CHAPTER 2

LITERATURE SURVEY

2.0 INTRODUCTION

This chapter starts with a brief overview of the problem of missing data in developing countries. Then, the concepts of entropy, EM and ANN techniques are presented.

2.1 MISSING DATA AND DEVELOPING COUNTRIES

The existence of adequate, accurate and timely data is a vital element for supporting development planning, implementation and program monitoring (Sadowsky, 1989). In the area of water resources planning and management, complete data sets are required on many variables such as rainfall, streamflow, evapotranspiration and temperature, etc. Unfortunately, records of hydrological processes are usually short and often have missing observations (Feldman, 1972). The inadequacy of hydrological data; i.e. observation error, shortness of the sample size, etc can severely affect the reliability of the design (Akiri, 1972). Harmancioglu et al. (1999) stipulated that developed countries could sometime suffer from what he calls “data-rich but information-poor” networks. However, the problem of missing data becomes sometime even worse in developing countries or Third World.

The existence of gaps might be attributed to a number of factors such as interruption of measurement because of equipment failure, effects of extreme natural phenomena such as hurricanes or landslides of human-included factors such as wars and civil unrest, mishandling of observed data by field personnel, or accidental loss of data files in the computer system (Elshorbagy et al., 2000a) and most of the old data for developing countries were lost due to an inexistent database storage (Medeiros, et al., 2002). In South Africa, for example, 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, due to the limited financial resources, even a minimum national network could not be achieved according to the meteorological network density ration (Balek, 1972).

Gaps are also due to the political instability, i.e. everlasting war in the Democratic Republic of Congo is only one case among others. In the context of the Zambezi basin, Balek (1972) pointed out that one of the basic problems is the collection of all existing data; the basin's boundaries are not identical with political boundaries, thus precipitations are of short length (i.e. 7 years). The author proposed that the non-uniformity of the observing periods have to be eliminated for the data to be used.

Developing countries generally lagged in the use of new technologies to process their statistical data (Sadowsky, 1989). Yet the needs are just great; they need to achieve a viable statistical data processing capability if they are to provide, on a continuous and sustained basis, the essential statistical information needed for development planning and administration in their countries (Sadowsky, 1989). This author gives some special problems impeding development in statistical data processing and data banks for developing countries: physical infrastructure, human resources infrastructure, financial poverty, technology transfer assistance. For more details, the reader is referred to that author.

In Brasilia, the information from the CERB institution came out from the water supply systems implanted by the referred institution: for most wells, waters were analyzed just once during the evaluation of the supply system implementation (Medeiros, et al., 2002). Another analysis from an identical system happens only if a requested repair is required by the same manager, due to the low frequency of analysis; not all sampling points have coordinates and many wells are abandoned with their water pumps being stolen (Medeiros, et al., 2002).

It is very rare from the literature survey to have a fixed range for the missing hydrological data proportion such that a data series can used. However, attempts were made in some cases and that depends on the case at hand. Deficiencies in hydrological data series vary from 5 % to 10 % in the case of runoff data and up to 25 % in the case of oceanic storm surges (Panu et al., 2000). However, Gyau-Boakye (1994) mentioned that the daily runoff records (in Ghana) considered for the Nabogo basin and the Ayensu basin (from 1962 to 1980) and for the Tano basin (from 1962 to 1980) contained respectively 9.04%, 8.30% and 13. 94% of missing data. These missing data were assumed randomly distributed over the entire record ranging from one day to a year. This assumption was also made by Makhuvha (1997a, 1997b), on a case study for rainfall patching in South Africa. Zucchini et al. (1984), in his study for rainfall patching in South Africa, pointed out that so long as the number of estimated values is small relative to the length of the target record the bias in the estimated values will be negligible, but this is not the case if a large proportion of values is estimated. Thus, it becomes difficult to determine what proportion of missing values alternative methods should be used because this depends on a number of factors such as the multiple-correlation between the selected control records and the target record.

Zucchini et, al.(1984) gave a very rough rule: no more than 25 % of the target rainfall record should be estimated using the regression methods (that he proposed), unless the multiple correlation is greater than about 0.90 in which case the target record can be extended without introducing much bias. Midgley et al., (1994) gave a report which contained the revised of appraisal of the surface water of South Africa where for example the proportion of monthly rainfall patched for stations used in drainage “D”and “F” varies from 0.6 % to 65.2%. It should be noted that 69 years usable for 65.2 % (Rondawel station) data patched and 25 years usable for 0.6% patched (Bundu station). Ilunga (2002a) used rainfall data whose proportion of missing was in the range between 1.68 % and 46.21 %, for the intervening catchments between Mtera reservoir and Kidatu reservoir (in the Rufiji of Tanzania). Makhuvha (1997b) compared EM and PEM (that he called Pseudo-EM) by conducting a Monte Carlo simulation where various proportions

of one site data (rainfall) were hidden purposely; these proportions of artificial missing data varied from 10 to 20 %. Ilunga and Stephenson (2002b) used 20 % of artificial missing data at a potential target station (of the Orange River system) for assessing hydrological data infilling techniques using entropy approach.

Most hydrological models do not tolerate missing observations and thus, data interpolation (infilling) techniques have evolved to deal with incomplete data sets. The problem of missing data estimation in hydrology has been treated by several authors among others Elshorbagy et al. (2000a, 2000b); Panu et al. (2000), Elshorbagy et al. (2001), Gyau-Boakye (1994), Hirsch (1979, 1982). The only work for which the methodology was tested specifically on developing countries is Gyau-Boakye's (1994). The present study develops merely a methodology based mainly on three concepts reputedly known as dealing with missing data viz; Entropy, ANNs and EM. The methodology was tested on selected catchments in South Africa.

2.2 HYDROLOGICAL INFORMATION

2.2.1 Value of information in hydrology

2.2.1.1 Types of data

Information as the usefulness of data is more important in any discipline. Yevyevich (1972) gives different types of data: (a) historic data or chronological data, or observations of processes in time; (b) the field data observations along lines, or observations hydrologic phenomena across areas or space; (c) the third is laboratory and field experimental data related to hydrology acquired by methods similar to data obtained in hydraulic and (d) the fourth type is the simultaneous measurements of two or more random variables in order to establish a relationship among these variables, mainly for the purpose of transferring statistical information among variables.

A distinction is often useful and necessary between a true, virgin and observed value of any hydrologic variable; the true value of any observation is never known because of the data obtained, through inevitable errors in observation, are not

exact values. Hence, streamflow processes are considered to be stochastic process because of the natural, or inherent, randomness apparent in the observed streamflow traces (Vicens et al., 1975). The virgin value is the value produced by unchanged conditions of an environment; data are or either unpredictable natural or man-made significant changes in hydrologic environments. The observed value is available as the result of various surveys, recordings, or experiments; this value is usually published hydrologic services. For example precipitation is normally the most variable hydrological element over a territory, and its characterization is most commonly needed for water balance studies and for floods forecasting (Rodriguez-Itube and Meija, 1974).

2.2.1.2 Levels of information

There is always an amount of uncertainty associated with data that engineers and planners have to use for water resources problems and it is this uncertainty that causes the questions of how much information is enough and what kind of data one needs deal within real-life problems. The answer will always depend on the particular objectives that are being pursued, and this why is it is so difficult to provide for example guidelines for the design of data collection programs (Rodriguez-Itube and Meija, 1974).

The U.S. Office of Water Data Coordination has defined three levels of information concerning the network design (Rodriguez-Itube and Meija, 1974). The level 1 is to provide a base level of information for wide regional or national planning to be used for resource inventory and as background information for the design of more intensive and specific network systems. The level 2 concerns networks called to provide general water resources planning data, and level 3 is restricted to data collection programs for specific planning and managing activities. Levels 1 and 2 are called to provide regional estimation type of data. The level of information, on the other hand is connected with accurate data of local as well as a regional nature, which are gathered and analyzed for use of with a specific system design.

Klemes (1977) examined the value of information in optimization of storage reservoir operation (e.g. search for optimal release rules). In this context, it should be noted that the term "information level" as used implies both the adoption of a set of assumptions about the structure of the input (annual flows) process and the availability of a set of sample of certain number of processed parameters. For example in level 1, the input, which is supposed to comprise the complete of hydrologic information is being considered stationary, ergotic and purely random process having finite mean and variance. Nothing is made about the marginal type of its distribution. In level 2, the input is defined as in level 1, but its distribution is assumed to be normal; thus a wrong distribution is being used and fitted by method of moments. The available information is given in terms of the mean and the variance. In level 3, the input is defined as in level 2, but its marginal distribution is assumed to be lognormal. In this case, this distribution was shown to be the correct one and is fitted with the aid of the normal model applied to the log transformed of input variable. However, in the absence of information about distribution of floods and the economic losses associated with the design of flood reduction measures, it could be shown that the use of normal distribution to represent the distribution of floods was generally better than either: Gumbel, log-normal or Weibull distribution (Slack et al., 1975).

2.2.2 Information measures of hydrological variables

2.2.2.1 Traditional statistical methods

The term *"traditional"* is only used here to make a difference between the usual or current statistical methods and the entropy concept.

Traditionally, the hydrological variable information content (e.g. rainfall, streamflow, etc.) is mostly measured by the variance; the higher the variance the greater the measurement error of the variable. The variance of a given set of data gives a measure of the variability of the data with respect to the mean (Yevjevich, 1972); for example more gauging stations will be needed (Krstanovic and Singh, 1992a).

Harmancioglu, et al. (1994) and Krstanovic and Singh (1992a) used the term current techniques to mean traditional statistical methods. As an example, in the analysis of empirical data, the variance has been often interpreted as measure of uncertainty and as revealing gain or loss of information (Singh, 1998c).

Another measure of information is the cross-correlation, amongst records (e.g. rainfall) at nearby sites (Krstanovic and Singh, 1992a; Chow, 1964). At this stage, the cross-covariance matrix helps generally examine the space dependency between hydrological variables while the auto-covariance matrix will therefore determine the time dependency (Krstanovic and Singh, 1992b).

Generally the majority of the current techniques is based on the classic correlation and regression theory, which basically constitutes a means of transferring information in space and time (Krstanovic and Singh, 1992a; Yevjevich, 1972; Harmancioglu and Yevjevich, 1987). The use of regression theory in transfer of information has some justification; 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 (Harmanciogu, et al. 1994). Thus, how much information is transferred by regression under specific assumptions has to be evaluated with respect to the amount of information that is actually transferable. On may refer to Harmancioglu et al. (1987) for the definition of terms “transferred information” and “transferable information”. The correlation coefficient cannot take care of arbitrarily relation between coordinates and classes (Battiti, 1994).

The traditional methods suffer also where information is insufficient (missing data), e.g. case of most developing countries. For example, the variance is not the appropriate measure of uncertainty (information content) if the sample size is small (Singh, 1998c). Some time both control and target stations are chosen arbitrarily in the regression analysis (French et al., 1992). The major difficult associated with these current methods (e.g. in network design) is related to the lack of precise definition for information. They either do not give a precise

definition of how information is measured, or they try to express it intuitively in terms of other statistical parameters like standard error or variance. Although current methods stress the distinction between data and information, a direct link between them has not yet been established (Harmancioglu et al., 1994).

Harmancioglu et al. (1994) summarizes the shortcomings of the current (traditional) methods within the context of water quality network design (nonetheless, this can extended to other fields): (a) a precise definition of "information" contained in the data and how it is measured is not given; (b) the value of data is not precisely defined, and consequently; existing networks are not "optimal" either in terms of information contained in these data or in terms of getting the data; (c) the method of information transfer in space and time is restrictive; (d) cost–effectiveness is not emphasized in certain aspect of monitoring; (e) the flexibility of the network in responding to new monitoring objectives and conditions is not measured and not generally considered in the evaluation of existing or proposed networks.

Since 70's hydrologists tried to find another way of applying theoretic information (entropy concept) as information measure, which was used to alleviate many of the above shortcomings of the existing network design methods. The entropy concept was applied to many disciplines such as water quality modeling (Singh, 1998a), in rainfall network design (Krstanovic and Singh, 1992 a and 1992b), in river flow network design (Yang and Burn, 1994), in water quality monitoring design (Harmancioglu et al.1994, 1999). The entropy concept was then applied to many other fields, among others; water resources (Singh and Florentino, 1992; Amorocho and Espildora, 1973), in sediment yield calculation (Singh and Krstanovic, 1987), in flood frequency analysis (Singh, 1988; Sonuga, 1972 and 1976; Jowitt, 1979), in streamflow forecasting (Krstanovic and Singh, 1991), in hydraulics (Chiu, 1987), in groundwater resources management and planning in developing countries (Mogheir and Singh 2002), in environmental and water resources (Singh, 1998c), etc.

2.2.2.2 Entropy concepts

2.2.2.2.1 Preamble

Entropy originated from physics. In 1872, Boltzman defined entropy as a measure of the degree of ignorance as to the true state of a thermodynamic system. Thus, he defined entropy a mathematical quantity sometimes described as disorder. Since the pioneering work of Shannon and Weaver (1949), much attention has been focused on the use of entropy and energy dissipation rate relationships in environmental and water resources engineering. Entropy can be also considered as a measure of the degree of uncertainty or disorder associated with a system. These features have been mathematically formulated in the theory of entropy by Shannon and Weaver (1949) and the principle of maximum entropy (POME) by Jaynes in 1957. This is also repeated in Jaynes (1982). Since then, entropy concepts could find a wide range of application in hydrology and water resources as mentioned in the previous section.

Engineering decisions are made frequently with less than perfect information. Such decisions may often be based on experience, professional judgment, rules of thumb, safety factors or probabilistic methods. Although probabilistic methods allow for more explicit and quantitative accounting of uncertainty, their major difficulty stems from the availability of limited data. The entropy concept enables determination of the least biased probability distributions with limited data. Singh (1998c) recommended the application of entropy to developing countries as they suffer very often from insufficient data. The entropy concept does not assume the variables to be normal unlike in the classic correlation coefficient (Chapman, 1985).

Generally, entropy can be viewed in three different but related contents and is hence typified by three forms: Thermo-dynamical entropy, statistical-mechanical entropy, and information-theoretic entropy. In water resources (hydrology), the most frequently used form is the information theoretic entropy by Shannon and Weaver (1949); thus it has got a great appeal in this field. Singh and Florentino

(1992) established an analogy between the hydrologic system and the thermodynamic system.

2.2.2.2.2. Formulation of entropy in hydrology

According to the entropy concept as defined in communication (or information) theory, the term “information content “ refers to the capability of signals to create communication. The problem is the generation of correct communication by sending a sufficient amount of signals, leading neither to any loss nor to repetition of information. Application of engineering principles to the problem of data collection calls for a minimum number of signals to be received to obtain the maximum amount of information. Redundant information does not help, reduce the uncertainty further; it only increases the costs of obtaining data. In the case of redundant information for example an existing monitoring network should be reduced and in the case of shortage of information, the existing network should be expanded (Mogheir and Singh, 2002). These considerations represent the essence of the field of communications and therefore hold equally true for hydrologic data sampling, which is essentially communicating with the natural system. Since the reduction of the uncertainty by means of making observations is equal to the amount of information gained, the entropy criterion indirectly measures the information content of a given series of data (Harmancioglu and Yevjevich, 1987).

