An entropy approach to data collection network design

Yujuin Yang*, Donald H. Burn

Department of Civil Engineering, University of Manitoba, Winnipeg, Man., Canada R3T 2N2

(Received 1 October 1992; revision accepted 8 October 1993)

Abstract

A new methodology is developed for data collection network design. The approach employs a measure of the information flow between gauging stations in the network which is referred to as the directional information transfer. The information flow measure is based on the entropy of gauging stations and pairs of gauging stations. Non-parametric estimation is used to approximate the multivariate probability density functions required in the entropy calculations. The potential application of the approach is illustrated using extreme flow data from a collection of gauging stations located in southern Manitoba, Canada.

1. Introduction

Many applications within the field of water resources rely on the availability of hydrometric data at a variety of locations. Hydrometric data are required for the efficient planning, design, and operation of virtually all water resource systems, including hydroelectric plants, water supply reservoirs, recreation and fishery facilities, and flood control structures, to name but a few. The diversity of uses and users of hydrometric data, both present and future, make the design of a data collection network an important consideration. The importance of this task is likely to increase as the available resources for data collection activities decrease in response to budgetary cutbacks. It is thus essential that a systematic approach to data collection network design be utilized to ensure that network expansion or rationalization considers the varied uses and users of the hydrometric data.

A methodology for data collection network design must consider the information content of each gauging station, or potential gauging station, in the network and must also consider the users and uses of the data collected at each of the stations. A station

* Corresponding author.
with a high information content would generally be given priority over other stations with lower measures of information content. The information content of a station must, however, be balanced with the site-specific uses and users of the data collected at a station. Thus, a station which is used by only one user might be given a lower priority than a station that has many diverse users. A data collection network design framework that considers the above-noted issues was developed by Burn and Goulter (1991). Burn and Goulter also provide a summary of some of the techniques that have been used in the past to design data collection networks.

In recent years, the entropy concept, which is borrowed from communication theory, has been adopted into hydrometric data collection network studies. In communication theory, entropy measures the uncertainty of a random event, or in other words, the information contained in the event, through the observations (signals) of it. A series of observations of an uncertain event contain more information about the event itself than that of a less uncertain event do. The information transferred among information emitters (predictor stations) and the information receivers (predicted stations), can be measured by the term 'mutual information'. Hydrometric gauging stations provide observations about random hydrometric events. This information is contained in the observations. The information observed at different sites can be inferred, to some extent, from observations at other sites. It is these features that provide the basis of the adoption of entropy into hydrologic studies. Entropy is now used in measuring the information content of observations at a gauging station, and the mutual information is used in measuring the inferred information or equivalently the 'information transmission'. As such, entropy and mutual information possess advantages relative to other measures of association in that they provide a quantitative measure of: (1) the information at a station; (2) the information transferred and lost during the transmission; (3) a description of the relationships among stations according to their information transmission characteristics.

Caselton and Husain (1980) introduced the entropy concept into a hydrometric network study. They computed the information transmissions, based on the entropy concept, and selected stations with the maximum information transmission. Chapman (1986) studied the application of entropy in various cases involving the use of different assumed distribution functions, different types of flow data, and also considered different units of entropy. Harmancioglu and Yevjevich (1987) used entropy to measure the information transmission among stations on the same river.

A difficulty preventing the more widespread use of entropy is the prerequisite of representing the spatial structure of a hydrometric event using a multivariate probability distribution. In most of the above-mentioned works and others (e.g. Husain (1989)), a continuous distribution function has been assumed and the entropy is calculated based on this. This approach has several disadvantages. First, because the options for distribution functions are limited (by knowledge or computational restrictions), an inappropriate distribution function may be selected, perhaps as a result of the limited sample size available to characterize the multivariate distribution. Secondly, this approach involves the comparison of different probability distribution functions, which represents another subjective aspect of the problem.
Finally, this approach requires a high level of skill and considerable experience with multivariate distribution functions.

A possible solution to these difficulties is the application of the non-parametric estimation of the density distributions. Rosenblatt (1956) introduced the concept of non-parametric estimation (NPE) for the estimation of probability density functions. NPE entails estimating the values of an unknown density function instead of the form of the function itself and requires no prior knowledge of that functional form. Parzen (1962) subsequently developed a kernel estimator for the one-dimensional density estimation problem. Cacoullos (1966) extended this method into the multivariate regime. Roussas (1969) dealt with time series random variables in non-parametric estimation.

