IWRRC Technical Reports
Indiana Water Resources Research Center

8-1-1975

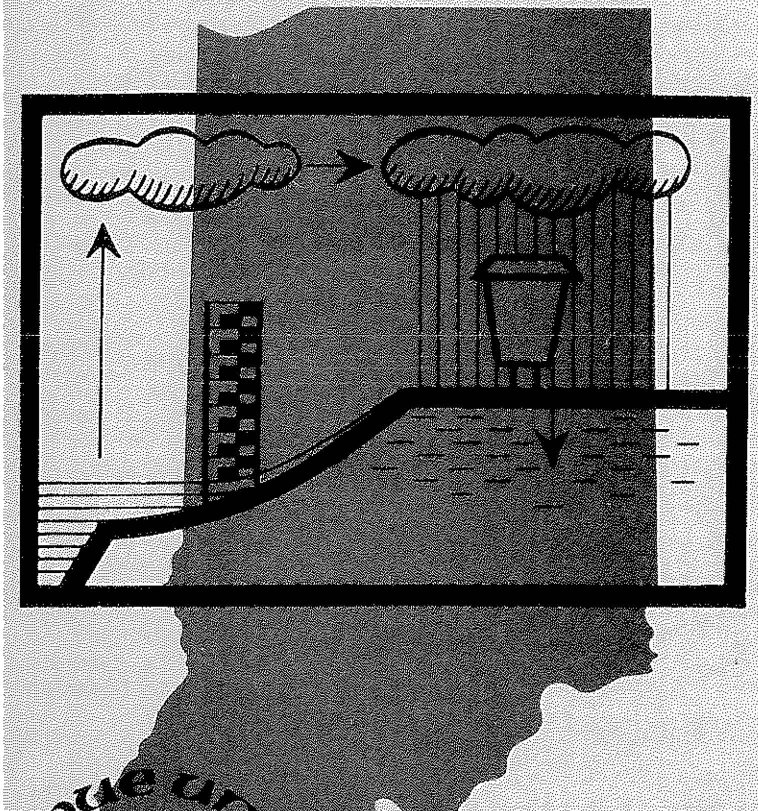
The Stochastic and Chronologic Structure Of Rainfall Sequences--Application To Indiana

M. L. Kavvas

J. W. Delleur

TECHNICAL REPORT NO. 57

THE STOCHASTIC AND CHRONOLOGIC STRUCTURE OF RAINFALL SEQUENCES - APPLICATION TO INDIANA



by

M. Levent Kavvas

Jacques W. Delleur

August 1975

PURDUE UNIVERSITY
WATER RESOURCES RESEARCH CENTER
WEST LAFAYETTE, INDIANA

Water Resources Research Center
Purdue University
West Lafayette, Indiana

THE STOCHASTIC AND CHRONOLOGIC STRUCTURE OF RAINFALL SEQUENCES - APPLICATION TO INDIANA

BY

M. Levent Kavvas and Jacques W. Delleur

The work upon which this report is based
was supported in part by funds provided by
the United States Department of the Interior,
Office of Water Research and Technology, as
authorized by the Water Resources Research Act
of 1964 (PL 88-379 as amended).

Period of Investigation: September 1971 - August 1975
Final Report for OWRT-B-036-IND
Matching Fund Agreement 14-31-0001-3585

Purdue University Water Resources Research Center
Technical Report No. 57
August 1975

ACKNOWLEDGMENTS

This research was supported in part by the Office of Water Research and Technology under matching fund grant OWRR-B-036-IND, in part by the Victor M. O'Shaughnessy Scholarship Fund and in part by Purdue University.

The authors wish to express their appreciation to Dr. L. Cote of Statistics Department at Purdue University for his valuable advice, to Drs. Dan Wiersma, Director of the Water Resources Research Center and J. F. McLaughlin, Head of the School of Civil Engineering for their assistance in the administration of the project.

ABSTRACT

PART I

This part is concerned with the point stochastic analysis of the daily rainfall occurrences in Indiana. The point statistical analysis was performed utilizing some statistical functions and some statistical tests of hypotheses. First, the analysis of trends in the daily rainfall counting process was performed. It was seen that there are cyclicities both in the first and the second moments of the point stochastic process. Physically meaningful annual and 15-day cycles were found to be significant. There is also a slight downward trend in the rate of daily rainfall occurrence in Indiana. The data were homogenized under the independent counting increments assumption and the Poisson model was tested by formal statistical tests and by statistical functions. The model was rejected for the daily rainfall counting process in Indiana. From the behavior of the spectrum and the variance-time function of the daily rainfall counts a clustering of the daily rainfalls in terms of storms is apparent. The Neyman-Scott cluster process was constructed in the time dimension to model this physical persistence. Physical concepts were attached to various components of the model. It is shown that the model fits the data quite well. Due to its very flexible spectral structure and to its physical interpretation of the various components of the rainfall occurrence, the Neyman-Scott cluster model deserves further investigation in different climates of the world.

PART II

The second part of this report is concerned with the time series analysis of the monthly and the annual rainfall sequences at various stations in the Midwestern United States.

First, a theoretical and empirical analysis of the removal of cyclicities was performed on the monthly rainfall series. It was seen that differencing cannot be used for the generation purposes although it removes the cyclicities in the data. Standardization, although it introduces some spurious nonstationarities into the data, is an acceptable method for the generation purposes.

A spectral and a variance-time analysis of the ARIMA family of the hydrologic time series models was done to study their long range dependence characteristics. It was seen that when the models are in the ARMA(p,q) family, they asymptotically end up in the Brownian domain. Therefore, in the strict mathematical sense, they cannot preserve the long range dependence characteristics in the form of Hurst's law for the variance. The ARIMA (1,d,1) family of models, on the other hand, yield infinite variance and are nonstationary in their generating forms.

The nonseasonal ARMA models were applied to either differenced or standardized monthly rainfall square roots. The ARMA(1,1) model, fitted to the standardized monthly square roots, emerged as the best model in terms of the statistical diagnostic tests.

The seasonal multiplicative ARIMA $(1,0,0) \times (1,1,1)_{12}$ passed all the goodness of fit tests on the seasonally differenced monthly rainfall square roots.

Forecasting of the monthly rainfall square roots was carried out by the use of ARIMA (0,0,0), ARIMA (1,0,1), ARIMA $(1,1,1)_{12}$ and ARIMA $(1,0,0) \times (1,1,1)_{12}$ models. Among these models the ARIMA (1,0,1) or the

ARIMA (0,0,0) models can preserve both the monthly means and the monthly standard deviations. On the other hand, the seasonal models could only preserve the monthly means. Therefore, they are inconvenient for hydrologic forecasting purposes.

For the annual rainfall series a white noise model with normally distributed residuals is adequate and can be used for the hydrologic simulation purposes in the Midwestern United States.

TABLE OF CONTENTS

Page

LIST OF TABLES

PART I

PART II

LIST OF ILLUSTRATIONS

Diagram Page

Map

LIST OF NOTATIONS

PART I

Latin Symbols

Symbol	Meaning
$C(x)$	Coefficient of variation of the interarrival times x
$C(\tau)$	Coefficient of variation function
D_ℓ	Two-sided Kolmogorov-Smirnov statistic
D_ℓ^+, D_ℓ^-	One-sided Kolmogorov-Smirnov statistics
$E[N_t^0]$	Renewal function of the ordinary renewal process $\{N_t^0\}$
$f_x(s)$	Laplace transform of $f_x(x)$
$f_1(\omega)$	Spectrum of the rainfall interarrival times
$\hat{f}(\omega)$	Estimate of $f(\omega)$
$g(\omega)$	Spectrum of counts
$g_+(\omega)$	Spectrum of counts on positive frequencies
$\tilde{g}_+(\omega)$	Periodogram estimate of $g_+(\omega)$
$g_{N+}(\omega)$	Normalized periodogram estimate of $g_+(\omega)$
$G_p(z)$	Probability generating function of the rainfall generating mechanisms
$G_\nu(z\|u)$	Probability generating function of the number of rainfalls in a storm that has its origin at u
$G_{N_{t_1}}(z)$	The probability generating function of the number of rainfalls in the time interval $(0,t_1)$
$H_T(\omega)$	Fourier-Stieltjes integral of N_t in the interval $(0,T)$
I_J, $I(\omega_J)$	Periodogram estimate of a time series
ℓ	Cutoff lag of the autocorrelation function in the spectral density estimation
ℓ_n	Moran's statistic
$L\{z\}$	Laplace transform of z
$L_t^{-1}(s)$	Inverse Laplace transform of s
$m(t)$	Mean rate of daily rainfall counts
$m_f(\tau)$	Intensity function
$\overline{m}_f(\tau)$	Unbiased estimate for the intensity function of the daily rainfall counting process
$M(t)$	Mean function for the counting process $N(t)$

$M_f(t)$	Expectation of N_t^f under stationarity
N_n	Number of rainy days in n-day period
N_t	Number of rainfall occurrences in (0,6)
N_t^f	Number of rainfall occurrences in the interval (0,t) which starts with an occurrence at 0 but does not include it
N_{1J}	The contribution of rainfalls to the interval $(0,t_1)$ from any prior interval I_J
$\{\Delta N_t\}$	The differential counting process
$P_1(u)$	The probability that a rainfall from a storm that has its origin at u will occur in $(0,t_1)$
$P_{iJ}^{(n)}$	Transition probability from i-th to J-th state in n-th stage
R^2	Explained variance
$S(n)$	n-day precipitation amount
t_i	Observed time to the i-th event
T	Forward recurrence time
T_r	Return period
$T(u)$	First passage time of the precipitation amount u
U	Random variable for the position of the storm origin
$U_{(i)}$	Normalized cumulative periodogram
$V(t)$	Variance-time function
$\overline{V}(J\delta)$	Estimate of the variance-time function
w_ℓ^2	Cramer-Von Mises statistic
W_ℓ^2	Anderson-Darling statistic
X	Interarrival time
$\{X_i\}$	The process of interarrival times
X_ν	Precipitation amount in ν rainy days
X_i, Y_i	Indicator random variables to show whether the i-th rainfall in a storm whose origin is at u falls into $(0,t_1)$ and $(0,t_2)$ respectively
$z(x)$	Harard function
Z	Number of rainfalls in a thunderstorm in the compound Poisson model

Greek Symbols

Symbol	Meaning
α, h_0	Rate of occurrence of rainfall generating mechanisms
$\gamma_+(u)$	The covariance density of the differential counting process
Γ_t	The counting random variable of the rainfall generating mechanisms
Δ	Transition matrix of first-order Markov chain
η_n	Shows whether n-th day is dry or wet

λ	Rate of occurrence
λ_{i_J}	Transition rate of the process from state i to state J
$\lambda(t)$	Rate of occurrence of a time-dependent process
$\lambda_\tau(t)$	Estimate of the mean rate of daily rainfall counts
Λ	Transition rate matrix
ν	Number of rainfalls in a storm
$\nu(u)$	Number of rainfalls in a storm that has its generating mechanism at the time position u
$\hat{\rho}_J$	Estimate of the autocorrelation function of the intervals at lag J
τ	Homogeneous time scale
T	Time position of a rainfall in a storm
$\psi(z,t)$	Probability generating function

PART II

Latin Symbols

Symbol	Meaning
ARMA(p,q)	p-th order autoregressive, q-th order moving average mixed model
ARIMA(p,d,q)	p-th order autoregressive, q-th order moving average integrated mixed model with d-th difference
ARIMA(P,1,Q) x (p,d,1)	Seasonal autoregressive-moving average model for the one-lag seasonally differenced and nonseasonally d-th differenced time series
$A(\tau_1,\tau_2)$	Autocorrelation function of the hydrologic time series
$\{a_t\}$	Residual series
$C(\bar{\phi},\bar{\theta})$	Covariance matrix of the least squares estimates of the ARIMA (p,d,q) model
$c(f_J)$	Estimate of the spectral distribution at frequency f_J
$dz(\omega)$	Stationary, uncorrelated, complex random increments with $E[z(\omega)] = 0$
$e_t(\ell)$	Forecast error for the lead time
$f(\omega)$	Absolutely continuous component of the spectral density function
$f_{ARMA}(\omega)$	Spectral density function of the ARMA(p,q) process
$f_c(\omega)$	Spectral density function of the generation scheme
$f_\Delta(\omega)$	Spectral density function of the nonseasonally one-lag differenced hydrologic time series
$f_{\Delta 12}(\omega)$	Spectral density function of the seasonally 12-lag differenced monthly hydrologic time series
h_k	Impulsive response function
$H(\omega)$	Fourier transform of the impulsive response function
$L(\bar{\phi},\bar{\theta},\sigma_a)$	Log likelihood function
ℓ_α	The magnitude of the α-th discontinuity in the spectral distribution function
Q_α	Random variables mutually uncorrelated for $\alpha = 0, 1, \ldots, r$

$R_i^{(k)}$	Covariance function of the k-th differenced time series at the lag i
R_J	Covariance of the stationary random component of the one-lag differenced hydrologic series
R_J^{12}	Covariance of the seasonally 12-lag differenced monthly hydrologic time series at the J-th lag
r	Half period of the circularly stationary time series of period length 2r+1
r_J	Covariance of the stationary random component of the hydrologic series at J-th lag
$S(\omega)$	Spectrum of the stationary time series
$S^{(k)}(\omega)$	Spectrum of the k-th differenced series
$S(\overline{\phi},\overline{\theta})$	Unconditional sum of squares
$s_{Y,J}$	Standard deviation of the square root transformed rainfall series for the month J
$V(\tau)$	Variance-time function at the time lag τ
$W(t)$	Brownian motion
$\{X_J\}$	Periodic hydrologic time series
$\{Y_J\}$	Second order stationary hydrologic time series
$\overline{y}_J$	Mean of the square root transformed rainfall series for the month J
$\hat{Y}_t(\ell)$	Forecasted, square root transformed, periodic monthly rainfall series
$\{z_t\}$	Standardized series
$\hat{z}_t(\ell)$	Forecasted, square root transformed, standardized monthly rainfall series

Greek Symbols

Symbol	Meaning
γ_k	Covariance function for the seasonal ARIMA model at the k-th lag
ΣX_J	Nonseasonally one-lag differenced monthly hydrologic time series
$\Delta_{12}X_J$	Seasonally 12-lag differenced monthly hydrologic time series
$\Delta^{(k)}X_J$	k-th difference of $\{X_J\}$
δ	Sampling time interval
$\Theta_Q(B^{12})$	Seasonal MA operator of the ARIMA (P,1,Q) x (p,d,q) model
$\theta(B)$	MA operator of the ARMA (p,q) model
$\lambda(\omega)$	Spectrum of the stationary random component of the original monthly hydrologic time series $\{X_J\}$
ρ_J	Autocorrelation function at the J-th lag
$\hat{\rho}_k$	Autocorrelation function estimate at the k-th lag
$\hat{\sigma}_a^2$	Residual variance estimate
$\hat{\sigma}_{Y,t}(\ell)$	Standard error of the forecast $\hat{Y}_t(\ell)$
$\hat{\sigma}_{z,t}(\ell)$	Standard error of $\hat{z}_t(\ell)$
Φ_α	Random phases mutually uncorrelated for $\alpha = 0, 1, \ldots, r$

$\Phi_p(B^{12})$	Seasonal AR operator of the ARIMA (P,1,Q) x (p,d,q) model
$\phi(B)$	AR operator of the ARMA (p,q) model
$\hat{\phi}_{\ell\ell}$	Partial autocorrelation estimate at the lag ℓ
χ	Chi-square statistic
ψ_J	Weights in the calculation of the standard error of the forecast
ω	Frequency
ω_α	Discrete frequency equal to $\alpha/(2r+1)$

GENERAL INTRODUCTION TO PARTS I AND II

THE PLACE OF THE REPORT IN THE FRAMEWORK OF WATER RESOURCES DEVELOPMENTS

The development of the water resources of a certain region of the world requires a number of analytical planning methodologies. These include the construction of models for the evaluation of the water resources systems on the basis of the achievements of certain stated goals under various environmental constraints. Once the objectives and the constraints of the development are stated, the model analysis leads to a set of hydraulic structures (such as dams, irrigation canals, flood zones, power plants, groundwater pumps, aqueducts), with their levels of output for the given available water. The analysis also leads to operational procedures for attaining the outputs which best fulfill the stated objectives.

Two types of techniques employed for model analysis are (1) the simulation approach, and (2) the mathematical programming approach. For the multipurpose, multistructure water resources system the interactions among the system components usually are very complex and may require not only numerical computations from algebraic expressions, but also decision rules based on logical relationships among the design variables (hydraulic structures, and their target outputs for various water needs). Furthermore, the employment of random numbers may be necessary in the description of the interactions among the design variables. This complexity is not amenable to a rigorous solution by straightforward mathematics. Therefore, the simulation approach is often preferred. The simulation of a complex water resources system makes use of mathematical representations and logical expressions which represent the various complex physical interrelations among the system components. This simulation enables the planner to obtain an optimal or near optimal combination of hydraulic structures, water target levels (for irrigation, water supply, energy and various other outputs), storage capacities for various uses, and an operation policy for the management of water through the system.

A simulation model is made up of two parts; (1) the system components, and (2) the operation. The operations constitute the model's representation of the internal interactions existing among the system components. The rules defining these interactions are naturally functions of the system components. These rules form the operation policy of the water resources system. The system components are of two types; (1) the state variables that describe the state of the system at any time, (2) the physical functions and constants which do not vary in time. The physical functions for a water resources system are, for example, the flood routing equations, the head-capacity curves for hydropower, the head-storage curves, and various other functions of the time-invariant relationships among the elements of the system. A certain combination of the state variables describes a particular water resources system that will yield a specific response to specified inflows to the system.

Hufschmidt and Fiering (1966) further classify the state variables of a water resources system as (1) the physical facilities - the sizes of the various hydraulic structures in service, (2) the operation policy parameters - the rules for the management of water through the system such as the storage allocations for flood control, dead storage and rules for releasing and routing the water. The state variables of the simulation model are the design variables of the water resources system. All the components of this physical system have their limits, such as the reservoir storage allocations and aqueduct capacities. These limits are induced by the availability of the water, the demand for water services and by the development measures for adjusting the supply to the demand.

The role of hydrology in the water resources system analysis through simulation is (1) to provide the inflows to the system once the design variables of the system are set for a simulation run, and (2) to provide those physical limits for the design variables, which are induced by the extreme natural conditions and by the risk levels the community is willing to take. The inflows are routed through the water system which is already bounded by the physical constraints, and a response in terms of the physical outputs and economical and social benefits is obtained. For each pattern of inflow the combinations of design variables can be varied, and a response surface can be obtained after many trials of simulation runs. From the response surface an optimal design in terms of the stated objectives can be identified as the combination of those design variables which best fulfill the objectives within the physical limits of the design variables.

The first part of this report addresses itself to the second role of hydrology in the water resources planning: the extreme hydrologic phenomena. Floods and draughts determine the critical periods in the operation of a water resources system. They are of vital importance in fixing the physical limits of the hydraulic structures, the limits of the target outputs for water demands, and the limits of the operation policy for the predetermined risk levels. A drought of long duration would necessitate very large reservoir capacities to meet the irrigation, municipal and industrial demands of a community which desires a small risk. The problem is of vital importance in the dry climates where lives and many economical activities depend basically on water. The countries in the Middle East are examples of this situation. An extreme flood may necessitate very high dams, large empty reservoir spaces, and expensive flood zoning measures. Therefore, the need for the precise calculation of the stochastic structure of these extreme phenomena becomes apparent.

The rainfall occurrences in a certain region decide the drought and the flood characteristics of that region. Therefore, the rainfall occurrence phenomenon has to be properly modeled. The rainfall occurrence is the end result of some complex processes in the atmosphere. These complex processes may be modeled to a certain extent by mathematical expressions. However, the internal relationships among the very diverse atmospheric components are too complicated to be solely described in mathematical terms. The knowledge about the behavior of the atmospheric components towards the production of rainfall is insufficient. This uncertainty and the above described complexity can most simply be modeled by considering the rainfall occurrences as a stochastic process.

The rainfall record in the particular region under study is just one realization in the stochastic process of the rainfall occurrences. The floods and the droughts that are expected to occur in the history of the rainfall time series may not be represented in the brief record at hand. The physical limits of the design variables that are calculated through this single realization are obviously not representative and may lead to catastrophes. It is necessary to construct a model for the point stochastic process of the rainfall occurrences in order to obtain the probabilities of practical importance in fixing the physical constraints for the design of the system. Once these probabilities are known, the planner can determine the risk measures corresponding to the dimensions of the hydraulic structures, the target water outputs, and the operation policy parameters.

The classical water resources system simulation studies such as the Harvard Water Program *Maass et al.*, (1962) used synthetic streamflows as the hydrologic inflow to the system. The historical streamflow is the hydrologic inflow to the system. But similar to the rainfall record the historical streamflow record is just one realization in the stochastic process of the streamflow time series. Just one realization obviously cannot represent all of the physical characteristics of the streamflows. The critical periods of floods and droughts that are expected to occur in the history of the streamflow time series may not be represented in the brief record at hand. The optimal design, derived through the water resources system response to a single realization of a stochastic process, can only be good for this single realization which may never repeat itself. No information can be derived for the response to other equally likely realizations, that is, to equally likely streamflow patterns which have the same statistical characteristics as those of the historical streamflow sequence. It is the job of the hydrologist to generate synthetic streamflows which preserve the probability structure of the historical streamflow sequence and which create the critical flood and drought conditions to be expected from a sufficiently long hydrologic record. These equally likely sets of hydrologic sequences will enable the planner to analyze his system performance under a great variety of conditions so that he can construct a more thorough response surface for the design variable combinations. The more thorough the response surface, the more reliable the optimal design will be. For the accomplishment of this objective the hydrologists have constructed stochastic models which preserved the mean, the variance and the autocorrelation structure of the historical streamflow record and have generated streamflow time series 500 years or more in length. The justification of preserving the mean and the variance was that the range of the cumulative departures from the mean, which in turn specify the design reservoir capacities, could best be estimated in terms of these two statistics (*Hufschmidt and Fiering*, 1966).

Since the time series model, to be employed for the generation of synthetic hydrologic sequences, is fitted to the autocorrelation function, or, equivalently, to the spectrum of the stationarized historical hydrologic record, the first practical problem is the removal of the time trends and circularities from the historical data. In the first section of the second part of this report various operations for the removal of circularities are analyzed to assess their properties. The time series model is then fitted to the stationary part of the historical data. The stationarity and the invertibility conditions of the ARMA(p,q) family of time series models are studied and the physical meanings of these conditions are attached.

A very important concept to be considered is the long range dependence since the hydrologist generates sequences of the time span of 500 years or more. The classical studies of *Hurst* (1951, 1956, 1965) pointed to the long range dependence in the hydrologic records. Long-range dependence is quite an important concept for the water resources system design since it affects the storage capacity requirements. The long-range dependence corresponds to the low frequency components of the spectrum. It can also be analyzed through the variance-time curve. In the first section of part II the long-range dependence properties of the current hydrologic time series models will be analyzed by the use of their spectral and variance-time properties to produce the necessary conditions to simulate the long-range dependence effect. The fit to the spectrum or to the autocorrelation function does not guarantee a good fit at the low frequencies, nor does it guarantee the preservation of the various spectral moments. Therefore, the fit to the spectrum does not guarantee the preservation of the long-range dependence properties of the hydrologic record. A study is needed to incorporate these properties into the models.

Many parts of the world do not have sufficiently long historical streamflow records to be used for the synthetic streamflow generation. In the underdeveloped regions of the world the historical streamflow records are often non-existent. In such cases the hydrologist will have to generate synthetic rainfall sequences from the historical rainfall records that often exist and are quite long. Then, by either using the convolution techniques or by the physical watershed models he can obtain the synthetic streamflow sequences from the synthetic rainfall sequences. Furthermore, the development of the agricultural resources of a region requires the prediction of the dry and wet weather sequences in order to determine the irrigation policies. The time series models and the forecasting procedures for the annual and the monthly rainfall time series will be developed in the second section of the second part of this report.

PART I

POINT STOCHASTIC ANALYSIS OF DAILY RAINFALLS IN INDIANA

CHAPTER 1 - INTRODUCTION

1.1 A SURVEY OF THE STOCHASTIC MODELS ON THE RAINFALL OCCURRENCE PHENOMENON

Hydrologists and meteorologists have been fitting stochastic models to the rainfall occurrence data since the early 1900s. The purpose of these stochastic models is the simulation of the dry and wet sequence probabilities. The knowledge about the probabilities of the dry and wet period lengths, and of the number of rainfall occurrences renders better decisions on the agricultural policies and the water resources system operations. The stochastic models applied for the daily rainfall occurrences may be broadly classified into three groups; (1) the Poisson models for the independent rainfall counting increments, (2) the Markov chain models which account for the persistence in the rainfall counting increments, (3) the alternating renewal models for the independent, alternating wet and dry periods.

The application of Poisson models for the rainfall occurrences dates back to the work of *Grant* (1938). He fitted a simple Poisson model for the number of excessive rainfalls occurring in any single year in the Midwest. The simple Poisson model for the number of occurrences N_t in the time interval (o,t) is given as

$$P[N_t = n] = (\lambda t)^n e^{-\lambda t}/n!$$

with the rate of occurrence λ being the sole parameter. *Thom* (1959) agreed with the results of *Grant* when he fitted a simple Poisson model to the number of excessive rainfalls in a year in Davenport, Minneapolis, and Omaha, which are again Midwestern cities. *Shane* (1964) applied the compound Poisson distribution to the analysis of the rainfall records in the U. S. *Lobert* (1967) used Poisson distributed daily rainfall occurrences in the wet periods of his alternating renewal process. He applied his model to the Allier basin in France with quite satisfactory results. *Todorovic and Yevjevich* (1969) treated the daily rainfall occurrences with the non-homogeneous Poisson model which may be described as:

$$P[N_t = n] = \exp\left\{-\int_0^t \lambda(\tau)d\tau\right\} \cdot \left[\int_0^t \lambda(\tau)d\tau\right]^n \cdot \frac{1}{n!}$$

with the time-dependent rate of occurrence $\lambda(\tau)$ being its sole parameter. They fitted the probability distributions derived from the model to Durango and Ft. Collins, Colorado, Austin, Texas, and Ames, Iowa. They concluded that the number of storms in a time interval is Poisson distributed if the storms are properly defined. Two definitions that were used for the storm were: (1) each rainy day is treated as an individual storm event, whether or not it is preceded or followed by a rainy or non-rainy day, and (2) each storm is identified as an uninterrupted sequence of rainy days. They observed the annual cyclicities in the rate of daily rainfall occurrences but failed to account for the longer physical periodicities or the long-term trends cited by *Mitchell* (1964).

Duckstein et al. (1972) fitted simple Poisson distribution for the number of rainfalls caused by the summer precipitation of the continental thunderstorm or of the local convective type in Arizona. These rainfalls were of short duration and high spatial variance. The basic assumption of the Poisson model is the independence of the counting increments. In order to fit a Poisson model to the rainfall counts there should be no dependence among the rainfall counts. This may be true for the rainfalls which are caused by thunderstorms whose life cycle is in the order of hours (*Petterssen*, 1969). During the summer season there is a low pressure system over Nevada, Arizona, Northern Mexico and Southern California. However, this is overlaid by a high-level anticyclone with strong subsidence and, therefore, the clouds are absent. Occasionally, this protection is removed and the summer showers occur. If these summer showers are of short duration, then the study of *Duckstein et al.* (1972) makes physical sense. However, the mechanism of a thunderstorm consists of a cluster of thunderclouds. When the downdraft spreads sufficiently far from the mother cloud, the upward motion ahead of the downdraft often results in the formation of a new cell. Thus the thunderstorm group will be replenished on its front while the old cells at the back dissipate. Due to the clustering effect, the life span of the whole group of thunderstorms will be much longer than the life of the individual thunderstorms (*Petterssen*, 1969). The presence of a "cluster" of thunderstorms could cause dependence in the rainfall counts. This may be the physical explanation of the Markov chain model of *Smith and Schreiber* (1973) for the thunderstorm rainfall occurrences in Arizona.

The dependence of the rainfall occurrences was shown as early as in 1916 by *Newnham* in the data of the British Isles. The persistence of the dry day sequences at San Francisco, given by *Jorgensen* (1949), and the persistence in the wet sequences all over the world, given by *Jennings* (1950), indicate the dependence in the rainfall occurrences. The simplest model to account for the dependence is the first-order Markov chain.

Gabriel and Neumann, in a sequence of papers (*Gabriel and Neumann*, 1957, *Gabriel*, 1959, and *Gabriel and Neumann*, 1962) formulated a homogeneous, first-order Markov chain for the daily rainfall occurrences in Tel-Aviv. Their basic assumption was that the probability of rainfall on any day depended only on whether the previous day was wet or dry. For a two-state first-order Markov chain $\{\eta_n ; n \geq 0\}$ the transition probabilities may be expressed as:

$$P[\eta_{n+1} = J \mid \eta_n = i] = p_{iJ}^{(n)} \qquad i,J = 0,1$$

That is, the two states that are considered here are 0 and 1, denoting, respectively, the dry and the wet states. In a homogeneous Markov chain

$$p_{iJ}^{(n)} = p_{iJ}^{(0)} = p_{iJ}$$

that is, the transition probabilities stay the same at each stage. Therefore, the transition matrix of a homogeneous two-state first-order Markov chain with 0,1 states becomes

$$\Delta = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

where $p_{00} + p_{01} = 1$, $p_{10} + p_{11} = 1$. Then there are only two parameters for the two-state Markov chain. They will be taken as $p_1 = p_{11} = 1 - p_{10}$, and $p_0 = p_{01} = 1 - p_{00}$. The n-step transition matrix $\Delta^{(n)}$ for a homogeneous two-state first-order Markov chain is equal to Δ^n so that the recurrence relation

$$p_{iJ}^{(n)} = p_{i1}^{(n-1)} \; p_{1J} + p_{i0}^{(n-1)} \; p_{0J}$$

follows. Using this relation and some combinatorics *Gabriel* (1959) derived the probability of exactly s wet days among n days following a wet day as

$$P[N_n = s \mid \eta_0 = 1] = p_1^s (1-p_0)^{n-s} \sum_{c=1}^{c_1} \binom{s}{a} \binom{n-s-1}{b-1} \left(\frac{1-p_1}{1-p_0}\right)^b \left(\frac{p_0}{p_1}\right)^a$$

$$\text{where } c_1 = \begin{cases} n + \tfrac{1}{2} - |2s - n + \tfrac{1}{2}| & \text{if } s < n \\ 0 & \text{if } s = n \end{cases}$$

and a and b are the least integers not smaller than $(\tfrac{1}{2})(c-1)$ and $(\tfrac{1}{2})c$, respectively. Gabriel also derived the probability of exactly s wet days following a dry day as

$$P[N_n=s|n_0=0] = p_1^s(1-p_0)^{n-s} \sum_{c=1}^{c_0} \binom{s-1}{b-1} \binom{n-s}{a} \left(\frac{1-p_1}{1-p_0}\right)^a \left(\frac{p_0}{p_1}\right)^b$$

where $c_0 = n + \tfrac{1}{2} - |2s - n - \tfrac{1}{2}|$ if $s > 0$

0 if $s = 0$

and a and b are defined as above. Therefore, the probability of s wet days in n days is

$$P[N_n=s] = P[N_n=s|n_0=1]\ P[n_0=1] + P[N_n=s|n_0=0]\ P[n_0=0]$$

The probability of a wet spell of length k is $(1-p_1)p_1^{k-1}$ and the probability of a dry spell of length m is $p_0(1-p_0)^{m-1}$. The lengths obey the geometric law. *Gabriel and Neumann* (1962) established that the Markov chain model agreed with the Tel-Aviv daily rainfall observations at the 5% significance level. By establishing that the proportions of wet days, given the previous day's weather, are independent of the weather two or more days earlier at the 5% significance level, they showed that first-order, homogeneous, two-state Markov chain is adequate for modeling the Tel-Aviv daily precipitation occurrences. Considering that Tel-Aviv is basically under the influence of the Mediterranean regime, the model makes physical sense.

Caskey (1963) fitted the first-order, two-state Markov chain to precipitation occurrences in Denver, Colorado. He observed the annual cyclicity within the year and employed different transition probabilities for each season. He considered four different Markov chain models for four different seasons. *Weiss* (1964) fitted the Markov chain model to the daily rainfall data in Kansas City, Fort Worth, Montreal, San Francisco and Moncton. He showed that the model fits the sequences of wet or dry days in records of various length and for climatically different areas. He observed the seasonal variations. However, the most important feature of his study is the testing of the presence of secular trends in the daily rainfall data. In table 5 of his paper the monthly precipitation probabilities P[Wet|Dry] for two 25-year periods for Kansas City and Fort Worth are given. Although *Weiss* concludes that "there is relative secular stability of the probabilities," the differences are large enough to warrant further research on long term trends. Actually *Mitchell* (1964) pointed to the physical periodicities of 11-year sunspot cycles, the biennial cycles and 80-90 year cycles in the meteorological time series. *Feyerherm and Bark* (1964) fitted higher order Markov chains to the precipitation data in the north central U. S. when the first-order Markov chain proved to be imperfect. *Feyerherm et al.* (1965) fitted a non-homogeneous first-order Markov chain to the daily rainfall occurrence data of 11 locations in Indiana. The transition probabilities were calculated for every week of the year. However, no goodness of fit test results were given. A persistence model makes physical sense for Indiana since the state is basically under the influence of the Atlantic cyclone regime. During the winter the zone that separates polar continental air from an intrusion of arctic air from the north passes through Indiana in the east-west direction. This cyclone belt persists during the winter and causes extensive precipitation. In the summer time the belt moves north to Canada (*Petterssen,* 1969). However, the disturbances of the polar front to the south cause summer showers and occasional thunderstorms. These thunderstorms are basically of two types; (a) scattered type air-mass thunderstorms of short duration, and (b) frontal thunderstorms in clusters whose life cycle is much longer than the one for the scattered type (*Newman,* 1975).

During the winter time due to the long memory of the rainfall producing cyclone belt, there is reason to believe that the probability of rainfall on any day may not only depend on the previous day but on the further past. *Wiser* (1965) stated, with slight modification, that

> there are ... several sets of data which have been reported, which are not described properly by the simple Markov chain model. Among these may be mentioned several of the results given by *Newnham* (1916) for the British Isles, sequences of dry days at San Francisco by *Jorgensen* (1949) ... and the results cited as fitting a higher order Markov chain given by *Feyerherm and Bark*

> (1964) in the midwestern states. *Green* (1965) stated that the fit of the simple Markov model was unsatisfactory for several of the cases cited by *Weiss* (1964). The consistency of the manner in which the discrepancies occur is indicative that there may be a more general probability model, of which the Markov chain model is a special case, which will describe in a more suitable way the behavior of wet and dry sequences.

In table 1 of his paper *Wiser* cited several locations where the deviations of the simple Markov model from the data are highly significant. *Wiser* proposed four modified Markov probability models. From his goodness of fit tests the Polya urn model emerges as the most successful in describing the contagious cases where the persistence extends over a prolonged period. The extension of the persistence corresponds to the overdispersion of the number of occurrences. This corresponds to the upward curvature of the log-survivor function of the rainfall occurrences and will be described in the later chapters. The curvature of the log-survivor function of the dry days in San Francisco (*Jorgensen,* 1949) is shown in the figure 1 of *Wiser.* The straight line predicted by the Markov model was clearly unsatisfactory. On the other hand, the Polya urn model successfully fitted the empirical log-survivor function. The Polya urn model consists of a single urn containing 'wet' and 'dry' balls. After a ball is randomly drawn and its state is noted, it is replaced with D other balls of the same state. *Wiser* introduced the modification that the number of draws with added balls will not exceed a specified number c. After c draws the urn acts as a Bernoulli urn with the number of 'wet' and 'dry' balls being constant in the following draws.

Feyerherm and Bark (1967) analyzed the sequences of wet and dry days in Indiana, Iowa and Kansas under the light of *Wiser's* (*Wiser*, 1965) and their earlier (*Feyerherm and Bark*, 1964) findings. They have concluded that a rainy spell is more likely to terminate after at least two wet days than after one wet day during early spring. They pointed to the inadequacy of the first-order Markov chain for simulating the probabilities of long sequences, "especially for prolonged dry spells when a different set of meteorological forces may be operative." In accordance with their point, at Iquique, in northern Chile, four years have passed without rain (*Petterssen,* 1969). It is not possible to simulate such a long dry period with the geometric memory of the simple Markov chain.

Romanof (1972) fitted non-homogeneous Markov chains of first and second order for a sequence of wind observations in Bucharest, Rumania. He compared these models with the independent increment (binomial) model. He concluded that the Markov models are far better than the independent increment model in the simulation of the occurrence probabilities. The second-order Markov chain led to an improvement of the results over the first-order chain. However, the first-order Markov model was considered satisfactory in most cases.

Todorovic and Woolhiser (1971) derived the distribution of the total n-day precipitation amount S(n) as

$$P[S(n) \leq x] = P[N_n = 0] + \sum_{\nu=1}^{n} P[X_\nu \leq x] \; P[N_n = \nu]$$

where N_n is the number of rainy days and X_ν is the total amount of precipitation in ν rainy days. However, they made some fundamental assumptions in their derivation. These assumptions were: (a) $\{N_n = 0\}$, $\{N_n = 1\}$, ..., $\{N_n = k\}$ represent a countable partition of the sample space, (b) if ξ_ν denotes the daily precipitation value of the ν-th rainy day, ξ_1, ξ_2, ..., ξ_n are independent, identically distributed random variables with finite mean and variance, (c) the information concerning which k days in the n-day period were wet and which (n-k) days were dry, does not contribute anything to the knowledge concerning the corresponding precipitation quantity X_k. Among the three assumptions the last one is doubtful since due to the cyclicities both in the rainfall quantities and occurrences, the position of the rainy day within the year will definitely tell something about the corresponding rainfall quantity. However, the above probability distribution was derived under the basic assumption of stationarity. As long as the n-day time interval is approximately stationary, assumption (c) seems satisfactory. *Todorovic and Woolhiser* gave the probability of the first passage time, T(u), to the precipitation amount u as

$$P[T(u) \leq n] = P[S(n) > u].$$

They fitted the first-order Markov chain to the daily rainfall occurrences in Austin, Texas for the month of May and obtained satisfactory results. Assuming that X_ν is gamma distributed, and the daily rainfall counts obey the first-order Markov chain, they derived the explicit distribution of S(n). They fitted this model to the n-day rainfall amount in Austin. However, the fit is unsatisfactory. Later, *Woolhiser et al.* (1972) regionalized the parameters for the n-day precipitation amount in eastern Colorado.

Smith and Schreiber (1973) tested the hypothesis of the sequential independence against a first-order Markov chain alternative for the daily rainfall sequences in the southwestern U. S. for the time interval from June to September. They showed that the Markov model described the thunderstorm rainfall activity better than the independent increment model of *Duckstein et al.* (1972) for the same region. They showed that the transition probabilities vary within the season as well as during the year. Their harmonic analysis indicated 4 day and 11 day periodicities. However, their analysis was in the time-series sense while a spectral analysis of the rainfall counts had to be undertaken. The cumulative distribution fits of the number of wet days per season by the Binomial and Markov models in figures 10 to 12 in their paper showed the inadequacy of these models in describing the counting phenomenon. *Smith and Schreiber* stated that they "do not wish to conclude that thunderstorm season daily rainfall occurs as a simple Markov chain. It has not been demonstrated, for example, that the order of dependency may not change within the season or that a second or higher-order chain may not be superior to a simple chain." They pointed to the variation of the meteorological conditions from year to year, and to the importance of the description of the annual variance.

Crovelli (1972) used the finite-state continuous-time Markov chain to model the precipitation process. Finite-state continuous-time Markov chain is defined as a process which at any time is in one of a finite set of mutually exclusive, collectively exhaustive states and whose time between transitions is random and dependent only on the currently occupied state. A continuous-time precipitation process may be considered as a continuous-time Markov chain model. There may be two states, state 1: wet or storm and state 2: dry. Let λ_{12} represent the transition rate of the process from wet to dry and λ_{21} represent the transition rate of the process from dry to wet. The transition rate matrix Λ where (*Crovelli*, 1972)

$$\Lambda = \begin{bmatrix} -\lambda_{12} & \lambda_{12} \\ \lambda_{21} & -\lambda_{21} \end{bmatrix}$$

describes the two-state continuous-time Markov chain. The transition rate λ_{i_J} of the process from state i to state J is defined as

$$\lambda_{i_J} = \lim_{\Delta t \to 0} \frac{P[X(t+\Delta t) = J | X(t) = i]}{\Delta t}$$

Following *Grace and Eagleson* (1966), *Crovelli* assumed that the storm duration is exponentially distributed with mean $1/\lambda_{12}$, while the time between storms is exponentially distributed with mean $1/\lambda_{21}$. If the time between storms and the storm duration are assumed independent, the two-state continuous-time Markov chain yields the alternating renewal process which was employed earlier by *Green* (1964), *Grace and Eagleson* (1966), and *Lobert* (1967).

Green (1964) proposed an alternating renewal process with exponential durations for the dry and wet spells for the rainfall occurrence data of Tel-Aviv. He showed that the model fitted the rainfall data for Tel-Aviv adequately and represented certain conditional probabilities better than the simple Markov chain of *Gabriel and Neumann* (1962). *Grace and Eagleson* (1966) first separated the time series of point rainfall observations into statistically independent events and then applied the alternating renewal process with Weibull distributed storm durations for the data of the northern U. S. However, this procedure artificially distorted the originally dependent point rainfall process. They applied the model to short time increment rainfall in the order of minutes and hours. However, even in the hourly scale there is a very strong diurnal cyclicity in the rainfall occurrences (*Mitchell*, 1964). This type of a model is quite inconvenient for the incorporation of the physical cyclicities since it is basically made up of two renewal processes (the wet and the dry periods) alternating with each other (*Cox*, 1962). The basic characteristic of a renewal process is that the collection of the random variables $\{T_1, T_2, \ldots\}$ are identically distributed. This means stationarity. An alternating renewal process can account for the periodicity only if it is made up of some J components following each other in cyclic order. However, this would tremendously complicate the general representation of the model, its estimation, and its application. *Lobert* (1967) applied the alternating renewal process with the exponential storm interarrival times and exponential storm durations, to daily rainfall data in France. In order to account for the annual cyclicity he separated his data into months. He then applied the alternating renewal process to each month.

The shortcomings of the previous work done on the stochastic modeling of the rainfall occurrence process may be stated as follows:

1. The constructed stochastic models are black box models. They do not contain the physical components of the rainfall phenomenon.
2. The work that has been done relies heavily on the circular stationarity for the calibration of the model parameters. The effect of the biennial and longer term cycles (*Mitchell*, 1964) was ignored in the practical applications. No rigorous approach for the analysis of the physically sound cyclicities in the rainfall counting process was attempted. A statistical methodology for the detection and the calibration of the cyclicities and trends in the rainfall counting process is needed.
3. Although the dependence in the rainfall occurrences was accounted by many authors, a rigorous identification of this structure through the correlogram or the spectrum of the rainfall counts was not undertaken. Hydrologists and meteorologists tried to justify the first-order Markov persistence by overfitting the data by higher order models. The work of *Wiser* (1965) should be mentioned as an exception. The log-survivor function was effectively used by *Wiser* for the identification of dependence. However, a methodology for the identification and the estimation of the dependence structure is needed.
4. The stochastic models reported in this survey only considered the occurrence of the "product" of a complex meteorologic process. The physical dependence characteristics, associated with the various types of rainfall generating mechanisms were not utilized. This was due to the fact that all the reported models were one-level models. The rainfall occurrence process is at least a two-level process. The occurrence of the rainfall generating mechanisms may be considered as the primary level process. The rainfall generating mechanisms are the fronts, the thunderstorm clouds, etc. In the second level there is the rainfall occurrence. The dependence of the rainfall occurrences is basically due to the dependence of the rainfall generating mechanisms over the concerned area. The development of a two-level stochastic model may utilize the meteorologic knowledge about the characteristics of the rainfall generating mechanisms for the simulation of the probabilities of the dry and wet sequences. Such a model could form a first bridge between the deterministic meteorologic facts and the characteristics of the random wet and dry sequences.
5. Except for the pioneering work of *Wiser* (1965), the long-term dependence that was recorded in the dry and wet sequences around the world was not dealt with. A flexible stochastic model that can account for the various lengths of dependence is needed. The wet period duration of 2 years observed at Cherrapunji, India (*Jennings*, 1954) and the dry period duration of 4 years observed at Iquique, Chile (*Pettersen*, 1969) cannot be explained by the geometric memory of the popular first-order Markov chain.

6. The stochastic models were verified by their goodness of fit to the marginal probability distributions obtained from the data. The goodness of fit to the correlation structure is equally important since a stochastic process is a collection of random variables.

1.2 THE FRAMEWORK OF THE POINT STOCHASTIC ANALYSIS OF THE DAILY RAINFALL

Part I will be concerned with the stochastic analysis of the daily rainfall occurrences. The methodology for the statistical analysis of the occurrences will be the point stochastic analysis developed by *Cox and Lewis* (1966), *Parzen* (1967), *Lewis et al.* (1969), *Vere-Jones* (1970) and others. For a point stochastic process the occurrences should be instantaneous points on the time axis. Since the rainfall is observed at equal sampling intervals, a convenient sampling interval has to be chosen. As the time sampling interval for the rainfall phenomenon is decreased, the rainfall occurrences eventually become points on the time axis so that the occurrences appear to be instantaneous. The minimum length of the time sampling interval is controlled by the computer storage capacity, and by the goodness of approximation of the stochastic point process which describes the probability structure of the number of events in certain intervals. Another criterion for the selection of the time interval is the preservation of the characteristics of the phenomenon the planner wants to consider. As the time interval is decreased, the approximation to the point process may improve but the number of intervals increases and the storage requirements increase. In order to accommodate the storage requirements short lengths of record are usually taken. The shorter the record the more difficult it is to observe the nonstationary characteristics of the hydrologic phenomenon. The preservation of the long term trend characteristics of the record may become impossible.

In this report a sampling interval of one day was utilized for the analysis of the rainfall occurrences. The daily sampling interval was large enough so that the nonstationarity effects could still be observed. This interval was short enough so that the daily rainfall occurrences could be considered as a stochastic point process where events occur singly or in small groups at the instants of time.

However, there is the question of whether the whole day interval should be considered as wet when there is rain on that day. The distribution of the rainfall during the day was not given in the data. Therefore, if there was any rain on a day above the depth of .01 in. the rainfall was assumed to occur in the middle of the day. This assumption rendered a closer approximation to a point process where the occurrences should be instantaneous. Under this assumption a wet period of n-days consists of n consecutive rainfall points placed at one-day intervals. If the whole day was considered wet when there was rainfall at any time during that day, this alternative would not yield instantaneous occurrences on the time axis, especially when there are n consecutive rainy days. This, in turn, would make the point stochastic analysis on the rainfall counts quite difficult.

1.3 THE DATA

The daily rainfall occurrences were analyzed for 17 stations in the state of Indiana. The names of the 17 analyzed stations and their corresponding identification numbers are given in Table 1-1. The locations of these rainfall stations are shown on the MAP-1 of Indiana. As is seen from the map, most of the stations were taken on the east-west air front that governs the weather of Indiana during the winter. This front is the main cause of persistence in the daily rainfall occurrences in Indiana.

The Indiana daily rainfall data were obtained from the U. S. Weather Bureau in the form of magnetic tapes. The record length was 10 years covering the period 1950 through 1959. Due to the computer capacity limitations only the first seven years of the data were analyzed.

The missing data points were filled from the Climatologic Data Publications of the U. S. Weather Bureau whenever it was possible. For the cases where no record could be found in the publications the missing daily rainfall values were fitted by multiple linear regression utilizing the data of the neighboring stations.

For further details on the data the reader is referred to Appendix A.

1.4 OBJECTIVES OF THE PART I

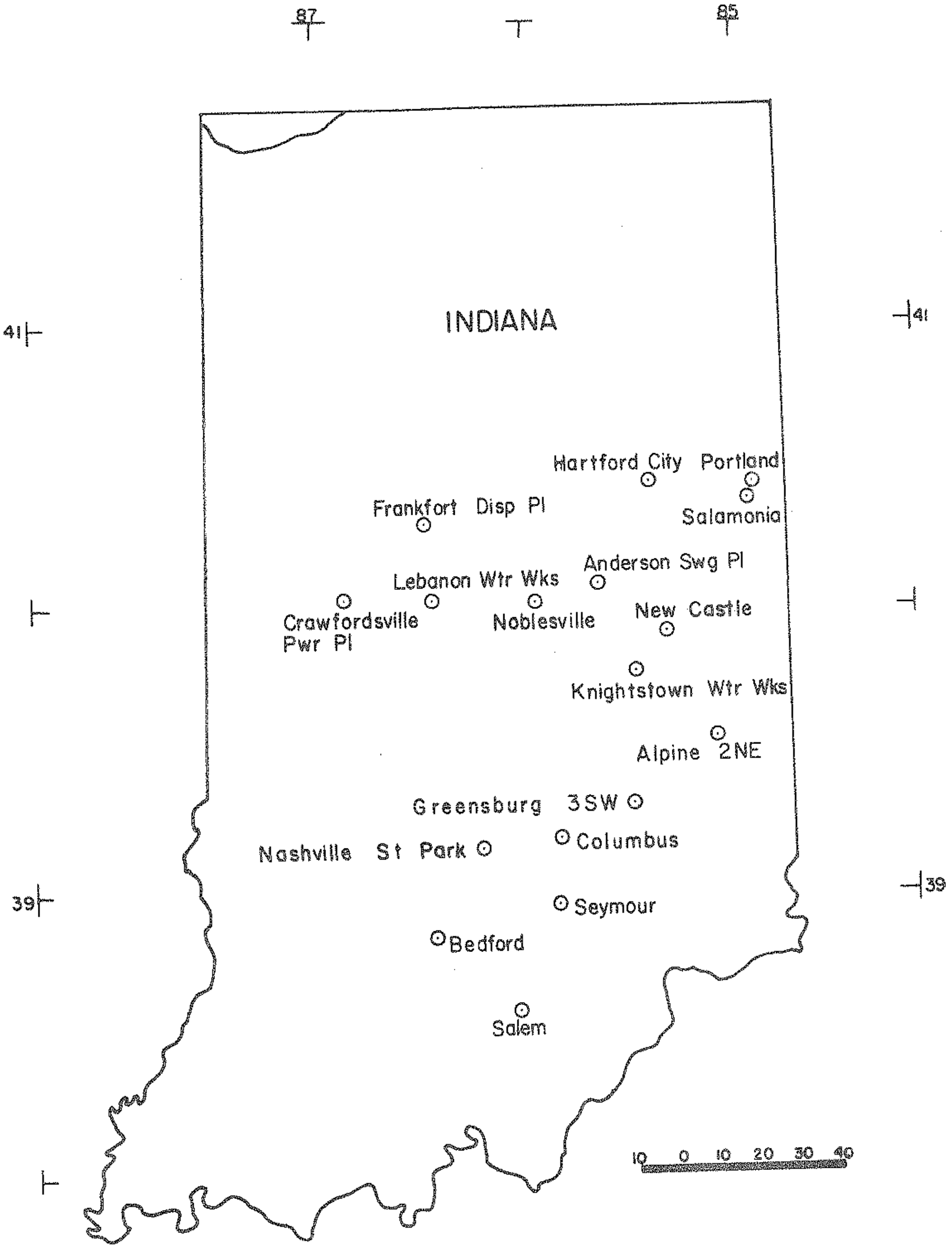
In the general introduction to parts I and II of this report and in section 1.2 the place of the stochastic analysis of daily rainfalls in the water resources developments, and the short comings of the current rainfall stochastic models are discussed. In the light of these discussions and the framework, the objectives of the first part of this report may be stated as follows:

1. To apply the point statistical analysis to the daily rainfall occurrence process in Indiana in order to
   a. detect the periodicities and trends in the daily rainfall occurrences,
   b. detect the dependence structure in the daily rainfall occurrences,
   c. select the proper point stochastic model for the daily rainfall occurrences.
2. To construct a point stochastic model for the daily rainfall occurrences in the light of the results of the point statistical analysis and based on physical assumptions.

TABLE 1-1

THE RAINFALL STATIONS IN INDIANA USED FOR THE ANALYSIS
OF THE DAILY RAINFALL OCCURRENCES
YEARS 1950-1959

STATION	IDENTIFICATION NUMBER
Alpine 2NE	0132
Anderson Quartz Plant	0177
Bedford	0545
Columbus	1747
Crawfordsville Power Plant	1882
Frankfort Disposal Plant	3082
Greensbury 3SW	3547
Hartford City	3777
Knightstown Water Works	4642
Lebanon Water Works	4908
Nashville State Park	6056
New Castle	6164
Noblesville	6338
Portland	7069
Salamonie	7747
Salem	7755
Seymour	7935



MAP I. RAINFALL STATIONS USED FOR ANALYSIS

CHAPTER 2 - ANALYSIS OF TRENDS IN THE DAILY RAINFALL OCCURRENCES

The analysis of trends in the hydrologic point processes gives the water resources system planner an indication of the long term and cyclic climatological changes through which he can decide on the long term and the seasonal strategies. The establishment of long term linear or curvilinear time trends would necessitate the construction of a nonstationary stochastic model and a homogenization procedure through which the model could be tested in the stationary domain. The presence of long term time trends could seriously affect the design procedures based on the probabilities derived from the stationarity assumption.

Two approaches will be utilized in the analysis of trends in the daily rainfall counts. The first will be a graphical method where the behavior of several statistical functions will be analyzed as functions of time. The second method will be a test of stationarity hypothesis of the interarrival times based on the assumption that the intervals are independent.

2.1 THE GRAPHICAL METHODS

2.1.1 Number of Rainy Days Versus Cumulative Time

The first graphical method to be considered is the cumulative plot of the total number of rainfall occurrences, called the "interval number," versus the total time in days to the last occurrence, denoted by "cumulative time." The daily rainfall counts in 17 stations in Indiana were examined for trends in the mean rate of occurrence of the daily rainfall using the described plots. Plots for the stations 0132, 3082, 3777, 4642, 6056 and 7747 are given in figure 1. The slope of the plot at any time is the inverse of the mean rate of daily rainfall occurrence at that time. These plots show that the mean rate of daily rainfall occurrence is decreasing with time, raising the possibilities that either there are long climatological cycles where the period 1950-59 from which the data was taken, is on the dipping portion of the cycle, or the climate is gradually drying. To answer this speculation an analysis of the behavior of the mean rate of occurrence for a longer time is needed.

2.1.2 The Mean Rate of Daily Rainfall Occurrence

The mean rate of daily rainfall counts, m(t), was estimated by the statistic $\lambda_\tau(t)$ where for the interval $(t, t+\tau)$,

$$\lambda_\tau(t) = \frac{n(t,t+\tau)}{\tau} \qquad (2.1)$$

starting at an arbitrary time and taking equal intervals of time length τ, which, in this study, was taken as one month. The starting time t is an integer multiple of τ, and $n(t, t+\tau)$ is the number of rainy days in $(t, t+\tau)$. It was assumed that inside each interval $(t, t+\tau)$ the process of daily rainfall counts $N(t, t+\tau)$ is stationary. Under this stationary assumption it can be shown that the above statistic, which is the ratio of the number of rainfall counts in $(t, t+\tau)$ and the time interval τ, is unbiased. That is (*Cox and Lewis,* 1966),

$$E[\lambda_\tau(t)] = \frac{E[N(t,t+\tau)]}{\tau} = m(t) \qquad (2.2)$$

In figure 2 the estimated mean rates of occurrence are shown as functions of time. There is a very strong yearly periodicity and a downward trend in the graphs for the stations 0132, 3082, 3777, 4642, 6056 and 6338 chosen out of the 17 stations in Indiana for demonstration. For a stationary point process the mean rate of occurrence should plot as a straight horizontal line. However, there are obvious cyclicities and a long term time trend in the mean rate of occurrence of daily rainfall in Indiana. Therefore, the point process under consideration is nonhomogeneous. The results of the mean rate of occurrence agree with those of the cumulative plot of the daily rainfall counts versus time. In figure 2 the harmonic fits to the sample mean rates of occurrence are shown. Visually the fits are quite good. The analytical details of these fits will, however, be discussed in the next chapter on the homogenization of the daily rainfall counts process.

2.1.3 Intensity Function

One can define N_t^f as the number of rainy days in the interval (0,t) which starts with an occurrence at 0 but does not include it. This is the counting process associated with the process of interarrival times $\{X_i\}$ such that

$$P[N_t^f < n] = P[X_1 + \dots + X_n > t] \qquad n = 1, 2, \dots \tag{2.3}$$

under stationarity. Let $M_f(t)$ be the expectation of N_t^f under stationarity. Then the intensity function $m_f(t)$ is (*Cox and Lewis,* 1966)

$$m_f(\tau) = \frac{dM_f(\tau)}{d\tau} = \lim_{\Delta t \to 0} \frac{\text{Prob[event in } (t+\tau,\ t+\tau+\Delta t) | \text{event at } t]}{\Delta t} \tag{2.4}$$

where event at t is an arbitrary event in the stationary process. In the renewal theory $m_f(\tau)$ will be called the renewal density function. Since $M_f(t)$ is the expectation of N_t^f,

$$M_f(t) = \sum_{r=1}^{\infty} P[N_t^f \geq r] = \sum_{r=1}^{\infty} P[X_1 + \dots + X_r \leq t] \ . \tag{2.5}$$

Then,

$$m_f(t) = \frac{dE[N_t^f]}{dt} = \sum_{r=1}^{\infty} f_{X_1} + \dots + {}_{X_r}(t) \ . \tag{2.6}$$

For a renewal process, that is, for a process with independent identically distributed interarrival times $\{X\}$, denoting the Laplace transform of Z by L{Z}, its inverse by L^{-1} and the Laplace transform of $f_X(x)$ by $f_X(s)$,

$$L\{m_f(t)\} = \sum_{r=1}^{\infty} L\{f_{X_1} + \dots + {}_{X_t}(t) = \sum_{r=1}^{\infty} [f_X(s)]^r$$

$$= \frac{f_X(s)}{1 - f_X(s)}$$

and

$$m_f(t) = L^{-1}\left\{\frac{f_X(s)}{1 - f_X(s)}\right\} \ . \tag{2.7}$$

If the daily rainfall counts is assumed to be a homogeneous Poisson process with parameter λ,

$$f_X(s) = \lambda/(\lambda+s)$$

$$L^{-1}\left\{\frac{f_X(s)}{1 - f_X(s)}\right\} = \lambda$$

$$m_f(t) = \lambda \ . \tag{2.8}$$

Therefore, if the above assumption is true, the intensity or the renewal density function of the daily rainfall counts should be a straight, horizontal line. If the daily rainfall counts is assumed to be a process with gamma-distributed interarrivals with the integer index k and parameter λ,

$$f_X(s) = \left(\frac{\lambda}{\lambda+s}\right)^k \quad , \quad \frac{f_X(s)}{1 - f_X(s)} = \frac{1}{\left(1 + \frac{s}{\lambda}\right)^k - 1} . \tag{2.9}$$

If k=2,

$$\frac{f_X(s)}{1 - f_X(s)} = \frac{\lambda^2}{s(s+2\lambda)}$$

and

$$m_f(t) = L^{-1}\left(\frac{\lambda^2}{s(s+2\lambda)}\right) = \frac{\lambda}{2}(1 - e^{-2\lambda t}) \tag{2.10}$$

for the intensity of the gamma-distributed interarrivals with $E(X) = 2/\lambda$. Therefore, for this case the intensity or the renewal density will have the exponential shape with the asymptote $\lambda/2$. A point to consider in the calculation of the intensity function is that one starts at a certain time t where an event occurs and computes the derivative of the expected number of events $m_f(\tau)$ for the interval $(t+\tau, t+\tau+\Delta t)$ for Δt small. However, if the counting process is nonhomogeneous, the intensity function $m_f(\tau)$ will depend on its starting point t. In the intensity function computations of this study, since the hypothesis of stationarity is tested, all the intensity functions are functions of τ and not of the starting time t. However, any periodicity or trend should be apparent as function of τ and the intensity function should yield valuable information about the nonhomogeneities in the daily rainfall counts process. Starting with a rainy day, the daily rainfall counting process will be of the sort shown in Diagram 1. $(0,t_n)$ is divided into equal intervals of length α. There will be t_n/α intervals in $(0,t_n)$. Starting at each event t_i and

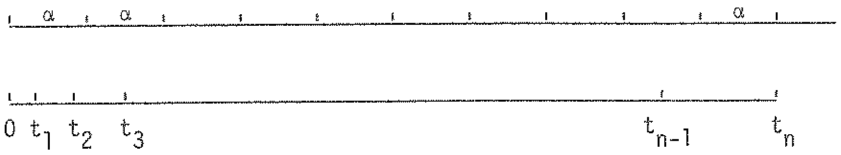


Diagram 1 Counting Setup No. 1

moving onwards until t_{n-1}, the number of events in each interval $(r\alpha, r\alpha+\alpha)$, r = 0, 1, ..., are counted for each setup which is identified by its starting point t_i. That is, if the counting starts at t_1, the intervals will be as shown in Diagram 2. That is, the first interval of length α will be the interval

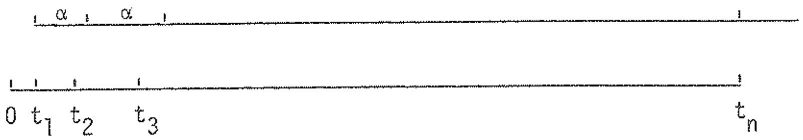


Diagram 2 Counting Setup No. 2

having its starting point at t_1. If the counting starts at t_3, the intervals will be as shown in Diagram 3.

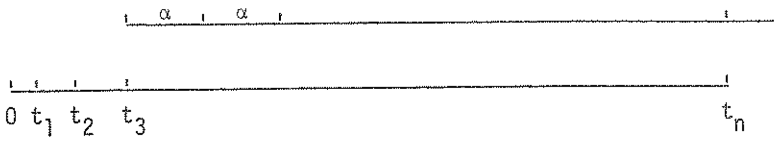


Diagram 3 Counting Setup No. 3

There will be n different counting setups due to the n starting points 0, t_1, t_2, ..., t_{n-1}. For each setup the number of events in the first, second, etc. intervals are counted. Then the sums S_r of the events which fall into $(r\alpha, r\alpha+\alpha)$ for r = 0, 1, ..., are formed by the addition of the number of events which fall into $(r\alpha,\ r\alpha+\alpha)$ for each of the setups. By dividing S_r by n the estimate $\tilde{m}_f(r\alpha)$ is obtained. In this procedure the intervals $(r\alpha,\ r\alpha+\alpha)$ for r large, will be averaged on a very few cases, especially when α is large. Therefore, there will be high variation for $\tilde{m}_f(\tau)$ when τ is large and near the end of the rainfall record. The analytical expression for the smoothed estimates by the above procedure is (*Cox and Lewis,* 1966)

$$\hat{m}_f(r\alpha + \frac{1}{2}\alpha) = \frac{1}{\alpha}\int_{r\alpha}^{(r+1)\alpha} \tilde{m}_f(u)du \qquad (2.11)$$

where $\tilde{m}_f(u) = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\delta(t_{i+j} - t_i - u)$ where $\delta(u)$ is the Dirac's delta function.

In (2.11) the intervals where $t_i + r\alpha > t_n$ do not contribute anything to $\hat{m}_f(t)$ and the $\hat{m}_f(t)$ is biased. In an analogous fashion to *Cox and Lewis* (1966) the unbiased estimate for the intensity function of the daily rainfall counting process is

$$\bar{m}_f(r\alpha - \frac{1}{2}\alpha) = \frac{t_n}{t_n - \alpha\left(r - \frac{1}{2}\right)}\hat{m}_f(r\alpha - \frac{1}{2}\alpha)\ . \qquad (2.12)$$

The graphs of $\bar{m}_f(\tau)$ versus τ, obtained by a computer program of *Lewis et al.* (1969), are shown in figure 3 for the stations 0132, 3082, 3777, 4642, 6056 and 7747. "Time interval" on the abcissa of these plots denotes the time span in days from the first event in 1950 on. The yearly cyclicity is clearly seen in the intensity function. Although the downward trend is also seen, 800 days for which $\bar{m}_f(\tau)$ is calculated is too short a time to claim a long term trend. The high sample fluctuations of the estimate $\bar{m}_f(\tau)$ for large τ prevented the calculation of $\bar{m}_f(\tau)$ for very long periods to access the long term time trend. Intensity function and the mean rate of occurrence are closely related functions. However, the intensity function is estimated on the counts which start with an arbitrary count, and is associated with N_t^f, while the mean rate of occurrence is estimated on counts which start with arbitrary time t, and is associated with $N(t,\ t+\tau)$, the stationary counting process in $(t,\ t+\tau)$.

2.1.4 Variance-time Function

The variance-time curve V(t) of the counting process N_t is defined as

$$V(t) = Var(N_t) = E(N_t^2) - E^2(N_t)\ . \qquad (2.13)$$

For a stationary process, the variance-time curve can be expressed in terms of the mean rate of occurrence m, and the intensity function $m_f(u)$ as (*Cox and Lewis,* 1966)

$$V(t) = mt + 2\int_0^t\int_0^V m(m_f(u) - m)dudv\ . \qquad (2.14)$$

Differentiating V(t),

$$V'(t) = m[1 + 2\int_0^t (m_f(u) - m)du] \qquad (2.14a)$$

and differentiating V'(t),

$$V''(t) = 2m(m_f(t) - m)\ . \qquad (2.15)$$

Expression (2.15) gives a simple relation between the mean rate of occurrence, intensity function and the second derivative of the variance-time curve which will be useful in later computations. Taking the Laplace transform of both sides of (2.14),

$$L_s\{V(t)\} = \frac{m}{s^2} + \frac{2m\ L_s\{m_f(t)\}}{s^2} - \frac{2m^2}{s^3}\ . \tag{2.16}$$

Denoting $L_s\{V(t)\}$ by $V^*(s)$ and $L_s\{m_f(t)\}$ by $m_f^*(s)$,

$$V^*(s) = \frac{m}{s^2} + \frac{2m\ m_f^*(s)}{s^2} - \frac{2m^2}{s^3}\ . \tag{2.17}$$

This general expression for any stationary point process can be used to obtain the variance-time properties for any type of a renewal process. For a renewal process it was earlier shown that

$$m_f^*(s) = f_x(s)/(1 - f_x(s))\ .$$

The Laplace transform of the variance-time function can be expressed in terms of the Laplace transform of the interarrival times, $f_x(s)$, as

$$V^*(s) = \frac{m}{s^2} + \frac{2m\ f_x(s)}{s^2(1 - f_x(s))} - \frac{2m^2}{s^3}\ . \tag{2.18}$$

$V(t)$ for a homogeneous Poisson process with parameter λ follows from (2.18) as

$$V^*(s) = \frac{\lambda}{s^2} + \frac{2\lambda f_x(s)}{s^2(1 - f_x(s))} - \frac{2\lambda^2}{s^3}\ . \tag{2.19}$$

Since the interarrival times are exponential with parameter λ, $f_x(s) = \lambda/(\lambda+s)$. Then

$$V^*(s) = \lambda/s^2\ ,\ \text{and}\ V(t) = \lambda t\ . \tag{2.20}$$

Therefore, the variance-time curve for a Poisson process is a straight line with slope λ. In the case of stationarity $\lambda = 1/E(x)$, where $E(x)$ is the interarrival time. Therefore, if the process is homogeneous Poisson

$$V(t) = t/E(x)\ . \tag{2.21}$$

$V(t)$ for the renewal process with gamma distributed interarrival times with $E(x) = 2/\lambda$ can be obtained as follows:

$$\text{Since } f_x(s) = \left(\frac{\lambda}{\lambda+s}\right)^2\ ,\ \text{then } V^*(s) = \frac{m}{s^2} + \frac{2m}{s^2}\cdot\frac{\lambda^2}{s(s+2\lambda)} - \frac{2m^2}{s^3}\ ,\ \text{where } m = \lambda/2\ .$$

Taking the inverse Laplace transform,

$$V(t) = \frac{\lambda}{2}\,t + \lambda^3 \int_0^t \int_0^\alpha \int_0^\tau e^{-2\lambda u}\, du\,d\tau\,d\alpha - \lambda^2 t^2/4\ , \tag{2.22}$$

$$V(t) = \frac{\lambda}{4}\,t + \frac{1}{8} - \frac{1}{8}\,e^{-2\lambda t}\ .$$

The variance-time curve is estimated by a computer program of *Lewis et al.* (1969) in the standard way through the use of a moving average procedure devised by *Cox and Smith* (1953). $V(t)$ is a moving average over the possible intervals of length t. Assume that the total length of the series is T. Since t is a section in T, take $k = T/t$. The rainfall series can be divided into intervals of length δ such that $t/\delta = j$. If the number of rainfall occurrences in the i-th interval of length δ is denoted by n_i, then n_i's in j consecutive blocks can be added to yield

$$\begin{aligned} S_1 &= n_1 + \dots + n_j \\ S_2 &= n_2 + \dots + n_{j+1} \\ S_{jk-(j-1)} &= n_{jk-(j-1)} + \dots + n_{jk} \end{aligned}$$

where each S is the number of rainy days in an interval of length t. Then V(t) can be estimated from the sum of squares of the moving sums S. Under the Poisson assumption the unbiased estimate of the V(t) is given as (*Cox and Smith*, 1953)

$$\tilde{V}(j\delta) = \frac{3}{3K(K-j)+j^2-1}\left\{\sum_{i=1}^{K} S_i^2 - \frac{1}{K}\left(\sum_{i=1}^{K} S_i\right)^2\right\} \tag{2.23}$$

where K = jk-(j-1). The variance of this estimate for the Poisson case was calculated by *Cox and Smith* (1953). It was advised that δ should be chosen at most one half the smallest interval for which the variance-time estimate is to be computed, so that there is a negligible decrease in the variance of the estimate. For the non-Poisson case little is known about the sampling properties of the estimate. According to *Cox and Lewis* (1966) the bias can be shown to be negligible for t less than about one-fifth of T. Therefore, variance-time curve for the daily rainfall counts was computed up to a time which was one-fifth of the whole record length. This corresponds to approximately two years. Two years is not a long time to notice the development of the transient in the variance behavior in the daily rainfall occurrences. As is seen in the figure 4 for the sample stations 0132, 3082, 3777, 4642, 6056, 7747 out of the 17 analyzed stations in Indiana, the linear asymptotic portion of the curve which is going to be important in the estimation of the spectrum of counts, is not developed. However, just the two years' behavior of the variance-time curve already indicates the clear periodicity in the variance of the daily rainfall occurrences. When compared to the theoretical variance-time curves for the homogeneous Poisson process which are also shown on these figures, the variance-time function of the daily rainfall process indicates an overdispersion, a clustering of events, since an upward deviation from the regular Poisson variance-time function would indicate a coefficient of variation greater than unity. This is a very important result since it indicates a grouping mechanism in the daily rainfall occurrence process while also indicating the nonhomogeneity of the process. As is seen from expression (2.14) the variance-time function is the double integral of the intensity function, which, in turn, will be shown to be the Fourier transform of the spectrum of counts. Furthermore, the estimates of V(t) for different t are highly correlated (*Cox and Lewis*, 1966) so that the conclusions based on V(t) should be taken with caution. Due to these two reasons the transient effects for short intervals can best be interpreted by the spectrum of counts of the daily rainfall.

2.1.5 The Spectrum of Daily Rainfall Counts

The spectrum of counts, $g(\omega)$, for a stationary point stochastic process is expressed as (*Cox and Lewis*, 1966),

$$g(\omega) = \frac{m}{2\pi} + \frac{m}{2\pi}\int_{-\infty}^{+\infty} \{m_f(\tau) - m\}\, e^{-i\omega\tau}\, d\tau\ , \quad -\infty < \omega < +\infty\ . \tag{2.24}$$

If the spectrum of counts is defined only for the positive frequencies,

$$g_+(\omega) = 2g(\omega) = \frac{m}{\pi} + \frac{m}{\pi}\int_{-\infty}^{+\infty} \{m_f(\tau) - m\}\, e^{-i\omega\tau}\, d\tau\ , \quad \omega \geq 0\ . \tag{2.25}$$

A very important feature to note is that $g_+(\omega)$ is not periodic since the corresponding function in the time domain, $m_f(\tau)$, is a function in continuous time. Therefore, there is no boundary on the extent of the spectrum of counts. The spectral value at the origin $g_+(0^+)$, as it is approached from the right, becomes

$$g_+(0^+) = \frac{m}{\pi} + \frac{m}{\pi}\int_{-\infty}^{+\infty} (m_f(\tau) - m)d\tau \qquad (2.26)$$

and making use of (2.14a) one obtains

$$g_+(0^+) = \frac{1}{\pi} V'(t)\big|_{t=\infty} . \qquad (2.27)$$

Using (2.25) and recalling the Laplace transform definition

$$g_+(\omega) = \frac{m}{\pi} [1 + m_f^*(i\omega) + m_f^*(-i\omega)] . \qquad (2.28)$$

For any renewal process;

$$g_+(\omega) = \frac{m}{\pi}\left\{1 + \frac{f_x(i\omega)}{1 - f_x(i\omega)} + \frac{f_x(-i\omega)}{1 - f_x(-i\omega)}\right\} . \qquad (2.29)$$

Using expression (2.29) the theoretical spectrum of counts for any type of renewal process can be derived.

i. Spectrum of counts for the homogeneous Poisson process with parameter λ.

For this case $f_x(i\omega) = \lambda/(\lambda+i\omega)$. The $m_f^*(i\omega)$ becomes $\lambda/i\omega$. The expression (2.29) for the spectrum of counts takes the form

$$g_+(\omega) = \frac{\lambda}{\pi} (1 + \frac{\lambda}{i\omega} - \frac{\lambda}{i\omega}) ,$$

so that

$$g_+(\omega) = \lambda/\pi , \ \omega \geq 0 .$$

Therefore, the positive spectrum of counts for a homogeneous Poisson process is a constant horizontal line.

ii. Spectrum of counts for the renewal process with gamma distributed interarrivals where $E(x) = 2/\lambda$:

Spectrum of counts for this case was derived by *Cox and Lewis* (1966) as

$$g_+(\omega) = \frac{\lambda}{2\pi}\left[\frac{\omega^2 + 2\lambda^2}{\omega^2 + 4\lambda^2}\right] , \ \omega \geq 0 .$$

Positive spectrum of counts increases monotonically from $\lambda/4\pi$ to $\lambda/2\pi$.

For the ordinary time series the periodogram is estimated by (*Cote,* 1973)

$$I_j = \frac{1}{n}\left|\sum_{\alpha=0}^{n-1} X_\alpha e^{-i2\pi\frac{\alpha j}{n}}\right|^2 .$$

In an analogous fashion to the time series, the spectrum of counts for the stationary counting process N(t) can be estimated as follows (*Cox and Lewis,* 1966);

let;

$$H_T(\omega) = \frac{1}{\sqrt{\pi T}}\int_{t=0}^{T} e^{it\omega}\, dN(t) .$$

Since n events occur at times $t_1, t_2, \ldots, t_n$ in (0,T) for the counting process N(t),

$$H_T(\omega) = \frac{1}{\sqrt{\pi T}}\sum_{j=1}^{n} e^{it_j\omega} = \frac{1}{\sqrt{\pi T}}\sum_{j=1}^{n} \mathrm{Cos}(t_j\omega) + i\sum_{j=1}^{n} \mathrm{Sin}(t_j\omega)$$

where $\sum_{j=1}^{n} \mathrm{Cos}(t_j\omega)$ will be called $A_T(\omega)$ and $\sum_{j=1}^{n} \mathrm{Sin}(t_j\omega)$ will be called $B_T(\omega)$. In analogy with the above periodogram of the time series, the periodogram of the counting process, $\tilde{g}_+(\omega)$, will be

$$\tilde{g}_+(\omega) = H_T(\omega)\,\overline{H_T(\omega)} = \frac{1}{\pi T}\sum_{j=1}^{n}\sum_{s=1}^{n} e^{i\omega(t_j-t_s)} = \frac{1}{\pi T}\{A_T^2(\omega) + B_T^2(\omega)\}\ . \tag{2.32}$$

For a homogeneous Poisson process with rate λ the distributional properties of $\tilde{g}_+(\omega)$ are derived by *Cox and Lewis* (1966) as follows:

$$\lim_{T\to\infty} P[\tilde{g}_+(\omega) > y] = e^{-\pi y/\lambda} \qquad , \frac{\omega T}{2\pi} = 1,\ 2,\ \ldots \tag{2.33}$$

$$E[\tilde{g}_+(\omega)] = \frac{\lambda}{\pi} \qquad , \frac{\omega T}{2\pi} = 1,\ 2,\ \ldots$$

$$= \frac{\lambda}{\pi} + \lambda^2 T\left\{\frac{\operatorname{Sin}\left(\frac{1}{2}\,\omega T\right)}{\frac{1}{2}\,\omega T}\right\}^2 \quad \text{otherwise, for } \omega > 0 \tag{2.34}$$

$$\operatorname{Var}[\tilde{g}_+(\omega)] = \frac{\lambda^2}{\pi^2}\left(1 + \frac{1}{\lambda T}\right) \qquad \frac{\omega T}{2\pi} = 1,\ 2,\ \ldots \tag{2.35}$$

$$\operatorname{Corr}[\tilde{g}_+(\omega_1),\ \tilde{g}_+(\omega_2)] = \frac{1}{1+\lambda T} \qquad \frac{\omega_1 T}{2\pi} = 1,\ 2,\ \ldots\ \frac{\omega_2 T}{2\pi} = 1,\ 2,\ \ldots, \quad \omega_1 \neq \omega_2\ . \tag{2.36}$$

From (2.34) it follows that, under the Poisson assumption, $\tilde{g}_+(\omega)$ is unbiased if $\omega T/2\pi$ is a positive integer. From (2.36) if follows that, under the Poisson assumption, the correlation between the two spectrum of counts estimates decreases as the record length T increases.

Due to the absence of a simple linear representation of the sequences, the distributional properties of $\tilde{g}_+(\omega)$, the estimate of the spectrum of counts under the non-Poissonian conditions was not derived. However, it was shown that (*Cox and Lewis,* 1966)

$$\lim_{T\to\infty} E[\tilde{g}_+(\omega)] = g_+(\omega) \qquad \text{for } \omega > 0\ . \tag{2.37}$$

That is, for a large sample size the spectral estimates are approximately unbiased even for a non-Poisson point process. The bias at the integer values of $\omega T/2\pi$ for $\omega > 0$, will be the smallest among all the values of ω. Therefore, the spectrum of counts estimate is still quite valid for any point process besides Poisson.

The normalized spectrum of counts for the point process is obtained by multiplying $\tilde{g}_+(\omega)$ by π/λ, the inverse of the asymptotic standard deviation of $\tilde{g}_+(\omega)$. Under the Poisson process the normalized spectrum of counts estimate $\tilde{g}_{N^+}(\omega_j')$ will have exponential distribution with (*Cox and Lewis,* 1966)

$$E[\tilde{g}_{N^+}(\omega_j')] \simeq 1$$

$$\operatorname{Var}[\tilde{g}_{N^+}(\omega_j')] \simeq 1 + \frac{1}{n}$$

$$\operatorname{Corr}\{\tilde{g}_{N^+}(\omega_{j_1}')\ \tilde{g}_{N^+}(\omega_{j_2}')\} \simeq \frac{1}{1+n}\ ,\ \omega_{j_1}' = \frac{2\pi j_1}{n},\ w_{j_2}' = \frac{2\pi j_2}{n}\ ,\ j_1 \neq j_2 \tag{2.40}$$

The normalized spectra of the daily rainfall counts for the 17 stations in Indiana were computed by a computer program (*Lewis et al.,* 1969) at the integer values $\omega_j = \omega T/2\pi = 1,\ 2,\ \ldots$ to avoid the bias. On the abscissa of the plots in figure 5 the frequency index j corresponds to the period 2556/j days since 7 years of data was analyzed. The normalized spectrum was then smoothed by averaging groups of k consecutive values. Since

$$P[\tilde{g}_{N^+}(\omega_j') > y] = e^{-y} \;, \; y > 0 \;,$$

then

$$E\, e^{iu\omega_j'} = \frac{1}{1-iu} \;, \quad E\, e^{iuk \frac{\omega_j' k}{k}} = \frac{1}{(1-iu)^k}$$

and

$$E\, e^{iu2k \frac{\omega_j' k}{k}} = 1/(1-i2u)^k \;. \tag{2.41}$$

Therefore, the smoothed normalized spectrum $\tilde{g}_k(\omega_j')$, obtained by averaging groups of k consecutive normalized spectrum values $\tilde{g}_{N^+}(\omega_j')$, when multiplied by 2k, has chi-squared distribution with 2k degrees of freedom.

That is

$$2k\tilde{g}_k(\omega_j') \sim \chi^2_{2k}$$

under the Poisson hypothesis. Once the distribution of the normalized, smoothed spectrum of counts estimates is known, the 99% confidence limits for the spectrum of counts under the Poisson null hypothesis are formed. The spectrum of counts analysis was carried for the 17 rainfall stations in Indiana for the daily rainfall counts process. The 99% confidence limits were constructed for each of the stations so as to detect the significant periodicities. However, the spectra of the daily rainfall counts not only showed significant periodicities but also the dependence structure in the daily rainfall occurrences. Due to this dependence structure, which will be explained in terms of the theoretical spectrum of counts of a dependence model in a later chapter, there were too many frequencies outside the 99% confidence limits. This fact is shown on the spectra of the daily rainfall counts for the sample stations 0132, 3082, 3777, 4642, 6056 and 7747 in figure 5. Since the dominant yearly periodicity is at j=7, corresponding to 2556/7 = 365 days, and since the spectral estimates were smoothed by averaging consecutive groups of 20 estimates, the highest spectral value appears at the origin. However, the value at the origin is not estimated since it is highly biased. Although the estimates averaged in groups of 20 show a dependence structure analogous to that of the autoregressive process in the time series analysis, there are still significant periodicities imbedded into this dependence mechanism. In order to determine these periodicities, averaging in smaller groups of estimates is needed. Therefore, the spectrum of daily rainfall counts estimates were averaged in consecutive groups of 5 in order to detect these periodicities. The periodicity analysis based on the spectral estimates, smoothed in five-member groups and twenty-member groups is given in Table 2-1. As an example to what is meant by five-member group smoothing, the spectrum of daily rainfall counts for the five-member group averaged estimates are shown for the stations 0132, 0545 and 3082 in Figure 6. By comparing the spectra of counts for 0132 in figures 5 and 6 and the spectra of counts for 3082 in figures 5 and 6 the dominant yearly periodicity is clearly seen. The striking fact derived from the Table 2-1 is that a periodicity of 11.6-16 days is significant in 13 out of 17 cases analyzed. The period of 2556 days, significant in 4 out of 17 cases, is thelength of the record being analyzed. Therefore, rather than considering it as a significant period it is believed to indicate to the long term time trends, which were shown to exist in the other graphical analyses. Another interpretation is that the high value is due to being the first value of the spectrum of a dependence mechanism.

Although the spectrum of daily rainfall counts is very important in the detection of the significant periodicities in the daily rainfall counts process, the results based on this analysis should be taken with caution. The theory for the spectrum of counts is based on stationarity. It is also acceptable for the circular stationarity in analogy with the time series. However, it could be seen from the earlier graphical analysis that there is a significant long term trend in the daily rainfall counts in Indiana.

Secondly, the significant periods were obtained under the hypothesis that the process is Poisson. The significant spectral values may be due to the fact that the daily rainfall counts is a non-Poisson process, and may just be the deviations from the Poisson hypothesis. In order to construct the spectrum of daily rainfall counts it had to be assumed that the long term time trends do not have an important effect in the seven years of daily rainfall record under study. The 99% confidence limits were constructed on the assumption that the daily rainfall counts is a Poisson process, based on the earlier works of *Lobert* (1967), of *Todorovic and Yevjevich* (1969), of *Duckstein et al.* (1972) and others on the daily rainfall counts process. Due to the existing dependence structure, as is seen from the spectra of counts, the detection of periodicities is still in a state of art since a general theory for the spectra of the dependent point stochastic processes does not exist.

Besides the graphical techniques for the analysis of nonhomogeneity in the daily rainfall occurrences, there are also the tests of the homogeneity hypothesis under certain conditions. In the next section these tests will be considered.

2.2 TESTS FOR THE STATIONARITY OF THE INTERARRIVAL TIMES IN THE DAILY RAINFALL OCCURRENCE PROCESS

Based on the assumption that the intervals between the daily rainfall occurrences are independent, the following tests are devised.

