3. Data and methodology

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever matter may be.
- Lord Kelvin, W. T., 1866.

Chapter 3 outlines the research method, data used, their sources and method of analysis. The schematic framework of the research is discussed. The Chapter focuses on the geodetic and model simulated data sub-sampled at the geodetic VLBI stations and the NWP grid cells over the SHADOZ network. In addition, the non-parametric techniques used in data analyses are also described.

3.1 Introduction

In this thesis, investigation of tropospheric delays due to geodetic WV and WV fluctuations over the Southern Africa region by geodetic and simulation data are reported. The stochastic behaviour of local WV time series is investigated by use of Auto-Regressive Moving Average (ARMA). In addition, the multi-scale variability and scaling behaviour of WV is studied in the time-frequency domain (wavelets) as well as using a data adaptive (noise assisted) methodology (i.e. EMD methods). All these methods take into account the inherent nonlinearity and nonstationary characteristics based on the local time scales of the data. This chapter describes the sources and different types of data that were used in the present research work. In addition, methods used to pre-process these data records are briefly described. In the analysis section, a general and brief description of the mutual information concept, often used in information theory, is discussed and its linkage to the correlation paradigm is presented. Further, for the purpose of studying the scaling behaviour in the WV fluctuations a general description of the wavelet transform, DFA and HHT techniques are also presented. Specific applications of each of these methods, which have been presented in

various international conferences, peer reviewed and published are presented in the subsequent chapters.

3.2. Research methodology

In order to investigate the nature of WV fluctuations over Southern Africa, this research was undertaken from three important viewpoints as depicted in Figure 3.1. Firstly, geodetic data (VLBI and GPS ZTD and delay gradients) at the HartRAO fiducial geodetic station were used to compute a long time series of geodetic WV. Troposphere gradients, VMF and WV derived from ECMWF data were used to investigate the nature of stochastic processes in the time series. Thereafter, the parameters of the ARMA model that characterise the stationarity of WV were adaptively estimated from geodetic tropospheric delay time series.

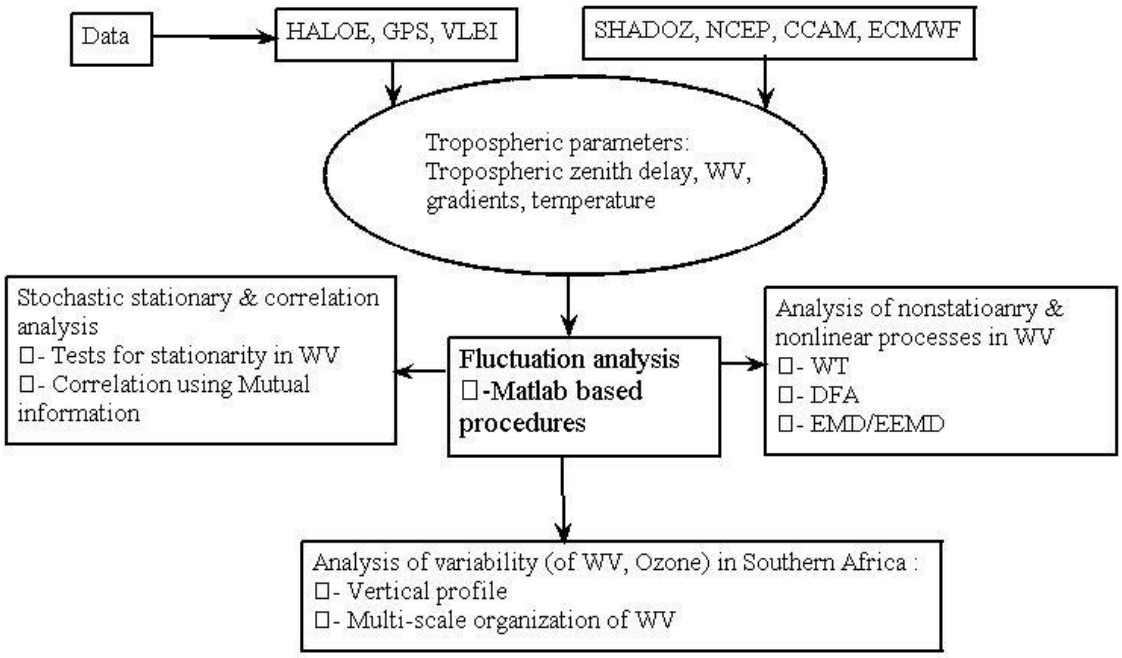

Figure 3.1. Flow diagramme of the research framework.

Secondly, the multi-scale structure of WV in the low- and mid-tropical Africa is investigated by use of *in situ* radiosonde observations of the SHADOZ station network comprising of Ascension, Irene (South Africa), Reunion (Reunion) and Nairobi (Kenya) and the numerical model simulations for the period from 1998 to 2006. Thereafter, the vertical model of

tropospheric WV over southern Africa is developed by use of radiosonde and NCEP/NCAR reanalysis data. Finally, the scaling properties of WV fluctuations were identified and measured using wavelet-based and DFA (DFA is described by Chen *et al*., 2002) approaches. Wavelet analyses could be viewed as a microscope and telescope. This is due to the high-frequency/low-time resolution in the low-frequency part and low-frequency/high-time resolution in high-frequency part characteristic of this technique. In addition, it uses a predefined wavelet basis; the mother wavelet. This implies that the wavelet analysis results are limited by the mother wavelet. Nevertheless, this nonparametric estimate of scaling behaviour is computationally efficient (e.g., the discrete wavelet transformation) and is robust because it has low variance and negligible bias. Further, the Hilbert-Huang transforms, HHT (Huang *et al*., (1998)) and the EEMD reported by Zhaohua and Huang, (2009) were used to adaptively analyse the nonlinear and nonstationary processes in WV. The HHT is built on the assumption that any data set consists of different, simple and intrinsic modes of oscillations (ranging from low to high frequency) that are derived from the observations objectively (adaptively). As a result, this methodology is suited for presenting the WV distributions (derived from observations) in time-energy-frequency distributions.

