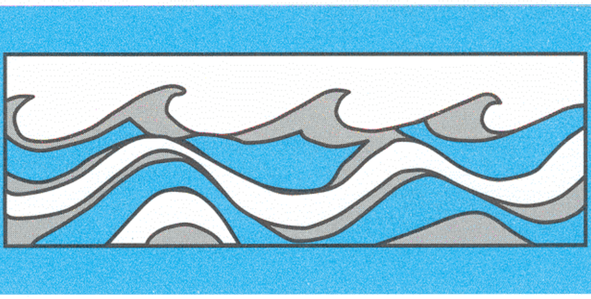
University of Washington
Department of Civil and Environmental Engineering

PHYSICAL CONSIDERATIONS IN THE ANALYSIS AND SYNTHESIS OF HYDROLOGIC SEQUENCES

K. Malcolm Leytham



Water Resources Series
Technical Report No. 76
June 1982

Seattle, Washington
98195

Department of Civil Engineering
University of Washington
Seattle, Washington 98195

PHYSICAL CONSIDERATIONS IN THE ANALYSIS AND SYNTHESIS OF HYDROLOGIC SEQUENCES

K. Malcolm Leytham

Water Resources Series
Technical Report No. 76

June 1982

Charles W. Harris Hydraulics Laboratory
Department of Civil Engineering
University of Washington
Seattle, Washington 98195

PHYSICAL CONSIDERATIONS IN THE ANALYSIS AND SYNTHESIS OF HYDROLOGIC SEQUENCES

by
K. Malcolm Leytham

Technical Report No. 76

June 1982

Project Completion Report: "Drought in the Pacific Northwest: Relationships between Atmospheric Pressure Patterns and Fall/Winter Precipitation" and "Model Evaluation"

Project Period: October 1, 1980 - September 30, 1982

Principal Investigator: Stephen J. Burges, Professor of Civil Engineering, University of Washington

OWRT Project Numbers: A-107-WASH; A-115-WASH

OWRT Agreement Numbers: 143-34-10E-3990-5304
143-34-10E-3991-3072

ABSTRACT

Consideration of qualitative relationships between precipitation and atmospheric circulation over the west coast of North America suggests that inter-station precipitation relationships are nonlinear with higher cross correlations during drought than normal or wet periods. This indicates that current multi-site stochastic models, which assume linear inter-station relationships, may underestimate the areal extent of drought.

The presence of nonlinear inter-station relationships is confirmed by analysis of monthly precipitation data. Evaluation of the performance of a simple multi-site stochastic model shows that current methods may seriously distort spatial drought characteristics. These difficulties occur in the synthesis of data both at widely separated sites (separation > 1000 km) and in high dimensionality problems (seven sites) where the maximum separation is less than 400 km.

The nature of the nonlinearities suggests that precipitation data may be modeled better using mixture models with precipitation during wet and dry periods drawn from different statistical distributions. Drought conditions are associated with meridional atmospheric circulation and wet conditions with zonal flow suggesting that atmospheric pressure data may be useful for classifying precipitation into wet and dry populations. Unfortunately, analysis of pressure and precipitation data failed to reveal useful quantitative relationships, and no objective method was found to classify precipitation data.

An investigation of univariate mixture models was undertaken including a study of the small sample properties of maximum likelihood parameter estimates for a mixture of two normal distributions. This study showed that while the parameters estimated from small

unclassified samples are unreliable, the estimated quantiles compare favorably with those estimated using classified data. Thus the ability to classify data from a mixture distribution is not necessarily important for hydrologic applications.

Analysis of long rainfall records from southern British Columbia showed that while mixture distributions fitted the data well, they are too complex to be justified for single-site precipitation modeling.

The use of multi-variate mixture distributions for multi-site precipitation modeling was explored. Such models are capable of preserving the marginal distributions and cross correlation structure of data which could not be modeled adequately using conventional models. Multi-variate mixture models also allow explicit recognition of the widespread nature of drought.

ACKNOWLEDGMENTS

This report was originally prepared as the author's Ph.D. dissertation. Many people provided encouragement, information and ideas in the course of this study. Special recognition is given to: Steve Burges and Dennis Lettenmaier for many discussions on topics related to this study; Doug Martin for introducing the author to the statistical literature on mixture models; Harry Edmons for providing the 500 mb data tapes and the software which made unreadable tapes readable; Gary Schaeffer of the Pacific Weather Center, Vancouver for helpful discussions on precipitation data; Jacques Faigenblum for a unique combination of good humor and computer resources; and Ed Albaugh of Crippen Consultants, Seattle for use of his organization's drafting equipment.

The report was typed and edited by Diana Leytham, without whom this work would not be seen in its present immaculate form.

The work upon which this publication is based was supported in part by funds provided by the Office of Water Research and Technology (Project Nos. A-107-WASH and A-115-WASH), U.S. Department of the Interior, Washington, D.C. through the Washington Water Research Center as authorized by the Water Research and Development Act of 1978.

TABLE OF CONTENTS

LIST OF FIGURES

Page

Page

Page

LIST OF TABLES

1.0 INTRODUCTION

One of the major determinants in the design and operation of water resource systems is the frequency and severity of drought. Historically, engineers and hydrologists have used observed streamflow sequences for design purposes. The observed sequences, however, are unlikely to be repeated in the future life of the system. Moreover, the short records available are unlikely to characterize adequately the infrequent drought events which control system design to such a large extent.

In response to these problems, hydrologists have developed or adapted a number of mathematical techniques for creating equally likely artificial streamflow sequences which they hope are representative of future flows. These techniques generally assume that the historic record is a sample from a time series which is stationary over the period of interest (i.e., from the beginning of the historic record to the end of the economic life of the project in question.) Sample statistics from the historic record are used to estimate the true parameters of the time series. For most practical purposes this has meant assuming either that the historic record gives the true parameters directly or that the true parameters may be obtained by simply correcting the sample statistics for small sample bias.

The methods for creating artificial streamflow sequences have been the subject of considerable research in the past decade, and the field of "stochastic hydrology" has risen to some prominence in the study of water resources. However, the techniques developed thus far have a number of shortcomings. I believe that the most significant of these is that the techniques are purely statistical with little or no physical basis and with no appeal to those mechanisms which control the natural generation of streamflow. A second area of concern is that most work until recently has been directed toward single-site applications, i.e., generation of data at a point. Despite such problems, these techniques are being used to model drought, a regional phenomenon, which can often be related to anomalies in the regional or even global atmospheric circulation.

The physical linkages between drought and atmospheric circulation are particularly clear in the Pacific Northwest. There high pressure ridges can act to prevent frontal systems from entering the area, steering them, as in the drought of 1977, to the north. The result is a pronounced rainfall deficit which can affect an extensive region as in 1977 from southern Alaska to southern California.

Although our ultimate interest may be in streamflow, there are obviously numerous important situations where the ability to synthesize rainfall data is of interest, e.g., in regional planning for irrigation. Moreover, deterministic simulation techniques exist for transforming rainfall into streamflow. Thus, the more general case of rainfall generation has been used as a surrogate for streamflow synthesis here.

The purpose of the research presented here is to explore techniques for using our knowledge of large-scale atmospheric circulation to improve the methods available for generating seasonal or annual synthetic rainfall sequences. The work considers conditions only on the western seaboard of North America, from southern California to the Gulf of Alaska, with particular emphasis on the Pacific Northwest, i.e., the state of Washington and southern British Columbia.

Past and present developments in the field of stochastic hydrology are reviewed in Chapter 2. Chapter 3 discusses the role of the atmospheric sciences in water resources planning and reviews observational studies relating precipitation to various features of the large-scale atmospheric circulation. The known qualitative relationships between precipitation and patterns of atmospheric circulation suggest that inter-station precipitation relationships are nonlinear with higher cross correlations during drought than during wet or normal periods. This indicates that current multi-site stochastic models which assume linear inter-station relationships may under estimate the areal extent of severe drought, particularly in applications involving large spatial scales.

An analysis of multi-site precipipation data is carried out in Chapter 4 and the performance of a simple multi-site stochastic model is evaluated. The results of this evaluation confirm that current multi-site models may seriously distort the spatial characteristics of droughts, and an alternative modeling approach based on sampling from mixed distributions is suggested.

The use of mixed distributions implies a classification of the data into two or more populations. One potentially attractive approach is to classify precipitation data into wet and dry populations based on the prevailing type of atmospheric circulation. An analysis of concurrent precipitation and atmospheric pressure data is presented in Chapter 5. The results of Chapter 5 demonstrate some weak relationships between precipitation and various indices of atmospheric circulation, but no method was found to permit an objective classification of the precipitation data based on the pressure data.

The parameters of simple mixture models can, however, be estimated in the absence of exogenous information suitable for classifying the data. In Chapter 6 the characteristics of simple univariate mixture models are explored with particular emphasis on the small sample properties of the maximum likelihood estimates of the parameters. For the record lengths generally available in the water resources field, the maximum likelihood estimates are shown to be unreliable, and no conclusive evidence is found to justify the use of mixtures in single-site applications.

A simple multivariate mixture model is discussed in Chapter 7. A method of parameter estimation based on a subjective classification of the precipitation data is suggested, and the performance of the model is explored in a number of situations. Finally, concluding remarks and recommendations for additional work are given in Chaper 8.

2.0 REVIEW OF STOCHASTIC HYDROLOGY

In this chapter I will discuss the philosophy of artificial data generation and recent developments in the field as they affect this research. The discussion is restricted to seasonal or annual models as shorter time interval models are not considered suitable for the problems addressed here. It should be noted that the great emphasis in research thus far has been placed on (1) development of models which replicate the moments and correlation structure of the historic time series and (2) statistical concerns such as estimation of small sample bias and evaluation of estimators. Physical considerations or attempts to understand further the natural generation mechanisms have been limited.

2.1 Model Development

Truly optimal design and operation of water resources systems could be readily achieved if the future streamflows were known with certainty. Since we will probably never be able to predict future flows accurately over an extended period of time, the question arises as to what flows should be used for design and operational purposes.

It is clear that the historic flow sequence at a site is only one of infinitely many possible and plausible sequences, and that use of the historic flows for design may be quite inappropriate. However, let us assume that the historic sequence is a sample from a stationary time series. The historic data may be used to estimate the parameters of the underlying stochastic process. A data generation scheme which preserves these parameters can then be used to create many equally likely streamflow sequences. Use of many such sequences in design enables us to determine, for example, the optimal project configuration. This is usually taken to be that configuration which maximizes some measure of expected economic return. Such sequences also allow us to make statements about the reliability of the system, the distribution of expected benefits and so on.

The first computer based data generation scheme was a single-site, lag-one, Markov model now known in the water resources field as the Thomas-Fiering model (Thomas and Fiering 1962). Subsequent development of this approach led to a multi-season, multi-site, lag-one, Markov model which was designed to preserve the mean, variance, skew, and lag-zero and lag-one autocorrelation and cross correlations of the observed data (Matalas 1967).

From a practical standpoint this basic technique was not improved upon until recently. However, it was recognized that Markov models may not adequately characterize the low frequency effects of natural time series (Burges and Lettenmaier 1977). That is, the low flows synthesized by Markov models do not exhibit the long-term persistence of natural streamflows and many other natural time series including rainfall. Such long-term persistence was the subject of intensive study by Hurst (1951) who found that the rescaled range of natural time series is proportional to the length of the record raised to a power greater than 0.5.

$$R_n/S_n = (n/2)^h \quad ; \; 0.5 \leq h \leq 1$$

where R = range of cumulative departures from the mean
S = standard deviation
n = record length
h = exponent (Hurst coefficient)

The rescaled range can be regarded as a measure of the length of runs for which a time series is above or below some reference level. Analysis of a great number of natural time series gave a mean value for the Hurst coefficient of 0.72 with standard deviation of 0.09. Markov models in contrast give a coefficient which tends asymptotically to 0.5.

The first model which preserved the Hurst coefficient was the Fractional Gaussian Noise Model developed by Mandelbrot and Van Ness (1968) and Mandelbrot and Wallis (1969). Further development of this model led to the computationally more efficient Fast Fractional Gaussian Noise (FFGN) model (Mandelbrot 1971). This is a single-site model which is expensive to run and is unable to preserve certain practical combinations of lag-one correlation and Hurst coefficient (Lettenmaier and Burges 1977).

Comparison of Markov models and the FFGN model using Monte Carlo simulation (Wallis and Matalas 1972, Burges and Lettenmaier 1977) demonstrated that for a given reliability level, the FFGN leads to larger reservoir design storages than the corresponding Markov model. The differences in storage, determined using the two models, become increasingly large as demand levels are increased and as the specified reliability is increased. The practical significance of these findings is, however, in doubt. As will be discussed in the next section, estimation of the Hurst coefficient from the short records available is unreliable, and the justification for using long-term persistence models such as FFGN is open to question. Moreover, although Markov models have a Hurst coefficient which tends asymptotically to 0.5, for short record lengths expected values of the Hurst coefficient have been found in the range of 0.7 (Hipel 1975).

The Fractional Gaussian Noise model implies a process with infinite memory, and for some time infinite memory was put forward as an explanation of the Hurst phenomenon. Since then, two other classes of models have been shown to preserve Hurst coefficients. These are the Broken Line model of Mejia, et al. (1972, 1974) and more significantly, the class of ARMA models (O'Connell 1971). The ARMA models are of particular interest in that they are the subject of an extensive literature (Box and Jenkins 1976) and in that there has been some success in application to multi-site problems (e.g. Ledolter 1978 and Lettenmaier 1980).

A variety of other approaches to data generation have been developed though none have met with universal approbation. Useful reviews of the subject matter are given by Fiering and Jackson (1971) and more recently by Jackson (1975a) and Lawrance and Kottegoda (1977).

2.2 Parameter Estimation and Model Evaluation

The models described in the previous section have been used to maintain the low-order moments (mean, variance, skew), low-order correlations (primarily lag-one) and in some cases the Hurst coefficient of the underlying hydrologic series. The topic of parameter estimation has received a great deal of attention in the literature. There are several well-recognized problems.

First and foremost, the parameters we wish to preserve are not those of the historic record but those of the future flow series. A convenient escape from this problem is to assume that the series is stationary, at least from the start of the historic record to the end of the project's economic life. This historic record is then simply a sample with a typical length of less than 50 years from some underlying stationary series.

The difficulties in parameter estimation are now reduced to the slightly more tractable problems of sampling error. The simplest approach to the problem is to assume that the statistics of the underlying stochastic process are those of the short historic record. However, the statistics of a short sample from a stochastic process are known, in general, to be biased. The cumulative distribution functions of the mean, variance, and skew of independent samples of various lengths and from various underlying distributions have been obtained through Monte Carlo simulation by Wallis et al. (1974). This work allows one to estimate the small sample bias of statistics obtained by sampling from most of the important probability distributions used in hydrology. The small sample bias of the lag-one correlation has also been investigated by Monte Carlo sampling from a Markov model with Gaussian marginal distributions (Wallis and O'Connell 1972).

Correction of the historic statistics for small sample bias has been suggested by a number of researchers. However, recent work by Stedinger (1980) shows that a correction for bias increases the mean square error of estimation of certain parameters. This suggests that use of the biased historic statistics may result in more reliable parameter estimates.

Small sample considerations are also of importance in evaluating synthetic traces. The models discussed in the previous section only maintain the historic statistics in synthetic series of infinite length. For applications in water resources planning, synthetic traces of from 30-50 years in length are of interest. As before, the statistics from these short synthetic traces are, in general, biased. This again suggests that the parameters of the generation scheme should be adjusted to produce unbiased statistics from the synthetic traces.

Within the limits discussed above, the models developed do preserve the moments and correlations they were intended to preserve. Restrictions on the ranges of applicability of some of the models have been reported, however. For example, as noted earlier the computationally useful form of FGN, Fast Fractional Gaussian Noise (FFGN) cannot preserve certain combinations of Hurst coefficient and lag-one correlation. There are similar restrictions on ARMA models (Lettenmaier and Burges 1977).

Another important question raised in the literature is which statistics to preserve for water resources analyses and under what conditions. It is generally agreed that the mean, variance, skewness, and lag-one correlations should be preserved.

However, as indicated earlier, there is continuing controversy as to whether the Hurst coefficient should or should not be preserved for streamflow generation. Burges and Lettenmaier (1977) used FFGN and Markov models to study the effects of parameter uncertainty on the non-failure sizing of storage reservoirs for a variety of demand conditions. Their results indicated that although correct modeling of the Hurst coefficient may be important for high demand levels, for a wide range of practical storage problems, accurate modeling of the Hurst effect is not important.

Klemes et al. (1981) made a preliminary evaluation of the role of the Hurst effect in reservoir design by comparing the differences in expected reservoir reliability achieved using synthetic traces generated by a Markov model and by a Broken Line model. The differences were found to be very small in comparison to the uncertainties that might be expected in the reliability estimates arising from either short historic records (parameter uncertainty) or from uncertainties in model identification. A similar but more extensive investigation by Burges and Lettenmaier (1982) served to confirm these findings. In fact, estimation of the Hurst coefficient from streamflow records of the lengths typically available is inherently unreliable. Wallis and O'Connell (1973) state that "in many regions of the world there is entirely insufficient hydrological data to make a reliable estimate of long-term persistence."

Two schools of thought appear to be arising with regard to modeling long-term persistence. In one school are those who feel that long-term persistence is of such potential import in the behavior or operation of water resources systems that it should be accounted for in analysis even though parameter estimates are uncertain and models maintaining long-term persistence are complex and expensive to use. The other school appears to be of the opinion that because of the modeling and estimation difficulties outlined above, long-term persistence should not be included in data generation but may more appropriately be accounted for in project design by use of some unspecified safety factor.

While these two approaches address the practical difficulties of reservoir design, it should be recognized that the potential importance of long-term persistence has not yet received detailed study in a wide variety of other planning situations or in other fields related to the hydrologic sciences (e.g., regional planning for dry-land agriculture). The more academic problem of explaining the Hurst effect in terms of some causal mechanism (discussed in Section 2.3) also remains unsolved. Thus it seems prudent to at least recognize that some of the current data generation schemes do not exhibit the long-term correlation structure thought to exist in natural time series.

Although emphasis has been placed on preserving the moments and correlation structure of the underlying distribution, it should be recognized that the ultimate purpose of stochastic hydrology is to create streamflow sequences which will preserve those features of the time series controlling the variables of interest (e.g., reservoir reliability). For reservoir design this suggests that stochastic hydrology should place greater emphasis on accurate replication of the low flow features of the time series.

Hirsch (1979) compared reservoir reliabilities predicted using a variety of Markov and ARMA models. One result was that for the time series under study, models correctly preserving the moments of the distributions did not perform as well as those models which only preserved the logs of the moments. The latter group of models apparently better preserved the low flow portion of the cumulative distribution function of the time series.

Model evaluation on the basis of replication of features such as reservoir reliability has received comparatively little attention in the literature, and conclusions regarding the true efficacy of the various models cannot yet be made. However, from Hirsch's work there is an indication that the more popular models currently available may not adequately characterize low-flow features irrespective of their abilities to preserve long-term persistence. It appears that more attention needs to be paid to preserving the full marginal cumulative distribution function of the historic data.

To summarize the present situation, parameter estimation and model identification remain difficult problems for all classes of models. In particular, estimates of the correlation and Hurst coefficients are unreliable for the short instrument records currently available. The necessity for preserving the Hurst coefficient in modeling remains a controversial question, and it may be that more appropriate measures of the correlation structure and long-term persistence should be developed.

2.3 Physical Considerations

With the exception of appeals to infinite memory, the above mentioned models are prescriptive, that is, they attempt to preserve features of the geometry of the time series without considering the physical mechanisms which give rise to the structure of the time series.

Klemes (1974) brought fresh insights to the field by demonstrating that the Hurst phenomenon is not necessarily an indication of infinite memory. In particular, Klemes used numerical experiments to show that the Hurst phenomenon can also be caused by non-stationarity of the mean and by random walks with one absorbing barrier. (The latter case is used as an analog for flow from a semi-infinite storage reservoir.) The idea of non-stationarity of the mean as a cause of the Hurst phenomenon was supported by analysis of long precipitation records from the east coast of the United States (Potter 1976).

Potter mentions (for the first time in the water resources literature that I am aware of) the possibility that long-term persistence in the precipitation records could be related to long-term persistence in atmospheric circulation. Subsequent reanalysis of these data (Potter 1979), however, showed that much of the non-stationarity could be ascribed to non-homogeneities in the record caused by shifts in the location of the recording station. So despite the obvious possibilities, it has thus far not been possible to establish a correspondence between the Hurst phenomenon and non-stationarities in recorded data or indeed to find any other causal mechanism.

Subsequent to Klemes' work, Boes and Salas (1978) proposed a general mixture model for shifting means. They showed that Klemes' shifting mean model was a special case of the more general model and they also demonstrated that Klemes' model and the ARMA(1,1) process have identical correlation structures. The shifting mean models of both Klemes and Boes and Salas are unfortunately quite complex, and it is unlikely that they will ever be used in other than theoretical studies because of the difficulties of model identification and parameter estimation using observed time series.

In attempts to improve the low flow characteristics of synthetic streamflows at a single site, Jackson (1975b) investigated the use of Markov mixture models to control explicitly the lengths of the synthetic low flow periods. While this model is still prescriptive in that it attempts to reproduce the geometry of the time series without appealing to any physical mechanism, it does recognize that low and high flows do not necessarily come from the same distribution. Although this approach may be an improvement from a conceptual viewpoint, it suffers from parameter estimation problems. The simplest model considered by Jackson made use of six parameters in comparison to three parameters for a comparable Thomas-Fiering model.

In a related paper Jackson (1975c) investigated the use of birth-death models to introduce differential persistence (i.e., the observed feature that streamflows are more highly correlated at low flows than at high flows) in the synthetic traces. Jackson's approach was based on a simple phenomenological model displaying differential persistence which related streamflow to groundwater storage. The reasoning used in the model was that in periods of low flows, with a relatively empty aquifer, a larger proportion of rainfall goes into groundwater storage than in high-flow periods with a relatively full aquifer. Groundwater storage thus has a more pronounced buffering effect at low flows than at high flows giving rise to differential persistence.

Although the above models do attempt to introduce some physical basis to synthetic generation, they assume that the correlation structure of the flows arises solely from the dynamics of the groundwater storage or the catchment. Rainfall data are treated simply as serially independent random variables and the possible role of atmospheric processes in determining correlation structure is not addressed.

An alternative approach to introducing the dynamics of the catchment into stochastic hydrology is through the use of rainfall-runoff models. A rainfall-runoff model is any mathematical technique for transforming rainfall over a river catchment to streamflow at points within the catchment. A wide variety of such models exist, varying in complexity from a simple mass balance model such as that used by Jackson (1975c) to complex deterministic conceptual models such as descendents of the Stanford Watershed Model (e.g., Hydrocomp 1976).

The rainfall-runoff model acts as a complex filter and integrator of the rainfall events occurring over the basin. The more complex models purport to represent the physical processes, or properties which control streamflow in the catchment. Thus in principle they should be able to represent accurately the role of the catchment in streamflow generation even under severe conditions. The accuracy of transformation in these models depends to a large extent on the climate of the area under study, and on the adequacy of the input data. For example, in an area with great spatial variation in rainfall, agreement between simulated and recorded flows may be poor. Simulation is also often poor in areas where snowmelt is a major component of the streamflow. Experience in a number of other climatic regimes (e.g., Mediterranean and temperate maritime) shows that simulation can be very accurate, particularly on a seasonal basis. This is especially true for simulation of low-flow conditions even when these conditions are more severe than those experienced during the calibration period (Hydrocomp 1978). The accumulated experience does suggest that rainfall-runoff models can accurately reproduce the low flow characteristics of streamflow though there appears to be little formal quantitative evidence to support this.

Use of a rainfall-runoff model in synthetic flow generation simply involves the generation of synthetic rainfall traces followed by application of the model to complete the rainfall-runoff transformation. Although in principle this process is straightforward, it does require a considerably greater investment of effort than direct streamflow generation. Rainfall-runoff modeling frequently requires considerable input of both personnel and computational resources in data collection and validation, calibration of the model against historic data, verification of the calibration and, of course, in the production runs themselves. The most serious economic argument against the approach is in the additional labor involved. Although computational time is also dramatically increased, the continuing drop in computational costs, especially on mini-computers indicates that this will be a less important consideration in the future.

The rainfall-runoff approach to streamflow generation has been used in a small number of studies thus far. Hydrocomp (1978) used a multi-site, Markov model to generate monthly rainfall at seven sites in the Rio Paranaiba catchment in Minas Gerais, Brazil. The rainfall was transformed to streamflow using a previously calibrated model of the catchment; the resulting streamflow was to be used in reservoir operation studies where the primary concern was in low flow conditions. A comparison of the pro's and con's of direct streamflow generation against rainfall generation followed by a rainfall-runoff transformation is given by Leytham and Franz (1980). This work, based on the previously cited study by Hydrocomp indicated that one of the principal problems in the latter approach is the difficulty of assessing the accuracy of the transformation. Although the conceptual catchment model may be beneficial in introducing some physical aspects of flow generation, it is not clear what magnitude of model errors or rainfall input errors can be allowed before direct streamflow generation becomes the better approach. Another problem not mentioned by Leytham and Franz is the difficulty of preserving long-term persistence in the multi-site rainfall generation. As discussed earlier, this is of potential significance in both reservoir design and operation studies. Severe drought is a manifestation of an anomaly in atmospheric circulation and the structure of the multi-site Markov model (Matalas 1967) may not be able to represent the true temporal or spatial nature of drought.

Another application of rainfall generation followed by a rainfall-runoff transformation is presented by Wilson, et al. (1979). The principal purpose of this work was to investigate the influence that the spatial distribution of storm rainfall has on the outflow hydrograph (specifically the peak flow and flood volume) from a small catchment. The synthetic rainfall was generated using a model developed by Bras and Rodriguez-Iturbe (1976). This model departs radically from those discussed earlier in Section 2.1. It was developed from earlier work by Mejia and Rodriguez-Iturbe (1974b) in which rainfall synthesis is based on the addition of harmonics of random frequencies sampled from the radial spectral density function of the rainfall time series. The model and its subsequent development is designed to synthesize storm events. It considers the direction and speed of storm movement and also permits inclusion of a radially symmetric spatial correlation function appropriate for the class of storm being synthesized, e.g., cyclonic or convective storms. Time between storms and storm durations are assumed to follow exponential distributions.

Although approaches such as that of Bras and Rodriguez-Iturbe have attempted to introduce some simple physical concepts into the generation of synthetic storm rainfall events, and their corresponding storm hydrographs, no parallel work appears to have been done in relation to data generation at larger time intervals for drought events. It is well established that droughts are the results of large scale anomalies in the atmospheric circulation and consideration of such circulation in data generation would appear to be a logical and necessary step toward improving the currently available tools.

3.0 ATMOSPHERIC DYNAMICS IN WATER RESOURCES PLANNING

The review in Chapter 2 has identified a number of deficiencies in current techniques in stochastic hydrology. For drought studies the principal difficulties lie in accurate characterization and synthesis of severe drought. Traditionally, synthetic data generation of either rainfall or streamflow has relied almost exclusively on parameter estimation from the historic rainfall and streamflow instrument records. The parameter estimation problems are exemplified by the large sampling variability of both the lag-one autocorrelation (Wallis and O'Connell 1972) and the Hurst coefficient (Wallis and Matalas 1970). The generation techniques currently employed also suffer from a number of problems. These include the continuing problems of preserving the low-lag correlation and Hurst coefficient in multi-site generation, the apparent inability of models to preserve appropriate low flow characteristics, and the difficulties of ensuring that synthetic low frequency events are plausible.

The stochastic methods currently employed are little more than "black boxes." The techniques are blind to our knowledge of both the processes giving rise to the time series of interest and the known physical limits on the processes. As a simple example, many stochastic methods (e.g., the Thomas-Fiering model) can generate negative rainfall amounts or excessively large rainfall amounts which are physically unreasonable. The question I wish to answer in this research is whether or not our knowledge of the atmospheric processes controlling precipitation can be used to better define the probability distribution of precipitation and so aid in the planning process.

3.1 Related Cross-Disciplinary Research

Researchers in water resources and the atmospheric sciences do not have a strong tradition of co-operation even though it is clear that much work in the atmospheric sciences may have profound implications for water resources planning. Only in one area, the study of severe storms, has there been a reasonable transfer of knowledge leading to the development of hydrometeorology as a specialty field for such problems as spillway design. Other areas of concern in the atmospheric sciences, such as climate reconstruction and theories of atmospheric circulation, may have received recognition by hydrologists and water resources planners, but little has been done to incorporate this knowledge in practice. The apparent assumption inherent in the majority of work by hydrologists is that the historic instrument record contains all necessary information pertaining to the relevant meteorological variables.

Increased use of the atmospheric sciences in water resources planning has recently been advocated in a seminal paper by Kilmartin (1980). Kilmartin drew attention to the anomalous period of instrument record on which many water resources projects are based. He pointed out that the streamflow record over much of the world is rarely more than 60 years in length, and commonly the record is less than 20 years. The last 80-100 years in many parts of the northern hemisphere have been, however, among the warmest and wettest in the past 1000 years with a temperature maximum in about 1940. Much of the available streamflow record may therefore have been affected by what appear to be unusual climatic conditions.

While climate change of this nature is not the concern of this research, the fact that the recent climate may not be representative of the future has serious implications in water resources planning. For stochastic hydrology, the implication is that historic streamflow and rainfall are by themselves inadequate indicators of future conditions. Kilmartin reviewed possible tools and techniques that could be useful in incorporating climatic fluctuations into water resources planning and emphasized the importance of considering the atmospheric processes that might give rise to such fluctuations. In particular he stressed that severe low streamflow events must stem from severe precipitation deficits, which in turn are related to and maintained by "severe, persistent anomalous behavior in regional, if not global, atmospheric circulation."

Kilmartin proceeded to review a number of potential relationships between global circulation patterns and descriptors of climate, such as rainfall, with emphasis on techniques for back-extension of hydrometeorologic records, Of particular interest is Kilmartin's reference to unpublished work in which regression relationships between monthly rainfall in Indonesia and surface pressures at Darwin, Honolulu and Taiwan were used for the back extension of rainfall data and the fill-in of missing data. Rainfall in Indonesia is influenced by the Southern Oscillation which, in the cited study, was characterized by surface pressures at the stations mentioned. Unfortunately, a detailed description of this work is not available.

Kilmartin's paper appears to be the only detailed discussion in the water resources literature of the potential role of atmospheric dynamics in water resources planning. The paper was a review and was, of necessity, qualitative in nature. It should, however, provide considerable stimulus to researchers in water resources as it points out a number of areas of valuable cross-disciplinary research in water resources and the atmospheric sciences.

3.2 Atmospheric Circulation and Drought

Numerous methods for describing the onset and severity of drought are available in the water resources field ranging from indices of soil moisture deficits to measures of inability to meet water demands (Yevjevich, et al. 1978). Many definitions are dependent on the use to which the water resource is put and quite often are area specific; clearly a drought in Arizona is hardly comparable to drought in the Pacific Northwest. For the purpose of this study the term drought will be used loosely. The main aim is to derive better measures for the lower tail of the probability distributions of spatial and temporal rainfall fields. In principle all other measures of drought can be derived from this information, and the difficulty of a general definition of drought is avoided.

The mechanisms and patterns of atmospheric circulation which give rise to drought are area specific. The principal features of the atmospheric circulation in the mid-latitudes are the circum-polar westerly flow and the associated large amplitude synoptic scale disturbances which manifest themselves in the ridges and troughs in the atmospheric pressure fields. The planetary circulation is highly asymmetric with the highest winds and the largest temperature and pressure gradients occurring along a fairly narrow band in the vicinity of the tropospheric jet stream. The path taken by the jet stream varies from day to day and from season to season. The ridges and troughs form a circum-planetary system of waves which generally move in an easterly direction but may remain stationary or may move more slowly in a westerly direction. The wave lengths of the circulation are externally forced by topographic barriers such as the Rockies and by large thermal gradients such as occur along the sea-ice boundaries in the high latitudes.

Synoptic scale disturbances, with which frontal type rain storms are associated, develop as a result of instabilities in the jet stream flow and are steered by the jet stream. Thus the track of the jet stream, which can be inferred from contour maps of geopotential height in the upper atmosphere, gives an indication of the areas most likely to receive frontal rainfall. More significantly from the point of view of this study, the location of ridges of high pressure in the upper atmosphere and at the surface, indicate areas around which the major frontal systems are steered. A basic introduction to atmospheric dynamics is given in texts such as Wallace and Hobbs (1977) and Holton (1979).

A number of observational studies have been made relating precipitation patterns to the position of the jet stream. In one of the earliest such studies, Starrett (1949) analysed precipitation patterns over the United States relative to the position of the geostrophic west wind maxima at 300 mb. Starrett's analysis, covering the period October 1946 to May 1947, demonstrated that the position of the precipitation maxima tends to coincide with the position of the 300 mb wind speed maxima. This work also indicates that the influence of the jet stream on precipitation extends at least 5 degrees latitude (about 500 km) either side of the wind speed maxima. This figure is in reasonable agreement with both the width of the core of greatest wind speeds in the jet stream and with the scale of typical frontal storm systems.

Since the 1940's understanding of both the jet stream and its relationship to synoptic-scale disturbances has greatly improved. An excellent description of the structure and development of synoptic scale disturbances is given by Palmen and Newton (1969). Observations of mid-latitude cyclonic storms show that they generally develop as frontal waves on the equatorward side of the jet stream later crossing under the jet stream to the poleward side. The fully developed disturbances reaching the west coast of the U.S.A. from the Pacific are thus generally slightly poleward of the jet· stream.

The relationship between regional patterns of precipitation and atmospheric circulation is well illustrated by a comparison of conditions on the west coast of the United States for the months December 1960 (Gelhard 1961) and February 1961 (Stark 1961). Figure 3.1 shows the mean 700 mb height contours over the western hemisphere north of

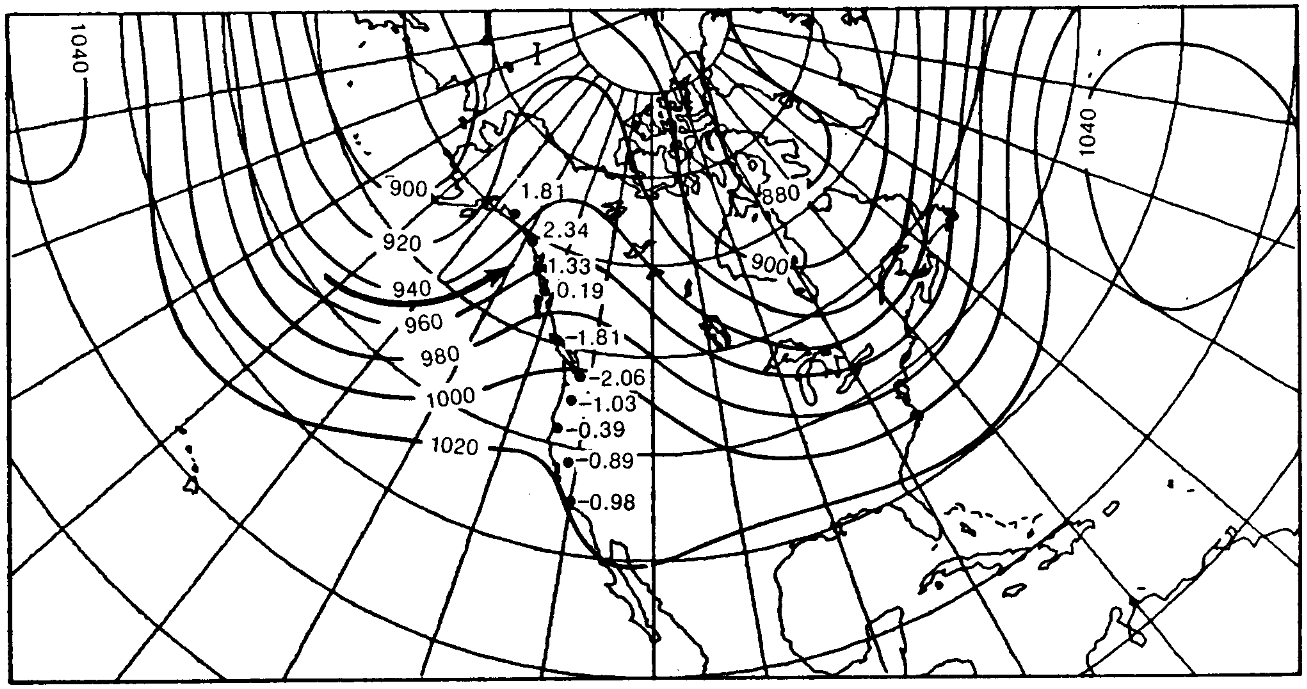
20 degrees north and the departures from normal of precipitation at selected sites along the west coast of the United States for December 1960. Figure 3.2 shows the comparable conditions for February 1961.

Although the jet stream cannot be identified at the 700 mb level, December 1960 (Figure 3.1) saw the development of a strong ridge of high pressure along the west coast with meridional flow and a persistent northerly jet stream track steering cyclonic storms into the Gulf of Alaska. As a consequence precipitation along the south coast of Alaska was substantially above normal whereas the Pacific Northwest was much drier than normal.

In contrast to December 1960, February 1961 (Figure 3.2) had strong zonal flow with a persistent jet stream track somewhat to the south of its mean position. This resulted in a succession of frontal systems crossing the Pacific Northwest bringing near record precipitation to the area. Precipitation in Alaska and California remained near normal.

The climate and weather of the mid-latitudes is thus associated to a large extent with the position of the jet stream track and with the position, movement and amplitude of the troughs and ridges in the atmospheric pressure field. Anomalous conditions in the atmospheric circulation are reflected in anomalous climatic or weather conditions. It is the persistence in departures from the normal pattern of circulation which are the primary cause of drought and which are thus of primary interest in water resource planning.

For a detailed study of the relationships between circulation and precipitation, I will concentrate on conditions along the west coast of North America with particular emphasis on drought in the Pacific Northwest, i.e., from northern Oregon to southwest Alaska. This region was chosen for a number of reasons. Firstly, from a technical point of view, the climate of the area and its relation to atmospheric



——960—— Mean 700 mb height countours (tens of feet)
● 1.02 Standardized precipitation
→ Principal storm track

Figure 3.1 Mean 700 mb height contours and departures of precipitation from monthly normal for December 1960 (adapted from Gelhard 1961)

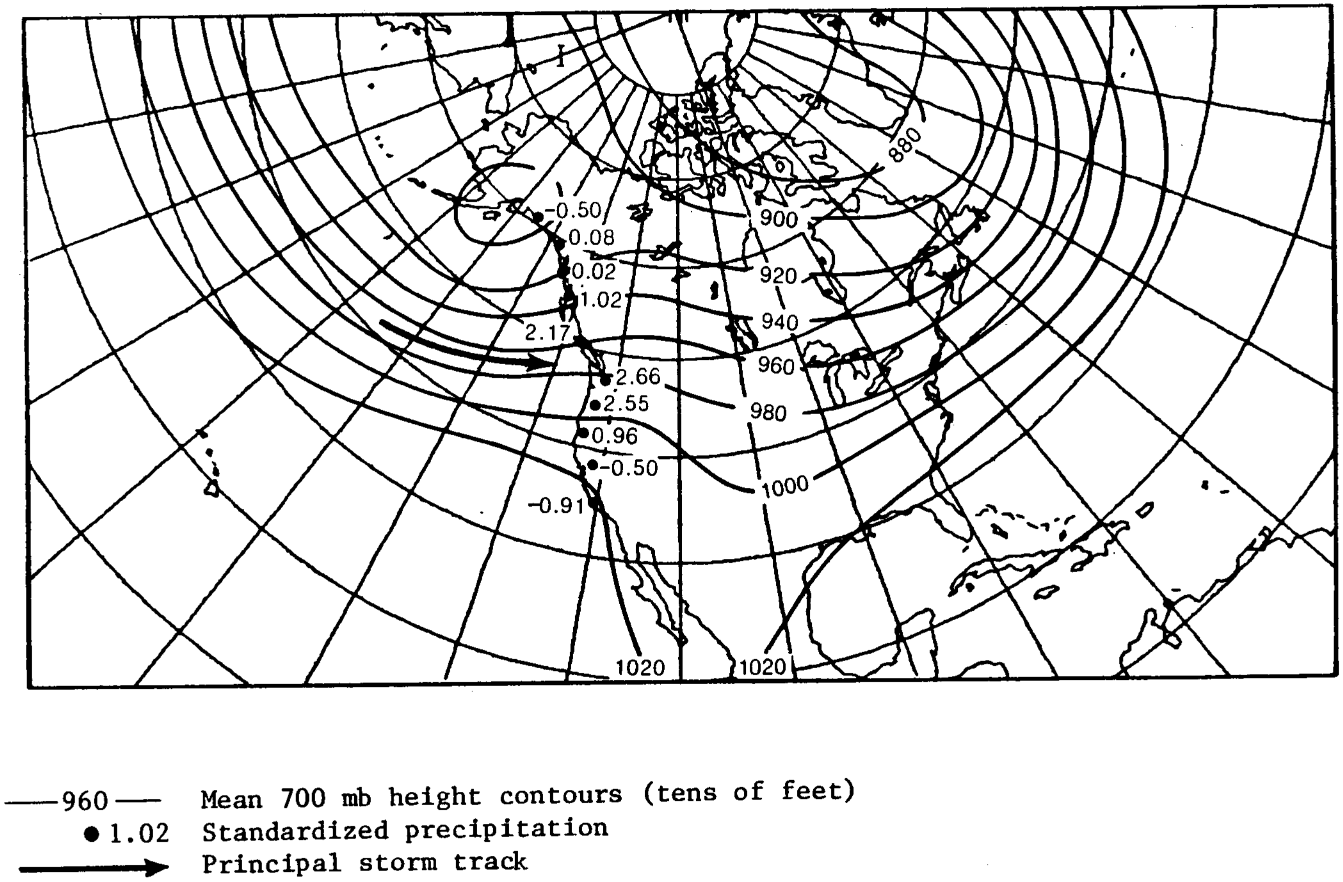


Figure 3.2 Mean 700 mb height contours and departures of precipitation from monthly normal for February 1961 (adapted from Stark 1961)

circulation is well understood in a qualitative sense. The climate is also relatively homogeneous, i.e., probably more than 95 percent of the precipitation in the region is associated with frontal systems, and the flow is predominately westerly from a single source of moisture. Finally, relatively high quality precipitation data are available in the area. This is particularly true of southern British Columbia where a number of long (90 year), high quality rainfall records are available.

In considering drought on the west coast, it is instructive to examine historic events. Of particular interest are analyses of recent drought years by Namias (1978a, 1978b) and Edmon (1980). Namias' principal interest in his papers is in attempting to relate anomalies in sea-surface temperatures to climatic anomalies. Although complex global interactions between ocean and climate probably exist, findings thus far are controversial and must be regarded as tentative. However, Namias' papers demonstrate the qualitative relationships between drought and atmospheric circulation. Namias (1978b) analyzes the 700 mb geopotential height field for the winter of 1975-1976 and shows the expected relationship between a high pressure ridge over the California coast and substantial rainfall deficits in California. The winter 1975-1976 drought extended from about Los Angeles to Southern Oregon. Rainfall in Washington, British Columbia, and southern Alaska were somewhat above normal indicating a persistent northerly jet stream track.

Edmon (1980) undertakes an extensive analysis and comparison of the structure of climatic anomalies over the northern hemisphere during the winters 1976-1977 and 1977-1978. On the west coast of the U.S.A. the precipitation patterns during the two winters differed markedly. During 1976-1977, drought extended from southern California to southern British Columbia, and wetter and warmer than normal conditions prevailed over northern British Columbia and southern Alaska. In comparison, the winter of 1977-1978 had unusually heavy rainfall in

California and extremely dry conditions from southern Washington to the Gulf of Alaska. Edmon again shows the relationship between the conditions described above and the track of the jet stream. During the winter of 1976-1977 the jet stream persistently followed a northerly track causing drought in the U.S.A. and during the winter of 1977-1978, it followed a southerly track. Thus as pointed out earlier, persistent deviations of the jet stream from its normal position and the dynamics of high pressure blocks are crucial in determining climatic fluctuations on the west coast.

Meteorologists have recognized for some time that certain patterns of weather tend to exhibit persistence or tend to recur. This has led to the development of classifications of weather types such as that presented for North America by Elliott (1951). The reasons for the establishment and persistence of particular circulation patterns are of course of great interest since the ability to forecast circulation for some months ahead, even qualitatively, would be of considerable value in long-range weather forecasting.

The high pressure blocks that recur over the central and east north Pacific and north Atlantic Oceans are of particular interest in this respect. Rex (1950, 1951) undertook an extensive study of blocking situations over the Atlantic and Pacific using approximately 14 years of data. For blocks over the Atlantic, Rex gave estimates for the duration of blocking action, for seasonal variation in blocking activity and for the location of blocks. The analyses showed an average duration of a block as about 17 days with an extreme duration of 41 days. These figures are however, highly dependent on the exact definition used for a block. They do not necessarily give a good indication of the duration of dry periods particularly since Rex did not consider the time interval between blocks. Rex compared precipitation patterns over Europe during periods of blocking with precipitation patterns obtained for periods of strong zonal flow, and as we would expect from the earlier discussion, showed rainfall substantially below normal downstream of a block and substantially above normal in regions subject to strong zonal flow.

A similar study of blocking activity over the cental north Pacific has been carried out by White and Clark (1975) using 20 years of monthly mean 700 mb data. They showed blocking activity with durations in extreme cases of about three months but again did not explicitly consider the interval between blocks.

Various mechanisms have been suggested to explain the initiation and maintenance of blocks. For example Rex (1950) attempted to use a mechanism similar to a hydraulic jump as an analogy in explaining the development of blocks. In a more general sense several decades of work in the atmospheric sciences (e.g. Namias 1975) show complex global interactions between sea surface temperatures and planetary circulation, but the true cause/effect relationships between sea surface temperatures and features such as blocks are still unknown, and the present prognosis for accurate long-range forecasting is poor (Newell 1979).

3.3 Atmospheric Dynamics and Stochastic Hydrology

The qualitative relationships between atmospheric circulation and patterns of precipitation discussed in Section 3.2 have a number of implications for stochastic hydrology, particularly in terms of characterizing inter-station precipitation relationships. As noted earlier drought over a particular region implies a persistent jet stream track steering frontal systems around the area affected by drought and bringing unusually heavy precipitation to some other distant region. Thus physical considerations indicate negative cross correlations at large distance, at least for the severe events of most interest in water resources planning.

The inter-station relationships are complicated both by the time scales of interest and by the extent of the region under study. It is clear that drought is a much larger scale phenomenon than the frontal systems which bring rain to the west coast. It is also clear that drought is a much more persistent phenomenon than the transitory frontal systems. We should therefore expect inter-station cross correlations to be higher during periods of drought than during normal or wet periods. This is perhaps obvious for 5-day precipitation depths at stations 100 km apart, for example; but it is not obvious for the monthly or annual intervals of principal interest in stochastic hydrology.

Consideration of the qualitative precipitation/circulation relationsips may also provide a subjective tool for assessing the reasonableness of synthetic multi-site precipitation sequences. We must be able to assign a plausible circulation pattern to any synthetic sequence. If this is not possible, then there seems to be no justification for using such a sequence for design purposes.

The reasons for the onset and maintenance of the anomalous atmospheric conditions associated with drought are clearly beyond the scope of this research. The concept of multiple equilibria states, however, currently under investigation in the atmospheric sciences provides a potentially attractive framework for incorporating consideration of atmospheric dynamics into water resources planning.

In attempts to explain the observed persistence of weather patterns, such as droughts, meteorologists have suggested that the atmosphere may exist in one of a number of possible equilibrium or quasi-stationary states. Numerical experiments by Charney and DeVore (1979) show that topographic forcing (i.e. interactions between atmospheric circulation and the earth's topography) may produce two possible equilibrium states, one a predominantly meridional flow with a strong wave component and a relatively weaker zonal component; the other a predominantly zonal flow with a weak wave component. It is hypothesized that the atmosphere exists in one of its possible multiple-equilibria states until large enough changes occur in external forcing (e.g., thermal gradients) to flip the circulation into another state.

The previous discussions have already shown that zonal flow is associated with wet conditions and meridional flow with dry conditions. There is thus an indication that rainfall over the Pacific Northwest may be represented as coming from two distributions conditioned on the state of the atmospheric circulation; wet conditions prevailing with zonal flow and dry conditions with meridional flow. This scenario together with the conjectured nonlinearities in the cross correlation structure therefore suggests that point and regional rainfall may be modeled better by sampling from mixed distributions.

The following chapters attempt to put the qualitative arguments of this section on a more quantitative basis. In the next chapter, the cross correlation structure of precipitation along the west coast is investigated in detail and in Chapter 5 quantitative relationships between atmospheric circulation and precipitation are explored. The use and characteristics of mixture models are discussed in Chapters 6 and 7.

4.0 ANALYSIS OF MULTI-SITE PRECIPITATION DATA AND EVALUATION OF CURRENT MULTI-SITE STOCHASTIC MODELS

One of the primary motivations for this work is my belief that the current generation of stochastic models is unable to represent the true spatial nature of drought especially over large regions. These difficulties can be attributed directly to the inadequacy of the linear cross correlation as a measure of inter-station relationships.

An unquestioning belief in the value of the cross correlation coefficient for expressing inter-station relationships seems to have developed in the water resources field. Whereas considerable effort has been expended in studying the autocorrelation function for single hydrologic time series, particularly in relation to the Hurst phenomenon, little work has been done in assessing the adequacy of the traditional cross correlation coefficients for annual or monthly time series where inter-station distances are large (in excess of 200 km).

It is fairly well established that the cross correlation of rainfall depths in convective and cyclonic storms can be represented by some kind of symmetric function which decays exponentially with distance from the storm center (e.g. Eagleson 1967). Similar assumptions have been made for annual and monthly rainfall depths. An extensive analysis of inter-station correlations for annual rainfall has been made by Caffey (1965) for stations in the U.S.A and Canada up to 2000 km apart. Caffey's analyses show the cross correlation decaying to zero with distance and "fluctuating randomly about zero beyond some distance." The impression given in this and other work is that negative correlations have no physical meaning in rainfall data; they are simply an outcome of sampling variability. Indeed there is a tendency, based on sample size considerations, to regard negative or small positive (<0.3) correlations as indicative of independence. Certainly the statistical tests available indicate that such correlations are generally not significantly different from zero. For example, with a sample size of 100 from a bivariate normal distribution, a correlation coefficient of ±0.3 is not significantly different from zero at the 10 percent level. However, as will be shown in this chapter, small or negative correlations may be indicative of a more complicated nonlinear structure which may be related qualitatively to features of the large-scale atmospheric circulation.

4.1 Analysis of Multi-site Annual and Monthly Precipitation Data

Monthly and annual precipitation data from seventeen stations along the west coast from southern California to the Gulf of Alaska were obtained for analysis. The stations are listed in Table 4.1 and their locations are shown on Figures 4.1 and 4.2. In all cases a record length of 32 years was used from October 1947 through September 1978. This period was chosen to cover a whole number of water years and to coincide with the available 500 mb geopotential height data used in Chapter 5.

Basic statistics for the monthly data from selected stations are shown in Table 4.2 and the monthly cross correlation coefficients are shown in Table 4.3. The variation of cross correlation with distance from selected base stations is shown for January data in Figure 4.3. Basic statistics for the annual (water year) data are shown in Table 4.4 and the annual cross correlation coefficients are shown in Table 4.5. The variation of annual cross correlation with distance is shown in Figure 4.4 for selected stations.

The cross correlations at both monthly and annual time intervals show a consistent pattern of variation with distance as typified by Figures 4.3 and 4.4. The most notable feature of these figures is the consistent negative correlation at large distances. Negative correlations along the west coast may be explained by considering the large scale circulation patterns discussed in the previous chapter.

Table 4.1 Precipitation Stations Used in Multi-site Analyses

Station Number*	Station Name	Latitude °N	Longitude °W	Months with One or More Days Missing**
503665	Homer WSO	59 38	151 30	Mar-Apr 1973
502177	Cordova FAA	60 30	145 30	None
509941	Yakutat	59 31	139 40	Oct-Dec 1947
508503	Sitka Magnetic Obs	57 03	135 20	Apr 1978
500352	Annette WSO	55 02	131 34	Oct-Dec 1947
1026270	Port Hardy A	50 41	127 22	None
1108447	Vancouver Int A	49 11	123 07	None
1018610	Victoria Gonzales Hts	48 25	123 22	None
457507	Sedro Wooley	48 30	122 14	Nov 1964, Feb 1976
457773	Snoqualmie Falls	47 33	121 51	Nov 1976
456894	Rainier Longmire	46 45	121 49	Dec 1972, Jul 1978
454201	Kid Valley	46 22	122 37	Jan 1950, Jun-Jul 1975
451276	Centralia	46 43	122 57	None
352709	Eugene WSO	44 07	123 13	None
042910	Eureka WSO	40 48	124 10	None
042294	Davis Exp Farm	38 32	121 46	None
047905	Santa Barbara FAA	34 26	119 50	None

*Six digit stations numbers correspond to U.S. National Weather Service stations. The seven digit numbers correspond to stations operated by Environment Canada.