2.2.1 Test of Trend in the Rate of Occurrence of Daily Rainfall When the Poisson Process is Assumed to be the Underlying Stochastic Model

If the rate of daily rainfall occurrence $\lambda(t)$ is considered to have the function form $\lambda(t) = e^{\alpha+\beta t}$, the test is for the null hypothesis $\beta=0$ against the alternative $\beta\neq 0$. Given the number of occurrences n, the positions of the events in a Poisson process are independently, uniformly distributed over $(0,t_n)$. Then $S = \sum_{i=1}^{n} t_i/n$ has the distribution of the sum of n independent uniform random variables. *Cramer* (1946) showed that the statistic

$$U = S - \frac{t_n}{2} \ /t_n/\sqrt{12n}$$

is asymptotically standard normal as $n\to\infty$. By this statistic the centroid of the observed times to events t_i is compared to the mid-point of the period of observation. A positive value of U means that the centroid of events is greater than the mid-point of $(0,t_n)$ and the rate of occurrence is increasing with time. The results of this test are given in Table 2-2 for 17 stations in Indiana. The results of the test show that the rate of daily rainfall occurrence, under the Poisson assumption, is decreasing in Indiana. β is significantly different from zero in 12 out of 17 cases, showing the nonhomogeneity in the daily rainfall occurrences in Indiana.

2.2.2 Homogeneity of Variance Test Using Bartlett's Statistic

k independent samples of sizes n_i, i = 1, ..., k, can be formed by grouping the successive interarrival times into k groups. If it is assumed that these samples are taken from normal populations of mean μ_i and variance σ_i, and if Bartlett's modification of the likelihood ratio statistic where the independent sample sizes n_i are replaced by the degrees of freedom $\nu_i = n_i-1$, is made, the modified likelihood ratio becomes (*Kendall and Stuart,* 1961),

$$\ell* = \frac{L(x|H_0)}{L(x|\Omega)} = \prod_{i=1}^{k} \left(\frac{s_i^2}{s^2}\right)^{\nu_i/2} \tag{2.42}$$

where x is the interval length, and Ω is the parameter space. Then

$$s_i^2 = \frac{1}{\nu_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \ , \ s^2 = \frac{1}{\nu} \sum_{i=1}^{k} \nu_i \, s_i^2 \ , \ \nu = \sum_{i=1}^{k} \nu_i \ .$$

Using (2.42), $$-2 \log \ell* = \nu \log s^2 - \sum_{i=1}^{k} \nu_i \log s_i^2$$

and (*Kendall and Stuart*, 1961), $$-2 \log \ell* / \left(1 + \frac{1}{3k-3} \left[\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\nu} \right] \right) \sim \chi_{k-1}^2 \ . \qquad (2.43)$$

Therefore, the null hypothesis H_0, $$H_0 \colon \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

can be tested by the use of the χ^2 statistic. This test gives information about the second-order stationarity of the interarrival times. However, it should be remembered that the test is only approximate since the normality and the independence of the intervals are assumed. In table 2-3 the homogeneity of variance test results are given for 17 stations in Indiana. All the stations have significant nonhomogeneity in the variance of the daily rainfall occurrences at 1% level.

2.3 SUPPLEMENTARY FIGURES

Plots of the number of rainy days vs. cumulative time, of the mean rate of daily rainfall occurrence, of the intensity function, of the variance-time function, of the spectrum of daily rainfall counts are given for stations 0177, 0545, 1747, 1882, 3547, 4908, 6164, 6338, 7069, 7755, 7935 in figures 1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B, and 5A, 5B.

TABLE 2-1

SPECTRUM OF DAILY RAINFALL COUNTS ANALYSIS FOR THE DETECTION OF SIGNIFICANT PERIODICITIES AT 1% LEVEL FOR THE ESTIMATES SMOOTHED IN CONSECUTIVE FIVE-MEMBERS-GROUPS.

Station	Periods Significant at 1% (days)
0132	2556, 365, 16, 14.8, 3.7
0177	365, 183, 69, 54.4, 5.7
0545	2556, 365, 71, 16, 12.7
1747	365, 16, 14.9, 114
1882	365, 150, 14.9, 14, 7.7
3082	365, 23.2, 16, 14, 10.2
3547	2556, 365, 150, 116, 16, 14, 12.5, 11.6, 8, 6, 3.9
3777	365, 14.6, 12.2, 8.6, 4.4
4642	365, 16, 14.9, 7.7
4908	365, 11.4, 8.6, 5, 3.9
6056	2556, 365, 122, 16, 10.2, 9.6
6164	365, 7.6, 6.3
6338	365, 20.3, 16, 14.9, 14, 8.6
7069	365, 14.9, 14, 11.6, 7.7, 6
7747	365, 67.3, 9.5, 8.6
7755	365, 150.4, 29.4, 16, 14.7, 12.6
7935	365, 16, 14.9, 12.6, 5.7

TABLE 2-2

TEST OF TREND IN THE RATE OF OCCURRENCE (UNDERLYING PROCESS ASSUMED POISSON)

Station	Cramer's Statistic U	Significant at 5%
0132	-2.339	Yes
0177	-2.289	Yes
0545	-1.108	No
1747	-2.243	Yes
1882	-2.004	Yes
3082	-3.182	Yes
3542	-3.65	Yes
3777	-1.764	No
4642	-3.628	Yes
4908	-2.826	Yes
6056	-4.314	Yes
6164	-1.469	No
6338	-2.066	Yes
7069	-2.782	Yes
7747	.4027	No
7755	-3.644	Yes
7935	- .822	No

$U \sim N(0,1)$

TABLE 2-3

HOMOGENEITY OF VARIANCE TEST FOR STATIONARITY OF THE INDEPENDENT INTERARRIVAL TIMES USING BARTLETT'S STATISTIC

Station	Group Size	Degrees of Freedom	Hom. of Variance Statistic	Significant at 1%
7935	6	107	349.016	Yes
	18	35	212.674	Yes
	48	12	97.152	Yes
	120	4	30.556	Yes
7755	18	39	234.63	Yes
	48	14	80.97	Yes
	120	5	43.48	Yes
7747	18	42	238.20	Yes
	48	15	159.08	Yes
	120	5	31.28	Yes
7069	18	43	263.29	Yes
	48	15	155.09	Yes
	120	5	40.81	Yes
6338	18	41	234.39	Yes
	48	14	95.31	Yes
	120	5	13.11	No
6164	18	42	241.34	Yes
	48	15	147.04	Yes
	120	5	30.82	Yes
6056	18	38	615.22	Yes
	48	13	517.32	Yes
	120	4	288.72	Yes
4908	18	40	201.12	Yes
	48	14	108.33	Yes
	120	5	40.47	Yes
4642	18	42	226.20	Yes
	48	15	147.09	Yes
	120	5	44.40	Yes
3777	18	41	263.52	Yes
	48	14	99.26	Yes
	120	5	41.94	Yes
3082	18	43	269.07	Yes
	48	15	129.94	Yes
	120	5	54.05	Yes
1882	18	31	153.00	Yes
	48	11	93.56	Yes
	120	3	33.22	Yes
1747	18	41	277.33	Yes
	48	14	186.93	Yes
	120	5	70.46	Yes
0545	18	41	274.86	Yes
	48	15	185.90	Yes
	120	5	64.47	Yes
0177	18	38	495.96	Yes
	48	13	259.48	Yes
	120	4	49.85	Yes
0132	18	43	336.01	Yes
	48	15	148.16	Yes
	120	5	51.08	Yes

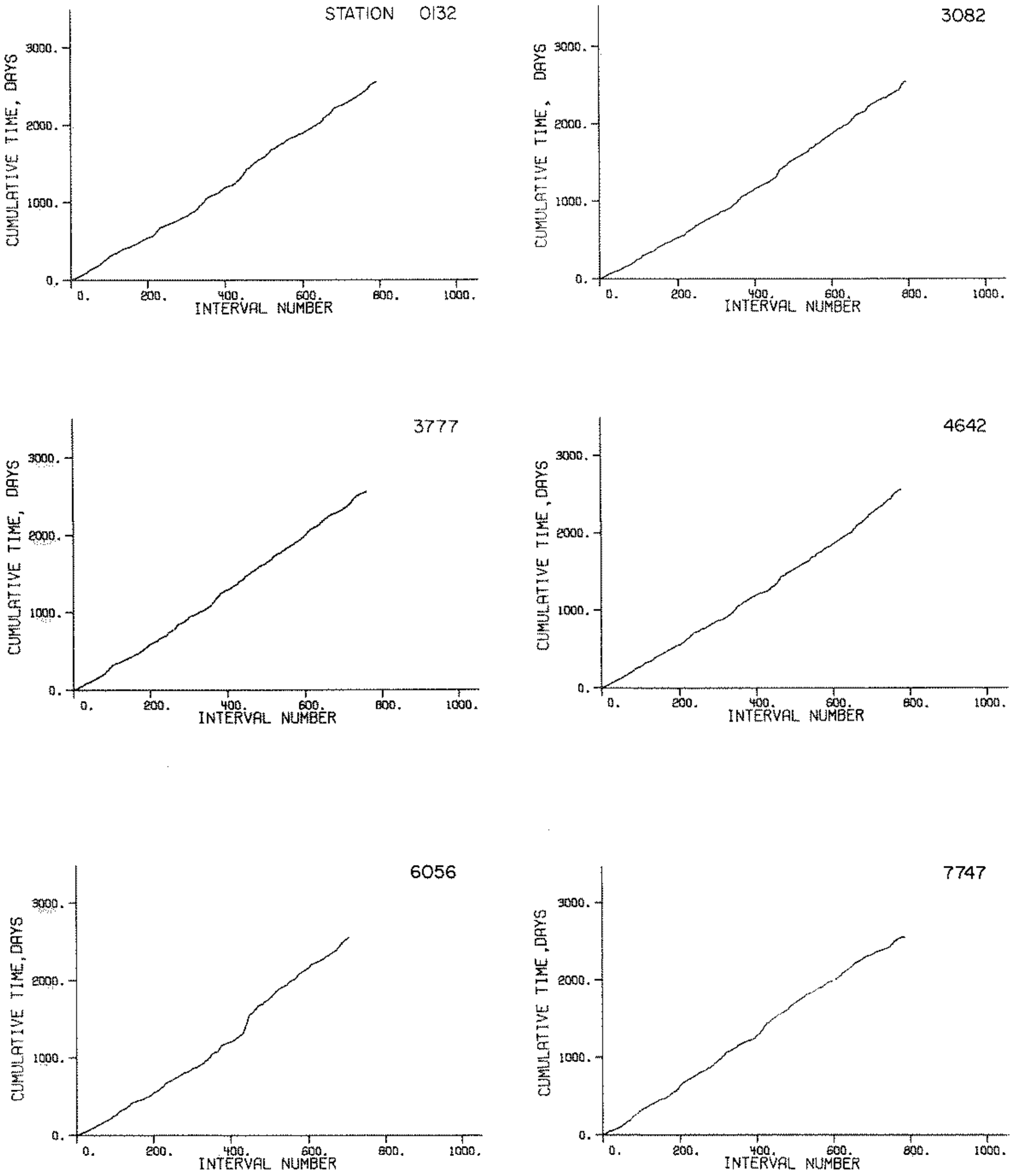


FIG. 1— CUMULATIVE TIME IN DAYS TO EVENTS VS
EVENT NUMBER

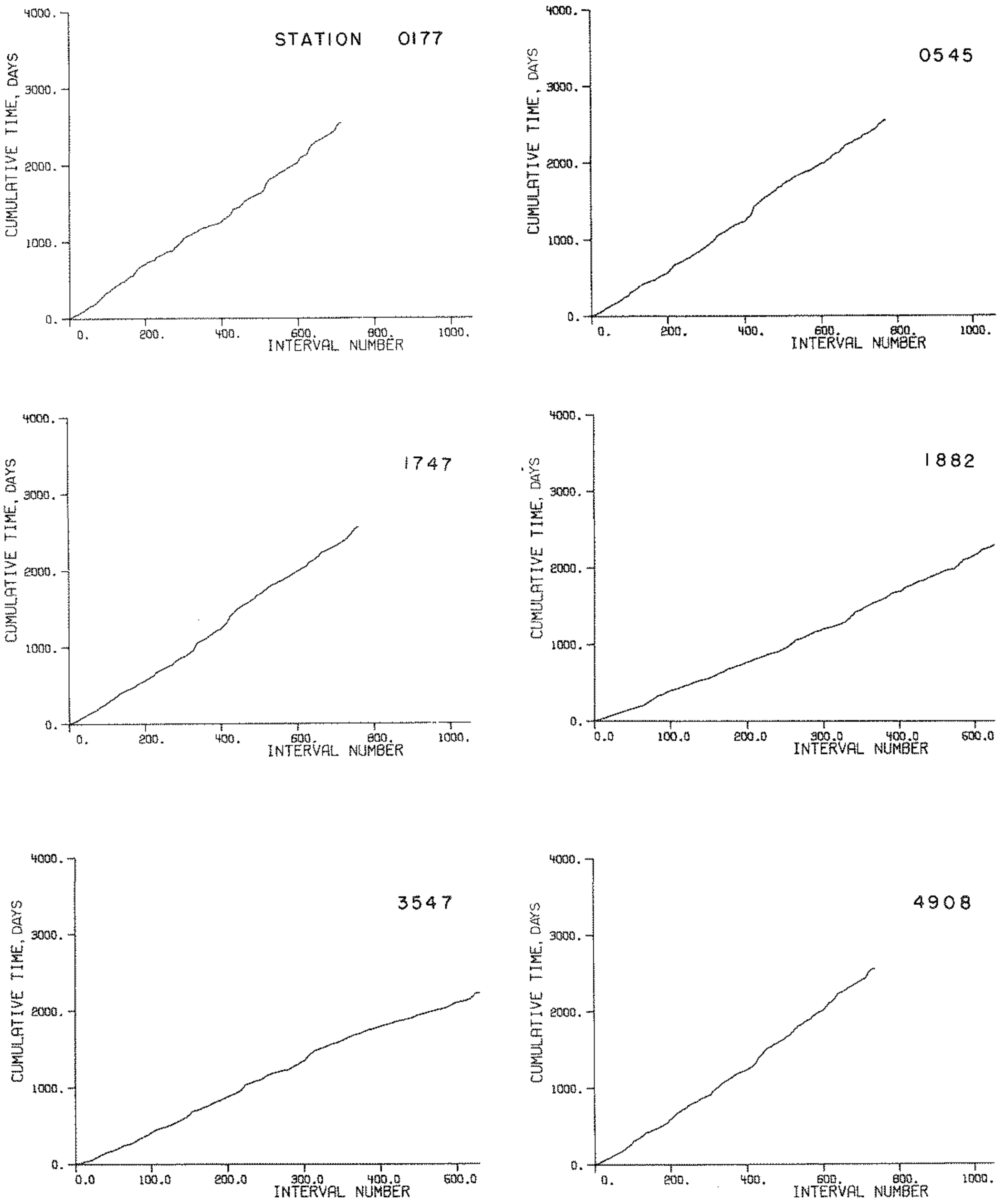


FIG. IA — CUMULATIVE TIME IN DAYS TO EVENTS VS EVENT NUMBER

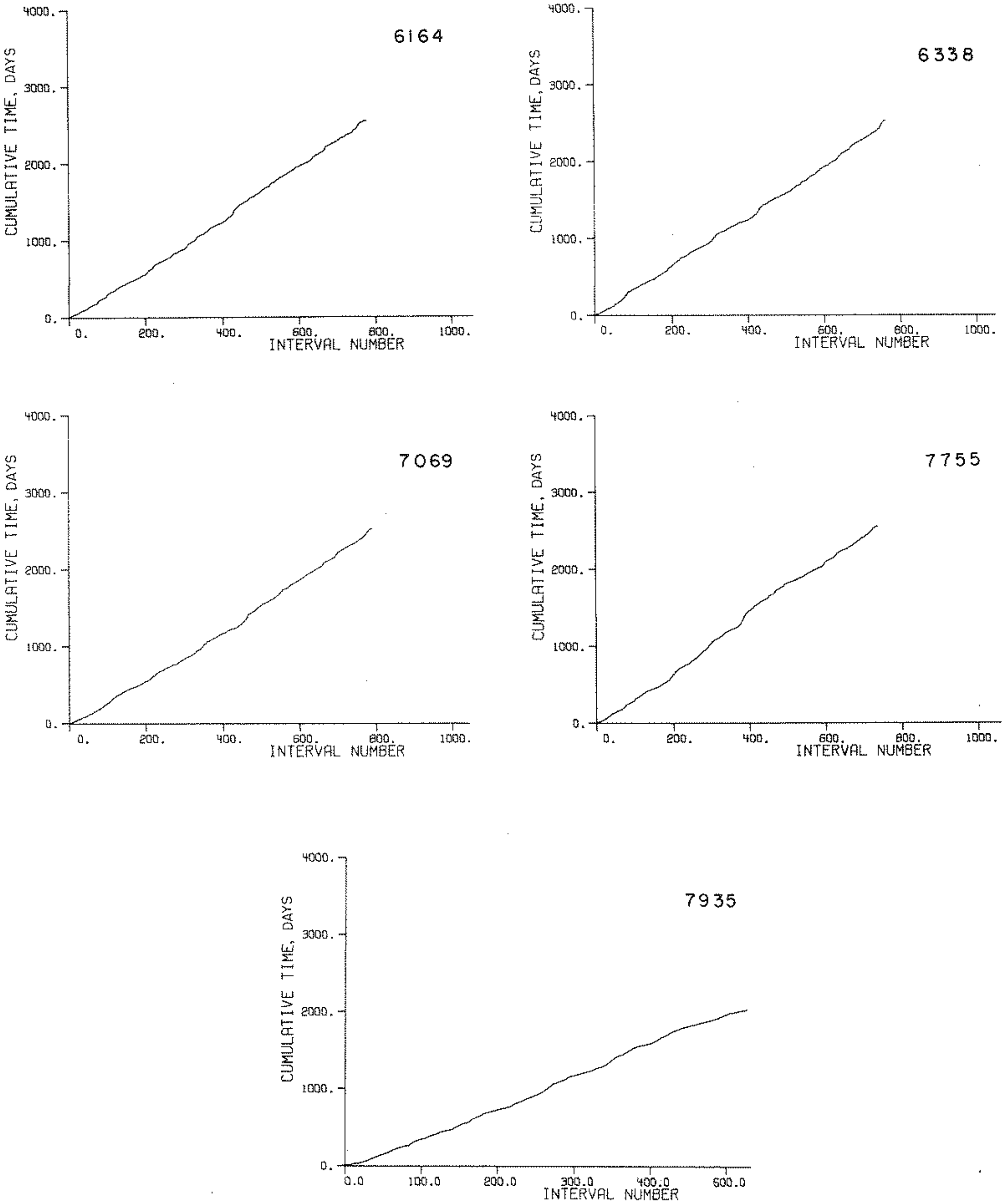


FIG. 1B —CUMULATIVE TIME IN DAYS TO EVENTS VS EVENT NUMBER

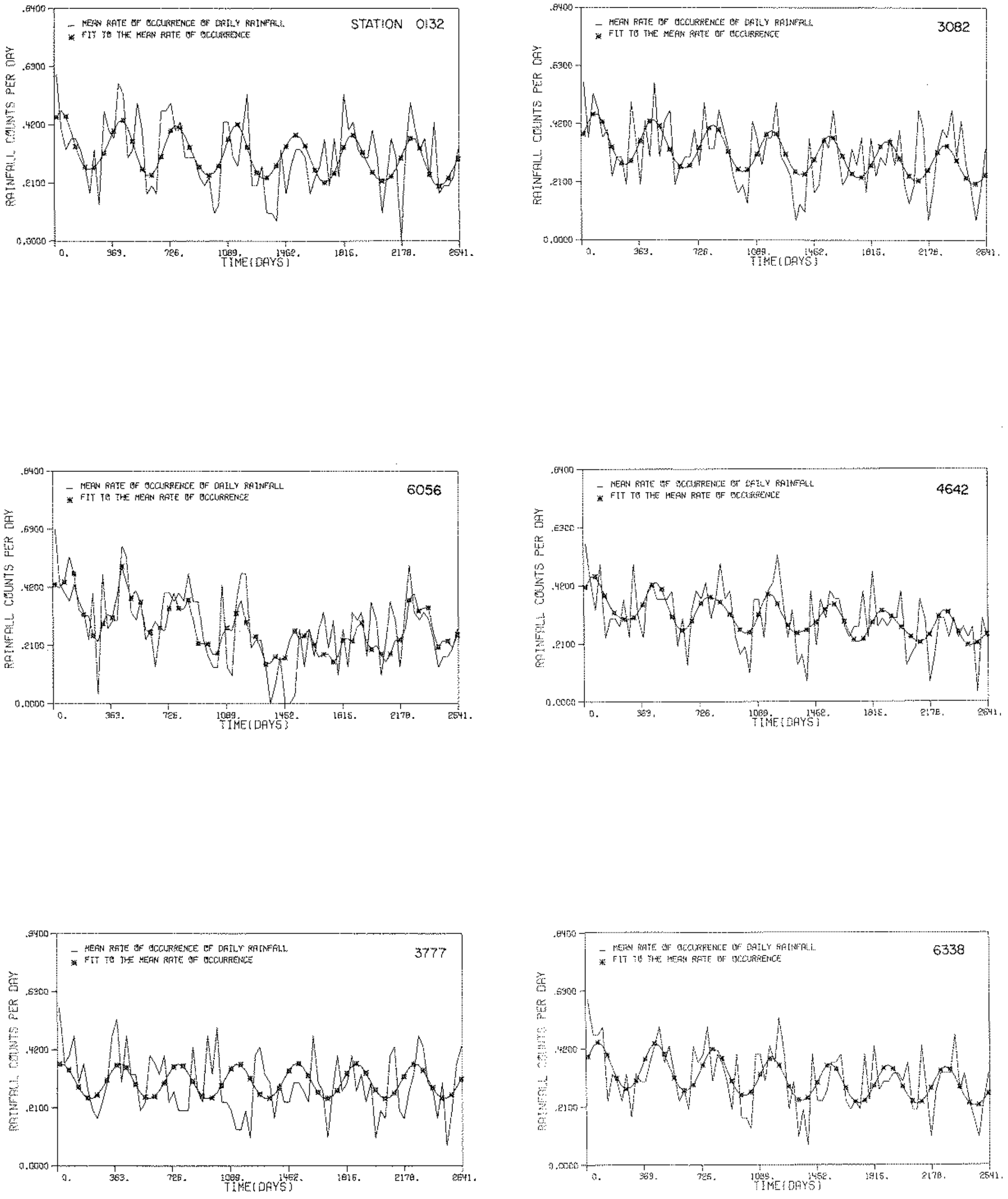


FIG. 2- RATE OF OCCURRENCE FUNCTION FOR DAILY RAINFALL

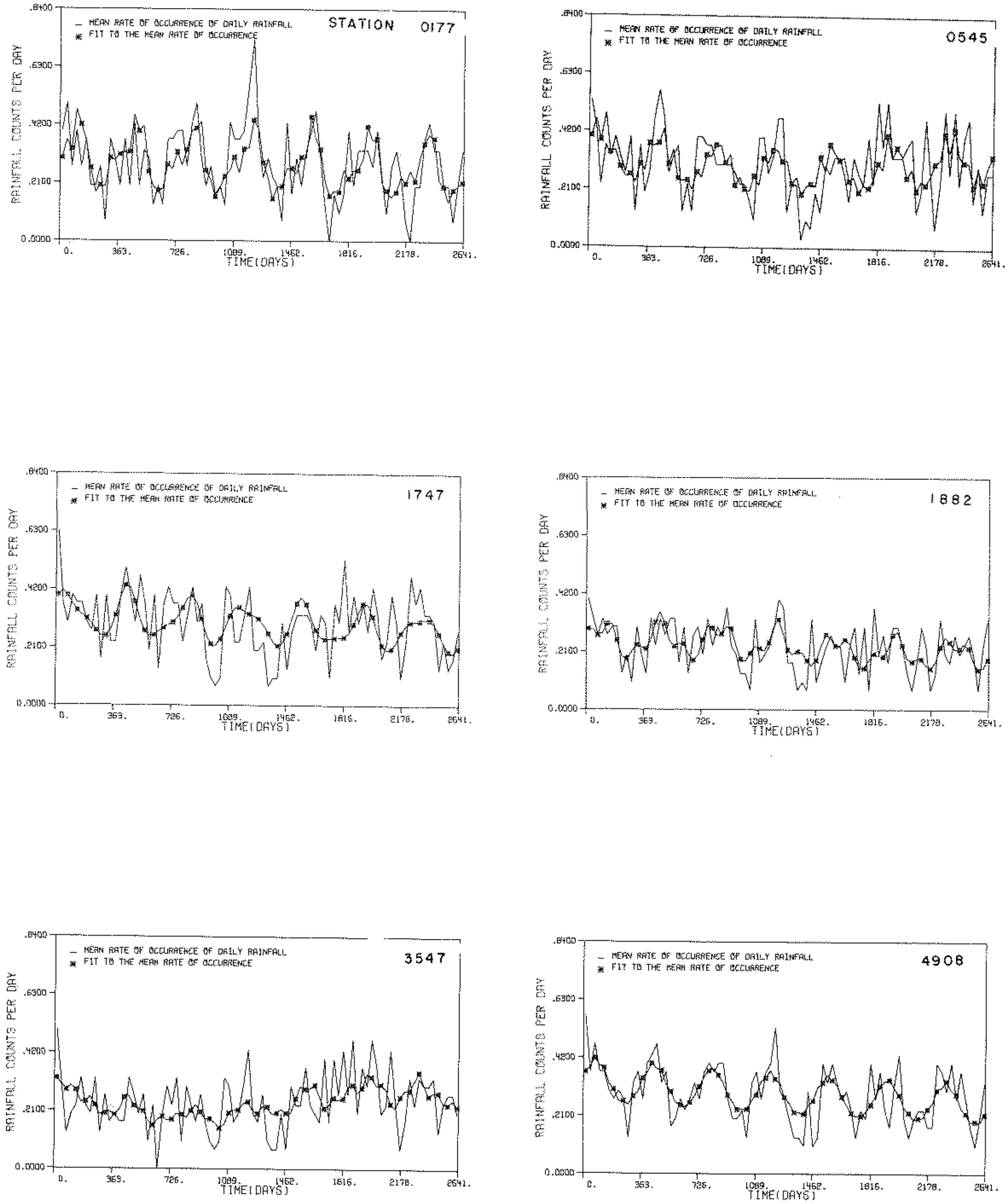


FIG. 2A—RATE OF OCCURRENCE FUNCTION FOR DAILY RAINFALL

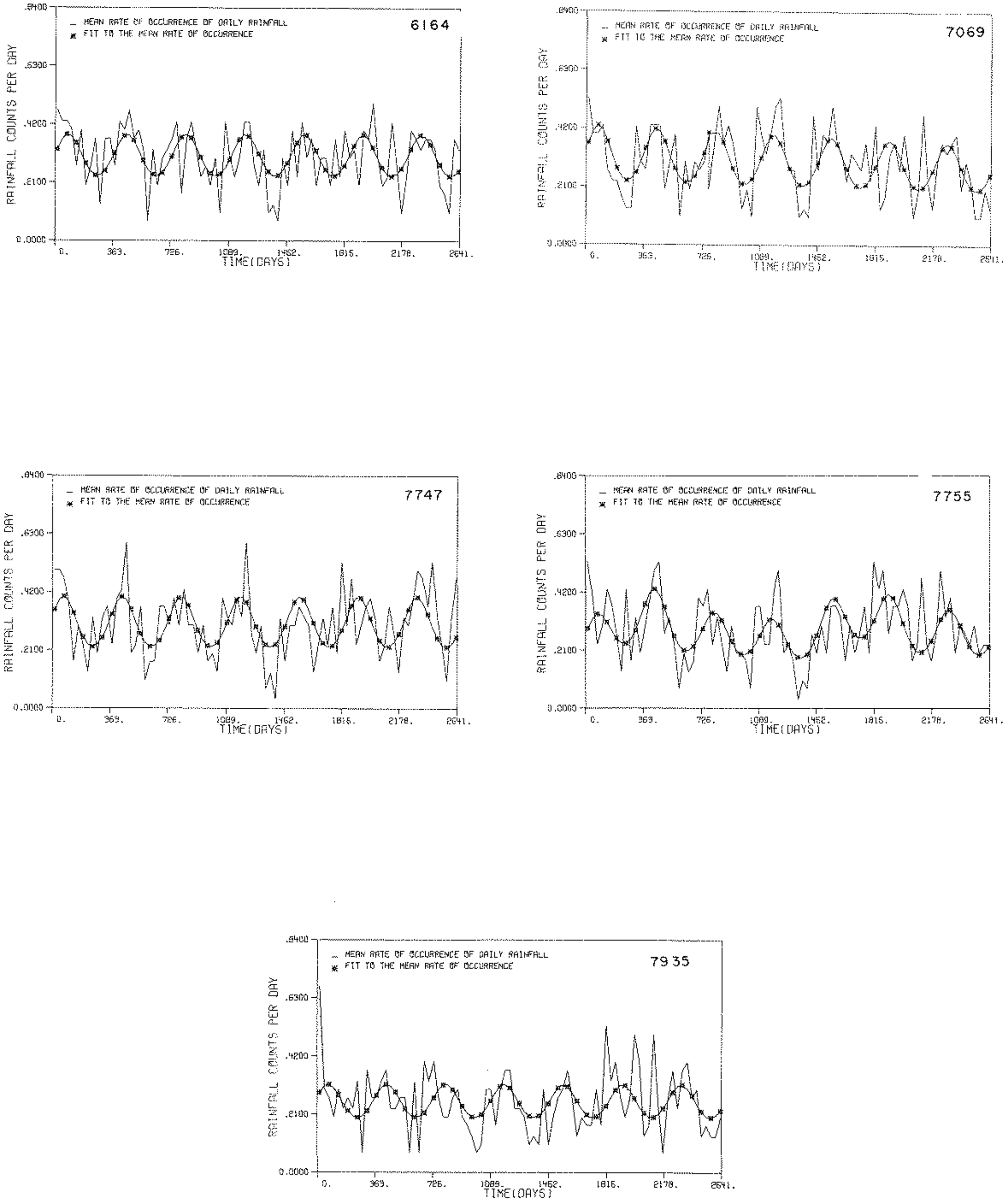


FIG. 2B—RATE OF OCCURRENCE FUNCTION FOR DAILY RAINFALL

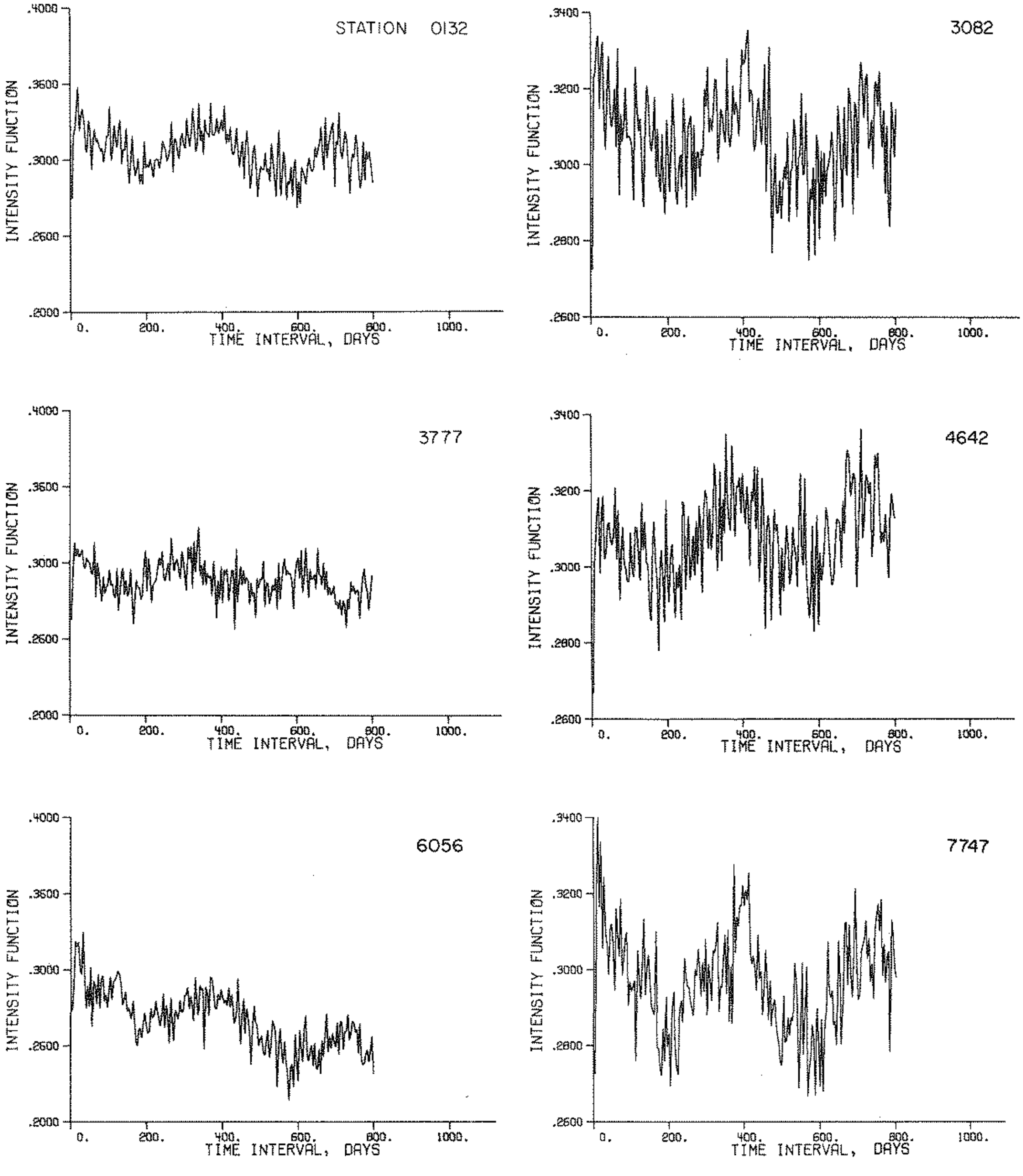


FIG. 3- INTENSITY FUNCTION OF DAILY RAINFALL

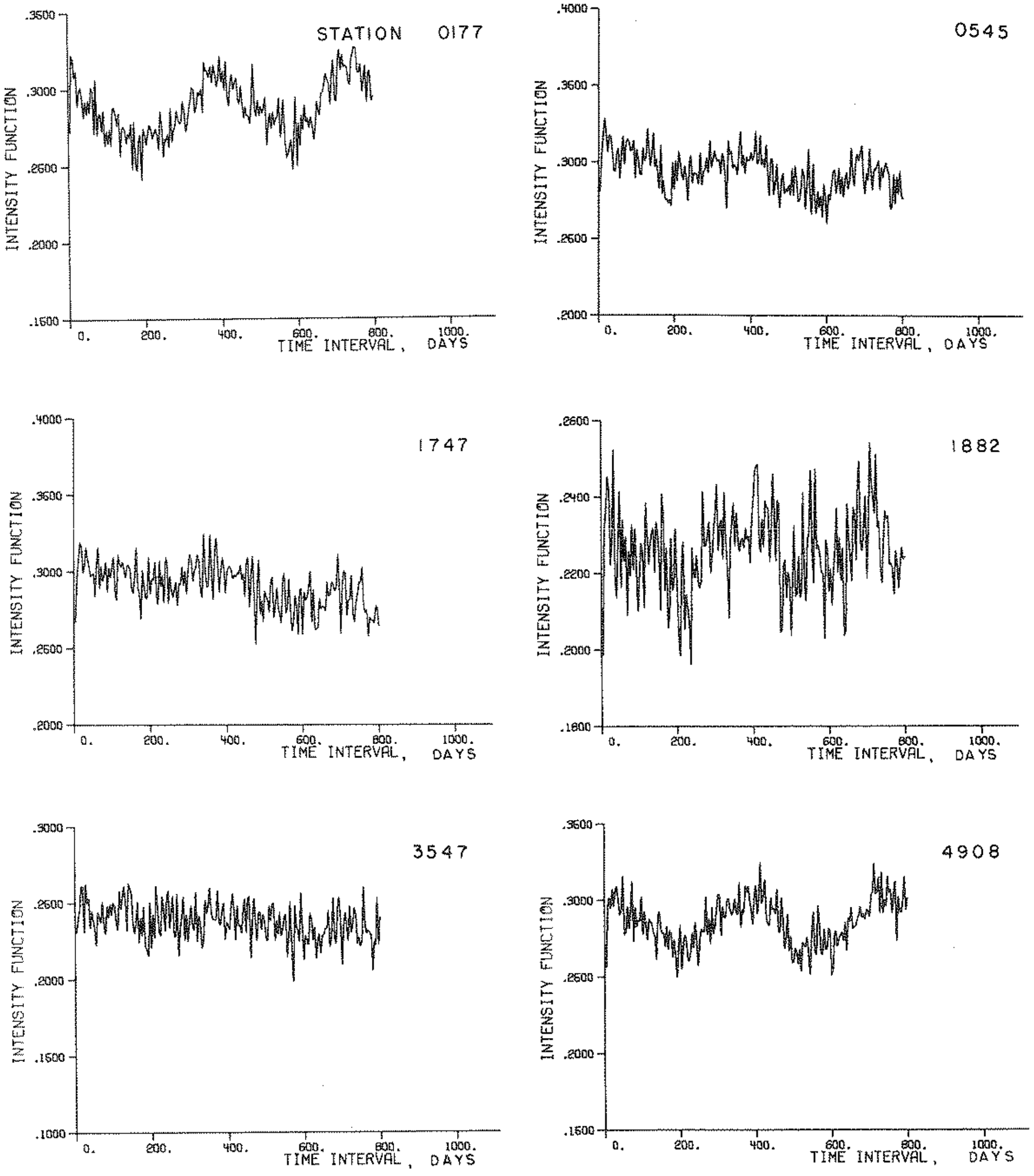


FIG. 3A – INTENSITY FUNCTION OF DAILY RAINFALL

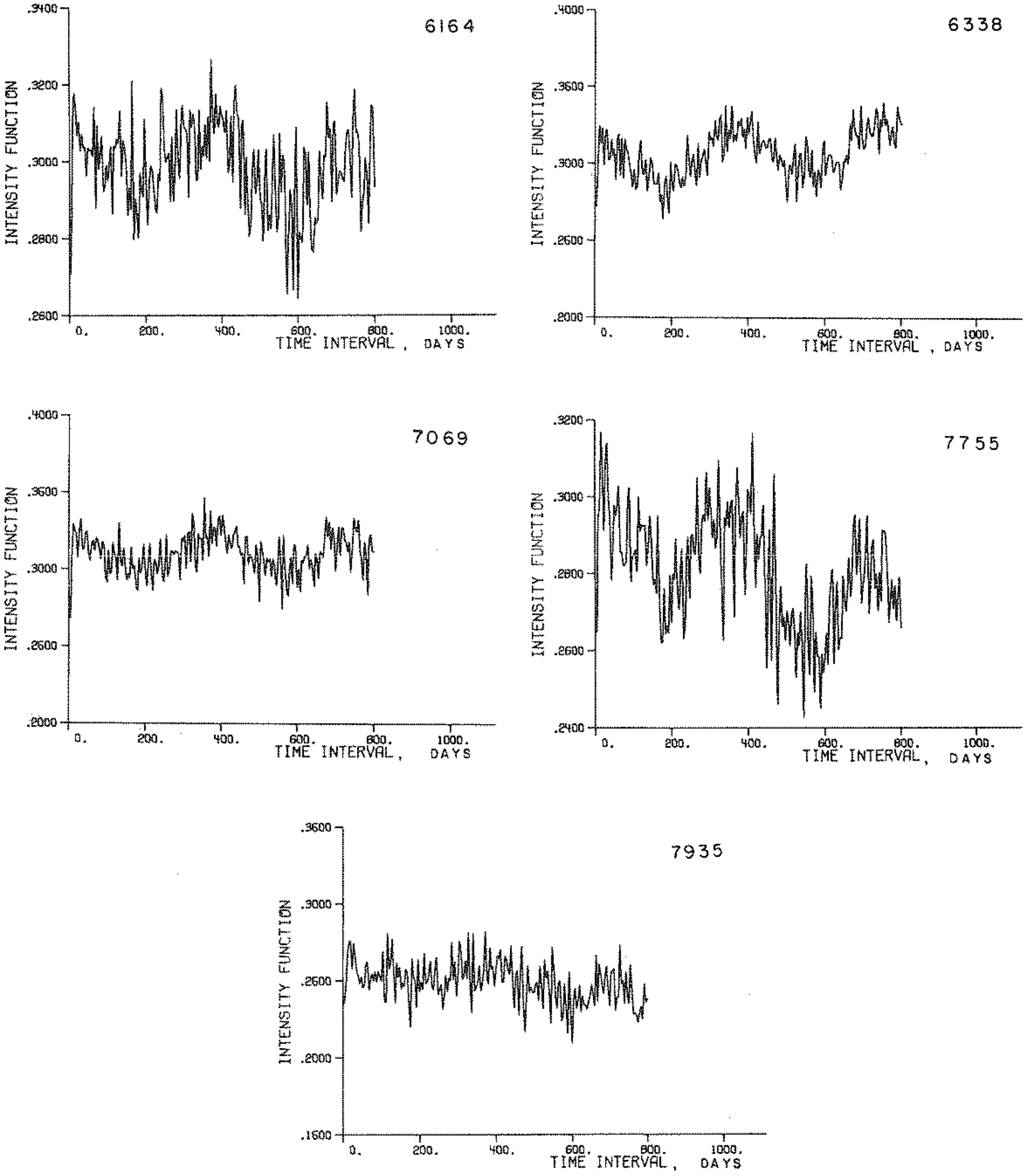


FIG. 3B —INTENSITY FUNCTION OF DAILY RAINFALL

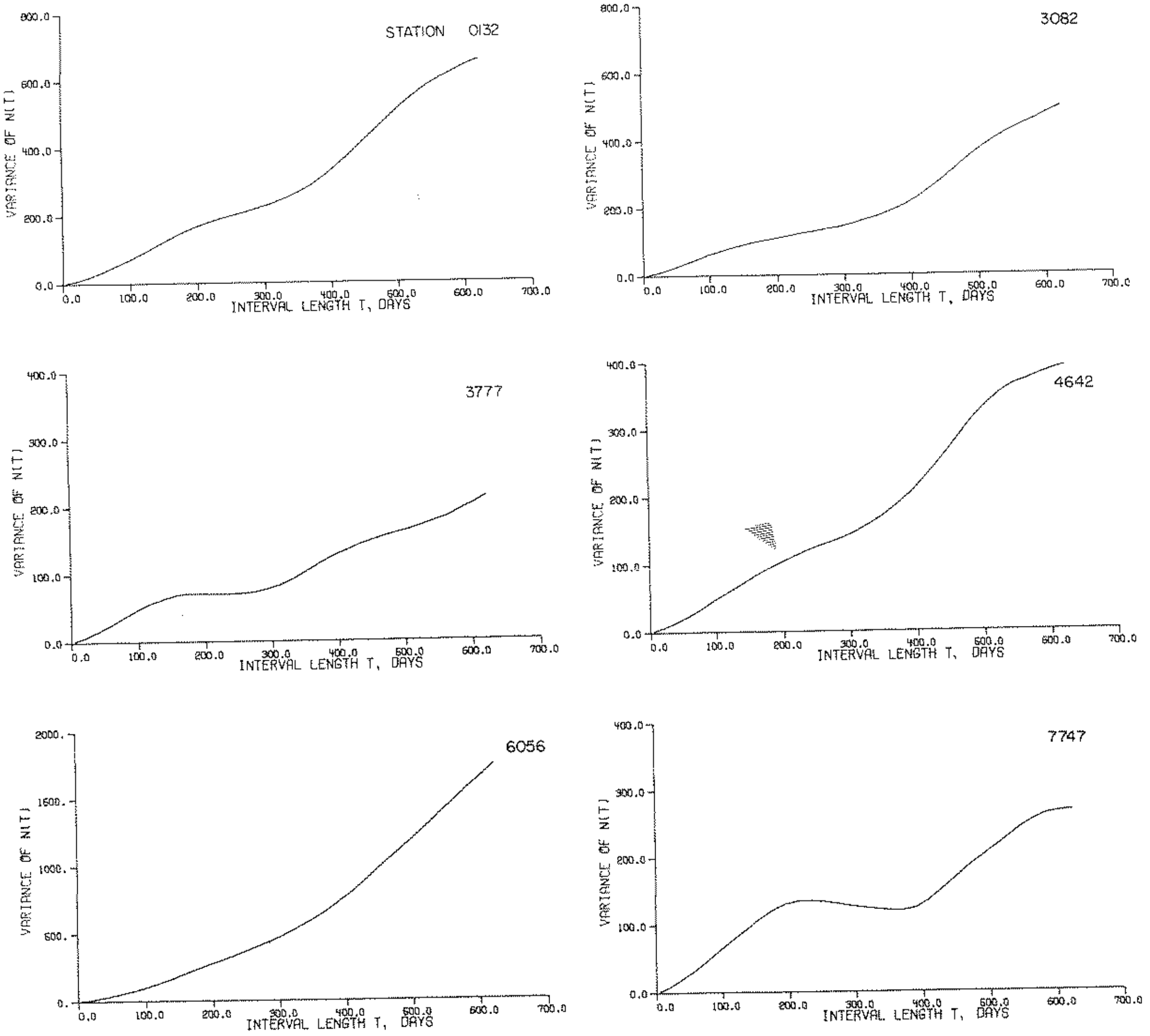


FIG. 4– VARIANCE TIME CURVE FOR COUNTS OF DAILY RAINFALL

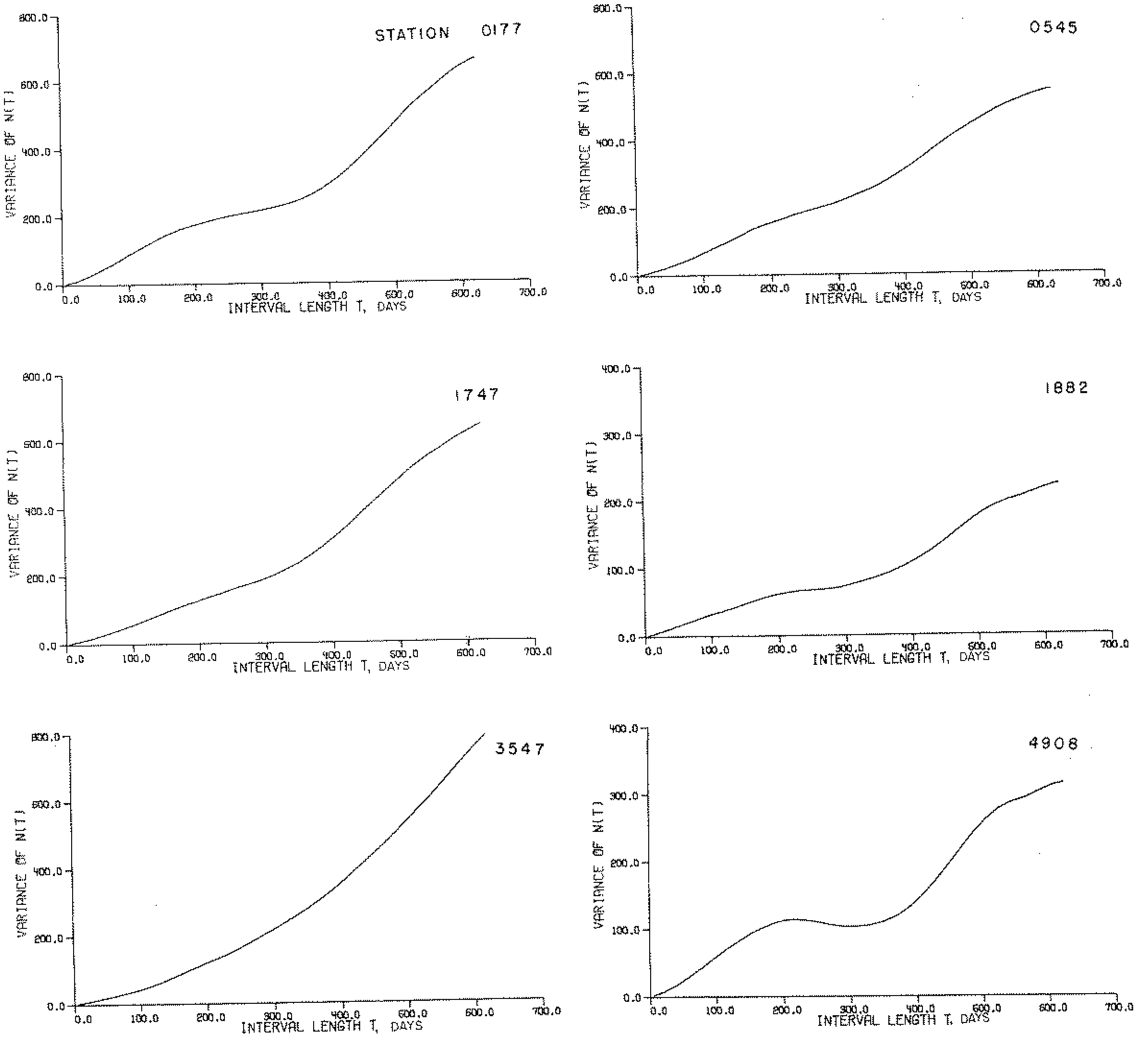


FIG. 4A – VARIANCE TIME CURVE FOR COUNTS OF DAILY RAINFALL

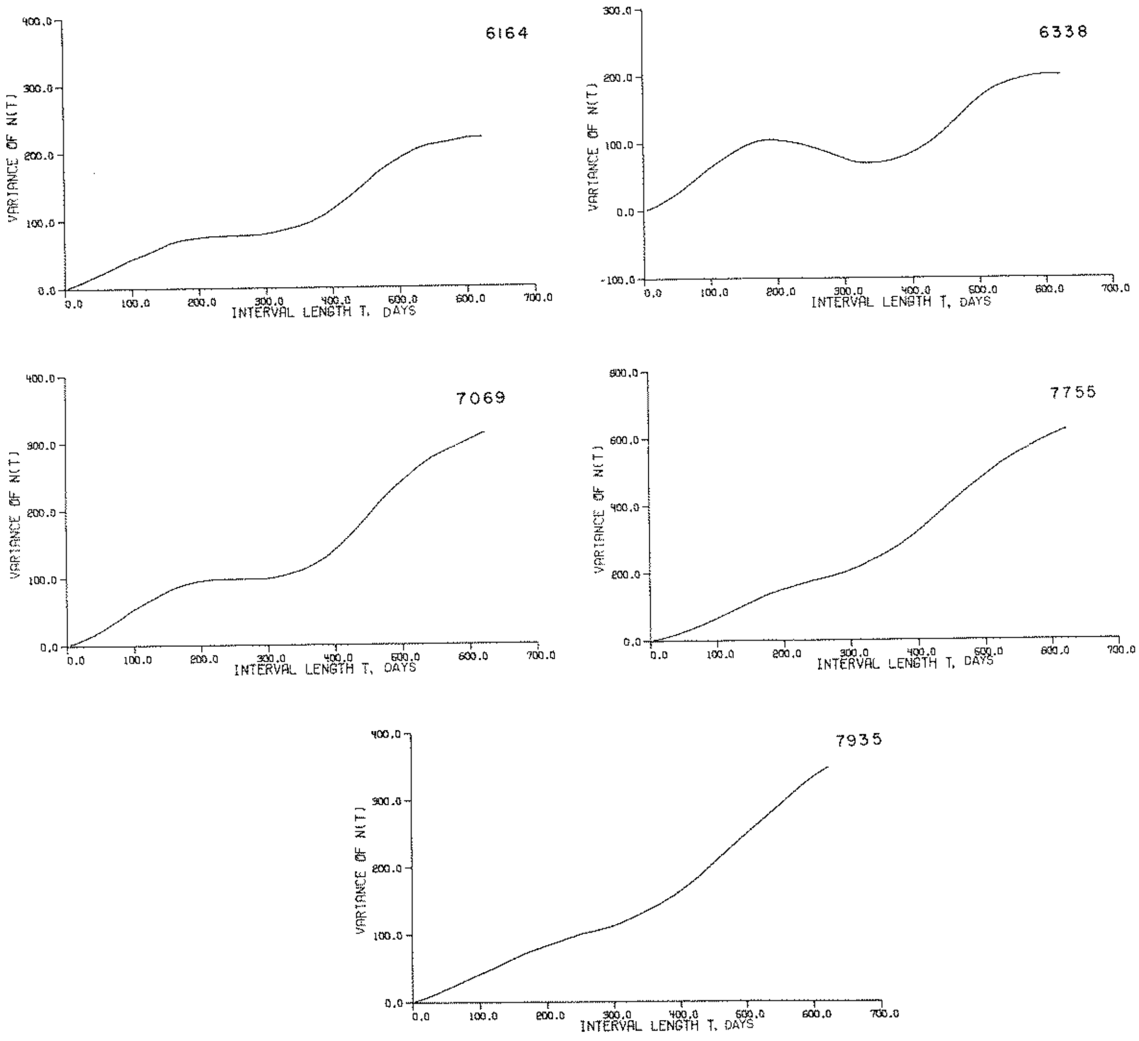


FIG. 4B — VARIANCE TIME CURVE FOR COUNTS OF DAILY RAINFALL

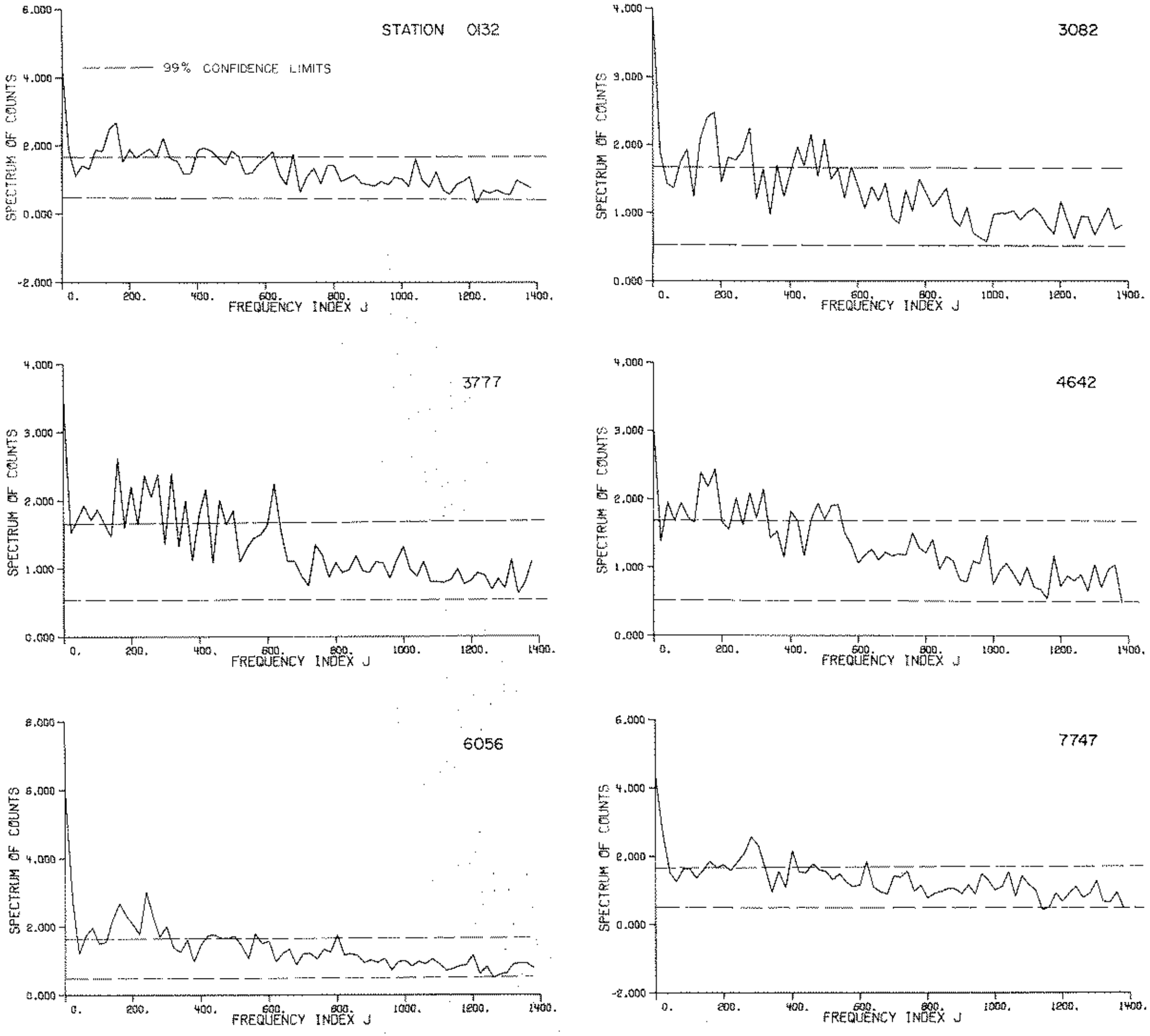


FIG. 5– SPECTRUM OF DAILY RAINFALL COUNTS SMOOTHED BY 20

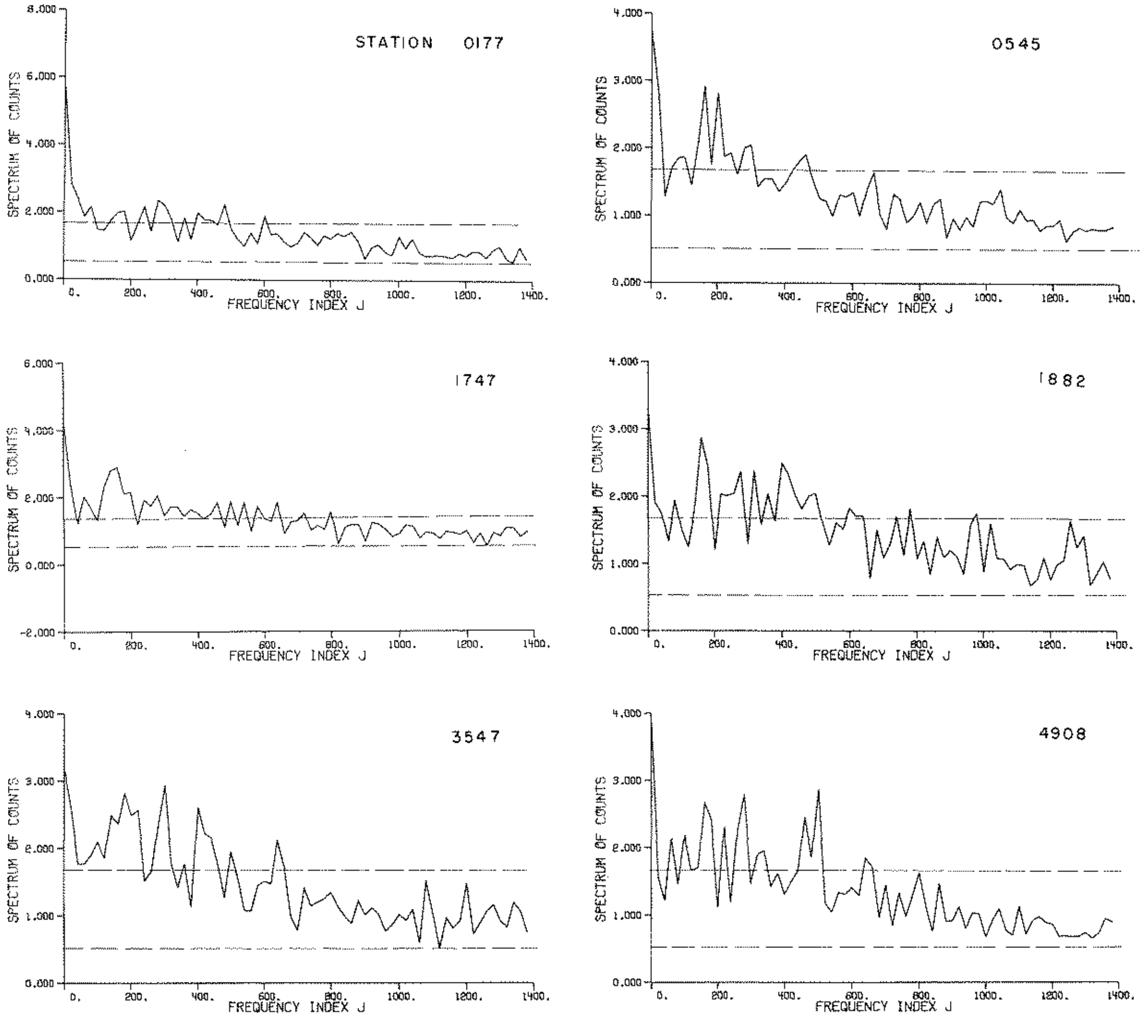


FIG. 5A—SPECTRUM OF DAILY RAINFALL COUNTS SMOOTHED BY 20

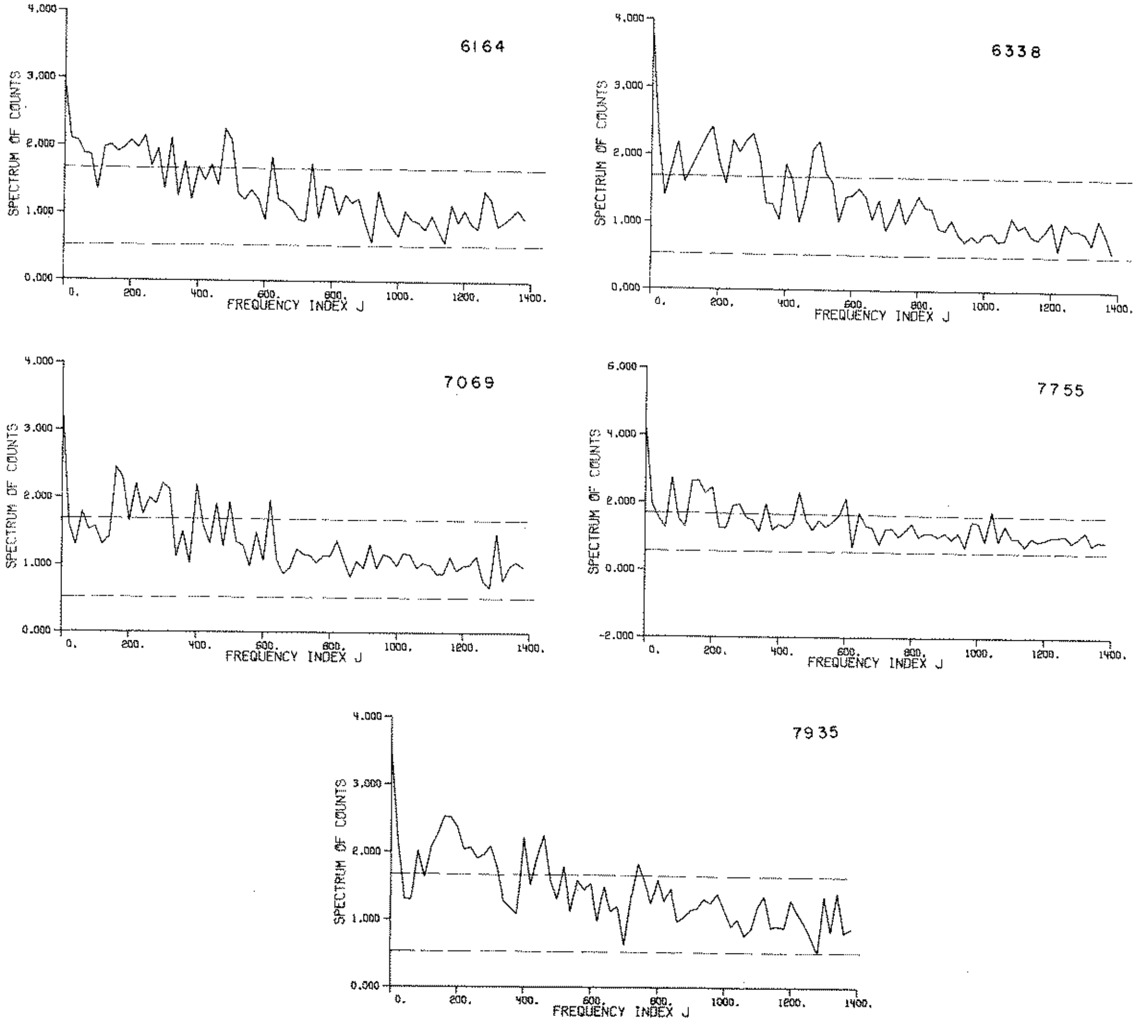


FIG. 5B—SPECTRUM OF DAILY RAINFALL COUNTS SMOOTHED BY 20

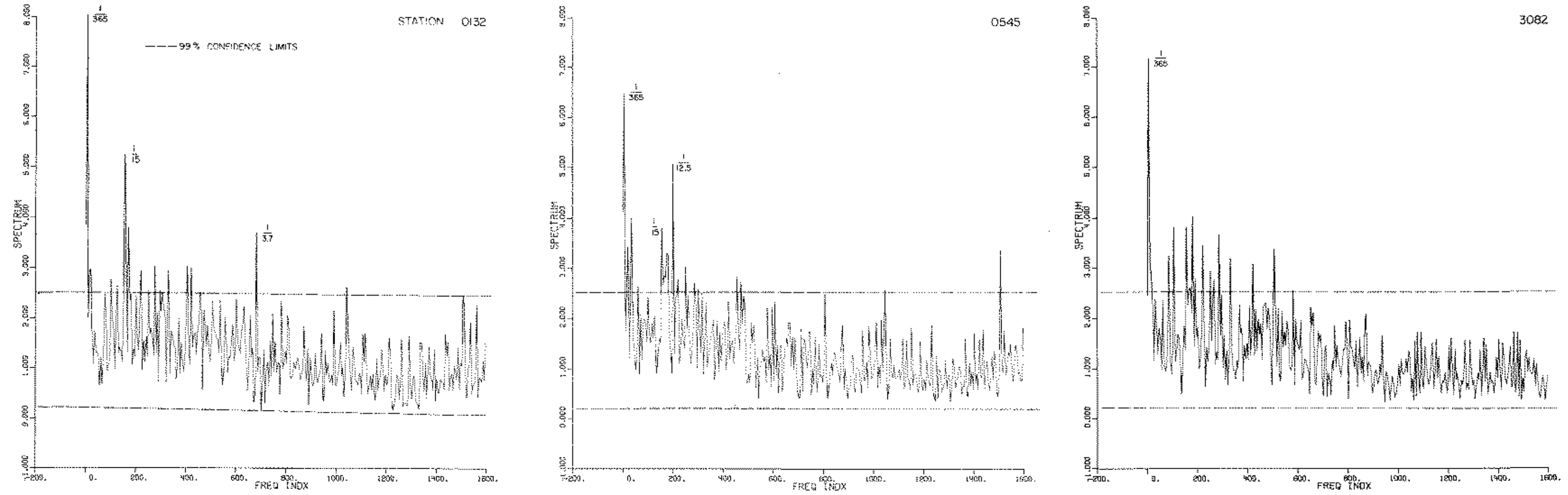


FIG. 6- SPECTRUM OF DAILY RAINFALL COUNTS SMOOTHED BY 5

CHAPTER 3 - HOMOGENIZATION OF THE DAILY RAINFALL COUNTS PROCESS

3.1 THE PURPOSE OF HOMOGENIZATION OF THE DAILY RAINFALL COUNTS PROCESS

As it was seen in the analysis of time trends for the daily rainfall occurrences, there are long term and cyclic time trends in the daily rainfall counts process. The fitting of a stochastic point process model to such a nonstationary phenomenon would necessitate the employment of the statistical tests of hypothesis. However, almost all the classical tests of hypothesis are deviced for the stationary domain. Therefore, even if a nonstationary point stochastic model was constructed for the underlying natural process, it could only be tested in its stationary form. From the practical point of view, the stationary form of the model could be employed for small intervals of time where stationarity of the natural process could be safely assumed. This would mean that in the case of daily rainfall occurrences if a stationary stochastic point process model is to be employed, the length of the time span in which this model could satisfactorily work would at most be four or five months. This span of time would be based on the assumptions that the effect of the long term time trend is minimal in such a small length of time and that the only significant periodicity is the yearly cycle.

As was discussed in the literature survey in section 1.2 the simple Poisson process was used by various hydrologists to model the rainfall occurrence phenomenon. The basic advantage of the Poisson model is its simplicity, especially when it is considered that the practicing hydrologist is not a statistician. The second advantage of the Poisson model is that there is a well developed and simple theory for its nonstationary form, the nonhomogeneous Poisson process, which has the rate of occurrence function as its sole parameter. Therefore, the homogenization scheme will be employed first under the assumption that the daily rainfall counts process can be modeled by a nonhomogeneous Poisson process. Once the data are homogenized, then the tests of the Poisson hypothesis will be employed in the stationary domain. The homogenization scheme will then be employed for more general cases with the purpose of testing the existence of dependence in the daily rainfall occurrence phenomena. The homogenized data will also be employed for the calibration of the parameters of the stochastic models in the stationary domain so as to make inferences about the properties of the underlying rainfall process.

3.2 HOMOGENIZATION BASED ON THE HYPOTHESIS THAT THE DAILY RAINFALL COUNTS PROCESS IS NONHOMOGENEOUS POISSON

3.2.1 The Homogenization Method

The homogeneous Poisson process is a renewal process with exponentially distributed interarrival times. A counting process $N(t)$, $t \geq 0$ is a homogeneous Poisson process if it satisfies the following five axioms (*Parzen*, 1967) :

Axiom 1 $N(0) = 0$.

Axiom 2 $\{N(t),\ t \geq 0\}$ has independent increments. That is $E[Z^{N(t+h)}] = E[Z^{N(t+h)-N(t)}]\ E[Z^{N(t)}]$.

Axiom 3 In any interval h there is a positive probability that an event will occur, no matter how small the interval is. From this axiom it follows that, for a constant occurrence rate λ of $N(t)$, $\lim_{h\to 0} \frac{P[N(t+h) - N(t) \geq 1]}{h} = \lambda$.

Axiom 4 In sufficiently small intervals, at most one event can occur. That is $\lim_{h\to 0} \frac{P[N(t+h) - N(t) \geq 2]}{P[N(t+h) - N(t) = 1]} = 0$.

Axiom 5 $N(t)$ has stationary increments. That is, for $t > s \geq 0$ and $h > 0$, the random variables $N(t) - N(s)$ and $N(t+h) - N(s+h)$ are identically distributed.

These five axioms lead to the probability generating function $\psi(z;t)$ of the Poisson process (*Parzen*, 1967).

$$\psi(z;t) = e^{\lambda t(z-1)} \ , \ |z| < 1 \qquad (3.1)$$

for the constant rate of occurrence λ.

If the first four axioms are preserved but the fifth axiom of stationary increments is changed as to have (*Parzen,* 1967)

$$\lim_{h\to 0} \frac{1 - P[N(t+h) - N(t) = 0]}{h} = \lambda(t)$$

for a time varying rate of occurrence $\lambda(t)$, the counting process $N(t)$ satisfying the new five axioms becomes the nonhomogeneous Poisson process. The probability generating function $\psi(z;t)$ of this process is (*Parzen,* 1967),

$$\psi(z;t) = \exp\{(z-1) \int_0^t \lambda(\tau)d\tau\} \; . \tag{3.2}$$

Therefore, the nonhomogeneous Poisson process has only one parameter, its rate of occurrence function $\lambda(t)$. Any time trends in the daily rainfall counts process would be explained in terms of the behavior of the rate of occurrence $\lambda(t)$ once it is assumed that the underlying stochastic structure is nonhomogeneous Poisson. If a method is devised to transform the natural process into a new domain where the rate of occurrence $\lambda(t)$ is a constant, this method would homogenize the nonhomogeneous Poisson process.

The mean function $M(t)$ of the daily rainfall counting process $N(t)$ is

$$M(t) = \int_0^t \lambda(u)du \tag{3.3}$$

where $\lambda(u)$ is the rate of occurrence of the nonhomogeneous counting process. The mean function $M(t)$ can be assumed to be continuous and differentiable with the derivative equal to $\lambda(t)$. Since $M(t)$ is the integral of $\lambda(u)$, a nonnegative function, it is nondecreasing. Actually $\lambda(u)$ is nonzero in the case of daily rainfall occurrences as is seen from the figure 2, and may be approximated by a continuous function of time. Then the mean function $M(t)$ can be taken to be strictly increasing in the case of daily rainfall occurrences. In order to homogenize the rate of occurrence function $\lambda(t)$, the whole process should be transformed into a new process with a time scale τ where, for a constant unit rate of occurrence,

$$1\cdot\tau = M(t) \tag{3.4}$$

so that τ is a strictly increasing function of t. On the new time scale the rainfall counting process will have the mean function τ and the rate of occurrence equal to unity. Combining equation (3.3) and (3.4),

$$\tau(t) = \int_0^t \lambda(u)du$$

and taking derivative of the both sides

$$\frac{d\tau(t)}{d\tau} = \frac{dt}{d\tau} \cdot \frac{d\int_0^t \lambda(u)du}{dt} \; ,$$

from which one obtains

$$\frac{dt}{d\tau}\lambda(t) = 1$$

and

$$d\tau = \lambda(t)dt \; . \tag{3.5}$$

For very small intervals equation (3.5) can be approximated by the difference equation

$$\Delta\tau = \lambda(t)\Delta t \; . \tag{3.6}$$

Therefore, the interarrival times between the rainfall occurrences are either stretched or squeezed according to the time-varying rate of daily rainfall occurrence to obtain a constant rate of rainfall occurrence of the new time scale τ. Since the rate of rainfall occurrence in the new time scale τ is unity, a homogeneous Poisson process of unit rate is obtained by the transformation (3.6).

In order to carry out the homogenization of the daily rainfall counts process under the Poisson hypothesis, the rate of daily rainfall occurrence has to be modeled and calibrated. This will be the topic of the next section.

3.2.2 The Estimation of the Rate of Daily Rainfall Occurrence Under the Nonhomogeneous Poisson Hypothesis

The sample estimates for the mean rate of daily rainfall occurrence were obtained in the analysis of trends. In figure 2 it can be seen that there is cyclicity and a downward long term trend in the mean rate of daily rainfall occurrence. A model for this type of behavior could be

$$\lambda(t) = \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \sum_{i=1}^{r} R_i \text{ Sin}(\omega_i'' t + \theta_i) \tag{3.7}$$

if it is assumed that there are r significant cycles and the dominant long term trend is quadratic. However, this form for the rate of occurrence would not ensure the positivity of $\lambda(t)$, a condition that has to be satisfied. The two ways to ensure this condition would be (i) to square the right side of (3.7), that is

$$\lambda(t) = \{\alpha_1 + \alpha_2 t + \alpha_3 t^2 + \sum_{i=1}^{r} R_i \text{ Sin } (\omega_i'' t + \theta_i)\}^2 , \tag{3.8}$$

and (ii) to take the right side of (3.7) to an exponential power, that is

$$\lambda(t) = \exp\{\alpha_1 + \alpha_2 t + \alpha_3 t^2 + \sum_{i=1}^{r} R_i \text{ Sin}(\omega_i'' t + \theta_i)\} \tag{3.9}$$

Case 1: When there is a single significant cycle and no long term time trend in the rate of daily rainfall occurrences. This corresponds to assuming that $\lambda(t)$ is of the form

$$\lambda(t) = \exp\{\alpha + R \text{ Sin}(\omega'' t + \theta)\} \tag{3.10}$$

or $\lambda(t)$ is of the form

$$\lambda(t) = \{\alpha + R \text{ Sin}(\omega'' t + \theta)\}^2 . \tag{3.11}$$

Lewis (1970) obtained a closed form solution for the maximum likelihood estimates of e^{α}, R and θ using the form (3.10). For a nonhomogeneous Poisson process in (0,T] the joint density that the number of events is n, and the times to events are $(t_1, \ldots, t_n)$ is

$$f(t_1, \ldots, t_n; n) = \lambda(t_1) \exp\left(-\int_0^{t_1} \lambda(u)du\right) \cdot \lambda(t_2) \exp\left(-\int_{t_1}^{t_2} \lambda(u)du\right) .. \lambda(t_n) \exp\left(-\int_{t_{n-1}}^{t_n} \lambda(u)du\right) \cdot \exp\left(\int_{t_n}^{T} \lambda(u)du\right) . \tag{3.12}$$

Therefore,

$$f(t_1, \ldots, t_n; n) = \prod_{i=1}^{n} \lambda(t_i) \exp\left(-\int_0^T \lambda(u)du\right) . \tag{3.13}$$

Using the open form of $\lambda(t)$ such that

$$\lambda(t) = \exp\{\alpha + R_s \text{ Sin } \omega'' t + R_c \text{ Cos } \omega'' t\} , \tag{3.14}$$

Lewis (1970) obtained the log likelihood function for the nonhomogeneous Poisson process as

$$\ln L(t_1, \ldots, t_n; n) = n\alpha - e^{\alpha} TI_0(R) + R \operatorname{Cos}(\theta) \cdot \sum_{i=1}^{n} \operatorname{Sin}(\omega'' t_i) + R \operatorname{Sin}(\theta) \cdot \sum_{i=1}^{n} \operatorname{Cos}(\omega'' t_i) \quad (3.15)$$

where $R = (R_s^2 + R_c^2)^{1/2}$ and $\theta = \tan^{-1}(R_c/R_s)$ and $I_0(R)$ is the modified Bessel function of the first kind of zero order which is expressed as

$$I_0(R) = \int_0^T \exp\{R \operatorname{Sin}(\omega'' t + \theta)\}dt$$

where $\omega'' = 2\pi J/T$. By differentiating the log likelihood function with respect to α, R, θ and setting it equal to zero, the maximum likelihood estimates of α, R, and θ are obtained as (*Lewis*, 1970, 1972)

$$\hat{\theta} = \tan^{-1} \frac{A_T(\omega'')}{B_T(\omega'')} \quad (3.16)$$

where $A_T(\omega'')$ and $B_T(\omega'')$ are the normalized periodogram components (Chapter 2);

$$e^{\hat{\alpha}} = \frac{n}{T} \frac{1}{I_0(\hat{R})} \quad (3.17)$$

and $\hat{R}$ is the solution of the equation

$$\frac{1}{n} \{A^2(\omega'') + B^2(\omega'')\}^{1/2} = \frac{I_1(\hat{R})}{I_0(\hat{R})} . \quad (3.18)$$

$I_1(\hat{R})$ is the modified Bessel function of the first kind of order 1, and is the derivative of $I_0(\hat{R})$. The ratio $I_1(\hat{R})/I_0(\hat{R})$ increases monotonically from 0 to 1 as $\hat{R}$ goes from 0 to ∞. The very important result based on the maximum likelihood estimates for the parameters of the rate of occurrence function $\lambda(t)$, is that, under the assumption of a single significant cycle and no longer term time trend, $\lambda(t)$ can be directly estimated from the spectrum of daily rainfall counts process when the process is nonhomogeneous Poisson. The rate of daily rainfall occurrence $\lambda(t)$ is easily calibrated by knowing the observation interval (0,T), the number of occurrences, n, in (0,T) and the counts spectrum estimate at the frequency ω'' of the significant cycle.

In the rainfall stations 3777, 6164, 7747 and 7935 the yearly periodicity was much greater than the other significant cycles. Therefore, for these four stations expression (3.10) was used for modeling the rate of daily rainfall occurrence and the parameters α, θ and R were estimated from the spectra of the daily rainfall counts. The calibrated models for the rate of occurrence for these stations is given in Table 3-1. Once the rate of daily rainfall occurrence $\lambda(t)$ is determined, then for this case, the homogenization procedure takes the form

$$\Delta\tau = \exp\{\alpha + R \operatorname{Sin}(\omega'' t + \theta)\}\Delta t . \quad (3.19)$$

The results of the homogenization of the rainfall data in stations 3777, 6164, 7747 and 7935 will be discussed in a later chapter.

Case 2: When there are long term time trends and more than one significant cycle in the daily rainfall counts process:

This case would correspond to a $\lambda(t)$ of the form

$$\lambda(t) = \exp\{\alpha_1 + \alpha_2 t + \sum_{i=1}^{r} R_i \operatorname{Sin}(\omega_i'' t + \theta_i)\} , \quad (3.20)$$

or to the form (3.9) given earlier. The exponential forms (3.20) and (3.9) are quite convenient in the handling of the curvilinear long term time trends. In the form (3.20) it is assumed that there are r significant cycles and a curvilinear long term time trend. For the form (3.20) the log likelihood function under the nonhomogeneous Poisson hypothesis becomes

$$\ln L(t_1, \ldots, t_n; n) = \sum_{j=1}^{n} \{\alpha_1 + \alpha_2 t_J + \sum_{i=1}^{r} R_i \ \mathrm{Sin}(\omega_i'' t_j + \theta_i)\} - \int_0^T \exp\{\alpha_1 + \alpha_2 t + \sum_{i=1}^{r} R_i \ \mathrm{Sin}(\omega_i'' t + \theta_i)\}dt \ . \quad (3.21)$$

The solution of the maximum likelihood estimates requires the integration of the exponential integral in (3.21). However, no simple analytical integration of this integral could be found. Therefore, the estimation of the parameters of the forms (3.20) and (3.9) were done by the least squares method.

In this study a computer program written by *Marquardt* (1966) was used for the least squares estimation of the parameters of the rate of daily rainfall occurrence model (see Appendix C).

As is seen in Table 2-1, the spectral analysis of the daily rainfall counts process yielded more than one significant periodicity for each of the 17 stations analyzed. Also the presence of a long term time trend was established from the graphical analyses of trends in the rainfall counts, and from the test of trend in the rate of rainfall occurrence under the Poisson hypothesis. Therefore, the forms (3.20) and (3.9) of the model of the rate of daily rainfall occurrence function, $\lambda(t)$, are appropriate. These exponential forms can take into account both the linear and the curvilinear long term time trends. Models (3.20) and (3.9) were fitted to the rate of daily rainfall occurrence data by Marquardt's algorithm with various combinations of significant cycles detected by the spectral analysis of counts. The models were compared for their R^2 values, where $R^2 = \Sigma(\hat{Y}_i - \bar{Y})^2/\Sigma(Y_i - \bar{Y})^2$ in which $\hat{Y}$, Y and $\bar{Y}$ are the estimated, the actual and the mean value of the dependent variable in the regression, respectively. The models which yielded the highest R^2 values for each station and model (3.10) constructed for the case of a single significant cycle and no long term trend are given in Table 3.1 for each of the 17 stations studied. The fitted rate of occurrence functions and the sample rate of occurrence functions are shown in Figure 2 for stations 0132, 3082, 3777, 4642, 6056 and 6338 in Indiana. Model (3.9) yielded a better fit than model (3.20) in the case of stations 0545, 3082, 6056 and 6338 as can be seen from Table 3-1. It is interesting to note that for stations 0545, and 6056 the seven-year period which was found significant in the spectral analysis of counts corresponds to the total time of observation limited to seven-years by the computer storage size. The quadratic term, introduced to the rate of occurrence model, accounts for this period and verified the earlier speculation that a seven-year periodicity is artificial and only shows the effect of the long term time trends.

Once the rate of occurrence function $\lambda(t)$ for the case 2 of several periodicity and long time trends is determined, the homogenization scheme given by equation (3.6) is rewritten as

$$\Delta\tau = \exp\{\alpha_1 + \alpha_2 t + \alpha_3 t^2 + \sum_{i=1}^{r} R_i \ \mathrm{Sin}(\omega_i'' t + \theta_i)\}\Delta t \ . \qquad (3.22)$$

The daily rainfall counts process for the stations that fall into the second case was homogenized through the use of this scheme. The results of the homogenization are analyzed in Section 3.4.

