Chapter 3 Data, Methodology and the Theoretical Approach in the Analysis of Time Series

3.1 Data

The meteorological data used in this study are obtained from two different sources. The first set of data, the NCEP/NCAR Reanalysis Monthly Mean data, is obtained from the National Centre of Environment Prediction (NCEP) in the United States of America. The other set of data, the Upper Air Observation (UAO) data, is obtained from the Malaysian Meteorological Service in Malaysia.

3.1.1 The NCEP/NCAR Reanalysis Monthly Mean Data

The NCEP/NCAR Reanalysis Monthly Mean data contains at least 41 years of complete monthly mean global data stored in a 2.5°×2.5° gridded format from the year 1958 to 1998. The length of the data is considered appropriate for the study of QBO and TBO as both oscillations have typical periods of 22-29 months. The meteorological parameters being used are the 10 levels of upper air zonal wind, the 500-hPa geopotential height and the SST. The 10 levels of upper air are at 925, 850, 700, 500, 300, 250, 200, 100, 50 and 20 hPa. Each of these data fields is either an analyzed field or an average/accumulation of forecast produced by a model. The accuracy of these data depends very much on the observational data coverage, and thus these data fields may not resemble the actual atmosphere accurately. Only data at grid points within a rectangular box covering Malaysia and its adjacent seas are considered. This box covers an area bounded by 100°E-120°E longitudinally and 0°-7.5°N latitudinally. Thus a total of 36 grid point values are being used to calculate the area-average of each data type, except SST, to represent the Malaysian region as shown in Figure 3.1. Only 21 out of the 36 grid points are located at sea, thus the area-average of SST is computed by using these 21 grid point values. These monthly area-averages are computed by using the Grid Analysis and Display System (GrADS) software.

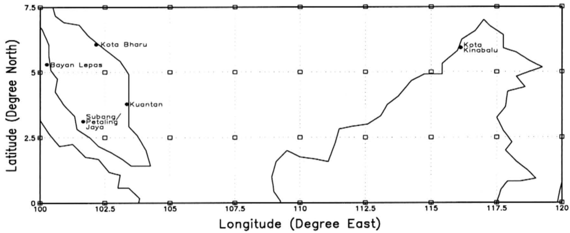

Figure 3.1 Map showing the 36 grid points (open squares) of the NCEP/NCAR Reanalysis Monthly Mean data where meteorological parameters are being extracted to compute the monthly mean area-averages for the Malaysian region. Close circles show the five locations of the Upper Air Observation stations where tropopause heights are extracted.

3.1.2 The Upper Air Observation (UAO) Data

The UAO data consists of daily upper air data obtained from soundings at eight locations in Malaysia twice a day at 0000 and 1200 UTC. Only one data field, the tropopause height, is being extracted from this data. The data series consisting of monthly mean tropopause height are obtained by averaging the daily values extracted at 0000 UTC from four stations in Peninsular Malaysia and one station in East Malaysia. The locations of these five stations are shown in Figure 3.1. The remaining three stations have data series that are either too short or incomplete. The lengths of these five data series are different as each station started its operation at different time. Table 3.1 shows the starting dates and the corresponding lengths of these data series of the five UAO stations. All UAO data are only available up to the year 1996.

Station name	Start date (month, year)	Length of data series
Bayan Lepas	January, 1967	30 years
Subang / Petaling Jaya	July, 1971	25 years 6 months
Kota Bharu	January, 1973	24 years
Kuantan	September, 1971	25 years 4 months
Kota Kinabalu	July, 1968	28 years 6 months

Table 3.1 Table showing the start dates and the lengths of the monthly mean tropopause height data series of the five UAO stations in Malaysia.

3.2 Data Filtering

Most, if not all, meteorological data have an obvious annual cycle, which is of no interest in this study. Another irrelevant feature is the long-term trend exhibited by most meteorological parameters. Thus all data series are filtered to remove both the annual cycle and the long-term trend before any analysis is carried out.

