The Stochastic Generation of Rainfall Time Series

Paul S.P. Cowpertwait

Thesis submitted for the degree of Doctor of Philosophy

University of Newcastle upon Tyne

September 1991

ABSTRACT

The purpose of this project was to propose and validate a stochastic rainfall time series model for the UK, where the model is to be applied to the design of sewer systems.

After reviewing the literature, the Neyman-Scott Rectangular Pulses model was selected as being potentially suitable for the project. Some mathematical properties for the model were derived, and used to fit the model to 10 years of hourly rainfall time series. The model performed well, and so could be used with reasonable confidence for the remaining part of the project.

A full investigation was carried out to find an optimum combination of historical rainfall statistics to be used to fit the model to hourly rainfall time series. A method of fitting the model to daily rainfall time series was also required. It was found that the hourly rainfall statistics used to fit the model to the hourly rainfall time series could successfully be predicted from daily rainfall statistics.

Regression equations were developed so that the mean and variance of the maximum daily rainfalls could be predicted using the parameters of the model. These regression equations were included in the fitting procedure when the model showed a poor fit to the historical daily maxima, so that the model was then able to closely match the historical maxima.

The model was fitted to rainfall data taken from 112 sites scattered throughout the UK. The parameters of the model were regressed on site characteristics (e.g. altitude, distance from coast, etc), so that the model could be used to generate hourly rainfall time series at sites lacking in data.

Finally, a method of disaggregating the generated hourly rainfall time series to 5 minutely time series was developed and tested.

For Sarah, Lydia and Louise

CONTENTS PAGE

LIST OF FIGURES PAGE

LIST OF TABLES PAGE

ACKNOWLEDGEMENTS

I would like to thank my supervisor and colleague, Enda O'Connell, for his support and guidance throughout the project. Thanks also to Andrew Metcalfe for his statistical advice and helpful comments on the text, and to John Mawdsley for his hydrological advice and support during the project. I would also like to thank Trevor Cox for his guidance in Cluster Analysis (Chapter 5). Paul Baldwin and Valerie Isham are gratefully acknowledged for their help with some of the mathematics in Chapter 3. The Computing Services staff at Newcastle University were also of considerable help; thanks to Martin Lamont (Micro computers and software), Dave Surtees (Statistical Packages), Chris Woodford (Numerical Methods), and Dave Sedgwick (Micro computers). In addition, thanks to Ted Edmonds of the Civil Engineering Department for his help with micro-computer software.

I would also like to thank John Threlfall for supporting me during my visits to the WRc. Finally, I am grateful to Ian Clifforde and John Threlfall for providing me with insight into storm sewer modelling, and the Water Research Centre (WRc Engineering, Swindon) for providing the financial support which made the project possible.

AUTHOR'S DECLARATION

The author hereby declares that the largest part by far of the material offered has not been submitted for a degree or other qualification in this or any other university/college.

The exception is the following part, which is based on an MSc dissertation (*Quinn*, 1991) for which the author provided the ideas and supervision: Section 5.8.5 of Chapter 5.

PREFATORY NOTES

Some parts of Chapter 3 and Appendix A have recently been published in Water Resources Research (see *Cowpertwait,* 1991).

After the above paper was published and thesis first submitted, an error was noticed in Chapter 3, Section 3.3. The error was in expression (3.6), which previously read:

$$p_t(h) = (1 - e^{-\beta t} + e^{-\beta(t+h)})(1 - \beta(e^{-\beta t} - e^{-\eta t})/(\eta-\beta)) \times \exp\left\{-\mu\beta(e^{-\beta t} - e^{-\eta t})/(\eta-\beta) - \mu e^{-\beta t} + \mu e^{-\beta(t+h)}\right\} \qquad (3.6)$$

This has been replaced by:

$$p_t(h) = \left\{ e^{-\beta(t+h)} + 1 - (\eta e^{-\beta t} - \beta e^{-\eta t})/(\eta-\beta) \right\} \times \exp\left\{-\mu\beta(e^{-\beta t} - e^{-\eta t})/(\eta-\beta) - \mu e^{-\beta t} + \mu e^{-\beta(t+h)}\right\} \qquad (3.6)$$

The correct version is presented in the thesis. However, the incorrect expression above did appear in many of the programs used for the project. Therefore, comparisons were needed to see whether the consequences of this error would have any practical effect on the results. These comparisons showed that this error could be neglected for the work described in this thesis (details of the comparisons made are provided in Appendix K).

CHAPTER 1

INTRODUCTION AND BACKGROUND TO THE PROBLEM

1.1 RAINFALL TIME SERIES FOR STORM OVERFLOW ASSESSMENT

In designing a sewer system, an engineer requires a rainfall input for the model of the system. Using historical rainfall data, a 'design' storm can be constructed for a given return period (see, for example, *Arnell et al* (1984) for some mathematical details on design storms). Traditionally, drainage engineers have used design storms to design new sewer systems. However, the Sewerage Rehabilitation Manual (1986) recognises that sewerage capital expenditure is now directed towards upgrading existing sewer systems, and consequently there is a need to understand the overall performance of the existing system.

Most sewer systems use Storm Sewage Overflows (SSOs). These divert sewage to local rivers when the system becomes overloaded due to heavy rainfall. SSOs may operate many times each year. Therefore, when the overall impact of pollution on the receiving river is under investigation, the use of design storms of long return periods is unsuitable. Furthermore, design storms are inappropriate when antecedent conditions within an existing sewer system need to be modelled. For example, the pollution impact (due to a storm) on the receiving river may be worse if the storm follows a dry period, when low river flows offer reduced dilution and pollutant concentrations are high due to in-sewer deposition.

Motivated by the inadequacies of the design storm approach, *Henderson* (1986) developed two Time Series Rainfalls (TSRs) for the UK. Rainfall stations were grouped according to whether they lied to the East or West of a dividing line proposed by *Wigley et al* (1984). For each region 'typical' minutely TSRs were selected for each month, using data taken from the sites lying in the region. The selected monthly TSRs were concatenated to produce a typical year of minutely TSR for each region. Thus the engineer could select one of the two TSRs depending on the location of the site under investigation (either East or West). *Henderson's* typical years of TSRs needed improvement for two main reasons:

i) TSRs of more than 1 years duration are required to evaluate a sewer systems performance under more extreme rainfall events, allowing for antecedent rainfall.

ii) The TSRs for each region are not accurate enough for many sites (particularly Northern sites), as they were developed from only two or three stations per region, all lying in the south of England.

The purpose of this project is to improve upon the typical years of TSRs by developing a regionalised stochastic rainfall model that can be used to generate minutely rainfall time series for <u>more</u> than one year for <u>any</u> location in the UK.

1.2 THE DATABASE OF RAINFALL STATIONS

Before the project began, some rainfall data were already available on the mainframe computers at Newcastle University and the Water Research Centre (WRc), Swindon, UK. Other data were bought from the Meteorological Office, Bracknell, UK.

Hourly data were chosen from urban areas, that were not already covered by the available data. This seemed appropriate for the hourly data as they are more expensive than daily data (about 24 times the cost of daily data), and the stochastic rainfall model was going to be used mainly in urban areas.

The remaining daily data were sampled from a Meteorological Office catalogue of rainfall stations. A random sampling procedure was adopted to avoid systematic bias which could effect the results of the regionalisation procedure, which is discussed in Chapter 6. One station was selected randomly from each page of the catalogue, except those pages which had stations in the N-E and S-W of the UK - these areas were already covered by data held at Newcastle University. This procedure seemed appropriate as i) the Meteorological Office catalogue is ordered geographically so that a good spatial coverage of station data was anticipated, and ii) the number of pages in the catalogue was approximately equal to the number of daily station data that could be afforded.

If areas of the UK were not covered by the sampled stations, a station was subjectively chosen within each such area. To compensate, a previously selected station had to be removed. This was achieved by choosing an area well covered, and then randomly deleting a selected station from this area. The complete database of rainfall station data used for the project is listed below in Table 1.1, and Figure 1.1 shows the location of the stations on a map of the UK.

Table 1.1

Rainfall Stations used for the Project

Station Number	Station Name	Type	Yrs	Alt(m)	East Grid	North Grid	No. on Figure 1.1	
1525	Howick Hall	D	92	34	4246	6177	1	(NU)
5349	Cockle Park	D	73	99	4200	5912	2	(NU)
10057	Bellingham	D	21	258	3808	5911	3	
15812	Haydon Bridge	D	19	82	3839	5605	4	
19121	Newcastle	D	21	78	4240	5647	5	
22164	Tunstall Res	D	75	221	4064	5407	6	(NU)
24724	Durham	D	25	102	4267	5415	7	
28106	Hury Res	D	39	261	3967	5193	8	(NU)
43941	Dalton Holme	D	25	34	4965	4452	9	
55659	Thirsk	D	19	35	4438	4818	10	
64718	Askham Bryan	D	25	32	4551	4477	11	
76203	Farnley Hall	D	25	123	4246	4324	12	
81698	Ingbirchworth	D	25	260	4213	4056	13	
82759	Norton Less	D	25	90	4348	3836	14	
91196	Codsall	D	25	125	3870	3028	15	
* 101202	Hollinsclough	D	24	291	4066	3666	16	
108124	Bakewell	D	21	149	4206	3692	17	
115306	Blackbrook Res	D	90	107	4456	3178	18	(NU)
119914	Barnstone	D	24	32	4736	3349	19	
123815	Bevercotes	D	21	24	4696	3735	20	
131736	Scawby Hall	D	25	24	4968	4057	21	
142319	Hackthorne Hall	D	25	32	4992	3825	22	
161255	Thorpe Malsor	D	23	107	4830	2795	23	

	166114	Weston St Mary	D	20	3	5275	3184	24	
	174566	Cardington	D	25	29	5061	2463	25	
	191591	Southerly	D	25	0	5612	2932	26	
	215823	Ormesby	D	25	7	6468	3152	27	
	225557	Cock Station	D	22	55	5964	2364	28	
	232671	Writtle	D	25	35	5677	2066	29	
	238055	Langton Gdns	D	25	27	5537	1875	30	
	240202	Luton	D	25	137	5064	2217	31	
	252556	Buckland	D	24	92	4342	1981	32	
	261604	Dancersend	D	19	198	4906	2089	33	
	269756	Wolverton	D	18	94	4560	1596	34	
	275443	Cippenham	D	25	22	4948	1794	35	
	275574	Windsor	D	90	21	4979	1754	36	(NU)
	283875	West Byfleet	D	23	27	5031	1613	37	
	287196	Merton	D	25	15	5240	1688	38	
	289129	Greenwich	D	25	7	5387	1776	39	
	297340	East Farleigh	D	25	9	5735	1535	40	
	306250	Stonegate	D	22	70	5652	1288	41	
	321314	Cobnor House	D	25	4	4792	1023	42	
	324462	West Tisted	D	19	180	4650	1293	43	
	336376	Boscombe Down	D	25	126	4172	1403	44	
	348847	Dorchester	D	25	95	3684	905	45	
	352316	Forde Abbey	D	64	70	3359	1051	46	(NU)
	354295	Feniton Court	D	58	67	3109	994	47	(NU)
	354864	Exmouth	D	79	66	3027	819	48	(NU)
	355363	Exeter Airport	D	47	32	3001	933	49	(NU)
	356262	Honeymead	D	48	381	2797	1392	50	(NU)
	388933	Okehampton	D	26	372	2585	928	51	
	403490	Durleigh Res	D	25	14	3275	1363	52	
	407349	Rodney Stoke	D	25	40	3488	1501	53	
	417634	Barrow Gurney	D	25	91	3537	1679	54	
	435388	Weston Park	D	25	113	3806	3108	55	
	448545	Rugby	D	25	117	4507	2749	56	
	455775	Bretforton Manor	D	21	40	4092	2438	57	
*	477662	Tafolog	D	20	274	3277	2297	58	
	490228	Pontypridd	D	23	101	3072	1906	59	
	497134	Swansea	D	25	10	2642	1923	60	
	501684	Llandovery	D	25	69	2765	2353	61	
	508283	Orielton Field	D	24	60	1953	1991	62	

	519580	Trawscoed	D	25	63	2674	2736	63	
	549265	Mount Pleasant	D	25	153	3256	3663	64	
	550167	Crosshill Res	D	20	65	3280	3843	65	
	557448	Appleton Res	D	21	30	3602	3845	66	
	561463	Cold Greave	D	22	255	3967	4124	67	
	562991	Heaton Park	D	25	99	3826	4043	68	
	565151	Nether Alderley	D	21	100	3845	3765	69	
	575383	Great Harwood	D	25	204	3722	4327	70	
	587408	Ferry House	D	25	44	3390	4956	71	
	588702	Poaka Beck Res	D	90	156	3240	4781	72	(NU)
	604039	Geltsdale	D	25	229	3575	5537	73	
	623619	Maxwelton House	D	22	107	2820	5896	74	
	656041	Shotts Res	D	27	247	2880	6613	75	
	660285	Abbotsinch	D	22	5	2480	6667	76	
	795076	Fasnaktle	D	25	80	2314	8288	77	
	805389	Inverness	D	25	4	2668	8462	78	
	840573	Old Meldrum	D	14	110	3809	8275	79	
	859107	Dundee	D	25	45	3422	7318	80	
	876839	Cardney House	D	25	107	3051	7452	81	
	888816	Pitreavie	D	25	40	3117	6848	82	
	893230	Argaty	D	21	76	2739	7032	83	
	902952	Samuelston	D	25	64	3486	6711	84	
	914568	Hawick	D	25	96	3512	6156	85	
	968133	Belfast	D	25	5	9380	5250	86	
	969771	Tullynacross	D	25	23	9282	5800	87	
	1584	Boulmer	H	12	23	4253	6142	88	(WRC)
	2245	Leeming	H	8	32	4306	4890	89	
	4913	Filton	H	7	59	3600	1805	90	
	9142	Aldergrove	H	7	68	1450	5300	91	
	96893	Elmdon	H	10	98	4167	2841	92	
	117626	Watnall	H	10	117	4503	3456	93	
	174062	Bedford	H	7	85	5049	2597	94	
	221992	Wattisham	H	10	89	6026	2514	95	
*	235389	Basildon	H	8	12	5737	1907	96	(WRC)
*	236428	Shoeburyness	H	7	2	5961	1878	97	(WRC)
	301114	Manston	H	20	44	6335	1666	98	(WRC)
	346474	Hurn	H	14	10	4117	978	99	
	355363	Exeter	H	10	32	3001	933	100	(WRC)
	433710	Shawbury	H	10	72	3553	3220	101	

547371	Moel-Y-Croes	H	5	263	3194	3699	102	(WRC)
564419	Ringway	H	10	75	3821	3849	103	
577267	Blackpool	H	10	10	3316	4316	104	(WRC)
606336	Carlisle	H	7	26	3384	5603	105	
841496	Dyce	H	7	65	3877	8127	106	
885313	Leuchars	H	7	10	3468	7209	107	
899407	Turnhouse	H	10	35	3159	6739	108	
* 18536	Chopwell Wood	M	5	136	4136	5580	109	(NU)
* 32821	Harpington Hill	M	6	90	4336	5267	110	(NU)
**125842	Finningley	M	15	10	4659	3989	111	(NU)
246690	Hampstead	M	35	137	5262	1863	112	(NU)
260991	Abingdon	M	32	69	4479	1991	113	(WRC)
271432	Farnborough	M	31	69	4867	1544	114	(WRC)
309038	Hastings	M	9	45	5809	1094	115	(WRC)
309902	Herstmonceaux	M	6	18	5645	1099	116	(WRC)
383478	St Mawgan	M	10	103	1873	642	117	(NU)
492325	Rhoose	M	18	65	3066	1678	118	(NU)
* 567423	Aigburth	M	8	12	3384	3852	119	(WRC)
660628	East Kilbride	M	3	178	2638	6535	120	(NU)

Key:-

*=incomplete, **=corrupt,

D=daily, H=hourly, M=minutely,

NU = data held at Newcastle University, WRC = data held at WRc.

N.B. The incomplete data (i.e. data containing many missing values) or the corrupt data were not used in the Project. Most of the data contained some missing values, which were usually taken as zero (see Appendix H for a discussion of the treatment of the missing values).

1.3 LAYOUT OF THE THESIS

This Section describes the overall layout of the thesis.

Chapter 2 contains a literature review on some of the most recent developments in rainfall time series modelling. From the review, the most promising model is identified, though judgement on the suitability of the model is reserved until the results of a preliminary analysis are obtained in Chapter 3. Chapter 2 also contains a review of recent literature that identifies homogeneous rainfall regions for the UK. From this review, some regions were selected as being suitable for the future sampling requirements of the project.

In Chapter 3 some theoretical properties for the model are developed, and an initial investigation into the performance of the selected stochastic model is carried out. The results in Chapter 3 showed that the selected model was worth persisting with for the remainder of the project.

There are many ways of fitting the selected stochastic rainfall model to historical hourly rainfall data. The purpose of Chapter 4 is to find an optimum fitting procedure for the model by choosing the best combination of historical statistics to estimate the parameters of the model.

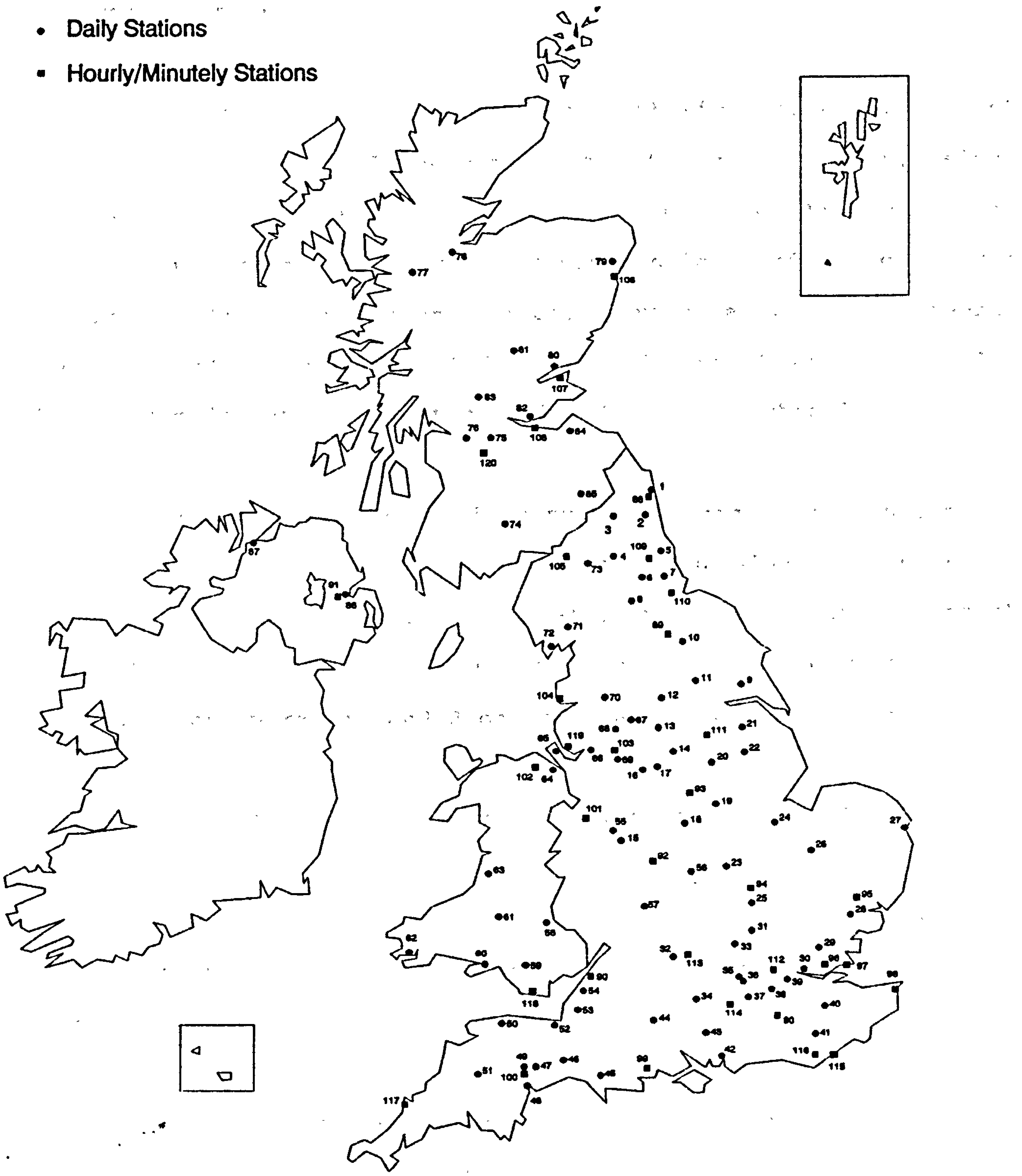
To produce a regionalised model use must be made of the available daily rainfall data, which are less expensive and more readily available than hourly data. The purpose of Chapter 5 is to find a suitable way of fitting the model to daily data. The Chapter concludes by fitting the stochastic rainfall model to five long records of daily data, and comparing the extremes generated by the model with those of the historical records.

The aim in Chapter 6 is to develop a regionalised stochastic rainfall model. In Chapter 6 the parameters of the stochastic model are estimated for each station-month, and these estimates are then regressed on site characteristics (e.g. altitude), so that the model can be used at sites lacking rainfall data.

In Chapter 7 a method of disaggregating hourly rainfall data into minutely data is proposed and tested.

Finally, some overall conclusions and directions for future research are given in Chapter 8.

Figure 1.1
Rainfall Stations used for the Project



CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

Rainfall is the result of complex atmospheric processes. Attempts have been made to model these processes deterministically, using a knowledge of atmospheric physics. However, the physics are not completely understood, so that deterministic rainfall models tend to be of little practical value, particularly in engineering design problems (see *Cho* (1985), and *Cho and Chan* (1987)). This has lead modellers to treating rainfall as a stochastic process.

With the rapid development of computing science and technology in the last decade, the potential for assessing and comparing stochastic rainfall models has greatly increased. Much work on the theoretical development of rainfall models has been completed, although more work is needed on model assessment and validation.

Stochastic rainfall models can be divided into four classes:

1) Temporal/single site Models. These model rainfall at a single site, without attempting to spatially distribute the rainfall over a catchment area.

2) Spatial/Field Models. These model the distribution of rainfall over a large spatial area, without attempting to model the rainfall time series for periods exceeding the storm duration. They are usually based on data taken from raingauges scattered

throughout the area in which the rain is to be modelled. The distribution of the rain between gauges is inferred using the spatial model.

3) Mult-site models. These model rainfall at more than one site using data taken from gauges at the sites. They do not attempt to infer rainfall patterns between sites.

4) Spatial-temporal Models. These model both the rainfall time series for long periods, and the spatial distribution of the rainfall over a catchment area.

The purpose of this project is to produce a stochastic rainfall time series model for the UK. Therefore, attention is focused on (1) above, i.e the temporal modelling of rainfall.

2.2 SOME SIMPLE STOCHASTIC MODELS

The models described in this Section have lead the way to the development of more complex stochastic rainfall models.

Temporal rainfall can be described by two sequences of random variables. The first sequence models the occurrence of rainfall events within successive time intervals, and the second sequence associates rainfall depths with each rainfall occurrence.

2.2.1 A simple Binomial model for rainfall occurrences

First consider a sequence of random variables: Y_1, Y_2, ..., Y_n, where Y_i denotes the rainfall depth on the ith day. Note, any discrete increment, other than days, could be used (e.g. hours or minutes), but it makes easier reading to fix on some particular increment.

Now let the sequence $\{X_i\}$ be defined by:

$$X_i = 1 \text{ if } Y_i > 0$$
$$X_i = 0 \text{ if } Y_i = 0$$

i.e. $X_i = 1$ if ith day is wet,

and $X_i = 0$ if ith day is dry,

and let $N_n = \sum_1^n X_i$ be the number of wet days in the n day sequence.

Perhaps the simplest stochastic model is that which assumes the X
are independent and identically distributed random variables with

$$pr\{X_i = 1\} = p \text{ and } pr\{X_i = 0\} = 1-p = q$$

It then follows that N_n is a Binomial B(n,p) random variable, i.e

$$pr\{N_n = k\} = \binom{n}{k} p^k (1-p)^{n-k} \qquad [0 \le k \le n]$$

For applications of this model the reader is referred to *Smith an Schreiber* (1973).

The Binomial process has not been used extensively for rainfal modelling. However, its continuous time counterpart, the Poisso Process, has received much attention.

2.2.2 A simple Poisson model for rainfall occurrences

Let N(t) denote the number of rainfall occurrences in th continuous time interval (0,t), where the rainfall events ar assumed to be point occurrences. Then, under the assumption tha rainfall occurrences follow a Poisson Process with rate λ,

$$pr\{N(t) = k\} = (\lambda t)^k e^{-\lambda t}/k! \qquad [k = 0, 1, 2, \ldots]$$

For an application of the Poisson Process to modelling rainfal the reader is referred to *Todorovic and Yevjevich* (1969) *Todorovic and Woolhiser* (1976), *Eagleson* (1978), o *Rodriguez-Iturbe et al* (1984).

2.2.3 A simple Markov Chain model for rainfall occurrences

The main drawback with the above two models is the assumption that rainfall occurrences are independent events. Since this assumption cannot usually be retained (e.g. see *Kavvas and Delleur* (1975), or *Todorovic and Woolhiser* (1976)), a more popular approach has been to model the occurrence process $\{X_i\}$ with Markov chains. The probabilistic structure of a Markov chain is completely determined by its transition matrix and an initial probability distribution (e.g. see *Feller* (1968)). Let the transition matrix be denoted by:

$$\underline{P} = \begin{bmatrix} q_0 & p_0 \\ q_1 & p_1 \end{bmatrix},$$

where: $p_0 = pr\{X_i = 1 | X_{i-1} = 0\}$, $p_1 = pr\{X_i = 1 | X_{i-1} = 1\}$,
$q_0 = 1 - p_0$, $q_1 = 1 - p_1$ $[i = 1, 2, ...]$.

i.e. p_0 is the probability that the ith day is wet given the (i-1)th day is dry, etc.

An expression for $pr\{N_n = k\}$ was found by *Gabriel* (1959).

The transition probabilities for the Markov chain rainfall model can be estimated using the equivalent proportions taken from the historical rainfall data. For example, to estimate p_1, the proportion of wet days with the previous day wet could be used. Alternatively, the historical wet/dry spell sequences could be used (refer to *Waymire and Gupta* (1981) for details).

For applications of this model the reader is referred to *Gabriel and Neumann* (1962), *Smith and Scheiber* (1963), and *Todorovic and Woolhiser* (1976).

2.2.4 Modelling the depth of rain

Given that a rainfall event occurs, the depth of rainfall can then be modelled. In the discrete case, perhaps the simplest approach is to assume that the depth of rain on wet days is independent of the depths on previous days and follows an Exponential distribution (e.g. see *Todorovic and Woolhiser* (1976)). In the continuous case, the simplest approach may be to associate an instantaneous depth of rain with each rainfall occurrence, where the depths of rain are again assumed to be independent Exponential random variables (e.g. see *Rodriguez-Iturbe et al* (1984)). In the discrete case, the parameter of the Exponential distribution can be obtained by the Method of Moments, using the mean depth of rain on wet days. In the continuous case, the aggregated properties of the model are needed because rainfall data are usually only available as historical records of discrete time series. These aggregated properties have been found for the simple Poisson model of occurrences with an Exponential distribution for the depth of rain (see *Rodriguez-Iturbe et al* (1984) and *Parzen* (1967) - this model is discussed in more detail in Section 2.4 of this Chapter).

If the Exponential distribution provides a poor fit to the depth process, a more complex distribution (e.g. the Gamma or Weibull) can be used (e.g. see *Eagleson* (1978)). Again, the method of moments could be used in the fitting procedure.

2.2.5 Summary

In summary, two approaches have been used for representing rainfall time series:

(i) The Discrete Time Series approach, in which a discete time increment (e.g. a day or an hour) is used for the time series model. Typically, a Markov chain is fitted to the sequence of wet and dry spells, together with some distribution for the amount of rain captured in the wet intervals.

(ii) The Point Process approach, which uses a continuous time model for the occurrence of the rainfall events, and associates some random amount of rain with each event.

Recent advances using these two approaches will now be considered. For other reviews on rainfall modelling, the reader is referred to *Foufoula-Georgiou* and *Georgakakos* (1988), or *Waymire and Gupta* (1981).

2.3 A REVIEW OF SOME RECENT DISCRETE TIME SERIES MODELS

Stern and Coe (1984) used a model with Markov chains fitted to the occurrence of rainy days, and a Gamma distribution fitted to the amount of rain captured on wet days. To take account of seasonal effects curves were fitted to the transition probabilities throughout the year. The mean of the Gamma distribution was also allowed to depend on whether rain had occurred on previous days, and the necessity of this was checked using standard statistical tests. Depending on the result of these tests, for a given site the model used between 20 and 50 parameters with seasonal effects taken into account. Examples of the performance of the model were given, using data taken from a 53 year record in Morogoro, Tanzania, and a 37 year record taken from Irbid, Jordan. The performance of the model was shown to be good for the intended application, which was agricultural planning.

Foufoula-Georgiou and Lettenmaier (1987) developed and used the Markov Renewal Model of daily rainfall occurrences, with a mixed Exponential distribution for the amount of rain captured on wet days. They used 2 Geometric distributions to model the inter-arrival times (in number of days) of the rainfall events. Events were classified as either primary or secondary, where a primary event corresponded to the arrival of a front, and a secondary event corresponded to the occurrence of rainfall within the same frontal system. The model thus exhibits clustering as a result of this dependence structure. The model uses 4 parameters for the occurrence process and then an additional 3 for the depth process, before seasonal effects are taken into account. They tested the model using 15 years of daily data taken from

Snoqualmie Falls, Washington, by comparing plots of various statistics. They also compared the Method of Moments with the Maximum Likelihood Method for estimating the parameters of the model and deduced that the latter gave better results. The data were divided into 5 seasons and the parameters were estimated for each season using the maximum likelihood method. A comparison was made between the model values and the actual values of the mean and standard deviations of the daily totals for each season. A good fit was evident, although no formal statistical tests were made.

Smith (1987) developed the Markov Bernoulli Model for daily rainfall occurrences, which is a generalisation of Markov chains and Bernoulli trial point processes. This occurrence model uses 9 parameters, and includes a seasonal structure. The Maximum Likelihood method was recommended for parameter estimation. The model was compared with standard Markov chain and Bernoulli models using 10 years of daily data taken from Washington, DC. By classifying a day as 'wet' if more than a threshold of 0.1 inch of rain fell on the day, the model was shown to be preferrable to the standard Markov chain and Bernouilli trial models. For large thresholds (1 inch) the Bernoulli trial model was found to provide the best fit, and for small thresholds (0.01 inch) a Markov chain model gave the best fit. It was concluded that the choice of model should depend upon the intended application.

2.4 RECENT POINT PROCESS MODELS

In this section several point process models will be considered. It will be found helpful if some of the models are defined prior to the discussion.

2.4.1 The Poisson White Noise model.

Rainfall occurrences are assumed to occur in continuous time according to a Poisson process. The depth of rain, associated with each rainfall occurrence, is a random variable, and is assumed to occur as an instantaneous 'burst'. Depths are aggregated to intervals which match historical records, for the purpose of parameter estimation. Expressions for the second order moments of these aggregated depths are well known in the literature (see *Parzen* (1967), or *Rodriguez-Iturbe et al* (1984)).

2.4.2 The Poisson Rectangular Pulses model

In the literature this model is sometimes referred to as the Rectangular Pulses (Markovian) model. The model assumes that rainfall events arrive according to a Poisson process and that each rainfall event has a random duration and intensity associated with it. The intensity is assumed to be constant throughout the duration of the rainfall event. The intensity and duration are often taken to be Exponentially distributed.

2.4.3 The Neyman-Scott White Noise model

The Neyman-Scott (N-S) point process was first used by *Neyman* (1939) in entomology and bacteriology population growth modelling, and subsequently used by *Neyman and Scott* (1958) to model the spatial variation of galaxies in the Universe. *Kavvas and Delleur* (1975) first used the model for representing rainfall events. They derived the probability generating function for the occurrence process, and fitted the model to daily rainfall sequences in Indiana. More recently, the second order moments of the aggregated process have been found (*Rodriguez-Iturbe et al* (1984)). The N-S White Noise model assumes that, with any rainfall event, there exists some generating mechanism, often called the STORM ORIGIN, from which rain cells arise. The generating mechanism could be regarded as passing fronts or some other criteria for convective storms. It is assumed that the storm origins arrive according to a Poisson process, and that the number of cells associated with each origin is a random variable. Furthermore, the waiting times for the rain cells, after the storm origin, are independent and identically distributed random variables (usually Exponential). With each cell is associated an instantaneous rainfall burst of random depth.

2.4.4 The Neyman-Scott Rectangular Pulses model

The model definition is as for the N-S White Noise model with the exception that each rain cell has a random duration and intensity associated with it, instead of instantaneous depths of rain. The second order moments for the aggregated process have been found by

Rodriguez-Iturbe et al (1987a), under the following assumptions: i) the waiting times after the storm origin for the cell origins are independent (of other cell and storm origins) Exponential random variables, and ii) the cell durations are independent Exponential random variables.

2.4.5 The Bartlett-Lewis Rectangular Pulses model

The Bartlett-Lewis (B-L) Rectangular Pulses model is similar to the N-S Rectangular Pulses model. Storm origins arrive according to a Poisson process, and a random number of rain cells are associated with each origin. The duration of the storm is a random variable with some probability density function, usually Exponential. Cells are then generated between the origin and the end of the storm, with the time interval between cell origins following some continuous probability distribution, e.g. Exponential. The duration and intensity of the cells are independent of all other cells and storm origins. *Rodriguez-Iturbe et al* (1987a) derived the second order moments of the aggregated process, and an expression for the probability of an arbitrary interval being dry.

2.4.6 Discussion and comparison of various Point Processes

Cox and Isham (1980) outlined the general theory of point processes. Their book includes a discussion of cluster point processes, such as the N-S and B-L models.

Waymire and Gupta (1981) provided a mathematical summary of the theory of Point Processes, with emphasis given to the probability generating functional (a generalisation of the probability generating function - see *Cox and Isham* (1980) for details) of the counting process of rainfall occurrences. Using the probability generating functional for the N-S model, they derived, in a more direct manner, some of the results found by *Kavvas and Delleur* (1975).

Rodriguez-Iturbe et al (1984) compared the Poisson White Noise model, the Poisson Rectangular Pulses model, and the N-S White Noise model. Their main interest was in comparing the performance of the models at the hourly and daily levels of aggregation. Equations were derived for the aggregated processes and were used to fit the models to 27 years of data taken from Denver over a 1 month period (May 15 to June 16), and 11 years of data taken from Agua Fria for the months of April and September. For each model, comparisons were made between parameters estimated at the hourly level and parameters estimated at the daily level. It was clear that the parameters estimated for the N-S model at the hourly level were close to the parameter estimates at the daily level, which was not observed for the other two models. The correlograms for the N-S and Poisson Rectangular Pulses models were also given. From these it could be seen that both of these models fitted well at the daily level, but only the N-S model fitted well at the hourly level.

Valdes et al (1985) compared the same three models with a spatial-temporal model proposed by *Waymire et al* (1984). The purpose was to test the feasibilty of approximating a spatial-temporal process, i.e. rainfall, with a simple stochastic time series models. They found that only the N-S White Noise model preserved the average storm duration and the average depth of rain captured over the duration of a single storm. They also performed an extreme value analysis by fitting a Gumbel distribution to the extremes at the hourly and daily level for the Poisson Rectangular Pulses model, the N-S White Noise models, and the spatial-temporal model. It was found that neither model adequately reproduced the extreme values obtained from the spatial-temporal model, the N-S White Noise model underestimating the extremes when fitted at the hourly level, and over estimating the extremes when fitted at the daily level (the Poisson Rectangular Pulses model behaved in an opposite way, with a poorer fit).

Perhaps motivated by the inadequacies of the N-S White Noise model, *Rodriguez-Iturbe et al* (1987a) developed the N-S and Bartlett-Lewis (B-L) Rectangular Pulses cluster models for representing rainfall. In their paper the aggregated second order moments were derived for each of the models. In addition, some further properties were found for the B-L model, for example, the probability of an arbitrary interval being dry.

Rodriguez-Iturbe et al (1987b) presented a detailed empirical analysis of rainfall data taken from Denver, Colorado. They compared the performance of the Poisson Rectangular Pulses, and the N-S and B-L Rectangular Pulses cluster models. Their paper showed that the N-S and B-L models were able to preserve rainfall statistics at various levels of aggregation (from 1 hour to 24 hours). The extreme values (up to return periods of about 20 years) for the models were plotted on Gumbel probability paper against the return period, and these plots compared favourably with the equivalent plot for the historical data. The probabilities of zero rain were also compared for 1, 6, 12, and 24 hour time intervals. At first sight the cluster models (N-S and B-L) appeared to over estimate the probability for the larger intervals. However, an improvement was found when the threshold for a dry historical interval was increased from zero to a small upper bound, so that an interval was classified as dry if the amount of rain captured in the interval fell below the threshold. No difference in the performance of the N-S and B-L model was found. The Poisson Rectangular Pulses model had a much poorer fit to the historical data than the cluster models.

To improve the fit of the N-S model to the historical proportion of dry days, *Entekhabi et al* (1989) proposed a modified N-S model by allowing the cell duration to vary for each storm according to a Gamma distribution (which introduced an additional parameter into the model). They found that the proportion of dry days given by the modified N-S model compared favourably to the historical records.

2.5 OTHER RAINFALL MODELS

Swartenbroekx (1987) wrote a document on applying the "Point-Rainfall Generator" (*Marien and Vandewiele,* 1986) on a mainframe computer. The model presented was able to generate rainfall time series for intervals from 10 minutes upwards. The model used 19 parameters when seasonalised to 2 seasons: summer and winter. It was found that the amount of data available (13.5 years) was insufficient to build a model with more refined seasonality. The model was shown to compare favourably with extreme value statistics for return periods of up to 100 years. Other statistics were not compared, because the purpose of the model was to simulate severe storm conditions.

Ormsbee (1989) proposed two rainfall disaggregation models. The first model disaggregates historical hourly rainfall time series to discrete time series of 20 minute intervals, and the second model disaggregates hourly time series to discrete time series of intervals for any chosen length from 1 to 30 minutes. The performance of the second model was assessed by comparing the predicted peak discharge flows (from a watershed) when using the model with the observed flows, and with flows predicted by assuming a Uniform distribution of rainfall over the hour. The disaggregation model showed an improvement when compared with the Uniform model, but under estimated peak flows when compared with the historical data.

Acreman (1990) developed a model to generate hourly rainfall data for Farnborough, UK. The historical rainfall time series were divided into wet and dry spells, and the Exponential and Pareto distributions (respectively) were fitted to the spell lengths. In addition, a Gamma distribution was fitted to the total volume of rain captured in the wet spells. The model had two seasons (Summer and Winter) and 22 parameters. Some of the model parameters were dependent on the season, and others were constant throughout the year. Using the model, data were simulated, and the mean simulated monthly totals were compared with the mean historical monthly totals. The results showed that the simulated monthly totals were consistently greater than the historical monthly totals for December to April, and consistently less than the historical totals for July to October. This model could possibly be improved by introducing more than two seasons. However, when fitting a wet/dry spell model, it is desirable to have as few seasons as possible to reduce the problem of the spells over-lapping the seasons, and small numbers of spells (e.g. wet spells in summer seasons) which leads to high sampling variability. A further problem with modelling wet/dry spells is the ambiguous definition of a wet/dry spell. For example, should a long sequence of wet hours with one central dry hour be treated as one or two events? Various definitions of spell lengths are available in the literature (e.g. see *Yen and Chow* (1980), or *Restrepo and Eagleson* (1982)), but any definition is likely to be subjective in some way.

2.6 THE CHOICE OF RAINFALL TIME SERIES MODEL

The rainfall model required must be able to generate rainfall time series from about 5 minute intervals upwards. There are two approaches which could be taken:

i) A model could be fitted to daily rainfall time series and the generated daily rainfall data disaggregated empirically to the required level. The choice of model would probably be a Markov Renewal model, as developed by *Foufoula-Georgiou and Lettenmair* (1987), or some other Markov chain type model, such as the one developed by *Stern and Coe* (1984). However, in the literature, only one model (*Hershenhorn and Woolhiser* (1987)), which was developed empirically for daily data located in the USA, could be found on disaggregating daily rainfall time series, and so this approach was not favoured.

ii) A continuous time model could be fitted to hourly rainfall data. The required stochastic time series could then be obtained by aggregation, if higher than hourly time steps are required. As hourly rainfall models have been tested in the literature, at various levels of aggregation, this approach was favoured. If the continuous time model failed to perform well at increments less than 1 hour, the hourly series would need to be disaggregated, and for this a model such as the one developed by *Ormsbee* (1989) could be used (see Chapter 7).

The model selected was the Neyman-Scott Rectangular Pulses Cluster model, as developed by *Rodriguez-Iturbe et al* (1987a). This model seemed the most promising for the following reasons:

(a) The model was shown to preserve historical rainfall statistics at various levels of aggregation. This included extreme values up to return periods of about 20 years (see *Rodriguez-Iturbe et al* (1987b)).

(b) The model requires only 5 parameters to be estimated, for all levels of aggregation of the hourly time series. To simplify the regionalisation of the model, it seemed important that the model should have as few parameters as possible.

(c) The model has a realistic physical structure, i.e. the incorporation of rain cells which are known to exist in actual rainfall events (e.g. see *Amorocho and Wu*, 1976, or *Shaw*, 1982). This makes interpretation of the parameters of the model easier.

(d) The Neyman-Scott model (White Noise or Rectangular Pulses) performed better than (or as well as) other rainfall models in the literature.

The Bartlett-Lewis Rectangular Pulses model also satisfies (a)-(c) above, and no difference in the performance of the two models has yet been found. However, the Neyman-Scott model has appeared in the literature on rainfall modelling since 1975 (*Kavvas and Delleur*, 1975), and has been presented in many more hydrology journals. Although the extra attention given to the Neyman-Scott model may not be fully justified, it does provide some support for choosing the Neyman-Scott model in preference to the Bartlett-Lewis model.

The literature on rainfall modelling contains only a few papers on fitting hourly rainfall models to data, and no papers were found on fitting the N-S model throughout the year. Therefore, judgement on the suitability of the N-S model was reserved until some further work in fitting the model had been completed. The reasons listed above provided sufficient motivation to test the performance of the model against historical rainfall data.

2.7 THE REGIONAL VARIATION OF RAINFALL IN THE UK

2.7.1 Introduction

The stochastic model must be capable of generating rainfall time series at any location in the UK. It was anticipated that the model parameters would be linked to certain regions of the country as well as criteria, e.g. altitude, within each region. The purpose in this Section was to decide upon an appropriate division of the UK into homogeneous precipitation regions.

The selected regions could also be used as strata for sampling schemes. The need for the sampling schemes will become more apparent in subsequent Chapters. For the present, note that there are many rainfall station data available for the analyses (listed in Table 1.1), and so at times it will be found convenient to take a sample of these. A stratified sampling scheme based on precipitation regions would ensure a good spatial coverage of station data.

Two papers (*Wigley et al* (1984), and, *Gregory,* (1975)) and one report (*Dales and Reed* (1989)) are considered to determine whether the regions described therein are suitable for the project. The two papers are reviewed in some depth and Principle Component Analysis, which is used in both papers, is judged to be satisfactory as a method for grouping the rainfall stations.

2.7.2 Summary of the papers

Dales and Reed (1989) used Wiltshire's G-Point statistical test, applied to annual 1-day maxima, to adapt initial regions similar to those given by *Jackson and Larke* (1975). The final choice divided the country into 11 homogeneous regions based on at least 40 years of data for each of 401 rainfall stations. It was noted that the G-Point test still produced significant values for 2 of the regions (at the 5% and 1% levels). By looking at the values of the test statistic it could be seen that many of the values were nearly significant and some were significant at the 10% level. However, the G-Point statistical test involves folding the distribution, which amplifies differences, and is therefore likely to be highly sensitive.

The regions proposed by *Dales and Reed* (1989) were judged to be unsuitable for the project for the following reasons:

(i) The statistics being considered (annual 1-day maxima) probably did not cover general regional differences in rainfall. In particular, the altitude of the rainfall station was not taken into consideration in the statistical tests. This means that the inferred regions were likely to be a result of regional differences in rainfall patterns which could be attributed to the altitude of the stations within the region, rather than more general rainfall patterns rainfall (e.g. frequency of events in the region, etc.).

(ii) The regions were selected prior to the statistical tests, and so other regions not considered could have been 'better'.

(iii) Many of the results for the statistical tests were significant or almost significant.

Gregory (1975) considered various methods of dividing the UK into homogeneous regions using a 70 year record (1881-1950) of annual rainfall data taken from 50 stations. Four approaches were taken:

(i) A graphical method, which involved plotting 10-year running means and (subjectively) drawing regions. In conclusion, Gregory stated that "Regional coherence was apparent, and possible simple circulation causes could be inferred, although full interpretation of boundaries was neither obvious nor clear". Thus, these regions were not selected, because the method seemed too subjective, and failed to produce physically identifiable boundaries.

(ii) A linkage analysis was carried out between all possible pairs of stations over the 70 year record. This approach produced regions which appeared to have some physical interpretation, for example a division across the Pennines. However, it appeared that some of the boundaries for the regions could have no objective basis, because of lacking station data

(iii) Principle component analysis (PCA). Gregory produced 5 regions based on a PCA. The regions were based on the loadings on the first and second principle components, which accounted for about 63% of the variance in the data. Gregory found that the sites could be grouped into 5 catagories:

(A) Stations for which component I was major (i.e. accounted for more variance than the other components), but was not dominant, and component II was negative.

(B) Stations for which component I was dominant (i.e. comprised of over 50% of the variance for that site, with a loading greater than 0.7), and component II was negative.

(C) Stations for which component I was dominant, and component II was positive.

(D) Stations for which component I was major, and component II was positive.

(E) Stations for which component II was major, and component I was positive.

These 5 groups of stations were seen to occupy regionally discrete units, although some of the boundaries for the regions must have been subjective because of lacking station data.

(iv) Factor Analysis. Gregory also rotated the solution of the PCA in order to maximise the loadings on as many stations as possible. This essentially meant that Gregory treated the principle components as factors in a factor analysis (FA). The new 'factors' that appeared were not the principle components, and have no obvious interpretation. Gregory stated that "...the PCA results were used (despite some theoretical objections)", of which the "objections" may be this lack of interpretation. Many statisticians regard FA as inappropriate for most practical problems. For example, *Chatfield and Collins* (1980) list a page of drawbacks on FA, as well as many critical quotations on the method, and finally end with the remark: "...we recommend that FA should not be used in most practical situations".

Wigley et al (1984) used a PCA on 55 stations, mainly located in Great Britain, for the 110 year period 1861-1970. They based their choice of homogeneous regions on contour maps of the loadings on the first four principle components for annual and seasonal data. These maps suggested that the UK could be divided into 5 homogeneous regions (see figure 2.1).

They then selected a station from each of the regions and found the correlation between the annual totals of the selected station and the remaining stations. Contour maps of the correlations were drawn. From these maps it was evident (although not suprising) that the selected stations had high correlations with stations lying in the same region.

Wigley et al's PCA was preferred to *Gregory's* PCA for the following reasons:

a) Their rainfall stations gave a better spatial coverage of the UK.

b) They considered longer time periods (110 year records compared with 70 year records).

c) They considered the loadings on more principle components (four compared with two).

d) They performed a separate PCA for monthly, seasonal, and annual data (Gregory only considered annual data).

Therefore, the regions proposed by *Wigley et al* (1984) were chosen for the project. It should be mentioned that any method involving the drawing of boundaries on maps is likely to be subjective in some way. However, the regions selected do seem physically realistic (e.g. there is an NE/NW division which corresponds to the well known 'rain shadow' effect of the Pennines) and can easily be classified (as North-East, North-West, Central, Southern, and South-West (see Figure 2.1)).

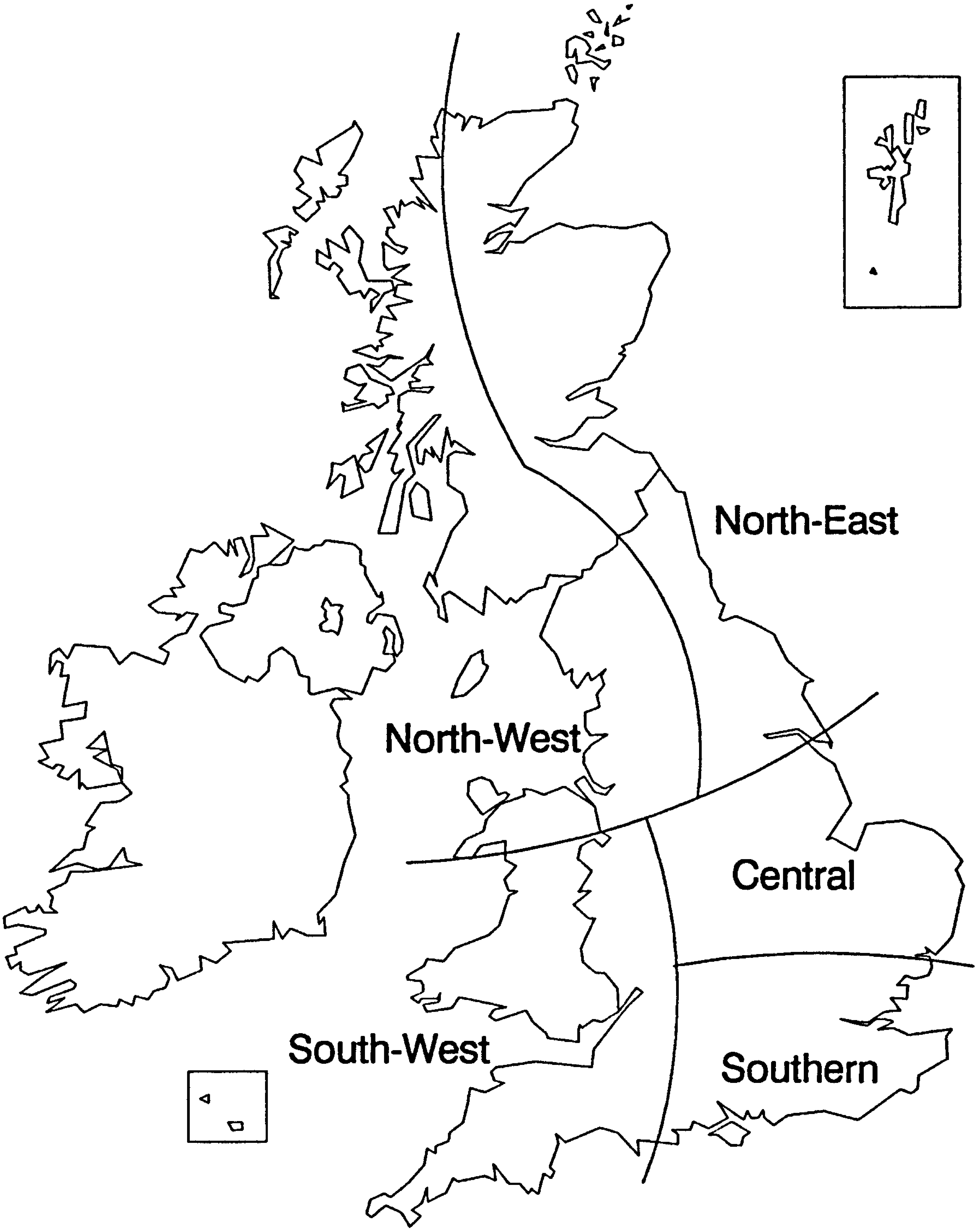


Figure 2.1
Homogeneous Regions proposed by Wigley et al (1984)

CHAPTER 3

AN INVESTIGATION OF THE NEYMAN-SCOTT RECTANGULAR PULSES RAINFALL MODEL

3.1 INTRODUCTION

This Chapter extends the Neyman-Scott Rectangular Pulses cluster model for simulating rainfall time series. Several important properties have previously been found for the model, for example, the expectation and variance of the amount of rain captured in an arbitrary time interval (*Rodriguez-Iturbe et al* (1987a)). In this Chapter, some further properties for the model are derived, e.g. the probability of an arbitrary interval of any chosen length being dry. In applications this is a desirable property to have, and is often used in fitting stochastic rainfall models to historical data.

As a preliminary investigation, the model is fitted to 10 years of hourly data taken from Blackpool, UK. The results indicate that the performance of the model is good, so that the model can be used with confidence for the remainder of the project.

3.2 DEFINITION OF THE MODEL

It is assumed that with any rainfall event there exists a generating mechanism, called the STORM ORIGIN, which may be passing fronts or some other criteria for convective storms, from which RAIN CELLS arise.

Furthermore it is assumed that:

(i) the storm origins arrive according to a Poisson process with rate parameter λ (per hour),

(ii) each storm origin generates a random number C of rain cells. To ensure at least one rain cell follows any given storm origin, C-1 will be distributed as a Poisson random variable, with ν as the mean number of cells per storm (i.e. $E(C-1) = Var(C-1) = \nu-1$).

(iii) the waiting time after the storm origin of each rain cell is exponentially distributed with parameter β (per hour),

(iv) the duration of each rain cell is exponentially distributed with parameter η (per hour),

(v) the intensity (in mm per hour) of each rain cell is constant throughout its duration and is exponentially distributed with parameter ξ (hour per mm).

(vi) the total intensity at any instant in time is the sum of the intensities due to all active cells at that instant.

(vii) the intensity, duration and the waiting time after the storm origin of any rain cell are independent of each other and other rain cells.

The parameters of the model can be summarised by:

$1/\lambda$ = mean time between storm origins,

$1/\beta$ = mean waiting time for cells after the storm origin,

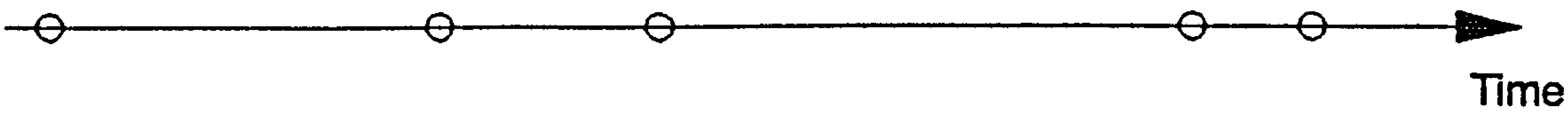
ν = mean number of rain cells per storm,

$1/\eta$ = mean cell duration and

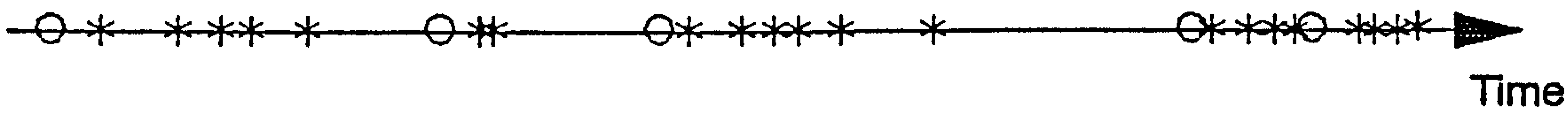
$1/\xi$ = mean cell intensity.

A schematic of the model is given in Figure 3.1.

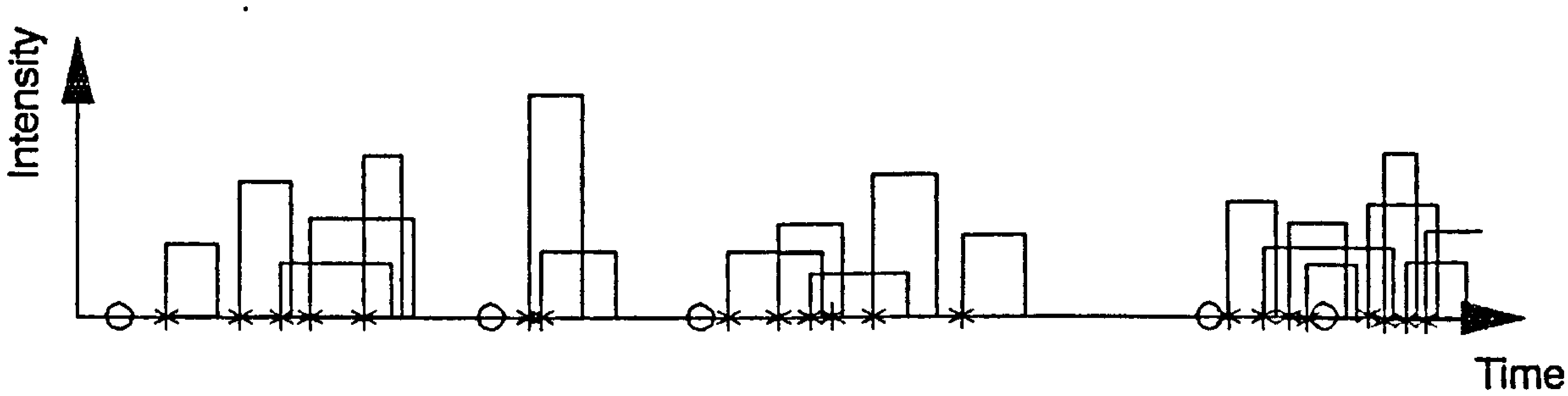
Storm origins arrive according to a Poisson Process



Each origin generates a random number of rain cells
with cell origins at *



The intensity and duration of each rain cell follow exponential
distributions - the intensity is constant throughout the duration



The total intensity at any point in time is the sum of the intensities
of all active rain cells at that point

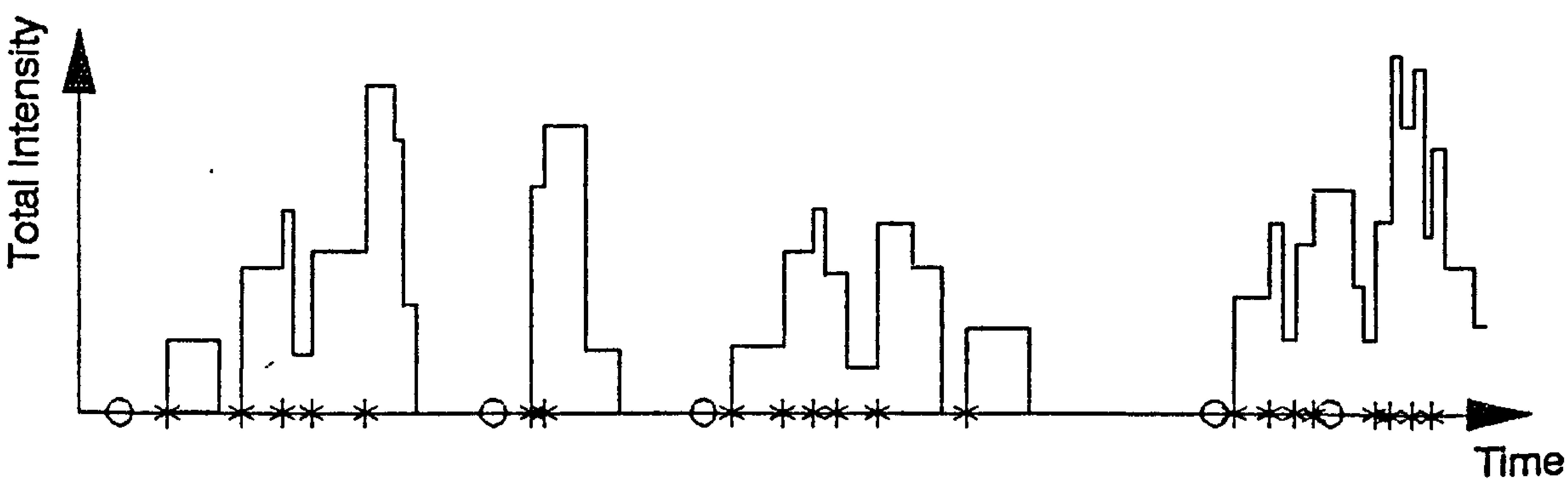


Figure 3.1 A Schematic of the Neyman-Scott Model

3.3 THE PROBABILITY THAT AN ARBITRARY INTERVAL IS DRY

Initially consider a single storm, ignoring the effects of any other storm (i.e. cells due to other storms). The storm origin will be taken at time zero and arbitrary time intervals [t,t+h] after the storm origin will be analysed (so t,h > 0).

$$\text{pr(k cell origins in [t,t+h] | C=n)} = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{if } k \le n,$$
$$= 0 \quad \text{if } k > n,$$

where p = p(t,t+h) = pr(a cell arrives in [t,t+h])

$$= \int_t^{t+h} \beta\, e^{-\beta t}\, dt = e^{-\beta t} - e^{-\beta(t+h)}$$

(N.B. p is a function of t and t+h - the brackets will sometimes be omitted for ease of notation)

Hence,

$$\text{pr(k cell origins in [t,t+h])} = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k}\, \text{pr(C=n)}$$
$$= \mu^{k-1} p^k (\mu - \mu p + k)\, e^{-\mu p}/k! \qquad (3.1)$$

by taking C-1 distributed Poisson with mean μ. (For convenience, ν - 1, in part (ii) of the model definition, is taken as μ).

An immediate consequence of (3.1) is:

pr(no cell origins in [t,t+h])

$$= (1 - e^{-\beta t} + e^{-\beta(t+h)}) \exp\{-\mu e^{-\beta t} + \mu e^{-\beta(t+h)}\} \qquad (3.2)$$

Let $F_W(x)$ denote the distribution function of the waiting time W after t for cells with origins in [t,t+h], i.e. let $F_W(x)$ be the probability that a cell has its origin in [t,t+x] given that the cell origin lies in [t,t+h]. Then $F_W(x)$ is given by:

$$F_W(x) = \frac{e^{-\beta t} - e^{-\beta(t+x)}}{e^{-\beta t} - e^{-\beta(t+h)}} \qquad \text{(for } x \le h) \qquad (3.3)$$

Now if D is the duration of any given cell, D is distributed $\exp(\eta)$, i.e. $f_D(x) = \eta e^{-\eta x} (x \ge 0)$. Hence the waiting time after t for the end of the rain cell is W+D, and thus the density of S = W+D is the convolution $f_W * f_D$, where $f_W = dF_W/dx$. The density $f_S(x)$ of S is:

$$f_S(x) = \frac{\beta\eta(e^{-\beta x} - e^{-\eta x})}{(\eta-\beta)(1-e^{-\beta h})} \qquad \text{if } 0 \le x \le h,$$

$$= \frac{\beta\eta(e^{(\eta-\beta)h} - 1)\, e^{-\eta x}}{(\eta-\beta)(1-e^{-\beta h})} \qquad \text{if } x > h \qquad (3.4)$$

From equation (3.4),

$$\text{pr}(S > t) = \int_t^{\infty} f_S(x)dx = \frac{\beta(e^{-\beta t} - e^{-\eta t})}{(\eta-\beta)(1-e^{-\beta t})}$$

and $\text{pr}(S \le t) = 1 - \text{pr}(S > t) = \alpha$, say.

Thus, pr(no rain in [t,t+h])

$$= \sum_{n=1}^{\infty} \sum_{k=0}^{n} \text{pr(k cell origins in (0,t) and all k cells terminate before t, and n-k cell origins in } (t+h,\infty) \mid C=n) \, \text{pr}(C=n)$$

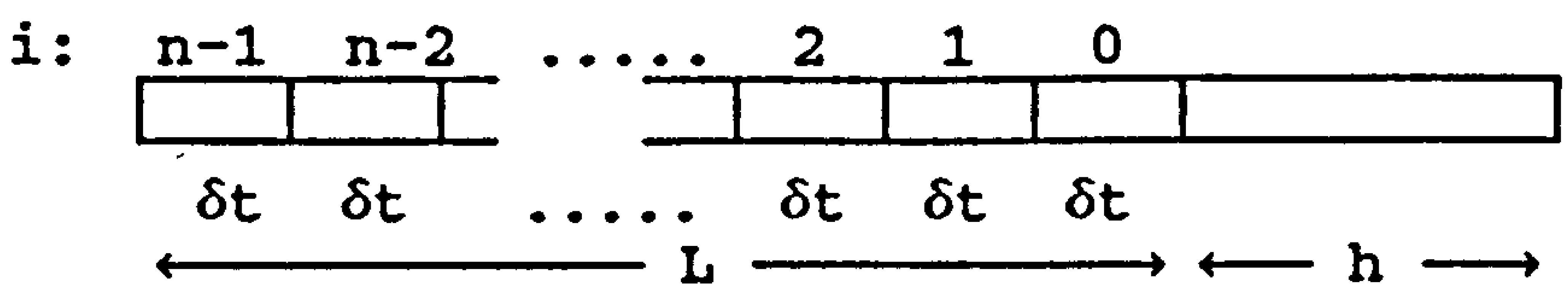
$$= \sum_{n=1}^{\infty} \sum_{k=0}^{n} \binom{n}{k} (1 - e^{-\beta t})^k \, e^{-\beta(t+h)(n-k)} \, \alpha^k \, \mu^{n-1} \, e^{-\mu} / (n-1)!$$

$$= \sum_{n=1}^{\infty} \left\{ \frac{\mu^{n-1}}{(n-1)!} \, e^{-\mu} \, e^{-\beta n(t+h)} \sum_{k=0}^{n} \binom{n}{k} \left(\alpha(1 - e^{-\beta t}) \, e^{\beta(t+h)} \right)^k \right\}$$

$$= e^{-\mu} \left(1 + \alpha(1 - e^{-\beta t}) e^{\beta(t+h)} \right) e^{-\beta(t+h)} \times \exp\left(\mu e^{-\beta(t+h)} + \mu\alpha(1 - e^{-\beta t}) \right)$$

$$= \left(e^{-\beta(t+h)} + 1 - (\eta e^{-\beta t} - \beta e^{-\eta t})/(\eta-\beta) \right) \times \exp\left(-\mu\beta(e^{-\beta t} - e^{-\eta t})/(\eta-\beta) - \mu e^{-\beta t} + \mu e^{-\beta(t+h)} \right) \qquad (3.6)$$

Now consider the general case. An arbitrary interval of length h is dry if it is dry due to storm origins in the interval and it is dry due to storm origins preceeding the interval. Consider a large interval of length L preceeding the interval of length h. Divide this large interval into n smaller subintervals of length δt:



(n.δt = L)

The probability that a storm origin arrives in an interval of length δt is $\lambda\delta t + o(\delta t)$. Interval h is dry, due to subinterval i of L, either if no storm origin is in i, with probability $1 - \lambda\delta t$

$+ o(\delta t)$, or a storm origin is in i, with probability $\lambda\delta t + o(\delta t)$, AND there is no rain in h due to the storm origin in interval i. The probability of no rain in h due to a storm origin in interval i is $p_{i\delta t}(h)$ (from (3.6)).

Hence the probability that h is dry due to storms in L is:

$$\lim_{\delta t \to 0} \Big((1 - \lambda\delta t + \lambda\delta t p_0(h)) \times (1 - \lambda\delta t + \lambda\delta t p_{\delta t}(h)) \times .. \\ ... \times (1 - \lambda\delta t + \lambda\delta t p_{(n-1)\delta t}(h)) \Big)$$

Therefore the probability that h is dry due to all storms preceeding h is given by:

$$\lim_{\delta t \to 0} \lim_{L \to \infty} \prod_{i=0}^{n-1} (1 - \lambda\delta t + \lambda\delta t p_{i\delta t}(h))$$

$$= \exp \left(- \lambda \int_0^\infty [1 - p_t(h)] \, dt \right) \tag{3.7}$$

Now divide the interval h into n smaller intervals of length δt. Interval h will be dry due to storms arriving in h with probability:

$$\lim_{\delta t \to 0} \Big((1 - \lambda\delta t + \lambda\delta t p_0((n-1)\delta t)) \times (1 - \lambda\delta t + \lambda\delta t p_0((n-2)\delta t)) \times .. \\ ... \times (1 - \lambda\delta t + \lambda\delta t p_0(\delta t)) \times (1 - \lambda\delta t + \lambda\delta t p_0(0)) \Big)$$

$$= \lim_{\delta t \to 0} \prod_{i=0}^{n-1} (1 - \lambda\delta t + \lambda\delta t p_0(i\delta t))$$

$$= \exp\left(- \lambda \int_0^h [1 - p_0(t)]\ dt \right)$$

$$= \exp\left(- \lambda \int_0^h [1 - e^{-\beta t} \exp\{-\mu + \mu e^{-\beta t}\}]\ dt \right)$$

$$= \exp\left(- \lambda h + \lambda(1 - \exp\{-\mu + \mu e^{-\beta h}\})/(\beta\mu) \right) \tag{3.8}$$

So the probability that an arbitrary interval of length h is dry, $\phi(h)$ say, is given by:

$$\phi(h) = \exp\left(-\lambda h + \frac{\lambda}{\beta\mu} (1 - \exp\{-\mu + \mu e^{-\beta h}\}) - \lambda \int_0^\infty [1 - p_t(h)]\ dt \right) \tag{3.9}$$

Hence $\phi(24)$ represents the proportion of dry days as predicted by the model. This is used to fit the model in Section 3.4.

Some further mathematical expressions for the Neyman-Scott Rectangular Pulses model (e.g. an approximation to $\phi(h)$ of equation (3.9)) were found and are given in Appendix A. These expressions may be of some use in future research. However, they were not needed in this project, and so the details of their development are omitted from this Chapter.

3.4 FITTING THE MODEL

3.4.1 Parameter estimation procedure

A detailed investigation into ways of estimating the parameters of the model will be given in the next Chapter. The purpose in this Chapter was to develop and test the model, and to decide on the suitability of the model for the remainder of the project. Consequently, only one parameter estimation procedure was considered, a procedure which seemed to make the most use of the available theoretical expressions.

The parameter estimation procedure used the following equations derived by *Rodriguez-Iturbe et al* (1987a), as well as expression (3.9) of the previous Section.

$$E(Y_i^{(h)}) = \lambda\mu_c h/(\eta\xi) \tag{3.10}$$

$$\mathrm{Var}(Y_i^{(h)}) = \frac{\lambda(\mu_c^2 - 1)[\beta^3 A_1(h) - \eta^3 B_1(h)]}{\beta\ \xi^2\ \eta^3\ (\beta^2 - \eta^2)} + \frac{4\ \lambda\ \mu_c\ A_1(h)}{\xi^2\ \eta^3} \tag{3.11}$$

$$\mathrm{Cov}(Y_i^{(h)}, Y_{i+k}^{(h)}) = \frac{\lambda(\mu_c^2 - 1)[\beta^3 A_2(h,k) - \eta^3 B_2(h,k)]}{\beta\ \xi^2\ \eta^3\ (\beta^2 - \eta^2)} + \frac{4\lambda\mu_c A_2(h,k)}{\xi^2\ \eta^3} \tag{3.12}$$

where:-

$Y_i^{(h)}$ = total rainfall in interval i of length h,

μ_c = E(C) = the mean no. of cells per storm,

$A_1(h) = \eta h - 1 + e^{-\eta h}$, $B_1(h) = \beta h - 1 + e^{-\beta h}$,

$A_2(h,k) = \frac{1}{2}(1 - e^{-\eta h})^2\ e^{-\eta h(k-1)}$, $B_2(h,k) = \frac{1}{2}(1 - e^{-\beta h})^2 e^{-\beta h(k-1)}$.

It will sometimes be found convenient to adopt the following notation:

$\mu(h) = E(Y_i^{(h)})$,
$\gamma(h) = Var(Y_i^{(h)})$,
$\rho(h,k) = Cov(Y_i^{(h)}, Y_{i+k}^{(h)})/Var(Y_i^{(h)})$ - the autocorrelation function,

The parameter estimation procedure was based on minimising a sum of squares, where the squared terms were differences between selected functions of the model parameters and the equivalent historical statistics taken from the rainfall data. With this kind of procedure more than 5 equations for the 5 unknown parameters can be selected, in an attempt to fit many more of the historical statistics. Ideally the selected sum of squares would give a minimum of zero, but in practice this puts too much demand on the model so that a value close to zero has to be accepted. The model functions selected for the minimisation procedure are given below, where the expected amount of rain captured in 1 hour as predicted by the model was substituted in equations (3.11) and (3.12), the purpose being to reduce the number of parameters requiring estimation within the procedure. The squared terms within the procedure were weighted to ensure that large valued historical statistics did not dominate the procedure.

Model functions used in minimisation procedure:
i) $\phi(24)$, ii) $\gamma(1)$, iii) $\gamma(6)$, iv) $\gamma(24)$, v) $\rho(1,1)$, vi) $\rho(6,1)$, and vii) $\rho(24,1)$.

The sum of squared differences between the above functions and their equivalent historical values were minimised using a quasi-Newton based algorithm (NAG routine EØ4JAF). The parameters were estimated for each month using 10 years of hourly data taken from Blackpool, UK.

To illustrate the order of magnitude of the parameter estimates, the values obtained for a summer and winter month are shown in Table 3.1 below, from which it can be seen that the parameter estimates seem physically realistic and are of a similar order of magnitude to those obtained by *Rodriguez-Iturbe et al* (1987b).

Table 3.1

Parameter Estimates for January and July

	λ	β	η	μ_c	ξ
	hr^{-1}	hr^{-1}	hr^{-1}	cells/storm	hr/mm
January	0.0149	0.0540	1.03	9.40	1.19
July	0.0136	0.0998	1.51	3.59	0.506

3.4.2 The Model's performance

For each month 10 years of hourly data were simulated using the model. The differences between historical and simulated statistics were calculated, standardised and plotted (see Figures 3.2 - 3.6). To interpret the plots note that each is given a label which refers to the statistic being tested. For each month this statistic is found for both the historical and simulated series, and the mean values calculated over the 10 year period. Lines are drawn on the plots to indicate the approximate location of the 5% significance level under a standard t-test. For example, in Figure 3.2, the label is monthly totals, so that the points plotted are the values of the t-statistics found by taking the difference in the mean monthly totals of the historical and simulated time series and dividing by an estimate of the standard error.

Most values lie within the bounds and vary about the zero line indicating that the model is performing well (Figures 3.2-3.6). The difference between the historical and simulated hourly autocorrelations for June is probably statistically significant. However, overall the model seems to be performing well and so this problem will not be considered here.

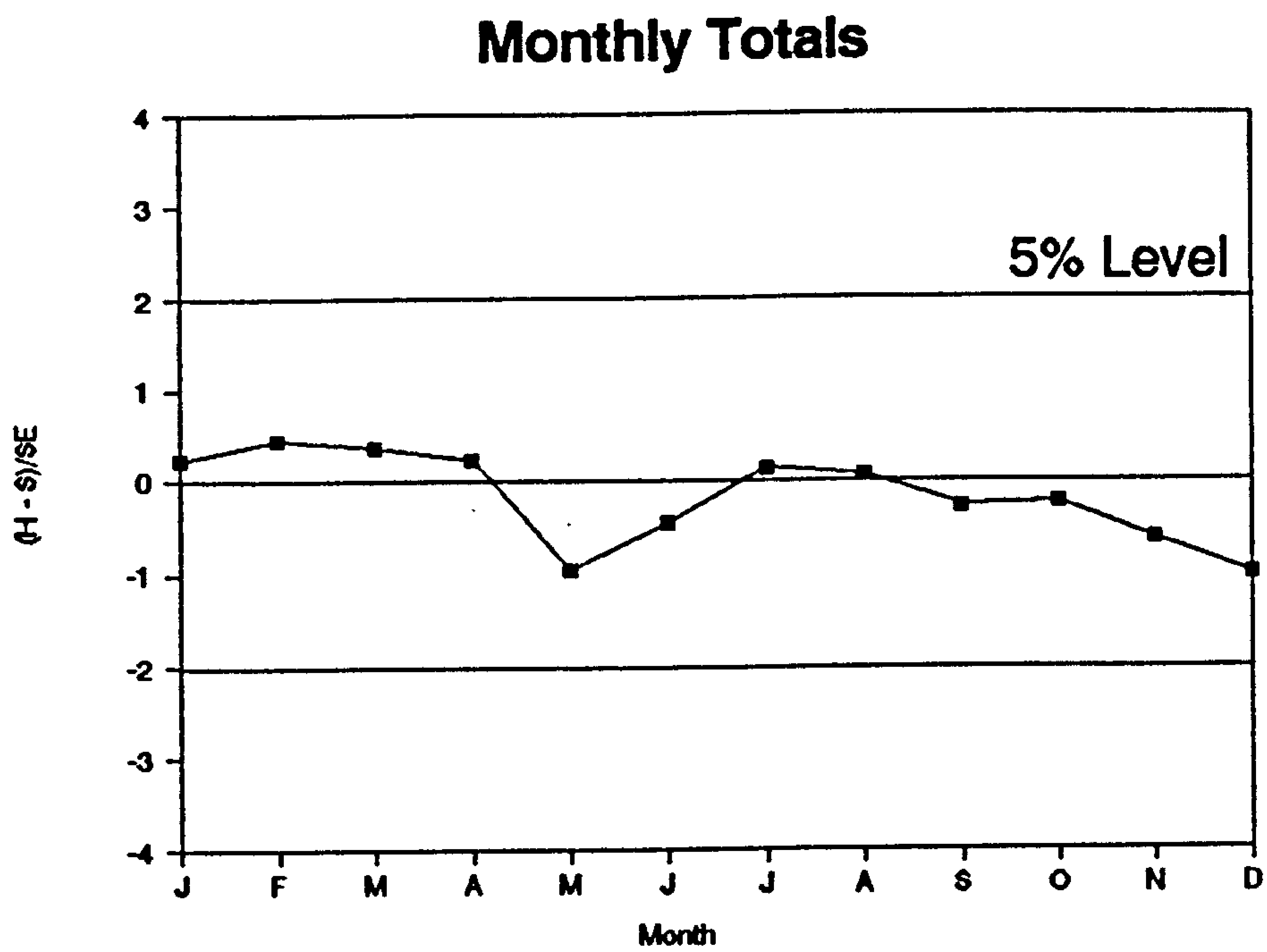


Figure 3.2

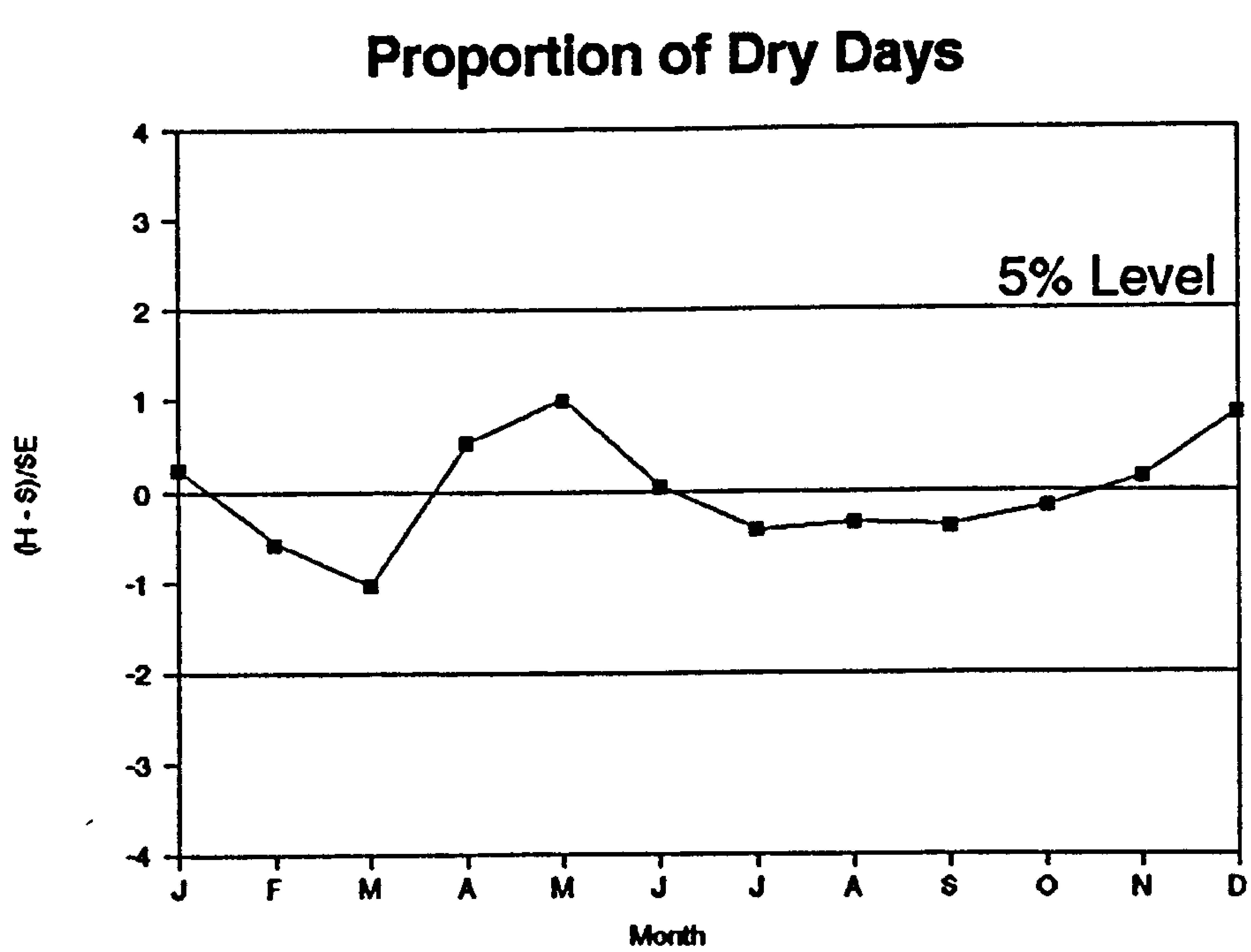


Figure 3.3

(H = Historical value, S = Simulated value,
SE = estimate of standard error)

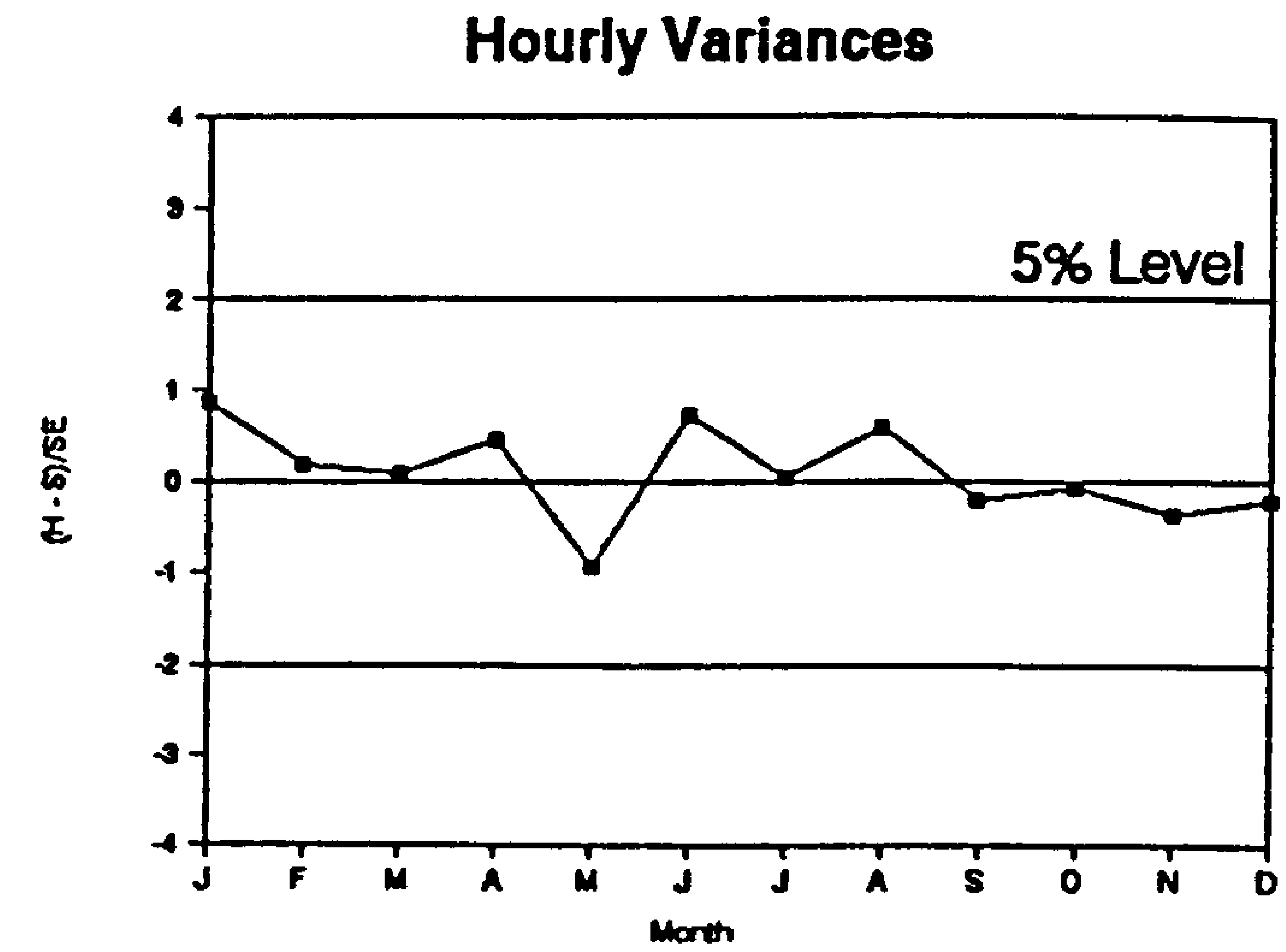


Figure 3.4(a)

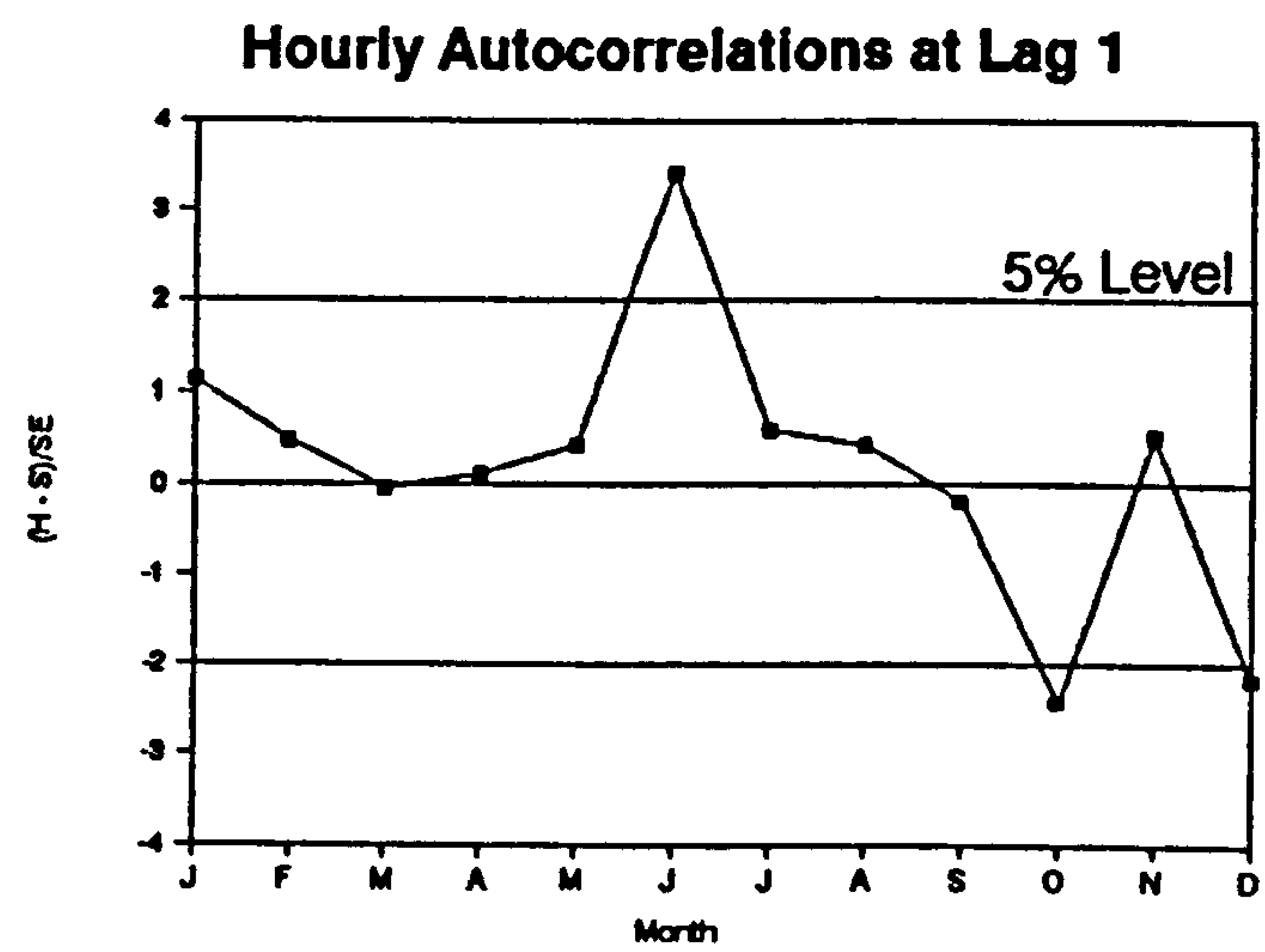


Figure 3.4(b)

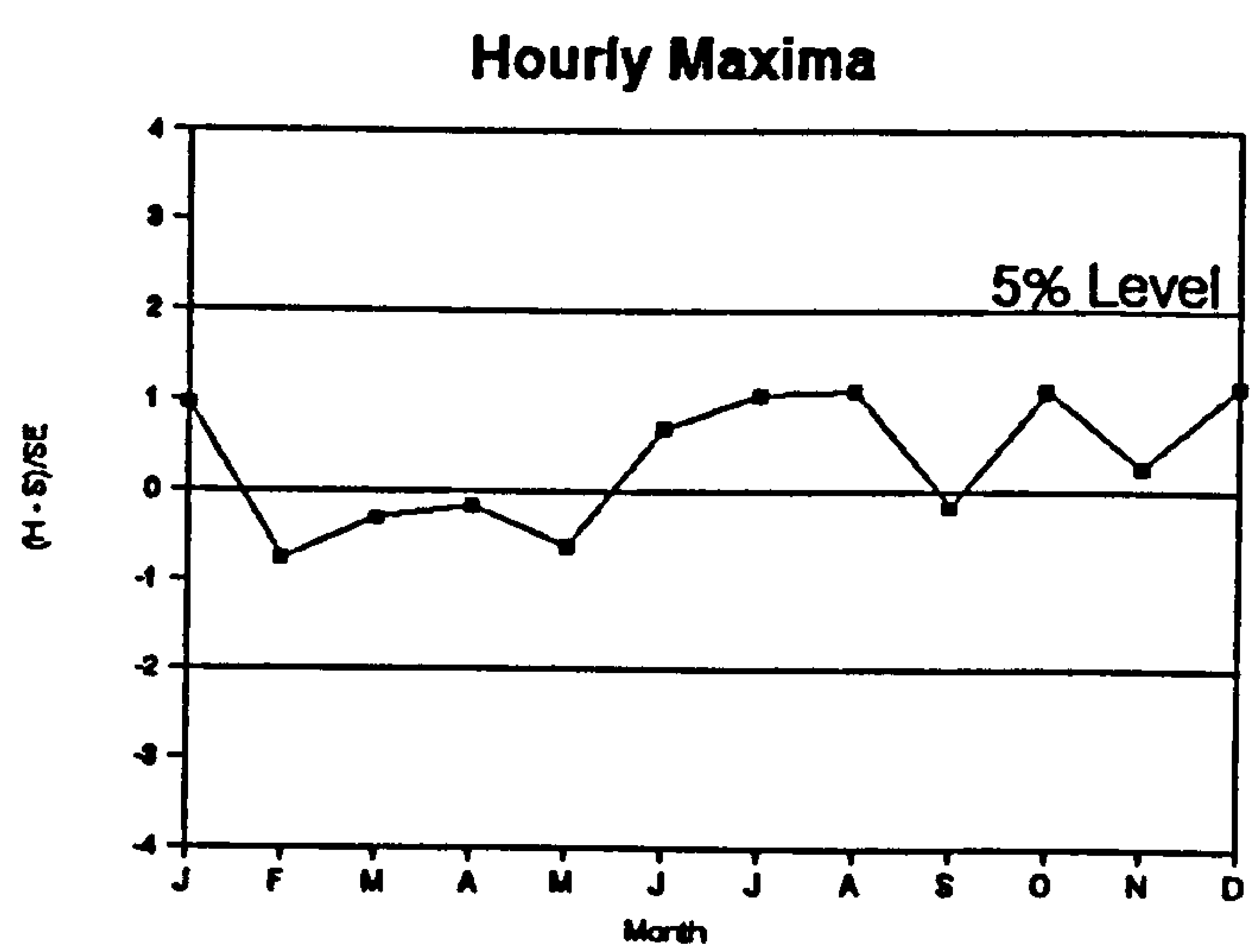


Figure 3.4(c)

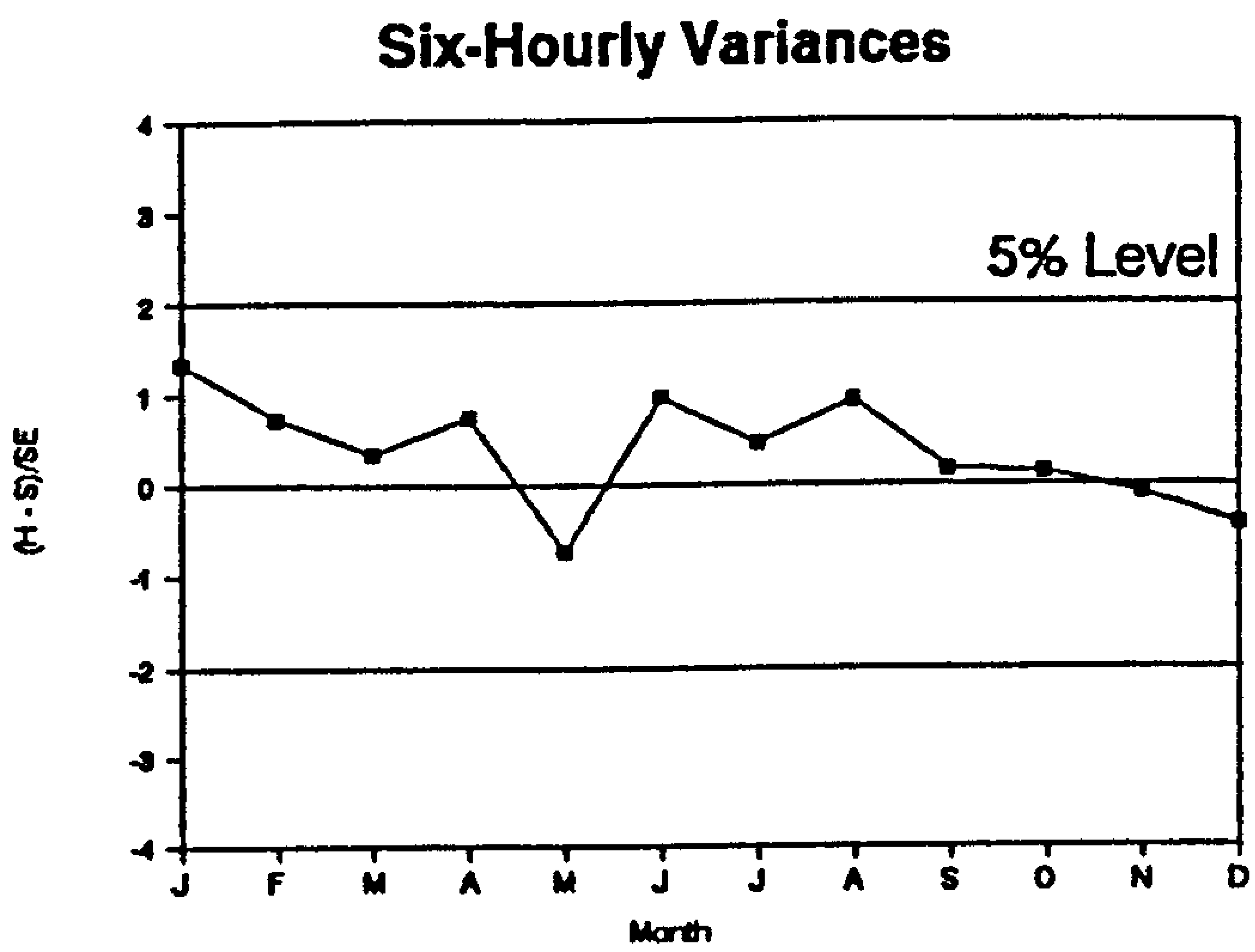


Figure 3.5(a)

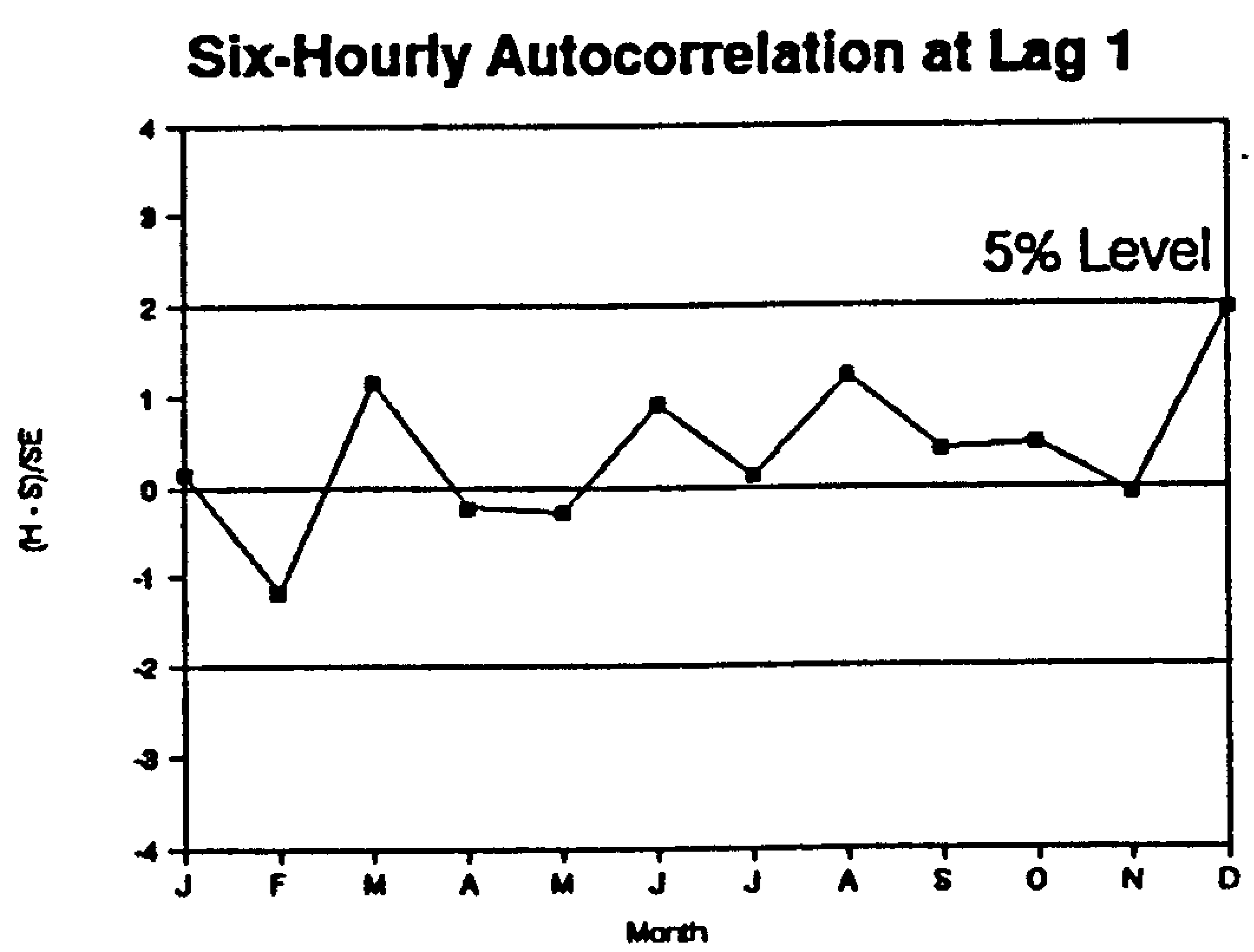


Figure 3.5(b)

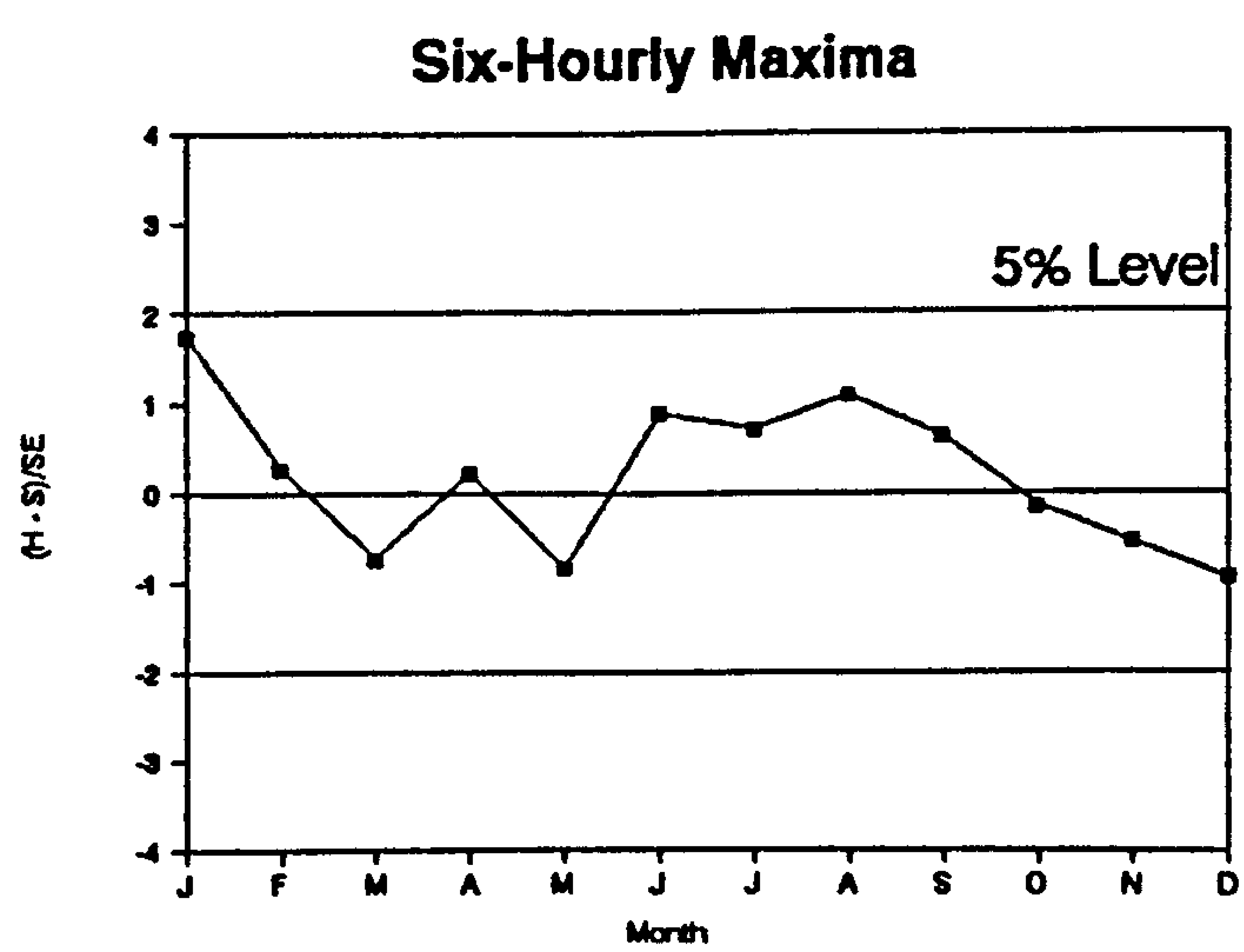


Figure 3.5(c)

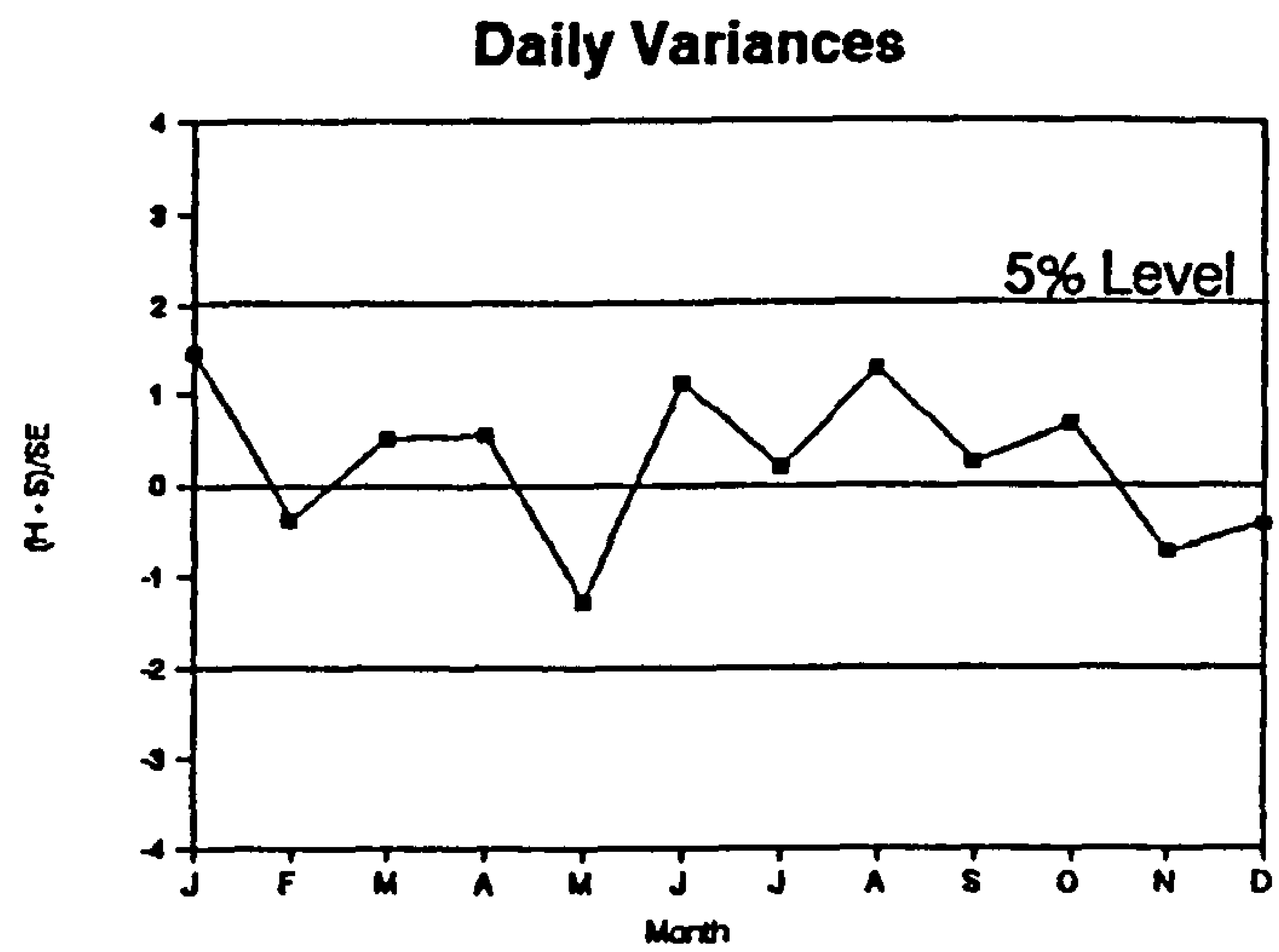


Figure 3.6(a)

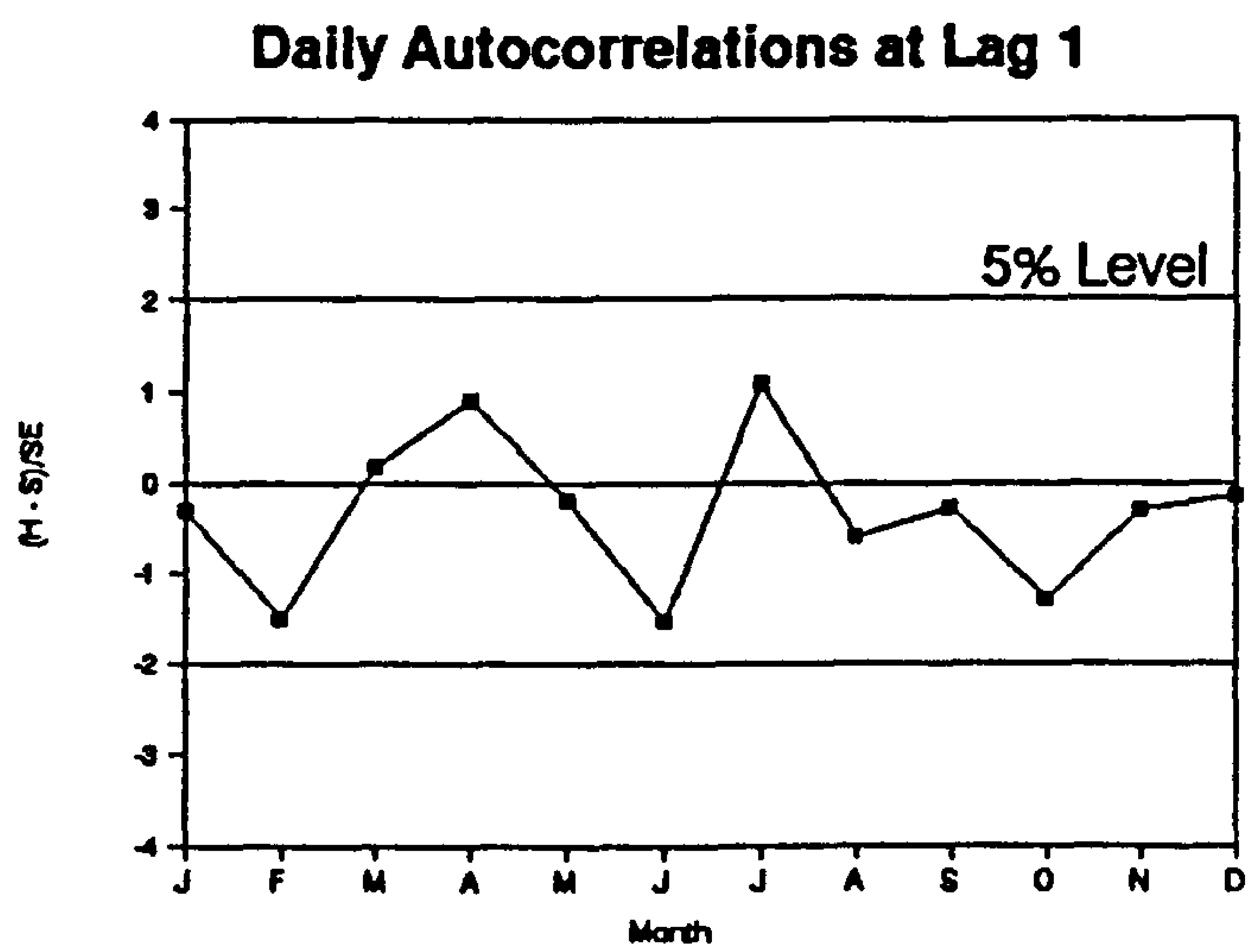


Figure 3.6(b)

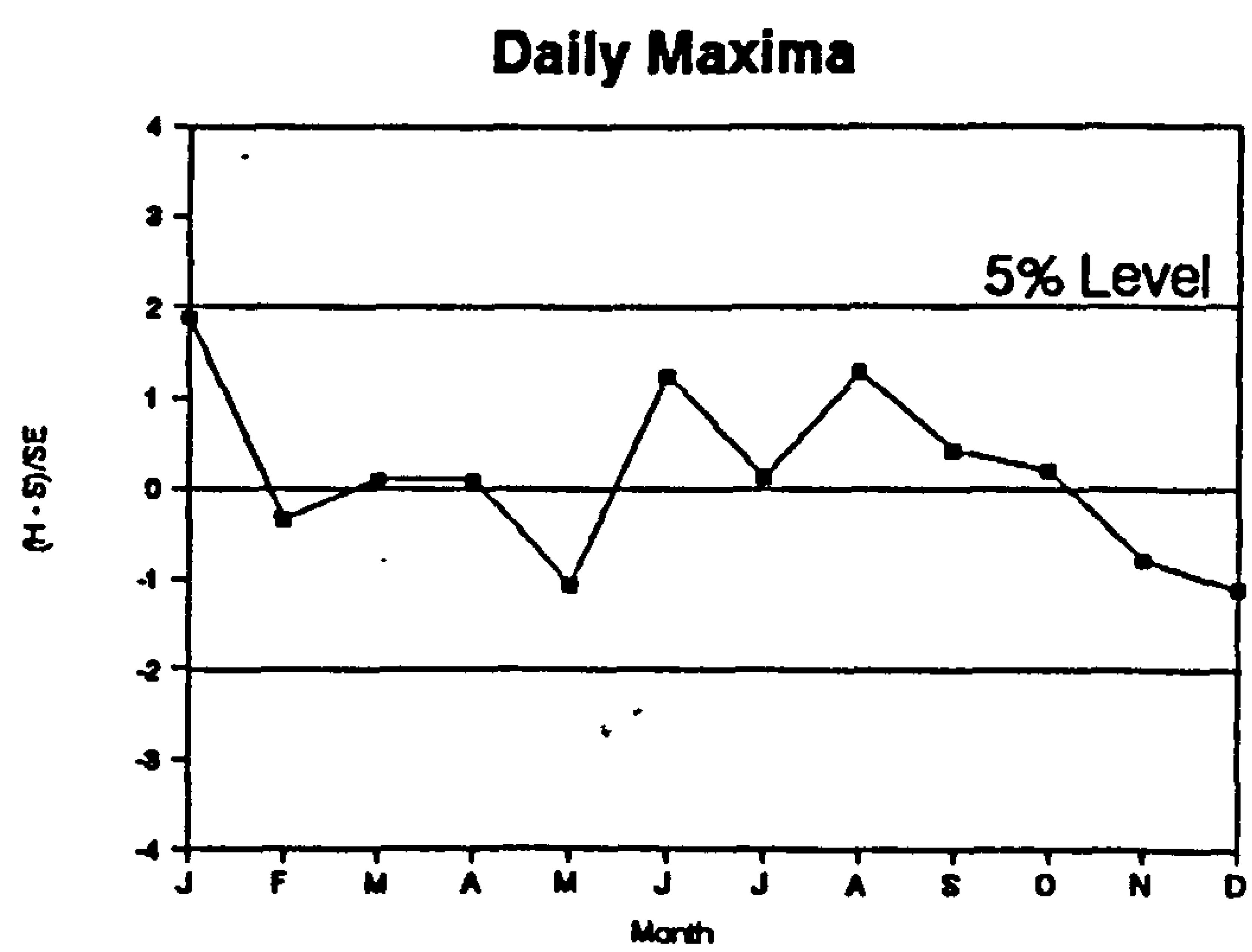


Figure 3.6(c)

CHAPTER 4

FITTING THE NEYMAN-SCOTT RECTANGULAR PULSES MODEL TO HOURLY RAINFALL TIME SERIES

4.1 INTRODUCTION

In the previous Chapter the first two moments, that is the mean, variance and autocovariance, of the aggregated Neyman-Scott (N-S) Rectangular Pulses model were given, so that a suitable method of parameter estimation could be based on the Method of Moments.

The Method of Moments involves equating observed moments, for example the sample mean, with their equivalent in the population, which are theoretical functions of the model parameters. With the N-S Rectangular Pulses Model there are 5 unknown parameters to be estimated from the rainfall data. Five model functions could be selected and equated to their equivalent historical statistics estimated from the rainfall data, and these equations solved as a set of simultaneous equations. Alternatively, a sum of squared terms could be minimised, where each term consists of a difference between the theoretical function of the model parameters and the equivalent historical sample statistic. The two methods should produce equivalent parameter estimates when the least squares method uses the same five model functions/historical statistics as the simultaneous equations. However, the least squares method has the advantage of being able to use more model functions/historical statistics to estimate the parameters, so that the model may be able to match more of the historical statistics. The number of historical statistics that the model can match is an indication of how good the model is. Naturally there will be a limit to this number, but by using the least squares method, historical and model statistics of importance can be made as close as possible, so that the parameter estimates under this method will give more than 5 model statistics that are almost equal to their historical equivalents, rather than 5 model statistics that are exactly equal to their historical equivalents (which would be the result of solving 5 simultaneous equations, assuming an exact solution existed).

In this Chapter a suitable way of fitting the Neyman-Scott Rectangular Pulses model to historical hourly data is sought. One approach to this problem would be to select historical statistics that are of practical importance and to use a sum of squares to measure how well the model fits the selected historical statistics when using different combinations of statistics/model functions in the fitting procedure. This approach is particularly suitable when all model functions of importance are known. However, a less formal approach was adopted in this Chapter because some comparisons (e.g. dry spell sequences) could only be made after simulation. Furthermore, initially, not all model functions of importance were known (e.g. the transition probabilities were derived after the first simulation study (Section 4.4.5)). To begin with, it was thought that the autocorrelation function could be used in the fitting procedure to ensure that the model captured the dependency inherent in historical rainfall events. However, this function was found to be inappropriate when modelling summer dry spell sequences, which needed to be modelled for the intended application, and so an alternative was sought by deriving transition probabilities for wet/dry spells. For convenience, these transition probabilities are derived in the next Section so that they can be included with the other model functions.

4.2 NOTATION AND MODEL FUNCTIONS

The following notation will be used for the functions of the model parameters:

$\mu(h)$ = mean of h hourly time series,

$\gamma(h)$ = variance of h hourly time series,

$\gamma(h,\tau)$ = lag τ autocovariance of h hourly time series,

$\rho(h,\tau)$ = lag τ autocorrelation of h hourly time series,

$\phi(h)$ = proportion of dry intervals for h hourly time series,

$\phi_{W|W}(h)$ = proportion of wet h hourly intervals with previous h hourly interval wet (called wet given wet transition probability),

$\phi_{D|D}(h)$ = proportion of dry h hourly intervals with previous h hourly interval dry (called dry given dry transition probability).

