CHAPTER 3

METHODOLOGY

3.1 SUMMARY

The current study presents a simple methodology by combining different approaches and coping with missing (limited) hydrological data using the theories of entropy, artificial neural network (ANN) and expectation-maximization (EM) techniques. The entropy concept is known as a versatile tool. Hence, this concept is firstly used for quantifying information content of hydrological variables (e.g. rainfall or streamflow). The same concept (through the directional information transfer index, i.e. DIT) is used in the selection of base/subject gauge. Finally, the directional information transfer index is extended to the evaluation of the hydrological data interpolation (infilling) technique performance (i.e. ANN and EM techniques). Thus, the validity of these data infilling techniques with respect to the different gap durations can be defined through entropy concept. The results were discussed.

The results from DIT values were crosschecked with other criteria such as statistical and graphical. Nonetheless, the DIT notion has the feature of being a non-dimensionally informational index. The notion of DIT could enable to compare data infilling techniques on different catchment areas. The data interpolation (infilling) techniques viz. artificial neural networks (ANNs) and expectation maximization EM techniques (e.g. existing methods applied and not yet applied in hydrology) and their new features have been also presented.

The methodology in this study is simply expressed into a model named ENANNEX since **en**tropy concept, **a**rtificial **n**eural **n**etworks and **ex**pectation maximization techniques were used. The summary of the flow chart of the model is given in Figure 3.1. The details of the model will be given in the next section.

The methodology of this study was applied to annual mean flow series; annual maximum flows, annual total rainfall; and 6-month flow series (means) or seasonal mean flows of selected catchments in the drainage region "Orange" of South Africa. A brief description of the data availability will be given at the end of this chapter.

To arrive at the different objectives, the proposed model containing different modules performs the following tasks:

- Checking whether the time-series of each gauging station is independent is performed. This eases the entropy calculations and the EM techniques as explained so far in the previous chapter.

- Transforming of data (to follow the normality assumption), *if necessary*, using Box Cox transformation families, is done. This transformation makes possible the EM theory, which is easily developed for normal distributions. The normality test is carried out much easier for normal distributions than for others. If the original (raw) data are approximately normal, transformation is not necessary and one can move to next step.

- Computing of marginal entropy for information contained in each gauging station is performed. The conditional entropy is computed to define the uncertainty remaining in one gauging station giving the information at the other. Combining the marginal entropy and the conditional entropy, the transinformation (T) or mutual information is defined. Therefore, the directional information transfer index (DIT) can be known. Thus, the determination of the base and the subject gauges within a station pair is based on the value of T (DIT). Hence, the station pair is selected if it satisfies the entropy criterion (i.e. its DIT value should be above some threshold value).

- Applying EM and ANNs algorithms to fill in the missing data is performed. The following techniques: standard EM, EM-with momentum (MEM1), Expectation constrained maximization (ECM1), Expectation Constrained Maximization-Either (ECME1). In this study, MEM2 and MEM3, which are the second version and the third

version of MEM1 are also introduced respectively. The effect of the momentum term on ECM1 and ECME1 is also carried out. Thus ECM2, which is a second of ECM1 is used. ECME2 and ECME3, which are the second and the third versions of ECME1 are also formulated respectively. For ANNs, the following algorithms are performed either in sequential or batch learning: standard backpropagation (BP) algorithm, BP with momentum (MBP), Variable Learning Rate (VLR), Generalized BP (GenerBP), Quick backpropagation (QBP), Golden Search BP (GoldSBP). In the BP technique, the Mac Laurin power series (order 1 and order 2 derivatives) are also used to approximate the different activation functions (e.g. sigmoid) in the hidden layer. Thus, McL1BP and McL2BP techniques, incorporating Mac Laurin power series order 1 and order 2 derivatives respectively, are formulated as other versions of the standard BP.

- Selection of the best technique to fill in the gaps is based on the value of T(DIT) between simulated and observed values. The DIT notion is clearly introduced in this study as a model selection criterion. The higher the value of T(DIT) the better the model. Some threshold value for DIT is set (e.g. the minimum amount of uncertainty that can be removed from the observed data series at the target gauge by applying a given technique). The technique is selected when its DIT is greater than (or equal to) this threshold value. The reduction in uncertainty at the target gauge (before and after infilling) as defined by Panu (1992) is extended, in this study, to cases of assessing flow simulation models. An illustration of this extension is given in Chapter 7.

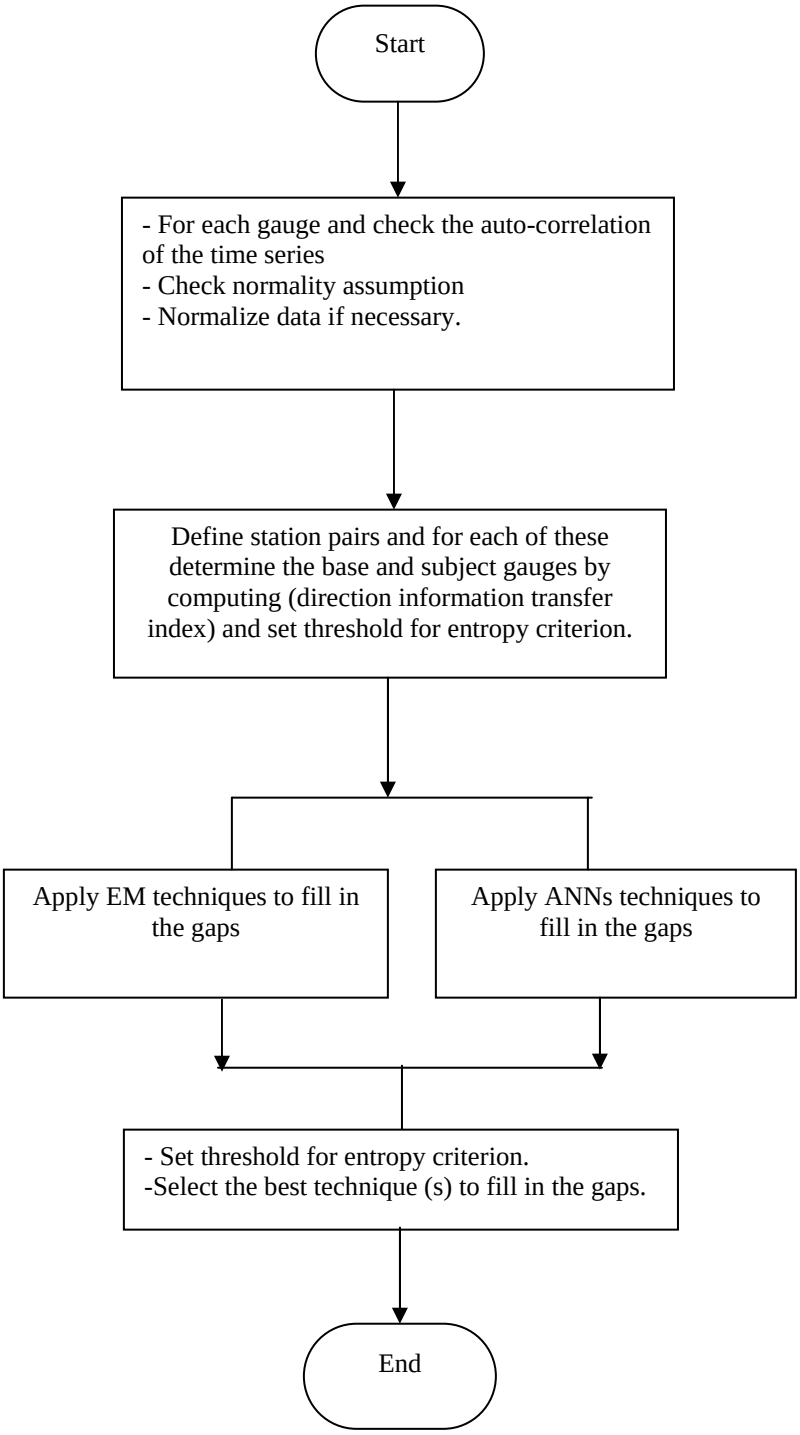


Figure 3.1 Summary of ENANNEX model

In order to crosscheck the results from entropy calculations, other statistical criteria such as mean square error of predictions (RMSE), relative mean error (RME), volumetric error (VE) are computed. To make the comparison fair, ANNs and EM are applied to the data after transformation, *if necessary*.

The transformation back to the original data can be possible through the inverse of mathematical transformations. In case where this operation introduces biased (e.g. negative values), conclusions can still be drawn on transformed data.

3.2 MODEL ASSUMPTIONS, SPECIFICATIONS AND JUSTIFICATIONS

This thesis is dealing with the data "interpolation" or "infilling", where data before and after the gaps are available. The missing data fall into a category where significant data are missing (e.g. consecutive observations). In this case, the missing data are considered important enough to deserve developing a technique that estimates them accurately as possible. At the same time data gaps are too short to have significant damaging effect on the structure of the whole records (Elshorbagy et al., 2000a).

Since the missing values under this category occur more often in developing countries in general, this category should be first the focus in this study. This case could happen as a result of interruption of measurements because of prolonged equipment failure, stopping the measurements at some stations for any raisons (e.g. budget limitations) and resuming them after some timed and accidental loss of data files. For example in South Africa, the overwhelming majority of gaps are caused by temporary absence of observers, the cessation of measurement or absence of observations prior to the commencement of measurements (Makhuvha et al., 1997a).

Different situations will be taken into consideration, namely:

(a) gaps with different durations;

(b) gaps in different regimes;

(c) gaps in different climatic zones;

This consideration was necessary, because it is well known to model builders that hydrological models may fit the data points differently e.g. monthly data, seasons, annual, etc.

The problem of missing data is analyzed here by exploiting the information transfer (through entropy approach) from nearby gauges. It is emphasized in this study that the techniques (i.e. EM and ANN techniques) used, assume that some records (rainfall or streamflow) are available. They are not designed to create a streamflow record where absolutely no record is available. The choice of these techniques depends on the available data (information), which is limited for developing countries in general.

The following data interpolation (infilling) techniques were considered for the objectives of this study:

-Standard BP (with its new features, for example McL1BP –Mac Laurin first order derivative BP and McL2BP- Mac Laurin second order derivative BP)

-BP-with momentum (MBP)

-Variable Learning Rate (VLR)

-GenerBP (GenerBP)

-Quick BP (QBP)

-Standard EM

-Momentum EM (MEM1)

-Second version and third version of the Momentum EM (MEM2 and MEM3)

-Expectation Conditional Maximization (ECM1) and its version ECM2

-Expectation Constrained (conditional) Maximization-Either (ECME1) and its second version (ECME2) and third version (ECME3).

