Nonparametric and Parametric Estimation Of Wave Statistics And Spectra

Ocean Engineering

Hidekatsu Yamazaki, Ph.D.
John B. Herbich, Ph.D., P.E.
Department of Civil Engineering
Texas A&M University

TEXAS A&M SEA GRANT COLLEGE PROGRAM
86-202

NONPARAMETRIC AND PARAMETRIC ESTIMATION OF WAVE STATISTICS AND SPECTRA

Prepared by

Hidekatsu Yamazaki, Ph.D.
and
John B. Herbich, Ph.D., P.E.

Ocean Engineering Program
Civil Engineering Department

TAMU-SG-86-202

COE Report No. 279

October 1985

This work is partially supported through Institutional Grant NA83AA-D-00061 to Texas A&M University by the National Oceanic and Atmospheric Administration's Sea Grant Program, Department of Commerce.

A three-volume set of monthly wave characteristics data is available to complement this report. This set is available for $25.00 from the address given below.

TAMU-SG-86-202
300 September 1985
NA83AA-D-00061
R/ME-2

$10.00
Additional copies available from:
Marine Information Service
Sea Grant College Program
Texas A&M University
College Station, Texas 77843-4115

ABSTRACT

A nonparametric bivariate density estimation technique is developed employing tensor product B-splines to provide a concise wave data summary. Most of the existing nonparametric techniques involve a certain level of subjectivity in the choice of smoothing parameters. A criterion based on the least squares concept is proposed to remove the subjective choice of smoothing parameters. Numerical experiments, in which random variables are generated from a known bivariate independent normal distribution and the modified Longuet-Higgins distribution, show that the technique reproduces the population density functions well. However, due to lack of the shape preserving property of B-splines, the positivity of the density function cannot be guaranteed.

An alternative spectral estimation procedure is proposed, extending the idea of Bretschneider (1959). The alternative spectrum is the second moment of the wave height of the joint probability density function (pdf) in terms of the frequency domain, and is named the PDF spectra. Comparison of the latter with other spectral estimators such as the FFT spectral window estimator and the autoregressive spectral estimator shows good agreement.

The nonparametric joint pdf provides a concise representation of long-term wave data from which one can obtain not only the usual wave statistics, but the wave spectra as well. That is, the wave spectrum is simply a subset statistical function contained in the bivariate pdf for wave height and period.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

List of Figures (Continued)

FIGURE Page

List of Figures (Continued)

FIGURE Page

INTRODUCTION

Objectives

To meet the increasing need for accurate wave climatology along the U.S. coast, both federal and nonfederal organizations have established wave information systems. The amount of wave data gathered has been increasing exponentially; however, it seems that the data analysis techniques have not developed at the same pace to meet the user needs and applications.

Wave data analysis is usually performed employing wave spectral analysis or wave statistics methods, sometimes referred to as the "wave by wave" method. Most of the wave information systems use the former technique; however, a long term data representation in terms of spectra does not seem to be adequate. In view of this shortcoming, this research project had the following specific objectives:

1. to develop a technique representing wave data in a concise form for both the short- as well as the long-term,
2. to develop an alternative spectral estimation technique making use of the joint probability density function of wave heights and periods,
3. to compare available spectral analysis techniques with an alternative technique.

Present Status of Wave Data Analysis

The wave data representation must meet the user requirements and needs for applications in the coastal areas. During the spring of 1982, NOAA's Coastal Wave Program and the American Society of Civil Engineers hosted a series of regional coastal wave workshops. The general needs and concerns of wave information were summarized by Edge and Moore (1982) as follows:

1. there is an urgent need for real time wave data analysis and forecasting primarily for safety and efficiency of offshore operations, recreational boating, and commercial and sports fishing,
2. validation of numerical modeling techniques used in forecasting and hindcasting is handicapped by poor wave data bases,
3. analysis of shoreline behavior and sediment transport to predict and mitigate adverse beach erosion or coastal hazard conditions require better wave data to reduce the level of uncertainty present in existing technology, and
4. the design and maintenance of coastal and offshore structures need comprehensive statistical and physically descriptive summaries of wave occurrence and form.

There is an obvious need to improve wave data collection techniques as well as analysis of observational data. An ideal system for real-time wave data information must be developed, involving the expertise of systems engineers, ocean engineers, and oceanographers. Another crucial problem that must be solved to meet the current wave

information needs is the lack of a comprehensive data representation technique. A single wave data observation, usually of a 20-minute duration, provides a large amount of information. Two methods are commonly employed to summarize wave data; i.e., wave statistics and wave spectral analysis. Existing wave information systems use the latter technique. A monthly data base collected for example, four times per day, contains about 120 observations. A series of wave spectra, while informative, does not provide a complete wave data summary.

Statistical considerations of ocean waves were originated by Sverdrup and Munk in 1942 in connection with the need for wave forecasts during World War II. They introduced the idea of representative wave height, called "the significant wave height". The significant wave height is the average of the 1/3 highest waves during an observation. Putz(1952) conducted a systematic analysis of wave data to obtain the empirical wave height and period distributions. During the same year, Longuet-Higgins(1952) derived a theoretical distribution of wave heights. He assumed that the ordinate of sea surface was a linear Gaussian process with a narrow banded spectrum. This leads to the well known Rayleigh distribution for heights. When the individual waves are defined in terms of the zero up-crossing method, the observed wave heights agree well with the Rayleigh distribution (Goodnight and Russell,1963; Collins,1967; Goda,1974a). However, several authors report that the wave heights in the higher exceedance probability range do not agree well with the theory. The wave heights in this range are important for the estimation of extreme

wave conditions, for example in the design of offshore structures. The discrepancy between the observations and theory has been reported in a series of studies by Forristall(1978), Longuet-Higgins(1980), and Tayfun(1980, 1981). Tayfun(1981) lists a number of reasons for the discrepancy, 1) the non-linear, non-Gaussian characteristics of the sea surface, 2) the effect associated with wide-band spectra, and 3) wave-breaking in deep and shallow water depths.

Sea waves must be characterized not only by wave height but also by wave period. Putz(1952) proposed an empirical wave period distribution which has the form of a Weibull distribution. Bretschneider (1959) analyzed a vast amount of wave data and found the wave length to be distributed according to the Rayleigh distribution. Applying the linear wave theory, Bretschneider developed an empirical distribution of wave period. The first theoretical derivation for the wave period distribution was given by Longuet-Higgins(1975) in connection with the joint distribution of wave heights and periods. The wave period distribution has not been studied in as much detail as the wave height distribution. According to Goda(1974b), the wave periods are distributed differently depending upon spectral shape.

The wave height and period are not physically independent, thus, they must be considered as jointly distributed random variables. Bretschneider (1959) studied a special case in which the wave heights and periods were distributed independently and hence were simply a product of marginal distributions. Longuet-Higgins (1975) derived a theoretical joint distribution but, unfortunately, the linear

correlation coefficient for this distribution is equal to zero due to an infinite variance of the period. This implies that the wave height and period are uncorrelated random variables. Recently, Kimura (1981) proposed an empirical joint distribution using the bivariate Weibull distribution based on a simulation study. Goda (1978) classified patterns of joint distribution in terms of the cross frequency histogram using the correlation coefficient. The joint distribution contains both the marginal wave height and period distributions. However, the marginal distributions are not necessarily descriptive of the joint variablity, and there is a need to pursue further research on the joint distribution of wave heights and periods.

The distributions mentioned above are of the parametric density family, which involve a few parameters to describe the variability of random variables. However, the important detailed features of variability, such as bimodal peaks, may not be characterized by a simple parametric distribution. An alternative, to prevent misleading analysis, is the application of nonparametric density estimation to the wave data. This technique has not been applied by those working in the field. However, nonparametric density estimation is a major research field in modern statistics. Several techniques have been proposed. Tapia and Thompson(1978) provide a useful guide and extensive bibliography. An excellent summary and updated information may be found in Bean and Tsoko(1980). A special effort was made in this study to develop a nonparametric density estimation technique which would provide a concise wave data representation.

Spectral analysis is the most popular method of wave data investigation. Due to the development of the Fast Fourier Transform (FFT) by Cooley and Tukey (1965), the computational effort of spectral estimation was considerably reduced as compared to the classic Blackman-Tukey method (1958), often referred to as the auto-covariance method. Most present wave data information systems use the FFT method because of the tremendous amount of data that must be analyzed. Spectral analysis has caused some confusion among engineers. A certain level of statistical knowledge is necessary to avoid misleading interpretations and applications; however, sophisticated statistical theorems are generally not required. Extensive research has been conducted in modern spectral analysis during the last decade. This research clarifies and summarizes available techniques for practical applications.

The major research area to be solved is the relationship between wave statistics and wave spectrum. Depending on the application of wave information, some applications call for wave spectra, while others require a single representation of wave height and period. Very little research has been done in connection with this question. Goda(1974a) attempted to estimate wave statistics from spectral information. However, it seems almost impossible to investigate the joint structure of wave height and period variablity using spectral information alone.

WAVE DATA COLLECTION ACTIVITIES

General

In 1977 a project was initiated at Texas A&M University to establish a Wave Data Bank for the Texas Coast (Herbich and Jensen, 1978). An extensive search for recorded data was made by contacting federal and state agencies as well as industries. Much of the data were found to be in possession of the petrochemical industries and not available for public use.

A complete set of analog-type wave data observed at the Flagship Pier in Galveston (May 1976 to May 1977) was obtained from the U.S. Army Corps of Engineers. Selected data were analyzed and published by Herbich and Watanabe(1980a).

Another set of wave measurements was obtained from the Data Office of the National Oceanic and Atmospheric Administration (NOAA) for two buoys located in the central part of the Gulf of Mexico. The long-term wave and meteorological statistics were analyzed and published in two volumes (Herbich and Yamazaki, 1984).

In conjunction with this project, two Waverider buoys were deployed along the Texas Coast. Both buoys were anchored in approximately 60 feet of water near offshore platforms. A significant amount of wave data was obtained for Waverider B, which was located near Port Mansfield. Only very limited data was obtained from Waverider C due to loss of equipment and malfunction of the recorder.

The Waverider buoy located near Port Mansfield was in a direct path of hurricane Allen. The buoy survived the hurricane and measurements of wave data were analyzed and compared with the hindcasted data computed from the meteorological observation (Herbich and Watanabe, 1980).

Wave Data for Galveston, Texas

A complete year of wave records observed at the Flagship Pier was obtained from the U.S. Army Corps of Engineers. The data cover a period from May 1976 to April 1977. Since the data obtained were in analog form on a strip chart, the data had to be digitized manually by an optical digitizer. Spectral analysis was performed on the sample data; the auto-convariance method employing the Hanning lag window was used for the analysis. Significant wave heights were estimated from the variance of surface elevation of each record. The selected data were published by Herbich and Watanabe (1980), and are reproduced in Table 1 and 2.

NDBO Wave Data

The NOAA Data Buoy Office (NDBO) operates a number of buoys on the continental shelf and in the deep ocean off the U.S. coast. The moored buoys have onboard data aquisition and reporting systems, commonly referred to as payloads. A payload includes meteorological and oceanographic sensors, an electronic system for data acquisition, data processing and formatting, as well as a communication system to relay the formatted data to shore (Steel, 1978).

TABLE 1 — Summary of results for Galveston gage - peak wave records

Date	H_{sig} (ft)	T_{peak} (sec)
1-5 May/76	5.26	5.6
4-5 May/76	4.73	6.4
1-13 May/76	3.05	4.4
4-26 May/76	11.12	6.0
3-5 June/76	1.94	5.6
3-7 June/76	3.19	5.6
3-24 June/76	3.80	6.0
2-25 June/76	3.29	6.4
2-3 July/76	4.17	6.9
3-9 July/76	2.79	6.4
1-10 July/76	3.18	6.4
1-13 July/76	3.60	6.9
2-6 Aug/76	2.97	4.4
1-11 Aug/76	2.69	3.5
2-19 Aug/76	2.99	5.6
2-20 Aug/76	3.78	6.0
1-1 Sept/76	4.11	3.5
1-14 Sept/76	4.44	5.2
1-20 Sept/76	3.71	5.6
1-27 Sept/76	3.65	4.2
1-5 Oct/76	4.50	6.0
1-23 Oct/76	4.70	6.0
1-28 Oct/76	3.96	6.4
2-29 Oct/76	7.09	7.6
2-8 Nov/76	5.13	6.0
2-13 Nov/76	5.77	6.9
3-19 Nov/76	3.99	5.2
3-26 Nov/76	3.13	6.9
2-6 Dec/76	5.12	6.4
1-14 Dec/76	3.59	6.9
1-25 Dec/76	3.27	7.6
2-31 Dec/76	6.08	6.9
1-8 Jan/77	2.50	4.6
3-15 Jan/77	5.11	6.4
2-23 Jan/77	3.83	4.9
3-30 Jan/77	4.98	6.0
4-2 Feb/77	4.92	6.9
3-7 Feb/77	4.83	5.2
1-11 Feb/77	4.44	6.0
3-18 Feb/77	2.53	5.6
4-5 Mar/77	5.60	6.9
4-11 Mar/77	5.00	6.4
3-18 Mar/77	4.02	6.0
2-28 Mar/77	8.26	6.9
3-3 Apr/77	4.99	7.6
3-12 Apr/77	4.92	7.6
3-20 Apr/77	4.44	6.9
4-28 Apr/77	3.57	4.9

TABLE 2 — Summary of results for Galveston gage - random wave records

Date	H_{sig} (ft)	T_{peak} (sec)
4-3 May/76	3.89	---
1-7 May/76	4.38	6.4
1-11 May/76	2.93	4.9
5-11 May/76	2.42	---
4-2 June/76	3.18	6.0
1-11 June/76	1.92	---
4-11 June/76	2.82	---
2-24 June/76	3.87	5.6
4-12 July/76	3.52	6.0
2-19 July/76	3.05	---
3-27 July/76	2.87	4.4
2-30 July/76	3.06	---
3-3 Aug/76	1.37	---
1-4 Aug/76	1.82	---
1-9 Aug/76	2.88	---
1-25 Aug/76	2.14	---
2-1 Sept/76	5.91	---
2-7 Sept/76	3.91	---
2-23 Sept/76	2.42	---
3-29 Sept/76	5.37	---
3-5 Oct/76	3.64	6.0
1-8 Oct/76	2.14	---
1-20 Oct/76	1.83	---
4-23 Oct/76	2.79	5.6
4-12 Nov/76	5.58	5.6
2-16 Nov/76	3.33	---
2-21 Nov/76	3.19	---
4-30 Nov/76	3.03	---
4-10 Dec/76	2.79	6.4
3-14 Dec/76	2.49	6.4
2-20 Dec/76	2.13	---
1-22 Dec/76	3.54	---
3-5 Jan/77	3.82	7.6
1-20 Jan/77	2.43	---
4-23 Jan/77	3.17	5.6
3-27 Jan/77	2.20	5.2
1-4 Feb/77	1.30	7.6
3-6 Feb/77	2.41	4.9
2-9 Feb/77	4.04	6.4
2-16 Feb/77	2.04	---
4-6 Mar/77	2.86	6.9
2-15 Mar/77	1.37	5.6
3-19 Mar/77	2.86	6.0
1-27 Mar/77	3.24	4.4
3-1 Apr/77	4.00	4.6
3-4 Apr/77	2.41	6.9
1-18 Apr/77	2.94	6.0
4-27 Apr/77	3.18	3.6

Two buoys, WMO-42001 and 42002, have been collecting meteorological and wave data in the central part of the Gulf of Mexico in approximately 6,000 feet of water. The location of the buoys is shown in Fig. 1. Data covering the period December 1979 to May 1982 were obtained from NDBO for these two buoys.

The data obtained are stored on a 9-track, 1600 bpi, unlabelled, ANSI/ASCII, FB, LRECL=120, and block size=4800 magnetic tape. The data cover geophysical conditions, meteorological information and wave information. Observations are made every hour. The data also provides spectral information. Table 3 shows the number of recorded data per month.

Significant wave height $H_{1/3}$ and the average wave period $\overline{T}$ are estimated using the following relations,

$$H_{1/3} = 4.0\sqrt{m_o}$$

$$\overline{T} = \sqrt{\frac{m_o}{m_2}}$$

where

$$m_o = \int_o^\infty S(f)df$$

$$m_2 = \int_o^\infty f^2 S(f)df$$

S(f) = one-sided power spectrum

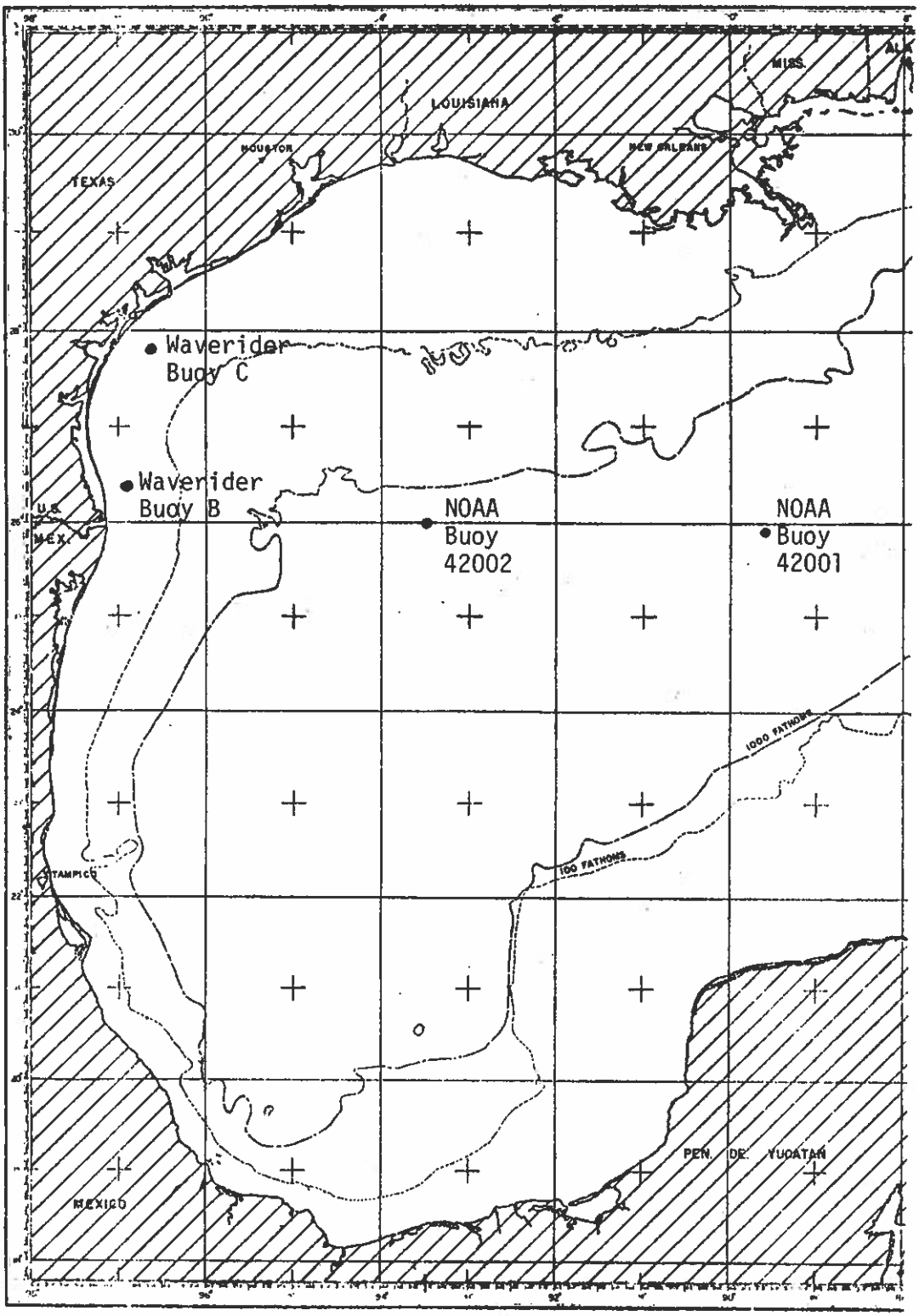

Figure 1. The location of NOAA 42001, 42002, TAMU Waverider buoy B and C. (adapted from Department of Oceanography map, Texas A&M University)

Table 3. Number of recorded wave information (NDBO wave data)

Date	Buoy WMO-42001	Buoy WMO-42002
79 Dec	647	730
80 Jan	648	744
80 Feb	286	696
80 Mar	0	744
80 Apr	648	711
80 May	744	742
80 Jun	720	720
80 Jul	743	744
80 Aug	744	293
80 Sep	720	714
80 Oct	744	732
80 Nov	720	0
80 Dec	691	744
81 Jan	0	732
81 Feb	0	618
81 Mar	0	736
81 Apr	0	720
81 May	1	739
81 Jun	716	716
81 Jul	743	743
81 Aug	743	742
81 Sep	342	711
81 Oct	692	720
81 Nov	720	720
81 Dec	736	735
82 Jan	739	739
82 Feb	670	655
82 Mar	742	728
82 Apr	717	706
82 May	742	742

Waverider Buoy Data

Since only a limited amount of wave data were available, a decision was made to establish a wave monitoring system along the Texas

coastline. Due to the high probability of equipment loss by either natural causes or commercial fishing operations in open water, many petrochemical companies were contacted concerning the deployment of Waverider buoys near their offshore platforms. In this manner the buoys could be closely monitored by the personnel either living on or visiting the platforms at frequent intervals. Mobil Oil Company showed interest and agreed to participate in the project.

The first deployment of a Waverider (Buoy A) was made on July 19, 1979, near Port Mansfield. It disappeared after a deployment of only three days (Herbich and Watanabe, 1979). In spite of extensive efforts to locate the missing buoy, it could not be found. The second Waverider (Buoy B) was deployed on February 14, 1980, at the Harena Platform near Port Mansfield. The third Waverider (Buoy C) was deployed on September 16, 1980, at Matagorda 487 Platform near Port O'Connor, Texas. Both Waveriders were anchored in approximately 60 feet of water. Both Waveriders have been transmitting wave data to the receiving stations on the respective platforms. Buoy C disappeared in October 1981. Another Waverider (Buoy D) was deployed at the same location as Buoy C. Buoy B survived more than two years until it disappeared in April 1983. Waverider Buoy B provided long-term wave observations. The locations of Buoy B and C are shown in Fig. 1.

Although only limited direct evidence could be found, the buoys were probably lost because of shrimping and pipe-laying operations close to the platforms.*

*Buoy C was recovered in damaged condition by the U.S. Coast Guard in 1985. It apparently had been hit by the propeller of a larger ship.

The Dima receiving unit (MARK II) records the wave data on the digital cassettes in a complimentary non-return to zero (CNRZ) format. The data density on a cassette is 615 bit/inch. One word consists of 3 bytes, where one byte is 4 bits. Therefore one word is 12 bits, and it can be represented as three hexidecimal numbers.

The smallest value that can be represented is

/0000/0000/0000/

which corresponds to a water elevation of -2048 cm. The maximum value that can be represented is

/1111/1111/1111/

which corresponds to a water elevation of +2047 cm. One data file contains 64 words. The standard 2882-foot cassette tape with 615 bit/inch holds 54 records of 20 minutes measurement period at a sampling rate of 0.5 sec.

Observations were made four times per day (i.e. 03:00, 09:00, 15:00, and 21:00 hrs) for 20 minutes at a sampling rate of 0.5 second. Each wave record contains approximately 2400 data points.

Datawell recommended a Datel cassette reader unit (LPR-16). Major difficulties encountered were the compatibilities between the Dima unit and Datel unit, and the data transformation from cassette to computer compatible form. LPR-16 reads the bit pattern in the opposite direction. A computer program (INVERT) was developed to convert the direction of the read-out. By means of a parallel interface, LPR-16 was connected to a PDP-11 minicomputer for transfering the data to a 9-track magnetic tape. The data were then analyzed on an Amdahl 470

V/6,V/8 Computer at Texas A&M's Data Processing Center (DPC). It was found recently that LPR-16 was unstable in performing the transformation of data. Mr. Chris Noynaert of DPC has been assisting in the data conversion process. PDP-11, which belonged to DPC, was transferred to another institute, and only five cassettes were successfully read by PDP-11.

A CP/M microcomputer (Balcones) was employed to perform the data conversion. LPR-16 was connected with CP/M by means of a serial interface. The data had to be transferred from a cassette to a floppy disk by this method before it could be analyzed on an Amdahl computer. At the present time, the LPR-16 is being connected to an IBM PC through a serial interface to transfer the data to disk, and the data analysis is being performed directly in the PC. There appear to be less problems in the transfer of data with this arrangement. The total conversion time of one cassette was increased to about 5 hours. The available data for potential users are given in the secondary volume of this report. The original data are available on request by contacting the Ocean Engineering Program at Texas A&M University.

DATA ANALYSIS

General

The Texas Wave Information System (TWIS) has been developed to analyze both NDBO and Waverider data. TWIS can analyze data in different ways depending on the data source. NDBO data comprised hourly meteorological and oceanographical data, and included the significant wave height and average period. TWIS was developed to provide comprehensive long-term wave information. The NDBO data were summarized on a monthly basis.

Waverider data provide a major part of data source for the coastal area of Texas. Efforts were made to develop a computer program package which would analyze the data in a comprehensive manner. Spectral analysis was performed employing several different methods for a selected data set. The computer package used was ARSPID, which was developed by Newton (1983) in the Institute of Statistics at Texas A&M University. ARSPID performs two different spectral analyses, the window spectral estimation and the autoregressive spectral estimation. It was proposed to develop an alternative way to estimate spectra by means of a joint probability density function of wave heights and periods, which is discussed in a separate chapter. A comparison of spectral estimators is also described in a separate chapter.

NDBO Data Analysis

The Statistical Analysis System (SAS) has been employed in various fields to handle statistical data. SAS provides a number of statistical analysis packages and computer graphic presentations. SAS can be used either by those having an extensive experience in programming or by those unfamiliar with the process. A SAS program was developed to analyze the NDBO data on a monthly basis. A FORTRAN program (called NODA) was developed to transfer data from NOAA data format to SAS data form. The following selected values are used for the monthly summary:

Selected variables	Symbol used
i) Air Temperature (C°)	AIRTEMP
ii) Sea Temperature (C°)	SEATEMP
iii) Barometric pressure (mb)	BAROMTR
iv) Wind speed (m/sec)	WINDSPED
v) Wind direction (in degree from true north)	WINDDRC
vi) Significant wave height (m)	SIGWVHT
vii) Average Wave period (sec)	AVWVPR

NDBO data include spectral information, however, it was not presented in the monthly summaries. An example of a monthly summary is presented for December 1979 for 42001 data. Table 4 shows the monthly wave statistics summary ,where N is the number of sample observations. The monthly histogram of wind speed is given in Fig. 2. The monthly histogram of wind direction can be found in Fig. 3. Fig. 4. shows the

monthly histogram of significant wave height, and the corresponding average wave period variation is shown in Fig. 5. The cross frequency of significant wave height and average period is given in Table 5.

Waverider Buoy Data

Cassette reader LPR-16 reads wave data stored on cassettes by means of a serial interface. LPR-16 was connected to a CP/M microcomputer (Balcones). A computer program WYLTERMC in ASSEMBLER was developed by the personnel at DPC (TAMU) to perform data transformation. Wave data were read from a cassette and stored on a flopy disc. WYLTERMC transferred the data from a flopy disc to Amdhal V6/V8. The data were temporarily stored on an Amdhal data file. The stored data are in a hexidecimal form and, since the data were read in the opposite direction with opposite bit pattern, it had to be "inverted". The program INVERT alters the data sequence order and the bit pattern, and provides the data in a suitable form for analysis.

The first step in data analysis was to separate data set sequence into a standard data format, and to ascertain the quality of recorded data. This process was performed manually. The Waverider's recorder provides the selfcheck of the data obtained. The standard data format consists of the title of data, the date and time, the number of data points, the data format, and the surface elevation data. The example of standard format is as follows:

```
WAVERIDER B CASSETTE #3 3 PM MARCH 13, 1980
 2310    (16F5.0)

        Surface elevation data
```

Table 4. Monthly wave statistics summary for Buoy 42001, December 1979

AIRTEMP	AIR TEMPERATURE (C)
SEATEMP	SEA TEMPERATURE (C)
BAROMTR	BAROMETER (MB)
WINDSPED	WIND SPEED (M/SEC)
WINDDRC	WIND DIRECTION (DEGREES FROM TRUE NORTH)
SIGWVHT	SIGNIFICANT WAVE HIGHTS (M)
AVWVPR	AVERAGE WAVE PERIODS (SEC)

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	RANGE	SKEWNESS	KURTOSIS
AIRTEMP	647	22.496	2.257	17.300	28.100	8.800	-0.477	-0.879
SEATEMP	647	25.674	0.578	24.800	27.300	2.700	0.023	-0.784
BAROMTR	647	1018.847	4.298	1005.700	1028.100	22.400	-0.585	0.685
WINDSPED	647	6.741	3.003	0.200	14.800	14.600	-0.148	-0.728
WINDDRC	647	135.300	102.285	0.300	359.800	359.500	0.942	-0.315
SIGWVHT	647	1.230	0.607	0.200	3.300	3.100	0.480	0.199
AVWVPR	647	4.873	0.959	2.600	7.600	4.800	0.338	-0.275

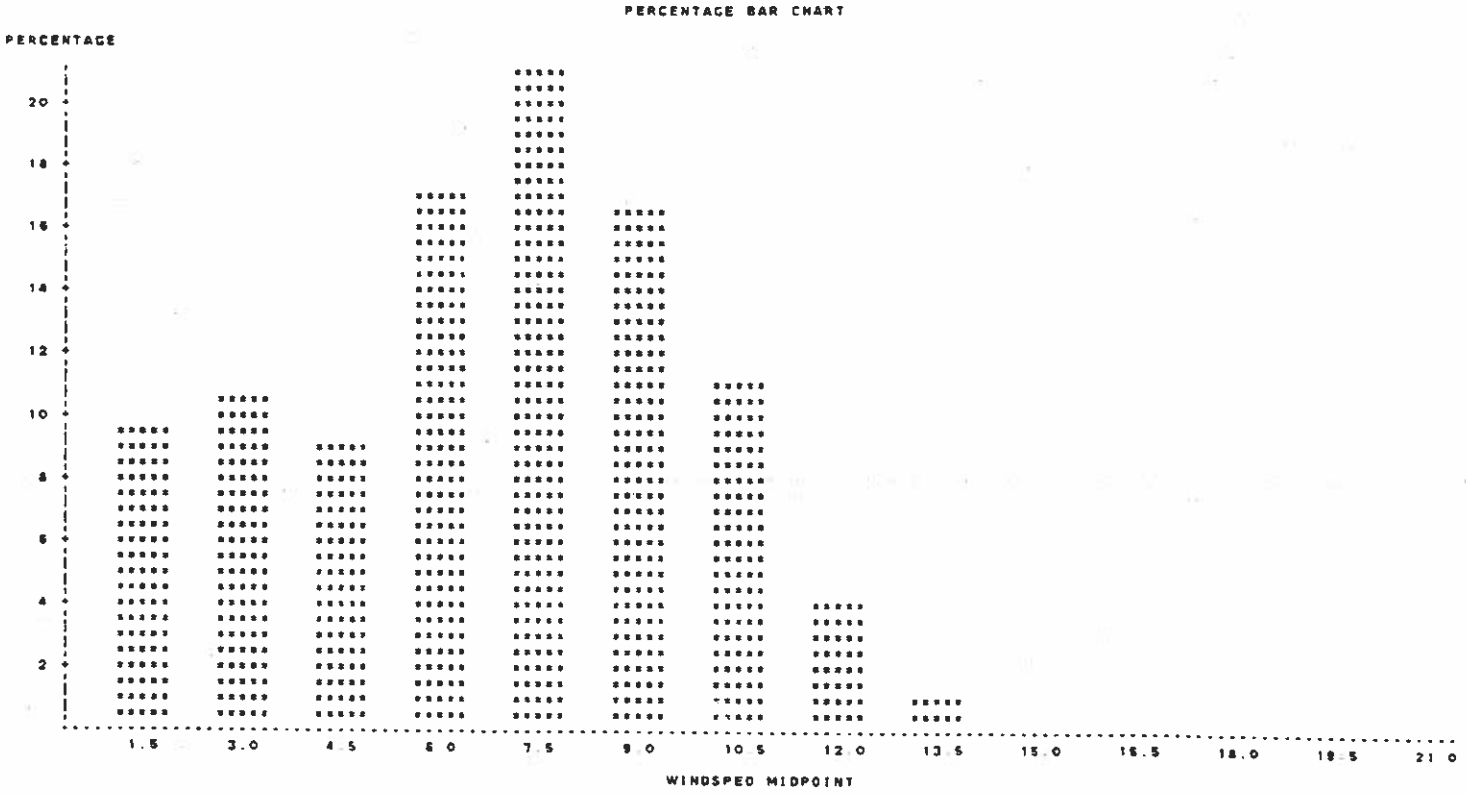

Figure 2. Histogram of wind speed for Buoy 42001, December 1979

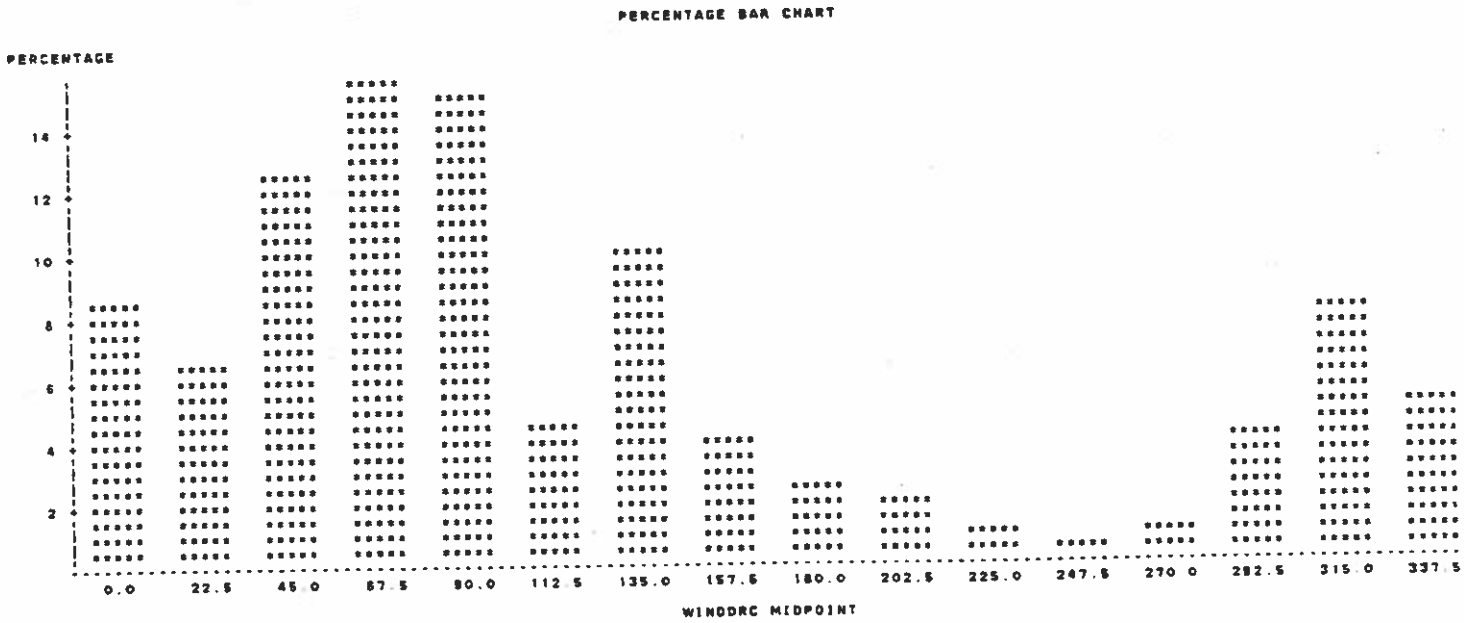

Figure 3. Histogram of wind direction for Buoy 42001, December 1979 in degree from true north

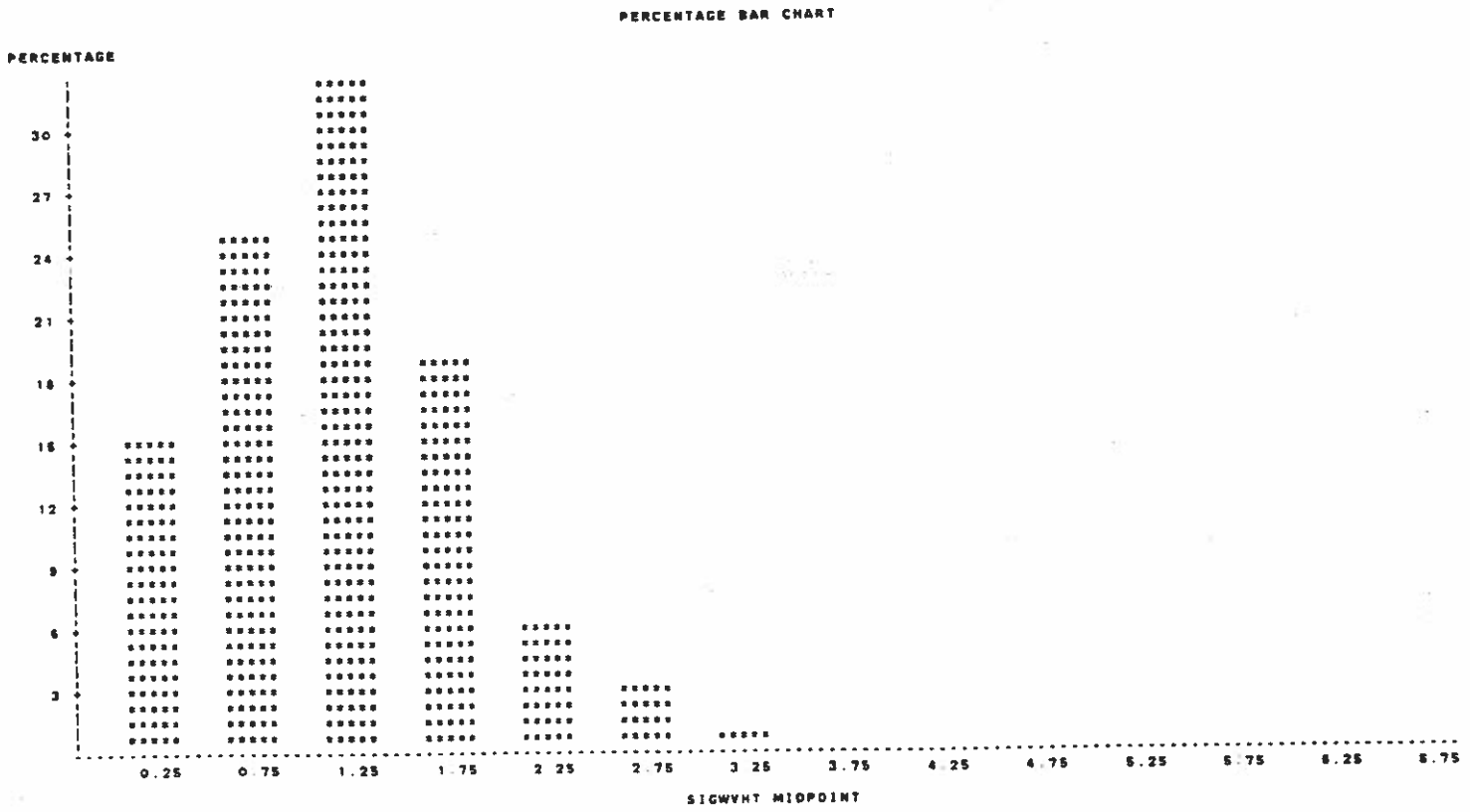

Figure 4. Histogram of significant wave height (m) for Buoy 42001, December 1979

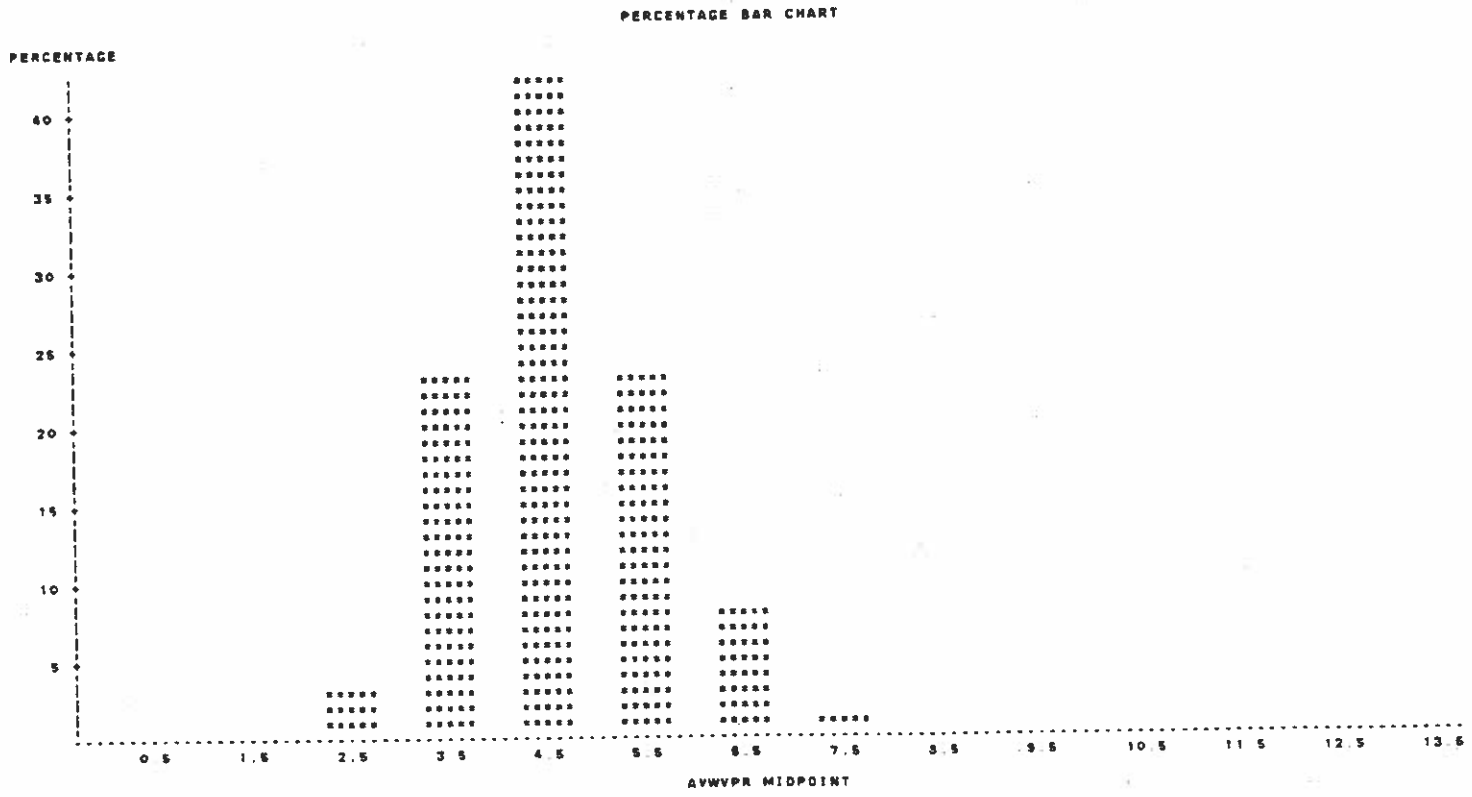

Figure 5. Histogram of average wave period (sec) for Buoy 42001, December 1979

Table 5. Cross frequency of significant wave height and average period for December 1979

TABLE OF AVWVPR BY SIGWVHT

AVWVPR / SIGWVHT	.	<=0.5	<=1.0	<=1.5	<=2.0	<=2.5	<=3.0	<=3.5	TOTAL
FREQUENCY / PERCENT / ROW PCT / COL PCT									
	1	0	0	0	0	0	0	0	.
		.	.	.	.	.	.	.	
		.	.	.	.	.	.	.	
		.	.	.	.	.	.	.	
<=3.0	0	20	0	0	0	0	0	0	20
		3.09	0.00	0.00	0.00	0.00	0.00	0.00	3.09
		100.00	0.00	0.00	0.00	0.00	0.00	0.00	
		20.62	0.00	0.00	0.00	0.00	0.00	0.00	
<=4.0	0	21	68	58	5	0	0	0	152
		3.25	10.51	8.98	0.77	0.00	0.00	0.00	23.48
		13.82	44.74	38.16	3.29	0.00	0.00	0.00	
		21.65	41.98	27.88	4.17	0.00	0.00	0.00	
<=5.0	0	56	68	81	51	6	0	0	272
		8.66	10.51	12.52	8.43	0.93	0.00	0.00	42.04
		20.59	25.00	29.78	22.43	2.21	0.00	0.00	
		57.73	41.98	38.84	50.83	15.00	0.00	0.00	
<=6.0	0	0	26	58	46	6	1	0	147
		0.00	4.02	10.51	7.11	0.93	0.15	0.00	22.72
		0.00	17.69	46.26	31.29	4.08	0.68	0.00	
		0.00	16.05	32.69	38.33	15.00	5.88	0.00	
<=7.0	0	0	0	1	8	28	15	0	52
		0.00	0.00	0.15	1.24	4.33	2.32	0.00	8.04
		0.00	0.00	1.92	15.38	53.85	28.85	0.00	
		0.00	0.00	0.48	6.67	70.00	88.24	0.00	
<=8.0	0	0	0	0	0	0	1	3	4
		0.00	0.00	0.00	0.00	0.00	0.15	0.46	0.62
		0.00	0.00	0.00	0.00	0.00	25.00	75.00	
		0.00	0.00	0.00	0.00	0.00	5.88	100.00	
TOTAL		97	162	206	120	40	17	3	647
		14.89	25.04	32.15	18.55	6.18	2.63	0.46	100.00

The first line can be read by means of A FORMAT of 20A4, the first number on the second line is the number of data points. The data should be read by the specified format on the second line. The data were stored on a disc in the Amdhal computer in a file called WRBB80.MR13P3. The selected wave data from waverider buoy B were used for comparing wave spectral estimators; the results will be discussed in a later chapter. It was not intended to estimate the spectra for each data set for the following reasons:

1) one data set should be evaluated using different estimators to determine which provides a reasonable solution for the wave data,
2) the massive amount of data may not provide a comprehensive data representation, and will be cost prohibitive.

