This article was downloaded by: [Montana State University Bozeman]
On: 14 August 2014, At: 09:44
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Hydrological Sciences Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/thsj20

Group-based estimation of missing hydrological data: I. Approach and general methodology

AMIN A. ELSHORBAGY [a] , U. S. PANU [b] & S. P. SIMONOVIC [c]

[a] Civil and Geological Engineering Department , University of Manitoba , Winnipeg, Manitoba, R3T 5V6, Canada E-mail:

[b] Civil Engineering Department , Lakehead University , Thunder Bay, Ontario, P7B 5E1, Canada E-mail:

[c] Department of Civil and Environmental Engineering , University of Western Ontario , London, Ontario, N6A 5B9, Canada

Published online: 25 Dec 2009.

To cite this article: AMIN A. ELSHORBAGY , U. S. PANU & S. P. SIMONOVIC (2000) Group-based estimation of missing hydrological data: I. Approach and general methodology, Hydrological Sciences Journal, 45:6, 849-866, DOI: 10.1080/02626660009492388

To link to this article: http://dx.doi.org/10.1080/02626660009492388

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms

& Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

Group-based estimation of missing hydrological data: I. Approach and general methodology

AMIN A. ELSHORBAGY
Civil and Geological Engineering Department, University of Manitoba, Winnipeg, Manitoba R3T 5V6, Canada
e-mail: umelshor@cc.umanitoba.ca

U. S. PANU
Civil Engineering Department, Lakehead University, Thunder Bay, Ontario P7B 5E1, Canada
e-mail: umed.panu@lakeheadu.ca

S. P. SIMONOVIC
Department of Civil and Environmental Engineering, University of Western Ontario, London, Ontario N6A 5B9, Canada

Abstract In this first paper in a set of two, the problem of estimating missing segments in streamflow records is described. The group approach, different from the traditional single-valued approach, is proposed and explained. The approach perceives the hydrological data as sequence of groups rather than single-valued observations. The techniques suggested to handle the group approach are regression, time series analysis, partitioning modelling, and artificial neural networks. Pertinent literature is reviewed and background material is used to support the group approach. Implementation and comparisons of models' performance are deferred to the second paper.

L'approche de groupe pour l'estimation des données hydrologiques manquantes: I. Présentation et méthodologie

Résumé Dans ce premier de deux papiers, nous décrivons le problème de l'estimation de suites de données manquantes dans les archives de débits. Nous présentons et expliquons l'approche de groupe, différente des approches traditionnelles focalisées sur l'estimation de valeurs singulières. Cette nouvelle approche conçoit les données hydrologiques comme des suites de groupes plutôt que comme des suites d'observations singulières. Les techniques susceptibles de la servir sont: la régression, l'analyse des séries chronologiques, la segmentation et les réseaux de neurones artificiels. Nous présentons une revue de littérature d'où nous avons tiré des arguments en faveur de la promotion de l'approche de groupe. L'implémentation et l'évaluation de l'approche de groupe font l'objet du second papier.

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

INTRODUCTION

In the area of water resources planning and management, complete data sets are required on many variables such as rainfall, streamflow, evapotranspiration and temperature. Unfortunately, records of hydrological processes are usually short and often have missing observations. The existence of data gaps might be attributed to a number of factors such as interruption of measurements because of equipment failure, effects of extreme natural phenomena such as hurricanes or landslides or of human-induced factors such as wars and civil unrest, mishandling of observed records by field personnel, or accidental loss of data files in the computer system. Deficiencies in hydrological data series vary from 5 to 10% in the case of runoff data and up to 25%

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

in the case of oceanic storm surges (Panu *et al.*, 2000). Most hydrological models do not tolerate missing observations and thus data in-filling techniques have evolved to deal with incomplete data sets.

Attracted by the importance of estimating missing data, hydrological researchers have adopted and developed various models and techniques to deal with the problem. Not only are efforts devoted to extending short records by adding lengthy segments of estimated data, but also attention is given to the gaps of short duration. One can find in the literature cases in which sophisticated techniques are used for estimation of a single missing observation (e.g. Griffith *et al.*, 1985). In water resources, the commonly used techniques for estimation of missing data are based on regression analysis, time series analysis, artificial neural networks and interpolation techniques. The diversity of the above-mentioned techniques does not necessarily indicate diversity in the approach. A major commonality exists in most of the applications of these techniques; that is, any hydrological time series record is perceived as a sequence of single-valued observations irrespective of the time scale of the data. In hydrological data (e.g. streamflows), it can be noted that annual or seasonal data might be independent, while monthly or weekly data of the same river have significant levels of autocorrelations. Different techniques are employed for modelling annual and monthly data; however, in both cases (year 1, year 2, ... or month 1, month 2, ...) observations are treated, in the literature, as single-valued entities that have interrelations modelled at one stage.

In this paper, it is argued that hydrological data can have a hierarchical type of structure that needs to be modelled in more than one stage. For example, when different months and different seasons are recognized in the same time series, months within each season can be modelled in the first stage and inter-season relationships can be modelled in a second stage. This approach is identified here as "group approach". The underlying difference between the traditional single-valued approach and the group approach is that in the latter, data are perceived as groups that have inter- and intra-relationships. When the monthly flows are autocorrelated and the seasonal flows are independent, the flows can be described as independent groups (seasons) of correlated elements (internal single-valued monthly observations). In the same manner, one can have independent groups of independent elements or correlated groups of correlated elements. It is worth mentioning that available techniques such as regression and time series analysis can be modified to cope with the group approach. This is the first of two papers in a set, as follows:

- This paper outlines the theory of the group approach. First, the existing techniques using traditional single-valued approach for estimation of missing data are reviewed. Then, the group approach is presented along with the methodology for data in-filling.
- A companion paper (Elshorbagy *et al.*, 2000) details the models used for estimating the missing values using the group approach and compares their relative performance in terms of accuracy and ease of operation.