3.3 HOMOGENIZATION OF THE DAILY RAINFALL COUNTS PROCESS UNDER THE ASSUMPTION THAT THERE IS A DEPENDENCE STRUCTURE UNDERLYING THE COUNTING PROCESS

It was discussed in Section 3.2.1 that the counting process with independent increments is a Poisson process. If the results of the statistical tests in the stationary domain disprove the Poisson hypothesis, then the increments of the counting process are dependent. That is

$$E[z^{N(t+h)}] \neq E[z^{N(t+h)-N(t)}] \ E[z^{N(t)}]$$

when there is dependence in the counting process increments. The two models that could explain the dependence structure in the counting increments are (i) the nonhomogeneous Markov chain model (*Parzen,* 1967) and

TABLE 3-1

MODELS FOR THE REPRESENTATION OF DAILY RAINFALL OCCURRENCE RATE FUNCTION $\lambda(t)$

Sta.		Cycles in the Model	R^2
0132	$\hat{\lambda}(t) = .344\ \mathrm{Exp}\{-.0001t - .284\ \mathrm{Sin}(.017t-90.4) - .034\ \mathrm{Sin}(.391t+5.68)\}$	365,16,2.7	.31
0177	$\hat{\lambda}(t) = .301\ \mathrm{Exp}\{-.00009t - .349\ \mathrm{Sin}(.017t+3.3) + .188\ \mathrm{Sin}(.0344t+2.94) + .109\ \mathrm{Sin}(.09t+4.08)\}$	365,183,69	.39
0545	$\hat{\lambda}(t) = .351\ \mathrm{Exp}\{- .0003t - .00000014t^2 - .25\ \mathrm{Sin}(.017t+3.42) - .12\ \mathrm{Sin}(.09t+1.86) - .033\ \mathrm{Sin}(.49t+2.17)\}$	365,71,12.7	.29
1747	$\hat{\lambda}(t) = .328\ \mathrm{Exp}\{-.00009t - .236\ \mathrm{Sin}(.017t+3.34) + .097\ \mathrm{Sin}(.39t+1.8)\}$	365,16	.27
1882	$\hat{\lambda}(t) = .255\ \mathrm{Exp}\{-.0001t - .22\ \mathrm{Sin}(.017t+3.04) - .137\ \mathrm{Sin}(.042t-2.13) - .042\ \mathrm{Sin}(.422t+2.12)\}$	365,150,14.9	.29
3082	$\hat{\lambda}(t) = .397\ \mathrm{Exp}\{-.00039t - .0000001t^2 - .246\ \mathrm{Sin}(.017t+3.12)\}$	365	.33
3547	$\hat{\lambda}(t) = .26\ \mathrm{Exp}\{-.000075t + .27\ \mathrm{Sin}(.0024t+2.8) - .133\ \mathrm{Sin}(.017t+2.94) + .0715\ \mathrm{Sin}(.447t+.917)$		.22
	$+ .0766\ \mathrm{Sin}(.78t+2.28)\}$	2556,365,14,8	.22
3777	$\hat{\lambda}(t) = .30\ \mathrm{Exp}\{+ .21\ \mathrm{Sin}(.017t+1.09)\}$	365	
4642	$\hat{\lambda}(t) = .36\ \mathrm{Exp}\{-.00016t - .21\ \mathrm{Sin}(.017t+3.32) - .052\ \mathrm{Sin}(.39t-6.85)\}$	365,16	.33
4908	$\hat{\lambda}(t) = .33\ \mathrm{Exp}\{-.00012t - .3\ \mathrm{Sin}(.017t+3.23) + .026\ \mathrm{Sin}(.73t+3.88) + .077\ \mathrm{Sin}(1.24t+.14)\}$	365,8.6,5	.51
6056	$\hat{\lambda}(t) = .4\ \mathrm{Exp}\{-.00066t + .0000002t^2 - .26\ \mathrm{Sin}(.017t+3.15) + .12\ \mathrm{Sin}(.059t+1.02) - .06\ \mathrm{Sin}(.62t-10.1)\}$	365,122,10.2	.44
6164	$\hat{\lambda}(t) = .3\ \mathrm{Exp}\{.247\ \mathrm{Sin}(.017t+.133)\}$	365	
6338	$\hat{\lambda}(t) = .285\ \mathrm{Exp}\{.0002t - .00000011t^2 - .303\ \mathrm{Sin}(.017t+3.37)\}$	365	.35
7069	$\hat{\lambda}(t) = .358\ \mathrm{Exp}\{-.00013t + .24\ \mathrm{Sin}(.017t+.247) + .03\ \mathrm{Sin}(.42t-.748)\}$	365,14.9	.38
7747	$\hat{\lambda}(t) = .3\ \mathrm{Exp}\{.298\ \mathrm{Sin}(.017t+.356)\}$	365	
7755	$\hat{\lambda}(t) = .268\ \mathrm{Exp}\{.00002t - .29\ \mathrm{Sin}(.017t+3.54) + .14\ \mathrm{Sin}(.214t+2.79) - .067\ \mathrm{Sin}(.428t+7.99)\}$	365,29.4,14.7	.34
7935	$\hat{\lambda}(t) = .251\ \mathrm{Exp}\{.236\ \mathrm{Sin}(.017t+.36)\}$	365	

(ii) the cluster processes (*Neyman and Scott*, 1958). This study will address itself to the modeling of the rainfall counts process by the cluster model in the case that the Poisson model for the daily rainfall counts process is disproved by the statistical tests in the stationary domain.

It will be shown in a later chapter that the rate of occurrence function for the stationary Neyman-Scott cluster model is

$$\lambda = \frac{dE[N_t]}{dt} = \alpha E(\nu) \tag{3.23}$$

where α is the rate of occurrence of the rainfall generating mechanisms or the rate of occurrence of the "parent" process while ν is the cluster or the "group" size or the number of first-generation offsprings. The transformation (3.6) that was specifically designed for the nonhomogeneous Poisson process would remove the time trends in the first moment of the cluster model since

$$E[N_t] = \alpha E(\nu)t \tag{3.24}$$

for the stationary cluster model. However, as it will be seen in the detailed analysis of the cluster model in a later chapter, nothing can be said about the removal of trends in the higher moments. However, even if only the first moment is homogenized, the spectrum of counts can be used effectively for the calibration and the testing of the stationary form of the Neyman-Scott cluster model. It will be shown in the detailed analysis of the Neyman-Scott model that under the assumption that the cluster members are negative-exponentially distributed from the origin of the cluster, the spectrum of counts takes the form

$$g_+(\omega) = \frac{\alpha E(\nu)}{\pi} + \frac{\alpha E(\nu^2 - \nu)}{\pi} \cdot \frac{\theta^2}{\theta^2 + \omega^2} \qquad \omega \geq 0 \tag{3.25}$$

where θ and ν are parameters of the storm structure while α is the rate of occurrence of storms. If it is further assumed that only the parent process or "the process of the rainfall generating mechanisms" is nonhomogeneous while the process of the cluster members or "the process of the rainfalls in a storm" is stationary, then only the rate of occurrence of the parent process, α, will be time-dependent in the spectrum of counts. In this case, once the rate of parent occurrence, $\alpha(t)$, is homogenized, the spectrum of counts can be effectively used for the testing of the assumptions and the calibration of the model.

For a complete homogenization scheme under the Neyman-Scott cluster assumption, the nonhomogeneous form of the model should be known. Since this is lacking, the best one can do is to analyze the results of the transformation (3.6) by the use of various statistical functions introduced in the analysis of trends.

3.4 RESULTS OF THE HOMOGENIZATION OF THE DAILY RAINFALL COUNTS PROCESS

The daily rainfall counts data was homogenized by the use of equation (3.6) with Δt = 1 day. The homogenization scheme reduced the interarrival times in all of the 17 stations analyzed. The record length was reduced from 2556 days on the average to 750 days on the average. The new time scale τ does not have a physical time meaning. However, if there are any periodicities or trends left in the data, they can be detected by the following graphical analysis and by the tests of trend hypothesis discussed in the next subsection. Finally the change in the moments of the interarrival distribution through the use of (3.6) will be given and the consequences of the transformation will be discussed.

3.4.1 The Graphical Analysis of Homogenization

3.4.1a Number of rainy days versus cumulative time

In the analysis of time trends in the daily rainfall occurrences it was seen in fig. 1 that the plots of the total number of daily rainfall occurrences, denoted by "interval number," versus the total time in days to the last occurrence, denoted by "cumulative time," had generally increasing upward slopes with time.

The increase in the upward slope indicated a decrease in the rate of daily rainfall occurrence. After homogenization, using the scheme (3.6) with the rate of occurrence models listed in table 3-1 for the corresponding stations, it was seen that in the transformed time scale, τ, the slopes became constant. Since the slope of the plot is the inverse of the mean rate of daily rainfall occurrence $\lambda(t)$, these plots showed that the scheme (3.6) effectively homogenizes the rate of rainfall occurrence. Figure 7 for the rainfall stations 0132, 3082, 3777, 4642, 6056 and 7747, when compared to fig. 1, show the change in slope, especially in the stations 3082 and 4642. Since the number of rainy days versus cumulative time is a cumulative plot, the changes in the rate of daily rainfall are smoothed out, and the effects of the transformation (3.6) on $\lambda(t)$ are not clearly apparent.

3.4.1b Intensity function

Figure 8 shows the intensity functions for the homogenized daily rainfall counts data for stations 0132, 3082, 3777, 4642, 6056 and 7747. After homogenizations the "time interval" in the abcissa loses its physical time meaning. However, it is important to note that the original time scale was modified so that for example 800 in the new time scale corresponds to 800X2556/750 or 2726 days in the natural time scale. It was shown in equation (2.8) that for the homogeneous Poisson process the intensity function is equal to the constant rate of rainfall occurrence. The data were homogenized so as to obtain a rate of rainfall occurrence equal to unity. Therefore, the sample intensity function should be close to unity for all values of τ.

When fig. 8 is compared to fig. 3 for the corresponding stations, it will be seen that the cyclicities are, in general, effectively removed. However, a cyclicity still appears in the intensity function for the station 0132 corresponding to about 600 x 2556/750 or 2045 days in the physical time scale. This would correspond to roughly 5.6 years, an artifact. The reason for this artifact is that the values of $\bar{m}_f(\tau)$ for τ large have very high variations because as τ is increased $\bar{m}_f(\tau)$ is estimated from fewer and fewer points. This effect is clearly seen for large values of τ in fig. 8 since all the data length was utilized for the computation of $\bar{m}_f(\tau)$. In all the cases studied the intensity function was close to unity, expecially in the beginning for small values of τ. It is difficult to conclude from the intensity function whether the deviations from the horizontal line $\bar{m}_f(\tau) = 1$ show deviations from the Poisson hypothesis or they are just due to the increase in the variance of the $\bar{m}_f(\tau)$ estimates with increasing τ.

3.4.1c Variance-time function

Variance-time curves for the rainfall stations 0132, 3082, 3777, 4642, 6056 and 7747 are shown in fig. 9. The interval length is in terms of the homogenized process time scale τ. The variance-time function was computed up to a time τ_0 which was one-fifth of the time length of the homogenized data, in order to avoid the increased variance due to the decrease in the number of degrees of freedom in the estimates. As is seen from the figures, the cyclicity in the variance-time function of the original data was removed. However, the original variance-time function was distorted.

For a homogeneous Poisson process it was shown earlier that $V(\tau) = \lambda\tau$. The mean function for the rainfall occurrence, $E[N_\tau]$, would again be $\lambda\tau$ under the homogeneous Poisson hypothesis. Therefore, the coefficient of variation function $C(\tau) = V(\tau)/E[N_\tau]$ can be used to draw inferences about the dispersion of the homogenized daily rainfall occurrences. In the case of the Poisson process $C(\tau) = 1$. If $C(\tau) > 1$, this would indicate to an overdispersion of the occurrences with respect to the Poisson process and would indicate to a grouping of the rainfall events, of the clustering. If $C(\tau) < 1$, the occurrences are underdispersed and too regular when compared to the Poisson case. $C(\tau)$ is measured indirectly by comparing the $V(\tau)$ function plotted for the homogenized daily rainfall counts data with the theoretical $V(\tau)$ function for the Poisson case plotted on the same graph. Since in the scheme (3.6) τ was taken to be unity, $V(\tau) = \tau$ under the Poisson hypothesis. If the $V(\tau)$ for the homogenized data is above the straight line $V(\tau) = \tau$, then $C(\tau) > 1$. If the $V(\tau)$ for the homogenized data coincides with $V(\tau) = \tau$, $C(\tau) = 1$, and if it is below

$V(\tau) = \tau$, $C(\tau) < 1$.

The most common case is for the station 0132 where $C(\tau) > 1$. This clustering effect was also noticed in the stations 0132, 0177, 0545, 1747, 3777, 6056, 6338, 7069, 7755 and 7935 out of the 17 stations whose data was homogenized. This clustering of the rainfall occurrences in the form of storms around the storm-causing mechanisms can be explicitly modeled with the Neyman-Scott stochastic cluster model. This model will be described in detail in a later chapter.

In fig. 3 for the station 6056, the plot of $V(\tau)$ vs. τ may be approximated by a straight line with a slope greater than unity. The variance-time function for such a case would correspond to a compound Poisson process where a random number of rainfall events occur at the instants of a simple Poisson process. This would correspond to the thunderstorm activity where the interarrival times are much longer than the size of the storm. A stochastic process $\{N(\tau), \tau \geq 0\}$ for the rainfall counts is a compound Poisson process if it can be represented by

$$N(\tau) = \sum_{i=1}^{n(\tau)} Z_i \ , \ \tau \geq 0 \qquad (3.26)$$

where $\{n(\tau), \tau \geq 0\}$ is a Poisson process counting the thunderstorms up to the time τ while Z_i is the number of rainfalls at the i-th thunderstorm. If the mean rate of occurrence of the thunderstorms is α, the mean function and the variance-time function of the homogenized daily rainfall occurrence data are (*Parzen*, 1967),

$$E[N(\tau)] = \alpha E(Z)\tau \ , \qquad (3.27)$$

$$Var[N(\tau)] = \alpha E(Z^2)\tau \ , \qquad (3.28)$$

so that $C(\tau) = E(Z^2)/E(Z)$, a constant. In the analysis of the Neyman-Scott cluster model, it will be seen that the compound Poisson process is a special case of the cluster model.

In fig. 9 for station 4642 it is seen that the homogenized daily rainfall counts are underdispersed. The same occurs for stations 1882, 3082 and 4908. This behavior of the variance-time function can be explained by the equation (2.22) for the $V(\tau)$ for an ordinary renewal process with gamma distributed interarrival times, X, with $E(X) = 2/\lambda$. However, the intensity function for this model, which was given in equation (2.10), does not quite correspond to the sample intensity function for the station 4642, given in the fig. 8. The underdispersion may also be due to an over-removal of the dominant yearly periodicity in the daily rainfall counts process. This hypothesis will be strengthened when the counts spectra of the homogenized data is analyzed in the subsequent section.

Only stations 3547, 6164 and 7747 satisfied the Poisson hypothesis in terms of the variance-time behavior. The Poisson case is shown on the fig. 9 for the Station 7747.

From equation (2.27) the asymptotic slope of the variance-time curve gives the value of the spectrum of the rainfall counts at the origin. Therefore, the asymptotic slope of the variance-time curve yields important information about the long range dependence properties of the daily rainfall counts process. In the time series analysis of the hydrologic data the relationship between the autocorrelation structure of the time series and the variance-time curve was established. An important statistic in *Hurst's* (1951, 1956, 1965) analysis of the long range dependence in geophysical time series was the variance-time curve. Likewise the variance-time curve for the daily rainfall counts can explain the long range dependence in the rainfall occurrences.

In the calibration of a stochastic model of the daily rainfall counts the long range dependence can be preserved by fitting the asymptotic slope of the variance-time curve. However, the data for the analysis of the daily rainfall counts in Indiana were limited to 7 years due to the computer storage limitations. In the homogenization process the time scale of the data was further reduced, and in the variance-time function estimation only one-fifth of the homogenized data could be used in order to avoid increased variance. The span for which the variance-time function could be calculated is too short to have the transient in the data to die out and for observing the asymptotic linear slope accurately. A record of much larger size is thus necessary to observe the long range dependence in the daily rainfall counts by means of the variance-time curve of the homogenized data.

3.4.1d Spectrum of the homogenized daily rainfall counts

The counts spectrum of the homogenized data is the most important statistical function for the verification of the removal of the periodicities of the scheme of equation (3.6), for the verification of the Poisson hypothesis, for the detection of the correlation structure in the daily rainfall counts process, and for the calibration of the stochastic model that can explain the correlation structure. The spectrum of counts is defined as the Fourier transform of the covariance density $\gamma_+(u)$ of the differential counting process $\{\Delta N_t\}$. The differential counting process is defined by *Cox and Lewis* (1966) as the counting process where ΔN_t is the number of events in $(t, t+\Delta t)$, that is, $\Delta N_t = N_{t+\Delta t} - N_t$ as $\Delta t \to 0$. It is an instantaneous process with value zero almost everywhere, except at the points of the random occurrences. Then the definition of $\gamma_+(u)$ is that

$$\gamma_+(u) = \lim_{\Delta t \to 0} \frac{\mathrm{Cov}\{\Delta N_t, \Delta N_{t+u}\}}{(\Delta t)^2} \tag{3.29}$$

so that the spectrum of counts $g(\omega)$ is defined as (*Cox and Lewis,* 1966)

$$g(\omega) = \frac{m}{2\pi} + \frac{1}{2\pi}\int_{-\infty}^{+\infty} \gamma_+(u)\, e^{-iu\omega} du \qquad -\infty < \omega < \infty \tag{3.30}$$

where $m/2\pi$ is the contribution of the covariance density at lag zero. *Cox and Lewis* (1966) show that

$$\gamma_+(u) = m\{m_f(u) - m\} \tag{3.31}$$

so that the expression (3.30) becomes identical to (2.24). Expression (3.30) shows that the spectrum of counts explains the correlation structure of the counting process $\{N(t)\}$ through the use of the differential counting process $\{\Delta N(t)\}$. This property will be used to make inferences about the dependence structure of the homogenized daily rainfall counts process.

The spectrum of counts analysis for the Poisson hypothesis was done for the homogenized daily rainfall counts data. In all of the 17 stations, there were significant deviations from the Poisson hypothesis. For the Poisson hypothesis $g_+(\omega) = \lambda/\pi$. The normalized spectrum of counts $\pi g_+(\omega)/\lambda$ is thus a constant horizontal line with the ordinate equal to unity. The 99% confidence limits for the Poisson hypothesis were constructed for the estimates smoothed in consecutive groups of 20. The smoothed spectral estimates consistently showed an exponential decay from lower to higher frequencies. They consistently deviated from the Poisson hypothesis, even at 1% level. The spectra of the homogenized daily rainfall counts for stations 0132, 3082, 4642, 6056 and 7747 are shown on fig. 10 together with the theoretical spectra of the Poisson and cluster models and the 99% confidence limits for the Poisson hypothesis.

The immediate conclusion that can be drawn from these spectra is that the daily rainfall counts process does not have independent increments. There is a definite correlation structure in the homogenized daily rainfall counts process. The Poisson hypothesis of the independent rainfall counts should be rejected. A stochastic model that can explain the correlation structure in the daily rainfall counts process is necessary. The Neyman-Scott cluster process is such a model, which not only can explain the dependence structure but also can physically describe the grouping of the rainfall events in the form of storms and the occurrence of the storm-generating mechanisms.

If the fig. 10 is compared to its counterpart fig. 5, it will be seen that the yearly periodicity is significantly reduced in all the cases by the homogenization scheme (3.6). In fact, the spectral component corresponding to the yearly periodicity was over-removed in the cases of the stations 3082 and 4642. Since the estimates were smoothed in consecutive groups of 20, the first value of the spectrum after the origin corresponds to the contributions from the yearly cycle and the longer cycles. It is seen that this contribution is over-removed for stations 3082 and 4642.

It was speculated in the discussion of the variance-time function results that the homogenized daily rainfall counts process for the stations 3082 and 4642 could be explained by an ordinary renewal process with gamma distributed interarrival times, X, and with $E(X) = 2/\lambda$. The spectrum of counts for this case is given as (*Cox and Lewis,* 1966).

$$g_+(\omega) = \frac{\lambda(\omega^2 + 2\lambda^2)}{2\pi(\omega^2 + 4\lambda^2)}, \qquad \omega \geq 0$$

so that $\pi g_+(\omega)$ increases monotonically from $\pi g_+(0^+) = \lambda/4$ to $\pi g_+(\omega) = \lambda/2$. However, the spectra of the homogenized daily rainfall counts are decreasing rather than increasing. Therefore, the ordinary renewal process model is rejected. The second possibility of the over-removal of the yearly cyclicity seems plausible from the spectra of the fig. 10.

In the variance-time analysis of the homogenized data the compound Poisson process was considered for the station 6056. It will be seen in the analysis of the Neyman-Scott cluster model that the spectrum of the homogenized daily rainfall counts satisfying the compound Poisson process hypothesis is a constant horizontal line. The spectrum of counts for the station 6056, shown on the fig. 10 is exponentially decaying rather than being a horizontal line. Therefore, the homogenized daily rainfall counts at the station 6056 should be modeled with the more general Neyman-Scott model which has the exponentially decaying spectrum of counts.

3.4.1e Relative frequency histogram and the log-survivor function for the rainfall interarrival times

The relative frequency histograms of the interarrival times between the daily rainfall occurrences were constructed for the 17 stations in Indiana. If there are n observed x_i's, $i = 1, \ldots, n$, the relative frequency histogram $f_X(x_i)$ is calculated by

$$f_X(x_i) = \frac{\text{number of } x_i\text{'s}}{n}, \quad i = 1, 2, \ldots . \tag{3.32}$$

Since the relative frequency histogram was constructed for the daily rainfall data, the probabilities are concentrated on the days. The times between the integer days are naturally void. This is seen in fig. 11 for the stations 0132, 3082, 3777, 4642, 6056 and 7747 respectively. The shape of the relative frequency histogram, ignoring the voids, is of the exponential type, which agrees with the Poisson hypothesis of the exponentially distributed interarrivals.

A second way to look at the probability distribution of the interarrivals is through the log-survivor functions, $\ln P[X > x]$. In the case of the Poisson hypothesis,

$$\ln P[X > x] = -\lambda x \tag{3.33}$$

which is a straight line with slope $-\lambda$. The log-survivor function was calculated from the natural logarithm of $1 - F_n(x)$ where (*Lewis et al.,* 1969)

$$\begin{aligned} F_n(x) &= 0 && x < x_{(1)} \\ &= \frac{i}{n} && x_{(i-1)} \leq x < x_{(i)} \qquad i = 2, 3, \ldots, n \\ &= 1 && x_{(n)} \leq x . \end{aligned} \tag{3.34}$$

In (3.34) $x_{(i)}$ denotes the i-th order statistic in the observed sample. The log-survivor functions were estimated by a computer program of *Lewis et al* (1969), for the 17 stations in Indiana. Sample log-survivor functions are shown on the fig. 12 for the stations 0132, 3082, 3777, 4642, 6056 and 7747. Except for the outliers, almost all of the log-survivor functions are straight lines with negative slopes, satisfying the Poisson hypothesis.

After the data were homogenized, the relative frequency histograms and the log-survivor functions for the interarrival times of the rainfall counts process in the 17 stations in Indiana were computed. The sample relative frequency histograms of the interarrivals for the homogenized daily rainfall occurrences are shown on fig. 13 for the stations 0132, 3082, 3777, 4642, 6056 and 7747. Except for the outliers, no voids are left in the relative frequency histograms. However, there is a shift from the negative exponential shape to, probably, the Weibull distribution for which the log-survivor function is

$$\ln P[X > x] = -\left(\frac{x-\varepsilon}{v-\varepsilon}\right)^k \quad , \; x \geq \varepsilon$$

for the parameters v and k such that $v \geq \varepsilon$ and $k > 1$. Among the 17 stations analyzed stations 0132, 0177, 0545, 1747, 3777, 6056, 6338, 7069, 7747, 7755 and 7935 yielded log-survivor functions which were convex from the theoretical Poisson log-survivor function. This is seen in fig. 14 for the stations 0132, 3777, 6056 and 7747. The hazard function z(x) is defined as

$$z(x) = f_X(x)/P[X > x]$$

for the interarrival times X. The relationship between the log-survivor function $\ln P[X > x]$ and z(x) is (*Cox and Lewis,* 1966)

$$\ln P[X > x] = -\int_0^x z(u)du$$

so that

$$\frac{d}{dx} \ln P[X > x] = -z(x)$$

and

$$\frac{d^2}{dx^2} \ln P[X > x] = -\frac{dz(x)}{dx} \tag{3.35}$$

Therefore, a concave log-survivor function would correspond to a monotone non-increasing hazard function z(x). A monotone non-increasing hazard implies that the coefficient of variation of X is greater than unity (*Watson and Wells,* 1961). *Cox and Lewis* (1966) have shown that

$$V'(\infty) = \frac{C^2(X)}{E(X)}$$

where C(x) is the coefficient of variation of the interarrival times X while $V'(\infty)$ is the asymptotic slope of the variance-time curve of the counting process N(t). For a convex log-survivor function

$$V'(\infty) \geq \frac{1}{E(x)} \; . \tag{3.36}$$

However, 1/E(x) is the constant slope of the variance-time function for the Poisson case. Therefore, for a convex log-survivor function the variance-time curve V(t), would be above t/E(x) for large t. The striking fact is that, except for station 7747 all the other stations with convex log-survivor functions yielded variance-time curves which lay above the theoretical Poisson variance-time function t/E(x). This suggests a clustering of the rainfall occurrences. However, the convexity is not too pronounced in the sample log-survivor functions. The functions lie comfortably close to the theoretical log-survivor function $\ln P[X > x] = -x$, for the exponentially distributed interarrival times.

3.4.2 Some Statistics on Homogenization

The effect of the homogenization scheme (3.6) on the daily rainfall counts data can be seen from the change in the mean, standard deviation, coefficient of skewness and the coefficient of variation. If the homogenization scheme (3.6) performs properly, the mean interarrival time of the homogenized data should be close to 1.

The statistics of the original data for the 17 stations are shown in the table 3.2. The table shows that the coefficients of variation in all the 17 stations are greater than 1, the value for the coefficient of variation of a Poisson process. Therefore, the data shows the grouping of the rainfall events. The coefficients of skewness are all positive, showing skewness to the right. This is expected for the exponentially distributed interarrivals for the Poisson hypothesis.

Table 3-3 shows the statistics for the homogenized daily rainfall data for the 17 Indiana stations. All the mean-arrival times are very close to unity. Therefore, the rate of rainfall occurrence is unity, as it is supposed to be after the transformation (3.6). The coefficients of variation are also quite close to unity implying that the interarrival times of the homogenized daily rainfall counts are negative exponentially distributed.

3.4.3 Statistical Tests on Homogenization

Two tests for the stationarity of the interarrival times between rainfall occurrences were employed in the analysis of trends. The same tests are used for analyzing the effects of the homogenization of the rainfall data by the scheme (3.6). In order to use these tests the independence of the interarrival times have to be assumed.

3.4.3.1 Test of trend in the rate of homogenized daily rainfall occurrences under the Poisson hypothesis

The reader is referred to Section 2 of Chapter II for the discussion of this test. The results of the test, given in table 3-4, show that the trend in the rate of rainfall occurrence is removed by the homogenization scheme (3.6).

3.4.3.2 Homogeneity of variance test for the interarrivals of the homogenized daily rainfall occurrences

The reader is referred to Sec. 2 of Chapter II for a discussion of this test. The results, given in table 3-5, show that the variance of the interarrival times is still nonhomogeneous, although it is considerably reduced when the results are compared to the variance homogeneity statistics of the original data in table 2-3.

The homogenization scheme (3.6) which homogenized the rate of occurrence in all cases, would also homogenize the variance if the stochastic model underlying the daily rainfall counts process was Poisson. The verification of the removal of trends in the rate of occurrence, coupled by the rejection of the homogenization of the variance of the interarrivals, leads to the conclusion that the stochastic model of the daily rainfall occurrences is not Poisson.

3.5 SUPPLEMENTARY FIGURES

Plots of the number of rainy days versus cumulative time, of the intensity function, of the variance-time function, and of the counts spectrum for the homogenized daily rainfall data for the stations 0177, 0545, 1747, 1882, 3547, 4908, 6164, 6338, 7069, 7755 and 7935 are given in figures 7A, 7B, 8A, 8B, 9A, 9B and 10A, 10B. The relative frequency histogram and the log-survivor function for the original daily rainfall interarrival times and the relative frequency histogram and the log survivor function for the homogenized daily rainfall interarrival times for the above stations are given in figures 11A, 11B, 12A, 12B, 13A, 13B and 14A, 14B.

TABLE 3-2

STATISTICS FOR THE DAILY RAINFALL INTERARRIVAL TIMES

Station	Mean	Standard Dev.	Coefficient of Variation	Skewness Coefficient
0132	3.2232	3.5854	1.1124	3.2998
0177	3.5563	4.5706	1.2852	4.4738
0545	3.3152	3.6897	1.1132	2.7794
1747	3.3587	3.6449	1.085	2.5958
1882	4.3562	4.6453	1.066	2.6325
3082	3.2031	3.4257	1.0678	2.9172
3547	4.1512	4.7224	1.1377	2.9085
3777	3.3356	3.5637	1.0621	3.1592
4642	3.2938	3.5329	1.0726	2.7681
4908	3.4648	3.7263	1.0755	2.4559
6056	3.6271	5.3424	1.4732	7.8724
6164	3.2811	3.6215	1.1041	4.2645
6338	3.3499	3.6525	1.0902	2.8726
7069	3.1962	3.3032	1.0335	2.7592
7747	3.2615	3.4794	1.0668	2.7644
7755	3.4742	3.815	1.093	2.6724
7935	3.9247	4.2371	1.0796	2.2415

TABLE 3-3

STATISTICS FOR THE INTERARRIVAL TIMES OF THE HOMOGENIZED DAILY RAINFALL DATA

Station	Mean	Standard Dev.	Coefficient of Variation	Skewness Coefficient
0132	.9979	1.0426	1.0448	3.2651
0177	.999	1.1292	1.1298	3.7726
0545	.9994	1.042	1.0427	2.3751
1747	.998	1.0127	1.0147	2.0878
1882	.9907	.9577	.9667	1.8155
3082	1.0001	.97	.97	2.076
3547	.9999	1.1069	1.107	3.333
3777	1.0216	1.0422	1.020	2.726
4642	.9987	.9964	.9977	2.210
4908	.9982	.9815	.9833	1.973
6056	.9949	1.269	1.2754	6.592
6164	.9973	1.0311	.0409	2.518
6338	.9996	.990	.9903	1.8812
7069	.9939	1.0012	1.010	2.0838
7747	.9994	.9818	.9823	2.2123
7755	.9859	1.0068	1.0212	2.3033
7935	.9984	1.024	1.026	2.0695

TABLE 3-4

TEST OF TREND IN THE RATE OF RAINFALL OCCURRENCE FOR THE HOMOGENIZED DATA
(Underlying process is assumed to be Poisson)

Station	Cramer's Statistic U	Significant at 5%	Station	Cramer's Statistic U	Significant at 5%
0132	.246	No	4908	.1507	No
0177	- .008	No	6056	- .0417	No
0545	- .553	No	6164	- .8979	No
1747	.1291	No	6338	.082	No
1882	.344	No	7069	.323	No
3082	.07	No	7747	1.0585	No
3547	- .095	No	7755	- .259	No
3777	-1.593	No	7935	- .35	No
4642	.0595	No			

U ~ N(0,1)

TABLE 3-5

HOMOGENEITY OF VARIANCE TEST FOR THE INDEPENDENT RAINFALL INTERARRIVALS FOR THE HOMOGENIZED DATA

Station	Hom. of Variance Statistic	Degree of Freedom (k)	Significant at 1%
0132	228.24	43	Yes
	100.15	15	Yes
	42.01	5	Yes
0177	268.100	38	Yes
	135.62	13	Yes
	10.34	4	No
0545	180.92	41	Yes
	106.85	15	Yes
	56.51	5	Yes
1747	166.69	41	Yes
	94.44	14	Yes
	55.73	5	Yes
1882	61.95	31	Yes
	31.66	11	Yes
	7.63	3	No
3082	131.35	43	Yes
	52.34	15	Yes
	14.76	5	No
3547	152.62	33	Yes
	70.26	11	Yes
	42.7	4	Yes
4642	121.19	42	Yes
	46.75	15	Yes
	12.05	5	No
4908	87.45	40	Yes
	40.31	14	Yes
	20.84	5	Yes
6056	400.11	38	Yes
	290.86	13	Yes
	175.99	4	Yes
6164	156.20	42	Yes
	72.45	15	Yes
	29.17	5	Yes
6338	117.87	41	Yes
	28.77	14	Yes
	6.05	5	No
7069	143.66	43	Yes
	66.35	15	Yes
	16.96	5	Yes
7747	128.75	42	Yes
	58.54	15	Yes
	24.13	5	Yes
7755	140.05	39	Yes
	65.15	14	Yes
	40.68	5	Yes
7935	153.33	35	Yes
	58.88	12	Yes
	16.34	4	Yes