Expressions for $\mu(h)$, $\gamma(h)$ and $\gamma(h,\tau)$ in terms of the parameters of the Neyman-Scott Rectangular Pulses Model were derived by *Rodriguez-Iturbe et al* (1987a), and are given below:

$$\mu(h) = E(Y_i^{(h)}) = \lambda\mu_C\mu_X h/\eta \tag{4.1}$$

For $\tau = 0$,

$$\gamma(h) = \mathrm{Var}(Y_i^{(h)}) = \lambda\eta^{-3}(\eta h - 1 + e^{-\eta h})\{2\mu_C E(X^2) + E(C^2-C)\mu_X^2\beta^2/(\beta^2-\eta^2)\} - \lambda(\beta h - 1 + e^{-\beta h})E(C^2-C)\mu_X^2/\{\beta(\beta^2-\eta^2)\} \tag{4.2}$$

For $\tau \geq 1$,

$$\gamma(h,\tau) = \mathrm{Cov}(Y_i^{(h)},Y_{i+\tau}^{(h)}) = \lambda\eta^{-3}(1 - e^{-\eta h})^2 e^{-\eta(\tau-1)h}\{\mu_C E(X^2) + \tfrac{1}{2}E(C^2-C)\mu_X^2\beta^2/(\beta^2-\eta^2)\} - \lambda(1 - e^{-\beta h})^2 e^{-\beta(\tau-1)h}\,\tfrac{1}{2}E(C^2-C)\mu_X^2/\{\beta(\beta^2-\eta^2)\} \tag{4.3}$$

where $Y_i^{(h)}$ = total rainfall (in mm) in interval i of length h, X = cell intensity (in mm per hour), and C = the number of cells per storm. Following Chapter 3, C − 1 will be distributed as a Poisson random variable with $E(C) = \mu_c = \nu$ and $E(C^2-C) = \nu^2-1$ in equations (4.1)-(4.3) above. [Note also that $E(X) = \mu_x = 1/\xi$, and $E(X^2) = 2/\xi^2$ in the above equations when the cell intensity follows an Exponential distribution with parameter ξ].

An expression for $\phi(h)$ was derived in Chapter 3, and is given below:

$$\phi(h) = \exp\left(- \lambda h + \frac{\lambda}{\beta(\nu-1)} (1 - \exp\{1 - \nu + (\nu-1)e^{-\beta h}\}) - \lambda \int_0^\infty [1 - p_h(t)]dt \right) \quad (4.4)$$

where:

$$p_h(t) = \left(e^{-\beta(t+h)} + 1 - (\eta e^{-\beta t} - \beta e^{-\eta t})/(\eta-\beta) \right) \times \exp\left(-\mu\beta(e^{-\beta t} - e^{-\eta t})/(\eta-\beta) - \mu e^{-\beta t} + \mu e^{-\beta(t+h)} \right).$$

Expressions for the transition probabilities $\phi_{D|D}(h)$ and $\phi_{W|W}(h)$ have not previously been given. However, they follow immediately from equation (4.4) as the following shows.

Consider an interval of length 2h (in hours) denoted by [0,2h].

$\Pr\{[0,2h] \text{ dry}\} = \phi(2h)$

$\Pr\{[0,h] \text{ dry}\} = \Pr\{[h,2h] \text{ dry}\} = \phi(h)$

$$\Rightarrow \phi_{D|D}(h) = \Pr\{[h,2h] \text{ dry} \mid [0,h] \text{ dry}\} = \phi(2h)/\phi(h) \quad (4.5)$$

Now $\Pr\{[0,h] \text{ or } [h,2h] \text{ wet}\} = \Pr\{[0,2h] \text{ wet}\} = 1 - \phi(2h)$

Also $\Pr\{[0,h] \text{ or } [h,2h] \text{ wet}\}$

$= \Pr\{[0,h] \text{ wet}\} + \Pr\{[h,2h] \text{ wet}\} - \Pr\{[0,h] \text{ wet and } [h,2h] \text{ wet}\}$

$= 1 - \phi(h) + 1 - \phi(h) - \Pr\{[0,h] \text{ wet and } [h,2h] \text{ wet}\}$

$= 2 - 2\phi(h) - \Pr\{[0,h] \text{ wet and } [h,2h] \text{ wet}\}$

$$\Rightarrow \Pr\{[0,h] \text{ wet and } [h,2h] \text{ wet}\} = 2 - 2\phi(h) - 1 + \phi(2h)$$
$$= 1 - 2\phi(h) + \phi(2h)$$

$$\Rightarrow \phi_{W|W}(h) = \Pr\{[h,2h] \text{ wet} | [0,h] \text{ wet}\} = \frac{1 - 2\phi(h) + \phi(2h)}{1 - \phi(h)} \quad (4.6)$$

For an h hourly rainfall time series (either historical or simulated), the following notation will be used:

Mh = mean,
Vh = variance,
ACVh = lag 1 autocovariance,
ACh = lag 1 autocorrelation,
PDh = proportion of dry intervals,
WWh = proportion of wet intervals preceded by a wet interval,
DDh = proportion of dry intervals preceded by a dry interval.

For example, PD24 is the proportion of dry days found in the (historical or simulated) rainfall time series record, and M24 and V24 are the mean and variance respectively of the amount of rain captured in a day.

For the purpose of defining the parameter estimation procedure (Section 4.3) it is convenient to define the following two sets:

1) the set of model functions: $\mathbb{F} = \{\mu(1),\ \gamma(h),\ \rho(h,1),\ \phi_{W|W}(h),\ \phi_{D|D}(h),\ \phi(h) : h = 1, 3, 6, 12, 24\}$,

2) the set of rainfall time series statistics:
$\mathbb{K} = \{M1, Vh, ACh, WWh, DDh, PDh : h = 1, 3, 6, 12, 24\}$.

The following points may be noted concerning the above sets:
i) there is a one to one correspondence between $\mathbb{F}$ and $\mathbb{K}$ given by:
$\mu(1) \rightarrow M1$, $\gamma(h) \rightarrow Vh$, $\rho(h,1) \rightarrow ACh$, $\phi_{W|W}(h) \rightarrow WWh$, $\phi_{D|D}(h) \rightarrow DDh$,

$\phi(h) \rightarrow PDh$, i.e. each model function has an equivalent statistic, which can be found using historical rainfall time series data,

ii) $N(\mathbb{F}) = N(\mathbb{K}) = 26$,

iii) members of the above sets will be selected for the purpose of fitting and testing the model. Other model functions/statistics could be included as members of the sets. However, the members given above are a reasonable choice given that hourly and daily data are available and the model is required to match h hourly historical rainfall time series for h between 1 and 24 hours (subsets of the above were also used by *Rodriguez-Iturbe et al* (1987b) to fit the Neyman-Scott and Bartlett-Lewis Rectangular Pulses Models to historical rainfall data from Denver, Colorado).

4.3 THE PARAMETER ESTIMATION PROCEDURE

4.3.1 Definition

Let $f_i = f_i(\lambda,\beta,\eta,\nu,\xi) \in \mathbb{F}$ be a selected model function and let $\hat{k}_i \in \mathbb{K}$ be the equivalent statistic taken from the historical rainfall record. Note that $\hat{k}_i$ is an estimate of a population value k_i, i.e. it is assumed that the statistic $\hat{k}_i$ comes from a population in which the historical time series is one realisation. Different letters (f and k) are used to reflect the possibility that there may be some inadequacy in the model due to k_i lying in an infeasible region of f_i, i.e. there may be no solution to $f_i(\lambda,\beta,\eta,\nu,\xi) = k_i$.

Suppose m ($\leq$26) such functions and estimates are selected. Then the parameter estimation procedure is defined by:

Minimise: $$S = \sum_{i=1}^{m} (1 - f_i/\hat{k}_i)^2 \tag{4.7}$$

Subject to: $lb(\lambda) = 0 < \lambda < ub(\lambda)$, $lb(\beta) = 0 < \beta < ub(\beta)$,

$lb(\eta) = 0 < \eta < ub(\eta)$, $lb(\nu) = 1 < \nu < ub(\nu)$,

$lb(\xi) = 0 < \xi < ub(\xi)$,

where the lb(.) and ub(.) are lower and upper bounds respectively (these are required later in this Chapter), and $\hat{k}_i \neq 0$.

A suitable FORTRAN routine available for minimising S is provided by the Numerical Algorithms Group (NAG), and is called EØ4JAF. NAG provide several routines to find local minimums, but the routine selected is one of the few routines that enable the user to set upper and lower bounds on the parameters. As all the parameters for the N-S model must be positive it is essential to be able to put lower bounds on the parameters. The need for the upper bounds appears later in this Chapter.

Note that, in the parameter estimation procedure (4.7), a ratio between model function and historical statistic is used. This ensures that large historical statistics do not dominate the procedure. However, for the squared term containing the historical hourly mean, a weighting of 10 is applied to ensure that the historical hourly mean is almost matched exactly by the model, so that, on average, the volume of rain generated by the model is about the same as that in the historical data.

The purpose in this chapter is to find suitable functions $f_i \in \mathbb{F}$ and equivalent statistics $\hat{k}_i \in \mathbb{K}$ for the parameter estimation procedure (4.7), so that the parameter estimates generated by (4.7) can be used with the rainfall model to simulate hourly rainfall time series that match historical rainfall time series of any increment from 1 hour upwards.

4.3.2 Method of fitting the model and calculation of historical statistics

Initially the model will be fitted month by month, and statistics for each month used in the parameter estimation procedure, i.e. to begin with the model will have 60 parameters in total, 5 for each month. This number could possibly be reduced in later Chapters if, for example, it is possible to fix one or more parameters seasonally or throughout the year without detriment to the model's goodness of fit.

Suppose there are N years of historical rainfall time series data. To estimate historical moments of order 2 (e.g. V1 or AC1) for one of the months, the sample values from the rainfall data for the month will be used using the overall sample mean, i.e. the sample mean of all the h hourly values for the month in the N year period. To illustrate, suppose we wish to estimate the parameters of the model for January, using M1, V1, V24, AC1, AC24 and PD24 in (4.7). Then, in order to calculate V1 from the historical rainfall time series, the following would be used:

$$V1 = \sum_{i=1}^{N} \sum_{j=1}^{31\times24} (y_{ij} - \bar{y})^2 \ / \ (N \times 31 \times 24 - 1),$$

where

$$\bar{y} = \sum_{i=1}^{N} \sum_{j=1}^{31\times24} y_{ij} \ / \ (N \times 31 \times 24)$$

and y_{ij} = amount of rain captured in interval j, year i, where i = 1,...,N. With the autocorrelations a mean value over the total number of years will be used in order to avoid a carry over of the last hour in one January to the first hour in the next January of the following year. So in order to calculate the lag 1 autocovariance (ACV1) of the historical hourly rainfall time series, the following would be used:

$$ACV1 = \left[\sum_{i=1}^{N} \sum_{j=1}^{31\times24-1} (y_{ij} - \bar{y})(y_{ij+1} - \bar{y}) \;/\; (31 \times 24 - 1) \right] / N,$$

where again $\bar{y}$ is an estimate of the overall population mean.

4.3.3 Assessing the solution and testing the performance of the model

Clearly it would be desirable if the minimum of S in (4.7) was equal to zero. However, in practice, when more than 5 functions are used (m>5 in (4.7)), this puts too much demand on the model so that a value close to zero has to be accepted. The question remains as to what value of S is acceptable.

Let the solution to (4.7) be $(\hat{\lambda},\hat{\beta},\hat{\eta},\hat{\nu},\hat{\xi})$. Note that if S is much greater than zero then $|f_i(\hat{\lambda},\hat{\beta},\hat{\eta},\hat{\nu},\hat{\xi}) - \hat{k}_i| >> 0$ for some (or all) i = 1, ..., m. Now $|f_i(\hat{\lambda},\hat{\beta},\hat{\eta},\hat{\nu},\hat{\xi}) - \hat{k}_i|$ is the error in term i. Hence the absolute percentage error will be:

$$100 \times \frac{|f_i(\hat{\lambda},\hat{\beta},\hat{\eta},\hat{\nu},\hat{\xi}) - \hat{k}_i|}{\hat{k}_i}$$

Therefore after solving (4.7) the percentage error in each term can be found and a decision made as to whether the error is acceptable or not. This will depend on the sampling variability of the rainfall data, which is discussed at the end of this Section.

Using the above parameter estimation procedure the parameters can be estimated for the model at any given site for which data are available, and rainfall time series can then be generated for that site. Statistics can be extracted from both the historical and simulated time series and compared. The statistics to be used to assess the performance of the model will be: (i) the monthly totals T, (ii) the variance of 1, 6, 12, and 24 hourly time series (V1, V6, V12, V24), (iii) lag 1 autocorrelations of 1, 6, 12, and 24 hourly time series (AC1, AC6, AC12, AC24), (iv) the maximum amount of rain per month per year for 1, 6, 12, and 24 hourly time increments, and (v) the proportion of dry days PD24. This choice of statistics is frequently used in the literature (see for example *Rodriguez-Iturbe et al* (1987b)). To some extent this choice is subjective, but on the other hand it is difficult to see why other time series increments would provide a better base on which to assess the performance of the model. In addition frequency plots of dry spell durations (in days) will be used to compare the historical and simulated time series.

For each year the statistics will be extracted from each month for both the historical and simulated time series. Assuming that monthly historical statistics are independent from one year to the next, t-tests will be used to test whether the simulated statistics are significantly different from the historical values.

To illustrate, suppose the statistic under comparison is monthly total T. The monthly totals will be found for each year for all months for both the historical and simulated rainfall time series. Let $T_{ij}^{(h)}$ and $T_{ij}^{(s)}$ (i = 1, ..., N = total years; j = 1, ..., 12) denote the monthly total for month j year i for the historical and simulated time series respectively. Let $\bar{T}_j^{(h)}$ and $\bar{T}_j^{(s)}$ be the mean monthly totals over the N year period, and let S_j^2 be a pooled estimate of the variance. Then

$$t_j = \frac{\bar{T}_j^{(h)} - \bar{T}_j^{(s)}}{S_j \sqrt{2 / N}} \qquad (j = 1, \ldots, 12)$$

follows a t-distribution with 2N - 2 degrees of freedom. Strictly speaking using the pooled variance S_j^2 assumes that $Var(T_{ij}^{(h)}) \approx Var(T_{ij}^{(s)})$. However, initially it will be found convenient to ignore the possibility of unequal variances and use t_j to compare the simulated statistics with their equivalent historical statistics, without reading too much into significance levels.

4.3.4 The sampling variability to be expected in rainfall data

To obtain an estimate of the sampling variability to be expected in a rainfall record, the longest records (Poaka Beck (90 years), Exmouth (70 years), Windsor (90 years), Blackbrook (90 years) and Howick Hall (90 years)) from each of the 'Wigley' regions of Chapter 2 (Figure 2.1) were divided into 10 year periods. For each station-month (a total of 12×5) in each period some key daily statistics (daily means, daily variances, proportion of dry days, wet and given wet transition probabilities) were found, and the mean and standard deviations of the key statistics (over the periods) evaluated for each station-month. For example, for each of the 9 periods of 10 years of data for Windsor-January the proportion of dry days were found, and the mean and standard deviation of these 9 estimates of the proportion of dry days also found. The standard deviations were divided by the means to obtain the coefficient of variation (CV) for each key statistic for each station-month. For each key statistic the mean CV (over all the station-months) was found, and these values are given in Table 4.1 below, together with an estimate of the mean CV for a 20 year record (= mean CV for 10 year record ÷ $\sqrt{2}$).

Table 4.1
Estimates of the coefficient of variation
for each key daily statistic for 10 and 20 year records

Key statistic	Mean CV (for a 10 year record)	SE of Mean (for a 10 year record)	Estimate of Mean CV for a 20 year record
M24	16%	0.5	11%
V24	31%	1.	22%
PD24	12%	0.5	8.5%
WW24	8.2%	0.3	5.8%

Table 4.1 gives a rough guide to the error that will be acceptable when fitting the model (i.e. whether the percentage error of Section 4.3.3 is acceptable). For example, with a 20 year record of rainfall data we should require the percentage error between the historical daily variance and the daily variance predicted by the model to be less than 22% for most of the parameter sets generated by (4.7) (about 70% of the parameter sets obtained from the fitting procedure (4.7) will be required to predict a daily variance (which is a function of the model parameters) that is within 22% of the historical daily variance that was used in the fitting procedure).

4.4 ASSESSING THE DISTRIBUTION FOR THE CELL INTENSITY

4.4.1 Using no specified distribution for the cell intensity

It can be seen in equations (4.1) - (4.3) that the cell intensity X has no specified distribution. We are therefore free to choose a distribution for X. A distribution frequently used for rainfall intensity is the Exponential distribution. This distribution has the advantage over other distributions of only having a single parameter ξ, in which case $\mu_X = 1/\xi$ and $E(X^2) = 2/\xi^2$ in equations (4.1) - (4.3).

Before using the Exponential distribution it is worth considering how well the model fits with no specified distribution for the cell intensity, i.e. treating μ_X and $E(X^2)$ as parameters in the fitting procedure (4.7). If the result of not specifying a distribution showed a much better fit than using an Exponential distribution, other distributions (e.g. Gamma or Weibull) could then be selected and tested against each other and the distribution giving the best fit used to model the cell intensity.

4.4.2 The statistics extracted from the Manston data set

The Manston (in Kent) data set is suitable for carrying out an initial investigation on fitting the model as it is a long (20 year) hourly record with no missing values.

A program was written to break this data set into 12 files corresponding to months. A further program was written to extract the statistics to be used in fitting the model from the monthly files. The results of running this program (i.e. the statistics to be used in fitting) are given in Table 4.2.

Table 4.2

Estimates of statistics to be used to fit the model

(taken from the Manston data set)

Month	M1 (mm)	V1 (mm^2)	AC1	V6 (mm^2)	AC6	V24 (mm^2)	AC24	PD24
Jan	0.063	0.082	0.54	1.2	0.23	7.5	0.08	0.52
Feb	0.052	0.068	0.49	1.0	0.33	6.5	0.25	0.58
Mar	0.054	0.071	0.54	1.2	0.27	7.4	0.19	0.57
Apr	0.055	0.082	0.55	1.4	0.27	7.9	0.13	0.60
May	0.056	0.120	0.43	1.7	0.19	8.1	0.11	0.60
Jun	0.062	0.174	0.39	2.3	0.30	14.	0.27	0.68
Jul	0.061	0.204	0.46	3.1	0.34	21.	0.09	0.70
Aug	0.067	0.386	0.33	4.5	0.32	36.	0.07	0.71
Sep	0.093	0.530	0.47	9.1	0.51	61.	0.25	0.64
Oct	0.080	0.208	0.46	3.1	0.37	21.	0.32	0.62
Nov	0.088	0.153	0.55	2.4	0.34	17.	0.15	0.54
Dec	0.066	0.097	0.54	1.6	0.27	9.5	0.07	0.56

4.4.3 The results of using no specified distribution for the cell intensity

The results of using the statistics in Table 4.2 in the parameter estimation procedure are given in Tables 4.3a and 4.3b. Table 4.3a gives the parameter estimates obtained, and Table 4.3b gives the absolute (i.e. positive valued) percentage error in using these parameters in the model equations. The means (over months and over statistics) are also given in Table 4.3b as they give an indication of where the model is failing to match the historical statistics. For example, looking at the means in the right hand column it can be seen that the mean value for August (10%) is greater than all the other mean values in that column which indicates that the poorest fit is in August, or more generally in the summer. Similarly, the worst fit in the historical statistics is indicated by looking at the bottom row of means (in this case the variance of the 6 hourly time series).

Table 4.3

The results of not fixing a distribution for the cell intensity

(a)
The parameter estimates (generated using (4.7))

Month	λ hour^{-1}	β hour^{-1}	η hour^{-1}	ν	E(X) mm/hour	Var(X) (mm/hour)^2
Jan	0.0230	0.182	1.25	2.97	1.14	0.568
Feb	0.0078	0.066	1.59	27.4	0.39	0.530
Mar	0.0104	0.075	1.15	9.47	0.63	0.671
Apr	0.0120	0.127	1.22	10.7	0.53	0.680
May	0.0149	0.107	1.64	3.26	1.87	2.69
Jun	0.0050	0.055	2.13	19.6	1.37	4.23
Jul	0.0047	0.219	3.17	583.	0.07	0.33
Aug	0.0038	0.288	99.6	200.	8.66	0.665
Sep	0.0025	0.086	2.22	0.13	0.052	0.389
Oct	0.0047	0.054	1.85	73.4	0.427	1.37
Nov	0.0112	0.131	1.55	74.1	0.164	0.306
Dec	0.0133	0.285	1.98	323.	0.030	0.048

(b)
Percentage errors (between model and historical statistics)

Month	M1	V1	AC1	V6	AC6	V24	AC24	PD24	mean
Jan	0.0	2.8	1.9	3.0	2.1	6.4	0.3	0.0	2.1
Feb	0.0	0.8	0.6	6.0	1.8	4.7	0.5	0.0	1.8
Mar	0.0	4.6	3.5	3.7	1.3	1.3	0.0	0.0	1.8
Apr	0.0	2.3	1.7	4.6	0.1	2.0	0.3	0.0	1.4
May	0.0	3.8	2.2	8.6	0.4	3.8	0.3	0.0	2.4
Jun	0.0	0.2	0.2	5.9	2.0	5.0	0.5	0.0	1.7
Jul	0.0	0.8	0.4	0.2	5.5	5.0	3.9	13.	3.6
Aug	0.1	1.8	1.5	12.	16.	19.	11.	22.	10.
Sep	0.1	1.7	0.8	12.	17.	3.6	2.1	22.	7.3
Oct	0.0	0.7	0.3	5.3	2.7	5.4	0.6	0.0	1.9
Nov	0.0	0.9	0.6	2.3	1.5	3.4	0.3	0.0	1.1
Dec	0.0	2.5	2.1	1.7	3.1	3.2	2.6	2.9	2.3
Mean	0.0	1.9	1.3	5.4	4.5	5.2	1.9	4.9	3.2

4.4.4 Using an Exponential distribution for the cell intensity

Tables 4.4a and 4.4b give the results of using the statistics in Table 4.2 in the parameter estimation procedure, where now the functions (equations (4.1) - (4.3)) use an Exponential distribution for the cell intensity X.

It can be seen in Table 4.4a that some of the parameter estimates (ξ and η for August, and β for November) are much different in magnitude than the same parameter estimates for other months. This may cause problems when trying to seasonalise the model, so, to overcome this, bounds were placed on β, η, and ξ and the parameters re-estimated, the results being given in Tables 4.5a and 4.5b.

It can be seen in Table 4.5a that the parameter estimates seem physically realistic, and exhibit some seasonal variation (e.g. in λ - the rate of storm arrival). With the possible exception of the proportion of dry days, the improvement in fit obtained by using no specified distribution on cell intensity (Table 4.3) does not seem to out weigh the advantage of using the Exponential distribution (Table 4.5), which uses only 1 parameter. Hence the Exponential distribution will be used to model the cell intensity.

Table 4.4

The results for an Exponential distribution
for the cell intensity
(with no upper bounds in the fitting procedure (4.7))

(a)
The parameter estimates

Month	λ hour^{-1}	β hour^{-1}	η hour^{-1}	ν	$1/\xi$ mm / hour
Jan	0.0218	0.191	1.30	4.55	0.82
Feb	0.0080	0.063	1.55	11.4	0.87
Mar	0.0106	0.073	1.13	6.82	0.84
Apr	0.0125	0.122	1.17	5.76	0.91
May	0.0147	0.108	1.65	3.75	1.66
Jun	0.0050	0.055	2.12	11.9	2.23
Jul	0.0047	0.225	3.51	19.0	2.42
Aug	0.0038	0.290	463.	20.6	392.
Sep	0.0024	0.091	2.67	26.7	3.86
Oct	0.0048	0.051	1.80	16.5	1.81
Nov	0.0071	0.007	0.49	6.71	0.90
Dec	0.0164	0.248	1.41	6.81	0.83

(b)
Absolute percentage errors

Month	M1	V1	AC1	V6	AC6	V24	AC24	PD24	mean
Jan	0.0	3.2	1.1	3.3	4.2	5.6	0.4	0.4	2.5
Feb	0.1	0.1	0.9	7.3	5.0	1.9	0.1	8.2	3.0
Mar	0.0	4.1	4.0	4.2	0.4	2.3	0.2	3.3	2.3
Apr	0.0	1.8	2.5	5.0	2.5	0.8	0.2	3.9	2.1
May	0.0	4.0	2.0	8.3	0.4	4.2	0.4	1.6	2.6
Jun	0.0	0.0	0.2	6.3	3.0	4.1	0.3	4.3	2.3
Jul	0.1	0.4	0.2	0.0	4.4	4.9	2.7	22.	4.4
Aug	0.1	2.2	2.0	12.	16.	19.	11.	26.	11.
Sep	0.2	4.1	1.6	12.	12.	5.7	0.6	39.	9.4
Oct	0.1	1.4	0.2	6.9	5.8	2.3	0.1	12.	3.6
Nov	0.0	8.9	34.	13.	23.	7.1	3.8	24.	14.
Dec	0.1	2.8	5.3	3.0	9.9	6.0	4.5	9.4	5.1
Mean	0.1	2.7	4.5	6.8	7.3	5.4	2.0	13.	5.2

Table 4.5

The results for an Exponential distribution for cell intensity (with lb(β) = 0.05, ub(η) = ub(ξ) = 4.0 in (4.7))

(a)

The parameter estimates

Month	λ hour^{-1}	β hour^{-1}	η hour^{-1}	ν	$1/\xi$ mm / hour
Jan	0.0218	0.191	1.30	4.55	0.82
Feb	0.0080	0.063	1.55	11.4	0.87
Mar	0.0106	0.073	1.13	6.82	0.84
Apr	0.0125	0.122	1.17	5.76	0.91
May	0.0147	0.108	1.65	3.75	1.66
Jun	0.0050	0.055	2.12	11.9	2.23
Jul	0.0047	0.225	3.51	19.0	2.42
Aug	0.0047	0.264	3.22	11.5	4.00
Sep	0.0024	0.091	2.67	26.7	3.86
Oct	0.0048	0.051	1.80	16.5	1.81
Nov	0.0124	0.121	1.38	9.62	1.02
Dec	0.0164	0.248	1.41	6.81	0.83

(b)

Absolute Percentage Errors

Month	M1	V1	AC1	V6	AC6	V24	AC24	PD24	mean
Jan	0.0	3.2	1.1	3.3	4.2	5.6	0.4	0.4	2.5
Feb	0.1	0.1	0.9	7.3	5.0	1.9	0.1	8.2	3.0
Mar	0.0	4.1	4.0	4.2	0.4	2.3	0.2	3.3	2.3
Apr	0.0	1.8	2.5	5.0	2.5	0.8	0.2	3.9	2.1
May	0.0	4.0	2.0	8.3	0.4	4.2	0.4	1.6	2.6
Jun	0.0	0.0	0.2	6.3	3.0	4.1	0.3	4.3	2.3
Jul	0.1	0.4	0.2	0.0	4.4	4.9	2.7	22.	4.4
Aug	0.3	11.	27.	4.9	21.	25.	12.	22.	16.
Sep	0.2	4.1	1.6	12.	12.	5.7	0.6	39.	9.4
Oct	0.1	1.4	0.2	6.9	5.8	2.3	0.1	12.	3.6
Nov	0.1	0.1	1.8	0.0	5.2	7.8	1.9	11.	3.5
Dec	0.1	2.8	5.3	3.0	9.9	6.0	4.5	9.4	5.1
Mean	0.1	2.7	3.9	5.1	6.2	5.9	1.9	12.	4.7

4.4.5 Simulating rainfall time series using an Exponential distribution for the cell intensity

A simulation program (see Appendix B) was written to simulate hourly rainfall time series using the Neyman-Scott Rectangular Pulses model. Using the parameter estimates in Table 4.5a the simulation program was used to generate 20 years of hourly data. A further program was written to enable the historical and simulated time series to be compared and tested using the t-tests described in Section 4.3.3. For each statistic under comparison, the t-ratios were plotted against the month (see Figures 4.1 - 4.11 for some selected examples and Appendix C for all the plots).

To compare dry spell sequences another program was written and the results of running this program for the months of July, August and September are given in Figure 4.12 (the results for the other months are given in Appendix C). The frequency given in the plots is obtained by counting over a fixed period in time (in this case 20 years) instead of a standardised number of dry days. The comparison is made in this way because the engineer is interested in the return period of, say, a dry spell of over 25 days, when considerable bacteria will have built up within a sewage system. Furthermore, lower bounds (lb) are also used to define 'dry' days, i.e. a day is said to be dry if less than lb millimetres of rain fell (these bounds ensure that small traces of rainfall, sometimes found in historical records, count as dry days).

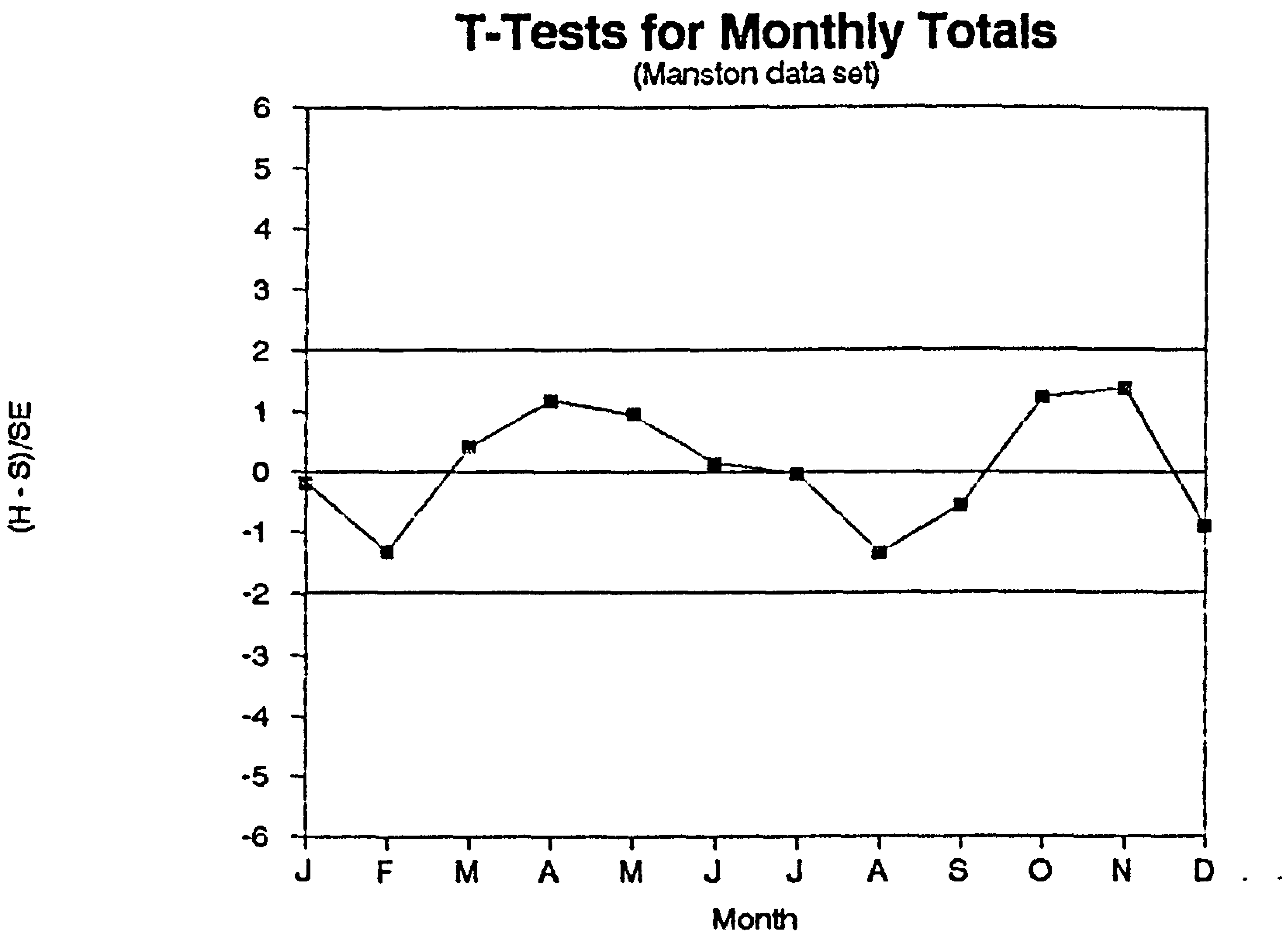


Figure 4.1

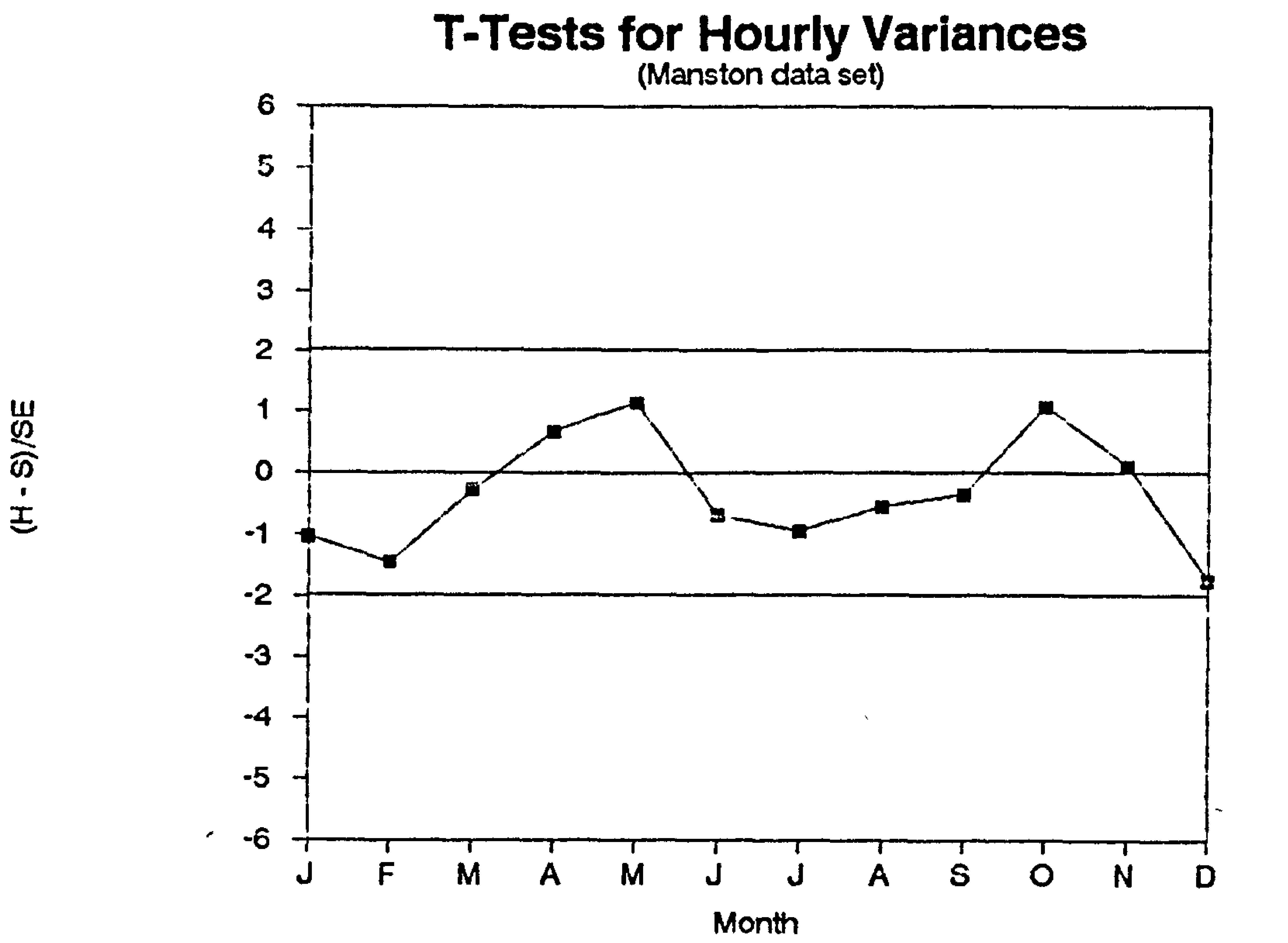


Figure 4.2

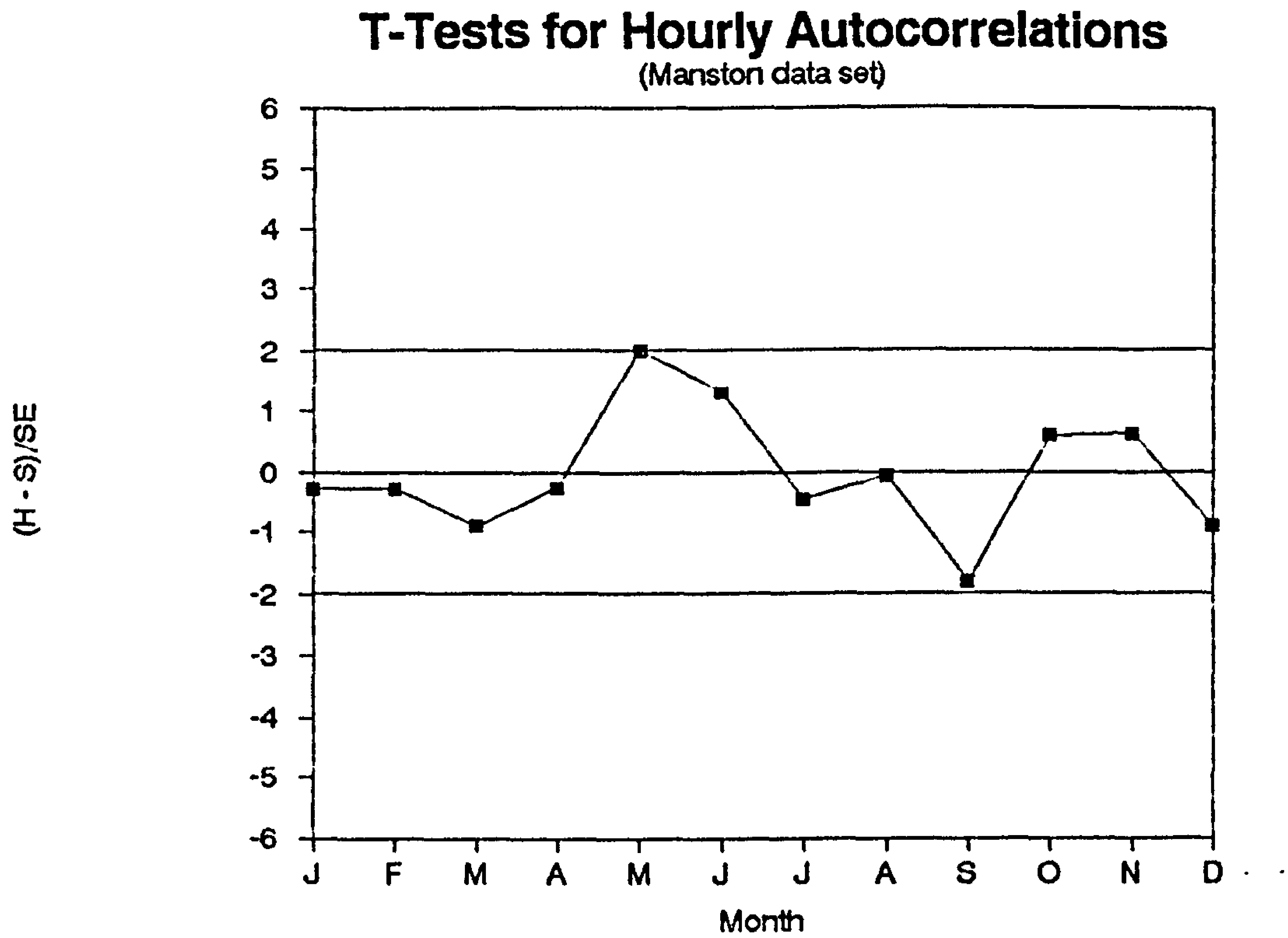


Figure 4.3

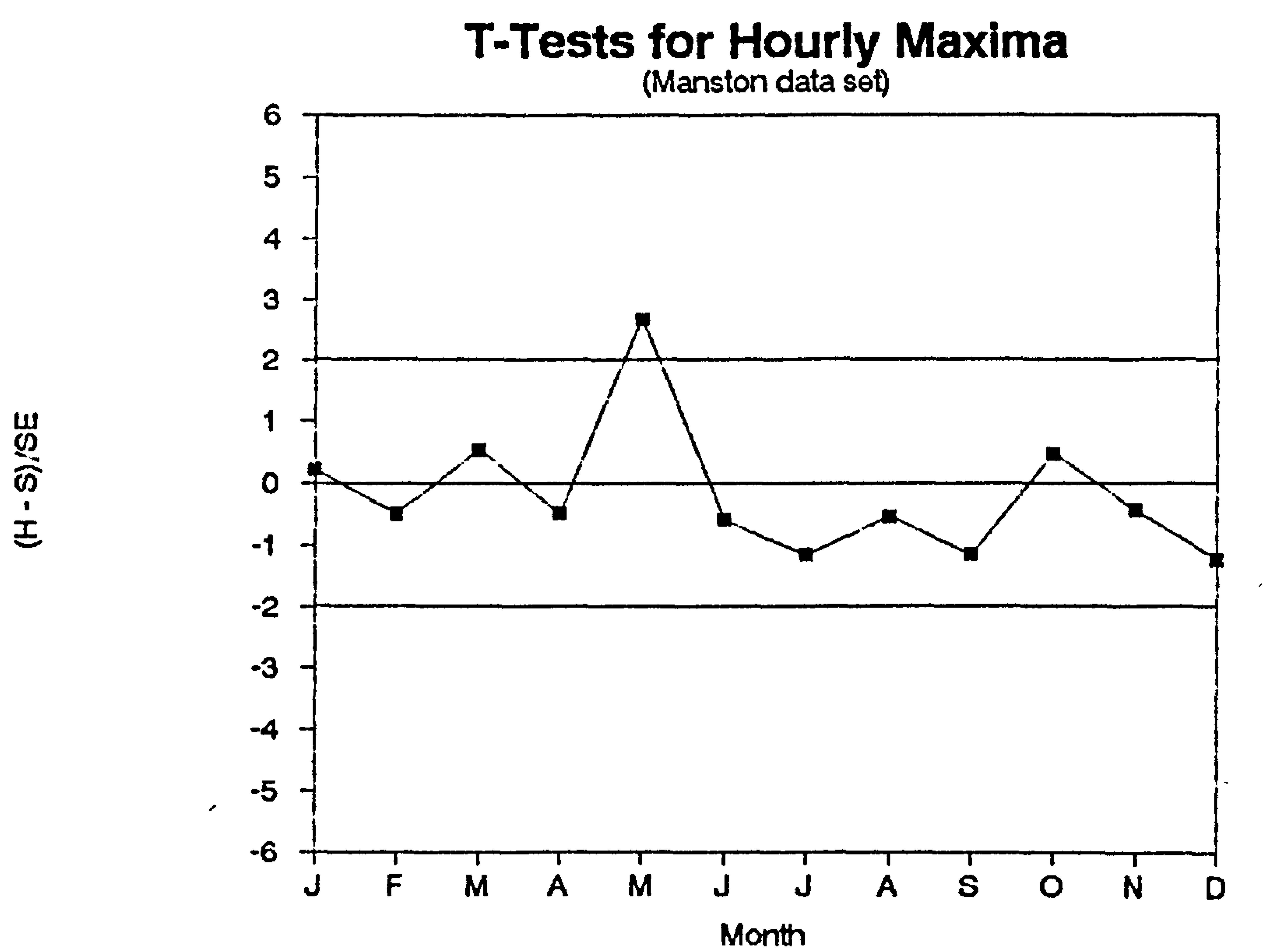


Figure 4.4

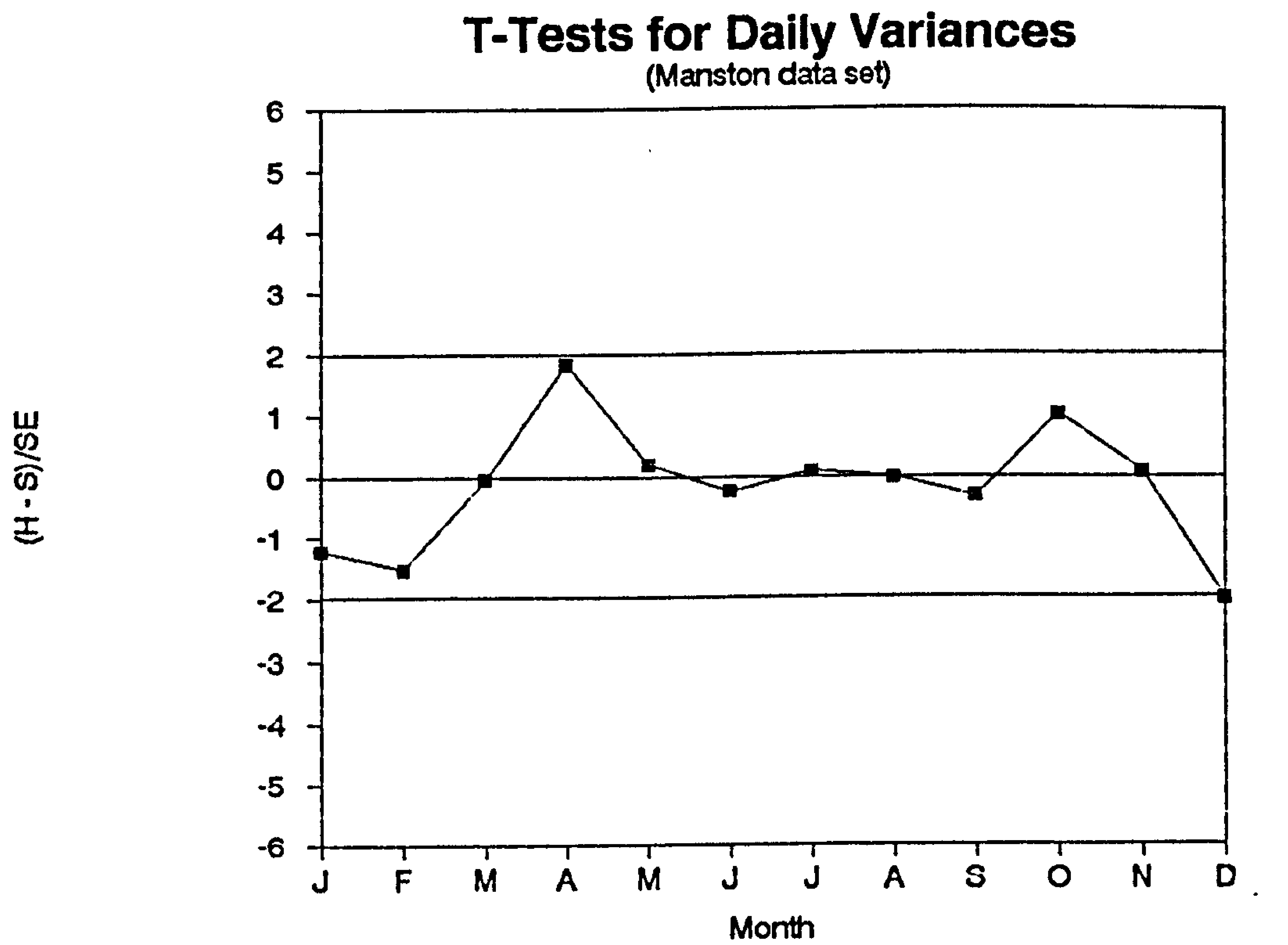


Figure 4.5

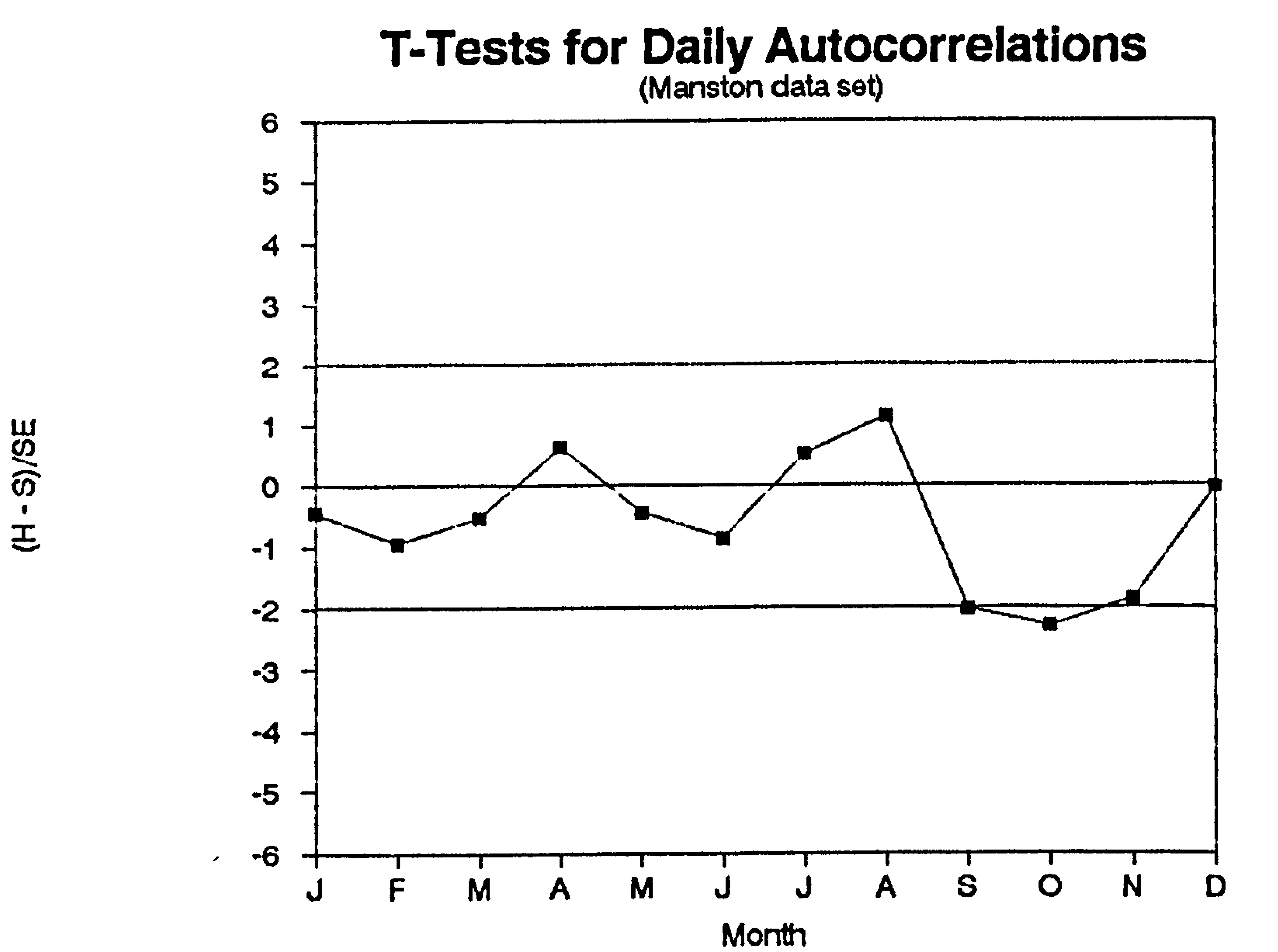


Figure 4.6

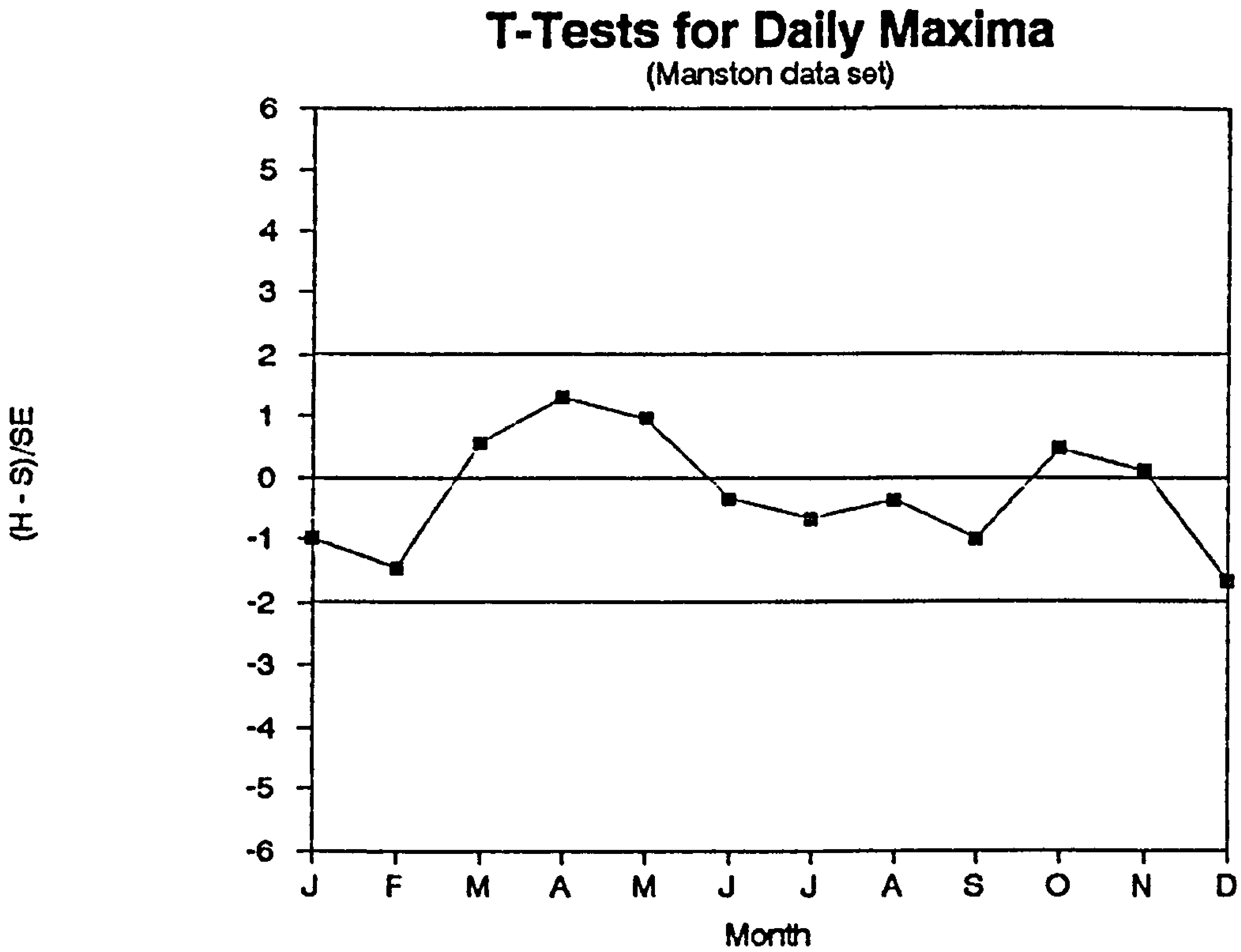


Figure 4.7

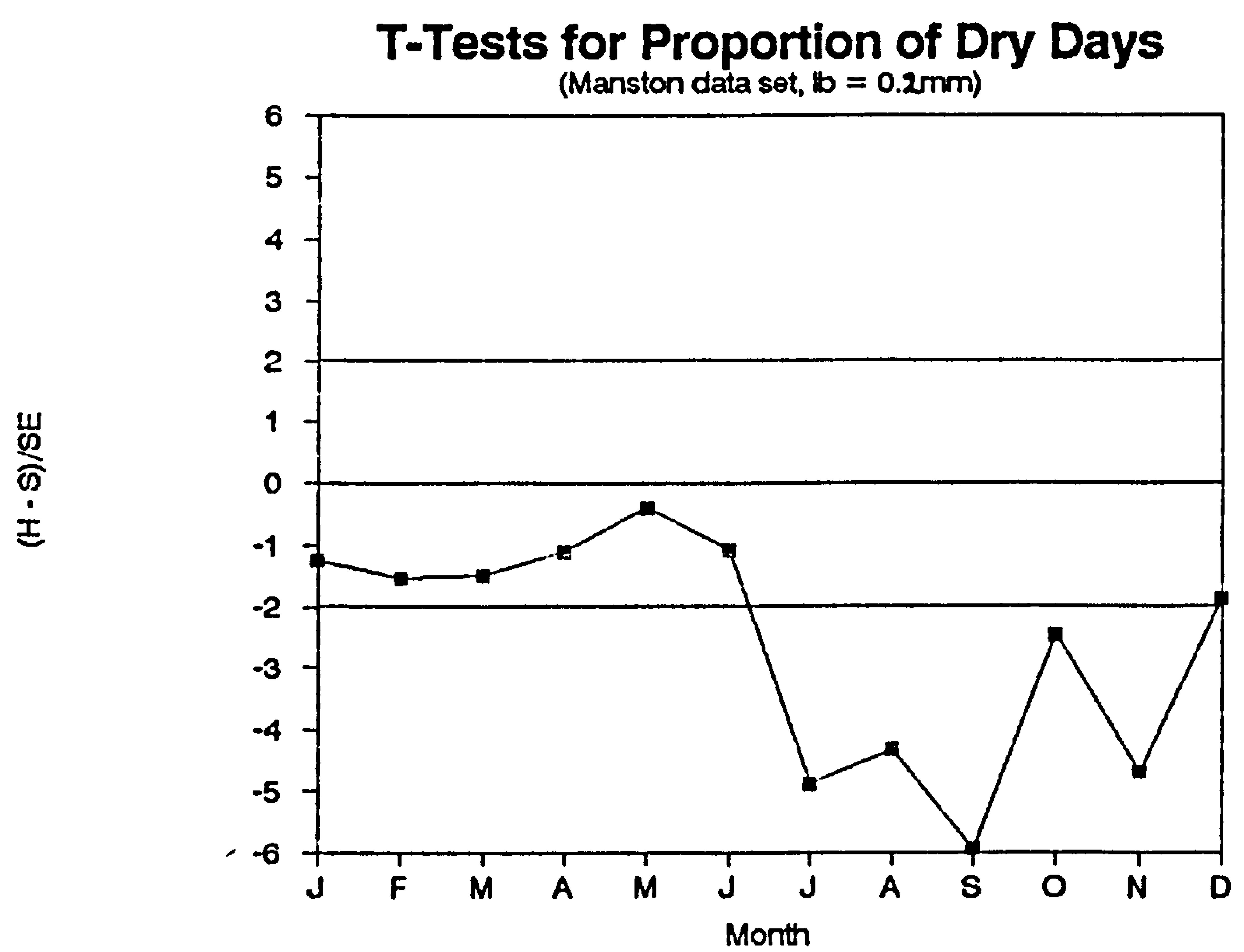


Figure 4.8

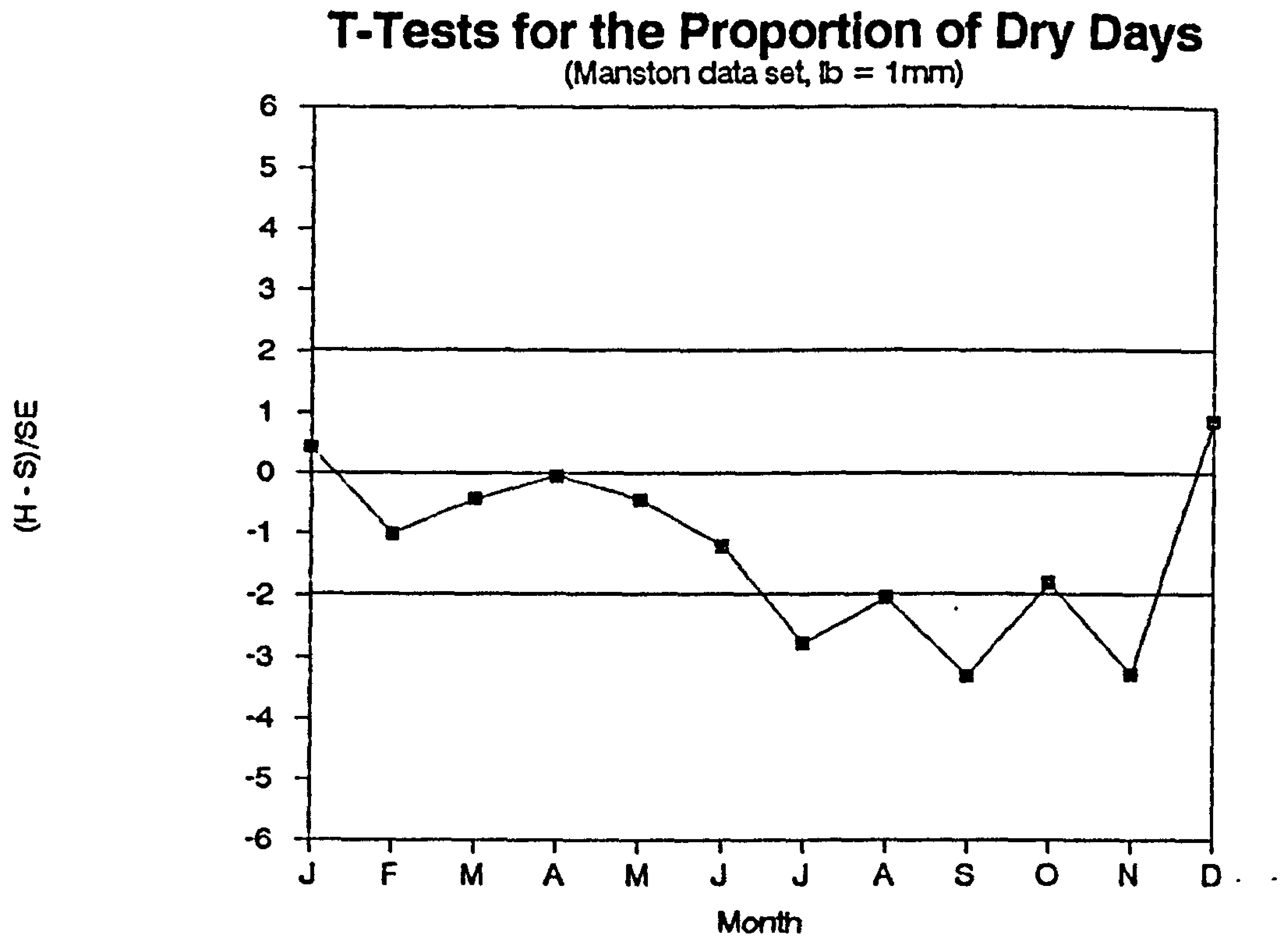


Figure 4.9

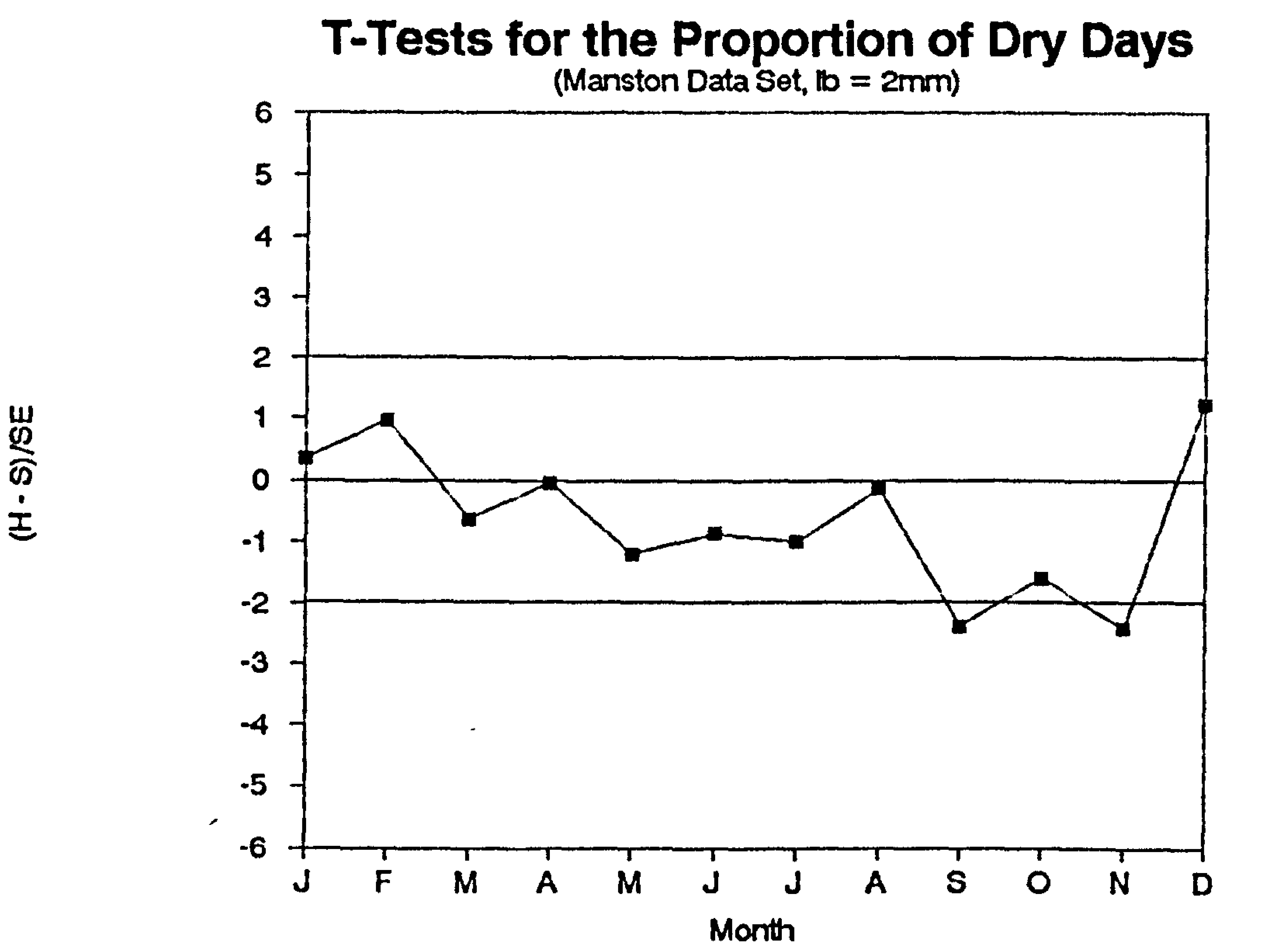


Figure 4.10

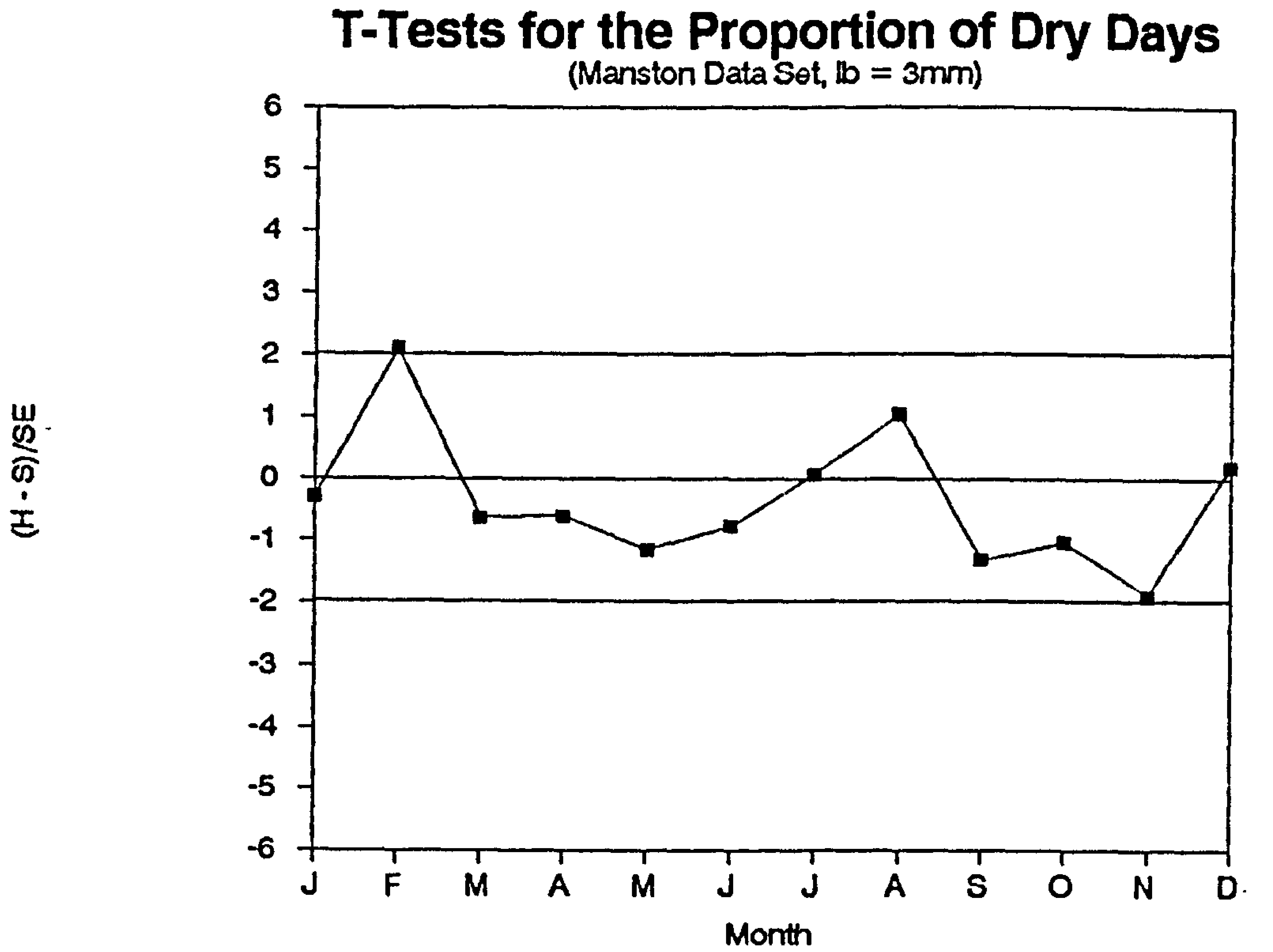

T-Tests for the Proportion of Dry Days
(Manston Data Set, lb = 3mm)
(H - S)/SE
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
J
F
M
A
M
J
J
A
S
O
N
D
Month

Figure 4.11

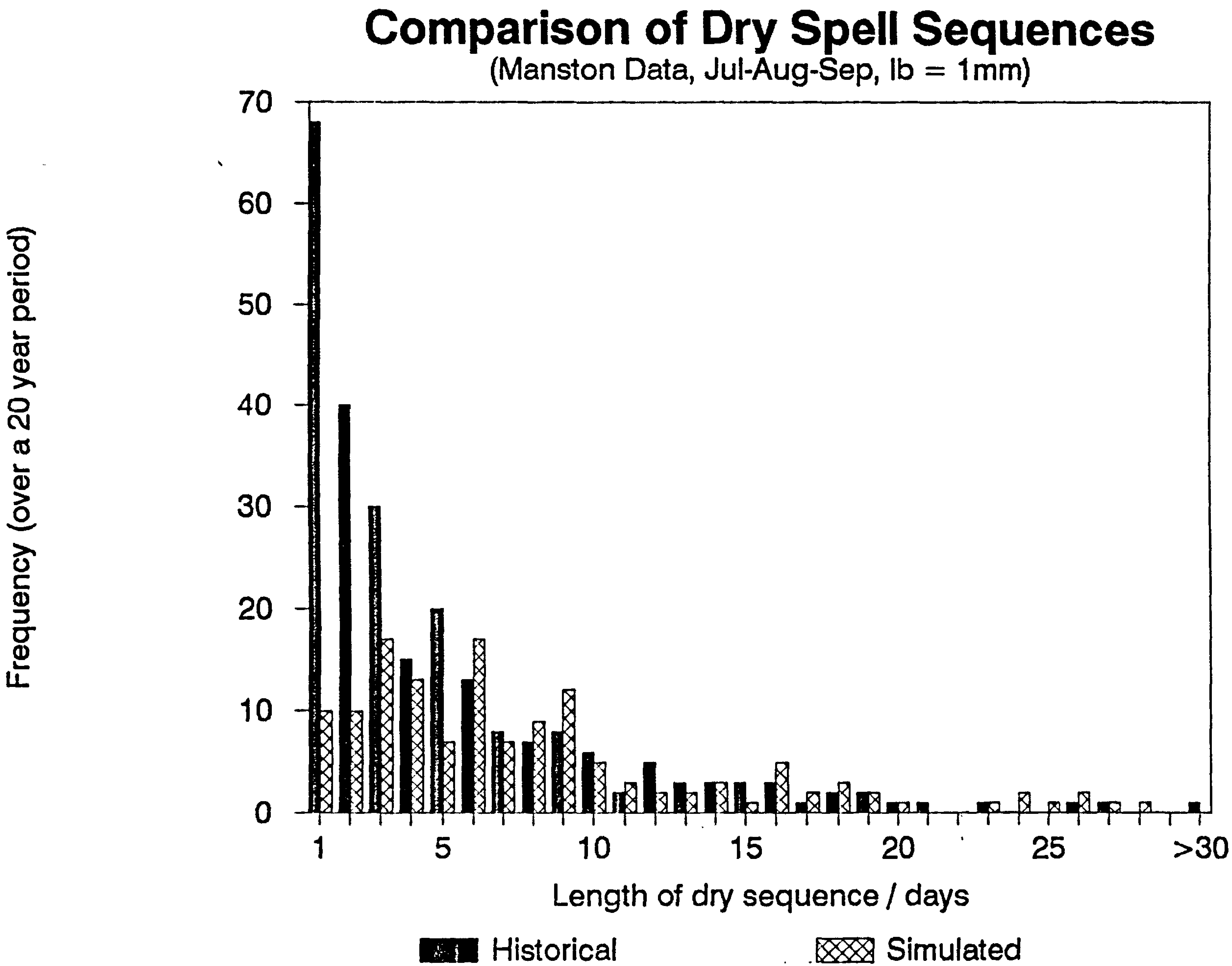

Comparison of Dry Spell Sequences
(Manston Data, Jul-Aug-Sep, lb = 1mm)
Frequency (over a 20 year period)
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of dry sequence / days
Historical
Simulated

Figure 4.12

4.4.6 Conclusions

In Appendix C it can be seen that most of the t-ratios vary naturally about the zero line indicating that, overall, the model is performing well. With so many tests being made the occasional significant result is expected, like, for example, the hourly maxima for May (see Figure 4.4). In fact, about 1 in 20 significant results at the 5% level (which is approximately indicated by the two lines at ±2 in the Figures) are expected for those tests which are independent.

Figure 4.8 shows that the model is tending to over-estimate the proportion of dry days (where a day is defined to be dry if less than 0.2mm of rain fell), particularly in the summer months. The problem looks less drastic if the bound (lb) for a dry day is set at 1, 2, or 3mm (Figures 4.9 - 4.11). Although the 3mm bound gives acceptable results for the statistical tests, from a practical point of view if 3mm of rain fell in a short space of time (which may happen in the summer months) there may be considerable runoff to a sewage system. Hence the results are regarded as practically and statistically significant, and so an improvement will be sought.

The model is also showing a poor fit to summer dry spell sequences (Figure 4.12), particularly to short dry spells. This again may have practical implications with regard to bacterial build up in sewage systems over summer dry spells, and so an improvement is clearly necessary.

4.5 AN ATTEMPT TO IMPROVE THE MODEL'S FIT TO DRY SPELLS

4.5.1 A seasonal model

The model's fit to the daily dry spell sequences may improve if the daily transition probabilities (equations (4.5) and (4.6) with h = 24 hours) are used in the fitting procedure (4.7). However, introducing these probabilities into the fitting procedure increases the number of percentage errors that need to be examined in assessing the model's fit to the historical statistics. To reduce this number a seasonal model will be adopted. The initial seasonal model proposed is not meant to be the best possible choice of seasonal model, and indeed it will be found that the seasonal model does not satisfactorily describe the seasonal variation in the historical data. The proposed initial seasonal model is based on the historical mean monthly totals. The seasonal effect s_j (j = 1,...,12) for the historical time series of monthly totals is given by:

$$s_j = \sum_{i=1}^{N} T_{ij} / N \quad - \quad \sum_{i=1}^{N} \sum_{j=1}^{12} T_{ij} / (12 \times N) \qquad (4.8)$$

where T_{ij} = monthly total for month j, year i.

A plot of the seasonal effect for the Manston data set is given in Figure 4.13, from which it seems reasonable to group the months into four seasons in the following way:

Season 1 (Winter) : Dec, Jan;

Season 2 (Spring) : Feb, Mar, Apr, May;

Season 3 (Summer) : Jun, Jul, Aug;

Season 4 (Autumn) : Sep, Oct, Nov.

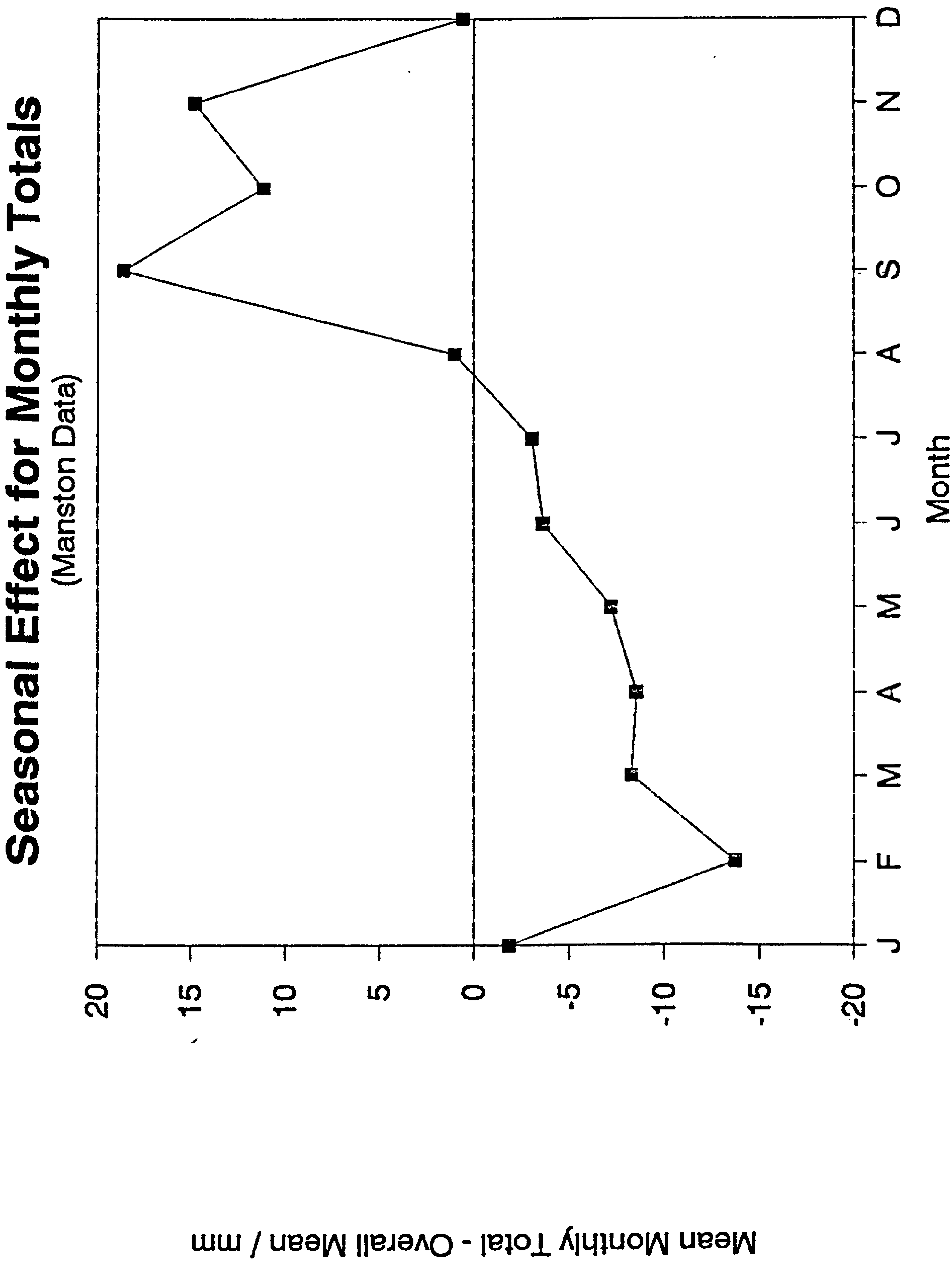

Seasonal Effect for Monthly Totals
(Manston Data)
Mean Monthly Total - Overall Mean / mm
20
15
10
5
0
-5
-10
-15
-20
J
F
M
A
M
J
J
A
S
O
N
D
Month

Figure 4.13

4.5.2 Determining an optimum combination of historical statistics for the fitting procedure

Some statistics that might be used to fit the model are given in Table 4.6. Tables 4.7b - 4.15b give the percentage errors on these statistics when using a subset of these statistics in the fitting procedure (4.7), and Tables 4.7a - 4.15a give the parameter estimates obtained when using this subset (asterisked in Tables 4.7b - 4.15b).

In Table 4.7 it can be seen that the percentage error in the model's fit to WW24 is about 27% for season 3. This Season was poorly matched by the model in Section 4.4 and so the fitting procedure was repeated with WW24 included. Other historical statistics that were poorly matched were: WW6, WW12, AC12, AC24, and PD24. WW6 will be used in the fitting procedure as it may be of practical importance, with reference to storm durations. It was anticipated that using WW24 and WW6 in fitting would preserve WW12, but this was not the case (see below).

In Table 4.8 it can be seen that WW24 is matched well, but that WW12 needs to be used in the fitting procedure. Some statistics, for example PD24 and WW6, that are being used in the fitting procedure are still not being preserved by the model, but an improvement will be found later as other statistics are introduced (or removed) from the fitting procedure.

When WW12 is introduced in the fitting procedure (see Table 4.9), an improvement in the model's fit to WW12 can be seen, although this is at the expense of the model's fit to some of the other historical statistics, for example AC6.

Table 4.6

Statistics taken from the Manston Data Set

	Season			
Statistic	1	2	3	4
M1	0.064	0.054	0.063	0.087
V1	0.090	0.086	0.256	0.296
AC1	0.540	0.493	0.372	0.483
PD1	0.890	0.910	0.940	0.900
DD1	0.957	0.964	0.976	0.963
WW1	0.653	0.639	0.619	0.668
V3	0.529	0.483	1.204	1.668
AC3	0.361	0.355	0.336	0.468
PD3	0.819	0.849	0.898	0.838
DD3	0.902	0.917	0.944	0.916
WW3	0.559	0.534	0.507	0.564
V6	1.417	1.321	3.306	4.839
AC6	0.248	0.254	0.321	0.452
PD6	0.736	0.777	0.846	0.770
DD6	0.836	0.859	0.904	0.860
WW6	0.543	0.508	0.472	0.535
V12	3.577	3.297	8.774	13.51
AC12	0.148	0.158	0.257	0.311
PD12	0.666	0.716	0.795	0.708
DD12	0.750	0.792	0.850	0.802
WW12	0.504	0.477	0.414	0.522
V24	8.479	7.472	23.66	32.71
AC24	0.071	0.161	0.109	0.242
PD24	0.542	0.585	0.696	0.601
DD24	0.629	0.697	0.772	0.709
WW24	0.564	0.571	0.476	0.569

Table 4.7

Using M1, V1, AC1, V6, AC6, V24, PD24, and DD24 in the fitting procedure (4.7)

(a)
Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1} (mmh^{-1})
1	0.0188	0.159	1.22	4.91	0.85
2	0.0136	0.130	1.43	5.49	1.05
3	0.0041	0.085	2.41	11.4	3.29
4	0.0051	0.092	1.96	16.1	2.13

(b)
Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	0.4	0.7	1.3	3.1	1.4
V1	*	2.8	2.7	1.2	3.0	2.4
PD1		1.1	0.4	1.3	2.5	1.3
DD1		0.4	0.5	0.2	0.9	0.5
WW1		0.1	4.1	9.2	0.8	3.5
AC1	*	3.6	3.2	1.1	0.1	2.0
V6	*	1.1	5.6	1.6	11.	4.9
PD6		0.5	0.2	4.3	9.3	3.6
DD6		3.7	3.9	5.7	10.	5.9
WW6		17.	23.	41.	35.	29.
AC6	*	1.9	3.6	3.0	16.	6.1
V12		2.6	5.8	2.8	12.	5.8
PD12		2.6	3.1	9.3	20.	8.8
DD12		9.6	8.2	11.	20.	12.
WW12		15.	18.	47.	29.	27.
AC12		14.	20.	13.	6.0	13.
V24	*	3.9	1.1	6.9	3.5	3.9
PD24	*	0.3	3.9	16.	29.	12.
DD24	*	4.9	4.1	18.	29.	14.
WW24		4.9	0.8	27.	14.	12.
AC24		30.	30.	88.	9.9	40.
Mean		5.8	6.8	15.	13.	10.

* = used in fitting

Table 4.8

Using M1, V1, AC1, V6, AC6, WW6, V24, PD24, WW24, and DD24 in the fitting procedure (4.7)

(a)
Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1} (mmh^{-1})
1	0.0185	0.247	1.29	5.51	0.81
2	0.0150	0.159	1.32	4.47	1.06
3	0.0064	0.162	2.27	6.79	3.33
4	0.0068	0.128	1.65	9.81	2.18

(b)
Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	0.6	0.8	0.4	1.8	0.9
V1	*	4.0	5.3	6.6	3.4	4.8
PD1		0.2	0.2	1.7	3.3	1.4
DD1		0.5	0.1	0.5	1.2	0.5
WW1		6.2	3.0	7.4	1.7	4.6
AC1	*	6.1	7.5	5.9	5.8	6.3
V6	*	2.8	1.1	3.7	5.3	3.2
PD6		4.0	1.5	5.3	10.	5.3
DD6		5.9	3.9	5.3	9.6	6.2
WW6	*	15.	17.	29.	26.	22.
AC6	*	6.9	11.	24.	31.	18.
V12		0.3	2.8	2.7	11.	4.2
PD12		1.8	1.7	6.6	13.	5.8
DD12		6.5	4.2	8.5	14.	8.3
WW12		15.	23.	37.	24.	25.
AC12		9.9	1.0	31.	22.	16.
V24	*	4.1	0.6	15.	8.5	7.1
PD24	*	0.1	0.8	12.	21.	8.7
DD24	*	2.1	0.1	11.	19.	8.2
WW24	*	2.4	1.4	3.3	2.6	2.4
AC24		9.5	46.	10.	41.	27.
Mean		4.9	6.3	11.	13.	8.8

* = used in fitting

Table 4.9

Using M1, V1, AC1, V6, AC6, WW6, WW12, V24, PD24, WW24, and DD24 in the fitting procedure (4.7)

(a)
Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1} (mmh^{-1})
1	0.0188	0.284	1.28	5.52	0.79
2	0.0164	0.211	1.30	4.19	1.02
3	0.0080	0.191	2.03	4.92	3.28
4	0.0084	0.135	1.35	6.55	2.17

(b)
Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	0.7	0.9	0.3	1.5	0.8
V1	*	3.9	5.4	8.6	5.4	5.8
PD1		0.0	0.4	1.8	3.6	1.4
DD1		0.7	0.2	0.5	1.1	0.6
WW1		8.2	0.2	9.5	4.0	5.5
AC1	*	8.1	10.	9.7	12.	9.9
V6	*	4.0	0.7	4.8	2.9	3.1
PD6		5.1	2.6	4.9	9.8	5.6
DD6		6.0	4.0	4.4	8.4	5.7
WW6	*	12.	14.	18.	18.	16.
AC6	*	11.	16.	35.	40.	26.
V12		0.6	2.1	4.7	12.	4.7
PD12		2.9	0.5	5.3	11.	5.0
DD12		6.2	3.2	6.5	12.	6.9
WW12	*	11.	15.	24.	17.	17.
AC12		18.	16.	45.	34.	29.
V24	*	4.9	2.1	19.	12.	9.6
PD24	*	0.7	0.5	8.9	17.	6.9
DD24	*	1.3	3.1	6.8	15.	6.6
WW24	*	0.1	4.2	4.9	1.9	2.8
AC24		20.	59.	32.	51.	40.
Mean		6.0	7.6	12.	14.	9.9

* = used in fitting

In Table 4.10 hourly statistics are omitted from the fitting procedure (with the exception of M1, which can be found from monthly (or seasonal) totals). It is clear that there is an improvement in the model's fit to 6, 12 and 24 hourly historical statistics, but this is at the expense of the model's fit to the historical hourly statistics. At some stage it must be decided if the improvement is great enough to warrant fitting the model using h hourly historical statistics, where h is greater than 1 hour, and then disaggregating simulated h hourly time series data to obtain time series of smaller time steps. After some further experiments on the statistics to be used in fitting, the performance of a 3 hourly time series model (i.e. a model fitted using h hourly historical statistics where h ≥ 3 hours) will be assessed.

Having found a considerable improvement in Table 4.10, the lag 1 autocorrelations for the 12 and 24 hourly time series are re-introduced into the fitting procedure to see whether their presence is detrimental to the model's fit to 6, 12, or 24 hourly statistics (Table 4.11). Looking at Table 4.11 there appears to be no major change to the model's fit to the h (≥6) hourly statistics.

In Table 4.12 hourly statistics are re-introduced in the fitting procedure so that an objective comparison can be made with Table 4.11. Table 4.12 shows that it is the presence of the 1 hourly statistics in the fitting procedure that throws out the model's fit to other h (≥6) hourly statistics of importance, for example PD24 for Seasons 3 and 4.

Table 4.10

Using M1, V6, AC6, WW6, WW12, V24, PD24, WW24, and DD24 in the fitting procedure (4.7)

(a)
Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1}(mmh^{-1})
1	0.0213	0.595	0.56	4.28	0.40
2	0.0196	0.290	0.66	2.89	0.63
3	0.0160	11.15	0.31	1.32	0.95
4	0.0103	0.028	0.27	2.13	1.07

(b)
Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	0.1	0.7	1.1	0.8	0.7
V1		22.	18.	43.	39.	31.
PD1		1.5	1.0	1.3	0.7	1.1
DD1		2.0	0.8	0.9	1.9	1.4
WW1		27.	16.	29.	22.	23.
AC1		44.	46.	116.	76.	70.
V6	*	3.7	2.0	2.7	6.0	3.6
PD6		6.3	2.4	1.3	7.3	4.3
DD6		5.3	3.0	0.5	5.0	3.4
WW6	*	4.4	8.1	3.7	1.1	4.3
AC6	*	0.9	1.2	8.6	3.2	3.5
V12		2.8	2.3	4.4	3.2	3.2
PD12		3.4	1.6	2.1	5.4	3.1
DD12		3.2	0.3	2.9	4.1	2.6
WW12	*	0.5	4.8	6.4	2.0	3.4
AC12		25.	20.	33.	14.	23.
V24	*	3.6	1.6	9.2	1.4	4.0
PD24	*	1.7	5.0	7.8	3.6	4.5
DD24	*	4.7	10.	12.	3.3	7.5
WW24	*	3.7	7.1	10.	1.8	5.6
AC24		29.	64.	29.	27.	37.
Mean		9.2	10.	15.	11.	11.

* = used in fitting

Table 4.11

Using M1, V6, AC6, WW6, WW12, AC12, V24, PD24, WW24, DD24, and AC24 in the fitting procedure (4.7)

(a)
Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1}(mmh^{-1})
1	0.0204	0.225	0.76	3.56	0.66
2	0.0055	0.006	0.49	6.98	0.70
3	0.0118	0.098	0.38	1.65	1.23
4	0.0064	0.018	0.24	3.34	1.00

(b)
Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	2.1	0.2	0.0	0.8	0.8
V1		8.0	23.	41.	42.	28.
PD1		0.7	1.5	0.2	0.6	0.8
DD1		0.8	1.2	0.9	2.0	1.2
WW1		12.	9.6	24.	24.	17.
AC1		27.	49.	111.	78.	67.
V6	*	7.1	1.1	4.3	8.2	5.2
PD6		3.4	5.4	2.9	8.0	4.9
DD6		4.5	2.6	2.0	6.1	3.8
WW6	*	9.8	9.5	1.5	6.7	6.8
AC6	*	2.9	1.2	11.	3.6	4.6
V12		6.5	0.3	6.5	3.4	4.2
PD12		0.1	0.8	1.0	7.2	2.3
DD12		3.9	2.1	1.3	6.7	3.5
WW12	*	12.	5.8	5.0	4.5	6.8
AC12	*	3.9	2.8	20.	4.2	7.6
V24	*	2.7	2.2	4.7	3.6	3.3
PD24	*	4.4	0.3	0.8	8.0	3.4
DD24	*	2.6	4.6	2.7	9.9	5.0
WW24	*	3.3	8.6	6.5	4.1	5.6
AC24	*	3.1	6.9	5.5	14.	7.3
Mean		5.7	6.6	12.	12.	9.0

* = used in fitting

Table 4.12

Using M1, V1, AC1, V6, AC6, WW6, WW12, AC12, V24, PD24, WW24, DD24, and AC24 in the fitting procedure (4.7)

(a)
Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1} (mmh^{-1})
1	0.0185	0.218	1.18	4.86	0.84
2	0.0136	0.063	0.97	3.42	1.09
3	0.0062	0.140	2.05	6.19	3.34
4	0.0055	0.070	1.35	9.64	2.25

(b)
Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	1.5	3.2	0.6	0.8	1.5
V1	*	4.6	6.4	8.3	4.9	6.0
PD1		0.0	0.9	1.8	3.5	1.6
DD1		0.4	0.1	0.5	0.9	0.5
WW1		4.4	6.8	9.9	7.0	7.1
AC1	*	8.0	18.	8.9	9.5	11.
V6	*	4.3	4.2	5.4	4.7	4.6
PD6		3.6	0.3	5.2	9.1	4.6
DD6		5.4	0.7	5.1	8.7	5.0
WW6	*	14.	2.5	26.	23.	16.
AC6	*	5.6	20.	25.	33.	21.
V12		2.0	0.5	1.5	11.	3.7
PD12		1.0	5.8	6.3	11.	6.0
DD12		6.3	0.9	8.4	14.	7.4
WW12	*	16.	23.	38.	30.	27.
AC12	*	4.6	10.	28.	7.2	12.
V24	*	1.8	4.1	13.	5.2	6.1
PD24	*	1.1	7.9	12.	19.	10.
DD24	*	2.0	0.1	11.	22.	8.9
WW24	*	3.8	13.	6.6	16.	9.9
AC24	*	1.5	12.	0.3	8.3	5.5
Mean		4.3	6.7	11.	12.	8.4

* = used in fitting

Table 4.13 is given to show that using just 1 and 24 hourly statistics in the fitting procedure produces poor results in the model's fit to historical 6 and 12 hourly statistics (see, for example, WW6 and WW12)

Tables 4.10 and 4.11 have shown that the model can be made to fit h (≥6) hourly statistics reasonably well, with most percentage errors less than 10. Table 4.14 shows the results of introducing 3 hourly statistics into the fitting procedure. It can be seen that introducing these statistics is at the expense of the model's fit to some of the 12 and 24 hourly statistics, for example WW12, although the results do seem quite good. To see if the results can be improved the lag 1 autocorrelations for the 12 and 24 hourly time series are removed from the fitting procedure and the results are shown in Table 4.15, an improvement being evident in, for example, PD24, WW24, and DD24.

Table 4.13

Using M1, V1, AC1, V24, PD24, WW24, and DD24 in the fitting procedure (4.7)

(a)
Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1} (mmh^{-1})
1	0.0180	0.245	1.41	6.19	0.82
2	0.0142	0.175	1.55	5.71	1.04
3	0.0063	0.168	2.41	7.24	3.38
4	0.0112	0.107	1.23	4.02	2.40

(b)
Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	0.1	0.2	0.9	1.0	0.6
V1	*	3.8	2.4	7.0	9.7	5.7
PD1		0.4	0.3	1.7	3.7	1.5
DD1		0.4	0.2	0.5	0.7	0.5
WW1		6.5	1.1	6.4	13.	6.6
AC1	*	3.4	2.0	4.0	9.6	4.7
V6		1.2	5.7	4.1	4.5	3.9
PD6		3.9	1.4	5.4	7.2	4.5
DD6		6.2	4.8	5.4	5.2	5.4
WW6		17.	23.	30.	3.7	18.
AC6		4.8	5.3	22.	52.	21.
V12		0.8	6.2	2.0	17.	6.4
PD12		1.9	0.9	6.9	5.5	3.8
DD12		7.2	5.6	8.7	6.8	7.0
WW12		16.	26.	37.	10.	22.
AC12		7.3	4.5	31.	44.	21.
V24	*	4.9	3.6	14.	19.	10.
PD24	*	0.6	1.3	13.	6.3	5.2
DD24	*	3.4	1.9	11.	7.1	5.9
WW24	*	2.9	1.2	2.9	0.8	1.9
AC24		6.8	46.	11.	52.	29.
Mean		5.0	6.8	11.	13.	8.9

* = used in fitting

Table 4.14

Using M1, V3, AC3, V6, AC6, WW6, WW12, AC12, V24, PD24, WW24, DD24, and AC24 in the fitting procedure (4.7)

(a) Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1} (mmh^{-1})
1	0.0199	0.216	0.90	3.89	0.73
2	0.0053	0.007	0.57	7.45	0.78
3	0.0095	0.092	0.61	2.42	1.66
4	0.0053	0.022	0.35	4.72	1.23

(b) Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	2.1	0.5	0.7	0.7	1.0
V1		3.1	16.	26.	30.	19.
PD1		0.4	1.8	0.8	1.9	1.2
DD1		0.6	1.0	0.7	1.8	1.0
WW1		8.0	5.5	11.	17.	10.
AC1		20.	43.	86.	68.	54.
V3	*	3.2	1.0	4.3	7.5	4.0
PD3		0.8	3.6	1.9	5.2	2.9
DD3		2.1	1.4	1.3	3.5	2.1
WW3		12.	8.6	3.2	8.7	8.2
AC3	*	4.1	13.	24.	21.	16.
V6	*	6.0	1.4	7.7	0.0	3.8
PD6		3.3	5.1	3.3	8.6	5.1
DD6		4.6	2.2	2.4	6.1	3.8
WW6	*	11.	11.	2.4	4.0	6.9
AC6	*	1.9	9.2	18.	15.	11.
V12		4.4	0.0	2.6	0.9	2.0
PD12		0.2	0.1	1.8	7.8	2.5
DD12		4.5	2.0	3.2	8.2	4.5
WW12	*	14.	7.8	16.	9.5	12.
AC12	*	4.4	1.9	32.	16.	14.
V24	*	0.6	1.9	11.	3.3	4.1
PD24	*	3.8	1.1	1.9	10.	4.2
DD24	*	1.3	5.4	2.3	14.	5.9
WW24	*	4.1	11.	2.4	11.	7.2
AC24	*	1.9	3.8	3.3	10.	4.7
Mean		4.7	6.1	10.	11.	8.1

Table 4.15
Using M1, V3, AC3, V6, AC6, WW6, WW12, V24, PD24, WW24, DD24 and AC24 in the fitting procedure (4.7)

(a) Parameter estimates

Season	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ^{-1} (mmh^{-1})
1	0.02028	0.3189	0.9160	4.495	0.6400
2	0.01843	0.2353	0.8306	3.126	0.7714
3	0.01097	0.1137	0.6536	2.193	1.7132
4	0.01160	0.0611	0.3831	2.269	1.2780

(b) Absolute percentage errors

Statistic		Season 1	2	3	4	mean
M1	*	1.0	1.4	0.4	1.0	0.9
V1		5.5	9.2	24.	27.	17.
PD1		0.4	0.2	0.8	1.9	0.9
DD1		1.0	0.5	0.6	1.7	1.0
WW1		14.	9.2	8.8	15.	12.
AC1		23.	34.	82.	65.	51.
V3	*	2.3	2.3	5.3	5.2	3.8
PD3		1.6	0.8	1.8	4.9	2.3
DD3		3.0	1.6	1.1	3.1	2.2
WW3		17.	11.	0.6	5.4	8.5
AC3	*	6.2	6.0	18.	16.	12.
V6	*	5.6	3.0	7.2	0.9	4.2
PD6		5.1	2.5	3.0	7.9	4.6
DD6		5.4	3.2	1.9	5.1	3.9
WW6	*	9.0	9.1	1.0	1.2	5.1
AC6	*	8.2	8.9	23.	21.	15.
V12		2.7	1.7	0.8	1.9	1.8
PD12		2.3	1.3	1.0	6.1	2.7
DD12		4.4	0.9	2.0	5.0	3.1
WW12	*	6.6	8.6	10.	0.1	6.4
AC12		22.	19.	38.	28.	27.
V24	*	3.4	1.2	13.	0.8	4.7
PD24	*	1.6	3.6	0.0	5.2	2.6
DD24	*	2.3	7.8	0.8	4.5	3.8
WW24	*	0.8	5.9	2.5	2.5	2.9
AC24	*	25.	62.	13.	39.	35.
Mean		6.9	8.2	10.	11.	8.9

4.5.4 Testing the performance of a 3 hourly time series model

Using the parameter estimates given in Table 4.15a rainfall time series were simulated using the simulation program (Appendix B). The t-tests were carried out in the same way as in Section 4.4, with the addition of tests on 3 hourly statistics. The results of of some selected examples of the t-tests are given in Figures 4.14 - 4.23 (the results for all the t-tests are given in Appendix D). A substantial improvement in the model's fit to the historical proportion of dry days is evident in Figures 4.22 and 4.23 (compare with Figures 4.8 and 4.9 respectively).

As might be anticipated many of the 1 hourly statistics are not preserved by the model, particularly the lag 1 hourly autocorrelations (Figure 4.16). Also the month of November is consistently incorrectly matched by the model. This is almost certainly due to non-stationarities present within the seasons of the model. Looking at Figures 4.25 and 4.26 the presence of non-stationarities is evident in season 3 (Sep, Oct, Nov) where clearly September has a much greater variance in the daily time series than both October and November, which leads to over-estimation of the variance, and consequently the maxima, for these two months.

Looking at Figure 4.20 it can be seen that the model matches the lag 1 daily autocorrelations of the historical time series (with the exception of November), within sampling error. This is of particular interest as they were not used in the fitting procedure, and suggests that perhaps a better procedure may be to omit autocorrelations from the parameter estimation procedure, and use transition probabilities instead. This will be considered in the next Section.

The main interest in this Section was to see whether the introduction of WW24 and DD24 in the fitting procedure improved the model's fit to the daily historical dry spell sequences. Figure 4.24 shows the dry spell frequency plot for July, August and September (the plots for the other months are given in Appendix D), from which it is clear that the model is now performing well.

