EPiC Series in Engineering
Volume 3, 2018, Pages 2293–2300
HIC 2018. 13th International Conference on Hydroinformatics

Spatial-temporal evaluation of rain-gauge network based on entropy theory

Wenqi Wang[1], Dong Wang[1], Vijay P. Singh [2] and Yuankun Wang[1]
[1] Department of Hydrosciences, School of Earth Sciences and Engineering, State Key Laboratory of Pollution Control and Resource Reuse, Nanjing University, Nanjing 210023, P.R. China
[2] Department of Biological and Agricultural Engineering, Zachry Department of Civil Engineering, Texas A & M University, College Station, TX77843, USA
dz1629008@smail.nju.edu.cn, wangdong@nju.edu.cn

Abstract

Ground-based rain-gauge stations are the most direct sources of precipitation data. The evaluation of rain-gauge network is essential and important for water management. One of the most popular methods for design of hydrometric network including rain-gauge network is information theory. Entropy concepts from information theory has been widely adopted and applied in rain-gauge network design. In this paper, spatial-temporal evaluation of rain-gauge network located in Shanghai, China will be performed based on entropy theory. The transinformation-distance (T-D) spatial model is applied under three different sampling frequencies. Weekly precipitation data fits the T-D model best. In addition, the representative network is evaluated to be suitable according to the result.

1 Introduction

Design of a suitable and optimal rain-gauge network is a crucial and challenging task in hydrological community. A reliable rain-gauge network should provide efficient precipitation data for decision-making related to hydrological monitoring and water resources management. In general, as Singh [1] has pointed out, a framework for network design or evaluation should consider several factors, including: 1) objectives of sampling, 2) variables to be sampled, 3) locations of measurement stations, 4) frequency of sampling, 5) duration of sampling, 6) uses and users of data, and 7) socio-economic considerations. Due to the practical and socio-economic complexity, the research community has not yet formed a standardized methodology for a proper network design or evaluation process. Considering the evaluation of a network, we often focus on two modes including space evaluation and time evaluation. The former often refers to number and location of gages and the latter refers to time interval or sampling frequency. Therefore, the establishment of an effective rain-gauge

network generally requires a comprehensive understanding of the information a station can provide both spatially and temporally.

In the past several decades, there has been considerable progress in the application of entropy theory to hydrometric network design ([2,3,4]). Entropy concept has been adopted to explain the inherent information content within a monitoring network system. The common principle is gradually formed to have the maximum amount of information. On the other hand, stations should not have too much sharable information (defined as transinformation), which means stations in the network should be as independent as possible to avoid information redundancy. For rain-gauge stations, Husain [5] proposed that transinformation at any given location can be assumed as a function of the distance to other rain-gauge stations. The function can be represented as a transinformation-distance (T-D) curve model. Also, as gamma distribution is commonly used to describe precipitation series, researchers conclude that exponentially decreasing functions provide a more effective means to estimate the transinformation of rain-gauge stations. The use of exponential T-D model makes it possible to approximate the information transfer in ungauged sites. To some extent, it avoids the step to interpolate observation data in ungauged sites and still realizes the network design task.

Although entropy-based network evaluation methods have been developed and extended greatly during the past several decades, limited applications have been made in China. Our study was developed for the rain-gauge network evaluation in Shanghai, which is a highly developed large city located in Yangtze River delta in China. The main goal of this study is to: (1) (Temporal) evaluate the impact of sampling frequency on rain-gauge network; (2) (Spatial) evaluate spatial coverage of the network with T-D model; (3) conclusive assessment of the representative rain-gauge network.

2 Methodology

Shannon [6] laid the mathematical foundation of entropy as a measure of information or uncertainty. The entropy theory based methods for designing hydrometric network mainly apply several entropy concepts, such as marginal entropy, joint entropy, mutual information, total correlation, etc.

2.1 Entropy concept

Marginal entropy of a variable X is defined as:

$$H(X) = H(p_1, p_2, p_3, \dots, p_N) = -\sum_{i=1}^{N} p_i \log p_i \quad (1)$$

where p_i ($0 \leqslant p_i \leqslant 1$) is the probability of occurrence x_i (i=1,2,…,N)) and $\sum p_i=1$.

For bivariate case (X,Y) with the joint probability of x_i and y_j denoted as $p(x_i, y_i)=p_{ij}$, i=1,2,…,N; j=1,2,…,M, the joint entropy between them can be defined as:

$$H(X,Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p_{ij} \log p_{ij} \quad (2)$$

Mutual information describes the amount of information transmission or shared between two random variables, which is represented as:

$$T(X,Y) = \sum_{i=1}^{N}\sum_{j=1}^{M} p_{ij} \log\left(\frac{p_{ij}}{p_i p_j}\right) \quad (3)$$

where p_i, p_j is the probability of occurrence x_i, y_j (i=1,2,…,N; j=1,2,…,M).

