MODELS FOR OCEAN WAVES

by

PETER BUTTON

THESIS

Submitted in fulfilment of the requirements
for the degree of

MASTER OF SCIENCE

In the department of
Mathematical Statistics
University of Cape Town.

SUPERVISOR: Professor W Zucchini

September 1988.

ACKNOWLEDGEMENTS

I am indebted to the following people who have offered valuable assistance in the preparation of this thesis:

Professor Walter Zucchini, for his enthusiastic supervision and encouragement throughout the preparation of this thesis, my sincere thanks.

Mr. J. Rossouw, Department of Coastal Engineering, University of Stellenbosch and Dr. F. Shillington, Department of Oceanography, University of Cape Town, for valuable comments and criticism.

Mrs. Tib Cousins, for help with TEX related problems.

TABLE OF CONTENTS

CHAPTER 1.

INTRODUCTION.

Ocean waves represent an important design factor in many coastal engineering applications. Although extreme wave height is usually considered the single most important of these factors there are other important aspects that require consideration. These include the probability distribution of wave heights, the seasonal variation and the persistence, or duration, of calm and storm periods.

If one is primarily interested in extreme wave height then it is possible to restrict one's attention to events which are sufficiently separated in time to be effectively independently (and possibly even identically) distributed. However the independence assumption is not tenable for the description of many other aspects of wave height behaviour, such as the persistence of calm periods. For this one has to take account of the serial correlation structure of observed wave heights, the seasonal behaviour of the important statistics, such as mean and standard deviation, and in fact the entire seasonal probability distribution of wave heights. In other words the observations have to be regarded as a time series.

The variable of interest in this study is not wave height itself but a representative wave height statistic, $Hm0$, which is an estimate of what is called the *significant wave height* which is defined as *the average of the one third largest waves during an observation period,* and is consequently sometimes denoted by $H_{\frac{1}{3}}$. Also of interest is the variable, Tz , which is an estimate of the zero up-crossing wave period.

Briefly the quantities $Hm0$ and Tz are obtained as follows (for a more detailed account see Appendix A):

Waves can be defined using the zero up-crossing method where a wave height is defined to be the range of water elevation between successive zero up-crossings. The time between these crossings is then the wave period. This is illustrated in Figure 1.A where wave height and period are depicted. These wave records are compiled using data collected using a wave rider buoy, positioned at various locations off the Southern Cape coast, that registers instantaneous water surface elevation at a fixed time interval Δt . A time interval of 0.5 seconds was used for the data in this study.

The *significant wave height* is then estimated by $Hm0 = 4 * \sqrt{m0}$ where $m0$ is

the zero moment of the spectral density function that is estimated from the wave record data. The spectral density, $S(f)$, is a density of the frequencies of the waves in a wave record. The n^{th} moment is given by

$$m_n = \int_0^\infty f^n S(f) df$$

In practice this integral is evaluated using numerical integration.

The variable Tz is the zero crossing wave period and is calculated by dividing the duration of the wave record by the number of times the record crosses the mean line in an upward direction.

Tz is estimated by

$$Tz = \sqrt{\frac{m0}{m2}}$$

where $m0$ and $m2$ are the zero and second moments of the spectral density respectively.

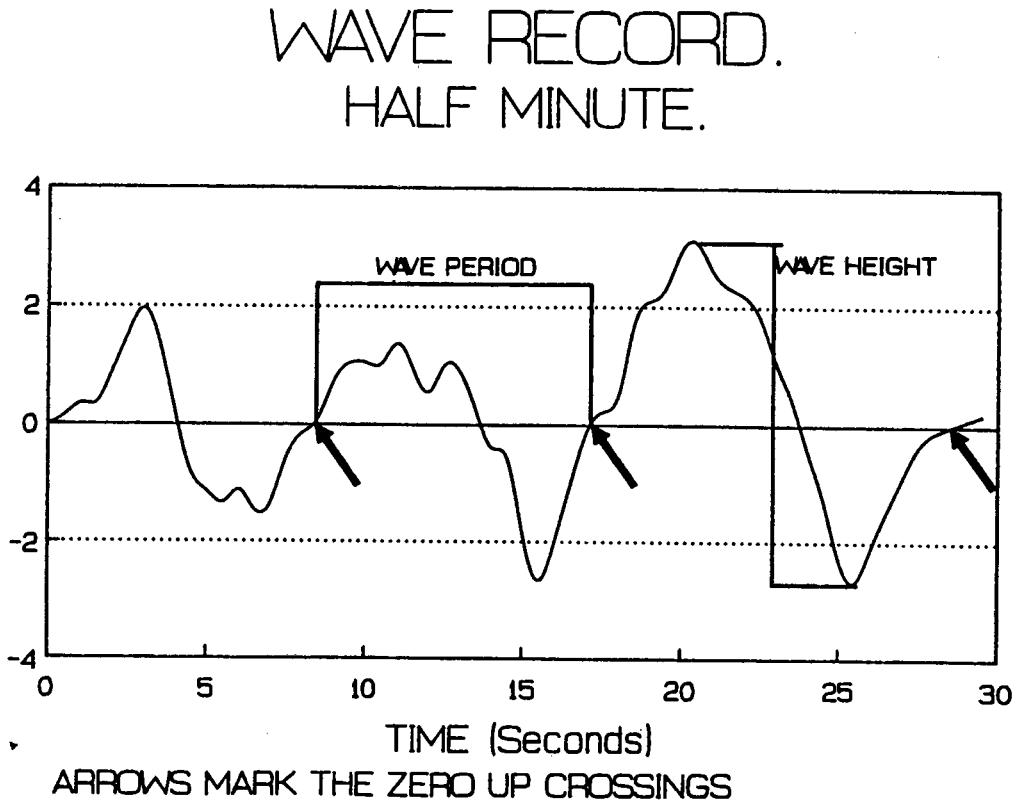


FIGURE 1.A

An aim of this study was to attempt to model the $Hm0$ process (and possibly the bivariate $Hm0, Tz$ process) by means of a time series model. Such models provide a concise description of the patterns that might exist in the $Hm0$ series.

Once we are satisfied that the model is an adequate representation of the real $Hm0$ series it can be used to generate artificial $Hm0$ sequences of arbitrary length which can then be used to calculate various quantities of interest. Although the model contains the information to estimate these quantities it is very difficult to derive analytical solutions. However these quantities can be found by simulation.

For many purposes artificial $Hm0$ sequences are more useful than the original historical records. Firstly they are free of the typical imperfections, such as incorrect recordings and missing observations which sometimes occur in real data sets. Secondly, the historical records available are often quite short and therefore only reflect a limited amount of information about what could occur. It is sometimes argued that, as the parameters of the model have to be estimated from the historical record, the artificial sequences generated by the model are no more than complicated extrapolations of the historical record. However a good stochastic model contains more than the information which can be extracted from a single historical record. It contains information in the form of assumptions about ocean waves which are based on our general knowledge about the behaviour of ocean waves derived from observations at other locations and from theory. For example, it is reasonable to assume that certain average properties of ocean wave variables are periodic and vary smoothly with time. Such assumptions give ocean wave models structure which may not be evident in a single short historical record.

A preliminary analysis confirmed that both $Hm0$ and Tz processes are indeed seasonal and are substantially serially correlated. It was found that both series exhibit relatively low signal-to-noise ratios, that is the deterministic components of the series are quite small compared to the stochastic components. Consequently it is particularly important in this application to model the stochastic component of the series accurately so as to capture the character of the variability in each of the series. The unusual distribution of this component made this the most difficult aspect of the series to model.

Several models for describing the $Hm0$ series were considered, four were inves-

tigated in detail. The seasonal components of the series were modelled using the convenient (and parsimonious) trigonometric functions, that is truncated Fourier series representations of the mean value and standard deviation functions. The autocorrelation structure was described by an Autoregressive process of order 1 (AR(1)). Essentially the four models differ in how the residual component is modelled, that is which probability density function is used to fit the independently distributed residuals.

Having fitted the four models to the observations the relative merits of the fits were assessed. No one model was found to be optimal in terms of all the criteria which are relevant in this application. In selecting a suitable model it is thus necessary to first decide which aspect of the fit is regarded as most important. This applies to both the $Hm0$ and Tz series.

We also attempted to model the bivariate $(Hm0, Tz)$ process as a bivariate time series, but met with only limited success. The strong (statistical) dependence between the two series can be modelled, at least approximately, but it was found difficult to simultaneously preserve all the properties of the observed bivariate process. A complicating factor is that there is a bound on the $Hm0, Tz$ combinations which are physically possible. In statistical terms we are dealing with random variables whose (bivariate) density function is bounded in some imprecisely specified way. Consequently although it is possible to model the two time series individually and even to preserve their average cross correlation, it is extremely difficult to incorporate these bounds into the model. Nevertheless one of the two models considered does provide a useful indication of the behaviour of the bivariate series and may be accurate enough for many practical purposes.

The thesis is organized as follows: Chapter 2 gives a brief review of the literature on models for ocean wave statistics. We discuss the reason why the existing models are not applicable to our observations. Chapter 3 deals with the preliminary statistical analysis of the data and includes the identification of a significant seasonal component and of an autoregressive structure. Chapter 4 deals with a method of density estimation which we required in the following chapters.

In Chapter 5 we give the details of the models that we used in this study. The details include parameter estimation, interpretation of results and overall fit of the

model to the data. Chapter 6 contains the details of the algorithm used to generate $Hm0$ sequences using the various models.

In Chapter 7 we compare various aspects of the models, including the artificial $Hm0$ sequences from each of them. We then chose a model whose overall performance seems to be better than the others and in Chapter 8 we conduct a more rigorous validation of the model. In particular we checked that all the important properties of the original series are preserved in the model. In Chapter 9 we consider various possibly applications of the artificially generated $Hm0$ series.

Chapter 10 deals with our attempt at modelling the bivariate $Hm0, Tz$ process. A brief summary of the study and the main conclusions reached are given in Chapter 11.

CHAPTER 2.

LITERATURE REVIEW

In this chapter a brief review is given of the literature concerned with statistical models for ocean waves. This includes various derived wave height and wave period distributions and also a discussion on the relevance and effect of the spectral width parameter, where this parameter gives an indication of the range of frequencies present in a wave record. For an ocean swell that resembles a simple sine wave the range of frequencies is very small (narrow banded) whereas if the ocean swells are particularly turbulant then the range of frequencies is large (wide banded). The applicability of the available theory to the problem in this study is also discussed. There is also a portion of the literature that deals with the extreme value prediction or estimation.

Extreme value analysis and related upper tail analysis has been used to estimate quantities of the high wave height distribution, such as the fifty year return value, and these methods are commonly employed in practice, see for example Carter et al. (1986). However the ability to model this aspect of ocean wave statistics was not a primary aim of this study. The subject is well covered in the literature.

Although the particular problem of attempting to model the $Hm0$ process, (and joint $Hm0, Tz$ process), as a time series has not been addressed there is a some literature dealing with ocean wave statistics. The following is a brief summary of the current status of wave data analysis.

Sverdrup and Munk (1942) introduced the idea of a representative wave height parameter, called the *significant wave height.* This *significant wave height* was defined as *the average of the one third highest waves during an observation period* (sometimes denoted $(H_{\frac{1}{3}})$).

In 1952 Longuet-Higgins (using theory developed for noise in electrical circuits by Rice (1945)) derived a theoretical distribution for wave heights. He assumed that the height of the sea surface was a linear Gaussian process with a narrow banded spectrum. This leads to the Rayleigh distribution for wave heights. This assumption that the process is a linear and stationary allows one to describe the surface elevation $\eta(t)$ as a Fourier series:

$$\eta(t) = \sum_{n=1}^{\infty} a_n \cos(2\pi f_n t + \varepsilon_n)$$

where a_n, f_n and ε are the amplitude, frequency and phase respectively. a_n and f_n have certain probability distribution in the interval $[0, \infty)$, and ε_n is uniformly distributed in the interval $[0, 2\pi)$. Thus $\eta(t)$ can be considered as an infinite sum of random variables.

Several authors (viz. Goodnight and Russel (1963),Collins (1967),Goda (1974)) confirm that observed wave heights agree well with the Rayleigh distribution when individual waves are defined in terms of the zero up-crossing method. However, some authors have reported that larger wave heights do not agree well with the Rayleigh distribution. This discrepancy between the observation and theory has been reported in studies by Forristall (1978) and Tayfun (1980). Reasons given for the discrepancy include the non-linear, non-Gaussian characteristics of the sea surface and the effect associated with wide-band spectra.

Goda (1974) concluded the following with respect to the Rayleigh distribution:

"The zero up-crossing wave heights strictly do not belong to the class of Rayleigh distributions but they can be sufficiently well approximated for practical applications. The applicability of the Rayleigh distribution for the zero up-crossing wave heights and existence of the relation of wave statistical parameters are not influenced by the spectral shape nor the spectral width parameter."

The first theoretical derivation for the wave period distribution was given Longuet-Higgins (1975) in connection with the joint distribution of wave heights and periods.

Longuet-Higgins extended the Rayleigh law to obtain the theoretical joint probability density function of wave heights and periods. The normalized wave height X, and normalized period Y, whose means are equal to 1.0, have the following joint probability density function:

$$f_{XY}(x, y) = \frac{\pi x^2}{4\nu} \exp\left[\frac{-\pi}{4} x^2 \left(1 + \frac{(y-1)^2}{\nu^2}\right)\right]$$

for

$$x \geq 0, -\infty < y < \infty$$

where ν is a spectral width parameter given by

$$\nu = 0.866 * \text{Interquartile Range}(\tau)$$

where $\tau =$ zero up crossing period.

The marginal distributions were obtained by intergrating $f_{XY}(.,.)$ with respect to x or y.

$$f_X(x) = \frac{\pi}{2} x \exp[-\frac{\pi}{4}x^2], \qquad x \geq 0$$

$$f_Y(y) = \frac{1}{2\nu(1+\frac{(y-1)^2}{\nu})^{\frac{3}{2}}}, \qquad -\infty < y < \infty$$

It has been shown that the correlation coefficient is zero, hence, X and Y are considered as linearly uncorrelated random variables. However, the product of the marginal distributions is not the joint probability density function, therefore X and Y are not independent. Since wave height and period are not physically independent and have been observed as correlated random variables, this is not a desirable feature of the model. Another undesirable property of the Longuet-Higgins (1975) joint distribution is that the period has a negative range.

Chakrabuti and Cooley (1977) confirmed the practical applicability of the Longuet-Higgins (1975) distribution provided the spectrum is a narrow band single peak spectrum.

Yamazaki and Herbich (1985) make use of the Longuet-Higgins (1975) joint probability density function but with a modification which removes the negative range from the wave period. The modified distribution is then given by

$$f_{XY}(xy) = \frac{\pi x^2}{2\nu} \exp\left[-\frac{\pi}{4}x^2\left(\frac{1+(y-\frac{1}{y})^2}{\nu^2}\right)\right] \qquad \text{for } x \geq 0, y \geq 0$$

This density function behaves as the Longuet-Higgins distribution and it does not have a negative range. However, Cov $[X,Y]$ is zero, therefore the correlation coefficient is always zero.

Longuet-Higgins (1983) derive a theoretical probability density for the joint distribution of wave periods and amplitudes which has the following properties:

(1) The distribution is asymmetric, in accordance with observation; and

(2) It depends only on the three lowest moments, ($m0, m1, m2$), of the spectral density function.

The spectral width parameter is defined by $\nu = \left(\frac{m0*m2}{m1^2+1}\right)^{\frac{1}{2}}$. This definition of ν avoids the use of $m4$ (which depends on the behaviour of the spectrum at higher frequencies) was used in the Cavanie (1976) definition of the spectral width parameter ϵ where $\epsilon^2 = \left(\frac{1-m2^2}{m0*m4}\right)$.

Amoung the properties of this model is that the total distribution of wave heights is slightly non-Rayleigh, and that the interquartile range of the conditional wave period distribution tends to zero as the wave amplitude diminishes.

Longuet-Higgins (1983) states: "the Longuet-Higgins (1975) theoretical expression for the joint distribution of the period and amplitude of sea waves, which was based on a narrow-band approximation applied to the well known linear theory of Gaussian noise, gave a fairly good fit to wave data with a narrow spectrum but did not account for the asymmetry in the distribution of wave periods which is commonly observed in wave spectra with a broader bandwidth".

The Longuet-Higgins(1983) joint density is given by

$$f_{RT}(r,t) = \left(\frac{2}{\pi^{\frac{1}{2}}\nu}\right)\frac{r^2}{t^2}\exp\left(-r^2\left[1+\frac{(1-\frac{1}{t})^2}{\nu^2}\right]\right)L(\nu)$$

where R =normalized wave amplitude, T =normalized wave period, and $L(\nu)$ is a normalization factor introduced to take account of the fact that we consider only positive values of T .

A narrow band hypothesis is adopted, namely $\nu^2 \leq 1.0$ (in practice it is assumed (by Longuet-Higgins (1983)) that $\nu^2 \leq 0.36$). The effect of broadening the spectrum is to reduce the "most probable" joint values of the wave period and amplitude, and also to reduce their probabilty density (when $\nu \leq 1.0$).

Longuet-Higgins (1983) concludes by noting that this theoretical joint distribution of wave periods and amplitudes gives a reasonably good fit to some typical data, and that it depends only on the low order parameter ν .

Often however the spectral shape has a broad band spectral width and sometimes has multiple peaks and the aforementioned density functions never exhibit the mul-

timodal peaks. Yamazaki and Herbich (1985) suggest various non-parametric estimates using both series and kernel estimators.

A note on spectral width parameters.

Various spectral width parameters have been used in the literature:

$$\epsilon^2 = 1 - \left(\frac{Tc}{Tz}\right)^2$$

where Tz estimated by $Tz = \sqrt{\frac{m0}{m2}}$ and Tc estimated by $Tc = \sqrt{\frac{m2}{m4}}$ and therefore $\epsilon^2 = 1 - \frac{m2^2}{m0m4}$ (Cavinie (1976)).

One can think of the significance of this parameter as follows:

If the wave components cover a wide range of frequencies, the long waves will carry short waves on top of them and there will be many more crests than zero crossings, so that Tc will be much smaller than Tz and ϵ will be nearly one. If, on the one hand, there is a simple swell which contains only a narrow range of frequencies, each crest will be associated with a zero crosing, so that Tc will be approximately equal to Tz and ϵ will be nearly zero.

Longuet-Higgins (1975) uses $\nu = 0.866 *$ Interquartile Range(τ) where τ is the zero up-crossing period.

Longuet-Higgins (1983) uses $\nu = \left(\frac{m0m2}{m1^2-1}\right)^{\frac{1}{2}}$

An indication as to what is considered narrow banded is given by Longuet-Higgins (1983) where the spectra is considered narrow banded if $\nu << 1.0$ or $\nu^2 \leq 0.36$. Longuet-Higgins(1983) also states that Cavanie (1976) data (with an $\epsilon = 0.865$) is considered broad banded.

Chakrabuti and Cooley (1977) have data with ϵ values ranging from 0.57 to 0.8 and it is considered broad banded.

For the data used in this study we get (when using the Longuet-Higgins (1983) definition for ν):

$$\text{mean } \nu = 9.407 \text{ and variance } = 3.003$$

When using $\epsilon^2 = \left(1 - \left(\frac{Tc}{Tz}\right)^2\right)$ we get a mean $\epsilon = 0.881$.

Clearly the data set used in this study cannot be considered narrow banded.

Initially we thought it might be possible to express $Hm0$ in terms of wave amplitude and then base the distribution of $Hm0$ on the theory developed by Longuet-Higgins (1983). Unfortunately we do not have the actual wave amplitude measurements and we do not know how well $Hm0$ estimates the significant wave height, $(H_{\frac{1}{3}})$. In addition we are not absolutely sure that the distribution of actual wave heights is Rayleigh. The main point here is that the Longuet-Higgins (1983) joint amplitude and period distribution is for narrow band spectra. An exploratory data analysis reveal that the data we have is far from narrow banded. The measuring device used in the Longuet-Higgins studies resided slightly under the water level and therefore did not record the higher frequency component and tended to infer that the waves fall into the "narrow band" category. The data used in this study was collected using a measuring device that remained on the sea surface and recorded the higher frequency component.

Thus we were not able to apply any of the theory available in the literature and new models had to be developed.

CHAPTER 3.

EXPLORATORY DATA ANALYSIS

This chapter is a summary of the exploratory analysis which was carried out on the data. The purpose of this analysis was to identify the main patterns in the two time series $Hm0$ and Tz, and then to examine each of the patterns separately. The patterns and features found at this stage of the study determine the type of components which would be required to model the data. In particular it was established that any adequate model for these series would have to incorporate seasonal fluctuations, serial correlations as well as an unusual distribution for the residual process. The latter proved to be by far the most difficult component to model and required a special study in itself.

In exploring different ways to cope with the non-standard residual processes we found that the spectral moments $m0$ and $m2$ (or more precisely their logarithm) were easier to model than $Hm0$ and Tz. Mathematically the two pairs $(m0, m2)$ and $(Hm0, Tz)$ are equivalent in the sense that they are simple transformations of each other

$$Hm0 = 4\sqrt{m0} \qquad Tz = \sqrt{\frac{m0}{m2}}$$

However the residual processes associated with time series models for $Hm0$ and Tz are not symmetric whereas those associated with $\ln m0$ and $\ln m2$ were found to be approximately symmetric. Thus one of the conclusions reached as a result of the exploratory analysis was that it would be more convenient to construct models for $m0$ and $m2$ rather than $Hm0$ and Tz. By transforming back from $(\ln m0, \ln m2)$ to $(Hm0, Tz)$ these models can then be used to describe the behaviour of $Hm0$ and Tz.

We now consider the various components of the data series in more detail.

Seasonal structure.

Initially the data were grouped into the following four seasons:

Summer Months: December, January, February

Autumn Months: March, April, May

Winter Months: June, July, August

Spring Months: September, October, November

The sample statistics of each of these seasons are given in Table 3.1 for $Hm0$ and Tz and in Table 3.2 for $m0$ and $m2$ where we had the following number of observations for each season: Summer (2036), Autumn (2095), Winter (1942), and Spring (1886).

It is clear that in all cases both the mean and standard deviation vary over the year. Approximate statistical tests of significance confirmed this. (We note that since the data are serially correlated the independence assumption required for such tests are not met. To circumvent this we took sub-samples of the observations which were sufficiently far apart to be approximately independently distributed.)

In particular the means and standard deviations (for the $Hm0$ series) are larger during the Winter months than during those of Summer. For the Tz series the means are largest during Winter and smallest during the Spring whereas the standard deviations are largest during Autumn and smallest during Summer. For both the $m0$ and $m2$ series the means and standard deviations are largest during Winter and smallest during Summer.

Having established the existence of seasonal fluctuations in the time series the next problem is to find a way to model this. One possibility is to divide the year into a fixed number of 'seasons', such as the four considered above, and then to fit a separate mean and standard deviation to each of them. This approach has the disadvantage of introducing discontinuities into the model; the mean and standard deviation changes abruptly at the end of each season. This is unsatisfactory from the point of view that we would not expect there to be such abrupt changes in reality. Secondly this would introduce non-stationarity in the residual process, and finally it would be difficult to decide how many seasons to use and how to define them since they need not all be of equal length.

A preferable approach is to allow the mean and standard deviation to fluctuate smoothly over the year. This can be conveniently achieved by making use of a truncated Fourier series expansion of the mean and standard deviation.

Let μ_t and σ_t, $t = 1, 2, \ldots$ denote the mean and standard deviation function of the time series, y_t, where we assume that these functions are periodic with peroid equal to one year. In this study we let $t = 1$ for June 1^{st} at 0600 hours, $t = 2$

for June 1^{st} at 1200 hours and so on.

Then for example μ_t can be represented (exactly) by:

$$\begin{aligned}\mu_t = \alpha_0 &+ \alpha_1 \cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2 \sin\left(\frac{2\pi}{1460} * t\right) \\ &+ \alpha_3 \cos\left(\frac{2\pi}{1460} * 2t\right) + \alpha_4 \sin\left(\frac{2\pi}{1460} * 2t\right) \\ &+ \\ &\vdots \\ &+ \alpha_{1457} \cos\left(\frac{2\pi}{1460} * 729t\right) + \alpha_{1458} \sin\left(\frac{2\pi}{1460} * 729t\right) \\ &+ \alpha_{1459} \cos\left(\frac{2\pi}{1460} * 730t\right)\end{aligned}$$

$$\begin{aligned}= \alpha_0 &+ \sum_{i=1}^{729} \left(\alpha_{2i-1} \cos\left(\frac{2\pi}{1460} * it\right) + \alpha_{2i} \sin\left(\frac{2\pi}{1460} * it\right)\right) \\ &+ \alpha_{1459} \cos\left(\frac{2\pi}{1460} * 730t\right)\end{aligned}$$

Thus the 1460 values $\mu_1, \mu_2, \ldots, \mu_{1460}$ (being the means of the 4 records for each of the 365 days) can be reparameterized in terms of the 1460 parameters $\alpha_0, \alpha_1, \ldots, \alpha_{1459}$.

The point of this reparameterization is that many natural series (see for example Zucchini & Adamson (1984) and Brandao (1986)) vary about a smooth function which approximately follows the pattern of a cosine function. Many of the high frequency coefficients in the above expression are approximately zero and we can write

$$\mu_t = \alpha_0 + \sum_{i=1}^{L} \left(\alpha_{2i-1} \cos\left(\frac{2\pi}{1460} * it\right) + \alpha_{2i} \sin\left(\frac{2\pi}{1460} * it\right)\right)$$

where L can be quite small (even as low as $L = 1$). Thus by reparameterizing the series in this way we can achieve smoothness and also very effectively reduce the number of parameters which need to be estimated, an important consideration in achieving stability of the final model.

The same type of approximation can be applied to the standard deviation

$$\sigma_t = \beta_0 + \sum_{i=1}^{L'} \left(\beta_{2i-1} \cos\left(\frac{2\pi}{1460} * it\right) + \beta_{2i} \sin\left(\frac{2\pi}{1460} * it\right)\right)$$

The question of how to select values for L and L' will be considered later. For the purposes of removing the seasonal component in order to continue with the exploratory analysis we selected $L = L' = 1$, i.e. we have

$$\mu_t = \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

$$\sigma_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi}{1460} * t\right) + \beta_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

The next problem was to obtain approximate estimates of the parameters $\alpha_0, \alpha_1, \alpha_2$ and β_0, β_1 and β_2 . When fitting the full model these parameters are estimated using maximum likelihood, but at this stage we were not yet in a position to make distributional assumptions about the residual process and a least squares procedure was used instead. Since the standard deviation also fluctuates seasonally it is necessary to apply weighted least squares rather than ordinary least squares. One has to minimize

$$\sum_{t=1}^{n} {e_t}^2 = \sum_{t=1}^{n} \left(\frac{y_t - \mu_t}{\sigma_t}\right)^2$$

with respect to the parameters. We note that this is a non-standard problem in this application. Firstly we are estimating the parameters of both the mean and standard deviation simultaneously, and secondly there are missing observations in the data.

Besides the gaps in the data set we also omitted observations taken at times other than at $t = (0600; 1200; 1800; 2400)$ hours. Observations on February 29^{th} are also omitted. In total there were 7880 valid observations.

It is not possible to obtain an explicit solution to the above minimization problem. A numerical solution had to be devised instead. (This is discussed in Appendix B.)

The parameter estimates for the series $m0$ and $m2$ are the following

For $m0$

$$\alpha_0 = 0.5117$$

$$\alpha_1 = 0.0608$$

$$\alpha_2 = 0.1107$$

$$\beta_0 = 0.1759$$

$$\beta_1 = 0.0712$$

$$\beta_2 = 0.0716$$

For $m2$

$$\alpha_0 = 0.0107$$

$$\alpha_1 = -0.00016$$

$$\alpha_2 = 0.0020$$

$$\beta_0 = 0.00005$$

$$\beta_1 = 0.00001$$

$$\beta_2 = 0.00002$$

For the purposes of interpretation it is convenient to write the mean function in a phase and amplitude representation.

For $m0$ we now have a mean curve of the following form:

$$\widehat{\mu}_t = 0.5117 + 0.0608 \cos\omega t + 0.1107 \sin\omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

We then obtain values for R and ψ (the amplitude and phase) $R = 0.1262$ and $\psi = -1.0685$. (Details of Phase and Amplitude formula are given in Appendix D.)
The mean curve can be written as

$$\widehat{\mu}_t = \widehat{\alpha_0} + R\cos(\omega(t-\psi))$$

$$= 0.5117 + 0.1262 \cos(\omega(t+1.0685))$$

These features are clearly shown in Figure 3.A and give an indication of the extent of the seasonal component of the data.

This has the following interpretation:

The maximum value for the mean curve is $\alpha_0 + R = 0.6377$ which occurs on the 6^{th} of August. The minimum value for the mean curve is $\alpha_0 - R = 0.3854$ which occurs on the 30^{th} of January.

For the $(m2)$ series we have:

$$\begin{aligned}\widehat{\mu}_t &= 0.0107 - 0.00016\ \cos\omega t\ +\ 0.0020\ \sin\omega t \\ &= 0.0107 + 0.002\ \cos(\omega(t + 1.6506))\end{aligned}$$

Figure 3.B shows the extent of the seasonal component of this series. This has the following interpretation:

The maximum value for the mean curve is $\alpha_0 + R = 0.0127$ which occurs on the 3^{rd} of September. The minimum value for the mean curve is $\alpha_0 - R = 0.0087$ which occurs on the 5^{th} of March.

For the variance curve we have a model of the following form (for the $m0$ series):

$$\begin{aligned}\widehat{\sigma}_t &= 0.1759 + 0.0712\ \cos\omega t\ +\ 0.0716\ \sin\omega t \\ &= 0.1759 + 0.1009\ \cos(\omega(t + 0.7881))\end{aligned}$$

This has the following interpretation: (see Figure 3.C)

The maximum value for the mean curve is $\beta_0 + R = 0.2768$ which occurs on the 15^{th} of July. The minimum value for the mean curve is $\beta_0 - R = 0.0749$ which occurs on the 14^{th} of January.

For the $(m2)$ series we have a variance curve of the following form:

$$\begin{aligned}\widehat{\sigma}_t &= 0.00005 + 0.00001\ \cos\omega t\ +\ 0.00002\ \sin\omega t \\ &= 0.00005 + 0.00002\ \cos(\omega(t - 1.107))\end{aligned}$$

This has the following interpretation:

The maximum value for the mean curve is $\beta_0 + R = 0.000072$ which occurs on the 27^{th} of March. The minimum value for the mean curve is $\beta_0 - R = 0.000028$ which occurs on the 26^{th} of September. No plot is given of this curve since there is very little variation of the values.

It is quite easy to extend this to more complicated means and standard deviations, i.e. with more terms, and in fact this was done at a later stage but with a negligible improvement in the model.

Serial Correlation.

Having found preliminary estimates of the seasonal components of the model we now consider the serial correlation structure. To do this we consider the residual process

$$e_t = \left(\frac{y_t - \mu_t}{\sigma_t}\right) \qquad t = 1, 2, \ldots$$

where μ_t and σ_t are then

$$\mu_t = \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

and

$$\sigma_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi}{1460} * t\right) + \beta_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

By construction, the process e_t, $t = 1, 2, \ldots$ is approximately stationary. To identify a suitable model for this process we examine the autocorrelation and partial autocorrelation structure. The formula for estimating the autocorrelation and partial autocorrelation functions (acf and pacf respectively) can be found in Box-Jenkins (1970). The acf and pacf of each of the series of residuals for $m0$ and $m2$ are given in Figures 3.D, 3.E, 3.F and 3.G and in Tables 3.3 and 3.4.

We then have to decide on the appropriate time series model for the data and we do this by examining the behaviour of the acf and pacf. Box-Jenkins (1970) make use of model identification techniques which identify a process as an auto-regressive of order one (AR(1)) if the acf exhibits 'exponential decay' and the only significant non-zero value of the pacf is that for lag 1. When examining the acf and pacf (Figures 3.D,E,F,G) it can be seen that this is exactly what we have and the process can be identified as an AR(1).

The acf and pacf for the series

$$e_t = \left(\left(\frac{y_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{y_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right)$$

(where ϕ is estimated by the serial correlation coefficient for lag 1) reveal a series of independent e_t values, which confirm that the processes are AR(1). These values are given in Table 3.5 and Figures 3.H and 3.I for the $m0$ data and in Table 3.6 and Figures 3.J and 3.K for the $m2$ data.

Residual process.

So far, the exploratory analysis had revealed that the processes we are considering are of the form:

$$e_t = \left(\left(\frac{y_t - \mu_t}{\sigma_t} \right) - \phi \left(\frac{y_{t-1} - \mu_{t-1}}{\sigma_{t-1}} \right) \right)$$

where the e_t are independently and identically distributed with zero mean. Furthermore it could be assumed that the parameter function μ_t and σ_t can be parsimoniously modeled using a truncated form of the Fourier representation. It remained to find a suitable distribution for the residuals, e_t, $t = 1, 2, \ldots$.

Although the e_t are (by construction) independently distributed and stationary they are not normally distributed. In fact they are not even symmetrically distributed. Various transformations were applied and it was found that the ln transformation on the original data, yielded residuals which were approximately symmetrically distributed. These distributions are shown in Figures 3.L and 3.M.

Although it was possible to achieve symmetry by using a transformation we were unable to find a transformation that yielded residuals which were both symmetric and normally distributed. A number of the standard (and several non-standard) transformations were applied but none yielded the desired distribution. In retrospect, this is not surprising, it would be indeed fortuitous to find a single transformation (on the original data) which acheived both objectives.

Had the time series with which we are dealing been dominated by seasonal variation, that is if the residuals had been relatively small, then it would have been not unreasonable to simply ignore the fact that the residual distributions are non-normal. It would have had little effect on the final models which particular distributions were fitted to the residuals. However our particular series have a quite low "signal-to-noise ratios" thereby making it essential that the residual distributions are modelled accurately.

The conclusion we drew from this part of the preliminary analysis was that it would be necessary to develop special models for this type of time series, in particular the error distribution would have to be modelled using some flexible family of distributions, since it could not be assumed that any of the standard two parameter distributions (such as lognormal, gamma, Weibull etc) would be sufficiently flexable to adequately represent the observed distribution of the residuals. We therefore decided to apply a transformation which would yield symmetrically distributed residuals and then to make a study of the ways of modelling these. The Elphinstone (1985) family of models, to be discussed in the next chapter offered a methodology to go about flexibly dealing with these unusually distributed residuals.

Summary

The main findings of this preliminary analysis were that a model for the time series $m0$ and $m2$ would have to be of the form:

$$Y_t = \ln(m0_t)$$

$$X_t = \ln(m2_t)$$

where

$$\mu_t = \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

$$\sigma_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi}{1460} * t\right) + \beta_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

Then

$$e_t = \left(\left(\frac{y_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{y_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right)$$

(and similarly for the X_t series) where the e_t are independently and identically distributed with zero mean, and that the error process would need to be modelled using some flexible family of distributions.

TABLE 3.1

Season		$Hm0(m)$	Tz(seconds)
WINTER:	Mean	2.909	7.160
	Standard Deviation	1.075	1.542
	Maximum	8.020	12.740
	Minimum	0.500	1.700
SPRING:	Mean	2.807	7.670
	Standard Deviation	1.051	1.558
	Maximum	8.670	14.780
	Minimum	0.340	1.720
SUMMER:	Mean	2.408	6.520
	Standard Deviation	0.796	1.442
	Maximum	8.600	13.190
	Minimum	0.200	1.880
AUTUMN:	Mean	2.585	7.330
	Standard Deviation	1.062	1.688
	Maximum	10.800	15.740
	Minimum	0.200	1.700

TABLE 3.2

Season		$m0$	$m2$
WINTER:	Mean	0.594	0.012
	Standard Deviation	0.457	0.008
	Maximum	4.020	0.068
	Minimum	0.015	0.0003
SPRING:	Mean	0.561	0.012
	Standard Deviation	0.458	0.007
	Maximum	4.698	0.057
	Minimum	0.007	0.0001
SUMMER:	Mean	0.402	0.009
	Standard Deviation	0.030	0.005
	Maximum	4.622	0.048
	Minimum	0.002	0.00003
AUTUMN:	Mean	0.490	0.009
	Standard Deviation	0.451	0.006
	Maximum	7.290	0.063
	Minimum	0.002	0.00002

TABLE 3.3

AUTOCORRELATIONS AND PARTIAL AUTOCORRELATIONS FOR THE *m0* SERIES.

