result of MOOP would be a set of solutions that would include at least the set of circles presented in Figure 2A, B and C. These solutions are non-dominated, because there is no other set of three circles that increment the contained area (row 2) without increasing the overlapped area (row 3).

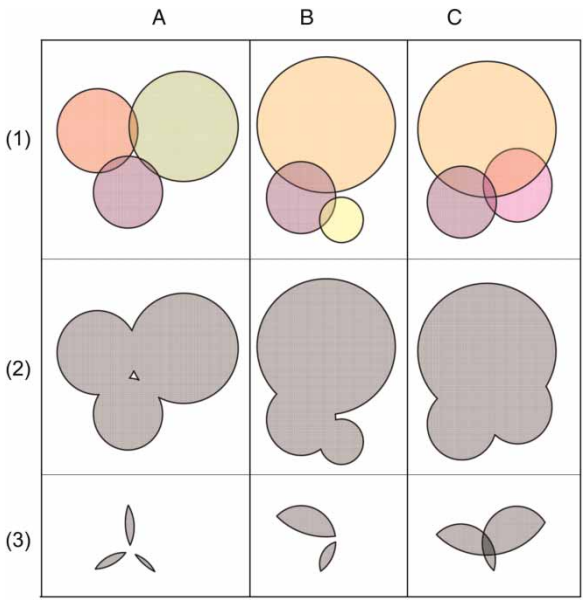

Figure 2 | Solutions of the problem in Figure 1 using MOOP. Adapted from Alfonso *et al.* (2010c).

MOGA has been successfully used to solve water-related optimization problems (see e.g., Barreto *et al.* 2009; Alfonso *et al.* 2010a). A series of experiments were carried out to identify the optimal set of points for monitoring. In this paper, NSGA-II, an elitist, non-dominated sorting genetic algorithm for multi-objective optimization (Deb *et al.* 2002) is used, which utilizes Simulated Binary Crossover (SBN) and Polynomial Mutation as genetic-related operations. SBN consists of a weighted average between parents, which allows the use of single-point binary-coded crossover principle in real-coded problems. Polynomial mutation is a real-coded method that consists of generating a child by varying a parent with an amount that is proportional to the product of the difference between the maximum and minimum limits of the variable under study and a random factor with a polynomial probability distribution. The reader should see Deb *et al.* (2002) and Deb & Agrawal (1994) for further details.

Rank-based greedy algorithm

A greedy algorithm that picks the best information-related gauge at a time is developed next. The idea is to rank all the potential places to locate the monitors according to the separately considered variations in Joint Entropy and in Total Correlation caused by the selection of a new monitor. The algorithm requires the location to be selected for the first monitor. This could be for the monitor with the highest reduction in uncertainty, as suggested by Krstanovic & Singh (1992). However, starting with the monitor that has the highest information content does not guarantee that the final set of monitors is the most informative, as is shown by the results below. The second monitor is then chosen from the remaining set of monitors in such a way that it provides either the highest increment in Joint Entropy or the lowest increment in Total Correlation with respect to the first monitor.

The solution to the problem in Figure 1 using the rank-based method is presented in Figure 3, where the dashed circle is considered as the starting point of the algorithms to maximize the contained area (A), and to minimize the overlapped area (B). These solutions can be seen to be the extremes of the Pareto set obtained with the MOOP approach.

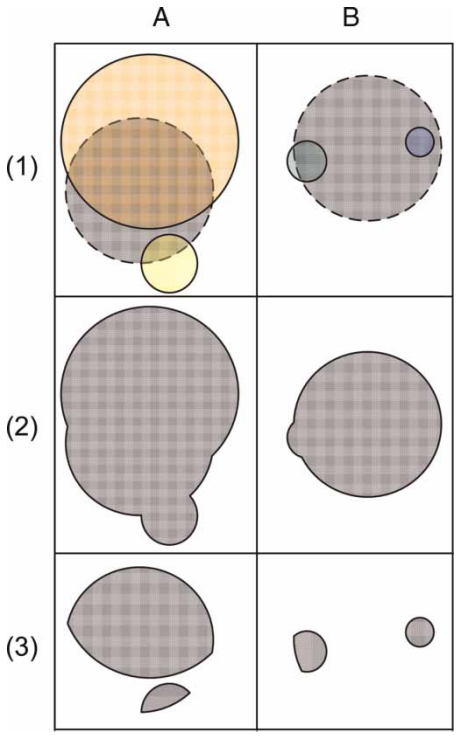

Figure 3 | Solutions of the problem in Figure 1 using greedy algorithm independently for maximizing covered area (A) and for minimizing overlapped area (B). The dashed circle represents the starting point of the algorithm.

For the remainder of this paper the method that represents case A will be named Joint Entropy Ranking, while the method that represents case B will be named Total Correlation Ranking. The flowcharts of the algorithms to be used in both situations are shown in Figure 4.

THE MAGDALENA RIVER

The multi-objective optimization method and the rank-based algorithm are applied to the monitoring network of the Magdalena River, the main river of Colombia, which runs for about 1,540 km from South to North through the western half of the country to the Caribbean Sea. About 70% of the Colombian population lives in this catchment.

Tributaries on the Magdalena River play an important role in many respects. The watershed and its main tributaries cover 257,400 km^2, which corresponds to 24% of the total surface of the national territory. The main tributaries, located mainly in the middle reach, have an important influence on the river's behaviour in terms of discharge. Figure 5 presents the location of the most important

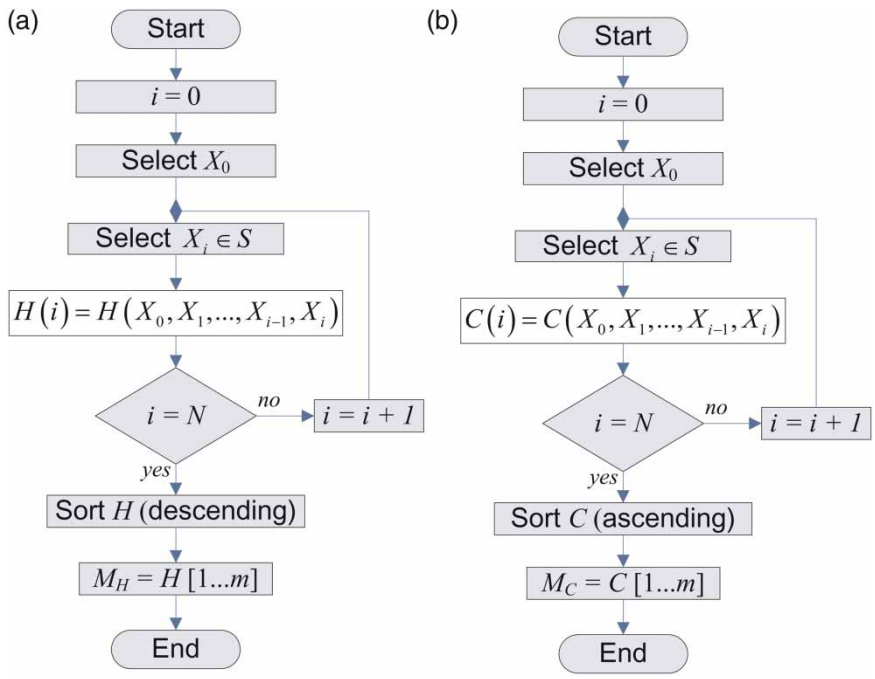

Figure 4 | Flowchart rank-based greedy algorithm for Joint Entropy (a) and Total Correlation (b).

