Hydrologic Model Performance Evaluation Applying the Entropy Concept as a Function of Precipitation Network Density

Mariano Hernandez, Leonard J. Lane, Jeffry J. Stone, J. Guillermo Martinez , and Mary Kidwell:
USDA- ARS Southwest Watershed Research Center and the University of Arizona, 2000 East Allen Rd., Tucson, AZ 85719 USA

Abstract Data from the Walnut Gulch Experimental Watershed are used with a distributed hydrologic model to examine the effects of decreasing the number of raingauges in the network on simulated runoff. When modeling a watershed using distributed models, the density of the network of raingauges can drastically affect the model results and calibration. The entropy concept is used as a measure of information content in distributed data from a raingauge network. This concept is used in this study to assess how much loss of information about simulated runoff occurs as a results of subtracting one raingauge successively from the existing 10 gauge network while maximizing the information contained in the remaining network. The experimental watershed is subdivided into a fine resolution to capture the spatial variability of the major driving processes affecting runoff at the watershed scale. The optimal information transmission is calculated for each set of combinations of raingauges and each set is used as rainfall input in the hydrologic model. The performance of the hydrologic model is assessed by computing the entropy for different network densities and by comparing annual runoff simulations with observed data for an 8 year period. The value of the optimal information transmission increases with increasing the number of raingauges up to a number of raingauges that no longer captures new information and the model results are not further improved. The methodology may be applied when economic or physical constraints affect existing hydrological networks and the number of stations need to be reduced. However, this study demonstrated that the entropy concept and a distributed hydrologic model are required to determine the minimum raingauge network required for runoff simulation.

1. INTRODUCTION

Numerous papers have been published in the past few years concerning the optimum recording of water resources data. Many of these studies have been concerned with sampling data other than precipitation data, but they are all pertinent, at least in part, to establishing or determining optimum raingauge densities. Obviously, before one can establish an optimum raingauge network or determine an optimum raingauge density, the limits and requirements for the measurement of the input (rainfall) and output (runoff) to the system must be defined. Almost all runoff from small rangeland watersheds in the southwestern United States is the result of intense thunderstorm rainfall, and the variability of this rainfall is an important runoff-influencing factor in such areas where high intensity rainfall dominates watershed hydrology (Osborn, 1983). The climate of the region may be a consideration in determining the optimum raingauge density. If rainfall is primarily low intensity over relatively large areas, a few widely spaced gauges may be satisfactory. If most of the precipitation is from high intensity thunderstorm cells of limited areal extent, as is the case in much of the southwest US, more raingauges may be needed for the same areal coverage. In addition to climatic differences, the goals of the data collection program must be taken into account, e.g., if only mean annual rainfall is needed for a particular study, sparse networks may be in order. If definition of the individual thunderstorm cells and the spatial variation in intensities within these cells is wanted, much more intense gauging is obviously necessary. Osborn et al. (1972) reported that to correlate daily rainfall and runoff, a 2.59 km^2 watershed with a length-width ratio of 4 requires a network of three recording raingauges. For watersheds of approximately 0.5 km^2 or less, the optimum network for rainfall-runoff correlation is one recording raingauge. They pointed out that, generally, the number of gauges required per unit area decreases as the watershed size increases up to about 26 km^2. A network of gauges located at 2.4-km intervals is necessary to adequately correlate the thunderstorm rainfall and runoff for watersheds of greater than 26 km^2.

In this paper, an attempt is made to assess how much loss of information about simulated runoff may occur as a result of successively reducing one raingauge from the existing network in the Walnut Gulch Experimental Watershed in Tombstone, Arizona while maximizing the performance of the distributed hydrologic model. The concepts of entropy and transinformation are used to determine the optimal combination of raingauges retained for each raingauge network configuration. The distributed model along with each set of raingauge densities are used to assess the effects of raingauge network density on annual runoff volume and peaks.