Non-parametric estimation has been adopted and applied by several hydrologists. Adamowski (1985) employed NPE in flood frequency analysis. Using NPE, he estimated the frequency distribution of flood events. Schuster and Yakowitz (1985) also used a method called parametric/non-parametric estimation (P-NPE) to estimate flood frequencies. This method was thought to be good for overcoming the problem with NPE of obtaining poor estimates in the tails of the distribution. The resulting distribution functions were then used in station regionalization. If two stations had similar density functions they were grouped together. The resulting distribution functions were also compared with those derived from parametric analysis. It was concluded that P-NPE was much better if the underlying distribution was incorrectly assumed in a parametric analysis and was still fairly good in comparison with parametric analysis if the underlying distribution was correctly assumed. Adamowski (1989) also compared parametric estimation with NPE and his conclusion was that NPE was slightly better in terms of the integral mean square error (IMSE). Adamowski and Feluch (1991) applied NPE to estimate the regression relationship between groundwater levels and streamflow. In this case, a two-dimensional density function was required.

The intention of this paper is to apply the entropy concept to the analysis of data collection networks. The procedure developed, which can be incorporated into a data collection network design framework, includes two phases. The first phase entails the regionalization of the network using the entropy concept. It involves the estimation of two-dimensional density functions for streamflow gauging station pairs from a collection of gauging stations by means of the non-parametric estimation. A directional information transfer index (DIT) is then defined and calculated for all pairs. The regionalization is subsequently performed based on the DIT values. The second phase is to select representative stations for homogeneous subregions resulting from the first phase.

The remainder of this paper is organized in the following manner. The next section provides a brief overview of data collection network design. A review of some of the relevant aspects of entropy theory is presented next. This is followed by a presentation of the non-parametric estimation technique. The application of these two techniques to streamflow data for a case study located in southern Manitoba follows. Observations regarding the application of the procedures presented herein to data col-
lection network design are then presented. The final section contains conclusions and recommendations for further work.

2. Data collection network design framework

Burn and Goulter (1991) presented a two-phase approach to data collection network design. The approach was primarily intended for the rationalization of an existing data collection network, but could, with modifications, be used for more general network design problems. The first stage of the approach uses a hierarchical clustering technique to form groups of hydrometric gauging stations with similar flow characteristics. A measure of association, based on the correlation of different components of the flow regime, is used to form a similarity matrix defining the similarity of each station to every other station. This similarity matrix is then used as the basis for forming groups of stations (regionalization) where the group formed at any stage results from the amalgamation of the two most similar groups. After each amalgamation step, the similarity matrix is updated and the procedure is repeated until a stopping criterion is satisfied. The output from this first stage of the network design procedure is a number of groups of hydrometric gauging stations with each group containing a collection of stations with similar data characteristics.

The second stage of the procedure involves the selection of a single station, or multiple stations, from each group for retention in the final data collection network. The intent is to identify unique sources of information that are not contained in any of the other stations in the network. The station selection stage is an inherently heuristic process that must consider the information content of each station as well as the users and uses of the data collected at each station. There is thus a need to strike a balance between a mathematical measure of information content and the practical user-oriented concerns that are relevant for the specific data collection network.

Within the data collection network design framework outlined above, there are numerous opportunities to incorporate alternative approaches to the components of the basic framework. For example, alternative measures of association between hydrometric stations could result in improved grouping of the stations if the measure of association used more closely reflects the information transfer capabilities of a station. As well, the station selection process, which requires, as one component, a measure of the relative information content of the stations in a group, could benefit from an enhanced measure of this characteristic of the stations.

