THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

CONVECTIVE RAINFALL REGIONS IN PUERTO RICO

By

MATTHEW M. CARTER

A Thesis submitted to the
Department of Meteorology
in partial fulfillment of the
requirements for the degree of
Master of Science

Degree Awarded:
Summer Semester, 1995

The members of the Committee approve the thesis of Matthew M. Carter defended on July 6, 1995.

James B. Elsner
Professor Directing Thesis

Jon E. Ahlquist
Committee Member

Henry Fuelberg
Committee Member

Dedication

This thesis is dedicated to my parents, who not only opened my eyes to the world of educational and career opportunities, but also provided me with the freedom to choose my path in life. My love always.

Acknowledgements

I am grateful to my advisor, Dr. Elsner, for his support and guidance, without which this work would not have been accomplished. I am further grateful to my committee members, Dr. Jon Ahlquist and Dr. Henry Fuelberg, for their insight on significance testing and convective rainfall predictors. Dr. LaSeur was a valuable resource on convective forcing in the tropics as well as offering input on statistical matters.

I would also like to acknowledge the assitance of Dr. Kevin Kloesel in retrieving hourly rainfall data from NCDC. Joey Comeaux and Will Spangler, both from NCAR, assisted in locating surface and upper air data.

Contents

List of Tables

List of Figures

Abstract

Geographic regions of covariability in hourly precipitation over the island of Puerto Rico are exposed using factor analysis. It is argued that the data are consistent with a common factor model when an orthogonal rotation is applied to the factor loading matrix. We suggest that Puerto Rico can be divided into six regions with each region having a similar covariance structure of summer season convective rainfall. These six regions can be further grouped into a western area and an eastern area based on contrasting diurnal rainfall signatures. This study is believed to be one of the first attempting geographic regionalization of precipitation on a convective scale.

We attempt to construct a ordinary least-squared linear regression model to forecast for twelve hour precipitation for the eastern area. By progressively adding a component to the predictor variable vector based on optimal correlation, we obtain a model equation that no longer improves forecasts after five predictors. We construct our linear regression model using an out of sample approach. The linear regression equation we obtain from these five predictor variable components is more accurate for predicting daytime convective rainfall than is climatology or persistence.

Chapter 1

Introduction

Puerto Rico is an island territory of the United States located in the Caribbean Sea bounded by 65.6° W to 67.25° W and 17.9° N to 18.5° N. The Commonwealth of Puerto Rico contains a brick-shaped main island, measuring 180 kilometers by 65 kilometers, and five nearby smaller islands. Although only 8,897 square kilometers in area, smaller than the state of Connecticut, Puerto Rico is topographically diverse (Figure 1). Much of the interior of the main island is a mountain range, with its spine running east-west along its length. The highest peak in this range, Cerro de Punta, is 1,338 meters (4,389 feet) above sea level. Foothills comprise the area surrounding this central range, and give way to a coastal plain in the northern and southern parts of the island (Pico 1974). In the east and the west, foothill valleys extend finger-like to the sea. There exists a divide between the central range, or Cordillera Central, and a small but steep range to the northeast. This isolated range, Sierra de Luquillo, contains the peak El Yunque, which acts as an important convection point for summertime rainfall.

Puerto Rico's varying topography over a small area in the Caribbean Sea lead to stark land, sea, and air interactions between the months of May and September. Easterly trade

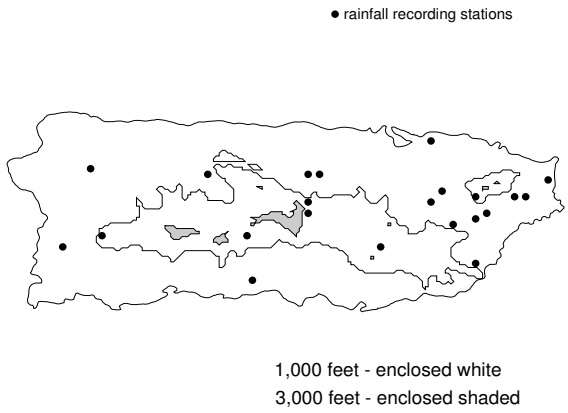


Figure 1: Topography of Puerto Rico. Enclosed white region is elevation between 1,000 feet and 2,999 feet. Enclosed shaded region is elevation above 3,000 feet.

winds in the tropics prevail over Puerto Rico during much of the year. During the summer, San Juan, which lies on a narrow peninsula on the northern coast, experiences a wind from the east, northeast, or southeast about sixty percent of the time.

Vertical motion associated with the sea breeze often leads to rainfall. Easterly trades reinforce the sea breeze on the eastern coast of the island, and may penetrate several kilometers inland. On the western coast, the sea breeze must overcome the prevailing easterlies in order to move ashore (Riehl 1954).

In addition, the rugged interior topography induces rain. On the windward side of mountains, moist air is forced up the slope where it cools and condenses leading to orographic precipitation. Interior mountain peaks also provide focal points of convection. Insolation heats the peaks much faster than the valleys below, which are shaded by the mountains (Pico 1974). Resulting towering cumulus and cumulonimbus clouds are common over the interior during summer days. Morevoer, differential heating between the mountains and

the valleys leads to daytime upslope winds as the peaks heat and force air upward. Nocturnal drainage flow occurs as the air cools and sinks down the slopes at night (Ray 1986). Interaction between prevailing easterlies, sea breezes, and upslope winds may force rainfall in the interior of the island.

Waves that originate off the west coast of Africa and propagate in the tropical trade belt also bring rainfall to Puerto Rico. These are primarily summertime phenomena, and number between fifty and seventy annually. Not all waves bring rain to Puerto Rico, and some may in fact inhibit mesoscale diurnal rainfall effects. Other waves develop into depressions, tropical storms, and hurricanes, which bring copious amounts of rain.

The tropical upper-tropospheric trough (TUTT) is a climatological feature that exists over the tropical Atlantic during the summertime. Cold core, upper level cyclones (Kousky and Gan 1981, Kelley and Mock 1982) often originating along the TUTT axis can influence rainfall over Puerto Rico (Frank 1970). These upper level lows with their direct thermal circulation and absence of dissipative mechanisms can last for several days to weeks. Precipitation amounts associated with these lows are proportional to their vertical extent, which on occasion reach down to the surface (Frank 1970).

Fassig (1916) conducted one of the earliest studies of rainfall in Puerto Rico. He compared the duration, frequency, and intensity of rainfall in San Juan with that of a mid-latitude city, Baltimore. He found that tropical rain is of shorter duration and greater frequency than mid-latitude rain. Excessive rainfall events (greater than one inch per hour, or two and a half inches per day), occur with greater frequency at San Juan, and last for a longer time, than in Baltimore. For such events, however, Fassig reported that Baltimore

exhibited higher rainfall rates than did San Juan.

Ray (1928) studied hourly rainfall frequencies in San Juan for the period 1905 to 1927. During the summer months, he reported a primary maximum in the hourly rainfall frequency in the late afternoon, with a secondary maximum during the early morning hours. This closely mirrors what is revealed in our data record, as we shall see in Chapter Two. During the winter, Ray found that the frequency maximum occurs for the duration of the overnight hours, while there is a pronounced minimum during the afternoon. For rainfall amounts, there is considerably less hour to hour variation during the winter than during the summer, although the rainfall amount extrema correspond to the frequency extrema temporally.

In this study, we seek to develop prediction algorithms for forecasting convective rainfall in Puerto Rico. Tropical convective rainfall is on the scale of a few kilometers at most and may miss an eight inch diameter rain gauge, despite the fact that it is raining nearby. The features that force this rainfall may exist over a small region surrounding the gauge, and force convective precipitation over an entire summer. It is useful to identify these regions so that we may capture common forcing mechanisms among the stations. In this paper, we accomplish a regionalization of Puerto Rico based upon its summertime convective rainfall using an analysis-of-variance technique known as common factor analysis. Because this study concerns small spatial and temporal scales of tropical rainfall, results may be of value in calibrating precipitation measurements taken from the Tropical Rainfall Measuring Mission (TRMM) satellite (Simpson 1988) currently planned for launch in 1997.

Chapter 2

Data

2.1 Hourly Rainfall Data

The National Climatic Data Center (NCDC) maintains records for twenty-two stations in Puerto Rico that record rainfall on an hourly basis (Figure 1). San Juan contains the most extensive hourly rainfall record on the island. Its data record began on 1 January, 1967. All of the other stations began their hourly rainfall records in either 1971 or 1973. So that all stations have uniform data record length, the data record for this study begins on 1 July, 1973 and ends on 30 June, 1988.

Diurnal convective and sea breeze rainfall events are most prevalent during the summer months. African easterly waves pass over the island during the summer months and cold core upper level cyclones are most numerous during this time. Summer rainfall regimes in Puerto Rico are either forced on the mesoscale, or propagate over the island, generally speaking, from the east. Rare exceptions are hurricanes and tropical storms, which will be discussed shortly. During winter, mid-latitude systems contribute to rainfall during the months of October through April. For the purpose of this study, which is to find common regions of

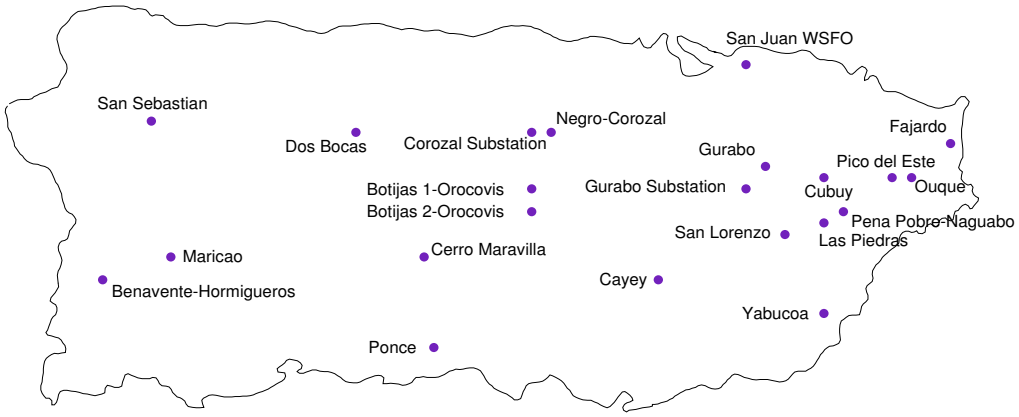


Figure 2: Stations in Puerto Rico that record rainfall on an hourly basis.

rainfall variability over Puerto Rico based upon summertime influences, we restrict interest to the months of May through September. The total length of each station's data record is 55,080 hours.

Hourly rainfall amounts are in tenths of an inch, except for San Juan and Benavente-Hormigueros, which have a resolution of one-hundredth of an inch. Missing hourly rainfall amounts are neglected from the data set. Hurricanes, tropical storms, and named tropical depressions are removed from the data set as well. Appendix A lists these storms and the corresponding hours that are removed from the data set. Developed tropical cyclones are associated with intense rainfall. Since regionalization is attempted with an eye toward prediction, such storms are omitted for the following two reasons: 1) These storms are monitored with particular attention by the National Hurricane Center. 2) Considerable skill exists in several statistical and dynamical models for forecasting tropical cyclones. The predictive scheme developed in this study for rainfall in Puerto Rico is a linear regression

model. Because statistical and dynamical hurricane models have long exceeded the standard of linear regression, it is felt that inclusion of tropical cyclones in this study is unnecessary.

2.1.1 Rainfall Climatology

Appendix B shows hourly rainfall climatologies for each of the stations in the data set. The maximum rainfall amount that occurred at each hour during the months of May through September is plotted as the top curve. All of the rainfall amounts for each hour are ordered greatest to least. The smallest value for each hour is zero. The second curve represents the 99.9th percent value for each hour, the third curve the 99th percent value, and the bottom curve the 97th percent value. Note that the values on the abscissa reach 30. Hours 24 through 30 are "wraparound" times, and are the same as hours midnight through 6 a.m. This reveals any continuity in the rainfall climatology over the nighttime hours.

The curves begin to flatten by the 97th percentile. This shows how quickly rainfall amounts fall from the maximum values. Rainfall events of one inch or more occur infrequently, and the most common rainfall amount is zero, as we shall see later. In many of the station records, over ninety percent of the hourly rainfall values are zero.

2.1.2 Rainfall Frequency

Appendix C shows the frequency of rainfall events at each hour for the same data set. Again, hours 24 through 30 on the abscissa are wraparound times. A rainfall event is any hour that a station records rainfall greater than a trace. The number of rainfall events for each hour is plotted by the black curve.

Two principal types of curve behavior emerge from these graphs: 1) Stations with low amplitude maxima that exhibit small hourly variation in the frequency of events (e.g. San Juan WSFO), and 2) stations that exhibit large amplitude, mid-afternoon maxima with relative minima in the early morning and overnight hours (e.g. Dos Bocas). The low amplitude stations lie in the eastern two thirds of the island, from San Lorenzo eastward. The exceptions are Cubuy, Gurabo, and Gurabo Substation which lie in the shadow of El Yunque, a strong orographic rainmaker. Stations west of San Lorenzo exhibit high amplitude afternoon maxima. Many of these stations lie near Cordillera Central where strong convective forcing takes place. The dichotomy of the rainfall frequency curves is an important consideration in regionalizing the island, as we shall see in Chapter 5.

2.2 Surface and Upper Air Data

Hourly surface data for San Juan used in developing a prediction algorithm were retrieved from the National Center for Atmospheric Research (NCAR) via the Cray Y-MP8/864 supercomputer Shavano. In developing a prediction scheme, we chose a representative sample of randomly selected hours from the years 1977 through 1981. We retrieved all hourly surface data for the period 1 May, 1977 through 30 September, 1988 from NCAR data set ds472.0.

The San Juan Weather Service Office is the only station in Puerto Rico that has an extensive upper-air data record. These data are collected twice daily, at 0000 UTC (8 p.m. AST) and 1200 UTC (8 a.m. AST) and are found in NCAR data set ds390.1. Our data set

consists of all upper air soundings for the same time period as the surface data set.

Chapter 3

A Factor Analysis Model

Factor analysis is a process by which we attempt to describe a correlation structure of several variables in terms of factors. Factors are for the most part intangible entities that may not be observed, but represent regions in which several variables are a) highly correlated with each other, and b) uncorrelated with other variables. In this study, each of the factors represents a region of rainfall, and each of the variables represents a station. By applying factor analysis to the correlation matrix, we seek to identify $m < p$ (where p is the number of stations) regions of rainfall that contain stations that exhibit a similar rainfall pattern with each other and a different pattern with all other stations (Johnson and Wichern 1982).

In practice factor analysis is not this clean. A station may be highly correlated with two separate groups of stations or not correlated to any. In such instances (and they exist in this study as we shall see later), ultimate determination of the regions must fall on somewhat subjective means. Factor analysis provides an objective guide to regionalization, not the final map itself.

Following the discussion by Johnson and Wichern (1982), each factor is broken into two components, common factors, $f_1 \ldots f_m$, which represent the unobservable explanation

for the groups of high correlation, and specific factors, $\epsilon_1 \ldots \epsilon_p$, where p is the number of additional sources of variation (stations). The relationship between the common factors, the specific factors, and the observations is given as

$$\begin{aligned} X_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \epsilon_1 \\ X_2 - \mu_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2m} f_m + \epsilon_2 \\ &\vdots \qquad\qquad\qquad \vdots \\ X_p - \mu_p &= \lambda_{p1} f_1 + \lambda_{p2} f_p + \cdots + \lambda_{pm} f_m + \epsilon_p \end{aligned} \tag{1}$$

If i ranges from 1 to p, and j ranges from 1 to m, then X_i is the rainfall vector for station i, μ_i is the mean, and λ_{im} is the loading for vector X_i on factor m. Common factor number j is f_j, and ϵ_i is the specific factor for station vector X_i. In matrix format, the relationship given by Equation Set 1 is

$$\mathbf{X} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{F} + \boldsymbol{\epsilon}, \tag{2}$$

where it is assumed that the expected values of $\boldsymbol{\epsilon}$, $\mathbf{F}$, and $\mathbf{X} - \boldsymbol{\mu}$ are zero. It is also assumed that each common factor has unit variance, the common factors are independent of each other, and the common factors and the specific factors are mutually independent.

In contrast to the commonly employed principal component analysis, factor analysis starts with the assumption of an underlying basic model for the data. This model is given by Equation Set 1. The factor loadings may be described in terms of the population covariance matrix by the following relation,

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}, \tag{3}$$

where $\mathbf{\Sigma}$ is the $p \times p$ population covariance matrix, $\mathbf{\Lambda}$ is the $p \times m$ matrix of factor loadings relating the common factors f_i's to the observed variables x_i's, and $\mathbf{\Psi}$ is the $p \times p$ matrix of covariances of the specific factors ϵ_i's. Since we are assuming a common factor model for the factor analysis, the specific factors are also assumed to be independent of one another so the matrix $\mathbf{\Psi}$ is diagonal.