Power spectral analyses are first carried out with annual cycles and long-term trends being removed. In order to enhance the QBO and TBO features, another moving-average filter is used to filter all data series before further power spectral analyses are carried out.

3.2.1 Annual Cycle Filtering

Since most data series contain substantial trends, it is inadequate to just simply calculate the average of each month and compare it with the overall average figure, either as a difference or as a ratio, to eliminate the seasonal fluctuations (Chatfield, 1975). A common and effective way to eliminate the annual cycle is to calculate the moving average

$$\bar{x}_t = \frac{\tfrac{1}{2}x_{t-6} + x_{t-5} + x_{t-4} + x_{t-3} + x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2} + x_{t+3} + x_{t+4} + x_{t+5} + \tfrac{1}{2}x_{t+6}}{12}. \quad (1)$$

Note that the sum of the weights is 1. That is,

$$\sum_{i=t-6}^{t+6} w_i = 1. \quad (2)$$

A simple moving average cannot be used as this would span 12 months and would not be centred on an integer value of t. A simple moving average over 13 months cannot be used as this would give twice as much weight to the month appearing at both ends.

3.2.2 Long-Term Trend Filtering

A special type of filtering, which is particularly useful for removing a trend, is simply to difference a given time series until it becomes stationary (Chatfield, 1975). For non-seasonal data, first-order difference is usually sufficient to attain apparent stationary, so that the new series $\{y_1, \ldots, y_{n-1}\}$ is formed from the original series $\{x_1, \ldots, x_n\}$ by

$$y_t = x_{t+1} - x_t. \quad (3)$$

3.2.3 The Moving-Average Filter

When applied to a time series, a moving average of any sort will have the effect of altering the amplitude and often the phase as well, of the variations in the series, but to an extent that differs for each wavelength component of variation (WMO, 1971). In this way, the moving average serves to change the spectrum of the original series, just as a coloured filter changes the optical spectrum of a light ray passing through it. Hence, a moving average can be regarded as a mathematical filter, and changing the manner in which the successive values of the series are weighted to obtain the average can alter the frequency response of the filter.

Consider symmetrical moving averages, that is, to filters in which the weighting of successive terms of a series varies symmetrically both backward and forwards from a central weight. Symmetry of this kind is necessary, but not always sufficient; to preclude shifts of phase in a series when operated on by the filter. Mathematically we may express such a filter by the equation

$$\bar{y}_t = \sum_{i=-n}^{n} w_i \, y_{t+i} \tag{4}$$

where w_i is the weight by which the value of the series i units removed from t is multiplied. The length of this filter is $\mathbf{2}n+\mathbf{1}$ time units.

The filter defined by equation (4) has a response function that depends on frequency. If frequency is arbitrarily expressed in cycles per interval between successive observations in the time series, the response function can be written as

$$R(f) = w_o + 2\sum_{k=1}^{n} w_k \cos 2\pi fk \,. \tag{5}$$

The response function measures the amplitude of variation in the time series after filtering relative to that before filtering, for any given frequency of variation, f.

Inasmuch as in the study of climatic fluctuations we are interested only in a particular range of wavelengths of time series. The response function of a moving-average filter has one desirable characteristic; namely, it passes the range of certain required wavelengths with the least diminution of amplitude and filters out a large proportion of the variation at other wavelengths. The fact that the responses of some moving average filters become appreciably negative in certain ranges of wavelength can introduce serious difficulties of interpretation when the results of using these particular filters are examined.

It is relatively easy to modify the weights in a moving average to prevent the response from ever becoming negative. Admittedly the use of a moving average in which the weights are unequal greatly increases the labour involved in computing it. However, the improvement in the result is often well worth the added labour.