The early works of theoritic entropy in hydrology dated around 1970’s (Sonuga 1972,1976; Amorocho and Espildora, 1973).

In information theory, the definition of entropy can be traced in the following argument (Shannon and Weaver, 1949). It is in fact the theoretic entropy. Imagine in fact the outcome of a process with N equally probable outcomes is known to all but a single person. The number of binary questions (i.e. question with yes or no) that need to be asked from the person in the know to ascertain the true outcome is given by:

$$I = -\log_2\left(1/N\right) \tag{2.1}$$

Where $1/N$: is the probability of positive identification after a single question when all outcomes are equally likely; *I*: is the minimum amount of information needed to obtain a positive identification of the outcome.

In general, whether or not all the outcomes are equally likely, the information needed, now called entropy is defined as the expectation of *I*, e.g. $E\{I\}$ or:

$$H(X) = E\{I\} = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.2}$$

Where equation (2.2) is the discrete form of the entropy for a random variable x; i = 1, 2, 3,…, N; p_i is the probability of occurrence of the event i.

Intuitively, the larger the amount of information required identifying the outcome, the greater a prior uncertainty of this outcome. Thus, a series of observations of an uncertain event contains more information about the event itself than that of a less uncertain event does. Hence, entropy is an indication of uncertainty represented by the probability distribution. Expression (2.3) can be written as

$$H(X) = -K\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.3}$$

where K: is a function of the base used or the scale factor (bits for base 2, napiers for base e, decibels for base 10). So, this definition holds only numbers of outcomes, which are countable and equal to some integer.

Considered as a measure of the amount of chaos or lack of information about a system, if complete information is available, i.e. if there is a pure state, the entropy is zero. Otherwise it is greater than zero. The entropy can be viewed as a

measure of ignorance about the system described in classical sense by a probability distribution. Indirectly, it measures the information about the system. Hence a scalar is assigned to a probability distribution.

It can be shown that the value of $H(X)$ is maximum when all variate values x_i are equally likely, that is, when the outcome has maximum uncertainty. In this case the entropy becomes

$$H_{\max}(X) = \log N \tag{2.4}$$

The ratio of the actual (marginal entropy) to the maximum entropy is called the relative entropy to the source. The more the uncertain the outcome is, the closer the relative entropy is to unity.

One minus the relative entropy is the redundancy. This is the fraction of the structure of the message; which is determined not by the free choice of the sender (station), but rather than by the accepted statistical rules governing the use of the symbols in question (Shannon and Weaver, 1949). It is sensibly called redundancy, for this fraction of message is in fact redundant in something close to the ordinary sense; that is to say, this fraction of the message is unnecessary (and hence repetitive or redundant) in the sense that if it were missing, the message would still be essentially complete, or at least could be complete.

Given two random hydrological variables X and Y the joint entropy of X and Y in the discrete form is given by (note that the marginal entropy of Y is $H(Y)$):

$$H(X,Y) = \sum_{i=1}^{N}\sum_{j=1}^{M} p(x,y)\log\frac{1}{p(x,y)} \tag{2.5}$$

where $p(x,y)$ is the joint probability of x and y.

Equation (2.5) represents the common uncertainty of their measured records X and Y. $H(X)$ can be interpreted as the information about the statistic at site X and $H(Y)$ is the information about the statistic at site Y.

The conditional entropy of X given Y can be regarded as the uncertainty of X given Y or the loss of information and it is expressed by:

$$H(X/Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x,y)\log p(x/y) \qquad (2.6)$$

where $p(x/y)$ is the conditional probability of x given y.

For two discrete random variables X, Y, the following expression defines then the amount of transferred information from X to Y; which is represented by the mutual information or the transinformation given by the following equation

$$T(X,Y) = H(X) - H(X/Y) \qquad (2.7)$$

Equation (2.7) can be viewed as the reduction in uncertainty of X, i.e. $H(X)$, due to the knowledge of Y or the information inferred by X about Y (Amorocho and Espildora, 1973). The transinformation is another entropy measure that measures the redundant information between X and Y (Ozkul et al., 2000).

The minimum transinformation has been used as a criterion in the networking design and evaluation, e.g. in rainfall network design (Krstanovic and Singh, 1992a, 1992b; Al-Zahrani and Hussein, 1998), in river flow network design (Yang and Burn, 1994), in water quality monitoring design (Harmancioglu et al.1994); groundwater design (Mogheir and Singh, 2002).

For a multivariate records (e.g. rainfall / river flow), the multi-dimensional joint entropy for n gauging stations in a region represents the common uncertainty of

their measured records. For more details the reader is referred to Krstanovic and Singh (1992a, 1992b).

The continuous form of entropy is mostly used in formal analysis while in actual numerical sense; the discrete form should be used. However, the latter form implies a much more complex calculation than the former one (Singh, 1998b; Amorocho and Espildora, 1973). The selection of the interval size for estimating a discrete bivariate distribution seems to be arbitrary. This can lead to different values of entropy. Amorocho and Espildora (1973) showed that both continuous form and discrete form could lead approximately to the same answer for entropy and thus, they recommended the continuous form for formal analysis since its computation is straightforward and lesser demanding.

For a continuous random variable X, equation (2.2) or the Shannon entropy becomes:

$$H(X) = -\int_a^b f(x)\log f(x)dx \qquad (2.8)$$

Under the following normality condition

$$\int_a^b f(x)dx = 1 \qquad (2.9)$$

where $f(x)$ is the probability distribution function of the random variable and $H(x)$ is the marginal entropy of the random variable which describes the information contained in x.

In the context of hydrology, the random variable can be streamflow, rainfall, etc… The conditional entropy of X given Y in a continuous form becomes

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

(2.10)

where $f(x/y)$ is the conditional probability density of x given y; $f(x)$ and $f(y)$ are the marginal probability density of x and y respectively.

It follows that:

$$T(X,Y) = -\int_{-\infty}^{+\infty} f(x)\log[f(x)]dx + \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x)f(x/y)*\log[f(x/y)]dxdy \tag{2.11}$$

Equation (2.11) can be written as:

$$T(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy \tag{2.12}$$

where $f(x,y)$ is the joint probability distribution of x and y.

For continuous multivariate distributions, the reader can be referred to Krstanovic and Singh (1992a).

Yang and Burn (1994) showed that equation (2.8) measures the relative information with $-\log\Delta x$ serving as the datum when Δx approaches zero (e.g. range of X is divided into N intervals of width Δx). As a measure of relative information, $H(X)$ can be positive, zero or negative and therefore the information lost, can be also negative, zero or negative. Therefore the information lost $H(X/Y)$ (or H_{lost}) can be also positive, zero or negative since it is a part of the information $H(X)$ and bounded from above by $H(X)$. But negative information or negative entropy lost has no physical meanings although both cases are mathematically possible. This difficulty arises from the use of the relative coordinate for which the origin is set at $-\log\Delta x$. If H and H_{lost} are considered in absolute coordinate in which, the origin is set at minus infinity, then

both $H(X)$ and $H(X/Y)$ are no longer negative and regain their normal physical meanings.

The mutual information T is symmetric, i.e. $T(X,Y) = T(Y,X)$. If the two stations are statistically independent of each other so that no information is mutually transmitted, then $T(X,Y) = 0$. When the two stations are functionally dependent, the information at one site can be fully transmitted to another site with no loss at all. Subsequently $T(X,Y) = H(X)$. In case where $0 \le T(X,Y) \le H(X)$ is the two stations are between totally independent and fully dependent. Thus $T(X,Y)$ is a measure of dependency between stations and $H(X)$ is seen to be the upper limit for $T(X,Y)$. Note that great values of $T(X,Y)$ correspond to great amounts of information transmitted from X to Y.

$T(X,Y)$ does not assume variables to be normal like in the classic correlation coefficient (Harmanciouglu and Yevjevich, 1987) and it does not depend on coordinates (Yang and Burn, 1994). The correlation coefficient cannot take care of arbitrarily relation between coordinates and classes while the mutual information does (Battiti, 1994). It should be noted that the entropy concept does not provide any means to transfer information but it can only measure whether all the transferable information is transferred via a model (e.g. regression, etc.).

The concept of T although indicating the dependency of two stations has been criticized by Yang and Burn (1994) as being not good index of the dependency since its upper bound varies from site to site. In the following section, the original definition of mutual information is improved.

2.2.2.2.3 Directional information transfer index (DIT)

The directional information transfer index was introduced because of the above criticism about transinformation. In order to normalize the upper bound of the mutual information, the original definition of mutual information has been altered to a directional information transfer index (Yang and Burn, 1994). The DIT notion

was initially defined for streamflow network design. The directional information transfer index (DIT) is given by

$$DIT = T/H = (H - H_{lost})/H$$
$$= 1 - H_{lost}/H \qquad (2.13)$$

DIT is physically the fraction of information transferred from one site to another. As $H_{lost} \le H$, thus DIT will range from zero to unity when $0 \le T \le H$.

$DIT = 0$ when no information is transmitted (i.e. independent situation).

$DIT = 1$ corresponds to the case where a fully dependent situation with no information is lost.

$0 \prec DIT \prec 1$ is a situation between independent and fully dependent.

If DIT_{xy} is the fractional information inferred by station X about Y and if DIT_{yx} the fractional information inferred by station Y and X, thus DIT_{xy} is not necessarily equal to DIT_{yx} since $DIT_{xy} = T/H(X)$ for station X will not be in general equal to $DIT_{yx} = T/H(Y)$.

A direct application of DIT is the regionalization of the network. Two related stations should be arranged in the same group since the hydrometric event patterns represented by then are strongly dependent and consequently information can be mutually inferred between them. Thus DIT_{xy} and DIT_{yx} should be high. If neither DIT is high, then the 2 stations should remain in separate groups. If only DIT (say DIT_{xy}) is high, then the station Y, whose information can be predicted by X, can join station X if station Y does not belong to any another group; otherwise it stays in its own group. The DIT is based on the stations essential connection, and is thus distinguished from the traditional similarities measures, such as the correlation coefficient (Yang and Burn, 1994). In traditional methods, the connections between the stations are quantified by similarities, which may be based on one of numerous measures of associations.

2.2.2.2.4 Principle of maximum entropy (POME)

Janes, for the first time in 1957, formulated the principle of maximum entropy (POME). Later, this principle was widely used in hydrology and water resources, e.g. Amorocho and Espildora (1973); Sonuga (1972, 1976); Chapman (1985), etc.

According to the POME, as quoted by Singh (1998a); when making inferences based on incomplete information, the probability distribution to be drawn must have the maximum entropy permitted by the available information expressed in the form of constraints. In other words, this distribution results in minimally prejudiced assignment of probabilities on the basis of given information.

The POME-based distribution is favored over those with less entropy among those, which satisfy experimentally given constraints called "testable information". Thus, entropy defines a kind of measure on the space of probability distributions. Intuitively, distributions of higher entropy represents more disorder, are smoother, are more probable, are less predictable, or assume less. So the POME-base distribution is maximally non-committal with regard to missing information and does not require invocation of ergodic hypotheses. According to POME, the probabilities should be assigned by maximizing the entropy function.

Mathematically the POME is expressed through the following:

Maximize $H(X) = -\int_a^b f(x)\log f(x)dx$ (2.14)

under the m linearly independent constraints $C_i's$

$C_i = \int_a^b y_i(x) f(x)dx$, i = 1, 2,…,m. (2.15)

and the normality condition

$C_0 = \int_a^b f(x)dx = 1$ (2.16)

Where y_i are some functions whose average over $f(x)$ is specified.

Then, it can be shown that the maximum of $H(X)$ subject to the conditions in equation (2.15) and (2.16) is given by the distribution:

$$f(x) = \exp(-\lambda_0 - \sum_{i=1}^{m} \lambda_i y_i(x)) \tag{2.17}$$

where $\lambda_{i's}$ are the Lagrange multipliers and can be determined from (2.13) to (2.17) with the normality condition.

Indeed, most frequency distributions produced in real experiments are maximum distributions (Singh, 1998a). In other words, these distributions are the least biased. Singh and Fiorentino (1992) showed that the distribution derived by maximization of entropy, is the one that maximizes the likelihood, that is, a relationship between the Shannon entropy and the maximum likelihood. When a POME-based distribution departs statistically significantly from an experimental one, it provides a conclusive evidence of the existence of new constraints that were not taken into account in the calculation.

The strength of POME is that this principle provides the most efficient procedure by which, if unknown constraints exist, they can be discovered (Singh, 1998a).

The POME-based approach has several advantages (Singh and Fiorentino, 1992):
(1) It requires little data or information
(2) Availabity of information can be expressed in a variety of different ways (information may be available in terms of moments, bounds, points value, mean, variance, probability, etc).
(3) The derived distribution is most unbiased and consistent with the available information.

Using the POME, Shannon and Weaver (1949) and Singh (1998b) derived a univariate normal distribution under specific conditions. Singh and Kristanovic (1987) used the POME to derive a bivariate normal distribution.

John and Rodney (1980) derived an axiomatic derivation of the principle of maximum entropy and the principle of minimum cross entropy; however the two principles were shown to be equivalent. This was confirmed by Singh (1998a).

Since entropy is a measure of uncertainty or chaos, and variance is a measure of variability, the connection between them is of interest. In general, an explicit relation between entropy and variance does not exist but does exist in the case of specific distributions (Singh, 1998a). For example, using the principle of POME, when the standard deviation and mean of a random variable X are supposed to be known, it can be shown that (Shannon and Weaver, 1949; Singh, 1998a, 1998b):

$$H(X) = \ln\left[s_x (2\pi e)^{0.5}\right] \tag{2.18}$$

where s_x is the standard deviation of the hydrological variable X computed from the available data. Thus the entropy can be seen as another measure of dispersion- an alternative to variance and this suggests that it is possible to determine the variance whenever it is possible to determine the entropy measures, but the reverse is not necessarily true (Singh, 1998c).

Under the assumption that if the marginal distributions of X and Y are normal, and if their joint distribution is normal too, hence, using the POME, it can be shown that (Amorocho and Espildora, 1973)

$$H(X,Y) = \log(2\pi)^{1/2} \left|\Sigma\right|^{1/2} \tag{2.19}$$

$$H(X/Y) = \ln\left\{\sigma_x \left[2\pi e(1-R^2)^{1/2}\right]\right\} \tag{2.20}$$

$$T(X,Y) = -\frac{1}{2}\ln(1-R^2) \tag{2.21}$$

Where R is the correlation coefficient of the gauge X with Y.

$|\Sigma|$ is the determinant of the covariance matrix and σ_x is the variance of the random variable x. The essential condition for existence of the entropy is the positive-definiteness of the covariance matrix (cross correlation matrix).