Although transinformation indicates the dependence (sharing information) of two variables, it is not suitable for describing dependence because its upper bound varies from station to station (varying from 0 to marginal entropy, min{H(X), H(Y)}). Therefore, directional information transfer index (DITI, [7]) is defined as:

$$\mathrm{DITI}(X,Y) = \frac{T(X,Y)}{H(X)} \quad (4)$$

$$\mathrm{DITI(Y,X)} = \frac{T(X,Y)}{H(Y)} \quad (5)$$

Similarly, transinformation can be normalized by defining information transfer index (ITI, [8]), which indicates the standardized transinformation from one station to another. We choose to use ITI in our T-D model for computation simplicity since DITI is not symmetric.

$$\mathrm{ITI} = \frac{T(X,Y)}{H(X,Y)} \quad (6)$$

For the estimation of discrete entropy, histogram method is commonly used to discretize variables. The estimation method is especially preferred due to its simplicity for calculation and understanding. However, histogram method is sensitive to bin size for frequency approximation. A suitable bin size has to be determined before estimating probability distribution and entropy measures. In this study, we use bin size suggested by Sturges [9] as:

$$NC = 1 + \log_2(\mathrm{N}) \quad (7)$$

where NC is number of classes and N is the number of observations for the variable. This method has been used in several studies for determining bin size [10,11,12].

2.2 Network evaluation model based on entropy

The transinformation between pair of stations has been shown to be affected by the distance between stations. Previous studies proposed an exponential curve that can fit the T-D relationship, since transinformation decreases with increasing distance and reaches a relatively stable minimum value at a threshold distance (D_{thres}) [8,10,12,13,14]. Here we use ITI as transinformation measurement for the construction of T-D model. After using all pairs of transinformation and distance between stations to fit T-D exponential curve, the optimal distance between rain-gauge stations can be determined to evaluate the network spatially. According to Mogheir et al. [15], if the distances between stations are less than the threshold distance, then there is available transinformation (redundant information) between stations. On the contrary, if the distances between stations are more than the threshold distance, then the transformation between stations is less than the minimum transinformation value (not enough information). So the adequate information available between stations is achieved when the distances between stations equal the threshold distance. This is very useful for evaluate and redesign the network spatially. In addition, data obtained from different sampling intervals will be separately used for constructing T-D model. Then we can compare the fitting performances and suggest a more reasonable sampling strategy temporally.

3 Materials

3.1 Study area and data

Shanghai (30°40′ ~31°53′ N, 120°52′ ~122°12′ E) is located on the west coast of the Pacific Ocean, along the eastern Asian continent, in the front of the Yangtze River delta. By the end of 2003, Shanghai covered an area of 6340.5 square kilometers, accounting for 0.06% of the total area of the country. It belongs to Huangpu River system, which is the most representative plain river system in Taihu Lake basin. The area has dense river network with average elevation of 2.5 to 5.0 meters, just like the bottom of the basin. Shanghai has north subtropical maritime monsoon climate with abundant sunshine and precipitation and its wet season extends from May to September.

In total, 47 rain-gauge stations in Shanghai (except Chongming district) were chosen for case study. We include two stations located in Kunshan, Jiangsu Province in the network system since they are very close to other stations. The spatial distribution of all rain-gauge stations is shown in Figure 1. For integrity, daily precipitation data of ten years (from 2006 to 2015, total length of the series is 3652)

is obtained in this study. Three sampling frequencies (daily, weekly and monthly) were obtained from the original data for temporal analysis. The sample size for daily, weekly and monthly rainfall is 3652, 521 and 120, respectively.