Lag	AC	PAC
1	0.837	0.830
2	0.673	-0.074
3	0.504	-0.103
4	0.360	-0.023
5	0.256	0.028
6	0.178	-0.026
7	0.123	-0.007
8	0.080	0.020
9	0.044	-0.019
10	0.009	-0.019
11	-0.015	-0.033
12	-0.035	-0.005

TABLE 3.4

AUTOCORRELATIONS AND PARTIAL AUTOCORRELATIONS FOR THE *m2* SERIES.

Lag	AC	PAC
1	0.830	0.830
2	0.654	-0.083
3	0.490	-0.033
4	0.357	-0.031
5	0.265	0.008
6	0.205	0.048
7	0.162	0.006
8	0.122	0.004
9	0.085	-0.013
10	0.048	-0.016
11	0.024	-0.013
12	0.012	0.040

TABLE 3.5

AUTOCORRELATIONS AND PARTIAL AUTOCORRELATIONS FOR THE e_t $(m0)$ SERIES.

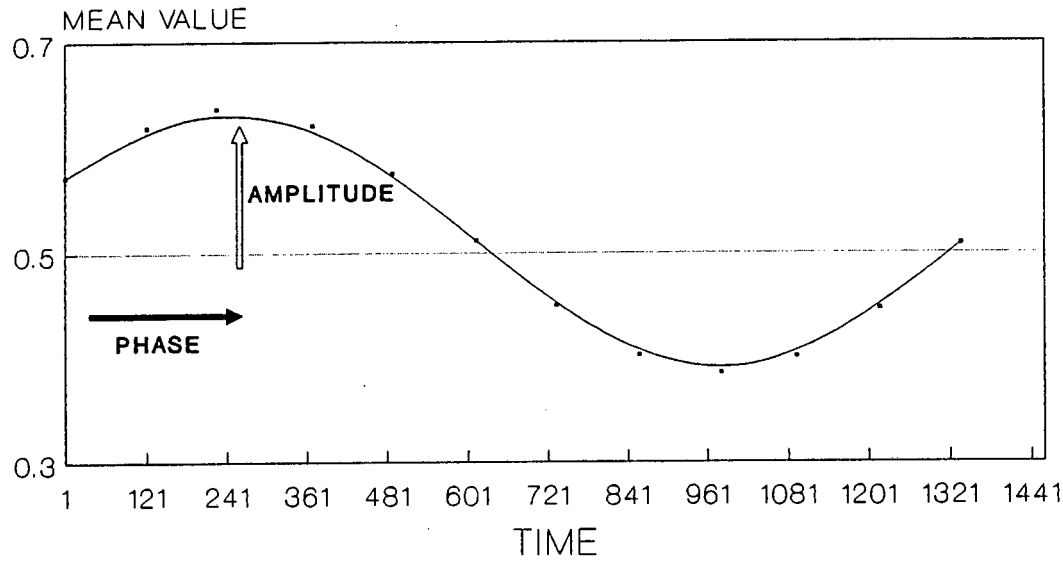
Lag	AC	PAC
1	0.080	0.080
2	0.045	0.039
3	-0.011	-0.018
4	-0.062	-0.062
5	-0.068	-0.058
6	-0.046	-0.031
7	-0.034	-0.024
8	-0.026	-0.024
9	-0.015	-0.018
10	0.005	0.001
11	-0.009	-0.017
12	-0.001	-0.007

TABLE 3.6

AUTOCORRELATIONS AND PARTIAL AUTOCORRELATIONS FOR THE e_t $(m2)$ SERIES.

Lag	AC	PAC
1	0.055	0.055
2	-0.006	-0.009
3	-0.017	-0.016
4	-0.052	-0.051
5	-0.079	-0.074
6	-0.025	-0.018
7	-0.011	-0.012
8	0.009	0.005
9	0.014	0.005
10	0.007	-0.001
11	-0.043	-0.048
12	0.014	0.018

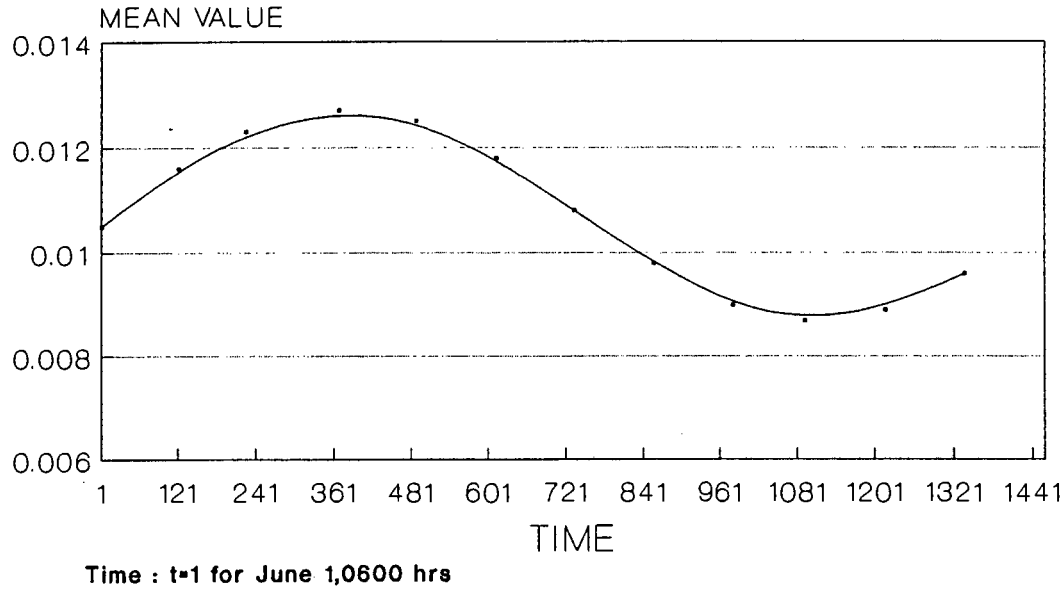
PERIODICITY OF MEAN CURVE (M0)



Time : t=1 for June 1,0600 hrs

FIGURE 3.A

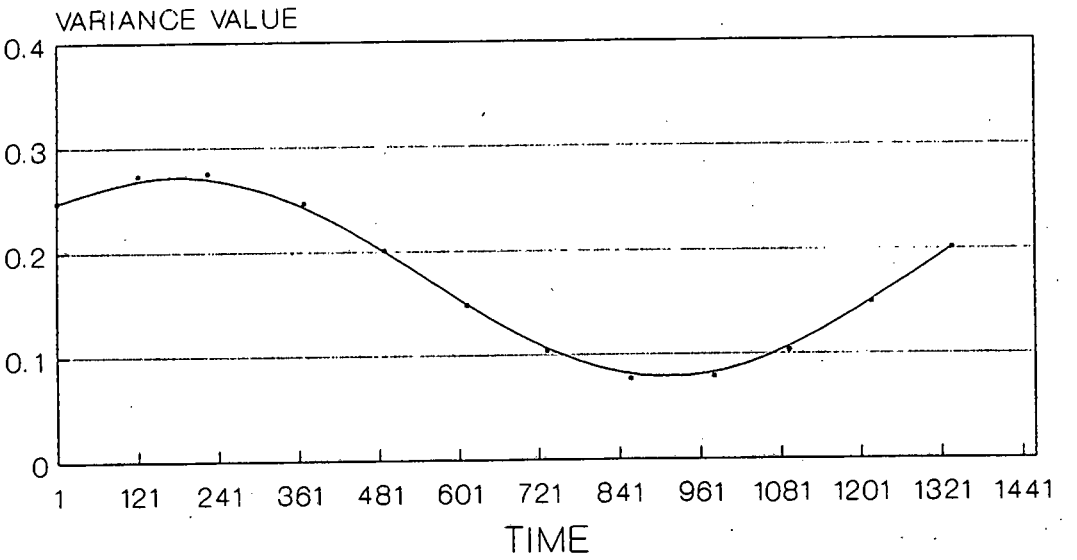
PERIODICITY OF MEAN CURVE (M2)



Time : t=1 for June 1,0600 hrs

FIGURE 3.B

PERIODICITY OF VARIANCE CURVE (M0)



Time : t=1 for June 1,0600 hrs

FIGURE 3.C

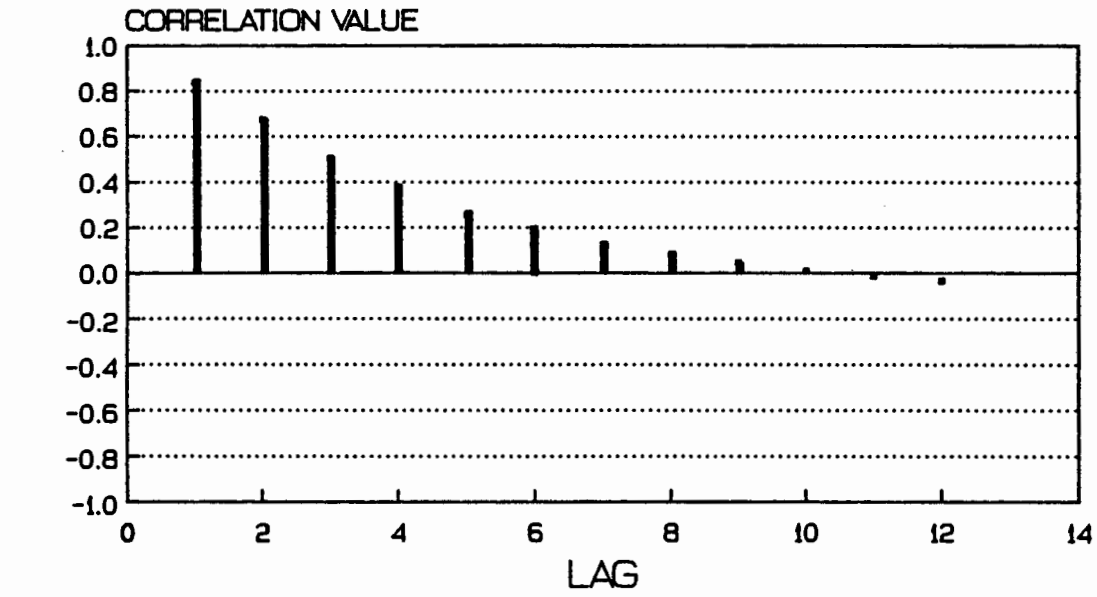


FIGURE 3.D

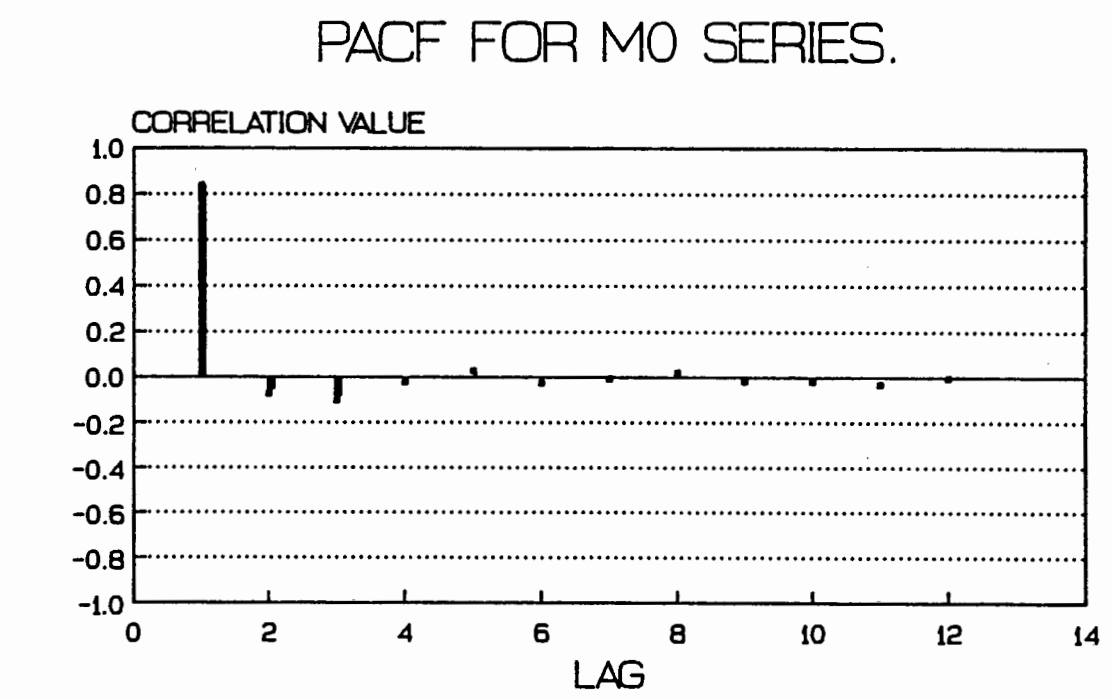


FIGURE 3.E

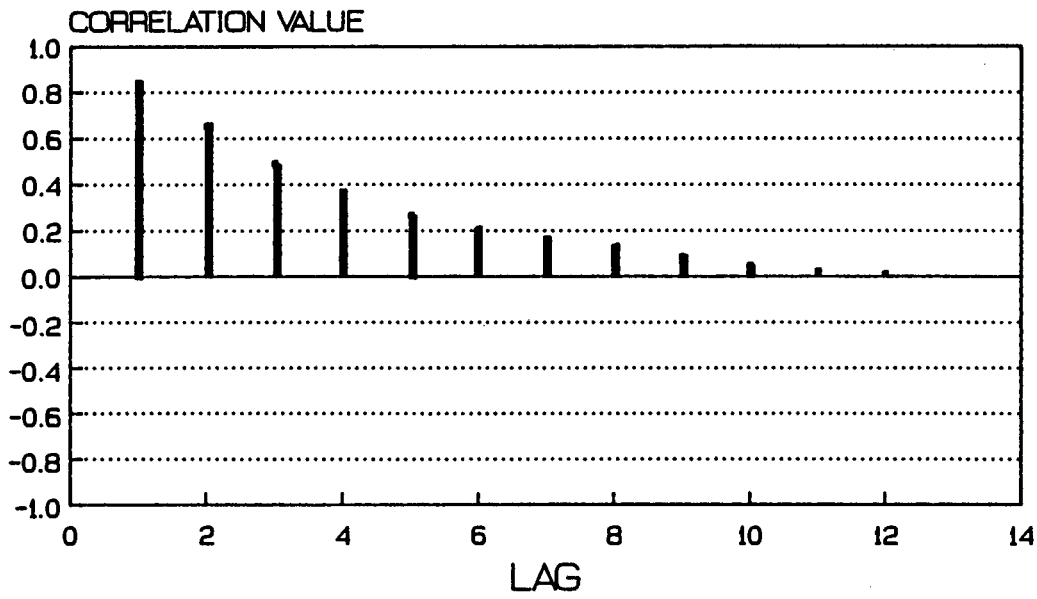


FIGURE 3.F

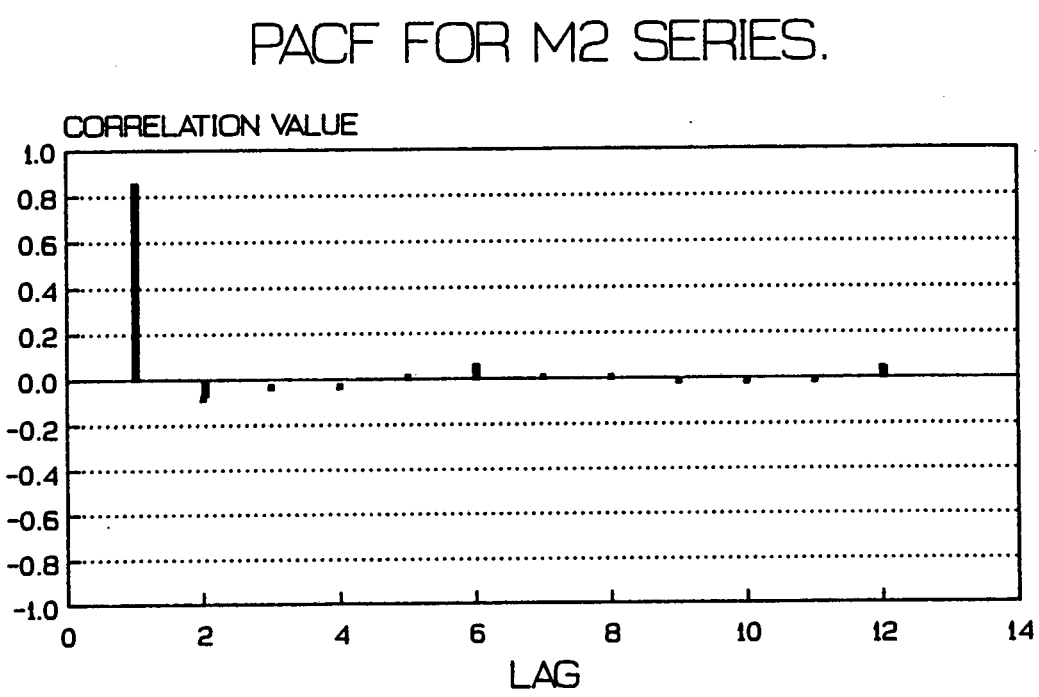


FIGURE 3.G

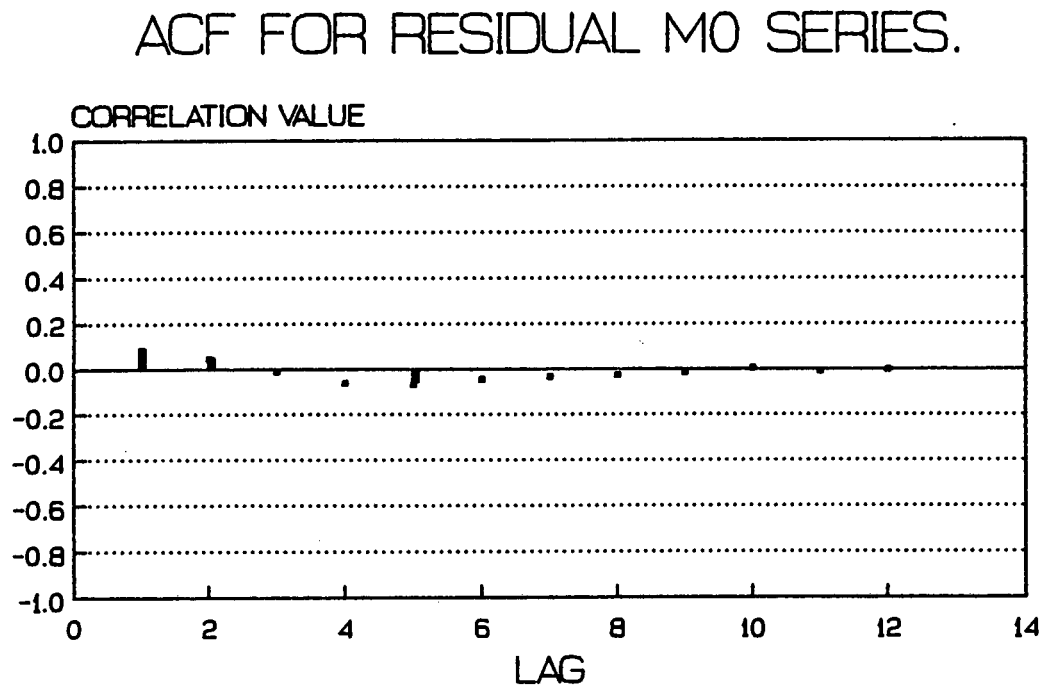


FIGURE 3.H

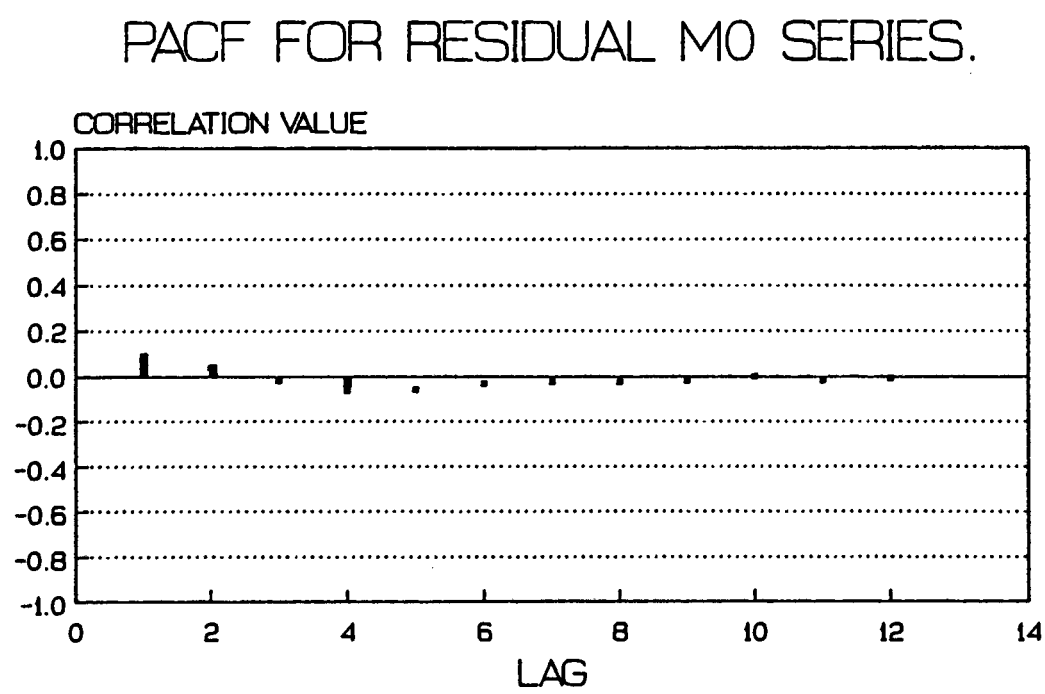


FIGURE 3.I

ACF FOR RESIDUAL M2 SERIES.

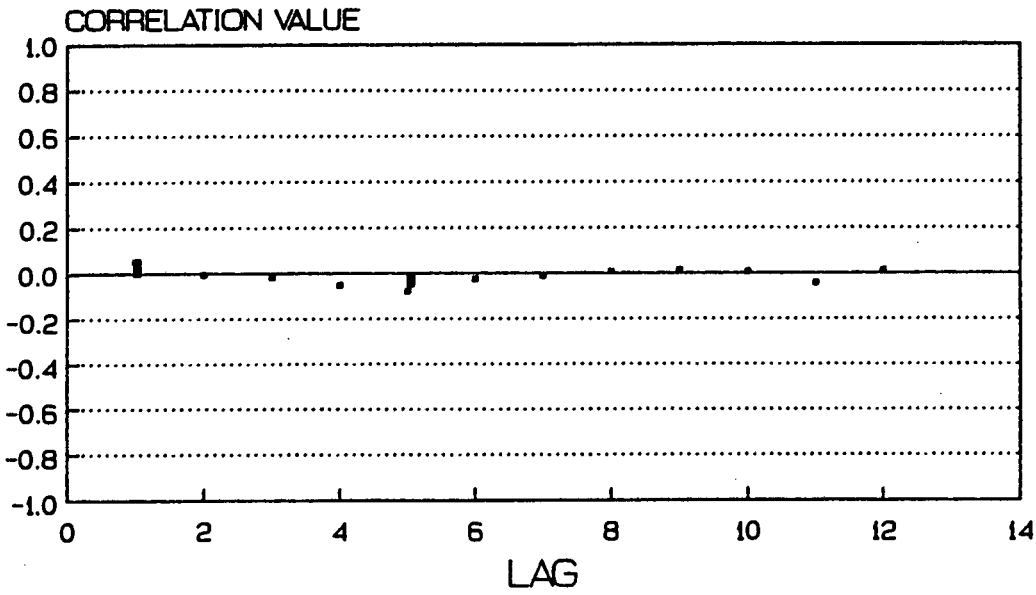


FIGURE 3.J

PACF FOR RESIDUAL M2 SERIES.

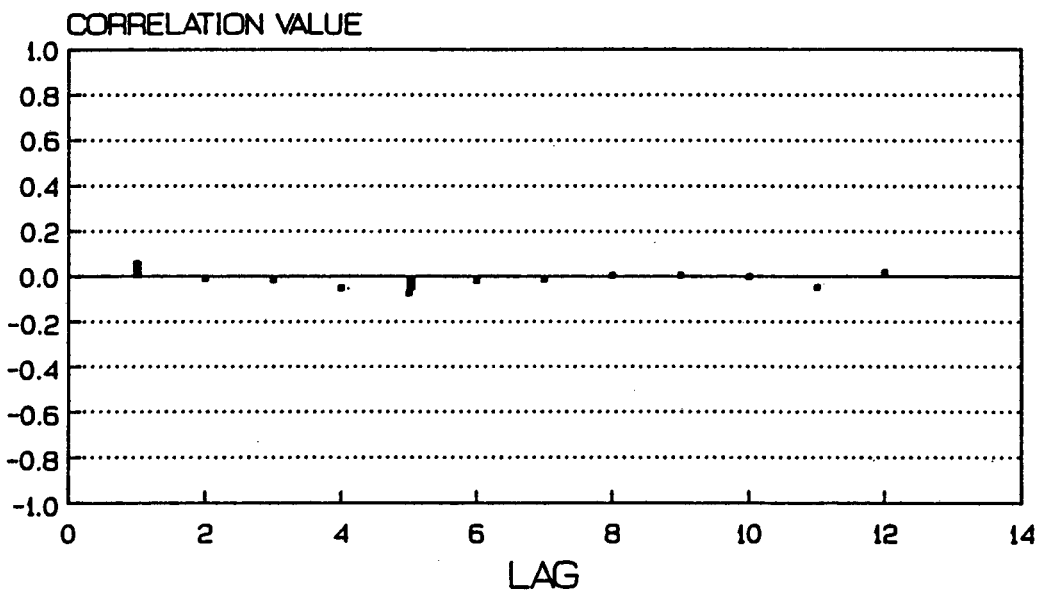


FIGURE 3.K

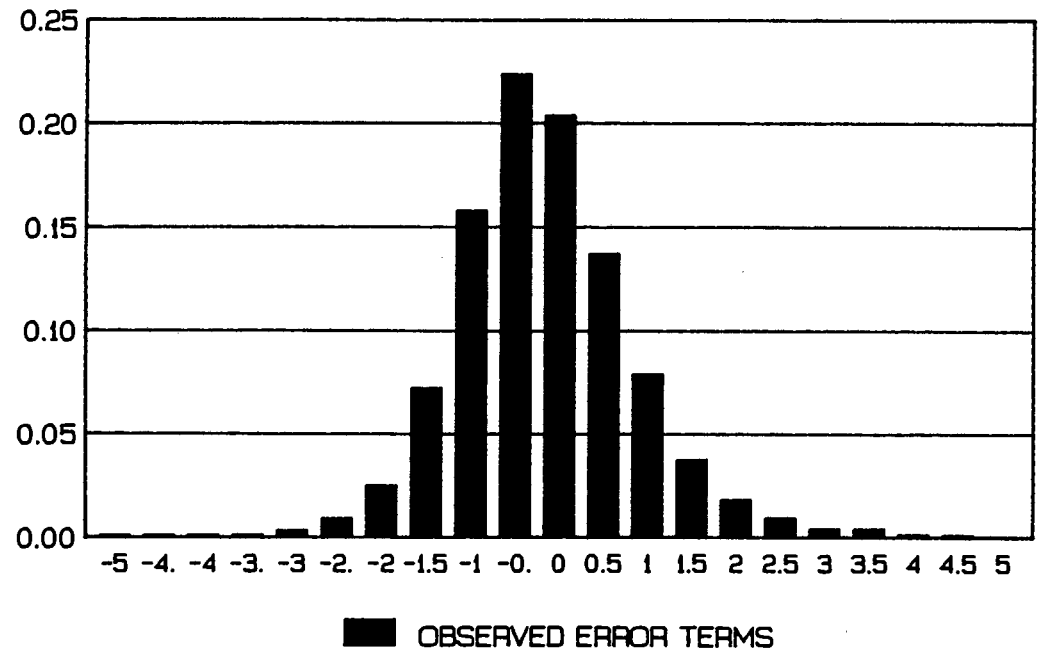


FIGURE 3.L

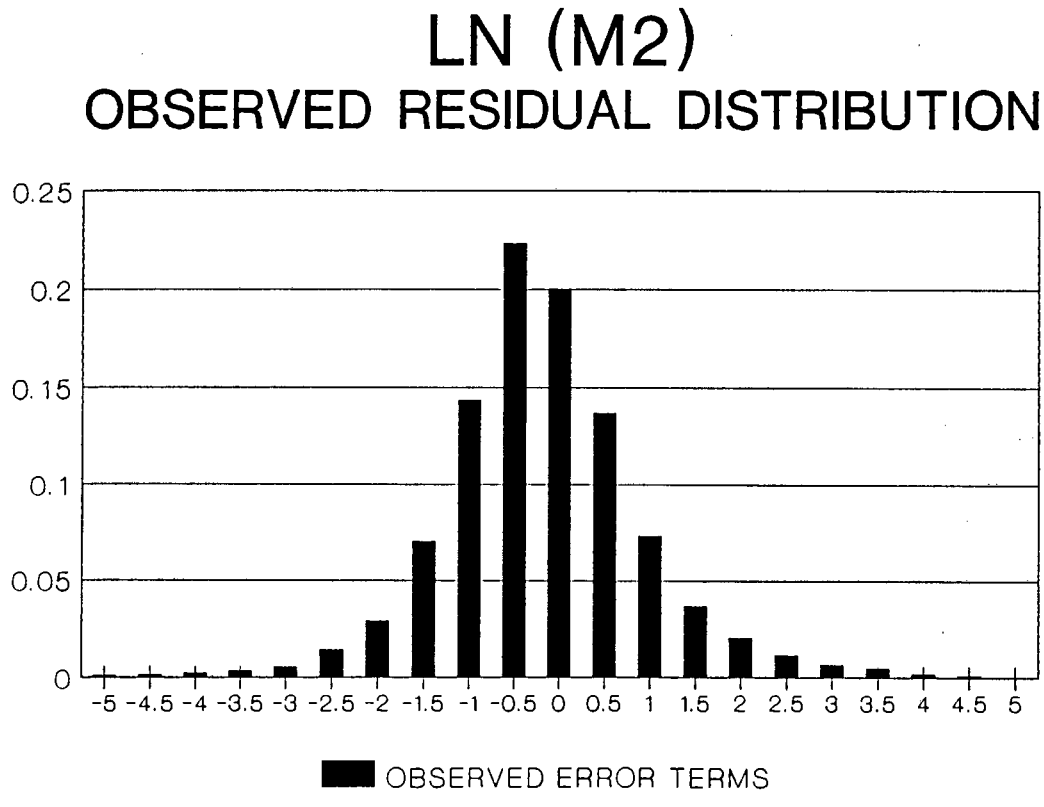


FIGURE 3.M

CHAPTER 4.

ELPHINSTONE'S FAMILY OF TARGET DISTRIBUTION MODELS

The exploratory analysis summarized in the last chapter established that in order to usefully model the time series $\ln m0$ and $\ln m2$ it is necessary to find a flexible family of distributions for the process of residuals. The techniques introduced by Elphinstone were found to be suitable for this purpose. In this chapter we give an outline of the those aspects of the methodology which apply to our application. A full account of the theory is given in Elphinstone (1983) and (1985).

Suppose that we wish to fit a distribution to a sample of independently distributed realisation of some unknown distribution function $F(.)$, and with corresponding probability density function $f(.)$. One begins by choosing a convenient distribution function $H(.)$, called the target distribution, with corresponding density function $h(.)$.

The method consists of successively approximating $F(x)$ by a transformation of the target distribution:

$$F_k(x;\theta) = H[g_k(x;\theta)] \qquad k = 1, 2, \ldots$$

where $g_k(x;\theta)$ is defined by equation (4.2). These coefficients are estimated from the data. The approximation is improved by increasing the degree of the polynomial, thereby introducing additional parameters into the model. In this way it is possible to transform a standard distribution to a rich variety of non-standard distributions.

The number of parameters which are required to achieve an acceptable approximation depends, of course, on the unknown, $F(.)$ and also on the choice of the target distribution $H(.)$. The "closer" these two distributions are the more rapidly this method will yield a sufficiently accurate approximation. The question of how large one should make k will be discussed later.

In our application we considered two alternatives of the target distribution, namely Double Exponential and Normal.

We consider now some of the details involved in applying this theory:

A positive polynomial $P_k(t;\Lambda)$ can be defined as

$$\begin{aligned} P_k(t;\Lambda) &= \alpha(\lambda) \qquad \text{for k} = 0 \\ &= \alpha(\lambda)\prod_{j=1}^{k}\left[1 - 2\lambda_{j1}t + ({\lambda_{j1}}^2 + \beta(\lambda_{j2}))t^2\right] \qquad \text{for k} > 0 \end{aligned} \tag{4.1}$$

where $\alpha(\lambda)$ and $\beta(\lambda_{j2})$ are functions of λ and λ_{j2} respectively such that $\alpha(\lambda) \geq 0$ and $\beta(\lambda_{j2}) \geq 0$.

The positive polynomial $P_k(t;\Lambda)$ can be used to construct a non-decreasing polynomial $g_k(x;\theta)$ of order $m = 2k+1$ as:

$$\begin{aligned} g_k(x;\theta) &= \xi + \int_0^x P_k(t;\lambda)dt \\ \text{where } \theta' &= (\xi, \Lambda') \\ &= (\xi, \lambda, \lambda_{11}, \lambda_{12}, \lambda_{21}, \ldots, \lambda_{k2}) \end{aligned} \tag{4.2}$$

It has been shown that $g_k(x;\theta)$ can be used to approximate any continuous monotonic function so that an approximation to the unknown distribution $F(x)$ will be obtained as:

$$F_k(x;\theta) = H[g_k(x;\theta)]$$

Alternatively, the derivative of the above gives

$$f_k(x;\theta) = h[g_k(x;\theta)]P_k(x;\Lambda)$$

and can be used to approximate the unknown density

$$f(x) = \frac{dF(x)}{dx}$$

and where $P_k(x;\Lambda)$ is the previously defined positive polynomial, (equation (4.1)). When θ is estimated by $\hat{\theta}$ on the basis of the observed data then

$$f_k(x;\hat{\theta}) = h[g_k(x;\hat{\theta})]P_k(x;\Lambda)$$

will be the estimate of the working model that best approximates $f(x)$ and where $h(.)$ is the target density function.

The scale parameter $\alpha(\lambda)$ must be positive for the polynomial transformation to be monotonically increasing. It is achieved in this study by setting

$$\alpha(\lambda) = \lambda \qquad \text{where } \lambda \geq 0$$

The function $\beta(\lambda_{j2})$ is required to be positive and this is achieved by setting

$$\beta(\lambda_{j2}) = \lambda_{j2} \qquad \lambda_{j2} \geq 0$$

In order to apply the method of maximum likelihood the distribution estimator needs to be differentiated to obtain the density. The negative log likelihood is thus minimized to obtain the parameter estimates.

Thus the procedure is to minimize L with respect to the elements of θ, where L is given by

$$\begin{aligned} L &= -\ln \prod_{i=1}^{N} f_k(x_i;\theta) \\ &= -\sum_{i=1}^{N} \ln\left(f_k(x_i;\theta)\right) \end{aligned}$$

where x_i, $i = 1,\ldots,N$ are the N observed sample points.

A number of methods are given for selecting the most appropriate index k for any target distribution estimator are given in Elphinstone (1985).

In this study the Akaike Information Criterion, (AIC), and the Schwartz Information Criterion, (SIC), were applied.

For AIC we use that k which minimizes $AIC(k)$ where

$$AIC(k) = -2 * \ln(\text{maximum likelihood}) + 2p$$

where $p = 2k + 2$ is the number of parameters in the model.

