EPiC Series in Engineering
Volume 3, 2018, Pages 1162–1170
HIC 2018. 13th International Conference on Hydroinformatics

Entropy based multicriterion evaluation for rainfall monitoring networks under the impact of discretization

Heshu Li[1*], Dong Wang[1*], Vijay P. Singh[2], Yuankun Wang[1]

[1]Key Laboratory of Surficial Geochemistry, Ministry of Education, Department of Hydrosciences, School of Earth Sciences and Engineering, State Key Laboratory of Pollution Control and Resource Reuse, Nanjing University, Nanjing, P.R. China
[2]Department of Biological and Agricultural Engineering, Zachry Department of Civil Engineering, Texas A & M University, College Station, TX77843, USA
wangdong@nju.edu.cn, dz1629006@smail.nju.edu.cn

Abstract

Rainfall monitoring networks provide fundamental input for hydrological models. Entropy, as a measure of uncertainty or information, is widely used in network evaluation or optimization. Computing entropy requires data discretization with methods like floor function, whereas the parameter selected is crucial and influential. This paper proposed an entropy based multicriterion method for evaluation of rainfall monitoring networks. Two indexes, separately account for information content and redundancy, were integrated with ideal point method. Values of the objective function were then computed to rank the stations and identify the significant ones. To find out the effect of discretization, parameter of the floor function was altered to get different schemes. A rainfall monitoring network containing 95 stations in the western Taihu Lake basin of China was analyzed as case study. Results showed that stations in the northern hilly area are more prominent in the network. Impact of the parameter in floor function is non-negligible as it determines entropy values, including its ranging scale and distribution pattern. Location of the stations rank extremely high and low also varies. As discretization process has an impact on the evaluation, it should be carefully used and sensitivity analysis is needed to avoid subjectivity and arbitrariness.

1 Introduction

A well-designed rainfall monitoring network can reflect the spatial-temporal variability of precipitation variables adequately and provide essential information for hydrological models (Yeh and

Chen, 2011; Xu et al., 2015). In a context of climate and land use changes, the requirement of optimal network design keeps increasing. Mishra and Coulibaly reviewed the methodological development in surface water network design by summarizing six commonly used methods, which are (1) statistically based methods, (2) information theory-based methods, (3) user survey approach, (4) hybrid method, (5) physiographic components, and (6) sampling strategies (Mishra and Coulibaly, 2009).

Entropy, as defined in information theory, directly defines information and quantifies uncertainty (Shannon, 1948). Among others, entropy based methods have been widely used in hydrometric network optimization. A fundamental consideration of network optimization is using a minimum number of gauges to gather abundant and accurate information, that is, the stations selected should contain as much information as possible and share less transinformation in the meanwhile. To this end, entropy is coupled with multiobjective method for the design or evaluation of networks (Krstanovic and Singh, 1992a, 1992b; Yang ans Burn, 1994; Mogheir at al., 2006; Alfonso et al., 2010, 2013; Li et al., 2012; Samuel et al., 2013; Leach et al., 2015; Su and You, 2014). Alfonso et al. used WMP approach for water level monitoring network design in polders, which also introduced quantization method for discretizing time series and the concept of total correlation (Alfonso et al., 2013). Li et al. presented a maximum information minimum redundancy (MIMR) criterion for designing hydrometric networks and compared it with WMP (Li et al., 2012). Alfonso et al. used ranking method to find more informative and less redundant monitoring network configurations (Alfonso et al., 2013). Samuel et al. proposed a combined regionalization and dual entropy-multiobjective optimization (CRDEMO) method for determining minimum network (Samuel et al., 2013). Leach et al. explored the impacts of hydrologic characteristics on the spatial variability of hydrometric networks using the dual entropy-multiobjective optimization (DEMO) model (Leach et al., 2015). Xu et al. used an entropy-based multi-criteria method to resample the rain gauge networks with different gauge densities [Xu et al., 2015].

To compute the entropy terms, floor function is frequently used to discretize the hydrometric time series, which avoids the choice of parametric distribution to fit the continuous data (Ruddell and Kumar, 2009). For instance, Alfonso et al. applied floor function to transform water level series into "pulses" of discrete information. The quantized series are "noise-free" in the sense that high-frequency, low-amplitude water changes can be filtered out (Alfonso et al., 2010). However, parameter of the floor function, which can be seen as the minimum dimensional unit of the data, remains inexplicit to determine. By now, few studies have investigated the impact that this discretization procedure have on the design or evaluation of hydrometric networks. One exploration was conducted by Li et al., in which floor function and merging method were applied to compute the multivariate joint entropy, total correlation and transinformation, and a sensitivity analysis was performed (Li et al., 2012). As discretization has a direct impact on the entropy values computed, this study aims to further investigate this impact by selecting different parameter values in the floor function and compare the corresponding evaluation results.

