ENTROPY BASED ASSESSMENT OF HYDROMETRIC NETWORK USING NORMAL AND LOG-NORMAL DISTRIBUTIONS

N. Vivekanandan

Central Water and Power Research Station, Pune, Maharashtra, India

ABSTRACT

Establishment and maintenance of a hydrometric network in any geographical region is required for planning, design and management of water resources. Setting up and maintaining a hydrometric network is an evolutionary process, wherein a network is established early in the development of the geographical area; and the network reviewed and upgraded periodically to arrive at the optimum network. This paper presents the methodology adopted in assessing the hydrometric network using entropy theory adopting normal and log-normal probability distributions. The technique, involving computation of marginal and conditional entropy values, is applied to the upper Bhima basin up to Ujjani reservoir for illustrative purposes; and results presented. The derived optimum hydrometric network for the basin is evaluated based on WMO guidelines for minimum density of hydrometric network.

Keywords:

Entropy, Log-normal, Marginal, Network, Stream flow, Transinformation

1. INTRODUCTION

Hydrological information system (HIS) for a region provides requisite data for planning, design and management of water resources and related research activities; thus enabling informed decision-making. A hydrometeorological network includes subsystem for measuring stream flow, precipitation, groundwater, etc.; and provides water-data for HIS. For a river basin, the hydrometric network for measurement of stream flow characteristics forms a subsystem of hydrometeorological network [1]. Setting up and maintaining a hydrometric network is an evolutionary process, wherein a network is established early in the development of the geographical area; and the network reviewed and upgraded periodically to arrive at the optimum network. For network optimization, approaches commonly used include statistical approaches, user-survey technique, hybrid method and sampling strategies [2]. Statistical approaches for hydrometric network optimization range from clustering techniques, spatial regression methods in generalized least square framework and entropy-based methods.

Entropy theory quantifies the relative information content for the hydrometric network, and has the advantage that it needs only stream flow data for evaluation. The method facilitates network design by quantifying the marginal contribution of each data collection node to the overall information provided by the network using an index termed marginal entropy. Probability distributions such as gamma, normal and log-normal are commonly used for computation of entropy values while optimising hydrometric network [3-7]. In the present study, 2-parameter

normal (N2) and log-normal (LN2) distribution is used for computation of entropy values for the stations under consideration. Transinformation index measures the redundant or mutual information between stations, and is computed from marginal and conditional entropy values to derive the optimum network. The methodology adopted in assessing the hydrometric network of upper Bhima basin up to Ujjani reservoir using entropy theory is briefly described in the ensuing sections.

2. METHODOLOGY

The main objective of evaluating a data collection network is to identify the stations that are producing redundant or repeated information; which can be measured quantitatively by computing transinformation or the redundancy produced by each station in the network.

2.1 Concept of Entropy

A quantitative measure of the uncertainty associated with a probability distribution, or the information content of the distributions termed Shannon entropy [8] can be expressed as:

$$H(X) = -k\sum p_i \ln(p_i) \quad \ldots (1)$$

Here, H(X) is the entropy corresponding to the random variable X; k is a constant that has value equal to one, when natural logarithm is taken; and p_i represents the probability of i^{th} observation.

2.2 Marginal Entropy

Marginal entropy for the discrete random variable X is defined as:

$$H(X) = -k\sum_{i=1}^{N} p(X_i)\ln(p(X_i)) \quad \ldots (2)$$

where, $p(X_i)$ is the probability of i^{th} random variable X, which is computed by either N2 or LN2 distribution, and N is the number of observations. The probability density functions of normal and log-normal distributions are expressed by:

$$f(X;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}, \; X>0 \text{ (Normal)} \quad \ldots (3)$$

$$f(X;\alpha,\beta) = \frac{1}{X\beta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{LN(X)-\alpha}{\beta}\right)^2}, \; X>0 \text{ (Log-normal)} \quad \ldots (4)$$

Here, μ and σ are parameters of the N2 distribution. Similarly, α and β are parameters of the LN2 distribution. The marginal entropy H(X) indicates the amount of information or uncertainty that X has. If the variables X and Y are independent, then the joint entropy [H(X,Y)] is equal to the sum of their marginal entropy values and given by:

$$H(X,Y) = H(X) + H(Y) \quad \ldots (5)$$

If the variables are stochastically dependent, then the joint entropy is less than its total entropy [9].