2. METHODOLOGY

Amorocho and Espildora (1973) used the concept of entropy, as introduced by Shannon (1948) and Shannon and Weaver (1949), to characterize uncertainty in hydrologic data. Chapman (1986) extended this concept to a measure of uncertainty in hydrologic data and a means to reduce that uncertainty through the application of a model. Harmancioglu and Yevjevich (1987) applied the concept of entropy in transferring hydrologic information between river points, using the bivariate or multivariate, normal and lognormal distributions. Krstanovic and Singh (1988) applied the entropy approach to space and time evaluation of rainfall networks in Louisiana. Space and time dependencies amongst raingauges were examined by auto-covariance and cross-covariance matrices. Husain (1989) used the entropy concept to estimate regional hydrologic uncertainty and information at both gauged and ungauged grids in a basin using rainfall data. His results show that the entropy method presents a convenient means of evaluating an optimum spatial design with respect to both the number and location of gauging stations.

2.1 Concepts of Information Transmission and Entropy

We consider a discrete random variable X which can take values $x_1,x_2,\dots,x_n$, with probabilities $p_1, p_2,\dots,p_n$; $P(X=x_1)=p_1$, $P(X=x_2)=p_2,\dots,P(X=x_n)=p_n$. P(x) is the probability distribution of X, satisfying

$$P(x) = (p_1, p_2,\dots,p_n);\quad \sum_{i=1}^{n} p_i = 1;\quad p_i \geq 0,\ i= 1,2,..n$$

There are several definitions of the concept of information. Here, we will give the traditional definition (Shannon, 1948). The information contained in the random variable X is given by

$$\mathrm{Inf}[x_1,\dots,x_m] = -\ln [p(x_1,\dots, x_m)] \qquad (1)$$

where the logarithm can be taken with arbitrary base b>1. When the logarithm is taken to the base b=2, the unit of entropy scale is called a "bit"; when the natural logarithm to base e is taken, the unit is called a "nit". According to Eq. (1), the more informative the random vector X, the less probable it is to occur. Closely related to the concept of information is the notion of entropy, defined as follows (Shannon, 1948).

$$H(X) = -\sum_{i=1}^{n} p_i \ln p_i \qquad (2)$$

where p_i is the probability of event i.

If X is continuous then

$$H(X) = -\int_0^{\infty} f(x)\ln f(x)dx \qquad (3)$$

where f(x) is the probability density function (pdf) of X. If two random variables are dependent then the Shannon entropy of the joint distribution is

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ln p(x_i,y_j) \qquad (4)$$

where $p(x_i,y_j)$ is the joint probability density function. If X and Y are continuous then

$$H(X,Y) = -\int_0^{\infty}\int_0^{\infty} f(x,y)\ln f(x,y)dxdy \qquad (5)$$

where f(x,y) is the joint probability density function (pdf) of X and Y.

The amount of information on a random variable X in another variable Y is defined by

$$T(X;Y) = H(X) + H(Y) - H(X,Y) \qquad (6)$$

In the field of Information Theory equation (6) is referred to as the Transinformation of the channel. The physical meaning of equation (6) is that while observing a natural process Y one receives valuable information about another process X that is dependent on Y.

Let X represent rainfall measured at a station with events $(x_1,x_2,\dots,x_N)$ with the probability of occurrence of the j^{th} event denoted by $p(x_j)$. The average entropy can be calculated with N events using equation (2) as:

$$H(X) = -\sum_{j=1}^{N} p(x_j)\ln p(x_j) \qquad (7)$$

Similarly, the average joint entropy in a region with "m" stations with hydrologic variables $(X_1,X_2,\dots,X_m)$ can be extended according to equation (4) as :

$$H(X_1,X_2,..,X_m) = -\sum_{j=1}^{N} p(x_j^1,x_j^2,..,x_j^m)\ \ln p(x_j^1,x_j^2,..,x_j^m) \qquad (8)$$

where $(X_1,X_2,\dots,X_m)$ are the ensemble of "m" precipitation variables measured at "m" stations, and $p(x_j^1,x_j^2,..,x_j^m)$ is the joint probability of occurrence of the j^{th} event at "m" stations.

