EPiC Series in Engineering
Volume 3, 2018, Pages 2178–2186
HIC 2018. 13th International
Conference on Hydroinformatics

Optimizing the selection of cross-sections using Information theory: a case in the Magdalena River, Colombia

Sara Alonso[1], Elena Ridolfi[2], Chiara Biscarini[3] and Leonardo Alfonso[1]
[1]IHE-DELFT, Westvest 7, Delft 2611 AX, Netherlands
[2]University of Perugia, DICA, via G. Duranti, Perugia 06125, Italy
[3]UNESCO Chair in Water Resources Management and Culture, University for Foreigners of Perugia, Perugia 06123, Italy
alons2@unesco-ihe.org, elena.ridolfi@unipg.it,
chiara.biscarini@unistrapg.it, l.alfonso@un-ihe.org

Abstract

Accurate flood propagation and inundation models are crucial in flood risk assessments. For fast flowing rivers such as the Magdalena River (Colombia) with high vulnerability and exposure rates is even more essential. Indeed, floods in Magdalena River account for 90% of the damages and 70% of the causalities in Colombia. River cross-sectional information (i.e. their number and spacing) must be optimally selected to properly capture river's hydraulic behaviour. Optimization is a powerful tool for doing such selection often necessary to increase the efficiency of field works and decrease model simulation time. A methodology based on the entropy concept provides interesting results in agreement with those proposed in literature. The optimization method proposes the use of two concepts belonging to information theory: the joint entropy and total correlation. Total correlation quantifies the redundancy of cross-sections; joint entropy provides their information content. This approach is applied to a reach of the Magdalena River. This study analyses the interrelation between the location of the optimal set of cross-sections and the hydraulic behaviour of the Middle-Magdalena River. Further work considers the evaluation of model performance with the optimized cross-sections, where no negative impacts on the reliability of flood profiles with respect to the original model are expected.

Keywords: river cross-sections, entropy, hydraulic modelling, information theory, Magdalena River, optimization.

G. La Loggia, G. Freni, V. Puleo and M. De Marchis (eds.), HIC 2018 (EPiC Series in Engineering, vol. 3), pp. 2178–2186

1 Introduction

Due to social, environmental and economic impacts, river flood studies need to be reliable. More sophisticated and complex models (two- and three-dimensional) have been developed with this objective. However, due to the availability of high-quality resolution topographic data from new topographical survey techniques, also simple one dimensional or quasi-two-dimensional hydraulic models (e.g. SOBEK 1D2D) accurately simulate the flood inundation extent with small implementation costs (Castellarin, Di Baldassarre, Bates, & Brath, 2009). Accurate definition of river geometry is essential for a proper description of river's hydraulic behaviour. Thus, river cross-sectional spacing is crucial to build a reliable hydraulic model. The intuitive approach uses a large number of cross-sections that increases the model performance but requires an onerous process of data collection and updating (Ridolfi et al., 2014). Therefore, the definition of the optimal number of cross-sections is an important step in hydraulic modelling. Optimal is defined as the minimum number of cross-sections necessary to achieve a good accuracy of the corresponding hydraulic model. For this purpose, here the problem is posed as a multi-objective optimization problem (MOOP). This problem is solved finding the optimal set of cross-sections which total information is maximum (i.e. maximum entropy, ME) and which redundancy is minimum (i.e. minimum total correlation, TC). Both ME and TC are defined in the information theory framework proposed by Ahammad, Daskalakis, Etesami, & Frome, (2004). The information theory has been applied in many fields as communication, biochemistry, statistics, gambling, geography or economics. In water resources management, to detect the potential water resources availability (Kawachi, Maruyama, & Singh, 2001) or the occurrence of extreme events as droughts (Mishra, Özger, & Singh, 2009), and to cluster homogeneous sites (Ridolfi et al., 2016). Another example is its application for optimizing water level monitoring networks (Alfonso, Ridolfi, Gaytan-Aguilar, Napolitano, & Russo, 2014) or groundwater-monitoring networks (Mogheir & Singh, 2002). An overview for the optimization of several types of monitoring networks applying the theory of information could be found in Keum, Kornelsen, Leach, & Coulibaly, (2017).