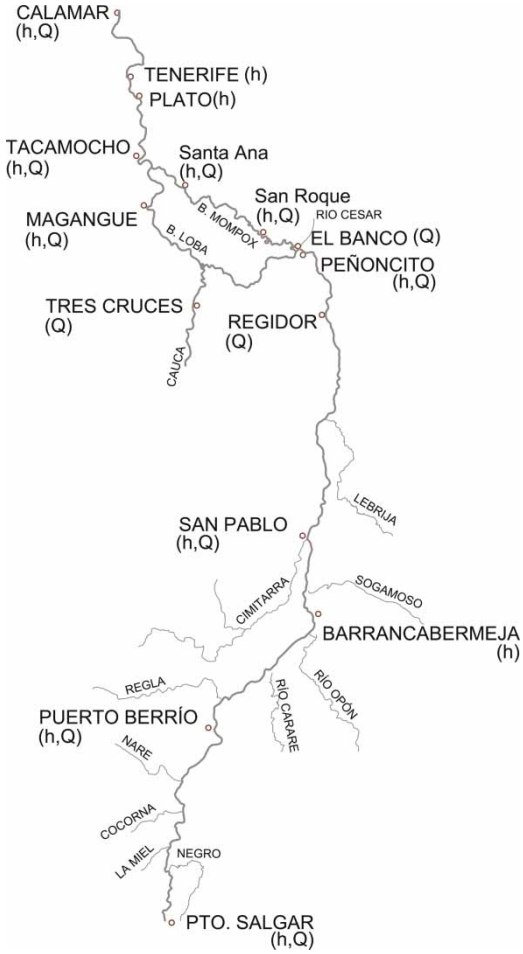

Figure 5 | Available hydrologic data records of discharges (Q) and water levels (h) at river stations for 1995.

cities near the river, the main tributaries and the available discharge and water level records for 1995 for the downstream half of the river.

Wetlands

In terms of slope, the Magdalena River is divided into three main regions, namely steep, medium and gentle. In the division between the medium and gentle regions, the river suffers an abrupt slope change. This transition, together with the particular geologic, geomorphologic and drainage characteristics of the area, generates an inner delta that consists of several hundred interconnected water bodies. These wetlands are also connected to the main river, with important exchanges of sediment and water. The area is a natural reservoir that absorbs the peak flows of the river. Additionally, the river divides into two branches called the Loba Branch (the main one) and the Mompox Branch.

The existing water-level gauges for the river were placed initially to support decision-making concerning local problems in the main populated areas, related to flood control and navigation, while keeping operation and maintenance costs low. However, from a global perspective, the information collected by these gauges is limited to supporting decision and policy-making for navigation, flood control and other issues at other points of the river. Therefore, insight is needed on the design of a new monitoring network while evaluating the existing network in terms of its information content.

Model-generated data

A 1D-hydrodynamic model was built in order to generate the time series of discharge to be used for the information theory analysis. The model includes 811 km of the river, from Puerto Salgar to Calamar, discretized by 4,058 computational points placed approximately every 200 m. Similarly, the Mompox branch is a natural diversion of 216 km simulated by 1,080 computational points. The hydrological data used corresponds to 1995, for which the most complete data records at the tributaries and hydrologic stations on the river were found. Small tributaries (Velazquez, Palagua, Ermitaño, Baul, Diña Juana, Pontoná, Claro del Sur, Balcanes and Pescado), for which no data are available and are therefore not included in the model, in total contribute about 125 m^3/s on average, which represents only 1.7% of the river's total discharge. Figure 6 shows the historic daily mean discharges of the tributaries to the Magdalena River and the mean discharges for the tributaries during the year 1995. These were included as boundary conditions for the model tributaries (Negro, Carare, Opón, Sogamoso, Lebrija, Miel, Cocorná, Nare, Regla, Cimitarra, Cesar and Cauca). The upstream boundary condition is the discharge series at Salgar and the downstream boundary is the water level series at Calamar. Discharges obtained by means of rating curves at the stations Berrío, Regidor, Peñoncito, El Banco, Magangué, Tacamocho, and Santa Ana were used to calibrate the model by reproducing water levels for the stations of Berrío, Barranca, San Pablo, Regidor, El Banco, Magangué, Tacamocho, Plato, Tenerife and Santa Ana. The model runs with a time step of 10 min.

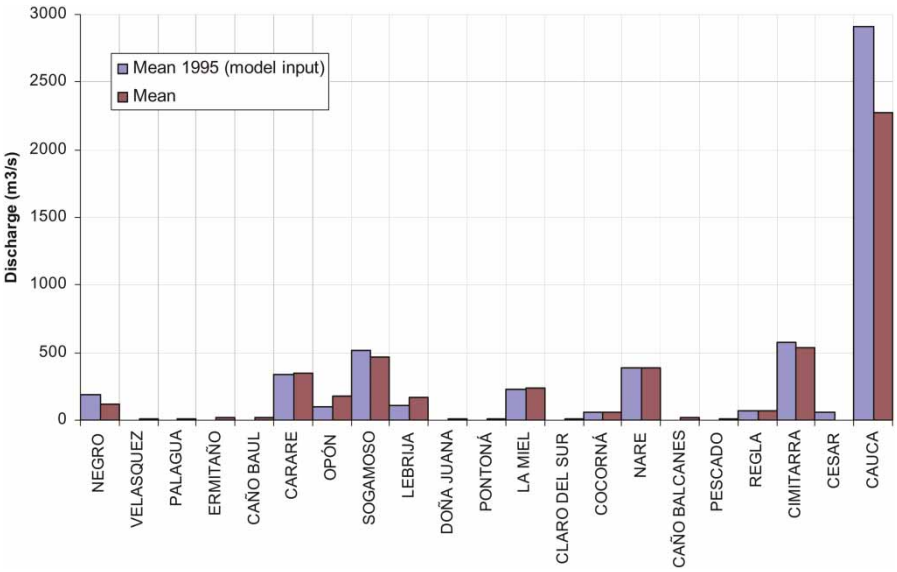

Figure 6 | Mean discharge of tributaries of the Magdalena River and mean discharges for the year 1995, used as model inputs.

The cross section information corresponds to bathymetries surveyed during the year 2000, associated flood plains were added to complement the cross sections using satellite elevation data HydroSHEDS (Lehner *et al.* 2006). It must be noted that these 5 years of difference between bathymetry information and hydrologic information imply additional uncertainty in the discharge outputs. However, the quantization method allows for using these outputs so that comparable performance of the method can be obtained.