TWIS provides wave statistics based on the zero up-crossing method. Program ETATH produces the wave height and period data from the original surface elevation data. The variance of surface elevation can be used to estimate the wave statistics, such as the average wave height HA, and the significant wave height H3

$$HA=2.507\ \eta_{rms}$$

$$H3=4.004\ \eta_{rms}$$

where η_{rms} is the root mean square of surface elevation

The average wave period TA may be estimated by dividing total length of observation by the number of zero up-crossing waves. The statistics of surface elvation $\{\eta_i\}$ are estimated such that the variance VR is

$$VR = \overline{\eta^2} = \frac{1}{N}\sum_{i=1}^{N} \eta_i^2$$

and

$$\eta_{rms} = \sqrt{\overline{\eta^2}}$$

where

$$\overline{\eta} = \frac{1}{N}\sum_{i=1}^{N} \eta_i = 0$$

The skewness SKW is

$$SKW = \frac{1}{N}\sum_{i=1}^{N} \eta_i^3/\eta_{rms}^3$$

The Kurtosis KRT is

$$KRT = \frac{1}{N}\sum_{i=1}^{N} \eta_i^4/\eta_{rms}^4$$

The above symbols are used in summarized data, which are presented in other volumes of the annual report.

Due to a malfunction of the magnetic cassettes in the Dima unit, some data were only available on analog strip chart records. The Tucker method (1963) is a simple method to estimate the root mean square of surface elevation from the analog form of the wave data. To fill gaps of missing data on magnetic cassettes, the Tucker method (1963) was used for the wave statistics; however, the skewness and the kurtosis are not estimated by the method employed. Wave statistics by means of the Tucker method are indicated with "*" in the list of available data. It should be noted that those data sets with "*" are not as reliable as the data sets without "*".

The zero up-crossing wave heights and periods are stored on Amhdal having a similar data format with the surface elevation data. An example of wave heights and periods is as follows:

```
WAVERIDER B CASSETTE #34.1 3 PM AUGUST 11 1981
NW 438     (12F6.2)
                .
                .
    Wave heights and periods data
                .
```

The first line is the title of data set. 438 in the second line is the number of waves in this data set. Next the data format (FORTRAN) used is shown. The data are given on the following third line. Each data set is coupled with a wave height and a wave period in the following sequence. The discussion of zero up-crossing method follows in the next chapter.

A program WVSTAT estimates the statistical wave characteristics also by the zero up-crossing method. WVSTAT also estimates a primitive joint probability density function of wave height and period and, marginal distributions of wave heights and periods. An example output of WVSTAT is shown in Table 6 and Figures 6, 7 and 8.

Table 6. Wave statistics of an example output of WVSTAT

```
WAVERIDER B CASSETT #34.1 3 PM AUGUST 11 1981

NUMBER OF WAVES =   438

STATISTICAL WAVE CHARACTERISTICS
CORRELATION COEFFICIENT OF WAVE HEIGHTS AND PERIODS      0.6368
VARIANCE OF WAVE HEIGHTS                                 0.0108
VARIANCE OF WAVE PERIODS                                 1.4875
SECOND MOMENT OF WAVE HEIGHT                             0.0583

AVERAGE       WAVE      HA=      0.21     TA=      2.83   OBS=438
SIGNIFICANT   WAVE      H3=      0.33     T3=      3.58   OBS=146
1/10 TH       WAVE     H10=      0.41    T10=      3.70   OBS= 43
MAXIMUM       WAVE     HMX=      0.56    TMX=      3.76   OBS=  1

                       VHA=      0.23
                       VH3=      0.36
                       VTA=      2.64
                    VHA/HA=    1.0574
                    VH3/H3=    1.0912
                     T3/TA=    1.3654
                    T10/TA=    1.4091
                    TMX/TA=    1.4313
```

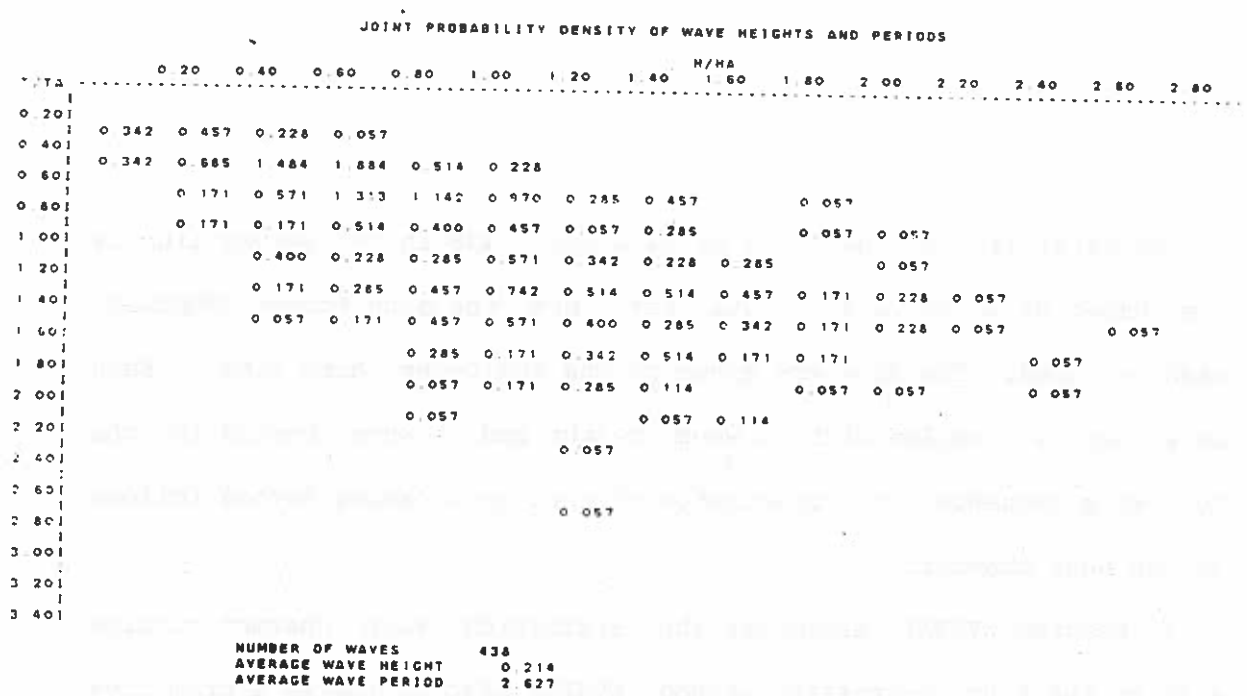

Figure 6. Joint probability density function of an example output of WVSTAT

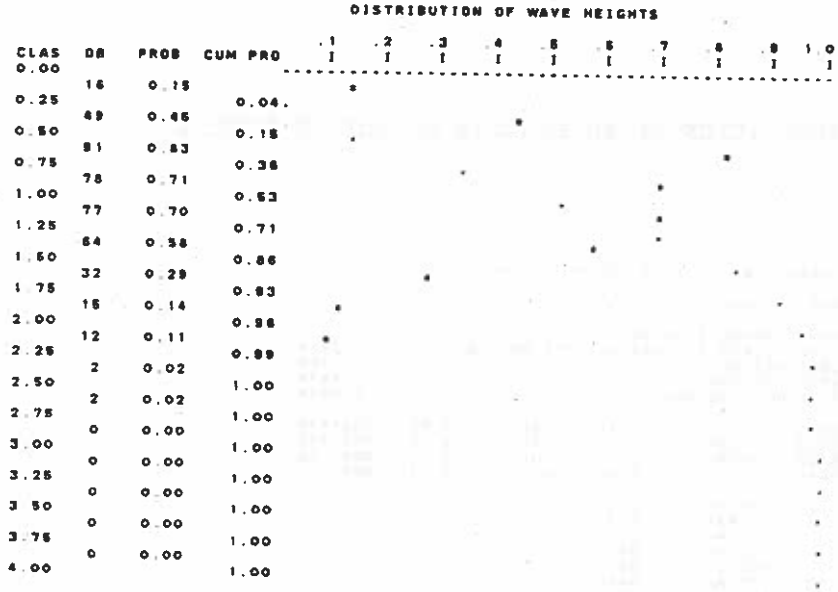

Figure 7. Distribution of wave heights of an example output of WVSTAT

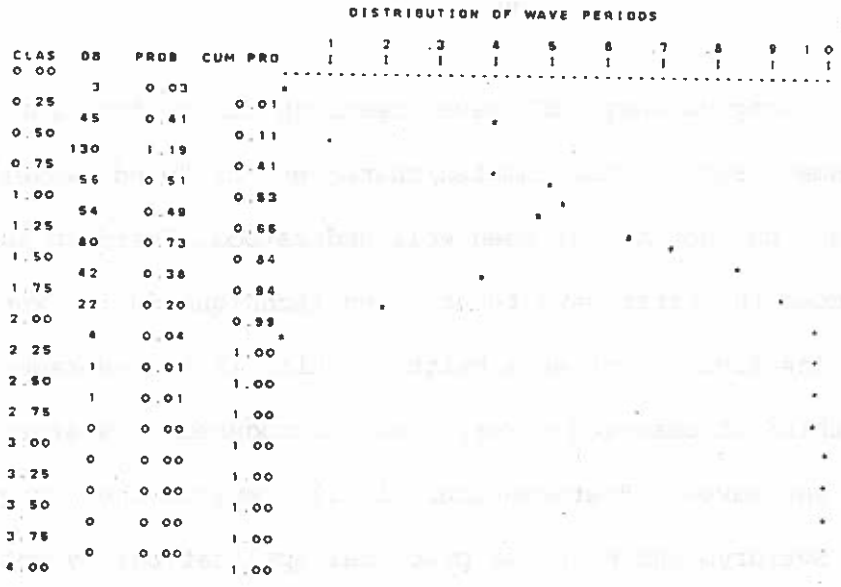

Figure 8. Distribution of wave periods of an example output of WVSTAT

A technique was developed to estimate the joint pdf of wave height and period nonparametrically. A program PDFSPEC performs a nonparametric estimation of the joint pdf. An alternative method was proposed to estimate the wave spectra by means of the joint pdf. The spectral estimation is named PDF spectral estimation. PDFSPEC also performs the estimation of the PDF spectra. A detailed discussion of the joint pdf is given in chapter 5, and a theoretical argument of the PDF spectra is provided in chapter 7.

PARAMETRIC MARGINAL DISTRIBUTION OF WAVE HEIGHTS AND PERIODS

General

Despite a long history of wave research since the 19th Century (Kinsman,1965), the complex character of "wind waves", or simply "waves," has not as yet been well understood. Sverdrup and Munk (1947) developed the first wave forecasting technique during World War II in 1942. The significant wave height, which is the average of the highest one-third of observed waves, was introduced to statistically characterize the waves. Bretschneider (1951) revised the forecasting technique of Sverdrup and Munk for practical applications by potential users. This is known as the S.M.B. method, and this method is still used as a simple forecasting technique.

The first effort to develop an empirical wave height distribution was made by Putz (1952). He discovered that the wave height distribution could be fitted by a Gamma type distribution quite well. About the same time, Longuet-Higgins (1952) derived a theoretical distribution of wave heights, known as the Rayleigh distribution, under the assumption of narrow band wave spectra and linearity. Both the Putz distribution and the Rayleigh distribution fitted observed data well (Bretschneider 1959). The Putz distribution is a two parameter family, and the Rayleigh distribution contains only one parameter. Watter (1953) and Bretschneider (1959) checked the Rayleigh distribution using observed wave data. Although they did not perform

statistical tests, the Rayleigh distribution appeared to be adequate for practical applications. Goodnight and Russell (1963) showed that the Rayleigh distribution did not satisfy the chi-square hypothesis test. However, they concluded that the Rayleigh distribution could be practically accepted in moderate ranges. Collins (1967) also showed that the Rayleigh distribution appeared to be a good fit for observed data.

Because of the rather cumbersome calculations involved in wave statistics, Tucker (1963) proposed a simple method to estimate the wave statistics based on the assumption that waves followed the Rayleigh distribution. The method is still useful when the data obtained are only in analog form.

The determination of each individual wave height from analog data represents a statistical sampling problem. Basically two methods are in existence: i.e., the zero crossing method and the crest-trough method. Wilson and Baird (1972) made a comparison of the zero crossing method for wave height with the crest-trough method for wave height. They also employed the Tucker method to supply an alternative definition. The zero crossing method for wave height exhibited a better fit to the Rayleigh distribution than the other methods. Goda (1974a) also confirmed that the zero up-crossing wave height appeared to fit the Rayleigh distribution well. He also discussed the effect of sampling interval on the wave height distribution. Goda (1974b) 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 spectral shape nor the spectral width parameter.

Wu (1973) found that the zero up-crossing wave height showed slightly lower values than those predicted by the Rayleigh distribution, but he also concluded that, "in general, the data show no significant deviation from the Rayleigh distribution." Though the practical applications of the Rayleigh distribution have been confirmed by many researchers, the theory predicts higher wave heights for larger waves in a sea state (Nolte and Hsu 1979). Forristall (1978) proposed an alternative distribution, which fitted the observation data better by suppressing this over-prediction. The distribution function is a two parameter Weibull distribution in which the parameters have to be estimated empirically. Longuet-Higgins (1980) noted that lesser agreement between observed data and the theory is due to some misunderstanding of the originally proposed distribution. He found the Rayleigh distribution fitted Forristall's data (1978) well by suitably adjusting the parameter of the distribution empirically.

The wave height distribution is not the only one of practical and theoretical interest; the waves are characterized also by heights and periods. Putz (1952) made an early study on the distribution of wave periods, using the crest-to-crest method. He proposed a Gamma type

distribution. Bretschneider (1959) found the wave length distribution was similar to the Rayleigh distribution. Applying linear wave theory, he proposed a modified Rayleigh distribution for the wave periods. Wu (1973) found the frequency distribution of (T/T_{rms}) , (where T_{rms} is the root mean square of wave period T), which indicated good agreement with the Rayleigh distribution. The other methods of period distribution have been discussed in conjunction with the joint behavior of wave heights and periods. Despite the abundance of wave statistics, the studies of wave period distribution have not been successful. Goda (1974b) found empirical relations of wave period parameters, which were compared with the results of the present study indicating similar relations.

A basic knowledge of general statistics is required to perform wave data analysis, therefore basic probability notions are given in Appendix B to assist the reader.

Wave Sampling Method

There are two methods that are used most commonly to analyze wave records in terms of individual waves. In the first method, the wave height is defined as the vertical distance between the succeeding crest and trough, and the wave period is defined as a time period between a succeeding crest to crest. In the second method, called the zero up-crossing method, the wave height is defined as the maximum excursion between two successive upward crossings of the mean water

elevation, and the wave period of the zero up-crossing method is the time period between the successive zero up-crossing points. The zero down-crossing wave can be defined similarly. Both definitions, the first method and the second method, give the same result as the spectrum approaches a line spectrum, which is a pure harmonic wave. However, since the real surface elevation has a finite spectrum band width, the two definitions produce different results. As the spectrum width becomes wider, the difference can be expected to be significant. According to the Rayleigh distribution, the significant wave height $H_{1/3}$, which is the average of the highest one third of observed waves, can be approximated as

$$H_{1/3} \doteq 4\eta_{rms} \qquad (1)$$

where η_{rms} is the root mean square of surface elevation η. The significant wave period $T_{1/3}$ is the average period of the highest one third of the waves.

Wilson and Baird (1972) compared the significant wave heights defined by the crest to trough method, the zero up-crossing method, and the above equation. They found the deviation between the crest to trough $H_{1/3}$ and $4\eta_{rms}$ was larger than the deviation between the zero up-crossing $H_{1/3}$ and $4\ \eta_{rms}$.

The second method was employed for the following reasons:

1) The zero crossing wave height seems to fit the Rayleigh distribution well (Goda,1974a), and

2) The zero crossing method is a low-pass filtering operator. It may be regarded that the successive wave profile is replaced by a random sequence of pure harmonics having equivalent wave height and period (Iwagaki and Kimura,1976).

There still remains the choice of either the zero up-crossing or the zero down-crossing method. Statistical testing was conducted for selected wave parameters to determine whether the zero up-crossing method and the zero down-crossing method were statistically different. In March 1980, 104 consecutive data sets, including seven missing observations, were used for this purpose.

The following wave parameters of wave height H and period T were investigated from sample observations for each method.

1) The mean wave height

$$\overline{H} = E[H] \tag{2.1}$$

2) The mean wave period

$$\overline{T} = E[T] \tag{2.2}$$

3) The variance of wave height

$$\sigma_H^2 = Var[H] \tag{2.3}$$

4) The variance of wave period

$$\sigma_T^2 = Var[T] \tag{2.4}$$

5) The correlation coefficient of H and T

$$\rho_{HT} = \frac{cov[H,T]}{\sigma_H \sigma_T} \tag{2.5}$$

6) The significant wave height

$H_{1/3}$ =the mean wave height of exceedance probability 1/3 (2.6)

7) The significant wave period

$T_{1/3}$ =the mean wave period associated with $H_{1/3}$ (2.7)

8) The 1/10-th wave height

$H_{1/10}$=the mean wave height of exceedance probability 1/10 (2.8)

9) The 1/10-th wave period

$T_{1/10}$=the mean wave period associated with $H_{1/10}$ (2.9)

10) The maximum wave height

H_{max} =the maximum wave height for the observation period (2.10)

11) The maximum wave period

T_{max} =the wave period associated with H_{max} (2.11)

The sample parameters were estimated using N coupled wave height and period samples, i.e. $\{H_i, T_i\}$, i=1,..,N, as follows

1) The sample mean of wave height

$$\hat{\bar{H}} = \frac{1}{N} \sum_{i=1}^{N} H_i \tag{3.1}$$

2) The sample mean of wave period

$$\hat{\bar{T}} = \frac{1}{N} \sum_{i=1}^{N} T_i \tag{3.2}$$

3) The sample variance of wave height

$$S_H^2 = \frac{1}{N} \sum_{i=1}^{N} (H_i - \hat{\bar{H}})^2 \tag{3.3}$$

4) The sample variance of wave period

$$S_T^2 = \frac{1}{N} \sum_{i=1}^{N} (T_i - \hat{\bar{T}})^2 \tag{3.4}$$

5) The sample correlation coefficent

$$\hat{\rho}_{HT} = \frac{\sum_{i=1}^{N} (H_i - \hat{\bar{H}})(T_i - \hat{\bar{T}})}{\sqrt{\sum_{i=1}^{N} (H_i - \hat{\bar{H}})^2 \sum_{i=1}^{N} (T_i - \bar{T})^2}} \tag{3.5}$$

6) The sample significant wave height

$\hat{H}_{1/3}$=the average INT(N/3) higher ordered statistics of H (3.6)

where INT(.) is the largest integer

7) The sample significant wave period

$\hat{T}_{1/3}$ =the average wave period associated with $\hat{H}_{1/3}$ (3.7)

8) The sample 1/10-th wave height

$\hat{H}_{1/10}$ =the average INT(N/10) higher ordered statistics of H (3.8)

9) The sample 1/10-th wave period

$\hat{T}_{1/10}$ =the average wave period associated with $\hat{H}_{1/10}$ (3.9)

10) The sample maximum wave height

$\hat{H}_{max}$ =the largest sample wave height (3.10)

11) The sample maximum wave period

$\hat{T}_{max}$ =the wave period of $\hat{H}_{max}$ (3.11)

The question addressed here was whether the joint pdf of H and T for the zero up-crossing and that of the zero down-crossing method were the same. However, no statistical test was available to the authors to answer this question. Therefore statistical tests were conducted on selected parameters, such as ρ_{HT}, σ_H, and σ_T, under the hypothesis that the estimated parameters are drawn from the same density function. Let α_ρ, α_H, and α_T be equal to:

$$\alpha_\rho = \rho_{HT}^{(U)} - \rho_{HT}^{(D)} \tag{4.1}$$

$$\alpha_H = \sigma_H^{2(U)} - \sigma_H^{2(D)} \tag{4.2}$$

$$\alpha_T = \sigma_T^{2(U)} - \sigma_T^{2(D)} \tag{4.3}$$

where the supercripts (U) and (D) denote the method used, namely the zero up-crossing method and the zero down-crossing method respectively.

The hypotheses to be tested are as follows:

i) $H_o : E[\alpha_\rho]=0, \quad H_A : E[\alpha_\rho]\neq 0$ (5.1)

ii) $H_o : E[\alpha_H]=0, \quad H_A : E[\alpha_H]\neq 0$ (5.2)

iii) $H_o : E[\alpha_T]=0, \quad H_A : E[\alpha_T]\neq 0$ (5.3)

where H_o is the null hypothesis and H_A is the alternative hypothesis. The estimated $\hat{\alpha}_\rho$, $\hat{\alpha}_H$, and $\hat{\alpha}_T$ based on the respective sample parameters are shown in Figures 9-11. The seven missing data among the consecutive observations were simply replaced by zero. These figures imply each series of $\hat{\alpha}_\rho$, $\hat{\alpha}_H$ and $\hat{\alpha}_T$ are possibly time-dependent random variables. Ordinary hypothesis testing assumes the sample must be a result of independent random sampling. If the sequence of $\hat{\alpha}$ were considered as a normal stochastic process, then N consecutive observations of $\hat{\alpha}.$, (a vector $\underset{\sim}{\alpha}.$) have an N - dimensional multinormal distribution the pdf of which is denoted by

$$f_{\underset{\sim}{\alpha}.}(\underset{\sim}{\alpha}.) = \text{Normal}\ (\underset{\sim}{\mu}, \underset{\approx}{\Sigma}) \tag{6}$$

where $\underset{\sim}{\mu}$ is a mean vector, and $\underset{\approx}{\Sigma}$ is the covariance matrix. The mean of $\underset{\sim}{\alpha}.$, denoted $\underset{\sim}{\bar{\alpha}}.$, has the univariate normal distribution, i.e:

$$f_{\underset{\sim}{\bar{\alpha}}.}(\underset{\sim}{\bar{\alpha}}.) = \text{Normal}\ (\mu, \frac{1}{N^2}\ \underset{\sim}{1}^T\ \underset{\approx}{\Sigma}\ \underset{\sim}{1}) \tag{7}$$

where $\mu=E[\underset{\sim}{\bar{\alpha}}.]$ and $\underset{\sim}{1}$ is the unit vector.

The variance of $\underset{\sim}{\bar{\alpha}}$ can be approximated by the following equation (Priestley, 1981)

$$\frac{1}{N^2}\underset{\sim}{1}^T\ \underset{\approx}{\Sigma}\ \underset{\sim}{1} \doteq \frac{2\pi S(0)}{N} \tag{8}$$

where $S(.)$ is the autoregressive spectrum of order P, and

$$S(0) = \frac{1}{2\pi} \frac{\sigma^2}{\sum_{j=1}^{P} |a_j|^2} \qquad (9)$$

σ^2 = the residual variance

a_j = the autoregressive spectrum coefficients

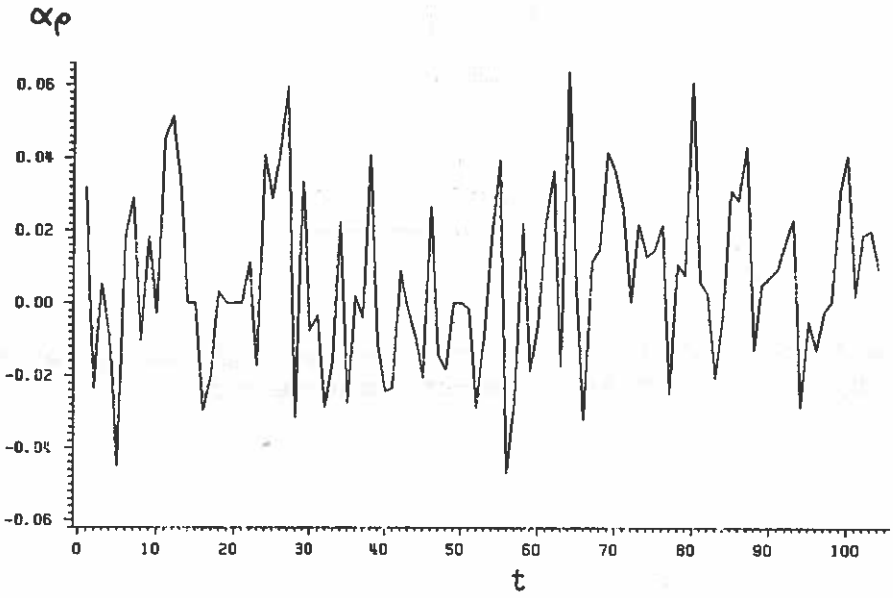

Figure 9. Time history of the difference of correlation coefficients (by means of zero up-crossing and zero down-crossing methods)

The ordinary t-test can be performed on the proposed hypothesis:

1) The difference of correlation coefficient:

The spectrum analysis indicated that $\{ \alpha_\rho \}$ appeared to be white noise. Therefore, each observation can be considered as an independent random sample. The hypothesis testing is essentially the ordinary t-test. The t-value is obtained as

$$t = 2.646 > t_{.975,103} = 1.984$$

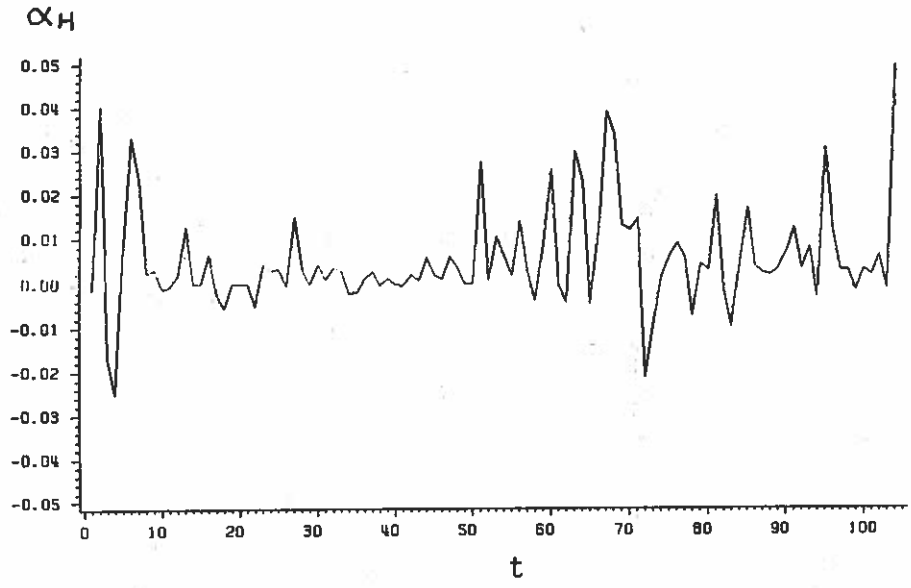

Figure 10. Time history of the difference of wave height variance (by means of zero up-crossing and zero down-crossong methods)

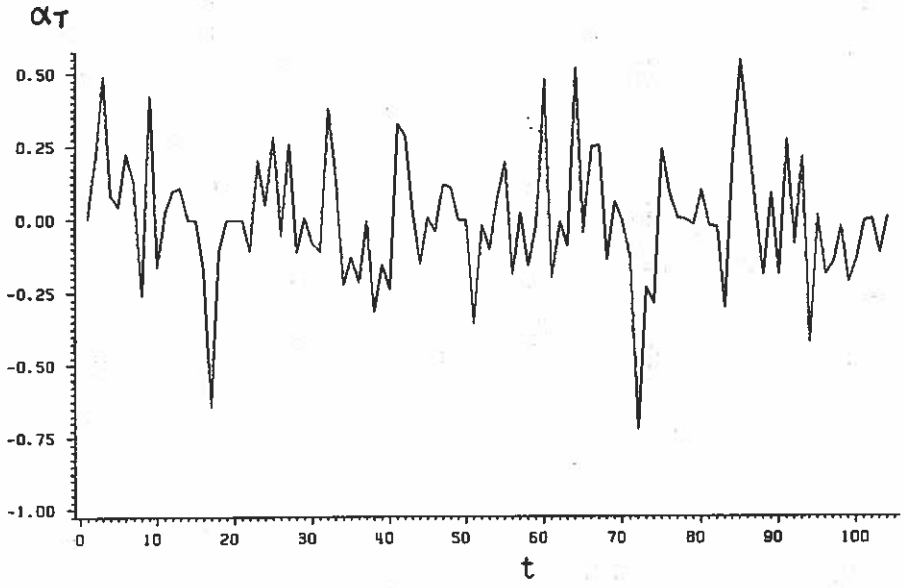

Figure 11. Time history of the difference of wave period variance (by means of zero up-crossing and zero down-crossing methods)

Then the hypothesis should be rejected at the 5 percent significant level.

2) The difference of sample variance of H :

Based on the autoregressive model criteria, namely AIC and CAT, which will be introduced in the following chapter, the series can be described by a second order autoregressive process, such that

$$\alpha_H(t) = 0.1460\alpha_H(t-1) - 0.1972\,\alpha_H(t-2) + \varepsilon(t)$$

$$t = 3,4,\ldots,N$$

where $\varepsilon(t)$ is white noise. Using this model, the standard deviation of α_H is estimated as

$$S_H = 1.069 \times 10^{-3}$$

The corresponding t value is

$$t = 5.104 > t_{.975,103} = 1.984$$

This value indicates that the hypothesis should be rejected at the 5 percent significance level.

3) The difference of sample variance of T :

The series of $\{\alpha_T\}$ is again shown to be white noise, and the t-value obtained is

$$t = 0.0534 < t_{.975,103} = 1.984$$

Hence the hypothesis cannot be rejected at 5 percent significance level.

An inspection of other estimated parameters showed that the average wave heights and periods of two methods were in agreement up to the second decimal, and as the number of samples used became smaller, the parameters deviated further. The maximum wave had a larger deviation

than the other parameters. This tendency merely explains the sampling problem. The above results imply that two methods, namely the zero up-crossing method and the zero down-crossing method, do not produce the same wave statistics in detail. However, the author could not reach a conclusion on whether both methods are equivalent. Since the zero up-crossing method has been widely used by previous researchers, this method was employed to define the individual waves in the present study.

Marginal Distribution of Wave Heights

Longuet-Higgins (1952) was the first to study the theoretical distribution of wave heights, which is known as the Rayleigh distribution. A brief review of the Rayleigh distribution is given below.

The assumption that the linear stationary process allows one to describe the surface elevation $\eta(t)$ as a Fourier series

$$\eta(t) = \sum_{n=1}^{\infty} a_n \cos(2\pi f_n t + \varepsilon_n) \qquad (10)$$

where a_n, f_n and ε_n are the amplitude, frequency and phase respectively. a_n and f_n have certain probability distributions 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. According to the central limit theorem, the distribution of $\eta(t)$ may be approximated by the normal distribution having a mean value

$$E[\eta(t)] = 0 \tag{11}$$

and variance

$$Var[\eta(t)] = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} a_i^2 = \frac{1}{2} a_{rms}^2 \tag{12}$$

where a_{rms} is the root mean square amplitude.

Assuming a constant frequency f_o, then rewriting equation (10) as

$$\eta(t) = R(t)\cos(2\pi f_o t + \phi(t)) \tag{13}$$

where

$$R(t) = \sqrt{A_c^2(t) + A_s^2(t)} \tag{14}$$

$$\phi(t) = \tan^{-1}\left(\frac{A_s(t)}{A_c(t)}\right) \tag{15}$$

$$A_c(t) = \sum_{n=1}^{\infty} a_n \cos(2\pi f_n t - 2\pi f_o t + \varepsilon_n) \tag{16}$$

$$A_s(t) = \sum_{n=1}^{\infty} a_n \sin(2\pi f_n t = 2\pi f_o t + \varepsilon_n) \tag{17}$$

Assume that the spectrum has a single narrow frequency band having a peak at frequency f_o, so that $R(t)$ is only significant around f_o. Then $R(t)$ is a slowly varying function which represents the envelope of $\eta(t)$. The cosine term of (13) is a carrier wave. The wave height H may be taken as 2R. The joint distribution of $R(t)$ and $\phi(t)$ can be obtained by applying the probability transformation on the joint distribution of $A_c(t)$ and $A_s(t)$. It can be shown that the joint pdf of $A_c(t)$ and $A_s(t)$ is an independent bivariate normal distribution, having the following parameters:

$$E[A_c(t)] = E[A_s(t)] = 0 \tag{18.1}$$

$$Var[A_c(t)] = Var[A_s(t)] = \frac{1}{2} a_{rms}^2 \tag{18.2}$$

$$cov[A_c(t), A_s(t)] = 0 \tag{18.3}$$

Performing the probability transformation, the joint pdf of R and ϕ can be derived as

$$f_{R\phi}(R,\phi) = \frac{R}{\pi a_{rms}^2} \exp[-\frac{R^2}{a_{rms}^2}] \tag{19}$$

Therefore the pdf of R is

$$f_R(R) = \int_0^{2\pi} f_{R\phi}(R,\phi)d\phi = \frac{2R}{a_{rms}^2} \exp[-\frac{R^2}{a_{rms}^2}] \tag{20}$$

Assuming H=2R is true, the pdf of H can be derived as

$$f_H(H) = \frac{H}{2a_{rms}^2} \exp[-\frac{H^2}{4a_{rms}^2}] \tag{21}$$

Let m_n be n-th moment of frequency spectrum $S\eta(f)$ of the surface elevation $\eta(t)$:

$$m_n = \int_0^{\infty} f^n S\eta(f)df \tag{22}$$

The lowest moment m_o has a value

$$m_o = \eta_{rms}^2 \tag{23}$$

where η_{rms} is the root mean square of η.

If the surface elevation is represented by (10), the following relation is true:

$$m_o = \eta_{rms}^2 = \frac{1}{2} a_{rms}^2 \tag{24}$$

Then equation (21) becomes;

$$f_H(H) = \frac{H}{4\eta^2_{rms}} \exp\left[-\frac{H^2}{8\eta^2_{rms}}\right] \qquad (25)$$

Longuet-Higgins (1980) showed that the nonlinearity and finite band width of spectra introduce an inequality in equation (24). He found that the nonlinear effect appeared in terms of the spectrum band width parameter ν . The ratio of $a^2_{rms}/2\eta^2_{rms}$ is expressed as

$$\frac{a^2_{rms}}{2\eta^2_{rms}} = 1 - 0.734\,\nu^2 \qquad (26)$$

where ν^2 is the spectrum width parameter, such that

$$\nu^2 = \frac{m_o m_2}{m_1^2} - 1 \qquad (27)$$

It has been found that the Rayleigh distribution, in the form of equation (25), overestimates higher wave heights. Forristall (1978) proposed an empirical Weibull distribution, which has two parameters, to overcome this problem. Longuet-Higgins compared the Rayleigh distribution with the Weibull distribution. It was found that the "one parameter" Rayleigh distribution fits the data as well as the "two parameters" Weibull distribution, if $a^2_{rms}/2\eta^2_{rms}$ is taken as 0.925. The significance of this result is that with suitable choice of the ratio of $a^2_{rms}/2\eta^2_{rms}$, the Rayleigh distribution is still applicable in the entire wave height range.

Several important wave height statistics are derived in terms of a_{rms} . The mean wave height $\overline{H}$ is

$$\bar{H} = E[H] = \int_0^\infty H\, f_H(H)dH = \int_0^\infty \frac{H^2}{2a_{rms}^2} \exp[-\frac{H^2}{4a_{rms}^2}]dH \qquad (28)$$
$$= \sqrt{\pi}\, a_{rms}$$

The second moment of H is denoted as $\overline{H^2}$, and is given as

$$\overline{H^2} = E[H^2] = \int_0^\infty H^2 f_H(H)dH$$
$$= \int_0^\infty \frac{H^3}{2a_{rms}^2} \exp[-\frac{H^2}{4a_{rms}^2}]dH = 4\, a_{rms}^2 \qquad (29)$$

The average height of exceedance probability 1/N has been extensively used to characterize a complex sea state, especially N=3 is well known as the significant wave height. Let X be a normalized wave height:

$$X = \frac{H}{a_{rms}} \qquad (30)$$

The pdf of X is transformed as

$$f_X(x) = \frac{1}{2}x\exp -[\frac{1}{4}\, x^2] \qquad (31)$$

The exceedance probability of 1/N is given as

$$\int_{X_N}^\infty f_X(x)d_x = \frac{1}{N} \qquad (32)$$

Hence

$$\exp[-\frac{1}{4}\, X_N] = \frac{1}{N}$$

X_N is obtained as

$$X_N = 2(\ell nN)^{\frac{1}{2}} \qquad (33)$$

Therefore $X_{1/N}$, the average height of exceedance probability 1/N, can be calculated as

$$X_{1/N} = \frac{\int_{X_N}^{\infty} f_X(x)d}{\int_{X_N}^{\infty} f_X(x)dx} \qquad (34)$$

$$= X_N + \sqrt{\pi N} - 2N \operatorname{erf}(X_N/2)$$

where $\operatorname{erf}(y) = \int_o^y e^{-t2}dt$, which is the error function.

The significant wave height can be obtained by setting N=3:

$$X_{1/3} = \frac{H_{1/3}}{a_{rms}} = 2.314 \qquad (35)$$

$$\therefore \quad H_{1/3} = 2.8314\ a_{rms}$$

Under the assumption of linearity, a_{rms} is

$$a_{rms} = \sqrt{2}\ \eta_{rms} \qquad (36)$$

Then $H_{1/3}$ can be expressed in terms of η_{rms} , i.e:

$$H_{1/3} \doteq 4.004\ \eta_{rms} \qquad (37)$$

which has been extensively used in many wave data programs. However, actual observations indicate that the above equation slightly overpredicts the significant wave height. Goda (1974b) found the factor of 3.79 for the data obtained in Nagoya Port:

$$H_{1/3} \doteq 3.79\ \eta_{rms} \qquad (38)$$

Forristall (1978) obtained the value of 3.77 based on the data obtained from several Hurricanes:

$$H_{1/3} \doteq 3.77\ \eta_{rms} \qquad (39)$$

The ratio of $H_{1/3}/\eta_{rms}$ was examined employing 97 data sets in March 1980, and 54 data sets in August 1981 obtained by the Waverider buoy near Port Mansfield, Texas. Each data set involves approximately 400 individual waves defined by the zero up-crossing method. The statistical summary of these data are shown in Tables 7 through 10; the following labels are used in the tables:

CORR(HT)= correlation coefficient of H and T

H3 = Significant wave height

HA = Average wave height

VH3 = 4.00 η_{rms}

VHA = 2.51 η_{rms}

TMX = Maximum wave period

T10 = 1/10-th wave period

T3 = Significant wave period

TA = Average wave period

data obtained in March 1980 show:

$$H_{1/3} = 3.83\ \eta_{rms}(\pm 0.1025) \tag{40}$$

and, data obtained in August 1981 show:

$$H_{1/3} = 3.76\ \eta_{rms}(\pm 0.1520) \tag{41}$$

where the values shown in parenthesis are 95 percent upper and lower confidence bands for each factor. The frequency diagrams of $H_{1/3}/\eta_{rms}$ are shown in Figures 12 and 13.

Using the Longuet-Higgins empirical ratio of $a_{rms}/\sqrt{2}\eta_{rms}$ =0.925, $H_{1/3}$ is expressed as

$$H_{1/3} = 3.70\ \eta_{rms} \qquad (42)$$

Table 7. Characteristic summary of wave parameters Part 1 (data obtained in March 1980).

LABEL	N	MEAN	STANDARD DEVIATION	SKEWNESS	KURTOSIS
CORR(HT)	97	0.691	0.046	-0.153	0.396
H3/VH3	97	0.957	0.013	-0.814	0.682
HA/VHA	97	0.961	0.026	-0.594	0.627
H3/RMS(ETA)	97	3.833	0.052	-0.814	0.682
HA/RMS(ETA)	97	2.410	0.065	-0.594	0.627
TMX/T3	97	1.011	0.116	0.046	1.107
T10/T3	97	1.021	0.045	2.817	13.871
TMAX/TA	97	1.278	0.162	0.320	0.582
T10/TA	97	1.291	0.091	1.116	2.111
T3/TA	97	1.264	0.068	0.251	1.745

Table 8. Characteristic summary of wave parameters Part 2 (data obtained in March 1980).

LABEL	N	STD ERROR OF MEAN	RANGE	MINIMUM VALUE	MAXIMUM VALUE
CORR(HT)	97	0.005	0.252	0.566	0.818
H3/VH3	97	0.001	0.067	0.915	0.983
HA/VHA	97	0.003	0.136	0.874	1.010
H3/RMS(ETA)	97	0.005	0.270	3.665	3.935
HA/RMS(ETA)	97	0.007	0.340	2.192	2.532
TMX/T3	97	0.012	0.680	0.658	1.338
T10/T3	97	0.005	0.340	0.952	1.292
TMAX/TA	97	0.016	0.857	0.849	1.706
T10/TA	97	0.009	0.538	1.097	1.635
T3/TA	97	0.007	0.438	1.030	1.468

Table 9. Characteristic summary of wave parameters Part 1 (data obtained in August 1981).

LABEL	N	MEAN	STANDARD DEVIATION	SKEWNESS	KURTOSIS
CORR(HT)	54	0.612	0.071	0.220	-0.488
H3/VH3	54	0.940	0.019	-1.270	2.895
HA/VHA	54	0.962	0.023	-0.784	1.300
H3/RMS(ETA)	54	3.762	0.078	-1.270	2.895
HA/RMS(ETA)	54	2.413	0.058	-0.784	1.300
TMX/T3	54	1.021	0.150	2.842	14.340
T1O/T3	54	1.022	0.034	1.168	2.726
TMAX/TA	54	1.268	0.221	3.282	17.346
T1O/TA	54	1.269	0.098	1.136	1.102
T3/TA	54	1.241	0.069	0.856	0.570

Table 10. Characteristic summary of wave parameters Part 2 (data obtained in August 1981).

LABEL	N	STD ERROR OF MEAN	RANGE	MINIMUM VALUE	MAXIMUM VALUE
CORR(HT)	54	0.010	0.308	0.462	0.770
H3/VH3	54	0.003	0.102	0.870	0.972
HA/VHA	54	0.003	0.120	0.892	1.011
H3/RMS(ETA)	54	0.011	0.409	3.483	3.891
HA/RMS(ETA)	54	0.008	0.300	2.235	2.535
TMX/T3	54	0.020	1.060	0.750	1.810
T1O/T3	54	0.005	0.174	0.965	1.139
TMAX/TA	54	0.030	1.534	0.957	2.491
T1O/TA	54	0.013	0.441	1.127	1.567
T3/TA	54	0.009	0.290	1.140	1.430

The values in equations (38)-(42) show surprisingly good agreement. It seems reasonable to suggest a value of 3.8 for the factor. This number agrees with the number suggested by Goda, et. al. (1976):

$$H_{1/3} = 3.80\ \eta_{rms} \qquad (43)$$

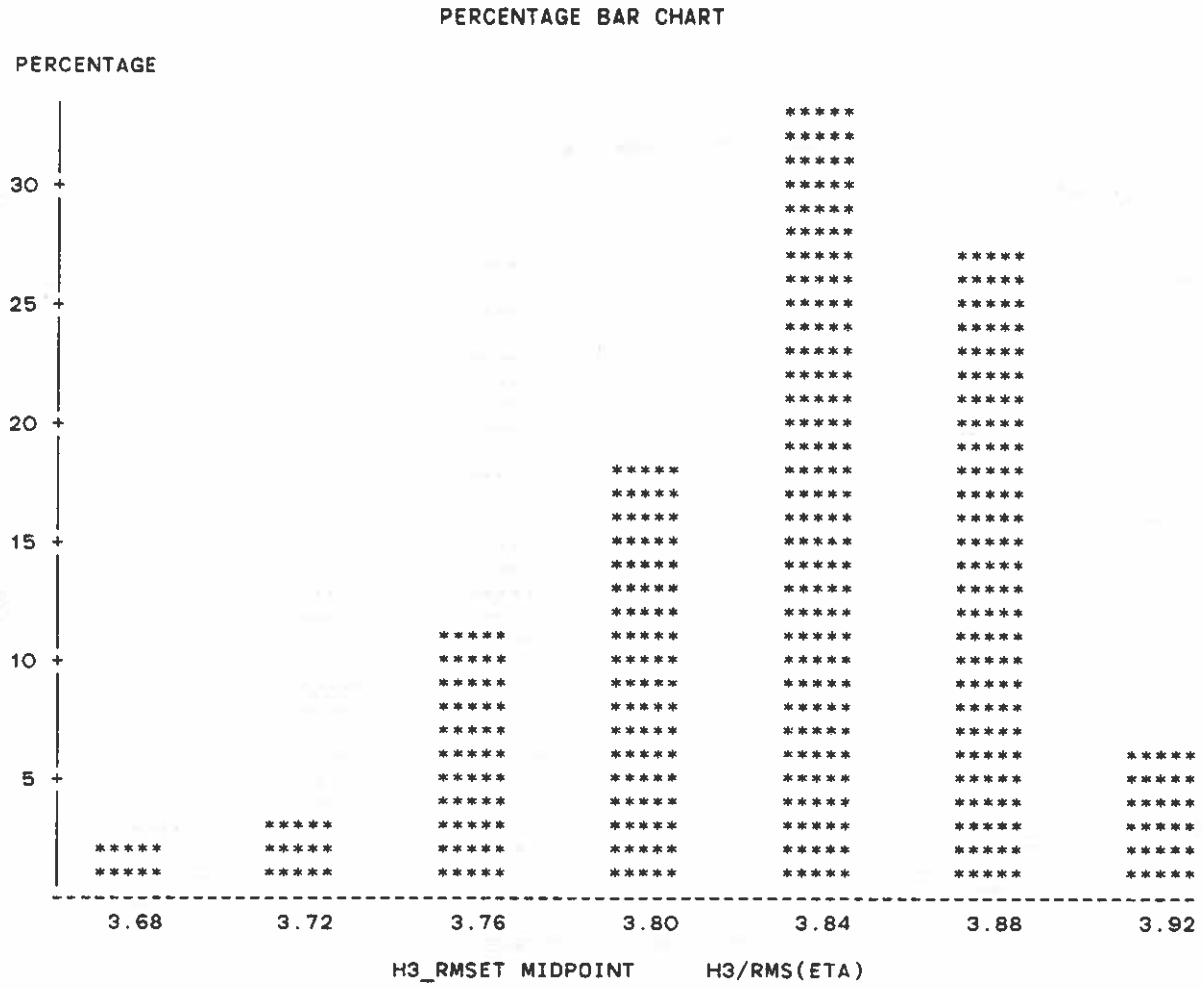

Figure 12. Frequency diagram of $H_{1/3}/\eta_{rms}$, data obtained by Waverider buoy B near Port Mansfield in March 1980.