3. Entropy concepts

The Shannon entropy for a continuous random variable, \( X \), is defined as (Lathi, 1968)

\[
H(X) = - \int_{-\infty}^{\infty} f(x) \log[f(x)] \, dx
\]  

(1)
where \( x \) is an observation of \( X \), \( f(x) \) is the probability density function of \( X \), and \( H(X) \) is the entropy measure that describes the information contained in \( X \). If two random variables \((X, Y)\) are considered, then the transmitted information from \( X \) to \( Y \) is represented by the mutual information given as (Lathi, 1968)

\[
T(X; Y) = H(X) - H(X|Y)
\]

and

\[
H(X|Y) = \int_{-\infty}^{\infty} f(y)H(X|y)dy = - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y)f(x|y) \log[f(x|y)]dxdy
\]

where \( H(X|Y) \) denoted as \( H_{\text{lost}} \) describes the loss of information, \( f(x|y) \) is the conditional probability density of \( x \) given \( y \), and \( f(x) \) and \( f(y) \) are the marginal densities of \( x \) and \( y \), respectively. It then follows that

\[
T(X; Y) = - \int_{-\infty}^{\infty} f(x) \log[f(x)]dx + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y)f(x|y) \log[f(x|y)]dxdy
\]

Following some manipulations, \( T \) can be expressed as

\[
T(X; Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy
\]

In the context of hydrology, the random variable \( X \) can be a statistic such as the annual extreme value of streamflow, \( x \) is then the observation of \( X \), and \( X \) and \( Y \) are statistics at two different sites. The entropy, \( H(X) \), can thus be interpreted as the information about the statistic at site \( X \) and \( H(Y) \) the information at site \( Y \). The mutual information is the information content inferred by \( X \) about \( Y \), or vice versa.

The entropy defined by Eq. (1) measures the relative information, with \(- \log(\Delta x)\) serving as the datum when \( \Delta x \) approaches zero (Lathi, 1968), and \( \Delta x \) is the division interval of the \( X \) domain. As a measure of the relative information, \( H(X) \) can be positive, zero, or negative and therefore the information lost, \( H_{\text{lost}} \), can also be positive, zero, or negative since it is a part of the information \( H(X) \) and bounded from above by \( H(X) \). Negative information or negative information lost have no physical meanings although both cases are mathematically possible. This difficulty arises from the use of a relative coordinate for which the origin is set at \(- \log(\Delta x)\). If we consider \( H \) and \( H_{\text{lost}} \) in an absolute coordinate in which the origin is at minus infinity, then both \( H \) and \( H_{\text{lost}} \) are no longer negative and regain their normal physical meanings. At this time, it is assumed that the problem is considered in such an absolute coordinate to establish the legitimacy of DIT, which will be introduced below. The measures necessary to satisfy this assumption in practical applications will be discussed in the section on the application of the methodology.

The mutual information, \( T \), is symmetric (i.e. \( T(X; Y) = T(Y; X) \)), and non-negative (readers are referred to Lathi (1968) or Reza (1973) for proofs). A zero value occurs when two stations are statistically independent of each other so that no information is mutually transferred. When two stations are functionally dependent, the information at one site can be fully transmitted to another site with no loss at all.
Subsequently, $T$ is equal to $H$. Any other situation between totally independent and fully dependent leads to a value of $T$ between zero and $H$. Larger $T$ values correspond to greater amounts of information transferred. In this regard, $T$ is an indicator of the capability of the information transmission and the degree of dependency of two stations.

Although $T$ indicates the dependency of two stations, it is not a good index of the dependency since its upper bound varies from site to site. To normalize it, the original definition of mutual information is altered to a directional information transfer index (DIT)

$$DIT = \frac{T}{H} = \frac{H - H_{lost}}{H} = 1 - \frac{H_{lost}}{H}$$

The physical meaning of DIT is the fraction of information transferred from one site to another. It is easy to see that DIT ranges from zero to unity when $T$ varies from zero to $H$. The zero of DIT corresponds to an independent situation where no information is transmitted. A value of unity for DIT corresponds to a fully dependent situation where no information is lost. Any other value of DIT between zero and 1 denotes a situation between independent and fully dependent.

It is noticed that DIT is no longer symmetrical, since $DIT_{xy} = T/H(X)$ for station $X$ will not in general be equal to $DIT_{yx} = T/H(Y)$ for station $Y$. $DIT_{xy}$ now describes the fractional information inferred by station $X$ about $Y$, while $DIT_{yx}$ is the fractional information inferred by station $Y$ about $X$. Between two stations of one pair, the station with higher DIT value should be given higher priority to be kept because of its higher capability of inferring (predicting) the information at the other site.