The wetland system was divided into four main areas modelled as large offline storages that are connected to the river. Although in reality the river-wetland connections occur through small natural canals, these connections were modelled as bidirectional weirs due to the scarcity of data. However, the use of weirs proved to simulate the hydrodynamic effects of these connections fairly well. The Appendix (available online at http://www.iwaponline.com/jh/015/066.pdf) shows details of the calibration results for the stations San Pablo and El Banco, the latter being affected by an upstream connection to wetland W2. A detailed description of the modelling and calibration process of the Magdalena River model, as well as its limitations can be found in Alfonso (2010, p. 51).

The selection of the parameter a in Equation (5) is done in such a way that all of the tributaries included in the model have a significant information content. That is, a big value of a would make small discharges insignificant in terms of information content because the rounding nature of the expression would transform the discharge series to a series with constant values. On the other hand, a too small value of a would make every discharge in the river have a similar information content, which is not convenient for our analysis. For these reasons, a was assumed to have a value of 200 m^3/s in order to include all the available discharges of the tributaries. A sensitivity analysis to show how results may change because of the selection of the parameter a is included at the end of this paper.

RESULTS

The model output includes discharge time series for 181 calculation points along the main river (Loba branch) and 31 points for the Mompox branch. These time series are quantized using Equation (4), adopting a value of 200 m^3/s for a, which corresponds to the mean discharge of the smallest tributary with available data; so only the effects of inputs with this magnitude are captured by the entropy analysis.

The pre-design of the monitoring network for the Magdalena River was done using Information Theory to select a limited number of points from the 181 discharge points on the main channel where gauge devices are worth placing.

As a first insight, the marginal entropy of each calculation point was estimated using Equation (1), and a map of the entropy for the Magdalena River was prepared (Figure 7).

Analysis of the entropy map

Before presenting the solutions of Equation (6) for the design of the monitoring network using the methods described above, the entropy maps (Figure 7(a)) obtained for the discharge time series are compared with the map of discharges (Figure 7(b)). Firstly, entropy increases at points where the tributaries discharge into the river (see for example the rivers Miel, Negro, Nare, Sogamoso, Cimitarra, Cauca and the convergence of the branches Mompox and Loba in Figure 4(b)). The rivers Opón and Carare do not show any increment in entropy, due to their relatively low influence in terms of discharge. The Mompox branch shows the same entropy along its channel, because there are no tributaries that flow into it. It is interesting to see that the lowest value of entropy occurs in this branch. This is because the discharge in this branch ranges from 400 to 1,000 m^3/s, so when applying Equation (5) with $a = 200\ m^3/s$ the resulting quantized series have only four unique values in the frequency analysis and therefore only four sums are required to evaluate Equation (1).

Secondly, entropy decreases when the wetlands interact with the river. As mentioned above, the wetlands act as a complex system of reservoirs that absorb the peak flows of the Magdalena River. For this reason, the discharge time series tends to be smooth, the range between minimum and maximum discharge is lowered and therefore entropy diminishes. Indeed, entropy is continuously increasing from upstream to downstream, until the wetland W_1, just before the inflow of the Lebrija River. From this point to El Banco, entropy remains constant because no additional inflows exist. However, a big change in entropy takes place at El Banco, reducing it to the values reported after the inflows of the rivers Miel and Negro. This change is due to the connection of the Magdalena River to the wetland W_2 (Figure 4), and the bifurcation of the main river into the Mompox and the Loba branches. It is clear that, after the bifurcation, the Loba branch contains, on average, a similar flow to that at San Pablo, or about 3,800 m^3/s, highlighting again the effects of the wetlands. On the other hand, the effect of the wetland W_3 on the entropy map is opposite that which is observed due to the first wetlands W_1 and W_2, since it adds entropy to the river. This effect is due to a third, minor bifurcation called Chicagua (not shown in Figure 7), which interacts with the so-called Momposine Depression, the 'island' formed by the Mompox and Loba

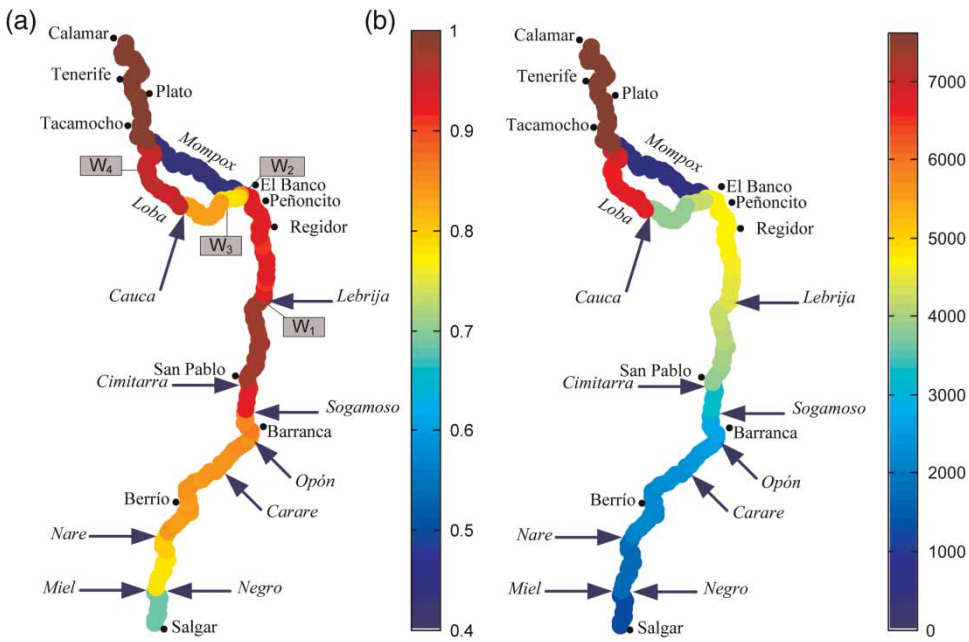

Figure 7 | Entropy map for $a = 200\ m^3/s$ in bits (a) and mean discharge map for 1995 in m^3/s (b), for the Magdalena River.

branches, in which water is discharged during high water level events in the river. For the year 1995, the inflow was mainly from the depression to the Magdalena and therefore it was acting, on average, as an additional tributary. Although this additional discharge is not significant for the river (see Figure 7(b)), between El Banco and Cauca inflow), it makes a difference in terms of entropy (see Figure 7(a)).

Thirdly, in the middle of the Loba reach, the biggest tributary, the Cauca River, flows into the Magdalena, greatly increasing its flow and also its entropy, which recovers some of the entropy absorbed by the wetlands W_1 and W_2. Additionally, although the wetland W_4 acts in a similar way to W_1 and W_2, its effect is not apparent in the discharge map or the entropy map, because the influence of this zone is driven by the significant inflow from the Cauca River.

Finally, at the point where the branches Loba and Mompox converge, the entropy is a maximum. As there are no additional inflows or wetlands, this value remains constant until the most downstream point in Calamar.