Consider a case where a single station "p" is to be retained from a dense network with "m" stations. The criterion for selecting a single station is based on maximization of information transmitted by the station about the region, which is based on the information additivity assumption (Husain, 1989). Under this assumption, the information transmitted by a station about a set of station locations is equivalent to the summation of the information transmitted by that station about each individual station location. It can be mathematically expressed as:

$$\text{MAX} \;\; T(X_1, X_2, .., X_m; X_p) =$$

$$\text{MAX} \;\; \sum_{i=1}^{m} T(X_i; X_p) =$$

$$H(X_p) + \sum_{i=1}^{m-1} T(X_i; X_p) \qquad (9)$$

$$(i \neq p) \qquad (i = 1,2,..,m)$$

where $T(X_1,X_2,\ldots,X_m;X_p)$ is the total information transmitted by station "p" about the region, and $T(X_i;X_p)$ is the information transmitted by station "p" about individual station "i" and is equal to Eq. (6). Details about the entropy and information transmission concepts and their computations are discussed by Caselton and Husain (1980).

The selection of a set of "q" stations (k, r,…,q) from a dense network of "m" stations is based on the information maximization principle (Husain, 1989)

$$\text{MAX} \;\; T(X_1, X_2, .., X_m; X_k, X_r, .., X_q) =$$

$$\text{MAX} \;\; \sum_{i=1}^{m} T(X_i; X_k, X_r, .., X_q) =$$

$$\text{MAX} \begin{bmatrix} H(X_k) + H(X_r) + \ldots H(X_q) + \\ \sum_{i=1}^{m-q} \sum_{j=1}^{q} T(X_i; X_j) \end{bmatrix} \qquad (10)$$

Taking various combinations of (k,r,..,q), the information transmission is calculated. The combination of these "q" stations that gives maximum information is retained.

3. CASE STUDY: THE WALNUT GULCH EXPERIMENTAL WATERSHED

3.1 General Description of the Study Area

Walnut Gulch is an ephemeral tributary of the San Pedro River, with the confluence near Fairbank, Arizona. The watershed is operated by the USDA-ARS Southwest Watershed Research Center in Tucson, Arizona. For a detailed description of the study area see Renard et al. (1993). The watershed is approximately 149 km^2 in size, with elevations ranging between 1190 and 2150m AMSL. Based on records from 1956-80 (Osborn, 1983), annual precipitation varied from 170 mm in 1956 to 378 mm in 1977; summer rainfall varied from 104 mm in 1960 to 290 mm in 1966; winter precipitation varied from 25 mm in 1966-67 to 233 mm in 1978-79. The annual precipitation falls during two distinct periods with greatly different characteristics. Winter precipitation, amounting to about one-third of the annual total, occurs as rain or snow of wide areal extent and low intensity. Small amounts of runoff have been recorded from winter storms. Most of the remaining two-thirds of the precipitation falls during July, August, and September as a result of intense convective thunderstorms of limited areal extent. Practically all runoff results from this type of storm. An extensive network of raingauges and runoff measuring devices distributed across the watershed allows the quantification of temporal and spatial variability in rainfall and runoff events, which can be highly variable both in timing and volume. Runoff from sub-watersheds is measured with a variety of gauging structures including broad-crested V-notch weirs, H-flumes, and supercritical flow structures. A network of 85 recording gauges is in place to measure rainfall.

One sub-watershed (Fig.1) was selected to assess how much information transmission loss on runoff may occur as a result of successively subtracting one raingauge from the raingauge network while maximizing the information contained in the remaining network. The sub-watershed is approximately 8.24 km^2 in size. Vegetation within the watershed is representative of the transition zone between the Chihuahuan and Sonoran deserts, and consists primarily of shrub-steppe and grassland rangeland vegetation.

Nine raingauges are located within the sub-watershed and one raingauge outside the area (Fig.1).

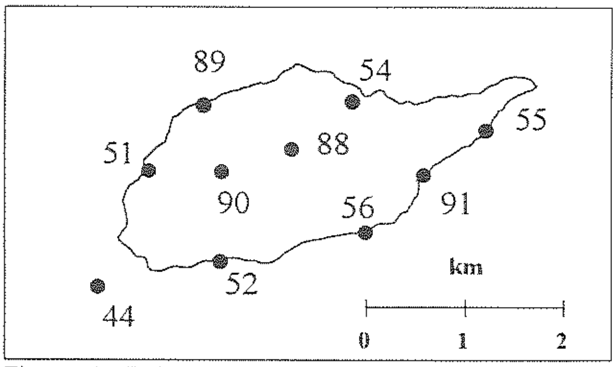

Figure 1. Raingauge network for the nested sub-watershed.