Based on this value, the following relation is derived:

$$\frac{a_{rms}}{\sqrt{2}\eta_{rms}} = 0.949 \qquad (44)$$

The physical interpretation of this equation is that the nonlinear effect reduces the total energy by 5 percent, which is a contribution

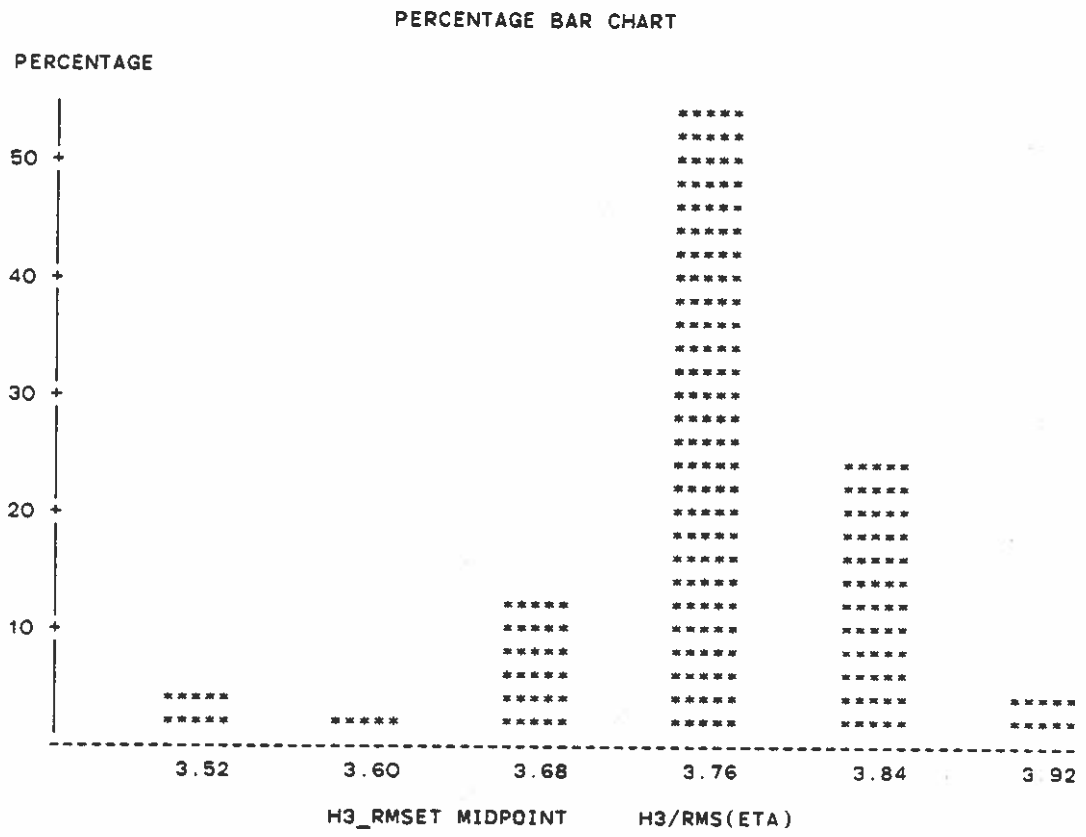

Figure 13. Frequency diagram of $H_{1/3}/\eta_{rms}$, data obtained by Waverider buoy B near Port Mansfield in August 1981.

of kinetic energy. Longuet-Higgins (1980) showed the potential energy was not altered by either the linear or the nonlinear process. Forristall (1978), Nolte and Hsu (1979), and Tayfun (1981) claimed other causes of discrepancies between the field observations and equation (25) were due to "effects associated with wide-band spectra"

and "wave breaking effect." Goda(1974a) stated that "the applicability of the Rayleigh distribution for the zero up-crossing wave heights is not influenced by the spectral shape, nor by the spectral width parameter." In spite of the wide band spectrum of the August 1981 data, the Rayleigh distribution seems applicable. A slight modification of the parameter should be made as suggested by equation (44). The Rayleigh distribution can be expressed in terms of H using equation (29).

$$f_H(H) = \frac{\pi H}{2\bar{H}^2} \exp\left[-\frac{\pi}{4}\frac{H^2}{\bar{H}^2}\right] \tag{45}$$

This equation involves neither η_{rms} nor a_{rms}. The exceedance probability P(x) of the normalized wave height x=H/H has a very simple form

$$P(x) = \exp\left[-\frac{\pi}{4}x^2\right] \tag{46}$$

Having obtained the mean wave height by means of the zero up-crossing method, the estimated H can be used in the Rayleigh distribution without making any modification. In practice, η_{rms}, is commonly determined by

$$\eta_{rms}^2 = \frac{1}{N}\sum_{i=1}^{N}(\eta_i - \bar{\eta}^2) \tag{47}$$

where $\bar{\eta}$ is the average of $\{\eta_i\}$.

If it is desired to estimate the wave parameters from η_{rms}, then the parameter of the Rayleigh distribution, equation (45), should be estimated by

$$\bar{H} = 0.947\sqrt{2\pi}\eta_{rms} \tag{48}$$

Marginal Distribution of Wave Periods

In contrast with the Rayleigh distribution for the wave height, which has been successfully employed in the practical applications, the marginal distribution of wave period has not been successfully developed from theory to-date. Several empirical distributions have been proposed. The first effort was made by Putz (1952). Based on 25 data sets, it was found that the wave period fitted the Gamma type distribution well. The proposed cdf is

$$F_T(T) = \frac{1}{\Gamma(P)} \int_o^u X^{P-1} e^{-x} dx \qquad (49)$$

where

$$u = P\left[1 + \frac{T-\bar{T}}{S_T}\right]$$

$$S_T = \left\{\frac{1}{N} \sum_{i=1}^{N} (T_i-\bar{T})^2\right\}^{\frac{1}{2}}$$

$$P = \frac{4 \; S_T^3}{\frac{1}{N} \sum_{i=1}^{N} (T_i-\bar{T})^3}$$

where $\Gamma(.)$ is the Gamma function, and T is the average wave period. However, only swell data were used, therefore Equation (49) is not applicable to the short period waves, such as wind waves.

Bretschneider (1959) found that the Rayleigh distribution fitted the wave length L well, based on the empirical data, i.e:

$$f_\lambda(\lambda) = \frac{\pi}{2} \lambda e^{-\frac{\pi}{4}\lambda^2} \qquad (50)$$

where λ is the normalized wave length, i.e:

$$\lambda = \frac{L}{\bar{L}}$$

where $\bar{L}$= the average wave length

According to linear wave theory, the wave length is proportional to the square of the wave period. Bretschneider defined the following relation

$$\lambda = k\tau^2 \tag{51}$$

where k is a constant, and τ is the normalized wave period, i.e.

$$\tau = \frac{T}{\bar{T}}$$

where $\bar{T}$= the average wave period.

After appropriate probability transformation, the pdf of τ can be obtained as

$$f_\tau(\tau) = 2.7\tau^3 e^{-0.675\tau^4} \tag{52}$$

It is easy to show that this pdf has the standard form of the Weibull distribution, i.e.:

$$f_\chi(x) = ab\chi^{b-1}\exp[-ax^b],\ 0 \le x < \infty \tag{53}$$

for a=0.675 and b=4

The power transformation of the Gamma distribution is known as the Weibull distribution. Therefore the empirical results obtained by Putz and Bretschneider may suggest that the Weibull distribution is a viable candidate for explaining wave period variability.

Several other types of wave period distributions have been proposed in conjunction with the joint probability density function of wave height and period. These will be introduced in the next chapter.

Despite the lack of theoretical background on the wave period variability, several wave period parameters were investigated based on two consecutive data records: 97 data sets in March 1980 and 54 data sets in August 1981. The following parameters were investigated:

For average wave period $\overline{T}$, the notation used is TA,

for significant wave period $T_{1/3}$, the notation used is T3,

for 1/10-th wave period $T_{1/10}$, the notation used is T10,

for maximum wave period T_{max}, the notation used is TMX.

The zero up-crossing method was adopted for the definition of individual waves, as discussed in the previous section. Since the magnitude of each of the parameters is not of particular interest in the description of wave period variability, the following ratios are discussed: $T_{max}/T_{1/3}$, $T_{1/10}/T_{1/3}$ and $T_{1/3}/\overline{T}$. The frequency histograms are shown in Figures 14 - 19, and Tables 7 - 10 give summaries of statistics for these ratios. It should be noted that the means of $T_{max}/T_{1/3}$ and $T_{1/10}/T_{1/3}$ are very close to 1.0, in other words, T_{max}, $T_{1/10}$ and $T_{1/3}$ may be practically taken as single values. The same relation was proposed by Goda (1974b). The mean of $T_{1/3}/\overline{T}$ obviously differs from 1.0. The overall mean of $T_{1/3}/\overline{T}$ is obtained as

$$\overline{T_{1/3}/\overline{T}} = 1.255 \qquad (54)$$

Goda (1974b) obtained $T_{1/3} \doteq 1.07\bar{T}$ for deep water waves, and $T_{1/3} \doteq 1.23\bar{T}$ for shallow water waves. Taking into account the observation site depth of this study, Goda's data and the present study agree very well. The following approximate relations are suggested:

$$T_{max} \doteq T_{1/10} \doteq T_{1/3} \doteq 1.2\ \bar{T} \qquad (55)$$

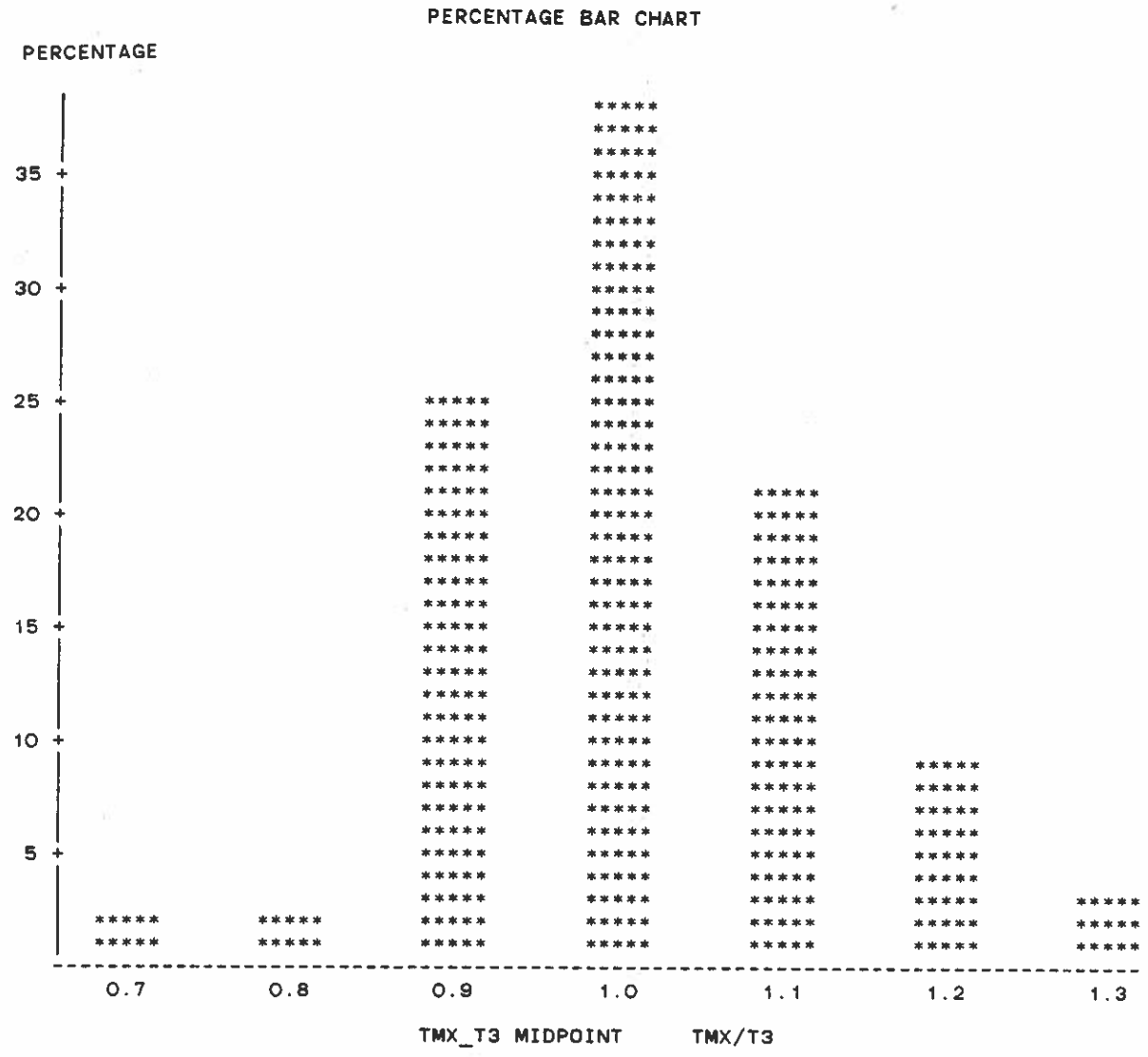

Figure 14. Frequency diagram of $T_{max}/T1_{/3}$, data obtained by Waverider buoy B near Port Mansfield in March 1980.

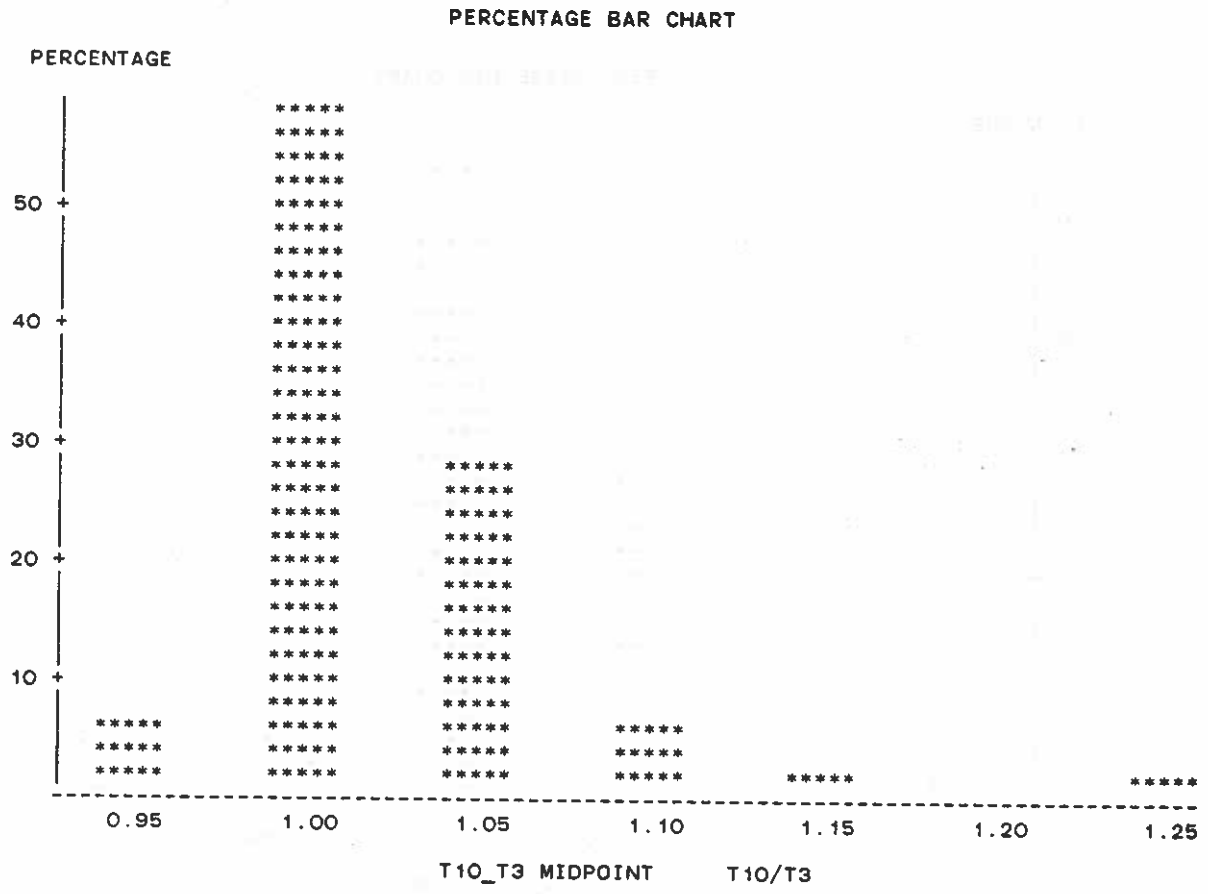

Figure 15. Frequency diagram of $T_{1/10}/T_{1/3}$, data obtained by Waverider buoy B near Port Mansfield in March 1980.

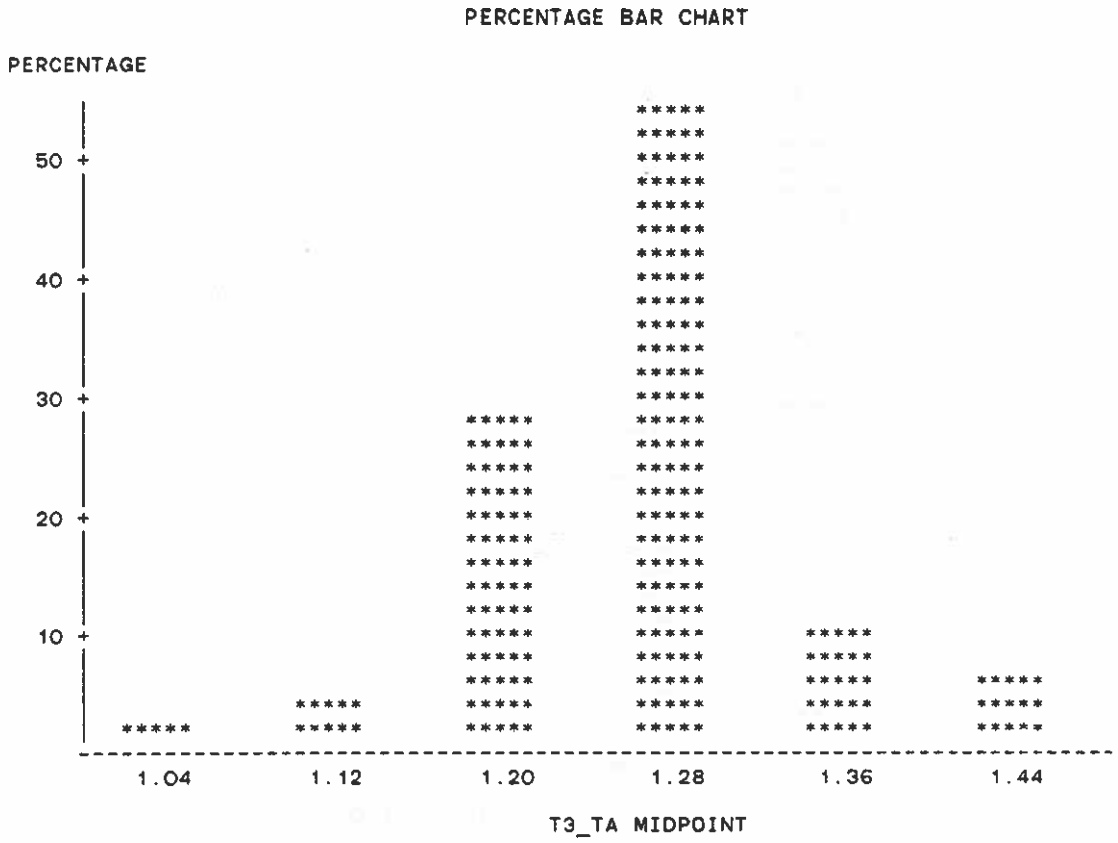

Figure 16. Frequency diagram of $T_{1/3}/\overline{T}$, data obtained by Waverider buoy B near Port Mansfield in March 1980.

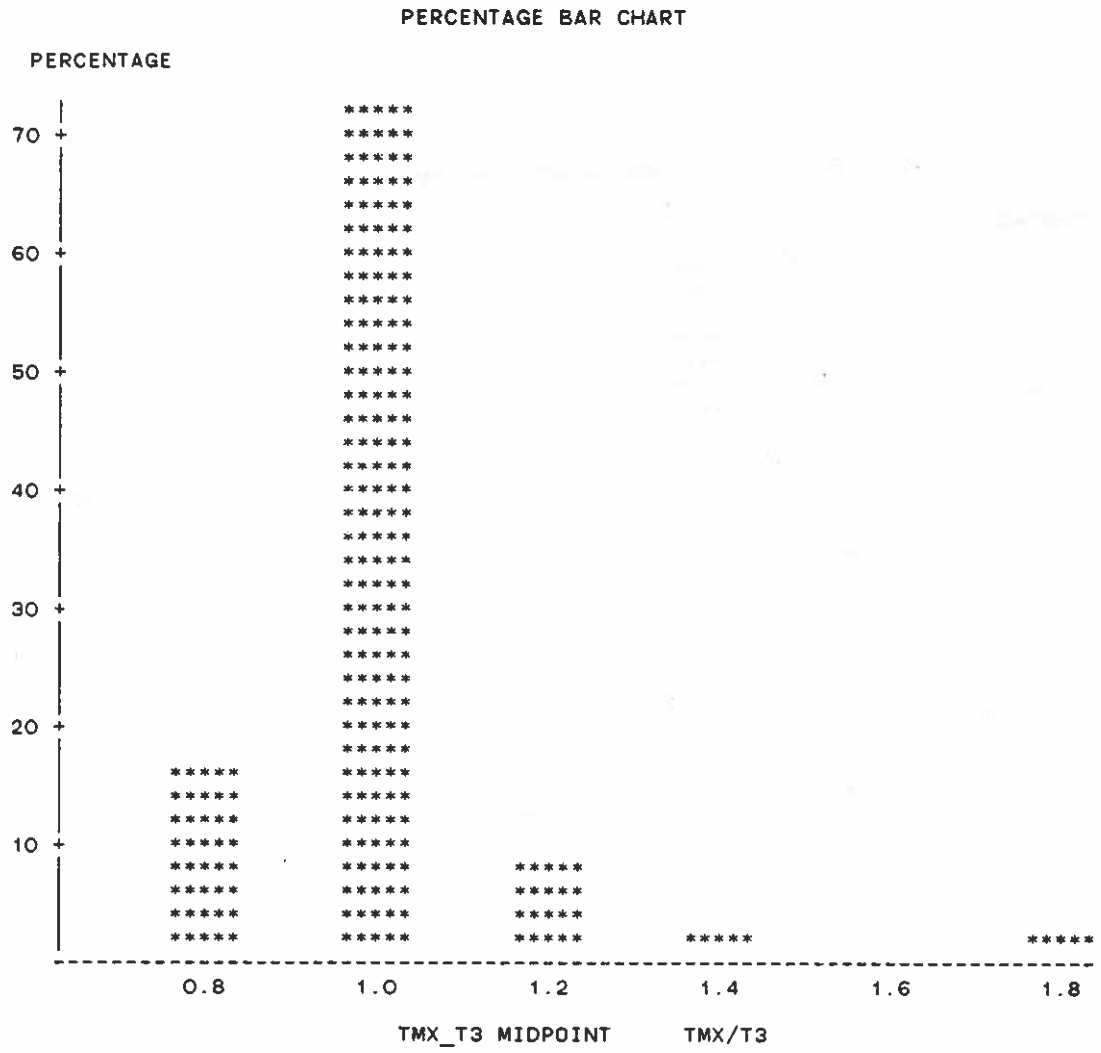

Figure 17. Frequency diagram of $T_{max}/T_{1/3}$, data obtained by Waverider buoy B near Port Mansfield in August 1981.

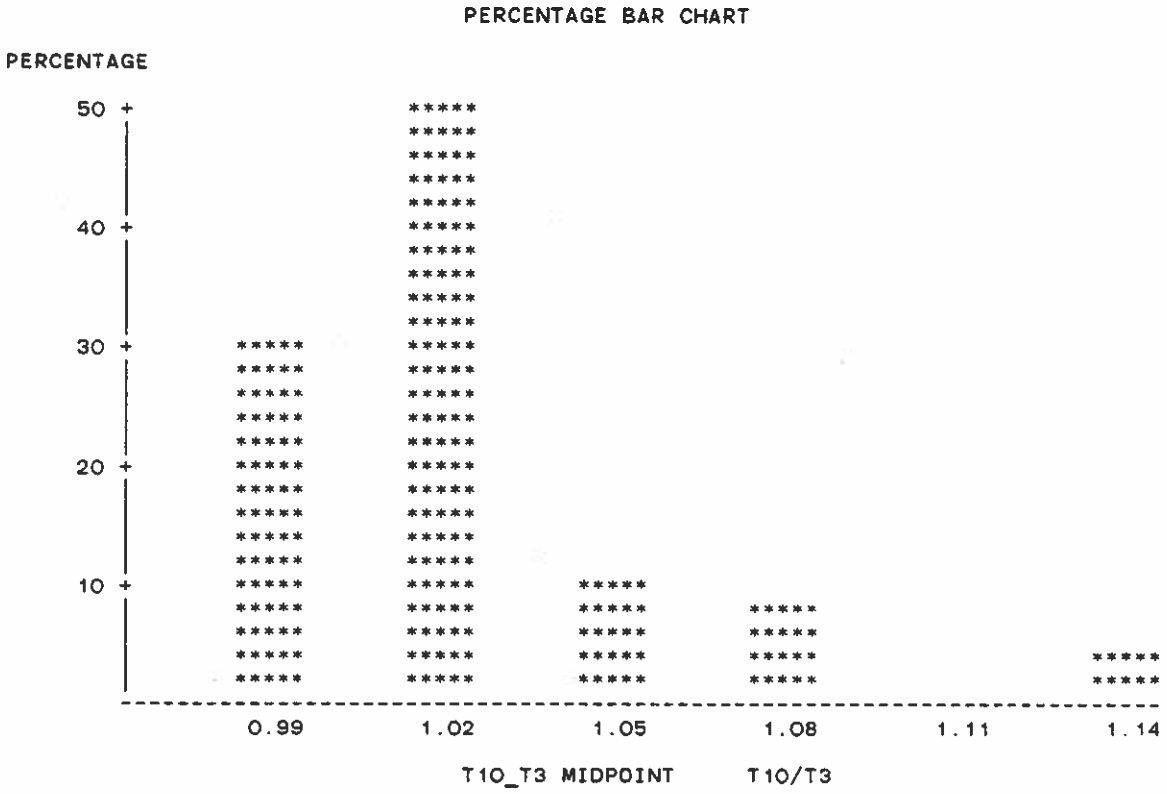

Figure 18. Frequency diagram of $T_{1/10}/T_{1/3}$, data obtained by Waverider buoy B near Port Mansfield in August 1981.

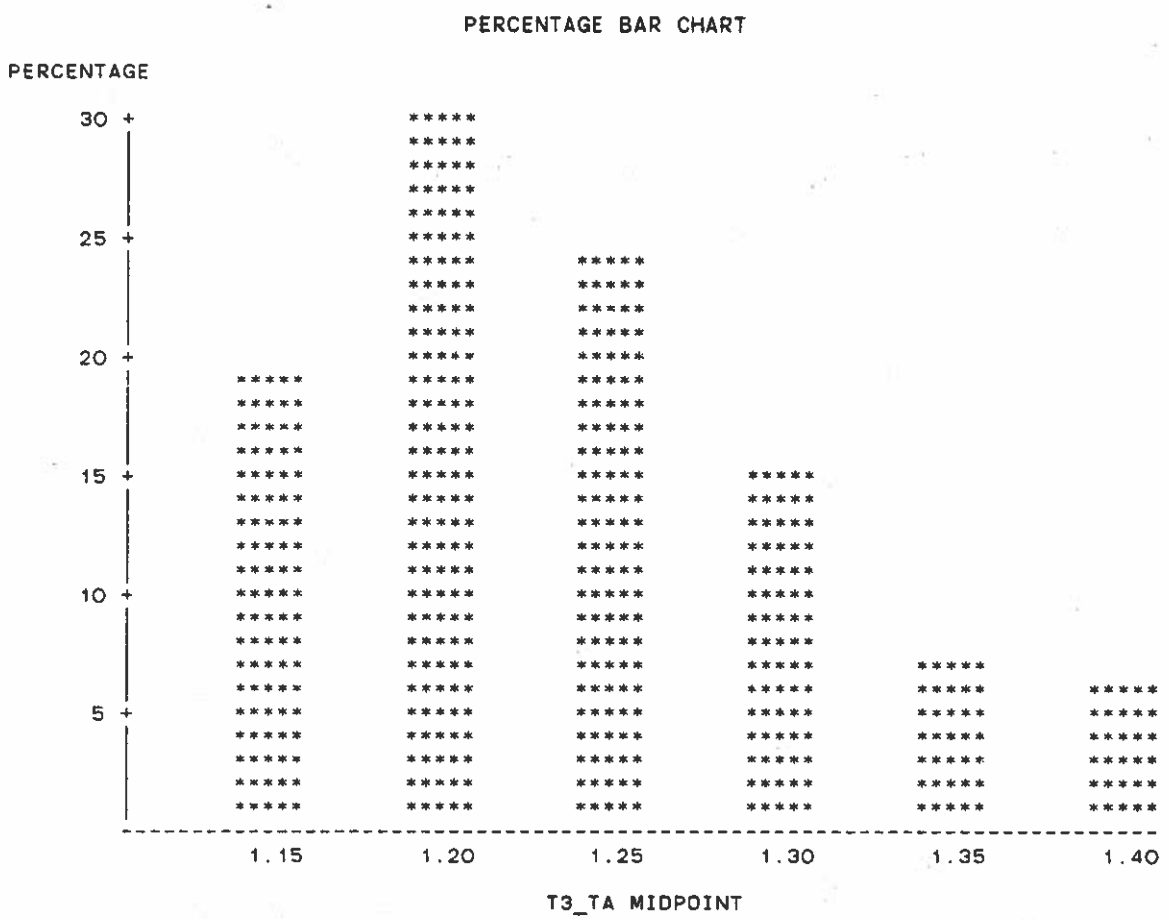

Figure 19. Frequency diagram of $T_{1/3}/\overline{T}$, data obtained by Waverider buoy B near Port Mansfield in August 1981.

JOINT DISTRIBUTION OF WAVE HEIGHTS AND PERIODS

General

Bretschneider (1959) found that the marginal distributions of H and T followed the Rayleigh distributions. Assuming that H and T were independent random variables, he proposed a joint pdf whose form was a product of two independent Rayleigh distributions. However, H and T are jointly distributed random variables, and as waves travel a long distance (swell), the wave heights and periods tend to be well correlated. Battjes (1969) examined the statistical properties of the bivariate Rayleigh distribution, which was originally derived by Uhlenbeck (1943) and Rice (1944,1945) in the communication field. The results have not been widely applied in wave studies.

The theoretical derivation of the joint pdf of wave heights and periods was addressed by Longuet-Higgins (1975) under the assumption of a linear process and narrow band spectrum. The theoretical distribution of T is symmetrical with respect to $T=\overline{T}$, the mean wave period, and the distribution extends to the negative range. Chakrabarti and Cooley (1977) demonstrated that the actual distribution of T exhibited asymmetric behavior. Venezian (1983) modified the Longuet-Higgins distribution, to eliminate the negative wave period range. Tayfun (1983) examined the effect of the spectrum band width on the Longuet-Higgins distribution. Kimura (1981) found that a bivariate Weibull distribution fitted the simulated data

well, and the correlation of H and T was related to the spectral band width.

The studies mentioned above are considered as parametric density estimation. Nonparametric density estimation is a major topic in modern statistics. Extensive research has been conducted by mathematical statisticians. Several different techniques have been proposed, but there still remains the question of which technique is superior. Nonparametric density estimation involves smoothing parameters for any type of technique. A major difficulty that arises in various techniques is finding an objective way to choose the smoothing parameters. Spline functions in nonparametric density estimation were applied and a criterion was proposed for objectively choosing the smoothing parameters.

The purpose of this chapter is to introduce available parametric joint pdf and nonparametric density estimation techniques. Since the major objective of this study was to develop a technique for summarizing a large amount of wave data in concise form, much time was spent on developing a nonparametric density estimation by means of the spline function. Thus, the parametric joint pdf was not examined in detail.

The Parametric Model of the Joint Probability Density Function

An extensive literature survey was conducted to seek the available information on joint pdf of wave heights and periods. A limited amount

of research has been conducted on this type of problem. This section will introduce several useful joint pdf of wave heights and periods.

Bretschneider (1959) proposed a joint pdf of wave heights and periods which is a product of marginal distributions, namely

$$f_{XY}(x,y) = 1.35\pi x e^{-\frac{\pi}{4}x^2} y^3 e^{-0.675y^4} \tag{56}$$

where $x = \frac{H}{\bar{H}}$

$y = \frac{T}{\bar{T}}$

H, $\bar{H}$, T and $\bar{T}$ are the wave height, the mean wave height, the wave period and the mean wave period, respectively. The marginal distributions are:

$$f_X(x) = \frac{\pi}{2} e^{-\frac{\pi}{4}x^2} \tag{57}$$

$$f_Y(y) = 2.7y^3 e^{-0.675y^4} \tag{58}$$

Since $f_{XY}(x,y)=f_X(x)f_Y(y)$, the correlation coefficient is always zero. This is not realistic, since the wave heights and periods appear correlated.

Battjes (1969) examined the bivariate Rayleigh distribution, which was originally developed by Uhlenbeck (1943). The bivariate Rayleigh distribution of two random variables U and V, whose means are 1.0, is expressed as

$$f_{UV}(u,v) = \frac{\pi^2}{4} \frac{uv}{1-k^2} \exp\left[-\frac{\pi}{4} \frac{u^2+v^2}{1-k^2}\right] I_o\left(\frac{\pi}{2} \frac{k}{1-k^2} uv\right) \tag{59}$$

for $u \geq 0$ and $v \geq 0$

where k is a shape factor which lies between 0 and 1, and $I_o(.)$ is the modified Bessel function of zero-th order. This distribution has not been successfuly applied for the joint variabilty of wave heights and periods. The examination of this distribution was not pursued. Recently Kimura (1981) examined the applicability of the bivariate Weibull distribution. Since the Rayleigh distribution is a special case of the Weibull distribution, Kimura's work represents a generalized version of previous density estimation. In fact, most of the empirical density functions have been proposed for the marginal pdf of wave heights by virtue of the Weibull distribution. The bivariate Weibull pdf of two random variables U and V is written as

$$f_{UV}(u,v) = \frac{mn}{4ab(1-k^2)} u^{m-1} v^{n-1} \exp\left[-\frac{1}{2(1-k^2)}\left(\frac{u^m}{a} + \frac{v^n}{b}\right)\right] \times I_o\left(\frac{k^2}{a^{1/2} b^{1/2}(1-k^2)} u^{m/2} v^{n/2}\right) \tag{60}$$

for $u \geq 0$ and $v \geq 0$

where m, n, a, b and k are the shape parameters, and $I_o(.)$ is the modified Bessel function of zero-th order. Evaluating the mean of U and V, the respective means are expressed as

$$E[U] = 2^{1/m} a^{1/m} \Gamma\left(\frac{1+m}{m}\right) \tag{61}$$

$$E[V] = 2^{1/n} b^{1/n} \Gamma\left(\frac{1+n}{n}\right) \tag{62}$$

It may be convenient to take the means equal to 1.0, then the random variables are treated as normalized by their respective means. This gives:

$$a = \tfrac{1}{2}\left[\Gamma\left(\frac{1+m}{m}\right)\right]^{-m} \tag{63}$$

$$b = \tfrac{1}{2}\left[\Gamma\left(\frac{1+n}{n}\right)\right]^{-n} \tag{64}$$

It is easy to show (59) is a special case of (60); when m=n=2, (60) reduces to (59). The marginal distributions of U and V are:

$$f_U(u) = \frac{m}{2a}\, u^{m-1} \exp\left[-\frac{u^m}{2a}\right] \tag{65}$$

$$f_V(v) = \frac{n}{2b}\, v^{n-1} \exp\left[-\frac{v^n}{2b}\right] \tag{66}$$

Kimura made a series of random wave simulations in a wave tank to obtain the empirical joint pdf of wave heights and periods, taking U as the wave height and V as the wave period. The shape factors were selected to fit the observed data. He found the shape factor m was close to 2.0. Note m=2.0 corresponds to the Rayleigh distribution. He also stated the shape factor m correlated with the skewness of surface elevation. The shape factor, n, varies depending on the spectral shape, and has a strong association between n and the spectral width parameter. The correlation coefficient has a close connection with the spectral width parameter. Kimura concluded that all the parameters could be estimated from the spectral width parameters. However, this may be possible only if the spectra have a single peak.

Longuet-Higgins (1975) extended the Rayleigh law to obtain the theoretical joint pdf of wave heights and periods. The normalized wave height X and period Y, whose means are equal to 1.0, have the following joint pdf:

$$f_{XY}(x,y) = \frac{\pi x^2}{4\nu} \exp\left[-\frac{\pi}{4}x^2\left\{1 + \frac{(y-1)^2}{\nu^2}\right\}\right] \tag{67}$$

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

where v is a spectral width parameter. The marginal distributions are obtained by integrating $f_{XY}(.,.)$ with respect to x or y.

$$f_X(x) = \frac{\pi}{2} x \exp[-\frac{\pi}{4}x^2], \quad x \geq 0 \tag{68}$$

$$f_Y(y) = \frac{1}{2\nu[1 + \frac{(y-1)^2}{\nu}]^{3/2}}, \quad -\infty < y < \infty \tag{69}$$

It can be shown that the correlation coefficient is zero; hence, X and Y are linearly uncorrelated random variables. However, the product of marginal distributions is not the joint pdf, therefore X and Y are not independent. Since wave height and period have been observed as highly correlated random variables, this is not a desirable feature. Another undesirable property of the Longuet-Higgins distribution is that the period has a negative range. Venezian, Bretschneider and Jagannathan (1980) describe a modification which removes the negative range from the wave period. The modified Longuet-Higgins distribution is

$$f_{XY}(x,y) = \frac{\pi x^2}{2\nu} \exp[-\frac{\pi}{4}x^2 \{1 + \frac{(y-1/y)^2}{\nu^2}\}] \tag{70}$$

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.

The shape parameter for the Longuet-Higgins distribution may be evaluated from spectral analysis. Chakrabati and Cooley (1977) confirmed the practical applicability of the Longuet-Higgins

distribution provided the spectrum is a narrow band single peak spectrum. However, in many cases, the spectral shape is a broad spectral width and sometimes has multiple peaks. The multiple spectral peaks may suggest the multimodal joint pdf of wave heights and periods. The aforementioned density function never exhibits the multimodal peaks. Thus, nonparametric density estimation should be taken into consideration as an alternative.

Nonparametric Density Estimation

Nonparametric estimation may be categorized into two groups; a) series estimators and b) kernel estimators. The series estimator approximates a pdf, which may or may not be known, in terms of an infinite series. The kernel density estimator is a natural extension of the classical histogram estimator. Parzen (1979) proposed an alternative method to estimate an unknown density extending the idea of the autoregressive spectra.

Series Estimators

A mathematical theorem states that an analytic function (such as a pdf) can be expanded into an infinite Maclaurin's series.

$$f(x) = \sum_{j=0}^{\infty} a_j x^j \qquad (71)$$

where

$$a_j = \frac{f^{(j)}(0)}{j!}$$

and $f^{(j)}(0)$ is the j-th derivative of $f(x)$ evaluated at $x=0$. It is possible to evaluate the coefficients $\{a_j\}$ if the function $f(x)$ is known. In general, if a finite number of samples $\{x_1, x_2, \ldots, x_N\}$ is available, then it is impossible to estimate $\{a_j\}$ from the samples. However, it may be possible to estimate a finite number of coefficients, say m+1, of the following series:

$$f_m(x) = \sum_{j=0}^{m} \hat{a}_j x^j \tag{72}$$

The Gram-Charlier series has the above form. The coefficients $\{\hat{a}_j\}$ are usually estimated from the sample moments. Huang and Long (1980) fitted the probability density of surface elevation employing the Gram-Charlier series, and showed the skewed nature of the surface elevation. Unfortunately the estimated pdf exhibits an undesirable feature of negativity.

A more general expansion of the series estimator is given by

$$f(x) = \sum_{j=-\infty}^{\infty} c_j \phi_j(x) \tag{73}$$

where $\{c_j\}$'s are real valued constants, and $\{\phi_j(x)\}$ are real or complex functions. One would like to estimate the $f(x)$ for $-\infty<x<\infty$, however, the samples have a finite domain, say $[a,b]$. Therefore, the nonparametric density estimators usually have a finite support $[a,b]$. The estimated $f(x)$ is a truncated density of true density, say $g(x)$, such that

$$f(x) = \frac{g(x)}{\int_a^b g(x)dx} \tag{74}$$

The support of f(x) in the equation (73) is [a,b]. A general strategy to approach this type of problem is to employ a complete orthonormal set. A complete orthonormal set $\{\phi_j(x)\}$ defined in a finite domain [a, b] with respect to a weighting function w(x) has the property

$$\int_b^a \phi_j(x)\phi_k(x)w(x)dx = \delta_{jk} \qquad (75)$$

where $$\delta_{jk} = \begin{cases} 1.0 & \text{if } j=k \\ 0.0 & \text{otherwise} \end{cases}$$

The harmonic function and the orthogonal polynomials, such as the Jacobi, Legendre, Chebyshev and Laguerre polynomials, are usually applied.

Kronmal and Tarter (1976) used the complex harmonic function for $\{ \phi_j \}$ to estimate a density function:

$$\phi_j = e^{-2\pi ijx} \qquad (76)$$

where $i^2 = -1.0$

Hence the estimate $\hat{f}_x(x)$ of the pdf $f_x(x)$ for the sample $\{x_1, x_2, \ldots, x_N\}$ can be expressed as

$$\hat{f}_x(x) = \sum_{m=-\infty}^{\infty} \hat{C}_m e^{-2\pi imx} \qquad (77)$$

where $$\hat{C}_m = \frac{1}{N} \sum_{\ell=1}^{N} e^{-2\pi imx_\ell}$$

They suggest $$\hat{C}_m = 0 \text{ if } \hat{C}_m \hat{C}_{-m} < \frac{2}{N+1}$$

where N is the number of samples. Large values of m produce a wiggly form of $\hat{f}(x)$. Kronmal and Tarter suggest the maximum order of m=10. Recently, Woodfield (1982) successfully applied a similar technique to estimate bivariate density functions in the quantile domain. He applied an objective way to choose the optimal order m based on the Akaike Information Criterion, AIC (Akaike,1974).

Kernel Estimators

Most kernel estimators are based on the idea of a histogram. Rosenblatt (1956) introduced this type of a technique, and Parzen (1962) proposed the detailed theoretical justification. The Rosenblatt estimator of a given sample $\{x_1, x_2, \ldots, x_N\}$ is

$$\hat{f}_N(x) = \frac{\#\text{ of sample points in } (x-h, x+h)}{2Nh} \tag{78}$$

where h is a real constant, which should be a function of sample size and population density. Note that

$$\hat{f}_N(x) = \frac{\tilde{F}_N(x+h) - \tilde{F}_N(x-h)}{2h} \tag{79}$$

where $\tilde{F}_N(x)$ is the empirical distribution function

$$\tilde{F}_N(x) = \frac{\#\text{ of sample points} \leq x}{N} \tag{80}$$

Scott and Thompson (1983) proposed an extension of the histogram estimator which they called the averaged shifted histogram. The estimator is the average of successive adjacent bins of shifted histograms. The extension to the multivariate problem is relatively simple. They successfully applied this technique to represent the

multidimensional data in terms of the density function. It should be noted that $\hat{f}_N(x)$ mentioned above is a discontinous step function, therefore, it does not have a continuous derivative. To overcome this shortcoming, the smoothing operator, the kernel K(u), may be introduced

$$\hat{f}_N(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-y}{h}\right) d\tilde{F}_N(y)$$

$$= \frac{1}{hN} \sum_{j=1}^{N} K\left(\frac{x-x_j}{h}\right) \qquad (81)$$

It can be shown when K(u)=1/2 for $|u|<1.0$, and K(u)=0 otherwise, the equation (81) is equivalent to (79). Several useful kernels are shown in Tapia and Thompson (1978, p.60). The kernel K(u) must satisfy the following properties (Woodfield, 1982):

i) $\sup_x |K(x)| < \infty$

ii) $\int_{-\infty}^{\infty} |K(x)| dx < \infty$.

iii) $\lim_{x \to \infty} |xK(x)| = 0$

iv) $\int_{-\infty}^{\infty} K(x)dx = 1.0$

Similar to the difficulty that a series estimator has in choosing the optimal order, the kernel estimator faces the problem of finding objectively a suitable window width, (or bin width), h. Most of the kernel estimators have employed some sort of subjective way to adapt the window width. Depending on the kernel used, negative values may appear in the estimated density function. However, the positivity of density function is constrained by the development of some kernel estimators.