For SIC we use that k which minimizes $SIC(k)$ where

$$SIC(k) = -\ln(\text{maximum likelihood}) + 0.5p \ln N$$

The Schwartz Information Criterion has the effect of introducing a larger penalty than the Akaike Criterion whenever there are more than 8 data points, resulting in lower order models being selected.

For $k = 0$ and $k = 1$ the models have the following form:

For k=0:

$$P_k(t;\Lambda) = \lambda$$

$$g_k(x;\theta) = \xi + \lambda x$$

For k=1:

$$\begin{aligned} P_k(t;\Lambda) &= \lambda[1 - 2\lambda_{11}t + {\lambda_{11}}^2 t^2 + \lambda_{12}t^2] \\ &= [\lambda - 2\lambda\lambda_{11}t + \lambda{\lambda_{11}}^2 t^2 + \lambda\lambda_{12}t^2] \\ g_k(x;\theta) &= (\xi + \lambda x - \lambda\lambda_{11}x^2 + \frac{1}{3}\lambda{\lambda_{11}}^2 x^3 + \frac{1}{3}\lambda\lambda_{12}x^3) \\ &= [\xi + \lambda x + (-\lambda\lambda_{11})x^2 + (\frac{1}{3}\lambda{\lambda_{11}}^2 + \frac{1}{3}\lambda\lambda_{12})x^3] \end{aligned}$$

It is usually an advantage to calculate (and then use) the coefficients of the polynomial expansion of $g_k(x;\theta)$.

For $k = 0$

The coefficient of x^0 : is given by ξ

The coefficient of x^1 : is given by λ

For $k = 1$

The coefficient of x^0 : is given by ξ

The coefficient of x^1 : is given by λ

The coefficient of x^2 : is given by $(-\lambda\lambda_{11})$

The coefficient of x^3 : is given by $(\frac{1}{3}\lambda{\lambda_{11}}^2 + \frac{1}{3}\lambda\lambda_{12})$

In this study the coefficient of x^0 is denoted $P0$, the coefficient of x^1 is denoted by $P1$ etc. and these are regarded as the parameters of the distribution (rather than ξ, λ, λ_{11} etc. ...) .

CHAPTER 5.

MODELS

INTRODUCTION

The exploratory analysis which we carried out established a number of facts concerning the behaviour of the time series $m0$ and $m2$. Both series vary seasonally and exhibit a short term persistence which can be described by an $AR(1)$ process. We also established that the residual process is unusual and somewhat difficult to model. Aside from the fact that the variance of the residuals varies seasonally, their distribution is asymmetric. By transforming the original series using logs we obtain approximately symmetric residuals but these are not normally distributed and consequently one cannot apply the standard models. Recall that since the signal-to-noise ratio of the time series with which we are dealing is relatively low it is important to model the process of residuals accurately in order to adequately represent the properties of the original series.

Several different models had to be fitted to the data in the attempt to accurately model the residual process. This chapter gives details relating to the four which came closest to providing an adequate fit. However none of these can be considered ideal since no one model is clearly superior to the others in terms of all the pertinent criteria, for example the preservation of seasonal behaviour, the relative likelihood properties, the relative frequency and size of extremes and so on. In the end we had to weigh up the relative importance of these criteria in order to decide which model to recommend.

The first two models we describe in this chapter are based on the Double Exponential (or Laplace) distribution and the second two make use the Elphinstone(1985) method described in the previous chapter. (Details of the Double Exponential distribution are given in the Appendix C). The model descriptions include the derivation of the likelihood equations, the parameter estimates and maximum likelihood value for the various distributions. For the Elphinstone (1985) models the issue of choosing the appropriate index k is also considered. The μ_t and σ_t terms vary seasonally and the times that these curves reach their minimum and maximum values are also given.

The likelihood equations are derived to enable us to estimate the parameter values. We can either maximize the log likelihood equation or minimize the negative log liklihood equation with respect to these parameters. Even in the Normal case exact likelihood estimates are not easy to obtain and it is for this reason that conditional likelihood estimates are used in this study. We condition on the first observation and since there are 7880 observations the difference between exact and conditional estimates is negligible, (see Box Jenkins, 1970).

All parameter estimation was performed using NAG library routines making use of Newton-Raphson based iterative methods. These programs required considerable computing effort on a main frame computer. To avoid using excessive computer time the estimates were taken as "final" when the iterations yielded negligible improvement in the maximum likelihood values.

The likelihood models are based on all available observations. It would have been more conveient to leave out some observations or alternatively to 'patch' some of the missing values. However the first of these two adjustments would have wasted some of the data while the second would have introduced a bias. After estimating the various parameters of these models we then generate the artificial $Hm0$ values (described in next chapter) and compare the results from each of the models to the observed data.

The first distribution that we used to model the residual process was the Double Exponential, $DE(\lambda)$, (Model A) since this seemed to be the shape of the residual process distribution. However the generated data from this model yielded an $Hm0$ variance that was slightly too large (see Chapter 7) and we then tried to model the residual process using $DE(\lambda_t)$, Model B, where we model the variance within the residual distribution but had very similar results with the generated data. The next approach was to make use of the method developed by Elphinstone (Chapter 3) since it was thought that we needed a better approximation to the distribution of the residual terms. We began by using a $N(0;1)$ target density (Model D) since this model was well developed in Elphinstone (1985). This model generated data that seemed adequate but the error distribution was not particulary well approximated. An alternative target density, $DE(1)$, was then attempted (Model C). This provided a good fit but produced unsatisfactory generated values. The details

of the comparison of the generated data of the four models are given in Chapter 7.

5.1 MODEL A

MODEL A: Double Exponential error distribution.

The distribution of the standardized residual is given by

$$f(e_t) = \frac{1}{2\lambda}\exp -(|e_t|/\lambda)$$

where

$$e_t = \left[\left(\frac{X_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right]$$

is the standardized residual (or error term) where we use X_t to denote the values of the time series, that is X_t is used to represent $\ln(m0)$ (or $\ln(m2)$), and

where

$$\mu_t = \alpha_0 + \alpha_1\cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2\sin\left(\frac{2\pi}{1460} * t\right)$$

and

$$\sigma_t = \beta_0 + \beta_1\cos\left(\frac{2\pi}{1460} * t\right) + \beta_2\sin\left(\frac{2\pi}{1460} * t\right)$$

We require the joint likelihood of the X_t process and therefore need to make a transformation where the Jacobian of the transformation is given by:

$$\frac{de_t}{dX_t} = \frac{1}{\sigma_t}$$

and therefore

$$|J| = |\frac{1}{\sigma_t}|$$

and we then have

$$f(X_t/X_{t-1}, \underline{\alpha}, \underline{\beta}, \phi, \lambda) = \frac{1}{2\lambda}\exp\left(-|\left[\left(\frac{X_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right]|/\lambda\right)\frac{1}{\sigma_t}$$

The conditional likelihood ($L(.)$) is then given by, where we condition on the first observation:

$$L(\underline{\alpha},\underline{\beta},\phi,\lambda;X_2,\ldots,X_n/X_1)=\prod_{t=2}^{n} f(X_t/X_{t-1},\underline{\alpha},\underline{\beta},\phi,\lambda)$$

$$=\left(\frac{1}{2\lambda}\right)^{n-1}\exp-\left(\sum_{t=2}^{n}\left(\left|\left[\left(\frac{X_t-\mu_t}{\sigma_t}\right)-\phi\left(\frac{X_{t-1}-\mu_{t-1}}{\sigma_{t-1}}\right)\right]\right|/\lambda\right)\right)*\prod_{t=2}^{n}\left(\frac{1}{\sigma_t}\right)$$

The conditional negative log likelihood (- ln $L(.)$) is then given by:

$$-\ln L(.)=(n-1)\ln 2\lambda+\left(\sum_{t=2}^{n}\left(\left|\left[\left(\frac{X_t-\mu_t}{\sigma_t}\right)-\phi\left(\frac{X_{t-1}-\mu_{t-1}}{\sigma_{t-1}}\right)\right]\right|/\lambda\right)\right)+\sum_{t=2}^{n}\ln(\sigma_t)$$

The negative log likelihood was minimized using Newton-Raphson based NAG library routines and the following results were obtained:

TABLE 5.1.1

RESULTS OF PARAMETER ESTIMATION PROCEDURES.

	ln($m0$)	ln($m2$)
Negative Log Likelihood:	3507.56	3727.92
Parameters:		
$\hat{\alpha}_0$	-1.1474	-4.9096
$\hat{\alpha}_1$	0.0268	-0.0888
$\hat{\alpha}_2$	0.1727	0.1587
$\hat{\beta}_0$	0.3479	0.3344
$\hat{\beta}_1$	0.0020	0.0168
$\hat{\beta}_2$	0.0010	-0.0135
$\hat{\lambda}$	0.8251	0.8849
$\hat{\phi}$	0.8856	0.8741

So the probability density function of X_t , given X_{t-1} , is then given by:

$$\hat{f}(X_t/X_{t-1},\underline{\hat{\alpha}},\underline{\hat{\beta}},\hat{\phi},\hat{\lambda}) = \frac{1}{2\hat{\lambda}}\exp\left(-|\left[\left(\frac{X_t-\hat{\mu}_t}{\hat{\sigma}_t}\right)-\phi\left(\frac{X_{t-1}-\hat{\mu}_{t-1}}{\hat{\sigma}_{t-1}}\right)\right]|/\hat{\lambda}\right)\frac{1}{\hat{\sigma}_t}$$

where

$$\hat{\mu}_t = \hat{\alpha}_0 + \hat{\alpha}_1\cos\left(\frac{2\pi}{1460} * t\right) + \hat{\alpha}_2\sin\left(\frac{2\pi}{1460} * t\right)$$

and

$$\hat{\sigma}_t = \hat{\beta}_0 + \hat{\beta}_1\cos\left(\frac{2\pi}{1460} * t\right) + \hat{\beta}_2\sin\left(\frac{2\pi}{1460} * t\right)$$

Interpretation of Model A.

The Mean function.

The earlier parameterization, being linear in the parameters, is convenient for the purposes of estimation, but for the purposes of interpretation it is more convenient to write this in terms of a phase-amplitude representation:

We have a mean function of the following form:

$$\hat{\mu}_t = \hat{\alpha}_0 + \hat{\alpha}_1\cos\omega t + \hat{\alpha}_2\sin\omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

The formula relating $(\hat{\alpha}_1,\hat{\alpha}_2)$ to (R,ψ), the amplitude and phase, are given in Appendix D. For the $\ln(m0)$ series we have:

$$\hat{\alpha}_0 = -1.1474$$

$$\hat{\alpha}_1 = 0.0268$$

$$\hat{\alpha}_2 = 0.1727$$

We note that the resulting estimates of R and ψ, namely $\hat{R} = 0.1747$ and $\hat{\psi} = -1.416$ constitute (conditional) maximum likelihood estimates of the corresponding parameters.

This has the following interpretation: (see Figure 5.1.A):

The maximum value for the mean curve is $\alpha_0 + R = -0.9726$ which occurs on the 21^{st} of August. The minimum value for the mean curve is $\alpha_0 - R = -1.3221$ which occurs on the 19^{th} of February.

For the $\ln(m2)$ series we have:

$$\widehat{\alpha}_0 = -4.9096$$

$$\widehat{\alpha}_1 = -0.0888$$

$$\widehat{\alpha}_2 = 0.1587$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.1818$ and $\psi = -2.0809$.

This has the following interpretation: (see Figure 5.1.B):

The maximum value for the mean curve is $\alpha_0 + R = -4.7278$ which occurs on the 28^{th} of September. The minimum value for the mean curve is $\alpha_0 - R = -5.0914$ which occurs on the 30^{th} of March.

Variance function.

We have a variance function of the following form:

$$\widehat{\sigma}_t = \widehat{\beta}_0 + \widehat{\beta}_1 \cos \omega t + \widehat{\beta}_2 \sin \omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

For the $\ln(m0)$ series we have:

$$\widehat{\beta}_0 = 0.3479$$

$$\widehat{\beta}_1 = 0.0020$$

$$\widehat{\beta}_2 = 0.0010$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.0022$ and $\psi = -0.4636$.

This has the following interpretation: (see Figure 5.1.C):

The maximum value for the variance curve is $\beta_0 + R = 0.3501$ which occurs on the 26^{th} of June. The minimum value for the variance curve is $\beta_0 - R = 0.3457$ which occurs on the 26^{th} of December.

For the $\ln(m2)$ series we have:

$$\widehat{\beta}_0 = 0.3344$$

$$\widehat{\beta}_1 = 0.0168$$

$$\widehat{\beta}_2 = -0.0135$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.0215$ and $\psi = 0.6769$.

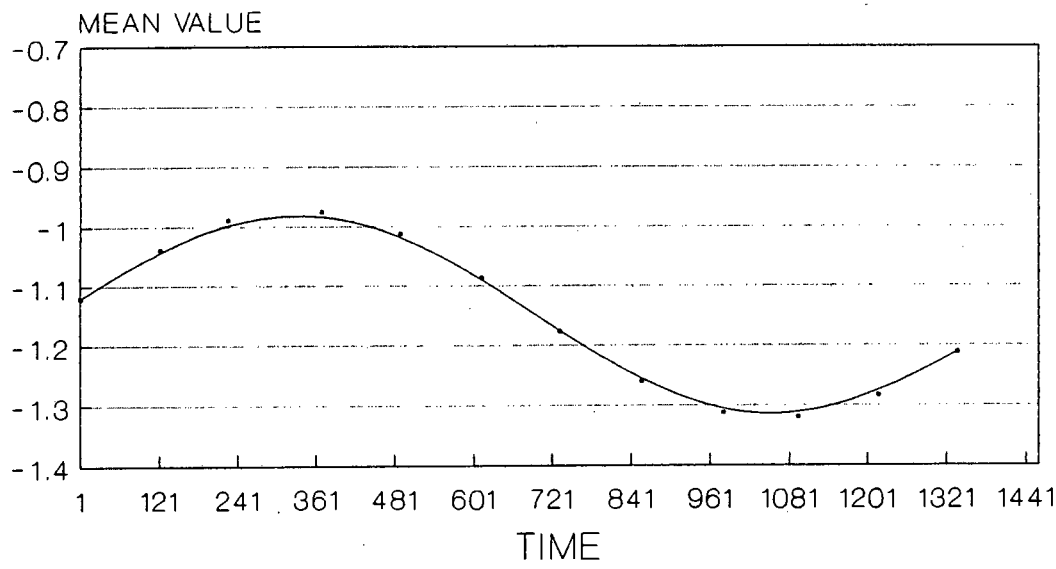
This has the following interpretation: (see Figure 5.1.D):

The maximum value for the variance curve is $\beta_0 + R = 0.3559$ which occurs on the 21^{st} of April. The minimum value for the variance curve is $\beta_0 - R = 0.3129$ which occurs on the 21^{st} of October.

Error Distribution

Finally in this section the plots are given of the comparison between the observed error terms when using this model and the distribution used to model them. These are given in Figures 5.1.E & 5.1.F for the $\ln m0$ series and in Figures 5.1.G & 5.1.H for the $\ln m2$ series.

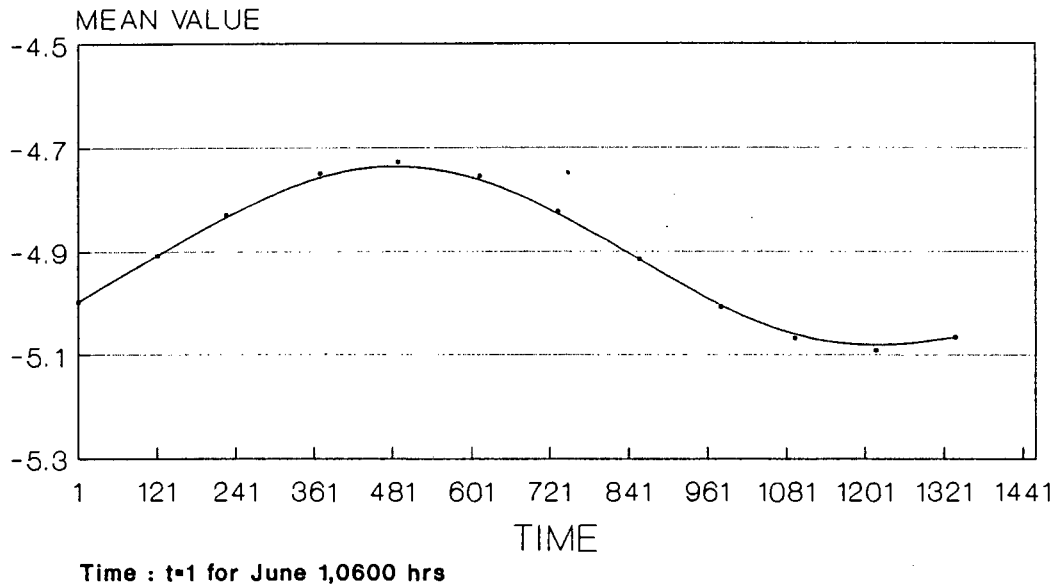
MODEL A (LN (M0))
PERIODICITY OF MEAN CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.1.A

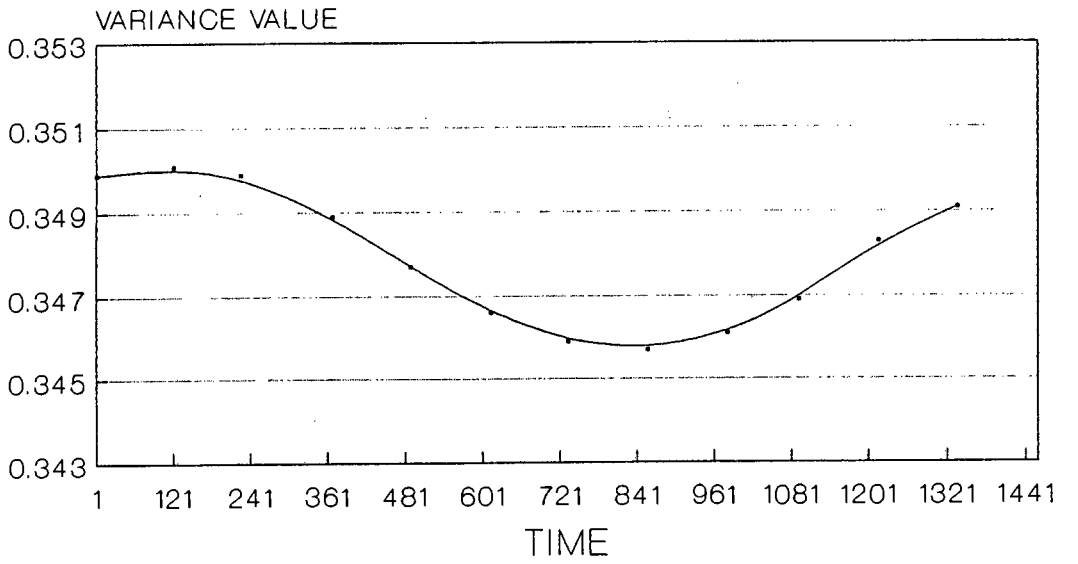
MODEL A (LN (M2))
PERIODICITY OF MEAN CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.1.B

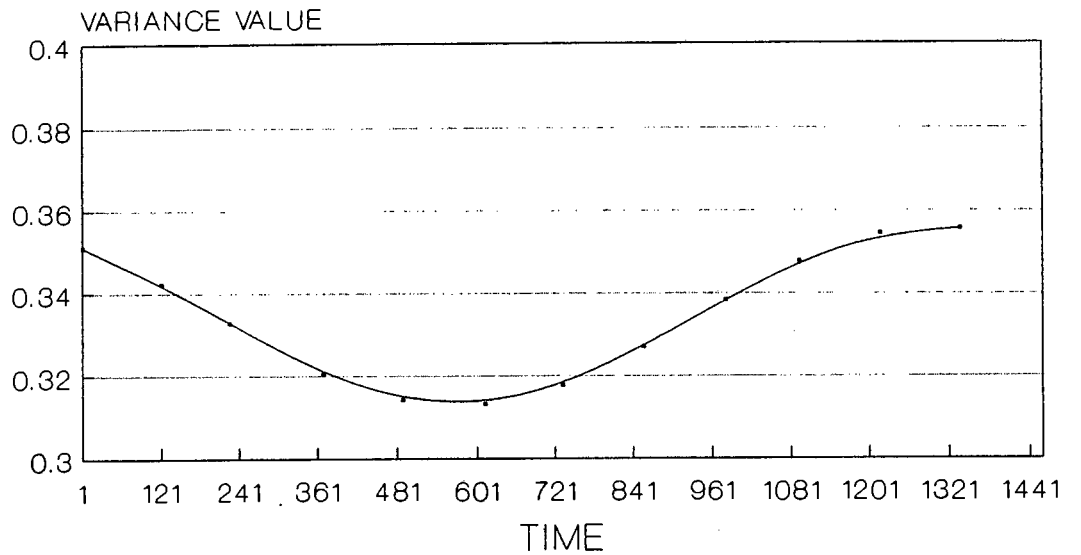
MODEL A (LN (M0))
PERIODICITY OF VARIANCE CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.1.C

MODEL A (LN (M2))
PERIODICITY OF VARIANCE CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.1.D

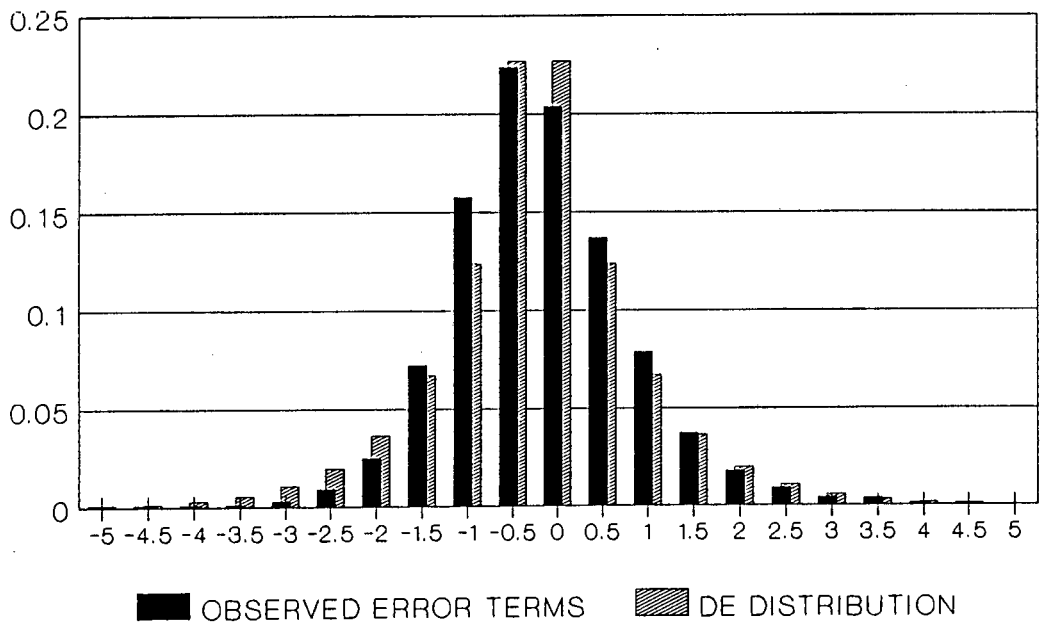


FIGURE 5.1.E

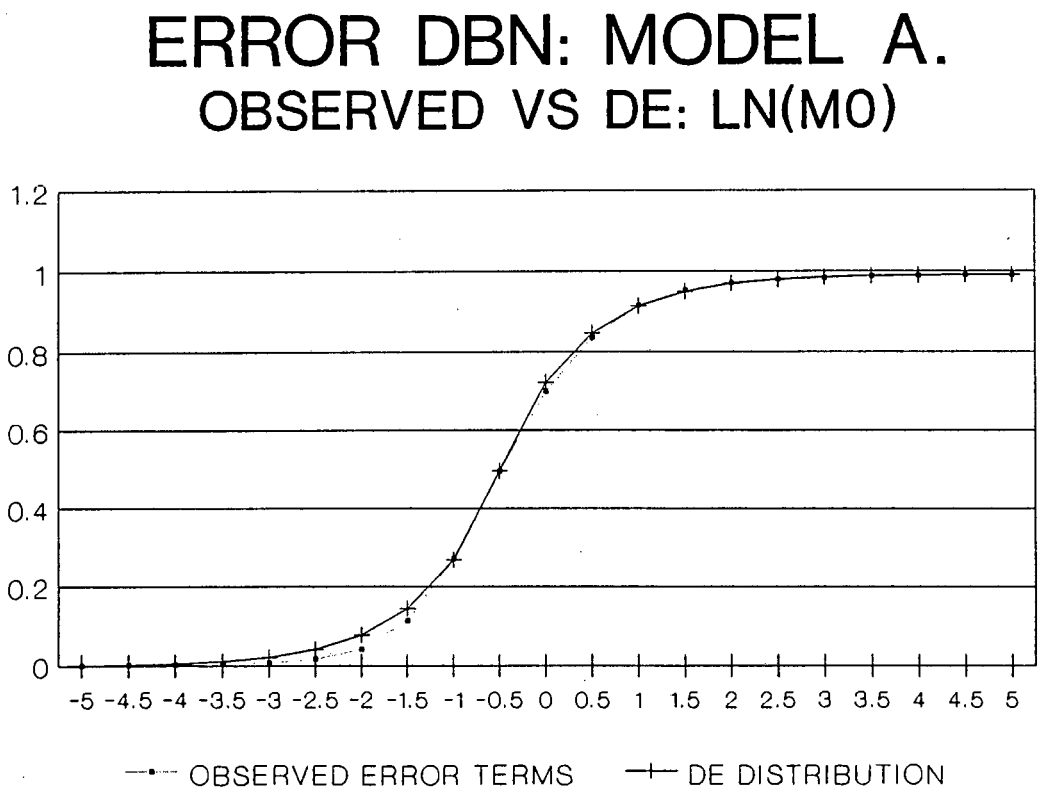


FIGURE 5.1.F

ERROR DBN: MODEL A.
OBSERVED VS DE: LN(M2)

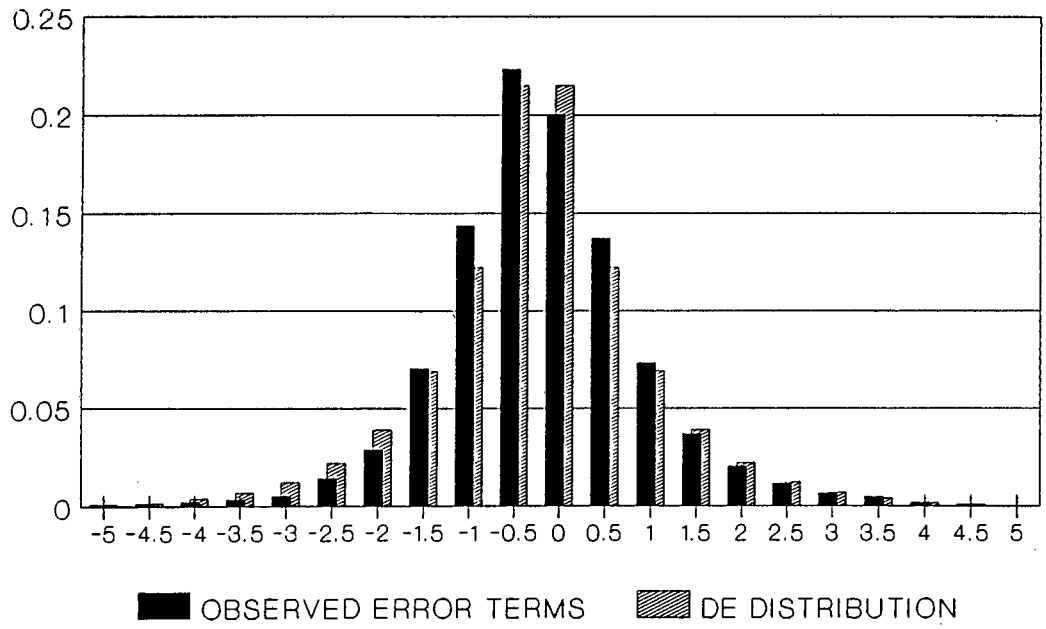


FIGURE 5.1.G

ERROR DBN: MODEL A.
OBSERVED VS DE: LN(M2)

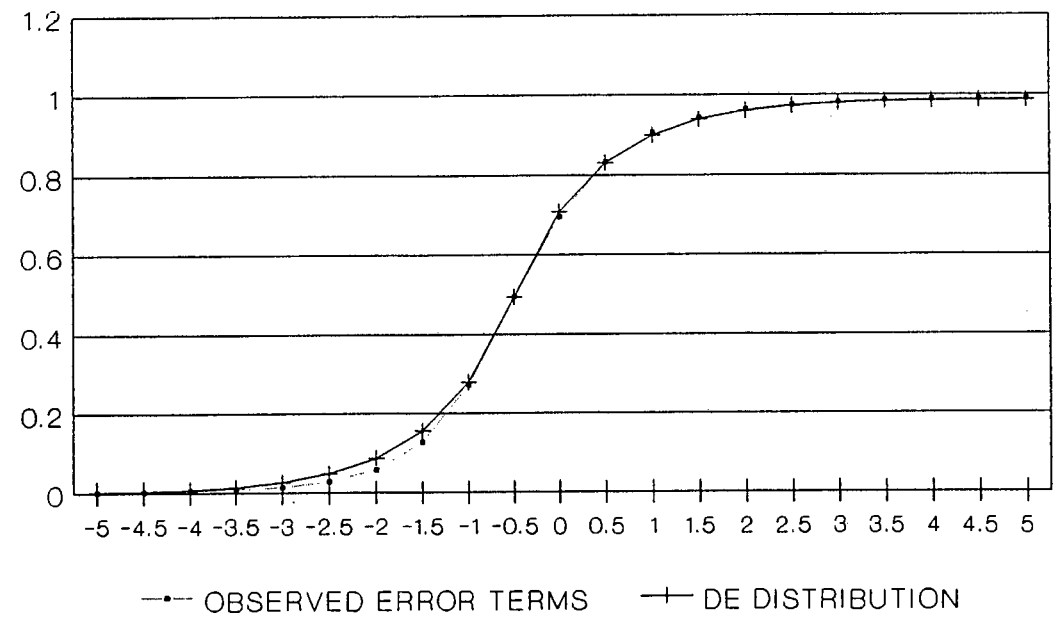


FIGURE 5.1.H

5.2 MODEL B

MODEL B: Model the error distribution using the Double Exponential distribution but include the seasonall variation in the λ term and not in the σ_t term as before.

The distribution of the standardized residuals is given by:

$$f(e_t) = \frac{1}{2\lambda_t}\exp -(|e_t|/\lambda_t)$$

where the standardized residual is given by

$$e_t = [(X_t - \mu_t) - \phi(X_{t-1} - \mu_{t-1})]$$

where we use X_t to denote the values of the time series, that is X_t is used to represent $\ln(m0)$ (or $\ln(m2)$) and where

$$\mu_t = \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

and

$$\lambda_t = \lambda_0 + \lambda_1 \cos\left(\frac{2\pi}{1460} * t\right) + \lambda_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

We require the joint likelihood of the X_t process and therefore need to make a transformation where the Jacobian of the transformation is given by:

$$\frac{de_t}{dX_t} = 1$$

therefore

$$|J| = |1|$$

then

$$f(X_t/X_{t-1}, \underline{\alpha}, \underline{\lambda}, \phi) = \frac{1}{2\lambda_t}\exp\left(-|[(X_t - \mu_t) - \phi(X_{t-1} - \mu_{t-1})]|/\lambda_t\right)$$

The conditional likelihood ($L(.)$) is given by, where we condition on the first observation:

$$L(\underline{\alpha}, \underline{\lambda}, \phi; X_2, \ldots, X_n / X_1) = \prod_{t=2}^{n} f(X_t / X_{t-1}, \underline{\alpha}, \underline{\lambda}, \phi)$$

$$= \left(\prod_{t=2}^{n} \left(\frac{1}{2\lambda_t} \right) \right) \exp \left(- \left(\sum_{t=2}^{n} (|[(X_t - \mu_t) - \phi(X_{t-1} - \mu_{t-1})]|/\lambda_t) \right) \right)$$

The conditional negative log likelihood, (- ln $L(.)$), is then given by:

$$- \ln L(.) = \sum_{t=2}^{n} \ln (2\lambda_t) + \left(\sum_{t=2}^{n} (|[(X_t - \mu_t) - \phi(X_{t-1} - \mu_{t-1})]|) \right)$$

The negative log likelihood was minimized using Newton-Raphson based NAG library routines and the following results were obtained:

TABLE 5.2.1

RESULTS OF PARAMETER ESTIMATION PROCEDURES.

	ln($m0$)	ln($m2$)
Negative Log Likelihood:	3507.57	3727.93
Parameters:		
$\hat{\alpha}_0$	-1.1486	-4.9093
$\hat{\alpha}_1$	0.0245	-0.0886
$\hat{\alpha}_2$	0.1728	0.1583
$\hat{\lambda}_0$	0.2871	0.2953
$\hat{\lambda}_1$	0.0016	0.0153
$\hat{\lambda}_2$	0.0006	-0.0126
$\hat{\phi}$	0.8858	0.8738

So the probability density function of X_t , given X_{t-1} , is given by:

$$\hat{f}(X_t / X_{t-1}, \hat{\underline{\alpha}}, \hat{\underline{\lambda}}, \hat{\phi}) = \frac{1}{2\hat{\lambda}_t} \exp \left(-| \left[(X_t - \hat{\mu}_t) - \hat{\phi}(X_{t-1} - \hat{\mu}_{t-1}) \right] | / \hat{\lambda}_t \right)$$

where

$$\hat{\mu}_t = \hat{\alpha}_0 + \hat{\alpha}_1 \cos \left(\frac{2\pi}{1460} * t \right) + \hat{\alpha}_2 \sin \left(\frac{2\pi}{1460} * t \right)$$

and

$$\widehat{\lambda}_t = \widehat{\lambda}_0 + \widehat{\lambda}_1 \cos\left(\frac{2\pi}{1460} * t\right) + \widehat{\lambda}_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

Interpretation of Model B.

The Mean function.

We have a mean function of the following form:

$$\widehat{\mu}_t = \widehat{\alpha}_0 + \widehat{\alpha}_1 \cos\omega t + \widehat{\alpha}_2 \sin\omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

For the $\ln(m0)$ series we have:

$$\widehat{\alpha}_0 = -1.1486$$

$$\widehat{\alpha}_1 = 0.0245$$

$$\widehat{\alpha}_2 = 0.1728$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.1745$ and $\psi = -1.429$.

This has the following interpretation: (see Figure 5.2.A):

The maximum value for the mean curve is $\alpha_0 + R = -0.9741$ which occurs on the 21^{st} of August. The minimum value for the mean curve is $\alpha_0 - R = -1.3223$ which occurs on the 20^{th} of February.

For the $\ln(m2)$ series we have:

$$\widehat{\alpha}_0 = -4.9093$$

$$\widehat{\alpha}_1 = -0.0886$$

$$\widehat{\alpha}_2 = 0.1583$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.1814$ and $\psi = -2.081$.

This has the following interpretation: (see Figure 5.2.B):

The maximum value for the mean curve is $\alpha_0 + R = -4.7270$ which occurs on the 28^{th} of September. The minimum value for the mean curve is $\alpha_0 - R = -5.0907$ which occurs on the 30^{th} of March.

The Variance function.

We have a lambda function of the following form:

$$\widehat{\lambda}_t = \widehat{\lambda}_0 + \widehat{\lambda}_1 \cos \omega t + \widehat{\lambda}_2 \sin \omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

For the $\ln(m0)$ series we have:

$$\widehat{\lambda}_0 = 0.2871$$

$$\widehat{\lambda}_1 = 0.0016$$

$$\widehat{\lambda}_2 = 0.0006$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.0017$ and $\psi = -0.3587$.

This has the following interpretation: (see Figure 5.2.C):

The maximum value for the λ curve is $\lambda_0 + R = 0.2888$ which occurs on the 20^{th} of June. The minimum value for the λ curve is $\lambda_0 - R = 0.2854$ which occurs on the 20^{th} of December.