3.3. Data

Troposphere parameters (N, WV, ZTD and delay gradients) that were analysed and presented in this thesis were derived from geodetic, radiosonde, other space-borne measurements and NWP model simulations. For clarity, Figure 3.1 depicts the data, processing and analysis methods that have been used to study the fluctuations of troposphere parameters (ZTD and WV).

Geodetic data

The central theme in geodetic processing is to derive the delay observable which has position information of the geodetic receiver and the source of the radio signal. For geodetic VLBI, the delay observable also has the structure information of the radio source. To derive this information with high accuracy, the troposphere contribution to the delay observable must be removed. This thesis addresses *an inverse problem*:

a) The results of actual geodetic observations are used to assess and compute the inherent properties of the fluctuating troposphere parameters that characterise the tropospheric structure and dynamics.

b) It also addresses the effect of the atmosphere on the geodetic observations, through the use of actual observations and numerical simulations of meteorological parameters.

The IVS for Geodesy and Astronomy provides tropospheric products such as zenith total delay and zenith wet delay, $\tau_{ztd/zwd}^{atm}$ for all IVS-R1 and IVS-R4 sessions since January 2002 (Schuh and Boehm 2003). All available VLBI observations are processed by the IVS ACs with three main analysis software packages, OCCAM (maintained by the Institute of Applied Astronomy, Russia), CALC/SOLVE (maintained by NASA Goddard Space Flight Centre, GSFC) and Steel Breeze (maintained by Main Astronomical Observatory-MAO, the National Academy of Sciences, Ukraine). The corresponding products such as $\tau_{ztd/zwd}^{atm}$ are transferred to the IGG (Institute of Geodesy and Geophysics, Vienna University of Technology, Austria) for comparison and combination. The motivation for combining the tropospheric parameters is to average out the systematic differences in $\tau_{ztd/zwd}^{atm}$ arising from the use of the different analysis software packages using different parameterisation and models, such as the thresholds of outlier detection, or elevation cut-off angles. For further details, please refer to Schuh and Boehm (2003) and Heinkelmann *et al*., (2007). The combined long time-series of $\tau_{ztd/zwd}^{atm}$ is determined from all geodetic VLBI sessions and can conveniently be obtained from all IVS data centres (see, ftp://cddis.gsfc.nasa.gov/vlbi/ivsproducts/trop).

The geodetic delay and other derived parameters such as troposphere gradients, WV, mean atmospheric temperature and VMF derived from the ECMWF were obtained from IGG. The data is archived at http://mars.hg.tuwien.ac.at. The archive consists of files which contain a record of the global geodetic VLBI, GPS and DORIS station names. The temporal resolution for troposphere parameters archived is six hours corresponding to the NWP model simulations. Since our concern is to assess the local and regional fluctuations of troposphere WV, we study WV (and those parameters that influence WV) variability over a geodetic station; HartRAO-South Africa (see Figure 3.2).

Figure 3.2. Fiducial geodetic site: Hartebeesthoek Radio Astronomy Observatory.

In addition to geodetic VLBI data, troposphere delays derived from GNSS observations were also analysed. The International GNSS Service (IGS) which was formally established in 1993 by the IAG, began routine operations in January 1994 (Beutler *et al.*, 1999). The IGS operates a global network of more than 350 permanent GPS tracking stations, each equipped with a GPS receiver that provides raw GPS tracking data in a Receiver Independent Exchange (RINEX) data format. All available near-real time global IGS observation data are transmitted to the global IGS data centres from where a combined tropospheric product (e.g., weekly files containing $\tau_{ztd/zwd}^{atm}$ in a 2-h time interval from the IGS tracking stations and archived for instance at ftp://cddis.gsfc.nasa.gov/gps/products/trop_new) can be downloaded. Other data sets used in VLBI and GPS processing are presented in Table 3.1.

Table 3.1. Data products and sources used in VLBI analysis.

Geodetic data	Description	Source
Ephemerides & Earth orientation parameters	• Current series: EOP 0504 (Standish, 1990)	http://hpiers.obspm.fr/eop-c
Atmospheric loading coefficients	• Tidal and non-tidal atmospheric tides (Petrov and Boy, 2004)	- http://www.ecgs.lu/atm (tidal S_1/S_2 sine and cosine components)
Thermal deformation	• Thermal deformation of VLBI antennas (Wresnik *et al*., 2005)	- http://mars.hg.tuwien.ac.at~vlbi/thermal
Ocean loading	• Ocean tide model GOT00.2 based on the global ocean tide model from TOPEX/ Poseidon Altimetry/ GOT99.2 (Ray, 1999)	Source, H.G. Scherneck

Numerical prediction model simulations, satellite and Ozonesonde data

NWP model simulated pressure, temperature, and specific and relative humidity fields derived from the reanalysis project of the NCEP / NCAR (Kalnay *et al*., 1996) in the United States (US)); which is the NCEP/NCAR data set were also used to compute WV. These data sets were obtained from NASA's website at http://www.cdc.noaa.gov. In addition, surface temperature measurements based on the automatic weather stations over the HCB were provided by the South Africa Weather Service (SAWS).