3.2.3.1 The Binomial Low-Pass Filter

A simple moving-average filter that has greatly improved response characteristics can be achieved by setting the weights in equation (4) proportional to the ordinates of a Gaussian probability curve, or what amounts to essentially the same thing, the weights may be set proportional to the binomial coefficients

$$c_k = \frac{\eta!}{k!(\eta - k)!} \tag{6}$$

which for large η closely approximate to the Gaussian curve ordinates. The appropriate factor of proportionality is easily determined by the condition that in equation (4) the weights should all add up to one. That is,

$$\sum_{i=-n}^{n} w_i = 1 . \tag{7}$$

The choice of η depends on individual application. A reasonably good choice of η in an actual application can be determined by setting the cut-off wavelength equal to about 4-6 standard deviations of the Gaussian curve or the binomial distribution involved. The standard deviation, σ, of the binomial distribution is given by

$$\sigma = \frac{\sqrt{\eta}}{2}. \tag{8}$$

This type of weighted moving average has a response function that is equal to unity at infinite wavelengths, and that trails off asymptotically to zero with decreasing wavelength. In particular, for this binomial weighted moving average, the response function is approximately

$$R(f) = \cos^{\eta} \pi f. \tag{9}$$

A moving average with this sort of response is commonly known as a low-pass filter.

3.2.3.2 The Band-Pass Filter

It is often useful to design a moving average that will filter out not only the shorter wavelengths in a series but also the longer wavelengths as well. More particularly, one may wish to study variations in a series that lie in a rather narrow range of wavelengths only and exclude all other wavelengths both longer and shorter. It is not difficult to design a filter of this kind, by calculating an appropriate set of weights that will tune the filtered output to the range of wavelengths desired. Such a filter is known as a band-pass filter.

One might suppose that the ideal band-pass filter is one in which the frequency response is unity for a narrow band of wavelengths, and drops abruptly to zero on both sides of this band. However, it is difficult to design such a sharp band-pass filter theoretically. In order to do so, a system of weights that fluctuated rapidly in sign and magnitude from one term to the next is needed, and a very large total number of weights are required as well. The application of such a filter to short time series would leave many original values near both ends of the series unrepresented by filtered values, and the labour of computing the filtered values would be very great.

It is easier to design a band-pass filter in which the frequency response decreases rather gradually on both sides of the central frequency of special interest, but preserves the phase and amplitude of the latter with excellent fidelity. A band-pass filter having this property is typified by a relatively small total number of weights, and by weight values that vary only slowly from one term to the next. Such a filter may be constructed by the "differenced low-pass filter" method.

Suppose that a time series y_t is subjected to two low-pass filtering by either the Gaussian ordinate or binomial coefficient method. By defining $\bar{y}_t$ as the series smoothed by the short- period filter and $\hat{y}_t$ as the series smoothed by the long-period filter, there will be a range of wavelengths for which the frequency response $R(f)$ of the short-period filter is large but for which the frequency response $R'(f)$ of the long-period filter is small. If, then, by subtracting $\hat{y}_t$ from $\bar{y}_t$ a new series will be generated in which only those wavelengths of variation in the original series y_t that lie between a certain frequency range will be emphasised.

Subtracting a series after being smoothed by one low-pass filter from the same series after being smoothed by another low-pass filter, the result is the same as if a band-pass filter has filtered the original series only once. The frequency response $R''(f)$ of the latter filter can in fact be represented by the difference of $R(f)$ and $R'(f)$. That is,

$$R''(f) = R(f) - R'(f), \tag{10}$$

and merely differencing the corresponding weights of the two low-pass filters can derive the band-pass filter itself. That is to say if

$$\hat{y}_t = \sum_{i=-n'}^{n'} w_i' \, y_{t+i} \tag{11}$$

and, by applying equation (9),

$$R'(f) = \cos^{\eta'} \pi f \,, \tag{12}$$

then, the frequency response of the band-pass filter becomes

$$R''(f) = \cos^{\eta} \pi f - \cos^{\eta'} \pi f \tag{13}$$

and the band-pass filter is simply subtracting equation (11) from equation (4). That is,

$$\tilde{y}_t = \bar{y}_t - \hat{y}_t = \sum_{i=-n'}^{n'} (w_i - w_i') y_{t+i} \,. \tag{14}$$

However, renormalization to satisfy equation (7) may be required.