For the $\ln(m2)$ series we have:

$$\widehat{\lambda}_0 = 0.2953$$

$$\widehat{\lambda}_1 = 0.0153$$

$$\widehat{\lambda}_2 = -0.0126$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.0198$ and $\psi = 0.6889$.

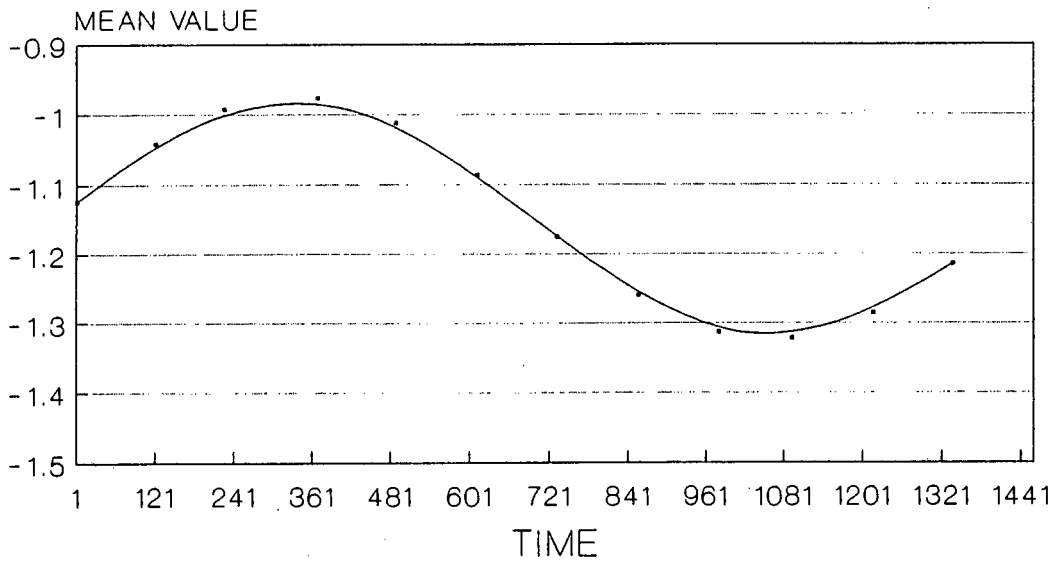
This has the following interpretation: (see Figure 5.2.D):

The maximum value for the λ curve is $\lambda_0 + R = 0.3151$ which occurs on the 20^{th} of April. The minimum value for the λ curve is $\lambda_0 - R = 0.2754$ which occurs on the 20^{th} of October.

Error Distribution

Finally in this section the plots are given of the comparison between the observed error terms when using this model and the distribution used to model them. These are given in Figures 5.2.E & 5.2.F for the $\ln m0$ series and in Figures 5.2.G & 5.2.H for the $\ln m2$ series.

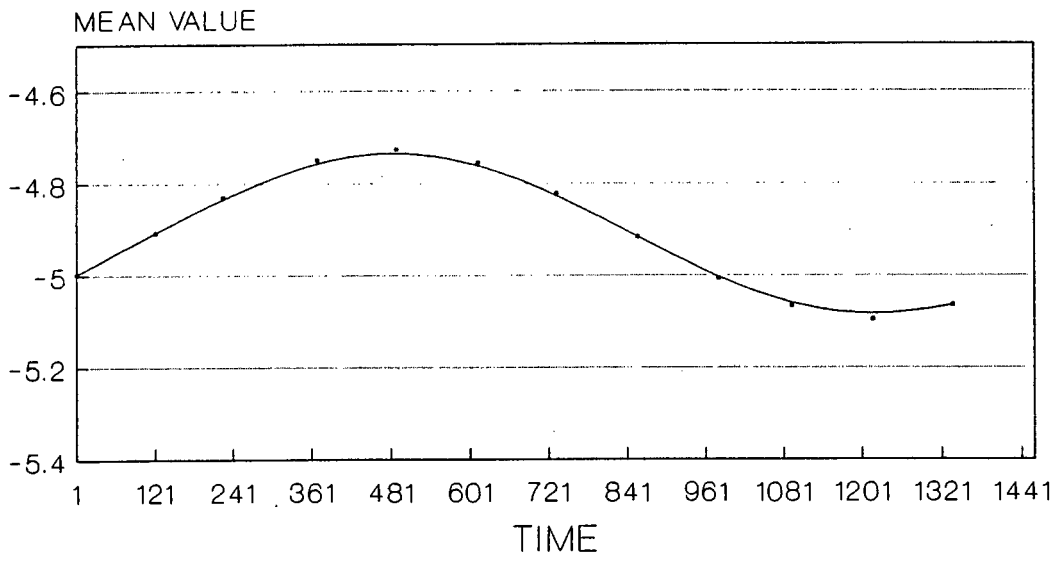
MODEL B (LN (M0))
PERIODICITY OF MEAN CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.2.A

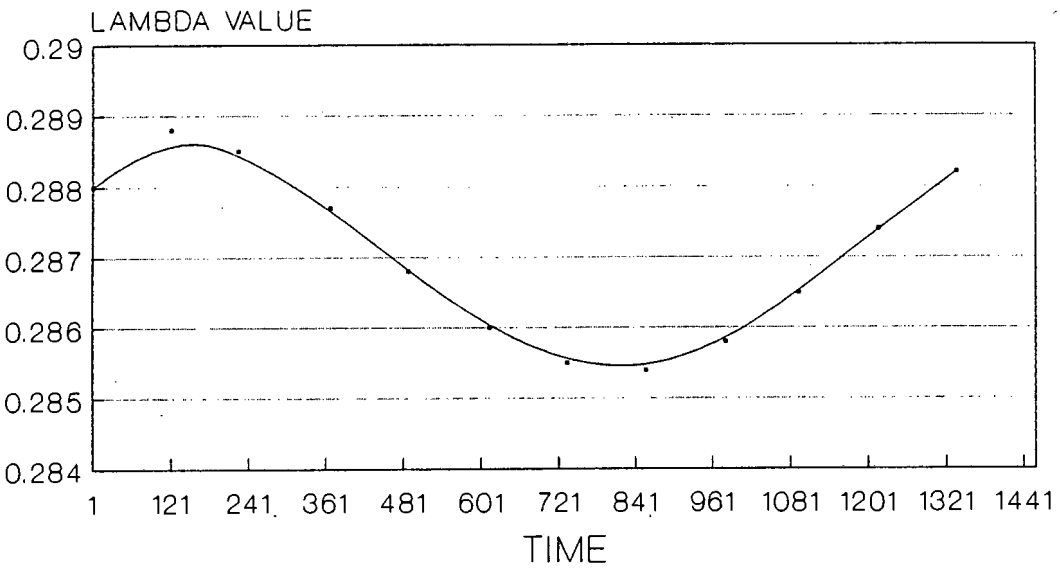
MODEL B (LN (M2))
PERIODICITY OF MEAN CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.2.B

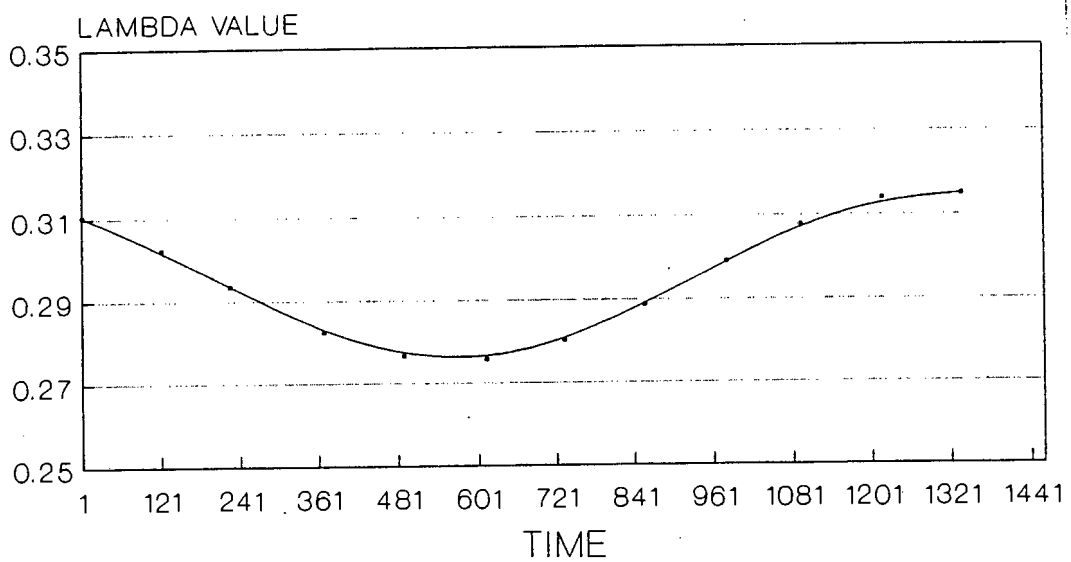
MODEL B (LN (M0))
PERIODICITY OF LAMBDA CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.2.C

MODEL B (LN (M2))
PERIODICITY OF LAMBDA CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.2.D

ERROR DBN: MODEL B.
OBSERVED VS DE: LN(M0)

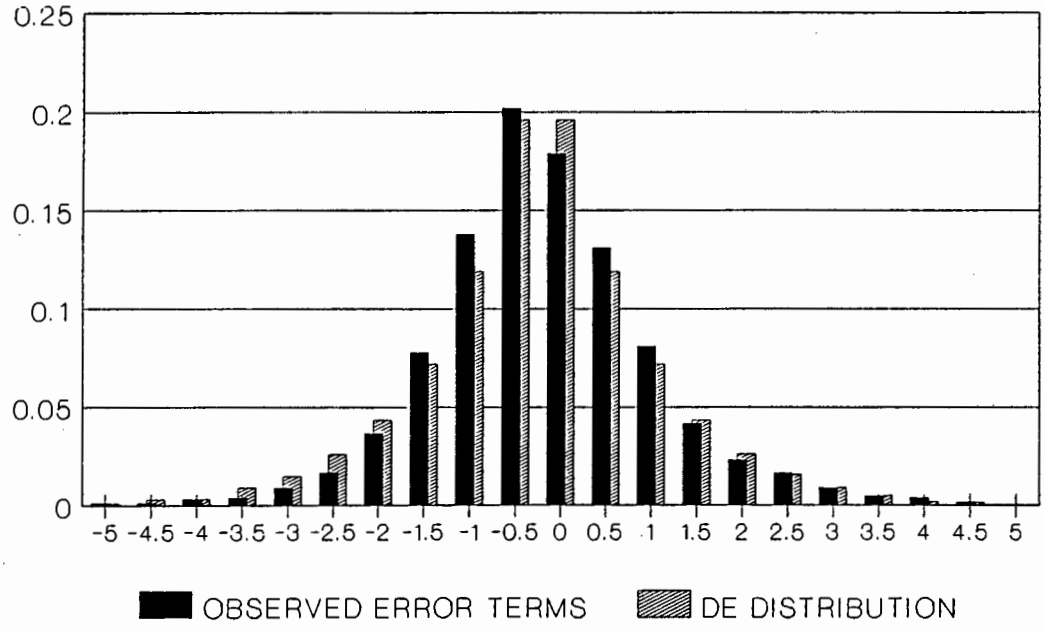


FIGURE 5.2.E

ERROR DBN: MODEL B.
OBSERVED VS DE: LN(M0)

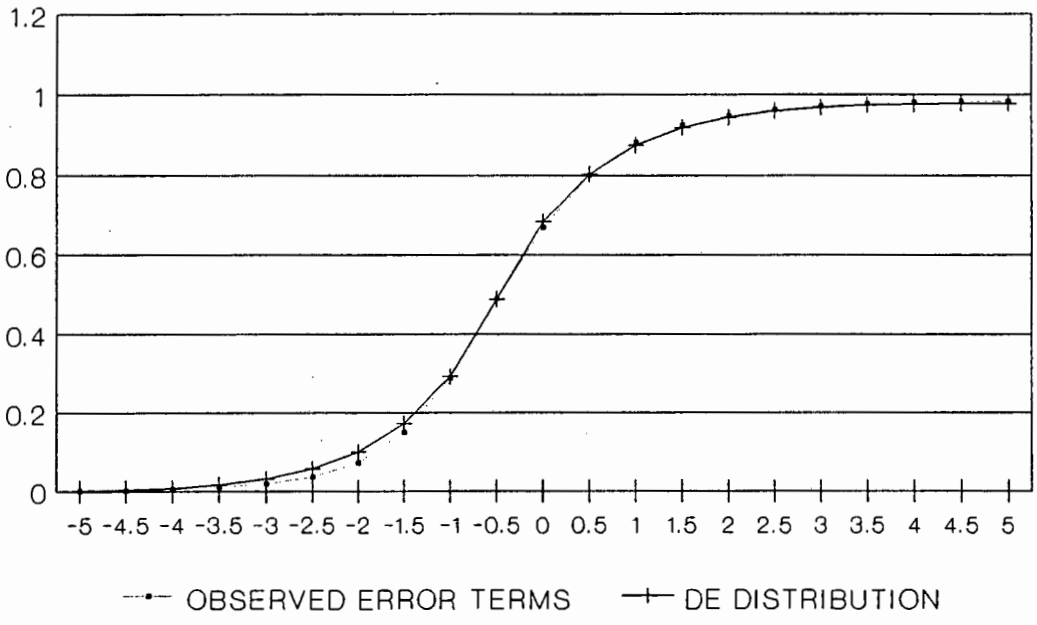


FIGURE 5.2.F

ERROR DBN: MODEL B.
OBSERVED VS DE: LN(M2)

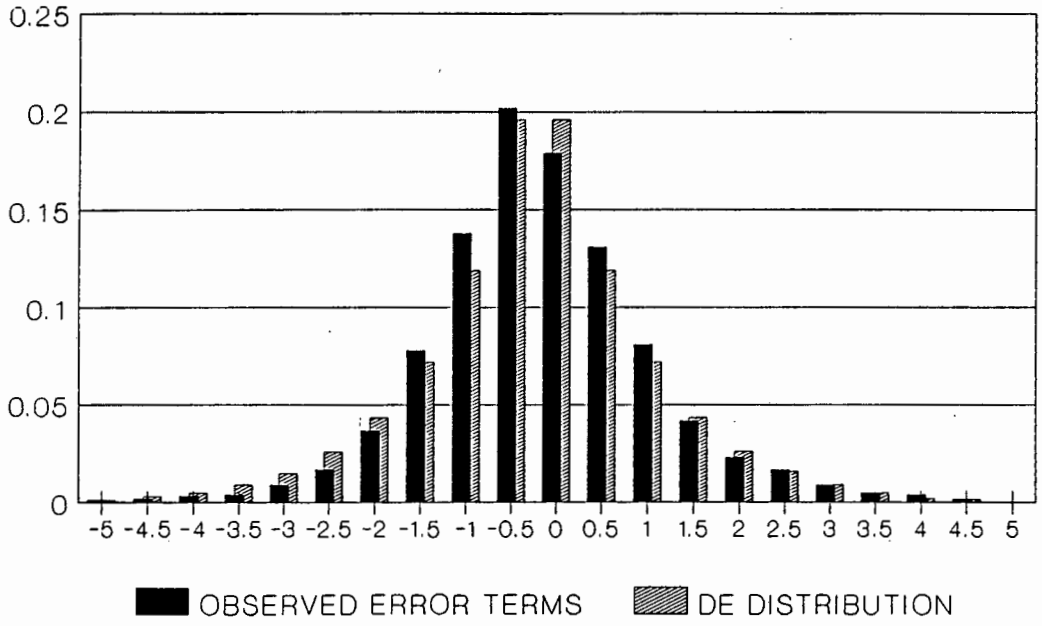


FIGURE 5.2.G

ERROR DBN. MODEL B.
OBSERVED VS DE: LN(M2)

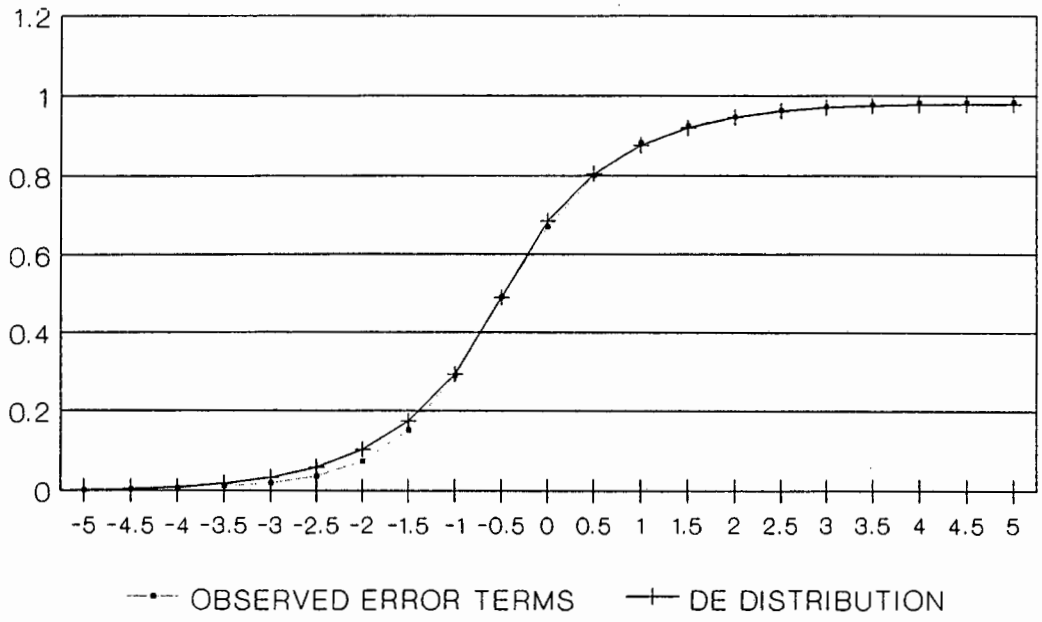


FIGURE 5.2.H

5.3 MODEL C

MODEL C: Model the error distribution using the Elphinstone (1985) method with a unit Double Exponential target distribution. Refer to Chapter 4 for further details of this method.

The distribution of the standardized residuals is given by:

$$f(e_t) = \frac{1}{2}\exp -(|\xi + \theta e_t|)\theta$$

(for $k = 0$ using Elphinstone notation).

where

$$e_t = \left[\left(\frac{X_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right]$$

is the standardized residual and where we use X_t to denote the values of the time series, that is X_t is used to represent $\ln(m0)$ (or $\ln(m2)$), and where

$$\mu_t = \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

and

$$\sigma_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi}{1460} * t\right) + \beta_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

We require the joint likelihood of the X_t process and therefore need to make a transformation where the Jacobian of the transformation is given by:

$$\frac{de_t}{dX_t} = \frac{1}{\sigma_t}$$

and therefore

$$|J| = |\frac{1}{\sigma_t}|$$

and we then have

$$f(X_t/X_{t-1}, \underline{\alpha}, \underline{\beta}, \phi, \xi, \theta) = \frac{1}{2}\exp\left(-|\xi + \theta\left[\left(\frac{X_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right]|\right)\theta\frac{1}{\sigma_t}$$

and call this distribution $ELPHDE_{k=0}$

The conditional likelihood ($L(.)$) is then given by, where we condition on the first observation:

$$L(\underline{\alpha},\underline{\beta},\phi,\lambda;X_2,\ldots,X_n/X_1)=\prod_{t=2}^{n} f(X_t/X_{t-1},\underline{\alpha},\underline{\beta},\phi,\xi,\theta)$$

$$=\left(\frac{1}{2}\right)^{n-1}\exp-\sum_{t=2}^{n}\left(|\xi+\theta\left[\left(\frac{X_t-\mu_t}{\sigma_t}\right)-\phi\left(\frac{X_{t-1}-\mu_{t-1}}{\sigma_{t-1}}\right)\right]|\right)\theta^{n-1}\prod_{t=2}^{n}\left(\frac{1}{\sigma_t}\right)$$

The conditional negative log likelihood (- $\ln L(.)$) is then given by:

$$-\ln L(.)=(n-1)\ln 2+\sum_{t=2}^{n}\left(|\xi+\theta\left[\left(\frac{X_t-\mu_t}{\sigma_t}\right)-\phi\left(\frac{X_{t-1}-\mu_{t-1}}{\sigma_{t-1}}\right)\right]|\right)$$
$$-(n-1)\ln\theta+\sum_{t=2}^{n}\ln(\sigma_t)$$

The above equations are where the Elphinstone index $k=0$. The equations are calculated for $k=0,1,2,3,4$ and the resulting negative log likelihood values are given in the following table.

TABLE 5.3.1

RESULTING NEGATIVE LOG LIKELIHOOD VALUES.

	$\ln(m0)$	$\ln(m2)$
$k=0$	3508.14	3729.02
$k=1$	3506.79	3728.88
$k=2$	3501.20	3721.10
$k=3$	3501.00	3720.20
$k=4$	3500.50	3718.04

Then as detailed in the Elphinstone chapter, a decision must be made as to what value for index k is the most appropriate. As stated in that chapter we use both the AIC and SIC to give an indication as to what value k should be. The following table gives the AIC and SIC values for both the $\ln(m0)$ and $\ln(m2)$ series.

TABLE 5.3.2

	$\ln(m0)$		$\ln(m2)$	
	SIC	AIC	SIC	AIC
$k=0$	3517.11	7020.28	3737.99	7462.04
$k=1$	3524.74	7021.59	3746.82	7465.76
$k=2$	3528.11	7014.59	3747.92	7454.20
$k=3$	3536.80	7018.00	3756.09	7456.40
$k=4$	3545.36	7021.00	3762.89	7456.08

Therefore, when using the AIC, we choose $k=2$ for both the $\ln(m0)$ and $\ln(m2)$ series since $AIC(k)$ is at a minimum when $k=2$. If we base our decision on the SIC we would choose $k=0$ since $SIC(k)$ is at a minimum when $k=0$.

The following table gives the results of the parameter estimation procedures, using $k=2$ and the six parameters $P0, P1, \ldots, P5$ are those relating to the Elphinstone method.

TABLE 5.3.3

RESULTS OF PARAMETER ESTIMATION PROCEDURES.

	$\ln(m0)$	$\ln(m2)$
Negative Log Likelihood:	3501.20	3721.10
Parameters:		
$\widehat{\alpha}_0$	-1.1139	-4.8974
$\widehat{\alpha}_1$	0.0558	-0.0944
$\widehat{\alpha}_2$	0.1912	0.2265
$\widehat{\beta}_0$	0.2345	0.2542
$\widehat{\beta}_1$	0.0014	0.0120
$\widehat{\beta}_2$	0.0006	-0.0126
$\widehat{\phi}$	0.8862	0.8701
$\widehat{P0}$	0.0087	0.0092
$\widehat{P1}$	0.8181	0.8626
$\widehat{P2}$	-0.0022	-0.0013
$\widehat{P3}$	-0.0019	-0.0020
$\widehat{P4}$	0.000006	0.000002
$\widehat{P5}$	0.000002	0.000002

Therefore the probability density function of X_t , given X_{t-1} is:

$$\hat{f}(X_t/X_{t-1},\underline{\hat{\alpha}},\underline{\hat{\beta}},\hat{\phi},\widehat{P0},...,\widehat{P5}) = \frac{1}{2}\exp\left(-|\widehat{P0}+\widehat{P1}e_t+\widehat{P2}{e_t}^2+\widehat{P3}{e_t}^3+\widehat{P4}{e_t}^4+\widehat{P5}{e_t}^5|\right)$$
$$\left(\widehat{P1}+\widehat{P2}e_t+\widehat{P3}{e_t}^2+\widehat{P4}{e_t}^3+\widehat{P5}{e_t}^4\right)\hat{\theta}\frac{1}{\hat{\sigma}_t}$$

and this distribution is called $ELPHDE_{k=2}$

where

$$e_t = \left[\left(\frac{X_t-\hat{\mu}_t}{\hat{\sigma}_t}\right)-\hat{\phi}\left(\frac{X_{t-1}-\hat{\mu}_{t-1}}{\hat{\sigma}_{t-1}}\right)\right]$$

$$\hat{\mu}_t = \hat{\alpha}_0+\hat{\alpha}_1\cos\left(\frac{2\pi}{1460}*t\right)+\hat{\alpha}_2\sin\left(\frac{2\pi}{1460}*t\right)$$

and

$$\hat{\sigma}_t = \hat{\beta}_0+\hat{\beta}_1\cos\left(\frac{2\pi}{1460}*t\right)+\hat{\beta}_2\sin\left(\frac{2\pi}{1460}*t\right)$$

Interpretation of Model C

The Mean function

We have a mean function of the following form:

$$\hat{\mu}_t = \hat{\alpha}_0 + \hat{\alpha}_1\cos\omega t + \hat{\alpha}_2\sin\omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

For the $\ln(m0)$ series we have:

$$\hat{\alpha}_0 = -1.1139$$

$$\hat{\alpha}_1 = 0.0558$$

$$\hat{\alpha}_2 = 0.1912$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.1991$ and $\psi = -1.2868$.

This has the following interpretation: (see Figure 5.3.A):

The maximum value for the mean curve is $\alpha_0 + R = -0.9147$ which occurs on the 13^{th} of August. The minimum value for the mean curve is $\alpha_0 - R = -1.3130$ which occurs on the 12^{th} of February.

For the $\ln(m2)$ series we have:

$$\widehat{\alpha}_0 = -4.8974$$

$$\widehat{\alpha}_1 = -0.0944$$

$$\widehat{\alpha}_2 = 0.2265$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.2453$ and $\psi = -1.9656$.

This has the following interpretation: (see Figure 5.3.B):

The maximum value for the mean curve is $\alpha_0 + R = -4.652$ which occurs on the 21^{st} of September. The minimum value for the mean curve is $\alpha_0 - R = -5.142$ which occurs on the 23^{rd} of March.

The Variance function.

We have a variance function of the following form:

$$\widehat{\sigma}_t = \widehat{\beta}_0 + \widehat{\beta}_1 \cos\omega t + \widehat{\beta}_2 \sin\omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

For the $\ln(m0)$ series we have:

$$\widehat{\beta}_0 = 0.2345$$

$$\widehat{\beta}_1 = 0.0014$$

$$\widehat{\beta}_2 = 0.0006$$

We then obtain values for R and ψ , (the amplitude and phase), and $R = 0.0015$ and $\psi = -0.4048$.

This has the following interpretation: (see Figure 5.3.C):

The maximum value for the variance curve is $\beta_0 + R = 0.2360$ which occurs on the 23^{rd} of June. The minimum value for the variance curve is $\beta_0 - R = 0.2329$ which occurs on the 22^{nd} of December.

For the $\ln(m2)$ series we have:

$$\widehat{\beta}_0 = 0.2542$$

$$\widehat{\beta}_1 = 0.0120$$

$$\widehat{\beta}_2 = -0.0126$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.0174$ and $\psi = 0.8097$.

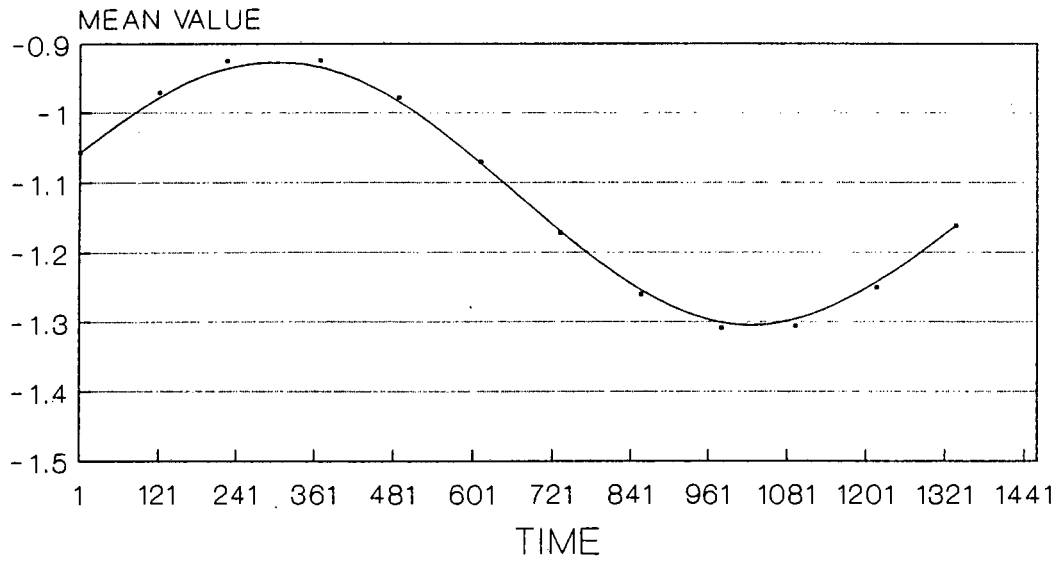
This has the following interpretation: (see Figure 5.3.D):

The maximum value for the variance curve is $\beta_0 + R = 0.2716$ which occurs on the 13^{th} of April. The minimum value for the variance curve is $\beta_0 - R = 0.2368$ which occurs on the 13^{th} of October.

Error Distribution

Finally in this section the plots are given of the comparison between the observed error terms when using this model and the distribution used to model them. These are given in Figures 5.3.E & 5.3.F for the $\ln m0$ series and in Figures 5.3.G & 5.3.H for the $\ln m2$ series.

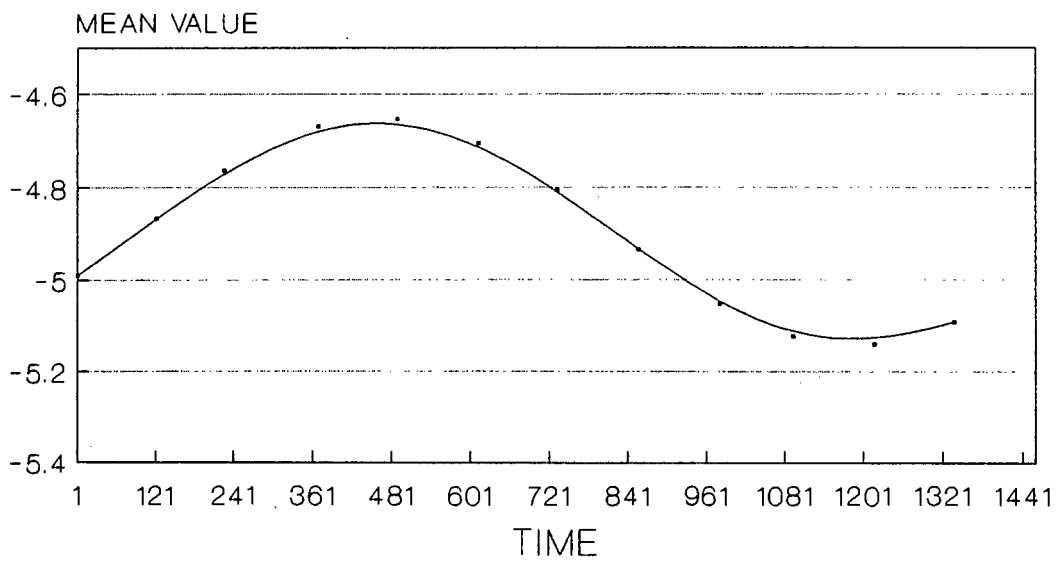
MODEL C (LN (M0))
PERIODICITY OF MEAN CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.3.A

MODEL C (LN (M2))
PERIODICITY OF MEAN CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.3.B

MODEL C (LN (M0))
PERIODICITY OF VARIANCE CURVE

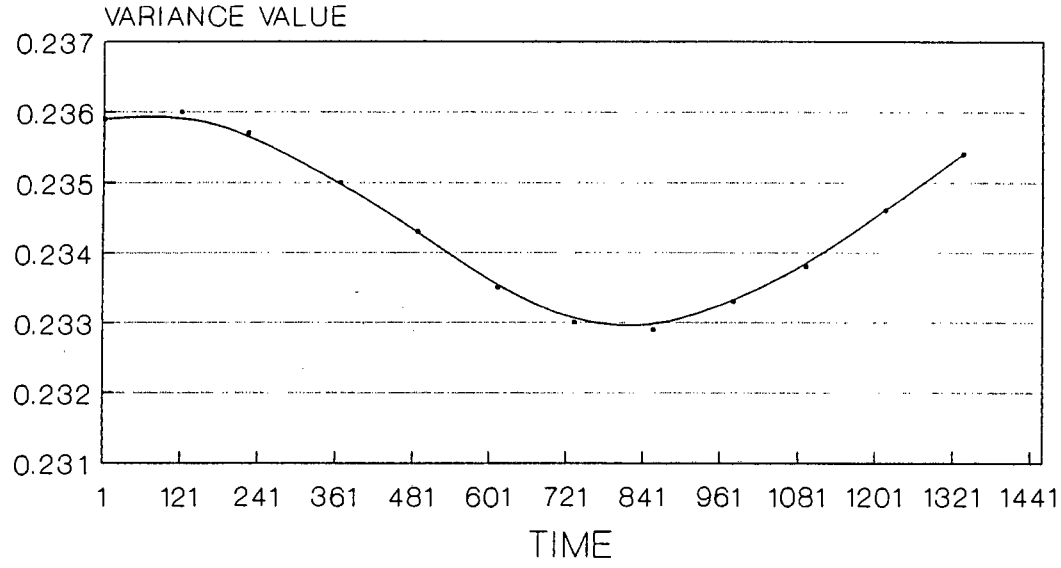


FIGURE 5.3.C

MODEL C (LN (M2))
PERIODICITY OF VARIANCE CURVE

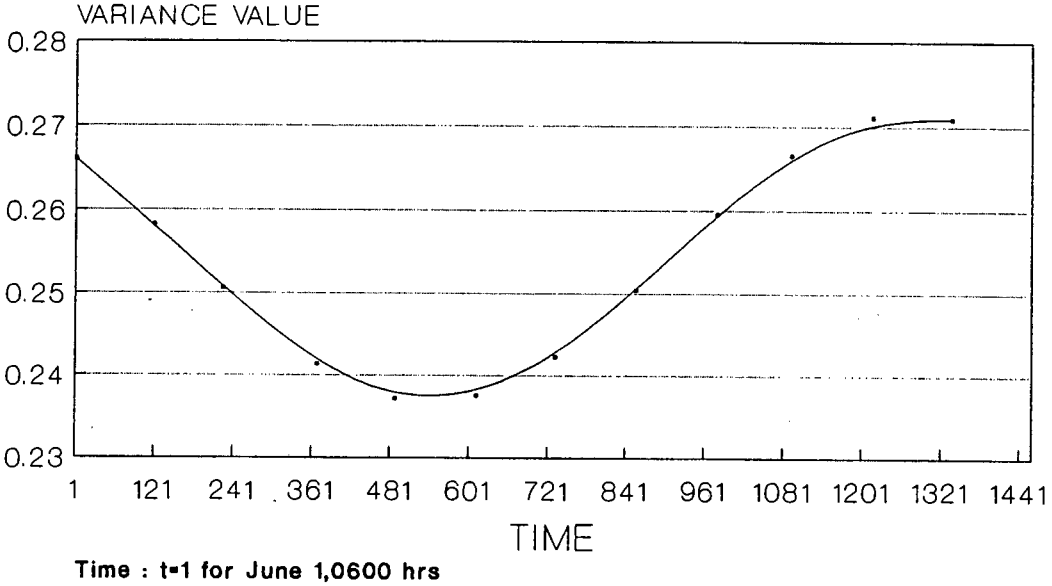


FIGURE 5.3.D

ERROR DBN: MODEL C.
OBSERVED VS ELPHDE(k=2): LN(M0)