2 Methodology

Shannon developed the Theory of Information with the aim of studying communication and all its abstraction. In this way, he defined the concept of entropy as a measure of information content in a message. In hydrology, entropy is defined as the uncertainty about the knowledge of the state of a given system. For any random variable (RV) X, the marginal entropy is:

$$H(X_i) = E[-\log_2 p(X_i)] = \sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{1}$$

where $P[X = x_i] = p_i$ is the probability that X assumes the value of x_i and n is the length of X. Units are represented in bits. The random variable X is the water-level time series at a specific cross-section. If the water levels at one location are the same over the time, the uncertainty to select randomly a water level is zero, and therefore the entropy is also zero. On the contrary, when all the water levels are different, the entropy is maximum.

For estimating the probability, the water levels are divided into classes (bins). The selection of bins size is important as the number of bin partitions increase the entropy. There are several studies evaluating which is the optimal bin size. Some authors proposed a proportional bin size instead of a fixed one such a Chapman, (1986). In this case study, each bin width equals 0.20 m, selected on the basis of the hydraulic model sensitivity consistently with Ridolfi et al., (2014). The marginal probability of each RV is estimated dividing the frequency associated with each class (i.e., how many values of the time series fall in each class interval) by the number of elements of the time series itself. To estimate the joint probability of *N* RVs, the RVs have to be agglomerated in pairs as fully described by Alfonso, Lobbrecht, & Price, (2010).

Considering *N* stochastically dependent RVs, their joint entropy (JH) is:

$$JH = H(X_1, X_2, \dots, X_N) = -\sum_{i=1}^{n} \dots \sum_{i_N=1}^{N} p_{i,\dots,i_N} \log_2(p_{i,\dots,i_N}) \tag{2}$$

where $p_{i,\dots,i_N}$ is the joint probability of *N* variables. The grouping property of mutual information, as reported by Kraskov, Stögbauer, Andrzejak, & Grassberger, (2003), is used to cluster the $X_1, X_2, \dots, X_N$ variables in a new RV to estimate the joint entropy of the *N* RVs.

Finally, the dependency and redundancy between *N* RVs is estimated through the total correlation (TC). It is defined as the difference between the sum of the marginal entropy of the RVs and their joint entropy:

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

The multi-objective optimization problem (MOOP) is presented in the following equations:

$$Min\ (TC) = Min\{TC(X_1, X_2, \dots, X_N)\} \tag{4}$$

$$Max\ (JH) = Max\{H(X_1, X_2, \dots, X_N)\} \tag{5}$$

To solve the MOOP, a Non-dominated Sorting Genetic Algorithm (NSGA-II) is used as proposed in Deb, Pratap, Agarwal, & Meyarivan, (2002). This genetic algorithm combines the parent and offspring until the size of the population is reached, and the best solution is selected.

3 Case study

The Magdalena basin is the South America’s fifth largest basin and is located in Colombia. The basin drains an area of 257000 km^2 which corresponds to 22.8% of the Colombian territory (Figure 1). In the basin are settled about 75% of Colombian inhabitants with a total of 728 municipalities. The Magdalena River originates in the southern Andean Mountains and ends in the Caribbean Sea at the city of Barranquilla with a total length of 1563 km. Its average discharge is 7100 m^{3}s^{-1} and has two rainy seasons from March to May and from October to December. However, the Magdalena River shows an important variability of its inter-annual flow regime due to a strong seasonally natural phenomena such an El Niño-Southern Oscillation (ENSO) (Higgins, Restrepo, Ortiz, Pierini, & Otero, 2016). The average precipitation is estimated in 2000 mm varying from 800 mm to 5000 mm, and its elevation from 0 to 5700 meters above the sea level. The Magdalena River is one of the top ten world-class river in sediment load transport, approximately 184 MT/yr (Restrepo & Escobar, 2016), showing a variability mainly associated with flashy peak events. This transport has increased in the last years associated to land use changes, deforestation and climate change. The variability in the flow and sediment transport increases the complexity of the river system making difficult to capture its hydraulic behaviour and implying recurrent and costly field work. Therefore, the selection of the optimal sections in the Magdalena River is a step required for future cost-efficient updates of its hydraulic model.

