THE USE OF EMPIRICAL MODE DECOMPOSITION (EMD) AND VARIABLE LENGTH BOOTSTRAP (VLB) FOR STOCHASTIC RAINFALL GENERATION

Nyaga Job Muriithi

A research report submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Masters of Science in Engineering.

Johannesburg, 2014

TABLE OF CONTENTS

DECLARATION

I declare that this research report is my own unaided work. It is being submitted for the Degree of Masters of Science to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other University.

……

(Signature of Candidate)

……day of .....……year……

day *month* *year*

ABSTRACT

This Research Report sets out to find out how the use of Empirical Mode Decomposition (EMD) for block selection impacts on the performance of the Variable Length Bootstrap (VLB) stochastic rainfall generator. Empirical Mode Decomposition (EMD), a relatively new data-adaptive approach, decomposes a time series into a group of component time series' termed Intrinsic Mode Functions (IMFs) that are considered to quantify the impact of the multiple physical processes that affect the variability in the original time series. Therefore using IMFs may be better than the subjective method currently used in the VLB for block determination. The performance of the resulting model is tested by comparing historic with generated rainfall statistics using a 10-site rainfall generator problem. The hybrid EMD-VLB model is further compared with the standard VLB model using 8 statistics. The EMD-VLB generator is found to replicate the statistics at par with the VLB generator on a monthly time scale while the standard VLB model performs better on a yearly time scale.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Professor John Ndiritu who introduced me to the stochastic modelling field and whom I have enjoyed working with. My gratitude is expressed to him because he was always available to discuss and provide a way forward for the challenges I faced in the research.

Special thanks go to my family, some of whom through my research study, particularly developed interest in modelling from the various disciplines they are in, and my gratitude unto all of them is expressed for their all rounded support.

I also thank the University of the Witwatersrand whose staff, colleagues were an encouragement through their various inputs that assisted to make this research a success.

I finally would like to thank the Water Research Commission (WRC) of South Africa, for providing both administrative and financial support hence making the project that entailed a lot of concentration in its various stages quite enjoyable.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ACC	Annual cross correlation
ASC	Annual serial correlation
EMD	Empirical Mode Decomposition
VLB	Variable Length Bootstrap
WRC	Water Research Commission

1. INTRODUCTION

1.1 Background

In the hydrological cycle, the movement of water is seen to follow a well understood deterministic path, but the magnitude and timing of the various processes (e.g. extreme rainfalls) that constitute the cycle are partly stochastic due to irregularities in the atmospheric circulation (Shaw, 1994). This is evident in studies of rainfall time series data that exhibit non stationarity properties despite the prevalence of ergodicity in the data. General climatological studies in support of this include global precipitation (Tsonis, 1996; Peel et al., 2009), temperature and stratospheric ozone (Diodato and Bellochi, 2010; Yang et al., 2010) and Pacific mean sea-level pressure (Trenberth, 1990). These examples demonstrate that although certain means and periodicities of the climatic variables are expected, they exhibit large variabilities that are not straightforwardly predictable hence necessitating the use of stochastic hydrology.

Water resources are planned and designed for the future and the available historic hydrologic records are highly unlikely to be replicated during the life of the water resource system. They however provide a plausible sample of the many possibilities that could be expected. Stochastic hydrology enables the generation of ensembles of artificial hydrologic sequences to enable assessment of the broad range of hydrologic conditions that could occur. This enables a statistical evaluation of system reliability that cannot be achieved with a single historic sequence. The artificial sequences however need to possess the statistical characteristics of the historic sequence for the assessment to be realistic. This is a non-trivial task that has been researched on for several decades (examples include Yevjevich (1987; 1972), Koutsoyiannis (2000) and Pegram (2003)).

For monthly water resources assessment in Southern Africa, streamflow stochastic generators have been used much more extensively than rainfall generators. However according to Ndiritu and Nyaga (2014), applying stochastic rainfalls rather than streamflows may yield many advantages since rainfall is the main input in the hydrologic cycle and probabilistic analysis can be included more realistically and easily in the analysis of catchment hydrological processes. The impact of climate variability/change on basin hydrology and water resources can be studied with more ease with rainfall rather than streamflow (Ndiritu and Nyaga, 2014). Among the needs for rainfall data is rainfall-runoff modelling to produce monthly flows for the estimation of water yield from large catchments (Srikanthan et al., 2002), and modelling of rainfall dependent activities such as irrigation and sediment transport.

Stochastic hydrology has been dominated by both parametric and non-parametric data generation approaches and the complexity as well as ease of applicability, influences the choice of the models to apply. Parametric and linear approaches are typically characterized by high sensitivity of model results to model parameters (Rajagopalan et al., 1999), and this has led to the development of non-linear and non-parametric approaches for hydrological prediction. In addition, parametric approaches assume that data can be fitted into a specific probability distribution with temporary fixed parameters but this stationarity no longer serves as the default assumption for water infrastructure planning and management (Milly et al., 2008; Peel and Blöschl, 2011). While exceptions to the use of large number of parameters in stochastic parametric models exist (Koutsoyiannis, 2001; Koutsoyiannis et al., 2008; Ndiritu, 2011a), a majority of those approaches become computationary intensive due to the complexities that characterize them. Although many models that are complex and computationally-intensive are considered to provide reliable results, water resources planners and managers often prefer models that are easier to understand and use for practical short and long term planning of water resources systems.

Several stochastic models exist and daily (and sub-daily) rainfall generators form the majority of these (Wilks, 1998; Rajagopalan and Lall, 1999; Srikanthan et al., 2002; Sharma et al., 2003; Clark et al., 2004; Mehrotra et al., 2006; Mehrotra and Sharma, 2007; Wang and Nathan, 2007; Eisinger and Wiegand, 2008; Kim et al., 2008; Mehrotra and Sharma, 2009; Srikanthan and Pegram, 2009; Wang et al., 2011). For the planning and management of water resources, a monthly time step is often (though not always) adequate and modelling at this time step is much simpler than at a daily time step. The need for effective stochastic monthly rainfall generators can therefore not be overstated. Though there is limited literature on monthly rainfall generators, notable ones include those by Sharma et al. (2002), Yates et al. (2003) , Ünal et al. (2004), Brissete et al. (2007), Wang and Nathan (2007) and Serinaldi and Kilsby (2012). The scope for research and development of effective and efficient rainfall generators at the monthly time step is therefore wide.

Most of the climatic processes originate from dynamic physical processes, and the resulting climatic time series are usually non-linear and non-stationary thus requiring adaptive modelling methods. Since an important goal of hydrological data analysis is to unearth the physical insights and implications hidden in the non-stationary and non-linear data, adaptive approaches need to be applied (Huang and Wu, 2008). In search of a stochastic monthly rainfall generator that is simple, reliable and possesses data adaptive properties, two relatively new methods; Empirical Mode Decomposition (EMD) pioneered by Huang et al. (1998) coupled with the Variable Length Bootstrap (Ndiritu, 2011) stand out. The two approaches represent a shift from complex and computationally intensive methods and are conceptually well grounded. Empirical Mode Decomposition (EMD) is a non-linear, data-adaptive, step wise procedure that is capable of breaking down a historic rainfall time series into rainfall amounts that are segmented within their respective time scales. Quantification of the proportion of hydro-climatic time series variations at different time scales due to fluctuations occurring at those time scales (monthly, seasonal,

annual inter-annual and inter-decadal) is particularly important for the sustainable management of land and water resources systems (Peel et al., 2005). Decomposition in this context means the breaking down of a composite into separate and simpler constituents (Victor, 2012). This segmentation results into datasets herein referred to as Intrinsic Mode Functions (IMFs) and a trend that are realized after decomposition. The VLB pioneered by Ndiritu (2011a,b) is a non-parametric monthly streamflow generator that has the ability to overcome one of the greatest limitations of a classical bootstrap; being able to generate extreme values of data beyond (higher or lower than) those in the historic record. As stated earlier, since rainfall amounts and their corresponding occurrence times are unlikely to be repeated in future, stochastic simulated sequences are therefore needed as long as the generated sequences replicate the statistical characteristics of the historic data. The VLB, in addition overcomes another common problem in stochastic generators; preservation of monthly serial correlation between end of one year and the beginning of the next (Ndiritu, 2011a) by using simple approaches. The VLB obtains sampling blocks fairly subjectively by terminating the blocks during the dry periods of the time series. It is likely that block selection can be improved by first decomposing the time series using EMD and then selecting the block start and termination locations based on the characteristics of the IMFs. EMD and the VLB can therefore complement each other to obtain a more robust stochastic generator. The EMD method is comprehensively described and applied in the various stages of this research report while a detailed description of the adaptation of the VLB generator for rainfall generation can be found in Ndiritu and Nyaga (2014).