Results using multi-objective optimization approach

The MOOP posed in Equation (5) is solved using the Non-Dominated Sorted Genetic Algorithm, NSGA-II (Deb *et al*. 2002), for which the number of population individuals and the number of generations must be specified, as well as the number of decision variables (number of monitors to be placed along the river). From the practical viewpoint, each decision variable is an integer between 1 and the number of computational points of the model (181 in this case), and therefore the decision variables could be transformed into chromosomes of 8 bits. However, as NSGA-II with real-coding of variables was used, the decision variables were rounded up to the nearest integer in order to get N numbers between 1 and 181, each one being related to a particular geographical location with its own time-series.

In order to perform a sensitivity analysis of these parameters, a number of experiments were carried out, in which five different populations (P) and generations (G) were tested with the following combinations (P, G): (50, 20), (50, 50), (100, 20), (100, 50), (200, 50), a task that was carried out for a number of gauges from 6 to 9. The probabilities of crossover and mutation were fixed at 90 and 10% respectively, following common values used in literature. A sensitivity analysis of these parameters to evaluate the efficiency of the algorithm for this particular problem is ongoing research and will be reported elsewhere. The final solution was determined by selecting the best solutions (those with high Joint Entropy and low Total Correlation) from the five obtained Pareto fronts. For comparison purposes, these solutions have been included in each Pareto front in Figure 8 as black dots.

From Figure 8 it can be observed that the increment in the number of decision variables translates into a small increment in joint information and into a significant increase in redundant information. This means that new monitors will not add much more information content compared to what can be deduced from fewer monitors. This can be equivalent to the selection of more than three circles in the problem presented in Figure 1, which implies that a large overlapping area is added with a marginal increment in the covered area.

Figure 9 presents the locations of the solutions A, B, C and D shown in Figure 5 for the case of nine decision variables on the entropy map of the Magdalena River. The redundancy of the solutions is evident, especially in the upstream part of the river. Additionally, Figure 10 presents the location of the solutions with the highest value of Joint Entropy for six, seven, eight and nine monitors.

Several conclusions can be drawn from Figure 9. First, in general, the monitors are located where significant changes in entropy take place. Second, redundancy is reduced by adding monitors upstream, which do not add extra information content. This is equivalent to selecting the three smallest circles in Figure 1, for which there is no overlap. In contrast, the joint information increases as more downstream monitors are added, with the consequent increment in their dependency on each other. This confirms the trade-off between both information measurements H and C. Finally, monitors are always selected at the Momposine Depression, especially where the wetlands have connections with the Magdalena River and where the Cauca River discharges to it. The complex hydraulic conditions ensure that the discharge changes continuously along the river, leading to the increase of information content.

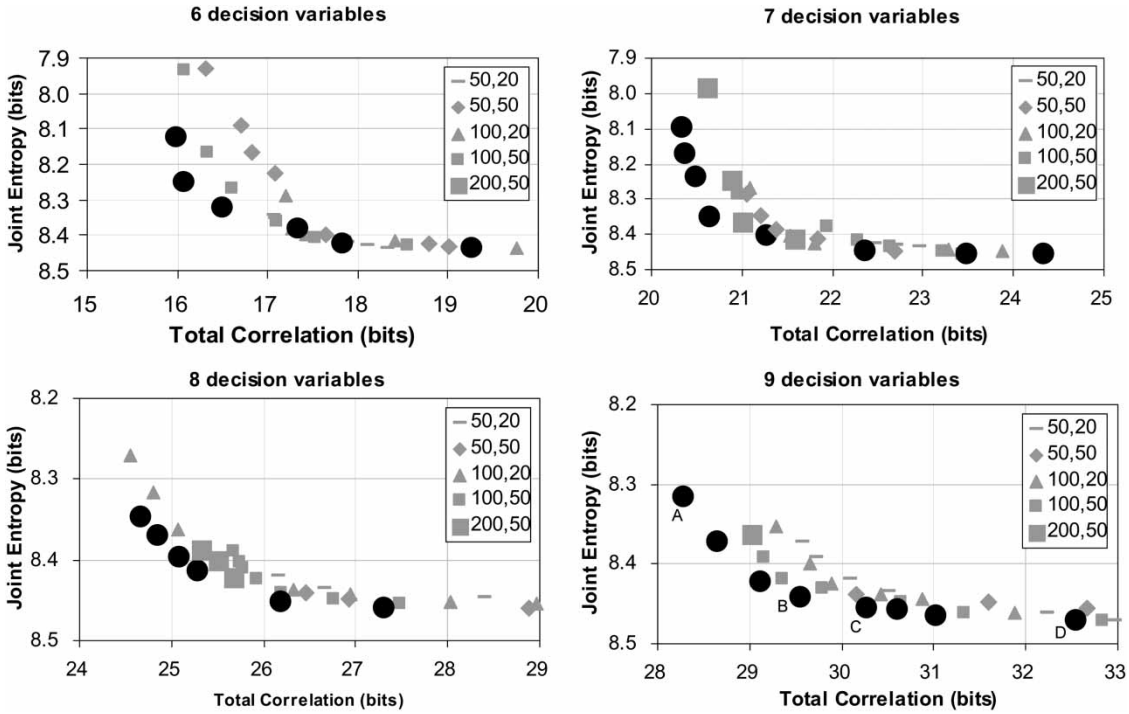

Figure 8 | Solutions for multi-objective optimization approach. Black dots form the best Pareto front obtained by selecting the best points of the five combinations (P, G). Points A, B, C and D are selected for further analysis for nine decision variables.

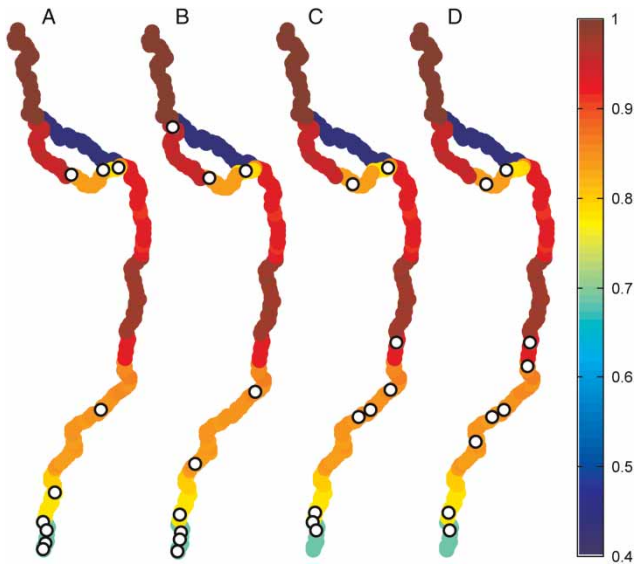

Figure 9 | Location of selected solutions A, B, C and D of Figure 5 for nine decision variables.