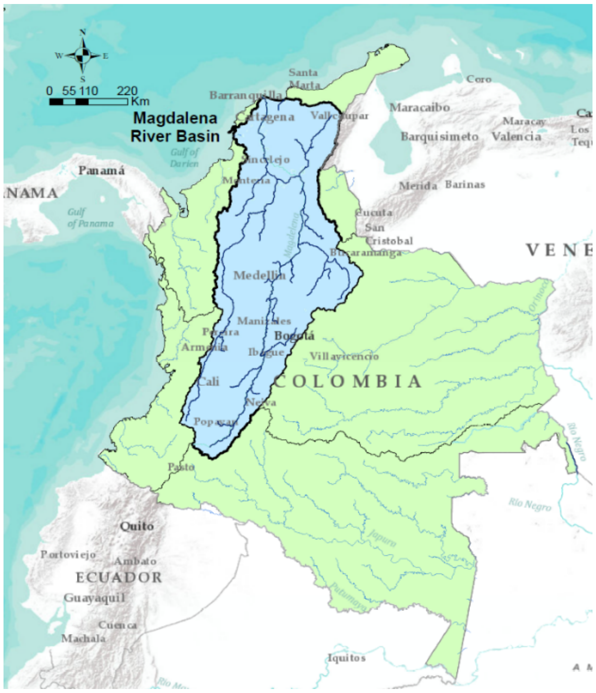

Figure 1: Location of the Magdalena River Basin in Colombia

4 Hydraulic model

The hydraulic model of the Magdalena River was developed in SOBEK 1D2D by the Research centre of the Magdalena River (CIRMAG) in cooperation with The Netherlands Institute of Applied Research in Water Sector (Deltares) for a 1507 km length. The SOBEK 1D2D solves the de Saint Venant equations in 1D inside the river channel and in 2D when the river banks are overtopped. The optimization method is applied to 575 km of the Middle Magdalena River between the water gauges of Puerto Salgar (i.e. the most upstream cross-section) and Regidor (i.e. the most downstream cross-section) as shown in Figure 2. Due to the complex geometry of the Magdalena River, its length is divided in two branches. The main branch was modelled with cross-sections each 200m —on average— of river length, meanwhile the other branch (with lower capacity) was modelled using cross-sections at the junctions. In the model, surveyed cross-sections and synthetic trapezoidal cross-sections were used. More details about the model can be found in Sanchez Lozano et al., (2015). The water levels and discharges were simulated at 925 locations —calculation points— and calibration has been partly made against observed data at seven water gauges —Puerto Salgar (upstream) to Regidor (downstream). However, there are still sources of uncertainty that need further analysis (e.g. DEM precision, cross-sections, embankments, etc.). In the Magdalena River, water gauge datum levels (WGDL) are a source of uncertainty because water gauges were not set at the same reference level after flood events, accidents and other interventions. Several Colombian institutions have issued reports claiming for the "right" WGDL, with differences between centimetres up to few meters (Hodžić, 2016) . WGDL differences at Sitio Nuevo ranged up to 7 m, meanwhile at La Gloria is only 2 cm. The water levels used in this study case are the result of 655 simulations ran with a randomly generated WGDL from a Monte Carlo Simulation in a studied carried out by Hodžić, (2016). The input data for the model dated from January 2010 until December 2010. However, the water levels used as input for the optimization process were water levels simulated from January 2010 until April 2010.