**Refers to data missing in the period Oct 1947 to Sep 1978.

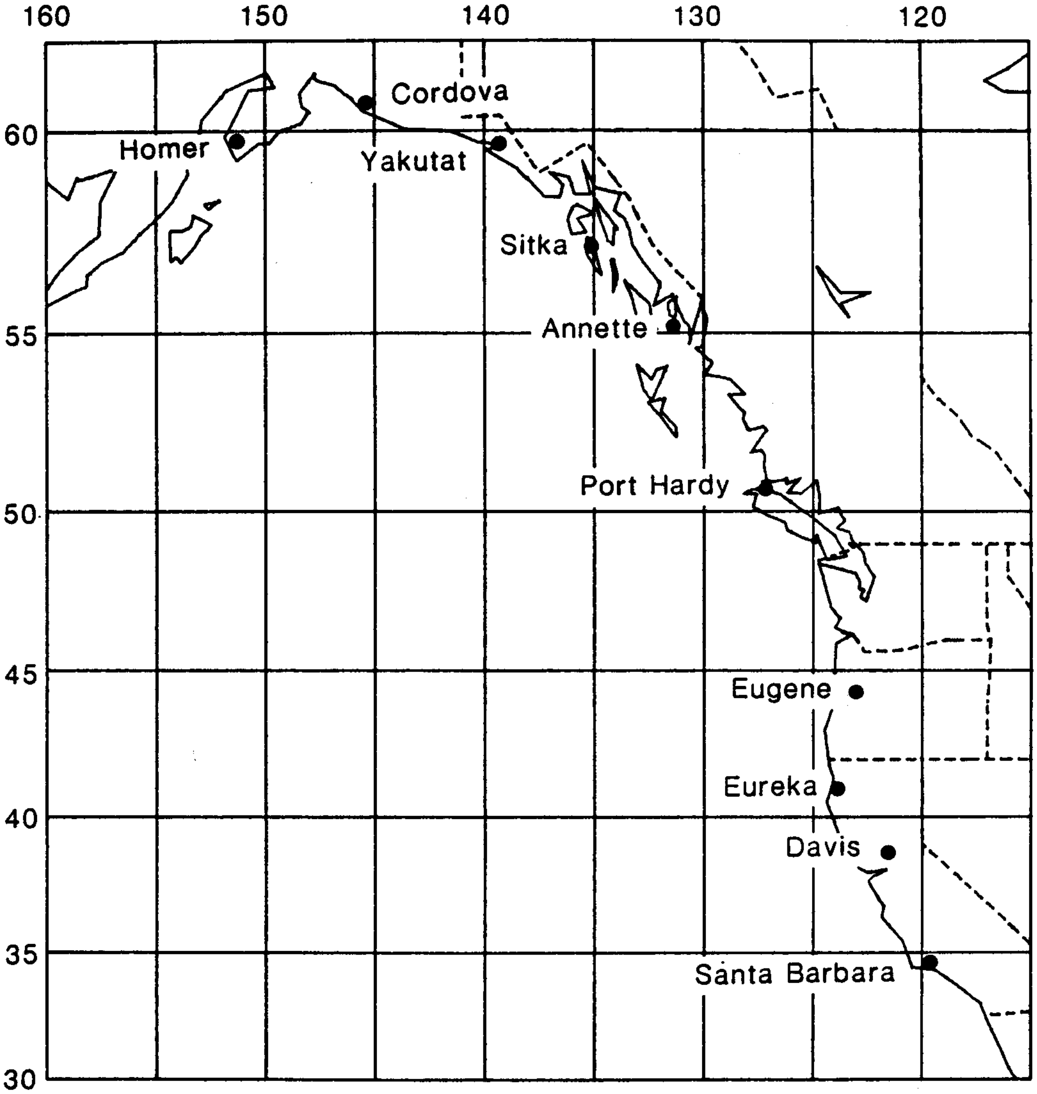


Figure 4.1 Location of precipitation stations (excluding Washington and Southern British Columbia)

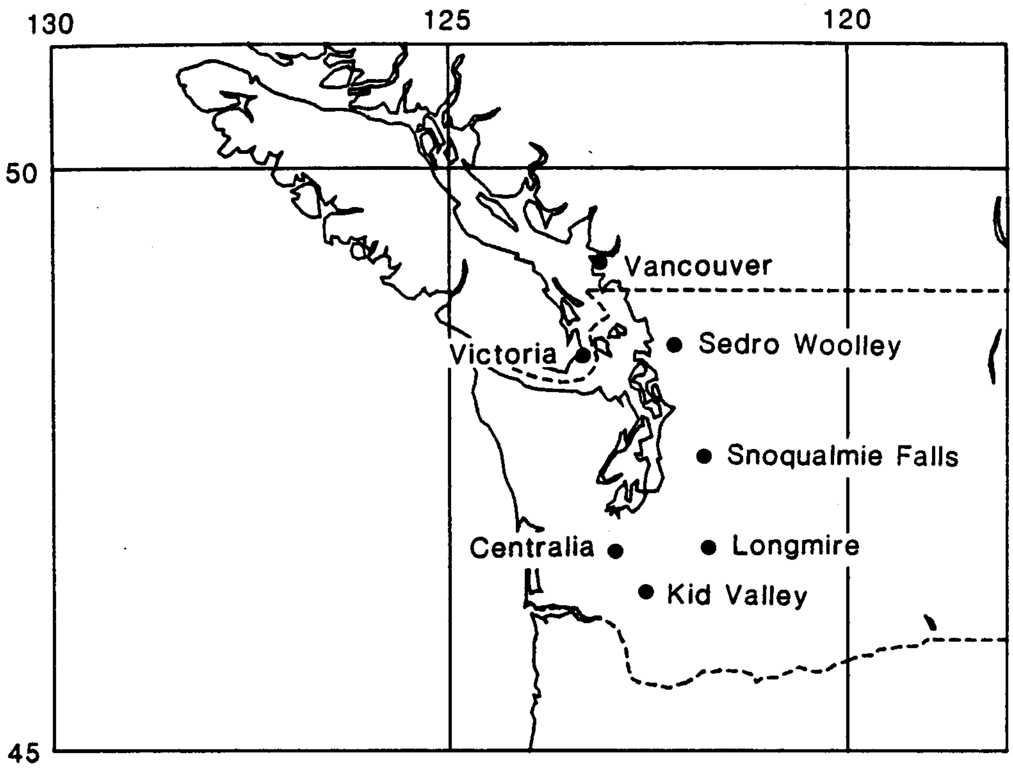


Figure 4.2 Location of precipitation stations (Washington and Southern British Columbia)

Table 4.2 Basic Statistics for Monthly Precipitation Data (in mm)
(1947-1978)

SITE HOMER 503665

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	82.3	64.5	58.9	39.9	44.2	31.8	31.2	25.5	26.7	39.0	58.8	73.1
STDEV	37.1	48.8	45.4	27.8	33.8	21.2	18.3	16.5	19.2	23.6	28.8	31.7
SKEW	1.72	1.31	1.08	1.10	1.12	1.23	.96	.46	.78	.51	.68	.52
MAXIMUM	217.2	218.2	199.4	111.8	142.7	102.6	88.6	58.4	70.1	96.3	133.4	136.9
MINIMUM	35.6	2.0	3.0	9.1	4.1	5.3	0.0	2.0	2.3	4.1	11.9	21.1

SITE CORDOVA 502177

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	311.9	219.6	182.8	137.9	156.8	135.6	141.0	150.2	126.8	166.3	198.4	348.7
STDEV	125.6	160.7	90.4	87.1	109.6	80.8	75.6	69.1	63.1	94.4	94.8	158.5
SKEW	.92	1.33	1.13	.84	1.61	.72	.16	.76	-.09	1.57	.87	.47
MAXIMUM	676.4	777.0	474.2	360.9	481.8	315.2	309.1	344.4	245.9	451.6	465.1	704.1
MINIMUM	145.5	37.1	47.5	16.3	27.2	20.3	2.0	48.5	17.5	43.4	65.3	109.7

SITE YAKUTAT 509941

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	485.9	392.5	318.7	249.7	241.9	230.2	209.8	227.9	139.9	204.8	248.6	409.5
STDEV	183.3	239.6	129.9	130.3	149.5	105.4	90.3	112.6	95.3	120.4	124.9	129.1
SKEW	.92	1.55	1.14	.63	1.85	.47	.52	.54	1.35	1.02	.52	.77
MAXIMUM	937.0	1114.6	708.4	609.1	814.1	479.0	485.6	481.3	412.5	545.8	531.6	742.4
MINIMUM	169.7	110.7	146.8	40.4	61.5	52.3	19.1	69.3	17.3	43.2	72.4	229.9

Table 4.2 Continued

SITE SITKA 508503

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	358.6	279.4	250.2	213.6	182.1	177.8	147.6	124.2	94.4	117.2	174.4	294.2
STDEV	121.8	110.3	92.7	102.3	86.2	72.5	59.0	57.4	51.5	62.4	87.3	119.6
SKEW	.45	.28	.61	.37	.72	.47	.27	.70	.92	1.30	1.06	.84
MAXIMUM	688.6	579.6	461.8	398.8	417.6	345.9	272.8	256.5	218.4	304.5	463.0	592.3
MINIMUM	156.5	68.6	125.2	65.0	47.5	35.8	34.5	35.1	27.4	47.2	61.2	115.8

SITE ANNETTE 500352

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	449.8	322.5	317.1	257.3	252.3	223.0	225.5	163.7	130.1	125.7	190.6	257.9
STDEV	159.4	149.1	133.2	124.1	96.6	125.9	93.0	84.8	56.4	60.7	113.5	97.8
SKEW	.70	.87	.69	.35	.05	1.52	.87	.61	.34	.69	1.10	-.02
MAXIMUM	885.7	713.5	734.1	525.5	458.7	598.7	542.3	372.9	264.2	275.6	526.3	430.8
MINIMUM	224.0	91.7	111.3	19.6	18.5	76.2	31.2	39.9	24.1	14.2	18.0	64.0

SITE PORT HARDY 1026270

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	253.7	246.5	268.3	212.7	159.6	139.3	105.9	69.6	68.5	53.4	74.4	125.2
STDEV	97.7	100.2	84.0	98.0	62.7	52.8	46.3	35.2	37.6	34.2	44.6	57.2
SKEW	.59	1.31	.46	.27	.82	.00	.03	.61	.54	.74	.49	.38
MAXIMUM	486.7	573.5	441.0	426.4	339.7	264.4	189.2	159.9	170.3	125.8	167.0	260.4
MINIMUM	84.7	125.1	110.8	63.4	45.6	32.1	23.6	14.3	8.9	2.6	14.8	25.1

Table 4.2 Continued

SITE VICTORIA 1018610

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	70.5	96.9	119.4	112.4	79.4	50.2	32.1	20.1	18.8	13.9	22.5	32.1
STDEV	40.0	46.3	43.7	54.4	40.6	27.2	16.7	13.7	12.8	10.7	17.1	20.0
SKEW	1.19	.45	.92	1.07	.54	.44	.76	2.28	.73	.68	.77	.57
MAXIMUM	200.6	187.5	247.1	293.0	180.0	120.5	77.6	78.3	50.5	35.9	65.7	77.5
MINIMUM	22.9	16.2	41.1	20.4	8.1	9.0	2.4	2.1	1.3	0.0	.3	5.1

SITE VANCOUVER 1108447

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	121.0	146.8	179.5	152.1	117.5	102.1	60.3	52.3	43.5	31.9	42.4	62.2
STDEV	61.9	55.6	40.3	57.6	49.2	38.6	27.4	26.4	24.0	22.5	28.2	37.2
SKEW	1.10	.17	.80	-.33	1.07	.52	.42	1.08	.66	.65	.48	.56
MAXIMUM	287.6	282.0	300.2	260.7	277.2	186.5	119.5	120.8	114.2	81.3	104.4	143.6
MINIMUM	35.2	33.3	102.2	18.3	48.2	42.3	13.0	10.6	4.7	0.0	2.8	.3

SITE SNOQ. FALLS 457773

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	150.6	216.0	240.2	231.0	168.4	154.8	115.5	81.5	76.7	38.0	48.1	81.4
STDEV	73.7	80.0	59.4	102.0	70.6	57.5	40.6	36.9	45.4	25.2	35.1	46.9
SKEW	.64	-.34	-.29	.28	.78	.48	-.37	.39	.84	.57	.70	.62
MAXIMUM	332.0	361.2	357.1	496.1	325.9	288.5	191.5	167.9	220.2	97.0	126.0	197.9
MINIMUM	39.9	33.8	118.1	50.5	77.5	25.4	20.3	22.9	7.4	0.0	.3	1.5

Table 4.2 Continued

SITE CENTRALIA 451276

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	114.9	170.2	196.9	186.3	129.9	119.8	76.0	50.0	44.5	22.7	37.1	54.1
STDEV	52.1	68.2	62.3	86.7	52.5	47.1	32.2	27.1	23.4	18.6	31.5	32.1
SKEW	.69	-.10	-.21	-.10	.60	-.04	-.23	.95	.23	.86	.74	1.20
MAXIMUM	248.2	297.4	317.0	352.8	269.5	210.6	135.9	119.4	89.4	68.1	105.4	170.9
MINIMUM	24.6	31.5	64.8	25.9	31.2	8.4	9.1	4.8	4.3	.8	0.0	.3

SITE EUGENE 352709

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	99.6	171.9	203.2	212.7	130.5	130.5	64.9	49.6	32.4	8.3	21.2	36.2
STDEV	68.3	96.7	117.7	101.8	64.2	66.0	37.8	29.3	26.9	13.7	28.1	22.1
SKEW	1.44	1.14	1.01	-.33	.56	.69	.88	.43	1.56	2.71	2.87	.46
MAXIMUM	321.6	520.2	533.1	376.7	294.1	316.5	166.6	112.8	120.9	66.8	147.1	87.6
MINIMUM	15.7	30.5	31.5	28.2	21.8	20.1	13.2	7.4	0.0	0.0	0.0	0.0

SITE EUREKA 042910

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	80.1	140.7	157.3	182.8	128.8	133.6	71.4	39.8	15.8	3.5	8.8	22.9
STDEV	65.1	85.2	72.7	91.7	65.3	59.0	54.3	36.3	16.0	6.3	14.2	20.7
SKEW	1.88	1.19	.23	.23	.32	.17	1.61	1.83	1.45	3.11	1.96	1.24
MAXIMUM	331.2	421.1	301.5	353.6	274.3	272.5	271.3	153.7	65.3	30.5	50.3	85.1
MINIMUM	7.1	7.1	13.2	41.4	30.5	31.2	7.9	.8	0.0	0.0	0.0	0.0

Table 4.2 Continued

SITE DAVIS 042294

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	27.9	53.7	80.3	94.5	64.8	52.7	31.2	10.6	3.7	.7	1.2	5.3
STDEV	38.2	47.0	60.7	69.1	56.9	35.8	30.2	14.4	5.6	2.8	3.2	9.9
SKEW	2.92	.95	1.42	.67	1.13	.46	1.08	2.11	1.84	4.87	3.16	2.59
MAXIMUM	201.4	174.5	301.5	246.1	230.6	120.7	104.4	64.3	24.1	16.0	13.0	46.5
MINIMUM	0.0	0.0	4.3	5.8	.5	2.0	0.0	0.0	0.0	0.0	0.0	0.0

SITE SANTA BARBARA 047905

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	9.3	49.9	60.1	86.3	77.4	58.5	33.3	6.7	.8	.8	.5	5.9
STDEV	13.3	49.9	51.7	75.2	84.8	58.4	36.6	12.0	1.6	4.2	1.4	19.2
SKEW	2.17	.91	.38	1.16	1.47	1.48	1.33	1.99	3.14	5.36	3.47	4.39
MAXIMUM	61.0	175.8	148.8	311.2	345.9	256.0	146.1	46.5	7.9	23.9	6.6	104.9
MINIMUM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 4.3 Cross Correlations for Monthly Precipitation Data
(1947-1978)

MONTHLY CORRELATION MATRIX OCT

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.40	.03	-.09	-.30	-.24	-.42	-.36	-.31	-.28	-.10	-.14	.06	-.01
CORDOVA	.40	1.00	.41	.34	-.18	-.39	-.65	-.65	-.70	-.71	-.59	-.45	-.16	-.05
YAKUTAT	.03	.41	1.00	.52	.39	.18	-.30	-.16	-.21	-.32	-.38	-.27	-.14	-.12
SITKA	-.09	.34	.52	1.00	.37	-.13	-.29	-.24	-.31	-.33	-.56	-.47	-.26	-.24
ANNETTE	-.30	-.18	.39	.37	1.00	.31	.11	.02	.11	.10	-.12	-.35	-.27	-.30
PORT HARDY	-.24	-.39	.18	-.13	.31	1.00	.53	.64	.43	.50	.31	.09	-.09	.03
VICTORIA	-.42	-.65	-.30	-.29	.11	.53	1.00	.81	.78	.85	.60	.45	.01	-.08
VANCOUVER	-.36	-.65	-.16	-.24	.02	.64	.81	1.00	.78	.77	.53	.39	.07	-.02
SNOQ. FALLS	-.31	-.70	-.21	-.31	.11	.43	.78	.78	1.00	.83	.73	.54	.11	-.10
CENTRALIA	-.28	-.71	-.32	-.33	.10	.50	.85	.77	.83	1.00	.75	.55	.20	-.08
EUGENE	-.10	-.59	-.38	-.56	-.12	.31	.60	.53	.73	.75	1.00	.77	.37	.04
EUREKA	-.14	-.45	-.27	-.47	-.35	.09	.45	.39	.54	.55	.77	1.00	.51	.15
DAVIS	.06	-.16	-.14	-.26	-.27	-.09	.01	.07	.11	.20	.37	.51	1.00	.15
SANTA BARBARA	-.01	-.05	-.12	-.24	-.30	.03	-.08	-.02	-.10	-.08	.04	.15	.15	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX NOV

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.75	.63	.47	.33	-.31	-.44	-.50	-.55	-.53	-.50	-.33	-.38	-.09
CORDOVA	.75	1.00	.82	.55	.12	-.24	-.47	-.46	-.60	-.65	-.61	-.42	-.42	-.20
YAKUTAT	.63	.82	1.00	.57	.27	-.14	-.40	-.47	-.41	-.51	-.64	-.53	-.49	-.27
SITKA	.47	.55	.57	1.00	.45	.01	-.22	-.17	-.23	-.33	-.56	-.52	-.60	-.31
ANNETTE	.33	.12	.27	.45	1.00	.15	.14	.07	-.09	-.07	-.46	-.46	-.61	-.46
PORT HARDY	-.31	-.24	-.14	.01	.15	1.00	.52	.50	.23	.25	.03	-.09	-.35	-.44
VICTORIA	-.44	-.47	-.40	-.22	.14	.52	1.00	.78	.73	.77	.28	.02	-.18	-.37
VANCOUVER	-.50	-.46	-.47	-.17	.07	.50	.78	1.00	.62	.62	.37	.27	.00	-.27
SNOQ. FALLS	-.55	-.60	-.41	-.23	-.09	.23	.73	.62	1.00	.91	.55	.31	.03	-.28
CENTRALIA	-.53	-.65	-.51	-.33	-.07	.25	.77	.62	.91	1.00	.65	.43	.13	-.18
EUGENE	-.50	-.61	-.64	-.56	-.46	.03	.28	.37	.55	.65	1.00	.80	.53	.11
EUREKA	-.33	-.42	-.53	-.52	-.46	-.09	.02	.27	.31	.43	.80	1.00	.64	.27
DAVIS	-.38	-.42	-.49	-.60	-.61	-.35	-.18	.00	.03	.13	.53	.64	1.00	.60
SANTA BARBARA	-.09	-.20	-.27	-.31	-.46	-.44	-.37	-.27	-.28	-.18	.11	.27	.60	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX DEC

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.82	.61	.14	.32	.07	-.21	-.24	-.22	-.31	-.23	-.07	-.02	-.22
CORDOVA	.82	1.00	.56	.11	.23	-.03	-.27	-.22	-.00	-.14	-.02	.01	.07	-.18
YAKUTAT	.61	.56	1.00	.48	.52	.16	-.12	.03	-.44	-.53	-.34	-.28	-.39	-.35
SITKA	.14	.11	.48	1.00	.76	.36	.16	-.24	-.32	-.45	-.49	-.34	-.49	-.36
ANNETTE	.32	.23	.52	.76	1.00	.50	-.01	-.09	-.24	-.50	-.50	-.44	-.51	-.50
PORT HARDY	.07	-.03	.16	.36	.50	1.00	.08	-.02	-.26	-.41	-.45	-.36	-.34	-.28
VICTORIA	-.21	-.27	-.12	.16	-.01	.08	1.00	.41	.40	.22	-.10	-.14	-.32	-.14
VANCOUVER	-.24	-.22	.03	-.24	-.09	-.02	.41	1.00	.40	.27	.05	-.06	-.29	-.22
SNOQ. FALLS	-.22	-.00	-.44	-.32	-.24	-.26	.40	.40	1.00	.74	.44	.20	.19	.18
CENTRALIA	-.31	-.14	-.53	-.45	-.50	-.41	.22	.27	.74	1.00	.74	.60	.50	.43
EUGENE	-.23	-.02	-.34	-.49	-.50	-.45	-.10	.05	.44	.74	1.00	.70	.62	.54
EUREKA	-.07	.01	-.28	-.34	-.44	-.36	-.14	-.06	.20	.60	.70	1.00	.67	.53
DAVIS	-.02	.07	-.39	-.49	-.51	-.34	-.32	-.29	.19	.50	.62	.67	1.00	.62
SANTA BARBARA	-.22	-.18	-.35	-.36	-.50	-.28	-.14	-.22	.18	.43	.54	.53	.62	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX JAN

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.82	.56	.39	.34	-.35	-.36	-.17	-.44	-.51	-.54	-.56	-.31	-.25
CORDOVA	.82	1.00	.79	.58	.39	-.35	-.47	-.39	-.58	-.58	-.63	-.57	-.34	-.25
YAKUTAT	.56	.79	1.00	.77	.50	-.14	-.45	-.36	-.50	-.47	-.58	-.65	-.37	-.36
SITKA	.39	.58	.77	1.00	.55	.16	-.21	-.23	-.33	-.36	-.47	-.60	-.44	-.48
ANNETTE	.34	.39	.50	.55	1.00	.25	-.12	.23	-.13	-.18	-.23	-.32	-.30	-.40
PORT HARDY	-.35	-.35	-.14	.16	.25	1.00	.46	.60	.52	.38	.23	.10	-.18	-.25
VICTORIA	-.36	-.47	-.45	-.21	-.12	.46	1.00	.66	.83	.77	.51	.40	-.13	-.20
VANCOUVER	-.17	-.39	-.36	-.23	.23	.60	.66	1.00	.69	.63	.44	.31	.08	-.11
SNOQ. FALLS	-.44	-.58	-.50	-.33	-.13	.52	.83	.69	1.00	.91	.82	.58	.07	-.09
CENTRALIA	-.51	-.58	-.47	-.36	-.18	.38	.77	.63	.91	1.00	.85	.66	.14	-.08
EUGENE	-.54	-.63	-.58	-.47	-.23	.23	.51	.44	.82	.85	1.00	.74	.29	.09
EUREKA	-.56	-.57	-.65	-.60	-.32	.10	.40	.31	.58	.66	.74	1.00	.41	.32
DAVIS	-.31	-.34	-.37	-.44	-.30	-.18	-.13	.08	.07	.14	.29	.41	1.00	.74
SANTA BARBARA	-.25	-.25	-.36	-.48	-.40	-.25	-.20	-.11	-.09	-.08	.09	.32	.74	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX FEB

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.59	.51	.41	.12	-.04	-.31	-.20	-.26	-.30	-.09	-.12	-.19	-.00
CORDOVA	.59	1.00	.81	.52	.19	-.09	-.32	-.36	-.30	-.32	-.21	-.41	-.32	-.16
YAKUTAT	.51	.81	1.00	.71	.47	.10	-.19	-.21	-.26	-.36	-.27	-.44	-.41	-.35
SITKA	.41	.52	.71	1.00	.60	.38	.00	-.24	-.16	-.31	-.24	-.39	-.50	-.48
ANNETTE	.12	.19	.47	.60	1.00	.66	.13	.18	.01	-.08	-.12	-.36	-.41	-.52
PORT HARDY	-.04	-.09	.10	.38	.66	1.00	.40	.40	.34	.28	.10	-.16	-.36	-.51
VICTORIA	-.31	-.32	-.19	.00	.13	.40	1.00	.69	.90	.80	.54	.13	-.32	-.40
VANCOUVER	-.20	-.36	-.21	-.24	.18	.40	.69	1.00	.66	.75	.56	.26	-.15	-.31
SNOQ. FALLS	-.26	-.30	-.26	-.16	.01	.34	.90	.66	1.00	.82	.50	.11	-.24	-.37
CENTRALIA	-.30	-.32	-.36	-.31	-.08	.28	.80	.75	.82	1.00	.76	.39	-.10	-.18
EUGENE	-.09	-.21	-.27	-.24	-.12	.10	.54	.56	.50	.76	1.00	.63	.13	-.06
EUREKA	-.12	-.41	-.44	-.39	-.36	-.16	.13	.26	.11	.39	.63	1.00	.52	.40
DAVIS	-.19	-.32	-.41	-.50	-.41	-.36	-.32	-.15	-.24	-.10	.13	.52	1.00	.77
SANTA BARBARA	-.00	-.16	-.35	-.48	-.52	-.51	-.40	-.31	-.37	-.18	-.06	.40	.77	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX MAR

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.49	.34	-.02	-.10	-.08	-.34	-.24	-.31	-.32	-.14	-.24	.02	.06
CORDOVA	.49	1.00	.82	.14	.02	-.20	-.25	-.15	-.34	-.22	-.10	-.37	-.15	-.12
YAKUTAT	.34	.82	1.00	.40	.15	-.18	-.18	-.19	-.41	-.29	-.25	-.56	-.29	-.20
SITKA	-.02	.14	.40	1.00	.37	.22	.10	-.14	-.08	.02	-.11	-.34	-.40	-.34
ANNETTE	-.10	.02	.15	.37	1.00	.49	.34	.31	.17	.41	.15	-.01	-.55	-.40
PORT HARDY	-.08	-.20	-.18	.22	.49	1.00	.39	.28	.42	.53	.23	.14	-.30	-.18
VICTORIA	-.34	-.25	-.18	.10	.34	.39	1.00	.68	.73	.63	.31	.09	-.45	-.38
VANCOUVER	-.24	-.15	-.19	-.14	.31	.28	.68	1.00	.84	.72	.58	.24	-.24	-.30
SNOQ. FALLS	-.31	-.34	-.41	-.08	.17	.42	.73	.84	1.00	.82	.59	.32	-.23	-.27
CENTRALIA	-.32	-.22	-.29	.02	.41	.53	.63	.72	.82	1.00	.65	.41	-.38	-.41
EUBENE	-.14	-.10	-.25	-.11	.15	.23	.31	.58	.59	.65	1.00	.61	-.01	-.29
EUREKA	-.24	-.37	-.56	-.34	-.01	.14	.09	.24	.32	.41	.61	1.00	.32	-.01
DAVIS	.02	-.15	-.29	-.40	-.55	-.30	-.45	-.24	-.23	-.38	-.01	.32	1.00	.75
SANTA BARBARA	.06	-.12	-.20	-.34	-.40	-.18	-.38	-.30	-.27	-.41	-.29	-.01	.75	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX APR

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.47	.42	.14	.16	.11	.18	-.08	.05	-.13	-.25	-.29	-.30	-.20
CORDOVA	.47	1.00	.68	.23	.18	.11	-.16	-.06	-.19	-.32	-.10	-.20	-.31	-.14
YAKUTAT	.42	.68	1.00	.46	.06	.16	-.06	-.15	-.21	-.37	-.29	-.31	-.47	-.27
SITKA	.14	.23	.46	1.00	.13	.10	.03	-.15	.01	-.18	-.21	-.37	-.40	-.22
ANNETTE	.16	.18	.06	.13	1.00	.51	-.05	-.09	-.10	-.13	-.19	-.42	-.37	-.10
PORT HARDY	.11	.11	.16	.10	.51	1.00	.38	.32	.37	.31	-.03	-.26	-.34	-.27
VICTORIA	.18	-.16	-.06	.03	-.05	.38	1.00	.37	.53	.43	.01	-.02	-.02	.00
VANCOUVER	-.08	-.06	-.15	-.15	-.09	.32	.37	1.00	.44	.52	.57	.25	.11	-.06
SNOQ. FALLS	.05	-.19	-.21	.01	-.10	.37	.53	.44	1.00	.72	.43	.21	.06	-.01
CENTRALIA	-.13	-.32	-.37	-.18	-.13	.31	.43	.52	.72	1.00	.57	.42	.17	.06
EUGENE	-.25	-.10	-.29	-.21	-.19	-.03	.01	.57	.43	.57	1.00	.63	.42	.29
EUREKA	-.29	-.20	-.31	-.37	-.42	-.26	-.02	.25	.21	.42	.63	1.00	.75	.57
DAVIS	-.30	-.31	-.47	-.40	-.37	-.34	-.02	.11	.06	.17	.42	.75	1.00	.79
SANTA BARBARA	-.20	-.14	-.27	-.22	-.10	-.27	.00	-.06	-.01	.06	.29	.57	.79	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX MAY

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.35	.14	-.02	-.05	-.23	-.02	-.08	.28	.24	.27	.17	-.01	.16
CORDOVA	.35	1.00	.73	.37	.07	-.19	.01	-.21	-.14	-.22	-.11	-.21	-.05	.11
YAKUTAT	.14	.73	1.00	.62	.28	-.09	-.12	-.25	-.25	-.38	-.34	-.36	-.22	.04
SITKA	-.02	.37	.62	1.00	.46	-.06	-.14	-.41	-.45	-.55	-.63	-.43	-.24	-.10
ANNETTE	-.05	.07	.28	.46	1.00	.18	-.11	-.29	-.25	-.23	-.37	-.14	-.23	-.12
PORT HARDY	-.23	-.19	-.09	-.06	.18	1.00	.20	.28	.12	.17	-.24	-.03	-.17	-.35
VICTORIA	-.02	.01	-.12	-.14	-.11	.20	1.00	.69	.60	.60	.14	.14	.38	-.20
VANCOUVER	-.08	-.21	-.25	-.41	-.29	.28	.69	1.00	.74	.68	.24	.15	.25	-.14
SNOQ. FALLS	.28	-.14	-.25	-.45	-.25	.12	.60	.74	1.00	.70	.45	.42	.32	-.06
CENTRALIA	.24	-.22	-.38	-.55	-.23	.17	.60	.68	.70	1.00	.46	.37	.33	.15
EUGENE	.27	-.11	-.34	-.63	-.37	-.24	.14	.24	.45	.46	1.00	.74	.42	.26
EUREKA	.17	-.21	-.36	-.43	-.14	-.03	.14	.15	.42	.37	.74	1.00	.27	.16
DAVIS	-.01	-.05	-.22	-.24	-.23	-.17	.38	.25	.32	.33	.42	.27	1.00	.55
SANTA BARBARA	.16	.11	.04	-.10	-.12	-.35	-.20	-.14	-.06	.15	.26	.16	.55	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX JUN

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.59	.63	.51	.12	-.26	-.18	-.26	-.40	-.27	-.25	-.38	-.21	.02
CORDOVA	.59	1.00	.68	.55	-.10	-.37	-.44	-.38	-.30	-.31	-.11	-.21	-.07	.04
YAKUTAT	.63	.68	1.00	.85	.27	-.27	-.24	-.38	-.27	-.36	-.26	-.30	-.13	.16
SITKA	.51	.55	.85	1.00	.31	-.23	-.35	-.49	-.38	-.49	-.32	-.38	-.23	.15
ANNETTE	.12	-.10	.27	.31	1.00	.32	.39	.32	.06	-.18	-.27	-.06	-.30	.12
PORT HARDY	-.26	-.37	-.27	-.23	.32	1.00	.42	.49	.43	.18	.01	.09	-.23	.15
VICTORIA	-.18	-.44	-.24	-.35	.39	.42	1.00	.76	.54	.46	.11	.24	.10	.02
VANCOUVER	-.26	-.38	-.38	-.49	.32	.49	.76	1.00	.47	.52	.18	.43	-.14	.05
SNOQ. FALLS	-.40	-.30	-.27	-.38	.06	.43	.54	.47	1.00	.62	.51	.53	.27	.18
CENTRALIA	-.27	-.31	-.36	-.49	-.18	.18	.46	.52	.62	1.00	.46	.52	.21	-.10
EUGENE	-.25	-.11	-.26	-.32	-.27	.01	.11	.18	.51	.46	1.00	.63	.27	-.08
EUREKA	-.38	-.21	-.30	-.38	-.06	.09	.24	.43	.53	.52	.63	1.00	.20	-.13
DAVIS	-.21	-.07	-.13	-.23	-.30	-.23	.10	-.14	.27	.21	.27	.20	1.00	-.00
SANTA BARBARA	.02	.04	.16	.15	.12	.15	.02	.05	.18	-.10	-.08	-.13	-.00	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX JUL

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.52	.49	.51	.02	-.07	.10	.04	.24	.06	.16	.13	.06	-.10
CORDOVA	.52	1.00	.43	.50	.02	-.27	-.17	-.32	-.11	-.24	-.13	-.14	-.12	-.08
YAKUTAT	.49	.43	1.00	.52	.35	-.22	-.38	-.27	-.27	-.28	-.09	-.02	-.22	-.00
SITKA	.51	.50	.52	1.00	.39	-.21	-.11	-.16	-.13	-.27	-.10	-.16	-.02	.19
ANNETTE	.02	.02	.35	.39	1.00	.22	-.16	-.07	-.33	-.20	-.07	.07	-.18	.38
PORT HARDY	-.07	-.27	-.22	-.21	.22	1.00	.37	.56	.33	.54	.39	.62	.15	-.01
VICTORIA	.10	-.17	-.38	-.11	-.16	.37	1.00	.85	.74	.61	.12	.13	.32	.19
VANCOUVER	.04	-.32	-.27	-.16	-.07	.56	.85	1.00	.70	.61	.26	.39	.36	.03
SNOQ. FALLS	.24	-.11	-.27	-.13	-.33	.33	.74	.70	1.00	.76	.36	.20	.39	-.05
CENTRALIA	.06	-.24	-.28	-.27	-.20	.54	.61	.61	.76	1.00	.63	.49	.36	.06
EUGENE	.16	-.13	-.09	-.10	-.07	.39	.12	.26	.36	.63	1.00	.72	.38	-.08
EUREKA	.13	-.14	-.02	-.16	.07	.62	.13	.39	.20	.49	.72	1.00	-.01	-.07
DAVIS	.06	-.12	-.22	-.02	-.18	.15	.32	.36	.39	.36	.38	-.01	1.00	-.04
SANTA BARBARA	-.10	-.08	-.00	.19	.38	-.01	.19	.03	-.05	.06	-.08	-.07	-.04	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX AUG

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.70	.42	.12	-.02	-.27	-.30	-.37	-.37	-.26	-.30	-.18	-.01	.01
CORDOVA	.70	1.00	.69	.53	.16	-.14	-.41	-.39	-.46	-.42	-.38	-.24	-.08	-.15
YAKUTAT	.42	.69	1.00	.63	.28	-.30	-.35	-.34	-.37	-.44	-.28	-.35	.05	-.22
SITKA	.12	.53	.63	1.00	.38	-.01	-.20	-.12	-.36	-.38	-.33	-.32	-.21	-.36
ANNETTE	-.02	.16	.28	.38	1.00	.03	-.14	-.19	-.28	-.43	-.19	-.22	-.15	-.36
PORT HARDY	-.27	-.14	-.30	-.01	.03	1.00	.42	.68	.41	.34	.12	.21	-.01	-.24
VICTORIA	-.30	-.41	-.35	-.20	-.14	.42	1.00	.76	.77	.64	.53	.56	.39	-.04
VANCOUVER	-.37	-.39	-.34	-.12	-.19	.68	.76	1.00	.75	.67	.41	.50	.28	-.01
SNOQ. FALLS	-.37	-.46	-.37	-.36	-.28	.41	.77	.75	1.00	.78	.71	.57	.43	.15
CENTRALIA	-.26	-.42	-.44	-.38	-.43	.34	.64	.67	.78	1.00	.61	.54	.23	.46
EUGENE	-.30	-.38	-.28	-.33	-.19	.12	.53	.41	.71	.61	1.00	.71	.34	.04
EUREKA	-.18	-.24	-.35	-.32	-.22	.21	.56	.50	.57	.54	.71	1.00	.47	-.12
DAVIS	-.01	-.08	.05	-.21	-.15	-.01	.39	.28	.43	.23	.34	.47	1.00	-.13
SANTA BARBARA	.01	-.15	-.22	-.36	-.36	-.24	-.04	-.01	.15	.46	.04	-.12	-.13	1.00

Table 4.3 Continued

MONTHLY CORRELATION MATRIX SEP

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.63	.41	.16	-.12	-.41	-.49	-.49	-.50	-.50	-.25	-.20	-.09	.24
CORDOVA	.63	1.00	.58	.38	-.05	-.47	-.30	-.31	-.35	-.34	-.19	-.12	-.19	.13
YAKUTAT	.41	.58	1.00	.33	.15	-.34	-.31	-.34	-.54	-.37	-.20	-.20	-.26	.40
SITKA	.16	.38	.33	1.00	.37	.07	-.12	-.23	-.15	-.13	.15	-.06	-.08	-.05
ANNETTE	-.12	-.05	.15	.37	1.00	.34	-.07	-.12	-.03	.09	.22	.09	.05	.25
PORT HARDY	-.41	-.47	-.34	.07	.34	1.00	.44	.45	.37	.52	.67	.27	-.05	-.09
VICTORIA	-.49	-.30	-.31	-.12	-.07	.44	1.00	.85	.63	.54	.29	.07	.23	-.13
VANCOUVER	-.49	-.31	-.34	-.23	-.12	.45	.85	1.00	.79	.67	.42	.34	.35	-.06
SNOQ. FALLS	-.50	-.35	-.54	-.15	-.03	.37	.63	.79	1.00	.71	.47	.33	.36	-.05
CENTRALIA	-.50	-.34	-.37	-.13	.09	.52	.54	.67	.71	1.00	.81	.66	.19	.08
EUGENE	-.25	-.19	-.20	.15	.22	.67	.29	.42	.47	.81	1.00	.71	.07	.09
EUREKA	-.20	-.12	-.20	-.06	.09	.27	.07	.34	.33	.66	.71	1.00	.35	-.09
DAVIS	-.09	-.19	-.26	-.08	.05	-.05	.23	.35	.36	.19	.07	.35	1.00	-.03
SANTA BARBARA	.24	.13	.40	-.05	.25	-.09	-.13	-.06	-.05	.08	.09	-.09	-.03	1.00

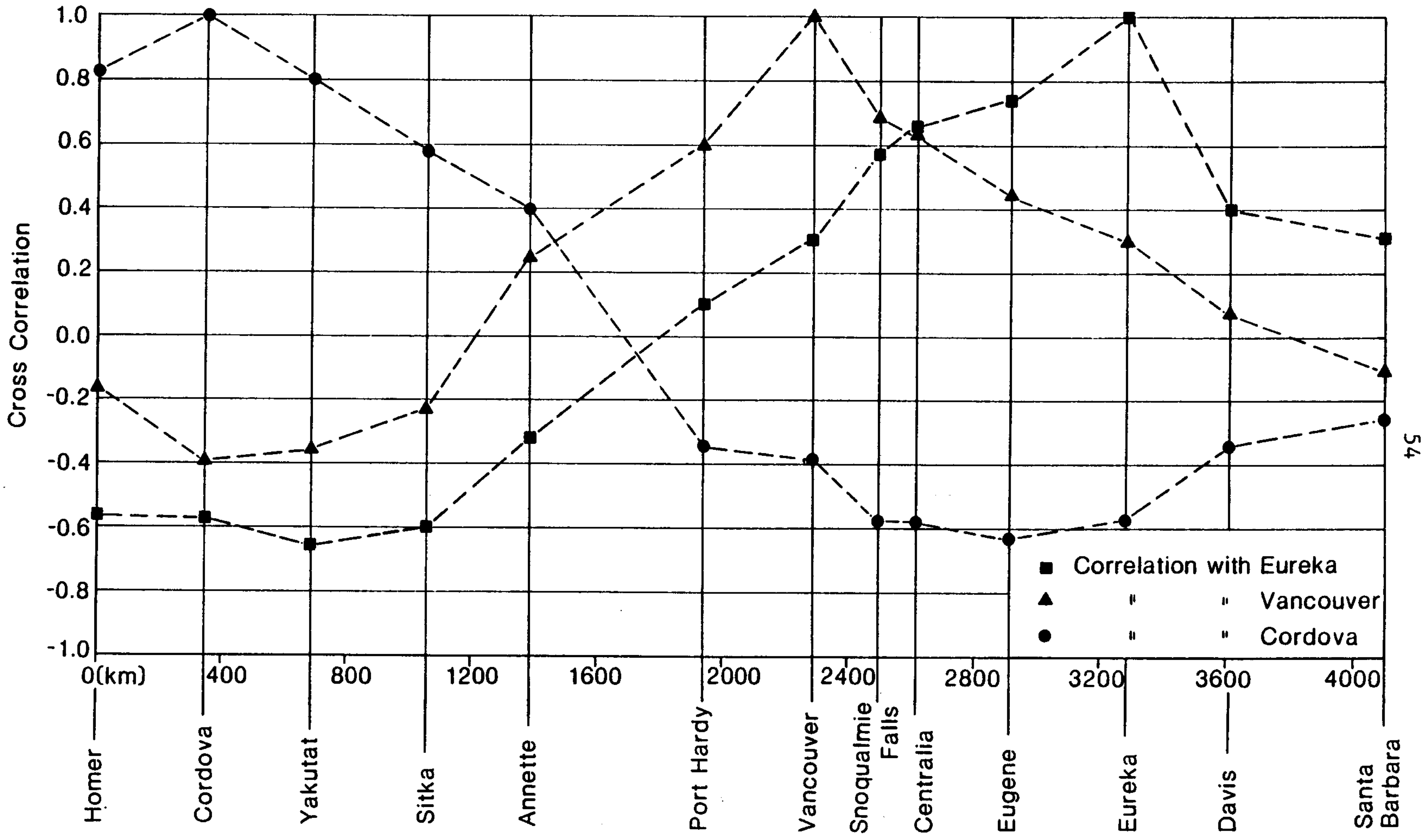


Figure 4.3 Variation of January cross correlation with distance

Table 4.4 Basic Statistics for Annual Precipitation Data (in mm)
(1947-1978)

	MEAN	STDEV	SKEW	MAX	MIN	CORR (LAGS 1-4)			
HOMER	575.8	135.3	.53	895.6	301.3	-.15	-.03	-.20	-.13
CORDOVA	2276.1	409.0	.22	3227.3	1466.0	-.23	.11	-.25	-.21
YAKUTAT	3359.4	688.3	.38	4820.8	2118.0	.26	.21	-.05	.08
SITKA	2413.8	365.0	.10	3193.0	1624.0	.10	.15	-.03	-.22
ANNETTE	2915.4	636.7	1.10	4647.5	2028.5	.50	.24	.21	.27
PORT HARDY	1777.1	262.9	.50	2399.7	1313.9	-.00	.01	.01	.27
VICTORIA	668.3	145.2	-.18	903.2	376.2	-.20	-.09	-.03	.28
VANCOUVER	1111.6	164.0	.35	1485.3	840.0	-.39	.05	.00	.08
SNOQ. FALLS	1602.0	276.6	-.05	2092.2	1028.8	-.35	-.06	.09	.05
CENTRALIA	1202.6	187.2	-.41	1541.4	759.0	-.50	.01	-.05	.11
EUGENE	1161.2	251.8	.67	1884.1	628.4	-.24	.32	.03	.17
EUREKA	985.4	184.8	-.65	1296.7	487.0	-.09	.19	-.24	-.06
DAVIS	426.7	154.2	.58	717.5	194.0	-.32	.02	-.45	.23
SANTA BARBARA	389.5	172.2	1.21	937.6	176.1	-.25	.08	-.23	.20

Table 4.5 Cross Correlations for Annual Precipitation Data (1947-1978)

	HOMER	CORDOVA	YAKUTAT	SITKA	ANNETTE	PORT HARDY	VICTORIA	VANCOUVER	SNOQ. FALLS	CENTRALIA	EUGENE	EUREKA	DAVIS	SANTA BARBARA
HOMER	1.00	.71	.53	.47	.18	-.10	-.30	-.23	-.17	-.21	-.22	-.19	-.26	-.28
CORDOVA	.71	1.00	.58	.43	-.02	-.13	-.41	-.48	-.38	-.38	-.43	-.32	-.26	-.29
YAKUTAT	.53	.58	1.00	.57	.28	.11	-.20	-.13	-.21	-.25	-.35	-.43	-.29	-.27
SITKA	.47	.43	.57	1.00	.36	-.02	-.03	-.14	-.05	-.28	-.44	-.40	-.30	-.37
ANNETTE	.18	-.02	.28	.36	1.00	.05	.20	-.02	-.07	-.02	-.38	-.27	-.19	-.23
PORT HARDY	-.10	-.13	.11	-.02	.05	1.00	.44	.53	.41	.33	.07	-.15	-.32	-.25
VICTORIA	-.30	-.41	-.20	-.03	.20	.44	1.00	.73	.77	.64	.17	.11	-.43	-.53
VANCOUVER	-.23	-.48	-.13	-.14	-.02	.53	.73	1.00	.83	.80	.53	.18	-.21	-.24
SNOQ. FALLS	-.17	-.38	-.21	-.05	-.07	.41	.77	.83	1.00	.80	.58	.24	-.32	-.46
CENTRALIA	-.21	-.38	-.25	-.28	-.02	.33	.64	.80	.80	1.00	.69	.45	-.02	-.21
EUGENE	-.22	-.43	-.35	-.44	-.38	.07	.17	.53	.58	.69	1.00	.58	.29	.15
EUREKA	-.19	-.32	-.43	-.40	-.27	-.15	.11	.18	.24	.45	.58	1.00	.47	.22
DAVIS	-.26	-.26	-.29	-.30	-.19	-.32	-.43	-.21	-.32	-.02	.29	.47	1.00	.75
SANTA BARBARA	-.28	-.29	-.27	-.37	-.23	-.25	-.53	-.24	-.46	-.21	.15	.22	.75	1.00

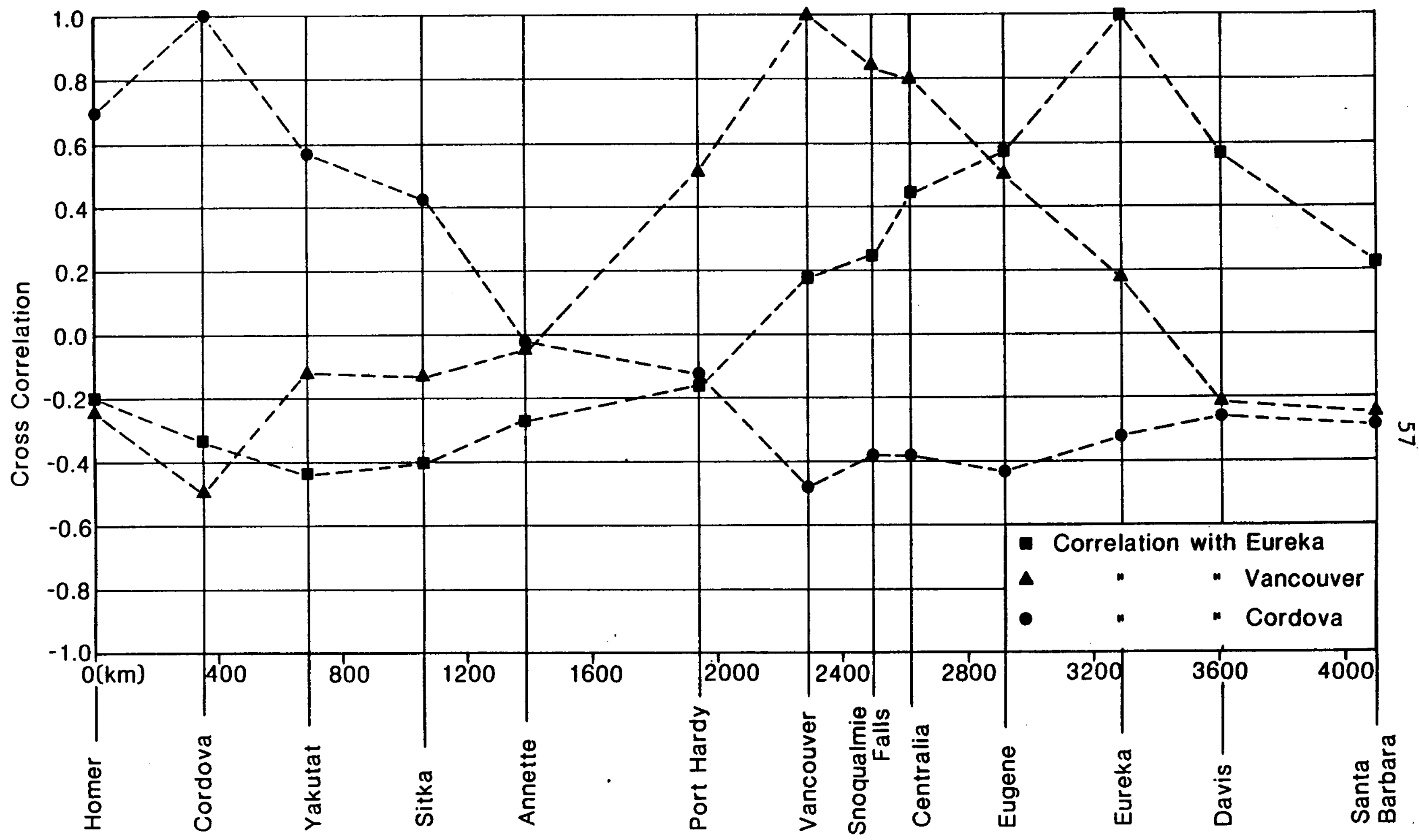


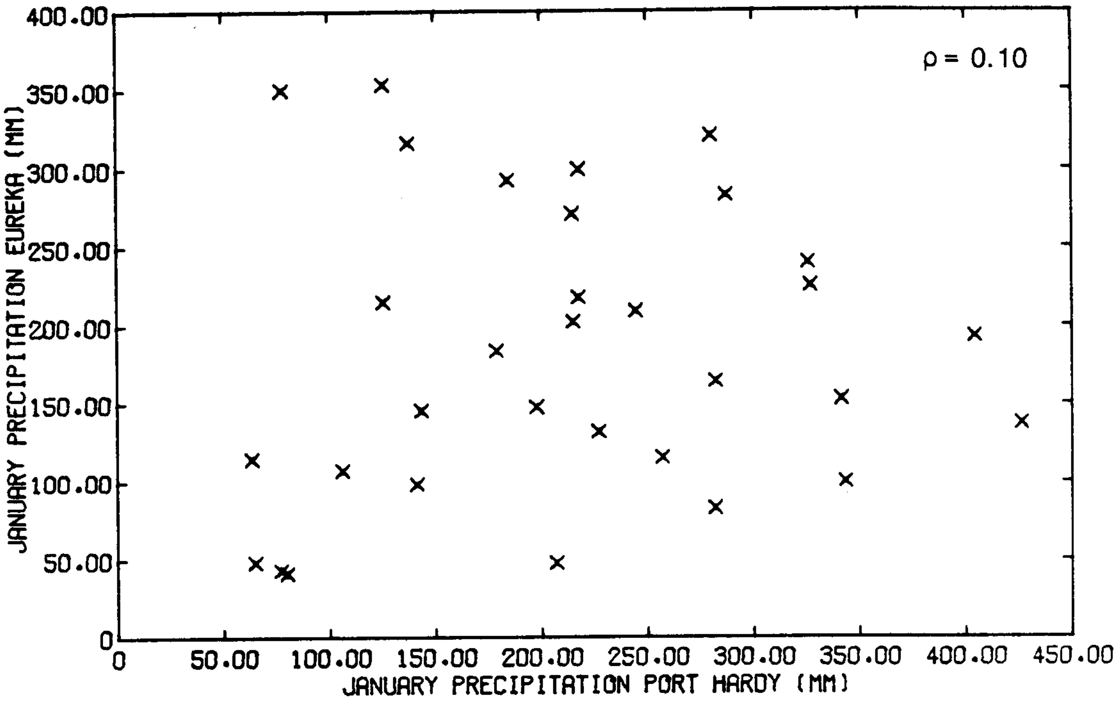
Figure 4.4 Variation of annual cross correlation with distance

For example, very dry conditions at Eureka are generally associated with a persistent northerly jet stream track and hence unusually wet conditions in the Gulf of Alaska. Thus negative correlations have a physical basis and are not necessarily a feature of sampling variability.

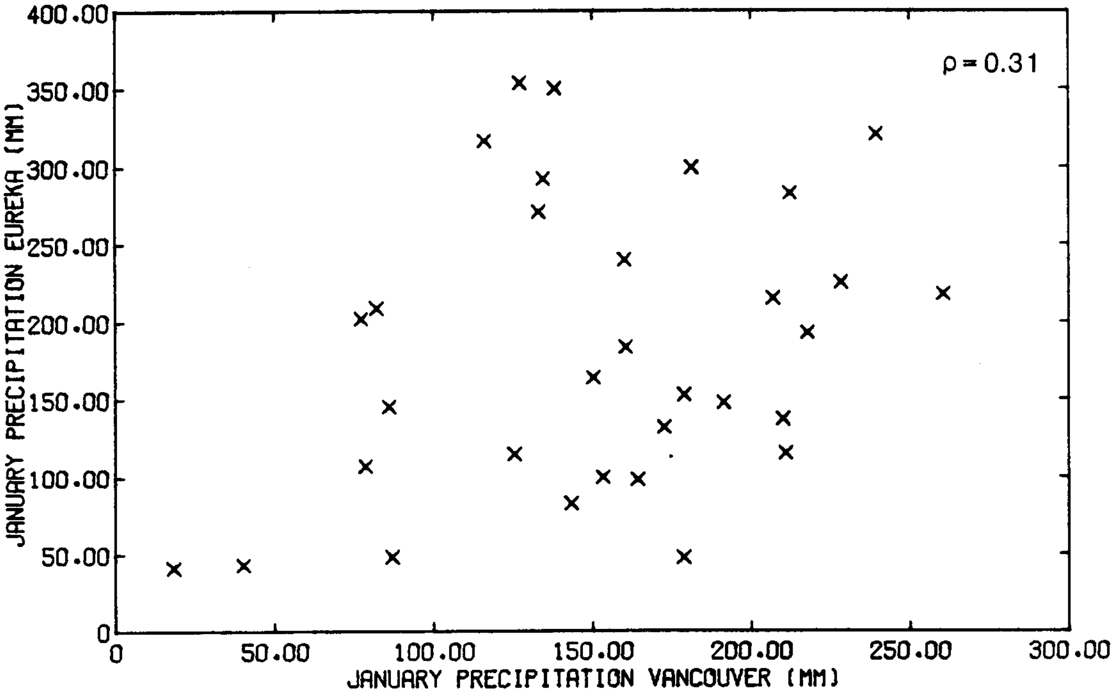
Another aspect of interest in Table 4.3 is comparison of cross correlations in winter and summer months. For example a comparison of January and July cross correlations shows that in general cross correlations in July are weaker than those in January and the variation of cross correlation with distance does not show the same degree of consistency. This is particularly true for cross correlations with stations in California. Again this feature can be explained qualitatively by considering the large scale atmospheric circulation for the winter and summer months. Circulation in summer is generally much weaker than in the winter months; zonal upper-level winds are weaker; rainfall is associated with small frontal systems and the exit region for the jet stream is ill-defined. A further complicating feature, especially for the California stations, is the large proportion of months with no rainfall.

Returning to the January data we have noted that negative cross correlations at large distances are physically reasonable. However, as pointed out earlier, the inter-station relationships may have a complex nonlinear structure and the simple cross correlation may not be an appropriate measure of dependence. This problem is illustrated in scatterplots for the January and the annual data shown in Figures 4.5 and 4.6.

Figure 4.5a shows the scatterplot of January data for Port Hardy vs. Eureka (separation 1300 km). The cross correlation is 0.10. However, this value is greatly influenced by the widespread droughts of January 1949, 1963, and 1977 which appear as a cluster of points in

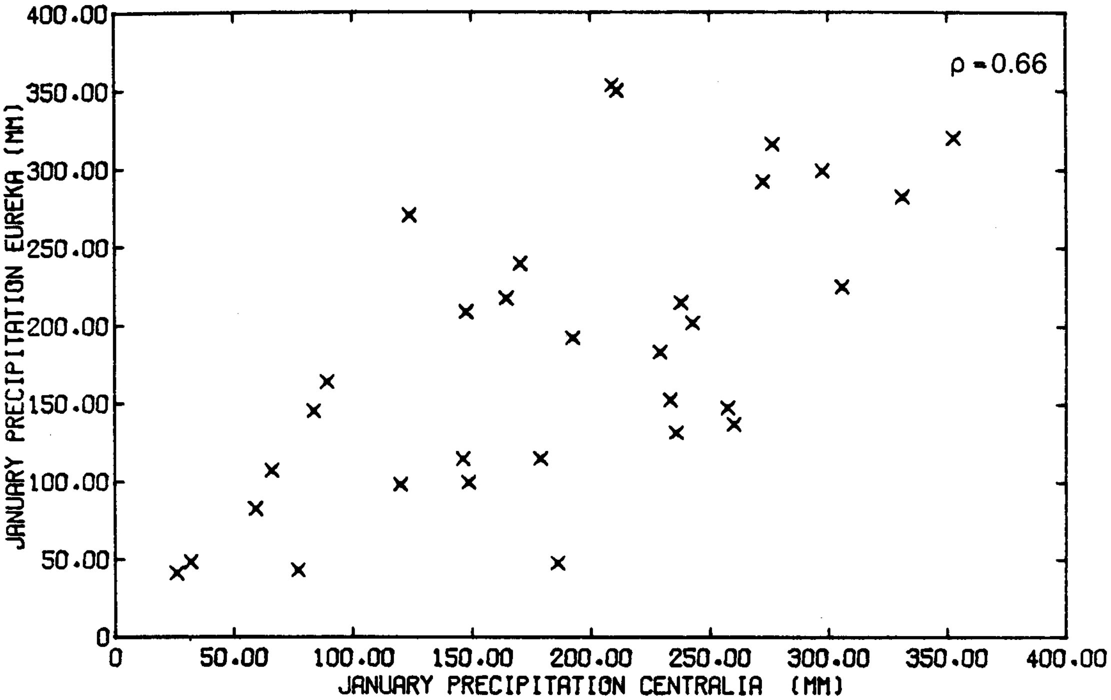


(a) Port Hardy vs. Eureka

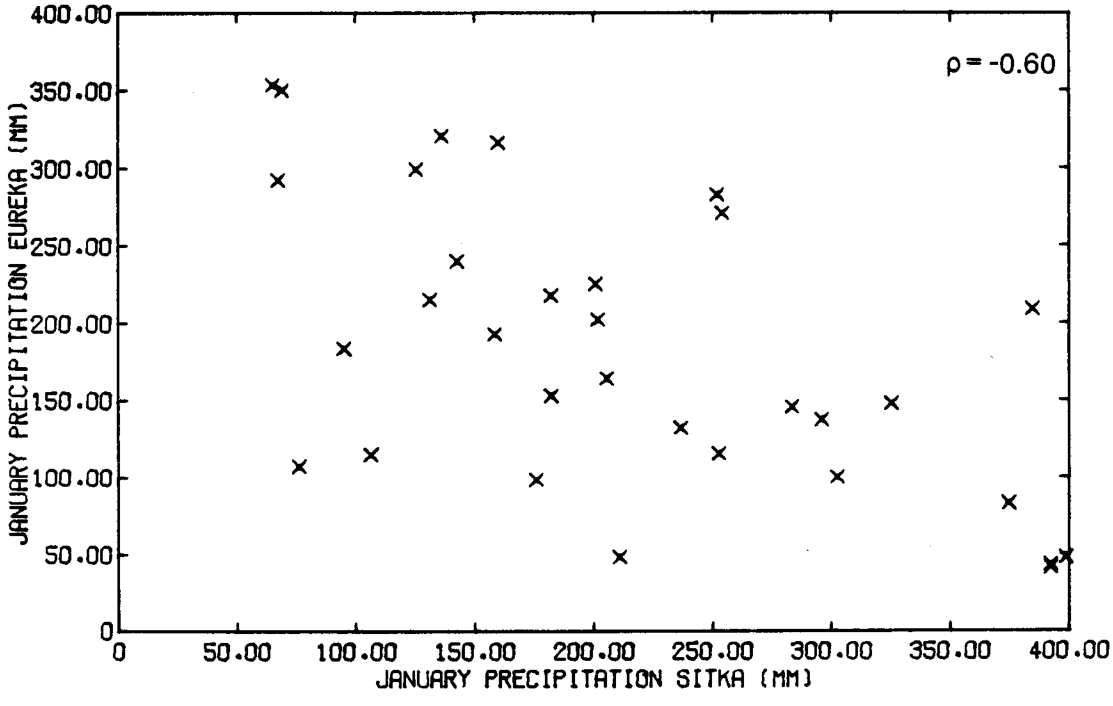


(b) Vancouver vs. Eureka

Figure 4.5 Scatterplots for January precipitation data

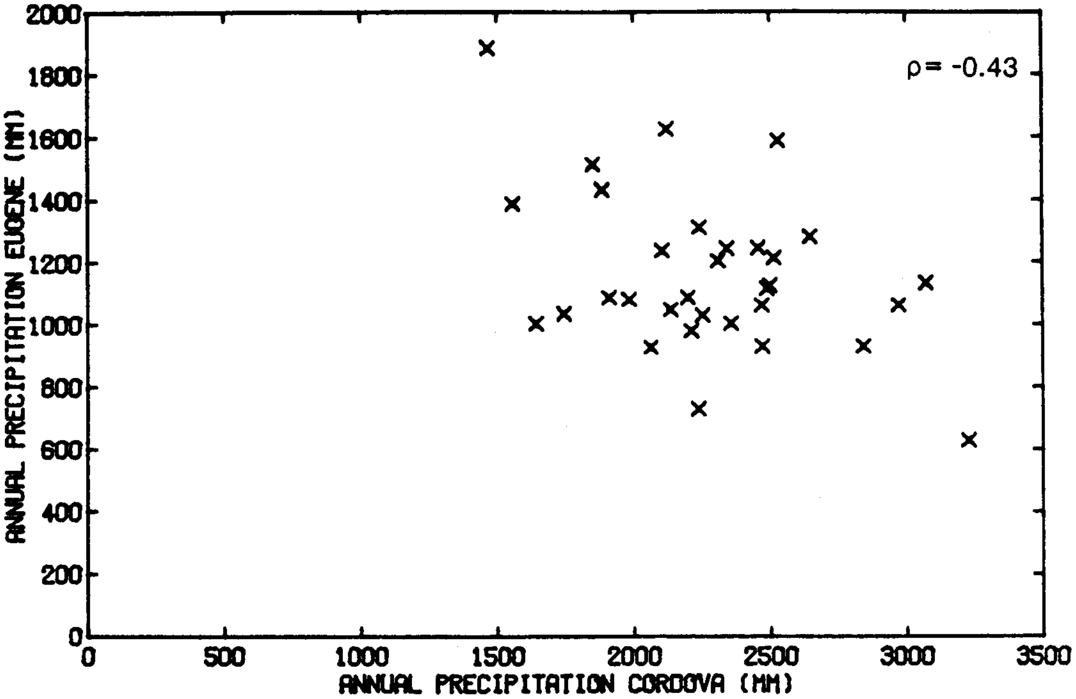


(c) Centralia vs. Eureka

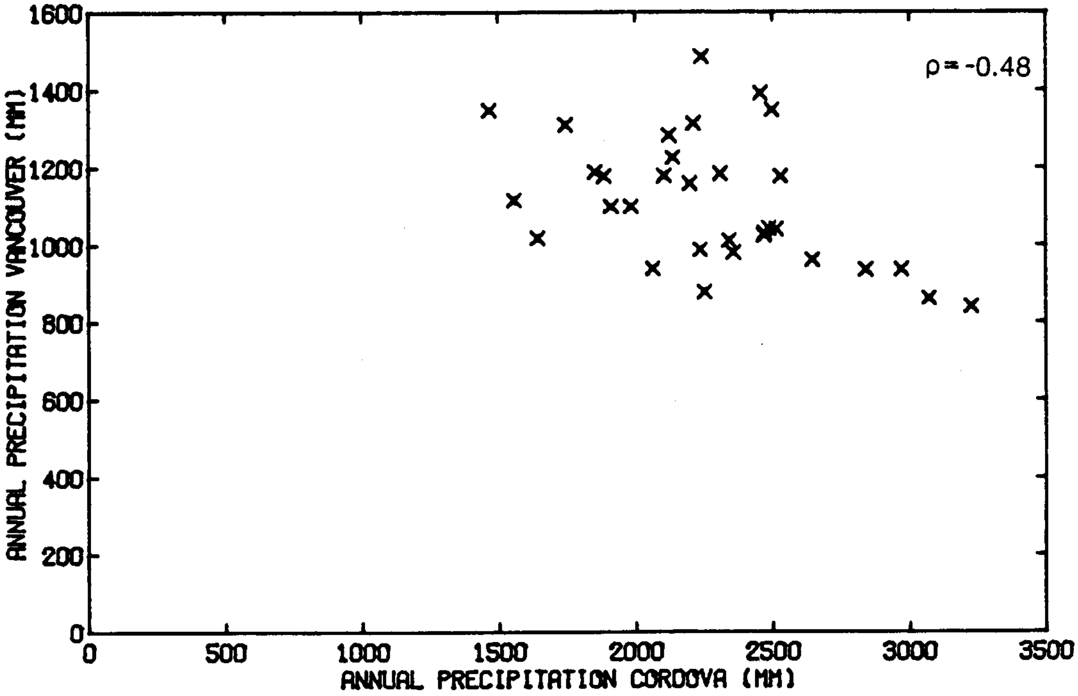


(d) Sitka vs. Eureka

Figure 4.5 Continued



(a) Cordova vs. Eugene



(b) Cordova vs. Vancouver

Figure 4.6 Scatterplots for annual precipitation data

the bottom left of the plot. Elimination of these points would give a slightly negative cross correlation.

A similar feature, though not as pronounced, is apparent in the scatterplot for Vancouver vs. Eureka (separation 1000 km) in Figure 4.5b. As the separation between stations decreases the assumption of linearity between rainfall depths at the two stations improves. Figure 4.5c shows the scatterplot for Centralia vs. Eureka and at this separation (700 km) the assumption of linearity would seem reasonable.

It is evident in Figures 4.5a to 4.5c that the scale of meteorologic phenomenona is important in determining the nature of inter-station relationships. Frontal systems on the west coast may affect stations up to 500 km apart whereas drought may affect stations up to 2000 km apart. In Figure 4.5a (station separation 1300 km) rainfall depths during normal or wet conditions at either station are essentially uncorrelated at the monthly interval. This is because (1) station separation is large in comparison to the size of frontal systems. (2) Frontal systems pass through the region in three to five days thus the rainfall in one month may be made up of contributions from a number of storms. (3) Apparently random fluctuations in the position of the jet stream track can move frontal systems several hundred kilometers north or south of the mean monthly position without greatly affecting the predominant pattern of circulation. In contrast, severe drought often extends from central California to southern Alaska and thus drought conditions may be expected to occur simultaneously at both Eureka and Port Hardy.

The inter-station relationships appear to become even more complicated as station separation is further increased. Figure 4.5d shows the January scatterplot for Sitka vs. Eureka (separation 2200 km). In this case, as discussed in Chapter 3, drought in California is associated with a persistent northerly jet stream track and wet conditions in the Gulf of Alaska. Similarly a persistent southerly jet stream track brings wet conditions to the Pacific Northwest and northern California and predominantly dry conditions in the Gulf of Alaska. Thus extreme wet or dry conditions are negatively correlated but it appears that normal conditions are essentially uncorrelated. Very similar phenomena appear in the annual (water year) scatterplots for Eugene vs. Cordova (Figure 4.6a, separation 2500 km) and Vancouver vs. Cordova (Figure 4.6b, separation 1900 km).

The scatterplots shown here demonstrate nonlinearities in the inter-station rainfall relationships, which can be attributed qualitatively to features of the large scale atmospheric circulation. The significance of these nonlinearities in stochastic hydrology is, however, unclear and is explored in the next section.

4.2 Scale and Dimensionality Problems in Multi-site Stochastic Precipitation Models

As indicated earlier, all current multisite stochastic models use the simple linear cross correlation to express interstation dependence. A typical approach to multi-site data generation is to transform the historic data at individual sites such that the distributions of the transformed data are normal. It is then assumed that the transformed data are in fact from a multivariate normal probability distribution and an appropriate scheme is adopted to sample from the multivariate population maintaining a suitable serial correlation structure. The inverse transform is then applied to the synthetic data at the individual sites to return to the natural distributions.

Irrespective of the details of the generation scheme, I suggest that the models currently in use for monthly or annual multi-site rainfall generation may suffer from two possible deficiencies. I will term these the "scale" effect and the "dimensionality" effect.

The scale effect was illustrated in Section 4.1. The term refers to the problems of modeling drought at widely separated points where joint occurrences of drought are to be expected but where normal or wet conditions are essentially independent. The averaging of the correlation structure in wet and dry periods will result in synthetic drought sequences which are less severe than those found in nature.

The problem of dimensionality is also related to the large areal extent characteristic of drought, and is concerned with the difficulties of ensuring that severe synthetic droughts occur concurrently at all necessary points in the study area. The problem is perhaps best illustrated by means of a rather artificial example.

Suppose that we are interested in generating data at a number of sites in a relatively small region such as western Washington. Further, suppose for simplicity that all sites have zero cross correlation. Now define a severe drought as a situation where rainfall is more than one standard deviation below the mean at any site. We are interested in the joint probability of severe drought at 2, 3, 4,...,n sites. Assuming a multivariate normal distribution we have:

$$\begin{aligned}
&\text{P(drought at 1 site)} \approx 0.16 \\
&\text{P(drought at 2 sites)} \approx 0.16^2 \approx 0.026 \\
&\text{P(drought at 3 sites)} \approx 0.16^3 \approx 0.004 \\
&\ldots \\
&\text{P(drought at n sites)} \approx 0.16^n
\end{aligned}$$

In this situation the probability of joint occurrences of drought at multiple sites rapidly becomes extremely small, whereas we know that drought frequently affects large regions, and the probability of joint low events at many sites in such a region should not be negligible.

This example is clearly unrealistic in that I have assumed zero cross correlations. This assumption was made because of the difficulty of calculating the required joint probabilities for more realistic correlation matrices. However, the point I wish to make is that even given a realistic correlation structure, there is no mechanism in current models to ensure that drought occurs concurrently at multiple points. I conjecture that this again results in synthetic drought sequences which are less severe in areal extent than those found in nature.

The purpose of this section is to evaluate the multi-site characteristics of synthetic drought sequences by comparing synthetic with historic data and to determine the extent and severity of the scale and dimensionality problems referred to above.

4.2.1 Evaluation of Scale Effects

Evaluation of the scale effect in multi-site modeling was undertaken using the January monthly precipitation record of 32 observations from Eureka and Port Hardy. The location of these sites was shown in Figure 4.1.

January data alone were used to simplify the evaluation and also because the primary concern in the study is in winter precipitation. Most major water resource facilities on the west coast rely heavily on snowmelt from winter precipitation and thus deficits in winter precipitation are of principal interest. The basic statistics for the January data at Port Hardy and Eureka were shown in Tables 4.2 and 4.3 and a scatterplot of the data was shown in Figure 4.5a. The serial correlation for the January data at the two sites is also of interest and this is shown for lags of 1 through 4 in Table 4.6.