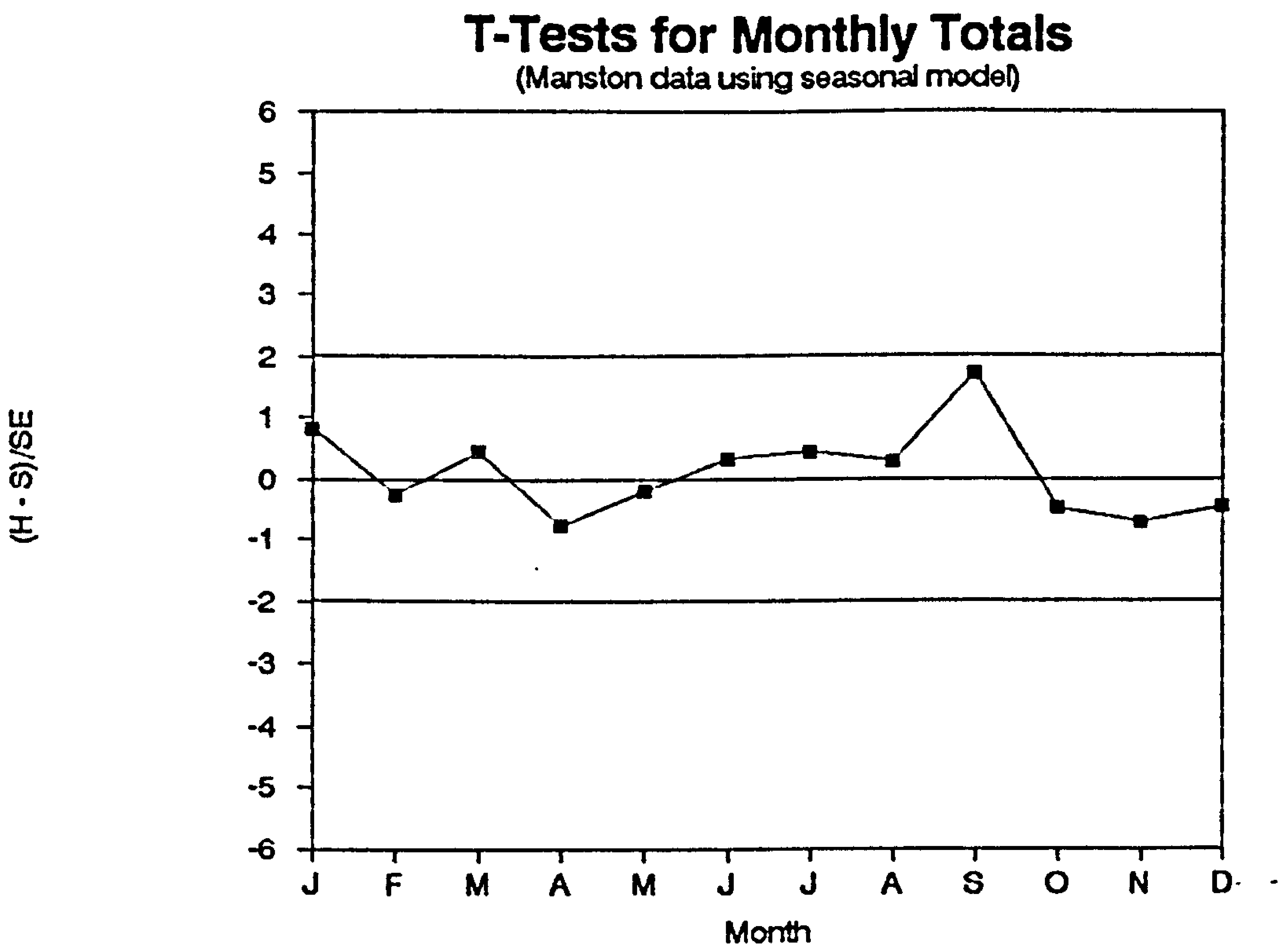


Figure 4.14

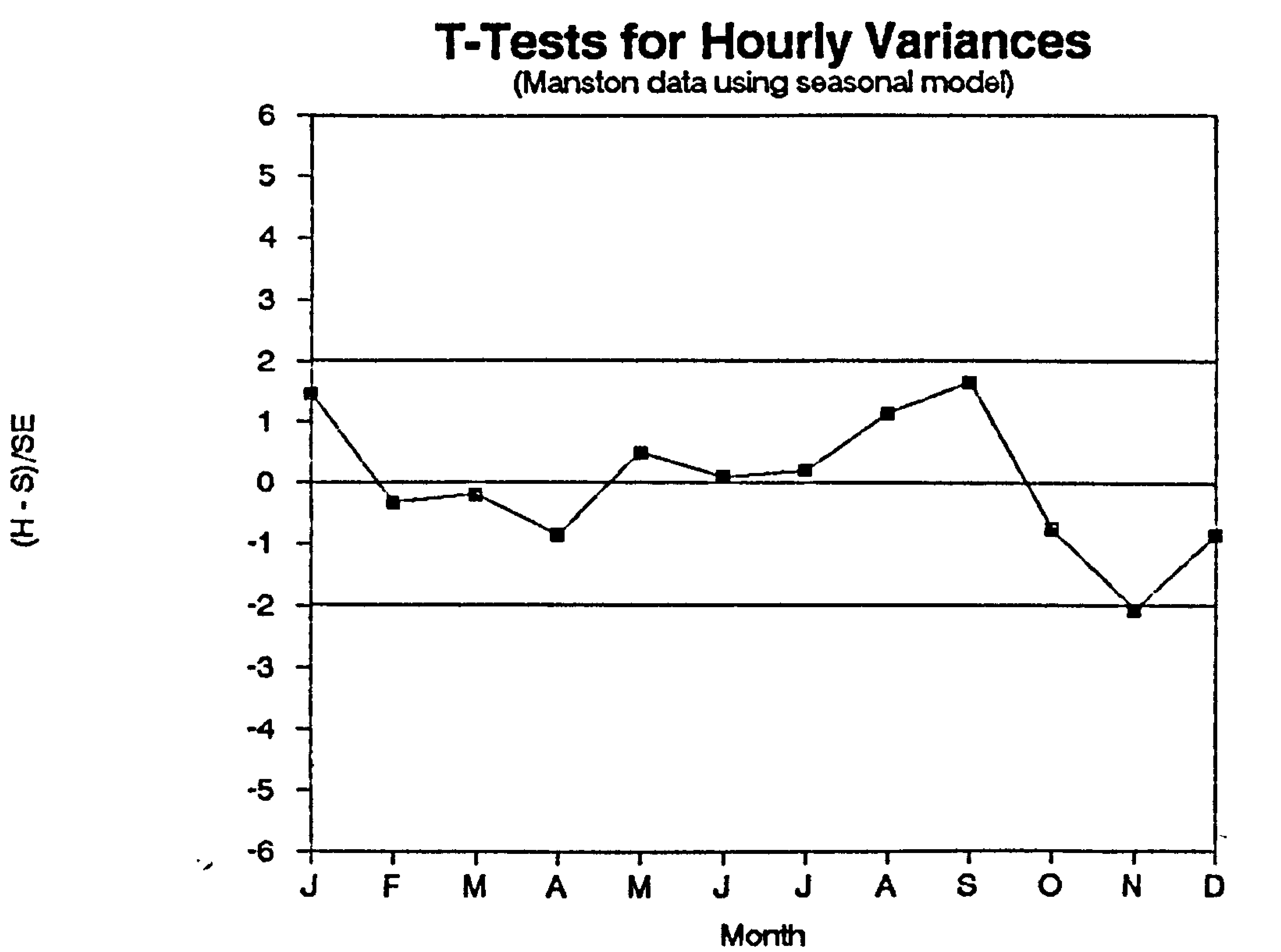


Figure 4.15

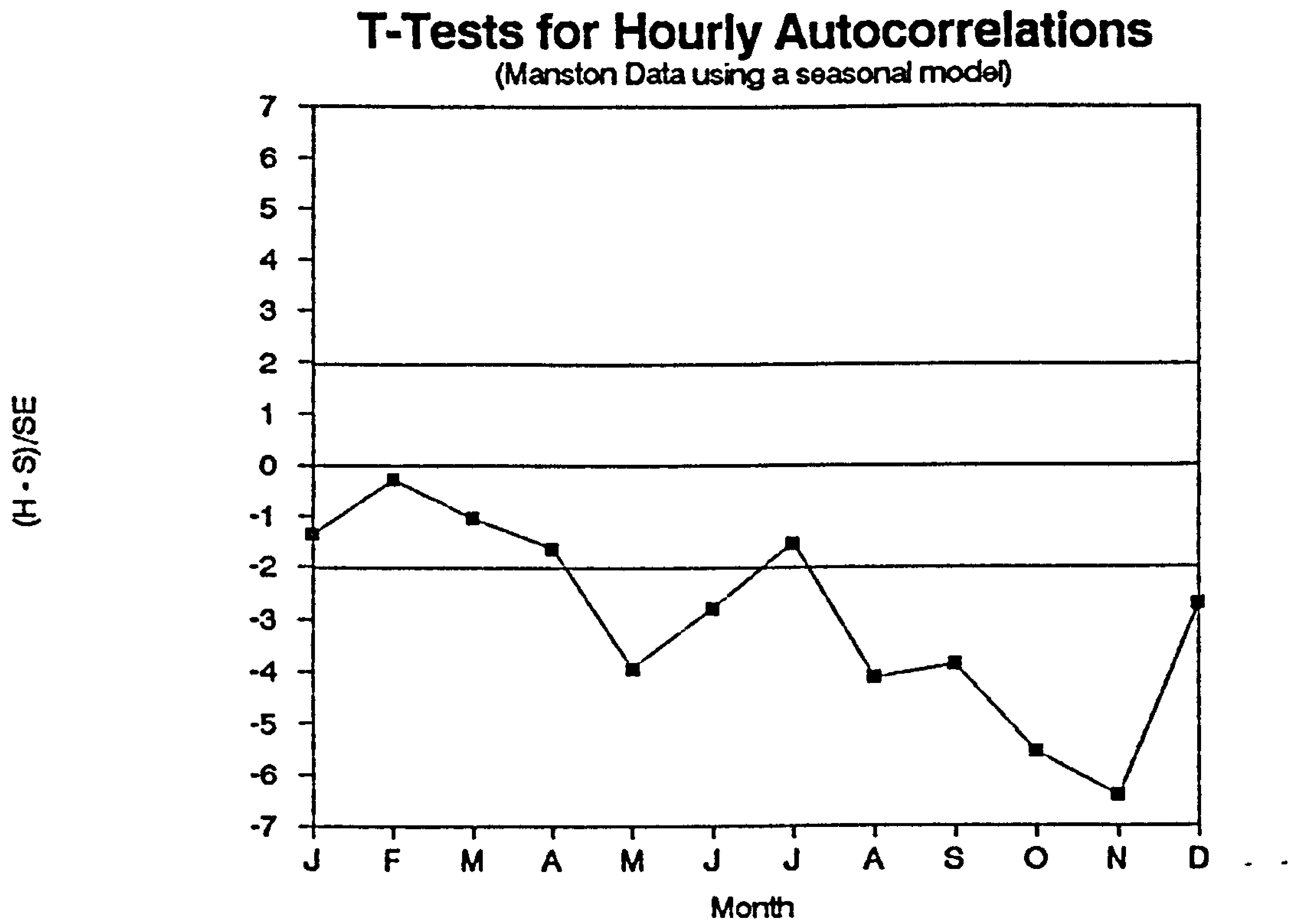


Figure 4.16

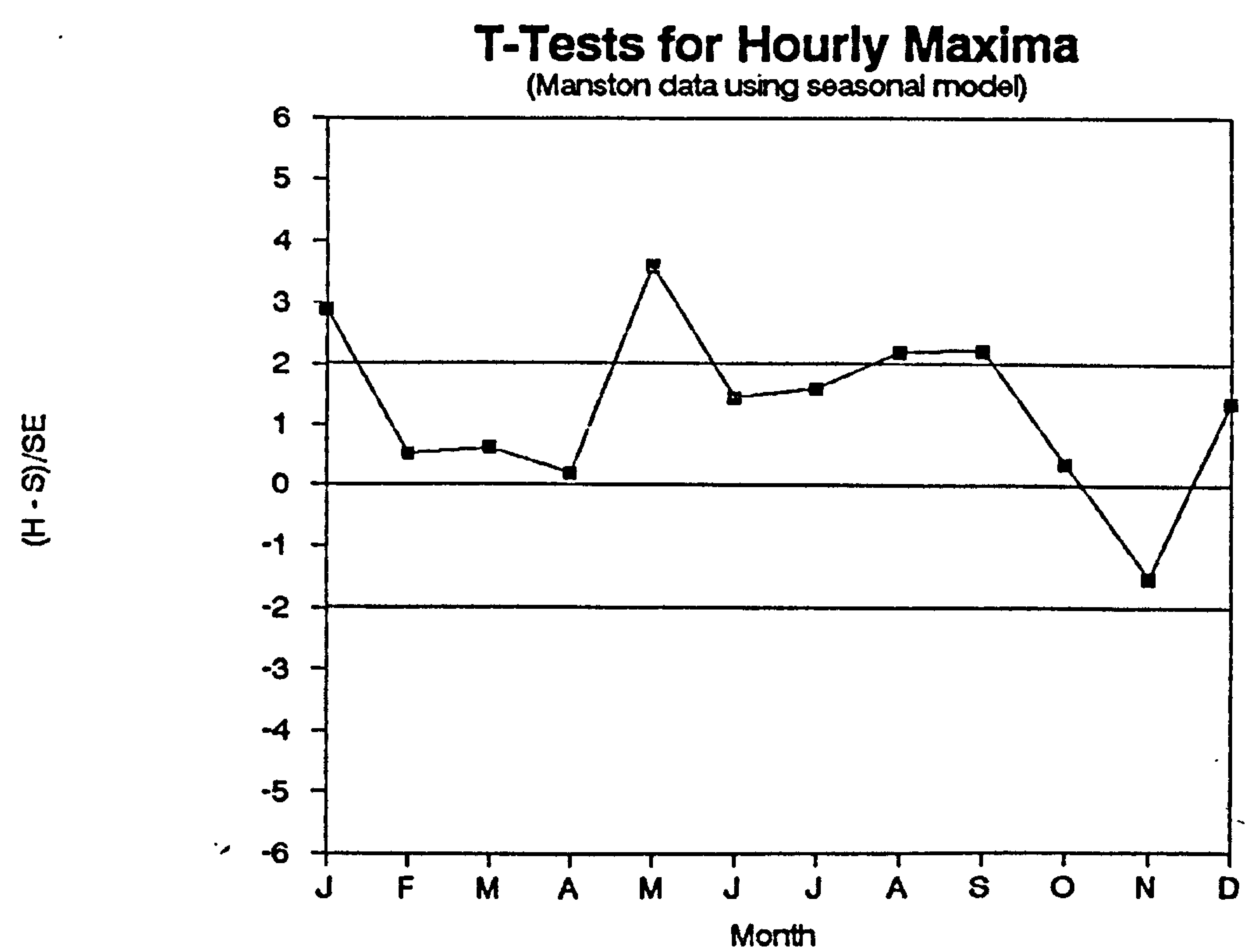


Figure 4.17

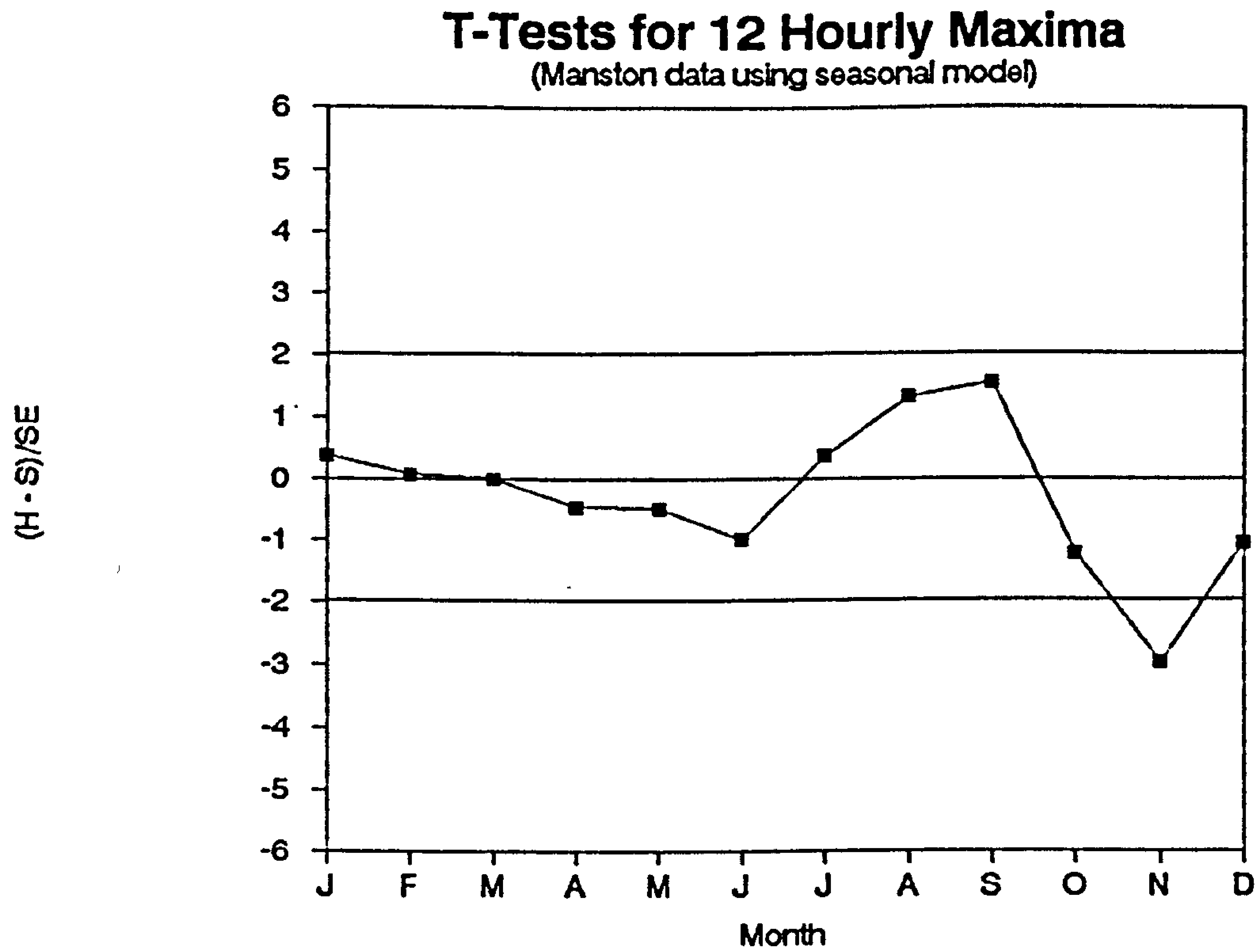


Figure 4.18

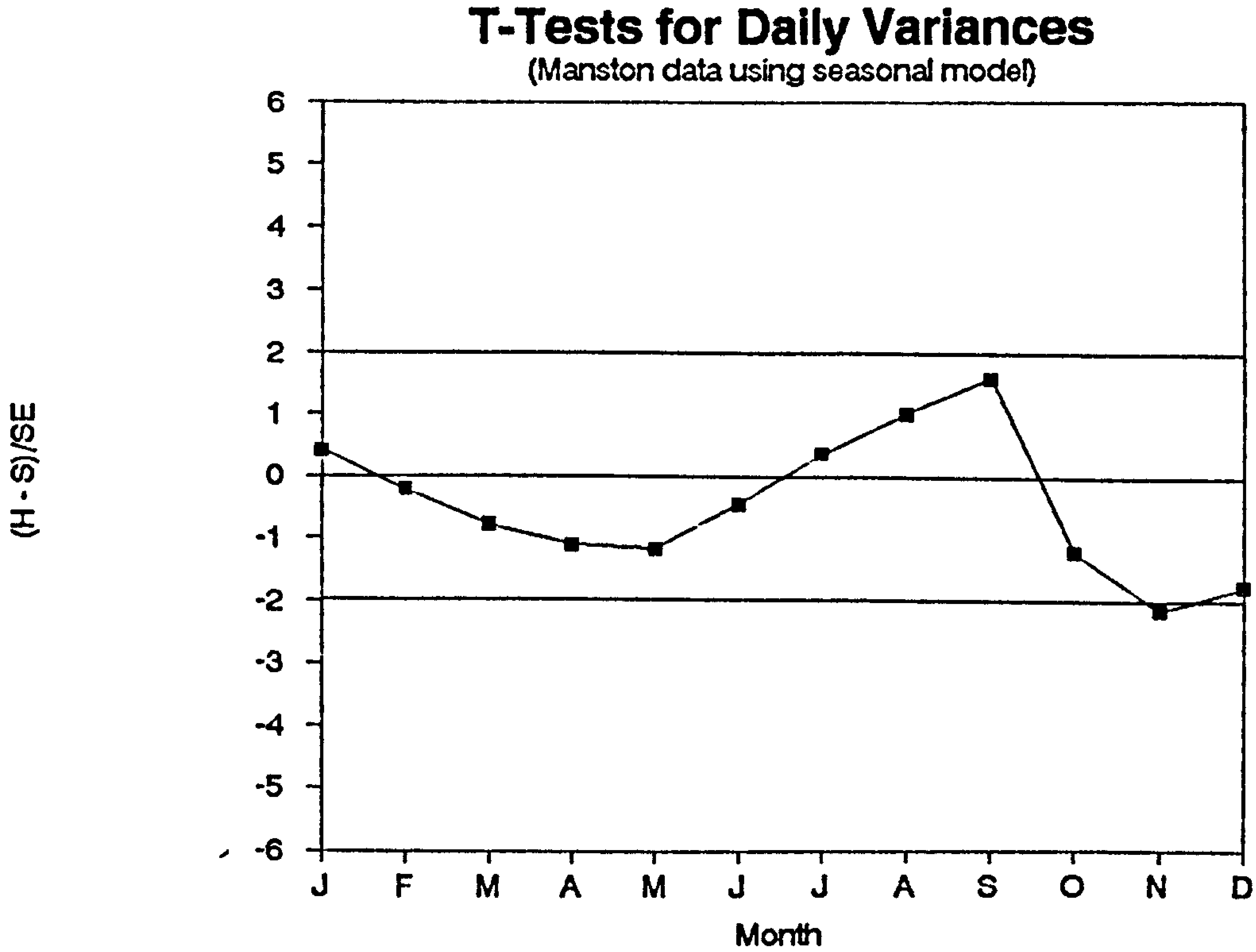


Figure 4.19

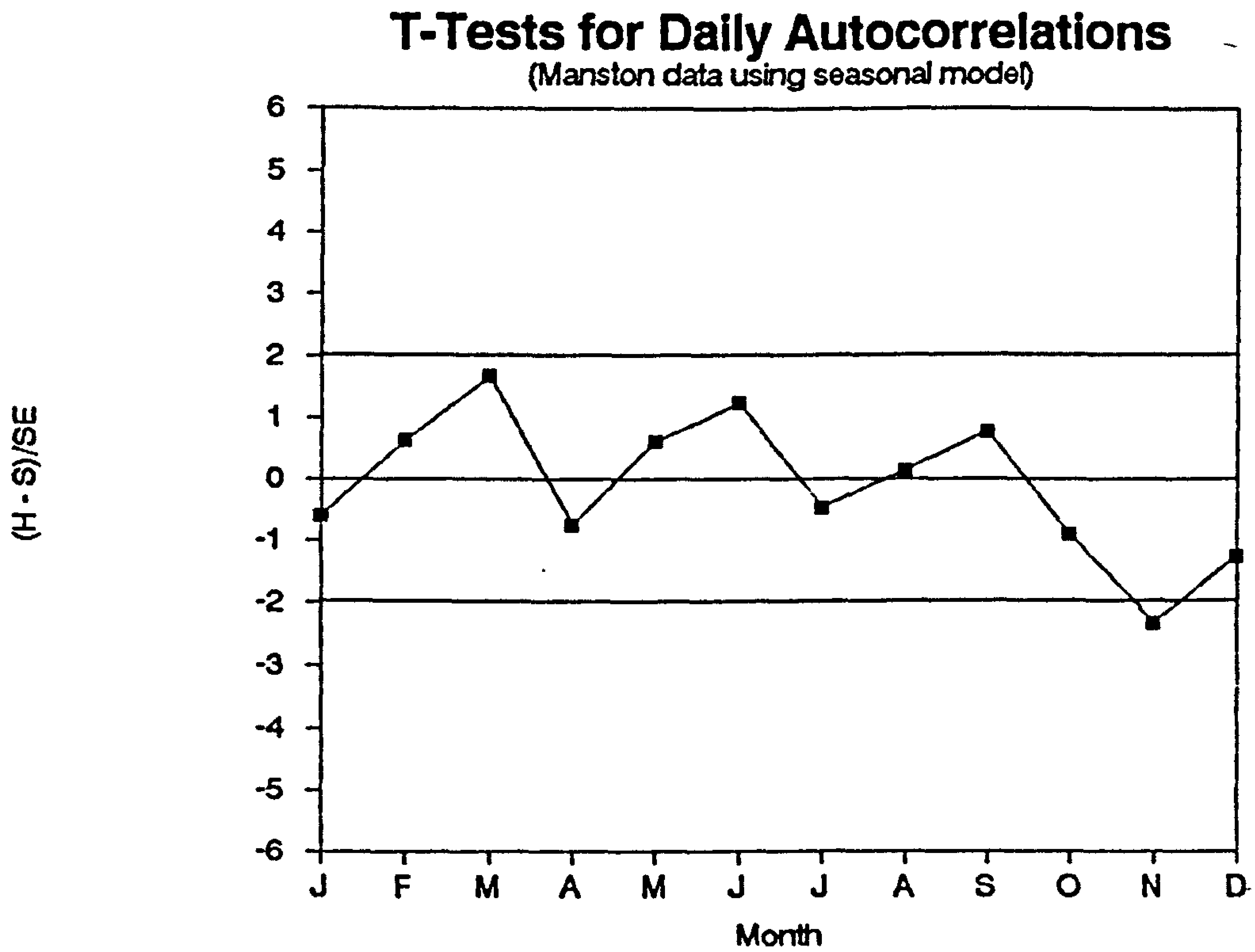


Figure 4.20

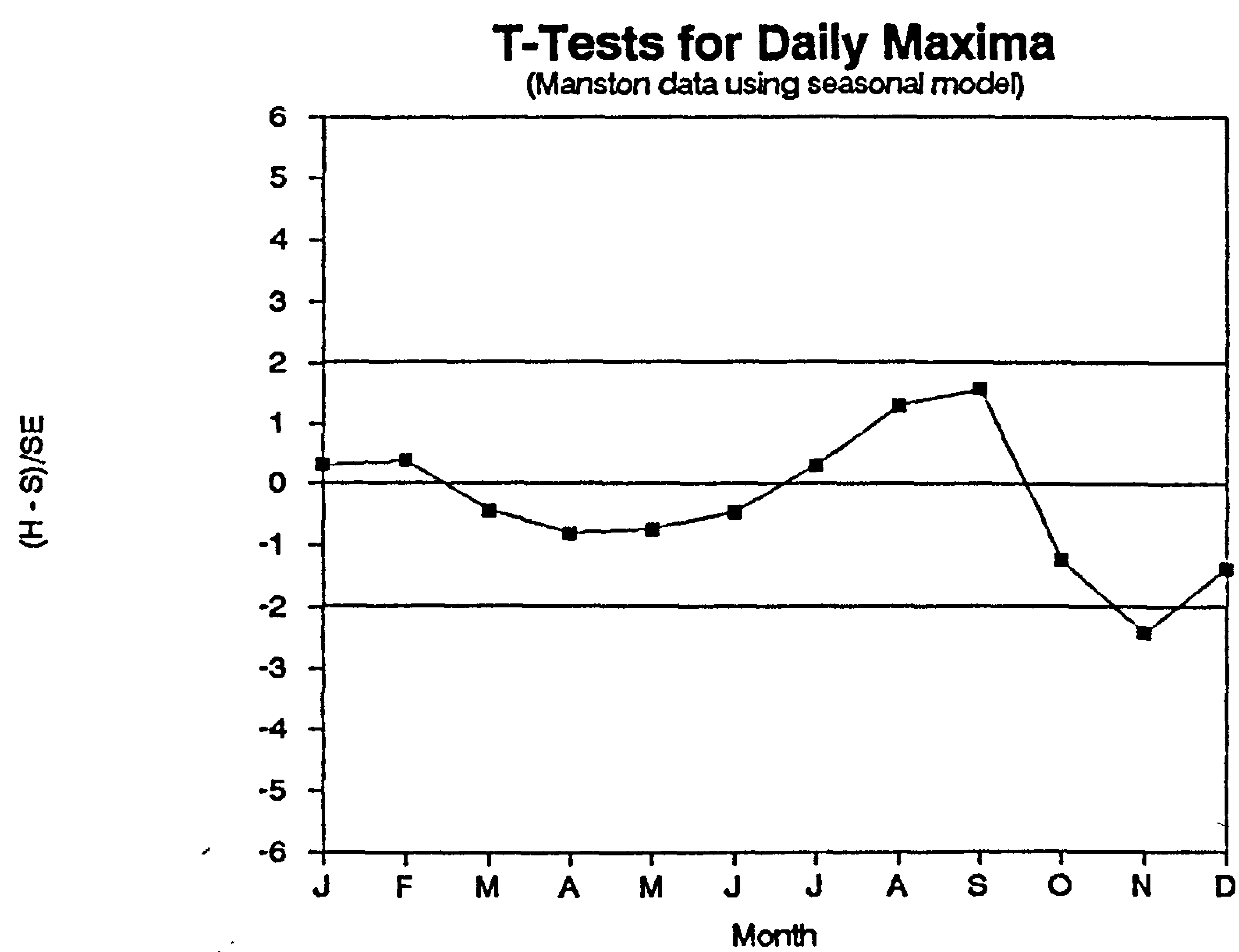


Figure 4.21

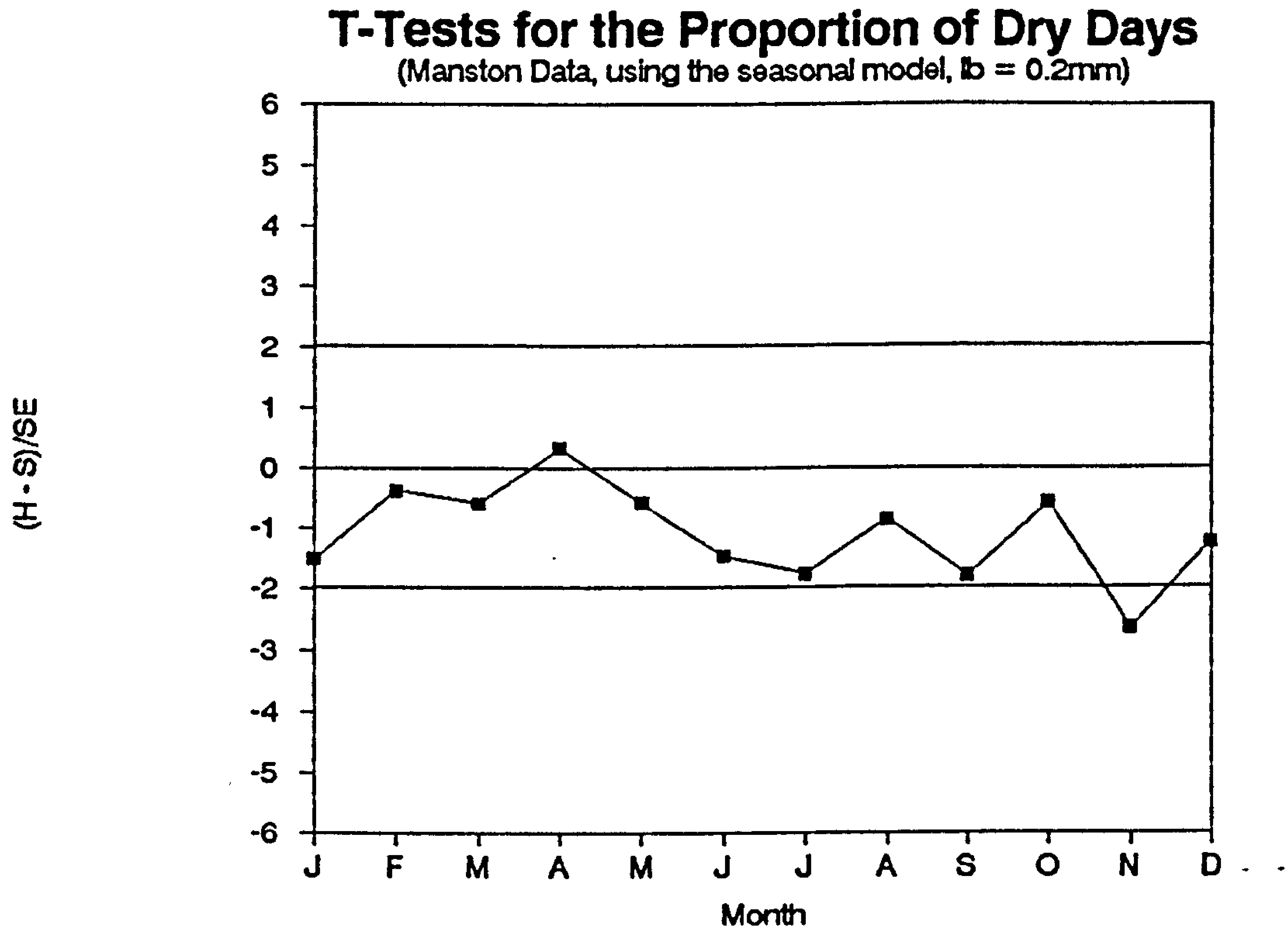


Figure 4.22

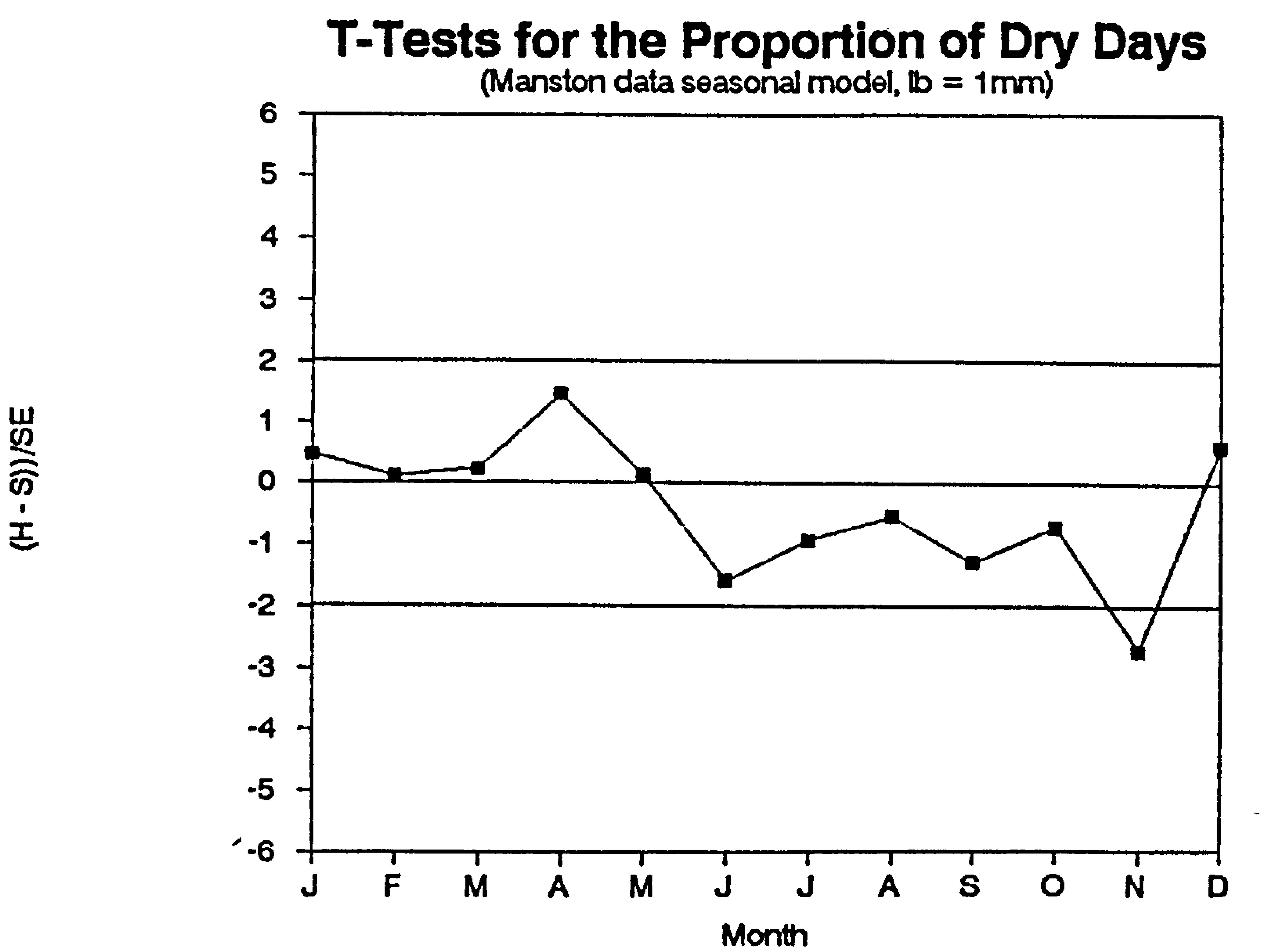
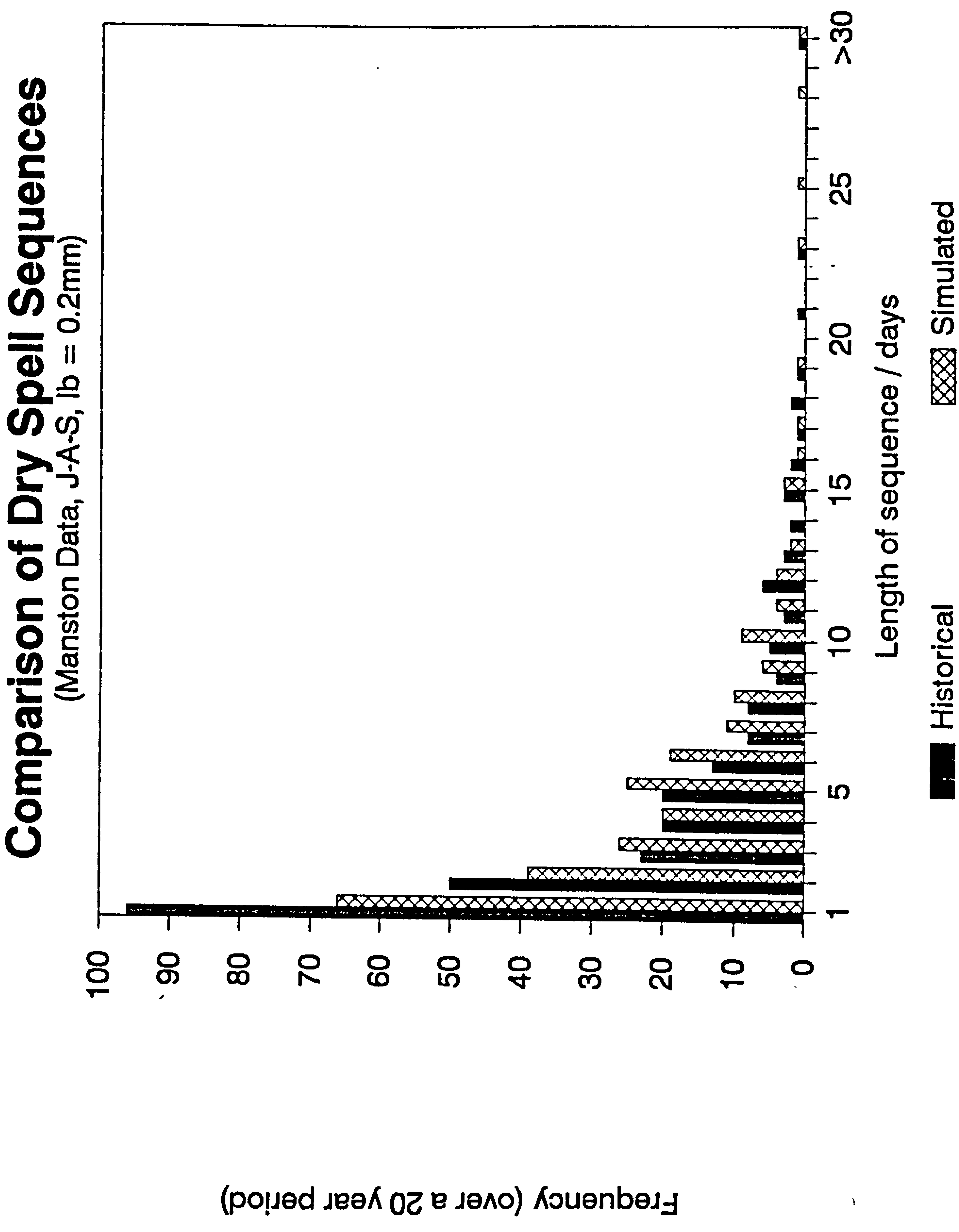


Figure 4.23



Comparison of Dry Spell Sequences
(Manston Data, J-A-S, lb = 0.2mm)
Frequency (over a 20 year period)
0
10
20
30
40
50
60
70
80
90
100
1
5
10
15
20
25
>30
Length of sequence / days
Historical
Simulated

Figure 4.24

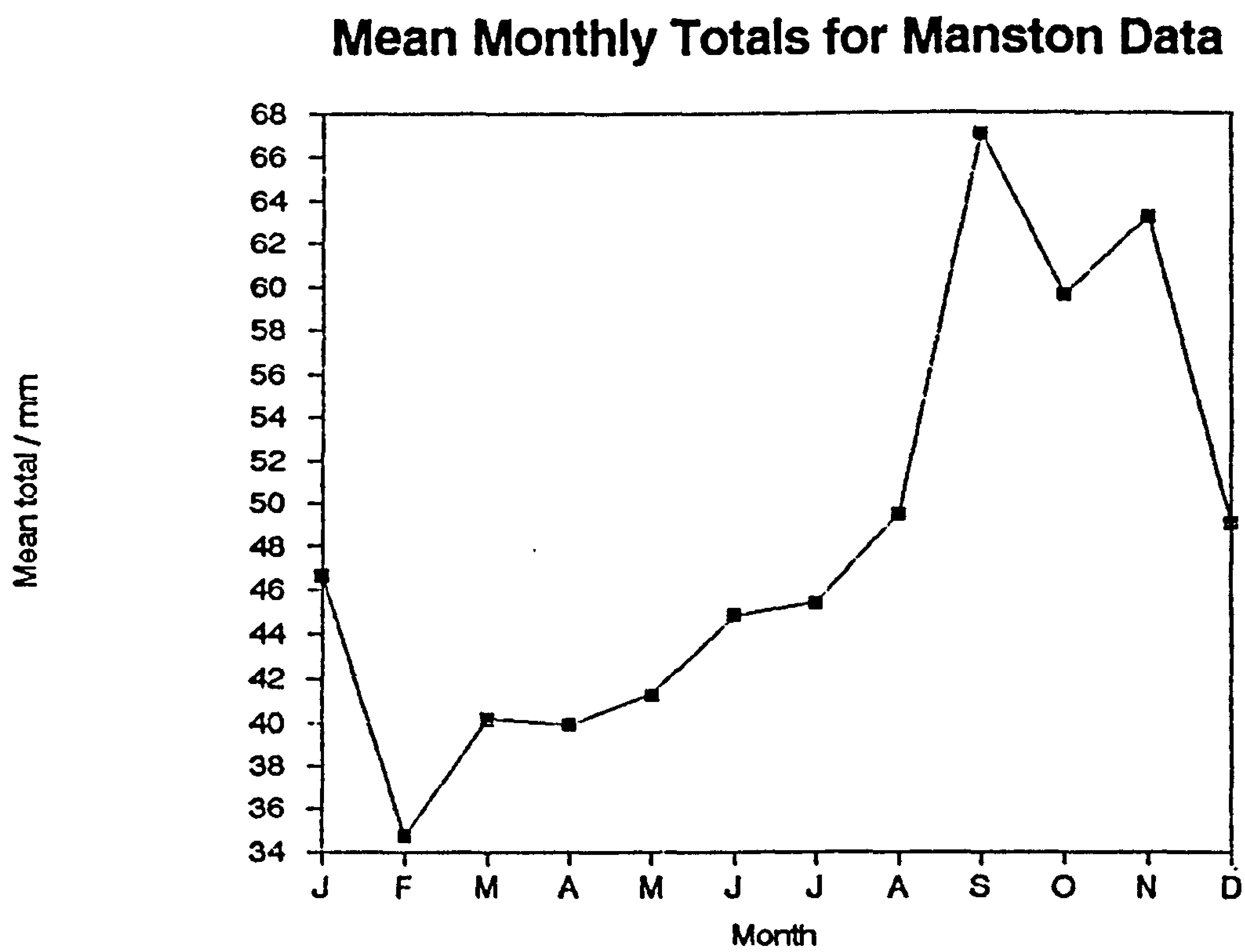


Figure 4.25

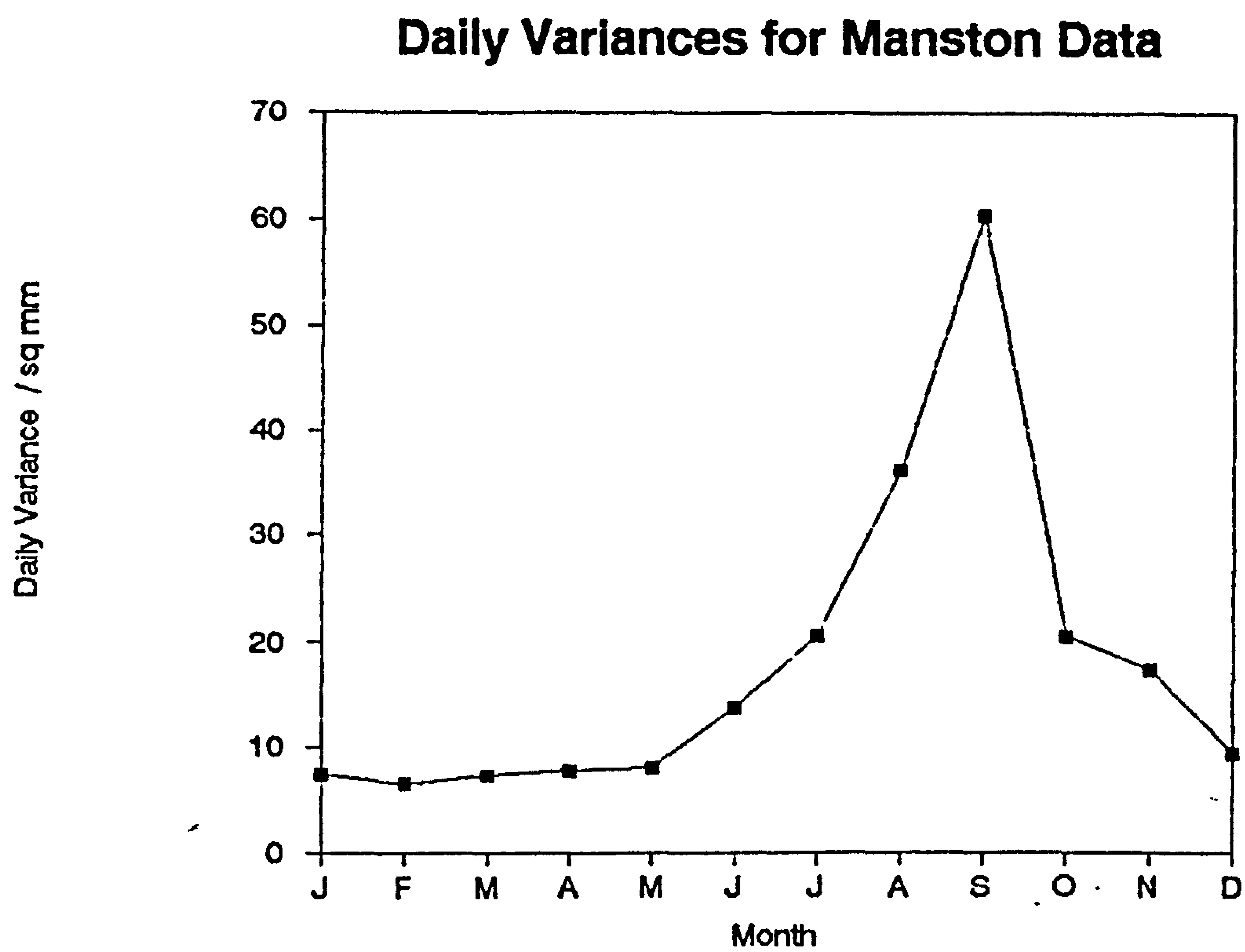


Figure 4.26

4.6 AN IMPROVED HOURLY TIME SERIES MODEL

4.6.1 A further assessment of the statistics to be used in fitting the model

In Section 4.5 many different combinations of statistics were considered for use in the parameter estimation procedure (4.7), and consequently a simplified seasonal model was found convenient for the comparisons. In this Section the simplified seasonal model will be abandoned as (i) less combinations of statistics will need to be considered in fitting the model, and (ii) non-stationarities were found within season 3. In Chapter 6 a revised seasonal/regional model will be developed.

In Section 4.5 it was also found that the lag 1 autocorrelations of the historical 12 and 24 hourly time series were matched (within sampling error) by the model, even though they were not used in the fitting procedure. In this Section we explore the possibility of excluding all autocorrelations from the fitting procedure, and using the transition probabilities as an alternative (equations (4.5) and (4.6)).

Before dispensing with the seasonal model it is worth comparing the results of not using autocorrelations in the fitting procedure with those obtained in the previous Section. Comparing the percentage errors obtained in Table 4.16 with those obtained in Table 4.9(b), it can be seen that using WW1 and WW6 instead of AC1 and AC6, in the fitting procedure improves the model's fit to the daily statistics, particularly for seasons 3 and 4. As might be expected the model no longer matches the overall value of AC1, but this will no longer be of concern as the historical autocorrelations are going to be excluded from the fitting procedure.

Table 4.16
Absolute percentage errors
when using M1, V1, WW1, V6, WW6, V12, WW12, V24, PD24, DD24, and WW24 in the parameter estimation procedure

Statistic		Season 1	2	3	4	mean
M1	*	1.2	1.0	0.1	1.1	0.9
V1	*	4.2	2.5	5.8	3.3	3.9
PD1		0.4	0.5	1.5	3.5	1.5
DD1		0.9	0.5	0.5	1.3	0.8
WW1	*	9.2	4.1	2.9	0.4	4.2
AC1		13.	21.	57.	41.	33.
V3		6.6	8.0	19.	12.	11.
PD3		2.2	1.5	2.4	5.9	3.0
DD3		2.8	1.7	1.0	2.6	2.0
WW3		13.	8.4	7.5	7.1	8.9
AC3		6.9	9.4	8.5	16.	10.
V6	*	6.3	4.4	13.	7.1	7.7
PD6		5.5	3.2	3.6	8.4	5.2
DD6		4.9	3.2	2.0	4.8	3.7
WW6	*	5.1	5.8	4.3	5.5	5.2
AC6		20.	23.	38.	44.	31.
V12	*	0.9	0.2	2.2	4.0	1.8
PD12		2.2	0.7	1.6	6.3	2.7
DD12		3.3	0.4	2.3	5.4	2.8
WW12	*	3.3	4.2	8.5	1.0	4.3
AC12		33.	33.	47.	43.	39.
V24	*	6.4	2.3	14.	6.6	7.3
PD24	*	2.7	3.4	0.8	5.7	3.2
DD24	*	4.4	8.9	0.6	5.1	4.7
WW24	*	1.6	8.1	4.8	2.3	4.2
AC24		35.	69.	26.	49.	45.
Mean		7.5	8.7	11.	11.	9.5

The seasonal model will now be abandoned and comparisons made on a monthly basis. The percentage errors for each month and each statistic will be considered without taking the absolute value, so that a negative percentage error will imply the model is under-estimating the historical statistic and a positive percentage error will imply the model is over-estimating the historical statistic. It should be mentioned that by fitting the model on a monthly basis, it is being assumed that the monthly data are independent from one month to the next, which seems a reasonable assumption for rainfall data (this assumption is examined more closely in Appendix J).

Table 4.17 gives the percentage errors when fitting the model using only daily historical statistics (V24, PD24, DD24, WW24, and M24/24 = M1) in the fitting procedure. As might be anticipated, the model fails to match many of the historical statistics at smaller time steps, e.g. the historical hourly variances. Hence, when it comes to fitting the model to an historical record of daily data (i.e. if there are no hourly data available), it may be better to use an estimate of the hourly statistics needed in the fitting procedure rather than using only the daily statistics.

Table 4.18 gives the percentage errors when using the daily statistics, and the variances and wet given wet transition probabilities of the 1, 3, 6, and 12 hourly time series in the fitting procedure. Comparing Table 4.18 with Table 4.17 an improvement in the model's fit to the historical variances and wet given wet transition probabilities is evident, so that it is advisable to use these historical statistics in the fitting procedure.

In Table 4.19 the 1, 3, 6, and 12 hourly historical wet given wet transition probabilities have been omitted from the fitting procedure. In this Table it is evident that the historical hourly wet given wet transition probability (WW1) is still reasonably well matched by the model despite it being omitted from the fitting procedure. This suggests that, when fitting the model to daily data, an estimate of the historical WW1 may not be required. It is also evident that the model tends to over-estimate the wet given wet transition probabilities and under-estimate the lag 1 autocorrelations for the 3, 6, and 12 hourly historical time series. This suggests that both the within storm variability and storm durations (of less than 1 day) will be slightly greater on average for the data generated by the model than the data of the historical record. However, this difference is unlikely to be of practical significance as the historical hourly and daily transition probabilities are nearly matched by the model. Hence, if only daily data are available, estimates of the 1, 3, 6, and 12 hourly wet given wet transition probabilities may not be required.

In conclusion, the fitting procedure for hourly rainfall data will use the historical hourly mean (M1) and the historical 1, 3, 6, and 12 hourly variances and transition probabilities (V1, V3, V6, V12, WW1, WW3, WW6, WW12) along with the historical daily statistics (V24, PD24, WW24, DD24). However, if only daily data are available, estimates of the 1, 3, 6, and 12 hourly wet given wet transition probabilities (WW1, WW3, WW6, WW12) could probably be left out of the fitting procedure without much detriment to the performance of the model.

Table 4.17

Percentage errors when using only the daily statistics in the fitting procedure

		Month											
		J	F	M	A	M	J	J	A	S	O	N	D
*	M1	0	0	0	0	0	0	0	1	2	0	0	0
	V1	8	11	23	4	-20	-13	2	-38	-34	4	36	19
	PD1	0	1	1	-1	-1	0	1	1	3	2	3	1
	DD1	1	0	0	-0	0	0	1	1	2	1	1	1
	WW1	5	0	-3	1	9	3	10	27	20	-3	-2	5
	AC1	6	16	3	3	34	56	49	143	72	35	8	8
	V3	8	17	17	3	-12	4	26	7	-13	19	38	17
	PD3	1	2	1	-1	-1	0	3	3	6	3	5	3
	DD3	3	2	2	1	1	1	1	2	4	1	3	3
	WW3	15	10	8	13	19	13	-9	12	10	-3	2	10
	AC3	1	-23	-21	-18	20	10	-12	89	8	-10	-19	-18
	V6	12	11	10	-2	-12	8	21	20	-10	15	34	11
	PD6	4	3	3	1	1	2	4	6	10	5	7	7
	DD6	6	5	4	4	4	3	1	4	6	3	7	5
	WW6	17	18	11	20	22	19	-18	-1	-6	0	7	1
	AC6	1	-30	-19	-7	15	-22	-36	5	-28	-36	-41	-27
	V12	11	11	2	-4	-10	7	1	25	-13	5	21	3
	PD12	1	-0	-1	-0	-2	1	2	6	11	1	4	4
	DD12	8	5	4	6	3	2	2	6	7	3	5	3
	WW12	20	23	16	25	25	15	3	10	-9	10	2	-2
	AC12	-12	-44	-20	24	70	-44	-21	-40	-33	-39	-54	-31
*	V24	0	0	0	0	0	0	0	-4	-9	0	0	0
*	PD24	-0	0	-0	1	0	0	-0	9	13	0	-0	-0
*	DD24	0	-0	0	-1	-0	-0	0	6	10	-0	0	0
*	WW24	-0	0	-0	1	-0	0	0	-3	-7	0	-1	-1
	AC24	-23	-68	-53	-25	-37	-73	-7	122	-33	-64	-64	-26

* = used in fitting,

- implies the model under-estimates the historical statistic,

+ implies the model over-estimates the historical statistic.

Table 4.18

Percentage errors when using variances, transition probabilities and daily statistics in the fitting procedure

	J	F	M	A	M	J	J	A	S	O	N	D
* M1	-2	-2	-2	-2	-1	-1	-0	1	2	-1	-3	-2
* V1	4	-0	0	1	0	-6	-6	-18	-6	-7	-1	2
PD1	1	1	-0	-0	1	0	1	2	4	2	1	0
DD1	1	1	0	0	0	-0	1	1	2	1	1	1
* WW1	9	5	5	4	1	-4	11	1	5	-0	7	6
AC1	10	27	23	17	23	44	45	83	48	39	23	19
* V3	6	11	7	6	4	7	13	16	9	9	9	7
PD3	3	2	1	0	1	0	2	3	7	3	3	2
DD3	3	2	1	1	1	0	1	2	4	1	2	2
* WW3	10	7	6	4	2	2	-1	-6	-2	0	7	8
AC3	-6	-19	-2	-13	-2	-7	-14	7	-22	-5	-2	-8
* V6	8	6	6	3	-0	7	8	10	2	7	12	4
PD6	6	4	2	1	3	1	4	5	11	4	4	4
DD6	5	4	2	2	2	1	2	3	7	3	5	3
* WW6	3	7	4	5	1	6	-6	-4	-5	2	10	2
AC6	-24	-39	-14	-18	-17	-38	-35	-41	-49	-35	-33	-18
* V12	2	3	-0	-1	-4	3	-9	2	-9	-2	3	-1
PD12	1	-2	-4	-2	-1	-2	1	4	13	-0	0	0
DD12	3	-0	-1	0	-1	-1	2	4	10	2	1	1
* WW12	5	7	9	8	3	3	9	14	-0	8	4	2
AC12	-45	-61	-25	-8	8	-58	-22	-60	-42	-42	-52	-18
* V24	-12	-10	-3	-1	3	-8	-10	-26	-7	-7	-14	-3
* PD24	-5	-6	-7	-6	-3	-5	0	5	18	-2	-7	-6
* DD24	-9	-12	-8	-11	-8	-7	-1	3	14	-3	-6	-3
* WW24	-4	-5	-1	-5	-8	-4	-4	-9	-3	-3	4	5
AC24	-53	-80	-58	-48	-62	-80	-26	31	-36	-69	-63	-9

* = used in fitting,

- implies the model under-estimates the historical statistic,

+ implies the model over-estimates the historical statistic.

Table 4.19

Percentage errors when using the variances and daily statistics in the fitting procedure

	J	F	M	A	M	J	J	A	S	O	N	D
* M1	-0	0	0	0	0	0	0	1	2	0	-0	-0
* V1	-0	-2	-2	-1	0	-1	-11	-1	-7	-4	-4	-2
PD1	-1	-0	-1	-1	0	0	1	2	4	1	-0	-0
DD1	0	0	0	0	0	-0	1	1	2	0	1	1
WW1	6	5	7	4	-2	-5	14	-2	6	-2	9	13
AC1	3	16	19	10	16	19	50	-12	51	26	20	21
* V3	-1	4	3	2	1	2	11	5	9	6	4	4
PD3	-0	0	-0	-1	1	0	2	4	7	2	1	2
DD3	3	2	2	1	1	1	2	3	4	2	3	3
WW3	20	19	18	14	7	16	5	36	-4	8	17	17
AC3	3	-17	-1	-13	-2	-8	-7	15	-20	-9	1	-2
* V6	3	0	2	-1	-3	2	8	2	3	4	8	3
PD6	3	2	2	1	2	2	4	7	11	4	4	6
DD6	7	6	5	4	3	4	3	7	6	5	8	5
WW6	23	26	17	20	13	26	1	46	-11	13	24	7
AC6	9	-22	-3	-5	-3	-24	-27	-3	-49	-28	-22	-13
* V12	3	2	-2	-2	-4	1	-7	4	-9	-4	2	-1
PD12	0	-1	-1	0	-1	1	2	10	13	2	2	3
DD12	9	7	5	6	2	4	4	11	8	6	7	4
WW12	26	29	19	24	19	20	19	58	-7	18	16	3
AC12	-2	-37	-11	24	50	-43	-8	-31	-47	-30	-38	-18
* V24	-5	-6	-2	2	6	-5	-7	-18	-8	-6	-12	-3
* PD24	-0	1	0	1	-0	2	3	19	16	4	-0	0
* DD24	2	2	1	-1	-2	2	3	15	11	4	5	2
* WW24	2	3	1	0	-2	2	1	5	-9	2	6	2
AC24	-13	-65	-50	-26	-43	-72	-9	89	-44	-62	-50	-12

* = used in fitting,

- implies the model under-estimates the historical statistic,

+ implies the model over-estimates the historical statistic.

4.6.2 The t-tests

Using the parameter estimates in Table 4.20 below, 20 years of hourly rainfall data were simulated using the simulation program (Appendix B). As with the previous Section, t-ratios were plotted against the month (some selected plots are given in Figures 4.27 - 4.36 and the complete set of plots are provided in Appendix E). From these plots it is evident that the overall performance of the model is good.

Table 4.20

Parameter estimates obtained when using V1, V3, V6, V12, V24, WW1, WW3, WW6, WW12, WW24, DD24, PD24 in the fitting procedure

Month	λ (h^{-1})	β (h^{-1})	η (h^{-1})	ν	ξ (h/mm)
Jan	0.0242	0.4546	1.2041	4.2445	1.3882
Feb	0.0206	0.2795	0.9567	3.1822	1.3572
Mar	0.0189	0.1413	0.7312	2.6734	1.3117
Apr	0.0182	0.1986	0.8258	2.9188	1.1790
May	0.0192	0.2847	1.2637	2.6089	0.7196
Jun	0.0143	0.2199	1.1270	2.9734	0.6129
Jul	0.0117	0.1514	0.6868	1.9997	0.5582
Aug	0.0085	0.1055	0.8883	2.4030	0.3424
Sep	0.0055	0.0500	0.6088	3.5915	0.3397
Oct	0.0135	0.1306	0.8447	3.5659	0.7181
Nov	0.0226	0.2781	0.7768	3.2201	1.0969
Dec	0.0202	0.2338	0.8911	3.4145	1.1941

Many of the statistics for the month of June were found to be significant in the tests. However, with so many t-tests being performed one would expect 1 in 20 significant values (out of those tests that are independent) at the 5% level even if the model fitted the historical data perfectly. In the case of June, the monthly total t-ratio was close to the 5% level. This value almost certainly occurred by chance, because in the fitting procedure the model is made to match the historical 1 hourly means (M1) almost exactly (and hence match the historical monthly totals). With the h hourly simulated means being consistently less than their historical equivalents, it follows that the simulated variance and maxima will probably also be consistently less than their historical equivalents. This highlights the importance of matching the historical hourly mean, M1, almost exactly in the fitting procedure (4.7).

Figure 4.29 provides evidence to suggest that the model tends to overestimate the lag 1 hourly autocorrelations. This was not surprising as AC1 was omitted from the fitting procedure. However, the differences between the historical and simulated lag 1 hourly autocorrelations are unlikely to be of practical significance, because the largest mean difference (H - S in the plots of the t-ratios) was close to zero (-0.15 for October (see Figure E.29 in Appendix E) - correlations of magnitude less than 0.2 were regarded as not practically significant).

4.6.3 The dry spells and proportion of dry days

The frequency plot of the dry spell sequences for July, August and September is given in Figure 4.37 (the plots for the other months are provided in Appendix E). Comparing Figure 4.37 with Figure 4.12, it is evident that using the daily transition probabilities in the fitting procedure improves the model's fit to the historical dry spells. Furthermore, by comparing Figures 4.35 and

4.36 with Figures 4.8 and 4.9 respectively, it is evident that there is an improvement in the model's fit to the historical proportion of dry days.

4.6.4 An extreme value for August

Some difficulty occurred with the month of August where the model consistently under-estimated the mean variance and mean maxima of all the historical h hourly rainfall time series, as well as the standard deviation of the h hourly variances and maxima (see Figures E.27 - E.60 in Appendix E). To find the reason for this, the maxima were plotted on a histogram (see Figure 4.39), from which it can be seen that in distribution the model's fit to the historical data is quite good. There was, however, one historical value (46.1mm) which was much greater than all the others, and it seemed likely that this value would make the historical variances and maxima greater than the simulated variances and maxima for August. This very unusual value probably has a high return period, so at this stage the results will not be regarded as significant. The extreme values will be considered in more detail in Chapter 5, where the model will be fitted to the five longest records of historical daily rainfall data.

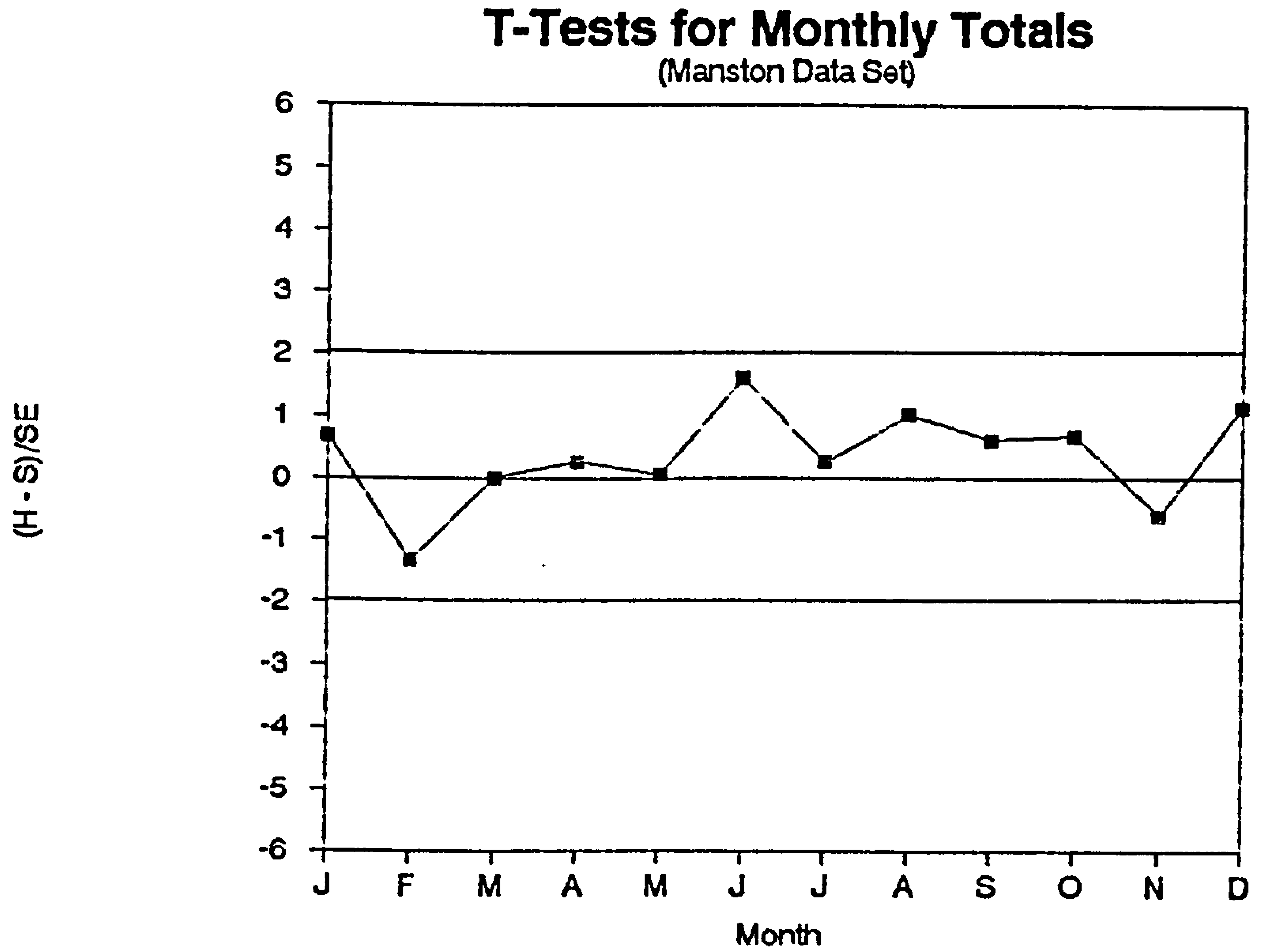


Figure 4.27

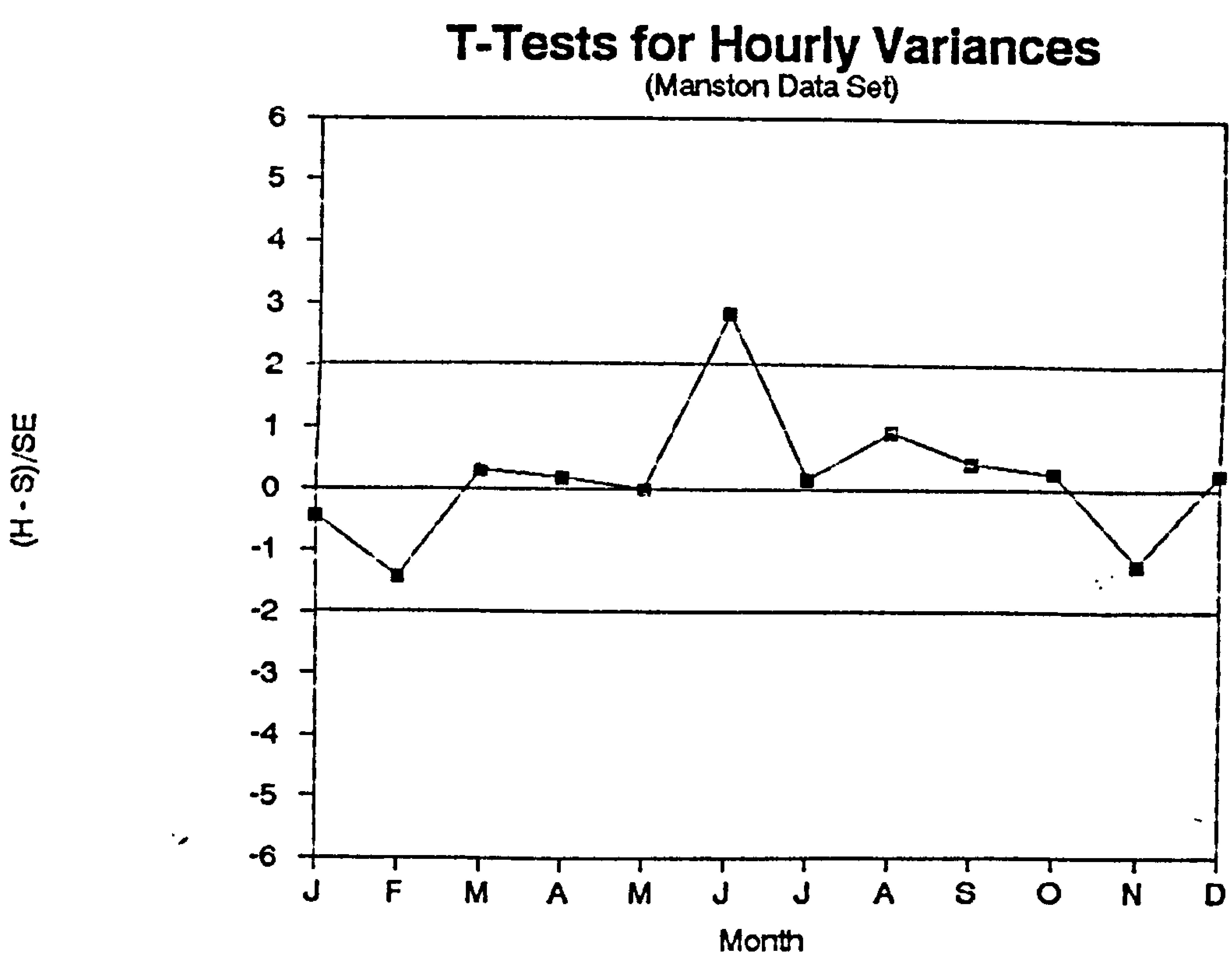


Figure 4.28

H = Historical value, S = Simulated value,
SE = Standard Error

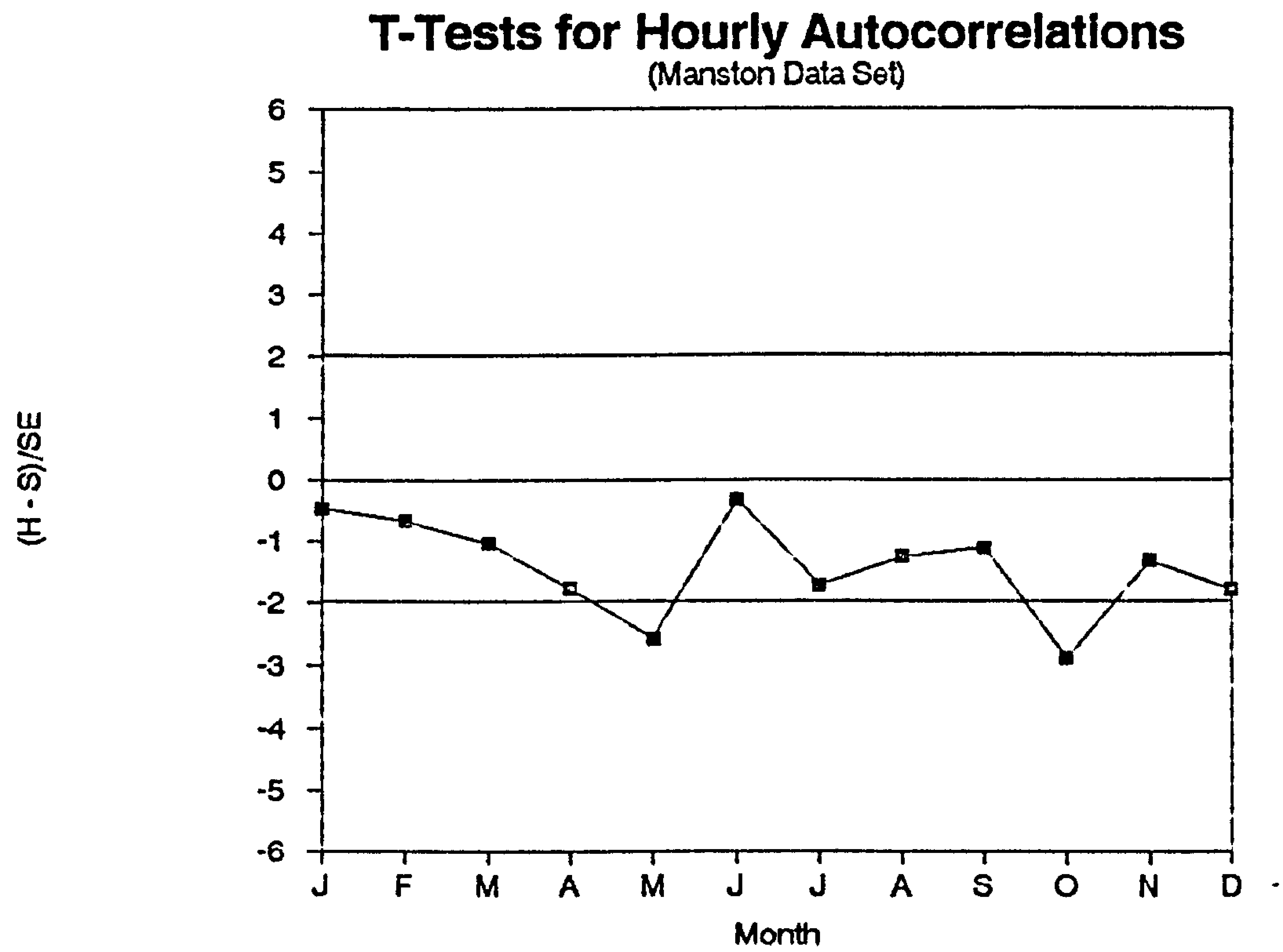


Figure 4.29

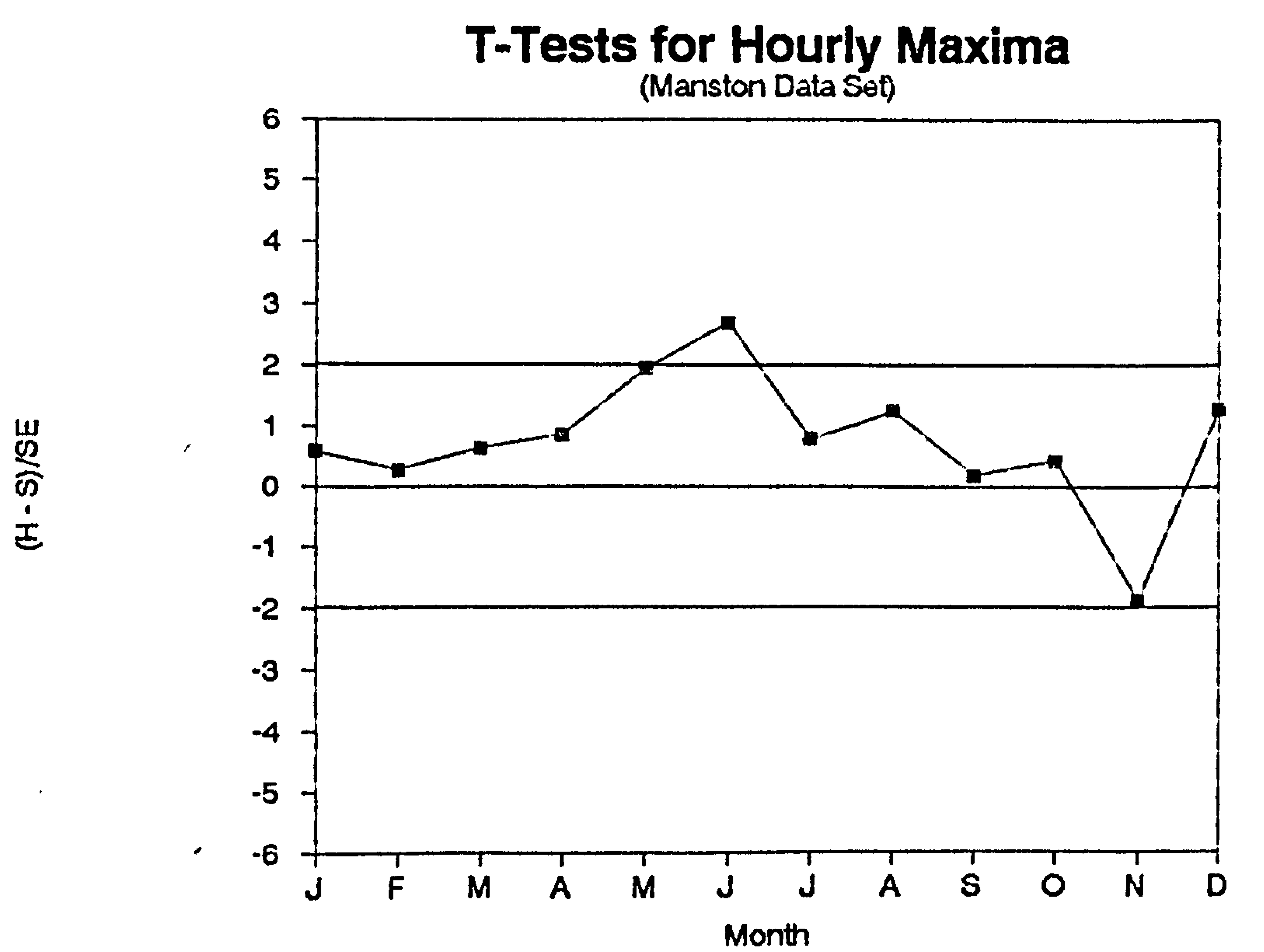


Figure 4.30

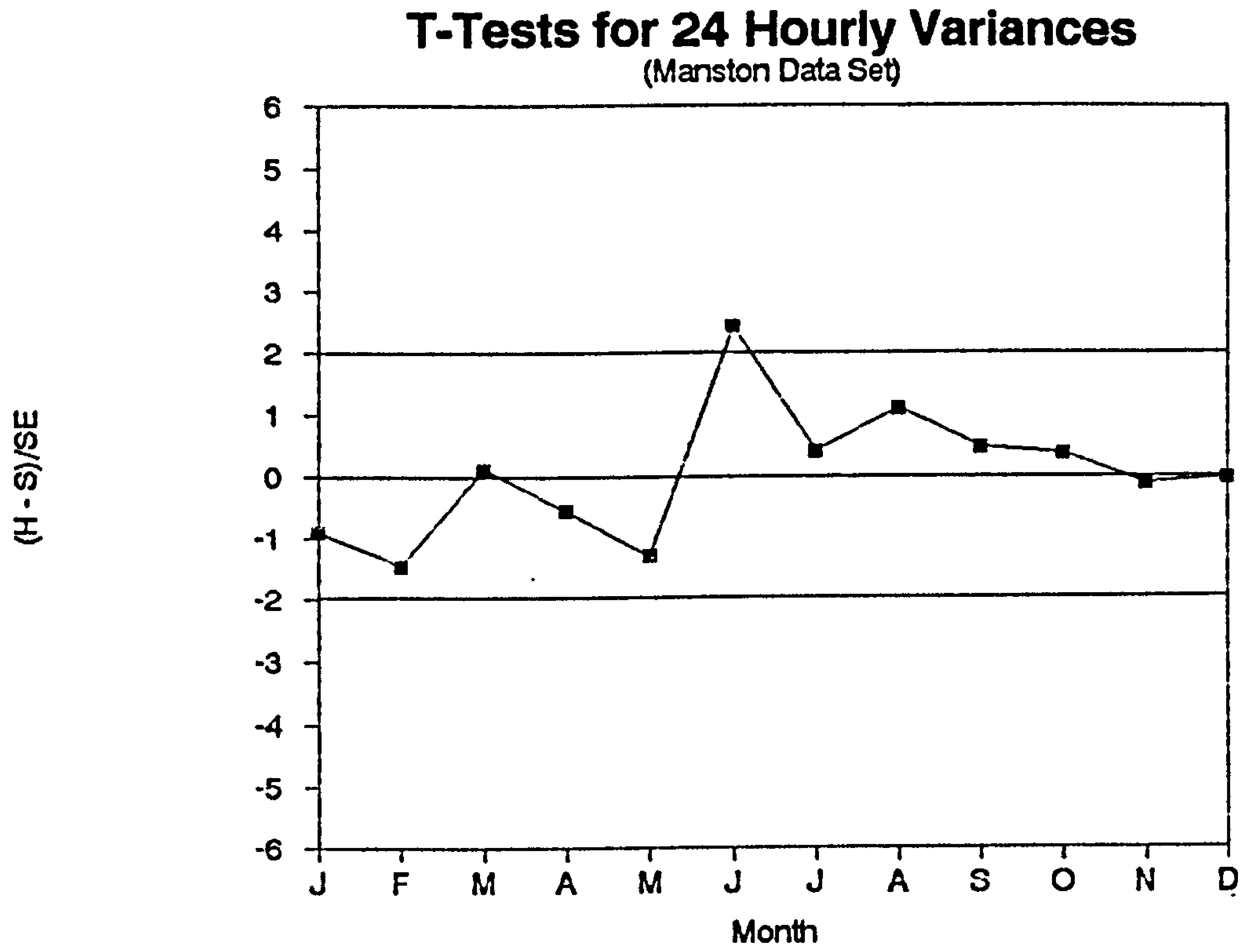


Figure 4.31

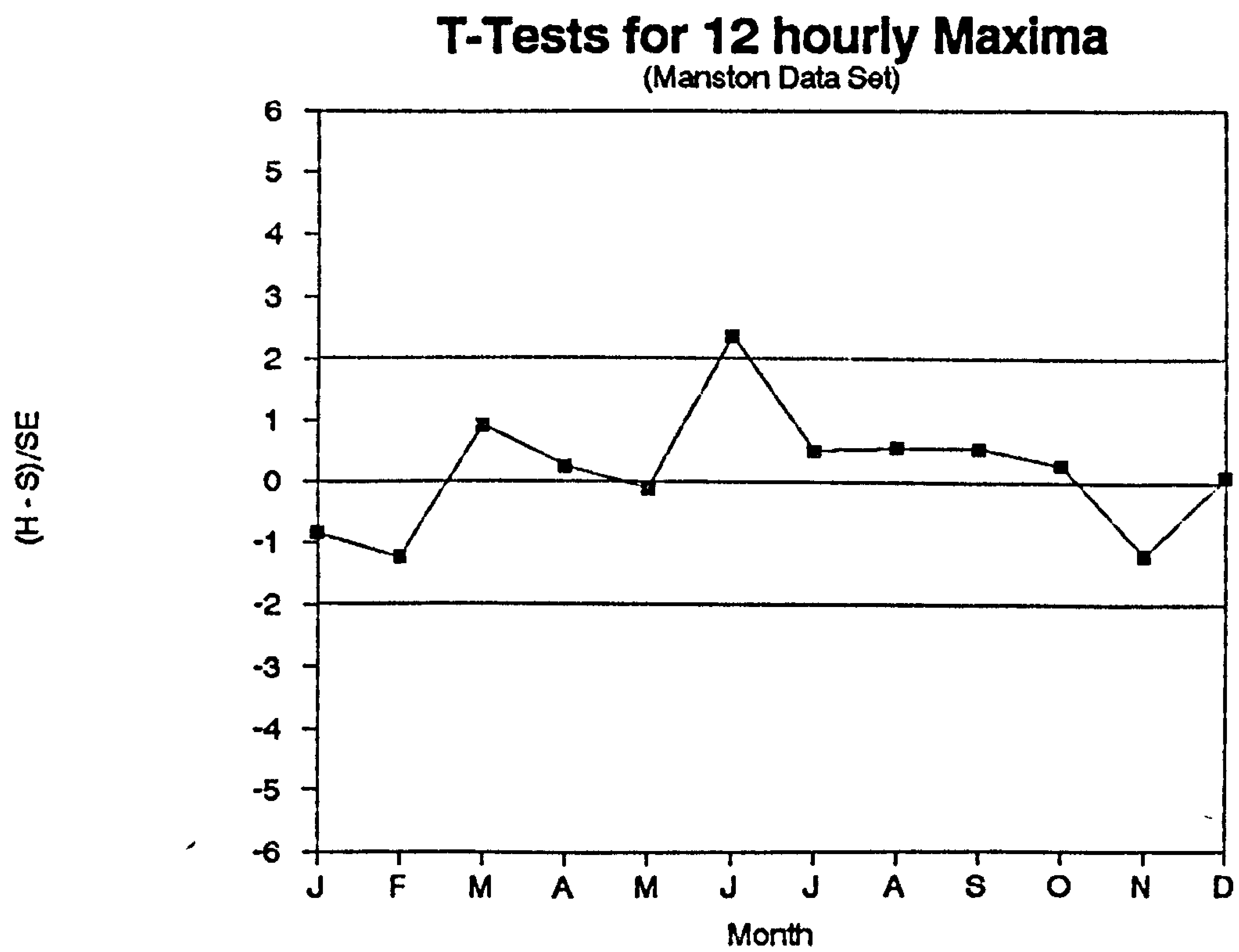


Figure 4.32

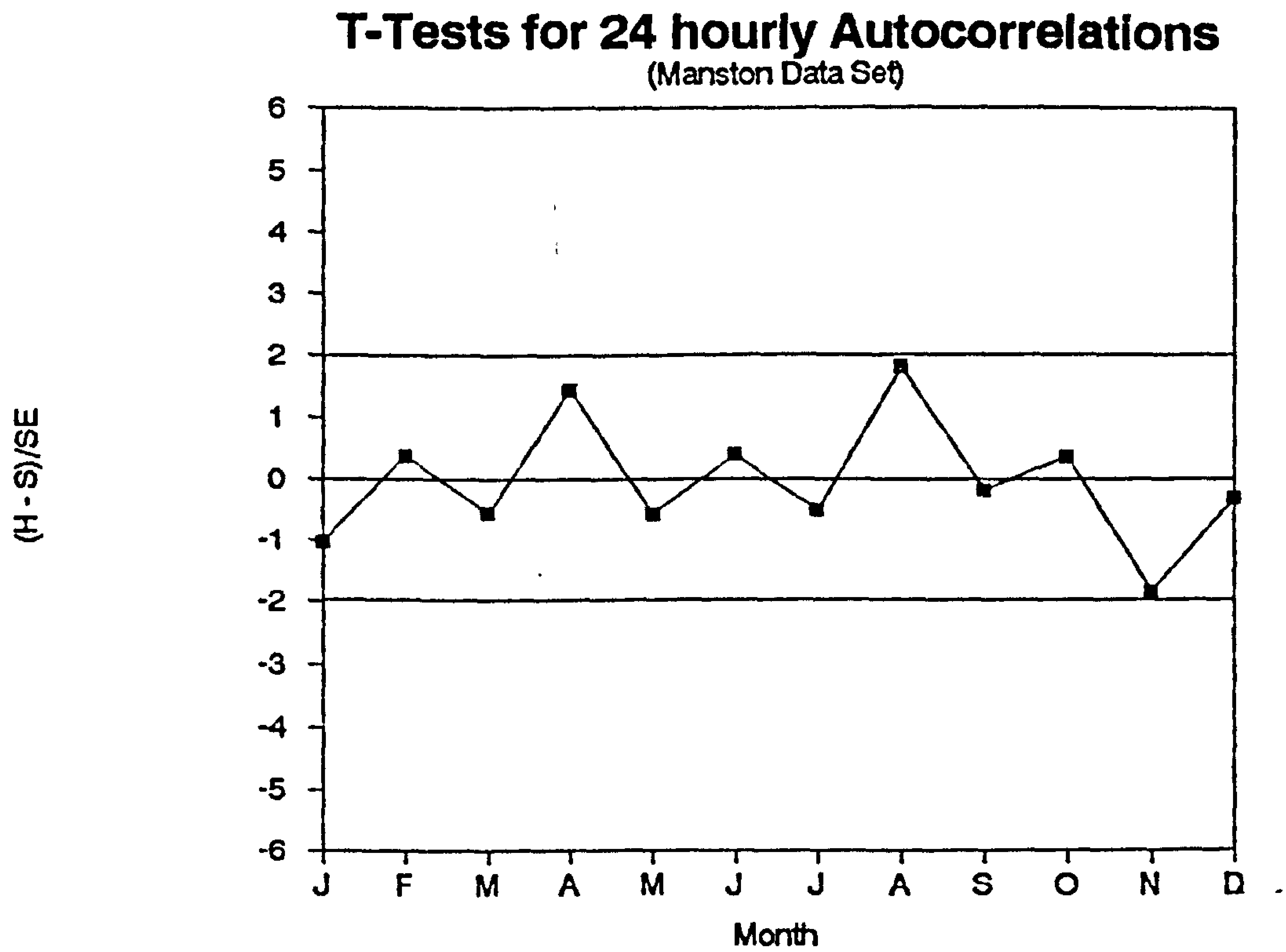


Figure 4.33

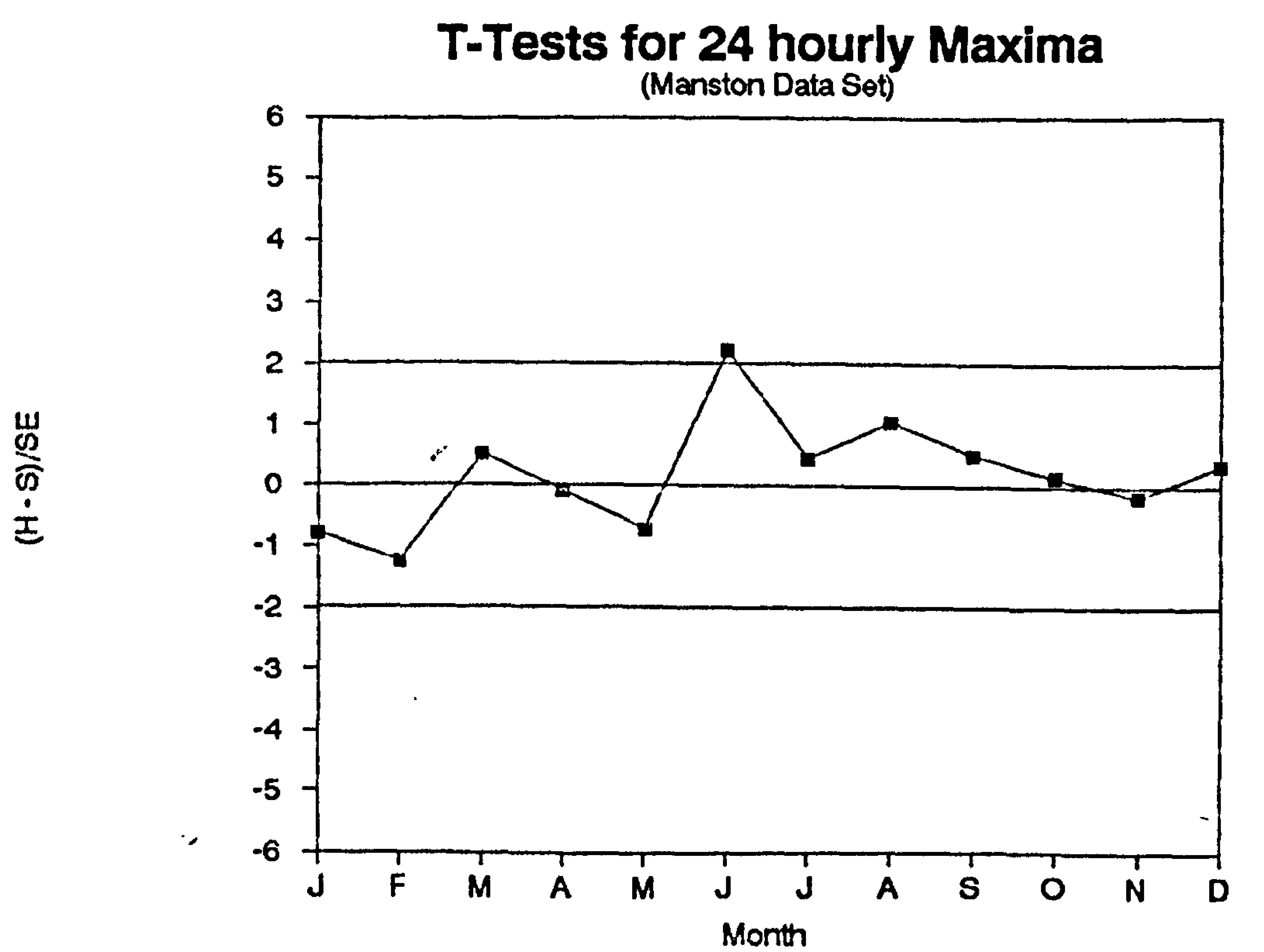


Figure 4.34

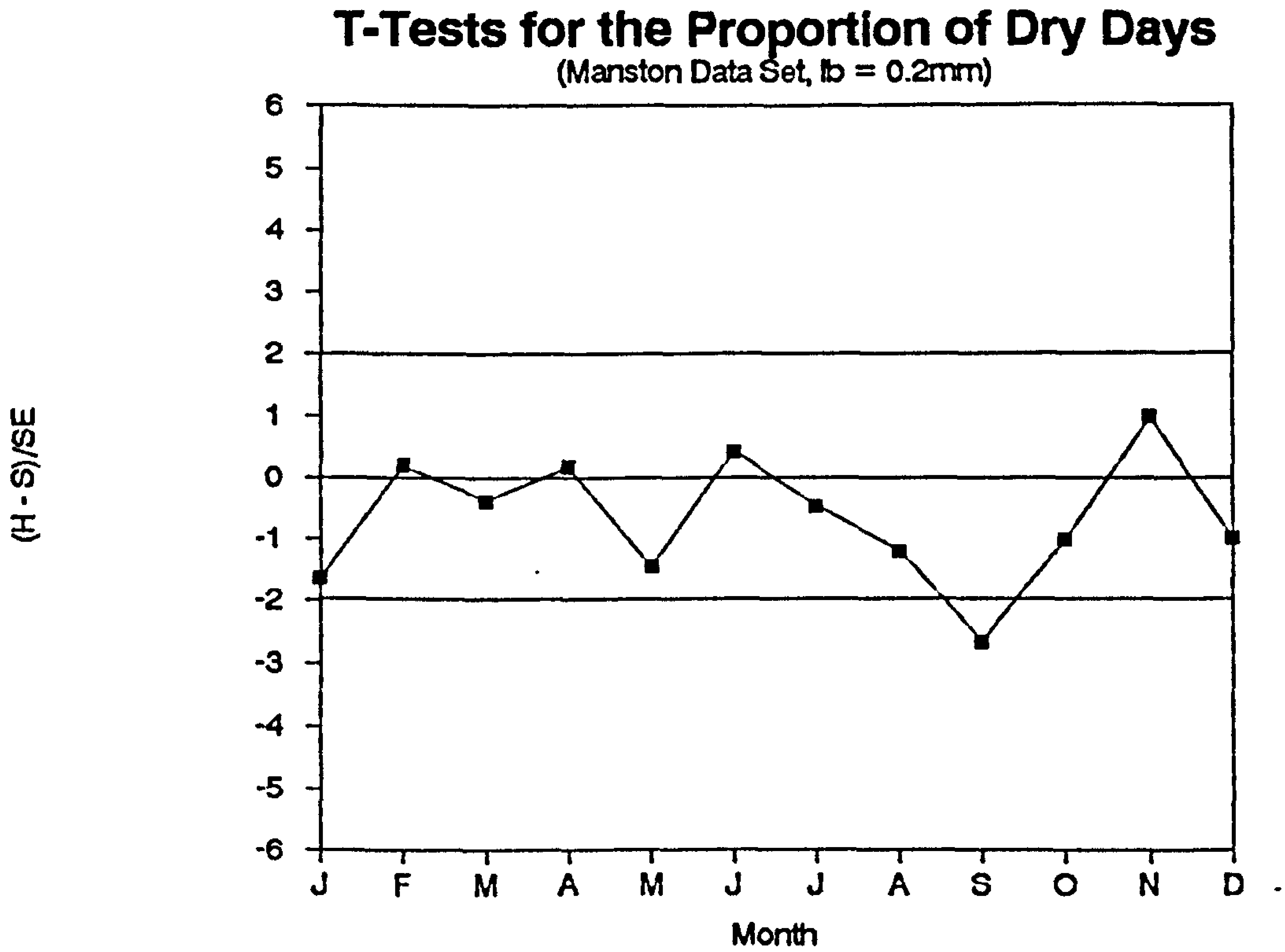


Figure 4.35

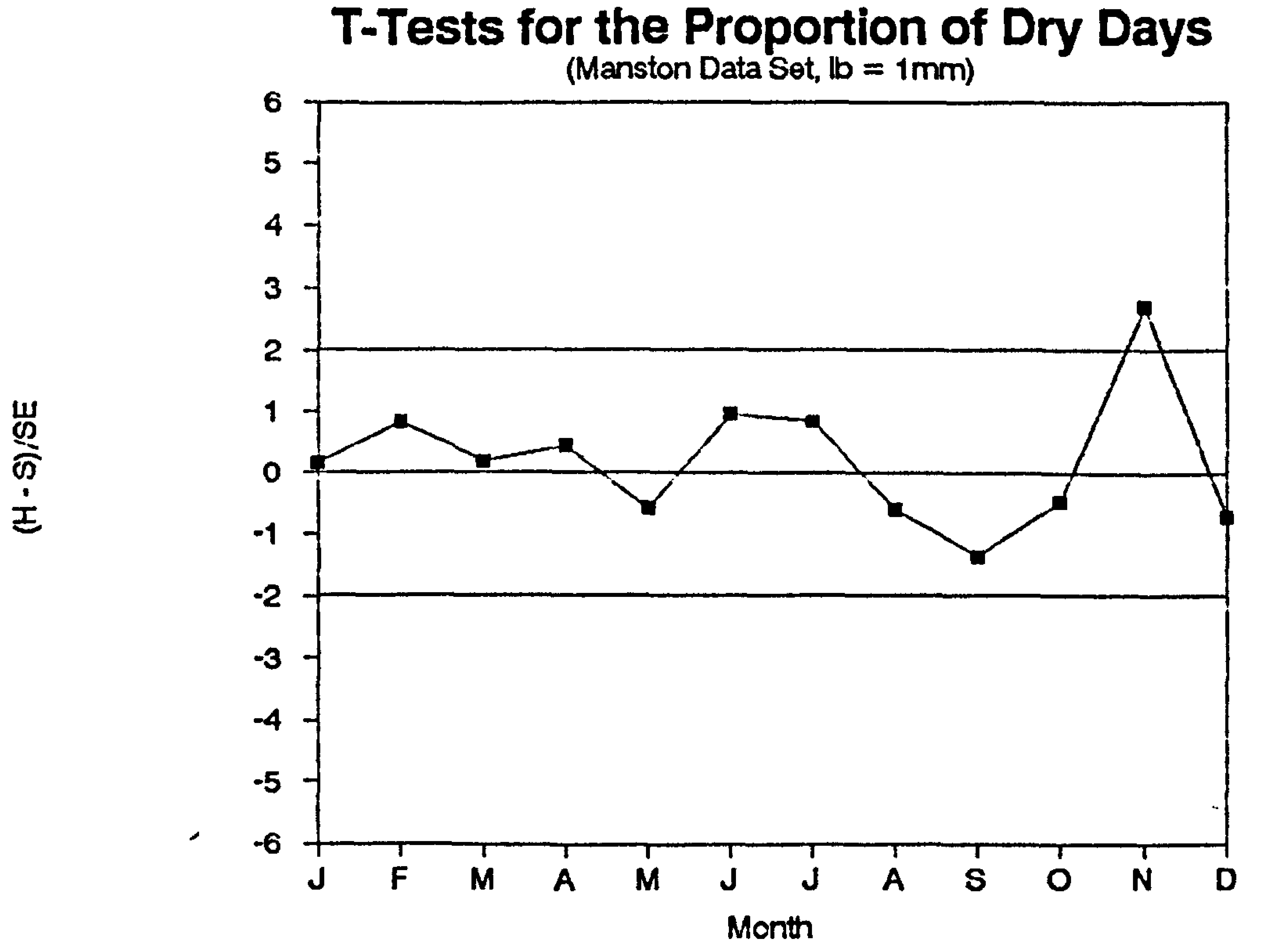


Figure 4.36

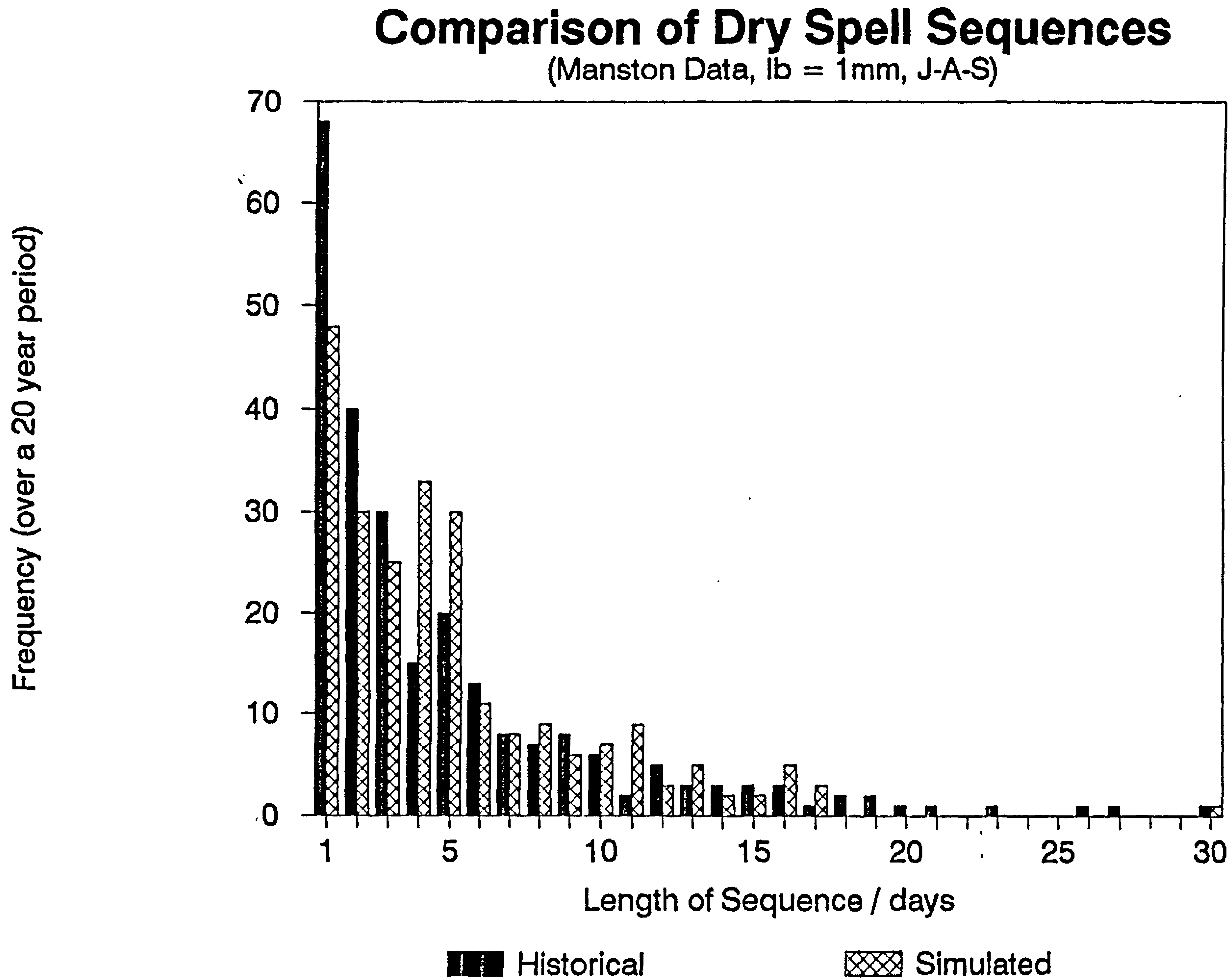
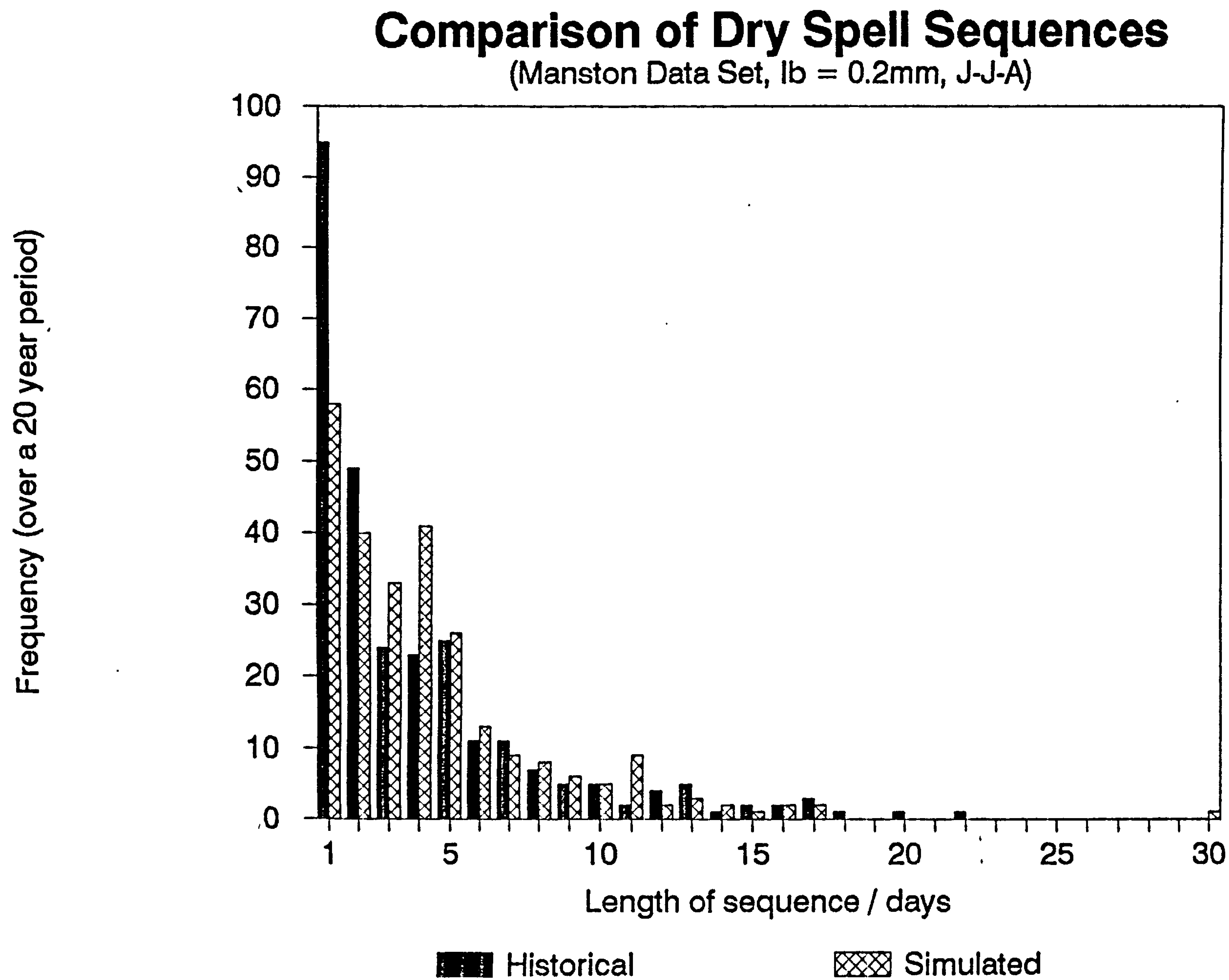

Comparison of Dry Spell Sequences
(Manston Data, lb = 1mm, J-A-S)
Frequency (over a 20 year period)
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of Sequence / days
Historical
Simulated

Figure 4.37



Comparison of Dry Spell Sequences
(Manston Data Set, lb = 0.2mm, J-J-A)
Frequency (over a 20 year period)
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of sequence / days
Historical
Simulated

Figure 4.38

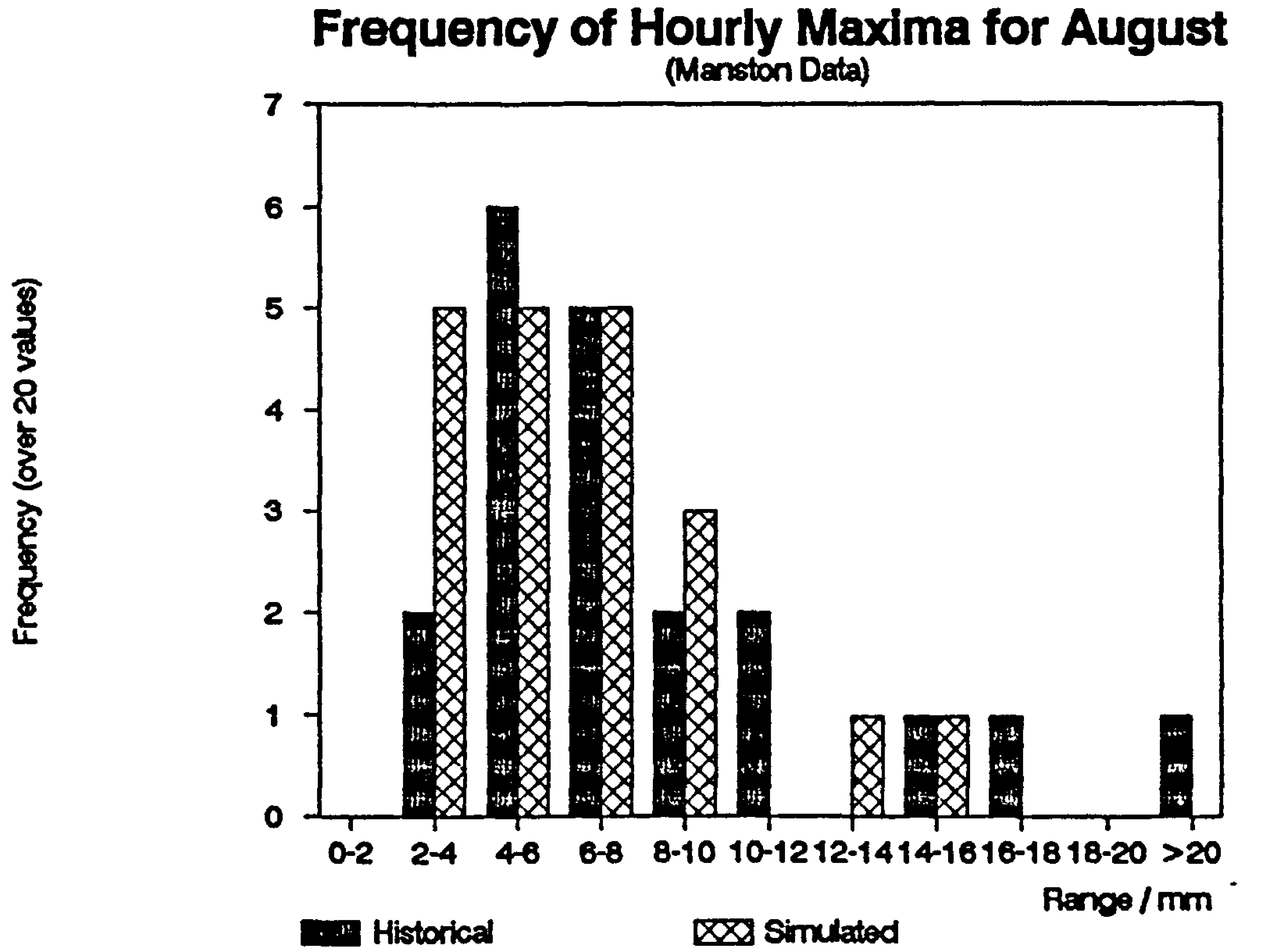

Frequency of Hourly Maxima for August
(Manston Data)
Frequency (over 20 values)
7
6
5
4
3
2
1
0
0-2
2-4
4-6
6-8
8-10
10-12
12-14
14-16
16-18
18-20
>20
Range / mm
Historical
Simulated

Figure 4.39

4.7 SOME FURTHER WORK IN MODEL VALIDATION

To some extent the model has been validated in the process of developing the fitting procedure. For example, comparisons have been made between historical and simulated statistics that were not used to fit the model (e.g. the mean and standard deviations of the maximum h hourly rainfalls (for h = 1, 3, 6, 12, and 24)). In this Section, some more comparisons are made between historical and simulated statistics that were not used in the fitting procedure.

4.7.1 The proportions of hourly rainfalls exceeding certain bounds

For each month of both the simulated and historical hourly rainfall time series the proportion of rainfall exceeding bounds between 0 and 10mm were found for each year. The means and standard deviations of these values were found for both the historical and simulated time series, and plotted against the month (see Figures 4.40 - 4.53).

From Figures 4.42 - 4.53 it is evident that the model follows the mean and standard deviations of the historical proportions of rainfall greater than 1mm, 2mm, 3mm, 4mm, 5mm and 10mm closely (probably within sampling error). However, in Figures 4.40 and 4.41 it is evident that the model tends to under-estimate the mean and standard deviation of the historical proportion of wet hours (i.e. hours with rainfall greater than 0mm). A possible explanation for this is that the model does not generate very light rainfall (i.e. drizzle), which would be recorded in the historical rainfall record. Drizzle is unlikely to be of importance in the designing of a sewage system, as it can easily be lost through evaporation, and so, overall, the results are regarded as satisfactory.

4.7.2 Time series plots of daily rainfall data

Figures 4.54 - 4.57 were prepared to illustrate visually the historical and simulated daily time series. January and July were selected to be representative of Winter and Summer respectively. For the historical and simulated time series the months for each year were concatenated to form a 20 year record of Januarys and Julys. Note that the order in which the years appear should be ignored, because, for example, the simulated January time series for year 1 is not meant to represent the historical January time series for year 1 (the overall simulated January series, however, is meant to represent the overall historical January series).

The simulated time series (Figures 4.55 and 4.57) compare favourably with their corresponding historical time series (Figures 4.54 and 4.56 respectively), with the exception that the simulated January series has more extreme values (i.e. daily rainfall exceeding 20mm) than the historical January series. No conclusions will be drawn about the extreme values (i.e. whether more extreme values occurred for the model by chance) in this Chapter as a more complete extreme value analysis of the model will be presented in the next Chapter when the model will be fitted to the longest records of daily data.