The traditional approach is that records at a site are sometimes in-filled by exploiting inter-station correlation of streamflow with a base (control) station having a long-term record (Hirsch, 1979 and 1982; Elshorbagy et al. 2001). This method ignores gauged flows/rainfall at many other potentially important stations, which could be used for filling

in some of the missing record. It is not always true that the short record should be considered as target station (Zucchini et al., 1984).

The methodology here selects the base station(s) from among several in a region for filling in missing data, if its (they) directional information transfer index value(s) in (a) station pair(s) is (are) at least above some threshold value (entropy criterion).

However, where interpolation or infilling or estimation of missing data is considered, minimizing the squared error (difference between estimated and true values) is the overriding objective as in Panu et al. (2000) and in Elshorbagy et al. (2000a). However, Panu (1992) emphasized that the intent of any infilling activity is to produce a time-series which, when considered as a whole, possesses statistical characteristics indistinguishable to those of historical records for the gauging station. Most importantly these statistics should be maintained for design-oriented purposes such as reservoir design, determination of reservoir operating policies (Makhuvha et al., 1997a, Zucchini et al. 1984). Thus, this study does not neglect this aspect. The objective here is to estimate missing data (few consecutive observations) in a way that minimizes the error (difference between actual and estimated values), however the statistical requirements should be fulfilled.

The EM theory is particularly simple and of useful interpretation when the complete data have a distribution from the regular exponential family (Dempster et al., 1977; Little et al., 1987). Specifically, the normal distribution is simple from its theoretical aspect (Makhuva et al., 1997 a, 1997 b; Little et al., 1987; Ibrahim, 1991). This particular form of distributions was the focus in this study. Moreover, the computations of entropy for multivariate normal distributions are simpler than for other types of distributions. It is also much easier to carry out the normality test in the case of multivariate distributions.

The theory on multivariate normal distributions assumes the realizations (e.g. rainfall, streamflow) to be serially independent. The foregoing assumption is needed here to mainly avoid a bias that can arise if the gaps in the sequence of observations occur in

runs. While this assumption should be checked in each instance, the methods (EM) to be used will not lead to nonsensical estimates even if the assumption is slightly violated (Makhuva et al., 1997a; Little et al., 1987, Dempster et al., 1977).

Monthly stream flow/rainfall data may display large serial correlation; in that event, one may want to consider other flow regimes such as annual mean flows, maximum flows series, annual total rainfall. In the case of rainfall, it is well known that rainfall sequences are not serially independent. However, this serial dependence is generally both short-term and quite weak, so much so that, for practical purposes, sequences of annual and even monthly rainfall can be regarded as being serially independent. In the case of streamflow both daily and monthly records are usually considered as being dependent and this applies sometime to seasonal flows (Zucchini et al., 1984) and in many cases annual flows are assumed to be independent. It was found convenient in this study that the auto-correlation of the time-series be checked for seasonal mean flows, annual mean flows, annual maximum flows and annual total rainfall whenever carrying out entropy computations and EM techniques. Thus, the independence of data within the time series will be checked by computing for example the first order auto-correlation coefficient.

The methodology in this study is such that only cases of independent realizations (e.g. rainfall, streamflow regimes) are considered. This consideration is drawn from the fact the EM techniques and entropy computations developed in this thesis are based on the normality assumption.

The method for carrying out entropy computations for multivariate (bi-variate) discrete variables is very complicated and somewhat subjective in terms of selection of the time-step (Chapman, 1985). For that reason, distributions to fit the hydrological data are assumed to be continuous for simplicity and based on the fact that entropy computations for continuous distributions lead approximately to the same results as when discrete hydrological variables are used (Amorocho and Epilsdora, 1973). This consideration was used to ease the computations for entropy. If the marginal distribution of each hydrological variable in a station pair is normal, it is also assumed that their joint

distribution and their conditional distribution are normal too. This consideration is based on practical purposes as outlined in the literature survey where it may be good enough and easy to test marginal normality.

If the raw data are not normal, appropriate transformations can be used. An alternative and convenient strategy to make the data follow approximately the normality assumption was to use the Box-Cox family of transformations, which are the most popular. Thus, for example the regime of data (e.g. annual mean flows, etc) could be referred to these data after transformation, only in cases where the inverse transformation (back) to original data could be susceptible to introduce biased (e.g. negative) estimated values.

The gaps (missing data) are taken to be missing at random (MAR), in other words the mechanism of missing data does not depend on the missing values. In South Africa, it is believed that it not unreasonable to assume that the three “missing mechanisms” (i.e. temporary absence of observers, the cessation of measurements or the absence of observations prior to the commencement of measurements) are not affected by rainfall depths during the relevant periods (Makhuvha et al., 1997a). The same applies to loss of records. Temporary failure of measuring devices also occurs and it is conceivable that, in some instances, failure could be related to the rainfall depth, for example damage by storms. (In the case of streamflow measurements, floods are more likely to damage measuring devices). However, the proportion of data that goes missing due to this cause is, at least in South Africa, negligibly small. South Africa is mentioned here as the methodology was tested on its selected catchments.

The methodology is developed for rainfall (streamflow) station pairs since the notion of directional information transfer index is used in the selection of potential predicted (target or subject) station and predictor (base or control) station for network design (Yang and Burn, 1994). The same notion in this thesis was therefore extended to data interpolation (infilling) technique performance assessment.

In contrast with physical hydrology, the model in this thesis is referred to as a systems investigation model regarded as being concerned with problems subject only to the constraints imposed by the available data and not subject to " physical considerations" (Minns and Hall, 1996). The selection of gauging stations (e.g. streamflow gauges or rainfall stations) to form the base stations set is not based on the climatic conditions (i.e. size and seasonal correspondence), the length of streamflows and the climatological records or other physio-geographic characteristics. However, this selection is based on entropy approach (i.e. DIT notion). In this study, the information available is only rainfall (streamflow) at a certain station with its neighboring rainfall (streamflow) stations. Unlike in typical rainfall-runoff modeling, the notion of transferability of information is applied here only to "similar" stations, i.e. rainfall stations with its nearby sites; the same applies to streamflow gauging stations.

Time series analysis for infilling missing data (Panu et al. 2000) is not part of the current study as the theory of entropy and of EM techniques for exponential families assume independent hydrological variables.

Although implementation of the algorithm involves the estimation of the missing values, the main focus of the literature of EM techniques is on the model parameters (Little and Rubin, 1987). Nonetheless, it can be also dedicated to missing values like in Makhuvha (1997 a, 1997 b). Buck's method was used to start the EM techniques.

Two categories of techniques, EM and ANNs are used in the analysis. The selected streamflow gauges and rainfall stations (Midgley et al., 1994) were complete and thus exhibited no gaps. However, for testing the methodology outlined so far, some gaps were created artificially on the data set. These artificial gaps were then interpolated (infilled) using the techniques outlined above and comparison was made between the estimated values and the historical data. Different durations (sizes) of "artificial gaps" (i.e. 6.7 %, 13.4 etc.) are created in the time series to simulate the gaps in the hydrological data. In this way, it was possible for the techniques to make use of the available information

before and after the missing data. The two categories of techniques in the model are designed as follows:

a) The testing data are unseen during parameter estimation and assumed to be missing. Thus, few patches of consecutive observations are removed from the data set of the potential target station to test the estimation of the procedures. Different gap durations (length of missing data) were considered, e.g. 6.7 %, 7.4%, 13.4%, 20% and 30%.

b) For any EM technique, the estimated value of an observed event is the observed value itself and the model parameters are estimated several times on the complete set (i.e. observed plus estimated) until convergence. To make the comparison (similarity) fair between EM and ANNs, the ANN techniques are trained over the concurrent data first as in Kuligowski and Barros (1998) and finally the observed values remain intact; only the missing values are estimated, similarly to Bennis et al. (1999).

c) The different techniques are trained to represent the different rainfall/streamflow regimes of the data (annual mean flows, annual total rainfall and annual maximum flows, 6-month flow series (means), etc).

The selection of the best technique(s) to fill in the data gaps is based on entropy criteria as follows:

-Computing the value of transinformation T between observed and simulated values of the target gauge is performed. The transinformation can be used when techniques are to be compared on the same data or same catchment area as pointed out in the literature review. The higher the value of T, the better the model. In addition to the above selection entropy criterion, the notion of directional information transfer index (DIT) is explicitly introduced here for model selection. Recall that Yang and Burn (1994) introduced DIT (which is always positive and between 0 and 1 or between 0 % and 100 %) as a generalization of T. The DIT was initially defined for hydrological network design and therefore defined the dependency between stations, hence a criterion for regionalization. Yet Chapman (1985) introduced already the ratio of the transinformation to the marginal

entropy as criterion for assessing techniques on different catchments. In Chapman's case the marginal entropy should be positive. In case of negative values, the value of marginal entropy can be considered in absolute coordinate in which the origin is set to minus infinity. Then, the marginal entropy is no longer negative and regains its normal physical meaning. Hence, in this study, the DIT was used and extended to technique performance assessment since it is a generalization of the transinformation (T). The higher the value of DIT, the better the technique. The DIT notion shows the amount of uncertainty removed from the subject station via a given technique. DIT was also used to compare technique selection on different catchments unlike T.

It should be noted that the percentage of reduction in the value of entropy of the subject gauge station after infilling of missing values as defined by Panu (1992), see equation 2.90 of Chapter 2 can be also used. However in this thesis, the same formula could be extended to cases where absolutely no data exist at the subject site. In this situation, simulation models were used to estimate the flows at that site from nearby sites and were assessed by an extended expression of Panu's formula. Chapter 7 illustrates this case.
Most important formulas from Chapter 2 were repeated in the current chapter. The reader is referred to the previous chapter for more details.

3. 3 MODEL DEVELOPMENT

The characteristics (e.g. independent realizations) in events (rainfall or streamflow) are examined to develop data interpolation (infilling) procedure for these events. Independent realizations are considered as a starting point for filling in streamflow data, for example in the study led by Khalil (2001); Goodier and Panu (1994). In the following, two categories of hydrological data interpolation (infilling) techniques are encompassed in the model: EM and ANN techniques. Through this methodology, the best technique(s) and then a comparison between the different techniques in terms of their performance can be made. For that, entropy approach is used in technique selection. The methodology has merely been translated into a model named ENANNEX, which is summarized in Figure 3.1. The details of this model are given in the form of a flow chart in Figure 3.2.