Table 4.6 January Serial-correlations at Port Hardy and Eureka

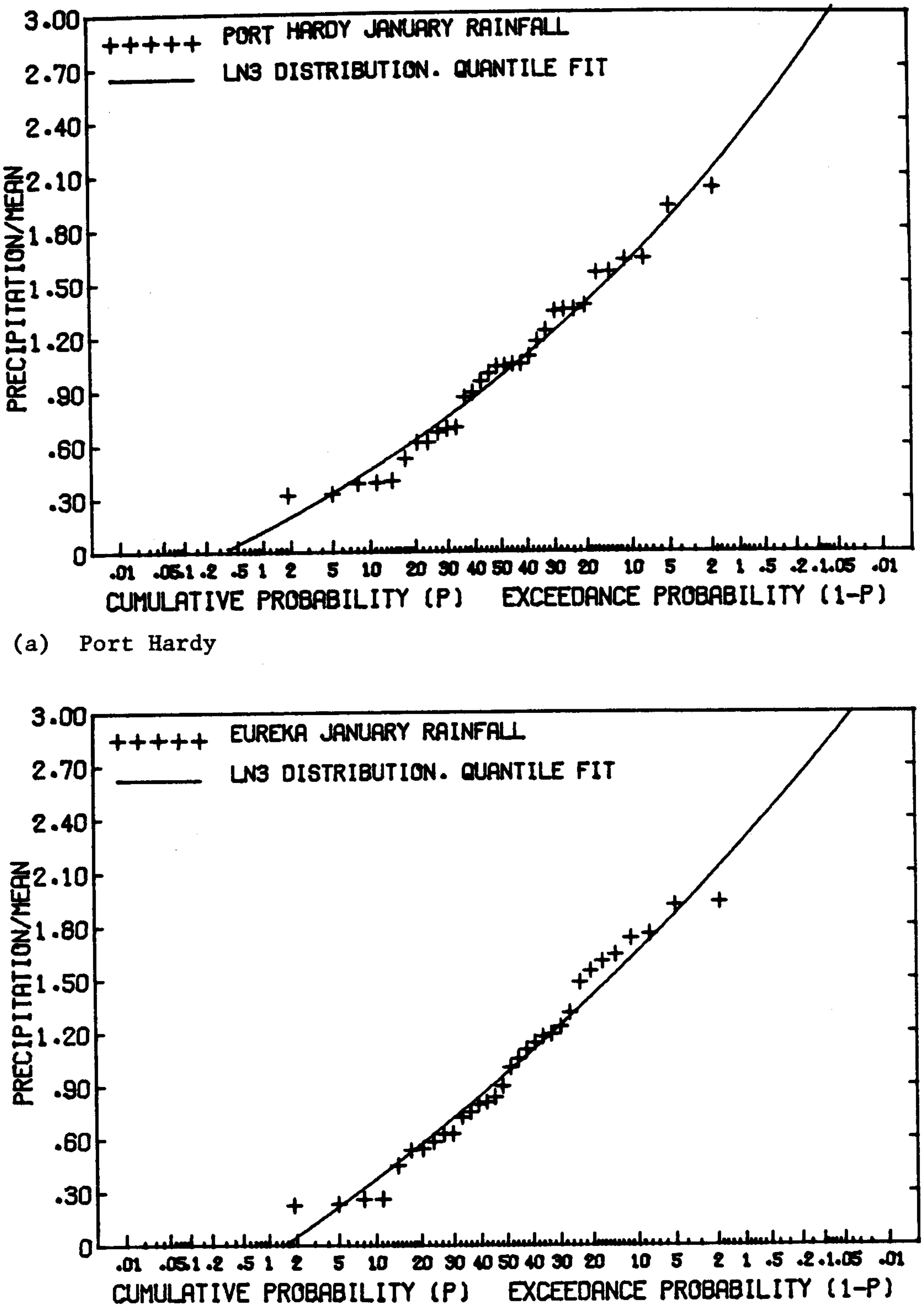
	Lag			
Site	1	2	3	4
Port Hardy	0.185	-0.053	-0.004	-0.135
Eureka	0.189	0.351	0.140	0.096

The serial-correlations for both the January data and annual data are small. Note that the lag two January correlation at Eureka of 0.351 is probably the result of sampling variability as there is no obvious reason why the lag two correlation should be greater than the lag one correlation. Moreover, our principal interest is in the spatial characteristics of the synthetic series in expectation. Thus the January data, after suitable transformation, will be assumed to be independent samples from a bivariate normal population.

Cumulative distribution functions for the Port Hardy and Eureka data are shown in Figure 4.7 along with fitted three parameter log normal (LN3) distributions. The LN3 parameters were chosen to provide a good fit over the lower quantiles to ensure that low events at the individual sites would be represented properly in the synthetic data. The LN3 were fitted by the mixed maximum likelihood/quantile method suggested by Stedinger (1980) with some adjustment of the sample quantiles by eye. The parameters of the fitted distributions are given in Table 4.7.

Table 4.7 LN3 Parameters for January Precipitation Data at Port Hardy and Eureka

Station	Transformation	μ_y	σ_y	a
Port Hardy	$y = \log(x - a)$	7.784	0.0436	-2187
Eureka	$y = \log(x - a)$	7.314	0.0676	-1317



(a) Port Hardy

(b) Eureka

Figure 4.7 CDF's for January precipitation data

The LN3 distributions give a reasonable fit to the data at both sites, thus the transformations

$$y = \log(x-a) \tag{4.1}$$

where y = transformed data $N(\mu_y,\sigma_y)$
a = shift parameter
x = raw data LN3 (μ_x,σ_x,a)

allow the data to be treated as samples from a bivariate normal. The correlation coefficient of 0.1 was estimated by moments from the raw data (Table 4.3). The correction to the cross correlation to account for the log transformation (Mejia and Rodriguez-Iturbe 1974a) was found to be negligible.

Since the serial correlation structure is of no interest, the generation scheme in the log domain can take the particularly simple form (Matalas 1967):

$$\underline{y}_t = \underline{B}\,\underline{\varepsilon}_t + \underline{\mu}_y \tag{4.2}$$

where $\underline{y}_t$ = (n x 1) matrix containing synthesized data in the transformed domain at time t with the normal distribution $N(\mu_y,\sigma_y)$
$\underline{\mu}_y$ = (n x 1) matrix containing the means of the series in the transformed domain
$\underline{\varepsilon}_t$ = (n x 1) matrix whose elements are independent identically distributed samples from the normal N(0,1) distribution
$\underline{B}$ = (n x n) matrix such that
$\underline{BB}^T = \underline{M}$
$\underline{M}$ = lag-zero covariance matrix in the transformed domain
n = number of sites

The matrix $\underline{B}$ was found from the relationship $\underline{BB}^T = \underline{M}$ using a Choleski decomposition (Pinder and Gray 1977). The inverse transform applied to $\underline{y}_t$ returns synthetic data to the natural domain.

The procedure used for evaluating the multi-site characteristics of the above model was through Monte Carlo simulation. For Port Hardy and Eureka 3200 years of synthetic January data were generated. This corresponds to one hundred 32-year sequences (recall that the historic record was 32 years in length). The 10, 15, 20, 25,... 50 percent quantiles for the two synthetic sequences were determined and then joint occurrences were counted in which both sites had rainfall less than or equal to their respective 10, 15, 20... 50 percent quantiles. These counts were divided by 100 to give an estimate of the expected number of joint occurrences in a period of 32 years. These data and the corresponding data from the 32 year historic record are shown in Figure 4.8.

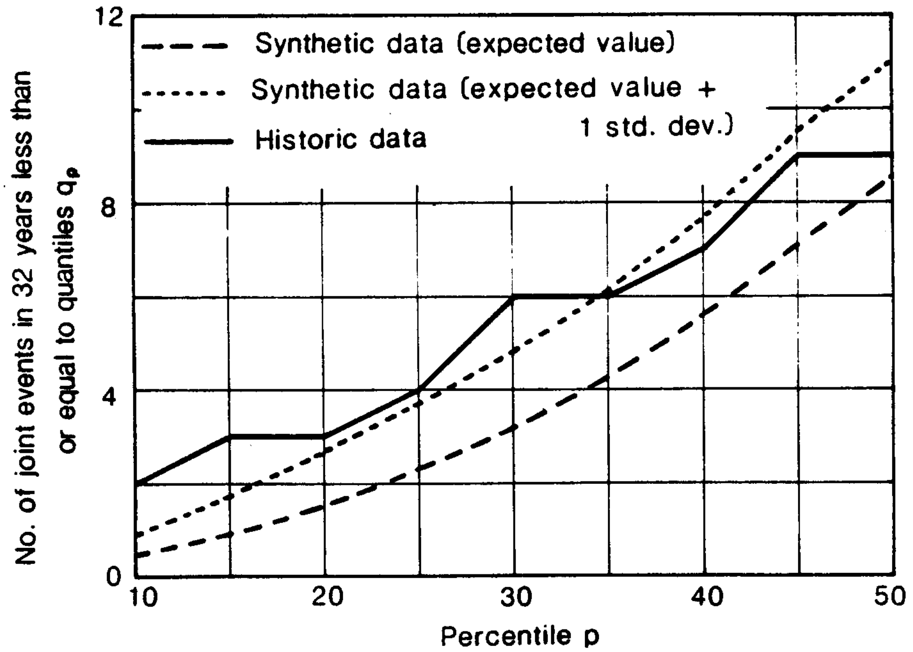


Figure 4.8 Occurrences of joint low events at Port Hardy and Eureka for historic and synthetic January data

The dotted line in the figure shows the expected number of synthetic events in 32 years plus one standard deviation and gives an indication of the variability in the number of synthetic joint occurrences in a 32-year period. Since the distribution of the number of joint events in 32-year periods of synthetic data is highly skewed at the low quantiles, the standard deviation as a measure of variability is somewhat misleading. Consequently, the actual numbers of joint events in the one hundred 32-year periods comprising the synthetic record are given in Table 4.8 for various quantile levels. For example, the table shows that of the one hundred 32-year periods in the synthetic sequences there were 60 periods with no events in which rainfall at both sites was concurrently below the 10 percent quantile and 37 periods in which there was one such event.

Table 4.8 Occurrences of Joint Low Events in January Synthetic Record for Port Hardy and Eureka

Percentile		Number of 32-year Synthetic Sequences Having n Joint Events Less Than or Equal to Quantiles q_p								
p	n	0	1	2	3	4	5	6	7	8
10		60	37	3						
15		36	41	20	2	1				
20		18	34	30	13	3	2			
25		10	22	24	26	12	2	3	0	1

For comparative purposes Figure 4.9 shows a similar plot for joint occurrences where both sites had rainfall greater than their respective 10, 15, 20,..., 50 percent quantiles. The sensitivity of the result in Figure 4.8 to changes in the cross correlation coefficient was investigated by repeating the experiment with values of the cross correlation of 0.2 and 0.4. The results of these experiments are shown in Figure 4.10.

Figures 4.8 and 4.10 demonstrate, as expected, that the model used cannot produce droughts as severe in areal extent as those found in the historic record from Port Hardy and Eureka. The statistical significance of these results is very difficult to determine in view of the low power of statistical tests, the short record available to us, and possible errors in fitting the marginal distributions. However, in expectation we would like the synthetic data to replicate reasonably closely the joint characteristics of the historic data. The model clearly fails to meet this goal. A useful measure of the lack of fit is the return period of severe joint events. The historic data indicates that joint events less than the 10 percent quantile have a return period of approximately 16 years. The synthetic data in contrast indicates a return period of 74 years. This is a major discrepancy from any viewpoint and could clearly have an important effect on projects with inter-connected components separated by large distances (> 1000 km). This of course raises the question of the practical significance of the result. Water and hydro power transfers currently take place between northern and southern California involving distances in excess of 1000 km. Hydropower transfers along the Pacific Intertie between Washington and southern California involve distances in excess of 1800 km. Similar hydropower transfers between southern Alaska and the lower states of the U.S.A. will become feasible in the future. Thus the results presented here have a future practical significance for the design of large projects.

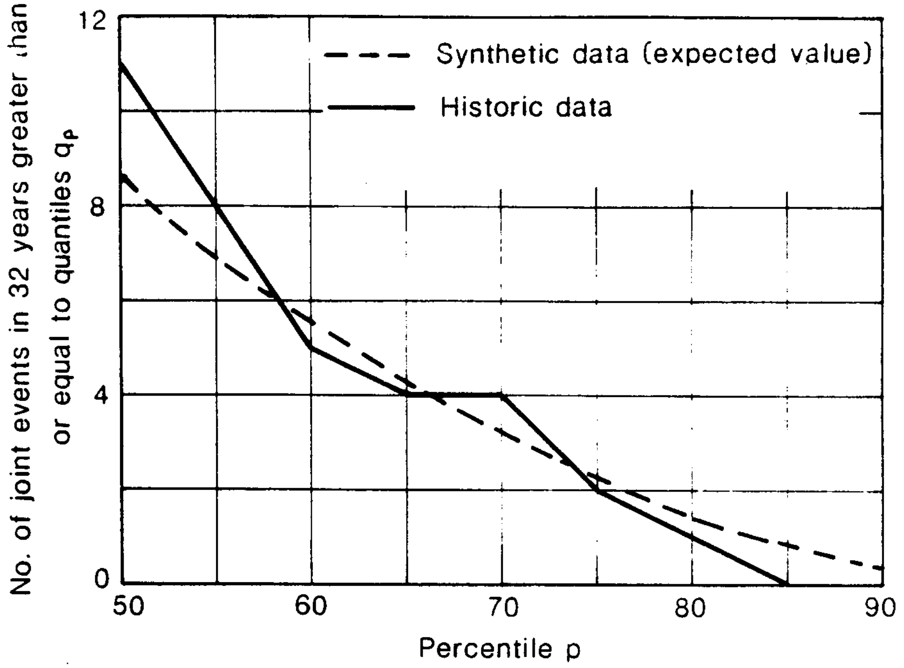


Figure 4.9 Occurrences of joint high events at Port Hardy and Eureka for historic and synthetic January data

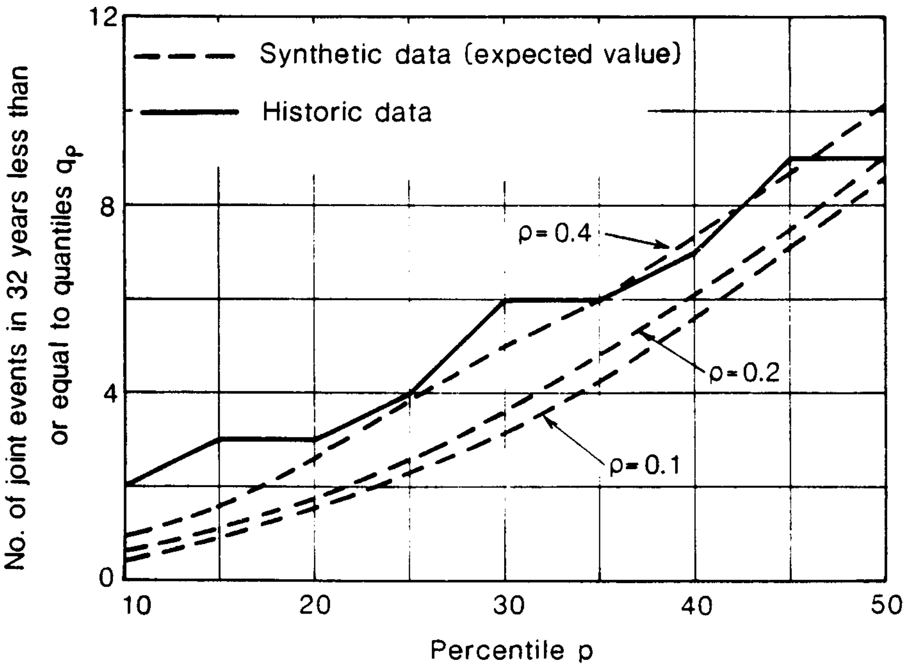


Figure 4.10 Sensitivity of occurrences of synthetic joint low events to changes in cross correlation

4.2.2 Evaluation of Dimensionality Effects

Evaluation of the conjectured dimensionality effect in multi-site modeling was again undertaken using 32 years of monthly January data from a network of seven sites in western Washington and southern British Columbia. These sites were Victoria, Vancouver, Sedro Woolley, Snoqualmie Falls, Longmire, Kid Valley and Centralia. The locations of the sites are shown in Figure 4.2. The basic statistics for the January data are given in Table 4.9 and the January cross correlation matrix is shown in Table 4.10. Inspection of Table 4.10 and Figure 4.2 shows that at this scale, there is no consistent variation of cross correlation with distance except that stations very close together (e.g. Centralia and Kid Valley) have much higher cross correlations than other stations. Scatterplots for selected pairs of stations are shown in Figures 4.11. These, in common with scatterplots for the other stations, show no obvious patterns of nonlinearity. The January serial correlations for the seven sites are all small (between -0.1 and +0.1) and thus, as in Section 4.2.1 the January data, after suitable transformation, will be assumed to be independent samples from a multivariate normal population.

Cumulative distribution functions for the seven sites are shown in Figure 4.12 along with fitted three parameter log normal (LN3) distributions. As in Section 4.2.1 the LN3 distributions were fitted using the mixed maximum likelihood/quantile method (Stedinger 1980). The parameters of the fitted distributions are given in Table 4.11.

After the appropriate LN3 transformation the data may be assumed to be samples from a multivariate normal and the serial independence of the data again allows use of the simple generation scheme:

$$\underline{y}_t = \underline{B}\,\underline{\varepsilon}_t + \underline{\mu}_y$$

Table 4.9 Basic Statistics for January Precipitation Data (in mm) at Seven Sites in the Pacific Northwest 1947-1978

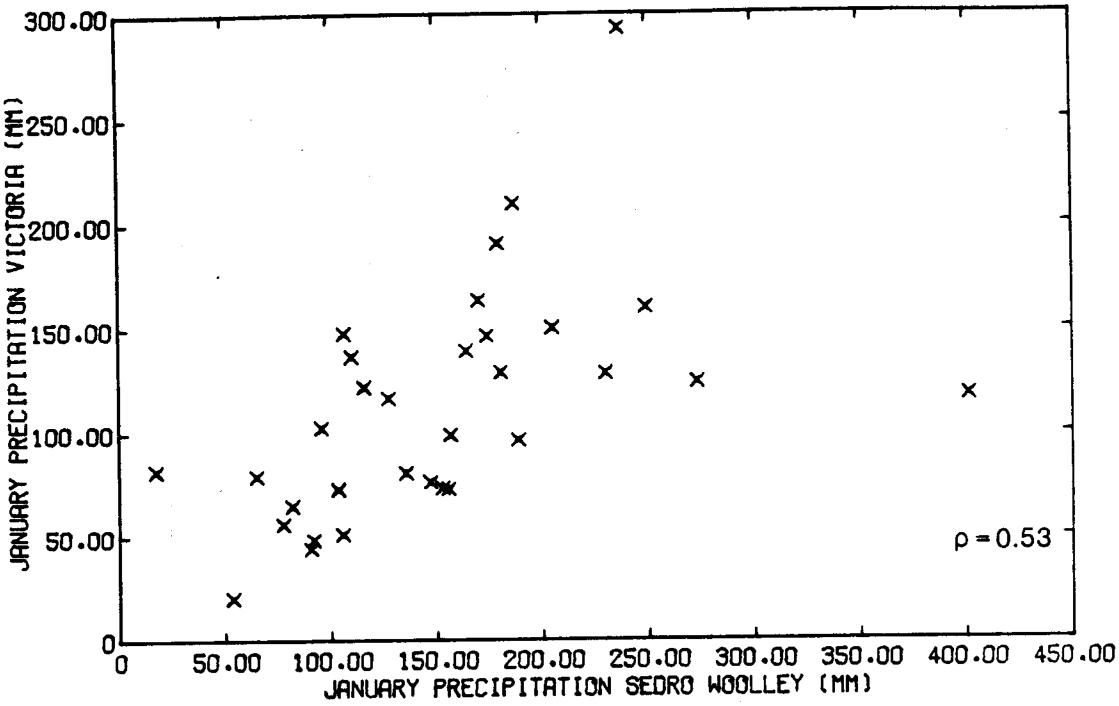
Station	Mean	St. Dev.	Skew	Maximum	Minimum
Victoria	112.4	54.4	1.07	293.0	20.4
Vancouver	152.1	57.6	-0.33	260.7	18.3
Sedro Woolley	151.0	73.8	1.11	401.3	17.5
Snoq. Falls	231.0	102.0	0.28	496.1	50.5
Longmire	346.4	176.6	0.24	681.5	59.7
Kid Valley	220.6	110.3	0.13	457.7	32.5
Centralia	186.3	86.7	-0.10	352.8	25.9

Table 4.10 Cross Correlations for January Precipitation Data at Seven Sites in the Pacific Northwest 1947-1978

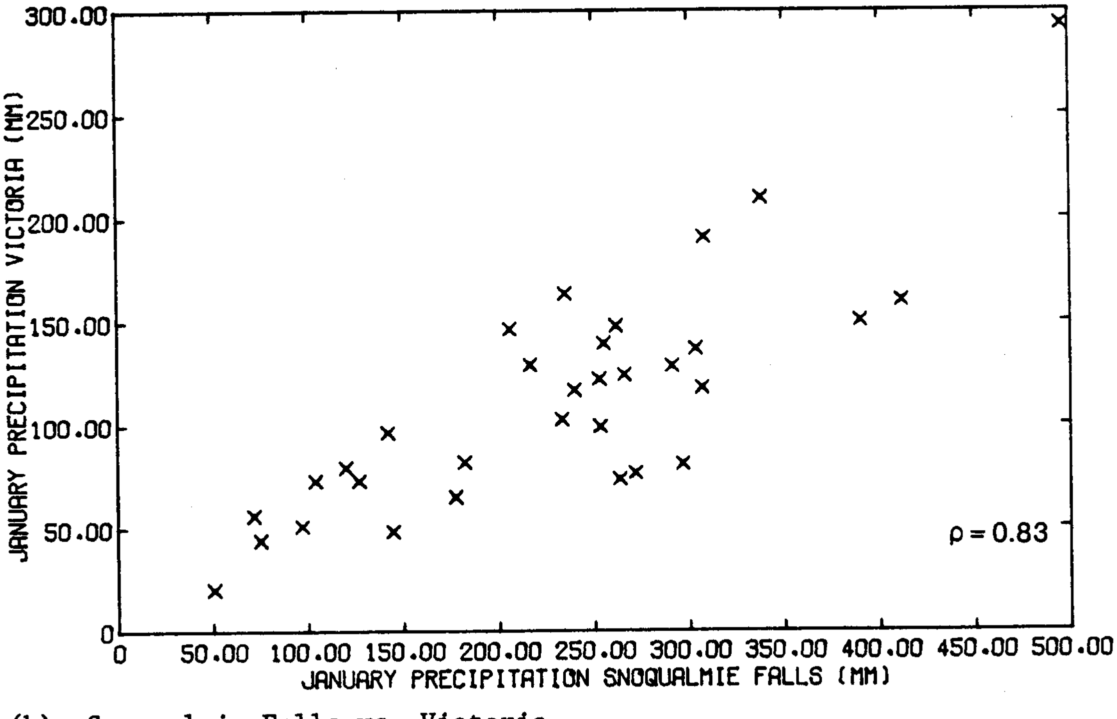
	Victoria	Vancouver	Sedro Woolley	Snoq. Falls	Longmire	Kid Valley	Centralia
Victoria	1.00	.66	.53	.83	.75	.75	.77
Vancouver	.66	1.00	.59	.69	.53	.61	.63
Sedro Woolley	.53	.59	1.00	.62	.72	.62	.60
Snoq. Falls	.83	.69	.62	1.00	.90	.92	.91
Longmire	.75	.53	.72	.90	1.00	.91	.89
Kid Valley	.75	.61	.62	.92	.91	1.00	.96
Centralia	.77	.63	.60	.91	.89	.96	1.00

Table 4.11 LN3 Parameter for January Precipitation Data at Seven Sites in the Pacific Northwest

Station	Transformation	μ_y	σ_y	a
Victoria	y = log(x - a)	5.305	0.252	-95.5
Vancouver	y = log(a - x)	5.852	0.163	504.7
Sedro Woolley	y = log(x - a)	5.661	0.239	-144.7
Snoqualmie Falls	y = log(x - a)	7.089	0.045	-971.6
Longmire	y = log(x - a)	7.314	0.116	-1165
Kid Valley	y = log(x - a)	7.951	0.0388	-2620
Centralia	y = log(a - x)	8.419	0.0191	4719

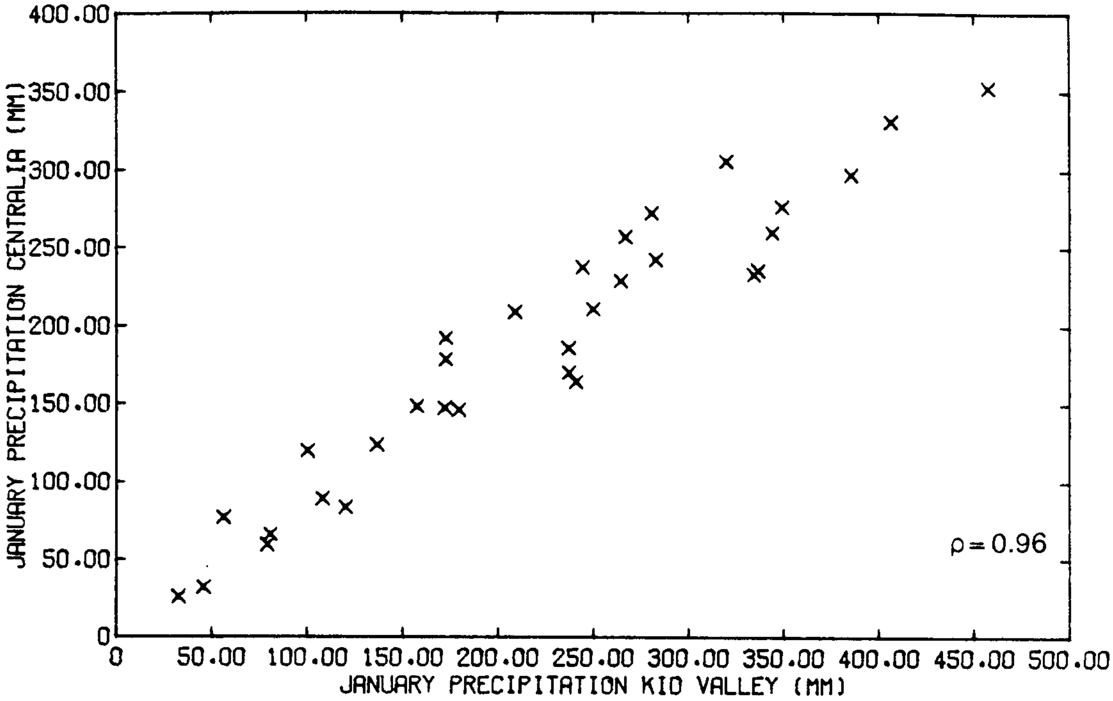


(a) Sedro Woolley vs. Victoria



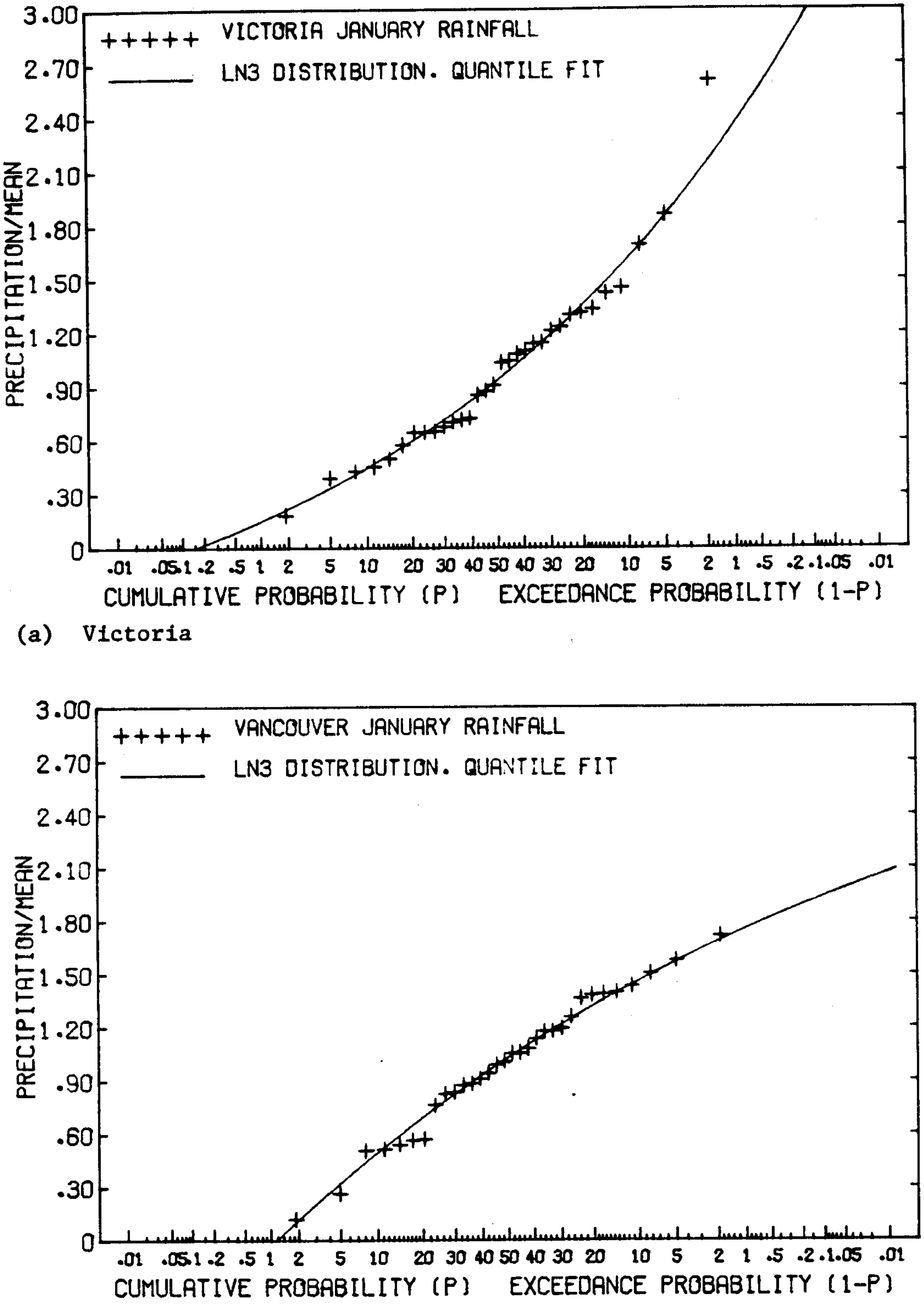
(b) Snoqualmie Falls vs. Victoria

Figure 4.11 Scatterplots for January precipitation data



(c) Kid Valley vs. Centralia

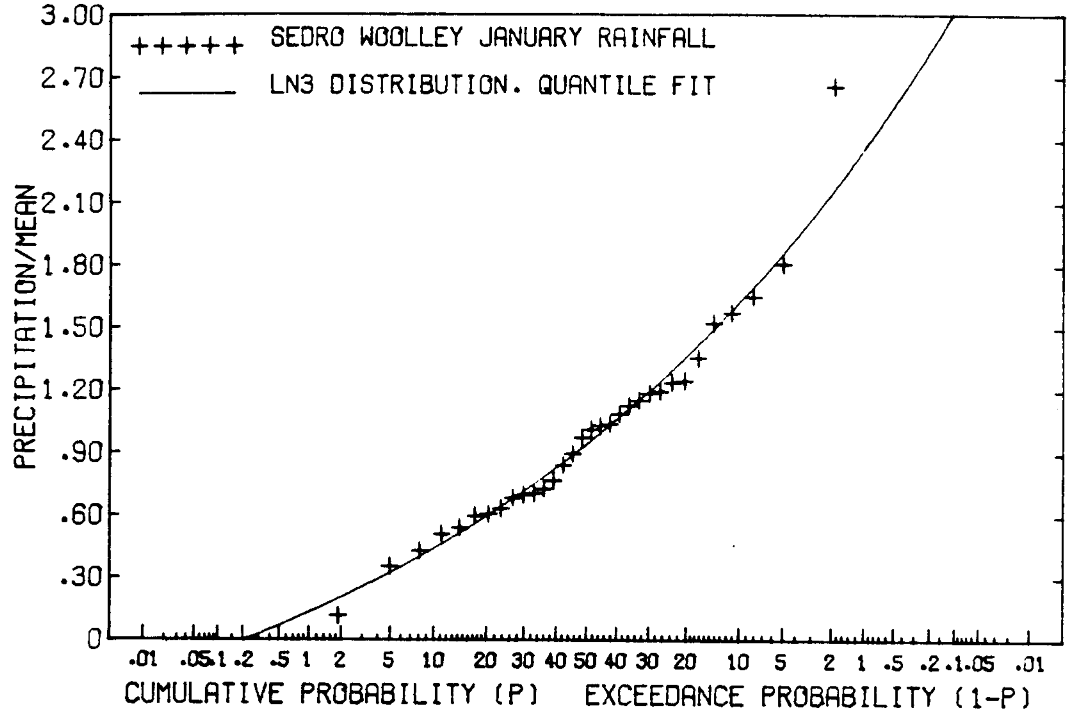
Figure 4.11 Continued



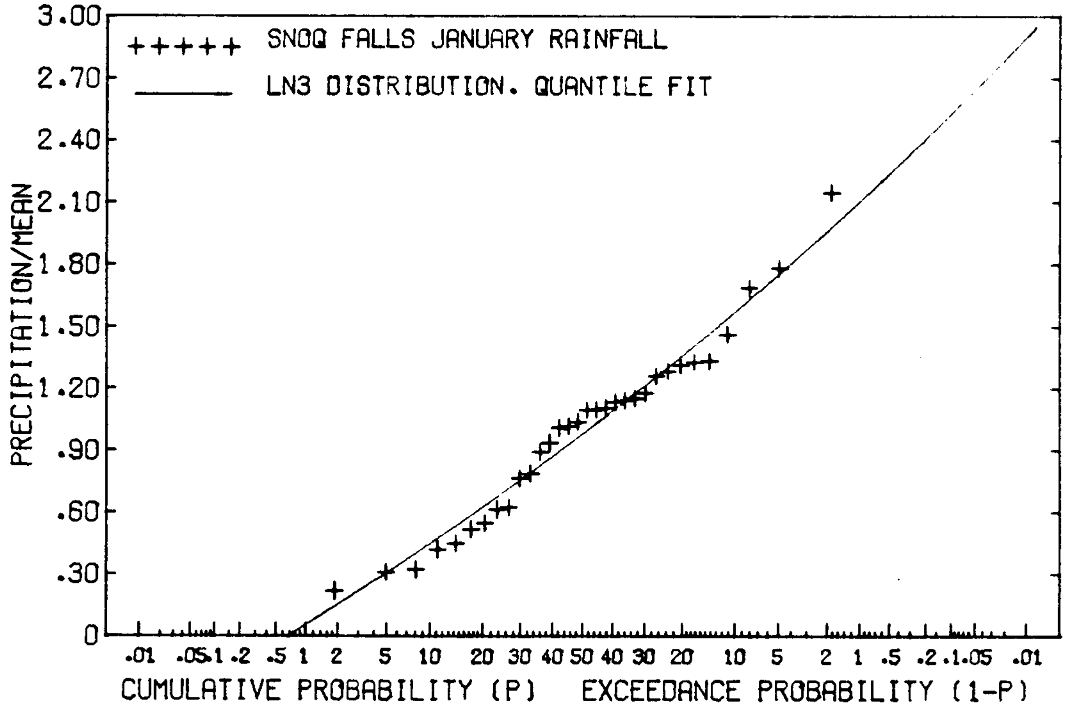
(a) Victoria

(b) Vancouver

Figure 4.12 CDF's for January precipitation data

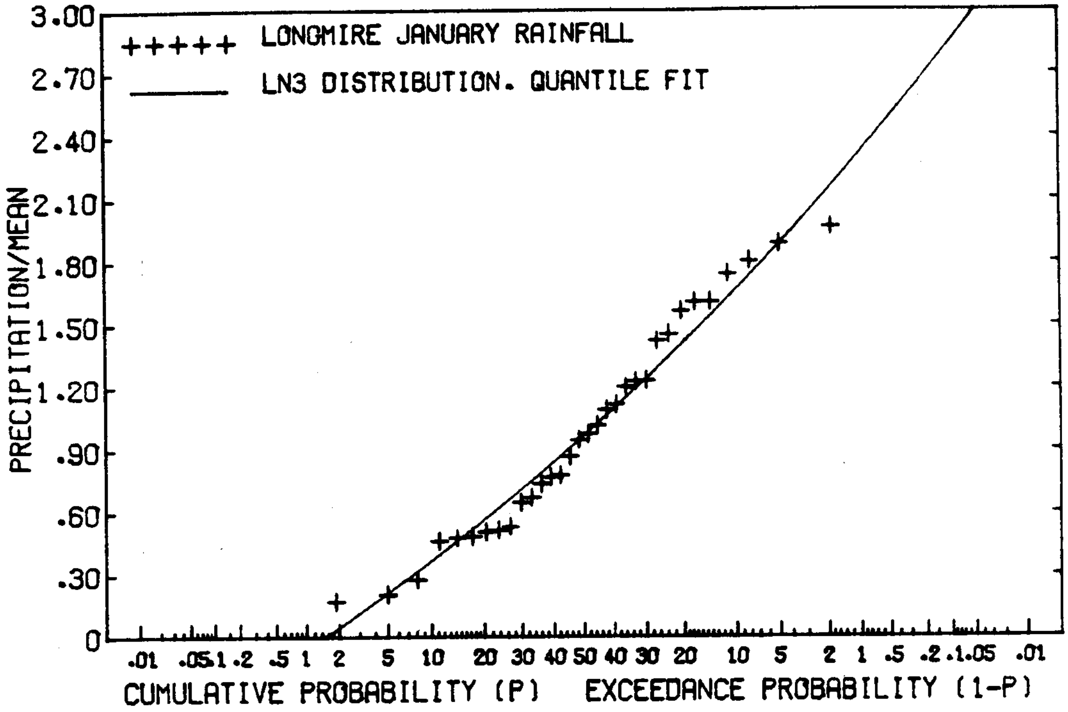


(c) Sedro Woolley

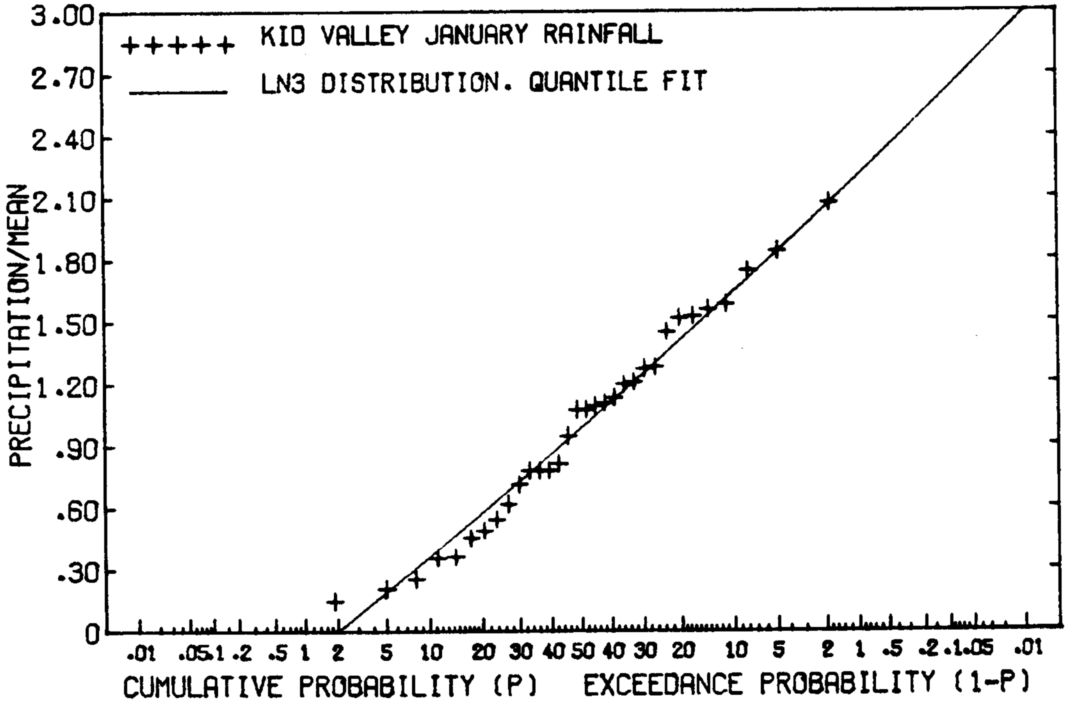


(d) Snoqualmie Falls

Figure 4.12 Continued

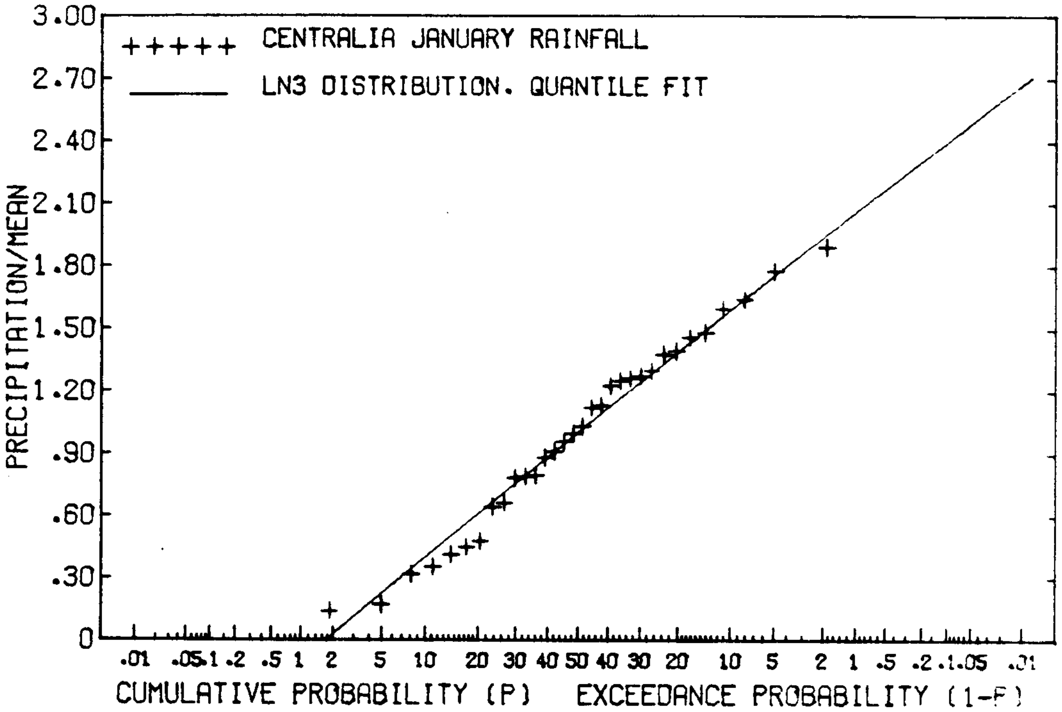


(e) Longmire



(f) Kid Valley

Figure 4.12 Continued



(g) Centralia

Figure 4.12 Continued

where the symbols are as defined in Section 4.2.1. Corrections to the cross correlation matrix to account for the log transformation were again found to be small with the largest correction being 0.024.

The Monte Carlo procedure for evaluating the "dimensionality" effect was very similar to that described in Section 4.2.1. For the seven sites 3200 years of synthetic January data were generated. The 10, 15, 20, 25,..., and 50 percent quantiles for the synthetic sequences at the individual sites were determined and then joint occurences were counted in which all seven sites had synthetic rainfall less than or equal to their respective 10, 15, 20,..., 50 percent quantiles. These counts were divided by 100 to give an estimate of the expected number of joint ocurrences in a period of 32 years. These data and the corresponding data from the historic record are shown in Figure 4.13. The actual count of the number of joint events in the one hundred 32-year periods comprising the synthetic record is also given in Table 4.12. The dotted line in Figure 4.13 shows the expected number of synthetic events in 32 years plus one standard deviation and gives an indication of the variability in the number of synthetic joint occurrences in a 32-year period.

Table 4.12 Occurrences of Joint Low Events in January Synthetic Record for Seven Sites in the Pacific Northwest

Percentile		Number of 32-year Synthetic Sequences Having n Joint Events Less Than or Equal to Quantiles q_p							
p	n	0	1	2	3	4	5	6	7
10		63	32	4	0	1			
15		37	44	15	2	2			
20		21	38	22	12	5	0	0	1
25		10	18	33	25	8	3	2	1

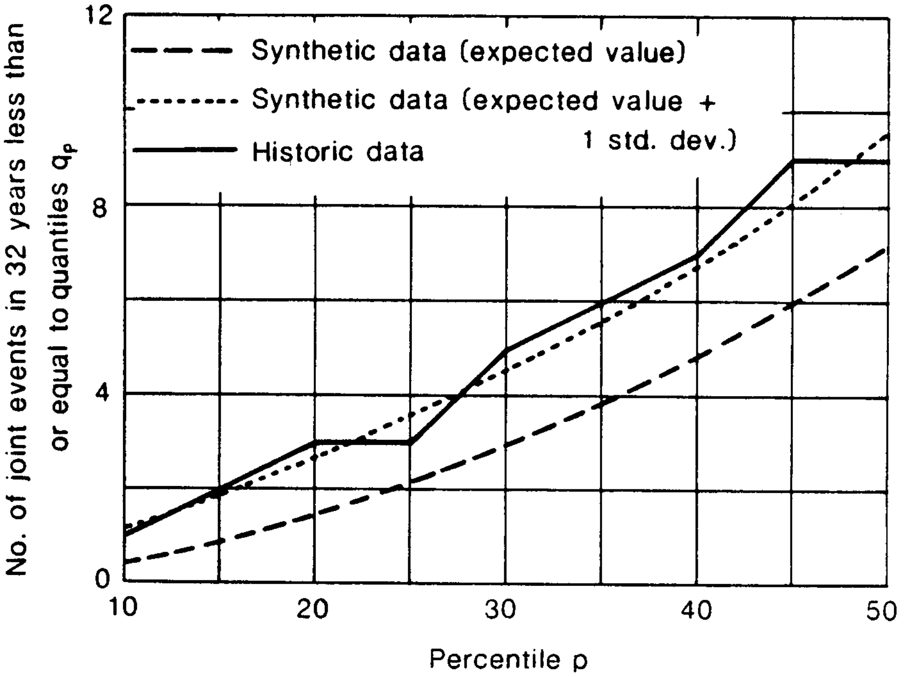


Figure 4.13 Occurrences of joint low events at seven sites in the Pacific Northwest

Figure 4.14 shows a similar plot for joint occurrences where all sites had rainfall greater than their respective 10, 15, 20,... 50 percent quantiles. Figure 4.13 shows that the model used does not, in expectation, produce droughts as severe as those encountered in practice. However, the significance of the result is again very difficult to assess particularly because of the high dimensionality of the problem and the short historic record.

The occurrence of joint low events at the 21 possible combinations of two sites is shown for the historic and synthetic data in Table 4.13. The data show an under simulation of joint low events at approximately 80 percent of the possible combinations of sites.

Although the under simulation for combinations of two sites is slight, the degree of under simulation increases as the number of sites in the analysis is increased. It is however unclear whether the increase in under simulation is a direct result of inabilities to reproduce correctly joint events at each pair of sites, or whether the hypothesized dimensionality effect does in fact occur.

The sensitivity of the result in Figure 4.13 to changes in the cross correlation matrix was investigated by increasing all cross correlations less than 0.75 to 0.75 and repeating the experiments. The results are shown in Figure 4.15. Even these relatively large increases in values of the cross correlation coefficients failed to rectify completely the problem of under simulation either at pairs of sites or at all seven sites. As may be noted from Figure 4.14, such adjustments to the correlation matrix must be treated with caution since they further detract from the model's ability to reproduce joint high events.

The practical implications of these results are again difficult to assess but would appear to depend on a number of features including the dimensionality of the problem, the values of the cross correlation

Table 4.13 Occurrences of Joint Low Events at Pairs of Sites in the Pacific Northwest

Percentile		Expected Number of Joint Events in 32 Years Less Than or Equal to Quantiles q_p																				
p	Sites	1/2	1/3	1/4	1/5	1/6	1/7	2/3	2/4	2/5	2/6	2/7	3/4	3/5	3/6	3/7	4/5	4/6	4/7	5/6	5/7	6/7
10	Synthetic	1.47	1.05	1.88	1.60	1.67	1.68	1.24	1.53	1.13	1.35	1.38	1.34	1.57	1.35	1.33	2.25	2.34	2.22	2.27	2.17	2.63
	Historic	1	1	3	2	2	2	2	2	3	2	2	2	2	2	1	3	3	2	3	3	3
15	Synthetic	2.36	1.85	3.06	2.67	2.70	2.82	2.24	2.55	2.03	2.28	2.26	2.34	2.65	2.16	2.22	3.58	3.66	3.52	3.57	3.46	3.92
	Historic	2	2	4	3	3	3	3	3	3	3	3	2	2	2	2	4	4	4	4	4	5
20	Synthetic	3.58	2.98	4.36	3.92	3.83	3.93	3.25	3.61	2.91	3.26	3.30	3.35	3.79	3.28	3.21	4.83	4.91	4.98	4.79	4.69	5.35
	Historic	5	5	6	5	6	7	4	5	5	4	5	4	4	4	4	6	5	6	5	5	6
25	Synthetic	4.70	3.92	5.56	5.10	5.07	5.24	4.37	4.88	3.97	4.38	4.43	4.35	4.97	4.27	4.23	6.19	6.44	6.42	6.23	6.17	6.99
	Historic	5	5	6	6	7	7	4	5	5	5	5	4	5	4	4	7	7	7	7	7	8

Site Key:

1 Victoria	2 Vancouver	3 Sedro Woolley	4 Snoqualmie Falls
5 Longmire	6 Kid Valley	7 Centralia	

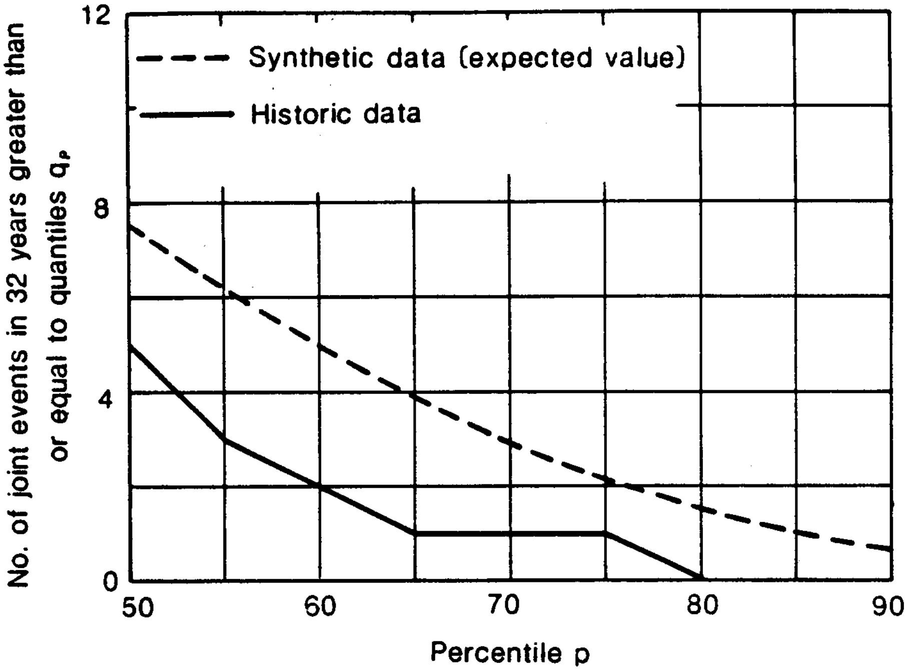


Figure 4.14 Occurrences of joint high events at seven sites in the Pacific Northwest

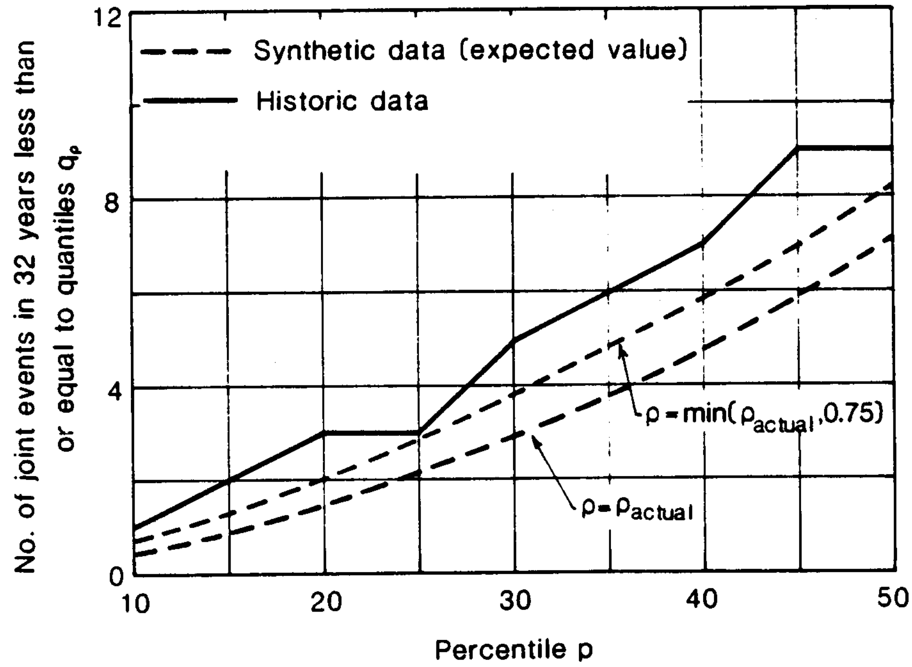


Figure 4.15 Sensitivity of occurrences of synthetic joint low events at seven sites to changes in cross correlation

coefficients and the degree of development of the water resource. Thus the problems outlined in this section may have little or no effect on a study of say seven sites on the Skagit River but could have serious implications for a study of seven widely separated sites in the Columbia basin.

The most disturbing feature of the results presented here is that very few of the synthetic sequences contained more joint drought occurrences than the historic sequence. For example, Figure 4.13 and Table 4.12 show that at the 20 percent quantile, only 6 percent of the synthetic sequences had more occurences of joint low events than the historic data. This could result in either a substantial underdesign of water resource facilities, or equivalently a substantial overestimate of the reliability of existing facilities. A true assessment of these difficulties can only be made by studying the design or operation of actual water resource systems. This task is beyond the scope of the current research.

4.3 Concluding Remarks

Consideration of large scale atmospheric circulation patterns and analysis of a limited amount of historic data both tend to support the contention that inter-station precipitation relationships are nonlinear and that the current suite of stochastic models do not adequately represent the spatial characteristics of drought. It will be recalled that the analyses are based on only 32 years of January data from a limited number of sites. Thus it could be argued that little has been shown other than that the spatial characteristics of the selected data were not reproduced by the simple multi-site generation scheme. From a statistical viewpoint, the hypothesis that the selected data come from the model in Equation 4.2, probably could not be rejected at any reasonable significance level. However, this ignores three important considerations:

(1) Standard statistical tests cannot make use of the qualitative information relating atmospheric conditions to nonlinearities in the interstation relationships.

(2) Statistical tests are notoriously weak in conditions such as those encountered in this work.

(3) The results of this work indicate that use of the conventional linear models may lead to underdesign of water resource facilities.

It is clear at least for the two site case of Port Hardy and Eureka that currently available models seriously under simulate joint drought occurrences. The principal difficulty with existing models is that there is no mechanism to control directly the areal extent of drought. As mentioned in Chapter 3 one possible approach to overcome this problem is through the use of multivariate mixture models with conditioning on the state of atmospheric circulation. The possibility of using atmospheric pressure data for this purpose is explored in the next chapter.

5.0 RELATIONSHIPS BETWEEN PRECIPITATION AND ATMOSPHERIC CIRCULATION

In the previous chapters I have made use of known qualitative relationships between precipitation and patterns of atmospheric circulation. In this chapter I attempt to put these relationships on a more quantitative footing by analyzing concurrent precipitation and 500 mb geopotential height data. In particular, I wish to determine if evidence exists to support the hypothesis that precipitation has a mixed distribution which may be conditioned on the state of the atmospheric circulation (i.e. zonal or meridional circulation).

As in the previous chapters I will restrict my attention to the west coast of the U.S.A. with particular emphasis on conditions in the Pacific Northwest. The analysis will consider relationships between atmospheric circulation and precipitation both at a point and over an extensive area along the west coast.

5.1 Data Requirements and Data Sources

Two types of data are of interest--precipitation data and atmospheric pressure data. The latter are appropriate for inferring the nature of atmospheric circulation in the mid-latitudes.

5.1.1 Precipitation Data

The precipitation data obtained for this study were described in Chapter 4. Although the analyses in Chapter 4 were restricted to monthly and annual data, daily data were obtained on magnetic tape for a number of the stations. This allows analysis at shorter intervals such as daily or 5-day as is deemed necessary.

5.1.2 Pressure Data

Basic meteorologic data comprising temperature, pressure, relative humidity, etc., are collected and processed by the National Meteorologic Center (NMC) and the National Center for Atmospheric Research (NCAR). Data are obtained from a wide variety of sources including upper air soundings, shipping, aircraft, and satellites. The data series available from NMC and NCAR are summarized by Jenne (1975).

It is clear from basic meteorological theory and observations that the atmospheric circulation is driven by and can be inferred from the three dimensional pressure gradient fields in the atmosphere (see for example, Wallace and Hobbs 1977 and Holton 1979). NMC uses models of the atmospheric circulation with the observed meteorologic data to produce best estimates of pressure data at various levels on an octagonal grid of 1977 points covering the whole of the northern hemisphere, north of about 15 degrees north. The octagonal NMC grid is shown in Figure 5.1. Pressure data available include surface pressure (mb) and the geopotential height on constant pressure surfaces at 1000 mb, 850 mb, 700 mb, 500 mb, 100 mb and at other levels. The data are thus available to describe the complete three dimensional structure of the atmospheric pressure field.

For a study of this nature such detail is not necessary. As discussed in Section 3.2 the features of primary interest are the track of the jet stream and the general pattern of circulation as it affects conditions at the ground. Surface pressures and the direction and magnitude of surface winds are both influenced by surface friction and topography. Moreover, surface pressure distributions cannot be used to detect the jet stream except perhaps by following the movement of frontal systems. The jet stream is generally at about the 300 mb level and at that level the track follows the line of maximum geopotential height gradient on that pressure surface. Unfortunately at

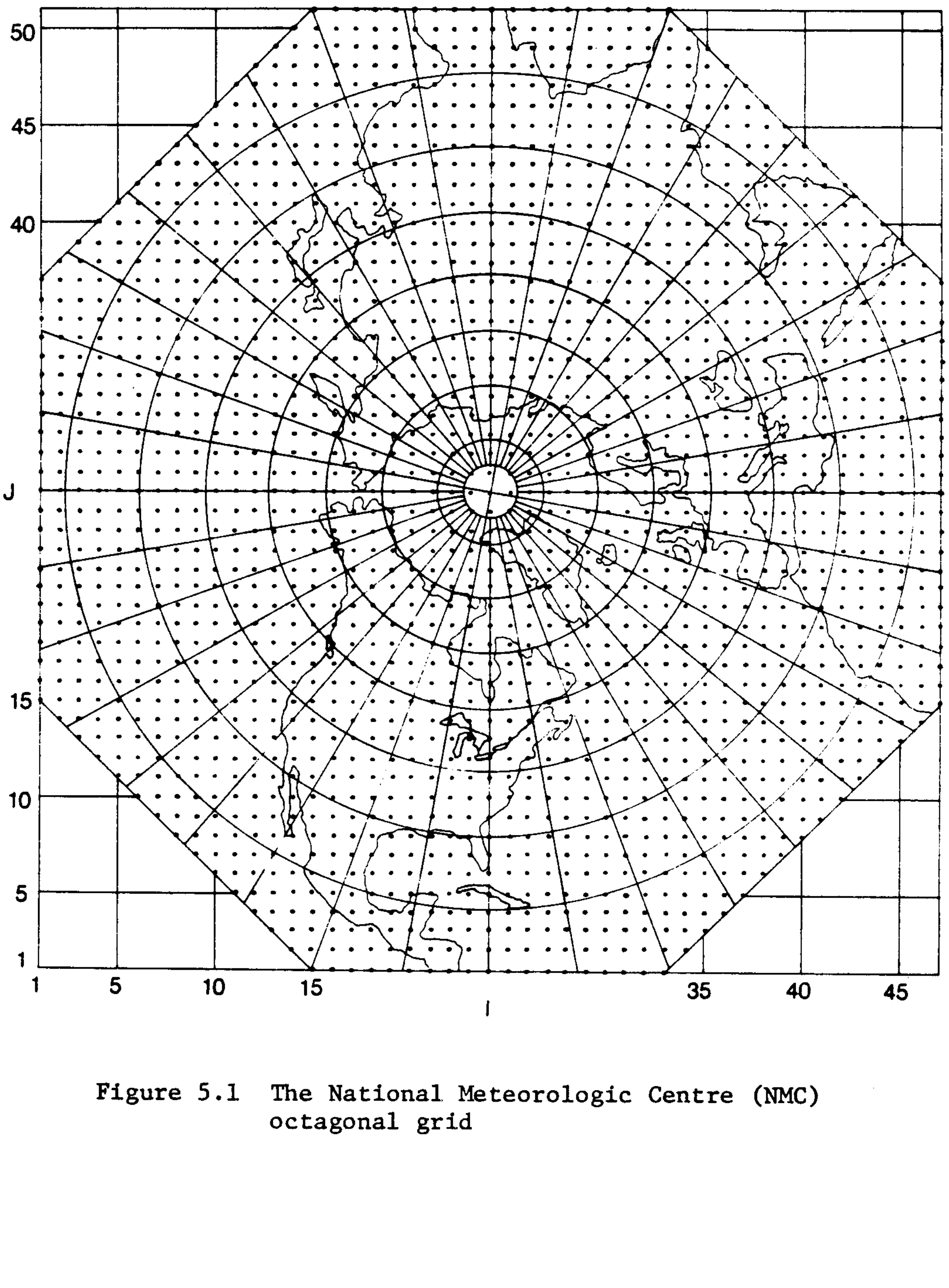


Figure 5.1 The National Meteorologic Centre (NMC) octagonal grid

300 mb the pattern of important surface features is lost, and it becomes difficult or impossible, for example, to follow the movement of frontal systems. A suitable compromise is the frequently used 500 mb data, i.e. data giving the geopotential height of the 500 mb pressure surface. This can be used both to infer the approximate position of the jet stream and to track the movement of major disturbances. Thus the 500 mb data are probably the single most useful data series for indicating the general pattern of atmospheric circulation; these data series have been used extensively in meteorological research. Additionally, the 500 mb series is the longest upper level pressure series available with continuous records dating from January 1946.