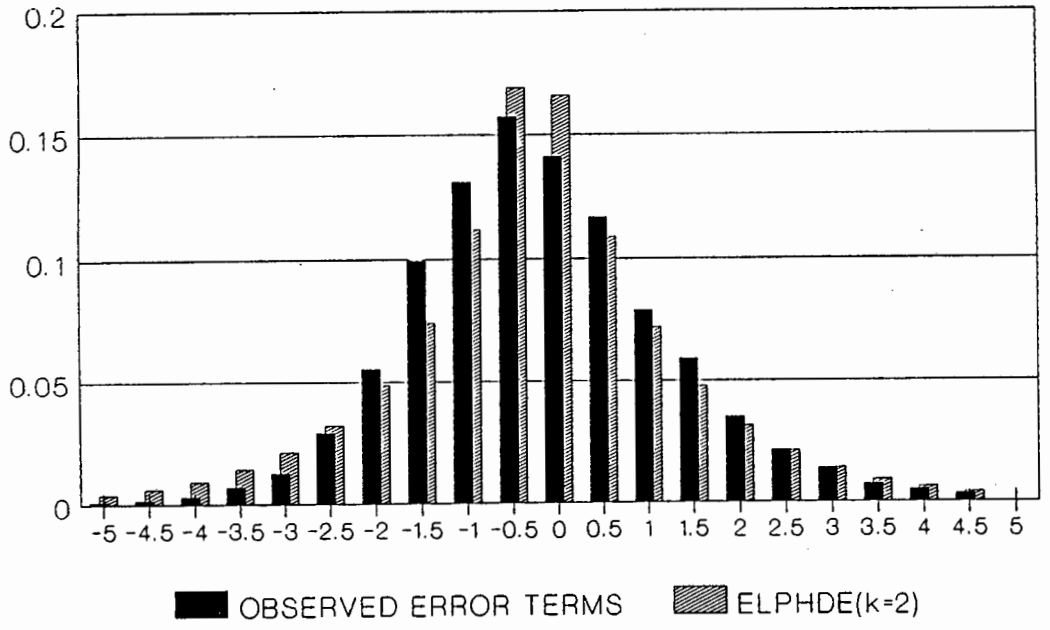


FIGURE 5.3.E

ERROR DBN: MODEL C.
OBSERVED VS ELPHDE(k=2): LN(M0)

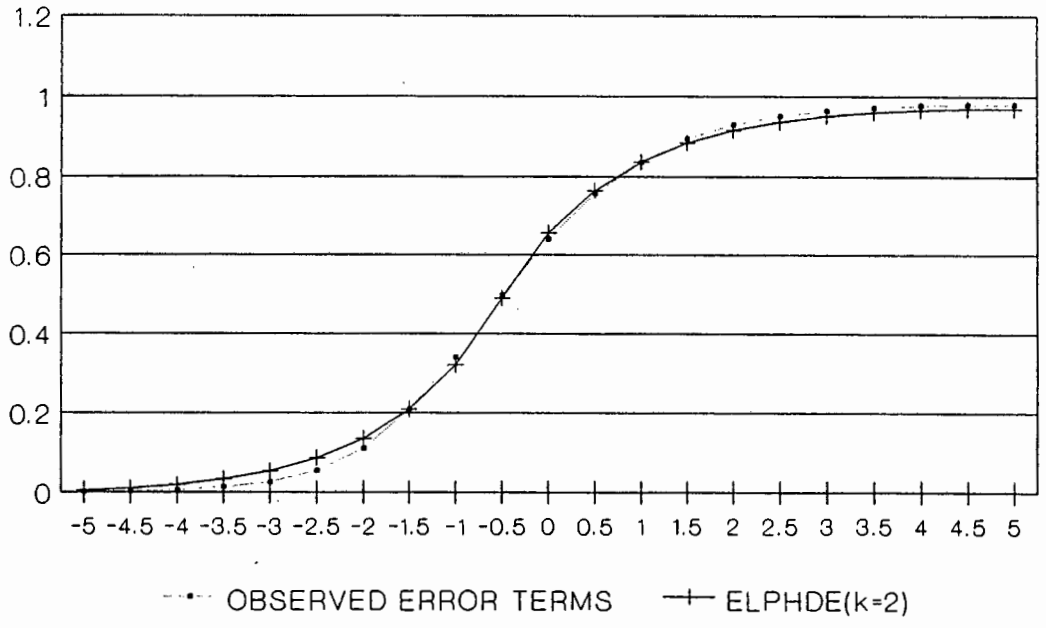


FIGURE 5.3.F

ERROR DBN.: MODEL C.
OBSERVED VS ELPHDE(k=2): LN(M2)

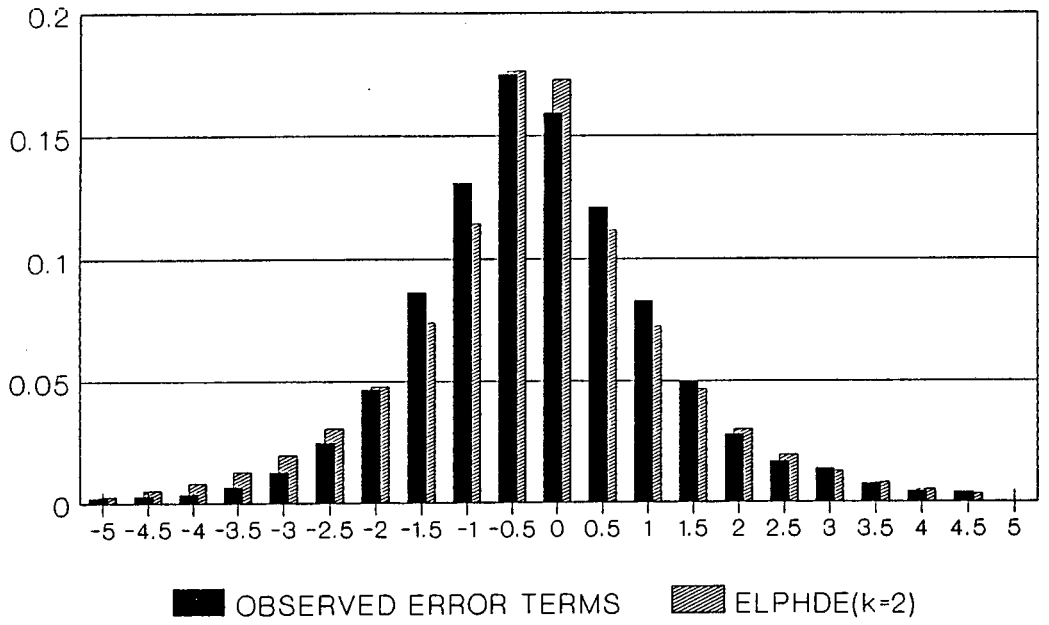


FIGURE 5.3.G

ERROR DBN: MODEL C.
OBSERVED VS ELPHDE(k=2): LN(M2)

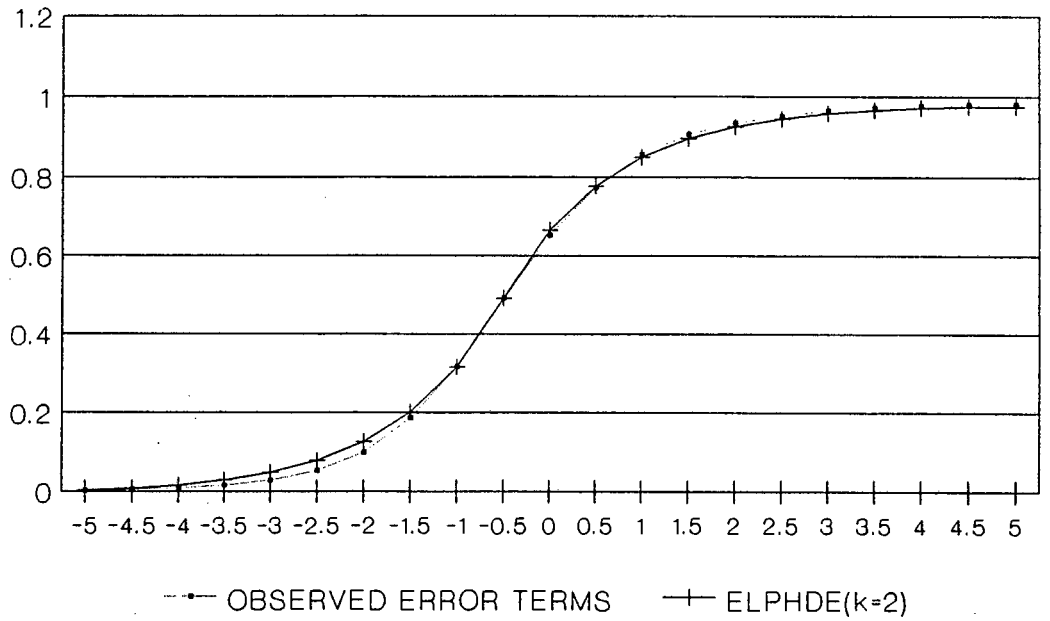


FIGURE 5.3.H

5.4 MODEL D

MODEL D: Model the error distribution using the Elphinstone (1985) method with a unit Normal target distribution.

The distribution of the standardized residuals is geven by:

$$f(e_t) = \frac{1}{\sqrt{2\pi}}\exp -\frac{1}{2}((\xi + \theta e_t)^2)\theta$$

(for $k = 0$ using Elphinstone notation)

where

$$e_t = \left[\left(\frac{X_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right]$$

and where we use X_t to denote the values of the time series, that is X_t is used to represent $\ln(m0)$ (or $\ln(m2)$) and where

$$\mu_t = \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

and

$$\sigma_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi}{1460} * t\right) + \beta_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

We require the joint likelihood of the X_t process and therefore need to make a transformation where the Jacobian of the transformation is given by:

$$\frac{de_t}{dX_t} = \frac{1}{\sigma_t}$$

and therefore

$$|J| = |\frac{1}{\sigma_t}|$$

and we then have

$$f(X_t/X_{t-1}, \underline{\alpha}, \underline{\beta}, \phi, \xi, \theta) = \frac{1}{\sqrt{2\pi}}\exp -\frac{1}{2}\left(\xi + \theta\left[\left(\frac{X_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right]\right)^2 \theta\frac{1}{\sigma_t}$$

and call this distribution $ELPHN_{k=0}$

The conditional likelihood ($L(.)$) is then given by, where we condition on the first observation:

$$L(\underline{\alpha},\underline{\beta},\phi,\xi,\theta;X_2,\ldots,X_n/X_1)=\prod_{t=2}^{n} f(X_t/X_{t-1},\underline{\alpha},\underline{\beta},\phi,\xi,\theta)$$

$$=\left(\frac{1}{\sqrt{2\pi}}\right)^{n-1}\exp-\frac{1}{2}\left(\sum_{t=2}^{n}\left(\xi+\theta\left[\left(\frac{X_t-\mu_t}{\sigma_t}\right)-\phi\left(\frac{X_{t-1}-\mu_{t-1}}{\sigma_{t-1}}\right)\right]\right)^2\right)$$

$$\theta^{n-1}\prod_{t=2}^{n}\left(\frac{1}{\sigma_t}\right)$$

The conditional negative log likelihood (- $\ln L(.)$) is then given by:

$$-\ln L(.)=(n-1)\ln(\sqrt{2\pi})+\frac{1}{2}\sum_{t=2}^{n}\left(\xi+\theta\left[\left(\frac{X_t-\mu_t}{\sigma_t}\right)-\phi\left(\frac{X_{t-1}-\mu_{t-1}}{\sigma_{t-1}}\right)\right]\right)^2$$

$$-(n-1)\ln\theta+\sum_{t=2}^{n}\ln(\sigma_t)$$

The above equations are where the Elphinstone index $k=0$. The equations are calculated for $k=0,1,2,3,4$ and the resulting negative log likelihood values are given in the following table.

TABLE 5.4.1

RESULTING NEGATIVE LOG LIKELIHOOD VALUES.

	$\ln(m0)$	$\ln(m2)$
$k=0$	4637.01	5127.02
$k=1$	4614.70	5101.88
$k=2$	3730.26	4124.10
$k=3$	3729.00	4123.20
$k=4$	3727.40	4120.04

Then as detailed in the Elphinstone chapter a decision must be made as to what value for index k is the most appropriate. As stated in that chapter we use both the AIC and SIC to give an indication as to what value k should be. The following table gives the AIC and SIC values for both the $\ln(m0)$ and $\ln(m2)$ series.

TABLE 5.4.2

	$\ln(m0)$		$\ln(m2)$	
	SIC	AIC	SIC	AIC
$k=0$	4645.98	9278.02	5135.99	10258.04
$k=1$	4632.64	9237.40	5118.96	10211.76
$k=2$	3757.17	7472.52	4151.01	8260.20
$k=3$	3764.88	7474.00	4159.08	8262.40
$k=4$	3772.26	7474.80	4164.90	8260.08

Therefore, when using the AIC, we choose $k=2$ for both the $\ln(m0)$ and $\ln(m2)$ series since $AIC(k)$ is at a minimum when $k=2$. If we base our decision on the SIC we would choose $k=0$ since $SIC(k)$ is at a minimum when $k=0$.

The following table gives the results of the parameter estimation procedures, using $k=2$ and the six parameters $P0, P1, \ldots, P5$ are those relating to the Elphinstone method.

TABLE 5.4.3

RESULTS OF PARAMETER ESTIMATION PROCEDURES.

	$\ln(m0)$	$\ln(m2)$
Negative Log Likelihood:	3730.26	4124.10
Parameters:		
$\widehat{\alpha}_0$	-1.0741	-4.8644
$\widehat{\alpha}_1$	0.0497	-0.1183
$\widehat{\alpha}_2$	0.2215	0.1364
$\widehat{\beta}_0$	0.1822	0.1906
$\widehat{\beta}_1$	0.0040	-0.0115
$\widehat{\beta}_2$	-0.0022	-0.0435
$\widehat{\phi}$	0.8227	0.8356
$\widehat{P0}$	-0.0893	-0.04345
$\widehat{P1}$	.5020	0.4036
$\widehat{P2}$	-0.0024	-0.0021
$\widehat{P3}$	-0.0014	-0.0013
$\widehat{P4}$	0.000005	0.000005
$\widehat{P5}$	0.000001	0.000002

The probability density function of X_t , given X_{t-1} is then given by:

$$\hat{f}(X_t/X_{t-1},\underline{\hat{\alpha}},\underline{\hat{\beta}},\hat{\phi},\widehat{P0},...,\widehat{P5}) = \frac{1}{\sqrt{2\pi}}\exp -\frac{1}{2}\left(\widehat{P0}+\widehat{P1}e_t+\widehat{P2}{e_t}^2+\widehat{P3}{e_t}^3+\widehat{P4}{e_t}^4+\widehat{P5}{e_t}^5\right)^2 \times\left(\widehat{P1}+\widehat{P2}e_t+\widehat{P3}{e_t}^2+\widehat{P4}{e_t}^3+\widehat{P5}{e_t}^4\right)\frac{1}{\hat{\sigma}_t}$$

where this distribution is called $ELPHN_{k=2}$

and where

$$e_t = \left[\left(\frac{X_t-\hat{\mu}_t}{\hat{\sigma}_t}\right)-\hat{\phi}\left(\frac{X_{t-1}-\hat{\mu}_{t-1}}{\hat{\sigma}_{t-1}}\right)\right]$$

where

$$\hat{\mu}_t = \hat{\alpha}_0+\hat{\alpha}_1\cos\left(\frac{2\pi}{1460}*t\right)+\hat{\alpha}_2\sin\left(\frac{2\pi}{1460}*t\right)$$

and

$$\hat{\sigma}_t = \hat{\beta}_0+\hat{\beta}_1\cos\left(\frac{2\pi}{1460}*t\right)+\hat{\beta}_2\sin\left(\frac{2\pi}{1460}*t\right)$$

It was then noted that one of the practical features of this model was that when generating $Hm0$ values the model using $k = 1$ yielded an $Hm0$ probability distribution that was closer to the observed $Hm0$ distribution than when using the model with $k = 2$.

Interpretation of Model D.

The Mean function.

We have a mean function of the following form:

$$\widehat{\mu}_t = \widehat{\alpha}_0 + \widehat{\alpha}_1 \cos\omega t + \widehat{\alpha}_2 \sin\omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

For the $\ln(m0)$ series we have:

$$\widehat{\alpha}_0 = -1.0741$$

$$\widehat{\alpha}_1 = 0.0510$$

$$\widehat{\alpha}_2 = 0.2226$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.2283$ and $\psi = -1.3455$.

This has the following interpretation:

The maximum value for the mean curve is $\alpha_0 + R = -0.8457$ which occurs on the 16^{th} of August. The minimum value for the mean curve is $\alpha_0 - R = -1.3024$ which occurs on the 15^{th} of February.

When using this model we also examined the effect of including more terms in the mean function, i.e. using a mean function of the form:

$$\widehat{\mu}_t = \widehat{\alpha}_0 + \widehat{\alpha}_1 \cos\left(\frac{2\pi}{1460}t\right) + \widehat{\alpha}_2 \sin\left(\frac{2\pi}{1460}t\right) + \widehat{\alpha}_3 \cos\left(\frac{2\pi}{1460}2t\right) + \widehat{\alpha}_4 \sin\left(\frac{2\pi}{1460}2t\right)$$

There was however no significant improvement in the maximum likelihood value and a neglible change in the other parameter estimates. The estimates for $\widehat{\alpha}_3$ and $\widehat{\alpha}_4$ given by: $\widehat{\alpha}_3 = -0.0025$ and $\widehat{\alpha}_4 = 0.0041$.

Figure 5.4.A shows both mean functions plotted together and as can be seen it is difficult to detect any significant difference between the two.

For the $\ln(m2)$ series we have:

$$\widehat{\alpha}_0 = -4.8644$$

$$\widehat{\alpha}_1 = -0.1183$$

$$\widehat{\alpha}_2 = 0.1364$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.1805$ and $\psi = -2.2852$.

This has the following interpretation: (see Figure 5.4.B):

The maximum value for the mean curve is $\alpha_0 + R = -4.683$ which occurs on the 10^{th} of October. The minimum value for the mean curve is $\alpha_0 - R = -5.044$ which occurs on the 11^{th} of April.

The Variance function.

We have a variance function of the following form:

$$\widehat{\sigma}_t = \widehat{\beta}_0 + \widehat{\beta}_1 \cos\omega t + \widehat{\beta}_2 \sin\omega t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

For the $\ln(m0)$ series we have:

$$\widehat{\beta}_0 = 0.1822$$

$$\widehat{\beta}_1 = 0.0040$$

$$\widehat{\beta}_2 = -0.0022$$

We then obtain values for R and ψ (the amplitude and phase), and $R = 0.0045$ and $\psi = 0.5028$.

This has the following interpretation: (see Figure 5.4.C):

The maximum value for the variance curve is $\beta_0 + R = 0.1867$ which occurs on the 1^{st} of May. The minimum value for the variance curve is $\beta_0 - R = 0.1777$ which occurs on the 1^{st} of November.

For the $\ln(m2)$ series we have:

$$\widehat{\beta}_0 = 0.1906$$

$$\widehat{\beta}_1 = -0.0115$$

$$\widehat{\beta}_2 = -0.0435$$

We then obtain values for R and ψ (the amplitude and phase) and $R = 0.0449$ and $\psi = 1.8292$.

This has the following interpretation: (see Figure 5.4.D):

The maximum value for the variance curve is $\beta_0 + R = 0.2355$ which occurs on the 13^{th} of February. The minimum value for the variance curve is $\beta_0 - R = 0.1456$ which occurs on the 14^{th} of August.

Error Distribution

Finally in this section the plots are given of the comparison between the observed error terms when using this model and the distribution used to model them. These are given in Figures 5.4.E & 5.4.F for the $\ln m0$ series and in Figures 5.4.G & 5.4.H for the $\ln m2$ series.

MODEL D (LN (M0))
PERIODICITY OF MEAN CURVES

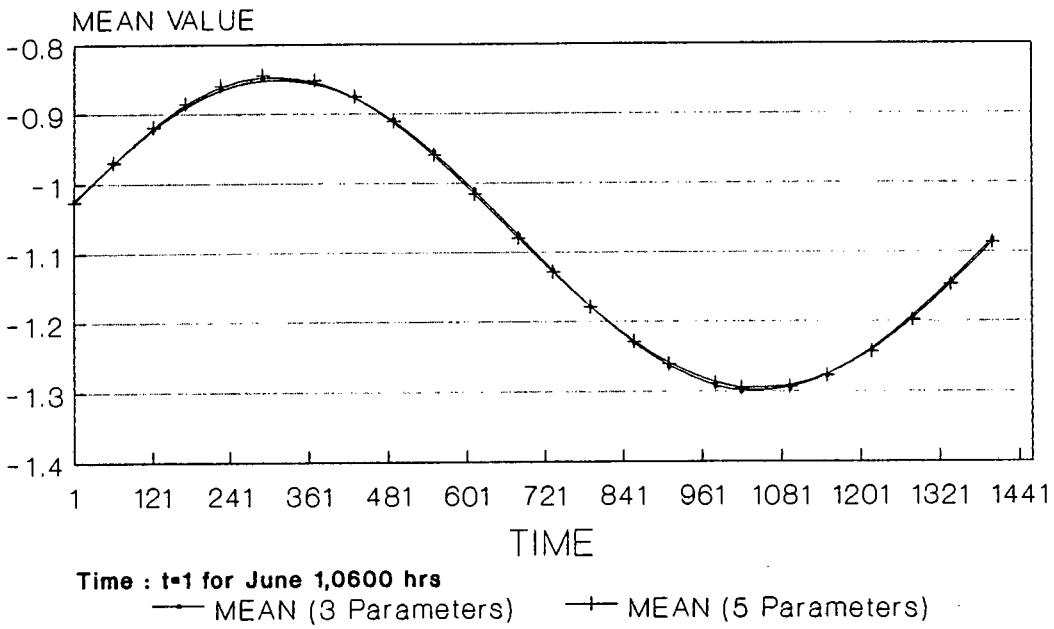


FIGURE 5.4.A

MODEL D (LN (M2))
PERIODICITY OF MEAN CURVE

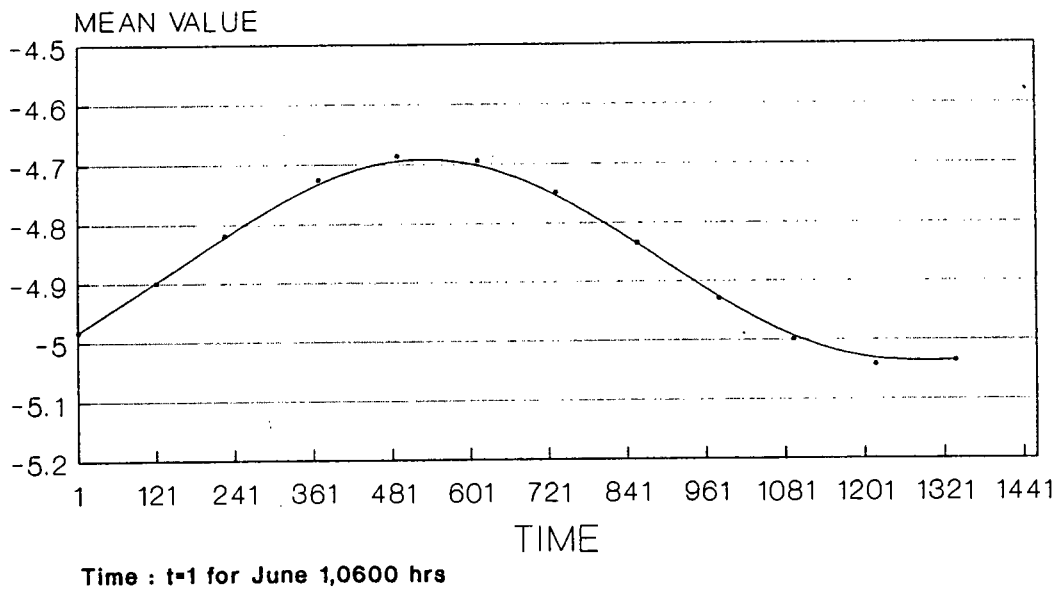
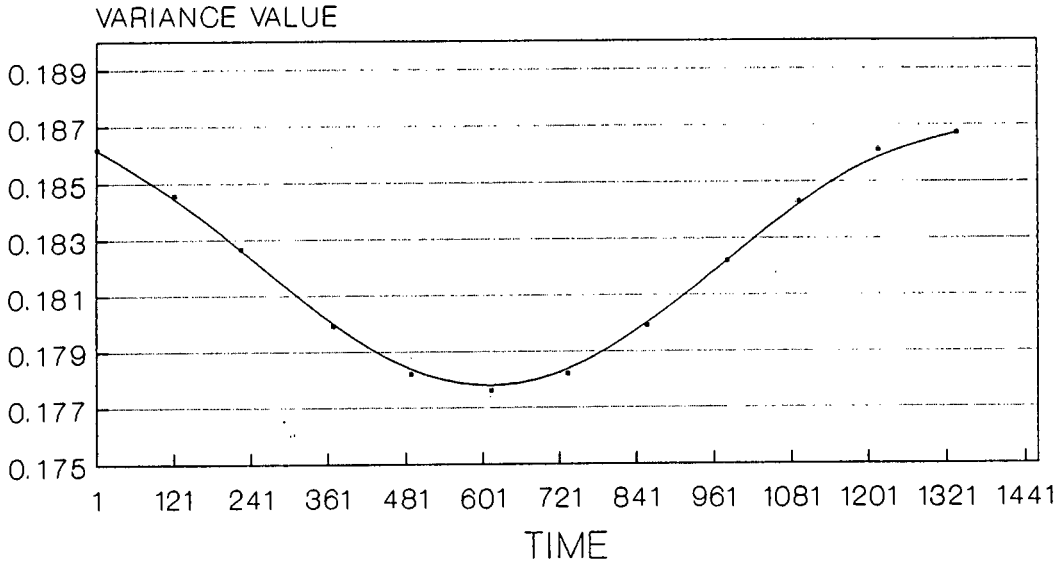


FIGURE 5.4.B

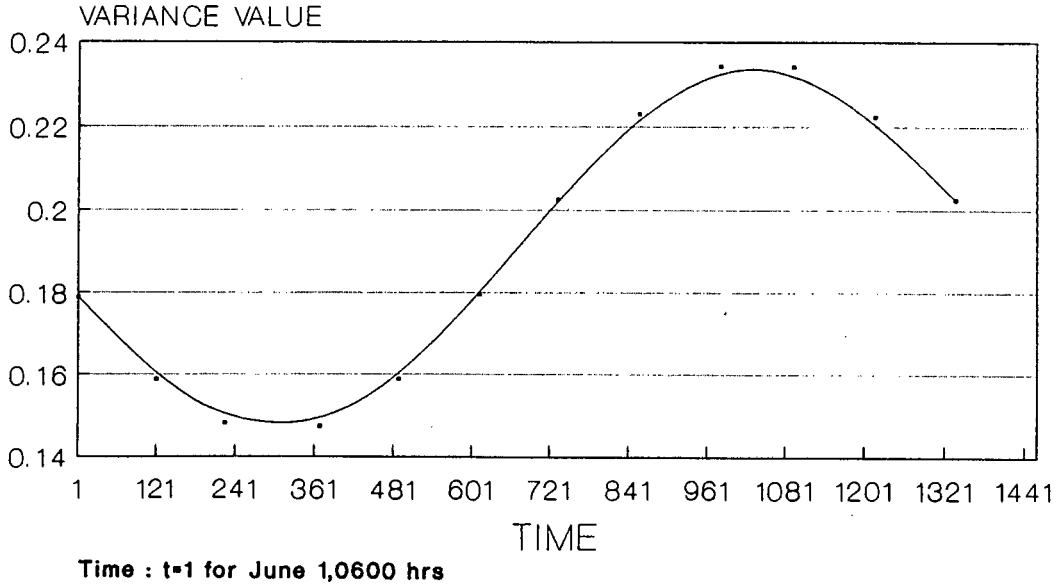
MODEL D (LN (M0))
PERIODICITY OF VARIANCE CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.4.C

MODEL D (LN (M2))
PERIODICITY OF VARIANCE CURVE



Time : t=1 for June 1,0600 hrs

FIGURE 5.4.D

ERROR DBN: MODEL D.
OBSERVED VS ELPHN(k=2): LN(M0)

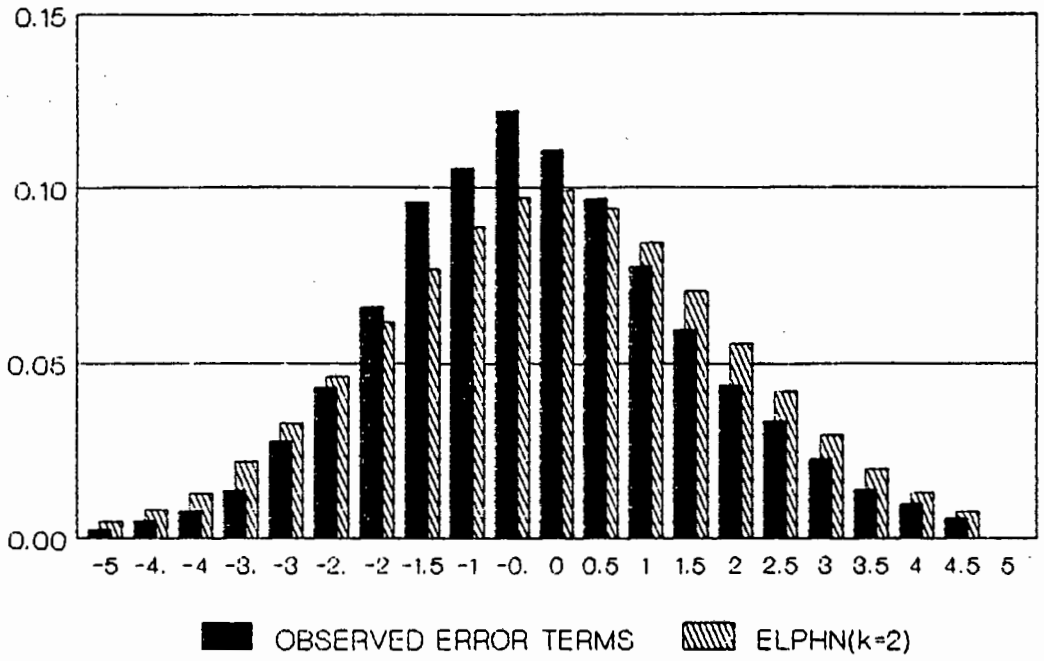


FIGURE 5.4.E

ERROR DBN: MODEL D.
OBSERVED VS ELPHN(k=2): LN(M0)

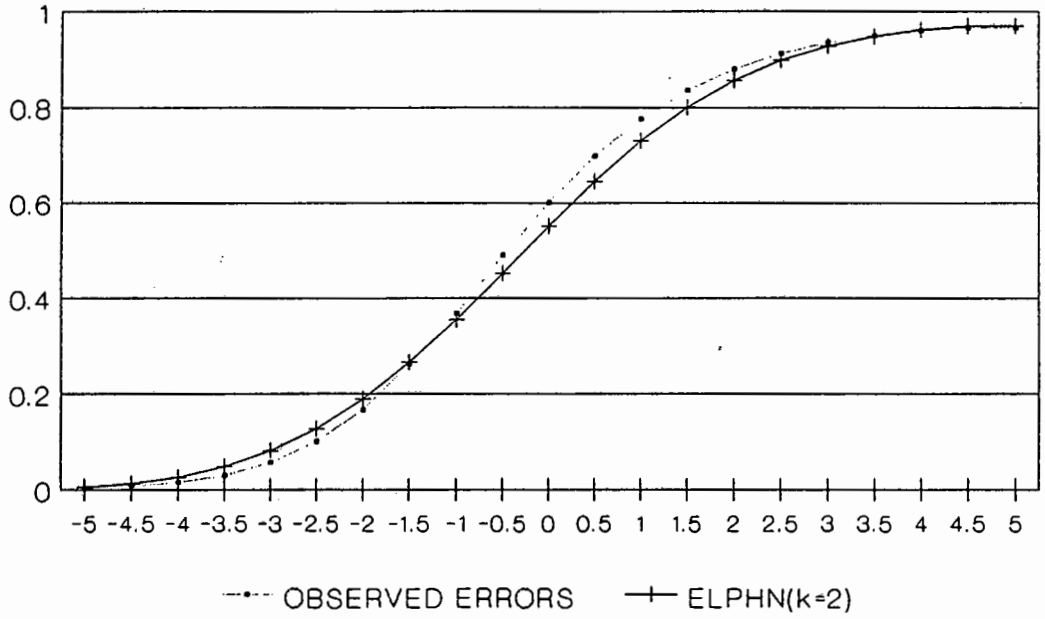


FIGURE 5.4.F

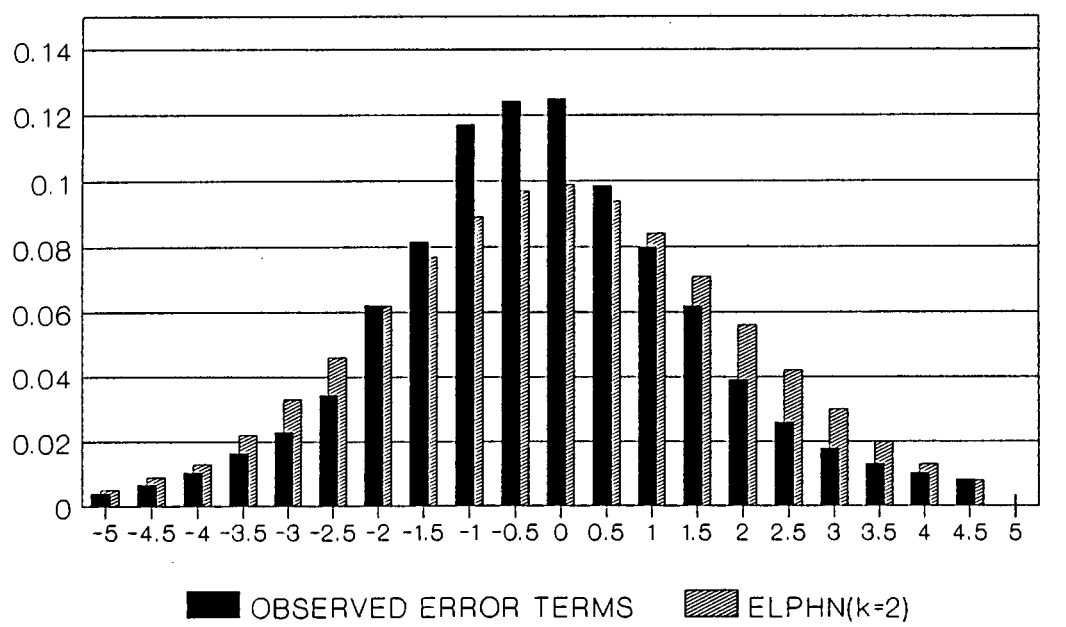


FIGURE 5.4.G

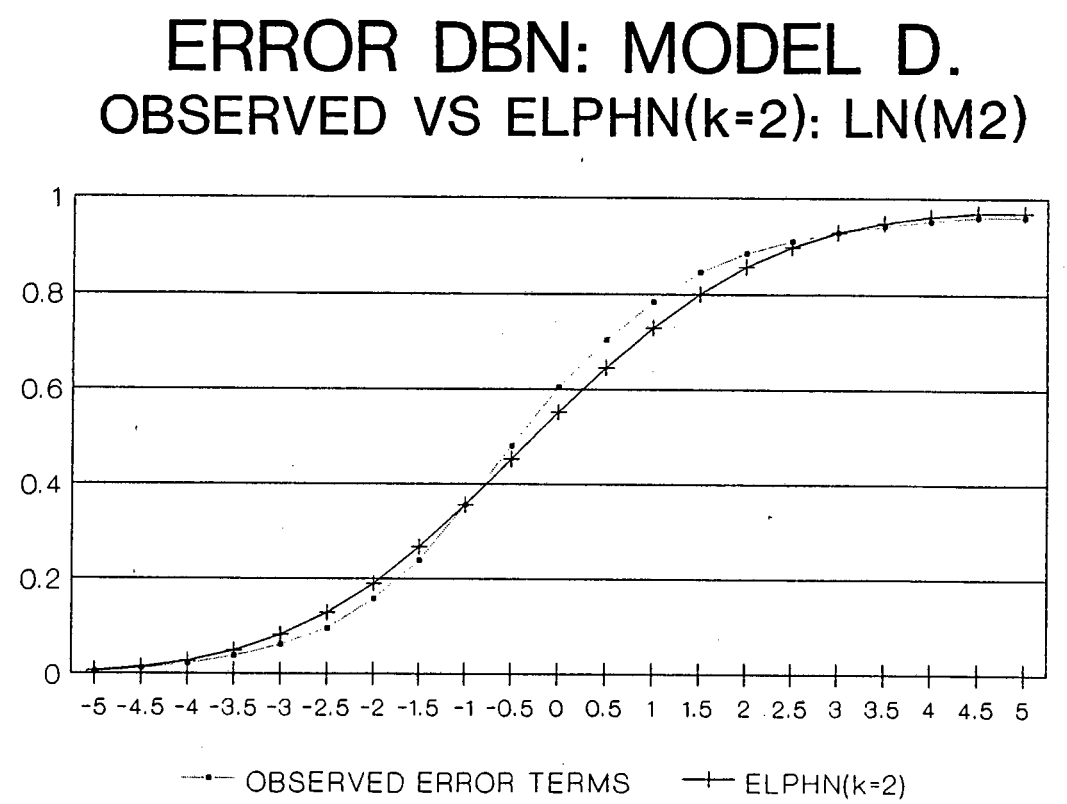


FIGURE 5.4.H

CHAPTER 6

$Hm0$ GENERATION ALGORITHM.