Hom. of Variance Statistic ~ X_k^2

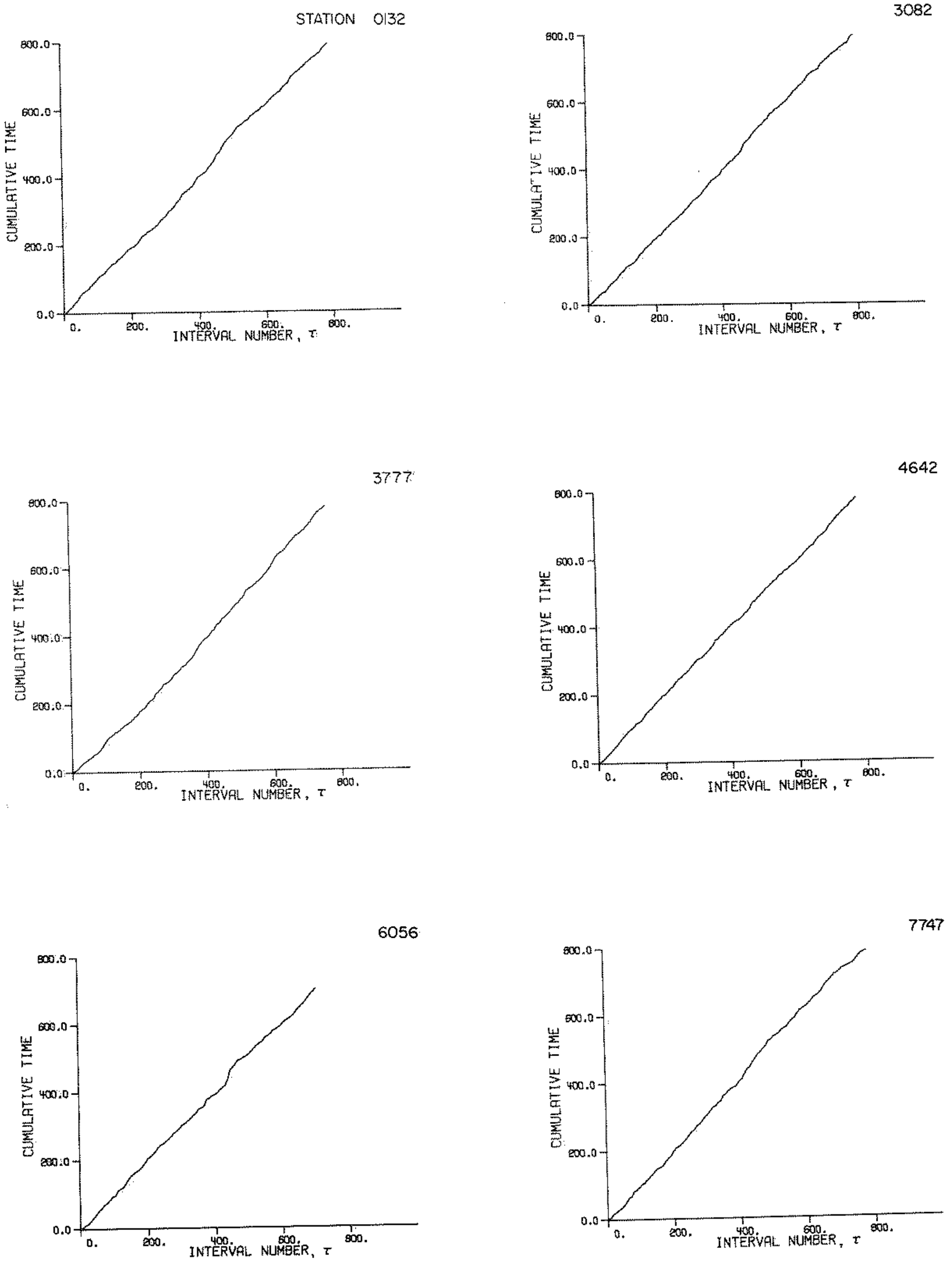


FIG.7- CUMULATIVE RESCALED TIME VS EVENT NUMBER

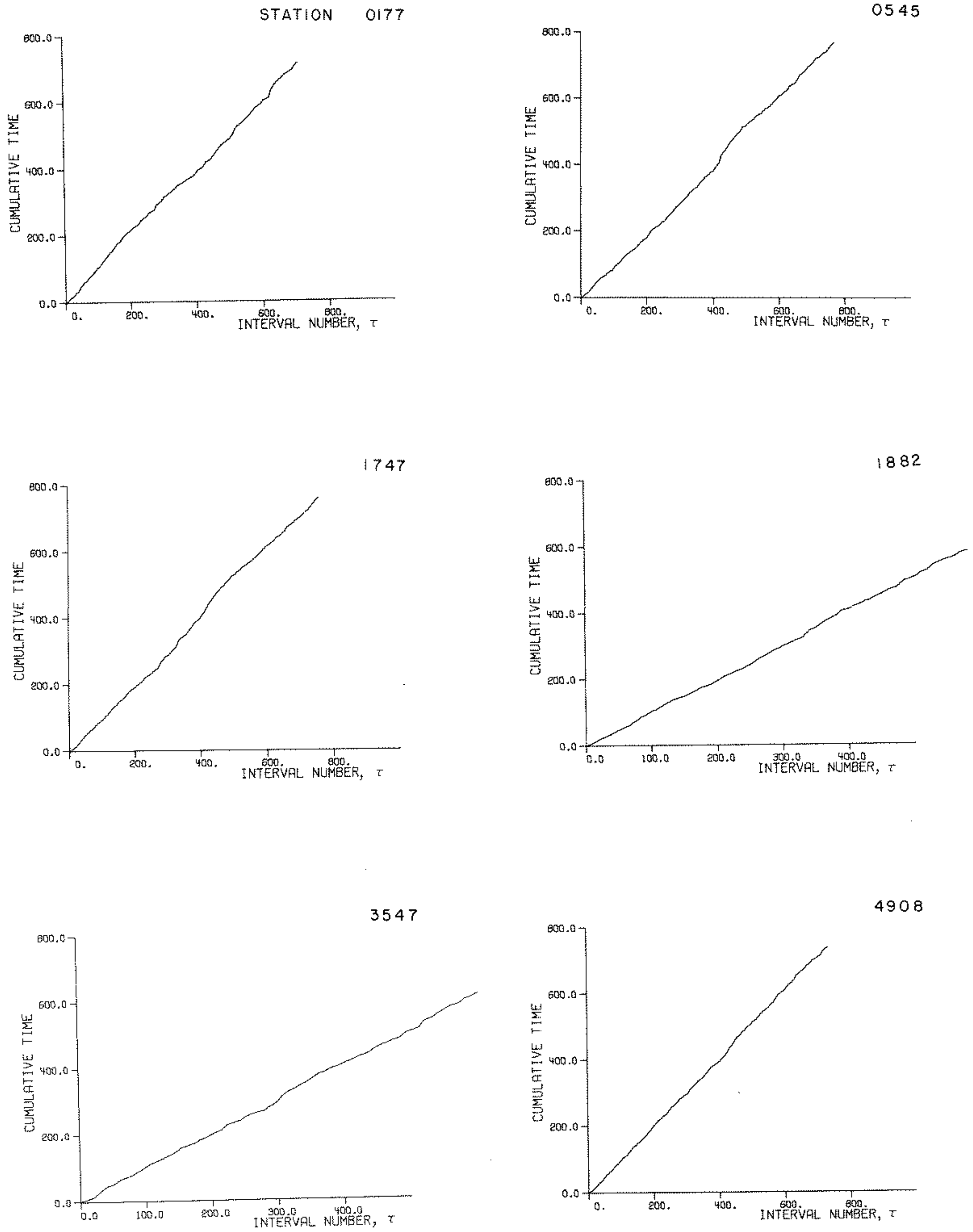


FIG. 7A — CUMULATIVE RESCALED TIME VS EVENT NUMBER

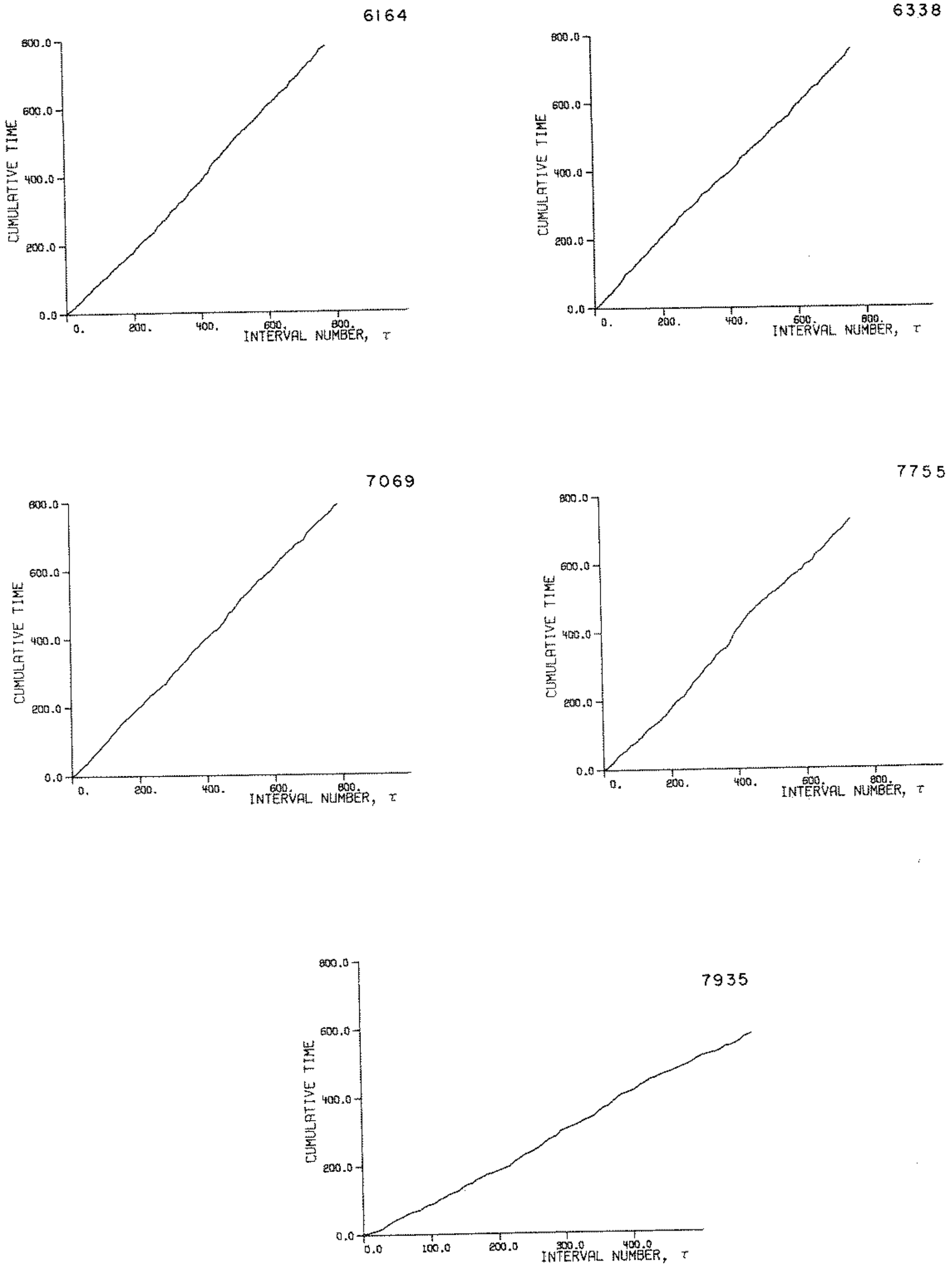


FIG. 7B - CUMULATIVE RESCALED TIME VS EVENT NUMBER

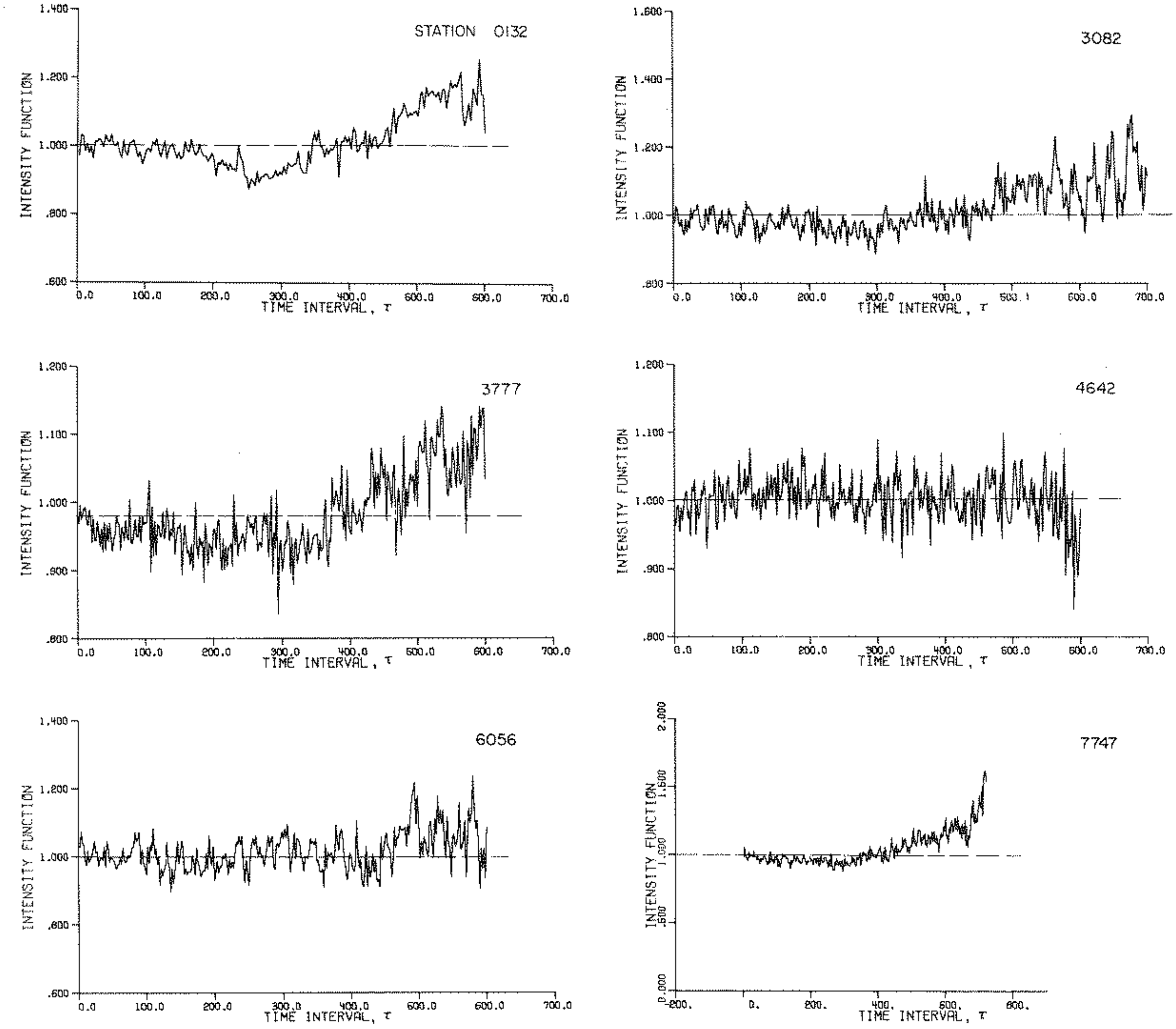


FIG. 8- INTENSITY FUNCTION OF HOMOGENIZED DAILY RAINFALL

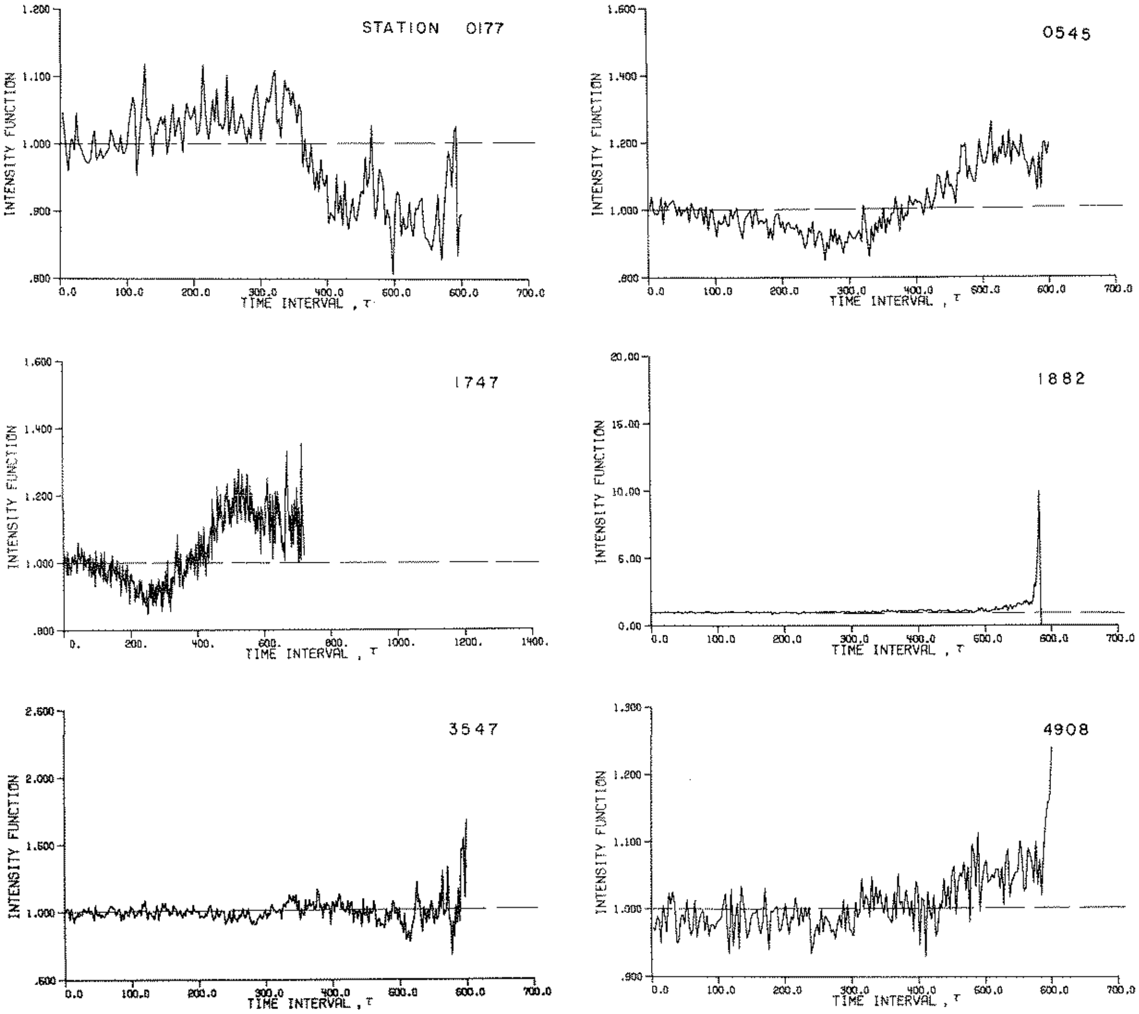


FIG. 8A—INTENSITY FUNCTION OF HOMOGENIZED DAILY RAINFALL

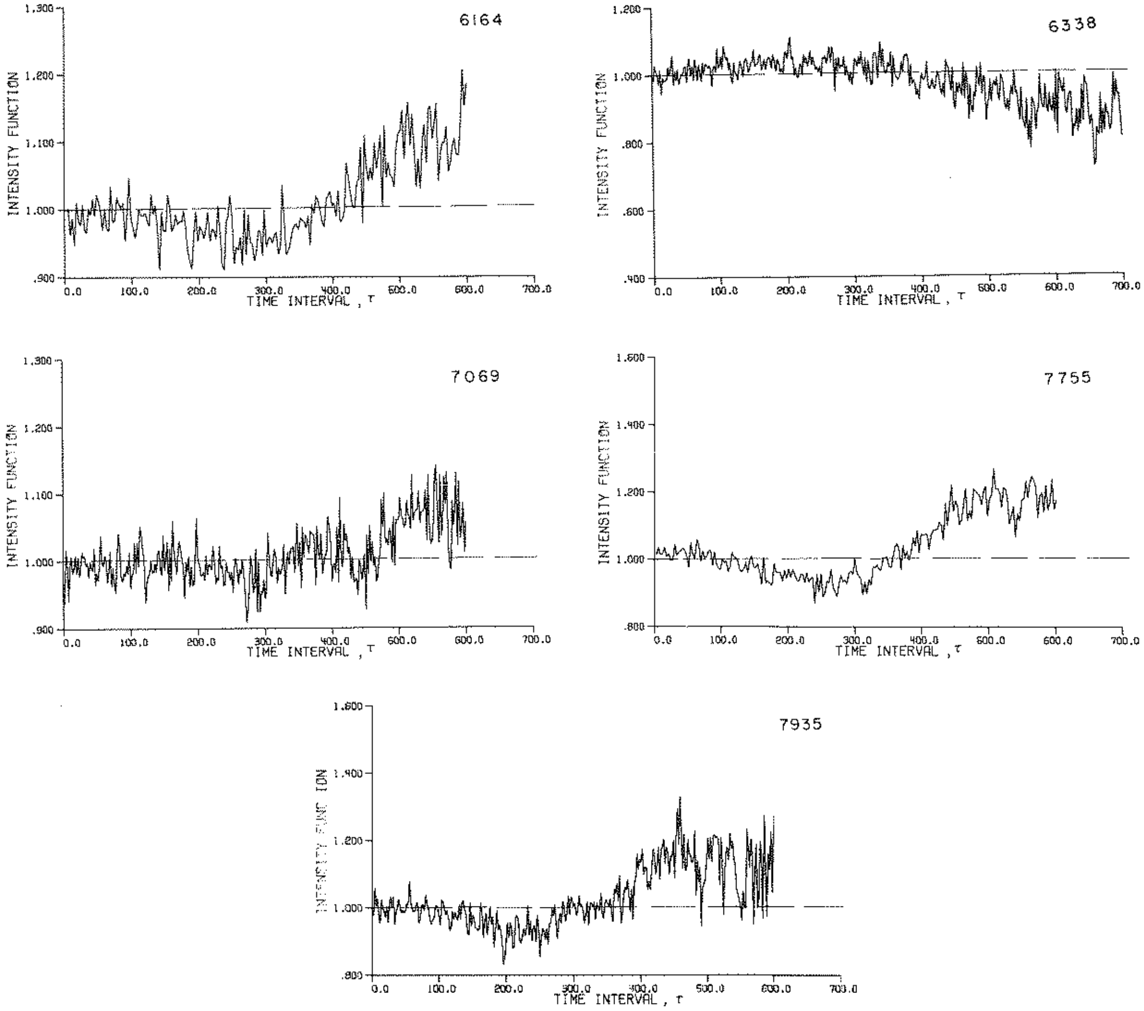


FIG. 8B—INTENSITY FUNCTION OF HOMOGENIZED DAILY RAINFALL

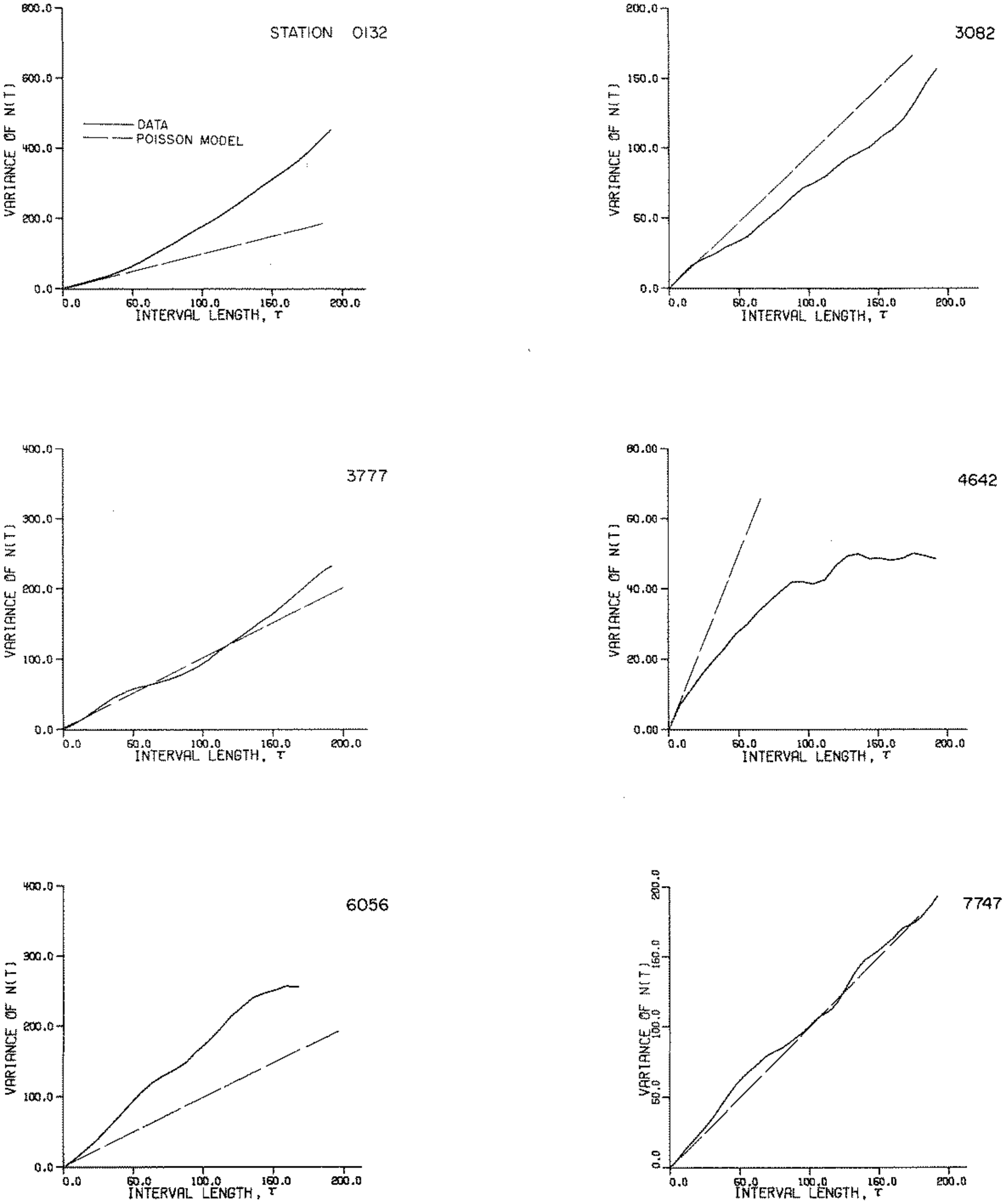


FIG. 9 – VARIANCE – TIME CURVE FOR HOMOGENIZED DAILY RAINFALL COUNTS

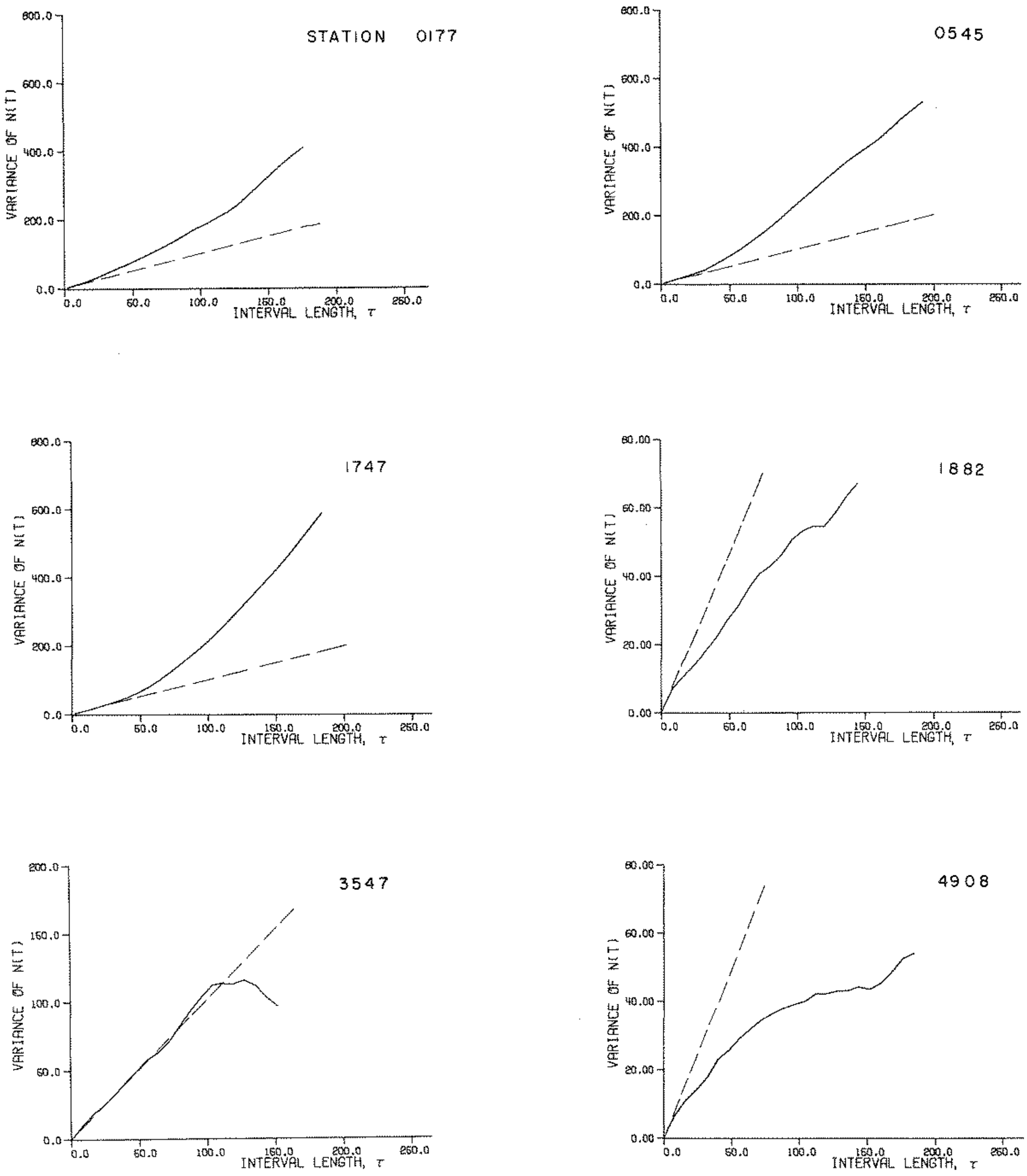


FIG. 9A — VARIANCE — TIME CURVE FOR HOMOGENIZED DAILY RAINFALL COUNTS

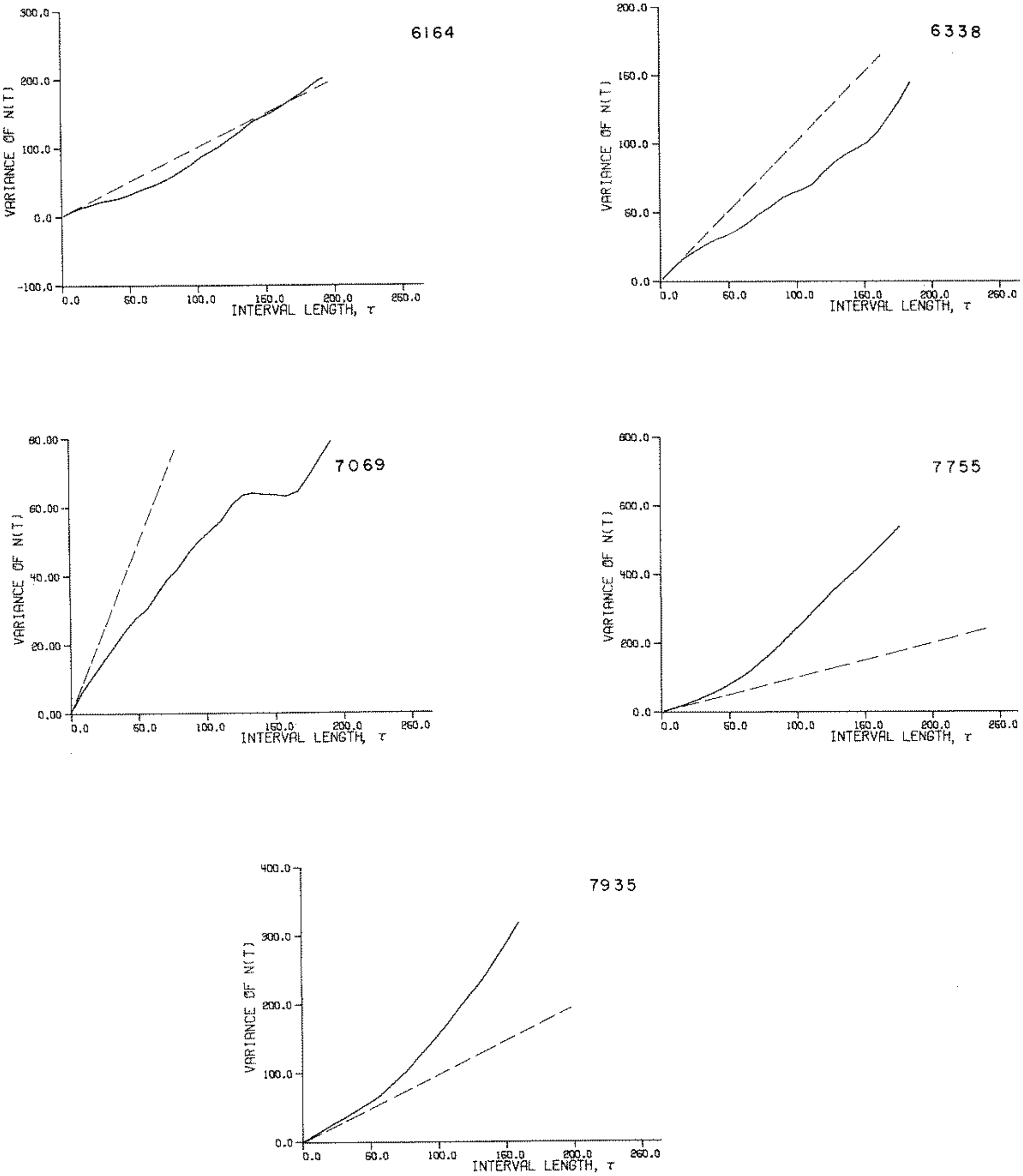


FIG. 9B — VARIANCE—TIME CURVE FOR HOMOGENIZED DAILY RAINFALL COUNTS

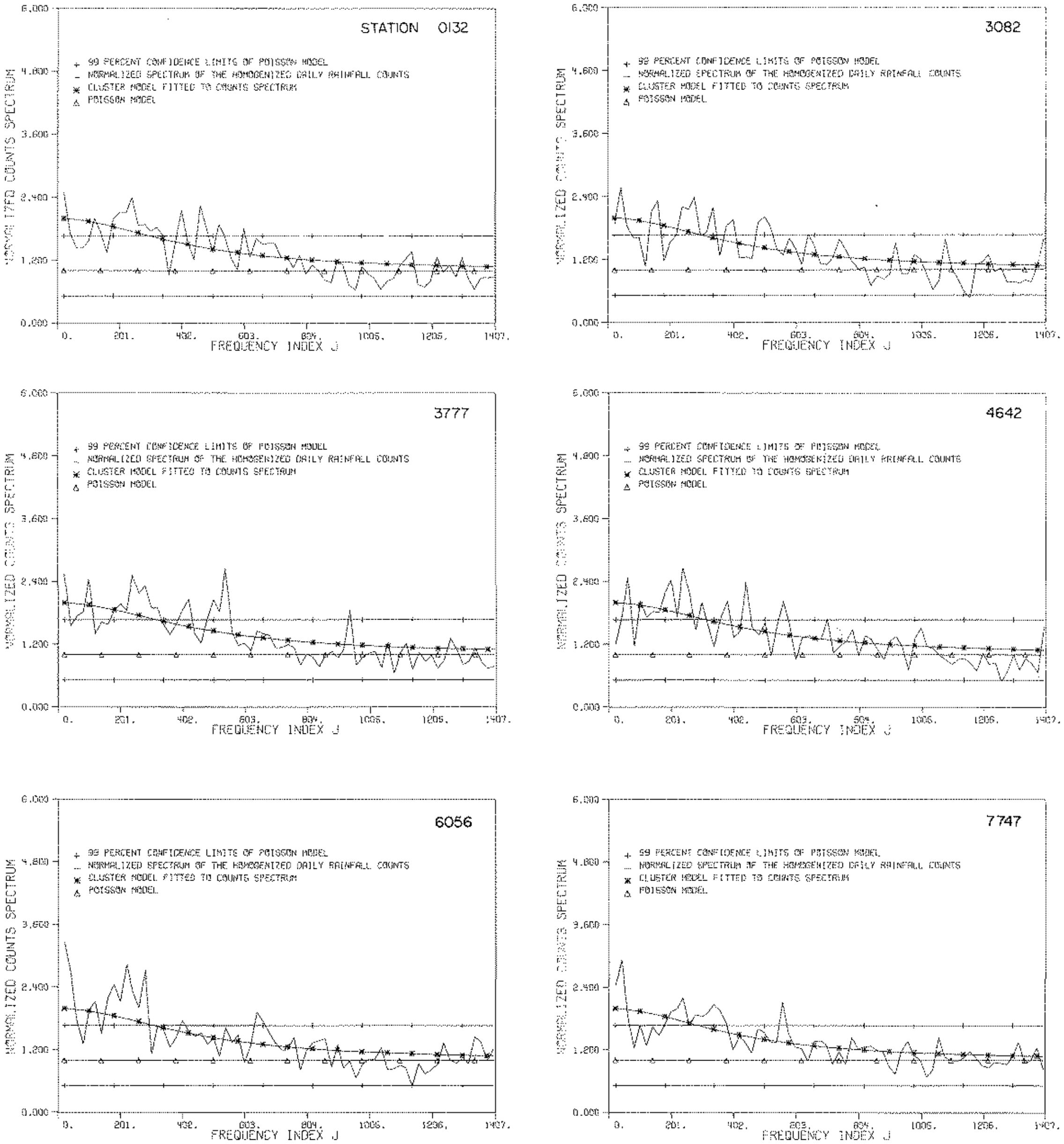


FIG. 10— NORMALIZED SPECTRUM OF THE HOMOGENIZED DAILY RAINFALL COUNTS AND OF FITTED POISSON AND CLUSTER MODELS

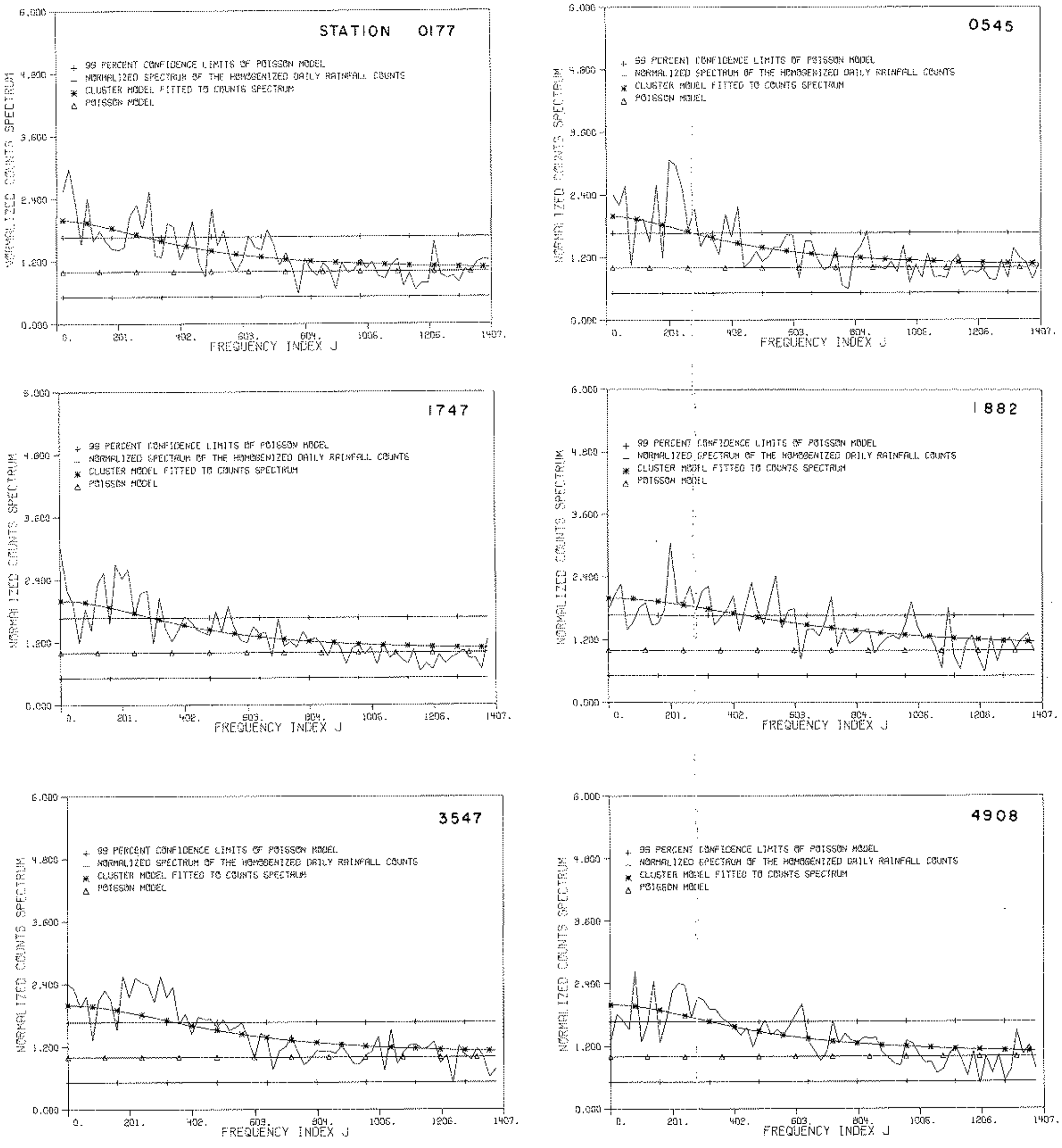


FIG. 10A— NORMALIZED SPECTRUM OF THE HOMOGENIZED DAILY RAINFALL COUNTS AND OF FITTED POISSON AND CLUSTER MODELS

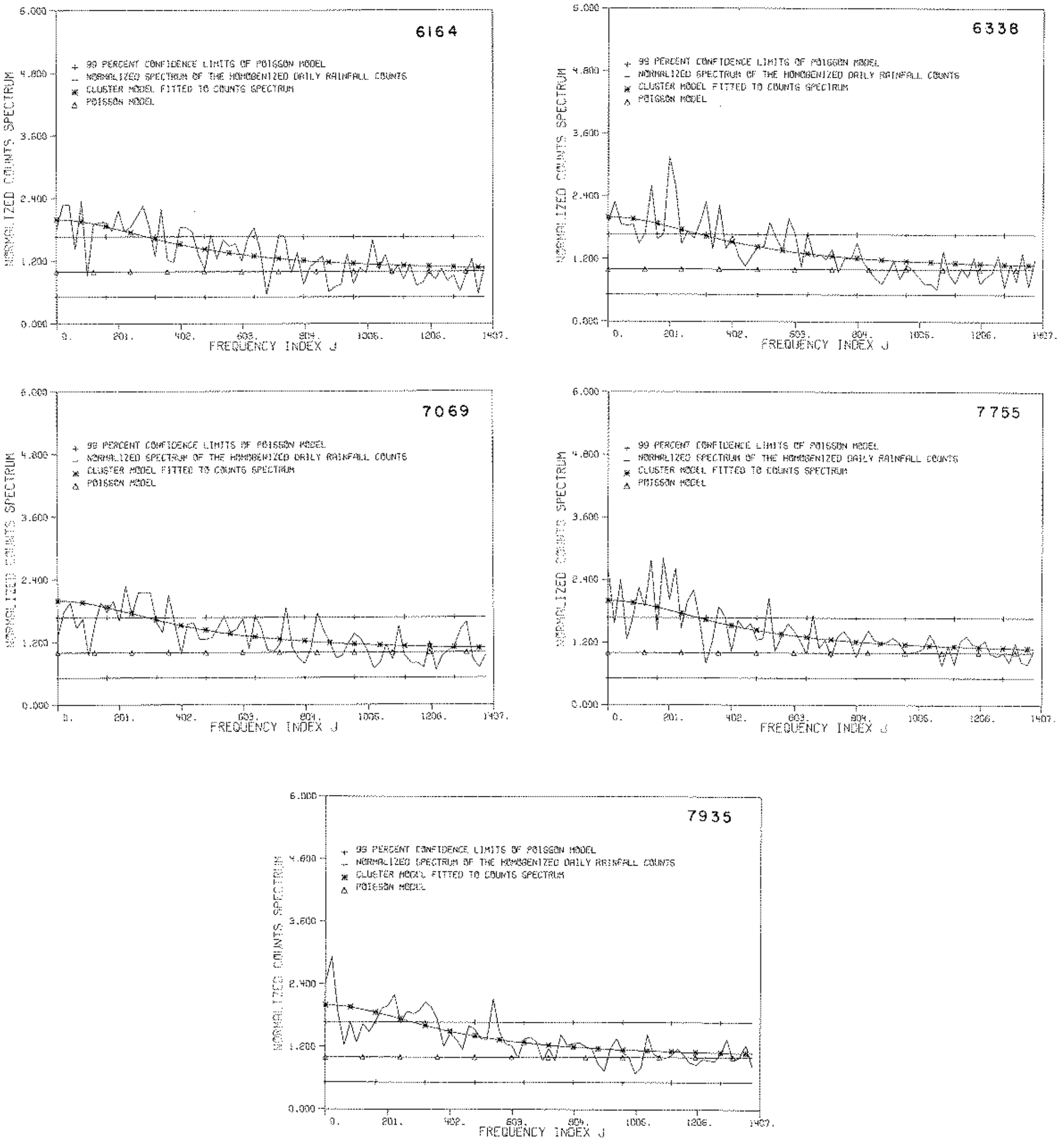


FIG. 10B — NORMALIZED SPECTRUM OF THE HOMOGENIZED DAILY RAINFALL COUNTS AND OF FITTED POISSON AND CLUSTER MODELS

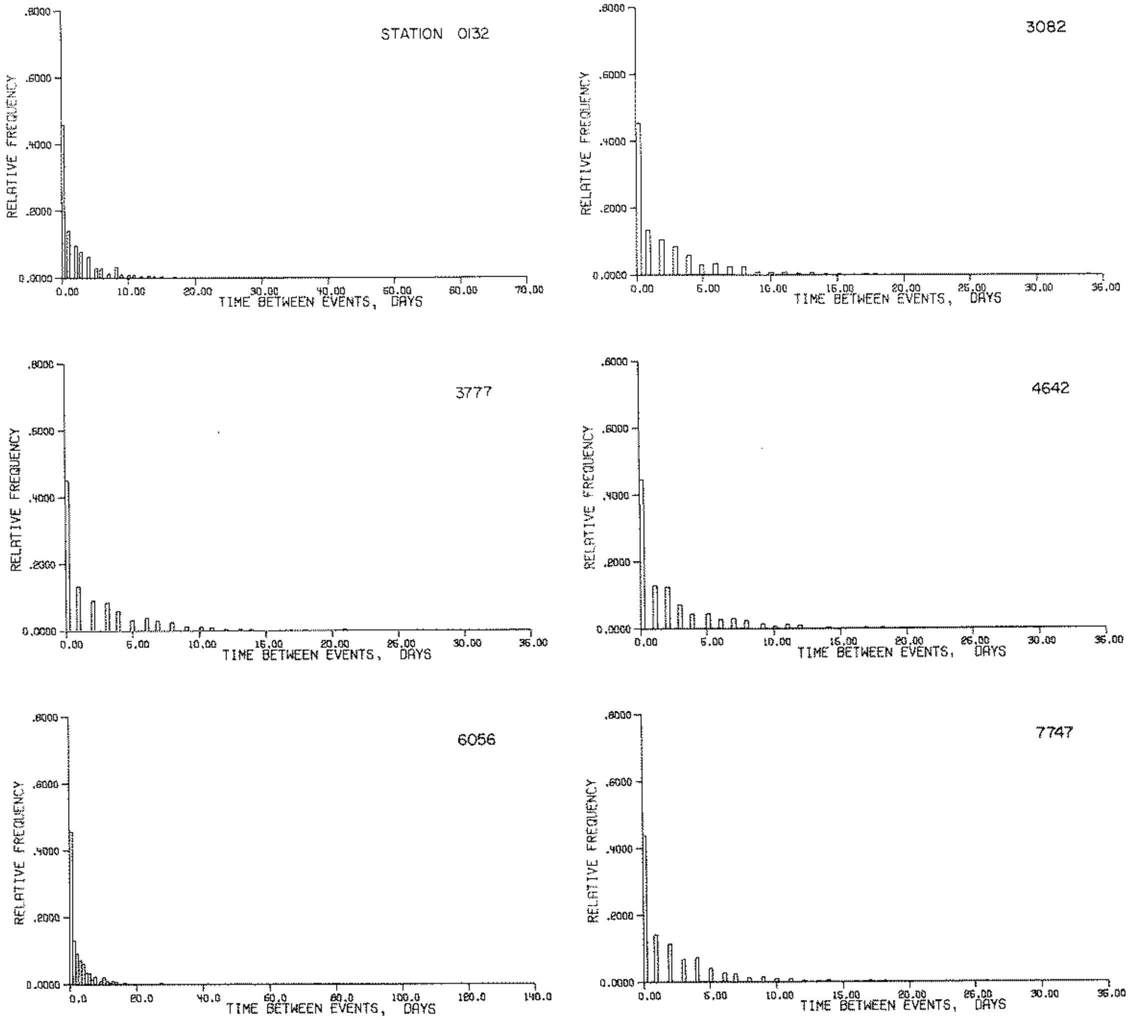


FIG. 11 – RELATIVE FREQUENCY HISTOGRAM FOR DAILY RAINFALL INTER-ARRIVAL TIME

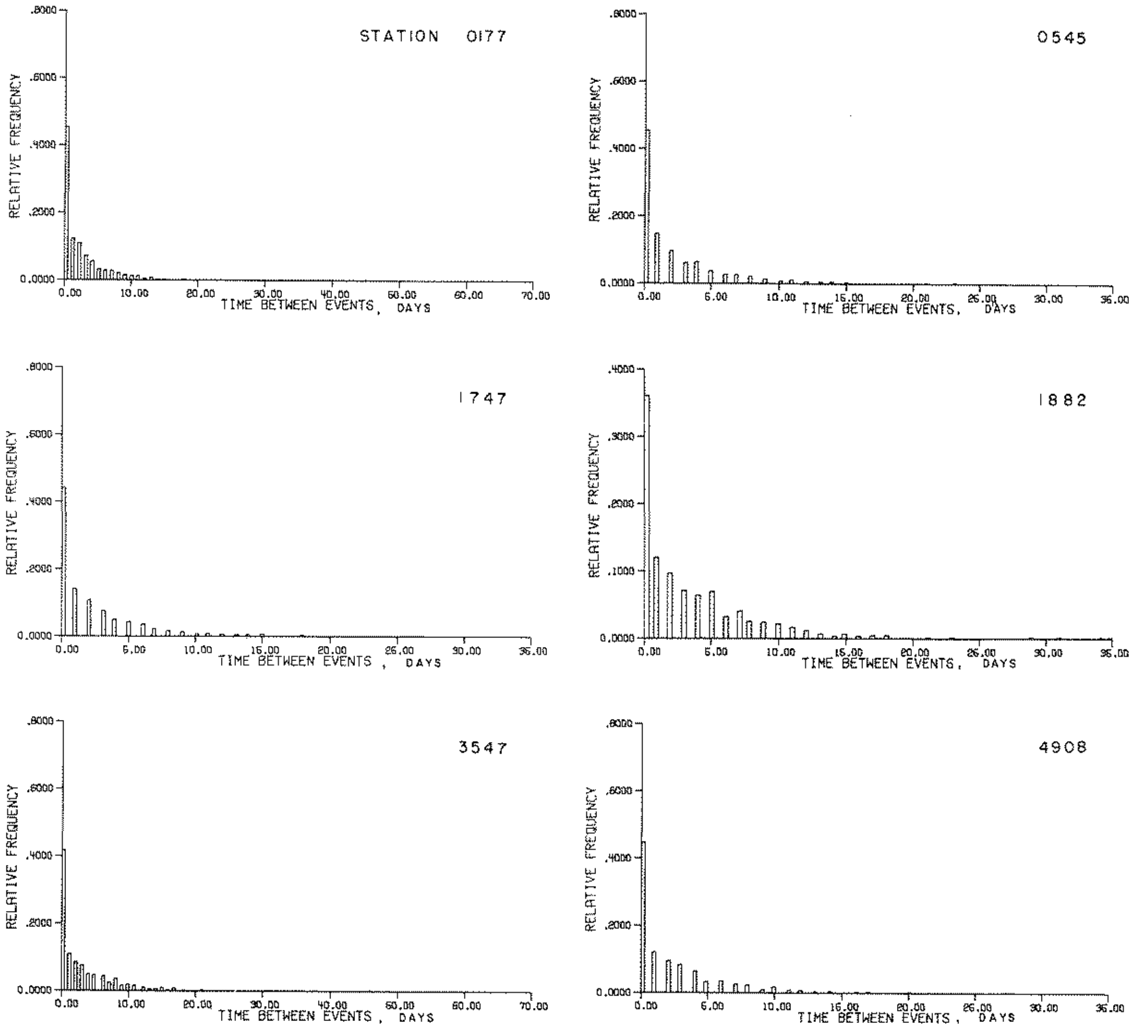


FIG. 11A – RELATIVE FREQUENCY HISTOGRAM FOR DAILY RAINFALL INTER-ARRIVAL TIME

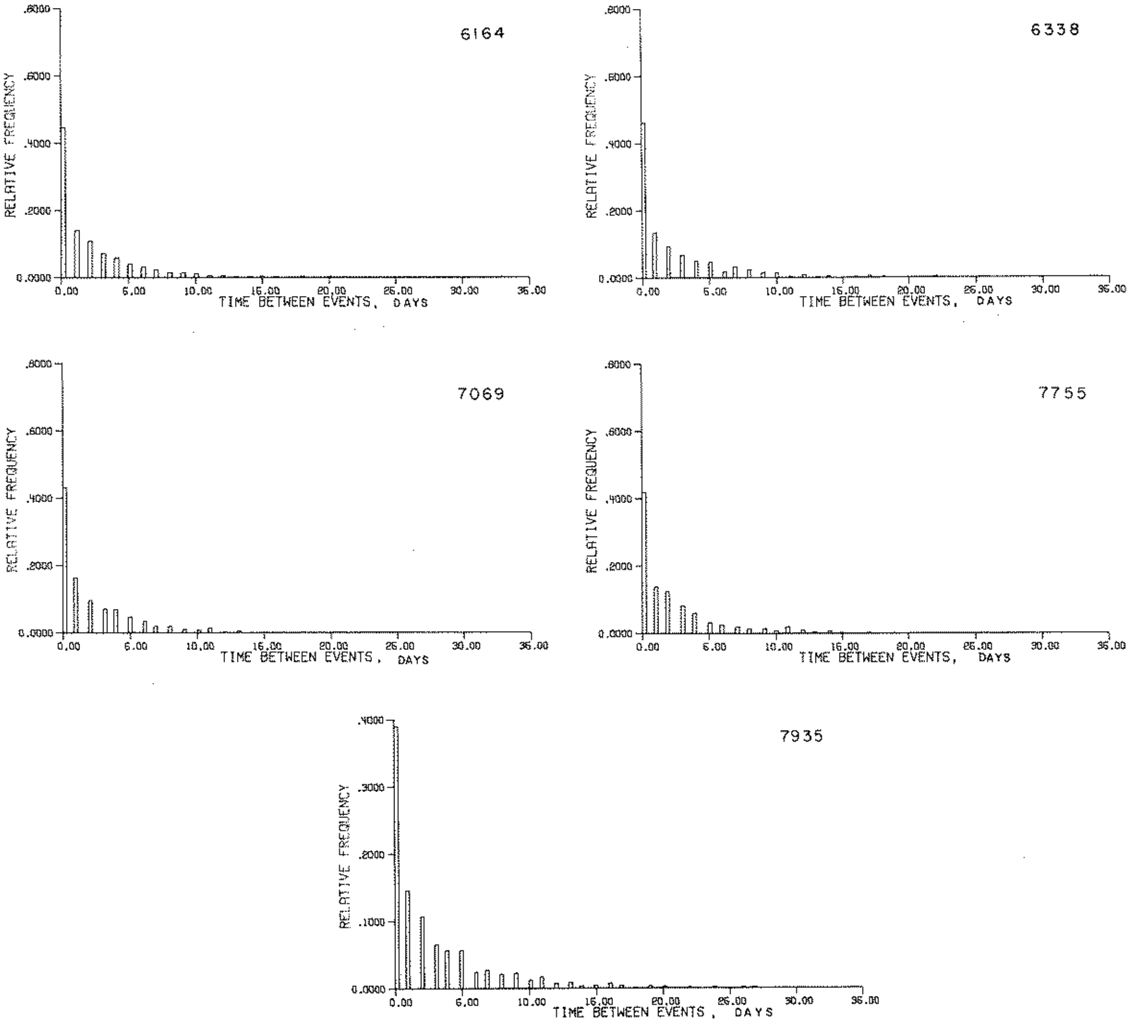


FIG. 11B – RELATIVE FREQUENCY HISTOGRAM FOR DAILY RAINFALL INTER-ARRIVAL TIME

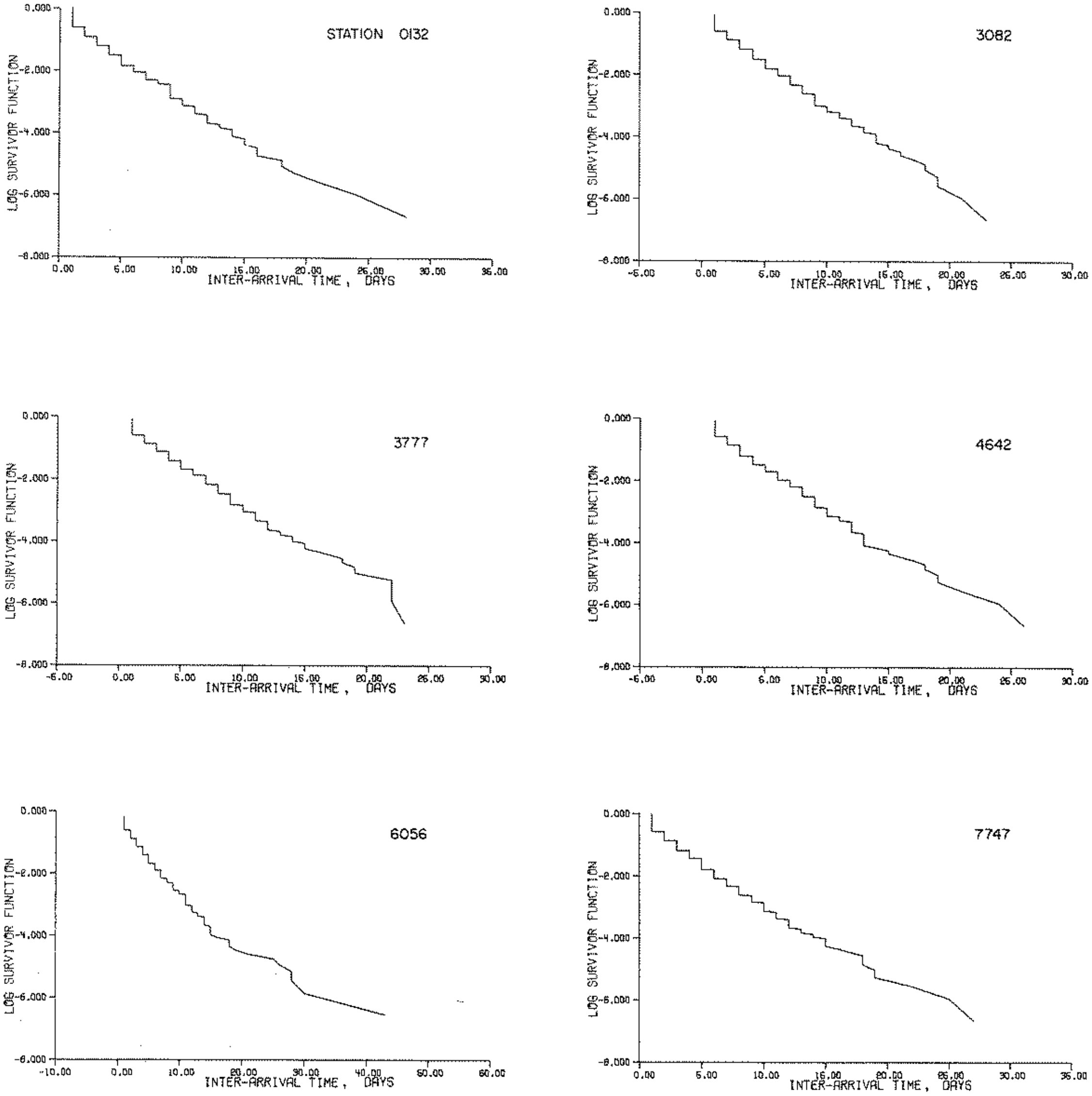


FIG. 12- LOG-SURVIVOR FUCTION

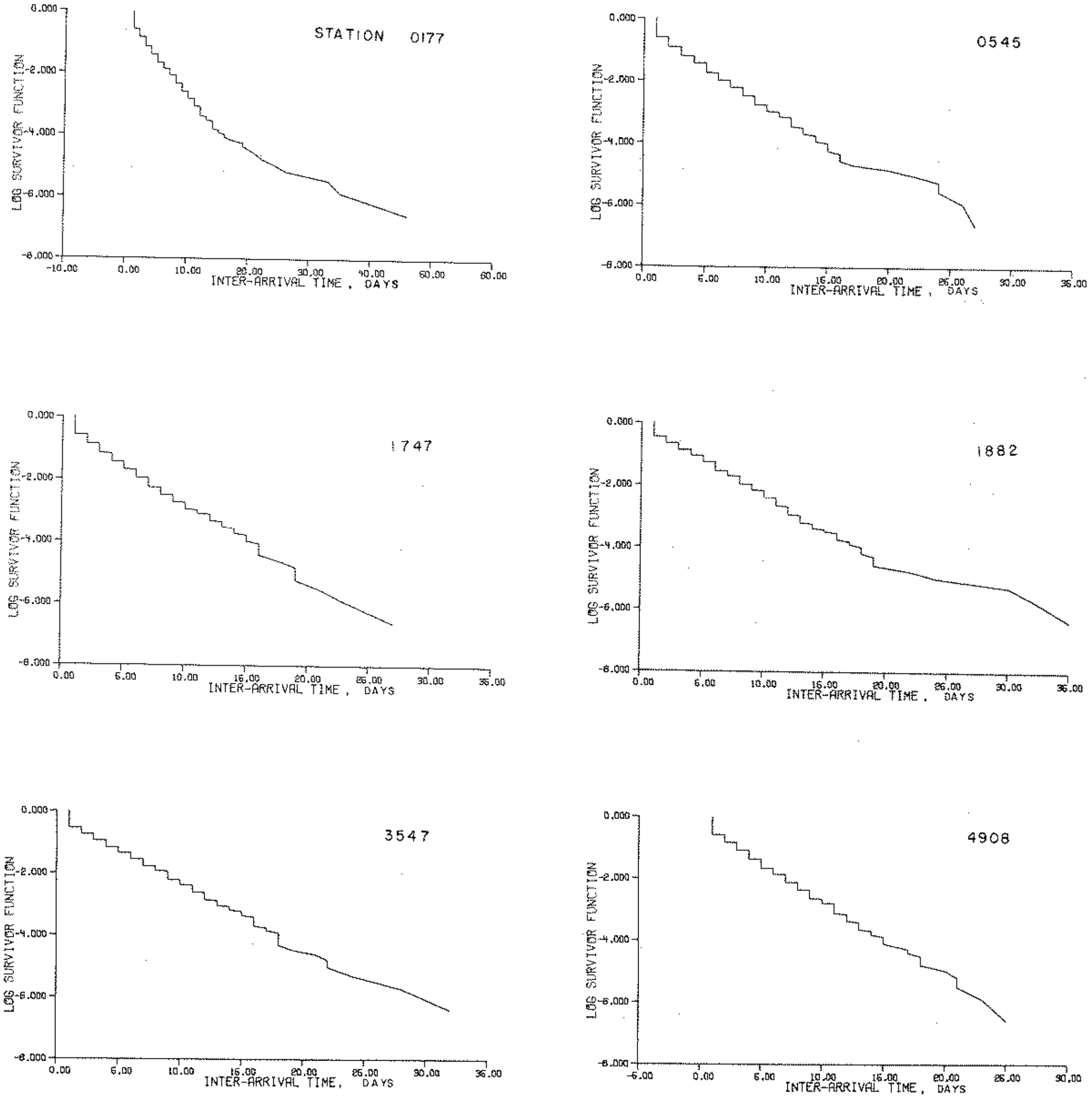


FIG. 12A - LOG-SURVIVOR FUNCTION

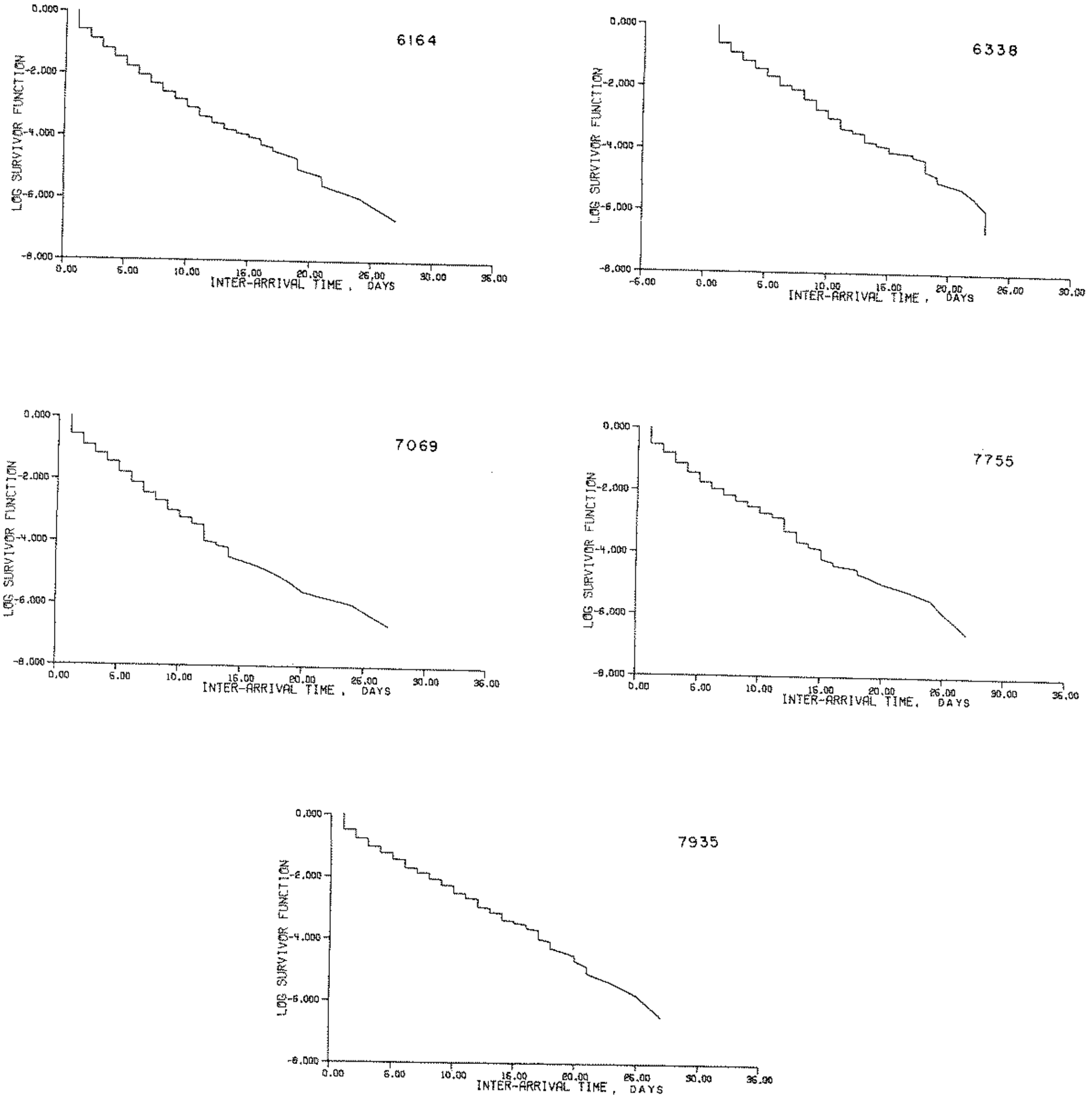


FIG. 12B – LOG-SURVIVOR FUNCTION

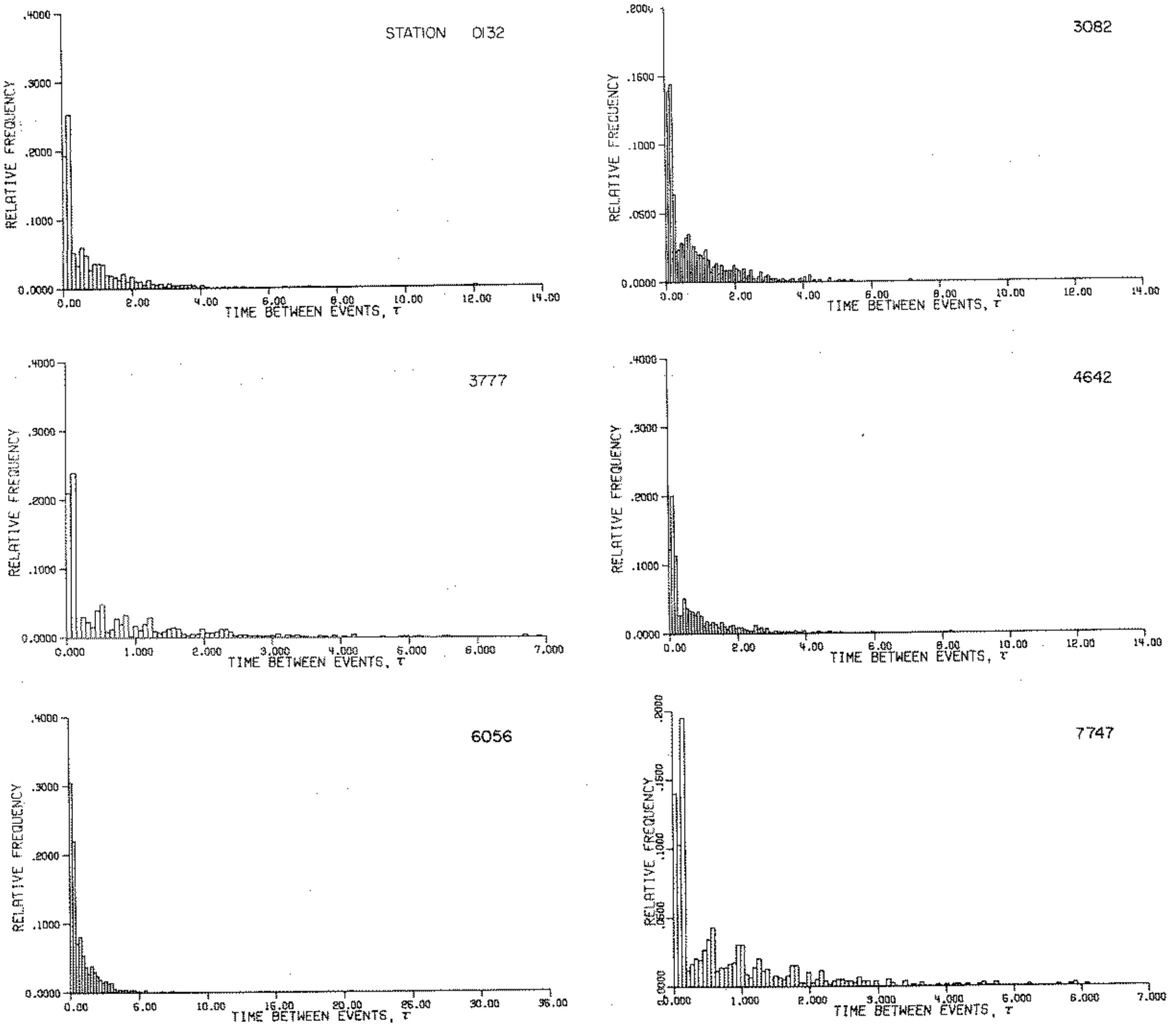


FIG. 13- RELATIVE FREQUENCY HISTOGRAM FOR HOMOGENIZED DAILY RAINFALL INTER-ARRIVAL TIMES

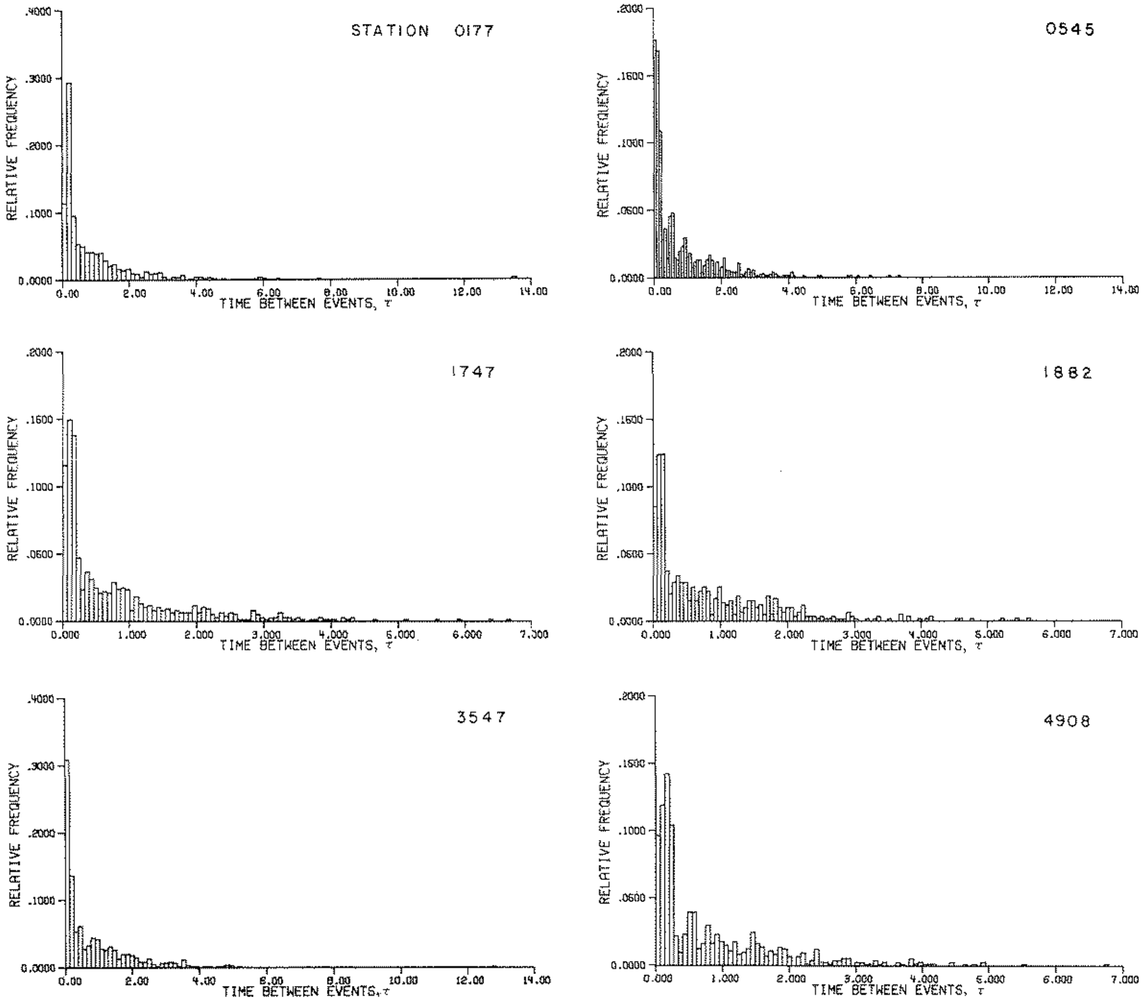


FIG. 13A—RELATIVE FREQUENCY HISTOGRAM FOR HOMOGENIZED DAILY RAINFALL INTER-ARRIVAL TIMES

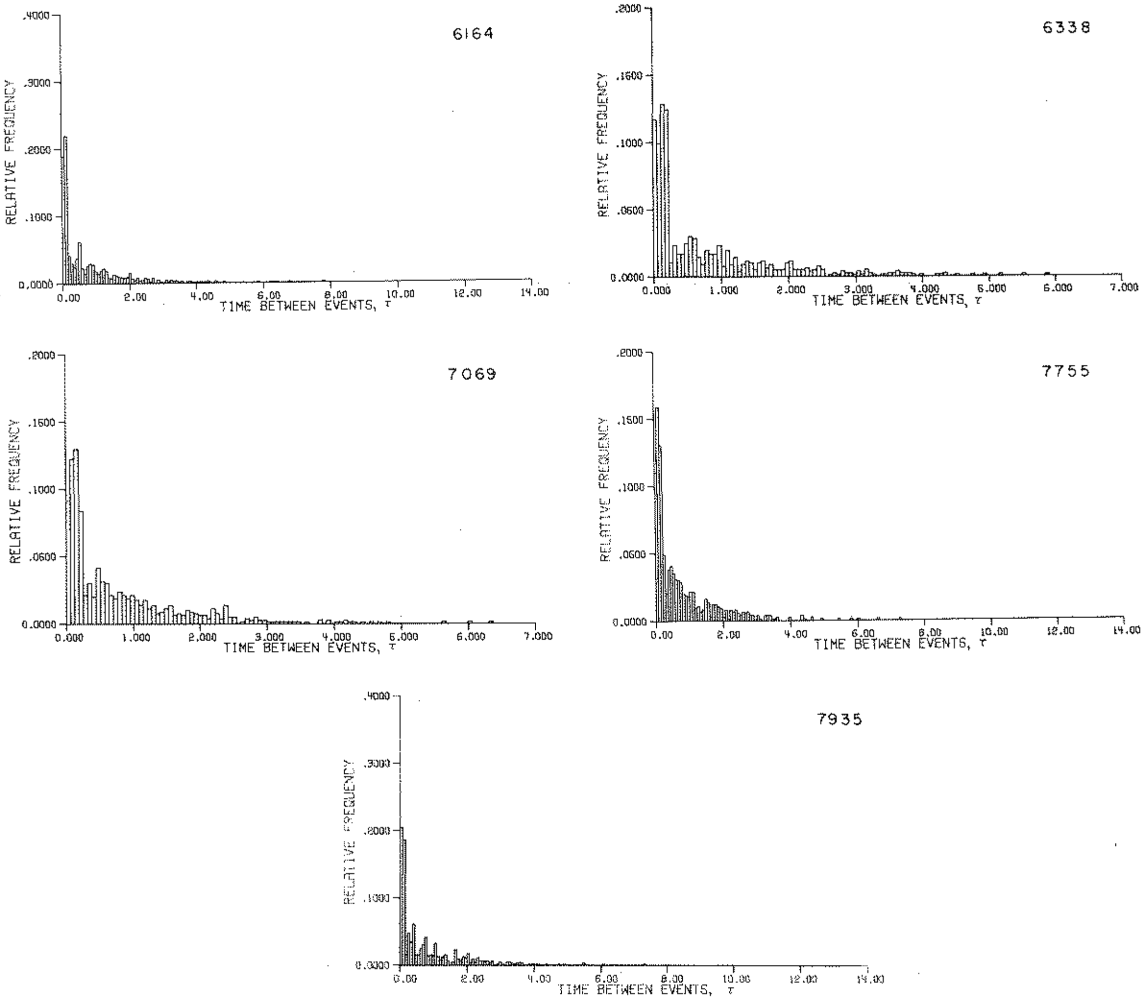


FIG. 13B – RELATIVE FREQUENCY HISTOGRAM FOR HOMOGENIZED DAILY RAINFALL INTER-ARRIVAL TIMES

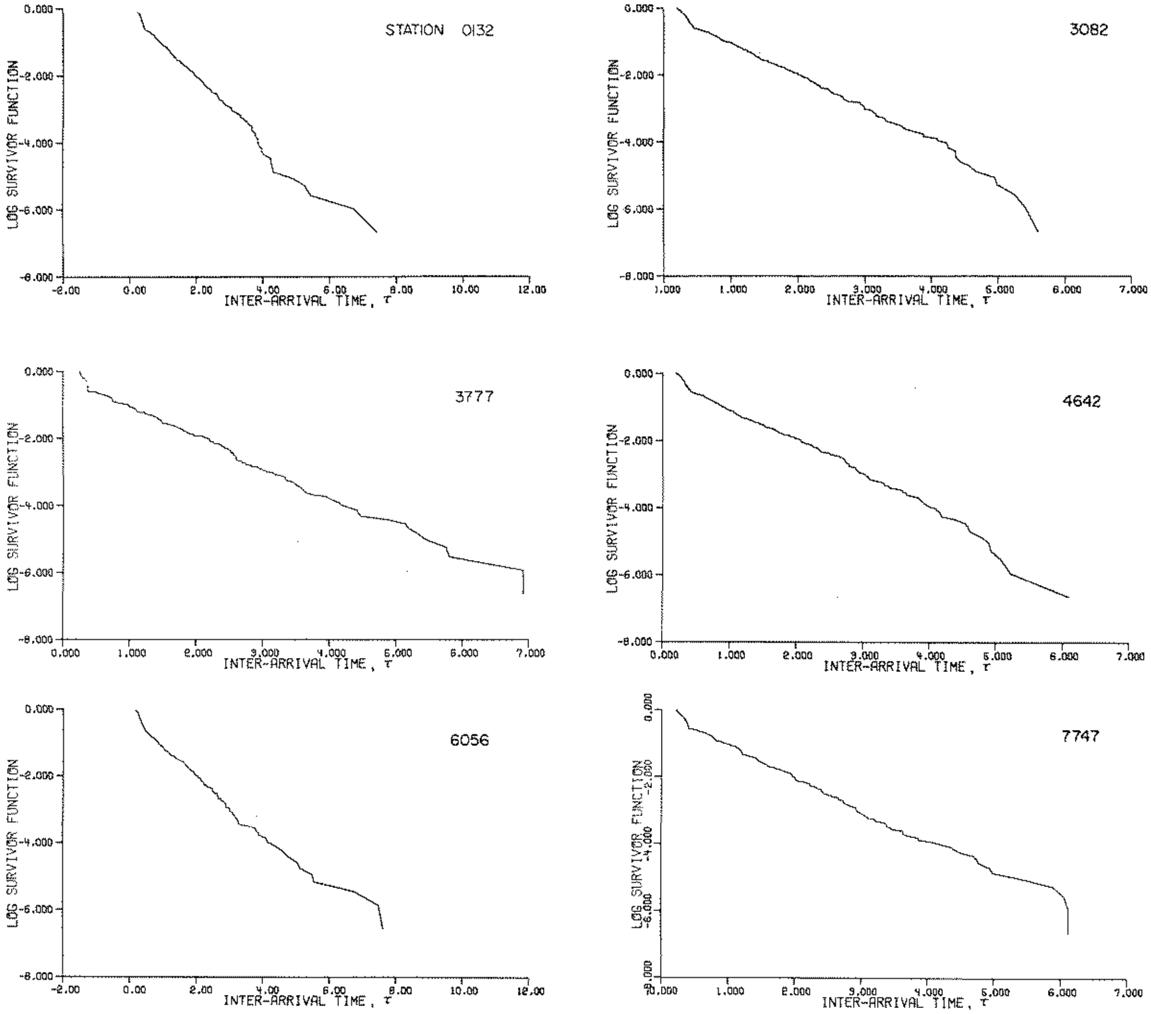


FIG. 14– LOG SURVIVOR FUNCTION OF THE HOMOGENIZED DAILY RAINFALL COUNTS

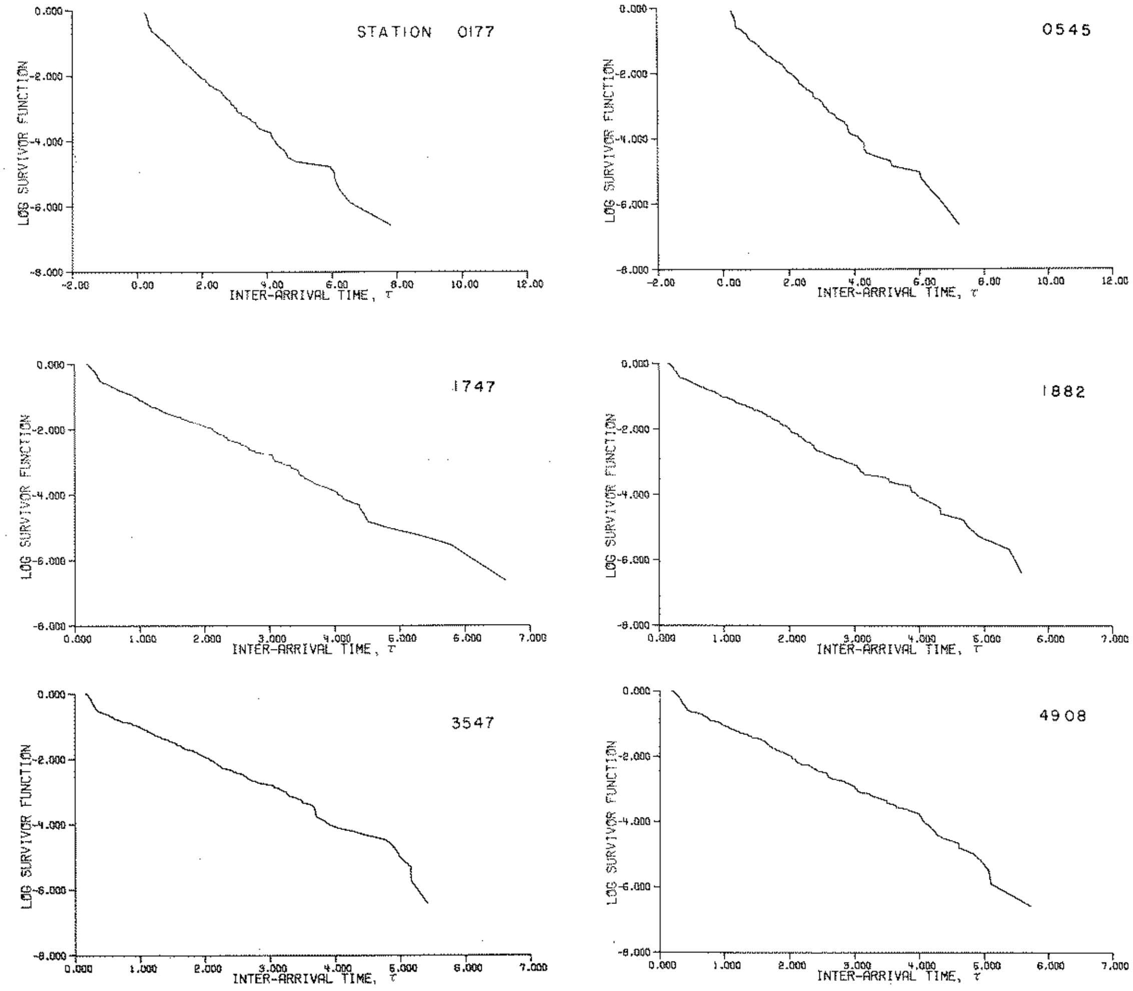


FIG. 14A–LOG SURVIVOR FUNCTION OF THE HOMOGENIZED DAILY RAINFALL COUNTS

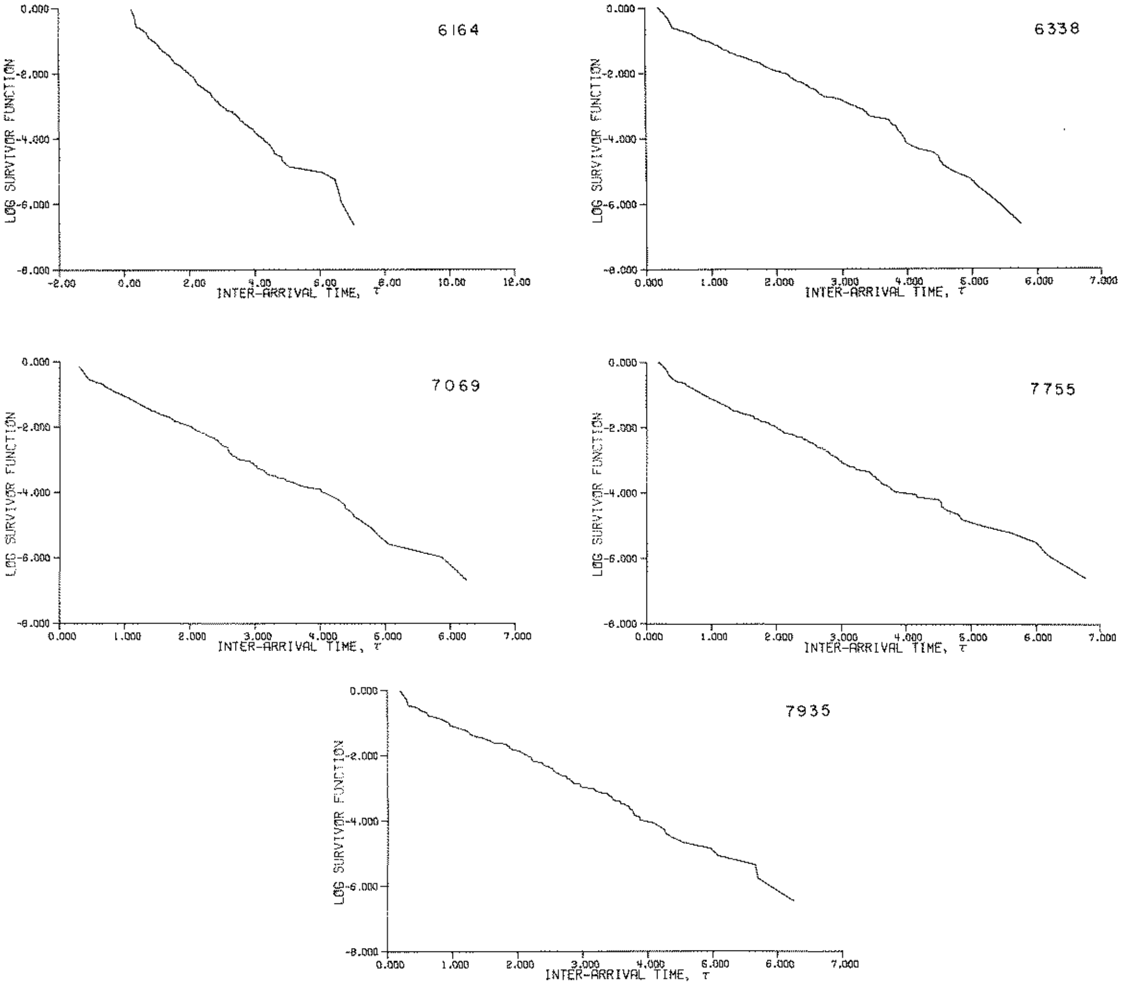


FIG. 14B—LOG SURVIVOR FUNCTION OF THE HOMOGENIZED DAILY RAINFALL COUNTS

CHAPTER 4 - TESTS FOR THE HYPOTHESIS THAT THE DAILY RAINFALL COUNTS PROCESS IS POISSON

4.1 TEST TO BE ACCOMPLISHED

Once the data are homogenized by the transformation (3.6), formal statistical tests of the Poisson hypothesis can be performed. Already, in the section dealing with the homogenization results, the null hypothesis of the independent counting increments was tested against the alternative of dependence in the increments through the use of the homogenized daily rainfall counts spectrum. In this section two types of testing will be accomplished. These are (a) the tests of the null hypothesis of independent interarrival times between the rainfall occurrences and (b) the general distribution-free tests of the Poisson hypothesis.

Since the transformation (3.6) can homogenize the whole process only under the Poisson assumption, the tests of the interarrival time independence will only be considered under the Poisson framework. For an ordinary renewal process it can be shown (*Cox,* 1962) that the Laplace transform of the renewal function, $E[N_t^0]$, is

$$L[E(N_t^0)] = H_0^*(s) = \frac{f_X^*(s)}{s[1 - f_X^*(s)]}$$

so that $E[N_t^0]$ depends specifically on the type of interarrival p.d.f. f(x) whose Laplace transform is $f_X^*(s)$. The interarrival time p.d.f. may have as many parameters as desired. These parameters may be time-dependent and there is no guarantee that the transformation (3.6) may homogenize these parameters. As will be seen below, the interarrival time independence tests are based on the autocorrelation and the spectrum of the interarrival time series which depend on the second moments of the interarrivals. The transformation (3.6) can homogenize the second moments of the interarrival times only in the case of the Poisson process. Therefore, the results based on the autocorrelation and the spectrum of the interarrival times of the homogenized daily rainfall counts are valid only under the Poisson hypothesis.

4.2 TESTS OF THE INTERARRIVAL TIME INDEPENDENCE UNDER THE POISSON HYPOTHESIS

4.2.1 Test Based on the Autocorrelation Coefficients of the Interarrival Times

The autocorrelation function of the interarrival times is calculated as

$$\hat{\rho}_J = \frac{n}{n-J} \frac{\sum_{i=1}^{n-J} (x_i - \bar{x})\,(x_{i+J} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{4.1}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. If n, the number of interarrivals in the data, is large, and provided that the interarrival time distribution is not highly skewed, then under the null hypothesis $\hat{\rho}_J = 0$, J = 1, 2, ..., the $\hat{\rho}_J$ can be considered to be normally distributed with zero mean and the standard deviation equal to $1/\sqrt{n-J}$ (*Cox and Lewis,* 1966). The correlation between the two autocorrelation coefficients at different lags is given by (*Bartlett,* 1955)

$$\mathrm{corr}(\hat{\rho}_J, \hat{\rho}_{J+h}) \quad \frac{\sum_{i=-\infty}^{\infty} \rho_i\, \rho_{i+k}}{\sum_{i=-\infty}^{+} \rho_i^2} \,. \tag{4.2}$$

For the independent intervals,

$$\mathrm{corr}(\hat{\rho}_J, \hat{\rho}_{J+k}) \approx O(1/n) \tag{4.3}$$

which shows that the autocorrelation estimates are correlated. Therefore, the tests based on the estimates of the autocorrelation coefficients of the interarrival times are not too reliable. The autocorrelation functions of the rainfall interarrival times for the homogenized data are shown in fig. 15, for the stations 0132, 3082, 3777, 4642, 6056 and 7747 for the positive lags. In table 4-1, $\hat{\rho}_J\sqrt{n-J}$ for $J = 1, \ldots, 5$, are given for the 17 stations analyzed. In 8 out of 17 cases studied some autocorrelation coefficient among the five lags was significant at the 5% level. Therefore, it is hard to conclude interval independence based on the autocorrelation function of the homogenized daily rainfall interarrival times.

4.2.2 Tests Based on the Spectrum of Interarrival Times

If the interarrival times sequence is considered as a time series, the spectral density function $f(\omega)$ for this sequence can be represented as

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} \rho_k \, e^{-ik\omega} \qquad -\pi \leq \omega \leq \pi \tag{4.4}$$

after the sequence is homogenized by the transformation (3.6). The spectral density is estimated by

$$f(\omega) = \frac{1}{2\pi} \sum_{J=-(n-1)}^{n-1} \hat{\rho}_J \,\mathrm{Cos}\, J\omega \quad , \quad -\pi \leq \omega \leq \pi \, . \tag{4.5}$$

For a process with uncorrelated interarrival times *Cox and Lewis* (1966) show that

$$E[\hat{f}(\omega)] \approx f(\omega)$$

$$\mathrm{Var}[\hat{f}(\omega)] \simeq [f(\omega)]^2 \;, \; 0 < \omega < \pi \qquad \mathrm{Var}[\hat{f}(\omega)] \approx 2[f(\omega)]^2 \;, \; \omega = 0, \pi$$

$$\mathrm{Cov}[\hat{f}(\omega_1), \hat{f}(\omega_2)] \approx 0, \; \omega_1 \neq \omega_2 \tag{4.6}$$

The spectral estimates $\hat{f}(\omega)$ are unbiased but inconsistent. To obtain consistency a spectral window is used and the estimate takes the form

$$\hat{f}(\omega) = \frac{1}{2\pi} \sum_{J=-\infty}^{+\infty} \lambda_J \, \hat{\rho}_J \mathrm{Cos}\, J\omega \;, \qquad -\pi \leq \omega \leq \pi$$

where λ_J is a weighting sequence to secure small variance and thereby consistency. However, as the variance of the estimate is decreased the bias is increased and there should be a trade-off between these two statistics. Among the various windows, Parzen's window where

$$\begin{aligned} \lambda_J = \lambda_{-J} &= 1 - \frac{6J^2}{\ell^2\left(1 - \frac{|J|}{\ell}\right)} \quad , \quad |J| \leq \ell/2 \\ &= 2\left(1 - \frac{|J|}{\ell}\right)^3 \quad , \quad \ell/2 < |J| \leq \ell \\ &= 0 \quad , \quad |J| > \ell \end{aligned}$$

in which ℓ is the cutoff lag of the autocorrelation function in the spectral density estimation, was used for the estimate of the spectral density in the computations. The expectation and variance of $f(\omega)$, using Parzen's window, are

$$E[\hat{f}(\omega)] \approx f(\omega) \;\cdot, \qquad \mathrm{Var}[\hat{f}(\omega)] \approx \frac{f^2(\omega)\ell}{2n}$$

where n is the total number of occurrences. For uncorrelated interarrival times the spectral estimates $\hat{f}(\omega)$ with the Parzen's window have the expectation and the variance,

$$E[\hat{f}(\omega)] \approx \frac{1}{2\pi} , \qquad Var[\hat{f}(\omega)] \approx \ell/8\pi^2 n . \tag{4.7}$$

The interarrival time spectra for the homogenized daily rainfall occurrences are shown on fig. 16 for the stations 0132, 3082, 3777, 4642, 6056 and 7747 in Indiana. The cutoff lag of the autocorrelation function in these spectral density estimates was $\ell = 40$. The "frequency index" on the abcissa of these plots denotes the integers $p = 0, 1, 2, \ldots, \left(\frac{n-1}{2}\right)$ corresponding to the frequencies $\omega_p = 2\pi p/n$. There is no definite structure on these spectra. Some spectral values which seem significant do not correspond to any physical time period since the abcissa scale corresponds to serial numbers of the occurrences. From these spectra one can check the statistical characteristics (4.6) for the uncorrelated intervals.

More formal tests are based on the periodogram estimates $I_n(\omega_p)$ where

$$I_n(\omega_p) = \frac{1}{2\pi n} \left| \sum_{J=1}^{n} X_J \, e^{iJ\omega_p} \right|^2 , \ -\pi \leq \omega_p \leq \pi , \ p = 1, 2, \ldots, \ \frac{n-1}{2} . \tag{4.8}$$

When the process of interarrival times {X} is made up of uncorrelated increments the periodogram estimates have the properties (*Cox and Lewis,* 1966)

$$\lim_{n\to\infty} P[I_n(\omega_p) \geq x] = \exp -\left\{\frac{x}{\sigma^2 f(\omega_p)}\right\} ,$$

$$E[I_n(\omega_p)] = \sigma^2 f(\omega_p) ,$$

$$\lim_{n\to\infty} Var[I_n(\omega_p)] = \sigma^4 \, f^2(\omega_p) ,$$

$$\lim_{n\to\infty} Cov[I_n(\omega_{p_1}), I_n(\omega_{p_2})] = 0 , \text{ for } 0 < \omega_p < \pi . \tag{4.9}$$

where σ^2 is the variance of the interarrival times $\{X_i\}$. As is seen from the above expressions $I_n(\omega_p)$ is not a consistent estimator. At $\omega_p = 0$;

$$E[I_n(0)] = \sigma^2 \, f(\omega_p) + \frac{n\mu^2}{2\pi}$$

where μ is the expectation of the $\{X_i\}$. Therefore, there will be large value at the origin when the process under consideration has a nonzero mean.

4.2.2.1 Test of trend in the periodogram

The formal tests of the interval independence have the null hypothesis H_0 that the periodogram estimates $I(\omega_p)$ for $-\pi < \omega_p < \pi$, and $\omega_p \neq 0$, have asymptotically uncorrelated exponential distributions with mean $\sigma^2/2\pi$. The alternative hypothesis H_A is that the periodogram estimates $I_n(\omega_p)$ have trends instead of being close to the horizontal line $I_n(\omega_p) = \sigma^2/2\pi$.

One can put $I(\omega_p)$, $p = 1, 2, \ldots, \ell$, where ℓ is the integer part of $\frac{n-1}{2}$, end to end and form an artificial counting process where the waiting times t_i are defined as

$$t_i = \sum_{j=1}^{i} I(\omega_J) . \tag{4.10}$$

One can test the time trends in the rate of occurrence of this artificial process in a manner as described in the section dealing with the tests of trend in the daily rainfall occurrences. Since the intervals are exponential and asymptotically uncorrelated, one can assume a Poisson model for the constructed artificial process. Then, for the test in sec. 2.2.1 the statistic S will become

$$S = \sum_{i=1}^{\ell} \sum_{J=1}^{i} I(\omega_J)/\ell \tag{4.11}$$

and

$$U = (S - \frac{1}{2} t_\ell)/t_\ell/\sqrt{12\ell}$$

where ℓ is the integer part of $(n-1)/2$. If there are trends in the rate of occurrence, this would mean that there are trends in the periodogram estimates $I(\omega_p)$. This test was performed for the 17 stations under investigation. The results are shown in the last column on table 4.1. Three out of 17 cases are significant at 5%.

4.2.2.2 Distribution-free tests of the interval independence on the homogenized daily rainfall occurrences

After the periodogram estimates are computed one can obtain the normalized cumulative periodogram values $U_{(i)}$ as

$$U_{(i)} = \sum_{J=1}^{i} I(\omega_J) / \sum_{J=1}^{\ell} I(\omega_J) \qquad i = 1, \ldots, \ell$$

where ℓ is the integer part of $(n-1)/2$. Under the renewal process hypothesis the periodogram estimates are asymptotically independent, and therefore, the normalized cumulative periodogram values $U_{(i)}$ become the order statistics from a uniformly distributed sample of size ℓ over (0,1) such that

$$P[U_i \leq u] = \begin{cases} 0 \ , & u \leq 0 \\ u \ , & 0 < u \leq 1 \\ 1 \ , & u > 1 \ . \end{cases}$$

This is the canonical form for the distribution-free tests of goodness-of-fit for the null hypothesis H_0;

$$H_0: \quad F_X(x) = F_0(x)$$

where $F_0(x)$ is the assumed distribution function (which is the uniform distribution in the case under study) and $F_X(x)$ is the unknown underlying probability distribution. $F_\ell(x)$ will denote the empirical distribution function of the random variable X of sample size ℓ. In the case under study ℓ is the integer part of $(n-1)/2$. The distribution-free tests are thoroughly treated by *Cox and Lewis* (1966). One can form three alternative hypotheses against H_0;

H_{A1}: $F_X(x) \neq F_0(x)$

H_{A2}: $F_X(x) > F_0(x)$

H_{A3}: $F_X(x) < F_0(x)$.

A test of H_0 against H_{A1} is a two-sided test of goodness-of-fit and H_0 versus H_{A2}, H_0 versus H_{A3} are one-sided tests of the goodness-of-fit. In relation to testing the uniformity of U_i's Kolmogorov-Smirnov and Anderson-Darling statistics will be considered.

i. Kolmogorov-Smirnov statistics;

One-sided statistics are

$$D_\ell^+ = \underset{-\infty<x<\infty}{\text{Sup}} \{F_\ell(x) - x\} = \max_{1\le i\le \ell} \{\frac{i}{\ell} - x_{(i)}\} \text{ and}$$

$$D_\ell^- = \underset{-\infty<x<\infty}{\text{Sup}} \{x - F_\ell(x)\} = \max_{1\le i\le \ell} \left\{x_{(i)} - \frac{(i-1)}{\ell}\right\}$$

where $x_{(i)}$ is the i-th order statistic from the random sample. The two-sided statistic is

$$D_\ell = \underset{-\infty<x<\infty}{\text{Sup}} |F_\ell(x) - x| = \max(D_\ell^+, D_\ell^-) .$$

The asymptotic distribution of one-sided statistics can be expressed as

$$\lim_{\ell\to\infty} P[D_\ell^+\sqrt{\ell} \le d] = \lim_{\ell\to\infty} P[D_\ell^-\sqrt{\ell} \le d] = 1 - e^{-2d^2} , \qquad d \ge 0 .$$

For a significance level α the asymptotic probability limit d_α such that $P[D_\ell\sqrt{\ell} > d_\alpha] = \alpha$, can be obtained from tables (*Cox and Lewis,* 1966), (*Box and Jenkins,* 1970). If the calculated value of D_ℓ is greater than d_α, the null hypothesis is rejected at the level α. In the case under study;

H_0: $F_U(u)$ is uniform in (0,1)

H_A: There is a trend in U_i, that is, U_i are not uniformly distributed in (0,1).

A two-sided Kolmogorov-Smirnov test follows;

$$D_\ell = \max\left\{\max_{1<i\le\ell} \left|\frac{i}{\ell} - U_{(i)}\right| , \max_{1<i<\ell} \left|U_{(i)} - \frac{(i-1)}{\ell}\right|\right\}$$

where D_ℓ^+ is the first maximum in the brackets and D_ℓ^- is the second maximum. If $D_\ell\sqrt{\ell} > d_\alpha$, or if $D_\ell^-\sqrt{\ell} > d_\alpha^-$, or if $D_\ell^+\sqrt{\ell} > d_\alpha^+$, then the hypothesis that the homogenized daily rainfall interarrival times are independent is rejected at the level α.

The Kolmogorov-Smirnov statistics are shown on the table 4-1. The interval independence is rejected in 5 out of 17 cases at the 5% level.

ii. Anderson-Darling Statistic;

It is seen above that the Kolmogorov-Smirnov statistics measure the differences between the empirical distribution $F_\ell(x)$ and the assumed probability distribution $F_0(x)$. Another way to measure the difference between the two distributions is through the Cramer-Von Mises statistic, (*Cox and Lewis,* 1966)

$$w_\ell^2 = \int_{-\infty}^{+\infty} \{F_\ell(x) - F_0(x)\}^2 \, dF_0(x)$$

which measures a mean square deviation between $F_\ell(x)$ and $F_0(x)$. This statistic is again distribution-free and gives a consistent test of the two-sided hypothesis. In computational form,

$$w_\ell^2 = \frac{1}{12\ell^2} + \frac{1}{\ell} \sum_{i=1}^{\ell} \left\{F_0(x_{(i)}) - \frac{(2i-1)}{2\ell}\right\}^2 .$$

Anderson and Darling (1952, 1954) gave a tabulation of

$$\lim_{\ell\to\infty} P[\ell w_\ell^2 \le w]$$

and the significance points for α = .05 and .01 are given in *Cox and Lewis* (1966).

$\{F_\ell(u) - u\}$ has mean zero and variance $u(1-u)/\ell$. Therefore, the maximum value of the variance happens at $u = 1/2$ and it is expected that D_ℓ and w_ℓ^2 are most sensitive to departures in the middle of the range (0,1) of u. In order to have w_ℓ^2 equally sensitive all through (0,1) *Anderson and Darling* (1952, 1954) weighted w_ℓ^2 by the reciprocal of the variance $u(1-u)/\ell$ to form the Anderson-Darling statistic W_ℓ^2 where

$$W_\ell^2 = \int_0^1 (F_\ell(u) - u)^2 \cdot \frac{\ell}{u(1-u)} du .$$

In the computational form

$$W_\ell^2 = -\ell - \frac{1}{\ell} \sum_{i=1}^{\ell} \{(2i-1) \log U_i + (2\ell - 2i+1) \log (1-U_i)\}$$

for testing the null hypothesis that U_i, $i = 1, \ldots, \ell$ divide the interval (0,1) at random. If the calculated value of W_ℓ^2 exceeds the asymptotic significance point at level α, then H_0 is rejected at this level. This, in turn, implies that the interarrival times of the homogenized daily rainfall counts are correlated. The Anderson-Darling test for the homogenized daily rainfall intervals rejected the interval independence hypothesis in 3 out of 17 cases at 5% level.

4.2.3 Results of the Tests for the Interval Independence Hypothesis

The tests considered in this section are all based on the large sample distributions and on certain approximations. The distribution-free tests on the periodogram of the intervals are tests against general alternative hypotheses. If there were specific alternatives, it would be possible to derive more powerful tests than the distribution-free tests. For a general alternative such as the dependence of the intervals one hopes that the distribution-free tests considered above will have a good power on the average. The tests were performed on the homogenized daily rainfall data which were obtained under the Poisson hypothesis. If the underlying process is not Poisson, the conclusions based on these tests are not necessarily valid.

The results of the tests of the interval independence hypothesis for the homogenized daily rainfall occurrences are given on the Table 4-1. Out of the 17 cases studied the interval independence is rejected in 5 cases by the distribution free tests. There are 3 more cases where some autocorrelation coefficient was found significant in the first five lags. However, in the majority of the cases the independence of the homogenized daily rainfall interarrival times is accepted by the tests constructed on the autocorrelation function and on the periodogram of the interarrival times. Based on these results it can be concluded that a point stochastic model with independent intervals can only approximate the daily rainfall counts process and cannot be taken as a general model.

4.3 TESTS OF THE POISSON HYPOTHESIS FOR THE HOMOGENIZED DAILY RAINFALL COUNTS USING THE DIRECT POISSON CHARACTERISTICS

4.3.1 The Uniform Conditional Test

In the analysis of the homogenized daily rainfall counts the series of occurrences are observed up to the r-th event where r will be taken to be n+1 for convenience. Consider the waiting times t_i, $i = 1, \ldots, n$ to the events in the interval $(0, t_{n+1})$. Given that n events have occurred in $(0, t_{n+1})$, the conditional p.d.f. that events occur at $t_1 \leq t_2 \leq \ldots \leq t_n$ is

$$\lambda e^{-\lambda t_1} \lambda e^{-\lambda(t_2-t_1)} \ldots \lambda e^{-\lambda(t_n-t_{n-1})} e^{-\lambda(t_{n+1}-t_n)} / e^{-\lambda t_{n+1}} (\lambda t_{n+1})^n/n! = \frac{n!}{t_{n+1}^n} , 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq t_{n+1}$$

where λ is the rate of occurrence. The joint p.d.f. of the random variables $U_{(i)} = t_i/t_{n+1}$, $i = 1, \ldots, n$, in $(0,t_{n+1})$ is $n!/t_{n+1}^n$ since they are the order statistics of n random variables $U_1, \ldots, U_n$ independently and uniformly distributed in $(0,t_{n+1})$. Therefore, the test can be based on the random variables $U_{(i)}$, $i = 1, \ldots, n$, given that n events have occurred in $(0, t_{n+1})$. This test is called the uniform conditional test for a Poisson process. $U_{(i)}$, $i = 1, \ldots, n$, can be considered as the order statistics from a random sample size n from a uniformly distributed population in (0,1) so that

$$P[U_i \leq u] = \begin{array}{l} 0 \ , \ u \leq 0 \\ u \ , \ 0 < u \leq 1 \\ 1 \ , \ u > 1 \ . \end{array}$$

This is the canonical form of the distribution-free tests discussed in the previous section. Therefore, Kolmogorov-Smirnov and Anderson-Darling statistics are used to test the Poisson hypothesis by employing the order statistics $U_{(i)}$.

These tests are shown on the table 4-2 for the 17 stations under investigation in Indiana. The empirical distribution function $F_n(u)$ of the U_i's is proportional to the graph of the cumulative number of events against cumulative time to the last event which was considered in the trend analysis of the daily rainfall occurrences. Therefore, the distribution-free tests based on U_i's will be sensitive basically to trend alternatives in the data. However, the data was homogenized as to remove the trends under the Poisson hypothesis. It was seen in the graphs of the number of rainy days versus cumulative time in fig. 7 that the trend which the $U_{(i)}$'s are testing, is effectively removed. The trend test for the rate of occurrence on the homogenized daily rainfall data did not show any time dependence either. Therefore, it is reasonable to expect that the uniform conditional tests based on Kolmogorov-Smirnov and Anderson-Darling statistics will accept the Poisson hypothesis. As seen from the test results on the table 4-2, the Poisson hypothesis for the daily rainfall counts process was accepted in all of the 17 stations.

If the null distribution $F_0(x)$ was parametric, there is no known modification of the distribution-free tests based on the Kolmogorov-Smirnov and Anderson-Darling statistics to account for the estimation of the parameters from the data. *Durbin* (1961) transformed the observations for eliminating the parameters so that the distribution-free tests can be safely used. In the testing of Poisson hypothesis by Kolmogorov-Smirnov and Anderson-Darling tests the null distribution $F_0(x)$ is already free of parameters so that the advantage of Durbin's modification in this case is not clear. The condition that the distribution-free tests based on Durbin's modification of the data to be more powerful than the same distribution-free tests based on the untransformed data is that the coefficient of variation $C(X)$ of the interarrival times X should be greater than unity (*Durbin,* 1961). This condition is strictly true if there are no evolutionary trends in the data. It is also expected that the tests based on Durbin's modification have more power for $C(X) < 1$ in the uniform conditional test of the Poisson hypothesis (*Cox and Lewis,* 1966). In the homogenized daily rainfall data the coefficient of variation for the interarrival times is greater than unity in 11 out of 18 cases, as is seen in the table 3-3. Therefore, in these 11 cases Kolmogorov-Smirnov and Anderson-Darling tests based on Durbin's modification of the data will be more powerful than their counterparts based on the untransformed data. Following the empirical evidence of *Cox and Lewis* (1966) the distribution-free tests using Durbin's modification are more powerful than their counterparts based on the untransformed data when the null Poisson hypothesis is tested against stationary general alternatives. This is the case in the homogenized daily rainfall counts process where the trends are removed under the Poisson hypothesis.

Durbin's modification of the data corresponding to the uniform conditional test statistic U_i is described by Cox and Lewis (1966). Let $X_1, \ldots, X_{n+1}$ be the homogenized daily rainfall interarrival times under the Poisson assumption. Then X_i, $i = 1, \ldots, n+1$, such that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n+1)}$ are formed from X_i, $i = 1, \ldots, n+1$. It can be shown that $\{X_{(i)} - X_{(i-1)}\}$ for $i = 1, \ldots, n+1$ are independent, exponentially distributed with mean $1/\{\lambda(n+1-i+1)\}$. Then the variables $(n+2-i) \cdot \{X_{(i)} - X_{(i-1)}\}$ will be the interarrival times in a Poisson process with the rate of occurrence λ. As was shown by Durbin (1961), the variables $U'_{(i)}$ such that

$$U'_{(i)} = \frac{1}{t_{n+1}} \{X_{(1)} + X_{(2)} + \ldots + (n+2-i)X_{(i)}\}, \ i = 1, \ldots, n$$

will be the order statistics of the random variables U_i' which are independent, uniformly distributed in (0,1). Therefore, Kolmogorov-Smirnov and Anderson-Darling tests can again be employed with the $U_{(i)}'$, $i = 1, \ldots, n$. The results of Kolmogorov-Smirnov and Anderson-Darling tests based on Durbin's modification are given in the table 4-2. The tests reject the Poisson hypothesis in all of the 17 cases under investigation. The test results are in agreement with the tests of the Poisson hypothesis by the spectra of counts and the variance-time curves which strongly reject the Poisson model for the daily rainfall counts process.

4.3.2. Tests of Poisson Hypothesis Against Renewal Alternatives

The only test which will be considered in this section is Moran's test. This is a test of the hypothesis of independent, exponentially distributed interarrival times against the alternative of independent gamma distributed interarrival times. Based on the likelihood function of n independent, gamma distributed interarrival times Moran (1951) derived the asymptotically most powerful test statistic ℓ_n for the above hypothesis as

$$\ell_n = \frac{2n(\log \bar{x} - \frac{1}{n}\sum_{i=1}^{n} \log x_i)}{1 + (\frac{n+1}{6n})}$$

where, under the null hypothesis, ℓ_n is approximately χ^2 distributed with (n-1) degrees of freedom. The divisor in the above expression for ℓ_n was introduced by Bartlett to improve the approximation to the chi-squared distribution. The test results for the homogenized daily rainfall counts are given in the table 4-2 and favor the exponentially distributed interarrivals against the gamma alternative in all of the 17 cases.

4.4 SUPPLEMENTARY FIGURES

The autocorrelation functions and the spectra for the intervals between homogenized daily rainfall counts for the stations 0177, 0545, 1747, 1882, 3082, 3547, 4908, 6164, 6338, 7069, 7755 and 7935 are shown on the figures 15A, 15B, and 16A, 16B.

TABLE 4-1

TEST OF THE INDEPENDENCE OF HOMOGENIZED DAILY RAINFALL INTERARRIVAL TIMES

	Test Based on the Autocorrelation Function of the Interval Time Series					Tests on the Periodogram of Interarrival Times				
						Distribution-free Tests				Test of Trend
						Kolmogorov-Smirnov Test			Anderson-Darling Test	
Station	$\hat{\rho}_1\sqrt{n-1}$	$\hat{\rho}_2\sqrt{n-2}$	$\hat{\rho}_3\sqrt{n-3}$	$\hat{\rho}_4\sqrt{n-4}$	$\hat{\rho}_5\sqrt{n-5}$	$D_n^+\sqrt{\ell}$	$D_n^-\sqrt{\ell}$	$D_n\sqrt{\ell}$	W_n^2	U
0132	-2.117+	2.347	- .639	1.084	- .211+	1.37+	.20	1.37+	3.38+	-2.08+
0177	- .443	1.553	.716	1.016	.681	.71	.61	.71	.71	- .24
0545	- .157	1.154	1.476	1.571	.351	.61	.83	.83	.80	.13
1747	-1.396	2.019+	1.118	- .364	2.104+	1.46+	.61	1.46+	1.83	-1.10
1882	-1.054	1.908	- .04	.232	-1.939	1.10	.45	1.10	1.49	-1.03
3082	- .249	.049	2.561	.528	.511	.68	.73	.73	.85	.16
3547	- .271	- .359	.831	1.256	-1.282	.65	.58	.65	.40	-1.6
3777	- .4506	- .221	.836	- .610	- .325	.55	.45	.55	.37	- .10
4642	- .796	- .856	-1.776	.004	- .583	.74	.17	.74	.92	- .73
4908	-1.726	-1.238	.340	2.249+	-1.693	1.28+	.20	1.28	1.91	-1.55
6056	- .934	1.226	.771	- .189	.824	.93	.38	.93	.89	- .68
6164	-2.737++	2.334+	.232	- .157	1.263	1.83++	.22	1.83++	4.38++	-2.57++
6338	-1.234	.894	.642	2.339+	.034	1.02	.43	1.02	1.38	-1.21
7069	-2.769++	1.28	- .573	-1.167	.430	1.67++	.23	1.67++	4.27++	-2.67++
7747	.214	1.312	.902	1.249	- .236	.46	.95	.95	.77	.36
7755	- .457	.250	.77	- .13	2.148+	.75	.46	.75	.57	- .24
7935	- .844	.560	- .783	2.367+	- .679	.86	.36	.86	1.20	- .90

n = sample size
ℓ = integer part of (n-1)/2
+ = significant at 5% level
++ = significant at 1% level

TABLE 4-2

TESTS FOR POISSON PROCESSES
(Homogenized data)

		Tests before Durbin's Modification		Tests based on Durbin's Modification		
		Kolm.-Smir. Stat.	Anderson-Darling Stat.	Kolm.-Smir. Stat.	Anderson-Darling Stat.	Moran Stat.
Station	n	$\sqrt{n}\,D_n$	W_n^2	$\sqrt{n}\,D_n$	W_n^2	ℓ_n
0132	793	.99	.816	5.7++	34.52++	604.49
0177	719	.818	.788	4.517++	28.65++	594.83
0545	771	1.05	1.31	6.089++	35.21++	616.23
1747	761	1.08	1.147	5.35++	29.7++	611.17
1882	587	.649	.687	3.627++	13.28++	497.31
3082	797	.656	.557	5.83++	33.3++	590.5
3547	616	.608	.328	4.39++	21.5++	568.7
3777	762	1.19	2.21	6.59++	34.71++	603.12
4642	776	.44	.1974	5.62++	31.8++	595.45
4908	738	.521	.412	4.977++	26.6++	580.34
6056	705	.859	.666	4.155++	27.8++	627
6164	779	1.005	1.08	6.34++	35.1++	608
6338	763	.718	.74	5.55++	30.4++	642
7069	800	.582	.38	4.28++	22.4++	562
7747	784	1.178	1.996	6.236++	33.5++	579.46
7755	736	.969	1.32	4.968++	26.64++	584
7935	651	.833	1.247	5.06++	21.54++	561.73

Significance Points:

	α=.05	α=.01
$\sqrt{n}\,D_n$	1.358	1.628
W_n^2	2.492	3.857

$$\ell_n \sim \chi^2_{n-1} \rightarrow \lim_{n\to\infty} \sqrt{2\ell_n} \sim N(\sqrt{2n-1},\ 1)$$

++ Statistic significant at 1% level

+ Statistic significant at 5% level

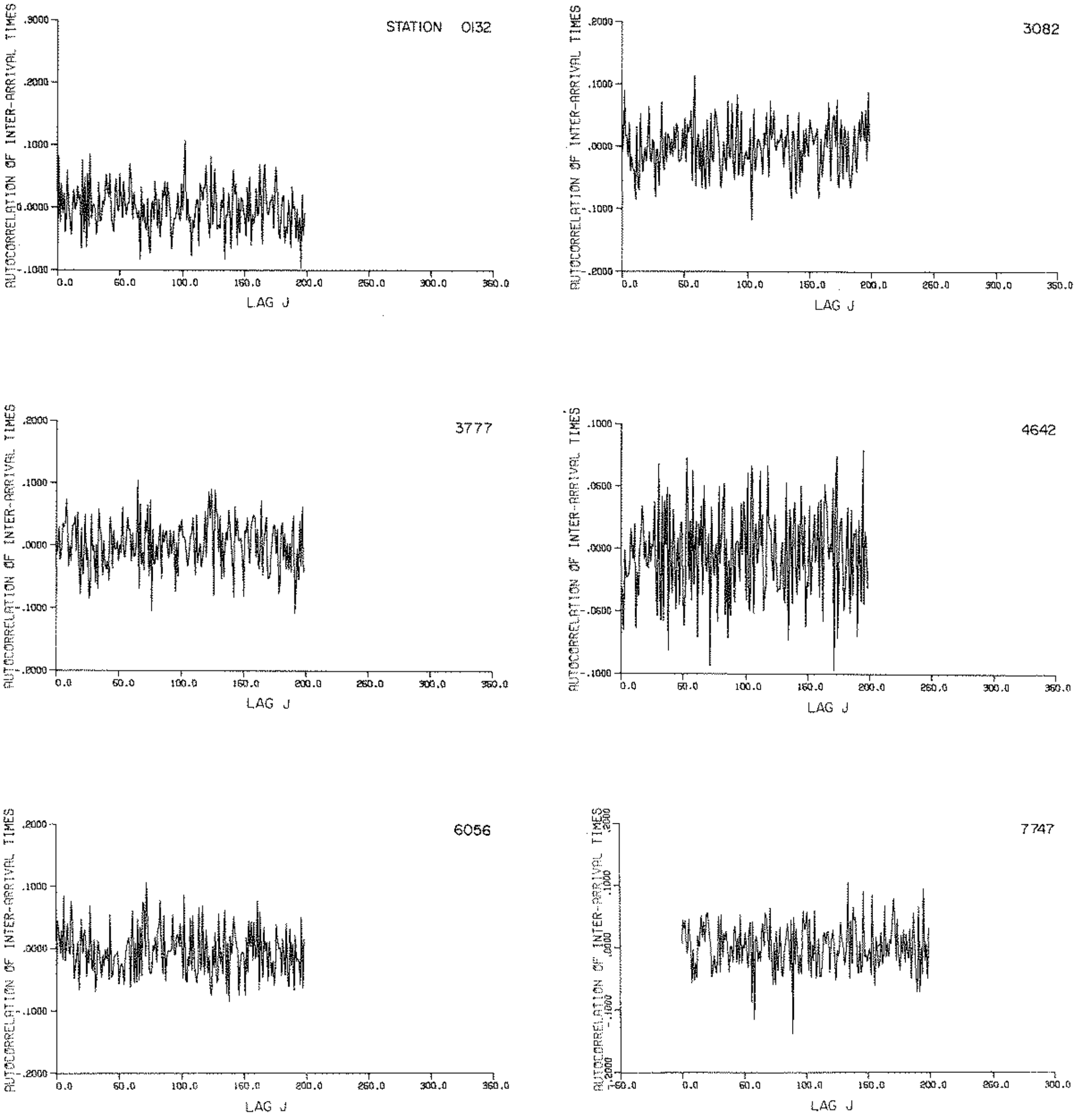


FIG. 15- AUTOCORRELATION FUNCTION FOR THE INTERVALS BETWEEN HOMOGENIZED DAILY RAINFALL COUNTS

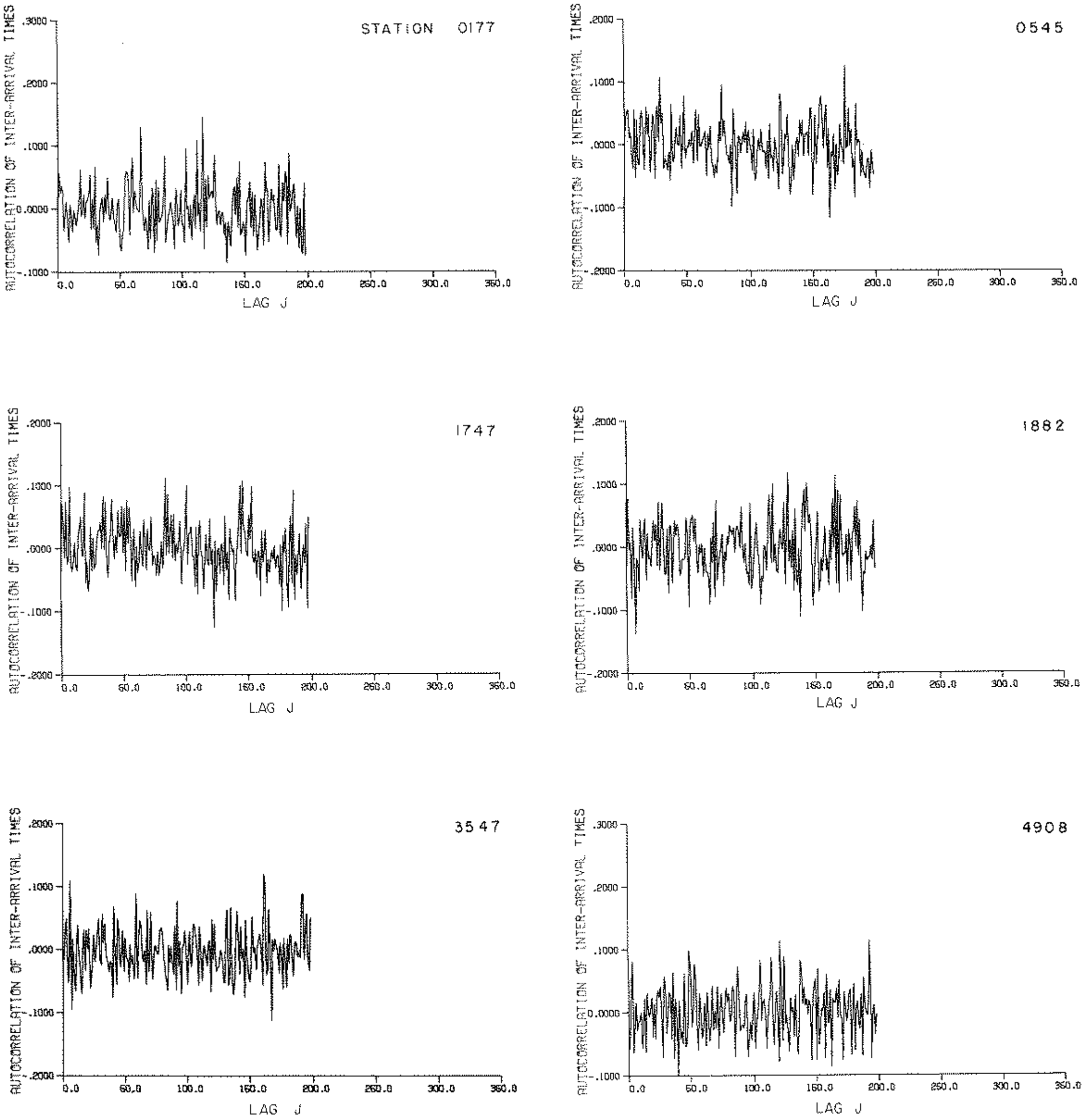


FIG. 15A—AUTOCORRELATION FUNCTION FOR THE INTERVALS BETWEEN HOMOGENIZED DAILY RAINFALL COUNTS

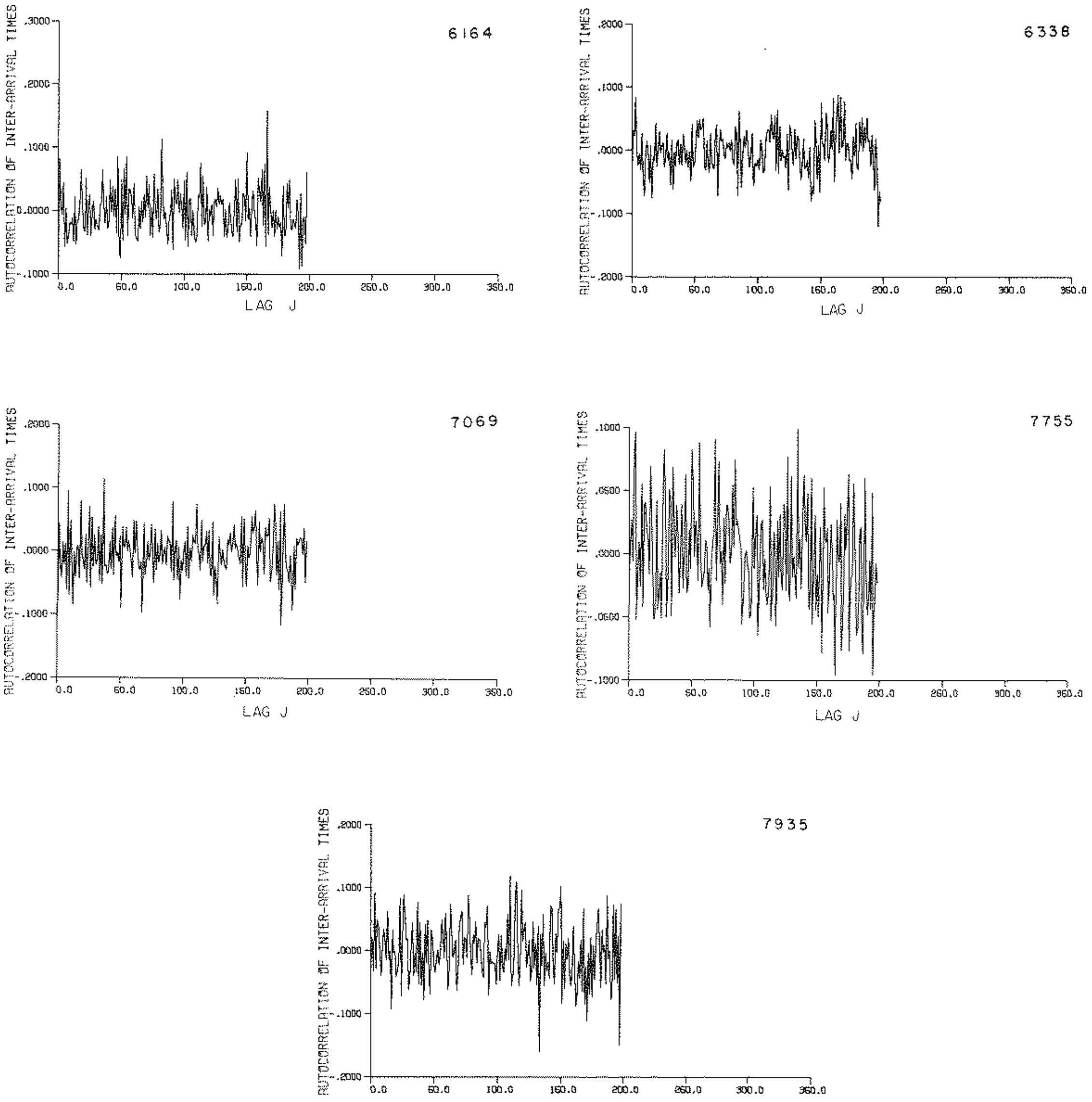


FIG. 15B—AUTOCORRELATION FUNCTION FOR THE INTERVALS BETWEEN HOMOGENIZED DAILY RAINFALL COUNTS

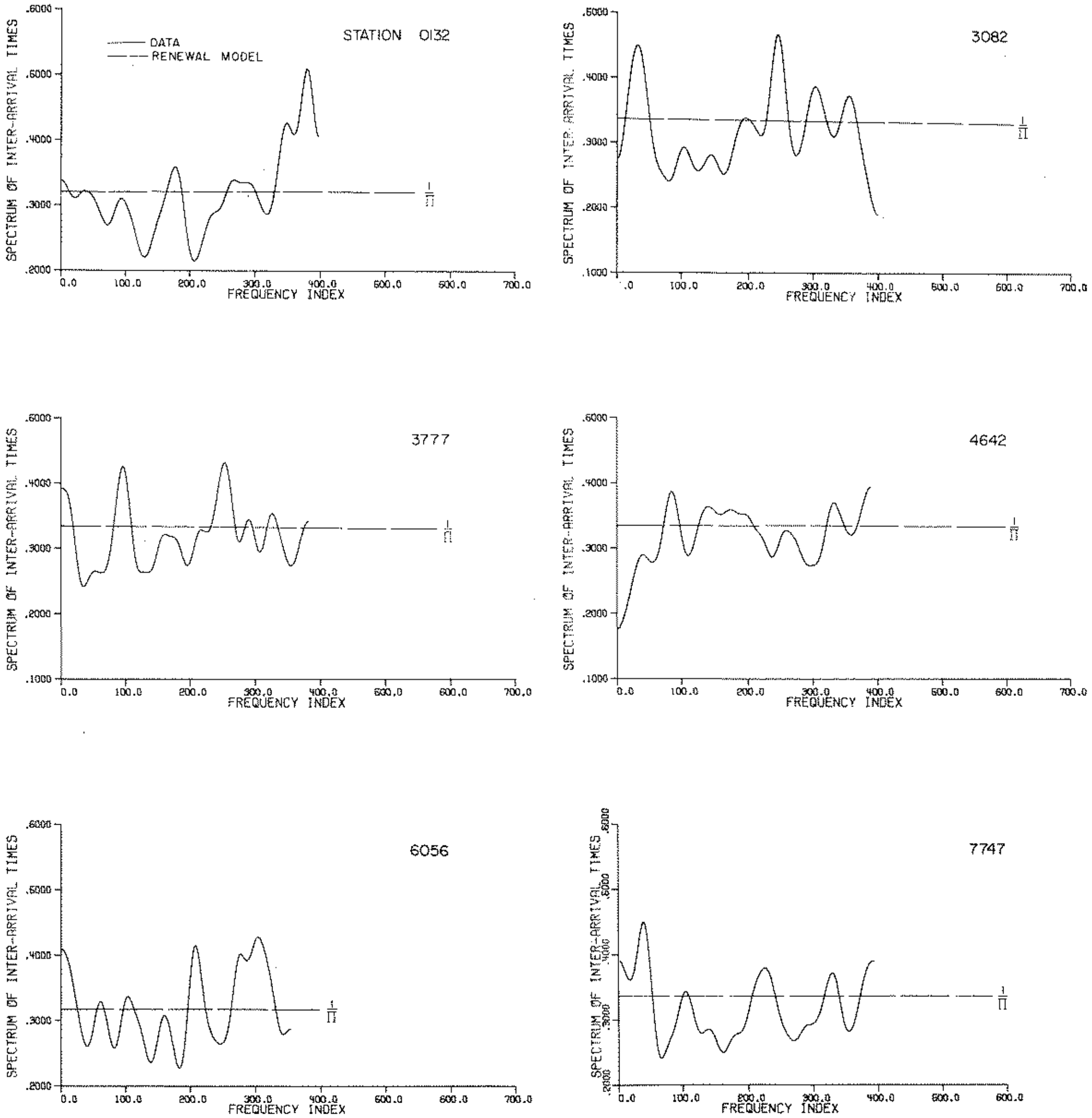


FIG. 16-SPECTRA FOR THE INTERVALS BETWEEN HOMOGENIZED DAILY RAINFALL COUNTS

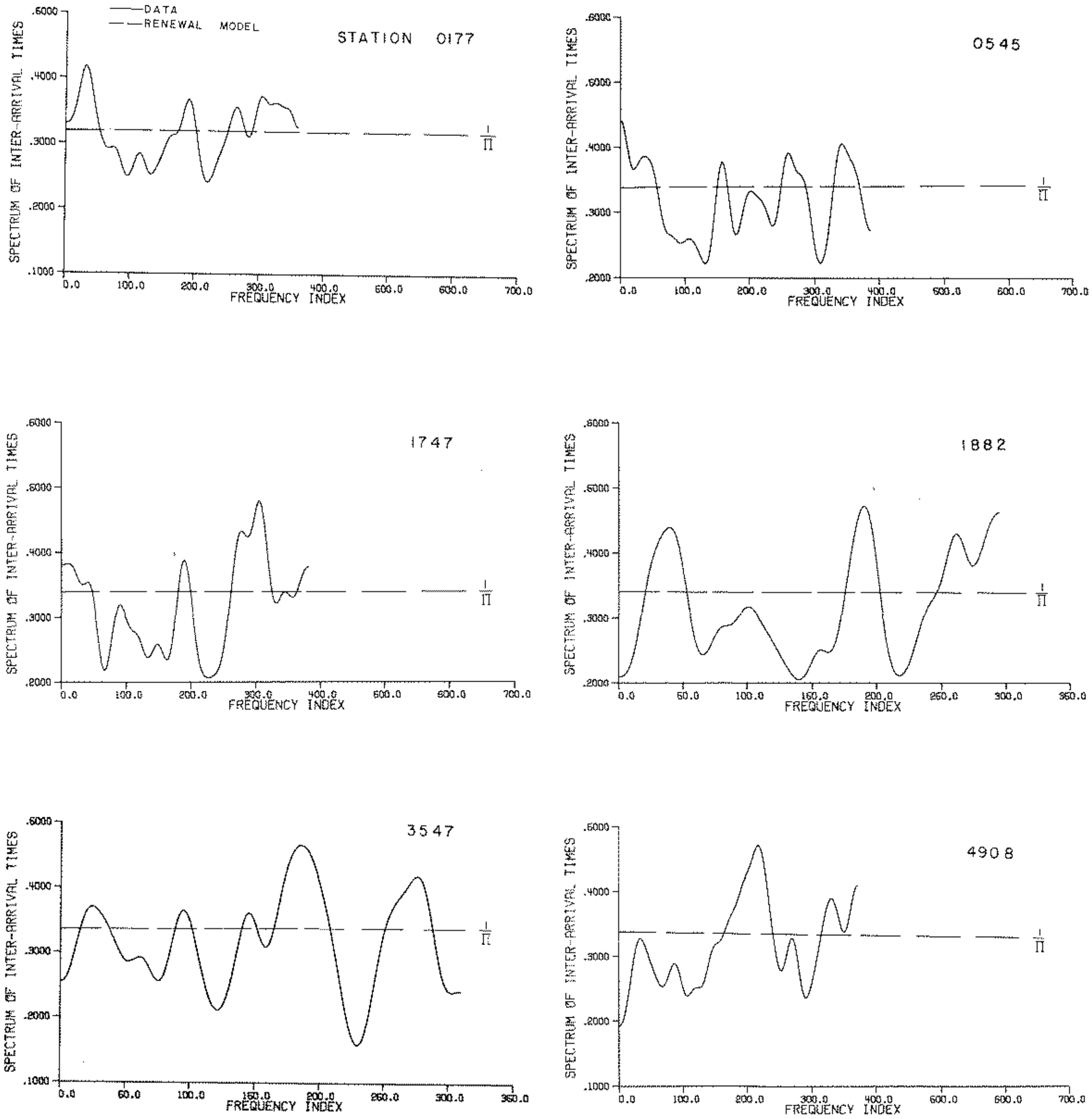


FIG. 16A–SPECTRA FOR THE INTERVALS BETWEEN HOMOGENIZED DAILY RAINFALL COUNTS

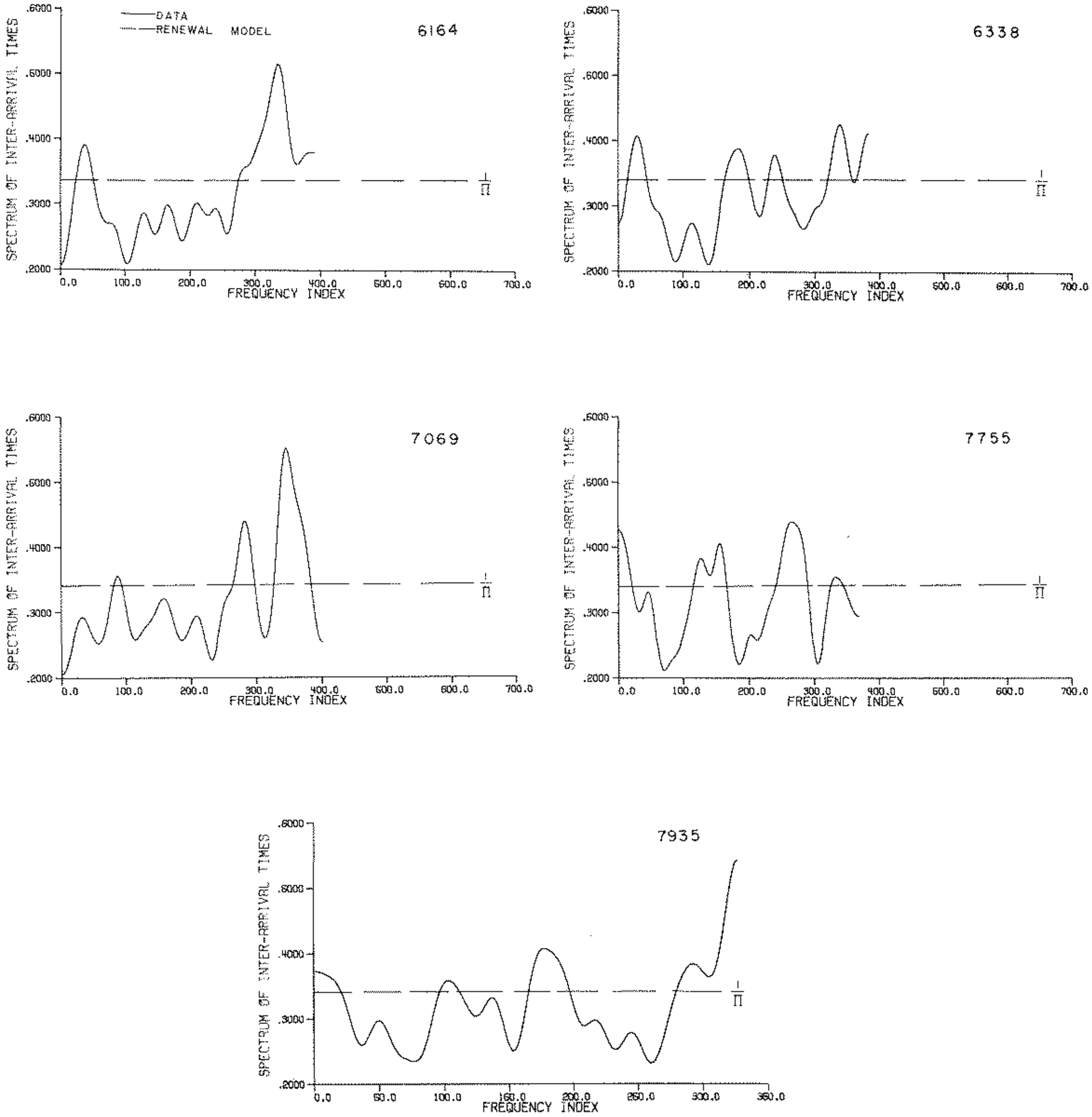


FIG. 16B–SPECTRA FOR THE INTERVALS BETWEEN HOMOGENIZED DAILY RAINFALL COUNTS

CHAPTER 5 - DISCUSSION OF THE RESULTS FROM THE ANALYSIS OF THE HOMOGENIZED DAILY RAINFALL DATA

Once the data were homogenized by transformation (3.6) under the Poisson assumption the homogenized daily rainfall counts process was analyzed in Chapters 3 and 4 by several statistical functions and by several statistical tests of hypothesis.