2.3 Conditional Entropy

The conditional entropy of Y on X is defined by:

$$H(Y/X) = H(X,Y) - H(X) \quad \ldots (6)$$

Conditional entropy value becomes zero, if the value of one variable is completely determined by the value of other variable. If the variables are independent, then $H(Y/X) = H(Y)$ [10].

2.4 Transinformation Index

Transinformation is the form of entropy that measures the redundant or mutual information between variables. Transinformation represents the amount of information, which is common to two stochastically dependent variables X and Y. The transinformation between X and Y is defined as:

$$T(X, Y)=H(X)+H(Y)-H(X, Y)=H(Y)-H(Y|X) \quad \ldots (7)$$

For independent X and Y, T(X,Y) = 0. The value of T(X, Y) is known as transinformation index [11].

2.5 Steps Involved in Computation of Entropy Values

In practice, the existing sampling sites of a hydrometric network can be arranged in the order of information content. In the ordered list thus obtained, the first station is the one where the highest uncertainty about the variable occurs, and the subsequent stations serve to reduce the uncertainty further [12-13]. The steps involved in selecting the best combination of stations using entropy theory are as follows:

i) Let the data collection network under review, consists of M monitoring stations. The data series of the variable of interest at each station ($X_1,X_2,\ldots,X_M$) is represented by X_{ij} , where 'i' denotes the station identification number (i=1, 2,…, M) and 'j' is for time period (j=1,2,…,N). The data length at all stations is assumed to be equal to N. The best fitted multivariate joint probability density function for the subset (X_1, $X_2,\ldots,X_M$) of M monitoring stations is selected.

ii) The marginal entropy of the variable $H(X_i)$ (i=1,2,…,M) for each station is calculated. The station with the highest marginal entropy is denoted as the first priority station $Pr(Xz_1)$. This is the location, where the highest uncertainty occurs about the variable and hence information-gain will be highest from the observations recorded at this site.

iii) This station $Pr(Xz_1)$ is coupled with every other (M-1) stations in the network to compute transinformation $T(X_i, Pr(Xz_1))$ with $X_i \neq Pr(Xz_1)$, i=1, 2,…M; and to select that pair, which gives the least transinformation. The station that fulfils this condition is marked as the second priority location $Pr(Xz_2)$.

iv) The pair $(Pr(Xz_1), Pr(Xz_2))$ is coupled with every other (M-2) station in the network to select a triplet with the least transinformation $T(X_i; Pr(Xz_1), Pr(Xz_2))$. The same procedure is continued by successively considering combinations of three and more stations, and selecting the combination that produces the least transinformation.

Finally, all M monitoring stations ($X_1,X_2,\ldots,X_M$) can be ranked in priority order to get ($Pr(Xz_1),Pr(Xz_2),\ldots,Pr(Xz_M)$).

v) It is possible to terminate the above process early, before carrying out for all M stations by selecting a particular threshold transinformation value as the amount of redundant information to be permitted in the network, such that sampling of the variable may be stopped at the stations that exceed the threshold to get optimum number of stations, which is less than M.

In the above procedure, the benefits for each combination of sampling sites are measured in terms of least transinformation or the highest conditional entropy produced by that combination. The above procedure helps to assess network configurations with respect to the existing stations. If new stations are to be added to the system, their locations may be selected again on the basis of the entropy theory by ensuring maximum gain of information.