The vertical profile of WV model for southern Africa reported in this thesis was derived from Ozonesonde data sets based on the SHADOZ network (i.e. Nairobi - Kenya, Malindi - Kenya, Irene – South Africa, Reunion - Reunion and Ascension–Ascension Islands stations); the data is archived at http://croc.gsfc.nasa.gov/shadoz/. Refer to Thompson *et al*., (2003) for a detailed and technical description of the data sets. In addition, data from the HALOE on board the upper atmosphere research satellite were from HALOE server.

3.4. Data pre-processing

Both Geodetic data and those derived from model simulations utilised in the present research exhibit inherent systematic biases. For instance, geodetic data sets are often acquired by use of geodetic instruments that have measurement biases. In addition, bias contribution from external environment such as changing weather systems is a known problem among the space geodesy community. Furthermore, the difference in the analysis schemes employed by different software packages is known to bias geodetic parameters such as geodetic WV. On the other hand, model simulated data sets are often constrained by the inherent parameterization schemes.

While the analysis strategies employed in this research are robust and some-worth data adaptive, the data sets considered were pre-processed before analysis. A static transformation function was applied to most of the data sets analysed in the thesis in order to ensure a symmetric frequency distribution and also to obtain a more Gaussian-like shape. This transformation is particularly important in assessing the stationary processes in the geodetic data. In addition, the periodic cycles and polynomial trends were estimated and subtracted from the original data sets as a method of disintegrating the time series into inherent components from which the stochastic characteristics of each component would be deduced. In particular, the data sets were transformed using Box-Cox transformation (Box and Cox, 1964) while second order statistics were adaptively subtracted from the data using the Wessel *et al*, (2000) adaptive filtering methodology.

3.4.1 Box-Cox transformation

The Box-Cox transformation is a non-linear static transformation function which converts non-normal datasets to a set of data which approximates a Gaussian distribution. Though the Box-Cox transformation is a family of power transformation, in this thesis, a geodetic data record $\{Y_j\}_{j=1,2,\cdots,N}$ for which $Y_j > 0 \quad \forall j \in \{1,\cdots,N\}$ is Box-Cox transformed by Equation(49),

$$Y'(\lambda) = \begin{cases} \dfrac{\left(Y^{\lambda} - 1\right)}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases} \tag{49}$$

The power parameter, λ is often selected based on the maximising logarithm of the likelihood function given by Equation (50),

$$f(Y,\lambda) = \frac{N}{2}\ln\left[\sum_{j=1}^{N}\frac{(Y_j(\lambda)-\bar{Y}(\lambda))^2}{N}\right] + (\lambda-1)\sum_{j=1}^{N}\ln(Y_j) \tag{50}$$

where geometric mean, $\bar{Y}(\lambda) = \frac{1}{N}\sum_{j=1}^{N} Y_j(\lambda)$.

3.4.2. Estimation of periodic cycles and adaptive filtering

Geodetic tropospheric data exhibit periodic components. A periodic component can be obtained by estimating the mean and the variance for a particular time span over some reference epoch, T_{ref} in the cycle. If the geodetic data record has p cycles of length N, according to Hipel and Mcleod (1994), the mean can be calculated from Equation (51);

$$\bar{\mu}_{T_{ref}} = \frac{1}{p}\sum_{k=1}^{p} Y_{T_{ref},j} \quad T_{ref} = 1,2,\cdots,N \tag{51}$$

where T_{ref} is the reference time epoch and k indexes the successful cycles. Note that $N=365$ for daily measurements over one year cycle. The variance is given by Equation (52);

$$\sigma^2_{T_{ref}} = \frac{1}{p-1}\sum_{k=1}^{p}\left(Y_{T_{ref},k} - \bar{\mu}_{T_{ref}}\right)^2, \quad T_{ref} = 1,2,\cdots,N \tag{52}$$

The normalised anomalies time series can be calculated by Equation (53)

$$Y'_{T_{ref},k} = \frac{Y_{T_{ref},k} - \bar{Y}_{T_{ref}}}{\sigma_{T_{ref}}}. \tag{53}$$

The main objective in the current analysis of geodetic data is to investigate the characteristics of the fluctuations of the WV and the nature of the underlying processes that drive this variability. However, some amount of noise is always expected to be embedded in the geodetic data records. Analysis of such data records in the presence of noise often fail to give the required accurate spatial-temporal structures of interest to the space geodesy community.

It is therefore necessary to exclude the artifacts, systematic and manual errors by use of a robust *cleaning* tool. A robust platform for denoising the data which is used in this thesis is based on Wessel *et al.*, (2000). The advantage of using this methodology stems from the fact that the filter coefficients are spontaneously adapted in the event of the sudden changes in the time series.