In this study we compute the factor loadings as

$$\mathbf{\Lambda} = \mathbf{\Gamma}\mathbf{\Delta}^{-\frac{1}{2}}, \tag{4}$$

where $\mathbf{\Gamma}$ and $\mathbf{\Delta}$ are the eigenvectors and eigenvalues, respectively, of the dispersion matrix $(\mathbf{R} - \mathbf{\Psi})$. The matrix $\mathbf{R}$ is the $p \times p$ sample correlation matrix computed from the data set consisting of p=22 stations and n=55,080 hours for each station. Although the common factor model given by Equation 1 applies for loading matrices described by a population covariance matrix $\mathbf{\Sigma}$ (Equation 3), it also applies for loadings described by a population correlation matrix. We standardize $\mathbf{X}$ and approximate population correlation matrix $\boldsymbol{\varrho}$ with sample correlation matrix $\mathbf{R}$ (Johnson and Wichern 1982). Loading matrix $\mathbf{\Lambda}$ is a $p \times m$ matrix, $\mathbf{\Gamma}$ is a $p \times m$ matrix, and $\mathbf{\Delta}$ is an $m \times m$ diagonal matrix. By choosing $m = p$, $\mathbf{\Gamma}$ and $\mathbf{\Delta}$ are square matrices composed of all p eigenvectors and eigenvalues respectively. For $m < p$, $\mathbf{\Gamma}$ and $\mathbf{\Delta}$ are rectangular matrices that reveal the eigenvectors and eigenvalues for the first m modes.

Individual elements of sample correlation matrix $\mathbf{R}$ are given by

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}, \tag{5}$$

where the s_{ik} is the sample covariance between stations i and k and is given by

$$s_{ik} = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k), \tag{6}$$

where x_i and x_k are rainfall amounts at hour j for stations i and k respectively, and where $\bar{x}_i$ and $\bar{x}_k$ are the respective sample means of stations i and k. Here we try two approaches to estimate the factor loadings with nearly identical results. The first is noniterative, and it assumes the specific variance for each station is equal to a constant times the squared multiple correlation coefficient. The other uses an unweighted least-squares iterative technique to refine these initial estimates.

The latter approach is shown in the following example. $\mathbf{\Lambda\Lambda^T} + \mathbf{\Psi}$ will be a $p \times p$ product matrix for all $m \leq p$. For $m = 1$, suppose our $p \times p$ correlation matrix is,

$$\begin{bmatrix} 1 & .4 & .6 \\ .4 & 1 & .3 \\ .6 & .3 & 1 \end{bmatrix}$$

We set this matrix equal to $\mathbf{\Lambda\Lambda^T} + \mathbf{\Psi}$, as given by Equation 3

$$\begin{aligned} 1 = \lambda_{11}^2 + \psi_1 \quad .40 = \lambda_{11}\lambda_{21} \quad .60 = \lambda_{11}\lambda_{31} \\ 1 = \lambda_{21}^2 + \psi_2 \quad .30 = \lambda_{21}\lambda_{31} \\ 1 = \lambda_{31}^2 + \psi_3 \end{aligned}$$

We can pick any pair of equations that share a loading. We choose,

$$\begin{aligned} .60 &= \lambda_{11}\lambda_{31} \\ .30 &= \lambda_{21}\lambda_{31} \end{aligned}$$

so that

$$\begin{aligned} \lambda_{21} &= \left(\frac{.3}{.6}\right)\lambda_{11} \\ &= 0.5\lambda_{11} \end{aligned}$$

Substituting this expression for λ_{21},

$$\begin{aligned} .40 &= 0.5\lambda_{11}\lambda_{11} \\ .80 &= \lambda_{11}^2 \\ \pm.894 &= \lambda_{11} \end{aligned}$$

Now we can solve for ψ_1

$$\begin{aligned} \psi_1 &= 1 - \lambda_{11}^2 \\ &= 1 - 0.8 \\ &= 0.2 \end{aligned}$$

The other loadings can be determined through substitution. This approach is iterative in that the International Mathematical and Statistical Library (IMSL) subroutine FACTR, which performs the actual calculations, uses a numerical algorithm to solve for the loadings in Equation 3.

Our factor analysis model bears resemblance to the popular principal component analysis (Dyer 1975, White et al. 1991, and Lyons and Bonell 1994). The difference between the

two methods is that factor analysis allows for specific factors (or specific variance), while principal component analysis does not. Perhaps the best way to understand specific variance is through an example. A certain type of forcing may influence rainfall variability for a group of stations. These stations have this forcing in common. There may be an additional, very localized, forcing on one of the stations in this group. In principal component analysis, the variance explained by this forcing is distributed among all of the stations, both inside and outside of the group. This variance is found in all of the loadings. This may lead to a poor regionalization of stations since the sum of localized effects will be spread over all modes. In factor analysis, the variance due to the localized forcing goes into the specific variance. The variance due to that localized forcing does not manifest itself on any of the loadings. The variance the stations share is called communality. Stations that have communality comprise a region and represent a single factor. Unlike principal component analysis, the specific variance is excluded from common factors and is not hidden in the loadings. Optimally, we would like to have the number of factors such that the specific variance is minimized; that is, we would like to have as much of the variance as possible be explained by communality. This difference between factor analysis and principal component analysis may only be superficial for variables such as monthly sea level pressures or 50 kPa heights, but could be important for variables, like rainfall, where very local effects can have a significant influence on individual station variability.

Chapter 4

Procedures

4.1 Correlation Matrix

Since the correlation matrix forms the backbone of our study, it is presented in Table 1. With the exception of Ponce correlated with San Sebastian, all of the correlation coefficients are positive. We note that $\bar{x}_i$ is very small for all i since each station vector is dominated by zeros (Table 2).

In fact, all $\bar{x}_i$'s are smaller than the resolution of the rain gauge. Since all values of x_{ij} and x_{kj} are positive or zero, values of $(x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)$ will only be negative when one station records no rainfall and not the other. If station i is reporting zero rainfall, values of $(x_{ij} - \bar{x}_i)$ are of an order of magnitude smaller than rain gauge resolution, or $O \sim 10^{-2}$ inches, and negative. When a station records rainfall, this value is of an order of at least 10^{-1} inches, and is positive. So when one station records rain while the other remains dry, it contributes negatively to the summation in the covariance equation (Equation 6). When summed over the entire record, we may find that the equation gives a negative value if one station has many zeros and a small mean, while the other has few zeros and a larger mean.

Table 1: Correlation matrix for all twenty-two stations in Puerto Rico, July 1973 through June 1988.

1	1.000	.075	.063	.065	.056	.018	.066	.019	.026	.027	.037	.040	.105	.060	.058	.010	.081	.273	.072	.062	.031	.026
2	.075	1.000	.309	.129	.085	.062	.478	.078	.067	.111	.064	.184	.080	.296	.136	.058	.108	.053	.080	.038	.048	.102
3	.063	.309	1.000	.133	.105	.080	.257	.108	.098	.177	.065	.183	.120	.161	.106	.054	.082	.054	.110	.033	.076	.107
4	.065	.129	.133	1.000	.268	.360	.084	.193	.443	.298	.262	.244	.074	.056	.074	.173	.231	.033	.120	.028	.180	.246
5	.056	.085	.105	.268	1.000	.249	.096	.147	.217	.146	.220	.143	.126	.072	.053	.132	.332	.069	.081	.003	.128	.110
6	.018	.062	.080	.360	.249	1.000	.055	.176	.459	.221	.427	.231	.052	.039	.017	.192	.240	.008	.114	.004	.237	.151
7	.066	.478	.257	.084	.096	.055	1.000	.110	.052	.107	.079	.125	.117	.432	.113	.082	.100	.073	.072	.034	.071	.077
8	.019	.078	.108	.193	.147	.176	.110	1.000	.215	.565	.184	.135	.032	.092	.071	.304	.125	.020	.053	.032	.205	.156
9	.026	.067	.098	.443	.217	.459	.052	.215	1.000	.341	.302	.178	.057	.034	.036	.204	.201	.016	.107	.015	.202	.193
10	.027	.111	.177	.298	.146	.221	.107	.565	.341	1.000	.198	.145	.044	.085	.063	.232	.126	.030	.078	.032	.176	.167
11	.037	.064	.065	.262	.220	.427	.079	.184	.302	.198	1.000	.196	.049	.054	.018	.186	.275	.026	.109	.005	.255	.127
12	.040	.184	.183	.244	.143	.231	.125	.135	.178	.145	.196	1.000	.090	.118	.063	.098	.175	.023	.191	.008	.175	.171
13	.105	.080	.120	.074	.126	.052	.117	.032	.057	.044	.049	.090	1.000	.056	.072	.045	.130	.130	.097	.055	.037	.052
14	.060	.296	.161	.056	.072	.039	.432	.092	.034	.085	.054	.118	.056	1.000	.111	.077	.084	.057	.050	.024	.049	.094
15	.058	.136	.106	.074	.053	.017	.113	.071	.036	.063	.018	.063	.072	.111	1.000	.037	.049	.078	.030	.149	.020	.060
16	.010	.058	.054	.173	.132	.192	.082	.304	.204	.232	.186	.098	.045	.077	.037	1.000	.147	.020	.056	.018	.198	.140
17	.081	.108	.082	.231	.332	.240	.100	.125	.201	.126	.275	.175	.130	.084	.049	.147	1.000	.116	.110	.009	.145	.113
18	.273	.053	.054	.033	.069	.008	.073	.020	.016	.030	.026	.023	.130	.057	.078	.020	.116	1.000	.063	.045	.042	.005
19	.072	.080	.110	.120	.081	.114	.072	.053	.107	.078	.109	.191	.097	.050	.030	.056	.110	.063	1.000	.001	.090	.066
20	.062	.038	.033	.028	.003	.004	.034	.032	.015	.032	.005	.008	.055	.024	.149	.018	.009	.045	.001	1.000	.008	.041
21	.031	.048	.076	.180	.128	.237	.071	.205	.202	.176	.255	.175	.037	.049	.020	.198	.145	.042	.090	.008	1.000	.087
22	.026	.102	.107	.246	.110	.151	.077	.156	.193	.167	.127	.171	.052	.094	.060	.140	.113	.005	.066	.041	.087	1.000

1 - Benavente-Hormigueros
2 - Botijas 1
3 - Botijas 2
4 - Cubuy
5 - Gurabo
6 - Las Piedras
7 - Negro-Corozal
8 - Ouque
9 - Pena Pobre-Naguabo
10 - Pico del Este
11 - San Lorenzo
12 - Cayey
13 - Cerro Maravilla
14 - Corozal Substation
15 - Dos Bocas
16 - Fajardo
17 - Gurabo Substation
18 - Maricao
19 - Ponce
20 - San Sebastian
21 - Yabucoa
22 - San Juan WSFO

Station	Percent zeros	Mean	Station	Percent zeros	Mean
Benavente-Hormigueros	79.5	.009	Cayey	80.7	.007
Botijas 1	72.3	.007	Cerro Maravilla	78.7	.011
Botijas 2	81.0	.007	Corozal Substation	87.7	.007
Cubuy	75.8	.012	Dos Bocas	89.4	.010
Gurabo	86.2	.008	Fajardo	87.8	.008
Las Piedras	92.1	.010	Gurabo Substation	92.8	.008
Negro-Corozal	90.8	.007	Maricao	89.1	.013
Ouque	83.5	.011	Ponce	94.4	.004
Pena Pobre Naguabo	78.6	.011	San Sebastian	78.0	.014
Pico del Este	60.8	.019	Yabucoa	91.2	.010
San Lorenzo	87.7	.011	San Juan WSFO	92.3	.006

Table 2: Percent of hours that stations recorded no rainfall, and the mean hourly rainfall in inches

This is the occurrence between San Sebastian and Ponce. The covariance between these two stations is a very small negative value.

Stol (1972) and Sharon (1974) prescribe removing all hours in which none of the stations records rainfall. If it is raining somewhere on the island at a particular hour, then that hour is considered to be "potentially rainy." We constructed a new $(n \times p)$ data matrix that includes only potentially rainy hours. Table 3 shows that the percent of hours with no rainfall over the entire island decreases from between two to nine percent for each station from the original data set. The mean for each station more than doubles, but still lies an order of magnitude smaller than gauge resolution. For this reason, we consider all of the hours in our data set, not just potentially rainy ones. The covariance structure of our factor analysis does not differ significantly from an analysis performed using only potentially rainy hours, as we shall see in Chapter 5. Again, all values of $(x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)$ are positive

Station	Percent zeros	Mean	Station	Percent zeros	Mean
Benavente-Hormigueros	76.2	.023	Cayey	75.9	.018
Botijas 1	70.4	.019	Cerro Maravilla	74.0	.029
Botijas 2	77.9	.019	Corozal Substation	84.0	.020
Cubuy	71.0	.030	Dos Bocas	85.5	.026
Gurabo	82.5	.020	Fajardo	84.0	.020
Las Piedras	84.7	.027	Gurabo Substation	88.5	.020
Negro-Corozal	87.2	.019	Maricao	84.5	.034
Ouque	75.6	.030	Ponce	93.3	.010
Pena Pobre Naguabo	72.2	.029	San Sebastian	71.8	.037
Pico del Este	53.0	.050	Yabucoa	84.6	.027
San Lorenzo	80.0	.028	San Juan WSFO	82.0	.016

Table 3: Data are for potentially rainy hours. Percent of hours that stations recorded no rainfall, and the mean hourly rainfall in inches

except at an hour when station i records rainfall and station k does not.

4.2 Upper Bound on the Number of Factors

A key decision to make in any factor analysis (also in principal component analysis) is how many factors are necessary to best describe the covariance relationships among the variables. Since there are no optimal procedures for doing this across all applications of factor analysis, we approach the problem with an emphasis on trial and error. To get started, however, we use a Monte Carlo procedure that provides an upper bound on the number of statistically significant factors.

The Monte Carlo approach begins by determining the spectrum of eigenvalues ($\mathbf{\Delta}^{\text{data}}$) for the matrix $\mathbf{C}^{\text{data}} = \mathbf{R}^{\text{data}} - \mathbf{\Psi}^{\text{data}}$ shown as a scree plot in Figure 3. (The superscript

"data" refers to the original data set.) As is typical, the first several eigenvalues explain a large portion of the total variance with latter ones explaining a decreasing amount. We next generate twenty-two surrogate rainfall records by randomly permuting the 55,080 hours within each station and compute, as before, a correlation matrix $\mathbf{R}^{\text{surr}}$. Then, as was done with the original data, we determine the spectrum of surrogate eigenvalues ($\mathbf{\Delta}^{\text{surr}}$) from the matrix $\mathbf{C}^{\text{surr}} = \mathbf{R}^{\text{surr}} - \mathbf{\Psi}^{\text{surr}}$. Repeating the entire procedure 100 times gives us a distribution of surrogate eigenvalues, and we choose the magnitude of the 5th largest eigenvalue for each mode as the 95% significance level (Overland and Preisendorfer 1982, Elsner and Tsonis 1991).

This 95% significance level is shown as the dashed line in Figure 3. The leading nine original data eigenvalues exceed this significance level and thus provide an upper bound on the number of factors to allow in our final analysis. Since scrambling the rainfall amount for each hour destroys the serial correlation in the data, the significance level represents white noise. Because we are only concerned with using these results as a guide, a more elaborate test against red noise was not considered.

In summary, having nine eigenvalues exceed the significance level indicates that we should choose no more than nine factors in our analysis. In other words, each rainfall recording station does not by itself represent a unique rainfall region. Stations may be grouped into regions, as long as the number of regions does not exceed nine. The significance testing provides an important "first guess" as to how many factors we should consider in our analysis.

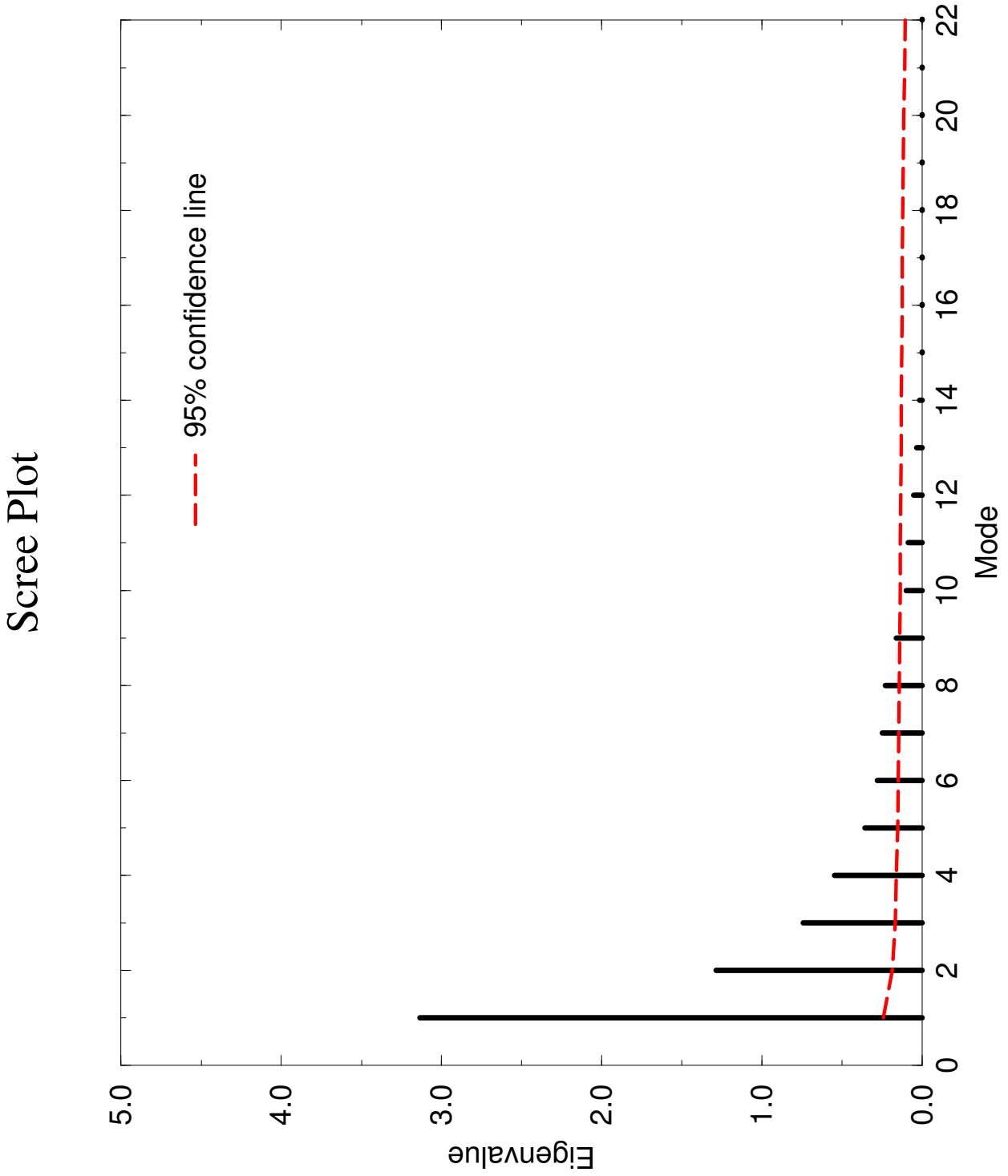


Figure 3: Scree plot showing the leading sixteen eigenvalues of the matrix $\mathbf{R} - \mathbf{\Psi}$. The dashed line represents the 95% significance line from a Monte Carlo simulation of white noise.