The statistical test on the trend of the rate of rainfall occurrence showed that the rate of rainfall occurrence is homogenized. This result is supported by the plots of the number of rainy days versus cumulative time shown in fig. 7 and by the plots of the rainfall counts intensity functions shown in fig. 8. However, the homogeneity of variance test on the homogenized daily rainfall interarrival times rejected the variance homogeneity of the daily rainfall interarrival times. The variance-time function of the homogenized daily rainfall counts also strongly rejected the variance homogeneity by significantly deviating from the Poisson case in all cases. This behavior cast a shadow on the utility of the Poisson process for modeling the daily rainfall counts. The behavior of the spectra of the homogenized daily rainfall counts was completely different from the theoretical counts spectra of the Poisson process in all of the analyzed stations. The deviations from the Poisson hypothesis were significant at 1% level in all of the rainfall stations.

The behavior of the spectra and of the variance-time curves of the homogenized daily rainfall counts showed a dependence structure in the counting increments which can be explained by the clustering of the rainfall occurrences around a rainfall-generating mechanism. It will be seen in the following chapter that the Neyman-Scott cluster model has the same type of theoretical behavior as is observed in the spectra and in the variance-time functions of the homogenized daily rainfall counts in Indiana. The log-survivor functions of the homogenized daily rainfall interarrival times showed convexity in 11 out of 17 cases studied and their behavior was consistent with the results of their variance-time and counts spectrum counterparts.

The tests of the independence of homogenized daily rainfall interarrival times accepted the independence hypothesis in the majority of the cases. However, five cases showed dependence of the rainfall intervals and three cases were doubtful. The uniform conditional tests of the Poisson hypothesis based on Durbin's modification of the homogenized daily rainfall data strongly rejected the Poisson hypothesis in all of the 17 cases.

Based on the behavior of the statistical functions and on the results of the statistical hypothesis tests, the Poisson process cannot be accepted as a model for the daily rainfall counts process. There is definite dependence in the counting increments of the daily rainfall occurrences and point stochastic models that can explain this behavior are necessary.

CHAPTER 6 - CLUSTER MODEL FOR THE RAINFALL COUNTS PROCESS

6.1 DESCRIPTION OF THE CLUSTER MODEL

From the statistical tests of the Poisson hypothesis, and from the behavior of the homogenized daily rainfall counts spectrum and of the variance-time curve, it was concluded that the counting increments of the homogenized daily rainfall occurrences were dependent and cannot be modeled by a Poisson process. What is needed is a point stochastic model that can explain the dependence in the counting increments of the daily rainfall occurrences. The Neyman-Scott cluster model (*Neyman and Scott,* 1958) can explain that dependence. This model assumes that the rainfall events occur in clusters which are the storms and that the occurrence of rainfall events in a certain time interval (0,T) is not only caused by the occurrence of the rainfall generating mechanisms in the time interval (0,T) but is also due to the rainfall generating mechanisms which occurred prior to the interval (0,T). This is due to the fact that a rainfall that belongs to a storm that was generated by a mechanism which had occurred before (0,T) may still fall in the interval (0,T) due to the memory of the storm generating mechanism.

A storm generating mechanism is any meteorological condition that would cause rainfall. For example, when a hot and humid air front meets a cold air front and rises, the humidity in the warm air front starts to condense and a rainfall-generating mechanism is born. As long as this mechanism exists over a region, there is a probability that it may cause rainfalls in that region. The existence of this mechanism for a long time would correspond to a long memory or a long range dependence. If the mechanism exists for a very short time and causes a thunderstorm as is seen in the arid regions, the memory is very short and the process may be approximated by independent counting increments.

The rainfall cluster model is described under the following 5 assumptions:

1. The rainfalls occur in the form of clusters on the time axis.
2. The cluster origins in the counting process $\{N_t\}$ of the rainfall occurrences are the positions of the occurrences of the rainfall generating mechanisms. The occurrences of the rainfall generating mechanisms is another counting process $\{\Gamma_t\}$. The cluster origins on the time axis are random and quasi-uniform. (In the time dimension context a distribution is quasi-uniform if the following condition is satisfied: considering two non-overlapping intervals I_1 and I_2 of equal length Δt on the time axis, the probability distribution of the number of cluster centers in I_1 is the same as the probability distribution of the number of cluster centers in I_2 and these two distributions are independent. The probability generating function of the rainfall generating mechanisms will be denoted by $G_\Gamma(z)$.)
3. To each cluster origin there corresponds a storm of rainfalls forming the cluster. The structure of the storm (the cluster) which has its origin at the occurrence time u of the rainfall generating mechanism is determined by the number of rainfalls $\nu(u)$ within that storm and by the time positions T of the rainfalls within that storm.
4. The storm sizes $\nu(u)$ will be mutually independent random variables that may depend on the storm origin u. The sizes $\nu(u)$ will also be independent of all other variables of the stochastic process. Their probability generating function will be denoted by $G_\nu(z|u)$.
5. The rainfalls within a storm are treated as the members of a cluster. Given the time position u of the storm origin, the time positions T of the rainfalls within a storm are independent identically distributed random variables with the conditional probability density function $f_T(\tau-u)$, depending only on the distance $d = \tau-u$ from the storm origin. The time positions T are also independent of all the other variables of the stochastic process. A schematic description of the model is given in Diagram 4.

From the above description of the stochastic model it becomes immediately obvious that the model is a

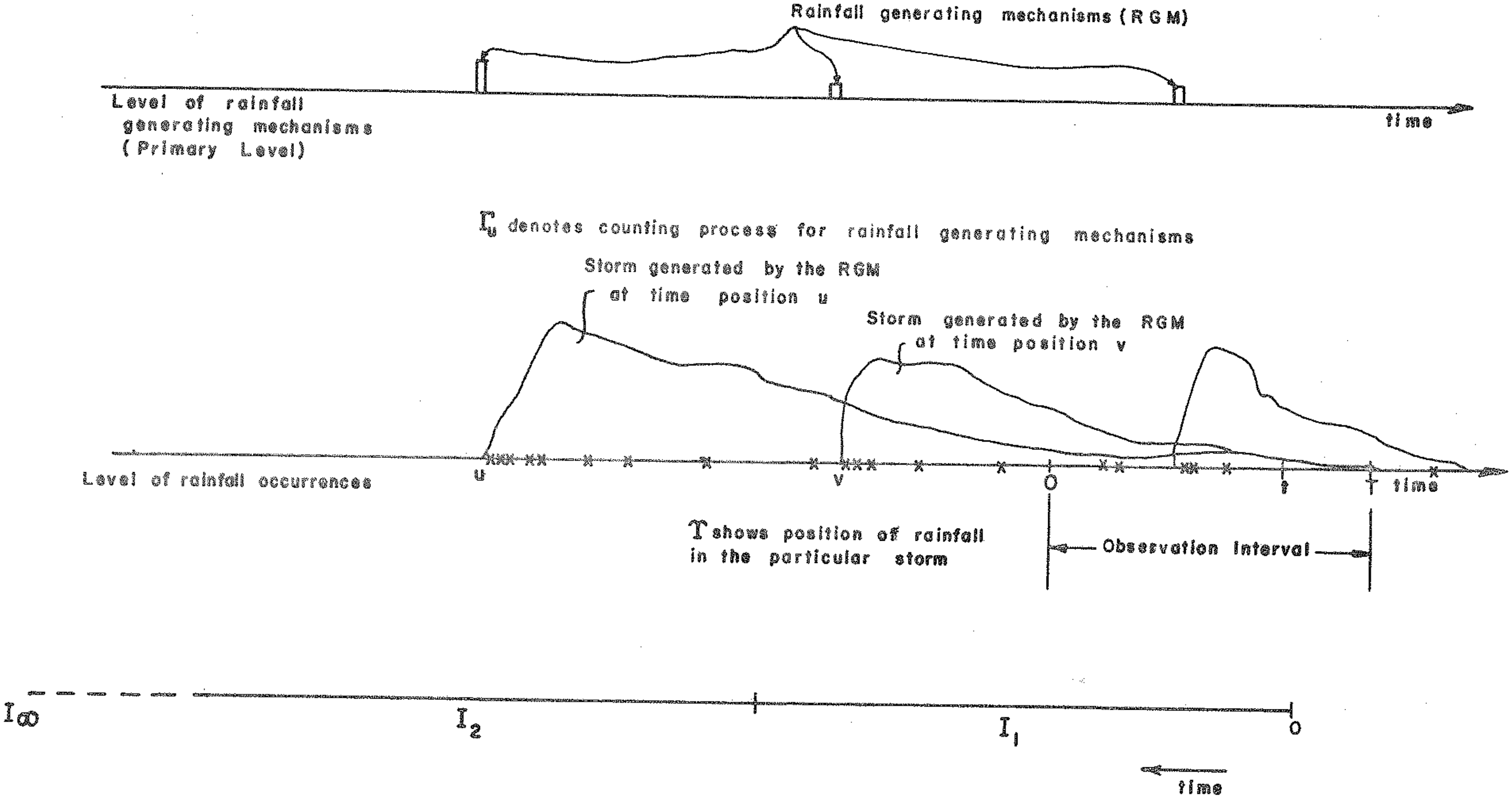

Rainfall generating mechanisms (R G M)
Level of rainfall generating mechanisms (Primary Level)
time
Γ_u denotes counting process for rainfall generating mechanisms
Storm generated by the RGM at time position u
Storm generated by the RGM at time position v
Level of rainfall occurrences
u
v
0
t
time
τ shows position of rainfall in the particular storm
Observation interval
I_∞
I_2
I_1
0
time

Diagram 4. A Schematic Description of the Cluster Model for the Rainfall Occurrences homogeneous one and can be fitted to the homogenized daily rainfall occurrences. It seems possible that the model may be extended to the nonhomogeneous domain (*Moyal,* 1972). However, at the present state of the point stochastic processes almost all the statistical work was done for the homogeneous domain and no established method is available for the fitting of a nonhomogeneous dependence structure by the point stochastic models. Therefore, the verification of the rainfall cluster model will be handled in the homogeneous domain.

6.2 DEVELOPMENT OF THE NEYMAN-SCOTT CLUSTER MODEL

From the above description of the Neyman-Scott cluster model for the homogenized daily rainfall occurrences it follows that the model under consideration is a compound stochastic counting process with (a) a primary or a parent process - the rainfall generating mechanisms which are at the cluster origins, and (b) a secondary or the first generation offspring process - the rainfalls within the particular generated storm. This compound stochastic counting process is constructed in terms of three random variables; (1) the counting random variable Γ for the occurrences of the rainfall generating mechanisms in time, (2) the random variable $\nu(u)$ that counts the number of rainfalls in a storm that has its origin at the time point u, (3) the random variable T, for the time positions of the rainfalls within a storm.

The random variable which interests the hydrologist in the daily rainfall counting process is the counting random variable N_t. Let the interval of observation of the daily rainfall occurrences at a certain station be $(0,T)$. Let N_{t_1} be the number of homogenized daily rainfall occurrences in $(0,t_1)$ and let N_{t_2} be the number of homogenized daily rainfall occurrences in $(0,t_2)$. It is assumed that $0 < t_1 < t_2 < T$ so that the intervals $(0,t_1)$ and $(0,t_2)$ are overlapping. Standing at the time point t_2 and looking at the infinite past, divide $(-\infty, t_2)$ into the time intervals I_J, $J = 1, \ldots, \infty$ of equal length Δt. Δt may be any time length in the homogenized daily rainfall occurrence analysis since the domain of analysis is homogeneous. Each interval I_J may contribute to the number of rainfalls in regions $(0,t_1)$ and $(0,t_2)$ according to the number of rainfall generating mechanisms in I_J and the probabilities $p_1(u) = \int_0^{t_1} f_T(\tau-u)d\tau$ and $p_2(u) = \int_0^{t_2} f_T(\tau-u)d\tau$ that a rainfall whose rainfall generating mechanisms is at the time position u in the time interval I_J, will fall into regions $(0,t_1)$ and $(0,t_2)$ respectively. It is obvious that the nature of the probability density function $f_T(\tau-u)$ is going to determine the memory of the rainfall occurrences. N_{t_1} and N_{t_2} may, therefore ,be represented as the sums of the contributions from each of the non-overlapping regions I_J, $J = 1, \ldots, \infty$. At each region I_J there are γ_J rainfall generating mechanisms so that there will be γ_J storm origins in I_J. Therefore, the contribution from I_J may further be divided into contributions from the rainfall generating mechanisms in I_J. Therefore, N_{t_1} and N_{t_2} may be represented as double summations. Since the cluster origins were assumed to be quasi-uniform the contribution of any rainfall generating mechanism to N_{t_1} or to N_{t_2} is independent of the contribution of any other rainfall generating mechanism. However, the contributions to N_{t_1} and N_{t_2} from the same rainfall generating mechanism in the interval I_J may be dependent. Therefore, N_{t_1} and N_{t_2} may be written as the sum of independent components as

$$N_{t_1} = \sum_{J=1}^{\infty} \sum_{\ell=0}^{\gamma_J} N_{1J\ell} \quad \text{and} \quad N_{t_2} = \sum_{J=1}^{\infty} \sum_{\ell=0}^{\gamma_J} N_{2J\ell}$$

where $N_{1J\ell}$ is the number of rainfalls contributed to $(0,t_1)$ from the ℓ-th rainfall generating mechanism in the interval I_J and $N_{2J\ell}$ is the number of rainfalls contributed to $(0,t_2)$ from the ℓ-th rainfall generating mechanism in I_J. The contribution to $(0,t_1)$ from any interval I_J will be

$$N_{1J} = \sum_{\ell=0}^{\gamma_J} N_{1J\ell}$$

and the expression for N_{2J} is similar. Due to quasi-uniformity the bivariate random variables (N_{1J}, N_{2J}), J = 1, ..., will be mutually independent. Let $G_{N_{t_1},N_{t_2}}(z_1, z_2)$ be the bivariate probability generating function of (N_{t_1}, N_{t_2}), that is

$$G_{N_{t_1},N_{t_2}}(z_1, z_2) = \sum_J \sum_i P[N_{t_1} = i, N_{t_2} = J] z_1^i z_2^J .$$

Due to the independence of (N_{1J}, N_{2J}), $J = 1, \ldots, \infty$, one can write

$$G_{N_{t_1},N_{t_2}}(z_1, z_2) = \prod_{J=1}^{\infty} G_{N_{1J},N_{2J}}(z_1, z_2) \tag{6.1}$$

The probability generating function for the bivariate random variable (N_{1J}, N_{2J}) is

$$\begin{aligned} G_{N_{1J},N_{2J}}(z_1, z_2) &= E\left[z_1^{N_{1J}} z_2^{N_{2J}}\right] \\ &= E\left[E\left(z_1^{N_{1J}} z_2^{N_{2J}} | \gamma_J\right)\right] \\ &= \sum_{r=0}^{\infty} \sum_{n_2=0}^{\infty} \sum_{n_1=0}^{\infty} P[N_{1J} = n_1, N_{2J} = n_2 | \gamma_J = r] z_1^{n_1} z_2^{n_2} P[\gamma_J = r] \\ &= \sum_{r=0}^{\infty} P[\gamma_J = r] \sum_{n_2=0}^{\infty} \sum_{n_1=0}^{\infty} P\left[\sum_{\ell=0}^{r} N_{1J\ell} = n_1, \sum_{\ell=0}^{r} N_{2J\ell} = n_2 | \gamma_J = r\right] z_1^{n_1} z_2^{n_2} . \end{aligned}$$

Since $(N_{1J\ell}, N_{2J\ell})$, $\ell = 1, \ldots, \gamma_J$, are independent, identically distributed as $(N_{1J\ell}, N_{2J\ell})$ and are also independent of γ_J,

$$G_{N_{1J},N_{2J}}(z_1, z_2) = \sum_{r=0}^{\infty} P[\gamma_J = r] \, G^r_{N_{1J\ell},N_{2J\ell}}(z_1, z_2) . \tag{6.2}$$

It can be shown that (*Neyman and Scott*, 1952, 1972) for quasi-uniform rainfall generating mechanisms the probability generation function $G_\Gamma(z)$ is of the form

$$\begin{aligned} G_\Gamma(z) &= \exp\left[-\Delta t\left(h_0 - \sum_{i=1}^{\infty} h_k z^k\right)\right] \\ &= \exp\{-\Delta t \, h(z)\} \end{aligned} \tag{6.3}$$

where $h_0 = \sum_{k=1}^{\infty} h_k < \infty$. Therefore,

$$G_{N_{1J},N_{2J}}(z_1, z_2) = \exp\{-\Delta t \, h(G_{N_{1J\ell},N_{2J\ell}}(z_1, z_2)) \quad . \tag{6.4}$$

It follows from (6.1) and (6.4) that

$$G_{N_{t_1},N_{t_2}}(z_1, z_2) = \exp\{-\Delta t \sum_{J=1}^{\infty} h[G_{N_{1J\ell},N_{2J\ell}}(z_1, z_2)] . \tag{6.5}$$

Now the problem is to express $G_{N_{1J\ell},N_{2J\ell}}(z_1, z_2)$ in terms of the probability generating function of $\nu(u)$, the number of rainfalls generated by a rainfall generating mechanism whose time position is at u. This can be done by the use of the conditional expectations as follows;

$$G_{N_{1J\ell},N_{2J\ell}}(z_1, z_2) = E\left[z_1^{N_{1J\ell}} z_2^{N_{2J\ell}}\right] = E\left[E\left(z_1^{N_{1J\ell}} z_2^{N_{2J\ell}}|U\right)\right] \quad (6.6)$$

where U is the random variable for the position of the storm origin. The indicator random variables X_i and Y_i are introduced to show whether the i-th rainfall in a storm whose origin is at U falls into $(0,t_1)$ and $(0,t_2)$ respectively. Therefore, given the number of rainfalls $\nu(u)$ in this particular storm

$$N_{1J\ell} = \sum_{i=1}^{\nu(u)} X_i \ , \ N_2 = \sum_{i=1}^{\nu(u)} Y_i \quad .$$

Since X_i and Y_i are independent, identically distributed random variables,

$$G_{N_{1J\ell},N_{2J\ell}}(z_1, z_2|U) = G_\nu[G_{X_i,Y_i}(z_1, z_2|U)] \quad (6.7)$$

where $G_{X_i,Y_i}(.,.)$ is the probability generating function of (X_i, Y_i). Opening $G_{X_i,Y_i}(z_1, z_2|U)$,

$$G_{X_i,Y_i}(z_1, z_2|U) = E\left[z_1^{X_i} z_2^{Y_i}|U\right]$$

$$= \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} P[X_i = a, Y_i = b|U] z_1^a z_2^b$$

$$= P[X_i = 0, Y_i = 0|U] + P[X_i = 0, Y_i = 1|U]z_2 + P[X_i = 1, Y_i = 0|U]z_1$$

$$+ P[X_i = 1, Y_i = 1|U]z_1 z_2 \ .$$

In the case under study $(0,t_1)$ is included in $(0,t_2)$. Therefore,

$$P[X_i = 1, Y_i = 1] = \int_0^{t_1} f(\tau-u)d\tau = p_1$$

$$P[X_i = 1, Y_i = 0] = 0$$

$$P[X_i = 0, Y_i = 1] = \int_{t_1}^{t_2} f(\tau-u)d\tau = p_2$$

$$P[X_i = 0, Y_i = 0] = 1 - p_1 - p_2 \ .$$

The $G_{X_i,Y_i}(z_1, z_2|U)$ can be written as

$$G_{X_i,Y_i}(z_1, z_2|U) = 1 - p_1(1 - z_1z_2) - p_2(1 - z_2) \quad (6.8)$$

Then from (6.6), (6.7), (6.8),

$$G_{N_{1J\ell}, N_{2J\ell}}(z_1, z_2) = E(G_\nu[1 - p_1(1 - z_1z_2) - p_2(1 - z_2)]) \ . \quad (6.9)$$

The probability density of the position u of the storm origin in any interval I_J is $1/\Delta t$ due to quasi-uniformity of the storm origins (cluster origins, or the time positions of the rainfall generating mechanisms). Then, from (6.9),

$$G_{N_{1J\ell},N_{2J\ell}}(z_1, z_2) = \int_{I_J} \frac{1}{\Delta t} G_\nu[1 - p_1(1 - z_1z_2) - p_2(1 - z_2)]du \ . \quad (6.10)$$

Substituting (6.10) into (6.5)

$$G_{N_{t_1},N_{t_2}}(z_1, z_2) = \exp\left\{-\Delta t \sum_{J=1}^{\infty} h\left(\frac{1}{\Delta t}\int_{I_J} G_\nu[1 - p_1(1 - z_1z_2) - p_2(1-z_2)]du\right)\right\} \qquad (6.11)$$

It can be shown that (*Neyman and Scott,* 1958) (11) obtains the form

$$G_{N_{t_1},N_{t_2}}(z_1, z_2) = \exp\left\{\int_{-\infty}^{T} h(G_\nu[1 - p_1(1 - z_1z_2) - p_2(1 - z_2)])du\right\} \qquad (6.12)$$

Expression (6.12) is the general probability generating function for the homogenized daily rainfall counts N_{t_1} and N_{t_2} in the intervals $(0,t_1)$ and $(0,t_2)$. One can then make assumptions about the particular form of the process of rainfall generating mechanisms, the distribution of $\nu(u)$, and the distribution of τ. Assume that the counting process of the rainfall generating mechanisms obey the Poisson law. Then, from (6.3),

$$h(z) = - h_0 + h_1 z = -h_0(1-z)$$

and (6.12) becomes)

$$G_{N_{t_1},N_{t_2}}(z_1, z_2) = \exp\left\{-h_0 \int_{-\infty}^{T} (1 - G_\nu[1 - p_1(1 - z_1z_2) - p_2(1 - z_2)])du\right\} \qquad (6.13)$$

where h_0 becomes the rate of occurrence of the rainfall generating mechanisms in time. The random variable that is of practical concern to the hydrologist is not (N_{t_1}, N_{t_2}) but just N_{t_1} or just N_{t_2}, the number of rainfall occurrences in $(0,t_1)$ and $(0,t_2)$ respectively. The probability generating function of the number of occurrences N_{t_1} in the time interval $(0, t_1)$ can be obtained by using arguments similar to above. The pgf of N_{t_1} can be shown to be

$$G_{N_{t_1}}(z_1) = \exp\left\{-h_0 \int_{-\infty}^{T} (1 - G_\nu[1 - (1-z_1)\ p_1(u)])du\right\} \qquad (6.14)$$

where $p_1(u) = \int_0^{t_1} f(\tau-u)d\tau$ is the probability that a rainfall from a storm that has its origin at u will occur in $(0,t_1)$. In (6.14) it is again assumed that the rainfall generating mechanisms obey the Poisson law.

6.3 APPLICATION OF THE NEYMAN-SCOTT CLUSTER PROCESS IN HYDROLOGY

The Neyman-Scott cluster process for the rainfall occurrences emerges as a general point stochastic model for the rainfall point processes that envelopes the models with independent counting increments as its special cases. The generalized Poisson process (the compound Poisson process) and the simple Poisson process that were fitted to the rainfall occurrences in different climatic regions of the world by various hydrologists will be shown to be the special cases of the Neyman-Scott cluster model for the rainfall occurrences in the form of (6.14). The concept that will yield the special cases is the memory of the rainfall process which is manifested by the memory of the rainfall generating mechanisms. This memory is expressed quantitatively by $p_1(u)$, the probability that a rainfall from a storm that has its origin at u, the occurrence time of the particular rainfall generating mechanism that has caused the storm, will occur in $(0,t_1)$. As the time position u is placed further and further to the past from the interval $(0,t_1)$ the probability for a rainfall, generated by a mechanism at u, to occur in $(0,t_1)$ approaches zero. If the rainfall generatin mechanism had an infinite memory, there would always be a positive probability that a rainfall, generated by this mechanism, would occur in $(0,t_1)$ even if the storm origin u was at the infinite past. Of course, this is a highly hypothetical situation.

In the case of processes with independent increments, the memory of the rainfall process is zero. The probability density function $f_T(\tau-u)$ becomes a delta function that picks up the value at $\tau=u$, so that

$$p_1(u) = \int_0^{t_1} f(\tau-u)d\tau = 1 \text{ if } u \in (0,t_1)$$
$$= 0 \text{ otherwise.}$$

Then (6.14) will become

$$G_{N_{t_1}}(z_1) = \exp\left\{-h_0 \int_0^{t_1} (1 - G_\nu[1 - (1-z_1)]du\right\}$$

$$= \exp\{-h_0\ t_1[1 - G\ (z_1)]\}\ . \qquad (6.15)$$

However, (6.15) is just the pgf of the generalized Poisson process or the compound Poisson process (*Parzen,* 1967, or *Feller,* 1968). A generalized Poisson process may be interpreted as an integer-valued counting process with stationary independent increments where there exists a probability that a random number ν of events may occur simultaneously at a given time, given that at least one has occurred. The random number ν of events may be $\nu = 1, 2, \ldots, \infty$. In a generalized Poisson process a group of events occur at the instants of occurrence of an event in the primary process which is a simple Poisson process. In the rainfall counting process under the generalized Poisson model the rainfall generating mechanisms will occur according to simple Poisson law and at the moment a mechanism occurs it will instantaneously generate a number ν of rainfalls at that moment. Consequently, the memory of the rainfall process will be zero under the generalized Poisson or the compound Poisson model. The generalized Poisson process may be used to model the thunderstorm rainfall activity in arid regions. (*Duckstein, et al.*, 1972).

Finally, when the number of rainfalls ν, generated at the instant of occurrence of the rainfall generating mechanism, is equal to unity, then $P(\nu=1) = 1$ and

$$G_\nu(z) = \sum_{J=0}^{\infty} P[\nu=J]z^J = z\ ,$$

and from (6.15)

$$G_{N_{t_1}}(z_1) = \exp\{-h_0\ t(1-z_1)\}\ . \qquad (6.16)$$

Expression (6.15) is the pgf of the simple Poisson process with the rate of occurrence h_0. Thus, the two-level process is reduced to a one-level simple Poisson process when the rainfall generating mechanisms obey the simple Poisson law and generate just one rainfall at the instant of their occurrence. The Poisson model was employed by *Lobert* (1967) in modeling the rainfall occurrence for regions in France and by *Todorovic and Yevjevich* (1969) in modeling the rainfall occurrence for the stations in Colorado, Texas and Iowa in U.S.A.

The Neyman-Scott cluster model in the point stochastic processes may play a role analogous to that of the ARMA family in the time series analysis. In both types of analysis the models are classified according to the existing persistence in the analyzed natural process. Lack of dependence among the time series values would eventually reduce the ARMA family to the white noise process while the lack of dependence in the counts would reduce the cluster model to the generalized Poisson process.

6.4 STATISTICS FOR THE NEYMAN-SCOTT CLUSTER MODEL

It is possible to derive all the moments of a point stochastic process from its probability generating function. In this section various important statistical functions will be obtained for the Neyman-Scott cluster model through its probability generating function.

The expectation of the number of rainfalls in $(0,t_1)$ is obtained as

$$E(N_{t_1}) = dG_{N_{t_1}}(z_1)/dz_1|_{z_1 = 1} \; .$$

Using (6.14) as the pgf of N_{t_1},

$$E(N_{t_1}) = \exp\{-h_0 \int_{-\infty}^{T} (1 - G_\nu[1 - (1-z_1)\, p_1(u)])du\} \cdot h_0 \int_{-\infty}^{T} \frac{dG_\nu\{\cdot\}}{dz_1} du|_{z_1=1} \; .$$

Taking

$$G_\nu\{\cdot\} = \sum_{J=0}^{\infty} p_J\{1 - (1-z_1)\, p_1(u)\}^J$$

where $p_J = P[\nu=J]$, so that

$$dG_\nu(\cdot)/dz_1|_{z_1=1} = \sum_J J p_J\{1 - (1-z_1)\, p_1(u)\}^{J-1}\, p_1(u)|_{z_1=1} = p_1(u)\, E(\nu) \; ,$$

$E(N_{t_1})$ is obtained as

$$E(N_{t_1}) = h_0\, E(\nu) \int_0^{t_1} \int_{-\infty}^{\tau} f(\tau-u)dud\tau = h_0\, t_1\, E(\nu) \; . \qquad (6.17)$$

Expression (6.17) is the mean-time function of the cluster model (6.14). The rate of rainfall occurrence under the Neyman-Scott model becomes

$$\lambda(t_1) = dE(N_{t_1})/dt_1 = h_0\, E(\nu) \; . \qquad (6.18)$$

Therefore, the rate of rainfall occurrence under the Neyman-Scott model is the product of the rate of rainfall generating mechanism occurrences and the expected number of rainfalls generated by each mechanism.

The variance-time function for the rainfall counts may again be derived from (6.14). In terms of the derivatives of the pgf of N_{t_1} evaluated at unity, the variance of the number of rainfall counts in $(0,t_1)$ becomes

$$\mathrm{Var}(N_{t_1}) = G'_{N_{t_1}}(z_1) + G''_{N_{t_1}}(z_1) - [G'_{N_{t_1}}(z_1)]^2|_{z_1=1}$$

and after some manipulations

$$\mathrm{Var}(N_{t_1}) = h_0\, t_1\, E(\nu) + h_0\, E(\nu^2 - \nu) \int_{-\infty}^{T} p_1^2(u)du \qquad (6.19)$$

where $\int_{-\infty}^{T} p_1^2(u)du = 2 \int_0^{t_1} \int_0^x \int_{-\infty}^y f_T(x-u)\, f_T(y-u)du\, dy\, dx$. As is seen from (6.19), the variance-time function does not depend on the explicit distribution of ν, the number of rainfalls in a storm, but only on its first two moments. Only the pdf $f_T(\tau-u)$ of the rainfall positions with respect to the storm origin has to be defined in order to obtain explicit expressions for the variance-time function of the rainfall counts under the Neyman-Scott cluster model.

The covariance function of the rainfall counts in two non-overlapping intervals τ time units apart may be needed to test the independence in the counting increments. For this purpose the bivariate pgf of (N_{t_1}, N_{t_2}) is used. However, $G_{N_{t_1},N_{t_2}}(z_1, z_2)$ in (6.13) was for the overlapping intervals $(0,t_1)$ and $(0,t_2)$. If two non-overlapping intervals $(0,t_1)$ and $(t_1+\tau,\ t_1+\tau+t_2)$ of lengths t_1 and t_2 are considered, the pgf of (N_{t_1}, N_{t_2}) takes the form

$$G_{N_{t_1}, N_{t_2}}(z_1, z_2) = \exp\{-h_0 \int_{-\infty}^{T} (1 - G_\nu[1 - (1-z_1)\, p_1(u) - (1-z_2)\, p_2(u)])du\} \qquad (6.20)$$

where $p_1(u) = \int_0^{t_1} f_T(x-u)dx$ and $p_2(u) = \int_{t_1+\tau}^{t_1+\tau+t_2} f_T(x-u)dx$. Then the covariance function $\mathrm{Cov}[N_{t_1}, N_{t_2}; \tau]$ for lag τ is given as

$$\mathrm{Cov}[N_{t_1}, N_{t_2}; \tau] = \left.\frac{\partial^2 G_{N_{t_1}, N_{t_2}}(z_1, z_2)}{\partial z_1\, \partial z_2}\right|_{\substack{z_1=1\\ z_2=1}} - \left[\left.\frac{\partial G_{N_{t_1}, N_{t_2}}(z_1, z_2)}{\partial z_1}\right|_{\substack{z_1=1\\ z_2=1}} \cdot \left.\frac{\partial G_{N_{t_1}, N_{t_2}}(z_1, z_2)}{\partial z_2}\right|_{\substack{z_1=1\\ z_2=1}}\right] .$$

After the manipulations, the covariance function for the rainfall counts in two non-overlapping intervals τ time units apart is obtained as

$$\mathrm{Cov}[N_{t_1}, N_{t_2}; \tau] = h_0\, E[\nu^2 - \nu] \int_{-\infty}^{T} p_1(u)\, p_2(u)du \ . \qquad (6.21)$$

As is seen from (6.21) as long as $p_1(u)$ and $p_2(u)$ are different from zero, there is a correlation among the non-overlapping counting intervals in a Neyman-Scott cluster model. Therefore, the dependence structure in the rainfall counts can be explained by this model.

The spectrum of counts for the Neyman-Scott cluster model may be obtained by taking the Fourier transform of one-half the second derivative of the variance-time function. It is given as (*Vere-Jones*, 1970),

$$g_+(\omega) = \frac{1}{\pi} [h_0\, E(\nu) + h_0\, E(\nu^2 - \nu) \left| \int_{-\infty}^{\tau} f_T(\tau-u)\, e^{i\omega(\tau-u)} du \right|^2] \ , \ \omega > 0 \ . \qquad (6.22)$$

Expressions (6.17), (6.18), (6.19), (6.21) and (6.22) are general expressions for the statistical functions studied. Only the primary process of the rainfall generating mechanisms was specified to be Poisson. Furthermore, the above expressions show that the statistical functions under study need only that the distribution of the rainfall positions with respect to their storm origin be specified. Since the variance-time function, the covariance function and the spectrum of counts measure the dependence in the rainfall counting increments, the measurement of the persistence in terms of $f(\tau-u)$ becomes clear. The specific form of $f(\tau-u)$ is going to decide the memory of the cluster model and the structure of the storm. One only needs to know the first two moments of ν, the number of rainfalls in a storm, to completely specify the dependence structure of the process.

In the following, the distribution of the rainfalls from the storm origin is assumed to be negative exponential. That is

$$f_T(\tau-u) = \theta e^{-\theta(\tau-u)} \ , \ \tau > u$$
$$= 0 \ , \ \text{otherwise.}$$

Once this assumption is made the following expressions for the variance-time function and the spectrum of counts are obtained from (6.19) and (6.22),

$$\mathrm{Var}(N_t) = h_0\, E(\nu^2)t + \frac{h_0\, E(\nu^2 - \nu)}{\theta} e^{-\theta t} - \frac{h_0\, E(\nu^2 - \nu)}{\theta} \qquad (6.23)$$

and

$$g_+(\omega) = \frac{1}{\pi} \left[h_0\, E(\nu) + h_0\, E(\nu^2 - \nu) \cdot \frac{\theta^2}{\theta^2 + \theta^2}\right] , \ \omega > 0 \ . \qquad (6.24)$$

The rate of rainfall occurrence and the variance-time function for the rainfall counts under the compound

Poisson process were given earlier. The spectrum of the rainfall counts under the compound Poisson model (which corresponds to thunderstorm activity) follows directly from (6.22) as

$$g_+(\omega) = \frac{1}{\pi} h_0 \, E(\nu^2) \ . \qquad (6.25)$$

Therefore, the spectrum of counts for the compound Poisson model is a horizontal line as in the case of simple Poisson model. This behavior is in analogy with the spectral density function of a white noise time series. In the counts spectrum a horizontal line indicates uncorrelated counting increments whereas a horizontal spectral density function indicates uncorrelated time series.

6.5 CALIBRATION OF THE NEYMAN-SCOTT CLUSTER MODEL

In order to test the goodness of fit of the Neyman-Scott cluster model for the daily rainfall counting process, the model was fitted to the counts spectrum and the log-survivor function of the homogenized daily rainfall occurrences. The fit to the rainfall counts spectrum was performed to see how well the cluster model can preserve the dependence structure of the daily rainfall counts. The fit to the log-survivor function was performed to see how well the cluster model can preserve the empirical probability distributions related to the daily rainfall occurrences. The fit to the log-survivor function of the homogenized daily rainfall counts by the cluster model will be discussed in the next section which deals with the practical probabilities of the dry and wet sequences.

Analyzing expressions (6.18), (6.23) and (6.24), if it is assumed that the only nonhomogeneity in the daily rainfall counts process is the nonhomogeneity of the rainfall generating mechanisms, then the homogenization scheme (3.6) is approximately valid for the second order moments of the counts process because the rate of occurrence, the counts spectrum and the variance-time function are all linear functions of h_0, the rate of rainfall generating mechanism occurrence in the homogeneous domain. However, as it can be seen from (6.23), $Var(N_t)$ is a nonlinear function of time t. Unless θ is large, a linear homogenization scheme such as (3.6) will not homogenize the second order moments of the process. Therefore, the homogenization scheme developed under the nonhomogeneous Poisson hypothesis is only approximate under the Neyman-Scott cluster model for the daily rainfall counts. The explicit homogenization scheme under the cluster model hypothesis can only be derived from the nonhomogeneous form of the Neyman-Scott cluster model which is not available. Further research will deal with the construction of the nonhomogeneous form of the Neyman-Scott model. Nevertheless, as was discussed in the homogenization results, the scheme (3.6) effectively removed the time trends in the rainfall occurrence data. Therefore, based on the empirical results obtained from the homogenized data, the theoretical counts spectrum and the theoretical log-survivor function for the Neyman-Scott cluster model were fitted to the homogenized data.

In order to preserve the correlation structure of the homogenized daily rainfall counts process the stochastic model for the process has to be fitted to the counts spectrum since the counts spectrum is the Fourier transform of the covariance density of the counts process. The spectrum of the homogenized daily rainfall counts behaved very consistently in the 17 cases analyzed. Even in the cases of stations 3082 and 4642 where the yearly cyclicity was over-removed the counts spectrum was affected only at the very narrow region near the origin. The general behavior of the higher frequencies, corresponding to the short term dependencies, remained very consistent through all the cases. Therefore, the logical way for the calibration of the model parameters for the preservation of the correlation structure of the homogenized daily rainfall counts is to fit the expression (6.24) of the Neyman-Scott cluster model counts spectrum to the counts spectrum of the homogenized daily rainfall counts data. However, as is seen from the table 6-1, the parameter values calibrated from the spectral fit are unrealistic in that the pairs of estimates for $E(\nu^2)$ and $E(\nu)$ for each case yield negative variance. However, this is basically due to the fact that in the nonlinear regression scheme employed for the spectral fits no constraint was put on $E(\nu^2)$ and $E(\nu)$. In table

6-1 the ratio of $E(\nu^2)$ to $E(\nu)$ is consistently around 2. The normalized spectrum to which the cluster model is fitted can be obtained from (6.24) as

$$\frac{\pi g_+(\omega)}{h_0\, E(\nu)} = 1 + \left[\frac{E(\nu^2)}{E(\nu)} - 1\right] \frac{\theta^2}{\theta^2 + \theta^2} \qquad \omega > 0 \; .$$

Using this expression the ratio $E(\nu^2)/E(\nu)$ was set equal to 2 and the cluster model was again fitted to the homogenized daily rainfall counts spectra for the stations 0132, 3082, 3777, 4642, 6056 and 7747. The results are given in the figs. 6.1 through 6.6 and in the table 6-2. As is seen from the figures, the cluster model satisfactorily fits the empirical spectra of the rainfall counts. $\hat{\theta}$ values calibrated from the counts spectra fits are between 3 and 4. A value between 3 and 4 shows a rapid decay of the pdf $f_T(\tau\text{-}u)$. Therefore, the memory of the homogenized daily rainfall counts process is very short. The $\hat{\theta}$ values for the stations of 0177, 0545, 1747, 1882, 3547, 4908, 6164, 6338, 7069, 7755 and 7925 are given in table 6-2A.

It was established earlier that $\qquad g_+(0^+) = \frac{1}{\pi} V'(t)|_{t=\infty}$

so that the origin value of the counts spectrum is equal to the asymptotic slope of the variance-time function. Therefore, the asymptotic behavior of the variance-time function of the homogenized daily rainfall counts would determine the long range dependence structure of the rainfall counts process. However, due to this relation between the variance-time function and the counts spectrum, the over-removal of the low frequencies drastically affects the behavior of the variance-time function. The diversion of the variance-time behavior in the stations 3082 and 4642 from the general behavior as exemplified in the fig. 9 by the station 0132, is due to the over-removal of the yearly periodicity in the daily rainfall counts. Therefore, the variance-time curves for different stations were quite inconsistent as was seen in the fig. 9. In order to observe the asymptotic behavior of the variance-time function a long record is needed. In this study the record length was 10 years and was not adequate for the observation of the asymptotic behavior. Therefore, the Neyman-Scott model was not fitted to the variance-time function of the homogenized daily rainfall counts data.

The distribution function for $\nu(u)$, the number of rainfalls in a storm, is not discussed in this section since it was not essential for the correlation structure of the point process. It will be discussed in the next section dealing with the practical applications of the cluster model to hydrologic problems. In the next section various practical probabilities will be derived and the pdf of $\nu(u)$ will be determined.

TABLE 6-1

CLUSTER MODEL PARAMETERS CALIBRATED FROM THE SPECTRUM OF THE HOMOGENIZED DAILY RAINFALL COUNTS

Station	$\hat{E}(\nu)$	$\hat{E}(\nu^2)$	$\hat{\theta}$
0132	24.1	48.48	3.37
3082	24.2	48.13	3.44
4642	24.2	48.16	3.60
6056	22.44	52.82	2.72
7747	23.95	48.83	2.77

TABLE 6-2

CALIBRATION OF THE CLUSTER MODEL PARAMETER θ WHEN $E(\nu^2)/E(\nu)$ IS CONSTRAINED TO BE 2

Station	$\hat{\theta}$	R^2
0132	3.410	.5817
3082	3.398	.5211
3777	3.709	.5427
4642	3.670	.5103
6056	4.003	.5168
7747	3.337	.5754

TABLE 6-2A

CALIBRATION OF THE CLUSTER MODEL PARAMETER θ
WHEN $E(\nu^2)/E(\nu)$ IS CONSTRAINED TO BE 2

Station	$\hat{\theta}$	R^2
0177	3.276	.5198
0545	3.275	.5314
1747	3.523	.5897
1882	6.840	.4403
3547	5.122	.6333
4908	3.995	.5110
6164	3.405	.5206
6338	3.513	.5368
7069	3.322	.3836
7755	3.677	.5229
7935	3.879	.5613

CHAPTER 7 - PROBABILITIES RELATED TO DRY AND WET SEQUENCES UNDER THE NEYMAN-SCOTT CLUSTER MODEL FOR THE STATIONARY DAILY RAINFALL OCCURRENCES

7.1 SEASONAL NEYMAN-SCOTT CLUSTER MODEL

In chapter 6 it was stated that the present form of the Neyman-Scott cluster model is homogeneous. In the time domain this statement means that the point stochastic model is stationary. Therefore, the present form of the Neyman-Scott cluster model can be of practical use only if the daily rainfall occurrences can be considered stationary. For a time length of one season the model can be used effectively for the computation of wet and dry sequence probabilities. When a time interval of one season is considered, then within the year fluctuations of the daily rainfall occurrences can probably be ignored. The long-term time trend effects can also be ignored for such time length as three months. Therefore, one can calibrate the Neyman-Scott cluster model for four different seasons, each season having different parameters.

An important point to consider for the separation of the time interval into seasons is the memory of the daily rainfall counts process. The memory of the process should be very short so that the storm generating mechanisms that have occurred, say in the spring, do not cause rainfalls in the summer months. As can be seen from the results of the spectral fit of the cluster model to the rainfall data, the parameter θ of the pdf $f_T(\tau-u)$ is between 3 and 4 so that the memory of the daily rainfall occurrences is quite short. Therefore, the storm generating mechanisms that have occurred prior to the seasonal time interval under consideration have a negligible effect on the occurrence of rainfalls during that season. Consequently, it is possible to divide the time into seasonal intervals and consider the Neyman-Scott cluster model for each season separately.

For any season the random variables of practical hydrologic concern for the design and the operation of water resources from the point of view of the daily rainfall occurrences would be (1) the time length of a dry period from an arbitrary time origin to the occurrence of the first rainfall event, (2) the time length of a dry period from a rainfall occurrence to the next rainfall occurrence. (3) the return period of the rainfalls, (4) the duration of a wet period, and (5) the number of rainfalls in an interval (0,t). The probabilities related to the above random variables will be derived through the pgf of the Neyman-Scott cluster model for the stationary daily rainfall counts process.

7.2 THE TIME LENGTH OF A DRY PERIOD FROM AN ARBITRARY ORIGIN TO THE FIRST DAILY RAINFALL OCCURRENCE

This time length is known as the "forward recurrence time" in the statistical literature (*Cox,* 1967), and it will be denoted by the letter T. The probability of T exceeding t days is of practical hydrologic concern since it will inform the hydrologist about the risk of a drought whose length exceeds t days when the hydrologist looks at the future from an arbitrary time origin onwards.

The probability $P(T > t)$ is equivalent to $P[N_t = 0]$, that is, the probability of no rainfall occurrence in the time interval (0,t) when the time origin is set at 0. The probability $P[N_t = 0]$ can be expressed in terms of the pgf of N_t as

$$P[N_t = 0] = G_{N_t}(0) = \exp\{-h_0 \int_{-\infty}^{T} (1 - G_\nu[1 - p(u)])du\} \; . \qquad (7.1)$$

where $p(u)$ is the probability of a rainfall whose origin is at u and h_0 is the rate of generating mechanism occurrence. In order to obtain an explicit expression for (7.1), the distribution of the number of rainfalls in a storm, ν, has to be assumed. Basically due to its simplicity, the geometric distribution will be assumed to be the law of ν. The probability generating function of the geometric distribution is (*Dwass,* 1969),

$$G(z) = \frac{pz}{1 - (1-p)z} \, , \; |z| < 1 \tag{7.2}$$

so that there is only one parameter, p. Due to the change in the behavior of p(u) in the regions $(-\infty, 0)$ and $(0,t)$ the pgf of N_t is expressed as

$$G_{N_t}(0) = \exp -h_0 \int_{-\infty}^{0} (1 - G_\nu\{1 - e^{\theta u}(1 - e^{-\theta t})\})du + \int_0^t (1 - G_\nu[e^{-\theta(t-u)}])du \tag{7.3}$$

where
$$G_\nu[1 - e^{\theta u}(1 - e^{-\theta t})] = \frac{p[1 - e^{\theta u}(1 - e^{-\theta t})]}{1 - (1-p)\,[1 - e^{\theta u}(1 - e^{-\theta t})]} \, ,$$

and
$$G_\nu[e^{-\theta(t-u)}] = \frac{pe^{-\theta(t-u)}}{1 - (1-p)\, e^{-\theta(t-u)}} \, .$$

Doing the integrations in (7.3), $G_{N_t}(0)$ is obtained as,

$$G_{N_t}(0) = e^{-h_0 t} \left\{\frac{1 - (1-p)\, e^{-\theta t}}{p}\right\}^{-h_0/\theta}$$

so that
$$P[T > t] = P[N_t = 0] = e^{-h_0 t} \left\{\frac{1 - (1-p)\, e^{-\theta t}}{p}\right\}^{-h_0/\theta} \, , \; t > 0 \, . \tag{7.4}$$

Expression (7.4) informs the hydrologist about the risk of a dry period whose length may exceed t days when the hydrologist investigates the future from the present (which corresponds to the time origin 0). From the plot of dry period length t versus P[T > t], using the eq. (7.4), one can obtain the dry period lengths corresponding to different risk levels.

7.3 PROBABILITY FOR THE LENGTH OF A DRY PERIOD BETWEEN TWO RAINFALL OCCURRENCES

This time length is known as the "interarrival time" in the statistical literature (*Parzen,* 1967), and it will be denoted by the letter X. The probability of X exceeding t days will inform the hydrologist about the risk of a drought whose time length exceeds t days when the hydrologist looks to the future from a rainy day onwards. It can be shown that (*Khinchin,* 1955)

$$P[X > t] = -\frac{1}{h_0\, E(\nu)} \frac{d}{dt} G_{N_t}(0) \tag{7.5}$$

when N_t, the number of rainy days in (0,t), is modeled by the Neyman-Scott cluster process. The derivative of $G_{N_t}(0)$ can be expressed by

$$\frac{dG_{N_t}(0)}{dt} = G_{N_t}(0) \cdot \frac{d \ln G_{N_t}(0)}{dt} \, . \tag{7.6}$$

The probability, p(u), of a rainfall whose origin is at u, to occur in (0,t) is

$$p(u) = \int_0^t e^{-\theta(\tau-u)}d\tau \, , \text{ for } u \leq 0$$

and
$$p(u) = \int_u^t \theta e^{-\theta(\tau-u)}d\tau \, , \text{ for } u > 0 \, . \tag{7.7}$$

From (7.3) it follows that

$$\frac{d \ln G_{N_t}(0)}{dt} = h_0 \int_{-\infty}^{0} \frac{\partial G_\nu[1 - p(u)]}{\partial t}\, du + h_0 \int_0^t \frac{\partial G_\nu[1 - p(u)]}{\partial t}\, du + [1 - G_\nu(1 - p(t))] \qquad (7.8)$$

where $1 - G\ (1 - p(t)) = 0$. After some integrations and manipulations on (7.8) the derivative of $\ln G_{N_t}(0)$ is obtained as,

$$\frac{d \ln G_{N_t}(0)}{dt} = h_0 \frac{G_\nu(e^{-\theta t}) - 1}{1 - e^{-\theta t}} . \qquad (7.9)$$

The expression for $G_\nu(e^{-\theta t})$ under the geometric law for ν, is

$$G_\nu(e^{-\theta t}) = \frac{pe^{-\theta t}}{1 - (1-p)\, e^{-\theta t}} . \qquad (7.10)$$

Combining (7.10), (7.9), (7.6), (7.5) and (7.4), $P[X > t]$ is obtained as

$$P[X > t] = e^{-h_0 t}\left[\frac{p}{1 - (1-p)\, e^{-\theta t}}\right]^{h_0/\theta + 1} , \; t > 0 . \qquad (7.11)$$

Considering the complexity of the cluster model with respect to the simple Poisson model it is quite satisfying to obtain simple probability expressions (7.11) and (7.4) for the dry periods.

There are only 3 parameters to be calibrated for the modeling of the daily rainfall counting process under the Neyman-Scott cluster model. These three parameters, h_0, p, and θ, govern not only the nature of the dry periods but also the nature of the wet periods as will be seen later. Given a certain risk level, the corresponding dry period length between two rainy days can be obtained from a plot of the dry period length t versus $P[X > t]$.

7.4 THE RETURN PERIOD

The return period, in days, of the rainfall events is the statistic that is most often used in the hydrologic problems dealing with the dry periods. Especially, in the arid regions of the earth, the designer wants to know the expected dry period in the region with which he is concerned. In the rainfall occurrence process the return period is the expected time length between two consecutive rainfall events. If the return period of the rainfall occurrences at a certain region is denoted by T_r, then

$$T_r = E(X) \qquad (7.12)$$

where X is the time between two consecutive rainfall occurrences. $E(X)$ can be expressed as

$$E(X) = \int_0^\infty P[X > t]dt . \qquad (7.13)$$

From the eq. (7.5),

$$E(X) = - \frac{1}{h_0\, E(\nu)} \int_0^\infty \left\{\frac{d}{dt} G_{N_t}(0)\right\} dt ,$$

$$= - \frac{1}{h_0\, E(\nu)} [P[T > t]]_{t=0}^{\infty}$$

where T is the time length from an arbitrary origin to the first rainfall occurrence. Then

$$E(X) = 1/[h_0\, E(\nu)] ,$$

and the return period T_r under the Neyman-Scott cluster model for the daily rainfall occurrences, is obtained as

$$T_r = \frac{1}{h_0 \, E(\nu)} \ . \tag{7.14}$$

Therefore, the return period is just the reciprocal of the rate of rainfall occurrence under the Neyman-Scott cluster model. In fact it follows from (7.5) that for all stationary point stochastic processes the return period of an event of certain magnitude is equal to the reciprocal of the rate of occurrence of that event.

7.5 LOG-SURVIVOR FUNCTION OF THE DAILY RAINFALL OCCURRENCE UNDER THE NEYMAN-SCOTT CLUSTER MODEL

The log-survivor function is the logarithm of $P[X > x]$. Therefore, the probability distribution of the rainfall interarrival times can be fitted by the use of the log-survivor function. Since the Poisson model has negative-exponentially distributed interarrival times with some parameter α, the theoretical log-survivor function for the Poisson model is

$$\ln P[X > t] = -\alpha t \quad , \quad t > 0$$

which is a straight line with the negative slope $-\alpha$. Thus, the log-survivor function has the distinct advantage over the probability distribution function that the deviation of the Poisson model from the empirical data for the interarrival probabilities can be detected very clearly through the deviation of the data from the linearity in the logarithmic domain.

In chapter 3 on the homogenization results it was observed that the log-survivor functions were convex, indicating to a clustering of the rainfall occurrences. A good way to compare the capabilities of the simple Poisson and the cluster models for their preservation of the probability distribution of the rainfall interarrival times is to fit these two models to the empirical log-survivor function obtained from the homogenized daily rainfall occurrence data.

The log-survivor function for the Neyman-Scott cluster model is obtained from the expression (7.11) as

$$\ln P[X > t] = -h_0 t + \left(\frac{h_0}{\theta} + 1\right) \ln \left[\frac{p}{1 - (1-p)\, e^{-\theta t}}\right] \ , \ t > 0 \tag{7.15}$$

Using the $\hat{\theta}$ values estimated from the rainfall counts spectra fits and taking h_0 and p as the free parameters, the theoretical log-survivor function of the Neyman-Scott model was fitted to the log survivor function of the homogenized daily rainfall counts in the stations 0132, 0177, 0545, 3777, 4642, 6056 and 7747. The regression results are given in the table 7-1. The log-survivor functions for the Poisson model, the Neyman-Scott cluster model and the homogenized daily rainfall occurrence data were plotted together for the stations 0132, 0545, 3777, 4642, 6056 and 7747 as shown in the fig. 17. As is seen from the figure, the cluster model can fit the log-survivor function of the rainfall data better than the Poisson model. The difference is especially conspicuous on the longer interarrival times. For station 4642 the case of the "underdispersion" of the rainfall counts with respect to the Poisson model is shown. As was discussed earlier, the underdispersion that is noticed in the stations 3082 and 4642 is due to the over-removal of the long periodicities in the data. The underdispersion is very unrealistic since it leads to a process more regular than the Poisson and can be modeled neither by the simple Poisson nor by the cluster model. The presence of the underdispersed cases, although they are few, points to the need of a more rigorous homogenization scheme that should be based on the nonhomogeneous form of the Neyman-Scott cluster model. Similar results were obtained for the stations 1474, 1882, 3082, 3547, 4908, 6164, 6338, 7069, 7755 and 7935. The results are given in the table 7-1A and in the figures 17A and 17B.

From the results of the log-survivor function fits, the Neyman-Scott cluster model emerges as a good model for the preservation of the probability law of the stationary daily rainfall occurrence interarrival times. Once the cluster model parameters are calibrated through the spectral and the log-survivor function fits the practical probabilities of the dry periods can be obtained from the expressions (7.11) and (7.4).

7.6 DURATION OF A WET PERIOD

In the observation of a hydrologic stochastic process the continuous time axis is divided into equal time units of length Δt. In the case under study Δt was chosen to be one day. A wet period of length $J\Delta t$ will be defined as the period of J consecutive rainy intervals of time length Δt followed by a dry interval of length Δt. The probability of a wet period of length $J\Delta t$ from an arbitrary time origin 0 is

$$P[\text{Wet period duration} = J\Delta t] = P[N_{J\Delta t} = J \cap N_{(J+1)\Delta t} = 0] \tag{7.16}$$

provided that the interval size Δt is small enough so that at most one event may occur in Δt.

The solution of (7.16) is not a trivial one since there is a quite general dependence structure governing the increments of the rainfall counts. A convenient way to solve (7.16) is through the introduction of the multivariate pgf $\phi(z_1, z_2, \ldots, z_J, z_{J+1})$ such that

$$\phi(z_1, z_2, \ldots, z_{J+1}) = G_{N_{\Delta t}, N_{2\Delta t}, \ldots, N_{(J+1)\Delta t}}(z_1, \ldots, z_{J+1}) \,. \tag{7.17}$$

The interval $(0, (J+1)\Delta t)$ is divided into $(J+1)$ non-overlapping intervals of equal length Δt and a counting random variable $N_{i\Delta t}$ is assigned to each interval $[(i-1)\Delta t, i\Delta t]$. For Δt small enough so that at most one rainfall can occur in Δt,

$$\begin{aligned} P[\text{Wet period duration} = J\Delta t] &= P[N_{\Delta t}=1, N_{2\Delta t}=1, \ldots, N_{J\Delta t}=1, N_{(J+1)\Delta t}=0] \\ &= \partial^J\phi(\cdot)/\partial z_1 \partial z_2 \ldots \partial z_J|z_1=0, z_2=0, \ldots, z_J=0, z_{J+1}=0 \,. \end{aligned} \tag{7.18}$$

The multivariate pgf $\phi(z_1, \ldots, z_{J+1})$ of the Neyman-Scott cluster model for the non-overlapping intervals can be shown to be (*Neyman and Scott,* 1958),

$$\phi(z_1, z_2, \ldots, z_{J+1}) = \exp\left\{-h_0 \int_{-\infty}^{T} \left(1 - G_\nu\left[1 - \sum_{i=1}^{J+1} (1-z_i)\, p_i(u)\right]\right) du\right\} \,, \tag{7.19}$$

where $p_i(u) = \int_{(i-1)\Delta t}^{i\Delta t} f(\tau-u)d\tau$, $i = 1, 2, \ldots, J+1$. The presence of dependence in the rainfall counting increments complicates the calculation of the wet period duration tremendously. However, it is still possible to obtain the wet period duration probabilities although the calculations are lengthy.

As an example, the probability of a wet period duration to be equal to $2\Delta t$, is calculated as follows:

$$\begin{aligned} P[\text{Wet period duration} = 2\Delta t] &= \partial^2\phi(z_1,z_2,z_3)/\partial z_1 \partial z_2|z_1=0,z_2=0,z_3=0 \\ &= \frac{\partial 2}{\partial z_1 \partial z_2} \exp\left\{-h_0 \int_{-\infty}^{T} (1-G_\nu[1-(1-z_1)p_1(u)-(1-z_2)p_2(u)-(1-z_3)p_3(u)])du\right\}\Big|_{z_1=0,z_2=0,z_3=0} \\ &= \exp\left\{-h_0 \int_{-\infty}^{T} \left(1-G_\nu\left[1 - \sum_{J=1}^{3} (1-z_J)p_J(u)\right]\right) du\right\}\left[h_0 \int_{-\infty}^{T} \frac{\partial G_\nu[\cdot]}{\partial z_1}\, du\right]\left[h_0 \int_{-\infty}^{T} \frac{\partial G_\nu[\cdot]}{\partial z_2}\, du\right] \\ &\quad + \exp\left\{-h_0 \int_{-\infty}^{T} (1 - G_\nu[\cdot])du\right\}\left[h_0 \int_{-\infty}^{T} \frac{\partial^2 G\,[\cdot]}{\partial z_1 \partial z_2}\, du\right]\Bigg|_{z_1=0,z_2=0,z_3=0} \,, \end{aligned} \tag{7.20}$$

where $$\frac{\partial G_\nu[]}{\partial z_1}\Big|_{z_1=0,z_2=0,z_3=0} = \frac{p p_1(u)}{[p + (1-p) \sum_{J=1}^{3} p_J(u)]^2}$$

$$\frac{\partial G_\nu[\,]}{\partial z_2}\Big|_{z_1=0,z_2=0,z_3=0} = \frac{pp_2(u)}{[p + (1-p)\sum_{J=1}^{3} p_J(u)]^2}$$

$$\frac{\partial^2 G_\nu[\,]}{z_1 \;\; z_2}\Big|_{z_1=0,z_2=0,z_3=0} = \frac{2p(1-p)\; p_1(u)\; p_2(u)}{[p + (1-p)\sum_{J=1}^{3} p_J(u)]^3}$$

and $p_1(u) = \int_0^{\Delta t} \theta e^{-\theta(\tau-u)}d\tau,\; p_2(u) = \int_{\Delta t}^{2\Delta t} \theta e^{-\theta(\tau-u)}d\tau,\; p_3(u) = \int_{2\Delta t}^{3\Delta t} \theta e^{-\theta(\tau-u)}d\tau \quad \text{if } -\infty < u < 0\,,$

$p_1(u) = \int_u^{\Delta t} \theta e^{-\theta(\tau-u)}d\tau,\; p_2(u) = \int_{\Delta t}^{2\Delta t} \theta e^{-\theta(\tau-u)}d\tau,\; p_3(u) = \int_{2\Delta t}^{3\Delta t} \theta e^{-\theta(\tau-u)}d\tau \quad \text{if } 0 < u < \Delta t$

$p_1(u) = 0 \quad ,\; p_2(u) = \int_u^{2\Delta t} \theta e^{-\theta(\tau-u)}d\tau,\; p_3(u) = \int_{2\Delta t}^{3\Delta t} \theta e^{-\theta(\tau-u)}d\tau \quad \text{if } \Delta t < u < 2\Delta t$

$p_1(u) = 0 \quad ,\; p_2(u) = 0 \quad ,\; p_3(u) = \int_{2\Delta t}^{3\Delta t} \theta e^{-\theta(\tau-u)}d\tau \quad \text{if } 2\Delta t < u < 3\Delta t\,.$

7.7 NUMBER OF RAINFALL OCCURRENCES IN A TIME INTERVAL (0,t)

The probability of a number J of rainfall occurrences in (0,t) can be obtained from the J-th derivative of the pgf of the Neyman-Scott cluster model. That is,

$$P[N_t = J] = \frac{1}{J!}\, \partial^J G_{N_t}(z)/\partial z^J\big|_{z=0} \qquad (7.21)$$

where $G_{N_t}(z) = \exp\{-h_0 \int_{-\infty}^{t} (1 - G_\nu[1 - (1-z)\, p(u)])du\}$. As an example, the probability of 2 rainfall occurrences in (0,t) will be calculated. Using (7.21),

$$P[N_t = 2] = \frac{1}{2}\, \partial^2 G_{N_t}(z)/\partial z^2\big|_{z=0}$$

$$= \frac{1}{2} \exp\left\{-h_0 \int_{-\infty}^{t} \frac{p(u)}{p + (1-p)\; p(u)}\, du\right\} \cdot \left\{\left[h_0 \int_{-\infty}^{t} \frac{pp(u)}{p + [(1-p)\; p(u)]^2}\, du\right]^2 + h_0 \int_{-\infty}^{t} \frac{2p(1-p)\; p^2(u)}{[p + (1-p)\; p(u)]^3}\, du\right\} \qquad (7.22)$$

where

$$p(u) = \int_0^t \theta e^{-\theta(\tau-u)}d\tau \quad \text{for } -\infty < u < 0$$

$$= \int_u^t \theta e^{-\theta(\tau-u)}d\tau \quad \text{for } 0 < u < t$$

$$= 0 \qquad \text{for } u > t\,.$$

The terms in (7.22) are expressed as

$$-h_0 \int_{-\infty}^{t} \frac{p(u)}{p + (1-p)\; p(u)}\, du = I_1 + J_1$$

$$h_0 \int_{-\infty}^{t} \frac{pp(u)}{[p + (1-p)\ p(u)]^2}\, du = I_2 + J_2, \text{ and } h_0 \int_{-\infty}^{t} \frac{2p(1-p)\ p^2(u)}{[p + (1-p)\ p(u)]^3}\, du = I_3 + J_3 ,$$

where

$$I_1 = \frac{-h_0}{\theta(1-p)} \ln \frac{1 - (1-p)\ e^{-\theta t}}{p}$$

$$I_2 = \frac{-h_0}{\theta(1-p)} \left[\frac{1}{1 - (1-p)\ e^{-\theta t}} - \frac{1}{p}\right]$$

$$I_3 = \frac{2ph_0}{\theta(1-p)} \left[\frac{-1}{1 - (1-p)\ e^{-\theta t}} + \frac{p}{2[1 - (1-p)\ e^{-\theta t}]^2} + \frac{1}{2p}\right]$$

$$J_1 = -h_0 t - \frac{h_0 p}{(1-p)} \ln \frac{p}{1 - (1-p)\ e^{-\theta t}}$$

$$J_2 = \frac{h_0 p}{\theta} \left[- \frac{1}{1-p} + \theta t + \frac{p}{(1-p)\ [1 - (1-p)\ e^{-\theta t}]} - \ln \frac{p}{1 - (1-p)\ e^{-\theta t}}\right]$$

$$J_3 = \frac{2h_0 p(1-p)}{\theta} \left[\frac{1}{2}\, \frac{(1-p)^2}{p^2} + \theta t - \frac{1}{2} \cdot \frac{2 - (1-p)\ e^{-\theta t}}{1 - (1-p)\ e^{-\theta t}} + \ln \frac{[1 - (1-p)\ e^{-\theta t}]}{p}\right]$$

$$- \frac{2h_0 p}{\theta} \left[\frac{1}{p^2} - \frac{1}{[1 - (1-p)\ e^{-\theta t}]^2}\right]$$

$$+ \frac{2h_0 p}{\theta(1-p)} \left[\frac{1-2p}{2p^2} + \frac{1}{1 - (1-p)\ e^{-\theta t}} - \frac{1}{2[1 - (1-p)\ e^{-\theta t}]^2}\right] .$$

From the theoretical and applied analysis of the Neyman-Scott cluster model for the daily rainfall occurrences it can be concluded that the model not only preserves the probability distributions obtained from the data but also the dependence structure that is present in the stationary daily rainfall counts process. However, in its present form the model is still of limited use since it is stationary. A non-homogeneous form of the Neyman-Scott cluster model needs to be constructed so as to model the long-term trends, the within-the-year cyclicities and the dependencies in the daily rainfall counts process, whose presence is established through the point stochastic analysis of the daily rainfall occurrence data in the 17 stations in the state of Indiana.

TABLE 7-1

CLUSTER MODEL PARAMETERS OBTAINED FROM THE FIT TO THE LOG-SURVIVOR FUNCTION OF THE HOMOGENIZED DAILY RAINFALL OCCURRENCES

Station	$\hat{h}_0$	$\hat{p}$	R^2
0132	.95899	.94224	.9863
0177	.86275	.85043	.9887
0545	.90137	.89329	.9907
3777	.93061	.93928	.9927
4642	1.01848	.99964	.9890
6056	.82674	.80731	.9867
7747	.94417	.93489	.9929

TABLE 7-1A

CLUSTER MODEL PARAMETERS OBTAINED FROM THE FIT TO THE LOG-SURVIVOR FUNCTION OF THE HOMOGENIZED DAILY RAINFALL OCCURRENCES IN THE REMAINING 11 STATIONS

Station	$\hat{h}_0$	$\hat{p}$	R^2
1747	.975	.959	.9927
1882	1.05	1.00	.9945
3082	1.35	.95	.9837
3547	.99	.982	.9885
4908	1.02	1.00	.9885
6164	.90	.90	.9914
6338	1.02	1.00	.9892
7069	1.04	1.00	.9948
7755	.95	.93	.9935
7935	.98	.97	.9932

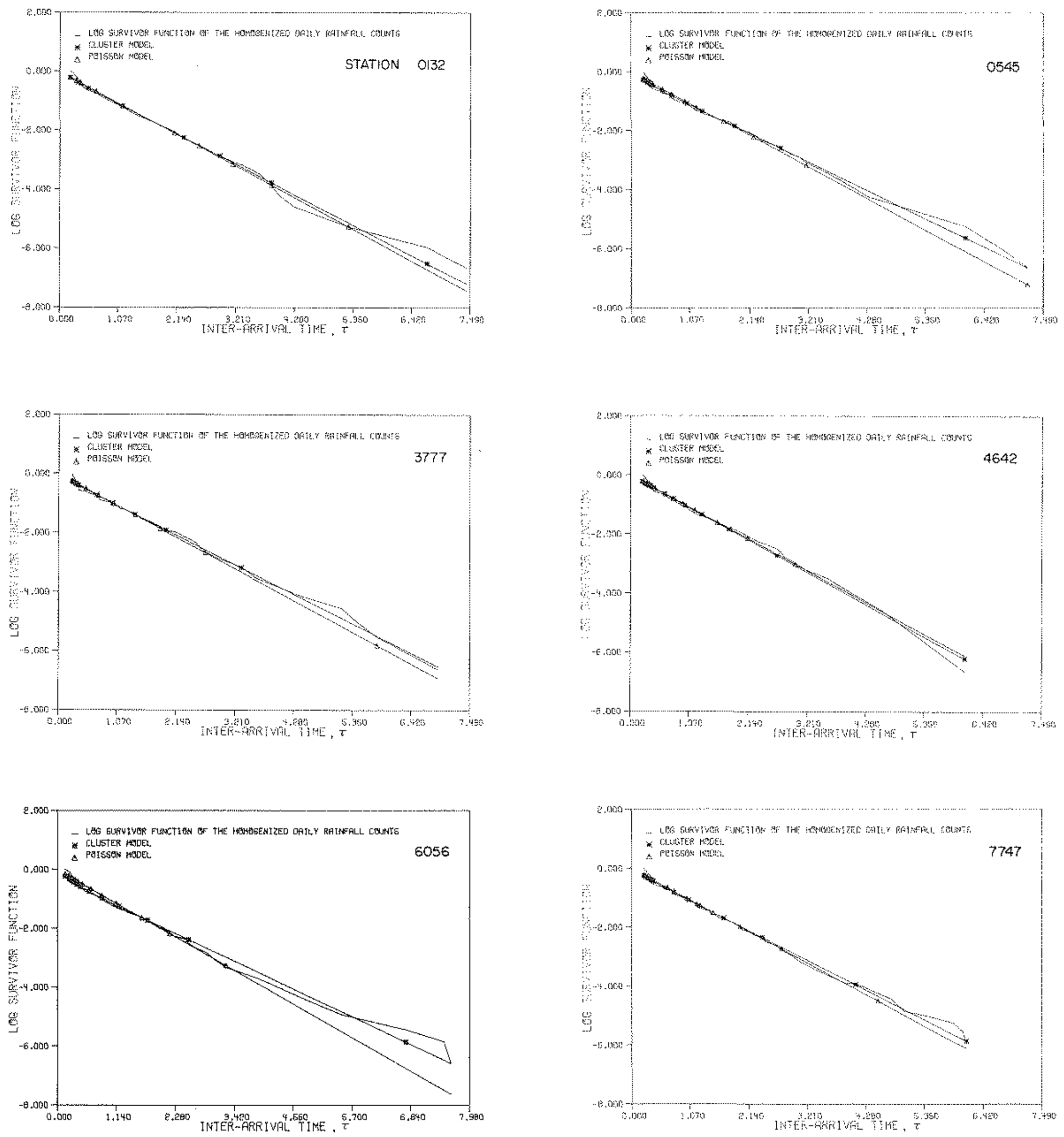


FIG. 17- THE FIT TO THE LOG SURVIVOR FUNCTION OF THE HOMOGENIZED DAILY RAINFALL COUNTS BY THE CLUSTER AND POISSON MODELS

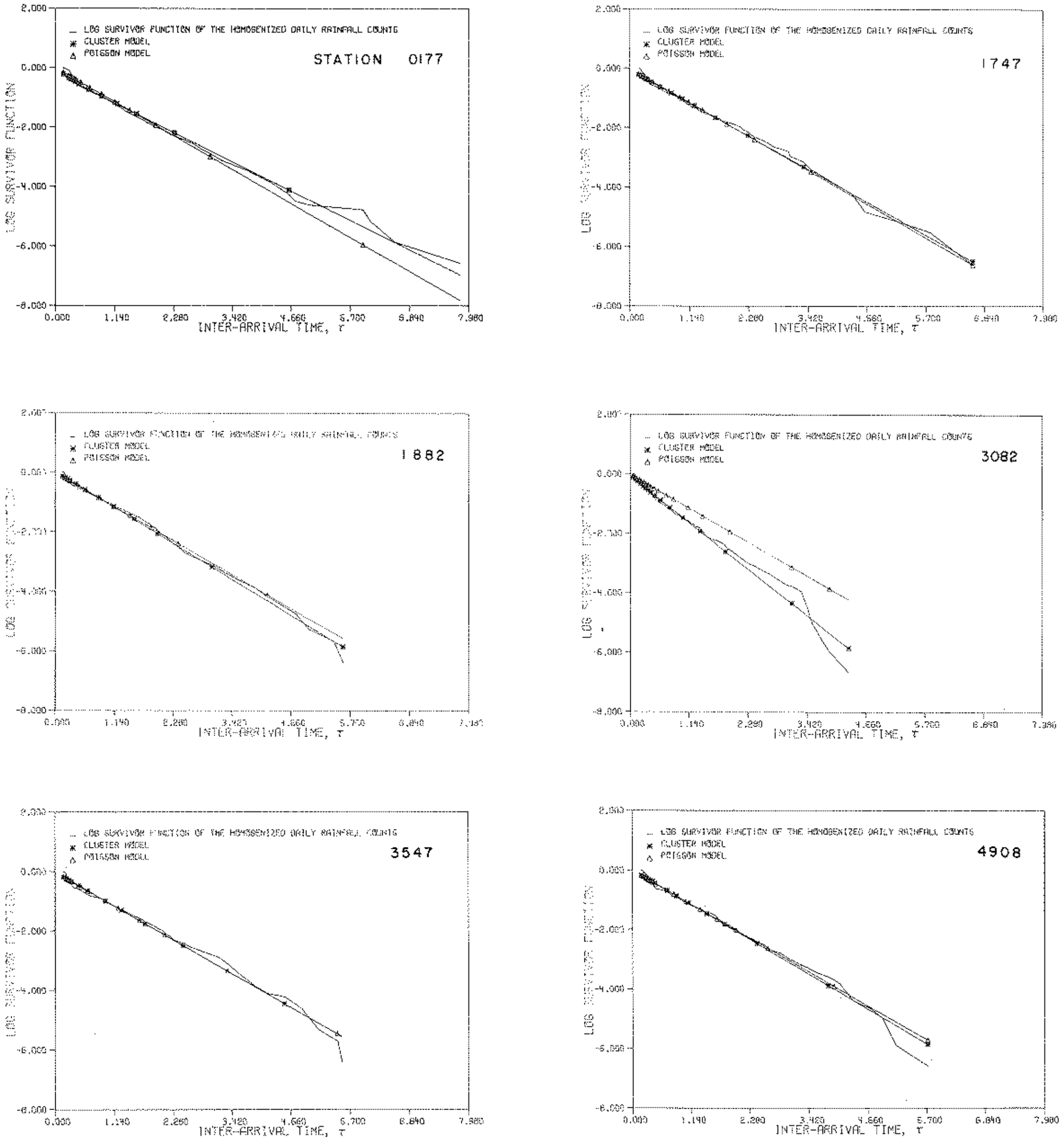


FIG. 17A—THE FIT TO THE LOG SURVIVOR FUNCTION OF THE HOMOGENIZED DAILY RAINFALL COUNTS BY THE CLUSTER AND POISSON MODELS

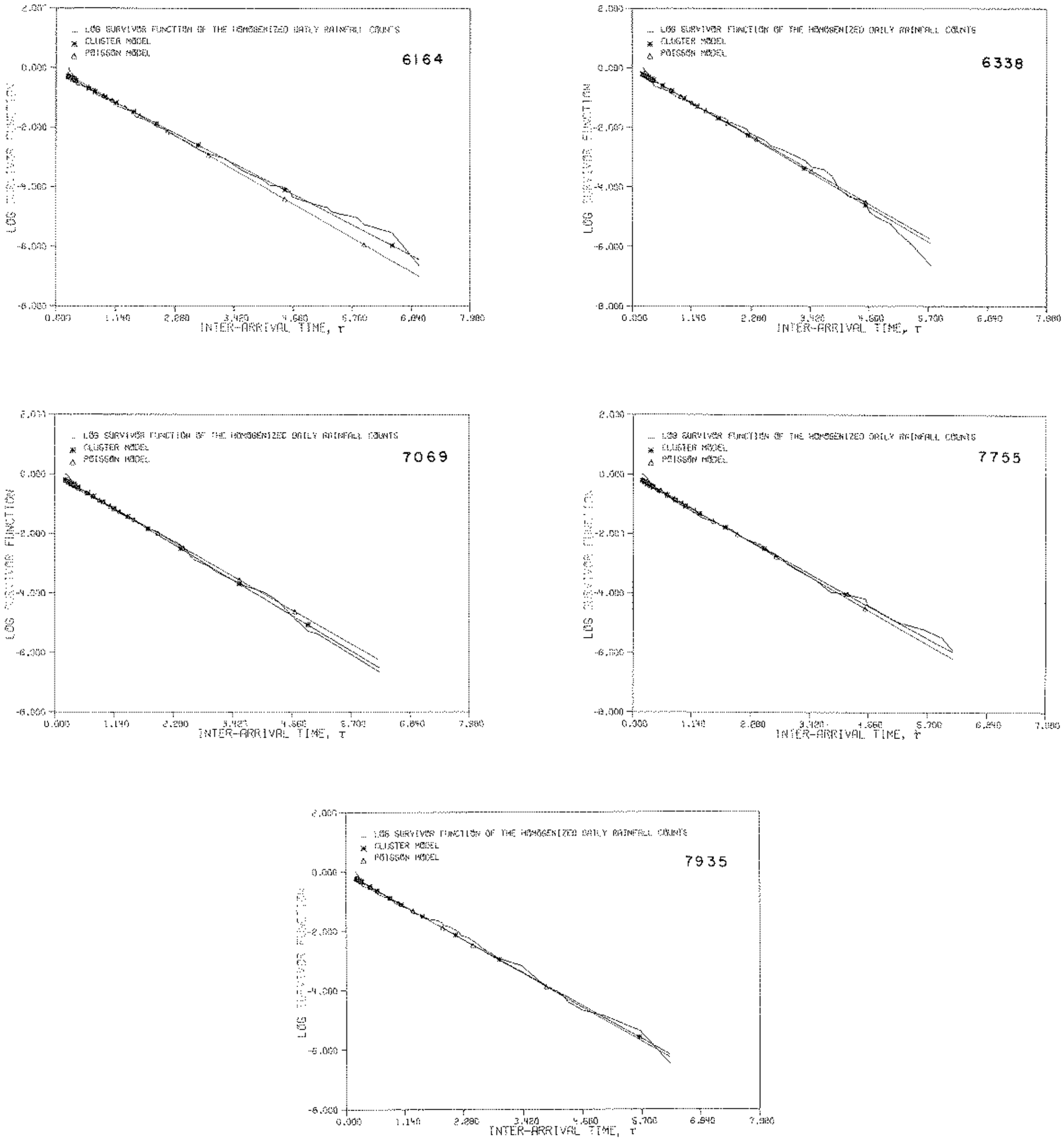


FIG. 17B — THE FIT TO THE LOG SURVIVOR FUNCTION OF THE HOMOGENIZED DAILY RAINFALL COUNTS BY THE CLUSTER AND POISSON MODELS

CHAPTER 8 - DISCUSSION OF THE RESULTS

8.1 LONG-TERM TRENDS AND CYCLICITIES IN THE DAILY RAINFALL OCCURRENCES

The plots of the cumulative daily rainfall counts versus time indicated a slight downward trend in the mean rate of daily rainfall occurrences in Indiana. This trend may be interpreted as the dipping portion of the 80 to 90 year cycle or of the 11 year cycle due to the systematic variation in the sunspot numbers (*Mitchell*, 1964). These plots extend to two years in time. What is interpreted as a downward trend can also be interpreted as a two year cycle which is very conspicuous in the tropics (*Mitchell*, 1964).