This chapter deals with the algorithms used to generate artificial $Hm0$ series. We have now obtained estimates for the parameters for each of the models and need to "re-generate" the process. This section gives details of how this was achieved. The models differ mainly in the e_t distribution, and the algorithms to generate random variables from each of these distributions is given. We then give the algorithm used for the generation of $Hm0$ values.

For Models A,C and D we have

$$e_t = \left[\left(\frac{X_t - \mu_t}{\sigma_t}\right) - \phi\left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right)\right]$$

and for Model B we have

$$e_t = [(X_t - \mu_t) - \phi(X_{t-1} - \mu_{t-1})]$$

where for all the models we have

$$\mu_t = \alpha_0 + \alpha_1 \cos\left(\frac{2\pi}{1460} * t\right) + \alpha_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

and for Models A,C and D we have

$$\sigma_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi}{1460} * t\right) + \beta_2 \sin\left(\frac{2\pi}{1460} * t\right)$$

Then making X_t the subject of the formula we have (for Models A,C and D)

$$X_t = \mu_t + \sigma_t \left(\phi\left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right) + e_t\right)$$

and for Model B

$$X_t = \mu_t + (\phi\,(X_{t-1} - \mu_{t-1}) + e_t)$$

For Model A $e_t \sim DE(\lambda)$

For Model B $e_t \sim DE(\lambda_t)$

For Model C $e_t \sim ELPHDE_{k=2}$

For Model D $e_t \sim ELPHN_{k=2}$

The algorithms for the generation of random variables, (e_t) , from the above distributions is given below.

Model A: $e_t \sim DE(\lambda)$

The algorithm used to generate a random variable (x) from the Double Exponential (DE) distribution for a given λ was simply the following, (see Appendix C for further DE details):

STEP 1: Generate U where ($U \sim U(0,1)$) i.e. Uniform [0;1];

STEP 2: If $U < 0.5$ then let $x = (\lambda * \ln(2*U))$ else let $x = (-\lambda * \ln(2*(1-U)))$

Model B: $e_t \sim DE(\lambda_t)$.

The algorithm used to generate random variables (x) from the $DE(\lambda_t)$ distribution:

STEP 1: Set up a vector of λ values for $t = 1, \ldots, 1460$ using the estimates of $\hat{\underline{\lambda}}$ given in Chapter 5.2.

STEP 2: Generate U where $U \sim U(0;1)$.

STEP 3: Using the appropriate λ_t , if $U < 0.5$ then let $x = \lambda_t * \ln(2 * U)$ else if $U \geq 0.5$ then let $x = -\lambda_t * \ln(2 * (1 - U))$

STEP 4: Either STOP or return to STEP 2 to generate more values.

Model C: $e_t \sim ELPHDE_{k=2}$

The algorithm used to generate a random variable (x) from the $ELPHDE_{k=2}$ distribution was simply the following, (see Appendix C for further DE details):

When using a Double Exponential target distribution we have, using notation from

the Elphinstone section, that

$$F(g_k(x;\theta)) = (\frac{1}{2})\exp(g_k(x;\theta)) \qquad g_k(x;\theta) \leq 0$$
$$= 1 - \frac{1}{2}\exp(-g_k(x;\theta)) \qquad g_k(x;\theta) \geq 0$$

Then

$$g_k(x;\theta) = \ln(2 * U) \qquad U \leq 0.5$$
$$= -(\ln(2 * (1 - U))) \qquad U \geq 0.5$$

For $k = 0$ $\quad g_k(x;\theta) = (P0 + P1x)$

Now if we let $A = \ln(2U)$ and $B = -(\ln(2 * (1 - U)))$ and then if

$U \leq 0.5$ then $P0 + P1x = A$ and $x = \left(\frac{A - P0}{P1}\right)$ else if

$U \geq 0.5$ then $P0 + P1x = B$ and $x = \left(\frac{B - P0}{P1}\right)$

For $k > 0$ generation requires the setting up of a two column vector of the form:

For a comprehensive range of x_i values column 1 of the vector contains the value $g_k(x_i;\theta)$ and column 2 contains the corresponding x_i value. The algorithm then proceedes as follows:

STEP 1: Generate U where ($U \sim U(0,1)$) i.e. Uniform [0;1];

STEP 2: If $U < 0.5$ then let $TEMP = (\ln(2 * U))$ else let $TEMP = (-\ln(2 * (1 - U)))$

STEP 3: Search through column 1 of the vector for the value in that column closest to $TEMP$.

STEP 4: The variable x then takes on that value in the corresponding column 2 of the vector. (Interpolation may also be required for further accuracy of the x value).

STEP 5:Either STOP or return to STEP 1 to generate further values.

Model D: $e_t \sim ELPHN_{k=2}$

The algorithm used to generate a random variable (x) from the $ELPHN_{k=2}$ distribution was simply the following:

For $k = 0$ $\quad g_k(x;\theta) = (P0 + P1x)$

Now if we let $A = \ln(2U)$ and $B = -(\ln(2 * (1 - U)))$ and then if

$$U \leq 0.5 \text{ then } P0 + P1x = A \text{ and } x = \left(\frac{A - P0}{P1}\right) \text{ else if}$$

$$U \geq 0.5 \text{ then } P0 + P1x = B \text{ and } x = \left(\frac{B - P0}{P1}\right)$$

For $k > 0$ generation requires the setting up of a two column vector of the form:

For a comprehensive range of x_i values column 1 of the vector contains the value $g_k(x_i;\theta)$ and column 2 contains the corresponding x_i value. The algorithm then proceedes as follows:

STEP 1: Generate N where ($N \sim N(0,1)$) i.e. Normal [0;1];

STEP 2: Let $TEMP = N$, then search through column 1 of the vector for the value in that column closest to $TEMP$.

STEP 3: The variable x then takes on that value in the corresponding column 2 of the vector. (Interpolation may also be required for further accuracy of the x value).

STEP 4:Either STOP or return to STEP 1 to generate further values.

Hm0 GENERATION ALGORITHM.

STEP 1. Set up μ_t and σ_t vectors for $t = 1,\ldots,1460$ using the appropriate model estimates for $\underline{\alpha}$ and $\underline{\beta}$

STEP 2. Initially set the variables $X_{t-1} = \mu_{1460}$, $\mu_{t-1} = \mu_{1460}$ and $\sigma_{t-1} = \sigma_{1460}$.

STEP 3. Generate the appropriate e_t . (Algorithms given).

STEP 4. For Models A,C and D let

$$X_t = \mu_t + \sigma_t \left(\phi \left(\frac{X_{t-1} - \mu_{t-1}}{\sigma_{t-1}}\right) + e_t\right)$$

and for Model B let

$$X_t = \mu_t + (\phi (X_{t-1} - \mu_{t-1}) + e_t)$$

STEP 5. Let $m0_t = \exp(X_t)$ (since our model is for $X_t = \ln(m0)$).

STEP 6. Let $Hm0_t = 4 * \sqrt{m0_t}$

STEP 7. Either RETURN to STEP 4 to generate more values or STOP.

The $Hm0$ generation is now used to generate artificial $Hm0$ series for each of the models and we compare the resulting series in the next chapter.

CHAPTER 7.

COMPARISON OF MODELS.

In this chapter various aspects of the models are compared. There are a number of possible criteria that we could use to compare the models and it is for us to decide which aspects are the most important. On a purely statistical basis we could use one of several goodness of fit criteria to decide which model fitted the observed data the best, or use a maximum likelihood criteria to decide on the best model.

For this study one of the main criteria was the ability to generate a $Hm0$ series which preserved selected properties of the original series and it was for this reason that this aspect was given greater emphasis.

A visual examination of how well the residual distribution was approximated, given in Chapter 5, revealed that the Double Exponential based distributions, including the DE target distribution when using the Elphinstone (1985) method, i.e. Models A,B and C, gave a better fit than Model D which used the Elphinstone (1985) method with a Normal target distribution.

The maximum likelihood comparison was performed using the negative log likelihood values -see Table 7.1. If we used this as a criterion we would choose that model which yielded the minimum negative log likelihood value. In this case it would be Model C for both the $\ln m0$ and the $\ln m2$ series. The "worst" in this case would once again be Model D which has the maximum negative log likelihood value.

TABLE 7.1

NEGATIVE LOG LIKLIHOOD VALUES.

Model	$\ln(m0)$	$\ln(m2)$
A	3507.56	3727.92
B	3507.57	3727.93
C	3501.20	3721.10
D	3730.26	4124.10

So when using either of the above criteria the DE based models and in particular

Model C would be regarded as "better" than Model D. However no indication is given as to how much "better".

We however view the generation aspect as being a very important part of this study and it is for this reason that we include a comparison of the generated seasonal $Hm0$ values from each of the Models to the observed $Hm0$ values.

These are given in Table 7.2 and it becomes clear that although there are similarities in the generated series the values from the DE based models yield slightly lower mean values, a slightly higher variance and maximum values that are cause for concern, when compared to the observed series.

TABLE 7.2

COMPARISON OF OBSERVED $Hm0$ TO MODEL GENERATED $Hm0$.

Season		OBSERVED	MODEL A	MODEL B	MODEL C	MODEL D
WINTER:	Mean	2.90	2.61	2.61	2.73	2.89
	Variance	1.15	1.17	1.16	1.61	1.23
	Maximum	8.02	9.01	9.00	15.34	9.96
	Minimum	0.50	0.78	0.78	0.26	0.57
SPRING:	Mean	2.80	2.49	2.49	2.64	2.75
	Variance	1.10	1.33	1.34	1.80	1.09
	Maximum	8.67	12.45	12.45	17.10	7.39
	Minimum	0.34	0.43	0.43	0.33	0.75
SUMMER:	Mean	2.40	2.33	2.33	2.39	2.40
	Variance	0.63	1.07	1.08	1.39	0.88
	Maximum	8.60	10.28	10.30	22.83	7.66
	Minimum	0.20	0.46	0.46	0.22	0.60
AUTUMN:	Mean	2.58	2.37	2 .36	2.55	2.57
	Variance	1.12	1.49	1.49	1.78	1.13
	Maximum	10.8	13.10	13.10	23.08	8.75
	Minimum	0.20	0.43	0.43	0.13	0.50

In our view all four models would be adequate in most respects but because the tails of the DE distribution are longer than those of the Normal distribution we tend to generate more extreme values (and of greater magnitude) when using these models.

We would therefore tentatively recommend Model D. Although the other three models are superior in turns of some criteria, they are less satisfactory in modelling the extremes and in preserving the seasonal properties of the original series.

In what follows we will restrict our attention to Model D. We show that this model does indeed preserve the main properties of the original series. However in making this tentative recommendation we emphasise that Model D is not superior to the other three model in all respects, and that the choice of model depends ultimately on the criteria which we choose to optimize.

CHAPTER 8.

VALIDATION OF MODEL D.

In this chapter we deal with the validation of the artifically generated $Hm0$ series when using Model D. If we are going to use these generated series for any type of application we must first ensure that the generated sequences preserve the properties of the original $Hm0$ series. For example, the monthly and seasonal means and standard deviations, the serial correlation structure, the persistence of calm and storm periods and in fact the entire probability distribution of the $Hm0$ values.

Before we start however it is important to note that there is a definite between-year variation in the observed $Hm0$ values, see Table 8.1. This problem, which has been mentioned in Carter et al (1986), presents difficulties in deciding whether the artificially generated $Hm0$ sequence is in fact "close" enough to the observed.

We note ,in the following Table, that the sample variance varies from 0.835 to 1.545, i.e. as much as 50 %, and the sample mean varies from 2.823 to 2.524.

TABLE 8.1

OBSERVED YEARLY $Hm0$ VALUES.

YEAR	SAMPLE SIZE	MEAN	VARIANCE	MAXIMUM
1978	686	2.823	1.091	8.910
1979	1243	2.684	1.115	8.250
1980	1318	2.709	0.889	6.930
1981	1391	2.663	0.941	6.600
1982	1292	2.640	0.835	6.930
1983	1420	2.620	1.115	7.590
1984	520	2.524	1.545	10.890

In Table 8.2 we give a comparison of the observed data to a number of generated sequences using Model D. The runs are for periods of 10 years and where the only changes to the generation algorithm are to the initial "seed" value used to start the random number generator for the error distribution.

TABLE 8.2

VARIOUS RUNS OF MODEL D GENERATED $Hm0$ VALUES.

Season		*OBSERVED*	RUN 1	RUN 2	RUN 3	RUN 4
WINTER:	Mean	2.909	2.880	2.892	2.857	2.880
	Variance	1.156	1.370	1.250	1.190	1.230
SPRING:	Mean	2.807	2.709	2.751	2.825	2.752
	Variance	1.105	1.050	1.210	1.250	1.090
SUMMER:	Mean	2.408	2.440	2.428	2.390	2.400
	Variance	0.635	0.850	0.850	0.712	0.888
AUTUMN:	Mean	2.585	2.617	2.568	2.498	2.572
	Variance	1.129	1.137	1.160	1.030	1.135
ALL YEAR	Mean	2.670	2.663	2.662	2.645	2.68
	Variance	1.040	1.130	1.150	1.140	1.090

It can be seen that all the runs are of the same magnitude and that if more observed data were available the statistics of the observed data would be more stable and a more usefull comparison to the artificial data could be made. This aspect is clearly illustrated in Figure 8.A (Seasonal Comparison) where we can see that the seasonal $Hm0$ means are very similar for the observed and generated series.

Figure 8.B shows a comparison of the monthly means for the observed and generated $Hm0$ series. One hundred years of generated data was used for to construct this Figure, hence the 'smoother' curve for the generated series. Overall though the generated values curve follows the observed values curve very closely.

The serial correlation structure for the observed values are given in Table 8.3, and Figure 8.C, and this structure seems to be fairly well preserved in the artificially generated sequence.

TABLE 8.3

OBSERVED VS GENERATED SERIAL CORRELATION STRUCTURES.

Lag	OBSERVED $Hm0$	GENERATED $Hm0$
1	0.847	0.816
2	0.714	0.665
3	0.579	0.545
4	0.457	0.449
5	0.353	0.370
6	0.272	0.308
7	0.212	0.263
8	0.161	0.222
9	0.128	0.189
10	0.093	0.164
11	0.069	0.149
12	0.054	0.138

For the next two comparisons we need to define "storm" periods and "calm" periods. The decision as to what $Hm0$ values to use is essentially arbitrary and in this study we defined a "storm" to be when $Hm0 \geq 4.5m$ and a "calm" to be when $Hm0 \leq 2.0m$

A mean value of 3.8, for example, for the calm day statistics (given in Table 8.4) is the mean number of successive observations where $Hm0 \leq 2.0m$. Recall that we have 4 observations per day so a value of 3.8 is almost a full day. (The length of calm and storm periods is sometimes referred to as the persistence of calms and storms.)

Similarly for storm days (given in Table 8.5) a mean value of 2.57 is the mean number of successive observations where $Hm0 \geq 4.5m$.

Once again we encounter the problem of a "small" data set where the effect of one very large storm (the May 1986 storm for example where $Hm0$ reached a value of 10.8 m) has a marked effect on both the mean and variance of the storm day

statistics for that month. This would seem to indicate an "unnatural" monthly variation which is not entirely realistic. It should also be noted here that there are gaps in the observed data series so we do not get a completely true reflection of the persistence of the original $Hm0$ series.

TABLE 8.4

OBSERVED VS GENERATED CALM DAY STATISTICS.

MONTH	OBSERVED $Hm0$		GENERATED $Hm0$	
	MEAN	VARIANCE	MEAN	VARIANCE
1	3.80	17.68	3.75	15.16
2	4.22	15.76	4.06	17.79
3	4.48	31.44	3.82	18.33
4	4.88	38.72	4.14	18.59
5	4.18	20.80	3.75	18.07
6	3.37	12.24	3.22	9.10
7	3.81	8.60	3.04	8.25
8	3.06	10.37	3.23	11.41
9	3.13	8.78	3.10	8.25
10	3.62	18.80	3.47	17.76
11	3.28	8.96	3.73	19.32
12	3.84	7.39	4.49	21.00

TABLE 8.5

OBSERVED VS GENERATED STORM DAY STATISTICS.

MONTH	OBSERVED $Hm0$		GENERATED $Hm0$	
	MEAN	VARIANCE	MEAN	VARIANCE
1	2.57	3.80	2.09	3.41
2	2.00	3.80	1.88	1.08
3	1.60	1.19	1.96	1.88
4	3.00	6.00	2.78	8.24
5	5.30	28.50	2.63	9.61
6	4.00	10.00	2.71	4.87
7	2.55	5.27	2.77	8.81
8	2.87	5.85	2.40	5.63
9	3.05	7.48	2.93	10.34
10	3.08	6.08	2.84	3.67
11	1.73	2.11	2.71	4.69
12	1.91	2.99	2.34	5.19

Finally in Figures 8.D and 8.E we compare the observed $Hm0$ distribution to that of the distribution of generated $Hm0$. Once again we note that the distributions are very similar.

From these comparisons we concluded that the artificially generated $Hm0$ series using Model D does in fact preserve the important properties of the original series, namely the seasonal and monthly means and variance, serial correlation structure and persistence.

SEASONAL COMPARISON
OBSERVED VS GENERATED HM0 MEANS

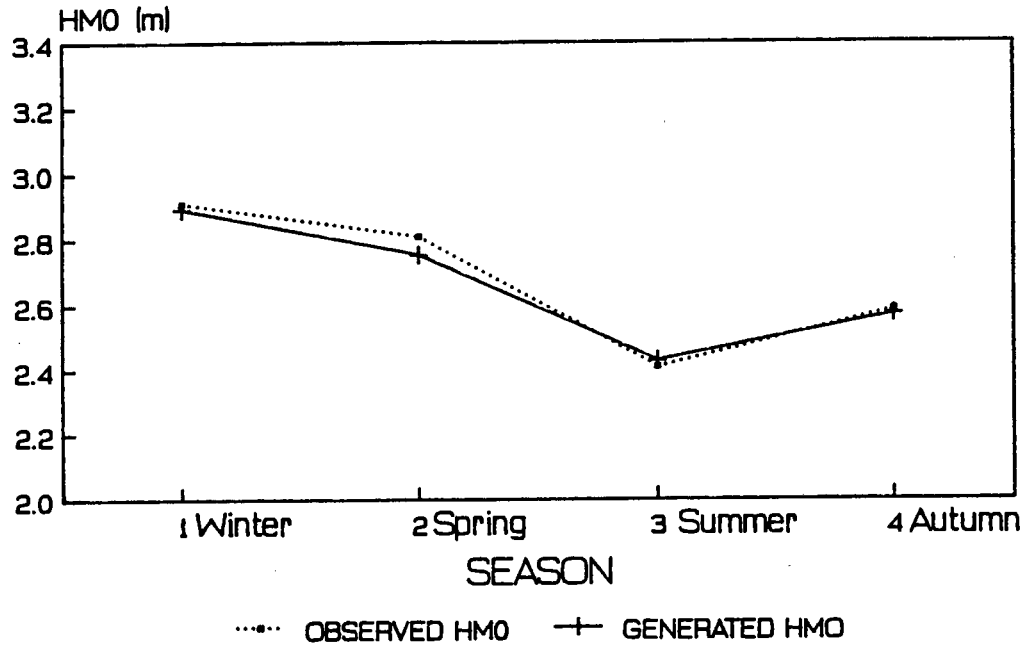


FIGURE 8.A

MONTHLY COMPARISON
OBSERVED VS GENERATED HM0 MEANS

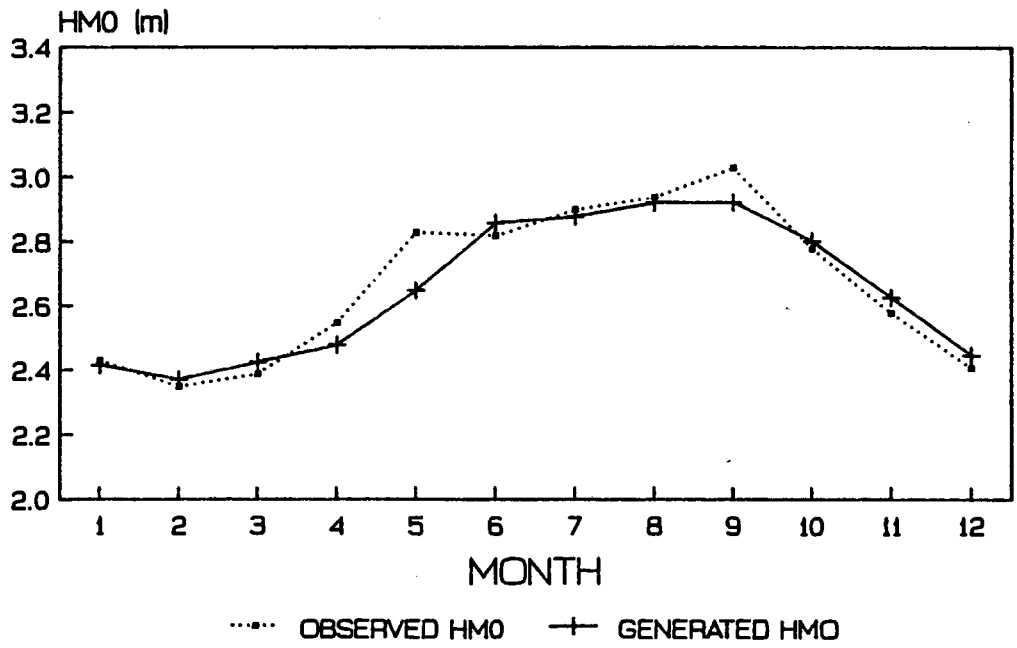


FIGURE 8.B

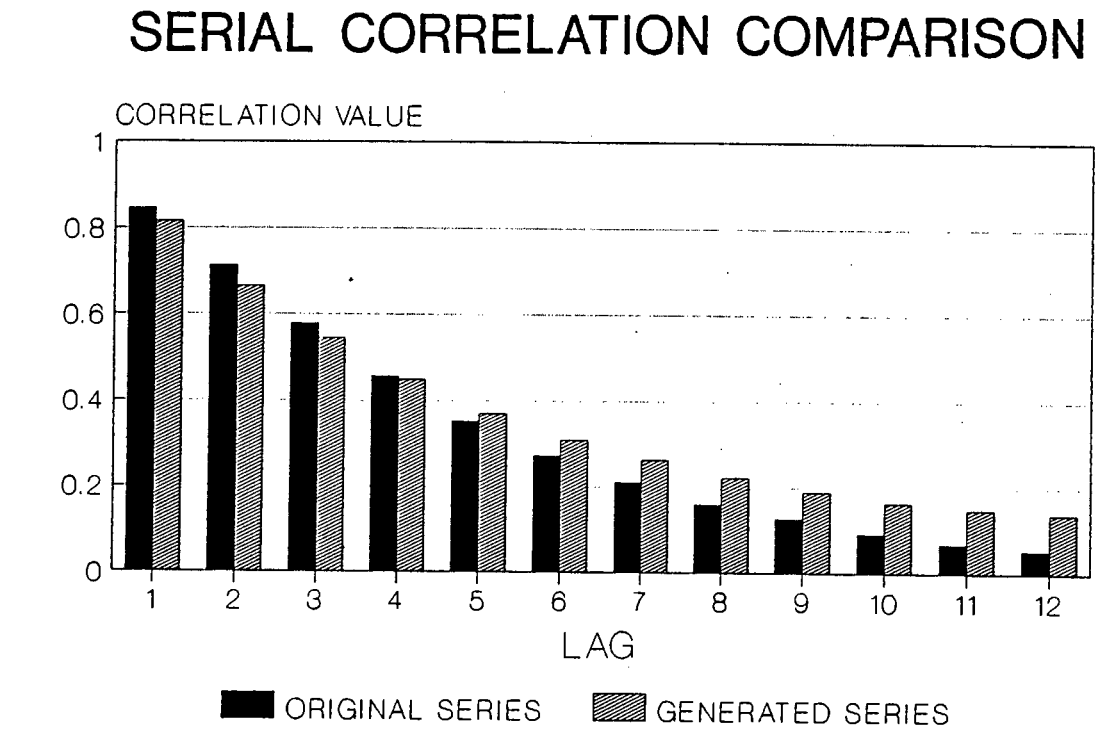

SERIAL CORRELATION COMPARISON
CORRELATION VALUE
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8
9
10
11
12
LAG
ORIGINAL SERIES
GENERATED SERIES

FIGURE 8.C

DISTRIBUTION OF HM0.
OBSERVED VS GENERATED HM0.

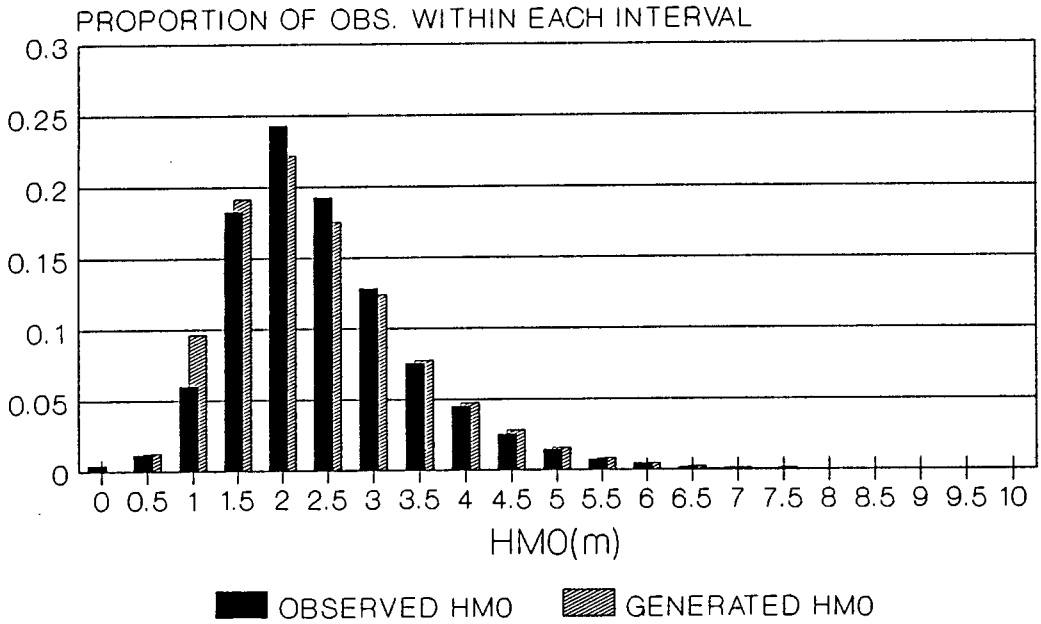


FIGURE 8.D

CUMULATIVE DBN. OF HM0.
OBSERVED VS GENERATED HM0.

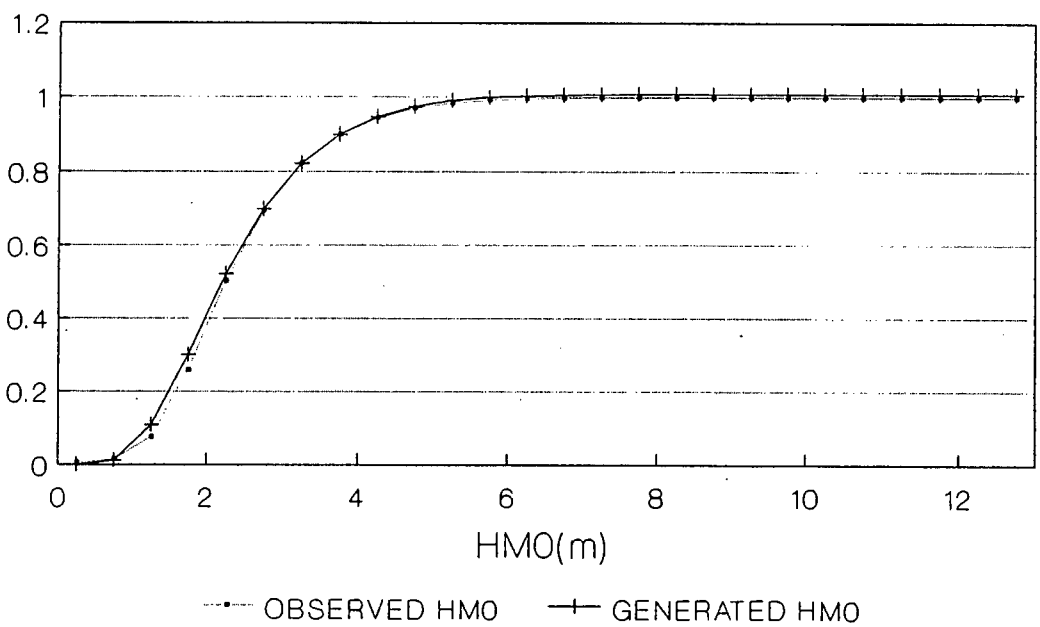


FIGURE 8.E

CHAPTER 9.

APPLICATIONS OF GENERATED $Hm0$ SERIES.

The point of having fitted a model to the $Hm0$ time series is to enable us to answer questions about the (stochastic) behaviour of the series. For example we might wish to estimate the mean $Hm0$ for a particular period, such as 10^{th} January to 25^{th} January, or the distribution of the $Hm0$ between these dates. The latter can be used for example to estimate quantities such as the size of the wave in this period which is exceeded only once in 100 years. Although the model contains the information to answer such questions it is very difficult to derive analytical solutions. Before the advent of cheap computing this would have rendered the model effectively useless for practical purposes. The way we can answer the question is by simulation, i.e. generating long sequences of $Hm0$ using the model. These sequences indirectly express all the properties of the model, and moreover they do so in a convenient form. We may simply regard the sequence as a long realisation of the 'real' data and can answer any question which we would have been able to answer had the real sequence been long enough for us to not have modelled it.

To estimate probabilities we simply regard the artificial $Hm0$ sequence that has been generated as a very long "real" $Hm0$ record. This can be done because the model used to generate the sequences, Model D in this case, preserves the properties of the real $Hm0$ sequence, for example, the monthly and seasonal means and variances, the serial correlation structure and in fact the entire $Hm0$ probability distribution.

Of course the generated sequence has the advantage of being without missing values and therefore the persistence statistics become more meaningful. A serious problem in estimating persistence statistics, such as the probability of having 4 consecutive calm days, arises when there are gaps in the data. For example suppose that there appears to be a run of consecutive calm days but there are missing observations in the run. It is not clear whether this event should be counted as a 4 day calm run or not.

Figures 9.A and 9.B are the persistence plots compiled using 100 years of artificially generated data. These plots are constructed simply by counting the number of times a particular event occurs during 100 year period, for example the number of

times that $Hm0$ was greater than 2.5 m for 3 consecutive days, and then using interpolation to connect these points to form a smooth curve.

When using Figure 9.A, and the 2.5 m curve for example, we can see that on average there are 5.2 times during the year that $Hm0 \geq 2.5m$ for 5 successive days. Similarly for Figure 9.B and using the 3.0 m curve we can see that on average there are 17 times a year that $Hm0 \leq 3.0m$ for 4 successive days.

Now suppose for example that you needed to estimate the probability that $Hm0 \leq 2.5m$ for 4 successive days and then determine the month during which this event was most likely to occur. This problem arises when, for example, a structure needs to be positioned at a place in the ocean where the operation takes 4 days and the $Hm0$ value may not exceed 2.5 m for those 4 days. These probabilities were calculated from the generated $Hm0$ sequence and are given in Figure 9.C. Once again the probabilities were calculated by counting the number of times an event occurred and then using these counts to estimate the probability of that event occurring.

From Figure 9.C we can see that if $Hm0 \leq 1.5m$ initially then the probability that $Hm0 \leq 2.5m$ for the following 4 days is (0.76) in January,...,(0.61) in August,..etc. From this Figure then we can see that the months February, March and possibly December would be the best months for this operation for all the given initial values of $Hm0$.

One can use the artificial sequence generated to estimate a variety of other quantities that may be of interest. Here are some other examples:

1. What is the probability of having $Hm0$ less than some value for 10 consecutive days in August ?

2. What is the probability of having $Hm0$ greater than some value between two specified dates ?

3. Which day (week,month, 50 day period,...) of the year has the highest (or lowest) probability of $Hm0$ remaining within some range of values ?

4. What is the average $Hm0$ for any given period of the year (eg. between 15^{th} May and 3^{rd} June) ? What is the corresponding standard deviation, probability distribution, median, 90% confidence interval ?

You can answer any of these and similar questions by simply treating the generated sequence as if it were a "real" $Hm0$ record.

The generated sequences can also be useful in estimating extreme $Hm0$ values. The methods usually employed for solving this type of problem involve the fitting of extreme distributions to the observed "storm" $Hm0$ values, or by extrapolation of existing annual maximum $Hm0$ values.

In this study we simply generated a 5000 year $Hm0$ sequence and kept a record of the annual maximum $Hm0$ values. The distribution of these annual maxima values is shown in Figures 9.D and 9.E.

In certain engineering applications it is useful to have an estimate of the "100 year" $Hm0$ value and this and other "T year" $Hm0$ values can be found by the following:

$F(x_T)$ defined to be the distribution function of the annual $Hm0$ maximum values. Then the "T year $Hm0$", x_T is the solution to

$$F(x_T) = 1 - (\frac{1}{T})$$

and therefore for the "100 year $Hm0$"

$$F(x_{100}) = \frac{99}{100}$$

$$x_{100} = F^{-1}\left(\frac{99}{100}\right)$$

Some of the resulting values are given in Table 9.1. These values are in close accordance with those estimated by Mr. J. Rossouw (1988) using various estimation techniques. Conservatively he has estimated the '10 year' $Hm0$ to be 10.0 m and the '100 year' $Hm0$ to be 12.0 m.