3.2 Model Description

The ARDBSN model (Renard et al., 1987) was developed by researchers at the USDA-ARS Southwest Watershed Research Center in Tucson, Arizona. The major processes included in the model are surface runoff, percolation, evapotranspiration, pond storage, and sedimentation. Since the model maintains a continuous water balance, complex basins are subdivided to reflect differences in evapotranspiration for various crops, soil, etc. Thus, runoff is simulated separately for each subarea and routed to obtain the total runoff for the basin. Since the model operates on a daily time step, surface runoff is simulated for daily rainfall using a modification of the SCS curve number method.

4. RESULTS

The ARDBSN model and ten different rainfall network density configurations were used to simulate annual runoff volume and annual peak runoff intensity. The model was calibrated using the maximum number of raingauges (Fig. 1) and observed annual runoff volume and peak runoff data collected between 1967 and 1974. The sub-watershed geometry was hydraulically represented by 5 channels and 12 overland flow elements. Results of the calibration of annual runoff volume and peak runoff intensity and the Nash-Sutcliffe (1970) efficiency coefficients are shown in Figs. 2 and 3.

The average entropy H(X) and the average joint entropy H(X,Y) of rainfall measured at each raingauge were computed using Eqs. (7) and (8), respectively, for the months of July, August, and September with 39 years of record. The analyses were repeated for the 8 years of record used in the calibration and the overall results were very similar to those presented below. Data were first converted to discrete form by subdividing, into 17 equal intervals of 5 mm of rainfall, the range between zero and the highest daily precipitation in the record (87mm). The discrete single and joint probabilities required to compute the entropy values were obtained by conducting a relative frequency analysis. No attempt was made in this paper to fit a theoretical distribution to the data. A preliminary analysis suggests that the data may fit a gamma or lognormal distribution. Once it is determined which distribution the data best fit, the average entropy and the average joint entropy can be computed using Eqs. 3 and 5, respectively.

Using the information maximization principle (Eqs. 9 and 10), the optimal locations for all n = 1 to 10 rainfall network density configurations were computed. Table 1 summarizes the optimal values of information and configurations.

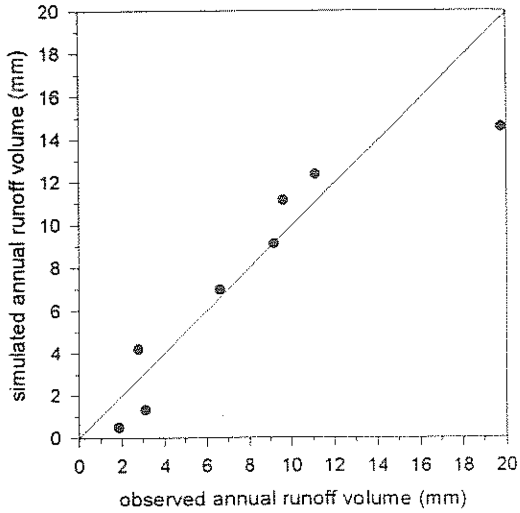

Figure 2. Calibration annual runoff volume for the period 1967 - 1974.

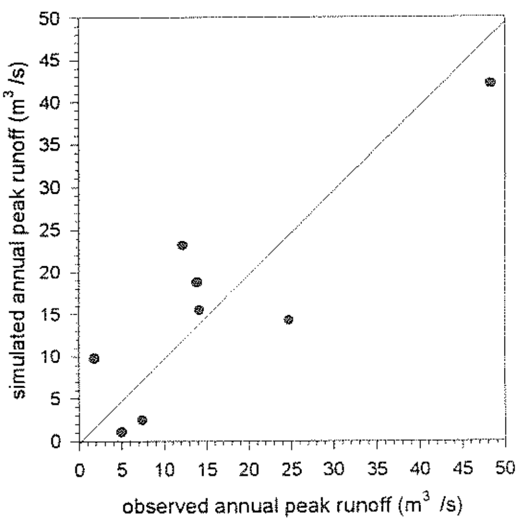

Figure 3. Calibration annual peak runoff for the period 1967- 1974.