Apart from the ordinary gap filling of data with missing data records, adaptive filtering proceeds via two important steps; the adaptive filtering and adaptive control procedures. In the adaptive filtering procedure, the adaptive second order properties such as the mean, μ_a^n and standard deviation, σ_a^n given in Equations (54) are computed from a reconstructed time series Y^{*k} (wherein the obvious errors such as gaps due to missing values) have been removed or filled.

$$\begin{aligned} \mu_a^k &= \mu_a(k-1) - q(\mu_a^{k-1} - Y^{*k} - 1) \qquad \text{k=1, 2, ... n} \\ \sigma_a^k &= \sqrt{\mu_a^k - \lambda_a^k} \end{aligned} \tag{54}$$

where $q \in \{0,1\}$ is the controlling coefficient and the adaptive second moment $\lambda_a^k = \lambda_a^{k-1} - q\left(\lambda_a^{k-1} - Y^{*k-1} \times Y^{*k-1}\right)$. Outliers are often identified using a filter constraint imposed on the raw data. The data point is an outlier if;

$$\begin{aligned} \left|Y^k - Y^{k-1}\right| &> \frac{\gamma Y^{k-1}}{100} + q_f \times \bar{\sigma}_a \\ \left|Y^k - Y^v\right| &> \frac{\gamma Y^v}{100} + q_f \times \bar{\sigma}_a \end{aligned} \tag{55}$$

where γ is the proportionality constant (~ 0.1), $q_f \times \bar{\sigma}_a$ is the generalised 3σ sigma rule. The last valid observation is denoted by Y^v while $\bar{\sigma}_a$ is the average of σ_a^n. A random number generated from Equation (56) is used for gap filling (replace all those values recognised as outliers).

$$\hat{\lambda} = \left\{ \mu_a^k - \frac{1}{2}\sigma_a^k, \mu_a^k + \frac{1}{2}\sigma_a^k \right\} \tag{56}$$

This gap filling procedure is used to avoid the false decreased variability that is often noticed after the adaptive filtering phase. In the adaptive control procedure, a percentage time series is build from the adaptively filtered time series. Thereafter, a new adaptive mean and standard deviation of the reconstructed percentage time series, $\left\{Y_k^{\%}\right\}$; $\forall k = 1, 2, \cdots, n$ of the adaptive filter and binomial-filtered series are calculated. Then, a constraint is imposed on the binomial-filtered series such that an outlier data point is detected using the following inequality:

$$\left|Y_a^{\%} - \mu_a^k\right| > q_{f_1} \times \sigma_a^k + \sigma_0. \tag{57}$$

Here, q_{f_1} and σ_0 are the filter coefficient and parameter that accounts for basic variability respectively. Equation (57) is introduced to dampen filtering errors due to minimal variability in the geodetic time series.

3.5. Data analysis strategies

The only links we have with the unexplained reality are the data and therefore the only way of investigating the underlying processes of any given phenomena is through data analysis, refer for example Lin *et al.*, (2009). Geodetic tropospheric parameter time series, such as tropospheric delay (and delay gradients), WV, tropospheric mean temperature and pressure consist of complex components which are manifestations of non-linear processes. The dynamics of the troposphere often evolve as a complex system with various spatio-temporal correlation scales that are either discrete (e.g., precipitation) or continuous (e.g., tele-connection patterns). These correlations often embed different components with, perhaps a wealth of unique statistical information about the interactions among the inherent tropospheric constituents: the geophysical signals. Traditional methods of determining characteristic time-frequency scales (e.g., Fourier and Principal Component Analysis) for each component involve decomposing the time series into component basis functions that satisfy two conditions; completeness of the basis and orthogonality. In terms of Fourier analysis, a given time series '$Y(t)$' is decomposed into global sinusoidal components of fixed amplitude a_j given by Equations (58) ,

$$Y(t) = \sum_{j=0}^{n} a_j e^{i\omega_j t};$$
$$a_j = \frac{1}{2\pi} \int_t Y(t) e^{-i\omega_j t} dt \qquad (58)$$

Equations (58) imply that the spectral amplitudes, a_j represent the energy contributed by a sinusoidal basis with frequency ω_j that spans the whole time series. The Fourier representation is most useful when the underlying geophysical process which causes variability in the time series is linear and therefore the superposition of the sinusoidal signals would make physical sense. As alluded to earlier, a_j remains time invariant thus $Y(t)$ is fairly constant. However, most of the geodetic time series do not meet this stationarity condition (they are non-uniform, non-linear and nonstationary). This would mean that the time series exhibits a broad spectral energy. In order to reconstruct the time series, global (e.g., harmonic) sinusoids are often required. Fourier transforms do not provide local features and therefore not suited for local description of the embedded dynamical structure of the observations.

3.5.1. Detrended fluctuation analysis

In this thesis, the presence or absence of random walk-type behaviour in troposphere WV is assessed using the DFA. The DFA methodology has been proven useful in revealing the extent of long-range correlations in diverse time series (e.g., Talkner and Weber, 2000; Király and Jánosi, 2005; Qian *et al*., 2008; Peña *et al*., 2009; Rybski and Bunde, 2009 and Varotsos *et al*., 2009). The DFA method is used to analyse WV fluctuations and also provide characteristics of the correlated stochastic components as well as effectively filtering out slow trends. The DFA approach handles nonstationary trends and also amplifies the intrinsic correlation structure of WV fluctuations of different time scales for analysis. The most important advantage of DFA over conventional methods such as autocorrelation and spectral analysis is that it has provision for the detection of intrinsic self-similarity that is embedded in the nonstationary WV. In the following, the general procedure of the DFA methodology is presented.

