4. Modelling the stochastic properties of WV time series

A mathematician is a device for turning coffee into theorems.

- Paul Erdős

In this Chapter, the stochastic properties of WV are investigated. Firstly, results reveal that the underlying process driving fluctuations in the monthly averaged tropospheric WV derived from geodetic VLBI measurements spanning 1999-2008 is non-linear and nonstationary. Furthermore, an ARMA(10, 9) could generally model the monthly averaged tropospheric WV (transformed to stationary) over the period 1999-2008.

4.1. Introduction

A time series represents a path (also called a realisation) of a stochastic process or a sequence of data-points measured at successive time intervals. The collection of variables indexed according to the order in which they are obtained in time, forms the basis for the statistical description of the data which might have inherent spatial-temporal fluctuations (Shumway and Stoffer, 2006). The purpose of time series analysis is to develop mathematical models that provide robust descriptions about the observables; and as a result obtain an understanding of the mechanism that generated the data records (e.g., the nature and structure of the underlying forces). Analysis of time series could also involve fitting models to the data for purposes of future predictions of the phenomena in question. In geodetic applications, tropospheric characteristics can be presented as a vector time series that hold geophysical information which is of interest to the space geodesy community. The tropospheric state at each moment in time can be modelled as a mapping function denoted as $\Re^3 \rightarrow \Re^N$; which assigns an n-dimensional (nD) vector of real tropospheric parameters such as temperature, pressure, refractivity, WV and tropospheric gradients to every point of the 3D space, herein referred to as a time series or data sequence.

The problems encountered in time series modelling and prediction dates back to the pioneering work of Yule in 1972 (Yule, 1972). Until 1970's, most of the research work on time series analysis concentrated on the use of parametric methods to describe underlying process in the observed data. In these parametric approaches, simple linear models are fitted to the data; see for example the text-book by Brockwell and Davis, (1996). Parametric

approaches such as the maximum likelihood estimate (MLE) are commonly used when there is sufficient prior knowledge that the model in the data set has a parametric form with unknown parameters i.e., that the model is from some parametric family-set $\{\mathbb{Z}_\theta, \theta \in \Theta\}$, where $\mathbb{Z}_\theta$ is a known parametric form with unknown parameter θ to be estimated. Though the parametric approaches have established appealing mathematical properties over time, they often impose unsound rigid structure upon the underlying process. To study nonlinear time series, nonparametric models such as the Multivariate Adaptive Regression Splines, (MARS) (Chen *et al*., 1997) and the EEMD (Zhaohua and Huang, 2009) that do not impose any structural assumptions have been developed to model underlying processes. Nonparametric models are often formulated based on the principle of *"letting data speak for themselves."* Nonparametric models are therefore useful when little information or when flexibility about the underlying model is required.

One aspect of investigating a time series involves finding appropriate models for the time series. Generally, for a given model, the central theme is to estimate the unknown quantities of the model based on discrete observations. The process often involves model identification, fitting and model diagnostics. Given some data, there are often an infinite number of models or hypothesis that fit the data equally well. As a result, there is no reason to prefer one model over another. Therefore, one is forced to make assumptions that lead to an inductive bias. In model selection, the model parameters are selected such that a model of optimal complexity for a given (finite) data is created. Such models are said to have the correct inductive bias. Additional details on model section can be found in the text-book by Burhan and Anderson (2002).

Time series models have been central in the study of some behaviour of a process or metric over a period of time. The application of time series models are manifold, the applications range from geophysical problems such as daily weather forecasts, electricity (Taylor, 2008) and astronomy (Subba and Priestly, 1997). In decisions that involve some element of uncertainty of future values, time series models been found to be robust methods of forecasting. Additionally, time series models can be used to understand the structure in the data and predict the future trends and patterns in the data. Bates, (1994) analysed the adjusted time series of global WV derived from satellite infra-red observations and reported the linkage between upper troposphere WV time series seasonal component and the monsoon circulations. In addition, the inter-annual variability of WV was found to be correlated to the

ENSO warm-cold events. Based on the notion that all of the power in the integrated WV variations is contained in a synoptic-scale motion of air-mass, Davis, (2001) investigated the statistics of integrated WV time series using GPS data in order to assess the large-scale weather systems. It was found that the power spectral density could be a robust estimator of the integrated WV than the structure function is over long time scales.

Kruger, (2006) examined the spatial-temporal variations of trends in daily extreme precipitation indices for 138 rainfall stations for the period 1910 to 2004 in South Africa and reported of some certain areas where significant changes in certain characteristics of precipitation amid the lack of real evidence of overall changes in precipitation over the past century. Documented studies on the spatial-temporal characteristics of WV assume that the variability of WV is driven by stationary processes. Despite all these areas of application as mentioned above, assessment of the stochastic and self-similar properties of tropospheric WV has remained unexplored. In addition, there is no literature known to the author that has reported on the model of WV variability.

To characterise the stochastic behaviour of WV fluctuations, a general Auto-Regressive Moving-Average (ARMA) time series model will be considered. In the analysis of WV fluctuations, an automatic algorithm could be used to estimate the appropriate model parameters such as the model order from the tropospheric WV and the estimated model could be used to investigate the nature of the underlying process that drives the variability of tropospheric WV.