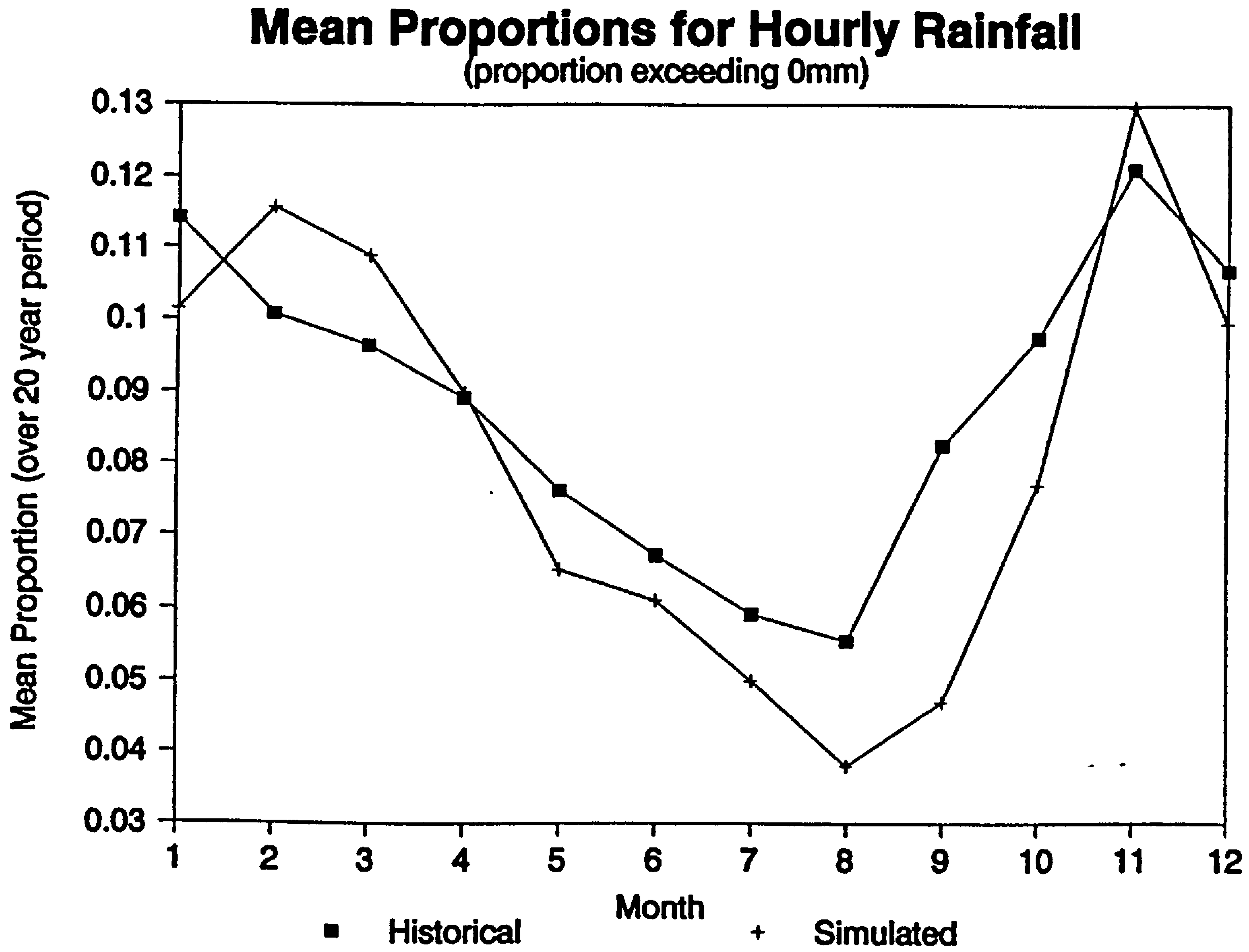


Figure 4.40

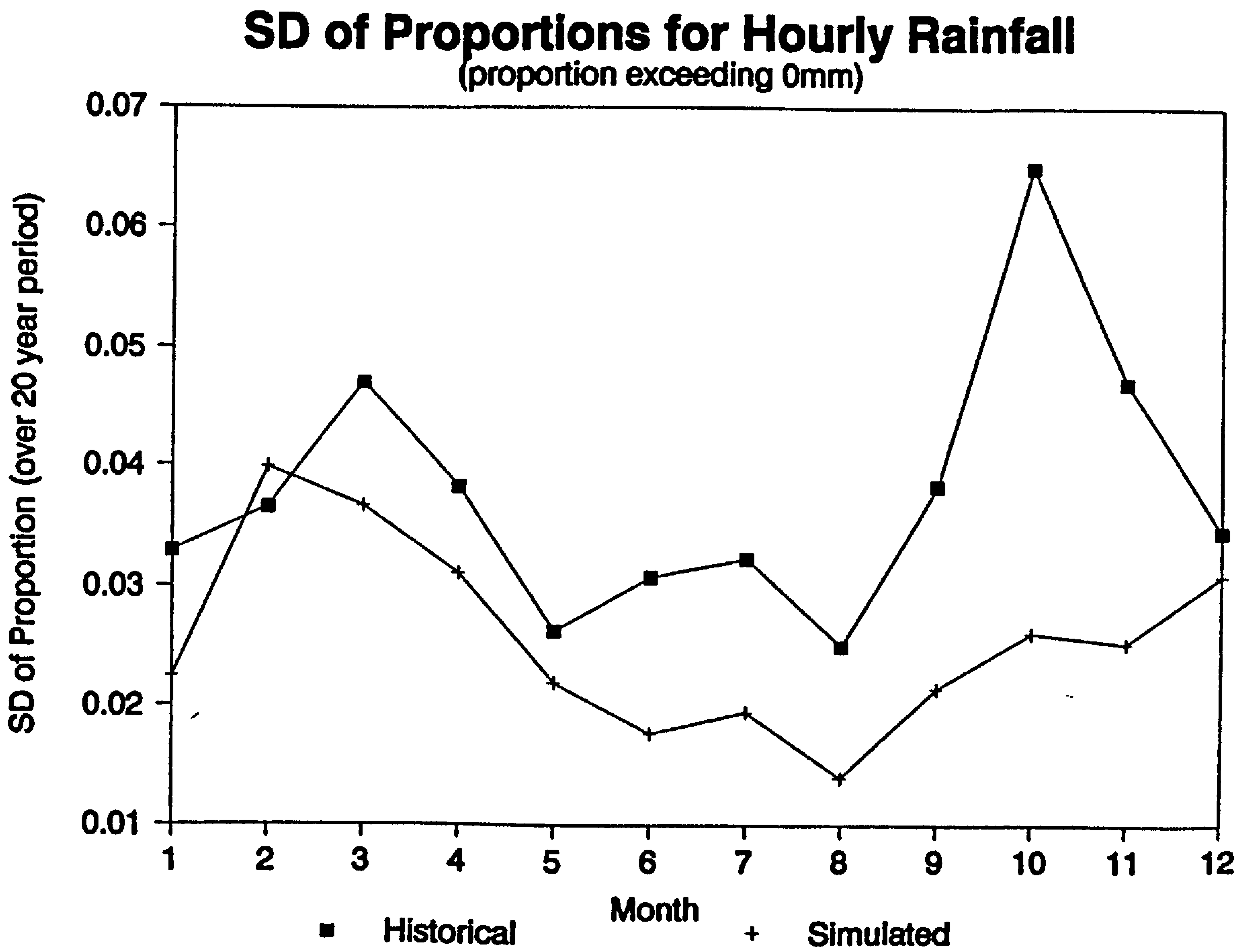


Figure 4.41

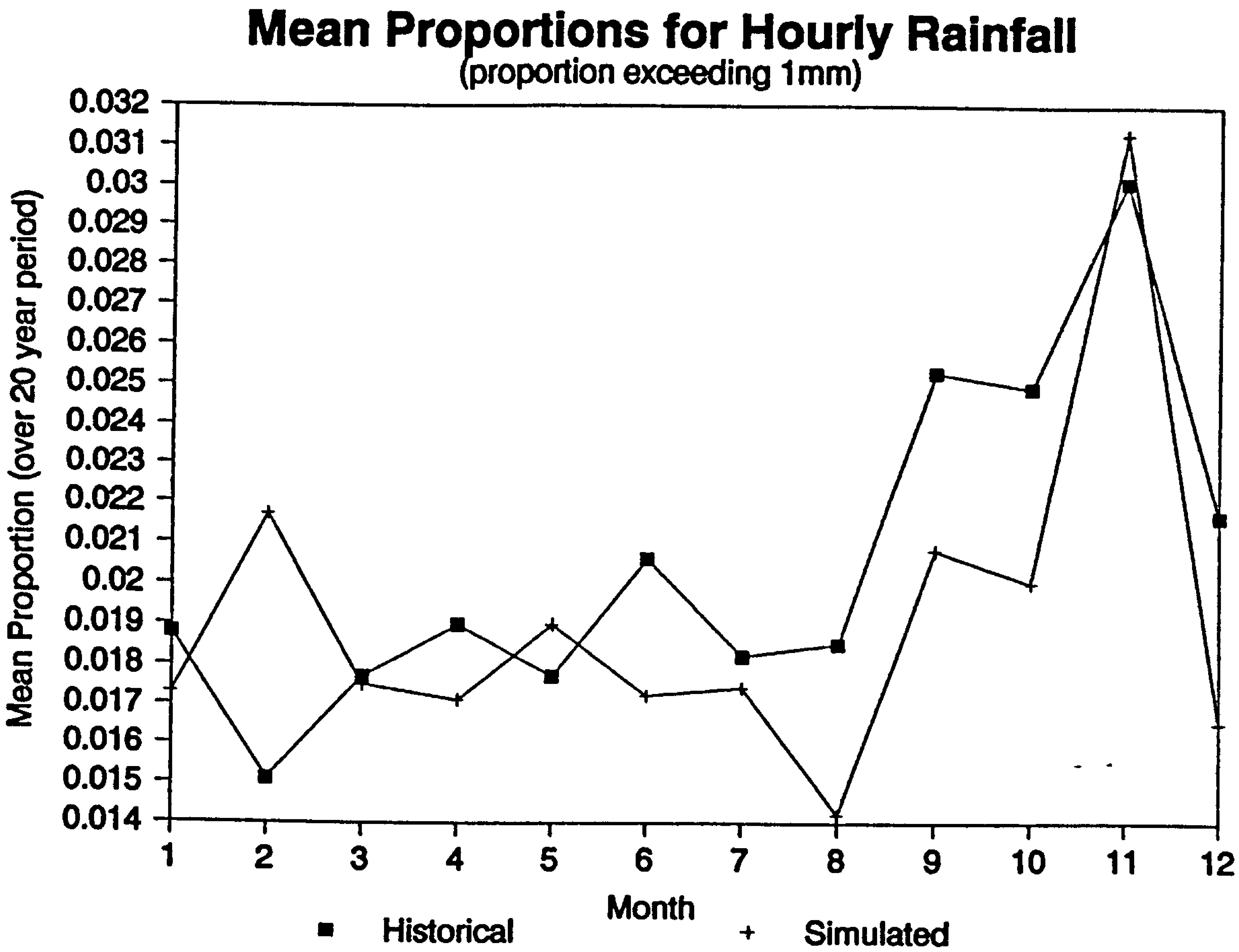


Figure 4.42

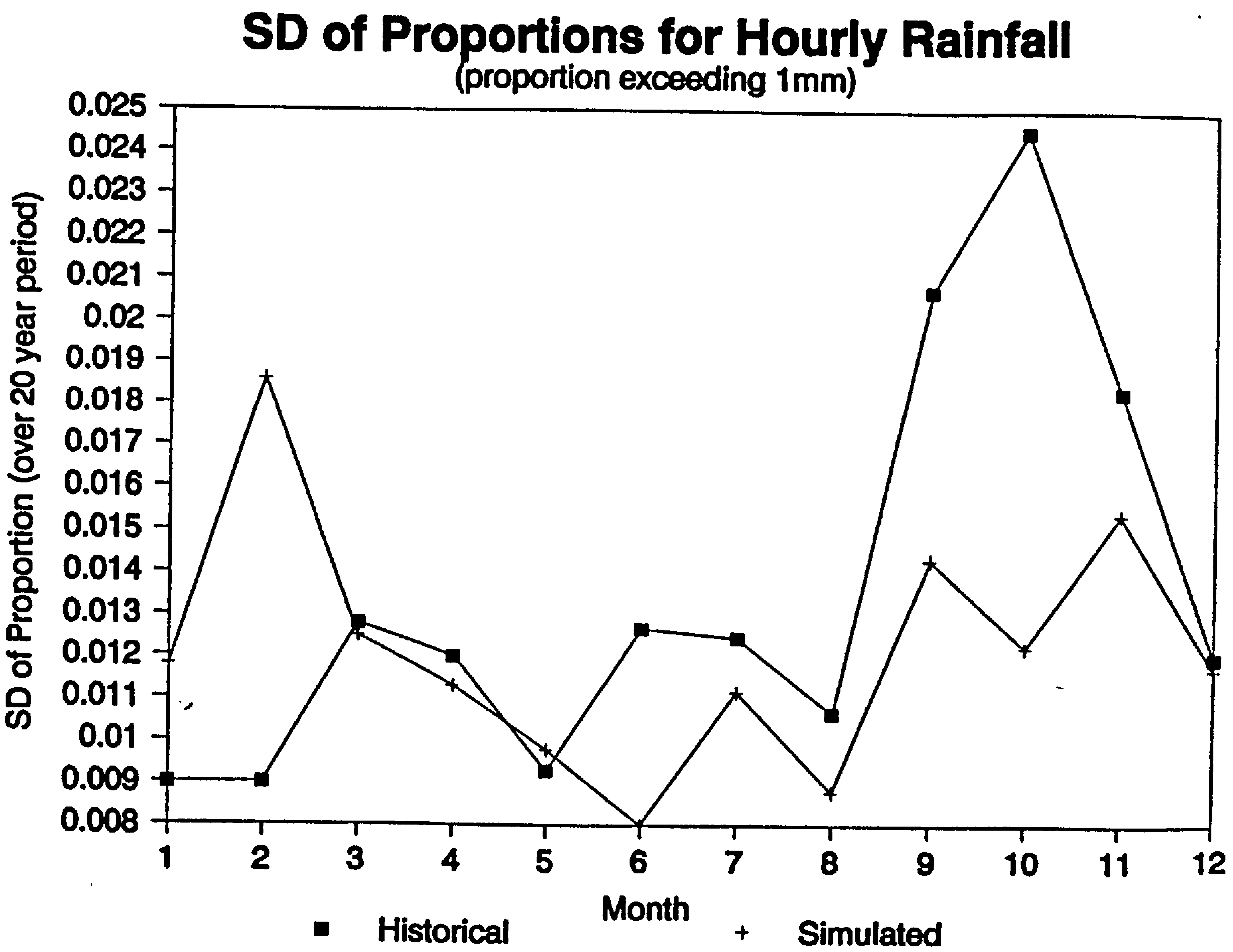


Figure 4.43

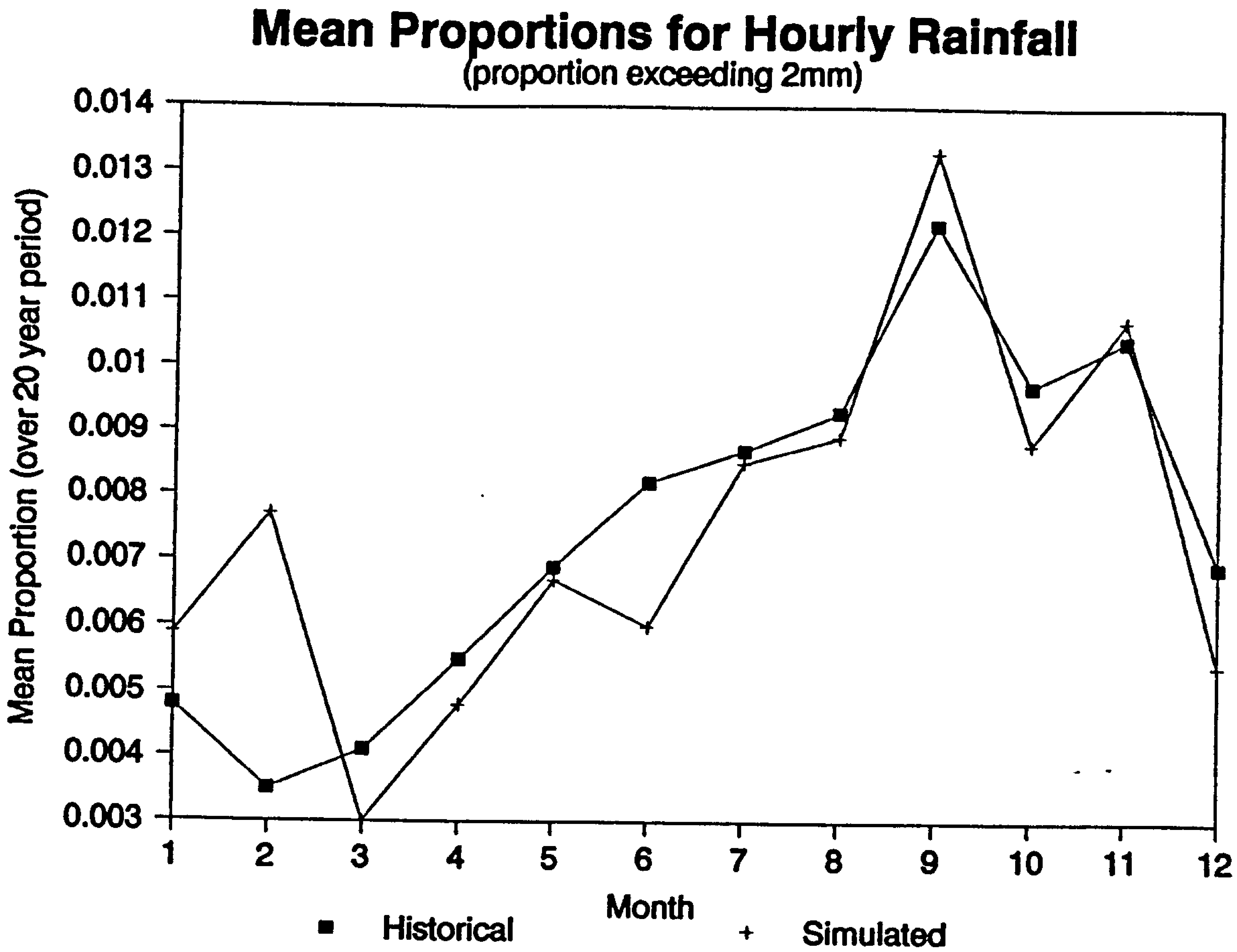


Figure 4.44

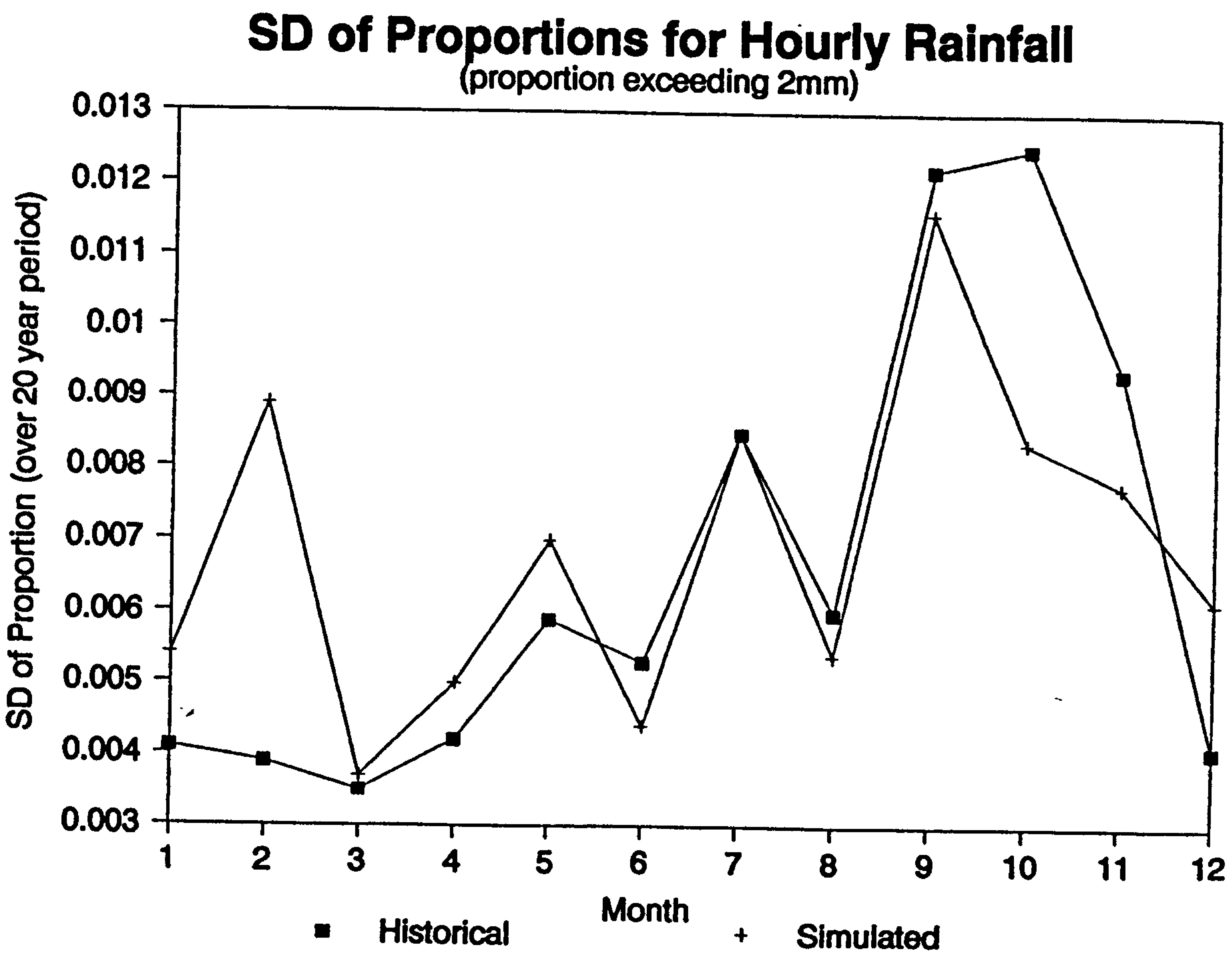


Figure 4.45

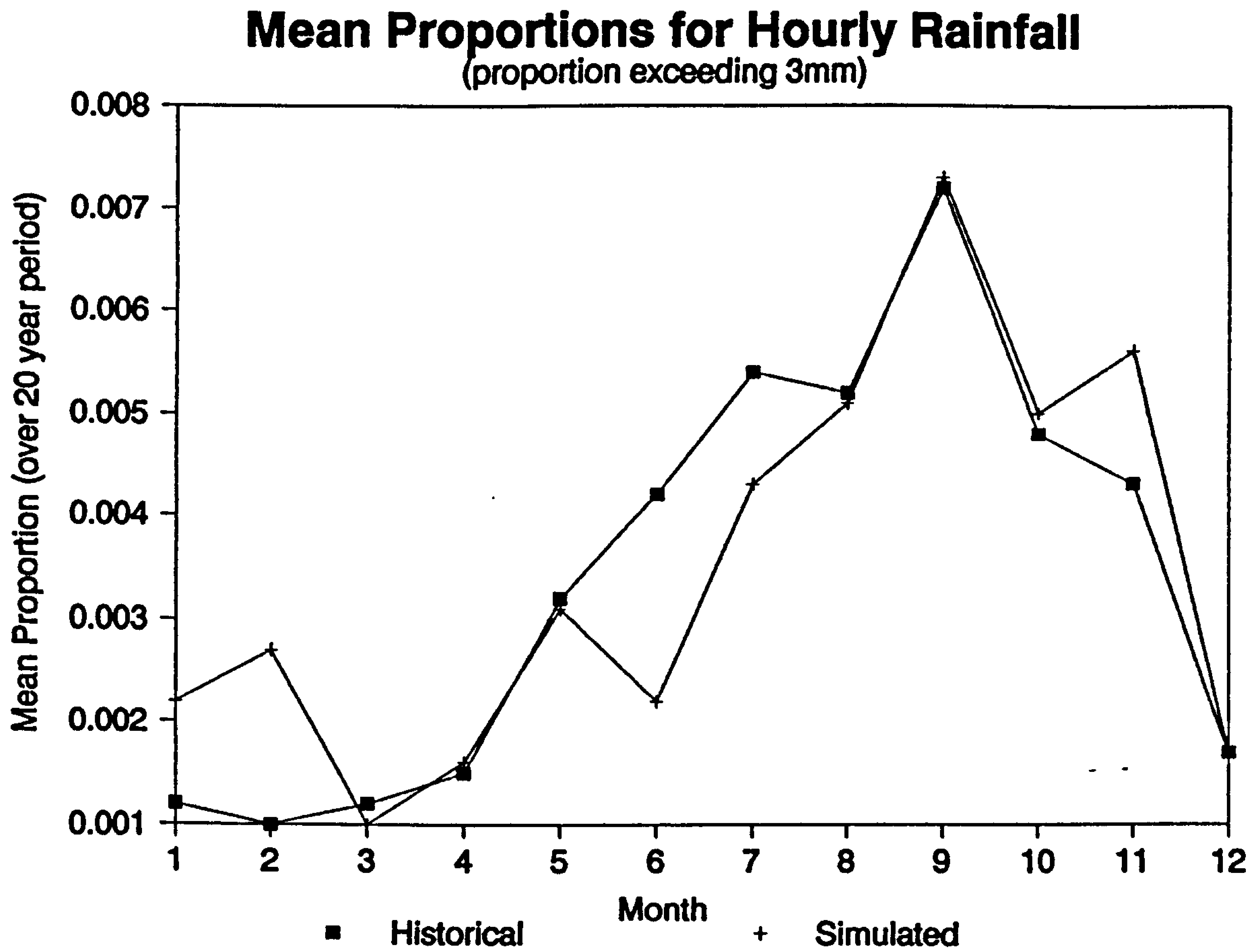


Figure 4.46

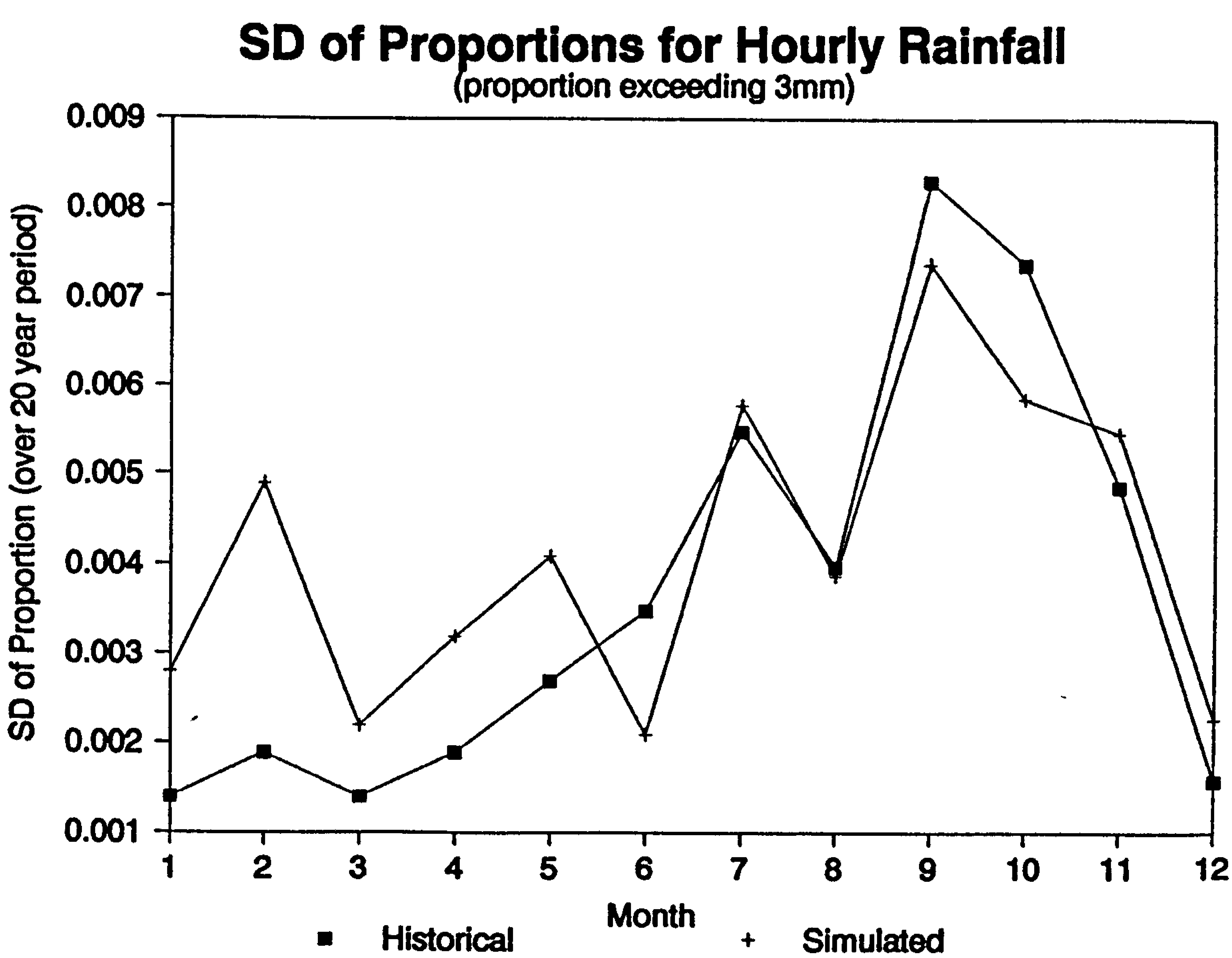


Figure 4.47

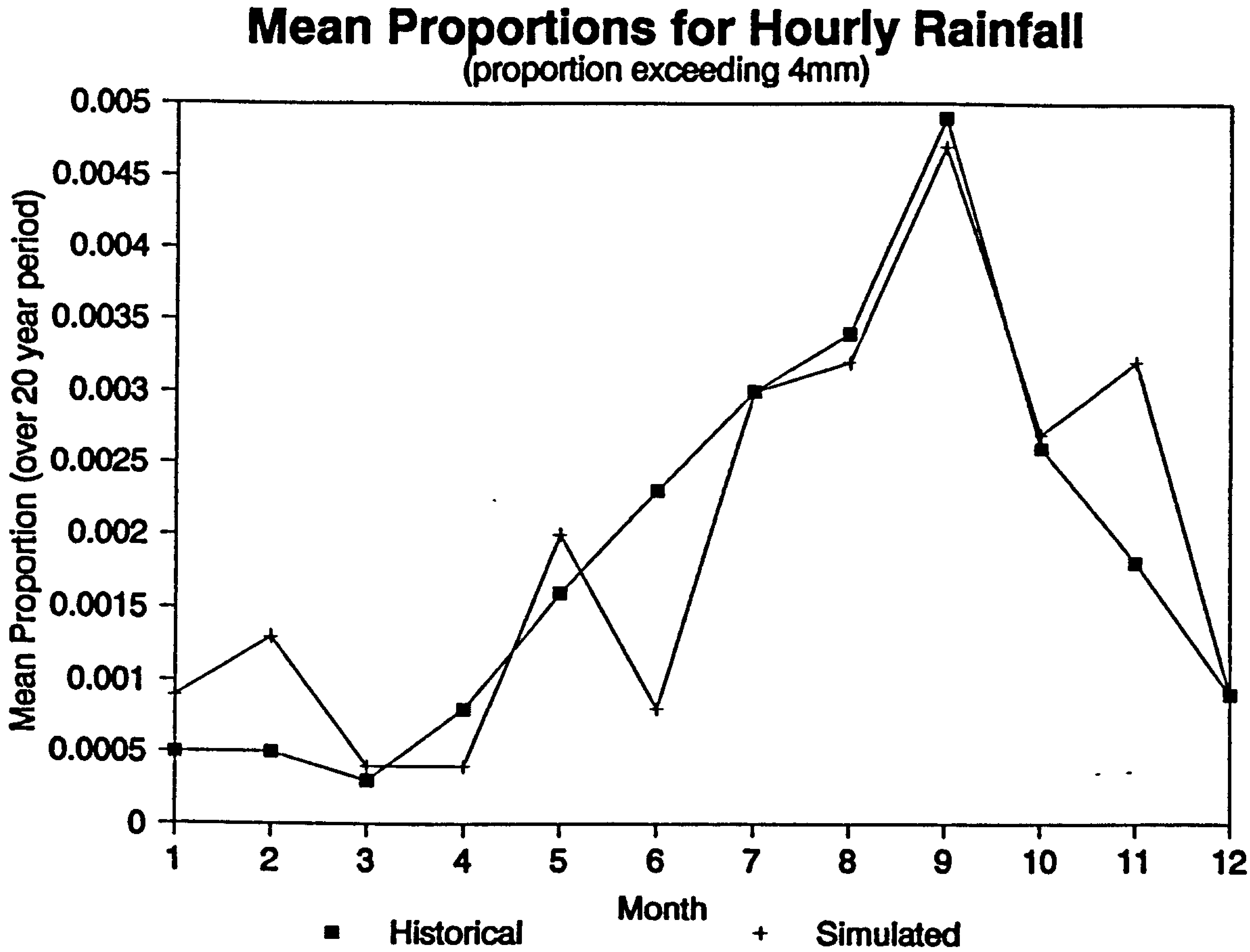


Figure 4.48

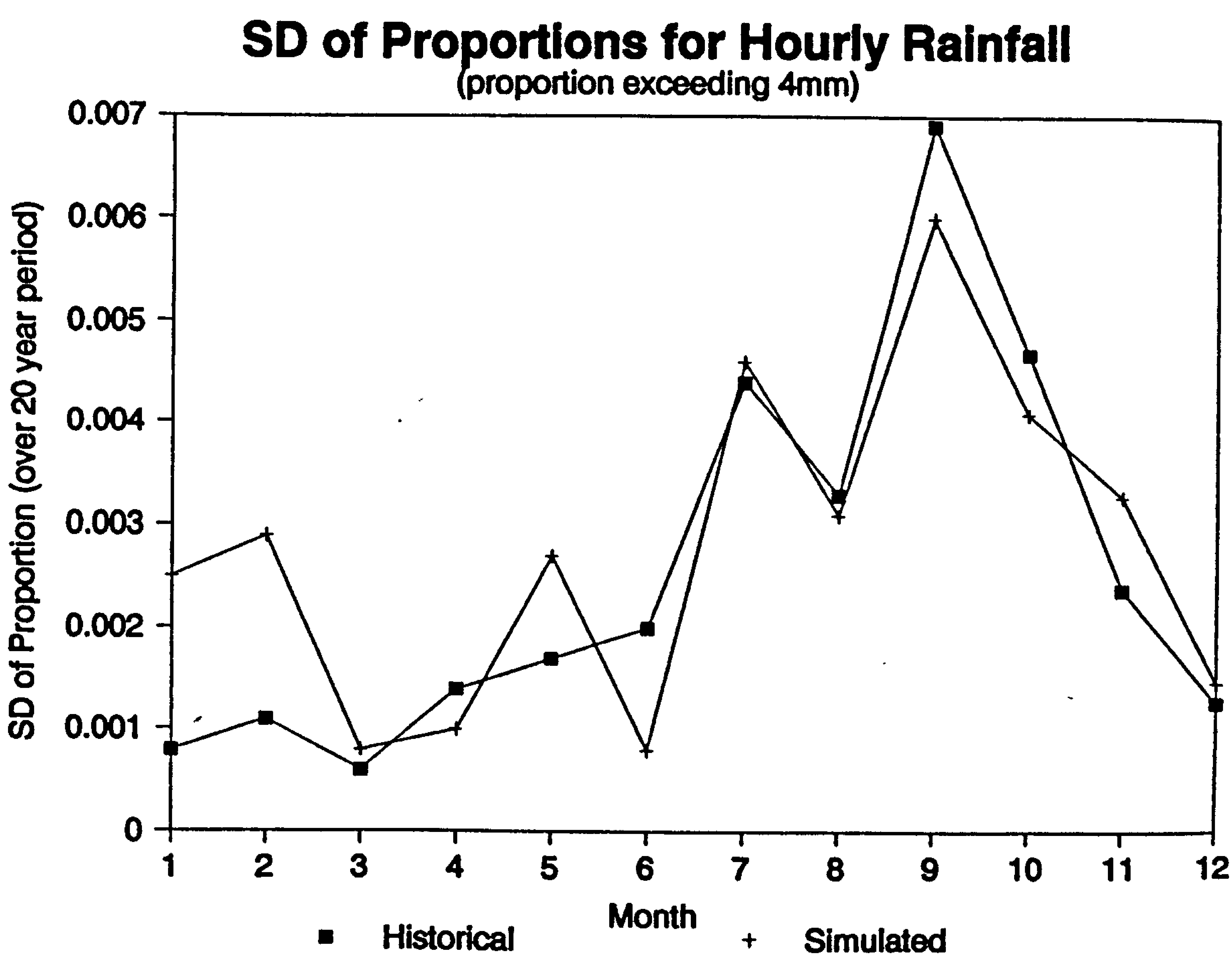


Figure 4.49

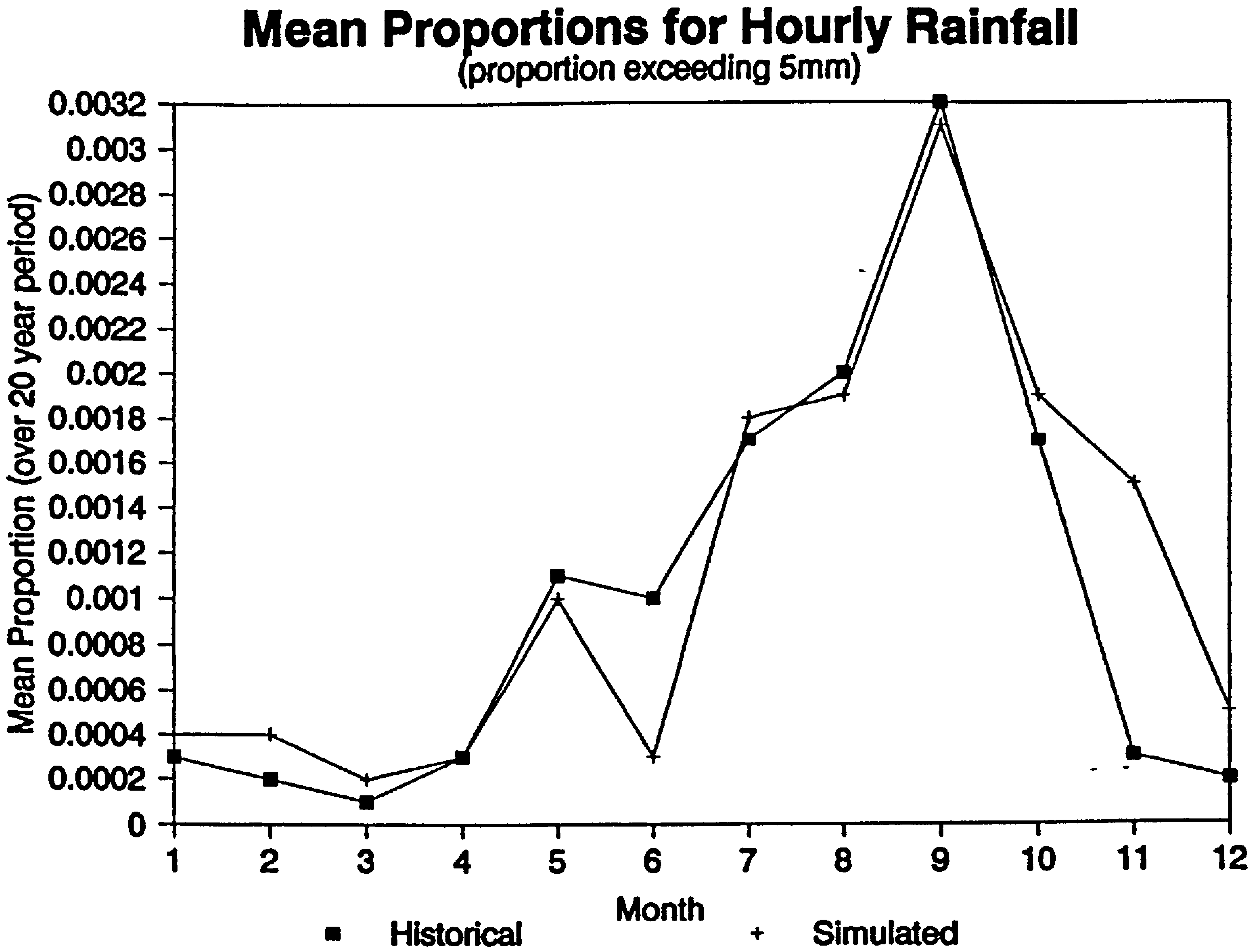


Figure 4.50

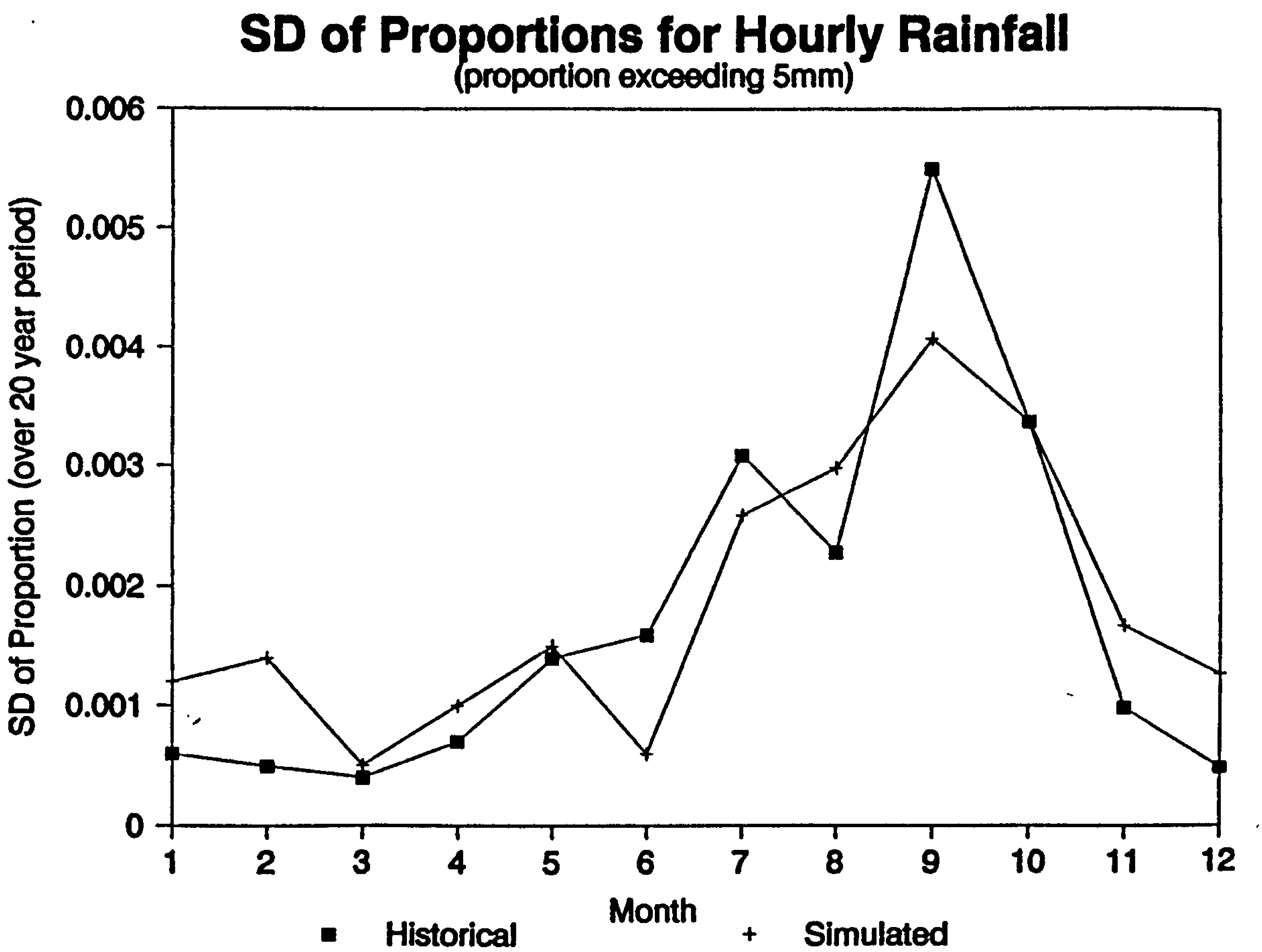


Figure 4.51

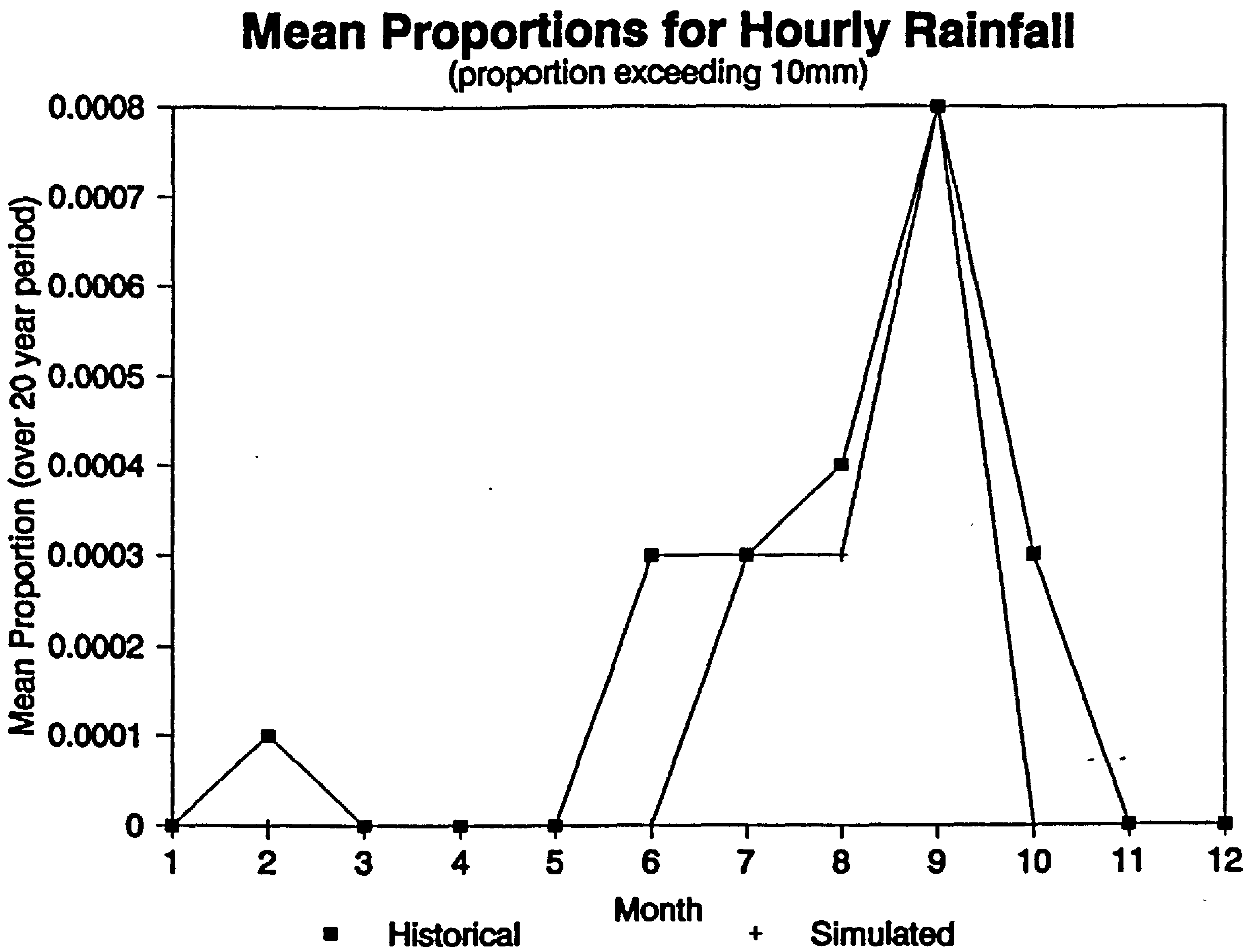


Figure 4.52

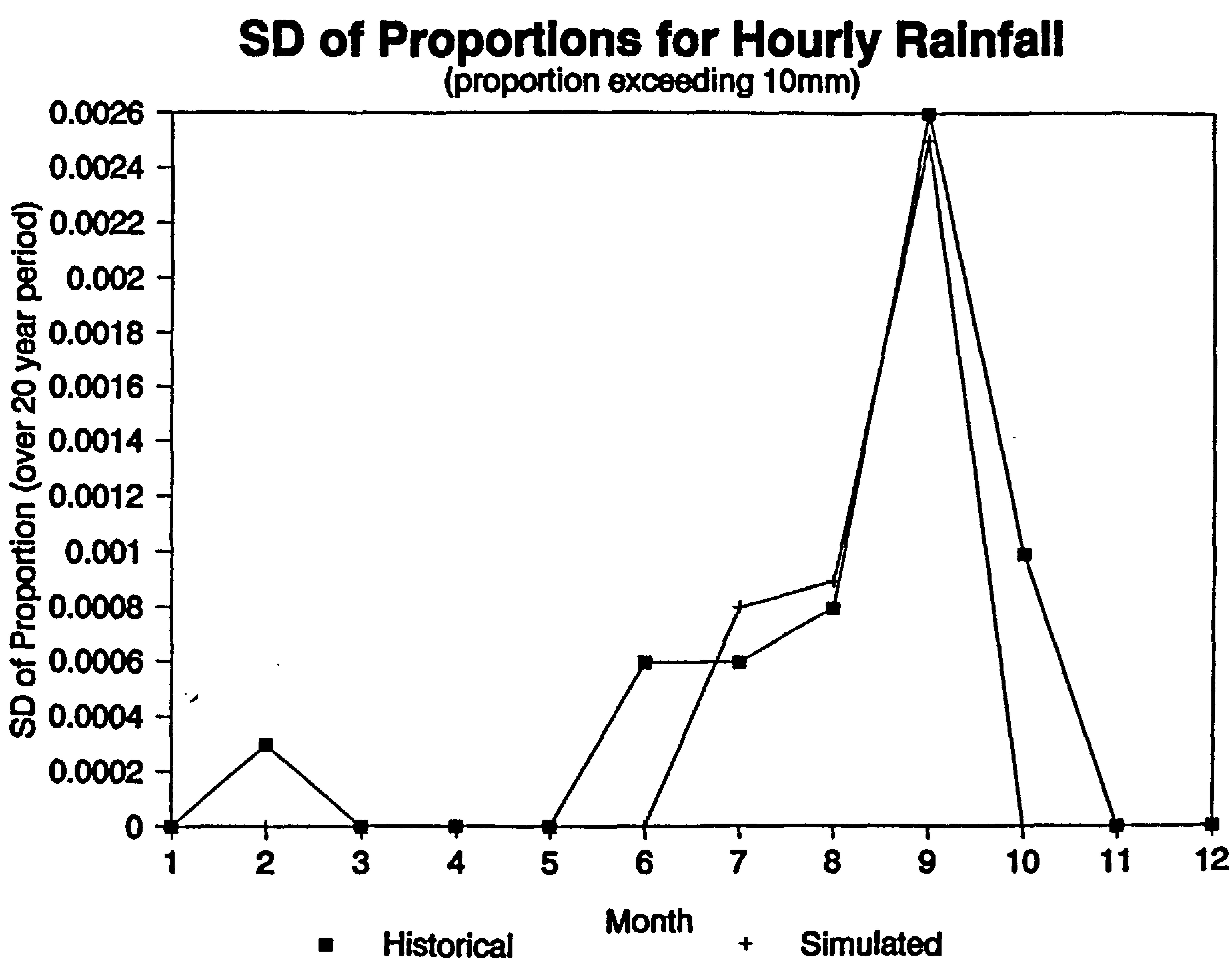


Figure 4.53

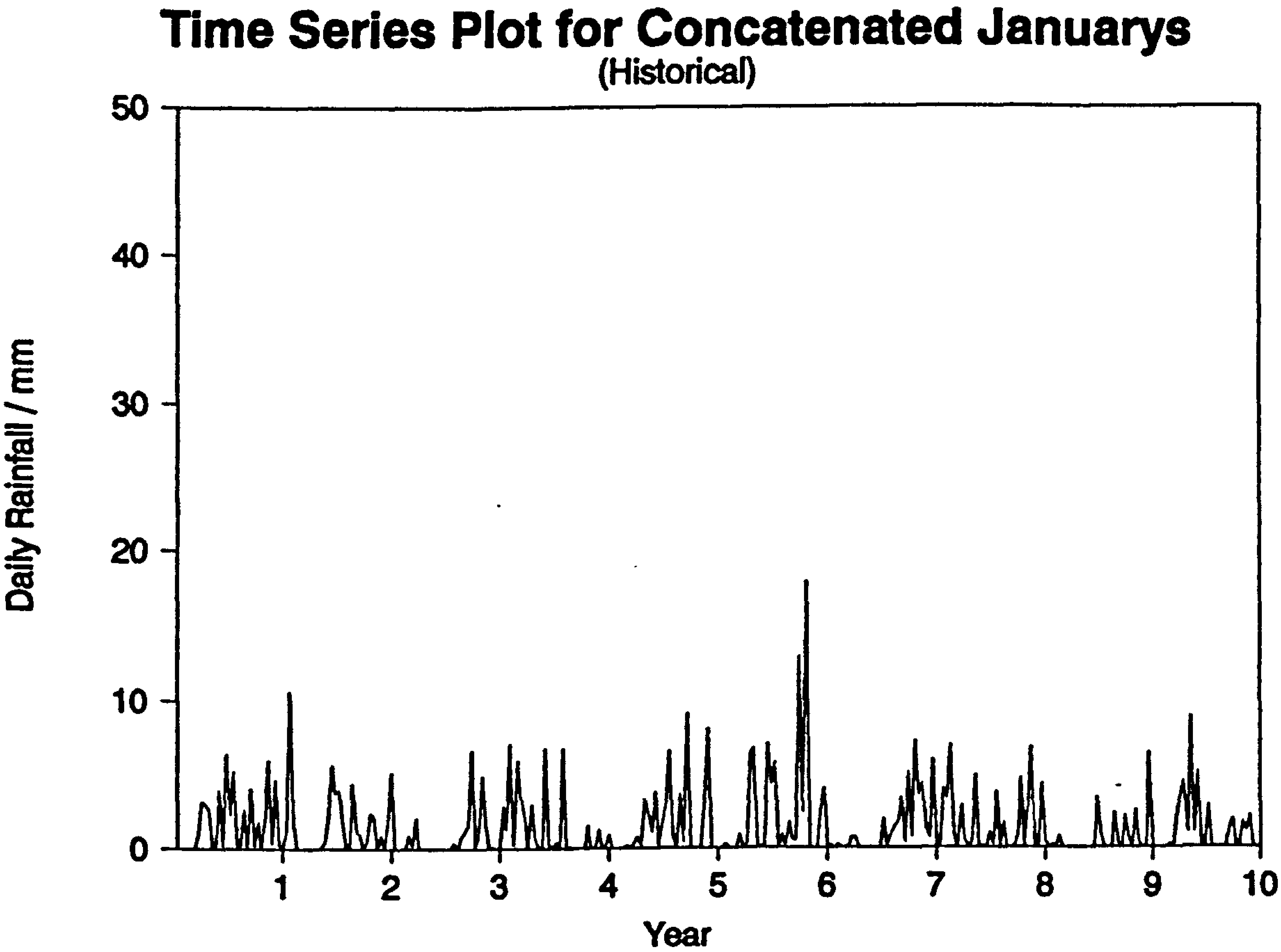


Figure 4.54(a)

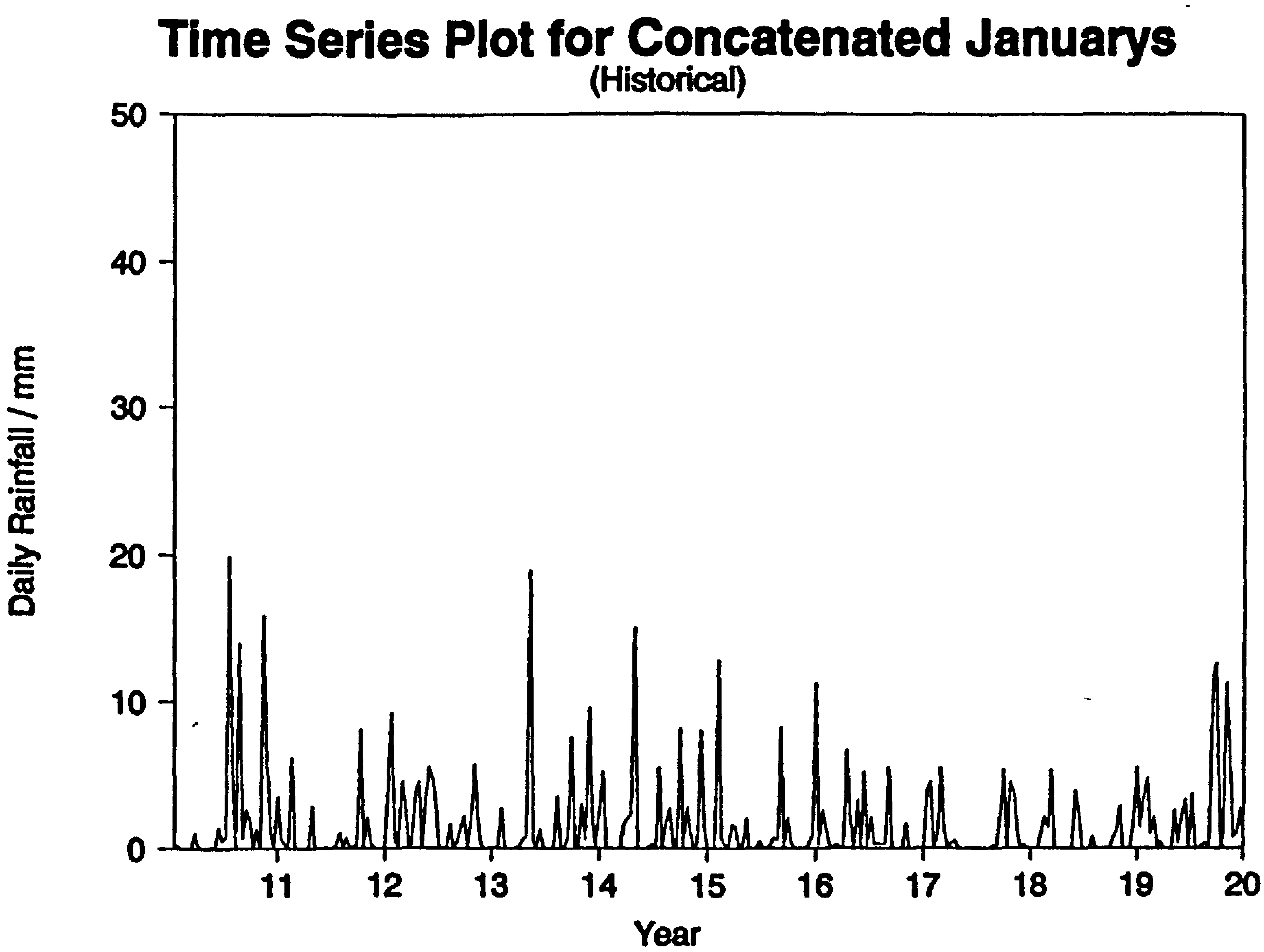


Figure 4.54(b)

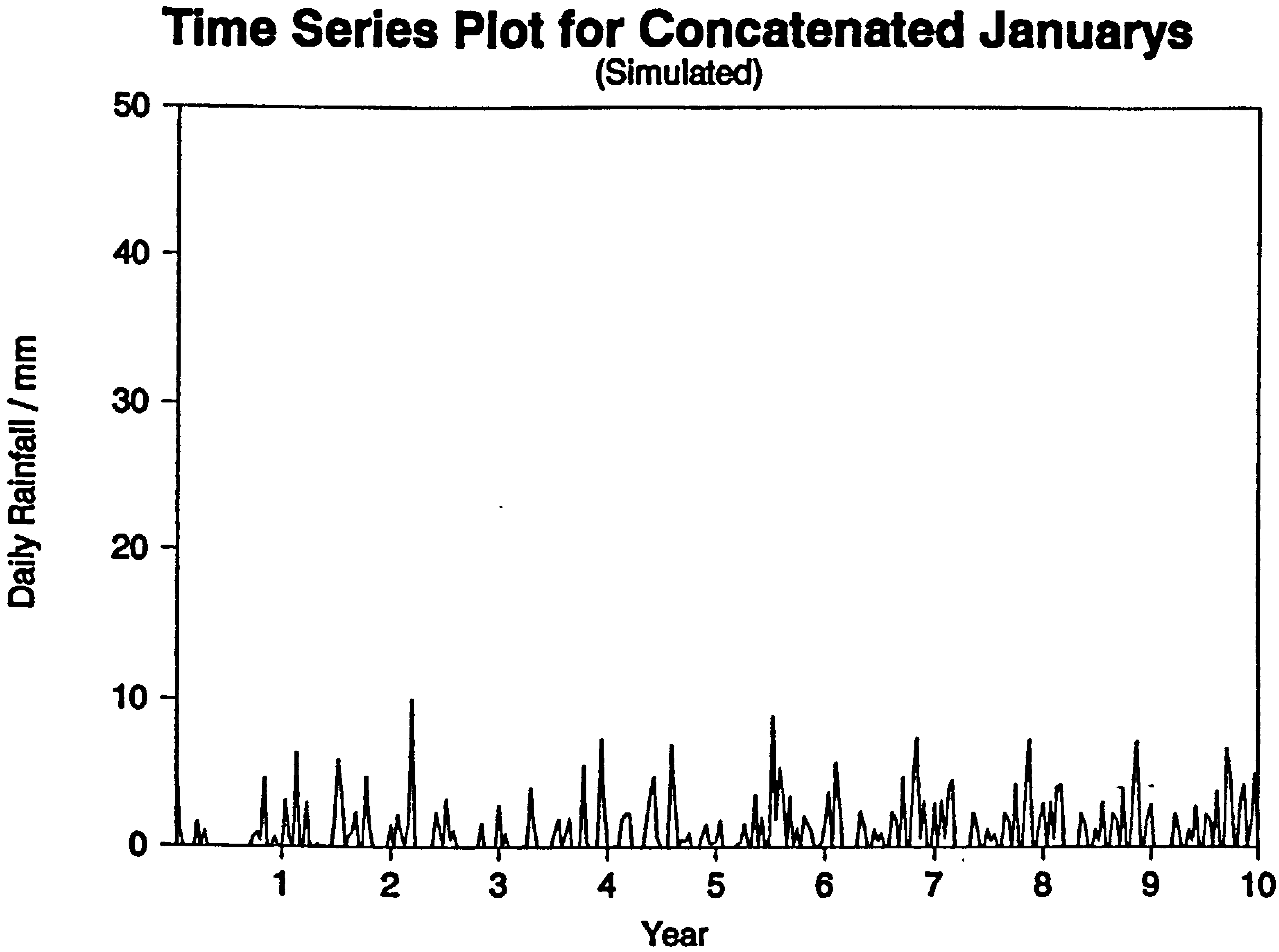


Figure 4.55(a)

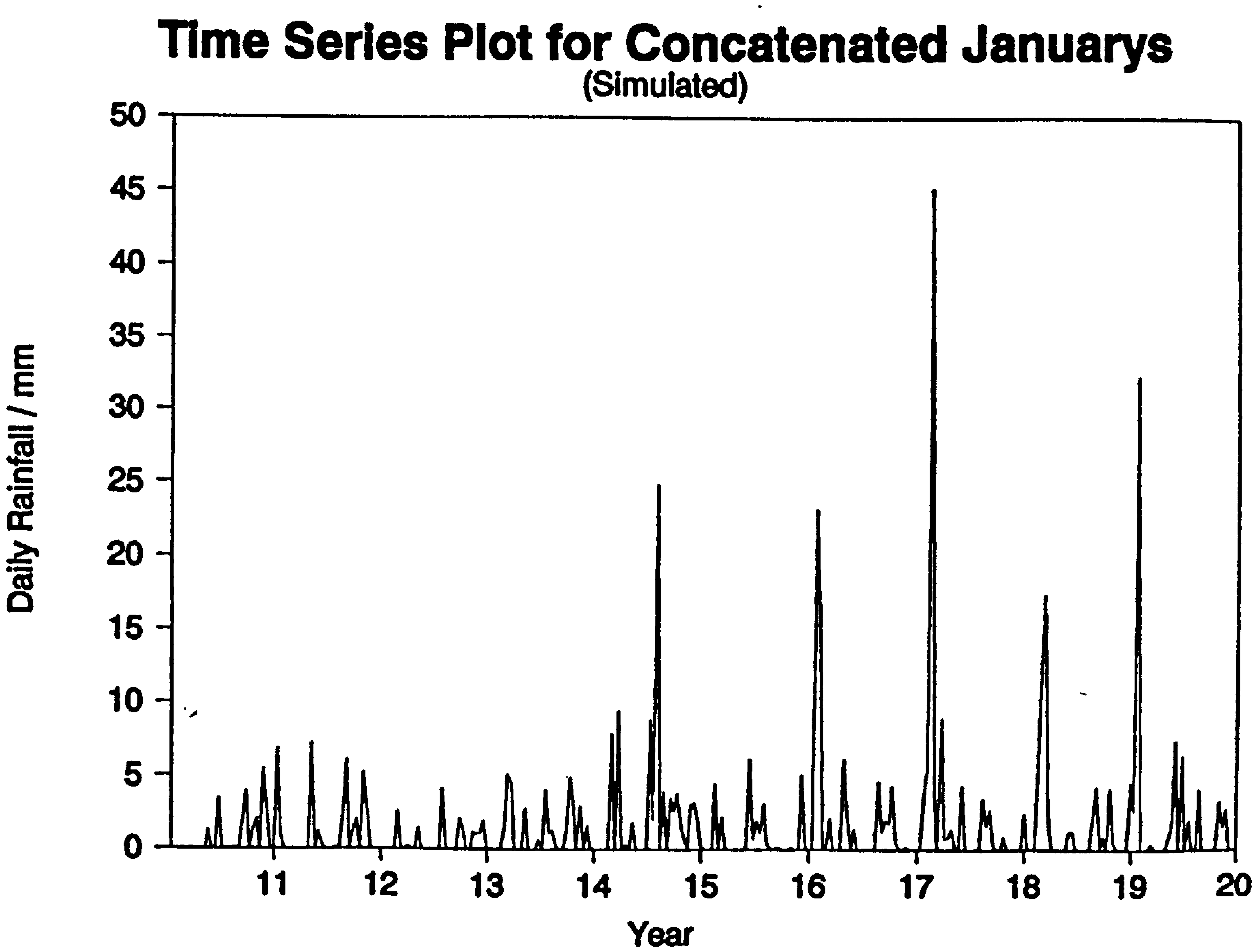


Figure 4.55(b)

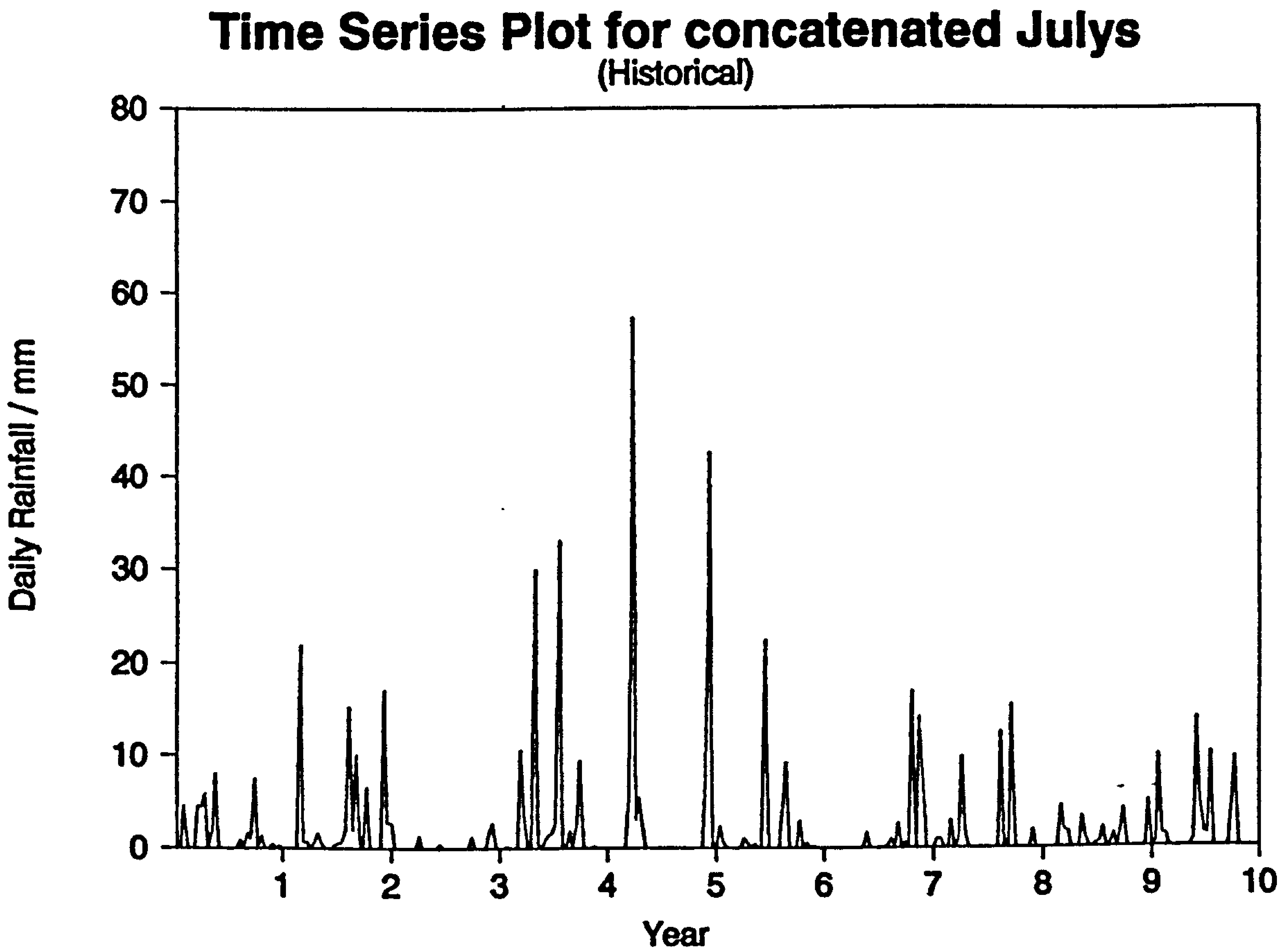


Figure 4.56(a)

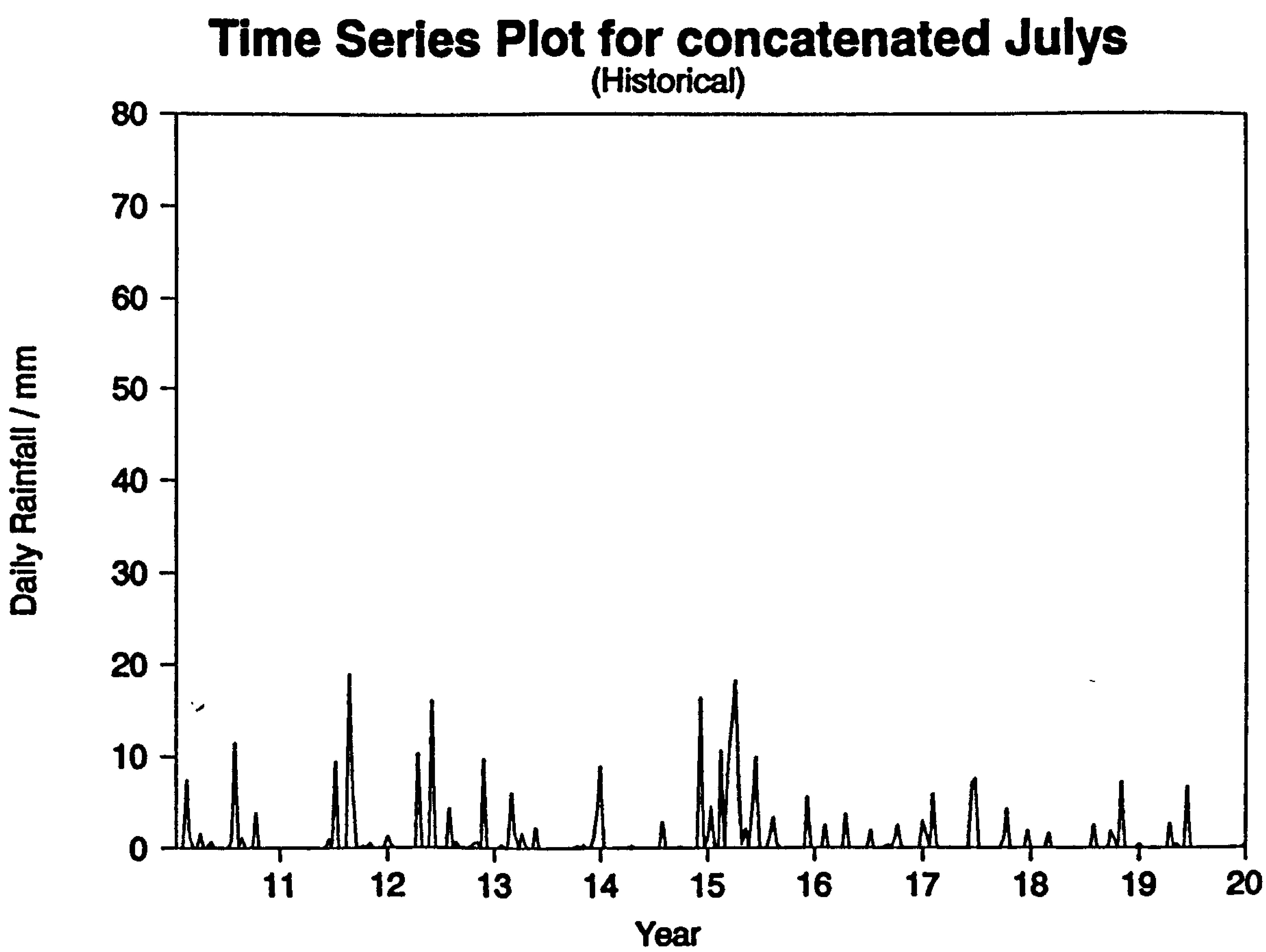


Figure 4.56(b)

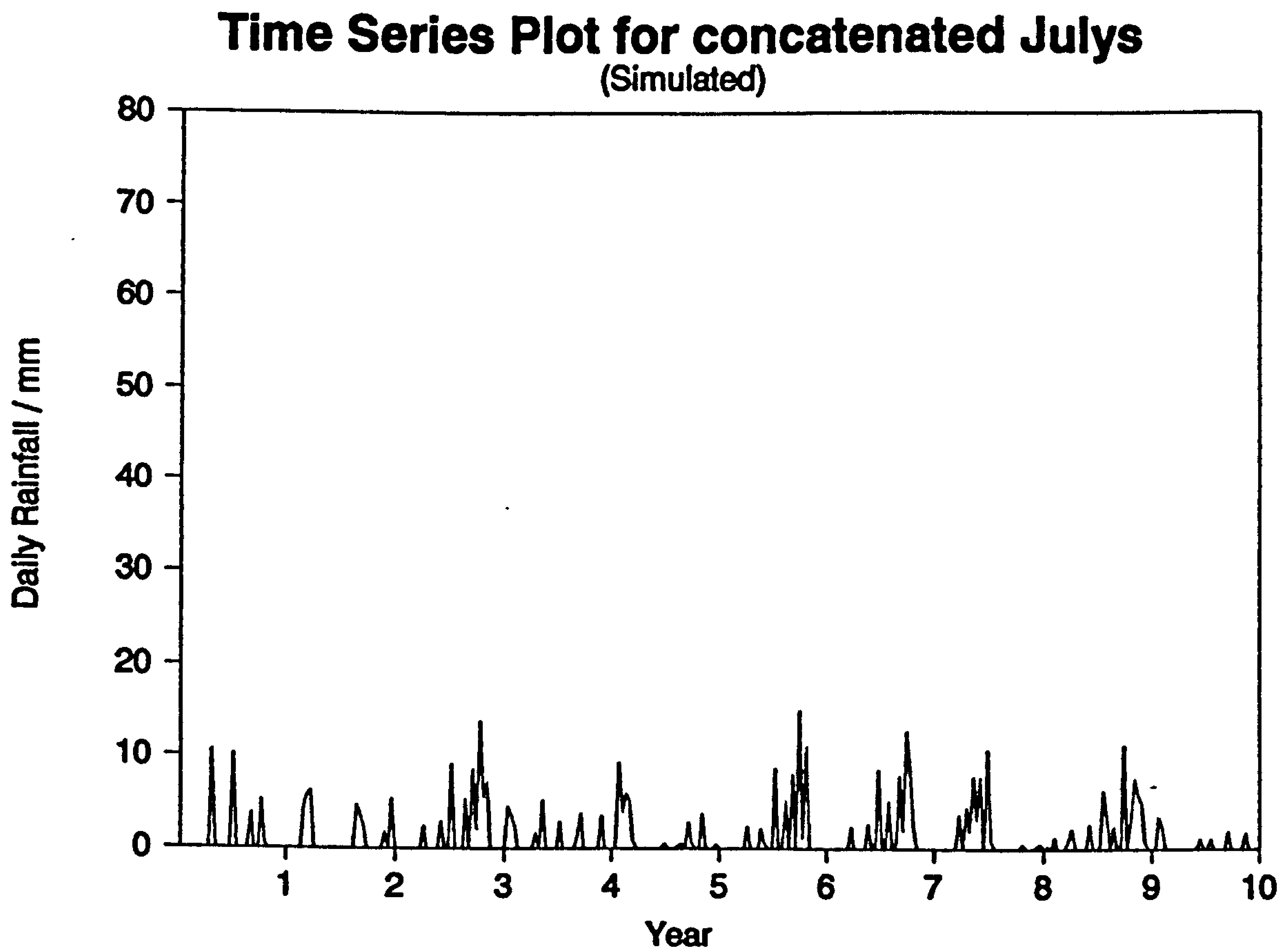


Figure 4.57(a)

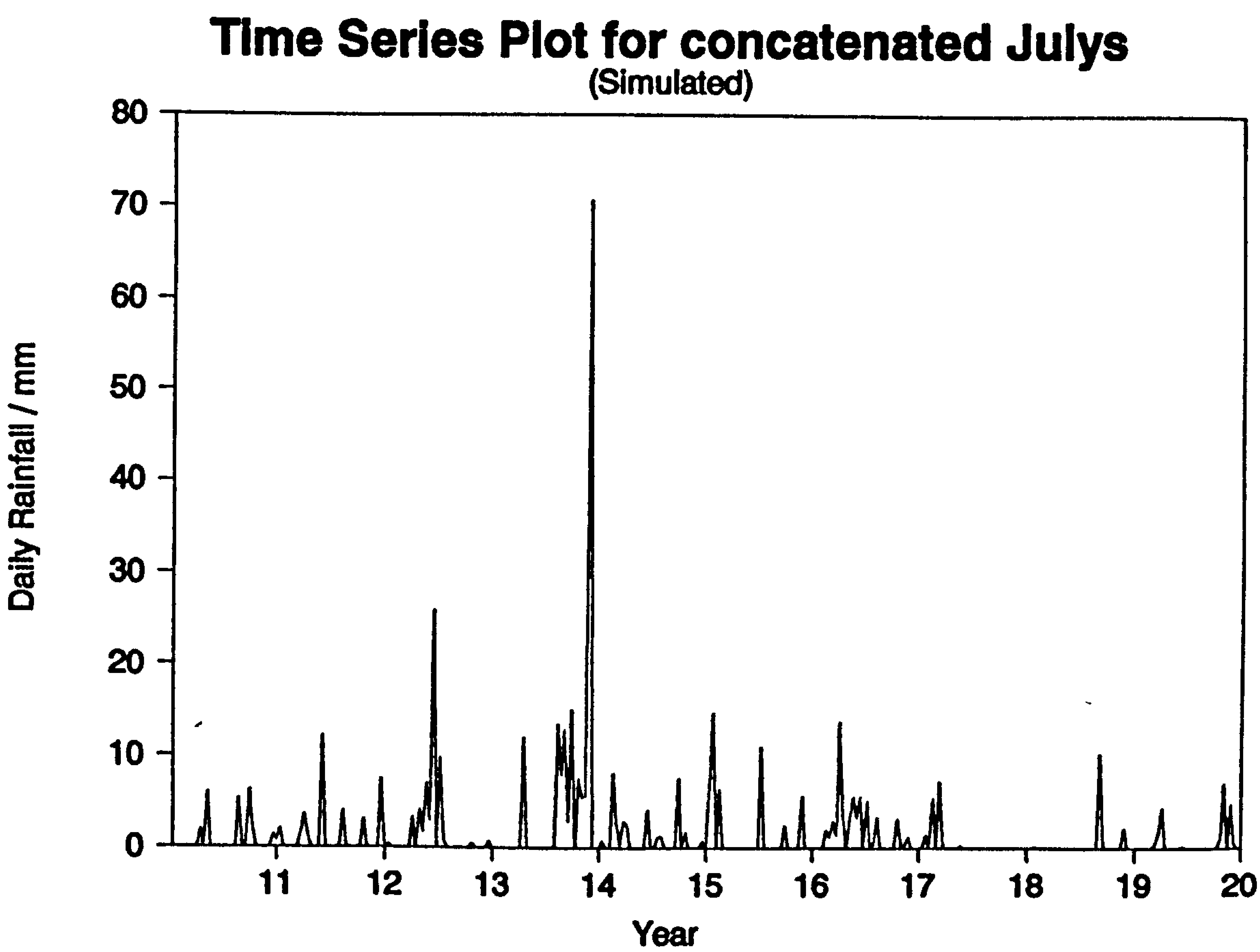


Figure 4.57(b)

4.8 TESTING THE PARAMETER ESTIMATION PROCEDURE ON OTHER STATIONS

From Sections 4.6 and 4.7 it is clear that model is performing well. However, the model has only been fitted to one data set (Manston), and so it now becomes necessary to test the performance of the model on other data sets.

Using the fitting procedure defined in the previous Sections the parameters were estimated for hourly rainfall stations scattered throughout the UK (see Figure 4.58). The percentage errors in the model statistics when using these parameter estimates are given in Tables 4.21-4.29, from which it is clear that the model is matching the historical rainfall statistics, probably within the sampling variability of the historical data (compare the percentage errors in Tables 4.21 - 4.29 with those of Table 4.1).

Figure 4.58
Hourly Rainfall Stations used to Test the Parameter Estimation Procedure

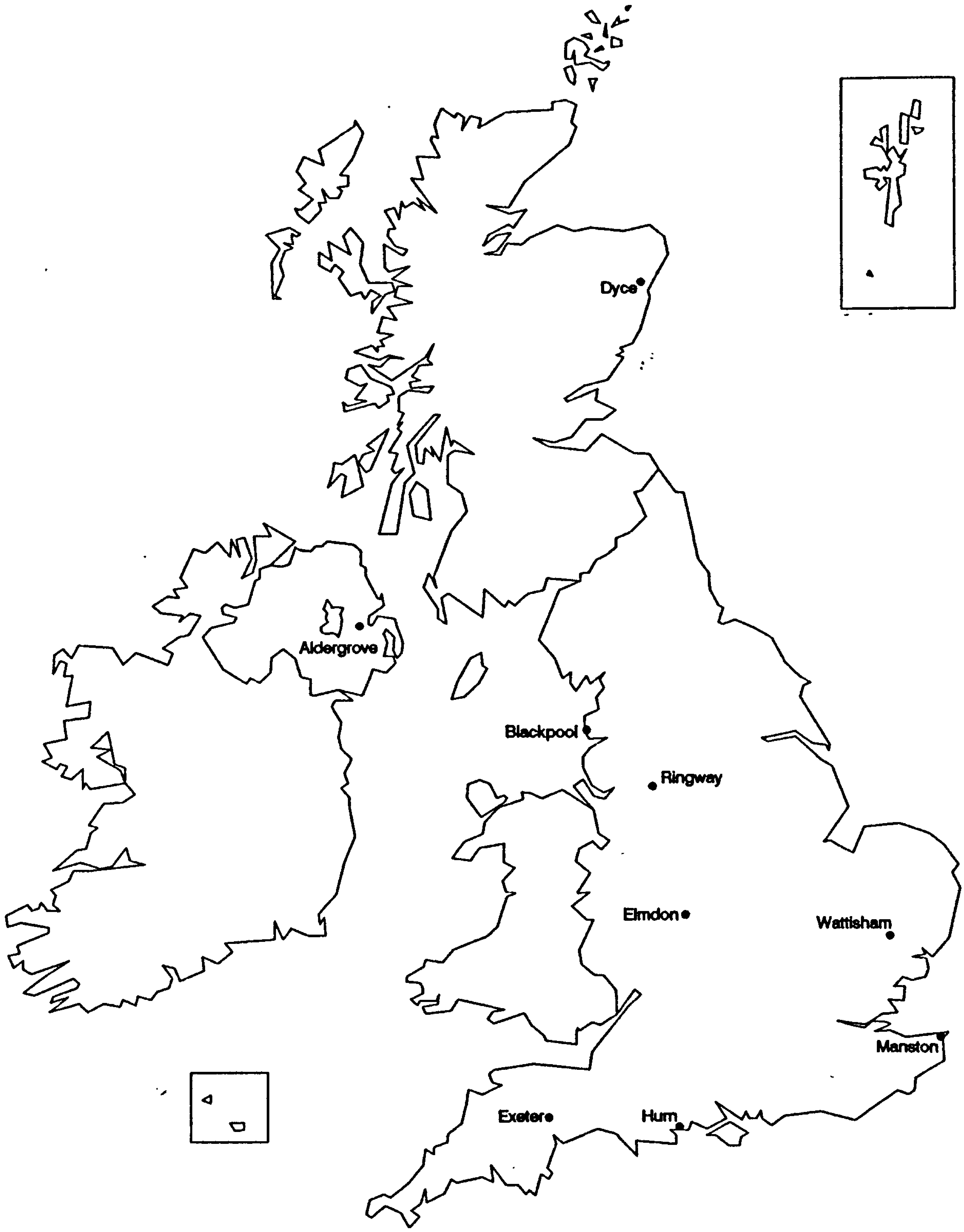


Table 4.21
Percentage Errors for Blackpool

Stats	J	F	M	A	M	J	J	A	S	O	N	D
											Month	
M1	-3	-3	-3	-3	-1	3	-0	-1	-1	0	-3	-2
V1	4	1	1	4	3	-10	-2	-1	-4	-2	1	-1
PD1	-1	-2	-1	-0	-0	4	1	1	1	1	-1	-1
DD1	0	-1	0	1	0	1	0	0	0	2	-0	1
WW1	4	0	6	10	5	-13	0	2	0	12	1	10
V3	4	7	10	10	4	-11	5	6	7	9	5	11
PD3	-1	-4	-0	1	1	7	2	1	2	5	-2	1
DD3	3	-0	2	2	1	3	1	1	1	3	1	3
WW3	11	11	9	11	8	-9	-1	1	-2	0	8	7
V6	11	11	9	11	8	-9	-1	1	-2	0	8	7
PD6	3	4	6	8	3	1	-1	2	5	4	0	6
DD6	3	-5	2	2	2	10	4	2	2	8	-0	5
WW6	8	1	5	2	3	5	3	2	4	5	6	5
V12	13	14	10	7	6	-7	3	4	7	-2	13	3
PD12	1	1	-1	-3	-1	1	-4	-6	-1	0	-0	-7
DD12	-2	-9	-2	-3	-0	9	0	-1	-2	-6	-0	-3
WW12	2	-1	6	-1	1	6	1	3	3	6	5	3
V24	2	3	-4	-5	-3	3	2	3	-1	-16	6	-3
PD24	-7	-13	-4	-9	-4	11	-2	-2	-2	1	-6	-5
DD24	-5	-13	-7	-15	-14	10	-3	-7	-5	3	-6	1
WW24	2	-4	0	-5	-13	-5	-3	-4	-1	3	0	2
Max	13	14	10	11	8	11	5	6	7	12	13	11
Min	-7	-13	-7	-15	-14	-13	-4	-7	-5	-16	-6	-7

- implies the model is under-estimating the statistic
+ implies the model is over-estimating the statistic

Table 4.22

Percentage Errors for Ringway (Manchester)

Stats	J	F	M	A	M	J	J	A	S	O	N	D
												Month
M1	-3	-3	-3	-2	-2	-0	-1	4	-3	-2	-2	-1
V1	3	-0	3	2	1	-1	-8	-10	0	2	0	1
PD1	-1	-0	-3	-1	1	2	1	5	-0	1	-1	-0
DD1	0	0	-1	0	0	1	0	2	0	1	1	1
WW1	6	5	5	7	3	-2	-9	0	5	3	10	8
V3	5	5	6	9	7	-1	4	4	9	4	10	7
PD3	-1	-0	-3	0	2	4	1	9	0	1	0	2
DD3	2	0	2	1	2	1	1	4	0	2	1	2
WW3	8	3	13	7	6	-3	1	-12	2	4	4	5
V6	4	10	7	3	5	4	2	-1	2	-1	2	6
PD6	1	0	-1	1	3	5	3	14	1	3	1	4
DD6	7	1	5	2	3	4	3	7	1	4	3	5
WW6	14	7	12	5	6	3	3	-13	5	6	5	5
V12	2	-0	-2	-2	-3	-4	3	-4	-3	-2	-3	-4
PD12	2	-4	-3	-2	1	2	1	15	-5	2	-2	0
DD12	7	1	6	0	2	3	4	12	2	9	6	3
WW12	7	14	11	9	2	5	9	-2	19	8	12	4
V24	-1	-4	-3	-3	-3	2	0	-14	5	5	-0	-4
PD24	-7	-10	-5	-5	-6	1	-2	22	-6	-4	-6	-4
DD24	-5	-10	-5	-10	-4	-6	-6	15	-13	-5	-4	0
WW24	2	-1	0	-6	3	-8	-9	-9	-4	-1	0	3
Max	14	14	13	9	7	5	9	22	19	9	12	8
Min	-7	-10	-5	-10	-6	-8	-9	-14	-13	-5	-6	-4

- implies the model is under-estimating the statistic

+ implies the model is over-estimating the statistic

Table 4.23

Percentage Errors for Elmdon (Birmingham)

Stats	J	F	M	A	M	J	J	A	S	O	N	D
	Month											
M1	-3	-2	-4	-2	-2	-2	0	2	-3	-3	-4	-3
V1	2	-0	5	1	-3	-1	-5	-12	-1	-1	3	-0
PD1	1	-1	-1	-1	1	0	1	3	0	-0	1	-2
DD1	1	1	1	1	0	1	0	1	0	1	1	0
WW1	13	11	9	10	-4	5	-4	-7	5	11	9	9
V3	8	8	9	9	7	10	4	3	7	14	11	12
PD3	3	-1	0	0	1	1	2	4	0	1	3	-1
DD3	2	-0	3	1	1	1	1	1	-0	1	3	1
WW3	3	2	15	3	6	5	-4	-10	-3	4	10	7
V6	8	7	8	8	10	9	-6	6	2	7	6	7
PD6	5	-1	3	1	2	2	4	6	-0	2	7	-0
DD6	5	1	7	3	3	2	3	3	1	3	7	3
WW6	5	11	13	10	11	6	7	-3	7	6	8	10
V12	-5	4	-2	0	-1	-4	2	-0	-0	-6	6	2
PD12	4	-3	3	-1	1	0	3	7	-3	0	6	-2
DD12	9	-0	11	2	2	2	2	8	1	7	12	4
WW12	11	6	10	10	6	7	-0	10	10	17	14	13
V24	-0	-9	-6	-8	-5	-10	3	-8	3	-3	-11	-8
PD24	-8	-10	-5	-8	-6	-4	-2	11	-11	-8	-9	-7
DD24	-7	-7	-9	-8	-8	-3	-3	3	-9	-9	-15	-7
WW24	2	5	-2	2	0	2	-6	-12	-1	-3	-7	4
Max	13	11	15	10	11	10	7	11	10	17	14	13
Min	-8	-10	-9	-8	-8	-10	-6	-12	-11	-9	-15	-8

- implies the model is under-estimating the statistic
+ implies the model is over-estimating the statistic

Table 4.24

Percentage Errors for Exeter

Stats	J	F	M	A	M	J	J	A	S	O	N	D
										Month		
M1	-0	-1	-2	-1	-1	-1	0	-0	0	-1	0	0
V1	1	1	4	-2	-2	-7	-7	-3	-6	-2	1	0
PD1	2	-0	0	0	1	1	0	1	2	2	2	2
DD1	1	1	1	-0	0	0	0	1	0	1	2	2
WW1	4	8	6	-4	-0	3	-6	8	-3	1	13	12
V3	3	5	5	3	8	13	3	9	5	5	8	5
PD3	5	1	2	-0	1	1	1	1	3	4	5	5
DD3	2	1	3	0	1	1	0	1	1	2	3	2
WW3	-4	1	7	3	4	-2	-5	-1	-5	1	-1	-2
V6	-4	1	-3	3	2	3	0	1	4	5	-0	1
PD6	6	2	5	1	3	2	1	2	4	7	8	7
DD6	5	3	5	1	2	2	1	1	3	6	4	5
WW6	4	4	6	9	4	4	-3	-7	4	5	-5	-1
V12	1	-4	-0	8	-6	-2	1	-9	-8	2	-4	-2
PD12	4	-0	-0	-4	-1	-2	-1	-0	2	4	3	3
DD12	5	2	2	-2	1	1	1	1	3	6	3	4
WW12	3	3	5	7	7	10	16	10	5	9	2	2
V24	1	0	3	-6	4	-6	3	0	2	-6	-10	-8
PD24	-1	-3	-4	-8	-5	-6	-3	-1	2	-1	2	2
DD24	-1	-2	-4	-14	-7	-6	-6	-1	-3	-6	3	2
WW24	-2	-0	2	-12	-2	-5	-20	-5	-5	-2	2	-1
Max	6	8	7	9	8	13	16	10	5	9	13	12
Min	-4	-4	-4	-14	-7	-7	-20	-9	-8	-6	-10	-8

- implies the model is under-estimating the statistic
+ implies the model is over-estimating the statistic

Table 4.25
Percentage Errors for Hurn

									Month			
Stat	J	F	M	A	M	J	J	A	S	O	N	D
M1	-3	-3	-3	-2	-3	2	-0	0	-1	-2	-3	-3
V1	1	4	6	-1	0	-7	-8	-9	-1	3	1	3
PD1	-0	-1	-1	-0	-0	3	2	2	1	-0	-0	0
DD1	1	0	1	0	0	1	1	1	1	1	1	1
WW1	7	8	10	2	5	-10	1	-2	7	7	9	5
V3	7	6	7	8	12	-8	13	8	8	6	10	3
PD3	1	-1	1	0	1	5	3	3	2	1	1	1
DD3	1	-0	4	0	1	2	2	1	1	2	2	2
WW3	3	4	15	0	4	-10	-1	-2	-1	6	8	8
V6	7	-1	2	1	8	-9	12	9	11	2	11	6
PD6	1	-1	4	-0	1	7	5	4	3	3	3	4
DD6	3	2	6	1	3	4	3	2	3	4	5	6
WW6	10	14	9	9	15	-1	4	-1	6	6	10	12
V12	2	0	-1	-1	0	-1	-6	-8	-7	-0	-1	3
PD12	2	-2	3	-1	-0	9	6	4	3	1	2	3
DD12	7	-0	7	1	1	6	5	5	5	6	8	6
WW12	11	3	6	9	6	0	1	8	8	9	14	7
V24	-4	5	-1	3	-10	13	-19	-4	-6	-1	-8	-3
PD24	-8	-15	-5	-9	-9	10	2	1	-5	-6	-12	-7
DD24	-6	-10	-7	-12	-13	4	3	-1	-5	-4	-8	-6
WW24	2	5	0	-1	-3	-12	3	-6	-0	1	2	2
Max	11	14	15	9	15	13	13	9	11	9	14	12
Min	-8	-15	-7	-12	-13	-12	-19	-9	-7	-6	-12	-7

- implies the model is under-estimating the statistic
+ implies the model is over-estimating the statistic

Table 4.26

Percentage Errors for Manston (in Kent)

Stats	J	F	M	A	M	J	J	A	S	O	N	D
											Month	
M1	-2	-2	-3	-2	-1	-1	-0	1	2	-1	-3	-2
V1	4	1	1	1	1	-5	-6	-18	-7	-6	0	3
PD1	1	1	-0	-0	1	0	1	2	4	1	1	0
DD1	1	1	0	0	0	-0	1	1	2	1	1	1
WW1	10	6	4	4	1	-5	10	-0	5	-1	7	6
V3	6	11	7	6	4	7	13	15	9	9	10	7
PD3	3	2	0	0	1	0	2	3	7	2	2	1
DD3	3	2	1	1	1	0	1	1	4	1	3	2
WW3	11	8	6	5	4	3	-1	-7	-2	1	9	9
V6	8	6	6	3	-0	7	8	10	2	7	12	4
PD6	6	4	1	1	3	1	3	4	11	4	4	4
DD6	5	4	2	2	2	1	2	3	6	3	5	3
WW6	3	7	5	6	1	7	-6	-4	-5	3	11	3
V12	2	3	0	-1	-3	3	-8	3	-9	-2	3	-1
PD12	2	-2	-4	-2	-2	-2	1	3	13	-0	0	0
DD12	3	-0	-1	0	-1	-1	1	4	9	2	2	1
WW12	4	7	10	8	2	2	9	14	-0	8	3	3
V24	-12	-11	-2	-1	3	-7	-9	-25	-7	-7	-15	-3
PD24	-4	-6	-8	-6	-4	-6	-1	4	17	-2	-7	-6
DD24	-9	-12	-8	-12	-9	-8	-2	2	14	-4	-6	-3
WW24	-4	-5	-0	-6	-8	-4	-4	-10	-3	-3	3	5
Max	11	11	10	8	4	7	13	15	17	9	12	9
Min	-12	-12	-8	-12	-9	-8	-9	-25	-9	-7	-15	-6

- implies the model is under-estimating the statistic
+ implies the model is over-estimating the statistic

Table 4.27
Percentage Errors for Wattisham (East Anglia)

	Month											
Stats	J	F	M	A	M	J	J	A	S	O	N	D
M1	-4	-4	-5	-5	-3	-1	-1	-0	-2	-1	-4	-5
V1	5	3	6	2	4	-6	-2	-10	-1	-6	2	8
PD1	1	-1	-1	-2	-1	1	1	2	1	2	2	1
DD1	1	0	0	-1	0	-0	1	1	1	1	1	2
WW1	11	5	9	0	7	-11	5	-0	5	3	4	18
V3	11	11	6	8	9	-2	11	13	12	12	10	11
PD3	3	-1	-0	-4	-0	1	2	2	2	3	3	5
DD3	5	1	2	-0	1	2	1	1	2	2	3	5
WW3	15	11	13	13	10	12	1	-5	6	-2	8	16
V6	6	0	8	6	3	-2	12	6	6	3	2	10
PD6	7	1	2	-4	0	4	3	3	5	4	6	11
DD6	9	2	7	2	1	3	3	2	2	4	6	9
WW6	12	9	16	18	6	3	4	2	-7	5	6	2
V12	0	3	2	8	-2	4	1	-2	-0	-1	-2	1
PD12	1	-6	-4	-9	-3	-0	-1	1	-1	0	-4	-2
DD12	5	-2	4	-3	-1	-2	1	2	2	1	-1	1
WW12	5	11	15	8	7	-1	7	7	17	2	12	9
V24	-6	0	-2	-1	-2	7	-23	-12	-12	-5	4	-10
PD24	-7	-14	-9	-16	-8	-4	-4	-0	-7	-5	-11	-11
DD24	-10	-13	-14	-10	-10	-19	-2	-3	-9	-3	-11	-6
WW24	4	2	-1	7	0	-20	2	-8	-10	1	1	15
Max	15	11	16	18	10	12	12	13	17	12	12	18
Min	-10	-14	-14	-16	-10	-20	-23	-12	-12	-6	-11	-11

- implies the model is under-estimating the statistic
+ implies the model is over-estimating the statistic

Table 4.28

Percentage errors for Dyce

Stats	J	F	M	A	M	J	J	A	S	O	N	D
											Month	
M1	-0	-2	-2	0	-2	-1	-1	-0	-2	-1	1	-2
V1	-2	2	2	2	-2	2	-1	-2	1	0	-5	7
PD1	0	0	-0	2	-0	2	2	2	-0	1	2	1
DD1	3	1	2	2	1	1	1	1	1	1	2	2
WW1	20	10	12	8	12	6	0	-2	10	5	11	16
V3	10	6	7	1	9	4	4	2	7	7	7	5
PD3	6	3	3	5	2	4	3	5	1	3	6	4
DD3	4	2	3	2	1	2	2	3	1	2	2	3
WW3	6	4	6	-4	1	4	-1	3	3	3	-4	5
V6	6	9	5	-3	6	7	-2	4	9	-1	4	-1
PD6	11	5	6	8	3	6	6	8	2	4	9	7
DD6	5	5	6	5	2	4	6	6	4	5	4	6
WW6	-4	4	5	-5	0	0	6	2	6	3	-4	2
V12	-8	-1	-2	-2	-5	-2	3	-4	-7	-3	-2	-6
PD12	2	-1	2	4	-4	1	1	5	-0	2	1	2
DD12	0	6	4	4	0	4	2	4	4	2	2	5
WW12	-2	15	4	-2	12	8	2	1	8	-0	1	6
V24	-6	-9	-3	-1	-0	-6	-0	1	-4	-2	-10	3
PD24	-2	-3	-5	1	-4	-2	-3	-0	-4	-2	4	-5
DD24	3	-5	-1	4	-5	-8	-3	-3	-2	-1	3	-3
WW24	5	-1	5	5	0	-5	1	-2	2	-1	-0	2
Max	20	15	12	8	12	8	6	8	10	7	11	16
Min	-8	-9	-5	-5	-5	-8	-3	-4	-7	-3	-10	-6

- implies the model is under-estimating the statistic
+ implies the model is over-estimating the statistic

Table 4.29

Percentage Errors for Aldergrove (Northern Ireland)

	Month											
Stats	J	F	M	A	M	J	J	A	S	O	N	D
M1	-3	-4	-2	-0	-0	-2	1	-1	-2	-0	-3	-3
V1	7	5	1	-3	-9	6	-10	3	1	3	6	9
PD1	-1	-1	-1	0	2	0	4	1	0	1	-1	0
DD1	1	1	1	2	0	1	1	1	1	2	1	2
WW1	8	15	8	21	-13	11	-11	10	12	13	15	16
V3	3	11	9	13	-2	5	-3	9	10	4	11	6
PD3	1	-0	-0	3	2	2	6	3	3	4	1	4
DD3	3	2	1	2	2	2	3	3	2	3	2	6
WW3	9	11	4	3	7	5	-7	5	1	1	5	20
V6	-4	9	4	10	3	-2	-6	-5	6	5	-1	1
PD6	5	1	3	6	4	3	10	7	6	9	3	9
DD6	9	4	7	4	3	2	8	5	5	7	4	13
WW6	7	11	8	1	3	2	2	-0	4	1	7	14
V12	3	-4	-4	-3	1	-5	-2	-3	2	-1	-4	-4
PD12	5	-1	3	4	4	2	7	4	4	6	3	11
DD12	8	2	15	5	3	2	6	7	9	12	6	5
WW12	2	8	9	4	-2	1	-1	3	13	8	6	-10
V24	4	-6	-3	-23	6	4	11	1	-13	-14	-1	-0
PD24	-7	-11	-3	0	-0	-5	6	-3	-1	4	-7	-4
DD24	-1	-7	-1	0	-11	-5	-6	-1	-4	-4	-3	-18
WW24	3	10	0	1	-13	4	-13	4	2	-3	5	-11
Max	9	15	15	21	7	11	11	10	13	13	15	20
Min	-7	-11	-4	-23	-13	-5	-13	-5	-13	-14	-7	-18

- implies the model is under-estimating the statistic
+ implies the model is over-estimating the statistic

4.9 CONCLUSIONS AND SUMMARY

The purpose of this Chapter was to find a suitable way to estimate the parameters for the rainfall model. Experiments were performed to test how well the model fitted the historical statistics. It was found that the daily transition probabilities were more suitable than autocorrelations within the fitting procedure, as they improved the model's fit to the historical dry spell sequences.

To further validate the model, comparisons were made between other simulated and historical statistics that were not used in the fitting procedure. For example, the historical and simulated mean and standard deviations of the maxima compared favourably, as did the simulated and historical mean and standard deviations of the proportion of hourly rainfalls above 1mm. However, the simulated and historical mean and standard deviations of the proportions of hourly rainfalls above 0mm did not compare so well, suggesting that the simulated data contained less light rainfall than the historical data. It was decided that this was unlikely to be of practical importance in simulating the hydraulic behavior of storm sewer systems.

In summary, the parameter estimation procedure for hourly rainfall data uses the following historical statistics: i) the mean of the hourly rainfall time series, ii) the variance of the h hourly time series (h=1,3,6,12,24), iii) the wet given wet transition probabilities of the h hourly time series (h=1,3,6,12,24), iv) the proportion of dry days, v) the daily dry given dry transition probability.

CHAPTER 5
FITTING THE MODEL TO DAILY RAINFALL TIME SERIES

5.1 INTRODUCTION

The purpose in this chapter is to find a suitable method of fitting the stochastic model to stations where only daily data are available. This is clearly desirable as most rainfall data are only available as daily time series records. Hence in order to propose a suitable regional and seasonal stochastic rainfall model use must be made of the many daily rainfall records available.

The Chapter concludes by fitting the model to the five longest records of daily rainfall data, and comparing the extremes generated by the model with those of the historical records.

5.2 THE STATISTICS USED IN FITTING

Recall from Chapter 4 that the historical statistics used to fit the model to the hourly rainfall data of each month were:

i) The variances of 1, 3, 6, 12, and 24 hourly time series (denoted V1, V3, V6, V12, and V24 respectively).

ii) The wet given wet and dry given dry transition probabilities from the daily time series.

iii) The proportion of dry days.

Also included in the fitting procedure for hourly data were the wet given wet transition probabilities from the 1, 3, 6, and 12 hourly historical time series (WW1, WW3, WW6, and WW12), but, as mentioned in Section 4.6.1, their inclusion is unlikely to be of practical benefit when fitting the model to daily rainfall data (as the historical WW1 was closely matched by the model even when omitted from the fitting procedure).

5.3 SOME PRELIMINARY INVESTIGATIONS

Given some daily rainfall data it would be convenient if the 1, 3, 6, and 12 hourly variances (V1, V3, V6 and V12) could be reliably predicted for each month, as these variances were shown to be needed in the fitting procedure for hourly data. A linear regression model based on the daily variances may be suitable (this idea came by noticing that the plots of the t-ratios for the variances had similar shapes for different levels of aggregation (e.g. compare Figures C.2, C.5, C.8 and C.11)). To see whether this would work a random sample of 2 stations was taken from each of the homogeneous regions found by *Wigley et al* (1984) (see Figure 5.1). The method of selecting the stations for each region was as follows:

i) The stations within the region were listed.

ii) For each station a random number (between 0 and 1) was generated.

iii) The stations with the highest two random numbers were selected for that region.

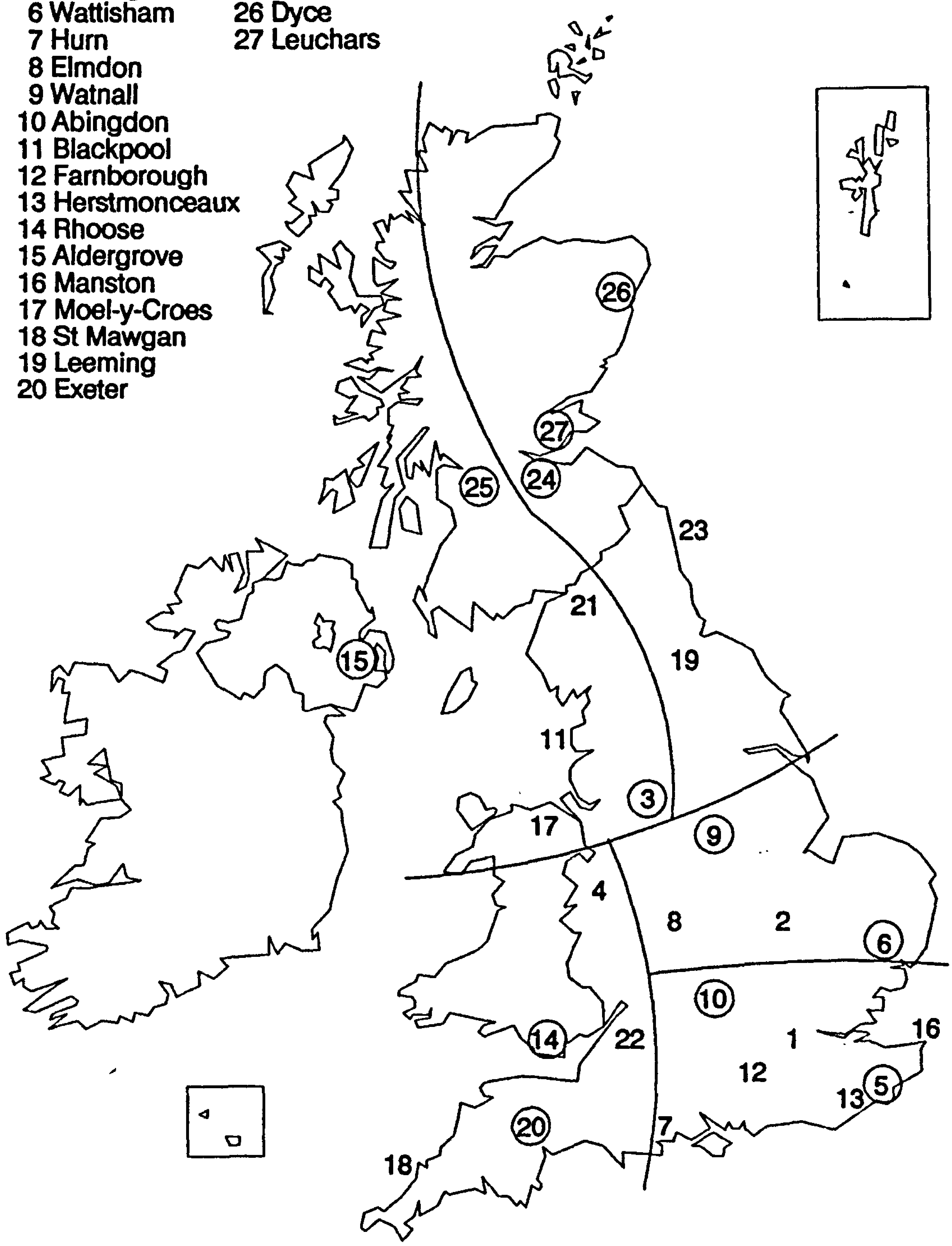
Two further stations were subjectively chosen, one from the North-East region (Dyce) and the other from the North-West region (Aldergrove), as these two stations were located away from the other stations in the same region (e.g. Aldergrove is located in Northern Ireland, whereas the other stations in the North-West region are located on the same land mass, i.e. England and Scotland).

For each member of the random sample, the hourly variances from each month were plotted against their corresponding daily variances (see Figures 5.2-5.13). From these Figures two conclusions were drawn. First a linear regression model of the form $y = \alpha + \beta x$ seemed appropriate for most of the stations, and secondly the variances of the residuals (about such a line) looked as if they could not be regarded as equal for all stations (e.g. compare Leuchars with East Kilbride). Hence any method of clustering these stations into suitable groups should not include the assumption of homogeneity of variance. For this reason some theoretical derivations are given below which add to statistical theory on Cluster Analysis.

Figure 5.1
Hourly Rainfall Stations used for Regression Analysis

(showing homogeneous regions proposed by Wigley et al (1984), and sampled stations (ringed))

1 Hampstead
2 Bedford
3 Ringway
4 Shawbury
5 Hastings
6 Wattisham
7 Hurn
8 Elmdon
9 Watnall
10 Abingdon
11 Blackpool
12 Farnborough
13 Herstmonceaux
14 Rhoose
15 Aldergrove
16 Manston
17 Moel-y-Croes
18 St Mawgan
19 Leeming
20 Exeter
21 Carlisle
22 Filton
23 Boulmer
24 Turnhouse
25 East Kilbride
26 Dyce
27 Leuchars