Analysis of 500 mb, 1000 mb and surface pressures by Blackmon, et al. (1979) has shown that the atmosphere over the Pacific coast of North America is generally equivalent barotropic. This in essence means that pressure perturbations in the upper atmosphere are in phase with those at the surface, e.g., a high pressure (low pressure) area at the 500 mb level overlies a high pressure (low pressure) area over the ground. This is significant for this project since it confirms that conditions in the upper atmosphere at 500 mb are indeed a good indication of what is happening at the surface vertically below. In other parts of North America, for example the Midwest, disturbances at the earth's surface are more loosely coupled to disturbances in the upper atmosphere so that the conditions at 500 mb may not accurately reflect conditions at the surface.

A second significant feature displayed by Blackmon et al. is the high correlation between the mean monthly 500 mb height and the mean monthly 1000-500 mb thickness over the west coast. In most areas of interest for this work the correlation coefficient exceeds 0.9. The difference in geopotential height between two pressure surfaces, referred to as the thickness, is directly related to the mean temperature of that layer. Thus the 500 mb height may turn out to be a useful indicator of the nature of precipitation (i.e., rain or snow) in addition to indicating the pattern of circulation.

The observed barotropic structure of the atmosphere largely eliminates the need for considering the three dimensional structure of the pressure field in assessing the circulation pattern. Data at a single level, above the influence of surface friction, such as the 500 mb level, provide adequate information for this study.

The 500 mb data set obtained for this study starts in January 1946 and runs through February 1979. The data are at a daily interval for the first nine years changing to twice daily in 1955. The data series is essentially complete with little missing data. The few data that are missing were filled in by linearly interpolating between data on succeeding and preceding days at the same grid point. The twice daily data were then averaged to produce a daily series with 33 years of data. The data actually used for analysis are from October 1946 through September 1978, i.e., 32 complete water years of data concurrent with the available precipitation data.

Obviously data for the entire hemisphere are not necessary in this study. The height of the 500 mb surface is only used to infer the nature of atmospheric circulation directly affecting the climate of the Pacific Northwest. Accordingly, only a small number of stations from the NMC grid were chosen for detailed analysis. Basic statistics for mean monthly 500 mb geopotential height data from selected stations are shown in Table 5.1 and monthly cross correlations for selected stations are shown in Table 5.2.

(text continues on p.103)

Table 5.1 Basic Statistics for Mean Monthly 500 mb Geopotential Height Data
(in meters)

SITE (13,15)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5785.9	5736.8	5699.3	5678.1	5666.0	5649.7	5676.0	5732.4	5802.3	5861.3	5852.3	5826.6
STDEV	31.3	43.1	47.0	45.8	52.7	46.7	55.2	32.9	25.9	21.2	19.1	21.5
SKEW	-.26	-.12	-.19	.02	-.49	-.04	-.97	-.40	-.13	-.31	-.56	-.14
MAXIMUM	5849.0	5814.0	5783.1	5756.4	5751.4	5744.2	5770.5	5793.4	5851.9	5898.0	5883.9	5864.4
MINIMUM	5706.3	5641.9	5602.0	5590.1	5522.3	5537.0	5499.0	5647.9	5754.1	5816.1	5799.1	5775.5

SITE (14,16)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5742.7	5678.7	5634.7	5613.9	5610.0	5588.2	5622.0	5686.6	5759.5	5828.9	5817.1	5795.9
STDEV	37.2	49.8	58.3	56.3	57.9	44.5	61.1	37.3	29.4	27.9	25.2	26.2
SKEW	-.14	.12	-.13	-.12	-.29	-.14	-.73	-.60	.09	-.55	-.23	.06
MAXIMUM	5806.4	5790.5	5726.7	5707.4	5702.7	5676.5	5705.9	5745.5	5817.6	5882.0	5865.4	5851.4
MINIMUM	5673.6	5564.1	5528.9	5507.5	5452.7	5475.5	5449.9	5588.8	5698.3	5765.9	5767.0	5742.5

SITE (15,17)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5675.2	5598.3	5549.3	5529.1	5540.3	5515.9	5560.3	5632.9	5704.5	5774.8	5768.6	5748.6
STDEV	46.4	58.9	64.9	69.9	57.7	41.2	56.7	39.0	36.3	33.8	29.8	37.3
SKEW	-.08	.11	-.06	-.30	-.02	.39	-.40	-.36	-.26	-.33	.19	-.07
MAXIMUM	5782.4	5722.7	5652.3	5649.8	5643.9	5603.3	5638.9	5710.9	5773.3	5842.6	5833.4	5820.6
MINIMUM	5576.1	5439.9	5437.3	5373.8	5396.7	5437.4	5428.8	5533.4	5628.5	5706.1	5712.0	5669.6

Table 5.1 Continued

SITE (16,18)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5595.4	5511.8	5460.5	5442.3	5466.9	5446.5	5502.3	5587.4	5654.4	5721.1	5721.0	5690.6
STDEV	53.0	63.7	64.8	77.6	57.3	43.8	43.9	40.6	42.1	34.3	32.8	44.2
SKEW	.03	.12	.05	-.31	.31	1.12	.03	.11	.26	.06	.30	-.08
MAXIMUM	5726.9	5658.7	5590.9	5580.2	5571.4	5561.5	5580.9	5679.3	5737.7	5797.3	5791.4	5771.4
MINIMUM	5474.6	5346.2	5335.7	5262.3	5353.8	5375.9	5418.2	5507.9	5589.3	5669.1	5670.5	5597.5

SITE (16,19)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5540.5	5461.2	5411.1	5403.8	5427.2	5412.3	5466.0	5555.5	5625.6	5691.6	5695.8	5657.7
STDEV	59.5	66.2	61.4	81.5	61.5	51.5	43.8	40.8	49.3	40.8	36.3	47.3
SKEW	-.00	.19	.17	.21	.21	1.40	.12	.10	.68	.11	-.01	.20
MAXIMUM	5676.5	5610.3	5546.3	5573.7	5548.5	5574.3	5543.7	5644.4	5723.2	5786.2	5757.2	5759.1
MINIMUM	5420.3	5306.1	5293.1	5235.7	5308.8	5338.6	5374.5	5474.3	5555.5	5620.0	5619.7	5553.5

SITE (16,20)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5480.5	5410.5	5361.7	5368.8	5385.1	5378.3	5428.7	5519.6	5597.1	5669.6	5673.9	5619.5
STDEV	64.7	63.5	57.3	82.8	64.4	59.4	45.6	41.5	54.2	44.8	42.1	52.2
SKEW	.23	.22	.22	.73	.05	1.46	-.04	.01	.85	.15	-.00	.75
MAXIMUM	5616.7	5537.9	5496.4	5580.0	5512.3	5582.1	5515.9	5602.4	5719.1	5763.1	5751.7	5744.8
MINIMUM	5373.9	5287.9	5253.9	5240.7	5255.1	5283.3	5327.4	5443.6	5517.6	5592.4	5578.8	5529.3

Table 5.1 Continued

SITE (16,21)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5420.0	5363.1	5314.3	5336.5	5342.4	5344.0	5393.1	5483.6	5572.3	5651.2	5652.9	5577.0
STDEV	66.4	57.2	54.7	85.7	64.7	66.3	50.1	42.4	54.1	43.3	46.8	56.4
SKEW	.56	.01	.06	1.03	.12	1.25	-.09	-.41	.80	.28	.28	.77
MAXIMUM	5566.4	5467.9	5433.9	5575.8	5480.4	5572.7	5500.0	5546.8	5694.5	5730.8	5754.1	5737.1
MINIMUM	5324.9	5231.7	5210.5	5223.9	5204.3	5234.7	5285.2	5391.5	5493.3	5577.1	5549.5	5487.2

SITE (16,22)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5369.7	5325.2	5275.2	5311.7	5304.7	5312.6	5363.6	5454.8	5556.4	5638.2	5632.6	5536.5
STDEV	64.7	53.1	56.4	89.3	63.9	70.4	56.0	43.2	49.7	36.3	47.4	56.7
SKEW	.74	-.23	-.25	1.19	.45	.81	.10	-.67	.55	.32	.35	.58
MAXIMUM	5533.7	5405.5	5372.5	5549.8	5480.2	5534.2	5481.6	5522.5	5657.1	5718.4	5735.6	5705.3
MINIMUM	5283.3	5214.6	5148.5	5189.8	5181.0	5195.4	5252.6	5360.8	5491.1	5566.3	5537.9	5423.0

SITE (16,23)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5333.4	5296.9	5247.5	5294.3	5274.9	5286.2	5340.4	5435.1	5547.3	5630.1	5613.0	5502.3
STDEV	59.8	57.8	61.8	91.6	64.0	70.5	61.9	45.5	42.6	31.0	45.0	51.8
SKEW	.76	.35	-.39	1.27	.56	.48	.48	-.40	.15	.01	.13	.33
MAXIMUM	5498.3	5409.9	5354.9	5530.0	5460.6	5475.0	5490.3	5519.2	5628.2	5701.6	5702.5	5647.5
MINIMUM	5237.6	5209.6	5111.2	5173.8	5169.3	5163.9	5237.2	5344.4	5464.0	5568.5	5536.2	5385.9

Table 5.1 Continued

SITE (15,24)

	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
MEAN	5325.7	5290.5	5256.1	5300.9	5265.0	5292.6	5337.7	5409.7	5528.4	5621.1	5612.8	5489.8
STDEV	54.1	78.0	78.3	100.4	70.6	71.0	76.2	55.4	41.9	36.2	45.4	57.5
SKEW	.29	1.06	-.33	1.19	.39	.75	1.11	-.11	-.10	.09	-.09	-.13
MAXIMUM	5455.3	5462.9	5382.6	5609.9	5436.3	5489.9	5580.5	5527.9	5622.0	5702.6	5694.1	5612.0
MINIMUM	5233.6	5196.3	5083.8	5151.0	5160.8	5182.9	5230.3	5300.9	5419.4	5547.4	5522.0	5345.0

Table 5.2 Cross Correlations for Mean Monthly 500 mb Geopotential Height Data

MONTHLY CORRELATION MATRIX OCT

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.86	.52	.22	.02	-.15	-.28	-.35	-.35	-.42
(14,16)	.86	1.00	.87	.64	.43	.21	.02	-.12	-.21	-.34
(15,17)	.52	.87	1.00	.92	.77	.58	.37	.18	.04	-.14
(16,18)	.22	.64	.92	1.00	.94	.81	.63	.45	.30	.07
(16,19)	.02	.43	.77	.94	1.00	.95	.84	.70	.56	.34
(16,20)	-.15	.21	.58	.81	.95	1.00	.96	.87	.76	.56
(16,21)	-.28	.02	.37	.63	.84	.96	1.00	.97	.90	.73
(16,22)	-.35	-.12	.18	.45	.70	.87	.97	1.00	.97	.84
(16,23)	-.35	-.21	.04	.30	.56	.76	.90	.97	1.00	.91
(15,24)	-.42	-.34	-.14	.07	.34	.56	.73	.84	.91	1.00

MONTHLY CORRELATION MATRIX NOV

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.91	.69	.48	.39	.28	.11	-.14	-.33	-.37
(14,16)	.91	1.00	.92	.77	.69	.57	.36	.02	-.30	-.46
(15,17)	.69	.92	1.00	.95	.90	.80	.59	.20	-.20	-.45
(16,18)	.48	.77	.95	1.00	.98	.91	.72	.35	-.07	-.38
(16,19)	.39	.69	.90	.98	1.00	.97	.83	.49	.07	-.27
(16,20)	.28	.57	.80	.91	.97	1.00	.94	.67	.27	-.08
(16,21)	.11	.36	.59	.72	.83	.94	1.00	.88	.56	.22
(16,22)	-.14	.02	.20	.35	.49	.67	.88	1.00	.88	.62
(16,23)	-.33	-.30	-.20	-.07	.07	.27	.56	.88	1.00	.89
(15,24)	-.37	-.46	-.45	-.38	-.27	-.08	.22	.62	.89	1.00

Table 5.2 Continued

MONTHLY CORRELATIØN MATRIX DEC

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.95	.87	.80	.73	.61	.43	.21	-.01	-.30
(14,16)	.95	1.00	.97	.90	.85	.72	.52	.25	-.01	-.33
(15,17)	.87	.97	1.00	.97	.93	.82	.62	.33	.04	-.32
(16,18)	.80	.90	.97	1.00	.98	.88	.68	.38	.09	-.33
(16,19)	.73	.85	.93	.98	1.00	.96	.81	.55	.27	-.18
(16,20)	.61	.72	.82	.88	.96	1.00	.94	.75	.49	.02
(16,21)	.43	.52	.62	.68	.81	.94	1.00	.93	.74	.30
(16,22)	.21	.25	.33	.38	.55	.75	.93	1.00	.94	.59
(16,23)	-.01	-.01	.04	.09	.27	.49	.74	.94	1.00	.80
(15,24)	-.30	-.33	-.32	-.33	-.18	.02	.30	.59	.80	1.00

MONTHLY CORRELATIØN MATRIX JAN

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.91	.68	.43	.25	.03	-.19	-.34	-.41	-.43
(14,16)	.91	1.00	.92	.75	.61	.40	.16	-.06	-.22	-.40
(15,17)	.68	.92	1.00	.94	.86	.69	.43	.18	-.04	-.34
(16,18)	.43	.75	.94	1.00	.96	.83	.60	.33	.08	-.29
(16,19)	.25	.61	.86	.96	1.00	.95	.78	.55	.30	-.09
(16,20)	.03	.40	.69	.83	.95	1.00	.94	.78	.57	.19
(16,21)	-.19	.16	.43	.60	.78	.94	1.00	.95	.81	.48
(16,22)	-.34	-.06	.18	.33	.55	.78	.95	1.00	.95	.72
(16,23)	-.41	-.22	-.04	.08	.30	.57	.81	.95	1.00	.88
(15,24)	-.43	-.40	-.34	-.29	-.09	.19	.48	.72	.88	1.00

Table 5.2 Continued

MONTHLY CORRELATION MATRIX FEB

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.96	.83	.56	.40	.19	-.05	-.25	-.40	-.40
(14,16)	.96	1.00	.94	.73	.57	.35	.11	-.12	-.29	-.36
(15,17)	.83	.94	1.00	.91	.79	.60	.36	.12	-.08	-.26
(16,18)	.56	.73	.91	1.00	.95	.82	.63	.41	.21	-.12
(16,19)	.40	.57	.79	.95	1.00	.95	.82	.64	.43	.07
(16,20)	.19	.35	.60	.82	.95	1.00	.95	.82	.64	.27
(16,21)	-.05	.11	.36	.63	.82	.95	1.00	.95	.82	.49
(16,22)	-.25	-.12	.12	.41	.64	.82	.95	1.00	.95	.68
(16,23)	-.40	-.29	-.08	.21	.43	.64	.82	.95	1.00	.83
(15,24)	-.40	-.36	-.26	-.12	.07	.27	.49	.68	.83	1.00

MONTHLY CORRELATION MATRIX MAR

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.92	.59	.17	-.04	-.21	-.36	-.48	-.53	-.51
(14,16)	.92	1.00	.84	.47	.28	.09	-.09	-.24	-.33	-.38
(15,17)	.59	.84	1.00	.85	.72	.54	.35	.17	.05	-.12
(16,18)	.17	.47	.85	1.00	.95	.82	.66	.50	.36	.11
(16,19)	-.04	.28	.72	.95	1.00	.96	.86	.72	.58	.32
(16,20)	-.21	.09	.54	.82	.96	1.00	.97	.87	.75	.50
(16,21)	-.36	-.09	.35	.66	.86	.97	1.00	.96	.88	.67
(16,22)	-.48	-.24	.17	.50	.72	.87	.96	1.00	.97	.80
(16,23)	-.53	-.33	.05	.36	.58	.75	.88	.97	1.00	.89
(15,24)	-.51	-.38	-.12	.11	.32	.50	.67	.80	.89	1.00

Table 5.2 Continued

MONTHLY CORRELATION MATRIX APR

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.95	.83	.53	.32	.08	-.12	-.22	-.25	-.27
(14,16)	.95	1.00	.95	.71	.50	.25	.02	-.12	-.21	-.29
(15,17)	.83	.95	1.00	.88	.71	.47	.20	.02	-.12	-.25
(16,18)	.53	.71	.88	1.00	.94	.74	.47	.26	.08	-.14
(16,19)	.32	.50	.71	.94	1.00	.92	.73	.53	.33	.07
(16,20)	.08	.25	.47	.74	.92	1.00	.93	.80	.62	.36
(16,21)	-.12	.02	.20	.47	.73	.93	1.00	.95	.84	.62
(16,22)	-.22	-.12	.02	.26	.53	.80	.95	1.00	.96	.80
(16,23)	-.25	-.21	-.12	.08	.33	.62	.84	.96	1.00	.92
(15,24)	-.27	-.29	-.25	-.14	.07	.36	.62	.80	.92	1.00

MONTHLY CORRELATION MATRIX MAY

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.91	.64	.24	.11	-.03	-.13	-.15	-.09	-.14
(14,16)	.91	1.00	.87	.51	.35	.16	-.03	-.14	-.18	-.22
(15,17)	.64	.87	1.00	.85	.72	.52	.27	.04	-.11	-.20
(16,18)	.24	.51	.85	1.00	.96	.81	.58	.32	.08	-.09
(16,19)	.11	.35	.72	.96	1.00	.94	.76	.51	.25	.04
(16,20)	-.03	.16	.52	.81	.94	1.00	.93	.75	.50	.27
(16,21)	-.13	-.03	.27	.58	.76	.93	1.00	.93	.74	.52
(16,22)	-.15	-.14	.04	.32	.51	.75	.93	1.00	.93	.73
(16,23)	-.09	-.18	-.11	.08	.25	.50	.74	.93	1.00	.88
(15,24)	-.14	-.22	-.20	-.09	.04	.27	.52	.73	.88	1.00

Table 5.2 Continued

MONTHLY CORRELATION MATRIX JUN

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.83	.34	-.09	-.25	-.31	-.31	-.26	-.14	.03
(14,16)	.83	1.00	.78	.35	.11	-.03	-.13	-.18	-.16	-.09
(15,17)	.34	.78	1.00	.83	.65	.47	.31	.15	.02	-.09
(16,18)	-.09	.35	.83	1.00	.94	.83	.69	.52	.31	.05
(16,19)	-.25	.11	.65	.94	1.00	.96	.87	.72	.49	.16
(16,20)	-.31	-.03	.47	.83	.96	1.00	.97	.86	.66	.31
(16,21)	-.31	-.13	.31	.69	.87	.97	1.00	.96	.81	.47
(16,22)	-.26	-.18	.15	.52	.72	.86	.96	1.00	.94	.65
(16,23)	-.14	-.16	.02	.31	.49	.66	.81	.94	1.00	.83
(15,24)	.03	-.09	-.09	.05	.16	.31	.47	.65	.83	1.00

MONTHLY CORRELATION MATRIX JUL

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.94	.67	.31	.14	.04	-.04	-.08	-.01	-.01
(14,16)	.94	1.00	.85	.54	.34	.21	.07	-.03	-.03	-.09
(15,17)	.67	.85	1.00	.87	.69	.48	.29	.16	-.01	-.09
(16,18)	.31	.54	.87	1.00	.93	.75	.57	.39	.11	-.03
(16,19)	.14	.34	.69	.93	1.00	.93	.79	.61	.26	.07
(16,20)	.04	.21	.48	.75	.93	1.00	.95	.78	.44	.17
(16,21)	-.04	.07	.29	.57	.79	.95	1.00	.92	.64	.32
(16,22)	-.08	-.03	.16	.39	.61	.78	.92	1.00	.84	.56
(16,23)	-.01	-.03	-.01	.11	.26	.44	.64	.84	1.00	.79
(15,24)	-.01	-.09	-.09	-.03	.07	.17	.32	.56	.79	1.00

Table 5.2 Continued

MONTHLY CORRELATION MATRIX AUG

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.92	.67	.31	.05	-.17	-.30	-.36	-.39	-.37
(14,16)	.92	1.00	.89	.54	.22	-.07	-.24	-.32	-.36	-.47
(15,17)	.67	.89	1.00	.84	.55	.24	.02	-.12	-.20	-.44
(16,18)	.31	.54	.84	1.00	.87	.62	.39	.21	.06	-.26
(16,19)	.05	.22	.55	.87	1.00	.92	.75	.58	.41	.08
(16,20)	-.17	-.07	.24	.62	.92	1.00	.95	.83	.68	.39
(16,21)	-.30	-.24	.02	.39	.75	.95	1.00	.96	.85	.60
(16,22)	-.36	-.32	-.12	.21	.58	.83	.96	1.00	.96	.75
(16,23)	-.39	-.36	-.20	.06	.41	.68	.85	.96	1.00	.83
(15,24)	-.37	-.47	-.44	-.26	.08	.39	.60	.75	.83	1.00

MONTHLY CORRELATION MATRIX SEP

	(13,15)	(14,16)	(15,17)	(16,18)	(16,19)	(16,20)	(16,21)	(16,22)	(16,23)	(15,24)
(13,15)	1.00	.77	.48	.29	.18	.04	-.09	-.19	-.23	-.21
(14,16)	.77	1.00	.90	.70	.55	.33	.10	-.08	-.21	-.35
(15,17)	.48	.90	1.00	.93	.80	.57	.31	.10	-.08	-.31
(16,18)	.29	.70	.93	1.00	.93	.72	.46	.24	.06	-.21
(16,19)	.18	.55	.80	.93	1.00	.92	.75	.57	.39	.11
(16,20)	.04	.33	.57	.72	.92	1.00	.94	.83	.69	.43
(16,21)	-.09	.10	.31	.46	.75	.94	1.00	.97	.87	.66
(16,22)	-.19	-.08	.10	.24	.57	.83	.97	1.00	.97	.80
(16,23)	-.23	-.21	-.08	.06	.39	.69	.87	.97	1.00	.90
(15,24)	-.21	-.35	-.31	-.21	.11	.43	.66	.80	.90	1.00

5.2 Precipitation at a Point and Atmospheric Circulation

The analyses in this section are aimed at exploring relationships between precipitation at a point and simple characteristics of the pressure field. The data used for analysis were daily precipitation data from Gonzales Heights, Victoria, British Columbia and Snoqualmie Falls, Washington and pressure data at a small number of points on the NMC grid surrounding Victoria. The locations of precipitation and pressure stations are shown in Figure 5.2. Victoria and Snoqualmie Falls were chosen for this analyis because of the high quality and reliability of their records. In particular the Snoqualmie Falls data are known to be representative of regional conditions over parts of the Cascade range (Rasmussen and Tangborn 1976). The pressure stations were chosen to allow a reasonable description of the upper level circulation affecting the area. Note that the pressure stations are referred to by their coordinates (I,J) on the NMC grid of Figure 5.1.

The time scale of analysis is a matter of some concern. Since the primary interest is in low frequency effects, viz., long periods of the order of months with below normal rainfall, analysis of daily data is probably not appropriate. As mentioned in Chapter 3, blocking highs have a mean duration of about 15 days and frontal systems pass through the Pacific Northwest at approximately 5-day intervals. Thus analysis of mean 5-day or monthly data may be more appropriate.

The most obvious relationship to investigate at the outset is that between the precipitation depth at a station and the geopotential height of the 500 mb surface at that site. Higher than normal pressures are generally associated with dry weather so it is expected that a weak inverse relationship may be present.

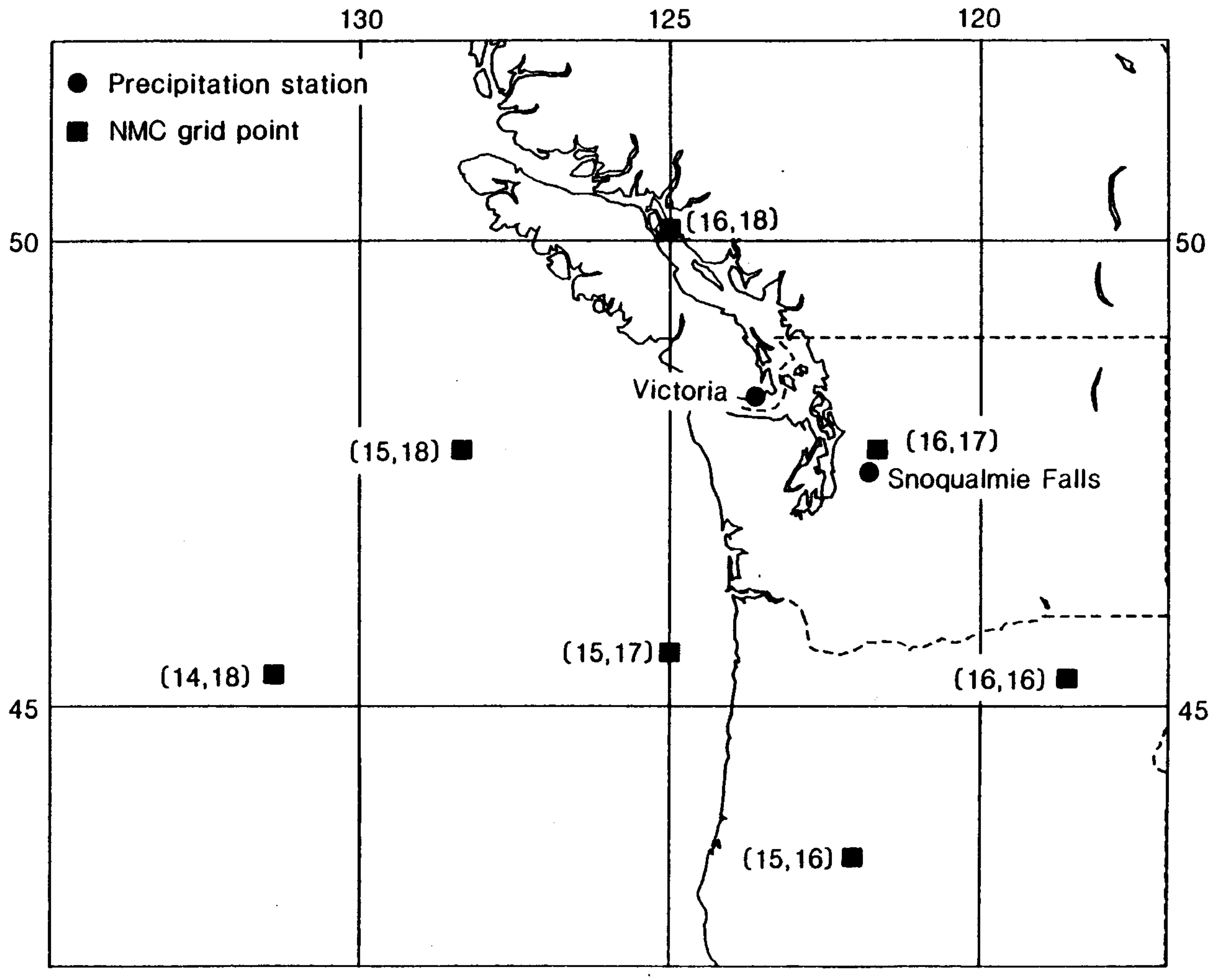


Figure 5.2 Location of data stations for analysis of concurrent point precipitation and atmospheric circulation

Figure 5.3 shows a scatterplot of 5-day January precipitation depths at Victoria against the mean 5-day 500 mb geopotential height at Victoria estimated by linear interpolation between data at grid points (16,17) and (15,18) as:

0.7*Z(16,17) + 0.3*Z(15,18)

where Z(I,J) = geopotential height on the 500 mb surface at NMC grid point (I,J)

The information in the scatterplot has been summarized by also plotting on Figure 5.3 estimates of the moving midmeans, moving lower semi-midmeans and moving upper semi-midmeans. Use of these moving statistics to summarize scatterplots was introduced by Cleveland and Kleiner (1975). The midmean is defined as the average of all observations between the quartiles and is a robust estimate of the mean. The lower semi-midmean is the midmean of all observations below the median and the upper semi-midmean is the midmean of all observations above the median. The semi-midmeans give an indication of the variability of the data.

The moving statistic curves were obtained using the procedures of Cleveland and Kleiner as follows: Given x_k, y_k, $k = 1,\ldots,n$, let x_k be in increasing order and define $x_{m(k)+1},\ldots, x_{m(k)+r}$ to be the r values of the abscissa which are closest to x_k in terms of absolute deviation.

Define $\tilde{x}_k = (x_{m(k)+1},\ldots,x_{m(k)+r})$

$\tilde{y}_k = (y_{m(k)+1},\ldots,y_{m(k)+r})$

Let $S_0(k) = \text{midmean } (\tilde{x}_k)$

$S_2(k) = \text{midmean } (\tilde{y}_k)$

$z_k = y_k - S_2(k)$

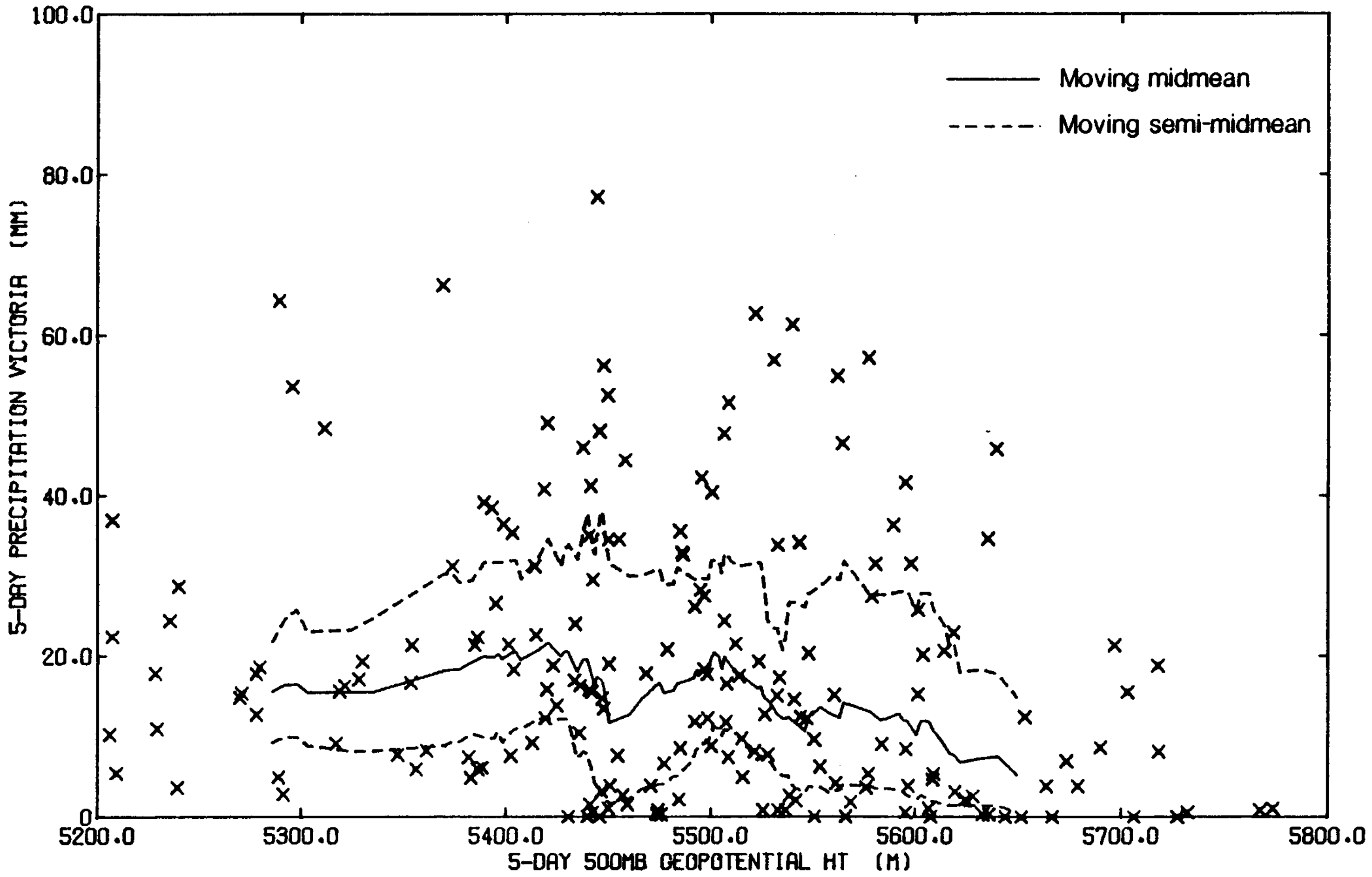


Figure 5.3 Five-day precipitation depths at Victoria vs. mean five-day 500 mb geopotential height at Victoria for January data

Define $\tilde{z}_k = (z_{m(k)+1},\ldots,z_{m(k)+r})$

Let $S_1(k)$ = lower semi-midmean $(\tilde{z}_k) + S_2(k)$
$S_3(k)$ = upper semi-midmean $(\tilde{z}_k) + S_2(k)$

Then the curves on Figure 5.3 are plots of
$S_1(k)$ vs $S_0(k)$ - moving lower semi-midmean
$S_2(k)$ vs $S_0(k)$ - moving midmean
$S_2(k)$ vs $S_0(k)$ - moving upper semi-midmean

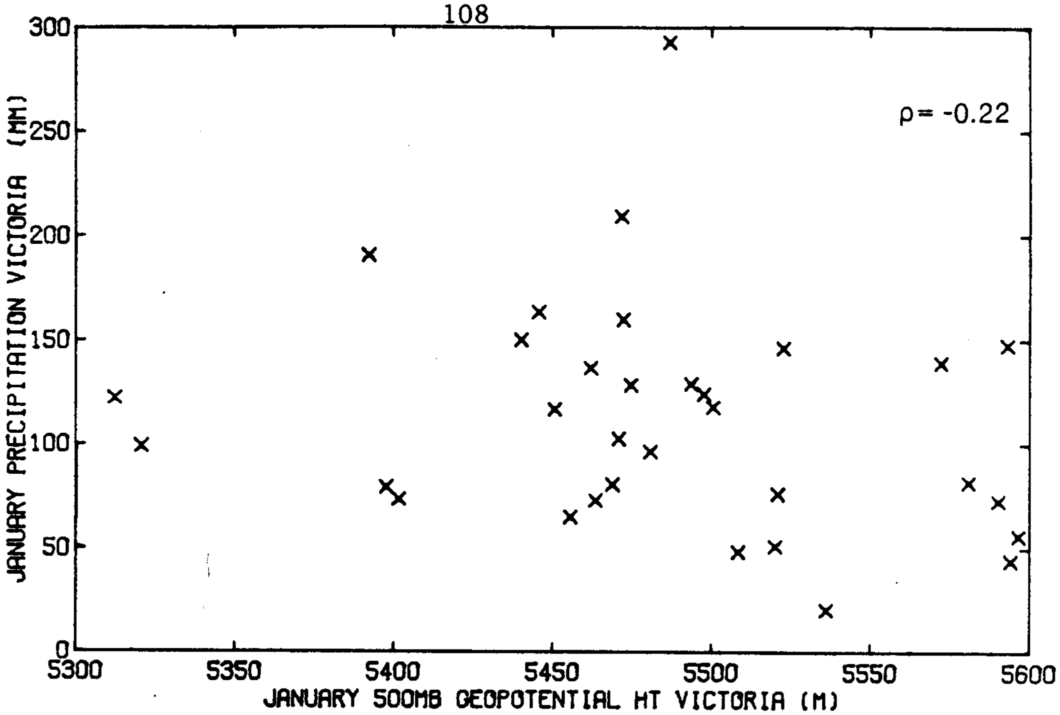
The smoothness of the curves depends on the value of r. In Figure 5.3 a value of $r = 29$ was used. A higher value would have resulted in smoother curves but would not have changed their general form. The moving statistics on Figure 5.3 show that for 500 mb geopotential heights less than 5500 m there is no relationship between pressure and precipitation. Above 5500 m, however, there does appear to be a slight inverse relationship with lower rainfall associated with higher pressures.

Figures 5.4a and 5.4b show scatterplots of monthly precipitation depths at Victoria against the mean monthly 500 mb geopotential height at Victoria for January and July respectively. As with the 5-day plots the geopotential height at Victoria is estimated as:

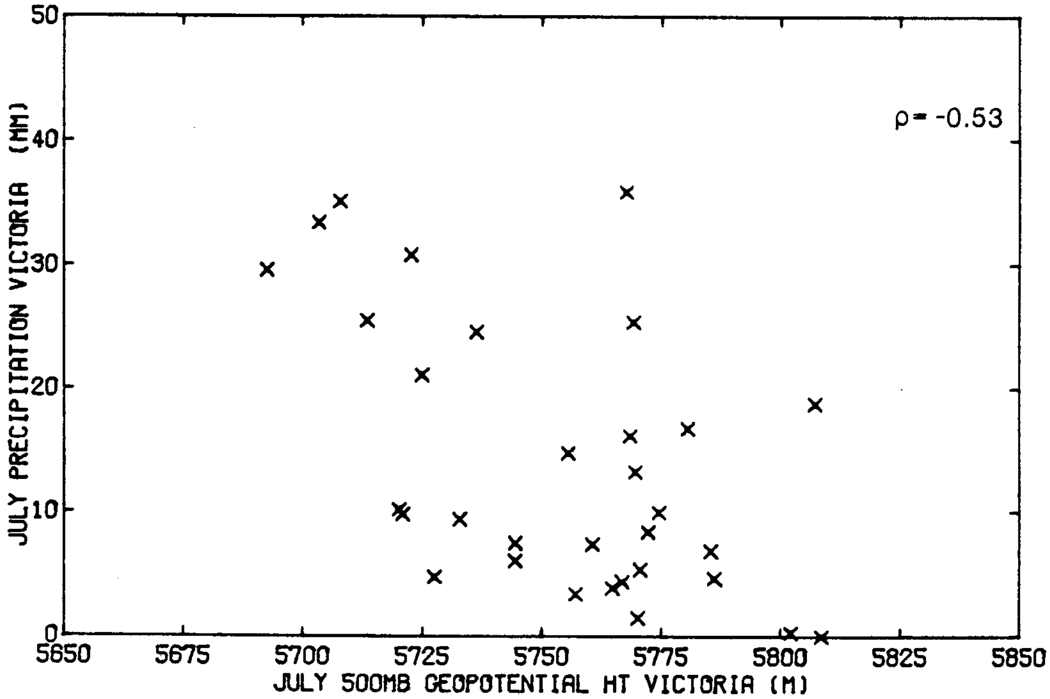
0.7*Z(16,17) + 0.3*Z(15,18)

where Z(I,J) = 500 mb geopotential height at (I,J)

The scatterplots show that in January (Figure 5.4a) there is essentially no relationship between precipitation and geopotential height, while in July (Figure 5.4b) there is a weak inverse relationship.



(a) January



(b) July

Figure 5.4 Monthly precipitation at Victoria vs. mean monthly 500 mb geopotential height at Victoria

Cross correlations of monthly data for Victoria precipitation against geopotential height (0.7*Z(16,17) + 0.3*Z(15,18)) and Snoqualmie Falls precipitation against geopotential height Z(16,17) are shown in columns 1 and 3 of Table 5.3 (see Figure 5.2 for station locations). The data indicate somewhat stronger inverse relationships for the Snoqualmie Falls data than for the Victoria data. This effect may be a result of the position of the Victoria site in the rainshadow of the Olympic Mountains. A second factor may be the linear interpolation used to estimate the 500 mb height at Victoria. The 500 mb surface is known, however, to be quite smooth and the method of interpolation is not believed to be particularly important. Another feature apparent from Table 5.3 is that correlations between precipitation and pressure are consistently higher in the summer and fall months (May through November) than in the months December through April. One possible explanation for this feature will become apparent later in this section.

The relationship between precipitation and pressure is clearly quite weak in this instance particularly in the important winter months. This is probably because the pressure at a point or the average pressure over a number of grid points does not indicate the nature of the atmospheric circulation. A better approach may be to assume that winds at the 500 mb level are geostrophic and are representative of the circulation in the troposphere. The approximate horizontal equations of motion for an air parcel can be written using Newton's Second Law (e.g. Wallace and Hobbs 1977):

$$\begin{aligned} \frac{du}{dt} - fv &= -\frac{1}{\rho}\frac{\partial p}{\partial x} + F_x \\ \frac{dv}{dt} + fu &= -\frac{1}{\rho}\frac{\partial p}{\partial y} + F_y \end{aligned} \qquad (5.1)$$

Table 5.3 Monthly Cross Correlations Between Precipitation and Characteristics of the 500 mb Geopotential Height Field

Month	Victoria Precipitation vs 0.7*Z(16,17)+0.3*Z(15,18)	Victoria Precipitation vs Z(15,17)-Z(16,18)	Snoqualmie Falls Precipitation vs Z(16,17)	Snoqualmie Falls Precipitation vs Z(15,16)-Z(16,17)	Snoqualmie Falls Precipitation vs Z(15,17)-Z(16,18)
Oct	-0.65	0.61	-0.72	0.56	0.46
Nov	-0.32	0.65	-0.57	0.66	0.58
Dec	0.02	0.43	-0.37	0.42	0.32
Jan	-0.22	0.68	-0.27	0.67	0.62
Feb	-0.19	0.57	-0.24	0.54	0.53
Mar	-0.08	0.56	-0.26	0.66	0.60
Apr	-0.11	0.03	-0.53	0.31	0.12
May	-0.26	-0.20	-0.72	-0.04	-0.15
Jun	-0.35	0.09	-0.53	0.17	-0.01
Jul	-0.53	-0.06	-0.68	0.18	-0.13
Aug	-0.47	-0.06	-0.56	0.13	0.13
Sep	-0.55	0.51	-0.72	0.66	0.50

where x = eastward distance,measured along a latitude circle

y = northward distance measured along a longitude circle

u = horizontal velocity in x-direction

v = horizontal velocity in y-direction

f = Coriolis parameter = $2\Omega \sin \phi$

where Ω = angular velocity of earths rotation

ϕ = latitude

ρ = density

p = pressure

F = frictional forces

For synoptic scale systems the acceleration terms are generally small compared with the other terms in the above expressions. Also at the 500 mb level, frictional terms are sufficiently small that they may be neglected leaving the geostrophic wind equation in isobaric coordinates at 500 mb is:

$$v_g = \frac{1}{f}\frac{\partial \Phi}{\partial x} = \frac{g}{f}\frac{\partial Z}{\partial x}$$

$$u_g = -\frac{1}{f}\frac{\partial \Phi}{\partial y} = -\frac{g}{f}\frac{\partial Z}{\partial y} \qquad (5.2)$$

where v_g, u_g = the northward and eastward components of geostrophic wind along the 500 mb pressure surface

Φ = the geopotential on the constant pressure surface

Z = geopotential height on the 500 mb pressure surface

Thus to a first approximation the gradient of geopotential height at 500 mb gives the direction and magnitude of winds at the 500 mb level.

The geostrophic approximations suggest a number of relationships worth investigation. An east-west zonal flow or a weak west-east zonal flow over the Pacific Northwest is generally associated with dry conditions and a strong zonal (i.e. west-east) flow with wet conditions. Thus a relationship might be expected between the north-south geopotential gradient $\partial\Phi/\partial y$ and precipitation. Large negative values of $\partial\Phi/\partial y$ indicating strong west-east flow should be associated with heavy precipitation and small negative or positive values of $\partial\Phi/\partial y$ with dry conditions.

Figure 5.5 shows a scatterplot of 5-day January precipitation depths at Victoria against the mean 5-day 500 mb geopotential height gradient at Victoria represented by the difference between geopotential heights Z(15,17) - Z(16,18).

As before, the information in the scatterplot is summarized in Figure 5.5 by plotting estimates of the moving midmeans and upper and lower semi-midmeans. The scatterplot shows an increase of precipitation depth with increasing geopotential height difference, but the scatter about the moving midmean curve is very large.

Figures 5.6a and 5.6b show scatterplots of monthly precipitation depths at Victoria against the mean monthly geopotential height difference (Z(15,17) - Z(16,18)) for January and July respectively. The scatterplots show that in January (Figure 5.6a) there is a weak relationship between precipitation and geopotential height difference while in July (Figure 5.6b) there is no apparent relationship.

Cross correlations of the monthly data for Victoria precipitation against the geopotential height difference (Z(15,17) - Z(16,18)) and Snoqualmie Falls precipitation against geopotential height difference (Z(15,16) - Z(16,17)) are shown in columns 2 and 4 of Table 5.3. The monthly cross correlations for Snoqualmie Falls precipitation against the geopotential height difference (Z(15,17) - Z(16,18)) are also shown in column 5 of Table 5.3.

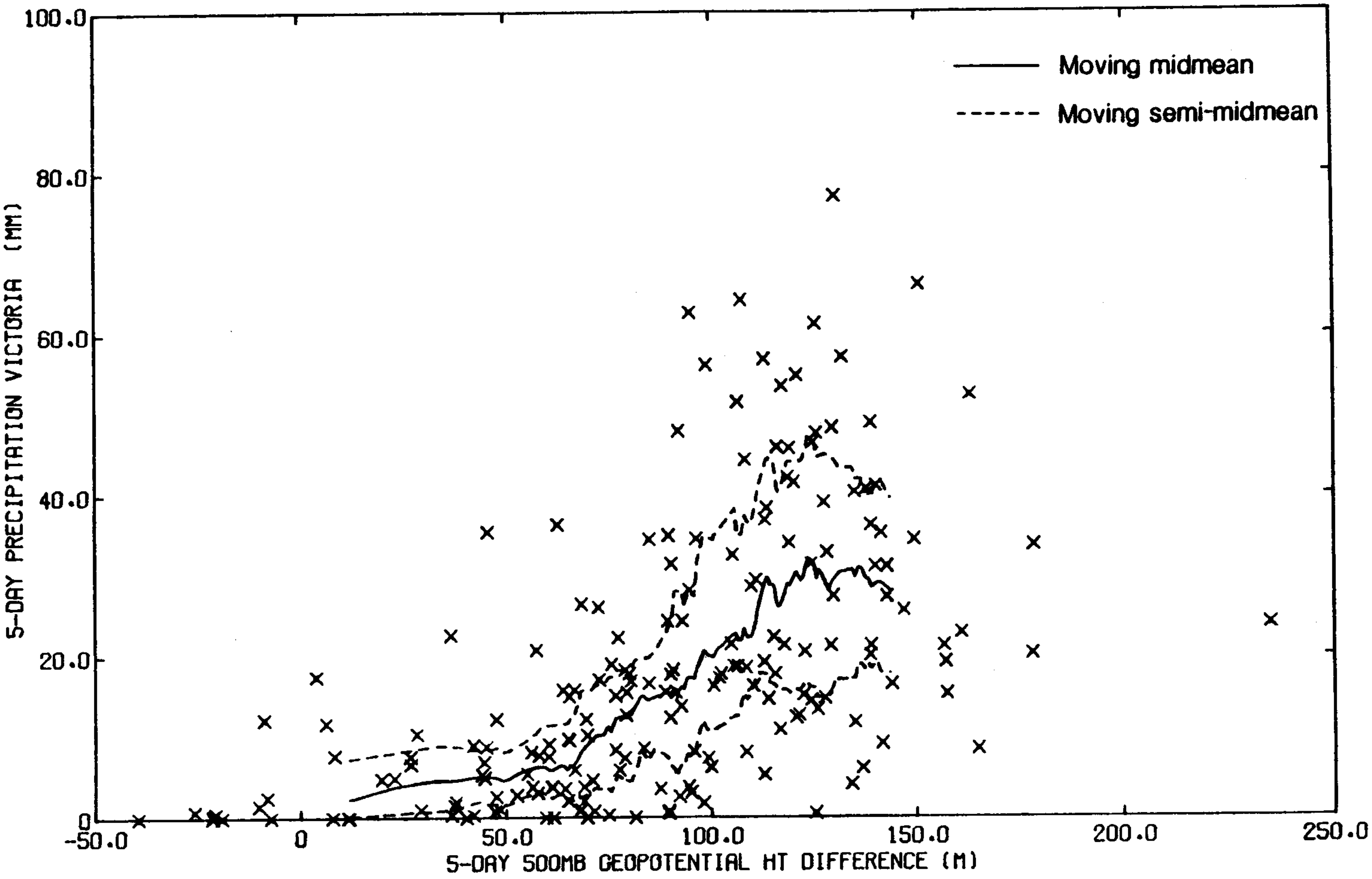
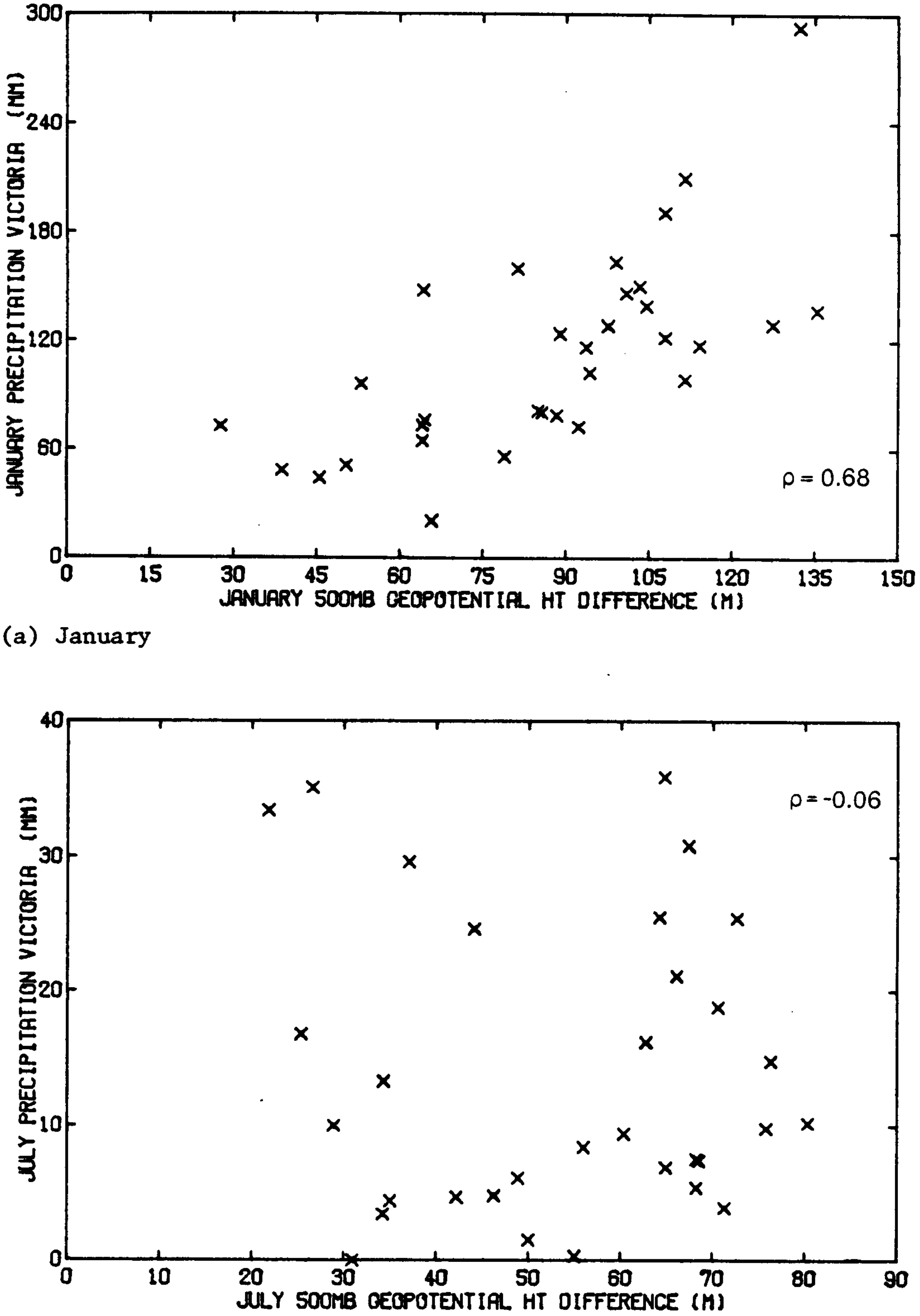


Figure 5.5 Five-day precipitation depths at Victoria vs. mean five-day 500 mb geopotential height difference Z(15,17) - Z(16,18) for January data



(a) January

(b) July

Figure 5.6 Monthly precipitation at Victoria vs. mean monthly 500 mb geopotential height difference Z(15,17)-Z(16,18)

These data show a correlation coefficient of about 0.6 for the months September through March. In April there is a very sharp drop in the correlation coefficient and there is apparently no relationship between precipitation and geopotential height difference from April through August.

It appears that in the winter months the strength of zonal circulation is more important in determining precipitation depths than the actual geopotential height. In the summer months, however, the geopotential gradient is a poor indicator of precipitation depth. This is probably a result of the general weakening of zonal circulation in the northern hemisphere summer. Summer rainfall depths appear to be tied more closely to the presence or absence of high pressure cells.

The information presented so far fails to demonstrate any strong relationship between precipitation and features of the pressure field. Since my principal interest is in use of the pressure data to differentiate between "wet" and "dry" periods, a more reasonable approach to analysis might be to compare characteristics of the pressure field during wet and dry periods in the historic rainfall data.

A "dry" period is defined here as any period of 5 days or more during which no more than 2 mm of rainfall fell on any one day. Any period of days not meeting these requirements is defined as a "wet" period. These definitions are by necessity somewhat arbitrary. The definition of the "dry" period is intended to identify periods of time during which the meteorologic mechanisms are such as to promote dry weather. Clearly a period of one or two dry days between the passage of frontal systems could easily occur during strong zonal flow when in principal at least there is no reason to expect dry conditions.

Since my principal interest is in winter precipitation, the daily precipitation data from Victoria for each year from November 15 to March 15 were split into dry and wet periods according to the above definition. Various characteristics of the geopotential height field concurrent with the wet and dry periods were then investigated.

For example, the mean geopotential height difference (Z(15,17) - Z(16,18)) for each wet or dry period in the winter record was determined. Histograms of the mean geopotential height differences during wet and dry periods are shown in Figure 5.7. The histograms show that in general dry periods are associated with lower geopotential height differences. There is, however,considerable overlap in the histograms and the inverse problem of identifying dry periods given data on the geopotential height difference could clearly not be solved with any degree of confidence.

Figure 5.7 gives no indication of the length of dry periods, only their occurrence. However, there is no apparent relationship between length of dry period and mean geopotential gradient. For example, dry periods of 5 days were associated with mean geopotential differences ranging between -3.5 m and 176.1 m and the two longest dry periods of 30 and 28 days had mean geopotential differences of 35.2 and 73.9 m respectively.

A more direct differentiation of zonal and meridional circulation is provided by the index:

$$I = \frac{|Z(14,18) - Z(15,17)| + |Z(15,17) - Z(16,16)|}{2\,|Z(15,17) - Z(16,18)|} \qquad (5.3)$$

The location of these geopotential sites are shown in Figure 5.2. It can be seen that a strong zonal flow would have large values of (Z(15,17) - Z(16,18)) (west-east flow component) and small values of $|Z(14,18) - Z(15,17)|$ and $|Z(15,17) - Z(16,16)|$ (north-south flow components).

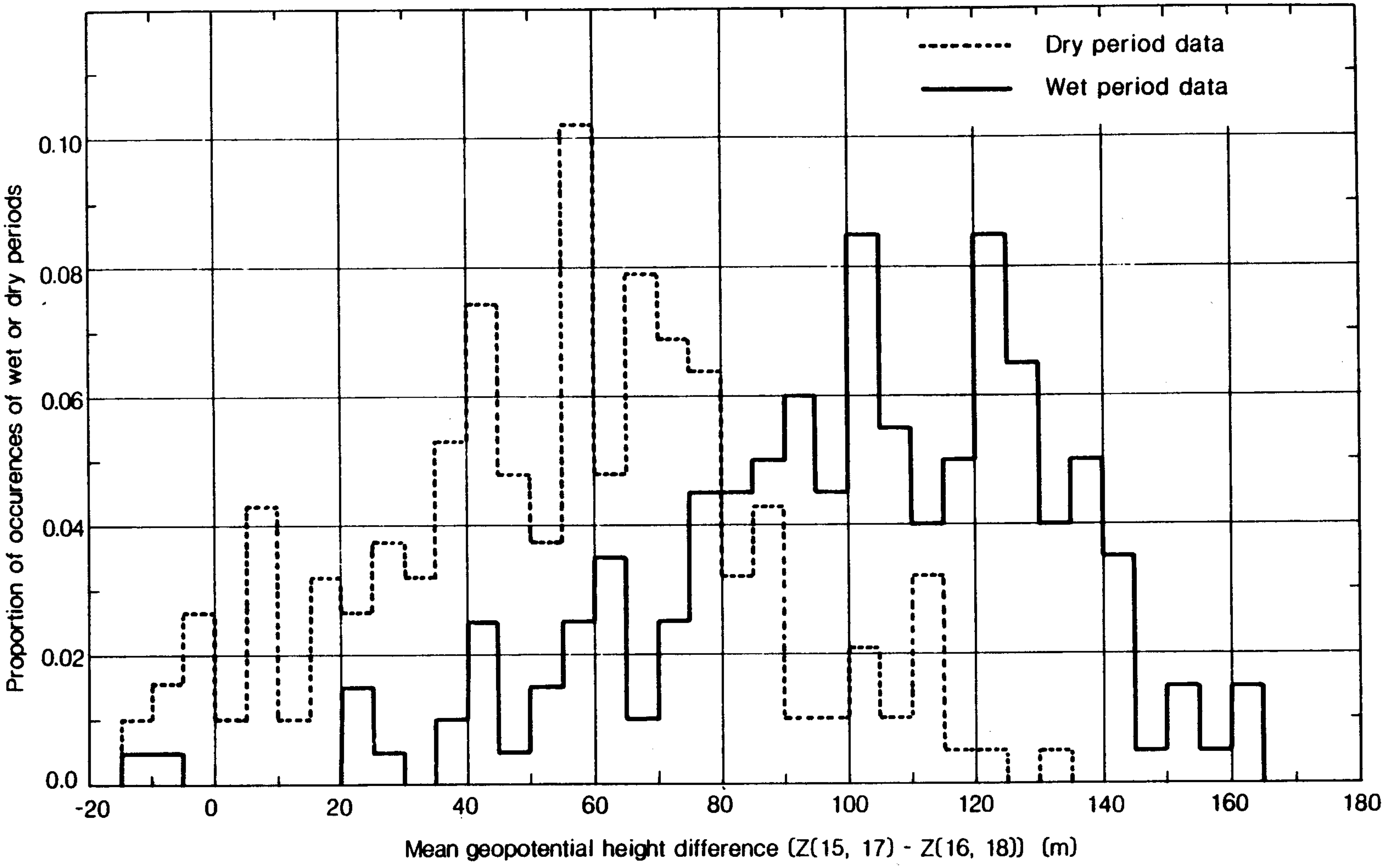


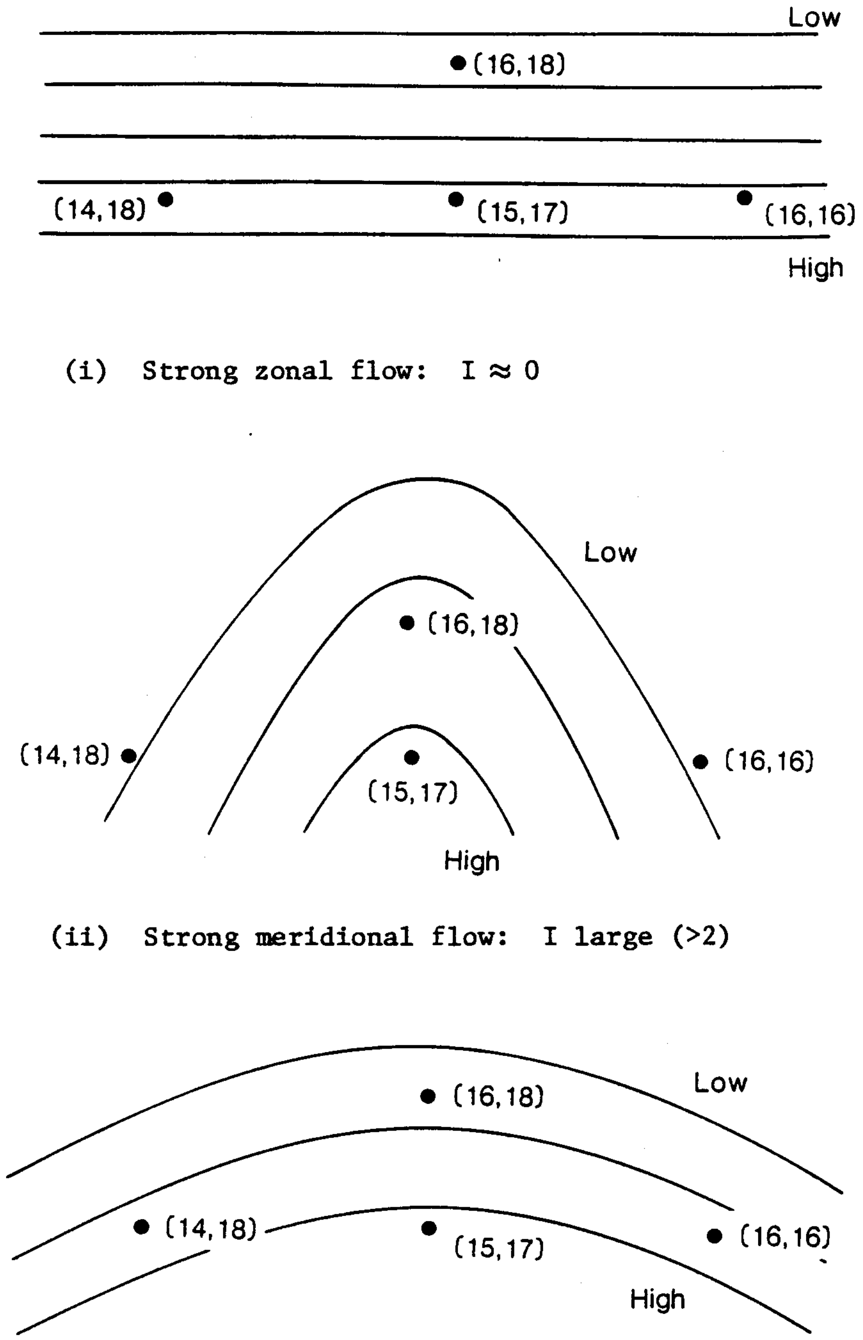
Figure 5.7 Histograms of mean geopotential height difference for wet and dry periods

The index values for flows with strong meridional components depend to a large extent on the exact form and position of the wave. However, in general, the numerator in the index would be relatively large and the denominator small in comparison to conditions prevailing during zonal flow. Strong meridional flows would thus tend to have large index values. A number of possible configurations for the geopotential height contours over the network of sites are sketched in Figure 5.8.