To find a more definite answer to these speculations the spectrum of the daily rainfall counts was computed for the 17 stations in Indiana. Besides the obvious annual cycle, the semi-lunar cycle of 15 days was quite conspicuous in the data. In 13 out of 17 cases a periodicity of 11.6 to 16 days was very significant in the spectrum of the daily rainfall counts. This periodicity of approximately 15 days has two physical interpretations; (a) the effect of the lunar synodical period of 29.53 days on earth, which emerges as a 15 day cycle, and which is also clearly shown in figure 11 of *Mitchell* (1964), and (b) the cycle of 15 days in the atmosphere which has been consistently observed in Indiana and is called an index cycle (*Newman*, 1975). This index cycle is due to the 15 day periodicity in the meridional and zonal air flows over Indiana.

In 12 out of 17 cases periodicites ranging from 3 to 9 days were observed. Indiana is under the influence of the Atlantic cyclone regime. The arctic air front that separates polar continental air from an intrusion of the arctic air from the north passes through Indiana in the winter time. This persistent front causes extensive and persistent precipitation. The life cycles of cyclone families over Indiana during the winter may have something to do with the 3 to 9 day cycles. During the summer the polar front moves north. However, the frontal disturbances still cause showers and thunderstorms in Indiana during the summer. These thunderstorms are of two types; (a) the line thunderstorms which pass over Indiana in 1 to 5 days and (b) the scattered type which passes over Indiana in 1 to 10 days (*Newman*, 1975). As was mentioned earlier, due to the clustering of the thunderclouds, the life span of a thunderstorm can extend to durations of the order of days (*Petterssen*, 1969). *Visser* (1944) has shown that the mean interarrival time of the rainfall occurrences in Indiana is approximately 3.3 days which explains the 3 to 4 day cycle. The 3 to 9 day periodicities highly vary among the different stations. Thus their effect may be neglected in the prediction of the daily rainfall occurrences.

The rate of occurrence functions for the daily rainfall counts were computed to observe the periodicities and the long-term trends in the first moment of the counting process. As is seen in figure 2, there are cyclicities and a slight downward trend. The exponential harmonic fits to the mean rate of occurrence function showed that there is a slight downward trend in the 7-year span analyzed. These results led to the rejection of the significance of a biennial cycle but strengthened the hypothesis of the presence of longer cycles due to sunspots. The results obtained from the intensity function confirmed the strong annual cycle in the daily rainfall occurrences. The periodicities in the second moment of the daily rainfall counting process were observed by the variance-time curve of counts. As is seen in figure 4 there is a distinct annual cycle in the variance of the daily rainfall counts. The neglect of the annual periodicity in the variance may lead to the construction of inadequate stochastic models. *Smith and Schreiber* (1973) pointed to the insufficient analysis of the counts variance as a cause of the inadequate fit of the Binomial and Markov models to the cumulative distribution of wet days per season in Arizona.

The detection of the cycles and the long-term trends is quite important for the model calibration.

The nonhomogeneous Markov chains previously applied to the daily rainfall data were calibrated under the assumption of circular stationarity with yearly circularity. This assumption excludes the effects of longer cycles which may be quite important for the prediction lengths longer than a year.

A statistical test of trend in the rate of daily rainfall occurrences based on Cramer's statistic confirmed the nonhomogeneity in the first moment of the daily rainfall occurrence process. However, the results of this test should be taken with caution since the stochastic process underlying the test is Poisson. A statistical test of variance homogeneity, using Bartlett's likelihood ratio statistic confirmed the nonhomogeneity in the second moment of the daily rainfall occurrence process. However, the results of this test should also be taken with caution since the samples used for the test are assumed to be normal.

8.2 DEPENDENCE STRUCTURE OF THE DAILY RAINFALL COUNTS

The important statistical functions for the analysis of the persistence of the daily rainfall counting process are the counts spectrum, the variance-time function of counts and the log-survivor function of the counts.

The counts spectrum can identify the covariance structure of the daily rainfall counting process since it is the Fourier transform of the covariance density function of the respective differential counting process. As was shown in Chapter 2, it is analogous to the spectrum of a time series $\{X_t\}$. The difference is that the counting random variable N_t replaces the time series value X_t. Although, a theory describing the general behavior of the dependent counting processes is not yet established, it is known that an independent increment counting process yields a horizontal spectrum. This is in analogy to the white noise spectrum of the time series analysis. It was seen in Chapter VI that the Neyman-Scott cluster model has a very general spectral form. The general appearances of the spectrum can easily explain the theoretical behavior of different types of dependence. For example, an exponentially decaying spectrum of the homogenized daily rainfall counts indicates a short memory dependence which could be modeled by a first-order Markov chain if discrete time steps were used in analysis. However, for the point stochastic analysis, a special form of the cluster model with exponentially distributed rainfall positions within a storm, can model the exponential, short-memory dependence.

The spectra of the Indiana daily rainfall counts data show a definite dependence mechanism which should be represented by a stochastic model that can preserve dependence. Through the preservation of the empirical counts spectrum the stochastic model can preserve the explicit covariance structure of the parameters of the model pertaining to the preservation of the dependence were calibrated by a least squares fit to the homogenized daily rainfall counts spectrum.

The asymptotic slope of the variance-time function of counts is equal to the origin value of the counts spectrum. Therefore, the long term dependence characteristics of the daily rainfall counting process can be obtained from the asymptotic characteristics of the variance-time function. Unfortunately, due to the computer storage limitations, the function could not be computed to a sufficient length so as to observe the asymptotic characteristics.

The variance-time function behavior in the homogenized (stationary) domain can yield valuable information about the dispersion characteristics of the daily rainfall counting process. When the empirical variance-time curve is above the theoretical variance-time function of the Poisson model, this means that the rainfall counts are overdispersed. Actually it can be shown through the cluster model that overdispersion may be interpreted as the grouping of the rainfalls in the form of storms. This was the case observed in 10 out of 17 stations. Therefore, the dependence, due to grouping or clustering, was identified from the variance time curve of the data.

The log-survivor function can be effectively used to assess the dependence structure of the dry days. In 11 out of 17 cases there was a slight convexity in the log survivor functions above the theoretical log-survivor function of the Poisson process. It was discussed in Chapter 3 that convexity implies overdispersion and thus, clustering of the rainfalls.

From the behavior of the counts spectrum, the variance time-function of counts, and the log-survivor function of counts it can be concluded that there is a short term dependence in the daily rainfall counts in Indiana. This dependence can be explained by the clustering of the rainfalls within a storm. Due to the comparatively short life of the storm generating mechanisms, the dependence is short and can be modeled by short memory stochastic models.

8.3. THE NEYMAN-SCOTT CLUSTER MODEL FOR THE POINT STOCHASTIC PROCESS OF THE DAILY RAINFALL OCCURRENCES

The stochastic models such as the simple Markov chain, the alternating renewal process, the Poisson process, etc., used for the daily rainfall occurrence process are black box models with very little physical meaning. They are one level models in that they are concerned only with the occurrence of the end product, the rainfall occurrence, of a complex atmospheric process. It is believed that valuable meteorologic knowledge about the rainfall generating mechanisms such as the cyclones or thunderstorms can be utilized in the prediction of the rainfall occurrences. This can be done by incorporating the occurrence process of the rainfall generating mechanisms as a primary process into the stochastic model of the rainfall occurrences. The actual occurrence of the rainfalls will then be at the secondary level and will be a product of the primary process. The application of the Neyman-Scott cluster model to the daily rainfall occurrence process is a first attempt is this direction. The Neyman-Scott cluster model for the rainfall occurrences is based on physical observations. This had not been done by the previous researchers. As a model of the point rainfall occurrences the Neyman-Scott cluster process contains the physical concepts of (a) the life length of a rainfall generating mechanism (e.g., a cyclone) that determines the storm duration and the dependence of the rainfall counts, (b) the storm structure that is defined by the number of rainfalls and their time positions within that storm.

In this research a storm is defined as the group of rainfalls that are generated by the same rainfall generating mechanism. Thus, due to the memory of cyclones over Indiana during the winter, two storms may overlap and contribute to the number of rainfalls in a time interval. The positions of the rainfalls within a storm are taken to be random in the cluster model. If the type of the rainfall generating mechanism could be identified, the meteorologic characteristics of the mechanism could be used to predict the time positions of the rainfalls. However, the identification of the mechanisms seems to be a difficult task in Indiana. The meteorologic mechanisms over the Great Lakes region are very complex during the winter. The directions of the Alberta and Colorado cyclone storms converge on the Great Lakes. There is also the northward movement of the storms formed over the Gulf of Mexico towards the Great Lakes. Then there are the disturbances due to the differential heating of the lakes and the land. All these effects combine to form a high frequency of precipitation during the winter (*Petterssen,* 1969).

Actually a three-level cluster process could also be formed (*Neyman and Scott,* 1958).where in the primary level there would be the occurrence of the Atlantic cyclone belt which dominates Indiana's weather during the winter time. This mechanism would generate the cyclones in the second level and the cyclones would generate the rainfalls on the third level. However, this is not a general description since there are other rainfall generating mechanisms besides the cyclones. A three-level model thus seems impractical for hydrologic applications.

The explicit spectral structure of the Neyman-Scott model is known. This spectrum, as given by the expression (6.21), is very general and can fit various types of persistence. As was pointed out by *Gabriel and Neumann* (1962), *Wiser* (1965) and others, the first-order Markov chain has a fixed geometric memory for the dry and wet sequences. *Jorgensen's* (1949) data of dry days at San Francisco and *Newnham's* (1916) data for British Isles are examples of the inadequacy of the simple Markov chain in fitting the various dependence structures seen in the dry and wet sequences. On the other hand, and especially for the dry sequences, the Neyman-Scott cluster model is very convenient. To preserve the probability structure of sequences which are more persistent than the first-order Markov chain's geometric memory, one may use a suitable probability distribution of the rainfall positions with respect to the storm origin since this distribution decides the form of the counts spectrum. This can be easily inferred from expression (6.21).

Then there are the extremely persistent dry and wet sequences. The two-year wet period at Cherrapunji, India (*Jennings*, 1950) during the years 1860 and 1861, or the four-year dry period at Iquique, Chile (*Petterssen*, 1969) cannot be modeled by anyone of the current stochastic models of the rainfall occurrence process. However, even this type of extreme phenomenon could be modeled by the Neyman-Scott cluster model by finding a suitable probability distribution for the positions of rainfalls. *Vere-Jones and Davies* (1966) showed that an inverse power law of the form

$$f_T(\tau-u) = \rho c^{\rho}/(c+\tau-u)^{\rho+1} \qquad c > 0,\ 0 < \rho < 0.5$$

yields a very sharp peak at the counts spectrum origin. This means that utilization of such an inverse power law for the distribution of the rainfall positions within a storm would enable the cluster model to preserve a highly dependent rainfall counting process.

The first step in the construction of a time series model involves the computation of the spectrum or of the autocorrelation function of the time series. Based on the behavior of these functions a suitable model is then selected and its parameters are calibrated so as to preserve the covariance structure. Once the model is constructed, the pdf of the white noise input is obtained.

An analogous methodology based on the general spectrum of the Neyman-Scott cluster process can be used for the modeling of the point stochastic counting process. In order to preserve the covariance structure of the rainfall counts the cluster model was first fitted to the counts spectrum. As is seen from (6.21) the pdf of the time positions, τ, of the rainfalls defines the shape of the counts spectrum. By using the normalized counts spectrum, the rate of occurrences of the rainfall generating mechanisms, h_0, was eliminated from the spectral expression. It was seen from a preliminary spectral fit that the ratio $E(\nu)^2/E(\nu)$ of the second moment to the mean of the number of rainfalls within a storm stays as a constant approximately equal to 2. Thus only the parameter θ of the exponential distribution of the rainfall positions was calibrated from the counts spectrum. As is seen from figure 13, the cluster model satisfactorily preserves the covariance structure of the homogenized daily rainfall counts.

The remaining parameters h_0 and p (the parameter of the geometric distribution for the number of rainfalls within a storm) were then calibrated from the log-survivor function using (7.14). The cluster model fits the log-survivor function of the daily rainfall occurrences very well as can be seen from the R^2 values in Table 7-1.

Since the rate of occurrence function was set to unity in the homogenization scheme (3.6), h_0 should be approximately equal to p. This condition is also satisfied as is seen in Table 7-1. However, there is a discrepancy between the $\hat{p}$ estimated from the log-survivor fit and the $\hat{p}$ which was assumed in the spectral fit. From the log-survivor fit $\hat{p}$ is around .9 while the assumed $\hat{p}$ was .66 in order to satisfy $(E(\nu^2)/E(\nu)) = 2$ under the geometric law. For $\hat{p}$ = .66 the normalized counts spectrum is

$$\frac{g_+(\omega)}{h_0 E(\nu)} = 1 + \frac{\theta^2}{\theta^2 + \omega^2}, \quad \omega > 0$$

while for $\hat{p}$ = .9 the normalized counts spectrum becomes

$$\frac{\pi g_+(\omega)}{h_o\ E(\nu)} = 1 + .35\ \frac{\theta^2}{\theta^2 + \omega^2}\ ,\ \omega > 0$$

so that the shape of the spectrum stays the same. The discrepancy may be due to the geometric law assumption for the number of rainfalls within a storm. Another possible cause may be the use of an approximate scheme for the homogenization of the rainfall process under the cluster hypothesis.

It should be emphasized that the application of the cluster model to the rainfall occurrence process is in the exploratory stage. Future work is needed for the investigation of different probability distributions for the rainfall positions and the rainfall numbers within a storm. The nonhomogeneous form of the cluster model is definitely needed in order to avoid the discrepancies that may occur from an approximate homogenization scheme.

In the application of the Neyman-Scott cluster model to the point rainfall occurrences on the time axis only the primary process is fixed to be Poisson. The secondary process of the rainfall occurrences is very general. The pdf's of the two random variables τ and ν that decide the storm structure are chosen according to the climatologic characteristics of the region under study. In this study τ was modeled by a negative exponential distribution and ν was modeled by a geometric distribution. Of course there are many possible distributions that could be used for the modeling of τ and ν. It was seen in Chapter VI that the cluster model easily reduces to the generalized Poisson model in the case that the rainfall generating mechanisms are memoriless or that the counting process has independent increments. The model is quite general so that it can be applied to various climatologic conditions.

The cluster model presented in this report is a complete model from hydrologic point of view. It not only accounts for the observed meteorologic facts but also preserves both the covariance structure and the marginal probabilities of the rainfall occurrence phenomenon.

It is to be emphasized that the Neyman-Scott cluster model is a point stochastic model. It models the occurrences of the point rainfall sequences on the continuous time axis. Since the rainfall observations are made at equi-spaced time intervals, assumptions are needed in order to analyze the rainfall occurrences as a point stochastic process. This may be a shortcoming. On the other hand if the raingages are recording the rainfall in continuous time, the rainfall counting process can be modeled directly by a point stochastic process.

8.4. APPLICATIONS OF THE PROPOSED METHODOLOGY FOR POINT STOCHASTIC PROCESSES OF DAILY RAINFALL OCCURRENCES

In this report the methodology for the identification and the calibration of the point stochastic processes developed by *Cox and Smith* (1953), *Neyman and Scott* (1958), *Bartlett* (1963), *Cox and Lewis* (1966), *Lewis et al.* (1969), *Lewis* (1970, 1971), *Vere-Jones* (1970) and various other statisticians is introduced to hydrology for the objective selection and calibration of the hydrologic point stochastic models. The methodology is also utilized for the detection of long term trends and cyclicities in the point hydrologic data.

Analogous to the approach used in the time series analysis the data are analyzed as an entity. First, the time trends in the data are identified through the use of the rate of occurrence, the intensity, the counts spectrum, and the variance time functions, and by the employment of approximate statistical tests of the trend in the rate of occurrence and in the variance. Once the trends are identified, a homogenization scheme is needed to transform the data from the nonstationary domain to the stationary domain. Such a homogenization scheme is developed under the Poisson hypothesis. Under the assumption that the point hydrologic process is Poisson, this scheme can be effectively used to stationarize the data and test the Poisson hypothesis by various statistical functions and tests in the stationary domain. However, a serious problem emerges when the Poisson hypothesis is rejected. The homogenization scheme (3.6) employed for Poisson hypothesis is only approximate for other processes. For example, in the case of the cluster model, the scheme (3.6) can homogenize only the first moment of the counting process. The second and higher moments of the process are only approximately homogenized.

The covariance structure of the point hydrologic process is obtained from the counts spectrum. The long range dependence characteristics of the point hydrologic process are identified from the asymptotic slope of the variance time curve. The general shape of the log-survivor function also gives valuable information about the persistence of wet and dry sequences. This function was already applied to rainfall counting process by *Wiser* (1965), and by *Smith and Schreiber* (1973). Once the explicit covariance structure of the point hydrologic counting process is identified, the knowledge about the theoretical counts spectra of the various point stochastic models enables the hydrologist to choose the most suitable model. The model whose counts spectrum behaves in the same manner as that of the empirical counts spectrum of the data is the most suitable one. The Neyman-Scott cluster model provides a general family of point stochastic models when different distributions are used for the rainfall positions, τ, and the number of rainfalls, ν, within a storm. It was shown earlier that generalized Poisson and the simple Poisson models are special cases of the cluster model. Thus the Neyman-Scott cluster model may assume a role in the point stochastic processes, analogous to the role of the mixed autoregressive-moving average (ARMA) family in the time series modeling.

A specific form of the Neyman-Scott cluster model with exponentially distributed rainfall positions and geometrically distributed rainfall numbers within a storm, was fitted to the homogenized daily rainfall counts in this study. From the results of the fits to the counts spectra and to the log-survivor functions, Neyman-Scott cluster model emerges as a very flexible and powerful model for the point stochastic process of the rainfall counts. However, as was discussed earlier, the model is stationary and, in its present state, can only be used to model the stationary intervals. A nonhomogeneous form of the Neyman-Scott model is necessary to preserve not only the stationary covariance structure and the marginal probability distributions of the daily rainfall counts but also the long-term trends and the cyclicities in the process.

The practical probabilities of dry period lengths from an arbitrary time origin, and from a rainy event are explicitly derived for the specific form of the cluster model used in this study. These expressions are simple to use and give the probability distributions for droughts. Since the theoretical log-survivor function for the specific form is derived, the cluster model can be calibrated in such a way as to specifically preserve the drought length probabilities. The return period of the rainfalls is the expected interarrival time between two consecutive rainfall occurrences. It is obtained directly from the log-survivor function fit since it is given by h_0 and E (ν) which are calibrated from the log-survivor function.

The probability distribution of the number of rainfall occurrences N_t within a period (o,t) is obtained directly from the probability generating function of the cluster model. However, the computations are lengthy and the utilization of the computer is necessary to obtain the explicit probability values. The same thing can be said about the probability distribution of the wet period durations. Due to the general dependence structure underlying the cluster model, the computations increase at a very fast rate with larger durations. The computer becomes a must for the computation of the practical probabilities of wet-period durations under the cluster model.

The same methodology is also applicable to floods exceeding a specified threshold level.

The cluster model can simulate the rainfall occurrences by first generating the primary process of the storm generating mechanisms and then the storms corresponding to each of the mechanisms. However, the rainfall quantities should be incorporated to the rainfall counting process in order to obtain a complete simulation scheme of the point rainfalls. In order to obtain the runoff sequences, a rainfall-runoff model is needed.

APPENDICES FOR PART I

APPENDIX A: THE DATA ACQUISITION

The daily rainfall data were acquired from the U. S. Weather Bureau in Asheville, N. C. The data were on magnetic tapes which were written in BCD, 7 track, 556 BPI, and even parity. The magnetic tapes contained not only the daily rainfall but also temperature, snow, wind and evaporation data. The record length was 74 characters and the precipitation data were in columns 23-26. When the precipitation data were missing the field was blank. The field had a code between 0001 and 9999 when the precipitation depth was 00.01 inches to 99.99 inches. 000X meant that the depth was less than 0.005 inches. The day was considered to be a rainy day only when the depth was greater than or equal to 0.01 inches.

The missing rainfall data on the magnetic tapes posed an important problem. The missing data were filled by hand from the climatological data publications of the U. S. Weather Bureau whenever it was possible. For the rest of the missing data a stepwise multiple linear regression was used. The three closest neighboring rainfall stations were selected and a multiple linear regression was run between the station with the missing data and the neighboring stations. The missing data for a particular date were calculated by the regression equation from the data of the neighboring stations.

A library of the daily rainfall data was formed for selected stations in Indiana, Ohio, Illinois and Kentucky for future use.

APPENDIX B: DESCRIPTION OF COMPUTER PROGRAMS FOR THE POINT STOCHASTIC ANALYSIS OF THE DAILY RAINFALL OCCURRENCES

The programs involved in the point stochastic analysis of the daily rainfall occurrences are mostly contained in the SASE IV computer program prepared by *Lewis et al.* (1969). This program, written for the IBM 360, was translated into CDC 6500 and modified so as to obtain CALCOMP plots of the various statistical functions. Due to the limitation of the computer capacity at Purdue University, the dimensions were reduced from a capability of 1999 events to a capability of 800 events. In its present form, when all the subroutines are used, the program requires a field length of 107,000 words and approximately 550 seconds for execution time. The input data consists either of $\{t_i\}$, the times to events, or of $\{X_i\}$, the times between events. The point stochastic analysis of the daily rainfall occurrences was done by the input $\{X_i\}$. There are two options for the analysis: (i) the total time of observation is fixed and the total number of events is random, (ii) the total number of events is fixed and total time of observation is random. In this report the first option was used. A homogenization scheme was incorporated into the main program of the SASE IV by the authors. The other programs used in the point stochastic analysis were (i) ROCC, the computer program which computes the rate of occurrence function of the daily rainfalls, and (ii) NONLINR, Marquardt's nonlinear regression program for the calibration of the cyclicities and trends in the rainfall data, and for the least squares fits to the counts spectrum and the log-survivor function of the homogenized daily rainfall counts. A short description of the programs involved in the point stochastic analysis of the daily rainfall occurrences will be given below. This description follows the order in which the rainfall data was analyzed. For a detailed description and listing of the SASE IV program the reader is referred to *Lewis et al.* (1969).

B.1 PROGRAMS FOR THE IDENTIFICATION OF LONG-TERM TRENDS AND CYCLICITIES IN THE POINT STOCHASTIC PROCESS OF DAILY RAINFALL OCCURRENCES

The programs involved in the trend detection are the subroutines TREND, DENS, VART, BART in the SASE IV, and the program ROCC for the computation of the rate of occurrence.

B.1.1 TREND

This subroutine employes both the graphical analysis and the statistical tests for the detection of trends in the daily rainfall counts. For the graphical analysis it plots the total number of occurrences versus the time to the last occurrence. The slope of the plot at any time is the inverse of the mean rate of daily rainfall occurrence at that time.

For statistical test of trend in the rate of occurrence the program computes U where

$$U = \left(\frac{\sum_{i=1}^{n} t_i}{nt_n} - \frac{t_n}{2}\right) / \ t_n/\sqrt{12n} \ .$$

In this expression t_i is the time to the i-th rainfall occurrence, t_n is the time to the n-th rainfall occurrence, and n is the total number of daily rainfall occurrences in (o,T). The total time T is fixed to be 10 years. A positive U value means that the rate of occurrence is increasing with time. Since the rate of occurrence is the slope of the mean time function, positive U would mean that there is an upward trend in the mean time of the daily rainfall occurrences. Negative U would imply a downward trend.

Bartlett's variance homogeneity statistic is computed (see section 2.2.b in the text) to detect any second-order non-stationarity in the interarrival times $\{X_i\}$ of the daily rainfall occurrences.

B.1.2 ROCC

This program computes the mean rate of daily rainfall occurrences by the statistic $\lambda_\tau(t)$ where, for the interval $(t,t+\tau)$ $\lambda_\tau(t) = \frac{n(t,t+\tau)}{\tau}$. For further description, the reader is referred to Section 2.1.b in the tex

B.1.3 DENS

This subroutine computes the intensity function of the daily rainfall occurrences and gives its graph. The intensity function of the daily rainfall counts is estimated by $\bar{m}_f(r\alpha - \frac{1}{2}\alpha)$ for intervals of length α where

$$\bar{m}_f(r\alpha + \frac{1}{2}\alpha) = \frac{t_n}{t_n - \alpha(r+\frac{1}{2})} \cdot \frac{S_r}{n\alpha}, \quad r = 0, 1, \ldots$$

such that S_r is the sum of events which fall into $(r\alpha, r\alpha + \alpha)$ for each of the n different counting set-ups (see section 2.2.c). The reason that $\bar{m}_f(\tau)$ has its first value at $\tau = \frac{1}{2}\alpha$ is that the data was read in terms of the interarrival times $\{X_i\}$ so that the first rainfall occurs at time zero.

B.1.4 VART

This subroutine computes the variance-time function of the daily rainfall counts and gives its graph. The variance-time function V(t) is computed as a moving average over the possible intervals of length t. Assume that T is the total observation time and take $k = T/t$. The rainfall series can be divided into interval of length δ such that $\frac{t}{\delta} = J$. If the number of rainfall occurrences in the i-th interval of length δ is denotec by n_i, then n_i's in J consecutive blocks can be added to yield

$$\begin{aligned} S_1 &= n_1 + \ldots + n_J \\ S_2 &= n_2 + \ldots + n_{J+1} \\ &\vdots \\ S_{Jk-(J-1)} &= n_{Jk-(J-1)} + \ldots + n_{Jk} \end{aligned}$$

where $K = Jk - (J-1)$. For further details the reader is referred to *Cox and Lewis* (1966) or to *Lewis et al.* (1969).

B.1.5 BART

This subroutine computes the spectrum of the daily rainfall counts. Letting $\alpha = (t_n - t_1)/(n-1)$, the periodogram components A(J) and B(J) are computed as

$$A(J) = \sum_{i=2}^{n} \text{Cos } JB \frac{(t_i - t_1)}{\alpha},$$

$$B(J) = \sum_{i=2}^{n} \text{Sin } JB \frac{(t_i - t_1)}{\alpha},$$

where B is taken as $2\pi/(n-1)$. The periodogram estimates I(J) are

$$I(J) = \frac{2}{n-1} \{A(J)^2 + B(J)^2\}$$

and the relation between the frequency index J and the frequency ω is given as $J = \omega T/2\pi$. The periodogram estimates are averaged in consecutive, nonoverlapping groups to obtain the spectral estimates. The program gives the periodogram values I(J) and the estimated counts spectrum smoothed over groups of 5, 10 and 20 points. A simultaneous plot of these spectra is also given in the subroutine output. For further details the reader is referred to *Bartlett* (1963), *Cox and Lewis* (1966), *Lewis et al.* (1969), or to section 2.1.3 in the text.

B.2. HOMOGENIZATION SCHEME

This scheme rescales the time intervals between the daily rainfall occurrences according to the general equation

$$\Delta\tau = \exp\{\alpha_1 + \alpha_2 t + \alpha_3 t^2 + \sum_{i=1}^{r} R_i \text{ Sin } (\omega_i t + \theta_i)\}\Delta t.$$

In this equation $\Delta\tau$ is the rescaled time increment and Δt is the original time increment. R_i, ω_i and θ_i are respectively the amplitude, the frequency, and the phase angle of the i-th significant periodicity. The long-term trends are treated by the terms $\alpha_2 t$ and $\alpha_3 t^2$. This scheme can completely remove the trends only in the case of the non-homogeneous Poisson process. For further details the reader is referred to section III.1 in the text.

B.3 PROGRAMS FOR THE IDENTIFICATION OF THE PERSISTENCE STRUCTURE IN THE HOMOGENIZED DAILY RAINFALL COUNTS

In the report the subroutines VART, BART and INTER of the SASE IV program were used for the identification of the dependence structure in the daily rainfall counts. Subroutines VART and BART were already discussed.

B.3.1 INTER

This subroutine makes a graphical and numerical analysis of the marginal distribution of the inter-arrival times $\{X_i\}$. It also calculates and plots the log-survivor function which can be used for the detection of clustering of the rainfall events. This is discussed in section III.4.1.e. The log-survivor function is computed from the natural logarithm of $1 - F_n(x)$ where the distribution function of the inter-arrival times, $F_n(x)$, is estimated as

$$\begin{aligned} F_n(x) &= 0 \qquad x < x_{(1)} \\ &= \frac{i}{n} \qquad x_{(i-1)} \leq x \leq x_{(i)} \text{ , } i = 2, 3, \ldots, n \\ &= 1 \qquad x_{(n)} \leq x \end{aligned}$$

where $x_{(i)}$ is the i-th order statistic in the observed sample. For further details the reader is referred to *Cox and Lewis* (1966).

B.4 PROGRAMS FOR TESTING THE POISSON HYPOTHESIS

As was discussed in section 3.4.1, the variance-time curve in the subrouting VART, and the counts spectrum in the subroutine BART already test the Poisson hypthesis. More formal tests of the Poisson hypothesis are given in the subroutines EXPO and DURB in SASE IV.

B.4.1 EXPO

In this subroutine the Poisson hypothesis is tested by testing whether the quantities

$$Y_i = t_i/t_n \text{ , } i = 1, 2, \ldots, n$$

are uniformly distributed. This test is called the uniform conditional test. However, this test is the canonical form of the distribution-free tests of goodness-of-fit. The program computes the one-sided and the two-sided Kolmogorov-Smirnov statistics, and the Anderson-Darling statistic for the uniform conditional test. For details of the distribution-free tests the reader is referred to section 4.2.2.

EXPO also computes Moran's statistic for testing the hypothesis of independent, exponentially distributed interarrival times against the alternative of independent gamma distributed interarrival times (see section 4.3.2 for further details).

B.4.2 DURB

In this subroutine the interarrival times $x_1, x_2, \ldots, x_n, x_{n+1}$ where $x_{n+1} = T - t_n$, are ordered. The order statistics x_i' such that

$$0 < x'_{(1)} \leq x'_{(2)} \leq \cdots \leq x'_{(n)} \leq x'_{(n+1)}$$

are obtained. Then the quantities $\omega_{(i)}$ such that

$$\omega_{(i)} = \frac{x'_{(1)}}{T} + \frac{x'_{(2)}}{T} + \ldots + \frac{x'_{(i-1)}}{T} + (n+2-i)\,\frac{x'_{(i)}}{T},\ i = 1, \ldots, n$$

are obtained. As in the case of y_i in EXPO, $\omega_{(i)}$ are also uniformly distributed under the null Poisson hypothesis, and are in the canonical form of the distribution-free tests. Therefore, the Kolmogorov-Smirnov and Anderson-Darling statistics are computed for $\omega_{(i)}$ and the uniformity of $\omega_{(i)}$ are tested. For further details on this statistic the reader is referred to *Cox and Lewis* (1966).

B.5.1 RHO

This subroutine computes the autocorrelation function of the interarrival times from the formula

$$\hat{\rho}_J = \frac{n}{n-J}\,\frac{\sum_{i=1}^{n-J} (x_i-\bar{x})\,(x_{i+J} - \bar{x})}{\sum_{i=1}^{n} (x_i-\bar{x})^2}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. If n, the number of interarrival times in the data, is large, and provided that the interarrival time distribution is not highly skewed, under the null hypothesis $\rho_J=0$, $J = 1, 2, \ldots$, the $\hat{\rho}_J$ are $N(0, 1/\sqrt{n-J})$. RHO prints the quantities $\sqrt{n-J}\ \hat{\rho}_J$ as a function of J and plots $\hat{\rho}_J$ against J. This plot is obtained by a minor modification of the original RHO. The null hypothesis of $\rho_J=0$ against general alternatives is tested using the computed $\sqrt{n-J}\ \rho_J$, $J = 1, 2, \ldots$. Acceptance of $\rho_J = 0$ implies that the interarrival times are uncorrelated.

B.5.2 SPEC

This subroutine computes the spectral density function, the periodogram and the cumulative periodogram of the interarrival times and tests the renewal process hypothesis. The spectral density estimates are computed by

$$\hat{f}(\omega) = \frac{1}{\pi} \sum_{J=-\infty}^{+\infty} \lambda_J\, \rho_J\, \mathrm{Cos}\ (J\omega),\ -\pi \leq \omega \leq \pi$$

where Parzen's window is used for λ_J. For uncorrelated interarrival times the spectral estimates with the Parzen's window have the expectation and the variance

$$E[\hat{f}(\omega)] \sim \frac{1}{2\pi}\ ,\ \mathrm{Var}\ [\hat{f}(\omega)] \simeq \frac{f^2(\omega)\ \ell}{2n}\ .$$

where ℓ is the integer part of $(n-1)/2$.
The subroutine plots the spectral estimates so that one can check whether the spectral density function is around $1/2\pi$ for the case of uncorrelated intervals.

More formal tests are based on the periodogram estimates p(J) where p(J) is computed by SPEC in the form

$$p(J) = \frac{1}{2\pi n \ \mathrm{Var}\ (X)} \left| \sum_{\ell=1}^{n} x_\ell \ e^{-2\pi i(\ell-1)J/n} \right|^2 \quad J = 0, 1, \ldots, n-1$$

where x stands for the interarrival time. The formal tests of the interval independence have the null hypothesis H_0 that the periodogram estimates p(J) have asymptotically uncorrelated exponential distributions with mean Var $(X)/2\pi$. Then one can form a Poisson process from the p(J) where the waiting times t_k are defined as

$$t_k = \sum_{J=1}^{k} p(J).$$

SPEC calls the subroutines EXPO and TREND to test this Poisson process.

SPEC computes the normalized cumulative periodogram P(J) of the interarrival times as

$$P(k) = \sum_{J=1}^{k} p(J) / \sum_{J=1}^{\ell} p(J) \ , \ k = 1, \ldots, \ell$$

where ℓ is the integer part of $(n-1)/2$. Under the renewal hypothesis the normalized cumulative periodogram values P(k) become the order statistics from a uniform distribution in (0,1). Since this is the canonical form of the distribution-free tests, Kolmogorov-Smirnov and Anderson-Darling statistics are computed and the renewal hypothesis is tested.

For further details the reader is referred to section IV.2.2.

B.6 PROGRAMS FOR THE MARGINAL PROBABILITY DISTRIBUTION OF THE INTERVALS BETWEEN POINT RAINFALL OCCURRENCES

The non-parametric estimate of the marginal probability distribution of the interarrival times is done by the subroutine TREND. This subroutine was discussed earlier with respect to the computation of the log-survivor function in section B.3. TREND also computes the empirical frequency histogram, the first four sample moments of the intervals between events, and sample coefficients of variation, skewness and kurtosis. The frequency histogram $f_X(x_i)$ is calculated by

$$f_X(x_i) = \frac{\text{number of } x_i\text{'s}}{n+1} \ , \ i = 1, 2, \ldots$$

The estimated mean $\hat{\mu}$ and the estimated variance $\hat{\sigma}^2$ are computed from

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ and } \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

The coefficient of variation is $\hat{\sigma}/\hat{\mu}$, the third sample moment estimate $\hat{\mu}_3$, is

$$\hat{\mu}_3 = \frac{n}{(n-1)\ (n-2)} \sum_{i=1}^{n} \ (x_i - \hat{\mu})^3$$

The skewness coefficient is estimated by $\hat{\mu}_3/\hat{\sigma}^3$. The fourth sample moment μ_4 is estimated as

$$\hat{\mu}_4 = \frac{n(n^2-2n+3)}{(n-1)\ (n-2)\ (n-3)} \sum_{i=1}^{n} (x_i - \hat{\mu})^4 - \frac{3(n-1)\ (2n-3)}{(n-1)(n-2)(n-3)} \hat{\mu}_2^2$$

and the kurtosis coefficient estimate is $\hat{\mu}_4/(\hat{\mu}_2)^2 - 3$.

APPENDIX C: MARQUARDT'S ALGORITHM

Consider a general nonlinear model to be fitted to the data of the dependent variable Y with the independent variables $x_1, x_2, \ldots, x_n$ and with the parameters $\theta_1, \theta_2, \ldots, \theta_k$. The model is of the form,

$$Y = f(x_1, \ldots, x_m; \theta_1, \ldots, \theta_k) + \varepsilon \qquad (C.1)$$

where ε is the error component. The basic assumptions for the least squares method will be that $E(\varepsilon) = 0$, $Var(\varepsilon)$ = a constant, and that the errors ε are uncorrelated. Since the independent variables x_i, $i = 1, \ldots, m$ are not random, the equation $E(Y) = f(x_1, \ldots, x_m; \theta_1, \ldots, \theta_k)$ will be the equation to be used for the least squares method. When there are n observations, there will be n equations of the form

$$Y_i = f(x_{1i}; x_{2i}, \ldots, x_{mi}; \theta_1, \ldots, \theta_k) + \varepsilon_i \quad i = 1, \ldots, n . \qquad (C.2)$$

The least squares problem is the estimation of the parameters $\theta_1, \ldots, \theta_k$ which will minimize

$$S(\vec{\theta}) = \sum_{i=1}^{n} \{Y_i - f(\vec{x}_1; \vec{\theta})\}^2 \qquad (C.3)$$

where $\vec{x}_i = (x_{1i}, \ldots, x_{mi})$ and $\vec{\theta} = (\theta_1, \ldots, \theta_k)$. Eq. (C.3) is differentiated with respect to θ and the derivatives are set equal to zero to find the least squares estimate $\hat{\vec{\theta}}$. The equations thus obtained are called the normal equations and take the form

$$\sum_{i=1}^{n} \{Y_i - f(\vec{x}_i, \vec{\theta})\} \left(\frac{\partial f_i}{\partial \theta_j}\right)_{\theta=\hat{\theta}} = 0 \quad , \quad j = 1, \ldots, k . \qquad (C.4)$$

However, when the model is nonlinear in θ's, the normal equations will also be nonlinear. Therefore, $f(\vec{x}_i, \vec{\theta})$ should be linearized so as to yield a system of linear equations from which $\hat{\vec{\theta}}$ can be solved. The linearization can be done by assuming an estimate $\vec{\theta}_0$ for the parameter vector and then by expanding $f(\vec{x}_i, \vec{\theta})$ and $\vec{\theta}_0$. Thus

$$f(\vec{x}_i, \vec{\theta}) = f(\vec{x}_i, \vec{\theta}_0) + \sum_{j=1}^{k} \left(\frac{\partial f(\vec{x}_i, \vec{\theta})}{\partial \theta_j}\right)_{\theta_j=\theta_{0_j}} \cdot (\theta_j - \theta_{0_j}) \quad , \quad i = 1, \ldots, n \qquad (C.5)$$

and

$$\varepsilon_i = Y_i - f(\vec{x}_i, \vec{\theta}_0) - \sum_{j=1}^{k} \left(\frac{\partial f(\vec{x}_i, \vec{\theta})}{\partial \theta_j}\right)_{\theta_j=\theta_{0_j}} \cdot (\theta_j - \theta_{0_j}) \quad , \quad i = 1, \ldots, n \qquad (C.6)$$

Letting $\left(\frac{\partial f(\vec{x}_i, \vec{\theta})}{\partial \theta_j}\right)_{\theta_j=\theta_{0_j}} = r^0_{j,i}$, and $(\theta_j - \theta_{0_j}) = \delta^0_j$, eqs. (C.6) reduce to

$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} - \begin{bmatrix} f(\vec{x}_1, \vec{\theta}_0) \\ \vdots \\ f(\vec{x}_n, \vec{\theta}_0) \end{bmatrix} - \begin{bmatrix} r^0_{11} & r^0_{21} & \cdots & r^0_{k1} \\ \vdots & \vdots & \cdots & \vdots \\ r^0_{1n} & r^0_{2n} & \cdots & r^0_{kn} \end{bmatrix} \cdot \begin{bmatrix} \delta^0_1 \\ \vdots \\ \delta^0_k \end{bmatrix} \qquad (C.7)$$

Eq. (C.7) in vectorial form is

$$\vec{\varepsilon} = \vec{Y} - \vec{f}_0 - \vec{R}_0 \cdot \vec{\delta}_0 \, . \tag{C.8}$$

The residual sum of squares $S(\vec{\theta})$ becomes equal to $\vec{\varepsilon}^T\vec{\varepsilon}$. The minimization of $S(\vec{\theta})$ with respect to $\vec{\theta}$ yields the estimate of $\vec{\delta}_0$ as

$$\vec{\delta}_0 = (\vec{R}_0^T \, \vec{R}_0)^{-1} \, \vec{R}_0^T(\vec{Y} - \vec{f}_0) \, . \tag{C.9}$$

In other words $\vec{\delta}_0$ is the solution of the equation

$$A\vec{\delta}_0 = \vec{g} \tag{C.10}$$

for $A = \vec{R}_0^T \, \vec{R}_0$ and $\vec{g} = \vec{R}_0^T(\vec{Y} - \vec{f}_0)$. The estimate $\vec{\delta}_0$ which will minimize the sum of squares $S(\vec{\theta})$, is the correction vector for the estimate $\vec{\theta}_0$ of the population parameters $\vec{\theta}$. The revised estimates $\vec{\theta}_1$ are taken as

$$\vec{\theta}_1 = \vec{\theta}_0 + \vec{\delta}_0 \, . \tag{C.11}$$

The iterative process continues until the solution converges in $\vec{\theta}$. However, it is possible that this iterative least squares scheme may diverge (*Draper and Smith,* 1966). The $\vec{\delta}_j$ values may be too big and may cause oscillation, increasing and decreasing the sum of squares values in different iterations.

Theoretically, the gradient or the steepest descent method converges to a solution. In this method one moves from the current trial values of $\vec{\theta}_j$ to a new value of $\vec{\theta}_{j+1}$ in the direction of the negative gradient of $S(\vec{\theta})$, that is, along the vector $\vec{d}_g$ where

$$\vec{d}_g = -\left(\frac{\partial S(\vec{\theta})}{\partial \theta_1}, \frac{\partial S(\vec{\theta})}{\partial \theta_2}, \ldots, \frac{\partial S(\vec{\theta})}{\partial \theta_k}\right) \, . \tag{C.12}$$

The iteration is continued until a convergence in $\vec{\theta}$ is obtained. The disadvantage of this model is that it may converge extremely slowly after a rapid progress (*Draper and Smith,* 1966). *Marquardt* (1963) devised a least squares estimation scheme which combined the best features of the linearization and the steepest descent methods. Any correction vector should be within 90° of $\vec{d}_g$. Otherwise, $S(\vec{\theta})$ will locally get larger. *Marquardt* (1963) noticed that the direction of the correction vector $\vec{\delta}$ obtained by the linearization procedure has an angle γ with $\vec{d}_g$ where $80° < \gamma < 90°$. So he devised a method for interpolating between $\vec{\delta}$ and $\vec{\delta}_g$ and determining an acceptable step size simultaneously. In his scheme the matrix A and the vector $\vec{g}$ of eq. (C.10) are scaled as to have

$$a^*_{ij} = a_{ij}/\sqrt{a_{ij}}\, a_{jj} \text{ and } g^*_j = g_j/\sqrt{a_{jj}} \tag{C.13}$$

where a^*_{ij} is the i,j-th element of A*, a_{ii} is i,i-th element of A, g^*_j is the j-th element of $\vec{g}^*$, and g_j is the j-th element of $\vec{g}$. Then for the p-th iteration the equation

$$[A^*_{(p)} + \lambda_{(p)}I]\vec{\delta}^*_{(p)} = \vec{g}^*_{(p)} \tag{C.14}$$

is solved for $\vec{\delta}^*_{(p)}$. The correction vector $\vec{\delta}_{(p)}$ is obtained by rescaling $\vec{\delta}^*_{(p)}$ by

$$\delta_j = \delta^*_j/\sqrt{a_{jj}} \, , \; j = 1, \ldots, k \tag{C.15}$$

and the revised parameters for the new iteration are obtained as

$$\vec{\theta}_{p+1} = \vec{\theta}_p + \vec{\delta}_p \, . \tag{C.16}$$

Then the iteration is continued until convergence is achieved. $\lambda_{(p)}$ in eq. (C.14) is a positive constant which is selected as to satisfy the condition

$$S(\vec{\theta}_{p+1}) < S(\vec{\theta}_p) \ . \qquad \text{(C.17)}$$

It is shown by *Marquardt* (1963) that as $\lambda \to \infty$ the angle between $\vec{d}_g$ and $\vec{\delta}$ approaches to zero. Therefore, in order to satisfy (C.17) large values of λ should be used. However, λ is taken small whenever the linearization method converges nicely.

PART II

TIME SERIES ANALYSIS OF THE MONTHLY AND ANNUAL RAINFALL SEQUENCES IN THE MIDWESTERN UNITED STATES

CHAPTER 1 - INTRODUCTION

1.1 THE FRAMEWORK OF THE TIME SERIES ANALYSIS OF THE MONTHLY AND THE ANNUAL RAINFALL SEQUENCES

The second part of this report is concerned with the time series analysis of the monthly and the annual rainfall occurrences. The ARIMA (p,d,q) family of models is applied to the hydrologic data and the parametric approach of *Box and Jenkins* (1971) is used as the methodology. However, the spectral analysis is employed in the theoretical assessment of certain properties of the ARIMA (p,d,q) time series models.

1.2 A SURVEY OF THE TIME SERIES ANALYSIS OF THE MONTHLY AND ANNUAL RAINFALL SEQUENCES AND THE APPLICATION OF ARIMA MODELS TO HYDROLOGY

Roesner and Yevdjevich (1966) analyzed the monthly rainfall sequences at 219 stations over various regions in the United States. They detected the annual cyclicity and removed it through standardization. The standardized monthly rainfall time series were tested with the null hypothesis that they are a White Noise sequence. This hypothesis was rejected at the 5% level in 52 out of the 219 stations casting doubts as to its adequacy. *Kisisel and Delleur* (1971) analyzed the square root transformed and standardized monthly rainfall sequences at twelve (12) watersheds in Indiana. They suggested that these sequences can be approximately modeled by White Noise. They also found that the probability distributions of the normalized square root transformed monthly rainfall sequences were approximately normal. *Shahabian* (1973) did a spectral analysis of the square root transformed and standardized monthly rainfall data of the Lower Ohio Tributaries. He found a significant first lag correlation coefficient in some of the cases and used a first order autoregressive process to model these cases. However, since in these cases with the large lag one serial correlation coefficients, the variance explained by the first order autoregressive process was very low, *Shahabian* concluded that the rainfall sequences may be approximated by White Noise processes.

The application of the ARIMA approach of *Box and Jenkins* (1971) is quite recent in Hydrology. *Carlson, MacCormick, and Watts* (1970) applied the nonseasonal ARIMA approach to four annual flow series selected from a large world-wide sample. *O'Connell* (1971) applied the ARIMA (1,0,1) process as a possible model for the long range dependence in the hydrologic sequences. *Moss* (1972) fitted an ARIMA (1,0,1) model for the annual streamflow sequences. ARIMA models were applied to the water temperature and the river flow time series on the Ohio River by *McMichael and Hunter* (1972) for the forecasting purposes. In a very elaborate application of the ARIMA models to monthly and yearly runoff sequences *McKerchar and Delleur* (1972) developed computer programs for a systematic analysis of the hydrologic time series. They successfully applied the nonseasonal and seasonal ARIMA models to monthly and annual runoffs in the Lower Ohio River Tributaries.

1.3 THE DATA

The watersheds selected for the time series analysis of the monthly and the annual data corresponded to those whose runoff stations were analyzed by *McKerchar and Delleur* (1972). The geographical region envelopes the midwestern states of Indiana, Illinois, Ohio and Kentucky. The region covers the tributaries of the lower Ohio river and is shown on the MAP-2.

The monthly rainfall data in each watershed was formed by the Thiessen polygon method from several point monthly rainfall data of the various stations at that particular watershed.

The annual data were formed by the summation of the corresponding monthly data.

Fifteen watersheds, given in Table 1-1, were used for the time series analysis of the monthly and the annual rainfall. Watershed areas ranged from 243 to 3955 square miles. The record lengths ranged from 468 to 684 months.

TABLE 1-1

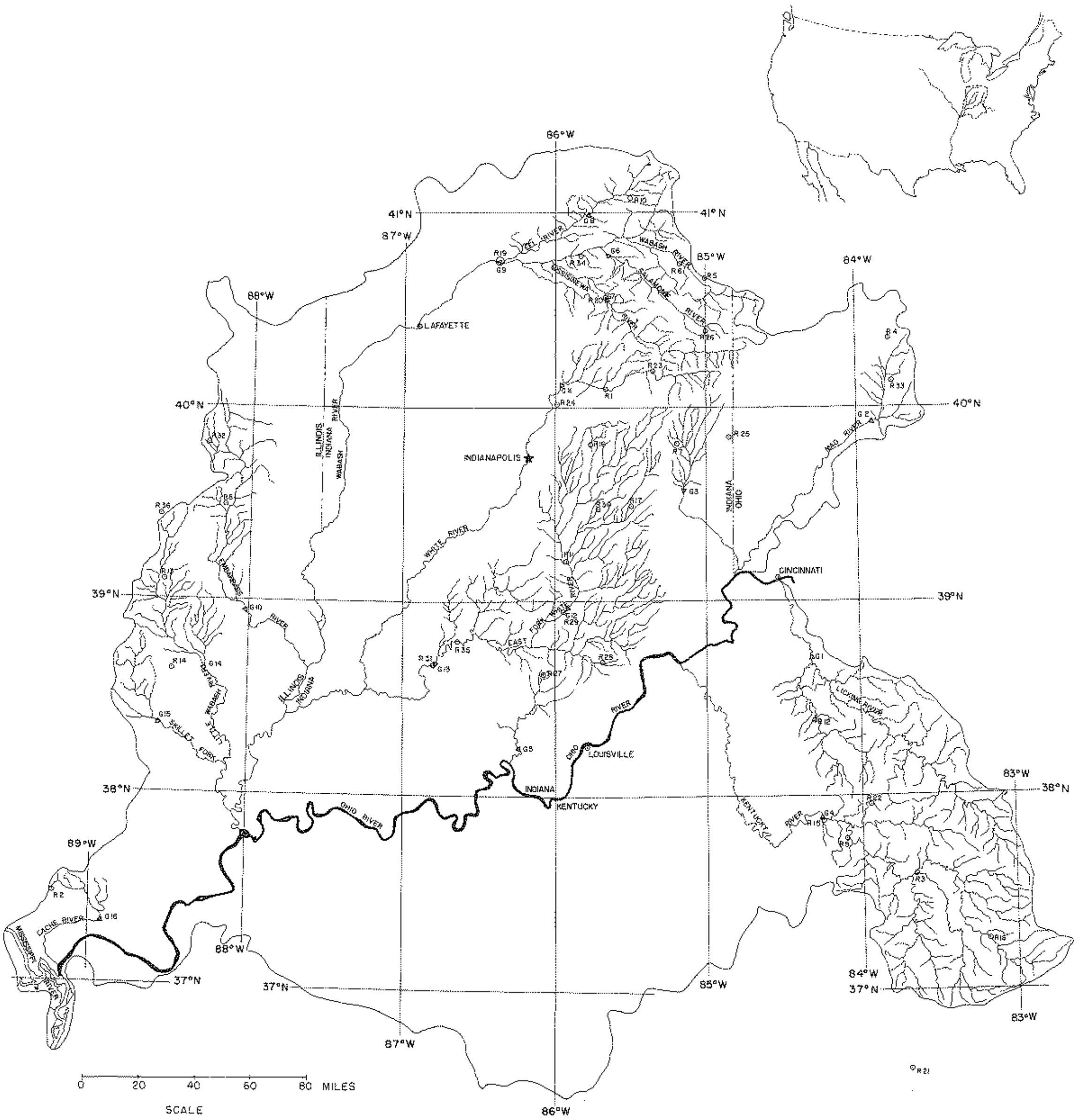
THE WATERSHEDS IN THE MIDWEST USED FOR THE ANALYSIS OF THE MONTHLY AND THE ANNUAL RAINFALL SEQUENCES

U.S.G.S. Identification No.	Identification Name	Drainage Area sq. mi.	Record length (months)
2535	Licking River at Catawaba, KY	3300	516
2695	Mad River near Springfield, OH	490	671
2750	Whitewater River near Alpine, IN	539	492
2840	Kentucky River at Lock 10 near Winchester, KY	3955	684
3030	Blue River near White Cloud, IN	461	468
3245	Salamonia River at Dora, IN	553	492
3265	Mississinewa River at Marion, IN	677	528
3280	Eel River at North Manchester, IN	416	576
3290	Wabash River at Logansport, IN	3751	528
3445	Embarras River at St. Marie, IN	1513	550
3485	White River near Noblesville, IN	814	492
3655	East Fork White River at Seymour, IN	2333	504
3795	Little Wabash River below Clay City, ILL	1134	684
3805	Skillet Fork at Wayne City, ILL	464	684
6120	Cache River at Forman, ILL	243	576

1.4 OBJECTIVES OF THE PART II

In the general introduction to parts I and II the place of the time series analysis of the monthly and the annual rainfall data in the water resources developments was discussed. In the light of that discussion, the objectives of the second part of this report may be stated as follows:

1. To study the various methods for the removal of periodicities in the monthly rainfall series to assess their properties.
2. To study the long range properties of the hydrologic time series models for the monthly rainfall sequences to determine their advantages and shortcomings in preserving the long range dependence.
3. To apply the nonseasonal and seasonal ARIMA models to the monthly and to the annual rainfall series to find the most suitable model for the generation and forecasting of monthly and annual rainfall sequences in the Midwestern United States.



MAP 2 LOCATION OF WATERSHEDS AND GAGING STATIONS

CHAPTER 2 - GENERAL THEORETICAL CONSIDERATIONS ON THE HYDROLOGIC TIME SERIES

In this chapter the theoretical problems of stationarity and the invertibility of the hydrologic time series models, the removal of cyclicities from the hydrologic time series, and the long range dependence behavior of the ARMA (p,q) family of the hydrologic stochastic models will be studied.

2.1 A DISCUSSION ON THE STATIONARITY OF THE HYDROLOGIC TIME SERIES MODELS

In stochastic hydrology the stationarity of the hydrologic time series is an important concept. Once the series are stationarized the stationary time series models can be fitted to these hydrologic time series.

An important concept to clarify is the physical meaning of the stationarity requirements on the hydrologic time series models. In many hydrologic studies it was seen that in order to preserve certain hydrologic characteristics the model parameters had to approach the stationarity boundaries. For example, in order to preserve the long range dependence observed in the hydrologic time series the ARMA (1,1) model parameters had to lie very close to their stationarity limits (*O'Connell*, 1971). The physical meaning of lying beside or on the stationarity boundary is yet not clarified. It will be the purpose of this section to clarify the physical meaning of the stationarity on the AR parameters and invertibility on the MA parameters of a general ARMA (p,q) model where it will be assumed that p > q.

A stochastic process is strictly stationary if its probability distribution function is invariant to shifts in time. In equation form, y_t, $t\varepsilon\tau$ is strictly stationary if and only if

$$F_{y_{t_1},\ldots,y_{t_n}}(y_1,\ldots,y_n) = F_{y_{t_1+u},\ldots,t_n+u}(y_1,\ldots,y_n)$$

for every n = 1, 2, ...; $t_1, \ldots, t_n\varepsilon\tau$; and $y_1, \ldots, y_n$.

A second order stationary (s.o.s.) process is one where the first two moments of the probability distribution of the process are invariant to time shift. This sense of stationarity is the most widely used concept in stochastic hydrology where many hydrologic time series are transformed so as to obey the normal probability law. Conformity with the second order stationarity means conformity with strict stationarity in the case of normal processes. In mathematical form y_t, $t\varepsilon\tau$ is s.o.s. if

1. $E[y_t] = \mu$, a finite constant,
2. $E[y_t^2] < \infty$,
3. $E[(y_t - \mu)(y_{t+k} - \mu)] = Cov(k)$; that is, the covariance function depends only on the lag k.

The ARMA (p,q) model may be defined by the relationship

$$\left(\sum_{J=0}^{p} \phi_J B^J\right) y_t = \left(\sum_{J=0}^{q} \theta_J B^J\right) a_t \tag{2.1}$$

where B is the backward shift operator. For further discussion it will be assumed that p > q. The first condition to be met for stationarity is that $E[y_t] = \mu$, a finite constant. Taking the expectation on both sides of the above equation,

$$\left(\sum_{J=0}^{p} \phi_J B^J\right) E[y_t] = \left(\sum_{J=0}^{q} \theta_J B^J\right) E(a_t). \tag{2.2}$$

Since the random shocks have zero mean it follows that

$$\left(\sum_{J=0}^{p} \phi_J B^J\right) E[y_t] = 0 \tag{2.3}$$

so that $E[y_t] = 0$

Thus the first condition for stationarity is that the variance of the process should be finite. The variance of the ARMA(p,q) model can be represented in terms of its spectrum by

$$\mathrm{Var}(y_t) = \sigma_a^2 \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{\left|\sum_{\alpha=0}^{q} \theta_\alpha e^{-i2\pi\omega\delta\alpha}\right|^2}{\left|\sum_{\beta=0}^{p} \phi_\beta e^{-i2\pi\omega\delta\beta}\right|^2} d\omega \tag{2.4}$$

where σ_a^2 is the variance of a_t. As for complex numbers

$$\frac{|A|}{|B|} = \left|\frac{A}{B}\right|,$$

for Var (y_t) to be finite one should have

$$\left|\sum_{\alpha=0}^{q} \theta_\alpha e^{-2\pi\omega\delta\alpha i} \Big/ \sum_{\beta=0}^{p} \phi_\beta \, e^{-i2\pi\omega\delta\beta}\right| < \infty. \tag{2.5}$$

The numerator is finite when θ_α, $\alpha = 0, \ldots, p$, are finite. Therefore, the finiteness of the variance requires that

$$1 \Big/ \sum_{\beta=0}^{q} \phi_\beta \, e^{-i2\pi\omega\delta\beta} \tag{2.6}$$

is a converging series. The polynomial in the denominator of (2.6) can be factored as

$$\phi(e^{-i2\pi\omega\delta}) = \phi_0(1 - c_1 e^{-i2\rho\omega\delta})(1 - c_2 e^{-i2\pi\omega\delta}) \ldots (1 - c_p e^{-i2\pi\omega\delta}),$$

or

$$\phi(e^{-i2\pi\omega\delta}) = \phi_0 \prod_{j=1}^{k} (1 - c_J e^{-i2\pi\omega\delta})^{m_J} \tag{2.7}$$

where m_J, $J = 1, \ldots, k$, are the multiplicities of the polynomial $\phi(e^{-i2\pi\omega\delta})$ which is assumed to have k distinct roots. The reciprocal of $\phi(e^{-i2\pi\omega\delta})$ can be expanded in partial fractions as

$$\frac{1}{\phi_0 \prod_{J=1}^{k} (1 - c_J e^{-i2\pi\omega\delta})^{m_J}} = \frac{1}{\phi_0} \sum_{J=1}^{k} \sum_{r=1}^{m_J} \frac{\alpha_{r_J}}{(1 - c_J e^{-i2\pi\omega\delta})^r} \quad \text{for} \quad q < p \tag{2.8}$$

where α_{r_J} are the coefficients in the partial fraction expansion. For the summation in (2.8) to converge each of the power series $1/(1 - c_J e^{i2\pi\omega\delta})^r$ must converge. This is possible for $|c_J e^{-i2\pi\omega\delta}| < 1$. However, as $|c_J e^{-i2\pi\omega\delta}| = |c_J|$, it is necessary that

$$|c_J| < 1 \tag{2.9}$$

for the series in (2.6) to converge. Since the roots of $\phi(e^{-i2\pi\omega\delta})$ are of the form $1/c_J$, $J = 1, \ldots, p$, they should all be outside the unit circle. Therefore, stationarity of the ARMA (p,q) model depends on the roots of $\phi(B) = 0$ to be outside the unit circle. Therefore, stationarity imposes restrictions only on the autoregressive coefficients.

Consider the case when $|c_J| > 1$ for any $J = 1, \ldots, k$, that is, when any root is inside the unit circle. Take the ARMA (1,0) model which has just this non-invertible root. This model may be written as

$$y_t = \frac{1}{\phi_0(1 - c_J B)} a_t, \qquad |c_J| \geq 1 \tag{2.10}$$

The model is, in conventional form, $(1 - c_J B)\, y_t = a_t$ when $\phi_0 = 1$. Thus, in conventional form

$$y_t = c_J \, y_{t-1} + a_t$$

$$y_{t+1} = c_J^2 \, y_{t-1} + c_J \, a_t + a_{t+1}$$

$$y_{t+2} = c_J^3 \, y_{t-1} + c_J^2 \, a_t + c_J \, a_{t+1} + a_{t+2}, \text{ etc.}$$

As is seen from this scheme, when $|c_J| > 1$, the effect of the past on the present value of the time series increases as the series move into the future. When $|c_J| = 1$ the effect of the past on the present value stays the same no matter how far into the future the series have moved. Both of the above cases contradict the hydrologic fact that the effect of the past on the present and future decreases as the series move to the future.

Although there were no restrictions on the Moving Average operator $\theta(B)$ of the ARMA (p,q) model under the stationarity conditions, the invertibility of $\theta(B)$ is required to assure hydrologic realizability. The invertibility of $\theta(B)$ is that the roots of $\theta(B) = 0$ should be outside the unit circle. The polynomial $\theta(B)$ may be written as

$$\theta(B) = \theta_0 \prod_{J=1}^{k} (1 - c_J B)^{m_J} \tag{2.11}$$

where m_J, $J = 1, \ldots, k$, are the multiplicities of the k distinct roots in $\theta(B)$. It follows from (2.11) and the previous discussion that $|c_J| < 1$, $J = 1, \ldots, k$, for invertibility of $\theta(B)$. Let one of the non-multiple roots be inside the unit circle. Take $|c_J| > 1$ for $J = \ell$ where $1 \leq \ell \leq k$.

$$\theta(B) = (1 - c_\ell B) \; \theta_0 \prod_{\substack{J=1 \\ J \neq \ell}}^{k} (1 - c_J B)^{m_J}$$

$$= (1 - c_\ell B) \; \theta_1(B)$$

The ARMA (p,q) model may be written as

$$\phi(B) \; y_t = \theta(B) \; a_t$$

Let $\phi(B)$ be invertible. Thus y_t can be written as

$$\frac{1}{1 - c_\ell B} \, y_t = \frac{\theta_1(B)}{\phi(B)} \; a_t \tag{2.12}$$

and

$$\frac{1}{1 - c_\ell B} \, y_t = (1 + c_\ell B + c^2 \, B^2 + \ldots) \; y_t$$

which is a divergent series for $|c_\ell| > 1$. Consider

$$\frac{1}{1 - c_\ell B} = - \frac{1}{c_\ell B} \cdot \frac{1}{1 - 1/c_\ell B}$$

$$= - \frac{1}{c_\ell B} - \frac{1}{c_\ell^2 B^2} - \frac{1}{c_\ell^3 B^3}$$

This series converge for $|c_\ell| > 1$. Thus

$$\frac{1}{1 - c_\ell B} \, y_t = \left(- \frac{1}{c_\ell B} - \frac{1}{c_\ell^3 B^3} - \cdots \right) \, y_t$$

$$= - \frac{1}{c_\ell} \, y_{t+1} - \frac{1}{c_\ell^2} \, y_{t+2} - \frac{1}{c_\ell^3} \, y_{t+3} - \cdots \tag{2.13}$$

converges. (2.13) says that for a root of $\theta(B)$ inside the unit circle the future values are used to generate the present value of y_t. This is not a realizable generation. Therefore, to have a hydrologically realizable generation all the roots of $\theta(B) = 0$ should be outside the unit circle.

2.2 PERIODICITIES IN THE MONTHLY HYDROLOGIC TIME SERIES

In this section the periodicities in the monthly hydrologic time series will be discussed. A spectral model for the periodic hydrologic time series will be introduced. Through the use of this spectral model the methods of differencing and monthly mean subtraction will be analyzed. The simulation of the hydrologic time series by the nonseasonal and the seasonal ARIMA models will be discussed.

2.2.1 Introduction

Due to the rotation of the earth around the sun, there is a yearly periodicity in the monthly hydrologic time series. This periodicity is manifested in the autocorrelation function which has the appearance of a sinusoidal function with a 12 month period and in the spectral density function which exhibits a discrete spectral component of the frequency 1/12 cycle per month. This periodicity is seen in Fig. 1 for the case of rescaled monthly rainfall data. The monthly hydrologic time series were rescaled by dividing each term of the series by the monthly standard deviation of the respective month to yield a constant variance throughout each series. This rescaling was done before the estimation of the sample spectrum and the fitting of the corresponding spectral model. It is also assumed that the non-stationarities due to long term time trends are removed before any operation. The following discussion applies to the rescaled trend-free time series.

The time series models currently used in hydrology are fitted to the stationary random component of the spectrum (or equivalently to the decaying part of the autocorrelation function). The hydrologist is thus faced with the problem of the removal of the circularly stationary component of the time series. This time series component corresponds to the discrete spectral component in the spectrum or to the sinusoidal periodic component in the autocorrelation function. The periodicity in the autocovariance is a function of the lag. The variance, being the value of the autocovariance function at lag zero, is not a function of the lag and is not subject to this periodicity.

The spectral model to be constructed and used in the following is strictly within the realm of the correlation theory stated by *Yaglom* (1962) and others. Therefore, our interest will be limited to the second-order stationarity which is obtained by a random function, $X(t)$, if it has a constant mean and if its autocorrelation function, $A(\tau_1, \tau_2)$, depends only on the lag difference $\tau_2 - \tau_1$.

It is worthwhile to emphasize that the periodic hydrologic time series considered here are not nonstationary. What is thought as the nonstationarity due to the yearly periodicity is really the circular stationarity as was developed by *Yaglom* (1962), *Hannan* (1960), *Cote* (1973) and others. The definition of circular stationarity can be stated for the monthly hydrologic time series as follows: the monthly hydrologic time series, $\vec{X}_0$, is circularly stationary if the multivariate probability distribution of $\vec{X}_0 = (x_{i_1}, x_{i_2}, \ldots, x_{i_{12}})$ is the same as the multivariate probability distribution of $\vec{x}_{0+12k} = (x_{i_1+12k}, \ldots, x_{i_{12}+12k})$ for $k = 1, 2, \ldots$. Therefore, the circular stationarity suggests that the probability distribution of the hydrologic quantity in a particular month is the same for the different years.

As a physical interpretation, the periodic hydrologic time series with a constant variance (i.e. rescaled) may be considered to be made up of two components: (a) the monthly means, comprising the circularly stationary component and (b) the deviations from the monthly means comprising the stationary random component. These two components correspond to the discrete and continuous parts of the spectrum of the series. Each part has its own spectral representation. A periodic time series $\{X_J\}$, comprised of a circularly stationary component and of a stationary random component, sampled at equispaced intervals, has the following spectral representation (see *Yaglom* (1962) or *Hannan* (1960)):

$$X_J = \sqrt{\alpha}\int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sqrt{f(\omega)}\, e^{i2\pi\delta J\omega}\, dZ(\omega) + \sqrt{\delta}\sum_{\alpha=0}^{r} \sqrt{\ell_\alpha Q_\alpha}\, \mathrm{Cos}(2\pi J\delta\omega_\alpha + \Phi_\alpha) \qquad (2.14a)$$

$$J = \ldots -1, 0, 1, \ldots$$

where δ = sampling time interval

ω = frequency, $-\frac{1}{2} \le \omega \le \frac{1}{2}$

$f(\omega)$ = absolutely continuous component of the spectral density function

$dZ(\omega)$ = stationary, uncorrelated, complex random increments with $E[Z(\omega)] = 0$

r = half period of the circularly stationary time series of period length $2r+1$

ℓ_α = the magnitude of the α-th discontinuity in the spectral distribution function

Q_α = random variables mutually uncorrelated for $\alpha = 0, \ldots, r$

ω_α = discrete frequency which is equal to $\alpha/(2r+1)$

Φ_α = random phase mutually uncorrelated for $\alpha = 0, \ldots, r$.

The first term on the right side of (2.14a) represents the stationary random component and is written in complex form since this is very concise. However, if this model is to be used for generation purposes the real form of the stationary random component can be written as (*Yaglom*, 1962)

$$X_{c,J} = \sqrt{\delta}\int_0^{\frac{1}{2}\delta} \sqrt{f(\omega)}\, \mathrm{Cos}(2\pi J\omega\delta)dZ_1(\omega) + \sqrt{\delta}\int_0^{\frac{1}{2}\delta} \sqrt{f(\omega)}\, \mathrm{Sin}(2\pi J\omega\delta)dZ_2(\omega) \qquad (2.14b)$$

where $Z_1(\omega)$ and $Z_2(\omega)$ are real random functions of the frequency ω with uncorrelated increments. However, the representation (2.14a) will be used since it is a more concise form and the physical behavior of the spectral density function is still clear under the various operations to be considered.

The stationary random component of the monthly hydrologic series can be interpreted as a process built up by elementary and mutually orthogonal oscillations with random amplitudes and random phases (Cramer and Leadbetter, 1967). So the stationary random deviations from the monthly means of the monthly hydrologic series can be modeled as a summation of spectral components, that is, an infinite number of uncorrelated harmonics with random phases and amplitudes.

The second term in the right-hand side of (2.14a) represents the circularly stationary (periodic) component of the time series. The seasonality of the monthly hydrologic time series, treated as a circularly stationary time series and assuming that there is a significant contribution at every multiple of the fundamental yearly frequency 1/12, is represented by (*Yaglom*, 1962; *Hannan*, 1960; *Cote*, 1973)

$$X_{d,J} = \sqrt{\delta}\sum_{\alpha=0}^{6} \sqrt{\ell_\alpha Q_\alpha}\, \mathrm{Cos}(2\pi J\delta\omega_\alpha + \Phi_\alpha), \qquad J = \ldots -1, 0, 1, \ldots$$

where $\omega_\alpha = \alpha/12\delta$. This representation has the physical interpretation that the circularly stationary seasonality of the hydrologic time series can be modeled as the superposition of seven mutually uncorrelated oscillations of different frequency with random amplitudes and phases. The physical reason for considering all the harmonics that are multiples of the fundamental yearly frequency 1/12 is that the circularity in nature may not follow an ideal cosine function with a yearly period. Irregularities or deviations from the shape of the cosine function with the yearly period will be seen as leakages in the frequencies that are multiples of 1/12 and can be easily modeled by considering the general superposition of all the possible harmonics for the case considered.

2.2.2 Nonseasonal Differencing

First, the nonseasonal first lag differencing of the monthly hydrologic series for the purpose of the removal of the yearly periodicity will be investigated. Differencing the original series $\{X_j\}$ once corresponds to obtaining the series ΔX_j such that

$$\Delta X_j = X_j - X_{j-1}$$

The spectral representation of the first differenced series, obtained by applying operation (2.16) to the spectral representation (2.14a) is:

$$\Delta X_j = \sqrt{\delta} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sqrt{f(\omega)}(e^{i2\pi j\delta\omega} - e^{i2\pi(j-1)\delta\omega})dZ(\omega) + \sqrt{\delta} \sum_{\alpha=0}^{6} \sqrt{\ell_\alpha Q_\alpha} \left[\mathrm{Cos}\left(2\pi j \frac{\alpha}{12} + \Phi_\alpha\right) - \mathrm{Cos}\left(2\pi j \frac{\alpha}{12} + \Phi_\alpha - 2\frac{\alpha\pi}{12}\right)\right]$$

or

$$\Delta X_j = \sqrt{\delta} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sqrt{f(\omega)}(2\ \mathrm{Sin}\ \pi\delta\omega)\ e^{i2\pi j\delta\omega}\ dZ(\omega) + \sqrt{\delta} \sum_{\alpha=0}^{6} \sqrt{\ell_\alpha Q_\alpha} \left[\mathrm{Cos}\left(2\pi j \frac{\alpha}{12} + \Phi_\alpha\right) - \mathrm{Cos}\left(2\pi \frac{\alpha}{12}(j-1) + \Phi_\alpha\right)\right] \quad (2.17)$$

The second term in the right-hand side of equation (2.17) shows that first differencing does not completely eliminate the periodic component although it significantly reduces it. The continuous spectral part of the differenced series is

$$f_\Delta(\omega) = f(\omega)\ (2\ \mathrm{Sin}\ \pi\delta\omega)^2, \qquad -\tfrac{1}{2}\delta < \omega < \tfrac{1}{2}\delta. \quad (2.18)$$

As is seen in equation (2.18), the original continuous spectral density $f(\omega)$ is distorted to yield $f_\Delta(\omega)$. Differencing wipes out the value of the original spectral density $f(\omega)$ at $\omega = 0$ and dampens it for $0 < |\omega| < 1/6\delta$ while it amplifies $f(\omega)$ for $1/6\delta \leq |\omega| \leq 1/2\delta$, introducing spurious high frequencies. From the hydrologic point of view differencing wipes out the long run properties while introducing spurious short run properties by magnifying the high frequency contributions.

The covariances r_j and R_j of the stationary random components of the hydrologic series and of the differenced series, respectively, for the j-th lag are related by:

$$R_j = -\Delta^2 r_{j+1} \quad (2.19)$$

and

$$R_0 = 2(r_0 - r_1). \quad (2.20)$$

Equations (2.19) and (2.20) show that the covariance of the stationary random part of the differenced series is distorted and that its variance depends on the magnitude of the variance and first lag covariance of the stationary random part of the original periodic hydrologic series.

2.2.3 Seasonal Differencing

The monthly hydrologic time series is differenced with a seasonality of 12 months using the operation

$$\Delta_{12}\ X_j = X_j - X_{j-12} \quad (2.21)$$

The spectral representation of the seasonally differenced series, obtained by applying the operation (2.21) to the spectral representation (2.14a), is:

$$\Delta_{12}\ X_j = \sqrt{\delta} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sqrt{f(\omega)}\ (e^{i2\pi\delta j\omega} - e^{i2\pi(j-12)\delta\omega})\ dZ(\omega) + \sqrt{\delta} \sum_{\alpha=0}^{6} \sqrt{\ell_\alpha Q_\alpha}\ \mathrm{Cos}(2\pi j\alpha/12 + \Phi_\alpha)$$

$$- \mathrm{Cos}(2\pi j\alpha/12 + \Phi_\alpha - 2\pi\delta 12\alpha/12\delta)$$

or

$$\Delta_{12}\ X_j = \sqrt{\delta} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} (2\ \mathrm{Sin}\ 12\rho\omega\delta)\ \sqrt{f(\omega)}\ e^{i2\rho\omega j\delta}\ dZ(\omega) \quad (2.22)$$

As is seen from expression (2.22) the periodic contribution is completely removed and the original continuous spectrum is distorted. That is, the spectral density $f_{\Delta 12}(\omega)$ of the stationary random series $\{\Delta_{12}\ X_j\}$ is

$$f_{\Delta 12}(\omega) = (2 \text{ Sin } 12\pi\omega\delta)^2 \ f(\omega). \tag{2.23}$$

Therefore, $f(\omega)$ is wiped out at $|\omega| = k/12\delta$, $k = 0, 1, \ldots, 6$ and amplified in between these frequencies. The spectral density $f_{\Delta 12}(\omega)$ obtained after 12 lag differencing has a sinusoidal shape which cannot be fitted by the conventional ARMA (p,q) family of mathematical models.

The covariances, r_j and R_j^{12}, for the j-th lag of the stationary random component of the hydrologic series and of the seasonally 12 lag differenced series, respectively, can be shown to be related by

$$R_j^{12} = -\Delta_{12}^2 r_{j+12}$$

$$R_0^{12} = 2[r_0 - r_{12}].$$

Therefore, the covariance function of the stationary part is distroted while the periodic part of the covariance function is removed since the discrete spectral part is removed.

2.2.4 Model Fitting and Simulation

As the nonseasonal first lag differencing of the monthly periodic hydrologic series significantly reduces the periodic component, a stationary ARMA (p,q) model can be fitted to the differenced hydrologic series ΔX_t. The general form of the model is (*Box and Jenkins*, 1971):

$$\phi(B)\ X_t = \theta(B)\ a_t \tag{2.25}$$

where B is the backward shift operator, thus $BX_t = X_{t-1}$ and $\Delta = 1\text{-}B$, $\phi(B)$ and $\theta(B)$ are polynomials in B of order p and q respectively, and a_t is a random variable. The ARIMA (p,1,q) process, represented by expression (2.25), can be defined as the stationary ARMA (p,q) model of the first differenced time series. The stationarity and invertibility conditions are that the roots of $\phi(B) = 0$ and $\theta(B) = 0$ should lie outside the unit circle. The spectral density of ARMA (p,q) process is (*Jenkins and Watts*, 1968):

$$f_{ARMA}(\omega) = \left|\sum_{\alpha=0}^{q} \theta_\alpha\ e^{-i2\pi\omega\delta\alpha}\right|^2 \Big/ \left|\sum_{\beta=0}^{p} \phi_\beta\ e^{-i2\pi\omega\delta\beta}\right|^2 \ \frac{\sigma_a^2}{R_0} \tag{2.26}$$

where R_0 is the variance of the process and σ_a^2 is the variance of the white noise input.

First differencing distorts the continuous spectrum of the original series X_j as shown in expression (2.18) and it is to this distorted spectral density, $f_\Delta(\omega)$, that the ARMA (p,q) model is to be fitted. The spectral density $f_\Delta(\omega)$ definitely does not represent the sample characteristics of the stationary component of monthly periodic hydrologic series. Therefore, the ARMA (p,q) model fitted to the autocovariance structure corresponding to $f_\Delta(\omega)$ is of no practical value for generation purposes since it really does not preserve the original spectral density $f(\omega)$ of the stationary random component of the periodic hydrologic series.

If an ARIMA (p,1,q) model is intended to be used for generation purposes, the transformation

$$f_c(\omega) = f_\Delta(\omega)/(2 \text{ Sin } \pi\omega\delta)^2 \tag{2.27}$$

where $R_0\ f_c(\omega)$ is the continuous spectrum of the generated series, has to be performed on the spectral density $f_\Delta(\omega)$ of the differenced series to retrieve the sample spectral properties of the stationary part of periodic hydrologic series. In the time domain the above operation corresponds to summing the differenced series ΔX_j; thus

$$X_j = \sum_{r=0}^{\infty} \frac{\theta(B)}{\phi(B)} a_{j-r} = \frac{\theta(B)}{\phi(B)} \sum_{r=0}^{\infty} a_{j-r}$$

where the right side of the latter expression is a divergent infinite series.

Combining (2.26) and (2.27) the spectral representation

$$X_{c,j} = \sqrt{\delta} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \left[\sqrt{\frac{\left|\sum\limits_{\alpha=0}^{q} \theta_\alpha \, e^{-i2\pi\omega\delta\alpha}\right|^2 \sigma_a^2}{\left|\sum\limits_{\beta=0}^{p} \phi_\beta \, e^{-i2\pi\omega\delta\beta}\right|^2 R_0}} \Bigg/ (2 \text{ Sin } \pi\omega\delta) \right] e^{i2\pi\omega\delta j} \, dZ(\omega) \tag{2.28}$$

is obtained for an ARMA (p,q) model fitted to the first difference of the periodic hydrologic series X_j. The variance of the $X_{c,j}$ generated by the above scheme is

$$\text{VAR}(X_{c,j}) = \delta \, R_0 \, \sigma_a^2 \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{f_{ARMA}(\omega)}{4 \text{ Sin}^2 \, \pi\omega\delta} \, d\omega \tag{2.29}$$

The spectral density at the origin for the ARMA (p,q) model is

$$f_{ARMA}(0) = \frac{\left(\sum\limits_{\alpha=0}^{q} \theta_\alpha\right)^2}{\left(\sum\limits_{\beta=0}^{p} \phi_\beta\right)^2} \cdot \frac{\sigma_a^2}{R_0} \tag{2.30}$$

If there is a perfect fit of the ARMA (p,q) model to the distorted spectral density $f_\Delta(\omega)$ of the differenced hydrologic series, then

$$f_{ARMA}(0) = f_\Delta(0) = 0. \tag{2.31}$$

The generation in the time domain would correspond to dividing the spectrum of the fitted ARMA (p,q) by $(2 \text{ Sin } \rho\omega\delta)^2$ in the spectral domain. Then the spectrum of the generated series would have the value $\frac{0}{0}$ at the spectral origin and would be undefined. Then the generation scheme with the ARIMA (p,1,q) model of perfect fit would be undefined. However, in general, the ARMA (p,q) models have few parameters and not enough degrees of freedom to render an exact fit. The spectral density of an ARMA (1,1) model at the origin is $f_{ARMA}(0) = (\sigma_a^2/R_0)(1 - \theta)^2/(1 - \phi)^2 = k$ which has a positive value when the stationarity and the invertibility conditions are satisfied. As in the region $(-\varepsilon, \varepsilon)$ around the spectral origin $\lim\limits_{\omega\to 0} \text{Sin } \pi\omega\delta = \pi\omega\delta$, the spectral density $f_c(\omega)$ of the generation scheme becomes

$$f_c(\omega) = \frac{k}{\pi^2\delta^2} \, \omega^{-2} \tag{2.32}$$

But, as the area under $1/\omega^2$ in the region $(0, \varepsilon)$ is infinite, $\text{Var}(X_{c,j}) = \infty$ and the conditions for the weak stationarity are not satisfied. Therefore, the ARIMA (p,1,q) model yields generation schemes with infinite variance when it does not satisfy the condition of zero spectral value at the origin.

It was shown earlier that the 12 lag seasonal differencing removes the periodic component of the hydrologic time series while distorting the continuous component of its spectrum. Expression (2.23) for $f_{\Delta 12}(\omega)$ implies that there is a 12-lag correlation in the series as well as a first lag type correlation. Seasonal ARIMA (P,1,Q) × (p,d,q) for the 12 month seasonality of the monthly hydrologic series can be written as (*Box and Jenkins*, 1971),

$$\phi_p(B) \, \Phi_P(B^{12}) \Delta^d \, \Delta_{12} \, X_t = \theta_q(B) \, \Theta_Q(B^{12}) a_t \tag{2.33}$$

where $\Delta^d = (1 - B)^d$, $\Delta_{12} = (1 - B^{12})$. Thus

$$\Delta^d \, \Delta_{12} \, X_t = \left[\theta_q(B) \, \Theta_Q(B^{12})/\phi_p(B) \, \Phi_P(B^{12})\right] a_t \tag{2.34}$$

is a stationary scheme which can be used for the simulation of $\Delta^d \Delta_{12} X_t$ which has the distorted spectral density $f_{\Delta^d\Delta 12}(\omega)$. However, $\{X_t\}$ is the series to be simulated. Summing (2.34)

$$X_t = \sum_{b_d=-\infty}^{t} \cdots \sum_{b_2=-\infty}^{b_3} \sum_{b_1=-\infty}^{b_2} \sum_{j=-\infty}^{0} \left[\theta_q(B)\ \Theta_Q(B^{12})/\phi_p(B)\ \Phi_p(B^{12})\right] a_{b_1+12_j} \qquad (2.35)$$

This is a divergent series with infinite variance. It increases without bounds as the simulation extends into the future. The effect of random shocks at the physically infinite past of X_t stays the same as the X_t values are extended into the future and the sum of random shocks increases pulling the series out of its mean level to erratic values.

Differencing, although very effective as a means for the removal of hydrologic periodicities, yields distorted spectra which are inconvenient for hydrologic simulation.