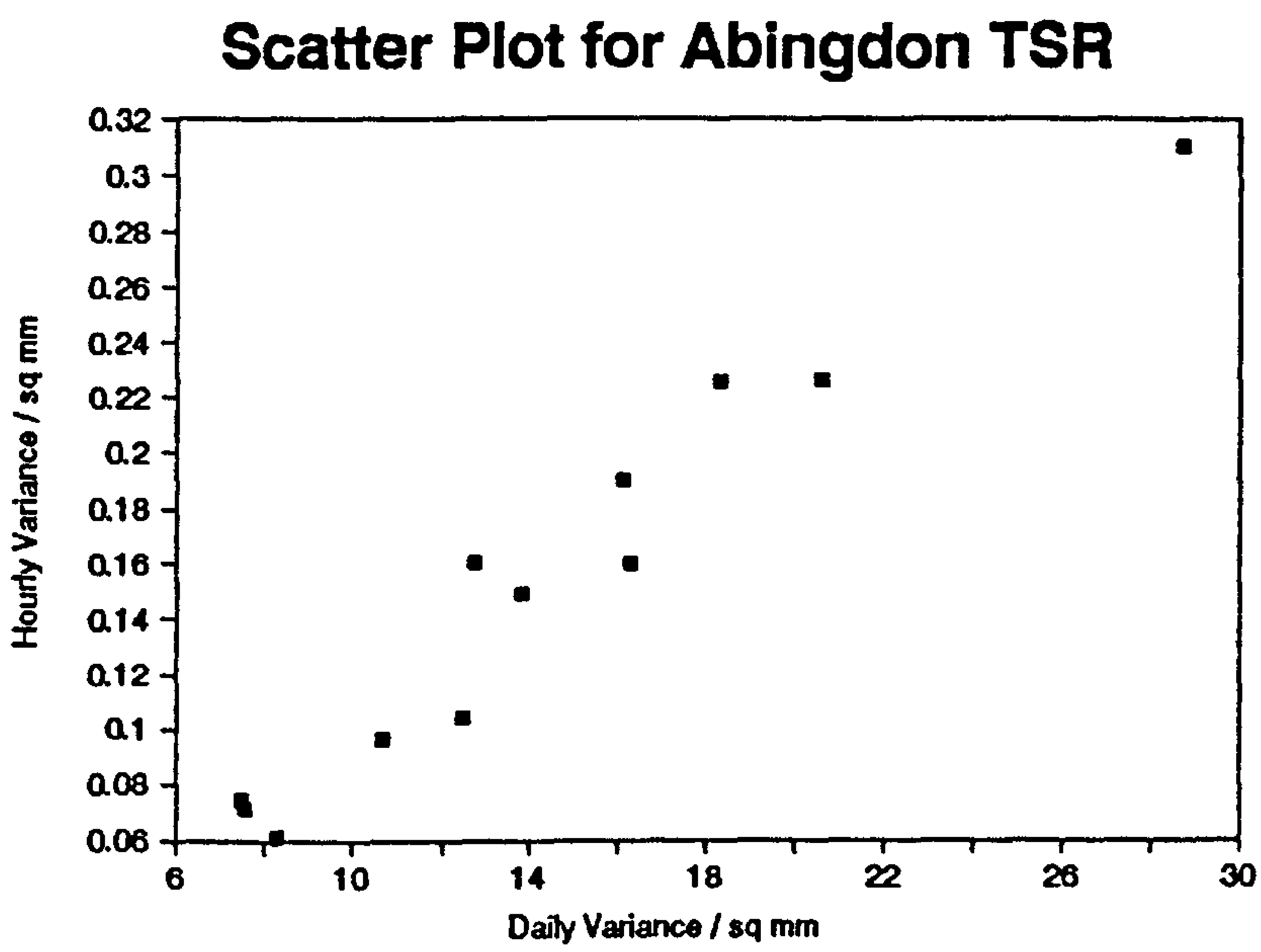


Fig. 5.2

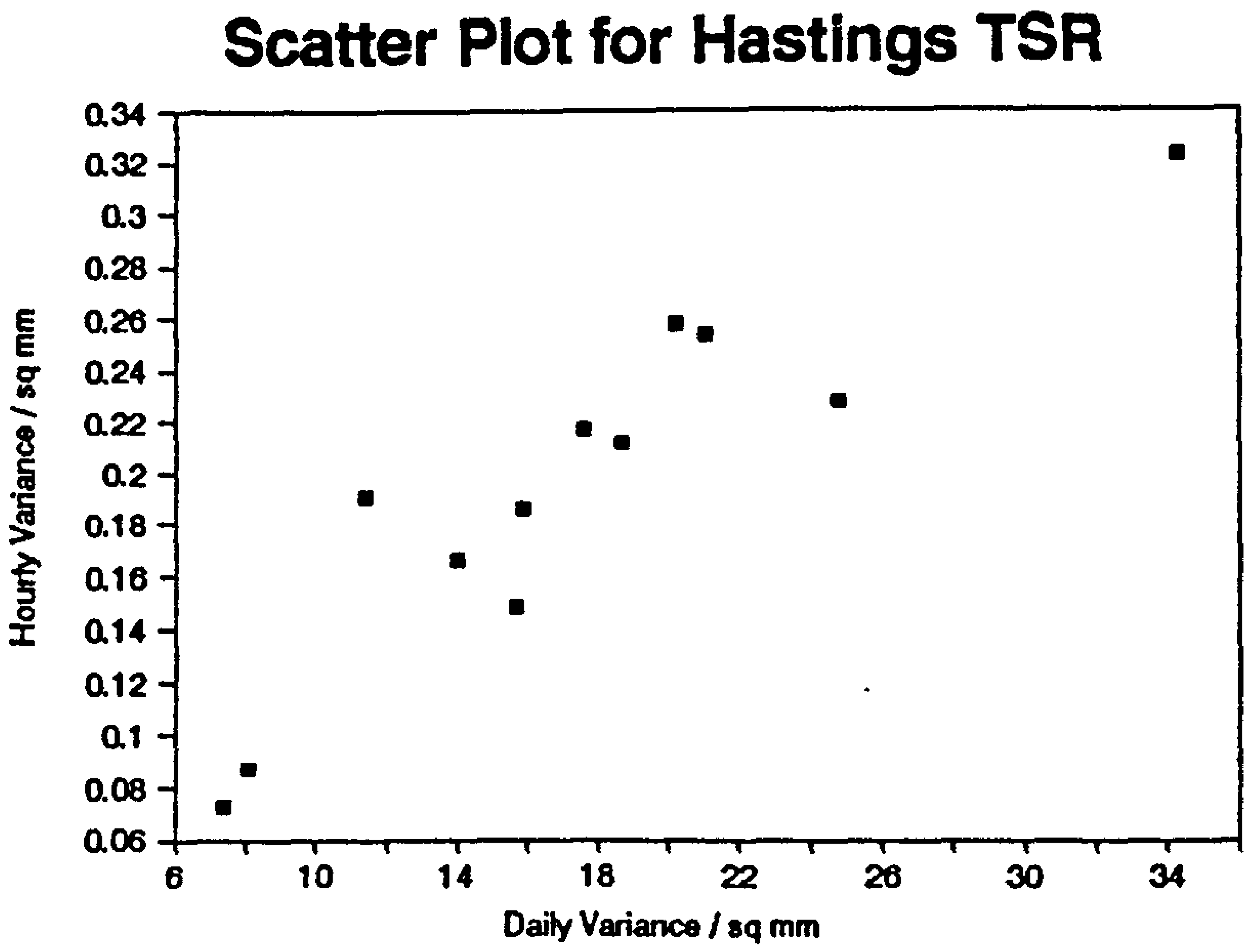


Fig. 5.3

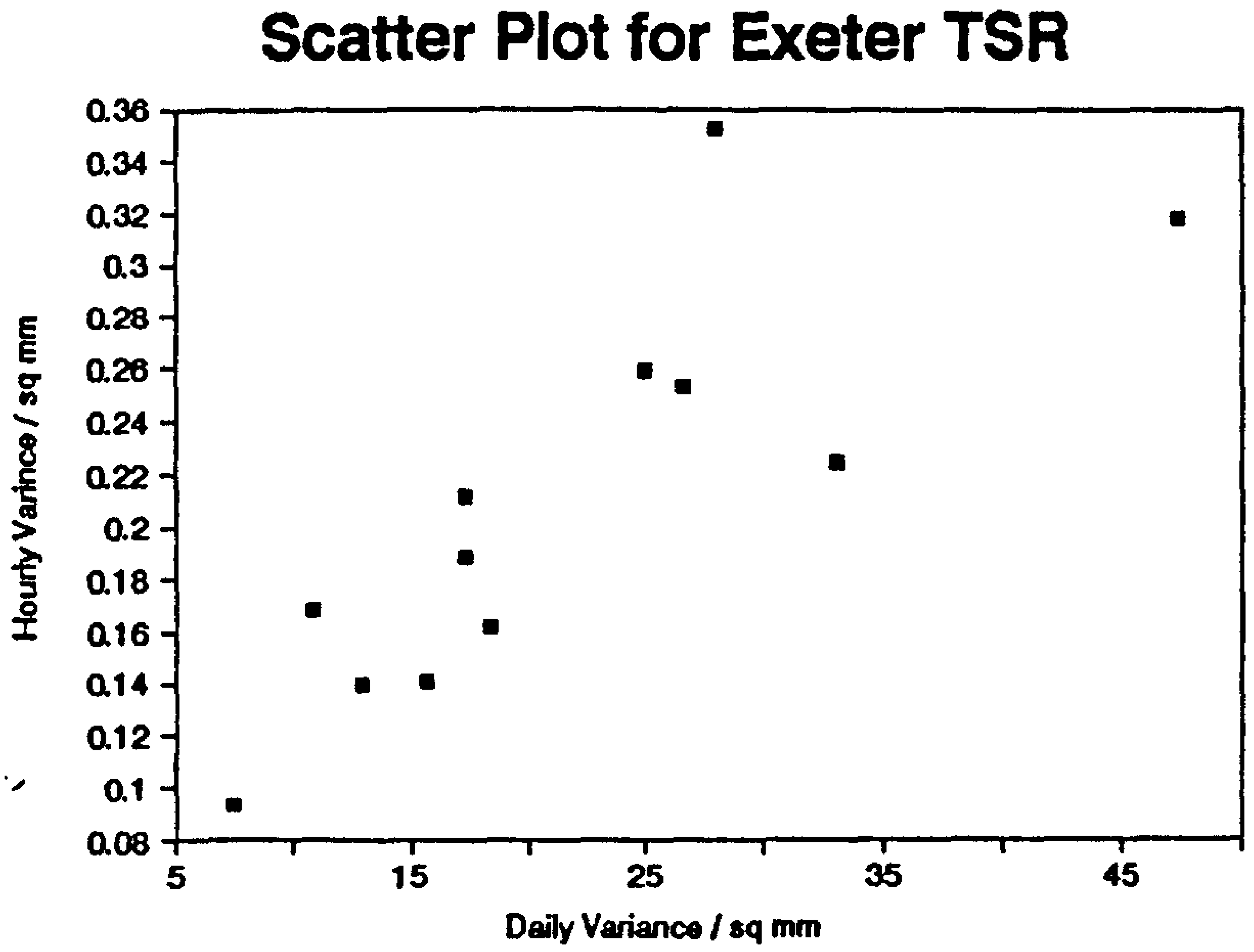


Fig. 5.4

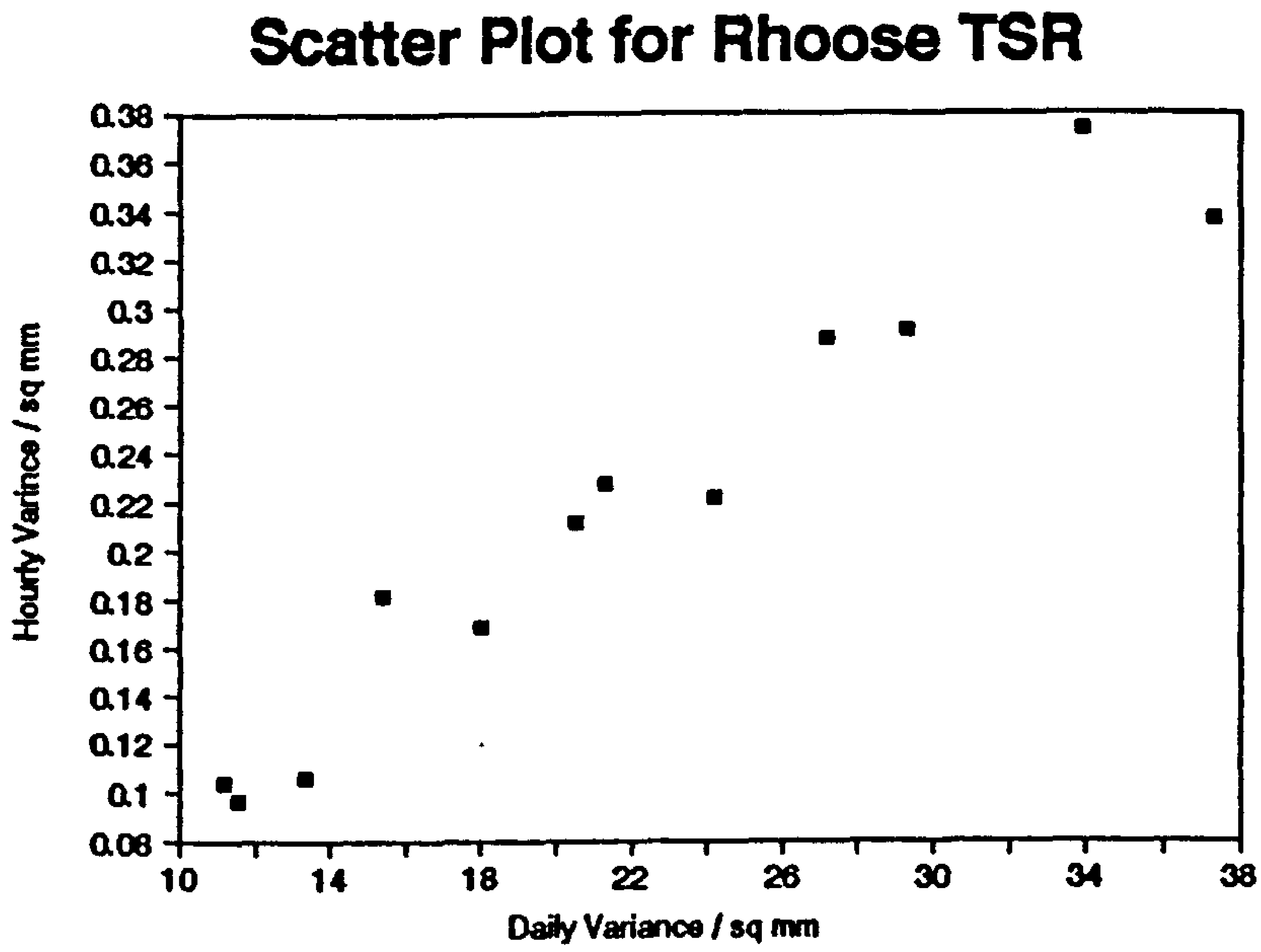


Fig. 5.5

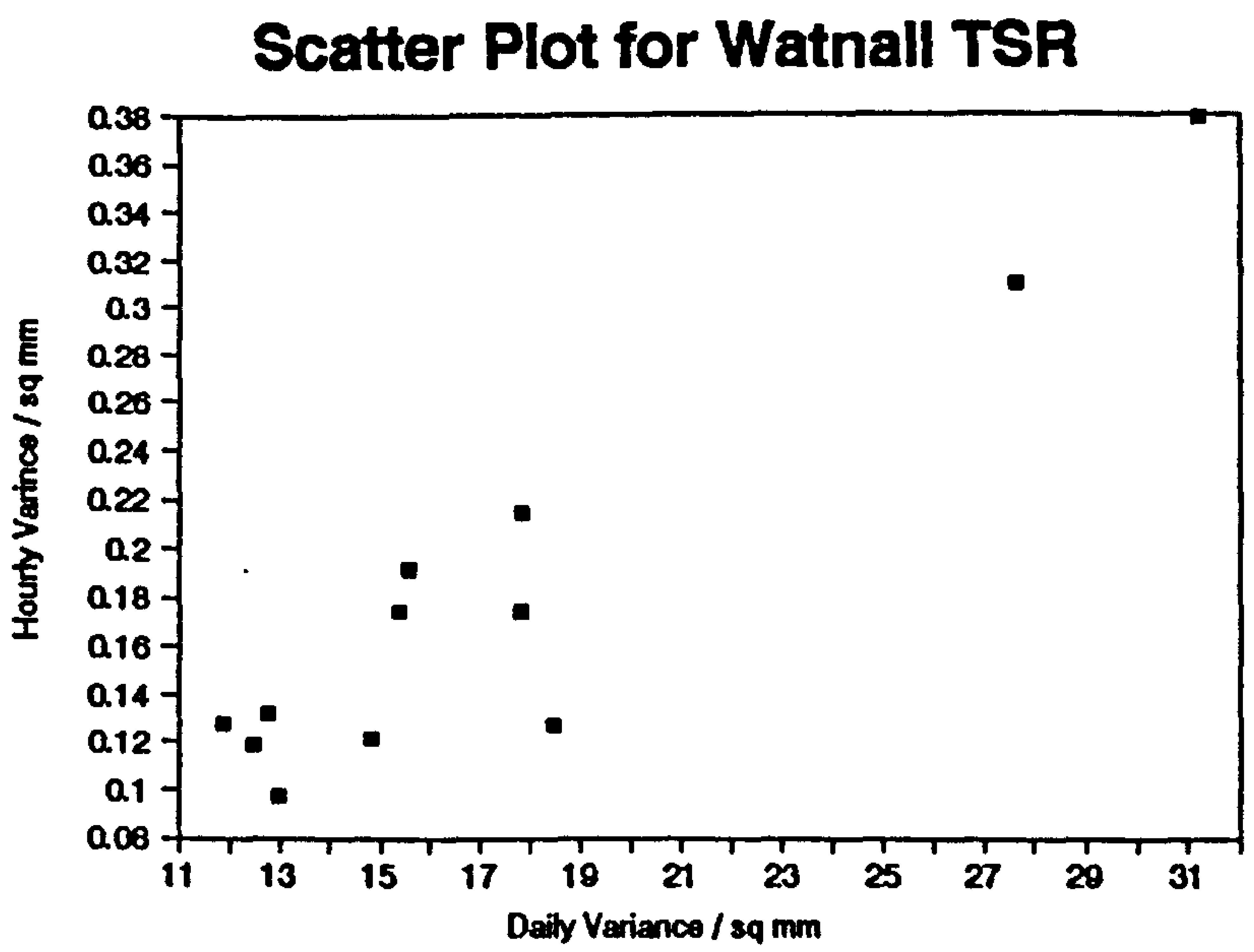


Fig. 5.6

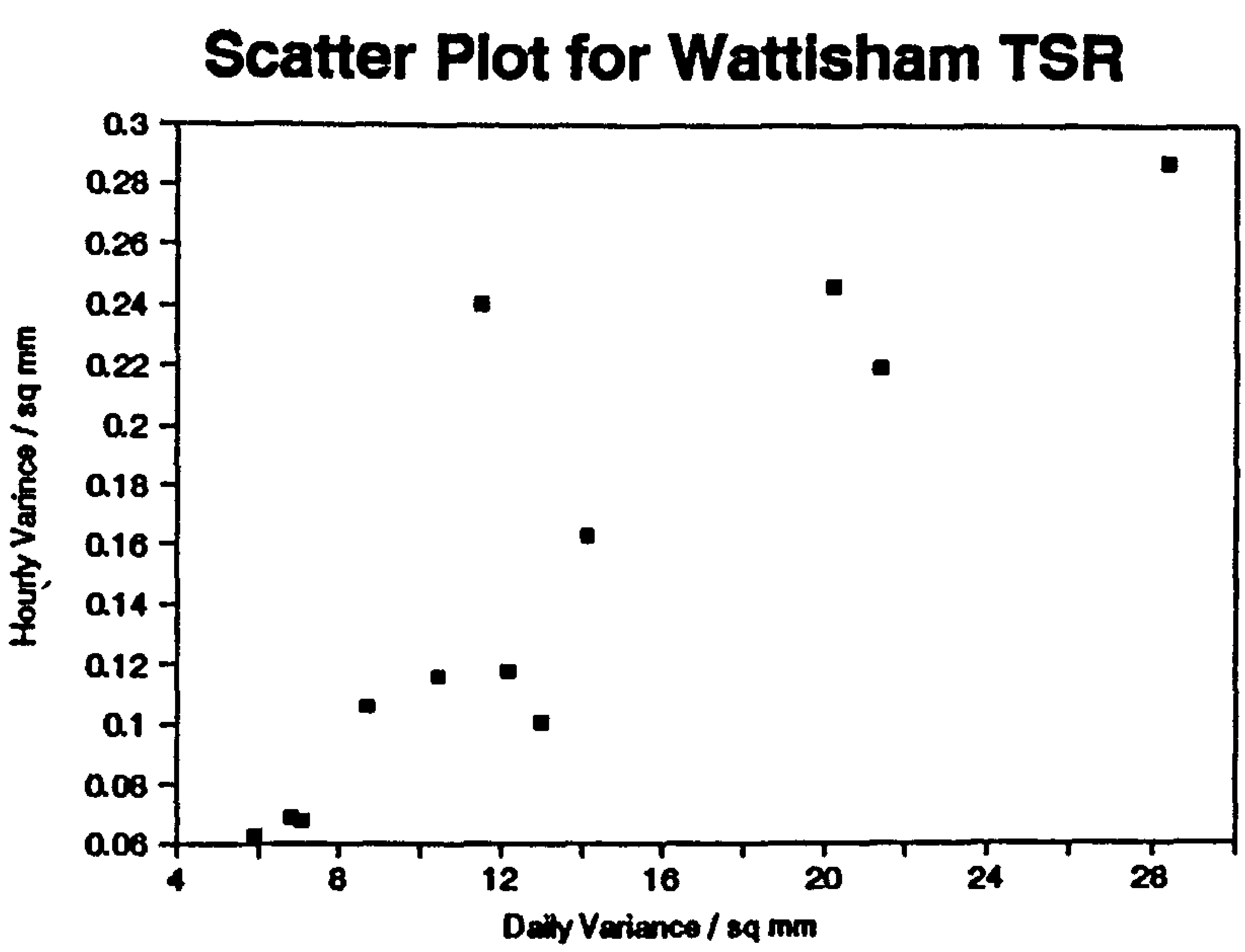


Fig. 5.7

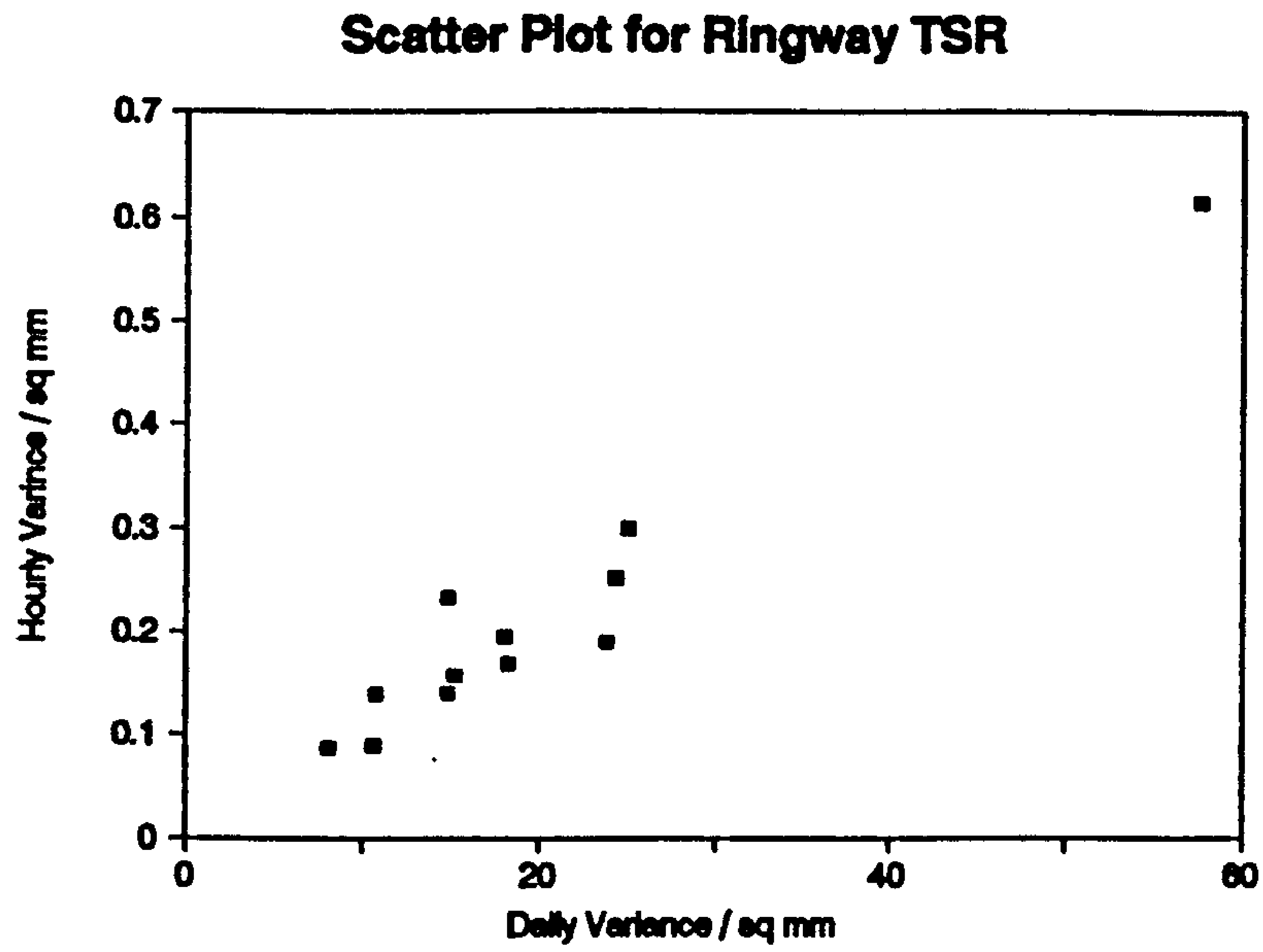


Fig. 5.8

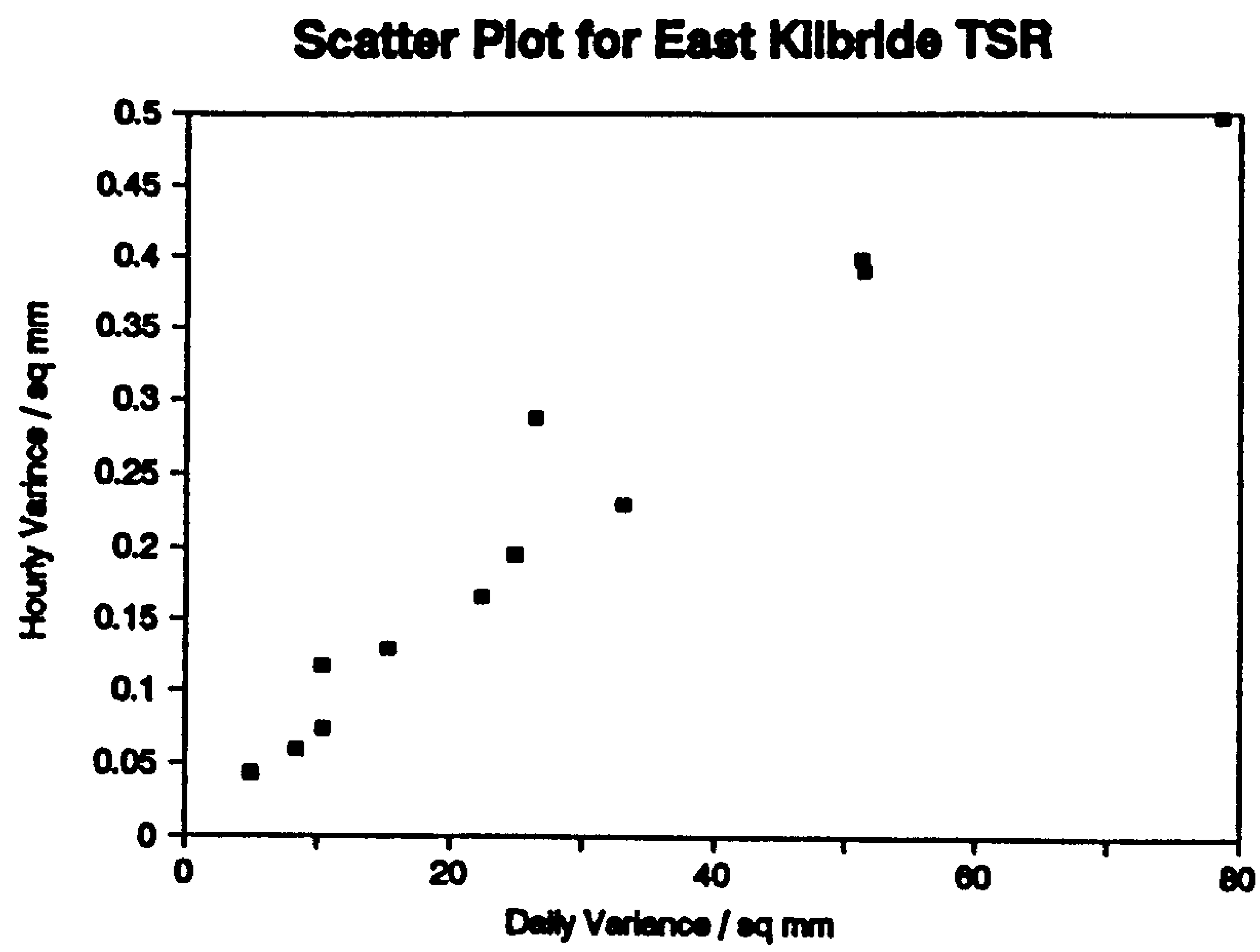


Fig. 5.9

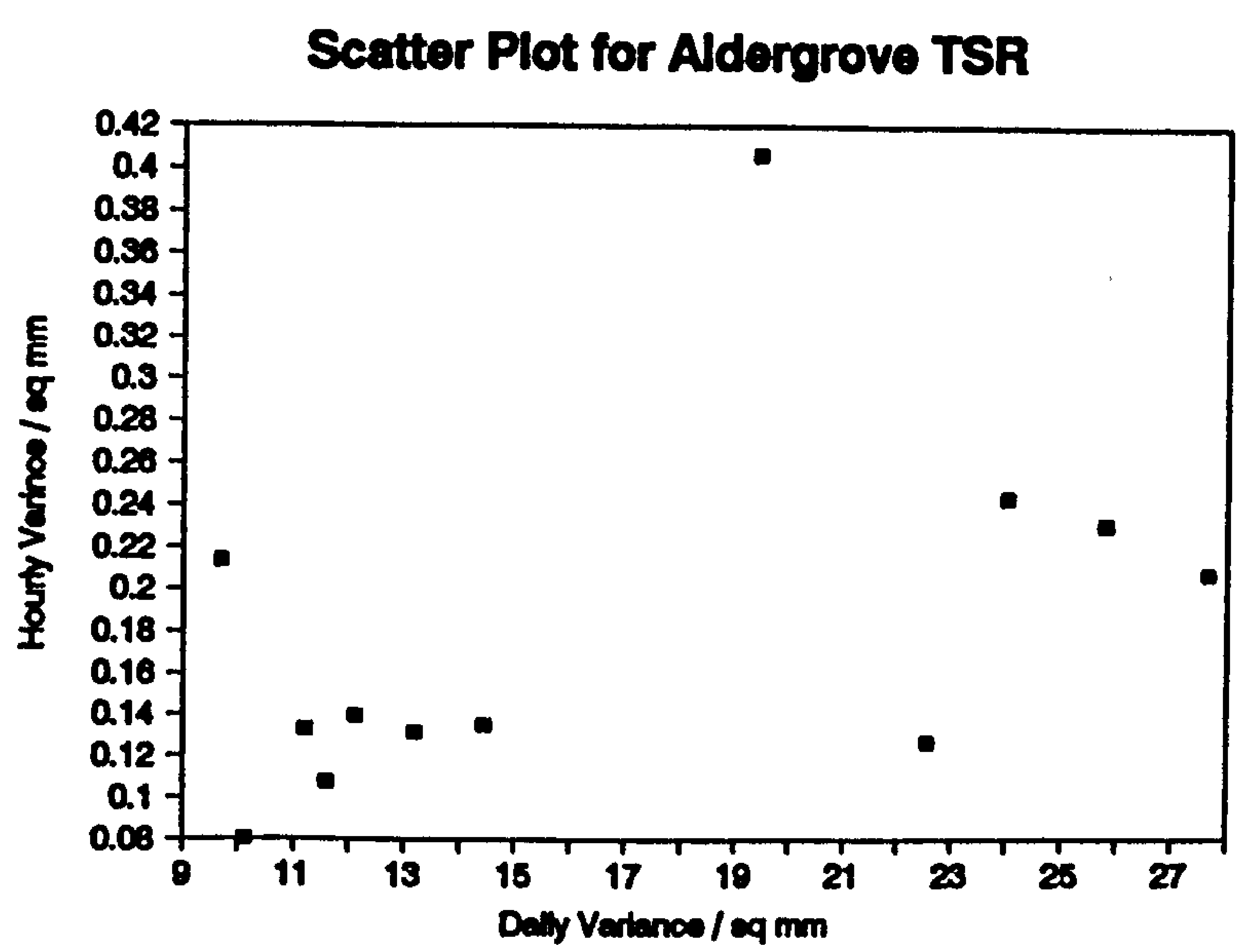


Fig. 5.10

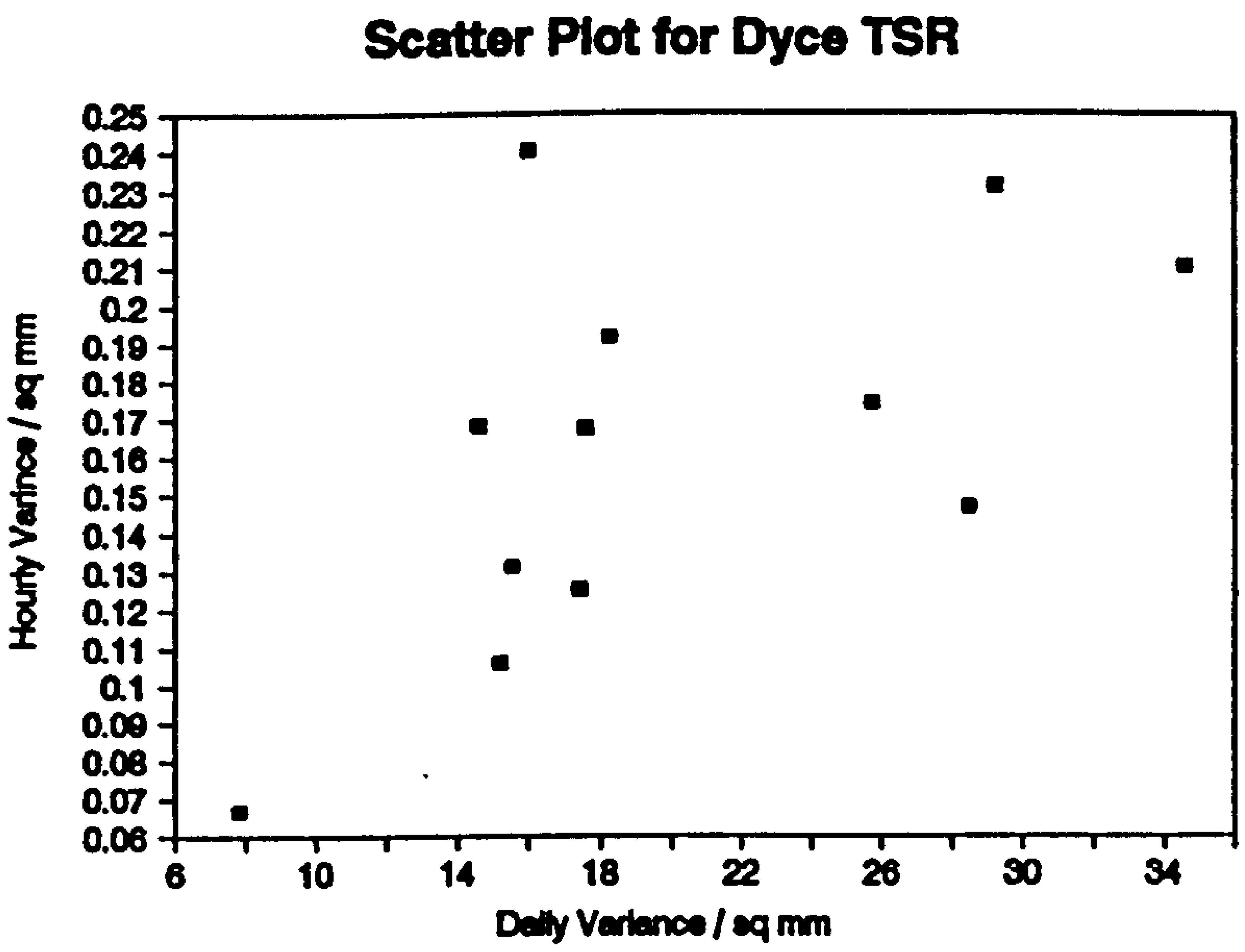


Fig. 5.11

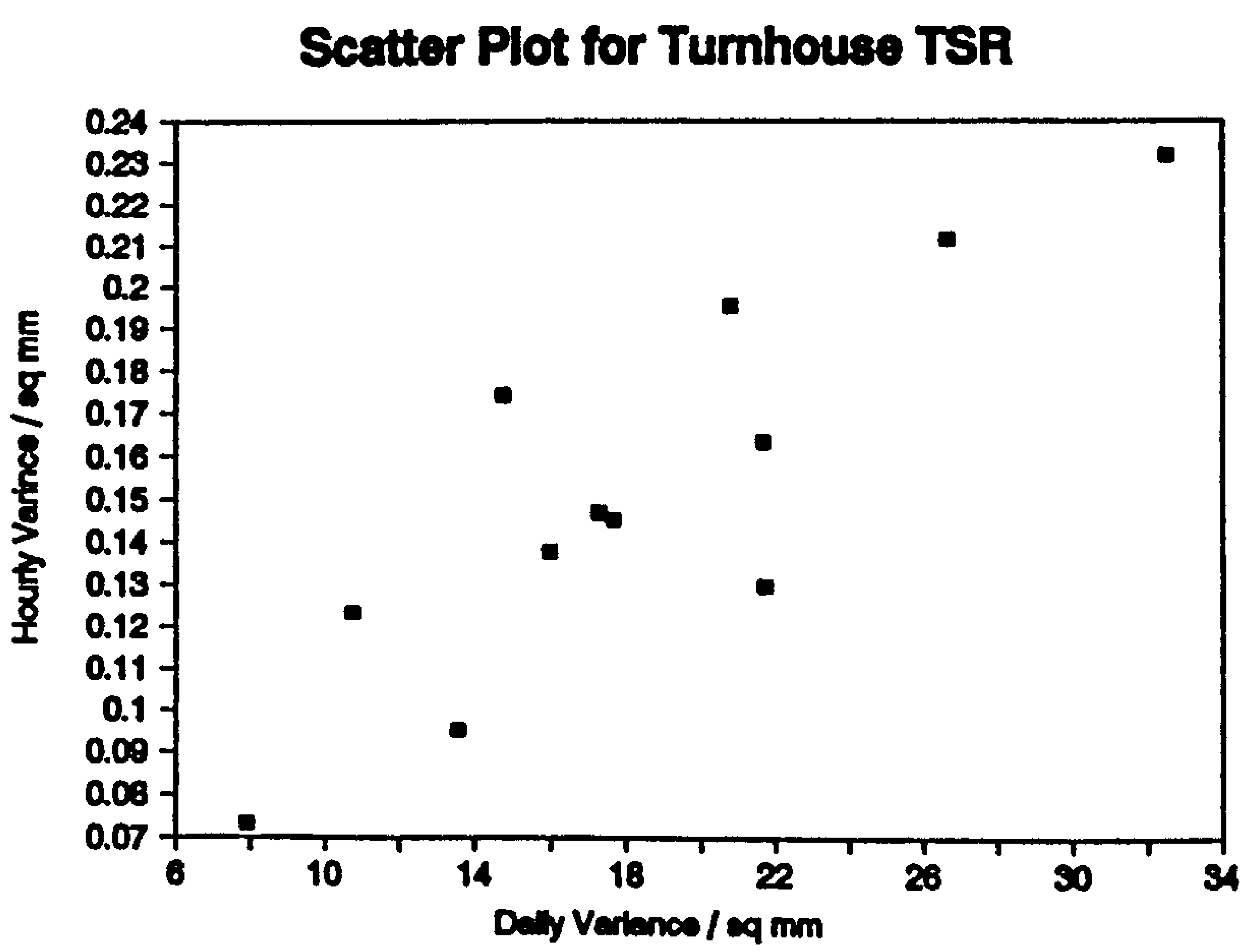


Fig. 5.12

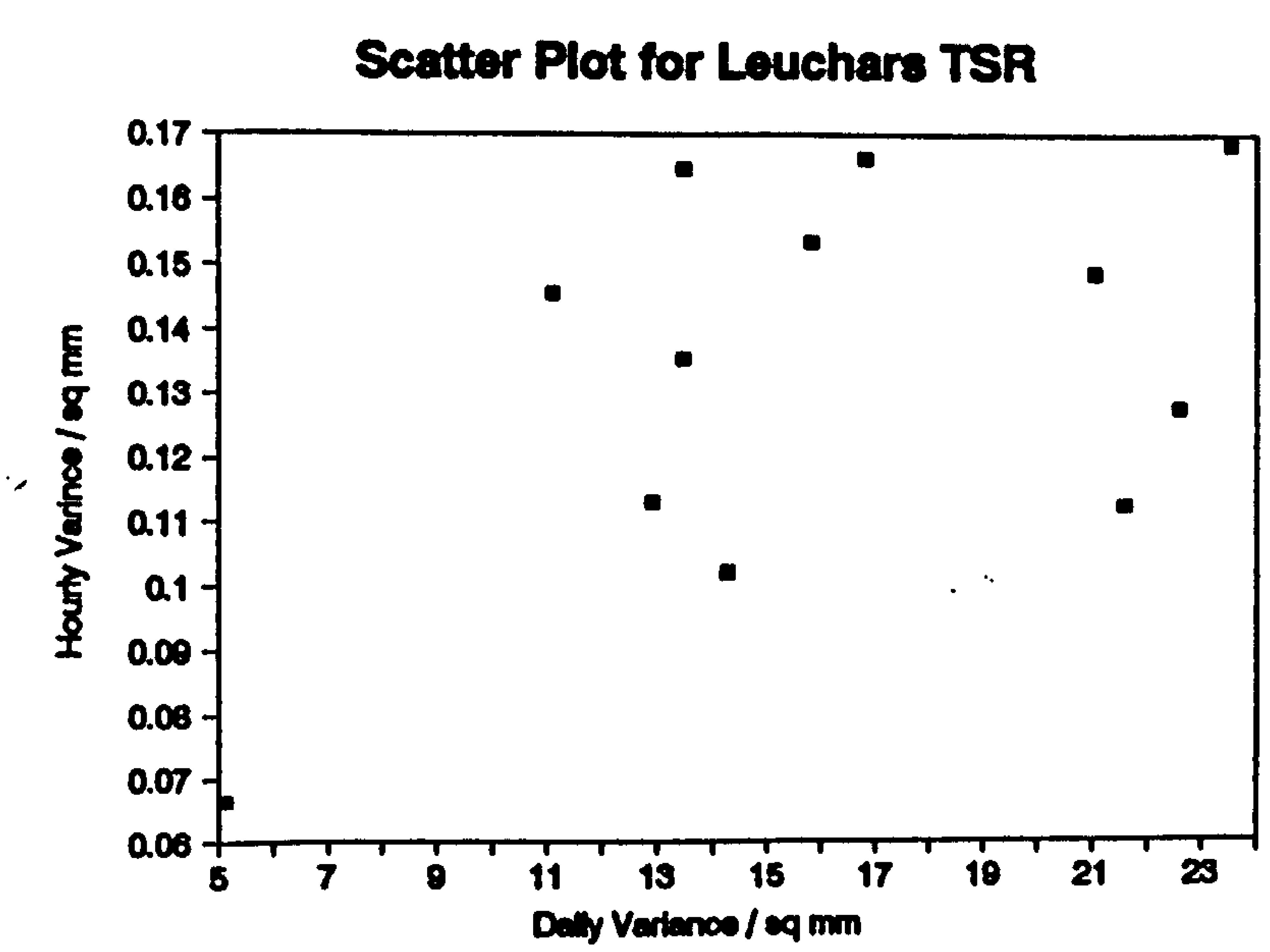


Fig. 5.13

5.4 CLUSTERING STATISTICS UNDER HETEROGENEITY OF VARIANCE

5.4.1 Introduction

The problem of testing the equality of population means under heterogeneity of variance is not new. The case of 2 population means (the Behrens-Fisher Problem) has been extensively researched (see, for example, *Fisher* (1935), *Welch* (1947), *Aspin* (1948) and *Cox and Jaber* (1990)). For the case of more than 2 population means, the problem is less well researched, but *Welch* (1951), *James* (1951), and *Jaber* (1984), offer statistics on which the test can be based. In this Section the *Welch* (1951) statistic for testing the equality of T population means under heterogeneity of variance is extended to the case of partitioning the means into non-overlapping homogeneous groups. Methods are available for partitioning population means when homogeneity of variance can be assumed (or the ratios of the population variances are known) and for this case the reader is referred to *Calinski and Corsten* (1985), *Cox and Spjotvoll* (1982) or *Scott and Knott* (1974).

5.4.2 An Extension of Welch's Statistic

Let a_i $(i = 1, \ldots, T)$ be statistical quantities normally and independently distributed with means μ_i and variances $\zeta_i\sigma_i^2$, where the ζ_i are known constants but nothing is known about μ_i and σ_i^2. Suppose that the data provide estimates s_i^2 of the σ_i^2 which are distributed respectively as $\chi_i^2\sigma_i^2/f_i$, where f_i is the number of degrees of freedom of χ_i^2. Suppose further that the s_i^2 are distributed independently of each other and of all the a_i. The first consideration will be the hypothesis H_0: $\mu_i = \mu$ $(i=1,2,...,T)$, i.e. that all the population means are equal.

If this hypothesis is rejected alternative hypotheses of the form:

$$H: \mu_i = \mu^{(1)} \ (i \in I_1),\ \mu_i = \mu^{(2)} \ (i \in I_2),\ ...,\ \mu_i = \mu^{(n)} \ (i \in I_n),$$

will be considered, i.e. that the population means can be partitioned into n groups of sizes $|I_1|$, $|I_2|$, ..., $|I_n|$ respectively, where I_1, I_2, ..., I_n are mutually exclusive and exhaustive subsets of $\{1, 2, ..., T\}$. This is the problem of cluster analysis under non-homogeneity of variance, of which little is known in the literature. In this Section Welch's result will be extended to cover the problem of clustering the population means into groups.

The results in this Section could be used in several different contexts. For example the a_i may be the sample means from T different normal populations, whose true means and variances are μ_i and $\zeta_i\sigma_i^2$ respectively, where, in this case, ζ_i = 1/(ith sample size). Alternatively the a_i could be the constant (or slope) parameters for T regression models, where the a_i are normally distributed with means μ_i and variances $\zeta_i\sigma_i^2$, where again ζ_i is a known constant from regression theory.

Welch (1951) considers the statistic $\sum w_i(a_i - \hat{a})^2$, where $w_i = \zeta_i^{-1}s_i^{-2}$, and $\hat{a}$ is the weighted average $(\sum w_i a_i)/\sum w_i$. Clearly Welch's statistic is measuring the overall departure of the a_i from this weighted average taking into consideration the sample variances associated with the a_i. The cumulant-generating function

K(t) to order $1/f_i$ of $\sum w_i(a_i - \hat{a})^2$ is given by:

$$K(t) = -\frac{1}{2}(T-1)\ln(1-2t) + \{2t(1-2t)^{-1} + 3t^2(1-2t)^{-2}\}\{\sum \frac{1}{f_i}(1-\frac{\omega_i}{\sum \omega_i})^2\} \quad (5.1)$$

(*Welch* (1951) equation (19)), where $\omega_i = \zeta^{-1}\sigma^{-2}$ (the population equivalent to w_i).

The cumulant-generating function for $G = (\nu_1 - 1 + A/\nu_2)F$, where F is distributed as $F(\nu_1 - 1, \nu_2)$, can also be found to order $1/\nu_2$ and is given by:

$$K_G(t) = -\frac{1}{2}(\nu_1 - 1)\ln(1-2t) + \frac{1}{\nu_2}(A + 2\nu_1 - 2)t(1-2t)^{-1} + \frac{(\nu_1^2 - 1)}{\nu_2}t^2(1-2t)^{-2}$$

(*Welch* (1951) equation (24)) (5.2)

Hence (5.1) and (5.2) are equivalent if:

$$T = \nu_1$$

$$(A + 2\nu_1 - 2)/\nu_2 = 2\sum_i \frac{1}{f_i}(1 - \frac{\omega_i}{\sum \omega_i})^2$$

and

$$\frac{(\nu_1^2 - 1)}{\nu_2} = 3\sum_i \frac{1}{f_i}(1 - \frac{\omega_i}{\sum \omega_i})^2$$

i.e. if:

$$1/\nu_2 = \frac{3}{(T^2 - 1)}\sum_i \frac{1}{f_i}(1 - \frac{\omega_i}{\sum \omega_i})^2 \quad (5.3a)$$

and

$$A/\nu_2 = \frac{2(T-2)}{(T+1)}\sum_i \frac{1}{f_i}(1 - \frac{\omega_i}{\sum \omega_i})^2 \quad (5.3b)$$

(*Welch* (1951) equation (26))

Hence $\sum w_i(a_i - \hat{a})^2/(T-1 + A/\nu_2)$ is approximately distributed as $F(T-1,\nu_2)$, where ν_2 and A are given in equations (5.3a) and (5.3b) respectively.

In order to extend Welch's result to the more general case of partitioning the population means into non-overlapping groups, consider the statistic P_j given by:

$$P_j = \sum_{I_j} w_i(a_i - \hat{a}_j)^2, \text{ where } \hat{a}_j = (\sum_{I_j} w_i a_i)/\sum_{I_j} w_i,$$

so that the summations are taken over all values in group j (j=1,2,...,n). For convenience the set symbol I_j may be dropped and the summation written as $\sum_{(j)}$. Let $n_j = |I_j|$ be the number of means in group j (for j = 1,...,n).

The cumulant-generating function $K_j(t)$ of P_j follows immediately from (5.1) and is given by:

$$K_j(t) = -\frac{1}{2}(n_j - 1)\ln(1-2t) + \{2t(1-2t)^{-1} + 3t^2(1-2t)^{-2}\}\ \{\sum_{(j)} \frac{1}{f_i}(1 - \frac{\omega_i}{\sum_{(j)} \omega_i})^2\} \tag{5.4}$$

Hence the cumulant-generating function of $P = \sum_{j=1}^{n} \sum_{(j)} w_i(a_i - \hat{a}_j)^2 = \sum_{j=1}^{n} P_j = K_P(t)$, say, is given by:

$$K_P(t) = \sum_{j=1}^{n} K_j(t) = -\frac{1}{2}(T-n)\ln(1-2t) + [2t(1-2t)^{-1} + 3t^2(1-2t)^{-2}] \sum_{j=1}^{n} \sum_{(j)} \frac{1}{f_i}(1 - \frac{\omega_i}{\sum_{(j)} \omega_i})^2 \tag{5.5}$$

Now (5.5) is equivalent to (5.2) if:

$$\nu_1 = T - n + 1$$

$$1/\nu_2 = 3\sum_{j=1}^{n} \sum_{(j)} \frac{1}{f_i}(1 - \frac{\omega_i}{\sum_{(j)} \omega_i})^2 / \{(T-n)(T-n+2)\} \qquad (5.6a)$$

and

$$A/\nu_2 = \frac{2(T-n-1)}{(T-n+2)} \sum_{j=1}^{n} \sum_{(j)} \frac{1}{f_i}(1 - \frac{\omega_i}{\sum \omega_i})^2 \qquad (5.6b)$$

Hence using (5.6a) and (5.6b) we have:

$$\frac{\sum_{j=1}^{n} \sum_{(j)} w_i (a_i - \hat{a}_j)^2 / (T-n)}{1 + \frac{2(T-n-1)}{(T-n)(T-n+2)} \sum_{j=1}^{n} \sum_{(j)} \frac{1}{f_i}(1 - \frac{\omega_i}{\sum_{(j)} \omega_i})^2} = W, \text{ say, is approximately distributed as } F_{T-n,\nu_2}$$

(5.7a)

where
$$\nu_2^{-1} = \frac{3 \sum_{j=1}^{n} \sum_{(j)} \frac{1}{f_i}(1 - \frac{\omega_i}{\sum_{(j)} \omega_i})^2}{(T - n)(T - n + 2)}$$

(5.7b)

The W statistic (5.7a) can be used to partition the population means into non-overlapping groups without assuming homogeneity of variance. In the next Section W will be used to group the regression parameters for each station into non-overlapping groups.

5.5 REGRESSING THE HOURLY VARIANCE ON THE DAILY VARIANCE

Let x_{ij} and y_{ij} be the daily and hourly variances (respectively) for month j (=1,...,12), station i (=1,...,27). For each rainfall station a linear regression model of the form:

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

was fitted using the method of least squares, where it was assumed that the ε_{ij} are independent Normal random variables with constant variance σ_i^2 for each station (an analysis of the residuals appears later in this Chapter).

The hypothesis $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_{27}^2$ was tested using the statistic:

$$\chi^2 = (Tr_i - 2T)\ \ln\left(\frac{\sum_{i=1}^{T}(r_i - 2)\hat{\sigma}_i^2}{Tr_i - 2T}\right) - \sum_{i=1}^{T}(r_i - 2)\ \ln\hat{\sigma}_i^2$$

where in this case T = total number of stations = 27, r_i = number of observations for station i (which was 12 for each station). This statistic is in common use in hydrology (e.g. see *Holder* (1985)) and is approximately χ^2_{T-1}. The value taken by χ^2 was 97.3 which is significant at the 0.1% level.

Although the hypothesis of homogeneity of variance was rejected, a standard ANOVA for regression was performed, so that some comparisons could be made between the results of hypotheses tests under the standard ANOVA and the results of hypotheses tests when using the W statistic of the previous Section. The results of calculating the relevant sums of squares for the ANOVA are given in Table 5.1.

Table 5.1 A Standard ANOVA for Regression Analysis

Source	SS	df	MS	F	0.1% level
Overall Regression	1.95	1	1.95	1091	11.
Difference in intercepts	0.115	26	0.00441	2.47	2.1
Difference in gradients	0.139	26	0.00534	2.99	2.1
Residual	0.482	270	0.00178		
Total	2.68	323			

In a regression analysis of this type the following hypotheses are of interest:

i) the hypothesis $\alpha_i = \alpha$ (i=1,...,27), i.e. all the intercepts are equal,

ii) the hypothesis $\beta_i = \beta$ (i=1,...,27), i.e. all the gradients are equal, and

iii) the hypothesis of no linear relationship.

Using Table 1 each of the above would be rejected at the 0.1% level. However, using the W statistic (equation (5.7a)) gives:

$W_1 = 1.2 \sim F_{26,18}$ under (i), and
$W_2 = 3.7 \sim F_{26,112}$ under (ii) above,
so that hypothesis (i) could be retained ($F_{26,18}(10\%) \simeq 1.8$).

For convenience, the common intercept α, was estimated by the mean value, i.e. $\hat{\alpha} = \sum_{i=1}^{27} \alpha_i/T$. The parameters β_i were then re-estimated for each station using the revised model:

$$y'_{ij} = y_{ij} - \hat{\alpha} = \beta_i x_{ij} + \varepsilon_{ij}.$$

The values of the estimates of β_i were then ordered and are shown in Figure 5.14.

The hypothesis H_0: $\beta_i = \beta$ (i=1,...,27) was again tested using the W statistic, which gave the value $W = 4.9 \sim F_{26,97}$ under H_0, and so H_0 is rejected at the 0.1% level.

Having rejected H_0 some groupings suggested by Figure 5.14 were tested[1]. The first partition tested was

$$H: \beta_i = \beta^{(1)} \text{ (i=1,...,22)}; \ \beta_i = \beta^{(2)} \text{ (i=23,...,27)}$$

For this grouping $W = 1.76 \sim F_{26,96}$ under H. This result is just significant at the 5% level. However, this hypothesis was retained for two reasons: a) the differences within each of these groups were probably not of much practical significance, and b) it could be seen in Figure 5.1 that this choice of groups gave stations that fall naturally into homogeneous regions, i.e. one group corresponds to the rainfall stations in England, Wales, and Northern Ireland, and the other group corresponds to stations in

[1] It should be mentioned that using the data to suggest possible groupings is not entirely satisfactory from a statistical point of view. However, this approach was adopted to find the smallest number of groups that could be regarded as homogeneous (which would reduce the problem, for engineers with sites near a boundary, of having to choose between regions). Regions could have been selected, prior to the statistical tests, based on physical grounds but this may have lead to a large number of regions due to the highly variable physical topography of the British landscape.

Scotland (together with Boulmer in the far north east of England). As it may be convenient to pool the Boulmer station (number 23) with the rest of the stations in England a further test was made:

$$H: \beta_i = \beta^{(1)} \ (i=1,\ldots,23); \ \beta_i = \beta^{(2)} \ (i=24,25,26,27)$$

was tested using the W statistic, and the result was W = 2.44 ~ $F_{26,96}$ under H. This was rejected at the 0.1% level, so that Boulmer was included with the rainfall stations in Scotland. The homogeneous regions for the regression models are shown in Figure 5.15.

A further test of the significance of this region was made by introducing an explanatory variable (N), which 'explained' the regions A and B, into a regression model for all the station-months, and testing the significance of this variable, i.e. the following regression model was fitted by the method of least squares:

$$y_{ij} = \alpha_0 + \alpha_1 x_{ij} + \alpha_2 N_{ij} + \varepsilon_{ij},$$

where x_{ij} is the daily variance for month j (=1, ..., 12) station i (=1, ..., 27), y_{ij} is the corresponding hourly variance, and N_{ij} is 1 if the ith station is in region B or 0 if the ith station is in region A.

Table 5.2 gives the least squares estimates of the parameters α_0, α_1, and α_2 of the above regression model, and the standard errors of these estimates. In addition, a t-ratio is given to test the null hypothesis that the explanatory variable has no effect, i.e. to test H_0: $\alpha_i = 0$ (i=0,1,2).

Table 5.2

The regression parameter estimates
with a regional explanatory variable included

Regression parameter	Least squares estimate	Standard error	t-ratio
α_0	0.04281	0.00539	8.0
α_1	0.00798	0.00025	31.
α_2	-0.04111	0.00647	-6.4

$$t_{0.1\%} = 3.7$$

From Table 5.2 it was evident that the effect of the regional variable N (corresponding to regression parameter α_2) is highly significant, and so the choice of the homogeneous regions A and B was retained. Furthermore, the coefficient of variation (= estimate of residual standard deviation ÷ mean of hourly variances) was 25%, which compared favourably to the coefficient of variation for the hourly variances in the Manston data set (which was 22% when averaged over the months), so that the residual variation could mainly be attributed to sampling variability rather than model inadequacy. It should perhaps be mentioned that the daily variances are also subject to sampling variability - the coefficient of variation for these lying somewhere between 20% and 30% for a 10 to 20 year record (refer back to Section 4.3.4 of Chapter 4 for the method by which the coefficients of variation were found).

It is worth noting that a similar choice of homogeneous regions was made in the Flood Studies Report (1975, Volume 2, Section 1.5), where two regions were proposed: i) Scotland and Northern Ireland and ii) England and Wales.

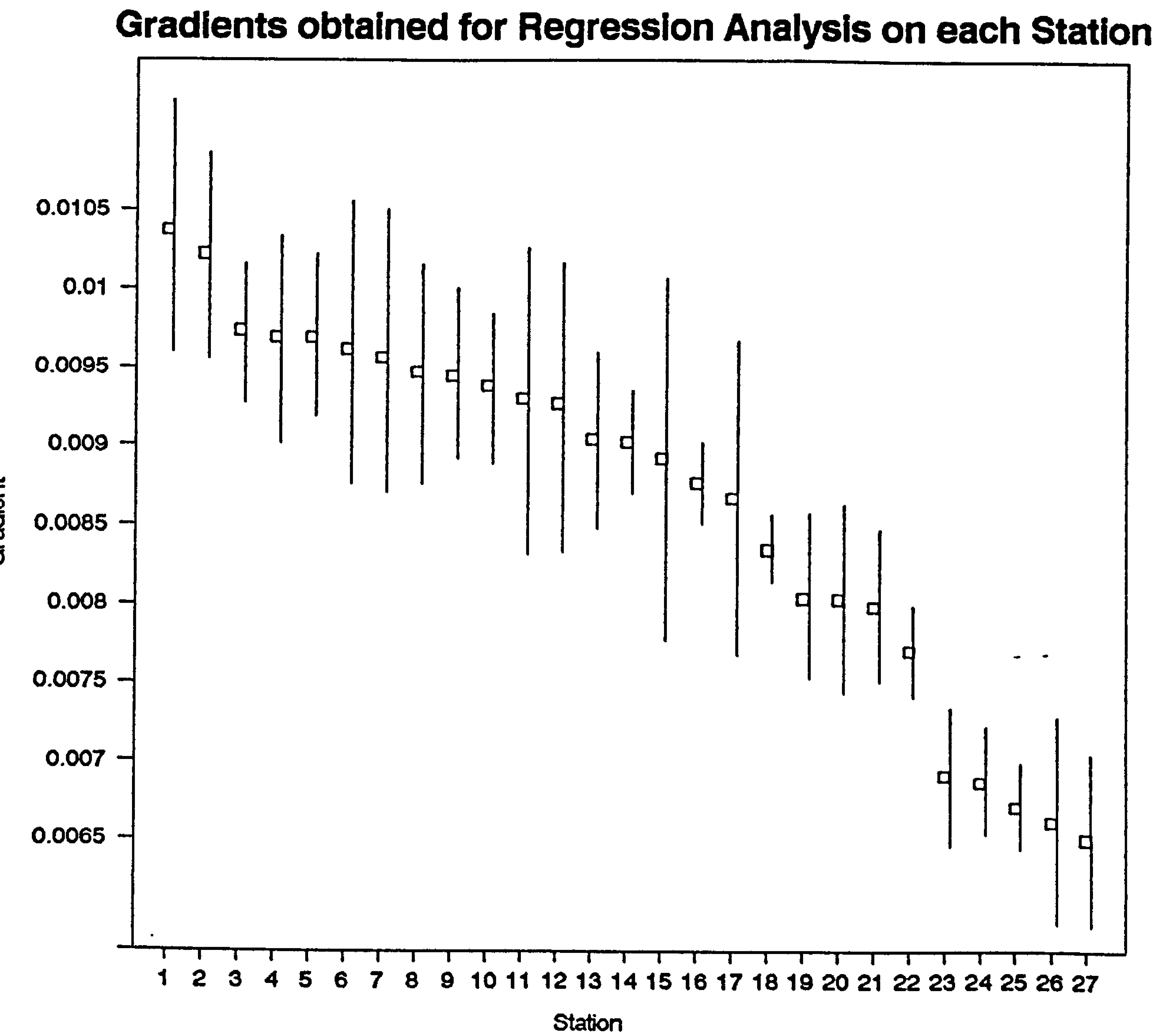


Figure 5.14
Gradients for the regression analysis
with 1 standard deviation shown either side

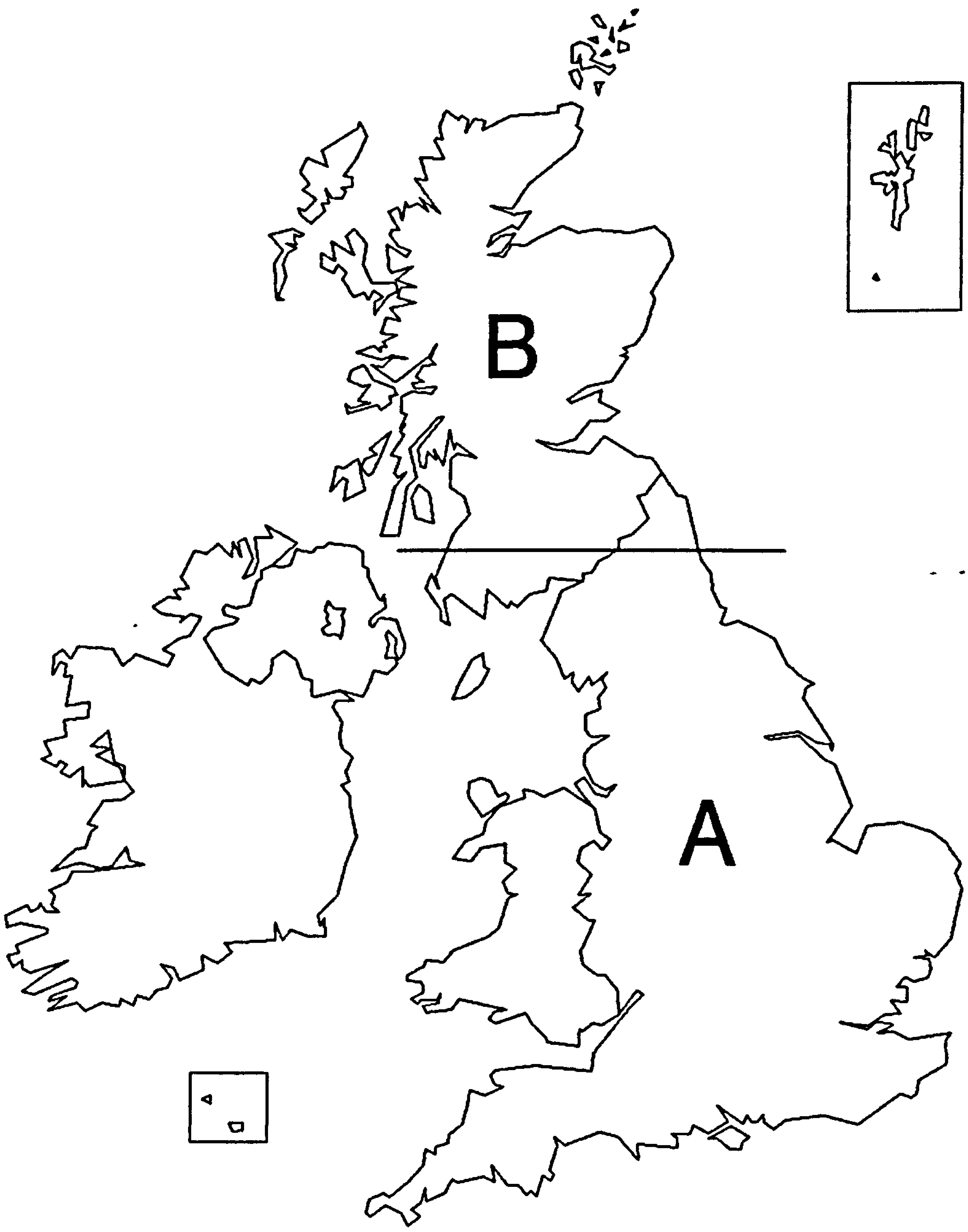


Figure 5.15
Homogeneous regions A and B
found in cluster analysis

5.6 REGRESSING V3, V6, AND V12 ON THE DAILY VARIANCE

Pooling the rainfall stations into the two groups suggested by the above analysis, and re-estimating α and β for each group, gave the regression models shown in Figures 5.16 to 5.19. The fitted lines had the following equations:

The hourly regression model (Figure 5.16):

Region A ($R^2 = 76\%$): $V1 = 0.03159 + 0.008597\ V24$ (5.8a)

Region B ($R^2 = 83\%$): $V1 = 0.03734 + 0.006205\ V24$ (5.8b)

The 3 hourly regression model (Figure 5.17):

Region A ($R^2 = 88\%$): $V3 = 0.1415 + 0.04931\ V24$ (5.9a)

Region B ($R^2 = 90\%$): $V3 = 0.1055 + 0.04524\ V24$ (5.9b)

The 6 hourly regression model (Figure 5.18):

Region A ($R^2 = 93\%$): $V6 = 0.2899 + 0.1426\ V24$ (5.10a)

Region B ($R^2 = 94\%$): $V6 = 0.01200 + 0.1493\ V24$ (5.10b)

The 12 hourly regression model (Figure 5.19):

Region A ($R^2 = 96\%$): $V12 = 0.4655 + 0.3874\ V24$ (5.11a)

Region B ($R^2 = 97\%$): $V12 = -0.1582 + 0.4174\ V24$ (5.11b)

To assess the validity of the above regression models a closer look at the residuals was needed.

Regressing the Hourly Variance on the Daily Variance

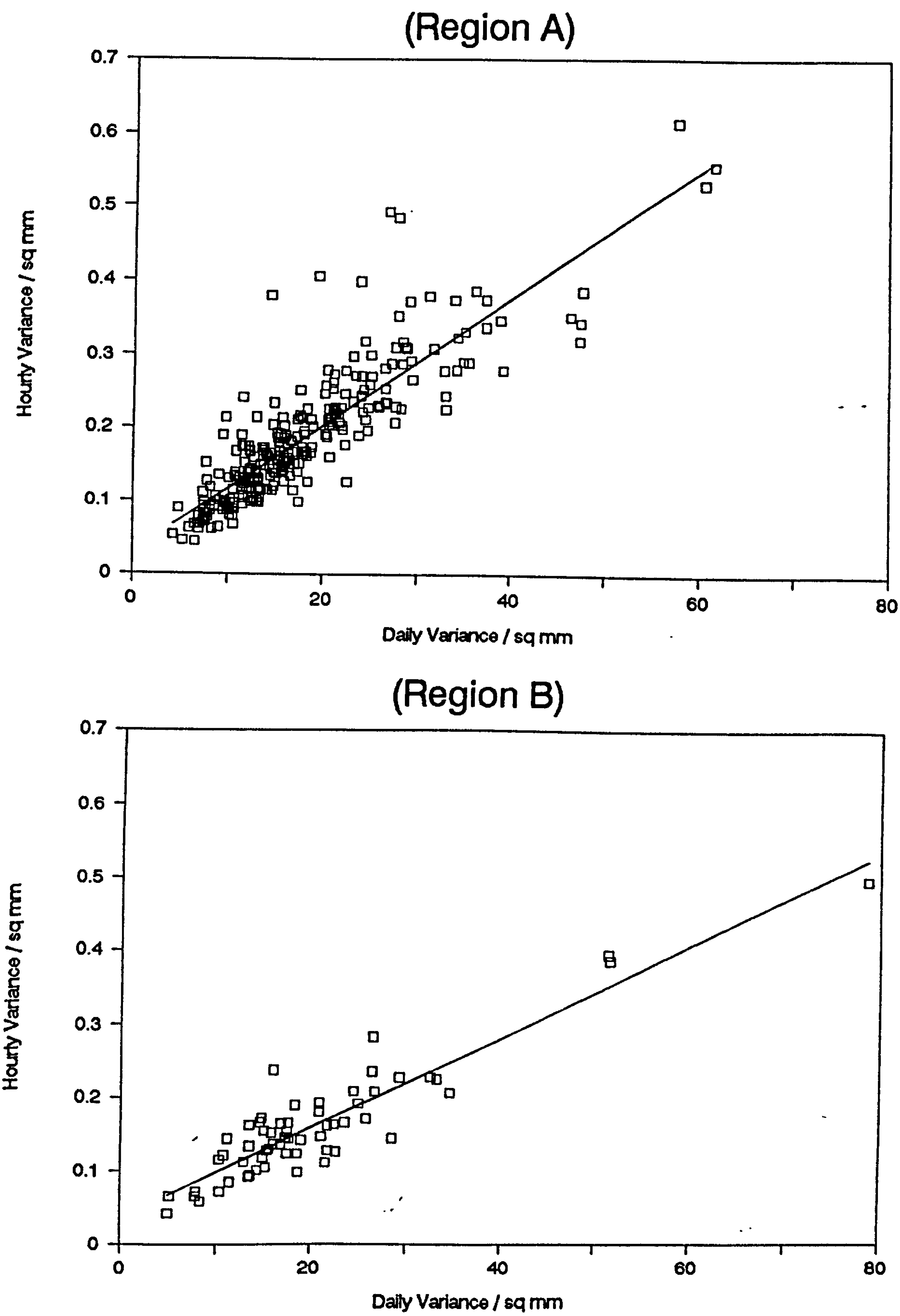
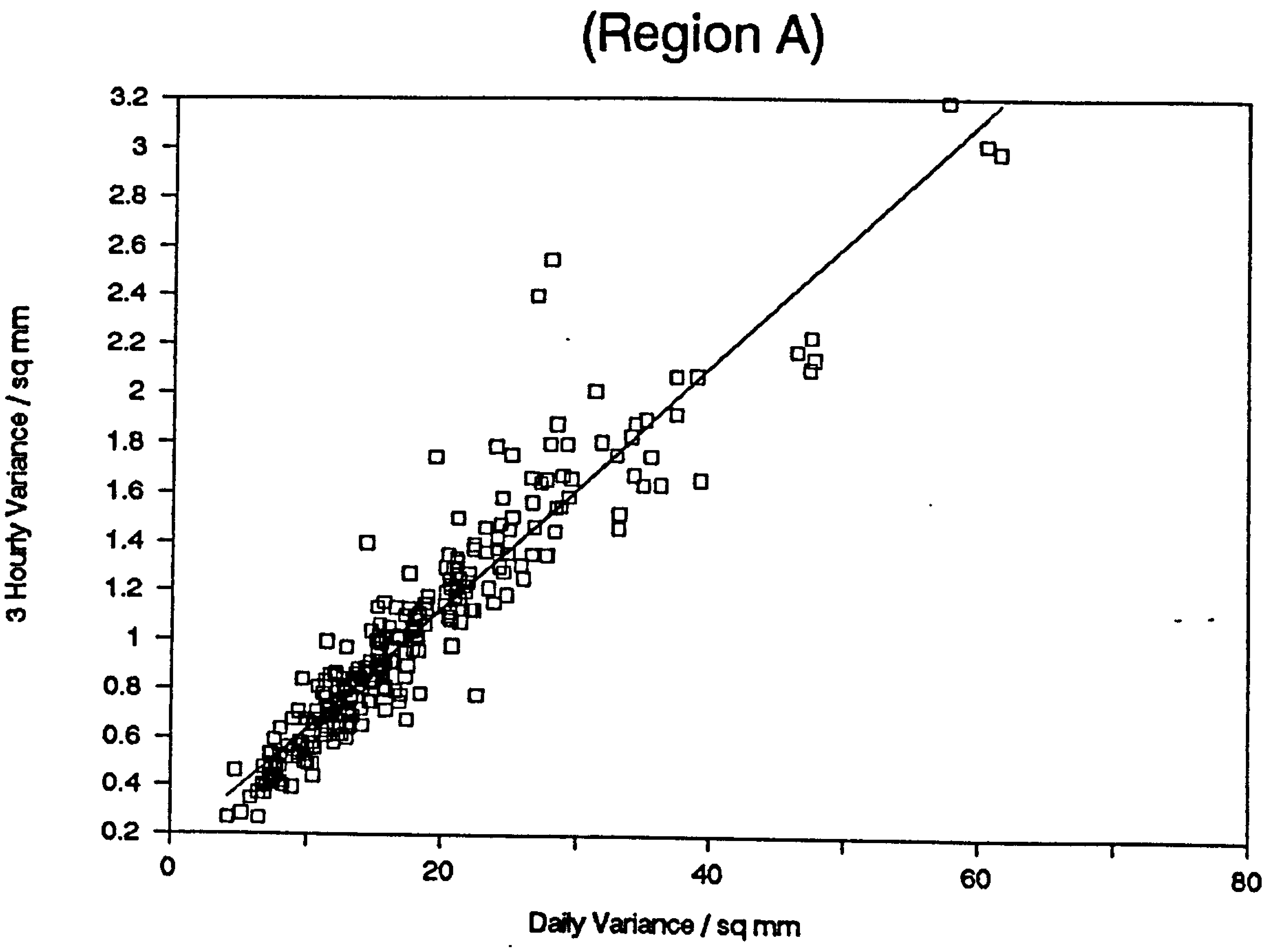


Figure 5.16
Regression plots for the hourly variance
for regions A and B

Regressing the 3 Hourly Variance on the Daily Variance



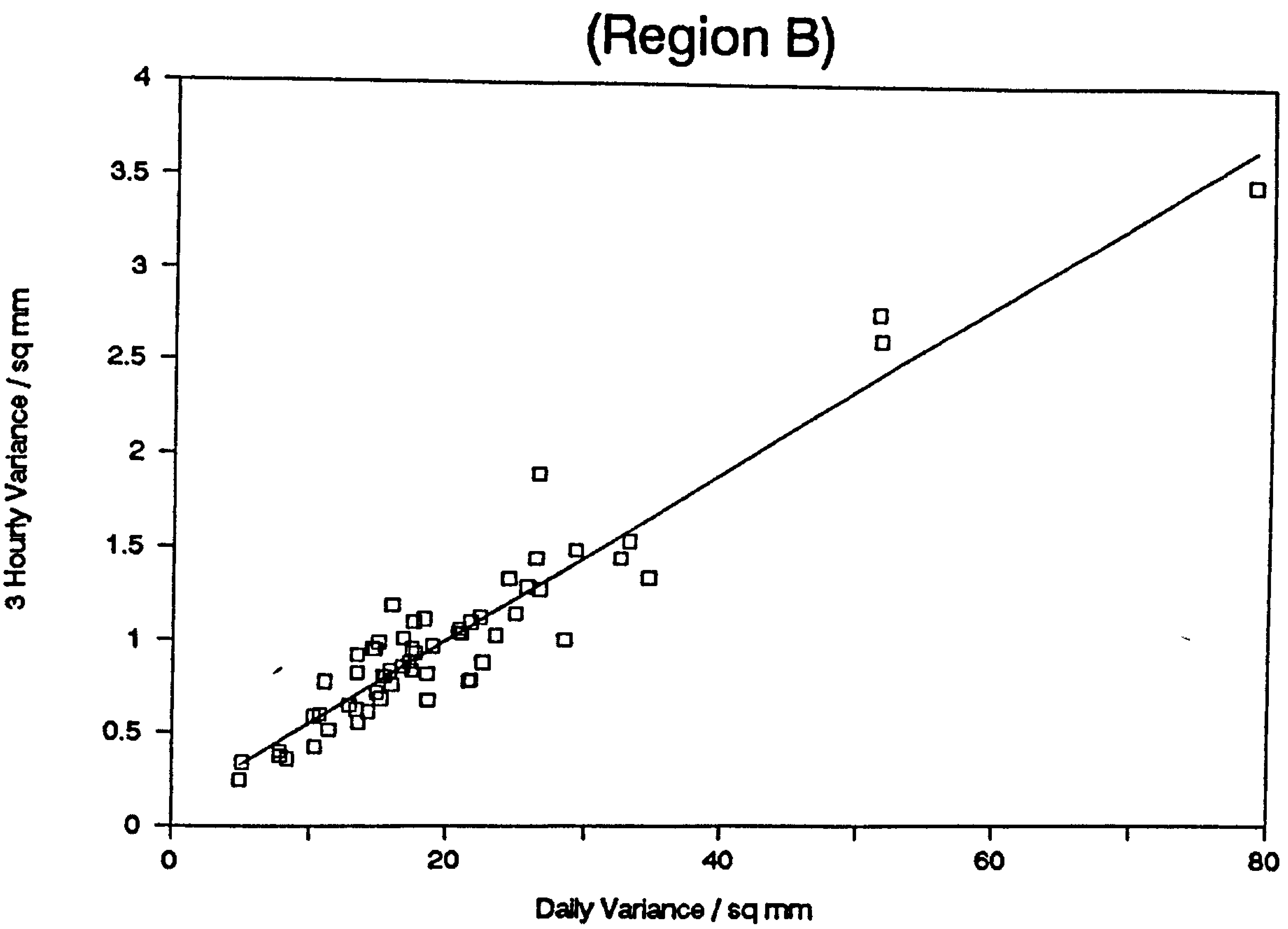
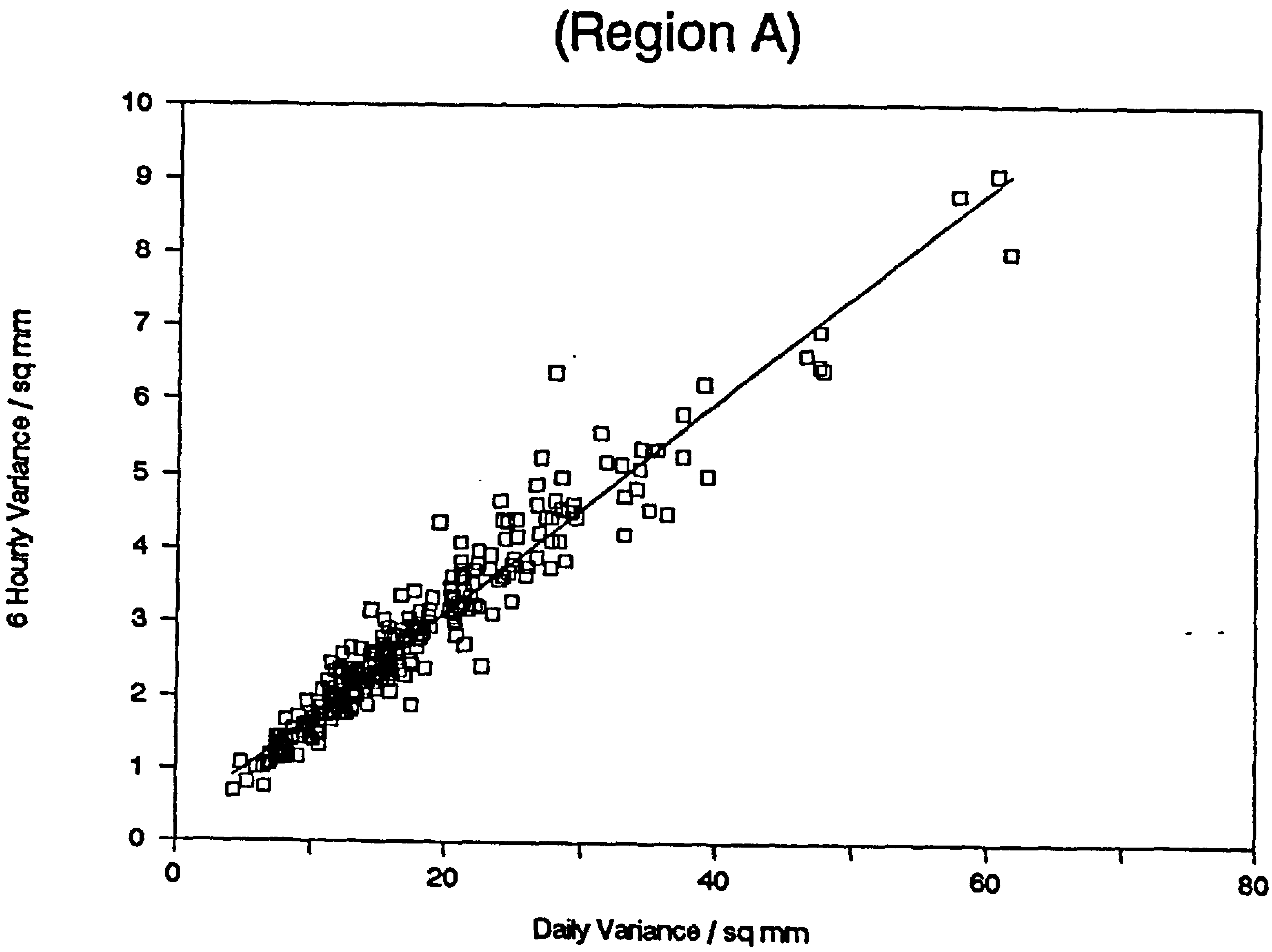


Figure 5.17
Regression plots for the 3 hourly variances
for regions A and B

Regressing the 6 Hourly Variance on the Daily Variance



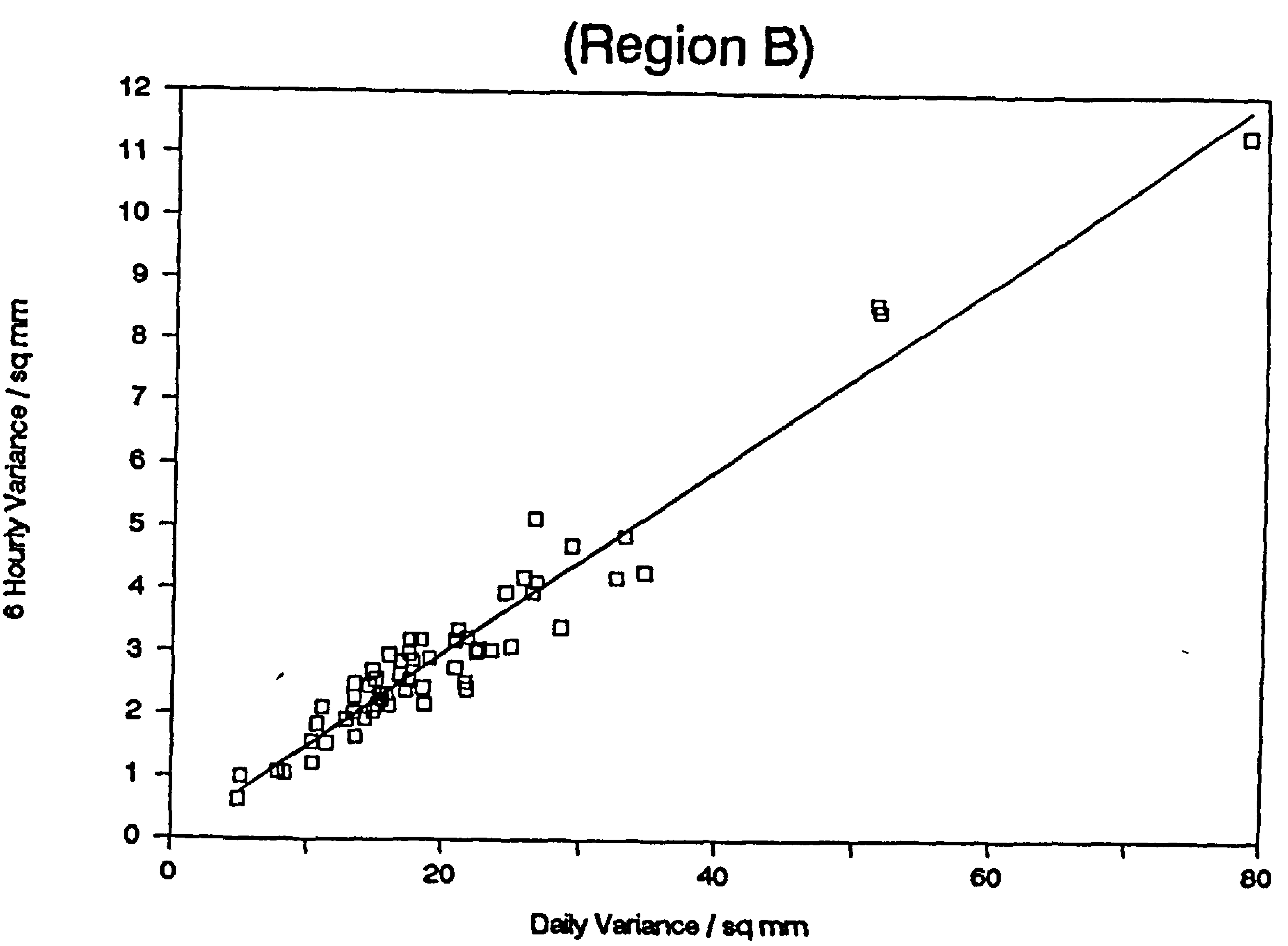
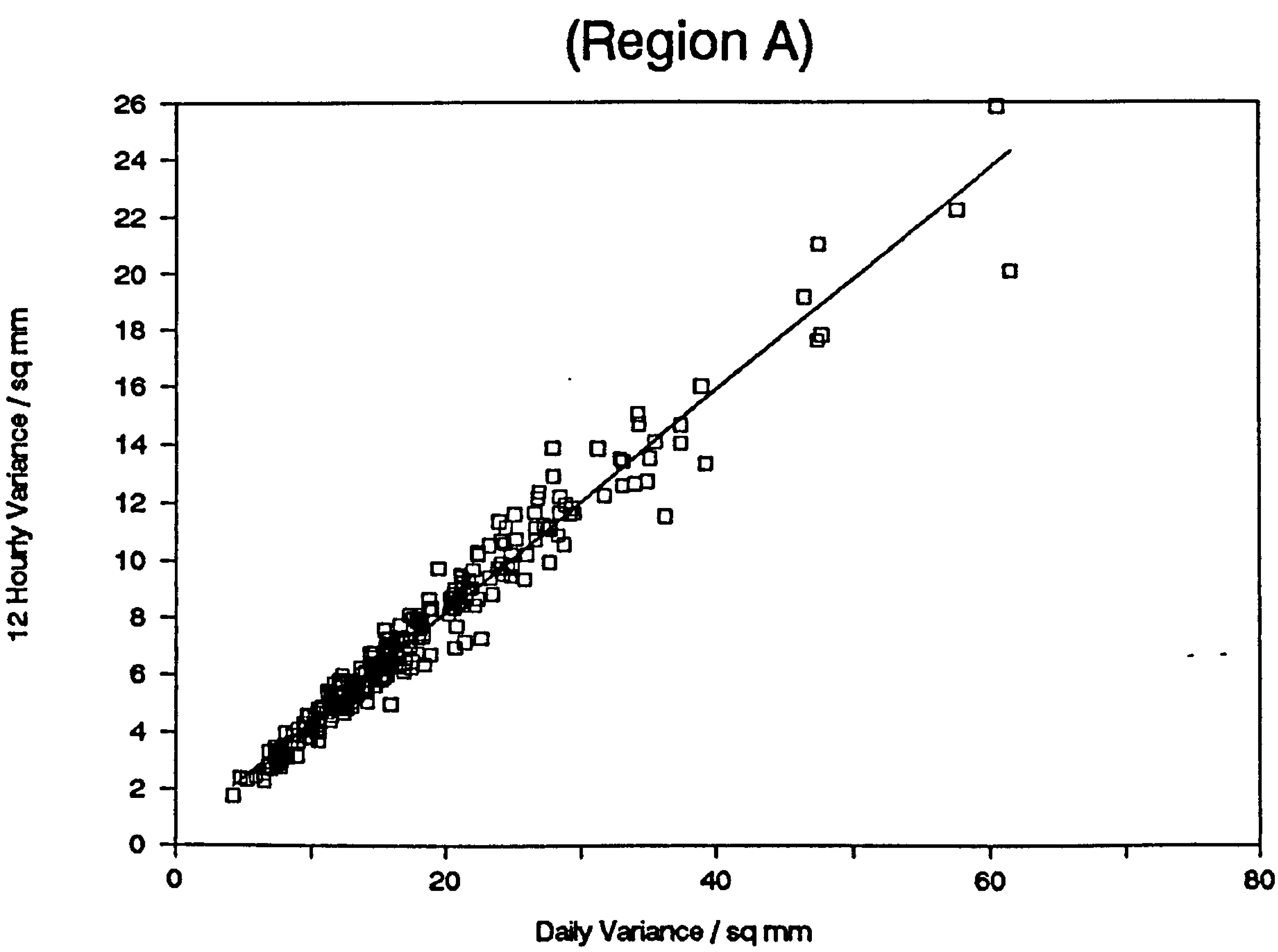


Figure 5.18
Regression plots for the 6 hourly variances
for regions A and B

Regressing the 12 Hourly Variances on the Daily Variances



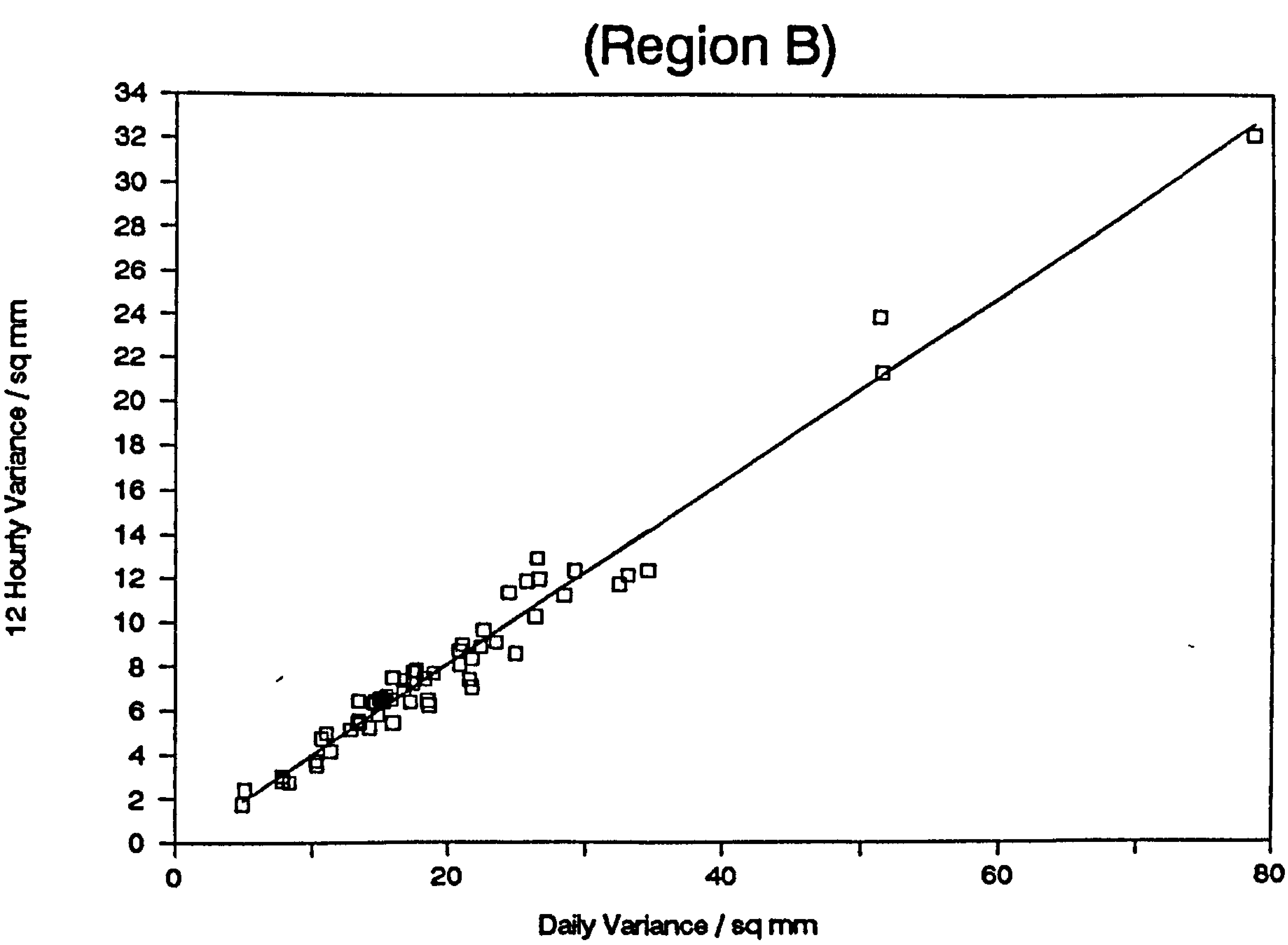


Figure 5.19
Regression plots for the 12 hourly variances
for regions A and B

5.7 AN ANALYSIS OF THE RESIDUALS

Plots of the residuals of the 1 hourly variances against the months and predicted variances are given in Figures 5.20 and 5.21. In Figure 5.21 it can be seen that there is no obvious dependency of the residuals on the predicted values. This implies that the results of the statistical tests are unlikely to be misleading. However, it was evident from Figure 5.20 that the residuals had some dependency on the month, which suggested that it might be worth including some harmonic as an explanatory variable. It was thought that the inclusion of a sine/cosine wave of 1 cycle per year (the first harmonic) in the regression model would probably remove the seasonal variation in the residuals. To include the first harmonic in the regression model, assume that the h hourly variances Vh are given by:

$$Vh = \alpha_0 + \alpha_1 V24 + \alpha_2 \cos(2\pi t/12) + \alpha_3 \sin(2\pi t/12) + \varepsilon$$

where $t = 1, 2, \ldots, 12$ ($1 \equiv$ Jan, $2 \equiv$ Feb, etc), $h = 1, 3, 6, 12$.

The parameters α_i of the above regression models were estimated by the method of least squares for both regions. Figures 5.22 and 5.23 show the 'improved' residual plots for the 1 hourly variances, and Table 5.3 shows the increase in R^2 for all h hourly variances. Standard t-tests were performed to test the hypotheses that the coefficients of the harmonic terms were zero. These hypotheses were all rejected at the 5% level for at least one of the harmonic terms in each regression model. However, as the increase in R^2 is only very slight (see Table 5.3), the inclusion of the harmonic terms is unlikely to be of much practical benefit.

Hence, the first regression models which do not include harmonic terms are preferred because of their simplicity. For users requiring the slight improvement in accuracy offered by the more complex harmonic regression models, the parameter estimates $\hat{\alpha}_i$ for the models are provided in Table 5.4.

Table 5.3

Improvement in R^2 obtained by introducing the first harmonic

Region	Dependent variable	R^2 excluding harmonic	R^2 including harmonic
A	V1	0.76	0.81
A	V3	0.88	0.90
A	V6	0.93	0.94
A	V12	0.96	0.96
B	V1	0.83	0.91
B	V3	0.90	0.93
B	V6	0.94	0.95
B	V12	0.97	0.97

Table 5.4

Parameter Estimates for the regression models with the first harmonic included

Region	Dependent Variable	Constant $\hat{\alpha}_0$	Coeff of Independent Variables V24 $\hat{\alpha}_1$	$\cos(\pi t/6)$ $\hat{\alpha}_2$	$\sin(\pi t/6)$ $\hat{\alpha}_3$
A	V1	0.047200	0.007731	-0.02413	-0.02258
A	V3	0.197625	0.046193	-0.06445	-0.08043
A	V6	0.420248	0.135363	-0.09569	-0.18519
A	V12	0.653956	0.376970	-0.15207	-0.26817
B	V1	0.044442	0.005852	-0.01705	-0.02604
B	V3	0.136823	0.043685	-0.07216	-0.11308
B	V6	0.072124	0.146325	-0.12098	-0.20635
B	V12	-0.15761	0.417416	-0.33580	-0.20346

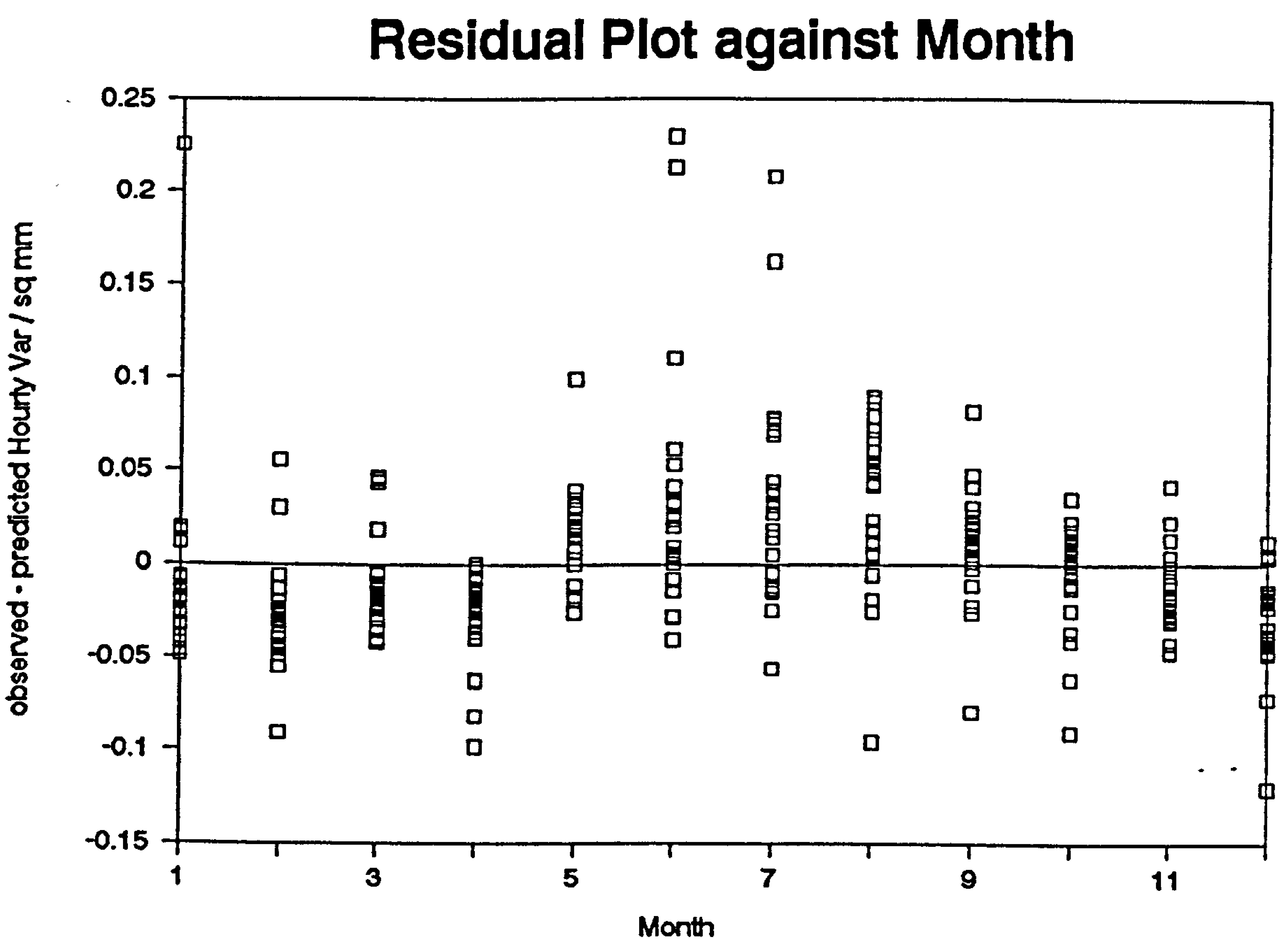


Figure 5.20

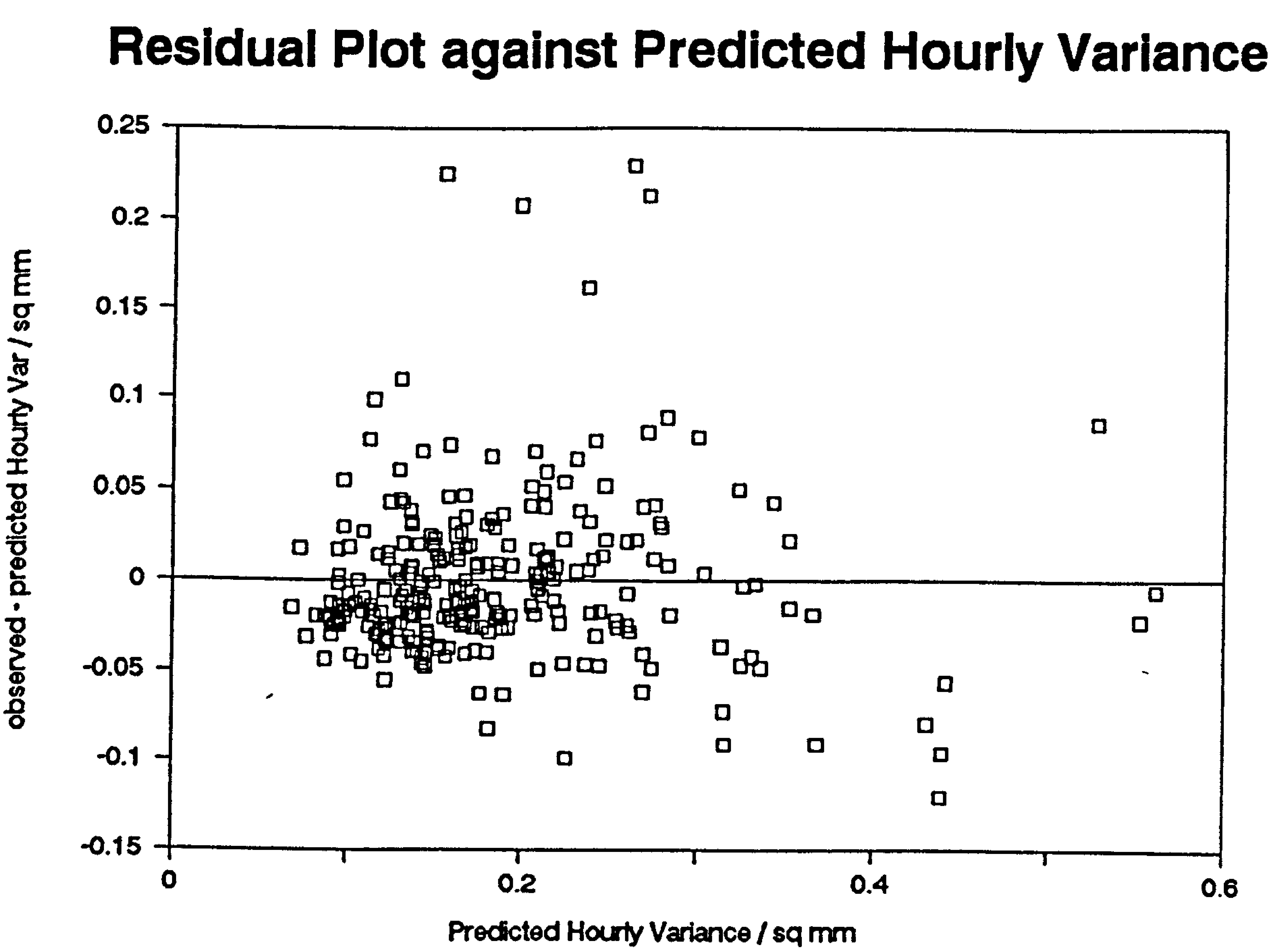


Figure 5.21

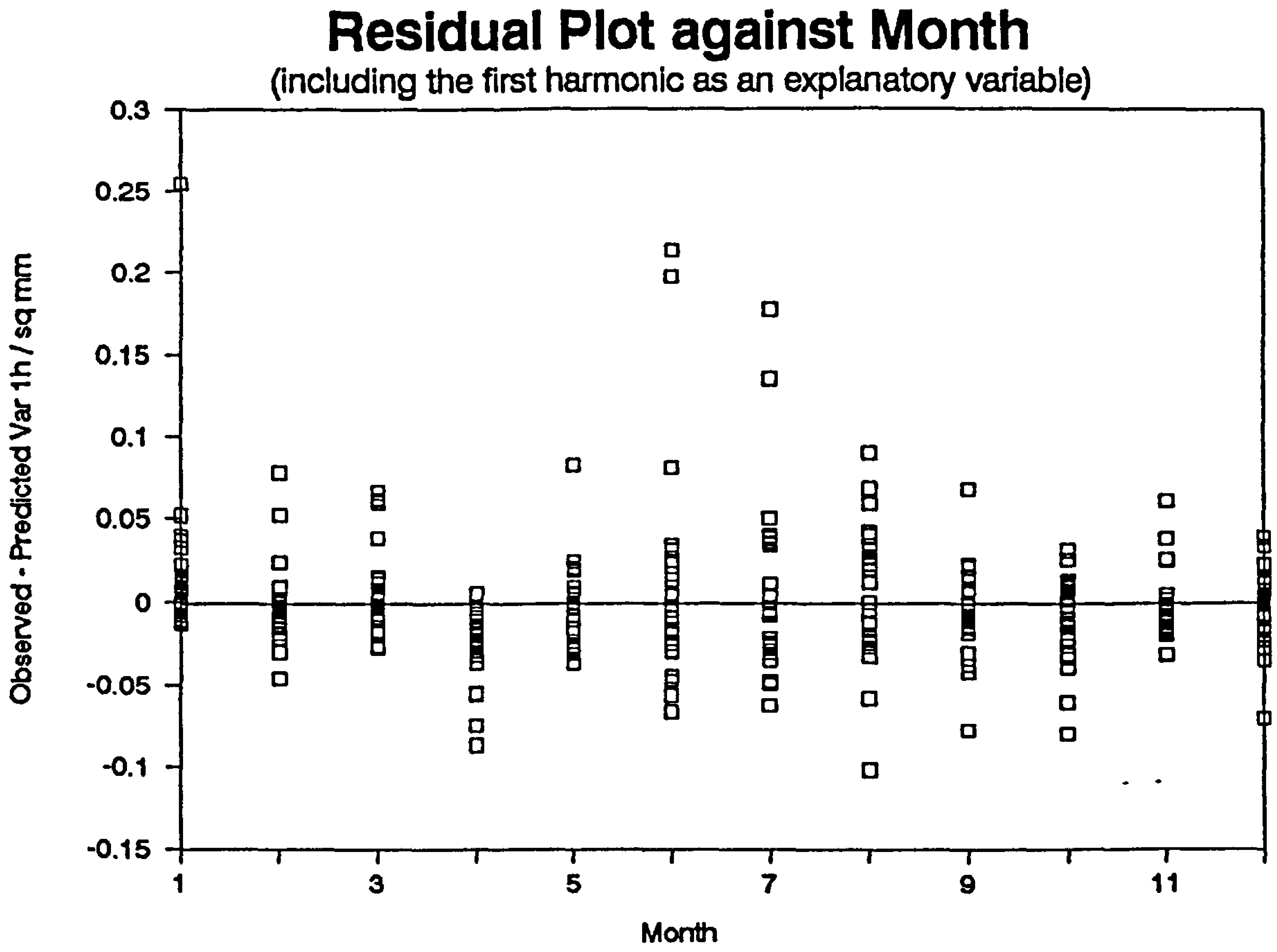


Figure 5.22

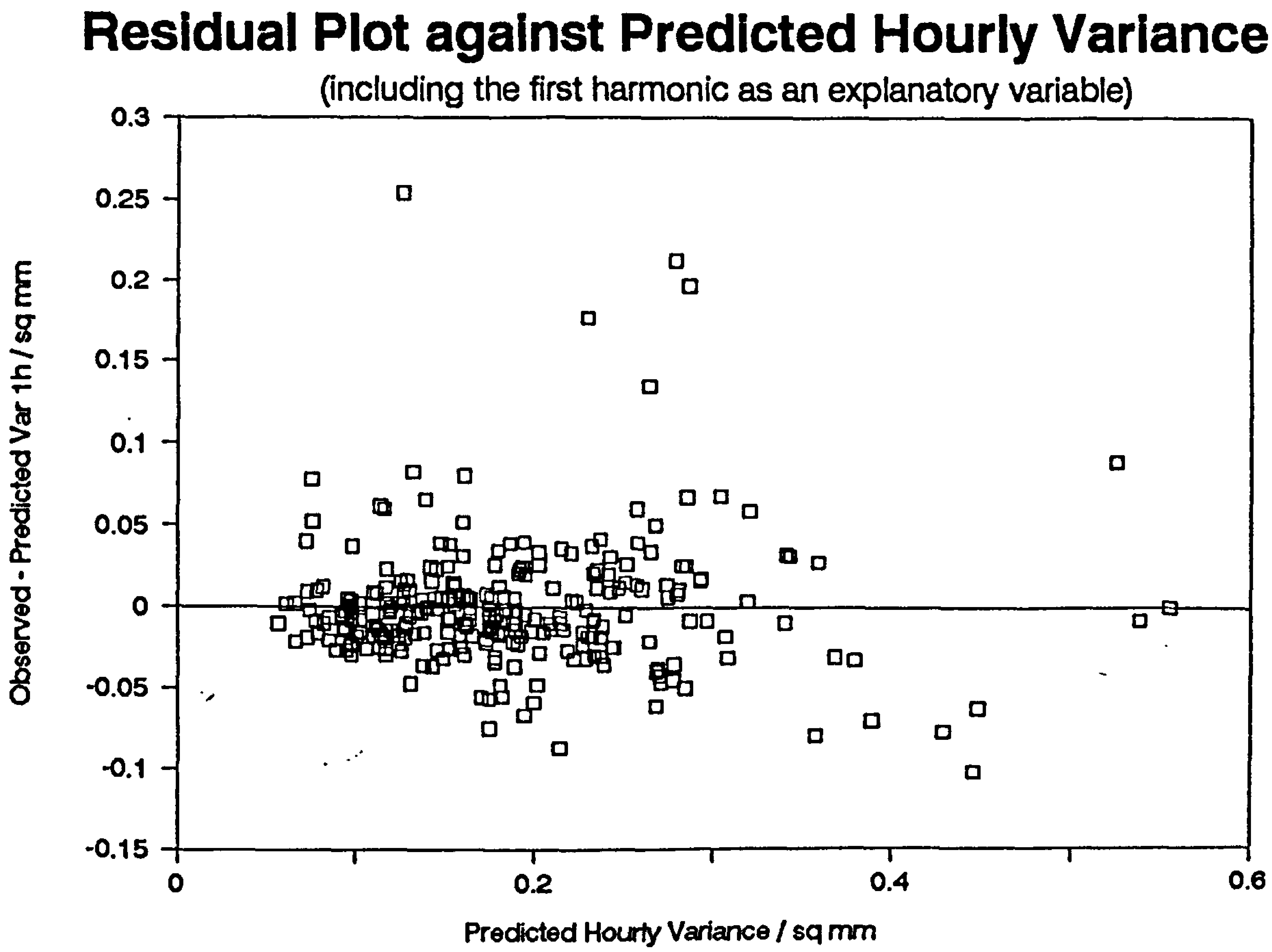


Figure 5.23

5.8 AN EXTREME VALUE ANALYSIS OF THE NEYMAN-SCOTT RAINFALL MODEL

5.8.1 Introduction

In the previous Sections, a method of fitting the Neyman-Scott model to daily rainfall data was given. In this Section the model is fitted to the longest daily records available in each of the 'Wigley' regions. Data are then simulated for January and July in each region, and maximum daily totals for historical and simulated data compared. The results of this Section show that the model has a tendency to under-estimate extreme values, mainly for return periods in excess of 10 years. However, this under-estimation may be compensated for by simulating for a longer period. A rough guide to the length of this period is given.

5.8.2 The Gumbel distribution

An analysis of extremes often involves fitting the Gumbel distribution to the data, or plotting the data on Gumbel probability paper. In this Section some properties of the distribution are reviewed. These properties can be found in the Flood Studies Report (1975).

The Gumbel Distribution has two parameters: i) a location parameter u, and ii) a shape parameter α.

If X is a Gumbel (α,u) variate, then X has pdf:

$$f_X(x) = \alpha^{-1} \exp\{-\alpha^{-1}(x-u) - e^{-\alpha^{-1}(x-u)}\} \qquad (5.12)$$

The distribution function $F_X(x)$ of X is given by:

$$F_X(x) = \exp\{-e^{\alpha^{-1}(x-u)}\} \quad (5.13)$$

The mean and variance of X are:

$$E(X) = u + \alpha\gamma \quad (5.14)$$

$$Var(X) = \pi^2\alpha^2/6 \quad (5.15)$$

where γ is Euler's constant (γ = 0.5772 to 4 d.p.).

The standardised variate Y is related to X by:

$$Y = (X - u)/\alpha \quad (5.16)$$

The variate Y has pdf:

$$f_Y(y) = \exp\{-y - e^{-y}\} \quad (5.17)$$

The distribution function of Y is:

$$F_Y(y) = \exp\{-e^{-y}\} \quad (5.18)$$

i.e. Y is a Gumbel (0,1) variate.

Let the return period for Y be T years. Then,

$$Y = -\ln(-\ln(1-T^{-1})) \quad (5.19)$$

(see Flood Studies Report (1975), Section 1.2.4)

$$\Rightarrow X = u - \alpha \ln(-\ln(1-T^{-1})) \qquad (5.20)$$

Let there be N years of daily rainfall data. Let $\{x_i\}$ be the ordered sequence of maximum daily rainfall for each year, so $x_1 \le x_2 \le \ldots \le x_N$. The x_i may be plotted against the standardised variates y_i, using:

$$y_i = -\ln(-\ln F_i), \qquad (5.21)$$

where $F_i = (i - 0.44)/(N + 0.12)$ (due to Gringorten (1963)).

If the x_i follow a Gumbel distribution then the (x_i, y_i) should lie on a straight line (from equation (5.16)). Deviations from a straight line indicate that some other distribution may be more appropriate. However, if only the last few points deviate from a straight line, then there is unlikely to be sufficient evidence against assuming a Gumbel distribution because of high sampling variability as i approaches N.

5.8.3 An analysis of maximum daily rainfalls

The longest record of data from each 'Wigley' region were selected for the analysis (see Table 5.5). For each of these stations the parameters of the Neyman-Scott model were estimated for January and July using the parameter estimation procedure of Chapter 4 with estimates of the 1, 3, 6, and 12 hourly variances given by equations (5.8) - (5.11) (see Table 5.6 for the parameter estimates). Data were simulated for each of these stations (the number of years of simulated data was equal to the number of years of historical data). The maximum daily rainfall for each year was found for both historical and simulated time series. These maxima were ordered and then plotted against the standardised Gumbel variate (see Figures 5.24 - 5.28).

Figures 5.24 - 5.28 show that the model tends to under-estimate the very extreme values. However, when re-designing an existing sewage system, the day to day performance of the system is more important than the performance of the system under extreme events (*Henderson* (1986)), and so the model is not regarded as inadequate for its intended purpose. If an engineer is interested in the performance of a sewage system under very extreme events (e.g. storms with return periods exceeding 10 years), then there are three choices available.

Firstly, the engineer can use the traditional design storm approach, where a storm profile is found from the historical data for a given return period (e.g. see *Arnell et al* (1984)). This approach does have its drawbacks (as pointed out in Chapter 1), but may still be useful in modelling extreme rainfall events.

Secondly, the model's fit to the historical maximum daily rainfalls may improve if an expression for the mean and variance of the maxima were included in the parameter estimation procedure (equation (4.7)) when fitting the model. However, the mean and variance of the maximum amount of rain captured in a day are not available as functions of the model parameters. An attempt to find these functions has proved too difficult mathematically (some workings towards these are given in Appendix A). An alternative approach is to use regression techniques to attempt to find the functions empirically. This approach is addressed in Section 5.8.5.

Thirdly, the engineer could simulate data for a longer period using the stochastic model. If this choice is made, the engineer will need some guide as to the number of years of simulated data that corresponds to the historical data, i.e. we need to determine how many years of simulated data are needed in order to obtain the same proportion of heavy storms, on average, as an historical record. This is approach is considered in the next Section.

Table 5.5

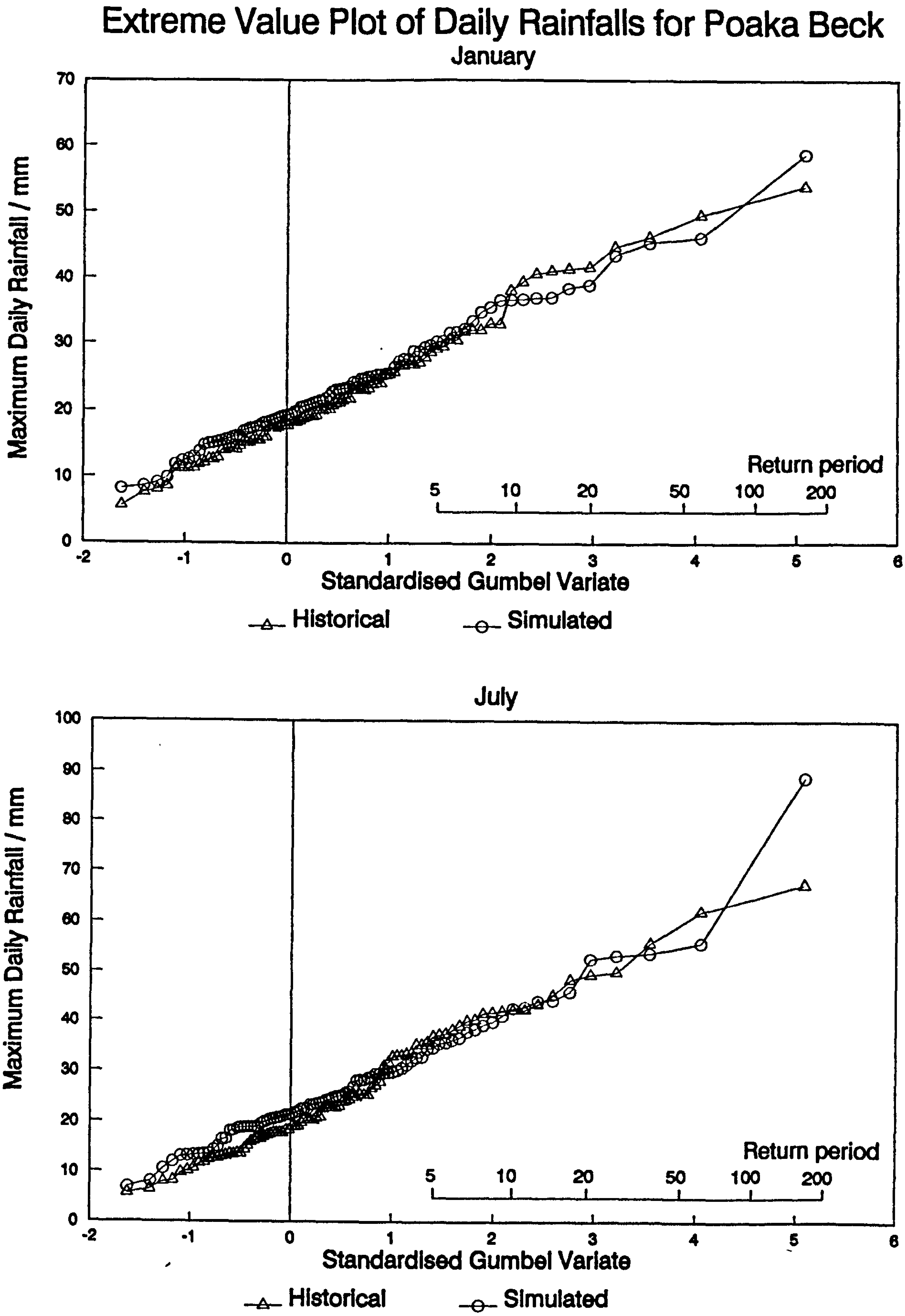
Long Daily Records used in Analysis

Station Number	Station Name	Years of Data	Altitude m	East Grid	North Grid	Region
1525	Howick Hall	92	34	4246	6177	NE
115306	Blackbrook	90	107	4456	3178	C
588702	Poaka Beck	90	156	3240	4781	NW
275574	Windsor	90	21	4979	1754	S
354864	Exmouth	74	66	3027	819	W

Table 5.6

Parameter Estimates for January and July

Station	Month	λ	β	η	ν	ξ
1525	January	0.0185	0.1275	0.828	4.60	1.294
	July	0.0163	0.1007	0.512	2.83	0.986
115306	January	0.0173	0.1796	1.432	6.78	0.976
	July	0.0127	0.1299	0.853	3.57	0.640
588702	January	0.0200	0.1023	1.181	8.16	0.840
	July	0.0181	0.1167	0.890	4.92	0.670
275574	January	0.0183	0.1325	1.215	4.90	0.974
	July	0.0115	0.1149	0.973	4.32	0.692
354864	January	0.0171	0.0854	0.958	5.22	0.843
	July	0.0091	0.0599	0.745	3.22	0.553



(a)

(b)

Figure 5.24
Extreme value plots for Poaka Beck (North-West region)

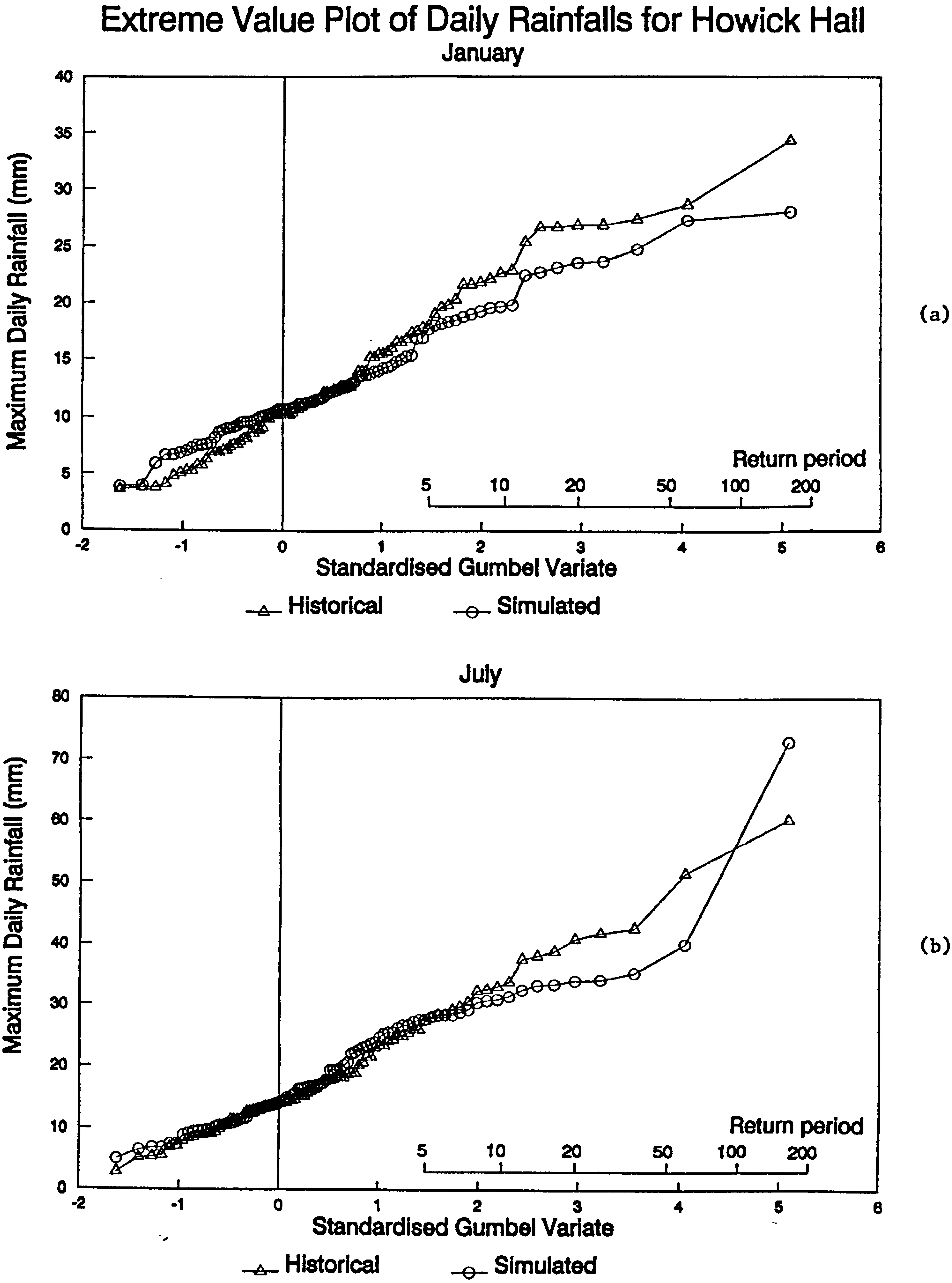
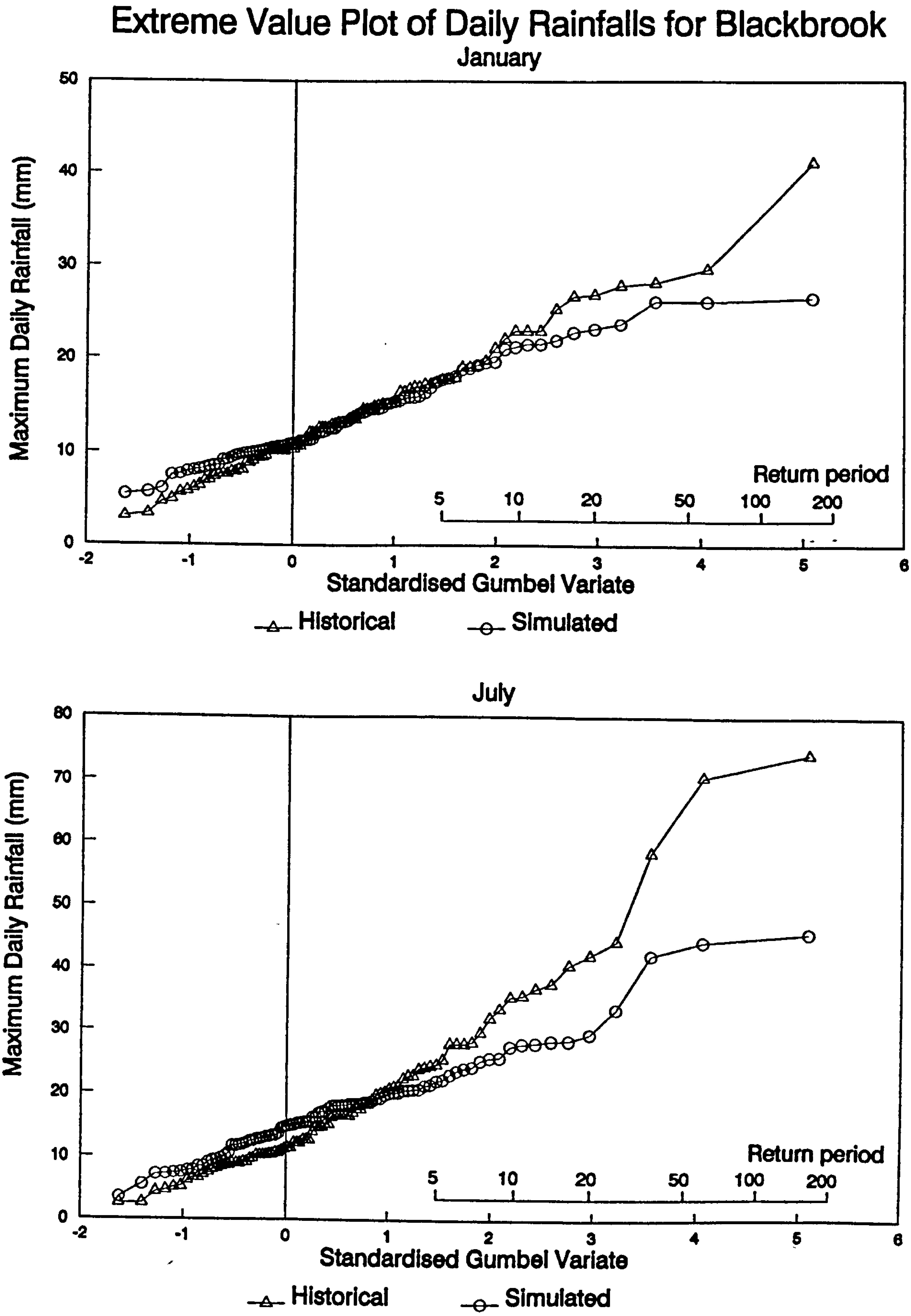


Figure 5.25
Extreme value plot for Howick Hall (North-East region)



(a)

(b)

Figure 5.26
Extreme value plot for Blackbrook (Central region)

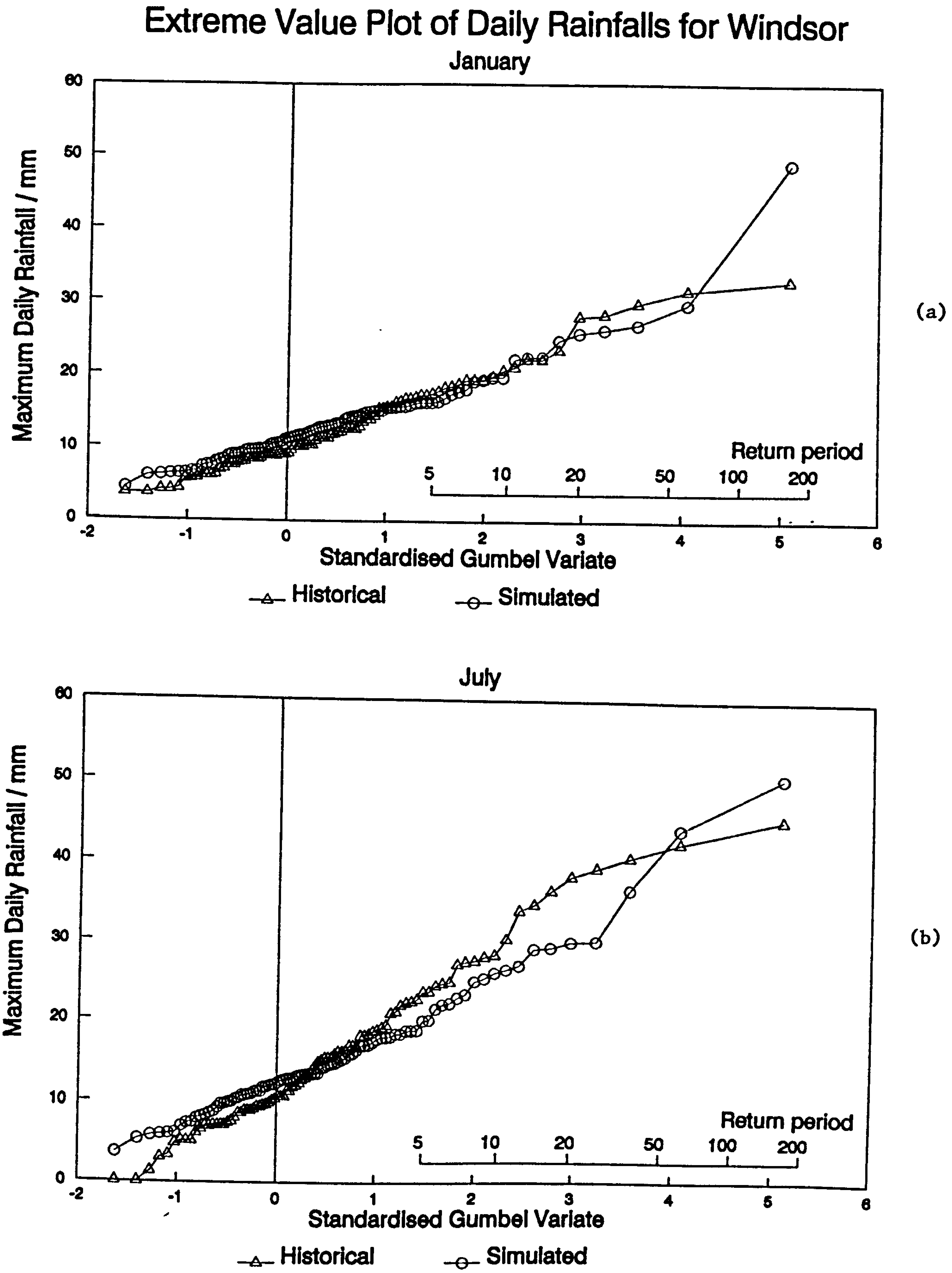
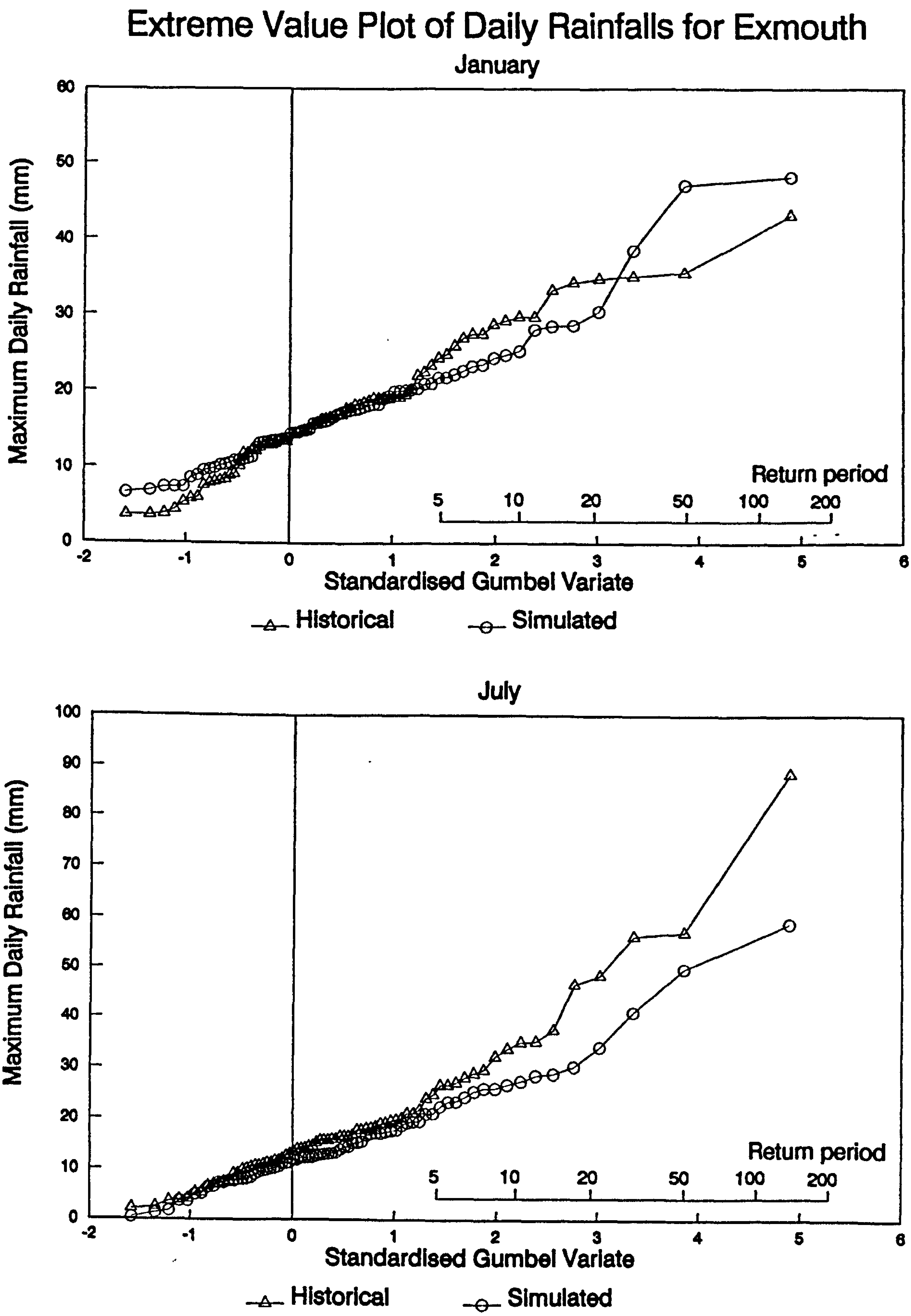


Figure 5.27
Extreme value plots for Windsor (Southern region)



(a)

(b)

Figure 5.28
Extreme value plots for Exmouth (South-West region)

5.8.4 Using the model to simulate for a longer period to capture the extreme rainfall events

Figures 5.24 - 5.28 show that the plotted points for both the historical and simulated data lie approximately on a straight line, allowing for the large sampling variability of the end points. It therefore seems reasonable to assume that the data come from Gumbel distributions.

The parameters (α, u) of the assumed Gumbel distributions were estimated by the Method of Moments by equating the observed mean and variances of the daily maxima with their equivalent in the population (equations (5.14) and (5.15) respectively). The Gumbel parameter estimates for both the historical and simulated maxima are given in Table 5.7 (denoted α_H, u_H and α_S, u_S respectively).

Table 5.7

Gumbel Parameter Estimates

Station	Month	Gumbel Parameter Estimate			
		u_H	α_H	u_S	α_S
Poaka Beck	January	18.0	7.65	19.6	7.08
Poaka Beck	July	19.8	10.0	21.6	9.61
Windsor	January	10.0	4.95	11.0	4.89
Windsor	July	12.1	7.65	12.2	6.33
Blackbrook	January	10.9	5.09	11.5	3.78
Blackbrook	July	12.5	10.3	14.1	6.11
Howick	January	10.3	5.16	7.8	3.95
Howick	July	14.6	8.32	15.1	10.76
Exmouth	January	13.7	6.66	14.0	6.15
Exmouth	July	12.7	11.0	11.3	8.00

Suppose there are N_H years of historical rainfall data. Of interest is the number of years N_S of simulated rainfall data, where the probability of storms with high return periods being found in N_S is the same as the probability of storms with high return periods being found in the historical data (where N_S is likely to be greater than N_H because the model has a tendency to under-estimate the extremes), i.e. we need to find N_S such that:

$$\text{pr}\{\text{at least one storm of return period} > T \text{ years in } N_S\} = \text{pr}\{\text{at least one storm of return period} > T \text{ years in } N_H\},$$

for a range of (high) T values and typical N_H values.

Values of N_S for different values of N_H and T would give the engineer a rough guide as to how many years of simulated rainfall data are equivalent to historical data of record lengths N_H. Let the historical maximum corresponding to storms of return period T be x_H, so that

$$x_H = u_H - \alpha_H \ln(-\ln(1-T^{-1})), \tag{5.22}$$

from equation (5.20), assuming Gumbel distributions for the maximum daily rainfalls. Hence, the probability p_H of obtaining a storm with return period exceeding T years in one year of historical data is given by:

$$p_H = \text{pr}\{X_H > x_H\} = 1 - F_{X_H}(x_H) = 1 - \exp\{-\exp\{\alpha_H^{-1}(x_H - u_H)\}\},$$

using equations (5.13) and (5.22). Also, the probability p_S of obtaining a storm of historical return period exceeding T years in one year of simulated rainfall data is given by:

$$p_S = pr\{X_S > x_H\} = 1 - F_{X_S}(x_H) = 1 - \exp\{-\exp\{\alpha_S^{-1}(x_H - u_S)\}\},$$

Thus the probabilities of there being at least one storm of return period in excess of T years in the historical and simulated data (of record lengths N_H and N_S respectively) are:

$1 - (1-p_H)^{N_H}$ and $1 - (1-p_S)^{N_S}$ respectively (assuming the data are independent from one year to the next).

Therefore, the required record length for the simulated data can be obtained by solving:

$$1 - (1-p_H)^{N_H} = 1 - (1-p_S)^{N_S}$$

$$\Rightarrow N_S = N_H \ln(1-p_H)/\ln(1-p_S) \qquad (5.23)$$

The values N_S were found for historical storms of return periods 5, 10, and 20 years, and historical record lengths of 10, 20, and 30 years (see Table 5.8). From Table 5.8 it was evident that, on average, the record length of the simulated data should be about 1.4 times the historical record length to obtain the right proportion storms with return periods in excess of 5 years, and about 1.8 and 2.3 times the historical record length to obtain the right proportion of storms with return periods in excess of 10 and 20 years respectively. However, the standard deviations of the N_S values are large, so these multiples (1.4, 1.8, and 2.3 times the historical length) can only be taken as rough guides.

Table 5.8

Simulated record lengths N_S equivalent to historical record lengths N_H

		Historcal record lengths N_H								
		5	10	20	5	10	20	5	10	20
		Return periods:								
Station	Mth	T > 20 yrs			T > 10 yrs			T > 5 years		
Poaka	Jan	5.1	10.2	20.3	4.8	9.6	19.2	4.5	9.1	18.1
Poaka	Jul	4.7	9.5	18.9	4.6	9.2	18.3	4.4	8.9	17.7
Windsor	Jan	4.2	8.5	16.9	4.2	8.4	16.8	4.2	8.3	16.7
Windsor	Jul	9.3	18.6	37.1	7.8	15.6	31.2	6.5	13.0	25.9
Blackbrk	Jan	11.8	23.5	47.1	9.2	18.4	36.7	7.1	14.2	28.3
Blackbrk	Jul	29.4	58.7	117.4	17.9	35.8	71.6	10.7	21.4	42.7
Howick	Jan	23.2	46.5	93.0	18.7	37.3	74.6	14.8	29.7	59.3
Howick	Jul	2.4	4.9	9.8	2.9	5.8	11.5	3.4	6.8	13.7
Exmouth	Jan	6.1	12.2	24.5	5.8	11.5	23.0	5.4	10.8	21.7
Exmouth	Jul	18.2	36.5	72.9	13.9	27.9	55.7	10.5	21.0	42.1
Mean		11.4	22.9	45.8	9.0	17.9	35.9	7.2	14.3	28.6
SD		9.2	18.4	36.7	5.8	11.7	23.3	3.7	7.4	14.8

5.8.5 An attempt to improve the model's fit to the historical maximum daily rainfalls using a regression model

The results of this Section summarise the work of *Quinn* (1991), for which the author of this thesis was the supervisor.

The model's fit to the historical maximum daily rainfalls may improve if an expression for the mean and variance of the maxima were included in the parameter estimation procedure (equation (4.7)) when fitting the model. However, the mean and variance of the maximum amount of rain captured in a day are not available as functions of the model parameters. An attempt to find these functions has proved too difficult mathematically (some workings are given in Appendix A). An alternative approach is to use regression techniques to attempt to find the functions empirically.

Rainfall data were simulated (for 90 years) using the simulation program (Appendix B) for 88 different combinations of model parameters (these combinations were chosen to give a good spread of the mean and variances of the maxima - see *Quinn* (1991) for details). For each of the 88 simulations the maximum daily rainfalls were found for each year, and the mean and variance of these maxima evaluated. The correlation matrix for the model parameters and the mean (μ) and variance (σ^2) of the maxima was found and is given in Table 5.9.