4.2. Basic concepts of time series analysis

If a random variable Y varies with time, then a simple time series, expressed as $Y_t=\{ y_t, \forall t=1, 2, ...\}$ (here $\forall t$ denotes for all integer valued time index) assumes that the measured data points are realizations of random processes (defined in the next section as) y_t that comprises of four components as shown in Equation(92) (Trömel and Schönmise, 2006);

$$Y_t=T_t+C_t+S_t+R_t. \qquad (92)$$

Equation (92) is similar to the model proposed by Li *et al.*, (2000) to investigate the presence of secular tectonic deformation fields and to distinguish between tectonically active and inactive regions in central Japan using GPS data. The additive model-components, T_t, C_t, and S_t in Equation (92) refer to the trend, non-random long term and short (seasonal) periodic

components respectively. The random variable R_t accounts for all the deviations from the ideal non-stochastic components. In many time series analysis strategies, the assumption that $E(R_t)$ exist or can also achieved by modifying one or more of the non-random components. Furthermore, if the underlying process is dominated by the growth component T_t only then:

$$\begin{aligned} E(Y_t) &= T_t, \\ &= f(t). \end{aligned} \tag{93}$$

Though the function f(t) is known, it is dependent on the unknown elements of the parameter space $\{\beta_1, \beta_2, \cdots, \beta_q\}$ such that,

$$f(t) \equiv f(t; \beta_1, \beta_2, \cdots \beta_q). \tag{94}$$

Using the ideal realisations, y_t the parameters, $\{\beta_1, \beta_2, \cdots \beta_q\}$ can be determined based on the least squares estimate, i.e.,

$$\sum_t \left(y_t - f(t; \beta_1, \beta_2, \cdots \beta_q) \right)^2 = \min_{\beta_1, \beta_2, \cdots \beta_q} \sum_t \left(y_t - f(t; \beta_1, \beta_2, \cdots \beta_q) \right)^2. \tag{95}$$

A feasible solution from the numerical problem given in Equation (95) entails;

$$y_t = f\left(t; \beta_1, \beta_2, \cdots \beta_q\right), \tag{96}$$

as the predictand of $y_t \; \forall t=1, 2, ...$; where t=1 is the current time. Therefore, the residuals in the realisations i.e., $\sim y_t - y_t$, possess the goodness of fit information of the model to the data.

4.3. Random variables

Over the probability space $\{\Omega, \mathcal{P}\}$, a discrete random variable Y is a function that maps a space of events Ω to the real axis $Y : \mathcal{P} \to \mathbb{R}$ by $\aleph \to Y(\aleph)$ where $\aleph \in \Omega$ is a particular elementary event. As a result, the probabilities $\mathcal{P}_Y(\aleph)$ map each $\aleph$ result between $0 \le \aleph \le 1$. A particular $\aleph$ with a probability $\mathcal{P}_Y(\aleph)$ is obtained based on the realisation of Y. The relative frequency $\mathcal{P} = \frac{N(\aleph)}{N}$ is the ratio between realisations of $\aleph$ and the total number of realisations N.

On the continuous space, $\mathcal{R}$, the definition of the random variable Y has the corresponding probability density function, $\rho_Y(y)$ with $y \in \mathcal{R}$. The probability distribution is therefore the integrated probability density function given in Equation (97)

$$P_Y(\lambda) = \int_{-\infty}^{\lambda} dy \rho_Y(y). \tag{97}$$

A value of $y < \lambda$ results from a realisation of Y with a probability $\mathcal{P}\{y|_y < \lambda\}$ based on the probability distribution function given by equation(98) where a typical example is one that exhibits a Gauss distribution with a density function given by Equation(99);

$$\mathcal{P}\left\{y|_y < \lambda\right\} = P_y(\lambda) \tag{98}$$

$$\begin{aligned} \rho_Y(y) &= N\left(\mu, \sigma^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{y-\mu}{2\sigma^2}} \end{aligned} \tag{99}$$

4.5. Stochastic processes

A stochastic process Y_t is a time dependent random variable with some dependency structure. A random process is often described by a probability density function $\rho(y_t)$ where, only the probability distributions are computed as compared to the deterministic case where a given outcome is determined by initial conditions and equations. A stochastic process Y_t is denoted as an indexed collection of random variable with $i = t_1$ indices specifying some time ordering in the discrete or continuous space. Two data records observed between t_1 and t_2 might exhibit linear dependency that could be described by an auto-covariance function given by;

$$Cov(t_1, t_2) = \left\langle \left(Y_1 - \langle Y_1 \rangle\right)\left(Y_2 - \langle Y_2 \rangle\right) \right\rangle.$$

If the joint probability distribution is time invariant, the stochastic process is strictly stationary for any moment of time $\left\{t_i, t_j\right\}$. A second-order (weakly) stationary process exhibits a constant mean and the auto-correlation function depends only on the lag.

One way to describe a stochastic process is to specify the joint probability distribution of $Y_{t_1},...,Y_{t_n}$ for any set of times $t_1,...,t_n$ and any value n. However, a simpler, more useful way of describing a stochastic process is to obtain the moments of the process which are the mean, variance and the auto-correlation function. Stochastic processes can be characterized by stationary, auto-covariance function a spectrum and ergodicity properties. Stationary processes play an important role in the analysis of time series. However many observed time series are nonstationary in nature. Therefore the stationary property (which means that $\rho(y_{t_i}, y_{t_j}, ...)$ is invariant) in real data is strictly rare. Instead, a first order stationary process exhibiting a time independent mean is more common. If a process also has time independent

variance and its covariance function is a function of time differences only, then the process is described as weakly stationary or second order stationary.

$$\begin{aligned} \left\langle Y_{t_i} \right\rangle &\equiv \mu \\ &= C_{\mu}, \\ \mathrm{Var}(Y_i) &\equiv \sigma^2 \\ &= C_{\sigma}, \\ \mathrm{Cov}(\tau) &= \mathrm{Cov}(Y_i(t_i), Y_j(t_j)). \end{aligned} \tag{100}$$

Here $C_{\mu,\,\sigma}$ are constants. The auto-covariance functions of a stationary process could be;

$$\mathrm{Cov}(\tau) \equiv \left\langle (Y_{t_1+\Delta t} - \left\langle Y \right\rangle)(Y_{t_1} - \left\langle Y \right\rangle) \right\rangle, \tag{101}$$

This auto-covariance function can be normalized to the variance to obtain the auto-correlation function, $C(t) = \left[\mathrm{Cov}(\tau_0)\right]^{-1} \mathrm{Cov}(\tau)$. Furthermore the spectrum $S(\omega) = \mathrm{Cov}(\tau)$ is frequency-domain equivalent of the auto-covariance. In order to represent a stochastic process, multiple realisations are often averaged. This gives rise to the power spectral density given by,

$$S = \lim_{T\to\infty} E\left[\frac{G_T(\omega)^2}{2T}\right], \tag{102}$$

where $G_T(\omega) = (2\pi)^{-0.5} \int_{-T}^{T} Y_T(t) e^{-i\omega t} dt$ is the Fourier integral of the stationary process $Y(t)$. The Fourier transform of the auto-covariance function is related to the spectral density as follows;

$$\mathcal{S}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} COV(\tau) e^{-i\omega t} d\tau. \tag{103}$$

Equation (103) implies that the spectral density and the auto-covariance function describe the linear dynamic properties of a process equivalently. Sometimes real data may exhibit a temporal mean and auto-covariance function that converges to a value equivalent to the average of a set of realisations (say for example, ensemble average). This property is referred to as the ergodicity of the underlying process.

