HYDROLOGICAL DATA INTERPOLATION USING ENTROPY

Masengo Ilunga

A thesis submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, in fulfilment of the requirements for the degree of Doctor of Philosophy

Johannesburg, 2004

DECLARATION

I declare that this thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

____

Masengo Ilunga

________________________ day of _______________ (year)______

ABSTRACT

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, 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 **en**tropy, **a**rtificial **n**eural **n**etworks (ANN) and **ex**pectation-maximization (EM) techniques. This methodology is simply formulated into a model named ENANNEX model. This study does not use any physical characteristics of the catchment areas but deals only with the limited information (e.g. streamflow or rainfall) at the target gauge and its similar nearby base gauge(s).

The entropy concept was confirmed to be a versatile tool. This concept was firstly used for quantifying information content of hydrological variables (e.g. rainfall or streamflow). The same concept (through directional information transfer index, i.e. DIT) was used in the selection of base/subject gauge. Finally, the DIT notion was also extended to the evaluation of the hydrological data infilling technique performance (i.e. ANN and EM techniques). The methodology was applied to annual total rainfall; annual mean flow series, annual maximum flows and 6-month flow series (means) of selected catchments in the drainage region D "Orange" of South Africa. These data regimes can be regarded as useful for design-oriented studies, flood studies, water balance studies, etc.

The results from the case studies showed that DIT is as good index for data infilling technique selection as other criteria, e.g. statistical and graphical. However, the DIT has the feature of being non-dimensionally informational index. The data interpolation

techniques viz. ANNs and EM (existing methods applied and not yet applied in hydrology) and their new features have been also presented. This study showed that the standard techniques (e.g. Backpropagation-BP and EM) as well as their respective variants could be selected in the missing hydrological data estimation process. However, the capability for the different data interpolation techniques of maintaining the statistical characteristics (e.g. mean, variance) of the target gauge was not neglected.

From this study, the relationship between the accuracy of the estimated series (by applying a data infilling technique) and the gap duration was then investigated through the DIT notion. It was shown that a decay (power or exponential) function could better describe that relationship. In other words, the amount of uncertainty removed from the target station in a station-pair, via a given technique, could be known for a given gap duration. It was noticed 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 at Bedford site (in South Africa). Two simulation models, viz. RAFLER and WRSM2000 models, were then assessed in this respect. Both models were found suitable for simulating flows at Bedford.

To my beloved son

Jonathan Kabeya

AKNOWLEDGEMENTS

First and foremost, I thank the Almighty Father for having provided the opportunity for me to undertake my PhD work.

I thank sincerely my parents Kavund David, Kabey Celestin, Mwambwa Scola and Mujing Ctherine for the love and support.

I wish to thank my supervisor and mentor, Professor David Stephenson for the cooperation, guidance, support and encouragement throughout the course of the research.

The collaboration from the members of the Water Systems Research Group is gratefully acknowledged.

My special thanks go to the Head of School, Professor Yunus Ballim for his encouragement and moral inspiration.

I owe a great deal to the University of Witwatersrand Financial Aids and Scholarships Office for financial assistance.

I really thank the friends and members of the Unilu Diaspora. I am greatly indebted, through friendship, to Mr. Kasongo Nyembwe for his constant help and encouragement in one way or the other.

Lastly but no means the least, I thank my relatives, brothers and sisters who have been encouraging my academic undertakings with constant prayers and moral inspiration.

LIST OF PUBLICATIONS

Following is a list of publications related to the research presented in this thesis.

Ilunga, M. and Stephenson, D. (2002b) Performance of hydrological data infilling techniques using entropy approach: backpropagation (BP), BP-with momentum and EM algorithms, *IET-IAHR International Conference Proceedings on Water the Blood of the Mankind*, Arusha, Tanzania pp.149-157.

Ilunga, M. and Stephenson, D. (2003a) Performance of hydrological data infilling techniques using entropy approach: expectation maximization algorithms, *11 th South African National Hydrology Symposium*, Port Elizabeth, South Africa, pp.1-6.

Ilunga, M. and Stephenson, D. (2003b) Entropy measures for comparing flow simulation models at Bedford site, *Journal of Hydrology*, under review.

Ilunga, M. and Stephenson, D. (2005) Infilling streamflow data using backpropagation (BP) artificial neural networks: Application of the standard BP and pseudo Mac Laurin power series BP techniques, *Water SA Journal*, Vol. 31, no. 2, pp. 171-176.

Ilunga, M. and Stephenson, D. (2004) Flow simulation model performance assessment using entropy approach, *Proceedings of the International Conference on Water Resources of Arid and Semi Arid Regions of Africa (WRASRA)*, Gaborone, Botswana, pp. 23-27.

CONTENTS

CHAPTER 3

CHAPTER 4

MODEL TESTING ON STREAMFLOWS:

CHAPTER 5

MODEL TESTING ON STREAMFLOWS:

CHAPTER 6
MODEL TESTING ON RAINFALL:
II. ANNUAL TOTALS……………………………………………..6.1

LIST OF FIGURES

Page

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

APPENDIX A

APPENDIX C

LIST OF TABLES

Page

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

APPENDIX B

APPENDIX C

GLOSSARY

ANN: Artificial neural network

BP (or Stand BP): Backpropagation or standard backpropagation

DIT: Directional information index

T: Transinformation

EM: Expectation maximization (or standard EM)

ECM1: Expectation constrained (conditional)

ECM2: Expectation constrained (conditional), version 2

ECM1-2: Expectation constrained (conditional) and
Expectation constrained (conditional) version 1

ECME1: Expectation constrained (conditional) either

ECME2: Expectation constrained (conditional) either, version 2

ECME3: Expectation constrained (conditional) and
Expectation constrained (conditional) version 3

ECME1-2-3: Expectation constrained (conditional) either,
Expectation constrained (conditional) either version 2 and
Expectation constrained (conditional) either version 3

MEM1: Momentum EM

MEM2: Momentum EM version 2

MEM3: Momentum EM version 3

MEMx-y-z: Momentum EM version x, Momentum EM version y and Momentum version z (x, y and z may take any of the values 1, 2 or 3)

Mm^3: Million-meter cube

GenerBP: Generalized backpropagation

GoldSBP (or GoldenBP): Golden Search backpropagation

McL1BP (or MacL1BP): pseudo Mac Laurin power series order 1 backpropagation

McL2BP: pseudo Mac Laurin power series order 2 backpropagation

QBP (or QPROBP): Quick backpropagation

VLR (or VLRBP): Variable learning rate backpropagation

NOTATION

Following are the notations that are applied frequently in the thesis (e.g. refer to results in tables):

acc: acceleration coefficient
u: momentum
lr: learning rate
k: number of iterations
Hyp.Tg.h.l: hyperbolic tangent hidden layer
Ratio mean: ratio of the mean of the observed series to the mean of the estimated series
Ratio variance: ratio of the variance of the observed series to the variance of the estimated series
Weight Cond: Weight condition