Chapter 3

Statistical Analysis of Rainfall Data

In this chapter we describe the preliminary data analysis based the rainfall measurements collected from the 14 stations over the 19-year period. As described in Chapter 2, rainfall measurements are accumulated into 73 five-day periods for each year, with an adjustment made for the extra day (February 29) occurring during leap years. In the first section of this chapter we show the rainfall measurements on leap year days, and we also investigate the pattern of missing data. The frequency distribution of the five-day totals is investigated in Section 3.2, with the aim of using a data transformation to reduce skewness. After further study of the most appropriate transformation, in Section 3.3 we compare the seasonal patterns and annual trends in the rainfall at each station.

3.1 Leap Year Days and Pattern of Missing Data

As described in chapter 2, we grouped the rainfall measurements into 5-day totals. This facilitated the statistical analysis by

(a) reducing the correlation between data on successive days,

(b) reducing the proportion of periods on which no rain fell, and

(c) reducing the number of measurements to be analysed in each year.

However, the comparison over several years is complicated by the fact that February 29 exists only in every fourth year. The method proposed for handling this anomaly is to adjust the total from February 25 to March 1 in a leap year by multiplying by 5/6 of the total. It is thus of interest to examine the pattern of rainfall at the 14 stations on February 29. Table 3.1 shows these measurements.

Since the end of February is in the middle of the dry season in Pattani Province, little rainfall is expected on February 29. The table shows that only small amounts of rainfall were recorded on February 29 in 1984, and no rainfall at all was recorded on this day in 1988, 1992, or 1996. However, in 2000 most stations recorded a substantial amount of rainfall on February 29. This shows that the February 29 measurements cannot be ignored, and that some kind of adjustment is desirable.

Station	*1984*	*1988*	*1992*	*1996*	*2000*
Pattani	0.05	0	0	0	19.8
Sai Buri	1.2	0	0	0	83.5
Yarang	4.2	0	0	0	1.5
Khok Pho	0	0	0	0	0
Panare	0.05	0	0	0	23.9
Mayo	0	0	0	0	17.6
Nong Chik	3.5	0	0	0	0
Yaring	0.05	0	0	0	50.3
Kapho	10.0	0	0	0	38.2
Mai Kaen	no record	0	0	0	68.2
Khok Pho (SHS)	21.1	0	0	0	0
Thung Yang Dang	no record	no record	0	0	16.1
Mae Lan	no record	no record	no record	0	0.05
Pattani Airport	0	0	0	0	0.7

Table 3.1: Rainfall Measurements on February 29

As described in Chapter 2, the same method was used to handle missing measurements that occurred within any 5-day period. This method involved computing the total for each 5-day period by multiplying the average rainfall during the period by 5. Thus, unless all five measurements were missing, an estimate could be obtained for each period.

Table 3.2 shows the percentage of periods for which data were missing at each station for each year, where the empty cells denote complete data and the dark shadings denote that 100% of the data in the year were missing. Note that rainfall measurements were not recorded in Station 9 (Kapho) before 1983, or before 1984 in Station 10 (Mai Kaen), or before 1989 in Station 12 (Thung Yang Dang), or before 1996 in Station 13 (Mae Lan). Thus, if these years are omitted, the percentages of missing data are quite small.

For each station, the percentages of 5-day periods for which all data are missing are shown in Table 3.3. For six of the stations there were no missing data. For the other eight stations some data were missing, with the percentages of periods having missing data ranging from 0.07% at Yarang to 6.49% at Pattani. The overall percent of periods for which the data were missing is 1.71%.

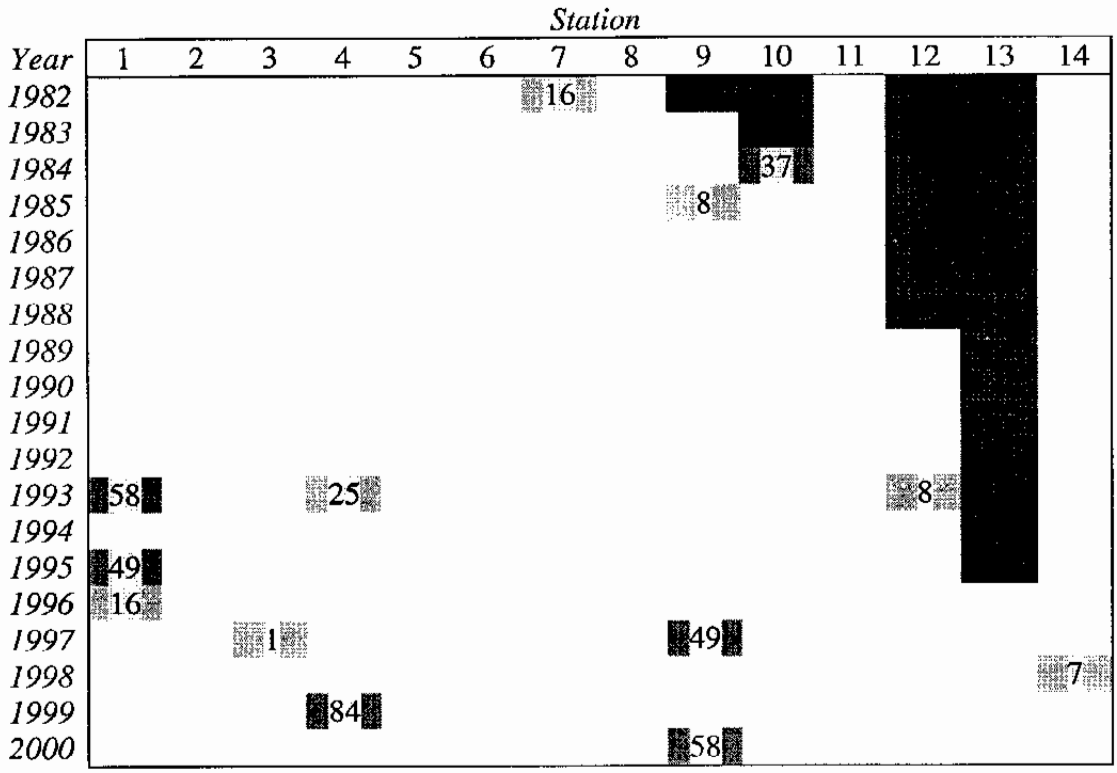

Table 3.2: Pattern of Missing Data by Station and Year

Station	*Percent Missing*
Pattani	6.49
Sai Buri	0
Yarang	0.07
Khok Pho	5.77
Panare	0
Mayo	0
Nong Chik	0.87
Yaring	0
Kapho	6.39
Mai Kaen	1.93
Khok Pho (SHS)	0
Thung Yang Dang	0.68
Mae Lan	0
Pattani Airport	0.36
Overall	**1.71**

Table 3.3: Overall Percentages of 5-day Periods with Missing Data at Each Station

3.2 Distribution of Five-Day Totals

Figure 3.1 shows the frequency distribution of 5-day total rainfall measurements over all 14 stations and 19 years. The distribution is clearly very highly skewed to the right.

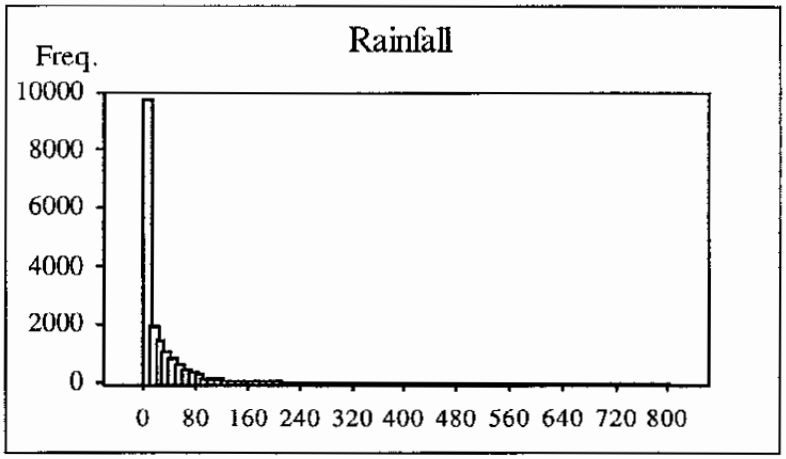

Figure 3.1: Distribution of Rainfall in 5-day Periods

The coefficients of skewness and kurtosis were 4.91 and 35.25, respectively. The mean was 24.97 mm and the standard deviation was 51.25 mm. The highest recorded 5-day total was 759.1 mm, at Station 10 (Mai Kaen) in period 68 (December 2-6) in 1987, and also at Station 11 (Khok Pho Self-Help Settlement) in the same period. (This could be due to an error in the data recording process.) Of the 17364 5-day periods when measurements were recorded, no rain at all was recorded on 5939 periods (34.2%), and less than 10 mm was recorded on 9765 (56.2%) periods.