2.2.5 Monthly Mean Subtraction

Consider monthly hydrologic time series $X_{p,\tau} = X_{12p+\tau}$ which have a periodicity of 12 months were $\tau = 1, \ldots, 12$ is the month and $p = 0, \ldots, N-1$ is the year. Taking expression (2.14a) derived earlier,

$$X_{12p+\tau} = \sqrt{\delta} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sqrt{f(\omega)}\ e^{i2\pi(12p+\tau)\delta\omega}\ dZ(\omega) + \sqrt{\delta} \sum_{\alpha=0}^{6} \sqrt{\ell_\alpha Q_\alpha}\ \mathrm{Cos}\left[2\pi(12p+\tau)\ \frac{\alpha}{12} + \Phi_\alpha\right] \qquad (2.36)$$

becomes the spectral representation for the monthly periodic hydrologic time series. The mean for a particular month τ can be expressed as

$$\overline{X}_\tau = \frac{\sqrt{\delta}}{N} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sqrt{f(\omega)}\ \frac{1 - e^{i2\pi 12\delta\omega N}}{1 - e^{i2\pi 12\delta\omega}}\ e^{i2\rho\tau\delta\omega}\ dZ(\omega)$$

$$+ \frac{\sqrt{\delta}}{N} \sum_{\alpha=0}^{6} \sqrt{\ell_\alpha Q_\alpha} \sum_{p=0}^{N-1} \mathrm{Cos}\left[2\pi(12p+\tau)\ \frac{\alpha}{12} + \Phi_\alpha\right] \qquad (2.37)$$

Consider the case of subtracting the monthly means from the original periodic hydrologic time series. Then, expressions (2.36) and (2.37) yield

$$X_{12p+\tau} - \overline{X}_\tau = \sqrt{\delta} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \sqrt{f(\omega)} \left[1 - \frac{1}{N}\ e^{-i\pi 12\delta\omega(2p+1-N)}\ \frac{\mathrm{Sin}\ 12\pi\delta\omega N}{\mathrm{Sin}\ \pi 12\delta\omega}\right] e^{i2\pi(12p+\tau)\delta\omega}\ dZ(\omega) \qquad (2.38)$$

which is solely a continuous spectral representation. Therefore, subtracting the monthly means removes the discrete spectral part representing the periodic component in the hydrologic time series. The spectral density $\lambda(\omega)$ of the series $X_{12p+\tau} - \overline{X}_\tau$ can be written as:

$$\lambda(\omega) = f(\omega)\left[1 - \frac{2}{N}\ \frac{\mathrm{Sin}\ 12\pi\delta\omega N}{\mathrm{Sin}\ 12\pi\delta}\ \mathrm{Cos}\ 12\pi\delta\omega(2p+1-N) + \frac{1}{N^2}\left(\frac{\mathrm{Sin}\ 12\pi\delta\omega N}{\mathrm{Sin}\ 12\pi\delta\omega}\right)^2\right] \qquad (2.39)$$

Therefore, although the first order periodicity represented by the discrete spectral contribution is completely removed, a different kind of nonstationarity is introduced into the monthly hydrologic time series when monthly means are subtracted. This is because $\lambda(\omega)$ is a function of the year p as well as the frequency. Analyzing expression (2.39) Sin $12\pi\delta\omega N$/Sin $12\pi\delta\omega$ has its peaks at $\omega = k/12$ where $k = 0, 1, \ldots, 6$. However, at exactly these values Cos $12\pi\delta\omega(2p+1-N)$ = Cos $12\pi\delta\omega(1-N)$. That is, at $\omega = k/12$, $k = 0, 1, \ldots, 6$, the spectral density $\lambda(\omega)$ is independent of p and, therefore, independent of time. For $\omega \neq k/12$, $k = 0, 1, \ldots, 6$, Sin $12\pi\delta\omega N$/Sin $12\pi\delta\omega$ decreases sharply, thereby minimizing the effect of Cos $12\pi\delta\omega(2p+1-N)$, and the nonstationarity effect due to year p is negligible. Therefore, subtracting the monthly means essentially removes the periodicity in the autocovariance function and yields a hydrologic time series which satisfies the weak second order stationarity conditions for the case under study.

2.2.6 Applications

The analytical results obtained through the manipulation of the spectral representation for the monthly periodic hydrologic series were compared to the spectral density estimates of the rescaled monthly rainfall square roots for 15 Indiana watersheds. Spectral density and the autocorrelation function estimates of the original square root transformed rescaled monthly rainfalls, of the 1-lag differenced rescaled monthly rainfall square roots, of the seasonal 12-lag differenced rescaled monthly rainfall square roots and of the rescaled monthly square roots after the monthly means are subtracted were obtained for these 15 Indiana watersheds. In the spectral density computations the Hamming Window was used for smoothing purposes. The results obtained consistently verified the analytical conclusions. As an example, the results for the watershed of the Mississinewa River at Marion, Indiana, identified as station 3265, are shown in Figure 18a through 21b. Figure 18a shows the spectral density estimates of the monthly rainfall square roots. In this figure the discrete periodic spectral component at the frequency of 1/12 cycles per month is clearly identified. Figure 18b shows a definite 12-monthly periodic component in the autocorrelation function for the rescaled monthly rainfall square roots. Figure 19a shows the spectral density of the 1-lag differenced rescaled monthly rainfall square roots. The spectral contribution at the origin and in its neighborhood are completely wiped out meaning that the long run properties of the sample are removed. At the frequency 1/12, which shows the contribution of the yearly periodicity, the spectral density is reduced from the original value of 1.05 to 0.20. Because of this significant reduction in the discrete spectral component, the periodicity in the autocorrelation is almost eliminated as can be seen in Figure 19b. On the other hand the high frequency contributions are magnified. A spurious significant autocorrelation is introduced at the first lag of the autocorrelation function estimate. This is explained by equations (2.17) and (2.19), and the explanation is given in the appendix as to the removal of the discrete spectral component. Figure 20a shows the spectral density estimate of the seasonally 12-lag differenced rescaled monthly rainfall square roots. The yearly spectral component at the frequency 1/12 is again wiped out. The spectral density function itself is almost periodic with the frequency period of 1/12. These features were observed through the analytical treatment also. In Figure 20b it is seen that the periodic component of the autocorrelation is introduced at 12th lag for the seasonally 12-lag differenced rescaled monthly rainfall square roots. This result is explained by equations (2.22), (2.24) and the mathematical appendix. Figure 21a shows the spectral density when only the rescaled monthly rainfall square root means are subtracted. At the yearly frequency 1/12 the discrete spectral contribution is effectively decreased from the original value of 1.05 to .500. The rather high value of .50 is explainable by equation (2.39) derived earlier. The autocorrelation function estimate corresponding to the spectral density of Figure 21a is shown in Figure 21b. Figure 21b shows the analytical conclusion, equation (2.38), that the periodicity within the autocovariance function is removed by simply subtracting the monthly means from the monthly rainfall square roots.

The rainfall data for each watershed were obtained by weighing the rainfall stations in that watershed by the Thiessen polygon method.

The modified Fisher test (*Andel and Balek*, 1971) for the detection of periodicities in the hydrologic time series was employed to test the effects of the removal of periodicities by the methods under consideration. However, the test proved to be insufficiently sensitive.

2.3 A SPECTRAL AND VARIANCE-TIME LOOK AT LONG RANGE DEPENDENCE

The long range dependence in the hydrologic time series is manifested by the fact that the extreme events may persist for a very long time like the "seven years of plenty and the seven years of famine" in the Bible or the four-year dry period at Iquique, Chile (*Pettersen*, 1969). In the design of water resources systems through simulation methods, the hydrologic sequence is synthesized for a very long time, usually extending from 500 to 1000 years. In a 1000 year long record it is very reasonable to expect very

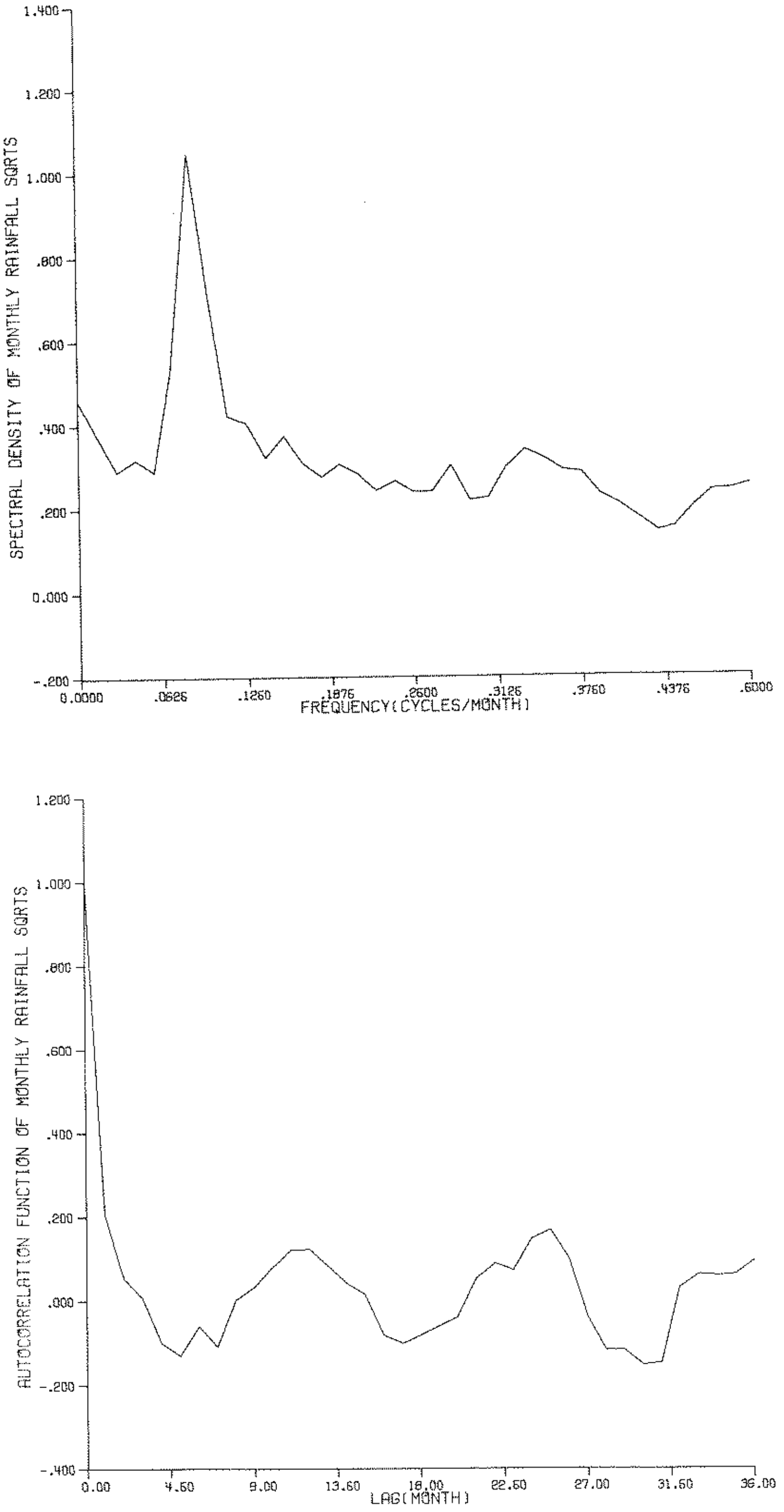


FIG. 18 (a,b) SPECTRAL DENSITY AND AUTOCORRELATION FUNCTIONS OF RESCALED MONTHLY RAINFALL SQUARE ROOTS

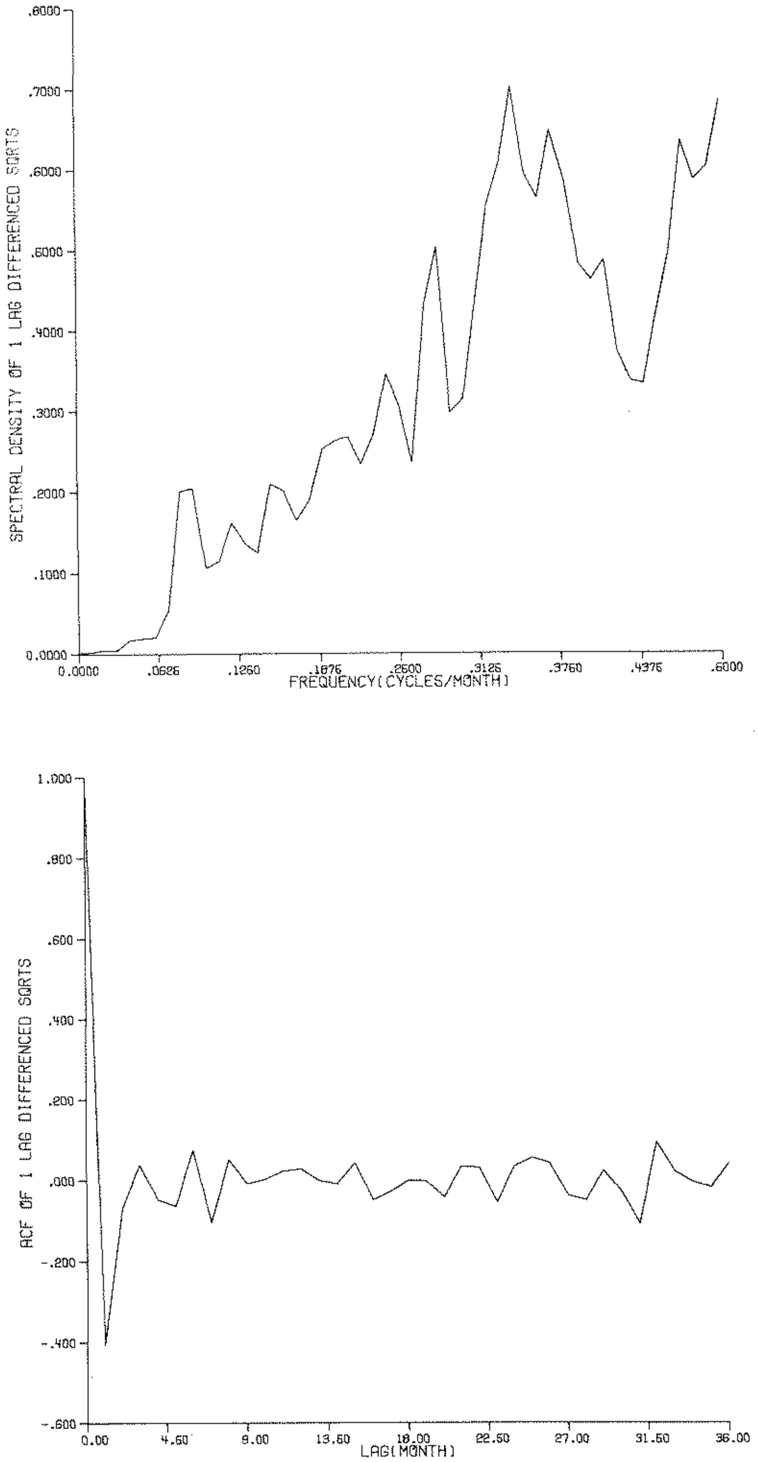


FIG. 19 (a,b) SPECTRAL DENSITY AND AUTOCORRELATION FUNCTIONS OF RESCALED I LAG DIFFERENCED MONTHLY RAINFALL SQUARE ROOTS

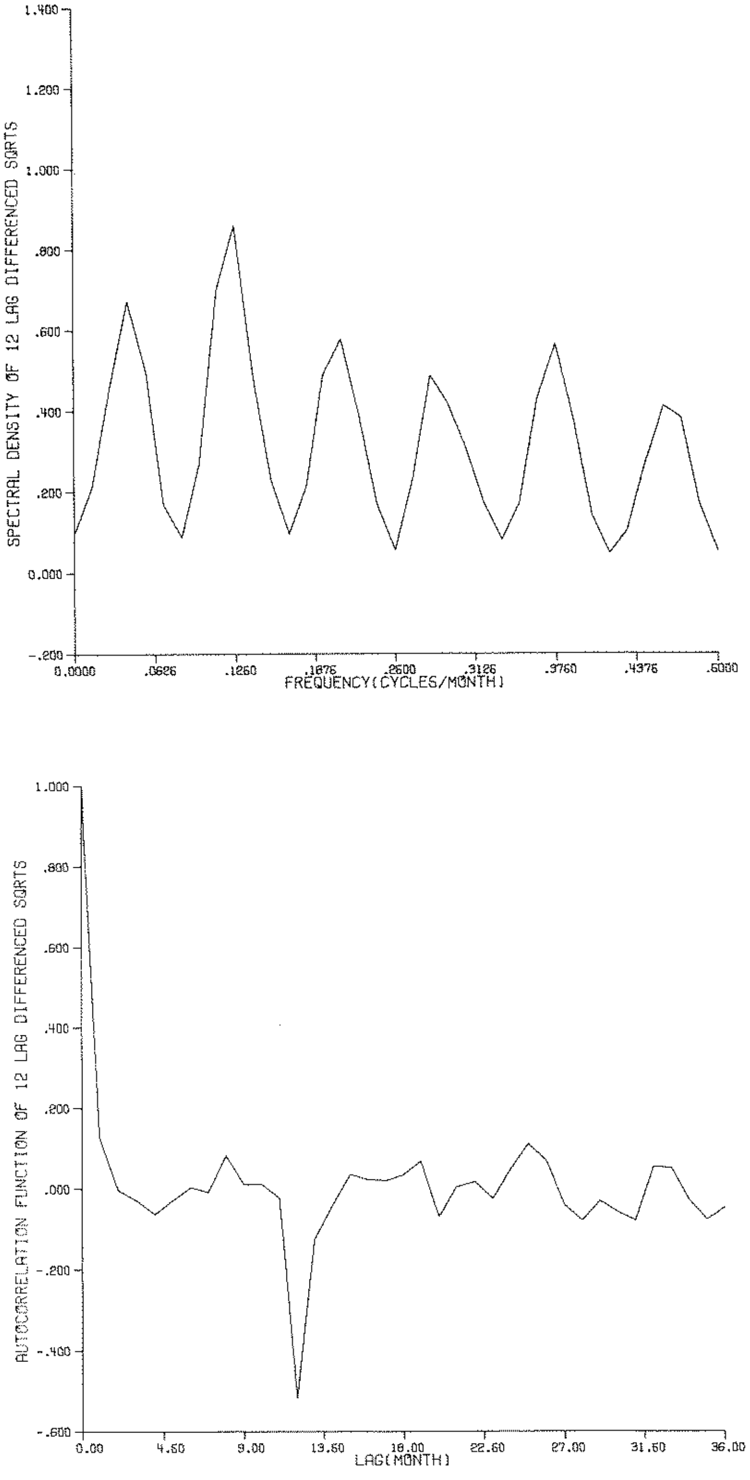


FIG. 20 (a,b) SPECTRAL DENSITY AND AUTOCORRELATION FUNCTIONS OF RESCALED 12 LAG DIFFERENCED MONTHLY RAINFALL SQUARE ROOTS

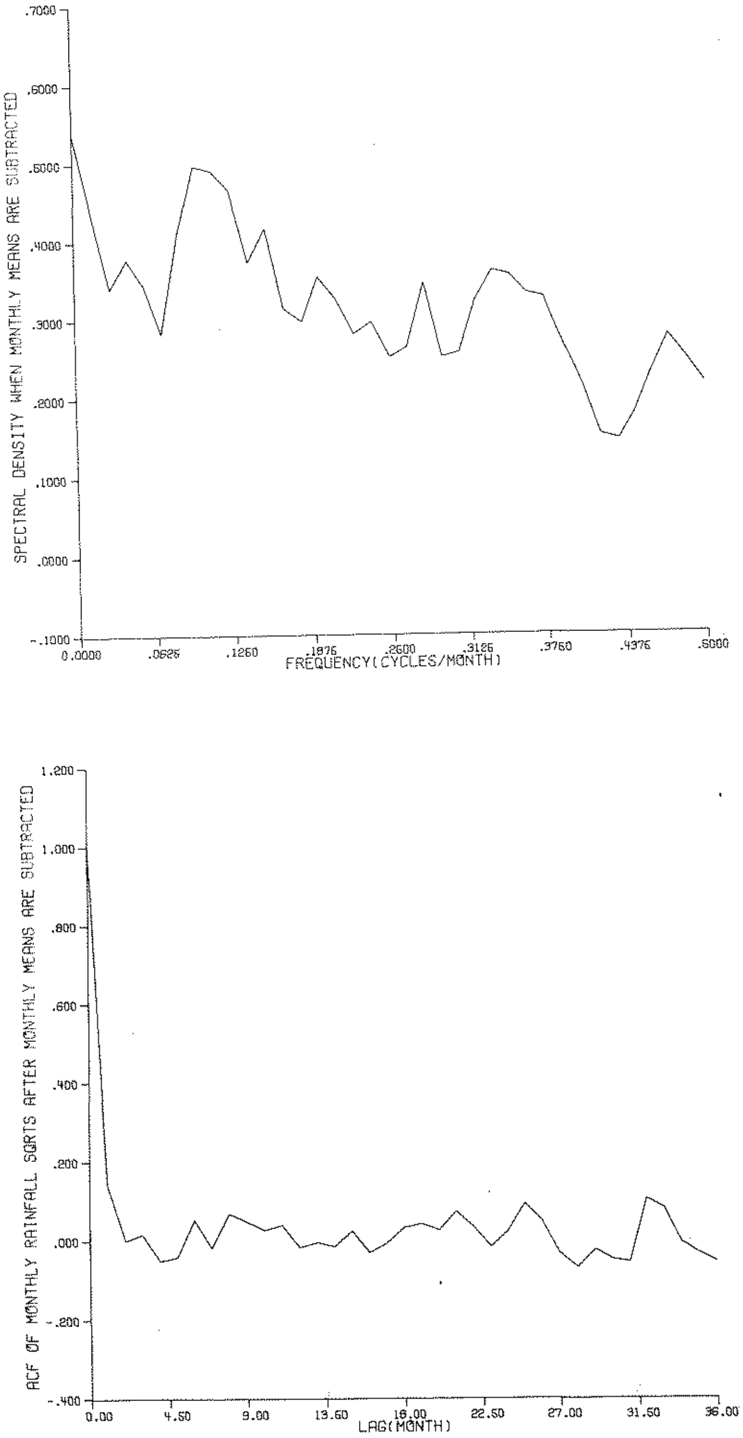


FIG. 21 (a,b) SPECTRAL DENSITY AND AUTOCORRELATION FUNCTIONS OF DIFFERENCES BETWEEN RESCALED MONTHLY RAINFALL SQUARE ROOTS AND RESCALED MONTHLY MEANS extreme precipitation and extraordinarily high river levels. *Mandelbrot and Wallis* (1968) labelled the occurrence of such very extreme phenomena as the Noah Effect. On the other hand the occurrence of very long periods of high precipitation and of droughts, as is indicated in the Bible, is to be expected, too. This persistence effect is called the Joseph Effect by *Mandelbrot and Wallis* (1968). It is evident that the occurrence of a very long period of drought or a very long period of flood would necessitate enormously large reservoir capacities. Therefore, simulation of the long range dependence effect is of vital importance to the water resources system planner. The immediate questions that the hydrologist, synthesizing the hydrologic record, has to answer are: (1) just how long the span of time dependence should be in order to simulate the extremes that are expected to occur within the life time of the hydraulic structures under study; (2) how to incorporate this span of dependence into the time series models to be used for synthesis; (3) can the time series model simulate the long range dependence facts observed in the long hydrologic records? In this section an attempt will be made to answer these questions mainly by the variance time and spectral analysis of the present time series models.

The τ law of the variance for the Brownian Motion domain is:

$$\text{Var } [W(t+\tau) - W(t)] = \text{Var} \sum_{u=t+1}^{t+\tau} X(u) = k\tau \tag{2.40}$$

where $\{X(u)\}$ is a white noise sequence with $E[X(u)] = 0$, and $W(t+\tau) = W(t) + \sum_{u=t+1}^{t+\tau} X(u)$. $W(t)$ is the Brownian Motion. The τ law of variance for the Brownian motion states that the variance of the partial sums from a white noise sequence $\{X(u), u = \ldots 1, 0, +1, \ldots\}$ is proportional to the time span of summation with a random proportionality constant k (*Parzen*, 1967).

Hurst (1951, 1956, 1965) studied a wide range of natural phenomena with very long records and observed the long range persistence of the "Joseph Effect" in terms of the variance-time function and the rescaled range. The Hurst law of variance can be stated as

$$\lim_{\tau\to\infty} \text{Var} \sum_{u=t+1}^{t+\tau} X(u) = k\tau^{2h} , \quad 0.5 < H < 1, \tag{2.41}$$

that is, the variance of partial sums of a hydrologic time series is asymptotically proportional to τ^{2H} where $0.5 < H < 1$. The observations of *Hurst* showed that the dependence structure in the hydrologic processes is very long, or physically infinite.

For the stationary random sequence $\{X(u), u = \ldots 1, 0, +1 \ldots\}$ the variance-time function can be expressed as

$$V(\tau) = \text{Var} \left[\sum_{u=1}^{\tau} X(u) \right] = \tau\, r_0 + 2 \sum_{\ell=1}^{\tau} (\tau - \ell)\, r_\ell \tag{2.42}$$

where r_ℓ is the covariance of $\{X(u)\}$ at the ℓ-th lag.
For a non-stationary random sequence the variance-time function $V(t)$ is expressed as

$$\begin{aligned} V(t) &= \text{Var } [X_1 + X_2, \ldots + X_t] \\ &= t \text{ Var } (X_1) + \sum_{i=1}^{t-1} \sum_{u=1}^{t-i} \text{Cov } (X_i, X_{i+h}) + \sum_{i=1}^{t-1} \sum_{u=1}^{t-i} \text{Cov } (X_{i+h}, X_i). \end{aligned} \tag{2.43}$$

Considering expression (2.42) of a stationary random sequence,

$$\lim_{\tau\to\infty} \sum_{\ell=1}^{\tau} \frac{1}{\tau} \ell r_\ell = 0 \quad \text{when} \quad \lim_{\ell\to\infty} r_\ell = 0 \quad \text{or when} \quad \sum_{\ell=1}^{\infty} r_\ell < \infty$$

So that

$$\lim_{\tau\to\infty} \text{Var} \left[\sum_{u=1}^{\tau} X(u) \right] = \tau \left[r_0 + 2 \sum_{\ell=1}^{\tau} r_\ell \right] \quad \text{if} \quad \sum_{\ell=1}^{\infty} r_\ell < \infty \tag{2.44}$$

The spectrum $S(\omega)$ of the stationary time series $\{X(u)\}$ is expressed as

$$S(\omega) = r_0 + 2 \sum_{\ell=1}^{\infty} r_\ell \, [e^{-i2\pi\omega\ell} + e^{i2\pi\omega\ell}] \Big/ 2$$

$$= r_0 + 2 \sum_{\ell=1}^{\infty} r_\ell \text{ Cos } 2\pi\omega\ell \tag{2.45}$$

Thus, the value of the spectrum at the origin, $S(0)$, is

$$S(0) = r_0 + 2 \sum_{\ell=1}^{\infty} r_\ell \tag{2.46}$$

From (2.44) and (2.46) it follows that

$$\lim_{\tau\to\infty} \text{Var} \left[\sum_{u=1}^{\tau} X(u) \right] = \tau \, S(0) \qquad \text{if} \qquad \sum_{\ell=1}^{\infty} r_\ell < \infty \tag{2.47}$$

The condition $\sum_{\ell=1}^{\infty} r_\ell < \infty$ may be written in terms of the spectrum and the variance of the time series as

$$\sum_{\ell=1}^{\infty} r_\ell < \infty \leftrightarrow \frac{S(0) - r_0}{2} < \infty \tag{2.48}$$

However, due to the second-order-stationarity (s.o.s.) conditions the variance r_0 of a stationary random sequence $\{X(u)\}$ is finite. Thus

$$\sum_{\ell=1}^{\infty} r_\ell < \infty \leftrightarrow S(0) < \infty, \tag{2.49}$$

that is,

$$\lim_{\tau\to\infty} \text{Var} \left[\sum_{u=1}^{\tau} X(u) \right] = \tau \, S(0) \qquad \text{if} \qquad S(0) < \infty. \tag{2.50}$$

When the variance of partial sums is asymptotically proportional to the time span τ as $\tau\to\infty$ the process ends up in the Brownian domain and, thus, has finite dependence. Therefore, the statistic to differ the processes with finite dependence from the processes with infinite dependence, observing expressions (2.49) and (2.50), is $S(0)$, the spectral value of the stationary time series at the zero frequency.

It follows from the above argument that the current hydrologic simulation models can be classified with the help of the spectrum $S(\omega)$ as:

1. Models with finite span of dependence and finite variance to simulate the short span hydrologic dependence. These would require

$$S(0) \text{ to be finite and}$$

$$\int_0^{\frac{1}{2}\delta} S(\omega)\, d\omega < \infty, \qquad 0 < \omega < \tfrac{1}{2}\delta.$$

2. Models with infinite span of dependence and finite variance to simulate not only the short range but also the long range dependence in terms of the Joseph effect. These would require

$$S(0) = \infty, \qquad \int_0^{\frac{1}{2}\delta} S(\omega)\, d\omega < \infty, \qquad 0 < \omega < \tfrac{1}{2}\delta.$$

3. Models which will yield infinite span of dependence and infinite variance to simulate not only the Joseph effect but also the Noah effect. These models would require either

$$S(0) = \infty, \qquad \int_0^{\frac{1}{2}\delta} S(\omega)\, d\omega = \infty$$

or

$$S(0) = \infty, \qquad \int_0^{\frac{1}{2}\delta} S(\omega)\, d\omega < \infty$$

and the white noise input to the hydrologic model to have marginal probability distributions with infinite variance. The first set of requirements in the models of type 3 would yield a nonstationary series while the second set of conditions would yield a stationary model with infinite span of dependence. In the second set the infinite variance of the generated time series would be obtained by the infinite variance of the noise input.

Mandelbrot (1969) observed that the spectra of the hydrologic processes exhibiting the Joseph effect, are J-shaped and proposed that such processes can be explained by assuming that $\lim_{\omega \to 0} S(\omega) = \infty$. Therefore, *Mandelbrot's* proposal takes the span of dependence of the simulation model to be infinite. He proposed the use of models with hyperbolic spectra, $S(\omega) = \omega^{\alpha-2}$, $1 < \alpha < 2$, to explain the Joseph effect. He called this proposal the "Hyperbolic spectrum hypothesis."

The presence of the Noah and Joseph effects together can cause significant variations in the different samples of a population since they correspond to extremely large values persisting for a long time. Since "the sample variances are enormously influenced by the precise values of the outliers" (*Mandelbrot*, 1969), it becomes impossible to estimate the population variance due to large variation of the sample variance. A hydrologic sample of "physically infinite" size would quite likely contain many Joseph and Noah effects, and sometimes both these effects together. This sample would yield infinite variance. *Mandelbrot* (1969) proposed the "Infinite Variance Hypothesis" which claims that the population variance of the generating process is infinite. This infinite variance hypothesis corresponds to having probability distributions with hyperbolic tails for the generating process. *Mandelbrot* (1969) proposed the use of stable probability distributions such as *Cauchy's* for this purpose. The infinite variance hypothesis corresponds to having a spectrum of the form

$$S(\omega) = c\omega^{-\alpha}, \qquad \alpha \geq 1 \tag{2.51}$$

locally around $-\varepsilon < \omega < \varepsilon$. This would yield not only infinite variance but also infinity at the spectral origin. Thus a model possessing a spectrum of one form (2.51) would be nonstationary.

For an ARMA (p,q) process it was discussed in the section on the periodicities that

$$S(0) = \left(\sum_{\alpha=0}^{q} \theta_\alpha\right)^2 \sigma_a^2 \Big/ \left(\sum_{\beta=0}^{p} \phi_\beta\right)^2 \tag{2.52}$$

Thus for any ARMA (p,q) model S(0) is finite and the ARMA (p,q) family of models is in the Brownian domain. It follows from (2.50) that the asymptotic variance-time behavior of an ARMA (p,q) model can be expressed as

$$\lim_{\tau\to\infty} \mathrm{Var}\left[\sum_{u=1}^{\tau} X(u)\right] = \tau\, S(0) = \tau\, \sigma_a^2 \left(\sum_{\alpha=0}^{q} \theta_\alpha\right)^2 \Big/ \left(\sum_{\beta=0}^{p} \phi_\beta\right)^2 . \tag{2.53}$$

The ARMA (1,1) model was suggested by *O'Connell* (1971) as a model that would preserve the long range dependence properties. It follows from (2.53), that the asymptotic behavior of the variance-time function for ARMA (1,1) is

$$\lim_{\tau\to\infty} \mathrm{Var}\left[\sum_{u=1}^{p} X(u)\right] = \tau\, \frac{(1-\theta)^2}{(1-\phi)^2}\, \sigma_a^2 \tag{2.54}$$

which simply shows that the model has finite dependence span. If the model is forced to preserve the *Hurst's* law for the variance, then the condition

$$\lim_{\tau\to\infty} \text{Var} \left[\sum_{u=1}^{\tau} X(u)\right] = k\tau^{2H}, \qquad \tfrac{1}{2} < H < 1 \tag{2.55}$$

has to be satisfied. From (2.42) it follows that for the ARMA (1,1) model

$$\text{Var} \left[\sum_{u=1}^{\tau} X(u)\right] = \tau\, r_0 + 2r_1 \sum_{\ell=1}^{\tau} (\tau - \ell)\, \phi^{\ell-1} \tag{2.56}$$

Expanding (2.56) and dividing both sides by τ,

$$\frac{1}{\tau} \text{Var} \left[\sum_{u=1}^{\tau} X(u)\right] = r_0 + 2r_1 \left(\frac{1 - \phi^{\tau}}{1 - \phi}\right) - 4r_1 \left(\frac{-\phi^{\tau} + \phi^{\tau+1}}{(1 - \phi)^2}\right) - 4r_1 \left(\frac{1 - \phi^{\tau}}{(1 - \phi)^2}\right) \frac{1}{\tau} \tag{2.57}$$

and, using (2.55)

$$r_0 + 2r_1 \left(\frac{1 - \phi^{\tau}}{1 - \phi}\right) - 4r_1 \left(\frac{-\phi^{\tau} + \phi^{\tau+1}}{(1 - \phi)^2}\right) - 4r_1 \left(\frac{1 - \phi^{\tau}}{(1 - \phi)^2}\right) \frac{1}{\tau} = k\tau^{2H-1} \tag{2.58}$$

should be satisfied by the ARMA (1,1) model for τ large. When τ is large

$$-\phi^{\tau} + \phi^{\tau+1} \approx 0$$

and

$$r_0 + 2r_1 \left(\frac{1 - \phi^{\tau}}{1 - \phi}\right) = r_0 + 2r_1 \left(\frac{1}{1 - \phi}\right) = S(0). \tag{2.59}$$

Then,

$$S(0) \approx k\tau^{2H-1} \qquad \text{for large } \tau \text{ and} \qquad \tfrac{1}{2} < H < 1. \tag{2.60}$$

Expression (2.60) gives the necessary condition for the ARMA (1,1) to preserve the *Hurst's* law for the variance. However, since S(0) is finite, condition (2.60) can not be satisfied by the ARMA (1,1) model. Therefore, the ARMA (1,1) model can not satisfy the *Hurst's* law for the variance. It follows from (2.58) that the only way for the ARMA (1,1) to satisfy the *Hurst's* law is by having time-varying parameters.

Denote the k-th difference of X_J by $\Delta^{(k)}X_J$ where

$$\Delta^{(k)}\, X_J = \Delta^{(k-1)}\, X_J - \Delta^{(k-1)}\, X_{J-1}. \tag{2.61}$$

For any lag i

$$\Delta^{(k)}\, X_{J-i} = \Delta^{(k-1)}\, S_{J-i} - \Delta^{(k-1)}\, X_{J-i-1}$$

and

$$\Delta^{(k)}\, X_{J-i}\, \Delta^{(k)}\, X_J = \Delta^{(k-1)}\, X_{J-i}\, \Delta^{(k-1)}\, X_J - \Delta^{(k-1)}\, X_{J-i-1}\, \Delta^{(k-1)}\, X_J$$
$$- \Delta^{(k-1)}\, X_{J-i}\, \Delta^{(k-1)}\, X_{J-1} + \Delta^{(k-1)}\, X_{J-i-1}\, \Delta^{(k-1)}\, X_{J-1} \tag{2.62}$$

Take Cov $[\Delta^{(i)}\, X_{J-i}\, \Delta^{(k)}\, X_J] = E[\Delta^{(k)}\, X_{J-i}\, \Delta^{(i)}\, X_J]$ since $E[\Delta^{k}X_J] = 0$. Denote Cov $[\Delta^{(k)}\, X_{J-i}\, \Delta^{(k)}\, X_J]$ by $R_i^{(k)}$ and the covariance function of the original time series $\{X_J\}$ at the ℓ-th lag by r_ℓ. From (2.62) it follows

$$R_i^{(1)} = \sum_{J=0}^{2} \binom{2}{J} (-1)^{J+1}\, r_{i+1-J}$$

$$R_i^{(2)} = \sum_{J=0}^{4} \binom{4}{J} (-1)^{J+2}\; r_{i+2-J}$$

$$\vdots$$

$$R_i^{(k)} = \sum_{J=0}^{2k} \binom{2k}{J} (-1)^{J+k}\; r_{i+k-J} \qquad (2.63)$$

Denote the spectrum for the k-th differenced series by $S^{(k)}(\omega)$. Then

$$S^{(k)}(\omega) = \sum_{\alpha=-\infty}^{+\infty} R_\alpha^{(k)}\; e^{i2\pi\omega\alpha\delta}, \qquad -\tfrac{1}{2}\delta \leq \omega \leq \tfrac{1}{2}\delta$$

$$= \sum_{\alpha=-\infty}^{+\infty} \sum_{J=0}^{2k} \binom{2k}{J} (-1)^{J+k}\; r_{\alpha+k-J}\; e^{-i2\pi\omega\alpha\delta}$$

$$S^{(k)}(\omega) = \lambda(\omega)\; (2 \text{ Sin } \pi\omega\delta)^{2k}, \qquad -\tfrac{1}{2}\delta \leq \omega \leq \tfrac{1}{2}\delta \qquad (2.64)$$

where $\lambda(\omega)$ is the spectrum of the original series $\{X_J\}$.
Following expression (2.64) and the argument given in the subsection on "model fitting and simulation" in the section 2.2 on the periodicities, an ARIMA (1,d,1) model, in its generating form, has the spectrum

$$S(\omega) = k\omega^{-2d} \qquad d = 1, 2, \ldots, \qquad -\varepsilon \leq \omega \leq \varepsilon \qquad (2.65)$$

for a random constant k. Thus $\lim_{\omega \to 0} S(\omega) = \infty$ and the ARIMA (1,d,1) family of models have, in their generating form, an infinite span of dependence and infinite variance. Thus it falls to the third class. The basic inconvenience of the ARIMA (1,d,1) family of models is that it is nonstationary in its generating form.

As a result of the above analysis it is seen that the ARMA (p,q) models have finite span of dependence. An important problem is the determination of the length of this span. This problem will be tackled by the help of the autocovariance structure of the ARMA (p,q) and the threshold concept of *Mandelbrot* (1969). For any ARMA (p,q) model the autocovariance function r_ℓ will decay to zero from the lag a = q+1 on either exponentially or in damped sinusoidal form (*Box and Jenkins*, 1971). Set some threshold $0 < s < 1$ such that the condition $|r_\ell / r_0| \leq s$ leads to r_ℓ approximately equal to zero. Then the lag T where the autocovariance r_ℓ passes $s\; r_0$ for the first time, becomes the span of dependence. For the first-order autoregressive model AR(1) the span of dependence T is

$$T = \frac{\ln s}{\ln|\phi|}, \qquad 0 < s < 1 \qquad |\phi| < 1. \qquad (2.66)$$

This is the span of initial transient where the *Hurst's* laws hold. For the ARMA (1,1) model, T is

$$T = \left[\ln \left[s \frac{1 + \theta^2 - 2\phi\theta}{|(1 - \phi\theta)(\phi - \theta)|} \right] \Big/ \ln |\phi| \right] + 1, \qquad |\theta| < 1, \qquad 0 < s < 1, \qquad |\phi| < 1 \qquad (2.67)$$

CHAPTER 3 - ANALYSIS OF THE MONTHLY RAINFALL DATA

3.1 APPLICATION OF NONSEASONAL MODELS TO MONTHLY RAINFALL SERIES

In this section the models fitted to the monthly rainfall series after the series are square root transformed and then stationarized by nonseasonal first lag differencing or by standardization, will be considered. There are basically three steps in fitting mathematical models to stationary time series. These steps are 1) identification of the model, 2) estimation of the parameters of the model, 3) diagnostic check of the fitted model. In the sections below each of these steps will be described in detail in the context of the analysis of the monthly rainfall series.

3.1.1 Identification

Identification of the time series model is done through the use of the autocorrelation function and of the partial autocorrelation function. The behavior of these functions is reverse of each other and because of this property they render a quick procedure for the identification of the model and its order. While the autocorrelation function of an AR(p) scheme tails off, the partial autocorrelation function for the same scheme has a cutoff after lag p. On the other hand, for an MA(q) process the autocorrelation function has a cutoff after lag q while the partial autocorrelation function tails off. For ARMA (p,q) model, if $q < p$, the autocorrelation function decays from (q-p+1) lag on, and for $q > p$ the autocorrelation function decays from lag (q-p+1) lag on while the partial autocorrelation function decays from lag zero on.

3.1.1a One-lag Differenced Series

Table 3-1 lists the estimated autocorrelation function $\hat{\rho}_k$ calculated by

$$\hat{\rho}_k = \frac{\frac{1}{N}\sum\limits_{t=1}^{N-k}(y_t-\bar{y})(y_{t+k}-\bar{y})}{\frac{1}{N-1}\sum\limits_{t=1}^{N}(y_t-\bar{y})^2} \qquad (3.1)$$

for the second order stationary time series $\{y_t\}$. The series $\{y_t\}$ were stationarized through the first lag nonseasonal differencing and then normalized. Table 3-2 gives the estimated partial autocorrelation function $\hat{\phi}_{\ell\ell}$ for the same series (*Box and Jenkins*, 1971). The $\hat{\phi}_{\ell\ell}$ were calculated by

$$\hat{\phi}_{\ell\ell} = \hat{\rho}_1 \qquad , \quad \ell = 1$$

$$\hat{\phi}_{\ell\ell} = \frac{\hat{\rho}_1 - \sum\limits_{J=1}^{\ell-1}\hat{\phi}_{\ell-1,J}\,\hat{\rho}_{\ell-J}}{1 - \sum\limits_{J=1}^{\ell-1}\hat{\phi}_{\ell-1,J}\,\hat{\rho}_J} \quad , \quad \ell = 2, 3, \ldots L \qquad (3.2)$$

and

$$\hat{\phi}_{\ell J} = \hat{\phi}_{\ell-1,J} - \hat{\phi}_{\ell\ell}\,\hat{\phi}_{\ell-1,\ell-J} \quad , \quad J = 1, 2, \ldots \ell-1$$

3.1.1b Standardized Series

Tables 3-3 and 3-4 give the autocorrelation and the partial autocorrelation functions for the monthly rainfall series which were stationarized through the standardization procedure.

3.1.1c Statistical Tests and Discussion

Since in the earlier studies by *Roesner and Yevdjevich*, (1966), it was concluded that the monthly rainfall time series is a white noise sequence, two statistical tests were applied to second order stationary time series to test the hypothesis of white noise. The first test is based on *Anderson's* statistic. The second test is *Box and Pierce's Portemanteau* lack of fit test.

If a stationary time series $\{y_t\}$ is independently, normally distributed, for a moderate sample size, N, *Anderson* (1942) showed that the first lag correlation coefficient, ρ_1, is normally distributed with mean $-1/(N-1)$ and variance $(N-2)/(N-1)^2$. He also noted that for any lag L the distribution of ρ_L is again normally distributed. For N quite large ρ_L can be assumed to be approximately normal with mean 0 and variance 1/N. Then a test for white noise can be devised at 5% level by considering the 95% confidence interval of ρ_L as

$$\rho_L\ (95\%) = \pm\ 1.96\ (1/\sqrt{N}) \tag{3.3}$$

and considering the null hypothesis: $\rho_L = 0$ for $L \geq 1$.

The second test is the *Portemanteau* lack of fit test utilized by *Box and Pierce* (1970). Taking the first L autocorrelations $\hat{\rho}_\ell(a)$, $\ell = 1, \ldots$ L of the residuals $\{\hat{a}_t\}$ from any ARIMA (p,d,q) process, *Box and Pierce* (1970) showed that if the fitted model is appropriate, then the statistic

$$Q = n \sum_{\ell=1}^{L} \hat{\rho}_\ell^2\ (\hat{a})$$

is approximately distributed as $\chi^2(\ell-p-q)$ where n = N-d is the number of y's used to fit the model. On the hypothesis that the model is a white noise, that is, an ARIMA (0,d,0) model, then the residuals $\{\hat{a}_t\}$ are the stationary time series $\{y_t\}$ and $\hat{\rho}_\ell(\hat{a})$ is the autocorrelation function of the stationary time series. Then, if ARIMA (0,d,0) is appropriate, the statistic Q will be approximately distributed as χ_L^2. The tests for a white noise model for the monthly rainfall series based on the above two statistics are given on tables 3-1 and 3-2.

From tables 3-1 and 3-2 it can be seen that the tests for the (0,1,0) white noise hypothesis fail at the 5% level. This is expected since the first lag differencing amplifies the high frequency components of the sample spectrum. As can be seen from table 3-1 all the autocorrelation functions have a cutoff after lag 1. On the other hand, all the partial autocorrelation functions tail off. This behavior suggests the theoretical behavior of an MA(1) model. Therefore, the ARIMA (0,1,1) model was selected for further considerations.

If tables 3-3 and 3-4 are analyzed, it can be seen that white noise hypothesis for square root transformed, standardized monthly rainfall series is accepted at 5% level by *Anderson's* test and at 10% level by the *Portemanteau* lack of fit test in 5 out of 15 cases. There was one case where the hypothesis was accepted by *Anderson's* test at 5% level but was doubtful at 10% with $\chi_{10}^2 = 16$ in the *Portemanteau* lack of fit test since the sample statistic Q was equal to 15.83. There was one case where *Anderson's* test was doubtful at 5% level since the confidence limit $\hat{\rho}_L$ (95%) was 1.0765 while there were three autocorrelation coefficients with the value 1.071. The autocorrelation coefficient at 33rd lag was .08 and the autocorrelation coefficient at first lag was .09. However, for the same case the *Portemanteau* lack of fit test was accepted at 10% level with $\chi_{10}^2 = 16.0$. Including the doubtful cases as acceptable, the white noise hypothesis was accepted in 7 out of 15 cases. Therefore, the white noise model is inadequate to explain the stationary component of the monthly rainfall time series. If the autocorrelation function is analyzed, it will be seen that the first lag autocorrelation coefficient is small but significant. This was tested by considering *Bartlett's* (1946) approximate statistic for the variance of the estimated autocorrelation function of a normal process given as

$$\mathrm{Var}\ [\hat{\rho}_\ell] \simeq \frac{1}{N} \sum_{v=-\infty}^{+\infty} \left\{\rho_v^2 + \rho_{v+\ell}\ \rho_{v-\ell} - 4\rho_\ell\ \rho_v\ \rho_{v-\ell} + 2\rho_v^2\ \rho_\ell^2\right\}$$

For a process where $\rho_v = 0$ for $v > k$

$$\text{Var } [\hat{\rho}_\ell] \simeq \frac{1}{N} \left\{1 + 2 \sum_{v=1}^{k} \rho_v^2\right\} \tag{3.4}$$

Therefore, the standard errors of the estimated autocorrelations based on the hypothesis that $\rho_\ell = 0$ for $\ell > 1$ will be

$$\hat{\sigma} \ [\hat{\rho}_\ell] \simeq \frac{1}{\sqrt{N}} \left\{1 + 2\rho_1^2\right\}^{\frac{1}{2}} \tag{3.5}$$

Quenouille (1949) showed that on the hypothesis that the process is autoregressive of order p the standard error of the estimated partial autocorrelation $\hat{\phi}_{kk}$ becomes

$$\hat{\sigma} \ (\hat{\phi}_{kk}) \simeq 1/\sqrt{n} \qquad k > p + 1 \tag{3.6}$$

Based on the hypothesis that $\rho_k = 0$ for $k > 1$, the standard deviation of $\hat{\rho}_\ell$ for $\ell > 1$ was calculated from (3.4) and the autocorrelation functions of the monthly time series were investigated for the values falling out of the range $\pm \hat{\sigma} \ (\hat{\rho}_\ell)$. This test was inconclusive since for almost all the investigated time series, with the exception of station 2840, there were enough $\hat{\rho}_\ell$, $\ell > 0$ where $|\hat{\rho}_\ell| > \hat{\sigma} \ (\hat{\rho}_\ell)$. However, since the distribution of $\hat{\rho}_\ell$ for the dependent series is not theoretically known, no further analysis could be applied. Howev r, on the hypothesis that $\rho_v = 0$ for $v > 0$, and the time series is a white noise the $\hat{\rho}_1$ was significant at 5% level in 10 out of 15 cases by *Andersons'* test. Therefore, the first lag autocorrelation coefficient was treated as significant and based on the general behavior of the sample autocorrelation function a cutoff was assumed after the first lag. Then the hypothesis that the monthly rainfall series is AR(1) was considered by the use of (3.6). Again the test was inconclusive since for all series except station 2840, there were $|\hat{\phi}_{kk}| > \hat{\sigma} \ (\hat{\phi}_{kk})$ for $k > 1$. Since the distribution of $\hat{\phi}_{kk}$ for dependent series was not known, a further analysis could not be pursued. If table 3-4 is analyzed, it is seen that the $\hat{\phi}_{kk}$ has a cutoff after lag 1 as a general behavior for the 15 time series considered.

The autocorrelation function of ARIMA (1,0,1) model decays from the first lag on while the partial autocorrelation function also decays from the first lag on. The sample autocorrelations generally had only $\hat{\rho}_1$ as significant and $\hat{\phi}_{11}$ is the only significant value for the autocorrelation. However, since $\hat{\rho}_1$ and $\hat{\phi}_{11}$ are small, the decay that should be observed is not observed. On the other hand neither AR(1) nor MA(1) can solely explain both the behavior of the autocorrelation and the partial autocorrelation functions. Therefore, the more general model (1,0,1) was considered for further studies of the standardized monthly rainfall time series.

3.1.2 Estimation

3.1.2a Initial Parameter Estimation

The initial estimates of the model parameters were obtained from the autocovariance structures of ARMA (p,q) models. For square root transformed first lag differenced monthly rainfall series ARIMA (0,1,1) was identified as a model. This model, written in open form

$$y_t = (1 - B) \ X_t = (1 - \theta B) \ a_t \tag{3.7}$$

has θ and σ_a^2 as its parameters. From the autocovariance structure of ARIMA (0,1,1)

$$\rho_1 = \frac{-\theta_1}{1 + \theta_1} \ . \tag{3.8}$$

TABLE 3-1

AUTOCORRELATION FUNCTIONS FOR SQUARE ROOT TRANSFORMED MONTHLY RAINFALL SERIES STATIONARIZED BY FIRST LAG DIFFERENCING AND TESTS FOR AIRMA (O,1,0) WHITE NOISE MODEL

STATION NUMBER / LAG	ESTIMATED AUTOCORRELATION FUNCTION																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2535	-.41	-.03	-.08	.09	-.10	.03	-.04	.05	-.05	.02	-.01	.09	-.02	-.01	-.04	.03	-.05	.04	-.05
2695	-.43	-.01	-.06	.03	-.03	-.01	-.02	.04	-.05	.04	-.03	.05	.01	.02	-.01	-.02	-.02	-.02	.00
3795	-.46	.01	-.07	.00	.03	-.01	.02	-.03	-.04	.00	.07	-.03	.03	-.03	.04	-.04	.04	-.04	-.01
3805	-.45	-.02	-.04	.00	.04	-.02	.02	.00	-.05	-.03	.10	-.03	.00	-.03	.07	-.06	.02	.00	-.03
3280	-.43	.02	-.10	-.01	.03	-.02	.00	-.02	.00	-.03	.02	.02	.07	-.06	.01	-.03	.02	-.02	-.02
3445	-.43	-.01	-.07	-.00	.02	.02	-.04	.01	-.03	-.03	.08	-.02	.06	-.03	.01	-.05	.05	-.07	.03
2840	-.41	-.06	-.03	.01	-.01	.02	.02	-.02	-.04	.03	-.02	.09	-.02	.01	-.06	.02	-.02	.03	-.01
3245	-.45	-.04	.04	-.06	.01	.01	-.06	.02	.02	-.01	-.01	.04	.02	-.04	.02	-.03	.01	.00	.00
3265	-.40	-.07	.04	-.05	-.06	.07	-.10	.05	.00	-.01	.03	.02	.02	-.03	.05	-.04	-.03	.01	-.01
6120	-.50	.03	-.03	.05	-.06	-.03	.11	-.08	.04	-.04	.05	-.06	.07	-.01	.00	-.01	-.04	.03	.02
3655	-.44	.00	-.03	-.01	-.05	.05	-.07	.09	-.04	-.02	.03	.03	.02	-.02	.02	-.05	.00	.01	-.03
2750	-.45	.01	.04	-.06	-.03	.04	.06	-.09	-.03	.04	.02	-.04	.02	-.01	.02	-.04	.01	-.10	.03
3485	-.46	.05	-.03	.02	-.04	.06	-.02	.01	.03	-.05	.09	-.04	.01	.02	-.01	.05	.02	-.05	.03
3290	-.44	.01	-.02	-.01	-.04	.03	.05	-.02	.02	.04	-.02	.06	.01	.09	-.05	.02	-.01	.03	-.02
3030	-.47	.07	.01	-.03	.05	-.01	.02	.04	-.09	.06	.05	-.02	.00	.04	-.10	.01	.06	.03	-.02

STATION NUMBER / LAG	ESTIMATED AUTOCORRELATION FUNCTION											TEST FOR WHITE NOISE	
	20	21	22	23	24	25	26	27	28	29	30	Anderson's Test at 5% Level	Portemanteau Lack of Fit Test at 5% with χ^2_{10}
2535	.08	-.08	.04	-.08	.10	-.02	.07	-.12	.10	-.03	-.01	FAILS	FAILS
2695	-.02	.02	.01	-.05	.03	.05	.03	-.05	.04	-.03	-.04	PASSES	FAILS
3795	.02	-.05	.08	-.09	.06	.05	-.04	.00	-.01	-.02	.07	FAILS	FAILS
3805	.04	-.03	.03	-.06	.05	.04	-.03	-.02	.01	.00	.03	FAILS	FAILS
3280	-.01	-.02	.03	-.08	.06	.03	-.03	-.02	.02	.05	-.09	FAILS	FAILS
3445	.00	-.04	.04	-.06	.06	.09	-.05	-.00	-.03	-.01	.05	FAILS	FAILS
2840	.05	-.08	.02	-.02	.06	.01	.01	-.06	.04	.02	-.04	FAILS	FAILS
3245	-.04	.06	-.01	.01	-.02	.04	-.02	.03	-.03	-.00	-.01	FAILS	FAILS
3265	-.04	.04	.03	-.06	.04	.03	-.01	-.06	.02	-.03	-.08	FAILS	FAILS
6120	-.30	.00	.03	-.04	.00	.09	-.10	.04	.03	-.03	-.01	FAILS	FAILS
3655	.05	-.05	.04	-.10	.09	.00	.06	-.04	-.02	-.01	.04	FAILS	FAILS
2750	.04	-.01	.03	-.05	.06	.03	.05	-.03	.01	-.02	.01	FAILS	FAILS
3485	.05	-.05	.02	.03	-.01	-.03	-.01	.03	.05	-.01	.02	FAILS	FAILS
3290	.02	.04	.01	-.03	-.06	-.03	.01	-.03	.02	.01	.04	FAILS	FAILS
3030	.07	.02	.03	-.04	-.01	-.02	-.01	.04	.01	-.02	.01	FAILS	FAILS

TABLE 3-2

PARTIAL AUTOCORRELATION FUNCTION FOR SQUARE ROOT TRANSFORMED MONTHLY RAINFALL SERIES STATIONARIZED BY FIRST LAG DIFFERENCING

STATION	RECORD	ESTIMATED PACF											
NUMBER	LENGTH	LAG = 1	2	3	4	5	6	7	8	9	10	11	12
2535	516	-.41	-.24	-.25	-.10	-.17	-.13	-.16	-.11	-.15	-.15	-.16	-.04
2695	671	-.43	-.24	-.22	-.14	-.14	-.15	-.17	-.12	-.18	-.15	-.18	-.14
3795	685	-.46	-.26	-.24	-.22	-.14	-.12	-.08	-.09	-.15	-.18	-.19	-.10
3805	684	-.45	-.29	-.26	-.22	-.14	-.14	-.09	-.06	-.12	-.19	-.08	-.08
3280	576	-.43	-.20	-.22	-.20	-.13	-.13	-.12	-.13	-.12	-.16	-.14	-.11
3445	660	-.43	-.24	-.23	-.20	-.15	-.10	-.13	-.11	-.14	-.21	-.12	-.13
2840	684	-.41	-.28	-.23	-.17	-.14	-.14	-.10	-.10	-.15	-.13	-.17	-.05
3245	492	-.45	-.30	-.17	-.18	-.15	-.11	-.17	-.16	-.12	-.11	-.13	-.08
3265	528	-.40	-.28	-.14	-.14	-.19	-.09	-.21	-.15	-.16	-.16	-.13	-.11
6120	576	-.50	-.29	-.24	-.13	-.16	-.22	-.07	-.11	-.07	-.11	-.17	-.12
3655	504	-.44	-.24	-.18	-.15	-.20	-.12	-.19	-.09	-.11	-.16	-.14	-.09
2750	492	-.45	-.25	-.23	-.17	-.19	-.17	-.15	-.12	-.16	-.14	-.16	-.08
3485	492	-.46	-.23	-.18	-.20	-.17	-.15	-.14	-.14	-.13	-.15	-.13	-.10
3290	528	-.44	-.25	-.24	-.23	-.19	-.21	-.15	-.12	-.14	-.11	-.13	-.14
3030	468	-.47	-.29	-.25	-.20	-.15	-.17	-.08	-.10	-.15	-.15	-.09	-.06

TABLE 3-3

AUTOCORRELATION FUNCTIONS FOR SQUARE ROOT TRANSFORMED STANDARDIZED MONTHLY RAINFALL SERIES
AND TESTS FOR ARIMA (0,0,0) WHITE NOISE MODEL

STATION NUMBER / LAG	ESTIMATED AUTOCORRELATION FUNCTION																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2535	.14	.03	.00	.01	-.05	-.01	.02	.02	.03	.01	.02	-.07	-.05	-.08	-.03	-.06	-.02	.02	.03
2695	.09	.00	-.04	-.01	.00	.03	.05	.07	.02	.03	.01	.03	.05	.03	.02	-.02	.00	.02	.03
3795	.05	.01	-.03	.05	.08	.04	.02	-.01	-.03	.02	.06	-.04	-.01	-.01	.04	.00	.01	-.05	-.02
3805	.06	.00	-.02	.03	.04	.01	.00	.00	-.07	.00	.08	-.03	-.03	-.01	.06	-.04	-.03	-.03	.04
3280	.12	.03	-.07	.00	.07	.07	.04	.04	.01	.01	.01	-.01	-.03	-.03	.00	-.01	.04	.03	.00
3445	.06	-.02	-.07	.02	.05	.05	.01	.02	-.02	-.00	.05	-.05	.03	.00	.02	.00	.03	-.03	.05
2840	.11	.01	.02	.01	-.01	-.04	.01	-.02	.01	.02	-.02	-.03	-.04	-.03	-.04	-.03	-.03	-.01	.02
3245	.14	-.02	-.01	-.06	-.05	.06	.01	.02	.04	.01	.04	-.02	.02	-.04	.02	-.02	-.02	.02	.00
3265	.14	.01	.01	-.05	-.05	.05	.00	.07	.04	.04	.05	-.03	.01	-.02	.05	-.03	.01	.04	.03
6120	.05	.07	.01	.04	-.04	.01	.12	-.01	.01	.02	.02	-.08	.05	.02	.00	-.03	-.03	.02	.01
3655	.13	.04	-.02	-.05	-.04	.03	.05	.09	.02	-.02	.04	-.02	.02	-.05	.00	-.06	.01	.01	.01
2750	.13	.05	-.04	-.07	-.04	.02	.05	.10	.02	.01	.01	-.01	.07	-.03	.04	-.03	-.04	-.03	.04
3485	.10	.00	.00	-.08	-.04	.02	-.02	.11	.00	.00	.02	-.06	.06	-.02	.05	-.01	-.02	.01	.01
3290	.14	.00	-.02	-.05	-.03	.09	.02	.07	.05	.01	.03	-.03	.00	.00	.03	-.03	.02	.04	.01
3030	.07	.00	-.03	.04	-.01	.04	.11	.01	.08	-.03	.03	-.03	.08	.01	.04	-.01	.06	-.01	.02

STATION NUMBER / LAG	ESTIMATED AUTOCORRELATION FUNCTION											TEST FOR WHITE NOISE HYPOTHESIS	
												Anderson's Test at 5% Level	Portemanteau Lack of Fit Test at 10% with χ^2_{10}
	20	21	22	23	24	25	26	27	28	29	30		
2535	.03	-.01	-.03	-.09	-.07	.00	.03	-.02	.04	-.01	-.07	REJECT	ACCEPT
2695	.01	.03	-.03	-.07	-.04	.07	.04	.01	.04	.00	-.02	DOUBTFUL	ACCEPT
3795	.02	.00	.03	-.08	-.02	.01	-.03	-.01	-.02	-.04	.01	REJECT	ACCEPT
3805	.02	.00	.00	-.06	.00	.03	.00	.01	.01	-.02	-.01	ACCEPT	ACCEPT
3280	.00	.00	.02	-.01	.03	.04	-.02	-.01	.04	.04	-.07	ACCEPT	REJECT
3445	.04	-.01	-.01	-.06	-.01	.06	-.04	-.04	-.04	-.03	-.01	ACCEPT	ACCEPT
2840	.02	-.03	-.02	-.04	-.04	.00	.01	.01	.04	.02	-.07	ACCEPT	ACCEPT
3245	.01	.06	.01	-.03	.02	.06	.05	-.01	-.05	-.03	-.08	ACCEPT	DOUBTFUL
3265	.01	.10	.04	-.02	.01	.09	.06	-.05	-.07	-.03	-.06	REJECT	REJECT
6120	-.05	-.02	-.01	-.04	-.05	.06	-.05	.03	.03	-.03	-.03	ACCEPT	ACCEPT
3655	.01	-.03	-.08	-.11	-.03	.01	.03	-.03	-.03	-.02	-.01	REJECT	REJECT
2750	.02	-.02	-.08	-.10	-.05	.03	-.01	-.04	-.03	-.04	.00	REJECT	REJECT
3485	.01	.07	-.02	-.03	-.01	.08	.04	-.07	-.08	-.05	-.04	REJECT	DOUBTFUL
3290	.00	.06	.02	-.04	.00	.07	.05	-.02	-.04	-.03	-.07	REJECT	REJECT
3030	.06	.00	-.06	-.07	-.07	-.07	.04	.04	.02	.04	.00	ACCEPT	ACCEPT

TABLE 3-4

PARTIAL AUTOCORRELATION FUNCTIONS FOR SQUARE ROOT TRANSFORMED STANDARDIZED MONTHLY RAINFALL SERIES

STATION NUMBER	RECORD LENGTH	LAG = 1	2	3	4	5	6	7	8	9	10	11	12
2535	516	.14	.01	-.01	.01	-.05	.01	.02	.01	.03	.00	.02	-.07
2695	671	.09	-.01	-.00	.00	.00	.03	.05	.06	.01	.04	.01	.03
3795	684	.05	.01	-.06	.06	.07	.04	.02	-.01	-.04	.02	.05	-.05
3805	684	.06	-.01	-.04	.04	.04	.01	.00	.00	-.07	.00	.08	-.04
3280	576	.12	.01	-.01	.01	.07	.05	.03	.04	.01	-.02	.01	-.01
3445	660	.06	-.02	-.03	.03	.04	.05	.01	.03	-.02	.00	.05	-.07
2840	684	.11	.00	.00	.00	-.01	-.04	.01	-.02	.01	.02	-.03	-.03
3245	492	.14	-.04	-.06	-.06	-.03	.07	-.01	.02	.03	.01	.04	-.04
3265	528	.14	-.01	.05	-.05	-.04	.06	-.02	.07	.02	.04	.04	-.04
6120	576	.05	.07	.03	.03	-.05	.01	.13	-.03	.00	-.02	.02	-.07
3655	504	.13	.02	-.04	-.04	-.02	.05	.04	.08	-.01	.02	.05	-.02
2750	492	.13	.03	-.06	-.06	-.02	.03	.05	.08	-.01	.01	.02	.00
3485	492	.10	-.01	.08	-.08	-.02	.02	.02	.11	-.03	.00	.02	-.04
3290	528	.14	-.03	.04	-.04	-.02	.10	.01	.07	.03	-.01	.04	-.04
3030	468	.07	-.01	-.05	.05	-.02	.04	.11	-.01	.09	.02	.02	-.02

Estimated first lag autocorrelation $\hat{\rho}_1$ is substituted into (3.8) to obtain the estimate of the MA component as

$$\hat{\theta} = -\frac{1}{2\hat{\rho}_1} \pm \left\{\frac{1}{4\hat{\rho}_1^2} - 1\right\}^{\frac{1}{2}}. \tag{3.9}$$

Only the solution that satisfies the invertibility condition $|\theta| < 1$, will be taken as the value of $\hat{\theta}$. The residual variance estimate $\hat{\sigma}_a^2$ is then obtained as

$$\hat{\sigma}_a^2 = \frac{\hat{\sigma}_y^2}{1 + \hat{\theta}_1^2} \tag{3.10}$$

For the square root transformed standardized monthly rainfall series ARIMA (1,0,1) model was identified. This model is

$$(1 - \phi_1 B)\, y_t = (1 - \theta_1 B)\, a_t\ . \tag{3.11}$$

Substituting the estimated autocorrelation coefficients into the first two values of the autocorrelation function of ARIMA (1,0,1) $\hat{\phi}$ and $\hat{\theta}$ can be obtained from

$$\hat{\rho}_1 = (1 - \hat{\theta}_1\phi_1)(\phi_1 - \hat{\theta}_1)/(1 + \hat{\theta}_1^2 - 2\phi_1\hat{\theta}_1)$$

$$\hat{\rho}_2 = r_1\hat{\phi}_1 \tag{3.12}$$

where $|\phi_1| < 1$ and $|\theta| < 1$ are the stationarity and invertibility conditions respectively. Once $\hat{\phi}_1$ and $\hat{\theta}_1$ are evaluated $\hat{\sigma}_a^2$ is calculated from

$$\hat{\sigma}_a^2 = [(1 - \hat{\phi}_1^2)\big/(1 + \hat{\theta}_1^2 - 2\hat{\phi}_1\hat{\theta}_1)]\ \hat{\sigma}_y^2\ . \tag{3.13}$$

The initial estimates of the parameters for both ARIMA (0,1,1) and standardized ARIMA (1,0,1) are given in table 3-5.

3.1.2b <u>Parameter Estimates from the Sum of Squares Surface</u>

For N = n+d observations assumed to be generated by an ARIMA (p,d,q) model *Box and Jenkins* (1971) give the Unconditional Log Likelihood Function as

$$L(\bar{\phi},\bar{\theta},\sigma_a) = f(\bar{\phi},\bar{\theta}) - n\ln\sigma_a - S(\bar{\phi},\bar{\theta})/2\sigma_a^2 \tag{3.14}$$

on the assumption that $a_t \sim N(0,\sigma_a^2)$. In (3.14) $f(\bar{\phi},\bar{\theta})$ is

$$f(\bar{\phi},\bar{\theta}) = \ell n\ 2\rho\left|M_n^{(p,q)}\right|^{\frac{1}{2}}$$

where $\left(M_n^{(p,q)}\right)^{-1}\sigma_a^2$ is the n × n covariance matrix of the stationary time series (*Box and Jenkins*, 1971). $S(\bar{\phi},\bar{\theta})$ is the unconditional sum of squares function given as

$$S(\bar{\phi},\bar{\theta}) = \sum_{t=-\infty}^{n} (a_t|\bar{\phi},\bar{\theta},\bar{y})^2\ . \tag{3.15}$$

However, expression (3.15) is just the least squares error criterion. *Box and Jenkins* (1971) state that for moderate and large n (3.14) is dominated by $S(\bar{\phi},\bar{\theta})/2\sigma_a^2$ and therefore the contours of the Log Likelihood function (3.14) can be approximated by the contours of $S(\bar{\phi},\bar{\theta})$. The ARIMA (p,d,q) parameters which are going to maximize the Log Likelihood function are obtained by minimizing the sum of squares of residuals.

Therefore, for a_t being normal, least squares estimates closely approximate the maximum likelihood estimates. The computer program for the estimation of ARIMA (p,d,q) parameters minimized the sum of squares of residuals and therefore used least squares estimation which is independent of the distribution of a_t.

For an ARIMA (p,d,q) process if y_t is the stationary series with $E[y_t] = 0$. That is, if $y_t = \nabla^d X_t$ where X_t is the original series, using the forward shift operator F such that $FX_t = X_{t+1}$, the ARIMA (p,d,q) is written as

$$[y_t] - \sum_{J=1}^{p} \phi_J [y_{t+J}] = [e_t] - \sum_{J=1}^{q} \theta_J [e_{t+J}] \tag{3.16}$$

where $[y_t]$ is the expectation of y_t conditioned on $\overline{\phi},\overline{\theta}$ and $\{y_t\}$, and $[e_t]$ is $E(e_t|\overline{\theta},\overline{\sigma},\omega)$. Since e_{-J}, $J = 0, 1, 2 \ldots$ are independent of $\{y_t\}$, $[e_{-J}] = 0$ for $J = 0, 1, 2, \ldots.$ Then using (3.16) $[\omega_{-J}]$, $J = 1, 2, 3, \ldots$ are backforecasted. Since there is a stationary autoregressive operator, $[y_{-J}]$ at or beyond some time $-T$ can be assumed zero. Then beginning at time $-T$, $[a_t] = E[a_t|\overline{\theta},\overline{\phi},\overline{\omega}]$ for $t = -T + 1, \ldots, 0, 1, \ldots$ are obtained from

$$[a_t] = \sum_{J=1}^{q} \theta_J \; [a_{t-J}] = [y_t] - \sum_{J=1}^{p} \phi_J [y_{t-J}] \tag{3.17}$$

where $[a_t] = 0$ for $t < (-T + 1)$. The unconditional sum of squares $S(\overline{\phi},\overline{\theta})$ is obtained from

$$S(\overline{\phi},\overline{\theta}) = \sum_{t=-T+1}^{n} [a_t]^2 \tag{3.18}$$

For given set of values of $[\phi_1, \ldots, \phi_p]$ and $[\theta_1, \ldots, \theta_q]$ sum of squares contours for the residuals can be plotted to see what combination of the parameter values yields the least sum of squares. In the identification process ARIMA (0,1,1) for the square root transformed then differenced monthly rainfall series and ARIMA (1,0,1) for the square root transformed then standardized monthly rainfall series were identified. The sum of squares surfaces of the residuals for the 15 monthly rainfall series in the state of Indiana were obtained for the ARIMA (1,1,1) and ARIMA (1,0,1) models. ARIMA (1,1,1) model was considered instead of ARIMA (0,1,1) for further analysis. The reason is ARIMA (1,1,1) is a more general form of ARIMA (0,1,1). As an example, the sum of squares surfaces of the residuals of ARIMA (1,0,1) and ARIMA (1,1,1) models for the monthly rainfall series of Station 3030 are shown on tables 3-6 and 3-7. A computer plot of the sum of squares contours of table 3-6 is shown in figure 22. The residual variance corresponding to the least squares estimates is obtained from

$$\hat{\sigma}_a^2 = \frac{1}{n - 2} S(\hat{\phi},\hat{\theta}) \tag{3.19}$$

where $\hat{\phi}$ and $\hat{\theta}$ are least squares estimates of the autoregressive and moving average parameters respectively. For judging the precision of the parameter-estimates the approximate confidence region can be determined by calculating the sum of squares contour bounding this region. This contour is given as (*Box and Jenkins*, 1971)

$$S(\phi,\theta) = S(\hat{\phi},\hat{\theta}) \; \{1 + \chi_\varepsilon^2(2)/n\} \tag{3.20}$$

where $1 - \varepsilon$ is the confidence level and $S(\hat{\phi},\hat{\theta})$ is the minimum sum of squares contour corresponding to the least squares estimates of the parameters. Approximate confidence region at 95% level is calculated and given in table 3-8. In table 3-6 the hatched region shows the approximate 95% confidence region for ARIMA (1,0,1) of the standardized monthly rainfall series at Station 3030. This region includes a large number of combinations of $(\hat{\phi},\hat{\theta})$ since the sum of squares surface is very flat. Such a flat surface actually stresses the need for calculating the sum of squares surface for the residuals of ARIMA (1,0,1) model of the monthly rainfall series, the least squares estimates for ϕ and θ greatly varied from one time series to another.

This is seen in table 3-8. No estimate for Station 6120 could be obtained since the surface was sloping downwards in the direction of the parameter increase and there was no local minimum smaller than the sum of squares value at the boundary $\phi = .5$, $\theta = +.4$ of the search region $-.5 \leq \phi \leq +.5$, $-.5 \leq \theta \leq +.5$. For fine parameter estimation $\hat{\phi} = .99$ and $\hat{\theta} = .99$ were taken as the initial values for the time series at station 6120. In table 3-7 the approximate 95% confidence region for ARIMA (1,1,1) is hatched. The region is quite small showing that the parameter estimates are stable. As is seen from the table 3-7 the least squares estimate for θ is equal to unity, that is, the ARIMA (1,1,1) model for the case of Station 3030 becomes

$$(1 - \phi B)(1 - B)X_t = (1 - B)a_t \tag{3.21}$$

which is equivalent to

$$(1 - \phi B)X_t = a_t \tag{3.22}$$

which is AR(1) model. This same behavior was observed at Stations 2750, 3245, and 3485 besides 3030. For fine estimation $\hat{\theta}$ was taken as .99.

3.1.2c Nonlinear Estimation

Once the approximate least squares estimates for the parameters of ARIMA (p,d,q) model are obtained from the sum of squares surface of the residuals, these estimates are used as the initial values of the nonlinear interative estimation of the parameters which minimizes the sum of squares of residuals. Therefore, the estimation problem reduces to the minimization of

$$S(\overline{\phi},\overline{\theta}) = \sum_{t=1-T}^{n} [a_t]^2$$

where

$$[a_t] = E[a_t|\overline{\phi},\overline{\theta},\overline{y}] \ . \tag{3.23}$$

$[a_t]$'s are nonlinear functions of the model parameters in ARIMA (p,d,q) model due to the presence of the MA component (*Box and Jenkins*, 1971). In order to use linear least squares technique for estimation $[a_t]$ are linearized by expanding $[a_t]$ about $[a_t|\overline{\phi}_0,\overline{\theta}_0,\overline{\omega}]$ in Taylor series and just considering the linear term; that is,

$$[a_t] = [a_t|\phi_0,\theta_0,y] + \sum_{i=1}^{p} (\phi_i - \phi_{i,0}) \ \partial[a_t]/\partial\phi_i\Big|_{\phi_i=\phi_{i,0}} + \sum_{J=1}^{q} (\theta_J - \theta_{J,0}) \ \partial[a_t]/\partial\theta_J\Big|_{\theta_J=\theta_{J,0}} \tag{3.24}$$

where $\overline{\phi}_0,\overline{\theta}_0$ are the vectors of initial parameter estimates for AR and MA components respectively. The partial derivatives are calculated from

$$\partial[a_t]/\partial\phi_i\Big|_{\phi_i=\phi_{i,0}} = \Big\{[a_t|\overline{y}, \phi_{1,0}, \cdots \phi_{i,0} + \delta_i, \cdots \phi_{p,0}, \theta_{1,0}, \cdots \theta_{q,0}]$$

$$- [a_t|\overline{\omega}, \phi_{1,0}, \phi_{1,0}, \theta_{1,0}, \ldots, \theta_{q,0}]\Big\}/\delta_i \tag{3.25}$$

where δ_i is the perturbation (*Box and Jenkins*, 1971).

Defining $\overline{\beta}$ as the vector whose (p+q) elements are the AR and MA parameters $\overline{\phi}$ and $\overline{\theta}$ and defining $x_{i,t}$ as $x_{i,t} = -\partial[a_t]/\partial\beta_i\Big|_{\beta_i=\beta_{i,0}}$, (3.24) becomes

$$a_t = a_{t,0} - (\beta_1 - \beta_{1,0})x_{1,t} - (\beta_2 - \beta_{2,0})x_{2,t} - \cdots - (\beta_{p+q} - \beta_{p+p,0})x_{p+q,t} \ . \tag{3.26}$$

Then,

$$\begin{bmatrix} a_{1-T} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{bmatrix} = \begin{bmatrix} a_{1-T,0} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ a_{n,0} \end{bmatrix} - \begin{bmatrix} x_{1,1-T} & \cdots\cdots\cdots & x_{p+q,1-T} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_{1,n} & \cdots\cdots\cdots & x_{p+q,n} \end{bmatrix} \begin{bmatrix} \beta_1 - \beta_{1,0} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_{p+q} - \beta_{p+q,0} \end{bmatrix}$$

$$A = A_0 - (X) \cdot (\overline{\beta} - \overline{\beta}_0) \tag{3.27}$$

The sum of squares function to be minimized was

$$S(\overline{\beta}) = A^T A$$

$$= \left(A_0^T - (\overline{\beta} - \overline{\beta}_0)^T X^T\right)\left(A_0 - X(\overline{\beta} - \overline{\beta}_0)\right) \quad . \tag{3.28}$$

Minimization of $S(\overline{\beta})$ will yield

$$(\overline{\beta} - \overline{\beta}_0) = (X^T X)^{-1} X^T A_0 \tag{3.29}$$

Since $[a_t]$ was not linear for the parameters of ARIMA (p,d,q), one multiple linear regression trial is not going to produce the least squares estimates. The parameter estimates found at the end of each iteration are substituted as the initial estimates for the next iteration. Once the convergence is reached, the covariance matrix of the least squares estimates of the ARIMA (p,d,q) parameters can be obtained as

$$\begin{aligned} c(\overline{\hat{\phi}},\overline{\hat{\theta}}) &= E[(X^T X)^{-1} X^T A_0][A_0^T X(X^T X)^{-1}] \\ C(\overline{\hat{\phi}},\overline{\hat{\theta}}) &= (X^T X)^{-1} X^T I \hat{\sigma}_a^2 X(X^T X)^{-1} \\ C(\overline{\hat{\phi}},\overline{\hat{\theta}}) &= (X^T X)^{-1}(X^T X)(X^T X)^{-1}\hat{\sigma}_a^2 \\ C(\overline{\hat{\phi}},\overline{\hat{\theta}}) &= (X^T X)^{-1}\, \hat{\sigma}_a^2 \end{aligned} \tag{3.30}$$

The standard errors of the parameter estimates are then obtained from the diagonal elements of the covariance matrix.

The least squares estimates through nonlinear estimation for the parameters of ARIMA (1,1,1) and ARIMA (1,0,1) models are shown on table 3-9. If the estimates from nonlinear estimation are compared with those from the sum of squares surface, the results agree quite well considering the flatness of the sum of squares surface. The results of the parameter estimation for ARIMA (1,1,1) model where $\hat{\theta} \geq .99$ for all cases suggests that differencing was not necessary since with $\hat{\theta} \approx$ 1.00 ARIMA (1,1,1) reduces to AR(1). That is,

$$(1 - \phi B)(1 - B)X_y = (1 - B)a_t \quad \text{is simply}$$

$$(1 - \phi B)X_t = a_t \tag{3.31}$$

which is AR(1) model. However, the sole purpose of experimenting with nonseasonal first lag differencing was for removing the periodic component of the time series. $\hat{\theta} \approx 1$ suggests that ARIMA (1,1,1) is not a suitable model for monthly rainfall series. Since the sum of squares surface of the residuals of ARIMA (1,0,1) model for standardized monthly rainfall series is very flat, it is not surprising that the standard errors

TABLE 3-5

PRELIMINARY ESTIMATES FOR MODEL PARAMETERS FOR ARIMA (0,1,1) AND ARIMA (1,0,1) MODELS

		ARIMA (0,1,1)		ARIMA (1,0,1)		
STATION NUMBER	RECORD LENGTH	$\hat{\theta}$	$\hat{\sigma}_a^2$	$\hat{\phi}_1$	$\hat{\theta}_1$	$\hat{\sigma}_a^2$
2535	516	.523	1.280	.15	0	.958
2695	671	.572	1.266	0	.1	.970
3795	684	.670	1.235	.11	.07	.980
3805	684	.626	1.315	0	-.05	.975
3280	576	.572	1.220	.22	.10	.965
3445	660	.572	1.280	-.15	-.30	.958
2840	685	.523	1.327	.04	0	.980
3245	492	.626	1.293	-.15	-.30	.958
3265	528	.500	1.272	.10	-.10	.950
6120	576	.999	.942	.999	.999	--
3655	504	.600	1.228	.20	.10	.970
2750	492	.626	1.230	.50	.40	.967
3485	492	.670	1.245	0	-.10	.970
3290	528	.600	1.193	0	-.12	.968
3030	468	.700	1.244	0	-.02	.980

TABLE 3-6

SUM OF SQUARES SURFACE FOR THE RESIDUALS OF ARIMA (1,0,1) MODEL FOR STANDARDIZED MONTHLY RAINFALL SERIES AT STATION 3030. APPROXIMATE 95% CONFIDENCE REGION IS HATCHED.