- Step 1: A fluctuating WV time series $Y_t \quad \forall t = 1, 2 \cdots T$ is integrated to determine the profile:

$$\chi_t = \sum_{t=1}^{i} (Y_t - \bar{Y}) \tag{59}$$

In Equation (59);

$$\bar{Y} = \frac{1}{T} \sum_{i=1}^{T} Y_t \tag{60}$$

- Step 2: χ_t is segmented into $K = \text{int}\left|\tau^{-1} T\right|$ non-overlapping time intervals, η_k of equal size τ where $k = 1, \cdots, K$. The above procedure is repeated from the other side of the series (from t=T, T-1, ..., T-(T-1)) in order to include all parts of the profile. This yields $2K$ segments.
- Step 3: For each of $2K$ segments, a local trend is calculated and a polynomial function of the form χ'_k is determined by the least-squares fit to the series. Thereafter, the variance is calculated using Equation (61)

$$F^2(\tau, k) = \frac{1}{\tau} \sum_{i=1}^{\tau} \left\{ \chi\left[(k-1) + i \right] - \chi'_k(i) \right\}^2; \tag{61}$$

for each segment, $k = 1, \cdots, K$ and

$$F^2(\tau, k) = \frac{1}{\tau} \sum_{i=1}^{\tau} \left\{ \chi\left[T - (k-K)\tau + i \right] - \chi'_k(i) \right\}^2, \tag{62}$$

for $k = K+1, \cdots, 2K$.

- Step 4: An m[th] order fluctuation is calculated by averaging each scale over all segments using Equation (63)

$$F_m(\tau) = \left\{ \frac{1}{2K} \sum_{k=1}^{2K} \left\{ F^2(\tau, k) \right\}^{m/2} \right\}^{\frac{1}{m}}. \tag{63}$$

In this report, m=2. Steps 2, 3 and 4 are repeated for several time scales in order to assess the dependence of $F(:)$ on the time scales.

- Step 5: The scaling behaviour of the WV fluctuations is then determined by analysing the log-log plots of $F(:)$ versus τ. Note that a power law relationship between F_m^τ and τ indicates the scaling with an exponent υ given by;

$$F_m^\tau \sim \tau^\upsilon. \tag{64}$$

Here, ν is a self-similarity parameter that represents the long-range power-law correlation in the data record. It is worth noting that if WV exhibits self-similar behaviour with $\nu > 0$ the fluctuations would grow with the window size in a power law way. This implies that the fluctuations on large observation windows exponentially grow faster than those with small windows. This would mean that WV fluctuations are unbounded. If $\upsilon = 0.5,$ the fluctuations are uncorrelated and are expected to be driven by processes that are a random walk and WV exhibit a Gaussian distribution; however, if $\upsilon < 0.5,$ the fluctuations are anti-correlated and for $\upsilon > 0.5,$ the signal is correlated. Processes exhibiting this behaviour have a power-law autocorrelation function expressed as;

$$C_\gamma = \left(Y_t Y_{t+\gamma}\right) \sim \gamma^{-\alpha}. \tag{65}$$

Here, $0 < \alpha < 1$. According to Talker and Weber, (2000), the relationship between the correlation exponents could be given by;

$$\frac{\alpha}{2} = 1 - \gamma \tag{66}.$$

3.5.2. Wavelet transform

The WT has been introduced and developed to study a large class of phenomena such as image processing, data compression, chaos, fractals, etc (Whitcher, 1998). Mallat, (1989) proposed a concept of multi-resolution analysis for constructing an orthonormal wavelet basis and further illustrated the wavelet multiresolution characteristic from the space aspect. As a

result, the works demonstrated the functions of wavelet theory in the frequency analysis of various data signals. Though recently developed, wavelets analyses techniques provide a powerful and insightful representation of the structure in data appropriate to both linear and nonlinear systems. The basic functions of the WT are related to the property of spatial-temporal-frequency localisation, contrary to what happens with trigonometric functions. The WT works as a mathematical microscope on a specific part of a signal to extract local structures and singularities. This makes the wavelets ideal for handling non-stationary and transient signals, as well as fractal-type structures.

Let $L^2(\mathbb{R})$ denote the two dimensional space of all square integral functions, $\varphi(t)$ with finite energy. If $\varphi(\mathrm{t}) \in L^2(\mathbb{R})$ is a fixed function, then the $\varphi(t)$ is said to be a wavelet if and only if its Fourier Transform (FT), $\hat{\varphi}(w)$ satisfies the permitted admissibility condition (also called complete reconstruction condition) given by Equation (67),

$$C_\varphi = \int_0^\infty \frac{\left|\hat{\varphi}(\omega)\right|^2}{|\omega|} d\omega, \\ < \infty. \tag{67}$$

Here, $\varphi(t)$ is the mother or basic wavelet. Equation (67) implies that the wavelet value is centred on the mean (see Equation(68))

$$\int_{-\infty}^{\infty} \varphi(t)dt = \hat{\varphi}(t), \\ = 0, \tag{68}$$

and therefore is oscillatory (some sort of a wave) as described by Daubechies, (1992), Mallat, (1999) and Qian, (2002).