Figure 3.2 shows a histogram of the rainfall after taking square roots of the 5-day totals. The distribution is still skewed, but less so. The skewness coefficient is 1.45 and the kurtosis is 2.87. The high frequency in the first bin corresponds to measurements between 0 and 0.5 on the square root scale, that is, between 0 and 0.25 mm. There were 7076 5-day periods on which the total rainfall was less than 0.25 mm, corresponding to 40.8% of the time.

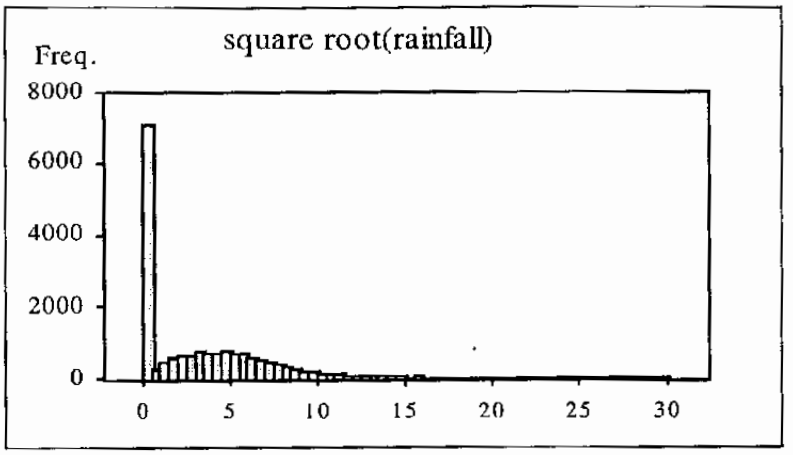

Figure 3.2: Distribution of Square Root of Rainfall in 5-day Periods

Figure 3.3 shows a histogram of the rainfall after taking base-10 logarithms of the 5-day totals. Since the logarithm of 0 cannot be taken, the transformation used is $\log(x+1)$ rather than $\log(x)$. This transformation has the advantage that zeros remain zeros after the transformation.

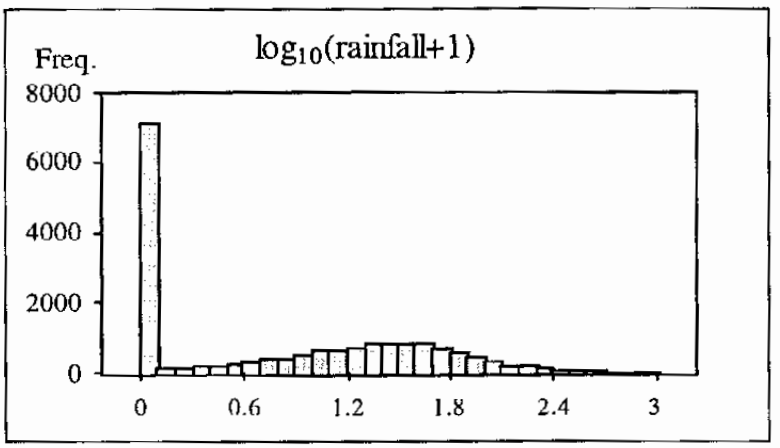

Figure 3.3: Distribution of Log(Rainfall+1) in 5-day Periods

This log transformation reduces the skewness coefficient further, to 0.29, and the kurtosis is negative (–1.34). The high frequency in the first bin corresponds to measurements between 0 and 0.1 on the transformed data scale, that is, between 0 and $10^{0.1}-1 = 0.26$ mm. There were 7101 5-day periods on which the total rainfall was less than 0.26 mm, corresponding to 40.8% of the time.

Note that there were 5939 5-days periods when no rainfall at all was recorded and no transformation can remove this spike in the histogram. Figure 3.4 shows a histogram of the log-transformed 5-day totals after omitting the periods on which no rain fell.

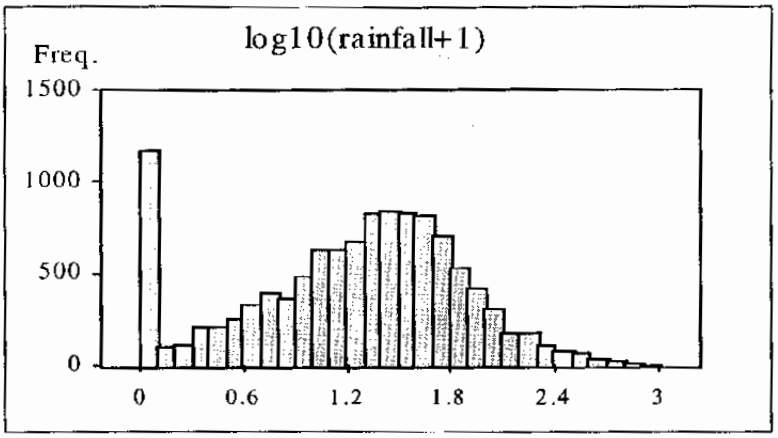

Figure 3.4: Distribution of Log(Rainfall+1), Omitting 5-day Periods of no Rain

3.3 Overall Seasonal and Regional Patterns

Table 3.4 shows the averages of the transformed 5-day rainfall totals at each of the 14 stations. This table also includes the averages for the raw (untransformed) totals, as well as the result of reversing the transformation after taking the averages of the transformed data. Note that this is similar to taking geometric means. It would be equivalent to taking geometric means if the constant 1 were not added to the data before taking logarithms.

	Averages (x = raw data)		
Station	ave[x]	y=ave[$log(1+x)$]	10^y-1
Pattani	19.31	0.685	3.84
Sai Buri	30.73	0.958	8.08
Yarang	23.08	0.779	5.01
Khok Pho	18.91	0.583	2.83
Panare	24.12	0.783	5.07
Mayo	28.37	0.862	6.28
Nong Chik	23.60	0.761	4.76
Yaring	22.69	0.690	3.90
Kapho	25.89	0.766	4.84
Mai Kaen	32.96	1.014	9.34
Khok Pho (SHS)	23.71	0.827	5.72
Thung Yang Dang	23.21	0.780	5.03
Mae Lan	35.46	0.978	8.52
Pattani Airport	24.97	0.922	7.37
Overall	**25.05**	**0.8064**	**5.40**

Table 3.4: Averages of Transformed Rainfall Totals in 5-day Periods at Each Station

Table 3.5 shows numerical summaries of the transformed 5-day rainfall totals for each period.

The rightmost column of the table gives the back-transformed means, obtained by applying the transformation $x = 10^y - 1$ to the means. The period with the highest rainfall was the 65th in the year, corresponding to November 17-21. The lowest rainfall was for period 10 (February 15-19).

Figure 3.5 shows a graph of this seasonal pattern. It can be seen clearly that the periods with high rainfall were around 60-70 in the year.