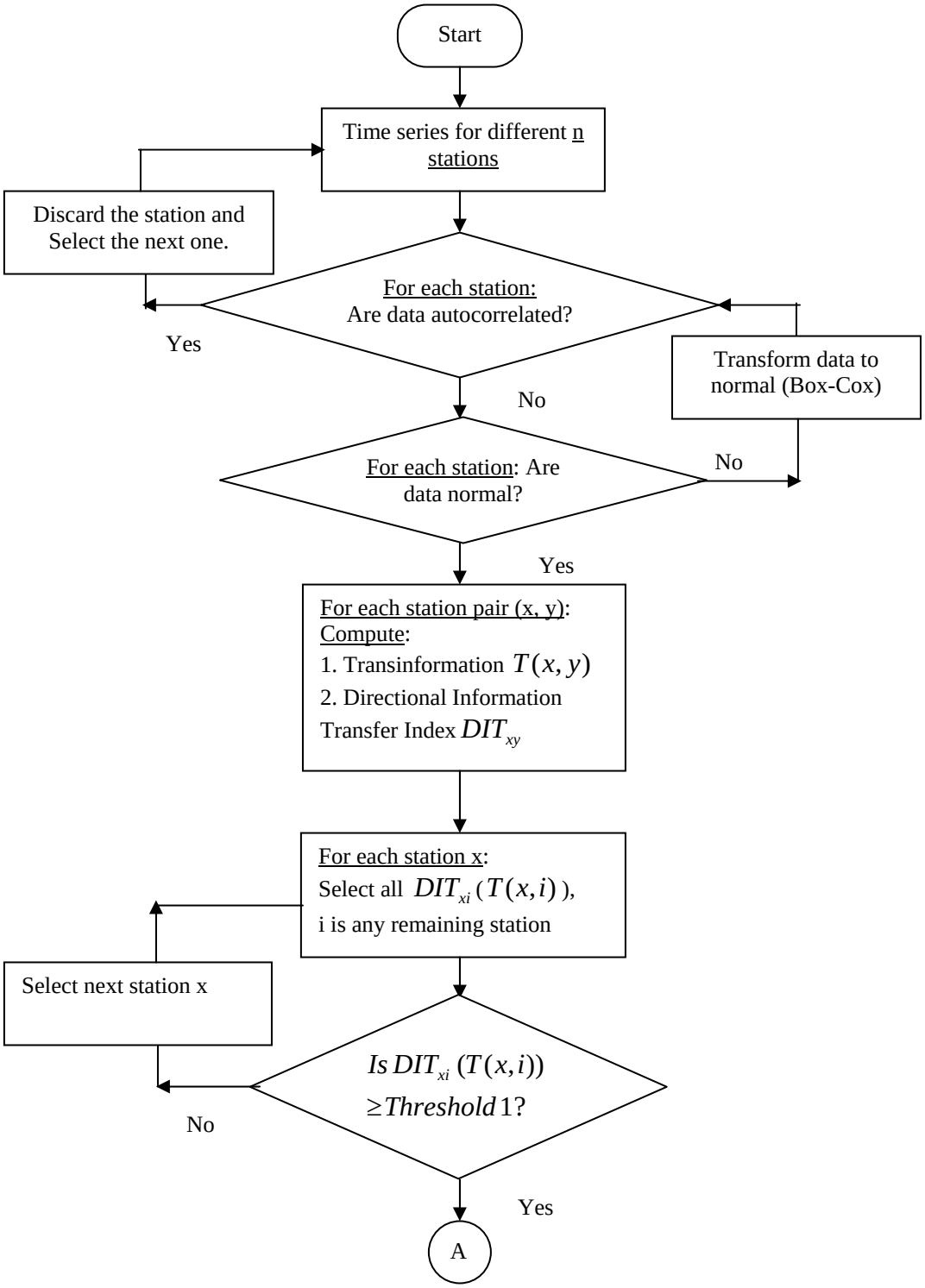
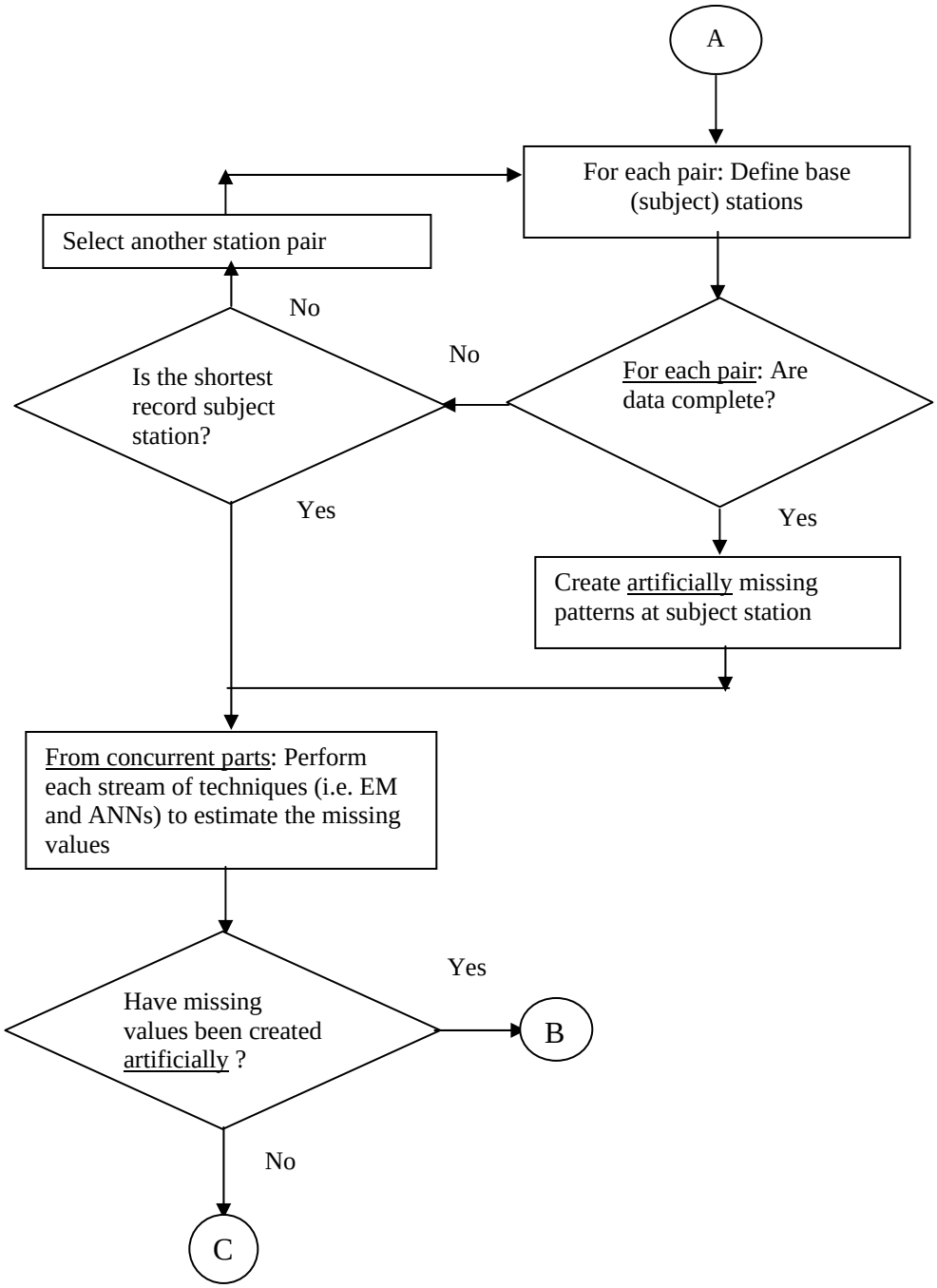
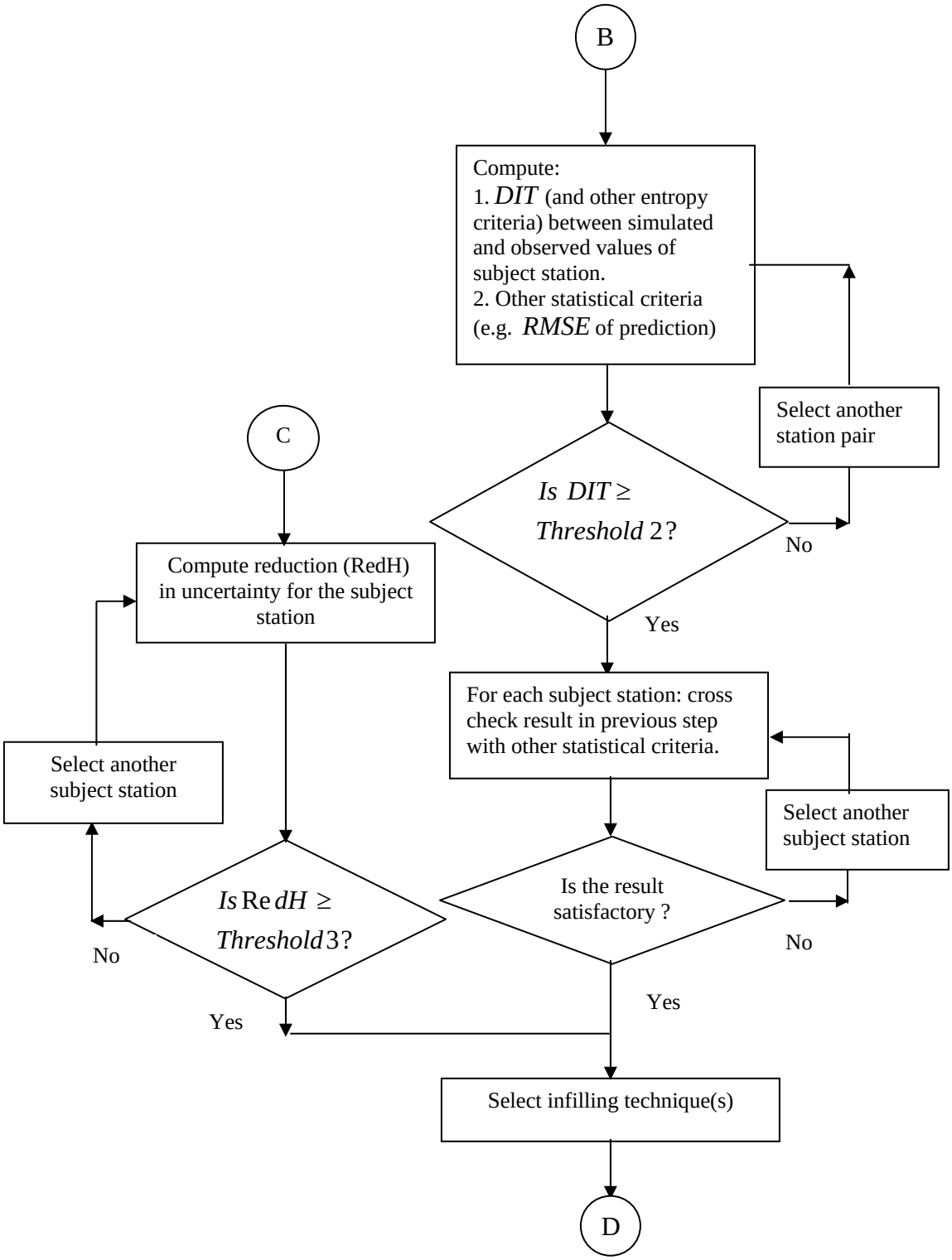
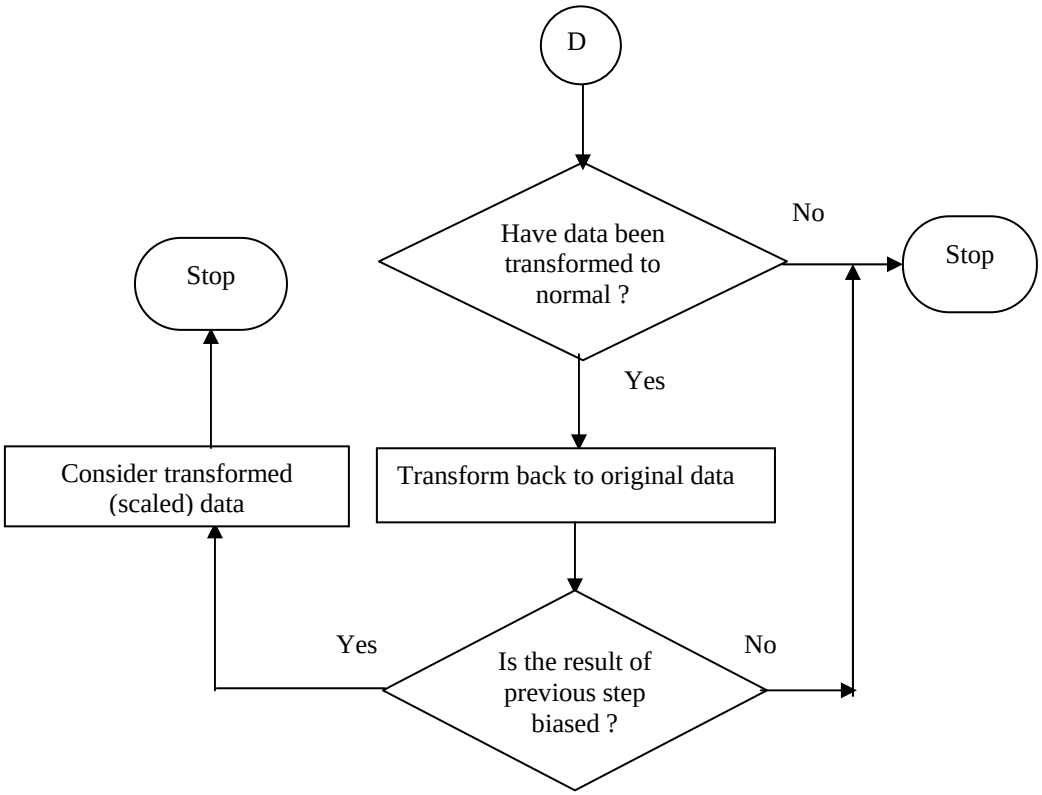