Table 1. Summary of Optimal Networks and Information Transmitted for n=1 to 10.

n	OIT* (nits)	Optimal raingauge configurations
1	4.53	88
2	7.89	88,90
3	10.61	88,90,91
4	12.35	52,88,90,91
5	13.44	52,54,89,90,91
6a	14.56	44,52,55,56,88,89
6b	14.96	44,51,54,55,56,90
7	15.14	44,52,54,55,56,88,89
8	15.19	44,51,52,54,55,88,89,91
9	15.22	44,51,52,54,55,56,89,90,91
10	15.27	44,51,52,54,55,56,88,89,90,91

* Optimal Information Transmission

In Table 1, two values of information transmission were reported for a six-raingauge network configuration 6a and 6b. Obviously, the highest value (14.96 nits) corresponds to the optimal information transmission and the second to the next highest value. Notice that in 6b (44,51,54,55,56,90) raingauge 52 is not included. On September 10, 1967 a major runoff-producing event occurred on Walnut Gulch (Osborn et al., 1980). Raingauge 52 measured 87 mm of rainfall in 1hr. 20 min. Nearby stations (44,56,90) (Fig. 1) registered 74, 53, and 45 mm of rainfall, respectively. It is clear that the spatial behavior of this extreme event was not captured with the optimal six-raingauge configuration (6b). The set yielding the second highest information transmission (6a) includes raingauge 52. As can be seen from Table 1, the gain or loss of information between the two sets is of 0.40 nits. While the difference may not be significant, the incorporation of station 52 plays an important role in the simulation of runoff volume and peak runoff. This illustrated the value of using a distributed hydrologic model in the analyses.

The response of the watershed to different network density configurations is summarized in Table 2 and Figs. 4 and 5. The results suggest that the information about the rainfall process in the sub-watershed can be characterized with seven raingauges, and that the reduction of the uncertainty by successively adding one more raingauge to the network is very small. The normalized increment of information from seven raingauges to ten is just 0.008. It is important to notice that the model performed poorly using the optimal six-raingauge configuration(6b). The efficiency coefficients for the (6b) configuration for both runoff volume and peak runoff were 0.32 and 0.18, respectively (Figs. 4 and 5). The maximum efficiency was reached with seven raingauges for runoff volume (Fig.4) and with five for peak runoff (Fig. 5). The model geometry complexity representing the watershed and parameters obtained during the calibration remained constant for all rainfall density configurations. It can be argued that by keeping the geometry representation of the watershed constant, the maximum efficiency of the model was reached at different network densities. Similarly, the decrease of efficiency of the model with 8, 9, and 10 raingauges in the network can be attributed to the level of model resolution representing the watershed. That is, to account for the small reduction of uncertainty by adding more raingauges to the network, a more complicated representation of the geometry is required.

Table 2. Summary of Model Performance under different rainfall network densities.

n	Normalized Optimal Information Transmission	Efficiency Coefficient for Runoff Volume	Efficiency Coefficient for Runoff Peak
1	0.296	0.40	0.59
2	0.517	0.19	0.45
3	0.695	0.36	0.55
4	0.809	0.79	0.78
5	0.880	0.82	0.79
6a	0.953	0.88	0.76
6b	0.979	0.32	0.18
7	0.992	0.88	0.75
8	0.995	0.85	0.66
9	0.997	0.78	0.65
10	1.000	0.80	0.67

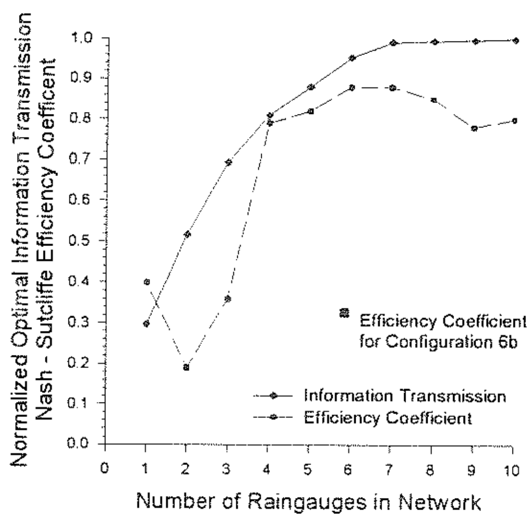

Figure 4. Normalized Optimal Information Transmission and Nash-Sutcliffe Efficiency Coefficient for Runoff Volume.