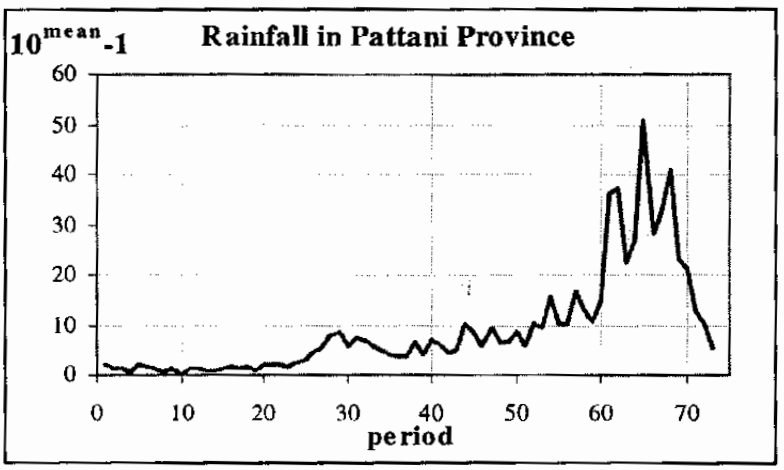

Figure 3.5: Plot of Back-transformed Rainfall Totals in 5-day Periods

period	Size	Mean	StDev	Skew	Kurt	Min	Med	Max	10^{mean}-1
1	241	0.458	0.608	0.963	-0.512	0.000	0.021	2.176	1.869
2	241	0.353	0.542	1.330	0.594	0.000	0.000	2.247	1.256
3	241	0.344	0.483	1.146	0.002	0.000	0.021	1.750	1.210
4	241	0.165	0.364	2.388	4.860	0.000	0.000	1.780	0.464
5	241	0.453	0.590	0.902	-0.611	0.000	0.021	2.321	1.838
6	241	0.427	0.589	1.055	-0.251	0.000	0.000	2.340	1.674
7	241	0.345	0.568	1.351	0.286	0.000	0.000	1.936	1.214
8	241	0.184	0.424	2.368	4.755	0.000	0.000	2.181	0.529
9	241	0.311	0.670	2.281	4.226	0.000	0.000	2.814	1.048
10	241	0.060	0.215	4.219	18.741	0.000	0.000	1.540	0.149
11	241	0.306	0.569	1.742	1.745	0.000	0.000	2.061	1.022
12	241	0.335	0.572	1.559	1.317	0.000	0.000	2.418	1.161
13	240	0.217	0.465	2.033	2.763	0.000	0.000	1.847	0.647
14	240	0.229	0.488	2.102	3.200	0.000	0.000	2.080	0.695
15	240	0.366	0.686	1.564	0.780	0.000	0.000	2.166	1.323
16	240	0.417	0.681	1.302	0.102	0.000	0.000	2.314	1.613
17	240	0.381	0.620	1.358	0.413	0.000	0.000	2.223	1.405
18	240	0.428	0.664	1.177	-0.230	0.000	0.000	2.127	1.681
19	240	0.210	0.477	2.266	3.889	0.000	0.000	2.107	0.621
20	240	0.492	0.717	1.000	-0.658	0.000	0.000	2.378	2.103
21	240	0.482	0.734	1.125	-0.402	0.000	0.000	2.288	2.035
22	240	0.507	0.650	0.801	-0.976	0.000	0.021	2.081	2.213
23	240	0.413	0.606	1.100	-0.355	0.000	0.000	1.927	1.588
24	240	0.539	0.685	0.746	-1.078	0.000	0.021	2.031	2.458
25	241	0.585	0.684	0.522	-1.430	0.000	0.021	1.977	2.848
26	241	0.725	0.694	0.275	-1.437	0.000	0.672	2.216	4.309
27	241	0.807	0.741	0.162	-1.530	0.000	0.886	2.235	5.410
28	241	0.937	0.682	-0.249	-1.390	0.000	1.121	2.166	7.652
29	241	0.973	0.702	-0.272	-1.343	0.000	1.199	2.267	8.397
30	241	0.820	0.734	0.064	-1.660	0.000	0.942	2.069	5.600
31	241	0.912	0.713	-0.075	-1.421	0.000	1.068	2.307	7.158
32	239	0.901	0.710	-0.104	-1.525	0.000	1.057	2.196	6.957
33	239	0.816	0.691	0.037	-1.484	0.000	0.908	2.268	5.544
34	239	0.772	0.654	-0.013	-1.615	0.000	0.906	1.894	4.921
35	239	0.707	0.678	0.199	-1.566	0.000	0.748	2.052	4.099
36	239	0.684	0.638	0.216	-1.398	0.000	0.724	2.141	3.828
37	239	0.678	0.709	0.432	-1.321	0.000	0.519	2.390	3.769
38	237	0.869	0.681	-0.085	-1.374	0.000	1.045	2.282	6.389
39	237	0.713	0.662	0.193	-1.456	0.000	0.785	2.239	4.169
40	237	0.900	0.699	-0.162	-1.529	0.000	1.111	2.268	6.943
41	237	0.845	0.683	-0.096	-1.520	0.000	0.987	2.027	5.999
42	234	0.753	0.691	0.071	-1.645	0.000	0.916	1.954	4.661
43	237	0.770	0.713	0.191	-1.488	0.000	0.813	2.216	4.889
44	237	1.048	0.638	-0.541	-0.971	0.000	1.204	2.135	10.172
45	237	0.977	0.691	-0.365	-1.380	0.000	1.210	2.123	8.485
46	237	0.825	0.651	-0.092	-1.484	0.000	0.978	2.057	5.680
47	237	1.015	0.740	-0.259	-1.437	0.000	1.175	2.163	9.361
48	237	0.867	0.668	-0.144	-1.460	0.000	1.025	2.109	6.355
49	237	0.883	0.719	0.012	-1.456	0.000	1.000	2.422	6.630
50	236	0.973	0.634	-0.418	-1.129	0.000	1.124	2.112	8.395
51	236	0.822	0.670	-0.074	-1.495	0.000	0.980	2.236	5.634
52	236	1.046	0.663	-0.564	-1.126	0.000	1.249	2.122	10.115
53	236	1.021	0.682	-0.368	-1.189	0.000	1.221	2.247	9.493
54	236	1.216	0.631	-0.813	-0.535	0.000	1.417	2.162	15.445
55	237	1.044	0.668	-0.486	-1.209	0.000	1.260	2.138	10.076
56	237	1.051	0.684	-0.324	-1.080	0.000	1.173	2.373	10.247
57	237	1.251	0.570	-0.896	0.057	0.000	1.360	2.286	16.810
58	237	1.152	0.634	-0.729	-0.674	0.000	1.328	2.279	13.184
59	237	1.071	0.685	-0.375	-1.096	0.000	1.258	2.324	10.765
60	237	1.197	0.607	-0.779	-0.448	0.000	1.356	2.083	14.738
61	237	1.572	0.571	-1.336	1.568	0.000	1.713	2.445	36.321
62	233	1.582	0.543	-1.211	1.727	0.000	1.660	2.465	37.172
63	233	1.367	0.634	-1.016	0.140	0.000	1.505	2.445	22.302
64	233	1.444	0.671	-0.752	0.014	0.000	1.569	2.734	26.786
65	233	1.712	0.717	-0.762	0.098	0.000	1.736	2.769	50.466
66	233	1.467	0.708	-0.855	-0.075	0.000	1.638	2.618	28.328
67	234	1.532	0.707	-0.741	-0.079	0.000	1.649	2.825	33.014
68	232	1.621	0.802	-0.798	-0.355	0.000	1.807	2.881	40.777
69	232	1.385	0.877	-0.556	-1.127	0.000	1.676	2.801	23.291
70	232	1.348	0.889	-0.507	-1.275	0.000	1.611	2.572	21.263
71	232	1.135	0.843	0.004	-1.281	0.000	1.209	2.673	12.645
72	232	1.053	0.754	-0.075	-1.193	0.000	1.152	2.530	10.295
73	232	0.791	0.829	0.587	-1.018	0.000	0.556	2.722	5.179

Table 3.5: Numerical Summaries of Transformed Rainfall Totals in 5-day Periods

3.4 One-way Analysis of Variance Results for each Station

In this section we compare the rainfall at each station over the different periods of the year. We can do this using the one-way analysis of variance model to compare the rainfall over the 73 5-day periods.

The one-way analysis of variance model makes two assumptions. The first assumption is that the population standard deviation is the same for each group. The second assumption is that the errors after fitting the model are normally distributed.

To assess the first assumption, we fitted the model to the complete set of 17364 measurements, and computed the standard deviation of the residuals for each of the 73 5-day periods. Figure 3.6 shows a plot of these standard deviations versus the means.