In order to emphasise the range of frequency that typically represents the QBO and TBO, a binomial band-pass filter with frequency response function as shown in Figure 3.2 is used.

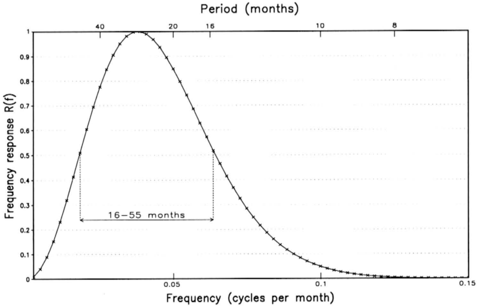

Figure 3.2 Frequency response function of the binomial band-pass filter used to emphasise periods that typically represent the QBO and TBO.

3.3 Methods of Analysis

3.3.1 Autocorrelation

As all meteorological data series used here are continuous series, persistence is typically characterized in terms of their serial correlations, or temporal autocorrelations. Autocorrelation means correlation of a variable with itself, so that temporal autocorrelation indicates the correlation of a variable with its own future and past values. Here, autocorrelations were computed as Pearson product-moment correlation coefficients.

The Pearson product-moment coefficient of linear correlation between two series x and y is the ratio of the sample covariance of the two series to the product of the two standard deviations (Wilks, 1995). That is,

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{s_x s_y}$$

$$= \frac{\dfrac{1}{n-1}\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]}{\left[\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-x)^2\right]^{1/2}\left[\dfrac{1}{n-1}\sum_{i=1}^{n}(y_i-y)^2\right]^{1/2}} \tag{15}$$

where n is the number of terms in the x as well as y series.

After some algebraic manipulation of the summations in the correlation coefficient, the computational form for the correlation coefficient becomes

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \dfrac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\left[\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]^{1/2}\left[\sum_{i=1}^{n} y_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]^{1/2}}. \tag{16}$$

The process of computing autocorrelation can be visualized by imagining two copies of the same series x being written, with one copy shifted by one unit of time to form $n-1$ pairs of lagged data. Autocorrelations are then computed by substituting these lagged data pairs into equation (15). Hence, the lag-1 autocorrelation is expressed by

$$r_1 = \frac{\sum_{i=1}^{n-1}[(x_i - \bar{x}_-)(x_{i+1} - \bar{x}_+)]}{\left[\sum_{i=1}^{n-1}(x_i - \bar{x}_-)^2 \sum_{i=2}^{n}(x_i - \bar{x}_+)^2\right]^{1/2}} \tag{17}$$

where the sample mean of the first $n-1$ values is denoted by the subscript "-" and that of the last $n-1$ values by the subscript "+".

While the lag-1 autocorrelation is the most commonly computed measure of persistence, it is also sometimes of interest to compute autocorrelations at longer lags as the two series are shifted by more than one time unit. Equation (17) can be generalized to the lag-k autocorrelation coefficient using

$$r_k = \frac{\sum_{i=1}^{n-k}[(x_i - \bar{x}_-)(x_{i+k} - \bar{x}_+)]}{\left[\sum_{i=1}^{n-k}(x_i - \bar{x}_-)^2 \sum_{i=k+1}^{n}(x_i - \bar{x}_+)^2\right]^{1/2}}. \tag{18}$$

Here the subscripts "-" and "+" now indicate sample means over the first and last $n-k$ data values respectively. As a time series is shifted increasingly relative to itself, there are progressively less overlapping data pairs to work with. As much data is lost at large lags, usually only the lowest few values of k will be of interest.

If the data series is sufficiently long, the overall sample mean will be very close to the subset averages of the first and last $n-k$ values. The overall sample standard deviation will also be close to the standard deviations of the two subset for the first and last $n-k$ values. Invoking these assumptions leads to the computational approximation

$$r_k \approx \frac{\sum_{i=1}^{n-k}[(x_i - \bar{x})(x_{i+k} - \bar{x})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n-k}(x_i x_{i+k}) - (n-k)\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{19}$$

Together, the collection of autocorrelations computed for various lags is called an autocorrelation function. Often autocorrelation functions are displayed graphically, with the autocorrelations plotted as a function of lag.