If the flex (also called the dilation) and translation transform is applied to the mother wavelet $\varphi(t)$, then $\varphi(t)$ can be decomposed into some wavelet series $\phi_{(a,b)}$ defined such that;

$$\varphi_{(a,b)}(t) = \mid a \mid^{-0.5} \varphi\left(\frac{t-b}{a}\right). \tag{69}$$

Here, $b \in \mathbb{R}$ is the translation parameter and $a \in \mathbb{R}^{+}(a \neq 0)$ is the flex/dilation or scale parameter (this is the scaling in frequency range). The normalisation factor $a^{0.5}$ ensures that $\phi_{(a,b)}$ has the same energy along all the scales. Given that tropospheric WV data sets are represented by finite number of observations or measurements, the orthogonal (discrete) wavelets associated with orthonormal bases of $L^2(\mathbb{R})$ are often appropriately used for their analysis. Therefore, WT is performed only on a discrete grid of the tropospheric WV over some dilation and translation. This implies that a and b parameters take only integral values, where in general terms, the expansion of the WV time series, $Y(t)$ can be expressed by Equation (70)

$$\mathrm{Y(t)}=\sum_{n}\sum_{m} Y_n^m \varphi_{m,n}(t) \tag{70}$$

From Equation (70) the orthonormal wavelet basis functions are related according to;

$$\varphi_{m,n}(t) = 2^{\frac{m}{2}} \varphi\left(2^m t - n\right), \tag{71}$$

where m and n are the dilation and translation indices respectively. Equation (71) is derived from equation (69) when $a = 2^{-m}$ and $b = 2^m \times n$. At any particular wavelet level m, the contribution of a time series could be given by Equation (72),

$$Y_m(t) = \sum_{n} Y_n^m \varphi_{m,n}(t) \tag{72}$$

The significance of Equation (72) is that it provides temporal behaviour of the time series within different scales as well their contribution to the total energy WV time series. As discussed in Qian, (2002), the wavelet function $\varphi(t)$ is related to scaling function $\phi(t)$ and scaling coefficients a_n^m.

For a given wavelet basis to be as representative as possible, some degree of regularity is often desired. This condition is met by wavelets that exhibit n vanishing moments;

$$\int_{-\infty}^{\infty} Y^k \varphi_n(y)dx = 0 \;;\; \forall k = 0,1,\cdots,n-1, \quad \text{and} \tag{73}$$

$$\int_{-\infty}^{\infty} Y^k \varphi_n(y)dx \neq 0; \;\; \forall k = n. \tag{74}$$

Equations (73) and (74) imply that a wavelet with n vanishing moments is orthogonal to polynomials up to order $n-1$. Note that the admissibility condition imposes the condition that a wavelet ought to have at least one vanishing moment. In general, a wavelet transform of $Y(t)$ with a wavelet $\varphi_n(y)$ and n vanishing moments is simply a smoothed version of the n^{th} derivative of $Y(t)$ on various scales.

Here, we have employed the Haar wavelet as the analysing signal where a set of non-continuous (and therefore non-differentiable) functions whose mother wavelet takes the form of:

$$\varphi(t) = \begin{cases} 1 & 0 \le t \le 0.5 \\ -1 & 0.5 \le t \le 1 \\ 0 & otherwise \end{cases} \tag{75}$$

with a scaling function $\phi(t)$ described as,

$$\phi(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & otherwise \end{cases} \tag{76}.$$

Important features of the Haar wavelet include:

a) the basis functions are often expressed by linear combinations: $\phi(2^0 t), \phi(2^1 t), \phi(2^2 t), \ldots, \phi(2^k t), \ldots$ and their shifted functions,

b) the constant functions, $\varphi(2^0 t), \varphi(2^1 t), \varphi(2^2 t), \ldots, \varphi(2^k t), \ldots$ and their shifted function form are used for approximations,

c) they exhibit the orthogonality, $\int_{-\infty}^{\infty} 2^m \varphi(2^m t - n)\varphi(2^{m_1} t - n_1)\, dt = \delta_{m,m_1}\delta_{n,n_1}$, where $\delta_{i,j}$ is the Kronecker delta and

d) the wavelet and scaling functions are related as shown in equations (77) ;

$$\begin{aligned} \varphi(t) &= \varphi(2t) + \varphi(2t-1), \\ \phi(t) &= \varphi(2t) - \varphi(2t-1). \end{aligned} \tag{77}$$

The Haar wavelet transform cross-multiplies a function against the Haar mother wavelet with various shifts and stretches which are derived from the Haar matrix. See Equation (78) for a 2 by 2 Haar matrix sample;

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{78}$$

The Haar WT therefore aids the sampling processing in which rows of the transform matrix act as samples of the finest resolution (this is the basis for multi-resolution analysis in wavelet methodology).

In the analysis of scaling behaviour in WV, nonparametric estimators (e.g. DFA described above) were considered in this thesis. These estimators are based on fitting a power-law on the n^{th} order moment of the data values themselves or of their variations as a function of some scale/lag parameter. The approach has however two presuppositions for scaling processes. For long memory processes:-

a) a statistically sufficient evidence that the relevant points on the curve do indeed represent a straight is required, and

b) that the line's slope is such that $0.5 < H < 1$, where H is the Hurst parameter.