3. APPLICATION

3.1 Study Area and Data Used

The methodology detailed above has been applied to optimize the hydrometric network for upper Bhima basin up to Ujjani reservoir. The basin is located in the western part of Maharashtra between 17^o 53' N and 19^o 24' N latitude and 70^o 20' E and 75^o 18' E longitude. The geographical area of the basin is 14712 km^2. Of the total geographical area under study, 25% is hilly and/or highly dissected, 55% plateau and 20% is remaining plain area [14].

Figure 1 gives a location map of upper Bhima basin up to Ujjani reservoir. From Figure 1, it may be noted that there are 14 stream gauge stations located in the upper Bhima basin up to Ujjani. From scrutiny of historical stream flow data, it was noted that two stations, namely Aamdabad and Pimple Gurav are having only four years of data, which was considered inadequate, and hence not considered for further analysis. Stream flow data recorded at the remaining twelve stations, namely Askheda, Budhawadi, Chaskaman, Dattawadi, Kashti, Khamgaon, Nighoje, Pargaon, Paud, Rakshewadi, Shirur and Wegre for the period 1994-2007 were used in deriving the optimum network.

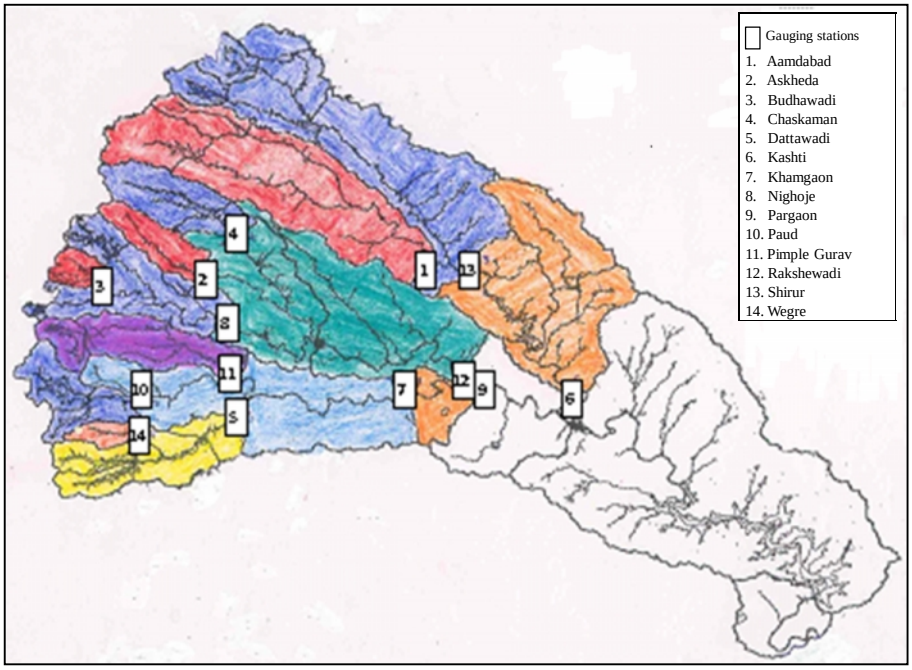

Figure 1: Location map of upper Bhima basin

4. RESULTS AND DISCUSSIONS

By applying the procedure detailed above, a computer program was developed and used to compute the transinformation index from marginal and conditional entropy. The program identify the first priority station based on marginal entropy; compute the conditional entropy with reference to first priority station; and arrange the stream gauge stations in order of priority based on transinformation index.

4.1 Computation of Marginal Entropy

Table 1 gives the summary statistics of annual average flow and marginal entropy values (using N2 and LN2 distributions) for twelve stream gauge stations of upper Bhima basin.