Figure 3.2 Flow chart for model development


A
For each pair: Define base (subject) stations
Select another station pair
No
No
Is the shortest record subject station?
For each pair: Are data complete?
Yes
Yes
Create artificially missing patterns at subject station
From concurrent parts: Perform each stream of techniques (i.e. EM and ANNs) to estimate the missing values
Yes
Have missing values been created artificially ?
B
No
C


B
Compute:
1. DIT (and other entropy criteria) between simulated and observed values of subject station.
2. Other statistical criteria (e.g. RMSE of prediction)
Select another station pair
C
Is DIT ≥ Threshold 2?
No
Yes
Compute reduction (RedH) in uncertainty for the subject station
For each subject station: cross check result in previous step with other statistical criteria.
Select another subject station
Select another subject station
Is RedH ≥ Threshold 3?
Is the result satisfactory ?
No
No
Yes
Yes
Select infilling technique(s)
D



The detailed flow chart is explained in the following, i.e. from step 1 to step 8

3.3.1 Step 1:Testing data series independence

The first step in developing the different techniques encompassed in the model involves the testing of structure independence of the hydrological time-series (e.g. rainfall or streamflow) for different stations of a given network. This consideration is based on the theory of entropy and EM techniques developed so far for exponential families, particularly for normal distributions. To check whether the time-series are independent, the first order auto-correlation coefficient is computed and the following test is performed:

The first order serial correlation computed from the sample is given by

$$r(1) = \left[\sum x_i x_{i+1} - (\sum x_i)^2 / n\right] \Big/ \left[\sum x_i^{\,2} - (\sum x_i)^2 / n\right] \tag{3.1}$$

and its value from the population is represented by $\rho(1)$.

The confidence limits (for a circular series) are given below

$$l = (-1 - z_{1-\alpha/2}\sqrt{n-2}) \big/ (n-1)$$

$$L = (-1 + z_{1-\alpha/2}\sqrt{n-2}) \big/ (n-1)$$

where l and L are respectively lower and upper limits, $z_{1-\alpha/2}$ is the critical value for different values of significance level α (Haan, 1977; Mason and Gunst, 1989).

If the calculated value $r(1)$ falls outside these confidence limits, the hypothesis that $\rho(1)$ is zero ($H_0 : \rho(1) = 0$ *versus* $H_a : \rho(1) \neq 0$) is rejected. H_0 is the null hypothesis and H_a is the non-null hypothesis. In other words, in case of rejection, the observations in the data series are auto-correlated otherwise these observations are independent.

So the techniques used in the model are concerned only with independent events (e.g. annual mean flows, seasonal flows, maximum annual flows, annual total rainfall, etc). In fact, the conclusions would be totally unrealistic if strongly auto-correlated raw data were to be transformed to follow approximately normal distributions, as model parameters would be determined from transformed (normal thus independent) data; not from the original (raw) data, which are actually serially correlated. Therefore, the model selects only the stations for which the time-series are tested to be independent.

A subjective choice can be made also at this stage as stipulated by Elshorbagy (2000a): "One can also assume a threshold value of 0.5 for the correlation coefficient below which the data series is considered to be insignificantly auto-correlated."

3.3.2 Step 2: Checking normality assumption

This step is to check whether the data are normal. In order to test the normality condition, the original data are ranked to their corresponding probabilities and one calculates the standard normal quantiles. Hence, the straightness of pairs, i.e. quantiles-ordered observations (Q-Q plot) is examined. The straightness of the Q-Q plot is measured by calculating the correlation coefficient of the points in the plot and the test of normality is based on the following consideration:

Formally, the hypothesis of normality at the level of significance α is rejected if the computed value of the correlation coefficient r falls below the appropriate value r^* (see, Johnson and Wichern, 1996). If the data do not exhibit signs of non-normality, transformations using Box-Cox power families are applied until the normality is achieved. These families of transformation are the mostly used in hydrology and include the important special cases of untransformed, inverse, logarithmic, square root, square and cubic (Yevjvich, 1972). The following strategy was used in this study:

$$y = \begin{cases} 1/x, & \lambda = 1(inverse) \\ \ln x, & \lambda = 0(\log arithmic) \\ \sqrt{x} & \lambda = 0.5(square\,root) \\ x, & \lambda = 1(untransformed) \\ x^2, & \lambda = 2(square) \\ x^3, & \lambda = 3(cubic) \\ (x^{\lambda} - 1)/\lambda, & if\ \lambda\ is\ not\ any\ of\ above\ values \end{cases} \tag{3.2}$$

For practical purposes as outlined in the literature survey and repeated in the previous section, it may be good enough and easy to test marginal normality.

If the original (raw) data follow approximately the normality assumption, thus the current step can be skipped. Recall that the EM techniques used here will not lead to nonsensical estimates even if the assumption is slightly violated (Makhuvha et. 1997a; Little et al., 1987, Dempster et al., 1977).

3.3.3 Step 3: Computation of transinformation (T) and directional information transfer index (DIT)

The third step is the computation of the mutual information (transinformation), i.e. T and subsequently the computation of the directional information index (DIT) in order to define for each station pair (x, y) the target (subject or predicted) gauge and base (control or predictor) gauge. Figure 3.3 depicts a bivariate case with missing data at one gauging station. Recall that DIT is based on the premises of station pairs. Computations are performed from the concurrent parts (i.e. missing data are excluded) of the gauging stations.

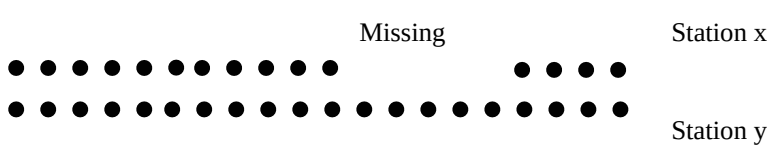


Figure 3.3 An example of missing data in station x (bi-series case)

Formulas (2.13) and (2.21) can be used as:

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

$$DIT = {T(X,Y)}/{H(X)}$$

Therefore the total number of values of T is the number of station pair combinations, as T is symmetric. However the total number of DIT is the number of station pair permutations, as DIT is non-symmetric.

3.3.4 Step 4: Determination of base/target gauge

The fourth step is the determination of the likely potential target station and base station. The procedure in this step is as follows: for each station i, one select all the values DIT for which information is inferred about a given station i by any other station j, i.e. DIT_{ji}. The possible values of DIT_{ji} are then compared to a threshold value, i.e. threshold1. The j stations to be retained as potential base stations (and i as subject station) are those; which satisfy the following condition:

$$DIT_{ji}(T_{ij}) \geq Threshold1 \qquad (3.3)$$

Yang and Burn (1994) assumed arbitrarily a threshold value of 0.35 to measure the association between gauging stations for network design using extreme flows.

In this study, the choice for Threshold1 values was made as follows:

-If none of the DIT values is above 0.2 (e.g. 20 % of information inferred by the potential base station about the subject station) thus no station pair is taken.

-In case some (all) values are between 0.2 and 0.3, thus threshold value is set to 0.2.

-In other cases where some (all) values are above 0.3, the threshold value is set to 0.3.

Although being subjective, the choice for the threshold as made above was found to produce reasonable results when testing the methodology (refer to chapters 4, 5 and 6).

Figure 3.4 depicts an example of a group of 5 stations where information is inferred about station 1. The remaining stations, e.g. 2, 3, 4 and 5 are potential candidates as base stations. Similar figures can be drawn for example where each of the remaining stations is considered to be the inferred (predicted or subject) station with respect to the others.