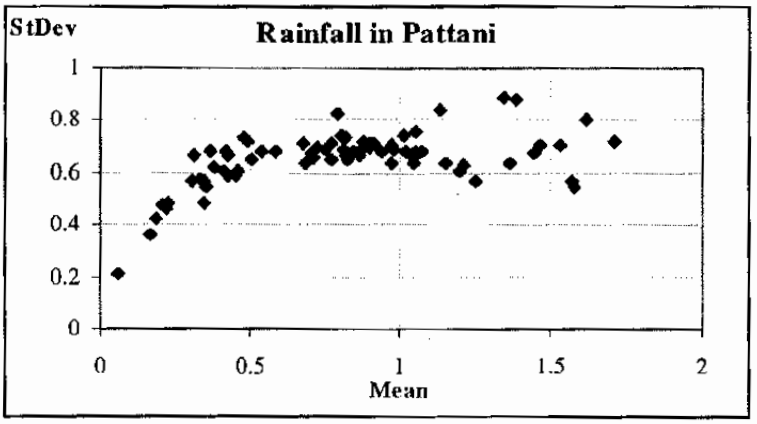

Figure 3.6: Scatter Plot of Standard Deviations vs Means after Fitting One-way Anova

Figure 3.6 shows that the standard deviations are approximately constant, except for small values of the mean.

To assess the normality assumption, Figure 3.7 shows a normal scores plot of the residuals after fitting the one-way anova model.

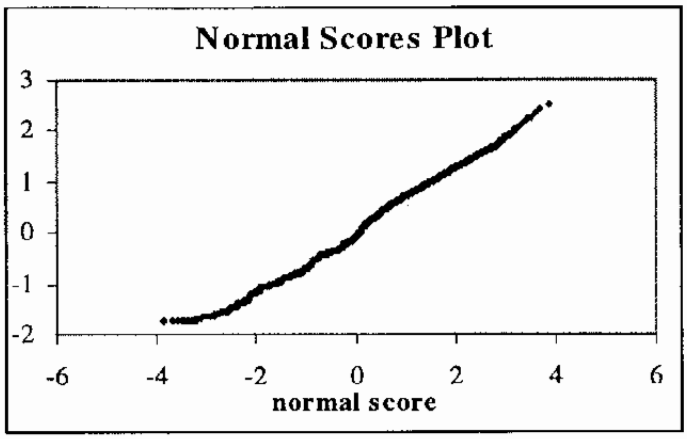

Figure 3.7: Normal Scores Plot of Residuals after Fitting One-way Anova

Figure 3.8 shows the results from fitting the one-way anova model to the transformed 5-day rainfall measurements for the Pattani (left panel) and Yaring (right panel) stations.

The tables at the top show the relevant results. These include the conventional statistics associated with the one-way anova model, namely (a) the residual sum of squares (*Resid SS*), (b) the r-squared goodness-of-fit statistic (*r-sq*), (c) the number of missing observations (if any), (d) the numbers of degrees of freedom in the numerator and the denominator (*df*), (e) the F-statistic (*F*), (f) its p-value (*p-val*), and (g) the standard deviation of the residuals (*s*). In addition, each table shows three statistics that are not always given in conventional one-way anova tables, namely, (h) the root-mean-squared difference between the group means (*rms*), (i) the half-width of a test-based confidence interval for the overall difference between the means (*CI/2*), and (j) the p-value for the Levene's test that each group has the same standard deviation (*Lev p*).

The figure also shows graphs of the means and their 95% confidence intervals corresponding to each of the 73 5-day periods, together with normal scores plots of the residuals.

log(1+rainfall): Pattani

Oneway Anova: Resid SS: 498.50 r-sq: 0.31 90 missing

factor	*df*	*F*	*p-val*	*s*	*rms*	*CI/2*	*Lev p*
Period	72/1224	7.657	0.0000	0.638	3.572	1.472	0.000

log(1+rainfall): Yaring

Oneway Anova: Resid SS: 579.58 r-sq: 0.32

factor	*df*	*F*	*p-val*	*s*	*rms*	*CI/2*	*Lev p*
Period	72/1314	8.639	0.0000	0.664	3.826	1.484	0.000

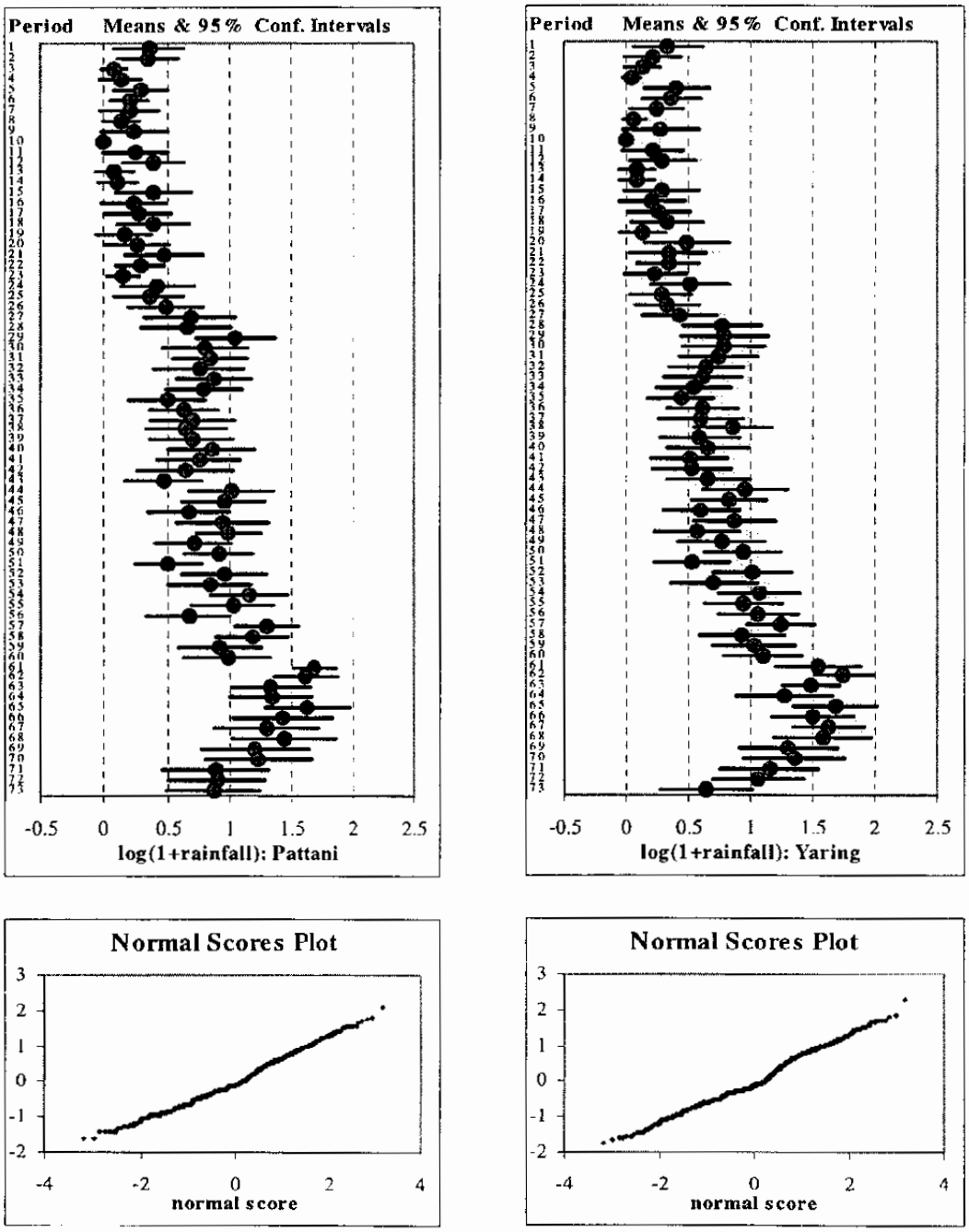

Figure 3.8: One-way Anova Results for Pattani & Yaring Rainfall

The residual sum of squares was 498.5 for Pattani and 579.6 for Yaring, corresponding to r-squared statistics of 0.31 and 0.32, respectively. As indicated earlier in Table 3.2, there were 90 periods of missing data at Pattani, but none at all at Yaring. The number of degrees of freedom in the numerator is always 72, one less than the number of periods. If no data were missing, the number of degrees of freedom in the denominator would be 1314 (the total number of periods, $73 \times 19 = 1387$, minus 73).