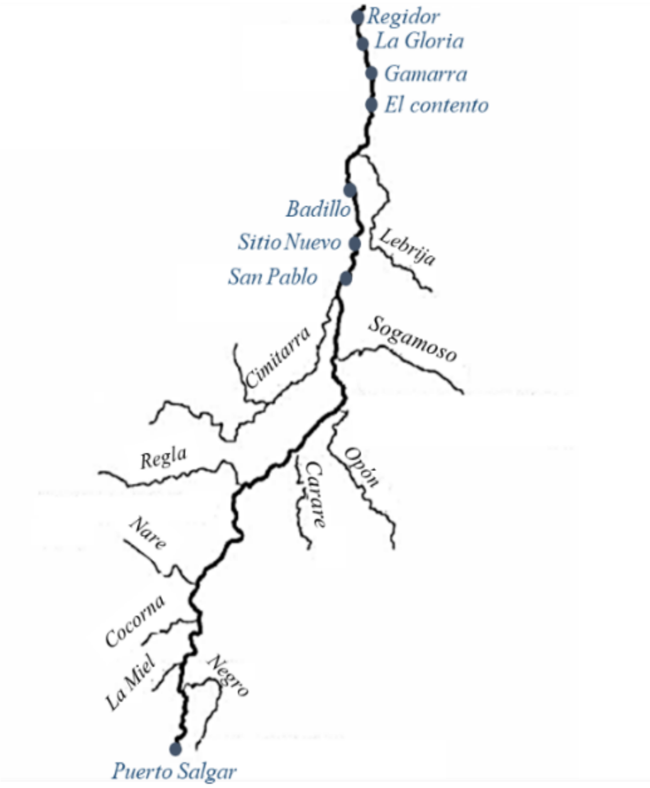

Figure 2: Middle Magdalena River with the seven water gauges for which are available observed data and the most downstream point (Puerto Salgar)

5 Numerical analysis and results

The optimization problem (i.e. Eqs. 4 and 5) was solved with different sets having a number of cross-sections ranging from 5 to 20. A Pareto front of the optimal solutions for different number of cross-sections is obtained (Figure 3). The optimal number of cross-sections ranges from 10 to 20. When more cross-sections are considered, the total correlation increases dramatically; while the gain in joint entropy is small. For instance, an increment from 10 to 12 sections leads to an increase of 0.16 bits in joint entropy and of 6.18 bits in total correlation. In other words, the increase of redundancy is much higher than the one in information content. From 10 to 20 sections, the joint entropy rises 2.5%, meanwhile the TC increases 60%. For this reason, it is not recommended to use more than 20 sections: above this number, the redundancy increases consistently without any relevant information gain. When the number varies from 5 to 9, the selection of a specific cross-sections set shows a dramatically difference in the amount of information gain (Figure 3).

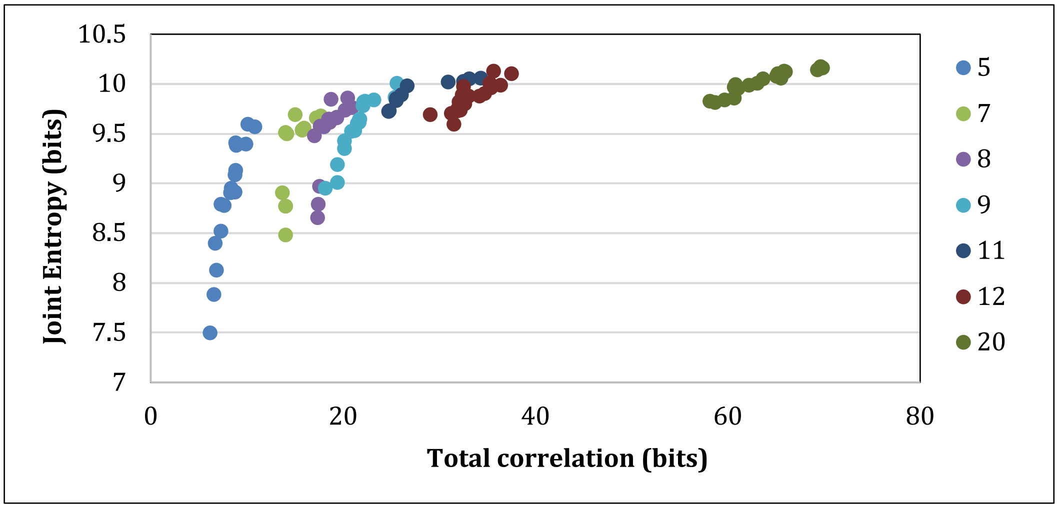

Figure 3: Pareto-optimal cross sections sets

6 Discussion and Conclusions

The hydraulic model of the Magdalena River is a complex braided river network with hundreds of reaches and with a total of 11 tributaries. Cross-section selection not only demand of a statistical analysis of the model outputs, but also a further evaluation of the physical interrelation between different model elements. Nevertheless, the statistical analysis of the theory of information brings the opportunity to relate the information content in the water levels with the hydraulic behaviour of the river. When the entropy is maximised and correlation is minimised, a crucial set of cross-sections among the total is selected. The location of optimal sets having 5, 6, 7, 8, 9, 10, 11, 12 and 20 cross-sections is presented in Figure 4.