1.2 Problem statement

The problem statement can hereby be summarized by the following points;

1. Scarcity of simple and robust stochastic models that can easily be used by water resources planners.
2. Presence of several parametric stochastic generators that are cumbersome due to the complexities involved in choosing model parameters that fit specific probability distributions of those generators. These distributions tend to represent linearity and non-adaptability of natural systems thus the need for data adaptive approaches that represent those systems.
3. Scarcity of rainfall stochastic models especially at the monthly time step.

1.3 Objectives of the study

1. To develop a simple, data-adaptive monthly stochastic rainfall generator by complementing the strengths of Empirical Mode Decomposition (EMD) and the Variable Length Bootstrap (VLB).

2. To assess the performance of the generator using representative rainfall data from South Africa.

1.4 Organization of the research report

The report presents a review of stochastic hydrologic generation that merits the choice of VLB and EMD for rainfall generation in chapter 2. Chapter 3 describes the EMD and VLB models and then presents the methodology that is used to develop the hybrid EMD-VLB generator. This is followed by chapter 4 which presents the results obtained from the developed generator and their comparison with the standard VLB generator. Chapter 5 then presents the conclusions and recommendations of the study.

LITERATURE REVIEW

2.1 Introduction

An overview of both Empirical Mode Decomposition (EMD) and the Variable Length Bootstrap (VLB) generator highlighting their respective modelling strengths has been presented in Chapter 1. Ndiritu and Nyaga (2014) reviewed stochastic hydrologic generation with the main aim of identifying a suitable non-parametric model for rainfall generation at the monthly time step. Five non-parametric models; the wavelet (Bayazit et al., 2001; Ünal et al., 2004; Wang et al., 2011), the Reordering model (Clark et al., 2004a, b; Mehrotra and Sharma, 2009), Nearest Neighbour (Lall and Sharma, 1996; Rajagopalan and Lall, 1999; Srikanthan et al., 2002; Yates et al., 2003; Mehrotra et al., 2006; Mehrotra and Sharma, 2006a; Prairie et al., 2006), Kernel Density approach (Sharma et al., 2003; Srikanthan et al., 2005; Wang and Ding, 2007; Mehrotra and Sharma, 2007, 2009) and the VLB (Ndiritu, 2011a, b) were identified. After a comparison of the five approaches on the basis of 5 criteria reproduced here as Table 2.1, the VLB streamflow generator was selected and adapted for monthly rainfall generation.

Table 2.1 A comparison of five approaches for nonparametric stochastic hydrologic generation (Ndiritu and Nyaga, 2014)

Criterion	Stochastic generation approach.				
	Wavelet	Reordering	Nearest neighbour	Kernel density	Bootstrap
Ability to preserve historic characteristics	If the simple Haar wavelet is used, then skewness is not preserved. If a more generalized wavelet is used, then within-year historic statistics are preserved. Long-term variability and persistence may not be preserved sufficiently	Preserves within-year statistics satisfactorily but does not include replication of long-term variability and persistence in its currently used forms.	Preserve within-year statistics adequately and have also been modified to replicate inter-annual dependence. Current forms of this method are not designed to replicate inter-decadal variability and persistence.	Preserve within-year statistics well and formulations for replicating inter-annual dependence have been formulated. Longer-term dependence is however not modeled in these approaches.	Preserve within-year statistics if the simple method of fragments is used in disaggregation. Use of a pair of weighted and perturbed fragments preserves most within-year statistics but over-estimates the minimum flow. The methods can preserve inter-annual and longer term statistics easily by the selection of long building blocks.
Ability to extrapolate beyond the range of historic data (to generate new data).	Has full ability to extrapolate beyond the historic values	Does not have this ability	Most of the formulations do not have this ability. A formulation with limited extrapolation ability has been developed.	Has full ability to extrapolate beyond the historic values	Most bootstrap methods do not have this ability but a bootstrap that has the full ability to extrapolate has been developed.
Limitations of applicability	If the Haar wavelet is used, the historic data is required to possess a normal distribution. More generalized wavelets do not require this.	May not generate effectively if there are many historic values that take on similar values (e.g. daily rainfall with many zeros).	No known limitation.	No known limitation.	No known limitation.
Possibility of generating negative values	The structure of the approach enables this possibility although this is not mentioned in the studies cited.	Not possible	Not possible	It is possible and an effective approach for dealing with this problem has been devised.	Not possible

Ease of use	The fundamentals of the approach and easy to grasp. The Haar wavelet is easy to understand but understanding more generalized wavelets may be more involving.	Method is easy to understand and set up. The length of the moving window for reordering is subjectively selected.	Generally easy to understand although the approach could be computation intensive. The number of neighbours and method of computing distance between data points are subjective	The method is complex and computationally intensive. There is subjectivity in the selection of the bandwidth and the type of kernel to use.	Bootstrap methods are generally easy to understand and apply. The selection of the block length is subjective. Where weighting and perturbation is done, the selection of the form of weighting and level of perturbation is also subjective.

Ndiritu and Nyaga (2014) revealed the VLB as robust in tests using two multiple-site rainfall generation problems. The VLB replicated most historic statistics very well and performed better than a recently developed parametric rainfall generator, PEGRAIM-W (Pegram, 2011). In spite of the successful performance of the VLB, there is still scope for improving the conceptual aspects of the model that could lead to even better performance. As stated in Chapter 1, this study aimed to find out if VLB performance could be improved if the block beginning and termination locations were based on EMD. In the standard VLB model, block start and termination is subjectively located at low flow periods of the time series. The strengths of EMD in frequency-time segmentation as well as its ability to identify and quantify trends (Radic et al., 2004) lead to its consideration for possible improvement of the VLB generator.

2.2 Empirical Mode Decomposition procedure

Although a detailed description of EMD will be carried out in the methodology, its effective review requires a brief introduction of its general aspects.

1. An original time series data is split into two, classified by the minimum and maximum values within the specified time steps in consideration. These extrema values are used to construct the upper and the lower envelopes of the plot joined together by cubic splines.
2. The differences between the two stated maxima in a given time step are used to obtain the mean loop in a process referred to as sifting. Each sifting results to a sequence referred to as an intrinsic mode function (IMF). The mean loop is then subtracted from the original signal (time series) to obtain an inner loop.

3. If in step 2 above the inner loop qualify to be an IMF, it is stored and then subtracted from the original time series, and the remainder signal is treated as the original signal for consecutive analysis.
4. If in step 2 above the inner loop does not qualify as an IMF, treat it as an original signal for more analysis (computations).
5. Steps 1 to 4 are repeated until several IMFs and a residual trend (the last IMF) are obtained from the original time series. The last IMF (that is referred to as the residual) is characteristic of a monotonic trend because no meaningful frequencies are obtained from it. It should however be noted that the residual might display significance if longer time series data is analysed where variations become more evident.

The different IMFs obtained represent various modes of an original time series that are separated by significantly different frequencies in different time periods. Sifting results into various sequences from an original time series that have different amplitudes and frequencies in different time periods. The more frequent occurrences (with smaller but sharper troughs) appear initially and the less frequent ones appear last on each IMF. Therefore in a rainfall time series, EMD allows us to compute the proportion of rainfall magnitudes variation in a time series that can be attributed to fluctuations in varying frequencies at different time scales (McMahon et al., 2008), where each individual IMF represents time sequences of almost similar frequencies different from the other IMFs; all derived from the original time series. The summation of the different IMFs will result into the original time series.

Mathematically this summation can be expressed by;

$$x(t) = \sum_{i=1}^{n} IMF\ (i) + \ e(n) \qquad (2.1)$$

where, $x(t)$ is the time series data being analyzed.

$IMF\ (i)$ is the Intrinsic mode functions at time t.

$e(n)$ is the Residual of the data set.

i=1,2,3,…,n is the number of extracted IMFs.

With this basic description of EMD, a review of EMD ultimately aimed at identifying an appropriate method of applying it with the VLB generator now follows.

2.3 Review of relevant studies on Empirical Mode Decomposition

The literature review identified 26 studies that were considered relevant to the current one. To ease the presentation of the review, it was decided to use a tabular format (Table 2.2) that informs for each study; the author(s), the main objective/s, the sources and main features (e.g. time scales) of the data used, the methodologies applied and the major findings from the study.

Although this is not exhaustive, it is considered to be adequately comprehensive for this study.