Table 5.9

Correlation matrix for model parameters

and mean and variance of the maximum daily rainfalls

	μ	σ^2	λ	β	η	ν	$1/\xi$
μ	1.00	0.77	0.25	-0.08	-0.37	-0.03	0.66
σ^2	0.77	1.00	0.04	-0.05	-0.54	-0.35	0.82

From Table 5.9 it is evident that λ and β show the least correlations with the other variables, and so may not be needed when predicting μ and σ^2 (this will be checked more formally later in this Section). In contrast, of the model parameters, $1/\xi$ (= mean cell intensity) shows the highest correlation with μ and σ^2 and so will probably be needed when predicting μ and σ^2. Table 5.9 also shows that μ and σ^2 are correlated to each other (a correlation of 0.77). For this reason the problem is best set up as a multivariate regression problem, where the dependent variables μ and σ^2 are treated as the coefficients of a dependent vector (as oppose to performing two separate multiple regressions on each of μ and σ^2), i.e. we will assume that:

$$\begin{pmatrix} \mu_i \\ \sigma_i^2 \end{pmatrix} = \begin{pmatrix} a_0 & a_1 & a_2 & \cdots & a_n \\ b_0 & b_1 & b_2 & \cdots & b_n \end{pmatrix} \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \\ \cdot \\ \cdot \\ \cdot \\ x_{ni} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix}$$

i.e. $\mu_i = a_0 + a_1x_{1i} + a_2x_{2i} + \ldots + a_nx_{ni} + \varepsilon_{1i}$ (5.24a)

and $\sigma_i^2 = b_0 + b_1x_{1i} + b_2x_{2i} + \ldots + b_nx_{ni} + \varepsilon_{2i}$ (5.24b)

where μ_i and σ_i^2 are the mean and variance of the maxima for the ith simulation (i = 1, ..., 88), ε_{1i} and ε_{2i} are the residuals due to the sampling error in estimating μ_i and σ_i^2 respectively, and x_{ji} is a product of the model parameters for the ith simulation. To reduce the number of combinations possible for the x_{ji}, only products of order 2 will be considered, i.e.

$$x_{ji} \in \{\lambda, \beta, \eta, \nu, \xi, \lambda\beta, \lambda\eta, \lambda\nu, \lambda\xi, \beta\eta, \beta\nu, \beta\xi, \eta\nu, \eta\xi, \nu\xi, \lambda^2, \beta^2, \eta^2, \nu^2, \xi^2\},$$

for j = 1, ..., n, and $x_{ji} \neq x_{ki}$ for all $j \neq k$. The significance of an explanatory variable x_{ji} will be tested using Wilk's lambda (Λ) (details of which can be found *Krzanowski* (1990)). The number of terms n appearing in equations (5.24a and 5.24b) will thus be the number of explanatory variables found to be significant using this test statistic.

The regression models given in equations (5.24a and 5.24b) were fitted by the Method of Least Squares. However, it was found that the variance of the residuals tended to increase as the predicted values increased (see *Quinn* (1991)). Hence, the natural logarithm of the mean and variance of the maxima was used in equations (5.24a and 5.24b), i.e. we now assume that:

$$\ln(\mu_i) = a_0 + a_1x_{1i} + a_2x_{2i} + \ldots + a_nx_{ni} + \varepsilon_{1i} \qquad (5.25a)$$

and $$\ln(\sigma_i^2) = b_0 + b_1x_{1i} + b_2x_{2i} + \ldots + b_nx_{ni} + \varepsilon_{2i} \qquad (5.25b)$$

(N.B. log transformations of hydrological data are frequently used in regression problems when the variance of the residuals fails to be constant (e.g. see *Holder,* 1985))

Although it was anticipated that products of the explanatory variables would be required in the final form of the regression equation (5.25), the first regression model fitted for the transformed data did not include any such products, i.e. the first model fitted only included the model parameters (λ, β, η, ν, ξ) as explanatory variables so that n = 5 in equation (5.25). The purpose behind this was to see if any of the parameters (and hence products of the parameters) could be left out of the regression model. Wilk's Λ was used to test the significance of each of the five explanatory variables (see Table 5.10).

From Table 5.10 it is evident that each of the variables β, η, ν, and ξ are going to be useful in predicting μ and σ^2. However, the F-ratio for the explanatory variable λ is not significant at the 15% level, and so λ could be left out of the regression model. Leaving λ out of the regression model reduces the number of explanatory variables that need to be considered, i.e. we need only consider the following terms: β, η, ν, ξ, $\beta\eta$, $\beta\nu$, $\beta\xi$, $\eta\nu$, $\eta\xi$,

$\nu\xi$, β^2, η^2, ν^2 and ξ^2, so that

$$x_{ji} \in \{\beta, \eta, \nu, \xi, \beta\eta, \beta\nu, \beta\xi, \eta\nu, \eta\xi, \nu\xi, \beta^2, \eta^2, \nu^2, \xi^2\}.$$

The parameters of the regression models (5.25a, 5.25b) were re-estimated (by the Method of Least Squares) where now all terms and products (except those involving λ) were included. The results of performing multivariate tests on these explanatory variables are given in Table 5.11.

From Table 5.11 it can be seen that the smallest F-ratio (which was 1.32 for the explanatory variable ν^2) fails to be significant at the 15% level, and so ν^2 will be left out of the regression model.

The parameters of the regression model were again re-estimated, where now the regression model excludes the variable ν^2. The significance of each explanatory variable was again tested using Wilk's Λ and the results of these tests are given in Table 5.12. From this Table it is evident that the smallest F-ratio (which was 1.79 for $\beta\nu$) is not significant at the 15% level, so that $\beta\nu$ will be left out of the regression model.

The parameters of the regression model were again re-estimated, where now the regression model excludes both the variables ν^2 and $\beta\nu$. The significance of each explanatory variable was again tested using Wilk's Λ and the results of these tests are given in Table 5.13. From this Table it is evident that the smallest F-ratio (which was 2.24 for β) is significant at the 15% level, so that all terms will remain in the regression model. The parameters estimates for the regression model, equations 5.25a and 5.25b, are given in Tables 5.14a and 5.14b respectively.

Table 5.10

Multivariate tests for the explanatory variables used in the first regression model

Explanatory variable	Wilk's Λ	F-ratio	Pr > F
λ	0.957	1.80	0.17
β	0.908	4.11	0.02
η	0.384	65.09	0.00
ν	0.431	53.52	0.00
$1/\xi$	0.190	172.31	0.00

Table 5.11

Multivariate tests on all explanatory variables and products up to order 2

Explanatory variable	Wilk's Λ	F-ratio	Pr > F
β	0.953	1.78	0.175
η	0.722	13.86	0.000
ν	0.815	8.19	0.001
$1/\xi$	0.814	8.23	0.001
$\beta\eta$	0.908	3.66	0.031
$\beta\nu$	0.926	2.88	0.063
β/ξ	0.804	8.75	0.000
$\eta\nu$	0.912	3.48	0.036
η/ξ	0.886	4.62	0.013
ν/ξ	0.935	2.51	0.089
β^2	0.601	23.88	0.000
η^2	0.824	7.70	0.001
ν^2	0.965	1.32	0.273
$1/\xi^2$	0.877	5.05	0.009

Table 5.12

Multivariate tests on explanatory variables where ν^2 has been left out of the model

Explanatory variable	Wilk's Λ	F-ratio	Pr > F
β	0.952	1.84	0.166
η	0.703	15.41	0.000
ν	0.808	8.66	0.000
$1/\xi$	0.813	8.39	0.001
$\beta\eta$	0.919	3.22	0.046
$\beta\nu$	0.953	1.79	0.174
β/ξ	0.810	8.55	0.001
$\eta\nu$	0.864	5.76	0.005
η/ξ	0.887	4.66	0.013
ν/ξ	0.937	2.45	0.094
β^2	0.604	23.91	0.000
η^2	0.817	8.18	0.001
$1/\xi^2$	0.893	4.38	0.016

Table 5.13

Multivariate tests on explanatory variables where ν^2 and $\beta\nu$ have been left out of the model

Explanatory variable	Wilk's Λ	F-ratio	Pr > F
β	0.943	2.24	0.114
η	0.676	17.77	0.000
ν	0.794	9.58	0.000
$1/\xi$	0.778	10.56	0.000
$\beta\eta$	0.933	2.65	0.077
β/ξ	0.803	9.09	0.000
$\eta\nu$	0.856	6.23	0.003
η/ξ	0.890	4.56	0.014
ν/ξ	0.939	2.39	0.099
β^2	0.576	27.29	0.000
η^2	0.823	7.97	0.001
$1/\xi^2$	0.865	5.80	0.005

Table 5.14a

Regression parameter estimates for dependent variable $\ln(\mu)$

Explanatory variable	Parameter estimate	Standard error
Constant	3.3140	0.3109
β	-4.9787	2.5371
η	-3.6510	0.6577
ν	0.5337	0.1239
$1/\xi$	-0.1334	0.2344
$\beta\eta$	3.5308	1.6475
β/ξ	4.1780	0.9783
$\eta\nu$	-0.2024	0.0585
η/ξ	0.7934	0.2618
ν/ξ	-0.0998	0.0454
β^2	-8.7195	1.2358
η^2	0.9952	0.2976
$1/\xi^2$	-0.0128	0.0498

R^2 = 92%, CV = 3.1%

Table 5.14b

Regression parameter estimates for dependent variable $\ln(\sigma^2)$

Explanatory variable	Parameter estimate	Standard error
Constant	2.8013	0.7324
β	3.9187	5.9762
η	-4.2610	1.5493
ν	0.3755	0.2918
$1/\xi$	2.4986	0.5522
$\beta\eta$	-2.7134	3.8808
β/ξ	1.8338	2.3045
$\eta\nu$	-0.1530	0.1379
η/ξ	0.0063	0.6167
ν/ξ	-0.0268	0.1069
β^2	-8.5777	2.9109
η^2	1.7607	0.7011
$1/\xi^2$	-0.4024	0.1174

R^2 = 93%, CV = 4.9%

Equations (5.25a) and (5.25b) were included in the parameter estimation procedure (4.7), where the parameters of the regression equations (5.25a) and (5.25b) are given in Tables 5.14a and 5.14b respectively. As an attempt is being made to simulate extreme rainfall events using the stochastic model, the squared terms containing the regression equations were both given a weight of 100 in the fitting procedure (4.7). Using this revised fitting procedure, the parameters of the Neyman-Scott Rectangular Pulses model were re-estimated for January and July for each of the long records of daily data (see Tables 5.15). In addition, the percentage errors between the historical and model statistics (which are functions of the model parameters) were found and are given in Table 5.16, from which it can be seen that the regression equations are matching the historical mean and variance of the maxima almost exactly (as expected). Furthermore, the model is also matching the other historical statistics used in the fitting procedure (so that the regression equations are unlikely to reduce the performance of the model in other ways (e.g. in the model's fit to the historical dry spell sequences)).

Data were then simulated for each of these stations using the parameter estimates in Table 5.15 (the number of years of simulated data was equal to the number of years of historical data). The maximum daily rainfalls for each year were found for both the historical and simulated time series. These maxima were ordered and then plotted against the standardised Gumbel variate (see Figures 5.29 - 5.33).

From Figure 5.29 - 5.33 it is evident that using the regression model (for the mean and variance of the maximum daily rainfalls) in the fitting procedure has improved the model's fit to the historical extreme values for most of the stations (e.g. compare Figures 5.31 with 5.26). However, for some station-months, where previously the model closely matched the historical maxima (e.g. Poaka Beck - July, Figure 5.24b), there is now more discrepancy between the historical and simulated maxima (compare Figure 5.24b with 5.29b). Therefore, it may be better to include the regression model in the fitting procedure only for the station-months where the model is failing to match the historical extremes.

Table 5.15

Parameter Estimates for January and July using the regression model in the fitting procedure

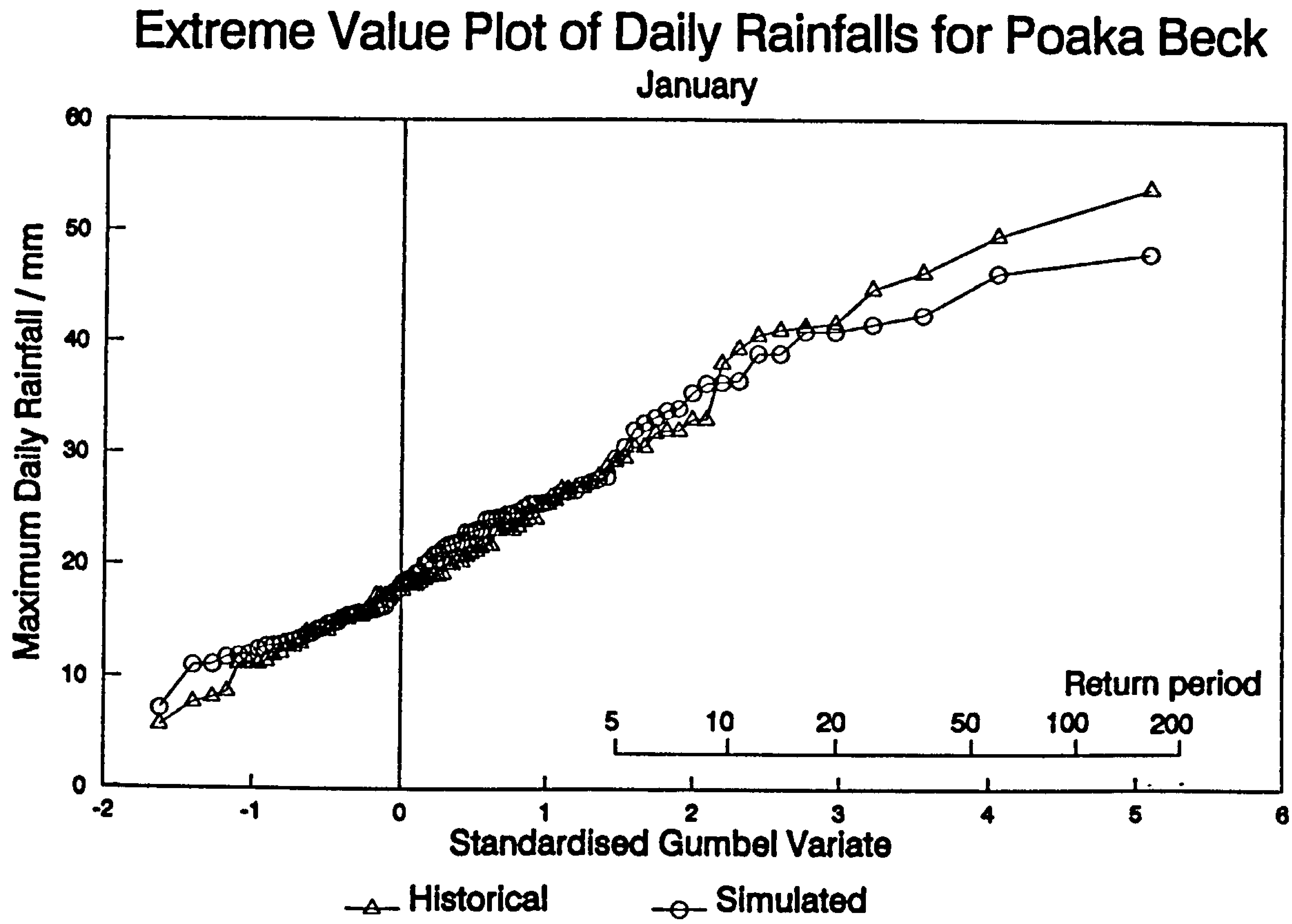
Station	Month	λ	β	η	ν	ξ
1525	January	0.0133	0.0615	1.1537	5.6377	0.8237
	July	0.0096	0.0500	0.8090	4.6269	0.6171
115306	January	0.0158	0.1434	1.7258	7.7702	0.8416
	July	0.0194	0.2498	0.6245	1.4267	0.5784
588702	January	0.0144	0.0602	1.3780	11.3792	0.7218
	July	0.0099	0.0500	1.1422	9.3752	0.5453
275574	January	0.0129	0.0762	1.3602	6.9279	0.8699
	July	0.0089	0.0500	0.8720	3.9516	0.5710
354864	January	0.0124	0.0521	1.1932	7.4178	0.6992
	July	0.0191	0.2764	0.5989	1.1916	0.5574

Table 5.16

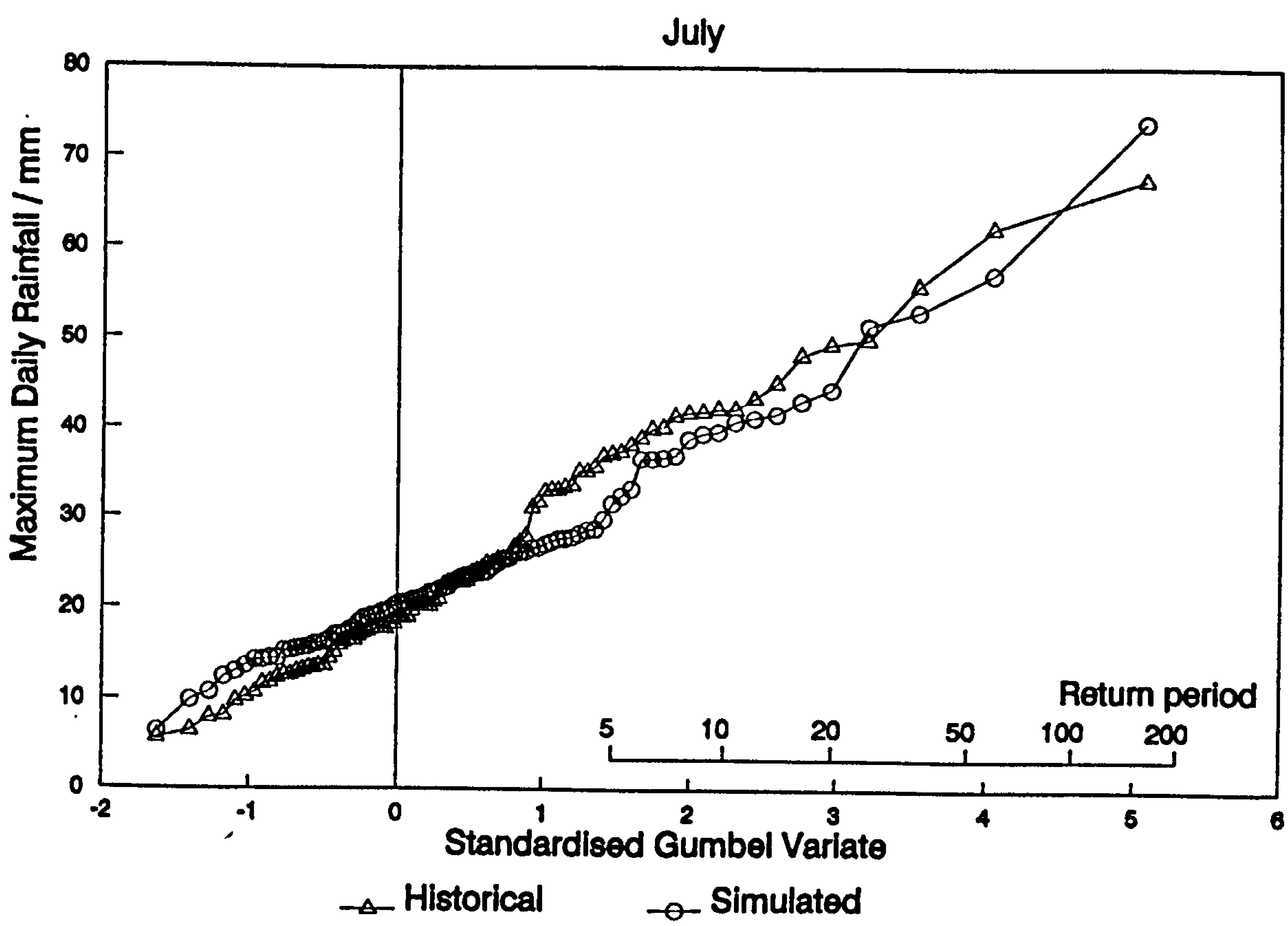
Percentage errors between model and historical statistics

		← Statistic →										
Station	mth	M1	V1	V3	V6	V12	V24	PD24	WW24	DD24	LNM	LNV
1525	Jan	-1	4	7	1	-3	0	1	11	11	0	-0
	Jul	-3	2	13	9	3	1	8	10	14	0	-1
115306	Jan	0	3	1	-3	-3	-1	1	6	4	0	-0
	Jul	-8	2	21	21	13	3	-4	-16	-13	-0	-1
588702	Jan	0	4	3	-3	-4	1	0	7	11	0	-0
	Jul	-0	5	7	0	-3	-1	5	14	19	0	-1
275574	Jan	-0	3	3	-2	-3	4	4	10	12	-0	-0
	Jul	-5	7	16	11	3	2	2	8	4	1	-1
354864	Jan	-0	6	7	-1	-5	-3	1	7	9	0	-0
	Jul	-4	-3	16	17	9	-1	-5	-26	-18	-0	-1

LNM, LNV = log of mean and variance of maximum daily rainfall respectively.

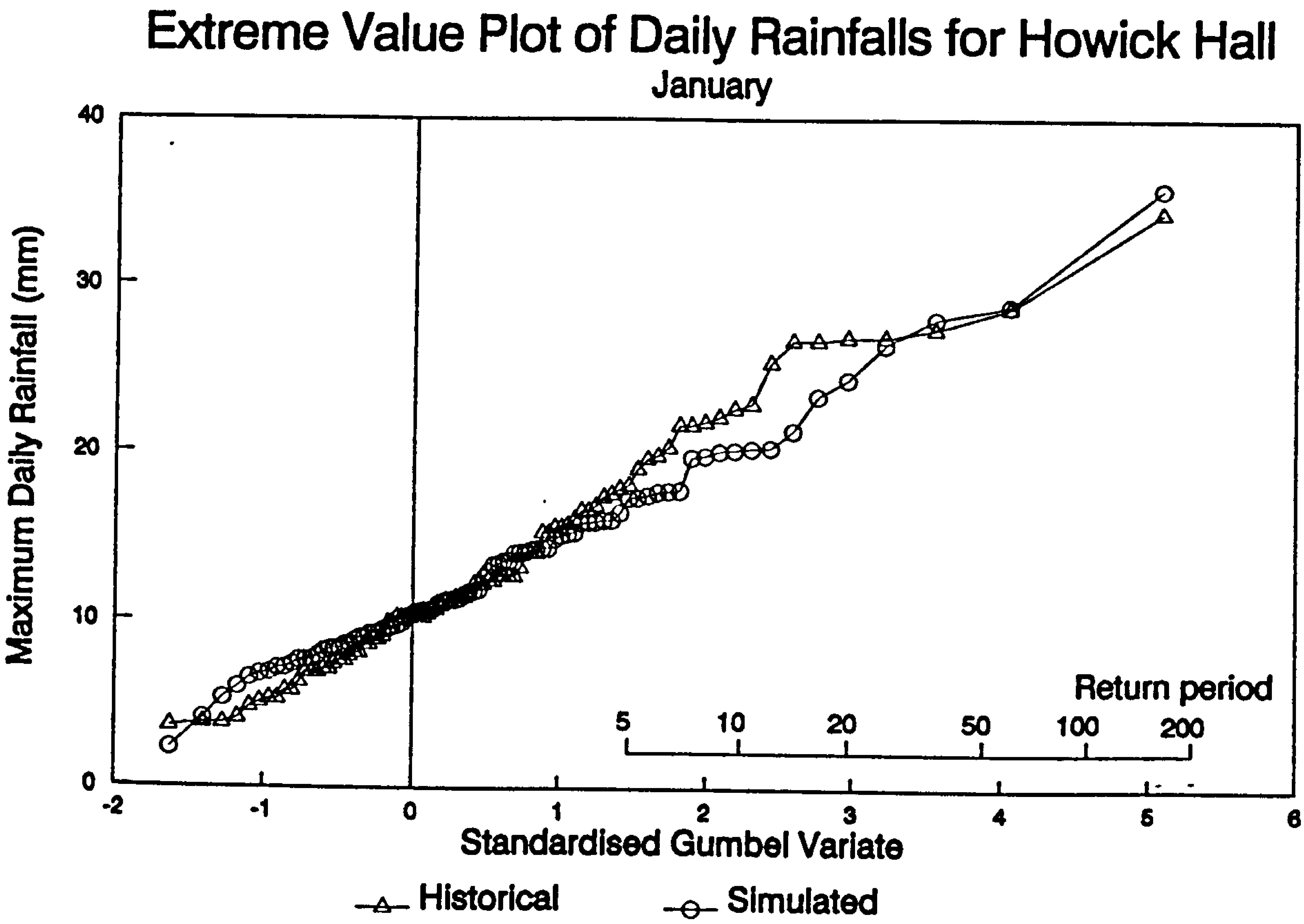


(a)

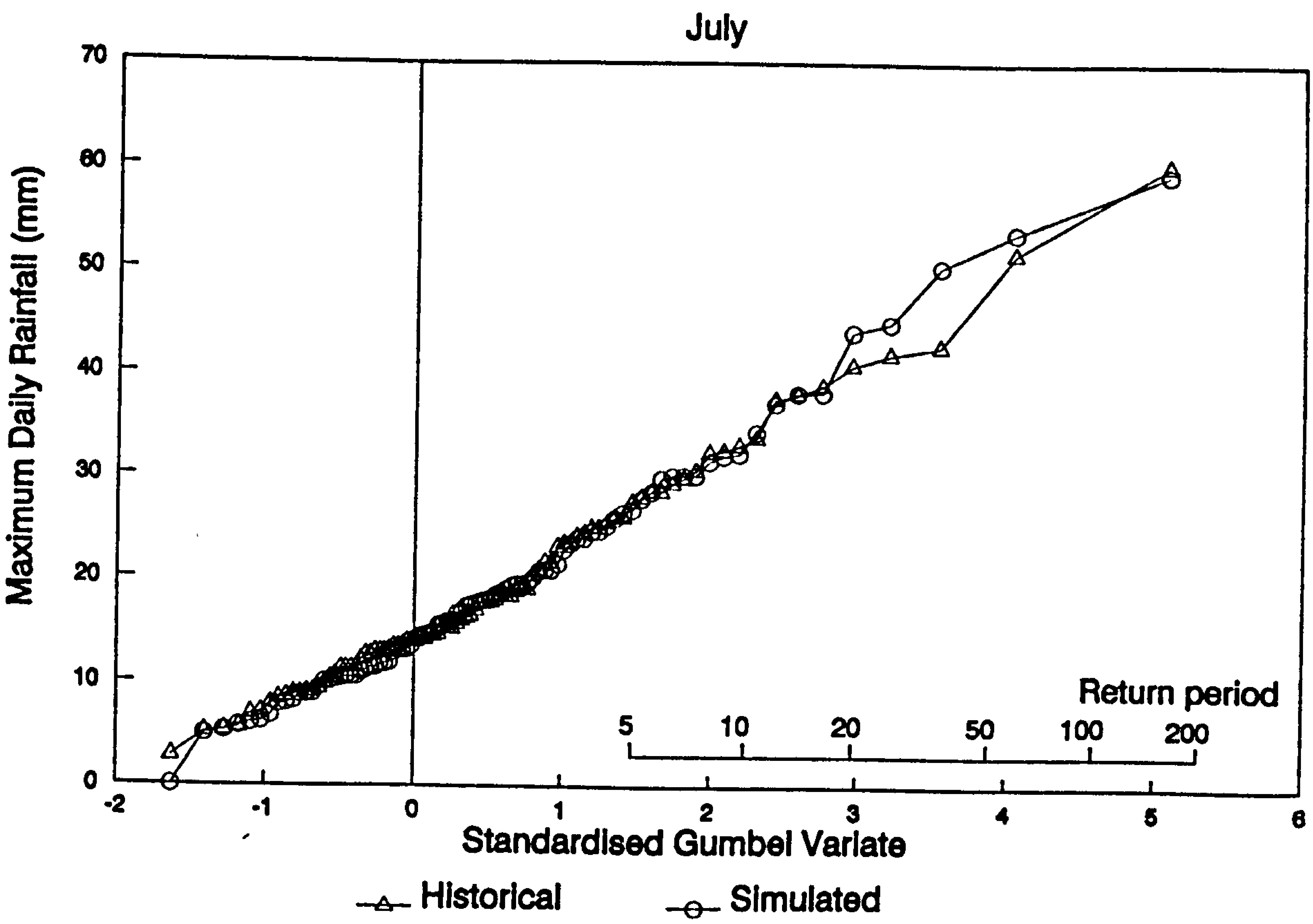


(b)

Figure 5.29
Extreme value plots for Poaka Beck
using the regression model in the fitting procedure

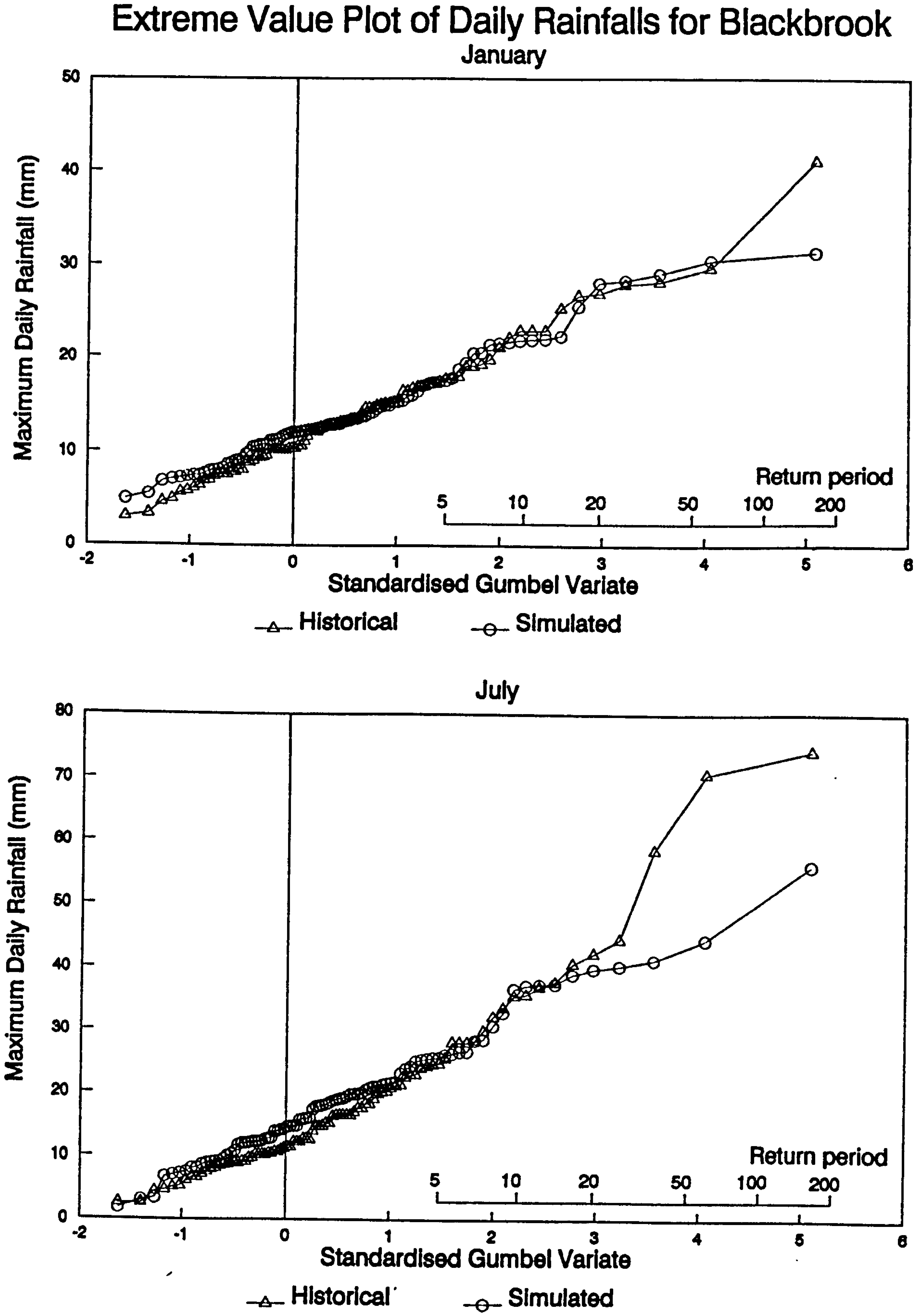


(a)



(b)

Figure 5.30
Extreme value plots for Howick Hall
using the regression model in the fitting procedure



(a)

(b)

Figure 5.31
Extreme value plots for Blackbrook
using the regression model in the fitting procedure

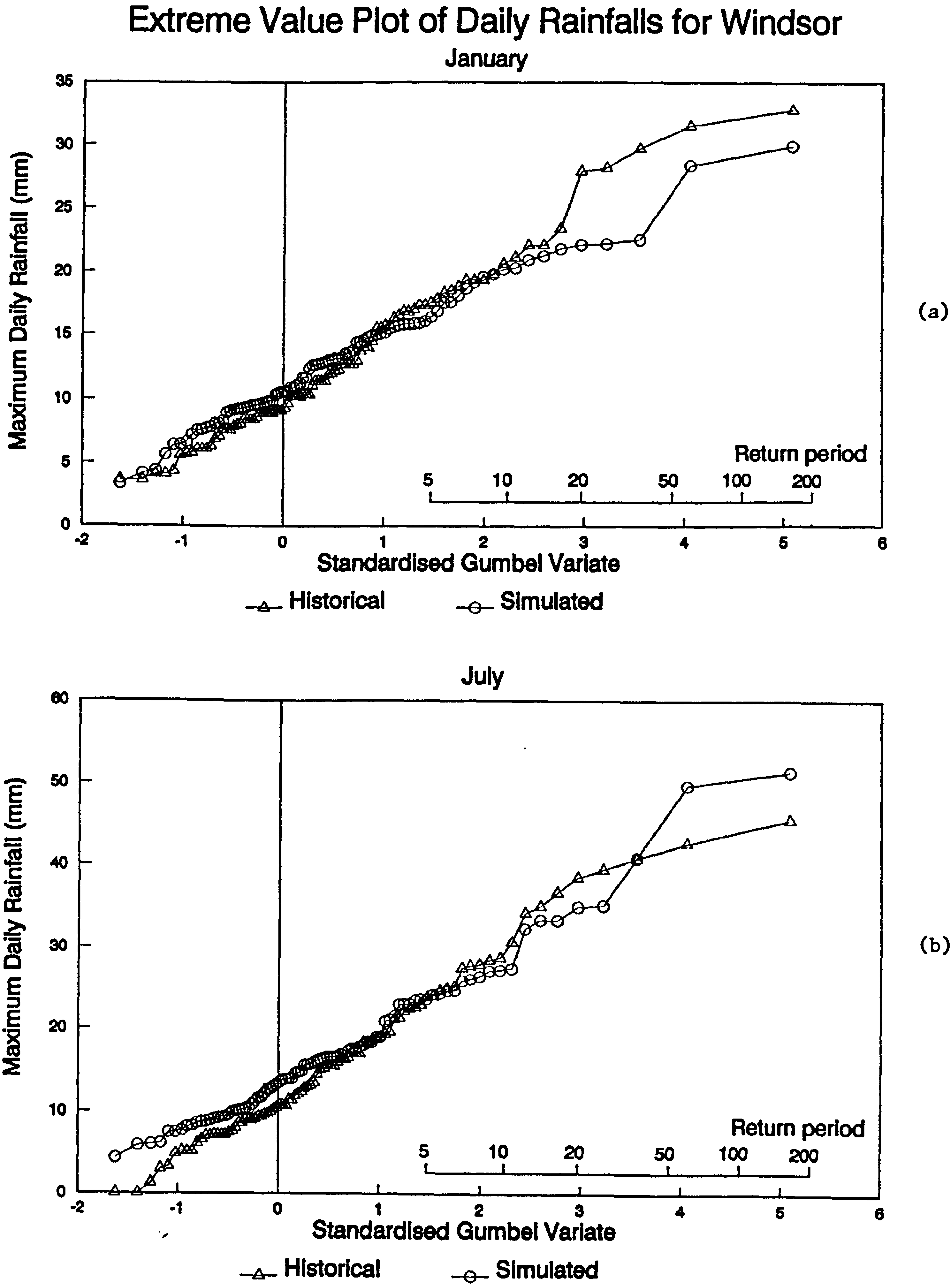
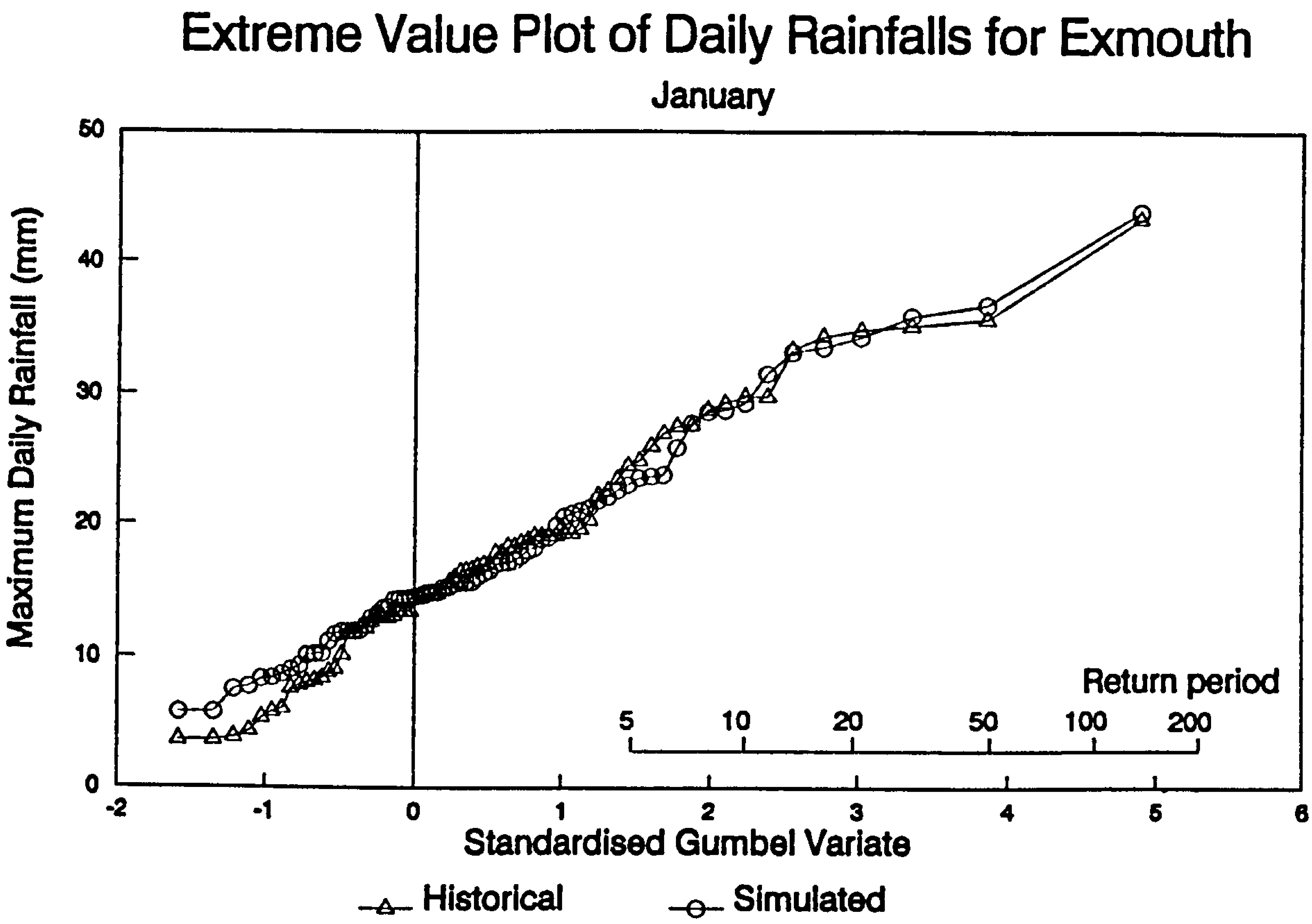
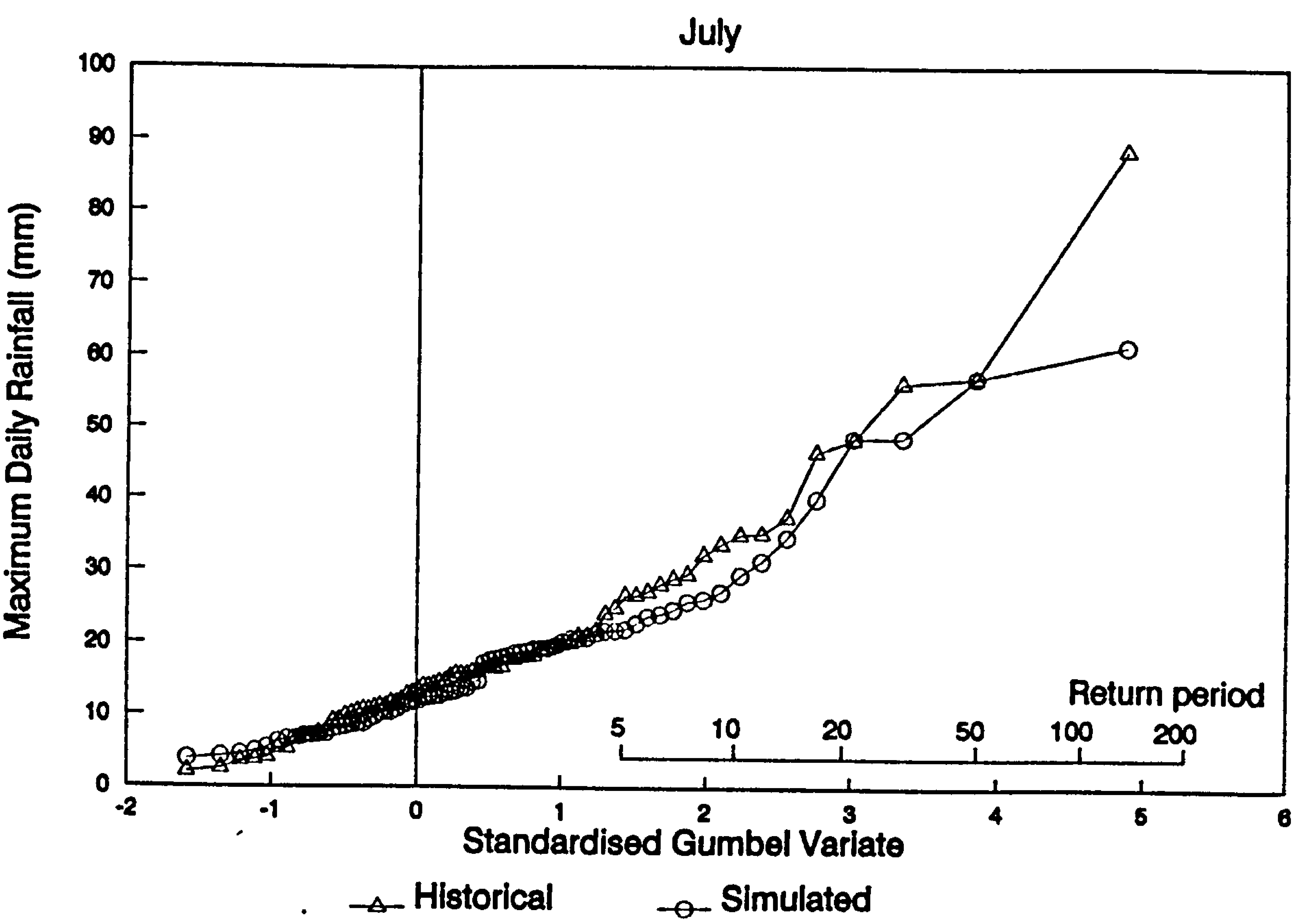


Figure 5.32
Extreme value plots for Windsor
using the regression model in the fitting procedure



(a)



(b)

Figure 5.33
Extreme value plots for Exmouth
using the regression model in the fitting procedure

CHAPTER 6
A REGIONALISED STOCHASTIC RAINFALL MODEL FOR THE UK

6.1 INTRODUCTION

The purpose of this Chapter is to find a regional/seasonal model that can be used to predict the parameters for the Neyman-Scott rainfall model at locations where rainfall data do not exist or are unavailable. An obvious approach to this problem is to attempt to regress the parameters of the model on location dependent variables that are likely to influence rainfall (e.g. altitude, distance from coast, etc). To allow for seasonal variation in the parameters harmonics can also be included in the regression model.

The parameters for the regression model will be estimated using a weighted least squares approach, where the weights will be equal to the number of years of rainfall data that were used to estimate the Neyman-Scott model parameters (such weights produce the best linear unbiased estimates for the parameters of the regression model).

For each month at each station the parameters of the Neyman-Scott model were estimated using the methods described in Chapters 4 and 5. It was found that 3 stations (out of 112) had one month in which the parameter estimates were of much greater magnitude than the typical parameter estimates obtained for the other station-months. These observations were not used in the analysis as they would unduly influence the least squares estimates in the regression analyses that follow. The total number of observations is therefore 12 × 112 - 3 = 1341. It was thought that these outlying observations were unlikely to reflect model inadequacy, but were probably a result of the minimisation procedure failing to reach a more typical local minimum, i.e. it was thought that the model could realise the same rainfall statistics with different parameter sets but that most of the time the parameter sets produced were of similar magnitude). To see whether this was the case the parameters for these station-months were estimated using the final regression model obtained in this Chapter, and the results are given in Appendix F.

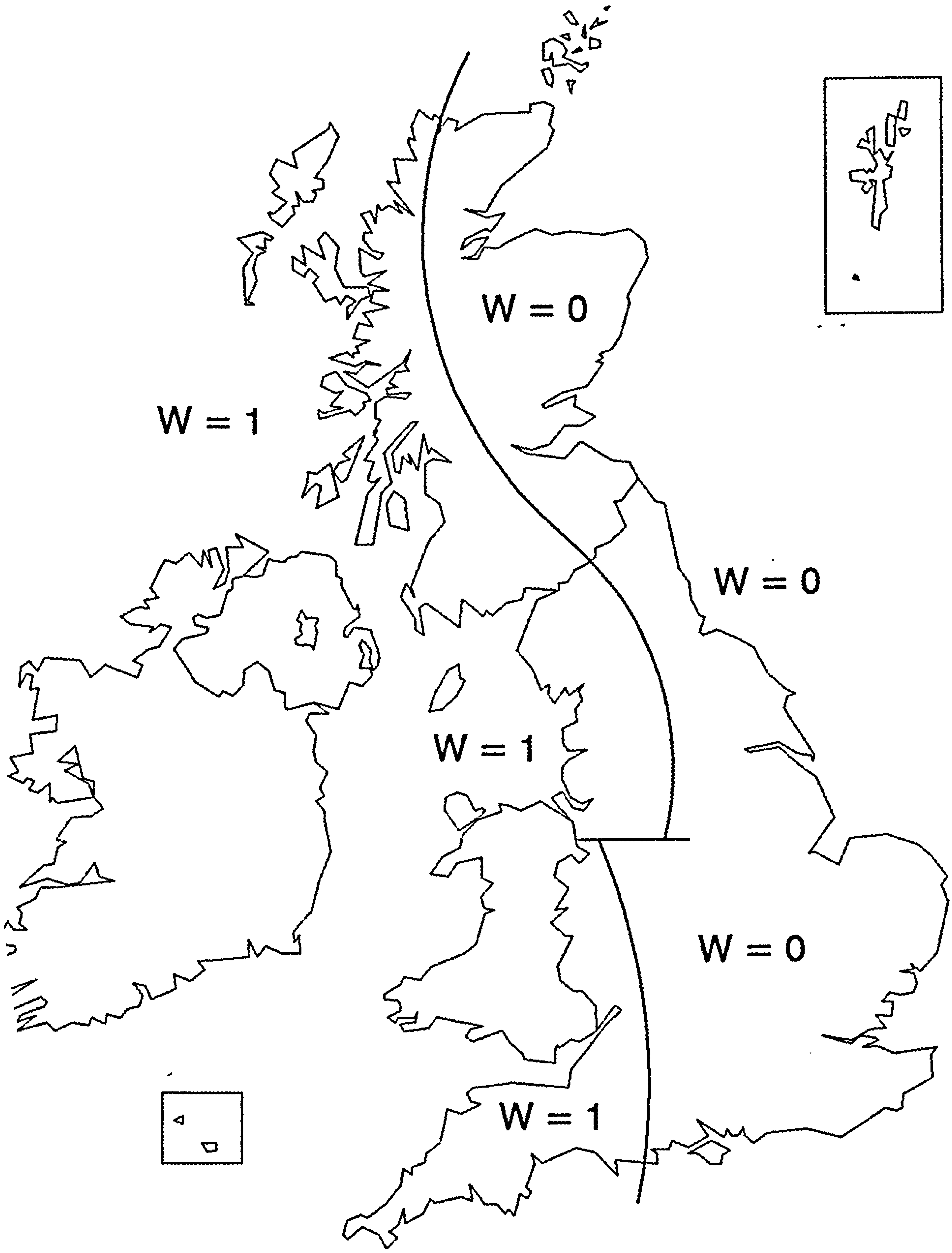
6.2 THE EXPLANATORY VARIABLES

The following variables were thought to be suitable for describing the regional variation of rainfall:

1) Altitude (A) in m×10,
2) North O/S Grid Reference (N) in km×10,
3) East/West effect (W),
4) Distance from Coast (C).

For each rainfall station A and N were known. For the East/West effect (W) the East-West dividing line found by *Wigley et al* (1984) (see Figure 6.1) was used (W was 1 if the station was in the west, otherwise W was zero). This dividing line seemed a good criterion on which to base an East/West effect as it corresponds to the well known 'rain shadow' effect of, for example, the Pennines.

Figure 6.1
East/West Effect (W)
(Using a dividing line proposed by Wigley et al (1984))



The distance from the coast (C) for each station was found by first measuring the distance D (in mm) on the O/S map used for this study, and then using the equation:

$$C = 1 + \text{trunc}\,(D/10) \qquad (6.1)$$

i.e. $C = 1, 2, 3, \ldots$

The variable C was used instead of D as no greater accuracy in 'Distance from coast' could be justified because the measurement of D was sometimes ambiguous (e.g. for station lying near an estuary).

To describe the seasonal effect the five harmonics (H1 - H5) given below were used:

$$H1 = A_1\cos(2\pi t/12) + B_1\sin(2\pi t/12),$$
$$H2 = A_2\cos(4\pi t/12) + B_2\sin(4\pi t/12),$$
$$H3 = A_3\cos(6\pi t/12) + B_3\sin(6\pi t/12),$$
$$H4 = A_4\cos(8\pi t/12) + B_4\sin(8\pi t/12),$$
$$H5 = A_5\cos(10\pi t/12) + B_5\sin(10\pi t/12),$$

where $t = 1, 2, \ldots, 12$ ($1 \equiv$ Jan, $2 \equiv$ Feb, ..., $12 \equiv$ Dec), so that H1 corresponds to 1 cycle per year, H2 to 2 cycles per year, etc. It will somestimes be found convenient to denote the cosine and sine components of H1 as C1 and S1 respectively, the cosine and sine components of H2 as C2 and S2 respectively, etc.

It was anticipated that not all of the harmonics would be needed in the final form of the regression model, but this was not the case as the contribution of each harmonic was found to be statistically significant (as is shown further on in this Chapter), suggesting a complex seasonal pattern.

6.3 A MULTIVARIATE REGRESSION PROBLEM

For each month of each station the parameters of the model were estimated using the methods described in the previous two Chapters. These parameter estimates together with the explanatory variables A, N, W, C, H1 - H5 were recorded in a data base. The number of stations used in the analysis was 112 (a few stations were omitted from the analysis because of incomplete or corrupt records) giving a total of 112 × 12 = 1344 observations. Three outlying observations were removed reducing the total to 1341 observations. The total number of explanatory variables was 14 (A, N, W, C, H1 - H5, counting 2 per harmonic), the number of model parameters is 5 (λ, β, η, ν, ξ), which gives a total of 19 variables in the regression model (so the complete data matrix has 1341 rows and 19 columns). In the least squares regression estimation procedures that follow each observation will be weighted by the number of years of data used to obtain that observation (thus producing the best linear unbiased estimates for the regression parameters).

One approach to the regression problem may be to perform a multiple regression analysis on each parameter separately. However, approaching the problem in this way would only be suitable if the parameters were uncorrelated random variables, with no relationship existing between them. To see whether the parameters were correlated the correlation matrix for the parameters estimates was found and is shown in Table 6.1.

Table 6.1

Correlation matrix for the parameter estimates

	$\hat{\lambda}$	$\hat{\beta}$	$\hat{\eta}$	$\hat{\nu}$	$\hat{\xi}$
$\hat{\lambda}$	1.0000	0.1790	-0.0504	-0.0075	0.5526
$\hat{\beta}$	0.1790	1.0000	0.5149	0.2254	0.2745
$\hat{\eta}$	-0.0504	0.5149	1.0000	0.6314	-0.1204
$\hat{\nu}$	-0.0075	0.2254	0.6314	1.0000	0.0686
$\hat{\xi}$	0.5526	0.2745	-0.1204	0.0686	1.0000

From the correlation matrix it can be seen that each row (or column) has at least one correlation exceeding 0.5 which is a significant correlation both practically and statistically. Hence a multiple regression approach that ignores the interaction between the parameters is unsuitable. Therefore the problem is best set up as a multivariate regression analysis. Letting $\underline{Y}$ be the 1341×5 matrix of parameter estimates, $\underline{X}$ be the 1341×15 matrix of observed explanatory variables (together with a column of 1's), $\underline{\gamma}$ be the matrix of regression parameters, and $\underline{\varepsilon}$ be the error matrix, then we assume:

$$\underline{Y} = \underline{X}\,\underline{\gamma} + \underline{\varepsilon} \tag{6.2}$$

Alternatively the regression model may be written as:

$$E(\lambda, \beta, \eta, \nu, \xi) = (1, A, N, W, C, C1, \ldots) \begin{pmatrix} \gamma_{\lambda 1} & \gamma_{\beta 1} & \cdots \\ \gamma_{\lambda A} & \gamma_{\beta A} & \cdots \\ \gamma_{\lambda N} & \gamma_{\beta N} & \cdots \\ \gamma_{\lambda W} & \vdots & \\ \vdots & & \end{pmatrix} \tag{6.3}$$

where the γ's are the regression parameters. Note that the parameter estimates (γ) obtained by the method of least squares applied to the multivariate problem are identical to the estimates that would be obtained with 5 separate multiple regression analyses on each model parameter (see, for example *Krzanowski* (1990)). Multivariate regression differs from multiple regression in the statistical tests on the explanatory variables that usually follow. In particular, in multivariate regression an explanatory variable is either in the model or out, whereas in multiple regression an explanatory variable may be in the regression model for one of the rainfall model's parameters and may be out of the regression model for another. In the multivariate regression problem the significance of an explanatory variable is tested by setting up null hypotheses of the form:

$$H_0: \gamma_{\lambda x} = \gamma_{\beta x} = \gamma_{\eta x} = \gamma_{\nu x} = \gamma_{\xi x} = 0,$$

where x is the explanatory variable (i.e. x = A, N, W, etc).
In testing the significance of the ith harmonic the following hypothesis would be used:

$$H_0: \gamma_{\lambda c_i} = \gamma_{\beta c_i} = \gamma_{\eta c_i} = \gamma_{\nu c_i} = \gamma_{\xi c_i} = \gamma_{\lambda s_i} = \gamma_{\beta s_i} = \gamma_{\eta s_i} = \gamma_{\nu s_i} = \gamma_{\xi s_i} = 0,$$

i.e. <u>both</u> the sine and cosine coefficients must be tested for each parameter simultaneously.

The hypotheses described above will be tested using Wilks' Lambda (Λ) - the details of this statistic are omitted as they are readily available in most texts on Multivariate Analysis (e.g. *Krzanowski* (1990))

6.4 FITTING THE MULTIVARIATE MODEL

The first model fitted was that described by (6.3) with all 5 harmonics included and no interaction terms included. The objective here is to dispense with as many terms as possible before considering interactions, thus reducing the number of interactions needing consideration. However, no terms could be removed from the model as the following analyses show.

Table 6.2 gives the results of an analysis of variance treating each dependent variable separately, i.e. the significance of 5 independent multiple regressions are tested separately. The F-ratios obtained in this Table are highly significant, from which we conclude that the choice of explanatory variables will give some success in predicting the parameters of the model.

To test the significance of each explanatory variable Wilks' Lambda (Λ) will be used. Table 6.3 gives the results of using this test statistic on each of the regional explanatory variables, and Table 6.4 gives the results for each of the seasonal (harmonic) variables. All the F-ratios in these Tables are highly significant indicating that all the explanatory variables could remain in the model.

Table 6.2
Analysis of variance for multiple regression treating each model parameter separately

Parameter	F Value	Prob>F	R-Squared
λ	142.108	0.0001	60%
β	22.384	0.0001	19%
η	30.874	0.0001	25%
ν	45.197	0.0001	32%
ξ	89.985	0.0001	49%

Table 6.3
Multivariate Statistical Tests using Wilks' Λ on regional variables

Variable	Wilks' Λ	F	Pr > F
A	0.7911	69.8	0.0001
N	0.5833	188.8	0.0001
W	0.6944	116.3	0.0001
C	0.9524	13.2	0.0001

Table 6.4
Multivariate Statistical Tests using Wilks' Λ on seasonal variables

Variable	Wilks' Λ	F	Pr > F
H1	0.2917	225.2	0.0000
H2	0.8996	14.4	0.0001
H3	0.9323	9.4	0.0001
H4	0.9686	4.3	0.0001
H5	0.9469	7.3	0.0001

6.5 FITTING THE REGRESSION MODEL WITH INTERACTIVE TERMS

As previously pointed out it will be necessary to consider interactions between terms. The total number of explanatory variables is 14, which means there are a total of $14 \times 13/2 = 91$ interactions. Interactions between the harmonic terms will not need to be considered (as two harmonics multiplied together will be equivalent to some other harmonic), and therefore the number of interactions needing consideration reduces to $91 - 10 \times 9/2 = 46$. The squares of the regional variables A, N, and C will also be considered. Note that $W^2 = W$ (as W takes the values 0 or 1 only), and so only W needs to be included. Thus the total number of terms of order 2 for consideration in the regression model is $46 + 3 = 49$.

The strategy for choosing which of these terms to include in the multivariate model was as follows. Five independent multiple regressions analyses on each of the model parameters were performed separately using a FORWARD model selection criterion. This selection criterion made the following steps:

1) The residual sum of squares R_1 was found for the multiple regression model which included only the 14 terms of order 1.

2) The residual sum of squares R was found for each of the 49 multiple regression models that had all terms of order 1 plus one term of order 2.

3) The difference in the residual sum of squares obtained in 1) and 2) divided by the residual mean square obtained in 2) gave an

F statistic from which the improvement in the models fit could be judged, i.e.

$$(R_1 - R)/(R/(1341-15-1)) \sim F_{1,1341-15-1}$$

4) If the largest of the 49 F-ratios in 3) was significant at the 1% level, then the term of order 2 was included in the model. (The data set was large so that terms significant at the 5% or 10% level were unlikely to be of practical significance).

5) Steps 1) to 4) were repeated with the significant term of order 2 included in step 1). The process continued until there were no F-ratios in step 4) significant at the 1% level.

The following terms of order 2 appeared in at least one of the FORWARD selected multiple regression models after the above 5 steps had been completed:

A^2, N^2, C^2, CW, AC, AN, NW, NH1, NH2, NH4, CH4, WH1, WH2, and WH4,

where NH1 = NC1 + NS1, NH2 = NC2 + NS2, etc (each harmonic interaction was jointly tested). This gives a total of 21 terms of order 2 for consideration in the multivariate model (counting 2 for each harmonic interaction).

Using all terms of order 2 with the 14 terms of order 1 the parameters (γ) for the multivariate model were estimated using the method of least squares. A multivariate test was then carried out on those terms (of order 2) that did not appear in any of the multiple regression models after step 5) above, the purpose being to see whether these terms could be dismissed from the multivariate model. The results are given in Table 6.5, from which it was decided to retain the null hypothesis, i.e. the remaining terms of order 2 did not significantly (at the 1% level) improve the multivariate model.

The parameters of the model were then re-estimated using only those terms of order 2 (as well as all the terms of order 1) in the FORWARD selected multivariate model. Multivariate tests were then performed on the interactive terms appearing in this model, the results being given in Table 6.6. From this Table it can be seen that the F-ratio for the interactive terms CH4, NH4, and AC are not significant at the 1% level, so that at least one of these terms could be removed from the model. To see whether more than one of these terms could be removed joint multivariate F-tests were performed, and the results are given in Table 6.7. From this Table it can be seen that CH4 and AC could be removed from the model and NH4 could be retained, giving a total of 32 explanatory variables in the multivariate regression model.

Table 6.5

Multivariate Test on Variables not selected for FORWARD models

Wilks' Λ	F	Pr > F
0.87	1.25	0.0260

Table 6.6

Multivariate tests for variables selected in the FORWARD Model

Variable	Wilks' Λ	F	Pr > F
AA	0.98	4.2	0.0008
NN	0.83	52.	0.0001
CC	0.94	16.	0.0001
CW	0.97	6.9	0.0001
AN	0.96	11.	0.0001
NW	0.96	10.	0.0001
AC	0.99	2.2	0.0539
NH1	0.92	12.	0.0001
NH2	0.97	3.9	0.0001
NH4	0.98	2.2	0.0169
WH1	0.86	21.	0.0001
WH2	0.98	2.3	0.0099
WH4	0.98	2.5	0.0052
CH4	0.99	0.99	0.4529

Table 6.7

Joint Multivariate tests for variables selected in the FORWARD Model

Variable	Wilks' Λ	F	Pr > F
NH4,CH4	0.97	1.7	0.0207
NH4,CH4,AC	0.97	1.8	0.0068
CH4,AC	0.98	1.4	0.1459

Using the 32 explanatory variables the parameters (γ) of the multivariate model were estimated using the Gentleman-Givens Method which reduces numerical inaccuracies caused by inverting a large data matrix. Tables 6.8 - 6.12 give estimates of these γ for each of the dependent variables (i.e. Neyman-Scott model parameters). The P-Values in Tables 6.8 - 6.12 are the values obtained under the null hypotheses that the explanatory variable is not needed in the multiple regression model. Although it would not be appropriate to use this test in building the multivariate model, it is useful in determining which explanatory variables are explaining most of the variation in the dependent variables.

Table 6.13 gives the results of an analysis of variance for the separate multiple regression on each dependent variable, and the R^2 values for each dependent variable. To reflect the multivariate nature of the model, the R^2 values for the mean daily rainfall $\mu(24)$ and the proportion of dry days $\phi(24)$ are also shown in this Table. These values give an overall indication of the performance of the regional model. It can be seen that the model has more success in predicting the proportion of dry days than the mean daily rainfall. To see this visually, a random sample of 2 daily stations was drawn from each of the Wigley regions and the mean daily rainfall and proportion of dry days plotted using the parameter estimates given by the regression model and the estimates obtained from the data (Figures 6.2-6.11). As the total number of station-months was large (about 1341), the effect on the regression estimates of any individual station-month is going to be negligible, so that the plots give some indication of the predictive capability of the multivariate regression model on sites not used in the analysis.

Table 6.8

Parameter Estimates for λ

Explanatory Variable	Parameter Estimate	Standard Error	P-Value
CONSTANT	0.0138742271	0.000386627	0.0001
A	-0.0000292872	0.000034466	0.3956
N	0.0012379447	0.000151387	0.0001
W	-0.0015956772	0.000350915	0.0001
C	-0.0004467139	0.000048667	0.0001
C1	0.0031267775	0.000187706	0.0001
S1	0.0020259476	0.000187767	0.0001
C2	0.0013255149	0.000187643	0.0001
S2	-0.0007516314	0.000187827	0.0001
C3	0.0006667431	0.000082061	0.0001
S3	0.0001296318	0.000082124	0.1147
C4	-0.0009089374	0.000187642	0.0001
S4	-0.0007417884	0.000187830	0.0001
C5	-0.0002951170	0.000082109	0.0003
S5	0.0006067482	0.000082077	0.0001
AN	0.0000160326	0.000004359	0.0002
NW	0.0000949378	0.000073960	0.1995
NC1	0.0000207914	0.000041344	0.6151
NS1	-0.0003689149	0.000041413	0.0001
NC2	-0.0001568978	0.000041325	0.0002
NS2	0.0000973809	0.000041432	0.0189
NC4	0.0001575037	0.000041324	0.0001
NS4	0.0000622454	0.000041433	0.1333
WC	0.0002205185	0.000046188	0.0001
WC1	-0.0004591073	0.000175230	0.0089
WS1	-0.0011349163	0.000175038	0.0001
WC2	-0.0000058556	0.000175032	0.9733
WS2	0.0001576206	0.000175235	0.3686
WC4	0.0006306925	0.000175031	0.0003
WS4	0.0004905837	0.000175236	0.0052
AA	0.0000014634	0.000000797	0.0665
NN	-0.0000965985	0.000016677	0.0001
CC	0.0000180567	0.000002370	0.0001

Table 6.9

Parameter Estimates for β

Explanatory Variable	Parameter Estimate	Standard Error	P-Value
INTERCEPT	0.1664563853	0.008288596	0.0001
A	0.0013254115	0.000738884	0.0731
N	-0.0057692188	0.003245468	0.0757
W	-0.0603957698	0.007522989	0.0001
C	0.0006293085	0.001043339	0.5465
C1	0.0216544008	0.004024086	0.0001
S1	0.0103808886	0.004025385	0.0100
C2	0.0121702678	0.004022737	0.0025
S2	0.0060989181	0.004026679	0.1301
C3	0.0018784352	0.001759247	0.2858
S3	0.0034074486	0.001760597	0.0532
C4	-0.0181397093	0.004022717	0.0001
S4	-0.0070715730	0.004026738	0.0793
C5	0.0007158923	0.001760269	0.6843
S5	0.0004736784	0.001759593	0.7878
AN	-0.0003826584	0.000093450	0.0001
NW	0.0087625467	0.001585570	0.0001
NC1	-0.0009945873	0.000886346	0.2620
NS1	-0.0010565827	0.000887823	0.2342
NC2	-0.0009539141	0.000885925	0.2818
NS2	-0.0002045539	0.000888217	0.8179
NC4	0.0029845048	0.000885916	0.0008
NS4	0.0004729344	0.000888245	0.5945
WC	0.0002024421	0.000990194	0.8380
WC1	-0.0262030553	0.003756615	0.0001
WS1	0.0006542113	0.003752505	0.8616
WC2	-0.0051184911	0.003752368	0.1728
WS2	-0.0070270555	0.003756731	0.0616
WC4	0.0071888144	0.003752361	0.0556
WS4	0.0042380513	0.003756754	0.2595
AA	-0.0000425965	0.000017085	0.0128
NN	0.0002211374	0.000357524	0.5363
CC	-0.0000264315	0.000050803	0.6030

Table 6.10
Parameter Estimates for η

Explanatory Variable	Parameter Estimate	Standard Error	P-Value
INTERCEPT	1.120781681	0.071061474	0.0001
A	0.001122746	0.006334746	0.8594
N	0.037355428	0.027824708	0.1797
W	-0.079393790	0.064497615	0.2186
C	0.022420413	0.008944962	0.0123
C1	0.150732113	0.034500109	0.0001
S1	0.280044997	0.034511246	0.0001
C2	-0.039089942	0.034488542	0.2572
S2	0.092941634	0.034522342	0.0072
C3	-0.038824199	0.015082734	0.0102
S3	0.062421691	0.015094311	0.0001
C4	-0.092888714	0.034488372	0.0072
S4	0.001687004	0.034522851	0.9610
C5	0.003297514	0.015091492	0.8271
S5	-0.029093928	0.015085703	0.0540
AN	-0.001751140	0.000801182	0.0290
NW	0.060569855	0.013593727	0.0001
NC1	-0.017020874	0.007599003	0.0253
NS1	-0.011488820	0.007611661	0.1314
NC2	0.005205398	0.007595394	0.4933
NS2	-0.013968988	0.007615043	0.0668
NC4	0.010698779	0.007595314	0.1592
NS4	-0.002184500	0.007615283	0.7743
WC	-0.015619471	0.008489328	0.0660
WC1	-0.099408399	0.032206977	0.0021
WS1	-0.075281177	0.032171735	0.0194
WC2	0.027058274	0.032170565	0.4005
WS2	-0.108030490	0.032207965	0.0008
WC4	-0.002374633	0.032170499	0.9412
WS4	-0.014673041	0.032208163	0.6488
AA	0.000011514	0.000146472	0.9374
NN	-0.012427648	0.003065201	0.0001
CC	-0.000772811	0.000435556	0.0762

Table 6.11

Parameter Estimates for ν

Explanatory Variable	Parameter Estimate	Standard Error	P-Value
INTERCEPT	4.4617138085	0.333521676	0.0001
A	0.1156717318	0.029731654	0.0001
N	-0.4410410415	0.130593169	0.0008
W	1.4487384049	0.302714701	0.0001
C	0.2774578297	0.041982507	0.0001
C1	0.7672839854	0.161923664	0.0001
S1	1.1945645510	0.161975933	0.0001
C2	-0.2909719128	0.161869374	0.0725
S2	0.0983670842	0.162028013	0.5439
C3	-0.2369340718	0.070789677	0.0008
S3	0.3310028941	0.070844014	0.0001
C4	-0.0872899688	0.161868577	0.5898
S4	0.0791741048	0.162030401	0.6252
C5	-0.0213272179	0.070830781	0.7634
S5	-0.0645265515	0.070803613	0.3623
AN	-0.0145270328	0.003760285	0.0001
NW	0.3601488366	0.063801136	0.0001
NC1	-0.0605166565	0.035665346	0.0900
NS1	-0.0754167745	0.035724757	0.0350
NC2	0.0851412034	0.035648409	0.0171
NS2	-0.0284699587	0.035740632	0.4258
NC4	0.0019231660	0.035648033	0.9570
NS4	-0.0338355615	0.035741759	0.3440
WC	-0.1607952261	0.039844022	0.0001
WC1	0.2802085898	0.151161020	0.0640
WS1	-0.0698251335	0.150995616	0.6438
WC2	-0.0501882673	0.150990125	0.7396
WS2	-0.5056188060	0.151165658	0.0008
WC4	-0.1999628607	0.150989815	0.1856
WS4	-0.2975571048	0.151166588	0.0492
AA	-0.0014028702	0.000687455	0.0415
NN	0.0435101503	0.014386288	0.0025
CC	-0.0113508555	0.002044251	0.00

Table 6.12

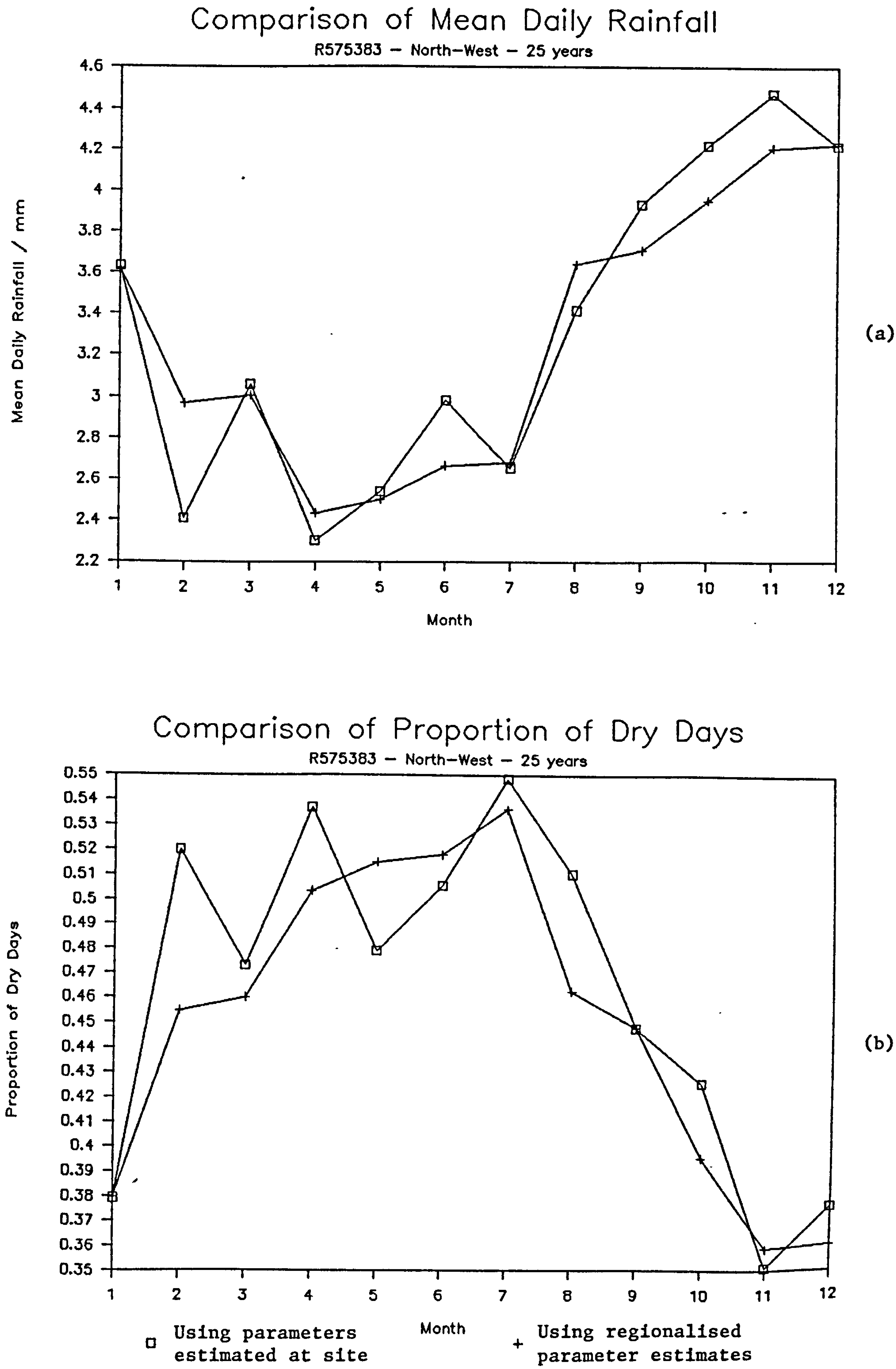
Parameter Estimates for ξ

Explanatory Variable	Parameter Estimate	Standard Error	P-Value
INTERCEPT	0.9064980664	0.028879845	0.0001
A	0.0003350582	0.002574482	0.8965
N	-0.0733628953	0.011308142	0.0001
W	-0.1240940733	0.026212250	0.0001
C	-0.0025967725	0.003635291	0.4752
C1	0.1443043564	0.014021069	0.0001
S1	0.1858651591	0.014025595	0.0001
C2	-0.0006757560	0.014016368	0.9616
S2	-0.0347332583	0.014030104	0.0134
C3	0.0106345485	0.006129721	0.0830
S3	0.0078405869	0.006134426	0.2014
C4	-0.0203850421	0.014016299	0.1461
S4	-0.0036819550	0.014030311	0.7930
C5	-0.0068851990	0.006133280	0.2618
S5	0.0091117818	0.006130928	0.1375
AN	0.0003634532	0.000325605	0.2645
NW	0.0093026815	0.005524579	0.0924
NC1	0.0004740334	0.003088284	0.8780
NS1	-0.0171389596	0.003093429	0.0001
NC2	0.0049464228	0.003086817	0.1093
NS2	0.0043278022	0.003094803	0.1622
NC4	0.0040753978	0.003086785	0.1870
NS4	-0.0020420496	0.003094901	0.5095
WC	0.0053914199	0.003450118	0.1184
WC1	-0.0888451903	0.013089125	0.0001
WS1	-0.0429460020	0.013074802	0.0010
WC2	-0.0046091109	0.013074327	0.7245
WS2	-0.0067888701	0.013089526	0.6041
WC4	0.0003232054	0.013074300	0.9803
WS4	0.0078137957	0.013089607	0.5506
AA	-0.0001229886	0.000059527	0.0390
NN	0.0150149053	0.001245718	0.0001
CC	0.0002258862	0.000177013	0.2021

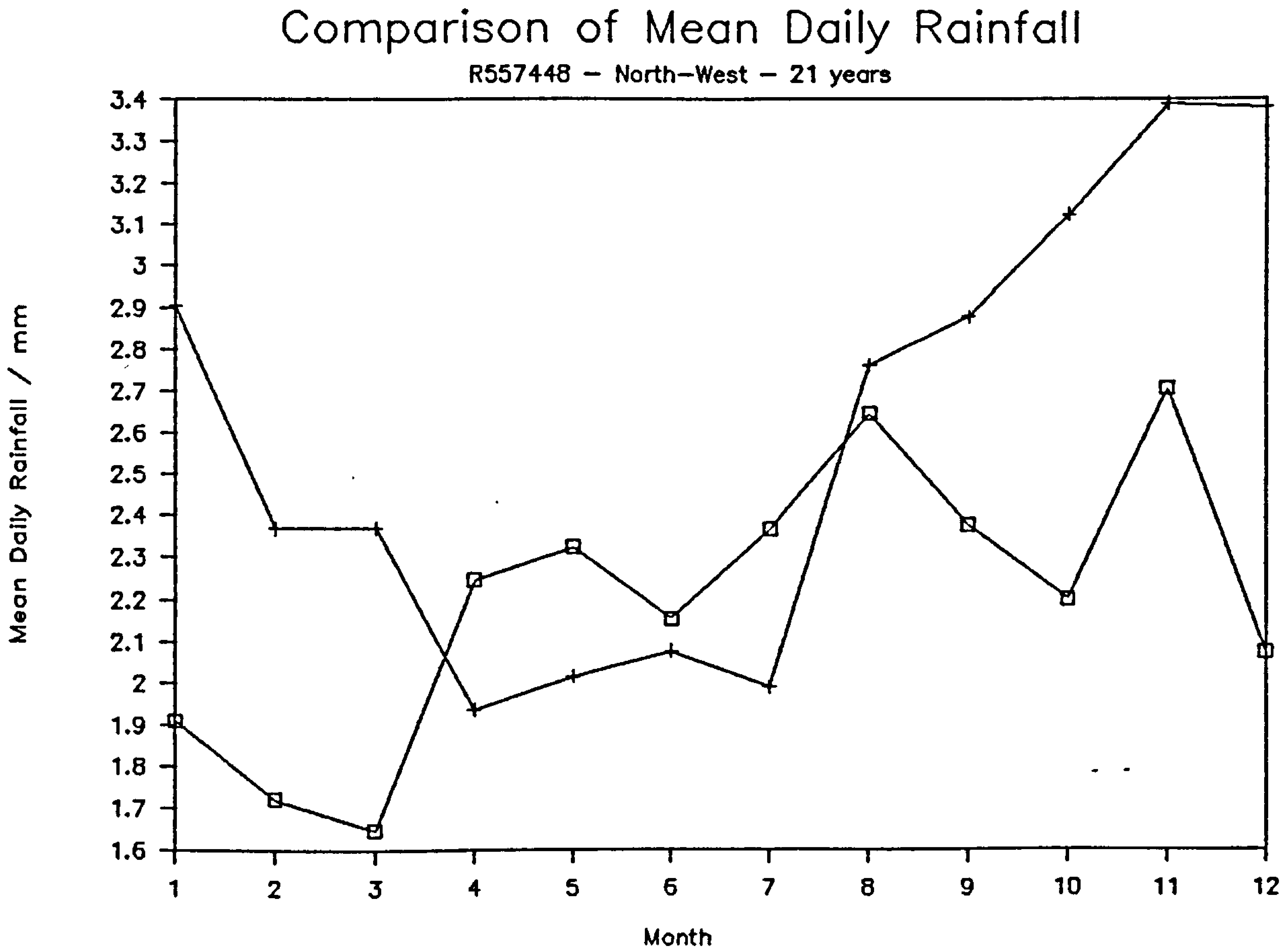
Table 6.13
Analysis of Variance
for the final regression model

Dependent Variable	F Value	Pr > F	R-Squared
λ	82.6	0.0001	67%
β	14.3	0.0001	26%
η	17.1	0.0001	29%
ν	24.5	0.0001	37%
ξ	54.2	0.0001	57%
$\mu(24)$	-	-	58%
$\phi(24)$	-	-	72%

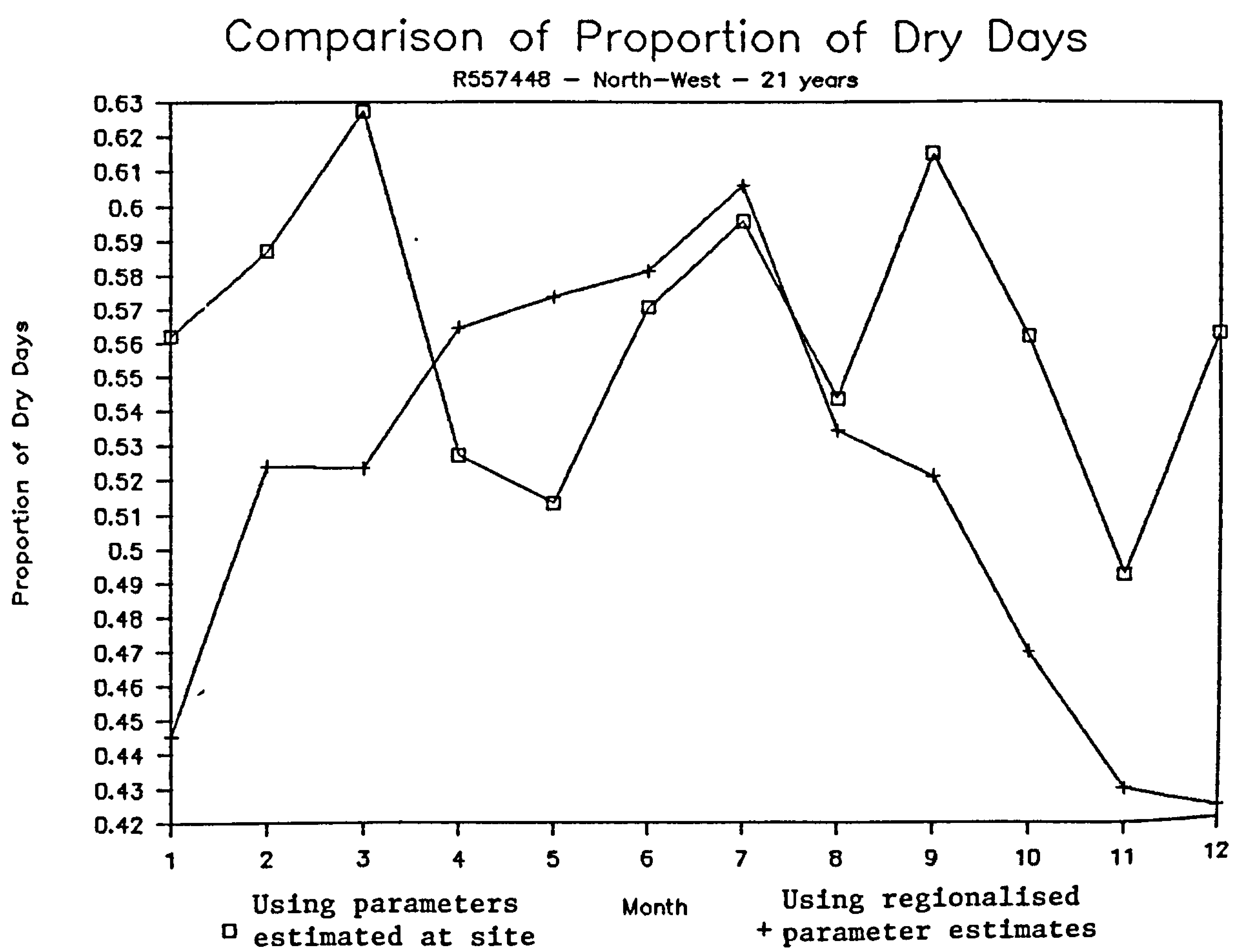
Note: In the above Table, $\mu(24)$ and $\phi(24)$ are the daily mean and proportion of dry days respectively given by the N-S model (recall from Chapter 4 that they are functions of the N-S model parameters.



Figures 6.2(a) and (b)

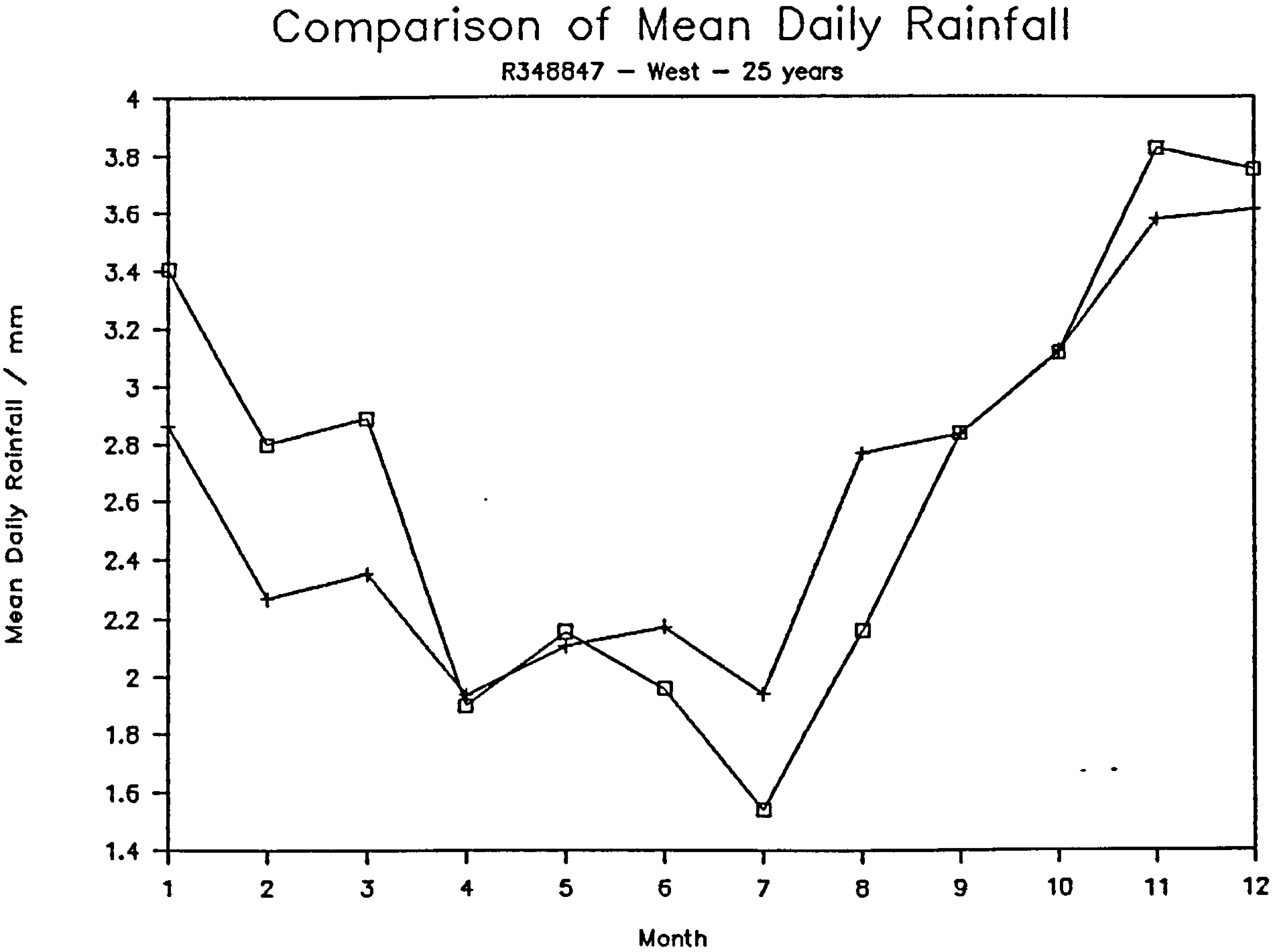


(a)

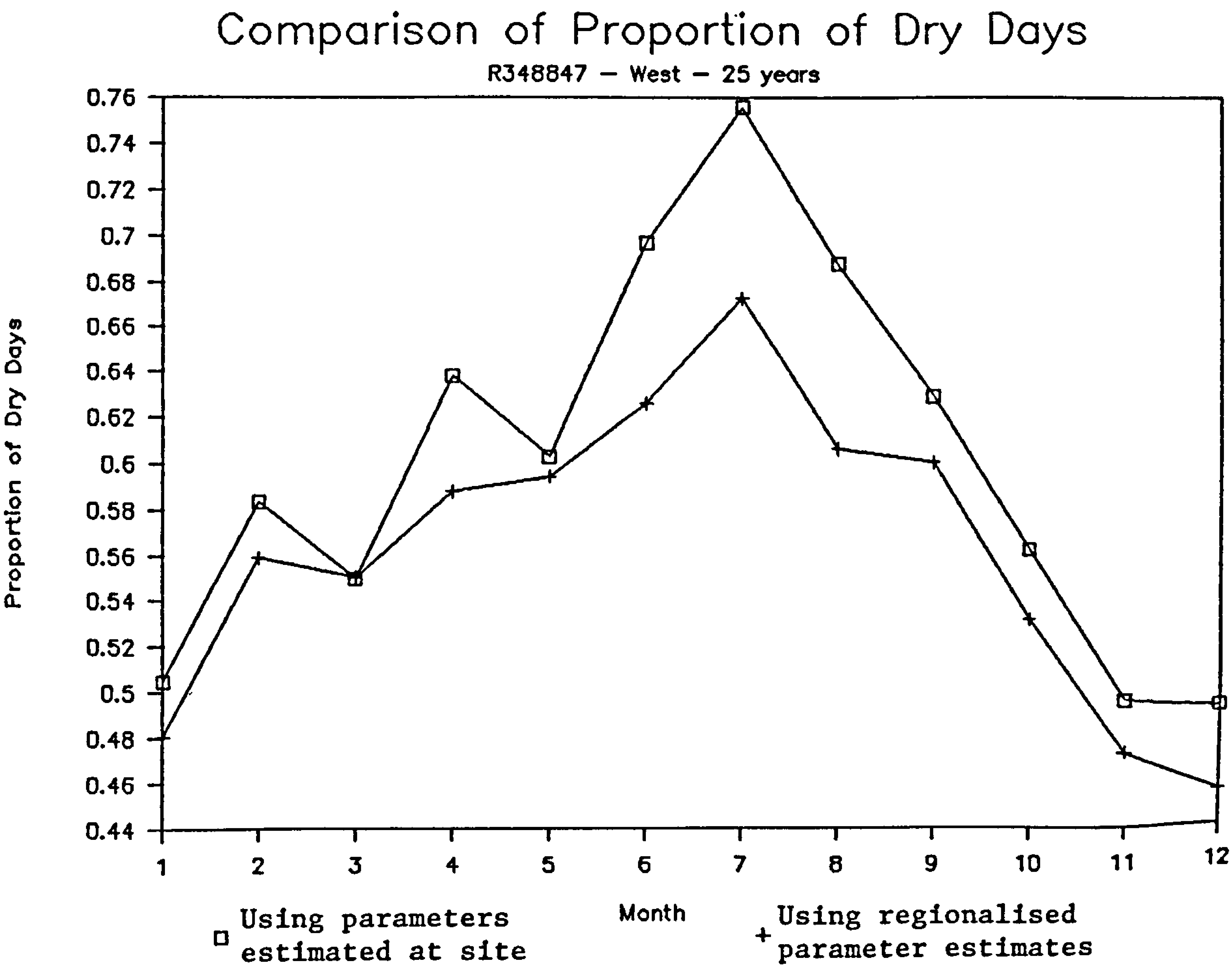


(b)

Figures 6.3(a) and (b)

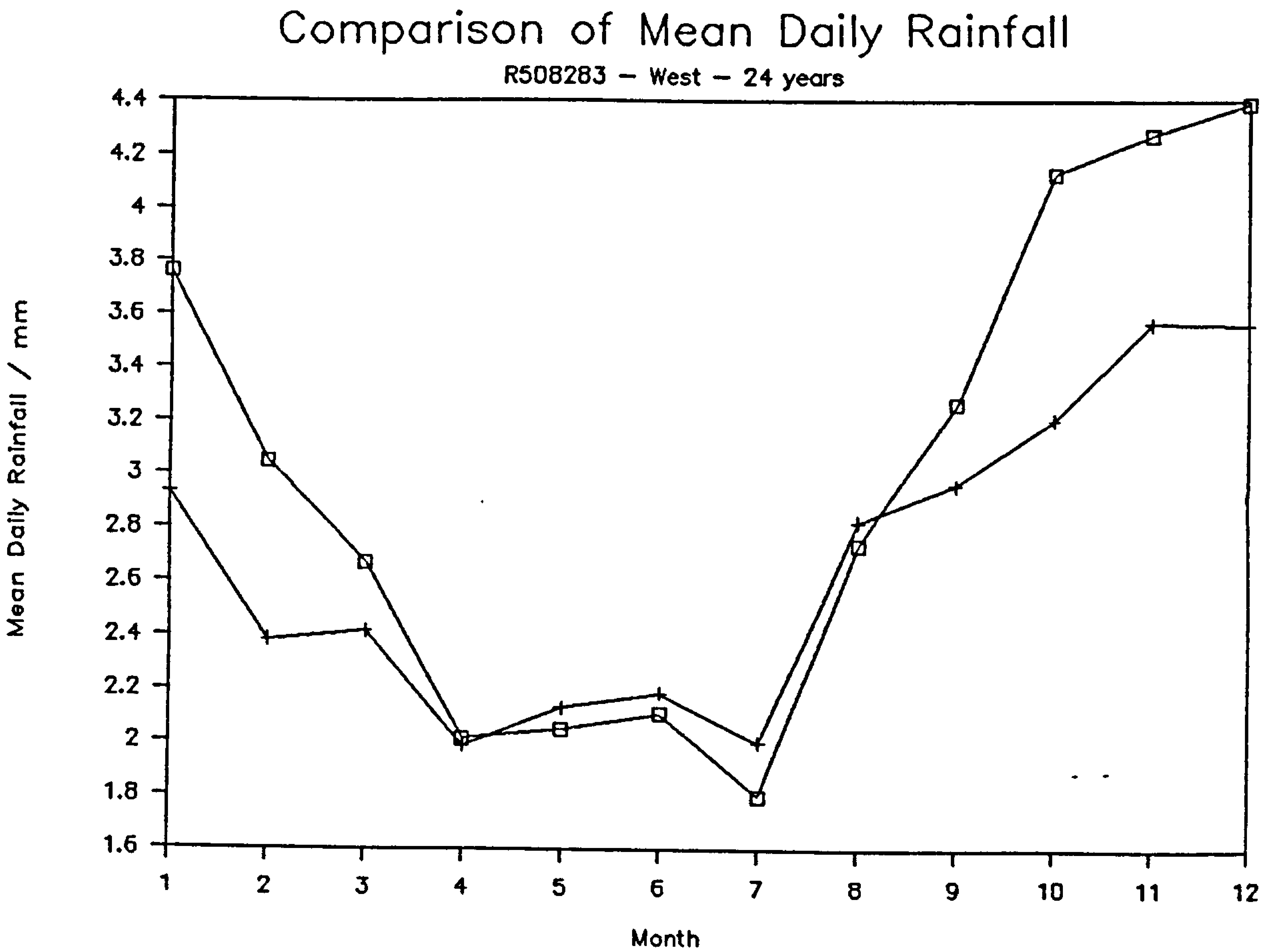


(a)

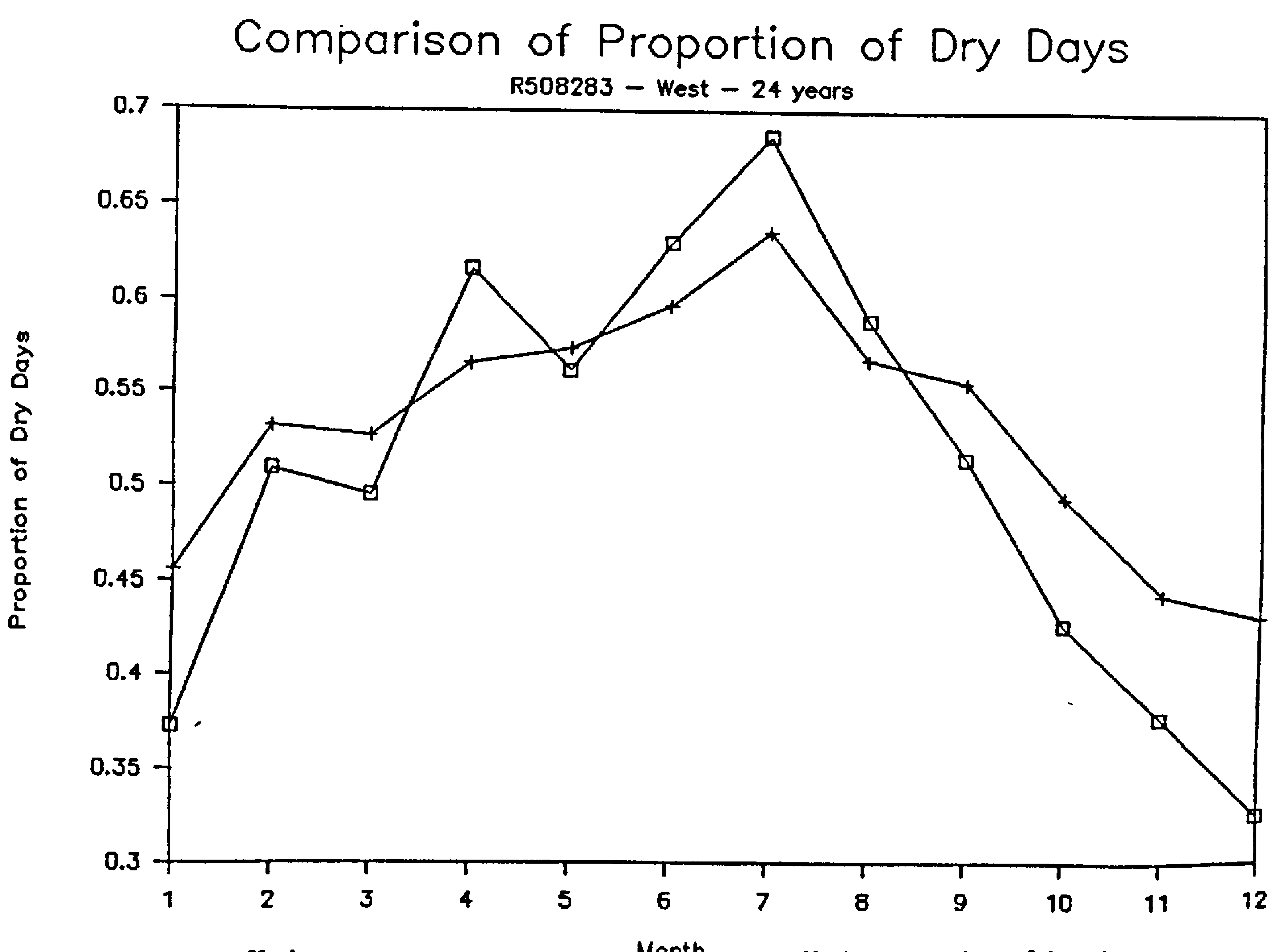


(b)

Figures 6.4(a) and (b)



(a)

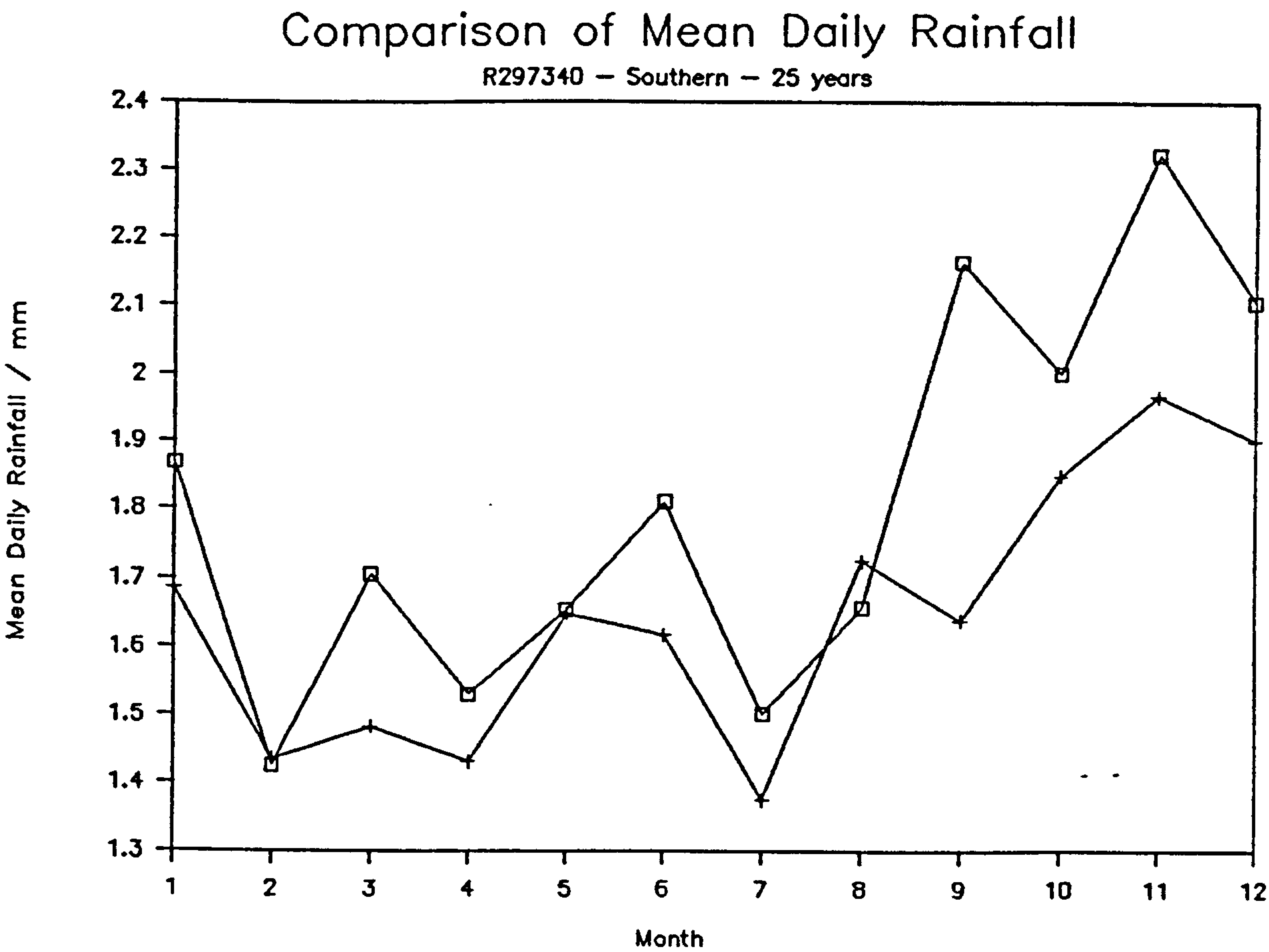


(b)

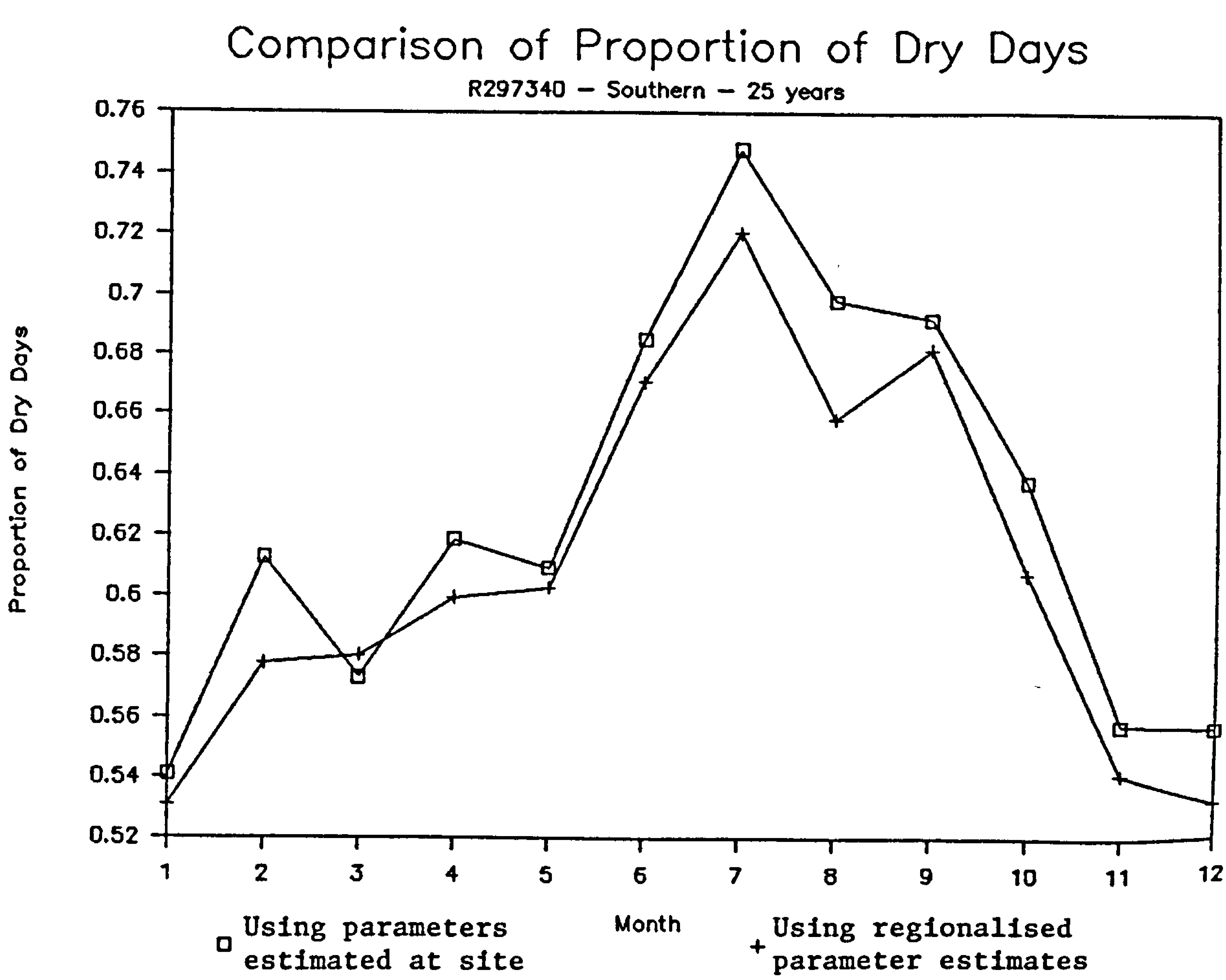
□ Using parameters estimated at site

\+ Using regionalised parameter estimates

Figures 6.5(a) and (b)

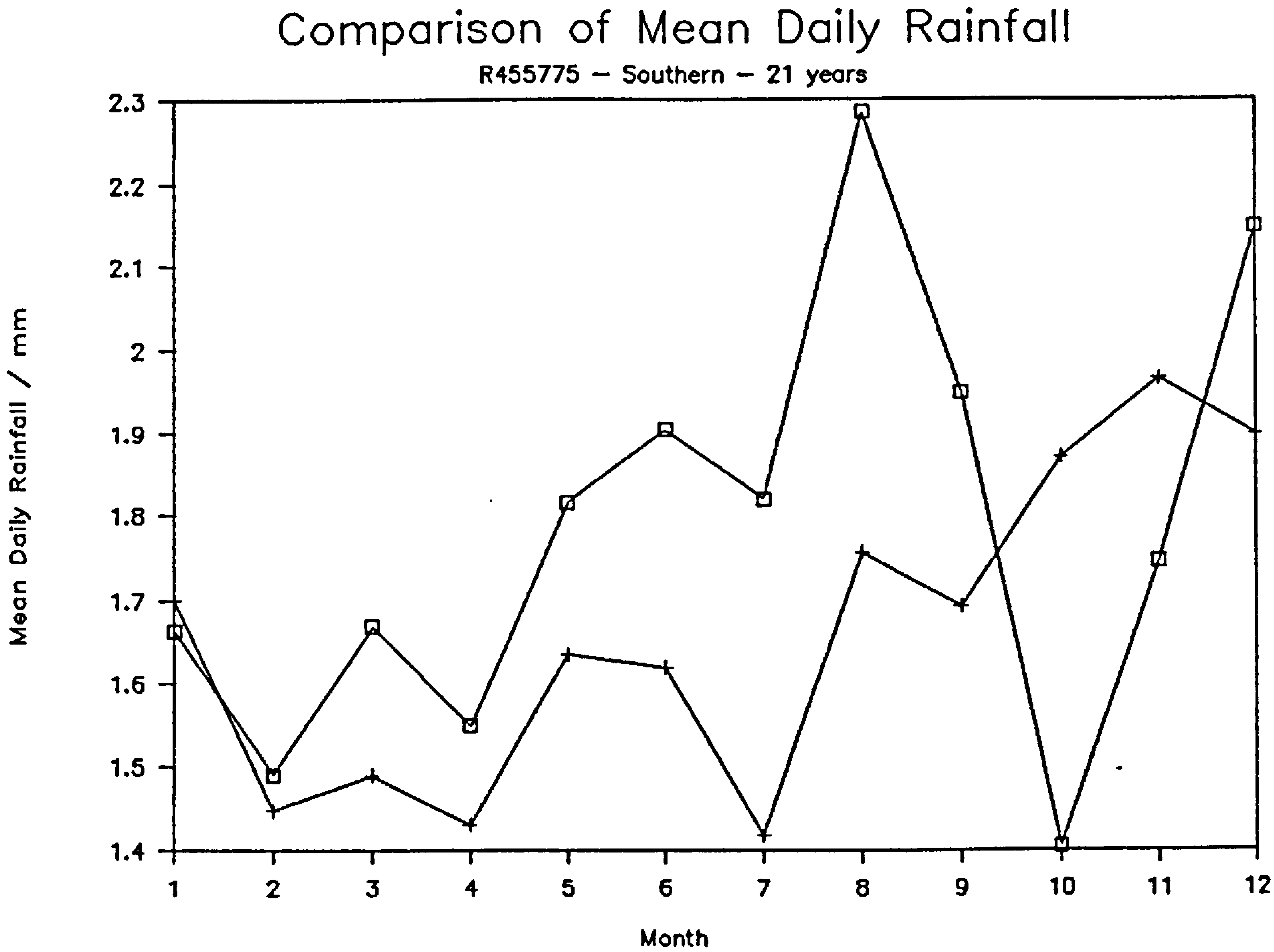


(a)

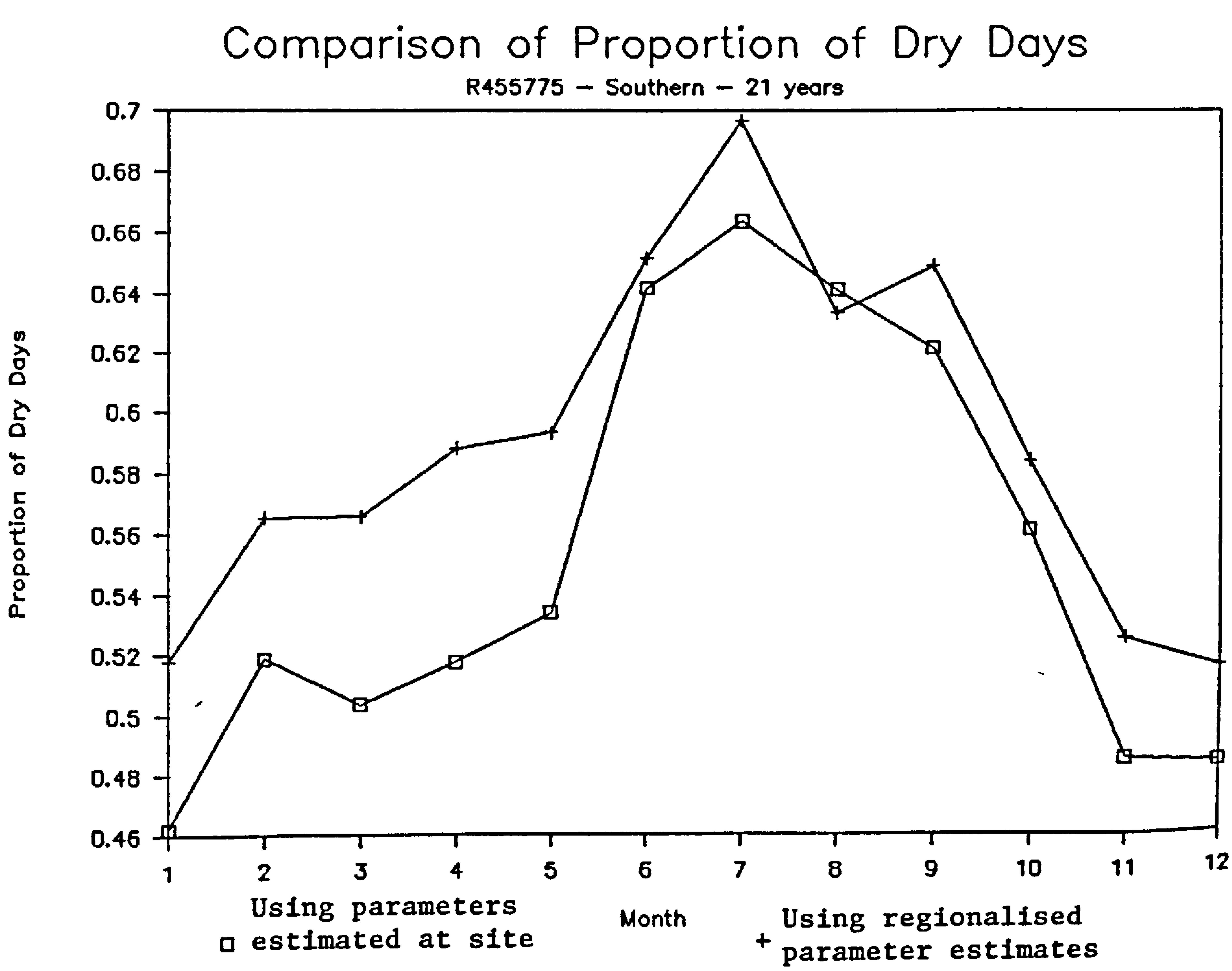


(b)

Figures 6.6(a) and (b)

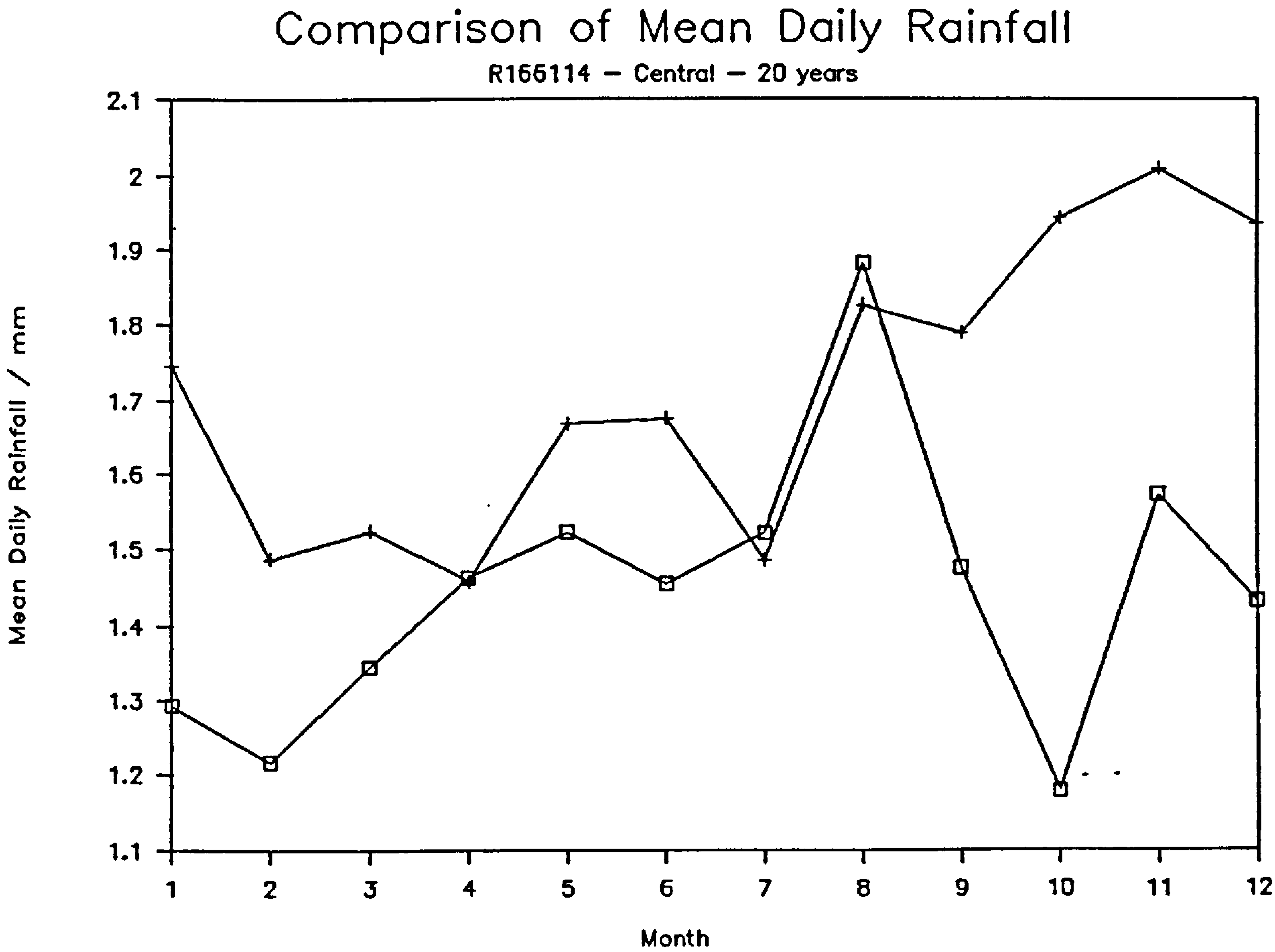


(a)