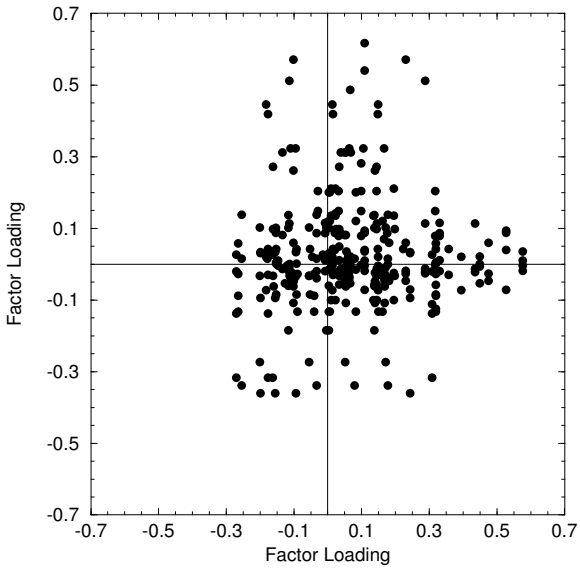


Figure 4: Composite of the pairwise plots of unrotated factor loadings from a factor analysis model with $m = 6$ factors.

4.3 Orthogonal Rotation

A useful way to diagnose a factor analysis is to examine pairwise plots of the factor loadings. As an example Figure 4 shows the composite of pairwise plots for m=6 ($\mathbf{\Lambda}_i$ versus $\mathbf{\Lambda}_j$; for $i = 1, m-1$ and $j = i, m$). In other words, the loading of a station on factor i is plotted against the loading of that station on factor j. Many of the points cluster either along the axis or near the origin indicating linear independence of the loadings (simple structure). There are, however, many points that lie off the axis indicating stations that are included in more than one common factor (complex structure). Because of the linear independence of common factors, interpreting the results from factor analysis (or principal component analysis) is easier when simple structure is present (Richman 1986).

To improve the simple structure, the loading matrix is multiplied by an orthogonal

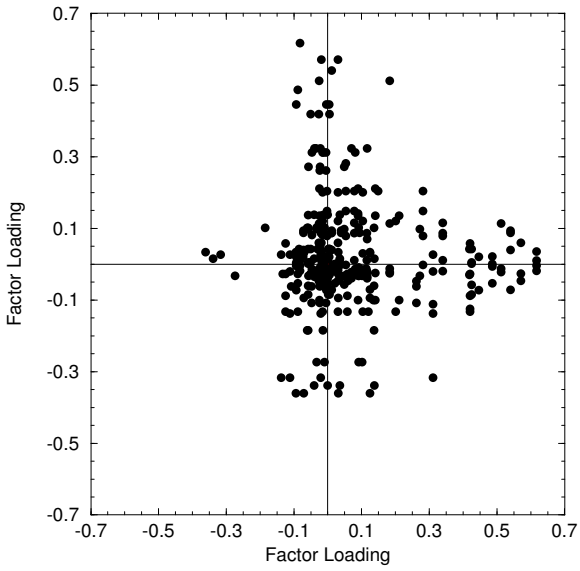


Figure 5: Composite of the pairwise plots of rotated factor loadings from a factor analysis model with $m = 6$ factors.

matrix $\mathbf{T}$. This linear transformation of the loading matrix has the property of conserving the inner product of the loading vectors (columns of $\mathbf{\Lambda}$) and geometrically represents a rigid rotation about the coordinate axis (Kreyszig 1993). We choose a varimax rotation and perform the calculations using IMSL (1987) subroutine FROTA.

Pairwise plots of the orthogonally rotated factor loadings for $m = 6$ are shown in Figure 5. The points better align along the coordinate axes indicating improved simple structure. This shows that the stations are loading on one particular factor and not on any others. There are exceptions, denoted by the stray points in quadrants one and four. These points are relatively few suggesting that a different type of rotation, such as oblique (Jolliffe 1986), will not be better in improving simple structure.

Chapter 5

Regionalization

5.1 Selecting the Number of Factors

Using the white noise significance test and orthogonal rotation of the previous section as guides we now proceed to regionalize Puerto Rico. We apply the factor analysis model with orthogonal rotation for each m in the interval 1 through 9 and carefully examine the factor loadings. We want to find an m for which all 22 stations optimally load on only one factor. No value of m between 1 and 9 allows each station to load on one and only one factor. Although no choice of m revealed perfect simple structure, $m = 6$ optimally loads the stations onto common factors, as we shall see later. Table 4 shows the common loadings for $m = 6$. These loadings are normalized by taking the absolute value of each factor loading for each station, summing these values for each station, dividing the original loadings by this sum, and multiplying by 100 (Equation 7)

$$\hat{\lambda}_{ij} = 100 \times \lambda_{ij} / \sum_{j=1}^{m} |\lambda_{ij}|. \tag{7}$$

It is the magnitude of the factor loadings that will determine the regions, since it is contributing to the common variance. Communality is defined by the sum of the squares of

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1	-1.89	0.18	-1.29	16.41	-1.12	2.25
2	0.41	19.15	-3.44	1.53	3.50	4.72
3	1.97	9.96	-0.10	3.09	5.55	10.95
4	15.77	-2.06	0.24	0.04	-3.16	2.25
5	3.38	-0.32	-1.58	3.99	-10.20	-1.72
6	11.54	-3.94	-0.95	-5.40	-11.80	1.40
7	-3.04	21.88	-0.91	1.37	0.37	-0.21
8	1.11	-0.67	27.26	-1.06	2.23	-2.47
9	15.58	-4.43	2.77	-3.45	-4.95	-1.45
10	6.79	-0.92	24.44	-0.86	4.21	-0.70
11	4.64	-2.52	1.48	-3.68	-13.44	1.83
12	4.30	2.48	-1.86	-0.86	-1.30	17.34
13	-0.85	1.74	-2.72	10.65	-1.23	5.26
14	-3.27	17.27	-0.29	0.82	0.00	-2.84
15	3.34	5.00	-1.15	8.26	5.02	-5.42
16	-0.19	-0.85	12.51	-1.88	-4.02	-3.33
17	1.30	-0.04	-1.91	5.40	-12.62	0.86
18	-3.45	0.14	-0.19	17.46	-2.72	1.07
19	-0.19	-0.53	-2.20	3.17	-0.48	16.75
20	3.82	1.06	-0.72	7.83	4.47	-7.14
21	-0.56	-2.06	6.54	-2.35	-6.89	5.42
22	12.62	2.80	5.49	0.43	-0.74	4.62

Table 4: Normalized common factor loadings from the common factor analysis after a varimax orthogonal rotation. Factor loadings are divided by the sum of the absolute value of loadings within each factor. This value is multiplied by 100 to give the normalized factor loading.

the factor loadings (Johnson and Wichern 1982), so negative factor loadings will contribute positively to the common variance. The importance of the sign is that within a particular factor, stations of common sign should be grouped. This will become apparent when we consider thresholds.

Since we are looking for stations that load heavily (large magnitude) on a particular factor, we can choose a threshold to define a primary loading as one that has a normalized magnitude of 7.50 or greater. Similarly, we define a secondary loading as having a magnitude

	Factor					
Station	1	2	3	4	5	6
1				**16.41**		
2		**19.15**				
3		**9.96**		5.55		**10.95**
4	**15.77**					
5					**-10.20**	
6	**11.54**			-5.40	**-11.80**	
7		**21.88**				
8			**27.26**			
9	**15.58**					
10	6.79		**24.44**			
11					**-13.44**	
12						**17.32**
13				**10.65**		5.26
14		**17.27**				
15		5.00		**8.26**	5.02	-5.42
16			**12.51**			
17				5.40	**-12.62**	
18				**17.46**		
19						**16.75**
20				**17.46**		
21			6.54		-6.89	5.42
22	**12.62**		5.49			

Table 5: Threshold loadings for orthogonally rotated, normalized six factor analysis. Boldface values are primary loadings with magnitudes of 7.50 or greater. The rest of the values are loadings with magnitudes between 5.00 and 7.49.

between 5.00 and 7.49. Anything below magnitude 5.00 is not considered sufficient to be included in the factor. Loadings that exceed 5.00 are shown in Table 5. Boldface values are primary factor loadings and the rest are secondary loadings. Note that within each factor column, all of the primary factor loadings are of the same sign.

As indicated by the separation of primary loadings in Table 4, and by the appearance of simple structure in Figure 5, there is good clustering (grouping) of stations into common factors. There are, however, some exceptions. Botijas-2 (station 3) and Las Piedras (station 6)

surpass the primary loading threshold on two separate factors. We dub these "freeloading" stations since they are free to load on more than one factor. Yabucoa (station 21) does not surpass the primary loading threshold for any factor. We call this a "homeless" station since it cannot be placed in any region based upon our primary threshold values. So for six factors, the sum of freeloading and homeless stations is three. We call this the "nonsingularity" sum. If every station loaded on one, and only one, factor this sum would be zero and would represent an ideal factor analysis for which little subjectivity would be required. As mentioned above, we applied factor analysis for $m = 1$ to 9, where m is the number of common factors. We found that $m = 6$ provided the smallest nonsingularity sum. The number of white noise significant modes gives us our upper bound for m, and the minimization of the nonsingularity sum gives us our exact number of factors in our regionalization of Puerto Rico.

The regionalization using six common factors is shown in Figure 6. Repeating, we have used the common factor model with an orthogonal rotation of the loading matrix and a minimization of the nonsingularity sum to achieve this map. The regions are divided based on primary loadings on each station. For the three nonsingular stations we have drawn the line on (or very close) to their locations.

We have tested the stability of this regionalization with respect to different temporal subdomains (not shown). This was done by repeating this factor analysis procedure on two separate data subsets. The first data subset consisted of only daytime hours and the second consisted only of potentially rainy hours. For both subsets the procedures resulted in a nearly identical regionalization as that given by the full data set. This lends confidence

Figure 6: Geographic regionalization of Puerto Rico based on a factor analysis (with an orthogonal rotation of the loadings) of summertime convective rainfall.

that the factor analysis we performed is representative of the underlying rainfall regions.

5.2 Diurnal Variability

Since our goal is to develop a prediction model for forecasting convective rainfall over the island, we examine the diurnal variability of precipitation in each of the six regions. We do this by considering the empirical probability of measurable precipitation for each hour. Figure 7 shows the frequency of rainfall (excluding rainfall from tropical cyclones) for each hour of the day for each of the six regions. Factors 2, 4, and 6 exhibit high amplitude maxima in the late afternoon and minima between midnight and 4 a.m. These three factors comprise the regions in the western two-thirds of the island. Factors 1, 3, and 5 are characterized by low amplitude maxima occurring in the early morning and minima taking place between 8 p.m. and midnight. These factors correspond to regions on the island's eastern third. Factors 1, 3, and 5 also show much less hourly variability than factors 2, 4, and 6. Based on their hourly frequency signatures, then, we can further separate the island into two larger regions: A western "super-region" and an eastern "super-region."

We could have initially divided Puerto Rico into two regions by choosing $m = 2$ and performing a factor analysis on the data. The normalized primary and secondary factor loadings for such an analysis are shown in Table 6. The nonsingularity sum for the $m = 2$ analysis is 12, compared to 3 for the $m = 6$ analysis. We can only place ten stations using our original threshold criteria. The rest must be placed subjectively. Even if we place stations that only satisfy the secondary loading criteria (normalized loadings greater than

	Factor	
Station	1	2
1		
2		**12.76**
3		**9.66**
4	**8.95**	
5	6.71	
6	**10.16**	
7		**14.13**
8	7.24	
9	**9.73**	
10	**8.49**	
11	**8.50**	
12	5.23	5.66
13		5.20
14		**10.40**
15		5.28
16	5.71	
17	6.74	
18		
19		
20		
21	5.82	1.28
22		

Table 6: Threshold loadings for orthogonally rotated, normalized two factor analysis. Boldface values are primary loadings with magnitudes of 7.50 or greater. The rest of the values are loadings with magnitudes between 5.00 and 7.49.

5.00), we still have six stations that cannot be placed by our objective procedures.

By choosing the number of factors that minimized the nonsingularity sum, we reveal six distinct regions of convective rainfall patterns. Upon examining the frequency of rainfall for the six regions (Figure 7), we see a dichotomy emerge. Had we begun with the assumption of dichotomy, the factor analysis would have not yielded six regions, and we would have had poor guidance in regionalizing most of the stations.

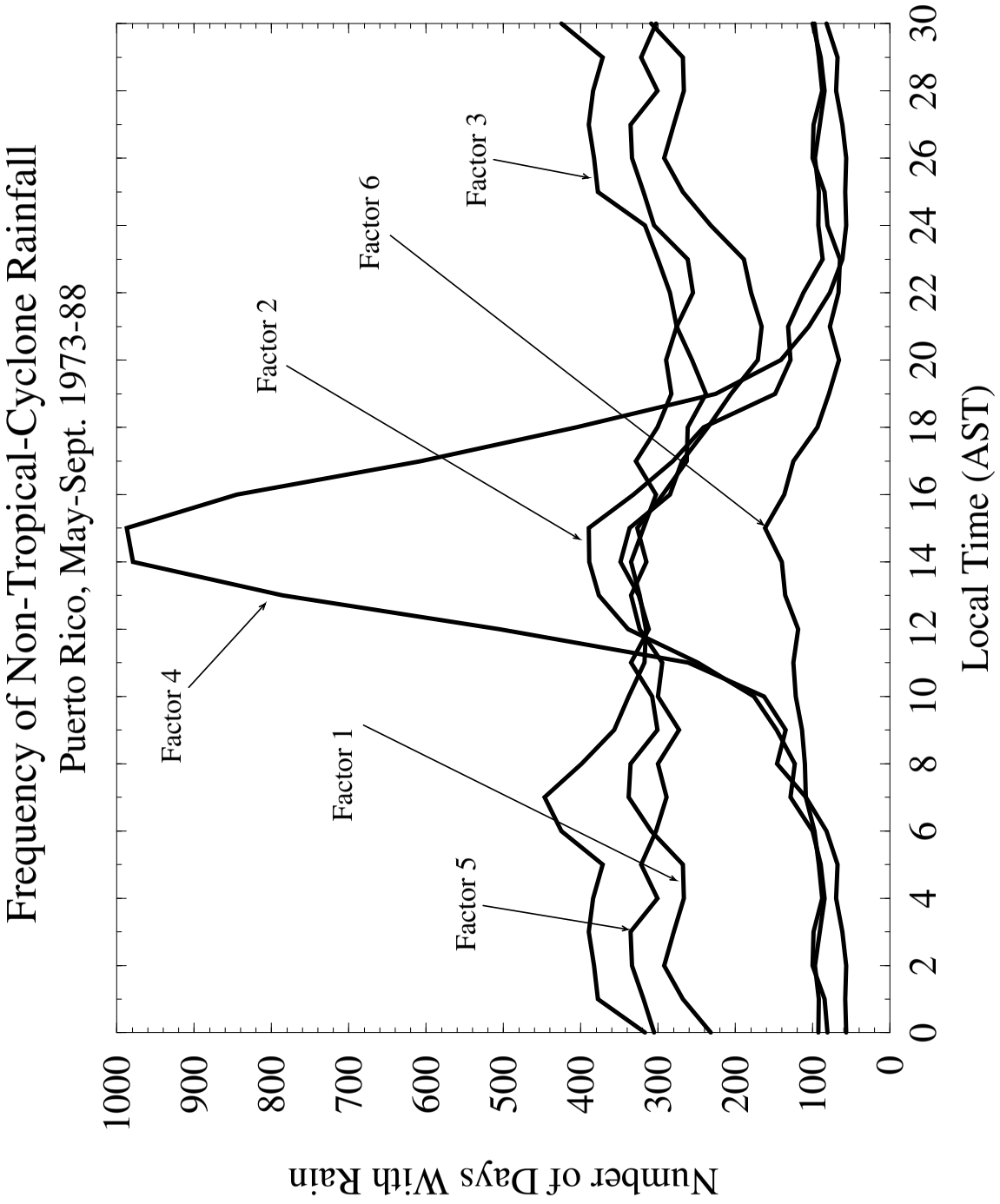


Figure 7: The number of rainfall events, July 1973 through June 1988, is plotted for each hour of the day. Hours 24 through 30 represent "wraparound" times corresponding to midnight through 6 a.m. A rainfall event is defined as any amount greater than a trace that was recorded at any station within a region. Rainfall associated with tropical cyclones is excluded.