TABLE 9.1

"T YEAR $Hm0$ VALUES

"YEAR $Hm0$"	$Hm0$ (m)
10	9.630
25	10.771
50	11.231
100	11.734
500	12.813
1000	13.183
3000	13.235
5000	13.300

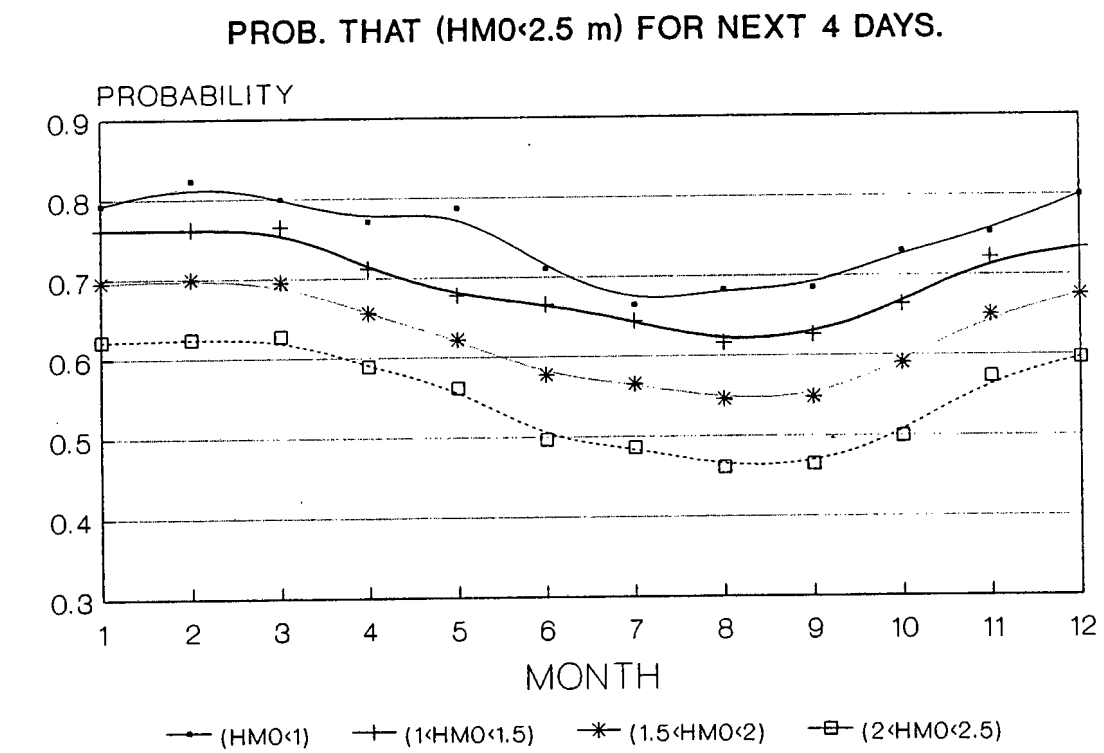


FIGURE 9.C HM0 STARTING VALUES.

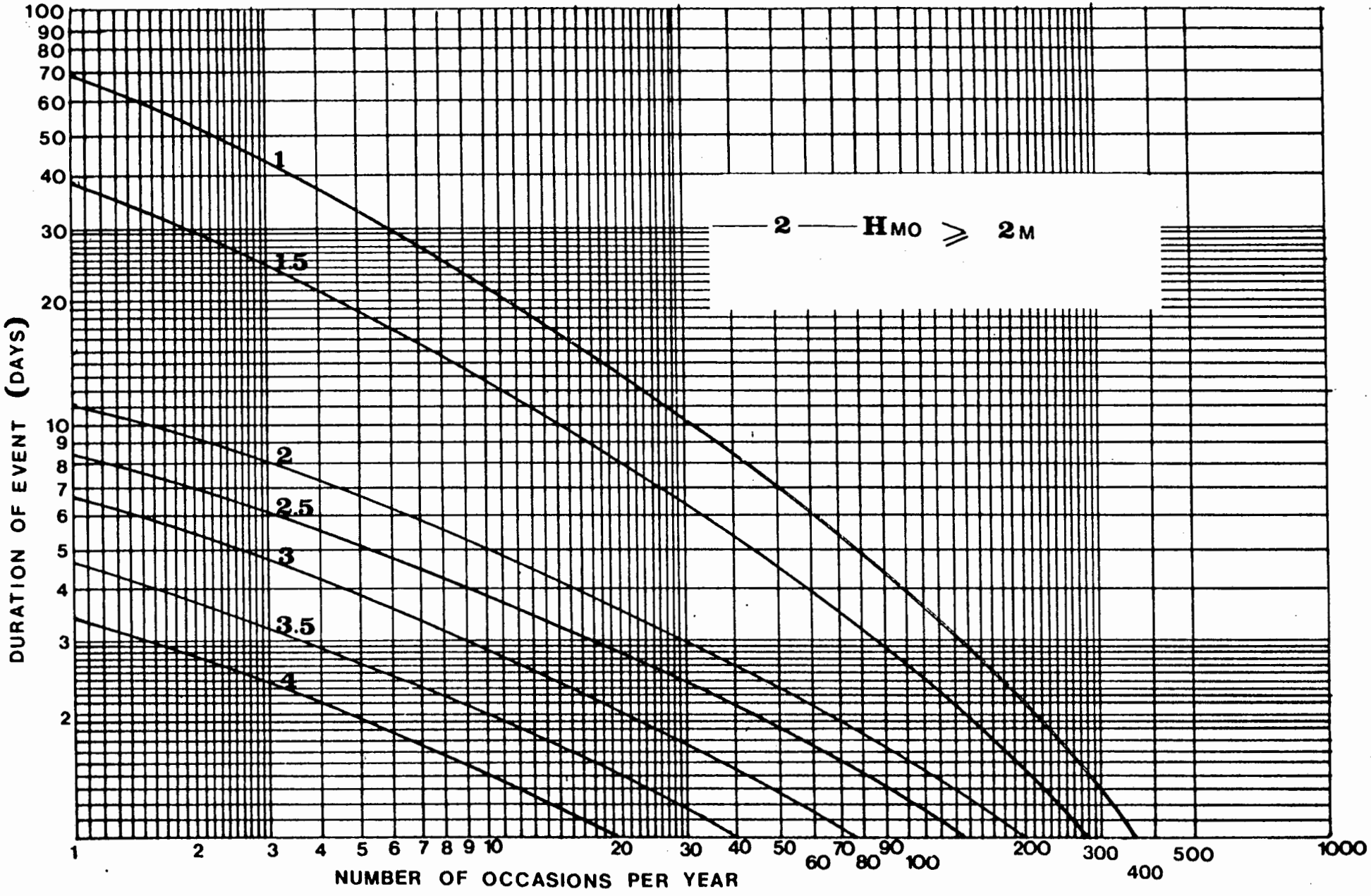


Figure 9.A **Persistence of Storms**

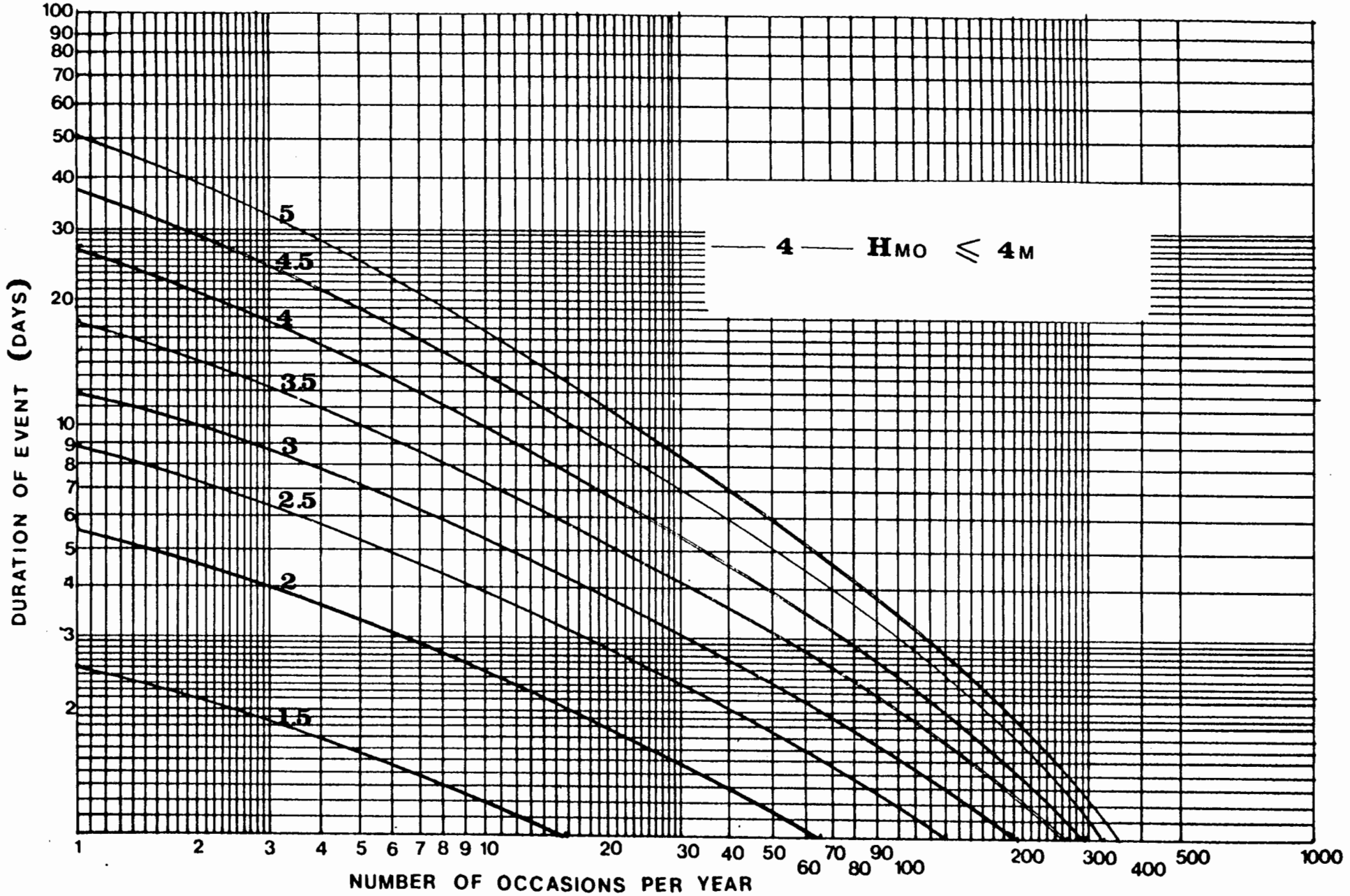


Figure 9.B **Persistence of Calms**

ANNUAL MAXIMUM HMO.
5000 YEARS OF GENERATED HMO.

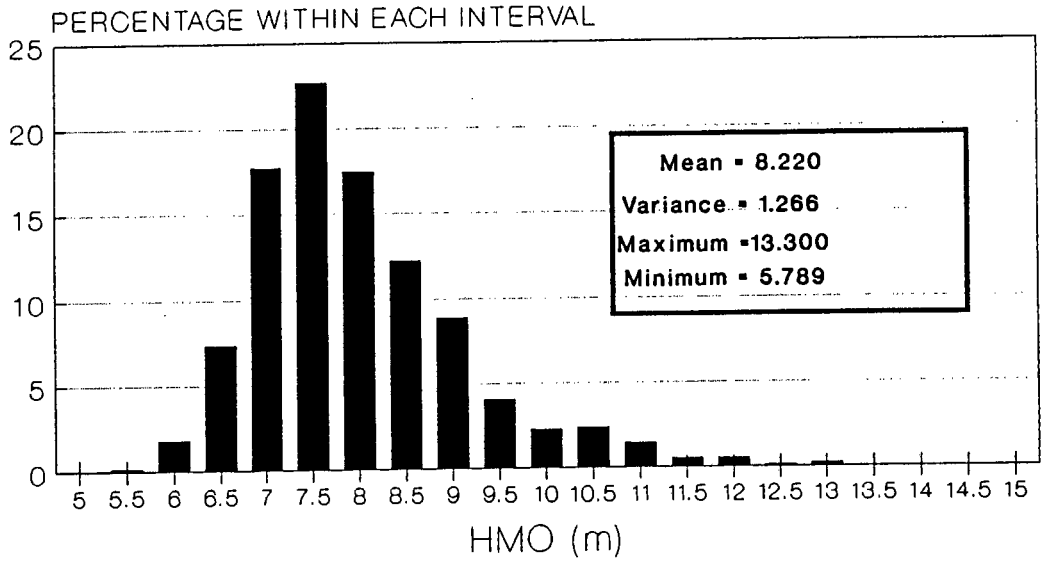


FIGURE 9.D

ANNUAL MAXIMUM HMO.
5000 YEARS OF GENERATED HMO.

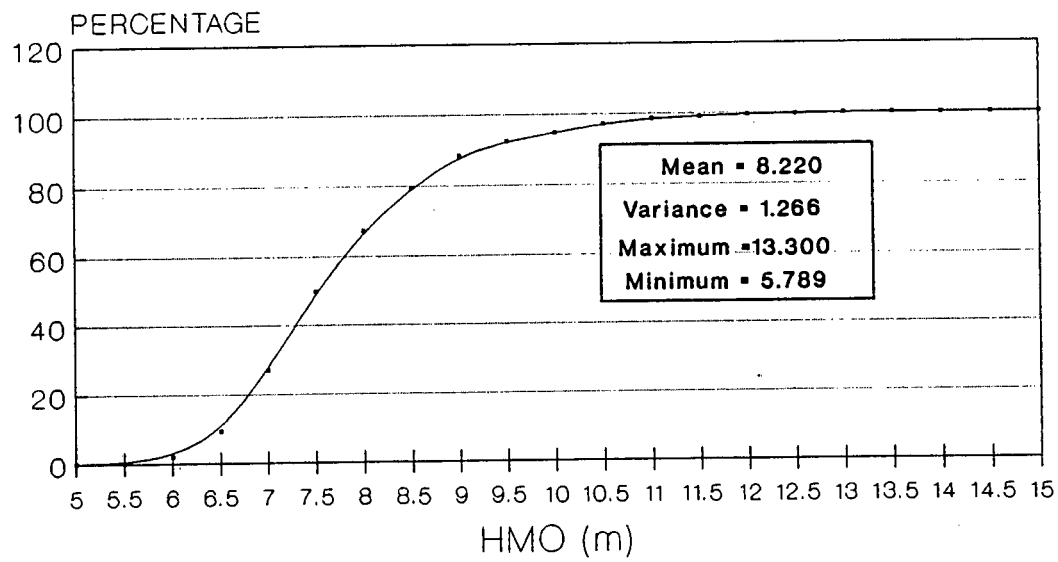


FIGURE 9.E

CHAPTER 10

THE JOINT DISTRIBUTION OF $Hm0$ AND Tz

In this chapter we describe two of the attempts which we made to model the joint distribution of $Hm0$ and Tz .

We note that since $Hm0 = 4 * \sqrt{m0}$ and $Tz = \sqrt{\frac{m0}{m2}}$ it is sufficient to model the joint distribution of $m0$ and $m2$, or alternatively $\ln m0$ and $\ln m2$. Now we have already constructed models for the marginal distributions of $\ln m0$ and $\ln m2$. If $\ln m0$ and $\ln m2$ were approximately independently distributed then their joint distribution would simply be given by the product of their marginals. However there is a substantial correlation (approximately 0.8) between these two series. The estimated correlation coefficient between $Hm0$ and Tz is approximately 0.4. Figures 10.A and 10.B are the bivariate scatter plots of $Hm0, Tz$ and $m0, m2$ respectively and give an indication of the correlation between each of the series. Even when we consider the correlation coefficient between the residuals of the $\ln m0$ and $\ln m2$ from their respective models, (which is the proper measure of cross correlation between two time series), we obtain an estimate of 0.7. Clearly these two series are not independently distributed.

The problem of modelling the joint distribution of the two time series is particularly difficult in our case since, as was pointed out in previous chapters the distributions involved are already rather complicated. In particular the distribution of the residual process of each of the series $\ln m0$ and $\ln m2$ are non-standard.

We considered several approaches to this problem. This chapter discusses the two which came closest to providing an acceptable model. Neither of these approaches are entirely satisfactory in that neither led to a model that preserved all the important properties of the data.

Indirect modelling of the joint distribution.

One approach to finding a model for the joint distribution of $\ln m0$ and $\ln m2$ is to model the bivariate distribution of their residuals. Since these residuals are supposed to be serially uncorrelated the problem is reduced to one of constructing a suitable bivariate distribution whose marginals conform to those which have been fitted to the individual residual series.

The problems associated with finding bivariate distributions with given marginals is well documented in the statistical literature (see Marshall and Olkin (1985), Stein et al. (1987) and Pei-Ling Lui et al. (1986)).

One of the problems is that there is often no unique or even 'natural' bivariate version of a given distribution. For even relatively simple distributions it is not always possible to obtain convenient expressions for estimators of the parameters (of the joint distribution). The marginal distributions with which we are dealing are particularly complicated in this respect and it is not clear how one could go about constructing a bivariate extension of them. The approach which we adopted is one of indirect construction. This approach is also suggested in Pei-Ling Lui et al. (1986).

Suppose that we wish to construct a bivariate distribution for the random variables X_1 and X_2 which is such that their marginal distribution functions are to be F_1 and F_2 respectively. An indirect way of doing this is to use a pair of random variables from some convenient distribution and then to transform these so that the marginal distributions have the required form. For example suppose that Z_1 and Z_2 are distributed according to the standard bivariate normal distribution with correlation coefficient ρ, i.e.

$$(Z_1, Z_2) \sim BN(0;0;1;1;\rho)$$

If we then define

$$X_1 = {F_1}^{-1}(\Phi(Z_1)) \quad \text{and} \quad X_2 = {F_2}^{-1}(\Phi(Z_2))$$

(where Φ is the distribution function of the standard univariate normal), then X_1 and X_2 have the required marginal distributions and are correlated. Their joint distributions might be quite complicated, or even intractable, but this does not present an insurmountable problem in applying this type of construction.

The problem that does need to be solved is that of estimating ρ, the parameter which gives the correlation between Z_1 and Z_2 and which indirectly determines the correlation between X_1 and X_2. We note that the correlation between X_1 and X_2 will not in general be equal to ρ.

Since the joint distribution function of X_1 and X_2 is intractable in our particular case where F_1 and F_2 are both rather complicated, there is no direct way of estimating ρ.

However we can note that each value of ρ gives rise to a corresponding value of the correlation coefficient between X_1 and X_2, say ρ^*. We can estimate ρ, using the method of moments, by finding that value of ρ which corresponds to the observed correlation coefficient between X_1 and X_2, namely $\widehat{\rho^*}$. The latter is computed from the data in the usual way. The corresponding estimate of ρ has to be found by Monte Carlo methods.

Essentially one needs a table relating ρ to ρ^*, and we give this in an abbreviated form, (see Tables 10.1 and 10.2).

Once ρ has been estimated we have in effect fitted a bivariate model to the random variables X_1 and X_2 which has the prescribed marginal distributions, namely F_1 and F_2.

Although the joint distribution of X_1 and X_2 is intractable it is very easy to generate variates from the joint distribution.

The procedure used to accomplish this was as follows:

1. We require an initial ρ correlation value.

2. Generate two independent $N(0;1)$ random variables: a_1 and a_2.

3. Use the Cholesky algorithm, (see Acton (1970)), to transform the independent $N(0;1)$ random variables into correlated $N(0;1)$ random variables, (b_1 and b_2), by setting:

$$b_1 = a_1 \text{ and } b_2 = (\rho * a_1) + ((\sqrt{1-\rho^2}) * a_2)$$

(We now have correlated $N(0;1)$ random variables b_1 and b_2 and can now transform them to random variables from the $ELPHN_{k=2}$ distribution if using Model D else continue).

4. Using the Standard Normal Distribution Function we set

$$c_1 = \Phi(b_1) \text{ and } c_2 = \Phi(b_2)$$

5. Then let

$$d_1 = F^{-1}(c_1) \quad \text{and} \quad d_2 = F^{-1}(c_2)$$

where $F^{-1}(.)$ is the Inverse Distribution Function for distributon of the residual terms that are being generated, i.e. our prescribed marginals.

The above estimation procedure was applied to estimate the values of ρ^* for the residuals of the bivariate distribution of the $\ln m0$ and $\ln m2$. The estimators for Models C and D are given in Table 10.1.

Having fitted a joint model to the residuals of the process, we are now in a position to generate artificial values for this model. These can be used to assess the fit of the joint model to the bivariate ($\ln m0, \ln m2$) and, particularly, to the bivariate $(Hm0, Tz)$ series.

TABLE 10.1 : Correlation Coefficient Values.

	OBSERVED	MODEL C			MODEL D		
INITIAL ρ	-	0.66	0.77	0.88	0.66	0.77	0.88
RESIDUAL ρ^*	-	0.64	0.75	0.87	0.66	0.77	0.87
$\ln(m0), \ln(m2)$	0.80	0.62	0.74	0.86	0.63	0.74	0.82
$Hm0, Tz$	0.39	0.52	0.46	0.39	0.34	0.24	0.048
Tz Variance	2.01	8.10	5.19	2.79	6.50	4.10	2.10

Summary of results given in above tables:

Model C: When using this model we can generate correctly correlated $Hm0, Tz$ terms and a Tz variance that although too large, is moving in the correct direction. The generated $Hm0$ values though, as discussed earlier, are slightly unsatisfactory and therefore the generated $Hm0, Tz$ distribution could not be used to good effect. (Recall that the Model C generated $Hm0$ values have a variance that is slightly too large and extreme values that are definately too large). If it were not for this fact the $Hm0, Tz$ bivariate distribution would be adequate.

TABLE 10.2: Correlation Coefficient Values.

	OBSERVED	MODEL C-D		
INITIAL ρ	-	0.66	0.77	0.88
RESIDUAL ρ^*	-	0.65	0.75	0.86
$\ln(m0), \ln(m2)$	0.80	0.62	0.74	0.85
$Hm0, Tz$	0.39	0.32	0.23	0.11
Tz Variance	2.01	7.60	5.20	3.15

Model D: This model performs well for $Hm0$ generation but not for $Hm0, Tz$. To get the correct $Hm0, Tz$ correlation value the variance of the generated Tz values is too large, and correspondingly getting the correct Tz variance means having a $Hm0, Tz$ correlation that is unacceptably small.

The next approach was to use the $Hm0$ generated using Model D, (for $m0$), and Model C for the $m2$ generation, since generated $Hm0$ is satisfactory when using Model D and Tz generation is "better" when using Model C, and we call it Model C-D. As can be seen from the results in Table 10.2 this arrangement is also unsatisfactory for the same reasons, namely that we obtain both the correct $Hm0, Tz$ correlation coefficient and a Tz variance of the correct magnitude but we cannot do so simultaneously.

Conditional approach.

Another approach attempted but which failed was to model the conditional distribution of the Tz given $Hm0$. We describe this attempt for completeness.

We mentioned the the models for $Hm0$ which we fitted in Chapter 5 and try to describe the conditional distribution of Tz for different values of $Hm0$. In effect the time series behaviour of Tz, including its serial correlation and seasonal structures, is only indirectly maintained by association with $Hm0$.

The strategy adopted to model the dependence of the conditional distribution of Tz given $Hm0$ was to first select a distributional form for the Tz and then to

express the parameters of the distribution as simple functions of $Hm0$.

We first divide the Tz values into those corresponding to certain intervals of $Hm0$ values, see Figures 10. C; D; E; F; G and H. Except for small $Hm0$ values (where the Tz values vary widely) the distribution of the Tz values have much the same shape. An increase in the mean and a decrease in variance of the Tz values was noted as the $Hm0$ values increased.

The distributional form that we selected was that of a $N(0;1)$ target using the Elphinstone (1985) method. In this application we also restrict k to be $k = 1$.

Using $N(0;1)$ target we have

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(g_k(x;\theta))^2) P_k(x;\theta)$$

And for k=1 we have:

$$P_k(x;\Lambda) = [\lambda - 2\lambda\lambda_{11}x + \lambda{\lambda_{11}}^2x^2 + \lambda\lambda_{12}x^2]$$

$$g_k(x;\theta) = (\xi + \lambda x - \lambda\lambda_{11}x^2 + \frac{1}{3}\lambda{\lambda_{11}}^2x^3 + \frac{1}{3}\lambda\lambda_{12}x^3)$$

$$= [\xi + \lambda x + (-\lambda\lambda_{11})x^2 + (\frac{1}{3}\lambda{\lambda_{11}}^2 + \frac{1}{3}\lambda\lambda_{12})x^3]$$

Figures 10. I; J; K and L give the parameter estimates for the above distribution for different $Hm0$ intervals, (we let Parameter 1 = ξ , Parameter 2 = λ , Parameter 3 = λ_{11} and Parameter 4 = λ_{12} .

This approach was not particularly successful since firstly the small parameter distribution did not fit the observed Tz data sufficiently well and secondly the parameters do not show a smooth fluctuation between each $Hm0$ interval. We therefore considered it impractical to continue with this approach.

In conculsion to the chapter we note that the construction of a joint model for $Hm0, Tz$ is difficult. This is because there are so many aspects of the data which need to be preserved using a small number of parameters.

We found that it is possible to tune the parameters of the model so as to preserve selected properties but were not able to preserve them all simultaneously.

The particular model we fitted is satisfactory in a number of respects but is inadequate in preserving all the properties.

BIVARIATE (Hm0,Tz) PLOT

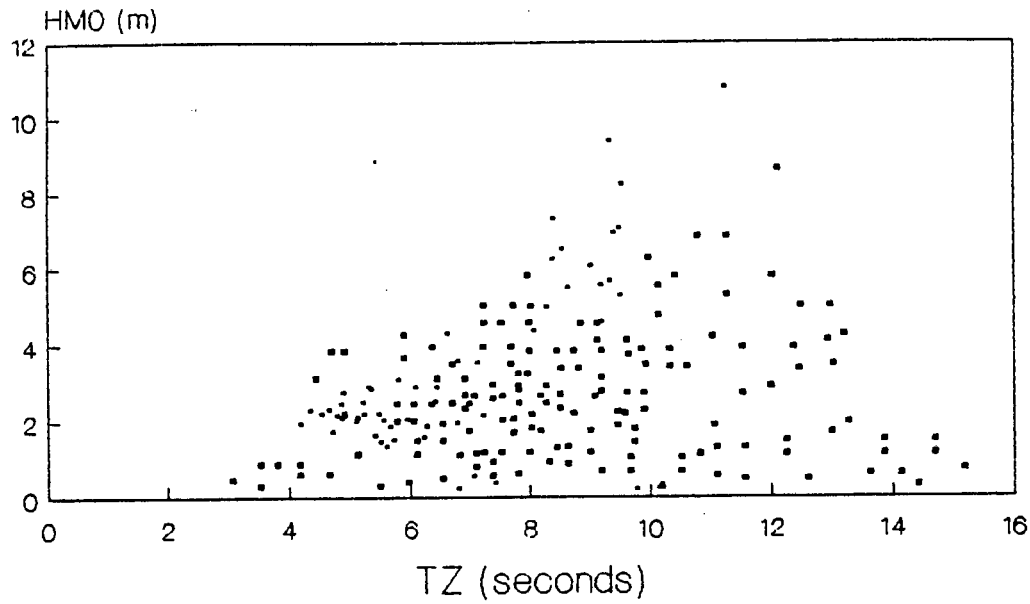


FIGURE 10.A

BIVARIATE (m0,m2) PLOT.