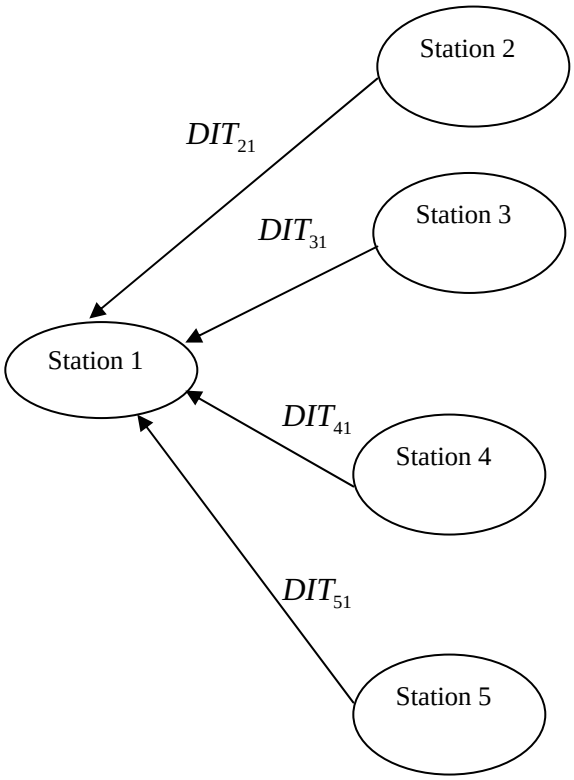


Figure 3.4 An example of potential base station candidates (e.g.2, 3, 4and 5) and a potential subject station 1.

3.3.5 Step 5: Creating artificial gaps for complete data sets

Before estimating the missing values, one has to see whether the selected station pairs have complete time-series or not (step 5). Thus, in case they are all complete, it was necessary to create artificially gaps with different durations, gaps in different seasons,

gaps in different climatic zones (e.g. 1 season, 2 seasons, 1 year, 2 years, etc) at the subject station. Hence, one could apply the methodology and proceed to the next step.

In the other case (e.g. data are not complete), it has to be checked whether the shortest record is really the subject (target) station. If this condition is satisfied, then one can proceed to the next step (i.e. interpolation or estimation of missing values). Otherwise, search for another station pair selected from the previous step. Recall that it is not always true that the short record should be considered as subject station (within its pair), when using entropy criterion (see, step 4), i.e. as long as the station (within its pair) satisfies condition (3.4), this specific station should be retained as base station.

3.3.6 Step 6: Filling in data by ANN and EM techniques

Step 6 is to apply the two categories (stream) of techniques, viz. EM techniques and ANNs and to make a comparison among these techniques and select the most suitable technique(s) for infilling (interpolating) the hydrological data. Standard techniques for EM and ANNs are incorporated into the model as well as some of their existing modifications. In addition, “new” versions for those techniques are also introduced in this study. These versions have been formulated intuitively. Then, the impact on the accuracy of the estimated values is also investigated.

3.3.6.1 ANN techniques

3.3.6.1.1 standard backpropagation (BP)

The algorithm has been given in the previous chapter and some features incorporated in the present model are explained in the following. The flow procedure of the standard PB is given below.

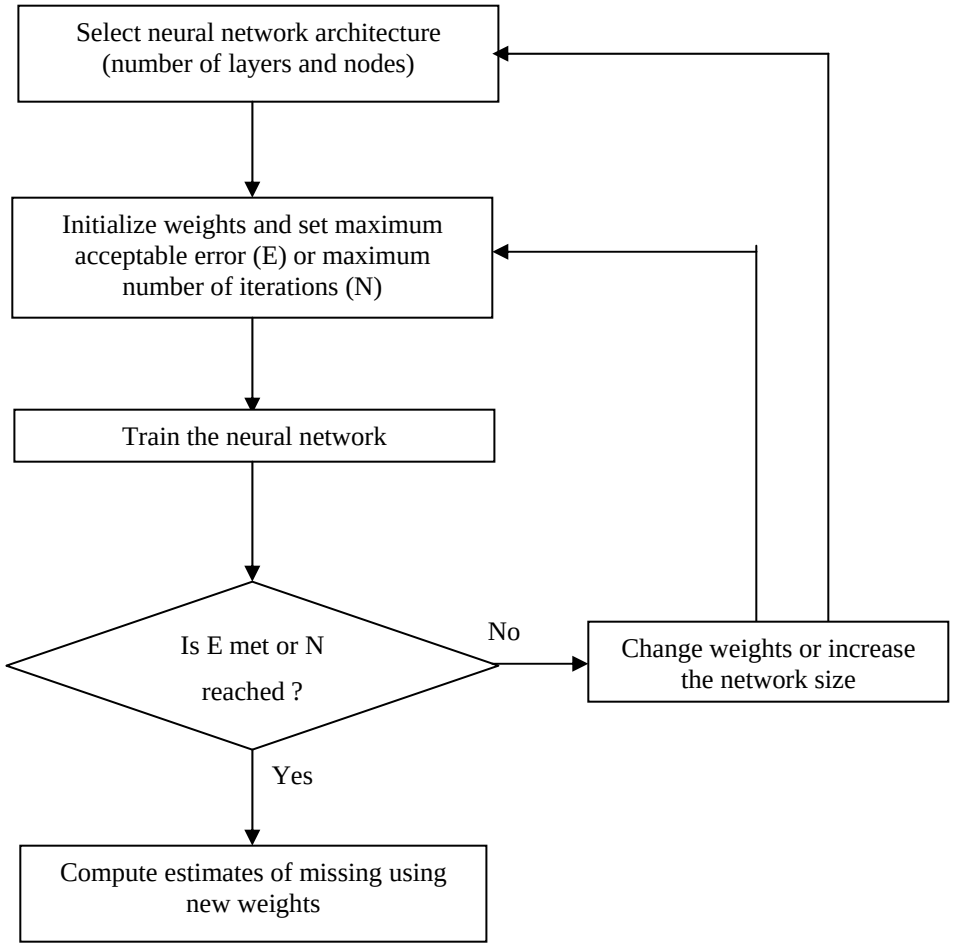


Figure 3.5 Standard backpropagation (BP) procedure

In Figure 3.5, the first step is the selection of the neural network architecture, starting from the simple to the complex one, e.g. 3 layers (1 hidden layer with 2 nodes, etc). Considering station pairs as outlined above, the number of nodes (in input layer) is the same as the number of nodes in the output layer, e.g. 1. The output layer may be non-linear (e.g. sigmoid, hyperbolic tangent or linear) while the input is always linear as no transformation occurs to the input nodes.

The second step in the figure above is the random initialization of the weights and the setting the error to a reasonable (acceptable) maximum value and the setting of number of

iterations to a maximum value N (e.g. N = 5000). This avoids an infinite loop. The error can be computed using formulas (2.75) as:

$$E = \sum_{k=1}^{P} E_p$$

The random initialization of the weights covers a wide range of values used by different authors both in hydrology applications and in other disciplines (Argawal and Singh, 2001; ASCE Task Committee, 2000; Patnaik et al., 1996; Freeman and Skapura, 1991; and others). The exiting ranges of initial weights found in the literature are different from one author to the other and are given by (-1.0, +1.0), (-0.1, +0.1), (-0.5, 0.5), (-0.3, 0.3), (0, +0.6), (0.6, +0.6), (-0.9, +0.9). *These ranges are among choices in this study*. Initial weights are retained if the convergence criterion is met, otherwise the weights are changed. Generally a 3-layered neural network is a starting point as this is the mostly used in hydrology as said in the previous chapter and it has been shown to be a general approximator (Zealand et el., 1999). This was also supported by Minns and Hall (1996), Lawrence (1996) and Raman and Sunilkumar (1995). At this step, output (input) may be normalized (scaled) as it has the advantage on the speed of the convergence of the system and it gives input equal importance and prevents premature saturation of the squash (sigmoid) function (Hines, 1997).

The model developed presents a choice among the mostly used techniques of scaling (standardization) and therefore decides on the optimum way of achieving this. The raison of this is that there are no fixed rules as to which approach should be used in particular circumstances and there has been very little research on the subject. Thus, the scaling technique by Dawson and Wilby (1998) could be used as follows

$$N_i = (R_i - Min_i)/(Max_i - Min_i) \tag{3.4}$$

$$N_i = R_i / \sqrt{SS_i} \tag{3.5}$$

where R_i is the real value to the node i; the subsequent standardized value calculated for node i; Min_i is the minimum value of all values applied to node i; Max_i is the maximum value of all the values applied to node i; and SS_i is the sum of squares of all values applied to node i.

The techniques proposed by Hines (1996) could be also used (e.g. linear scaling and a mean center unit variance scaling). The former scaling is given by

$$N_i = ((R_i - Min_i)/(Max_i - Min_i)) * 0.8 + 0.1 \tag{3.6}$$

where all the terms are already defined above and the latter scaling is as follows:

$$N_i = (R_i - \overline{X}r_i)/\sigma r_i \tag{3.7}$$

where $\overline{X}r_i$ and σr_i are the mean and the variance of all values applied to node i respectively. All other terms are already defined above.

It should be noticed that equation (3.6) was enough to be used in this study and to produce quite good results.

The mostly used activation functions in hydrology and other disciplines are incorporated in this model, e.g. sigmoid and hyperbolic tangent. Recall that these equations are differentiable everywhere for x values and are given by equations (2.62) and (2.64) as

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

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

Recall that for a 3-layered neural network, the update equations (2.71 and 2.72) for the output layer and the hidden layer are given in the previous chapter as:

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

$$w_{ji}{}^{h}(t+1) = w_{ji}{}^{h}(t) + \eta\delta_{pj}{}^{h} x_{i}$$

respectively.

The training of the neural network (i.e. finding the new weights) is done from the concurrent parts of data. The new weights are then used to compute the estimated values. The bias term for the ANN can be assumed to be zero as it use is optional (Freeman and Skapura, 1991). As said so far, the testing part can be unseen during the training and is assumed to be missing in case observed data were complete.

In case the sigmoid (or hyperbolic tangent) function is used, it is necessary to unscale the data (back to normality assumption), in such away one can perform different entropy calculations. Statistical performance criteria such RMSEp, RMEp and EVp were also performed to crosscheck the results.