If WV data (of length N) is assumed to be stationary, then a simple sample estimator of the mean, μ_Y of a second order process Y_t is a reasonable choice. However, as $N \to \infty$, μ_Y follows a normal distribution with;

$$\begin{aligned} E[Y] &\sim \mu_Y \\ \text{var}[Y] &\sim \frac{\sigma_Y^2}{N} \end{aligned} \tag{79}$$

If Y_t exhibits self-similar behaviour, the sample mean is asymptotically and normally distributed with μ_Y but the variance is expressed according to Beran, (1994) by Equation (80),

$$\sigma_{LRD}^2 = \frac{2C_r N^{\alpha}}{(1+\alpha)\alpha} \frac{1}{N} \tag{80}$$

where $\alpha \in [0,1]$ and $C_r \in [0,\infty]$ parameters describe self-similar behaviour. Equation (80) implies that the variance of the sample mean decreases with sample size at a slower rate than the classical one with the ratio $\sigma_Y^2 : \sigma_{LRD}^2 \to \infty$ quickly with N. Clearly, computing any confidence interval of μ_Y would be biased. A robust approach that is capable of handling self-similar behaviour in WV time series could therefore be required. The wavelet estimator is one of the methods chosen in this thesis. Using the wavelet based approach; scaling properties in the data sets could be detected, identified and quantified. This is because, the WT often utilises an analysing *a priori* mother wavelet such as the Haar basic wavelet; which has inherent scale invariant properties. These semi-parametric estimators are computationally efficient and allow data analysis of arbitrary length. In addition, the estimators can also detect and isolate the deterministic components (trends) in the data.

There are many classes of scaling processes (Beran, 1994). In this thesis, it is desirable to distinguish between self-similarity and long range dependent processes. Self-similar (SS) processes (e.g. Fractional Brownian motion) are stochastic processes that are invariant in distribution under suitable scaling of time and space. A stochastic process $(Y_t, t \geq 0)$ is SS with exponent $H \in \mathbb{R}$ of SS if and only if all $c > 0$,

$$\left(Y_{ct}, t \geq 0\right) \overset{d}{=} c^H \left(Y_t, t \geq 0\right) \tag{81}$$

where $\overset{d}{=}$ indicates an equality in the statistical and/or distribution sense. For Gaussian processes with finite variance (these processes exhibit stationary increments), the following properties hold:

1. If $H < 0$, then $Y_t = 0 \ \ \forall t \geq 0$,
2. If $H = 0$ and $(Y_t, t \geq 0)$ is continuous probability, then $\forall t \geq 0$, $P(Y_t = Y_0) = 1$, which implies that $H > 0$ for this particular SS processes.
3. If for some $0 \leq \gamma \leq 1$, $E\left[Y_{t=1}\right]^{\gamma} < \alpha$, then $0 < H < 1$.

Some processes exhibit inbuilt memory which is dependent upon widely separated values that are significant even across large time shifts. Such stochastic processes are referred to as Long-Range Dependent (LRD) and their autocorrelations decay to zero slowly such that their

sum does not converge. Processes with long memory (or LRD) are stationary processes and contain spectral density that satisfies,

$$\rho_{\nu} \sim c_f |\nu|^{-\alpha} \quad \text{as } \nu \to 0 \tag{82}$$

where $\alpha = 0 < \alpha < 1$ (this describes the quantitative nature of the scaling) and $c_f > 0$ (is the measure of the size of the LRD and has the dimension of variance). Equation (82) implies that the auto-covariance, $r_{\tau} = E\left[Y_t Y_{t+\tau}\right]$ satisfies,

$$r_{\tau} \sim c_r \tau^{\alpha-1} \text{ as } \tau \to 0 \tag{83}$$

where $c_r = c_f 2\Gamma(1-\alpha)\sin\left(2^{-1}\pi\alpha\right)$, and Γ is the Gamma function (Beran, 1994). Equations (82) and (83) imply that the covariances, r_{τ} decays slowly. Increments of finite variances of SS processes have LRD as long as $0.5 < H < 1$, where H and α are related through

$$\alpha = 2H - 1 \tag{84}$$

Based on the wavelet analysis framework, the wavelet coefficients $d_{j,k}$ represent the difference between the aggregated time series by factors 2^{j-1} and 2^j for a fixed scale j. In this regard, the underlying assumption is that $d_{j,k}$ are short-range correlated. Given that $d_{j,k}$ are the wavelet coefficients at octave j, and if the mother wavelet has M vanishing moments and that its Fourier transform is M differentiable at the origin and m_j is the number of wavelet coefficients available at octave j, then $d_{j,k}$ is second order stationary. Furthermore, $E\left[d_{j,k}\right]$ can be estimated as reported by Abry *et al.*, (2000) by;

$$\mu_j = \frac{1}{m_j}\sum_{k=1}^{m_j} d_{j,k}^2 \tag{85}$$

In addition, the estimator of the log $E\left[d_{j,k}\right]$ is,

$$s_j = \log \mu_j - \frac{1}{m_j \log 2} \tag{86}$$

where the last term of right-hand side cancels the bias contributed from the nonlinear component of log2. As a result, a plot of j versus s_j yields the log-scale diagramme as described by Abry and Veitch (1998).