This study focused on the evaluation of rainfall networks. An entropy based multicriterion method was first presented to identify stations with more information content and less redundancy. The impact of the quantification procedure was explored by altering the parameter value in the floor function. The method was applied to a rainfall monitoring network in the western Taihu Lake basin of China. The paper is organized as follows: Section 1 briefly reviewed the application of entropy and multiobjective methods in hydrometric network design, followed by the description of the data and study area in section 2. Section 3 introduced the entropy concept and discretization function, as well as the entropy based multicriterion method. Section 4 clarified the methodology through results and discussion of the case study. Finally, conclusions were given in section 5.ent.

2 Study area and data

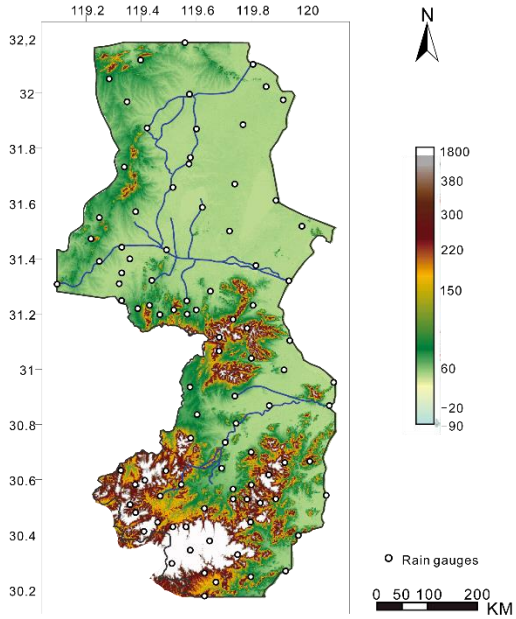

Figure 1 Terrain map of the study area

The study area is located in the subtropical monsoon climate zone in eastern coastal China, with a total area of around 13480 km^2. The landform here is a combination of mountains, hills and plain (figure 1). The Maoshan Mountain, the Jieling Mountain and the Tianmu Mountain distribute from north to south in the area with increasing heights. Elevation of the hilly and plain areas are 10~30 m and 6~8 m separately. The area is densely covered with lakes and interconnected waterways, providing plentiful water resources for agricultural irrigation and urban water supply. The annual mean temperature is 15~17℃. The annual precipitation is 1177 mm, mostly concentrated in May to September. A densely distributed rainfall monitoring network containing 95 stations was chose for this study, using the daily precipitation series data recorded in the year of 2006~2012.

3 Methodology

3.1 Entropy theory

Entropy, as defined by Shannon in the field of statistics, is a measure of the uncertainty of random variables, which is also regarded as the information content. Suppose X is a random variable with sample values x_i (i=1,2,…,n) and probability density function $p(x_i)$, the entropy of X is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (1)$$

The unit of entropy is bit. To measure the total uncertainty or information contained in two or more variables, joint entropy is defined. For the bivariate case, it is formulated as:

$$H(X,Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j)\, log_2\, p(x_i, y_j) \quad (2)$$

where $p(x_i, y_j)$ is the joint probability density of X and Y. Similarly for the multivariate case, joint entropy is:

$$H(X_1X_2, \dots, X_N) = -\sum_{i=1}^{n1}\sum_{j=1}^{n2}\cdots\sum_{k=1}^{nN} p(x_{1,i}, x_{2,j}, \dots, x_{N,k})\, log_2\, p(x_{1,i}, x_{2,j}, \dots, x_{N,k}) \quad (3)$$

where $p(x_{1,i}, x_{2,j}, \dots, x_{N,k})$ is the joint probability density function.

Mutual information is defined to estimate the shared information between two variables X and Y, and it can be interpreted as the reduction in the uncertainty of X given the knowledge of Y:

$$I(X,Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j)\, log_2 \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \quad (4)$$

To investigate multivariate dependence, total correlation is defined as:

$$C(X_1X_2, \dots, X_N) = \sum_{i=1}^{N} H(X_i) - H(X_1X_2, \dots, X_N) \quad (5)$$

It measures the total amount of information shared by all the variables X_1, X_2,…,X_N.