A stochastic process is called short-range correlated or Short-Range Dependent (SRD) if the auto-covariance function, $\rho_{SRD}(\tau)$ is summable and decays exponentially,

$$\sum_{\tau=-\infty}^{\infty} \rho_{SRD}(\tau) = \mathrm{const} < \infty. \tag{104}$$

Not all data records exhibit SRD, instead they posses the long-range dependence which can be described by an algebraic decay of the auto-covariance function; $\rho_{LRD}(\tau) \propto \tau^{\gamma}$. Such type

of processes are called long-memory processes or long-range correlated and their decay exhibits a diverging sum as shown in Equation (105)

$$\sum_{\tau=-\infty}^{\infty} \rho_{\mathrm{LRD}}(\tau) = \infty. \tag{105}$$

It is worth mentioning that long-memory processes are non-periodic stochastic processes. Dependence over infinite time lags due to the deterministic behaviour in the data record does not influence the long-range correlated datasets.

4.6. Geodetic parameters time series

Geodetic time series is derived from a sequence of geodetic data sets observed over a period of time and arranged according to observation time. There are numerous reasons to record and analyze the geodetic time series data. For instance, geodetic time series analysis could be vital in understanding the structure of the processes that generate the data as well as aid in the prediction of future values. The characteristic property of a time series is that, the data records are not independently generated; they exhibit a time dependent dispersion and are often governed by trend and may also exhibit periodicity. In geodetic applications, there are two steps of geodetic time series analysis. The first is identifying the nature of the process represented by the sequence of observations and, secondly prediction of the future values. Both of these goals require that a pattern of time series data is observed and described. Thereafter, the internal pattern (which could be the autocorrelation, trend or seasonal components) could then be interpreted and integrated to formulate models in geodetic time series that could be vital for tropospheric modelling which improves the accuracy of the geodetic delay observable.

In general, analysis of the geodetic time series reported in this thesis, focuses on the estimation and extraction of the deterministic (e.g. trend and seasonal) components in the geodetic (e.g. WV) data, see the first three terms in Equation(92). The components in geodetic WV time series could be used to determine the best model representing their variability. In the analysis of geodetic data, it is assumed that the bias term (which is contained in last term in Equation (92)) turns out to be a stationary random process. Thus, the theory of random processes can be used to find a satisfactory (and probabilistic) model for the bias term, analyze the properties of bias term and use it in conjunction with the deterministic components to predict the observed geodetic WV series. Alternatively, analysis of the observed geodetic WV series could be approached from Box and Jenkins paradigm

(Box and Jenkins 1970). In this model, the difference operators are applied repeatedly to the observed series until the differenced observations resemble a realisation of some stationary processes, from then on, the theory of stationary process is used model and analyse the stationary (and therefore the original) series.

4.6.1. Time series analysis of tropospheric WV

Strategies for quantifying the overall transient variability of WV are not straightforward, and a number of statistical approaches for modelling WV variability have been attempted. Gierens et al. (1997) used the measurements of about 2000 flights within the Ozone and WV by Airbus in-service aircraft (MOZAIC) program and confirmed that fluctuations of humidity and temperature from their local means could be characterized by occasional large fluctuations (i.e. heavy-tailed distributions) in the upper troposphere (at pressure levels 166 to 290 hPa on the general circulation model grid scale). It was found that the fluctuations could then be modelled by the Lorentz distribution rather than the Gauss distribution and this was due to large excursions in the fluctuations of humidity and temperature. Later, data from 3 years of MOZAIC measurements (this is data is described in Gierens et al. 1997) was used to determine the nature of the distribution law of WV; which plays a vital role in testing whether the hydrological cycle in climate models is adequately represented. It was reported that the frequency of occurrence of relative humidity greater than 100% decreased exponentially above ice saturation and that it decreases exponentially for the entire range of values in the lower stratosphere (Gierens *et al*. 1999). A stochastic source-sink model capable of producing such distributions was then formulated.

Data from NASA's Pacific exploratory mission in the tropics phase A, that was conducted between August and September 1996 was used to study the impact of human activity on tropospheric chemistry in the remote regions over the pacific (Cho and Newell, 2000). Based on the empirical multifractal formula for the structure function originally described by Pierrehumbert, (1996), an "anomalous scaling" or multi-fractality between 50 to 100 km horizontal range of the WV distribution was reported (Cho and Newell, 2000). From these findings, it was noted that while WV increase was statistically stationary, the transient WV fields did not exhibit the stationary properties. As a result, the probability distribution (e.g., the variance) of transients could therefore not be characterised accurately from a finite number of observations.

Recently, Jin *et al*., (2009) used the precipitable WV time series determined from co-located space geodetic techniques to quantify the systematic biases between VLBI and GPS in the 5-year co-located measurements. The reported results demonstrated that systematic biases in the geodetic data that describe the atmosphere systems and processes could be accounted for if the co-located observations are utilized. Due to the role played by atmospheric WV in Earth's atmospheric radiation budget, global hydrological cycles and global climate change (Suparta *et al*., 2009), these findings have important applications in weather forecasting, numerical weather prediction and climate change studies as discussed by Gettelman and Fu, (2008).