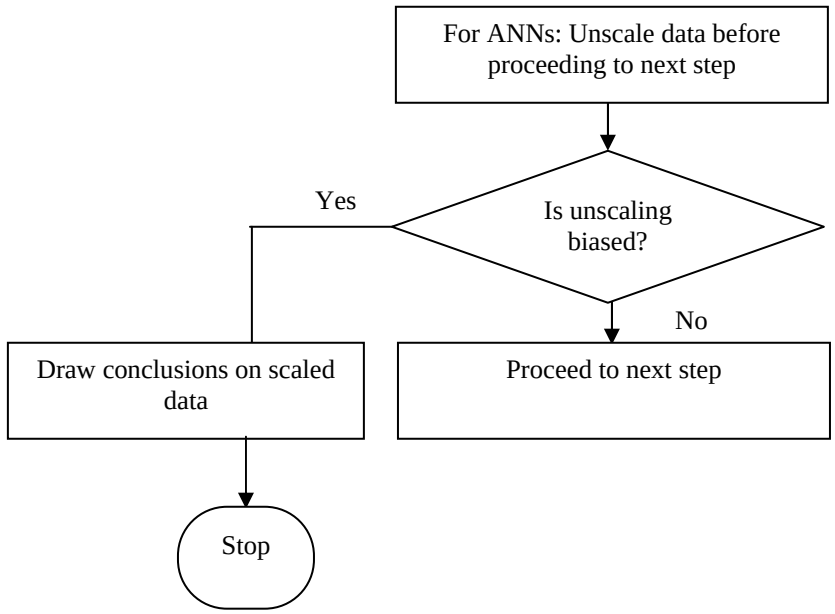


Figure 3.6 Unscaling for ANNs

The above remark applies to the rest of ANN techniques developed in this thesis.

3.3.6.1.2 Momentum Backpropagation (MBP)

This algorithm is explained in section 2.3.2.2.5.5.1. The procedure is similar as in the previous section but in " train neural network" step, the momentum term is added. Recall that the update equations (2.77) and (2.78) are given in the previous chapter as:

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

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

All the features for the standard BP can be used for MBP.

3.3.6.1.3 Variable Learning BP (VLR) module

The algorithm has been explained in section (2.3.2.2.5.5.2). The implementation is similar to the MBP algorithm. However, in "train neural network" step, a variable learning parameter is introduced according to equations 2.79 and 2.80.

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

$\eta = \eta * \rho \qquad (\rho \succ 1)$.

The weight update is accepted.

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

$\eta = \eta * \delta \qquad (\delta \prec 1)$

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. This a batch mode technique.

It is suggested in this study to use a wide range of the above parameters, i.e. δ, ρ, *including the above-proposed values. The additional features for BP are all included in the* VLR technique.

3.3.6.1.4 Generalized BP (GenerBP) module

The same applies as for VLR here but in "train neural network" step, the weight update is done according to equations 2.81 and 2.82 as

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

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

Ng et al. (1996) developed these equations only in the case of the squash (logistic) function, i.e. equation (2.62). These authors applied this technique to three different problems including XOR, 3-bit parity and 5-bit counting. No literature was found where the Generalized BP was used in hydrology or water related fields. *In this study, the Generlized BP technique incorporates also the hyperbolic tangent as activation function (see equation 2. 63). Several values for b could be tried. Only those values, which give an acceptable accuracy of the estimated values, could be retained.*

3.3.6.1.5 Quick backpropagation (QBP) module

The same applies here as for the standard BP technique, however in "train the neural network" step, the weight update is done according to equations 2.83 and 2.84 as:

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

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

$$\Delta\omega(t) = \alpha_1 \Delta w(t-1)$$

Before applying the QBP, the first values of weights were determined using the standard BP.

In this study, two options have been used for the QBP. The first option is named the Weight Condition (Weight Cond), where the update is done according following

$$
\begin{aligned}
&\textit{If} \quad Abs(\Delta\omega(t)) \prec \alpha_2 * Abs(\Delta w(t-1)) \quad \textit{Then} \\
&\quad \textit{If} \;\; s(t) \succ s(t-1) \qquad \textit{Then} \\
&\qquad \textit{Update according to equation } (2.83) \\
&\quad \textit{Else} \\
&\qquad \textit{Update according to equation } (2.85) \\
&\textit{End If}
\end{aligned}
$$

where α_2 is a factor greater than 1 and should be chosen heuristically and *Abs* means absolute value.

The second option is the Gradient Condition (Grad. cond) and is done as follows:

$$
\begin{aligned}
&\textit{If} \;\; s(t)) \prec s(t-1)) \quad \textit{Then} \\
&\qquad \textit{Update according to equation } (2.83) \\
&\quad \textit{Else} \\
&\qquad \textit{Update according to equation } (2.85) \\
&\textit{End If}
\end{aligned}
$$

3.3.6.1.6 Golden Search BP (GoldSBP) module

This is basically the linear search method. This is done by the procedure by locating the minimum of a function in a specific direction. This will involve two steps: interval location and interval reduction. The purpose of this interval location is to find some initial interval that contains a local minimum. The interval reduction step then reduces the size of the initial interval until the minimum is located to the desired accuracy.

In the interval location step, the error function is first evaluated at an initial point, which corresponds to the current values of the neural weights. The next step is to evaluate the function at a second point, which is at some distance from the initial point, along the first search direction. Then the error function is evaluated at new points, successively doubling the distance between points. This process stops when the function increases between two consecutive evaluations.

The next step in the linear search is interval reduction. This involves the function at points inside the interval was selected in the first step. Given an interval (a, b), if the function error E is evaluated at two different points c and d, the interval of uncertainty can be reduced. If $E(c) > E(d)$, then the minimum must occur in the interval (c, d), then the minimum must have occured in the interval (a, d). The Golden Section Search allows deciding how to determine the location of the interval points c and d. The Golden Section Search is designed to reduce the number of function error evaluations. The algorithm of this method can be for example can be traced in Hagan et al. (1996) and in Press et al. (1996).

3.3.6.1.7 Pseudo Mac Laurin order 1 BP and Mac Laurin order 2 BP (e.g. McL1BP and McL2BP) modules

These techniques are introduced for the first time. They are modifications to the standard BP, by approximating the sigmoid activation function by "pseudo" Mac Laurin power series order 1 and 2 derivatives. Thus, using the Mac Laurin series power series, the sigmoid (logistic) function of a variable x (e.g. scaled input data) can be approximated by:

$$f(x) = \frac{1}{1+e^{-x}} = \frac{1}{2 - x + \frac{x^2}{2} - \frac{x^3}{6} + \frac{x^4}{24} - \cdots} \tag{3.8}$$

The Mac Laurin power series (which is actually a particular case of a Taylor power series) approximate the function $f(x)$ when x approaches zero. In other words, for small values of x such that $0 < x <<< 1$, a good approximation of $f(x)$ can be achieved by a Mac Laurin power series. The Mac Laurin first order derivative approximation of equation (3.8) is given by:

$$f(x) \approx \frac{1}{2-x} \tag{3.9}$$

The derivative of equation (2.9) is given by:

$$f'(x) \approx \frac{1}{(2-x)^2} \approx (f(x))^2 \tag{3.10}$$

Equation (3.10) which, is an approximation to the first derivative of the logistic function can also be used in the weights update equations of the neural network.

The Mac Laurin second order derivative approximation of equation (3.8) is given by:

$$f(x) \approx \frac{1}{2-x+\frac{x^2}{2}} \tag{3.11}$$

The first derivative of equation (3.11) is given by

$$f'(x) \approx \frac{1-x}{(2-x+\frac{x^2}{2})^2} \approx (1-x)(f(x))^2 \tag{3.12}$$

Expression (3.12), which is an approximation to the first derivative of the logistic function, can also be used in the weights update equations of the neural network. Like the sigmoid function, equation (3.10) and (3.12) are also continuous, monotonic non-decreasing functions and differentiable on the interval of scaled input data or output data (0.1, 0.9). In the discussion of the results, sometimes the prefix “pseudo” is omitted.

For this study, no strict limitation on the range of values of x (e.g. x is greater than 0 but approaching 0) was set for the application of the Mac Laurin power series. However, the Mac Laurin power series approximation is just applied to an interval such that $0 < x < 1$, e.g. (0.1, 0.9) for scaled input and output data. That is why the prefix “pseudo” is introduced. The Mac Laurin (order 1 and order 2) approximation is done purposely for this interval just to evaluate the impact on the accuracy of the estimated missing values.

The reader would bear in mind that the formulas above don't have unit since they were applied to scaled streamflow or rainfall data (e.g. refer to equation 3.6).

3.3.6.2 Expectation maximization (EM) techniques

3.3.6.2.1 Standard EM (EM) module

The reader should remember that step 6 is still being processed, i.e. apply ANNs and EM techniques (see the detailed flowchart for model development).

Despite of the intensive use for problems involving incomplete data (Little and Rubin, 1987; and many others), the literature of EM techniques dealing with missing data still remains very sparse in hydrology and related fields. From the literature available, two studies involving missing data have been found in hydrology, i.e. Makhuvha (1997a, 1997b) and Kuczera (1987).

In fact, the latter considered the standard EM technique in a state-space framework compatible with the multi-site streamflow model and concluded that this technique is simple to implement and produces smoothed estimates of the missing data. The former deals with rainfall data patching and made a comparison between the standard EM and introduced a modification (to the standard EM) that he called pseudo-EM algorithm. Hence, a comparison of performance between the two techniques was conducted by computing the RMSE of the predictions.

This study does not include only the standard EM technique but includes its several modifications existing in the literature (not yet applied in hydrology) and encompasses other new features. The present research work takes also the opportunity of assessing the different EM techniques through entropy criterion.

As said before, the standard EM technique (EM) is the primitive form developed by Dempster et al. (1977). The EM algorithm and other modifications are particularly simple and useful interpretation when the complete data Y have a distribution from the regular

exponential family (Little et al., 1987). Thus, the normal distributions were specifically considered in this research work.

Since the EM technique is applied to the station pairs selected from step 5 and since the there is no missing values for the base gauging station (e.g. 1), its data values will remain unchanged throughout at any t-th iteration. Consequently the estimates of the mean and variance of that site (e.g. 1) will also remain unchanged throughout. The computations in the t+1 iteration of the estimates at the subject site (e.g. 2) is based on equations (2.47-2.50), that is

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

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

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

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

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

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

$$\beta_{21.1}{}^{(t+1)} = \sigma_{j2}{}^{(t+1)} / \sigma_{11} \qquad for \;\; j = 1$$

The above steps are repeated until reasonable convergence is achieved.

Convergence is normally judged by examining changes in individual components of $\theta = (\theta_1, \theta_2, ..., \theta_k)$ from one iteration to the next. The criterion of convergence used is as follows (Kuczera, 1987; Schaffer, 1994):

$$\left|\theta^{(t+1)}{}_j - \theta_j{}^{(t)}\right| \leq \varepsilon \left|\theta_j{}^{(t)}\right| \tag{3.13}$$

for $j = 1, 2, ..., k$ and a suitable ε (tolerance). In the present case, $\theta = (\mu_2, \sigma_{22}, \beta_{21.1})$.