3.2 Entropy based multicriterion method for network evaluation Section

In this study, the precipitation series recorded by each station in the rainfall monitoring network is viewed as a random variable and there are 95 of them in total. Due to the variety of topography, elevation, vegetation and meteorological conditions, the precipitation recorded by stations at different locations varies from each other, so too is the information content. For each station X_i, we use two indicators -- the entropy ratio RE_i and the correlation ratio RC_i to measure its information capacity and redundancy. Specifically,

$$RE_i = \frac{H(X_i)}{H(X_1X_2,\dots,X_N)} \quad (6)$$

and

$$RC_i = \frac{I(X_i,X_1)+I(X_i,X_2)+\cdots+I(X_i,X_{i-1})+I(X_i,X_{i+1})+\cdots+I(X_i,X_N)}{C(X_1X_2,\dots,X_N)} \quad (7)$$

where N is the total number of stations. RE_i is the proportion of the information contained by station X_i in the total information of the network. Larger RE_i value indicates that more information is contained and thus the station is more important in the network. Analogously, RC_i measures proportion of the information shared between station X_i and the other stations takes up in the total correlation of the network. Larger RC_i value means more overlapped information, i.e., more redundancy is contained.

In order to realize multi-criterion evaluation, we integrate these two indexes with the ideal point method in multi-objective optimization. By using the quadratic sum of deviations as the evaluation function, this method constructs a dispersion function between general solutions and the ideal solution. In this case, the objective function $Q(X_i)$ is formulated as:

$$G(X_i) = [(RE_i - RE_{max})^2 + (RC_i - RC_{min})^2]^{\frac{1}{2}} \quad (8)$$

where RI_{max}, RC_{min} are ideal solutions. For a specific station X_i, a smaller G value indicates that the two indexes are closer to the ideal point, meaning that it contains relatively more marginal information and less redundant information, and is more prominent in the network. On the contrary, a larger the G value indicates greater the distance from the ideal point and lower significance of a station.

3.3 Discretization

The discretization procedure is conducted in preparation for computing the entropy related variables. The most frequently used methods are histogram discretization and mathematical floor function. Since the bin size is subjectively determined and has a significant effect on the computation results, the histogram method remains questionable. Thus the floor function is used here, which converts the continuous value x to a quantized value x_q – the nearest lowest integer multiple of a constant a:

$$x_q = a\left\lfloor \frac{2x+a}{2a} \right\rfloor \quad (9)$$

Li et al. pointed out that the advantage of floor function is that it can incorporate physical considerations and there three general rules guiding the selection of a: (1) it should make the quantized value of each station distinguishable; (2) it should preserve the spatial and temporal variability of precipitation series; (3) evaluation or optimization results based on parameter a, e.g., the significant stations identified, should remain stable when a fluctuates within some interval (Li et al., 2012). Given that the value of parameter a directly determines entropy values and thus is crucial to the final decision, we applied floor function in the multicriterion method proposed and investigated this effect by altering the value of a and comparing the evaluation results, which will be illustrated in section 4.

4 Results

4.1 Entropy Values

In the discretization process, the floor function parameter directly determines the value of the entropy computed and thus affects network evaluation results. To have an initial idea of this effect, we first made a comparison among the entropy values computed with different parameter values. In this study, parameter a (Equation 9) was taken as 1, 5, 10, 20, 50 and 100. To find out the scale and distribution of the entropy values, the histogram of entropy computed with the 95 rainfall monitoring stations and the corresponding density curves were plotted in figure 2. According to it, as a ranges from 1 to 100, scales of the entropy values are 1.43-2.10, 0.83-1.22, 0.58-0.87, 0.36-0.58, 0.13-0.26 and 0.14-0.10 separately, i.e., entropy values decrease with the increase of a, as well as its ranges. In addition, as can be seen from the probability density curves, when a takes a small value, the curve is relatively flat, namely, the data is uniformly distributed. With the increase of a, the curve gradually becomes steep and the distribution tends to be concentrated. When a takes 100, obvious concentration is seen and a large amount of data is concentrated between 0.04 and 0.08, with a very small maximum difference of about 0.04.