A direct application of DIT is the regionalization of the network. If both $DIT_{xy}$ and $DIT_{yx}$ are high, the two related stations should be arranged in the same group since the hydrometric event patterns represented by them are strongly dependent and consequently information can be mutually inferred between them. If neither $DIT$ is high, then by the same reasoning they should remain in separate groups. If only one DIT (say $DIT_{xy}$) is high, then the station $Y$, whose information can be predicted by $X$, can join station $X$ if station $Y$ does not belong to any other group; otherwise it stays in its own group. The predictor station $X$ cannot enter station $Y$'s group in any circumstance since if station $X$ were to be discontinued the information at that site would be lost.

It has been shown that DIT can be both a measure of the information transmission capability and an indicator of the dependency of a station pair. The first feature is distinct from other measures of the relationships among stations that have previously been used, and the second feature renders this measure suitable for comparison with procedures that have used other measures as a basis for indicating association.

Through the use of directional information transfer, a new measure of the relationship between a pair of stations is created. This measure is based on the stations' essential connection, the information connection, and is thus distinguished from traditional similarity measures, such as the correlation coefficient. In a traditional method, the connections between the stations are quantified by similarities which may
be based on one of numerous measures of association. With different similarity metrics, the final results of the data collection network study (such as regionalization, station selection, etc.) may be different as a result of the subjective selection of the similarity metric. The use of a measure of the DIT can alleviate some of these difficulties in that only a single measure of association, which has a theoretical justification for use, is needed as a similarity metric. In the station selection process, a 'predicted station' should be removed first because it only receives information efficiently and does not predict information efficiently. When all remaining stations in the group have strong mutual connections with each other (i.e. both \( \text{DIT}_{xy} \) and \( \text{DIT}_{yx} \) are high), they can be further selected according to a criterion of S-DIT, which is defined as

\[
S - \text{DIT}_i = \sum_{j=1, j\neq i}^{m} \text{DIT}_{ij}
\]

where \( \text{DIT}_{ij} \) is the information inferred by station \( i \) about station \( j \), and \( m \) is the number of stations in the group. The station in each group with the highest value for S-DIT, in comparison with other members of the group, should be retained in the network. The regionalization and station selection will be demonstrated later in a case study.

A byproduct of the DIT procedure is referred to herein as the frequency response matrix

\[
f(x|y) = \frac{f(y|x)f(x)}{f(y)} = \frac{\int_{-\infty}^{\infty} f(y|x)f(x)dx}{\int_{-\infty}^{\infty} f(y|x)f(x)dx}
\]

The frequency response matrix helps in understanding the characteristics of DIT, as will be illustrated later. The frequency response matrix is actually a conditional probability distribution.

4. Non-parametric density estimation

Non-parametric estimation does not attempt to describe a density function by formula and parameters but rather by its point values everywhere in the domain. If the density values at every point are known, then the function is numerically known. Based on the work of Parzen (1962), the theory can be described as follows. Let \( x_1, x_2, ..., x_n \) be a sample of the random variable \( X \) and be independently and identically distributed with probability density function \( f(x) \). If \( f'(x) \) is the derivative of the distribution function almost everywhere and a continuous function, then at a given point \( x \), it can be estimated by a kernel estimator

\[
f_n(x) = \int_{-\infty}^{\infty} \frac{1}{h} K \left( \frac{x - x_j}{h} \right) dF_n \left( \frac{x - x_j}{n} \right) \approx \frac{1}{nh} \sum_{j=1}^{n} K \left( \frac{x - x_j}{h} \right)
\]

where \( f_n(x) \) is called the kernel estimator of \( f(x) \); the function \( K(.) \) is called a kernel;
the positive number \( h \) is a smoothing factor of the kernel, which is a function of the sample size \( n \) and the sample values; and \( F_n(.) \) is an estimate of the cumulative probability distribution function of \( X \). Parzen (1962) indicated that \( f_n(x) \) is an asymptotically unbiased estimator of \( f(x) \), under certain conditions. Cacoullos (1966) extended the above principles into multidimensional situations where \( X = X(X_1, X_2, ..., X_p) \). At a given point of \( X^0 = X^0(x_1, x_2, ..., x_p) \), the kernel estimator has the form of