Independently from the number of optimal cross-sections, the river reach between the tributaries Opon and San Bartolome accounts for the lower number of cross-sections selected, from 0 to 1. In the latter case, the only cross-section selected is located close to the Carare River. This tributary had a peak discharge of approximately 2400 $m^3.s^{-1}$ in December 2010, much higher than the peak observed at the Opon and San Bartolome Rivers (i.e. 760 and 580 m^3 s^{-1} in June 210, respectively). As the number of cross-sections increases from 5 to 20, the cross-sections are concentrated in the upstream and downstream parts of the Middle Magdalena River. The rationale of this can be due to the high concentration of tributaries in the river upper part. Though the discharge of such tributaries is not the highest one, they are placed close to each other. On the other hand, in the river lower part, the Magdalena split in two branches at three locations. This braided morphology causes high changes in in the water levels, and therefore increases the value of the entropy and the number of the cross-sections needed to better describe this behaviour. The cross-sections tend to concentrate close to the tributaries with highest discharge, the Sogamoso and La Miel Rivers. Both rivers observed a peak discharge in November 2010 of approximately 2600 and 1100 m^3 s^{-1}, respectively.

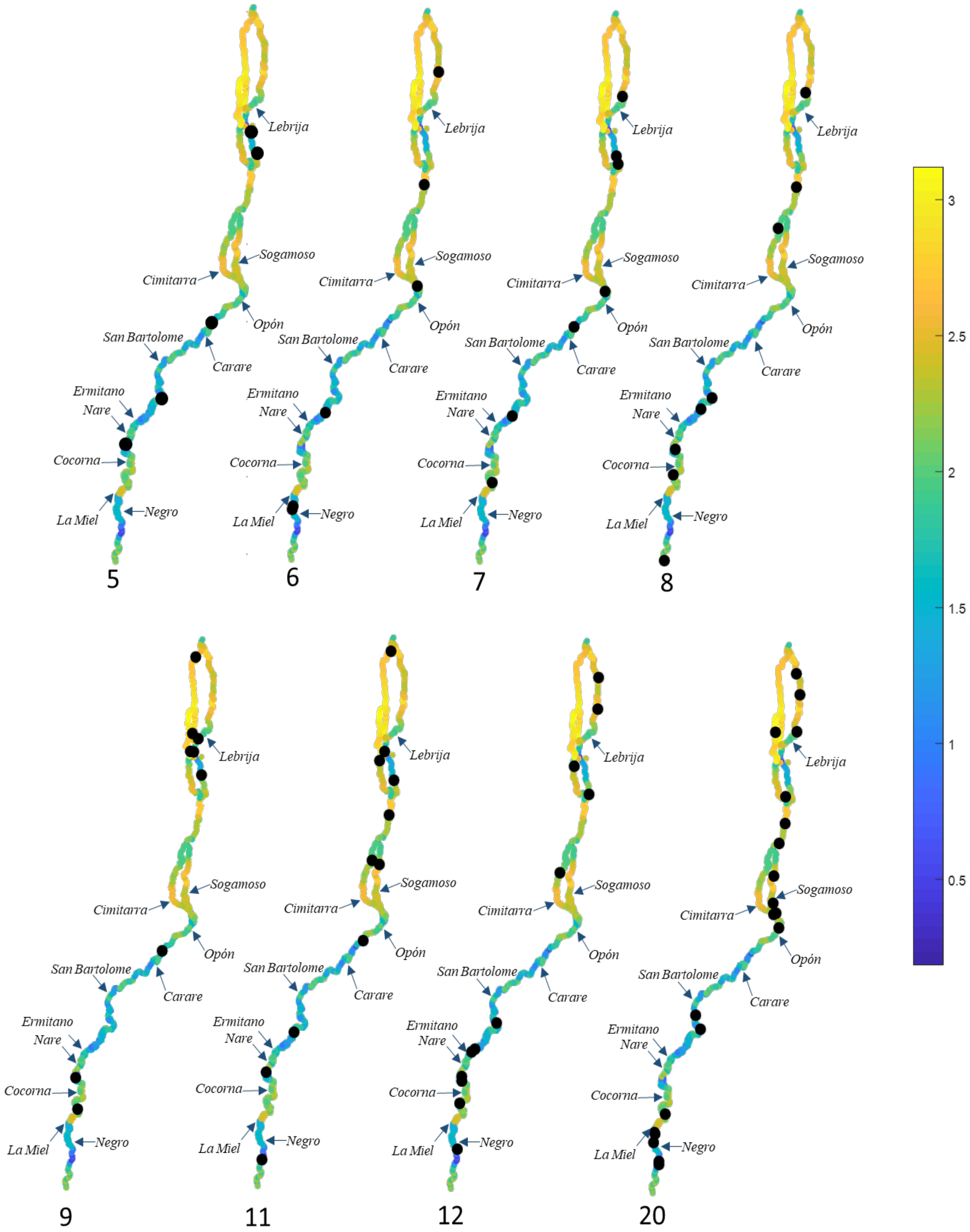