In a multivariate case (Kristanovic and Singh, 1992b), expression (2.29) can be written as

$$T((X_1, X_2, ..., X_{n-1}), X_N) = -\frac{1}{2}\ln(1 - R^2) \qquad (2.22)$$

where R is the multiple correlation between the independent variables $X_1, X_2, ..., X_{n-1}$ and the dependent variable X_n.

Formulas (2.18)-(2.22) can be applied to any distribution, which can be normalized (Chapman, 1985).

Ahmed and Gokhale (1989) gave the entropy calculations for multivariate distributions other than the multivariate normal distribution.

2.2.2.2.5 Prior distributions for entropy calculations

A number of frequency distributions, commonly employed in hydrology, have been using the POME-based methodology. Sonuga (1972) was probably the first to use POME in hydrologic frequency analysis; he essentially derived a normal distribution. Jowitt (1979) analyzed the extreme-value type I distribution. Examples demonstrating application of POME to the gamma, Pearson-type III, lognormal, and the log-Pearson type III distributions are described by Singh (1998b). Krstanovic and Singh (1992a) extended the POME –based method to multivariate distributions.

Most of the studies on rainfall network design and on river flow network design made use of a prior (assumed) distribution (e.g. normal, log-normal) to fit the

hydrological data in the calculation of entropy (Amorocho and Espildora, 1973; Chapman; 1985; Harmancioglu and Yevjevich, 1987, Yang and Burn, 1994). Normally a variety of distributions should be in general tried and the most likely adopted by applying decision theory techniques (Amorocho and Espildora, 1973).

Chapman (1986) showed that the lognormal distribution was better (than the gamma distribution) to fit the data and was subsequently used for entropy calculation. Singh and Krstanovic (1987) assumed a bivariate normal distribution in the derivation of a stochastic model sediment yield using the POME. In entropy calculations for univariate (bivariate and multivariate) cases, it is simpler to use univariate and multivariate normal distributions than other distributions such as gamma, Pearson type or Weibull. The normal distributions are known because of the complexity involved in application of entropy with other distributions (Krstanovic and Singh, 1992a). Yang and Burn (1994) criticized the normality assumption (for entropy calculation) in their formulation of the non-parametric estimation of probability density function. Although their criticism; they used the logarithmic transformation of the data in that formulation and they also used a *gaussian normal* kernel. So Yang and Burn (1994) recognized that normal and lognormal distributions could be used for entropy computation in the multivariate case because the description of multivariate probability density functions for other skewed density functions is very difficult.

The normal distributions are known as the most important widely used continuous probability distributions because of their early connections with the "theory of errors". Thus, many statistical techniques such as analysis of variance and test of certain hypotheses rely on the assumption of normality (Haan, 1977). In a strict sense, most hydrological variables cannot be normally distributed (Yevijevich, 1972). If the normality is not a viable assumption and if one ignores the normality check and proceeds as if the data were normally distributed, this could lead to incorrect conclusions (Johnson and Winchern, 1996). It should be noted that many continuous distributions could be approximated by the normal distribution for certain values of the parameters (Haan, 1977). Slack et al. (1975) showed that

when the information about the distribution of floods and economics losses associated with the design of floods retardation structures was lacking, it was better to use the normal distribution than other distributions such as Extreme Value, Weilbull, etc…

In addition, for random variables that have characteristically skewed distributions, the lognormal distributions could be used (Yevjevich, 1972; Yang and Burn, 1994). Because of its simplicity, readily available tables for evaluation and the fact that many hydrologic variables are bounded by zero on the left and positively skewed, the lognormal distribution has received wide usage in hydrology (Haan, 1977; Feldman, 1972, Amorocho and Espildora, 1973; Yang and Burn, 1994, Alley and Burn, 1983).

Because of the simplicity of multivariate normal distribution for the computation of entropy, Krstanovic and Singh (1992a) applied the Box-Cox transformation to rainfall data to follow approximately a multivariate normal distribution. The Box-Cox transformation family is mostly used for normality and is equivalent to the family of power transformations (Weisberg, 2001; Krzanowsky and Marriott, 1994; Mason et al., 1989; Johnson and Wichern, 1996).
Power transformations are defined only for positive values. However, this is not as restrictive as it seems, because a single constant can be added to each observation in the data set if some of the variables are negative (Johnson and Wichern, 1996).

The following is a Box-Cox family of transformation:

$$\text{(i)}\ y=\begin{cases}(x^{\lambda}-1)/\lambda & if\ \ \lambda \neq 0\\ \ln x & if\ \ \lambda = 0\end{cases} \qquad (2.23)$$

where λ is the transformation parameter and x should be strictly positive numbers; x represents the data before transformation.

This family of transformation is includes the important special cases of untransformed, inverse, logarithmic, and square and cubic root.
(Weisberg, 2001).

Given the observations $x_1, x_2, ..., x_n$ from a univariate process, the Box-Cox solution for the choice of an appropriate power λ is the one, which maximizes the following expression of the maximum likelihood (ML)

$$\ell(\lambda) = -\frac{n}{2}\ln s^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n}\ln x_i \qquad (2.24)$$

Where

$$s^2(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (2.25)$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (2.26)$$

Generally in hydrological studies, whatever the transformations of the variable, it has to be reminded that the principles of and assumptions regarding parameter estimation method (e.g. Least Squares, etc) apply to the transformed model, not to the original model (Haan, 1977). The method of analysis can be applied to the transformed data and, *if appropriate*, the results can be transformed back so that conclusions are presented in relation to the original units of measurement (Krstanowsky and Marriot, 1994). In other words, if inappropriate, the results cannot get transformed back and conclusions can be drawn on the transformed variables. McCuen et al. (1990) pointed out the following: for the logarithmic transformation (where appropriate), the correlation coefficient reflects the accuracy of the unbiased log space; not the accuracy of the original data; otherwise the results become biased. Thus, the logarithmically derived parameters are not necessarily unbiased estimators (McCuen et al., 1990).

Many other authors drew conclusions in term of parameter estimation (model efficiency) with regard to the transformed variables (Chapman, 1985; Hirsch, 1979, Hirsch 1982, Amorocho and Espildora, 1973; Krstanovit and Singh 1992 b; Singh and Krstanovit, 1987; Kilmartin and Peterson, 1972; Yang and Burn, 1994; Harmacioglu and Yevjevich, 1987). Attempts (to reduce the bias for example in the log transformation) were made and analytical solutions were then proposed (Zucchini et al., 1984; Yevjevich; 1972). However, the analytical solution could lead sometimes to unsatisfactory results (McCuen, 1990). For correcting sensibly the bias, a numerical method was also proposed in the case of a power model (original model); which leads to a log transformation through linearization (McCuen et al., 1990).

In some applications of ANN to water resources, conclusions could be drawn on the transformed variables from the model parameters obtained after transformation, e.g. scaling or standardization of variables (Minns and Hall, 1996; Deo and Thirumalaiah, 2000; Thirumalaiah and Deo, 2000; Shin and Salas, 2000, Abrahart et al., 1999). It is strongly believed that these authors thought in terms of bias introduced when transforming back to the original data, though they did not state it clearly in they studies

Generally, it is much easier to select appropriate transformations for the marginal distributions than for the joint distributions (Johnson and Wichern, 1996).

Simple normality tests (for a univariate distribution) are based on the variance, mean, skewness, etc. (Yevjevich, 1972; Haan, 1977; Krzanowsky and Marriot, 1994). However, the mostly used normality test for multivariate distributions is the plot of the Mahalannobis distance and the chi-squared distribution (Panu et al., 2000; Johnson and Wichern, 1996). In the following, this test is briefly described.

2.2.2.2.5.1 Normality test

Before embarking an analysis that makes distributional assumptions about the data, it is prudent to check that those assumptions are reasonable for the data

under consideration. The majority of techniques to be described rely on the assumption of multivariate normality of data, so testing this assumption is often an important aspect of initial data analysis. Very many tests exist for assessing univariate normality of a sample of values on a single variable. One possible approach in multivariate situation would be to test the marginal normality of each variable separately, using one of these statistics, i.e., coefficient of skweness and kurtosis (Krzanowsky and Marriott, 1994). However, this approach would ignore the covariance between the variables in its execution. Also marginal normality of all variables does not ensure joint normality of the ensemble (Krzanowsky and Marriott, 1994); while in practical applications it may be good enough and easy to test the marginal normality (Richard and Wichern, 1996).

2.2.2.2.5.1.1 Univariate distributions

In the case of univariate, it is important to add that the Q-Q plots are not particularly informative unless the sample size is moderate too large (Johnson and Wichern, 1996). The straightness of the Q-Q plot can be measured by calculating the correlation coefficient of the points in the plot and a powerful test of normality based on it. Formally, the hypothesis of normality at a level of significance is rejected if the value of the computed correlation coefficient falls below the appropriate critical value.

2.2.2.2.5.1.2 Multivariate distributions

One would like to check on the assumption of normality for all the distributions of 2, 3,…, p dimensions. However, as pointed out before, for practical applications it is usually sufficient to test the marginal normality. In the case of a multivariate normal distribution, each bivariate distribution would be normal and the contours of constant density would be ellipses (Johnson and Wichern, 1996). Thus, it is expected that the set of bivariate outcome x such that

$$(x-\mu)\Sigma^{-1}(x-\mu) \le \chi_2^{\,2}(0.5) \qquad (2.27)$$

has a probability 0.5. Thus, one should expect *roughly* the same percentage, 50% of ample observations to lie in the ellipse

$$\text{all x such that } (x-\bar{x})S^{-1}(x-\bar{x}) \leq \chi_2{}^2(0.5)$$

where μ has been replaced by its estimates $\bar{x}$ and Σ^{-1} by its estimates S^{-1}. If not the normality assumption is suspect.

Somewhat more formal method for judging the joint normality of a data set is based on th squared generalized distances or Mahalannobis distances. For more details the reader should be referred to Johnson and Wichern (1996)

2.2.2.2.6 Limitations of entropy theory

Despite the overwhelming advantages offered by the entropy theory, some limitations of the method must also be noted (Harmancioglu et al., 1994). The following limitations are done with much experience related to network design, however it can be extended to other cases. As the situation holds true for the majority of statistical techniques, a sound evaluation of network features by the entropy method requires the availability of reliable data. Applications with inadequate data often cause numerical difficulties and hence unreliable results. For example, when assessing spatial and temporal frequencies in the multivariate case, the major numerical difficulty is related to the proprieties of the covariance matrix and the determinant of the matrix is too small, entropy measures cannot be determined reliably since the matrix becomes ill-conditioned. This often occurs when the available sample size is very small. On the other hand, the question with respect to data availability is *"how many data would be considered sufficient?"* Particularly, it difficult to determine when a data record can be considered sufficient (Harmancioglu et al., 1994). The presence of gaps in data series puts limitations on entropy estimates particularly in the time domain such that temporal design cannot be realized after certain lags (Harmancioglu et al., 1999). The same difficulty extends to space/time design, which leads to unreliable results.

Another important point in entropy applications is that the method requires the assumption of a valid distribution-type. The major difficulty occurs here when different values of the entropy function are obtained for different probability distribution functions assumed for the same variable. On the other hand, the entropy method works quite well with multivariate normal and lognormal distributions. The mathematical definition of entropy is easily developed for other distributions in bivariate cases; however, the computational procedure becomes much more difficult when their multivariate distributions is considered. When such distributions are transformed to normal, the uncertainties in parameters need to be assessed. Shannon' s basic definition of entropy is developed for a discrete random variable, and the extension of this definition to the continuous case entails the problems of selecting the discretizing class interval to approximate probabilities with class frequencies. Different measures of entropy vary with class intervals such that each selected class constitutes a different base level or scale for measuring uncertainty. Consequently, the same variable investigated assumes different values of entropy for each selected class interval. It may take even on negative values, which contradict the positivity property of the entropy function in theory. However, this problem can be alleviated by working in absolute coordinates where the origin is set to minus infinity (Yang and Burn, 1994).

Despite these limitations, within the context of the development of a quantitative definition of information and the value of data, application of the concept of information in entropy theory has produced promising results in several disciplines (Harmancioglu et al., 1994; Singh, 1998c; Krstanovic and Singh, 1992 a, 1992 b; Singh and Mogheir, 2002 and many others).

2.2.2.2.7 Entropy and developing countries

As emphasized before, one of the main problems plaguing environmental and water resources development in developing countries is the lack of data or lack of sufficient data.

Very often, the data is missing or incomplete or are not of good of good quality or the record is not of sufficient length. As a result, more often that not, it is the data that dictates the type of model to be used and not the availability of modeling technology (Singh, 1998c). Subjective information such as professional experience, judgment and thumb or empirical rules has played a significant role in hydrologic practices in many developing countries. Conventional models do not have the capability to accommodate such subjective information, although such information may be good quality or high value.

The potential for application of the entropy theory is enormous in developing countries, for it maximizes the use of information contained in data however little it may be; it permits the use of subjective information and furthermore, it offers an objective avenue for drawing inferences as the model results (Singh, 1998a, 1998c). Thus, in the face of a limited data, the entropy theory can result in a reliable solution of the problem at hand; in addition the entropy-based modeling is efficient requiring relatively little computational effort and versatile in its applicability across many disciplines (Singh, 1998a).

2.3 EXISTING DATA INTERPOLATION (INFILLING) TECHNIQUES

2.3.1 Preamble

Generally, there are two basic problems in dealing hydrologic time series data. In the first case the time-series are of adequate time length but suffer from the presence of data gaps. Data interpolation (infilling), in this case, has been referred to as data augmentation. In the second case, the historic span of the data series is inadequate, and thus efforts are made to extend the historic time span to a desired one. This latter case of data infilling has commonly seen referred to as a data extension.

Panu et al. (2000) recall what other authors have suggested on missing values. The missing values can be viewed as belonging to three categories. Firstly, missing data values are of trivial importance; that is, in a long historic record only a few data are missing, and there are not consecutively distributed. In this case, simple

infilling methods such as infilling by simple average could be satisfactory. It is undesirable to encounter any peaks or extreme values in data gaps while using simple data infilling techniques. Secondly, fundamental data values may be missing; that is, a number of lengthy segments or many intermittent observations are missing; which renders data patterns or data structure unrecognizable from the available or remaining records. Beale and Little (1975) suggest that according to the current state of available techniques, any attempt to infill missing records in this case may be unreliable, and as such the entire record may be dropped from further consideration. Thirdly, significant data values may be missing; that is, a sequence of consecutive observations is missing. In this case the missing values quantitatively or qualitatively considered important and it is deemed useful to develop techniques and methods capable of estimating such missing values as accurately as possible. Further, the number of missing values in this case is considered too short to have any significant damaging effects on the data patterns or data structures of the whole record. The missing values under the third category occur more often (Panu, et al., 2000). The third category will be the focus of this study.

While modeling hydrological data for use in the design of the water resources, it is imperative that all the characteristics of hydrological times-series be considered. Thus, Elshobagy et al. (2000 a, 2000 b) and Panu et al. (2000) have evoked the notion of group and single valued-data. This discussion is not done here.