In each case the F-test gives a very small p-value (less than 0.00005), indicating that there is a seasonal pattern in the rainfall measurements. Levene's test also gives a very small p-value, indicating that there are also differences in the standard deviations of the rainfall at different periods during the year.

The seasonal patterns appear in the graphs of the means and their confidence intervals. These show that at both Pattani and Yaring the rainfall shows an increasing trend during the year, reaching a maximum after period 60 (around the middle of October), and then declining over the last six periods (during December). The patterns for Pattani and Yaring are quite similar.

Figures 3.9-3.14 show the corresponding results for the other 12 stations. The normal scores plot show a linear trend, suggesting that the normality assumption is reasonable for data from 11 stations except those from Khok Pho. The patterns for the other 12 stations are similar to those for Pattani and Yaring.
However, the pattern for Mae Lan has a wide confidence interval for every period.

log(1+rainfall): Khok Pho(SHS)

Oneway Anova: Resid SS: 577.15 r-sq: 0.28

factor	df	F	p-val	s	rms	CI/2	Lev p
Period	72/1314	7.092	0.0000	0.663	3.460	1.481	0.000

log(1+rainfall): Khok Pho

Oneway Anova: Resid SS: 638.64 r-sq: 0.19 80 missing

factor	df	F	p-val	s	rms	CI/2	Lev p
Period	72/1234	4.072	0.0000	0.719	2.929	1.656	0.000

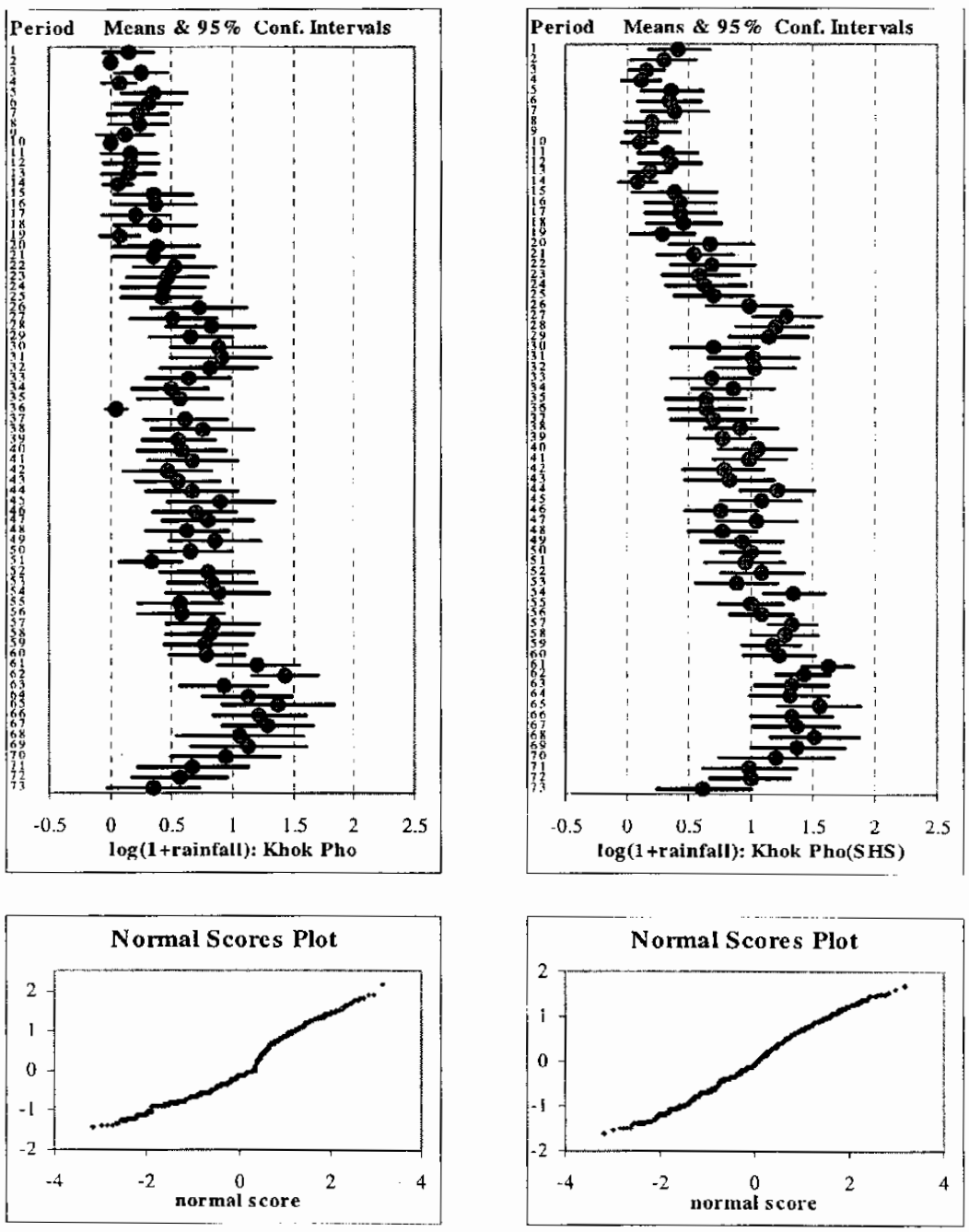

Figure 3.9: One-way Anova Results for Khok Pho & Khok Pho (SHS) Rainfall

log(1+rainfall): Nong Chik

Oneway Anova: Resid SS: 564.11 r-sq: 0.31 12 missing

factor	df	F	p-val	s	rms	CI/2	Lev p
Period	72/1302	8.030	0.0000	0.658	3.671	1.477	0.000

log(1+rainfall): Pattani Airport

Oneway Anova: Resid SS: 467.47 r-sq: 0.33 5 missing

factor	df	F	p-val	s	rms	CI/2	Lev p
Period	72/1309	8.926	0.0000	0.598	3.506	1.338	0.013