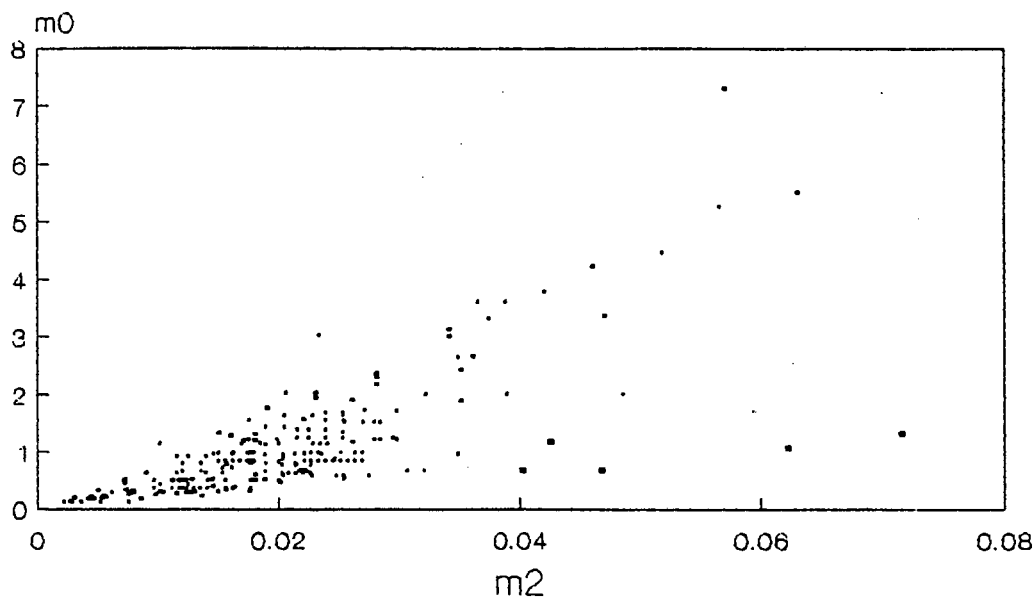


FIGURE 10.B

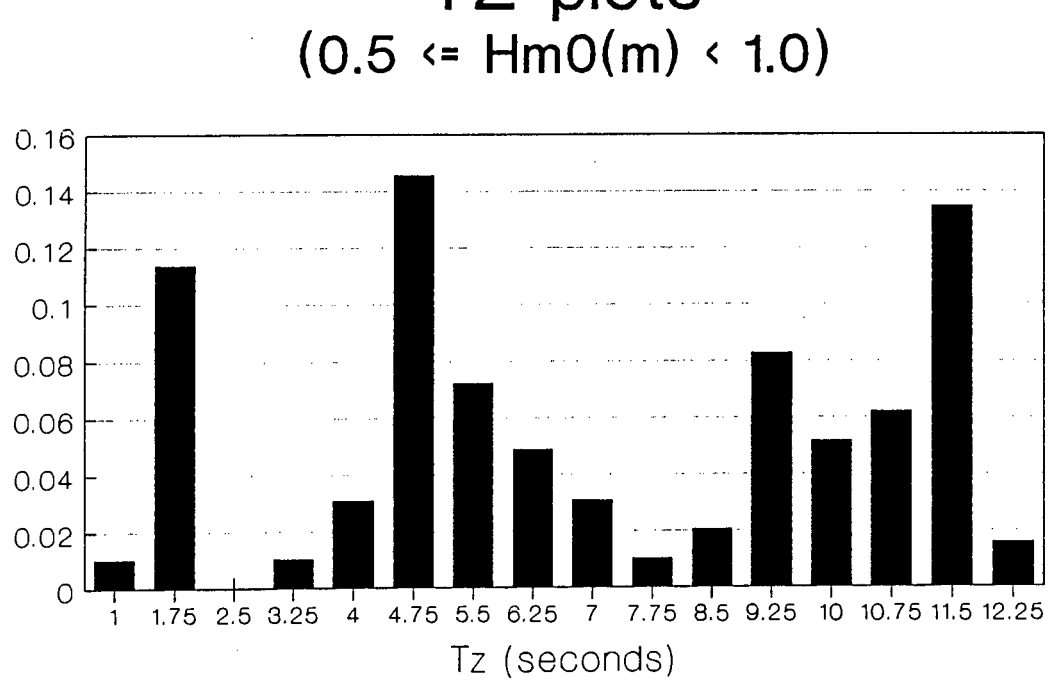


Figure 10.C

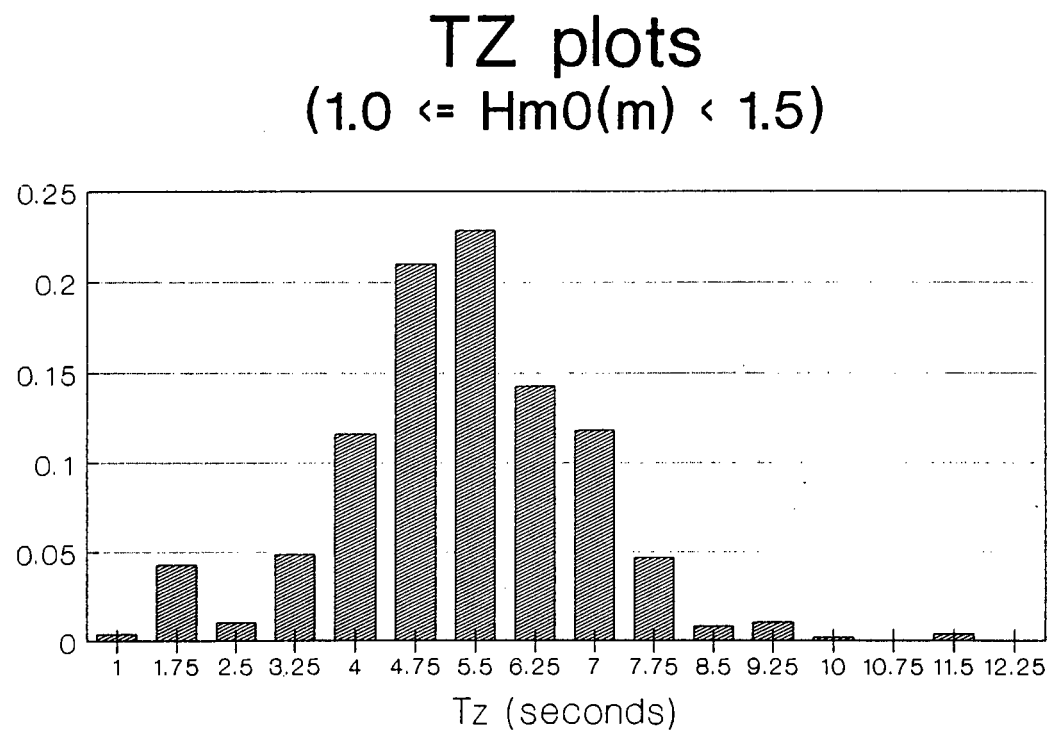


Figure 10.D

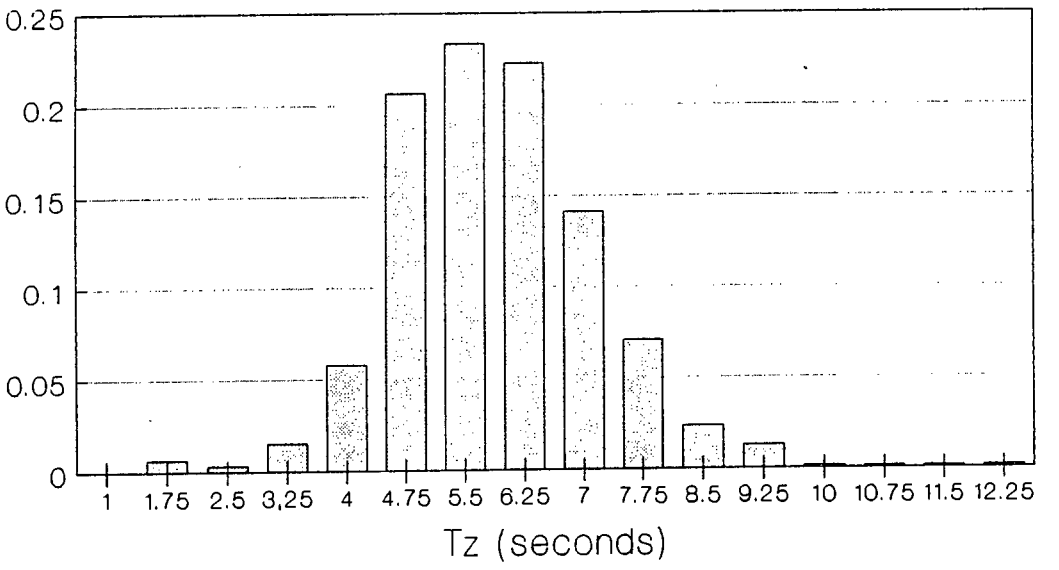


Figure 10.E

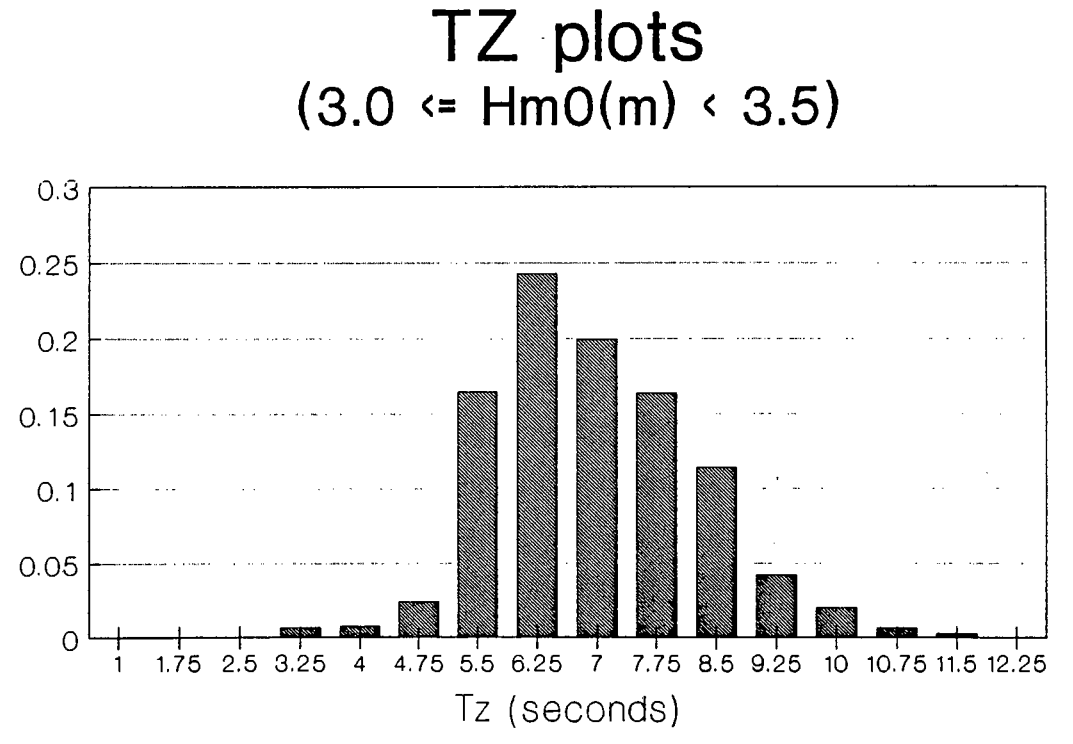


Figure 10.F

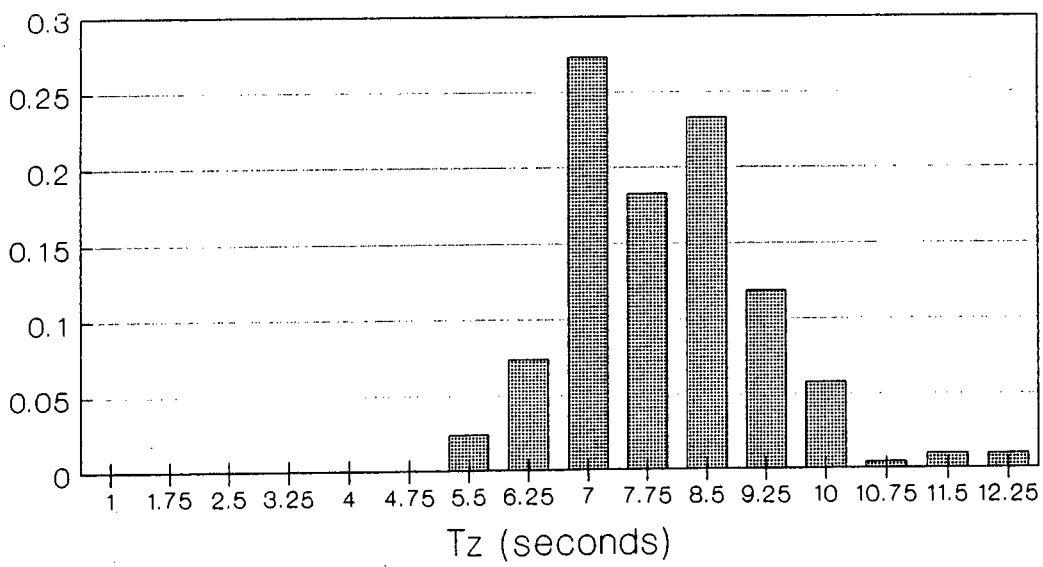


Figure 10.G

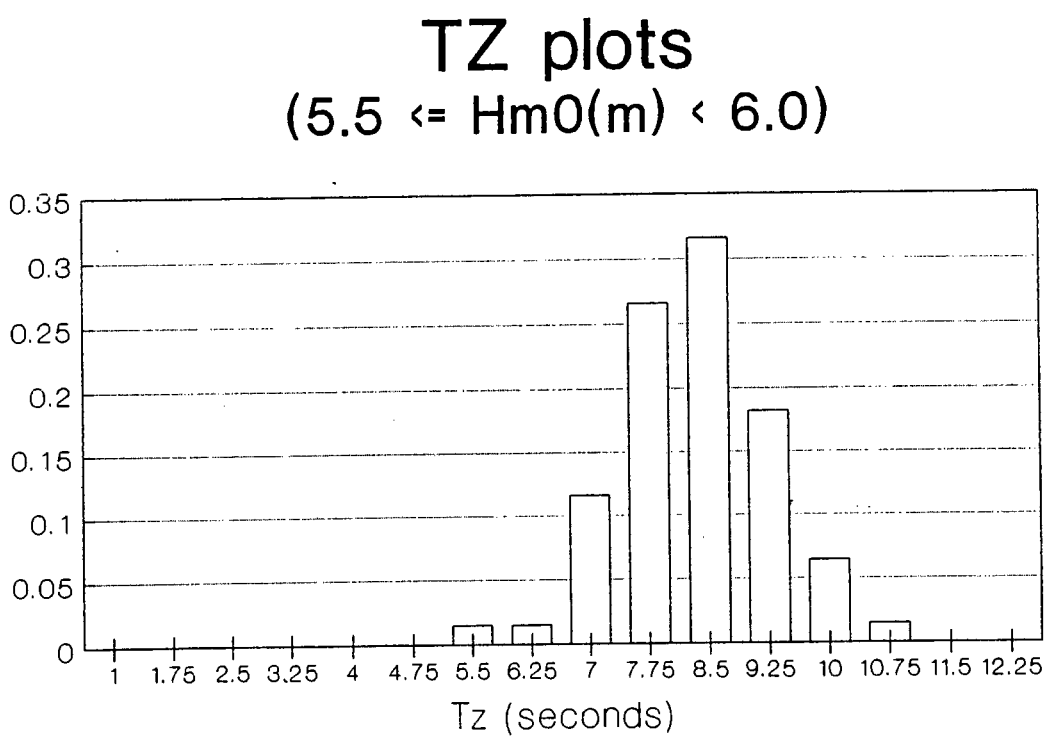


Figure 10.H

PARAMETER PLOT

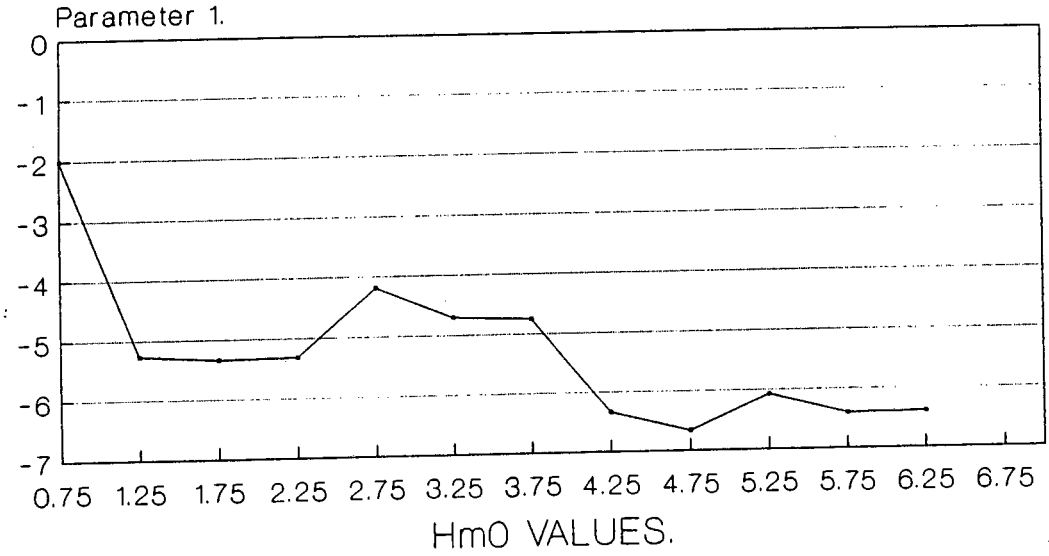


FIGURE 10.I

PARAMETER PLOT

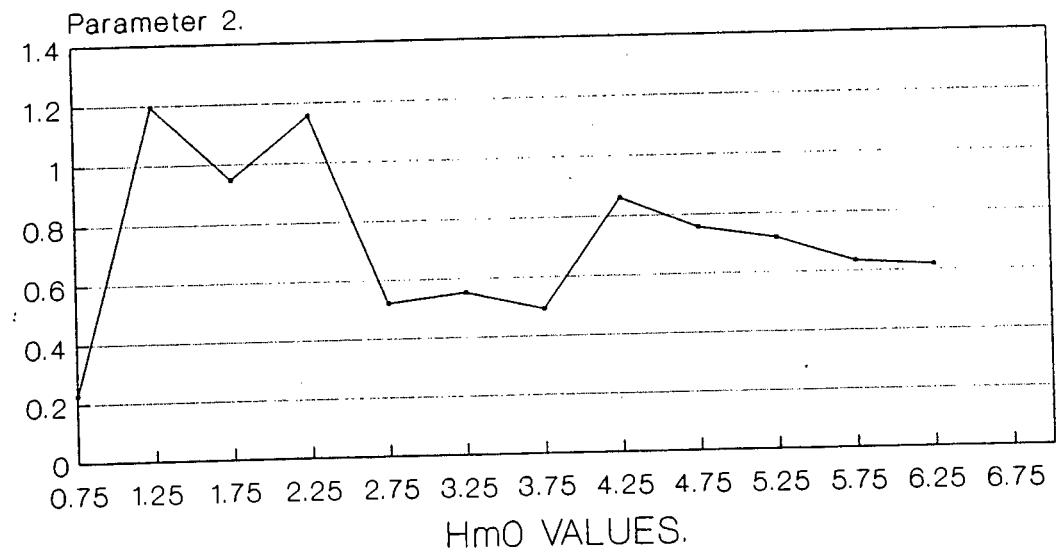


FIGURE 10.J

PARAMETER PLOT

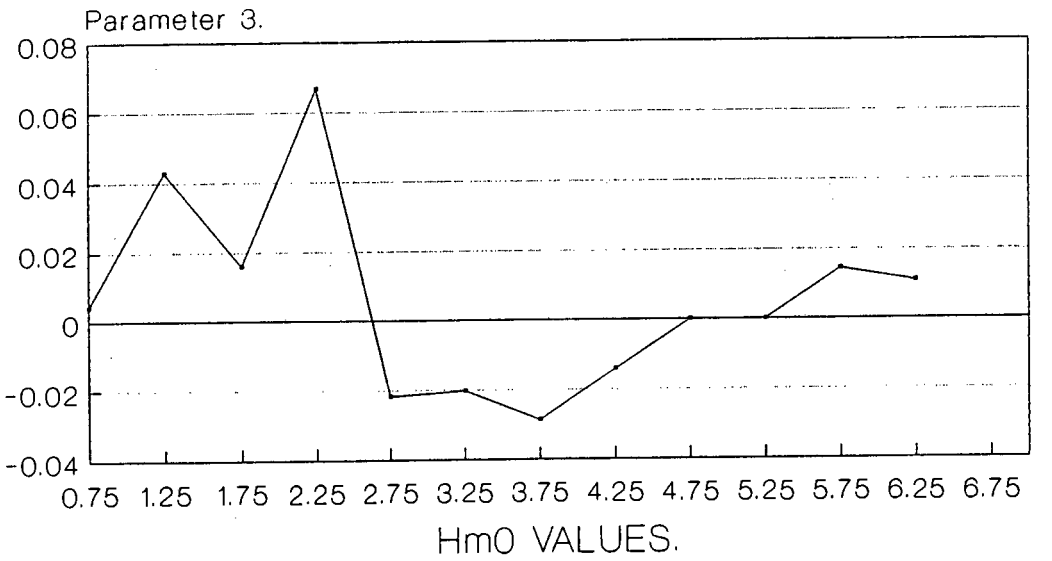


FIGURE 10.K

PARAMETER PLOT

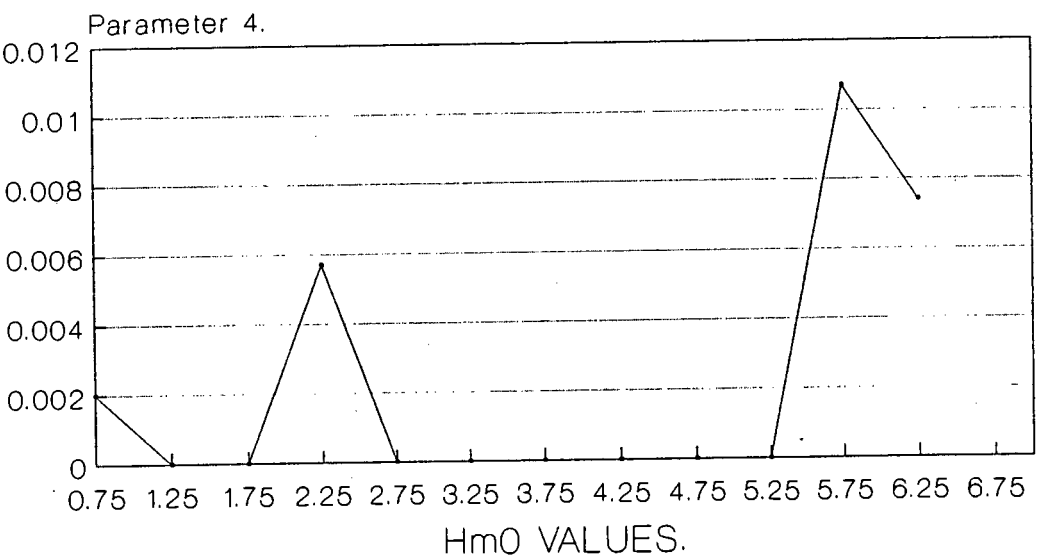


FIGURE 10.L

CHAPTER 11.

CONCLUSION.

We have shown that, at least for the particular data series considered here, the $Hm0$ process can be usefully modelled as a time series. We actually modelled the $\ln m0$ and $\ln m2$ processes since the distribution of the residual terms was approximately symmetric when using this transformation. The time series approach allowed us to incorporate the substantial seasonal and serial correlation components into the model.

The modelling of the residual process proved to be more difficult and we considered several models in this regard. Four were investigated in detail and after having fitted the models to the observed data the relative merits of the models were assessed. No one model was found to be optimal in terms of all the criteria which were relevant in this application. We tentatively selected a model based on the criteria we considered the most important. This model was then used to generate artificial $Hm0$ sequences from which various quantities of interest were estimated.

The attempt to model the bivariate $(Hm0, Tz)$ process as a bivariate time series proved to be less successful. Although we found it possible to model the two time series individually, and even to preserve the average cross correlation structure between them, we were unable to preserve all the important properties of the observed bivariate process simultaneously. Thus the problem of modelling the bivariate $(Hm0, Tz)$ process still requires attention. One approach to this problem would be to model $(Hm0, Tz)$ directly, rather than via the bivariate $(m0, m2)$ process. This would involve the modelling of the asymmetric bivariate error distribution. Here one would have to find a family of flexible bivariate distributions, perhaps a bivariate extension of the methods of Elphinstone (1985).

REFERENCES.

ACTON, F.S. (1970). *Numerical methods that work,* Harper and Row, New York.

BENDAT, J.S. and PIERSON, A.G. (1971). *Random Data: Analysis and Measurement Procedure,* John Wiley & Sons, New York.

BLOOMFIELD, P. (1976). *Fourier Analysis of Time Series: An Introduction,* John Wiley & sons, New York.

BOX, G.E.P. and JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control,* Holden-Day, San Fransisco.

BRANDAO, A. (1987). A stochastic model for daily climate, Unpublished M.Sc. thesis, University of Cape Town.

CARTER, D.J.T., CHALLENOR, P.G., EWING, J.A., PITT, E.G., SROKOSZ, M.A., TUCKER, M.J. (1986). Estimating wave climate parameters for engineering applications. *Offshore Technology Report. Oth 86 288,* The Institute for Oceanographic Sciences, Surrey, England.

CHAKRABARTI, S.K. and COOLEY, R.P. (1977). Statistical distribution of periods and heights of ocean waves. *Journal of Geophysical Research,* **82, No. 9,** 1363-1368.

COLLINS, J.I. (1967). Wave statistics from hurricane Dora. *Journal of Waterways and Harbours Div., American Society of Civil Engineers,* **93, WW2,** 59-77.

DRAPER, N. and SMITH, H. (1981). *Applied Regression Analysis,* John Wiley & Sons, New York.

ELPHINSTONE, C.D. (1983). A target distribution model for nonparametric density estimation. *Communications in Statistics A, Theory and Methods,* **12, No. 2,** 161-198.

ELPHINSTONE, C.D. (1985). A Method of distribution and density estimation. *Technical Report TWISK 421,* National Research Institute for Mathematical Sciences, CSIR.

FORRISTAL, G.Z., (1978). On the statistical distribution of wave heights in a storm. *Journal of Geophysical Research,* **83, C5,** 2353-2358.

GODA, Y. (1974 a). Estimation of wave statistics from spectral analysis. *Proceedings of International Symp. on Ocean Wave Measurement and Analysis,* **1** 320-337

GODA, Y. (1974 b). Investigation of the statistical properties of sea waves with field and simulation data. *Report of the Port and Harbour Res. Inst.,* **13, No. 1** 3-37.

GOODNIGHT, R.C. and RUSSEL, T.L. (1963). Investigation of the statistics of wave heights. *Journal of Waterways and Harbours Div., American Society of Civil Engineers,* **89, WW2,** 29-55.

HUANG, N.E. and LONG, S.R., (1980). An experimental study of the surface elevation probability distribution and statistics of wind-generated waves. *Journal of Fluid Mechanics,* **101, Part 1,** 179-200.

JOHNSON, N. and KOTZ, S. (1970). *Continuous Univariate Distributions 1 & 2,* John Wiley & Sons, New York.

LARSON, H.J. (1982). *Introduction to Probability Theory and Statistical Inference,* John Wiley & Sons, New York.

LEADBETTER, M.R., LINDGREN, G. and ROOTZEN, H. (1983). *Extremes and related properties of random sequences and processes,* Springer-Verlag, New York.

LONGUET-HIGGINS, M.S., (1952). On the statistical distribution of the heights of sea waves. *Journal of Marine Research,* **9, No. 3,** 245-266.

LONGUET-HIGGINS, M.S., (1957). The statistical analysis of a random moving surface. *Phil. Trans. Royal Society London,* **A 249,** 321-387.

LONGUET-HIGGINS, M.S., (1975). On the joint distribution of the periods and amplitudes of sea waves. *Journal of Geophysical Research,* **80,** 2688-2694.

LONGUET-HIGGINS, M.S., (1980). On the distribution of the heights of sea waves: some effects on nonlinearity and finite band width. *Journal of Geophysical Research,* **85,** 1519-1523.

LONGUET-HIGGINS, M.S., (1983). On the joint distribution of wave periods and amplitudes in a random noise. *Procedings Royal Society London,* **A 389,** 241-258.

MARSHAL, A.W. and OLKIN, I. (1985). A family of bivariate distributions generated by the bivariate Bernoulli distribution. *Journal of the American Statistical Association,* **80,** 332-338.

MOSTELLER, F. and TUKEY, J. (1977). *Data Analysis and Regression: A second course in Statistics,* Addison-Wesley, Reading Mass.

OTNES, R.K. and ENOCHSON, L. (1978). *Digital Time Series Analysis,* John Wiley & Sons, New York.

PEI-LING LIU and DER KIUREGHIAN, A. (1986). Multivariate distribution models with prescribed marginals and covariances. *Probabilistic Engineering Mechanics,* **1, No. 2,** 105-112.

RICE, S.O., (1944 & 1945). The mathematical analysis of random noise. *Bell Systems Technical Journal,* **23 & 24,** (282-332) & (46-156)

ROSSOUW, J. (1984). Review of existing wave data, wave climate and design waves for South African and South West African (Namibian) coastal waters. *CSIR Report T/SEA 8401,* National Research Institute for Oceanology, Stellenbosch, South Africa.

ROSSOUW, J. (1988). Personal Communication. Department of Coastal Engineering, University of Stellenbosch.

STEIN, G.Z., ZUCCHINI, W. and JURITZ, J.M. (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association,* **82, No. 399,** 938-944.

SVERDRUP, H.U. and MUNK, W.H. (1947). Wind, sea and swell: Theory of relations for forecasting. *U.S. Navy Hydrographic Office Publication,* **601**

TAYFUN, M.A., (1980). Narrow band nonlinear sea waves. *Journal of Geophysical Research,* **85 ,C3** 1548-1552.

YAMAZAKI, H. and HERBICH, J.B., (1985). Nonparametric and parametric estimation of wave statistics and spectra. *Ocean Engineering Program,* Civil Engineering Department, Texas University.

YAMAZAKI, H. and HERBICH, J.B., (1985). Determination of wave height spectrum by means of a joint probability density function. *Journal of Geophysical Research* **90, C2** 3381-3390.

ZUCCHINI, W and ADAMSON, P.T. (1984). The occurrence and severity of droughts in South Africa. *WRC Report No. 91/1/84,* Water Research Commission, Pretoria.

APPENDIX

Appendix A.

ROUTE OF DATA FROM OCEAN TO SPECTRAL MOMENTS.

This section gives a basic overview of the method used by National Research Institute of Oceanology (NRIO) to convert the wave-rider buoy data to the format that was used in this study, including the variables $Hm0$ and Tz. This algorithm is outlined for completeness; this study was based on the estimates of $Hm0$ and Tz supplied by NRIO.

The raw data consists of a sea surface elevation value taken every 0.5 seconds for a 20 minute period, this is taken as as one record, and a record is taken once every 6 hours where possible.

A record therefore contains 2400 data values. Each record takes the following route:

A search for outliers is carried out and any outliers found are replaced with interpolated values. This is followed by a test for any significant trend in the data and if a trend is found it is removed.

A taper is carried out on the first and last ten percent of the data with a cosine bell taper and an estimate of the power spectra is then obtained using Fast Fourier Transform (FFT) computations.

The moments of the power spectral density estimate and various summary statistics of the record are calculated.

The details of the above stages are now given.

If the absolute difference between two consecutive measurements is greater than a predefined value then the data value is replaced with an interpolated value. Starting and stopping problems sometimes occur with the wave rider buoy recording device. This can lead to apparent trends which may need to removed. A test for a significant trend in the record is thus performed. In this context trend is defined as any frequency whose period is longer than the record length. Trend removal is an important intermediate step in the digital processing of random data and should be given due consideration. If trends are not eliminated then large distortions can occur in the later processing of correlation and spectral quantities. In particular, trends

can completely nullify the estimation of low frequency spectral content. Thus, if there is a significant trend it is removed.

A smooth filter shape for FFT estimates to reduce leakage can be obtained by tapering the original random time series at each end. A 10 percent cosine taper was used at each end the data. The effect of tapering is to reduce the variance of tapered data relative to the original data. The purpose of tapering when viewed from its frequency domain is to supress large side lobes in the effective filter obtained with the raw transform. When viewed from the time domain, the object of tapering is to "round off" potential discontinuities at each end of the finite segment of the time history being analysed. The Fast Fourier Transform (FFT) algorithm is then used to compute estimates of power spectral density functions directly from the tapered data values. In principle, any sample size N can be handled, but in practice most programs are designed for digital records of length $N = 2^p$, where p is some integer. Hence data sequences must either be truncated or have zeros added to obtain the required number of data points. In the equations that follow, it is convenient to let $x(t)$ be defined over the time interval $(\frac{-T}{2}, \frac{T}{2})$. Then the finite range FFT folmula

$$X(f,T) = \int_0^T x(t) \exp(-j2\pi ft)\, dt$$

can be viewed as a transformation of an infinitely long record $y(t)$ defined over $(-\infty, \infty)$, multiplied by a finite length boxcar function $U_{T/2}(t)$ defined over $(-\frac{T}{2}, \frac{T}{2})$. That is

$$\begin{aligned} X(f,T) &= \int_{\frac{-T}{2}}^{\frac{T}{2}} x(t) \exp(-j2\pi ft)\, dt \\ &= \int_{-\infty}^{\infty} y(t) U_{\frac{T}{2}}(t) \exp(-j2\pi ft)\, dt \end{aligned}$$

where $y(t)$ is the same as $x(t)$ in the range $(-\frac{T}{2}, \frac{T}{2})$ and

$$\begin{aligned} U_{T/2}(t) &= 0 \qquad t < -\frac{T}{2} \\ &= 1 \qquad \frac{-T}{2} \le t \le \frac{T}{2} \\ &= 0 \qquad t > \frac{T}{2} \end{aligned}$$

The Fourier transform of $U_{T/2}(t)$ is given by

$$U_{\frac{T}{2}}(f) = T\left(\frac{\sin(\pi fT)}{\pi fT}\right)$$

and has its first zero crossing when $f = \pm\frac{1}{T}$, and represents the effective raw filter shape for FFT estimates.

Power Spectrum Estimates.

For a single record $x(t)$, a raw estimate of the power spectral density function at any frequency f is given by the formula

$$\tilde{G}_x(f) = \frac{2}{T}|X(f,T)|^2 \qquad (a)$$

Here, $T = Nh$ and then

$$X(f,T) = h\sum_{n=0}^{N-1} x_n \exp(-j2\pi fnh) \qquad (b)$$

At the usual FFT discrete frequency values

$$f_k = \frac{k}{T} = \frac{k}{Nh} \qquad k = 0,1,2,\ldots,k-1$$

the Fourier components are

$$\begin{aligned} X_k &= \frac{X(f_k,T)}{h} \\ &= \sum_{n=0}^{N-1} x_n \exp\left(\frac{-j2\pi kn}{N}\right) \end{aligned} \qquad (c)$$

Hence, from equations (a) and (c) , the power spectrum estimates becomes

$$\begin{aligned} \tilde{G}_k = \tilde{G}_x(f_k) &= \frac{2}{Nh}|X(f_k,T)|^2 \\ &= \frac{2h}{N}|X_k|^2 \end{aligned} \qquad (d)$$

The steps carried out during this stage of the data analysis can be summarized as follows: The data is truncated or patched with zeros so as to obtain a record length which is an integer power of 2. Secondly the resulting sequence is tapered and the X_k values of equation (c) are calculated for $k = 0,1,\ldots,N-1$. The $\tilde{G}_k$ of equation (d) for $k = 0,1,...,N-1$. Finally these estimates are adjusted for the scale factor due to tapering, e.g. by replacing $\tilde{G}_k$ by $\left(\frac{1}{0.875}\right)\tilde{G}$ since the cosine tapering was used.

If the spectrum is of a bandwidth limited white noise nature, then estimates at a frequency spacing of $\frac{1}{T}$ will be essentially uncorrelated. Hence if l neighbouring

frequency components of the spectral estimates are averaged, then the final smooth spectral estimate $\widehat{G}_k$, where the $\widehat{G}_k$ replaces the $\widetilde{G}_k$, is given by

$$\widehat{G}_K = \frac{1}{l}[\widetilde{G}_k + \widetilde{G}_{k+1} + \widetilde{G}_{k+2} + \ldots + \widetilde{G}_{k+l-1}]$$

The moments mi, $i = 0, \ldots, 4$ (where i is the i^{th} moment of the power spectral density) are found and various statistics of the particular wave record are calculated as functions of these moments, for example $Hm0 = 4\sqrt{m_0}$ and $Tz = \sqrt{\frac{m_0}{m_2}}$.

The program described above was not used in this study since we recieved our data in $Hm0$ and Tz form. This appendix is simply given for completeness.

Appendix B

WEIGHTED LEAST SQUARES METHODS.

It sometimes happens that some of the observations used in a regression analysis are "less reliable" than others. What this usually means is that the variances of the observations are not equal, in other words the variance matrix V is not of the form $I\sigma^2$ but is diagonal with unequal diagonal elements. When this event occurs it is necessary to amend the ordinary least squares estimator of the regression coefficients. (Draper & Smith, 1981).

The basic idea is to transform the observations so as to correct for the heteroscedasticity:

Suppose the model under consideration is

$$Y = X\alpha + \varepsilon$$

where $E(\varepsilon) = 0$, $V(\varepsilon) = V\sigma^2$, and $\varepsilon \sim N(0, V\sigma^2)$.

It is possible to find a unique non-singular symmetric P such that

$$P' P = P P = P^2 = V$$

If we premultiply the model by P^{-1} we obtain a new model:

$$P^{-1} Y = P^{-1} X \alpha + P^{-1} \varepsilon$$

or

$$Z = Q \alpha + f$$

where $f = P^{-1} \varepsilon$ and $E(f) = 0$

then $\hat{\alpha} = (X' V^{-1} X)^{-1} X' V^{-1} Y$

This (standard) weighted least squares estimator of α is a function of the variance V , and is applicable if V is known (at least up to a scalar multiple). However in our application V is not known and so this estimator is not directly applicable.

What we can assume is that the variance function is smooth and that it can be modelled using a periodic function of the form:

$$\sigma_i^2 = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

where

$$x_{0i} = 1$$
$$x_{1i} = \cos\left(\left(\frac{2\pi}{1460}\right) * i\right)$$
$$x_{2i} = \sin\left(\left(\frac{2\pi}{1460}\right) * i\right)$$

Now if the parameters β_0 ; β_1 and β_2 were known (up to a scalar multiple) then it would be possible to estimate the coefficients α . Since these coefficients are not known it is necessary to estimate both the α 's and β 's simultaneously. There is no analytic method of obtaining these estimates. The following iterative algorithm was applied to yield both the weighted least squares estimates of the α 's and estimates of the variance function:

An initial estimate of the α 's is obtained using the method of ordinary least squares, i.e.

$$\hat{\alpha} = (X'X)^{-1}X'Y$$

where the Y vector has the components $y_t = m0_t$ (or $= m2_t$) .

and where $X = [X_0\ X_1\ X_2]$

The algorithm then proceeds by successively estimating the parameters β (hence the variance matrix V) and then using these to obtain an improved weighted least squares estimate of the parameters α . The latter are then used to re-estimate the β 's and this cycle is repeated until convergence is achieved.

The coefficients β are estimated as follows. Using the available values of $\hat{\alpha}$ we compute the square residuals

$$y_i^* = (y_i - \hat{y}_i)^2 \qquad i = 1, 2, \ldots, n$$

where

$$\hat{y}_i = \hat{\alpha}_0 x_{0i} + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$$

The estimates of β are then obtained using the method of ordinary least squares, i.e.

$$\hat{\beta}^* = (X'X)^{-1}X'Y^*$$

These values have to be scaled. Recall that the variance needs only to be determined up to a scalar multiple in order to apply the method of weighted least squares to estimate the α 's. The rescaling needs to be done in order to avoid the (arbitrary) units of the variance from becoming too small which can lead to numerical problems, such as overflow on the computer. A convenient way to rescale the β is to set

$$\hat{\beta}_j^{\,0} = \frac{\hat{\beta}_j^{\,*}}{\hat{\beta}_0} \qquad j = 0,1,2$$

The fitted variances are then given by

$$\sigma_i^{\,2} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

and hence

$$\hat{V} = \begin{pmatrix} \sigma_1^{\,2} & & \\ & \ddots & \\ & & \sigma_n^{\,2} \end{pmatrix}$$

This estimate is then used to obtain an updated weighted least squares of the α 's:

$$\hat{\alpha} = (X'\hat{V}^{-1}X)^{-1} X'\hat{V}^{-1}Y$$

and the cycle is then repeated until the estimates of the α 's converge.

In practice about 5 iterations were needed to achieve convergence.

We note that this procedure is based on a model that assumes independence of the observations. Our data are not independently (or normally) distributed and hence the estimates cannot be expected to enjoy the optimality properties associated with least squares estimates under the usual assumptions. However we emphasise that this procedure was only used in the exploratory analysis to obtain approximate estimates of the seasonal cycles evident in the time series. These estimates were also used to provide starting values for the maximum likelihood procedure used to finally fit the models. The latter procedure takes account of the serial correlation and the fact that the data are not normally distributed.

Appendix C.

DOUBLE EXPONENTIAL (LAPLACE) DISTRIBUTION.

This section gives a brief summary of the details of the Double Exponential distribution, (see Johnson & Kotz (1970))

If the random variable x is Double Exponentially distributed then we say that

$$x \sim DE(\lambda)$$

and then x has the probability density function $f(x)$ where

$$f(x) = \frac{1}{2\lambda} \exp[-|x-\theta|/\lambda] \qquad \lambda > 0$$

but in our case $\theta = 0$ so we use

$$f(x) = \frac{1}{2\lambda} \exp[-|x|/\lambda] \qquad \lambda > 0 \tag{1}$$

The distribution function $F(x)$ is then given by

$$\begin{aligned} F(x) &= \frac{1}{2} \exp[-(-x)/\lambda] \qquad x \leq 0 \\ &= 1 - \frac{1}{2} \exp[-(x)/\lambda] \qquad x \geq 0 \end{aligned}$$

Given observed values of n mutually independent random variables $X_1, \ldots, X_n$ each with probability density function (1) the likelihood function is

$$-n \ln(2\lambda) - \frac{1}{\lambda} \sum_{j=1}^{n} |X_j - \theta|$$

Whether the value of λ is known or not, any value $\widehat{\theta}$ minimizing

$$\sum_{j=1}^{n} |X_j - \theta|$$

with respect to θ is the maximum likelihood estimate of θ .

If λ is unknown, a maximum likelihood estimate of λ is

$$\widehat{\lambda} = \frac{1}{n} \sum_{j=1}^{n} |X_j - \widehat{\theta}|$$

Given that in this study $\theta = 0$ the method for generating random variables from a Double Exponential distribution is as follows:

$$F(x) = \frac{1}{2}\exp[-(-x)/\lambda] \qquad x \leq 0$$
$$= 1 - \frac{1}{2}\exp[-(x)/\lambda] \qquad x \geq 0$$

Therefore a $DE(\lambda)$ random variate can be generated as follows:

Generate U (where $U \sim U(0;1)$) then if $(U < 0.5)$ then let $x = \lambda * \ln(2 * U)$ else if $(U \geq 0.5)$ then let $x = -\lambda * \ln(2 * (1 - U))$

Appendix D.

PHASE AND AMPLITUDE REPRESENTATION.

Consider a model of the following form:

(Bloomfield (1976)).

$$x_t = \mu + A\cos\omega t + B\sin\omega t + \varepsilon t$$

where $\omega = \left(\frac{2*\pi}{365*4}\right)$

then to find the corresponding amplitude and phase, R and ψ, we solve the equations:

$$A = R\cos\psi \quad \text{and} \quad B = -R\sin\psi$$

x_t can then be written with the following amplitude/phase representation:

$$x_t = \mu + R\cos(\omega(t-\psi))$$

Since R is non-negative, it follows that $R = (A^2 + B^2)^{\frac{1}{2}}$. The basic equation for ψ is $\tan\psi = -B/A$. However, the solution $\psi = \arctan(-B/A)$ gives the same value for $-A$ and $-B$ as for A and B. To achieve a one-to-one relationship between (A, B) and (R, ψ) it is usual (see Bloomfield (1976) to split up the domain of the arctan function and define ψ as follows:

$$\psi = \begin{cases} (Arctan(-B/A)) & A > 0 \\ (Arctan(-B/A)) - \pi & A < 0, B > 0 \\ (Arctan(-B/A)) + \pi & A < 0, B \leq 0 \\ (-\pi/2) & A = 0, B < 0 \\ (\pi/2) & A = 0, B < 0 \\ (arbitrary) & A = 0, B = 0 \end{cases}$$

where *Arctan* represents the principal value of the arctan function. To find the time where the function has maximum and minimum values, i.e. for peaks and troughs, we solve the following for t :

For peaks the argument $(\omega t + \psi)$ of the term $\cos(\omega t + \psi)$ vanishes, i.e.

$$(\omega t + \psi) = 0 \quad \text{and therefore} \quad t = -\frac{\psi}{\omega}$$

For troughs the argument $(\omega t + \psi) = \pi$ and therefore $t = \frac{\pi - \psi}{\omega}$