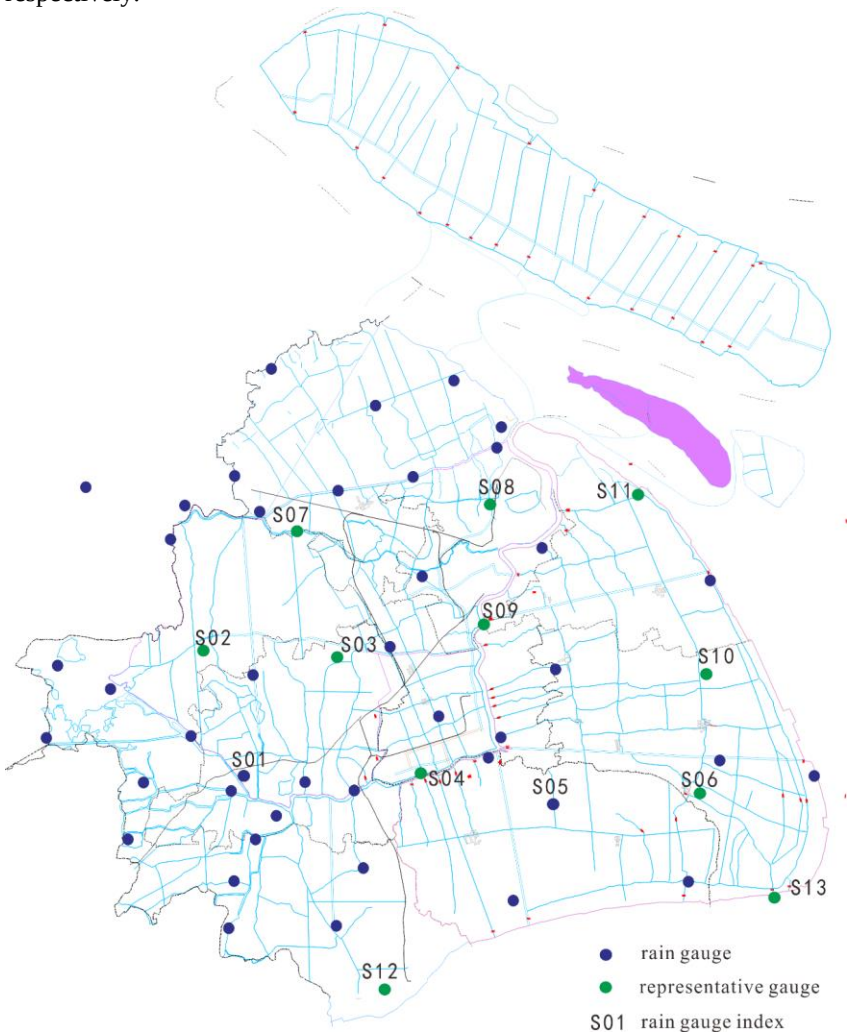

Figure 1: Study area and rain-gauge network

3.2 Representative rain-gauge network

To reduce computational burden, representative rain-gauge network is chosen for evaluation in our study, which is suggested by administration of water management. The rain-gauge network is composed of 13 representative rain-gauge stations (green spots shown in Figure 1). Generally, these representative stations spread out across the whole city. However, the whole network including 47 stations is still used for constructing T-D model as more data points will fit the curve better.

4 Results and Discussions

4.1 Temporal scale effect

To assess the impact of different temporal scales, we first provide marginal entropy under daily, weekly and monthly precipitation samples (shown in Figure 2). Marginal entropy would generally be higher with increasing sampling interval. So monthly data yielded the highest entropy values among three cases. This can be explained by the amplification effect as a larger sampling interval naturally bring higher variability in precipitation distribution. However, monthly data also brought more divergences from daily and weekly data in relative entropy values for different stations. For example, station S01 had the lowest marginal entropy among thirteen representative stations both for daily and weekly measurements. However, for monthly data station S06 was the least informative of the representative network. Similar results can be found in station S03, S09, S11, S13, etc. Generally, monthly precipitation can exhibit much more distinctive distribution patterns from relatively smaller temporal scales as daily or weekly intervals. Therefore, the network evaluation process might show different results under different temporal scales.

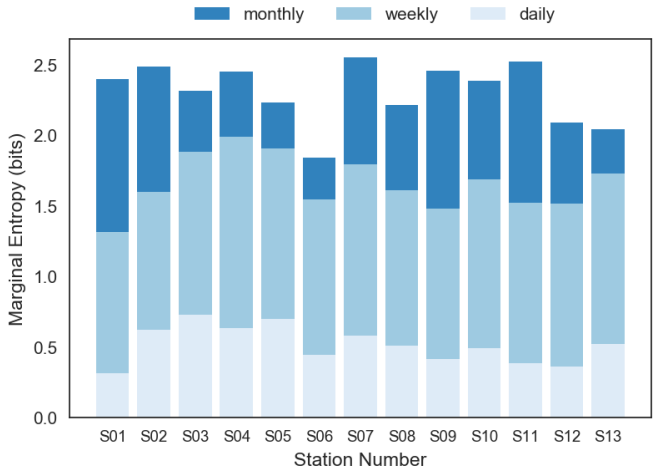

Figure 2: Marginal entropy of precipitation data for representative rain-gauge stations

4.2 Transinformation-distance model

In this study, the available data of all stations (47 stations) in the existing network are used for developing the T-D curves (Figure 3). In general, three T-D curves were similar with different sampling frequencies (daily, weekly and monthly) from the exponential fitting results. Furthermore, we find that the model of exponential fitting showed best result using weekly data (see in Table 1) with minimum SSE (2.04), RMSE (0.93) and maximum R-square (0.04). This may suggest that when using weekly precipitation data, the spatial relationship of dependence between rain-gauge stations can be best captured using exponential model and the distribution pattern could be clearest under weekly sampling interval.