Table 1: Summary statistics of annual average flow and indices of marginal entropy

S. No.	Gauging station	Summary statistics of annual average flow				Marginal entropy	
		N2		LN2			
		μ (10^2 m^3/s)	σ (10^2 m^3/s)	α	β	N2	LN2
1	Askheda	23.7	14.0	7.600	0.399	8.662	0.959
2	Budhawadi	15.9	9.5	7.187	0.427	8.278	0.993
3	Chaskaman	34.0	24.5	7.894	0.537	9.222	1.108
4	Dattawadi	52.7	63.3	7.732	2.581	10.173	**1.893**
5	Kashti	54.5	57.6	7.828	2.546	10.078	1.886
6	Khamgaon	179.3	115.5	9.648	0.279	10.774	0.781
7	Nighoje	81.8	35.1	8.928	0.173	9.582	0.541
8	Pargaon	348.6	220.3	10.280	0.389	**11.419**	0.946
9	Paud	32.0	25.8	7.835	0.461	9.274	1.032
10	Rakshewadi	120.0	81.9	9.268	0.210	10.429	0.638
11	Shirur	75.6	50.0	8.590	1.216	9.935	1.517
12	Wegre	24.0	11.2	7.693	0.184	8.438	0.572

μ and σ are mean and standard deviation of annual average flow for N2; α and β are mean and standard deviation of the log-transformed data of annual average flow for LN2.

4.2 Transinformation Index Matrix

Based on marginal entropy values, as given in Table 1, it was identified that the Pargaon is the first priority station when N2 distribution applied whereas Dattawadi is the first priority station when LN2 distribution applied. The first priority station was coupled with other 11 stations individually to identify the next priority station in order, and to compute the transinformation index. By using marginal entropy values obtained from probability distributions, joint and conditional entropy values were computed to arrive at a transinformation matrix for the stations under study. Tables 2 and 3 give the transinformation index matrix (using N2 and LN2) for the stations under study.

Table 2: Transinformation index matrix based on marginal entropy using N2 distribution

Station	Transinformation index matrix (T)										
	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Askheda	0.465	0.302	-	-	-	-	-	-	-	-	-
Budhawadi	0.821	0.714	0.710	0.697	0.697	0.704	-	-	-	-	-
Chaskaman	0.712	0.501	0.497	0.517	0.600	-	-	-	-	-	-
Dattawadi	0.660	0.799	0.848	0.987	1.613	1.450	1.457	1.538	1.731	1.671	-
Kashti	0.407	0.525	0.550	0.875	0.722	0.793	0.784	-	-	-	-
Khamgaon	0.733	0.527	0.557	0.648	1.146	1.157	1.329	1.411	1.398	1.698	1.687
Nighoje	0.763	0.425	0.460	0.481	-	-	-	-	-	-	-
Pargaon	-	-	-	-	-	-	-	-	-	-	-
Paud	0.929	0.920	1.005	1.108	1.252	1.289	1.546	1.360	1.314	-	-
Rakshewadi	0.281	-	-	-	-	-	-	-	-	-	-
Shirur	0.578	0.486	0.618	1.071	0.747	0.796	0.900	0.880	-	-	-
Wegre	0.626	0.345	0.373	-	-	-	-	-	-	-	-
Si (i=2, 3, 4,..... 12) indicate the steps involved in computation of transinformation index based on marginal entropy of first priority station, namely Pargaon. From the table, the station having the least transinformation index at each step is considered as the next priority station to Pargaon.											

Table 3: Transinformation index matrix based on marginal entropy using LN2 distribution