Figure 4: Marginal entropy at 925 calculation points of the Magdalena River and location of the optimal sets having a number of cross-sections equal to 5, 6, 7, 8, 9, 11, 12 and 20

The most selected cross-section in the optimal sets —a total of 20 times— is located downstream of the Lebrija River (Figure 5). Other two cross-sections selected up to 11 times are close to this most-repeated one. Results suggest that the high concentration of cross-sections in the lower part is due to the two-branch configuration of the river.

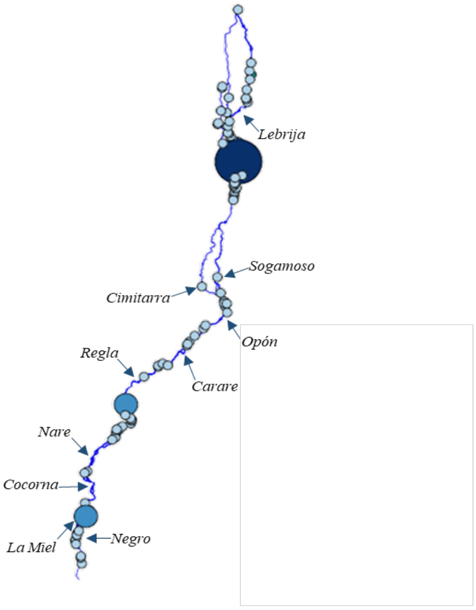

Figure 5: Frequency of selection of the 925 calculation points in the optimization process using the entropy concept

The existing model (with all cross-sections) highly overpredicted the water depths at the water gauges stations. This is proven with the Root Mean Square Error (RMSE) that has the highest value at Gamarra with 8.97 m and the minimum at Badillo with 0.34 m, considering the observed data from October to December 2010 (Hodzic, 2016). The future lines of research will be focused on the comparison of the RMSE obtained from the original model —with a total of 925 calculation points and 2627 cross-sections— with the one built using the cross-sections obtained from the optimization method. This model performance analysis requires a simplification of the river network and reduction of the river branches, adapted to the lower number of cross-sections.