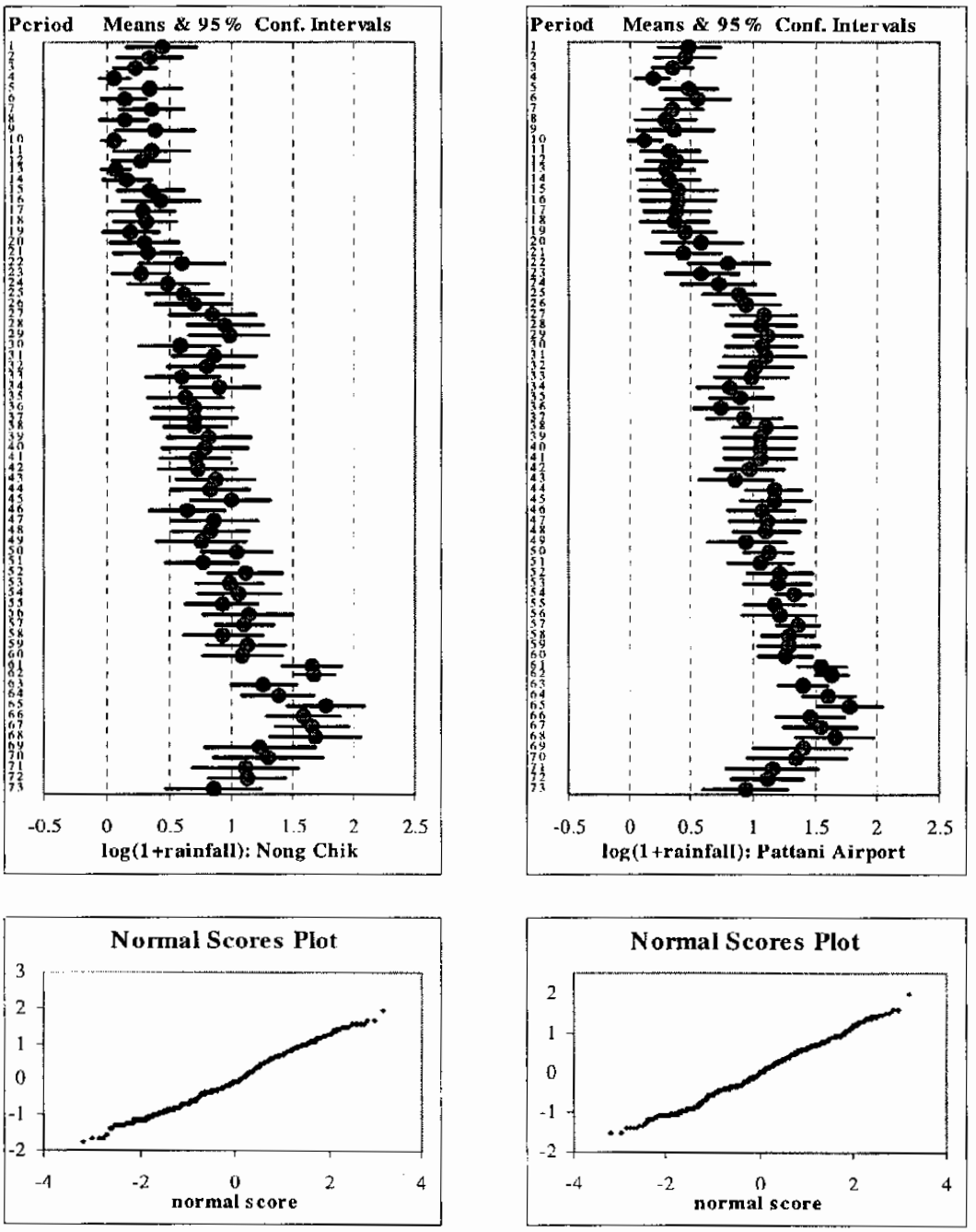

Figure 3.10: One-way Anova Results for Nong Chik & Pattani Airport Rainfall

log(1+rainfall): Kapho

Oneway Anova: Resid SS: 547.26 r-sq: 0.30 84 missing

factor	*df*	*F*	*p-val*	*s*	*rms*	*CI/2*	*Lev p*
Period	72/1157	6.847	0.0000	0.688	3.739	1.630	0.004

log(1+rainfall): Mai Kaen

Oneway Anova: Resid SS: 434.49 r-sq: 0.32

factor	*df*	*F*	*p-val*	*s*	*rms*	*CI/2*	*Lev p*
Period	72/1144	7.577	0.0000	0.616	3.548	1.471	0.024

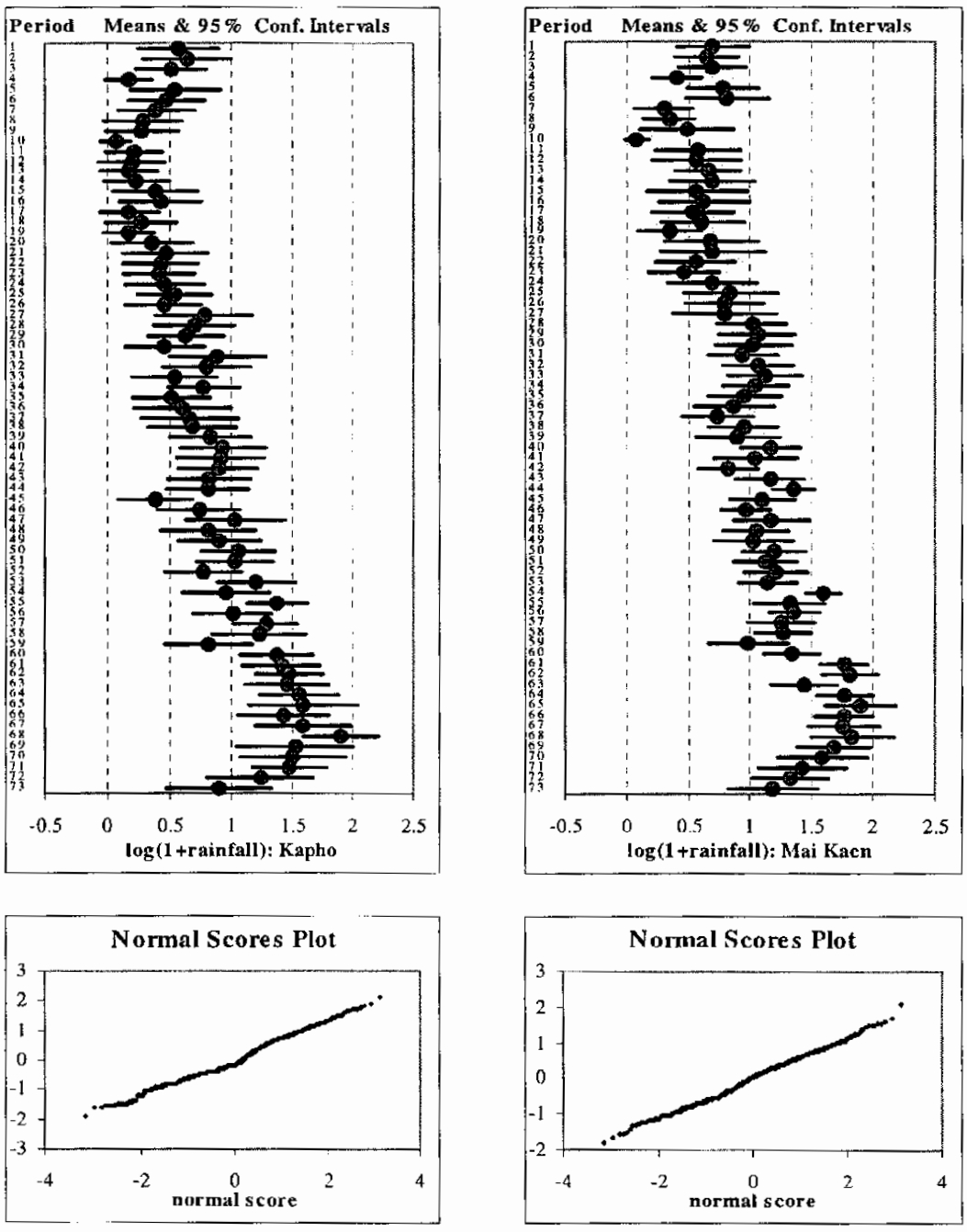

Figure 3.11: One-way Anova Results for Kapho & Mai Kaen Rainfall

log(1+rainfall): Thung Yang Dang

Oneway Anova: Resid SS: 354.84 r-sq: 0.30 6 missing

factor	*df*	*F*	*p-val*	*s*	*rms*	*CI/2*	*Lev p*
Period	72/797	4.685	0.0000	0.667	3.574	1.888	0.024

log(1+rainfall): Mae Lan

Oneway Anova: Resid SS: 148.96 r-sq: 0.39

factor	*df*	*F*	*p-val*	*s*	*rms*	*CI/2*	*Lev p*
Period	72/292	2.648	0.0000	0.714	4.441	3.157	0.985