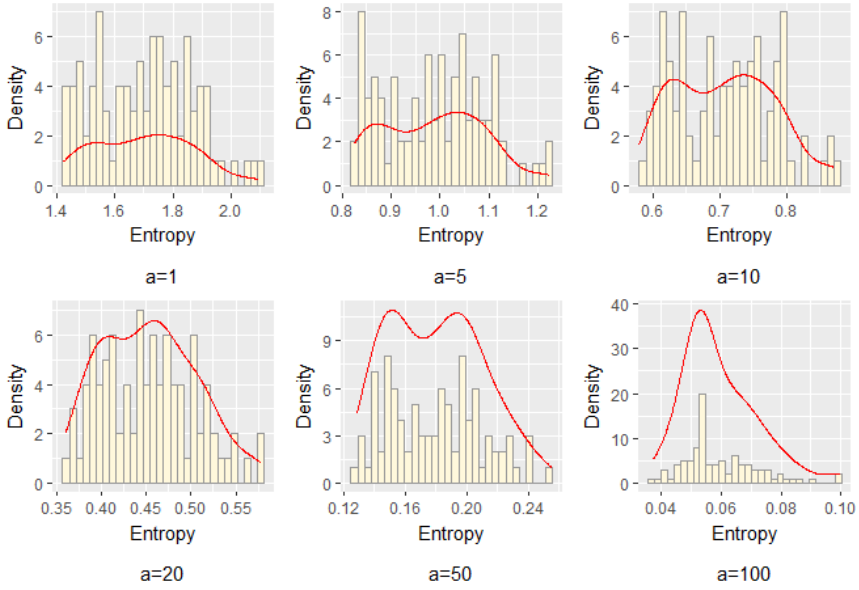

Figure 2 Distribution of entropy values computed under different floor function parameter values

This is because in the quantization process, all the data is rounded to integer multiple of a, and a large a will centralize the data to limited sets of bins. As a result, diversity of the data is reduced, causing low entropy values. Meanwhile, due to small number of bins, the probability of different stations having same grouping scenarios also increases, making entropy more concentrated. Therefore from the perspective of getting more distinguished entropy values, smaller a (no more than 10), which can distribute the entropy uniformly in a larger range and make the stations more distinguishable, is recommended.

4.2 Network evaluation

We evaluated the rainfall monitoring network of 95 stations with the entropy based multicriterion method defined in Equations 6-8. Taking a the same values as in section 4.1, the entropy ratio index RE and the correlation ratio index RC of each station were computed, and the objective function value G was obtained. Then the 95 stations were ranked accordingly, where stations with extreme low and high G values were identified as significant and non-significant stations. To be more intuitively, the obtained G values were regionalized with kriging interpolation and illustrated by contour maps (Figure 3). The top and bottom 10 stations were highlighted with yellow and red points respectively. It can be seen that the mountainous regions in the middle and southern area of the region exhibit high G values, while the hilly region in the northern area shows low G values. Note that the network is denser in the central and southern regions and sparser in the north, which may causing the difference. This indicates that to some extent, there exists information redundancy in the central and southern area, while in the northern area, additional stations are needed to incorporate more information.