Table 2.2 Brief description of relevant studies on Empirical Mode Decomposition

Author(s)	Main objectives	Data, source, length and time step	Methodology	Main findings
Zhao and Huang (2001)	To explore the effectiveness of mirror extending and circular spline function for Empirical Mode Decomposition	Same data in the examples provided in Huang et al (1998)	1)_Data decomposition by EMD.2)_ Determination of data envelopes by cubic spline fitting 3) _Mirroring at the end of data such that extrapolation of data by extension of data image symmetrical to the data is formed. 4)_ A connecting curve between the two mirrors with the original data altogether presents a circular pattern 5)_Extrapolation is done by extending data to the lower mirror but only the output from the upper mirror is used.	The method utilizes data to extrapolate and is thus data adaptive, reasonable and reliable and can be used for any kind of characteristic data.
Huang et al. (2003)	To identify a confidence limit for empirical mode decomposition and Hilbert spectral analysis	Daily Length of Day (LOD) dataset produced by Gross (2001) from 20th January 1962-6th January 2001 in a total of 14232 days.	Introduction of statistical measures of confidence limits tested on non-stationary and non-linear data and the experimentation of various stopping criterion	If mode mixing is observed to occur, determination of scales should be carried out so that each IMF contains results of one narrow time-scale range. The method increases more rigour of the EMD method thus making it more robust and more useful
Rilling et al. (2003)	To provide a step-wise insight on EMD and its algorithms.	Analysis on the use of sinusoidal Frequency Modulation components and Gaussian wave pockets.	1)_Sampling, interpolation and use of border effects in the cubic splines, over sampling and use of mirror symmetry to achieve smooth boundary conditions 2)_Determining the stopping criteria for sifting 3)_Performance elements in-terms of tones sampling and separation	Provides new insights on EMD and its use experimentally but a need for further studies devoted to theoretical approaches is required.
Coughlin and Tung (2004)	Investigating climate variability by the use of EMD.	Monthly averages of daily Global pressures from Jan 1749-Sept 2002	1)_Calculation of envelopes by cubic splines and end extensions of typical waves by nearest local extrema. 2) Decomposition of the sequences into 5 modes and a residue in which statistical tests of significance and noise distribution are conducted.	Each mode remained orthogonal to each other and has great significant interpretation in the climatic cycle from each other.

Radic et al. (2004)	The use of empirically decomposed intrinsic mode functions (IMFs) to analyze climatological data	EMD applied to a series of annual and seasonal averages of temperature, cloudiness, air pressure and annual and seasonal sums of global radiation and precipitation in Zagreb-Gric, Croatia, between1862-2002.	Analysis of the influence of the particular seasonal and annual averages as well as correlations. This is carried out by discarding the climatic noise and summing up the low frequency IMFs and the residual.	The decomposition in the IMFs with the associated time scales could be used in future climatic predictions.
Peel et al. (2005)	Identification of the prevalent issues in the application of EMD	8135 annual precipitation records around the world	1)_Undertaking sifting involving a tradeoff between under-sifting that leads to under-defined IMFs and over-sifting to produce smooth amplitudes but less physical meaningful IMFs. 2)_Comparison of 3 different end condition rules (mirror (Rilling et al (2003)),(average (Chiew et al, 2005)) and Szero (Coughlin and K. K. Tung, 2005) methods) tested on the data	The SZero rule that assumes that the slope of the spline is zero at the end points decomposes it into fewer IMFs and is recommended as the more efficient and physically meaningful end condition rule.
Sinclair and Pegram (2005)	Exploration of the effects of lower and higher frequency components of spatial rainfall data on temporal persistence by use of 2-dimensional EMD.	Analysis of a large set of 800 radar rainfall images in South Africa	1)_Decomposition of spatial rainfall data into its Intrinsic Mode 2-D Surfaces (IMS). 2)_Computation and removal of the least persistent IMS from the raw rainfall data. 3)_Decompositions are carried out until the monotone trend residual left is more persistent, and of low frequency, where the IMIS surface function is almost zero.	The decomposed spatial rainfall (into IMSs) is mutually orthogonal and adds up to the original data. The method successfully demonstrates that lower frequency components (with large spatial extent), of spatial rainfall exhibit greater temporal persistence than the higher frequency ones.
Wu and Huang (2005)	The use of white noise in data analysis to aid Empirical Mode Decomposition	Analysis of El Nino-Southern Oscillation (ENSO) between the western and Southern Pacific.from January 1870 to December 2002 provided by the Hadley Center for Climate Prediction and Research (Rayner et al., 1996)	Addition of white noise of finite amplitude to the original time sequences and then decomposition using EMD approach.	The approach separates signals of different scales without undue mode mixing.

Molla et al. (2006)	The use of EMD to establish the relation between rainfall variability and global warming.	15 year daily rainfall data for the years between 1989-2004,from the Agricultural Experimental Farm, Giridih, India.	EMD of the data and the identification of Instantaneous frequency (signal's frequency at every time instance), Hilbert spectrum (The joint distribution of the amplitude and frequency as a function of time), marginal Hilbert spectra (measure of total energy contribution from each frequency value), PDF of the IMFs, stationarity test as well as the completeness and orthogonality of the decomposition. Reconstruction of the original data.	Majority IMFs are normally distributed and hence they satisfy X^2 distribution. The study suggests that the recent global warming and decadal climate variability contribute to more extreme events and more frequent, floods and long lasting droughts.
Huang and Wu (2008)	A review of the Hilbert spectral analysis and the EMD processes in data adaptivity.	Remote Sensing Systems (RSS) T2, the channel 2 troposphere temperature of the microwave sounding unit (Mears et al, 2003) during various time steps.	1)_Brief explanation and the review of the construction of the IMF components. 2)_The optimum S stoppage criteria whereby zero crossings and extrema are equal or differ by 1 and stay the same for S consecutive times.	EMD offers a potentially viable method for non-linear ad non-stationary data analysis especially for time frequency representations. Mathematical foundations are required to make the method more rigorous, robust and friendlier.
McMahon et al. (2008)	The use of EMD to stochastically generate six monthly rainfall sequences that takes into account the natural climate phenomena	Six rainfall stations consisting of 135 years long. The study timescale is on a 6 monthly time step.	1)_Decomposition of a historical rainfall series by use of EMD.2)_Recombination of the decomposed series into two components-intra-decadal and inter-decadal time series.3)_Stochastic hybrid generation by use of Matalas (1967)-AR(1)-EMD multisite model 4) Comparison of the results with the traditional AR(1) models	Both the EMD and the traditional methods preserved the historical input parameters But the EMD generated more multi-year extreme rainfalls. EMD is a favourable method to study the effects of anthropogenic climate changes
Wu and Qu (2008)	Presentation of an improved method for restraining the end effect in EMD and its applications to the fault diagnosis of large rotating machinery.	Analysis of vibrational displacements with a sensitivity of 200Mv/mil, each data set consisting of 1024 data points sampled at a rate of 2000Hz	Comparison of the performances of cubic spline end conditions namely; 1)_Mirror Method (MM)_(Zhao and Huang,2001), 2)_Slope Based Method (SBM)_(Dätig and Schlurmann, 2004) and the 3)_Improved Slope Based Method (ISBM)_(Wu and Qu, 2008), by evaluating the orthogonality of the IMFs of several numerical simulated time series.	The ISBM improved the performance of the EMD method as compared to the other end condition methods and hence is recommended for the analysis of non-stationary, non-linear signals.

Zhang et al. (2008)	To provide a secondary segmentation algorithm of EMD to sift areas that do not satisfy the EMD criteria.	Yarn signal in the drawing frame is acquired from Uster-I yarn. Evenness-meter in a Textile mill.	1)_Subdivision of original data into segments .2)_The segmented IMFs are joined again and EMD is applied again on the joint segments to obtain the original IMF and a residue.	The algorithm reduces the computing time of EMD.
Xinxia et al. (2009)	The use of EMD and the RBF neural network prediction model for rainfall prediction	39 year Rainfall sequences between1956-1995, Handan city, China.	Decomposition of the historic time series into IMFs and a residual 2) phase-space reconstruction by use of RBF mode 3) Reconstruction of the series by adding the RBF prediction models to obtain the stochastic sequence	The model's prediction is superior to direct prediction as it accurately predicts rainfall significantly.
Sang et al. (2010)	The prediction of non-stationarity climate series based on EMD.	Monthly mean surface air temperature anomaly in the northern hemisphere (SATA) covering 1752 months	1)_Decomposition of the time series 2)_Use of the time index method (Yu et al ,1998) to inspect the non-stationarity of the series. 3)_Running prediction experiments for a non-linear and non-stationary signal	The EMD process effectively decomposes the non-linearity and non-stationarity hence improving the prediction skill. Prediction errors may arise from the end effects. A transformed time series in-terms of finite number of modes and low non-stationarity improves its prediction capability by a segregation technique.
Zheng et al. (2010)	The use of EMD in the prediction of agricultural drought trend.	58 years of precipitation data from 1951-2008 in Guangdong province, China	1)_Step-wise decomposition of the trend with different scales in signals thus generating series of data sequences with different characteristics. 2)_Use of precipitation anomalies percentages to calculate and analyze medium, heavy and mega disasters quasi-drought periods.	The analysis provides a reference for analysis and predictions of agricultural drought. The precipitation anomalies presented a rising trend from 1960 – 1980 with a maximum in 1989 which corresponds with the drought periods
Karagiannis and Constantinou (2011)	The processing of white Gaussian noise biomedical signals with EMD.	A sampling frequency of 1000Hz from electrocardiogram signals	Pre-processing by addition of white noise and EMD and carrying out several iterations to establish statistical samples	Preprocessing results into a reduced number of siftings and as well as computational time.