Hence convergence tolerance can be set or a maximum number of iterations should also be set. The impact on the accuracy of the missing values could be finally determined.

To start the iterations, one needs to give a value to $\theta^{(0)}$. However, the Buck'method could be used in this study for the starting value of $\theta^{(0)}$ (refer to equation 2.51, Chapter 2).

3.3.6.2.2 Momentum EM (MEM1) module

No literature is available where MEM technique is used in hydrology and water related fields. It has been taken the opportunity of applying it in this study. This algorithm has been described in section 2.3.2.6 of Chapter 2. The parameters are updated according to equation 2.58, that is,

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

For an iterative algorithm with a current incremental $\Delta\theta = \theta^{(t+1)} - \theta^t$, one can always modify the obtained $\theta^{(t+1)}$ into $\eta\theta^{(t+1)} + (1-\eta)\theta^t$. The momentum can be chosen heuristically to speed up the convergence (Melijson, 1989). This author did not propose

any specific value for the momentum. Xu (1997) proposed a value such that $\eta \succ 0.5$ while in this thesis a wide range of the values for the momentum were used and the optimum value could be the one which gives the accurate results of the estimated missing values. In the following, the second and third versions of the MEM1 module are proposed.

3.3.6.2.3 Second and third versions of the MEM algorithm (i.e. MEM2 and MEM3 modules)

In this thesis, a second version and a third one (i.e. MEM2 and MEM3) of the MEM1 technique are proposed. These two versions are more based on a consideration which is more intuitive than a strong mathematical basis. The starting point of these two modifications to the original MEM1 technique is that a second term is added to the update above equation 2.83 and thus the accuracy of the estimated values is investigated. This second term is taken to be proportional to the previous change in the parameter, i.e. $\Delta\theta^{(t-1)} = \theta^{(t)} - \theta^{(t-1)}$ in the update equation. In this case, the update equation can be written as follows:

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

$$\theta^{(t+1)} = \theta^{t} + \eta * \Delta\theta(t) + \alpha\Delta\theta(t-1) \tag{3.14}$$

where α is just a coefficient of proportionality.

The two versions are as follows:

<u>a) MEM2</u>

-For the iterative algorithm with the current increment $\Delta\theta = \theta^{(t+1)} - \theta^{t}$, the obtained regression coefficients are modified into $\eta\beta^{(t+1)} + (1-\eta)\beta^{(t)} + \alpha\Delta\beta^{(t-1)}$ while the other individual parameters (e.g. mean and variance) remain unchanged at the subject station.

b) MEM3

-For the iterative algorithm with the current increment, $\Delta\theta = \theta^{(t+1)} - \theta^{t}$ the obtained statistical parameters, e.g. variance, mean are modified into $\eta\mu^{(t+1)} + (1-\eta)\mu^{(t)} + \alpha\Delta\mu^{(t-1)}$ and $\eta\sigma^{2(t+1)} + (1-\eta)\sigma^{2(t)} + \alpha\Delta\sigma^{2(t-1)}$ respectively. However, the other individual parameters, e.g. regression coefficients are not modified.

In either case, the accuracy of the estimated values is investigated.

3.3.6.2.4 Expectation constrained maximization (ECM) module and its Version

As outlined in section 2.3.2.1.7, the M-step is replaced by a series of constrained (conditional) steps. In this study, it is not argued about the complexity of the complete-data maximum likelihood estimation to apply the ECM algorithm. Hence, it is believed that the CM are over small dimensional spaces, often they are simpler and, faster and more reliable than the corresponding full maximization called for in the M-step of the EM algorithm (Meng and Rubin, 1993). However, the most important thing is to apply it to hydrological missing data problem and to compare its performance with the standard EM or other data interpolation (infilling) techniques. It is reminded that no literature is available where the ECM technique has been applied to hydrology or related fields, specifically to problems dealing with missing hydrological data.

From Meng and Rubin (1993), the ECM algorithm performs as follows:

-Having the mean and variance of the target station, the regression coefficient β are estimated first through equation 2.15, which is the first CM-step

$$\beta^{(t+1)} = (\sum_{i=1}^{n}(y_{i2}{}^{(t+1)} - \hat{\mu}_2{}^{(t)})(y_{ij} - \hat{\mu}_j))/\sigma_j \quad for \quad j=1 \tag{3.15}$$

Given $\beta = \beta^{(t+1)}$, the <u>second CM-step</u> is carried out by computing the conditional maximum likelihood estimate of the other individual parameters, e.g. mean, variance and covariance are then estimated. The CM-steps are performed until convergence.

The algorithm described above can be seen as first version named ECM1 module. The second version (ECM2) of the above algorithm is base just on an intuitive consideration, i.e. is one can always re-compute $\beta^{(t+1)}$ using the result from the above-mentioned *second CM-step*.

3.3.6.2.5 Expectation Constrained Maximization Either (ECME1) module and its versions

No literature was found where this technique is used in hydrology or water related fields, specifically for data interpolation (infilling) problems. The ECME was explained in section 2.3.2.1.8 and leads basically to cases where CM-steps maximizes either the expected complete loglikelihood, as with ECM, or the actual likelihood function subject to the same constraints. A multi-cycle version of ECM which is obtained by performing a second E-step before the second CM-step to find the expected complete loglikelihood is an example of the ECME algorithm (Meng and Rubin, 1993).

The versions here are also based on intuitive considerations rather than strong mathematical proofs, similarly to the other versions introduced in the previous sections.

Version 1

This option is performed similarly as in the previous section both for E-step and CM-steps, however *each CM-step is preceded by an E-step.*

Version 2

The second option is to consider the introduction of the momentum term into this version of the ECME as set out above. At each CM-step of the iterative algorithm with the

current increment, $\theta^{(t+1)}$ the obtained parameter, e.g. variance, mean at the subject station are modified into $\eta\theta^{(t+1)} + (1-\eta)\theta^{t} + \alpha\Delta\theta^{(t-1)}$

Version 3

Performing once CM-step for the regression coefficient as above for each selected station pair is done. Then, the obtained value of the regression coefficient is used to perform several CM-steps (2, 3, 4, etc) with respect to the statistical parameters e.g. mean, variance, and covariance. Each CM-step is preceded by an E-step.

For each version, the accuracy of the estimated values is also investigated.

3.3.7 Step 7: Technique performance assessment

This step is to assess the performance of the techniques for interpolating (infilling) hydrological data gaps to achieve an optimum agreement between computed and observed data.

Step 7.1

According to the detailed flow chart depicted in Figure 3.1, in a case where the station with short record has been infilled and no missing values have been created artificially, the reduction in uncertainty at the predicted station, via a given technique, is the only criterion used here (Panu, 1992). It is defined by equation 2.90 as

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

This equation can be used for cases where the values of marginal entropies for the target station, e.g. H_{cc} and H_{comp} are positive before and after infilling respectively. A slight modification can be introduced in the above formula in case these values are negative. As negative information does not have any physical meaning, entropy values should be considered in absolute coordinates in which, the origin is set at minus infinity. Thus

entropy values are no longer negative and regain their physical meaning. Thus, in this study, the above formula can be modified into

$$\mathrm{Re}\,d = 100 * (H_{cc} - a - H_{comp} + a)/(H_{cc} + a)$$

which is equivalent to

$$\mathrm{Re}\,d = 100 * (H_{cc} - H_{comp})/(H_{cc} + a) \qquad (3.16)$$

where a is defined here as a translation parameter which enables to keep always the reduction in uncertainty between 0 and 100%.

Given a set of data, the computation of the reduction in uncertainty is carried out for all infilling techniques used in the previous step. Therefore, the only techniques, which are selected, are those for which $\mathrm{Re}\,d$ values satisfy the following condition

$$\mathrm{Re}\,d \geq Threshold\ 2 \qquad (3.17)$$

The value of Threshold2 was chosen in this study such that at least 30 % of uncertainty should be removed from the subject station via a technique, after the data series at this station has been infilled. According to Panu (1992), the values of reduction in uncertainty at the subject station via the different models he used ranged from 5% to 79%.

It should be noticed that runoff simulation models were assessed through the extension of Panu's formula (refer to Chapter 8) for cases where no data was absolutely available at the subject site.

Step 7.2

In cases where missing values have been created artificially, the following is performed: The entropy criterion could be performed using equations 2.86 and 3.19:

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

$$DIT_{OBS/EST} = T(X,Y) \big/ H(X) \tag{3.19}$$

In formula 3.19, the $DIT_{OBS/EST}$ is the directional information transfer index between the observed and the estimated series. In this study, the DIT notion as formulated by Yang and Burn (1994) is extended to models evaluation criterion. The DIT as generalization of the transinformation (T) was used for deciding on the dependency of station pairs. Here, the DIT notion is used as a generalization of the T for model performance evaluation since the transinformation (T) can be used for that purpose. DIT was used as positively non-dimensional information index, which varies between 0 and 1 (e.g. between 0% and 100% in terms of percentage). This was applied in recent papers (Ilunga and Stephenson, 2002; 2003a). Given a set of data, the DIT computation is carried out for all data infilling techniques used in the previous step. Therefore, the only selected techniques are those for which DIT values satisfy the following condition

$$DIT \geq Threshold\,3 \tag{3.20}$$

The value of Threshold3 was chosen in this study such that at least 30 % of uncertainty should be removed from the subject site via a technique. In a study led by Chapman (1987), the values of reduction in uncertainty for the different models ranged from 18 % to 46 %.

Since the DIT notion was used as a general criterion to compare technique performance on different data sets, it was possible to consider the same origin of coordinates for entropy values. In the present study, this origin could be set to the highest transinformation for the different techniques.

For each subject station, if entropy criterion is not fulfilled, then the technique is not selected and therefore cannot be used to fill in the missing data at that specific station. The same is repeated for each technique and every target station.

The following statistical criteria (see equations 2.91-2.93) are used just to crosscheck whether the results from the entropy criterion are reasonable. These criteria can be also performed after step 8.2.

(i) $Root\ Mean\ Square\ Errors\ of\ predictions\left(RMSEp\right) = \left[\dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{i}\right]^{1/2}$

$(ii)\ RME = \dfrac{1}{n}\sum_{i=1}^{n}\dfrac{\left|\hat{y}_i - y_i\right|}{y_i}$

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

These criteria are made on the predictions of the series and the summation can be done over those predictions.

A part from the numerical performance indicators above, the following graphical indicator has to be given below:

(iii) A scatter plot of the simulated (interpolated) versus observed data.

