Journal of Hydrology 555 (2017) 228–241

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

A two-phase copula entropy-based multiobjective optimization approach to hydrometeorological gauge network design

Pengcheng Xu [a], Dong Wang [a,*], Vijay P. Singh [b], Yuankun Wang [a], Jichun Wu [a], Lachun Wang [c], Xinqing Zou [c], Yuanfang Chen [d], Xi Chen [d], Jiufu Liu [e], Ying Zou [e], Ruimin He [f]

[a] Key Laboratory of Surficial Geochemistry, Ministry of Education, Department of Hydrosciences, School of Earth Sciences and Engineering, State Key Laboratory of Pollution Control and Resource Reuse, Nanjing University, Nanjing, PR China
[b] Department of Biological and Agricultural Engineering, Zachry Department of Civil Engineering,Texas A & M University, College Station TX77843, USA
[c] School of Geographic and Oceanographic Science, Nanjing University, Nanjing, PR China
[d] School of Hydrology and Water Resources, State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, PR China
[e] Nanjing Hydraulic Research Institute, Nanjing, PR China
[f] State Key Laboratory of Hydrology, Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing, PR China

ARTICLE INFO

Article history:
Received 21 July 2017
Received in revised form 23 September 2017
Accepted 26 September 2017
Available online 28 September 2017

Keywords:
Hydrometeorological gauge network
Copula entropy
Mutual information
Multiobjective optimization

ABSTRACT

Hydrometeorological data are needed for obtaining point and areal mean, quantifying the spatial variability of hydrometeorological variables, and calibration and verification of hydrometeorological models. Hydrometeorological networks are utilized to collect such data. Since data collection is expensive, it is essential to design an optimal network based on the minimal number of hydrometeorological stations in order to reduce costs. This study proposes a two-phase copula entropy- based multiobjective optimization approach that includes: (1) copula entropy-based directional information transfer (CDIT) for clustering the potential hydrometeorological gauges into several groups, and (2) multiobjective method for selecting the optimal combination of gauges for regionalized groups. Although entropy theory has been employed for network design before, the joint histogram method used for mutual information estimation has several limitations. The copula entropy-based mutual information (MI) estimation method is shown to be more effective for quantifying the uncertainty of redundant information than the joint histogram (JH) method. The effectiveness of this approach is verified by applying to one type of hydrometeorological gauge network, with the use of three model evaluation measures, including Nash–Sutcliffe Coefficient (NSC), arithmetic mean of the negative copula entropy (MNCE), and MNCE/NSC. Results indicate that the two-phase copula entropy-based multiobjective technique is capable of evaluating the performance of regional hydrometeorological networks and can enable decision makers to develop strategies for water resources management.

1. Introduction

Hydrometeorological stations provide basic data on precipitation, water quality, air quality, streamflow and groundwater. The collected hydrometric data is needed for planning, decision making, and operation and management of water resources systems, including design of dams and reservoirs, flood forecasting, risk assessment of regional freshwater resources, allocation of water resources, and development of water distribution systems. An optimally designed network should provide sufficient information for these needs. There are additional needs for network design, such as increasing and conflicting water demands, climate change, long-term benefits, and quantifying the influence of network density on the precision of hydrological models (Mishra and Coulibaly, 2009). It is therefore desirable to design a hydrometeorological network that takes into account varied uses and users of hydrometeorological data.

The methodology for determining an appropriate strategy for spatial sampling of hydrometeorological variables, such as rainfall/streamflow, depends on the pre-existing conditions of the gauge network in the river basin: (1) ungauged, (2) gauged with not enough rain gauges, or (3) a dense network exceeding the requirement (Dong et al., 2005). Perhaps the most common networks are designed for observations at discrete points in space and time and to estimate the characteristics of a continuous field or flux (Pardo-Iguzquiza, 1998). Considerable research has been done on the design and evaluation of monitoring networks in

* Corresponding author.
E-mail address: wangdong@nju.edu.cn (D. Wang).

https://doi.org/10.1016/j.jhydrol.2017.09.046
0022-1694/

Nomenclature

Notation

d represents the number of random variables or stations. For dataset 1, d is 13. For dataset 2, d is 16

n the length of the hydrometric time series; For dataset 1, n is 132. For dataset 2, n is 366

$H(X_i)$ marginal entropy of a random variable X_i; If random variable X_i stands for hydrometric time series data (rainfall or streamflow) collected at station i, then $H(X_i)$ can quantify the information retained by station i

$H(X_1, X_2, \ldots, X_d)$ the multivariate joint entropy of random variables collected at these d stations

$p(x_1, x_2, \ldots, x_d)$, $P(x_1, x_2, \ldots, x_d)$ the joint probability density function and distribution function of random variables collected at these d stations

$p_{X_i}(x_i), P_{X_i}(x_i)$ the marginal density function and the distribution function for the hydrometric time series data collected at station i