Sang et al. (2012)	The combined use of EMD and Maximum Entropy spectral analysis for period identification in hydrologic time series	9 hydrological series rainfall data from various locations in China.	1)_Decomposition of the analyzed series into components 2)_Removal of noise and trend from the decomposition results 3) separation of different deterministic components 4)_Application of MESA (Parker, 1975) into each component to identify trends.	Period identification avoids the influence of noise and trend. Removal of the influence of multi-scale characteristics of hydrologic series makes period identification more accurate and reliable.
Victor (2012)	Introducing Empirical mode Decomposition Method and its algorithms	Various data sources	1)_Introduction into the various terms as involved in Empirical mode decomposition. 2)_A brief explanation of the EMD algorithm 3)_Implementation of the EMD using a C++ programming language	Sets forth the essence of EMD and demonstrates the application of the method in data analysis. The method can as well be applied to linear and stationary sequences
Karthikeyan and Kumar (2013)	Assessment of the Wavelet and EMD coupled ARMA models in forecasting non-stationary hydrologic data.	Four site monthly streamflow volume data, two site rainfall data, USA.	1)_Decomposition of time series data by wavelet and EMD into simpler components 2)_ARMA models are fitted to calibrate and predict each component independently 3)_ Addition of component predictions to obtain series forecasts	With reasonable accuracy, wavelet has a better accuracy in predicting some maxima of the data at lesser time-steps
Tianlu and Zengli (2013)	To determine the effects of Envelope correction Method on EMD's end effect.	Experimental analysis from various simulation signals.	1)_Use of local maxima and minima as interpolation points to fit up and low envelopes.2)_ Regard the first maximum as the endpoint reference on the left.3)_Set the maximum value of signal where each maxima of it is an internal reference point.4)_Determine sub-waves by subtracting different maxima from the determined highest maximum point.	The method solves the problem of envelope distortion thus improving overall EMD accuracy.

Wang et al. (2013)	Incorporation of Ensemble EMD in a hybrid annual rainfall-runoff forecasting model	Annual rainfall series from the upper Longyangxia, the sub-water resources region between Longyangxia and Lanzhou and the natural annual runoff series in the Lanzhou station between 1956-2000.	1)_Decomposition of annual rainfall in a run-off model based on a Support Vector Machine (SVM) (Vapnik,1995;Wang et al, 2009).2)_Use of Particle Swarm Optimization (PSO)(Kenndey and Eberhart, 1995) to establish the parameters of SVM. 3)_Evaluation of the model's performance by its forecasting capability.4)_ Least squares regression and a three feed forward Artificial Neural Network (ANN) are the benchmark models.	PSI-SVM model based on the Ensemble EMD approach can significantly enhance rainfall-runoff forecasting.

The studies outlined in Table 2.2 on EMD demonstrate its wide applicability in many fields and reveals the robustness of this data driven technique. The acceptance of the method as a processing tool is stressed by the large number of applications in diverse areas including biomedical signal processing (Karagiannis and Constantinou, 2010; Dragomiretskiy and Zosso, 2013), machine fault diagnosis, (Gao et al., 2008; Wu and Qu, 2008), stock exchange (Qian et al., 2011) and hydrological time series analysis and prediction (Sinclair and Pegram, 2005; Molla et al., 2006; McMahon et al., 2008; Peel et al., 2009; Diodato and Bellocchi, 2010; Yang et al., 2010; Zheng et al., 2010; Sang et al., 2012; Karthikeyan and Kumar, 2013; Wang et al., 2013). There is limited literature on the use of EMD for stochastic hydrologic simulation of new sequences that can be used for time series prediction. Most EMD studies relate to; the improvement of the method (Rilling et al., 2003; Wu and Qu, 2008; Peel et al., 2009; Hofreiter and Trnka, 2011),comparison of the online and offline EMD(Hofreiter and Trnka, 2011) analysis and reviews of the algorithm (Huang et al., 2003; Huang and Wu, 2008; Victor, 2012) analysis of the resultant decomposed IMF segments (Coughlin and Tung, 2004; Radic et al., 2004; Sinclair and Pegram, 2005; Molla et al., 2006),detrending fluctuation analysis (Qian et al., 2011) and predictions of the resultant trends (Zheng et al., 2010).

The few notable studies in stochastic hydrologic simulations using EMD include McMahon et al. (2008), Xinxia et al. (2009), Yang et al. (2010), Wang et al. (2011) and Karthikeyan and Kumar (2013). These studies decompose a time series into various IMFs that are then fitted into auto-regressive models to construct new stochastic sequences. However, as explained in Chapter 1, the objective of this research is to apply a simple and robust non-parametric method and not a parametric generator.

EMD method identifies various IMFs having almost similar modes in frequency and time and thus the oscillations in a particular IMF are almost similar. Therefore an irregularity in any IMF can be attributed to faults or a random occurrence that happened in the course that are physically meaningful. This has

been useful in identifying faults in machinery (Wu and Qu, 2008) and identification of extreme droughts (Zheng et al., 2010).

In order to complement the strengths of EMD and VLB, coupling of both methods is proposed to come up with a more effective model because;

1. The EMD method is able to intuitively separate and quantify cyclic patterns of rainfall occurring at various inter and intra-decadal time scales. These patterns can be considered to be the result of physical processes that are not explicitly identifiable or well understood at the current state of knowledge. Identifying these patterns and basing VLB block on them could improve the generation of stochastic sequences.
2. The VLB is able to adequately replicate variability and so the multiple length blocks from the EMD will be a vital input into VLB's stochastic rainfall generator. It is considered likely that the variability identified and expressed by these EMD blocks will be replicated in the synthetic series by the hybrid model.

The choice of VLB to replicate observed variability is further demonstrated by its ability to adequately model climate-related change by Ndiritu and Nyaga (2014). Though the scope of the research does not include this, the robustness of the generator in modelling different hydrological components as opposed to generators that can only model the components that they were specifically designed for makes it ideal for this study.

2.4 Summary of the Literature Review

Although not much literature is available in the use of EMD for stochastic rainfall generation, applications in one field are generally applicable to others. One notable contribution in support of this is demonstrated by Wu and Qu (2008) in modelling end effects in cubic splines where findings derived from climate modelling as detailed by Huang et al. (1998), Rilling et al. (2003), Coughlin and Tung (2004), Dätig and Schlurmann (2004) and Chiew et al. (2005)

and are compared and effectively modified for use in the diagnosis of industrial machinery. This demonstrates the adaptivity of EMD as a data generation method and hence its growth as a suitable choice for modelling nonlinear and non-stationary data as shown in Table 2.2. It is in line with this that though EMD as a new data generation tool has not developed much, various studies listed in Table 2.2 will be used to enhance the development of the generator.

After obtaining adequate justification from literature to use EMD and VLB methods, the next step entails the development of the hybrid EMD-VLB generator. This is described in the next Chapter.

3 DEVELOPMENT OF THE HYBRID EMD-VLB GENERATOR

3.1 Introduction

Due to the fact that both EMD and VLB are data based, the development of the hybrid EMD-VLB model requires the use of data. This chapter therefore first describes the data used in the analysis. This is followed by a description of EMD and then the VLB with emphasis on the block identification (start and termination) procedure. A description of the hybrid EMD-VLB model is finally presented.