The log-scale diagramme of the coefficients of the WT was used to analyse WV fluctuations and investigate the presence of two most important self-similar behaviours; the LRD and SS. The wavelet-based estimator of the LRD and SS is based on the discrete wavelet transform, DWT. The analyses of the WV's LRD/SS and other derived parameters are based on the following procedure. Firstly, the data is discretely pre-filtered to eliminate outliers in the WV sequence. Thereafter, the DWT of the pre-filtered WV data series is computed and then the squares of the coefficients of WT are averaged. A linear regression on the log of the averaged coefficients of the WT (plotted on the y-axis) and the log of the scale (plotted on the x-axis) is fitted. In this regard, the log-scale diagramme was used to:

a) select the scale range where scaling is observed, and

b) estimate the scaling properties in the coefficients of the WV.

It is assumed that, a scaling phenomena could occur over a range of scales, $j = \{j_1, j_2\}$ and therefore for LRD processes, j_2 is infinite but j_1 is where the LRD begins (this value has to be selected). However, for SS processes, $j = \infty$ as j_2 remain infinite (Abry *et al*., 1999).

3.5.3. Hilbert-Huang transform

Geodetic data collection, pre-processing, analysis and visualisation of the inherent signal structure by use of DFA and WT methodologies often assume that the underlying processes are weakly stationary. Ideally, stationarity in geodetic data and tropospheric WV fluctuations cannot be guaranteed. In order to accommodate the inherent nonlinear and nonstationary properties of WV sequence, the reported research utilised an objective and flexible method that could describe the oscillatory events in WV fluctuations whose associated time-frequency characteristics evolve over time called the HHT. The HHT approach is able to deal with WV fluctuations in the multiple resolutions and therefore distinguishes different processes driving variability.

The gist of the HHT is the EMD whose basic concept involves the empirical identification of oscillatory modes in the data by means of the local extrema. The decomposition is based on the assumption that:-

a) the data must have at least two extrema,

b) there exists a characteristic time-scale defined by the time interval between consecutive extremes; and

c) the presence of an inflection point (no extreme) requires interpolation in order to obtain the extrema.

The EMD approach assumes that the target data set consists of different, simple and intrinsic modes of oscillation that need not be sinusoidal (e.g. slowly varying amplitude and phase), called *IMF*. Each *IMF* ought to satisfy two criteria (to resemble the generalised Fourier decomposition); a) an *IMF* may only have one zero between successful extrema; and b) an *IMF* ought to have zero local mean.

The EMD adaptive process is a recursive 'sifting' algorithm described by e.g., Huang *et al*., (1998) and Pegram *et al*., (2008). Given a time series $\{Y_t, t \geq 0\}$, the recursive 'sifting' procedure can be summarized as follows:

1. Take the input signal Y_{t-1} to decompose, where Y_0 is the original time series;

   1.1. identify the local extrema of Y_{t-1}

   1.2. construct the upper/lower envelope ($\Omega_{u,t}$ / $\Omega_{l,t}$ by interpolation

   1.3. approximate the local average envelope by $\mu_{\Omega_t} = 0.5\left(\Omega_{u,t} + \Omega_{l,t}\right)$

   1.4. extract the detail $d_{t,1} = Y_{t-1} - \mu_{\Omega_t}$

   1.5. If $d_{t,j}$ is an *IMF*, decompose Y_{t-1} into an *IMF* i.e. $IMF_t = d_{t,j}$ and the residual $Y_t = Y_{t-1} - IMF_t$. Otherwise repeat steps 1.1 through 1.5.

2. If Y_t has an implicit oscillation, set Y_t as an input signal and repeat from step1.

If the IMF_s are added together with the residual trend, the original signal is often recovered without any distortions or loss of information as shown in Equation (87),

$$Y_t = \sum_j \left(IMF_j\right) + Y_{res} \tag{87}$$

A key advantage of EMD is that *IMF*$_s$ can be transformed from the temporal-space to time-frequency space by applying the Hilbert Transform (HT) to each *IMF* component determined by Equation (88)

$$\begin{aligned} Y_t^H &= H\left[Y_t\right] \\ &= \frac{1}{\pi} pv \int_{-\infty}^{\infty} \frac{Y_\tau}{t-\tau} d\tau \end{aligned} \tag{88}$$

where pv is the Cauchy principal value or principal value of the singular integral. Note that Y_t and Y_t^H form a complex conjugate pair. Based on HT, the analytic signal is defined by,

$$\begin{aligned} z_t &= Y_t + iY_t^H \\ &= A_t e^{i\theta_t} \end{aligned} \tag{89}$$

where the instantaneous amplitude and phase are given by $A_t = \sqrt{Y_t^2 + Y_t^{H2}}$ and $\theta_t = A_t \tan\left(Y_t^{-1} Y_t^H\right)$ respectively. From Equation (89), the instantaneous frequency (which is also a function of time) of each *IMF* can be defined as

$$\omega_t = \frac{d\theta_t}{dt} \tag{90}$$

This implies that the HT of the n^{th} *IMF* components of Y_t can be written as:

$$Y_{n,t}^H = \sum_{j=1}^{n} A_{j,t} e^{i\int \omega_t^j dt}, \tag{91}$$

where $A_{j,t}$ is the amplitude of the analytic signal associated with j^{th} IMF. It is worth mentioning that the $\{A_{j,t}, \theta_{j,t}\}$can be projected on the time-frequency-energy (=$|A_{j,t}|^2$): forming the Hilbert-Huang spectrum. This spectrum has the same information as in the continuous WT reported in Torrence and Compo, (1998).