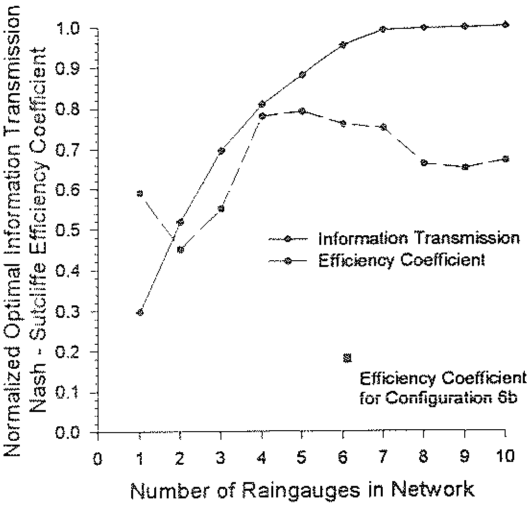

Figure 5. Normalized Optimal Information Transmission and Nash-Sutcliffe Efficiency Coefficient for Peak Runoff.

5. CONCLUSIONS AND RECOMENDATIONS

The entropy concept along with a continuous hydrologic model were used to assess the contribution of different raingauge network density configurations to the reduction of total uncertainty of watershed runoff volume and peak runoff. The entropy concept should not be used as the only criterion for reducing the number of stations in case that they do not contribute significantly to uncertainty reduction. A distributed simulation model should be used to assess the response of the watershed to the elimination of a certain station. A sensitivity analysis should be carried out to determine the effects of model geometry complexity on model efficiency to different optimal network densities.

6. REFERENCES

Amorocho, J. and B. Espildora. Entropy in the assessment of uncertainty in hydrologic systems and models. *Water Resour. Res.*, 9(6), 1511-1522, 1973.

Caselton, W. F. and T. Husain. Hydrologic networks: Information transmission. *Journal of the Water Resour. Planning and Management Division*, ASCE, Vol. 106, No. WR2, 503-520, July, 1980.

Chapman, T. G. Entropy as a measure of hydrologic data uncertainty and model performance, *Journal of Hydrology*, 85, 111-126,1986.

Harmancioglu, N. and V. Yevjevich. Transfer of hydrologic information among river points, *Journal of Hydrology*, 91, 103-118, 1987.

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

Krstanovic, P. F. and V. P. Singh, Application of entropy theory to multivariate hydrologic analysis, Vol 1, Water Resources Program, Department of Civil Engineering, Technical Report WRR9, 271 pp., Louisiana State University, Baton Rouge, Louisiana, 1988.

Nash, J. E. and J. V. Sutcliffe, River flow forecasting through conceptual models, I, A discussion of principles, *Journal of Hydrology*, 10, 282-290, 1970.

Osborn, H. B., Precipitation characteristics affecting hydrologic response of southwestern rangelands, USDA-ARS Agric. Rev. and Manuals, ARM-W-34, 55pp., 1983.

Osborn, H. B., L. J. Lane, and J. F. Hundley, Optimum gaging of thunderstorm rainfall in southeastern Arizona, *Water Resour. Res.*, 8(1), 259-265, 1972.

Osborn, H. B., E. D. Shirley, D. R. Davis, and R. B. Kohler, Model of time and space distributions of rainfall in Arizona and New Mexico, USDA-Agricultural Research Service, ARM-W-14, May, 1980.

Renard, K. G., E. D. Shirley, J. R. Williams, and A. D. Nicks, Hydrology Component: Water routing and Sedimentation, Chapter 3 in: SPUR Simulation Production and Utilization of Rangelands Documentation and User's Guide. U.S. Department of Agriculture ARS-63, 17-30, 1987.

Renard, K. G., L. J. Lane, J. R. Simanton, W. E. Emmerich, J. J. Stone, M. A. Weltz, D. C. Goodrich, and D. S. Yakowitz, Agricultural impacts in an arid environment: walnut gulch studies, *Hydrological Science and Technology*, American Institute of Hydrology, 9(1-4), 145-190, 1993.

Shannon, C. E., A mathematical theory of communication, *Bell System Tech. J.*, 27, 379-423 and 623-656, 1948.

Shannon, C. E. and W. Weaver, The mathematical theory of communication, The University of Illinois Press, Urbana, Illinois, 1949.