Boneva, Kendall and Stefanov (1971) proposed the "histospline" estimator which is a smoothed estimator of unknown probability density function $f(.)$ with cdf $F(.)$ having a finite support $[a, b]$. The histospline estimator $\hat{f}(x)$ is the mimimizer of

$$\int_a^b (\hat{f}(x))^2 dx \tag{82}$$

subject to

$$\int_{x_j}^{x_{j+1}} \hat{f}(t)dt = \tilde{F}_N(x_{j+1}) - \tilde{F}_N(x_j) \tag{83}$$

for $j=0,1,\ldots,m$

where $x_o = a$ and $x_{m+1} = b$, and $a = x_o, \ldots, x_{m+1} = b$ are an equal partition of $[a, b]$ for the histogram. The base function of the histospline is the deltaspline named by Boneva et. al. in their paper, which has the form of a piecewise cubic spline. The histospline estimator is in a class of C^2 function, which is an absolutely continuous function up to the second order derivative. The idea behind the deltaspline is to form a polynomial which acts like the Dirac delta function. It should be noted that the deltaspline has negative tails. Some confusion concerning histospline appears in Tapia and Thompson(1978). They stated that the histospline was a smoothing estimator of the empirical distribution function, but, in fact, it was a smoothing estimator of the histogram.

Since ordinary spline functions do not recover the monotonicity of the empirical distribution function without imposing the side constrains, the density estimator can become negative. Shape

preserving spline interpolation has become an important research area in the numerical analysis field. The B-splines have been extensively used in the present study; details concerning the spline functions are discussed in Appendix A.

The B-spline has a high computational efficiency and is easily extended to higher dimensions even though it lacks the shape preserving property. Let the domain of the spline estimator [a, b] be divided into k-1 equal intervals, such as

$$[a,b] = \bigcup_{j=1}^{k-1} [t_j, t_{j+1}] \tag{84}$$

and

$$h = t_{j+1} - t_j = \frac{b-a}{k-1}$$

The spline estimator is

$$\hat{F}(x) = \sum_{j=1}^{m} C_j B_j(x) \tag{85}$$

where m is the dimension of the B-spline, and

$$m = k+2$$

subject to the interpolation conditions

$$\hat{F}(t_j) = \tilde{F}_N(t_j) \tag{86}$$

$$\frac{d\hat{F}(a)}{dx} = \frac{d\hat{F}(b)}{dx} = 0 \tag{87}$$

where $\tilde{F}_N(\cdot)$ is the empirical distribution function. Similar to the other kernel estimators, the spline estimator changes the estimated density function in shape depending on the bin width h. An objective method to choose h is proposed. The optimal h for the sample $\{x_1, x_2, .., x_N\}$ may be obtained by minimizing the Objective Least square Norm (OLN), which is defined as

$$OLN = \begin{cases} \delta & \text{if } N-m-1>0 \\ \infty & \text{otherwise} \end{cases} \tag{88.1}$$

where

$$\delta = \frac{1}{N-m-1} \sum_{j=1}^{N} (\tilde{F}_N(x_j) - \hat{F}(x_j))^2 \tag{88.2}$$

The reason for subtracting 1 from the denominator is that the $\{x_j\}$ are normalized by the average value prior to estimates of density for the sake of computational efficiency. If the original sample is $\{y_j\}$, then the normalized sample $\{x_j\}$ is

$$x_j = \frac{y_j}{\bar{y}} \text{ for } j=1,2,\ldots,N \tag{89}$$

where

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{N} y_j$$

Therefore, the mean of a sample $\{x_j\}$ is 1.0. Having obtained the estimated pdf $f_X(.)$ of X, the pdf $f_Y(.)$ of Y can be estimated applying a simple transformation, i.e.

$$\hat{f}_Y(y) = \bar{y}\hat{f}_X(x) \tag{90}$$

The mean $\hat{\mu}_Y$ and variance $\hat{\sigma}_Y^2$ are straight forward.

$$\begin{aligned} \hat{\mu}_Y &= \int_{-\infty}^{\infty} y\hat{f}_Y(y)dy \\ &= \bar{y}\int_{-\infty}^{\infty} x\hat{f}_X(x)dx \\ &= \bar{y} \end{aligned} \tag{91}$$

$$\hat{\sigma}_y^2 = \int_{-\infty}^{\infty} (\bar{y}x - \hat{\mu}_Y)\hat{f}_X(x)dx$$
$$= \bar{y}^2\hat{\sigma}_X^2 \qquad (92)$$

Knowledge of $\hat{f}_X(.)$ is equivalent to knowledge of $\hat{f}_Y(.)$.

Joint pdf Estimation by Means of Tensor Product Splines

The bivariate extension of nonparametric density estimator by means of B-splines (Spline Density Estimator) will be introduced in this section. The technique was independently developed by the author, therefore much effort was spent on establishing a FORTRAN computer package. A similar technique using the tensor product B-spline was introduced by Bennett(1974), de Boor(1979) introduced an extremely efficient algorithm for the tensor product spline. Making use of de Boor's algorithm for the spline density estimator, the technique developed in this study is computationally much faster than Bennett's method.

First, it is necessary to define the empirical bivariate distribution function $\tilde{F}_N(x,y)$ of N pairs of samples $\{x_j, y_j\}_{j=1}^N$ for the random variables X and Y.

$$\tilde{F}_N(x,y) = \frac{\#\text{ of sample points} \le x \text{ and} \le y}{N} \qquad (93)$$

The function $\tilde{F}_N(.,.)$ has the same statistical properties as the univariate case and is a consistent and unbiased estimator of the population distribution function $F_{XY}(.,.)$. The bivariate histogram plays a basic role in the spline density estimator. Suppose N pairs of

samples $\{x_j,y_j\}_{j=1}^{N}$ exist in the real domain $[a_X, b_X]$ $[a_Y, b_Y]$, (which is a rectangle in the plane). Let the domain be divided into equal rectangular elements, such that

$$[a_X,b_X] = \bigcup_{j=1}^{k_X-1} [t_{j,X} - t_{j+1,X}] \tag{94.1}$$

$$[a_Y,b_Y] = \bigcup_{j=1}^{k_Y-1} [t_{j,Y} - t_{j+1,Y}] \tag{94.2}$$

and

$$h_X = t_{j+1,X} - t_{j,X} = \frac{b_X - a_X}{k_X-1} \tag{95.1}$$

$$h_Y = t_{j+1,Y} - t_{j,Y} = \frac{b_Y - a_Y}{k_Y-1} \tag{95.2}$$

The bivariate histogram density estimator is

$$\hat{f}_H(x,y;h_X,h_Y) = \frac{\#\text{ of samples in}[t_{j,X},t_{j+1,X}] \times [t_{j,Y},t_{j+1,Y}]}{h_X h_Y N} \tag{96}$$

where $t_{j,X} \le x \le t_{j+1,X}$

and $t_{j,Y} \le y \le t_{j+1,Y}$

The corresponding distribution function is

$$\hat{F}_H(x,y;h_X,h_Y) = \int_{a_X}^{x}\int_{a_Y}^{y} \hat{f}_H(u,v;h_X,h_Y)dudv \tag{97}$$

The population distribution function $F_{XY}(x,y)$ may be estimated by interpolating $\hat{F}_H(x,y;h_X,h_Y)$. The technique used here is the tensor product splines which is a tensor product of one dimensional B-splines. Let $\hat{F}_S(x,y;h_X,h_Y)$ be the smoothed distribution function of $\hat{F}_H(x,y;h_X,h_Y)$.

$$\hat{F}_S(x,y;h_X,h_Y) = \sum_{i=1}^{m_X}\sum_{j=1}^{m_Y} C_{ij}B_i(x)B_j(y) \tag{98}$$

where $m_X = k_X+2$

$m_Y = k_Y+2$

C_{ij}'s are constants

$B_i(.), B_j(.)$ are the cubic B-spline basis

The number of coefficients to be estimated is $m_X\, m_Y$; the coefficients are obtained by the following conditions:

$$\hat{F}_S(t_{i,X}, t_{j,Y}; h_X h_Y) = \hat{F}_H(t_{i,x}, t_{j,Y}; h_X, h_Y), \tag{99.1}$$

for $i=1,\ldots,k_X$ and $j=1,\ldots,k_Y$

$$\frac{d\hat{F}_S(a_X, t_{j,Y}; h_X, h_Y)}{dx} = \frac{d\hat{F}_S(b_X, t_{j,Y}; h_X, h_Y)}{dx} = 0, \tag{99.2}$$

for $j=1,\ldots,k_Y$

$$\frac{d\hat{F}_S(t_{i,X}, a_Y; h_X, h_Y)}{dy} = \frac{d\hat{F}_S(t_{i,X} b_Y; h_X, h_Y)}{dy} = 0, \tag{99.3}$$

for $i=1,\ldots,k_X$

$$\frac{d^2\hat{F}_S(a_X, a_Y; h_X, h_Y)}{dx\, dy} = \frac{d^2\hat{F}_S(b_X, b_Y; h_X, h_Y)}{dx\, dy} = 0, \tag{99.4}$$

$$\frac{d^2\hat{F}_S(a_X, b_Y; h_X, h_Y)}{dx\, dy} = \frac{d^2\hat{F}_S(b_X, a_Y; h_X, h_Y)}{dx\, dy} = 0 \tag{99.5}$$

The number of conditions provided above is $m_X\, m_Y = k_X k_Y + 2(k_X+k_Y)+4$, and the system of equations to obtain $\{ C_{ij} \}$ is completed. The matrix to be solved is a $(m_X \times m_Y) \times (m_X \times m_Y)$ matrix. In general, this is a large matrix so a vast amount of computational effort may be expected.

However, the computational load can be reduced by introducing de Boor's algorithm (de Boor,1979) which makes two small tridiagonal $(m_X \times m_X)$ and $(m_Y \times m_Y)$ matrices to be solved. This will be discussed in Appendix A.

It may be shown that $\hat{F}_S$ is a linear mapping of $\hat{F}_H$ in the spline space. The shape of $\hat{F}_H$ is significantly dependent on the choice of bin widths h_x and h_y, which are called the smoothing parameters. The wider they are, the smoother the estimated density function. The optimal smoothing parameters may be obtained by extending the idea of Objective Least square Norm (OLN) to two dimensions.

$$OLN = \begin{cases} \delta & \text{if } N-m-2>0 \\ \infty & \text{otherwise} \end{cases} \tag{100.1}$$

where

$$\delta = \frac{1}{N-m-2} \sum_{i=1}^{N} [\tilde{F}_N(x_i,y_i) - \hat{F}_S(x_i,y_i;h_x,h_y)]^2 \tag{100.2}$$

N is the number of samples

and m is the number of spline coefficients

To make programming easy, the sample variables are normalized by their average values. Suppose the original samples are $\{u_i, v_i\}_{i=1}^{N}$. The normalized samples $\{x_i, y_i\}_{i=1}^{N}$ can be obtained as follows:

$$x_i = \frac{u_i}{\bar{u}} \tag{101.1}$$

$$y_i = \frac{v_i}{\bar{v}} \tag{101.2}$$

where

$$\bar{u} = \frac{1}{N} \sum_{i=1}^{N} u_i \tag{101.3}$$

$$\bar{v} = \frac{1}{N} \sum_{i=1}^{N} v_i \tag{101.4}$$

Since the normalization process reduces two degrees of freedom, an additional number 2 is included in the penalizing factor. In total, one hundred different combinations of h_x and h_y are examined such that

$$h_x = 1/NIX \text{ for } NIX=1,2,\ldots,10 \tag{102.1}$$

$$h_y = 1/NIY \text{ for } NIY=1,2,\ldots,10 \tag{102.2}$$

It seems reasonable to check the proposed density estimation technique using the data from a known population density function. Figure 20 shows the scatter diagram of 500 samples generated from the independent bivariate normal distribution which is

$$f_{xy}(x,y) = \frac{1}{2\pi}\frac{1}{\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)\right] \tag{103}$$

where

$$\mu_x = 1.0$$

$$\mu_y = 1.0$$

$$\sigma_x = 1/5$$

$$\sigma_y = 1/5$$

Contour lines in Figure 20 are the probability density of generated samples. The true density function $f_{xy}(x,y)$ is depicted in Figure 21. The contour map of log values of the OLN for this data set is shown in Figure 22. A logarithmic scale is used to compress deviations of the OLN values. The peak of OLN appears at NIX =5, NIY =4. Based on these smoothing parameters, the best estimate of the

joint pdf for the proposed criteria is obtained and shown in Figures 23 and 24. The contour plots of the true and estimated densities were visually examined to check whether or not the proposed technique was acceptable. The location of the mode for both densities agrees well, but the peak of the estimated density is slightly less than the true density. Small negative values also appear in the estimated density. Over-all, the features of both densities agree well.

Another numerical evaluation of the density estimation technique was made generating 500 samples from the modified Longuet-Higgins distribution (equation (70), p.69). A series of numerical simulation runs was conducted for different values of the spectral width parameter. The true density function is shown in Figure 25, where the parameter is taken as 0.2. The true density function has a very narrow mode. A realization of 500 samples is depicted in Figure 26. Contour lines in Figure 26 show the true probability density of generated samples. Samples appear around Y=1.0, which corresponds to the mean of wave period. The estimated joint pdf is shown in Figures 27 and 28. The estimated pdf does not have a density peak as high as the true pdf; however, the basic feature of the true density function is recovered in the estimated pdf. The spline density estimator seems to work well and is acceptable, in a practical sense, even if it shows negative values of estimate.

The joint pdf of wave heights and periods was estimated employing the spline density estimator. A scatter diagram of a typical data set of wave heights and periods is produced from data obtained by a

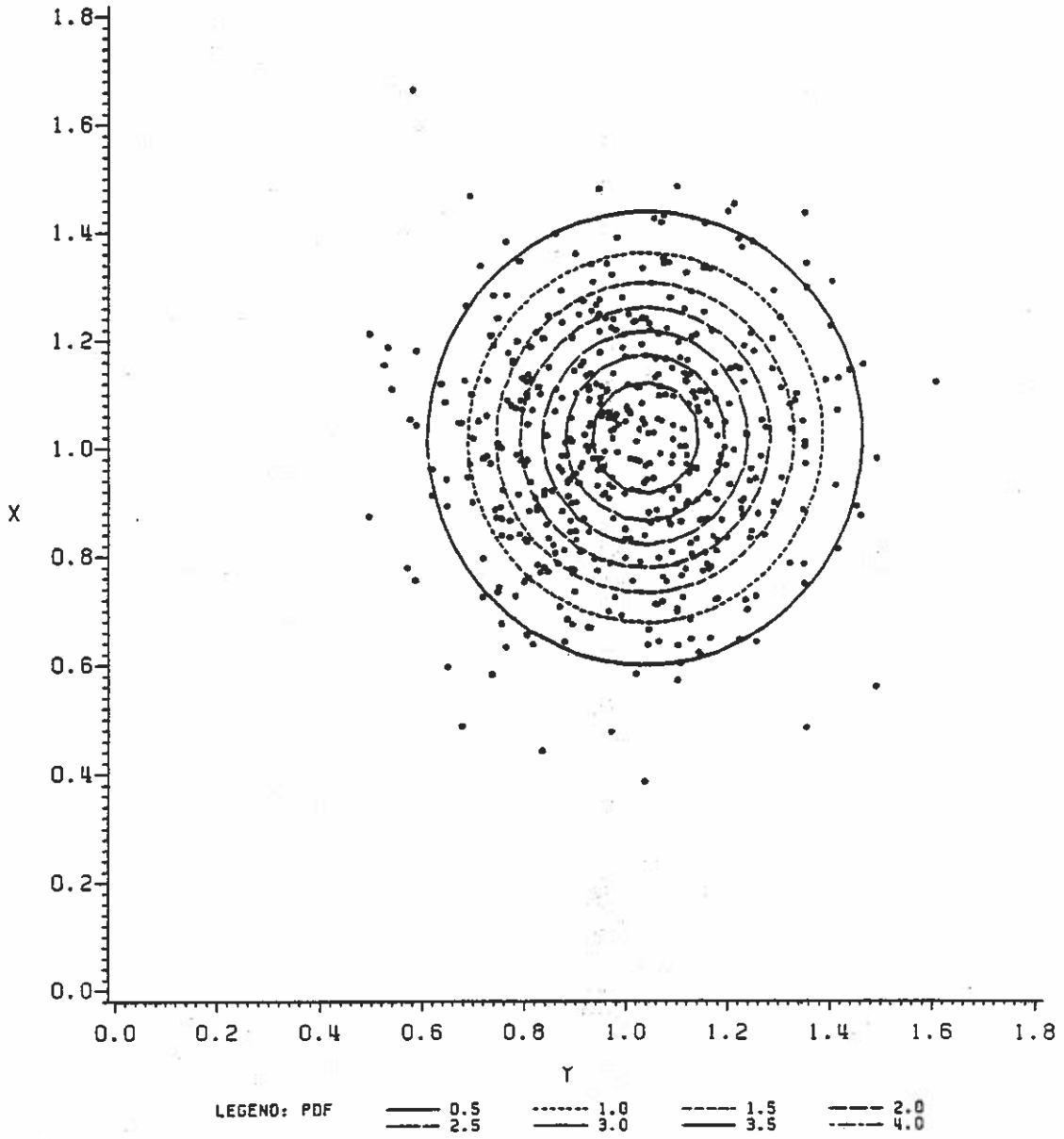

Figure 20. Scatter diagram of 500 samples generated from independent bivariate normal distribution. (Contour lines are the true probabilty density)

Waverider buoy at 3 p.m. on March 13, 1980, near Port Mansfield, Texas. There are 240 samples shown in Figure 29. The OLN (see Figure 30) chooses the optimal smoothing parameters NIH=2 and NIT=4,

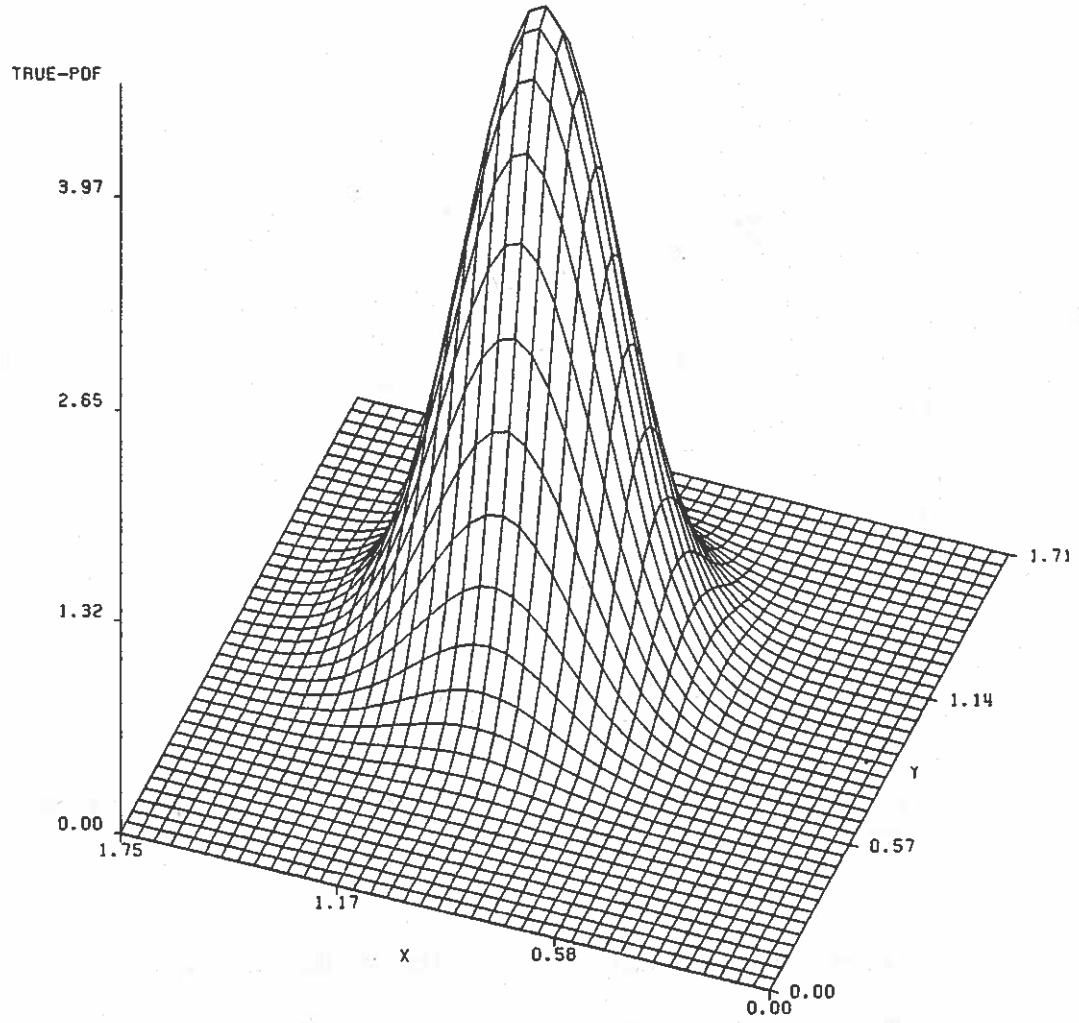

Figure 21. The true independent bivariate normal density function. (means are 1.0, variances are 0.04 and correlation coefficient is 0)

here NIH, NIT correspond to NIX and NIY respectively in the previous notation. The best estimate of joint pdf is shown in Figure 31. The mutimodal feature of the pdf can be seen in Figure 31. The contour

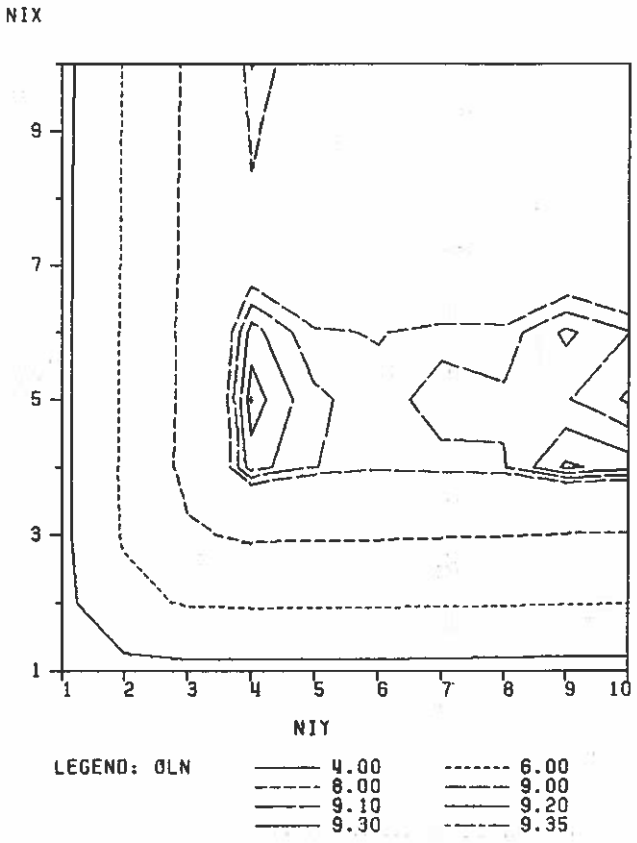

Figure 22. The contour map of -log(OLN) for the 500 generated samples. (The peak occurs at NIX=5, NIY=4)

plot of the estimated pdf clarifies the location of modes and is shown in Figure 32. The negative values may be replaced by zeros without significantly altering the result. For better visualization of the estimated joint pdf, the flat surface appearing in Figure 33 corresponds to the negative values. Comparing this data set with one obtained in a different season, Figure 34 shows the best estimate of joint pdf based on 319 samples obtained at 3 a.m., August 13, 1981, near Port Mansfield. The corresponding contour map is shown in Figure

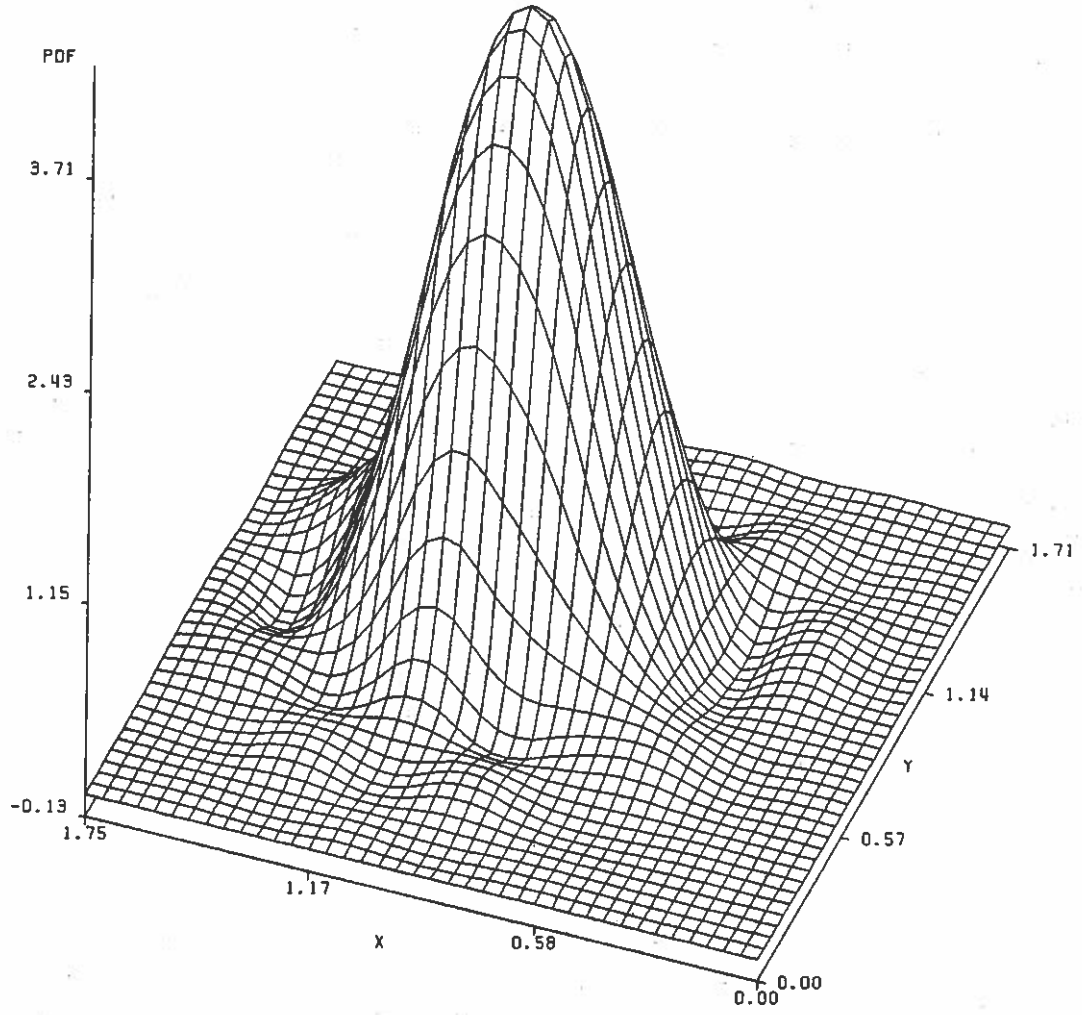

Figure 23. The estimated independent bivariate normal density function. (means are 1.0, variances are 0.04 and correlation coefficient is 0)

35. The multimodal peaks appear in the estimated density. The peaks possibly indicate the different wave generation sources. The joint pdf estimation for the data obtained in March 1980 and August 1981 was

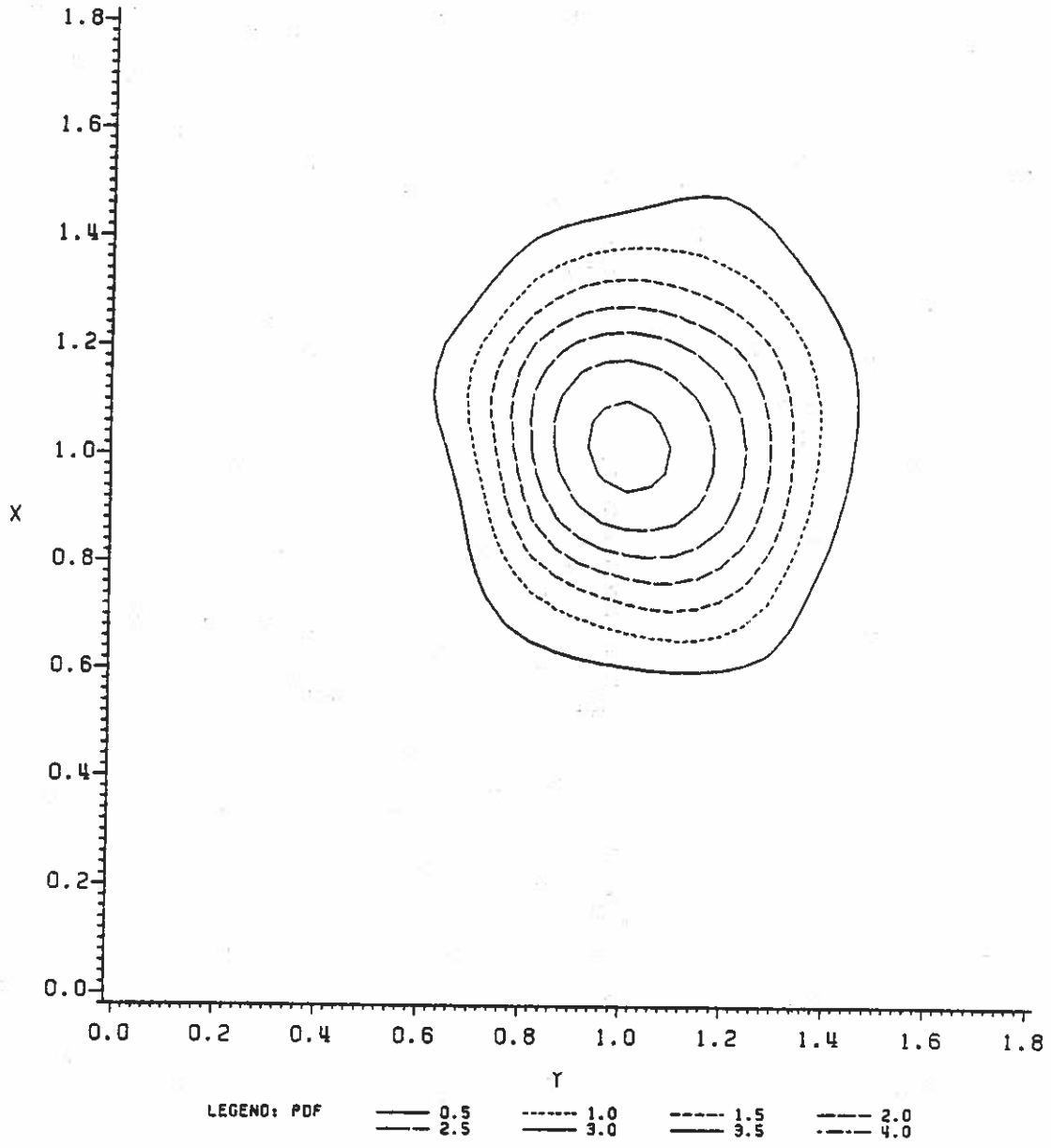

Figure 24. The contour plot of estimated independent bivariate normal density function. (means are 1.0, variances are 0.04 and correlation coefficient is 0)

also conducted. The multimodal peaks appeared in all cases, and the location of modes consistenly moves with time. Adjacent data sets,

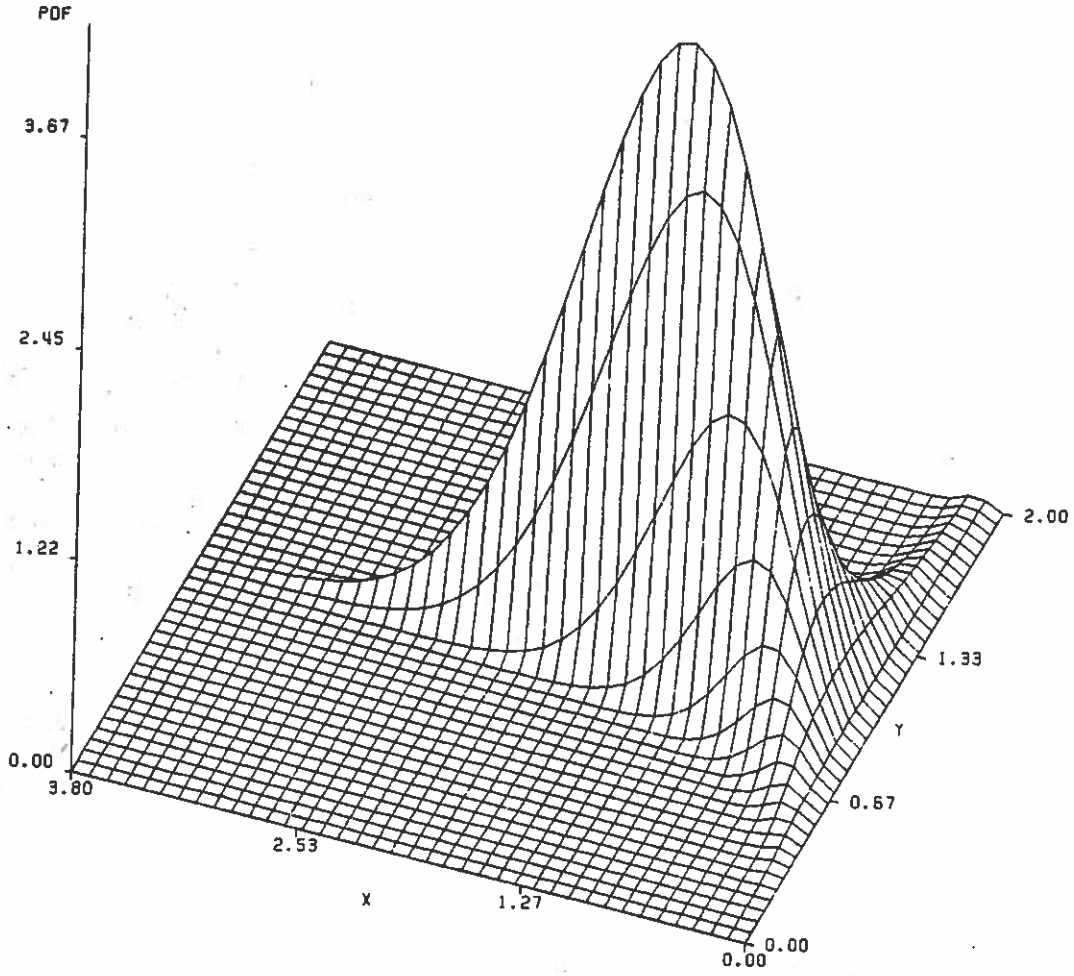

Figure 25. The true modified Longuet-Higgins density function. (the spectral width parameter is 0.2)

six hours apart, do not change the joint pdf dramatically for the March case, but the joint pdf changes significantly in some of the August data sequences. The reason for this may be due to noise from

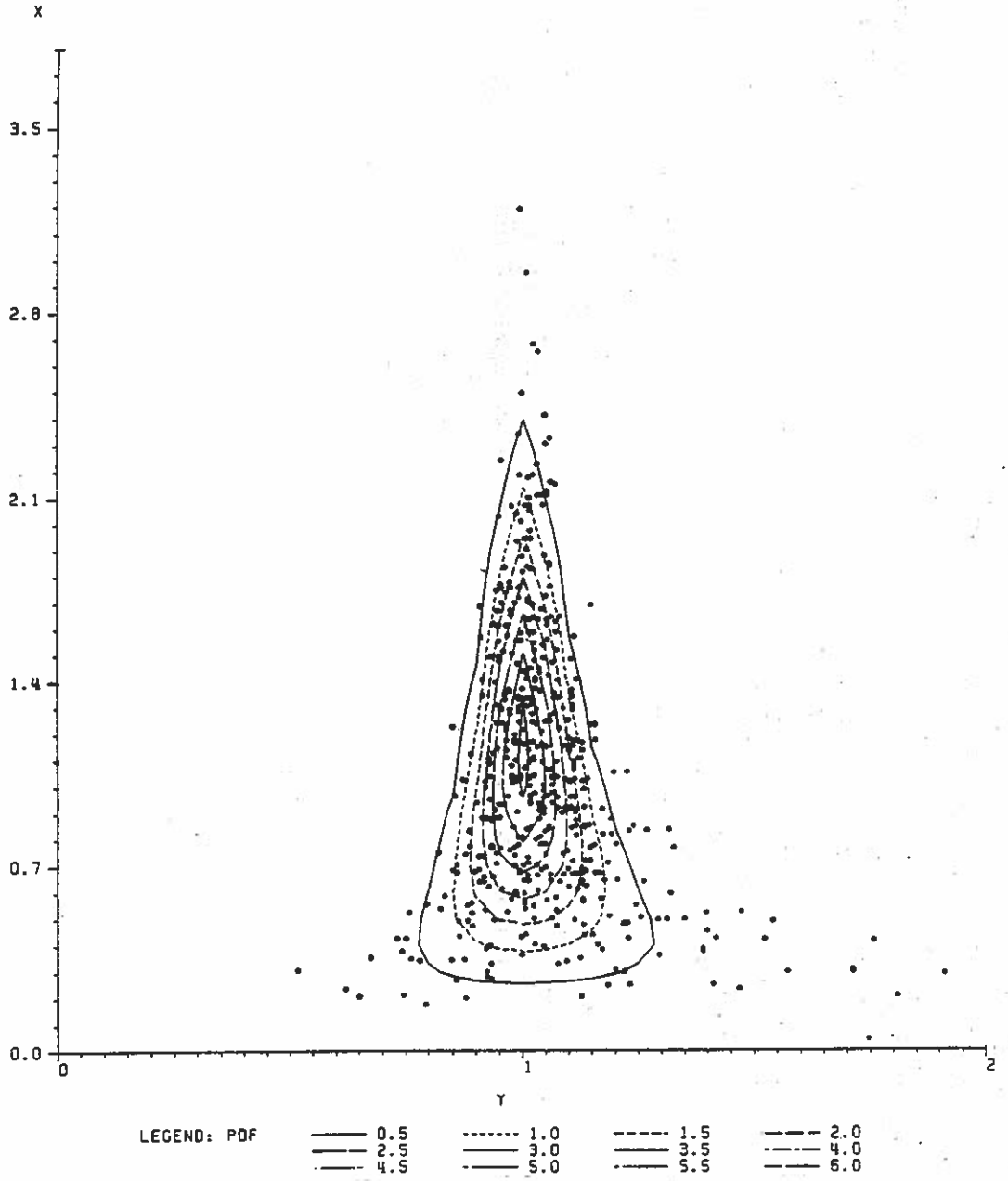

Figure 26. Scatter diagram of 500 samples generated from the modified Longuet-Higgins distribution. (Contour lines are the true probabilty density)

wave sampling. The noise would be more noticeable if the wave heights were not large enough to be measured by the device.