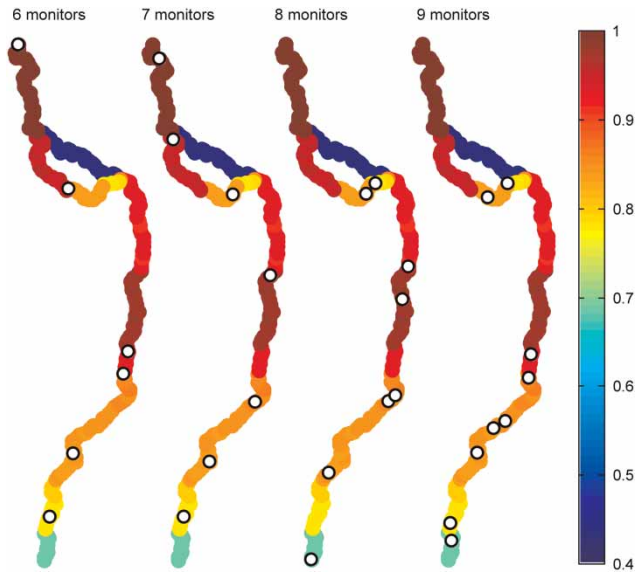

Figure 10 | Location of the most informative (and redundant) solution obtained for 6, 7, 8 and 9 monitors (the most right black dots of each Pareto front of Figure 5).

Moreover, Figure 10 shows the most informative solutions obtained for different numbers of monitors and how they are redistributed when an additional monitor is considered in the solution set. There is a regular distribution of monitors for six and seven decision variables, while for eight and nine decision variables there is a tendency to

locate the monitors upstream. This can be explained by recalling that all the tributaries, with the exception of the Cauca River, are located upstream (see Figure 5) so that the information collected at monitors located in this area provides insight into the state of the system downstream.

It is important to note that during the design process of a new monitoring network equifinality issues related to the fact of selecting a single solution from multiple optimal solutions given by the Pareto fronts might arise. The trade off between information content and redundancy may imply that the ideal network configuration would be located at the origin of any Pareto graph (e.g., Figure 14). One criterion to select a solution, then, is to choose the closest point to the origin, converting the problem to a single-objective optimization one. If two solutions provide similar information content, a rational decision-maker should choose one that has the least total correlation among the sites (thus avoiding the costly activity of collecting redundant information). In the case of having two or more solutions that are equivalent in both criteria, then additional, practical considerations must be taken into account to make a final selection, such as geographical convenience, accessibility, safety, electricity, etc. In any case, the selection of a final solution needs expert knowledge and is to be made generally by decision-makers (e.g., Ferreira *et al.* 2007), who may include extra, non-technical issues such as political aspects and social convenience.

Results using the rank-based greedy algorithm

The algorithms presented in Figure 4 were applied to the Magdalena River in order to find the location of m number of monitors. In order to analyse the evolution of the monitoring network, experiments were carried out for $m = 5$, 6, 7, 8 and 9. The algorithms were executed using as the starting point each of the 181 computational points of the model (which is equivalent to repeating the exercise in Figure 3 for a different starting circle), generating five matrices (one for each m) with size $181 \times m$.

Ranking by joint entropy

The solution for the monitors with the maximum Joint Entropy was selected from the previously generated matrix. The locations of the monitors are plotted in Figure 11.

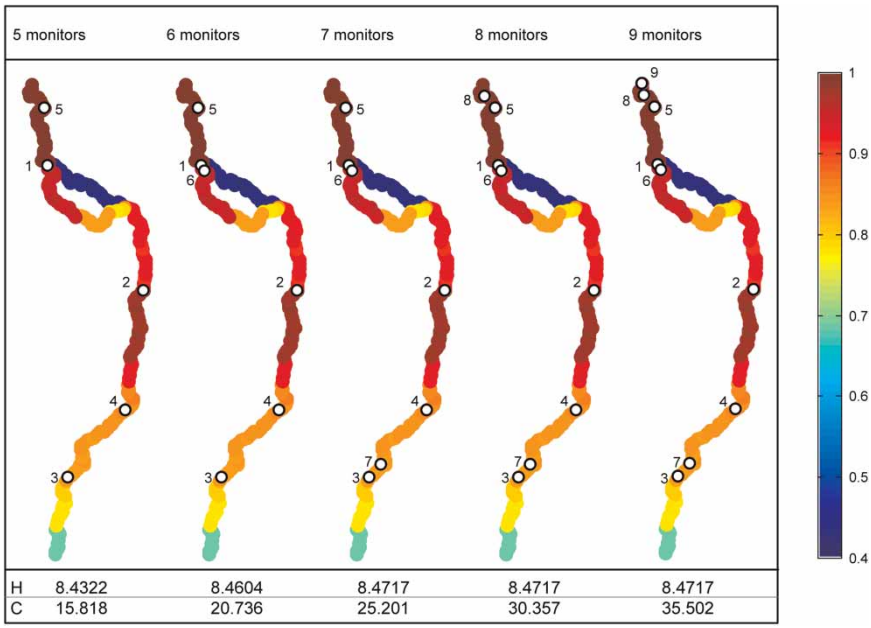

Figure 11 | Results obtained running the flowchart of Figure 1-left. Numbers represent the order in which each monitor was selected. The scale represents entropy (bits). A colour version of this figure is available in the online version of the paper, available at http://www.iwaponline.com/jh/toc.htm.

From Figure 11 it can be observed that regardless of the number of monitors, the first monitor is always located on the Loba branch, just before the convergence with the Mompox branch; this monitor, however, is not the one with the maximum information content (which is the second point from downstream to upstream). This means that starting with the monitor with the highest entropy does not guarantee that the final set of monitors has the maximum Joint Entropy.

The second monitor is located where the Lebrija River joins the Magdalena River at the connection to the wetland W_2; the third monitor, that adds the maximum Joint Entropy to the previous set of two monitors, is placed after the junction with the Nare River; the fourth monitor is located between the discharges to the Magdalena River of the rivers Carare and Opón and the fifth one is placed at the downstream part of the main river, completing a set of five monitors with a Joint Entropy value of $H = 8.4322$ bits. The solution for six monitors includes the same five locations selected in the same order with the sixth added at the place where the wetland W_3 is connected, incrementing the Joint Entropy of the set to $H = 8.4604$ bits. Similarly, the solution for seven monitors includes the previous six and adds the seventh at a location near the city of Berrío, downstream of the third monitor. This makes the Joint Entropy increase again to $H = 8.4717$.

Up to this point, every new monitor selected has increased the maximum information content possible with the previous set and this monitor has been unique at every step. However, 20 different candidates arise for the eighth monitor and none of them provides any additional information content to the set of seven monitors, implying that further monitors are redundant. The location of the eighth and ninth monitors (for which 148 different candidates arose) as shown in Figure 8 also confirms that these monitors are not worth selecting: they all congregate downstream repeating the information provided by the fifth monitor.

Ranking by total correlation

The same exercise was performed for the algorithm presented in Figure 1(b); the results are shown in Figure 12.

