CHAPTER 9

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

9.1 SUMMARY AND CONCLUSIONS

The problem of missing data, insufficient length of hydrological data series and poor quality is common in developing countries. This problem is much more prevalent in developing countries than it is in developed countries. This situation can severely affect the outcome of the water systems managers' decisions (e.g. reliability of the design, establishment of operating policies for water supply, development of water resources, etc). Thus numerous data interpolation (infilling) techniques have evolved in hydrology to deal with the missing data.

The current study presents merely a methodology by combining different approaches and coping with missing (limited) hydrological data using the theories of entropy, expectation-maximization (EM) and artificial neural networks (ANN) techniques. This methodology was simply formulated into ENANNEX model. The entropy concept was confirmed to be a versatile tool. This concept was firstly used for quantifying information content hydrological variables (e.g. rainfall or streamflow). The same concept (through directional information index, i.e. DIT) was used in selection of base/subject gauge. Finally, the DIT notion was extended to the evaluation of the hydrological data infilling technique performance (e.g. ANN and EM techniques). Thus the validity of these data infilling techniques with respect to the different gap sizes (duration) could be defined through entropy concept. The methodology was tested on annual mean flow series; annual maximum flows, annual total rainfall; and 6-month flow series (means) of selected catchments in the drainage region D "Orange" of South Africa. These data regimes can be considered as useful for design-oriented studies, flood studies, water balance studies, etc. The results were presented and discussed.

The results from the case studies showed that DIT notion is as good index for hydrological data infilling technique selection as other criteria. The results from entropy calculations were crosschecked with other criteria such as statistical and graphical. However, DIT has its own feature of being a non-dimensionally informational index. The DIT notion could enable to compare data infilling techniques on different catchment areas. The data infilling techniques viz. ANNs and EM (existing techniques applied and not yet applied in hydrology) and their new features have been also presented. The following feedforward ANN techniques were then used: standard BP, MBP, GenerBP, VLR, GoldSBP, QBP, McL1BP and McL2BP (refer to Chapter 3). The following EM techniques were used: standard EM, MEM1, MEM2, MEM3, ECM1, ECM2, ECME1, ECME2 and ECME3 (refer to Chapter 3).

This study showed that the standard techniques (i.e. standard BP and standard EM) as well as their respective variants could be selected in the hydrological missing estimation process. However, in some cases, the respective variants of the selected techniques could show higher capabilities (than the standard techniques) in the interpolation (estimation) of missing values. Both accuracy of the estimated values and statistical requirements at the subject gauge could be fulfilled. From this study, the relationship between accuracy of the estimated series and gap duration (at the subject station) was then investigated through the DIT notion. Generally, it was shown that a decay (power or exponential) function could better describe that relationship. Thus, for a given hydrological data infilling technique and for given gap duration, it was possible to predict the accuracy of the estimated values at the target station. From this research work, it was concluded that the performance of the different techniques depends on the gap duration at the target gauge, the station-pair involved in the missing data estimation and the type of the data regime.

This study showed also that it was possible, through entropy approach, to assess (preliminarily) model performance for simulating runoff data at a site where absolutely no record exist: a case study was conducted a Bedford (South Africa) site where a dam is proposed to be built in the future.

This study does not use any physical characteristics of the catchment areas but deals only with the limited information (i.e. streamflow/rainfall) at the target gauge and its nearby similar base station(s).

For annual mean flow data infilling case study, it was shown that the gauges in the station-pairs D1H003-D1H009 and D1H009-D1H003 could infer mutually information (contained in the data series) about one another. In other word, when one gauge could be considered as predicted gauge, the other one was the predictor gauge and vice-versa. Those 2 stations could finally be considered to fill in missing values mutually. Gauge D1H006 was selected as potential (predictor/predicted) station for both gauges D1H003 and D1H009. However, gauge D1H006 could not finally be considered in the interpolation process of the missing annual mean flows. The reason is that the entropy criterion (e.g. 30 % of uncertainty to be removed from the subject gauge) was not satisfied for the different data infilling techniques.

