DEVELOPMENT OF A METHODOLOGY FOR DATA COLLECTION NETWORK DESIGN

YU-JUIN YANG AND DONALD H. BURN
Department of Civil and Geological Engineering, University of Manitoba, Winnipeg, MB, Canada R3T 2N2

A data collection network design methodology is presented using entropy and a directional information transfer index (DIT). The methodology developed includes three stages. The first stage is to investigate the stochastic relationships amongst the gauging stations. The non-parametric estimation technique is used to estimate the joint frequency distribution for pairs of stations. The distributions are then used in the second stage to compute the directional information transfer (DIT) index. In the last stage, DIT is employed to assist with the network regionalization as a component of the station selection process. A streamflow network in southern Manitoba, Canada is used as a demonstration of the methodology.

INTRODUCTION

The design and operation of many water resource systems rely on hydrometric data such as rainfall and river flow. Hydrometric data are typically collected through the operation of a network of gauging stations. There are basically two tasks in data collection network design: network rationalization, which can involve shifting of resources of finances and manpower and possibly leading to network contraction or expansion; and network management, which deals with sampling issues such as when and how frequently samples should be taken at a gauging station. Methods of data collection network design have been widely studied and many technologies from other fields have been introduced into network design procedures. Some of the more widely used methods include the least square error approach and the clustering approach. However, both of these methods have shortcomings. The least square error approach is limited to linear or linearizable non-linear relations and leads to large computational burden. The results of the application of a clustering approach are affected by the selection of the association measure used to describe station similarity. The motivation of this paper is to develop a method which overcomes some of the disadvantages noted above.

The method to be developed herein is based on the entropy concept. Entropy, which is borrowed from communication theory, measures the information contained in a random

event, through observations of the event. The information transferred among information transmitters (predictor stations) and the information receivers (predicted stations) can be measured by the concept of 'mutual information'. Hydrometric gauging stations provide information about random hydrometric events. This information is contained in the observations. Entropy and mutual information can be used as the measures of the information either collected at gauging sites through the observations, or transferred, to some extent, to other sites. These features provide the basis of the adoption of entropy into hydrologic studies. 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; and (3) a description of the relationships among stations according to their information transmission characteristics.

Although entropy has been used in other areas of hydrology, the entropy concept has not been widely used in hydrometric data collection network studies. Application examples are Harmancioglu and Yevjevich (1987), Caselton and Husain (1980), and Husain (1989). Harmancioglu and Yevjevich (1987) used entropy to measure the information transmission among stations on the same river. Caselton and Husain (1980) developed a conceptual model employing the entropy concept for hydrometric network study. This model examines every potential configuration of a network, given the size of the network, and selects the one with the greatest information supply in terms of at-site and transmitted information. This model was applied to a rainfall network by Husain (1989). Since the objective in this work was to maximize the sum of at-site information for the retained stations and the information transferred from them, it does not systematically tend to retain the stations with high information transmission losses, or in other words, with high degree of independence. However, if these stations have high entropies, they should be retained so that the information at these sites will not be lost.

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 the above-mentioned works, continuous distribution functions were assumed and the calculation of entropies and the mutual information were based on them. This approach has several disadvantages. Firstly, 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 this difficulty is the non-parametric estimation of the density functions. 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