Chapter 6

A Rainfall Prediction Scheme

Here we attempt to build a prediction algorithm for daytime convective rainfall. On many conditionally unstable days, convective rainfall is on a small enough scale that it may occur in the vicinity of a recording station, but never reach the rain gauge. On another day with similar convective instability, the station may actually record rainfall. In both cases, rainfall occurred due to similar forcing, but in the data record, it only rained on one day. This aspect of diurnal convective rainfall makes it very difficult to predict. This is especially true in Puerto Rico during the summer because variation in such variables as temperature, dew point, and wind direction is small on a diurnal basis.

By regionalizing the island, we capture stations that exhibit similar rainfall signatures, which reflect shared convective forcing, and group them within a boundary. A convective rain shower that misses one gauge within a region may be recorded by another. Treating a small region as a single collection point is more reflective of the prevailing forcing.

We saw in Chapter 5 that the six regions of the island comprise larger "super-regions" based on their diurnal rainfall frequency signatures. San Juan lies within the eastern super-region, and we attempt to develop our prediction model for this area. For our predictor

variables, we use rainfall data for the eleven recording stations within the eastern super-region, and the surface-based and upper-air data from the San Juan Weather Service Forecast Office (WSFO). The San Juan WSFO is the station in Puerto Rico that has the most extensive hourly surface data record, and the only one that retrieves upper air data.

6.1 Ordinary Least-Squared Regression Model

We select a set of *a priori* predictor variables denoted by the vector $\mathbf{x}$. We also choose a dependent variable $\hat{y}$ to be the predicted outcome. We employ a linear regression model of the form

$$\hat{y} = f(x) \tag{8}$$

and

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n \tag{9}$$

where n is the number of predictors used in the regression, and β_i is the ordinary least-squared (OLS) determined coefficient for predictor variable x_i (Johnson and Wichern 1982). The least squares solution for the column vector of coefficients is $\beta = (X^T X)^{-1} X^T y$. X is the matrix of predictor variables, and the superscripts T and -1 denote matrix transpose and inverse respectively (Elsner and Schmertmann 1994).

6.2 Data Set and Selection of Variables

We begin by assembling a data set of 420 randomly selected hours from the period 1 May, 1977 through 30 September, 1981. We seek to predict daytime convective rain, so our

predictor data is taken at 1200 UTC (8 a.m. AST). This is the time at which we initialize the model. The dependent data is taken up to 0000 UTC (8 p.m. AST). Therefore, the predictor data set of 420 randomly selected hours will only include 8 a.m. hours between 1 May, 1977 and 30 September, 1981, and the dependent data occurs twelve hours later.

The surface-based components of the predictor variable vector **x** are

- **Dew point**. This measure of atmospheric moisture content at the surface has little diurnal variability. In the data record, the dew point is between 71° F and 79° F for 94 percent of all hours.

- **Temperature**. This surface variable also shows little day to day variation. For 98 percent of all hours, the temperature fell between 70° F and 90° F.

- **U and V wind components**. Easterly winds prevail over the island during the summer months. In the morning at San Juan, winds have a southerly component due to the land breeze propagating offshore. During the afternoon, the wind takes on a northerly component as a sea breeze develops and moves onshore (Ruffner and Bair 1978). Changes in wind speed and direction may have a significant effect on diurnal rainfall in Puerto Rico (Gere Gallup, personal communication).

- **Sea level pressure anomaly**. We average the sea level pressure for each hour of the day throughout the entire data period and then subtract the appropriate mean from each pressure record in the data set. For instance, the first record is for 19 August, 1980 at 8 a.m. AST. The average sea level pressure for 8 a.m. is subtracted from the sea level pressure for 19 August, 1980 at 8 a.m., and we get a sea level pressure

anomaly. In this way, the semi-diurnal pressure oscillation is removed from the data set.

- **Twelve hour sea level pressure anomaly tendency**. We take the twelve hour trend of the sea level pressure anomaly described above. For this data set, the tendency measured is for 8 p.m. previous day through 8 a.m. current day.

- **Past one hour rainfall total**. The rainfall amounts for all eleven rain gauges in the eastern super-region for the past hour are summed. For this data set, the recorded hour is 7 a.m. through 8 a.m.

- **Past three hour rainfall total**. The rainfall amounts for all eleven eastern stations are summed over the past three hours, 5 a.m. through 8 a.m.

- **Past twelve hour rainfall total**. Rainfall amounts are summed over the past twelve hours, 8 p.m. previous day through 8 a.m. current day.

- **Percent of stations recording rainfall, past one hour**. The percent of stations in the eastern super-region that reported rainfall between 7 a.m. and 8 a.m. is included. This predictor variable gives insight into the coverage of rainfall over the eastern third of the island during the past hour.

The upper air predictor variable is

- **70 kPa relative humidity**. The 70 kPa relative humidity value for the current 1200 UTC sounding is included in the independent variable vector $\mathbf{x}$. The atmosphere over Puerto Rico during summer is almost always convectively unstable (Noel LaSeur,

personal communication), yet it does not rain every day. A column of deep moisture may be critical in initiating rainfall in a convectively unstable atmosphere (Fuelberg and Biggar 1994). Relative humidity values of less than twenty percent are coded as a single, phantom value in the data record, so only values of twenty percent or greater contribute to the construction of the model. While we recognize that this may be a limiting factor on the optimization of our model's performance, particularly on very dry days, we shall see later that 70 kPa relative humidity is not selected as a predictor.

We choose $\hat{y}$, the variable for which we are predicting, to be one of the following:

- **Twelve hour rainfall total**. Rainfall between 8 a.m. (time of initialization) and 8 p.m. is summed over all stations in the eastern super-region.
- **Three hour rainfall total, nine hour lead**. We sum rainfall between 5 p.m. and 8 p.m. over all stations. The lead time of our forecast is nine hours, the elapsed time between 8 a.m. and 5 p.m.
- **Six hour rainfall, six hour lead**. We sum rainfall between 2 p.m. and 8 p.m. over all stations. The intent of six hour lead time is to predict convective rain over the afternoon hours.

6.3 Building a Linear Regression Model

In building a OLS model according to Equation 9 using a subset of predictor variables in $\mathbf{x}$, we begin by individually correlating each of the eleven components of $\mathbf{x}$ with twelve hour

Independent Variable	Correlation Coefficient
Dew point	-0.0960
Temperature	-0.1900
U wind component	0.0000
V wind component	0.0000
Sea level pressure anomaly	0.1329
12 hour sea level pressure anomaly tendency	0.1564
70 kPa relative humidity	0.1659
Past one hour rainfall total	0.2256
Past three hour rainfall total	0.2425
Past twelve hour rainfall total	0.2584
Percent of stations recording rainfall	**0.2747**

Table 7: Correlation coefficients of an out of sample OLS regression model with $n = 1$ predictor variable. The predictand is the twelve hour eastern super-region rainfall total ending at 8 p.m.

rainfall total, our initial choice for $\hat{y}$. This is done by building a linear regression equation with $n = 1$. We use an out of sample approach by removing one of the 420 randomly selected hours for which we from the predictor data set, and then build the model predicting for the $\hat{y}$ corresponding to the hour we removed. Table 7 shows the correlation coefficient between $\hat{y}$ and observed rain, $\bar{y}$, based on our out of sample, one variable OLS model.

The highest correlation coefficient in Table 7 is in boldface, and it corresponds to the percent of stations in the eastern super-region that recorded rainfall during the past hour (7 a.m. to 8 a.m.). This is the single best variable for predicting an eastern super-region twelve hour rainfall total ending at 8 p.m. We retain this variable as x_1, and regress the other predictors individually as x_2 for an OLS model based on $n = 2$ predictor variables. Again we use an out of sample approach, and generate correlation coefficients between $\hat{y}$ and $\bar{y}$. We then retain the combination of variables that yields the highest coefficient. This procedure is repeated for $n = 3$, $n = 4$, and so on, fixing an additional variable at each

Component	Independent Variable	Cumulative Corr. Coeff.
x_1	Percent of stations reporting rainfall	0.2747
x_2	Past twelve hour rainfall	0.3030
x_3	12 hour sea level pressure anomaly tendency	0.3243
x_4	Sea level pressure anomaly	0.3309
x_5	U wind component	0.3332

Out-of-sample correlation coefficient: 0.3332

Table 8: Components of optimal predictor variable vector $\mathbf{x_{opt}}$

step, until adding another predictor variable no longer increases the correlation coefficient from the previous iteration. In this way we perform a stepwise regression in building an OLS model to predict for twelve hour rainfall.

6.4 Results

By "optimizing" the OLS model in this manner, we find that the best vector of predictor variables for predicting $\hat{y}$ is given by $\mathbf{x_{opt}}$, with $n = 5$. These components are shown in Table 8. It may seem odd that while the u wind component, which has a "stand alone" correlation coefficient of 0.000 appears in $\mathbf{x_{opt}}$, 70 kPa relative humidity (correlation coefficient of 0.1659) does not. The predictor variables are not necessarily independent of each other (they are chosen *a priori*), so the influence of 70 kPa relative humidity on predicting twelve hour rain is assimilated in one or more components of $\mathbf{x_{opt}}$.

By stepwise regressing the components of $\mathbf{x_{opt}}$ against twelve hour super-region rainfall, we seek to predict for daytime convective rain. The components given in Table 8, however

may be indicative of synoptic scale easterly waves described in Chapter 1. All components in $\mathbf{x_{opt}}$ correlate positively with twelve hour rainfall. By looking at each component, we can see how they may contribute to predicting synoptic scale rainfall.

- **Percent of stations reporting rainfall**. If a large percentage of stations in the super-region reported rainfall during the past hour, a large scale easterly wave may be forcing rainfall over the entire island.

- **Past twelve hour rainfall**. Convective scale rain events occur on the order of an hour or two. Rainfall events that last the entire night (past twelve hours) may be on a synoptic scale.

- **12 hour sea level pressure anomaly**. Typical easterly waves resemble troughs that bow northward. Riehl (1954) noted precipitation associated with these waves often occurs on the eastern side of the trough axis, where the sea level pressure anomaly is not at a minimum.

- **12 hour sea level pressure anomaly tendency**. Rising anomalies occur east of the trough axis as the wave passes to the west. Again, this is where we expect to find rain associated with an easterly wave.

- **U wind component**. On the eastern side of the trough axis, the u wind component increases as the axis passes to the west. Similarly, the v wind component decreases. The addition of this predictor barely increases the cumulative correlation coefficient; it is the weakest of the five predictor components.

The components of the $\mathbf{x_{opt}}$ vector merely suggest the ability to predict for rainfall associated with passing easterly waves, as opposed to diurnal afternoon convection. Easterly waves may enhance or suppress convection, so the separation of days based on synoptic influence and non-synoptic influence is difficult. We take trepidation in identifying $\mathbf{x_{opt}}$ as synoptic predictors, especially in lieu of the small out of sample correlation coefficient.

In addition to calculating the correlation coefficient for the linear regression model based upon independent variables $\mathbf{x_{opt}}$, we also calculate the root mean squared (r.m.s.) error and the mean absolute error for the twelve hour, eleven station rainfall total. The r.m.s. error and mean absolute error are given by Equations 10 and 11 respectively.

$$\text{r.m.s. error} = \sqrt{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2/n} \tag{10}$$

$$\text{mean absolute error} = \sum_{i=1}^{n} |\hat{y} - \bar{y}| \tag{11}$$

The total number of observations is n, $\hat{y}$ is the predicted twelve hour, eleven station rainfall total, and $\bar{y}$ is the actual total. The r.m.s. error for twelve hour precipitation, 8 a.m. to 8 p.m., is 1.15 inches across all eleven stations. The mean absolute error for the same period is 1.49 inches. We also predict for the nine-hour-lead three hour total and the six-hour-lead six hour total. The predictor variable vector $\mathbf{x_{opt}}$ is the same for these forecasts as it is for the twelve hour forecast. We optimize only once based upon our choice of twelve hour rainfall for dependent variable $\hat{y}$.

For comparison, we determine a persistence forecast for twelve hour rainfall by summing rain amounts over eastern super-region stations for the past twelve hours. So the persistence forecast is the rainfall that fell between 8 p.m. previous day and 8 a.m. current day. We

use an out of sample approach, as we did in building the linear regression model. From the hourly rainfall frequency (Figure 7), there is little indication of a diurnal rainfall frequency for the three eastern regions. This suggests that the previous day's rainfall amount (8 a.m. to 8 p.m.) may be no better than overnight rainfall as a persistence forecast for the current day's rainfall total. For a $\hat{y}$ of twelve hour rainfall, and persistence as the sole predictor, the r.m.s. error is 1.22 inches, the mean absolute error is 1.58 inches, and the correlation coefficient is 0.2584 (Table 9). The linear regression model, whose components are given in Table 8, is more accurate than persistence for predicting rainfall during all three time periods.

We also determine a climatology forecast by taking the mean twelve hour rainfall across all eleven stations over the entire data set of randomly selected hours. Again, we use an out of sample approach by removing one of the hours from the predictor variable set. Each hour in the data set has a corresponding rainfall total for the previous twelve hours. It is these corresponding totals that we average to get a climatology prediction. Choosing $\hat{y}$ to be twelve hour rainfall, and climatology as the sole predictor, the r.m.s. error is 2.03 inches and the mean absolute error is 1.69 inches. Again, our simple linear regression model is more accurate at predicting rainfall for all three time periods. These forecast results are also shown in Table 9.

	9 hr. lead, 3 hr. total	6 hr. lead, 6 hr. total	12 hr. total
Linear Regression Model			
Root mean squared error	1.27	1.25	1.15
Mean absolute error	0.65	1.12	1.49
Correlation coefficient	0.1887	0.2134	0.3332
Persistence			
Root mean squared error	1.31	1.28	1.22
Mean absolute error	0.67	1.15	1.58
Correlation coefficient	0.1435	0.1858	0.2584
Climatology			
Root mean squared error	1.99	1.93	2.03
Mean absolute error	0.71	1.21	1.69

Table 9: OLS regression model results. All errors are in inches.

6.5 Cross Validation

Elsner and Schmertmann (1994) emphasize that for out of sample hindcasts, the subset of predictor variables must remain independent of the predicted (omitted) observation. In our study, we developed an OLS model through stepwise linear regression. Since we optimized our model in increments, we become dependent on the correlation coefficient with each step. In other words, we know that the percent of stations reporting rain is the best choice for x_1 because it gives us the highest correlation coefficient. This is knowledge we do not have in developing a model in real time since the predicted hour has not yet occurred.

Additional cross validation is probably warranted, though its further application is not straightforward. If we permute all eleven *a priori* independent variables in building a linear regression model, we may cross validate this algorithm by removing two hours for each permutation. The first hour is removed to cross validate the algorithm of permutations, and

an out of sample linear regression model produced by the algorithm predicts for the second hour removed. There exist millions of such permutations. Instead we choose a progressive, somewhat subjective method of building our forecast model based on maximum correlation. The cross-validation of our subjectively based algorithm is not easily conceptualized, even though our model is still developed out of sample. We recognize this as a caveat for calling our model building method truly cross validated.

Chapter 7

Summary and Conclusion

Daytime convective patterns in Puerto Rico during the summer may be divided into six rainfall regions through factor analysis. Since the leading six eigenvalues are significant with respect to white noise, there is evidence of physical mechanisms underlying these six factors. The six regions suggest important mechanisms that force precipitation over the island and indicate that the factor analysis model is sensitive to variations in weather regimes. For example, factors one and five are likely related to convection forced by El Yunque, an isolated mountain to the southeast of San Juan (LaSeur, personal communication), whereas factors two, four and six result from inland sea breeze penetration and/or sea breeze interaction with the interior mountains. Rather than attempting to develop a prediction scheme for a single station, we can build a model that will predict for a convective scale. Whereas a shower may miss a particular station, it will not miss the entire region.

These six regions may further be divided into an eastern region and a western region. We did this based upon hourly rainfall frequency for each of the six factors. The communality of regions within these two larger regions are descriptive of phenomena existing on differing time scales. This is an important consideration in developing a prediction scheme since the

western regions exhibit a strong diurnal frequency change while the eastern regions do not. By regionalizing Puerto Rico based upon its rainfall signature, we made the problem of predicting convective rainfall during the summer more tractable.

We built a prediction model by selecting from a group of *a priori* variables a subset of predictors that correlated with twelve hour rainfall total. Accounting for diurnal frequency differences described above, we incorporated only stations in the eastern super-region of the island into our data set. We chose the eastern super-region because San Juan, the major city and data collection site in Puerto Rico, lies within this part of the island. The simple linear regression model devised in this study is more accurate than both climatology and persistence for predicting daytime convective rainfall. Nevertheless, an r.m.s. error of 1.15 inches over the entire eastern super-region is an average of 0.10 inches per station. An out of sample correlation coefficient of 0.3332 is not an indicator of a highly reliable statistical forecast model.

This study is limited by the number of available rainfall and surface data stations over Puerto Rico, which fixes the spatial resolution of the analysis. A greater spatial resolution will likely alter the above results since the distribution of variance will reflect even more local scale phenomena. Although the eastern stations and western stations share common diurnal rainfall frequency patterns, there are subtle differences between each station. Within a super-region, amplitudes of maximum frequency may be out of phase with each other by an hour of two.

Additional surface data exists from the same NCAR data set, in particular for Roosevelt Roads, Ponce, and Ramey Air Force Base. These stations do not have the historical

extent that the San Juan WSFO does, and only Roosevelt Roads lies within the eastern super-region. Additional research into building a prediction model that incorporates both Roosevelt Roads and San Juan may be worthy of future investigation.