The statistical significance of the lag-one autocorrelation r_1 for the null hypothesis of randomness is tested. Before doing this test, the sample frequency distribution of the values of the series is examined. Since the frequency distributions of the meteorological data series approximate to the Gaussian distributions, and it is generally sufficient to base the significance test on a desired probability point of the Gaussian distribution when the data length n is greater than 40, the exact one-tail significance point test (WMO, 1971)

$$(r_1)_t = \frac{-1 \pm t_g\sqrt{n-2}}{n-1} \tag{20}$$

can be used. Here t_g is the value of the standard deviate in the Gaussian distribution corresponding to the desired significance point of r_1.

In the present study, the minimum data length used is the tropopause height data series from Kota Bharu station. This data length, after the annual cycle and trend have been removed, and the series further filtered by the band pass filter, is 236. To test the sample r_1 against the null hypothesis that r_1 is no larger than the value appropriates to randomness, the one-tailed 95% probability point of t_g is chosen. From an extracted table (Hald, 1952) of probability points of Gaussian normal distribution in Appendix B, the value of t_g is 1.645. By putting this value of t_g into equation (20) and using the plus sign in the numerator,

$$(r_1)_t = \frac{-1 + 1.645\sqrt{236-2}}{236-1} = 0.1028.$$

Therefore, if and only if the sample value of r_1 is larger than 0.1028, the series of the tropopause height from Kota Bharu station can be concluded to be non-random at the 95% significance level.

It is noted, from equation (20), that the value of $(r_1)_t$ is inversely proportional to the square root of n. Hence, for all other meteorological data series, which have data length longer than 236, the value of $(r_1)_t$ will be smaller than 0.1028.

In general, the one-tailed significance test used here is appropriate because most alternatives to randomness with which meteorological series are concerned would be expected to increase the value of r_1.

3.3.2 Pearson Cross Correlation

Analogous to autocorrelation, the general lag-k Pearson cross correlation coefficient between two data series x and y, from equation (18), is

$$(r_{xy})_k = \frac{\sum_{i=1}^{n-k}[(x_i - \bar{x}_-)(y_{i+k} - \bar{y}_+)]}{\left[\sum_{i=1}^{n-k}(x_i - \bar{x}_-)^2 \sum_{i=k+1}^{n}(y_i - \bar{y}_+)^2\right]^{½}}. \tag{21}$$

Here, the lagged data pairs are formed by shifting k units time the series y with respect to the series x. Hence, the subscripts "-" and "+" now indicate sample means over the first and last $n-k$ data values of the data series x and y respectively.

Similar to equation (19), for sufficiently long data series of x and y, the computational approximation becomes

$$(r_{xy})_k \approx \frac{\sum_{i=1}^{n-k}[(x_i-\bar{x})(y_{i+k}-\bar{y})]}{\left[\sum_{i=1}^{n-k}(x_i-\bar{x})^2\sum_{i=k+1}^{n}(y_i-\bar{y})^2\right]^{1/2}}. \tag{22}$$

As before, the graphical display of a collection of Pearson cross correlations computed for various lags is called the Pearson cross correlation function.

Lastly, the statistical significance of $(r_{xy})_1$ for the null hypothesis of randomness can be tested by using the exact one-tail significance point test similar to the one appropriate for testing the randomness of r_1.