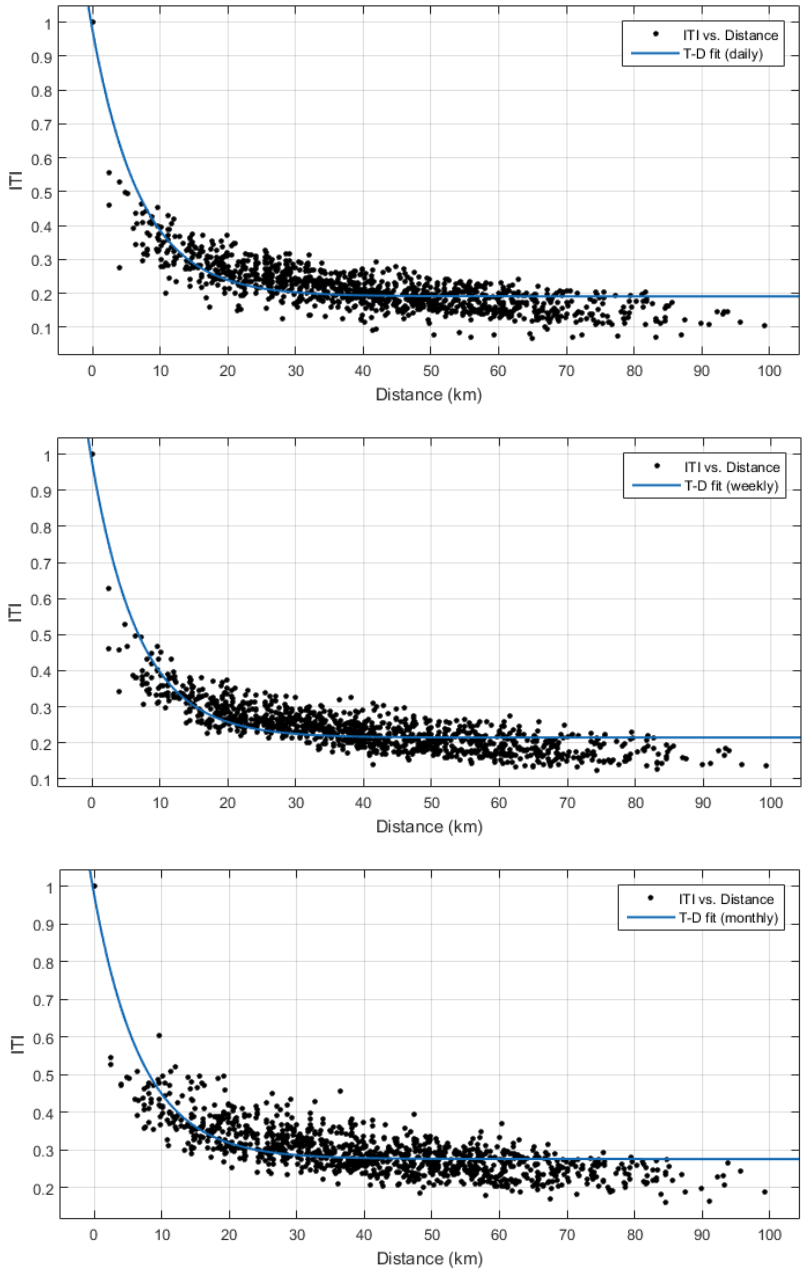

Figure 3: Transinformation-distance curve for different sampling intervals (from upper to lower using daily, weekly and monthly data)

Exponential model	Fitting equation	Goodness of fit		
		SSE	R-square	RMSE
daily data	ITI=0.79*exp(-0.14*D)+0.19	2.62	0.92	0.05
weekly data	ITI=0.76*exp(-0.14*D)+0.21	2.04	0.93	0.04
monthly data	ITI=0.70*exp(-0.14*D)+0.28	2.67	0.90	0.05

Table 1: The equations of T-D curves and goodness of fit for three sampling frequencies

In addition, we can find from three curves that the threshold distance for information transfer between stations is always around 30~40 kilometers in three sampling frequencies. This signified that the efficient information transmission distance should be no more than 40 kilometers (D_{thres}≈40km), which may serve as an evaluation criterion for representative network density. Meanwhile, if the distance between stations is less than 30 kilometers, the network may generate more information redundancy, which is also against the common principle of rain-gauge network design based on information theory.