Forecast guidance as it exists now consists primarily of the National Meteorological Center (NMC) Medium Range Forecast Model (MRF), satellite interpretation, and synoptic analyses. NMC attempted to develop Model Output Statistics (MOS) for San Juan, but it performed less accurately than the MRF during the summertime (Rafael Mojica, personal communication). The purpose of developing a linear regression model in this study is not to build a "stand alone" model for forecasting daytime rainfall during the summer, but to provide an additional tool at the discretion of forecasters. Indeed, this model is preliminary, not definitive. The value of the linear regression model is to show how the regionalization of Puerto Rico, based on its convective rainfall history, may improve the accuracy of forecasts.

Appendix A

List of Hurricane Hours Removed

Hurricanes, tropical storms, and named tropical depressions are removed from the data set for guidance considerations. Any such storm that passed within 500 kilometers of Puerto Rico is thought to possibly influence the island's weather on a synoptic scale and is consequently removed. Only hours for which the tropical cyclone is within 500 kilometers of Puerto Rico are extracted from the data set.

Tropical cyclone	Year	First hour removed	Last hour removed
Christine	1973	September 3, 0000 UTC	September 4, 1200 UTC
Fifi	1974	September 14, 1200 UTC	September 15, 1200 UTC
Eloise	1975	September 14, 0000 UTC	September 17, 1200 UTC
Emmy	1976	August 23, 1200 UTC	August 23, 0000 UTC
Juliet	1978	October 8, 0000 UTC	October 10, 1200 UTC
David	1979	August 29, 0000 UTC	August 31, 0000 UTC
Allen	1980	August 4, 1200 UTC	August 5, 1200 UTC
Gert	1981	September 7, 0000 UTC	September 9, 1200 UTC
Debby	1982	September 13, 1200 UTC	September 14, 0000 UTC
Arthur	1984	September 2, 1200 UTC	September 4, 1200 UTC
Gloria	1985	September 23, 0000 UTC	September 24, 0000 UTC
Danielle	1986	September 8, 0000 UTC	September 9, 1200 UTC
Emily	1987	September 21, 0000 UTC	September 23, 1200 UTC

Table 10: Hurricanes, tropical storms, and named tropical depressions removed from the data set.

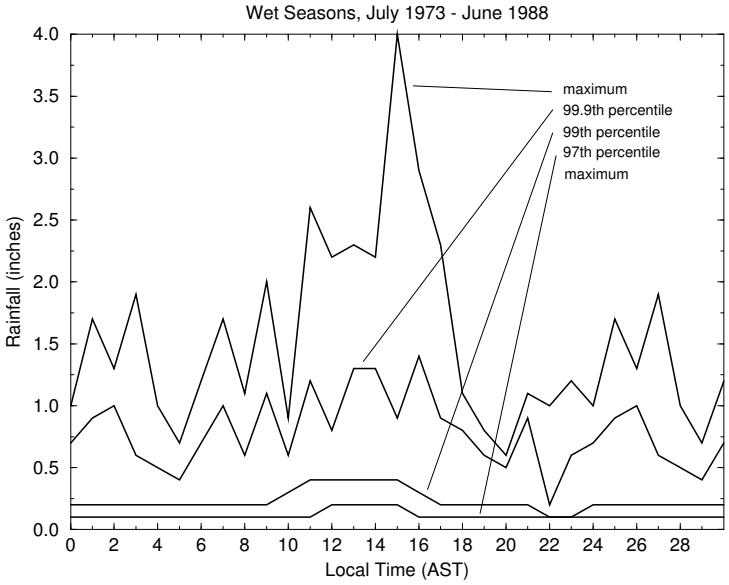
Appendix B

Hourly Rainfall Climatology

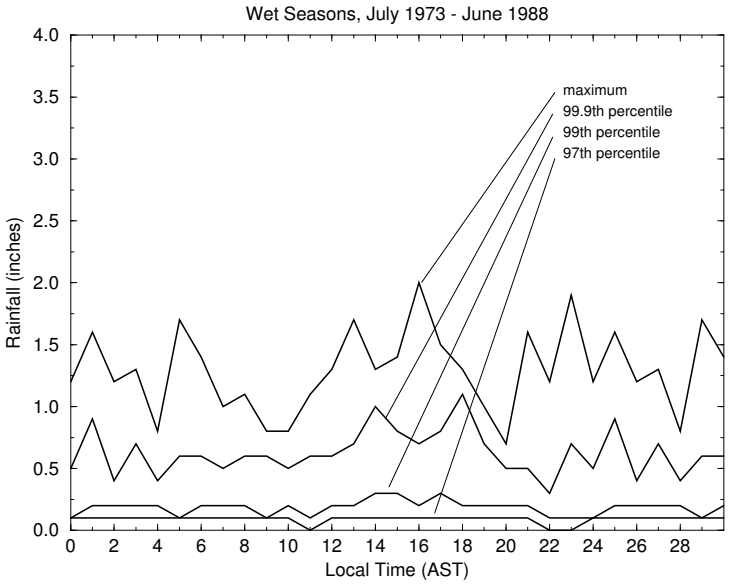
The hourly rainfall climatology for the twenty-two stations in Puerto Rico is shown in this appendix. Stations are arranged alphabetically within the two "super-regions." The first eleven stations represent the eastern third of Puerto Rico, and the second eleven stations represent the western part of the island.

Local time is shown on the abscissa, and rainfall amount is shown on the ordinate. Hours 24 through 30 represent "wraparound" times corresponding to midnight through 6 a.m. Within each hour, rainfall amounts are ordered greatest to least. The top curve is the maximum rainfall amount for each hour. The second curve is the 99.9th percent rainfall value, the third curve is the 99th percent rainfall value, and the bottom curve is the 97th percent rainfall value.

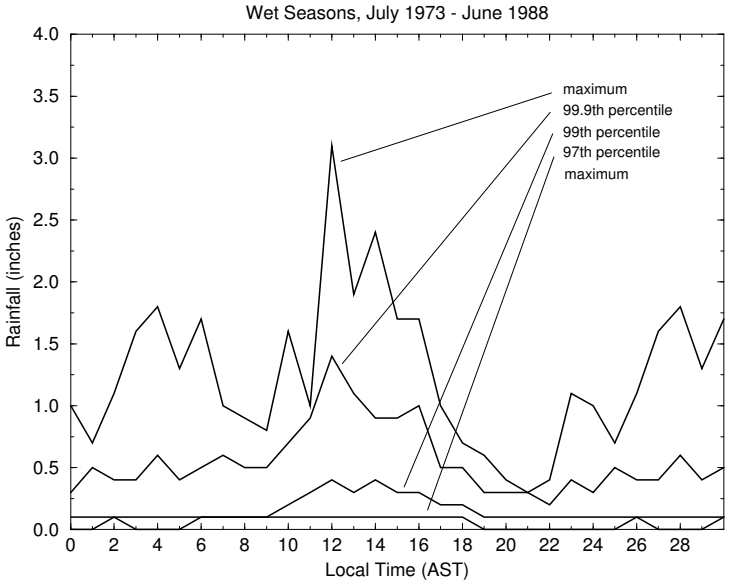
Rainfall Climatology: Cubuy



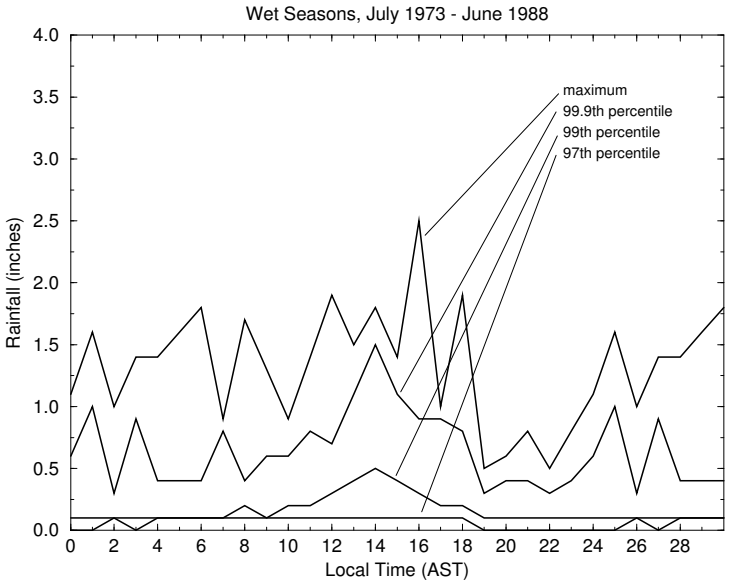
Rainfall Climatology: Fajardo



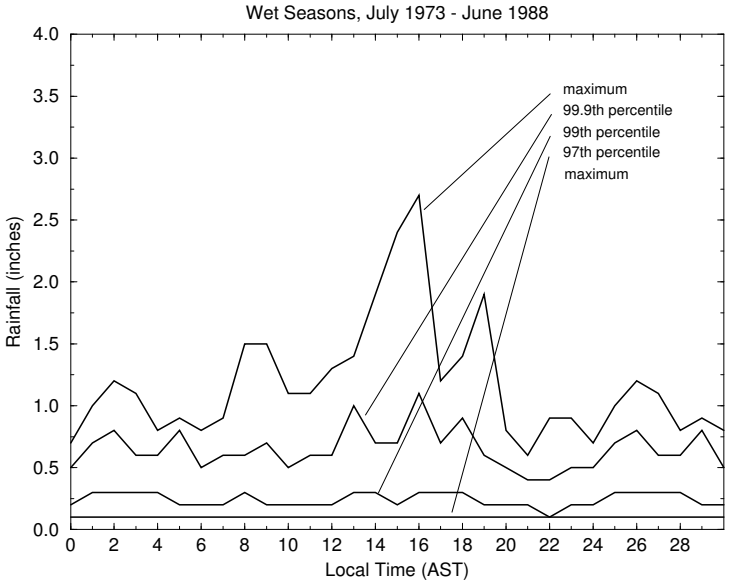
Rainfall Climatology: Gurabo



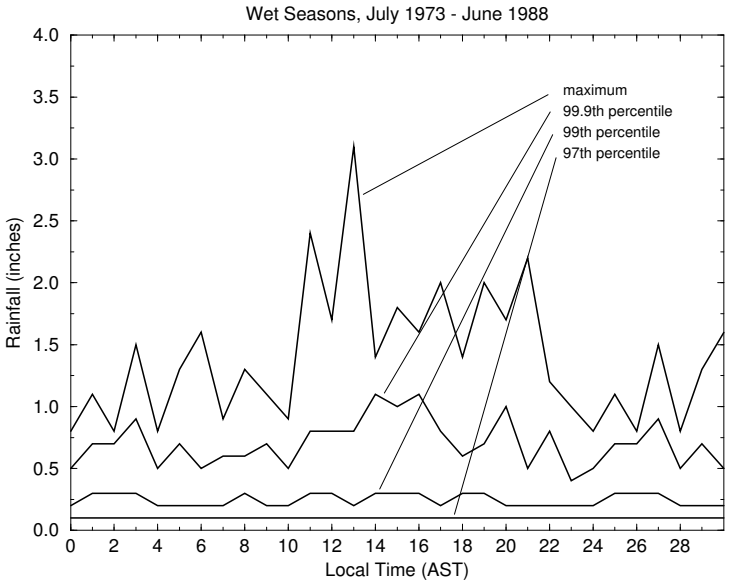
Rainfall Climatology: Gurabo Substation



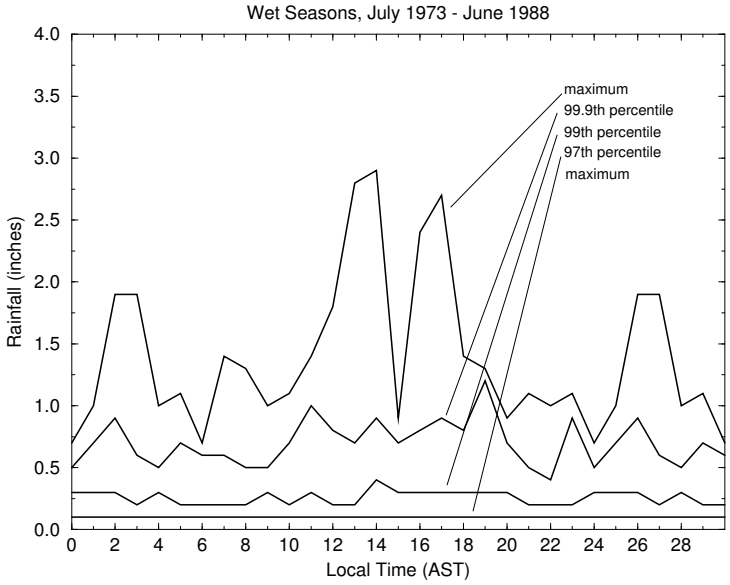
Rainfall Climatology: Las Piedras



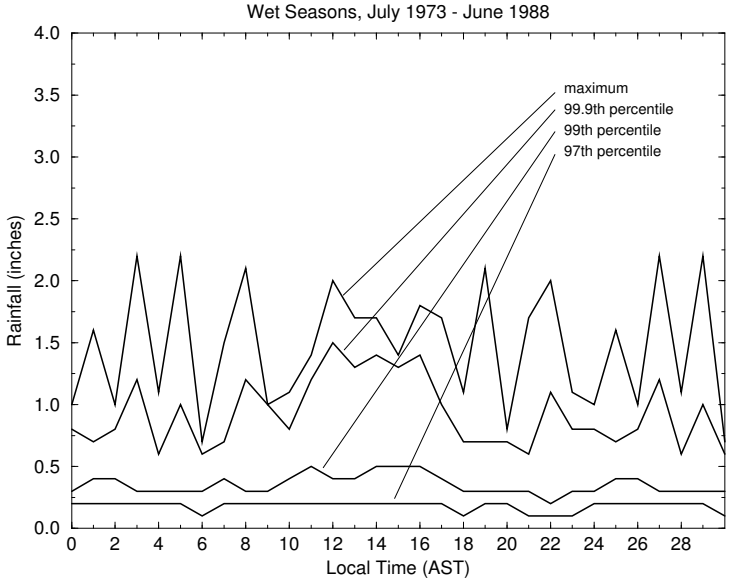
Rainfall Climatology: Ouque



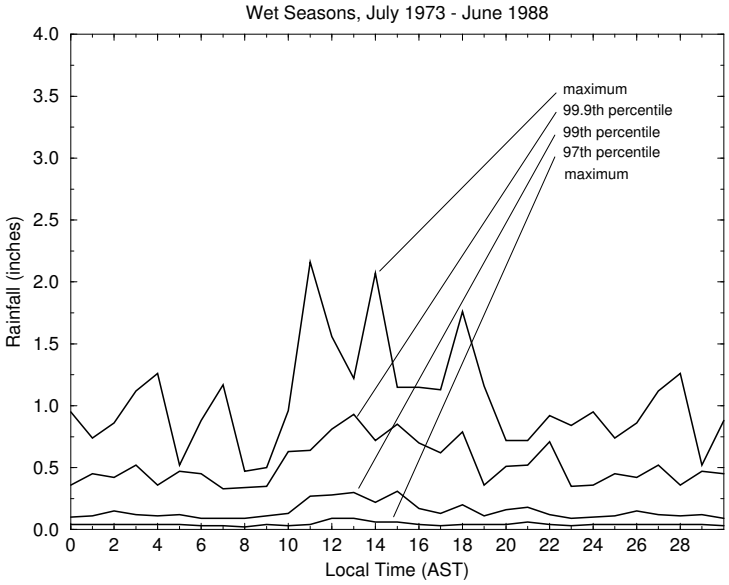
Rainfall Climatology: Pena Pobre-Naguabo