From Figure 12 it can be observed that the first monitor is located at the connection to the wetland W_3 and that subsequent monitors are located in the upstream part of the river, looking for points with very low information content. This is because one way of reducing the Total Correlation is

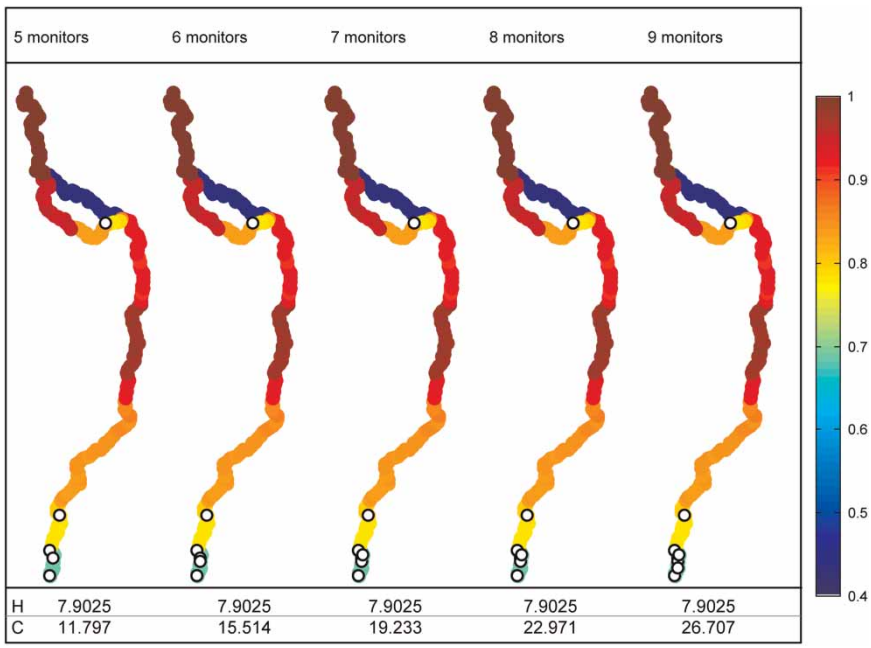

Figure 12 | Results obtained running the flowchart of Figure 1-right. Numbers represent the order in which each monitor was selected. The scale represents entropy (bits). A colour version of this figure is available in the online version of the paper, available at http://www.iwaponline.com/jh/toc.htm.

by adding random variables with very low (or null) entropy (Alfonso *et al*. 2010b).

Comparison with the existing monitoring stations

The monitoring network formed by the existing stations on the Magdalena River, with flow data available for the year 1995, was evaluated from an Information Theory perspective. The set of nine stations (Salgar, Berrío, San Pablo, Regidor, Peñoncito, El Banco, Magangué, Tacamocho and Calamar) has a value of Joint Entropy of $H = 8.3808$ bits and a value of Total Correlation of $C = 34.7464$ bits. The performance of this network can be compared to the results obtained for nine variables using the multi-objective optimization approach and by the ranking approach in the Joint Entropy – Total Correlation space (Figure 14). It is observed that the set is not optimal (there exist other solutions that give better Joint Entropy and Total Correlation values). In the figure it is evident that the multi-objective optimization algorithm is not able to reach extreme solutions as effectively as the ranking algorithms for the case of nine variables. This can be an indication that the algorithm must run for a longer period of time or that some adjustments in the probabilities of the genetic-related operations might be needed. However, it can be concluded that the ranking method is an effective way to complement the Pareto front at the extremes.

It must be pointed out that the existing monitoring network was built taking into consideration only local data needs at populated places (e.g., water levels for navigation activities and flood surveillance), without looking at the most convenient location for capturing the global behaviour of the river, and for this reason the method is limited to the evaluation of such a network.

Comparison with monitors located at tributaries

From the practical point of view, discharge measurements are part of the navigation studies for the Magdalena River. In order to determine the water balance, these measurements are taken before and after the most important tributaries and at bifurcations such as the Mompox and Loba branches. In order to evaluate these locations from the Information Theory perspective, the value of the marginal entropy before and after the inflows of the eight tributaries included in the model is presented in Figure 13. It can be noted that the information content always increases after every inflow, with the exception of the Lebrija River, whose discharge is produced near to the wetland W_1. Therefore, the straightforward conclusion is to place monitors after the tributaries in order to get the maximum information content of the river.

However, a comparison between monitors located according to this analysis and the optimal solutions obtained with the multi-objective optimization method for eight

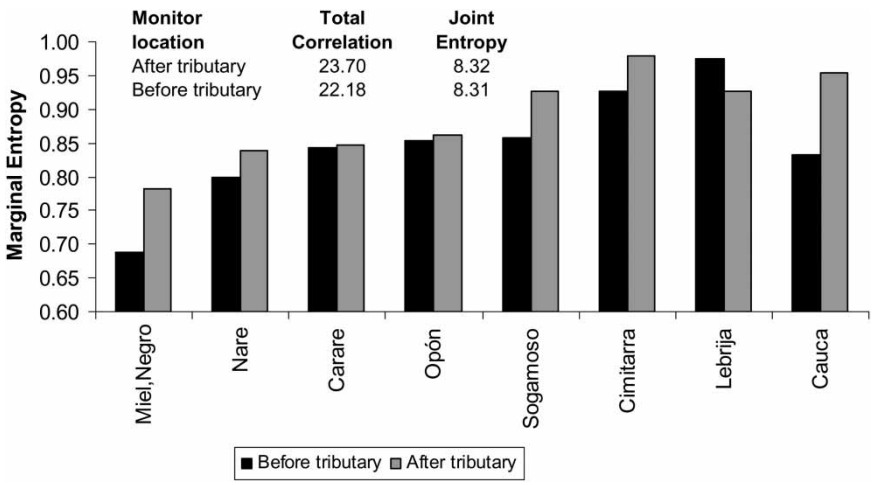

Figure 13 | Entropy values before, at and after the main tributaries.

decision variables (bottom-right of Figure 5), reveals that locating gauges at the tributaries is sub-optimal; this suggests that the effect of the wetlands, typically ignored in the measurement campaigns due to the difficulties of monitoring in such a vast wetland area, the hundreds of small connections between river and wetlands and the poor elevation data available, must be taken into account in order to understand the behaviour of the river better.

To make a general comparison, the resulting monitoring set located taking into consideration the eight tributaries of Figure 13 is included in the Total Correlation – Joint Entropy plane for nine variables in Figure 14. It can be observed that both sets (before and after the tributaries) have a slightly better value of Joint Entropy than the existing monitors. Naturally, the Total Correlation cannot be evaluated in this graph because it is very sensitive to the number of monitors in place.