4.3 Evaluation of representative network

With respect to the space-time evaluation of representative network, number and locations of the rain-gauge stations and sampling frequency are often considered. Generally, the average control area of a representative station is around 487.73 km^2 (calculated by dividing total area with the number of representative stations). If we assume that information transmission is isotropic in the study area, the control area ($S_{control}$) of a representative station can be approximately denoted as:

$$S_{control} = \pi D^2 \quad (8)$$

So the control distance (D) is estimated as 12.46 km, which is much lower than the threshold distance (D_{thres}), even lower than 30 km. From the discussion, we can conclude that the density of the representative network is rather enough from the perspective of transinformation. Moreover, the distance between stations might be overly close and the current representative rain-gauge network could produce more information that is redundant.

5 Conclusions

This study evaluated the rain-gauge network in Shanghai both spatially and temporally based on entropy theory. ITI was used to determine the spatial variability between rain-gauge stations, which defined the information affinity between stations. A transinformation-distance model was further adopted for constructing the spatial information transfer ability of the rain-gauge network. For temporal consideration, three sampling frequencies including daily, weekly and monthly precipitation data were compared. It was found that entropy values increased with larger sampling intervals and weekly data performed the best T-D fitting results. We used the threshold distance suggested by T-D model to evaluate the representative rain-gauge network. It is concluded that the representative network could be further redesigned and improved to avoid information redundancy based on transinformation analysis.

References

[1] V. P. Singh, Entropy theory and its application in environmental and water engineering. John Wiley & Sons, 2013.
[2] A.K. Mishra, P. Coulibaly, Developments in hydrometric network design: A review. Rev. Geophys. 47(2), 2009.
[3] J. C. Chacon-Hurtado, L. Alfonso, D. P. Solomatine, Rainfall and streamflow sensor network design: a review of applications, classification, and a proposed framework. Hydrol. Earth Syst. Sci., 21(6), 3071, 2017.
[4] J. Keum, K. Kornelsen, J. Leach, P. Coulibaly, Entropy applications to water monitoring network design: a review. Entropy, 19(11), 613. 2017.
[5] T. Husain, Hydrologic uncertainty measure and network design. Water Resour. Bull. 25 (3), 527–534, 1989.
[6] Shannon, C. E., A mathematical theory of communication, Bell Syst. Tech. J., 1948.
[7] Y. Yang, D.H. Burn, An entropy approach to data collection network design. J. Hydrol, 1994.
[8] Y. Mogheir, J. L. M. P. de Lima, V. P. Singh, Characterizing the spatial variability of groundwater quality using the entropy theory: I. Synthetic data. Hydrol Process. 18(11): 2165-2179, 2004.
[9] H. A. Sturges, The choice of a class interval, J. Am. Stat. Assoc., 21(153), 65-66, 1926.
[10] F. Masoumi, R. Kerachian, Optimal redesign of groundwater quality monitoring networks: A case study. Environ. Monit. Assess. 161(1-4), 247-257, 2010.
[11] E. Ridolfi, V. Montesarchio, F. Russo, F. Napolitano, An entropy approach for evaluating the maximum information content achievable by an urban rainfall network. Nat. Hazards Earth Syst. Sci. 11(7): 2075-2083, 2011.
[12] E. Ridolfi, M. Rianna, G. Trani, L. Alfonso, G. D. Baldassarre, F. Napolitano, F. Russo, A new methodology to define homogeneous regions through an entropy based clustering method. Adv. Water Resour. 96, 237-250, 2016.
[13] N. Mahjouri, R. Kerachian, Revising river water quality monitoring networks using discrete entropy theory: the Jajrood River experience. Environ. Monit. Assess. 175, 291–302, 2011.
[14] R. R. Owlia, A. Abrishamchi, M. Tajrishy, Spatial-temporal assessment and redesign of groundwater quality monitoring network: a case study. Environ. Monit. Assess. 172(1-4), 263, 2011.
[15] Y. Mogheir, J. L. M. P. de Lima, V. P. Singh, Characterizing the spatial variability of groundwater quality using the entropy theory: II. Case study from Gaza Strip. Hydrol. Process. 18(13), 2579–2590, 2004.