2.3.2 Assessment of data infilling methods and techniques

Researchers have been tackling the problem of missing data in different ways and from different perspectives as well. The definitions of “missing data” and the expressions that they have used to describe the in-filling process are no less diversified than the different techniques that they used (Panu et al.; Elshorbagy et al., 2000 a, 2000 b).

A group of researchers tackled the problem of intermediate missing data where data or observations before and after the missing observations are available

(Gyau-Boakye and Schultz, 1994; Bennis et al., 1997). The words *patching* (Hughes and Smakhtin, 1996; Makhuvha et al. 1997 a, 1997 b; Pegram; 1985; Zucchini, 1984) or *infilling* (Panu et al., 2000; Gayau-Boakye and Schultz, 1994, Khalil et al., 2001) could be used to express the in-filling in these cases. In cases where historic data are available from only one side of the gap, or in which the gap is so large that the historic data set is only considered bounded from one side, the methods have been called *data extension* by Hirsch (1979, 1982), Alley and Burns (1983), Hughes and Smakhtin (1996). The word synthesizing or interpolation (Simonovic, 1995) or estimation, which is use to indicate the patching type of infilling process (Berkowitz et al., 1992, Bennis et al., 1997; Knotters and van Walsum, 1997). *Data augmentation* is used especially when the overriding objective is to estimate the model parameters and the estimation of missing data comes as a direct result of applying the developed model (Vogel and Stedinger, 1985; Moran, 1974). The term such as *reconstruction* by Hirsch (1979) has been also used to indicate the estimation of missing values.

As said before, data can be categorized into 3 groups, viz data of trivial importance are missing (e.g. a few sparsely distributed, not consecutive, missing observations in a long historical record); fundamental data are missing (e.g. lengthy segments or many intermittent observations) where patterns or structure cannot be recognized from the remaining record and significant data are missing at the same time data gaps are too short to have significant damaging effect on the patterns and the structure of the whole records (Elshorbagy et al., 2000a).

Since the missing values under the third category occur more often in developing countries more often in general, this category should be first the focus in this paper. The available literature on estimation of missing hydrological data can be classified into single-valued approach and group approach (Elshorbagy, 2000 a, 2000 b and Panu et al., 2000).

Only research works closely to the topic of this thesis are reviewed hereafter to give a general view of the available literature on estimating of missing data (i.e.

expectation-maximization and artificial neural networks techniques). Time series methods for data infilling purposes are not part of the current study as the theory on entropy for normal distributions and the expectation maximization techniques for exponential families assume independent hydrological variables. Also, rainfall-runoff models used for extending or/and infilling runoff data as in Bennis et al. (1997) and Kachroo (1992a, 1992b), etc. are not thoroughly discussed here.

The overriding objective of record extension is to maintain statistical proprieties of the time–series (e.g. mean, variance, etc), like in Alley and Burns (1983), Hirsch (1979, 1982). The objective in this study is to estimate missing values (few consecutive observations, viz third category) in a way that minimizes the error (difference between actual and estimated values). However, at the same time statistical properties should be maintained. In the following, traditional regression methods are not discussed.

2.3.2.1 EM techniques

2.3.2.1.1 Background

The EM algorithm was first introduced formally by Dempster et al. (1977). The EM algorithm as formulated for the first time can be called Standard EM algorithm.

This technique is an iterative procedure where the E-step (expectation) adjusts the values of the sufficient statistics, given the incomplete data and the current values of the parameters. The M-step (maximization) solves the likelihood equations using the adjusted values of the sufficient statistics in a sample of complete data. The repeated applications of the E and M steps lead ultimately to the maximum likelihood-ML (Dempster et al., 1977; Little and Rubin; 1987; Schaffer, 1997).

Some modifications of the standard EM exist and were done for example by Makhuva (1997 a, 1997 b), Xu (1997), Meng and Rubin (1991), Jamshidian and Jennrich (1997), Jennrich and Sampson (1976), etc. For the past two decades the use of the expectation maximization (EM) algorithm has become intensive for

problems involving incomplete (missing) data on multivariate normally distributed variables (Little et Rubin, 1987, Laird and Ware, 1982). However, the literature on EM techniques dealing with missing data is very sparse in hydrology, a part from the studies led by Makhuvha (1997 a, 1997b) and Kuczera (1987).

Some of the fields of application of EM techniques are the following: missing data, categorical data analysis, finite mixture analysis, factor analysis, robust statistical modeling, variance-components estimations, survival analysis, repeated measures designs, tissue classes and regression analysis.

The repeated applications of the E and M steps lead ultimately to the maximum likelihood (Dempster et al., 1977). The fact that the missing data are estimated along the way is regarded as an artifact of the procedure. Thus, the EM algorithm copes *with little information (missing values)*. Although implementation of the algorithm involves the estimation of the missing values, the main focus of the literature is on the model parameters (Little and Rubin, 1987). Nonetheless, in this study, the missing values are of primary interest, not the model parameters.

The EM algorithm formalizes a relatively old ad hoc idea for handling missing data: (1) replace missing values by estimated values, (2) estimate parameters, (3) re-estimate the missing values assuming the new parameter estimates are correct, and so forth, iterating until convergence.

Suppose that one has a model for complete data Y, with associated density $f(Y/\theta)$ indexed by unknown parameter θ. One can write $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} represents the observed part and Y_{mis} denotes the missing values. In this chapter, for simplicity, it is assumed that the data are missing randomly (MAR), thus the mechanism of missing-data does not depend on the missing values and that the objective is to maximize the likelihood

$$L(\theta / Y_{obs}) = \int f(Y_{obs}, Y_{mis} / \theta) dY_{mis} \tag{2.28}$$

with respect to the parameter θ.

The distribution of the complete data can be found as

$$f(Y / \theta) = f(Y_{obs}, Y_{mis} / \theta) = f(Y_{obs} / \theta) f(Y_{mis} / Y_{obs}, \theta) \tag{2.29}$$

where the $f(Y_{obs} / \theta)$ is the density of the observed data Y_{obs} and $f(Y_{mis} / Y_{obs}, \theta)$ is the density of missing data given the observed data. The decomposition that corresponds to (2.49) is

$$l(\theta / Y) = l(\theta / Y_{obs}, Y_{mis}) = l(\theta / Y_{obs}) + \ln f(Y_{mis} / Y_{obs}, \theta)$$

The wish is to estimate θ by maximizing the incomplete-data likelihood $l(\theta / Y_{obs})$ with respect to θ for fixed Y_{obs}; this task, however, can be difficult to accomplish directly.

This expression can be re-rewritten as

$$l(\theta / Y_{obs}) = l(\theta / Y) - \ln f(Y_{mis} / Y_{obs}, \theta) \tag{2.30}$$

where $l(\theta / Y_{obs})$ is the observed loglikelihood to be maximized, $l(\theta / Y)$ is the complete-data loglikelihood, which is presumably relatively easy to maximize, and $f(Y_{mis} / Y_{obs}, \theta)$ is the missing part of the complete data loglikelihood. The expectation of both sides of (2.30) over the distribution of the missing data Y_{mis}, given the observed data Y_{obs} and a current estimate of θ, say $\theta^{(t)}$, is

$$l(\theta / Y_{obs}) = Q(\theta / \theta^{(t)}) - H(\theta / \theta^{(t)}) \tag{2.31}$$

where

$$Q(\theta/\theta^{(t)}) = \int [l(\theta/Y_{obs}, Y_{mis})] f(Y_{mis}/Y_{obs}, \theta^{(t)}) dY_{mis}$$

and

$$H(\theta/\theta^{(t)}) = \int [\ln f(Y_{mis}/Y_{obs}, \theta)] f(Y_{mis}/Y_{obs}, \theta^{(t)}) dY_{mis}$$

Note that

$$H(\theta/\theta^{(t)}) \le H(\theta^{(t)}/\theta^{(t)}) \qquad (2.33)$$

by Jensen's inequality (Little and Rubin, 1987).

Consider a sequence of iterates $\theta^{(0)}, \theta^{(1)}, ...,$ where $\theta^{(t+1)} = M(\theta^{(t)})$ for some function $M(.)$. The difference in values of $l(\theta/Y_{obs})$ at successive iterates is given by

$$l(\theta^{(t+1)}/Y_{obs}) - l(\theta^{(t)}/Y_{obs}) = \left[Q(\theta^{(t+1)}/\theta^{(t)}) - Q(\theta^{(t)}/\theta^{(t)})\right] - \left[H(\theta^{(t+1)}/\theta^{(t)}) - H(\theta^{(t)}/\theta^{(t)})\right] \qquad (2.33)$$

An EM algorithm chooses $\theta^{(t+1)}$ to maximize $Q(\theta/\theta^{(t)})$ with respect to θ. More generally, a generalized EM algorithm chooses $\theta^{(t+1)}$ so that $Q(\theta^{(t+1)}/\theta^{(t)}) \ge Q(\theta^{(t)}/\theta^{(t)})$. Hence, for any EM or Generalized EM algorithm, the change from $\theta^{(t)}$ to $\theta^{(t+1)}$ increases the loglikelihood.

2.3.2.1.2 The E-step and the M-step of EM

The steps (E and M) are generally easy to construct conceptually, to program for calculation, and to fit into computer program. The M step is particularly simple to

describe: perform the maximum likelihood estimation of θ just as if there were no missing data, that is, as if they had been filled in. Thus the M step uses the identical computational methods as ML estimation from $l(\theta / Y)$. The E step finds the conditional expectation of the "missing data" given the observed data and current estimated parameters, and then substitutes these expectations for the "missing data". The key idea of EM, which delineates it from the ad hoc idea of filling in missing values and iterating, is that "missing data" is not Y_{mis} but the functions of Y_{mis} appearing in the complete loglikelihood, that is $l(\theta / Y)$.

Specifically, let $\theta^{(t)}$ be the current estimate of the parameter θ. The E step of EM finds the expected loglikelihood $Q(\theta / \theta^{(t)})$ if θ were $\theta^{(t)}$:

$$Q(\theta / \theta^{(t)}) = \int l(\theta / Y) f(Y_{mis} / Y_{obs,} \theta = \theta^{(t)}) dY_{mis} \qquad (2.34)$$

The M – step of EM determines $\theta^{(t+1)}$ by maximizing this expected loglikelihood:

$$Q(\theta^{(t+1)} / \theta^{(t)}) \geq Q(\theta / \theta^{(t)}), \quad \text{for all } \theta. \qquad (2.35)$$

An advantage of this method is that it can be shown to converge reliably, in the sense that under general conditions, each iteration increases the loglikelihood $L(\theta / Y_{obs})$, and if $L(\theta / Y_{obs})$ is bounded, the sequence $L(\theta^{(t)} / Y_{obs})$ converges to a stationary value of θ. If $f(Y / \theta)$ is a regular exponential family and $l(\theta / Y_{obs})$ is bounded, then $\theta^{(t)}$ converge to a stationary point $\theta*$. Quite generally, if the sequence $\theta^{(t)}$ converges, it converges to a local minimum or saddle point of θ. A disadvantage of the algorithm is that its rate of convergence can be painfully slow if a lot of data are missing (Dempster et al., 1977).

To start the iteration, one needs to give $\theta^{(0)}$. In this case, one can compute for example $\theta^{(0)}$ using the imputing conditions means (Little et al., 1987).

2.3.2.1.3 EM theory for exponential families

The EM algorithm is particularly simple and useful interpretation when the complete data Y have a distribution from the regular exponential family (Little et al., 1987) defined by

$$f(Y/\theta^{(t)}) = b(Y)\exp(s(Y)\theta)/a(\theta) \tag{2.36}$$

Where θ denotes a $(d \times 1)$ parameter vector, $s(Y)$ denotes a $(1 \times d)$ vector of the *complete-data* sufficient statistics, and a and b are functions of θ and Y, respectively. This family contains among others, the normal, gamma, inverse Gaussian, binomial and Poisson distributions (Ibrahim, 1991). The E-step for (7.21) consists in estimating the complete-data sufficient statistics $s(Y)$ by

$$s^{(t+1)} = E(s(Y)/Y_{obs}, \theta^{(t)}) \tag{2.37}$$

The M step determines the new estimates $\theta^{(t+1)}$ of θ as the solution of the likelihood equations

$$E(s(Y)/\theta) = s^{(t)} \tag{2.38}$$

However, the normal distribution is much simpler from its theoretical aspect than other distributions (Makhuva, 1997 a; Little and Rubin, 1987; Ibrahim, 1991). For these families, the observed information can be seen as the difference between the unconditional and conditional variance of the complete-data sufficient statistics (Little and Rubin, 1987).

2.3.2.1.4 The bivariate case with data missing from one site only in the context of linear regression

Consider a dataset with variables Y_1 and Y_2 where Y_1 is observed units but Y_2 1, 2,..., n is observed only for units 1, 2,..., m with $m \prec n$. The missing data will

be MAR if the probability that Y_2 is missing does not depend on Y_2, although it may possibly depend on Y_1. Let y_{i1} and y_{i2} denote the values of Y_1 and Y_2, respectively, for unit i.

The assumption made here is that observation pairs (y_{i1}, y_{i2}) from a bivariate normal distribution are independently and identically distributed as $N(\mu, \Sigma)$, where μ, Σ are the mean vector and the covariance matrix such that $\mu = (\mu_1, \mu_2)$ and the determinant of Σ is given by

$$\det(\Sigma) = |\Sigma| = \begin{vmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{vmatrix} \tag{2.39}$$

The likelihood for a bivariate normal sample with m complete bivariate $((y_{i1}, y_{i2}); i = 1,2,..., m)$ and $n - m$ univariate observation $(y_{i1}, i = m+1,..., n)$ is given by:

$$l(\mu, \Sigma / Y_{obs}) = -\frac{1}{2} m \ln|\Sigma| - \frac{1}{2} \sum_{i=1}^{m} (y_i - \mu)\Sigma^{-1}(y_i - \mu)^T$$
$$-\frac{1}{2}(n-m)\ln\sigma_{11} - \frac{1}{2}\sum \frac{(y_{i1} - \mu_1)^2}{\sigma_{11}} \tag{2.40}$$

ML estimates of μ and Σ can be found by maximizing this function with respect to μ and Σ. The likelihood equations however do not have an obvious solution. The joint distribution of y_{i1} and y_{12} can be expressed as a factor of the marginal distribution of y_{i1} and the condition al distribution of y_{12} given y_{i1}:

$$f(y_{i1}, y_{i2} / \mu, \Sigma) = f(y_{i1} / \mu_1, \sigma_{11}) f(y_{i2} / y_{i1}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1}), \tag{2.41}$$

where $\hat{\mu}_1$ and $\hat{\sigma}_{11}$ are the mean and the variance of the n population (site) having the longest period of records.