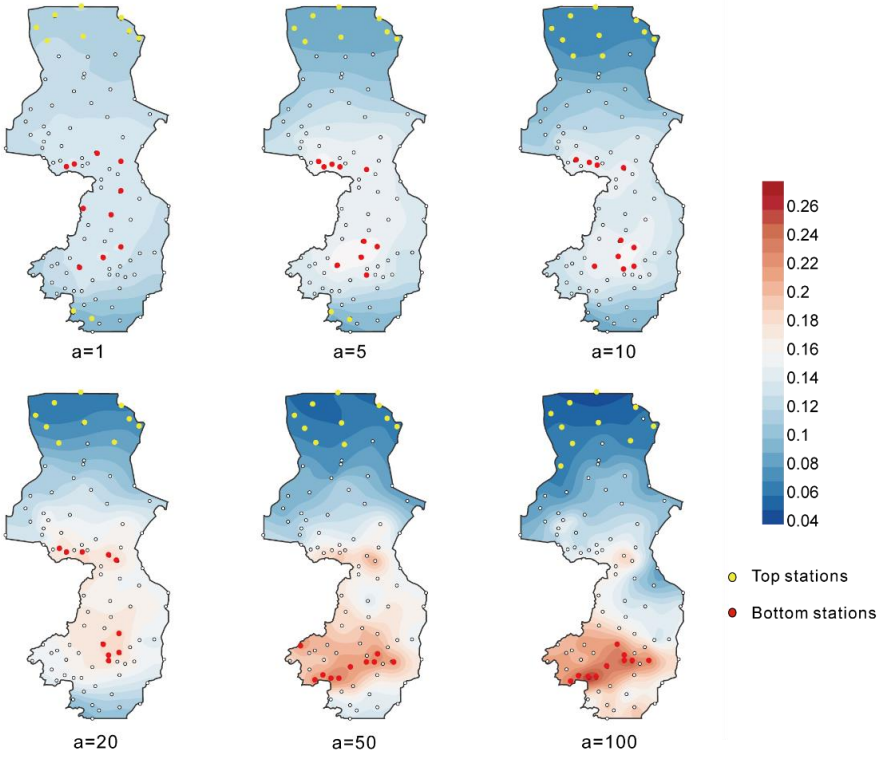

Figure 3 Contours of the evaluation function value G under different floor function parameter values

Moreover, unlike the entropy values, the ranges of G increases with the increase of a, which enlarges the difference between stations and areas. This indicates that although an increasing a will concentrate the entropy values, this impact is weaken by the proportion indexes and the quadratic sum of deviations (Equations 6-8). Even so, by comparing the locations of significant stations in the maps, it can be found that locations of the top 10 stations with extreme low G values almost keep unchanged, while those of the bottom 10 ones vary with a and some stations in the middle area can't be identified when it takes a large value. Therefore, to ensure that regions with high G values, meaning where redundant information exists, can be identified, a should not be taken too large.

Scheme No.	Number of Stations	RTH (%)					
		a=1	a=5	a=10	a=20	a=50	a=100
1	5	**64.96**	55.32	38.49	32.27	23.48	19.49
2	15	**82.60**	74.83	65.02	59.47	39.41	39.49
3	25	**90.01**	82.05	78.76	74.83	51.12	48.38
4	35	**93.06**	85.97	82.93	82.38	70.74	55.19
5	45	**94.94**	91.69	88.02	89.75	80.06	66.07
6	55	**96.25**	95.81	94.01	93.60	84.87	74.82
7	65	97.64	**97.68**	96.35	96.17	89.22	81.00
8	75	**98.65**	98.64	98.14	97.67	95.68	86.99
9	85	**99.56**	99.39	99.42	99.33	98.50	94.32
10	95	100.00	100.00	100.00	100.00	100.00	100.00

Table 1 Joint entropy percentage of optimization schemes

Scheme No.	Number of Stations	RTC(%)					
		a=1	a=5	a=10	a=20	a=50	a=100
1	5	3.51	2.27	2.52	2.33	2.16	**2.11**
2	15	13.44	11.39	10.79	10.18	9.58	**9.16**
3	25	23.80	21.68	21.03	19.93	17.83	**17.52**
4	35	34.22	31.73	31.07	29.87	**26.51**	26.93
5	45	45.31	42.88	41.77	40.90	37.11	**36.45**
6	55	55.95	54.56	53.78	52.78	48.07	**46.40**
7	65	67.21	65.98	64.78	64.43	59.66	**57.35**
8	75	78.20	76.98	76.48	76.04	72.17	**70.29**
9	85	88.98	88.29	88.10	87.52	85.19	**83.85**
10	95	100.00	100.00	100.00	100.00	100.00	100.00

Table 2 Total correlation percentage of optimization schemes

To further explore the influence of discretization on network evaluation, we analysed the network optimization schemes based on different a values. We selected the top stations in each sequence to form 10 "optimization schemes", with the numbers of stations ranging from 5 to 95 (at an interval of 5). For the selected stations in each scheme, the proportions that their joint entropy and total correlation take up in those of the original network, denoted as RTH and RTC, were computed and compared, as shown in tables 1 and 2 separately. As table 1 shows, RTH varies greatly with a, especially when the scheme containing a small number of stations. The smaller the a, the higher the RTH value and the better the optimization scheme. For all the schemes, most of the maximum values of RTH are obtained at a=1 with an exception at a=2, meaning that smaller a makes for schemes with more information content. The other index RTC performs reversely, tending smaller as a increases, as shown in table 2. Most of the minimum RTC are obtained at a=100. Smaller RTC indicates less

information redundancy and better optimization schemes. As can be seen, joint entropy and total correlation are two contradictory targets. If one is optimized, the other tends to be worsen. Improving the information content index by using smaller parameter a may naturally cause more redundancy. Conversely, the reduction of redundancy leads to the decrease of information content. This can also be used to guide the selection of a.