EXISTING TECHNIQUES FOR ESTIMATING MISSING DATA

Researchers have been tackling the problem of missing data in different ways and from different perspectives as well. Even their definitions of "missing data" and the expressions that they have used to describe the in-filling process are no less diversified

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

than the different techniques that they have used. A group of researchers tackled the problem of intermediate missing data where data or observations before and after the missing observations are available (e.g. Lettenmaier, 1980; Griffith *et al.*, 1985; Dax, 1985; Mott *et al.*, 1994; Gyau-Boakye & Schultz, 1994; Bennis *et al.*, 1997). The words *patching* (e.g. Hughes & Smakhtin, 1996; Makhuvha *et al.*, 1997; Pegram, 1997) or *filling* (e.g. Panu, 1992; Gyau-Boakye & Schultz, 1994) are used to express the in-filling in these cases. Others consider the cases in which data are available only from one side of the gap, or where the gap is so lengthy that the data set is considered bounded from one side only. Generally, the in-filling process in this case is called *extension* (e.g. Hirsch, 1982; Alley & Burns, 1983; Hughes & Smakhtin, 1996), *synthesizing* (e.g. Beauchamp *et al.*, 1989; Simonovic, 1995), or *estimation*, which is also used to indicate patching type of in-filling (e.g. Ben-Zvi & Kesler, 1986; Berkowitz *et al.*, 1992; Bennis *et al.*, 1997; Knotters & van Walsum, 1997). The word *augmenting* is used especially when the overriding objective is to estimate the model parameters and the estimation of the missing data comes as a direct result of applying the developed model (e.g. Fiering, 1962; Gilroy, 1971; Moran, 1974; Vogel & Stedinger, 1985). It should be noted that other words such as *reconstruction* (Hirsch, 1979) and *completing* (Dax, 1985) are also employed to denote estimating missing data. In the above-mentioned cases, single variable and multi-variable applications are found in the hydrological literature. This paper explores the first type: "in-filling" or "patching", where data before and after the gap are available.

Missing data can be categorized into three groups:

(a) Data of trivial importance are missing (e.g. a few sparsely distributed, not consecutive, missing observations in a long historical record). In this case a simple in-filling method such as in-filling by using series average or simple interpolation could be satisfactory. Peak or extreme values should not be encountered in the data gaps to be in-filled by the simple in-filling methods.

(b) Fundamental data are missing (e.g. lengthy segments or many intermittent observations) where data patterns or structure cannot be recognized from the remaining record. In this case, any attempt at in-filling such a record may be unreliable and therefore the whole record should be dropped according to the current available state of knowledge and techniques (Beale & Little, 1975).

(c) Significant data are missing (e.g. a segment of consecutive observations). In the latter case, the missing data are considered important enough (quantitatively or qualitatively) to deserve developing a technique that estimates them as accurately as possible. At the same time, the data gaps are too short to have significant damaging effect on the patterns and structure of the whole record (Elshorbagy *et al.*, 1999).

Since missing values under the third category occur more often, this category is the focus of this paper. This case could happen as a result of interruption of measurements because of prolonged equipment failure, stopping the measurements at some stations for any reasons (e.g. budget limitations) and resuming them after some time, and accidental loss of data files. Further, the available literature on estimation of missing hydrological data can be classified into single-valued approach (such as regression, time series analysis, interpolation, and artificial neural networks) and group approach. Only research works closely related to the topic of this paper are reviewed hereafter to give a general view of the available literature on estimation of missing hydrological data. The literature related to extending short hydrological records is

extensive. However, it is not presented here because the overriding objective of record extension is to maintain statistical properties of the time series (e.g. mean, variance, etc.). The objective here is to estimate missing segments (few consecutive observations) in a way that minimizes the error (difference between actual and estimated values).

Single-valued approach

Regression analysis Methods for analysing multivariate data with missing observations from one or more independent variables are presented by Beale & Little (1975). Their objective is mainly to develop a model that describes the data set. Multiple Nonlinear Standardized Correlation (MNSC) analysis is used by Simonovic (1995) for streamflow data in-filling, extension, and generation. Although the technique incorporates the nonlinearity, it is mentioned here because it is based on transforming data to their logarithms, replacing the logarithms by their empirical probabilities, and then replacing the probabilities by their standardized variables. This means that data are linearized before proceeding further with regression. The problem of patching missing rainfall data is described by Makhuvha *et al.* (1997): the Expectation Maximization (EM) algorithm and a modified version of it, which they name pseudo-EM, are the basis for their analysis. A missing or bad block of climate data is filled or replaced using a linear regression technique by Mott *et al.* (1994). The point of interest in their research is that the authors thought of every observation as a member of a group, but groups and groupings as exist in literature are not invoked in their proposed methods.

Time series analysis Following the forecasting approach, forward and backward estimates were derived for each missing value by Lettenmaier (1980). A single value out of the two estimates was calculated. The importance of considering both time and space is put into focus by Griffith *et al.* (1985). Space–time interdependencies are assumed to be characterized by a first-order Markov structure. The cases of single missing value and cluster of missing values are addressed in their paper. A trial for improving single variable and multivariable techniques for estimating missing hydrological data is made by Bennis *et al.* (1997). They propose the use of Kalman Filter (KF), combined with AR model, instead of ordinary least squares (OLS) technique to allow for time variant model parameters. Box-Jenkins transfer/noise models, known to be single-output models, are extended by Van Geer & Zuur (1997) to a multiple output model. The models are extended to unmeasured locations to estimate unmeasured values of groundwater.

Interpolation approach When the observed phenomenon is characterized by spatial configuration, like the case of a groundwater wells network, a simple function of the space using the historical observations can be formulated (Ben-Zvi & Kesler, 1986). Their model represents a space surface fluctuating parallel to itself on the time axis and is used to estimate any missing data. The constraint of parallel fluctuations is removed by the method proposed by Berkowitz *et al.* (1992). Their method is based on integration of two sources of information: prior estimates assuming stochastic beha-

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

viour of the system and online estimates assuming probability distribution function of field data conditioned on the system state. Another form of using the interpolation technique for estimating missing observations is presented by Hughes & Smakhtin (1996). It is based on the use of 1-day flow duration curves for each month of the year.

Pattern recognition An example of using pattern recognition technique with a single-valued data approach is the work of Zhang & Berndtsson (1991). They use pattern recognition as a useful technique when patterns in space and time need to be evaluated simultaneously. They handle the issue of soil water dynamics by introducing both a time lag and a spatial lag depending on the depth in the calculation of the covariances between consecutive pairwise combined time series of soil water content. It is possible to trace a pulse of infiltrated water through a correlation matrix representing the time and space that is used directly for interpolation and estimation of values at unmeasured locations.

Artificial neural networks (ANNs) The last decade has witnessed many applications of ANNs in water resources. These include, among others, rainfall forecasting (French *et al.*, 1992), multivariate modelling of water resources time series (Raman & Sunilkumar, 1995), modelling of rainfall–runoff processes (Hsu *et al.*, 1995), flow forecasting (Zealand *et al.*, 1999; Dawson & Wilby, 1998), and river level forecasting (See & Openshaw, 1999). The promising results due to the use of ANNs in water resources make them a feasible technique to be employed in this research. The literature related to the use of ANNs with missing or incomplete data is limited to a few papers. Kuligowski & Barros (1998) have used ANNs to estimate missing rainfall data and found that results compare favourably to regression and simple techniques such as arithmetic and distance-weighted averages of the values from nearby gauges. Outside the water resources domain, ANNs are used by Gupta & Lam (1996) to estimate missing values in a multivariate data set. The accuracy of the results is superior to that obtained by iterative regression analysis.