$f(y_{i1} / \mu_1, \sigma_{11})$ is the normal distribution with mean μ_1 and variance σ_{11} and

$f(y_{i2} / y_{i1}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})$ is the normal distribution with mean

$\beta_{20.1} + \beta_{21.1} y_{i.1}$

and variance $\sigma_{22.1}$. The parameter

$\phi = (\mu_1, \sigma_{11}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})^T$

is one-one monotone function of the original parameter

$\theta = (\mu_1, \mu_1, \sigma_{11}, \sigma_{12}, \sigma_{22.})^T$

of the joint distribution of y_{2i} given y_{i1}. The betas are regression coefficients of y_{2i} on y_{i1}.

Maximizing the likelihood corresponding to the two components in equation (2.40) and assuming the population parameters can be replaced by the sample parameters, i.e. $\mu = s$ will give the following respectively (Little et al., 1987):

$$\hat{\mu}_1 = (1/n)\sum_{i=1}^{n} y_{i1} \qquad (2.42)$$

$$\hat{\sigma}_{11} = (1/n)\sum_{i=1}^{n} (y_{i1} - \hat{\mu}_1)^2 , \qquad (2.43)$$

and

$$\beta_{21.1} = s_{12} / s_{11} \qquad (2.44)$$

$$\beta_{20.1} = \bar{y}_2 - \hat{\beta}_{21.1}\bar{y}_1 \tag{2.45}$$

$$\hat{\sigma}_{22.1} = s_{22} - s_{12}{}^2 / s_{11} = s_{22.1} \tag{2.46}$$

where $\bar{y} = (1/m)\sum_{i=1}^{m} y_{ij}$, $s_{ij} = (1/m)\sum_{i=1}^{m}(y_{ij} - y_j)(y_{ik} - \vec{y}_k)$ for j, k = 1, 2

These are initial values of the statistics. As there are no missing values for site 1, its values will remain unchanged throughout at any t-th iteration. Consequently the estimate of the mean and variance of site 1 will also remain unchanged throughout. These quantities need to be computed at t-th iteration

(i.e. $\beta_{21.1}{}^{(t)} = \sigma_{12}{}^{(t)} / \sigma_{11}{}^{(t)}$)

<u>*E-step*</u>

In the t+1 iteration, ones compute the conditional expectation vector sufficient parameters given the observed values and the current estimate of θ.

The linear terms are computed as follows:

$$E(y_{i2} / X, Y_{obs}, \theta^{(t)}) = y_{i2}{}^{(t+1)} = \begin{cases} y_{i2} & if\ \ y_{i2}\ \ is\ observed \\ \bar{y}_2{}^{(t)} + \hat{\beta}_{21.1}{}^{(t)}(y_{i1} - \bar{y}_1) & if\ \ y_{i2}\ \ is mis\sin g \end{cases} \tag{2.47}$$

The square terms are computed by:

$$E(y_{i2}{}^2 / X, Y_{obs}, \theta^{(t)}) = y_{i2}{}^{(t+1)^2} = \begin{cases} y_{i2}{}^2 & if\ y_{i2}\ is observed \\ y_{i2}{}^{2(t+1)2} + \sigma_{2i}{}^{2(t)2} & if\ y_{i2}\ is mis\sin g \end{cases} \tag{2.48}$$

<u>*M-step*</u>.

In the $t+$ iteration, one computes:

$$\hat{\mu}_2^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} y_{i2}^{(t+1)} \tag{2.49}$$

$$\hat{\sigma}_{2j}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}(y_{i2}^{(t+1)} - \hat{\mu}_2^{(t+1)})(y_{ij} - \hat{\mu}_j) \quad for \quad j=1 \tag{2.50}$$

In case of linear regression, the M corresponds to the least squares analysis on the original design (Little et al., 1987). Consequently the EM iterations can omit the M - step estimation of $\hat{\sigma}_{2j}^{\,2}$ and the E-step estimation of $E(y_i^{\,2} / XY, \theta^{(t)})$ and find $\beta_{22.1}$ by iteration after convergence and then calculate $\hat{\sigma}_{2j}^{\,2}$

As initial estimates for μ and Σ, here one could apply the usual maximum likelihood estimators to the set of complete concurrent data. The imputing conditional means (Buck's method) mentioned so far, can be used to estimate the missing values $\hat{y}_{i2}$ of y_{i2} as

$$\hat{y}_{i2} = \bar{y}_2 + \beta_{21.1}(y_{i1} - \bar{y}_1) \tag{2.51}$$

to start EM algorithm.

2.3.2.1.5 The bivariate case with data missing at both sites

The general pattern of missing data is as follows: Y_1 is observed but are missing Y_2, the second group of units has both Y_1 and Y_2 observed, and the third group of units has Y_2 observed but are missing Y_1. The log likelihood is not linear in the data, but rather is linear in the sufficient statistics (Little and Rubin, 1987):

$$s_1 = \sum_1^n y_{i1}\ ,\ s_2 = \sum_1^n y_{i2}\ ,\ s_{11} = \sum_1^n y^2{}_{i1}\ ,\ s_{22} = \sum_1^n y^2{}_{i2}\ ,\ s_{12} = \sum_1^n y_{i1}y_{i2}\ , \tag{2.52}$$

which are simple functions of the sample means, variances, and covariances.

The task at the E step is to find the conditional expectation given Y_{obs} and $\theta(\mu,\Sigma)$ of the sums in equation (2.52). For the group of units with both y_{i1} and y_{i2} observed, the conditional expectations of the quantities equal their observed values. For the group of units with y_{i1} observed but y_{i2} missing, the expectations of y_{i1} and y^2_{i1} equal their observed values; the expectations of y_{i2}, y^2_{i2} and $y_{i1}y_{i2}$ are found from the regression of the y_{i2} on y_{i1}.

$$E(y_{i1} / X, Y_{obs}, \theta^{(t)}) = y_{i1}{}^{(t+1)} = \beta_{20.1} + \beta_{21.1} y_{i1} \tag{2.53}$$

$$E(y_{i1}{}^2 / X, Y_{obs}, \theta^{(t)}) = y_{i1}{}^{(t+1)^2} = (\beta_{20.1} + \beta_{21.1} y_{i1})^2 + \sigma_{22.1} \tag{2.54}$$

$$E(y_{i2} y_{i1} / X, Y_{obs}, \theta^{(t)}) = y_{i1}{}^{(t+1)^2} = (\beta_{20.1} + \beta_{21.1} y_{i1}) y_{i1} \tag{2.55}$$

These quantities are computed if y_{i1} is observed, otherwise they are equal to zero. $\beta_{20.1}$, $\beta_{21.1}$ and $\sigma_{22.1}$ are functions of Σ corresponding to the regression of y_{i2} on y_{i1} as defined previously. For the units with y_{i2} observed and y_{i1} missing, the regression of y_{i1} and y_{i2} is used to calculate the missing contributions to the sufficient statistics. Having found the expectations of y_{i1}, y_{i2}, y^2_{i1}, y^2_{i2} and $y_{i1}y_{i2}$ for each unit in the three groups, the expectations of the sufficient statistics in (2.52) are found as the sums of these quantities over all n units. The M step calculates the usual moment-based estimators of μ and Σ from those filled-in sufficient statistics:

$$\hat{\mu}_1 = s_1 / n, \ \hat{\mu}_2 = s_2 / n, \tag{2.56}$$

$$\hat{\sigma}^2{}_1 = s_{11} / n - \hat{\mu}_1{}^2, \ \hat{\sigma}_2{}^2 = s_{22} / n - \hat{\mu}_2{}^2, \ \hat{\sigma}_{12}{}^2 = s_{12} / n - \hat{\mu}_2{}^2 \mu_2{}^2 \tag{2.57}$$

The EM algorithm for this problem consists of performing these steps iteratively until convergence.

Although the assumption of multivariate normality may appear restrictive, the methods discussed here can provide consistent estimates under weaker assumptions about the underlying assumptions (Little and Rubin, 1987). It should be noted that even the normality assumption is slightly violated, Mahkhuva et al. (1997 a) pointed out that the EM and the subset selection methods could not lead to nonsensical estimates. The independence assumption can be checked by computing the first order serial autocorrelation; this value can be tested if it is significant within predetermined confidence limits (Yevjevich, 1972; Makhuvha, 1997 a, 1997 b; Xu, 2002; Mkhandi et al. 2000)

The above-described method is the *standard* EM algorithm. Other variants of the EM algorithm exist. This will be set out from the next section.

It should be worth noting that other methods for handling missing data exist and are outlined in General FAQ #25 (1999) and are explained by Schaffer (1997). The methods are listwise or casewise data deletion, pairwise data deletion, mean substitution, Hot deck imputation, raw maximum likelihood methods, multiple imputation. General FAQ #25 (1999) mentioned that regression methods are somewhat better, but not as good as hot deck imputation or maximum likelihood approaches. The EM method falls somewhere in between: it is generally superior to listwise, pairwise data deletion, and mean substitution approaches, but it lacks the uncertainty component contained in the raw maximum likelihood and multiple imputation methods.

2.3.2.1.6 Momentum EM (MEM) algorithm

It is recalled that the EM algorithm is a general methodology for the maximum likelihood (ML). The EM algorithm, while being simple to implement and numerically very stable, is generally slow as repeatedly pointed out (Dempster et

al. 1977; Louis, 1982 and many others). The start of ways of improving the speed of EM is the fact that EM is a first order or linearly convergent algorithm (Xu, 1997). The variant described is the Momentum EM algorithm (MEM).

For an iterative algorithm with a current incremental in the parameter $\Delta\theta = \theta^{(t+1)} - \theta^t$, one can always modify the obtained $\theta^{(t+1)}$ into $\eta\theta^{(t+1)} + (1-\eta)\theta^t$ or

$$\theta^{(t+1)} = \theta^t + \eta * \Delta\theta, \quad \eta \succ 0 \tag{2.58}$$

Usually, this is called the momentum approach. The MEM is easily got by using equation (2.58) to modify the incremental $\Delta\theta = \theta^{(t+1)} - \theta^t$. The momentum has been considered to speed up the convergence (Melijson, 1989) with an appropriate η that is usually chosen heuristically. Xu (1997) demonstrated that the MEM could speed up the convergence of the EM algorithm if a suitable amount of momentum is added and the momentum term should be chosen at least with $\eta \succ 0.5$.

2.3.2.1.7 Expectation constrained maximization (ECM) algorithm

Two major reasons for the popularity of the EM algorithm are that its maximum step involves only complete-data maximum likelihood estimation, which is often computationally simple, and that its convergence is stable, with each iteration increasing the likelihood. When the associated complete-data maximum likelihood estimation itself is *complicated*, EM is less attractive because the M-step is computationally unattractive. In many cases, however, complete data maximum likelihood estimation is relatively simple when conditional on some function of the parameters being estimated. Thus, Meng and Rubin (1993) introduced the ECM algorithm, which takes the advantage of the simplicity of complete data conditional maximum likelihood estimation by replacing a complicated M-step of EM with several computationally simpler CM –steps.

In general, let $G = \{g_s(\theta);\ s = 1,...,S\}$ be a set of S pre-selected (vector) functions of θ. Starting with $\theta^{(0)}$, at the $t+1$ st iteration, t = 0, 1, 2,…, the ECM algorithm first performs the E-step and the S CM-steps instead of the M-step, where CM-steps are defined as follows. For s = 1, 2,…, S, find $\theta^{(t+s/S)}$ that maximizes $Q(\theta/\theta^{(t)})$ over subject $\theta \in \Theta$ to the constraint $g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})$ That is, for s = 1, 2, … S, the sth CM-step in the th iteration of ECM finds $\theta^{(t+s/S)}$ such that

$$Q(\theta^{(t+s/S)}/\theta^{(t)}) \geq Q(\theta/\theta^{(t)})$$
$$\text{for all } \theta \in \Theta_s(\theta^{(t+(s-1)/S)}) \equiv \left\{\theta \in \Theta : g_s(\theta) = g_s(\theta^{(t+(s-1)/S)}\right\}$$

(2.59)

Then the value of θ for starting the next iteration of ECM, $\theta^{(t+1)}$, is defined as the output of the final step of (2.59), that is $\theta^{(t+S/S)} \equiv \theta^{(t+1)}$

The following can be viewed as main convergence properties of ECM algorithm (Meng and Rubin, 1993):

-Any ECM is a Generalized EM. As a result of that any propriety established by Dempster et al. (1977) and Wu (1983) for GEM holds for ECM.

-Suppose that all conditional maximizations in (2.59) of ECM are unique. Then all limit points of any ECM sequence $\{\theta^{(t)},\ t \geq 0\}$ are stationary points of $L_{obs}(\theta/Y_{obs})$ if the G is space of filling at all $\theta^{(t)}$.

-Suppose that all conditional maximizations in (2.59) of ECM are unique. Then all limit points of any ECM sequence $\{\theta^{(t)},\ t \geq 0\}$ belong to the set

2.3.2.1.8 Expectation conditional maximization either (ECME1) algorithm

This algorithm can be viewed as a simple extension of the EM and ECM algorithm with *fast monotone convergence*. This algorithm shares with both EM and ECM the stable monotone convergence and basic simplicity of implementation relative to competing faster convergence methods and was

introduced by Liu and Rubin (1994). The basic idea is to replace the M-step of each EM iteration with a sequence of $S \succ 1$ conditional or constrained maximization, or CM, steps, each of which maximizes the expected complete-data loglikelihood found in the preceding E-step subject to constraints on θ, where the collection of all constraints is such that the maximization is over the full parameter space of θ. That is, each EM maximizes the expected complete-data loglikelihood over some function of θ, say $\theta_s = \theta_s(\theta)(s = 1,..., S)$, where θ_s span the θ space. A CM might be in a closed form or may require iteration, but because the CM maximizations are over small dimensional spaces, often they are simpler, faster and more reliable than the corresponding full maximization called for in the M-step of EM.

The same benefits of working in a lower space hold for maximizing the actual likelihood function subject to the same constraints. So the ECME as suggested by Liu and Rubin (1994) leads to cases where CM-steps maximize either the expected complete-data loglikelihood, as with ECM, or the actual likelihood function subject to the same constraints on θ.

Let $X \in \Im$ be the complete-data with density $f(X/\theta)$ and $Y \in \wp$ the observed incomplete data, where $\theta \in \Theta$, and $Y = Y(X)$ is a many-to-one mapping from $\Im$ to $\wp$. Also let $g(Y/\theta)$ denote the density of Y and $k(X/Y, \theta)$ the conditional density of X given Y; then

$$g(Y/\theta) = \int_{\Im(Y)} f(X/\theta) dX , \qquad (2.60)$$

where

$$\Im(Y) = \{X : X \in \Im, Y(X) = Y\}, \quad f(X/\theta) = g(Y/\theta)k(X/Y, \theta)$$

The objective is to find the maximum likelihood estimate of

$$L(\theta) \equiv \log g(Y/\theta) = Q(\theta/\theta') - H(\theta/\theta') \tag{2.61}$$

where

$$Q(\theta/\theta') \equiv E\{\log f(X/\theta)/Y, \theta'\}$$

is the expected complete-data loglikelihood, and

$$H(\theta/\theta') \equiv E\{\log k(X/\theta)/Y, \theta'\}$$

is the expected missing-data loglikelihood.