In this section, the analysis of geodetic WV time series is limited to, a) determining or transforming of the geodetic WV time series to stationarity, b) detecting seasonality using the autocorrelation, partial autocorrelation and automatic spectral plots, and c) deducing the inherent stationary model in the data. In the present analysis, geodetic WV data is assumed to exhibit a systematic pattern and random noise (error). In order to observe the pattern more clearly, some form of noise filtering is done by use of point-averaged smoothing. This methodology involves fitting some function or adjusting/correcting for the trend in the data records. To this end, an adaptive filtering algorithm reported in Wessel and Voss (2000) and described in detail in Section 3.4.2 is applied in smoothing the geodetic WV time series. For instance, Figure 4.1 depicts a sample wet tropospheric linear horizontal gradient data set that has been adaptively filtered in order to eliminate measurement or systematic errors. The filtering procedure employed here is based on the adaptive cumulates (the mean and standard deviation); used as filter coefficients which were adapted spontaneously during the computation process. This is a robust approach because it caters for the sudden changes in the time series. In the current filtering procedure, basic variability of zonal gradients is calculated using a binomial-7-filtering given in equation(106).

$$Y_t^{\Delta} = \frac{Y_{t-3} + 6Y_{t-2} + 15Y_{t-1} + 20Y_t + 15Y_{t-1} + 6Y_{t-2} + Y_{t-3}}{64} \qquad (106)$$

The adaptive mean and standard deviation are calculated and observations that are flagged anomalous based on the procedure described in Section 3.4. The reconstructed series is used to compute the basic variability, the new adaptive mean $\mu_t^{a'}$ and standard deviation $\sigma_t^{a'}$ which are then used to test for anomalous values using the inequality,

$$\left|Y_t^{\Delta} - \mu_t^{a'}\right| > \sigma_t^{a'} k_{f1} + \sigma^k. \qquad (107)$$

Here the filter coefficient vary as $2 \le k_{f1} \le 5$. The diurnal fluctuations of the zonal gradients are represented by a basic variability $\sigma^k \sim 365^{-1}$. Anomalous values are replaced with the respective values determined *a priori*. The filter coefficients take empirical constant values that range between 2.5 and 5.0 (these values are arbitrary determined). The reconstructed series plotted in Figure 4.1 used a filter coefficient that was arbitrary set to 3.0. In order to minimize the filtering biases, an adaptive basic variable of 0.011 was considered. This value was selected because the tropospheric linear horizontal gradients considered in this analysis are often sampled 4 times daily. The weighting factor of 365^{-1} was used to account for daily fluctuations in the observed series. As depicted in Figure 4.1, the unfiltered zonal gradients have anomalous values in the first and last quartiles. These values are not removed from the series. Instead, interval filtering described in Wessel *et al*., (1994) is used to adjust the anomalies. The interval filtering is done through the spontaneous adaptation of the filter coefficients due to the sudden changes in the observed series.

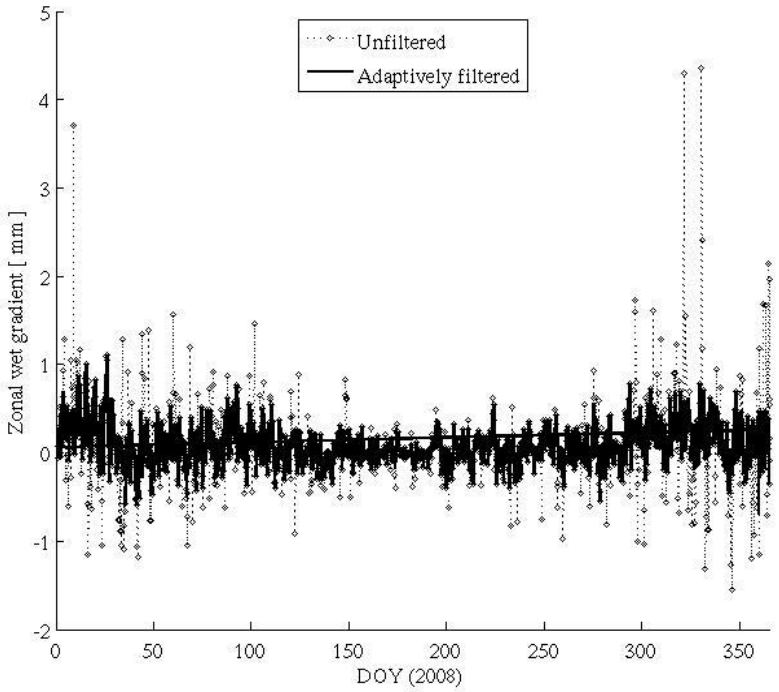

Figure 4.1 Adaptive filtering applied to the zonal linear horizontal wet gradient component observed over HartRAO. The linear zonal gradient has not been detrended.

4.6.2. Investigating stationarity in tropospheric geodetic WV

In this subsection, the stationarity in the geodetic tropospheric WV is examined and described by use of classical Box-Jenkins test. If the test for stationarity fails, then the geodetic WV ought to be transformed into a stationary time series before a model for stationarity is formulated. In the current analysis, stationarity in geodetic WV time series is identified by examining the behaviour of the sample auto-correlation function. The auto-correlation function measures the correlation between Y_t and $Y_{t+\tau_{acf}}$ (where τ_{acf} is time lag). In other words, the auto-correlation function of the geodetic WV time series describes the correlation (which also refers to the degree of dependence) that exists between geodetic WV time series and the same geodetic WV time series but lagged by 1, 2,.., τ_{acf} . The auto-correlation function is then depicted graphically by use of a histogram both quantitatively and qualitatively by, a) determining the period of the oscillation and b) by looking at the shape of the autocorrelation plot which gives some indication of the suitable model parameters of the time series model (e.g., autoregressive (AR), moving average (MA) or autoregressive-moving average (ARMA)). In most data sets, the auto-correlation coefficients are significant for a large number of time lags, τ_{acf}. However, due to the propagation of the auto-correlation at τ_{acf}, the auto-correlation at τ_{acf} >1 might not be explicit. As a result, the partial auto-correlation plot is used to ascertain whether auto-correlation at τ_{acf} *>1* indeed represents the auto-correlation at τ_{acf} =1. The partial auto-correlation plot is derived from the partial auto-correlation function. The partial auto-correlation plot often has a spike at τ_{acf} =1 which could imply that all the higher order auto-correlations are effectively explained by the auto-correlation at τ_{acf} =1.