The literature presented above follows the traditional methods of hydrological data analysis, which are based on information contained in individual data. These methods ignore information contained in and among groups of data. A few published reports that are available and which use the concept of extracting information from data groups are briefly presented below.

The group approach

A few published reports can be found in the area of water resources in which a data set is dealt with as a sequence of recognized groups. For example, three group types are identified in hourly water demand data by Shvarster *et al.* (1993). They use a time series analysis technique to model the intra-group structure. However, the data within each group are treated as single-valued observations. Both inter-group and intra-group structures are claimed to be provided by Panu & Unny (1980). A Markovian model is assumed to describe the inter-groups relations while each group type is divided into sub-groups. This further division helps in their objective of generating synthetic data but cannot be considered as intra-group modelling. No model is specified in their work

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

to describe the underlying mathematical relation that links the individual observations within each group. Their model structure would allow them to deal with a group as a whole but not with a partial group. It should be noted that the emphasis of Panu & Unny (1980) is the generation of synthetic streamflow data while Shvarster *et al.* (1993) are concerned with forecasting the hourly water demand. Also, the researches mentioned earlier (Dax, 1985; Mott *et al.*, 1994) indicate, implicitly, the existence and significance of groups in the hydrological data. A crude method of grouping is noticed in the work of Dax (1985). The author builds the interpolation technique for estimating missing data on the assumption that each year is composed of two seasons, rainy season with water levels monotonically increasing and dry season with water levels monotonically decreasing. Estimated data are supposed to follow the pre-assumed pattern, though data within each season are treated as single-valued observations. Khalil *et al.* (1998) indicate the possibility of in-filling missing data, using ANNs, based on the group approach. However, questions of how to employ the group approach for in-filling missing data and how to adapt the ANN to perform it are not clearly answered in their work.

Perceiving data as groups can give insight into the data structure and affect the results of further analysis using the data under consideration. For a long time, annual flood peaks have been assumed to be serially independent. Lately, it has been shown by Booy & Morgan (1985) that flood peaks on the Red River at Winnipeg, Canada, show clustering characteristics. As a result, flood risk is shown to be substantially higher than that estimated by conventional methods, which assume serial independence of the peak flows. This supports the grouping found by Panu & Unny (1980) in some annual streamflow data and the long-term persistence in annual streamflows shown by Hurst (1951, 1956).

PERCEPTION OF GROUPS IN HYDROLOGICAL DATA

Perceiving the hydrological data as groups can be considered different from the single-valued approach in the underlying algorithm and the modelling mathematics. The concept of data groups has been always recognized in the traditional single-valued approach. Assuming that the streamflows can be discretized into hourly units (we assume the hour as the smallest unit and flows are constant over the one hour duration), then 24 observations per day and 720 observations per month can be obtained. In hydrological models dealing with monthly flows, the 720 values of each month are aggregated in one value, which is the average of these values. So, when a monthly value is predicted or forecasted based on the previous monthly value, in essence, it is a group of 720 hourly values that are forecasted based on the previous group of 720 observations. Condensing the whole group in one value is viewed here as a way of simplifying the complexity of the needed mathematics when the multi-dimensionality of the groups is maintained. In this regard, it is argued that the concept of data groups is widely recognized but traded for simplicity in hydrology.

Another way of recognizing groups in hydrological literature is through the application of disaggregation models. As a first step, data are treated as aggregate values and, in a subsequent step, these aggregate values are disaggregated into multi-dimensional elements (Valencia & Schaake, 1973). The mathematics needed to deal

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

with the group vectors are simplified by dividing the modelling procedures into the above-mentioned two steps. Alternatively, if disaggregation model is used to estimate a missing season (group of values), an aggregate value of the season has to be estimated first and then disaggregated into its original elements. Performing these steps entails two stages of parameter estimation and hence more complications.

A clear diversion from the traditional single-valued approach in the stochastic hydrology literature is a work by Panu *et al.* (1978) based on concepts of pattern recognition. Their proposition was made for the sake of generating synthetic hydrological sequences. In generation, only the statistical properties of the series are important, and they are the parameters that should be preserved. The generated sequences could be far from the original data in terms of squared error, yet generation process can be successful as long as the statistical properties (mean, variance, skewness) are preserved. Where in-filling or estimating missing data is considered, minimizing the squared error (difference between estimated and true values) becomes the overriding objective. The proposition of Panu *et al.* (1978), which is briefly described and clarified here, can be modified for in-filling of missing data.

The hydrological time wave (HTW) form is a plot of a random variable, such as rainfall or discharge, *vs* time. Any section of this HTW, for example a section corresponding to a time duration of one season in a geophysical year, is an object. Thus the HTW consists of many objects. A collection of all such objects is considered to constitute a universe. The simplest example of a universe is that containing hydrological objects from two categories, namely, the dry season and the wet season, describing all the objects of the HTW. The HTW or any observed hydrological sequences of data or observations can be converted into a hydrological "signature" by joining the data points by a continuous line representing simple geometric figures. The characteristics of such a signature can be evaluated in terms of several different specified features. These features and the collection of such features can be considered equivalent to letters and words, respectively, in a written language. This interesting analogy with written language is made by Panu *et al.* (1978), because it is there that pattern recognition has found wide application:

"....Written language is an expression of human thought in signature. Such a signature contains a proper set of words arranged according to a specified grammar in order to provide proper meaning. It has been observed that any sequence of letters, as well as any sequence of words in a written language, is stochastic in character. It is groups of letters that provide a meaning to the stochastic written language. For example, individual letters, such as W, I, N, D, and O, carry no meaning. However, if they are considered as a group of letters (WINDO), the group not only provides a sense but also indicates that it is an incomplete group and the next most probable letter to complete the group is the letter W. In this case the occurrence of the letter W depends on the previous sequence of letters...".