SENSITIVITY ANALYSIS OF THE PARAMETER *A*

The selection of bin-size for probability estimation by means of frequency analysis is a well-known problem and for this reason the selection of the parameter a in Equation (4) may change the value of the entropy-related quantities. In order to analyse the implications of these changes, the entropy map presented in Figure 4 is redrawn for different values of a (see Figure 12). It can be observed that entropy values decrease when the value of a increases,

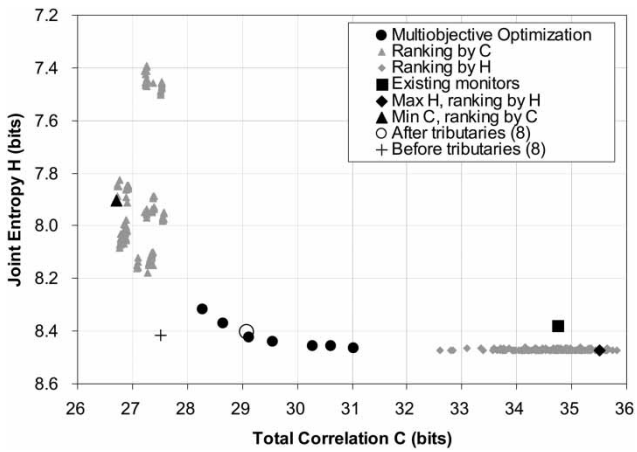

Figure 14 | Solutions obtained with the different methods in the Total Correlation – Joint Entropy plane.

because the number of bins for the frequency analysis are fewer, and therefore the number of sums required to assess Equation (1) is less. However, the relative value of the points with respect of the others in the same map is, in general, maintained regardless of the value of a. Therefore the expressions (2) and (3) yield numerically different values, but basically the same locations are obtained. This can be seen in Figure 15, where the zones with high entropy are always between the discharges of the tributaries Cimitarra and Lebrija and also after the convergence of the branches Mompox and Loba. On the other hand, the zones with low entropy are located in the Mompox branch and at the upstream part of the river, before the junctions with the rivers Miel and Negro. It can also be observed that in the wetlands zone the entropy changes in a similar way between maps. This implies that the resultant monitoring networks generated with the presented methods do not change significantly when changing the value of a. It must be noted, however, that an extreme, illogic value of a, such as $a = 1$ or $a = 100{,}000\ m^3/s$, leads to useless constant entropy maps. It is recommended, therefore, that this value should be set between the lowest and the highest mean flow of the incoming tributaries of interest.

CONCLUSIONS AND RECOMMENDATIONS

The entropy map for discharge in the Magdalena River shows that entropy increases at places where the tributaries flow into the river and diminishes at places where connections to the wetlands exist.

The series of experiments carried out above gives rise to the following conclusions:

- The selection of high-entropy points for monitoring leads to redundant monitors and the selection of low-entropy points generates a final set with low information content. The conflicting nature of these Information Theory quantities promotes the use of a multi-objective optimization approach. However, the selection of the final monitoring network selection is not straightforward if only the generated Pareto fronts are analysed, and it is still difficult to find an optimal solution that satisfies both criteria. In

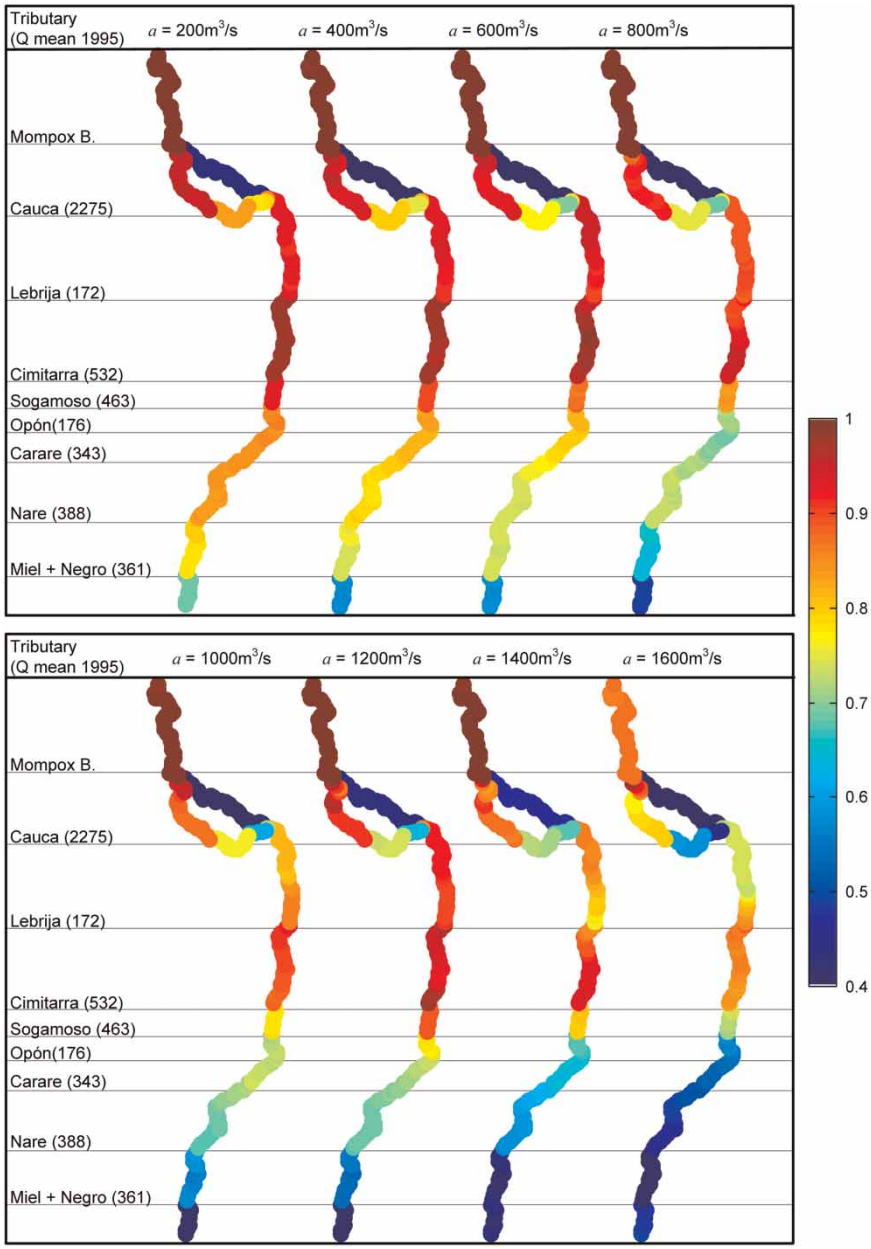

Figure 15 | Entropy maps for different values of a, Equation (5).

order to choose one point from the Pareto front, additional constraints are needed to determine the relative importance of joint entropy and total correlation. It is recommended that decision makers find these additional constraints by considering the requirements of water users.

- Seven monitors is the maximum number of monitors for which the Joint Entropy continuously increases along the Magdalena River, under the conditions with which the model was built. Additional monitors are fully redundant and do not add any further information content.
- The ranking-based methods are useful for finding the extremes of the Pareto fronts generated by the multi-objective optimization procedure and could be used in further research to normalize the information quantities and therefore to evaluate the solutions in a relative way.