The definition of ECME can be traced in the following:
The ECME is an iterative algorithm, $\theta^{(t)} \to \theta^{(t+1)}$, consisting of an E-step, which computes $Q(\theta/\theta^{(t)})$ as a function of θ and S constrained maximization steps indexed by s with input $\theta^{(t+(s-1)/S)}$ and output $\theta^{(t+(s-1)/S)}$. For $s \in \varphi_Q$, $Q(\theta^{(t+s/S)}/\theta^{(t)}) \geq Q(\theta/\theta^{(t)})$ for all θ satisfying $h_s(\theta) = h_s(\theta^{(t+(s-1)/S)})$; for $s \in \varphi_L$, $L(\theta^{(t+s/S)}) \geq L(\theta)$ for all θ satisfying $h_s(\theta) = h_s(\theta^{(t+(s-1)/S)})$; $\varphi_Q \cup \varphi_L = \{1, 2, ..., S\}$

Different algorithms, in the sense of different sample paths $\theta^{(0)}$, $\theta^{(1)}$,…are obtained for different orderings of the S steps. More precisely, the method of Jamshidian and Jennirich (1993) can be viewed technically as a special case of ECME where each CM-step maximizes the actual likelihood and the constrained functions corresponding to different conjugate linear combinations of the parameters across iterations.

When the ECME sequence of the loglikelihood values $\{L(\theta^{(t)})\}$ is bounded above, $L(\theta^{(t)})$ converges monotonically to a finite limit L^*. As with EM and ECM, the limit L^* is not necessarily a stationary value of $L(\theta^{(t)})$. If

$\hbar = \{h_s(\theta) : s = 1,..., S\}$ is space-filling at each iteration, then from Wu (1983), Meng and Rubin (1993) it follows that all limits points of any instance $\{\theta^{(t)}\}$ of an ECME algorithm are stationary points of $L(\theta)$, and $L(\theta^{(t)})$ converges monotonically to $L^* = L(\theta^*)$ for some stationary point θ^*.

The convergence of ECME sequence of likelihood values $\{L(\theta^{(t)})\}$ in general does not imply the convergence of the corresponding ECME sequence $\{\theta^{(t)}\}$ (Liu and Rubin, 1994).

A multi-cycle version of ECM (Meng and Rubin, 1993) is obtained by performing a second E-step before the second CM-step to find the expected complete-data loglikelihood given θ. In other words an E-step precedes each CM step. Since the second CM-step of ECME is for the actual likelikehood, multi-cycle in this case is the same as ECME.

2.3.2.2 Artificial neural networks (ANNs)

2.3.2.2.1 Background

ANNs were first developed in 1940's by McCulloch and Pitts who were inspired by a desire to understand the human brain and its functioning. The past two decades have witnessed a tremendous surge of interest in the application of artificial neural networks (ANNs) for a variety of purposes.

It 's recognized that Rumelhart is one of the pioneers who introduced the back propagation algorithm in ANN, around 1986 (i.e. gradient descent search optimization technique). This algorithm has become quite popular modeling technique in diverse areas such as bio-medical engineering, animal sciences, image processing, water resources engineering, electric engineering, computer science, acoustics, cybernetics, robotics, image processing, financing and others.

Since the early nineties, ANNs have been successfully used in hydrology related areas such as rainfall / runoff forecasting (Minns and Hall, 1996; Abrahart et al.,

1999; French et al., 1992; Deo and Thirumalaiah, 2000; Agarwal and Singh, 2001), in grass geographical information systems (Muttiah et al., 1998), streamflow forecasting, groundwater modeling, water quality modeling, rainfall-runoff modeling (Tokar and Markus, 2000), regional drought analysis (Shin and Salas, 2000), solute transport in soils (Wostern et al., 2001), in infilling streamflow data (Panu et al., 2000; Elshorbagy et al. 2000 a , 2000 b; Khalil et al. 2001). Nelson and Illingworth (1991) gives a list of possible applications for neural networks in general. ASCE Task Committee (2000 b) gave a summary of different hydrologic applications. However, the literature on ANN for streamflow interpolation (infilling) remains very sparse (Panu et al., 2000). The same applies to rainfall data.

There are two kinds of researchers, which may be identified within the field of neural network technology. There are those who are involved in the development of ANN themselves-attempting to find faster, smatter and more efficient ways of implementing ANN theory using computer software and hardware. The second area of research involves the application of neural networks. This area considers how ANNs can be applied to both new and existing domains. It is within this area of research that the ANN techniques fall; i.e. ANNs are viewed as hydrological data interpolation (infilling) techniques as EM techniques.
It is important that some general concepts related to NN are put forward.

2.3.2.2.2 Introduction to ANNs

An ANN is a kind of massive parallel connectionism originated from the research of the human neural system and consists of processing units (representing biological neurons), where each processing unit in each layer is connected to all processing in the adjacent layers representing biological synapses and dendrites. Strictly speaking the networks here should term “artificial” neural networks (ANNs) so as to distinguish them from the biological neural networks occurring in the brains of humans and other living organisms. However as there is no danger of confusion, the prefix “artificial “ can be used or not in what follows. ANNs have

been developed as a generalization of mathematical models of human cognition or neural biology. Their development is based on the following rules:

- Information processing occurs at many single elements (nodes, cells or neurons).
- Signals are passed between nodes through connection links.
- Each connection link has an associated weight that represents its connection strength.
- Each node typically applies a non-linear transformation called an activation function to its net input to determine its output signal.

2.3.2.2.3 Architecture of neural networks

Generally, the architecture for a given neural network can be described by specifying the number of layers, the number of neurons in each layer, each layer activation function, the number of inputs, the number of outputs, how the layers are connected to each other. The number of processing elements for the inputs and outputs layers depends on the number of inputs and outputs for the system. No general rule yet exists for determining of the number of elements to use in the inner or hidden layers (Nelson and Illingworth, 1991). However, if too many nodes are available in the hidden layers, it becomes hard for the network to make a generalization. Too few nodes available leads to an inability to form an adequate representation and to encode what the network thinks are the signification features of the input data. In this situation, the neural network “forgets” too easily (Watanabe, 1997). So hidden layers hold the key to more complex computation.

On way of classifying neural networks is by the number of layers (ASCE Task Committee, 2000a): single (Hopfield nets), bilayer (Carpenter/Grossberg adaptive resonance networks), and multi-layer (most backpropagation). Models using only two layers, directly mapping input patterns and this suffices when there is good similarity of input to output and the encoding provided by the external environment alone can perform the mapping.

ANNs can also be categorized based on the direction of information flowing and and processing. A network where outputs can be passed only to the next layer is said to be a feedforward network. In this category the nodes are generally arranged in layers, starting from a first input layer and ending at the final output layer. The nodes are connected in one layer to those in the next, but not to those in the same layer. Thus, the output of a node in a layer is only dependent on the inputs it receives from previous layers and the corresponding weights. While a feedback network would allow outputs to be inputs to the preceding layers; and a feed lateral connections would send some inputs to other nodes in the same layer (Kothari and Kwabena, 1996). Sometimes networks have closed loops; thus they are said to be recurrent; information flows through the nodes in both directions, from the input to the output side and vice versa. This is generally achieved by recycling previous network outputs as current inputs. Thus allowing the feedback. The network, in which every output from one layer is passed along to every node in the next layer, is said to be *fully connected*.

The application of any type of neural network depends on the problem at hand. For example, recurrent networks perform function such as automatic gain control or energy normalization and selecting a maximum in complex systems whilst feedback loops permits trainability and adaptability (Nelson and Illingworth, 1991). However, the feedforward networks are faster than feedback nets, because, one can set a solution with only one pass and guaranteed to reach stability. On the other hand, feedback networks must iterate over many cycles until the system stabilizes.

In most networks, the input (first) layer receives the input variables for the problem at hand. This consists of all quantities that can influence the output. The input layer is thus transparent and is a means of providing information to the network. The last or output layer consists of values predicted by the network and thus represents model output. The number of hidden layers and the number of nodes in each hidden layer is determined by trial-and-error procedure. The nodes within neighboring layers of the network are fully connected by links. A synaptic

weight is assigned to each link to represent the relative connection strength of two nodes at both ends in preceding the input-output relationship.

Figure 2.1 shows the configuration of a feedforward three-layer ANN. These kinds of ANNs can be used in a wide variety of problems, such as storing and recalling data, classification pattern, performing general mapping from input pattern (space) to output pattern (space), grouping similar patterns, or finding solutions to constrained optimization problems. In this figure, $X = (x_1, x_2, ..., x_n)$ is a system-input vector composed of a number of causal variables that influence system behavior, and Y is the system output vector composed of a number of resulting variables that represent the system behavior. In most hydrological applications, three-layered feedforward ANNs are used (Minns and Hall, 1996; Deo and Thurumalaih, 2000; Thurumalaih and Deo, 2000; French et al., 1992; Zealand et al., 1999) and they are thought to be universal approximators (Tateishi and Tamura, 1997). For that, this study will mainly focus on three-layered feed forward neural networks.

2.3.2.2.4 Training methods for neural networks

The ability to change the weights allows the processing element to modify its behavior in responses to its inputs; this is called learning or adaptation (Nelson and Illingworth, 1991). It is understood here that training is the way a neural network learns. It can be also understood as the process by means of which a network is taught to predict and interpret its informational environment (Freeman and Skapura, 1991).

The following types of training are briefly explained: supervised training; grade training and self organized training. Some researchers term supervised / unsupervised modes as learning modes (Nelson and Illingworth, 1991).

2.3.2.2.4.1 Supervised training

A supervised training implies a regimen where, at the same time, the network is presented the input vectors and the "desired" or "target" or "correct" output vectors. In this case, the network is told exactly what it should be generating as its output. With supervised training, it is necessary to train the neural network before it becomes operational. Usually in practical application, there are training pairs (Freeman and Skapura, 1991). The training pairs may be presented to the network in two different modes, namely pattern (sequential) mode and batch mode. In pattern mode or sequential mode, each time a single training pair is presented, learning takes place; while in batch mode learning takes place after all the training pairs have been presented to the network. One complete presentation of the entire training set is called epoch.

2.3.2.2.4.2 Grading training

It is a kind of the reinforcement or the performance of the neural network whereby a network receives a score or grade.

2.3.2.2.4.3 Self organized training

Self-organizing is the modification of many processing elements at once in response to the input vector. No grade or target is provided. Training session exercises the rules for learning as modifications take place throughout on the entire network system. It is as if the neural network is developing its own heuristic as they go through iterations.

In self-organizing (unsupervised) training, the network uses no external influence to adjust their weights (Nelson and Illingworth, 1991). Instead there is an internal monitoring of performance. The network looks for regularities or trends in the input signals, and makes adaptations according to the function of the network. At the present state of the art, unsupervised training is not well understood and is still the subject of much research; supervised training procedures, on the other hand,

have achieved a reputation for producing good results in practical applications and are gaining in popularity (Nelson and Illingworth, 1991).

Most of the hydrologic applications have used supervised training. That is, this training is used in this study.

2.3.2.2.5 Learning laws

Many learning laws are in common use. Most of the common laws are some sort of variation of the best known and oldest learning law, Hebb's Rule, or Hebb synapse. Research has continued, however, and new ideas are being tried. Some researchers have made the modeling of biological learning as their main objective; others are experimenting with adaptation of their perceptions of how nature handles learning. Unfortunately there is still a great deal one does not know about how learning happens, and experimental evidence is not easy to obtain. Learning is certainly more complex than the simplifications represented by the laws developed in theory. These learning laws include Hebb's rule (which is the oldest one), delta rule, steepest descent rule, Backpropagation (BP) learning. Kohonen's learning rule, Grossberg learning, drive-Reinforcement theory, stochastic learning, etc

In what follows the first three laws will be briefly described and much emphasis will be put on the backpropagation (BP) learning because it's widely applied in many disciplines; particularly in water resources engineering. For the rest of laws, details are given by Nelson Illingworth (1991).

2.3.2.2.5.1 Hebb's rule

Donald Hebb first introduced this learning rule in 1949. This basic rule simply states: "When a neuron stimulates another neuron at a time when the receiving cell is actively firing; the connection from the first cell to the second is strengthened."

This law was later found incomplete in the sense that it does not specify, for example how much the connection between neurons should increase, nor how to compute the activity of the two neurons. Also Hebb's statement of learning does not specify the exact conditions under which the connection should strengthen. Thus the neo-hebbian learning has been introduced. For example in 1960's Michael Cohen's and Grossberg tried to explain learning process by introducing two dynamical differential equations. One of the equations governing the activity change of an arbitrary network at a given instant in time and the other one governing the weight changes of an arbitrary connection in the network at any instant in time. More details are given by Maureen (1992).

2.3.2.2.5.2 Delta rule

This rule is also referred to as the Windrow–Hoff learning rule (Widrow and Hoff used in their ADALINE Model) and the Least Mean Square (LMS) learning rule, because it minimizes the mean square error. It is commonly used and based on the simple idea of continuously modifying the strengths of the connection to reduce the difference (delta) between the desired output value and the correct output value of the processing element.

2.3.2.2.5.3 Steepest (gradient) descent rule

The steepest or gradient decent rule is based on a mathematical approach of minimizing the error between the actual and the desired outputs. The weights are modified by an amount proportional to the first derivative of the error with respect to the weight. Pictorially, one could think of this procedure as descending along the curve to the bottom of a hyperboloidal surface is reached; once one reaches the bottom, the error is at its minimum. So the delta rule can be seen as an example of this rule. This rule is commonly used, even though it converges to a point of stability very slowly (Nelson and Illingworth, 1991; Freeman and Skapura, 1991).

2.3.2.2.5.4. Backpropagation (BP) learning law

This learning algorithm was proposed for the first time by Rumelhart in 1986 and is referred to as a generalization of the gradient descent technique to the network that contains hidden layers. In general terms, the root mean square (RMS) error signal computed at the output layer is used by the hidden layers to update their weights. The fact that the RMS is propagated backwards, as it will be seen later, gives the name of the algorithm. The backpropagation is the most widely used method in neural networks. It usually provides a benchmark for other methods.

The BP concept is an important concept, because high a percentage of all networks today employ this learning law (Nelson and Illingworth, 1991). In this, the term BP algorithm or technique will be much of use.

The BP is a supervised learning algorithm applied to a multi-layer feedforward network. In a multi-layer feedforward network, the nodes in the network can be divided into three or more layers. Nodes in the input layer receive the data (input vector). Example of a three-layered feedforward network is shown in Figure 2.1. The signals are carried along the connections to each of the other adjacent layer and can be amplified or inhibited through weights, w_i, associated with each connection. The nodes in the adjacent layer act as summation devices for the incoming (weighted) signals (Figure 2.2). The incoming signal is transformed into an output signal O_j, with the processing units by passing it through a threshold function. A common threshold function for the ANN is the sigmoid function defined as (Dawson and Wilby, 1998):

$$f(x) = \frac{1}{(1+\exp(-x))} \tag{2.62}$$

which provides an output in the range between 0 and 1. When using the sigmoid function, the output values are preferably scaled to fall into the range between 0 and 1 (Freeman and Skapura, 1991; Minns and Hall, 1996; Hines, 1997). Because of the form of the sigmoidal function, the network outputs will never reach the

values 0 and 1. Thus one can use the values 0.1 and 0.9 (Freeman and Skapura, 1991). Sajikumar and Thandaveswara (1999) suggested to scale both inputs and outputs values into a range FMIN (minimum) to FMAX (maximum), rather than 0 and 1 ($FMIN \succ 0$ *and* $FMAX \prec 1$) but he did not give any specific value.