Using the earlier definitions of wet and dry periods, values of the index I were determined for each wet or dry period in the winter record. Histograms of the index value during wet and dry periods are shown in Figure 5.9. As in Figure 5.7 the histograms for the wet and dry periods show considerable overlap. Extremely high index values, however, are always associated with dry periods.

Again Figure 5.9 gives no indication of the length of dry periods. However, there is no apparent relationship between dry period length and the index value.

Examination of concurrent monthly precipitation data at Victoria and mean monthly index values was disappointing. The monthly scatterplots showed little of the expected structure. The strongest relationship was found for the January data whose scatterplot is shown in Figure 5.10. The January data show that higher index values tend to be associated with lower rainfall but little else can be inferred from the data. Scatterplots for other winter months show very little structure at all.



(i) Strong zonal flow: I ≈ 0

(ii) Strong meridional flow: I large (>2)

(iii) Intermediate condition: 0<I<2

Figure 5.8 Hypothetical configurations of geopotential height contours over the Pacific Northwest

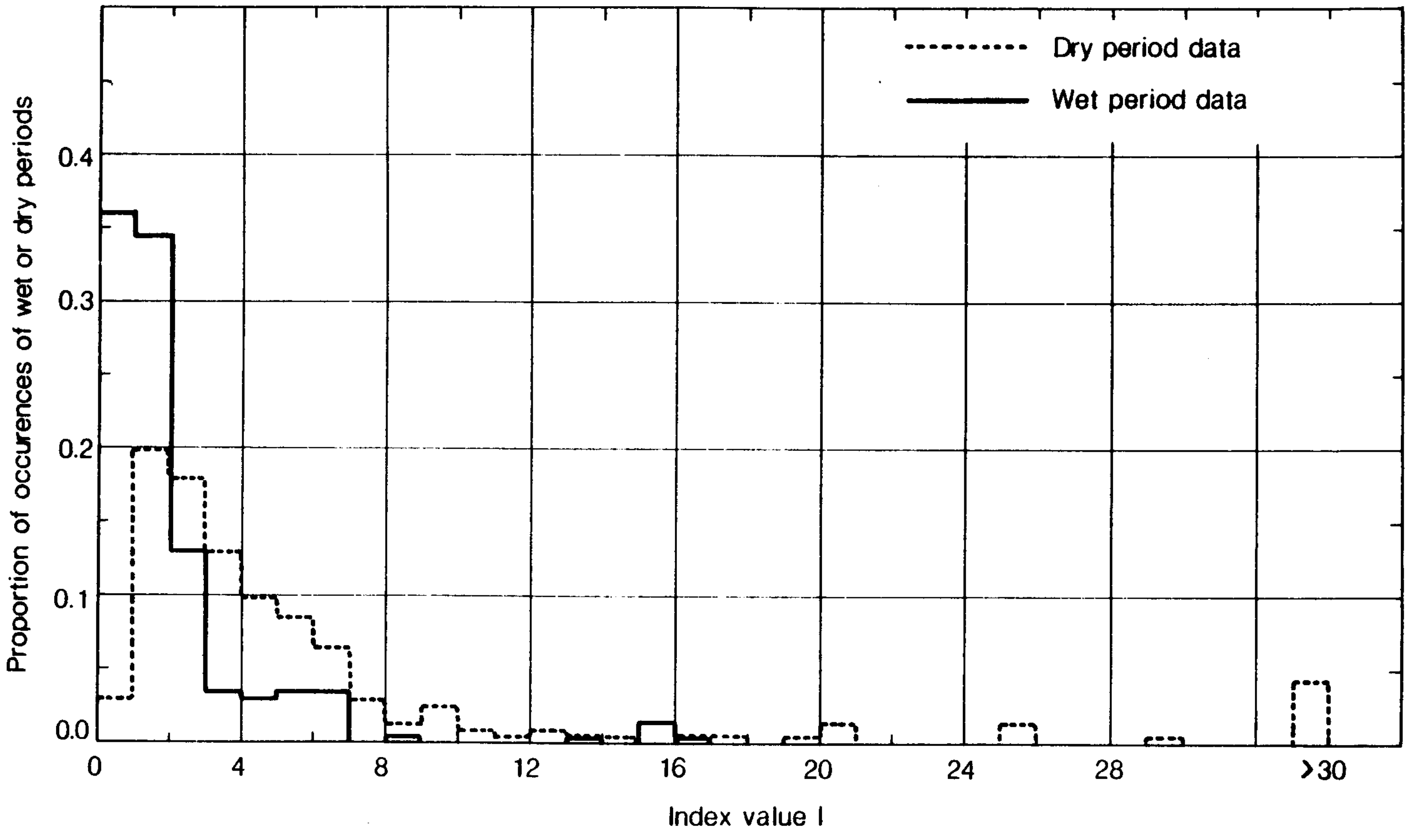


Figure 5.9 Histograms of index I (equation 5.3) for wet and dry periods

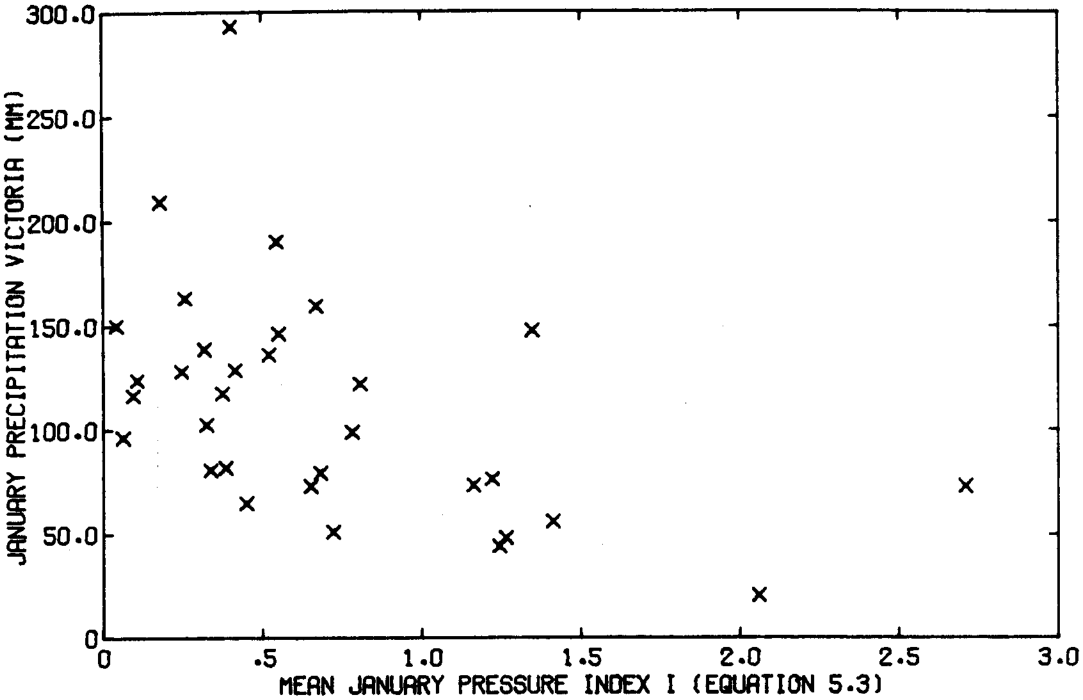


Figure 5.10 Monthly precipitation at Victoria vs. mean monthly index I (Equation 5.3) for January data

A number of other indices were investigated for differentiating between zonal and meridional flow. However, none performed better than the index described above. The main difficulty in using this or any other index is that in nature there is a continuum of flow conditions. Strong zonal and strong meridional flows can be easily distinguished but in the many intermediate cases it is clear that a simple zonal/meridional classification scheme is of little value. Moreover, circulation patterns are often too complex to be summarized by a simple index based on the geopotential at only three or four grid points. Both the daily analysis of Figure 5.9 and the monthly analysis confirm what is already well known, i.e., that strong zonal flow is generally associated with wet conditions and strong meridional flow with dry conditions. However, this analysis, in common with the other work presented in this section, fails to indicate a suitable approach for making reliable quantitative statements about pressure/precipitation relationships.

5.3 Regional Precipitation and Atmospheric Circulation

The analysis in this section is aimed at exploring relationships between regional precipitation patterns along the west coast from central California to the Gulf of Alaska and atmospheric circulation. The data used for analysis were monthly precipitation data from 12 sites along the west coast and mean monthly 500 mb geopotential height data from 10 sites just off the west coast. The pressure and precipitation stations used are shown in Figure 5.11. Again the pressure stations are referred to by their coordinates on the NMC grid (Figure 5.1). The precipitation stations shown on Figure 5.11 were all previously used in the analyses of Chapter 4 and basic statistics from data at these sites are shown in Tables 4.1 through 4.5.

The relationships between pressure and precipitation during severe drought are illustrated in the monthly pressure and precipitation profiles of Figures 5.12 through 5.14. The pressure profiles were constructed by plotting the pressure at a site against the distance from site (13,15) measured along the series of chords connecting the pressure grid points. The precipitation profiles were obtained by plotting the standardized monthly precipitation against the distance between site (13,15) and the perpendicular dropped from the precipitation station to the nearest chord joining two pressure stations.

Widespread drought affected the Pacific Northwest during the winters of 1949, 1963, and 1977. The pressure-precipitation profile for January 1977 (Figure 5.13) shows below normal precipitation from Davis to Annette Island (approximately 2300 km) and above normal precipitation from Yakutat to Homer. The region of above normal precipitation can be seen to be associated with pressure gradients which are much steeper than those over the area affected by drought. The inference, as discussed in Chapter 3, is that the jet stream is taking a persistent northerly track into the Gulf of Alaska steering

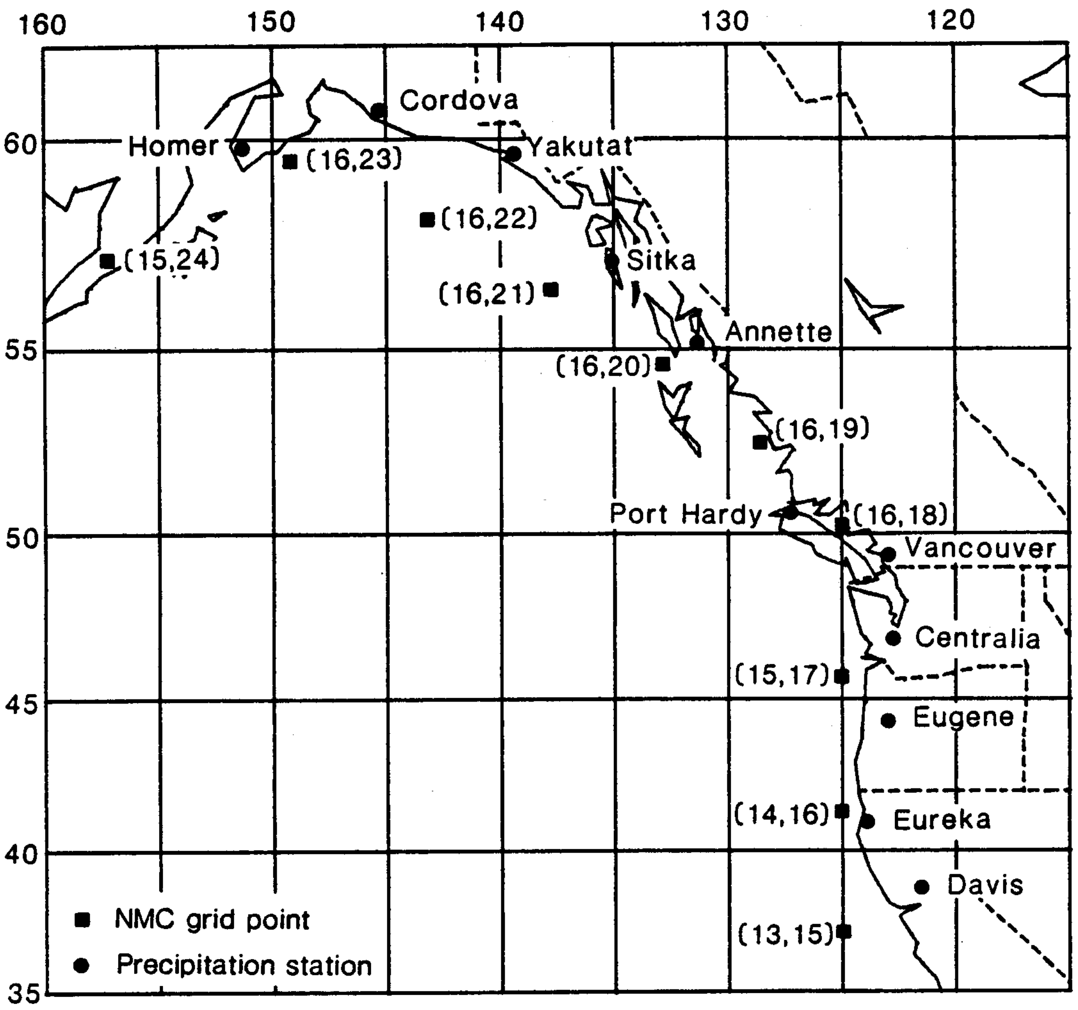


Figure 5.11 Location of data stations for analysis of concurrent regional precipitation and atmospheric circulation

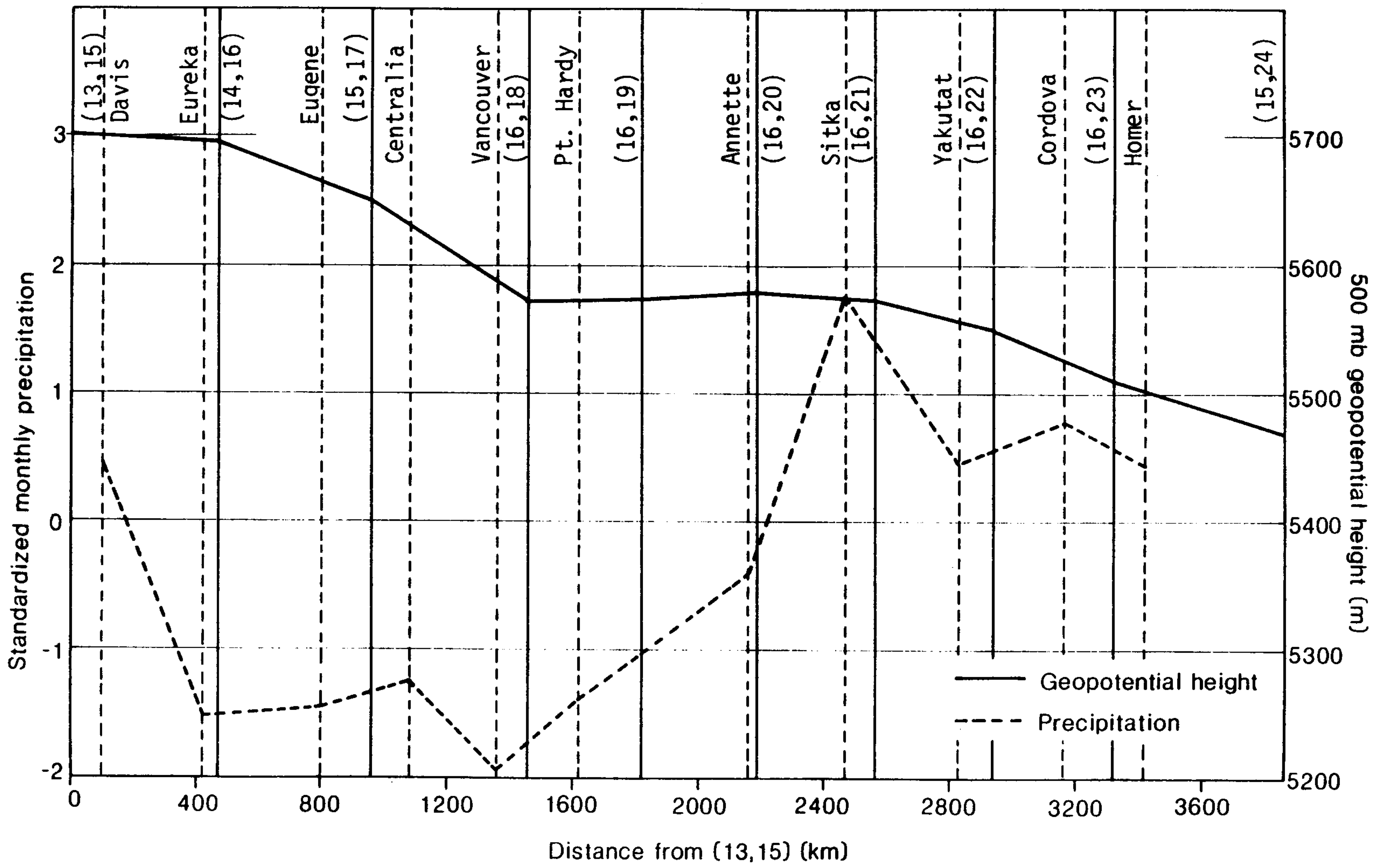


Figure 5.12 Precipitation and pressure profiles for January 1963

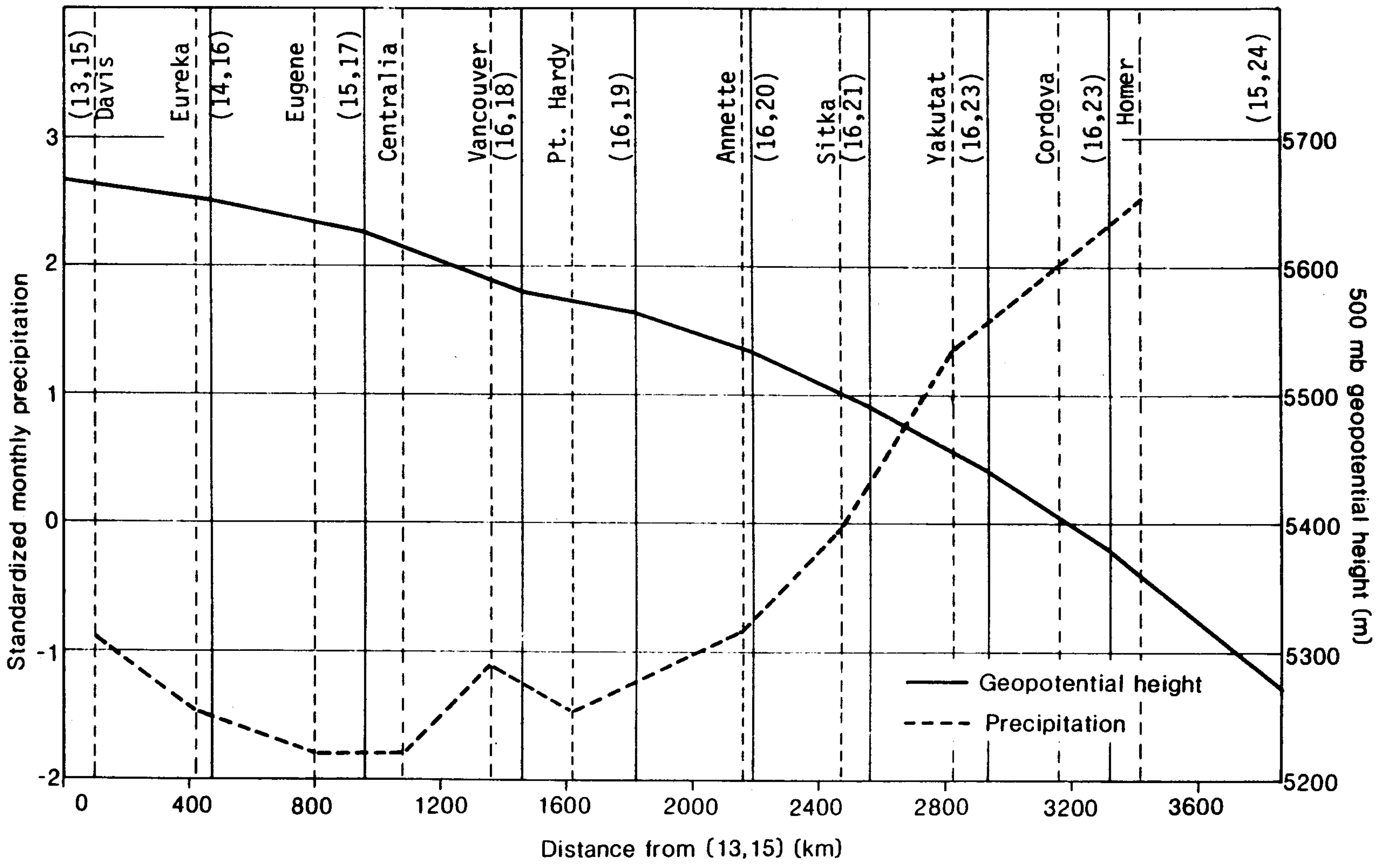


Figure 5.13 Precipitation and pressure profiles for January 1977

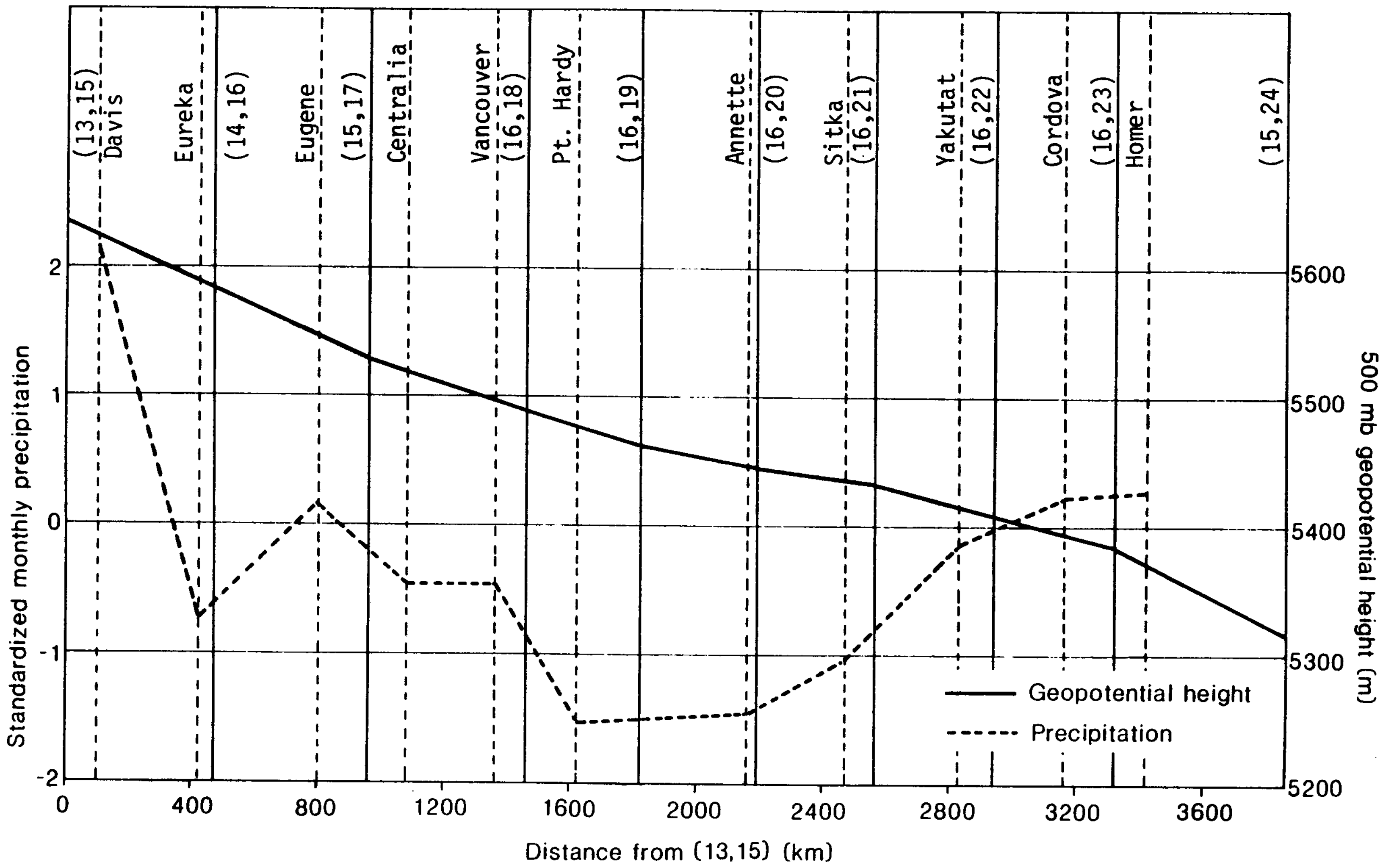


Figure 5.14 Precipitation and pressure profiles for January 1978

frontal systems over the Yakutat-Homer region and leaving the west coast south of Sitka unusually dry.

Unfortunately profiles for other periods are not as revealing as the January 1977 profiles. Figure 5.12 shows pressure and precipitation profiles for January 1963, a month with strong blocking in the eastern Pacific. The pressure profile in this instance is not an accurate reflection of the overall pressure pattern. O'Connor (1963) shows strong blocking between latitudes 35 and 55 degrees north in the eastern Pacific with the jet stream split into two branches; one branch flowing north over the eastern Aleutians (west of Homer) and the other branch flowing west to east over northern Mexico at about 25 degrees north. The profile in Figure 5.12 thus misses some of the important features of the actual pressure pattern even though the figure covers points up to 3600 km apart.

Figure 5.14 shows profiles for January 1978 which was similar in a number of respects to January 1963. As in January 1963 there was strong blocking for at least part of the month over the eastern Pacific with two distinct axes of wind speed maxima. One branch as in January 1963 flowed north into the Gulf of Alaska passing just west of Homer. The other axis followed a west-east track over southern Oregon. These conditions are reflected in Figure 5.14. The pressure profiles show a slight S-shape with flat gradients from Centralia to Sitka and relatively steeper gradients south of Centralia and north of Sitka. The precipitation profile shows dry conditions between Centralia and Sitka, near normal conditions north of Sitka and wet conditions south of Eureka.

The pressure/precipitation profiles such as Figures 5.12 to 5.14 could usually be interpreted with the assistance of the northern hemisphere geopotential height contour maps published in Monthly Weather Review. However, by themselves the profiles provide little useful qualitative or quantitative information.

The problems of interpreting the profiles are exacerbated by use of the monthly mean data. Figure 5.15 shows the pressure and precipitation profiles for January 1964. One possible interpretation of this figure is that strong zonal flow affected the entire coast from Davis to Yakutat with unusually heavy precipitation from Eugene to Annette Island. In fact the monthly weather report by Wagner (1978) shows that at the start of the month a high pressure ridge along the west coast forced storms into Alaska. Examination of the daily precipitation records shows that Sitka and Annette experienced heavy precipitation during the first 15 days of the month while Eugene and Eureka were fairly dry. Toward the middle of the month the ridge weakened and the storm track moved south. The latter part of the month saw heavy rainfall at Eugene and Eureka with relatively dry conditions in southern Alaska. Thus the mean monthly circulation in this, and a number of other instances, is a poor indication of the conditions actually prevailing during the month, and obviously limits the use of mean monthly data.

5.4 Concluding Remarks

The analyses of concurrent precipitation and geopotential height data presented in this chapter have demonstrated some weak relationships between precipitation depths and simple characteristics of the pressure field. These relationships tend to follow the known qualitative relationships between precipitation fields and the pressure patterns but are too weak to be of use in any quantitative setting. It is evident that the distinction between zonal and meridional flow is quite subjective and no simple characteristic of the pressure field has been found which would allow accurate classification of all data into zonal and meridional types. Only under extreme conditions is it possible to make a distinction between zonal and meridional circulation.

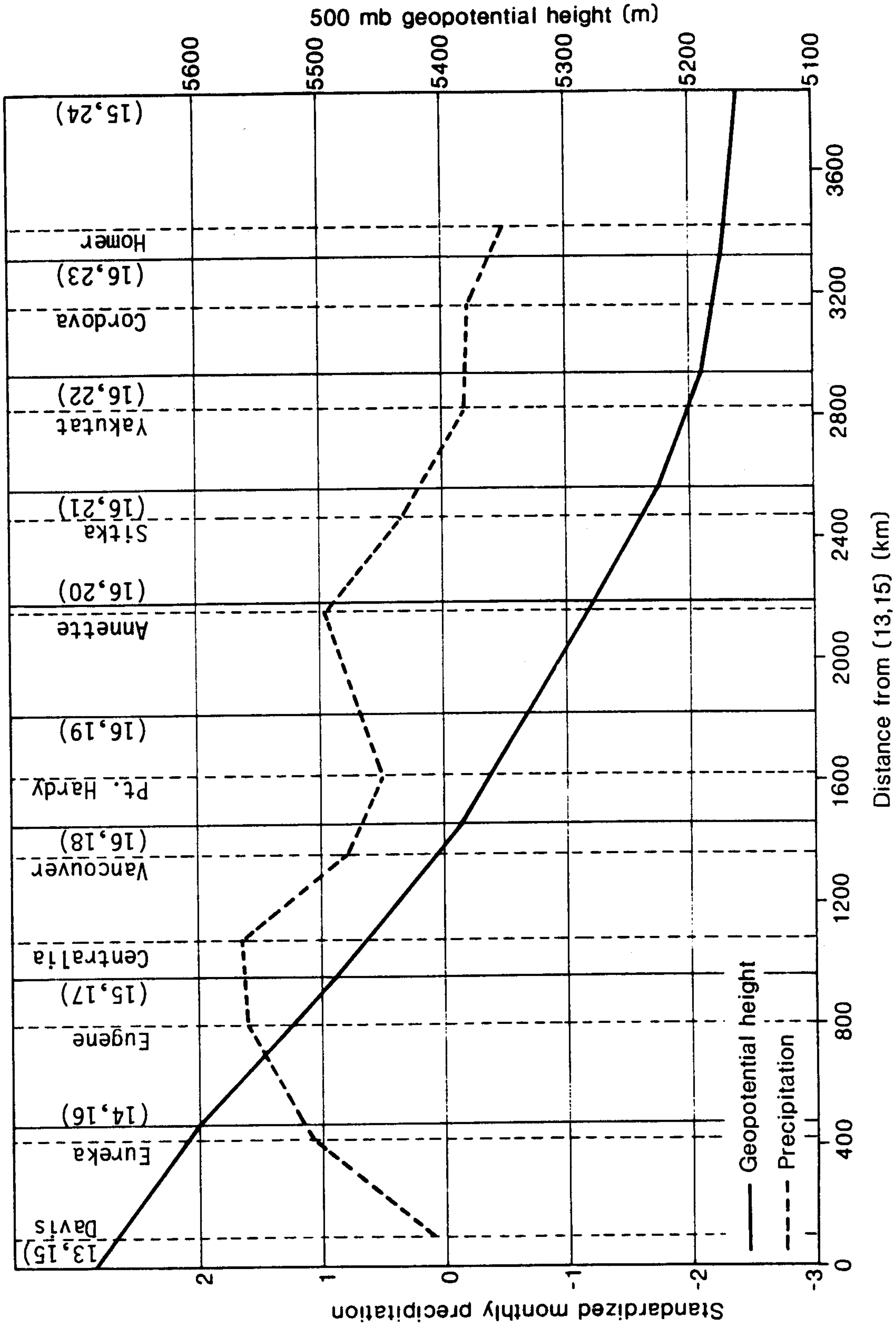


Figure 5.15 Precipitation and pressure profiles for January 1964

Attempts to relate precipitation depth to various indices of meridional flow were unsuccessful. However, the profiles of Figures 5.12 to 5.14 illustrate quite well the great north-south extent of drought associated with meridional flow over the Pacific Northwest. For example, the drought of 1977 extended at least 2400 km from southern California to southern Alaska. In contrast, many of the wettest months on record are associated with conditions affecting a very limited area. For example, the wettest January at Vancouver (January 1958) saw near normal conditions at the neighboring stations of Victoria, Centralia and Port Hardy. As has been shown in Figure 5.15, however, wet conditions can extend, on a monthly basis, over large areas but such conditions seem to be caused by changes in the latitude of the storm track during the month and are not a reflection of concurrent wet conditions on a daily basis.

It seems that at present there is no justification for using 500 mb geopotential data in a quantitative sense in stochastic hydrology. It may be possible to use pressure data to fill in missing rainfall records in regions with a very sparse network of rain gages (as has been done by Kilmartin 1980) but this consideration does not apply to the United States or Canada. Consideration of the qualitative aspects of atmospheric circulation does, however, seem to provide some potential for improving the tools currently in use. Certainly, as has been shown in Chapter 4, such considerations lead to the conclusion that inter-station precipitation relationships are nonlinear. Futhermore, an understanding of the qualitative relationships between precipitation patterns and atmospheric circulation allows a subjective assessment to be made of the plausibility of multi-site synthetic sequences, and as will be seen in the subsequent chapters, still provides a framework for what may be more realistic approaches to precipitation modeling.

Attempts to find a simple objective technique for separating precipitation data into dry and wet populations based on the state of atmospheric circulation were unsuccessful. This however, does not necessarily imply that precipitation data does not come from a mixed distribution. The lack of exogenous information for classifying the data into component populations greatly increases the difficulties of both identifying mixtures and estimating their parameters. Techniques for applying mixture distributions to unclassified data have been developed, however, and are discussed in the next chapter.

6.0 UNIVARIATE MIXTURE MODELS

Mixture models have been used in a number of fields in the past including genetics, telephone engineering, fisheries and water resources (see for example Fowlkes 1977). The basic concept underlying the development of mixture models is that a set of n observations $y = (y_1, y_2, \ldots y_n)$ are sampled from a finite set of m distributions, where the distribution from which each y_i is sampled may or may not have been observed and where the value m may or may not be known.

Most practical applications of mixture models have involved samples from a mixture of two normal, lognormal or possibly exponential distributions, where samples from either distribution can be associated with one of two physical states or phenomena. For example, Hosmer (1973a) used a mixture of two normal distributions to represent the lengths of 11 year old halibut. For any given age, it is known that the female halibut is on average larger than the male, thus lengths of 11-year old female and male halibut can be considered to be samples from two different distributions differentiated by sex. When a sample of halibut is obtained which cannot be classified by sex, then the lengths can be represented by a mixture model. The main points to note here are that there is a physical basis for using a mixture model and that objective exogenous information exists (i.e. sex) which, if observed, can be used to classify the sample accurately.

Another example of mixtures, discussed by Fowlkes (1977), concerns the distribution of the lengths of WATS telephone calls. In this case it was found that the data could be well represented as a mixture of two 2-parameter log-normal distributions. However, there was no objective exogenous information which could be used to classify the data. Since there is no objective physical basis for using mixtures in this case, it seems that application of a mixture model may be simply viewed as a curve fitting exercise, which could equally well have employed some other flexible multi-parameter distribution such as the Wakeby (Houghton 1978).

Mixed distributions have had limited application in the water resources field. Some examples of applications in hydrology are given by Hawkins (1974) and by Singh (1974). Hawkins demonstrated the use of a mixture of two normal distributions to represent the frequency curve of annual flood peaks. The justification for using a mixture model for representing flood peaks in certain geographical areas was that floods may be generated by a number of different meteorological phenomena such as spring snowmelt or hurricanes. It has been presumed, but not rigorously demonstrated, that floods arising from different hydrometeorological causes can be treated as samples from different distributions. In this particular example reasonably reliable exogenous information exists to classify data, but it seems that such data have not yet been used rigorously either to justify the model selection or to estimate model parameters.

Interest in the use of mixture models for water resources planning was aroused when Klemes (1974) and Potter (1976) demonstrated that time series exhibiting the Hurst phenomenon could be created by sampling from certain classes of mixture models. It is evident from this work that non-stationarity in the mean could provide an explanation for the long-term persistence which naturally occuring values of the Hurst coefficient imply. The use of mixture models by Klemes and Potter can be thought of either as a means for mimicking a particular autocorrelation structure or as tentative evidence that non-stationarity of the mean provides a physical explanation of the Hurst effect, depending on one's viewpoint. In later work Boes and Salas (1978) developed a general mixture model for shifting means. They demonstrated by use of numerical experiments that such models not only mimic the Hurst phenomenon but also have a correlation structure identical to an ARMA (1,1) process (Box and Jenkins 1976), as used by O'Connell (1971, 1974) to model the Hurst effect.

The principal reason for interest in replicating the Hurst effect in synthetic streamflows was the observation that the early Markov or autoregressive order one AR(1) models did not produce droughts as severe as those encountered in natural time series. Jackson (1975b) used a two state Markov mixture model to explicitly control the distribution of synthetic low flow periods. The justification for using a mixture model was the observation that low and high mean annual flows do not necessarily come from the same distribution. In this case the distinction between low and high flows is subjective and exogenous information for classifying the data do not exist. Thus Jackson's application may again be viewed largely as a curve fitting exercise in which the parameters of the Markov mixture model are chosen so that the generated data replicate certain statistics of observed drought sequences. Jackson did not investigate in any detail the marginal distributions of her mixture model.

One purpose of this introductory review was to establish that applications of mixture models can range from curve-fitting exercises to situations where there is a strong physical basis for using mixtures and where objective exogenous information exists for accurately classifying the data. As demonstrated by Hosmer (1973b), the availability of exogenous information to classify even a small part of a mixed sample may be of considerable value in estimating model parameters.

As was discussed in Chapter 3, there is tentative evidence suggesting that precipitation comes from a mixed distribution. Unfortunately, attempts to classify precipitation data into wet and dry distributions using features of the atmospheric circulation were unsuccessful. The analyses in this chapter investigate the possibility of using mixture models for single-site modeling of precipitation in the absence of exogenous information for classifying the data.

The remainder of this chapter explores the characteristics and use of mixture models in detail. Section 6.1 discusses the population characteristics of mixture models and demonstrates their great flexibility. Section 6.2 discusses techniques for identifying mixtures. Section 6.3 discusses methods of parameter estimation with emphasis on the use of maximum likelihood estimates. Section 6.4 explores the small sample characteristics of maximum likelihood parameter estimates and the chapter finishes with examples of the application of mixture models to represent annual and monthly rainfall data in the Pacific Northwest.

6.1 Characteristics of Mixture Distributions

Mixture models form a vast class of distributions. In this study I will be concerned exclusively with mixtures of two univariate normal distributions. As will be seen, even with this restriction, we have a large and flexible family of distributions.

6.1.1 Independent Mixture Models

Suppose we have a sequence of n observations $y = (y_1, y_2, \ldots, y_n)$ where each y_i is associated with one of two unobserved states. Let the sequence of states associated with y be represented by the outcome of n identical and independent Bernoulli trials such that the probabilities:

$$P(y_i \text{ is associated with state 1}) = p_1$$
$$P(y_i \text{ is associated with state 2}) = p_2 = (1-p_1)$$

Further assume that the y_i given the associated state are conditionally independent with normal densities:

$$g(y|\text{state } 1) = f_1(y) = (2\pi\sigma_1^2)^{-\frac{1}{2}} \exp[-(y-\mu_1)^2/2\sigma_1^2]$$
$$g(y|\text{state } 2) = f_2(y) = (2\pi\sigma_2^2)^{-\frac{1}{2}} \exp[-(y-\mu_2)^2/2\sigma_2^2] \qquad (6.1)$$

Then the unconditional marginal density of the y_i is the mixed distribution:

$$\begin{aligned} f(y) &= p_1f_1(y) + p_2f_2(y) \\ &= p_1f_1(y) + (1-p_1)f_2(y) \end{aligned} \qquad (6.2)$$

Denoting the overall mean and variance of the mixed distribution by μ and σ^2 respectively, the overall mean is:

$$\begin{aligned} \mu = E(y) &= \sum_{i=1}^{2} E(y|\text{state i})P(\text{state i}) \\ &= p_1\mu_1 + (1-p_1)\mu_2 \end{aligned} \qquad (6.3)$$

and the overall variance is:

$$\sigma^2 = E(y^2) - (E(y))^2$$
$$\begin{aligned} E(y^2) &= \sum_{i=1}^{2} E(y^2|\text{state i})\ P(\text{state i}) \\ &= p_1(\sigma_1^2 + \mu_1^2) + (1-p_1)(\sigma_2^2 + \mu_2^2) \end{aligned}$$

which gives

$$\sigma^2 = p_1(\sigma_1^2 + \mu_1^2) + (1-p_1)(\sigma_2^2 + \mu_2^2) - \mu^2 \qquad (6.4)$$

Expressions for higher moments are given by Cohen (1967). It is a simple exercise to show that the serial correlation coefficient for this model is zero for all lags.

6.1.2 Markov Mixture Models

In the independent mixture model of the previous section, the sequence of states was chosen by using Bernoulli trials. A wide variety of other techniques may be used for selecting states. One particularly attractive option is to represent changes of state by` means of a two-state Markov chain. Models of this type have been studied in some detail by Jackson (1975b).

The single-step transition matrix for the two-state chain of Figure 6.1 is:

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \tag{6.5}$$

where a = probability of transition from state 1 to state 2 in a single step.

b = probability of transition from state 2 to state 1 in a single step.

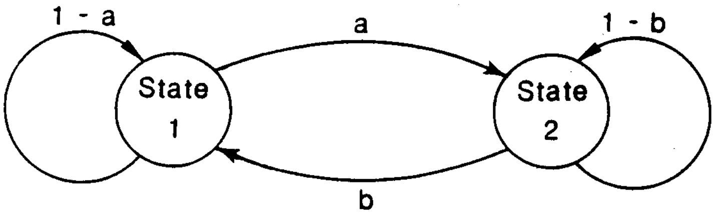


Figure 6.1 Two-state Markov chain

Use of this model requires specification of an initial or starting state. The probability of being in a particular state at successive steps or intervals can then be found by repeated applications of the transition matrix. After a large number of steps, the long-run or invariant distribution u is given (Feller 1968) by:

$$u = Pu \tag{6.6}$$

where $u = (u_1 \; u_2)$

u_i = long run probability of being in state i

$\sum_i u_i = 1$

Substituting for u and P in (6.6) gives:

$$(u_1 \; u_2) \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} = (u_1 \; u_2)$$

$$\text{hence} \quad = (u_1 \; u_2) = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}$$

$$= (p_1 \quad 1-p_1) \tag{6.7}$$

Thus the transition probabilities of the Markov chain are related to the state probabilities of the independent mixture model by:

$$p_1 = \frac{b}{a+b} \tag{6.8}$$

Given a particular state i, the observations y_i may be sampled in the same manner as for the independent mixture model, in which case the marginal densities of the observations for the two model types are identical.

The overall lag-one serial correlation for the model may be derived as follows:

Let x_i = random variable representing observation in interval i

$r_i \rightsquigarrow s_j$ = transition from state r in interval i to state s in interval j

The expectation

$$E(x_i x_{i+1}) = \sum_{r=1}^{2} \sum_{s=1}^{2} E(x_i x_{i+1} | r_i \rightsquigarrow s_{i+1}) \; P(r_i \rightsquigarrow s_{i+1})$$

$$= \mu_1^2 \, p_1(1-a) + \mu_1 \mu_2 p_1 a + \mu_2 \mu_1 (1-p_1) b$$

$$+ \mu_2^2 (1-p)(1-b) \qquad (6.9)$$

and the-lag one autocorrelation

$$\rho(1) = \frac{E(x_i x_{i+1}) - \mu^2}{\sigma^2} \qquad (6.10)$$

where μ^2 is given by Equation 6.3 and σ^2 by Equation 6.4.

The lag-two autocorrelation can be found in a similar manner by making use of the two-step transition matrix

$$P^2 = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

$$= \begin{pmatrix} (1-a)^2 + ab & (1-a)a + a(1-b) \\ b(1-a) + b(1-b) & ab + (1-b)^2 \end{pmatrix} \qquad (6.11)$$

Then

$$E(x_i x_{i+2}) = \sum_{r=1}^{2} \sum_{s=1}^{2} E(x_i x_{i+2} | r_i \rightsquigarrow s_{i+2}) \; P(r_i \rightsquigarrow s_{i+2})$$

$$= \mu_1^2 \, p_1 \left[(1-a)^2 + ab\right] + \mu_1 \mu_2 p_1 \left[(1-a)a + a(1-b)\right] +$$

$$\mu_2 \mu_1 (1-p_1) \left[b(1-a) + b(1-b)\right] + \mu_2^2 (1-p_1) \left[ab + (1-b)^2\right] \qquad (6.12)$$

with the lag-two autocorrelation

$$\rho(2) = \frac{E(x_i x_{i+2}) - \mu^2}{\sigma^2} \qquad (6.13)$$

Autocorrelations for higher lags can be found in a similar fashion. Note that the autocorrelation is a function of the transition probabilities a and b, and that once the lag one auto correlation is chosen, the remainder of the correlation function is fixed. Figure 6.2 shows the autocorrelation function for Markov mixture models with parameters $p_1 = 0.3$, $\mu_1 = -1.0$, $\sigma_1 = 1.0$, $\mu_2 = 1.0$, $\sigma_2 = 1.0$ and values of the transition probability a = 0.1, 0.2, 0.3. (Note that once a and p_1 are fixed b is given by the relationship $p_1 = b/(a+b)$). For comparative purposes, the correlation function for an AR(1) model with $\rho(1) = 0.39$ is also shown (for a = 0.1 the mixture model has $\rho(1) = 0.39$). Note that for the parameter sets used in Figure 6.2, the correlation function of the mixture model drops off much more slowly than that of the AR(1) model.

A wide variety of other approaches to mixture modeling are possible. Jackson (1975b), for example, suggests a Markov mixture model in which the observations in a given state are not independent but are represented by an AR(1) structure. The practicality of such complex models is doubtful, however, especially when one considers the difficulties encountered in model identification and parameter estimation for the independent mixture model.

6.2 Detection of Mixtures

The ease with which mixtures can be detected depends largely on the (unknown) parameters of the distribution and on the sample size available. Where the components of the mixture are widely separated the distribution function will generally be multi-modal and the detec-

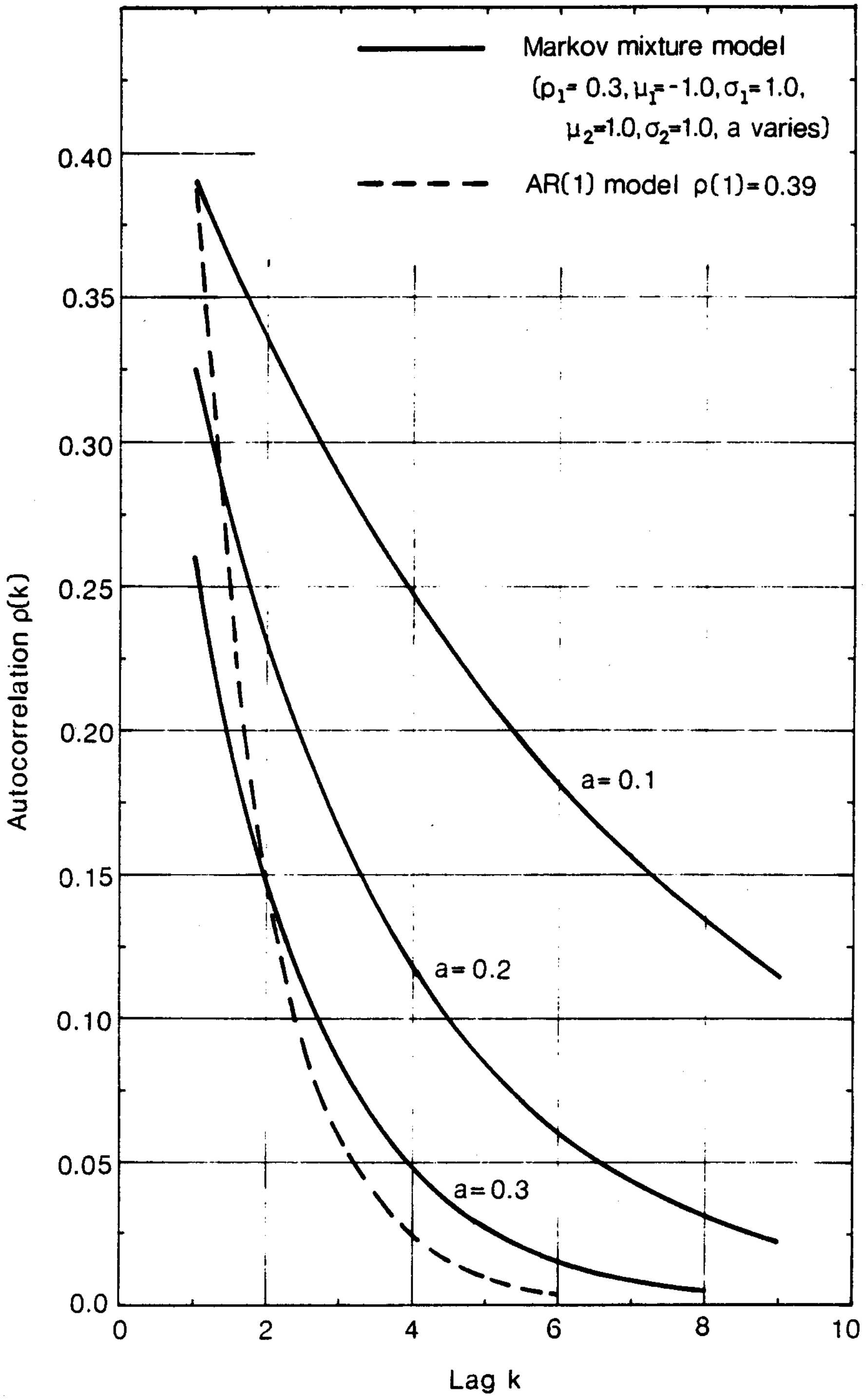


Figure 6.2 Autocorrelation functions for Markov mixture models

tion of mixtures presents no difficulty. However, if the components are not widely separated, the detection of mixtures and estimation of their parameters pose severe problems which increase with the complexity of the model.

Detection of mixtures of two normal distributions has been investigated in considerable detail by Fowlkes (1977, 1979). One useful diagnostic tool is the quantile-quantile (Q-Q) plot (Wilk and Gnanadesikan 1968). The theoretical Q-Q plot of the mixture quantiles plotted against standard normal quantiles is generally S-shaped, but the variety of such shapes ranges from a barely perceptible kink, in what is otherwise a straight line, to greatly transmogrified S-shapes, depending on the parameters of the mixture. Fowlkes (1977) shows the full variety of possible shapes in a compendium of Q-Q plots covering a wide range of parameter values.

The approximate shape of a mixture's Q-Q plot can be deduced by recognizing that at either extreme the plot is asymptotic to a straight line representing the Q-Q plot of the component distributions. This feature is useful for making initial parameter estimates as discussed by Fowlkes and as demonstrated in Section 6.5.

Theoretical Q-Q plots for two different mixtures and their component distributions are shown in Figures 6.3 and 6.4. (The parameter sets used for these examples were obtained from the rainfall data analyzed in Section 6.5).

One can expect sample Q-Q plots to closely resemble the theoretical Q-Q plots only when sample sizes are large (perhaps more than 500 observations for a subtle mixture). For small samples, the sampling variability of mixtures is known in general to be large, and sample Q-Q plots may not resemble the theoretical plots. In fact the sample Q-Q plot may not indicate the presence of a mixture. This is illustrated in Figure 6.5 which shows the Q-Q plot for a set of 81 observations sampled from the mixture of Figure 6.3. The method of sampling is described in Section 6.5. A sample size of 81 was used because this was the maximum length of the annual rainfall series obtained for use in this study.

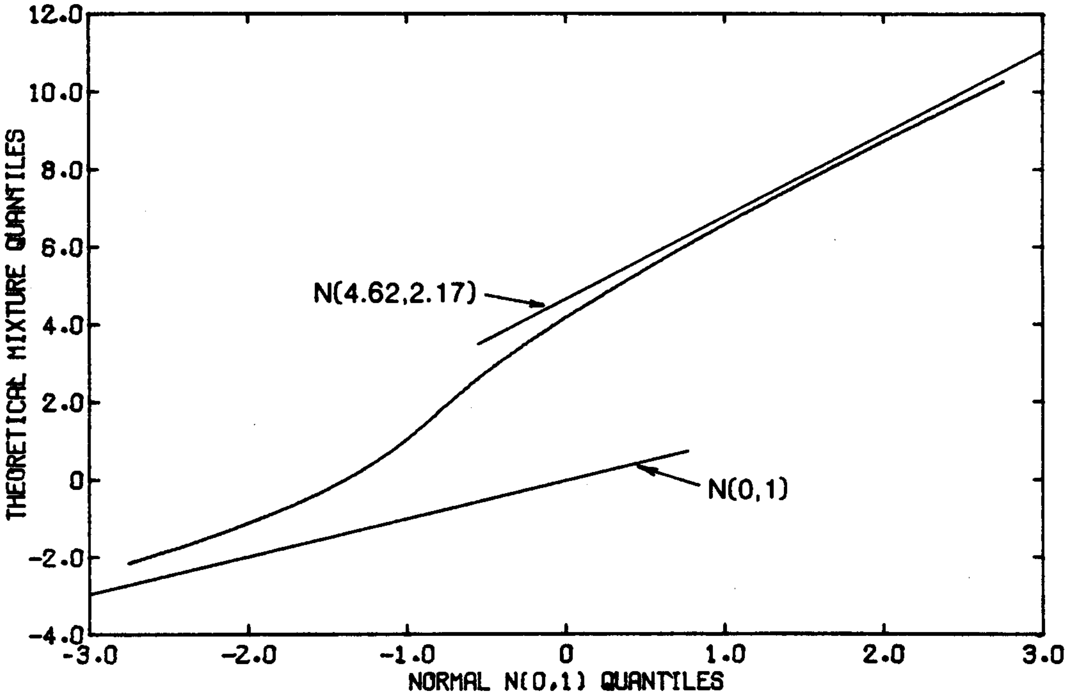


Figure 6.3 Theoretical Q-Q plot for a mixture of two normal distributions with parameters (0.14, 0.0, 1.0, 4.62, 2.17)

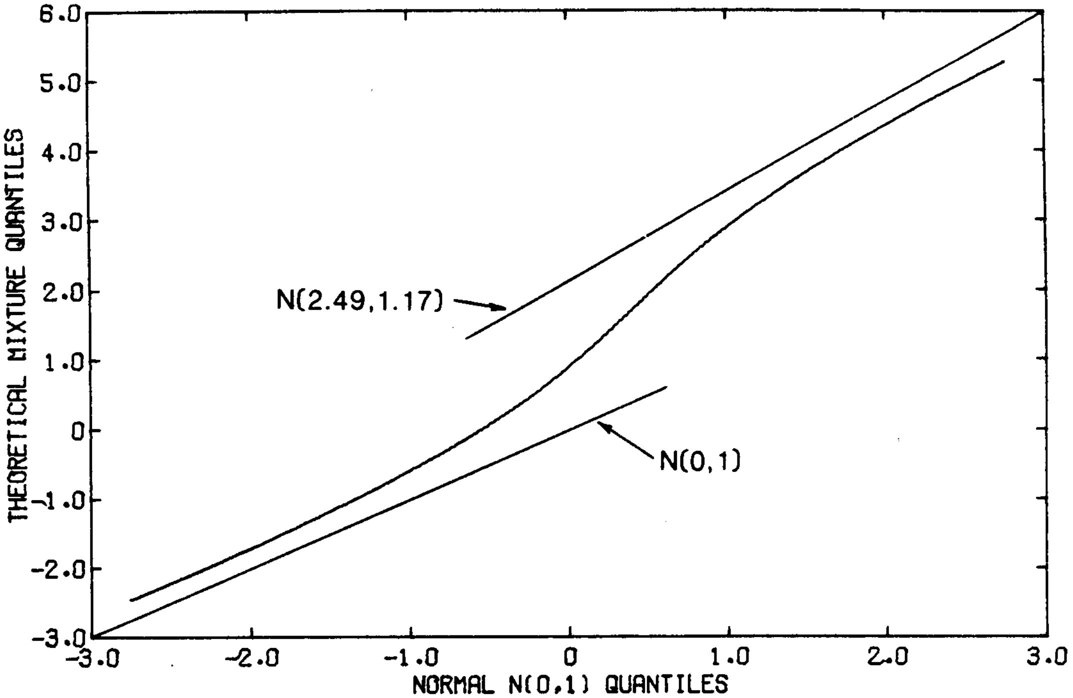


Figure 6.4 Theoretical Q-Q plot for a mixture of two normal distributions with parameters (0.57, 0.0, 1.0, 2.49, 1.17)

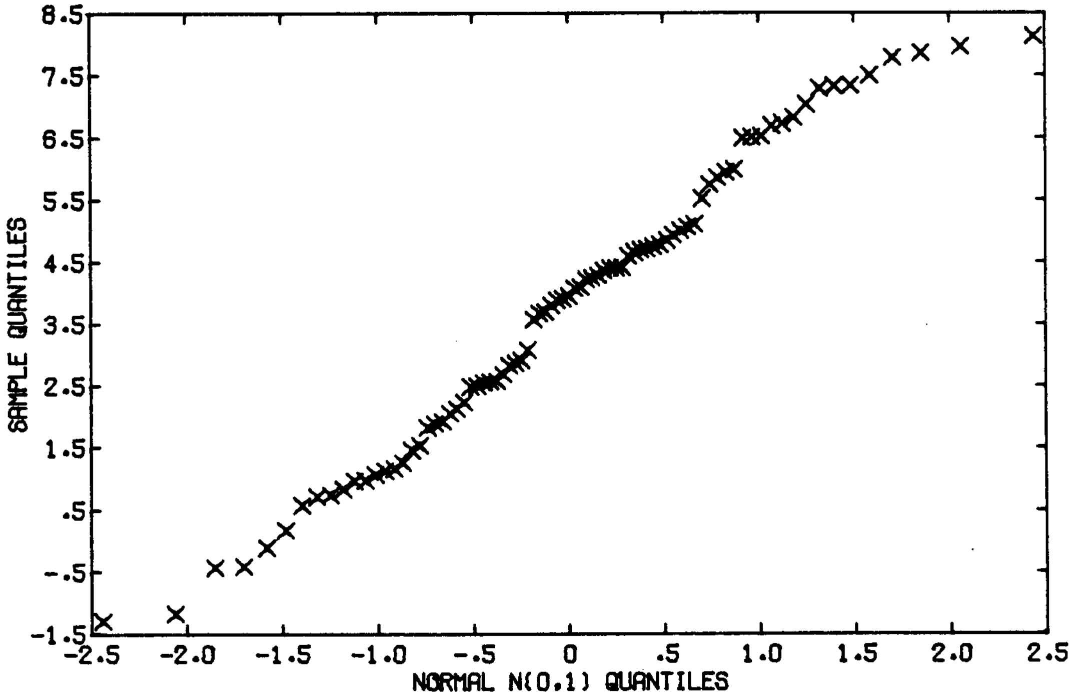


Figure 6.5 Sample Q-Q plot for a sample of size 81 from a mixture of two normal distributions with parameters (0.14, 0.0, 1.0, 4.62, 2.17)

Fowlkes (1977, 1979) presents another diagnostic tool which may be useful for detecting mixtures. This is a variant of the percentile-percentile plot named by Fowlkes, the Φ-p versus Q plot. For the sake of brevity this will be referred to here as a P-Q plot. The procedure for constructing the P-Q plot adapted from Fowlkes is as follows. Let

$$y_{(1)} < y_{(2)} \ldots < y_{(n)}$$

be ordered quantiles of a sample of size n. Then using the plotting formula of Cunane (1978) the corresponding sample percentiles are:

$$p_{(i)} = \frac{i - 0.4}{n + 0.2}$$

Let $\bar{y}$ and s_y represent the sample mean and sample standard deviation. The x coordinates of the P-Q plot are taken as $x_i = (y_{(i)} - \bar{y})/s_y$ and the y coordinates are:

$$N(x_i \mid 0,1) - p_{(i)}$$

where $N(x_i|0,1)$ is the cumulative distribution function for the normal N(0,1) distribution evaluated at x_i. The P-Q plot is thus a plot of the standard normal percentiles minus sample percentiles versus the standardized sample quantiles.

The P-Q plot is designed to be sensitive to departures from normality in the middle quantiles. The theoretical P-Q plot for a normal distribution is simply a horizontal line passing through zero. Theoretical P-Q plots for the two mixtures of Figures 6.3 and 6.4 are shown in Figures 6.6 and 6.7. The principal feature of the theoretical plot for the mixtures is the oscillation of the line about the zero ordinate in the middle quantiles and convergence of the line to zero for large positive and negative values of the abscissa.