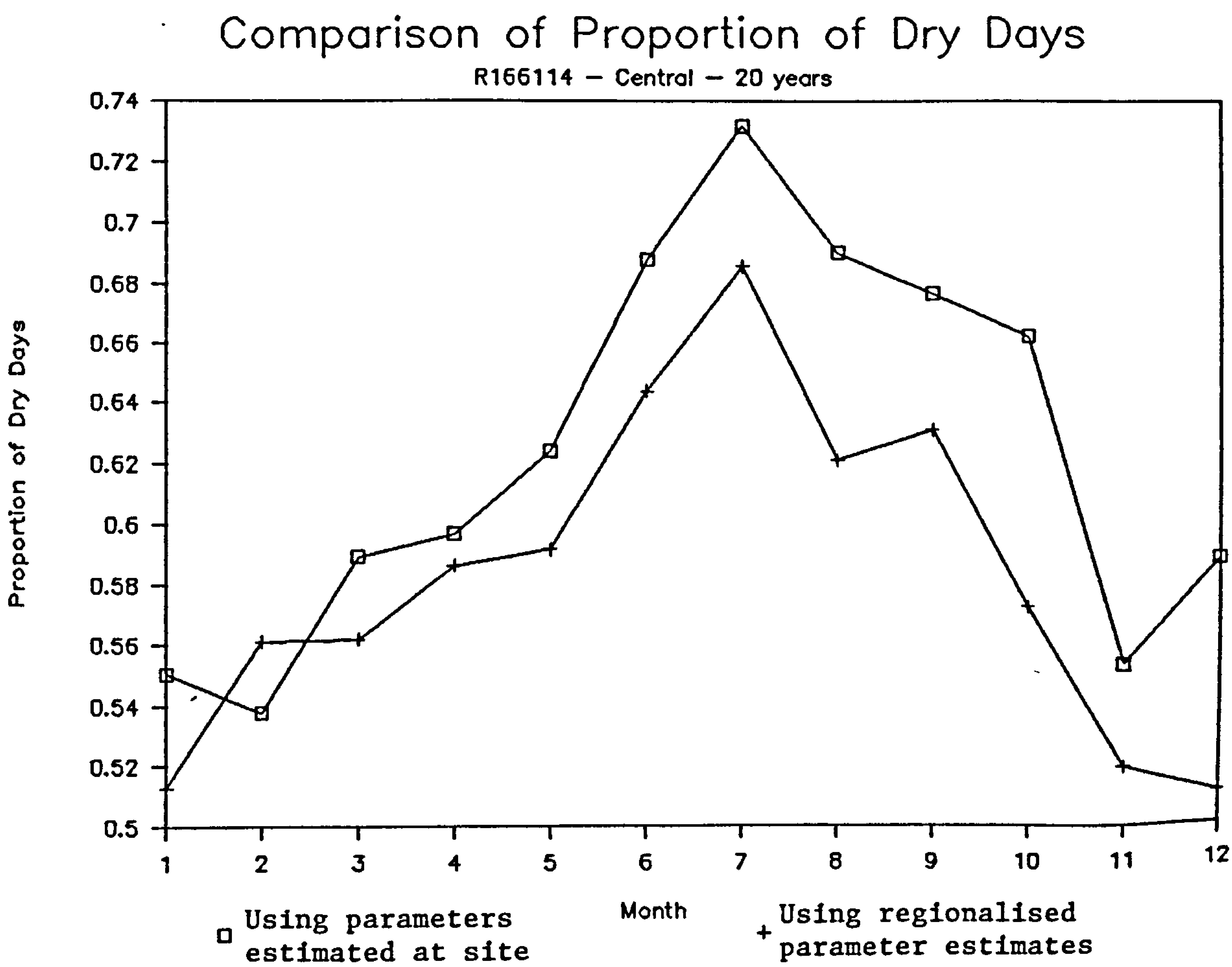


(b)

Figures 6.7(a) and (b)

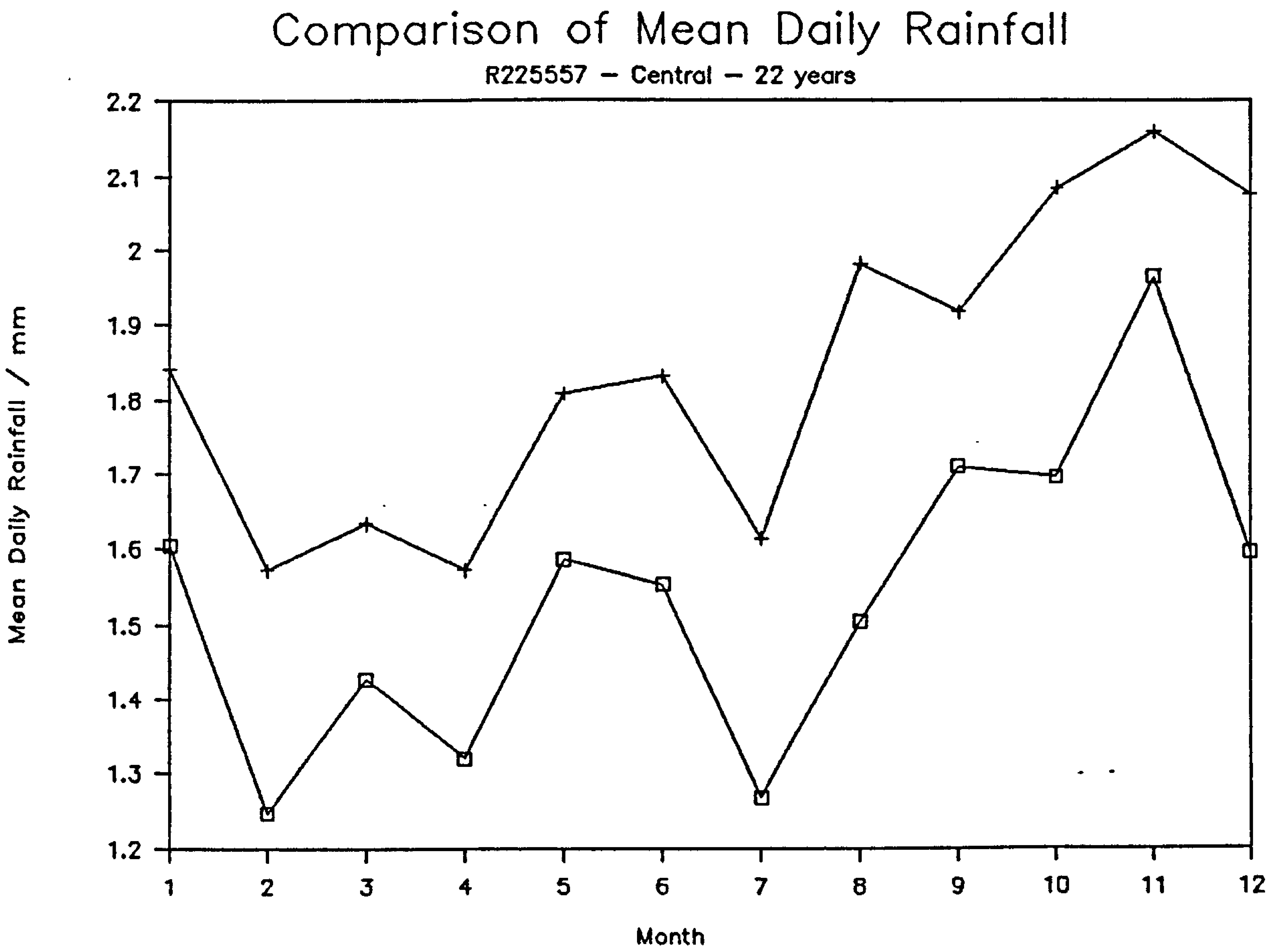


(a)

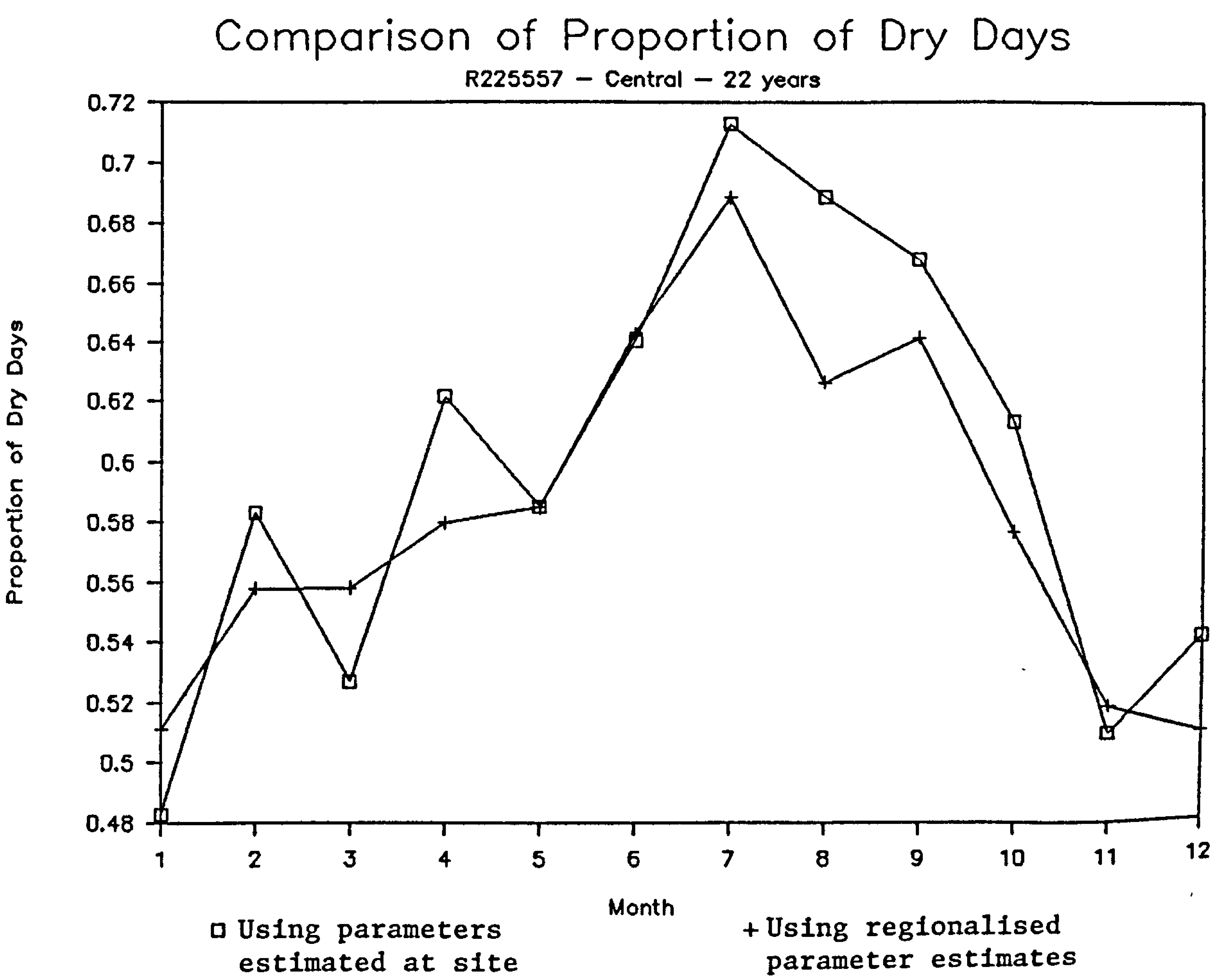


(b)

Figures 6.8(a) and (b)

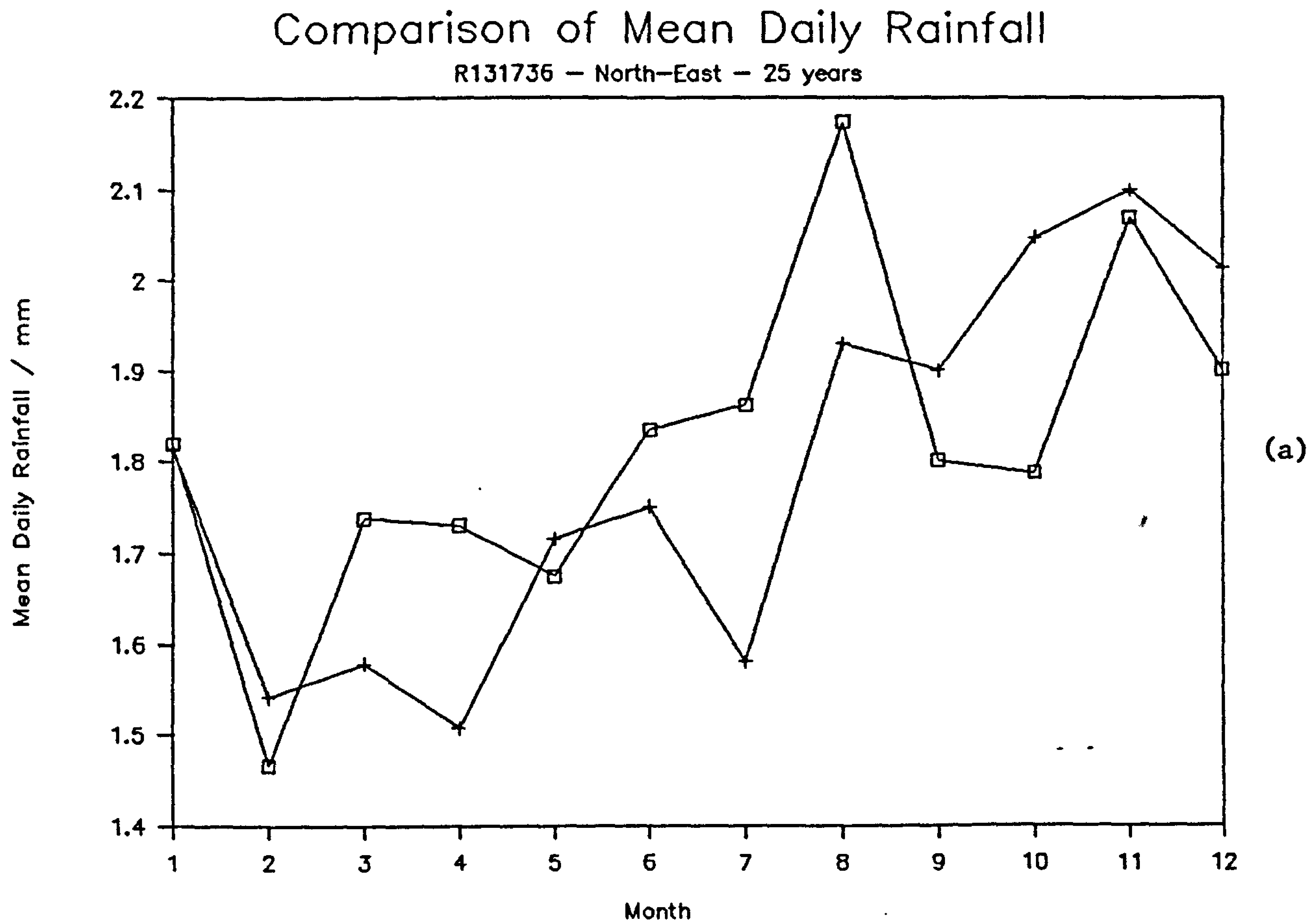


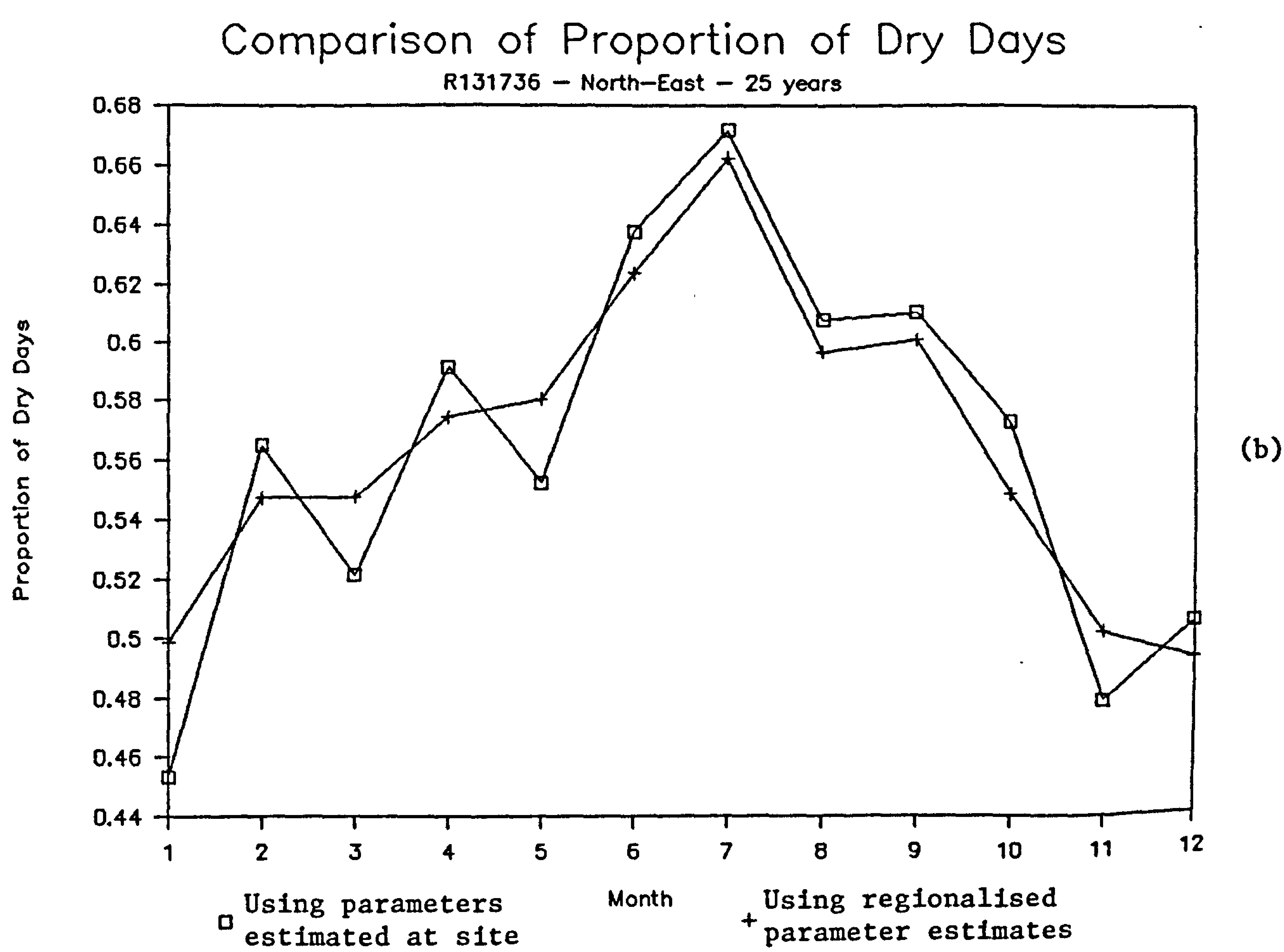
(a)



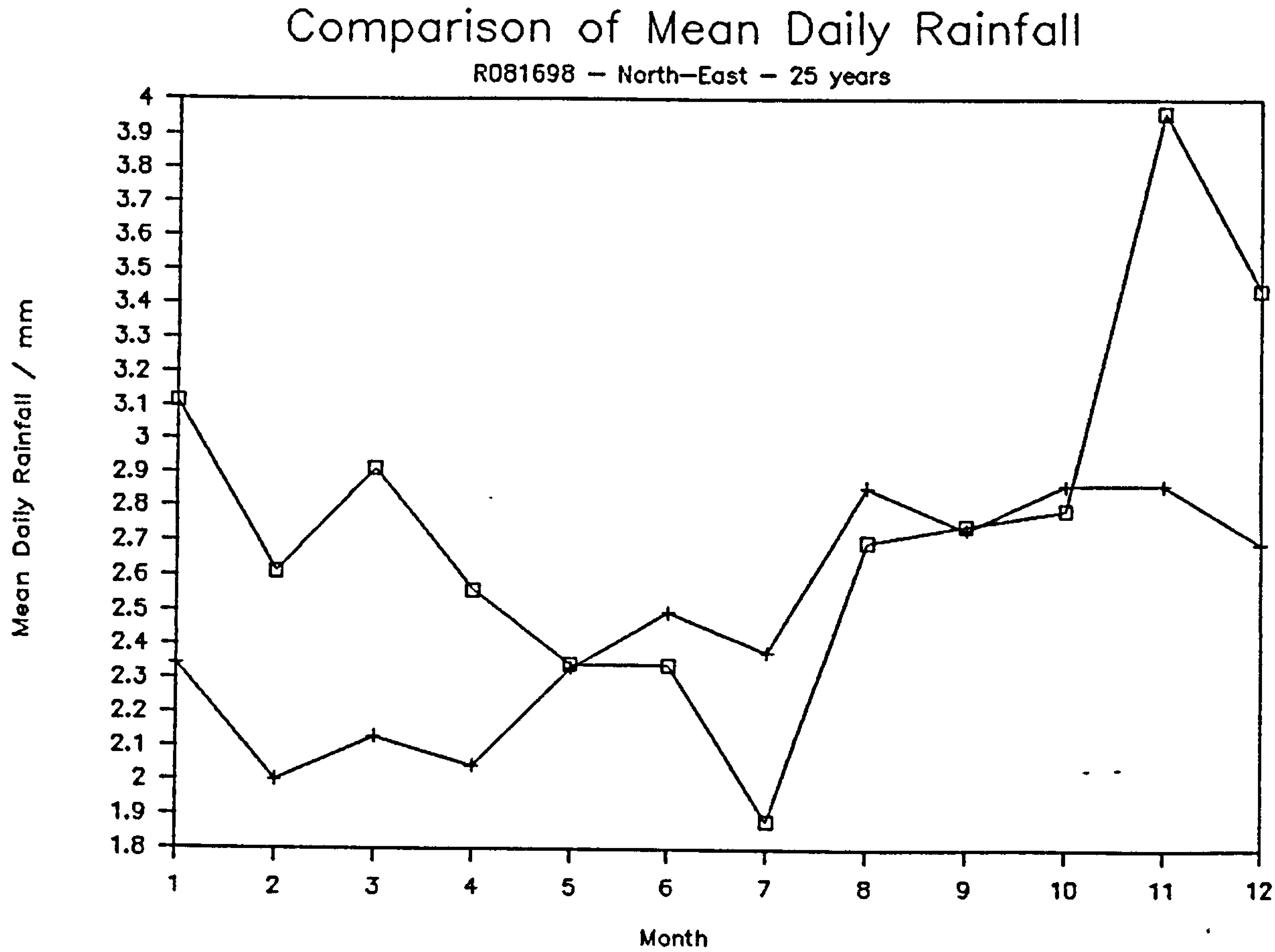
(b)

Figures 6.9(a) and (b)

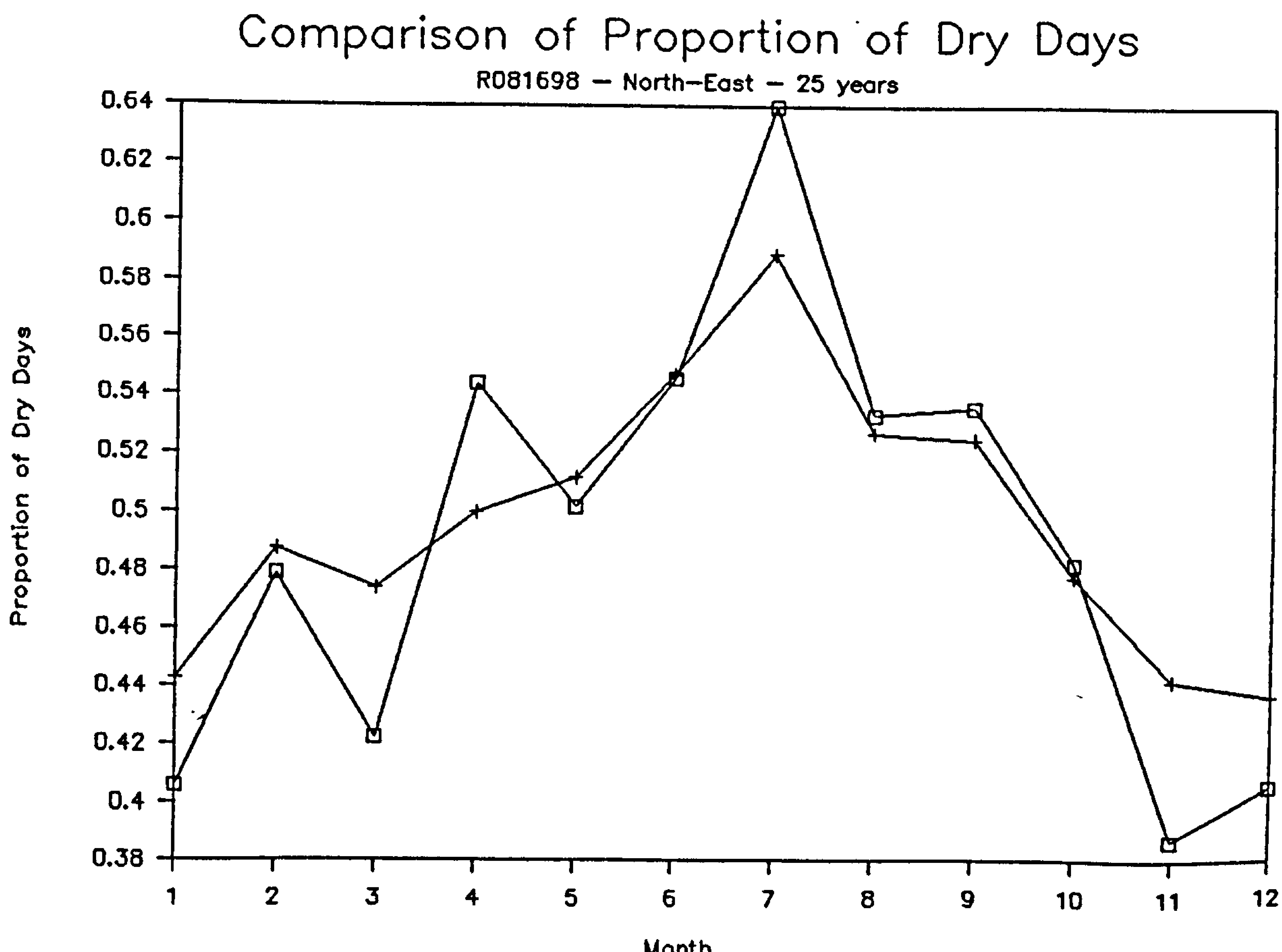




Figures 6.10(a) and (b)



(a)



(b)

□ Using parameters estimated at site

+ Using regionalised parameter estimates

Figures 6.11(a) and (b)

6.6 AN ANALYSIS OF THE RESIDUALS

Figures 6.12 - 6.16 give plots of the predicted N-S parameter value against the residual value. The plots indicate a slightly skewed distribution for the residual values. As there were so many station-months (a total of 1341) used in the analysis, it was concluded that the statistical tests were almost certainly giving valid results despite the slight skew in the residual distributions.

Some further plots of the residuals were made to see if, for example, the residuals depended on the North O/S Grid Reference. These plots are given in Appendix G, from which it could be seen that there was no obvious dependency of the residuals on location or time of year.

Lack of fit in one of the model parameters may be compensated for in another model parameter. The multivariate regression model given takes into account the correlation between the parameters of the model. The most straightforward way of testing the multivariate nature of the model was to select some key model expressions and evaluate these expressions with the actual parameter estimates based on the site data and the predicted parameter estimates given by the regression model, and then find the difference (i.e. residual = actual - predicted) in these values. The key expressions selected were the expected amount of rainfall captured in a day, denoted M24, the proportion of dry days, denoted PD24, and the variance of the amount of rain captured in a day, denoted V24. The percentage errors (100 × (actual - predicted)/actual) in these expressions were plotted against altitude, North Grid Reference, distance from coast, and month (see Appendix G). These plots indicated that there was no dependency of the percentage errors on location or time of year.

Residual Plot Against Predicted Lambda

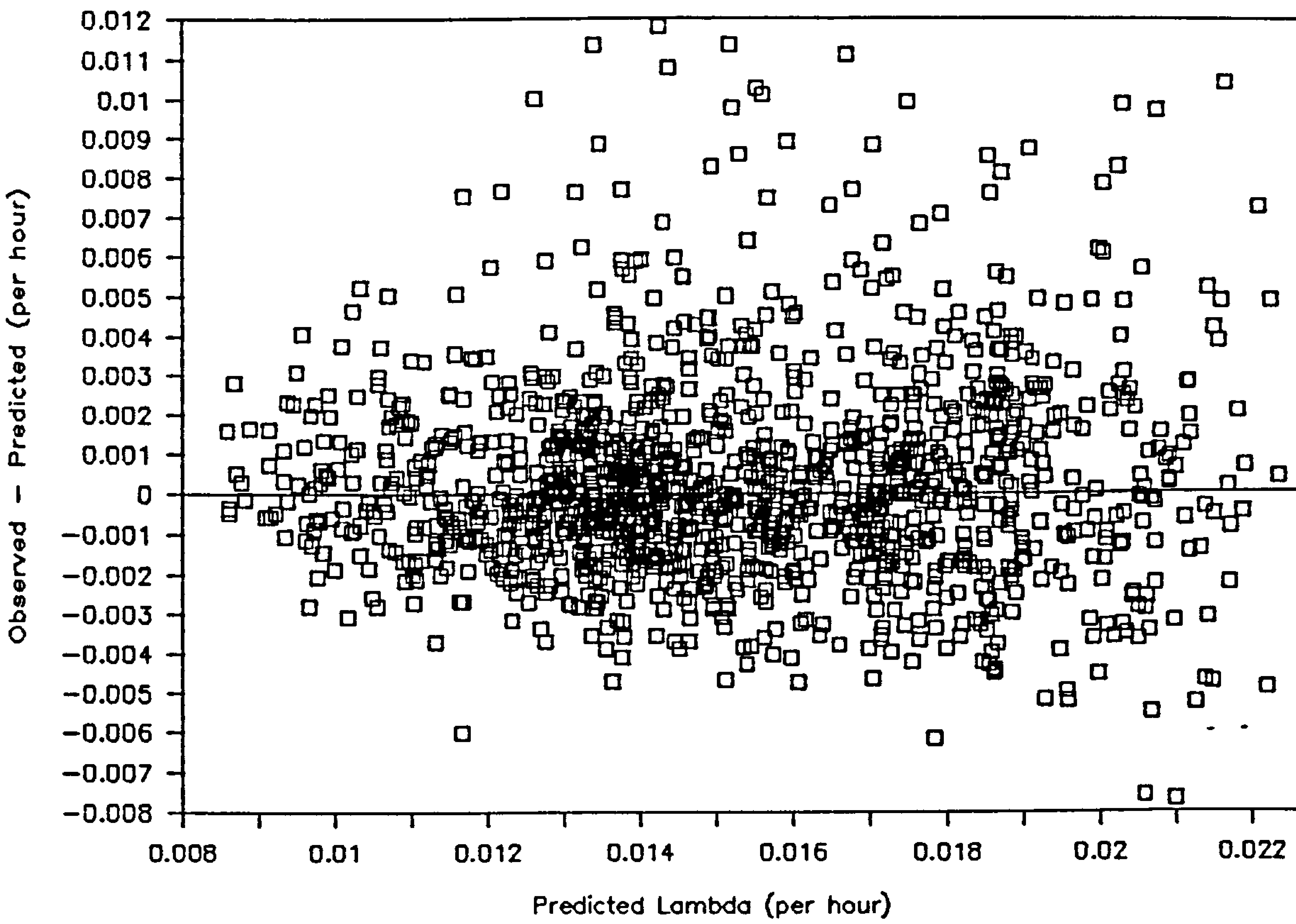


Figure 6.12

Residual Plot Against Predicted Beta

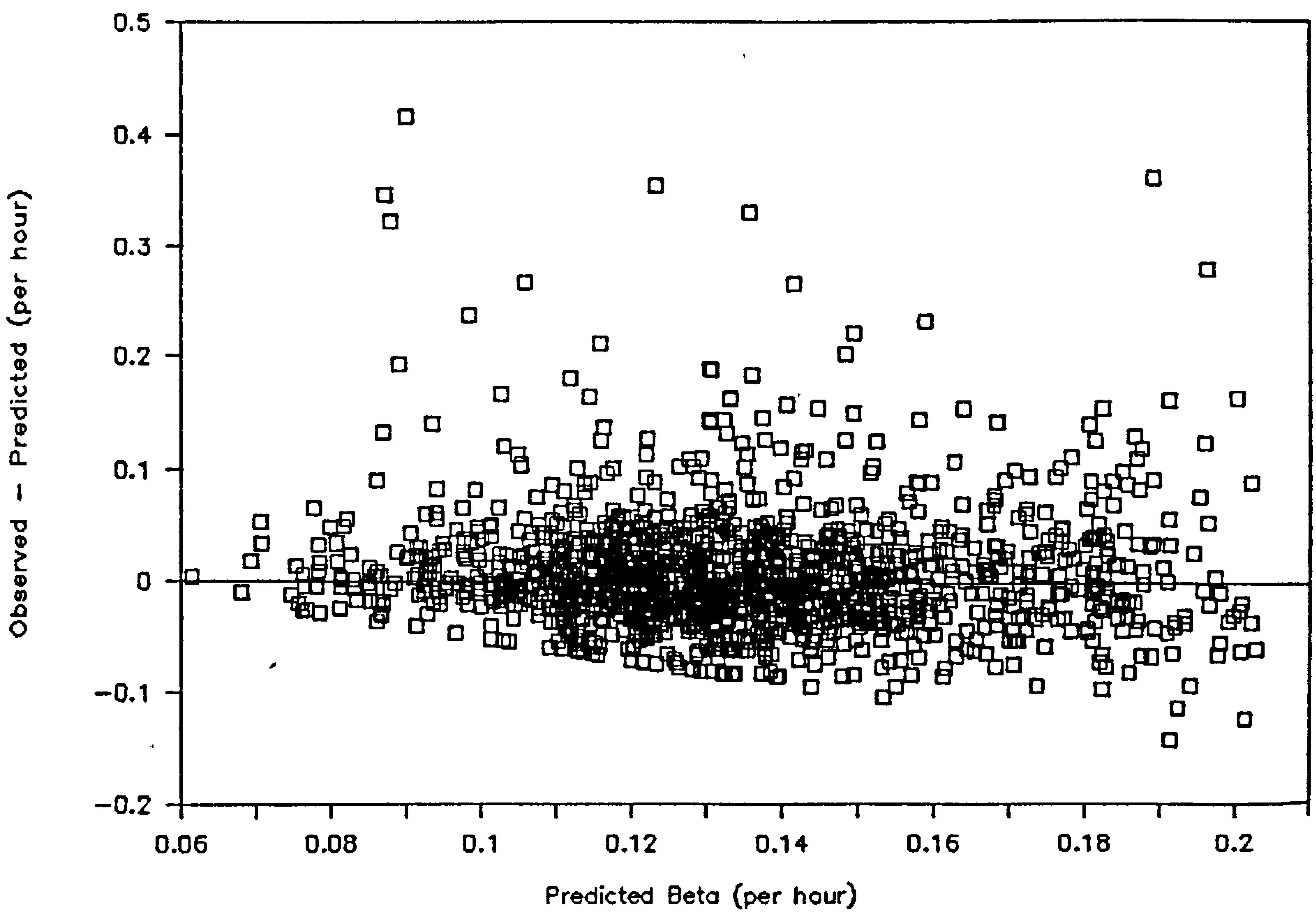


Figure 6.13

Residual Plot Against Predicted Eta

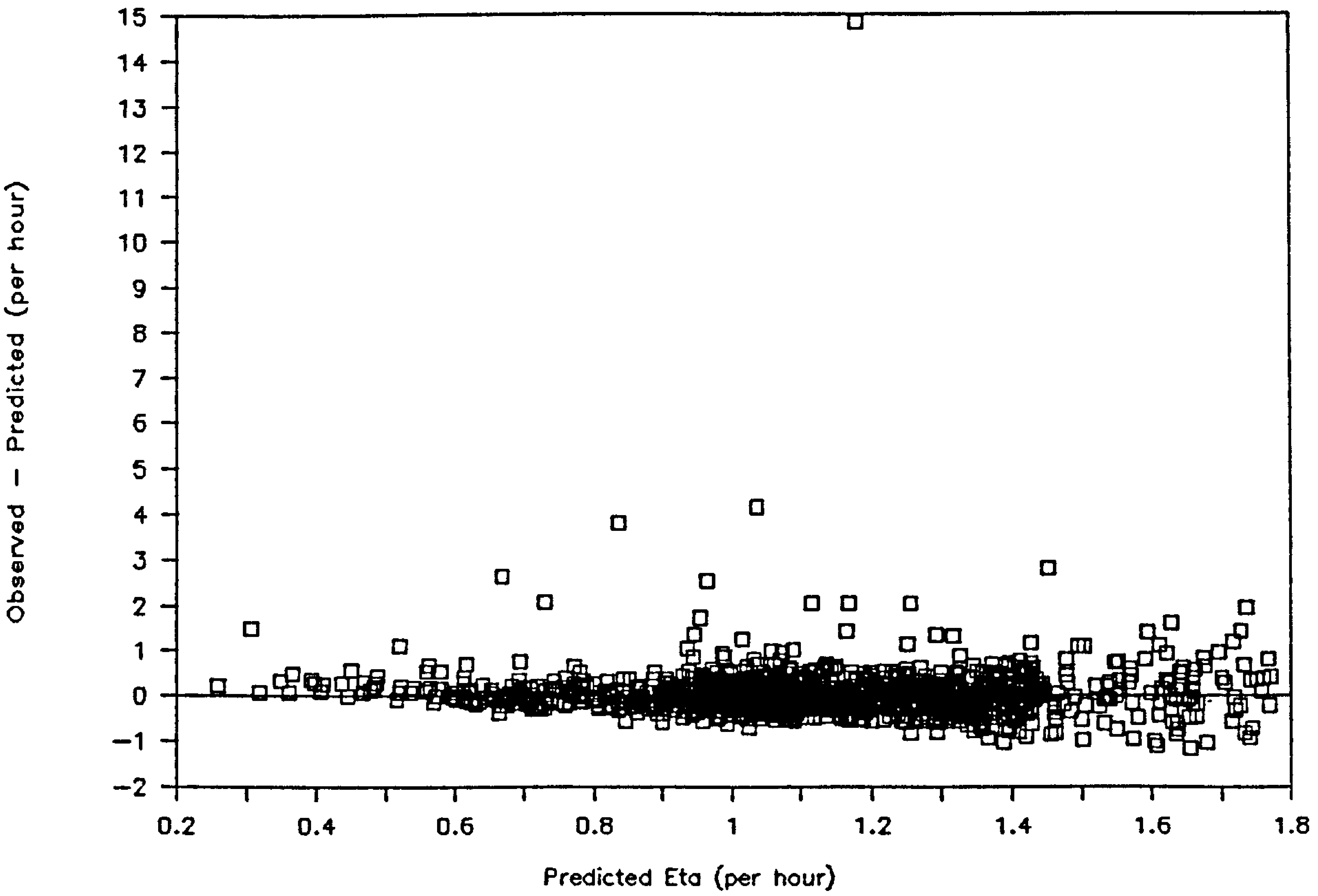


Figure 6.14(a)

Residual Plot Against Predicted Eta

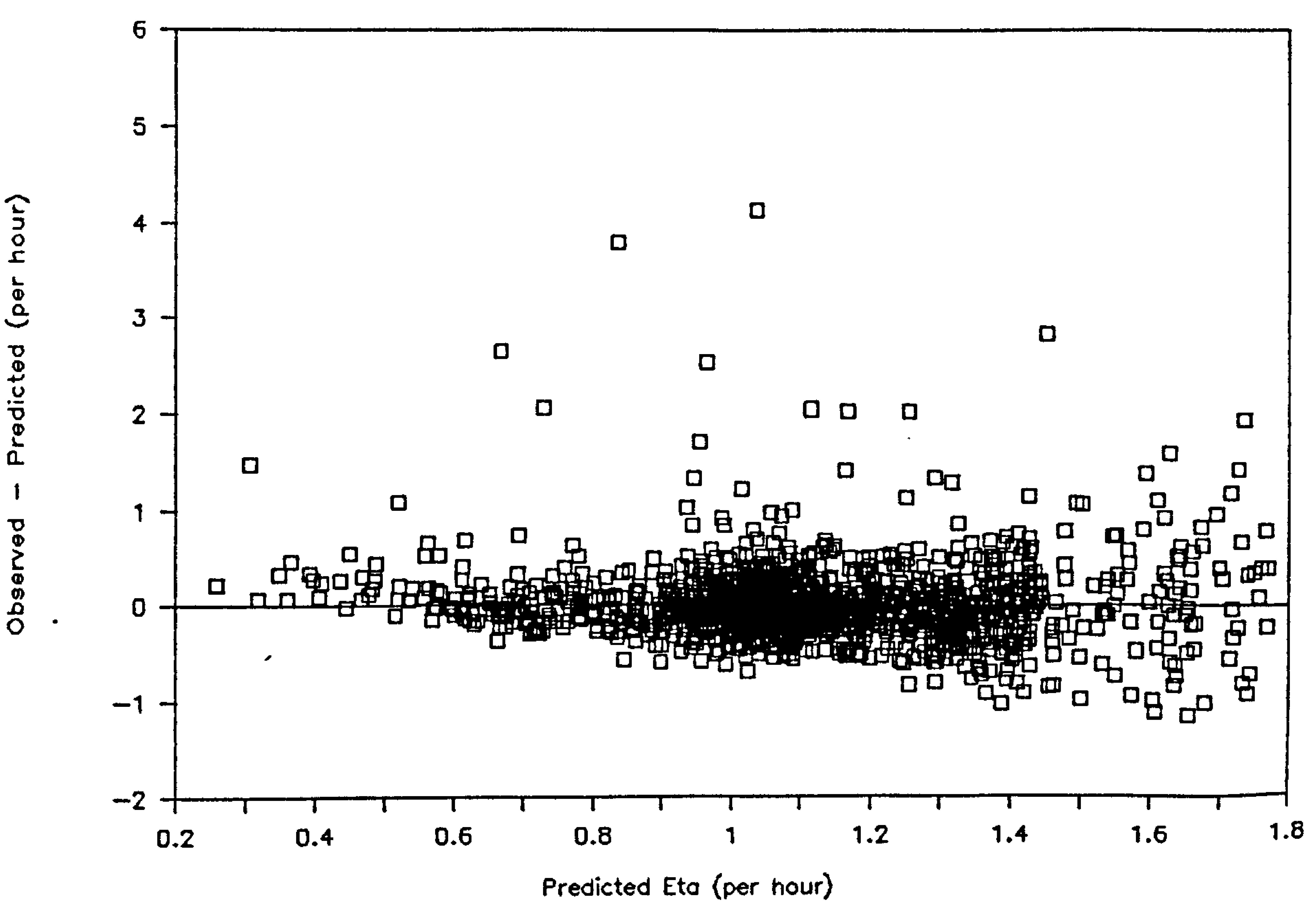


Figure 6.14(b)

Residual Plot Against Predicted Nu

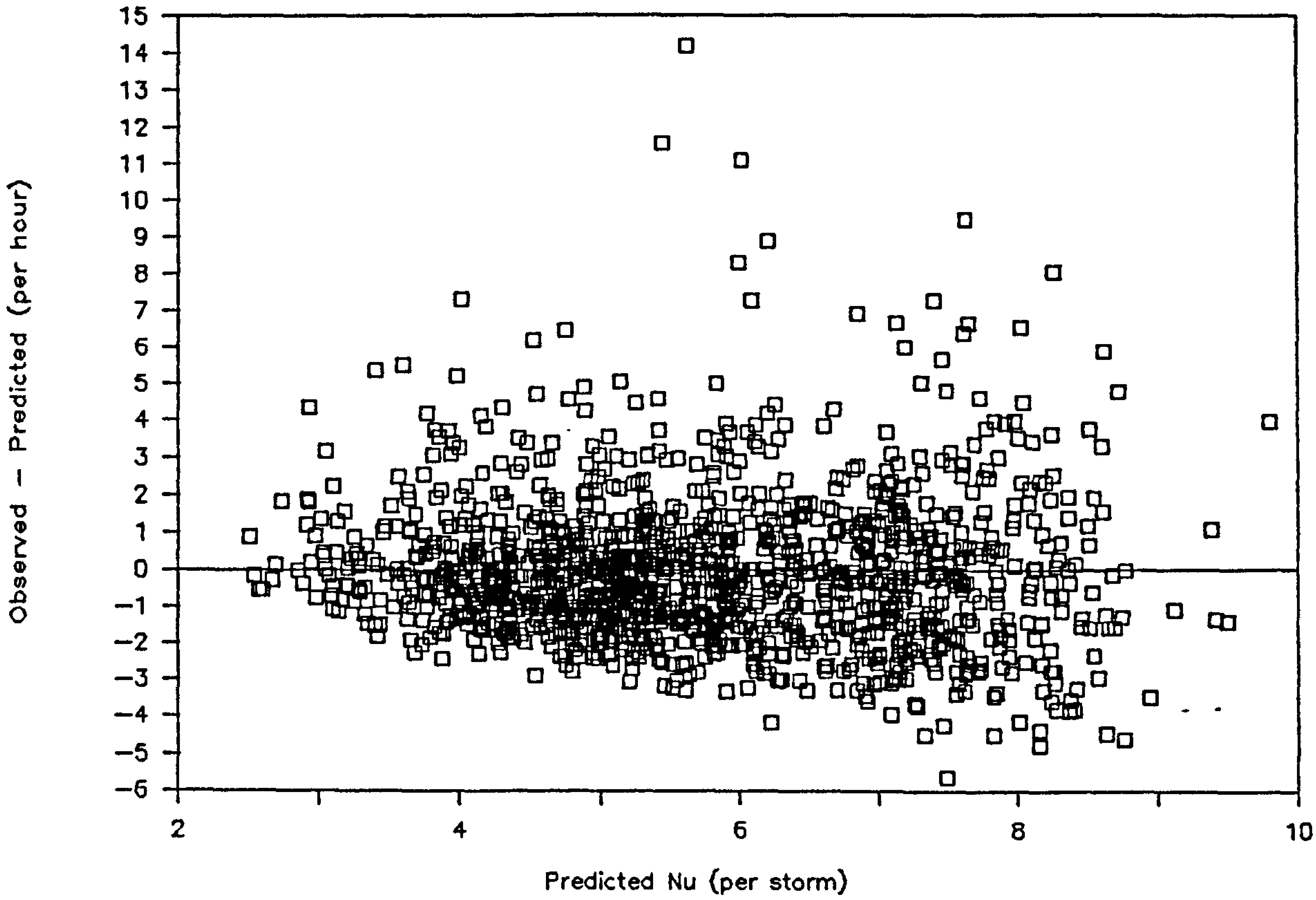


Figure 6.15

Residual Plot Against Predicted Xi

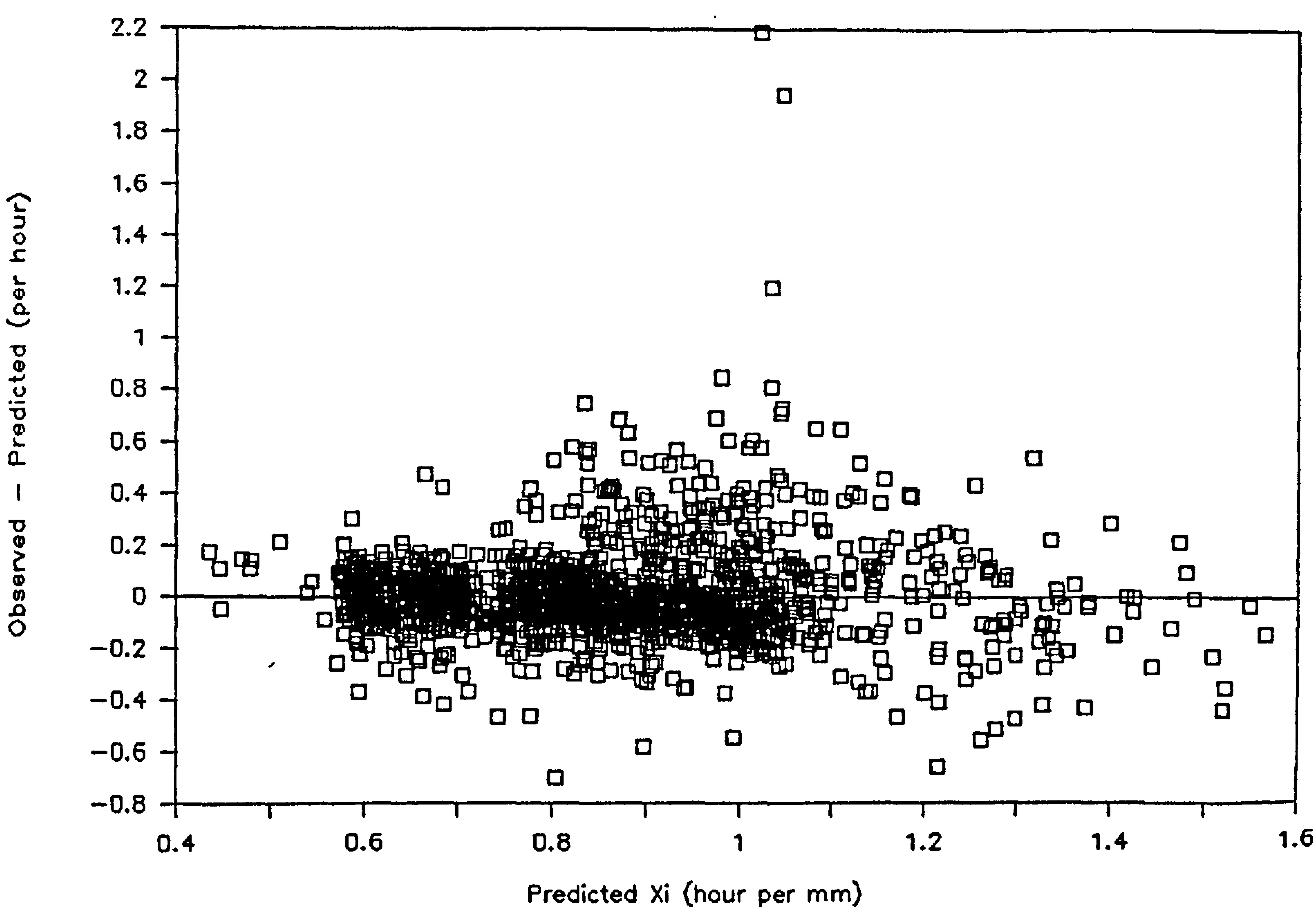


Figure 6.16

6.7 USING HISTORICAL DATA WITH THE MULTIVARIATE MODEL

At some sites rainfall data could be purchased or may already be available, in which case it would be preferable to use these data, together with the multivariate regression model, to estimate the parameters of the stochastic rainfall model. The engineer applying the stochastic model will need a guide to the accuracy of the regionalised parameter estimates compared to parameter estimates that could be obtained from historical data. This would enable an appropriate weighted parameter estimate to be evaluated using both parameter estimates given by the data and the regression model.

Suppose data for a particular month of the year are available from a long record (mn years) of historical daily rainfall. Divide these mn years, for the chosen month, into n groups consisting of that monthly record from m consecutive years. Let $\hat{f}_{ij}$ be the value of a model function, for example the proportion of dry days $\phi(24)$, for the jth year in the ith group ($j = 1, \ldots, m$; $i = 1, \ldots, n$). Such a value is obtained by first estimating the model parameters using data from the jth year in the ith group, and then substituting these parameter estimates into the model function.

Let $E(\hat{f}_{ij}) = \mu$, $Var(\hat{f}_{ij}) = \sigma^2$, and assume the $\hat{f}_{ij}$ are independent. Let $\bar{f}_i = \sum_{j=1}^{m} \hat{f}_{ij}/m$ be the mean value for the ith group, so that $Var(\bar{f}_i) = \sigma^2/m$.

Now the $Var(\bar{f}_i)$ can also be estimated from: $\sum_{i=1}^{n}(\tilde{f}_i - \bar{\tilde{f}})^2/(n-1)$, where $\tilde{f}_i$ is an estimate of the statistic using the m years of data in the ith group ($i = 1, \ldots, n$), and $\bar{\tilde{f}}$ is the sample mean of all these n estimates. Hence σ^2 can be estimated from: $m\sum_{i=1}^{n}(\tilde{f}_i - \bar{\tilde{f}})^2/(n-1)$. Note that $\tilde{f}_i$ is not exactly the same as $\bar{f}_i$, because $\tilde{f}_i$ is estimated using <u>all</u> m years of data, whereas $\bar{f}_i$ is an <u>average</u> over the m years. However, this difference is only slight and so has been ignored in the above treatment.

Four long records of daily rainfall data were selected from different regions of the country (the sites were located in

Exmouth (South-West), Poaka Beck (North-West), Howick Hall (North-East) and Windsor (South)). These records were divided into n 10-year groups (where n = 7 for Exmouth and n = 9 for the other stations). The parameters of the model were estimated for each of these 10-year groups for the months of January and July, the daily mean $\hat{\mu}(24)$ and proportion of dry days $\hat{\phi}(24)$ calculated, and $\hat{\sigma}^2$ found for both statistics.

As a separate exercise, the variances of the residuals of the regression model were estimated for both the daily mean and the proportion of dry days. Let s^2 denote this variance estimate (for either the daily mean or proportion of dry days), and let t be the number of years of data to be bought. Then, the estimate of the daily mean (or proportion of dry days) obtained using the regression model is equivalent to the estimate from t years of data if $s^2 = \hat{\sigma}^2/t$.

The t year values were found for January and July for each of the long records, and are shown in Table 6.14. On average, the accuracy in using the regression model is equivalent to using about 18 years of historical data to estimate the proportion of dry days, and equivalent to using about 5 years of historical daily data to estimate the daily mean. As monthly rainfall data are published regularly (*Meteorological Office Monthly Weather Report*), it may be advisable to use the regionalised model together with the published data. If the published monthly data are used, an adjustment in one or more of the model parameters would be needed. An obvious choice for this adjustment is ξ (=1/mean cell intensity), because the proportion of dry days and wet/dry spell lengths do not depend on this parameter. The parameter could be adjusted as follows.

Suppose the historical daily mean (obtained from the published monthly data) is denoted $\hat{m}$, and the regionalised parameter estimates (obtained from the regression model) are $\hat{\lambda}$, $\hat{\beta}$, $\hat{\eta}$, $\hat{\nu}$. Then an estimate for ξ using the daily mean is given by: $\hat{\xi} = 24\ \hat{\lambda}\ \hat{\nu}\ /\ (\hat{\eta}\ \hat{m})$. Here it has been assumed that $\hat{m}$ is obtained using data of record length much greater than 5 years (otherwise a weighted average would be preferable).

Using the above estimate for ξ for each of the long records, the t year values were found for the variances $\hat{\gamma}(1)$, $\hat{\gamma}(6)$ and $\hat{\gamma}(24)$ for both January and July (see Table 6.15). The mean and standard deviation of all the t year values in Table 6.15 were 19.4 and 10.1 respectively, which suggests that, on average, the regression model is equivalent to using about 20 years of historical daily data.

The regression model provides parameter estimates for the stochastic model that have a similar standard error to the parameter estimates obtained from fitting the model to about 20 years of daily data. The regression model was based on parameter estimates from 112 sites with typical record lengths of about 25 years. This suggests that the uncertainty in the parameter estimates from the regression model is more due to sampling variability of the rainfall data than to localised climatic anomalies (the former has a standard deviation of about twice the latter). If site data are available, it would be appropriate to combine parameter estimates made from these data with estimates from the regression model. Suppose daily site data are available for N years, then the best weighted average of the two parameter estimates is

$$w_1\hat{f}_{site} + w_2\hat{f}_{reg},$$

where $\hat{f}_{site}$, and $\hat{f}_{reg}$ are the parameter estimates based on site data and the regression model respectively, $w_1 + w_2 = 1$, and $w_2 = 20w_1/N$.

In practice, purchase of 20 years of site data could only be expected to reduce the standard deviation of the parameter estimate by a factor of about $1/\sqrt{2}$ and, unless such data are freely available or there is reason to suspect a micro-climate, it would bé more financially expedient to rely on the regression model. Even if the variation about the regression line is due to localised climatic anomalies, the weights would still be appropriate unless there is a particular reason to suspect a micro-climate for the specific site of interest.

Table 6.14

The number of years t of historical data needed to improve the estimates of the daily mean $\mu(24)$ and the proportion of dry days $\phi(24)$

Station	Month	t years	
		$\mu(24)$	$\phi(24)$
Windsor	January	2.4	24.6
	July	6.4	27.7
Poaka Beck	January	10.1	11.5
	July	3.3	1.5
Howick Hall	January	5.9	23.2
	July	2.3	13.9
Exmouth	January	7.0	12.9
	July	5.2	25.7
	Mean	5.3	17.6
	SD	2.7	9.1

Table 6.15

The number of years t of historical data needed to improve the estimates of the variances $\gamma(24)$, $\gamma(6)$, and $\gamma(1)$, and the proportion of dry days $\phi(24)$ using mean monthly totals to estimate ξ

Station	Month	t years			
		$\phi(24)$	$\gamma(24)$	$\gamma(6)$	$\gamma(1)$
Windsor	January	24.6	35.6	24.0	21.3
	July	27.7	14.4	11.6	7.7
Poaka Beck	January	11.5	38.9	27.1	20.3
	July	1.5	20.8	18.8	11.8
Howick Hall	January	23.2	11.9	10.7	6.9
	July	13.9	26.5	24.3	15.1
Exmouth	January	12.9	15.1	8.7	7.9
	July	25.7	40.0	41.3	20.1
	Mean	17.6	25.4	20.8	13.9
	SD	9.1	11.5	10.8	6.1

Overall mean = 19.4, overall SD = 10.1.

6.8 SOME FURTHER TESTS FOR THE REGIONAL MODEL

Some further tests were made by re-estimating the regression parameters with five records of hourly rainfall data excluded from the data matrix (so the total number of observations used was 1341 - 5×12 = 1281). The purpose here was to see how well the model fitted sites that do not have data (this was also discussed on p226, where 2 sites were randomly selected from each 'Wigley' region). The sites selected were: Farnborough, Manston, Rhoose, Ringway and Turnhouse. These sites were chosen because they had the longest records for different regions of the country. Using the re-estimated regression parameters the Neyman-Scott model parameters λ, β, ν, and η were estimated for these sites. The Neyman-Scott model parameter ξ was estimated using the estimates for λ, β, ν, η and the mean rainfall for each month, which was obtained using the site data (the mean rainfall for each month could also have been obtained from the Meteorological Office, Bracknell, UK, or by calculation using their *Monthly Weather Report* or *Rainfall* (which contains monthly totals for a large number of sites scattered throughout the UK)). Comparisons were then made between historical statistics at the sites and their equivalent statistic predicted using the regression estimates (together with the Neyman-Scott model functions). The following statistics were selected for the comparisons, which were made on a monthly basis: (i) the hourly variance, (ii) the daily variance, (iii) the proportion of dry days, and (iv) the dry given dry transition probability. The comparisons were made by plotting the historical and predicted values against the month. For convenience, these plots are presented in Appendix G with the other Figures for this Chapter (see Figures G.42 - G.61).

Looking through Figures G.42 - G.61 it can be seen that the regionalised model follows the historical statistics reasonably well. In particular, it can be seen that the regional model follows the historical hourly variances quite closely for most station-months, which provides some additional support for regressing the hourly variance on the daily variance (Chapter 5) as most of the stations (about 80%) in the regionalisation procedure had daily records and therefore used this regression. However, some differences are apparent (e.g. the variances for Manston-September - Figure G.47), which are unlikely to be due to the sampling variability of the rainfall data (e.g. the percentage difference between the predicted and historical daily variance for Manston-September was about 60% and the coefficient of variation for the daily variance of a 20 year record is about 22% (Table 4.1 of Chapter 4), so that such a difference is unlikely to occur by chance. The question remains as to whether such differences will have any practical consequences from the engineering sewer design viewpoint (this issue is addressed in Chapter 8, Section 8.3, where some recommendations are described so that the model can be further validated for its intended application).

CHAPTER 7
A DISAGGREGATION MODEL FOR UK HOURLY DATA

7.1 INTRODUCTION

The stochastic rainfall model generates raincells that have a rectangular shape (because the intensity of the rain is constant throughout the cell duration), which is physically unrealistic. To overcome this and thereby introduce more realistic storm profiles with greater 'within cell' variability, a disaggregation model for hourly rainfall time series is proposed and developed. The proposed disaggregation model is similar to that used by *Ormsbee* (1989), with the exception that the depth of rain per pulse burst is a parameter of the model. The model is fitted to 28 years of minute data taken from Farnborough, UK, and tested on 16 years of data from Rhoose, UK. The model performs well, and so can be used with some confidence at other locations.

7.2 DEFINITION OF DISAGGREGATION MODEL

The reader is referred to Chapter 2 for a non-mathematical review of Ormsbee's paper. In this Section, Ormsbee's 'Continuous Stochastic Disaggregation Model' will be reviewed.

Rainfall is assumed to occur in discrete pulses of amount δ. In Ormsbee's paper δ is taken to be 0.01 inches. The disaggregation model that we shall use, will be identical to that used by Ormsbee, with the exception that δ will be a parameter of the model.

Consider a sequence of 3 hours. Let x = depth of rain in 1st hour, y = depth in central hour, and z = depth in 3rd hour. The purpose of the model is to distribute the central depth of rain, over the hour, based on a knowledge of x, y, and z. Let $f(t)$ be the probability density function (pdf) for the time of occurrence of a single rainfall pulse in the central hour. The pdf is assumed to be composed of two line segments as shown in Figure 7.1. The pdf represented by these lines may be written:

$$f(t) = k^{-1}(x - (x - y)t/t^*) \qquad 0 \le t < t^* \qquad (7.1a)$$

$$f(t) = k^{-1}(y - (y - z)(t - t^*)/(60 - t^*)) \qquad t^* \le t \le 60 \qquad (7.1b)$$

(*Ormsbee* (1989) equation 16)

where t^* is the point where the line segments meet (see Figure 7.1), and k is a constant to be determined by integration. Using

$$\int_0^{60} f(t)\, dt = 1 \text{ gives:}$$

$$k = 30(y + z) - t^*(z - x)/2 \qquad (7.2)$$

The distribution function follows as:

$$F(t) = xt/k - (x - y)t^2/(2kt^*) \qquad [0 \le t < t^*] \qquad (7.3a)$$

$$F(t) = (y + x)t^*/(2k) + y(t - t^*)/k - (y - z)(t - t^*)^2/(2k(60 - t^*)) \qquad [t^* \le t \le 60] \qquad (7.3b)$$

Therefore, the pdf for a particular hour is dependent only on the rainfall depths in the 3 hour sequence and the time parameter t^*. Application of equation (7.1), to each of four different types of hourly rainfall sequences, produces a variety of pdfs as shown in Figure 7.2. The parameter t^* may also be expressed in terms of the hourly depths for each type of sequence (see Figure 7.2) using:

Type 1:

$$t^* = 60(x - y)/(x - z) \tag{7.4}$$

Type 2:

$$t^* = 60(y - x)/(z - x) \tag{7.5}$$

Type 3:

$$t^* = 60(x - y)/(x + z - 2y) \tag{7.6}$$

Type 4:

$$t^* = 60(y - x)/(2y - x - z) \tag{7.7}$$

(*Ormsbee* (1989) equations (19)-(23)).

Ormsbee developed equations (7.4)-(7.7) empirically after an examination of the mean distribution functions for the four different types of sequences for several historical records. Using equations (7.3)-(7.7) allows the distribution function to change from hour to hour, which is likely to reflect the changing dependency of rainfall within different hourly sequences.

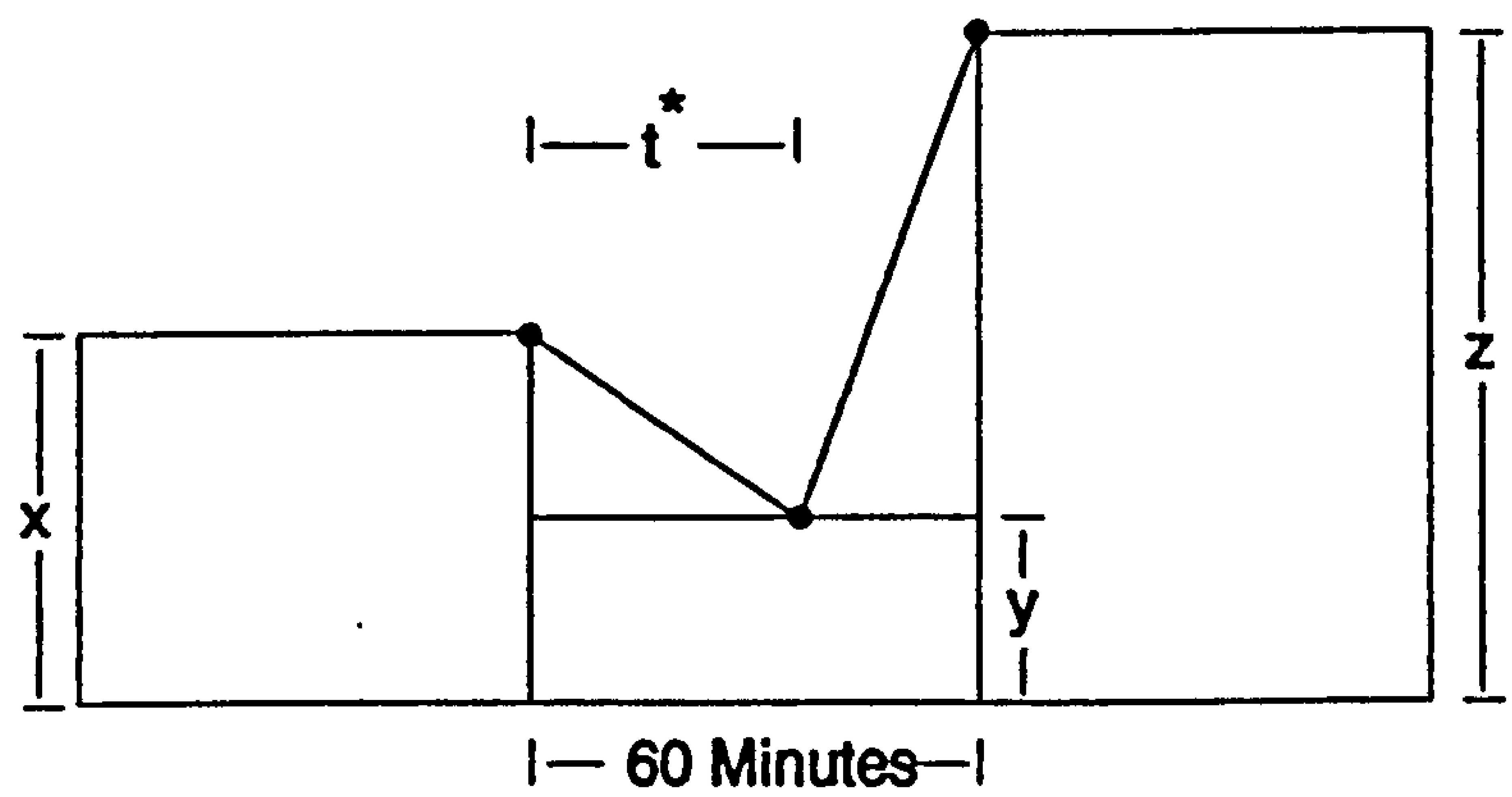


Figure 7.1
Example shape of the pdf for the disaggregation model

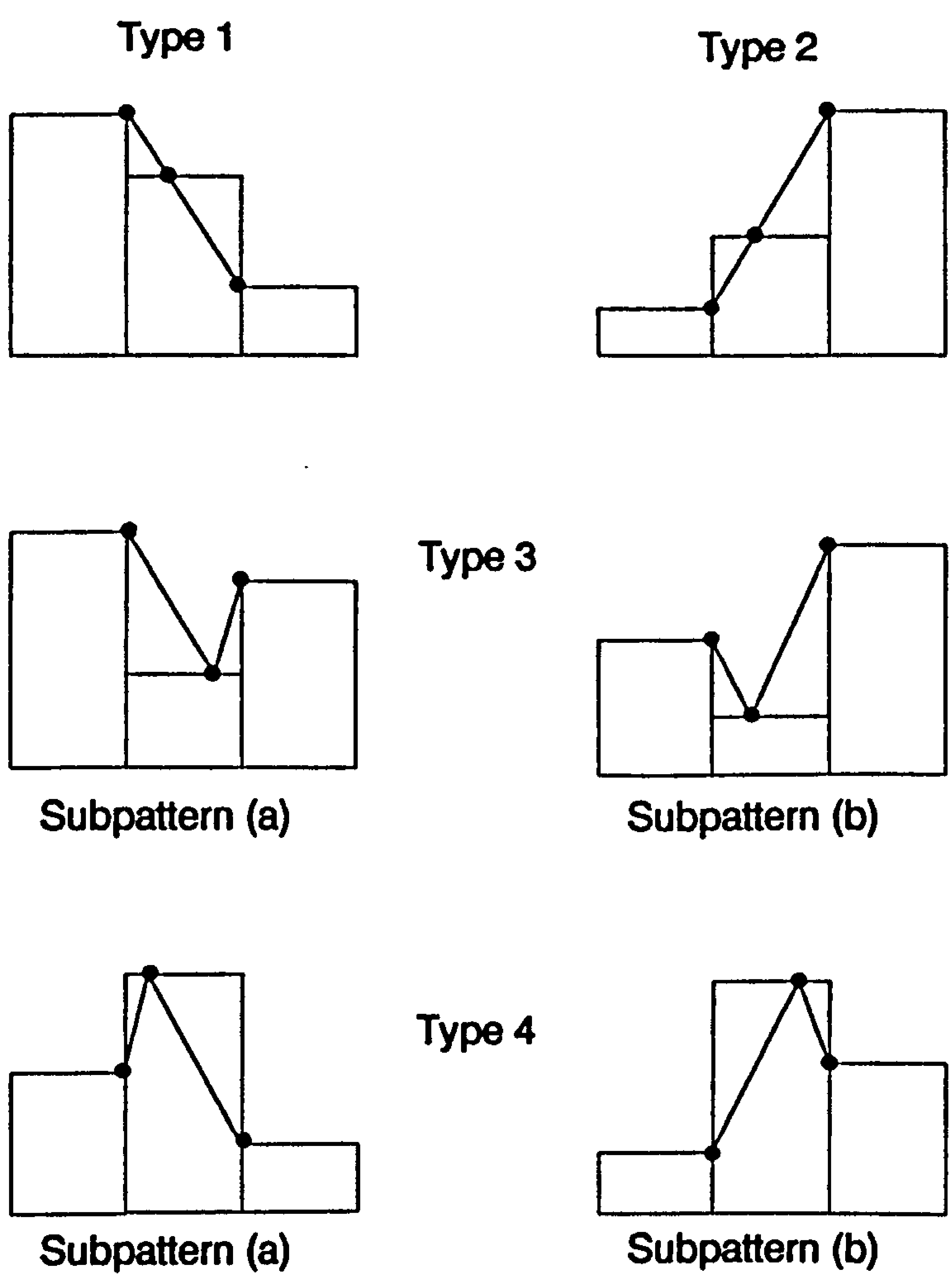


Figure 7.2
Example pdf shapes for different rainfall sequences

The steps involved in the disaggregation process can be summarised as follows (assuming δ has been estimated):

i) Determine the type of sequence (as represented in Figure 7.2).

ii) Disaggregate the central hour into T time intervals of m minutes, where T = 60/m, and m = desired disaggregation time interval (e.g. 1 or 5 minutes).

iii) Determine the probability associated with ith time interval (where i = 1, ..., T) using equations (7.3) - (7.7), and the following relationship:

$$pr\{\text{pulse in ith interval}\} = F(im) - F((i-1)m) \tag{7.8}$$

iv) Disaggregate the total depth of rain (y) in the central hour into N pulses of δ mm, where N = y/δ.

v) Assign each rainfall pulse (one at a time) to one of the T time intervals using the probabilities developed in step (iii) and a sequence of N Uniform (0,1) random numbers.

vi) Continue this process until all hours have been disaggregated.

7.3 FITTING THE DISAGGREGATION MODEL TO THE FARNBOROUGH DATA SET

In the previous Chapters it was found that the variances of the h hourly time series (h = 1, 3, 6, 12, and 24) were needed in the fitting procedure. It therefore seems reasonable to fit the disaggregation model using the variance of the 5 minutely time series. As no expression is available for the 5 minutely variance in terms of δ, the optimum value of δ has to be found iteratively. The method of estimating δ for any given month was as follows:

a) The historical 5 minutely variances were estimated for each year, and the mean of these estimates used to estimate the 'overall' historical 5 minutely variance.

b) The minute data were aggregated to obtain an hourly historical time series.

c) The hourly time series was disaggregated into fourteen 5 minutely time series using 14 different values for δ (0.01, 0.05, 0.1, 0.2, ..., 1.0, 1.5, 2.0 mm) and the mean model 5 minutely variances found for each of these series.

d) The model 5 minutely variances for each of these series were compared to the overall historical 5 minutely variance obtained in (a).

e) The two model 5 minutely variances that were closest to the historical 5 minutely variance were selected. Using these two variances and their corresponding δ values, an estimate $\hat{\delta}$ of optimum pulse depth was obtained by linear interpolation.

Table 7.1 gives the model 5 minutely variances and δ values for all months of the Farnborough data set, and Table 7.2 gives the $\hat{\delta}$ estimates obtained by linear interpolation. These estimates are plotted for each month in Figure 7.3, together with a fitted curve through the points. The values of δ obtained from the fitted curve (δ_f) were preferred to the point estimates $\hat{\delta}$ because of the large sampling variability of some of the point estimates (e.g. the historical 5 minutely variances for July have a standard error of the mean of 0.0015). The fitted curve was obtained by regressing the point estimates on the first two harmonics, i.e.

$$\delta_f = \alpha_0 + \alpha_1\cos(2\pi t/12) + \alpha_2\sin(2\pi t/12) + \alpha_3\cos(4\pi t/12) + \alpha_4\sin(4\pi t/12)$$

where t = 1, 2, ..., 12 (1 ≡ Jan, 2 ≡ Feb, etc), δ_f = fitted value, and the α_i are found by the method of least squares. The fitted δ_f are given in Table 7.2.

Using the fitted values δ_f, the historical 5 minutely variances were compared to the model variances (which had to be interpolated - the values are given in Table 7.2), and plotted in Figure 7.4. As expected, the model follows the historical 5 minutely variances within sampling variability.

The heaviest storms in January and July of the first 3 years in the Farnborough data set were selected to determine whether the disaggregation model gave a realistic storm profile (some time series plots are given in Appendix I). It could be seen that the disaggregated storm profiles tended to show a greater amount of 'within storm' variability than the original storm profiles (see Appendix I). In order to 'smooth' the disaggregated 5 minutely

Table 7.1

5 Minutely Variances for each Disaggregation Model

δ/mm	Month 1	2	3	4	5	6
0.01	0.00095	0.00078	0.00067	0.00083	0.00127	0.00114
0.05	0.00120	0.00099	0.00085	0.00103	0.00149	0.00134
0.1	0.00151	0.00126	0.00109	0.00128	0.00178	0.00157
0.2	0.00216	0.00181	0.00156	0.00182	0.00238	0.00209
0.3	0.00282	0.00235	0.00202	0.00228	0.00300	0.00260
0.4	0.00341	0.00289	0.00256	0.00285	0.00352	0.00306
0.5	0.00409	0.00345	0.00299	0.00329	0.00409	0.00360
0.6	0.00456	0.00394	0.00336	0.00369	0.00468	0.00405

Cont.

δ/mm	Month 7	8	9	10	11	12
0.01	0.00280	0.00219	0.00174	0.00191	0.00176	0.00121
0.05	0.00302	0.00245	0.00199	0.00218	0.00207	0.00149
0.1	0.00330	0.00278	0.00233	0.00252	0.00250	0.00182
0.2	0.00400	0.00347	0.00299	0.00317	0.00338	0.00252
0.3	0.00448	0.00421	0.00369	0.00383	0.00423	0.00328
0.4	0.00514	0.00491	0.00438	0.00461	0.00514	0.00396
0.5	0.00582	0.00561	0.00500	0.00529	0.00590	0.00460
0.6	0.00647	0.00630	0.00583	0.00601	0.00670	0.00534

Table 7.2

Interpolated and fitted δ

Month	Historical variance (sq mm)	Interpolated $\hat{\delta}$ (mm)	fitted δ_f (mm)	Model Variance* (sq mm)
1	0.00131	0.067741	0.071746	0.001334
2	0.00098	0.048095	0.060135	0.001044
3	0.00098	0.077083	0.057787	0.000887
4	0.00132	0.107407	0.096279	0.001261
5	0.00227	0.181666	0.185377	0.002292
6	0.00214	0.209803	0.290339	0.002550
7	0.00569	0.480882	0.352093	0.004823
8	0.00409	0.283783	0.334013	0.004448
9	0.00295	0.193939	0.251811	0.003352
10	0.00333	0.224242	0.158461	0.002900
11	0.00240	0.088372	0.099055	0.002491
12	0.00164	0.072727	0.078643	0.001679

* Interpolated

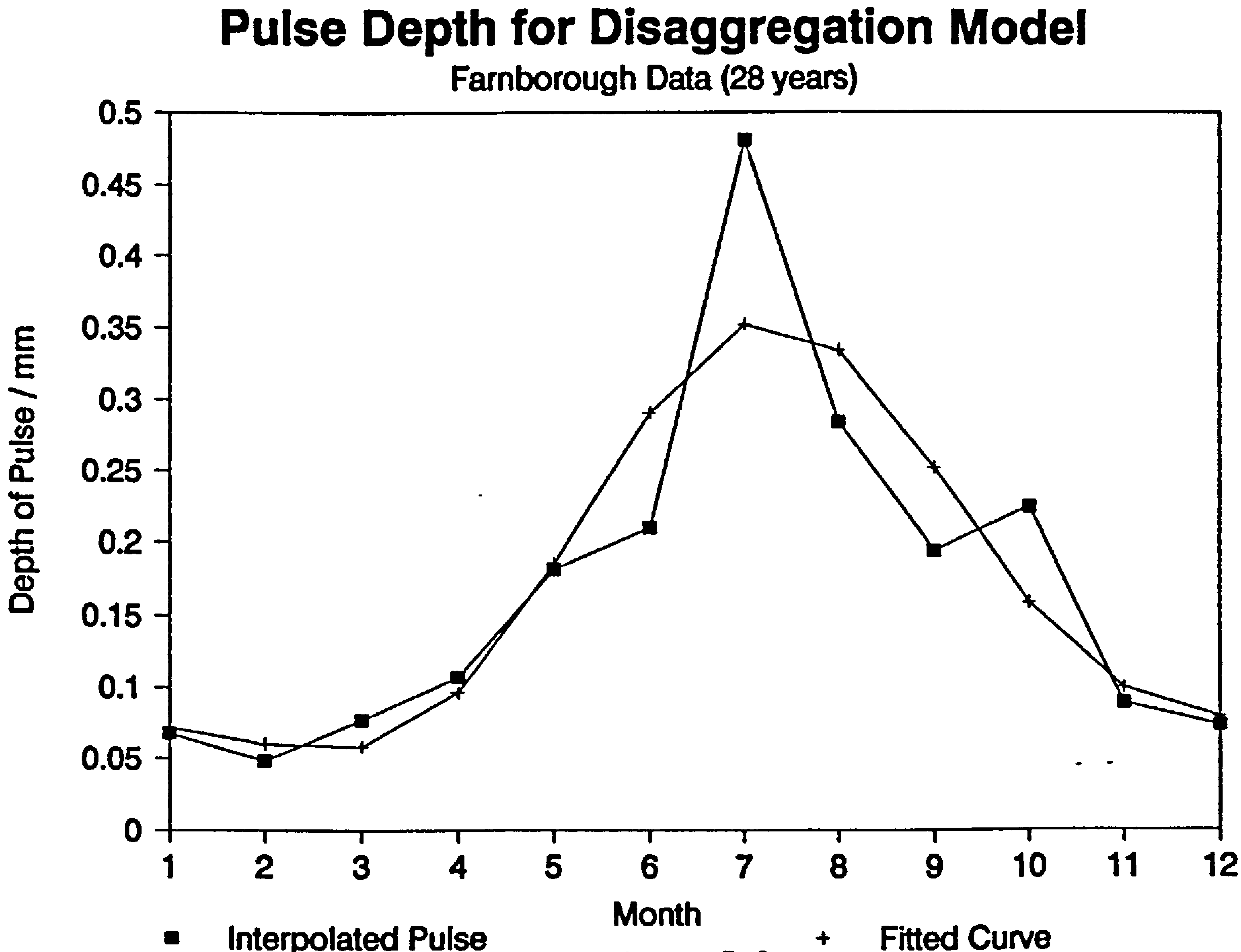


Figure 7.3

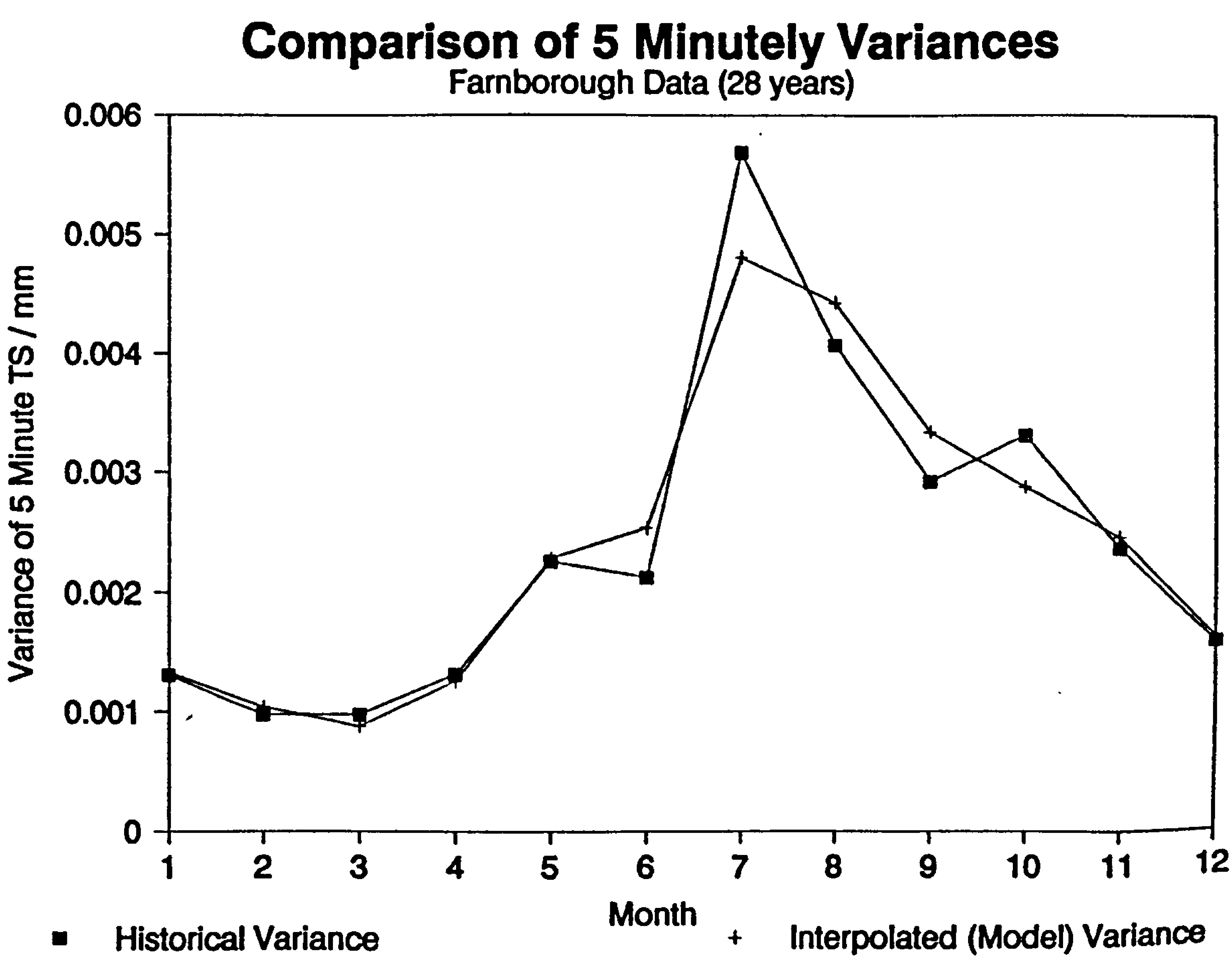


Figure 7.4

time series, a moving average of the disaggregated series could be used. To avoid over 'smoothing' the series a moving average process of order 1 was used, so that if the disaggregated time series is denoted by $\{z_i\}$, then the 'smoothed' disaggregated series is given by $\{z_i'\}$, where $z_i' = (z_{i-1} + z_i + z_{i+1})/3$. Some time series plots for this series are also given in Appendix I, from which it can be seen that the 'smooth' disaggregated series gives a more realistic storm profile.

7.4 TESTING THE MODEL ON THE RHOOSE DATA SET

Only a few sites available for the project had minute data. In order to test the disaggregation model, it is desirable to use a data set from another region. Only two sites were available in a region outside the Southern 'Wigley' Region: Rhoose (16 years) and St Mawgan (10 years) - both located in the South-West. The Rhoose data set was selected because it was a longer record.

The historical 5 minutely variances were found for each year and each calender month for the Rhoose data set, and the mean and standard deviation of these variances found for each month (see Table 7.3). The minute data were then aggregated to form an hourly rainfall time series. Using the fitted δ_f (in Table 7.2), the hourly time series were disaggregated, and the model 5 minutely variances found for each month and each year. The mean and standard deviations for these variances were also found (see Table 7.3) and plotted with the historical values obtained from the original 5 minutely series (see Figures 7.5 and 7.6). From the Figures it is evident that the disaggregation model is following the historical mean and standard deviations of the 5 minutely variances reasonably well.

Standard t- and F-tests were performed to see if the values were significantly different (see Table 7.3). Three of the tests showed values that were significant at the 5% level (e.g. F-test for December, t-test for September). However, the model seems to be following the seasonal trends in the data (Figures 7.5 and 7.6), and so these significant results did not cause concern. Perhaps a more complex disaggregation model, incorporating a regional structure, could be developed and tested. However, with the small number of sites with minute data available, this was not possible. Even if the data were available, and a regional disaggregation model developed, it is unlikely that such a model would improve the results sufficiently to make a practical difference (from an urban drainage viewpoint).

Table 7.3
Statistics for 5 minutely variances
for the Rhoose data set

Month	Historical Mean, sq mm	Model Mean, sq mm	Historical SD, sq mm	Model SD, sq mm	F-ratio	t-ratio
1	0.002006	0.001819	0.001275	0.001106	1.328831	-0.62605
2	0.00129	0.001175	0.001166	0.001029	1.282392	-0.41819
3	0.001075	0.001026	0.000594	0.000589	1.014981	-0.32928
4	0.001356	0.001492	0.000884	0.000741	1.421087	0.66764
5	0.002596	0.002834	0.004092	0.002957	1.914801	0.26676
6	0.003042	0.003821	0.002529	0.002371	1.137436	1.27144
7	0.005106	0.005671	0.006100	0.004121	2.191103	0.43366
8	0.004831	0.005905	0.004500	0.004376	1.057381	0.96710
9	0.004091	0.002777	0.002523	0.001791	1.984640	-2.40118
10	0.004236	0.003915	0.003398	0.002787	1.485864	-0.41425
11	0.003216	0.002840	0.002353	0.001473	2.552268	-0.76514
12	0.003019	0.002541	0.002140	0.001203	3.160639	-1.10003

$t_{5\%} = 2.1$, $F_{5\%} = 2.4$.

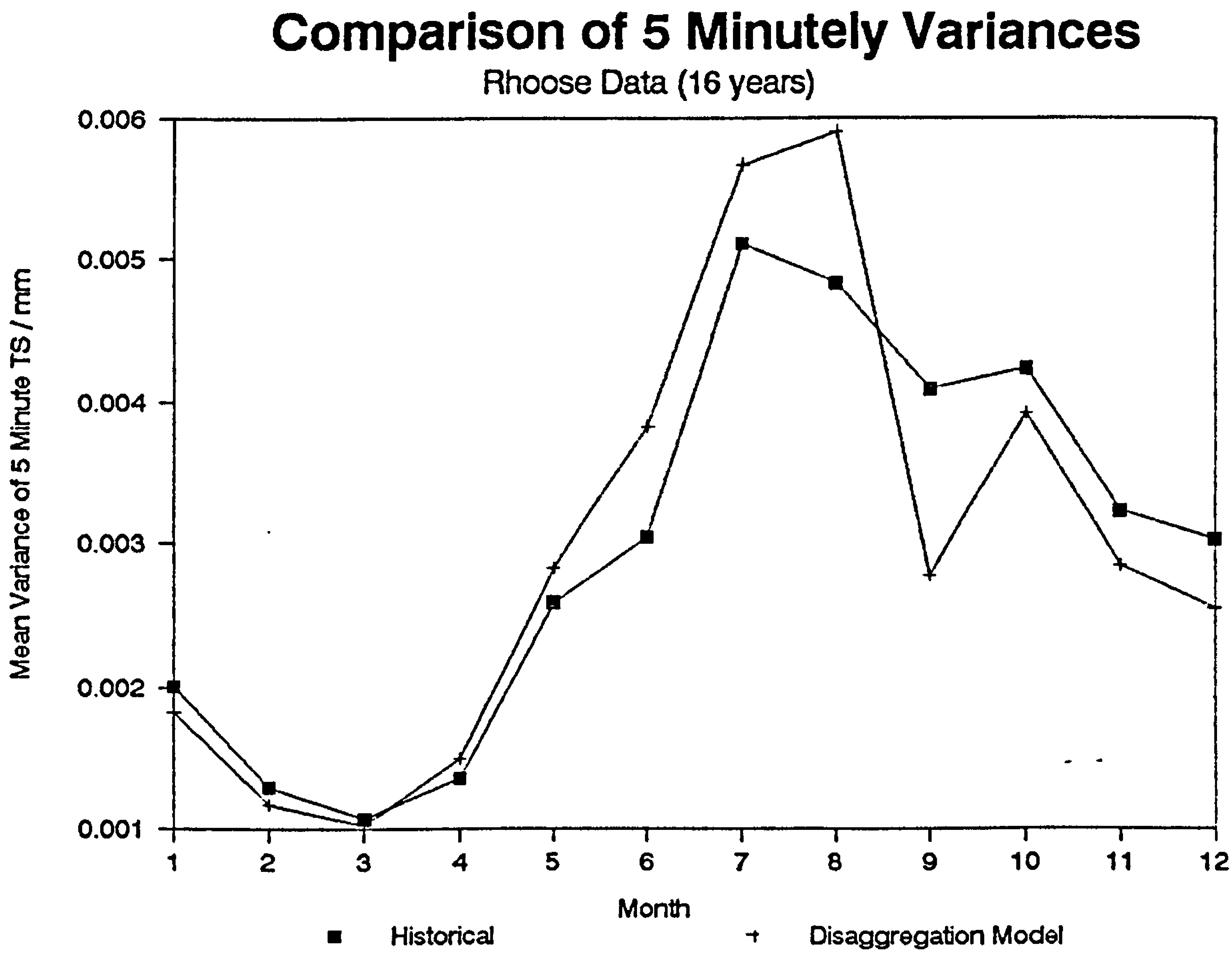


Figure 7.5

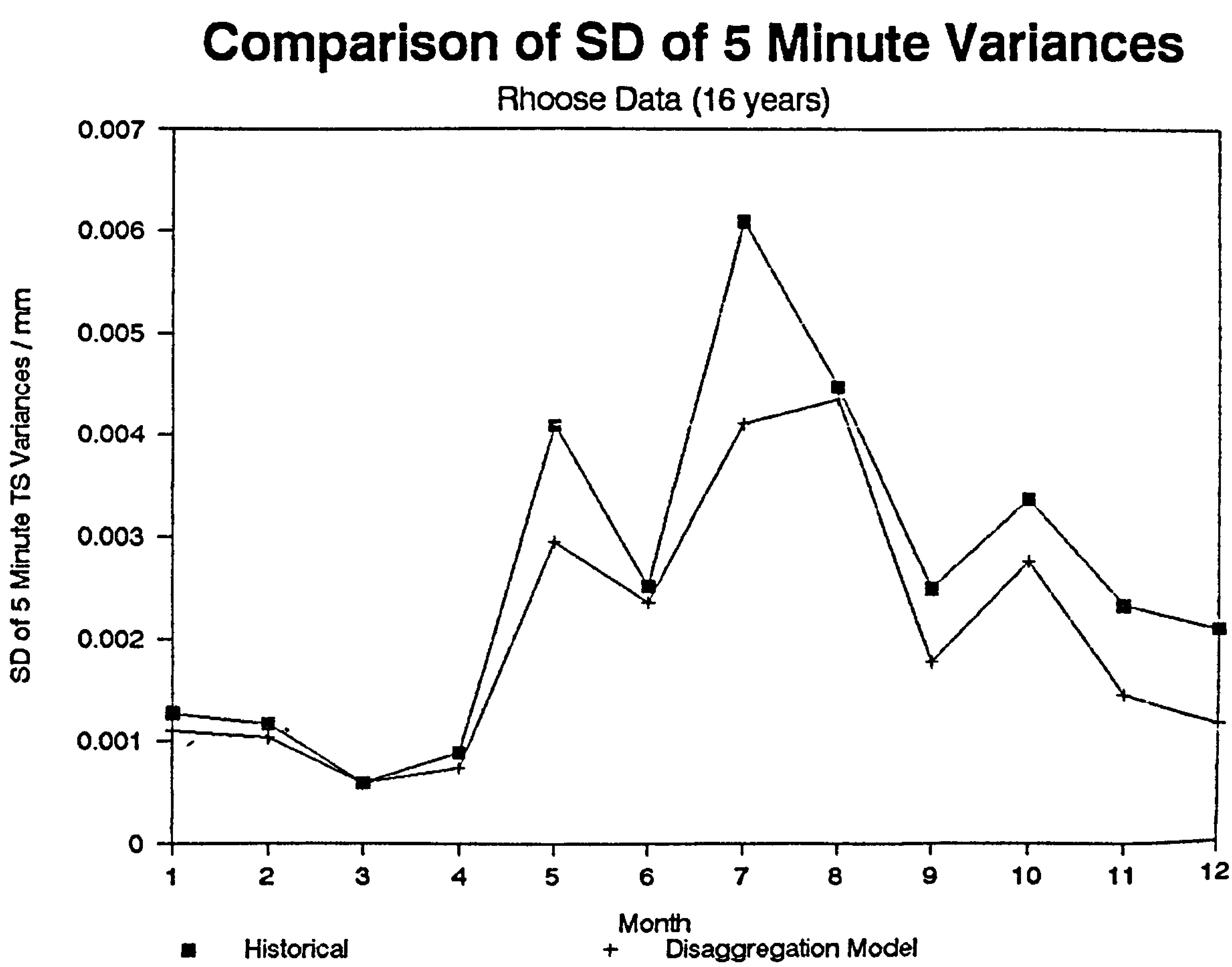


Figure 7.6

CHAPTER 8
SUMMARY AND CONCLUSIONS

8.1 SUMMARY

The purpose of this project was to propose and validate a regionalised stochastic rainfall time series model for the UK, where the model is to be applied to the design/upgrading of sewer systems.

The literature on time series modelling of rainfall was reviewed, and the Neyman-Scott Rectangular Pulses rainfall model was identified as being potentially suitable for the project. An expression for the probability of an arbitrary interval being dry was derived for the Neyman-Scott Rectangular Pulses model, which was then used, together with expressions found by *Rodriguez-Iturbe et al* (1987a), to fit the model to historical hourly rainfall data taken from a site in Blackpool. Hourly rainfall data were then simulated using a computer program (Appendix B) for the model. Statistics were extracted from the historical and simulated rainfall time series and compared. The results of these comparisons showed that the performance of the model was good, so that the model could be used with reasonable confidence for the remaining part of the project.

An investigation was carried out to find an optimum fitting procedure for the selected stochastic rainfall model. This investigation revealed that the summer dry spell sequences were poorly matched by the model, when the historical daily lag 1 autocorrelations for each month were included in the fitting procedure. However, when the historical autocorrelations were omitted from the fitting procedure and the transition probabilities used as an alternative, the model's fit to the historical summer dry spell sequences showed considerable improvement. Furthermore, the model was able to match (within sampling variability) the historical lag 1 autocorrelations for each month even though these had been omitted from the fitting procedure.

To further validate the model, comparisons were made between other simulated and historical statistics that were not used in the fitting procedure. For example, the historical and simulated mean and standard deviations of the maxima (of the 1, 3, 6, 12, and 24 hourly time series) compared favourably, as did the simulated and historical mean and standard deviations of the proportion of hourly rainfalls above 1mm. However, the simulated and historical mean and standard deviations of the proportions of hourly rainfalls above 0mm did not compare so well, suggesting that the simulated data contained less light rainfall than the historical data. It was decided that this was unlikely to be of practical importance in simulating the hydraulic behavior of storm sewer systems.

Two fitting procedures were recommended (where the model is fitted one month at a time):

1) For hourly rainfall data the recommended fitting procedure used the following monthly historical statistics to fit the model: i) the mean of the hourly time series, ii) the variances of the 1, 3,

6, 12, and 24 hourly time series, iii) the wet given wet transition probabilities of the 1, 3, 6, 12, and 24 hourly time series, iv) the proportion of dry days, and v) the dry given dry transition probability of the daily time series.

2) For daily rainfall data the recommended fitting procedure used the following monthly historical statistics to fit the model: i) the daily mean, ii) the wet given wet and dry given dry transition probabilities of the daily time series, iii) the proportion of dry days, and iv) the daily variance. In addition, estimates for the 1, 3, 6, and 12 hourly historical variances were recommended for use in the fitting procedure. These estimates were obtained in the following way.

Hourly rainfall stations were sampled from each of the regions proposed by *Wigley et al* (1984). For the sampled stations the hourly variances were plotted against the corresponding daily variances. From these plots two conclusions were drawn:

i) for each station a regression equation of the form:

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

could be used to predict the hourly variance given the daily variance (where y_{ij} and x_{ij} are the hourly and daily variances (respectively) for the jth month of the ith station, and ε_{ij} is the residual for the jth month of the ith station), and

ii) the variances of the residuals about such a line were substantially different for each plot. Therefore, any method of grouping the rainfall stations by the coefficients of the regression equations should not include the assumption of homogeneity of variance for the regression parameters. Hence, some further developments in Cluster Analysis were made and the W statistic proposed. This statistic could be used to group population means (e.g. the parameters of the regression models) into non-overlapping groups, without the requirement that the population variances are equal.

The W statistic was used to test the hypothesis that the constants α_i for the regression models for each station were equal, i.e. the hypothesis H_0: $\alpha_i = \alpha$ (for all i) was tested. This hypothesis was retained, and the regression model for each station revised, with the constant taken to be the mean of the constants for each station, i.e. for each station the following model was fitted:

$$y_{ij} = \bar{\alpha} + \beta_i x_{ij} + \varepsilon_{ij},$$

where $\bar{\alpha}$ is the mean of the α_i. The W statistic was then used to test hypotheses that the β_i could be partitioned into non-overlapping groups suggested by the geographical location of the rainfall stations. It was found that the β_i could be taken as constant within two non-overlapping groups: one group which corresponded to stations lying in Scotland and the far North-East of England and the other group which corresponded to stations lying in England, Wales and Northern Ireland. The hourly rainfall stations were then pooled into the 2 groups and the parameters of the regression models were re-estimated for each group. This gave two regression equations for the hourly variances. Similarly, by using the same two groups, regression equations were developed to enable the 3, 6, and 12 hourly variances to be predicted from the daily variances.

The Neyman-Scott Rectangular Pulses rainfall model was fitted to the five longest records available for the project, where each station was located in one of the 'Wigley' regions. The following monthly historical statistics were used to fit the model: i) the daily mean, ii) the wet given wet and dry given dry transition probabilities of the daily time series, iii) the proportion of dry days, and iv) the daily variance. In addition, the regression estimates for the 1, 3, 6, and 12 hourly historical variances were also used in the fitting procedure. For each station rainfall data were simulated for January and July. The historical and simulated maximum daily rainfalls for each year were found and plotted against the standardised Gumbel variate. These plots suggested that the model had a tendency to under-estimate the extreme rainfall events (particularly events with return periods in excess of 10 years). Regression equations were developed so that the mean and variances of the simulated maxima could be predicted given the Neyman-Scott model parameters. These regression equations were then included in the fitting procedure and the Neyman-Scott model parameters re-estimated for each station. Rainfall data were simulated for January and July using the revised parameter sets. The historical and (revised) simulated maximum daily rainfalls for each year were found and plotted against the standardised Gumbel variate. These plots were compared to the original plots. For the station-months where the simulated maxima had originally failed to match the historical maxima, it was evident that the model was now able to match the historical maxima (within sampling variability). However, for one station-month the revised simulated maxima gave a poorer fit to the historical maxima (which on the original plots had been well matched by the model). This suggested that the regression model for the mean and variances of the maximum daily rainfalls should only be used in the fitting procedure when the model fails to match the historical extremes.

The stochastic rainfall model was fitted to the remaining daily data. The historical statistics used in the fitting procedure for each station-month were: i) the daily mean, ii) the wet given wet and dry given dry transition probabilities of the daily time series, iii) the proportion of dry days, and iv) the daily variance. Furthermore, the regression estimates for the 1, 3, 6, and 12 hourly historical variances were also used in the fitting procedure.

The parameters for all station-months were then regressed on site characteristics (e.g. altitude, distance from coast, etc) so that the parameters of the Neyman-Scott rainfall model could be estimated at sites lacking rainfall data. On average, the accuracy of the parameter estimates obtained from the regression model was equivalent to using 20 years of daily data to estimate the model parameters (assuming the mean monthly totals are known for the site). However, the standard deviation of this average value was quite large, which implied that the regression model may sometimes give much poorer estimates of the Neyman-Scott model parameters than would be obtained if the model was fitted to 20 years of daily data. Conversely, if rainfall data are available at a site, the engineer may need to buy more than 20 years of daily data in order to be confident that the parameter estimates obtained when fitting the model to the bought daily data are better than those of the regression model.

The hourly rainfall time series generated by the computer program for the rainfall model will sometimes need to be disaggregated to 5 minutely time series, which are sometimes used as input to sewer system design programs (such as WASSP-SIM). The proposed disaggregation model is similar to the disaggregation model used by *Ormsbee* (1989), with the exception that the depth of rain per pulse is taken to be a parameter of the model. The parameter of the disaggregation model was estimated for each month of the Farnborough data set (28 years of minutely data). Using these parameter estimates, the disaggregation model was tested and shown to perform well on the longest record of minutely data in another precipitation region (Rhoose, 16 years). Due to the lack of available minutely rainfall data, it was not possible to test the model on any sites located in the North of Great Britain. However, most of the regional variation in rainfall can probably be found in hourly or daily time series, and so this should not cause concern.

8.2 CONCLUSIONS

1) The Neyman-Scott Rectangular Pulses model may be used with confidence to generate rainfall time series in the UK. For its intended application, discrepancies between historical and simulated rainfall should be checked to see whether they are of practical significance for the drainage engineer engaged in sewer system modelling (see Section 8.3).

2) When fitting the model to historical rainfall time series, it is preferable to use a procedure which minimises a sum of squares, rather than solving a set of simultaneous equations, where the squared terms are the difference between historical statistics and the equivalent function of the model parameters.

3) When fitting the model to historical rainfall time series (either hourly or daily), the model should be fitted on a monthly basis, i.e. the parameters of the model estimated for each month of the record.

4) When fitting the model to historical rainfall time series (either hourly or daily), the mean rainfall (for each month) should be used in the fitting procedure.

5) When fitting the model to historical rainfall time series (either hourly or daily), the daily transition probabilities (both wet given wet and dry given dry) should be used in the fitting procedure.

6) When fitting the model to historical hourly rainfall time series, the 1, 3, 6, 12, and 24 hourly variances and the 1, 3, 6, 12, and 24 hourly wet given wet transition probabilities should be used in the fitting procedure.

7) When fitting the model to historical daily rainfall time series, it is advisable to use an estimate of the historical 1, 3, 6, and 12 hourly variances in the fitting procedure.

8) When estimating the historical 1, 3, 6, and 12 hourly variances, it is advisable to use regression equations based on the 24 hourly variances. It is recommended that different regression equations are used for 2 different regions of the UK (Scotland/far North-East of England and England/Wales/Northern Ireland).

9) If the model is required to match extreme values found in an historical record of daily rainfall data, then the model should be fitted to the data using the recommended procedures described above, and the maximum daily rainfalls between the historical and simulated data compared on Gumbel probability paper. If the model shows a tendency to under-estimate the historical maximum daily rainfalls, then the regression equation for the mean and variance of the maximum daily rainfalls should be included in the fitting procedure.

10) For applications to sites with no data, the parameters of the Neyman-Scott Rectangular Pulses model can be estimated using the multivariate (regionalised) regression model based on site characteristics. The following points should be noted concerning the regionalised model:

a) On average, the standard error of the parameter estimates obtained from the multivariate (regionalised) regression model is equal to the standard error of the parameter estimates that would be obtained when fitting the model to about 20 years of historical daily rainfall time series taken from the site (assuming mean monthly totals are used in the fitting procedure).

b) If historical rainfall time series (either hourly or daily) are to be used at the site to fit the stochastic rainfall model, then a weighted average between the parameters estimated using the site data and the parameters estimated from the regionalised regression model should be used.

11) If required 5 minutely rainfall time series can be generated by disaggregating the simulated hourly rainfall time series. A parameter for the depth (in mm) of a pulse of rain was required when fitting the disaggregation model to the variances of the 5 minutely time series for each month of the Farnborough data set. This parameter was shown to vary seasonally. Using the same estimates of this parameter for each month, the disaggregation model can be used with reasonable confidence at other sites located in the UK.

8.3 DIRECTIONS FOR FUTURE RESEARCH

8.3.1 Testing the model using a sewer system simulation program

Before the model can be used with confidence for its intended application, it should be checked by running simulated rainfall time series through a sewer system model (such as WASSP-SIM), because discrepancies, which cannot always be attributed to chance, between historical and simulated data sometimes occur.

There are three questions which need to be answered before the model can be used with confidence for upgrading UK sewer systems:

1) How well does the Neyman-Scott Rectangular Pulses model perform for its intended application when fitted to site data.

2) How well does the regionalised version of the Neyman-Scott model perform?

3) How well does the disaggregation model perform?

Perhaps the most efficient and reliable approach would be to attempt to answer all three questions with one testing strategy. If the model fails the testing strategy the questions could then be answered one by one to identify weak areas of model performance. One possible testing strategy is now discussed and a less expensive alternative considered afterwards.

To answer (3) minute data are required. To answer (2) historical rainfall records of greater than 20 years duration (Section 6.7) would be needed at several sites known to have overflow problems. The three questions could then be answered by simulating 5 minutely rainfall data (for more than 20 years) using the regionalised stochastic model, together with the disaggregation model, for the sites with known overflow problems. The procedures used by drainage engineers for inputing rainfall time series to sewer system models[1] could then be applied to both the simulated and historical time series. The spills predicted using the historical series could then be compared to those predicted using the simulated series, and a decision made as to whether the difference (in spill volumes) is of practical significance for the engineer involved in upgrading a sewer system. This exercise should be performed on several sites with known overflow problems, so that any tendencies for the model to over- or under-estimate spill volumes could be found. In addition, the sites should be selected from several regions of the country to take account of different rainfall patterns for different locations.

If a tendency for the model to over- or under-estimate spill volumes is found, then measures can be taken to compensate. One such measure would be in the sampling procedure (Section 8.3.2), where rainfall events may be selected for input to the design program. When upgrading an existing sewer system, a common approach is to remove the most extreme events from the rainfall record, so that the day to day performance of the system can be assessed. If the model is found to under-estimate the spill

[1] This may mean that the sampling procedures discussed in the next Section would need to be developed before the testing strategy can take place.

volumes (which is more likely than over-estimation (Section 5.8.3)), then less of the extreme events should be removed to compensate. Obviously this would depend on how much, if at all, the model under-estimates the spill volumes.

The testing strategy, described above, requires several long historical records of minute data, and so may prove too expensive if these data have to be bought. A less expensive strategy would be to test the disaggregation model first, and then, if the model passes this test, use the disaggregation model to disaggregate both historical and simulated hourly time series so that the testing strategy above can be carried out. The disaggregation model could be tested as follows:

i) Select a site with known overflow problems, for which historical 5 minutely data are available.

ii) Aggregate the 5 minutely data to form an hourly rainfall time series.

iii) Using the disaggregation model, disaggregate the aggregated hourly rainfall time series to form a series of disaggregated 5 minutely data.

iv) Compare the spill volumes predicted using the disaggregated series with those predicted using the historical 5 minutely series.

v) Decide whether the difference in spill volumes is of practical significance for the engineer engaged in upgrading sewer systems.

vi) Repeat the above for several sites with known overflow problems.

For this test, a long record is not required but several sites should be considered.

8.3.2 A sampling procedure

Running rainfall time series through a sewer system simulation program, such as WASSP-SIM, is expensive on computing resources. Therefore, procedures need to be developed, so that the most significant rainfall events can be sampled from the generated time series. One approach would be to classify events according to length of dry spell (preceding the event), depth, and duration of event, and then use those events which are of most practical significance in the design program (see *Henderson* (1986) for a similar procedure).

8.3.3 Further research on using the Neyman-Scott Rectangular Pulses model to generate extreme events

In Chapter 5, Section 5.8.3, the model showed a tendency to under-estimate the extreme events, mainly for return periods in excess of 10 years. A solution to this problem was proposed by regressing the mean and variance of the simulated maxima on the model parameters and then including these regression equations in the fitting procedure, when the model showed a poor fit to the historical maxima. An alternative approach may be to allow the cell intensity to follow a distribution (e.g. the Gamma distribution) that has a longer 'tail' than the exponential distribution, and is thus likely to lead to more extreme events when simulating data using the model. This would require no further theoretical developments of the model, and would not require the fitting procedure to be adapted. This approach was not adopted in this thesis as parameter parsimony was required, mainly for the regionalisation procedure, and a distribution (such as the Gamma distribution) would introduce a further parameter into the model.

8.3.4 Generalising the model as a spatial-temporal process

The purpose of this project was to produce a stochastic rainfall time series model for the UK. Therefore, attention was focused on the temporal modelling of rainfall. However, it is also desirable, in engineering design problems, to model the spatial variation of the rainfall over a catchment area, because the concentration of rainfall in some locations can lead to high local run-off (e.g. see *Wilson et al* (1979), *Hamlin* (1983), *Nicks* (1982), or *Milly and Eagleson* (1988)). Therefore, further research is needed to generalise the time series model to a spatial-temporal model.

One approach may be to use a spatial (field) model (e.g. the Modified Turning Bands Model (*Mellor,* 1991)), and condition this model on the temporal time series model, i.e. generate rainfall time series using the temporal stochastic model and, when storms occur, distribute the rainfall over the catchment area using the spatial model.

An alternative approach may be to follow *Breckling* (1989) and use a 'directional' time series model. *Breckling* (1989) successfully applied a directional time series model to wind data taken from Fremantle, Western Australia. Further work could be carried out to see whether such a model could be successfully used for UK rainfall data.

APPENDICES

A: Some Further Theoretical Developments

B: A Simulation Program for the Neyman-Scott Rectangular Pulses Model (written in Pascal)

C: T-tests and Dry Spell Sequences when M1, V1, V6, V24, AC1, AC6, AC24, PD24 (monthly historical statistics) are used in the fitting procedure

D: T-tests and Dry Spell Sequences when M1, V3, AC3, V6, AC6, WW6, WW12, V24, PD24, WW24, DD24 (seasonal historical statistics) are used in the fitting procedure

E: T-tests, Dry Spell Sequences, and Comparisons of Mean and Standard Deviations of historical and simulated statistics when M1, V1, V3, V6, V12, V24, WW1, WW3, WW6, WW12, WW24, DD24, PD24 are used in the fitting procedure

F: A closer look at the 3 observations excluded from the multivariate (regional) regression analysis

G: Residual plots for the multivariate (regional) regression model

H: Treatment of missing values

I: A Comparison of Storm Profiles for the Historical 5 Minutely Time Series, the Disaggregated 5 Minutely Time Series and the 'Smooth' Disaggregated 5 Minutely Time Series

J: The Correlograms of the Monthly Mean Daily Rainfalls and Residual Series

K: The effect of the error in expression (3.6)

R: References

APPENDIX A: SOME FURTHER THEORETICAL DEVELOPMENTS

In this Section, some results are derived for single storms under the Neyman-Scott model. Therefore, as cells from one storm may overlap another storm, the results are not applicable in estimating the parameters of the model. However, they may be useful for future developments of the model.

Of interest is the location (relative to storm origin) of the most intense part of the storm. To find this, consider a single cell following a storm origin,

$$\begin{aligned}
\text{pr(cell active at time t)} &= \int_0^t \beta e^{-\beta u}\, e^{-\eta(t-u)}\, du \\
&= \beta e^{-\eta t} \left[\frac{1}{\eta-\beta}\, e^{(\eta-\beta)u} \right]_0^t \\
&= \beta e^{-\eta t} \left(e^{(\eta-\beta)t} - 1 \right) \\
&= \beta (e^{-\beta t} - e^{-\eta t})/(\eta-\beta) \\
&= A(t), \text{ say.}
\end{aligned}$$

The most intense part of the storm is most likely to occur where most of the rain cells occur. This will be the value of t which maximises A(t). Therefore, we proceed as follows:

$A'(t) = \beta(-\beta e^{-\beta t} + \eta e^{-\eta t})/(\eta-\beta) = 0 \Leftrightarrow \beta e^{-\beta t} = \eta e^{-\eta t}$

$\Leftrightarrow t = \frac{1}{\eta-\beta} \ln(\eta/\beta) = t_{max}$ - the most likely location of cell activity relative to the storm origin.

Hence, the probability of a rain cell being active at t_{max} is given by:

$$A_{max} = A(t_{max}) = \left(\frac{\beta}{\eta} \right)^{\eta/(\eta-\beta)} \qquad \text{(A.1)}$$

Now, let N be the number of cells generated from the storm origin, and assume N-1 is Poisson with mean $\mu = \nu - 1$. Also let N_{max} be the number of cells active at t_{max}.

$$pr(N_{max} = k \mid N=n) = \binom{n}{k} A_{max}^{k} (1-A_{max})^{n-k} \quad (k \le n), \tag{A.2}$$

i.e. $N_{max} \mid N=n \sim B(n, A_{max})$.

Therefore,

$$pr(N_{max} = k) =$$

$$pr(k \text{ cells active at } t_{max}) = \sum_{n=k}^{\infty} \binom{n}{k} A_{max}^{k} (1-A_{max})^{n-k} \; pr(N=n)$$

$$= \sum_{n=k}^{\infty} \binom{n}{k} A_{max}^{k} (1-A_{max})^{n-k} \mu^{n-1} e^{-\mu} / (n-1)!$$

$$= \sum_{n=k}^{\infty} \binom{n}{k} A_{max}^{k} (1-A_{max})^{n-k} \mu^{n-1} e^{-\mu} / (n-1)!$$

$$= A_{max}^{k} \mu^{k-1} \exp\{-\mu A_{max}\} [\mu(1-A_{max}) + k]/k!$$

$$= (\mu A_{max})^{k} \exp\{-\mu A_{max}\} [1 - A_{max} + k/\mu]/k! \tag{A.3}$$

The expectation and variance of the number of cells active at t_{max} can be derived using the results:

$$E(N_{max} \mid N=n) = n A_{max}, \text{ and} \tag{A.4}$$

$$E(N_{max}^{2} \mid N=n) = n A_{max} (1 - A_{max}) + (n A_{max})^{2} \tag{A.5}$$

(which are obtained from $N_{max} \mid N=n \sim B(n, A_{max})$)

Using (A.4),

$$E(N_{max}) = \sum_{n=1}^{\infty} n \, A_{max} \, pr(N=n)$$

$$= A_{max} \sum_{n=1}^{\infty} n \, \mu^{n-1} \, e^{-\mu} \,/\, (n-1)!$$

$$= A_{max} \, e^{-\mu} \sum_{n=0}^{\infty} (n+1) \, \mu^{n} \,/\, n!$$

$$= A_{max} \, (\mu + 1)$$

$$= \nu A_{max}, \tag{A.6}$$

as might be anticipated.

Similarly, using (A.5),

$$E(N^2_{max}) = \sum_{n=1}^{\infty} E(N^2_{max} \mid N=n) \; pr(N=n)$$

$$= A_{max} \, (1 - A_{max}) \, e^{-\mu} \sum_{n=1}^{\infty} n \, \mu^{n-1}/(n-1)! \; + \; e^{-\mu} \, A^2_{max} \sum_{n=1}^{\infty} n^2 \, \mu^{n-1}/(n-1)!$$

$$= \nu A_{max} \, (1 - A_{max}) \; + \; e^{-\mu} \, A^2_{max} \sum_{n=0}^{\infty} (n + 1)^2 \, \mu^n/n!$$

$$= \nu A_{max} \, (1 - A_{max}) \; + \; e^{-\mu} \, A^2_{max} \sum_{n=0}^{\infty} (n^2 + 2n + 1) \, \mu^n/n!$$

$$= \nu A_{max} \, (1 - A_{max}) \; + \; e^{-\mu} \, A^2_{max} \, [\mu(\mu+1)e^{\mu} + 2\mu e^{\mu} + e^{\mu}]$$

$$= \nu A_{max} \, (1 - A_{max}) \; + \; A^2_{max} \, [\mu^2 + 3\mu + 1]$$

$$= \nu A_{max} \, (1 - A_{max}) \; + \; A^2_{max} \, [\nu^2 + \nu - 1]$$

$$= (\nu A_{max})^2 \; + \; \nu A_{max} \; - \; A^2_{max}$$

Hence, $Var(N_{max}) = \nu A_{max} - A^2_{max}.$ (A.7)

Now let X_i be the intensity of ith cell that is active at t_{max} ($i=1,2,\ldots,N_{max}$), and I_{max} be the total intensity at t_{max}. Then,

$$I_{max} = \sum_{i=1}^{N_{max}} X_i$$

$$E(I_{max} \mid N_{max} = k) = E(\sum_{i=1}^{k} X_i) = k\,E(X_i) = k/\xi. \qquad (A.8)$$

Therefore,

$$E(I_{max}) = \sum_{k=0}^{N} E(I_{max} \mid N_{max} = k)\ pr(N_{max} = k)$$

$$= \sum_{k=0}^{N} k\ pr(N_{max} = k)/\xi \qquad \text{(using (A.8))}$$

$$= E(N_{max})/\xi = \nu A_{max}/\xi$$

$$= \frac{\nu}{\xi}\left(\frac{\beta}{\eta}\right)^{\eta/(\eta-\beta)} \qquad (A.9)$$

Now let X_i = duration of ith rain cell, and Y_i = intensity of ith rain cell ($i = 1, 2, \ldots, N$).

Let V_i = total volume of rain falling due to ith cell, so $V_i = X_i Y_i$, and let $V = \sum_i V_i$ = total volume of rain falling due to storm.

Under the model the X_i are iid $\exp(\eta)$, and the Y_i are iid $\exp(\xi)$.

Hence,

$E(V_i) = E(X_i Y_i) = E(X_i)E(Y_i) = 1/(\eta\xi)$, and

$E(V_i^2) = E(X_i^2)E(Y_i^2) = 4/(\eta^2\xi^2) \Rightarrow Var(V_i) = 3/(\eta^2\xi^2)$.

Therefore,

$$E(V|N=n) = n/(\eta\xi), \text{ and}$$

$$\begin{aligned}E(V^2|N=n) &= E(\{V_1 + V_2 + \dots + V_n\}^2) \\ &= \sum_i E(V_i^2) + \sum_{i\neq j} E(V_i)E(V_j) \\ &= 4n/(\eta^2\xi^2) + 2n(n-1)/(\eta^2\xi^2) \\ &= 2n(n+1)/(\eta^2\xi^2),\end{aligned}$$

and, $\text{Var}(V|N=n) = n(n+2)/(\eta^2\xi^2)$.