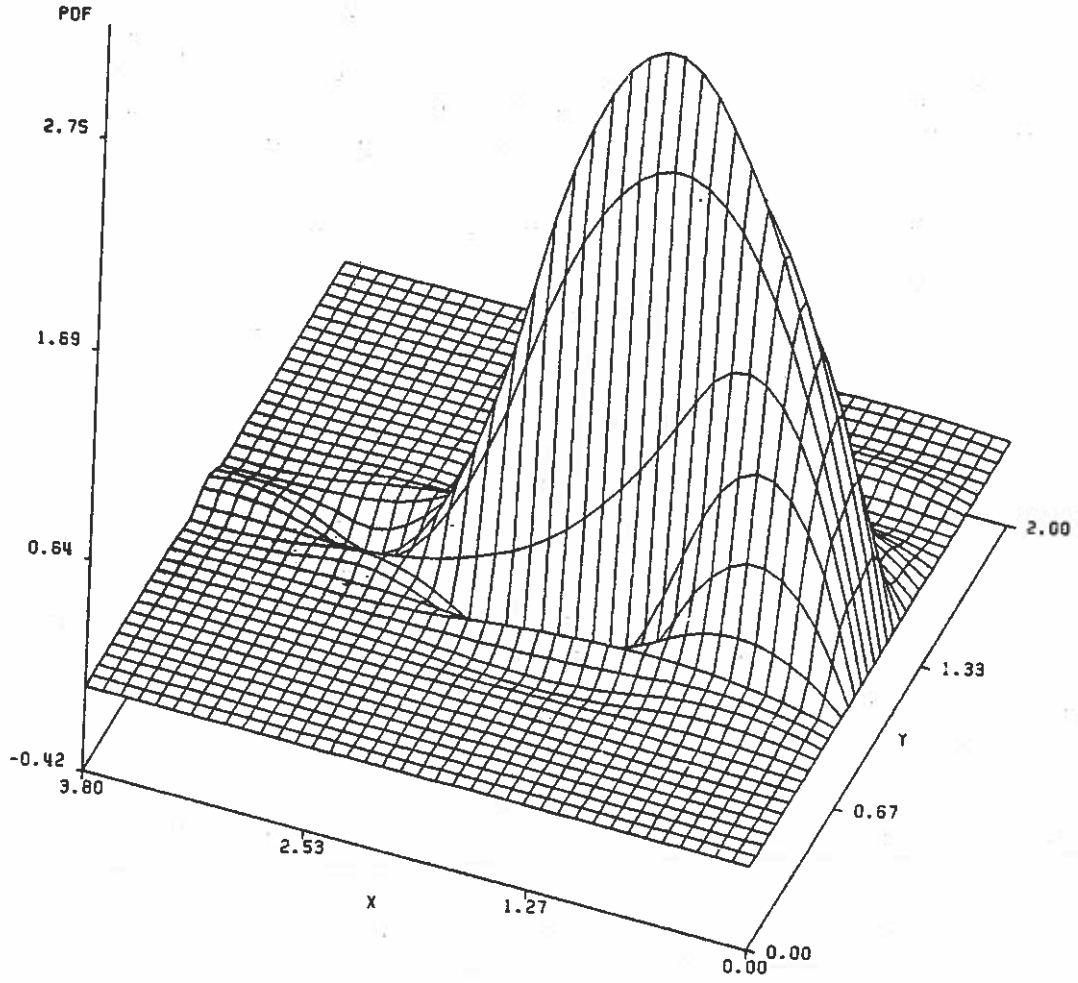

Figure 27. The estimate of modified Longuet-Higgins density function. (the spectral width parameter is 0.2)

The joint pdf which represents the long-term wave statistics may be obtained using a large sequence of data sets. Figures 36-41 show the estimated joint pdf based on 54 successive data sets, each of which contains approximately 300 waves. The estimated joint pdf of March

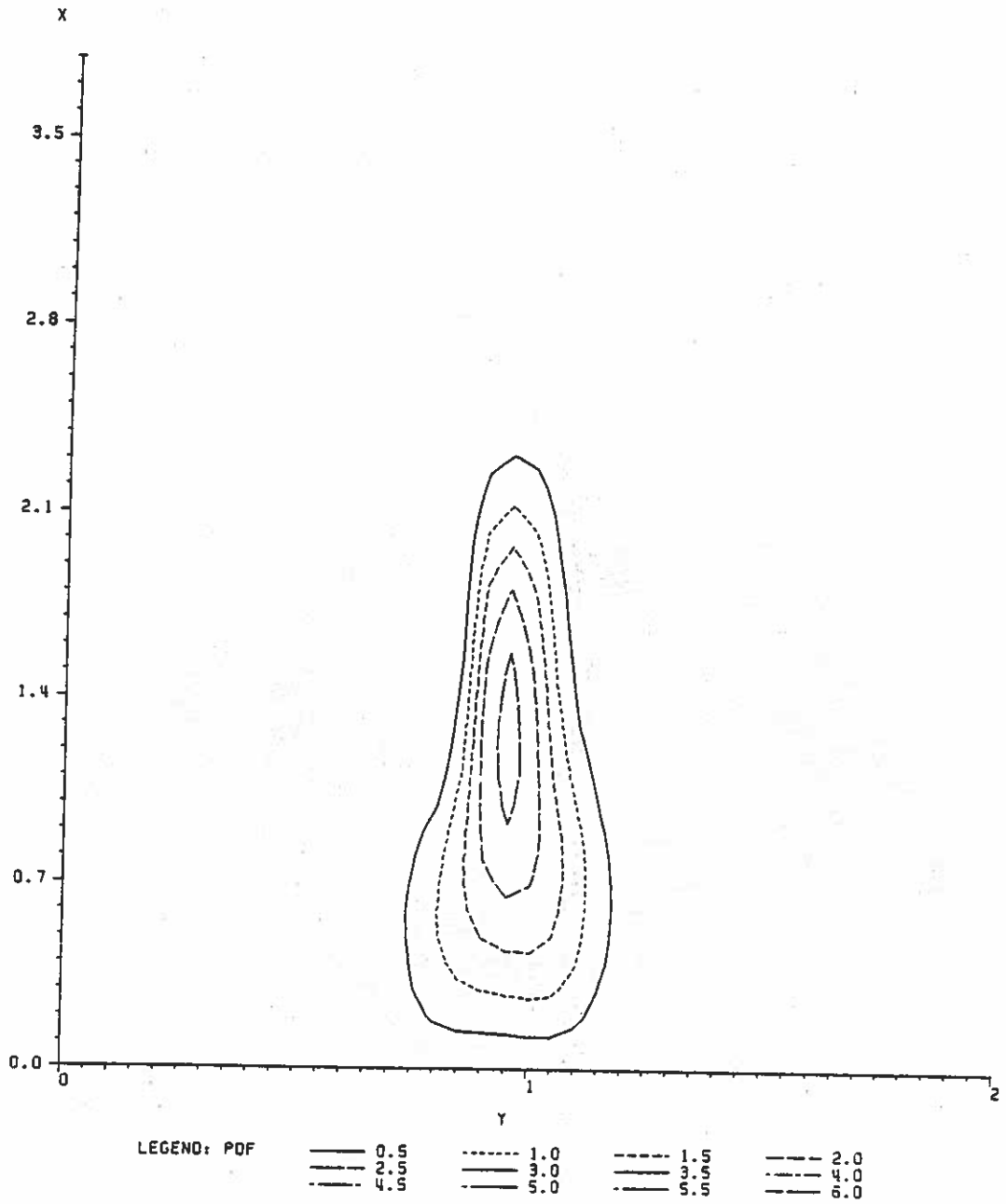

Figure 28. The contour plot of estimated density function. (the spectral width parameter is 0.2)

1980 and August 1981 was selected to characterize the wave statistics for each month. Note the single mode that appears in March and the double modes that appear in August. The larger peak in August appears

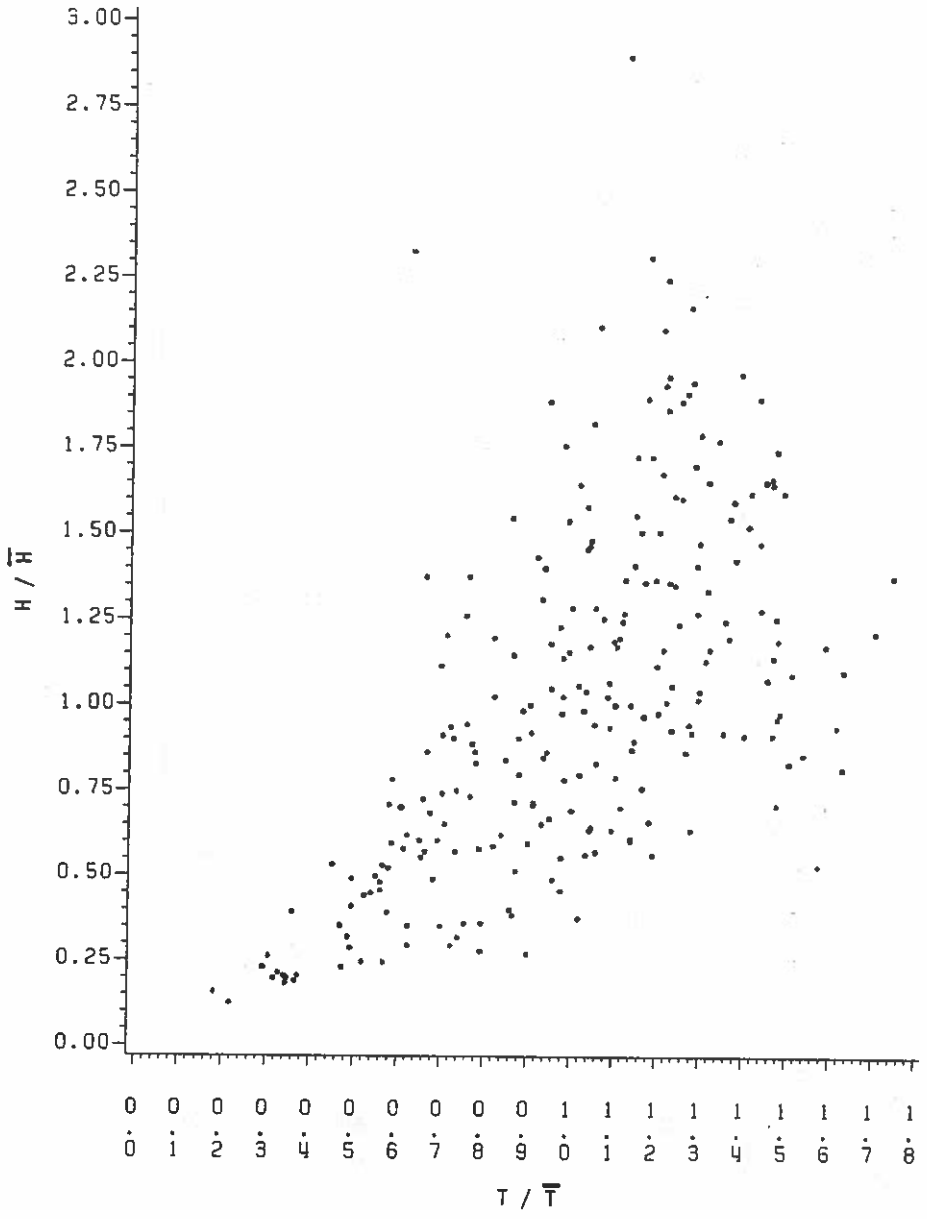

Figure 29. Scatter diagram of wave heights and periods samples. (Data obtained by Waverider buoy B at 3 p.m., March 13, 1980, near Port Mansfield, Texas)

at half the mean period. The second peak in August is located about the mean value of the period. It should be emphasized that the

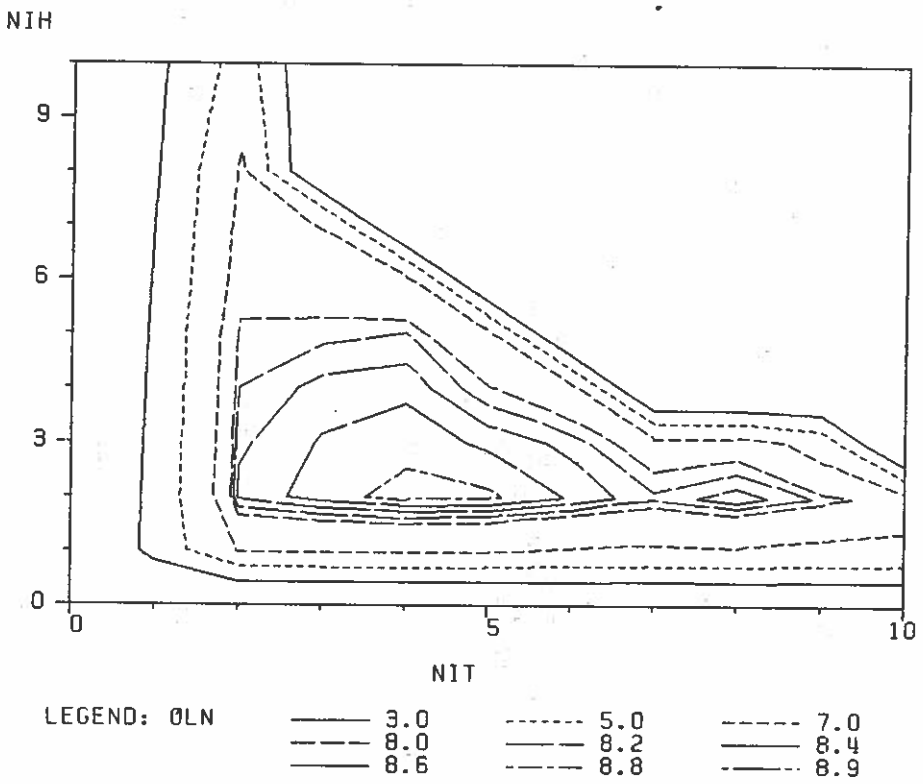

Figure 30. The contour map of -log(OLN) for sample wave heights and periods in Figure 29. (The peak occurs at NIH=2, NIT=4)

Longuet-Higgins distribution has a single mode which is located at the mean period; hence, it never shows the multimodal joint pdf features obtained by the proposed nonparametric density estimation technique.

Nonparametric density estimation is an excellent method for representing a large amount of long-term data in a concise form,

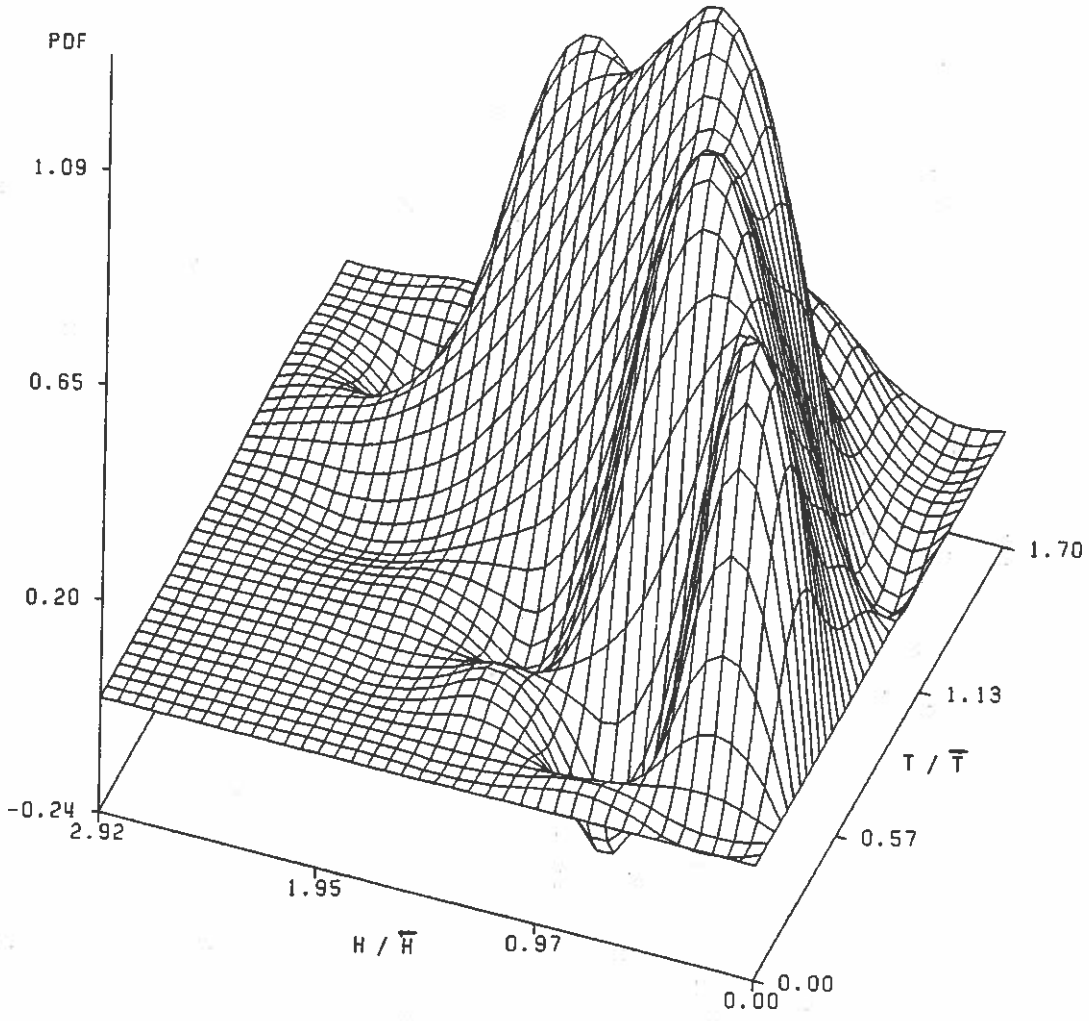

Figure 31. The best estimate of joint pdf for data obtained by Waverider buoy B at 3 p.m., March 13, 1980. (near Port Mansfield, Texas)

without masking the statistical characteristics of the data set. This was one of the major objectives of the study. The parametric density estimation may significantly alter the estimated density profile from the population density. Nevertheless, the direct practical application

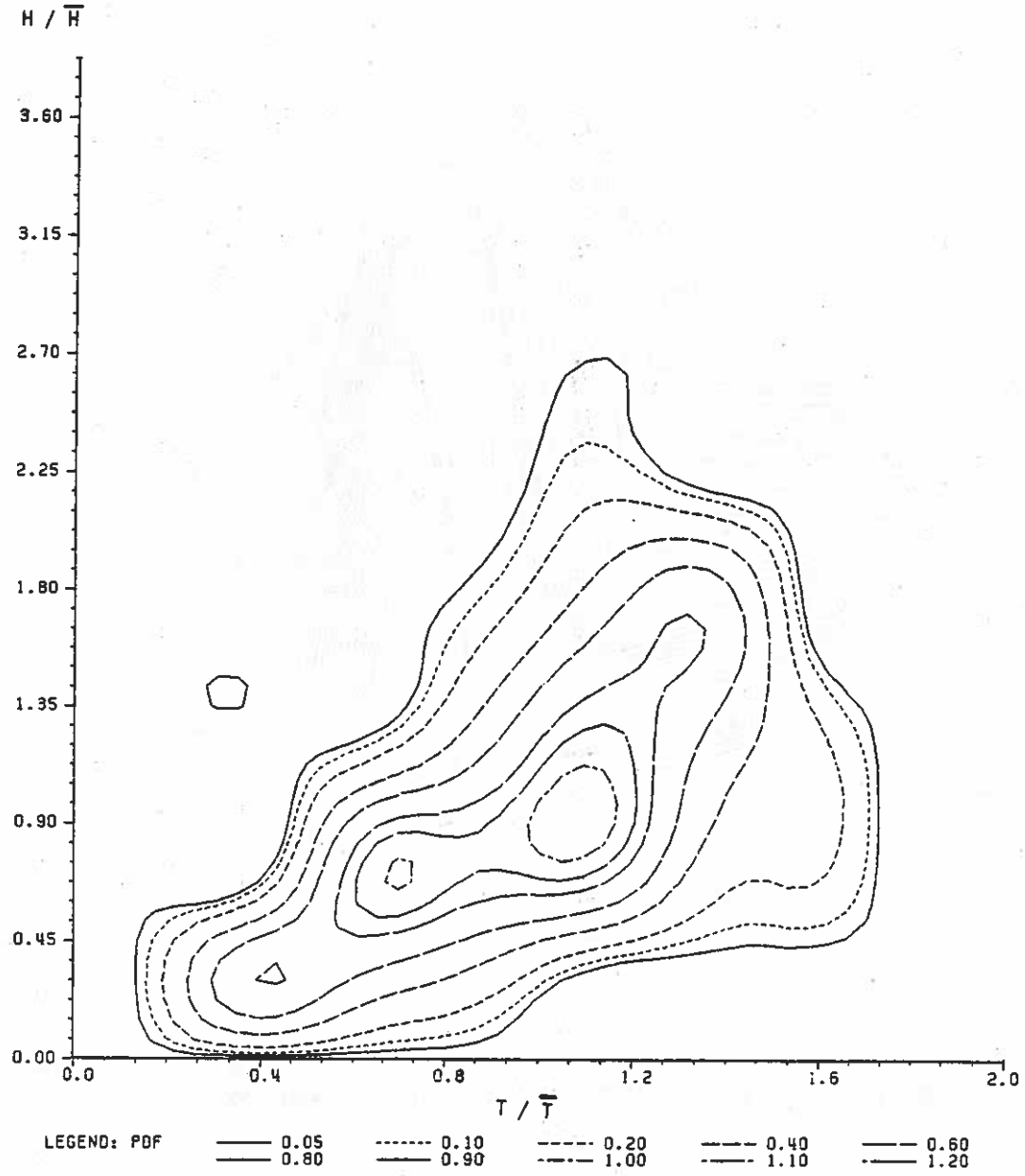

Figure 32. The contour map of best estimate of joint pdf for data obtained by Waverider buoy B at 3 p.m., March 13, 1980. (near Port Mansfield, Texas)

of the proposed technique may be difficult in the actual offshore

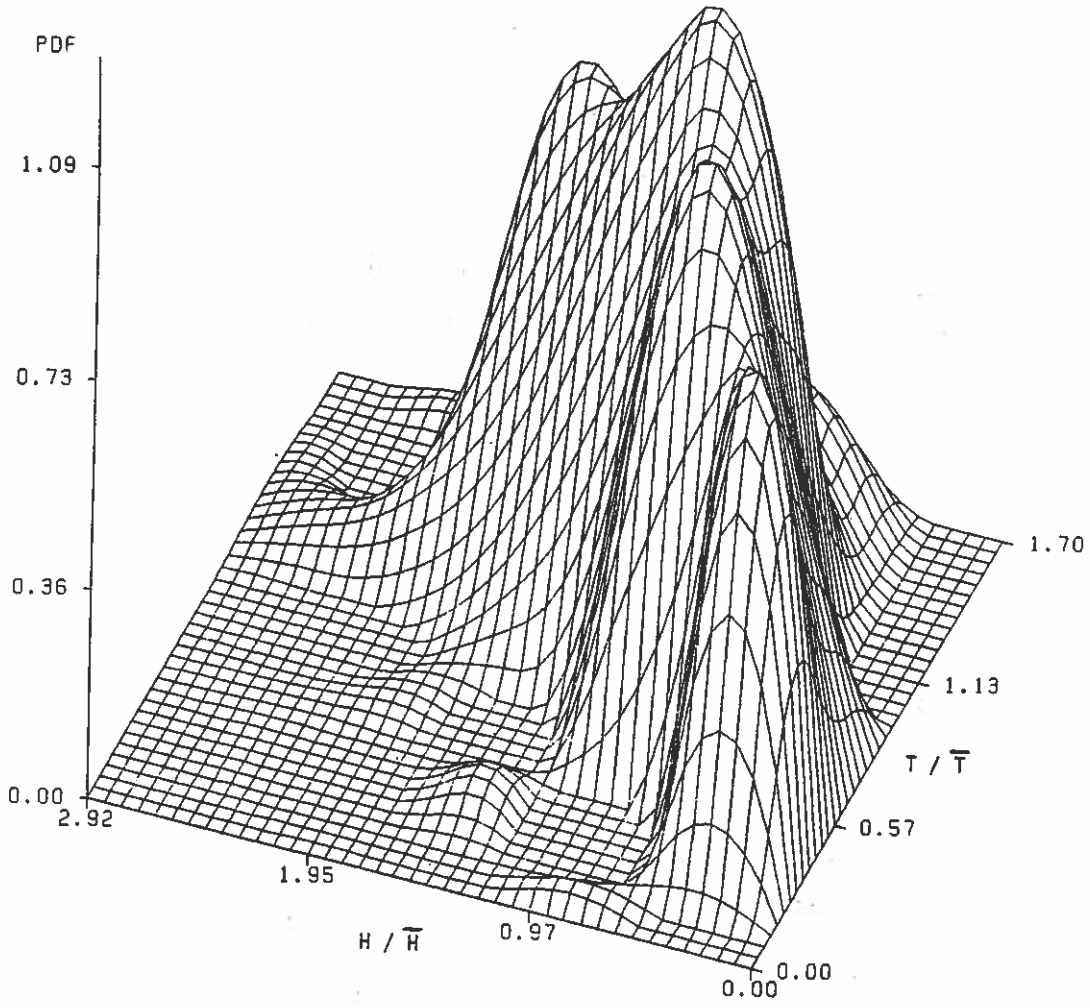

Figure 33. The best estimate of joint pdf. Negative values are replaced by zeros. (Data obtained by Waverider buoy B at 3 p.m., March 13, 1980, near Port Mansfield, Texas)

structural design. However, the visual impression of data representation can be achieved by the method presented. A question might be posed on how the joint pdf and spectra are related. Does the spectral information provide the joint pdf estimation? Is it possible

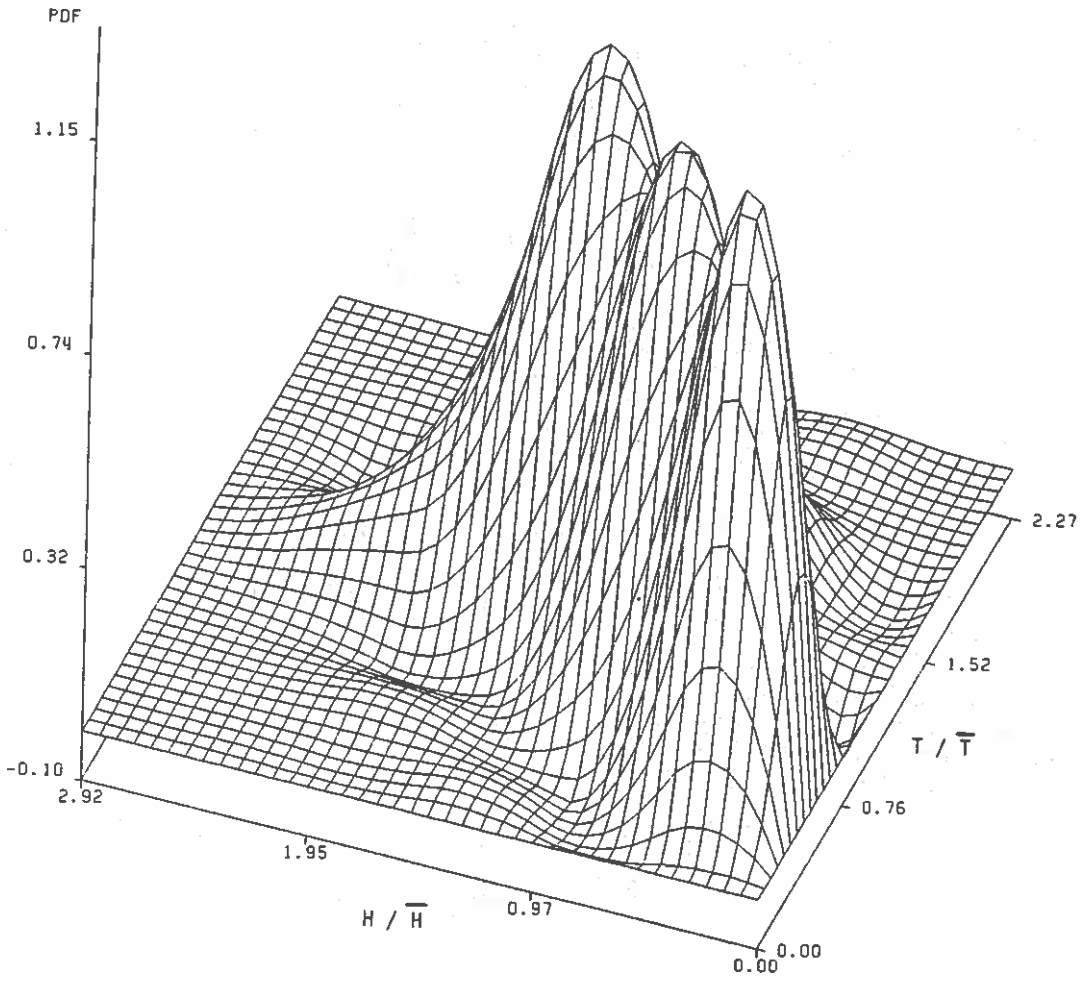

Figure 34. The best estimate of joint pdf for data obtained by Waverider buoy B at 3 a.m., August 13, 1981. (near Port Mansfield, Texas)

to estimate the spectra from the joint pdf information? The answer to the first question is "NO" unless the parametric density estimation is postulated. Longuet-Higgins (1975), Goda (1978), and Kimura (1981) have examined the correlation between the spectral parameters and the

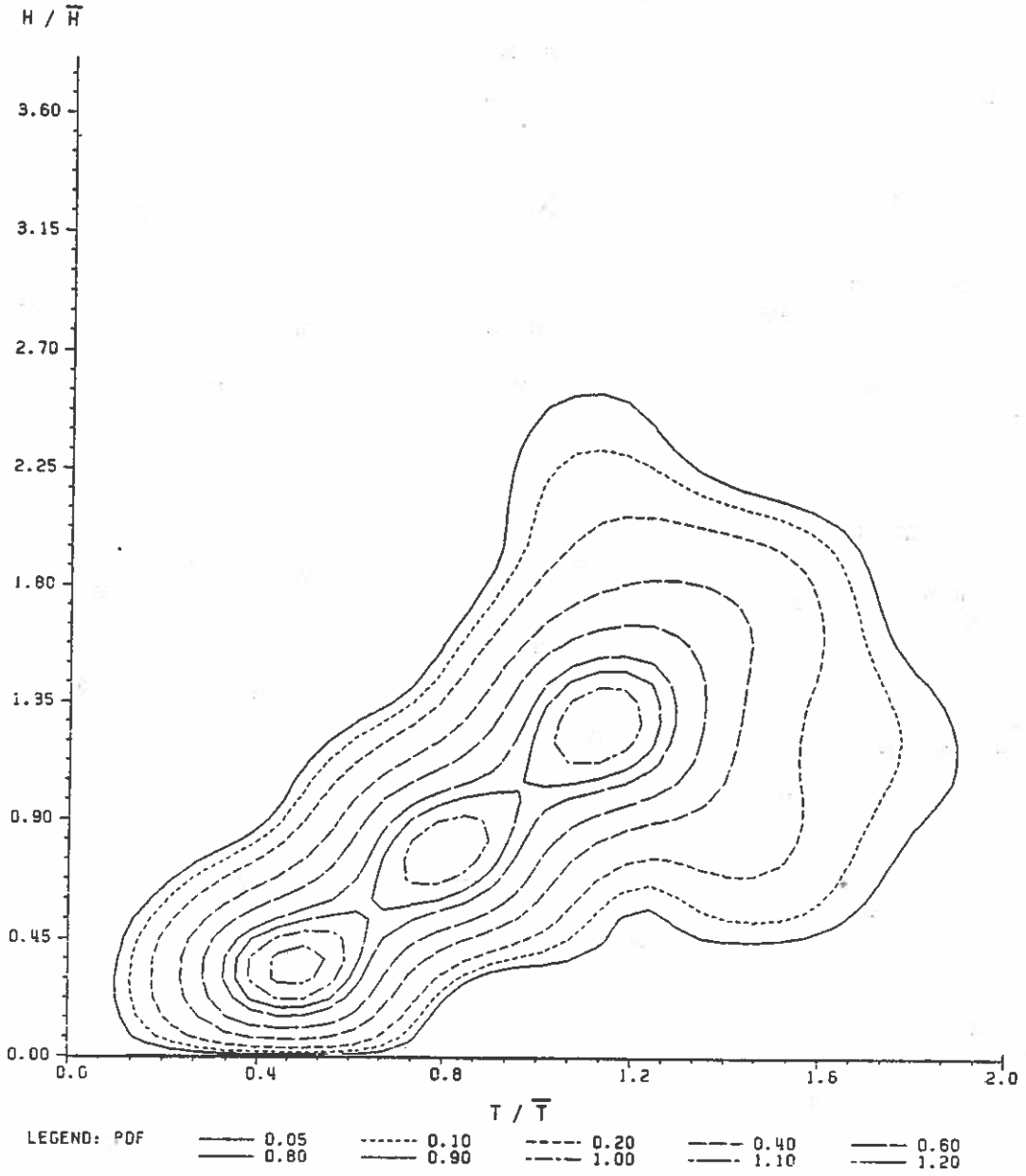

Figure 35. The contour map of best estimate of joint pdf for data obtained by Waverider buoy B at 3 a.m., August 13, 1981. (near Port Mansfield, Texas)

"parametric" joint pdf. The answer to the second question is "YES".

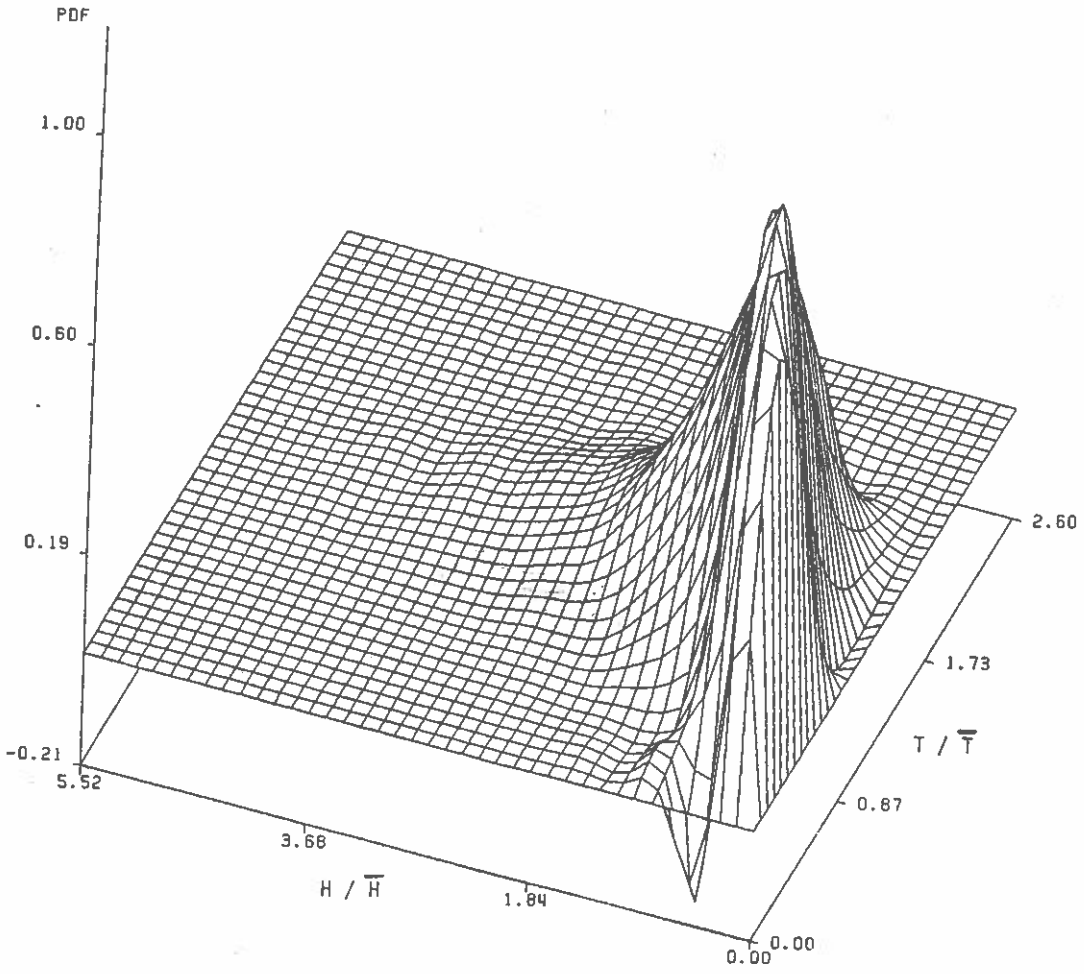

Figure 36. Monthly joint pdf of March 1980. (Data obtained by Waverider buoy B near Port Mansfield, Texas)

The author extended the idea of Bretschneider to estimate the spectra from the joint pdf. This will be discussed in a separate chapter.

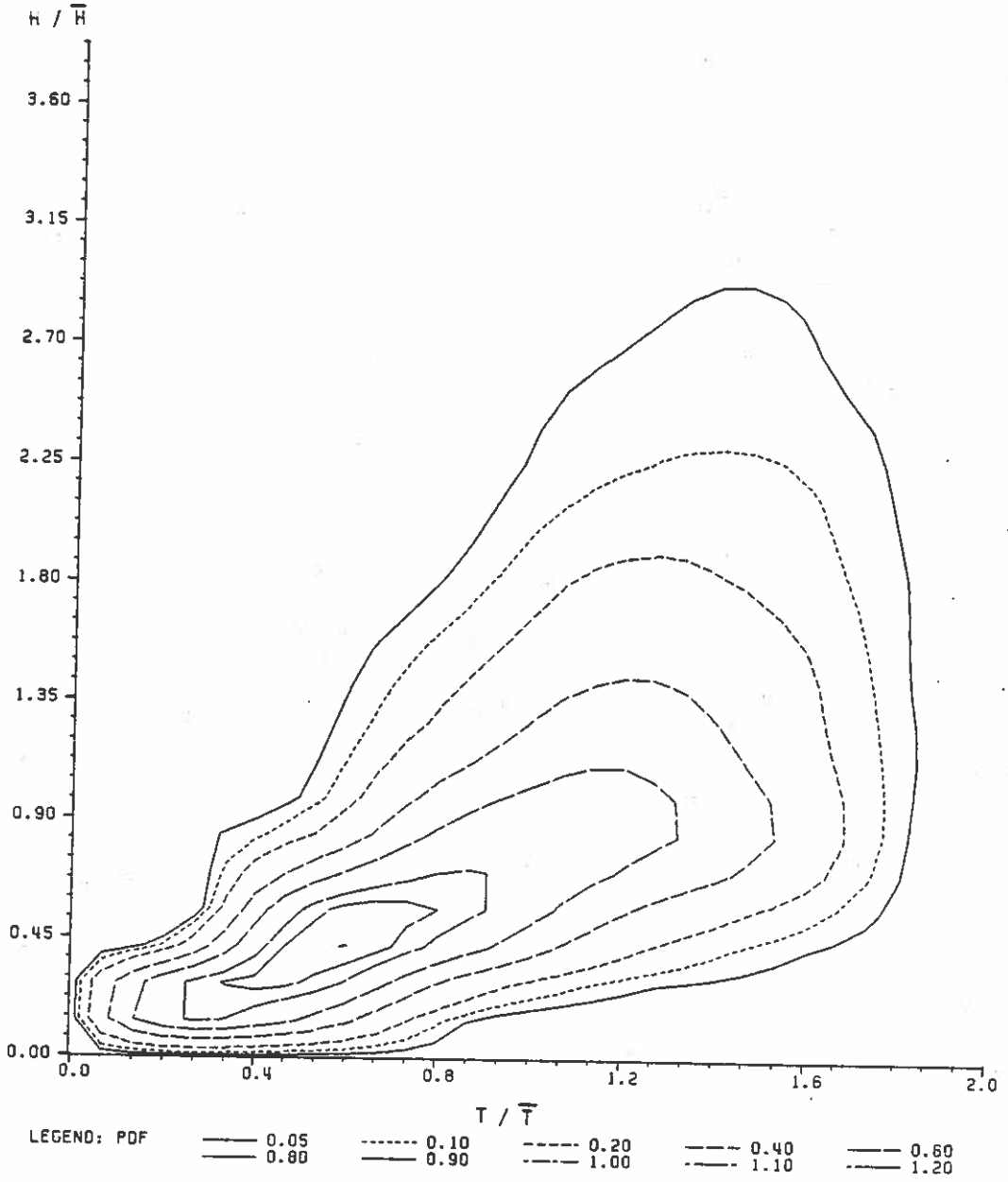

Figure 37. The contour map of monthly joint pdf of March 1980. (Data obtained by Waverider buoy B near Port Mansfield, Texas. Mode peak is at H=0.29 m, T=2.82 sec.)

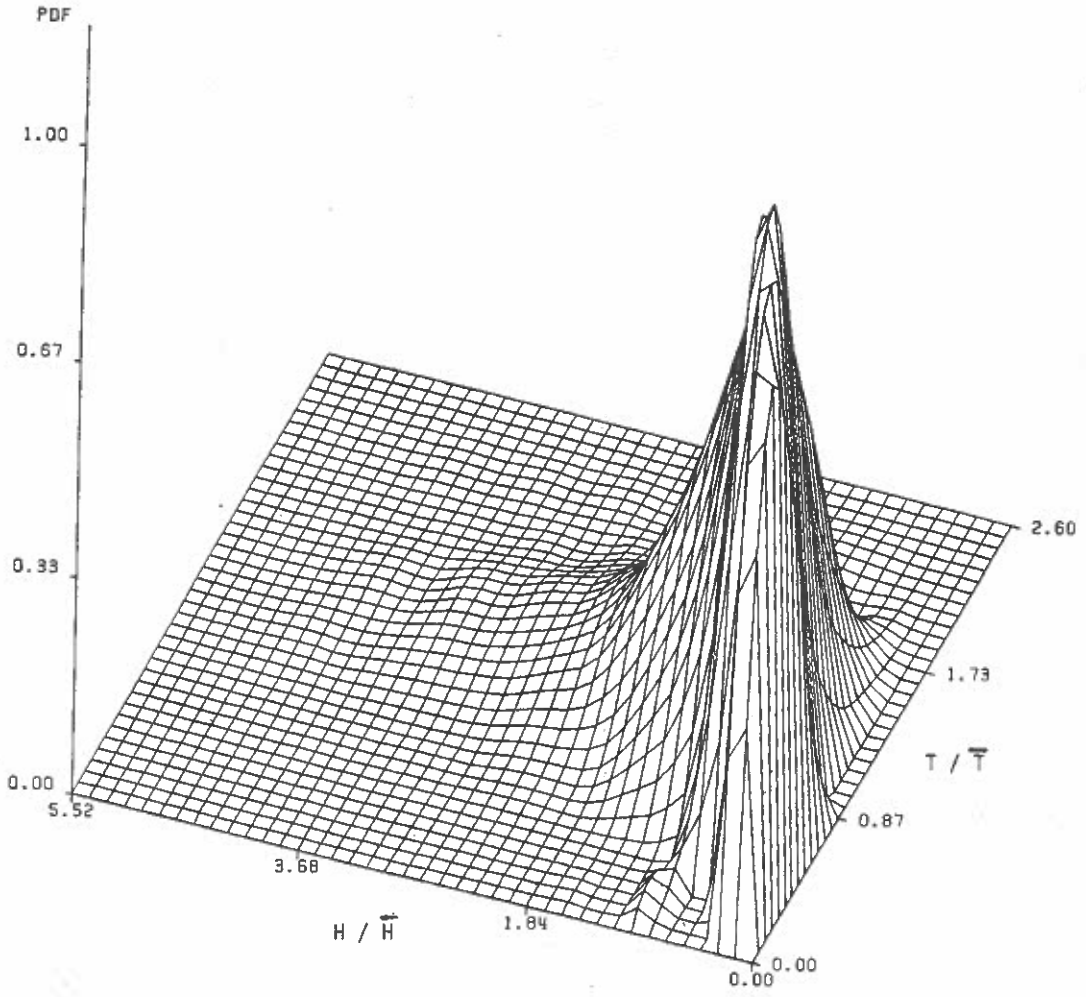

Figure 38. Monthly joint pdf of March 1980. Negative values are replaced by zeros. (Data obtained by Waverider buoy B near Port Mansfield,Texas)

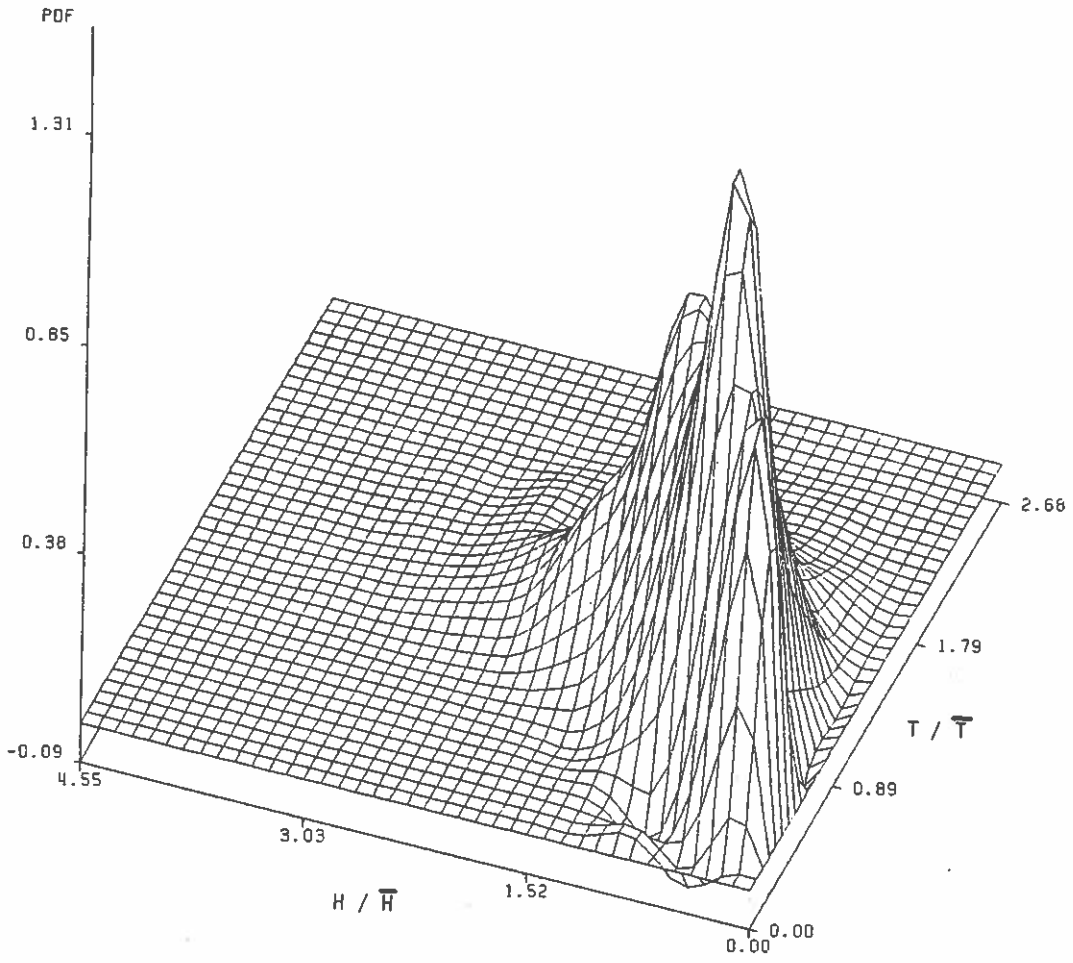

Figure 39. Monthly joint pdf of August 1981. (Data obtained by Waverider buoy B near Port Mansfield, Texas)

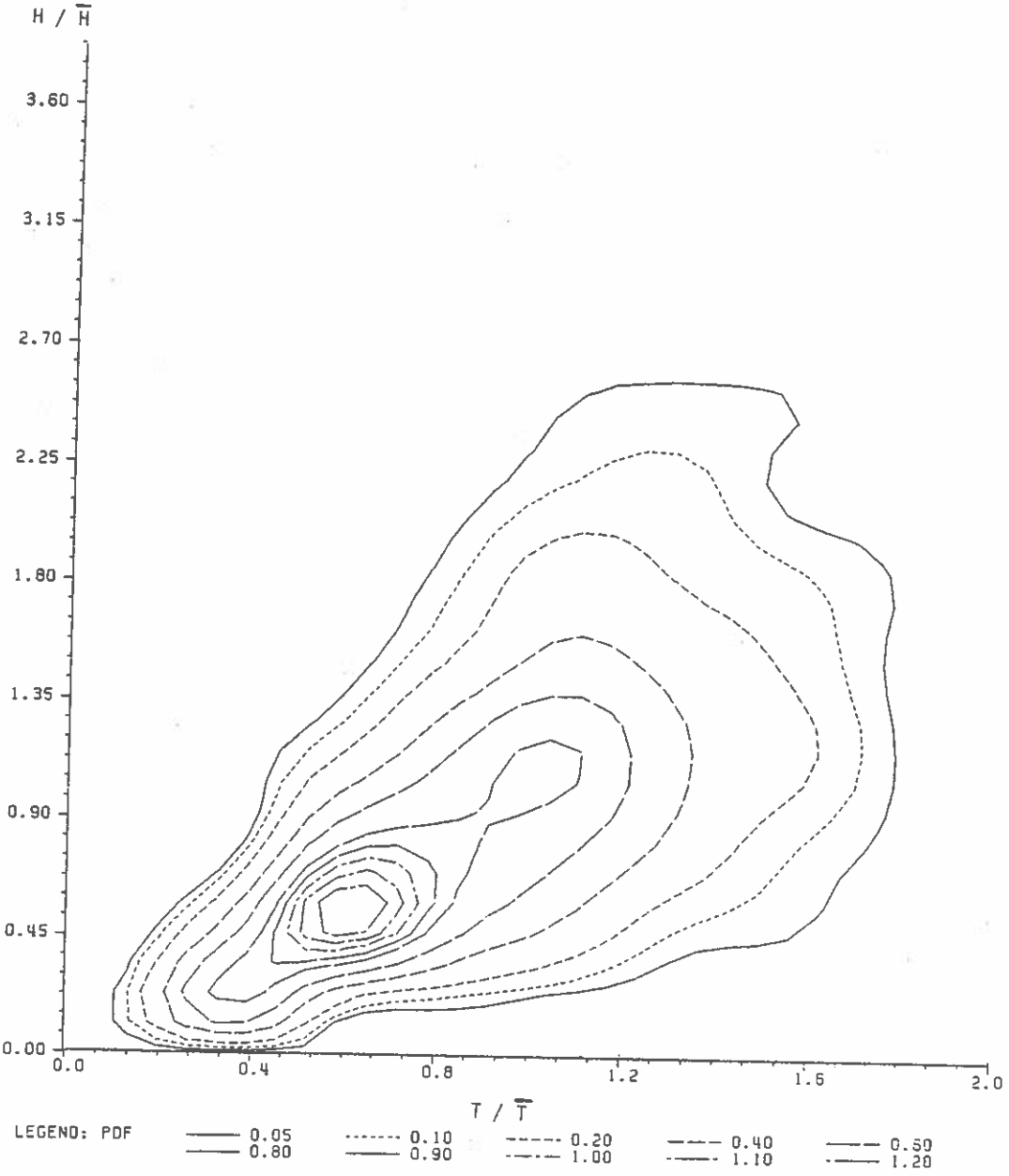

Figure 40. The contour map of monthly joint pdf of August 1981. (Data obtained by Waverider buoy B near Port Mansfield, Texas. Higher mode peak is at H=0.18 m, T=2.07 sec. Lower mode peak is at H=0.34 m, T=3.40 sec.)