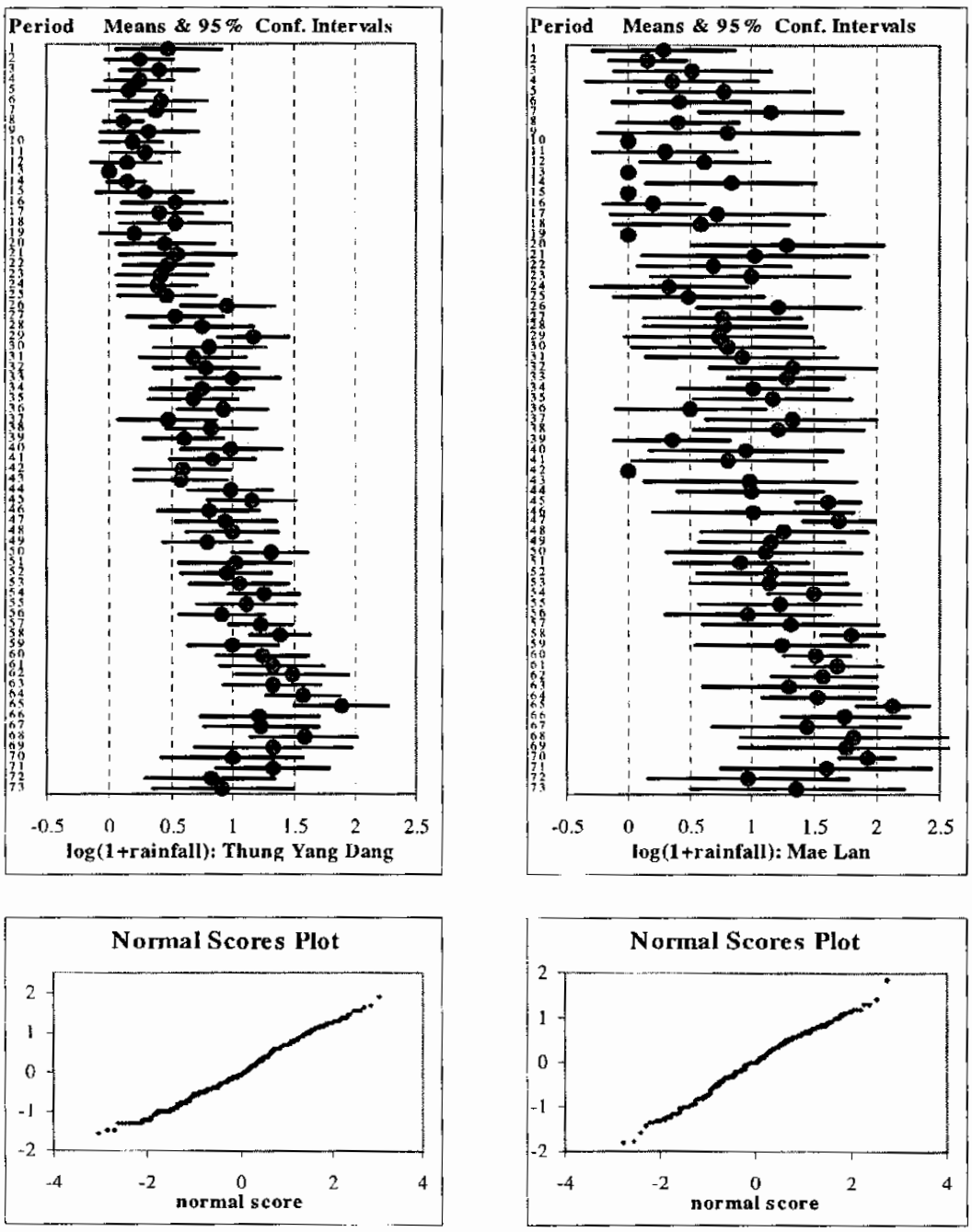

Figure 3.12: One-way Anova Results for Thung Yang Dang & Mae Lan Rainfall

log(1+rainfall): Panare

Oneway Anova: Resid SS: 526.48 r-sq: 0.32

factor	df	F	p-val	s	rms	CI/2	Lev p
Period	72/1314	8.680	0.0000	0.633	3.655	1.415	0.001

log(1+rainfall): Sai Buri

Oneway Anova: Resid SS: 527.137 r-sq: 0.31

factor	df	F	p-val	s	rms	CI/2	Lev p
Period	72/1314	8.199	0.0000	0.633	3.555	1.416	0.030

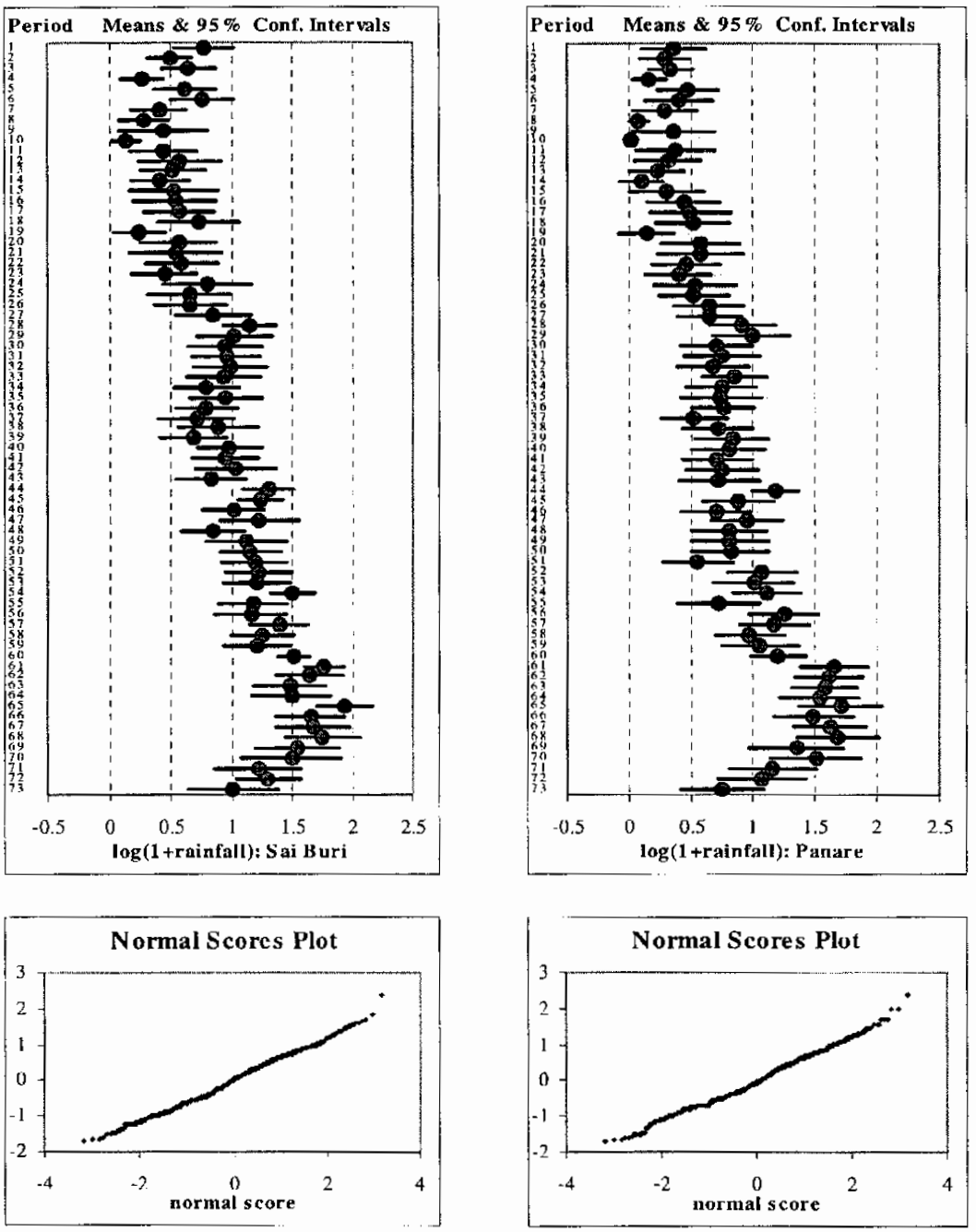

Figure 3.13: One-way Anova Results for Sai Buri & Panare Rainfall

log(1+rainfall): Yarang

Oneway Anova: Resid SS: 575.26 r-sq: 0.29 1 missing

factor	*df*	*F*	*p-val*	*s*	*rms*	*CI/2*	*Lev p*
Period	72/1313	7.520	0.0000	0.662	3.559	1.480	0.000

log(1+rainfall): Mayo

Oneway Anova: Resid SS: 606.746 r-sq: 0.31

factor	*df*	*F*	*p-val*	*s*	*rms*	*CI/2*	*Lev p*
Period	72/1314	8.287	0.0000	0.680	3.834	1.519	0.000