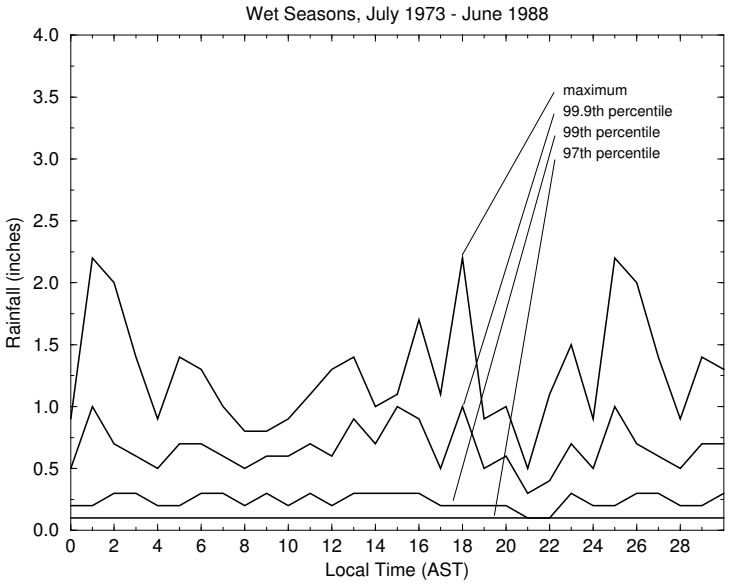
Rainfall Climatology: Pico del Este



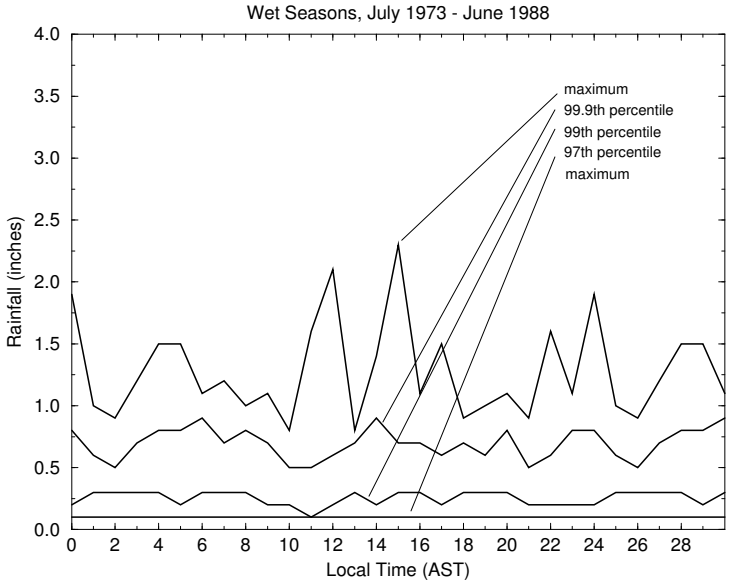
Rainfall Climatology: San Juan WSFO



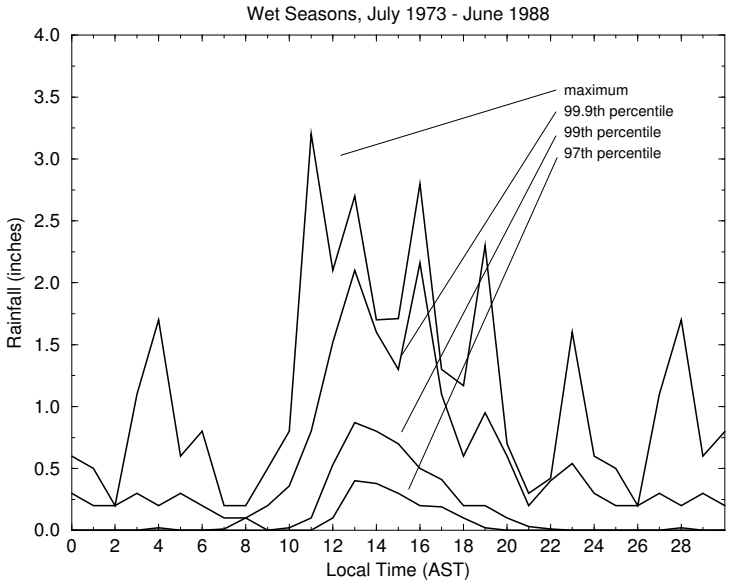
Rainfall Climatology: San Lorenzo



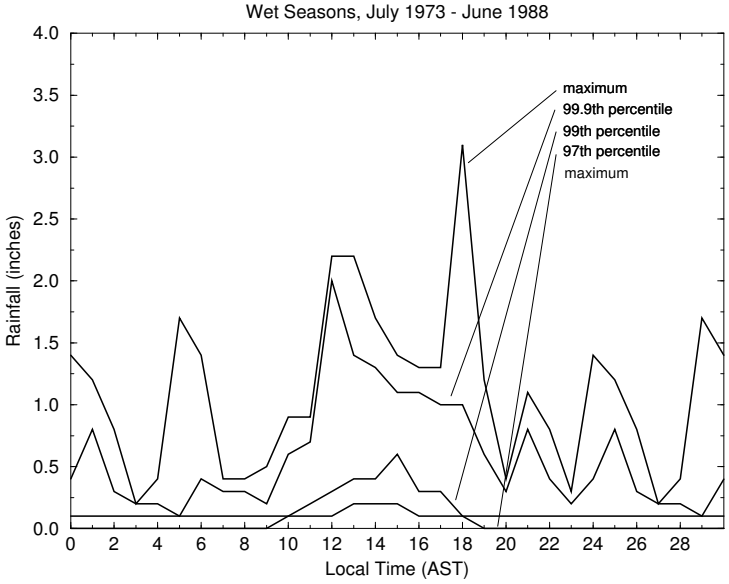
Rainfall Climatology: Yabucoa



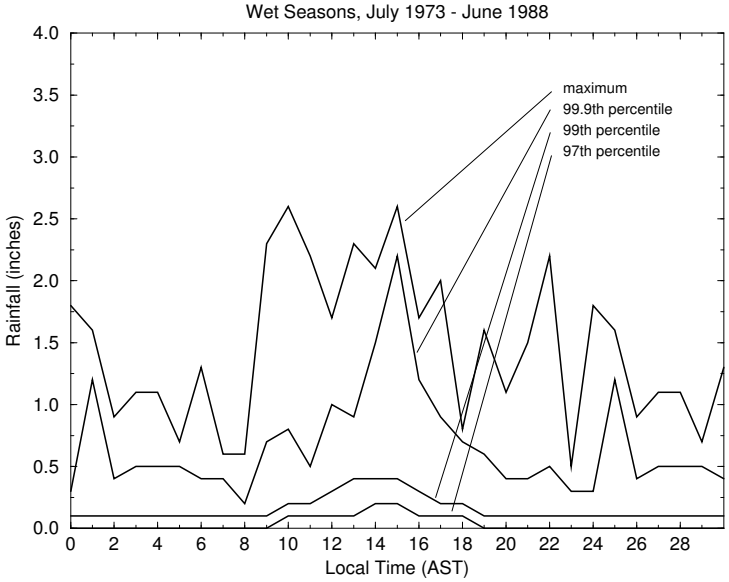
Rainfall Climatology: Benavente-Hormigueros



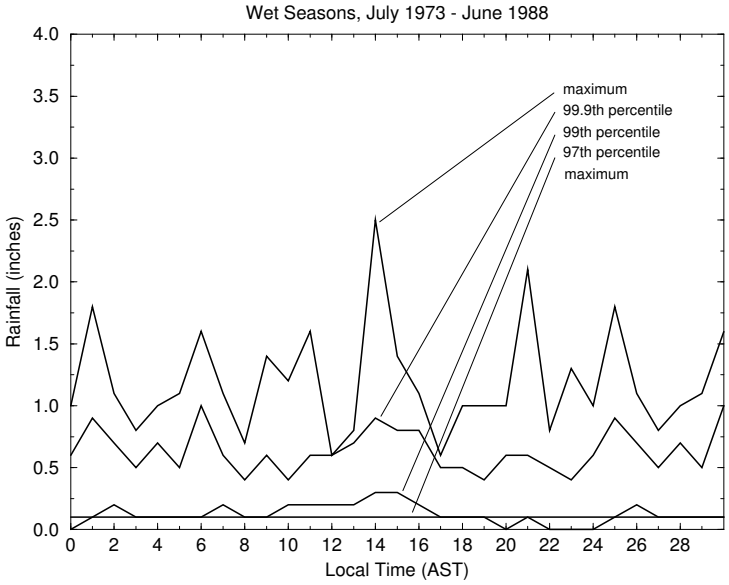
Rainfall Climatology: Botijas1



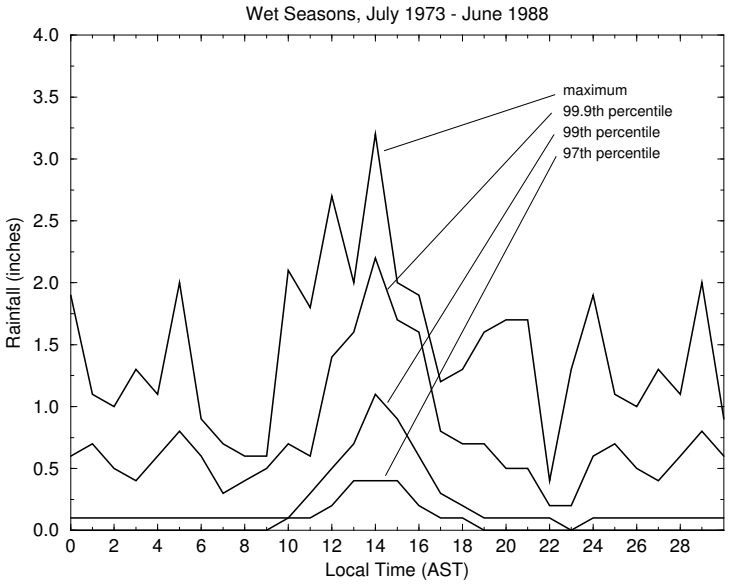
Rainfall Climatology: Botijas2



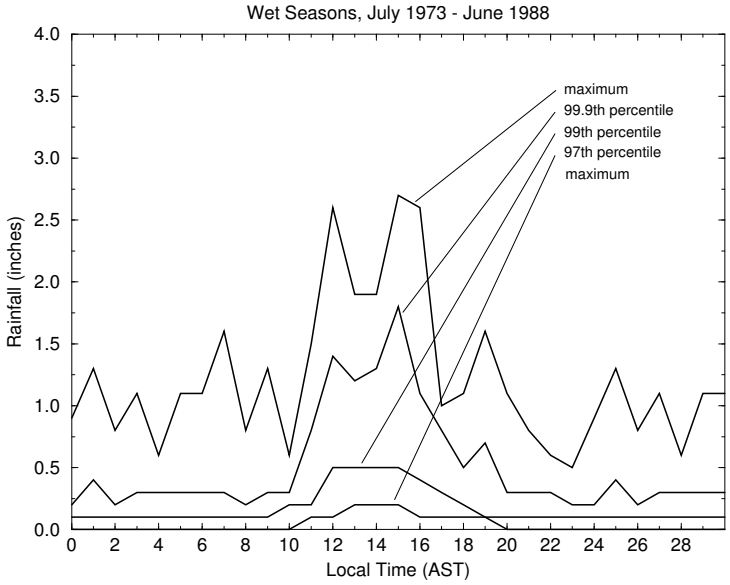
Rainfall Climatology: Cayey



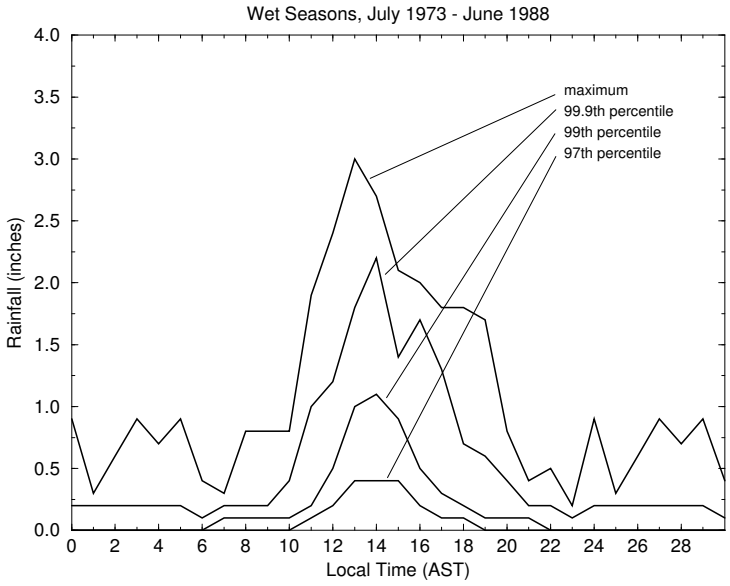
Rainfall Climatology: Cerro Maravilla



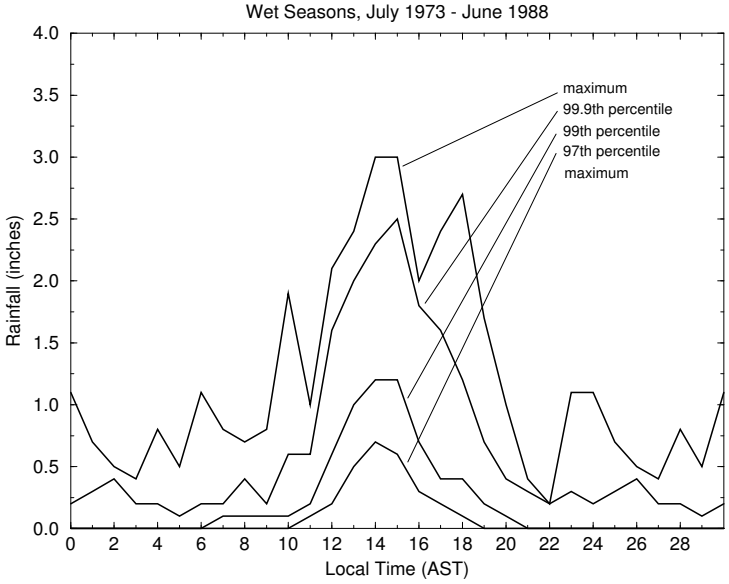
Rainfall Climatology: Corozal Substation



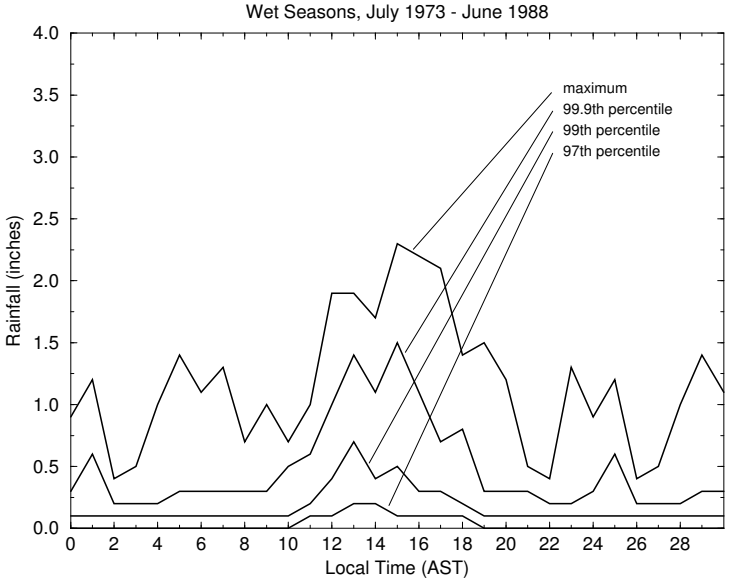
Rainfall Climatology: Dos Bocas



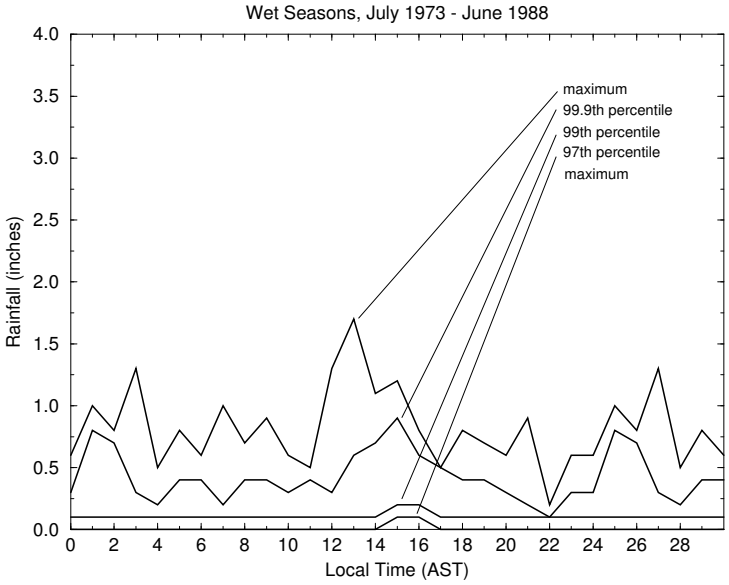
Rainfall Climatology: Maricao



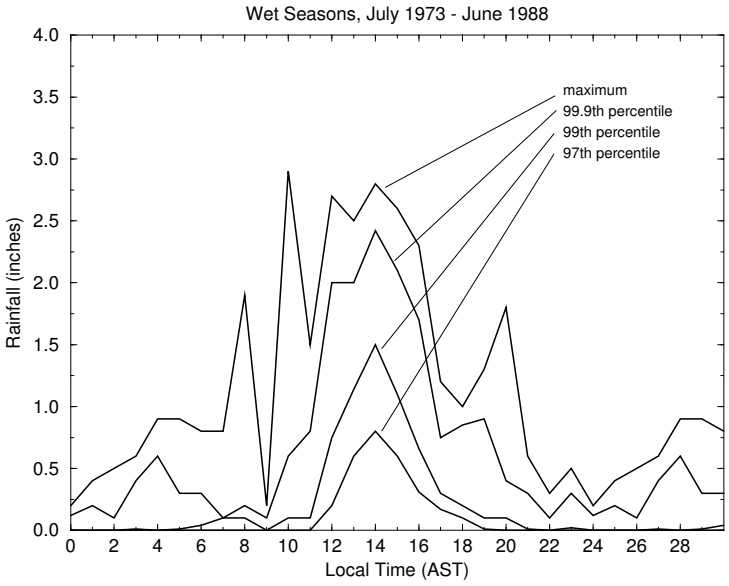
Rainfall Climatology: Negro-Corozal



Rainfall Climatology: Ponce



Rainfall Climatology: San Sebastian

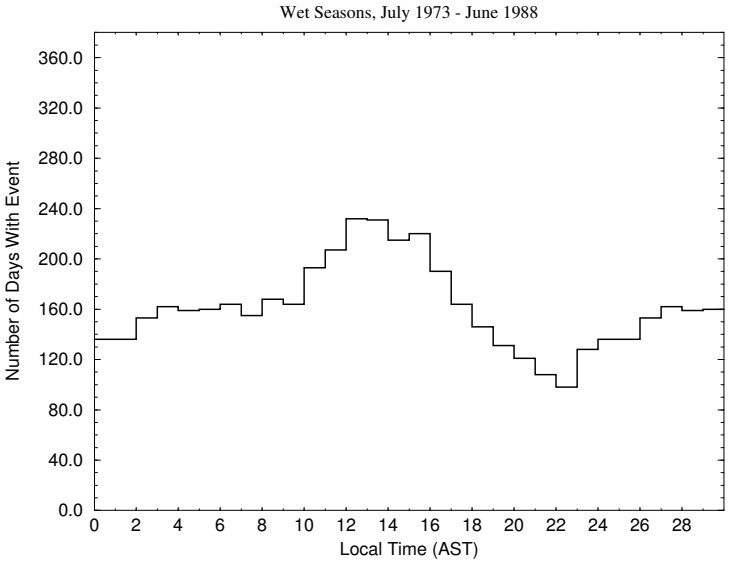


Appendix C

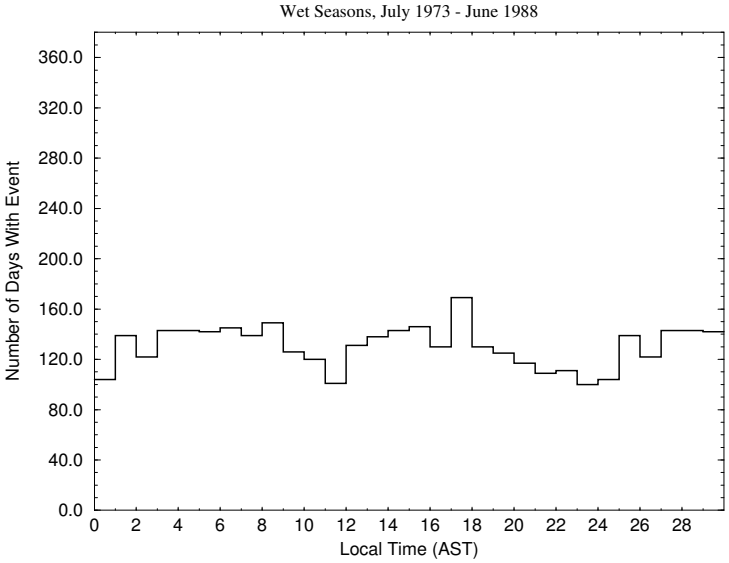
Frequency of Rainfall Events

The frequency of rainfall events for each of the twenty two rainfall recording stations in Puerto Rico is given in this appendix. A rainfall event is any hour in which that station recorded an amount greater than a trace. Local time is shown on the abscissa. Hours 24 through 30 represent "wraparound" times corresponding to midnight through 6 a.m. The number of events are shown on the ordinate.