Station	Transinformation index matrix (T)										
	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Askheda	0.192	0.197	-	-	-	-	-	-	-	-	-
Budhawadi	0.403	0.418	0.472	0.474	0.772	0.791	0.808	-	-	-	-
Chaskaman	0.153	0.203	0.254	-	-	-	-	-	-	-	-
Dattawadi	-	-	-	-	-	-	-	-	-	-	-
Kashti	0.869	0.952	0.963	0.964	0.994	1.072	1.147	1.313	1.519	2.032	3.055
Khamgaon	0.476	0.489	0.604	0.619	0.619	-	-	-	-	-	-
Nighoje	0.349	0.357	0.649	0.794	0.796	0.816	0.982	1.848	1.865	1.871	-
Pargaon	0.591	0.609	0.819	0.932	0.956	1.041	1.174	1.422	1.433	-	-
Paud	0.326	0.408	0.854	0.858	0.946	0.987	1.002	1.052	-	-	-
Rakshewadi	0.088	-	-	-	-	-	-	-	-	-	-
Shirur	0.246	0.286	0.397	0.424	-	-	-	-	-	-	-
Wegre	0.220	0.254	0.377	0.597	0.642	0.773	-	-	-	-	-
Si (i=2, 3, 4,..... 12) indicate the steps involved in computation of transinformation index based on marginal entropy of first priority station, namely Dattawadi. From the table, the station having the least transinformation index at each step is considered as the next priority station to Dattawadi.											

4.3 Computation of Redundant Information

Based on transinformation index values obtained from N2 and LN2 distributions, the amount of redundant information passed by different stations were computed and given in Tables 4 and 5.

Table 4: Contribution of redundant information by different stations (using N2 distribution)

Station	Transinformation index	Redundant information (%)
Rakshewadi	0.281	16.7
Askheda	0.302	17.9
Wegre	0.373	22.1
Nighoje	0.481	28.5
Chaskaman	0.600	35.6
Budhawadi	0.704	41.7
Kashti	0.784	46.5
Shirur	0.880	52.2
Paud	1.314	77.9
Dattawadi	1.671	99.1
Khamgaon	1.687	100.0

Table 5: Contribution of redundant information by different stations (using LN2 distribution)

Station	Transinformation index	Redundant information (%)
Rakshewadi	0.088	2.9
Askheda	0.197	6.4
Chaskaman	0.254	8.3
Shirur	0.424	13.9
Khamgaon	0.619	20.3
Wegre	0.773	25.3
Budhawadi	0.808	26.4
Paud	1.052	34.4
Pargaon	1.433	46.9
Nighoje	1.871	61.2
Kashti	3.055	100.0

After summarizing the results, the following observations are drawn from the study.

i) Khamgaon station provided 100% redundant information when N2 distribution applied. The amount of redundant information passed by the stations other than Paud and Dattawadi vary between about 17% and 52%.
ii) Kashti station provided 100% redundant information when LN2 distribution applied. The amount of redundant information passed by the stations other than Nighoje vary between about 3% and 47%.
iii) Based on marginal entropy value obtained from LN2 distribution, Dattawadi is the first priority station and hence not to be considered for discontinuation while optimizing the network though the redundant information passed by the station is about 99% while N2 distribution applied for optimizing the network.
iv) By considering the 100% redundant information, Khamgaon and Kashti stations are proposed for discontinuation from the existing network of upper Bhima basin.
v) In addition, Aamdabad and Pimple-Gurav stations are also proposed for discontinuation from the existing network because of inadequacy of stream flow data.
vi) The derived optimum network for the basin consists of ten stream gauge stations with 1471 km^2 per gauging station; which satisfies the WMO [15] recommended value of 1875 km^2 per station for minimum density of hydrometric network.
vii) However, it should be noted that this is a preliminary analysis and demonstrates the application of entropy theory for network evaluation, more analysis need to be done considering all the stations for final recommendations.

5. CONCLUSIONS

The paper presented a computer aided procedure for the assessment of hydrometric network using entropy theory adopting normal and log-normal probability distributions for upper Bhima basin up to Ujjani reservoir. The results of N2 distribution showed that the amount of redundant information passed by the stations other than Paud and Dattawadi vary between about 17% and 52%. Similarly, the amount of redundant information passed by the stations other than Nighoje and Kashti vary between about 3% and 47% when LN2 distribution applied for network optimization.. The study suggested that the two pairs of stations Khamgaon-Kashti (based on 100% redundant information) and Aamdabad-Pimple Gurav (based on inadequacy of stream flow data) may be considered for discontinuation from the existing hydrometric network of the basin. The paper presented that the derived optimum network of upper Bhima basin consists of ten stream gauge stations with 1471 km^2 per gauging station, which satisfy the WMO recommended value of 1875 km^2 per station for minimum density of hydrometric network. The results presented in the paper would be helpful to the stakeholders for decision making as regards network optimization in upper Bhima basin.