$u_1, u_2, \ldots, u_d$ the marginal distribution function of the hydrometric time series data collected at station 1,2,...,d respectively; $u_i = P_{X_i}(x_i)$

$C(u_1, u_2, \ldots, u_d)$, $c(u_1, u_2, \ldots, u_d)$ the d-dimensional copula distribution function, and copula density function

$I(X_i, X_j)$ the mutual information of two random variables X_i and X_j collected at station i and j

$TC(X_1, X_2, \ldots, X_d)$ the total correlation of random variables collected at station 1,2,...d respectively

$\hat{P}_i(x)$ the estimator of the marginal distribution

$CDIT_{ij}$ represents the fractional information inferred by station i about j

$CDIT_{ji}$ represents the fractional information inferred by station j about i

θ the copula parameter

TA_t the "true" areal mean rainfall, t stands for the tth observation of the hydrometric time series

SA_t the sampled areal mean rainfall (or streamflow) from the gauge combination

NSC the Nash–Sutcliffe Coefficient

$MNCE$ the mean negative copula entropy

surface water hydrology. The approaches which have been commonly used for network design can be summarized as (1) statistically based methods, (2) spatial interpolation techniques, (3) information theory-based methods, and (4) hybrid methods.

For over half a century, statistical regression techniques have been widely applied to locate gauging stations (Moss and Tasker, 1991). The most representative statistical method is the generalized least squares (GLS) method which maximizes regional information within a limited budget and time horizon (Moss and Tasker, 1991). Spatial interpolation techniques have been employed in the hydrometeorological field to get station data (Camera et al., 2014; Bechler et al., 2015). Geostatistical techniques and methods are employed to do spatial interpolation on the basis of characteristics which include global versus local, exact versus approximate, point-based versus areally-based, and the ability to consider covariates. The kriging methods, the most well-known spatial interpolation methods, are composed of two parts: spatial variation analysis using variograms and calculation of distance weights for interpolation (Goovaerts, 2000; Li and Shao, 2010; Adhikary et al., 2015; Zhang et al., 2014; Aalto et al., 2016). Goovaerts (2000) compared three multivariate geostatistical algorithms by incorporating a DEM (digital elevation model) into the spatial prediction of rainfall: simple kriging with varying local means, kriging with an external drift, and co-kriging. It was observed that prediction would improve if correlated secondary information, such as a DEM, was taken into account. This observation corroborated the findings of Creutin and Obled (1982). However, they were not able to capture nonlinear relationships or unambiguously identify the directionality of coupling if they were asymmetric, that is, if variable X influenced variable Y at one time scale, while variable Y influenced variable X at a different time scale. These limitations have prevented the use of correlation-based techniques for robust analysis of complex systems, where feedback is important, and hydrological frequency analysis should take non-stationarity into consideration (Liang et al., 2012).

The concept of entropy has been popular in the hydrometeorology and water resources literature since the 1970s (Singh, 1997). The fundamental basis in designing networks based on the entropy concept is that the gauging stations should have as little transinformation as possible, meaning that the stations should be independent of each other as far as possible. Transinformation is defined as mutual information in the entropy theory. Husain (1987) assumed bivariate and multivariate continuous distributions to determine entropy. Information transmission between station pairs was calculated for different cases of probability distributions and on the basis of information maximization, the optimum locations of stations to be retained were identified. Krstanovic and Singh (1992a,b) evaluated rainfall networks in Louisiana using joint entropy and transinformation for which multivariate distributions were determined using the principle of maximum entropy. Yang and Burn (1994) proposed directional information transfer index (DIT) for measuring the information flow between gauging stations in the network. Entropy-based assessment of water quality monitoring networks was done by Ozkul et al. (2000). Markus et al. (2003) used a hybrid combination of GLS and DIT index by including a function of the negative net information as a penalty function in the GLS. The weights of the hybrid combination were determined to maximize the average correlation with the results of GLS and DIT. They found that no matter how big differences in entropy values were when changing the bin size, the ranking of stations in terms of the difference between the information received and the information sent remained, in most cases, unchanged. The problem of the interval size was first analyzed by Amorocho and Espildora (1973), and was also used by Steuer et al. (2002) to calculate mutual information for discrete variables. Evaluating the effect of class interval size, Singh (1997) found that the entropy value decreased with increasing class interval size and increasing sampling interval. Al-Zahrani and Husain (1998) used Shannon entropy to evaluate an existing hydrological network located in the southwestern region of the Kingdom of Saudi Arabia and examined its suitability for providing maximum hydrological information.

Recent years have witnessed the adoption of hybrid methods. Chen et al. (2008) combined kriging with entropy to determine the optimum number and spatial distribution of rain gauge stations in catchments. Mishra and Coulibaly (2014) assessed the effect of seasonal streamflow information implementing the kernel density approach for estimating mutual information between gauging stations. Mahmoudimeimand et al. (2015) used an optimization model based on entropy and kriging using GIS for determining the number and location of rain gauges. The candidate stations were those with minimum variance of kriging error and maximum information entropy.