3.2 Data for model development and testing

The aim of the project is to develop a rainfall generator by utilizing data adaptive properties of EMD and the strengths of VLB. The VLB has been previously tested on 10 rainfall stations problem using rainfall stations that are widely-spaced over South Africa (Ndiritu and Nyaga (2014)). The stations are shown in Figure 3.1, and have been selected for the development of the hybrid model. This allows for easy comparison of the hybrid with the standard VLB as the VLB had been assessed using the same problem. The rainfall data is obtained from an extensive data base by Lynch (2003) and consists of a consecutive record of 93 years of observed monthly rainfall with minimum patching (averaging 3.5 %). The basic characteristics of the rainfall stations illustrated below in Figures 3.1 -3.5 and Tables 3.1- 3.2 are extracted from Ndiritu and Nyaga (2014). Table 3.1 highlights the basic statistics of the stations and Figure 3.2 shows the monthly rainfall distributions of the stations. From Figure 3.2, the presence of both dry and wet seasons is observed.

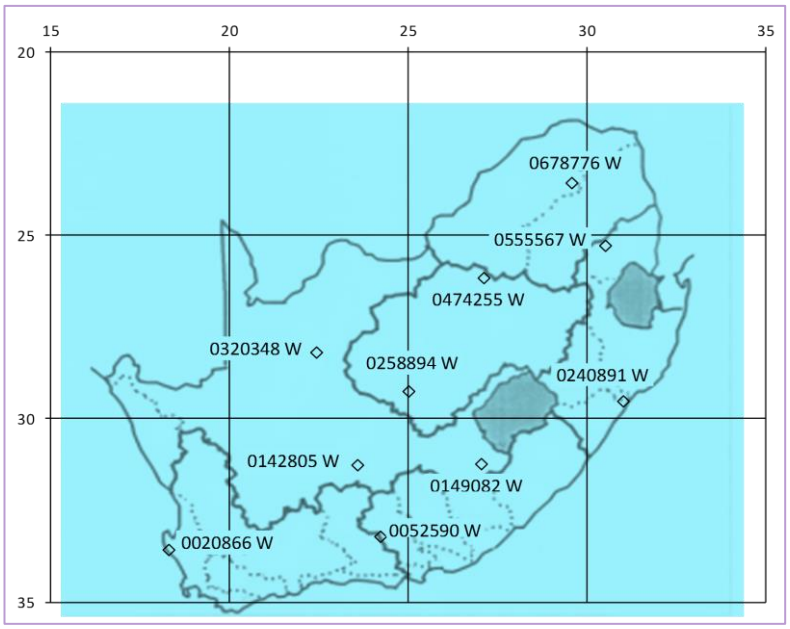

Figure 3.1 Location of selected stations in South Africa (Ndiritu and Nyaga, 2014)

Table 3.1 Basic statistics of rainfall stations (Ndiritu and Nyaga, 2014)

Station	0020866W	0555567W	0474255W	0320348W	0258894W	0678776W	0052590W	0142805W	0149082W	0240891W
Mean	605	830	579	325	394	843	238	320	588	995
Stdev	115	221	151	138	140	285	90	103	141	218
CV	0.19	0.27	0.26	0.42	0.35	0.34	0.38	0.32	0.24	0.22
Skewness	0.31	0.93	0.36	1.53	0.84	0.86	0.85	0.43	-0.01	0.6
Minimum	349	556	209	104	159	405	69	113	247	549
Maximum	857	1501	1061	959	793	1577	607	627	990	1741
% patching	0.5	2	7.1	0.6	4.4	6.3	8	0.3	2.4	3.4

CV: Coefficient of variation

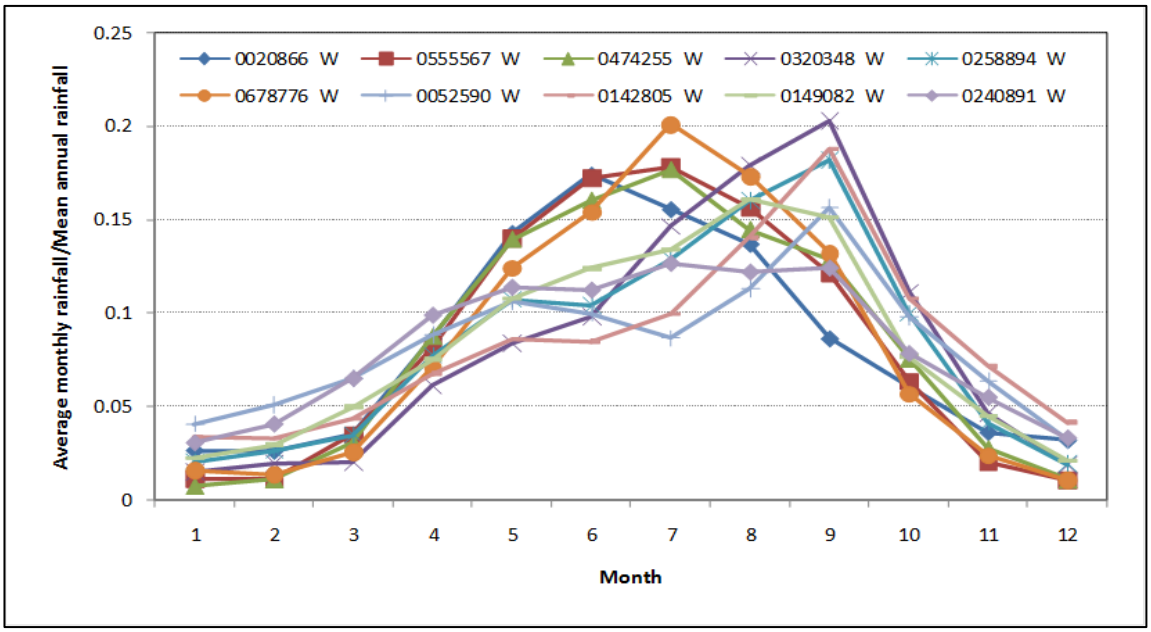

Figure 3.2 Monthly rainfall distributions of selected rainfall stations (Ndiritu and Nyaga, 2014). The hydrological year starting with January is applied for the winter region and the hydrologic year starting with July for the winter region.

Table 3.2 shows the cross correlations and the serial correlation coefficients of the rainfall stations. It can be seen that station 0020866W has a very low cross correlation with the other stations; with the highest being 0.16 with station 0240891W.This might be attributed to the station being situated in a winter rainfall zone while the other 9 are in the summer rainfall zone. The stations have very low annual serial correlation coefficients thus highlighting very weak annual temporal dependence structures even within each of the stations.

Table 3.2 -Cross correlation and serial correlation coefficients of annual rainfalls (Ndiritu and Nyaga, 2014)

	ACC										
Station	0020866W	0555567W	0474255W	0320348W	0258894W	0678776W	0052590W	0142805W	0149082W	0240891W	ASC
0020866W	1										0.09
0555567W	0	1									-0.06
0474255W	0	0.44	1								0.05
0320348W	-0.12	0.46	0.3	1							-0.08
0258894W	0	0.39	0.37	0.73	1						-0.15
0678776W	-0.05	0.76	0.32	0.46	0.42	1					-0.07
0052590W	-0.07	0.2	0.13	0.39	0.37	0.31	1				0.1
0142805W	-0.03	0.37	0.32	0.66	0.66	0.41	0.55	1			-0.07
0149082W	0	0.42	0.31	0.52	0.64	0.43	0.42	0.6	1		-0.16
0240891W	0.16	0.27	0.22	0.23	0.35	0.23	0.19	0.34	0.31	1	0.01

ACC – Annual cross correlation coefficient; ASC – Annual serial correlation coefficient

Figure 3.3 shows the monthly cross correlation coefficients for the 10 stations which follows a fairly distinct pattern in all the stations. Figure 3.4 shows the monthly serial correlations from which a distinct pattern is not observed. Therefore although there is a monthly spatial dependence structure within the stations, the monthly temporal dependence structure is minimal. This is further demonstrated by Figure 3.5 which shows the overall average monthly cross and serial correlations. It is noted that the highest annual serial correlation coefficient is 0.1 in station 0052590W the lowest being -0.16 in station 0149082-W. These properties are illustrated graphically in Figures 3.3-3.5.