In most time series analysis, there is no theoretical reference result that could be used to select and characterize a stochastic process. In this thesis, we consider a general model ARMA *{p, q}* and then automatically identify the appropriate model parameters hereafter, the model orders, *{p, q}* based on the underlying temporal pattern embedded in the time series data through:- a) identifying the appropriate time dependence by ensuring that the auto-covariance, auto-correlation and the partial auto-correlation described in the time series model represents a stationary stochastic process, b) ensuring that the filtered biases from the appropriate models are normally distributed, and c) ensuring that the model choice is parsimonious. After obtaining stationary geodetic WV time series, the sample auto-correlation function and partial correlation function could be used to identify a Box-Jenkins model that appropriately describes the geodetic WV time series. In order to assess the

specific statistical structure of geodetic WV sequence, a stochastic approach that models the dependent structure embedded in geodetic WV time series is considered. As a result, the Box-Jenkins statistical methodology briefly described in the following is used.

If the values of geodetic WV series are denoted by Y_t, Y_{t-1}, Y_{t-2}, then Box-Jenkins methodology is based on an ARMA *(p, q)* model given by Equation (108);

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} + U. \quad (108)$$

Here the a_t's are independent and identically distributed random shocks with a zero mean and a finite variance σ_a^2. Further, $\phi_i \ \forall i = 1,2,\ldots,p$ and $\theta_j, \ \forall j = 1,2,\ldots,q$ respectively denote the AR and MA coefficients while *U* is the model constant which is related to the mean of geodetic WV series. In the model characterized by Equation(108), the current geodetic WV series observation, Y_t is explained by a linear combination of the *p* previous observations, Y_{t-1}... Y_{t-p}, a linear combination of the *q* previous random shocks a_{t-1}..., a_{t-q} and a constant term *U*. The error term is given by a_t. If *q=0*, a pure AR (*p*) is derived and if *p=0*, the class of pure MA (*q*) is retrieved. The backward shift operator, *B* can be used such that $BY_t=Y_{t-1}$, a compact ARMA (*p, q*) model then becomes;

$$\phi(B)Y_t = U + \theta(B)a_t. \quad (109)$$

Here, $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$ represent the AR and MA operators respectively. For stationarity, $\phi(B)$ roots ought to lie outside the unit circle. The model is described by Equation (109) could be visualized in the sample and partial autocorrelation plot with characteristics given in Table 4.1.

Table 4.1 Sample AC and PAC model behaviour

Model	**SAC**	**SPAC**
MA	Cuts off at lag p	Tails off
AR	Tails off	Cuts off at lag p
ARMA	Tails off	Tails off

From Table 4.1, the plots of the SAC and SPAC could be visualised and a pure AR (*p*) and/or MA (*q*) processes established. However, estimates of (*p, q*) are not trivial from ARMA (*p, q*) processes. In the current analysis, a large number of candidate geodetic WV time series

models are computed and some statistical criteria (e.g., AIC (Akaike, 1969) or SIC in equations(110)), where μ_t and $\sigma^2=\sum_{t=1}^{T}\left(\mu_t\ T^{-1}\right)$ are the estimated residuals and variances respectively) is used to select a suitable model that is representative of the data. To this end, an ARMA (*p, q*) model is estimated based on the method of maximum likelihood. The likelihood function of the ARMA (*p, q*) model is non-linear in the unknown parameters and therefore non-linear optimization techniques are often used to solve for the unknown parameters.

$$\begin{aligned} \mathrm{AIC} &= \log\tilde{\sigma}^2+\frac{2}{T}(p+q+1); \\ \mathrm{SIC} &= \log\tilde{\sigma}^2+\frac{(p+q+1)\log T}{T}. \end{aligned} \tag{110}$$

In the current case study, WV time series data (for the period 1998 to 2008) derived from surface temperature measurements at HartRAO and numerical simulations of the ECMWF was used to investigate the stationarity properties. From the results, some indication of the broad correlation characteristics were averred from the sample auto-correlation and partial auto-correlation plots of the tropospheric WV time series as depicted in Figure 4.2. From the sample autocorrelation plot, it is evident that the line graphs exhibit damped oscillations which are the absolute sinusoidal components. These oscillations tail off slowly to zero, therefore indicating that the fluctuations in geodetic tropospheric WV time series are driven by non-stationary stochastic processes. In the present analysis, the auto-correlation function plot of the WV series exhibits spikes at several lags (e.g. at lags 1, 2, 4, 5, 6, 7, 8, 10, 11 and 12). This is expected since WV time series possess inherent diurnal, seasonal and trend components. One way to formulate a model for data with such auto-correlation pattern is to use the Space Time Auto-Regressive Moving Average (STARMA) models proposed by Pfeifer and Deutsch, (1980). However, the STARMA models do not take into account the embedded nonlinear behaviour that is representative of the underlying process. In particular, dynamical processes with unusual jumps cannot be effectively studied using STARMA models.