\[
f_n(X^0) = \frac{1}{n h_1...h_p} \sum_{j=1}^{n} K\left(\frac{x_1 - x_{j1}}{h_1}, ..., \frac{x_p - x_{jp}}{h_p}\right)
\]

where \( X^j = X^j(x_{j1}, x_{j2}, ..., x_{jp}) \) is the \( j \)th observation of \( X \) which is a \( p \)-dimensional variable. The components of \( X, (X_1, X_2, ..., X_p) \), could be mutually dependent or independent, but the observations of each component are still assumed to be independently and identically distributed. To simplify the structure of the kernel estimator, the kernel in Eq. (10) can be assumed to be a product function of the form

\[
K\left(\frac{x_1 - x_{j1}}{h_1}, ..., \frac{x_p - x_{jp}}{h_p}\right) = \prod_{i=1}^{p} K_i\left(\frac{x_i - x_{ij}}{h_i}\right)
\]

Practically, the \( h_i \) will be assumed to be equal to a unique \( h \), such that

\[
h_1 = h_2 = ... = h_i = ... = h_p = h
\]

It should be noted that if observations have a time series feature (i.e. correlated observations), the method of Roussas (1969) should be invoked.

Adamowski (1985) indicated that the choice of a kernel is not crucial for the performance of the model. Common forms of the kernel are presented by Parzen (1962) and Wertz (1978). Unlike the selection of the kernel, the selection of \( h \) is a very sensitive factor. It affects both the bias and the mean squared error (MSE) of the estimator and because of that, it has been widely discussed (Parzen, 1962; Wertz, 1978; Adamowski, 1985; Schuster and Yakowitz, 1985). All of the theoretical analyses of the optimal value of \( h \) express \( h \) as a function of \( f(x) \), which is unknown. Therefore, it is not practical to obtain an optimal value for \( h \) from these analyses. In practice, a widely used procedure is called 'cross-validation maximum likelihood' (Duin, 1976; Schuster and Yakowitz, 1985; Adamowski and Feluch, 1991). This procedure aims at the minimization of the integral mean square error of \( f_n(x) \). Duin (1976) indicated that this procedure for determining an appropriate \( h \) value was superior to other alternatives.

In the cross-validation procedure, \( h \) is selected to meet

\[
\max L(h) = \prod_{j=1}^{n} f_n(X_j; h) = \prod_{j=1}^{n} \frac{1}{h_1...h_p} \sum_{i=1, i\neq j}^{n} K(X_i; h)
\]

where \( f_n(X_j; h) \) is the estimated density value at \( X_j \), but with \( X_i \) removed. The principle of this procedure is simple: if \( h \) is appropriate, then \( f_n(X_j; h) \) must be a proper
Table 1
Summary of gauging stations

<table>
<thead>
<tr>
<th>Station identification number</th>
<th>Water Survey number of Canada</th>
<th>Name of station</th>
<th>Period of record</th>
<th>Length of record (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>05OC004</td>
<td>Pembina R. near Neche</td>
<td>1903–1988</td>
<td>81</td>
</tr>
<tr>
<td>2</td>
<td>05OB007</td>
<td>Pembina R. near Windygates</td>
<td>1962–1988</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>05OB023</td>
<td>Pembina R. below Crystal Cr.</td>
<td>1962–1988</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>05OA010</td>
<td>Pembina R. above Lorne Lake</td>
<td>1962–1988</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>05OA008</td>
<td>Pembina R. near Killarney</td>
<td>1959–1988</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>05OA007</td>
<td>Badger Cr. near Cartwright</td>
<td>1959–1988</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>05OA006</td>
<td>Long River near Holmfield</td>
<td>1959–1988</td>
<td>30</td>
</tr>
</tbody>
</table>

estimator and consequently the probability of the occurrence of the $X_j$ from $f_n(X_j; h)$ must be the highest.

5. Application

Seven gauging stations from the Pembina River basin in southern Manitoba, Canada were chosen to demonstrate the two-phase procedure developed above. Table 1 provides summary information regarding the gauging stations used in this application. Burn and Goulter (1991) also studied the hydrometric gauging stations in this basin. They provided a brief description of the basin and the available data network to which interested readers are referred. In the present work, the focus is on the development of the procedure and therefore only a relatively small subset of the entire data collection network is used to illustrate the methodology. Theoretically and realistically, there is no limit on the size of the network for which the procedure can be applied.