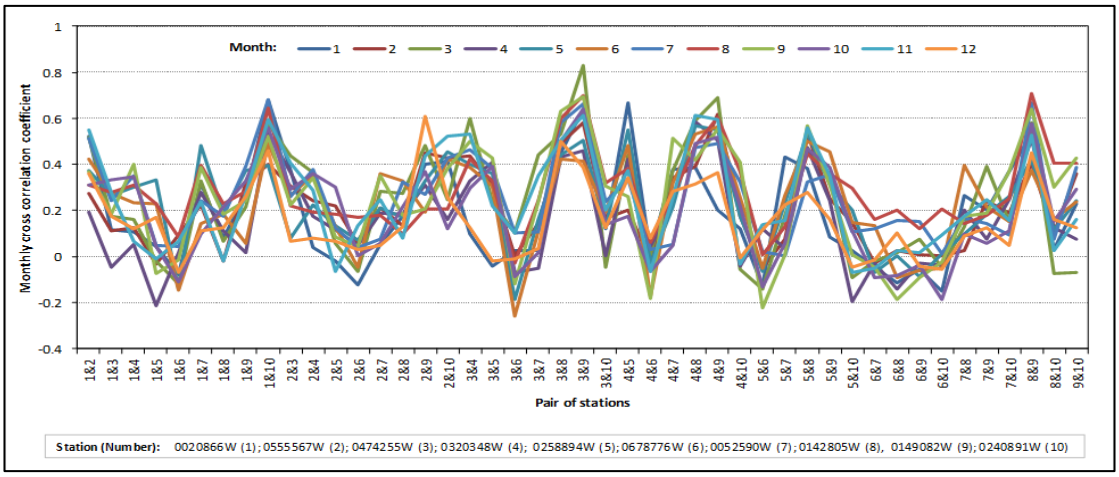

Figure 3.3 Monthly cross correlation coefficients for rainfall stations (Ndiritu and Nyaga, 2014). The x-axis denotes lag-1 monthly correlation.

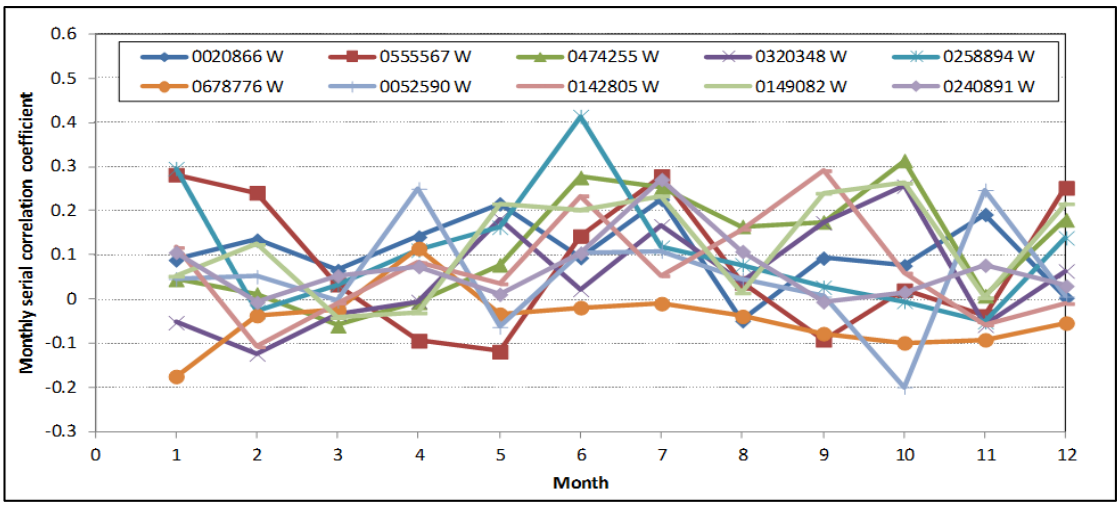

Figure 3.4 Monthly serial correlation coefficients for rainfall stations (Ndiritu and Nyaga, 2014)

Figures 3.3 and 3.5 can both be summarized by Figure 3.5. The average serial correlation in one month for all the stations is calculated and plotted as a single value for that month. This is carried out for all the months in the hydrological year. The same is done for the cross correlations.

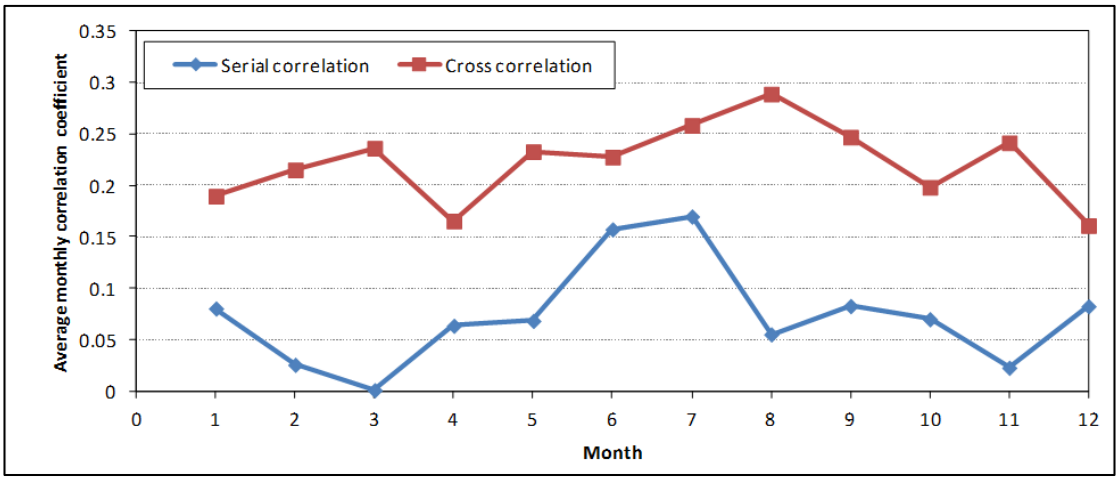

Figure 3.5 Average monthly cross and serial correlation coefficients of rainfall stations (Ndiritu and Nyaga, 2014).

From Table 3.2 and Figures 3.3-3.5, the low cross and serial correlation coefficients, then means that there is no requirement for the proposed generator to preserve them.

3.3 Empirical Mode Decomposition (EMD)

There have been several modifications in the EMD algorithm since the initial development by Huang et al. (1998) in various aspects of its methodology including the sifting process stoppage criteria, sifting modifications and in splines' end effects, but a consensus on the generalized algorithm still exists. A description of the EMD as used in this study now follows.

1. The input (original) time series x(t) is classified on the high values herein referred to as the maxima and the low values referred to as the minima as described in Chapter 2 (Section 2.2). The enclosure in between the maxima and the minima that is referred to as the envelope is defined by a cubic spline illustrated in Figure 3.6. Cubic splines are piece-wise polynomial approximations that are very widely used in fitting each successive pair of data points due to their smooth functions and hence their use as interpolation functions in EMD. Extrapolations are necessary to determine the cubic spline values at the end extrema at both ends of the time series and their poor determination leads to unrepresentative IMFs during sifting. A detailed description of cubic splines, their formulation, and the role they play in EMD's interpolations and extrapolations is presented in appendix A.

2. The mean of the minima and the maxima is calculated and plotted on the same graph as shown in figure 3.6 is given by;

$$mean, m_1 = \left(\frac{maxima + minima}{2}\right) \tag{3.1}$$

3. The mean is subtracted from the original time series to obtain h1, defined by;

$$h_1 = x(t) - m_1 \tag{3.2}$$

This defines the first sifting and h_1 is expected to satisfy the definition of an IMF but this is not the case since changing a local zero from a rectangular to a curvilinear coordinate system may introduce new extrema, and further adjustments are needed (Huang and Wu, 2008).

Steps 1-3 are carried severally on the residual until the detail signal $h_1(t)$ can be considered as an IMF in which;

i. The number of zero crossings (of the graph on which the loop is plotted on) and extrema must be zero or at least differ at most by one in the complete data sets. This is necessary in order to remove riding waves (Prah and Okine, 2008).

ii. The mean value of the envelope (defined by the local minima and the maxima of the new constructed loop) must be zero (Prah and Okine, 2008).

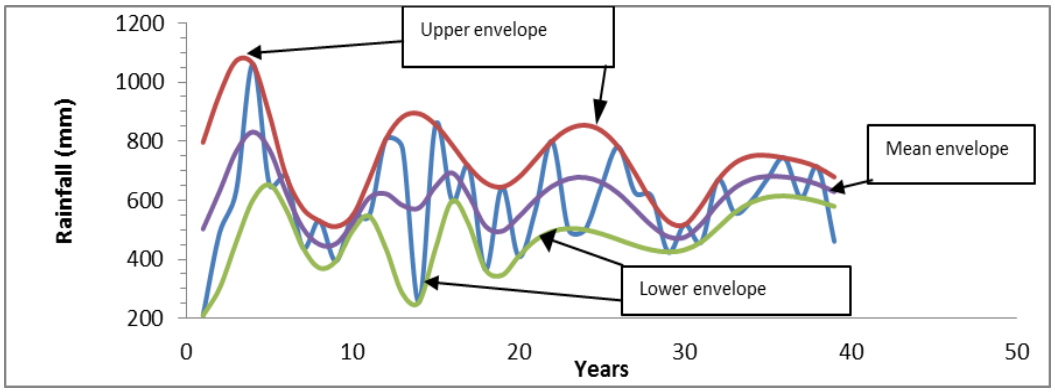

Figure 3.6 Illustration of cubic splines

Therefore in various iterations, h_1 is treated as model signal from which more iterations are derived from, so the next iteration will be defined by;

$$h_{1,1} = h_1 - m_{1,1} \qquad (3.3)$$

And after k iterations, (k is dependent on the series in consideration)

$$h_{1,k} = h_{1,(n-1)} - m_{1,k} \quad (3.4)$$

is realized if the detail signal can be considered as an IMF. $h_{1,k}$ becomes the first IMF and can be denoted by IMF_1. This is because there are k iterations that are carried out to obtain an IMF in a particular series in consideration.