In this research the above argument is accepted and also extended as follows:
If (WIN) is available, then the process of guessing whether it is complete or not is not an easy task. If it is not complete, then one letter could be missing. That letter might be D, E, G, S, or K. If more than one letter is assumed missing then the number of possibilities increases and might become an unmanageable process. Two major criteria can be considered to limit the number of possibilities and to facilitate the task. The first criterion is the group length and the second criterion is the existence of groups (words)

that come before and after the group (word) under consideration. If one knows that the group length is four units long and that only one letter is missing, one needs to consider the letters that come before and after the word (WIN). Assuming the following letters are forming the series:

TODAYISCOLDBUTTHE*WIN*-ISCALM (1)

If the grammar is learnt from the previous letters and sentences and as well from other related texts, this sequence of letters might be segmented and written as follows:

TODAY IS COLD BUT THE WIN- IS CALM (2)

In this case, the missing letter is, most probably, the letter *D*. The estimated letter depends on: first, segmenting the sequences into meaningful groups (words), second, knowing the length of the group where some data are missing, and third, defining the structure of the whole sentence before and after the group with missing data. These three steps cannot be conducted without learning the grammar of the language. Returning to the hydrological application on monthly streamflow data, the analogy is made as follows: the letters represent the monthly values, the words might represent seasons (few consecutive months), and the sentence might represent the whole year. If the HTW is regarded as a meaningful text, it is conceptualized that monthly observations are linked in a specific way to form the seasons, seasons are linked in a specific way to form the year, and also years might be linked in a specific way to form the whole series. The problem in hydrological series is that such a language of the time series and its grammar is neither well defined nor understood. Grammar, in this case, has to be deciphered and learnt from the whole text (time series under consideration) and also from any other available related texts (related time series). To help understand the terminology used in this paper, a brief explanation of all such terms follows.

A *segment* is a vector of consecutive observations. For example, in the case of monthly streamflows, January, February, March, and April can form *segment 1*; May, June, July, and August are elements of *segment 2*; and September, October, November, and December constitute *segment 3*. A segment can be referred to as *group, pattern vector* or *object*. Each segment may indicate different flow conditions (e.g. wet, dry, semi-wet, etc). A *segment type* is a collection of all segments from the entire record that have similar flow conditions (e.g. wet conditions). Words such as *class*, *pattern class*, *season*, *group type*, or *cluster* are used in literature of different domains. An *element* is the smallest unit of a segment, which is the monthly flow in the case of a monthly time series record. It may also be referred to as an *attribute*. The *dimensionality* represents the number of dimensions in the problem and is defined according to the number of elements in the segments (e.g. in the example herein, it is a four-dimensional problem). Based on these definitions, a 50-year record of monthly flows has (50*12) = 600 observations that are segmented into (600/4) = 150 *segments*. These four-dimensional segments are clustered into three pattern classes.

PROPOSED APPROACH AND METHODOLOGY

Every hydrological observation in a time series record is conceived as a member of a season or a group of observations. Such a concept requires the hydrological data record to be segmented into parts. These parts might be named segments, groups or seasons.

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

Each group consists of a number of consecutive observations. For example, consider a monthly streamflow record of 50 years: the entire data set includes (12*50) = 600 observations. Plotting these observations on a graph, one can observe that well-defined patterns, which can be proved by further investigations, exist in the data. So, the first step in the proposed methodology is to divide the data into segments: assume 150 segments are obtained out of the 600 observations. These segments need not be of equal length (i.e. an equal number of months in each segment is not required). Also, a month can be associated with different group members (months) every year. Implicitly this means that one month might fall in a wet season in one year, semi-wet in another year, dry in a third year, and so on. At the same time, having segments of different lengths means that the hydrological seasons are of different durations. The second assumption seems to be feasible and close to reality, while the first one is a problematic assumption for several reasons. The first reason is that historical observations show that every month has common hydrological properties, which make it a member of the same season every year. Therefore, considering every month to be a member of a specific fixed group repeating over the years not only sounds logical but also is a realistic assumption from the hydrological viewpoint. The second reason relates to the concept of modelling itself, where generalization and pre-determined pattern structure are sought and modelled to mimic reality as closely as possible. If a small percentage of the observations are not following the assumed pattern, they are considered as exceptions and are included in the error probability of the model. Consideration of the exceptions, or the minority of the observations, in formulating the model structure suggests that they may constitute a significant pattern in the data set. Referring to the example of 50 years of monthly data, many sets or replicates of 50 years data may be required in order to track the patterns of the minority. In other words, hundreds and probably thousands of years of data will be needed for the analysis, which imposes an obstacle that hinders any progress in the modelling approach. Since the objective of modelling in general, including the approach proposed herein, is to ensure that simplicity and closeness-to-reality can coexist in hydrological modelling, the assumptions are formulated as follows: (a) the segments are allowed to be of different lengths, and (b) definite types of groups are recognized and assumed to be repeating over the years. It should be noted that repeating groups share common statistical properties, not identical internal values.

After segmenting the data as a first step, the second step in the methodology is to prepare the segmented data for further analysis. The analyses or the models which adopt the single-valued approach in stochastic hydrology require that the data set satisfy some basic statistical characteristics. The major characteristics are normality, trend, seasonality and correlation structure. What has to be achieved in order to apply the proposed group approach is the identification of the statistical properties of the data groups rather than the individual observations. Stationarity, multivariate normality within the groups, seasonality, and independence or correlation functions among the groups should be checked before proceeding further with any analysis.

The third step of the proposed methodology is to model the data as a sequence of groups in order to in-fill the missing segments. Two cases of applications are addressed and are explained below.

Case 1—single series case: where only one time series is available for the analysis. It is the time series with data gaps that need to be filled. From now on, this time series

record is named as the *target time series*. In cases of streamflow data, it can be called "target river". Many scenarios of data gaps can occur in such a record (Panu, 1992). The cases of entirely and partially missing segment are assumed in this paper. A schematic diagram of the single series case is shown in Fig. 1.

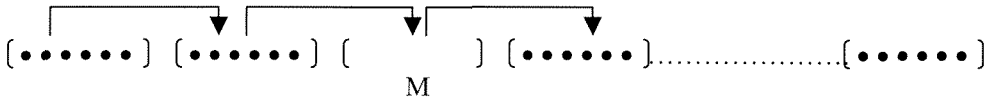

Fig. 1 Missing data indicated as M in target time series (single series case).

Case 2—multiple series case: where time series other than the target one are also available. These series are referred to as *reference time series*, or "reference rivers" in the case of streamflow data. They should be cross-correlated to the target series. The number of the reference time series could be one or more, and they need not to be restricted to the same hydrological variable of the target series (e.g. flows, precipitation, temperature, etc.). The bi-series case, which is a special case of the multiple series, is considered in these papers. A schematic diagram of the bi-series case is shown in Fig. 2.

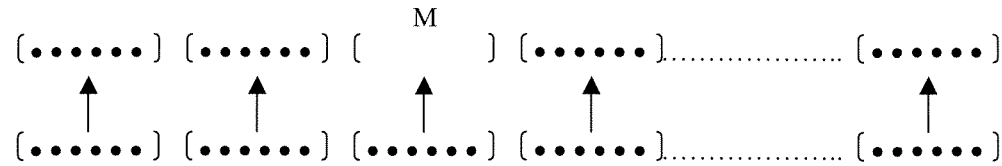

Fig. 2 Missing data indicated as M in target time series (bi-series case).

