CHAPTER 1

GENERAL INTRODUCTION

1.0 INTRODUCTION

For planning, management and effective control of water resources systems, a considerable amount of data on numerous hydrologic variables such as rainfall, streamflow, evaporation, temperature, etc. are required. These data stimulate the outcome of the water managers' decisions.

Very often in developing countries, hydrological data sequences at a given network have gaps (because of missing values) or are incomplete, or are not of good quality or are not of sufficient length. Although developed countries may encounter hydrological data problems with existing networks (Harmancioglu et al., 1999), it seems that this situation is more prevalent in developing countries than it is in developed countries. This can severely affect the reliability of the design, e.g. hydropower plant, construction of water storage, etc. (Akiri, 1992). For example, in South Africa, the overwhelming majority of gaps are caused by temporary absence of observers, the cessation of measurement or absence of observations prior to the commencement of measurement (Makhuvha et al. 1997a, 1997b). In Bolivia for example, due to the limited financial resources, even a minimum national network could not be achieved according to the Meteorological network density ration (Balek, 1992). Gaps are also due to the political instability, i.e. everlasting war in the Democratic Republic of Congo is only one case among others. These examples hold also for developing countries in general. Thus the need of appropriate data (e.g. rainfall, streamflow) infilling techniques is emphasized. Numerous data infilling (interpolation) techniques have evolved in hydrology to deal with missing data.

The present study develops merely a methodology by combining different approaches and coping with missing data using the theories of entropy, artificial neural network (ANNs) and expectation-maximization (EM) algorithms. None of the studies undertaken in hydrology has taken the opportunity of combining the above-mentioned three concepts by applying them to hydrological data infilling (interpolation) problems. In this study, the data infilling techniques viz. feedforward ANNs and EM (e.g. some of the existing techniques applied and not yet applied in hydrology) and their new features are also presented.

The entropy concept is used, in this research work, as a versatile tool. Firstly, the entropy concept is used to quantify the information content of hydrological variables (e.g. rainfall or streamflow). Secondly, the entropy concept (through the directional information transfer index, i.e. DIT) defined rigorously both selection of control (base)/target (subject) gauging stations. Finally, the DIT notion is extended to the assessment of hydrological data infilling technique performance. Thus, the validity of the data infilling techniques with respect to the different gap durations is defined through the entropy concept. A performance comparison between the different data interpolation techniques is achieved through the directional information transfer index. A case study was then conducted from rainfall stations and streamflow gauges of selected catchment areas of the primary drainage region D “Orange” of South Africa. These selected catchments belong to the secondary drainage systems D33 and D1 respectively.

The methodology was then tested on annual mean flows, annual maximum flows, 6-month flow series (means) and annual total rainfall for selected gauges. These data regimes can be regarded as useful for design-oriented studies, flood studies, water balance studies, etc. The performance of the different data infilling techniques was assessed through entropy concept. A relationship between gap size of missing values at the target gauge and accuracy, i.e. in terms of DIT, of the estimated series was investigated. It was also possible, through entropy concept, to assess runoff simulation models for areas with non-observed data at a target site.

1.1 PROBLEM STATEMENT

The problems of missing data, insufficient length of hydrological data series and poor quality are common in developing countries (Akiri, 1992; Feldman, 1992; Medeiros et al., 2002; Singh, 1998 a, 1998c). This situation can severely affect the outcome of the water systems managers' decisions (e.g. reliability of the design). Thus numerous data infilling (interpolation) techniques have evolved in various scientific disciplines to deal with the missing data.