3.3.3 Power Spectrum Analysis

3.3.3.1 Harmonic Analysis

In general, a data series x consisting of n points can be represented exactly, meaning that a harmonic function can be found that passes through each of the points, by adding together a series of $n/2$ harmonic functions (Wilks, 1995),

$$\begin{aligned} x_t &= \bar{x} + \sum_{k=1}^{n/2}\left[C_k\cos\left(\frac{2\pi kt}{n}-\phi_k\right)\right] \\ &= \bar{x} + \sum_{k=1}^{n/2}\left[A_k\cos\left(\frac{2\pi kt}{n}\right)+B_k\sin\left(\frac{2\pi kt}{n}\right)\right] \end{aligned} \tag{23}$$

where the $n/2$ terms in the summation are harmonics with frequencies

$$\omega_k = \frac{2\pi k}{n}. \tag{24}$$

When $k=1$, the first harmonic has a fundamental frequency $\omega_1 = 2\pi/n$.

In general, the coefficients in the second line of equation (23), A_k and B_k, corresponding to particular data series x_t can be found using multiple regression methods, but if the series is equally spaced in time and contains no missing values, these coefficients can be obtained more easily using

$$A_k = \frac{2}{n}\sum_{t=1}^{n} x_t \cos\left(\frac{2\pi kt}{n}\right) \tag{25a}$$

and

$$B_k = \frac{2}{n}\sum_{t=1}^{n} x_t \sin\left(\frac{2\pi kt}{n}\right). \tag{25b}$$

Having computed these coefficients, the amplitude, C_k, in the first line of equation (23), can be obtained, separately for each harmonic, by computing

$$C_k = \left(A_k^2 + B_k^2\right)^{1/2}. \tag{26}$$

Notice that equations (25) do not depend on any harmonic other than the one whose coefficients are being computed. This fact implies that the coefficients A_k and B_k for any particular harmonic can be computed independently of those for any other harmonic. It is a remarkable property of the harmonic functions that they are uncorrelated so that the amplitude and phase for the first or second harmonic are the same regardless of whether they will be used in an equation with the third, fourth, or any other harmonics. This remarkable attribute of the harmonic functions is a consequence of what is called the orthogonality property of the sine and cosine functions.

Since the relationships between harmonic predictor variables and the data series x_t do not depend on what other harmonic functions are also being used to represent the series, the proportion of the variance of x_t accounted for by each harmonic is also fixed. When this proportion is expressed as the R^2 statistic commonly computed in regression, the R^2 for the k^{th} harmonic is simply

$$R_k^2 = \frac{n/2\, C_k^2}{(n-1)\, s_x^2} \tag{27}$$

where s_x^2 is simply the sample variance of the data series. Notice that the strength of the relationship between the k^{th} harmonic and the data series is expressed entirely by the amplitude C_k. The phase angle ϕ_k is necessary only to determine the positioning of the cosine curve in time. Furthermore, since each harmonic provides independent information about the data series, the joint R^2 exhibited by a regression equation with only harmonic predictors is simply the sum of the R_k^2 values for each of the harmonics. If all the $n/2$ possible harmonics are used as predictors, then the total R^2 will be exactly 1. That is,

$$R^2 = \sum_{k=1}^{n/2} R_k^2 = 1 . \tag{28}$$

3.3.3.2 Spectral Analysis

Equation (23) says that a data series x_t of length n can be completely specified in terms of the n parameters of $n/2$ harmonic functions. Equivalently, it can be said that the data series x_t is transformed into new set of quantities A_k and B_k according to equations (25). For this reason, equations (25) are called the discrete Fourier transform. According to equation (28), this data transformation accounts for all the variation in the series x_t.

The foregoing suggests that a different way to look at a time series is as a collection of Fourier coefficients A_k and B_k that are a function of frequency ω_k, rather than as a collection of data point x_t measured as a function of time.

The characteristics of a time series that has been Fourier-transformed into the frequency domain are most often examined graphically, using a plot known as the power spectrum, or the periodogram. In simplest form, this plot of a spectrum consists of the squared amplitudes C_k^2 as a function of the frequencies ω_k. Equation (27) can be used to rescale the vertical axis numerically so that the plotted points become proportional to the squared amplitudes. As a result, the power spectrum becomes a plot of normalized spectral density against frequency. The horizontal axis of the power spectrum consists of $n/2$ frequencies ω_k if n is even, and $(n-1)/2$ frequencies if n is odd. The smallest of these is the fundamental frequency $\omega_1 = 2\pi/n$, which corresponds to the cosine wave that executes a single cycle over the n time points. The highest frequency $\omega_{n/2} = \pi$, called the Nyquist frequency, corresponds to the cosine wave that executes a full cycle over only two time intervals, and that executes $n/2$ cycles over the full data record. The Nyquist frequency depends on the time resolution of the original data series x_t, and imposes an important limitation on the information available from a spectral analysis.