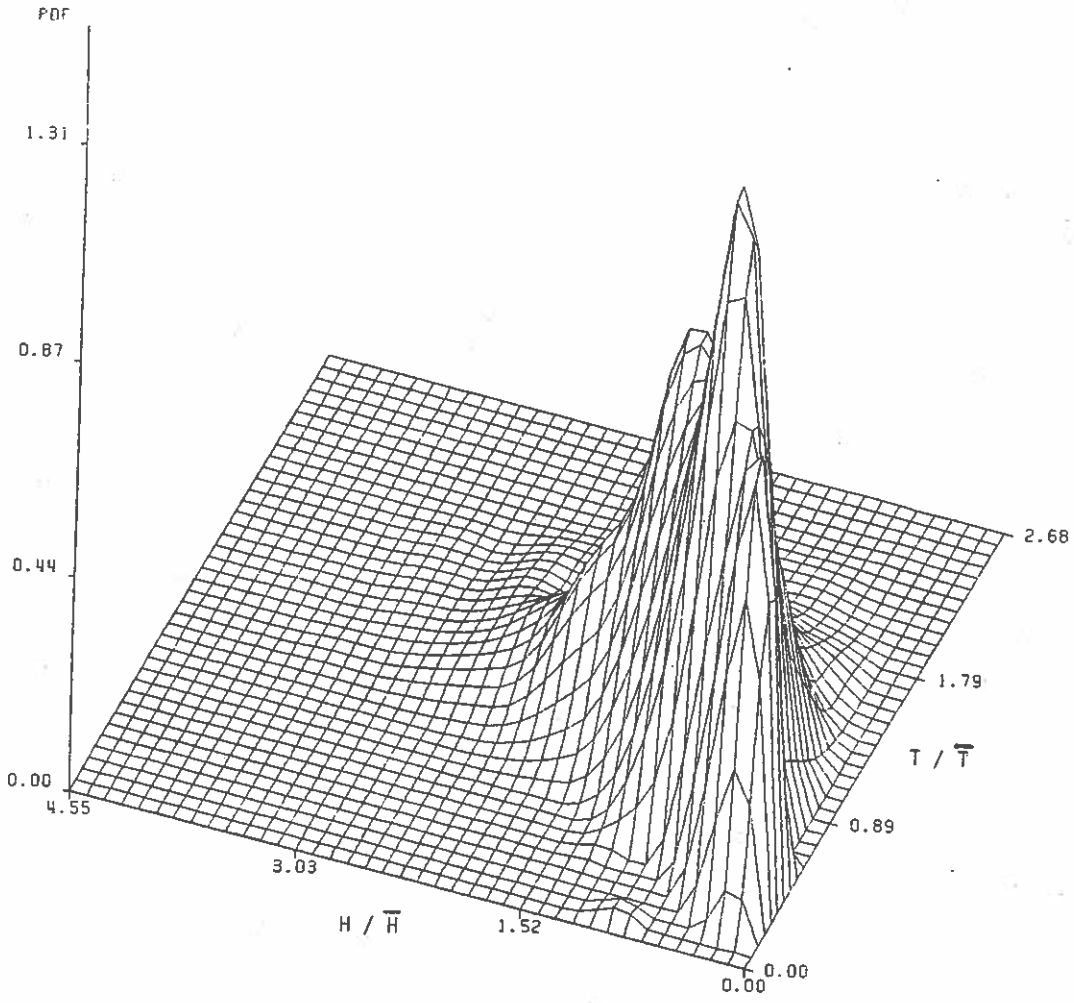

Figure 41. Monthly joint pdf of August 1981. Negative values are replaced by zeros. (Data obtained by Waverider buoy B near Port Mansfield, Texas)

EMPIRICAL WAVE SPECTRAL ESTIMATION

General

Spectral analysis has become a popular diagnostic method for the description of random phenomena, and reveals a detailed energy distribution as a function of frequency. It should be emphasized that spectral analysis is a "statistical" technique employed to investigate a "stochastic" process which is determined by a certain type of probabilistic law. Ocean waves are a random phenomena and exhibit very complex states. To investigate wave characteristics statistical approaches have to be employed. The statistical properties of ocean waves can be described in two ways: a) the wave statistics using the zero up-crossing method, and b) spectral analysis. The wave statistics are discussed in the previous chapter. This chapter will be devoted to introducing available spectral analysis techniques.

Extensive research on spectral analysis has been conducted in the last two decades, however, most of the theories were developed by statisticians, and involve the use of sophisticated statistical techniques. It is convenient to introduce modern empirical spectral analysis techniques without including the rigorous theoretical background. Spectral analysis may be classified into two categories: a) nonparametric, and b) parametric estimations. The well-known autocovariance method, commonly referred to as the Blackman-Tukey method (1958), is nonparametric in nature. Cooley and Tukey (1965)

introduced the Fast Fourier Transform (FFT) to reduce a large amount of computational effort of discrete Fourier transforms. The FFT has been extensively applied to the estimation of spectra. It can be shown that both methods; namely Blackman-Tukey method and FFT method, are theoretically equivalent. These techniques are called nonparametric spectral estimation. Parametric spectral estimation can be considered as a model identification of time series. Burg (1967) made an innovative study on spectral analysis employing the entropy concept. The method consists of essentially finding a linear whitening filter which maximizes the information entropy of data. Akaike (1969) and Parzen (1969) found the linear whitening filter of data employing a different approach. The linear whitening filter has the true spectrum, hence the spectral estimation can be made finding an adequate whitening filter, commonly referred to as Auto-Regressive (AR) spectral estimation.

The main objective of spectral analysis is to detect the best estimate of true spectrum from a sample observation. Chatfield (1980), Jenkins and Watts (1968) provide a detailed discussion. A more advanced discussion is in Priestley (1981). A concise summary of recent developments on time series analysis can be found in Newbold(1981).

A brief overview of the theoretical background of spectral analysis is discussed in section 2. Nonparametric estimation is presented in section 3. The AR estimator is outlined in section 4. A comparision of various estimators will be found in the next chapter.

A Brief Overview of Time Series Analysis

The theory of time series is discussed separately as discrete time series and continuous time series. Most data sampling is made in the form of digital reading with equal sampling time intervals. Since the computer input must be in a digital form, the continuous data in an analog form must be discretized before actual computations can take place. The discrete time series is of the main concern; namely

$$\{x(t),\ t=0,\pm1,\pm2,\ldots,\pm\infty\}$$

This is an infinite sequence of random variables. The mean of X(t) is denoted by

$$\mu(t) = E[x(t)] \tag{104}$$

The covariance, usually referred to as the autocovariance, is expressed as

$$\begin{aligned} R(t;\tau) &= \mathrm{cov}[x(t),x(t+\tau)] \\ &= E[(x(t) - \mu(t))(x(t+\tau) - \mu(t+\tau))] \qquad (105) \\ \tau &= 0,\pm1,\ldots,\pm\infty \end{aligned}$$

These parameters are all dependent on time (nonstationarity) and it is very difficult to deal with a nonstationary time series. It is extremely convenient to assume that the parameters are all independent of time and the means are all zero, namely:

$$\mu(t) = E[x(t)] = 0 \tag{106}$$

$$R(\tau) = E[x(t)x(t+\tau)] \tag{107}$$

The power spectral density function is a discrete Fourier transform of $R(\tau)$

$$S(f) = \sum_{\tau=-\infty}^{\infty} R(\tau)e^{-2\pi if\tau} \qquad (108)$$

where f is the frequency $-0.5 \leq f \leq 0.5$ and $i^2 = -1$. $f_N = 0.5$ is called the Nyquist frequency. Since $R(\tau) = R(-\tau)$, (108) can be expressed

$$S(f) = R(0) + 2\sum_{\tau=1}^{\infty} R(\tau)\cos 2\pi f\tau \qquad (109)$$

It is easy to show

$$\int_{-0.5}^{0.5} S(f)df = R(0) = E[x^2(t)] \qquad (110)$$

The total area underneath S(f), identically R(0), is the variance of the series, and is proportional to the mean energy of the process. The power spectra can also be defined in the following manner (Jenkins and Watts, 1968):

$$S(f) = \lim_{N\to\infty} \frac{1}{N} \left| \sum_{t=-N/2}^{N/2} x(t)e^{-2\pi ift} \right|^2 \qquad (111)$$

The power spectral density function is a counterpart of a pdf for ordinary statistics. As a counterpart of cdf, the power spectral distribution function $\Gamma(f)$ can be defined in the following form:

$$\Gamma(f) = \int_{-0.5}^{f} S(\theta)d\theta \qquad (112)$$

The autocovariance can be expressed in terms of $\Gamma(f)$.

$$R(\tau) = \int_{-0.5}^{0.5} e^{-2\pi if\tau} d\Gamma(f) \qquad (113)$$

Sometimes the autocorrelation function $\rho(\tau)$, which is defined as

$$\rho(\tau) = \frac{R(\tau)}{R(0)} \tag{114}$$

is used instead of $R(\tau)$. Then the corresponding spectral distribution function is upper bounded by 1.0. However, to avoid confusion, the power spectral density function in terms of $\rho(\tau)$ will not be further discussed. Technically there is a difference of the constant R(0) in practical applications.

It is apparent that $S(f)$ is an even function with respect to $f=0$ by inspecting equation (109). This is commonly called a two-sided spectrum. Since the negative frequency domain of $S(f)$ is completely the mirror image of the positive domain, it is not necessary to look at the negative frequency which has no physical meaning. Therefore, the power spectral density function is often represented by the one-sided spectrum

$$S_+(f) = 2S(f), \quad 0 \le f \le 0.5 \tag{115}$$

The linear filter theorem provides another way to define the power spectrum. A general class of linear processes is the AutoRegressive Moving Average process (ARMA). The ARMA process of order p and q is denoted by ARMA(p, q, $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$,σ^2),and defined

$$x(t) + \alpha(1)x(t-1) + \ldots + \alpha(p)x(t-p)$$

$$= \varepsilon(t) + \beta(1)\varepsilon(t-1)+\ldots+\beta(q)\varepsilon(t-q) \tag{116}$$

where $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$ are coefficient vectors

$$\underset{\sim}{\alpha} = [\alpha(1),\ldots,\alpha(p)]^T$$

$$\underset{\sim}{\beta} = [\beta(1),\ldots,\beta(q)]^T$$

$\varepsilon(t)$ = the independent identically distributed white noise having mean 0 and variance σ^2, denoted by WN($0,\sigma^2$)

It can be shown that the process X(t) has the true spectrum (see Jenkins and Watts 1968)

$$S(f) = \frac{\sigma^2 \left| 1 + \sum_{k=1}^{q} \beta(k) e^{-2\pi i f k} \right|^2}{\left| 1 + \sum_{j=1}^{P} \alpha(j) e^{-2\pi i f j} \right|^2}, \quad -0.5 \leq f \leq 0.5 \qquad (117)$$

Most time series can be expressed in terms of ARMA processes. Special classes of the ARMA process are the Auto-Regressive (AR) process of order P and the Moving Average process of order q which are denoted by AR($p,\underset{\sim}{\alpha},\sigma^2$) and MA($q, \underset{\sim}{\beta}, \sigma^2$). The MA process is always a stationary process. However, the AR process is only stationary when the roots of the polynomial

$$1 + \sum_{j=1}^{p} \alpha(j) Z^j = 0 \qquad (118)$$

lie outside the unit circle. Having assumed the stationarity of time series, the process may be expressed in terms of AR. Parzen (1982) assumes the stationary time series is AR(∞), the estimate of AR(∞) can be obtained approximating the sample process by a finite order of the AR equation. The spectrum of an AR(P) is

$$S(f) = \frac{\sigma^2}{\left| 1 + \sum_{j=1}^{P} \alpha(j) e^{-2\pi i f j} \right|^2}, \quad -0.5 \leq f \leq 0.5 \qquad (119)$$

Nonparametric spectral estimation is based on the definition of (109) or (111), and parametric spectral estimation is based on the definition (119).

Nonparametric Spectral Estimation

The theory of time series states that a process is an infinite sequence of random variables; however, practically it is only possible to observe a finite number of sample values, namely

$$\{x(t),\ t=1,2,\ldots,N\}$$

where N is taken to be an even integer for the sake of simplicity.

A natural estimator of the power spectra for a zero mean stationary process is

$$I(f) = \frac{1}{N}\left|\sum_{t=1}^{N} x(t)e^{-2\pi ift}\right|^2$$

$$= \frac{1}{N}\{(\sum_{t=1}^{N} x(t)\cos 2\pi ft)^2 + (\sum_{t=1}^{N} x(t)\sin 2\pi ft)^2\} \qquad (120)$$

which is usually called the periodogram.

The periodogram can be expressed in terms of a sample autocovariance function estimator

$$I(f) = \sum_{\tau=-(N-1)}^{N-1} \hat{R}(\tau)e^{-2\pi if\tau}$$

$$= \hat{R}(0) + 2\sum_{\tau=1}^{N-1} \hat{R}(\tau)\cos 2\pi f\tau \qquad (121)$$

where

$$\hat{R}(\tau) = \frac{1}{N}\sum_{t=1}^{N-\tau} x(t)x(t+\tau),\ \tau=0,1,\ldots,N-1 \qquad (122.1)$$

$$\hat{R}(\tau) = \hat{R}(-\tau) \qquad (122.2)$$

The estimator $\hat{R}(\tau)$ is an asymptotically unbiased estimator of $R(\tau)$. Another estimator of $R(\tau)$ which often appears in the literature is

$$\hat{R}'(\tau) = \frac{1}{N-\tau}\sum_{t=1}^{N-\tau} x(t)x(t+\tau),\ \tau=0,1,\ldots,N-1 \qquad (123.1)$$

$$\hat{R}'(\tau) = \hat{R}'(-\tau) \tag{123.2}$$

The estimator $\hat{R}(\tau)$ can be shown to be an unbiased estimator, however, $\hat{R}(\tau)$ has less mean square error (Jenkins and Watts, 1968). The covariance matrix (Toeplitz matrix) of $\hat{R}(\tau)$ is always positive definite which implies that the inverse of the covariance matrix always exists. This property becomes vital when autoregressive spectral estimation is considered. Hence, $\hat{R}(\tau)$ is recommended for the autocovariance estimator. It is a fact that $I(f)$ is a naive estimator that displays a very wiggly shape due to the large variance. The natural extension of the periodogram to overcome the shortcoming of large variance was proposed by Bartlett. He suggested subdividing a series into k pieces, then taking the average of each periodogram. Suppose $I_i(f)$ is a periodogram of an i-th subseries, say of length M, a smoothed spectral estimator $\hat{S}(f)$ may be obtained

$$\begin{aligned}\hat{S}(f) &= \frac{1}{k}\sum_{i=1}^{k} I_i(f) \\ &= \frac{1}{k}\sum_{i=1}^{k}\left(\hat{R}_i(0) + 2\sum_{\tau=1}^{M-1}\hat{R}_i(\tau)\cos 2 f\tau\right)\end{aligned} \tag{124.1}$$

where

$$\hat{R}_i(\tau) = \frac{1}{M}\sum_{t=1}^{M-\tau} x(M(i-1)+t)x(M(i-1)+t+\tau) \tag{124.2}$$

This is essentially identical to that performed for the whole series as follows:

$$\begin{aligned}\hat{S}(f) &= \sum_{\tau=-(N-1)}^{N-1} w_B(\tau)R(\tau)e^{-2\pi i f\tau} \\ &= \hat{R}(0) + 2\sum_{\tau=1}^{M} w_B(\tau)\hat{R}(\tau)\cos 2\pi f\tau\end{aligned} \tag{125}$$

where

$$w_B(\tau) = \begin{cases} 1 - \frac{\tau}{M} & |\tau| \le M \\ 0 & \text{otherwise} \end{cases}$$

The function $w_B(\tau)$ is called the Bartlett lag window. It should be noted how the lag window affects the estimated spectra. Mathematically, it is a weighted average of the periodogram. The variance of $I(f)$, $\mathrm{Var}[I(f)]$, is large and is independent of the number of samples N. On the other hand, the variance of $\hat{S}(f)$, $\mathrm{Var}[\hat{S}(f)]$, can be reduced by letting M be a small value. In other words, $\hat{S}(f)$ is more closely distributed around $E[\hat{S}(f)]$. However, the bias, $E[\hat{S}(f)]-S(f)$, becomes large as M decreases. Hence $\hat{S}(f)$ may differ from the true $S(f)$. In order to compromise the contradictory facts, it is necessary to choose a suitable truncation point M. The idea of the Bartlett lag window can be generalized by examining the properties of the lag window. The lag window can be characterized as

$$(1)\ w(0) = 1 \tag{126.1}$$

$$(2)\ w(\tau) = w(-\tau) \tag{126.2}$$

$$(3)\ w(\tau) = 0 \text{ for } \tau > M \tag{126.3}$$

As the power spectrum is the Fourier transform of the autocovariance function, the spectral window W(f) is the Fourier transform of the lag window

$$W(f) = \int_{-\infty}^{\infty} w(\tau) e^{-2\pi i f \tau} d\tau \tag{127.1}$$

also

$$w(\tau) = \int_{-\infty}^{\infty} W(f)e^{2\pi f\tau}df \qquad (127.2)$$

The spectral window satisfies the following conditions

$$(1) \int_{-\infty}^{\infty} W(f)df = 1 \qquad (128.1)$$

$$(2)\ W(f) = W(-f) \qquad (128.2)$$

The smooth spectral estimator can be expressed in terms of the spectral window.

$$\hat{S}(f) = \int_{-0.5}^{0.5} W(\lambda)I(f-\lambda)d\lambda \qquad (129)$$

Therefore $\hat{S}(f)$ is a convolution of the periodogram with respect to W(.). Since the spectral window is the Fourier transform of the lag window, the smaller M is, the wider the spectral window. Several lag windows have been proposed. A comparison of lag windows can be found in Neave (1972). Common types of lag windows and corresponding spectral windows are shown in Table 11. It can be seen that the rectangular spectral window has zeros at j/2M where j's are positive integers, and oscillate rapidly around zero. This is an undesirable feature since it may introduce a large negative peak. The Tukey window also exhibits a negative tail in the spectral window, but the negative contribution is considerably smaller than for the rectangular window. The Bartlett window is no longer used because its properties are inferior to the Tukey and Parzen windows (Chatfield, 1980). The Tukey and Parzen windows give very similar estimated spectra. The Parzen

window has a slight advantage of always being positive and therefore will not produce negative spurious estimates. Neave (1972) shows that the Parzen and Tukey windows are superior to most of the other proposed windows. However, the Parzen window is recommended in most cases.

A question of the optimal choice of a truncation point has not yet been solved in a satisfactory manner. Parzen (1964) suggested an evaluation of the effect of a number of different truncation points. Jenkins and Watts (1968) suggested trying three different values for truncation points and then observe the variation of estimated spectra. An empirical suggestion was made by Chatfield (1980) using a truncation point $M=2\sqrt{N}$ where N is the number of samples. Recently Wahba (1980) developed an objective optimum smoothing procedure for an estimate of the lag spectra based on smoothing the log periodogram with a smoothing spline. The method employed by her is rather complicated and has not become a popular algorithm in practical applications.

Another possible method to estimate spectra having less variance than the periodogram is the smoothing periodogram which simply groups the periodogram ordinate in sets of size m and calculates their average

$$\hat{S}(f) = \frac{1}{m} \sum_{j=1}^{m} I(f_j) \qquad (130)$$

Application of the FFT enables one to perform a quick calculation of the periodogram. The computational effort of spectral estimation can be reduced from the one based on the estimated autocovariance.

Table 11. Lag and Spectral Windows. (The table is reproduced from Jenkins and Watts(1968))

Description	Lag window	Spectral window
rectangular	$w_R(u) = \begin{cases} 1, & \lvert u\rvert \leqslant M \\ 0, & \lvert u\rvert > M \end{cases}$	$W_R(f) = 2M\left(\frac{\sin 2\pi fM}{2\pi fM}\right), \quad -\infty \leqslant f \leqslant \infty$
Bartlett	$w_B(u) = \begin{cases} 1 - \frac{\lvert u\rvert}{M}, & \lvert u\rvert \leqslant M \\ 0, & \lvert u\rvert > M \end{cases}$	$W_B(f) = M\left(\frac{\sin \pi fM}{\pi fM}\right)^2, \quad -\infty \leqslant f \leqslant \infty$
Tukey	$w_T(u) = \begin{cases} \frac{1}{2}\left(1 + \cos\frac{\pi u}{M}\right), & \lvert u\rvert \leqslant M \\ 0, & \lvert u\rvert > M \end{cases}$	$W_T(f) = M\left\{\frac{\sin 2\pi fM}{2\pi fM} + \frac{1}{2}\frac{\sin 2\pi M(f+\frac{1}{2}M)}{2\pi M(f+\frac{1}{2}M)} + \frac{1}{2}\frac{\sin 2\pi M(f-\frac{1}{2}M)}{2\pi M(f-\frac{1}{2}M)}\right\}$ $= M\left(\frac{\sin 2\pi fM}{2\pi fM}\right)\left(\frac{1}{1-(2fM)^2}\right), \quad -\infty \leqslant f \leqslant \infty$
Parzen	$w_P(u) = \begin{cases} 1 - 6\left(\frac{u}{M}\right)^2 + 6\left(\frac{\lvert u\rvert}{M}\right)^3, & \lvert u\rvert \leqslant \frac{M}{2} \\ 2\left(1 - \frac{\lvert u\rvert}{M}\right)^3, & \frac{M}{2} < \lvert u\rvert \leqslant M \\ 0, & \lvert u\rvert > M \end{cases}$	$W_P(f) = \frac{3}{4}M\left(\frac{\sin \pi fM/2}{\pi fM/2}\right)^4, \quad -\infty \leqslant f \leqslant \infty$

A measure of reliability for the spectral estimator can be expressed in terms of a confidence interval. Assuming the normal stationary process, meaning that each observation comes from a multivariate normal distribution, it can be shown that the periodogram has the form of the sum of the squared independent normal random variables. Hence the ratio $2I(f)/S(f)$ is distributed as a chi-square distribution with two degrees of freedom. The 95% confidence interval of a periodogram at an ordinate f is

$$\frac{2I(f)}{\chi^2_{2,0.025}} < S(f) < \frac{2I(f)}{\chi^2_{2,0.975}} \qquad (131)$$

where $\chi^2_{\alpha,\beta}$ is a chi-square value for the degree of freedom α at a level β significance. The above confidence interval is rather broad and is a reflection of a large variance of $I(f)$. A simple extension leads to the confidence interval of a smoothed spectral estimator. The 95% confidence interval can be expressed in the following form (Chatfield,1980):

$$\frac{\nu\hat{S}(f)}{\chi^2_{\nu,0.025}} < S(f) < \frac{\nu\hat{S}(f)}{\chi^2_{\nu,0.975}} \qquad (132)$$

where ν is a degree of freedom for the selected lag window. The values of ν for the Tukey and Parzen window are 2.67N/M and 3.71N/M respectively. The corresponding ν for the smoothed periodogram is 2m. The confidence interval around $\hat{S}(f)$ is usually much smaller than that of $I(f)$.

Parametric Spectral Estimation

The basic idea of the Auto-Regressive (AR) spectral estimation is simple. A mean zero stationary process $\{X(t)\}$ may be well approximated by a finite order P AR process

$$x(t) = -\alpha(1)x(t-1)-\alpha(2)x(t-2)-\ldots-\alpha(P)x(t-p)+\varepsilon(t) \qquad (133)$$

where $\varepsilon(t)$ is a white noise series with a mean of zero, and variance σ^2 . As stated earlier, an AR process has a true spectrum. Hence the methodology of the AR spectra is to estimate AR coefficients $\{\alpha(.)\}$ and the optimal order of the AR process P. Several methods have been proposed to estimate $\{\alpha(.)\}$. The Yule-Walker Estimate (YWE) has been extensively used for this purpose. Burg (1967) approached the spectral estimation from a different way commonly called the Maximum Entropy Method (MEM) of spectral analysis. The MEM spectral analysis is identical to fitting an AR model to the random process (Ulrych and Bishop, 1975). The only difference from the Yule-Walker Estimate is the method of estimating the AR coefficients. Newton (1983b) calls it the Burg algorithm rather than the MEM spectra. A comparison of YWE and MEM can be seen in Ulrych and Bishop (1975).

The order criteria of AR model plays another vital role of parametric spectral estimation. The first objective criteria was proposed by Akaike (1969) commonly called the Final Prediction Error (FPE). Akaike (1974) proposed another objective criteria based on maximum likelihood ideas, named the Akaike Information Criterion (AIC). Parzen (1977) approaches this problem in a slightly different

way. His criterion is named the CAT (Criterion for Autoregressive Transfer function).

Since a rigorous statistical investigation was not the purpose of this study, the following discussion will be made from a practical application point of view.

The Yule-Walker equations of AR($P,\underset{\sim}{\alpha},\sigma^2$) can be derived without much difficulty. Multiplying $X(t-\tau)$ on both sides of the AR equation and taking the expectation leads to

$$R(\tau) = -\alpha(1)R(\tau-1)-\ldots-\alpha(P)R(\tau-P) + \delta\tau\sigma^2 \qquad (134)$$

where $R(\tau) = E[x(t)x(t-\tau)]$

$$\delta t = \begin{cases} 1 & \tau=0 \\ 0 & \tau\neq 0 \end{cases}$$

Using the first P autocovariance estimators, $\hat{R}(\tau)$ in (122.1), the AR coefficients, $\{\hat{\alpha}\}$, may be estimated. The estimator $\{\hat{\alpha}\}$ can be obtained solving the following linear systems of equations:

$$\begin{bmatrix} \hat{R}(0) & \hat{R}(1) & \ldots & \hat{R}(P-1) \\ \hat{R}(1) & \hat{R}(2) & \ldots & \hat{R}(P-1) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \hat{R}(P-1) & \hat{R}(P-2) & \ldots & \hat{R}(0) \end{bmatrix} \begin{bmatrix} \hat{\alpha}(1) \\ \hat{\alpha}(2) \\ \cdot \\ \cdot \\ \cdot \\ \hat{\alpha}(P) \end{bmatrix} = \begin{bmatrix} \hat{R}(1) \\ \hat{R}(2) \\ \cdot \\ \cdot \\ \cdot \\ \hat{R}(P) \end{bmatrix} \qquad (135)$$

The $P \times P$ matrix on the left hand side is called a Toeplitz matrix, and is denoted by

$$\Gamma p = \text{TOEPL}(\hat{R}(0),\ldots,\hat{R}(p-1)) \tag{136}$$

since the first row contains the distinct elements in the matrix. The estimate of σ^2 is called the residual variance of order P, and is denoted by $\hat{\sigma}_p^2$

$$\hat{\sigma}_p^2 = R(0) + \hat{\alpha}(1)R(1)+\ldots+\hat{\alpha}(P)R(P) \tag{137}$$

Let $\{\hat{\alpha}_p\}$ be the AR coefficient estimate vector of order P, namely

$$\{\hat{\alpha}_p\} = [\hat{\alpha}_p(1),\hat{\alpha}_p(2),\ldots,\hat{\alpha}_p(P)]^T \tag{138}$$

Levinson (1947) provides a recursion relation between $\{\hat{\alpha}_p\}$ and $\{\hat{\alpha}_{p-1}\}$.

$$\hat{\alpha}_p(P) = -\frac{\left[\sum_{j=0}^{P-1} \hat{\alpha}_{p-1}(j)\hat{R}(P-j)\right]}{\hat{\sigma}_{p-1}^2} \tag{139.1}$$

$$\hat{\alpha}_p(j) = \hat{\alpha}_{p-1}(j) + \hat{\alpha}_p(P)\hat{\alpha}_{p-1}(P-j), j=1,\ldots,P-1 \tag{139.2}$$

$$\hat{\sigma}_p^2 = \hat{\sigma}_{p-1}^2[1 - \hat{\alpha}_p(P)]\ 1 - \hat{\alpha}_p{}^2(P) \tag{139.3}$$

and

$$\hat{\alpha}_1(1) = \frac{\hat{R}(1)}{\hat{R}(0)} \tag{139.4}$$

$$\hat{\sigma}_1^2 = \hat{R}(0)[1-\hat{\alpha}_1(1)]\ 1 - \hat{\alpha}_1{}^2(1) \tag{139.5}$$

Therefore $\{\hat{\alpha}_p\}$ can be quickly calculated without solving a matrix. The foregoing method is the YWE. It should be noted that the estimator $\hat{R}(\tau)$ considers values outside of the sample observations as being simply zero, i.e.:

$$\hat{R}(\tau) = \frac{1}{N}\sum_{t=1}^{N} y(t)y(t-\tau) \tag{140}$$

where $y(t)=0$ $\quad t \leq \tau$ $\quad$ and $\quad N<t+\tau$

$y(t) = x(t)$ $\quad$ otherwise

Newton (1983b) addresses this effect as a possible cause for the poor estimate of the AR coefficient in some circumstances, especially when the roots of polynomial

$$1 + \hat{\alpha}_p(1)Z + \ldots + \hat{\alpha}_p(P)Z^P = 0 \tag{141}$$

appear around the unit circle, the YWE's display significantly poor estimates. A numerical study of several AR coefficient estimation techniques may be found in Newton (1983b).

A careful inspection of Levinson's algorithm leads to the following: all $\hat{\alpha}_p(j)$ and $\hat{\alpha}_p^2$ can be obtained from $\hat{R}(0)$ and $\hat{\alpha}_p(P)$, $P = 1, \ldots, M$, where M is a maximum order of AR. $\hat{R}(0)$ is simply

$$\hat{R}(0) = \frac{1}{N} \sum_{t=1}^{N} x^2(t) \tag{142}$$

The Burg algorithm employs a different approach to estimate $\alpha_p(P)$. The estimator $\hat{\alpha}_p(P)$ is obtained by minimizing the backward and forward prediction error of AR(P). It is denoted as E_p , given by

$$E_p = \frac{1}{2} \frac{1}{N-P} \sum_{t=1}^{N-P} \Big[\{x(t) + \sum_{k=1}^{P} \alpha_p(k)x(t+k)\}^2 + \{x(t+p) + \sum_{k=1}^{P} \alpha_p(k)x(t+p-k)\}^2 \Big] \tag{143}$$

It can be shown that E_p achieves a minimum value with respect to $\alpha_p(P)$ taking

$$\frac{\partial E_p}{\partial \alpha_p(P)} = 0$$

Recalling the recursive relation of $\alpha_p(P)$, E_p can be expressed in the following form:

$$E_p = \frac{1}{2}\frac{1}{N-P} \sum_{t=1}^{P} [(B_{p-1}(t) + \alpha_p(P)F_{p-1}(t))^2 + (F_{p-1}(t)+\alpha_p(P)B_{p-1}(t))^2]$$

where

$$B_{p-1}(t) = \sum_{k=0}^{p-1} \alpha_{p-1}(k)x(t+k), \quad \alpha_{p-1}(0) = 1$$

$$F_{p-1}(t) = \sum_{k=0}^{p-1} \alpha_{p-1}(k)x(t+p-k), \quad \alpha_{p-1}(0) = 1$$

Hence the estimate of $\alpha_p(P)$ can be expressed as

$$\hat{\alpha}_p(P) = \frac{-2\sum_{t=1}^{N-P} B_{p-1}(t)F_{p-1}(t)}{\sum_{t=1}^{N-P} F_{p-1}(t) + \sum_{t=1}^{N-P} B_{p-1}(t)} \qquad (144)$$

The right hand side of (144) contains only $\{\hat{\alpha}_{p-1}(.)\}$, therefore $\hat{\alpha}_p(P)$ can be estimated recursively using Levinson's algorithm in the following manner:

1) $P=1$ $\quad B_o(t) = x(t)$

$$F_o(t) = x(t+1)$$

$$\hat{\alpha}_1(1) = \frac{-2\sum_{t=1}^{N-1} x(t)x(t+1)}{\sum_{t=1}^{N-1} x^2(t+1) + \sum_{t=1}^{N=1} x^2(t)}$$

2) $P=2$ $\quad B_1(t) = x(t) + \alpha_1(1)x(t+1)$

$$F_1(t) = x(t+2) + \alpha_1(1)x(t+1)$$

$$\hat{\alpha}_2(2) = \frac{-2\sum_{t=1}^{N-2} B_1(t)F_1(t)}{\sum_{t=1}^{N-2} F_1^2(t) + \sum_{t=1}^{N-2} B_1^2(t)}$$

$$\hat{\alpha}_2(1) = \hat{\alpha}_1(1) + \hat{\alpha}_2(2)\hat{\alpha}_1(1)$$

$$\hat{\sigma}_2^2 = \hat{\sigma}_1^2 [1-\hat{\alpha}_2^2(2)]$$

Therefore $\{\hat{\alpha}_p(1),\ldots,\hat{\alpha}_p(P)\}$ can be obtained using equations (144) and (139.2). Equation (139.3) gives the P-th order residual estimate $\hat{\sigma}_P^2$.

It is necessary to determine the optimal order of the AR model which adequately approximates the stationary time series $\{X(t)\}$. As stated earlier, three criteria are commonly employed for this purpose. Akaike's FPE, which is a similar expression of an unbiased residual variance etimator of AR order p for the stationary series of length N, is expressed as

$$FPE(P) = \frac{N+P+1}{N-P+1}\hat{\sigma}_P^2 \qquad (145)$$

and the order is determined as the value of p minimizing FPE(p). Akaike (1974) proposed an alternative order criterion, AIC, which has been extensively applied to various statistical problems. The order p is determined so as to minimize AIC(p) where:

$$AIC(P) = \ell n\,\hat{\sigma}_P^2 + \frac{2P}{N} \text{ for } P \geq 1 \qquad (146)$$

Parzen (1977) introduced another order of determination criterion, CAT for p>1 which is defined as

$$CAT(P) = \frac{1}{N}\sum_{j=1}^{P}(1 - \frac{j}{N})\hat{\sigma}_j^{-2} - (1 - \frac{P}{N})\hat{\sigma}_P^{-2} \qquad (147)$$

A comparison of these criteria is shown in Beamish and Priestley (1981). The AIC and CAT are more frequently adopted for the order criteria. It should be mentioned that both criteria provide almost always the same order for long time series.

The order determination of the AR spectral estimation is a counterpart of the choice of optimal truncation points for the window spectral estimator. The window estimator has a theoretical justification on the confidence interval; on the other hand, the AR spectral estimator lacked a way to estimate the confidence interval until recently. Newton and Pagano (1984) made a great breakthrough on this limitation. The basic idea is that the inverse of the AR spectra with a known order is a linear combination of asymptotically distributed normal random variables. A detailed discussion can be found in Newton and Pagano (1984).

PDF SPECTRAL ESTIMATION

General

The physical interpretation of power spectra is that the ordinate of a given frequency is approximately proportional to the energy of the frequency. The same type of function may be derived from a different point of view. Accordingly, an alternative method of spectral estimation is developed by means of the joint pdf of wave heights and periods. Thus this method is named the PDF spectra. The original idea was proposed by Bretschneider (1959) who introduced the summation function of the joint pdf.

The PDF spectra will be derived in the next section, and a comparison of the PDF spectra to selected empirical spectral estimators is shown in a subsequent section.

PDF Spectra

The nonlinear effect of surface elevation has been discussed in a series of studies by Longuet-Higgins (1963), Tayfun (1980), and Huang (1983). However, the nonlinear effect does not contribute significantly to the surface elevation process. The following discussion assumes that the surface elevation is a linear process, and statistics of the wave height follows the Rayleigh distribution. Using statistics of the Rayleigh distribution, the second moment of wave

height H and the variance of surface elevation η have the following relation;

$$\overline{\eta^2} = \frac{1}{8}\overline{H^2} \tag{148}$$

Define the second moment distribution function of H as follows:

$$\Gamma_H(T) = \int_0^T \int_0^\infty \xi^2 \, P_{HT}(\xi,\zeta)d\xi d\zeta \tag{149}$$

$$\text{and} \quad \overline{H^2} = \Gamma_H(\infty)$$

This function is the counterpart of the power spectral distribution function $\Gamma_\eta(f)$.

$$\Gamma_\eta(f) = \int_0^f S_\eta(t)\, dt \tag{150}$$

where $S_\eta(f)$ is the one-sided power spectra, i.e.:

$$S_\eta(f) = 2\int_{-\infty}^{\infty} R(\tau)e^{-2\pi i f\tau}d\tau \tag{151}$$

$$R(\tau) = \text{cov}[\eta(t),\eta(t+\tau)] \tag{152}$$

$$\text{and} \quad \overline{\eta^2}=\Gamma_\eta(\infty)$$

Another way of obtaining $S_\eta(f)$ is by differentiating $\Gamma_\eta(f)$ with respect to f, so that

$$S_\eta(f) = \frac{d\Gamma_\eta(f)}{df} \tag{153}$$

An analogy to the relationship above is the density function $S_H(\xi)$ which can be defined as

$$S_H(\xi) = \frac{d\Gamma_H(\xi)}{d\xi} = \int_0^\infty \xi^2 P_{HT}(\xi,\zeta)d\xi \tag{154}$$

The PDF spectra is defined by transforming $S_H(\xi)$ into the frequency domain such that

$$S_H(f) = \int_0^\infty (\frac{\xi}{f})^2 \, P_{HT}(\xi, \frac{1}{f}) d\xi \qquad (155)$$

Recalling equation (148), $S_\eta(f)$ and $S_H(f)$ have the following relation:

$$\Gamma_\eta(\infty) = \frac{1}{8} \Gamma_H(\infty)$$

Therefore

$$\int_0^\infty S_\eta(f)df = \frac{1}{8} \int_0^\infty S_H(f)df \qquad (156)$$

Assume that the following relation is true for any constant f_o

$$\int_0^{f_o} S_\eta(f)df = \frac{1}{8} \int_0^{f_o} S_H(f)df \qquad (157)$$

Then $S_\eta(f)$ and $S_H(f)$ differ only by a factor of 1/8,i.e.:

$$S_\eta(f) = \frac{1}{8} S_H(f) \qquad (158)$$

The PDF spectra require the estimation of the joint pdf $P_{HT}(\xi,\zeta)$, and therefore the nonparametric density estimation becomes vital. The joint pdf $P_{HT}(\xi,\zeta)$ is estimated employing the technique by means of B-splines discussed in a previous chapter.

The normalized wave height and period are

$$X = \frac{H}{\overline{H}}$$

$$Y = \frac{T}{\overline{T}}$$

The truncated joint cdf of X and Y is estimated by

$$\hat{P}_{XY}(x,y) = \sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} C_{ij} B_i(x) B_j(y) \qquad (159)$$

which has a finite support $[0, a\,]x[0, b\,]$. The joint pdf is estimated by differentiating (159) with respect to x and y, so that

$$\hat{P}_{XY}(x,y) = \sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} C_{ij} D_x B_i(X) D_Y B_j(y) \qquad (160)$$

where D_X, D_Y are differential operators with respect to x and y respectively. Making use of the probability transformation, $S_H(f)$ can be estimated as

$$\hat{S}_H(f) = \frac{f^2 \bar{H}^2}{\bar{T}} \sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} C_{ij} g_i D_Y B_j(y) \qquad (161)$$

where

$\bar{H}$, $\bar{T}$ are the mean of H and T

$f = 1/y\bar{T}$

$$g_i = \int_o^a x^2 D_X B_i(x)dx$$

Comparison of Spectral Estimators

In total, 50 wave data sets from a Waverider buoy placed near Port Mansfield during March 1980 and August 1981 were selected to examine several spectral estimators. The statistical summary of the data sets is shown in Tables 12 and 13.

There are several software packages for the general purpose of spectral analysis. SAS and IMSL have been widely used in the various problems. FESTSA of Brocks (1976) is designed for the purpose of processing geophysical data. It performs the window spectral

Table 12. Statistical summary of wave data from Waverider buoy B in March 1980. (data obtained near Port Mansfield, Texas)

NO	Data ID	$\overline{\eta^2}$	Skewness	Kurtosis	Number of Waves	Correlation Coefficient	$\overline{H^2}$
1	MR13P3	0.2526	-0.0652	2.8882	240	0.6219	1.9017
2	MR13P9	0.1274	-0.0201	2.8332	236	0.6855	0.9287
3	MR14A9	0.0565	0.0283	2.8870	258	0.6405	0.4300
4	MR15A3	0.1648	0.0509	3.1016	219	0.6870	1.2108
5	MR15P3	0.1326	0.0214	2.9019	216	0.6544	1.0052
6	MR16A3	0.1111	-0.0429	2.9314	238	0.7235	0.8157
7	MR16A9	0.1085	-0.0664	2.8120	246	0.7146	0.8147
8	MR16P9	0.1104	-0.0602	2.7200	230	0.7420	0.8333
9	MR17A9	0.1173	0.0172	2.9042	206	0.6455	0.8742
10	MR18A9	0.2128	-0.0324	3.1813	217	0.6866	1.5710
11	MR18P3	0.1626	-0.0559	2.7874	216	0.6586	1.2110
12	MR19A3	0.1407	0.0933	3.0375	200	0.7146	1.0311
13	MR19A9	0.1113	0.0213	3.0723	227	0.7188	0.8040
14	MR20A3	0.0568	0.0306	2.9440	235	0.6946	0.4206
15	MR20P9	0.0544	0.0548	2.8526	224	0.6645	0.4134
16	MR21A3	0.0816	-0.0271	3.0935	311	0.6591	0.5813
17	MR22A3	0.1206	-0.0237	3.1151	219	0.6690	0.8820
18	MR22A9	0.0818	-0.0605	2.9563	247	0.7428	0.5977
19	MR22P3	0.0588	-0.0931	2.9511	284	0.6991	0.4365
20	MR23A9	0.1137	0.0247	2.9711	261	0.6922	0.8203
21	MR24A3	0.1112	-0.0374	2.7320	201	0.6337	0.8553
22	MR24P9	0.1161	-0.0611	2.8709	238	0.6792	0.8674
23	MR25A3	0.0682	-0.0586	3.0317	252	0.6858	0.4877
24	MR26A9	0.1635	0.0137	2.8571	234	0.6519	1.2434
25	MR26P9	0.3598	0.0055	2.8811	193	0.6602	2.7236

estimation using both the Blackman-Tukey method and the FFT method. TIMESBOARD of Newton (1983a) is a FORTRAN package of 250 subroutines and 10 main programs. It is designed for general applications. ARSPID is the main program of TIMESBOARD which performs both the window spectral estimation and the autoregressive spectral estimation. A comparison of spectral estimation was made using ARSPID for 50 data sets. The following were the selected methods used in this study from available spectral estimation algorithms:

Table 13. Statistical summary of wave data from Waverider buoy B in August 1981. (data obtained near Port Mansfield, Texas)

NO	Data ID	$\overline{\eta^2}$	Skewness	Kurtosis	Number of Waves	Correlation Coefficient	$\overline{H^2}$
1	AG11P9	0.0191	-0.0880	3.0105	412	0.5657	0.1402
2	AG12A9	0.0206	-0.0261	3.0658	340	0.6870	0.1503
3	AG12P3	0.0127	0.0110	3.0178	343	0.6792	0.0918
4	AG13P9	0.0414	-0.0244	2.7689	327	0.6707	0.3025
5	AG14P9	0.0229	-0.0736	3.0123	334	0.7702	0.1590
6	AG15A3	0.0264	-0.0105	3.1803	340	0.7276	0.1847
7	AG15P3	0.0134	-0.0186	3.0895	362	0.7455	0.0875
8	AG15P9	0.0212	0.0015	2.9379	379	0.5711	0.1527
9	AG16A9	0.0104	0.0014	2.9133	343	0.6344	0.0971
10	AG17A3	0.0165	-0.1173	3.2099	426	0.5937	0.1160
11	AG17P9	0.0184	0.0191	2.8562	407	0.5633	0.1332
12	AG18P3	0.0111	-0.1638	3.1294	401	0.6770	0.0716
13	AG18P9	0.0174	-0.0779	2.8806	400	0.5203	0.1262
14	AG19A3	0.0140	-0.0393	3.1357	404	0.5496	0.1003
15	AG19P3	0.0045	-0.1433	2.7910	464	0.5209	0.0292
16	AG19P9	0.0124	-0.1388	3.0168	440	0.4957	0.0886
17	AG20A3	0.0095	0.0142	2.9012	362	0.5031	0.0728
18	AG20P3	0.0106	-0.0376	3.0065	450	0.5359	0.0749
19	AG21A9	0.0234	-0.0264	3.0849	310	0.5012	0.1823
20	AG21P9	0.0203	-0.0266	3.1772	383	0.5862	0.1460
21	AG22A9	0.0139	-0.0424	3.0153	368	0.5642	0.1013
22	AG23A3	0.0260	-0.0438	3.1142	327	0.6168	0.1929
23	AG24A9	0.0115	-0.0123	2.9510	328	0.6219	0.0860
24	AG24P3	0.0148	-0.0246	2.9340	367	0.6386	0.1063
25	AG24P9	0.0227	-0.0539	3.0351	299	0.6757	0.1619

1) FFT method: A periodogram is estimated by a FFT algorithm, then the Parzen window with a truncation point 80 is used to obtain the smoothed spectra.

2) AR method: The Yule-Walker algorithm is employed to estimate the autoregressive coefficients. The model identification is made by Parzen's CAT.