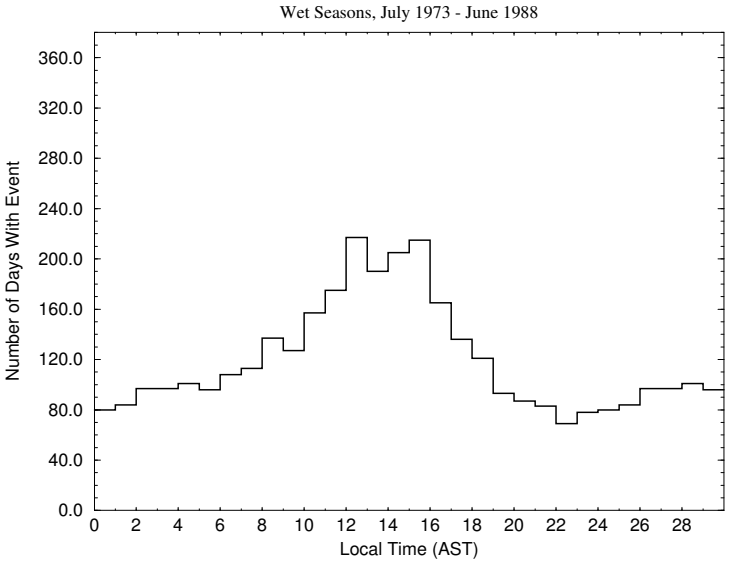
Frequency of Rainfall Events: Cubuy



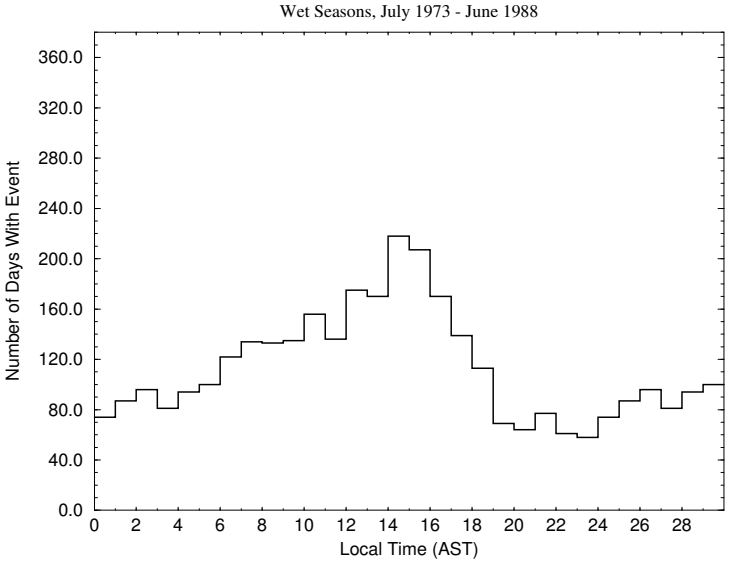
Frequency of Rainfall Events: Fajardo



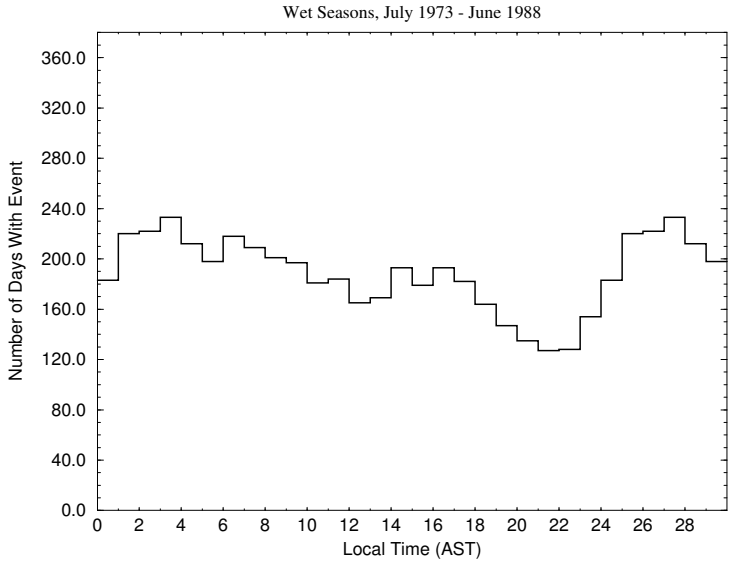
Frequency of Rainfall Events: Gurabo



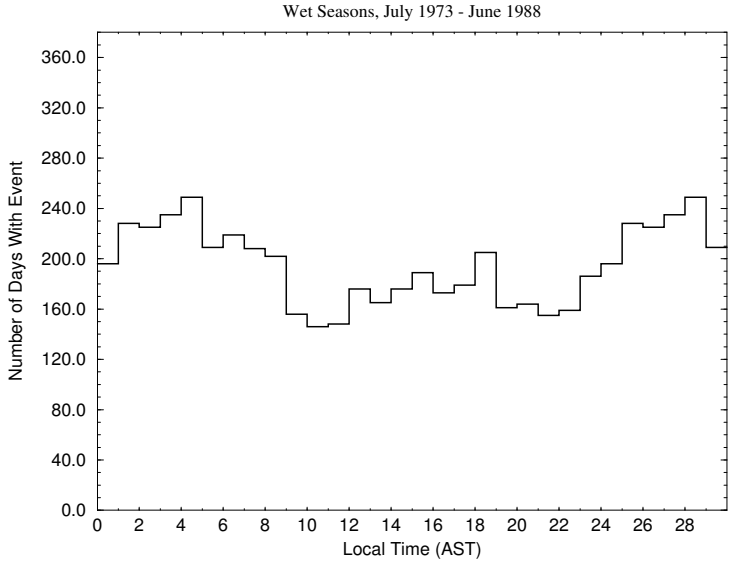
Frequency of Rainfall Events: Gurabo Substation



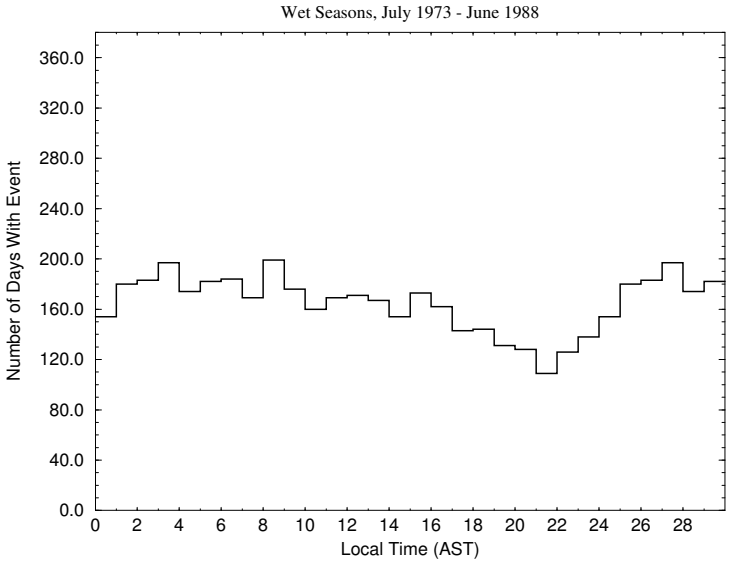
Frequency of Rainfall Events: Las Piedras



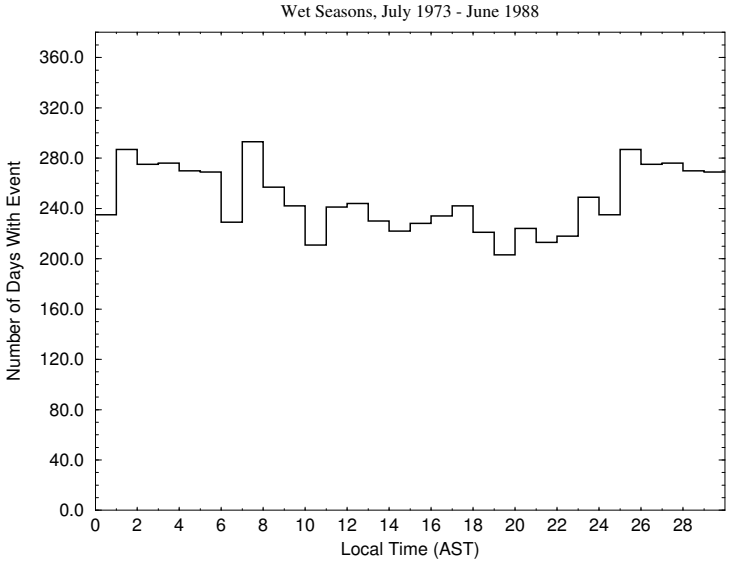
Frequency of Rainfall Events: Ouque



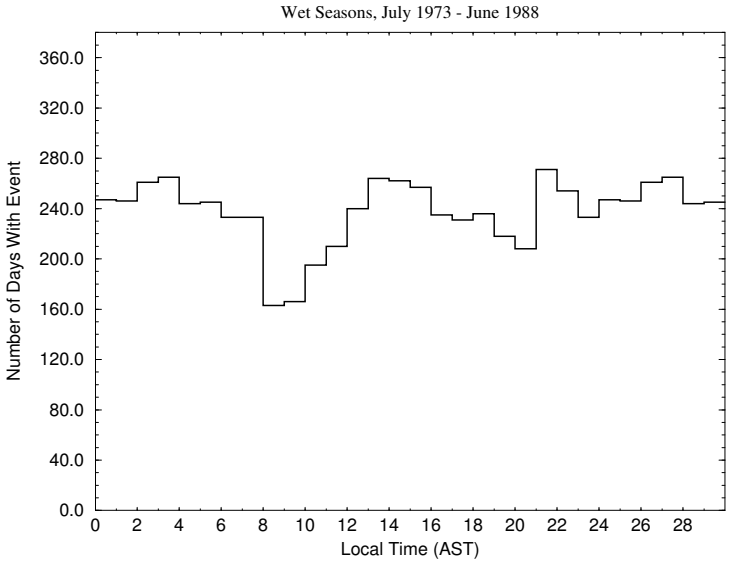
Frequency of Rainfall Events: Pena Pobre-Naguabo



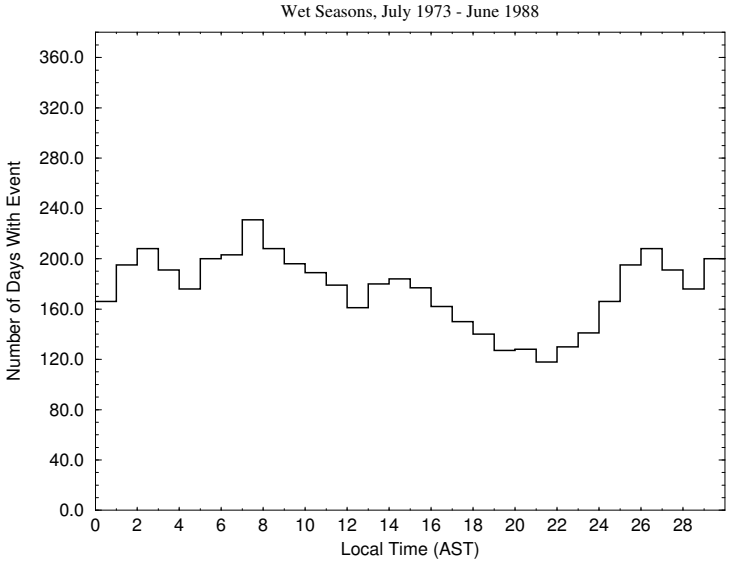
Frequency of Rainfall Events: Pico del Este



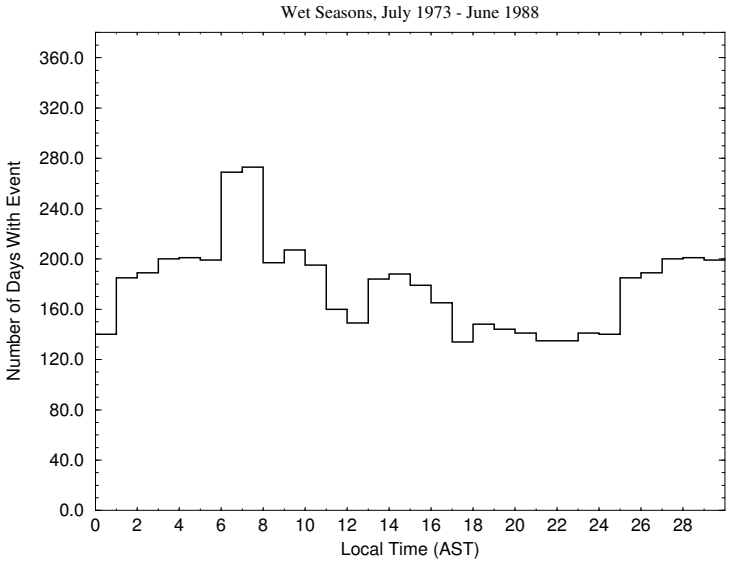
Frequency of Rainfall Events: San Juan WSFO



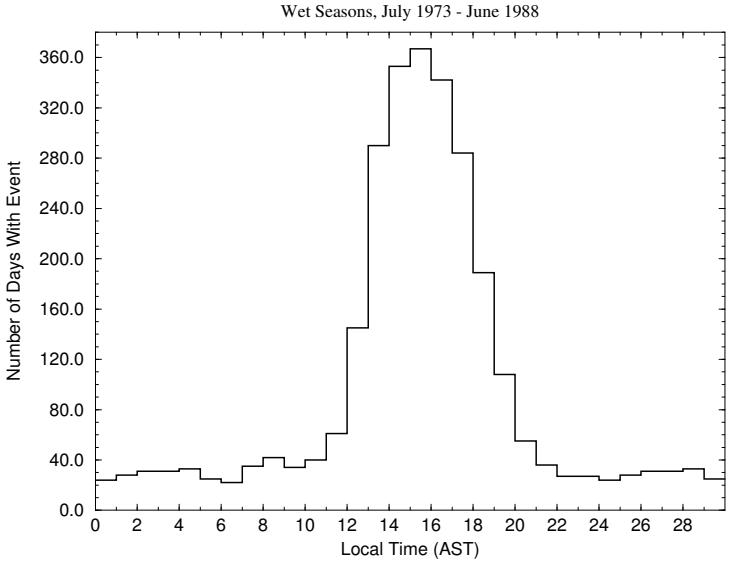
Frequency of Rainfall Events: San Lorenzo