The analogy between the traditional single-valued approach and the group approach is depicted in Figs 1 and 2. In the proposed group approach, the segment of consecutive observations replaces the single-valued observations. Therefore, techniques which are capable of modelling hydrological data according to the single-valued approach can be modified to cope with the group approach. Moreover, the techniques and the models that have been developed to deal with groups rather than single-valued data can also be employed for analysis of groups.

In the proposed methodology, each group is viewed as an object or a rigid segment of multiple elements (observations). These elements behave as a group in relation to other groups. In this regard, the intra-group structure needs not to be modelled separately from the inter-group model. Development of a model that treats the data as a sequence of multi-dimensional segments is considered satisfactory to handle inter and intra structures of data groups. At the same time this group character (as one object) is utilized in handling partial groups as will be shown in the companion paper.

Based on the previous discussion, statistical analyses of data groups play the major role in selecting the most appropriate technique for the analysis. In the single series case, the time series analysis can be employed after modifying the formulations to deal with groups. It should be noted that time series models will be applied to the single series case when autocorrelations among groups are considered significant. The ANNs and multivariate partitioning modelling (MPM) (Johnson & Wichern, 1988) are also

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

used for the single series case. The ANN technique can deal easily with groups with no required modifications, and the MPM is designed mainly to handle data vectors (groups). The details of these techniques are provided in a later section. It is worth mentioning that ANNs can handle both autocorrelated and independent groups. The conditions of significant cross-correlation and insignificant auto-correlation among groups are assumed for application of the proposed methodology in the case of multiple series analysis. As in the single series case, the techniques of ANNs and MPM are utilized. Further, the regression analysis technique lends itself to the analysis when cross-correlation between the *reference* and *target* time series is significant and groups are considered to be serially independent within each of the participating time series. According to the group approach, regression technique should be modified to handle the situation of multiplicity in both of predictors and response variables. It is found that multivariate multiple regression (MMR) (Johnson & Wichern, 1988) can be employed to deal with group regression.

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

DATA SEGMENTATION

As the first and most significant step in the group approach, data segmentation is further elaborated in this section. The analysis pertaining to groups is perhaps best accomplished through pattern recognition. The analogy, discussed earlier, between groups of hydrological observations and formal language makes pattern recognition a good candidate for segmentation and recognition of a definite number of classes in the hydrological data.

Segmenting the data into a definite number of repeating groups entails answering three questions. The first question is: What is an appropriate number of groups? The second is Where and how can the boundaries between different groups be drawn? and the third is How does one test or verify the performance of the segmenting technique? Through the hydrological literature, the first two questions are addressed through the use of periodogram as a tool for determining the number of seasons or harmonics in a year (Panu *et al.*, 1978; Khalil *et al.*, 1998). The segmenting of hourly water demand data is done by observation only in Shvarster *et al.* (1993). Additional elaborations concerning the grouping process are not considered in their work. The statistical method of bi-plot is used by Pegram (1997) to check the consistency of multiple sites or months to form a group. The third question has not been given enough attention.

The problem can be described in the pattern recognition context by considering each segment type as a pattern class (PC). The segments from the same type can be treated as pattern vectors (PV). So in the case of monthly streamflows, a segment (e.g. January, February, March, and April) is a PV where every month is an attribute of that vector. Pattern recognition may be regarded as a technique of discriminating the input data, not between individual patterns but between populations, via the search for features or invariant attributes among members of a population (Tou & Gonzalez, 1974). The features that represent differences between pattern classes are called inter-class features while the features that characterize attributes common to all patterns belonging to one class are called intra-class features. Although the analogy between the pattern recognition field and the problem at hand is feasible, one should bear in mind two important points. First, considering each month as an attribute of the PV implies the case of multivariate or multi-dimension, but all attributes are of the same unit of measurement

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

(flow, volume of water/time). Second, a learning set of data where the patterns are already of known classes is not available. Therefore, our hydrological problem will come under the umbrella of the unsupervised learning technique. Suppose that one is given a set of pattern vectors $\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N\}$ of unknown classification: a good way to let them cluster themselves in groups and get separated from each other is by calculating their distance from a fixed reference. The motivation for using distance functions as a clustering and/or classification tool follows naturally from the fact that the most obvious way of establishing a measure of similarity between pattern vectors, which are also considered as points in the Euclidean space, is by determining their proximity (Tou & Gonzalez, 1974). In some cases of the hydrological data observations with higher mean values tend to be of higher variance. If the flow value and the variance are taken as two criteria for segmenting the data, then streamflow data can be clustered according to their value and variance. It is in this sense that seasonal mean flow and variance can be used as inter-class features. By calculating the distance between every month and the time series mean value as a reference, observations might get clustered (e.g. January, February, etc. as PC1, May, June, etc. as PC2, and so on) for the N years. Hence the pattern vectors can be denoted as follows:

$$\boldsymbol{PV}_i^{(j)} = [x_1, x_2, ..., x_k] \tag{3}$$

where the superscript (j) refers to pattern class, the subscript (i) refers to the year, and k is the number of months in the class.

Moving to the question of how many classes should be used, one can say that outside the hydrology domain, this problem is recognized by some researchers. The difficulty inherent in estimating the number of classes is demonstrated by Dubes (1987). Many methods are proposed for this purpose, some of which depend on minimizing the squared error by calculating the distance from each PV to the centroid of the class. Criteria are developed for stopping rules or preventing more splitting of classes (Duda & Hart, 1973).

In the hydrological application herein the problem seems to be unique because the required segmenting entails also deciding on the dimensionality of the pattern vectors and the Euclidean space. By segmenting the monthly data, the dimensionality of the Euclidean space, which may be different for each pattern class, is determined. This takes us back to our proposition of calculating the distances from each month to the mean of the whole series. Plotting these distances and their variance can help easily segment them, visually, into classes.

After segmenting the time series into m pattern classes, PC1, PC2, PCm, these classes need to be checked in order to evaluate the segmentation process. The Euclidean distance measure equation (4) readily lends itself to the process because of its familiar interpretation as a measure of proximity. The Euclidean distance between two pattern vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is given by:

$$d_E = \| \boldsymbol{X}_1 - \boldsymbol{X}_2 \| = \sqrt{(\boldsymbol{X}_1 - \boldsymbol{X}_2)'(\boldsymbol{X}_1 - \boldsymbol{X}_2)} \tag{4}$$

where $(\boldsymbol{X}_1 - \boldsymbol{X}_2)'$ denotes the transpose of $(\boldsymbol{X}_1 - \boldsymbol{X}_2)$.