The extreme flow data (annual maximum flood series) from 1962 to 1988 were used in the computations that follow. Using a common period for all stations herein was to simplify the calculations. Actually, the common period is only required for every station pair, as with other methods dealing with measures of association between station pairs. It is the choice of the analyst to use a common period for all stations in the network (as is done in our case study) so that the accuracy of the analysis for each station pair is the same, but with some resources unused; or to use different periods for each station pair so that the maximum amount of resources are used but with varying accuracies for different station pairs. The joint density functions for all pairs are obtained by non-parametric estimation.

As was noted above, the selection of the form of the kernel function does not generally affect the estimation results significantly. There is, however, a concern in practice since the random variable of interest in our work, streamflow, cannot be negative. As such, the density function value for negative values of the random variable should be zero. For this reason, the original observations are logarithmically...
transformed and then a Gaussian kernel is chosen to form the estimator

$$K(x, y) = K_1(x)K_2(y) = \frac{1}{2\pi} \left[ \exp \left( -\frac{x^2}{2} \right) \right] \left[ \exp \left( -\frac{y^2}{2} \right) \right] = \frac{1}{2\pi} \left\{ \exp \left[ -\frac{(x^2 + y^2)}{2} \right] \right\}$$

(14)

where \( x = (x' - x_j)/h; y = (y' - y_j)/h; (x_j, y_j) \) is an observation point, and \((x', y')\) is the coordinate of any point in the field. All \( x' \) and \( y' \) (or \( x_j \) and \( y_j \)) are logarithmically transformed. The optimal \( h \)-values are derived from Eq. (13). After this the \( f_n(x, y) \) becomes

$$f_n(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} 1 \left\{ \exp \left[ -\frac{(x^2 + y^2)}{2} \right] \right\}$$

(15)

It was indicated earlier that entropy must be considered in an absolute coordinate to ensure physically meaningful values. Actually, if both \( H \) and \( H_{\text{lost}} \) are positive, all of the previous statements about DIT will be true, even if a relative coordinate system is used. There are, however, ways to enforce the satisfaction of this condition of DIT in the more general case. One approach is to eliminate the stations with comparatively very low (i.e. negative) entropies and (if necessary) to lower the datum to ensure that the remaining stations have both positive \( H \) and \( H_{\text{lost}} \). In hydrometric studies, if a station is found to have a very low entropy value (information content), then there is no point in keeping such a station since it does not provide much information. After
the exclusion of such stations, all others are compared based on a common (possibly nonzero) datum. A change of the datum does not change the mutual relations among stations.

Although the procedure is subject to certain conditions (the application condition of DIT, and the requirement of the sample being identically and independently distributed), these requirements are not difficult to satisfy in most applications.

6. Observations and discussion

Through Eq. (13), the preferred value for \( h \) is obtained for each station pair. The corresponding probability density functions for all station pairs are subsequently computed. From this process, two basic types of density functions are revealed. The first type is an essentially unimodal density function as illustrated in Fig. 1 using stations 1 and 5. In this figure, the two-dimensional density distribution is like a skewed bell. If the data were to be fitted with a parametric density function, the attempt to approximate the density function would likely be fairly successful. The second type is a multimodal density function as illustrated in Fig. 2 with stations 2 and 6. There are a few peaks in the field and the range of those peaks is not regular. For this type of density function, finding a proper parametric density function may be more difficult.
Table 2
Entropy of all stations

<table>
<thead>
<tr>
<th>Station</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>1.56</td>
<td>1.64</td>
<td>1.72</td>
<td>1.5</td>
<td>1.6</td>
<td>1.77</td>
<td>1.49</td>
</tr>
</tbody>
</table>

The integrations of Eqs. (1)–(5) are approximated by summations. With the density functions derived previously, the entropies for all stations and the DIT matrix are obtained and presented in Tables 2 and 3, respectively. The entropy table gives a measure of the total information at a station. The DIT matrix displays the fraction of information transmitted among the stations. The larger the DIT value is, the larger the fraction of the information that is transmitted (i.e. the better the information can be inferred for one station based on the information at the other station).