Table 14. Spectral analysis summary of wave data from Waverider buoy B in March 1980. (data obtained near Port Mansfield, Texas)

				AR		FFT		PDF			
NO	ID	AIC	CAT	$S(f_p)$	f_p	$S(f_p)$	f_p	$S(f_p)$	f_p	Ω_{AF}	Ω_{PF}
1	MR13P3	34	34	3.962	0.161	3.100	0.169	3.049	0.169	0.953	1.000
2	MR13P9	31	31	1.826	0.127	1.274	0.152	1.219	0.143	0.836	0.941
3	MR14A9	14	14	0.645	0.161	0.506	0.161	0.604	0.180	1.000	1.118
4	MR15A3	41	41	3.397	0.135	2.353	0.135	2.047	0.150	1.000	1.111
5	MR15P3	17	17	2.106	0.127	1.486	0.135	2.002	0.150	0.941	1.111
6	MR16A3	33	33	1.350	0.161	1.074	0.152	1.228	0.161	1.059	1.059
7	MR16A9	38	38	1.879	0.152	1.158	0.152	1.531	0.161	1.000	1.059
8	MR16P9	36	36	2.129	0.144	1.446	0.144	1.440	0.153	1.000	1.060
9	MR17A9	25	25	2.953	0.127	1.720	0.127	1.903	0.135	1.000	1.063
10	MR18A9	17	17	4.792	0.127	2.577	0.127	2.490	0.135	1.000	1.063
11	MR18P3	28	28	2.351	0.127	1.893	0.135	1.873	0.141	0.941	1.044
12	MR19A3	23	23	3.247	0.119	2.002	0.127	2.169	0.128	0.937	1.008
13	MR19A9	40	40	2.139	0.144	1.499	0.144	1.592	0.144	1.000	1.000
14	MR20A3	17	17	0.864	0.135	0.595	0.144	0.818	0.152	0.938	1.056
15	MR20P9	16	16	0.907	0.144	0.646	0.152	0.777	0.159	0.947	1.046
16	MR21A3	39	39	0.932	0.152	0.654	0.161	0.484	0.207	0.944	1.286
17	MR22A3	25	25	2.578	0.127	1.551	0.135	1.582	0.149	0.941	1.104
18	MR22A9	25	25	1.210	0.135	0.853	0.152	0.954	0.158	0.888	1.039
19	MR22P3	25	25	0.554	0.152	0.447	0.169	0.493	0.186	0.899	1.101
20	MR23A9	34	34	1.952	0.161	1.311	0.161	1.341	0.162	1.000	1.006
21	MR24A3	45	45	2.291	0.127	1.605	0.135	2.188	0.141	0.941	1.044
22	MR24P9	35	34	1.930	0.161	1.533	0.152	2.063	0.161	1.059	1.059
23	MR25A3	35	35	1.024	0.161	0.783	0.152	0.774	0.161	1.059	1.059
24	MR26A9	35	35	3.239	0.135	1.914	0.144	2.829	0.155	0.938	1.076
25	MR26P9	21	21	10.365	0.119	4.898	0.127	5.270	0.127	0.937	1.000

The authors developed a FORTRAN program to perform the PDF spectral estimation. Tables 14 and 15 show the spectral peak $S(f_p)$ and peak frequency f_p of each estimator. The orders of the AR model determined by AIC and CAT are also listed in the tables. The spectral peak frequency plays a significant role in the design of offshore

Table 15. Spectral analysis summary of wave data from Waverider buoy B in August 1981. (data obtained near Port Mansfield, Texas)

				AR		FFT		PDF			
NO	ID	AIC	CAT	$S(f_p)$	f_p	$S(f_p)$	f_p	$S(f_p)$	f_p	Ω_{AF}	Ω_{PF}
1	AG11P9	16	16	0.149	0.296	0.127	0.296	0.131	0.288	1.000	0.973
2	AG12A9	17	17	0.294	0.212	0.231	0.212	0.230	0.220	1.000	1.038
3	AG12P3	31	31	0.141	0.229	0.110	0.229	0.140	0.213	1.000	0.930
4	AG13P9	20	20	0.398	0.178	0.314	0.186	0.345	0.202	0.957	1.086
5	AG14P9	24	24	0.241	0.178	0.218	0.178	0.193	0.196	1.000	1.101
6	AG15A3	20	20	0.256	0.178	0.226	0.186	0.244	0.202	0.957	1.086
7	AG15P3	13	13	0.193	0.178	0.135	0.178	0.100	0.186	1.000	1.045
8	AG15P9	21	21	0.110	0.178	0.097	0.186	0.119	0.265	0.957	1.425
9	AG16A9	14	14	0.177	0.186	0.067	0.195	0.076	0.249	0.954	1.277
10	AG17A3	18	18	0.104	0.288	0.092	0.288	0.104	0.281	1.000	0.976
11	AG17P9	33	33	0.132	0.332	0.118	0.313	0.136	0.285	1.029	0.911
12	AG18P3	52	52	0.234	0.102	0.044	0.279	0.054	0.209	0.366	0.749
13	AG18P9	40	40	0.108	0.110	0.083	0.330	0.114	0.300	0.333	0.909
14	AG19A3	36	36	0.102	0.288	0.085	0.288	0.072	0.287	1.000	0.997
15	AG19P3	25	25	0.028	0.135	0.018	0.144	0.012	0.262	0.938	1.819
16	AG19P9	39	39	0.114	0.322	0.096	0.330	0.089	0.315	0.976	0.955
17	AG20A3	21	21	0.086	0.245	0.073	0.245	0.077	0.289	1.000	1.180
18	AG20P3	26	26	0.069	0.296	0.059	0.296	0.065	0.306	1.000	1.034
19	AG21A9	19	19	0.244	0.229	0.218	0.229	0.254	0.228	1.000	0.996
20	AG21P9	18	18	0.097	0.271	0.088	0.262	0.134	0.282	1.034	1.076
21	AG22A9	18	18	0.097	0.195	0.080	0.212	0.089	0.240	0.920	1.132
22	AG23A3	18	18	0.350	0.195	0.243	0.203	0.266	0.219	0.961	1.079
23	AG25A9	53	53	0.110	0.220	0.090	0.212	0.094	0.235	1.038	1.108
24	AG24P3	54	54	0.100	0.178	0.076	0.186	0.081	0.256	0.957	1.376
25	AG24P9	49	49	0.220	0.229	0.144	0.229	0.181	0.230	1.000	1.004

structures, and it may be useful to look at the deviation of estimated spectral peak frequencies among estimators. Ratios of AR spectral peak to FFT and PDF spectral peak to FFT are denoted as Ω_{AF} and Ω_{PF}, and are shown in the tables. The general tendency of estimators may be indicated by means of these values, which are as follows:

1) 25 wave data sets in March 1980

$\overline{\Omega}_{AF}$=0.966 ($\pm$0.054)

$\overline{\Omega}_{PF}$=1.061 ($\pm$0.063)

2) 25 wave data sets in August 1981

$\overline{\Omega}_{AF}$=0.935 ($\pm$0.179)

$\overline{\Omega}_{PF}$=1.090 ($\pm$0.210)

The number inside the parenthesis indicates corresponding standard deviation. The data in March 1980 imply that the AR spectral peak appears at a slightly lower frequency than the FFT spectral peak. On the other hand the PDF spectra tend to estimate slightly higher spectral peaks than the FFT spectral estimator.

Each spectral estimator is plotted at the same scale, and examples are shown in Figures 42, 43, and 44. In general, the PDF spectra seem similar to the FFT estimator, however, it has smoother features, and a sharper slope in the lower frequency. The AR estimator always exhibits higher and narrower peaks than the FFT. Figure 44 shows the worst case in which the PDF spectra significantly departs from the other estimators. It seems that the PDF spectra tend to shift the peaks to a higher frequency and provide a smoother shape over the entire frequency domain.

These values $\overline{\Omega}_{AF}$ and $\overline{\Omega}_{PF}$ of August 1981 imply a similar conclusion; however, the large standard deviation indicates that a careful evaluation should be carried out. The spectral shape of August 1981 is a wider spectra of higher frequency than data in March 1980. Some of the spectra show multi-spectral peaks. Figure 45 shows an example of a

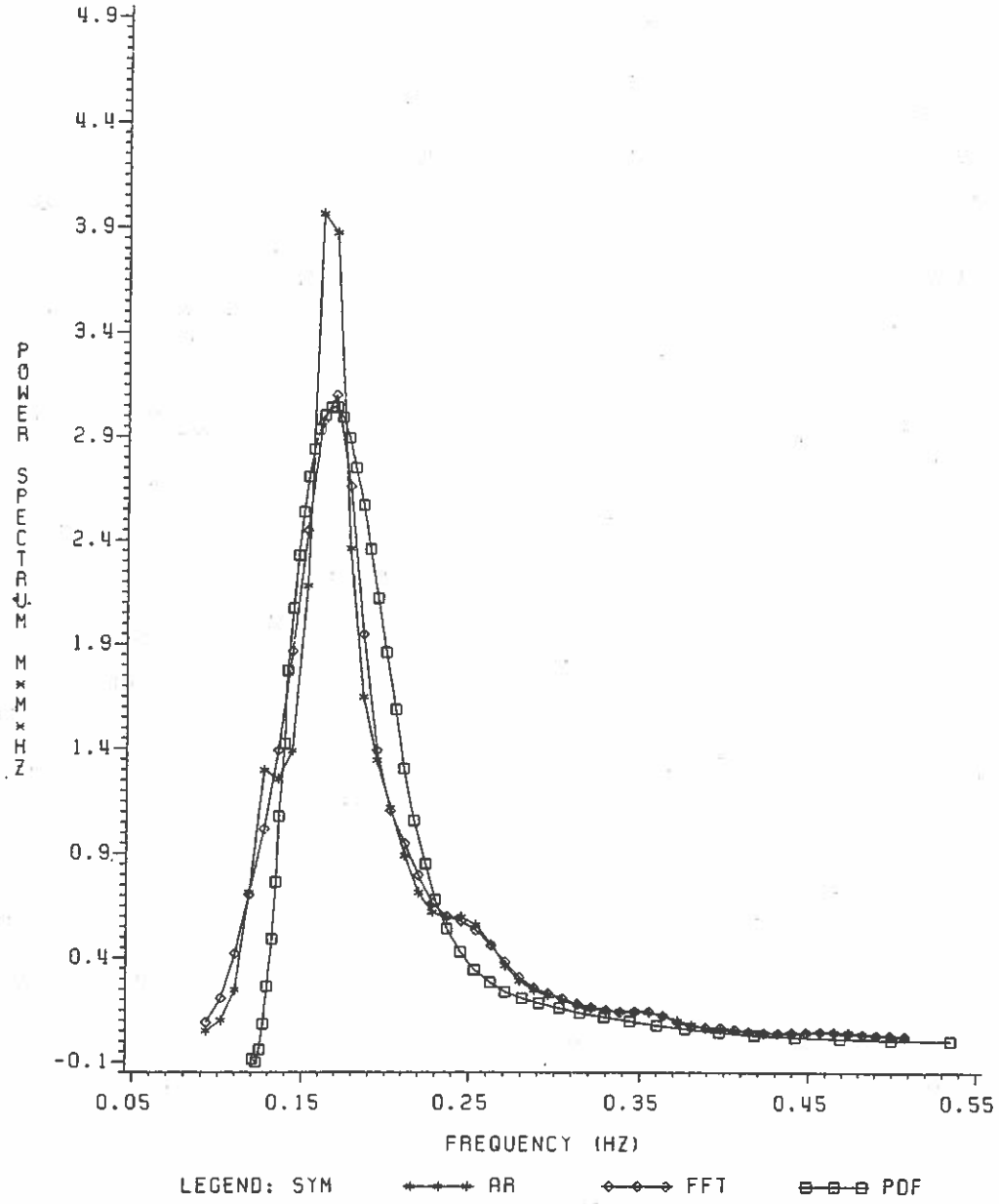

Figure 42. A comparison of spectral estimators, data ID MR13P3. (Data obtained by Waverider buoy B at 3 p.m. March 13, 1980, near Port Mansfield, Texas)

single spectral peak data set, where the PDF spectra again agree well

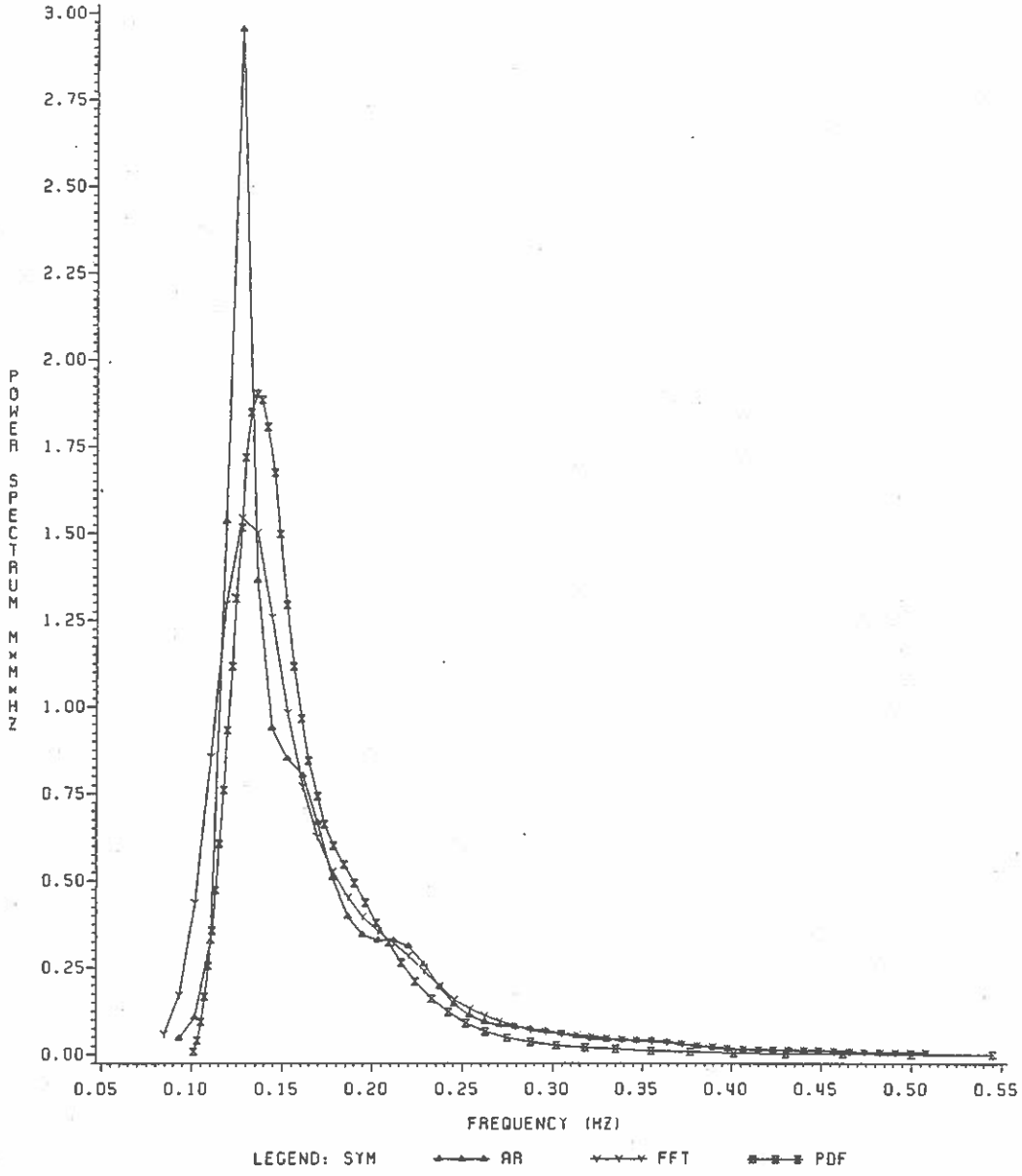

Figure 43. A comparison of spectral estimators, data ID MR17A9. (Data obtained by Waverider buoy B at 9 a.m. March 17, 1980, near Port Mansfield, Texas)

with the FFT spectra. On the other hand, Figures 46 and 47 show

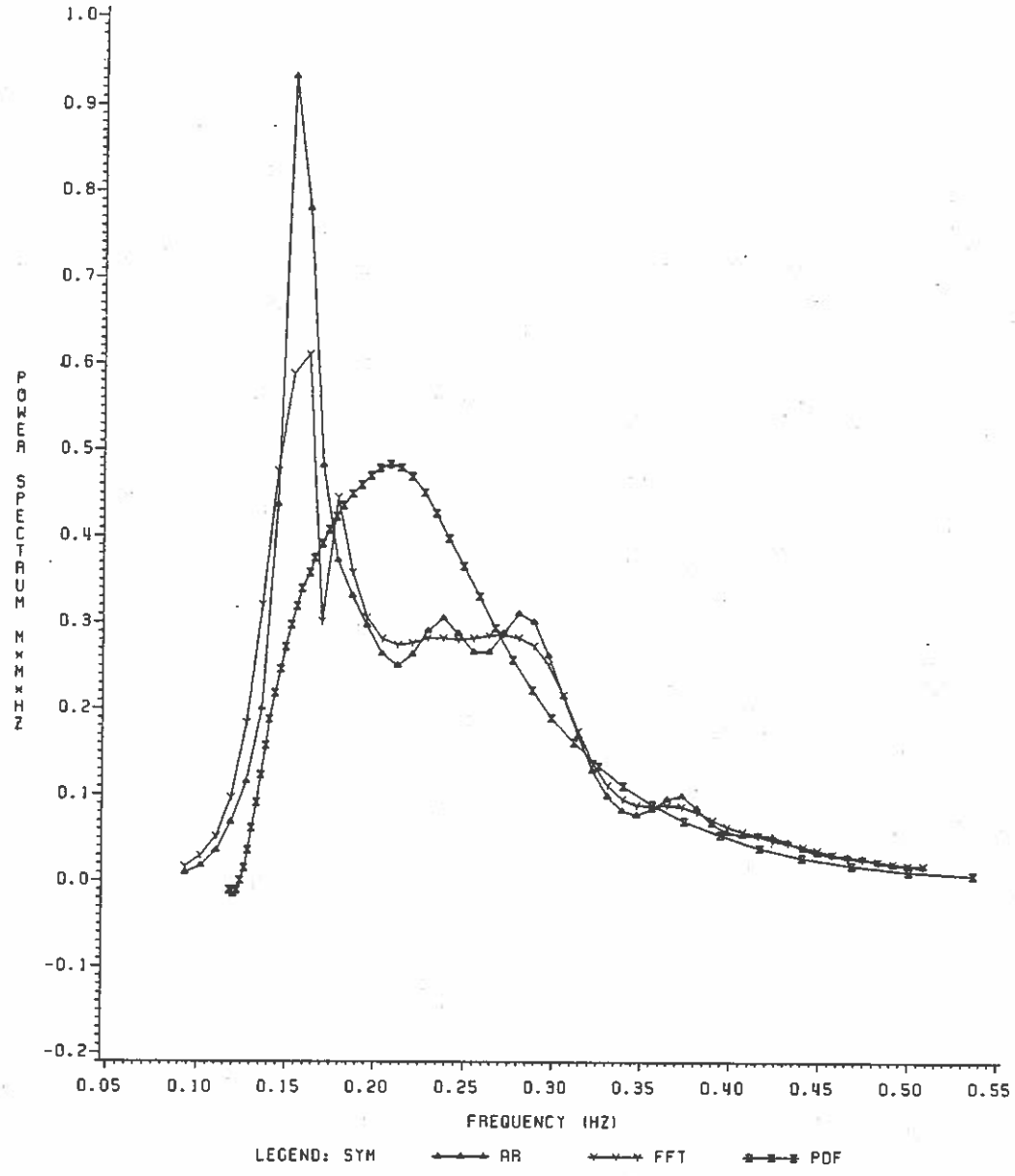

Figure 44. A comparison of spectral estimators, data ID MR21A3. (Data obtained by Waverider buoy B at 3 a.m. March 21, 1980, near Port Mansfield, Texas)

considerably different features among estimated spectra. The AR

spectra of data AG18P9 (Figure 46). show a very high peak in the low frequency range. FFT spectra also indicates the existence of spectrum peak near the low frequency. However, the PDF spectra do not show evidence of a lower frequency spectral peak. The sampling property of individual waves, namely the zero up-crossing method, does not allow the detection of the low frequency wave component, which reflects the lack of PDF spectra resolution in the lower frequency domain.

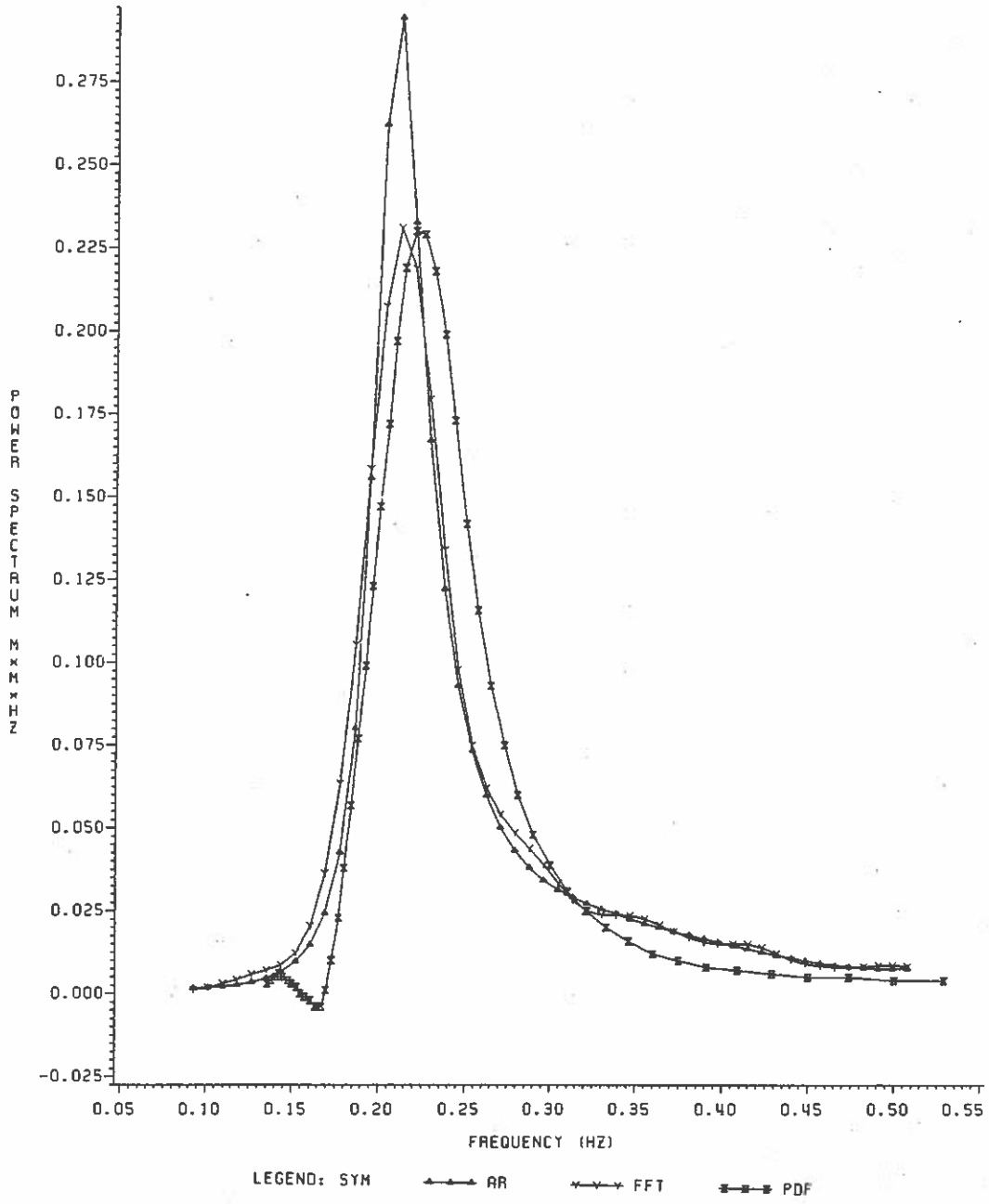

Figure 45. A comparison of spectral estimators, data ID AG12A9. (Data obtained by Waverider buoy B at 9 a.m. August 12, 1981, near Port Mansfield, Texas)

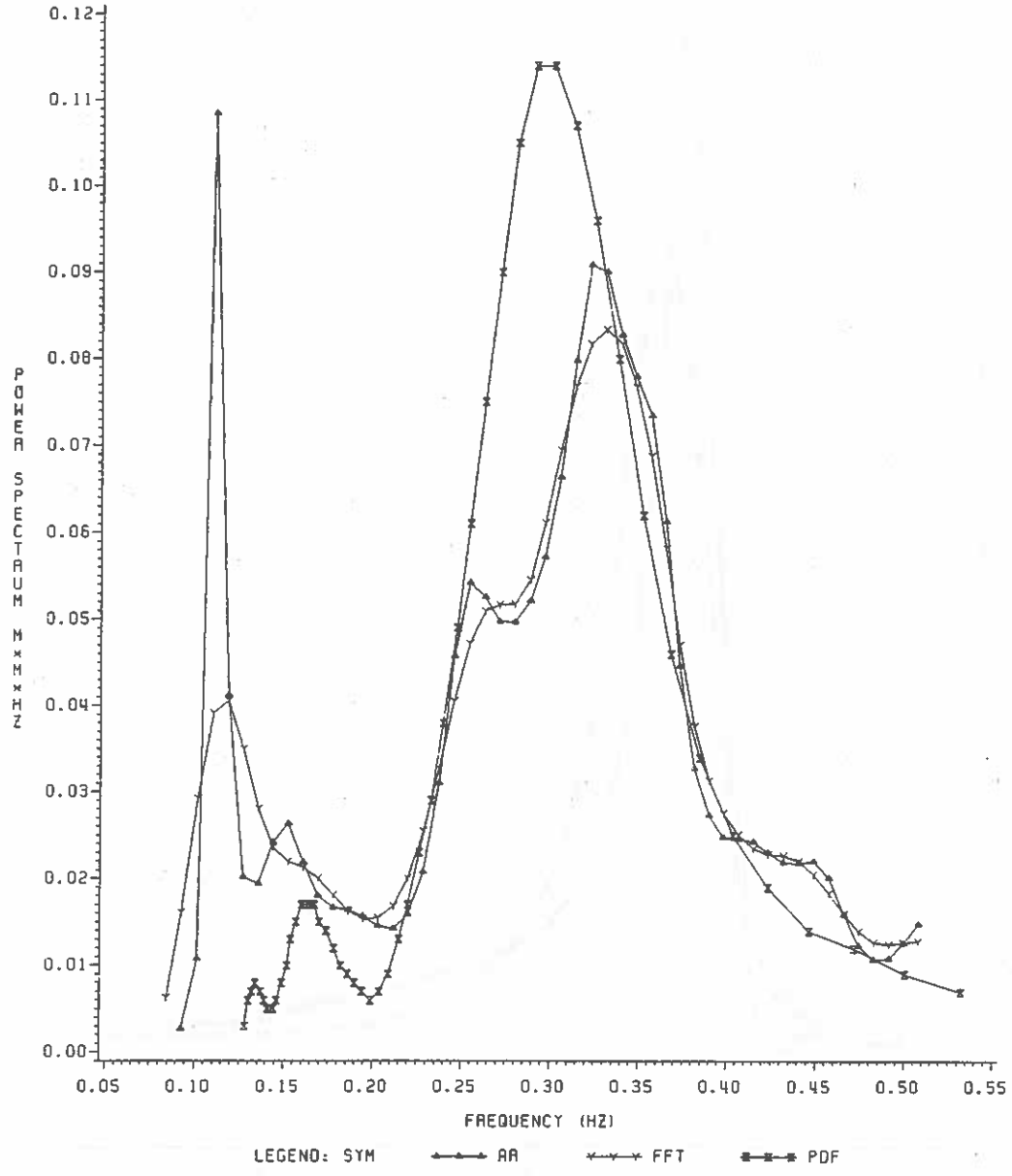

Figure 46. A comparison of spectral estimators, data ID AG18P9. (Data obtained by Waverider buoy B at 9 p.m. August 18, 1981, near Port Mansfield, Texas)

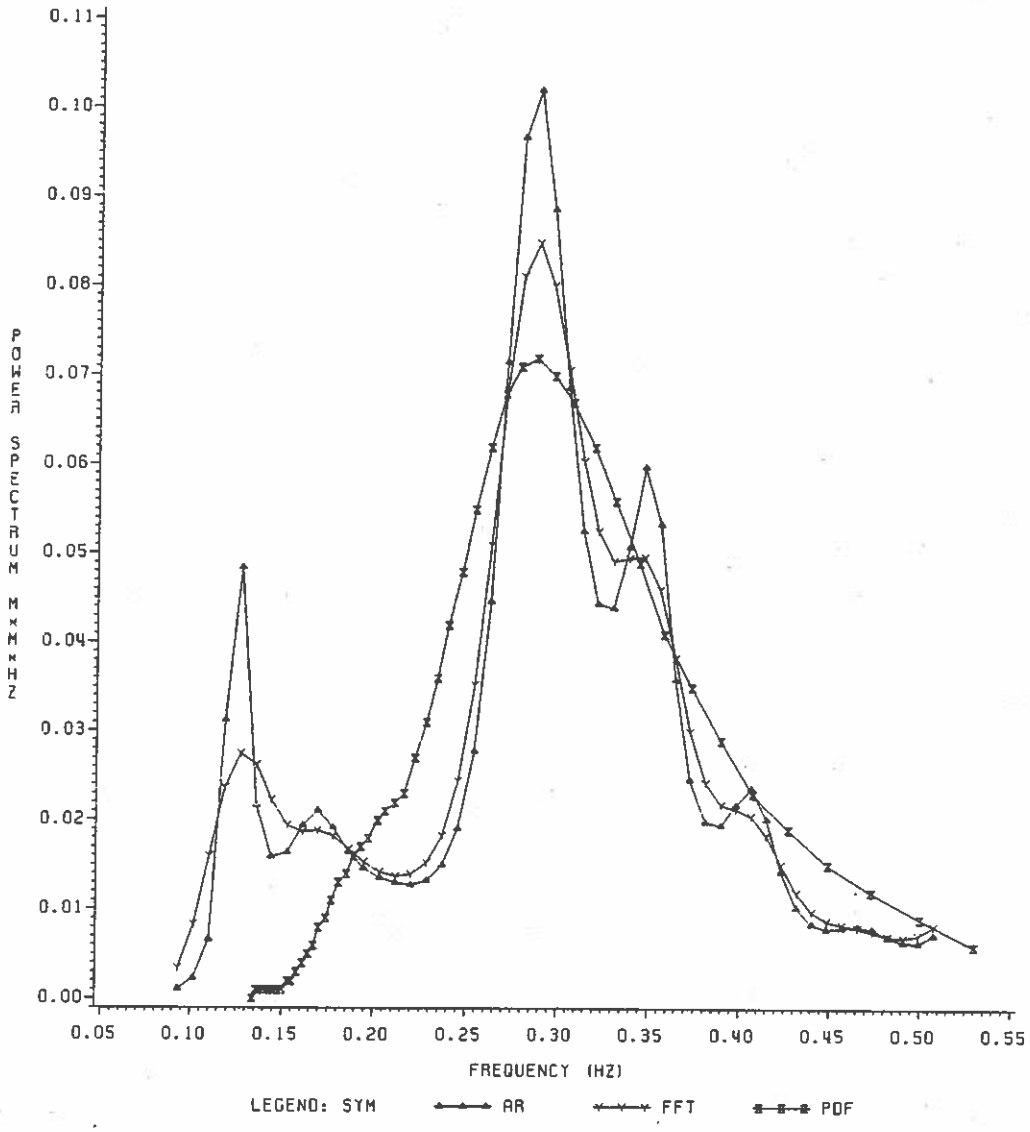

Figure 47. A comparison of spectral estimators, data ID AG19A3. (Data obtained by Waverider buoy B at 3 a.m. August 19, 1981, near Port Mansfield, Texas)

CONCLUSIONS

Conclusions of Present Study

An extensive literature survey was conducted to list available diagnostic tools of wave data analysis, namely wave statistics and wave spectra. Both strategies involve two approaches: nonparametric estimation and parametric estimation. Therefore the wave data analysis can be categorized as

1) Parametric Wave STATistic analysis (PWSTAT),

2) Nonparametric Wave STATistic analysis (NWSTAT),

3) Parametric Wave SPECtrum analysis (PWSPEC), and

4) Nonparametric Wave SPECtrum analysis (NWSPEC).

PWSTAT has the advantage that the pdf can be completely determined by estimating a few parameters. For example, the Rayleigh distribution of wave heights is a one parameter family. The parameter is usually either the variance of surface elevation $\overline{\eta}^2$ or the root mean square of wave heights $\overline{H}^2$. A slight modification of the Rayleigh distribution is a practical and acceptable way to describe the wave height statistics of not only moderate but higher wave heights, if sample wave heights exhibit a single probability density peak. Less computational effort is needed in estimating $\overline{\eta}^2$ than in estimating $\overline{H}^2$. Having assumed the population density is the Rayleigh distribution, the wave height statistics are determined. However, the sample wave heights cannot be characterized by a simple function, as

the sample distribution may exhibit multimodal density peaks. Therefore PWSTAT is not a good method to represent a summary of wave data.

Extensive research has been conducted in the last decade concerning nonparametric density estimation. However, most of the work has been done by the statisticians. Rigorous theoretical arguments make it difficult to apply the techniques to other fields. A special effort was made to develop a unique method of wave data analysis based on nonparametric estimation of the joint pdf of wave heights and periods employing tensor product B-splines. This method is a member of NWSTAT. Most of the existing nonparametric techniques lack an objective method for the choice of smoothing parameters. The Objective Least square Norm (OLN) is proposed for an objective choice of smoothing parameters for developed technique. Despite the lack of theoretical justification of OLN, a numerical examination shows that the proposed technique reproduces the population density very well.

The wave spectral analysis is the most popular diagnostic method of analyzing wave data. The Blackman-Tukey method and the FFT method are mathematically equivalent methods of NWSPEC. Because of significant computational efficiency, the FFT has been almost exclusively used for wave data analysis. A suitable spectral window and truncation point are required to estimate the power spectrum. The Parzen window has superior characteristics compared with other windows, and the Tukey window gives a result very close to the Parzen window. Both windows show almost the same result if equivalent window widths are chosen. An

objective method to choose the optimal truncation point has not successfully been developed to date. There are only empirically derived suggestions. Truncation points of 80-100 seem to work well for the wave data used in this study.

A great deal of research has been conducted in AR spectral estimation, which has become an important method of spectral estimation. The method consists of basically fitting an Auto-Regressive (AR) model to a time series. The wave spectra may be estimated parametrically employing the AR spectral estimator, PWSPEC. The AR model identification involves two stages: the estimation of AR coefficients and the determination of the optimal order of AR model. The Yule-Walker method is employed to estimate AR coefficients in this study. Recently Newton (1983b) reported that the Burg method was superior to the Yule-Walker method. The order determination was made by Parzen's CAT in this study. Comparison with AIC showed that both criteria agreed in 49 out of 50 wave data sets presented in this study. The result is consistent with the comments made by Parzen (1982). According to the present study, the wave data require a high order of the AR model. A deterministic harmonic component tends to increase the order of the AR model, therefore the high order AR model of wave data may imply the existence of deterministic harmonics in the data. Another interesting feature of the AR spectra shown in this study is that the AR spectra always have higher and sharper peaks than the FFT method. It has not yet been determined whether the AR spectra or the FFT spectra is superior for wave data analysis. However, the FFT method seems to achieve an adequate estimation of wave spectra.

The pdf spectra is classified in NWSPEC. The advantage of the pdf spectra is that the long-term wave spectra may be estimated from the joint pdf based on long-term data. The present study suggests that a concise long-term data representation may be achieved by either the pdf spectra or the nonparametric joint pdf.

A comparison of the FFT spectra, the AR spectra and the pdf spectra shows that the pdf spectra agree well with the others. However, consistent statistical properties of pdf spectra have not been identified in this study.

Recommended Future Studies

The proposed nonparametric density estimation does not possess positivity due to the non-shape preserving property of B-splines. Linear programming techniques may be applied to remove the limitation of subjecting the monotonicity condition on the spline coefficients. According to a recent personal communication, Schumaker is developing a density estimation technique by means of a spline by applying these linear programming techniques. However, the bivariate extension is not as simple as the regular tensor product splines and the computational effort is expected to increase significantly. Carlson and Fritsch (1981) developed a bivariate monotone spline interpolation which is a bicubic polynomial in a class C^1 of continuous function. Thus this method guarantees reproduction of the properties of a joint cdf. However, the spline is a C^1 class function and the estimated joint pdf,

obtained by differentiating the joint cdf, possesses discontinuous derivatives at knots. The estimated joint pdf should require, at least, a C^2class function. It is desirable to develop spline techniques which satisfy the condition of cdf and pdf.

Another approach is to transfer the data into some functional space, and apply the regular B-spline technique. However, this method may not guarantee the positivity of density functions.

The theoretical justification of the proposed nonparametric density estimation introducing the entropy concept should be conducted to investigate the statistical properties of the technique. The statistical properties of the pdf spectra should be investigated to make it more compatible with empirical spectral estimators. The confidence interval is an important diagnostic tool to indicate the reliability of the estimator. Intuitively, it is a function of the bin width of a histogram and the number of data in a bin.

APPENDIX A

SPLINE FUNCTIONS

General

Functional approximation arises in many fields. The true function may be estimated from observed data based on a certain type of approximation theory. The spline function has been extensively studied by mathematicians in the late 1960's to present. The B-spline is a most useful class of spline functions. The spline is a relatively new mathematical tool; however, Schoenberg (1946) suggested the idea of B-spline was known to Laplace. B-splines with equal space knots were first used by Schoenberg in 1946. A detailed historical note of spline functions is described by Schumaker (1981).

The major difficulty arises as an interpolation problem in ordinary polynomial interpolation. The ordinary polynomial interpolation does not possess convergence (Schumaker,1981). In other words, interpolated function does not approach the true function as the number of interpolation points increases. A superior property of B-splines is that it possesses convergence in interpolation.

The importance of B-spline is the basic recursion relation of the B-spline base which will be discussed in the next section. The relation was discovered by three authors simultaneously, Cox (1972), de Boor (1972) and Mansfield (1972 in Shumaker,1981). The recursion relation of the derivative of the B-spline base was found by de Boor (1972). Because it is easy to store, evaluate and manipulate data on a digital computer, B-splines have been used in a variety of fields.

B-splines

Due to the lack of convergence of an ordinary polynomial, it was decided to use smooth piecewise polynomials for approximation purposes. The polynomial splines are smooth piecewise polynomials defined in a closed interval. Let $[a,b]$ be a finite closed interval. The knots Δ are defined

$$\Delta = \{x_i\}_{i=1}^{k} \tag{162}$$

$$\text{with } a = x_o < x_1 < \ldots < x_k < x_{k+1} = b$$

Define a partition of Δ as follows

$$I_i[x_i, x_{i+1}) \quad \text{for } i=0,1,\ldots,k-1$$

$$\text{and } I_k[x_k, x_{k+1}]$$

Let m be a positive integer, corresponding to the order of the polynomial spline, and let $M=(m_1,..,m_k)$ be a vector of integers with $1<m_i<m$, $i=1,..,k$ corresponding to the multiplicities of knots. The space of the polynomial spline of order m with knots $\Delta = \{x_1, .., x_k\}$ of multiplicities $M=(m_1,.., m_k)$ is defined as

$$S(P_m;M;\Delta) = \{S;\ \text{there exist polynomials } S_o, S_1, \ldots, S_k \text{ in } P_m \text{ such that } S(x) = S_i(x) \text{ for } x \varepsilon I_i$$
$$i=0,..,k \text{ and } D^j S_{i-1}(x_i) = D^j S_i(x_i) \text{ for}$$
$$j=0,1,\ldots,m-1-m_i, i=1,\ldots,k\}$$

where P_m is the space of polynomials of order m, and D^j is the differential operator of order j.

For example, a cubic piecewise polynomial defined in a closed interval [0,1] with the knots

$$\Delta = \{0.1, 0.2, \ldots, 0.9\}$$

is in a spline space $S(.)$, such that

$$S(.) = \{S_o, .., S_q \text{ are cubic polynomials}$$
$$S(x) = S_i(x) \text{ for } x \varepsilon I_i \text{ and } D^j S_{i-1}(x_i)$$
$$= D^j S_i(x_i) \text{ for } j=0,1,2, i=1,\ldots,9\}$$

Here $k=9$, $m=4$, $\underset{\sim}{m}=(1,\ldots,1)$. This means $S(x)$ is a smooth piecewise cubic polynomial with continuous derivatives up to second order at the knots Δ.

A useful class of spline is the class of B-splines. The m-th order B-spline base is defined as follows:

$$B_i^m(x) = \begin{cases} (-1)^m (y_{i+m} - y_i)[y_i, \ldots, y_{i+m}](x-y)_+^{m-1} & \text{if } y_i \le x \le y_{i+m} \\ 0 & \text{otherwise} \end{cases} \quad (163)$$

where y_i's are called an extended partition $\tilde{\Delta} = \{y_i\}_1^{2m+k}$, which is defined as

$$y_i \le \ldots \le y_m \le a$$
$$b \le y_{m+k+1} \le \ldots \le y_{2m+k}$$

and

$$y_{m+1} \le \ldots \le y_{m+k} = x_1, \ldots x_1, \ldots, x_k, \ldots, x_k$$
$$\underbrace{\qquad}_{m_1} \qquad\qquad \underbrace{\qquad}_{m_k}$$

$$K = \sum_{j=1}^{k} m_j$$

The truncation function $(x-y)_+^m$ (de Boor 1978) is defined as

$$(x-y)_+^m = (x-y)^m (x-y)_+^o \tag{164}$$

where

$$(x-y)_+^o = \begin{cases} 0 & x<y \\ 1 & x \geq y \end{cases}$$

The m-th order divided difference for a function f is

$$[y_i,\ldots,y_{i+m}]f = \frac{D\begin{pmatrix} y_i,\ldots,y_{i+m} \\ 1,x,\ldots,x^{m-1},f \end{pmatrix}}{D\begin{pmatrix} y_i,\ldots,y_{i+m} \\ 1,x,\ldots,x^{m} \end{pmatrix}} \tag{165}$$

where D(.) is a determinant such that

$$D\begin{pmatrix} y_i,\ldots,y_{i+m} \\ 1,x,\ldots,x^{m-1},f \end{pmatrix} = \begin{vmatrix} 1 & y_i & \ldots & y_i^{m-1} & f(y_i) \\ 1 & y_{i+1} & \ldots & y_{i+1}^{m-1} & f(y_{i+1}) \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 1 & y_{i+m} & \ldots & y_{i+m}^{m-1} & f(y_{i+m}) \end{vmatrix} \tag{166}$$

$$D\begin{pmatrix} y_i,\ldots,y_{i+m} \\ 1,x,\ldots,x^{m} \end{pmatrix} = \begin{vmatrix} 1 & y_i & \ldots & y_i^{m-1} & y_i^m \\ 1 & y_{i+1} & \ldots & y_{i+1}^{m-1} & y_{i+1}^m \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 1 & y_{i+m} & \ldots & y_{i+m}^{m-1} & y_{i+m}^m \end{vmatrix} \tag{167}$$

If some of the y_i's are repeated numbers, derivatives must be taken of the corresponding row (Schumaker 1981).

The importance of B-spline is a recursion relation. The m-th order B-spline can be expressed as a sum of the (m-1)-st order B-spline

$$B_i^m(x) = \frac{x-y_i}{y_{i+m}-y_i} B_i^{m-1}(x) + \frac{y_{i+m}-x}{y_{i+m}-y_{i+1}} B_{i+1}^{m-1}(x) \tag{168}$$

Observing theat the first order B-spline is

$$B_i^1(x) = (-1)^1 (y_{i+1}-y_i)\ [y_i, y_{i+1}]\ (x-y)_+^0$$

$$= (-1)\ (y_{i+1}-y_i)\ \frac{\begin{vmatrix} 1 & (x-y_i)_+^0 \\ 1 & (x-y_{i+1})_+^0 \end{vmatrix}}{\begin{vmatrix} 1 & y_i \\ 1 & y_{i+1} \end{vmatrix}}$$

$$= (x-y_i)_+^0 - (x-y_{i+1})_+^0$$

$$= \begin{cases} 1 & \text{for} \quad y_i \le x \le y_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

Making use of (168), the m-th order B-spline can be recurrently evaluated from the first order B-spline. The recursion relation provides another property of B-splines. Since $B^m(x)$ can be obtained by the lower order of B-spline starting with the first order, B-spline $B^1(\chi)$ is always a positive function, a careful observation reveals that the m-th order B-spline is a positive function, i.e.

$$B_i^m(x) > 0 \qquad \text{for} \quad y_i < x < y_{i+m}$$

It can be proved that a sum of B-spline ordinates is a partition of unity (Schumaker, 1981)

$$\sum_{i=j+1-m}^{j} B_i^m(x) = 1 \tag{169}$$

An unknown function f can be approximated in terms of the m-th order B-spline with knots Δ, extended partition $\tilde{\Delta}$ and multiplicities M,

$$\delta f(x) \to S(x)$$

where δ is the spline operator. The dimension of the B-spline space is $m+K=m+\sum_{i=1}^{k} m_i$. Simple knots are extensively used in most practical applications. The dimension of the space for simple knots is $m+k$.

$S(x)$ can be expressed as a linear combination of B-splines

$$S(x) = \sum_{i=1}^{m+k} C_i B_i^m(x) \qquad (170)$$

A value of $B_i^m(x)$ at a given point x can be evaluated from (168). Therefore one does not actually deal with the divided difference. Remembering that $B_i^m(x)$ is a locally supported positive fuction, the matrix of the system of equations to be solved is a 2m+1 bounded matrix. It has been shown that this matrix is positive definite. The solution of this type of system of equations is known to be numerically stable.

Once C_i's are calculated, $S(x)$ at a given point should be calculated using the following relation rather than the recursion relation of $B_i^m(x)$ (Schumaker,1981)

$$S(x) = \sum_{i=1}^{n+j-1} C_i^{[j]}(x) B_i^{m-j+1}(x) \qquad (171)$$

for any $j=1,\ldots,m$

where $C_i^{[1]} = C_i, i=1,\ldots,n$

and the $C_i^{[j]}(x)$ can be computed recurrently by setting $C_{n+j}^{[j]}=C_o^{[j]}=0$, all j and using

$$C_i^{[j+1]}(x) = \begin{cases} \dfrac{(x-y_i)C_i^{[j]}(x) + (y_{i+m-j}-x)C_{i-1}^{[j]}(x)}{y_{i+m-j}-y_i} & \text{otherwise} \\ 0 & \text{if } y_{i+m-j}-y_i=0 \end{cases} \tag{172}$$

for $i=1,2,..,n+j$ and $j=1,2,...,m-1$

If $y_\ell \le x \le y_{\ell+1}$, then

$$S(x) = \sum_{i=1}^{n+m=1} C_i^{[m]}(x)B_i^1(x)$$

$$= C_\ell^{[m]}(x) \tag{173}$$

Therefore $S(x)$ can be evaluated by constructing the $C_i^{[j]}$ array. This method is computationally faster than (168).