The centroid $\underline{\mu}_i$ for each pattern class is calculated as follows:

$$\underline{\mu}^{(j)} = \frac{1}{N}\left[\boldsymbol{x}_1^{(j)}, \boldsymbol{x}_2^{(j)}, ..., \boldsymbol{x}_k^{(j)}\right]' \tag{5}$$

where j denotes the pattern class, k is the number of elements in the pattern vector (dimension of the class), and N is the number of pattern vectors (segments) in the class, which is the number of years (in the case of monthly data).

For calculating the similarity measure within the class, it might be argued that Mahalanobis distance can do a better job because it includes the class covariance matrix. Mahalanobis distance, which is shown in equation (6), differs from Euclidean by weighing each attribute of the vector according to its variance. The months of higher variance are given lower weight.

$$d_M = (\boldsymbol{X} - \underline{\mu})' \boldsymbol{C}^{-1} (\boldsymbol{X} - \underline{\mu}) \tag{6}$$

where $\boldsymbol{X}$ represents a pattern vector (segment), $\underline{\mu}$ is the mean vector, and $\boldsymbol{C}$ is the covariance matrix of a pattern population. In this segmentation process, the similarity of variances was considered, visually, in clustering the months together in groups. Therefore it is assumed that the differential effect of the monthly variances is insignificant within each class.

DIFFERENCE BETWEEN THE GROUP APPROACH AND EXISTING SEASONAL TIME SERIES ANALYSIS

It may be worthwhile to clarify the essential difference between the proposed group approach and the seasonal time series analysis models (e.g. periodic autoregressive moving average, PARMA model), especially when the word "season" may be used interchangeably with "group". The PARMA model has been proposed to overcome the problem of seasonality in the auto-correlation by developing a model with different parameters for each season (Salas, 1992). In that sense, there is a difference between the PARMA and the group approach. In the PARMA model, the word "season" represents one value rather than a vector, and adjacent observations are not considered as a group. For example, in order to apply the PARMA model to daily data where a season corresponds to one day, then 365 models are required. On the other hand, in the group approach, the number of models is significantly reduced (e.g. 52 models when groups corresponding to weeks are assumed). Therefore, the PARMA model is a member of the traditional single-valued approach.

PROPOSED GROUP-BASED TECHNIQUES AND MODELS FOR DATA IN-FILLING

Models of the time series analysis, regression, ANNs and MPM techniques are used to model the data groups and to estimate the missing segments. A description of each type of model follows.

Multivariate time series modelling

Multivariate analysis of time series started to receive attention as early as 1969 (Gnanadesikan, 1977). Since then, several multivariate models have been proposed in water resource literature. Different multivariate ARMA models are classified and

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

reviewed by Salas *et al.* (1985). The autoregressive model with constant parameters AR(p) (Salas *et al.*, 1980) is given below:

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + \ldots\ldots\ldots + \phi_p(y_{t-p} - \mu) + \varepsilon_t \quad (7)$$

where y_t is the time dependent variable, ε_t is the time independent series (error), which is independent of y_t, and normally distributed with mean zero and variance σ_ε^2, μ is the average value of the time series, and ϕ is the autoregressive coefficient. The group approach presented in this paper treats the single series as a multivariate problem. Therefore, the multivariate AR model can be the most appropriate form of AR model to be employed to handle data groups. The multivariate AR(1) model is given in equations (8) and (9):

$$\boldsymbol{Y}_t = \underline{\mu} + \underline{\sigma}\boldsymbol{Z}_t \quad (8)$$

$$\boldsymbol{Z}_t = \boldsymbol{A}_1 \boldsymbol{Z}_{t-1} + \boldsymbol{B}\underline{\varepsilon}_t \quad (9)$$

where $\underline{\mu}$ is a (k*1) matrix; $\underline{\sigma}$ is a (k*k) diagonal matrix; $\boldsymbol{Y}_t$, $\boldsymbol{Z}_t$, and $\boldsymbol{Z}_{t-1}$ are (k*1) matrices assuming that there are k series; and $\boldsymbol{A}$ and $\boldsymbol{B}$ are (k*k) parameter matrices. In the proposed group approach, a single time series with consecutive segments, each of which consists of k elements, is treated as k series. Figure 3 provides a schematic diagram for representing a single series of monthly flows using the group approach. Three segment types of four elements each are assumed for illustration. In this case, the problem of a single time series is viewed as similar to the case of a multivariate AR model with four time series. The data are arranged in a fashion that facilitates calculation of different correlation coefficients. This model is named the *autoregressive group* (ARG) model. In the same way of designing a periodic multivariate AR model (Salas *et al.*, 1980), a periodic ARG model (PARG) for each segment type is developed and applied in this study.

1	*Jan*	*May*	*Sep*	*Jan*	----------N
2	*Feb*	*Jun*	*Oct*	*Feb*	----------N
3	*Mar*	*Jul*	*Nov*	*Mar*	----------N
4	*Apr*	*Aug*	*Dec*	*Apr*	----------N

Fig. 3 Schematic diagram representing a single series in the group approach as a multivariate problem.

Multivariate partitioning model (MPM)

Assuming the case of single or multiple time series with two segment types (classes) $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \hline \boldsymbol{X}^{(2)} \end{bmatrix}$ be distributed as $N_p(\underline{\mu}, \boldsymbol{C})$ with $\underline{\mu} = \begin{bmatrix} \underline{\mu}^{(1)} \\ \hline \underline{\mu}^{(2)} \end{bmatrix}$, $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} \mid \boldsymbol{C}_{12} \\ \hline \boldsymbol{C}_{21} \mid \boldsymbol{C}_{22} \end{bmatrix}$, and $|\boldsymbol{C}_{22}| > 0$. Then the conditional distribution of $\boldsymbol{X}^{(1)}$, given

$\boldsymbol{X}^{(2)} = \boldsymbol{X}_2$, is normal with:

$$Mean = \underline{\mu}^{(1)} + \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\left(\boldsymbol{X}_2 - \underline{\mu}^{(2)}\right) \tag{10}$$

This result is given by Wichern & Johnson (1988) as a property of multivariate normal distribution. Note that the vertical and horizontal lines shown inside a matrix notation indicates that it is a partitioned matrix. The resultant mean of equation (10) is used as an estimate of the missing segment in these papers for both cases of single and bi-series. In the case of a single time series, $\boldsymbol{X}_1$ is considered the missing segment of the "target river" and $\boldsymbol{X}_2$ is the previous segment assuming lag-1 transition model. For the bi-series case, $\boldsymbol{X}_1$ is the missing segment of the "target river" and $\boldsymbol{X}_2$ is the concurrent segment of the "reference river".