The threshold function is chosen for mathematical convenience because it resembles a hard limiting step-function for extremely large positive and negative values of the incoming signal and also gives useful information about the threshold value. Furthermore, the sigmoid function has a very simple derivative that makes the subsequent implementation of the learning algorithm much easier. In its general form, equation (2.62) can be written as follows

$$f(x) = \frac{1}{1 + \exp(-ax + b)} \tag{2.63}$$

where a is the gain or scaling factor and b is the bias or the amount of translation of the sigmoidal transfer function on the x-axis and it has been shown that BP networks are equivalent to Fourier series approximation when these sigmoidal units are used (Muttiah et al., 1998). The factor a is also called the shape factor (Reddy and Wilamowski, 2000).

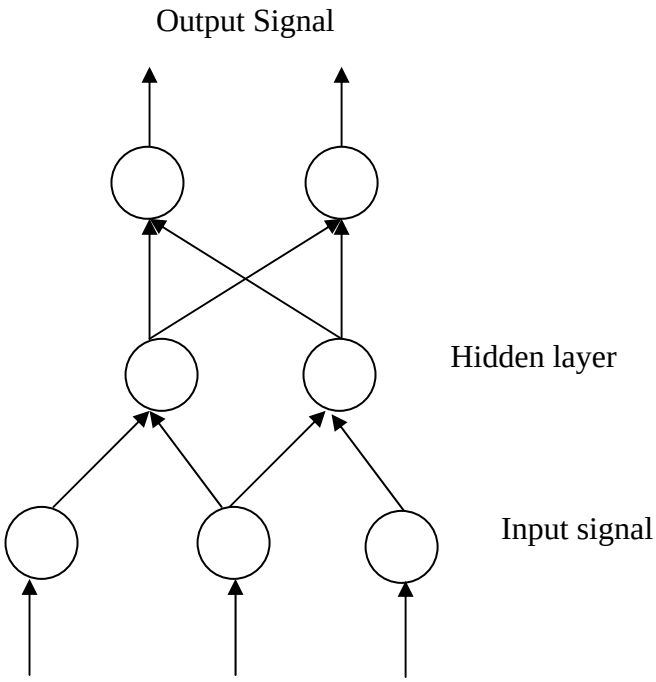

Figure 2.1 A three-layered feedforward ANN

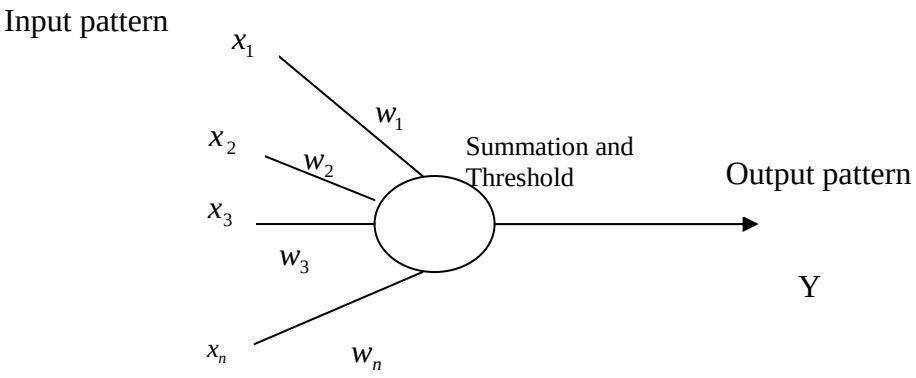

Figure 2.2 A typical ANN node

This output O_j is subsequently carried along the weight connections to the following nodes and the process is repeated until the signal reaches the output layer. The one or more layers of processing units located between the input and output layers have no direct connections to the outside world and are referred to as *hidden* layers. The output signal can be interpreted as the response of the ANN to the given input stimulus. The ANN can be trained to produce known or desired responses for given stimuli. Weights should be initialized to small, random values to the connections, as should be the bias numbers (Hines, 1997; Freeman and Skapura, 1991). Input is then introduced to the input layer and the resulting output is compared to the desired output signal. The input/output vectors can be scaled or normalized before the initialization of the training of the network. Normalization or scaling of input data and output data has the advantage on the speed of convergence of the system and it gives each input equal importance and prevents premature saturation of the activation function (Hines, 1997). The interconnection weights are then adjusted to minimize the error between the ANN output and the desired output. This process is repeated many times with many different input/output patterns until a sufficient accuracy for all the set has been obtained.

The adjustment of the interconnection weights during training employs a method as *error backpropagation* in which the weight associated with each connection is adjusted by an amount proportional to the strength of the signal in the connection

and the total measure of the error. The total error at the output layer is then reduced by redistributing this error value backwards through the hidden layers until the input layer is reached. The next input/output pattern is then applied and the connection weights readjusted to minimize this new error. In this way, the backpropagation algorithm is seen to be a form of gradient descent for finding the minimum value of the multi-dimensional error function. This procedure is repeated until all training data sets have been applied. The whole process is then repeated starting from the first data set once more and continued until the total error for all data set is sufficiently small and subsequent adjustments to the weights are inconsequential. The ANN is now said to have *learned* a relationship between input and training data sets. This way of learning is referred to as *pattern* or *sequential learning*. After the ANN has been trained on sufficiently large number of input-output pairs, it can then correctly *predict* all future input-output pairs, even for those inputs that the network has not seen previously: the ANN is said to have generalized. As said before, no fixed rules as to how many nodes should be included in hidden / layer/s (Freeman and Skapura, 1991; Agarwal and Singh, 2001, Nelson and Illingworth, 1991). If there are too few nodes in the hidden layer, the network may have difficulty to generalize. On the other hand if there are many nodes in the hidden layer, the network may take long to learn. The method normally suggested to reduce the complexity of network is to start with one hidden layer with number of nodes in hidden layer approximately equal to the double of the input nodes (Agarwal and Singh, 2001). The above description of this algorithm is the *standard backpropagation*.

In addition, other activation functions such as hard-limit, linear, etc. can be also used (Nelson and Illingworth, 1991; Demuth and Beale, 1998). Another activation function commonly used is the hyperbolic tangent function (Hines, 1997), which is given by

$$f(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{2.64}$$

It should be worth noting that weights should be initialized to small, random values to the connections, as should be the bias numbers before even the input is introduced to the input layer (Hines, 1997; Freeman and Skapura, 1991). Deo and Thirumalaiah (2000); Agarwal and Singh (2001) as well as Freeman and Skapura (1991) suggested that the initial values weights should be within the range (-0.5, 0.5). Agarwal and Singh (2001) mentioned also other ranges of the initial selection of weights, e.g. (-1.0, +1.0) and (-0.1, +0.1). On its side, the ASCE Task Committee (2000b) suggested that the weights and threshold values are assigned small random values initially, usually in the range (-0.3, 0.3). Patnaik et al. (1996) used the following ranges (0, +0.6), (-0.6, +0.6) and (-0.9, +0.9).

The sequential mode is the mostly used and is recommended in the ANN training process (Minns and Hall, 1996; Hines, 1997; Agarwal and Singh, 2001) for weights updating. Sometime the sequential training may be *more stochastic when the pattern are chosen randomly* and may reduce the chance of getting stuck in a local minimum (Hines, 1997), in this situation the batch learning mode can be tried. During the batch training all the patterns are processed before a weight is update and the process is governed by the error of the data having the highest error in the data domain (Agarwal and Singh, 2001).

The steepest descend method means, whenever one is on the surface of error function; one always goes in the steepest direction towards the down slope within the step size. Consequently, the solution often follows a zigzag path while trying to reach the minimum error position, which may slow down the training process (ASCE Task Committee, 2000a). The method guarantees that the algorithm will find the nearest local minimum. It will not be necessary to find a global minimum on the error surface despite the use of the learning rate. The performance of the BP algorithm, i.e. in function approximation, becomes unsatisfactory when gross errors are present in the training data, however the BP technique is popular for solving practical problems (Chen and Jain, 1994). Despite the error propagation does not guarantee convergence to an optimal solution since local minima may exist, it appears that *in practice* the standard back propagation leads to solutions in

almost every case (Lawrence et al., 1996; Minns and Hall, 1996, Raman and Sunilkumar, 1995). If a network reaches an acceptable solution from an error standpoint, it does not matter, whether the minimum is global or local (Freeman and Skapura, 1991). The step size, which is decided by the learning rate η (which is normally between 0 and 1), plays an important role in the convergence of error (Freeman and Skapura, 1991; Hines, 1997). A small step size ensures a low and smooth convergence with possibility of falling in local minimum. A large step size speeds up the convergence but may cause the network to become paralyzed/oscillate and further training does little/no convergence. French et al. (1992) worked within the range η = 0.01-0.1 and Freeman and Skapura (1991) recommended to use η = 0.05 – 0.25.

The standard BP algorithm of a single training vector in a sequential mode can be summarized as follows (Freeman and Skapura, 1991):

(i). Apply the input vector, $x_p = (x_{p1}, x_{p2}, ..., x_{pN})'$ to the inputs units.

where p is the pattern.

(ii). Calculate the net-input values to the hidden layer units:

$$net_{pj}^{\ h} = \sum_{i=1}^{N} w_{ji}^{\ h} x_{pi} + \theta_j^{\ h}, \qquad \text{h: hidden} \tag{2.65}$$

where $net_{pj}^{\ h}$ is the net input for i unit to hidden node j and $\theta_j^{\ h}$ is the bias term to node j.

(iii). Calculate the outputs i_{pj} from the hidden layer:

$$i_{pj} = f_j^{\ h}(net_{pj}^{\ h}) \tag{2.66}$$

(iv). Move to the output layer. Calculate the net-input $net_{pk}^{\ 0}$ values to each unit:

$$net_{pk}^{\ 0} = \sum_{j=1}^{L} w_{kj}^{\ h} i_{pj} + \theta_k^{\ 0} \tag{2.67}$$

(v). Calculate the outputs:

$$o_{pk} = f_k^{\,0}(net_{pk}^{\,0}) \tag{2.68}$$

(vi). Calculate the error terms for the output units:

$$\delta_{pk}^{\,0} = (y_{pk} - o_{pk})\, f_k^{\,0'}(net_{pk}^{\,0}) \tag{2.69}$$

(vii). Calculate the error terms for the hidden units (e.g. the error terms on the hidden units are calculated before the connection weights to the units have updated):

$$\delta_{pj}^{\,h} = f_j^{\,h'}(net_{pj}^{\,h}) \sum \delta_{pk}^{\,0} w_{kj} \tag{2.70}$$

(viii). Update weights on the output layer:

$$w_{kj}^{\,0}(t+1) = w_{kj}^{\,0}(t) + \eta \delta_{pk} i_{pj} \tag{2.71}$$

The learning rate is the proportionality factor between the weight change and the negative direction of the gradient. A learning rate is used to increase the chance of avoiding training process being trapped in a local minimum instead of global (ASCE Task Committee, 2000).

(ix). Update weights on the hidden layer:

$$w_{ji}^{\,h}(t+1) = w_{ji}^{\,h}(t) + \eta \delta_{pj}^{\,h} x_i \tag{2.72}$$

The order of the weight updates on an individual layer is not important. Nonetheless one has to be sure to calculate the error term for a training pattern p

$$E_p = 1/2 \sum_{k=1}^{N} \delta_{pk}{}^2 \qquad (2.73)$$

where N is the total number of output nodes and the factor 1/2 in (2.73) is normally added for mathematical differentiation purposes (Freeman and Skapura, 1991).

In any training, algorithm, the aim is to reduce the global error, *E* (Mean of Sum of Squared Errors (MSSE)), defined as (Thirumalaiah and Deo, 2000; Chen and Jain, 1994):

$$E = (1/P) \sum_{k=1}^{P} E_p \qquad (2.74)$$

where P is the total number of patterns.

On their side Freeman and Skapura (1991) define the global error, E (Sum of Squared Errors (SSE)) as follows:

$$E = \sum_{k=1}^{P} E_p \qquad (2.75)$$

(ASCE Committee Task, 2000 a) gave the same expression for global error but without the factor 1/2 when computing E_p. SSE remains the mostly used error function as pointed out by Hines (1997) and this happens in most hydrological applications.

The definition of the error term is usually determined by the user's experience and preference (ASCE Committee Task, 2000 b).

For practical considerations, it ‘s suggested sometime to remove the bias terms altogether: their use is optional (Freeman and Skapura, 1991; Demuth and Beale, 1998).

When it comes to consider pairs of gauging stations, the model becomes a single input-output model. The three-layered ANN can implement the transformation through the following function (Wei and Qing, 1991):

$$y(x) = \sum w_i \varphi(v_i, \theta_i, x_i) \qquad (2.76)$$

w_i *and* ϑ_i are respectively the weights to the hidden layer and the weights to the output layer.

To speed up the convergence for the BP algorithm, the following modifications of this algorithm in were proposed.

2.3.2.2.5.5 Other ANN techniques

2.3.2.2.5.5.1 BP algorithm with momentum (MBP)

Sometime the standard BP technique may be slow. Increasing η as the network decreases will often help to speed convergence by increasing the step size as the error reaches the minimum, but the network may bound around too far from the actual minimum value if η gets too large. In other words, if the learning rate is too large, learning can become unstable and errors may even increase (Demuth and Beale, 1998). In this case, the training process may not converge, instead either an oscillation, non-optimal solution is approached or no recognizable solution is developed (French et al., 1992). Thus another way of increasing the speed of convergence is to use a technique with momentum α. This has been analyzed by Phansalkar and Sastry (1994). The momentum factor can speed up the training in very flat regions of the error surface and help prevent oscillations in the weights (ASCE Task Committee, 2000a).

When calculating the weight change value, $\Delta_p w$ a fraction of the previous change is added. This additional term tends to keep the weights change going in the same direction, hence the term momentum. Referring to equations 2.71 and 2.72, the weights change equations on the output layer and hidden layer then become:

$$w_{kj}{}^{0}(t+1) = w_{kj}{}^{0}(t) + \eta\delta_{pk} i_{pj} + \alpha\Delta_p w_{kj}(t-1) \tag{2.77}$$

$$w_{ji}{}^{h}(t+1) = w_{ji}{}^{h}(t) + \eta\delta_{pj}{}^{h} x_i + \alpha\Delta_p w_{ji}(t-1) \tag{2.78}$$

where $\Delta_p w_{kj}$ and $\Delta_p w_{ji}$ are the weight change values corresponding to the output and hidden layers respectively. The momentum parameter α is usually set to positive values less than 1. The use of momentum term can also be optional as well as the use of bias term (Freeman and Skapura, 1991).