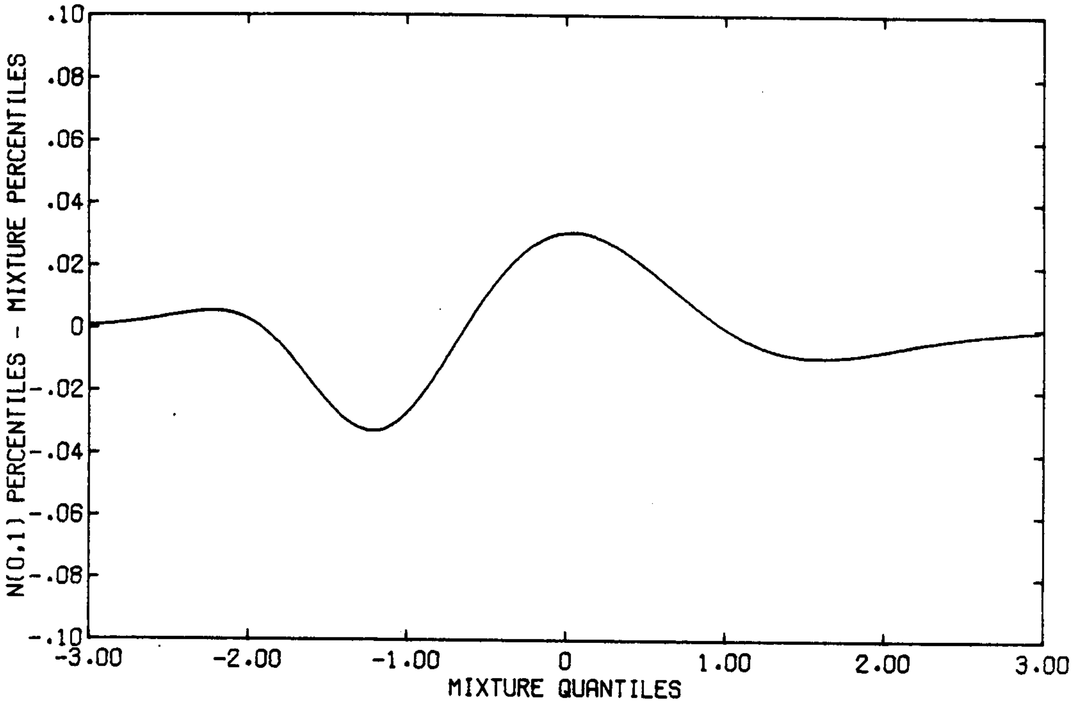


Figure 6.6 Theoretical P-Q plot for a mixture of two normal distributions with parameters (0.14, 0.0, 1.0, 4.62, 2.17)

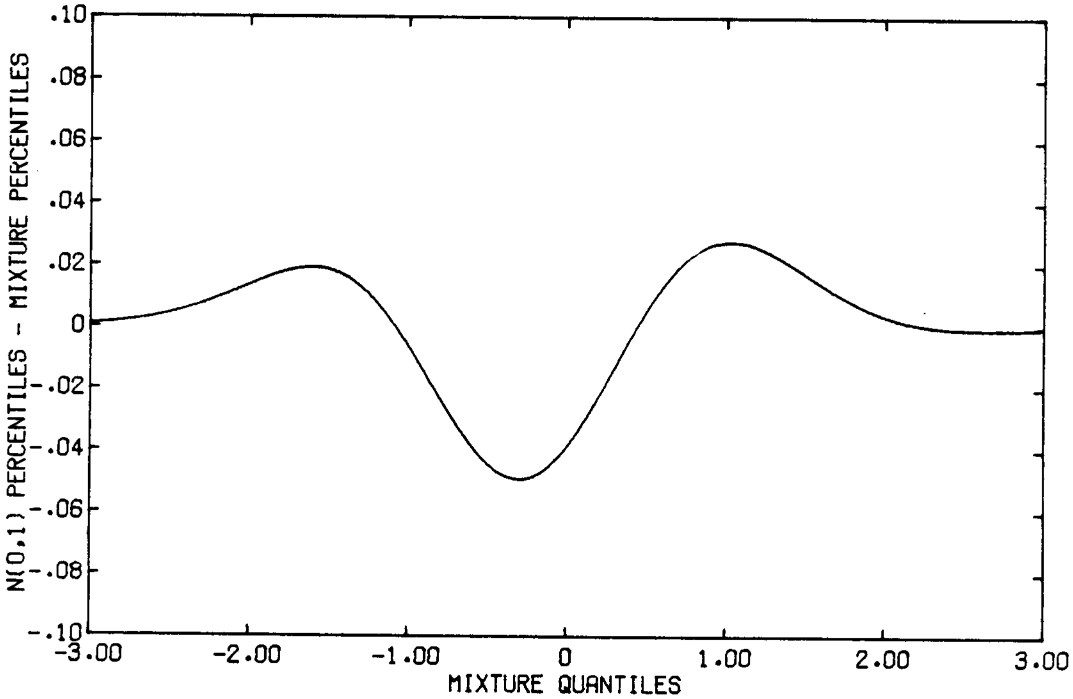


Figure 6.7 Theoretical P-Q plot for a mixture of two normal distributions with parameters (0.57, 0.0, 1.0, 2.49, 1.17)

Fowlkes (1977) explored the characteristics of both theoretical and sample P-Q plots in considerable detail, and demonstrated that, for large sample sizes, P-Q plots can identify successfully quite subtle mixtures. However, for the small samples usually available from hydrologic records, P-Q plots like the Q-Q plots, can exhibit pronounced variability. Again this is illustrated in Figure 6.8 for the set of 81 observations used to construct the Q-Q plot of Figure 6.5. The theoretical P-Q plot of the mixture from which this sample was taken is shown in Figure 6.6. The sample P-Q plot bears little resemblance to its theoretical counterpart, nor does it resemble any of the theoretical P-Q plots presented by Fowlkes for a variety of other distributions.

It is clear from this brief examination that Q-Q and P-Q plots are of limited value in detecting mixtures when sample sizes are small. It appears that for small samples, the assumption of a mixed distribution must be based on physical rather than statistical grounds. However, even if strong physical and/or statistical evidence exists for justifying the use of a mixture, estimation of the mixture parameters presents great difficulties. In the next section I discuss various approaches to parameter estimation and in Section 6.4 derive estimates for characterizing sampling variability.

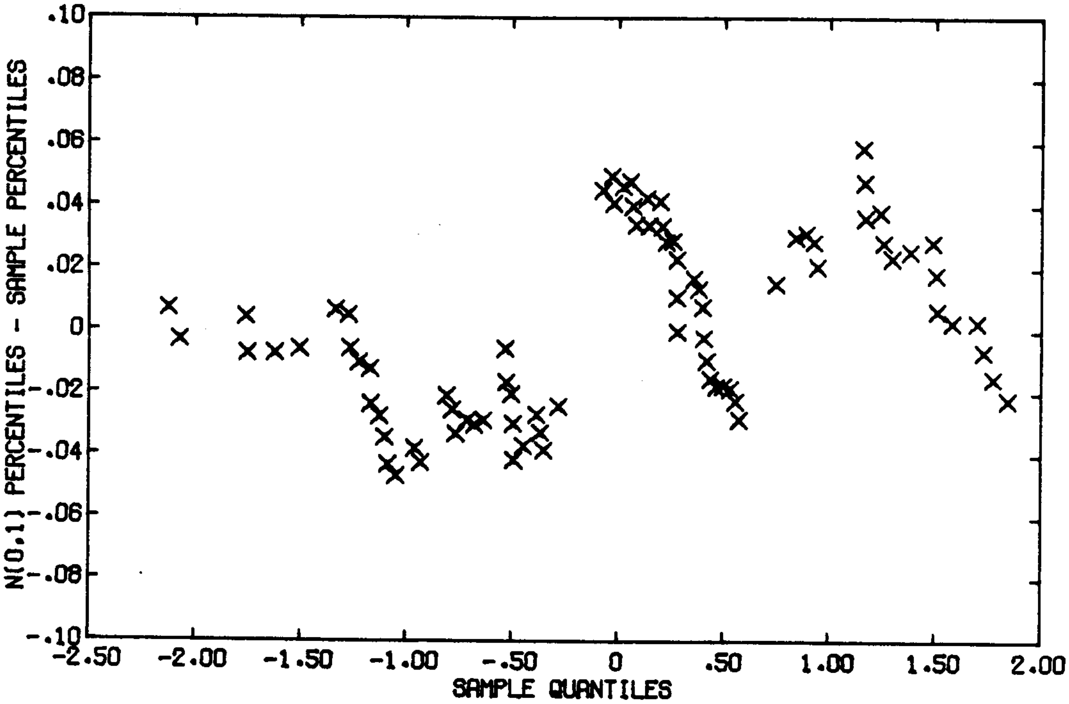


Figure 6.8 Sample P-Q plot for a sample of size 81 from a mixture of two normal distributions with parameters (0.14, 0.0, 1.0, 4.62, 2.17)

6.3 Parameter Estimation

Four principal methods for estimating the parameters of mixtures appear in the literature. These are:

(1) graphical techniques (Fowlkes 1977, 1979)

(2) method of moments (Cohen 1967)

(3) moment generating function method, MGF (Quandt and Ramsey 1978)

(4) maximum likelihood method (Hasselblad 1966, 1969; Hosmer 1973a; Dempster, et al. 1977)

Because of the computational difficulties involved much of the early work in estimating mixture parameters was devoted to graphical techniques. With improvements in computational facilities it became feasible to use moment estimators and maximum likelihood estimators and more recently researchers have suggested estimators based on fits to the theoretical moment generating function or characteristic function. A brief review of developments in the estimation of mixture parameters is provided by Fowlkes (1977).

This section briefly describes two methods for parameter estimation; a graphical method and the moment generating function method of Quandt and Ramsey (1978). The maximum likelihood approach is also discussed in detail. Work by Tan and Chang (1970) and others has demonstrated that the method of maximum likelihood is superior to the method of moments. Consequently, the method of moments will not be discussed.

6.3.1 Graphical Techniques

Graphical methods were the main tool for early work in estimating mixture parameters. The principal function of graphical techniques today, however, is to provide reasonable initial estimates of parameters for use with more powerful estimation techniques such as maximum likelihood.

The most successful graphical method for parameter estimation for mixtures of two normal distributions is based on the Q-Q plot of the sample quantiles plotted against standard normal quantiles. The idea behind this technique was hinted at in Section 6.2. Figures 6.3 and 6.4 showed theoretical Q-Q plots of mixture quantiles plotted against standard normal quantiles. The figures also showed the Q-Q plots of the component distributions of the mixture. As pointed out earlier the upper and lower limbs of the S-shaped Q-Q plot of a mixture are asymptotic to the straight Q-Q plots of its normal components. Given a sample from an unknown mixture, the means and variances of the constituent distributions can thus be estimated by fitting straight line asymptotes to the upper and lower limbs of the sample Q-Q plot. The means and variances are of course found from the slope and intercept of those lines.

The mixing proportion p_1 remains to be estimated from the sample. An approximate value for p_1 can be found by first determining the x-coordinate of the point of inflection of the sample Q-Q plot. It will be recalled that the x-coordinate is a standard N(0,1) quantile. The mixing proportion p_1 is then estimated by the value of the N(0,1) cumulative distribution function corresponding to that quantile.

It has been pointed out by Fowlkes (1977) that the above estimate of p_1 is generally biased with both the sign and magnitude of the bias being a function of the mixture parameters. The estimate is only unbiased for $p_1 = 0.5$ and $\sigma_1 = \sigma_2$, although the bias remains small for $0.3 \leq p_1 \leq 0.7$ and $\sigma_1 = \sigma_2$.

A variety of methods exist for fitting asymptotes to the sample Q-Q plot and for determining the point of inflection of the plot.

Rough estimates of the parameters can be obtained by simply fitting by eye; this is the method used in Section 6.5. Fowlkes (1977), however, has investigated the graphical approach to parameter estimation in great detail and presents a rather sophisticated technique which involves fitting a logistic distribution to the sample Q-Q plot. The reader is referred to Fowlkes for details of that method.

6.3.2 Moment Generating Function Estimators

The moment generating function estimator for mixture parameters was introduced by Quandt and Ramsey (1978) as an alternative to the maximum likelihood and moment estimators. The basis of the method is to determine that set of parameters which minimizes the sum of squares of differences between the theoretical and sample moment generating function. For the mixture distribution:

$$f(y) = p_1(2\pi\sigma_1^2)^{-\frac{1}{2}} \exp[-(y-\mu_1)^2/2\sigma_1^2]$$

$$+ (1-p_1)(2\pi\sigma_2^2)^{-\frac{1}{2}} \exp[-(y-\mu_2)^2/2\sigma_2^2] \qquad (6.14)$$

The moment generating function (MGF) is:

$$E(e^{\theta y}) = p_1 \exp(\mu_1\theta + \sigma_1^2\theta^2/2)$$

$$+ (1-p_1) \exp(\mu_2\theta + \sigma_2^2\theta^2/2) \qquad (6.15)$$

The MGF method from Quandt and Ramsey (1978) is as follows: Given a sample $y = (y_1, y_2, \ldots y_n)$ choose k values of $\theta, \theta_1, \theta_2, \ldots \theta_k$ in a small interval (a,b), $a<0<b$. For any value of θ, θ_j, the value of the moment generating function may be estimated by:

$$E(e^{\theta_j x}) = \frac{1}{n}\sum_{i=1}^{n} \exp(\theta_j y_i) \qquad (6.16)$$

The parameters of the mixture are then estimated by the set of parameters $(p_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ which minimizes:

$$S(\theta) = \sum_{j=1}^{k} \left[\left(\frac{1}{n} \sum_{i=1}^{n} \exp(\theta_j y_i) \right) - p_1 \exp(\mu_1 \theta_j + \sigma_1^2 \theta_j^2 / 2) - (1-p_1) \exp(\mu_2 \theta_j + \sigma_2^2 \theta_j^2 / 2) \right]^2 \qquad (6.17)$$

Choice of k and the values $\theta_1, \theta_2, \ldots \theta_k$ present difficulties in application of the MGF method. As general guidelines Quandt and Ramsey recommend using a value $k \geq 5$. The choice of θ_j, however, appears to be quite important and sensitive to the data being used. If the θ_j are too small the function to be mimimized (Equation 6.17) is quite flat causing difficulties in the precise determination of the minimum. If the θ_j are too large, then computational overflows can occur in the exponential terms of Equation 6.17. Optimal values for the θ_j are not currently available. Moreover, it should be noted from Equation 6.17 that the optimal θ_j will depend on the values of the data used to estimate the generating function.

Quandt and Ramsey performed a Monte Carlo study to compare the small sample properties of the MGF estimator and method of moments. The MGF estimator was shown to be superior to the method of moments, but as yet no equivalent comparison has been made with the maximum likelihood estimators which, as noted earlier, are also superior to the moment estimators. The MGF method was implemented for testing on an HP3000 Series III mini computer using the simplicial minimization technique presented by Nelder and Mead (1965). The method was tested on samples of size 100 taken from a relatively tractable mixture with population parameters (0.5, 0,1,4,1) for k=5 and θ values of -0.2, -0.1, 0.1, 0.2, 0.3. The starting parameters for minimizing Equation 6.17 were taken to be either the population parameters or some arbitrary set of parameters close to the population parameters.

Difficulties were encountered in a number of respects. Among the problems was a tendency for the search technique to converge occasionally on infeasible parameter values such as negative values for p_1. The objective function surface was also found to be almost horizontal in some instances, causing premature termination of the search procedure. In view of these difficulties and the absence of optimal values for θ and k, use of the MGF method was discontinued in favor of the better understood maximum likelihood method.

6.3.3 Maximum Likelihood Estimators

The maximum likelihood estimates for mixture parameters can be found by means of the iterative algorithm presented by Hasselblad (1966, 1969) and others as follows:

Let $y = (y_1, y_2, \ldots, y_n)$ be a sample of n independent observations from the mixture distribution

$$f(y_i|\phi) = p_1 f_1(y_i|\phi) + p_2 f_2(y_i|\phi) \tag{6.18}$$

where $0 < p_1 < 1$; $p_2 = 1-p_1$

$$f_j(y_i|\phi) = (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp[-(y_i-\mu_j)^2/2\sigma_j^2] \quad ;j=1,2$$

$$\phi = (p_1,\mu_1,\sigma_1,\mu_2,\sigma_2)$$

Let $\phi^{(v)} = (p_1^{(v)}, \mu_1^{(v)}, \sigma_1^{(v)}, \mu_2^{(v)}, \sigma_2^{(v)})$ be parameter estimates at iteration v with $\phi^{(0)}$ being a set of initial parameter estimates

At iteration v define weights

$$w_{ij}^{(v)} = \frac{p_j^{(v)} f_j(y_i|\phi^{(v)})}{\sum_{j=1}^{2} p_j^{(v)} f_j(y_i|\phi^{(v)})} \qquad ;\begin{array}{l} i=1,\ldots,n \\ j=1,2 \end{array} \tag{6.19}$$

then the parameter estimates at iteration (v+1) are:

$$p_j^{(v+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^{(v)} \qquad ;j=1,2 \qquad (6.20)$$

$$\mu_j^{(v+1)} = \sum_{i=1}^{n} w_{ij}^{(v)} y_i \Big/ \sum_{i=1}^{n} w_{ij}^{(v)} \qquad ;j=1,2 \qquad (6.21)$$

$$\sigma_j^{(v+1)} = \left[\left(\sum_{i=1}^{n} w_{ij}^{(v)} (y_i - \mu_j^{(v)})^2\right) \Big/ \sum_{i=1}^{n} w_{ij}^{(v)}\right]^{\frac{1}{2}} \qquad ;j=1,2 \qquad (6.22)$$

Iterative calculations are continued until some suitable convergence criteria are met.

Although the algorithm presented above always converges to a maximum on the likelihood surface, Hasselblad was unable to prove such convergence. More recently, however, Dempster, Laird, and Rubin (1977), herein referred to as DLR, proved that Hasselblad's iterative algorithm is an EM (expectation-maximization) algorithm and, as such, is guaranteed to converge to at least a local maxima. A detailed derivation of the above MLE's in terms of an EM algorithm has not, to my knowledge, appeared in either the statistics or water resources literature. Consequently, a detailed development of the MLE's is given in Appendix A. (It should be noted that DLR's work on the EM algorithms has wide applicability for computing maximum likelihood estimates from incomplete data, and its application to the estimation of mixture parameters is only one of many possible uses.)

The EM algorithm has a number of attractive features:

(1) The algorithm is guaranteed to converge to a local if not a global maximum. The speed of convergence is however a function of the separation of the mixture components and for components which are close together, the convergence may be extremely slow. DLR suggest a number of approaches for increasing the rate of convergence.

(2) Every iteration of the algorithm is guaranteed to increase the log-likelihood.

(3) The iterative estimates always yield valid parameter values, i.e., positive variance and mixing proportion between zero and one.

In addition, the iterative estimates have the usual attractive asymptotic properties of all MLE's. For further details and proofs associated with the EM algorithm the reader is referred to Appendix A and Dempster, Laird and Rubin (1977).

Potential problems in the use of MLE's are:

(1) The presence of singularities on the likelihood surface. In some situations the iterative algorithm will converge to parameter values associated with a singularity. (The variance and mixing proportion of one component goes to zero.) Although this may be a problem from a mathematical point of view, it has been found that in practice singularities do not present serious difficulties and valid MLE's generally can be found. The Monte Carlo experiments of Section 6.4 showed that difficulties with singularities increase as the sample size decreases and as the separation of components decreases. However, even with a sample size of

50 and with sampling from a relatively subtle mixture with parameters (0.3, 0, 1, 2, 1) it was found that singularities presented no serious difficulties. As pointed out by a number of investigators (e.g., Hosmer in the discussion of Quandt and Ramsey 1978), it appears that the problem of singularities on the likelihood surface has been exaggerated in the past.

(2) Like all other currently available estimation procedures, the EM algorithm does not guarantee convergence to a global maximum. A global maximum can only be ensured by an exhaustive search of the five dimensional likelihood surface.

(3) Little is known about the small sample properties of the MLE's. Since hydrometeorologic records are generally less than 50 years in length, the small sample properties of estimators are of great concern. Consequently, the small sample properties of the maximum likelihood estimates are investigated in the Monte Carlo study described in Section 6.4.

6.4 Small Sample Properties of the Maximum Likelihood Estimates of Mixture Parameters

In Section 6.3, I discussed the iterative EM algorithm for the maximum likelihood estimates of a mixture of two normal distributions. Although MLE's have attractive large sample properties, their small sample properties for mixtures have received comparatively little attention.

Hosmer (1973a, 1973b) conducted limited simulation studies of the small sample properties of MLE's. He reported pessimistic results indicating that parameter estimates are probably unreliable for sample sizes less than 300 and for parameter sets such that

$$|\mu_2 - \mu_1| < 3 \min (\sigma_1, \sigma_2)$$

Hosmer's results were based on Monte Carlo experiments with only ten samples of size 100 for each parameter set studied. In view of the large sampling variabilities reported by Hosmer, this sample is far too small to allow reliable estimates of the mean or variance of the mixture parameters. Accordingly this section presents the methods and results for a more extensive Monte Carlo study of the small sample parameter estimates. The work was done using a CDC CYBER 170-750 computer.

The parameter sets and sample sizes used in the Monte Carlo study are shown in Table 6.1. For each parameter set and sample size, 200 samples were generated as follows: For a sample of size n, n independent uniform pseudo-random numbers on the interval [0,1) were generated using the CDC RANF random number generator. These were then converted to normal N(0,1) independent pseudo-random numbers using the Box-Muller transformation (Box and Muller 1959) as presented by Forsythe, et al. (1972). For each N(0,1) random number, a further uniform random number was generated using RANF. If this uniform random number was greater than parameter p_1, the N(0,1) variate was converted to an $N(\mu_2,\sigma_2)$ variate, otherwise it was left as an N(0,1) variate. We have thus created a sample of size n from the mixture distribution:

$$f(x_i) = p_1 \, f_1(x_i) + (1 - p_1)f_2(x_i)$$

where $f_1 \sim N(0,1)$

$f_2 \sim N(\mu_2,\sigma_2)$

Table 6.1 Monte Carlo Experiments to Investigate the Characteristics of the Maximum Likelihood Estimates of the Parameters of Mixtures of Two Normal Distributions

Expt	Population Parameters					Starting Values					Sample Size	No. of Samples
	P_1	μ_1	σ_1	μ_2	σ_2	P_1	μ_1	σ_1	μ_2	σ_2		
1	0.3	0	1	3	1	0.3	0	1	3	1	100	200
2	0.3	0	1	3	1	0.3	0	1	3	1	50	200
3	0.3	0	1	2	1	0.3	0	1	2	1	50	200
4	0.3	0	1	3	1.5	0.3	0	1	3	1.5	100	200
5	0.3	0	1	3	2	0.3	0	1	3	1	50	200

Note: The starting values for the iterative solution of the MLE's were taken to be the population parameters except, inadvertently, in experiment 5. Experiment 5 was repeated for the first 20 samples using starting values (0.3, 0, 1, 3, 2). The MLE's obtained were essentially the same as those obtained with starting values (0.3, 0, 1, 3, 1).

This approach to sampling from a mixture distribution allows one to keep track of the component distribution from which each observation originated. For each sample we can thus determine the exact proportion of observations sampled from each distribution and estimate directly the mean and variance of the component distributions. This is equivalent to the situation in practical applications where reliable exogenous information exists to classify the data. Where such information exists, I will refer to the data as fully classified data, and to the estimated statistics as fully classified statistics.

The mixture parameters for the unclassified samples were estimated using the iterative maximum likelihood EM algorithm discussed in Section 6.3.3. The procedure adopted was as follows:

(1) Make an initial estimate of the parameter set

$$\phi^{(0)} = (p_1^{(0)}, \mu_1^{(0)}, \sigma_1^{(0)}, \mu_2^{(0)}, \sigma_2^{(0)})$$

For this study the population parameters were used for the initial parameter estimates.

(2) At iteration v of the procedure the parameter estimates are found as follows:

Define weights

$$w_{ij}^{(v)} = \frac{p_j^{(v)} f_j(y_i|\phi^{(v)})}{f(y_i|\phi^{(v)})} \qquad ;j=1,2 \text{ and } i=1,\ldots,n$$

then

$$p_j^{(v+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^{(v)} \qquad ;j=1,2$$

$$\mu_j^{(v+1)} = \sum_{i=1}^{n} w_{ij}^{(v)} y_i \Big/ \sum_{i=1}^{n} w_{ij}^{(v)} \qquad ;j=1,2$$

$$\sigma_j^{(v+1)} = \left[\left(\sum_{i=1}^{n} w_{ij}^{(v)} (y_i - \mu_j^{(v)})^2\right) \Big/ \sum_{i=1}^{n} w_{ij}^{(v)}\right]^{\frac{1}{2}} \qquad ;j=1,2$$

Repeat step 2 until one of the following conditions occurs:

(1) $p_1^{(v)}$, $p_2^{(v)}$, $\sigma_1^{(v)}$ or $\sigma_2^{(v)}$ drop below a value of 0.01. It is known (see e.g. Fowlkes 1977) that the likelihood surface contains singularities and in some instances the MLE's will converge to parameter values associated with those singularities. When this occurs, either the mixing proportion (p_1 or p_2) or one of the component variances (σ_1^2 or σ_2^2) will tend to zero. In such cases, in this simulation study, the sample was rejected and another sample generated. No attempt was made to repeat the estimation procedure for the rejected sample from another starting point. In the experiments conducted here, fewer than 5 percent of the samples were rejected for this reason.

(2) The parameter values converge. The iterative calculations were stopped if the absolute change in each parameter value in the previous iteration was less than 0.001, and if the absolute change in the log likelihood was less than 0.001. These

conditions are somewhat more stringent than those used by Hosmer (1973a), whose stopping condition was based solely on changes in the log likelihood. It was found in examining the log likelihood surface in one instance that flat local areas (shoulders) existed which were not high points or maxima on the likelihood surface. The more stringent stopping rule ensured that computations did not stop under such conditions.

(3) One thousand iterations have been performed. If convergence did not take place after 1000 iterations, the final parameter set was taken to be the parameter set at 1000 iterations. Again under the worst conditions encountered in these experiments, this situation arose in fewer than 5 percent of the samples. Since every iteration of the algorithm is guaranteed to increase the log-likelihood, it is believed that the final parameter set will be near optimal in most situations of this kind. There is, however, the possibility that the algorithm may have converged on a singularity if the iterations had been continued.

The results of the simulation study are summarized in Table 6.2, which shows the means and standard deviations of the parameters estimated from 200 samples. For comparative purposes, the means and standard deviations of the parameters estimated from the fully classified data are also shown. The population parameters are taken from Table 6.1.

The results in Table 6.2 show that in most instances, both the absolute bias and the variance of the parameter estimates increase as the sample size is decreased and as the separation of the component distributions is decreased. It should also be noted that the parameter estimates for the fully classified data are much more reliable than the parameters estimated from the unclassified mixture. For the sample sizes and parameter sets studied here, the variances of the parameters

Table 6.2 Small Sample Properties of the MLE's for Parameters of Mixtures of Two Normal Distributions

Expt		MLE					Fully Classified Statistics					Population Parameters					Sample
		p_1	μ_1	σ_1	μ_2	σ_2	p_1	μ_1	σ_1	μ_2	σ_2	p_1	μ_1	σ_1	μ_2	σ_2	Size
1	Mean	0.320	0.030	0.976	3.034	0.976	0.302	0.001	0.977	3.009	1.006	0.3	0	1	3	1	100
	St Dev	0.118	0.506	0.552	0.265	0.339	0.045	0.184	0.286	0.118	0.155						
2	Mean	0.342	0.109	1.028	3.056	0.889	0.299	0.028	1.051	2.990	0.981	0.3	0	1	3	1	50
	St Dev	0.0164	0.708	0.741	0.338	0.420	0.063	0.276	0.392	0.152	0.243						
3	Mean	0.365	-0.101	0.742	2.167	0.830	0.308	-0.029	0.951	2.026	1.004	0.3	0	1	2	1	50
	St Dev	0.235	0.774	0.591	0.483	0.442	0.063	0.268	0.389	0.173	0.224						
4	Mean	0.343	0.088	0.985	3.088	1.434	0.305	0.006	0.956	3.007	1.519	0.3	0	1	3	1.5	100
	St Dev	0.158	0.627	0.645	0.401	0.553	0.045	0.176	0.249	0.141	0.276						
5	Mean	0.392	0.207	1.047	3.322	1.646	0.309	-0.002	0.966	3.025	2.064	0.3	0	1	3	2	50
	St Dev	0.200	0.685	0.758	0.724	0.885	0.063	0.279	0.397	0.237	0.446						

estimated from the unclassified mixture are so large that the parameter estimates are probably of limited practical value.

In most water resources applications, it is not the small sample properties of the parameter estimates that are of most concern, but the properties of the quantile estimates. Of particular importance are the quantiles for the tails of the distribution.

Given estimates of the mixture parameters, the quantiles of the mixture can be computed by means of an iterative scheme illustrated in Figure 6.9. The approach for determining the mixture quantile corresponding to a percentile P is as follows (see Figure 6.9):

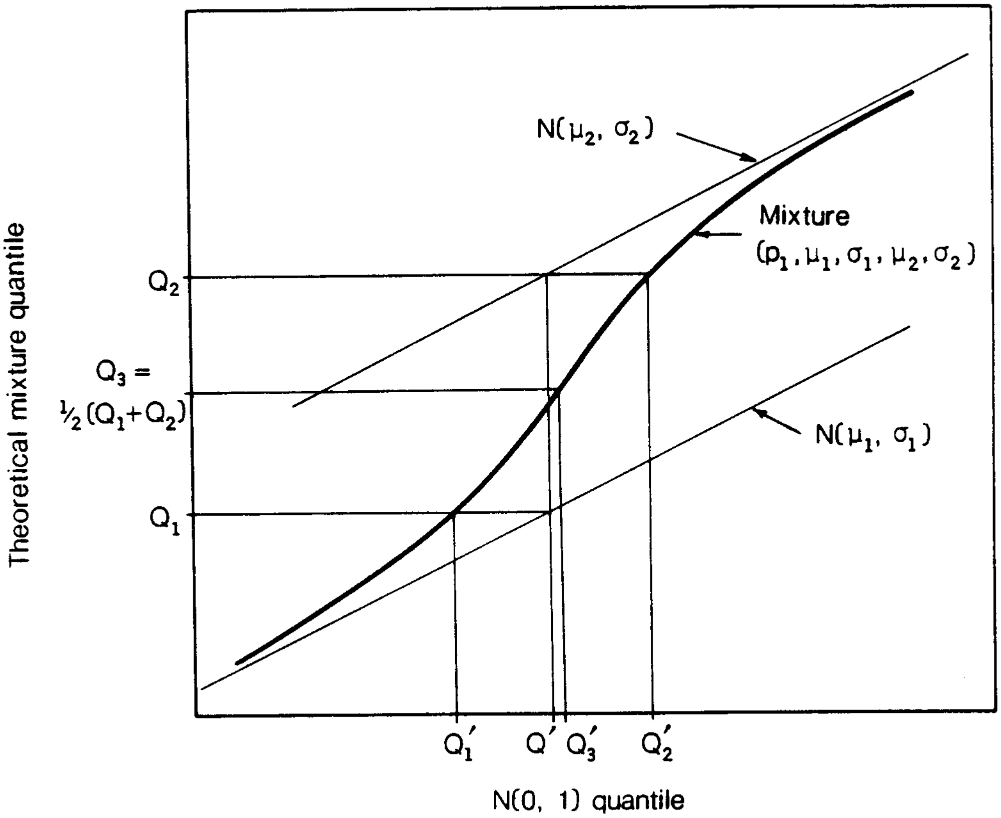


Figure 6.9 Iterative estimation of mixture quantiles

(1) Find the N(0,1) quantile Q' corresponding to P.

(2) Find the quantiles Q_1 and Q_2 of component normals $N(\mu_1,\sigma_1)$ and $N(\mu_2,\sigma_2)$ corresponding to the N(0,1) quantile Q'. The required mixture quantile is now bounded by Q_1 and Q_2.

(3) For the mixture quantiles Q_1 and Q_2 determine the percentiles P_1 and P_2 and hence the corresponding N(0,1) quantiles Q_1' and Q_2'.

(4) For the mixture quantile $Q_3 = 1/2(Q_1 + Q_2)$ find the percentile P_3 and the corresponding N(0,1) quantile Q_3'. Q' is now bounded by either Q_3' and Q_1' or Q_3' and Q_2'.

(5) For the case of Figure 6.9 Q' is bounded by Q_3' and Q_1' and thus the required mixture quantile is bounded by Q_3 and Q_1. Hence determine $Q_4 = 1/2(Q_1 + Q_3)$ and repeat steps (4) and (5) until convergence on Q' has been achieved.

Using this scheme quantiles were calculated for the mixture parameters estimated from both the fully classified and unclassified samples used in experiments 1 and 3 of Table 6.1. Quantiles at various percentile levels were calculated for all 200 samples used in these experiments. The means and variances of the quantile estimates are shown in Table 6.3.

Table 6.3 shows that quantiles estimated using fully classified samples have both a smaller bias and lower variance than those estimated using unclassified samples. However, the differences in the quantile estimates for the two types of sample are considerably smaller than the corresponding differences in the parameter estimates. This indicates that the MLE parameter estimates provide good fits to the sample distributions even though actual parameter estimates are quite variable. This feature may be explained by examining the parameter

Table 6.3 Properties of Quantile Estimates for Samples from a Mixture of Two Normal Distributions

		Quantiles from Parameters Estimated from Unclassified Sample		Quantiles from Parameters Estimated from Fully Classified Sample	
Percentile	Population Quantiles	Mean	St. Dev.	Mean	St. Dev.
Expt 1 (0.3, 0, 1,3,1)*					
0.01	-1.7939	-1.6995	0.3681	-1.7502	0.3354
0.05	-0.9610	-0.9123	0.2696	-0.9333	0.2617
0.10	-0.4311	-0.4060	0.2571	-0.4114	0.2500
0.20	0.4080	0.4237	0.3344	0.4179	0.3189
0.30	1.2935	1.2572	0.3632	1.2633	0.3652
0.40	1.9795	1.9491	0.2834	1.9525	0.2782
0.50	2.4424	2.4338	0.2152	2.4328	0.2007
0.60	2.8219	2.8238	0.1772	2.8195	0.1619
0.70	3.1805	3.1857	0.1556	3.1818	0.1432
0.80	3.5646	3.5684	0.1442	3.5683	0.1375
0.90	4.0588	4.0551	0.1490	4.0642	0.1459
0.95	4.4436	4.4330	0.1715	4.4499	0.1606
0.99	5.1257	5.1054	0.2371	5.1329	0.1977
Expt 3 (0.3, 0, 1, 2, 1)*					
0.01	-1.7947	-1.6070	0.4342	-1.7689	0.4454
0.05	-0.9694	-0.9496	0.3453	-0.9792	0.3357
0.10	-0.4640	-0.5082	0.3294	-0.4939	0.2965
0.20	0.2188	0.1511	0.3033	0.1729	0.2787
0.30	0.7296	0.6861	0.2858	0.6935	0.2615
0.40	1.1474	1.1273	0.2596	1.1307	0.2352
0.50	1.5120	1.5151	0.2337	1.5088	0.2147
0.60	1.8516	1.8733	0.2156	1.8558	0.2022
0.70	2.1928	2.2201	0.2110	2.2002	0.1975
0.80	2.5693	2.5872	0.2179	2.5773	0.2025
0.90	3.0602	3.0456	0.2337	3.0667	0.2238
0.95	3.441	3.3963	0.2594	3.4489	0.2480
0.99	4.1257	4.0185	0.3441	4.1268	0.3046

*See Table 6.1 for full details of experiment.

correlation matrices for the parameters estimated from the fully classified and unclassified samples of experiment 1 (Table 6.4). The correlations for the classified data are essentially zero, whereas the correlations for the unclassified data are relatively large. This indicates a much lower information content in the unclassified sample. It is clear that if a small change is made in the value of one parameter from the unclassified data, compensating changes can be made in the values of other parameters without greatly affecting the adequacy of the fitted distribution.

The results shown in Tables 6.2 and 6.3 indicate that while the ability to fully classify a sample is crucial for the estimation of parameters, it is of lesser significance when considering the overall fit of the distribution.

Before leaving this section a few comments should be made concerning the validity of the Monte Carlo experiments just described. The essential goal of the Monte Carlo method used here is to replicate the conditions and procedures used in estimating mixture parameters from a natural rather than synthetic data set. The principal problem is that, in general, the likelihood surface has multiple maxima (see e.g. Fowlkes 1977) and nèither the EM algorithm nor any other algorithm currently available can guarantee convergence to the global maximum. (As pointed out in Section 6.3, however, the EM algorithm does ensure convergence to a local maxima, which may or may not correspond to the global maximum.) In a practical application the analyst will generally make strenuous efforts to ensure that the final parameter estimates correspond to the global maximum. This would involve a detailed exploration of the likelihood surface and several applications of the estimation procedure from different starting points.

In using the Monte Carlo method to characterize the small sample properties such detailed investigation for each sample is not possible. Instead an implicit assumption is made that the population parameters

Table 6.4 Cross Correlations of Small Sample Parameter Estimates of Experiment 1

Unclassified Samples

	p_1	μ_1	σ_1	μ_2	σ_2
p_1	1.0	0.849	0.743	0.774	-0.684
μ_1	0.849	1.0	0.728	0.805	-0.716
σ_1	0.743	0.728	1.0	0.621	-0.502
μ_2	0.774	0.805	0.621	1.0	-0.766
σ_2	-0.684	-0.716	0.502	-0.766	1.0

Classified Samples

	p_1	μ_1	σ_1	μ_2	σ_2
p_1	1.0	-0.060	-0.004	-0.086	0.044
μ_1	-0.060	1.0	-0.049	0.014	0.033
σ_1	-0.004	-0.049	1.0	-0.058	0.087
μ_2	-0.086	0.014	-0.058	1.0	-0.080
σ_2	0.044	0.033	0.087	-0.080	1.0

represent a good starting point for the estimation procedure and that the final parameter estimates are in some sense good estimates. This however is not necessarily true. It is quite conceivable that a search starting from the population parameters may lead only to a local optimum and may result in quite unrealistic parameter estimates. Such difficulties will inflate the variance of the parameter estimates and the Monte Carlo study may give an overly pessimistic view of small sample variability.

Despite this problem, it is felt that the results of Tables 6.2 and 6.3 give a reasonable indication of the difficulties involved in estimating mixture parameters from small samples.

6.5 Representation of Annual and Monthly Precipitation Data by Means of Mixture Models

In this section, I present examples of the use of the techniques described in Section 6.2 and 6.3 in fitting mixture models to annual and monthly rainfall data. The examples here assume that no exogenous information exists to classify the data by state. However, it is assumed here that there is at least a qualitative justification for using mixture models based on the discussions in Chapters 3 and 4.

The discussions in Chapter 3 suggested that inter-station precipitation relationships are nonlinear. This was confirmed for widely separated stations in the work described in Chapter 4. It appears that cross correlations are higher during drought than during wet or normal periods. One way of modeling this feature is to treat the data as samples from a mixture of two multivariate distributions. This in turn suggests that the data at a site may be modeled by means of a univariate mixture model.

Chapter 3 also introduced the concept of multiple equilibria states by which it is assumed that the atmosphere can exist in one of two quasi-stable states. One state is associated with predominantly meridional circulation over the Pacific Northwest and the other state with predominantly zonal circulation. Meridional circulation, associated with a high pressure ridge over the area, leads to generally dry conditions, while zonal circulation is associated with wet conditions.

There is thus a tentative physical basis for representing rainfall by a mixture of two distributions; one distribution associated with a meridional state and generally low rainfall, the other associated with a zonal state and normal or wet conditions. Unfortunately, the detailed analysis of concurrent pressure and precipitation data in Chapter 5 failed to develop a suitable approach for differentiating objectively between meridional and zonal circulation. Hence it was not possible to show that precipitation comes from a mixed distribution, although the known qualitative association of dry conditions with meridional flow and wet conditions with zonal flow was confirmed.

The work in this section is an attempt to see whether the rainfall data indicate that use of a mixture model is appropriate. For this purpose 81 years of concurrent monthly rainfall data, from October 1891 to September 1979, were obtained for two high quality stations in southern British Columbia; Victoria Gonzales Heights and Agassiz CDA. The location of the two stations is shown in Figure 6.10. They were chosen for this study after consultation with staff of the Pacific Weather Center, Vancouver, B.C. because their records are among the longest available from the area; the records have few missing data; and the data are generally considered to be quite reliable. It should be noted also that the stations are close enough together that they are affected simultaneously by the same type of upper level circulation whether it be meridional or zonal.

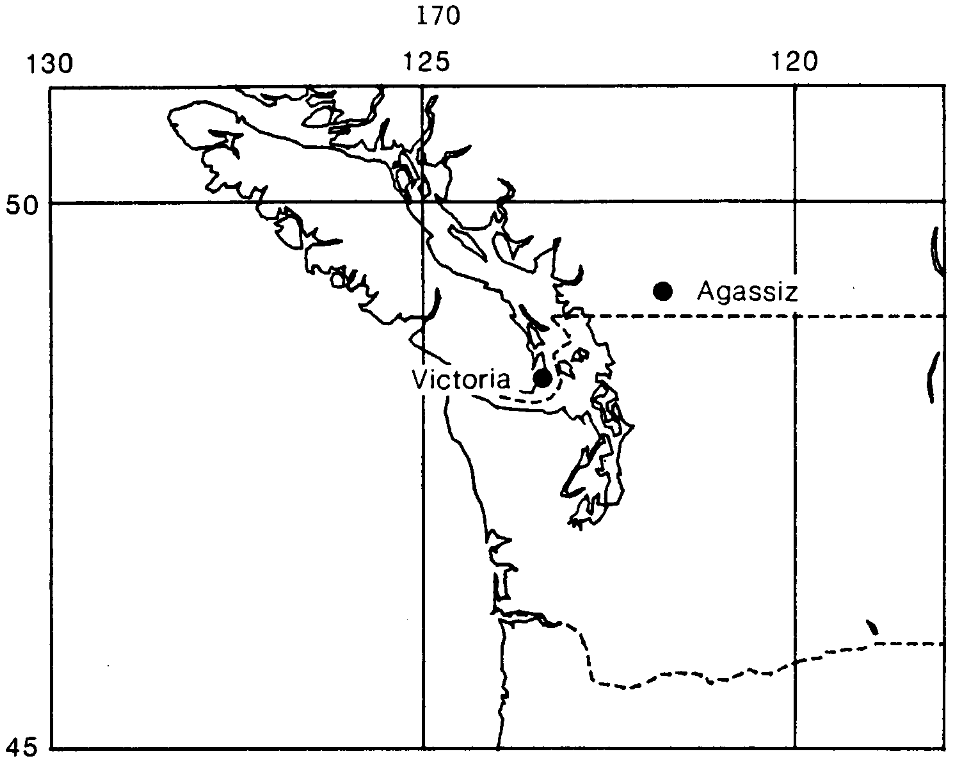


Figure 6.10 Location map

Annual data were obtained by summing monthly data over the water year October 1 to September 30. Some basic statistics for the annual and monthly data are given in Tables 6.5.

6.5.1 Analysis of Annual Precipitation Data, Victoria Gonzales Heights

The Q-Q plot of the quantiles of the standardized Victoria data plotted against standard N(0,1) quantiles is shown in Figure 6.11. The plot shows a slight S-shape suggesting a mixture distribution. Note, however, that if a mixture had not been expected apriori, one might have concluded that the data came from a normal distribution and attributed the departure from the theoretical straight line to sampling variability. (In this example the physical basis for assuming a mixture has not been well established. For the purposes of

Table 6.5 Basic Statistics for Monthly and Annual Precipitation Data (in mm), Victoria and Agassiz

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Annual
Victoria													
Mean	68.7	104.0	118.8	107.0	72.3	53.3	30.6	22.8	20.7	12.4	18.6	34.8	667.9
St. dev.	35.9	51.2	54.4	53.0	39.3	26.9	17.0	13.7	16.5	10.8	14.7	23.8	138.4
Skew	0.83	0.68	1.05	1.54	0.59	0.72	0.94	1.38	2.25	1.02	1.16	0.76	-0.15
Maximum	200.6	297.7	330.9	337.4	180.0	136.6	77.6	78.3	110.0	49.1	65.7	102.4	957.5
Minimum	13.3	16.2	15.0	20.0	8.1	9.0	2.4	2.1	0.6	0.0	0.0	2.6	361.0
Lag one correlation													0.09
Agassiz													
Mean	176.6	207.5	224.5	206.1	159.2	145.2	109.6	95.6	85.6	47.5	59.6	102.9	1619.7
St. dev.	82.4	90.9	90.1	103.3	80.0	63.6	48.2	45.1	49.1	29.2	44.5	63.6	323.6
Skew	0.40	0.66	0.13	0.65	1.00	0.59	1.19	0.56	1.04	0.47	1.04	0.89	0.26
Maximum	377.1	537.0	434.0	514.9	406.2	350.0	290.7	215.0	273.4	129.3	201.7	315.3	2368.5
Minimum	31.4	47.6	13.4	42.4	20.9	46.6	19.6	10.0	8.9	0.0	0.0	6.6	901.6
Lag one correlation													0.15

Note: Statistics based on 81 years of data from October 1, 1898 to September 30, 1979 illustration, however, I will assume that use of a mixture distribution is justified.)

The P-Q plot for the annual Victoria data is shown in Figure 6.12. The plot oscillates about the zero line on the ordinate, but the configuration of the plot does not conform closely to that of the theoretical P-Q plot for mixtures shown in Figure 6.6. Although in this case the sample P-Q plot alone would not necessarily indicate a mixture, the departure of the sample plot from the theoretical mixture configuration could again be ascribed to sampling variability.

Graphical estimation of mixture parameters from the Q-Q plot of the standarized Victoria data is illustrated in Figure 6.13. Straight lines asymptotic to the upper and lower limbs of the S-shape were fitted by eye, and the point of inflexion X_L was estimated by eye.

Figure 6.13 shows the point of inflexion at the N(0,1) quantile $X_L = -0.77$. An estimate of the mixing proportion p_1 is given by the corresponding N(0,1) percentile which is found from tables of the standardized normal distribution to be 0.22.

A first estimate of mixture parameters for the standardized Victoria data is thus

$$\phi = (p_1, \mu_1, \sigma_1, \mu_2, \sigma_2) = (0.22, -0.60, 0.70, 0.14, 0.90)$$

Given a variable y having a mixture distribution with parameter set

$$\phi = (p_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$$

then the transformed (standardized) variable

$$Z = (\frac{y-\mu}{\sigma})$$

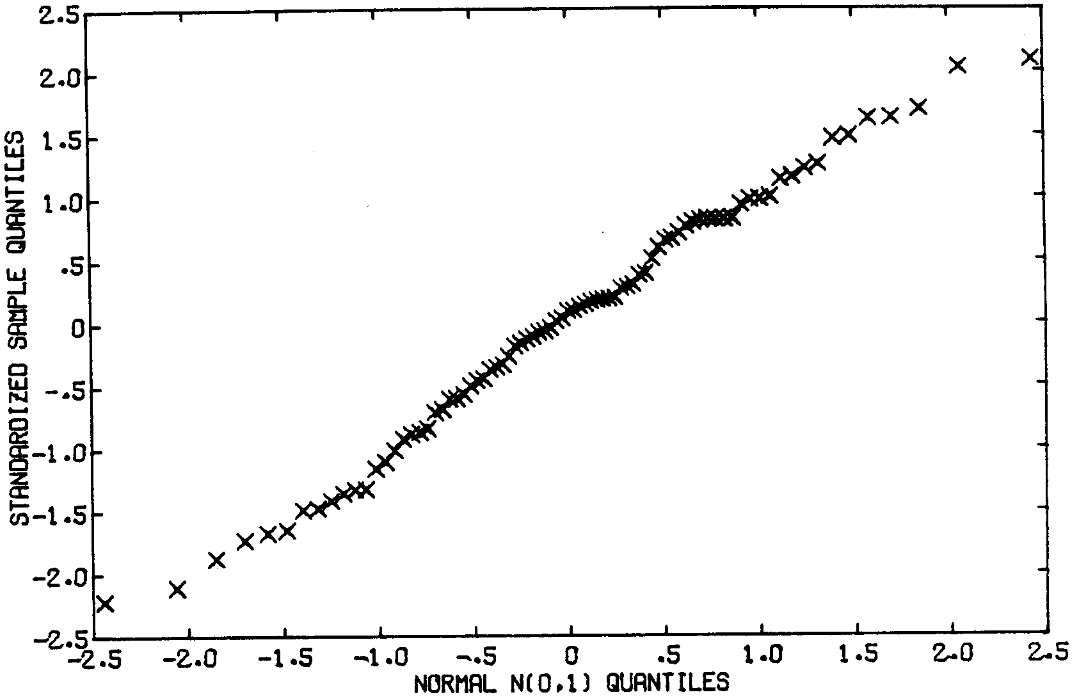


Figure 6.11 Q-Q plot of standardized annual precipitation at Victoria

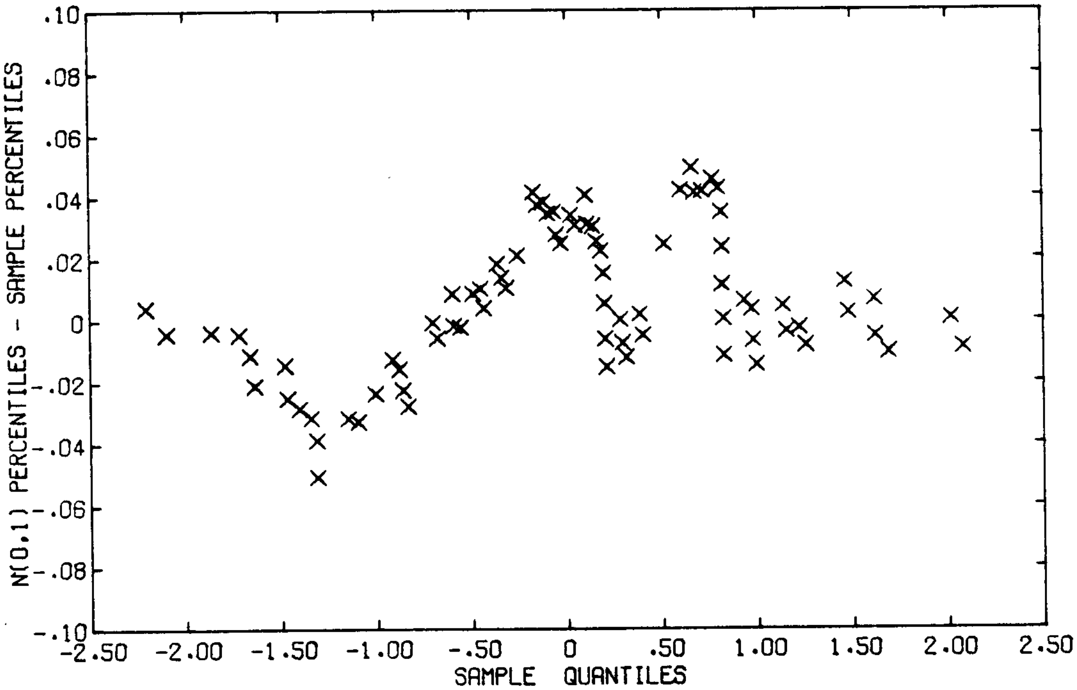


Figure 6.12 P-Q plot of annual precipitation at Victoria

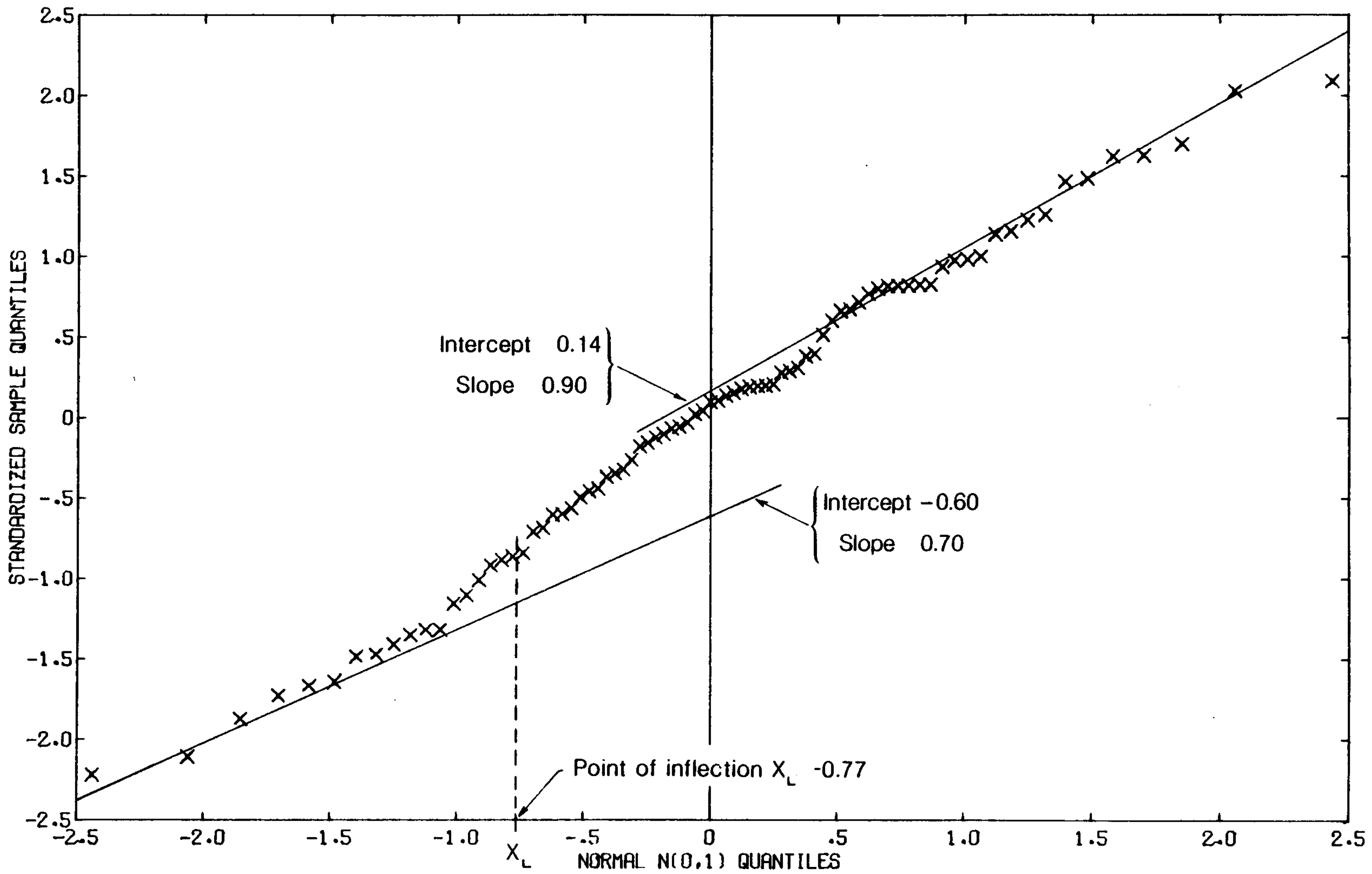


Figure 6.13 Graphical estimation of mixture parameters for annual precipitation at Victoria

also has a mixture distribution with parameters

$$\left(p_1, \frac{\mu_1-\mu}{\sigma}, \frac{\sigma_1}{\sigma}, \frac{\mu_2-\mu}{\sigma}, \frac{\sigma_2}{\sigma} \right)$$

The overall mean μ and standard deviation σ used to standardize the Victoria data were 667.9 mm and 138.4 mm respectively. Using the inverse transform $Y = Z\sigma + \mu$ gives the mixture parameters for the original Victoria data as (0.22, 585, 97, 687, 125).

These initial parameter extimates were used as a starting point for finding the MLE's of the parameters using the EM algorithm described in Section 6.3.3. Convergence was achieved in 177 iterations using the criteria described in Section 6.4 with a resolution of 0.01. The final parameter set was (0.139, 456, 53.2, 702, 115.5). To study the effect of the initial choice of parameters on the MLE's, parameters were estimated using the EM algorithm with a number of different initial or starting points. The starting points and resulting MLE's are shown in Table 6.6. The final MLE's were the same from each starting point giving us some confidence that the MLE corresponds to a global maximum rather than a local maxima.

Table 6.6 Effect of Starting Point on the MLE's for Annual Data from Victoria

Initial Parameter Set	Maximum Likelihood Parameters
(0.22, 585, 97, 687, 125)	(0.139, 456, 53.2, 702, 115.5)
(0.50, 569, 98.4, 766, 98.4)	(0.139, 456, 53.2, 702, 115.5)
(0.80, 670, 63.3, 670, 126.5)	(0.140, 456, 53.3, 702, 115.4)

The simple transformation

$$Z = \frac{Y - 456}{53.2}$$

converts the fitted distribution just obtained to the "standard" form of the mixture distribution with parameters (0.139, 0, 1, 4.62, 2.17). The theoretical Q-Q and P-Q plots for this distribution were shown in Figures 6.3 and 6.6 respectively and the Q-Q and P-Q plots for a sample of 81 observations drawn from this theoretical distribution were shown in Figures 6.5 and 6.8. Comparison of Figures 6.11 and 6.12 with Figures 6.3, 6.5, 6.6, and 6.8 again demonstrates the problem of sampling variability. The sample Q-Q plot for the Victoria data resembles the theoretical Q-Q plot for the fitted distribution, but the sample P-Q plot only bears a superficial resemblance to its theoretical counterpart.

The adequacy with which the data fits the mixture distribution is shown in Figure 6.14. This is a Q-Q plot of the quantiles of the standardized data plotted against the quantiles of the fitted mixture distribution. As would be hoped for a fit to a five parameter distribution, the sample Q-Q is essentially straight with only minor deviations from the straight line.

The mixture quantiles used in Figure 6.14 were computed by means of the iterative scheme illustrated in Figure 6.9 and described in Section 6.4.

A final set of figures worth studying are the weights used in deriving the MLE's of the mixture parameters via the EM algorithm. The weight w_{ij}, it will be recalled, is the posterior probability that, given y_i, observation i is from state j. The set of weights w_{i1}, i=1,...n corresponding to the MLE's is shown, together with the associated observations y_i, in Table 6.7. To help understand the meaning of the weights, Figure 6.15 shows the density functions for

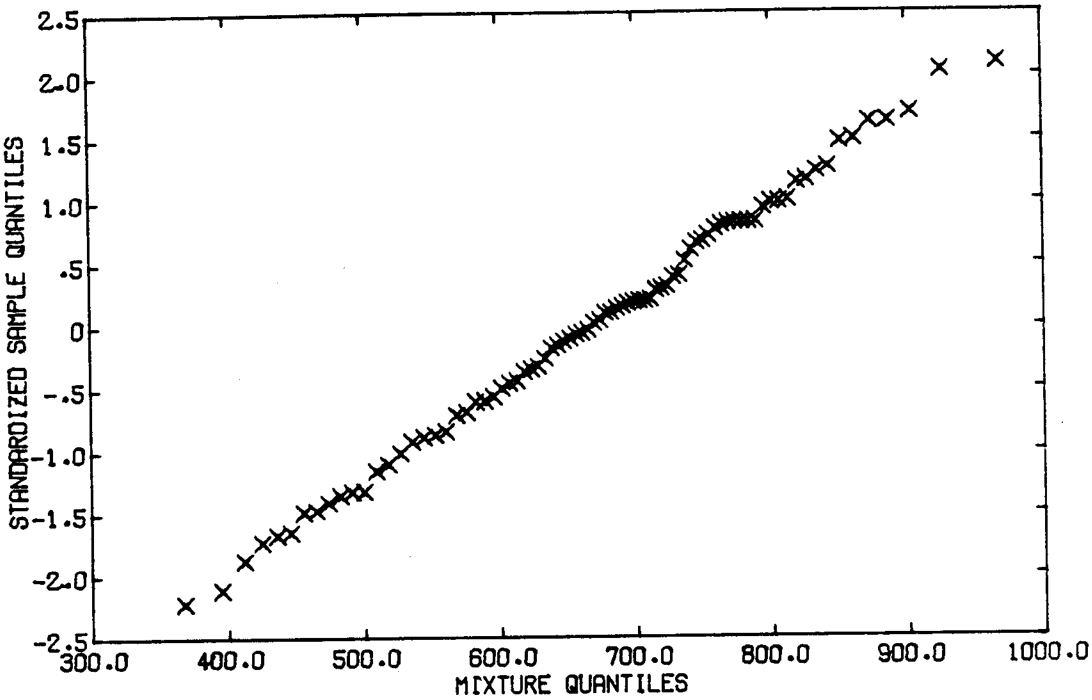


Figure 6.14 Sample Q-Q plot of standardized Victoria data plotted against quantiles of the mixture with parameters (0.139, 456, 53.2, 702, 115.4)

Table 6.7 Weights Associated with the MLE's of the Mixture Parameters for Annual Precipitation at Victoria

Year	Observation	Weight*	Year	Observation	Weight*
	y_i	w_{i1}		y_i	w_{i1}
1899	806.0	0.0	1940	711.7	0.0
1900	781.0	0.0	1941	515.6	0.409
1901	620.3	0.004	1942	464.0	0.744
1902	623.6	0.003	1943	585.6	0.029
1903	751.7	0.0	1944	472.8	0.706
1904	660.4	0.0	1945	545.8	0.174
1905	693.1	0.0	1946	697.3	0.0
1906	462.5	0.750	1947	643.6	0.001
1907	708.9	0.0	1948	842.0	0.0
1908	696.1	0.0	1949	663.7	0.0
1909	573.5	0.054	1950	892.8	0.0
1910	948.8	0.0	1951	759.9	0.0
1911	828.7	0.0	1952	528.4	0.301
1912	694.9	0.0	1953	631.9	0.002
1913	695.8	0.0	1954	779.1	0.0
1914	682.1	0.0	1955	658.7	0.0
1915	437.0	0.822	1956	781.3	0.0
1916	871.4	0.0	1957	804.4	0.0
1917	590.4	0.023	1958	485.2	0.638
1918	782.6	0.0	1959	893.5	0.0
1919	798.0	0.0	1960	689.8	0.0
1920	767.4	0.0	1961	723.5	0.0
1921	774.8	0.0	1962	570.3	0.063
1922	687.2	0.0	1963	599.7	0.013
1923	782.8	0.0	1964	803.2	0.0
1924	584.8	0.030	1965	541.1	0.205
1925	674.4	0.0	1966	605.1	0.010
1926	481.0	0.663	1967	739.6	0.0
1927	552.1	0.138	1968	826.0	0.0
1928	651.2	0.0	1969	617.4	0.005
1929	440.4	0.815	1970	428.9	0.835
1930	508.4	0.469	1971	548.6	0.157
1931	683.1	0.0	1972	836.0	0.0
1932	761.8	0.0	1973	408.6	0.857
1933	781.8	0.0	1974	707.2	0.0
1934	957.5	0.0	1975	607.4	0.009
1935	873.9	0.0	1976	903.2	0.0
1936	646.5	0.001	1977	376.2	0.860
1937	671.2	0.0	1978	485.5	0.636
1938	721.4	0.0	1979	361.0	0.849
1939	654.2	0.0			

*Weight w_{ij} is the posterior probability that given y_i, observation i comes from state j

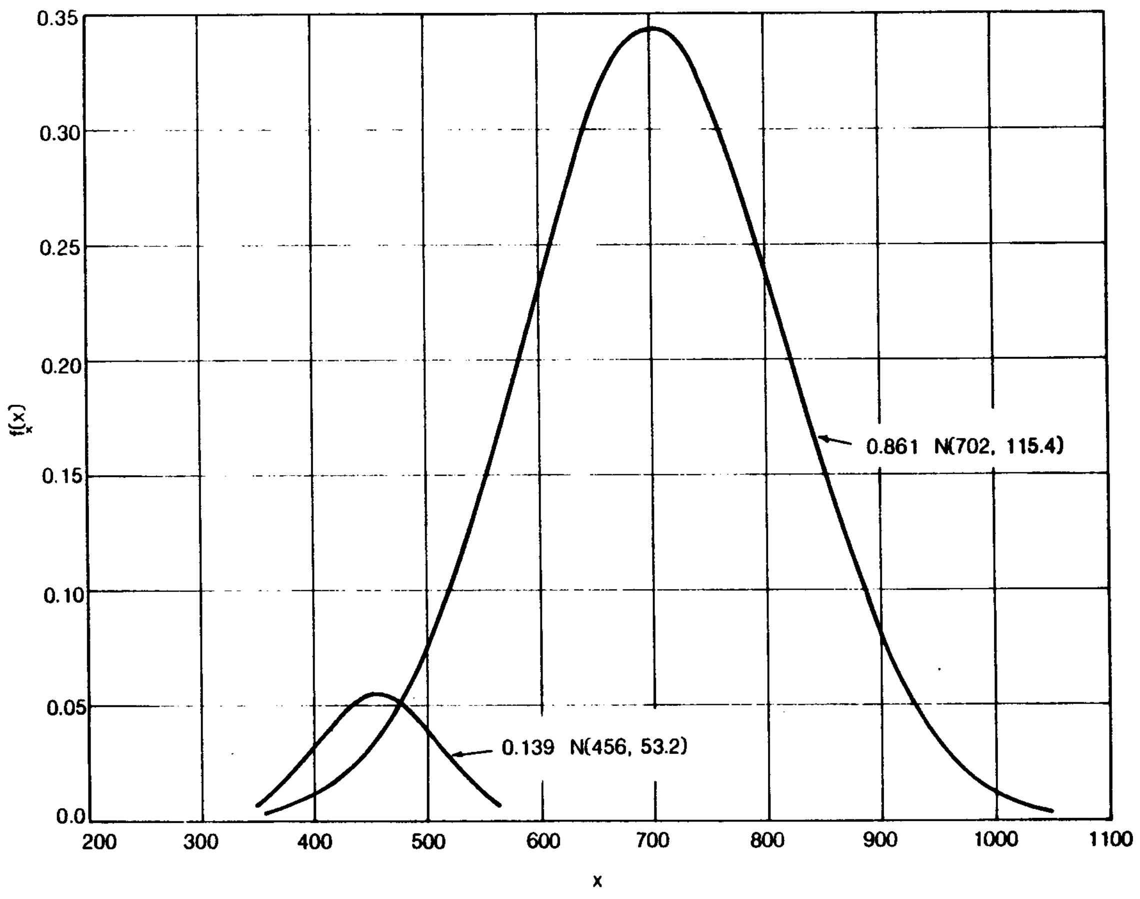


Figure 6.15 Component normal distributions scaled by mixing proportion for the mixture with parameters (0.139, 456, 53.2, 702, 115.4)

the component distributions making up the mixture scaled by the appropriate mixing proportions.