The entropies in Table 2 are all positive and so are the $H_{lost}$ values (since all DIT values are less than or equal to 1). The conditions for Eq. (6) to apply are therefore met. Although not a proof that this will occur in general, the positive entropy results support our feeling that a meaningful gauging station should have a positive entropy value.

Let us compare two extreme pairs of gauging stations. Pair 1 is composed of stations 1 and 2 corresponding to the largest DIT values (0.5 and 0.48), and pair 2 is composed of stations 1 and 5, with the smallest DIT values (0.15 and 0.14). Their frequency response matrices are displayed in Figs. 3 and 4, respectively. According to the DIT values, pair 1 is obviously more dependent than pair 2. The magnitude of DIT for the first pair is 3.5 times larger than the DIT for the second pair. This is consistent with the previous analysis that a more dependent pair should have larger DIT values. The great discrepancy in magnitudes of the DIT values also indicates that DIT is sensitive to the dependency of related stations. This feature is noteworthy in that it explores small variations in dependency so that the classification of stations using this measure may be more accurate.

A comparison of Figs. 3 and 4 enhances the understanding of the principle of the directional information transfer. DIT is totally built on the information transmission

Table 3
Directional information transfer (DIT)

<table>
<thead>
<tr>
<th>Station</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.18</td>
<td>0.15</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>0.48</td>
<td>1</td>
<td>0.41</td>
<td>0.17</td>
<td>0.15</td>
<td>0.32</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0.27</td>
<td>0.4</td>
<td>1</td>
<td>0.18</td>
<td>0.15</td>
<td>0.39</td>
<td>0.19</td>
</tr>
<tr>
<td>4</td>
<td>0.18</td>
<td>0.18</td>
<td>0.21</td>
<td>1</td>
<td>0.42</td>
<td>0.25</td>
<td>0.48</td>
</tr>
<tr>
<td>5</td>
<td>0.14</td>
<td>0.15</td>
<td>0.17</td>
<td>0.4</td>
<td>1</td>
<td>0.21</td>
<td>0.29</td>
</tr>
<tr>
<td>6</td>
<td>0.21</td>
<td>0.3</td>
<td>0.4</td>
<td>0.22</td>
<td>0.19</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>7</td>
<td>0.23</td>
<td>0.27</td>
<td>0.22</td>
<td>0.5</td>
<td>0.31</td>
<td>0.23</td>
<td>1</td>
</tr>
</tbody>
</table>
Fig. 3. Frequency response matrix of station 1 given 2.

Fig. 4. Frequency response matrix of station 1 given 5.
characteristics of a station pair. The transmission is based on the relation of the hydrometric events observed at the corresponding gauging stations. If these characteristics are favorable (i.e. response pulses are sharp, high and narrow, as can be seen in Fig. 3, so that the response at one station is clear for a given observation at the other), the relation is strong. Consequently, the DIT value is large, implying that the stations have a high degree of directional dependence. In contrast, poorer characteristics (i.e. the response pulses are opposite to those described above so that the response to a given observation is ambiguous) indicate a weaker relation. A station pair with such response characteristics is less dependent (Fig. 4). A large fraction of the information will be lost when information is transmitted between them. Other similarity measures do not necessarily reflect these response characteristics.

Figure 5. Frequency response matrix of station 3 given 7.

Table 4
Correlation coefficient ($R$)

<table>
<thead>
<tr>
<th>Station</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.908</td>
<td>0.801</td>
<td>0.74</td>
<td>0.722</td>
<td>0.818</td>
<td>0.772</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.959</td>
<td>0.782</td>
<td>0.681</td>
<td>0.898</td>
<td>0.878</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.852</td>
<td>0.752</td>
<td>0.956</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.908</td>
<td>0.851</td>
<td>0.942</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.834</td>
<td>0.807</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For a comparison with more conventional methods, a regression analysis was conducted for each station pair. Table 4 presents the correlation coefficient matrix from such an analysis. A traditional method of determining station dependence may be based on this matrix. The station pairs with a correlation coefficient, $R$, greater than a threshold value (say, 0.9) can be deemed as dependent, while others can be considered to have lower degrees of dependence. It is interesting to note discrepancies between the DIT matrix in Table 3 and the $R$ matrix in Table 4. In general, a high DIT corresponds to a high $R$, but a one-to-one correspondence between simple linear correlation and the information transmission relationship does not exist. A counter example occurs for stations 3 and 7 and stations 6 and 7. This implies that the two measures reflect different types of relations.