5 Conclusions

This study explored the entropy based multicriterion evaluation of rainfall monitoring networks. In particular, this study aimed to access the impact of the discretization process. Two indexes were presented to evaluate each station, an entropy ratio index accounting for the information content and a correlation ratio index measuring overlapped information between stations. The indexes were integrated with ideal point method to an objective function, which can be used to sort the stations by their significance in the network. Moreover, the impact of the discretization process on evaluation results was investigated by altering the parameter value in the floor function. A rainfall monitoring network containing 95 stations in the western Taihu Lake, China was used for analysis, which gives rise to the following conclusions:

• Value of the parameter in floor function directly determines the entropy value, including its range scales and distribution patterns. Generally, smaller parameter values lead to higher entropy, larger ranging scale and more uniform distribution. This naturally makes the stations more distinguishable for easier decision making;

• The results of network evaluation are also affected by discretization. Locations of the top and bottom stations ranked by the multicriterion objective function vary when the parameter is altered. Some bottom stations cannot be identified when extremely large parameter is taken;

• Changing parameter values result in different performances of network optimization schemes. Smaller parameter values tend to generate schemes with more information content, while larger values correspond to less redundancy.

Above all, entropy is effective at evaluating rainfall monitoring networks by quantizing information. The multicriterion method can efficiently identify significant stations based on multiple targets, such as information capacity and redundancy. Parameter of the floor function used for data discretization directly determines the entropy value, as well as its ranging scale and distribution, and impact the evaluation results. As a consequence, it should be carefully selected and sensitivity analysis is needed to avoid the impacts of subjectivity and arbitrariness.

References

Yeh, H. C., Chen, Y. C., Wei, C., Chen, R. H. (2011). *Entropy and kriging approach to rainfall network design.* Paddy & Water Environment. 9.3 343-355.

Xu, H., Xu, C., Sælthun, N. R., Zhou, B., Xu, Y. (2015). *Entropy theory based multi-criteria resampling of rain gauge networks for hydrological modelling - a case study of humid area in southern China.* Journal of Hydrology. 525.A: 138-151.

Mishra, A. K., Coulibaly, P. (2009). *Developments in hydrometric network design: A review*. Reviews of Geophysics. 47.2.

Shannon, C. E. (1948). *A mathematical theory of communication*. Bell System. Technical Journal. 27.3: 379-423.

Krstanovic, P. F., Singh, V. P. (1992a). *Evaluation of rainfall networks using entropy: I. theoretical development*. Water Resources Management. 6.4: 279-293.

Krstanovic, P. F., Singh, V. P. (1992b). *Evaluation of rainfall networks using entropy: II. application*. Water Resources Management. 6.4: 295-314.

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

Mogheir, Y., Singh, V. P., Lima, J. L. M. P. D. (2006) *Spatial assessment and redesign of a groundwater quality monitoring network using entropy theory, Gaza strip, palestine*. Hydrogeology Journal. 14.5: 700-712.

Alfonso, L., Lobbrecht, A., Price, R. (2010). *Information theory–based approach for location of monitoring water level gauges in polders*. Water Resources Research. 46.3: 374-381.

Li, C., Singh, V. P., Mishra, A. K. (2012). *Entropy theory-based criterion for hydrometric network evaluation and design: Maximum information minimum redundancy*. Water Resources Research. 48.5: 5521.

Alfonso, L., He, L., Lobbrecht, A., Price, P. (2013). *Information theory applied to evaluate the discharge monitoring network of the Magdalena River*. Journal of Hydroinformatics. 15.1: 211-228.

Samuel, J., Coulibaly, P., Kollat, J. (2013). *CRDEMO: Combined regionalization and dual entropy-multiobjective optimization for hydrometric network design*. Water Resources Research. 49.12: 8070-8089.

Leach, K. J. M., Kornelsen, C., Samuel, J., Coulibaly, P. (2015). *Hydrometric network design using streamflow signatures and indicators of hydrologic alteration*. Journal of Hydrology. 529: 1350-1359.

Su, H. T., You, J. Y. *Developing an entropy-based model of spatial information estimation and its application in the design of precipitation gauge networks*. Journal of Hydrology. 519(2014): 3316-3327.

Ruddell, B. L. & Kumar, P. (2009). *Ecohydrologic process networks: 1. Identification, Water Resources Research*. 45.3: 450-455.

Mcgill, W. (1954). *Multivariate information transmission.* Psychometrika, 19.2: 97-116.

Watanabe, S. (1960) *Information theoretical analysis of multivariate correlation.* IBM Journal of Research and Development, 4.1: 66-82.