Traditionally, statistical regression methods are the most commonly used for information transfer between two or more gauging stations and are subsequently used for filling in the gaps. In regression methods, the variance could be considered as a measure of information content of a given station and the cross-correlation coefficient is used to measure the information at nearby sites. The cross covariance matrix is used to examine the space dependency between different hydrological variables. However, regression approaches transfer of information on the basis of certain assumptions regarding the distributions of variables and the form of the transfer function such as linearity and non-linearity. The correlation coefficient cannot take care of arbitrary relation between the coordinates and classes. In addition, these measurements of information could suffer where information could be insufficient (thus missing data) i.e. cases of developing countries. Sometimes, the control station and target station are chosen arbitrarily. This led some researchers to seek other measures to quantify conveniently the information content transfer from one gauging station to another gauge by using the entropy approach (i.e. mutual information or directional information transfer index) as a *versatile tool where information is limited or incomplete*. The entropy approach was also suggested to assess hydrological model performance. It should be worth noting that entropy notion does not give any means of transferring information but it only measures whether all the information is transferable via a model (e.g., linear, non-linear, etc). Also, entropy within the context of hydrological data infilling (interpolation) has not been thoroughly exploited yet. However, among the regression methods, attention is drawn to particular techniques; i.e. expectation maximization (EM) techniques, which have been intensively used in the last two decades and is *reputed to cope with missing data*. These techniques were also used to map the relationship between input data and output data.

From the literature survey, it has been found that Mahkhuva (1997a, 1997b) and Kuczera (1987) applied EM techniques in hydrology for incomplete data cases. However, the application of these techniques in hydrology or water resources related fields remains very space. Beside the EM techniques, the artificial neural networks have been also used to map (learn) the relationship, through multi-layers, between the input and output. The ANNs are thought also to *cope with problems where information is limited or incomplete*. However, the literature on ANNs dealing with missing data is very sparse apart from the studies led by Panu et al. (2000), Khalil et al. (2001), Eshorbagy (2000a, 2000b).

It's now well understood that all the three concepts; i.e. entropy, EM and ANNs are reputedly known to cope with problems dealing with *missing data (little information)*. The current study appears to have taken the opportunity of combining these three concepts for specifically infilling (interpolation) data problems in hydrology. The present study is undertaken to develop a merely methodology where the abovementioned three concepts (accepting missing data) are used together with the choice of the appropriate technique (s) viz. EM or ANN for filling in hydrological data. The entropy approach (i.e. mutual information, directional information transfer index then introduced here explicitly) is compared with other criteria and used in the evaluation of data interpolation technique accuracy.

1.2 OBJECTIVES

The main objective of this study was to develop a simple methodology where the abovementioned three concepts (accepting little information/missing data) are used together with the choice of the appropriate technique (EM or ANNs) for interpolating (filling in) hydrological data. The methodology was translated into a model that took the name of ENANNEX as it encompasses the three concepts, viz. Entropy, Artificial Neural Networks and Expectation maximization.

Other objectives were:

- Determination of the information (uncertainty) contained in different sites (e.g. rainfall / streamflow gauges);
- Evaluation of information inferred from one station to the other (within station-pairs);
- Determination of dependency between stations within station pairs. Thus, the control (base) station and target (subject) station are known.
- In addition to EM and ANN techniques, introducing of modifications (which are essentially intuitive) in these existing techniques is also achieved.
- Estimation (interpolation) of the missing data;
- Assessment of the performance of different data infilling techniques (techniques selection). Thus, the amount of uncertainty removed from the target station (gauge) in a station-pair, via the different techniques, is determined.
- Investigation of the relationship between gap duration and accuracy of estimated (interpolated) data series, when applying the different techniques.
- Comparison of simulation model performance for areas with non-existing data.

1.3 LIMITATIONS AND SPECIFICATIONS

Only research works closely to the topic of this thesis will be reviewed to give a general view of the available literature on entropy concept and missing data interpolation (estimation or infilling) techniques, i.e. expectation-maximization and artificial neural network techniques. This study does not deal with data *generation* techniques nor data extension of short data but with data *interpolation* (infilling) techniques where the overriding objective is to minimize the difference between the observed and estimated data. However, the statistical requirements for the estimated (interpolated) series should be fulfilled, i.e. the variance and the mean before and after interpolation should be maintained. For example, Simonovic (1995) used the term interpolation. The interpolation process could be used to fill in the missing data (at a gauging site where some records do exist but the sequence has been interrupted) with the uninterrupted sequences of nearby sites. In this study, the computational efficiency for the different techniques is not the primary objective in the missing estimation process: it is a difficult concept to quantify absolutely because so much depends on clever coding and fast algorithms.