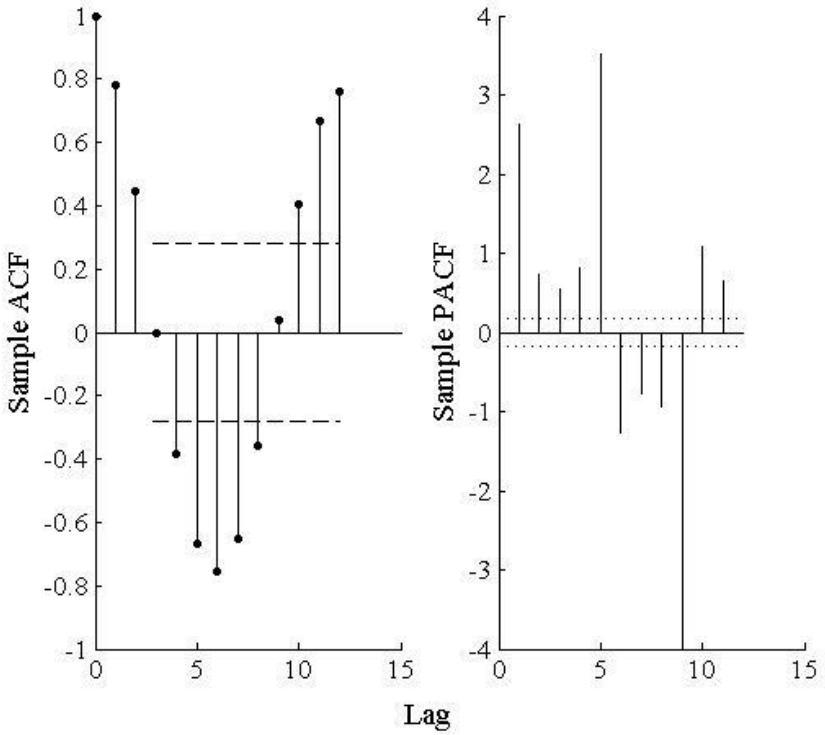

Figure 4.2 Sample autocorrelation and partial autocorrelation function for WV

To use ARMA models adequately, the WV time series must be stationary with respect to the mean and variance. Differencing between consecutive observations is one way to efficiently achieve stationarity. If the difference operator is defined as $\nabla = 1 - B$, such that $\nabla Y_t = Y_t - Y_{t-1}$ and the differencing goal is to stabilize the mean, then the model corresponding to the original WV series could be called Auto-Regressive Integrated Moving Average (ARIMA) model. An ARIMA (*p, d, q*) model is defined by the equation;

$$\phi(B)\nabla^d Y_t = U + \theta(B)a_t. \qquad (111)$$

Here, the d^{th} is the difference of the original series $\nabla^d Y_t$ is stationary and could be represented by a stationary ARMA (*p, q*) model.

The WV time series is often computed daily, monthly, quarterly or annually. Therefore WV series exhibit strong diurnal, seasonal or annual periodic fluctuations that often recur at pre-determined phases. In addition, it is expected that seasonal nonstationary features could be embedded in the series. For a WV time series of period k, (k =12 for

monthly data, and k = 36 for quarterly data), stationarity is often achieved by calculating the seasonally differenced series $Y_t - Y_{t-k}$. The seasonal difference operator of period k is denoted by $\nabla_k = 1 - B^k$ and therefore a seasonal ARIMA model could be defined by an equation of the form;

$$\phi(B)\nabla^d \nabla_k^D Y_t = U + \theta(B)a_t \qquad (112)$$

The model in Equation (112) takes into account seasonal and regular differencing due to the presence of trend components in the data.

In ARMA modelling applied in this research, the data series ought to be stabilised by using various transformations such as square root or the logarithmic transformations. A practical tool for the choice of the appropriate transformation which is based on the power transformation is the mean-range plot where the range of data is plotted against the mean of each seasonal period. For a detailed discussion on the mean-range plot, refer to Helfenstein, (1986). In the current analysis, the Box-Cox transformation was used to transform the data to a stationary time series. The Box-Cox transformation can be taken as a general time deformation process applied to WV series. This type of transformation is only in the time domain and therefore the notion of stationarity is restricted to a linear transformation. The resultant data series is then subjected to the two-sided Lilliefors and the Jarque-Bera goodness-of-fit test of composite normality which performs the normality test based on the hypothesis that the data in the WV comes from an unspecified normal distribution.

The Lilliefors test evaluates the hypothesis that the WV observations have a normal distribution with unspecified mean and variance, against the alternative that the WV observations do not have a normal distribution. This test compares the empirical distribution of WV with a normal distribution having the same mean and variance as WV. The test is similar to the Kolmogorov-Smirnov test, but since the parameters of the normal distribution are estimated from WV rather than specified *a priori*, the test becomes more data adaptive. From the test, if the result *H=1*, then the hypothesis that WV observations have a normal distribution is rejected. However, if *H =0*, then it implies that null hypothesis cannot be rejected. Additionally, the P-value is computed by interpolation into the Lilliefors simulation table. Table 4.2 illustrates that at 5% significance level, the result of the test is *H=0*. This indicates that the null hypothesis (i.e. the data are normally distributed) cannot be rejected at 7.2 % significance level. In addition, the Lilliefors test statistic of 0.0713 is smaller than the

cut-off value of 0.0745 at 5% significance level, therefore the hypothesis of normality cannot also be rejected.

Table 4.2. Test statistics of the Box-Cox transformed WV normality tests

Test type	**H**	**P-value**	**Statistic**	**Critical value**
Lilliefors	0.0	0.072	0.0713	0.0745
Jarque-Bera	0.0	0.055	5.3233	5.5782

Additionally, the Jarque-Bera test evaluates the hypothesis that WV observations have a normal distribution with unspecified mean and variance, against the alternative that WV does not have a normal distribution. The test is based on the sample skewness and kurtosis of WV. For a true normal distribution, the sample skewness should be near 0 and the sample kurtosis should be near 3. The Jarque-Bera test determines whether the sample skewness and kurtosis are unusually different than their expected values, as measured by a chi-square statistic. The Jarque-Bera test for normal distribution in WV results yielded similar result of $H=0$ obtained from the Lilliefors test and therefore the normality hypothesis at the 5% significant level and 5.5% P-value could not be rejected. The derived time series models could be used to analyze the power spectral density and covariance function of the stochastic WV observations. These models could be suitable for characterizing the spectral density of the random WV observations with known model type and order. However, it should be noted that the spectral characteristics of the WV observations are often unknown *a priori* and therefore a large number of candidate models ought to be computed.