An interesting finding is that the initial monitor used to start the algorithms in Figure 1 plays a significant role in the Joint Entropy and the Total Correlation of the final set. Also, starting with the point with the highest entropy does not guarantee that the final set of monitors has the maximum information content.

- Although the existing monitoring stations were placed individually to fulfil the requirements of the cities without assessing the network as a whole, the performance of this set yields acceptable information content but there is a high redundancy between monitors. Moreover, its performance is similar to what is obtained if the monitors are located following the location of the tributaries, as is normally done during monitoring campaigns.

ACKNOWLEDGEMENTS

This research has been founded by Delft Cluster and Delfland Waterboard. The authors want to thank CORMAGDALENA for allowing the use all the available information of the Magdalena River, and the Laboratory LEH-UN of the National University of Colombia and University of Norte for their collaboration in collecting and providing the data.

REFERENCES

Alfonso, L. 2010 Optimisation of Monitoring Networks for Water Systems: Information Theory, Value of Information and Public Participation. PhD Thesis, UNESCO-IHE/TU, Delft.

Alfonso, J. L., Jonoski, A. & Solomatine, D. P. 2010a Multiobjective optimization of operational responses for contaminant flushing in water distribution networks. *Journal of Water Resources Planning and Management* **136** (1), 48–58.

Alfonso, L., Lobbrecht, A. & Price, R. 2010b Information theory-based approach for location of monitoring water level gauges in polders. *Water Resources Research* **46** (3), W03528.

Alfonso, L., Lobbrecht, A. & Price, R. 2010c Optimization of water level monitoring network in polder systems using information theory. *Water Resources Research* **46**, W12553, 13.

Amorocho, J. & Espildora, B. 1973 Entropy in the assessment of uncertainty in hydrologic systems and models. *Water Resources Research* **9** (6), 1511–1522.

Barreto, W., Vojinovic, Z., Price, R. & Solomatine, D. 2009 A multi objective evolutionary approach to rehabilitation of urban drainage systems. *Journal of Water Resources Planning and Management* **136** (5), 547–554.

Chiu, C. L. & Chen, Y. C. 2003 An efficient method of discharge estimation based on probability concept. *Journal of Hydraulic Research* **41** (6), 589–596.

Chiu, C. L., Lin, G. F. & Lu, J. M. 1993 Application of probability and entropy concepts in pipe flow study. *Journal of Hydraulic Engineering, ASCE* **119** (6), 742–756.

Davar, Z. K. & Brimley, W. A. 1990 Hydrometric network evaluation: audit approach. *Journal of Water Resources Planning and Management* **116** (1), 134–146.

Deb, K. & Agrawal, R. B. 1994 Simulated binary crossover for continuous search space. *Complex Systems* **1** (9), 115–148.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002 A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6** (2), 182–197.

Ferreira, J., Fonseca, C. & Gaspar-Cunha, A. 2007 Methodology to select solutions from the pareto-optimal set: a comparative study. *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*. ACM, London, England, pp. 789–796.

Freni, G. & Mannina, G. 2011 The identifiability analysis for setting up measuring campaigns in integrated water quality modelling. *Physics and Chemistry of the Earth, Parts A/B/C* **42–44** (0), 52–60.

Husain, T. 1989 Hydrologic uncertainty measure and network design. *Water Resources Bulletin* **25** (3), 527–534.

Kraskov, A., Stögbauer, H., Andrzejak, R. G. & Grassberger, P. 2003 Hierarchical clustering based on mutual information. Arxiv preprint q-bio.QM/0311039.

Krstanovic, P. F. & Singh, V. P. 1992 Evaluation of rainfall networks using entropy: I. Theoretical development. *Water Resources Management* **6** (4), 279–293.

Loucks, D. P., van Beek, E. & Stedinger, J. R. 2005 *Water Resources Systems Planning and Management*. UNESCO - WL Delft Hydraulics, Paris.

Lehner, B., Verdin, K. & Jarvis, A. 2006 *HydroSHEDS Technical Documentation, V 1. 0*. WWF, Washington, DC. Available from: www.worldwildlife.org/hydrosheds.

Maruyama, T., Kawachi, T. & Singh, V. P. 2005 Entropy-based assessment and clustering of potential water resources availability. *Journal of Hydrology* **309** (1–4), 104–113.

McGill, W. J. 1954 Multivariate information transmission. *Psychometrika* **19** (2), 97–116.

Mishra, A. K. & Coulibaly, P. 2009 Developments in hydrometric network design: a review. *Reviews of Geophysics* **47**, RG2001.

Mishra, A. K., Özger, M. & Singh, V. P. 2009 An entropy-based investigation into the variability of precipitation. *Journal of Hydrology* **370** (1–4), 139–154.

Mogheir, Y. & Singh, V. P. 2002 Application of information theory to groundwater quality monitoring networks. *Water Resources Management* **16** (1), 37–49.

Mogheir, Y., Singh, V. P. & de Lima, J. 2006 Spatial assessment and redesign of a groundwater quality monitoring network

using entropy theory, Gaza Strip, Palestine. *Hydrogeology Journal* **14** (5), 700–712.

Moon, Y. I., Rajagopalan, B. & Lall, U. 1995 Estimation of mutual information using kernel density estimators. *Physical Review E* **52** (3), 2318–2321.

Moss, M. E. & Karlinger, M. R. 1974 Surface water network design by regression analysis simulation. *Water Resources Research* **10** (3), 427–433.

Moss, M. E. & Tasker, G. D. 1991 Intercomparison of hydrological network-design technologies. *Hydrological Sciences Journal/ Journal des Sciences Hydrologiques* **36** (3), 209–221.

Ruddell, B. L. & Kumar, P. 2009 Ecohydrologic process networks: 1. Identification. *Water Resour. Res.* **45**, W03419.

Sene, K. J. & Farquharson, F. A. K. 1998 Sampling errors for water resources design: the need for improved hydrometry in developing countries. *Water Resources Management* **12** (2), 121–138.

Shannon, C. E. 1948 A mathematical theory of communication. *Bell System Technical Journal* **27** (3), 379–423.

Shannon, C. E. & Weaver, W. 1949 The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.

Singh, V. P. 1997 The use of entropy in hydrology and water resources. *Hydrological Processes* **11** (6), 587–626.

Sun, S. & Bertrand-Krajewski, J. L. 2012 On calibration data selection: the case of stormwater quality regression models. *Environmental Modelling and Software* **35**, 61–73.

Watanabe, S. 1960 Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development* **4** (1), 66–82.

WMO 2003 The Second Report on the Adequacy of the Global Observing Systems for Climate in Support of the UNFCCC. Rep., WMO/TD No. 1143.

WMO 2008 *Guide to Hydrological Practices. Hydrology – From Measurement to Hydrological Information*. Publications Board WMO, Geneva.

Yang, Y. & Burn, D. H. 1994 An entropy approach to data collection network design. *Journal of Hydrology* **157** (1–4), 307–324.

First received 19 May 2011; accepted in revised form 31 May 2012. Available online 17 September 2012