Instead of using the angular frequency, a common alternative for the horizontal axis is to use the frequencies

$$f_k = \frac{k}{n} = \frac{\omega_k}{2\pi}. \tag{29}$$

Under this alternative convention, the allowable frequencies range from $f_1 = 1/n$ for the fundamental to $f_{n/2} = 1/2$ for the Nyquist frequency. The horizontal axis can also be scaled according to the reciprocal of the frequencies, or the periods

$$\tau_k = \frac{n}{k} = \frac{2\pi}{\omega_k} = \frac{1}{f_k}. \tag{30}$$

3.3.3.3 Smoothing the Power Spectrum

The spectral density $\hat{s}_k$ of a power spectrum can be smoothed by a number of different methods. In the "Hanning" method (WMO 1971), the smoothing formulae are

$$
s_k = \begin{cases} \frac{1}{2}(\hat{s}_k + \hat{s}_{k+1}), & k = 1 \\ \frac{1}{4}(\hat{s}_{k-1} + 2\hat{s}_k + \hat{s}_{k+1}), & k = 2,3,\ldots,m-1 \\ \frac{1}{2}(\hat{s}_{k-1} + \hat{s}_k), & k = m \end{cases} \tag{31}
$$

where $m \le n/2$, is the maximum number of lags.

Note that for any smoothing function applied to sample spectra, the increased smoothness and representativeness of the resulting spectra come at the expense of decreased frequency resolution. Essentially, stability of the sampling distributions of the spectra estimates is obtained by smearing spectral information from a range of frequencies across a frequency band. Hence smoothing across broader bands produces less erratic estimates, but hides sharp contributions made at particular frequencies. In practice, some compromise between stability and resolution is needed.

3.3.3.4 Tests of Statistical Significance applied to Power Spectra

To test for statistical significance of power spectra, a "null"-hypothesis continuum has to be fitted to the spectrum.

In general, if the lag-one serial correlation coefficient r_1 of the series does not differ from zero by a statistically significant amount, the series should be regarded as free from persistence. In this case, the appropriate "null" continuum is that of "white noise", or in other words, a horizontal straight line whose value is everywhere equal to the average of the values of all the raw spectral estimates in the computed spectrum.

On the other hand, if the lag-one serial correlation coefficient r_1 of the series differs from zero by a statistically significant amount, the lag-two and lag-three serial correlation coefficients, r_2 and r_3 respectively, have to be checked. If both r_2 and r_3 approximate to the exponential relation

$$r_k \approx r_1^{\,k}, \qquad k = 2,3,\ldots,\text{etc}, \tag{32}$$

then the appropriate "null" continuum should be assumed that of Markov "red noise" whose shape depends on the unknown value of the lag-one correlation coefficient for the population ρ. If the sample lag-one coefficient r_1 is an unbiased estimate of ρ, the following equation

$$S_k = \bar{s}\frac{1 - r_1^2}{1 + r_1^2 - 2r_1\cos(\pi k/m)} \tag{33}$$

for various choices of harmonic number k between $k = 1$ and $k = n/2$ can be evaluated. In equation (33) $\bar{s}$ is the average of all the raw spectral estimates $\hat{s}_k$ in the computed spectrum. The resulting values of S_k may then be plotted superposed on the sample spectrum, and a smooth curve passed through these values to arrive at the required "null" continuum. Notice that in the presence of trend or a slippage of the mean r_1 may overestimate the value of ρ. Under these circumstances, a better estimate of ρ may be obtained as the ratio r_2/r_1.