Time series methods (AR, ARIMA, etc) for data infilling purposes are not part of the current study. Also, rainfall-runoff models which can be used for extending or/and infilling runoff data (e.g. Kachroo, 1992 a, 1992 b; Bennis et al. 1997) are not thoroughly discussed here.

In contrast with physical hydrology, the model developed later on is referred to as a systems investigation model, which is regarded as being concerned with problems subject only to the constraints imposed by the available data (i.e. rainfall and streamflow) and not subject to “physical considerations” (Minns and Hall, 1996). The notion of transferability of information is applied to “similar” station pairs; i.e. a rainfall station with its nearby stations and a streamflow gauge with its neighboring gauges, unlikely typical rainfall-runoff modeling. Missing hydrological data will be considered as observable and occurring at random, thus the unobservable missing data will be ignored.

The term “traditional” statistical regression method is used only to make a difference with EM techniques (or entropy concept) due essentially to the iterative procedure of EM techniques (or the unique characteristic of entropy in theoretic information). However, EM techniques can be seen within the context of regression methods for the purpose of data infilling (interpolation). This study deals only with EM techniques as maximum likelihood approach. Other approaches such as raw maximum likelihood and multiple imputation techniques are not covered here.

Sometimes, the words: technique, algorithm and method will be used interchangeably. The same applies to the words interpolation, estimation and infilling. Only feedforward ANNs will be used in this thesis.

In this research work, the term “hydrological” data is more used than “hydrologic” data. Pegram (1985) used also “hydrological” data. However, both terms are interchangeable. Term like “best technique(s)” can be used for techniques found suitable for data interpolation according to the DIT notion. Thus, the ranking of techniques can be made according to, for example, the “first best”, “second best”, etc.

1. 3 LAYOUT OF THE THESIS

The rest of the thesis will be organized as follows:

Chapter 2 gives a detailed theory on entropy approach, artificial neural networks and expectation-maximization techniques. It also gives an overview of missing data in general and in developing countries. The information measures (transfer) are given according to both traditional the statistical methods and the entropy concept and a comparison is made. Data interpolation (infilling) techniques are briefly described, e.g. ANNs and EM techniques. This chapter presents also the different criteria (statistical, graphical and entropic) for model performance assessment.

Chapter 3 presents the research methodology. This includes merely a description of the model development, which translates the methodology. This model incorporates the measure of information and its transferability according to the entropy approach. It includes also some exiting hydrological data infilling techniques (applied and not applied in hydrology) as well as their new features. Data infilling technique criteria are also incorporated into the model. The data availability is also given here.

Chapter 4 gives the testing of the methodology to selected catchments of the Orange River system (specifically the secondary drainage D1) of South Africa: annual mean flows are used. The results, the analysis and the discussion are also presented. The performance of different streamflow data infilling techniques is discussed.

Chapter 5 gives the testing of the methodology to selected catchments of Orange River system (specifically the secondary drainage D1) of South Africa: annual maximum flows for gauges considered in the previous chapter are used. The results, the analysis and the discussion are also presented. The performance of different streamflow data infilling techniques is discussed.

Chapter 6 gives the testing of the methodology to selected catchments of Orange River system (specifically the secondary drainage D33) of South Africa: annual total rainfall series are used. The results, the analysis and the discussion are also presented. The performance of the different rainfall data interpolation (infilling) techniques is discussed.

Chapter 7 gives an assessment of runoff simulation models (i.e. RAFLER and WRSM2000 models) using entropy approach. This approach is applied to a site where non-observed data is available.

In Chapter 8, the BP technique is performed by approximating the sigmoid function by pseudo Mac Laurin power series order 1 and 2 derivatives. The impact of this approximation on the accuracy of the estimated values is investigated through entropy approach. The data regime used here is 6-month flow series (means) or seasonal mean flows of selected gauges from the secondary drainage D1.

Chapter 9 is devoted to summary, conclusions and recommendations.

The list of references and the appendices finalize the thesis.