The first IMF contains the shortest–scale oscillation and it is removed from the original time series to obtain the residue, $e(1)$by the expression,

$$e(1) = x(t) - IMF_1 \quad (3.5)$$

where, $e(1)$ is the residue after removal of the first IMF. This IMF_1 is characteristic of longer-period variations as compared to the other IMFs. More siftings are carried out from the residue $e(1)$, which is treated as new data of longer frequencies from where subsequent shorter time shorter frequency IMFs are extracted from.

Thus, after repeating the process n times, the following expressions are obtained.

$$e(2) = e(1) - IMF_1.$$

$$e(3) = e(2) - IMF_2 \ . \quad (3.6)$$

.

.

$$e(n) = e(n-1) - IMF_n \ .$$

The above process results into n IMFs and a residual $e(t)$ that is a monotonic function or a function that contains one extremum from which no more IMFs can be extracted. The realization of this monotonic function is the stoppage criterion. An elaborate flowchart of the EMD process is illustrated in Figure 3.7.

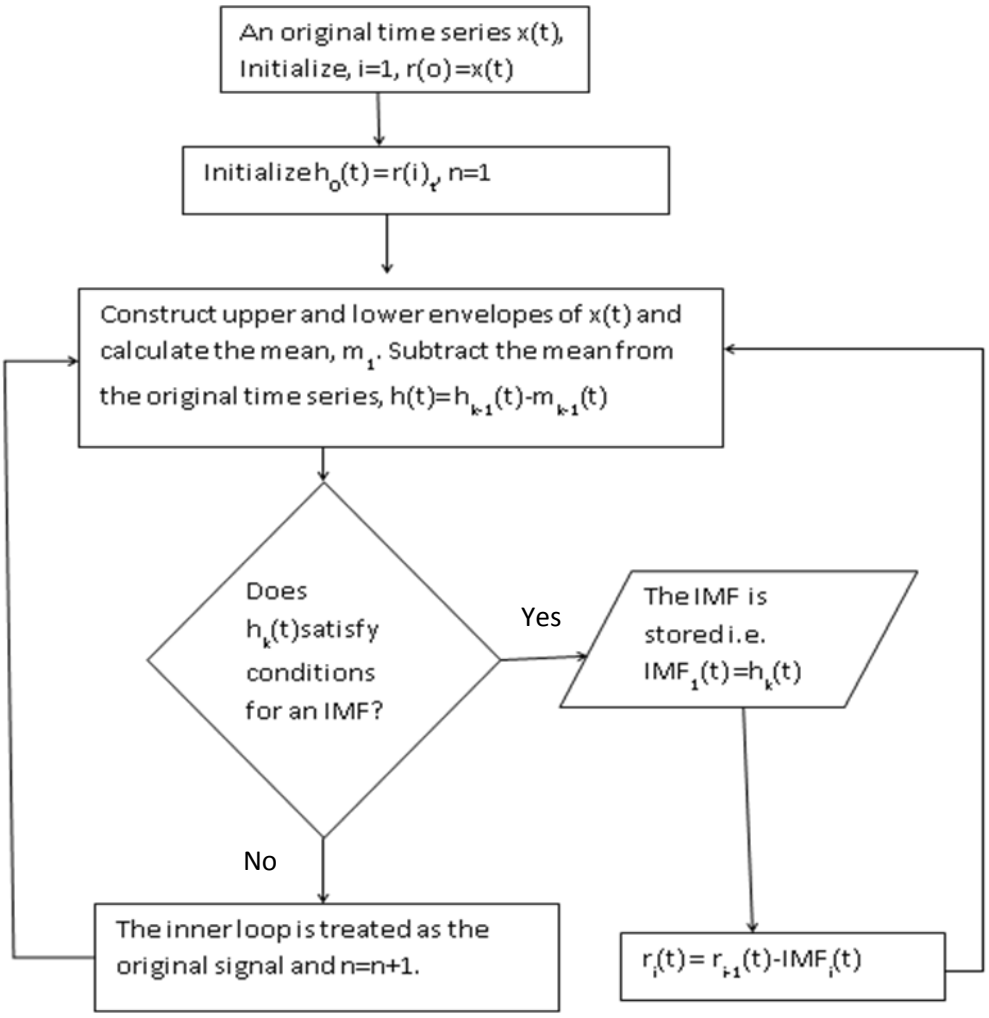

Figure 3.7 A flowchart of Empirical Mode Decomposition

An original time series together with its constituent four IMFs and a residue that result from a step-wise decomposition by EMD from station 0320348 W is illustrated in Figure 3.8. From the figure, both intra and inter-decadal fluctuations of rainfall are revealed. These can be considered to inform how various physical processes of different time scales affected the rainfall over the 93 year period. These will therefore be used in identifying block start and termination locations on the original time series.

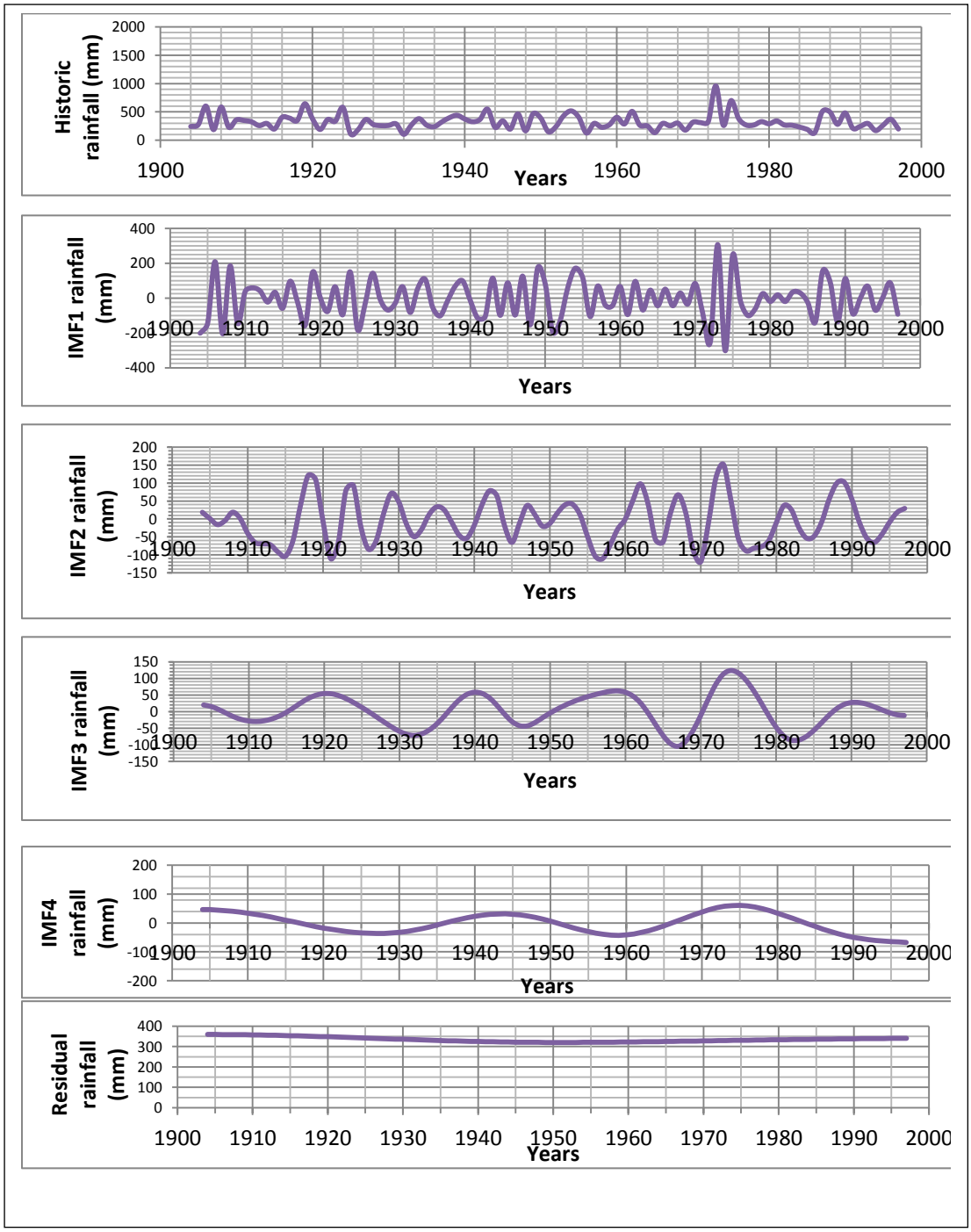

Figure 3.8 Decomposition of a rainfall time series from station 0320348W into 4 IMFs and a residue.