2.3.2.2.5.5.2 Variable learning rate BP (VLR) algorithm

In the previous sections, a fixed learning rate was used. When training a neural network iteratively, it is more efficient to use an adaptive learning rate (Hines, 1997; Hagan et al., 1996; Demuth and Beale, 1998; Patnaik et al., 1996). The learning rate can be thought of as the size of a step down of the error gradient. If very small steps are taken, one is guaranteed to find an error minimum, but this may take a very long time. Larger steps may result in unstable learning since one may step over a minimum. To speed training and still have stability, a heuristic method is used to determine the step size.

1. If training is "went well "(error decreased) then increase the step size.

$$\eta = \eta * \rho \qquad (\rho \succ 1). \tag{2.79}$$

Thus, the weight update is accepted.

2. If training is "went poor "(error increased) then decrease the step size.

$$\eta = \eta * \delta \qquad (\delta \prec 1) \tag{2.80}$$

Thus, the weight update is discarded.

Hines (1997) suggested $\rho = 1.1,\ \delta = 0.5$ while Demuth and Beale (1998) suggested these values be 1.05 and 0.7 respectively. Patnaik et al. (1996) did not give any specific value for these parameters.

2.3.2.2.5.5.3 Generalized BP (GenerBP) algorithm

The main reason for problems of the standard backpropagation is due to the derivative of the activation function (Ng et al., 1996). When the actual o_{pk} is approaching to either the extreme values, such as 0 or 1, the derivative of the activation function having the factor $o_{pk}(1-o_{pk})$ (where o_{pk} is the actual output of the k-th output neuron for the p-th pattern) will vanish, and the BP error signal will become very small (Ng et al., 1996). Thus the output can be maximally wrong without producing a large error signal. The algorithm can be trapped into local minima. Consequently the weight adjustment of the algorithm can be very slow or ever suppressed. Therefore a generalization on the derivative of the activation function (i.e. logistic) is proposed so as to improve the convergence of the learning process by preventing the error signal drop to a very small value.

The error signals for the output layer and hidden layer become now:

$$\delta_{pk}^{\ 0} = (y_{pk} - o_{pk})(f_k^{\ 0'}(net_{pk}^{\ 0}))^{1/b} \tag{2.81}$$

$$\delta_{pj}^{\ h} = (f_j^{\ h'}(net_{pj}^{\ h}))^{1/b} \sum \delta_{pk}^{\ 0} w_{kj} \tag{2.82}$$

where the activation function is sigmoid. In this case $b \geq 1$. For $b = 1$, one gets the BP algorithm. The effect of GenerBP is to change the slope of the sigmoid function in the two "tail" regions. For $b \succ 1$, error will be significantly enlarged when o_{pk} will approach a wrong value, the error signals will reflect the true error

($y_{pk} - o_{pk}$) more appropriately. This technique was applied to different problems including XOR, 3-bit parity and the 5-bit counting problems.

2.3.2.2.5.5.4 Quick backpropagation (QBP) algorithm

This algorithm appears to the one of the fastest algorithm reported in the literature (Alexander et al., 1994). Despite the name, quick backpropagation (Patnaik, 1996, Alexander et al., 1994) is not necessarily faster than standard BP, although it may prove significantly faster for some applications; i.e. multisensor data fusion for a single target scenario as detected by Airborne Track While Scan radar (Patnaik, 1996).

QBP works by making the (typically ill-founded) assumption that the error surface is locally quadratic, with the axes of the hyper-ellipsoid surface aligned with the weights. If this true, then the minimum of the error surface can be found after only a couple of epochs.

QBP is *batch-based* and it uses the following formula for weight updating (Patnaik et al., 1996)

$$\Delta\omega(t) = \frac{s(t)}{s(t-1) - s(t)} \Delta w(t-1) \tag{2.83}$$

where $s(t)$, $s(t-1)$ are current and previous values $\partial E / \partial w$.

The above formula is numerically unstable if $s(t)$ is very close to, equal, or greater than $s(t-1)$. In this case the weight formula becomes:

$$\Delta\omega(t) = \alpha_1 \Delta w(t-1) \tag{2.84}$$

Where α_1 is the accelerator coefficient.

On the other hand Alexander (1994) gave proposed the following weight update expression:

$$\Delta\omega(t) = \frac{s(t)}{s(t-1) - s(t)} \Delta w(t-1) + \eta s(t) \qquad (2.85)$$

The second term in expression (2.85), a gradient descent term based on the error gradient and some learning rate η, is added only when two consecutive error gradients have the same sign; otherwise it is omitted. Initialization of the learning process is accomplished by considering only the second term, because the first term makes no contributions during the first iteration.

In hydrology, i.e., the standard BP technique remains the most popular. Other techniques have been sparsely applied in this field, e.g. conjugate gradient method, Levemberg – Marquandt method (Deo and Thirumalaiah, 2000; ASCE, Committee Task, 2000 a; Thirumalaiah and Deo, 2000; Reddy and Wilamoski, 2000) while some others have not even probably been tried within the same field, specifically for problems of data interpolation (infilling). With regards to the hydrological data interpolation techniques, the literature of ANNs remains very sparse (Panu, 2000); if ANNs have been used, it ‘s mostly with the BP learning law.

2.3.2.2.6 Strengths and limitations of neural networks

The strengths have been summarized by ASCE Committee Task, 2000 a. Some of the features of neural networks that can be usefully employed in hydrology are: (a) neural networks are useful when the underlying problem is either poorly defined or not clearly understood and where information is little (missing hydrological data), (b) their application do not require a prior knowledge of the underlying process, (c) they are advantageous when specific solutions do not exist to the problem posed, (d) they can be able to recognize the relationship between the input and the output variables without explicit physical consideration, (e) they possess other inherent information-processing characteristics and once trained are easy to use.

The particular advantage of the ANNs even if the “exact” relationship between sets of inputs and outputs data is unknown but is acknowledged to exist, the network can be trained to learn that relationship, requiring no prior underlying assumptions (e.g. non-linear versus linear versus multiple regression) as in conventional methods (Minns and Hall, 1996; Muttiah et al., 1998; Corne et al., 1998; Abrahart et al., 1999). ANNs seek to learn patterns not to replicate the physical processes in transforming input to output (Minns and Hall, 1996) and they regarded as ultimate black-box models (Agarwal and Singh, 2001; Minns and and Hall, 1996). As opposed to conventional methods, the ANNs are thought to have the *ability to cope with the missing data* (limited data) and perhaps mostly important are able to generalize a relationship from the small subsets of data whilst remaining relatively robust in the presence of *noisy or missing inputs*, thus they can learn in response to changing environment (Dawson and Wilby, 1998; Corne et al., 1998; ASCE Committee Task, 2000 a). The neural networks have the capabilities of approximating functions and they are regarded as a generalization of regression analysis (Wostern et al., 2001). In situations where information is needed only at specific sites in a river basin and where adequate meteorological or topographic information are not available, specific and simple neural networks seem alternatives to apply (Deo and Thirumalaiah, 2000).

Although several studies indicate that ANNs have proven to be potentially useful tools in hydrology, their disadvantages should not be ignored (ASCE Committee Task, 2000 b). The success of an ANN application depends both on the quality and the quantity of data available. This requirement cannot go back far enough. Quite often the requisite data is not available and has to be generated by other means, such as another well-tested model. Even when long historic records are available, one is not certain that conditions remain homogeneous over the time span. Therefore, data sets recorded over a period that is relatively stable and unaffected by human activities are desirable. Yet another limitation of ANNs is in the lack of physical concepts and relations. The fact that there is no standardized way of selecting network architecture also received criticism. The choice of network archtecture, training algorithm, and definition are usually determined by

the user's experience and preference, rather than the physical aspect of the problem. Despite these shortcomings, ANNs remain greatly used in hydrology and resources related fields.

2.4 MODEL PEFORMANCE EVALUATION CRITERIA

It is the objective of the application of models for interpolating (infilling) hydrological data gaps to achieve optimum agreement between computed and observed data. The optimum should be specified by some criteria, which have to be formulated in mathematical terms and quantified with the aid of the relevant data. The criteria are chosen according to their suitability to a given study (Sajikumar and Thanveswara, 1999).

To evaluate the adequacy of the proposed models or techniques (i.e. EM and ANNs) as given previously, the evaluation of the models should be measured analytically and related to stable statistics. There are several model criteria for assessing the performance of the models. Several hydrological evaluation criteria available in the literature are applied in hydrology modeling (Panu et al., 2000; Agarwal and Singh, 2001; Hu et al., 2000; Gyau-Boaye and Schultz, 1994). Those criteria include the following: Transinformation (T), ratio between T and H ($R_{T/H}$), D (difference of marginal entropies), Mean Square Error (MSE), volumetric error (VE), relative mean square error (RME).

(i) Transinformation (T); Conditional entropy $H(X/Y)$; ratio $R_{T/H}$ (of $T(X,Y)=T$ to $H(X)=H$); D (difference of marginal entropies)

$$T(X,Y) = -\frac{1}{2}\ln(1-R^2) \tag{2.86}$$

$$H(X/Y) = H(X) - T(X,Y) \tag{2.87}$$

$$R_{T/H} = T \big/ H \tag{2.88}$$

$$D = H(Y) - H(X) \tag{2.89}$$

$$\mathrm{Re}\,d(\%) = (H_{cc} - H_{comp}) / H_{cc} \tag{2.90}$$

where H_{cc} and H_{comp} are entropy values before and after infilling the data series.

In the above equations, X and Y are *the observed and simulated values* respectively. R is the usual correlation coefficient of X and Y. Formula (2.86) is better used when models have to be compared on the same set of data or same catchment (Amorocho and Espildora, 1973; Chapman, 1985; Singh and Fiorentino, 1992). The higher the value of $T(X,Y)$, the better the model. Expression (2.87) can be used for assessing the performance of a given simulation model in terms of the degree of accuracy in terms of its predictions. Formula (2.88) can be used for evaluating models applied to different catchment areas (Chapman, 1985). The difference in expression (2.89) was proposed as a criterion for measuring the performance of models (Chapman, 1985); thus a big difference is interpreted as failure of the model to predict flow extremes (minima or maxima). On the other hand, a good model will have a very small difference. (Panu, 1992) defined equation (2.90) as the reduction in uncertainty at the subject station before and after infilling the data series.

$$\text{(i) } \textit{Mean Square Error}\,(MSE) = \left[\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{i} \right]^{1/2} \tag{2.91}$$

MSE is simply defined by (Agarwal and Singh, 2001) as the Root Mean Square Error (RMSE). The mean square error (MSE) shows the measure of mean residual variance summed over the given period (e.g., Argawal and Singh, 2001; Sajikumar and Thandavewera, 1999).

(ii) $Volumetric\ Error(EV) = \sum_{n=1}^{n}(\hat{y}_i - y_i) / \sum_{i=1}^{n}(y_i)$ (2.92)

This is the absolute prediction error.

(iii) $RME = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}$ (2.93)

A value of RME near zero implies that the model is providing a good estimate of the missing values (Panu et al., 2000). A higher relative error is indicative of greater deviation from the observed and vice versa.

(iv) A scatter plot of the simulated versus observed (Sajikumar and Thandavewera, 1999; Stephenson, 2002; etc).

2.5 CONCLUSION OF THE LITERATURE REVIEW

The problem of missing hydrological data is prevalent in developing countries. The proportion of missing values can vary from 0 % to about 65 % in some cases (Midgley et al., 1994). In other cases, the records are even inexistent. For application of different data interpolation (infilling) techniques, there is not yet a universal agreement on the range of missing data proportion.

Entropy approach is a versatile tool where hydrological information is little (limited or missing) since it maximizes the use of information in data, however, little it may be (Singh, 1998c). This concept only measures whether all transferable information is transferred via a model (e.g. regression, etc.) but it does not give any means to transfer information. Entropy can be used to measure the information content of a hydrological variable. Entropy is also used to measure hydrological model performance. Recall that entropy computations of both continuous and discrete distributions lead approximately to the same results.

But with the former distributions, the computation is more efficient than with the latter ones. Entropy within the context of hydrological data infilling is not yet fully exploited.

Within the context of regression methods, recent techniques (EM algorithm and its extensions) have been used intensively in problems dealing with missing data. From their iterative procedure, these techniques are reputedly known to dealing with little information unlike traditional regression methods. However, the literature on EM techniques dealing with missing data remains very sparse in hydrology and water resources related fields apart from articles published by Makhuvha (1997a, 1997 b) and Kuczera (1987).

Generally, the normality assumption of hydrological data makes easier the computation of both entropy (Kristanovic and Singh, 1992a, 1992b; Amorocho and Epilsdora, 1973 and others) and EM techniques for univariate and multivariate distributions than for other distributions (Makhuva, 1997a, 1997b; and others). The transformation of data to follow approximately a normal distribution is mostly made by the family of Box Cox transformations. Hence in hydrology, it often happens that model parameters are determined from the transformed variables not from the original data (Haan, 1977; Yevjevich, 1972; Chapman, 1985; Yang and Burn, 1994; etc). Conclusions should be drawn on the transformed data if the analytical (numerical) back transformation to original data can cause some biasness (McCuen et al., 1990). This remark is more general.

Besides the EM techniques, ANNs have been intensively used in hydrology and water resources related fields. These techniques reveal to be powerful tools in coping with missing data (or limited information). However, the literature on neural networks dealing with missing data remains vary sparse in hydrology apart from few researches led for example by Elshorbagy et al. (2000 a, 2000 b); Panu et al. (2000); Eshorbagy et al. (2001) and Khalil (2001). In hydrology, model parameters for ANNs are sometime computed from the transformed (e.g. scaled, standardized) data when using the sigmoid function. The conclusions are, in most

cases, made (for untransformed variables) from the parameters computed from the transformed variables. In this case, the model efficiency on untransformed variables is seen to be satisfactory (i.e. relatively high). However, other authors, in their hydrological studies, drew conclusions on model performance with regards to transformed data (Minns and Hall, 1996; Deo and Thirumalaiah, 2000; Thirumalaiah and Deo, 2000; Salas and Shin, 2000, Abrahart et al., 1999). It is strongly believed that these authors thought in terms of bias (e.g. negative values or unrealistic results) occurred during the mathematical back transformation to original data, although they did not state it. Therefore, in this case the model efficiency for transformed data and original (untransformed) data is not thought to be the same.

Several model evaluation criteria have been compiled in this chapter. The criteria are chosen according to their suitability to a given study (Sajikumar and Thanveswara, 1999).

In the light of the above, the literature with regard to the combined concepts of entropy, ANNs and EM remains sparse in hydrology, specifically for data interpolation (infilling) problems. This study takes the opportunity to combine the three concepts and to present merely a methodology for hydrological data infilling. The methodology as explained in the next chapter was tested specifically to some selected catchments of South Africa.