		-.5	-.4	-.3	-.2	-.1	0	.1	.2	.3	.4	.5
	-.5	456.000	455.853	467.086	489.698	523.689	569.060	625.810	693.940	773.449	864.337	966.605
	-.4	467.372	456.000	454.888	464.036	483.444	513.112	553.041	603.229	663.678	734.387	815.356
	-.3	488.255	467.333	456.000	454.255	462.098	479.529	506.549	543.157	589.353	645.137	710.510
	-.2	517.627	487.867	467.325	456.000	453.893	461.003	477.330	502.875	537.638	581.618	634.816
	-.1	555.748	517.203	487.730	467.329	456.000	453.743	460.559	476.446	501.406	535.438	578.542
$\hat{\theta}$	0	603.946	556.122	517.416	487.827	467.355	456.000	453.763	460.642	476.639	501.754	535.985
	+.1	664.732	606.479	557.772	518.310	488.194	478.424	456.000	453.922	461.189	477.803	503.762
	+.2	742.279	672.054	611.605	560.932	520.035	488.914	467.569	456.000	454.207	462.189	479.948
	+.3	843.518	758.482	683.902	619.779	566.111	522.899	490.144	467.844	456.000	454.612	463.680
	+.4	980.605	876.378	783.635	702.377	632.603	574.314	527.509	492.188	468.352	456.000	455.133
	+.5	1176.828	1045.972	928.176	823.441	731.767	653.154	587.601	535.110	495.679	469.309	456.000

TABLE 3-7

SUM OF SQUARES SURFACE FOR THE RESIDUALS OF ARIMA (1,1,1) MODEL FOR STANDARDIZED MONTHLY RAINFALL SERIES AT STATION 3030. APPROXIMATE 95% CONFIDENCE REGION IS HATCHED

		-.5	-.4	-.3	-.2	-.1	0	.1	.2	.3	.4	.5
	.0	669.093	669.584	686.854	720.903	771.729	839.334	923.718	1024.879	1142.820	1277.538	1429.035
	.1	657.819	649.560	656.712	679.265	717.220	770.576	839.334	923.494	1023.055	1138.017	1268.481
	.2	649.554	633.874	632.457	645.305	762.416	713.791	679.431	839.334	923.501	1021.933	1134.628
	.3	643.694	621.567	612.750	617.241	635.042	666.151	710.570	768.297	839.334	923.680	1021.335
	.4	639.715	611.907	596.597	593.783	603.466	625.646	660.323	707.496	767.167	839.334	923.998
$\hat{\theta}$	.5	637.028	604.174	583.116	573.855	576.391	590.723	616.852	654.778	704.500	766.019	839.334
	.6	634.751	597.443	571.316	556.370	552.606	560.023	578.621	608.401	649.362	701.505	764.829
	.7	631.558	590.439	559.944	540.074	530.828	532.207	544.209	566.836	600.087	643.962	698.462
	.8	626.248	582.027	547.908	523.891	509.976	506.163	512.452	528.843	555.336	591.932	638.629
	.9	620.929	573.973	536.642	508.934	490.850	482.390	483.555	494.343	514.746	544.792	584.452
	1.0	707.177	634.120	574.371	527.442	493.249	471.783	463.049	467.048	483.786	513.234	555.161

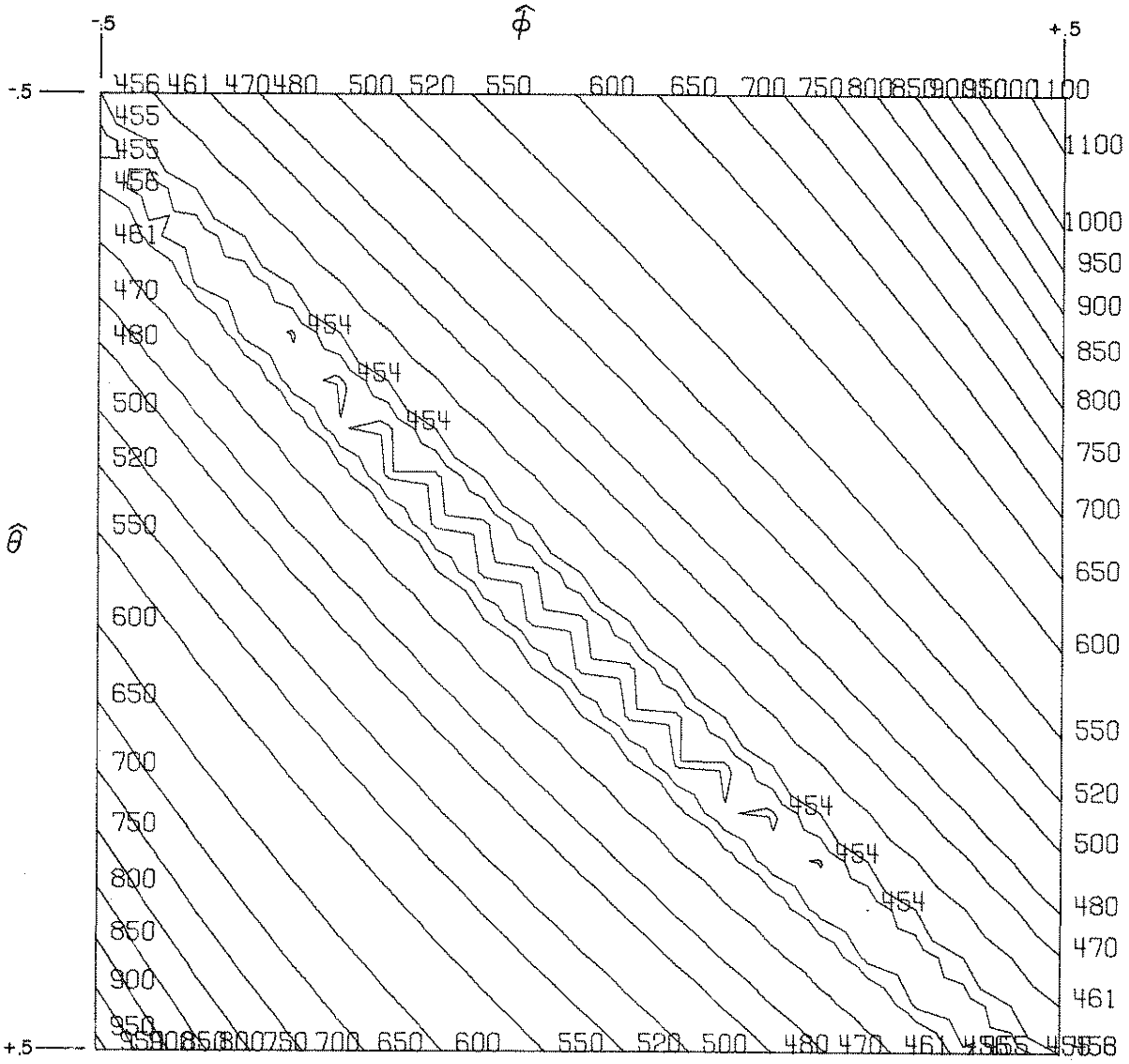


FIG. 22 SUM OF SQUARES SURFACE FOR THE RESIDUALS OF ARIMA (1,0,1) FITTED TO THE STANDARDIZED MONTHLY RAINFALL SERIES AT STATION 3030

TABLE 3-8

PARAMETER ESTIMATES FOR ARIMA (1,1,1) AND ARIMA (1,0,1) BASED ON SUM OF SQUARES SURFACE $S(\hat{\phi},\hat{\theta}) = \Sigma a_t^2(\hat{\phi},\hat{\theta})$

STATION NUMBER	RECORD LENGTH	$\hat{\phi}$	$\hat{\theta}$	$\hat{\sigma}_a^2$	$S(\hat{\phi},\hat{\theta})$	$S(\phi,\theta)$ FOR APPROX. 95% CONFIDENCE REGION	$\hat{\phi}$	$\hat{\theta}$	$\hat{\sigma}_a^2$	$S(\hat{\phi},\hat{\theta})$	$S(\phi,\theta)$ FOR APPROX. 95% CONFIDENCE REGION
2535	516	.1	.9	1.015	530.79	537	.2	.1	.960	494.85	500
2696	671	.1	.9	1.020	686.31	692	0	-.1	.965	654.53	658
3795	684	.1	.9	1.030	711.38	716	0	0	.990	672.00	676
3805	684	0	.9	1.050	716.39	722	0	.1	.984	671.12	676
3280	576	.1	.9	1.001	585.60	589	.2	.1	.905	556.08	579
3445	660	.1	.9	1.023	682.61	687	0	-.1	.970	646.95	660
2840	684	.1	.9	1.011	709.25	714	.1		.973	664.18	667
3245	492	.1	.99	1.040	488.32	494	-.2	-.3	.942	470.86	474
3265	528	.2	.9	1.001	534.81	539	0	-.1	.948	506.47	519
6120	576	0	.9	1.025	597.45	603					
3655	504	.1	.9	1.020	514.15	518	.3	.2	.960	483.83	487
2750	492	.1	.99	1.010	495.77	502	.2	.1	.960	472.19	477
3485	492	.2	.99	1.015	478.91	484	-.1	-.2	.970	475.29	479
3290	528	.2	.9	1.000	533.14	540	-.1	-.2	.960	506.18	511
3030	468	0	.99	1.030	463.05	470	0	-.1	.955	453.74	457

TABLE 3-9

FINAL LEAST SQUARES ESTIMATES OF THE PARAMETERS FOR ARIMA (1,1,1) AND ARIMA (1,0,1) MODELS

		ARIMA (1,1,1) MODEL				ARIMA (1,0,1) MODEL		
STATION NUMBER	RECORD LENGTH	$\hat{\phi}$ ± SE	$\hat{\theta}$ ± SE	$\hat{\sigma}_a^2$	Q(23D.O.F.)	$\hat{\phi}$ ± SE	$\hat{\sigma}_a^2$	Q(23D.O.F.)
2535	516	.188 ± .04	.998 ± 0	.971	39.17	.189 ± .33	.960	15.81
2695	671	.159 ± .04	.996 ± 0	.978	50.86	.04 ± .46	.977	15.09
3795	684	.105 ± .04	.991 ± 0	.997	35.40	.132 ± 1.0	.983	21.33
3805	684	.084 ± .04	.994 ± 0	1.000	24.29	.016 ± .81	.982	19.53
3280	576	.259 ± .05	.990 ± 0	.983	36.23	-.169 ± .37	.968	12.95
3445	660	.147 ± .04	.995 ± 0	.984	51.79	.055 ± .79	.982	18.12
2840	684	.156 ± .04	.996 ± 0	.980	32.60	-.099 ± .38	.974	8.97
3245	492	.103 ± .04	.992 ± 0	.994	24.71	.107 ± .30	.983	12.47
3265	528	.203 ± .04	.995 ± 0	.966	45.60	-.011 ± .32	.960	21.22
6120	576	.06 ± .04	.997 ± 0	1.000	18.67	.756 ± .30	.975	23.84
3655	504	.171 ± .04	.998 ± 0	.976	29.21	.248 ± .35	.963	20.53
2750	492	.176 ± .04	.999 ± 0	.982	22.63	.242 ± .36	.963	23.04
3485	492	.166 ± .01	.999 ± 0	.973	45.46	.044 ± .49	.969	22.05
3290	528	.214 ± .04	.996 ± 0	.961	45.82	-.013 ± .31	.960	16.48
3030	468	.109 ± .04	.999 ± 0	.994	16.89	-.012 ± .13	.970	23.40

χ_{23}^2 (95%) = 35.2 χ_{23}^2 (90%) = 32.0

of the parameter estimates of ARIMA (1,0,1) are large. Therefore, ARIMA (1,0,1) parameter values are very unstable and there is a large number of parameter combinations which will be inside the 95% confidence region as was seen in table 3-6 for Station 3030.

3.1.3 Diagnostic Checking

The two tests applied for the diagnostic checking of the fitted models were the Portemanteau Lack of Fit test and Cumulative Periodogram test applied to the residuals.

Portemanteau Lack of Fit test was utilized by *Box and Pierce* (1970) as an approximate test of the model adequacy. Considering the autocorrelation function of the residuals $\hat{\rho}_k(\hat{a})$ of the fitted ARIMA model and taking L lags such that the weights $\psi(B)$ in the model

$$y_t = \psi(B)a_t$$

will be small after the lag L, *Box and Pierce* (1970) showed that if the model is appropriate, the statistic Q such that

$$Q = n \sum_{k=1}^{L} \hat{\rho}_k^2(\hat{a})$$

is approximately distributed as $x^2(L-p-q)$. One can check the model adequacy by comparing it with the theoretical chi-square value for (L-p-q) degrees of freedom.

Cumulative Periodogram test is devised mainly for the detection of the periodicity in the residuals. The spectrum of white noise is $2\sigma_a^2$ if only the positive side of frequency axis 0 to ½ is considered. Then the cumulative spectral distribution of white noise will be a straight line from (0,0) to (.5,1). *Box and Jenkins* (1971) showed that for $I(f_i)$ defined as

$$I(f_i) = (2/n)\left[\left(\sum_{t=1}^{n} a_t \text{ Cos } 2\pi f_i t\right)^2 + \left(\sum_{t=1}^{n} a_t \text{ Sin } 2\pi f_i t\right)^2\right],$$

$(1/n) \sum_{i=1}^{J} I(f_i)$ is an unbaised estimate of cumulative spectrum. Therefore, the normalized cumulative periodogram $C(f_J)$ such that

$$C(f_J) = \frac{\sum_{i=1}^{J} I(f_i)}{ns^2}$$

is the estimate of spectral distribution. For a white noise series, the plot of $C(f_J)$ versus f_J will be scattered about a line joining (0,0) to (.5,1). *Box and Jenkins* (1971) state that "... periodicities in the a's would tend to produce a series of neighboring values of $I(f_J)$ which were large. These large ordinates would reinforce each other in $C(f_J)$ and form a bump on the expected straight line." The non-parametric *Kolmogorov-Smirnov* test can be used to test whether a_t series is white noise or not. If the straight line joining (0,0) to (.5,1) is denoted by S(f), then the *Kolmogorov-Smirnov* Statistic K is defined as

$$K = \max |S(f) - C(f_J)|$$

where K has a sampling distribution which was discussed in the part I of the report. Taking a certain confidence level ε, K_ε, such that

$$P[|S(f) - C(f_J)| \geq K_\varepsilon] = \varepsilon$$

can be found from tables. If $K < K_\varepsilon$, the hypothesis that the residuals of the fitted model is white noise is accepted at ε level.

Portemanteau Lack of Fit test is applied to ARIMA (1,1,1) and ARIMA (1,0,1) models taking L = 25 lags and considering significance levels of 10% and 5%. The Cumulative Periodogram test is applied to the above models with 95% confidence level. Results are shown on Table 3-10. ARIMA (1,0,1) model fitted to the standardized monthly rainfall series passes the Portemanteau Lack of Fit test at 5% and 10% levels in all cases. ARIMA (1,1,1) model passes the test only in 7 out of 15 cases.

ARIMA (1,0,1) passes the Cumulative Periodogram test in 14 out of 15 cases at 5% level while ARIMA (1,1,1) passes the Cumulative Periodogram test in 13 out of 15 cases. Therefore, standardization of time series effectively removes the periodic component of the monthly rainfall series. As a result of the applied tests, ARIMA (1,0,1) model fitted to the square root transformed standardized monthly rainfall series is an adequate model for the hydrologic phenomenon of monthly rainfall.

TABLE 3-10

DIAGNOSTIC CHECKS ON RESIDUALS

		ARIMA (1,1,1)		ARIMA (1,0,1)	
STATION NUMBER	RECORD LENGTH	PORTEMANTEAU STATISTIC Q	KOLMOGOROV-SMIRNOV STATISTIC K	PORTEMANTEAU STATISTIC Q	KOLMOGOROV-SMIRNOV STATISTIC K
2535	516	39.17	.045	15.81	.040
2695	671	50.86	.045	15.09	.045
3795	684	35.40	.055	21.33	.035
3805	684	24.29	.035	19.53	.040
3280	576	36.23	.030	12.95	.025
3445	660	51.79	.025	18.12	.055
2840	684	32.60	.030	8.97	.050
3245	492	24.71	.035	12.47	.030
3265	528	45.60	.040	21.22	.035
6120	576	18.67	.025	23.84	.035
3655	504	29.21	.025	20.53	.025
2750	492	22.63	.045	23.04	.050
3485	492	45.46	.090	22.05	.095
3290	528	45.82	.090	16.48	.040
3030	468	16.89	.030	23.40	.035

$Q = n \sum_{k=1}^{25} \hat{\rho}_k(a)$ $\qquad X_{23}^2$ (90%) = 32.0 X_{23}^2 (95%) = 35.2

$K = \max |S(f) - C(f)|$ $\qquad K_{.05} = \frac{1.36}{\sqrt{q}}$, $K_{.10} = \frac{1.22}{\sqrt{q}}$, $q = \left\lfloor \frac{n-1}{2} \right\rfloor$

3.2 ANALYSIS OF ARMA (1,1) AS A MODEL FOR THE STATIONARIZED MONTHLY RAINFALL SERIES IN INDIANA

The ARMA (1,1) model was successfully fitted to the standardized monthly rainfall series at various stations in Indiana in the previous section. In view of the hydrologic significance of the ARMA (1,1) model, its general properties need to be discussed in detail. In this section the stationarity conditions, the autocorrelation structure, the spectrum and the impulse response characteristics of the ARMA (1,1) model will be analyzed. In section 3.3, the long-range dependence characteristics of the ARMA (1,1) model were analyzed. The reader can refer to that section for the dependence, spectral and variance-time behavior of the ARMA (1,1) model for long time lags.

3.2.1 Stationarity and Invertibility Conditions

The ARMA (1,1) model is given as

$$y_t - \phi_1 y_{t-1} = a_t - \theta_1 a_{t-1} \tag{3.32}$$

or

$$(1 - \phi_1 B) y_t = (1 - \theta_1 B) a_t \tag{3.33}$$

The stationarity and the invertibility conditions for a general ARMA (p,q) model were analyzed in section 2.1. Since the root of $(1 - \phi_1 B) = 0$ should be outside the unit circle, $1/|\phi_1| > 1$ and $|\phi_1| < 1$. For hydrologically realizable generation invertibility of $(1 - \theta_1 B)$ is sought. This is possible for $|\theta_1| < 1$ since the power series expansion of $1/(1 - \theta_1 B)$ converges only for this region. Therefore, the parameter region of ARMA (1,1) is

$$-1 < \phi_1 < 1, \quad -1 < \theta_1 < 1.$$

3.2.2 Autocorrelation Structure

Box and Jenkins (1970) derive the autocovariance function of the ARMA (1,1) model for lags 0, 1, and u as

$$R_0 = \frac{1 + \theta_1^2 - 2\phi_1\theta_1}{1 - \phi_1^2} \sigma_a^2 \tag{3.34a}$$

$$R_1 = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 - \phi_1^2} \sigma_a^2 \tag{3.34b}$$

$$R_u = \phi_1 R_{u-1} = \phi_1^{u-1} R_1, \qquad u \geq 2 \tag{3.34c}$$

The autocovariance function decays (for the parameter space of stationary, realizable ARMA (1,1)) either exponentially as for positive ϕ_1 or in an oscillatory manner for negative ϕ_1.

The autocorrelation function is given by *Box and Jenkins* (1970) as

$$\rho_1 = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 + \theta_1^2 - 2\theta_1\phi_1} \tag{3.35a}$$

and

$$\rho_u = \phi_1^{u-1}, \qquad u \geq 2 \tag{3.35b}$$

The behavior of ρ_u in relation to the values of ϕ_1 and θ_1 was investigated by *O'Connell* (1971). He stated that the low frequency or long-range effects can be preserved by the proper selection of the value of ϕ_1

while the high frequency effects are controlled by the value of ρ_1. However, it was seen in section 3.3 that ARMA (1,1) cannot preserve the long-range variance time properties of *Hurst's* observations. ϕ_1 is the only parameter affecting the rate of the autocorrelation decay. As ϕ_1 approaches unity the decay rate of ρ_u decreases. However, as it was seen in section 2.2, when ϕ_1 becomes unity, the ARMA (1,1) model, in its generating form, yields infinite variance and generates erratic values.

3.2.3 The Spectrum of the ARMA (1,1)

The spectrum for an ARMA (1,1) model is given as (*Jenkins and Watts*, 1968)

$$S(\omega) = \sigma_a^2 \frac{|1 - \theta_1 e^{-i2\pi\omega}|^2}{|1 - \phi_1 e^{-i2\pi\omega}|^2} = \sigma_a^2 \frac{1 + \theta_1^2 - 2\theta_1 \text{ Cos } 2\pi\omega}{1 + \phi_1^2 - 2\phi_1 \text{ Cos } 2\pi\omega}, \quad -\tfrac{1}{2}\delta < \omega < \tfrac{1}{2}\delta \tag{3.36}$$

At $\omega = 0$, that is, at the spectral origin

$$S(0) = \sigma_a^2 \frac{(1 - \theta_1)^2}{(1 - \phi_1)^2} \qquad -1 < \theta_1 < 1, \quad -1 < \phi_1 < 1. \tag{3.37}$$

3.2.4 Impulsive Response Behavior of the ARMA (1,1)

The ARMA (1,1) model is basically a linear filter the transfer function of which is $(1 - \theta_1 B)/(1 - \phi_1 B)$ with a white noise input a_t and an output y_t. The memory of the ARMA (1,1) model can be analyzed by studying the behavior of its impulsive response h_k and by considering its Fourier transform $H(\omega)$, which is related to the spectrum by

$$S(\omega) = |H(\omega)|^2, \qquad -\tfrac{1}{2}\delta \leq \omega \leq \tfrac{1}{2}\delta$$

Taking the Fourier transform of both sides in eqn. (3.32),

$$\sum_{t=0}^{\infty} y_t e^{-i2\pi\omega t} - \phi \sum_{t=1}^{\infty} y_{t-1} e^{-i2\pi\omega t} = \sum_{t=0}^{\infty} a_t e^{-i2\pi\omega t} - \theta \sum_{t=1}^{\infty} a_{t-1} e^{-i2\pi\omega t}$$

Letting $\ell = t-1$, and denoting the discrete Fourier transform of y_t by $y(\omega)$, and that of a_t by $a(\omega)$, the previous equation becomes

$$y(\omega) - \phi \sum_{\ell=0}^{\infty} y_\ell e^{-i2\pi\omega(\ell+1)} = a(\omega) - \theta \sum_{\ell=0}^{\infty} a_\ell e^{-i2\pi\omega(\ell+1)}$$

or

$$y(\omega) - \phi\, e^{-i2\pi\omega}\, y(\omega) = a(\omega) - \theta\, e^{-i2\pi\omega}\, a(\omega)$$

or

$$y(\omega) = \frac{1 - \theta e^{-i2\pi\omega}}{1 - \phi e^{-i2\pi\omega}} = a(\omega)$$

from which

$$H(\omega) = \frac{1 - \theta e^{-i2\pi\omega}}{1 - \phi e^{-i2\pi\omega}} = \sum_{J=0}^{\infty} h_J e^{-i2\pi\omega J}.$$

Letting $P = e^{-i2\pi\omega}$, the previous equation becomes

$$\sum_{J=0}^{\infty} h_J P^J = \frac{1 - \theta P}{1 - \phi P} = \sum_{J=0}^{\infty} (1 - \theta P)\, \phi^J P^J$$

Expanding both sides of the above equation, one obtains

$$1 - \theta P - \phi P - \theta\phi P^2 + \phi^2 P^2 - \theta\phi^2 P^3 + \phi^3 P^3 - \theta\phi^3 P^4 + \phi^4 P^4 = h_0 + h_1 P + h_2 P^2 + h_3 P^3 + h_4 P^4 + \cdots \quad (3.38)$$

Equating the coefficients of P in (3.38),

$$\begin{aligned} h_0 &= 1 \\ h_1 &= -\theta \\ h_2 &= -\theta\phi + \phi^2 \\ h_3 &= -\theta\phi^2 + \phi^3 \\ h_k &= -\theta\phi^{k-1} + \phi^k \end{aligned} \quad (3.39)$$

Therefore, the impulsive response function of the ARMA (1,1) model is expressed as

$$h_k = 1 \qquad , \quad k = 0 \quad (3.40a)$$

$$= (-\theta\phi^{k-1} + \phi^k) \ , \quad k \geq 1 \quad (3.40b)$$

for $|\phi| < 1$, $|\theta| < 1$. This result leads to the casting of the ARMA (1,1) model in the form

$$y_t = a_t + \sum_{k=1}^{\infty} (-\theta\phi^{k-1} + \phi^k) \, a_{t-k} \quad (3.41)$$

For ϕ and θ significantly different from +1 the impulsive response function decays exponentially and the model has a finite memory. The rate of decay is dependent on both parameters ϕ and θ. However, ϕ is really the more important parameter since it is exponentiated to the power equivalent to the lag. If ϕ is positive, the decay will be slow and the memory will be longer for $\theta < 0$ than for $\theta > 0$. For ϕ negative the decay will be oscillatory.

3.3 APPLICATION OF SEASONAL MODELS TO MONTHLY RAINFALL DATA

It was shown earlier that 12-lag differencing removes the annual cycle in the monthly hydrologic time series. However, 12-lag differencing was shown to introduce periodicities to the continuous spectral density. These periodicities can be seen in figure 20 of the section 2.2 on differencing. However, these periodicities in the spectral density are not due to any discrete spectral components created by certain natural cycles.

Such a spectral density implies that there is a 12-lag relation in the series as well as a first lag type relation. 12-lag relation has the physical meaning that a particular month of this year is related to the same months in other years. For the 12-lag differenced series *Box and Jenkins* (1971) propose the model

$$\phi\ (B^{12})\ \nabla_{12}\ X_t = \Theta_a\ (B^{12})\ \alpha_t \quad (3.42)$$

to explain the dependence of the observation X_t taken at the particular month to the observations taken at the same month during previous years. Since there are 12 months, there will be 12 such models for each month. In (3.42) the symbol ∇_{12} means $(1 - B^{12})$ and $\phi_P\ (B^{12})$ is the seasonal autoregressive operator of P-th degree and $\Theta_a(B^{12})$ is the seasonal moving average operator of Q-th degree.

Although the observation of the monthly rainfall at a certain month, say May, is related to previous May rainfalls, it is also related to the other monthly rainfalls during the same year. Then "the error components, α_t, α_{t-1}, ..., in these models would not in general be uncorrelated," and "we would expect that α_t would be related to α_{t-1} and to α_{t-2}, etc." (*Box and Jenkins*, 1971). To take care of the serial correlation within the months *Box and Jenkins* (1971) introduced the model

$$\phi(B)\ \nabla^d\ \alpha_t = \theta(B)\ a_t \tag{3.43}$$

which explains the correlation among α_t, α_{t-1}, ... In expression (3.43) $\phi(B)$ is the autoregressive operator of p-th degree, $\theta(B)$ is the moving average operator of q-th degree, $\nabla^d = (1 - B)^d$ and a_t is white noise. Assuming that the parameters Φ and Θ obtained for each month are approximately the same, *Box and Jenkins* (1971), by combining (3.43) and (3.42), arrived at the general multiplicative model

$$\phi(B)\ \Phi_p(B^{12})\ \nabla^d \nabla_{12} X_t = \theta(B)\ \Theta_Q(B^{12}) a_t \tag{3.44}$$

for 12 month seasonality. This is the general seasonal model which was applied to monthly rainfall process. The ARIMA $(1,1,1)_{12}$ and the ARIMA $(1,0,0)\ X(1,1,1)_{12}$ models were successfully fitted to monthly rainfall series. Their autocovariance structures will be analyzed for identification purposes.

3.3.1 Seasonal ARIMA $(1,1,1)_{12}$ Model

This model is

$$(1 - \Phi B^{12})\ \nabla_{12} X_t = (1 - \Theta B^{12}) a_t \tag{3.45}$$

where a_t is white noise, $B^{12} X_t = X_t - X_{t-12}$ and $\nabla_{12} = 1 - B^{12}$. ARIMA $(1,1,1)_{12}$ is the short notation meaning that the series have 12 lag seasonality shown by subscript 12, have one seasonal AR parameter shown by the first 1 in the parentheses, are once 12 lag differenced shown by the second 1 in the parentheses, and have one seasonal MA parameter shown by the last 1 in the parentheses.

Denote $\nabla_{12} X_t$ by y_t. The y_t series is a stationary series. Expression (3.45) can be rewritten as

$$y_t = \Phi y_{t-12} + a_t - \Theta a_{t-12} \tag{3.46}$$

Multiplying both sides by y_{t-k} one obtains

$$y_{t-k}\ y_t = \Phi y_{t-k}\ y_{t-12} + y_{t-k}\ a_t - \Theta y_{t-k}\ a_{t-12}. \tag{3.47}$$

Denoting $\mathrm{Cov}[y_{t-k}\ y_t]$ by $\gamma(k)$ and $\mathrm{Cov}(y_{t-k}\ a_t)$ by $\gamma_{ya}(k)$ and remembering that $E[y_t] = 0$, the covariance structure is calculated as follows:

$$\gamma(k) = \Phi\gamma(k - 12) + \gamma_{ya}(k) - \Theta\gamma_{ya}(k - 12). \tag{3.48}$$

and

$$y_t\ a_{t+k} = \Phi y_{t-12}\ a_{t+k} + a_t\ a_{t+k} - \Theta a_{t-12}\ a_{t+k}$$

$$\begin{aligned} \gamma_{ya}(k) &= 0 & k &\geq 1 \\ &= \sigma_a^2 & k &= 0 \end{aligned} \tag{3.49}$$

$$y_t\ a_{t+k-12} = \Phi y_{t-12}\ a_{t+k-12} + a_t\ a_{t+k-12} - \Theta a_{t-12}\ a_{t+k-12}$$

$$\begin{aligned} \gamma_{ya}(k - 12) &= (\Phi - \Theta)\sigma_a^2 & k &= 0 \\ &= \sigma_a^2 & k &= 12 \\ &= 0 & &\text{otherwise} \end{aligned} \tag{3.50}$$

Then,

$$\begin{aligned} \gamma(k) &= \Phi\gamma(k - 12) + \sigma_a^2 - \Theta(\Phi - \Theta)\sigma_a^2 , && k = 0 \\ &= \Phi\gamma(k - 12) - \Theta\sigma_a^2 , && k = 12 \\ &= \Phi\gamma(k - 12) , && k \geq 13 \\ &= 0 , && 1 \leq k \leq 11 \end{aligned} \tag{3.51}$$

This is the general autocovariance structure of ARIMA $(1,1,1)_{12}$. In particular, from (3.50) and (3.51)

$$\begin{aligned} \gamma(0) &= ((1 - 2\Phi\Theta + \Theta^2)/(1 - \Phi^2))\sigma_a^2 \\ \gamma(12) &= ((1 - \Phi\Theta)(\Phi - \Theta)/(1 - \Phi^2))\sigma_a^2 . \end{aligned} \tag{3.52}$$

The stationarity condition of the model is

$$|\Phi| < 1$$

and invertibility condition of the model is

$$|\Theta| < 1.$$

From hydrologic point of view this model attributes all the significant correlation structure to the serial dependence of the same month on several years and assumes that this serial dependence is same for all the 12 months of the year, although the true correlation structure is distorted by 12 lag differencing.

3.3.2 Seasonal ARIMA $(1,0,0)$ X$(1,1,1)_{12}$ Model

For a monthly rainfall series $\{X_t\}$ this model is fitted in the form

$$(1 - \phi B)(1 - \Phi B^{12})\ \nabla_{12}X_t = (1 - \Theta B^{12})a_t \tag{3.53}$$

From hydrologic point of view this multiplicative model assumes that the serial dependence in the monthly rainfall series cannot be completely explained by ARIMA $(1,1,1)_{12}$ and there should be a correlation structure within the months of the same year. This second structure is introduced by correlating the error terms of ARIMA $(1,1,1)_{12}$ model by a nonseasonal ARIMA (p,d,q) model. In the case of monthly rainfall series a first order autoregressive model was found sufficient to explain this serial dependence.

Denoting the stationary $\nabla_{12}X_t$ by y_t, (3.53) can be written as

$$y_t = \phi y_{t-1} + \Phi y_{t-12} - \phi\Phi y_{t-13} + a_t - \Theta a_{t-12} \tag{3.54}$$

where ϕ is the nonseasonal AR operator, Φ is the seasonal AR operator and Θ is the seasonal MA operator. After some manipulations the general autocovariance structure of this multiplicative model is obtained as

$$\begin{aligned} \gamma(k) &= \phi\gamma(k - 1) + \Phi\gamma(k - 12) - \phi\Phi\gamma(k - 13) + [1 - \Theta(\Phi - \Theta + \phi^{12})]\sigma_a^2 , && k = 0 \\ &= \phi\gamma(k - 1) + \Phi\gamma(k - 12) - \phi\Phi\gamma(k - 13) - \Theta\phi^{12-k}\ \sigma_a^2 , && k = 1, \ldots, 12 \\ &= \phi\gamma(k - 1) + \Phi\gamma(k - 12) - \phi\Phi\gamma(k - 13) , && k \geq 13 \end{aligned} \tag{3.55}$$

Such a covariance structure is quite difficult to identify from a 12-lag differenced sample autocovariance function. The autocovariance structures of seasonal multiplicative models become impractically complicated when there are nonseasonal autoregressive components together with the seasonal components in the model. In such a case the usual procedure for identification and fitting is that first a non-multiplicative seasonal model of the form $(p,d,q)_s$ is identified and fitted, then a diagnostic check on the residuals is performed, and if there is a correlation structure in the residuals, an ARIMA (p,d,q) model is fitted to the residuals and a diagnostic check on the overall multiplicative model is performed.

3.3.3 Fit to the Seasonally Differenced Monthly Rainfall Data

Seasonal ARIMA models were considered on 12 lag differenced monthly rainfall series. Fitting procedure of the seasonal ARIMA models is the same as the procedure for nonseasonal ARIMA models. So as a first step the autocorrelation functions for 12 lag differenced series were obtained. These are shown on Table 3-11. As is seen from the table there is a significant correlation at lag 12 in all the cases. In 8 out of 15 cases there is quite a high correlation at first lag.

The drawback in the identification of the seasonally differenced series is that their general autocorrelation behavior is not thoroughly investigated. The only guide to the identification of a model for the sample autocovariance is the table of autocovariances for some seasonal models given by *Box and Jenkins* (1971). However, in the case of monthly rainfall time series none of the given autocovariances was suitable. The problem of fitting a seasonal model to the 12-lag differenced series became one of trial and error. From the Sections 3.3.1 and 3.3.2, it follows that the seasonal moving average operator Θ can singly explain the significant 12-lag autocovariance. However, to be safe, a seasonal autoregressive component Φ was added to have an overfitted model. Thus the seasonal ARIMA $(1,1,1)_{12}$ model was formed. The most efficient way for checking the model is fitting the model directly through the least squares estimation program with the diagnostic checks built into the program. Since this computer program utilizes a gradient search technique, the optimum model parameters are obtained when the iterations start with any combination of Φ and Θ values inside the region bounded by stationarity and invertibility conditions.

For the general Multiplicative Seasonal ARIMA (p,d,q) $X(P,D,Q)_s$ described as

$$\phi_p(B)\ \Phi_P(B^s)\ \nabla^d\ \nabla_s\ X_t = \theta_q(B)\ \Theta_Q(B^{12})a_t \qquad (3.56)$$

the least squares procedure for parameter estimation is similar to the one of nonseasonal ARIMA (p,d,q). The difference is introduced by the seasonal operators and seasonal differencing. Denoting $\nabla^d\ \nabla_s\ X_t$ by y_t and $(y_t - \bar{y})$ by $\bar{y}_t$ if just the seasonal part of ARIMA (p,d,q) x $(P,D,Q)_s$ is considered, the ARIMA $(P,D,Q)_s$ model will have the open form

$$(1 - \Phi_1\ B_s - \Phi_2\ B_s^2 - \cdots - \Phi_P\ B_s^P)\tilde{y}_t = (1 - \Theta_1\ B_s - \Theta_2\ B_s^2 - \cdots - \Theta_Q\ B_s^Q)\alpha_t \qquad (3.57)$$

where α_t is the residual at time t. The forward shift operator F is used to backforecast $\tilde{y}$'s in the form

$$[\tilde{y}_t] - \sum_{i=1}^{P} \Phi_i[\tilde{y}_{t+is}] + \sum_{J=1}^{Q} \Theta_J[n_{t+Js}] = [n_t] \qquad (3.58)$$

where brackets denote expectations conditioned on ϕ, θ, Φ, Θ, y. Consider that $[n_{-J}] = 0$ for J = 0, 1, 2, ... and $[\alpha_J] = 0$. The backforecasting starts by calculating $[n_{N-d-sD-sP}]$ obtained by setting the unknown $[n]$'s equal to zero. It is terminated at a point -T in time where y_{-T} becomes negligible. Then the form

$$[\tilde{y}_t] - \sum_{i=1}^{P} \Phi_i[y_{t-is}] + \sum_{J=1}^{Q} \Theta_J[\alpha_{t-Js}] = [\alpha_t] \qquad (3.59)$$

is used to calculate the residuals α_t of the proposed seasonal ARIMA $(1,1,1)_{12}$ model. Once the residuals α_t are calculated, the rest of the procedure for least squares parameter estimation is the same as of nonseasonal ARIMA (p,d,q).

For the case of least squares parameter estimation of the multiplicative seasonal ARIMA (p,d,q) x $(P,D,Q)_s$ model, ARMA (p,q) model

$$a_t = \alpha_t - \sum_{i=1}^{p} \phi_i \alpha_{t-1} + \sum_{J=1}^{q} \theta_J a_{t-J} \tag{3.60}$$

will be fitted to dependent $\{\alpha\}$ series. This would yield a two stage procedure for the calculation of residuals a_t of (3.56). First the residuals $\{\alpha\}$ of the seasonal ARIMA $(P,D,Q)_s$ are calculated by (3.58) and (3.59). Then (3.60) is used for obtaining $\{\alpha\}$ series.

In Table 3-12 the results of the fit of seasonal ARIMA $(1,1,1)_{12}$ model to the square root transformed monthly rainfall series by the direct use of least squares estimation program, are given. As can be seen from the results seasonal ARIMA $(1,1,1)_{12}$ model failed to explain the monthly rainfall series at Stations 3265, 3290, and 3485. These were the stations where the first lag autocorrelation coefficient was significant, as can be seen from Table 3-11. The autocovariance structure for seasonal ARIMA $(1,1,1)_{12}$ was derived earlier as

$$\gamma(0) = \frac{1 - 2\Phi\Theta + \Theta^2}{1 - \Phi^2} \sigma_a^2$$

$$\gamma(k) = 0 \qquad\qquad 1 \leq k \leq 11$$

$$\gamma(12) = \frac{(1 - \Phi\Theta)(\Phi - \Theta)}{1 - \Phi^2} \sigma_a^2 \qquad k \geq 13$$

$$\gamma(k) = \Phi\gamma(k - 12)$$

When the estimated autocorrelation functions in Table 3-11 are analyzed it is seen that seasonal ARIMA $(1,1,1)_{12}$ autocovariance structure explains the behavior of the estimated autocorrelation functions except at the first lag. When the first lag correlation coefficient was small ARIMA $(1,1,1)_{12}$ could still pass the Portemanteau Lack of Fit test.

In order to take account of the significant first lag which appeared at the Stations 3265, 3290, and 3485, ARIMA (1,0,0) was fitted to the residuals $\{\alpha\}$ of the seasonal ARIMA $(1,1,1)_{12}$ model. The results of this fit are given on Table 3-13. As is seen from this table, seasonal ARIMA $(1,0,0) \times (1,1,1)_{12}$ satisfactorily fits the square root transformed monthly rainfall series.

3.4 FORECASTING OF THE MONTHLY RAINFALL SERIES IN INDIANA

Although the water resources system design is based on the generated streamflow sequences, the hydrologic forecasting may be an important tool for the modification of the operation policies after the completion of the water resources development project. As the historical streamflow record is extended into the future the forecasting model can be updated and used for the modification of the operation policies to adjust to the gain of the hydrologic information towards a better accomplishment of the stated objectives. Another use for forecasting is for testing the performance of the candidate time series models to be used for generating synthetic hydrologic sequences. In this section, the monthly rainfall sequences will be forecasted by the candidate models and the model performance will be evaluated by comparing the forecasts with the historical hydrologic records.

An observation $z_{t+\ell}$ at time $t+\ell$ can be expressed as

$$z_{t+\ell} = \sum_{J=0}^{\ell-1} \psi_J \, a_{t+\ell-J} + \sum_{J=\ell}^{\infty} \psi_J \, a_{t+\ell-J} \tag{3.61}$$

The second summation on the right side is the minimum mean square error forecast of $z_{t+\ell}$ at the time origin t for the lead time ℓ, denoted by $z_t(\ell)$ (*Box and Jenkins*, 1971). The forecast error for the lead time ℓ, denoted by $e_t(\ell)$, becomes the first summation on the right side of (3.61). Since $E[a_{t+J}] = 0$ for $J \geq \ell$,

TABLE 3-11

ESTIMATED AUTOCORRELATION FUNCTIONS FOR SQUARE-ROOT TRANSFORMED THEN 12-LAG DIFFERENCED MONTHLY RAINFALL SERIES

STATION	RECORD	LAGS														
NUMBER	LENGTH	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2535	516	.15	.04	.01	.03	-.04	-.01	.02	.00	.03	.03	-.01	-.48	-.11	-.10	-.04
2695	671	.04	-.04	-.13	-.03	-.03	-.01	.03	.04	.08	.08	-.01	-.46	.02	-.01	.10
3795	684	.02	.01	-.03	.05	.05	.04	.00	-.04	-.01	-.01	.07	-.50	-.02	.00	.05
3805	684	.02	.00	-.02	.05	.05	.02	.01	-.03	-.05	-.01	.07	-.51	-.05	.00	.06
3280	576	.11	.06	-.09	-.01	.02	.00	-.03	.03	.07	-.04	-.05	-.50	-.03	-.02	.05
3445	660	.02	.00	-.06	.00	.03	.03	-.03	.00	.04	.01	.04	-.51	-.03	.02	.06
2840	684	.13	.01	.05	.03	.00	-.02	.01	-.04	.00	.02	-.05	-.49	-.09	-.04	-.06
3245	492	.30	.12	.04	-.01	.05	.06	.06	.12	.08	-.01	-.06	-.49	-.15	-.03	.00
3265	528	.13	.00	-.03	-.06	-.03	.00	-.01	.08	.01	.01	-.02	-.50	-.12	-.04	.03
6120	576	.01	.06	.01	.06	-.06	-.01	.11	-.01	.02	-.05	.01	-.51	.01	.02	-.03
3655	504	.10	.06	-.01	-.05	-.04	.01	.03	.10	.05	.02	.03	-.48	-.06	-.08	.01
2750	492	.05	.04	-.06	-.09	-.06	.00	.04	.07	.08	.05	.00	-.47	-.01	-.05	.06
3485	492	.39	.18	.08	.00	.04	.05	.04	.12	.04	-.03	-.08	-.42	-.18	.01	.03
3290	528	.31	.14	.06	-.02	.03	.07	.07	.12	.05	-.03	-.06	-.42	-.18	-.04	-.01
3030	468	.00	.00	-.09	.06	-.08	.03	.09	-.05	.10	.01	.07	-.48	-.01	-.03	.06

STATION	RECORD	LAGS														
NUMBER	LENGTH	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
2535	516	-.08	.01	.05	.05	.04	-.04	-.01	-.08	-.01	.01	.08	.00	.02	-.01	-.07
2695	671	-.02	-.02	.04	.02	-.07	-.03	-.08	-.06	-.06	.02	.05	-.03	.00	.04	-.04
3795	684	-.02	-.02	-.06	.00	.04	-.01	.04	-.09	.03	.04	-.03	-.03	-.05	-.02	.03
3805	684	-.06	-.05	-.02	-.01	.02	.00	.01	-.08	.03	.04	-.02	-.03	.01	.02	.00
3280	576	-.02	-.01	.03	.05	-.03	-.02	.02	.00	.03	.02	-.02	.00	.03	.01	-.07
3445	660	-.01	.02	-.05	.08	.01	-.05	.00	-.08	.04	.07	-.04	-.03	-.04	-.05	.01
2840	684	-.05	-.03	.05	.03	.05	-.06	-.01	-.03	-.02	.02	.04	.03	.05	.00	-.07
3245	492	.00	-.02	-.03	-.08	-.10	-.12	-.05	-.07	.04	.06	.00	-.01	-.04	-.04	-.03
3265	528	.02	.02	.03	.07	-.07	.01	.02	-.02	.05	.11	.06	-.04	-.07	-.03	-.05
6120	576	-.07	-.01	.03	-.03	-.03	-.05	.00	-.02	.01	.05	-.06	.04	.03	-.01	-.04
3655	504	.00	.03	-.01	.02	-.05	-.09	-.06	-.11	.01	.00	.05	-.04	-.05	-.02	.00
2750	492	.04	.04	.00	.03	-.01	-.11	-.09	-.10	.01	.01	.04	-.03	-.05	-.01	-.01
3485	492	.05	.01	.00	-.02	-.10	-.11	-.08	-.09	-.04	.04	-.05	-.07	-.09	-.07	-.05
3290	528	.01	-.01	-.04	-.07	-.11	-.12	-.06	-.10	-.01	.07	-.01	-.02	-.04	-.04	-.03
3030	468	-.07	.06	-.04	-.02	.03	-.05	-.01	-.07	.01	-.01	.05	-.02	.06	-.03	.05

TABLE 3-12

RESULTS OF LEAST SQUARES FIT OF SEASONAL ARIMA $(1,1,1)_{12}$
TO SQ-RT TRANSFORMED MONTHLY RAINFALL SERIES

STATION NUMBER	RECORD LENGTH	VARIANCE OF √‾ TRANSFORMED SERIES	SEASONAL ARIMA $(1,1,1)_{12}$ PARAMETER ESTIMATES			STATISTIC Q FOR PORTEMANTEAU LACK OF FIT TEST (28 D.O.F.)
			$\hat{\Phi}$ ± SE	$\hat{\Theta}$ ± SE	$\hat{\sigma}_a^2$	
2535	516	.235	-.044 ± .045	.960 ± .008	.203	28.14
2695	671	.244	-.006 ± .047	.954 ± .008	.222	26.98
3795	684	.291	-.006 ± .047	.960 ± .005	.266	23.38
3805	684	.342	-.051 ± .038	.960 ± .005	.325	19.15
3280	576	.229	-.020 ± .048	.944 ± .009	.204	26.14
3445	660	.279	-.060 ± .039	.960 ± .007	.243	25.57
2840	684	.342	-.051 ± .038	.960 ± .005	.325	19.155
3245	492	.150	.074 ± .045	.953 ± .007	.100	149.32
3265	528	.254	-.028 ± .044	.960 ± .008	.227	35.52
6120	576	.386	-.087 ± .042	.960 ± .007	.375	26.90
3655	504	.270	.008 ± .045	.950 ± .008	.255	33.87
2750	492	.279	.002 ± .050	.951 ± .008	.263	23.32
3485	492	.165	.118 ± .045	.963 ± .007	.110	180.24
3290	528	.765	.060 ± .044	.957 ± .007	.494	164.47
3030	468	.362	-.029 ± .047	.950 ± .008	.348	22.63

χ^2_{28} (90% = 37.9 χ^2_{38} (95%) = 41.3 $Q = n \sum_{i=1}^{30} \hat{\rho}_i^2(\hat{a})$

TABLE 3-13

RESULTS OF LEAST SQUARES FIT OF SEASONAL ARIMA $(1,0,0) \times (1,1,1)_{12}$
TO SQ-RT TRANSFORMED MONTHLY RAINFALL SERIES

STATION NUMBER	RECORD LENGTH	VARIANCE OF √‾ TRANSFORMED SERIES	SEASONAL ARIMA $(1,0,0) \times (1,1,1)_{12}$ PARAMETER ESTIMATES				STATISTIC Q FOR PORTEMANTEAU LACK OF FIT TEST (27 D.O.F.)
			$\hat{\phi}$ ± SE	$\hat{\Phi}$ ± SE	$\hat{\Theta}$ ± SE	$\hat{\sigma}_a^2$	
2535	516	.235	.138 ± .040	-.044 ± .045	.96 ± .008	.200	14.87
2695	671	.244	.078 ± .047	-.013 ± .047	.955 ± .009	.221	22.94
3795	684	.291	.051 ± .040	-.062 ± .040	.960 ± .005	.266	21.95
3805	684	.342	.057 ± .039	-.055 ± .039	.960 ± .005	.325	17.62
3280	576	.229	.143 ± .047	-.032 ± .048	.944 ± .009	.200	17.53
3445	660	.279	.066 ± .040	-.066 ± .040	.958 ± .007	.243	22.71
2840	684	.342	.107 ± .040	-.030 ± .040	.967 ± .006	.195	12.55
3245	492	.150	.387 ± .040	-.005 ± .040	.952 ± .007	.089	26.84
3265	528	.254	.144 ± .043	-.032 ± .040	.959 ± .008	.223	20.61
6120	576	.386	.047 ± .040	-.090 ± .040	.959 ± .007	.375	25.65
3655	504	.270	.129 ± .040	-.004 ± .040	.952 ± .007	.252	20.71
2750	492	.279	.0897± .040	-.004 ± .040	.951 ± .007	.261	17.68
3485	492	.165	.425 ± .040	.074 ± .040	.960 ± .008	.088	30.6
3290	528	.765	.403 ± .040	.003 ± .040	.961 ± .007	.414	30.0
3030	468	.362	.045 ± .047	-.034 ± .047	.951 ± .008	.348	21.95

χ^2_{27}(90% = 36.7 χ^2_{27} (95%) = 40.1 $Q = n \sum_{i=1}^{30} \hat{\rho}_i^2(\hat{a})$

$E[e_t(\ell)] = 0$ and the forecasts are unbiased. The variance of the forecast error is

$$\text{Var}\,[e_t(\ell)] = E[e_t^2(\ell)] = \sum_{J=0}^{\ell-1} \psi_J^2\, \sigma_a^2 \qquad (3.62)$$

Four models were fitted to the square roots of the monthly rainfalls. For the standardized series the white noise or ARIMA (0,0,0) and ARIMA (1,0,1) models were investigated, and for the seasonally 12 lag differenced series ARIMA $(1,1,1)_{12}$ and ARIMA (1,0,0) x $(1,1,1)_{12}$ models were fitted.

The standardized series $\{z_t\}$ have zero mean and unit variance. To retrieve the square root transformed monthly rainfall series which had the seasonal periodic component, a reverse of the standardization procedure is performed. Letting y_t be the rainfall square root at time t

$$\hat{z}_t(\ell) = (\hat{y}_t(\ell) - \bar{y}_j)/s_{y,j} \qquad (3.63)$$

where j is the month in a 12 month annual cycle. Then

$$\hat{y}_t(\ell) = \hat{z}_t(\ell)\; s_{y,j} + \bar{y}_j \qquad (3.64)$$

where $\hat{y}_t(\ell)$ is the forecasted square root transformed periodic monthly rainfall series, $s_{y,j}$ and $\bar{y}_j$ are the standard deviation and the mean of square root transformed rainfall series for the month j. Then the standard error of the forecast $\hat{y}_t(\ell)$ is

$$\hat{\sigma}_{y,t}(\ell) = \hat{\sigma}_{z,t}(\ell)\; s_{y,j} \qquad (3.65)$$

where $\hat{\sigma}_{z,t}(\ell)$ is the standard error of $\hat{z}_t(\ell)$ which can be obtained from (3.62).

The ARIMA (0,0,0) or the white noise model, for the square root transformed standardized monthly rainfall series, has the forecast function

$$\hat{z}_t(\ell) = 0, \qquad \ell \geq 1.$$

The forecasted square root transformed monthly rainfall series is the monthly means. That is,

$$\hat{y}_t(\ell) = \bar{y}_j \qquad (3.66)$$

with the standard error $s_{y,j}$ which is the standard deviation for the month j.

The ARIMA (1,0,1) model was also fitted to square root transformed, standardized monthly rainfall series. The forecasting function and the ψ-weights are given by *Box and Jenkins* (1971).

The seasonal ARIMA $(1,1,1)_{12}$ model was fitted to the square-root transformed and 12th lag differenced monthly rainfall series. The forecast function and the ψ weights are

$$\begin{aligned}
\hat{z}_t(\ell) &= (\Phi + 1)z_{t-12+\ell} - \Phi z_{t-24+\ell} - \Theta a_{t-12+\ell} && \ell = 1, \ldots, 12 \\
\hat{z}_t(\ell) &= (\Phi + 1)z_t(\ell-12) - \Phi z_{t-24+\ell} && \ell = 13, \ldots, 24 \\
\hat{z}_t(\ell) &= (\Phi + 1)z_t(\ell-12) - \Phi z_t(\ell-24) && \ell \geq 25.
\end{aligned} \qquad (3.67)$$

$$\begin{aligned}
\psi_0 &= 1 \\
\psi_{12} &= (\Phi + 1) - \Theta \\
\psi_J &= 0 && 1 \leq J \leq 11 \\
&= (\Phi + 1)\,\psi_{J-12} - \Phi\psi_{J-24} && J > 12
\end{aligned} \qquad (3.68)$$

Since the series were made stationary by 12th lag differencing, the forecast function for square root transformed periodic monthly rainfall series $\hat{y}_t(\ell)$ is simply $\hat{z}_t(\ell)$ and $\hat{\sigma}_{y,t}(\ell)$ is $\hat{\sigma}_{z,t}(\ell)$.

$$\hat{z}_t(1) = \phi z_t + (\Phi + 1)z_{t-11} - (\phi + \phi\Phi)z_{t-12} - \Phi z_{t-23} + \phi\Phi z_{t-24} - \Theta a_{t-11}$$

$$\hat{z}_t(\ell) = \phi\hat{z}_t(\ell - 1) + (\Phi + 1)z_{t+\ell-12} - (\phi + \phi\Phi)z_{t+\ell-12} - \Phi z_{t+\ell-24} + \phi\Phi z_{t+\ell-25} - \Theta z_{t+\ell-12}, \quad \ell = 2, \ldots, 12$$

$$\hat{z}_t(13) = \phi\hat{z}_t(12) + (\Phi + 1)\ \hat{z}_t(1) - (\phi + \phi\Phi)z_t - \Phi z_{t-11} + \phi\Phi z_{t-12}$$

$$\hat{z}_t(\ell) = \phi\hat{z}_t(\ell - 1) + (\Phi + 1)\ \hat{z}_t(\ell - 12) - (\phi + \phi\Phi)\ \hat{z}_t(\ell - 13) - \Phi z_{t+\ell-24} + \phi\Phi z_{t+\ell-25} \quad , \quad \ell = 14, \ldots, 24$$

$$\hat{z}_t(25) = \phi\hat{z}_t(24) + (\Phi + 1)\ \hat{z}_t(13) - (\phi + \phi\Phi)\ \hat{z}_t(12) - \Phi\hat{z}_t(1) + \phi\Phi z_t$$

$$\hat{z}_t(\ell) = \phi\hat{z}_t(\ell - 1) + (\Phi + 1)\ \hat{z}_t(\ell - 12) - (\phi + \phi\Phi)\ \hat{z}_t(\ell - 13) - \Phi\hat{z}_t(\ell - 24) + \phi\Phi\hat{z}_t(\ell - 25),\ \ell \geq 26 \qquad (3.69)$$

The forecast function $\hat{Y}_t(\ell)$ for the square root transformed periodic monthly rainfall series is simply $\hat{z}_t(\ell)$. $\hat{\sigma}_{Y;t}(\ell)$ is $\hat{\sigma}_{z;t}(\ell)$. The ψ weights for the model are

$$\begin{aligned}
\psi_0 &= 1 \\
\psi_J &= \phi^J \\
\psi_{12} &= \phi^{12} + (\Phi + 1) - \Theta & , \quad & J = 1, \ldots, 11 \\
\psi_{13} &= \phi\psi_{12} \\
\psi_J &= \phi\psi_{J-1} + (\Phi + 1)\psi_{J-12} - \phi(1 + \Phi)\psi_{J-13} & , \quad & J = 14, \ldots, 23 \\
\psi_{24} &= \phi\psi_{23} + (\Phi + 1)\psi_{12} - \phi(1 + \Phi)\psi_{11} - \Phi \\
\psi_{25} &= \phi\psi_{24} + (\Phi + 1)\psi_{13} - \phi(1 + \Phi)\psi_{12} \\
\psi_J &= \phi\psi_{J-1} + (\Phi + 1)\psi_{J-12} - \phi(1 + \Phi)\psi_{J-13} - \Phi\psi_{J-24} + \phi\Phi\psi_{J-25}, & & J > 25
\end{aligned} \qquad (3.70)$$

The $(1 - \varepsilon)$ confidence limits for the minimum mean square error forecast $\hat{z}_t(\ell)$ of the actual value $a_{t+\ell}$ are given by *Box and Jenkins* (1971) as

$$z_{t+\ell}\ (\pm) = \hat{z}_t(\ell) \pm u_{\varepsilon/2} \left\{1 + \sum_{J=1}^{\ell-1} \psi_J^2\right\}^{\frac{1}{2}} \hat{\sigma}_a \qquad (3.71)$$

where $u_{\varepsilon/2}$ is the deviate exceeded by $\varepsilon/2$ of the standard normal distribution. These will be the confidence limits for ARIMA (1,1,1), ARIMA $(1,1,1)_{12}$ and ARIMA (1,0,0) x $(1,1,1)_{12}$ models fitted to the square root transformed series, whenever the residuals are normally distributed. For the ARIMA (1,0,1) and ARIMA (0,0,0) with the forecasts expressed as in equation (3.64), the $1 - \varepsilon$ confidence limits for $y_{t+\ell}$ are given by

$$y_{t+\ell}(\pm) = \hat{y}_t(\ell) \pm (\hat{\sigma}_{z,t}(\ell) \cdot s_{y,j})u_{\varepsilon/2}\ . \qquad (3.72)$$

If the actual monthly rainfall data are to be forecasted, denote the monthly rainfall series by $X_{t+\ell}$. It can be shown that

$$\hat{X}_t(\ell) = \{\hat{y}_t(\ell)\}^2 + \{\hat{\sigma}_{Y,t}(\ell)\}^2 \qquad (3.73)$$

and

$$\hat{\sigma}_{X,t}(\ell) = 2\hat{\sigma}_{y,t}(\ell)\ [2\{\hat{y}_t(\ell)\}^2 + \{\sigma_{y,t}(\ell)\}^2]. \tag{3.74}$$

The variance of the differenced models may not reflect the seasonality in the variance of the actual rainfall square roots and consequently the corresponding rainfall forecasts obtained by equation (3.73) may be distorted. The models using the standardized values are free of this bias.

In figure 23 plots of the monthly rainfall square root forecasts, for ARIMA (0,0,0), ARIMA (1,0,1), ARIMA (1,0,0) x $(1,1,1)_{12}$ models and the observed rainfall square roots are given. As can be seen from these plots the forecasts of the three different models closely follow each other. The observed values are within the confidence limits of ARIMA (1,0,0) x $(1,1,1)_{12}$ model except the case of Station 3655. Figure 24 shows the monthly rainfall square root forecasts by ARIMA (0,0,0), ARIMA $(1,1,1)_{12}$ models and the observed rainfall square roots. ARIMA $(1,1,1)_{12}$ model closely follows the monthly means represented by ARIMA (0,0,0) model. The observed values are within the 95% confidence limits except in the case of Station 3655.

In Table 3-14 a comparison of the mean square errors of the monthly rainfall square root forecasts by the four models is shown. There is not a considerable difference among the models with respect to the mean square error.

From the results of forecasting, ARIMA (0,0,0) or the white noise model seems to be the most suitable for monthly rainfall forecasting since it is the simplest model. However, from the parsimony point of view ARIMA (1,0,0) x $(1,1,1)_{12}$ or ARIMA $(1,1,1)_{12}$ models have fewer parameters (3 and 2 respectively) than ARIMA (0,0,0) (24 parameters) but the forecasting functions and the ψ weights are considerably more difficult to determine for the seasonally differenced models.

TABLE 3-14

MEAN SQUARE ERRORS OF MONTHLY RAINFALL SQUARE ROOT FORECASTS
FOR THE LEADS UP TO 24 MONTHS AHEAD

STATION	M.S.E. FOR ARIMA (0,0,0)	M.S.E. FOR ARIMA (1,0,1)	M.S.E. FOR ARIMA $(1,1,1)_{12}$	M.S.E. FOR ARIMA (1,0,0) x $(1,1,1)_{12}$
3805	4.0181	4.0181	3.9438	3.9287
2840	2.5242	2.5164	2.5895	2.5822
3795	4.6190	4.6075	4.5236	4.5139
3445	4.6766	4.6833	4.8005	4.8174
2750	5.5863	5.5875	5.6947	5.6822
2695	1.9149	1.9141	1.8231	1.8274
3030	4.6694	4.6809	4.6631	4.6772
6120	4.3317	4.3245	4.5283	4.5252
3655	4.9881	5.049	4.9059	4.9591
2535	2.3104	2.3152	2.2675	2.1977
3280	5.086	5.0898	5.0766	5.0737
3265	1.8771	1.8663	2.0246	2.0647

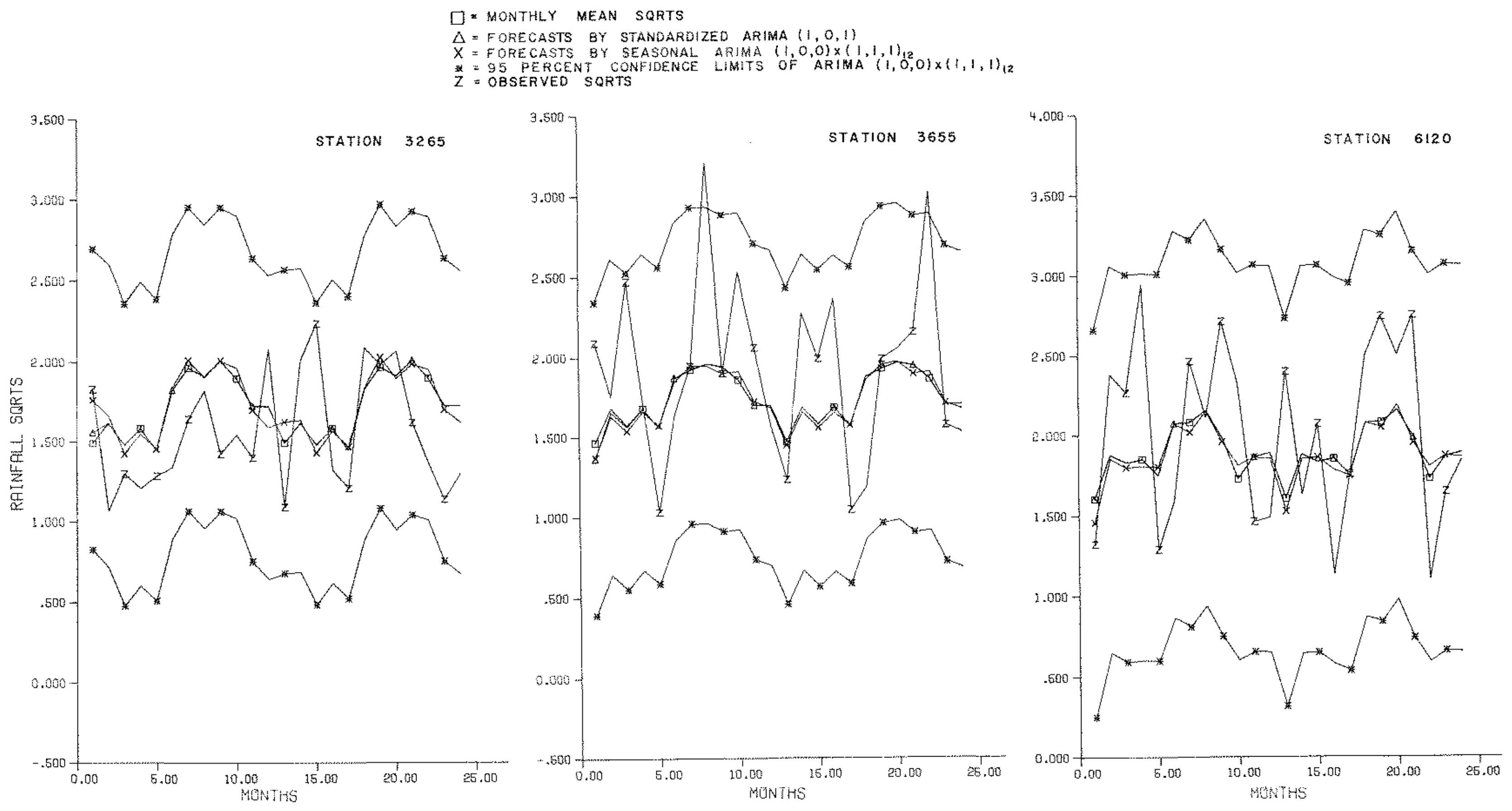


FIG. 23 MONTHLY RAINFALL SQUARE ROOT FORECASTS BY VARIOUS ARIMA MODELS

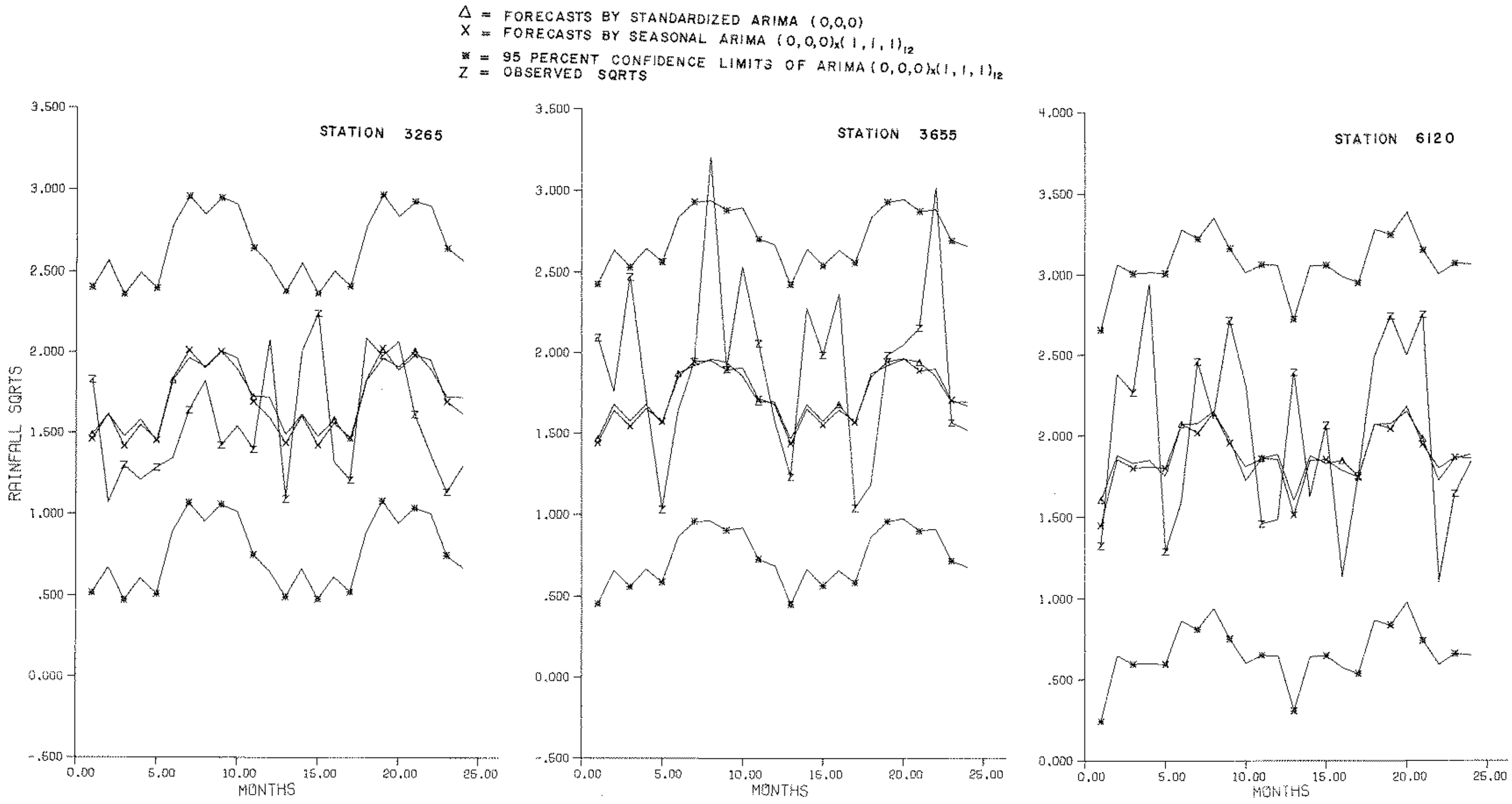


FIG. 24 MONTHLY RAINFALL SQUARE ROOT FORECASTS BY VARIOUS ARIMA MODELS

CHAPTER 4 - ANNUAL RAINFALL ANALYSIS

Annual rainfall series for the stations in Indiana were analyzed for their probability distributions and for the time series model of best fit.

White noise model was assumed for the series. The chi-square goodness of fit test was applied to the series for the test of normality. The sample distribution was classified into cells which, for convenience, had equal probability. The selection for the number of cells was arbitrary. However, *Markovic* (1965) had suggested the rule that the number of cells should be chosen in such a way that the average expected frequency of each cell is at least five. Since the sample sizes for this study varied from 38 to 57 years, the number of cells was taken as 8. Since equiprobable cells were considered, the expected value for each cell was

$$E = N/8 \tag{4.1}$$

where N was the sample size. Each cell j had a frequency $OBSC_J$ observed from the sample. The cell limits were found from the standard normal distribution $\Phi(\cdot)$ once the sample was standardized. The chi-square statistic was formed as

$$\chi = \sum_{J=1}^{8} (OBSC_J - E)^2/E. \tag{4.2}$$

This statistic is asymptotically chi-square distributed with seven degrees of freedom. In Table 4-1 the sample chi-square statistics for the standardized samples are given. In the next column the probabilities of exceeding the chi-square statistic are given. It is concluded that the annual rainfall series obey normal probability law.

Portemanteau lack of fit test was applied to the autocorrelation function of the actual series to test the white noise hypothesis. The results are shown on the Table 4-1. It is seen that all the series satisfy the white noise hypothesis.

As a result of the above analysis it is concluded that the annual rainfall series is a white noise process with the normal probability distribution.

TABLE 4-1

ANNUAL RAINFALL ANALYSIS

STATION NUMBER	RECORD LENGTH	CHI-SQUARE STATISTIC X	$P[\chi_7^2 > \chi]$	PORTEMANTEAU STATISTIC Q (30 D.O.F.)
3795	57	3.2105	.8649	22.47
3445	55	3.0364	.8816	23.56
2535	43	5.1860	.6373	18.30
3280	38	3.2632	.8596	12.52
3265	44	4.3636	.7371	15.18
6120	48	10.00	.1886	13.86
3655	42	6.7619	.4541	14.54
3805	57	10.2281	.1760	14.37
2840	57	2.6491	.9155	20.10

$Q = n \sum_{k=1}^{30} \hat{\rho}_k$ $\qquad \chi_{30}^2(90\%) = 40.3$ $\qquad \chi_{30}^2(95\%) = 43.8$

$\chi_7^2(90\%) = 12.0$ $\qquad \chi_7^2(95\%) = 14.1$

CHAPTER 5 - DISCUSSION OF THE RESULTS OF THE TIME SERIES ANALYSIS OF THE MONTHLY AND ANNUAL RAINFALL IN THE MIDWESTERN UNITED STATES

From the analysis of the differencing and the standardization as the methods for the removal of the circular stationarity in the monthly hydrologic time series it is seen that differencing, although very effective in the removal of the periodicities, distorts the spectral structure of the original monthly hydrologic series. This distortion causes inconveniences in the generation schemes of the ARIMA models fitted to the differenced series. Actually, because the contribution at the spectral origin is completely wiped out by differencing, it is impossible to regain the original spectral structure in the generation scheme. Therefore, differencing should be abandoned when the hydrologists are constructing models for the simulation purposes. From the forecasting point of view, seasonal ARIMA models, although they can preserve the mean, cannot preserve the variance. Therefore, even in forecasting they are quite ineffective.

Standardization, although it introduces some negligible nonstationarities, is seen to be an effective method in the removal of the circularly stationary part of the hydrologic time series. Furthermore, while removing the discrete spectral component, corresponding to the periodicities in the data, it distorts the stationary random component of the spectrum only slightly. The ARMA models fitted to the standardized monthly rainfall data can safely be used for generation since the original spectrum can be retrieved in their generation schemes. They are better than the seasonal ARIMA models for the forecasting purposes since they can preserve both the means and the standard deviations.

From the spectral and the variance-time analysis of the ARMA (p,q) models it is seen that these models asymptotically end up in the Brownian domain, and strictly speaking, cannot preserve the *Hurst's* law for the variance of the long range dependent time series. On the other hand, the ARIMA (1,d,1) family of models, although they can satisfy the infinite variance hypothesis of *Mandelbrot*, yield infinite variance in their generating forms. Due to this nonstationarity they are not suitable for the generation of the long range dependent hydrologic time series.

From the application of the nonseasonal ARIMA models to the monthly rainfall data the ARIMA (1,0,1) or the ARMA (1,1) model emerged as the most suitable model for the generation and the forecasting of the monthly rainfall series. This model passed the goodness of fit tests in all the cases studied. However, it should be remembered that this model is only good for the preservation of the short range dependent time series. As is shown in the long range dependence analysis, this model asymptotically ends up in the Brownian domain. However, for the practical purposes the time span of dependence that the ARMA (1,1) model preserves may be adequate. For this purpose the time span of dependence for the ARMA (1,1) model was derived in the section 1.3.

From the application of the seasonal multiplicative ARIMA models to the monthly rainfall data it was found that the ARIMA $(1,0,0) \times (1,1,1)_{12}$ model passed the goodness of fit tests in all the cases. However, this model cannot be used for generation and has only limited use in the forecasting of the monthly rainfall series since it cannot preserve the standard deviations.

The annual rainfall analysis showed that a normally distributed white noise sequence is adequate for the generation of the annual rainfall series. This result shows that the summation of the monthly rainfalls in the formation of the annual rainfall series filters out the circularity as well as any dependence structure that was present in the monthly rainfall series.

CONCLUSIONS FROM THE POINT STOCHASTIC AND THE TIME SERIES ANALYSIS OF THE RAINFALL SEQUENCES IN THE MIDWESTERN UNITED STATES

In Indiana the daily rainfall sequences were handled by the point stochastic analysis and the monthly and the annual rainfall sequences were handled through the time series analysis. The conclusions from the point stochastic analysis of the daily rainfall occurrences in Indiana may be stated as follows:

1. An objective methodology for the selection and calibration of a point stochastic model was developed and was successfully applied to the point daily rainfall occurrences in Indiana.
2. There is a slight downward long-term trend in the rate of daily rainfall occurrences in Indiana. The annual and the 15-day cyclicities are significant in the daily rainfall occurrences in Indiana.
3. There is a short memory dependence in the daily rainfall counts in Indiana.
4. The Neyman-Scott cluster model can preserve the dependence and the marginal probability characteristics of the point daily rainfall counting process in Indiana.
5. Through the physical concepts that are attached to the various components of the Neyman-Scott cluster model a first attempt is made in tying the meteorologic knowledge and the point stochastic analysis concerning the rainfall occurrence phenomenon.
6. The Neyman-Scott cluster model is most practical for the probability computations associated with the continental droughts. Although it is possible to obtain the practical probabilities associated with wet sequences, a computer program is necessary for the lengthy computations.

The conclusions from the time series analysis of the monthly and annual rainfall sequences in the Midwestern United States may be stated as follows:

1. The nonseasonal and the seasonal ARIMA models, fitted to the differenced monthly rainfall series, should not be used for simulation purposes. One cannot retrieve the original spectral structure of the monthly rainfall series through the generation schemes of these models.
2. ARMA (1,1) model, fitted to the standardized monthly rainfall series, is the most suitable model for the simulation of the monthly rainfalls.
3. ARMA (p,q) models end up in the Brownian domain when the span of dependence goes to infinity. Therefore, strictly speaking, they cannot preserve the long-range dependence characteristics of the hydrologic data.
4. For the forecasting purposes an ARMA (1,1) model or a white noise sequence on the standardized monthly rainfalls emerge as the most convenient models. The seasonal multiplicative ARIMA models, although they can preserve the monthly means, cannot preserve the monthly standard deviations and are inconvenient for the forecasting of the monthly rainfalls.
5. The annual rainfall sequences can be simulated by a white noise model whose random inputs are normally distributed.

BIBLIOGRAPHY

Andel, J. and Balek, J. "Analysis of Periodicity in Hydrological Sequences," *Journal of Hydrology*, Vol. 14, Oct. 1971, pp. 66-82, N. Holland Pub. Co., Amsterdam.

Anderson, R. L., "Distribution of the Serial Correlation Coefficient," *Ann. Math. Stat.*, Vol. 13, No. 1, 1942.

Anderson, T. W. and D. A. Darling, "Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes," *Ann. Math. Stat.*, Vol. 23, pp. 193-212, 1952.

_____, "A Test of Goodness of Fit," *J. Amer. Stat. Ass.*, Vol. 49, pp. 765-769, 1954.

Bartlett, M. S., "On the Theoretical Specification of Sampling Properties of Autocorrelated Time Series," *Jour. Royal Stat. Soc.*, B8, 27, 1946.

_____, *An Introduction to Stochastic Processes*. London: Cambridge University Press, 1955.

_____, "The Spectral Analysis of Point Processes," *J. R. Stat. Soc.*, B25, pp. 264-296, 1963.

Box, G. E. P. and Jenkins, G. M. *Time Series Analysis, Forecasting and Control*. Holden-Day: San Francisco, 1971.

_____ and D. A. Pierce, "Distribution of Residual Autocorrelations in Autoregressive-integrated moving Average Time Series Models," *Jour. Amer. Stat. Assoc.*, 64, 1970.

Carlson, R. F., A. J. A. MacCormick, and D. G. Watts, "Application of Linear Random Models to Four Annual Flow Series," *W. R. R.*, Vol. 6(4), pp. 1070-1078, 1970.

Caskey, I. E., Jr., "A Markov Chain Model for the Probability Occurrence in Interval of Various Length," *Monthly Weather Review*, Vol 91, pp. 289-301, 1963.

Cote, Louis, *Unpublished Lecture Notes on Time Series Analysis*. Stat. Dept., Purdue Univ., W. Lafayette, Indiana, 1973.

Cox, D. R. *Renewal Theory*. London: Methuen and Co. Ltd., 1962.

Cox, D. R. and P. A. W. Lewis, *Statistical Analysis of Series of Events*. London: Methuen and Co. Ltd., 1966.

Cox, D. R. and W. L. Smith, "The Superposition of Several Strictly Periodic Sequences of Events," *Biometrika*, Vol. 40, pp. 1-11, 1953.

Cramer, H., *Mathematical Methods of Statistics*. Princeton University Press, 1946.

_____ and Leadbetter, M. R., *Stationary and Related Stochastic Processes*. Wiley: New York, 1967.

Crovelli, R. A., "Stochastic Models for Precipitation," *Proc. Int. Symp. on Uncertainties in Hydrologic and Water Resource Systems*, Tucson, 1972.

Draper, N. R. and H. Smith, *Applied Regression Analysis*. New York: Wiley and Sons, Inc., 1966.

Duckstein, L., M. N. Fogel, and C. C. Kisiel, "A Stochastic Model of Runoff-Producing Rainfall for Summer Type Storms," *Water Res. Res.*, Vol. 8, No. 2, pp. 410-421, 1972.

Durbin, J., "Some Methods of Constructing Exact Tests," *Biometrika*, Vol. 48, pp. 41-55, 1961.

Dwass, M., *Probability Theory and Applications*. New York: W. A. Benjamin Inc., 1970.

Feyerherm, A. M. and L. D. Bark, "Statistical Methods for Persistent Precipitation Patterns," *Sixth Nat. Conf. on Agricultural Meteorology*, Lincoln, 1964.

Feyerherm, A. M., L. D. Bark, and W. C. Burrows, "Probabilities of Sequences of Wet and Dry Days in Indiana," *North Central Regional Research Pub. 161*, Tech. Bull. 139f, Kansas, 1965.

Feyerherm, A. M. and L. D. Bark, "Goodness of Fit of a Markov Chain Model for Sequences of Wet and Dry Days," *J. Appl. Meteor.*, Vol. 6, pp. 770-773, 1967.

Gabriel, K. R. and J. Neumann, "On a Distribution of Weather Cycles by Lengths," *Quart. J. R. Met. Soc.*, Vol. 83, pp. 375-380, 1957.

_____, "A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv," *Quart. J. R. Met. Soc.*, Vol. 88, pp. 90-95, 1962.

Gabriel, K. R., "The Distribution of the Number of Successes in a Sequence of Dependent Trials," *Biometrika*, Vol. 96, pp. 454-460, 1959.

Grace, R. A. and P. S. Eagleson, *The Synthesis of Short-time Increment Rainfall Sequences.* MIT Civil Eng. Dept. Hydrodynamics Lab., Rept. 91, Cambridge, Mass., 1966.

Grant, E. L., "Rainfall Intensities and Frequencies," *ASCE Trans.*, Vol. 103, pp. 384-388, 1938.

Green, J. R., "A Model for Rainfall Occurrence," *J. R. Stat. Soc., B*, Vol. 26, pp. 345-353, 1964.

_____, "Two Probability Models for Sequences of Wet and Dry Days," *Monthly Weather Review*, Vol. 93, pp. 155-156, 1965.

Hanna, E. J., *Time Series Analysis.* Science Paperbacks, Chapman and Hall, Ltd.: London, 1960.

Hufschmidt, M. M., and M. B. Fiering, *Simulation Techniques for Design of Water Resource Systems.* Cambridge, Mass.: Harvard University Press, 1966.

Hurst, H. E., "Long-term Storage Capacity of Reservoirs," *Trans. Amer. Soc. Civil Engrs.*, Vol. 116, pp. 770-808, 1951.

_____, "Methods of Using Long Term Storage in Reservoirs," *Proc. Inst. Civil Engrs.*, Vol. 1, pp. 519-543, 1956.

Hurst, H. E., R. P. Black, and Y. M. Simaika, *Long Term Storage, An Experimental Study.* London: Constable, 1965.

Jenkins, G. M. and D. G. Watts, *Spectral Analysis and its Applications*, Holden-Day: San Francisco, 1968.

Jennings, A. H., "World's Greatest Observed Point Rainfalls," *Monthly Weather Review*, Vol. 78, pp. 4-5, 1950.

Jorgensen, D. L., "Persistency of Rain and No-Rain Periods During the Winter at San Francisco," *Monthly Weather Review*, Vol. 77, pp. 303-307, 1949.

Kendall, M. G. and Stuart, A., *The Advanced Theory of Statistics*, Vol. 2, London: Griffin, 1961.

Khintchine, A. Y., *Mathematical Methods in the Theory of Queuing.* Moscow: Izdad, Akad. Nauk. (English Translation Griffin, 2nd Ed., 1969).

Kisisel, I. T. and J. W. Delleur, "An Analysis of Hydrologic Time Series and Generation Models of Synthetic Flows for Some Indiana Watersheds," Purdue Univ., *W. R. R. Center*, Tech. Rep. 19, Lafayette, Ind., 1971.

Lewis, P. A. W., A. M. Katcher, and A. H. Weis, *SASE IV - An Improved Program for the Analysis of Series of Events.* IBM Report, 1969.

Lewis, P. A. W., "Remarks on the Theory, Computation and Application of the Spectral Analysis of Series of Events," *J. Sound Vib.*, Vol. 12, pp. 353-375, 1970.

_____, "Recent Results in the Statistical Analysis of Univariate Point Processes," in *Stochastic Point Processes*, ed. by P. A. W. Lewis, New York: Wiley and Sons Inc., 1972.

Lobert, A. *Modele Probabiliste de Base pour les pluies dans le bassin de L'Allier.* Note 45/67, Chatou: C.R.E.C. HYD., 1967.

Mandelbrot, B. B., "Long-run Linearity, Locally Gaussian Process, H-Spectra and Infinite Variances," *International Economic Review*, 10(1), pp. 82-111, 1969.

_____ and J. R. Wallis, "Noah, Joseph and Operational Hydrology," *Water Resources Research,* 4(5), pp. 909-918, 1968.

_____ and _____, "Some Long-run Properties of Geophysical Records," *Water Resources Research,* 5(2), pp. 321-340, 1969.

Markovic, R. D., "Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff," Colorado State Univ. Hydrology Paper No. 8, 1965.

Marquardt, D. W., "An Algorithm for Least Squares Estimation of Nonlinear Parameters," *SIAM Journal,* Vol. 2, pp. 431-441, 1963.

_____, "Least Squares Estimation of Non-linear Parameters," A computer program in FORTRAN IV; IBM Share Library, No. 3094, 1966.

McKerchar, A. I. and J. W. Delleur, "Stochastic Analysis of Monthly Flow Data Application to Lower Ohio River Tributaries," Purdue Univ., *W. R. R. Center*, Tech. Rep. 26, Lafayette, Ind., 1972.

McMichael, F. C. and J. S. Hunter, "Stochastic Modeling of Temperature and Flow in Rivers," *Water Resources Research,* 8(1), pp. 87-98, 1972.

Mitchell, J. M., Jr., "A Critical Appraisal of Periodicities in Climate," in *Weather and Our Food Supply.* CAED Rep. 20, Ames, Iowa, Iowa State Univ. of Science and Tech., 1964.

Moran, P. A. P., "The Random Division of an Interval, Part II," *J. R. Stat. Soc., B,* Vol. 13, pp. 147-150, 1951.

Moss, M. E., *Serial Correlation Structure of Discretized Streamflow.* Ph.D. Thesis, Colorado State Univ., July 1972.

Moyal, J. E., "The General Theory of Stochastic Population Processes," *Acta Math.,* Vol. 108, pp. 1-31, 1962.

Newman, J. E., *Private Communication.* Purdue Univ., W. Lafayette, Ind., 1975.

Newnham, E. V., "The Persistence of Wet and Dry Weather," *Quart. J. Roy. Met. Soc.,* Vol. 42, pp. 153-162, 1916.

Neyman, J. E. and E. L. Scott, "A Theory of the Spatial Distribution of Galaxies," *Astrophysical Jour.,* Vol. 116, 1952.

_____, "A Statistical Approach to Problems of Cosmology," *J. Roy. Stat. Soc., B,* Vol. 20, pp. 1-43, 1958.

_____, "Processes of Clustering and Applications," in *Stochastic Point Processes,* ed. by P. A. W. Lewis, New York: Wiley and Sons, Inc., 1972.

O'Connell, P. E., "A Simple Stochastic Modelling of Hurst's Law," Paper presented at IASH International Symposium on Mathematical Models in Hydrology, Warsaw, Poland, 1971.

Parzen, E., *Stochastic Processes.* 3rd Ed., San Francisco: Holden-Day, 1967.

Petterssen, S., *Introduction to Meteorology.* 3rd Ed., New York: McGraw-Hill, 1969.

Quenouille, M. H., "Approximate Tests of Correlation in Time Series," *Jour. Royal Stat. Soc., B11,* 68, 1949.

Roesner, L. A., and V. M. Yevdjevich, "Mathematical Models for Time Series of Monthly Precipitation and Monthly Runoff," Colordao State Univ. Hydrology Paper 15, Ft. Collins, Colorado, 1966.

Romanof, N., "The Markov Chains Estimation Theory Application to Meteorology," *Proc. 5th Conf. Prob. Theory,* Brasov, 1972.

Shahabian, H. L., *Spectral Analysis and Its Applications to Hydrologic Time Series of Lower Ohio Tributaries.* M.S. Thesis, Purdue Univ., August 1973.

Shane, R. M., *The Application of the Compound Poisson Distribution to the Analysis of Rainfall Records.* M.S. Thesis, Cornell Univ., Ithaca, N. Y., 1964.

Smith, E. R. and H. A. Schreiber, "Point Processes of Seasonal Thunderstorm Rainfall. Part 1. Distribution of Rainfall Events," *Water Res. Res.,* Vol. 9, No. 4, pp. 841-884, 1973.

Todorovic, P. and V. Yevjevich, "Stochastic Process of Precipitation," C. S. U. Hydrology Paper No. 35, Fort Collins, 1969.

Todorovic, P. and D. A. Woolhiser, "Stochastic Model of Daily Rainfall," *Proc. of the USDA-IASPS Symposium on Statistical Hydrology*, Tucson, 1971.

Watson, G. S. and W. R. Wells, "On the Possibility of Improving the Mean Useful Life of Items by Eliminating Those with Short Lives," *Technometrics*, Vol. 3, pp. 281-298, 1961.

Weiss, L. L., "Sequences of Wet and Dry Days Described by a Markov Chain Model," *Monthly Weather Review*, Vol. 92, pp. 169-176, 1964.

Wiser, E. H., "Modified Markov Probability Models of Sequences of Precipitation Events," *Monthly Weather Review*, Vol. 93, pp. 511-516, 1965.

Woolhiser, D. A., E. Rovey, and P. Todorovic, "Temporal and Spatial Variation of Parameters of N-day Precipitation," *Proc. 2nd Int. Symp. Hydrology*, C. S. U., Ft. Collins, 1972.

Vere-Jones, D. and R. B. Davies, "A Statistical Survey of Earthquakes in the Main Seismic Region of New Zealand. Part II; Time Series Analysis," *N. Z. J. Geol. Geophysics*, Vol. 9, pp. 251-284, 1966.

Vere-Jones, D., "Stochastic Models for Earthquake Occurrence," *J. Roy. Stat. Soc., B*, Vol. 32, pp. 1-62, 1970.

Visser, S. S., *Climate of Indiana*. Bloomington, Ind.: Indiana Univ. Press, 1944.

Yaglom, A. M., *An Introduction to the Theory of Stationary Random Functions*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962.