7 References

Ahammad, P., Daskalakis, K., Etesami, O., & Frome, A. (2004). Claude Shannon and "A Mathematical Theory of Communication." *Relation*, 1–12. Retrieved from http://www.eecs.berkeley.edu/~christos/classics/shannon-report.pdf

Alfonso, L., Lobbrecht, A., & Price, R. (2010). Information theory-based approach for location of monitoring water level gauges in polders. *Water Resources Research*, *46*(10). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2009WR00810

Alfonso, L., Ridolfi, E., Gaytan-Aguilar, S., Napolitano, F., & Russo, F. (2014). Ensemble entropy for monitoring network design. *Entropy*, *16*(3), 1365–1375. Retrieved from http://www.mdpi.com/1099-4300/16/3/1365

Castellarin, A., Di Baldassarre, G., Bates, P. D., & Brath, A. (2009). Optimal Cross-Sectional Spacing in Preissmann Scheme 1D Hydrodynamic Models. *Journal of Hydraulic Engineering*, *135*(2), 96–105. Retrieved from http://ascelibrary.org/doi/10.1061/%28ASCE%290733-9429%282009%29135%3A2%2896%29

Chapman, T. G. (1986). Entropy as a measure of hydrologic data uncertainty and model performance. *Journal of Hydrology*, *85*(1–2), 111–126. Retrieved from http://www.mdpi.com/1099-4300/19/11/613/htm

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. Retrieved from https://www.iitk.ac.in/kangal/Deb_NSGA-II.pdf.

Higgins, A., Restrepo, J. C., Ortiz, J. C., Pierini, J., & Otero, L. (2016). Suspended sediment transport in the Magdalena River (Colombia, South America): Hydrologic regime, rating parameters and effective discharge variability. *International Journal of Sediment Research*, *31*(1), 25–35. Elsevier. Retrieved from http://dx.doi.org/10.1016/j.ijsrc.2015.04.003

Hodzic, I. (2016). Methodology to assess the influence of data uncertainties on modelling the Magdalena River, Colombia, (August). Retrieved from https://www.un-ihe.org/library.

Kawachi, T., Maruyama, T., & Singh, V. P. (2001). Rainfall entropy for delineation of water resources zones in Japan. *Journal of Hydrology*, *246*(1–4), 36–44. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022169401003559

Keum, J., Kornelsen, K., Leach, J., & Coulibaly, P. (2017). Entropy Applications to Water Monitoring Network Design: A Review. *Entropy*, *19*(11), 613. Retrieved from http://www.mdpi.com/1099-4300/19/11/613

Kraskov, A., Stögbauer, H., Andrzejak, R. G., & Grassberger, P. (2003). Hierarchical Clustering Based on Mutual Information. *Europhysics Letters*, *70*(2), 278. Retrieved from http://arxiv.org/abs/q-bio/0311039

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. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022169409001498

Mogheir, Y., & Singh, V. P. (2002). Application of information theory to groundwater quality monitoring networks. *Water Resources Management*, *16*(1), 37–49. Retrieved from https://link.springer.com/article/10.1023/A:1015511811686.

Restrepo, J. D., & Escobar, H. A. (2016). Sediment load trends in the Magdalena River basin (1980–2010): Anthropogenic and climate-induced causes. *Geomorphology*, *302*, 76–91. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169555X16311928.

Ridolfi, E., Alfonso, L., Di Baldassarre, G., Dottori, F., Russo, F., & Napolitano, F. (2014). An entropy approach for the optimization of cross-section spacing for river modelling. *Hydrological Sciences Journal*, *59*(1), 126–137. Taylor & Francis. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/02626667.2013.822640

Ridolfi, E., Rianna, M., Trani, G., Alfonso, L., Di Baldassarre, G., Napolitano, F., & Russo, F. (2016). A new methodology to define homogeneous regions through an entropy based clustering method. *Advances in Water Resources*, *96*, 237–250. Elsevier Ltd. Retrieved from https://www.sciencedirect.com/science/article/pii/S030917081630241X.

Sanchez Lozano, J. L., Ardila Camelo, F., Oliveros Acosta, J. J., Ramirez Morales, W. D., Cardona Almeida, C. A., Garay Bohorqez, C. I., Verschelling, E., et al. (2015). Hydraulic Modeling of Magdalena River Using Sobek. *Proceedings of the 36Th Iahr World Congress: Deltas of the Future and What Happens Upstream*, (July), 4630–4642. Retrieved from https://www.researchgate.net/publication/280293710_HYDRAULIC_MODELING_OF_MAGDALENA_RIVER_USING_SOBEK