To help understand this discrepancy, stations 3 and 7 are considered. The $R$-value for stations 3 and 7 is high ($R = 0.93$), but both DIT$_{37}$ and DIT$_{73}$ are low (0.19 and 0.22, respectively). According to the correlation coefficient, the two stations are likely to be placed in the same group but according to the information transmission, they should not be. The frequency response matrix of station 3 given 7 turns out to be comparatively flat, wide, and low, as shown in Fig. 5 (e.g. compare this with Fig. 3). There is a ridge across the whole field, which explains the observed statistical similarity, reflected by the high correlation coefficient $R$. But the flatness and wideness of the response pulses suggests that the response of station 3 to a given observation at station 7 will become ambiguous. It is then difficult to tell what the real response will be. Therefore, the observations at station 7 do not give much information regarding the response at station 3. In other words, the information transmission between them is not well built. A similar conclusion can be drawn by inspecting the frequency response matrix of station 7 given station 3. It can be concluded that DIT is, in this case, a better measure than $R$ because the DIT not only reflects the similarity between stations, as $R$ does, but also further reveals the information inferring relations between two stations.

An example of the application of the DIT values in the data collection network design process is in regionalization, or grouping, of stations. This is shown in Fig. 6. In the figure, a noteworthy information relation is denoted by a line with an arrow towards the information inferee (predicted) station. Only DIT values higher than a given threshold value appear on the graph. Others do not appear because the information transmission relation presented by the corresponding station pair is too weak. If the threshold value for DIT is chosen as 0.35, then according to the grouping principles stated previously, stations 1, 2, 3 and 6 will be considered as a group and stations 4, 5 and 7 are in another group (Fig. 6). A similar diagram according to the correlation coefficient $R$ is presented as Fig. 7 to be compared with Fig. 6. The threshold value of $R$ is 0.9. Figs. 6 and 7 result in identical groupings with the exception of the linkage between station 7 and stations 3 and 6 which only occurs for the grouping based on the correlation coefficient. The fallacy of these two relations has been explained above. It is this fallacy which confuses the regionalization and will finally result in an improper solution.

The DIT transmission measure can also be used in the station selection procedure wherein one gauging station must be selected for retention in the data collection
Fig. 6. Station grouping according to DIT.

Station 1  Station 2  Station 3
      |       |
  Station 4  Station 5  Station 6
          |       |
              Station 7

Fig. 7. Station grouping according to correlation coefficient, $R$. 

Station 1  Station 2  Station 3
      |       |
  Station 4  Station 5  Station 6
          |       |
              Station 7
network from a group of stations. Since all of the stations in each group are mutually strongly related (both DIT values of every pair are above the threshold), further selection may be accomplished by the S-DIT criterion. Consider, as an example, the group consisting of stations 1, 2, 3 and 6 (group 1). The S-DIT values for stations 1, 2, 3 and 6 are listed in Table 5. In Table 5 station 2 has the highest S-DIT value so that it is retained with the highest priority.

The number of groups is controlled by the threshold value of DIT. A higher threshold value will lead to a larger number of (but smaller) groups and consequently a larger number of stations to be retained. The result for a lower threshold value is just the opposite. Hence, one can control the number of stations remaining in the network by manipulating the threshold value of DIT.

7. Conclusions and recommendations

Entropy and directional information transfer have been applied to a pilot data collection network study. To reduce the error in determining the bivariate probability density functions, non-parametric estimation is employed to estimate the two-dimensional density functions that characterize the joint distribution of the extreme flow data at pairs of gauging stations.

The directional information transfer, while essentially a measure of the information transmission, is useful in a network study to measure the association between gauging stations. As such, this measure finds wide application in the regionalization and station selection processes of data collection network design.

Future work could explore further the utility of the entropy concept within a data collection network design framework. Possible avenues for investigation are the integration of DIT with other considerations in the station selection process. In addition, this work dealt with extreme flow data which represent but one of the flow regimes of interest in data collection network design. Methods of simultaneously considering multiple components of the flow regime could be examined as well.

8. References