Of particular interest in Table 6.7 is the fact that the maximum value for a weight w_{i1} is 0.860, for data from 1977. Thus even in a year with a persistent meridional circulation (see Monthly Weather Review, Volume 105), a purely probabilistic approach to parameter estimation suggests that data for that year could have been associated with the "wet" distribution, and perhaps zonal atmospheric flow, with probability 0.14.

A comparison of weights for years 1977 and 1979 is also instructive. Note that although 1979 is drier than 1977, the value of weight w_{i1} is lower. This implies that there is a greater probability that the drier year was associated with the "wet" distribution and zonal flow. Although 1979 was the driest year on record, examination of Figure 6.15 shows that for the current mixture distribution more extreme dry years would have even larger probabilities of association with a zonal state. In fact a hypothetical year with rainfall of only 300 mm would have approximately a 50 percent probability of being associated with either state. This is completely counter to the qualitative physical arguments used to justify a mixture distribution and implies either that the model is inappropriate or that the parameters estimated in the absence of exogenous information are unreliable.

6.5.2 Analysis of Annual Precipitation Data, Agassiz

An analysis similar to that described in the previous section for the Victoria data was performed for the annual precipitation data from Agassiz. Only the essential details and results of the analysis are described here.

Figure 6.16 shows the Q-Q plot for the standardized Agassiz data and Figure 6.17 shows the corresponding P-Q plot. The Q-Q plot shows definite departures from normality and, if one drops the two smallest values, then the sample plot closely resembles some of the theoretical Q-Q plots for mixtures shown by Fowlkes (1977). The sample P-Q plot, however, does not resemble the general configuration of the theoretical P-Q plots for mixtures. As with the Victoria data, a mixture distribution would probably not be used here unless a mixture had been expected a priori.

A mixture distribution was fitted to the Agassiz data using an approach indentical to that used for the Victoria data. The MLE's for the mixture parameters were (0.57, 1405, 197, 1895, 233) which corresponds to the "standard" parameter set (0.57, 0,1, 2.49, 1.18). Q-Q and P-Q plots for the theoretical mixture were shown in Figures 6.4 and 6.7, and a sample Q-Q plot of the quantiles of the Agassiz data plotted against the theoretical mixture quantiles is shown in Figure 6.18. Figure 6.18 shows close agreement between the data and the fitted distribution. However, as pointed out earlier, many other five parameter distributions might have fitted the data just as well.

Finally Table 6.8 shows the weights associated with the MLE's of the mixture parameters and Figure 6.19 shows the component distributions of the mixture scaled by the mixing proportions. Note that in this case the mixture is more widely separated than for the Victoria data making classification by state somewhat easier. The values of the weight are also consistent with the highest weights associated with the lowest rainfall values.

It had been assumed that rainfall at Victoria and Agassiz are both affected by the same large scale atmospheric features such that if rainfall at Victoria in a particular year was associated with zonal circulation then the rainfall at Agassiz in that year would also be

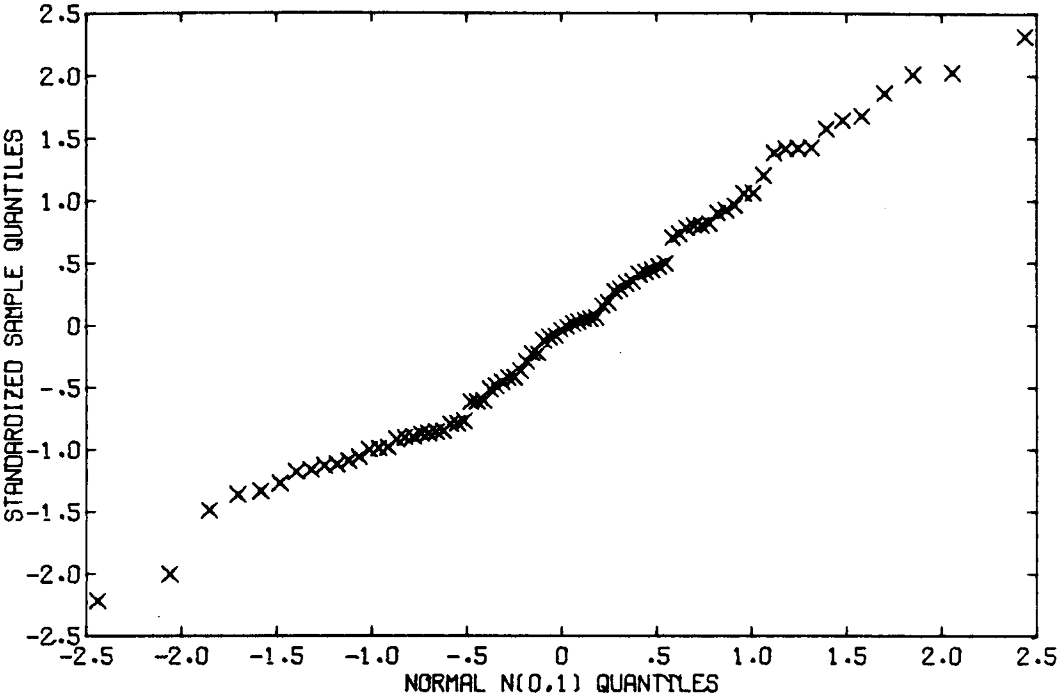


Figure 6.16 Q-Q plot of standardized annual precipitation at Agassiz

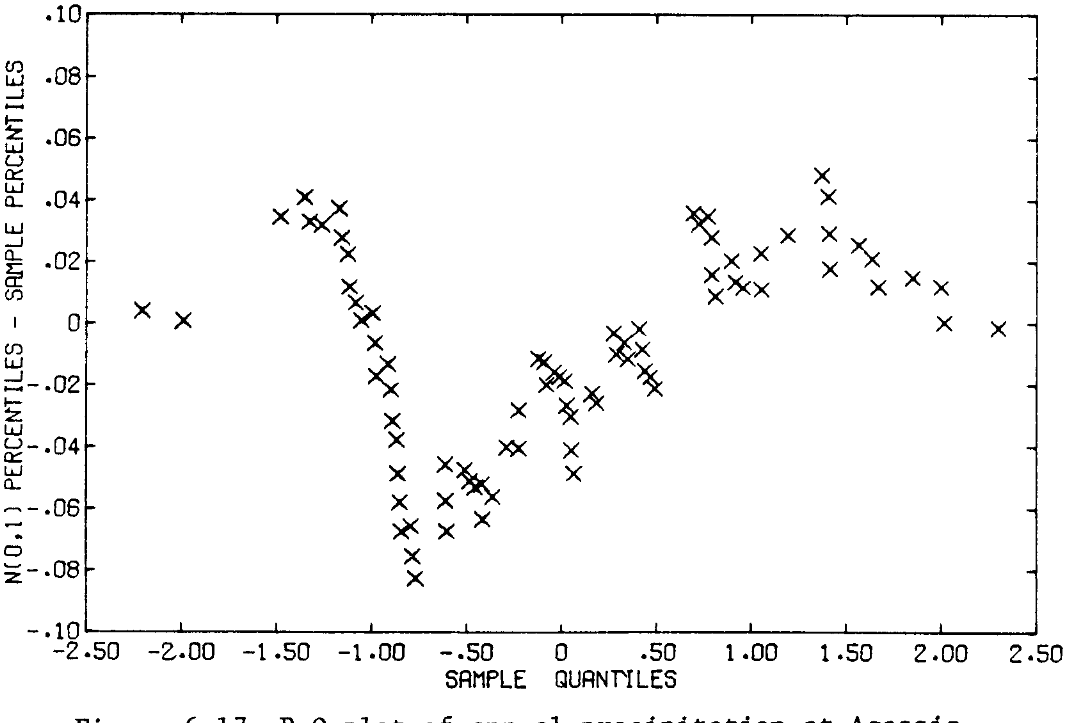


Figure 6.17 P-Q plot of annual precipitation at Agassiz

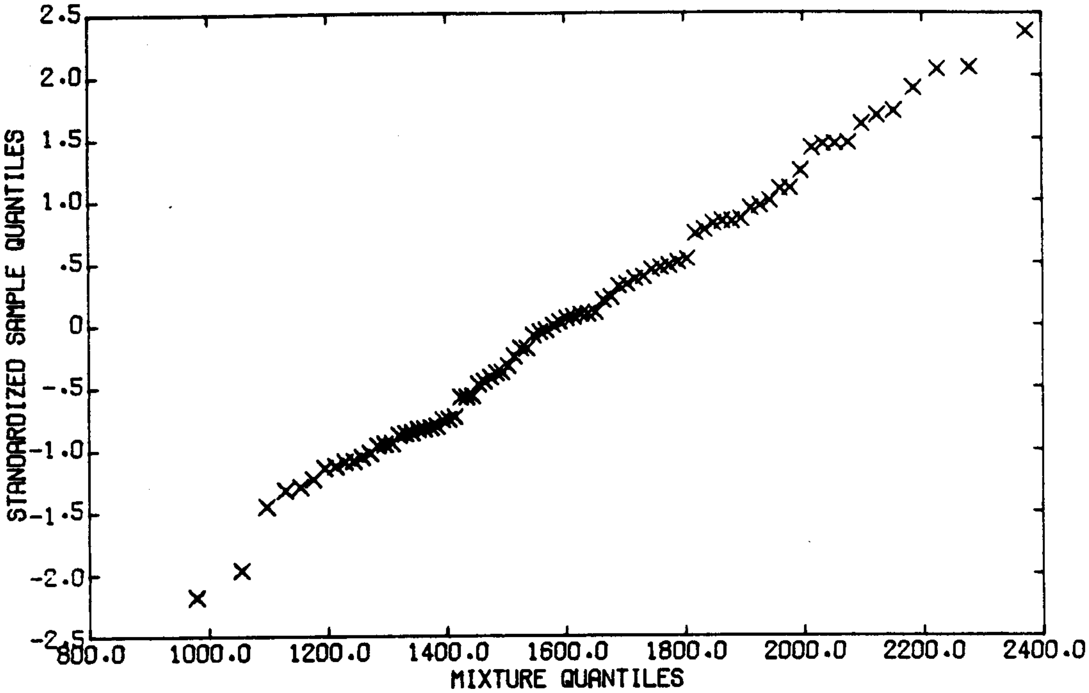


Figure 6.18 Sample Q-Q plot of standardized Agassiz data plotted against quantiles of the mixture with parameters (0.57, 1405, 197, 1895, 233)

Table 6.8 Weights Associated with the MLE's of the Mixture Parameters for Annual Precipitation at Agassiz

Year	Observation	Weight*	Year	Observation	Weight*
	y_i	w_{i1}		y_i	w_{i1}
1899	1461.9	0.891	1940	1640.5	0.574
1900	2066.5	0.007	1941	1420.9	0.923
1901	1299.7	0.972	1942	1256.4	0.980
1902	1364.1	0.952	1943	1593.5	0.689
1903	1580.3	0.718	1944	1253.9	0.980
1904	1296.2	0.972	1945	1423.3	0.921
1905	1635.3	0.588	1946	1733.2	0.325
1906	1361.4	0.953	1947	1483.1	0.870
1907	1771.7	0.235	1948	1877.7	0.078
1908	1209.0	0.986	1949	1545.9	0.783
1909	1243.1	0.982	1950	1846.4	0.111
1910	1636.5	0.585	1951	1871.4	0.084
1911	1327.0	0.965	1952	1337.1	0.962
1912	1752.4	0.278	1953	1615.8	0.637
1913	2276.2	0.0	1954	1855.6	0.101
1914	1628.2	0.606	1955	1679.5	0.468
1915	1369.8	0.95	1956	2078.7	0.006
1916	1962.1	0.028	1957	1545.9	0.783
1017	1484.0	0.869	1958	1266.9	0.978
1918	1930.9	0.041	1959	2129.6	0.003
1919	2077.8	0.006	1960	1713.4	0.376
1920	2386.5	0.0	1961	1779.4	0.219
1921	1917.7	0.049	1962	1501.2	0.849
1922	1757.6	0.266	1963	1342.3	0.960
1923	1525.5	0.816	1964	2162.9	0.002
1924	1671.9	0.489	1965	1420.4	0.923
1925	1911.2	0.053	1966	1624.8	0.615
1926	1238.6	0.982	1967	1877.4	0.078
1927	1345.4	0.959	1968	2152.6	0.002
1928	1138.5	0.992	1969	1727.2	0.340
1929	971.8	0.997	1970	1322.3	0.966
1930	901.6	0.998	1971	1708.8	0.388
1931	1179.3	0.989	1972	2270.8	0.0
1932	1883.4	0.073	1973	1188.0	0.988
1933	2080.3	0.006	1974	2008.3	0.015
1934	1962.9	0.028	1975	1453.9	0.898
1935	1762.4	0.255	1976	2222.2	0.001
1936	1589.5	0.698	1977	1329.6	0.964
1937	1338.7	0.961	1978	1471.2	0.882
1938	1276.3	0.976	1979	1301.9	0.971
1939	1606.9	0.658			

*Weight w_{ij} is the posterior probability that given y_i observation i comes from state j

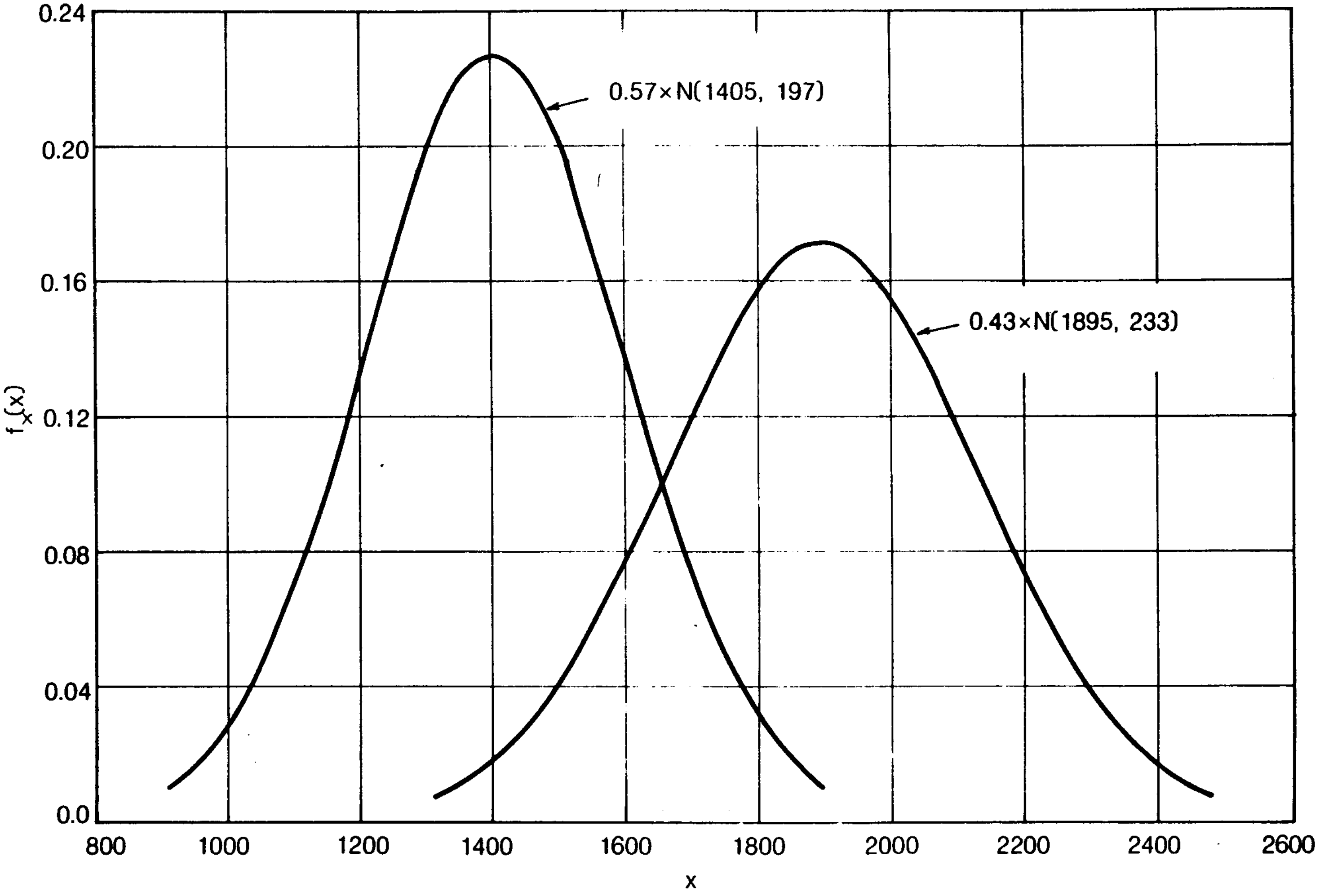


Figure 6.19 Component normal distributions scaled by mixing proportion for the mixture with parameters (0.57, 1405, 197, 1895, 233)

associated with zonal circulation. Comparison of Figure 6.15 and Figure 6.19 shows that this assumption is frequently violated, again indicating either that the model assumed is inappropriate or that the estimated parameters are inappropriate.

6.5.3 Analysis of Monthly Precipitation Data, Victoria Gonzales Heights

For the periods of drought I am primarily interested in, a pattern of meridional circulation will perhaps persist for a number of weeks or months rather than years. There is thus a danger in performing analysis on an annual time scale of aggregating rainfall associated with both meridional and zonal states. In principle then it appears that it should be somewhat easier to detect mixtures in monthly rather than annual rainfall data.

Basic statistics for monthly data from Victoria Gonzales Heights were shown in Table 6.5. Sample Q-Q and P-Q plots for the January data are shown in Figures 6.20 and 6.21 respectively. In this case, unlike the situation for annual data, neither the Q-Q nor the P-Q plot in any way suggests the presence of a mixture of two normal distributions. Rather, the Q-Q plot suggests perhaps a logarithmic distribution. To check for the presence of a mixture of two two-parameter lognormal distributions, Figures 6.22 and 6.23 show respectively the Q-Q and P-Q plot for the natural logs of the January data. Again neither plot suggests, in any way, the presence of a mixture.

Q-Q plots for the remaining months of the year have been plotted from the Victoria data but are not shown here. Only data for October and November exhibited the S-shaped Q-Q plot characteristic of a mixture of two normal distributions. The plots for the remaining months were qualitatively similar to those presented for the January data, i.e. convex, and in no way suggestive of the presence of a

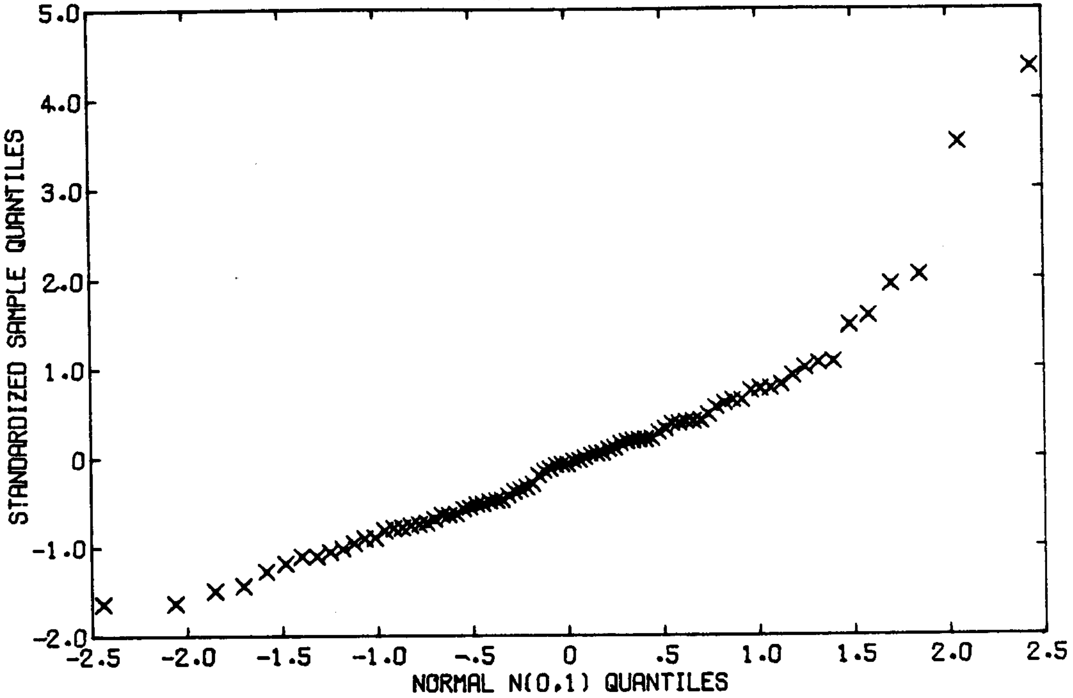


Figure 6.20 Q-Q plot of standardized January precipitation at Victoria

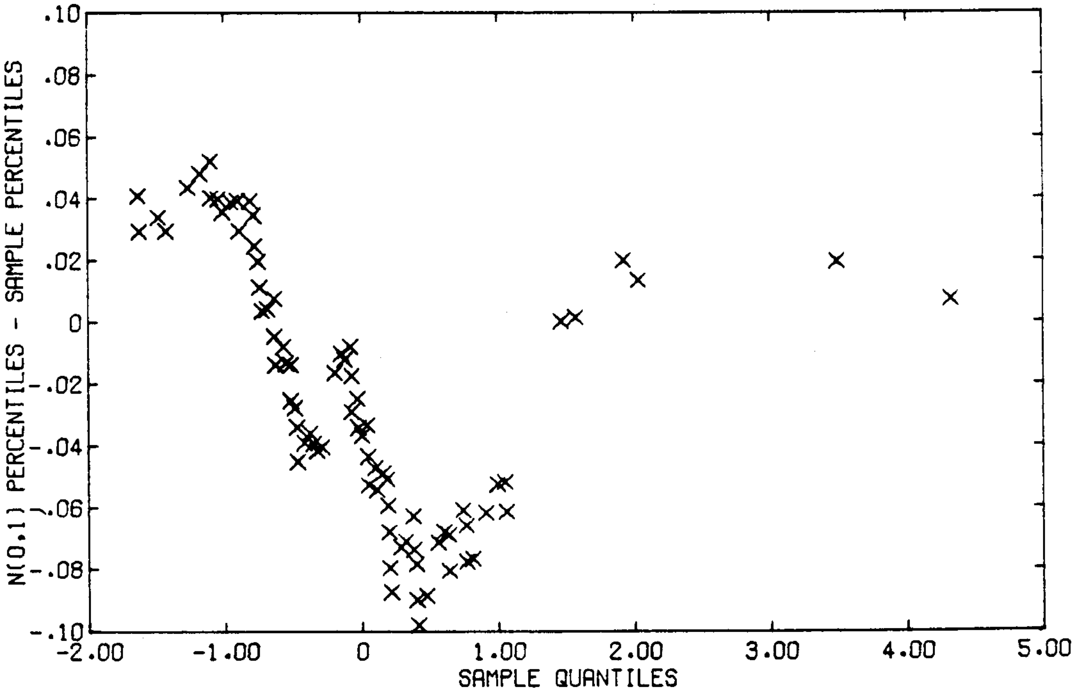


Figure 6.21 P-Q plot of January precipitation at Victoria

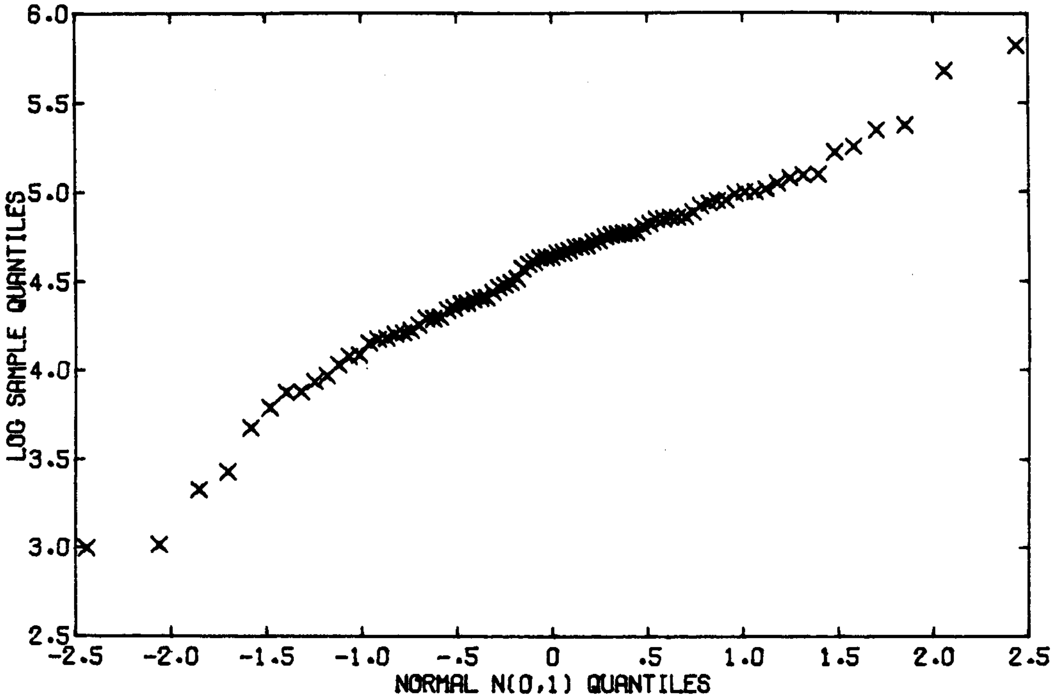


Figure 6.22 Q-Q plot of natural logarithms of January precipitation at Victoria

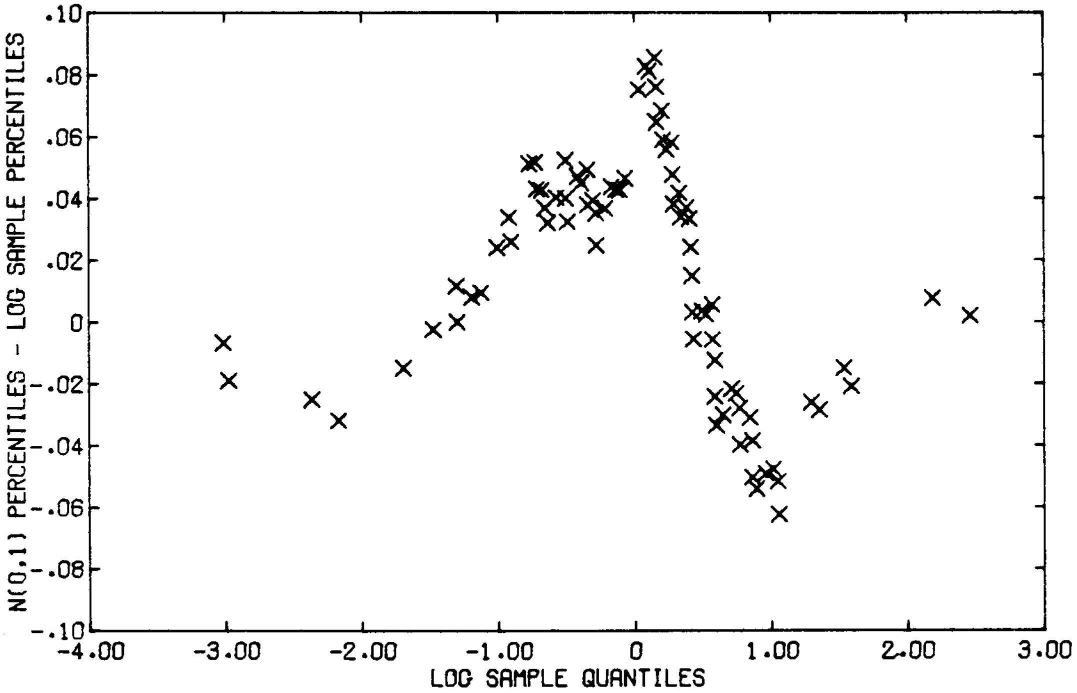


Figure 6.23 P-Q plot of natural logarithms of January precipitation at Victoria

mixture. Consequently, no attempts were made to fit mixture models to the monthly data.

6.6 Concluding Remarks

In the work described in this chapter I have investigated in some detail the properties of univariate normal mixture models, the estimation of mixture parameters, and the application of mixture models for representing annual and monthly precipitation data.

The precipitation records from Victoria and Agassiz are among the longest available in the area and are considerably longer than most records used in the water resources field. Even with these records, however, it was not possible to confirm the presence of mixtures. The Q-Q plots for the annual data showed slight S-shapes, suggesting that a mixture might be appropriate, but the departure from normality was small and could easily have been the result of sampling variability. The Q-Q plots for monthly data did not indicate the presence of mixtures, and the P-Q plots for the data failed to exhibit the configurations typical of the theoretical P-Q plots of mixtures.

In principal, as pointed out earlier, it should have been easier to detect mixtures from the monthly data than from the annual data. The failure to detect mixtures at the monthly interval reinforces the supposition that the slight S-shapes of the annual Q-Q plots were the result of sampling variability.

The inevitable conclusion from this work is that the use of univariate mixture models cannot be justified for modeling single-site precipitation data. The problems of modeling nonlinear inter-station precipitation relationships, however, remains. These problems are discussed in the next chapter.

7.0 MULTIVARIATE MIXTURE MODELS FOR PRECIPITATION SYNTHESIS

Models of multi-site hydrologic sequences are commonly based on the assumption that, after suitable transformations, the multi-site data can be regarded as samples from a multivariate normal distribution. This assumption is, however, generally an act of faith made, not because the data have been shown to come from a multivariate normal distribution, but because of the attractive computational features of the distribution. Of particular importance from the operational point of view is the fact that the marginal distributions are normal.

The usual approach to modeling has been first to fit a simple distribution, such as a normal or lognormal, to the data from each site. The choice of distribution is dictated by the requirement that application of a simple transformation should permit the transformed single-site data to be treated as samples from normal distributions. These normal distributions are of course the marginal distributions of the transformed multi-site data. Multi-site data in the transformed domain can hence be obtained by sampling from a multivariate normal distribution having the fitted marginal distributions and with an appropriate covariance matrix.

The primary concern in most modeling efforts to date has been to ensure that the theoretical marginal distributions fit the data adequately. The validity of assuming joint normality has received little or no attention. The work in Chapter 4 has shown, however, that under certain circumstances inter-station precipitation relationships cannot be modeled adequately using a multivariate normal distribution. The question now arises as to how multi-site sequences should be modeled to ensure that both the marginal distribution and the cross properties of the data are adequately represented.

It was seen in Chapter 4 that the difficulties in modeling the inter-station relationships could be ascribed to the occurence of higher cross correlations during drought than during wet or normal periods. Although mixture distributions were not found to be useful for single-site applications in the work described in Chapter 6, they remain attractive for multi-site work since they are perhaps the simplest form of model which incorporates a nonlinear structure in the inter-station relationships.

The work in this chapter uses mixtures of two multivariate normal distributions. This is a somewhat arbitrary choice of distribution from the point of view of representing the observed data. However, the distributions are not only computationally attractive, but the flexibility allows one to fit both the marginal distributions and the important features of the inter-station relationships.

7.1 Modeling Approach

Assume that the precipitation at n sites may be modeled by means of a mixture of two n-variate normal distributions with different means and different covariance matrices. Denote an observation from the mixture distribution by

$$\underline{y} = (y_1, y_2, \ldots y_n)$$
$$= (nx1) \text{ matrix of concurrent observations at n sites}$$

In a sequence of m independent and identically distributed observations suppose that each $\underline{y}$ may be associated with one of two possible states. Let the sequence of states be represented by the outcome of m identical and independent Bernouilli trials such that:

$$P(\underline{y} \text{ is associated with state 1}) = p_1$$
$$P(\underline{y} \text{ is associated with state 2}) = 1-p_1 = p_2$$

Further assume that the $\underline{y}$ given the associated state j are conditionally independent with n-variate normal densities.

$$g(\underline{y}|\text{state } j) = g_j(\underline{y})$$

$$= \frac{1}{(2\pi)^{n/2}(\det \underline{M}_j)^{\frac{1}{2}}} \exp\left\{-\tfrac{1}{2}(\underline{y}-\underline{\mu}_j)\underline{M}_j^{-1}(\underline{y}-\underline{\mu}_j)^T\right\}$$

$$;j=1,2 \qquad (7.1)$$

where $\underline{M}_j$ = (nxn) covariance matrix of the $\underline{y}$ given state j

$\underline{\mu}_j$ = (nx1) matrix containing the means of the $\underline{y}$ given state j = $(\mu_{ij}, \mu_{2j}, \mu_{3j},\ldots,\mu_{nj})$

Then the unconditional density of the $\underline{y}$ is the multivariate mixture distribution:

$$g(\underline{y}) = p_1g_1(\underline{y}) + (1-p_1)g_2(\underline{y}) \qquad (7.2)$$

and the unconditional marginal density of the y_i, i=1,..,n is in general the mixture of two univariate normal distributions:

$$f(y_i) = p_1f_1(y_i) + (1-p_1)f_2(y_i) \qquad (7.3)$$

where

$$f_j(y_i) = \frac{1}{(2\pi)^{\frac{1}{2}}\,\sigma_{ij}} \exp\left\{-\frac{(y_i - \mu_{ij})}{2\sigma_{ij}^{\ 2}}\right\} \cdot \qquad ;i=1,n \quad j=1,2 \qquad (7.4)$$

σ^2_{ij} = conditional variance of y_i given state j

μ_{ij} = conditional mean of y_i given state j

Sampling from the independent multivariate mixture distribution was undertaken using an extension of the scheme described in Section 4.2.1 as follows:

A uniform pseudo-random number U on the interval [0,1) was generated using the CDC RANF random number generator.

For $U \leq p_1$ the matrix of observations $\underline{y}$ given state 1 was generated as:

$$\underline{y} = \underline{B}_1\underline{\varepsilon} + \underline{\mu}_1 \tag{7.5}$$

where $\underline{\varepsilon}$ = (nx1) matrix whose elements are independent identically distributed samples from the normal N(0,1) distribution.

$\underline{B}_1$ = (nxn) matrix such that

$$\underline{B}_1\underline{B}_1^T = \underline{M}_1$$

$\underline{M}_1$ = lag-zero conditional covariance matrix of $\underline{y}$ given state 1

$\underline{\mu}_1$ = (nx1) matrix of conditional means of $\underline{y}$ given state 1

Similarly for $U > p_1$, $\underline{y}$ was generated as

$$\underline{y} = \underline{B}_2\underline{\varepsilon} + \underline{\mu}_2 \tag{7.6}$$

where the subscript 2 now indicates that the parameters of the scheme are conditioned on state 2.

The number of parameters in this multivariate mixture model is $(n^2 + 3n + 1)$ whereas the comparable multivariate normal (Equation 4.2) has $(n^2 + 3n)/2$ parameters.

Parameter estimation for univariate mixture models was discussed in Chapter 6 for unclassified data. Parameter estimation for multivariate mixture models clearly presents serious difficulties for unclassified data particularly in view of the short records commonly used in hydrologic work. Maximum likelihood estimates for the special case of parameters of multivariate mixtures with common but unknown covariance matrices have been discussed by Day (1969), but a more general maximum likelihood approach is not available to my knowledge.

Parameter estimation for the purposes of this work is based on a subjective classification of the data into two states, followed by estimation of the parameters conditioned on the state by means of moment estimators. The two states used in the examples in Section 7.2 are widespread drought (i.e. unusually dry conditions affecting all sites) and "normal" conditions (comprising all events not classified as drought). Further details of this approach are given in the following examples.

7.2 Applications of Multivariate Mixture Models

7.2.1 Port Hardy and Eureka

The monthly January data from Port Hardy and Eureka were used in Section 4.2.1 to demonstrate the difficulties of modeling monthly precipitation at widely separated sites using conventional models. In this section the same data are modeled by means of a mixture of two bivariate normal distributions in an attempt to improve the representation of the inter-station relationships. The locations of Port Hardy and Eureka were shown in Figure 4.1, basic statistics for the data were shown in Tables 4.2 and 4.3 and a scatterplot of the data was shown in Figure 4.5

The mixture parameters were estimated by first identifying observations associated with dry conditions at both sites. A dry condition was rather arbitrarily defined as any observation less than one-half a standard deviation below the mean. This definition gave six observations for which dry conditions prevailed at both sites; these observations were classified as coming from a drought state, state 1. The remaining 26 observations were associated with a normal state, state 2. The mean, standard deviation and cross correlation for observations from both states were then calculated by the method of moments. The resulting estimates are shown in Table 7.1.

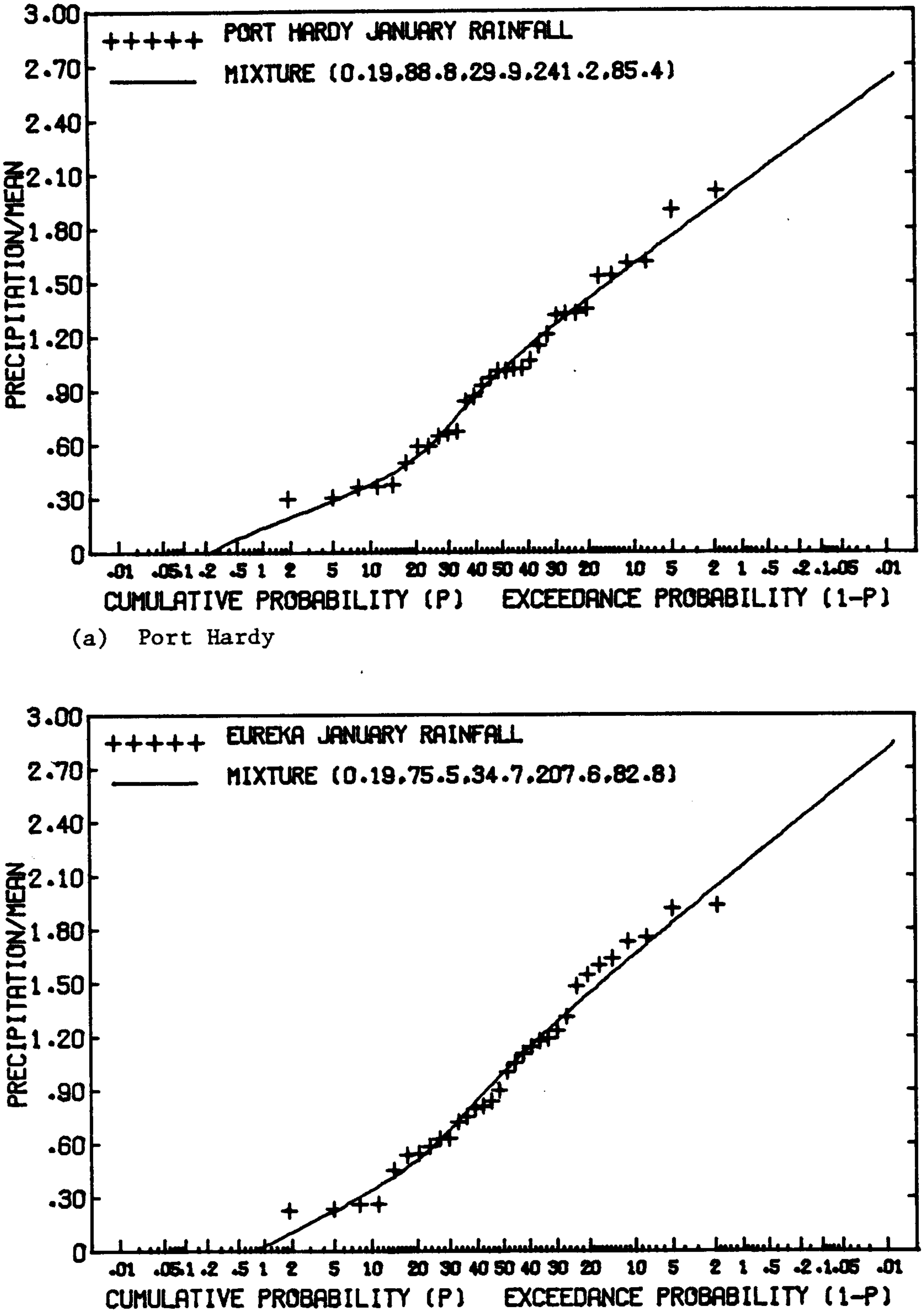
Table 7.1 Bivariate Mixture Parameters for January Precipitation Data at Port Hardy and Eureka

	No. Obs.	Mean (mm)	St. Dev. (mm)	Cross Correlation
State 1				
Port Hardy	6	88.8	29.9	0.40
Eureka	6	75.5	34.7	
State 2				
Port Hardy	26	241.2	85.4	
Eureka	26	207.6	82.8	-0.39
Mixing proportion p_1 =	0.19			

The difficulty of estimating parameters for the mixture model is obvious from the small sample associated with state 1 (drought state). The state 1 parameter estimates are unreliable, but the state 2 parameter estimates with 26 observations should be reasonably stable. Assuming that the mixing proportion p_1 is a reasonable estimate, the state 1 parameters only affect the lower tail of the marginal distribution, which is not well defined for any distribution. The use of mixture distributions in this case thus allows for adjustments in the shape of the lower tail of the marginal distribution by altering the relatively unreliable state 1 parameter estimates.

The cumulative distribution functions for the Port Hardy and Eureka data are shown in Figures 7.1a and 7.1b along with the marginal mixture distributions given by the parameters in Table 7.1. The mixture parameters provide quite a good fit to the observed data. The fit is usually as satisfactory as that for the LN3 distributions shown in Figure 4.7. The effect of changes in the mixture parameters on the marginal distribution of the Port Hardy data is shown in Figures 7.2a to 7.2c. Note that the parameters used are given in the annotation on the figures.

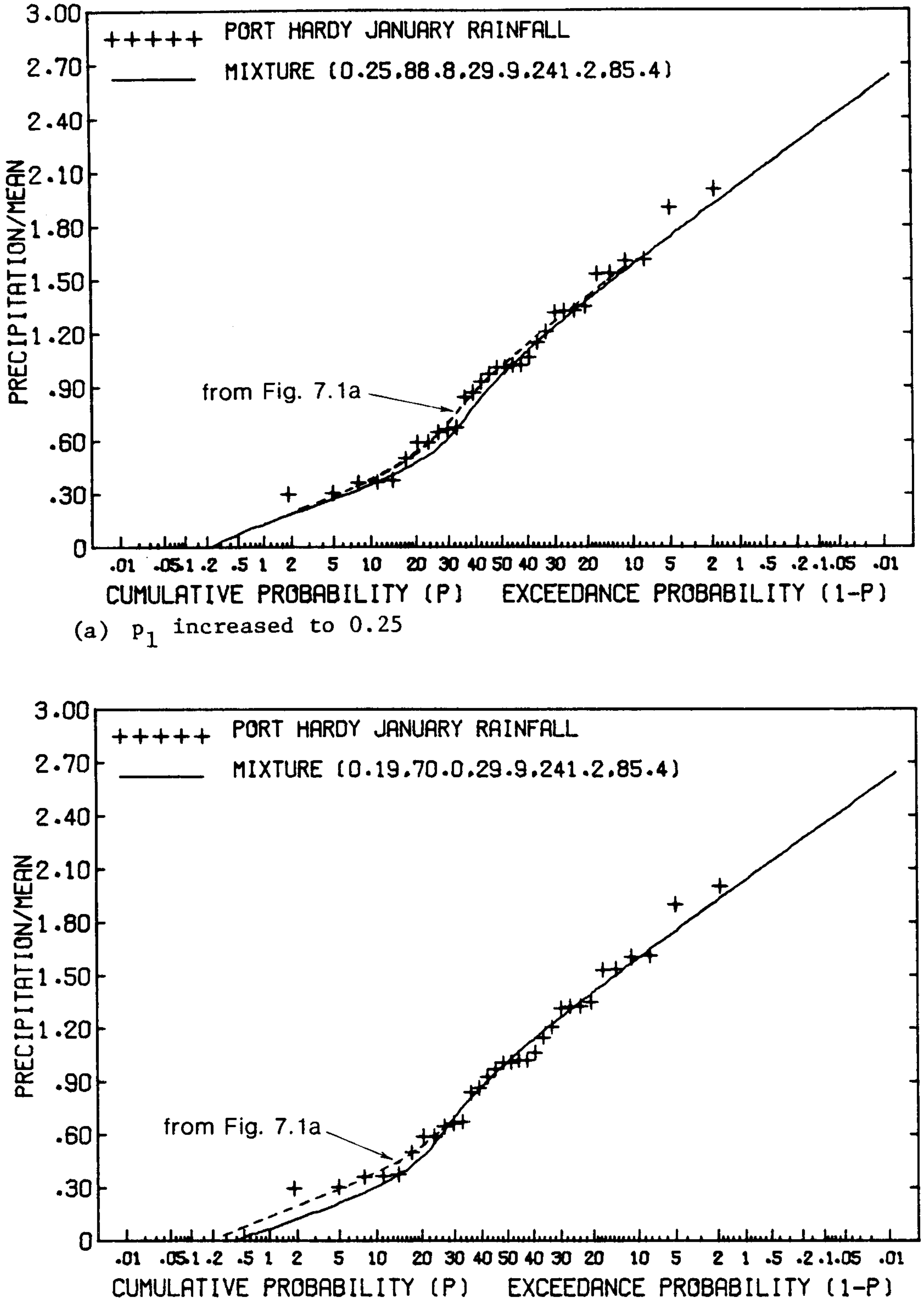
Using a similar approach to that of Section 4.2.1, the inter-station characteristics of the multivariate mixture model were investigated by means of Monte Carlo simulation. The generation scheme described in the previous section was used to create 3200 years of synthetic January data at Port Hardy and Eureka. The parameters used were those given in Table 7.1. The 3200-year synthetic sequence corresponds to one hundred 32-year sequences or one hundred sequences of the same length as the historic record. The 10, 15, 20,...50 percent quantiles for the synthetic sequences from Port Hardy and Eureka were determined and then joint occurrences were counted in which both sites had rainfall less than or equal to their respective 10, 15, 20,...50 percent quantiles. These counts were then divided by 100 to give an estimate of the expected number of joint occurrences in a period of 32 years. These data and the corresponding data from the historic record are shown in Figure 7.3. For comparative purposes Figure 7.4 shows a similar plot for joint occurrences where both sites had rainfall greater than their respective 10, 15, 20,...50 percent quantiles. As in Section 4.2.1, the actual numbers of joint low events in the one hundred 32-year periods comprising the synthetic record were recorded and are given in Table 7.2 for various quantile levels.



(a) Port Hardy

(b) Eureka

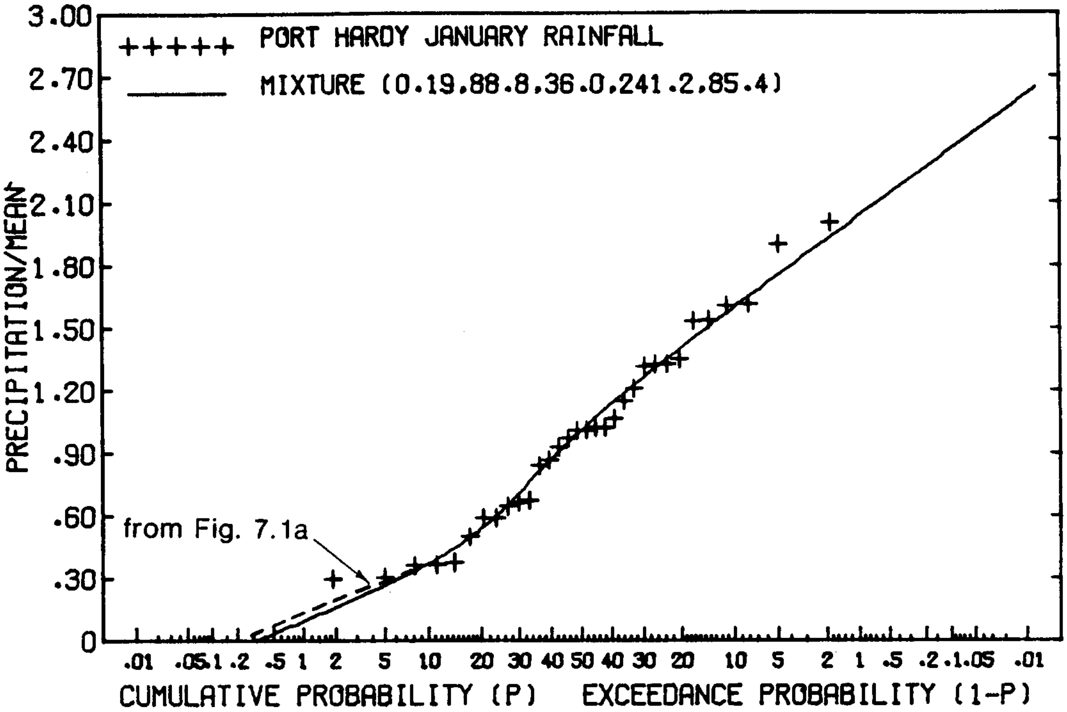
Figure 7.1 CDF's for January precipitation data



(a) p_1 increased to 0.25

(b) μ_1 reduced to 70.0

Figure 7.2 Effect of parameter changes on form of mixture CDF



(c) σ_1 increased to 36.0

Figure 7.2 Continued

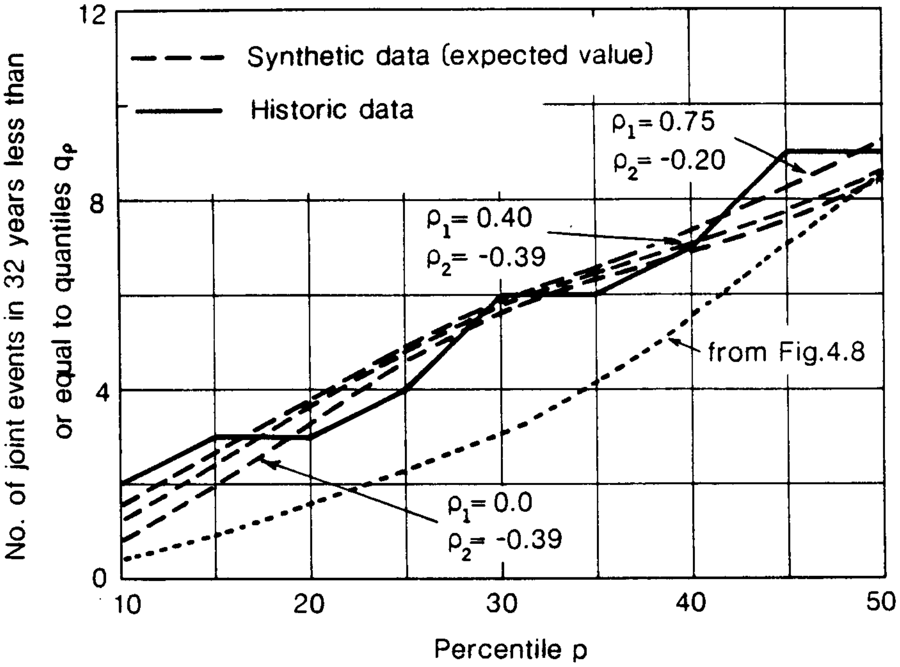


Figure 7.3 Occurrences of joint low events at Port Hardy and Eureka

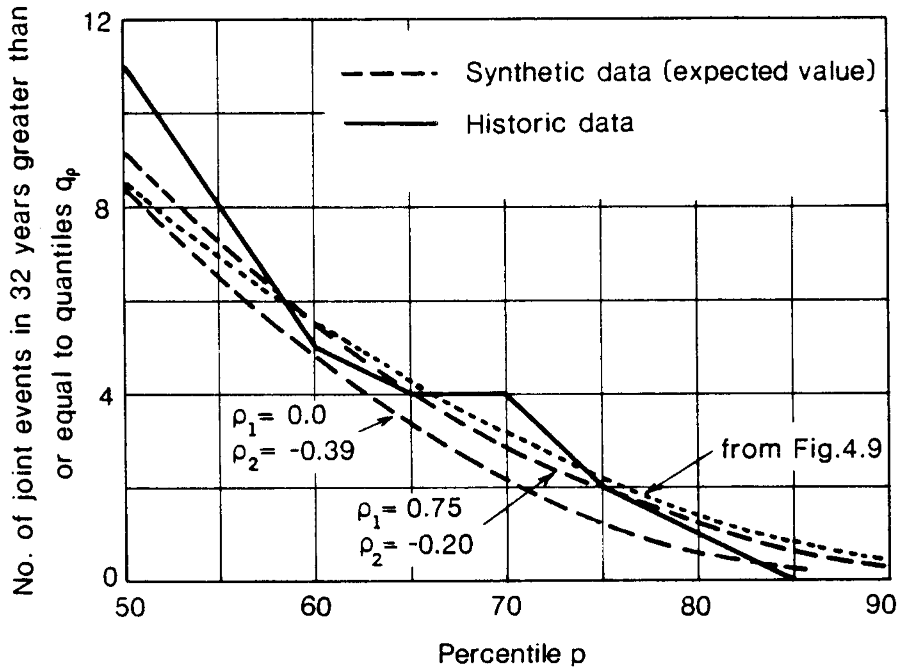


Figure 7.4 Occurrences of joint high events at Port Hardy and Eureka

Table 7.2 Occurrences of Joint Low Events in January Synthetic Record at Port Hardy and Eureka

Percentile		Number of 32-year Synthetic Sequences Having n Joint Events Less Than or Equal to Quantile q_p													Historic Joint Events*
p	n	0	1	2	3	4	5	6	7	8	9	10	11	12	
Expt. 1 ($\rho_1 = 0.4$, $\rho_2 = -0.39$)															
10		35	26	28	7	4									2
15		11	18	25	21	16	6	2	1						3
20		4	5	13	15	29	13	10	6	1	3				3
25		1	6	3	13	15	29	13	10	6	1	3			4
Expt. 2 ($\rho_1 = 0.0$, $\rho_2 = -0.39$)															
10		43	40	13	4										2
15		14	26	27	17	13	1	2							3
20		4	10	20	27	18	10	7	2	1	1				3
25		0	5	16	11	15	20	17	6	6	3	0	1		4
Expt. 3 ($\rho_1 = 0.75$, $\rho_2 = -0.20$)															
10		19	35	28	13	5									2
15		8	19	21	18	7	4	1							3
20		2	8	16	19	24	16	8	2	4	1				3
25		1	4	9	9	23	20	11	13	6	1	1	1	1	4

Note: Means and variances used are those given in Table 7.1.

*Refers to number of joint events in historic record less than or equal to quantile q_p

As mentioned earlier with the small sample available it is recognized that the state 1 parameter estimates in particular are quite unreliable. It is clear from Figures 7.1 and 7.2 that the state 1 mean and variance can be adjusted to give a subjective best fit over the lower quantiles of the CDF; it is really the form of the CDF that is of interest and not the parameter values per se. Similarly, it is not really the cross correlation coefficient that is of principal interest in water resources planning, rather it is the related distribution of joint events. Figures such as 7.3 and 7.4 give a more direct indication of both the cross properties of the data and the adequacy of the generation scheme. Consequently, both the state 1 and state 2 correlations may be regarded as parameters which should be selected and adjusted so that the synthetic and historic data in Figures 7.3 and 7.4 conform reasonably closely.

The effect of adjustments to the state 1 and state 2 cross correlations (ρ_1 and ρ_2) on the distribution of joint events was investigated by repeating the above experiments with correlations (ρ_1= 0.0, ρ_2= -0.39) and (ρ_1= 0.75, ρ_2= -0.20). The results are again plotted in Figures 7.3 and 7.4 and summarized in Table 7.2.

Figures 7.3 and 7.4 show that the multivariate mixture model is capable of adequately representing both the marginal distributions of the historic data and the inter-station precipitation relationships. Comparison of Figures 7.3 and 7.4 with Figures 4.8 and 4.9 and Table 7.2 with Table 4.8 demonstrates the superior ability of the multivariate mixture model vis-a-vis more conventional models in preserving the spatial characteristics of drought.

7.2.2 Pacific Northwest Data

Precipitation data from seven sites in the Pacific Northwest were used in Section 4.2.2 to evaluate the possible effect of dimensionality on the ability of conventional models to reproduce the observed joint occurence of drought at multiple points. It was found that current models tend to undersimulate joint drought occurence for the Pacific Northwest data.

In this section the same data (i.e. 32 years of monthly January data from Victoria, Vancouver, Sedro Woolley, Snoqualmie Falls, Longmire, Kid Valley and Centralia) are used with a multivariate mixture model to determine whether the additional flexiblity of the mixture models can improve significantly our ability to model the spatial characteristics of drought. The essential features of the data were discussed in Section 4.2.2 and the locations of the sites is shown in Figure 4.2.

The modeling approach adopted is similar to that used in the previous section. The model parameters were again estimated by classifying as drought events (state 1) those events for which all sites had precipitation less than a one-half standard deviation below the mean. This produced five observations from state 1 with the remaining 27 observations being classified as associated with state 2. The means and standard deviations of the state 1 and state 2 data are shown in Table 7.3 and the cross correlation matrices are shown in Table 7.4. Again it is recognized that the state 1 parameter estimates are probably unreliable because of the small sample available.

The CDFs for data from the seven sites are shown in Figures 7.5a to 7.5g along with the marginal mixture distributions with parameters given in Table 7.3. The same CDFs with fitted LN3 distributions were shown in Figure 4.12.