An automatic algorithm that estimates a suitable ARMA (p, q) model to the monthly averaged stationary tropospheric WV revealed that the 1999 to 2008 monthly tropospheric WV data could be modelled by AR (100), MA (20) and ARMA (10, 9). It can be seen that MA order of 20 is much lower than the selected AR order of 100 partly due to the fact that a high AR order model is often used as an intermediate parameter for the estimation of MA models. The optimal coefficient vector of the AR model consists of the set, {ϕ: 1.00, -0.72, -0.11, 0.32, -0.05, 0.11, 0.03, 0.13, 0.02, and -0.25} and the optimal coefficients of the MA model were found to be the vector set, {θ: 1.00, -0.25, -0.15, 0.12, -0.28, 0.27, 0.33}. The prediction error of the three best-selected models is estimated based on the measured and

given values of the residual in variance described by Broersen (2002). For MA and ARMA models, the prediction error is given by Equation (113);

$$e_p(m) = \{\delta(m)\}\frac{1+\frac{m}{N}}{1-\frac{m}{N}}. \qquad (113)$$

Here, m is the number of estimated parameters in the model and δ is the residual of the variance. For AR (*p*) models, the prediction error is given by Equation (114);

$$e_p(p) = \{\delta(p)\}\prod_{m=1}^{p}\frac{1+\frac{1}{N+1-m}}{1-\frac{1}{N+1-m}} \qquad (114)$$

Equation (114) is significantly different from Equation (113) for m > *0*. A single time series model, with selected model order and type, with the smallest estimate of e_p could therefore be easily selected. Using the estimated parameters of the selected model, the spectral density and other statistically significant details such as the second-order characteristics of WV observations could be inferred. The model error denoted as ME is defined as the measure of the accuracy of the estimated model which is the difference between the estimated model and a true stationary process. This measure is simply a scaled transformation of the one step ahead squared error of e_p. Based on the vector set $\{\phi, \theta\}$ of the tropospheric WV, a true stochastic stationary process was modelled and the difference between the stationary process and the estimated model from the WV data. An ME value of ~ 262 was obtained. However, the difference between the selected ARMA and AR models from the same WV data set was found to be ~ 15.

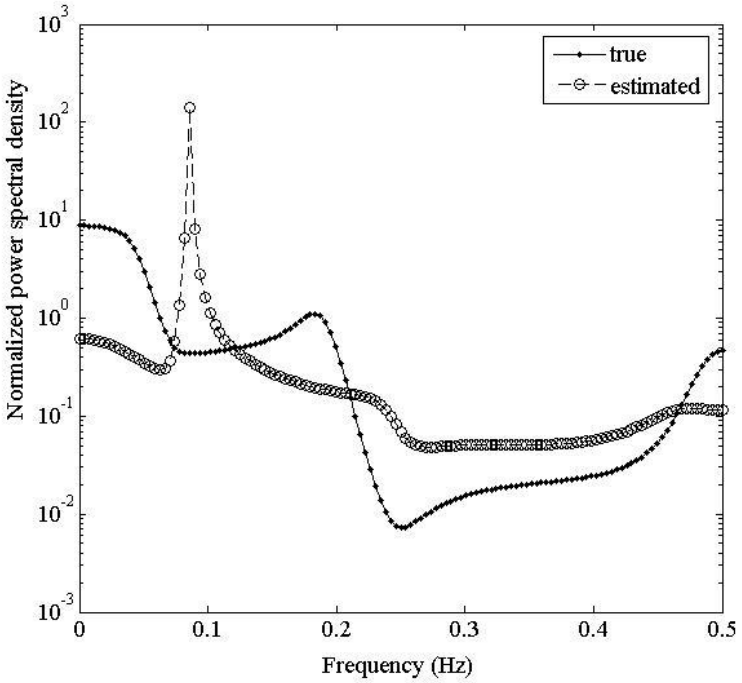

Figure 4.3. True and estimated power spectral density (log scale)

Figure 4.3 compares the power spectral densities of a true stationary process and the spectrum estimated from WV model on the log-scale. It is evident from the figure that the estimated model approximates to the true stationary process with subtle differences which are quantified by the ME. Furthermore, the estimated model accuracy as a function of model order and type is depicted by Figure 4.4. It is clear from Figure 4.4 that AR (50) and ARMA (10, 9) models have higher accuracy than the MA (20) based on the underlying processes in WV data. In all the estimated models, the accuracy increases with increase in the model order up to the $\{p, q\}$ of the ARMA model.