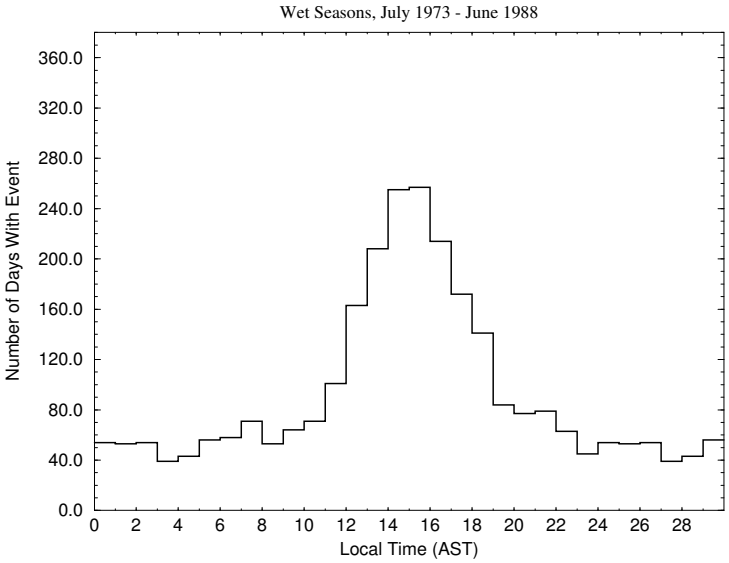
Frequency of Rainfall Events: Yabucoa



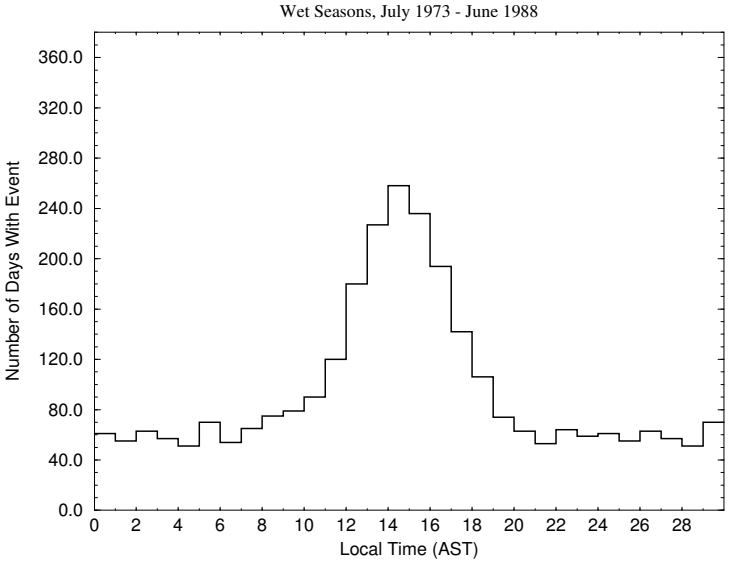
Frequency of Rainfall Events: Benavente-Hormigueros



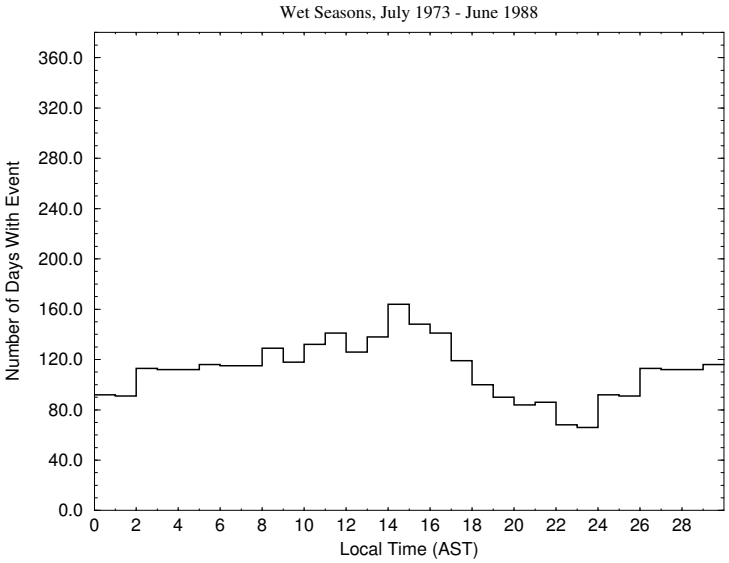
Frequency of Rainfall Events: Botijas1



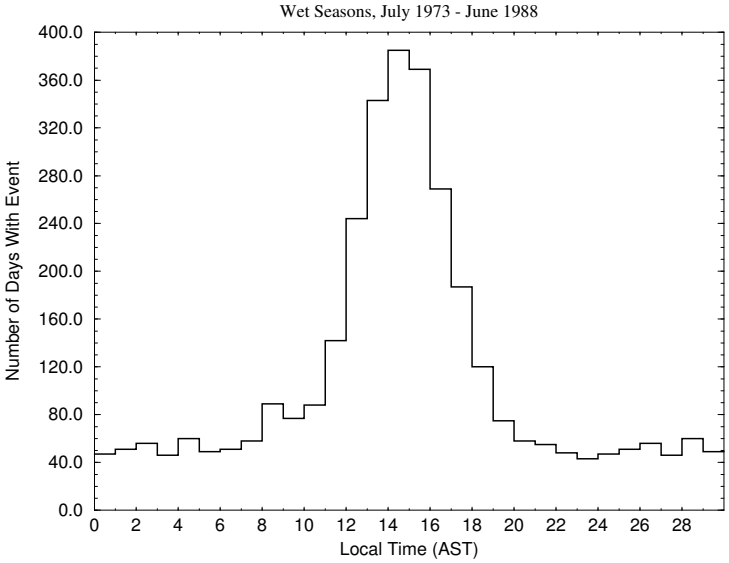
Frequency of Rainfall Events: Botijas2



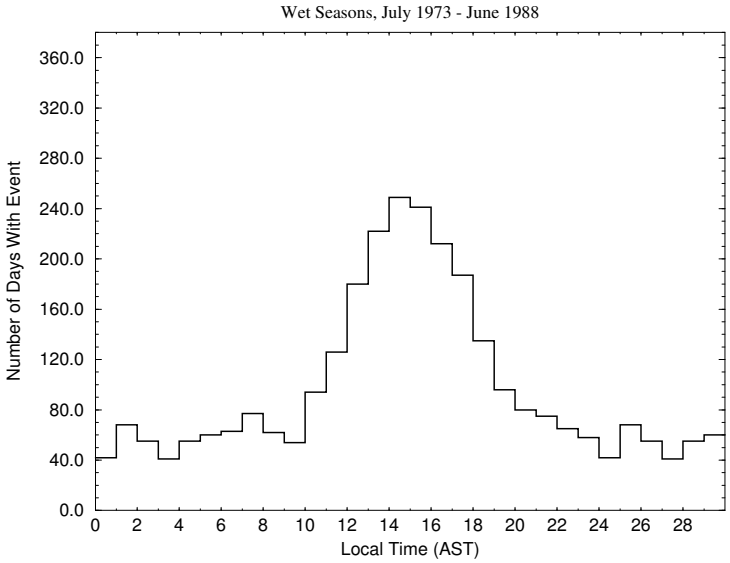
Frequency of Rainfall Events: Cayey



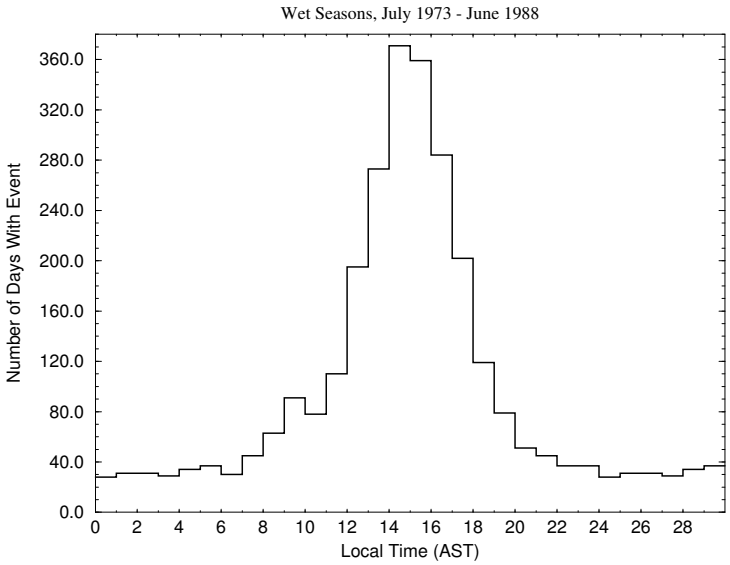
Frequency of Rainfall Events: Cerro Maravilla



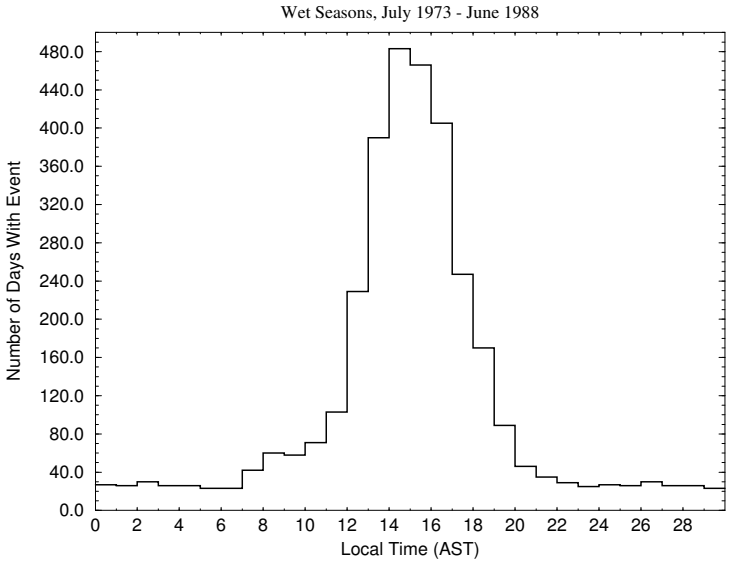
Frequency of Rainfall Events: Corozal Substation



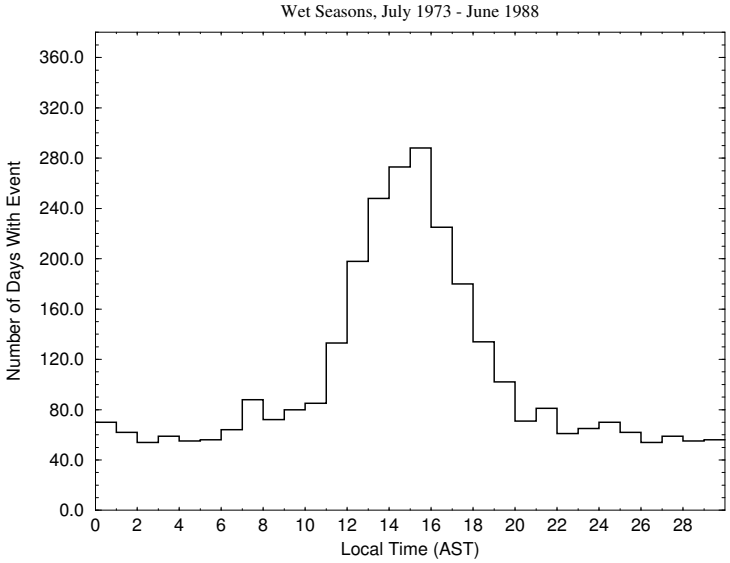
Frequency of Rainfall Events: Dos Bocas



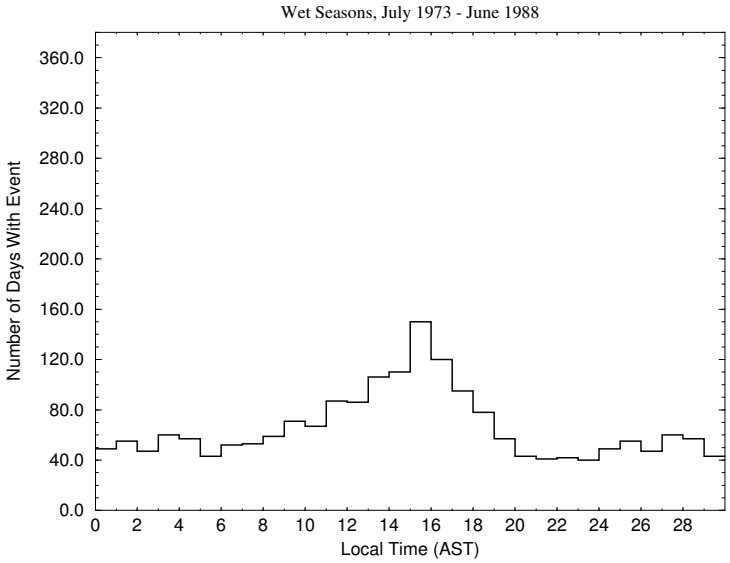
Frequency of Rainfall Events: Maricao



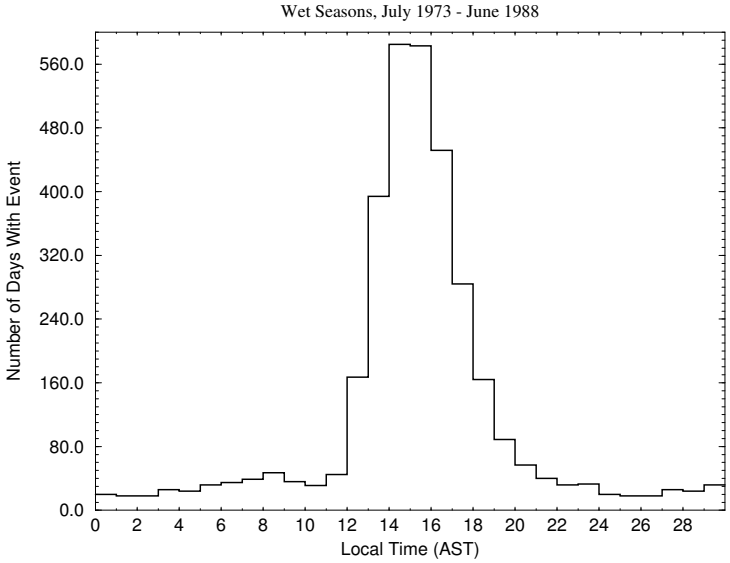
Frequency of Rainfall Events: Negro-Corozal



Frequency of Rainfall Events: Ponce



Frequency of Rainfall Events: San Sebastian



References

Dyer, T. G. J., 1975: The assignment of rainfall stations into homogeneous groups: An application of principal component analysis. *Quart. J. Roy. Meteor. Soc.,* **110**, 1005–1013.

Elsner, J. B., and A. A. Tsonis, 1991: A note on the spatial structure of the covariability of observed Northern Hemisphere surface air temperatures. *PAGEOPH* , **137**, 133–146.

Elsner, J. B., and C.P. Schmertmann, 1994: Assessing forecast skill through cross validation. *Wea. Forecasting* , **9**, 619–624.

Fassig, O. L., 1916: Tropical rains—their duration, frequency, and intensity. *Mon. Wea. Rev.,* **44**, 329–337.

Frank, N., 1970: On the Nature of Upper Tropospheric Cold Core Cylcones Over the Tropical Atlantic, Florida State University Ph.D Thesis, Tallahassee, Florida, 1–242.

Fuelberg, H., and D. Biggar, 1994: The preconvective environment of summer thunderstorms over the Florida panhandle. *Wea. Forecasting*, **9**, 316–326.

Hoffman, M., Ed., 1990: *World Almanac and Book of Facts.* Newspaper Enterprise Association, 960 pp.

Johnson, R., and D. Wichern., 1982: *Applied Multivariate Statistical Analysis*, Prentice Hall, Inc., 594 pp.

Jolliffe, I. T., 1986: *Principal Component Analysis*, Springer-Verlag, 271 pp.

Kelley, W. E. Jr., and D. R. Mock, 1982: A Diagnostic Study of Upper Tropospheric Cold Core Lows Over the Western North Pacific. *Mon. Wea. Rev.,* **110**, 471–480.

Kousky, V. E. , and M. A. Gan, 1981: Upper Tropospheric Cyclonic Voritces in the Tropical South Atlantic. *Tellus,* **33**, 538–551.

Kreyszig, E., 1993: *Advanced Engineering Mathematics*, John Wiley and Sons, Inc., 1294 pp.

Lyons, W. F., and M. Bonell, 1994: Regionalization of daily mesoscale rainfall in the tropical wet/dry climate of the Townsville area of north-east Queensland during the 1988–1989 wet season. *Int. J. of Climat.*, **14**, 135–163.

Overland, J. E., and R. W. Preisendorfer, 1982: A significance test for principal components applied to a cyclone climatology. *Mon. Wea. Rev.,* **110**, 1–4.

Pico, R., 1974: *The Geography of Puerto Rico*, Aldine Publishing Company, 439 pp.

Ray, C. L., 1928: Diurnal variation of rain at San Juan, P.R. *Mon. Wea. Rev.,* **56**, 140–141.

Ray, P., *Mesoscale Meteorology and Forecasting*, 1986: American Meteorology Society, 793 pp.

Richman, M. B., 1986: Review Article: Rotation of Principal Components. *J. Clim.,* **6**, 293–335.

Riehl, H., 1954: *Tropical Meteorology*, McGraw-Hill Book Company, Inc., 392 pp.

Ruffner, J., and F. Bair, Ed., 1978: *Climates of the States*, Gale Research Company, Detroit, 1158 pp.

Sharon, D., 1974: On the modelling of correlation functions for rainfall studies. *J. Hydrol.*, **22**, 219–224.

Simpson, J., R. F. Adler, and G. R. North, 1988: A proposed tropical rainfall measuring mission (TRMM) satellite. *Bull. Amer. Meteor. Soc.,* **69**, 278–295.

Stol, P. T., 1972: The relative efficiency of the density of rain-guage networks. *J. Hydrol.*, **15**, 193–208.

White, D., M. Richman, and B. Yarnal, 1991: Climate regionalization and rotation of principal components, *Int. J. of Climat.*, **11**, 1–25.

Biographical Sketch

Matthew M. Carter was born on 5 November 1970 in Lincoln, Nebraska. He graduated from John Adams High School, South Bend, Indiana, in June 1988. In June 1992, he graduated from Northwestern University in Evanston, Illinois with a B.A. in history and political science.