ACKNOWLEDGEMENTS

The author is grateful to the Director, Central Water and Power Research Station, Pune, for providing the research facilities to carry out the study. The author is thankful to the Chief Engineer, Water Resources Department, Government of Maharashtra, for supply of stream flow data of upper Bhima basin.

REFERENCES

[1] Yang Y. and Burn D.H. (1994); An entropy approach to data collection network design, Journal of Hydrology, Vol. 157 (4), pp 307-324.

[2] N. Vivekanandan (2012); Evaluation of Stream Flow Network using Entropy Measures of Normal and Lognormal Distributions, Bonfring Journal of Industrial Engineering and Management Science, Vol. 2 (3), pp 33-37.

[3] Ozkul S., Harmanucioglu N.B. and Singh P.V. (2000); Entropy based assessment of water quality monitoring networks, ASCE Journal of Hydrologic Engineering, Vol. 5(1), pp 90-99.

[4] Markus M., Knapp H.V. and Tasker G.D. (2003); Entropy and generalized lease square methods in assessment of the regional value of stream gauges, Journal of Hydrology, Vol. 283 (1-2), pp 107-121.

[5] Sarlak N. and Sorman U.A. (2006); Evaluation and selection of stream flow network stations using entropy methods, Turkey Journal of Engineering and Environmental Science, Vol. 30 (2), pp 91-100.

[6] Yoo C., Jung K. and Lee J. (2008); Evaluation of rain gauge network using entropy theory: Comparison of mixed and continuous distribution function applications, ASCE Journal of Hydrologic Engineering, Vol. 13 (4), pp 226-235.

[7] Jairaj P.G. and Remya A.R. (2009); Rainfall station network optimization using entropy method, Proc. of International Conference on Water, Environment, Energy and Society, New Delhi, 12-16 January 2009, pp 891-897.

[8] Shannon C.E. (1948); A mathematical theory of communication, The Bell System Technical Journal, Vol. 27, October issue, pp 625-656.

[9] N. Vivekanandan, S.K. Roy and A.K. Chavan (2012); Evaluation of Rain Gauge Network using Maximum Information Minimum Redundancy Theory, Journal of Scientific Research and Reviews, Vol. 1 (3), pp 96-107.

[10] N. Vivekanandan and Rahul S. Jagtap (2012); Optimization of Hydrometric Network using Spatial Regression Approach, Bonfring Journal of Industrial Engineering and Management Science, Vol. 2, Special issue, pp 56-61.

[11] Jordan A. Clayton and Jason W. Kean (2010); Establishing a multi-scale stream gauging network in the Whitewater river basin, Kansas, USA, Water Resources Management, Vol. 24 (13), pp 3641-3664.
[12] N. Vivekanandan and Rahul S. Jagtap (2012); Evaluation and Selection of Rain Gauge Network using Entropy, Journal of Institution of Engineers (Series A), Vol. 93 (4), pp 223-232.
[13] N Vivekanandan and M. Janga Reddy (2011); Evaluation of Streamflow Network Using Entropy Method, Proc. on 'Hydraulics and Water Resources (HYDRO 2011)' organized by ISH at SVNIT, Surat during 29-30 December, pp 1031-1036.
[14] Government of Maharashtra (GoM, 1999); Report of the Maharashtra water and irrigation commission,Vol. 5.
[15] World Meteorological Organization (WMO, 1994); Guide to Hydrological Practices, Report No. 168, Geneva