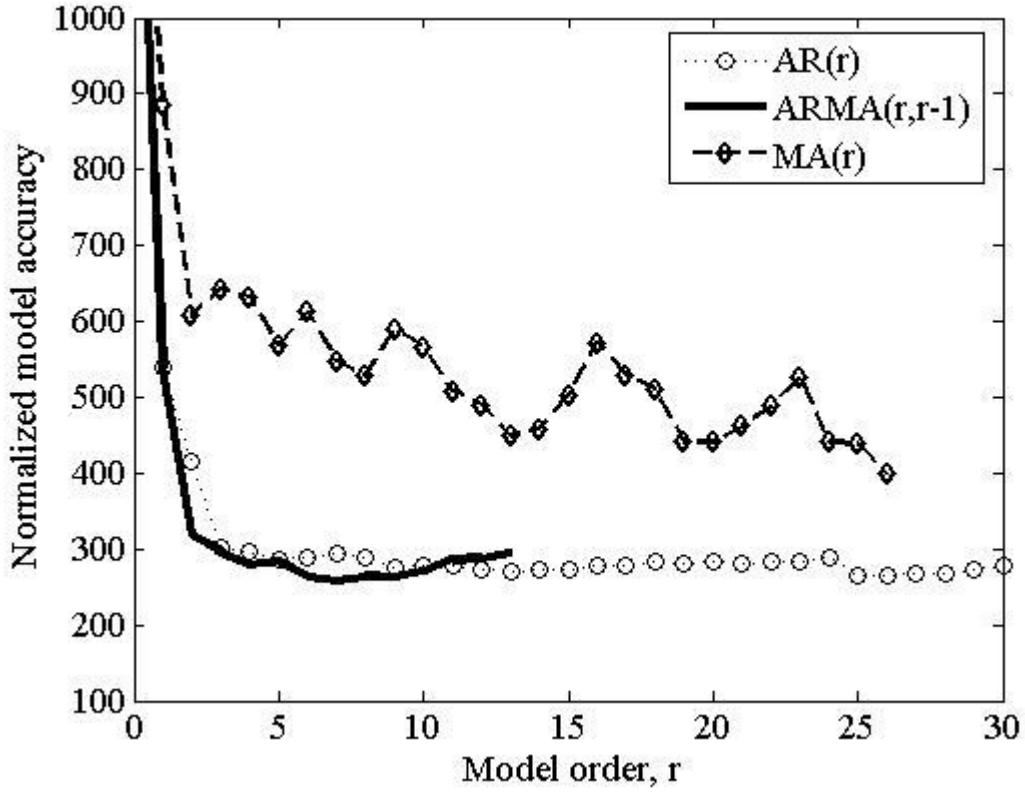

Figure 4.4. Estimated model accuracy as a function of the model type and order.

Assuming that the estimated model is representative of the data, the resultant spectrum could be robust and more accurate compared to the parametric spectrum which could be computed from, for instance the classical periodogram. This is evident from the left panel of Figure 4.5. As can be seen from the left panel of Figure 4.5, the spectral density of the periodogram do not have a smooth curve; this could be attributed to the distortion of the spectrum by convolution of the window function whose width equal the length of the WV. In addition, the periodic oscillations of WV are often treated differently in the discrete Fourier transform. For instance, if there are narrow spectral features (e.g. high frequency components in WV fluctuations), those narrow components will be broadened by the convolution. On the other side, there will be less broadening of the spectral peaks for a broad window function in the time domain but whose spectral main lobe is narrow in the frequency domain.

As depicted in Figure 4.5, the power spectral density based on the periodogram vary significantly from the true power spectral density of the WV series. A common approach (also called the Bartlett method) used to correct this inconsistency is to divide the data record into small subsets and compute the periodogram separately. Thereafter, the results are averaged over all the small records. Averaging periodograms reduces the variance in the

estimated power spectral density and therefore provides a better estimate of the spectral properties of the WV observations. Furthermore, the spectrum of the true stationary process (in the left panel) compares well with the spectrum of the true and the estimated time series (this is plotted in the right panel).

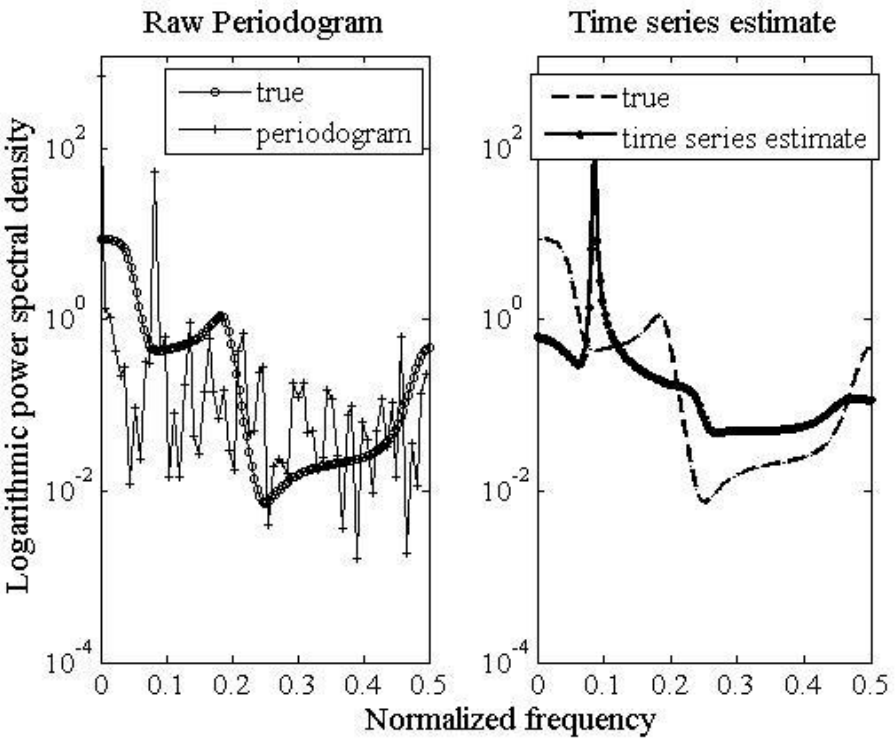

Figure 4.5. Periodogram and the spectra of the true and estimated time series.

4.7. Concluding remarks

Recent advances in atmospheric remote sensing have availed WV data from a variety of sources and sensors with improved spatial-temporal resolution. As a result, data sets that could be used to compute WV for investigating the structure and dynamics in the troposphere have increased. Further, the long time series of WV allows detailed studies on WV acting as a major component of the global hydrological cycle, as a greenhouse gas as well as the variability of WV at different spatio-temporal scales in the climate system. In this chapter, the stochastic behaviour of stationary WV fluctuations has been characterised using a general auto-regressive moving-average (ARMA) time series model. In the analysis, an automatic algorithm which estimates the appropriate model parameters has been used to formulate a model that is used to investigate the nature of the underlying processes that drives the

variability of tropospheric WV. In the present analysis, monthly averaged stationary tropospheric WV derived from geodetic VLBI measurements for the period from 1999 to 2008 is modelled by AR (100), MA (20) and ARMA (10, 9). The power spectral densities of a true stationary process and the spectrum estimated from WV model were compared on the log-scale. Results showed that the estimated model approximates to the true stationary process. Furthermore, the estimated model accuracy as a function of model order and type showed that AR (50) and ARMA (10, 9) models have higher accuracy than the MA (20) based on the underlying processes in WV data.