Finally, if it is found that the lag-one serial correlation coefficient r_1 of the series differs significantly from zero, but that the coefficients for higher lags do not bear an exponential relationship to r_1, then it is doubtful that simple Markov persistence is the dominant form of non-randomness in the series. In this event, the Markov "red noise" continuum described in the preceding paragraph may not be appropriate. Nevertheless, it is suggested that the Markov "red noise" continuum be computed as before. If this continuum is in fact erroneous, the discrepancy will be made apparent during the next step of the analysis of statistical significance and a more appropriate form of "null" continuum may suggest itself at that stage.

Once an initial choice of "null" continuum has been made, and plotted superposed on the spectrum, the consistency of the spectrum with the continuum can be evaluated by comparing the value of each spectral estimate $\hat{s}_k$ with the local value of the "null" continuum. If none of the spectral estimates deviates by a statistically significant amount from the continuum, the continuum does in fact approximate to the true spectrum of the population series. On the other hand, if one or more of the spectral estimates deviate significantly from the continuum, the continuum is not a satisfactory approximation to the true spectrum of the population series. In that case, the shape of the continuum is modified to one that fits satisfactory, or the magnitude of the discrepancy as well as the range of wavelengths involved be established and the form of non-randomness that it indicates to be present in the series be specified.

The statistic associated with each spectral estimate, on the basis of which the statistical significance of the deviation of that estimate from the "null" continuum can be determined, is the ratio of the magnitude of the spectral estimate to the local magnitude of the continuum. This ratio has been found (Tukey, 1950) to be distributed as chi-square divided by degree of freedom. The degree of freedom ν of each estimate of a spectrum that is based on a record length of n values and a maximum lag number of m is given by

$$\nu = \frac{2n - m/2}{m}. \tag{34}$$

Therefore, if the local value of the continuum approximates the true magnitude of the population spectrum, this value corresponds to the 50% point of the χ^2/ν distribution appropriate to the sample. The ratio of any sample spectral estimate $\hat{s}_k$ to its local value of the continuum is then compared with the critical percentage point levels of a χ^2/ν distribution for the proper value of ν, and this comparison establishes the level of statistical significance required.

In a given sample spectrum, the 95% point of the χ^2/ν distribution is the same for all spectral estimates $\hat{s}_k$. In other words, the 95% confidence limit for the "null" continuum is given by a second continuum whose value for any wavelength in the spectrum is equal to a certain fixed multiple (greater than one) of the value of the "null" continuum at that wavelength. Similarly, the 5% confidence limit for the "null" continuum is given by a third continuum the value of which for any wavelength in the spectrum is equal to a certain fixed fraction (less than one) of the value of the "null" continuum at that wavelength. The 95% and 5% points of the χ^2/ν distribution for the appropriate degree of freedom ν can be found by using the extracted table (Hald 1952) of probability points of χ^2/ν distribution in Appendix B.

Once the desired confidence limits of the continuum have been added graphically to the spectrum, it can be seen at a glance if any of the spectral estimates lie outside these limits, either above the 95% limit or below the 5% limit. If none do, then it is justified in concluding that the sample spectrum does in fact derive from a population the spectrum of which approximates to the "null" continuum. Should any spectral estimate be found to exceed the 95% confidence limit of the continuum, the decision as to what this means should properly depend on which of the two following circumstances apply.

(i) The spectral estimate in question corresponds in wavelength to an oscillation, such as the well-known biennial oscillation, suspected on a priori grounds that might be contained in this series. The problem of interpreting the statistical significance of the biennial oscillation in this particular spectrum is straightforward: the biennial oscillation is significant at or above the 95% confidence level.

(ii) The spectral estimate in question corresponds to wavelengths that have not seemed noteworthy in previous studies of other climatological time series and that are unrelated to any known physical processes either in the atmosphere itself or elsewhere in the earth's physical environment that might be expected to influence climate. Under these circumstances, it is required to base the judgement of statistical significance on a more stringent criterion than before.