From Figure 3.8, decomposition results into shorter length variabilities in the initial IMFs whose length increase in the successive IMFs.

3.4 The Variable Length Block (VLB) rainfall generator

The VLB rainfall generator (Ndiritu and Nyaga, 2014) is a variation and development from the streamflow generator (Ndiritu, 2011 a,b). The VLB partitions the observed time series into blocks of variable lengths unlike the traditional bootstrap that divides the time series into blocks of constant length. The blocks are then resampled with repetition to create the first synthetic annual time series of the desired length. The traditional bootstrap (e.g. Vogel and Shallcross (1996)) would take this series as the final one which implies that the generated sequences would only contain values from the observed record that are simply temporally re-ordered in different ways. Since the future is expected to have observations that will be at times higher and at other times lower than any of the values in the historic record, the traditional bootstrap is considered substantially limiting. The VLB uses a weighted averaging of the monthly fragments (monthly value/annual value) from different years to obtain perturbations on the annual rainfall to and thereby obtain annual values that are not in the historic record. These new annual values also exceed the historic extremes at times. Ndiritu and Nyaga (2014) provides the complete description of the VLB rainfall generator. Block determination is the main aspect of interest in this study and is now described in more detail.

3.4.1 Block selection by the VLB generator

The following steps describe how blocks of variable length are obtained from the historic time series.

i. A low-rainfall year is defined as that having an annual rainfall lower than that exceeded for a set proportion of time. This proportion is obtained as a random value from a uniform distribution within a specified range. For the rainfall generation by Ndiritu and Nyaga (2014) a range of 60-

90% exceedance was applied. The rainfall value corresponding to this proportion is obtained by a plotting position approach (e.g. the Weibull method).

ii. The minimum length of the block in years is decided and set (a value of 3 years has been found reasonable)

iii. Starting with the first year of the series, move forwards by a length equal to the minimum block length and then proceed at a yearly time step and locate the first low-rainfall year as defined in step i.

iv. Specify this year as the last year of the first block. Obtain the other blocks in a similar manner considering the following year as the new beginning of the time series and check that the last block also meets the minimum block length requirement.

Figure 3.9 from Ndiritu and Nyaga (2014) shows 16 blocks of variable length obtained using this method. Since the method uses randomly selected percentage exceedances (within a range) to obtain the low rainfall threshold (step i), the blocks obtained for each generation vary accordingly thereby creating the desired variability.

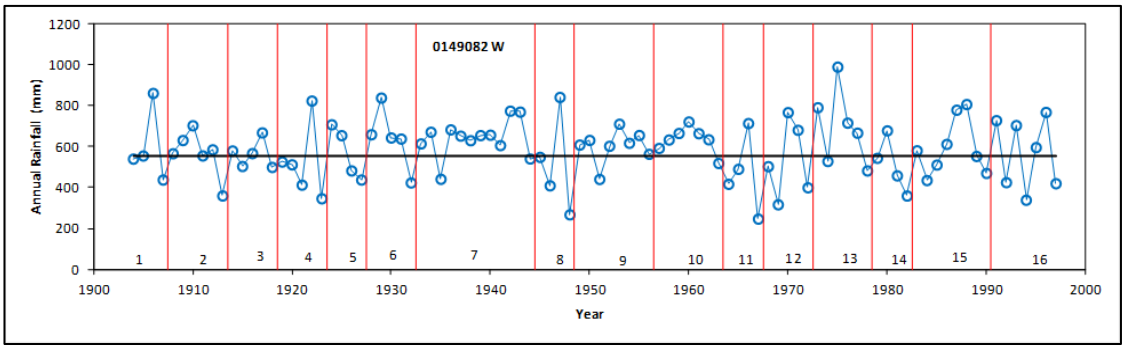

Figure 3.9 The generation of variable length blocks by the VLB generator. The black horizontal line defines the low rainfall threshold and the vertical red lines the termination and starting location of the blocks. The blocks are numbered as 1 to 16 above the x-axis (Ndiritu and Nyaga, 2014).

3.5 The proposed hybrid EMD-VLB model

The VLB block identification method described in Section 3.4.1 is reasonable as it allows the low rainfall periods to recombine in a large number of possibilities when the random resampling is done. The method is however subjective and EMD that is considered to implicitly identify the effects of the short-term (intra-decadal) and the longer term (inter-decadal) hydro meteorological processes on rainfall could form a basis for a more rational basis of block identification. This assumes that block termination using EMD will prevent identification of blocks at locations where the processes were still happening as could happen with the VLB method as described in Section 3.4.1. Considering the minima and the maxima of the IMFs (see Figure 3.8) as the locations where processes impacting on rainfall start and end, the minimum and maxima are therefore used as the locations to start and to terminate the blocks. This is illustrated in Figure 3.10 for an IMF of one of the 10 rainfall stations used in this study.

With these considerations, the EMD-VLB generator is formulated as follows;

1. Decomposition of the original historic time series by EMD to identify representative IMFs and residual trends.
2. Generation of variable length blocks based on the extrema of the different IMFs generated in step 1 as illustrated in Figure 3.10.

3. Random resampling with replacement of the blocks using the lengths obtained in step 2 above to generate initial synthetic sequences of various lengths. Resampling is weighted from various IMFs based on the average block length from each IMF. This is done indirectly by weighting in direct proportion to the number of blocks generated by each IMF. For example, the chance of obtaining a block from an IMF that produced 10 blocks is twice that of an IMF that produced 5 blocks. This prevents bias towards favouring the replication of the effects of longer-term processes at the expense of shorter-term processes. This is because a block takes up a period equal to its length in the generated rainfall.
4. Synthetic generation of stochastic sequences by the use of VLB rainfall generator.

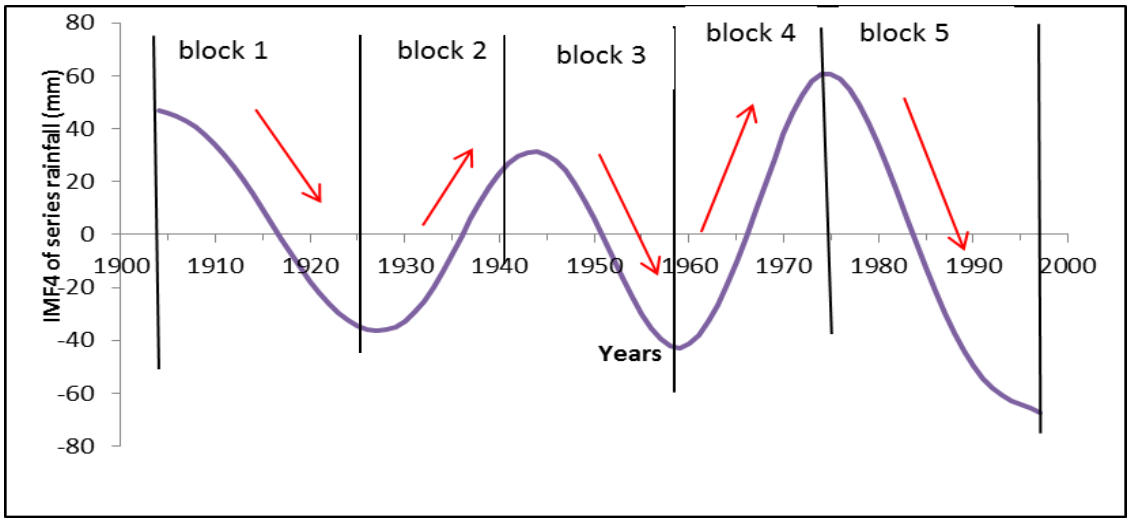

Figure 3.10 Illustration of block start and termination location for IMF4 of station 0320348W. It shows the creation of five blocks that begin and end at the crests and the troughs and proceed along the series as shown by the red arrows.