For annual mean flow series, the results for gap duration of 6.7 %, 13.4 % and 20 % at D1H003 (starting from 1965) were discussed as the different data interpolation techniques (i.e. standard BP and EM and their respective variants) could perform well. Therefore, the rest of the gap duration (i.e. 30 %) was not part of the discussion at this specific gauge. Nonetheless, these values were also used to investigate the relationship between gap duration and accuracy of estimated values for the different techniques. For D1H009 taken as target station, the whole range of gap duration (i.e. 6.7 %, 13.4 %, 20 % and 30 %) was discussed as the results for technique performance were shown to be satisfactory.

From DIT notion and for a given gap duration, it was shown that D1H003 is in general slightly better predictor for the estimation process of missing annual mean flows at D1H009 than when gauge D1H009 is used to fill in missing data at D1H003. Nonetheless D1H009 could be used to fill in the missing annual mean flows at D1H003.

It was also noticed that the DIT between observed and estimated values at the subject station generally decreases when the proportion of missing annual mean flow values increases. In other words, the proportion of information physically transferred by the knowledge of the estimated series into the process to make (the annual mean flow series at the subject gauge) better defined, will decrease as the gap duration increases at the subject station.

Using D1H009 as subject station, a decay power function could describe better (than exponential or linear) the relationship between the gap duration (i.e. range from 6.7 % to 30 %) and the technique accuracy (i.e. DIT). Considering D1H003 as subject gauge, a similar relationship could also be established. Thus, for a given technique, it was possible to find approximately the expected accuracy of the estimated values at the subject gauge when the gap duration is known (e.g. between 6.7 % and 30 %). It was noticed that an earlier start (e.g. 1963) or later start (e.g. 1970) for the gaps created on the records of the subject gauge did not have any substantial impact on the accuracy of the estimated values.

For annual maximum series interpolation, it was shown that the gauges in the station-pairs D1H003-D1H009 and D1H009-D1H003 could infer mutually (the information contained in the data series) about one another. When one gauge could be considered as predicted station, the other one was predictor station and vice-versa. Only the results for gap duration of 6.7 % (starting from 1965) annual maximum flows at either station were discussed as the performance of the different techniques could satisfy the entropy criterion. The results from the rest of gap duration didn’t give satisfactory results and were not, therefore, discussed. Nonetheless, these values were also used just to investigate the relationship between gap duration and accuracy of estimated values for the different techniques.

Using DIT as technique performance on different catchment areas, it could be shown that D1H003 was in general a better predictor for the estimation process of missing annual maximum flows at D1H009 than when gauge D1H009 is used to fill in data at D1H003.

It was also noticed that the directional information transfer index between observed and estimated values generally decreased when the proportion of missing annual maximum flows increased. In other words, the proportion of uncertainty removed by applying a data infilling technique will decrease as the gap duration increases at the subject station.

Generally, for annual maximum flows, a decay power function could describe better (than exponential or linear) the relationship between the gap duration (e.g. ranging from 6.7 % to 30 %) and the technique accuracy in terms of DIT. It was noticed that an earlier start (e.g. 1963) or later start (e.g. 1970) for the gaps created on the records of the subject gauge did not have any substantial impact on the accuracy of the estimated values.

For annual total rainfall series, it was shown that the gauges within each respective station-pair (0228170-0228495 and 0228495-0228458) could infer mutually information (contained in the data series) about one another. The results from station pairs 0228495-0228458 were the only ones to be discussed. The results for the station-pairs 0228170-0228495 and 0228495-0228170 were just depicted in Appendix C as the conclusions drawn from these results could be similar to those drawn from the other above mentioned two station pairs. For the station pair 0228495-0228458, the results for the gap duration of 7.6 % and 13.6 % (starting from 1935) in annual total rainfall at either station were the only ones to be discussed. This was done so, as the entropy criterion was satisfied with regard to the performance of the different techniques for those proportions of missing values. At the same time, the statistical requirements could be fulfilled at the subject station. The results from the rest of the gap duration (e.g. 19.7 % and 30.3 %) were not satisfactory results and were not therefore discussed. Nonetheless, these values (19.7 % and 30.3 %) were used in the investigation of the relationship between gap duration and accuracy of estimated values for the different techniques.