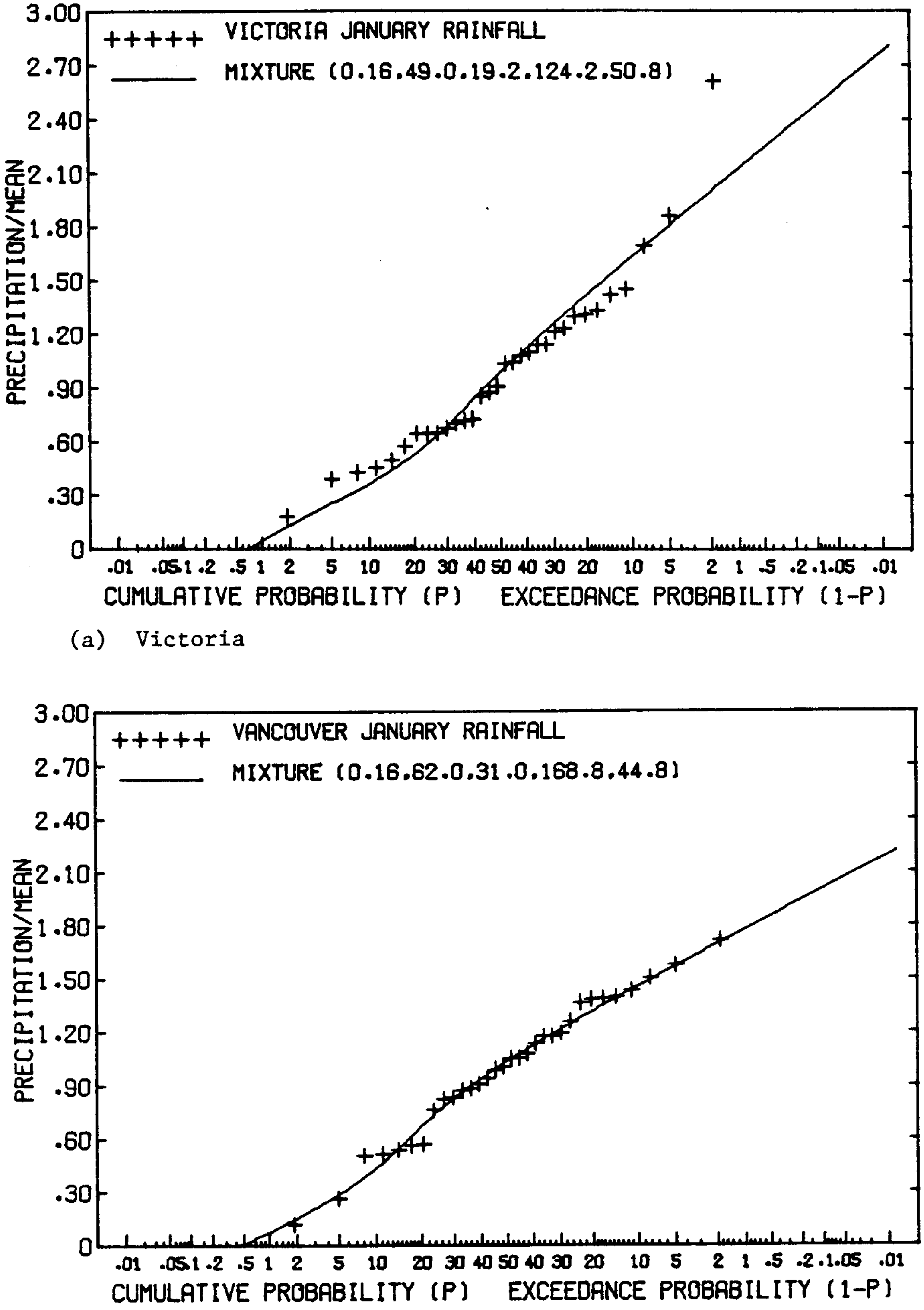
Table 7.3 Multivariate Mixture Parameters for January Precipitation Data at Seven Sites in the Pacific Northwest

	State 1		State 2	
	Mean (mm)	St. Dev. (mm)	Mean (mm)	St. Dev. (mm)
Victoria	49.0	19.2	124.2	50.8
Vancouver	62.0	31.0	168.8	44.8
Sedro Woolley	86.3	21.4	163.0	74.0
Snoq. Falls	79.6	21.4	259.0	85.0
Longmire	111.0	50.0	390.0	156.3
Kid Valley	67.8	33.5	249.0	95.3
Centralia	56.9	26.4	210.3	71.6

Table 7.4 Cross Correlations for January Precipitation Data at Seven Sites in the Pacific Northwest

State 1	Victoria	Vancouver	Sedro Woolley	Snoq. Falls	Longmire	Kid Valley	Centralia
Victoria	1.0	.60	.79	.86	.67	.55	.71
Vancouver	.60	1.0	.93	.81	.47	.63	.33
Sedro Woolley	.79	.93	1.0	.96	.73	.81	.62
Snoq. Falls	.86	.81	.96	1.0	.86	.83	.68
Longmire	.67	.47	.73	.86	1.0	.91	.79
Kid Valley	.55	.63	.81	.83	.91	1.0	.80
Centralia	.71	.33	.62	.68	.79	.80	1.0
State 2 Experiment 1[*]							
Victoria	1.0	.51	.42	.76	.65	.66	.68
Vancouver	.51	1.0	.49	.46	.23	.35	.35
Sedro Woolley	.42	.49	1.0	.52	.66	.53	.50
Snoq. Falls	.76	.46	.52	1.0	.85	.87	.85
Longmire	.65	.23	.66	.85	1.0	.86	.83
Kid Valley	.66	.35	.53	.87	.86	1.0	.94
Centralia	.68	.35	.50	.85	.83	.94	1.0
State 2 Experiment 2[*]							
Victoria	1.0	.51	.10	.76	.65	.66	.68
Vancouver	.51	1.0	.10	.46	.40	.40	.40
Sedro Woolley	.10	.10	1.0	.10	.10	.10	.10
Snoq. Falls	.76	.46	.10	1.0	.85	.87	.85
Longmire	.65	.40	.10	.85	1.0	.86	.83
Kid Valley	.66	.40	.10	.87	.86	1.0	.94
Centralia	.68	.40	.10	.85	.83	.94	1.0

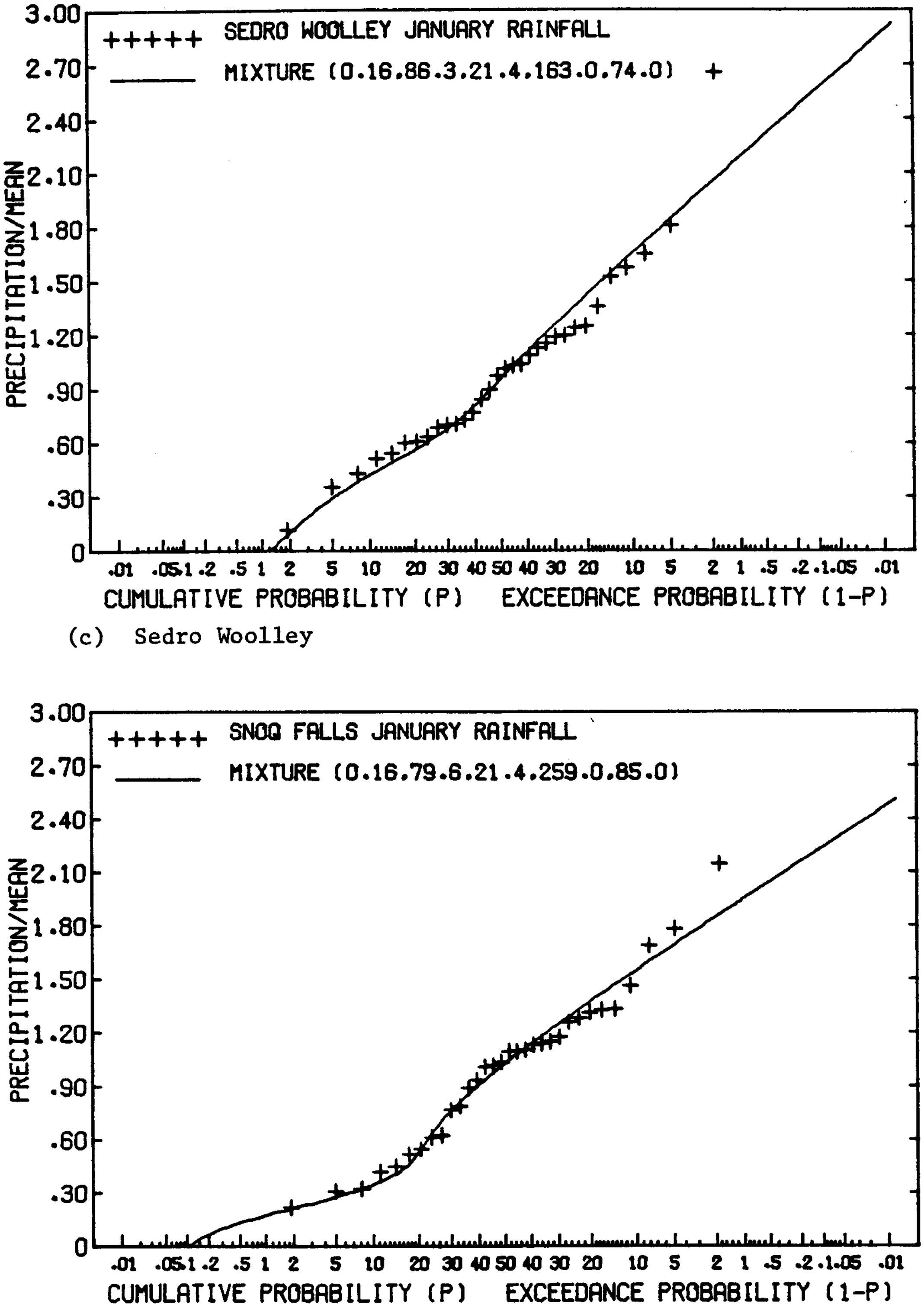
[*] See Figures 7.6, 7.7, 7.9, 7.10



(a) Victoria

(b) Vancouver

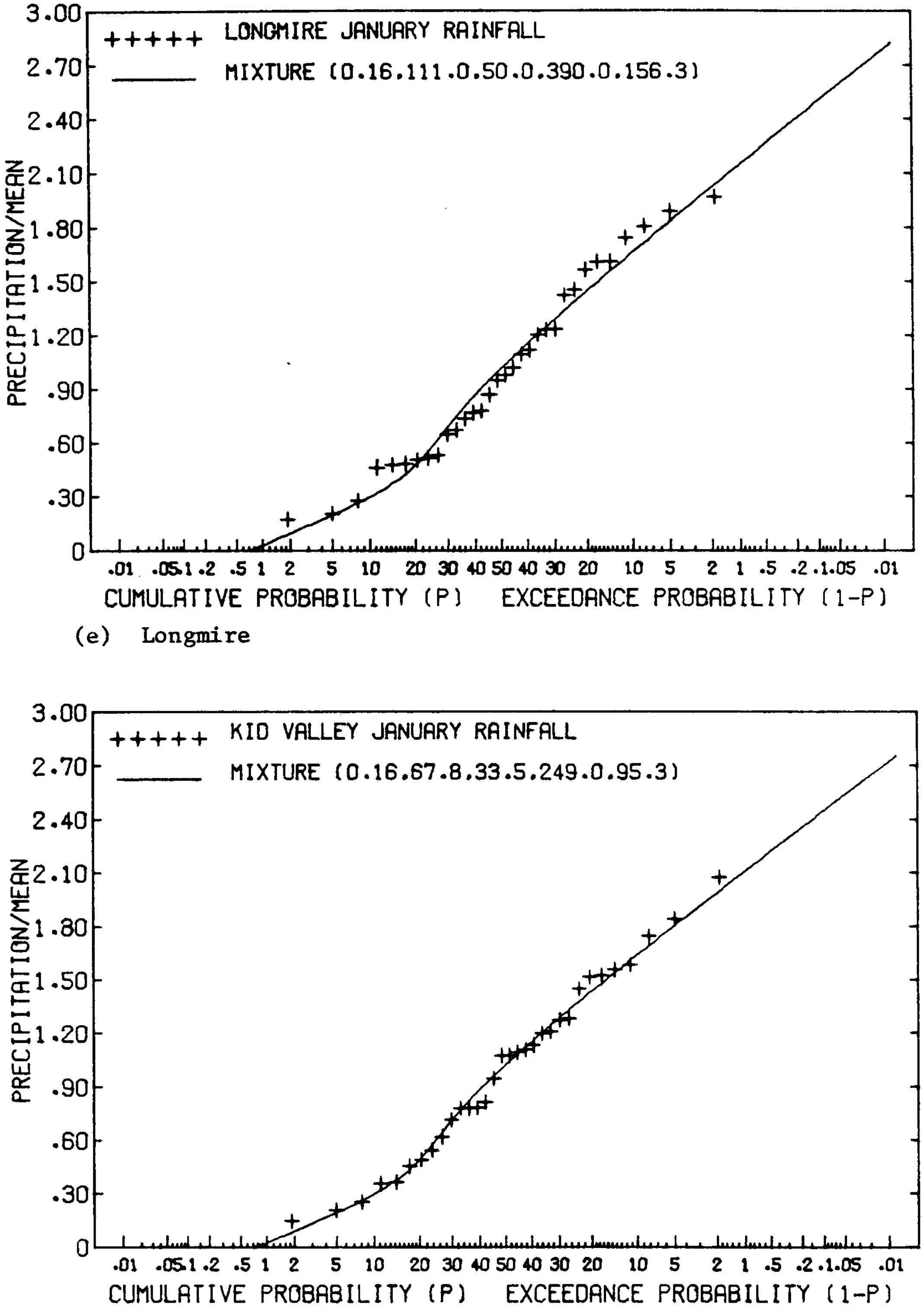
Figure 7.5 CDF's for January precipitation data



(c) Sedro Woolley

(d) Snoqualmie Falls

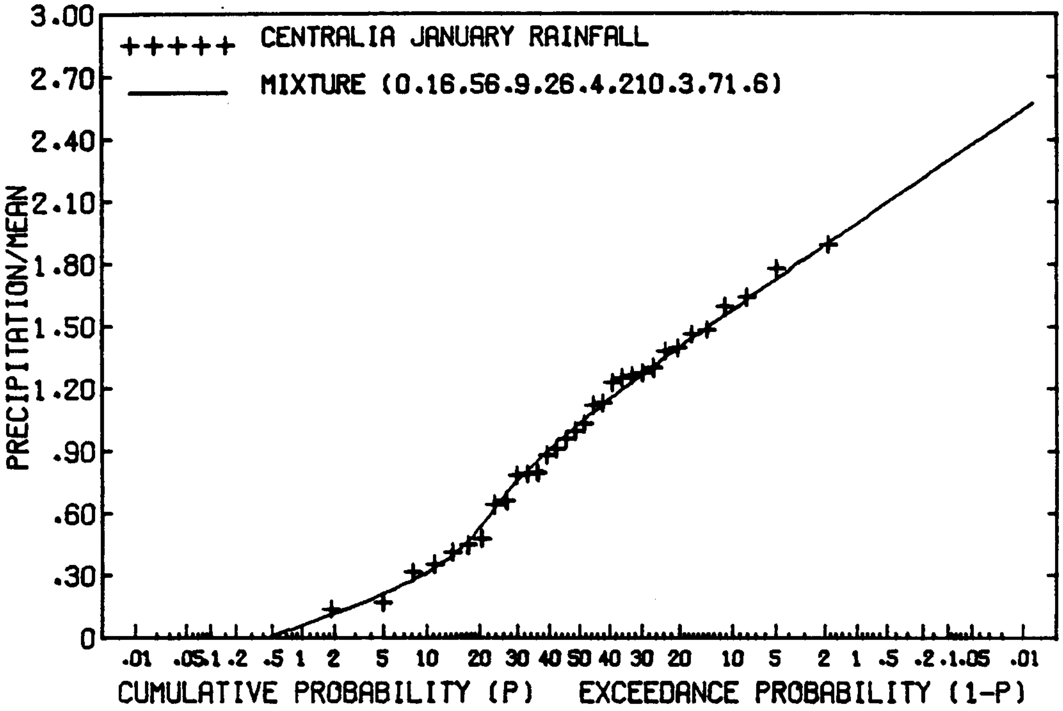
Figure 7.5 Continued



(e) Longmire

(f) Kid Valley

Figure 7.5 Continued



(g) Centralia

Figure 7.5 Continued

The mixture distributions provide a good fit at all sites except perhaps Victoria and Sedro Woolley. Even at these sites, the fit at the low and middle quantiles is adequate. The lack of fit arises mainly from the largest observation, and if this were ignored, the fit would be reasonable over all quantiles. With the exception of the Victoria and Sedro Woolley data, the mixture distributions provide fits which are as good as those provided by the LN3 distributions shown in Figures 4.12.

Following the approach described in the previous section, the multivariate mixture model was used to generate 3200 years of synthetic January precipitation data at the seven sites in question. The parameters used initially were those given in Tables 7.3 and 7.4. To find a real solution to the matrix equation $\underline{BB}^T = \underline{M}$, however, the correlation matrix $\underline{M}_1$ must be positive definite. Unfortunately this was not the case with the initial estimate of the state 1 correlation matrix $\underline{M}_1$ and some adjustments had to be made to the elements of the matrix. These adjustments were made using Fiering's (1968) eigenvalue modification scheme. The final adjusted positive definite matrix differed only slightly from the initial estimate of $\underline{M}_1$. The difficulties encountered with the state 1 correlation matrix are probably a direct result of the small sample size available and again indicate a lack of reliability in the parameter estimates.

The results of the Monte Carlo experiment are summarized using the procedure described in Chapter 4 (i.e. in terms of the frequency of occurence of joint events at all sites) in Figures 7.6 and 7.7, and the actual number of occurences of joint low events in the synthetic record are given in Table 7.5.

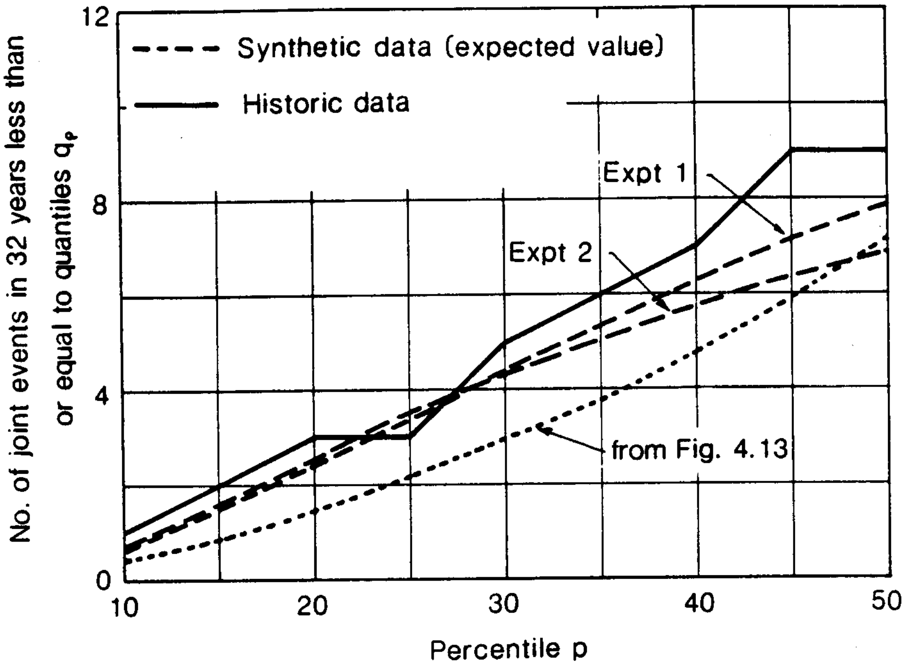


Figure 7.6 Occurrences of joint low events at seven sites in the Pacific Northwest

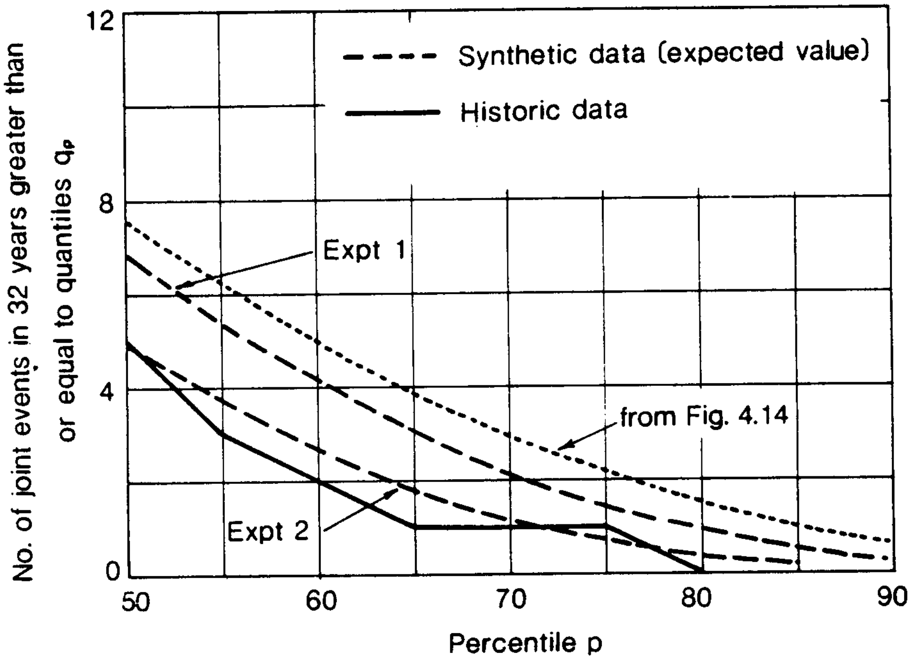


Figure 7.7 Occurrences of joint high events at seven sites in the Pacific Northwest

Table 7.5 Occurrences of Joint Low Events in January Synthetic Record at Seven Sites in the Pacific Northwest

		Number of 32-year Synthetic Sequences having n Joint Events Less than or Equal to Quantile q_p								
Percentile p	n	0	1	2	3	4	5	6	7	8
Run 1										
10		49	38	13						
15		22	33	25	16	3	1			
20		7	23	24	24	11	7	3	1	
25		3	11	17	25	22	11	6	2	3

Comparison of Figures 7.6 and 7.7 with Figures 4.13 and 4.14 and Table 7.5 with Table 4.12 indicates that use of the multivariate mixture model again results in a significant improvment in reproducing the observed frequency of widespread drought. The frequency of occurrence of joint synthetic high events also shows some improvement but is still over simulated.

In the previous section it was suggested that correlation coefficients be adjusted to ensure reasonable agreement between the observed and synthetic distributions of joint events. For the case of two sites only this is a relatively easy task. However, for seven sites this approach may be impractical. Clearly an inability to reproduce joint events at seven sites could be caused by difficulties with the modeling at only one of those sites. Thus to ensure correct modeling of joint events at multiple sites, it would seem necessary to investigate the occurrence of joint events at all possible pairs of sites (21 combinations for the seven site case).

A study of this nature showed that the high events occuring concurrently at Sedro Woolley and other sites were not being modeled correctly. A scatterplot of the January data at Sedro Woolley and Snoqualmie Falls (Figure 7.8) illustrates the nature of the problem. Apparently the correlation between rainfall depths at Sedro Woolley and other sites is small for high events. Since the Sedro Woolley data are thought to be reasonably representative of conditions over parts of the North Cascades (Rasmussen and Tangborn 1976), these problems are probably due to meteorological conditions rather than errors or inconsistencies in the data. The occurrence of joint low and high events at Sedro Woolley and Snoqualmie Falls for the historic and synthetic data is summarized in Figure 7.9 and 7.10.

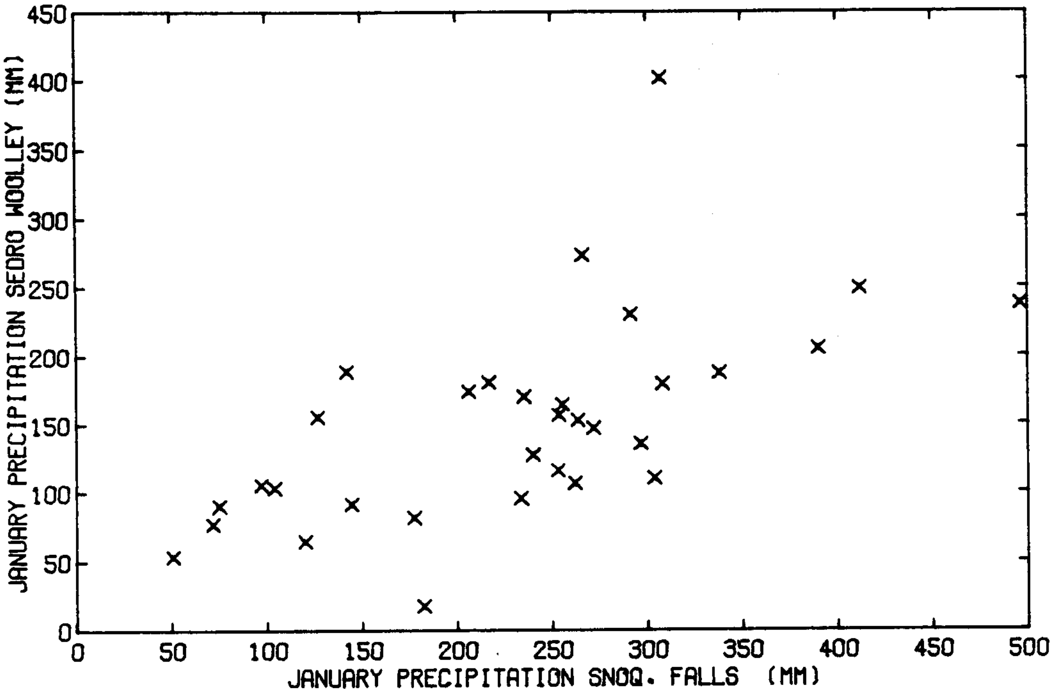


Figure 7.8 Scatterplot of January precipitation data, Sedro Woolley vs. Snoqualmie Falls

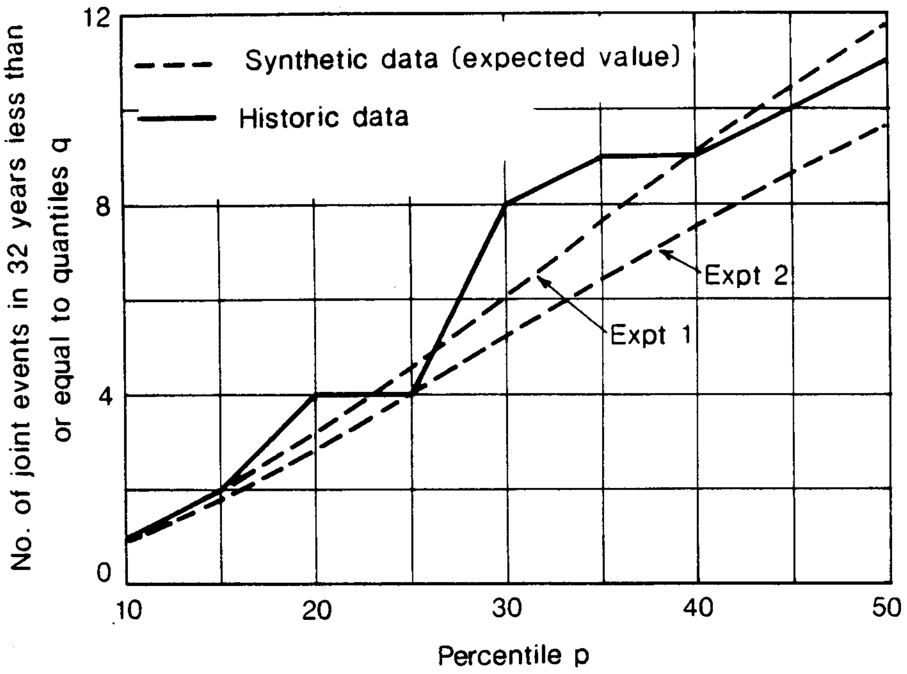


Figure 7.9 Occurrences of joint low events at Sedro Woolley and Snoqualmie Falls

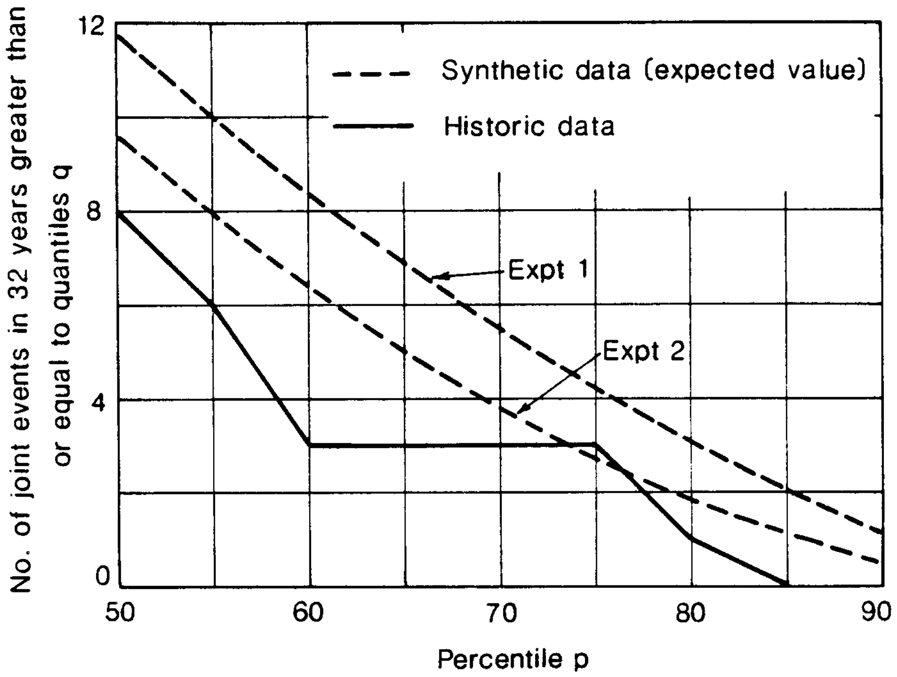


Figure 7.10 Occurrences of joint high events at Sedro Woolley and Snoqualmie Falls

A considerable improvement in the simulation of joint high events was achieved by reducing the state 2 correlation coefficients for the Sedro Woolley data to 0.1. Small adjustments were also made to the state 2 correlation coefficients for the Vancouver data. The adjusted correlation matrix is shown in Table 7.5, denoted as state 2 experiment 2.

The Monte Carlo experiments were repeated using the revised correlation matrix. Some additional minor adjustments were again necessary to ensure that the matrix was positive definite. The results obtained are summarized as before in Figures 7.6, 7.7, 7.9, and 7.10.

The substantial improvements in the simulation of joint high events at Sedro Woolley and Snoqualmie Falls (Figure 7.10) and at all seven sites (Figure 7.7) is obtained at the cost of some deterioration in the simulation of joint low events. A decision as to whether one should try to correctly model only the joint low events or joint events at all quantile levels depends on the nature of the problem being addressed. In some circumstances the joint low events may be of principal concern whereas in other situations joint events over a wider range of quantiles may be of interest. In the latter case some deterioration in the simulation of joint low events may be acceptable if the simulation were improved for other quantile levels.

The multivariate mixture model presented here is obviously useful in its ability to represent both the marginal distributions of the historic data and the observed cross properties of the data. The model has the additional attractive feature that the proportion of widespread drought is controlled directly by one of the model parameters thus introducing some degree of consistency into the parameter estimates.

8.0 SUMMARY AND CONCLUSIONS

8.1 Summary

The primary motivation for this work was the belief that consideration of the physical mechanisms involved in the hydrologic cycle could lead to improvements in the techniques currently used in stochastic hydrology. Of particular concern to water resources planning is the frequency, severity and areal extent of droughts, which on the west coast of North America can be related to anomalous conditions in large scale atmospheric circulation.

A brief review of the concepts and past developments underlying stochastic hydrology is given in Chapter 2. This review illustrates the general lack of physical bases in most work in this area. In particular, no consideration seems to have been given to the role of the large scale atmospheric processes associated with drought.

The known qualitative relationships between precipitation and patterns of atmospheric circulation are reviewed in Chapter 3. The most important qualitative association for water resources planning is between precipitation and the track of the jet stream. Persistent anomalous jet stream tracks lead to severe droughts such as those which occurred on the west coast during 1977 and 1978. The nature of the atmospheric circulation associated with drought suggests that inter-station precipitation relationships are nonlinear with higher cross correlations during drought than during wet or normal periods. This indicates that current multi-site stochastic models, which all assume linear inter-station relationships, may under simulate the areal extent of drought.

As noted in Chapter 3 both observational studies and theoretical work in the atmospheric sciences have led to the hypothesis that atmospheric circulation may exist in one of two quasi-stable states;

one state associated with meridional circulation and dry conditions and the other with zonal flow and generally wet conditions. This hypothesis and the previously noted nonlinear inter-station precipitation relationships indicate that rainfall may perhaps be modeled better using a mixture of two distributions; one distribution associated with meridional flow and dry conditions and the other associated with zonal flow and wet conditions.

Multi-site precipitation data from the west coast were analyzed in Chapter 4 in an attempt to detect nonlinearities in the inter-station relationships. Many existing multi-site models generate synthetic data by sampling from multivariate normal distributions. It is hypothesized that these models suffer from two deficiencies which have been termed the "scale effect" and "dimensionality effect". Both of these effects are related to the large areal extent of severe drought.

The scale effect refers to the potential inability of conventional models to reproduce correctly the observed frequency of extreme joint low events occurring concurrently at widely separated sites. This is illustrated by analyzing the January precipitation data from Port Hardy and Eureka (separation 1300 km). Attempts to model concurrent Port Hardy and Eureka January monthly data using a conventional model resulted in a substantial under simulation of joint drought events. From a practical standpoint this has serious implications for the design of projects such as those involving the long distance transfer of water or hydro power. It is clear that conventional models are unsuitable for multi-site modeling in such situations.

The dimensionality effect refers to the hypothesized inability of conventional models to reproduce correctly concurrent drought at multiple points in an area where drought at one point is invariably associated with drought at all other points. This is illustrated by modeling the January precipitation at seven sites in the Pacific Northwest. The conventional model gave a consistent under simulation of joint drought occurrences, but the discrepancies between the historic and synthetic data were not as significant as those encountered when modeling the joint Port Hardy and Eureka data. The dimensionality effect is potentially significant in the generation of multi-site synthetic data for studying the operation or design of large water resources systems such as the Columbia basin system.

Although the work in Chapter 4 is based on the analysis of a limited amount of data, it is clear that the assumption of linear inter-station relationships may be invalid in some multi-site studies. While conventional multi-site models may be appropriate in any particular situation, the assumptions of linearity should be checked and the consequence of nonlinearities in the observed data evaluated.

In an attempt to understand better the relationships between precipitation and circulation, an analysis of concurrent precipitation and 500 mb geopotential height data was undertaken. The results of this analysis presented in Chapter 5 generally confirm the previously known qualitative relationships between precipitation and circulation, but attempts to develop more detailed quantitative relationships were unsuccessful.

The hypothesis put forward in Chapter 3 that rainfall comes from a mixed distribution implies a classification of the data into two or more populations. It had been hoped to classify the data into wet and dry populations based on the prevailing type of atmospheric circulation, (i.e. meridional or zonal). However, the work in Chapter 5 showed that the distinction between zonal and meridional conditions is quite subjective except in extreme situations. As a result no useful method was found to classify the precipitation data by circulation type.

The parameters of mixture models can be estimated, however, in the absence of suitable information for classifying the data. The characteristics of simple univariate mixtures of two normal distributions were investigated in Chapter 6. Various methods for parameter estimation were explored with particular emphasis on maximum likelihood estimates. Although mixtures have been used in various fields for some time, little is known about the small sample properties of the MLE's. Consequently, a number of Monte Carlo experiments were performed to determine the mean and variance of the ML parameter estimates for both classified and unclassified samples of sizes 50 and 100. As had been expected, parameter estimates from the unclassified data showed considerably greater bias and variability than the estimates from the classified data. However, the quantiles of the distributions fitted to classified data proved to be only slightly better estimates than those quantiles obtained from the unclassified data. Since the interest in water resources is primarily in the quantiles of a distribution and not in the parameters per se, these findings indicate that the ability to classify the data from a mixture distribution is not necessarily important.

Attempts were made in Chapter 6 to fit mixture distributions to precipitation records from Victoria and Agassiz, these stations having some of the longest records available in the area. Unfortunately, again no evidence was found to justify the use of univariate mixture models for single-site precipitation modeling.

The use of multivariate mixture models for multi-site precipitation modeling was explored in Chapter 7. Such models are attractive for multi-site modeling because they are perhaps the simplest form of model which incorporate nonlinearities in the inter-station relationships. The model was applied to the Port Hardy and Eureka data used to evaluate the scale effect in Chapter 4. The performance of the multivariate mixture model was quite encouraging in that it was able to preserve both the marginal distributions of the data and the nonlinear inter-station relationships. Its performance was superior to that of the conventional model used in Chapter 4.

The multivariate mixture model was also applied to the seven site data used in Chapter 4. Although the model again performed better than its conventional counterpart, severe difficulties were encountered in estimating consistent (i.e. positive definite) correlation matrices, and the parameter estimation method is far from satisfactory.

8.2 Conclusions

The following conclusions regarding seasonal multi-site precipitation modeling may be drawn from this study:

(1) Both recorded data and physical considerations show that inter-station precipitation relationships are nonlinear for widely separated sites on the west coast of North America (separation greater than 1,000 km) with higher cross correlation during drought than during wet periods.

(2) Monte Carlo simulation shows that as a result of such nonlinearities, conventional multi-site models, which all assume linear inter-station relationships, could seriously under simulate the frequency and areal extent of drought at widely separated points.

(3) Evaluation of data synthesized by a conventional multi-site model at seven sites in the Pacific Northwest also revealed a tendency to under simulate widespread drought in high dimensionality problems where stations separation is not necessarily large. Such difficulties are attributed to the fact that droughts generally affect considerably larger areas than are affected by frontal systems.

(4) In many situations the conventional cross correlation coefficient is an inadequate descriptor of inter-station precipitation relationships. An alternative descriptor of inter-station relationships is suggested which is based on plots of the occurrence of joint events below (or above) specified quantile levels.

In exploring alternative approaches for the analysis and synthesis of precipitation sequences, it was hypothesized that precipitation may be treated as coming from a mixture of two distributions; one distribution associated with meridional atmospheric circulation and dry conditions, the other with zonal flow and wet or normal conditions. Analysis of concurrent precipitation and 500 mb geopotential height fields led to the following conclusions:

(1) Relationships between precipitation depths and simple characteristics of the 500 mb field are too weak to be of quantitative value in the field of stochastic hydrology. The data examined, however, do support the previously known qualitative relationships between precipitation and atmospheric circulation.

(2) There is no evidence in the geopotential height data to suggest that atmospheric circulation may be objectively classified into only two states: zonal and meridional.

(3) No basis was found for using 500 mb data to classify rainfall data into dry or wet states dependent on the type of atmospheric circulation.

A detailed study of the properties of univariate mixture models led to the following findings:

(1) Parameters estimated from small unclassified samples drawn from a mixture of two normal distributions exhibit great variability. Accurate parameter estimates from samples of the size available in hydrologic work are only possible if the observations in the sample can be classified by state.

(2) Despite the variability of the parameter estimates, the quantiles of mixture distributions fitted to small classified samples are only slightly more accurate than the quantiles of distributions fitted to unclassified data. Since most hydrologic work requires a knowledge of the quantiles of a distribution rather than the parameters per se, this suggests that the ability to classify observations from mixed distributions is not necessarily important in hydrologic work.

(3) Analysis of long (81 years) rainfall records from two sites in the Pacific Northwest failed to demonstrate the presence of mixture distributions. Although mixtures of two normal distributions fitted annual data well, the mixture was subtle, and a simpler 2 or 3 parameter distribution would have provided a satisfactory fit to the data.

Despite difficulties in justifying the use of univariate mixture distributions in single site precipitation synthesis, multivariate mixture distributions were found to have a number of attractive features for multi-site modeling:

(1) They are perhaps the simplest form of model which supports a nonlinear cross correlation structure.

(2) Experiments performed in this study showed that multivariate mixture models are capable of preserving both the marginal distributions and cross correlation structure of data which

could not be modeled adequately using more conventional models.

(3) Mutivariate mixture models allow explicit recognition of the wide spread nature of drought.

8.3 Implications and Recommendations

The work described in this report has potentially serious implications in water resources planning. I have shown that conventional multi-site stochastic models may greatly understate the areal extent of drought on the west coast of North America. Difficulties with existing models stem from their assumption that inter-station relationships are linear. A class of multivariate mixture models is explored which show great promise in its ability to preserve both the marginal distributions and cross properties of the historic data.

The difficulties with conventional models are especially evident in attempting to model concurrent drought events at widely separated sites on the west coast. This problem is of great significance if stochastic methods are to be used in the design or operation of schemes for the long-distance transfer of hydropower (for example, from northern British Columbia or southern Alaska to the southwest U.S.A.). In such situations multivariate mixture models provide an attractive and practical method for maintaining both the observed areal coverage of drought and the marginal distributions of the recorded data.

While this may appear to be an extremely limited application, the concepts explored in this dissertation are likely to be useful for large scale problems in other parts of the world. Serious nonlinearities in cross correlations may exist wherever drought occurs on larger scales than the rainfall producing mechanisms, and where there are distinct relationships between drought and large scale atmospheric circulation. Conditions such as these may occur, for example, over central Brazil where rainfall is related to the movement of the inter-tropical convergence zone (ITCZ).

During the northern hemisphere summer, the ITCZ generally lies several degrees north of the equator. During the fall, the ITCZ moves south usually lying over central Brazil during the rainy months of November through March. Failure of the ITCZ to move south over the region results in drought. Droughts over the Indonesian archipelago may be related to movement of the ITCZ in a similar manner.

While multivariate mixture models seem attractive for modeling events at widely separated sites, the justification for their use in more general multi-site applications is not as clear. As has been seen, there are severe parameter estimation problems but, more important is the practical significance of the possible discrepancies between natural multi-site sequences and the synthetic sequences produced by conventional multi-site models. Suppose for example that we wish to model rainfall at seven sites which are all affected concurrently by drought. Use of a conventional model may however produce synthetic drought at only five of the seven sites in the study area. The significance of this discrepancy depends on whether the two sites not affected by drought had substantially greater synthetic rainfall than normal or whether they were just marginally wetter than our arbitrary definition of drought. In the former case, multivariate mixture models may be of great value; in the latter case they may simply be irrelevant from a pactical viewpoint.

The value of using multivariate mixture models depends on the degree of nonlinearity in inter-station relationships. The investigation reported here studied monthly precipitation from a limited number of sites. Additional work is needed to identify more accurately situations in which nonlinear inter-station relationships present practical difficulties. In particular, work is needed to study the cross correlation structure of streamflow series, to extend multivariate mixture modeling to preserve serial correlation, and to improve methods of parameter estimation.

REFERENCES

Blackmon, M.L. et al. 1979. Geographical variations in the vertical structure of geopotential height fluctuations. Journal of the Atmospheric Sciences 36:2450-2466.

Boes, D.C., and Salas, J.D. 1978. Nonstationarity of the mean and the Hurst phenomenon. Water Resources Research 14(1):135-143.

Box, G.E.P., and Muller, M.E. 1959. A note on the generation of random normal deviates. Annals of Mathematical Statistics 29:610-611.

Box, G.E.P., and Jenkins, G.M. 1976. Times series analysis, forecasting and control. San Francisco: Holden-Day.

Bras, R.L., and Rodriguez-Iturbe, I. 1976. Rainfall generation: a nonstationary time-varying multi-dimensional model. Water Resources Research 12(3):450-456.

Burges, S.J., and Lettenmaier, D.P. 1977. Comparison of annual streamflow models. Journal of Hydraulics Division ASCE 103(HY9):991-1006.

Burges, S.J. and Lettenmaier, D.P. 1982. Reliability measures for water supply reservoirs and the significance of long-term persistence. In Decision making for hydrosystems: Forecasting and operation. Book II. Littleton: Water Resources Publications. Paper II.16.

Caffey, J.E. 1965. Inter-station correlations in annual precipitation and in effective annual precipitation. Colorado State University Hydrology Paper No. 6.

Charney, J.G., and DeVore, J.G. 1979. Multiple flow equilibria in the atmosphere and blocking. Journal of the Atmospheric Sciences 36:1205-1216.

Cleveland, W.S., and Kleiner, B. 1975. A graphical technique for enhancing scatterplots with moving statistics. Technometrics 17(4):447-454.

Cohen, A.C. 1967. Estimation in mixtures of two normal distributions. Technometrics 9(1):15-28.

Cunnane, C. 1978. Unbiased plotting positions--a review. Journal of Hydrology 37:205-222.

Day, N.E. 1969. Estimating the components of a mixture of normal distributions. Biometrika (56)3:463-474.

Dempster, A.; Laird, N.M.; and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Ser. B(39):1-38.

Eagleson, P.S. 1967. Optimum design of rainfall networks. Water Resources Research 3(4):1021-1033.

Edmon, H.J. 1980. A study of the general circulation over the northern hemisphere during the winter of 1976-1977 and 1977-1978. Monthly Weather Review 108:1538-1553.

Elliott, R.D. 1951. Extended range forecasting by weather types. In Compendium of meteorology, pp. 834-840. New York: American Meteorological Society.

Feller, W. 1968. An introduction to probability theory and its applications. Vol. 1. New York: John Wiley & Sons.

Fiering, M.B. 1968. Schemes for handling inconsistent matrices. Water Resources Research 4(2):291-297.

Fiering, M.B., and Jackson, B.B. 1971. Synthetic streamflows. Water Resources Monograph 1. Washington, D.C.: American Geophysical Union.

Forsythe, G.F.; Malcolm, M.A.; and Moler, C.B. 1972. Computer methods for mathematical computations. Stanford,California: Computer Science Department, Stanford University.

Fowlkes, E.B. 1977. Some methods for studying the mixture of two normal or log normal distributions. Unpublished Bell Telephone Laboratories technical memorandum.

Fowlkes, E.B. 1979. Some methods for studying the mixture of two normal (lognormal) distributions. Journal of the American Statistical Association 74:561-575.

Gelhard, R.H. 1961. The weather and circulation of December 1960. An unusually cold month in the United States. Monthly Weather Review 89(3):109-114.

Hasselblad, V. 1966. Estimation of parameters for a mixture of normal distributions. Technometrics 8:431-444.

Hasselblad, V. 1969. Estimation of finite mixtures of distributions from the exponential family. Journal of the American Statistical Association 64:1459-1471.

Hawkins, R.H. 1974. A note on mixed distributions in hydrology. In Proceedings of the symposium on statistical hydrology, Tucson, Arizona, August 31-September 2, 1971. pp. 336-344. Department of Agriculture, Agricultural Research Service Misc. Pub. No. 1275.

Hirsch, R.M. 1979. Synthetic hydrology and water supply reliability. Water Resources Research 15(6):1603-1615.

Hipel, K.W. 1975. Contemporary Box-Jenkins modeling in hydrology. Department of Civil Engineering, University of Waterloo, Ontario. Technical Report 75-4.

Holton, J.R. 1979. An introduction to dynamic meteorology. New York: Academic Press.

Hosmer, D.W. 1973a. On MLE of the parameters of a mixture of two normal distributions when the sample size is small. Communications in statistics 1:217-227.

Hosmer, D.W. 1973b. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics 29:761-770.

Houghton, J.C. 1978. Birth of a parent: The Wakeby distribution for modeling flood flows. Water Resources Research 14(6):1105-1109.

Hurst, H.E. 1951. Long-term storage capacity of reservoirs. Transactions ASCE 116:770-799.

Hydrocomp Inc. 1976. Hydrologic simulation programming operations manual. Palo Alto, California: Hydrocomp Inc.

Hydrocomp Inc. 1978. Hydrologic simulation of the Rio Paranaiba, Brazil. Palo Alto, California: Hydrocomp Inc.

Jackson, B.B. 1975a. The use of streamflow models in planning. Water Resources Research 11(1):54-63.

Jackson, B.B. 1975b. Markov mixture models for drought lengths. Water Resources Research 11(1):64-74.

Jackson, B.B. 1975c. Birth-death models for differential persistence. Water Resources Research 11(1):75-95.

Jenne, R.L. 1975. Data sets for meteorological research. National Center for Atmospheric Research. Boulder, Colorado. NCAR Technical Note 1A-111.

Kilmartin, R.F. 1980. Hydroclimatology--a needed cross-discipline. In Improved hydrologic forecasting: why and how. pp. 160-198. New York: American Society of Civil Engineers.

Klemes, V. 1974. The Hurst phenomenon: a puzzle? Water Resources Research 10(4):675-688.

Klemes, V.; Srikanthan R.; and McMahon, T.A. 1981. Long-memory flow models in reservoir analysis: What is their practical value? Water Resources Research 17(3):737-751.

Lawrance, A.J., and Kottegoda, N.T. 1977. Stochastic modeling of riverflow times series. Journal of the Royal Statistical Society Ser. A(140):1-47.

Ledolter, J. 1978. The analysis of multivariate time series applied to problems in hydrology. Journal of Hydrology 36:327-352.

Lettenmaier, D.P. 1980. Parameter estimation for multivariate streamflow synthesis. In Proceedings of the joint automatic control conference August 13-15, San Francisco, California. American Automatic Control Council.

Lettenmaier, D.P., and Burges, S.J. 1977. Operational assessment of hydrologic models of long-term persistence. Water Resources Research 13(1):113-124.

Leytham, K.M., and Franz, D.D. 1980. Techniques for the generation of long streamflow sequences. In Improved hydrologic forecasting: why and how. pp. 20-46. New York: American Society of Civil Engineers.

Mandelbrot, B.B., and Van Ness, J.W. 1968. Fractional Brownian motions, fractionalnoises and applications. Journal of the Society for Industrial and Applied Mathematics 10(4):422-437.

Mandelbrot, B.B. and Wallis, J.R. 1969. Computer experiments with fractional gaussian noises. Water Resources Research 5(1):228-267.

Mandelbrot, B.B. 1971. A fast fractional Gaussian noise generator. Water Resources Research 7(3):543-553.

Matalas, N.C. 1967. Mathematical assessment of synthetic hydrology. Water Resources Research 3(4):937-945.

Mejia, J.M., and Rodriguez-Iturbe, I. 1974a. Correlation links between normal and log-normal processes. Water Resources Research 10(4):689-690.

Mejia, J.M., and Rodriquez-Iturbe, I. 1974b. On the synthesis of random field sampling from the spectrum. An application to the generation of hydrologic spatial processes. Water Resources Research 10(4):705-711.

Mejia, J.M.; Rodriguez-Iturbe, I.; and Dawdy, D.R. 1972. Streamflow simulation 2: The broken line process as a potential model for hydrologic simulation. Water Resources Research 8(4):931-941.

Namias, J. 1975. Short period climatic variations. Collected works of J. Namias 1934 through 1974. San Diego, California: University of California.

Namias, J. 1978a. Multiple causes of the North American abnormal winter 1976-1977. Monthly Weather Review 106(3):279-295.

Namias, J. 1978b. Recent drought in California and Western Europe. Review of Geophysics and Space Physics 16(3):435-458.

Nelder J.A., and Mead, R. 1965. A simplex method for function minimization. The Computer Journal 7(4):308-313.

Newell, R.E. 1979. Climate and ocean. American Scientist 67:405-416.

O'Connell, P.E. 1971. A simple stochastic modeling of Hurst's law. In International symposium on mathematical models in hydrology July 26-31. Warsaw, Poland, International Association of Scientific Hydrology.

O'Connell, P.E. 1974. Stochastic modeling of long-term persistence in streamflow sequences. Rep. 1974-2. London: Hydrology Section, Department of Civil Engineering, Imperial College.

O'Connor, J.F. 1963. The weather and circulation of January 1963. One of the most severe months on record in the United States and Europe. Monthly Weather Review 91(4):209-217.

Palmen, E., and Newton, C.W. 1969. Atmospheric circulation systems. New York: Academic Press.

Pinder, G.E., and Gray, W.G. 1977. Finite element simulation in surface and subsurface hydrology. New York: Academic Press.

Potter, K.W. 1976. Evidence of nonstationarity as a physical explanation of the Hurst phenomenon. Water Resources Research 12(2):1047-1052.

Potter, K.W. 1979. Annual precipitation in the northeast United States: Long memory, short memory or no memory? Water Resources Research 15(2):340-346.

Quandt, R.E., and Ramsey, J.B. 1978. Estimating mixtures of normal distributions and switching regressions. Journal of the American Statistical Association 73:730-752.

Rasmussen, L.A., and Tangborn, W.V. 1976. Hydrology of the North Cascades Region, Washington (1) Runoff, precipitation, and storage characteristics. Water Resources Research 12(2):187-202.

Rex, D.F. 1950. Blocking action in the middle troposphere and its effects upon regional climate. I. An aerological study of blocking action. Tellus 2:196-211.

Rex, D.F. 1951. Blocking action in the middle troposphere and its effects upon regional climate. II. The climatology of blocking action. Tellus 2:275-301.

Singh, K.P. 1974. A two-distribution method for fitting mixed distributions in hydrology. In Proceedings of the symposium on statistical hydrology Tucson, Arizona, August 31-September 2, 1971. pp. 371-382. Department of Agriculture, Agricultural Research Service. Misc. Pub. No. 1275.

Stark, L.P. 1961. The weather and circulation of February 1961: An example of attenuation in the long-wave pattern. Monthly Weather Review 89(5):178-184.

Starrett, L.G. 1949. The relation of precipitation patterns in North America to certain types of jet streams at the 300 millibar level. Journal of Meteorology 6(5):347-352.

Stedinger, J.R. 1980. Fitting log-normal distributions to hydrologic data. Water Resources Research 16(3):481-490.

Tan, W.Y., and Chang, W.C. 1970. Comparisons of method of moments and method of maximum likelihood in estimating parameters of a mixture of two normal densities. Journal of the American Statistical Association 67:702-709.

Thomas, H.A., and Fiering, M.B. 1962. Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In Design of water resource systems. Maass, et al. Cambridge: Harvard University Press.

Wagner, A.J. 1978. Weather and circulation of January 1978. Monthly Weather Review 106(4):579-585.

Wallace, J.M., and Hobbs, P.V. 1977. Atmospheric Science: An introductory survey. New York: Academic Press.

Wallis, J.R., and Matalas, N.C. 1970. Small sample properties of H and K-estimators of the Hurst coefficient h. Water Resources Research 6(6):1583-1594.

Wallis, J.R., and O'Connell, P.E. 1972. Small sample estimation of ρ. Water Resources Research 8(3):707-712.

Wallis, J.R., and Matalas, N.C. 1972. Sensitivity of reservoir design to the generating mechanism of inflows. Water Resources Research 8(3):634-641.

Wallis, J.R.; Matalas, N.C.; and Slack, J.R. 1974. Just a moment! Water Resources Research 10(2):211-219.

Wallis, J.R., and O'Connell, P.E. 1973. Firm reservoir yield--how reliable are historic hydrological records. Hydrologic Sciences Bulletin 18(3):347-365.

White, W.B., and Clark, N.E. 1975. On the development of blocking ridge activity over the central North Pacific. Journal of the Atmospheric Sciences 32:489-502.

Wilk, M.B., and Gnanadesikan, R. 1968. Probability plotting methods for the analysis of data. Biometrika 55(1):1-17.

Wilson, C.G.; Valdes, J.B.; and Rodgriguez-Iturbe, I. 1979. On the influence of the spatial distribution of rainfall on storm runoff. Water Resources Research 15(2):321-328.

Yevjevich, V.; Hall, W.A.; and Salas, J.D. 1978. Drought research needs. In Proceedings of the conference on drought research needs. Colorado State University, Fort Collins, Colorado December 12-15, 1977. Fort Collins: Water Resources Publications.

APPENDIX A

MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS OF A MIXTURE OF TWO NORMAL DISTRIBUTIONS USING THE EM ALGORITHM

Making use of the development by Dempster, Laird and Rubin (1977), herein referred to as DLR, suppose that we have a sample of n observations $y = (y_1, y_2, \ldots, y_n)$ and that each y_i is associated with one of two unobserved states. We can thus define an unobserved vector $z = (z_1, z_2, \ldots z_n)$ where z_i is an indicator vector with components zero and one, the component equal to one indicating the unobserved state associated with y_i. The complete data can thus be defined as $x = (y,z)$ where y is the incomplete observed data.

A useful approach to specifying a mixture model is to first obtain the marginal distributions of the indicators z and then to specify the conditional distribution of the y_i given z_i. In the simplest type of model, as pointed out in Section 6.1.1, we can think of the z_i as being based on the results of identical and independent Bernoulli trials so that the z_i are drawn from the discrete distribution:

$$\begin{aligned} P(z_i = (1,0)) &= p_1 \\ P(z_i = (0,1)) &= p_2 = 1 - p_1 \end{aligned} \tag{A.1}$$

For mixtures of two normal distributions, the y_i given z_i are conditionally independent with the conditional distributions:

$$\begin{aligned} g(y|z_i = (1,0),\phi) &= (2\pi\sigma_1^2)^{-\frac{1}{2}} \exp[-(y-\mu_1)^2/2\sigma_1^2] \\ g(y|z_i = (0,1),\phi) &= (2\pi\sigma_2^2)^{-\frac{1}{2}} \exp[-(y-\mu_2)^2/2\sigma_2^2] \end{aligned} \tag{A.2}$$

Then for complete data **x** sampled from the joint density $f(x|\phi)$ depending on parameter set ϕ, we can write the <u>complete data</u> likelihood as:

$$f(x|\phi) = \prod_{i=1}^{n} z_i^T \Big(g(y_i|z_i = (1,0),\phi), g(y_i|z_i = (0,1),\phi) \Big) \; z_i^T(p_1,p_2) \tag{A.3}$$

or the complete data log-likelihood as:

$$\begin{aligned} L &= \log f(x|\phi) \\ &= \sum_{i=1}^{n} z_i^T \Big(\log g(y_i|z_i = (1,0),\phi), \log g(y_i|z_i = (0,1),\phi) \Big) \\ &\quad + \sum_{i=1}^{n} z_i^T(\log p_1, \log p_2) \end{aligned} \tag{A.4}$$

For brevity denote

$$G(y_i,\phi) = \Big(\log g(y_i|z_i = (1,0),\phi), \log g(y_i|z_i = (0,1),\phi) \Big) \tag{A.5}$$

so that

$$L = \sum_{i=1}^{n} \Big(z_i^T \, G(y_i,\phi) + z_i^T(\log p_1, \log p_2) \Big) \tag{A.6}$$

The EM algorithm comprises two steps, an E-step or expectation step, and an M-step or maximization step. The algorithm requires an initial choice of parameters say $\phi^{(0)}$, and the MLE's are then determined by cycling between E- and M-steps until some convergence criteria are met. Suppose that at iteration v we have a parameter set $\phi = \phi^{(v)}$, then in its most general form, the next iteration of the EM algorithm is defined as follows:

E-step: Compute $Q(\phi|\phi^{(v)}) = E(\log f(x|\phi)\ y,\phi^{(v)})$

M-step: Choose $\phi^{(v+1)}$ to be a value of ϕ which maximizes $Q(\phi|\phi^{(v)})$

$Q(\phi|\phi^{(v)})$ is the current expectation of the complete data log likelihood given observations y and the current parameters $\phi^{(v)}$. So rather than maximize the complete data log-likelihood, which is not known, we are instead maximizing the current conditional expectation of the complete data log-likelihood.

For the case of a mixture of two normals:

$$\begin{aligned} Q(\phi|\phi^{(v)}) &= E(\log f(x|\phi)|y,\phi^{(v)}) \\ &= E\left(\sum_{i=1}^{n} \log f(x_i|\phi)|y_i,\phi^{(v)}\right) \\ &= \sum_{i=1}^{n} E(\log f(x_i|\phi)|y_i,\phi^{(v)}) \end{aligned} \tag{A.7}$$

But $x_i = (y_i,z_i)$ so that given parameter set $\phi^{(v)}$:

$$\begin{aligned} P(x_i|y_i,\phi^{(v)}) &= P(y_i,z_i|y_i,\phi^{(v)}) \\ &= \frac{P(y_i|z_i,\phi^{(v)})\ P(z_i|\phi^{(v)})}{\sum_{\text{all } z_i} P(y_i|z_i,\phi^{(v)})\ P(z_i|\phi^{(v)})} \\ &= \frac{g(y_i|z_i,\phi^{(v)})\ P(z_i|\phi^{(v)})}{\sum_{\text{all } z_i} g(y_i|z_i,\phi^{(v)})\ P(z_i|\phi^{(v)})} \end{aligned} \tag{A.8}$$

Substituting from Equation A.7 and A.8:

$$Q(\phi|\phi^{(v)}) =$$

$$\sum_{i=1}^{n}\left[\sum_{\text{all } z_i}\left(z_i^T\, G(y_i,\phi) + z_i^T(\log p_1,\log p_2)\right) * \frac{g(y_i|z_i,\phi^{(v)})\; P(z_i|\phi^{(v)})}{\sum_i g(y_i|z_i,\phi^{(v)})\; P(z_i|\phi^{(v)})}\right] \qquad \text{(A.9)}$$

The M-step now consists of finding $\phi = \phi^{(v+1)}$ to maximize $Q(\phi|\phi^{(v)})$. Taking partial derivative of $Q(\phi|\phi^{(v)})$ with respect to the components of $\phi = (p_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and setting them to zero:

$$\frac{\partial Q}{\partial \mu_j} = \sum_{i=1}^{n}\left(\frac{g(y_i|r_j,\phi^{(v)})\; P(r_j|\phi^{(v)})}{\sum_j g(y_i|r_j,\phi^{(v)})\; P(r_j|\phi^{(v)})}\left(\frac{y_i - \mu_j}{\sigma_j^2}\right)\right) \qquad \text{(A.10)}$$

where $r_j = (1,0)$ for $j = 1$
and $r_j = (0,1)$ for $j = 2$

For brevity put

$$g(y_i|(1,0),\phi^{(v)}) = f_1(y_i|\phi^{(v)})$$

$$g(y_i|(0,1),\phi^{(v)}) = f_2(y_i|\phi^{(v)})$$

and define weights at iteration v

$$w_{ij}^{(v)} = \frac{p_j^{(v)} \; f_j(y_i|\phi^{(v)})}{\sum_{j=1}^{2} p_j^{(v)} \; f_j(y_i|\phi^{(v)})} \qquad ;\begin{matrix} j=1,2 \\ i=1,n \end{matrix} \tag{A.11}$$

Then

$$\frac{\partial Q}{\partial \mu_j} = \sum_{i=1}^{n} w_{ij}^{(v)} \left(\frac{y_i - \mu_j}{\sigma_j^2} \right) \qquad ;j=1,2 \tag{A.12}$$

Choosing $\mu_j = \mu_j^{(v+1)}$ to set $\partial Q/\partial \mu_j$ to zero gives

$$\mu_j^{(v+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(v)} y_i}{\sum_{i=1}^{n} w_{ij}^{(v)}} \qquad ;j=1,2 \tag{A.13}$$

Similarly for the

$$\frac{\partial Q}{\partial \sigma_j} = \sum_{i=1}^{n} w_{ij}^{(v)} \frac{1}{\sigma_j} \left(\left(\frac{y_i - \mu_j^{(v)}}{\sigma_j} \right)^2 - 1 \right) \qquad ;j=1,2 \tag{A.14}$$

which gives

$$\sigma_j^{(v+1)} = \left[\frac{\sum_{i=1}^{n} w_{ij}^{(v)} \; (y_i - \mu_j^{(v)})^2}{\sum_{i=1}^{n} w_{ij}^{(v)}} \right]^{\frac{1}{2}} \qquad ;j=1,2 \tag{A.15}$$

For the p_j, recalling that $p_2 = 1 - p_1$ and noting that $w_{i2} = 1 - w_{i1}$:

$$\frac{\partial Q}{\partial p_1} = \sum_{i=1}^{n} \left(\frac{w_{i1}^{(v)}}{p_1} - \frac{(1-w_{i1}^{(v)})}{1-p_1} \right) \qquad \text{(A.16)}$$

For $\partial Q/\partial p_1 = 0$ then

$$\sum_{i=1}^{n} w_{i1}^{(v)} \; (1-p_1^{(v+1)}) = p_1 \sum_{i=1}^{n} \; (1-w_{i1}^{(v)})$$

$$p_1^{(v+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{i1}^{(v)} \qquad \text{(A.17)}$$

Equations A.11, A.13, A.15 and A.17 together with an initial estimate of parameters $\phi^{(0)}$ provide an iterative solution to the maximum likelihood equations. Note that w_{ij} is the posterior probability that given y_i, observation i is from state j. For further details and proofs associated with the EM algorithm, the reader is referred to Dempster, Laird and Rubin (1977).