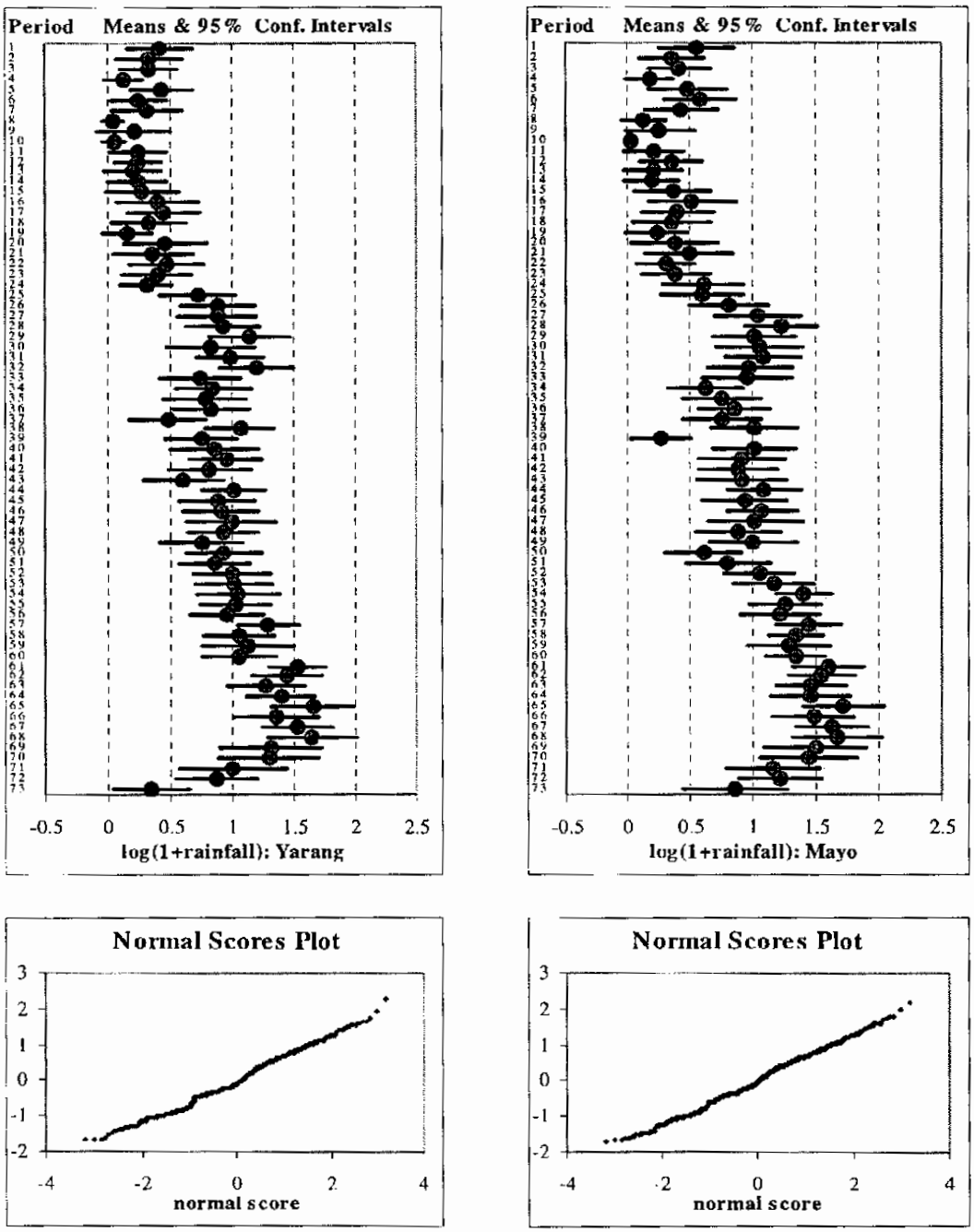

Figure 3.14: One-way Anova Results for Yarang & Mayo Rainfall

Finally, Figure 3.15 shows time series plots of the means of transformed rainfall in the 5-day periods at each station, based on the 19 years from 1982 to 2000.

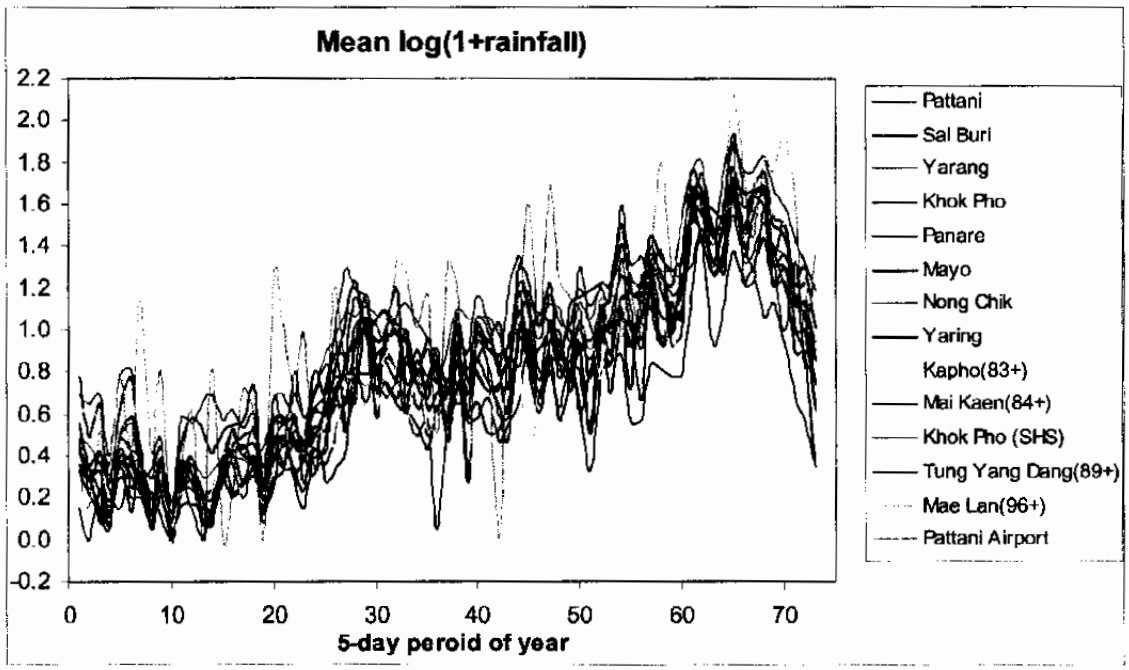

Figure 3.15: Time Series Plots of Means of Transformed Rainfall Totals in 5-day Periods

Figure 3.16 shows two-way anova result of the means of transformed rainfall in the 5-day periods at each station, based on the 19 years from 1982 to 2000.

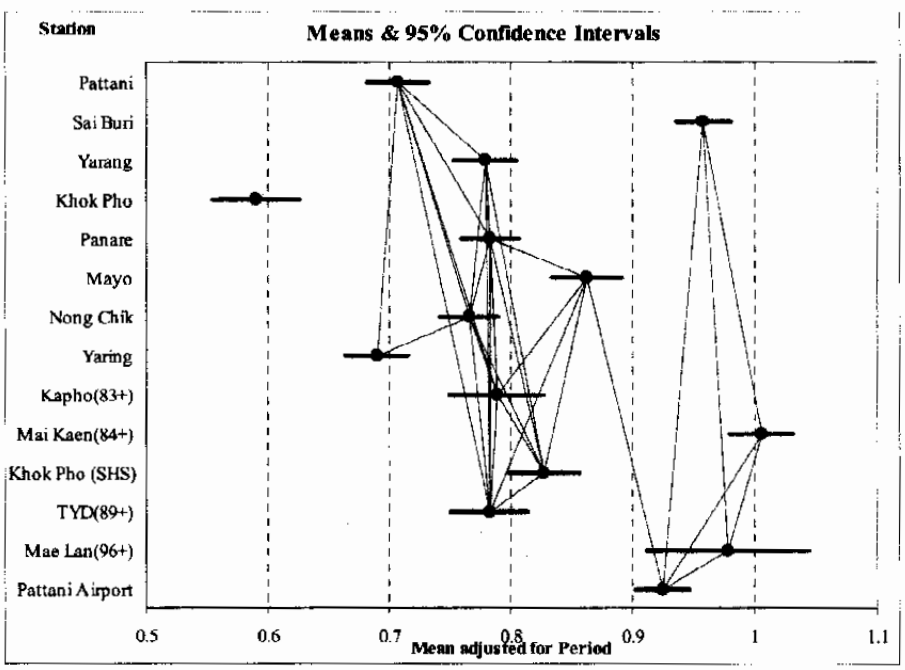

Figure 3.16: Two-way Anova Results of Mean 5-day Rainfall