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

Multivariate multiple regression (MMR)

Simple and multiple regression techniques are widely used in hydrology. In both cases of applications, the response variable is always a single dependent variable. The group approach requires a regression model that can handle multiple response variables as well as multiple predictors. The MMR is the statistical linear method for estimating values of one or more dependent variables from a collection of predictor (independent) variables (Johnson & Wichern, 1988). One such bi-variate MMR (multi-dimensional regression) model is presented as follows:

$$\begin{array}{ccccccc} \boldsymbol{Y} & = & \boldsymbol{Z} & & \underline{\beta} & + & \underline{\varepsilon} \\ (N*k) & & (N*(r+1)) & & ((r+1)*k) & & (N*k) \end{array} \tag{11}$$

where $\underline{\varepsilon}$ is the residual term, whose sum of squares and cross product is obtained as follows:

$$\underline{\hat{\varepsilon}}'\underline{\hat{\varepsilon}} = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} = \boldsymbol{Y}'\boldsymbol{Y} - \underline{\hat{\beta}}'\boldsymbol{Z}'\boldsymbol{Z}\underline{\hat{\beta}} \tag{12}$$

and k is the number of elements in each estimated response vectors $\boldsymbol{Y}$, N is the number of observations, $(r + 1)$ is the rank of the design matrix $\boldsymbol{Z}$, and the unknown parameter, β, is estimated as follows:

$$\underline{\hat{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \tag{13}$$

where $\boldsymbol{Y}$ is the estimated vector and has the following form:

$$\hat{\boldsymbol{Y}} = \boldsymbol{Z}\underline{\hat{\beta}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \tag{14}$$

The multivariate multiple regression equation is based on the assumptions that the expectation of errors is equal to zero, the errors are statistically independent, the variance of errors is constant, and the errors are normally distributed.

Artificial neural network (ANN) technique

ANN is a computing paradigm that may have various types of configurations employing different learning algorithms. The feed-forward neural nets with back

propagation (BP) learning algorithm are the most widely used neural networks (Freeman & Skapura, 1991). This study employs networks with three layers. The configuration of a neural network includes determining the number of hidden layers, the number of nodes in each of the hidden layers, and the connection weights. Details on the ANNs and the BP training algorithm (to estimate the network parameters) can be found in Freeman & Skapura (1991). It is worth mentioning that the ability of ANNs to establish a functional relationship between multiple inputs and outputs makes them a self-justified candidate for the group approach proposed in this study.

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

SUMMARY

A different approach from the traditional single-valued method of estimating missing values has been proposed and presented in this paper. This "group approach" is based on perceiving the hydrological time series as a sequence of groups of values rather than single-valued observations. The pertinent literature has been reviewed, and the methodology adopted to employ the group approach has been outlined. It has been shown that the concept of data grouping is recognized, both implicitly and explicitly, in the hydrological literature. For simplicity, in some cases it has been avoided after recognition. Proximity measures, from the pattern recognition domain, have been presented as representatives of the unsupervised learning technique to divide the data into segments and classes. Models representing four different techniques for modelling the streamflow data as groups have been briefly described. The multivariate AR(1) model has been manipulated to treat univariate time series according to the group approach. Accordingly, the autoregressive group ARG(1) and the periodic autoregressive group PARG(1) models have been developed. Techniques capable of treating data vectors, such as multivariate partitioning modelling (MPM) and multivariate multiple regression (MMR), have been employed to model the data groups. Artificial neural networks, with their flexibility in handling groups and single-valued observations, have been used for comparison because of their previously established utility in water resource applications.

Application of the segmentation process and implementation of different models along with comparisons regarding their relative performance are presented in a companion paper (Elshorbagy *et al.*, 2000).

Acknowledgement This research has been supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Alley, W. M. & Burns, A. W. (1983) Mixed-station extension of monthly streamflow records. *J. Hydraul. Engng ASCE* **109**(10), 1272–1284.

Beale, E. M. L. & Little, R. J. A. (1975) Missing values in multivariate analysis. *J. Roy. Statist. Soc. B.* **37**(1), 129–145.

Beauchamp, J. J., Downing, D. J. & Railsback, S. F. (1989) Comparison of regression and time series methods for synthesizing missing streamflow records. *Wat. Resour. Bull.* **25**(5), 961–975.

Bennis, S., Berrada, F. & Kang, N. (1997) Improving single-variable and multivariable techniques for estimating missing hydrological data. *J. Hydrol.* **191**, 87–105.

Ben-Zvi, M. & Kesler, S. (1986) Spatial approach to estimation of missing data. *J. Hydrol.* **88**, 69–78.

Berkowitz, B., Ben-Zvi, M. & Berkowitz, J. (1992) A spatial time-dependent approach to estimation of hydrologic data. *J. Hydrol.* **135**, 133–142.

Booy, C. & Morgan, D. R. (1985) The effect of clustering of flood peaks on a flood risk analysis for the Red River. *Can. J. Civil Engng* **12**, 150–165.

Dawson, C. W. & Wilby, R. (1998) An artificial neural network approach to rainfall–runoff modelling. *Hydrol. Sci. J.* **43**(1), 47–66.

Dax, A. (1985) Completing missing groundwater observations by interpolation. *J. Hydrol.* **81**, 375–399.

Dubes, R. C. (1987) How many clusters are best? An experiment. *Pattern Recognition* **20**(6), 645–663.

Duda, R. & Hart, P. (1973) *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, USA.

Elshorbagy, A., Panu, U. S. & Simonovic, S. P. (1999) Investigations into group-based data in-filling techniques. *Proc. Annual Conf. Can. Soc. Civ. Engng* (Regina, Saskatchewan, 2–5 June), 337–348.

Elshorbagy, A., Panu, U. S. & Simonovic, S. P. (2000) Group-based estimation of missing hydrological data: II. Application to streamflows. *Hydrol. Sci. J.* **45**(6), 867–880 (this issue).

Fiering, M. B. (1962) On the use of correlation to augment data. *J. Am. Statist. Assoc.* **57**(297), 20–32.

Freeman, J. A. & Skapura, D. M. (1991) *Neural Networks. Algorithms, Applications, and Programming Techniques*. Addison-Wesley, Massachusetts, USA.

French, M. N., Krajewski, W. F. & Cuykendall, R. R. (1992) Rainfall forecasting in space and time using a neural network. *J. Hydrol.* **137**, 1–31.