The derivative of the B-spline representation of $S(x)$ can be evaluated using the following recursion relation (de Boor 1972, Schumaker 1981). Suppose $1 \le d \le m$, then for all $y_m \le x < y_n$

$$D_+^{d-1}S(x) = \sum_{i=d}^{n} C_i^{(d)}B_i^{m-d+1}(x) \tag{174}$$

where $C_i^{(1)} = C_i, i=1,2,..,n$ and

$$C_i^{(j)} = \begin{cases} (m-j+1)\dfrac{C_i^{(j-1)}-C_i^{(j-1)}}{y_{i+m-j+1}-y_i} & \text{if } y_{i+m-j+1}-y_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{175}$$

For example, the first order derivative of the cubic B-spline can be obtained as

$$D_+S(x) = \sum_{i=2}^{n} C_i^{(2)}B_i^3(x)$$

where

$$C_i^{(2)} = 3\,\frac{C_i - C_{i-1}}{y_{i+3} - y_i}$$

The derivative of B-splines can be expressed as a function of the difference of the B-spline coefficients.

The indefinite integral of the m-th order B-spline can be expanded in terms of B-splines of the (m+1)-st order (Schumaker,1981). The integration can be considered as the antiderivative of a spline. The integration form of B-spline has a similar expression with (174). Let $D_{y_1}^{-1}S(x)$ be the antiderivative of S(x), then for all

$$D_{y_1}^{-1}s(x) = \int_{y_i}^{x} S(t)dt = \sum_{i=1}^{n} C_i^{(-1)} B_i^{m+1}(x) \qquad (176)$$

where

$$C_i^{(-1)} = \sum_{j=1}^{i} C_j \frac{y_{m+j} - y_j}{m}, \quad i=1,2,\ldots,n \qquad (177)$$

$y_1, y_2, .., y_n, .., y_{n+m}$ are the extended partition

The interpolation schemes of B-splines must be introduced for practical application purposes. Here the complete cubic spline interpolation scheme is discussed and is used in this study. Suppose data $\{z_i\}$ are given at $k+2$ distinct points $\{t_i\}$ for an unknown function f. The spline S(x), which interpolates these data points, is defined in a closed interval $[t_o, t_{k+1}]$. The knots Δ are taken as

$$x_i = t_i \quad \text{for} \quad i=1,2,\ldots,k$$

The extended partition $\tilde{\Delta}$ is

$$y_i = y_2 = y_3 = y_4 = t_o$$

$$y_{i+4} = t_i, \quad i=1,2,\ldots,k$$

$$y_{k+5} = y_{k+6} = y_{k+7} = y_{k+8} = t_{k+1}$$

The dimension of the space is $n=4+k$. Since the number of data points is $k+2$, two more conditions must be specified to complete a set of linear equations. If first order derivatives are known at both edges, the number of conditions coincides with the dimension of the space, and the scheme is called the complete spline interpolation. When the first order derivative of f is not known, one may take the second order derivative to be zero. This type of interpolation problem is known as natural spline interpolation. The elastic beam problem may be treated by natural splines.

Tensor Product Spline

The surface approximation problem has been studied by a number of mathematicians. A fairly complete summary of recent developments can be read in Schumaker (1982). A special case of the surface approximation problem is when the data are given at regular grid points. The most computationally efficient method is the application of the tensor product spline. The tensor product spline has the form of a tensor product of one dimensional polynomial splines (de Boor 1978, Schumaker 1981). An important feature of tensor product splines is that the algebraic properties of the one dimensional spline spaces are not altered in tensor product space.

Suppose the function $f(x,y)$ to be approximated is defined in a closed rectangle, such that

$$a_x \leq x \leq b_x$$

$$a_y \leq y \leq b_y$$

Let Δ_x and Δ_y be partition of a closed interval in x and y

$$\Delta_x = \{a_x = x_{x,0} < x_{x,1} < \ldots < x_{x,k_{x+1}} = b_x\} \tag{178.1}$$

$$\Delta_y = \{a_y = x_{y,0} < x_{y,1} < \ldots < x_{y,k_{y+1}} = b_y\} \tag{178.1}$$

and partition data sets $\tilde{\Delta}_x, \tilde{\Delta}_y$ are defined similar to the one dimensional spline

$$y_{x,1} \leq \ldots \leq y_{x,m_x} \leq a_x$$

$$y_{y,1} \leq \ldots \leq y_{y,m_y} \leq a_y$$

$$b_x \leq y_{x,m_x+k_x+1} < \ldots < y_{x,2m_x+k_x}$$

$$b_y \leq y_{y,m_y+k_y+1} < \ldots < y_{y,2m_y+k_y}$$

$$y_{x,m_x+1} \leq \ldots \leq y_{x,m_x+k_x}$$

$$y_{y,m_y+1} \leq \ldots \leq y_{y,m_y+k_y}$$

$$\tilde{\Delta}_x = \{y_{x,1}, \ldots, y_{x,2m_x+k_x}\} \tag{179.1}$$

$$\tilde{\Delta}_y = \{y_{y,1}, \ldots, y_{y,2m_y+k_y}\} \tag{179.2}$$

Associated multiplicity vectors are

$$M_x = (m_{x,1}, m_{x,2}, \ldots, m_{x,kx}) \tag{180.1}$$

$$M_y = (m_{y,1}, m_{y,2}, \ldots, m_{y,ky}) \tag{180.2}$$

Then the bivariate tensor product polynomial spline space is defined by

$$S_{xy} = S(P_{m_x}; M_x; \Delta_x) \quad S(P_{m_y}; M_y; \Delta_y) \tag{181}$$

where P_{m_x} is a m_x th order polynomial space and P_{m_y} is a m_y th order polynomial space.

Because of a computational efficiency, the B-spline is a most useful spline for practical purposes. Further discussion will be concentrated on the B-spline tensor product splines. The tensor product spline space may be expressed as

$$S_{xy} = \text{span } \{B_i^{m_x}(x)B_j^{m_y}(y)\}_{i=1\ j=1}^{m_x+k_x\ m_y+k_y} \tag{182}$$

where
$$K_x = \sum_{i=1}^{k_x} m_{x,i} \qquad 1 \le m_{x,i} \le m_x$$
$$K_y = \sum_{i=1}^{k_y} m_{y,i} \qquad 1 \le m_{y,i} \le m_y$$

It should be noted that the tensor product is bilinear.

A function $f(x,y)$ may be approximated by the tensor product B-spline.

$$\delta f(x,y) \to S(x,y)$$

where δ is a spline operator, and

$$S(x,y) = \sum_{i=1}^{m_x+k_x} \sum_{j=1}^{m_y+k_y} C_{ij} B_i^{m_x}(x) B_j^{m_y}(y) \tag{183}$$

The C_{ij}'s are coefficents, the dimension of space is $(m_x + k_x)x(m_y + k_y)$.

As a specific application of tensor product splines, the complete cubic tensor product spline with simple knots is discussed. The multiplicity vectors are

$$M_x = (1,\ldots,1)$$
$$M_y = (1,\ldots,1)$$

The order of the cubic spline is 4, therefore the dimension of the space is

$$\dim S = (4+k_x)(4+k_y)$$
$$= k_x k_y + 4(k_x+k_y) + 16$$

where k_x and k_y are the number of distinguished simple knots in x and y. The tensor product cubic B-spline base with simple knots is depicted in Figure 4 8.

Suppose the values of $f(x,y)$ are given at $\Delta_x \otimes \Delta_y$ (178.1),(178.2). Let Z_{ij} be the data, such that

$$Z_{ij} = f(x_i,y_j) \qquad \text{for } \begin{matrix} i=0,1,\ldots,k_x,k_x+1 \\ j=0,1,\ldots,k_y,k_y+1 \end{matrix} \qquad (184)$$

The approximated surface has the values

$$S(x_i,y_j) = Z_{ij} \qquad \text{for } \begin{matrix} i=0,1,\ldots,k_x,k_x+1 \\ j=0,1,\ldots,k_y,k_y+1 \end{matrix}$$

The number of conditions provided is

$$(k_x+2)(k_y+2) = k_x k_y + 2(k_x+k_y) + 4$$

Therefore the number of edge conditions to be added is

$$2(k_x+k_y) + 12$$

Suppose the first order derivatives are known at the edges. Let

$$\frac{\partial}{\partial x} f(x_o,y_i) = Z^x_{o,i} \qquad \text{for } i=0,1,\ldots,k_y+1$$

$$\frac{\partial}{\partial x} f(x_{k_x+1},y_i) = Z^x_{k_x+1,i} \qquad \text{for } i=0,1,\ldots,k_y+1$$

$$\frac{\partial}{\partial y} f(x_i,y_o) = Z^y_{i,k_y+1} \qquad \text{for } i=0,1,\ldots,k_x+1$$

$$\frac{\partial}{\partial y} f(x_i,y_{k_y+1}) = Z^y_{i,k_y+1} \qquad \text{for } i=0,1,\ldots,k_x+1$$

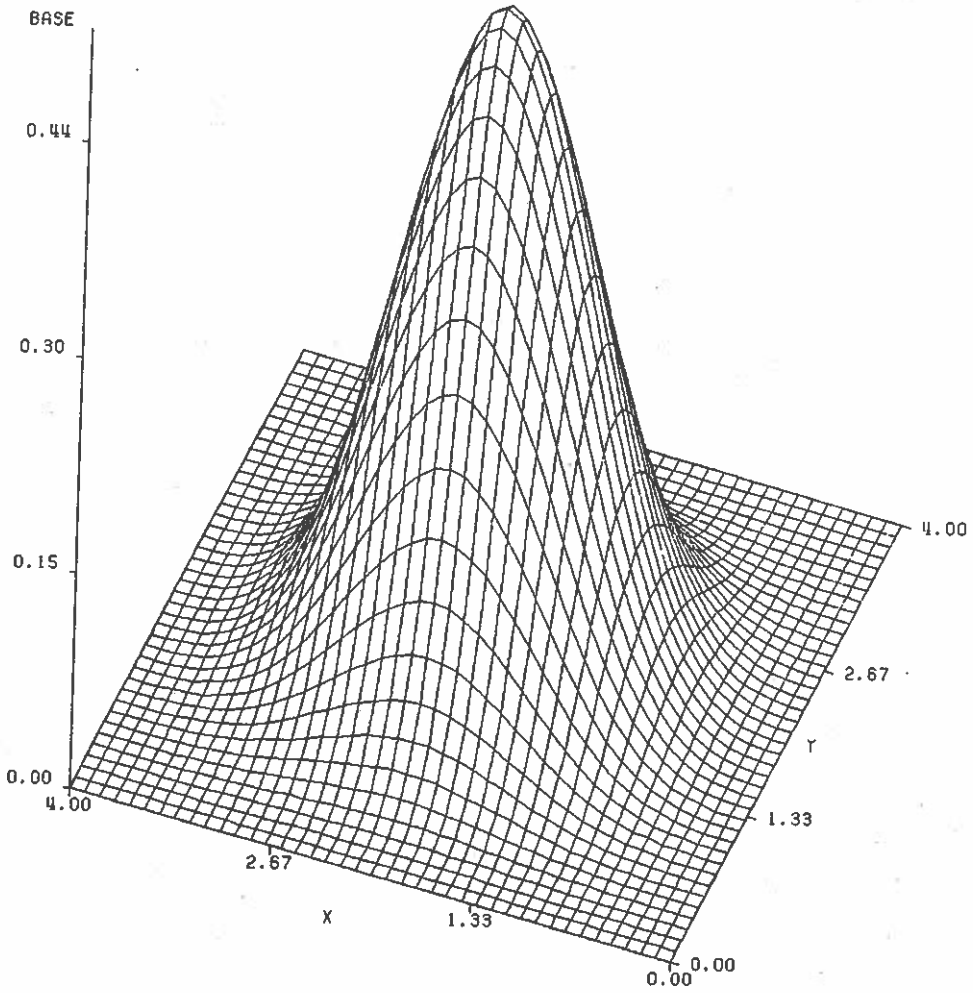

Figure 48. Tensor product cubic B-spline base, simple knots locate 0, 1, 2, 3 and 4 in X and Y

and

$$\frac{\partial^2}{\partial x \partial y} f(x_i, y_j) = Z^{xy}_{i,j} \qquad \text{for} \quad \begin{matrix} i=0, k_x+1 \\ j=0, k_y+1 \end{matrix}$$

The above condition completes the non-trivial set of linear equations. Since the order of the spline is specified (cubic), the upper script of $B_i(x)$ can be omitted without confusion. Let n_x and n_y be dimensions of space in x and y

$$n_x = k_x+4$$

$$n_y = k_y+4$$

$$S(x,y) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} C_{ij} B_i(x) B_j(y) \qquad (185)$$

de Boor (1979) introduced an efficient method to calculate the tensor product spline coefficents C_{ij}. Let V_{ij} be the given condition at x_i and y_i for $i = 1,2,..,n_x$ and $j=1,2,..,n_y$, the system of linear equations is

$$C_{11}B_1(x_1)B_1(y_1) + C_{12}B_1(x_1)B_2(y_1)+\ldots+C_{1n_y}B_1(x_1)B_{n_y}(y_1)+\ldots$$

$$\ldots+C_{n_x1}B_{n_x}(x_1)B_1(y_1)+ \quad \ldots \quad +C_{n_xn_y}B_{n_x}(x_1)B_{n_y}(y_1) = V_{11}$$

$$C_{11}B_1(x_1)B_1(y_2) + C_{12}B_1(x_1)B_2(y_2)+\ldots+C_{1n_y}B_1(x_1)B_{n_y}(y_2)+\ldots$$

$$\ldots+C_{n_x1}b_{n_x}(x_1)B_1(y_2)+ \quad \ldots \quad +C_{n_xn_y}B_{n_x}(x_1)B_{n_y}(y_2) = V_{12}$$

$$\vdots$$

$$C_{11}B_1(x_{n_x})B_1(y_{n_y}) + C_{12}B_1(x_{n_x})B_2(y_{n_y})+\ldots+C_{1n_y}B_1(x_{n_x})B_{n_y}(y_{n_y})+\ldots$$

$$\ldots+C_{n_x1}B_{n_x}(x_{n_x})B_1(y_{n_y})+\ldots+C_{n_xn_y}B_{n_x}(x_{n_x})B_{n_y}(y_{n_y})=V_{n_xn_y}$$

The above system can be expressed as

$$d_{11}B_1(x_1) + d_{12}B_2(x_1) + \ldots + d_{1n_x}B_{n_x}(x_1) = V_{11}$$
$$d_{21}B_1(x_1) + d_{22}B_2(x_1) + \ldots + d_{2n_x}B_{n_x}(x_1) = V_{12}$$
$$\vdots$$
$$d_{n_y1}B_1(x_1) + d_{n_y2}B_2(x_1) + \ldots + d_{n_yn_x}B_{n_x}(x_1) = V_{1n_y}$$
$$\vdots$$
$$d_{11}B_1(X_{n_x}) + d_{12}B_2(x_{n_x}) + \ldots + d_{1n_x}B_{n_x}(x_{n_x}) = V_{n_x1}$$
$$\vdots$$
$$d_{n_y1}B_1(x_{n_x}) + d_{n_y2}B_2(x_{n_x}) + \ldots + d_{n_yn_x}B_{n_x}(x_{n_x}) = V_{n_xn_y}$$

where

$$d_{ji} = C_{k1}B_1(y_j) + C_{i2}B_2(y_j) + \ldots + C_{in_y}B_{n_y}(y_j)$$

$i=1,2,\ldots,n_x$ and $j=1,2,\ldots,n_y$

A careful observation reveals that the above system is equivalent to n_y independent $n_x.n_x$ linear systems of equations. Considering that the original system of equation is $(n_x \times n_x)\times(n_y \times n_y)$ linear systems of equations, a significant reduction of computational effort can be expected. The above system of equations can be put into matrix forms as follows:

$$\begin{bmatrix} B_1(x_1) & B_2(x_1) & \ldots & B_{n_x}(x_1) \\ B_1(x_2) & B_2(x_2) & \ldots & B_{n_x}(x_2) \\ \vdots & \vdots & & \vdots \\ B_1(x_{n_x}) & B_2(x_{n_x}) & \ldots & B_{n_x}(x_{n_x}) \end{bmatrix} \begin{bmatrix} d_{11} \\ d_{12} \\ \vdots \\ d_{1n_x} \end{bmatrix} = \begin{bmatrix} V_{11} \\ V_{21} \\ \vdots \\ V_{n_x1} \end{bmatrix}$$

$$\begin{bmatrix} B_1(x_1) & B_2(x_{n_x}) & \ldots & B_{n_x}(x_1) \\ B_1(x_2) & B_2(x_2) & \ldots & B_{n_x}(x_2) \\ \vdots & \vdots & & \vdots \\ B_1(x_{n_x}) & B_2(x_{n_x}) & \ldots & B_{n_x}(x_{n_x}) \end{bmatrix} \begin{bmatrix} d_{21} \\ d_{22} \\ \vdots \\ d_{2n_x} \end{bmatrix} \begin{bmatrix} V_{12} \\ V_{22} \\ \vdots \\ V_{n_x2} \end{bmatrix}$$

$$\begin{bmatrix} B_1(x_1) & B_2(x_1) & \cdots & B_{n_x}(x_1) \\ B_1(x_2) & B_2(x_2) & \cdots & B_{n_x}(x_2) \\ \vdots & \vdots & & \vdots \\ B_1(x_{n_x}) & B_2(x_{n_x}) & \cdots & B_{n_x}(x_{n_x}) \end{bmatrix} \begin{bmatrix} d_{n_y 1} \\ d_{n_y 2} \\ \vdots \\ d_{n_y n_x} \end{bmatrix} = \begin{bmatrix} V_{1 n_y} \\ V_{2 n_y} \\ \vdots \\ V_{n_x n_y} \end{bmatrix}$$

Solving the coefficient matrix $\{d_{ij}\}$, the coefficient matrix $\{C_{ij}\}$ can be calculated by

$$\begin{bmatrix} B_1(y_1) & B_2(y_1) & \cdots & B_{n_y}(y_1) \\ B_1(y_2) & B_2(y_2) & \cdots & B_{n_y}(y_2) \\ \vdots & \vdots & & \vdots \\ B_1(y_{n_y}) & B_2(y_{n_y}) & \cdots & B_{n_y}(y_{n_y}) \end{bmatrix} \begin{bmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{1 n_y} \end{bmatrix} = \begin{bmatrix} d_{11} \\ d_{21} \\ \vdots \\ d_{n_y 1} \end{bmatrix}$$

$$\begin{bmatrix} B_1(y_1) & B_2(y_1) & \cdots & B_{n_y}(y_1) \\ B_1(y_2) & B_2(y_2) & \cdots & B_{n_y}(y_2) \\ \vdots & \vdots & & \vdots \\ B_1(y_{n_y}) & B_2(y_{n_y}) & \cdots & B_{n_y}(y_{n_y}) \end{bmatrix} \begin{bmatrix} C_{21} \\ C_{22} \\ \vdots \\ C_{2 n_y} \end{bmatrix} = \begin{bmatrix} d_{12} \\ d_{22} \\ \vdots \\ d_{n_y 2} \end{bmatrix}$$

$$\begin{bmatrix} B_1(y_1) & B_2(y_1) & \cdots & B_{n_y}(y_1) \\ B_1(y_2) & B_2(y_2) & \cdots & B_{n_y}(y_2) \\ \vdots & \vdots & & \vdots \\ B_1(y_{n_y}) & B_2(y_{n_y}) & \cdots & B_{n_y}(y_{n_y}) \end{bmatrix} \begin{bmatrix} C_{n_x 1} \\ C_{n_x 2} \\ \vdots \\ C_{n_x n_y} \end{bmatrix} = \begin{bmatrix} d_{1 n_x} \\ d_{2 n_x} \\ \vdots \\ d_{n_y n_x} \end{bmatrix}$$

It should be noted that the transpose of the $\{d_{ij}\}$ matrix is used solving for $\{C_{ij}\}$. The importance of this method is that it only

requires two small size inverse matrices and the matrices are bounded matrices. In particular, the matrices are tridiagonal matrices for the case of a complete cubic spline with simple knots. Once the $[\ C_{ij}\ ]$ matrix is obtained, the one dimensional B-splines algebraic properties can be used for the tensor product spline. The interpolation can be achieved as

$$S(x,y) = \sum_{k=1}^{n_x} \sum_{j=1}^{n_y} C_{ij} B_i(x) B_j(y)$$

$$= \sum_{i=\ell_x-3}^{\ell_x} \sum_{j=\ell_y-3}^{\ell_y} C_{ij} B_i(x) B_j(y) \qquad (186)$$

where

$$y_{x,\ell_x} \leq x < y_{x,\ell_x+1}$$

$$y_{y,\ell y} \leq y < y_{y,\ell_y+1}$$

Now

$$S(x,y) = \sum_{j=\ell_y-3}^{\ell_y} e_j(x)\ B_j(y) \qquad (187)$$

where

$$e_j(x) = \sum_{i=\ell_x-3}^{\ell_x} C_{ij}\ B_i(x)$$

APPENDIX B

BASIC PROBABILITY NOTIONS

The cumulative distribution function (cdf) of a random variable X, denoted $F_X(\cdot)$, is defined to be a function, which satisfies $F_X(\chi)=\text{Prob}[X\leq \chi]$, i.e. probability of X less than or equal χ, for every real number χ (Mood et. al. 1974). The cdf $F_X(\chi)$ has the following properties

1) $F_X(-\infty)=\lim_{\chi\to-\infty} F_X(\chi)=0$, and $F_X(\chi)=\lim_{\chi\to\infty} F_X(\chi)=1.0$

2) $F_X(\cdot)$ is a monotone, non-decreasing function; that is,

$F_X(a)\leq F_X(b)$ for $a \leq b$

3) $F_X(\cdot)$ is continuous from the right; that is

$$\lim_{0<h\to 0} F_X(\chi+h)=F_X(\chi)$$

The cdf $F_X(\cdot)$ can be defined uniquely for either discrete random variables or continuous random variables. However, both random variables, discrete or continuous, have different definitions of the probability density function (pdf). The pdf $f_X(\cdot)$ for a discrete random variable X has a value at distinct points $\chi_1, \chi_2, \ldots, \chi_n, \ldots$, such that

$$f_X(\chi) = \begin{cases} \text{Prob}[X=\chi_j] & \text{if } \chi=\chi_j,\ j=1,2,\ldots,\ \ldots \\ 0 & \text{otherwise} \end{cases} \qquad (188)$$

and possesses the following properties:

$$1) f_X(\chi_j)>0 \text{ for } j=1,2,\ldots, \qquad (189.1)$$

$$2) f_X(\chi)=0 \text{ for } \chi\neq\chi_j,\ j=1,2,\ldots, \qquad (189.2)$$

$$3) \sum_j f_X(\chi_j) = 1.0 \qquad (189.3)$$

On the other hand, the pdf $f_x(\chi)$ for a continuous random variable X has the following relationship with the cdf $F_x(\chi)$: if the $F_x(\chi)$ is sufficiently smooth,

$$f_x(\chi) = \frac{dF_x(\chi)}{d\chi} \tag{190}$$

and $f_x(\chi)$ satisfies

$$1) f_x(\chi) \geq 0 \quad \text{for all } \chi \tag{191.1}$$

$$2) \int_{-\infty}^{\infty} f_x(\chi)\, d\chi = 1.0 \tag{191.2}$$

A difficulty that arises in practical estimation of the pdf for a continuous random variable, is that only a finite number of sample observations from a pdf can be drawn. Density estimation is one of the main interests in the field of statistics. There are essentially two methods of estimating pdf, i.e., parametric estimation and nonparametric estimation. Nonparametric density estimation is given in chapter 3. It is necessary to discuss the expectations and moments of a random variable before proceeding to the discussion of parametric density estimation. The following discussion assumes that random variables are continuous. The mean of a random variable X is denoted by μ_x and is defined by

$$\mu_x = E[X] = \int_{-\infty}^{\infty} \chi f_x(\chi) d\chi \tag{192}$$

The mean μ_x describes where the values of the random variable X are centered. The variance of a random variable X is a measure of spread or dispersion of the density of X, is denoted by σ_x^2, and is defined by

$$\sigma_x^2 = Var[X] = E[(x-\mu_x)^2] = \int_{-\infty}^{\infty} (\chi-\mu_x)^2 f_x(\chi) d\chi \tag{193}$$

It can be easily shown σ_x^2 has an alternative expression, such that

$$\sigma_x^2 = E[(x-\mu_x)^2] = E[x^2] - \mu_x^2 \qquad (194)$$

The r-th moment of X about μ_x is denoted μ_r and is defined by

$$\mu_r = E[(x-\mu_x)^r] = \int_{-\infty}^{\infty} (\chi-\mu_x)^r f_x(\chi)d\chi \qquad (195)$$

The third and fourth moments often provide useful information on the distribution of a random variable. The third moment μ_3 measures asymmetry or skewness. The skewness is denoted β_1, and is defined by

$$\beta_1 = \frac{\mu_3}{\sigma_x^3} \qquad (196)$$

When a random variable has a normal distribution, β_1 is zero. But the converse is not true. The fourth moment is sometimes used as a measure of excess or kurtosis. The kurtosis is denoted β_2, and is defined by

$$\beta_2 = \frac{\mu_4}{\sigma_x^4} \qquad (197)$$

When a random variable has a normal distribution, β_2 is 3.0. But again converse is not true.

The expected value of a function of random variable X, say $g(\chi)$, is denoted $E[\,g(\chi)\,]$, and is defined by

$$E[g(\chi)\,] = \int_{-\infty}^{\infty} g(\chi)\, f_x(\chi)d\chi \qquad (198)$$

Another class of statistical parameters, quantiles, provides other measures of characterizing the density function. The q-th quantile of a random variable X is denoted $Q(q)$ and is defined as the smallest number χ satisfying $F_x(\chi) > q$, i.e.

$$Q(q) = F_x^{-1}(q) = \inf\{\chi : F_x(\chi) \geq q,\ 0 \leq q \leq 1.0\} \tag{199}$$

The function $Q(q)$ is sometimes called the quantile function of X.

The normal distribution is common in many application fields. A random variable X is said to be normally distributed if the pdf has the following form:

$$f_x(\chi) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\left(\frac{\chi-\mu_x}{\sigma_x}\right)^2\right], -\infty<\chi<\infty \tag{200}$$

where parameters μ_x and σ_x satisfy $-\infty<\mu_x<\infty$, $\sigma_x>0$. The normal distribution $f_x(\chi)$ is a two-parameter family. Given μ_x and σ_x, the density function can be obtained. It is easy to show the parameter μ_x is the mean of X, i.e. $\mu_x = E[X]$ and the parameter σ_x^2 is the variance of X, i.e. $\sigma_x^2 = Var[X]$. If one is willing to accept that a random variable X is normally distributed, the only task left is to estimate μ_x and σ_x of the density function. Suppose $\underset{\sim}{X}=[\chi_1,\chi_2,\ldots,\chi_N]^T$ is a random sample vector drawn from a normal density function having parameters μ_x and σ_x. The parameters μ_x and σ_x may be estimated by estimators $\theta_1(.)$ and $\theta_2(.)$. Let $\hat{\mu}_x$ and $\hat{\sigma}_x$ be estimated values of μ_x and σ_x, which are defined as

$$\hat{\mu}_x = \theta_1(\underset{\sim}{X}) \tag{201.1}$$

$$\hat{\sigma}_x = \theta_2(\underset{\sim}{X}) \tag{201.2}$$

The best estimator of μ_x is the sample mean, which is denoted by $\overline{X}$ and defined as

$$\hat{\mu}_x = \overline{X} = \frac{1}{N}\sum_{i=1}^{N} \chi_i \tag{202}$$

A popular estimator of σ_x is the sample variance estimator, which is denoted by S_x and defined as

$$\hat{\sigma}_x = S_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\chi_i-\bar{X})^2} \tag{203}$$

One may estimate the pdf using $\hat{\mu}_x$ and $\hat{\sigma}_x$. The estimated pdf is denoted by $\hat{f}_x(\chi)$:

$$\hat{f}_x(\chi) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}_x}\exp\left[-\left(\frac{\chi-\hat{\mu}_x}{\hat{\sigma}_x}\right)^2\right] \tag{204}$$

This type of density estimation is called parametric density estimation. In general, an m parameter $\alpha_1, \alpha_2, \ldots, \alpha_m$ family of density function for a random variable X is denoted by

$$f_x(\chi) = f_x(\chi;\alpha_1,\alpha_2,\ldots,\alpha_m) \tag{205}$$

The parametric estimation of the density function can be achieved by estimating $\alpha_1, \alpha_2, \ldots, \alpha_m$ by estimators $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_m$ if the random variable is drawn from the aforementioned density function. The estimated pdf is

$$\hat{f}_x(\chi) = \hat{f}_x(\chi;\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_m) \tag{206}$$

Sometimes a situation involves two random variables. The joint behavior of random variables X and Y is characterized by the joint cdf $F_{XY}(x,y)$ or the joint pdf $f_{XY}(x,y)$. The joint cdf $F_{XY}(x,y)$ is defined by

$$F_{XY}(x,y) = \text{Prob}[X \le x, Y \le y] \tag{207}$$

and has the following properties

1) $F_{XY}(-\infty,y\ ) = \lim_{x\to-\infty} F_{XY}(x,y) = 0$ for all y,

$F_{XY}(x,-\infty\ ) = \lim_{y\to-\infty} F_{XY}(x,y\ ) = 0$ for all x, and

$F_{XY}(\infty,\infty\ ) = \lim_{\substack{x\to\infty \\ y\to\infty}} F_{XY}(x,y) = 1.0$

2) If $x_1 \le x_2$ and $y_1 \le y_2$, then

$$\text{Prob}[x_1 \le X \le x_2, y_1 \le Y \le y_2] = F_{XY}(x_2,y_2\ ) - F_{XY}(x_2,y_1\ ) - F_{XY}(x_1,y_2\ ) + F_{XY}(x_1,y_1\ ) \ge 0$$

i.e. $F_{XY}(.,.\ )$ is a bivariate monotone increasing function.

3) $F_{XY}(.,.)$ is right continuous in each argument; that is

$$\lim_{0<h\to 0} F_{XY}(x+h,\ y) = \lim_{0<h\to 0} F_{XY}(x,y+h) = F_{XY}(x,y\ )$$

As in the univariate case, the joint pdf $f_{XY}(.,.)$ may be obtained by differentiating $F_{XY}(.,.)$ with respect to x and y if $F_{XY}(.,.)$ is sufficiently smooth:

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} \qquad (208)$$

It has similar properties to a univariate pdf, i.e.

1) $f_{XY}(x,y) \ge 0$ (209.1)

2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1.0$ (209.2)

The marginal probability density functions of bivariate random variables X and Y are denoted $f_X(x)$ and $f_Y(y)$, and obtained from

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy \qquad (210.1)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx \qquad (210.2)$$

It should be noted that $f_X(x)$ and $f_Y(y)$ do not characterize the joint behavior of X and Y, unless X and Y are independent, i.e:

$$f_{XY}(x,y) = f_X(x)f_Y(y) \qquad (211)$$

The conditional probability density function of Y given X= x, denoted by $f_{Y|X}(y|x)$, is defined by

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \qquad (212)$$

Similarly $f_{X|Y}(x|y)$ is the conditional probability density of X given Y= y, and is defined by

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \qquad (213)$$

The functions $f_{Y|X}(y|x)$, $f_{X|Y}(x|y)$ have the same properties as a univariate probability density function. The "linear" association of random variables X and Y is measured either by their covariance or their correlation coefficient. The covariance of X and Y is denoted by σ_{XY} and is defined as

$$\sigma_{XY} = \text{cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y)f_{XY}(x,y)dxdy \qquad (214)$$

The correlation coefficient of X and Y is denoted by ρ_{XY} and defined as

$$\rho_{XY} = \frac{\text{cov}[X,Y]}{\sigma_X\sigma_Y} \qquad (215)$$

The covariance σ_{XY} tends to measure the linear relationship of X and Y. However the actual value of σ_{XY} does not show the relative magnitude of linear association, because σ_{XY} depends on the variability of X and Y. The correlation coefficient ρ_{XY} has the following property,i.e.

$$|\rho_{XY}| \leq 1.0 \tag{216}$$

and the variabilities of X and Y do not alter ρ_{XY}. When ρ_{XY} has the value of zero, the random variables are said to be linearly uncorrelated.

The bivariate normal density function plays an important role among bivariate random variables. The bivariate normal pdf $f_{XY}(.,.)$ has the following form

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\{ - \frac{1}{2(1-\rho_{XY}^2)} \left[(\frac{x-\mu_X}{\sigma_X})^2 \right.$$

$$\left. - 2\rho_{XY}(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2 \right] \} \tag{217}$$

for $-\infty<x<\infty$, $-\infty<y<\infty$

where μ_X , μ_Y , σ_X , σ_Y ,and ρ_{XY} are parameters and it can be shown that those parameters are:

$\mu_X = E[X]$

$\mu_Y = E[Y]$

$\sigma_X^2 = Var[X]$

$\sigma_Y^2 = Var[Y]$

$$\rho_{XY} = \frac{\text{cov}[X,Y]}{\sigma_X \sigma_Y}$$

The marginal density functions have the form of univariate normal denstiy functions. It should be noted the bivariate normal distribution is independent, i.e. $f_{XY}(x,y)=f_X(x)f_Y(y)$, if, and only if, the random variables are linearly uncorrelated. However the general class of bivariate distributions can be dependent even if random variables are linearly uncorrelated.

REFERENCES

Akaike, H., Fitting autoregressive models for prediction, Ann. Inst. Statist. Math., 21, 243-247, 1969

Akaike, H., A new look at the statistical model identification, IEEE Trans. Autom. Contr., AC-19, 716-723, 1974

Battjes, J. A., Facts and figures pertaining to the bivariate Rayleigh distribution, Dept. of Civil Engineering, Delft University of Technology, 11p, 1969

Beamish, N., M. B. Priestley, A study of autoregressive and window spectral estimation, Appl. Stat., 30, No. 1, 41-58, 1981

Bean, S. J., C. P. Tsokos, Developments in nonparametric density estimation, Inter. Stat. Rev., 48, 267-287, 1980

Bennett, J. O., Estimation of multivariate probability density functions using B-splines, Ph. D. dissertation, Rice University, 164pp, 1974

Blackman, R. B., J. W. Tukey, The measurement of power spectra from the point of view of communication engineering, New York, Dover Publications, 1958

Boneva, L. I., D. G. Kendall, I. Stefanov, Spline transformations: Three new diagnostic aids for the statistical data-analyst, J. of Royal Stat. Soc., Series B33, 1-70, 1971

Bretschneider, C. L., Revised Wave Forecasting Curves and Procedure, Tech. Rep. No. HE-155-47, Inst. of Engineering Research, University of California, Berkeley, 1951

Bretschneider, C. L., Wave variability and wave spectra for wind generated gravity waves, U.S. Army Corps Engr. Beach Ero. Board, Tech. Memo, No. 113, 192p, 1959

Brooks, D. A., FESTSA (Fast and EaSy Time Series Analysis), Center for Marine and Coastal Studies, North Carolina State University, 1976

Burg, J. P., Maximum entropy spectral analysis, paper presented the 37th Annual International Meeting, Soc. of Explor. Geophys., Oklahoma, 1967

Carlson, R. E., Fritsch, F. N., Monotone piecewise bicubic Interpolation, submitted to SIAM J. Numer. Anal., 1981

Chakrabarti, S. K., R. P. Cooley, Statistical distribution of periods and heights of ocean waves, J. Geophys. Res., vol.82, No. 9, 1363-1368, 1977

Chatfield, C., The analysis of time series: an introduction, Chapman and Hall, London, 1980

Collins, J. I., Wave statistics from hurricane Dora, J. Waterways Harbors Div., Am. Soc. Civ. Engr., vol. 93, WW2, 59-77, 1967

Cooley, J. W., J. W. Tukey, An algorithm for the machine calculation of complex Fourier series, Mathematics of Computation, Vol. 19 No. 90, 291-301, 1965

Cox, M. G., The numerical evaluation of B-splines, J. Inst. Math. App., 10, 134-149, 1972

de Boor, C., On calculating with B-splines, J. Approximation Theory, 1, 219-235, 1972

de Boor, C., A practical guide to splines, Applied Math. Sci. 27, Springer-Verlag, New York, 1978

de Boor, C., Efficient computer manipulation of tensor products, ACM Trns. Math. Software, vol. 5, No. 2, 173-182, 1979

Edge, B. L., J. T. Moore, Results of the regional coastal waves workshops, Oceans 82, Marine Technology Society, Washington D.C., 809-813, 1982

Forristall, G. Z., On the statistical distribution of wave heights in a storm, J. Geophys. Res., 83, c5, 2353-2358, 1978

Goda, Y., Estimation of wave statistics from spectral analysis, Proc. Int. Symp. on Ocean Wave Measurement and Analysis, vol. 1, 320-337,1974 a

Goda, Y., Investigation of the statistical properties of sea waves with field and simulation data, Rep. of the Port and Harbour Res. Inst., vol. 13, No. 1, 3-37, 1974 b,(in Japanese)

Goda, Y., T. Tomotsuka, Y. Kishira, Theory and example of digital significant wave height indicator, Tech. Note of the Port and Harbour Res. Inst., No. 238, Ministry of Transport, Japan, 1-12, 1976,(in Japanese)

Goda, Y., The observed joint distribution of periods and heights of sea waves, 16-th Coastal Engineering, Hamburg, 227-245, 1978

Goodknight, R. C., T. L. Russell, Investigation of the statistics of wave heights, J. Waterways Harbor Div., Am. Soc. Civ. Engrs., vol. 89, No. WW2, 29-55, 1963

Herbich, J. B., R. Jensen, Wave Data Bank, Annual Report, Ocean Engineering Program, Texas A&M University, 1978

Herbich, J. B., R. K. Watanabe, Wave Data Bank for the Texas Coast, Annual Report, Ocean Engineering, Texas A&M University, 1979

Herbich, J. B., R. K. Watanabe, Wave Data Bank for the Texas Coast, TEES Technical Bulletin, Texas A&M University, July 1980

Herbich, J. B., H. Yamazaki, Monthly wave characteristics December 1979-May 1982 National Oceanographic Data Center Buoy No. 42001, No.42002, Tech. Rep., Ocean Engineering, Texas A&M University, to be published 1984

Huang, N. E., S. R. Long, An experimental study of the surface elevation probability distribution and statistics of wind-generated waves, J. Fluid Mech., vol. 101, part 1, 179-200, 1980

Iwagaki, Y., A. Kimura, A study on a characteristic of zero up-crossing wave length and celerity, Proc. 23th Japanese Conf. on Coastal Engg, 751-754, 1976

Jenkins, G. M., D. G. Watts, Spectral analysis and its applications, San Francisco, Holden-Day, 1968

Kimura, A., Joint distribution of wave heights and periods of random sea waves, Coastal Eng. in Japan, vol. 24, 77-92, 1981

Kinsman, B., Wind Waves, New Jersey, Prentices-Hall, 1965

Krinmal, R. A., M. E. Tarter, The estimation of probability densities and cummulatives by Fourier series methods, J. Am. Stat. Ass. 63, 925-952, 1976

Levinson, N., The Wiener (root mean square) error criterion in filter design and prediction, J. Math. Phys., 25, 261-278, 1948

Longuet-Higgins, M. S., On statistical distribution of the heights of sea waves, J. Mar. Res., 11(3), 245-266, 1952

Longuet-Higgins, M. S., The effect of nonlinearities on statistical distributions in the theory of sea waves, J. Fluid Mech., 17(3), 459-480, 1963

Longuet-Higgins, M. S., On the joint distribution of the periods and amplitudes of sea waves, J. Geophys. Res., 80, 2688-2694, 1975

Longuet-Higgins, M. S., On distribution of the heights of sea waves: some effects of nonlinearity and finite band width, J. Geophys. Res., 85, 1519-1523, 1980

Mood., A. M., F. A. Graybill, D. C. Boes, Introduction to the theory of statistics, McGraw-Hill Co., New York, 1974

Neave, H. R., A comparison of lag window generators, J. Am. Stat. Ass., vol 67, No. 337, 152-158, 1972

Newbold, P., Some recent developments in time series analysis, Inter. Stat. Rev., vol. 49, 53-66, 1981

Newton, H. J., An introduction to the methods of time series analysis in the time and frequency domain, Inst. Stat., Texas A&M University, 1983 a

Newton, H. J., M. Pagano, Computing for autoregressions, Computer Science and Statistics, Netherlands, North-Holland Pub Co., 113-118, 1983 b

Newton, H. J., M. Pagano, Simultaneous confidence bands for autoregressive spectra, Biometrika, vol. 1 1984

Nolte, K. G., F. H. Hsu, Statistics of larger waves in a sea state, J. Waterways Port Coastal Ocean Div., Am. Soc. Civ. Engr., vol. 105, WW4, 384-404, 1979

Parzen, E., On estimation of a probability density function and mode, Annal of Math. Stat., 33, 1065-1076, 1962

Parzen, E., An approach to empirical time series, J. Res. Nat. Bur. Standards, 68B, 937-951, 1964

Parzen, E., Multiple time series modeling, Multivariate Analysis II, P.R. Krishnaish, ed. New York, Academic Press, 389-409, 1969

Parzen, E., Multiple time series: determining the order of approximating autoregressive schemes, Multivariate Analysis IV, Krishnaiah, ed. North-Holland: Amsterdam, 283-295, 1977

Parzen, E., Nonparametric statistical data modeling, J. of Am. Stat. Ass., 74, 105-131, 1979

Parzen, E., Autoregessive spectral estimation, Tech. Rep. N-28, Inst. Stat., Texas A&M University, 48p, 1982

Priestley, M. B.,Spectral Analysis and Time Sereis, Academic Press, 1981

Putz, R. R., Statistical distribution for ocean waves, Tran. A. G. U., vol. 33, No.5, 685-692, 1952

Rice, S. O., The mathematical analysis of random noise, Bell Syst. Tech. J., 23: 282-332, 1944, 24: 46-156, 1945

Rosenblatt, M., Remarks on some nonparametric estimates of a density function, Annal. of Math. Stat., 27, 832-837, 1956

Schoenberg, I. J., Contributions to the problem of approximation of equidistant data by analytic functions, Part A: On the problem of smoothing graduation, a first class of analytic approximation formulae, Quart. Appl. Math., 4, 45-99, 1946

Schumaker, L. L., Spline functions: Basic theory, New York, John Wiley & Sons Inc., 1981

Schumaker, L. L., Fitting surface to scatter data, Technical Report, Department of Mathematics and Center of Numerical Analysis, University of Texas at Austin, 1982

Scott, D. W., J. R. Thompson, Probability density estimation in higher dimensions, Computer Science and Statistics, Netherlands, North-Holland Pub Co., 173-179, 1983

Steele, K., A. Johnson, Data buoy wave measurements, Ocean Wave Climate Symposium, vol. 8, 17pp, 1977

Sverdrup, H. U., W. H. Munk, Wind, sea and swell: Theory of relations for forecasting, U. S. Navy Hydrographic Office Pub., No. 601, 44pp, 1947

Tapia, R. A., J. R. Thompson, Nonparametric probability density estimation, The Johns Hopkins University Press, Baltimore, 1978

Tayfun, M. A., Narrow band nonlinear sea waves J. Geophys. Res., vol. 85, c3, 1548-1552, 1980

Tayfun, M. A., Distribution of crest to trough wave heights, J. Waterways Port Coastal Ocean Div., Am. Soc. Civ. Engrs. vol. 107, WW3, 149-158, 1981

Tayfun, M. A., Effect of spectrum band width on the distribution of wave heights and periods, Ocean Engr., No. 2, 107-118, 1983

Tucker, M. J., Analysis of records of sea waves, Proc. Inst. Civ. Engrs, London, vol. 26, No. 10, 305-316, 1963

Uhlenbeck, G. E., Theory of Random Processes, MIT Radiation Lab Rept., 454 pp, 1943

Ulrych, T. J., T. N. Bishop, Maximum entropy spectral analysis and autoregressive decomposition, Review of Geophysics and Space Physics, vol. 13, No. 1, 183-200, 1975

Venezian, G., C.L. Bretschneider, and S. Jagannathan, Cumulative Distribution of Forces on Structures Subjected to the Combined Action of Currents and Random Waves for Potential OTEC Sites, Vol. III, University of Hawaii, Look Lab Report 80-1, 244 pp, August 1980.

Wahba, G., Automatic smoothing of the log periodogram, J. of Am. Stat. Ass., vol. 75, No. 389, 122-132, 1980

Watters, J. K. A., Distribution of height in ocean waves, New Zealand Journal of Science and Technology, Sect. B, vol. 34, 408-422, 1953

Wegman, E. J., Nonparametric probability density estimation: I A summary of available methods, Technometrics, 14, 533-546, 1972

Wilson, J. R., W. F. Baird, A discussion of some measured wave data, Proc. 13th Coastal Engineering Conference, Vancouver B.C., 113-130, 1972

Woodfield, T. J., Statistical modeling of bivariate data, Tech. Rep. B-7, Inst. Stat., Texas A&M University, 263p, 1982

Wu, H. J., A discussion of the wave records obtained at weather station Papa, Pacific Ocean, Engineering Dynamics of the Coastal Zone, Sydney, 47-51, 1973