3.3.8 Step 8.1: Transformation back to original data

Step 8.1

Now the question that arises is to know whether raw data have been forced so far to follow approximately the normality assumption through Box-Cox transformation (see step 2). If that is the case, thus one has to untransform the current results through inverse of set of expression (3.2), i.e.

$$x = \begin{cases} 1/y, & \lambda = 1(inverse) \\ e^{y}, & \lambda = 0\left(anti\log arithmic\right) \\ y^{2} & \lambda = 0.5(square) \\ y, & \lambda = 1(untransformed) \\ \sqrt{y}, & \lambda = 2(square\ root) \\ \sqrt[3]{y}, & \lambda = 3(cubic\ \ square) \\ (1+\lambda y)^{1/\lambda}, & if\ \lambda\ \ is\ not\ \ any\ of\ \ above\ \ values \end{cases} \tag{3.21}$$

If this transformation back original data does not introduce any bias (e.g. negative values) in the estimates or eventually in the parameters, the results are considered to be satisfactory and the algorithm ends. Otherwise, conclusions can still be drawn on transformed variables as it has been done in many hydrological studies, e.g. Minns and Hall, (1996) and many others.

On transformation back to the initial space, one has to ensure that the infilled series contains no negative values.

Step 8.2

Steps 8.1 and 8.2 are mutually exclusive. In step 8.2, raw data followed the normality assumption and therefore, if this is the case, the algorithm is terminated and step 8.1 is not done.

After the missing parts have been estimated, a comparison between estimated (interpolated) and observed values, is made to judge the accuracy of the estimates. Although the overriding objective in infilling or estimating missing data is to minimize the squared error between observed and simulated data (Panu et al., 2000, Elshorbagy, 2000a), there are some requirements, namely, that the estimated values should not introduce systematic bias in the statistics of importance in the record (Zucchini et al., 1984). In particular, the mean and variance should not be systematically distorted by the estimates. Thus, the infilled series can be used for other purposes, say reservoir design, water resources development, reestablishment of reservoir operation rules, hydropower

development, estimation of some statistics of the annual yields of reservoir, flood control, planning of storage projects, evaluation of the severity and duration of hydrological extremes, etc. Hence, the potential applications are also taken into consideration and therefore a range of proportion of missing data is determined beyond the statistical properties (e.g. variance, mean). The degree of uncertainty is reduced so much a lot and this can affect the above-mentioned purposes (Stephenson, 2003).

3.4 DATA AVAILABILITY

3.4.1 Introduction

The methodology as explained in this chapter was tested on selected streamflow and rainfall gauges of South Africa (Midgley et al., 1994). These gauges belong to the Orange drainage system rivers. Rainfall and streamflow data of this drainage system used in Chapters 5, 6 and 8 are briefly described in this chapter, except for data used in Chapter 7. The hydrological data used in Chapter 7 are briefly explained in the same chapter. It was convenient to do so, as these data don't belong to the Orange system rivers.

3.4.2 Physical characteristics of the Orange drainage system rivers

The Orange River rises in the sovereign Kingdom of Lesotho, draining the Maluti Mountains and the western slopes of the Drankesberg range. Its major tributary is the Vaal.

Along the highest yields of sediments in the country are those encountered in the Cave sandstone formations of the upper Caledon and in the Kraai catchment as well as the southern watershed of the Orange River.

Listed in Table 3.1 are total areas of catchments of the Orange River System as well as catchment-averaged MAP (mean annual precipitation) and MAR (mean annual riverflow). Since the Orange drainage river system has known major developments in irrigation and hydropower), this table contains the summary of the information relating to irrigation, afforestation, and storage utilization in the Orange system rivers.

Table 3.1 Total catchment area, mean annual precipitation and mean annual riverflow of the primary Orange River system

Primary drainage region: Orange (D)					
Catchment area (km^2)	MAP (mm)	MAR (Mm^3)	Total irrigation area (km^2)	Total afforestation area (km^2)	Total reservoir capacity
409621	315	6987	746	-	9507

3.4.3 Secondary drainage river systems considered for this study

Some selected streamflow gauges and rainfall stations belonging respectively to the secondary drainage D1 and D33 of the Orange drainage system rivers (D) were considered for this study. The monthly data could be found from Midgley et al. (1994). The geographical location of the selected rivers is depicted in Figures A.1 and A.2 (refer to Appendix A) and Tables 3.2 and 3.3. The geographical location of the selected rainfall stations is depicted in Figure A.3 (refer to Appendix A) and Tables 3.4 and 3.5. The mean monthly rainfall data (in Table 3.4) were obtained by multiplying the MAP (in mm) by the monthly rainfall as percentage of MAP. Note that the rainfall stations were just coded and no names were given to them according to the report by Midgley et al. (1994).

The hydrological year starts in October and ends in September for the data used (i.e. rainfall and streamflow).

It has to be said that the subcatchment for the three gauges namely D1H003, D1H006 and D1H009 belong to the Upper Orange catchment. Gauge D1H003 named Aliwal North and D1H009 named Oranjedraai are both situated on the Orange River while D1H009 named Maghaleen is situated on the Kornetspruit River, which is a confluence with the Orange River. The distances between D1H003 and D1H009 and D1H006 and D1H009 are approximately 8.8 km and 2 km respectively. Gauge D1H003 is in the Aliwal North Town of South Africa and is at the latitude 30^0 40' 47" and the longitude 26 ^{0} 42' 45". Gauge D1H006 is in the Maghaleen Town of South Africa and is at the latitude 30^0 09' 37" and the longitude 27 ^{0} 24' 06". Gauge D1H009 is in the Free State at the latitude 30^0 20' 10" and the longitude 27 ^{0} 21' 34". The Orange River rises in the sovereign Kingdom

of Lesotho, less than 200 mm and flows West, with its wide sweeps and North, to the Atlantic. It drains, with its tributaries, an area estimated over 400, 000 m^2, passing through more than 12 degrees longitude or 750 m, a straight line from source to mouth. The headstreams of the Orange River are in the highest part of the Drankesberg range, the principal source, the Senku, rising, at an elevation of more than 10, 000 feet (or 305 m), on the South face of the Mont aux Sources in 28 ^{0} 48' East and 28 ^{0} 50' South. The headstreams are South East of the Senku source, in Champagne Castle, Giant's Castle and other heights of the Drankesberg. Rising in the inner slopes of the hills, these rivulets all join the Senku, which receives from the north several streams which rise in the Maluti Mountains. After a course of some 200 m, passing the South West corner of the Maluti Mountains, the Seku, already known as the Orange, receives the **Kornetspruit** (90 m), which rises in the Machaba Mountain. The Orange River here enters the Great inner Plateau of South Africa, which at **Aliwal North**, the first town of any size on the banks of the river, 80 m below the **Kornet Spruit** Confluence, has an elevation of 4300 feet (or 1300 m).

As is to be expected in a river that traverses practically the full width of the subcontinent, there is a wide range of topography. With its sources in the high mountains of Lesotho, the Orange River remains deeply incised in the interior plateau until it reaches the main irrigation areas from Buchuberg to Kakamas. Then at Aughrabies Falls, the river plunges into a deep canyon and winds its way through the broken country of the Richtersveld to emerge on a broad stretch of desert across which it meanders to the sea.

The soils of 3 gauges in the upper Orange River catchment are lithosols, solonetzic and montmorillonitic clays while along the lower Orange, the soils are mainly lithosols. The catchment for the three gauges in the upper Orange catchment is dominated in the by pure grassveld.

Rainfall is high in the upper Orange, MAP (mean annual precipitation) reaching 2000 mm in the Drankesberg and decreasing sharply to 400 mm at the confluence of the

Caledon with the Orange. Evaporation increases along with the Orange River from the 1300 mm in the east to 2700 mm at the confluence with the Molopo.

Table 3.2 Geographical location of selected rivers of the secondary drainage D1

Secondary drainage D1								
Gauge	**Name**	**River**	**Latitude**	**Longitude**	**Area (km^2)**	**Period of records used**	**Mean annual riverflow (Mm^3)**	**% of Missing**
D1H001	Diepkloof	Wonderboomspruit	$31^0$00’11’’	$26^0$21’11’’	2397	1924 –1953	45.51	0
D1H003	Aliwal North	Orange River	$30^0$40’47’’	$26^0$42’45’’	37075	1960 – 1989	5170.27	0
D1H004	Molteno	Stombergspruit	$31^0$24’00’’	$26^0$22’17’’	348	1924 – 1953	6.89	0
D1H006	Maghaleen	Kornetspruit	$30^0$09’37 ”	$27^0$24’06’’	2969	1960 – 1989	566.44	0
D1H009	Oranjedraai	Orange River	$30^0$20’10’’	$27^0$21’34’’	24550	1960 – 1989	4212.14	0

Table 3.3 Mean monthly flows for selected rivers of the secondary drainage region D1

Mean monthly flows (Mm^3)												
Gauge	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.
D1H001	1.57	3.02	2.99	4.22	6.51	11.75	4.97	3.69	1.04	0.51	2.92	2.30
D1H003	407.8	546.9	537.6	586.2	888.8	757.8	556.1	261.4	178.6	111.9	124.9	212.4
D1H004	0.24	0.77	0.71	0.74	0.89	1.45	0.86	0.66	0.12	0.19	0.05	0.21
D1H006	41.98	45.02	59.54	72.24	87.58	81.51	77.47	30.97	18.17	13.10	14.21	24.66
D1H009	355.9	459.9	429.6	493.5	710.8	603.5	429.3	188.2	134.1	92.45	109.0	206.0

Table 3.4 Geographical location of selected rainfall stations of the secondary drainage D33

Secondary drainage D33							
Gauge	**Section**	**Position**	**MAP (mm)**	**Latitude**	**Longitude**	**Period of records used**	**% of Missing**
0228170	288	170	341	$29^0$50’00’’	$24^0$36’00’’	1924 -1989	0
0228458	228	458	348	$29^0$38’00’’	$24^0$46’00’’	1924 -1989	0
0228495	228	495	376	$29^0$45’00’’	$24^0$47’00’’	1924 -1989	0

Table 3.5 Mean monthly rainfall for selected stations of the secondary drainage D33

	Mean monthly rainfall (mm)											
Station	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.
0228170	23.59	34.92	33.32	39.62	56.67	60.39	38.94	16.54	7.63	6.58	10.47	12.48
0228458	23.98	35.64	34.0	40.44	57.84	61.63	39.74	16.88	6.16	6.72	10.68	12.74
0228495	25.57	38.50	36.74	43.69	62.49	66.59	42.94	18.24	6.66	7.26	14.14	13.76