Gilroy, E. J. (1971) Reliability of a variance estimate obtained from a sample augmented by multivariate regression. *Wat. Resour. Res.* **6**(6), 1595–1600.

Gnanadesikan, R. (1977) *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley & Sons, Inc., USA.

Griffith, D. A., Haining, R. P. & Bennett, R. J. (1985) Estimating missing values in space-time data series. In: *Time Series Analysis: Theory and Practice* (ed. by O. D. Anderson, J. K. Ord & E. A. Robinson), 6. Elsevier Science Publishers BV, Amsterdam, The Netherlands.

Gupta, A. & Lam, M. S. (1996) Estimating missing values using neural networks. *J. Oper. Res. Soc.* **47**(2), 229–238.

Gyau-Boakye, P. & Schultz, G. A. (1994) Filling gaps in runoff time series in West Africa. *Hydrol. Sci. J.* **39**(6), 621–636.

Hirsch, R. M. (1979) An evaluation of some record reconstruction techniques. *Wat. Resour. Res.* **15**(6), 1781–1790.

Hirsch, R. M. (1982) A comparison of four streamflow record extension techniques. *Wat. Resour. Res.* **18**(4), 1081–1088.

Hsu, K. L., Gupta, H. V. & Sorooshian, S. (1995) Artificial neural network modeling of the rainfall–runoff process. *Wat. Resour. Res.* **31**(10), 2517–2530.

Hughes, D. A. & Smakhtin, V. (1996) Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. *Hydrol. Sci. J.* **41**(6), 851–871.

Hurst, H. E. (1951) Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civil Engrs* **16**, 770–808.

Hurst, H. E. (1956) Methods of using long-term storage in reservoirs. *Proc. Instn Civil Engrs* **1**, 519–543.

Johnson, R. A. & Wichern, D. W. (1988) *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey, USA.

Khalil, M., Panu, U. & Lennox, W. (1998) Estimation of missing streamflows: a historical perspective. *Proc. Annual Conf. Can. Soc. Civil Engrs* (Halifax, Nova Scotia, 10–13 June), 235–246.

Knotters, M. & Van Walsum, P. E. V. (1997) Estimating fluctuation quantities from time series of water-table depths using models with a stochastic component. *J. Hydrol.* **197**, 25–46.

Kuligowski, R. J. & Barros, A. P. (1998) Using artificial neural networks to estimate missing rainfall data. *AWRA* **34**(6), 1437–1447.

Lettenmaier, D. P. (1980) Intervention analysis with missing data. *Wat. Resour. Res.* **16**(1), 159–171.

Makhuvha, T., Pegram, G., Sparks, R. & Zucchini, W. (1997) Patching rainfall data using regression methods. 2. Comparisons of accuracy, bias and efficiency. *J. Hydrol.* **198**, 308–318.

Moran, M. A. (1974) On estimators obtained from a sample augmented by multiple regression. *Wat. Resour. Res.* **10**(1), 81–85.

Mott, P., Sammis, T. W. & Southward, G. M. (1994) Climate data estimation using climate information from surrounding climate stations. *Appl. Engng in Agric.* **10**(1), 41–44.

Panu, U. S., Unny, T. E. & Ragade, R. K. (1978) A feature prediction model in synthetic hydrology based on concepts of pattern recognition. *Wat. Resour. Res.* **14**(2), 335–344.

Panu, U. S. & Unny, T. E. (1980) Stochastic synthesis of hydrologic data based on concepts of pattern recognition, I. General methodology of the approach. *J. Hydrol.* **46**, 5–34.

Panu, U. S. (1992) Application of some entropic measures in hydrologic data infilling procedures. In: *Entropy and Energy Dissipation in Water Resources* (ed. by V. P. Singh & M. Fiorentino), 175–192. Kluwer, Dordrecht, The Netherlands.

Panu, U. S., Khalil, M. & Elshorbagy, A. (2000) Chapter 12 in: *Artificial Neural Networks in Hydrology* (ed. by R. S. Govindraju), 235–258. Kluwer, Dordrecht, The Netherlands.

Pegram, G. (1997) Patching rainfall data using regression methods. 3. Grouping, patching and outlier detection. *J. Hydrol.* **198**, 319–334.

Raman, H. & Sunilkumar, N. (1995) Multivariate modeling of water resources time series using artificial neural networks. *Hydrol. Sci. J.* **40**(2), 145–163.

Salas, J. D., Delleur, J. W., Yevjevich, V. & Lane, W. L. (1980) *Applied Modelling of Hydrologic Time Series*. Water Resources Publications, Littleton, Colorado, USA.

Salas, J. D., Tabios III, G. Q. & Bartolini, P. (1985) Approaches to multivariate modelling of water resources time series. *Wat. Resour. Bull.* **21**(4), 683–708.

Salas, J. D. (1992) Analysis and modeling of hydrologic time series. In: *Handbook of Hydrology* (ed. by D. R. Maidment), 19.1–19.72. McGraw-Hill, New York, USA.

See, L. & Openshaw, S. (1999) Applying soft computing approaches to river level forecasting. *Hydrol. Sci. J.* **44**(5), 763–778.

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014

Shvarster, L., Shamir, U. & Feldman, M. (1993) Forecasting hourly water demands by pattern recognition approach. *J. Wat. Resour. Plan. Manage. ASCE* **119**(6), 611–627.

Simonovic, S. P. (1995) Synthesizing missing streamflow records on several Manitoba streams using multiple nonlinear standardized correlation analysis. *Hydrol. Sci. J.* **40**(2), 183–203.

Tou, J. T. & Gonzalez, R. C. (1974) *Pattern Recognition Principles*. Addison-Wesley, Massachusetts, USA.

Valencia, D. R. & Schaake, J. C. (1973) Disaggregation processes in stochastic hydrology. *Wat. Resour. Res.* **9**(3), 580–585.

Van Geer, F. C. & Zuur, A. F. (1997) An extension of Box-Jenkins transfer/noise models for spatial interpolation of groundwater head series. *J. Hydrol.* **192**, 65–80.

Vogel, R. M. & Stedinger, J. R. (1985) Minimum variance streamflow record augmentation procedures. *Wat. Resour. Res.* **21**(5), 715–723.

Zealand, C. M., Burn, D. H. & Simonovic, S. P. (1999) Short term streamflow forecasting using artificial neural networks. *J. Hydrol.* **214**, 32–48.

Zhang, T. & Berndtsson, R. (1991) Analysis of soil water dynamics in time and space by use of pattern recognition. *Wat. Resour. Res.* **27**(7), 1623–1636.

Received 1 November 1999; accepted 26 June 2000

Downloaded by [Montana State University Bozeman] at 09:44 14 August 2014