Considering the values for DIT, the rainfall data inflling techniques could be also compared for different catchment areas. In general, it could be shown that gauge 0228458 was a good predictor in the estimation process of missing annual total rainfall at

0228495 and vice-versa. Generally, a decay power function could describe better (than exponential or linear) the relationship between the gap duration (from 7.6 % to 30% missing annual total rainfall) and the technique accuracy in terms of DIT. It was noticed that an earlier start (e.g. 1930) or later start (e.g. 1970) for the gaps created on the records of the subject gauge did not have an substantial impact on the accuracy of the estimated values. It was also noticed that the directional information transfer index between observed and estimated values generally decreased when the proportion of missing annual total rainfall increased. In other words, the proportion of uncertainty removed by applying a data infilling technique will decrease as the gap duration increases at the subject station.

For the 6-month flow series (means), e.g. seasonal mean flows, the results for 6.7 %, 13.3 %, 20 % and 30 % (starting from 1924) were discussed. It was shown that the gauges with the station pair (D1H001-DH1004) could mutually infer information about one another. This was possible with a threshold value of 30 % for DIT. However, gauge D1H001 was found to be potentially far better predictor than D1H004. Therefore, the former was taken as control gauge and the latter as target gauge. For this specific station-pair, the “pseudo” Mac Laurin power series approximation order 1 and 2 derivatives (to the sigmoid function for feedforward BP ANN) did not affect substantially the accuracy of the estimated values at gauge D1H004, when compared to the standard BP. It was also observed that an exponential function could describe a strong relationship between the gap size and the expected DIT for the pseudo Mac Laurin order 1 and 2 (i.e. McL1BP and McL2BP) and the standard BP. It was noticed that an earlier start (e.g. 1928) or later start (e.g. 1938) for the gaps created on the records of the subject gauge did not have any substantial impact on the accuracy of the estimated values.

Generally, increasing the number of data points (e.g. seven values for gap size at the subject gauging station) did not sensitively affect the relationship between accuracy and gap size for the different types of data regime.

The results from entropy calculations showed that both models (i.e. RAFLER and WRSM2000) could be used for simulating the annual total flows at a site where absolutely no record exist, i.e. Bedford. Nonetheless, RAFLER model could perform better (than WRSM2000 model).

9.2 RECOMMENDATIONS

From the above, it is recommended the following:

-More case studies for South Africa should be conducted and other data regimes (4-month flow data series, low flows, etc) should be also considered. More winter rainfall and summer rainfall catchments should be specifically targeted. Catchments of other developing countries could also be taken as case studies.

-The methodology in this study could be extended to cases where missing data are encountered at both sites, e.g. control and target sites. In these cases, a *"mixed"* interpolation (infilling) could be undertaken in a similar way of Alley and Burns (1983). Accordingly, the ANN and EM techniques as well as the entropy approach could be readjusted.

-The testing of *"mixed models"* for EM techniques should be conducted. This could take into account the unobservable factors in the missing data (i.e. rainfall or streamflows). Thus, the impact of unobservable factors on the accuracy of the estimated series could be investigated.

-Formulation of a notion similar to DIT, which can be applied to a group of three, four etc gauges so that the contribution for each predictor (control) station to the global uncertainty removed from the predicted (target) station, via a given technique, could be investigated in the missing data estimation process. A this specific stage, for example a target flow gauge with its "own" basin, then with streamflows in adjacent basins and then with rainfalls in and near its own basin should be considered. Recall that the DIT notion

as used in this thesis was applied only to station pairs according to Yang and Burn's definition (1994).

- Sensitivity analysis should be thoroughly conducted to investigate the impact of the different starts of missing values on the performance of the different hydrological data infilling techniques.

- Sensitivity analysis on the accuracy of the estimated data series using pseudo Mac Laurin power series (BP ANNs) of higher orders (e.g. 3, 4, etc) should be conducted. The sigmoid activation function as well as the hyperbolic tangent activation function could be used.