Hence,

$$E(V) = \sum_{n=1}^{\infty} n \, \text{pr}\{N=n\}/(\eta\xi) = \sum_{n=1}^{\infty} n \, (\nu-1)^{n-1}e^{-\nu+1}/ \{(n-1)!\eta\xi\},$$

and, assuming N-1 is Poisson with mean ν-1, this gives

$$\begin{aligned}E(V) &= (\eta\xi)^{-1} e^{-\nu+1} \sum_{n=0}^{\infty} (n+1)(\nu-1)^n/n! \\ &= (\eta\xi)^{-1} e^{-\nu+1} \nu e^{\nu-1} \\ &= \nu/(\eta\xi)\end{aligned} \tag{A.10}$$

$$\begin{aligned}E(V^2) &= \sum_{n=0}^{\infty} 2n(n+1)(\nu-1)^{n-1}e^{-\nu+1}/ \{(n-1)!\eta^2\xi^2\} \\ &= (2\nu(\nu-1) + 6(\nu-1) + 4)(\eta\xi)^{-2} \\ &= 2(\nu^2 + 2\nu - 1)(\eta\xi)^{-2}\end{aligned}$$

Therefore,

$$\begin{aligned}\text{Var}(V) &= 2(\nu^2 + 2\nu - 1)(\eta\xi)^{-2} - \nu^2(\eta\xi)^{-2} \\ &= (\nu^2 + 4\nu - 2)(\eta\xi)^{-2}\end{aligned} \tag{A.11}$$

APPENDIX B:

A SIMULATION PROGRAM FOR THE NEYMAN-SCOTT RECTANGULAR PULSES MODEL (WRITTEN IN PASCAL).

```
PROGRAM TimeSeries (input, output, data, randomdata, parset);

{This program simulates hourly rainfall time series using the
 Neyman-Scott Rectangular Pulses Model. References are made to
 Chapter 3}

{exit} LABEL 999;

CONST maxstorms {per month} = 30; maxcells {per storm} = 40;
      maxmins = 53280 {the maximum time in minutes in which
      storms will arise};
      maxyears  {the maximum number of years that can be
      simulated} = 100;
      maxhours = 888 {the maximum time in hours};
      correction {in days} = 6 {The correction allows for some
      over-lap of storms from one month to the next};

TYPE whole = 0..maxint; storms = 0..maxstorms; months = 1..12;
     cells = 0..maxcells; hours = 0..maxhours; minutes =
     0..maxmins; years = 1..maxyears;

VAR x1, x2, x3, x4, x5 {model parameters} : longreal;
    daysinmonth {the total number of days in the month},
    nlines {loop variable for number of lines written to
    output file},
    laststormorigin {the time at which the last storm origin
    occurs}: whole;
    laststorm {the number of storm origins occurring in a month}
    : storms;
    month : months; year {loop variable}, totalyears {the
    number of years of simulated data specified by the user}
    : years;
    totalhours {the total number of hours in which storm origins
    can arise = number of hours in month + number of hours
    in correction} : hours;
```

```
   totalmins {the total number of minutes in which storms can
   arise = total number of minutes in month + number of minutes
   in correction},
   previousnmins {the number of minutes in the preceeding month}
   : minutes;
   {see Model Definition, Chapter 3, for meaning of terms such
   as storm origin, cells, etc}
   stormorigin : ARRAY [storms] OF whole;
   totalcells : ARRAY [storms] OF cells;
   cellbegin, cellend : ARRAY [storms, cells] OF whole;
   cellintensity : ARRAY [storms, cells] OF longreal;
   cellduration : whole;
   hour {hourly rainfall time series} : ARRAY [hours] OF
   longreal;
   minute {minutely rainfall time series} : ARRAY [minutes] OF
   longreal;
   randomnum {a uniform random number between 0 and 1 read from
   random.dat},
   zero {0.00...} : longreal;
   data {the output file of simulated hourly rainfall time
   series},
   parset {the file containing the parameter estimates for a
   month},
   randomdata {the file of random numbers} : text;



PROCEDURE ReadParameters;

{Model parameters discussed in Section 3.2 of Chapter 3}
{x1 = lambda, x2 = beta, x3 = eta, x4 = nu, x5 = xi}

BEGIN
  reset (parset);
  read (parset, x1, x2, x3, x4, x5);
  x1 := 60 / x1;
  {the mean waiting time in minutes between two adjacent
   storm origins}
  x2 := 60 / x2;
  {the mean waiting time after the storm origin
```

```
   for the beginning of a rain cell}
  x3 := 60 / x3;
  {the mean cell duration}
  x4 := x4 - 1;
  {the mean number of cells per storm - 1. The -1 ensures that
   at least one rain cell follows a storm origin - see Model
   Definition, Section 3.2, Chapter 3}
  x5 := 1 / (x5*60)
  {the mean cell intensity in mm per minute}
END {PROCEDURE};



FUNCTION NegExp (mean : longreal) : longreal;

{generates an Exponential random variable}

BEGIN
  read (randomdata, randomnum);
  NegExp := -mean * ln (randomnum)
END {FUNCTION};



FUNCTION Poisson (mean : longreal) : integer;

{generates a Poisson random variable}

VAR sum : real; count : whole;

BEGIN
  sum := 0;
  count := 0;
  REPEAT
    sum := sum + NegExp (1);
    count := count + 1
  UNTIL sum > mean;
  Poisson := count - 1
END {FUNCTION};
```

```
PROCEDURE CellError;

{if this happens maxcells, in the CONST declaration statement,
 may need to be increased}

BEGIN
  writeln;
  writeln (' ****************************** ');
  writeln (' *            ERROR           * ');
  writeln (' *   total cells > maxcells   * ');
  writeln (' *      program aborted       * ');
  writeln (' ****************************** ');
  writeln;
  GOTO 999
END;


PROCEDURE WriteStorms;
{writes the rainfall data for the month to the output file}

LABEL 98;

VAR
  nhours {total hours in the month}, h {loop variable}, h1 {loop
  starting point}, h2 {loop end point} : hours;
  nmins {total minutes in the month}, m {loop variable}: minutes;
  cell {loop variable} : cells; storm {loop variable} : storms;
  m1 {loop starting point} , m2 {loop end point} : whole;

BEGIN
  nhours := daysinmonth * 24;
  nmins := 60 * nhours;

  {The difference between nhours and totalhours, and nmins and
   totalmins may need to be clarified. nhours or nmins are the
   total number of hours or minutes (respectively) in the
   month (e.g. 31*24 or 31*24*60). totalhours or totalmins
   are the total number of hours or minutes in which rainfall
   are generated, which equals nhours (or nmins) + correction
   (in hours or minutes) (e.g. 31x24+6x24 or 31x24x60+6x24x60)}
```

```
FOR m := 1 to totalmins DO minute [m] := 0;
{the minute time series starts with zero values throughout,
 and rainfall due to the rain cells are added to these values}
FOR storm := 1 to laststorm DO
  FOR cell := 1 to totalcells [storm] DO
  BEGIN
    m1 := cellbegin [storm, cell];
    m2 := cellend [storm, cell];
    IF (m1 > totalmins) OR (m2 > totalmins) THEN GOTO 98;

    {Exit to 98 would occur when a storm near the end of the
     month lasted longer than the correction time of 6
     days - this unlikely for the parameter estimates obtained
     for the UK rainfall data of this project}

    FOR m := m1 to m2 DO
      minute [m] := minute [m] + cellintensity [storm, cell];
      {the cell intensities are constant throughout the cell
       durations - refer to  Model Definition, Chapter 3}
    98:
  END;
FOR h := 1 to totalhours DO
{converts the minutely time series to an hourly time series}
BEGIN
  m1 := h * 60 - 59;
  m2 := h * 60;
  FOR m := m1 to m2 DO hour [h] := hour [h] + minute [m]
END;
nlines := 0;
REPEAT
  {writes the hourly data to TSR_1H.SIM, with 12 values per
   line}
  nlines := nlines + 1;
  h1 := nlines * 12 - 11;
  h2 := nlines * 12;
  FOR h := h1 to h2 DO
  write (data, hour [h]:6:2);
  writeln (data)
UNTIL nlines = nhours DIV 12;
FOR h := 1 to correction * 24 DO
```

```
  {this allows for the possibility of storms from one month
   carrying over to the next month}
    hour [h] := hour [nhours + h];
    FOR h := correction * 24 + 1 to totalhours DO
    hour [h] := 0
END;


PROCEDURE GenerateStorms;

{Generates storms on a monthly basis}

LABEL 99;

VAR i, j, k {loop variables}, cellno {the random number of
raincells generated} : whole; nmins {number of minutes in the
month} : minutes;

BEGIN
  nmins := daysinmonth * 24 * 60;
  FOR i := 1 to maxstorms DO
  {the loop will be exited when a storm origin occurs after
   the last day in the month}
  BEGIN
    IF i = 1
    THEN
      stormorigin [i] := laststormorigin - previousnmins
      {when year = month = 1 this is zero. Otherwise it is the
       time at which the first storm origin occurs}
    ELSE
      stormorigin [i] := stormorigin [i - 1] +
      round (NegExp (x1));
      {the time between adjacent storms is Exponential - see
       model definition, Chapter 3}
    IF stormorigin [i] > nmins THEN
    BEGIN
      laststorm := i - 1;
      {Storms will be generated up to, and including the
       (i - 1)th storm}
```

```
      laststormorigin := stormorigin [i];
      previousnmins := nmins;
      {the number of minutes in the month, which on the next
       iteration is the number of minutes in the previous month}
      {laststormorigin - previousnmins gives the position at
       which the first storm origin appears in the following
       month}
      GOTO 99;
    END;
    cellno := Poisson (x4) + 1;
    {the number of cells generated -1 is random
    and follows a Poisson distribution}
    IF cellno > maxcells THEN CellError;
    totalcells [i] := cellno;
    {stores the number of cells generated for ith storm}
    FOR j := 1 to totalcells [i] DO
    BEGIN
      cellbegin [i,j] := stormorigin [i] + round (NegExp (x2));
      {starting time in minutes of jth cell of ith storm}
      cellduration := round (NegExp (x3));
      cellend [i,j] := cellbegin [i,j] + cellduration - 1;
      {end time of jth cell for ith storm}
      cellintensity [i,j] := NegExp (x5)
      {intensity in mm per minute of jth cell of ith storm}
    END;
  END;
  99 : WriteStorms
END {PROCEDURE};



BEGIN {MAIN}
  zero := 0;
  assign (data, 'tsr_1h.sim');
  rewrite (data);
  assign (randomdata, 'random.dat');
  reset (randomdata);
  writeln;
  writeln
  ('This program simulates hourly rainfall time series,');
```

```
writeln
('for all months over a number of years given by the user.');
writeln ('Uniform random numbers are read from: RANDOM.DAT');
writeln ('Parameter estimates are read from: ???.PAR');
writeln ('Hourly data are written to: TSR_1H.SIM');
writeln;
write ('Enter the number of years of data required: ');
readln (totalyears);
writeln;
writeln
('Please wait. Simulating ',totalyears:1,' years of data...');
writeln;
writeln ('Years done:- ');
FOR year := 1 to totalyears DO
BEGIN
  FOR month := 1 to 12 DO
  BEGIN
    CASE month OF
    1  : BEGIN
           totalhours := 31 * 24 + correction * 24;
           daysinmonth := 31;
           IF year = 1 THEN
           {January, year 1, begins with a storm}
           BEGIN
             laststormorigin := daysinmonth * 24 * 60;
             previousnmins := laststormorigin
           END;
           IF year = 1
           THEN
             assign (parset, 'jan.par')
           ELSE
             close (parset, true);
             assign (parset, 'jan.par')
         END;
    2  : BEGIN
           totalhours := 28 * 24 + correction * 24;
           daysinmonth := 28;
           close (parset, true);
           assign (parset, 'feb.par')
         END;
```

```
3   : BEGIN
        totalhours := 31 * 24 + correction * 24;
        daysinmonth := 31;
        close (parset, true);
        assign (parset, 'mar.par')
      END;
4   : BEGIN
        totalhours := 30 * 24 + correction * 24;
        daysinmonth := 30;
        close (parset, true);
        assign (parset, 'apr.par')
      END;
5   : BEGIN
        totalhours := 31 * 24 + correction * 24;
        daysinmonth := 31;
        close (parset, true);
        assign (parset, 'may.par')
      END;
6   : BEGIN
        totalhours := 30 * 24 + correction * 24;
        daysinmonth := 30;
        close (parset, true);
        assign (parset, 'jun.par')
      END;
7   : BEGIN
        totalhours := 31 * 24 + correction * 24;
        daysinmonth := 31;
        close (parset, true);
        assign (parset, 'jul.par')
      END;
8   : BEGIN
        totalhours := 31 * 24 + correction * 24;
        daysinmonth := 31;
        close (parset, true);
        assign (parset, 'aug.par')
      END;
9   : BEGIN
        totalhours := 30 * 24 + correction * 24;
        daysinmonth := 30;
        close (parset, true);
```

```
          assign (parset, 'sep.par')
        END;
    10 : BEGIN
          totalhours := 31 * 24 + correction * 24;
          daysinmonth := 31;
          close (parset, true);
          assign (parset, 'oct.par')
        END;
    11 : BEGIN
          totalhours := 30 * 24 + correction * 24;
          daysinmonth := 30;
          close (parset, true);
          assign (parset, 'nov.par')
        END;
    12 : BEGIN
          totalhours := 31 * 24 + correction * 24;
          daysinmonth := 31;
          close (parset, true);
          assign (parset, 'dec.par')
        END
    END {CASE};
    totalmins := totalhours * 60;
    ReadParameters;
    GenerateStorms
  END;
 if year mod 10 - 1 = 0 then writeln;
 write (year:1, '  ')
 END;
 999:
END {MAIN}.
```

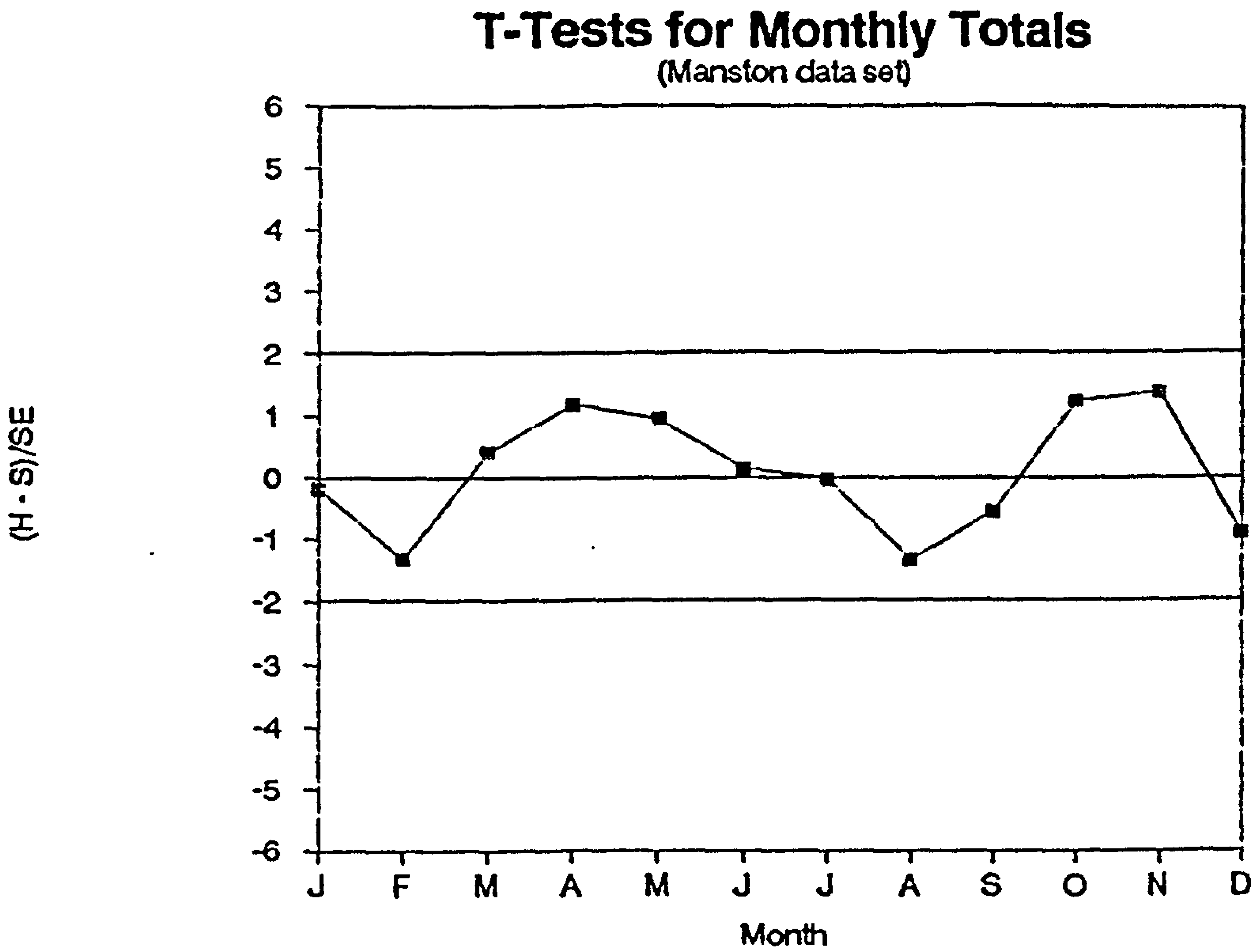


Figure C.1

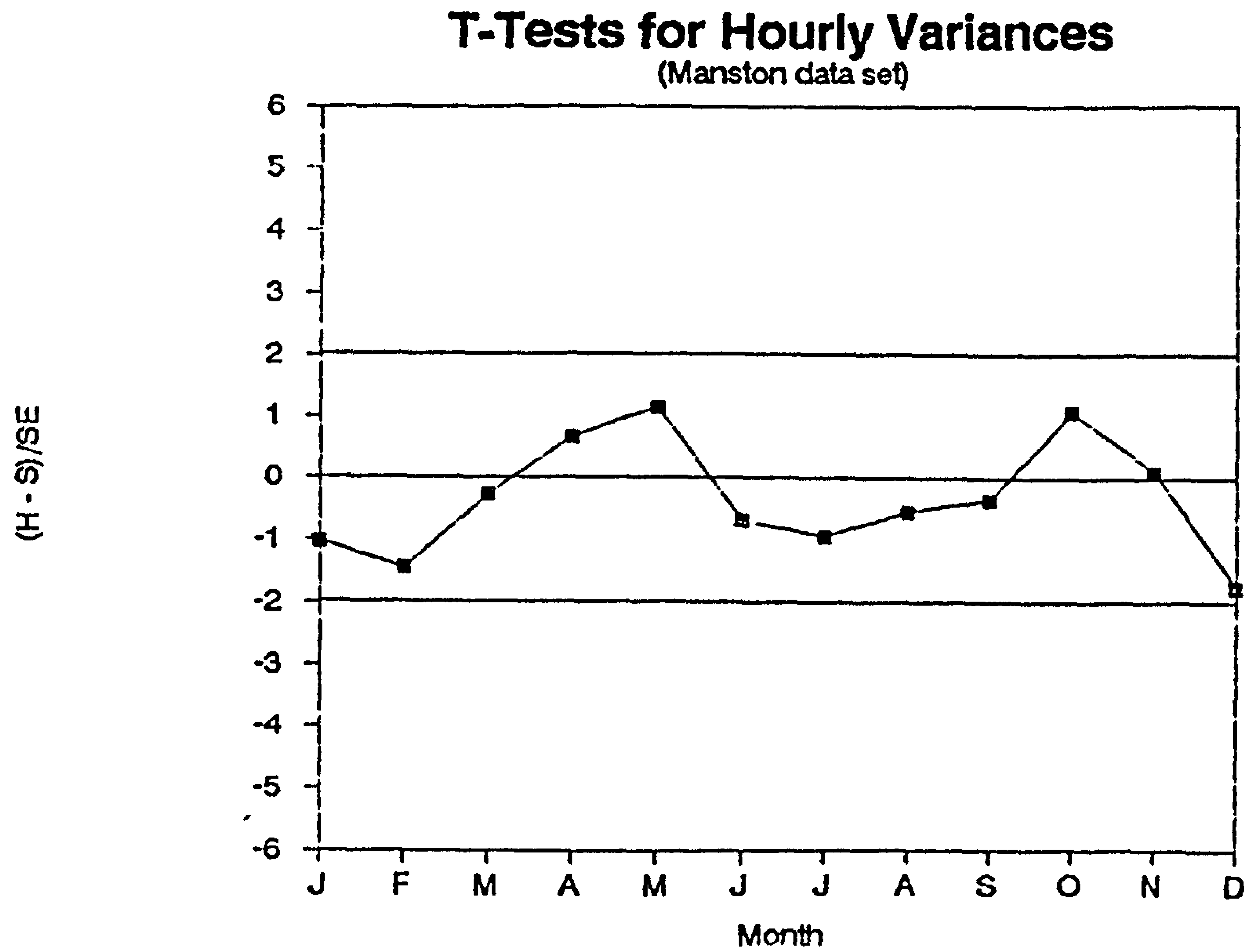


Figure C.2

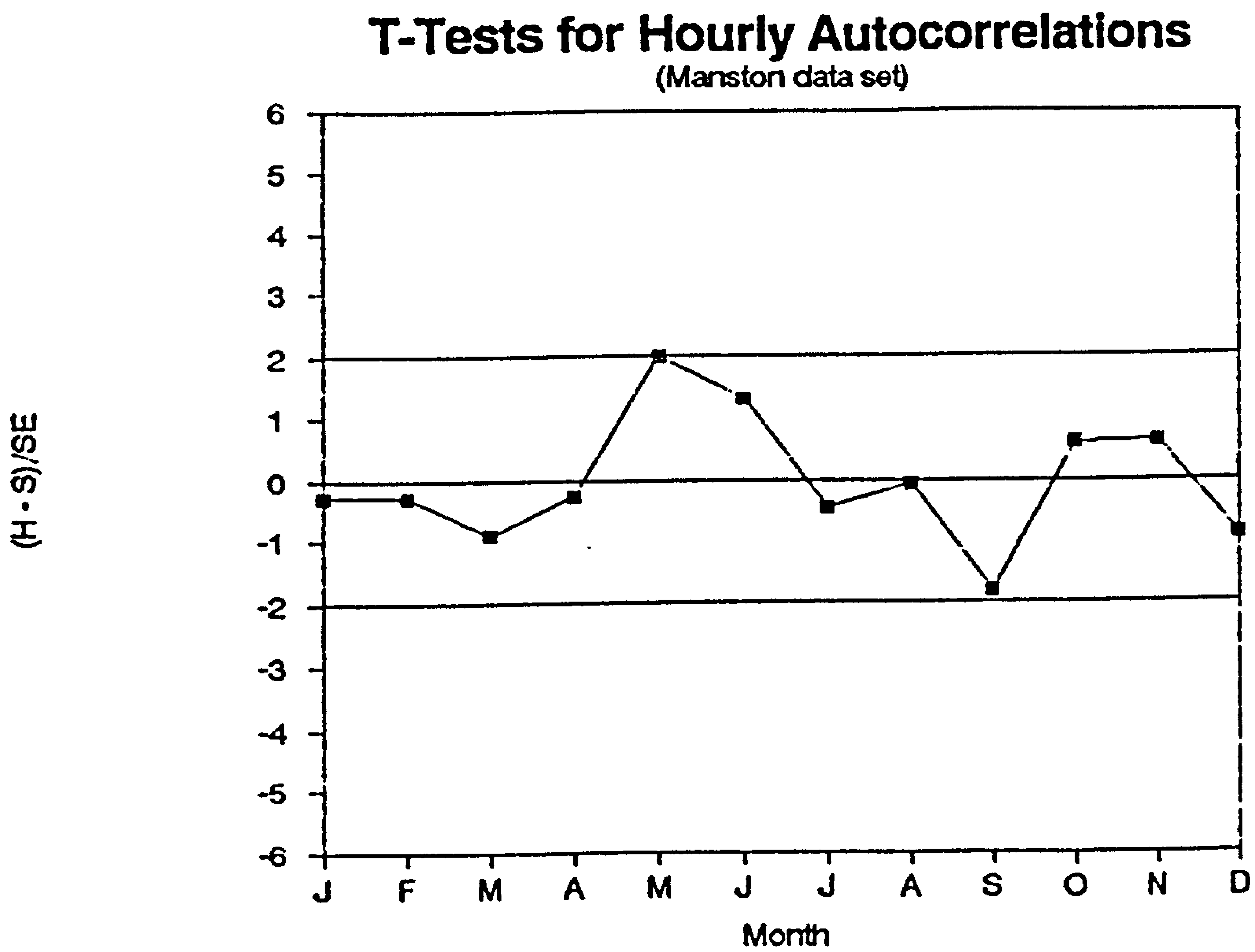


Figure C.3

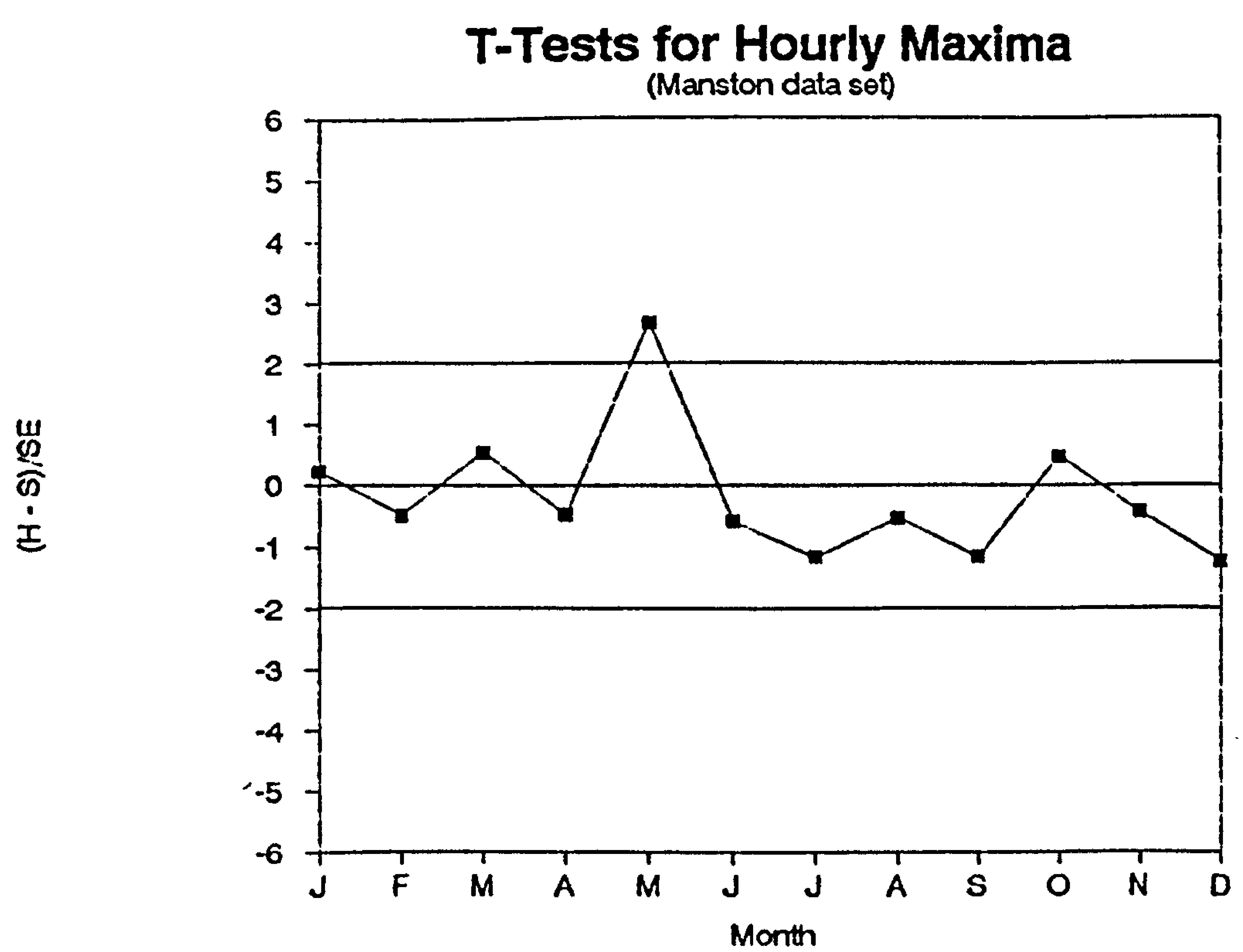


Figure C.4

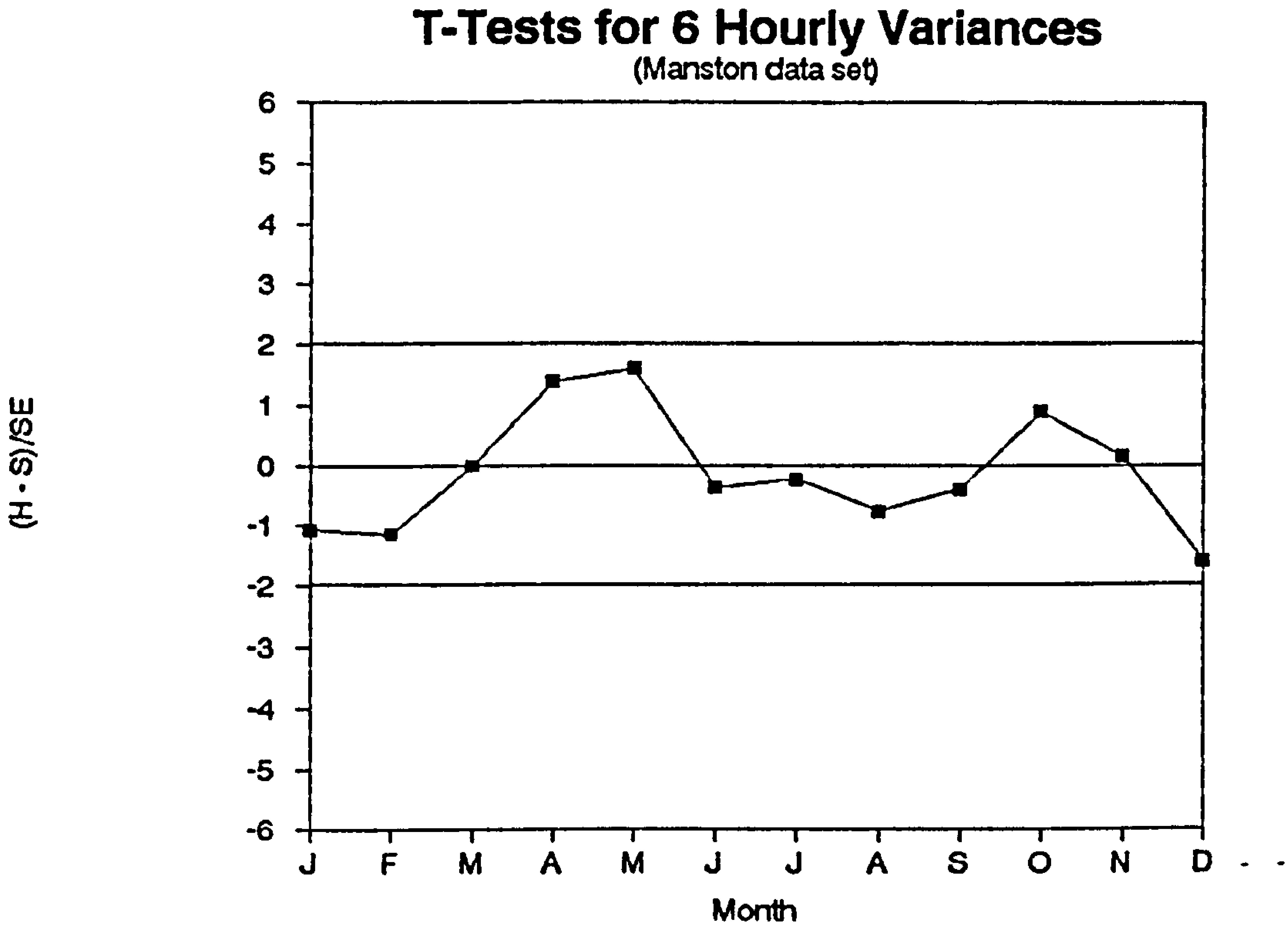


Figure C.5

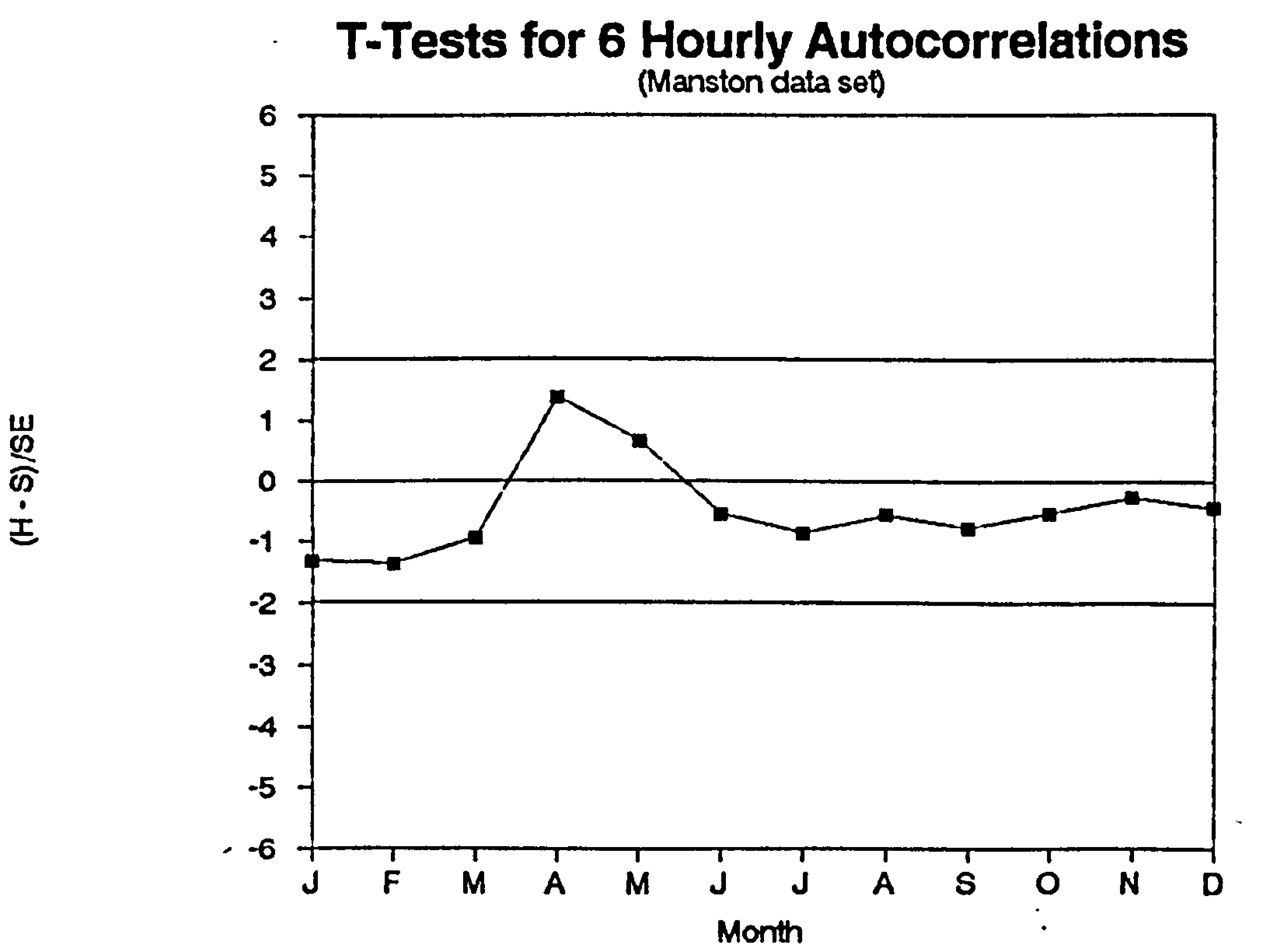


Figure C.6

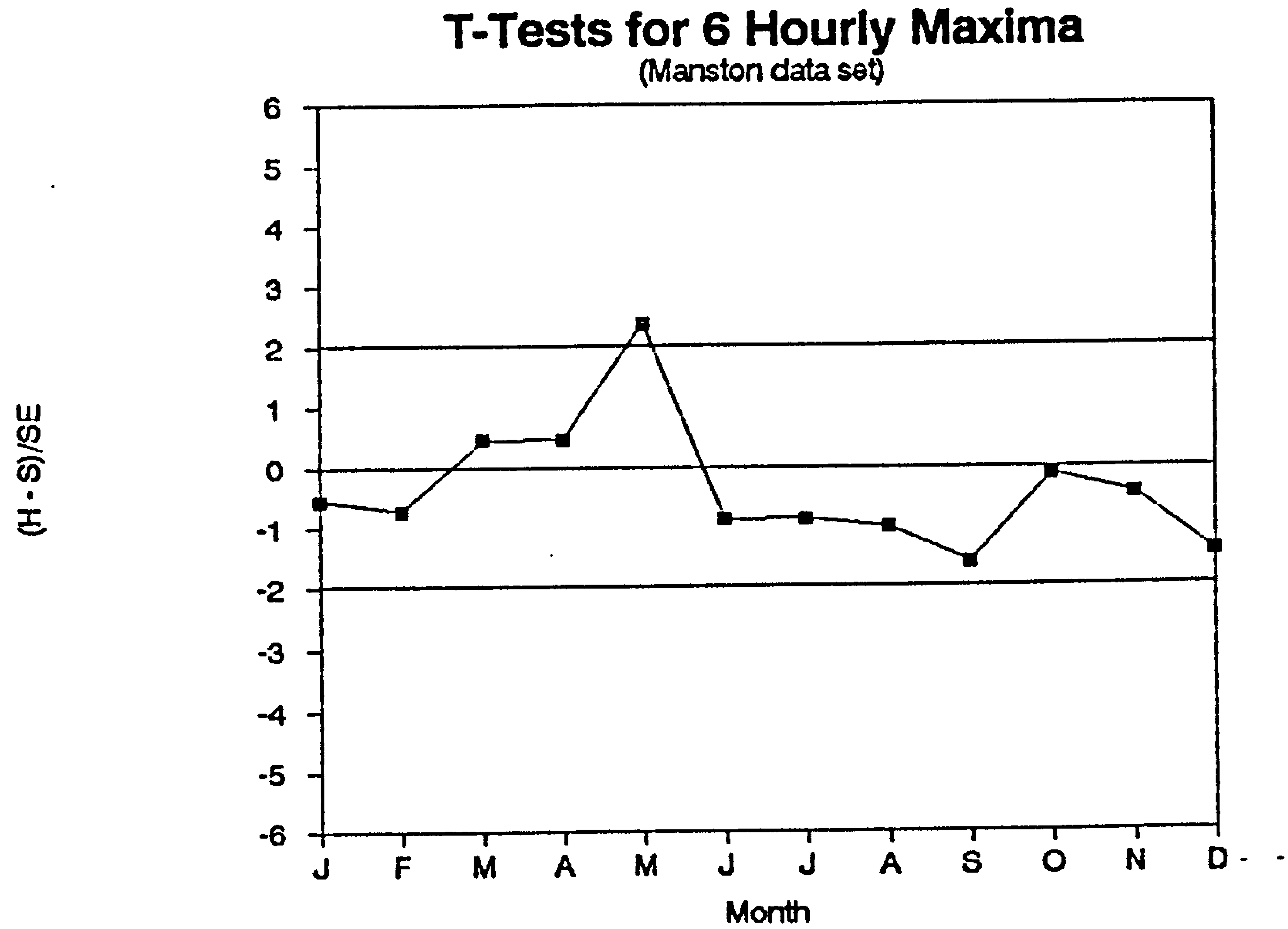


Figure C.7

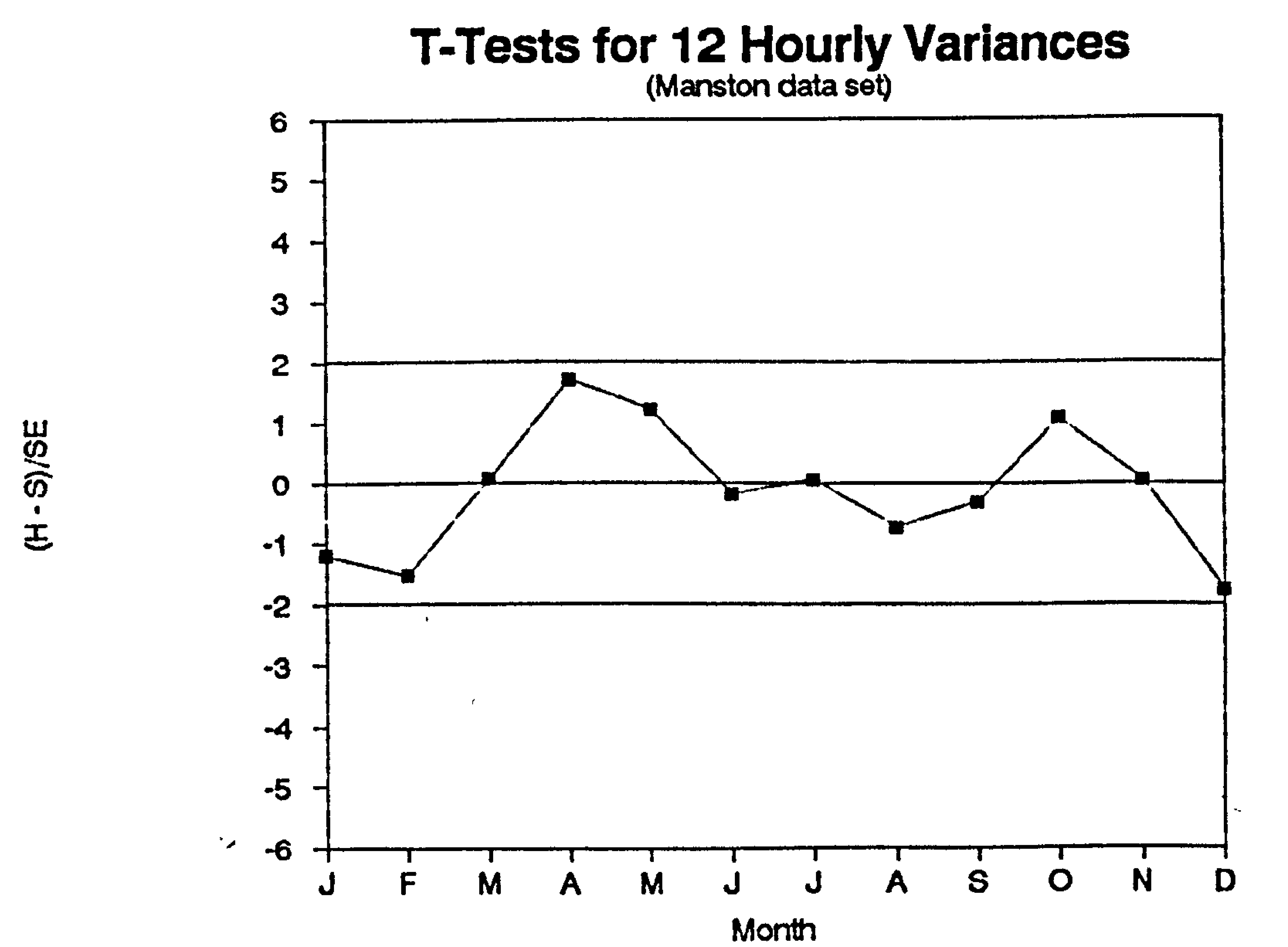


Figure C.8

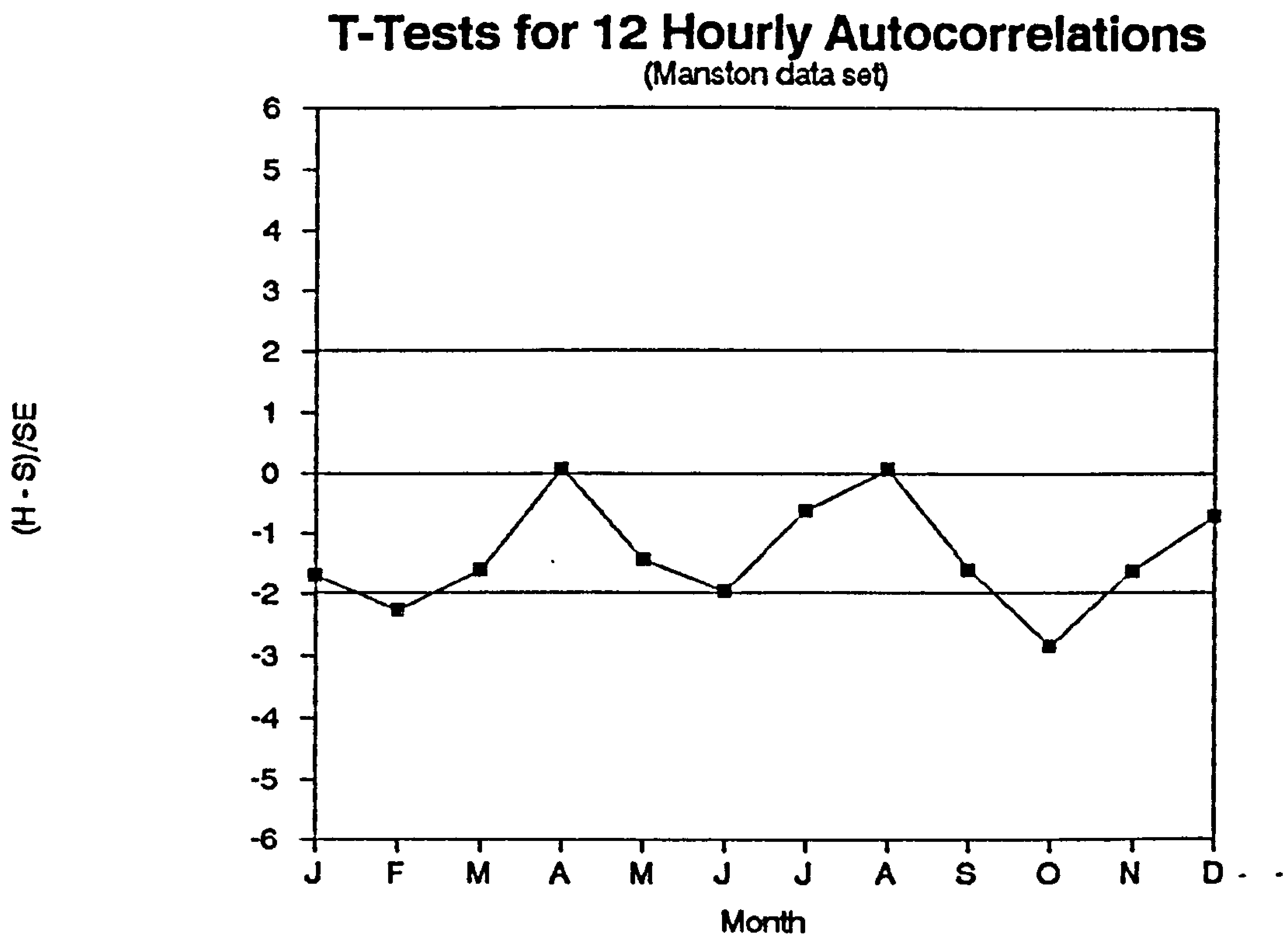


Figure C.9

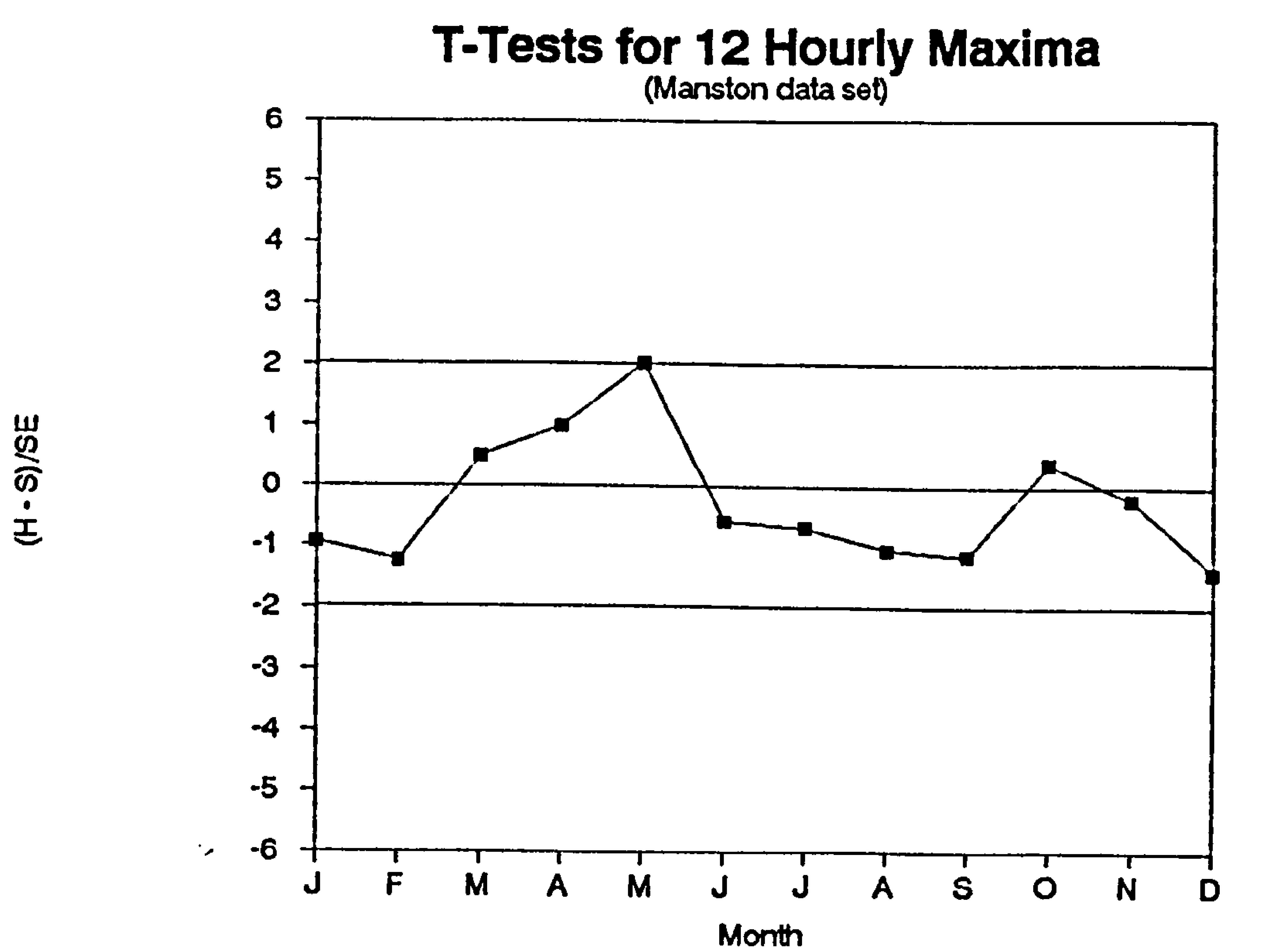


Figure C.10

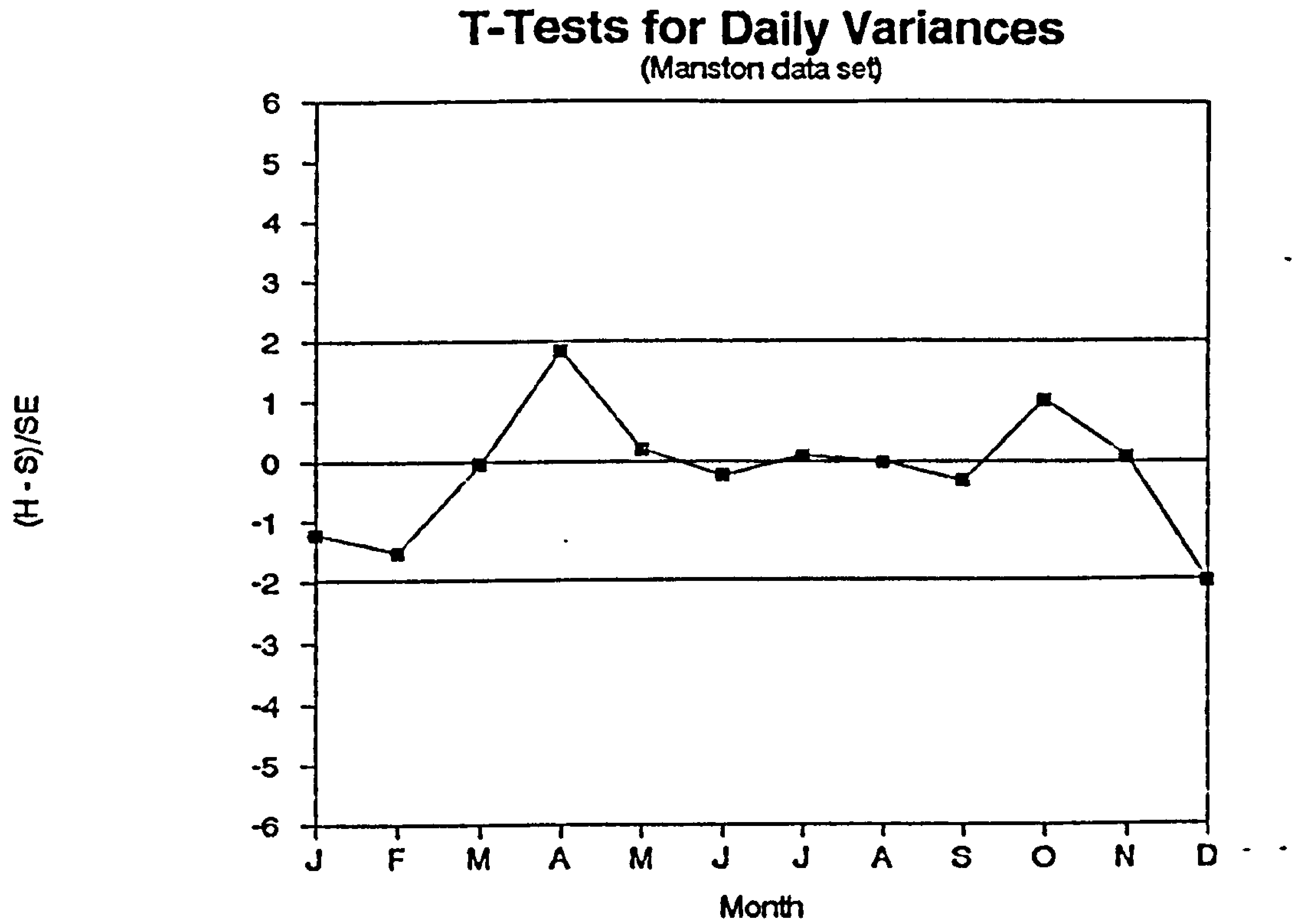


Figure C.11

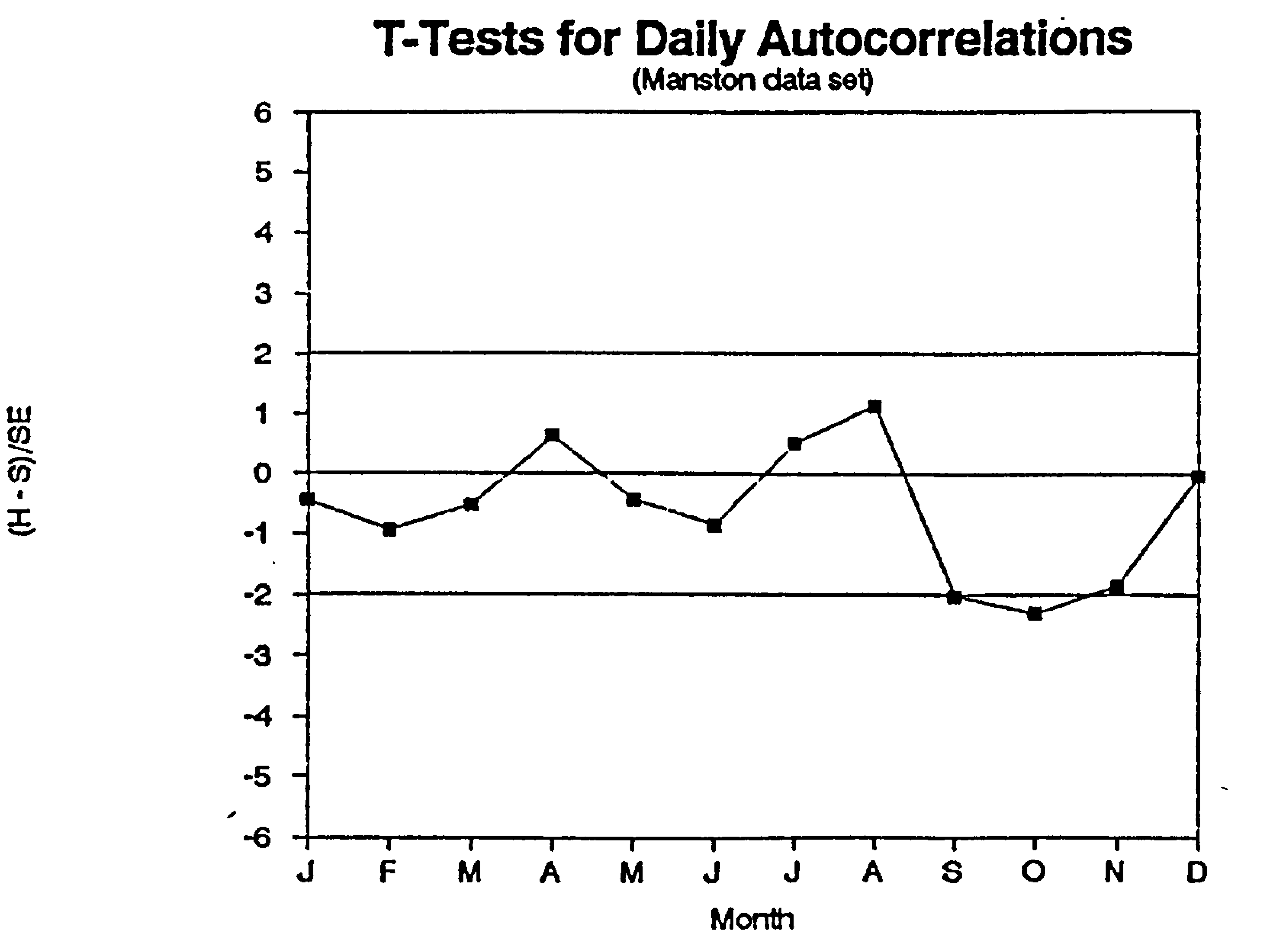


Figure C.12

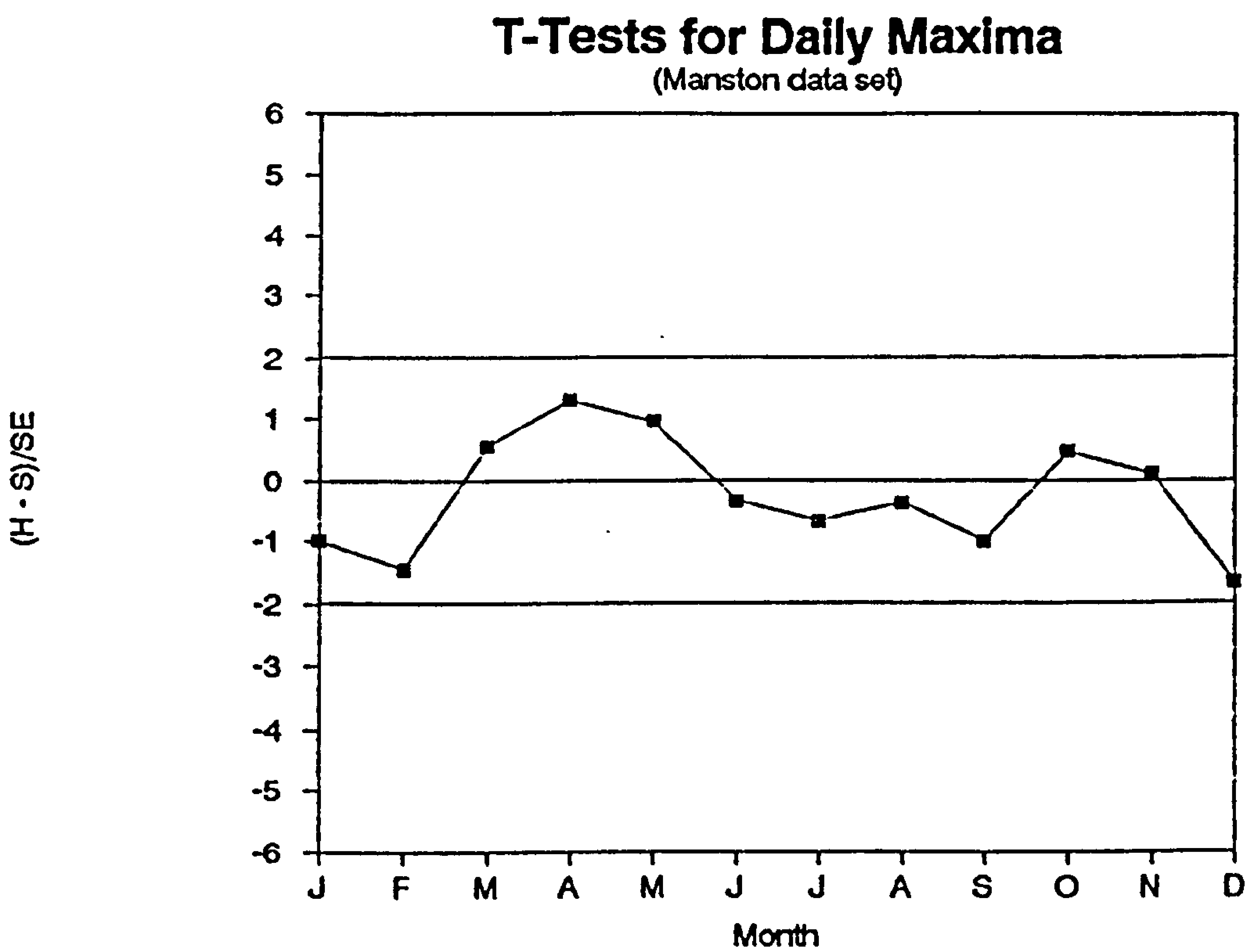


Figure C.13

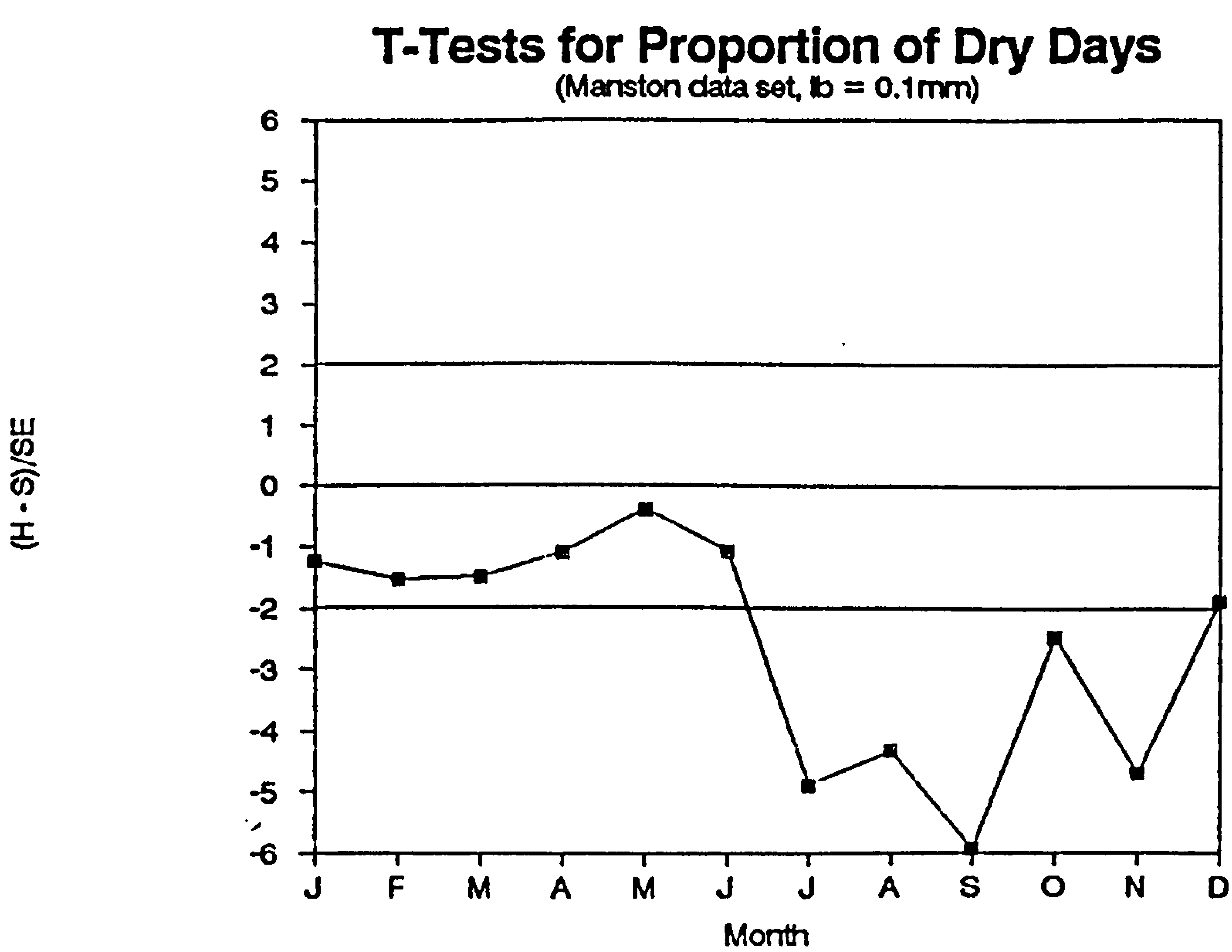


Figure C.14

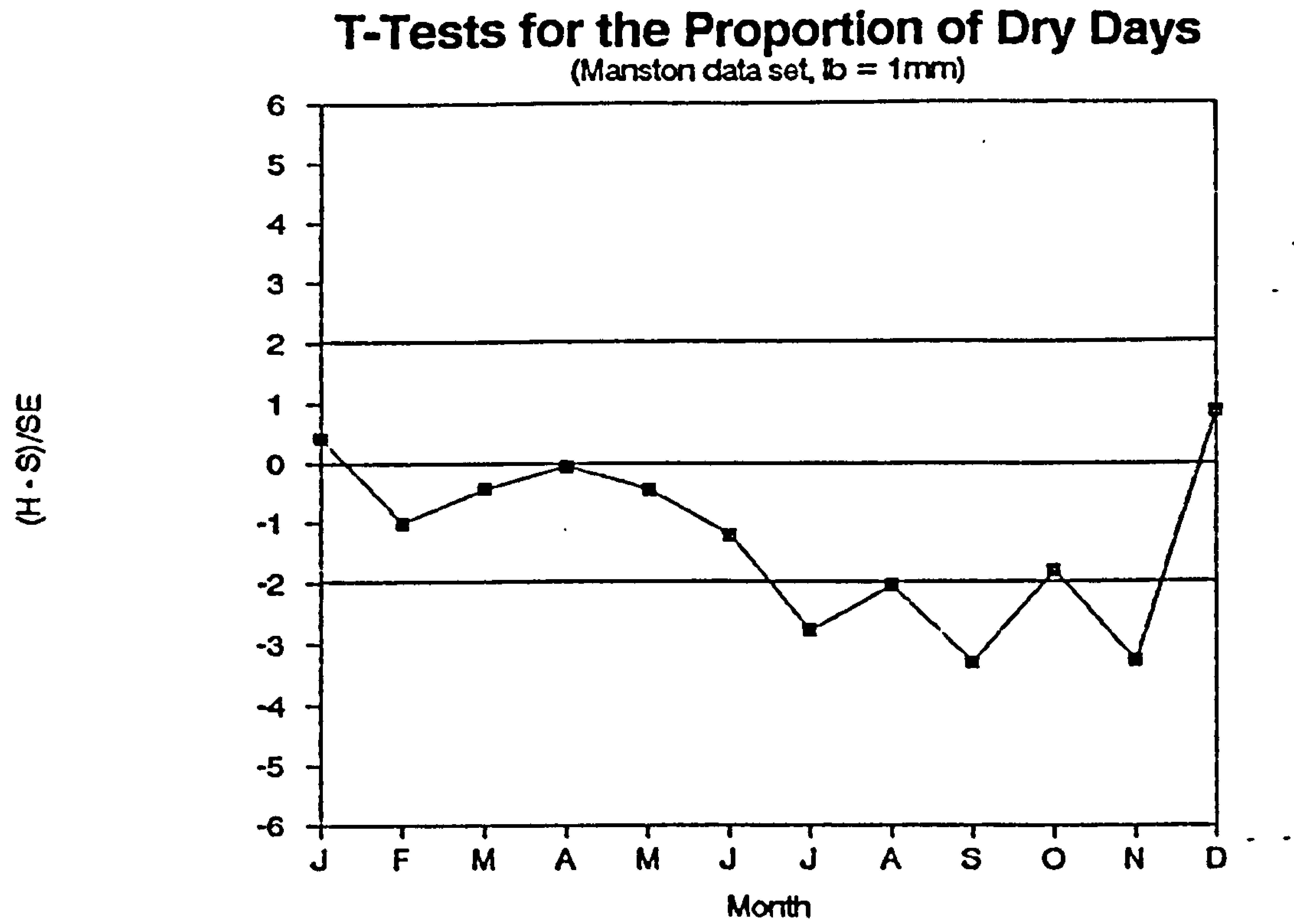


Figure C.15

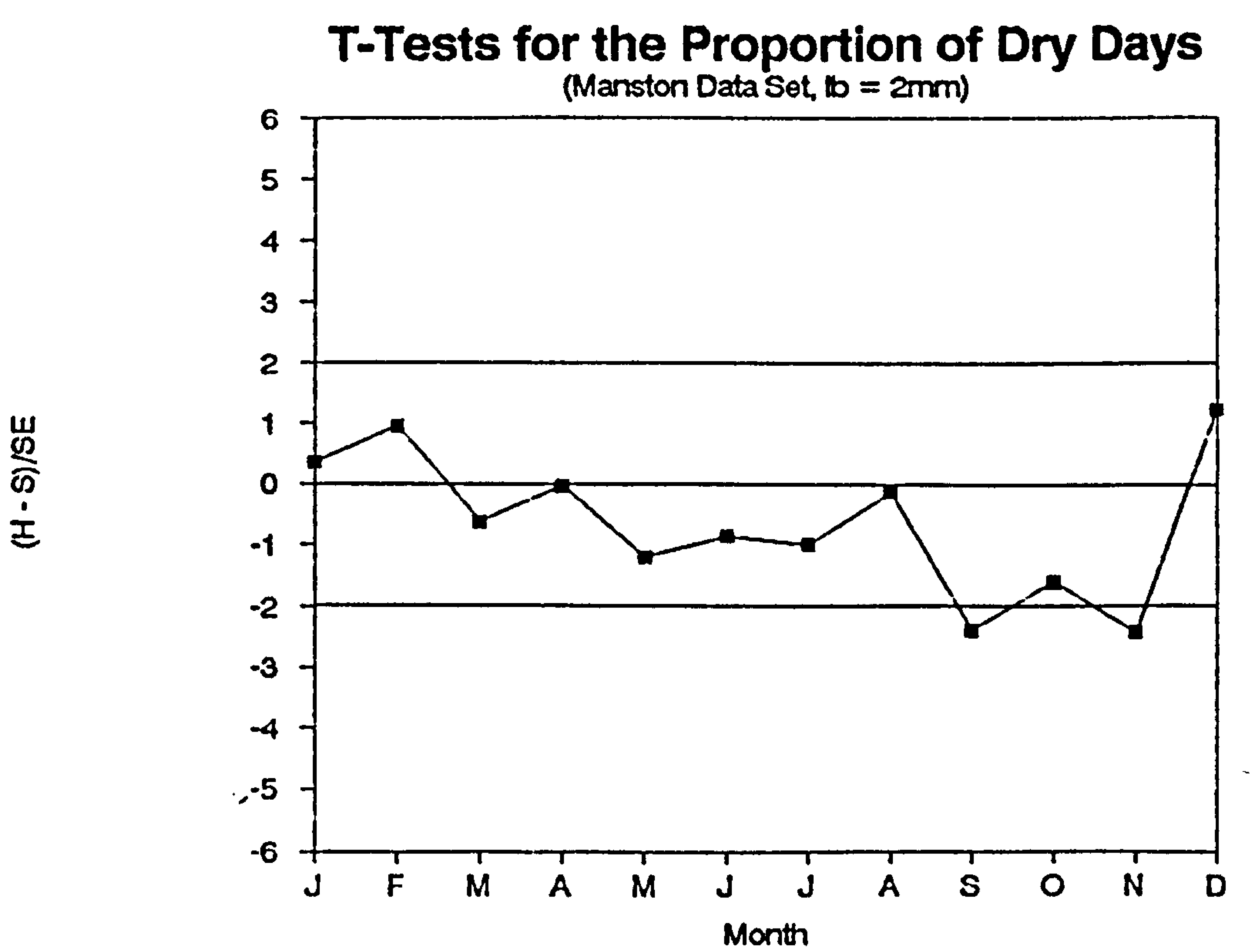
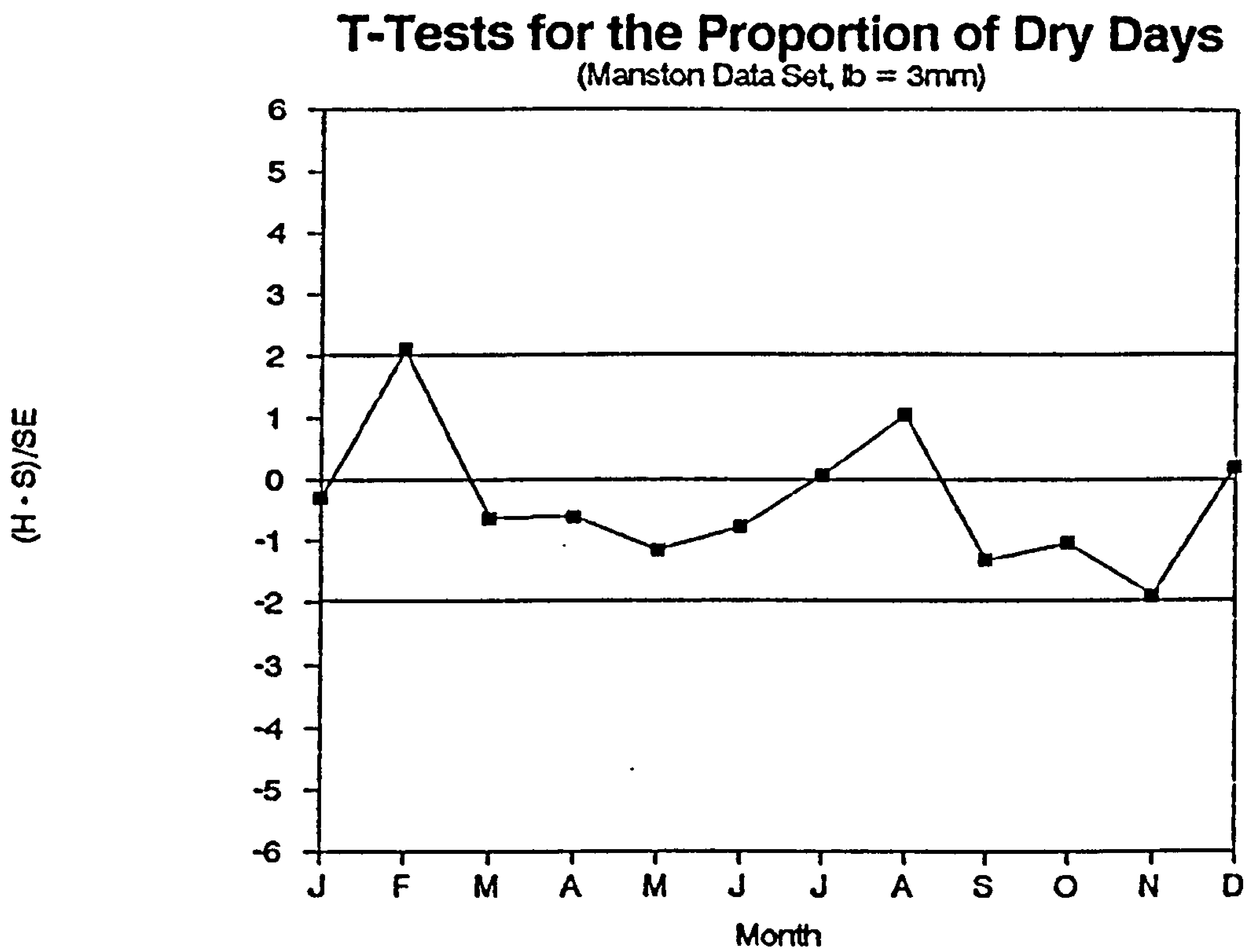


Figure C.16



T-Tests for the Proportion of Dry Days
(Manston Data Set, lb = 3mm)
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
(H - S)/SE
J
F
M
A
M
J
J
A
S
O
N
D
Month

Figure C.17

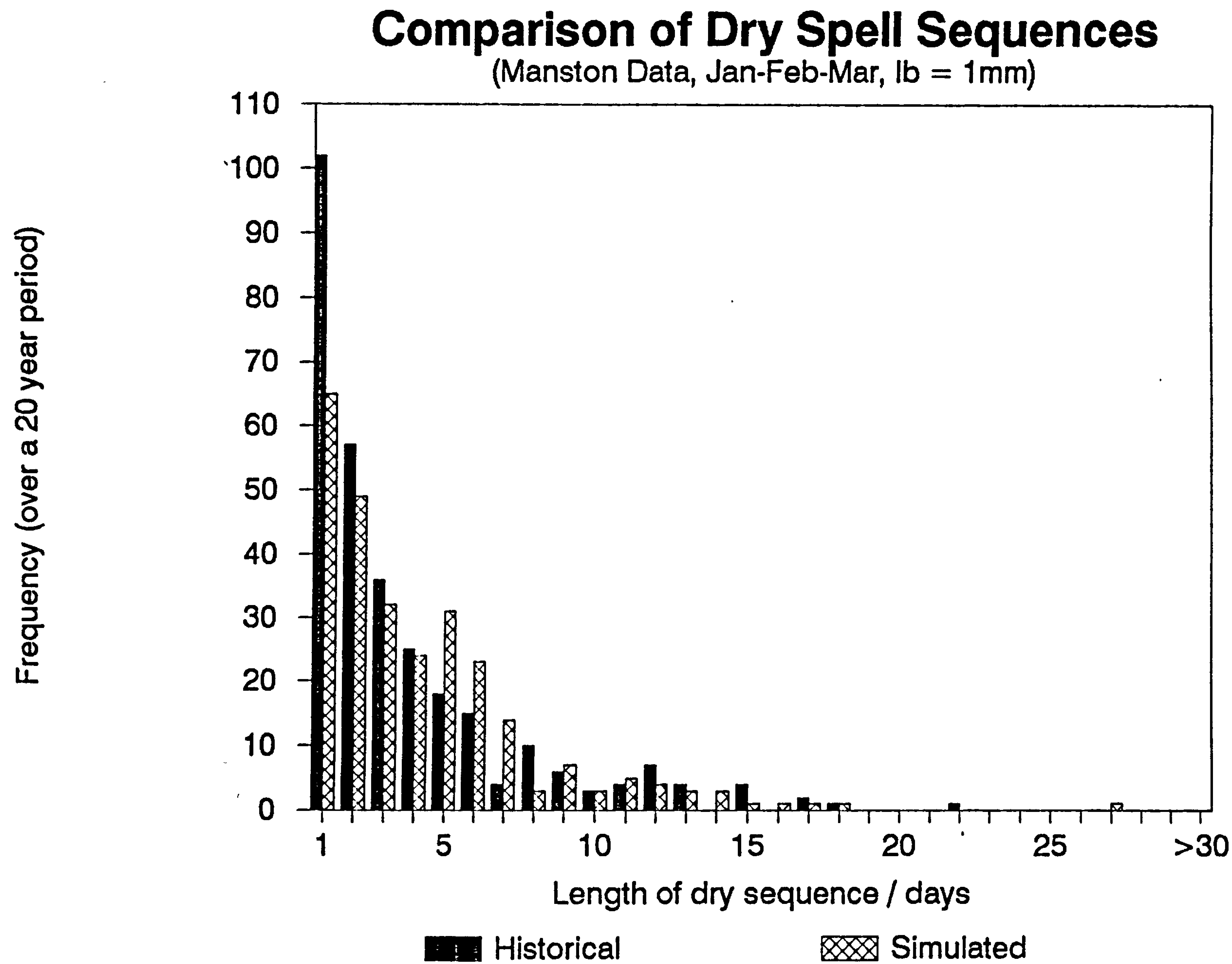

Comparison of Dry Spell Sequences
(Manston Data, Jan-Feb-Mar, lb = 1mm)
Frequency (over a 20 year period)
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of dry sequence / days
Historical
Simulated

Figure C.18

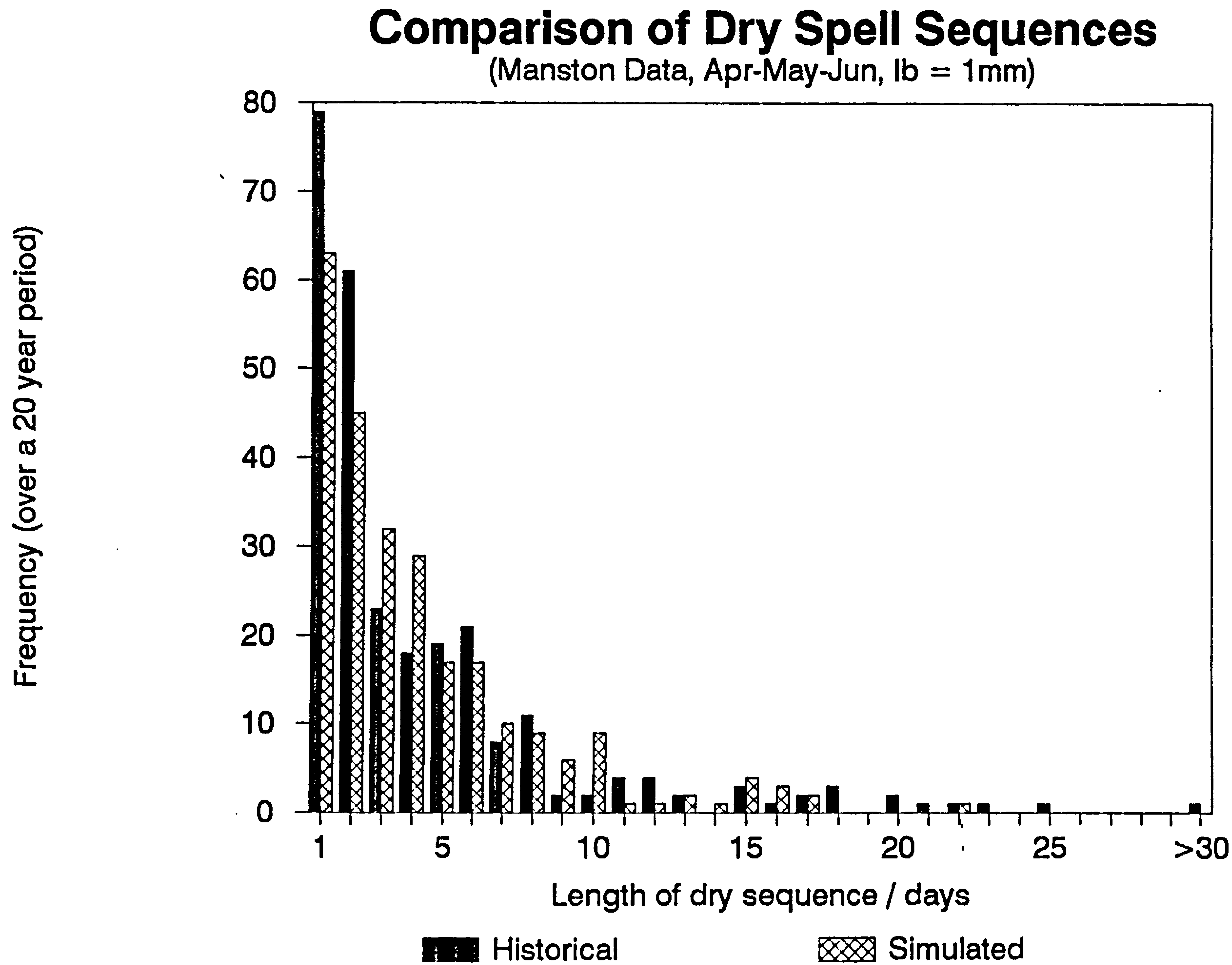

Comparison of Dry Spell Sequences
(Manston Data, Apr-May-Jun, lb = 1mm)
Frequency (over a 20 year period)
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of dry sequence / days
Historical
Simulated

Figure C.19

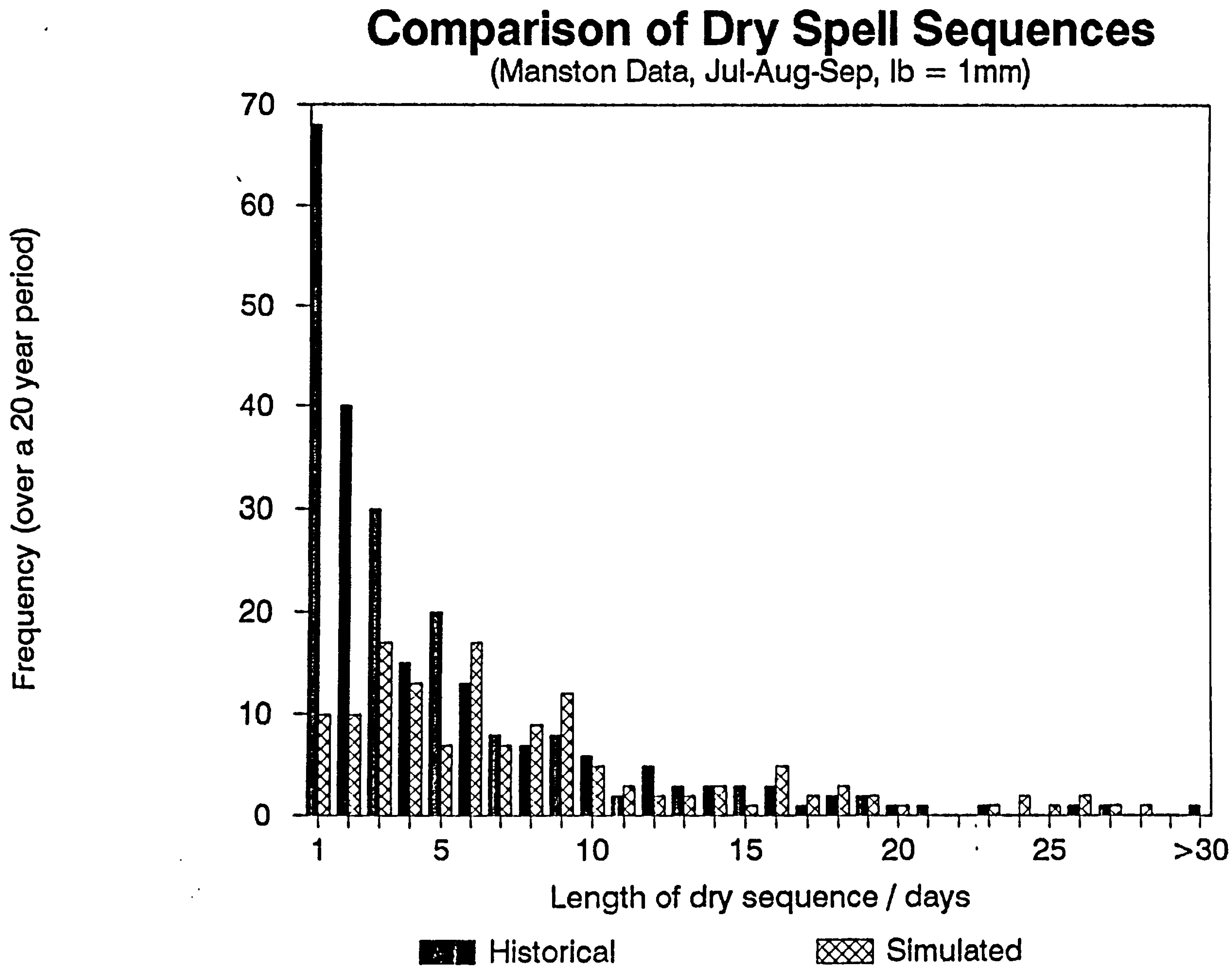
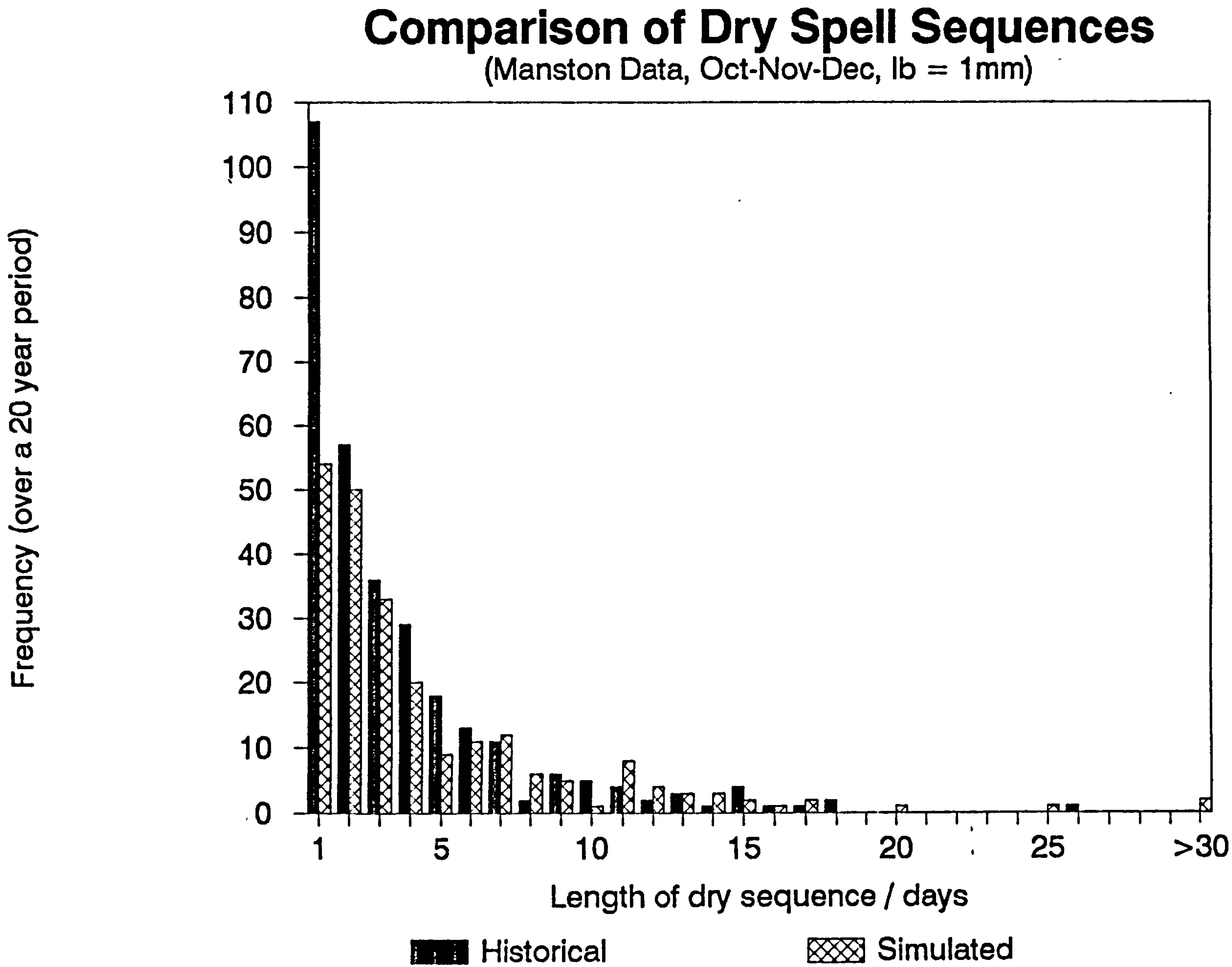

Comparison of Dry Spell Sequences
(Manston Data, Jul-Aug-Sep, lb = 1mm)
Frequency (over a 20 year period)
0
10
20
30
40
50
60
70
1
5
10
15
20
25
>30
Length of dry sequence / days
Historical
Simulated

Figure C.20



Comparison of Dry Spell Sequences
(Manston Data, Oct-Nov-Dec, lb = 1mm)
Frequency (over a 20 year period)
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of dry sequence / days
Historical
Simulated

Figure C.21

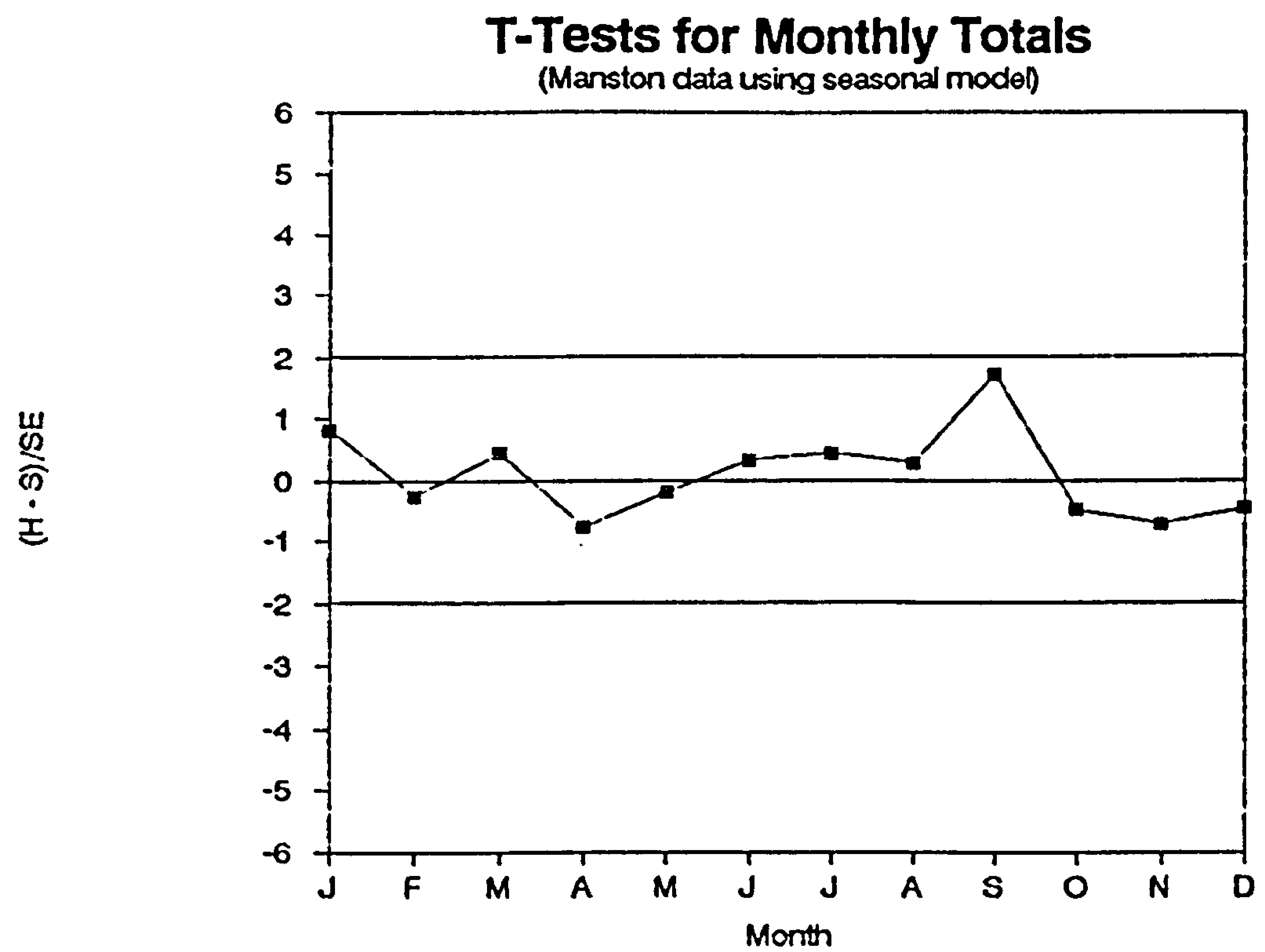


Figure D.1

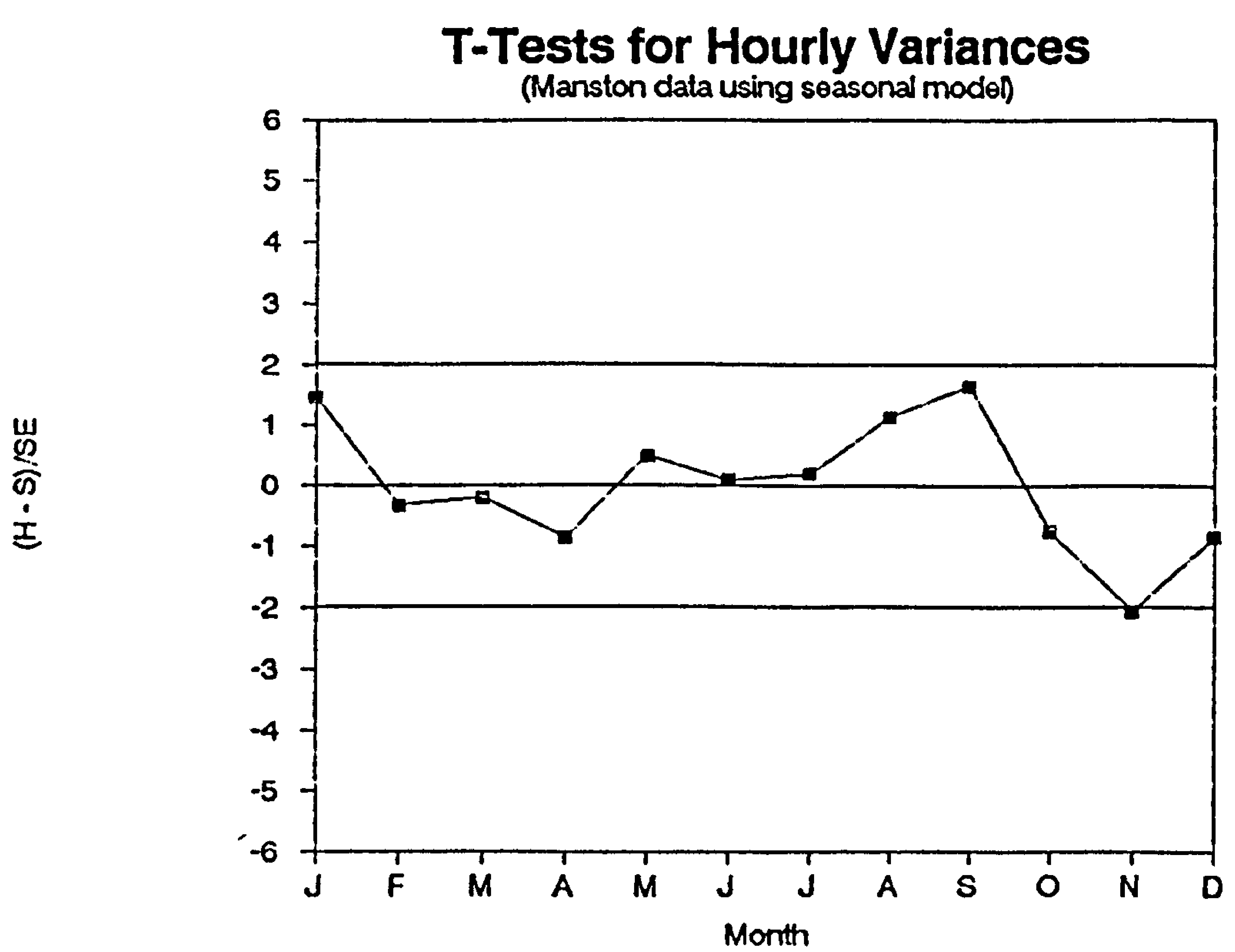


Figure D.2

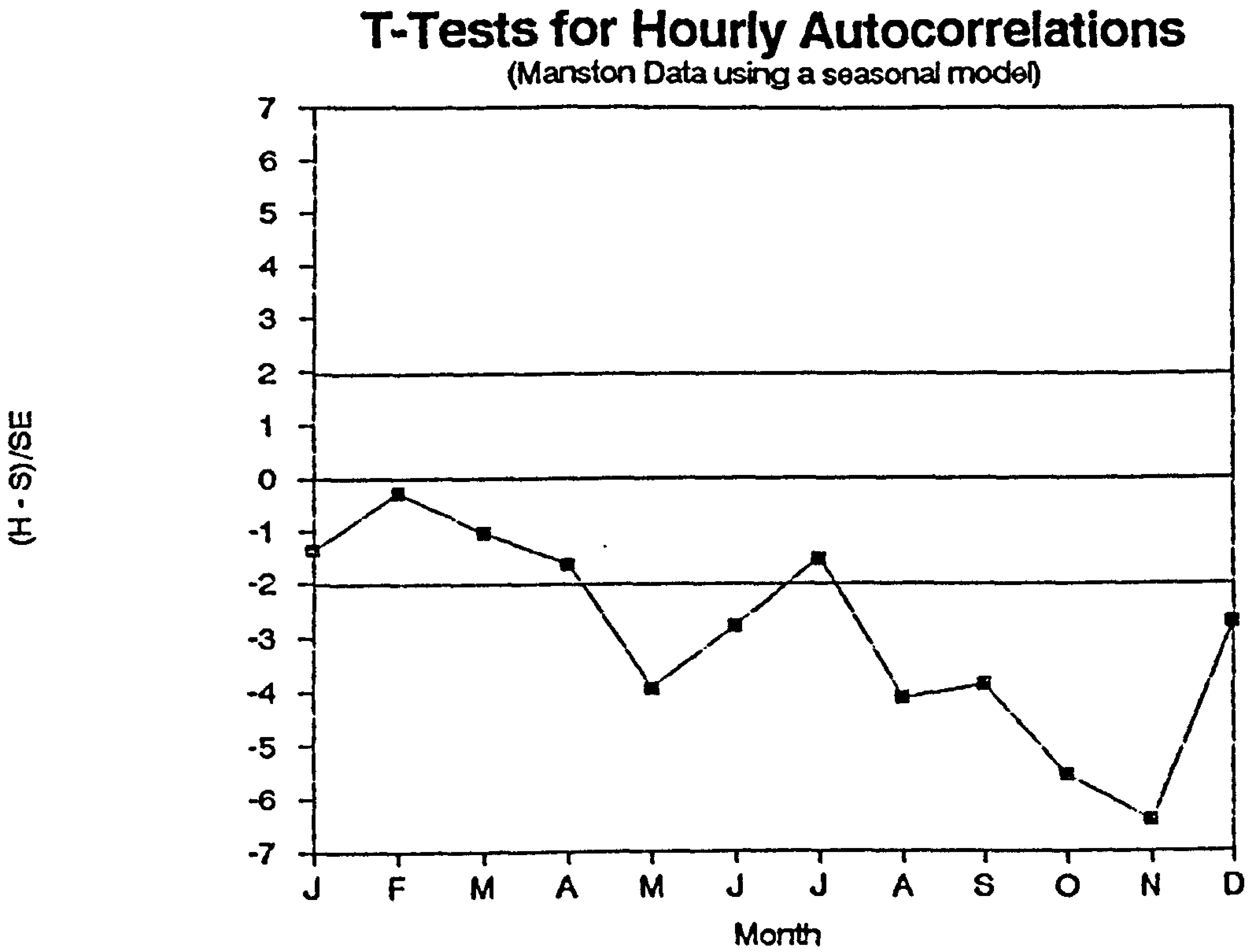


Figure D.3

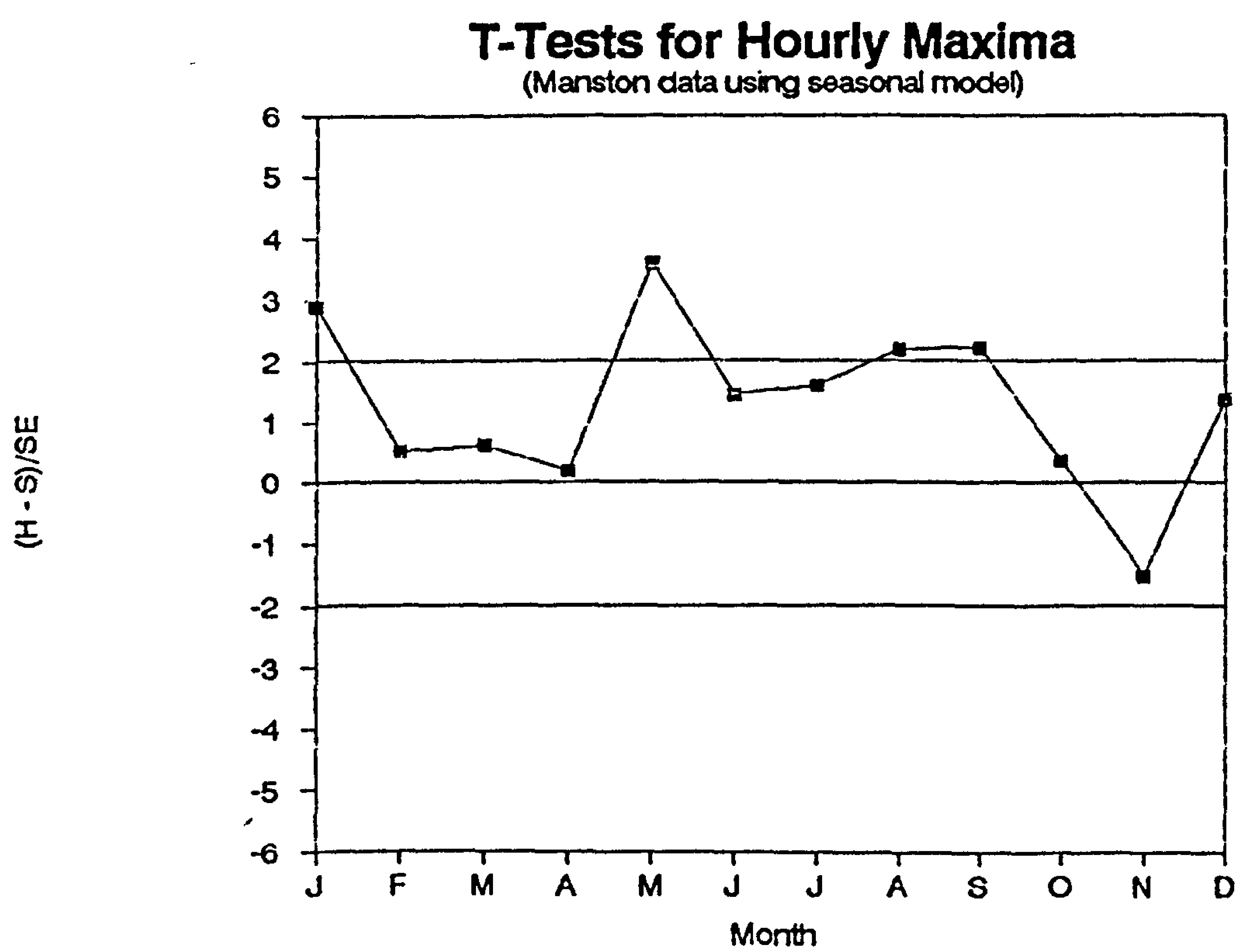


Figure D.4

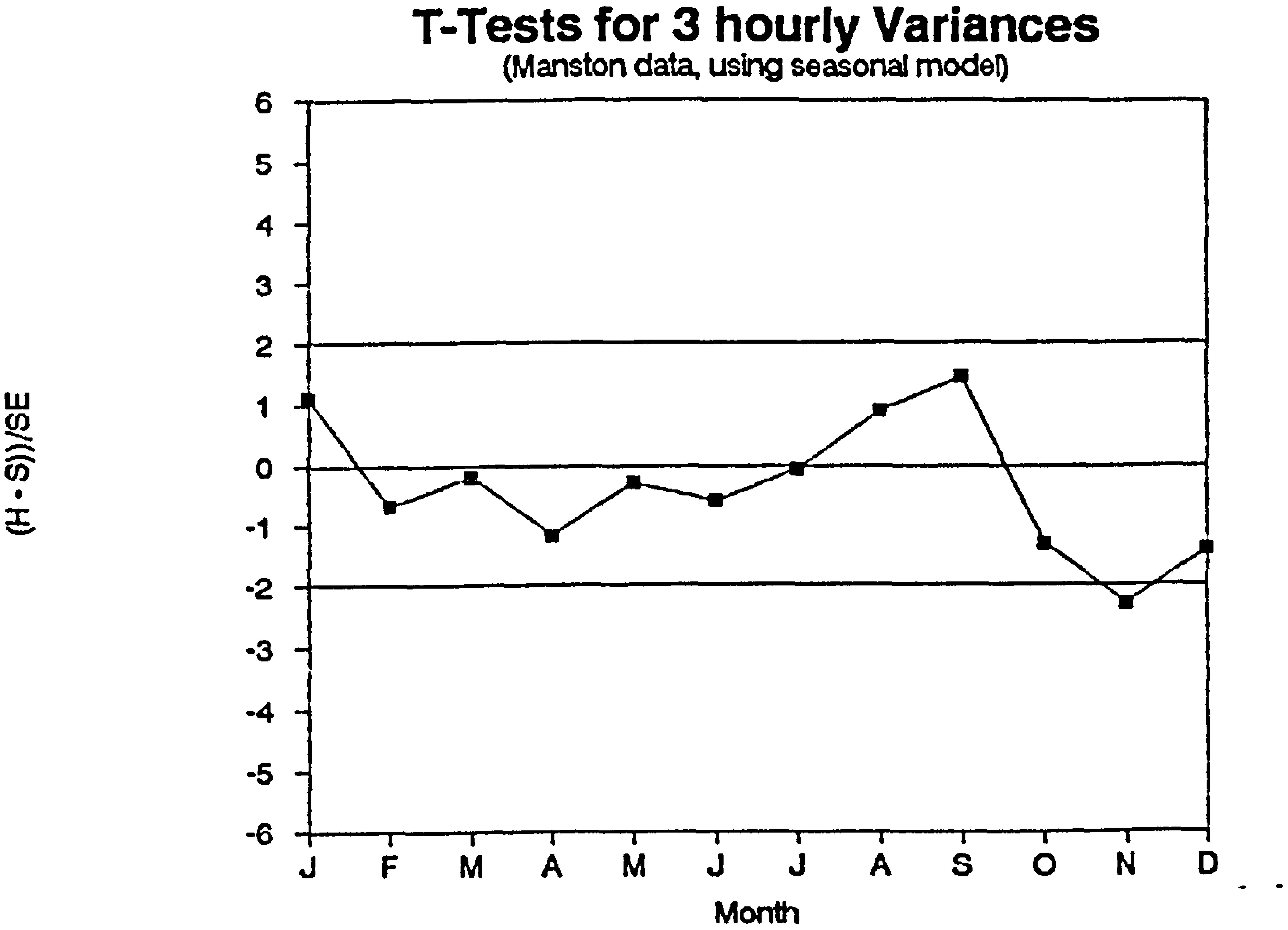


Figure D.5

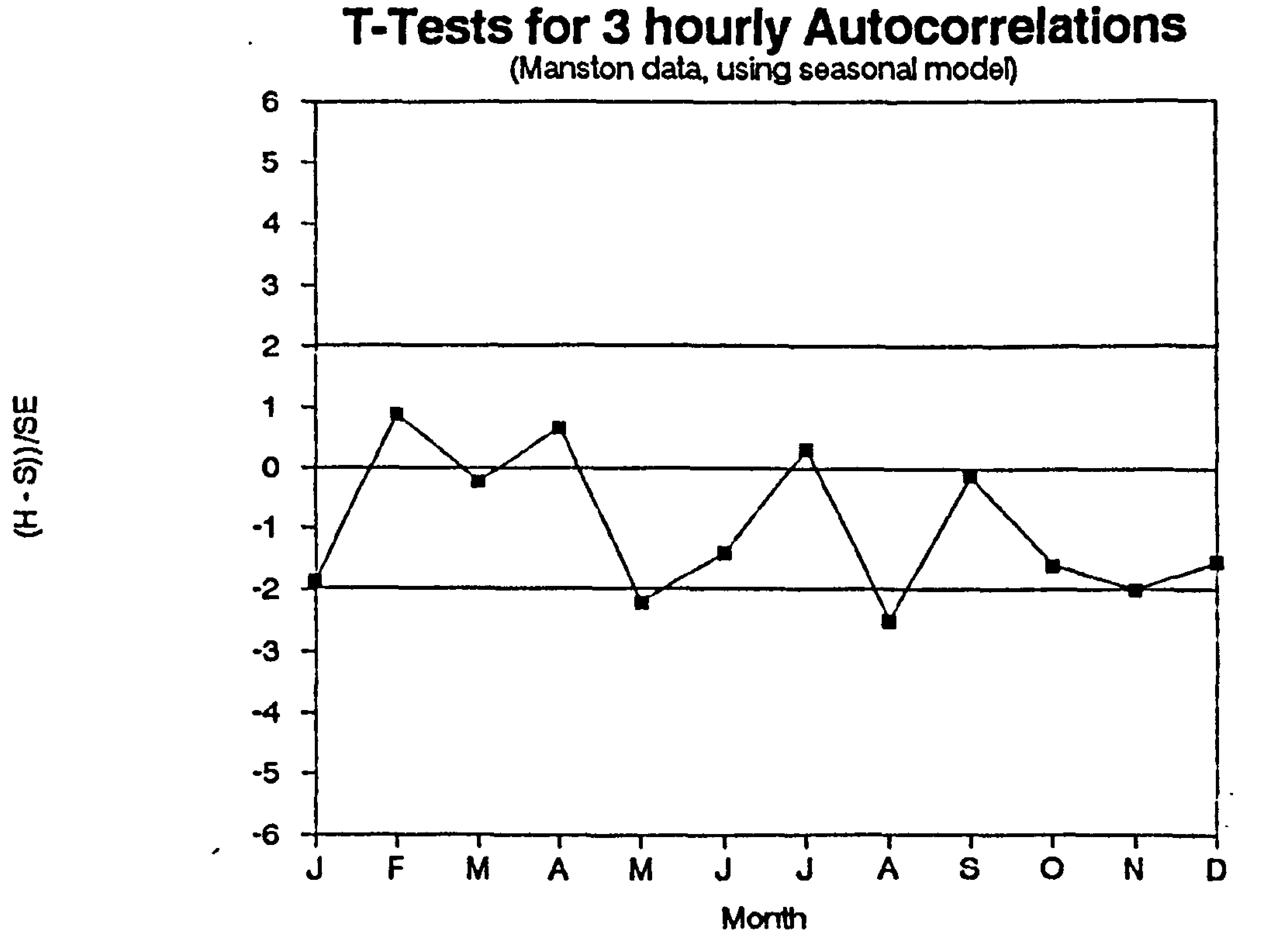


Figure D.6

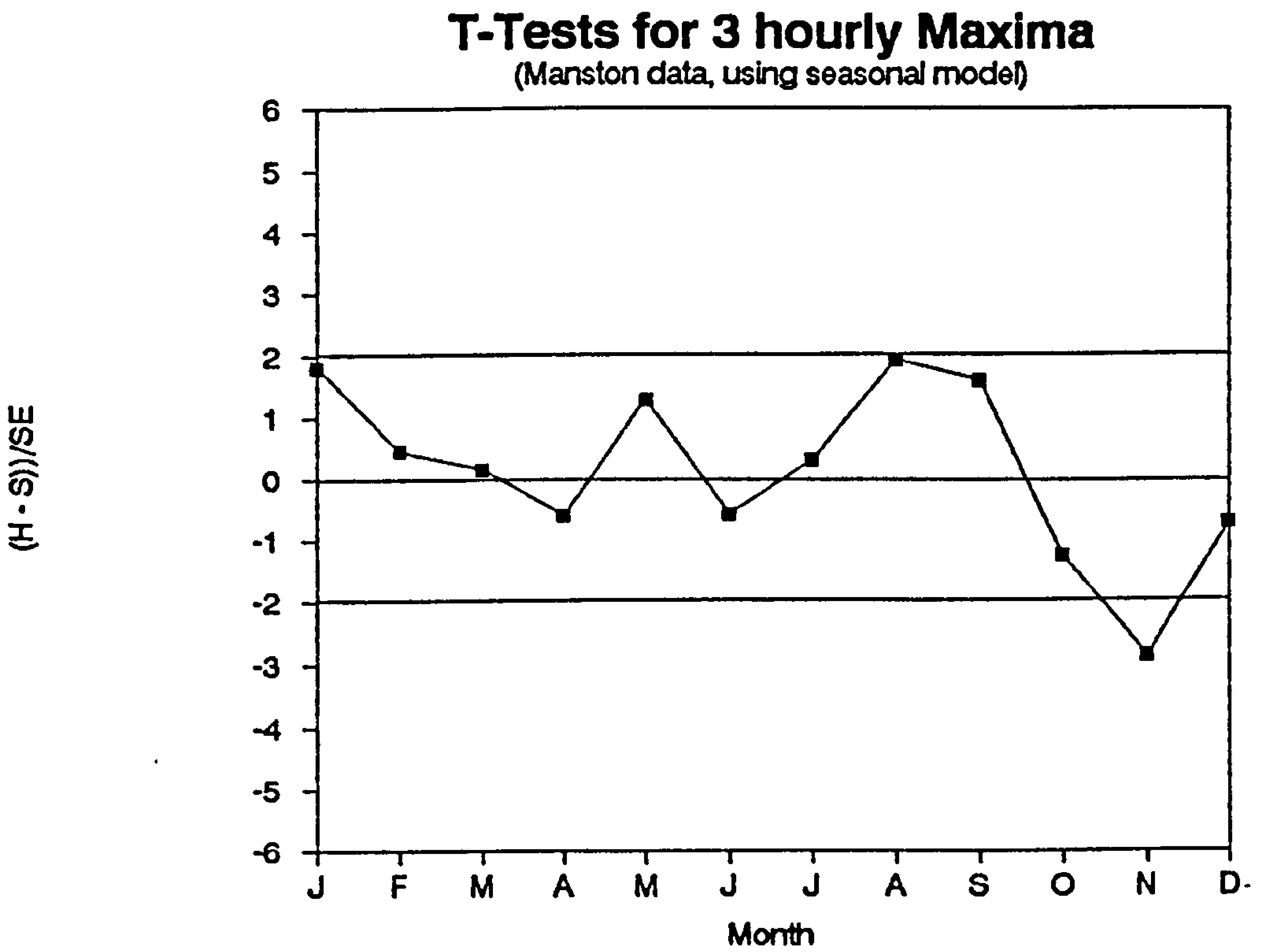


Figure D.7

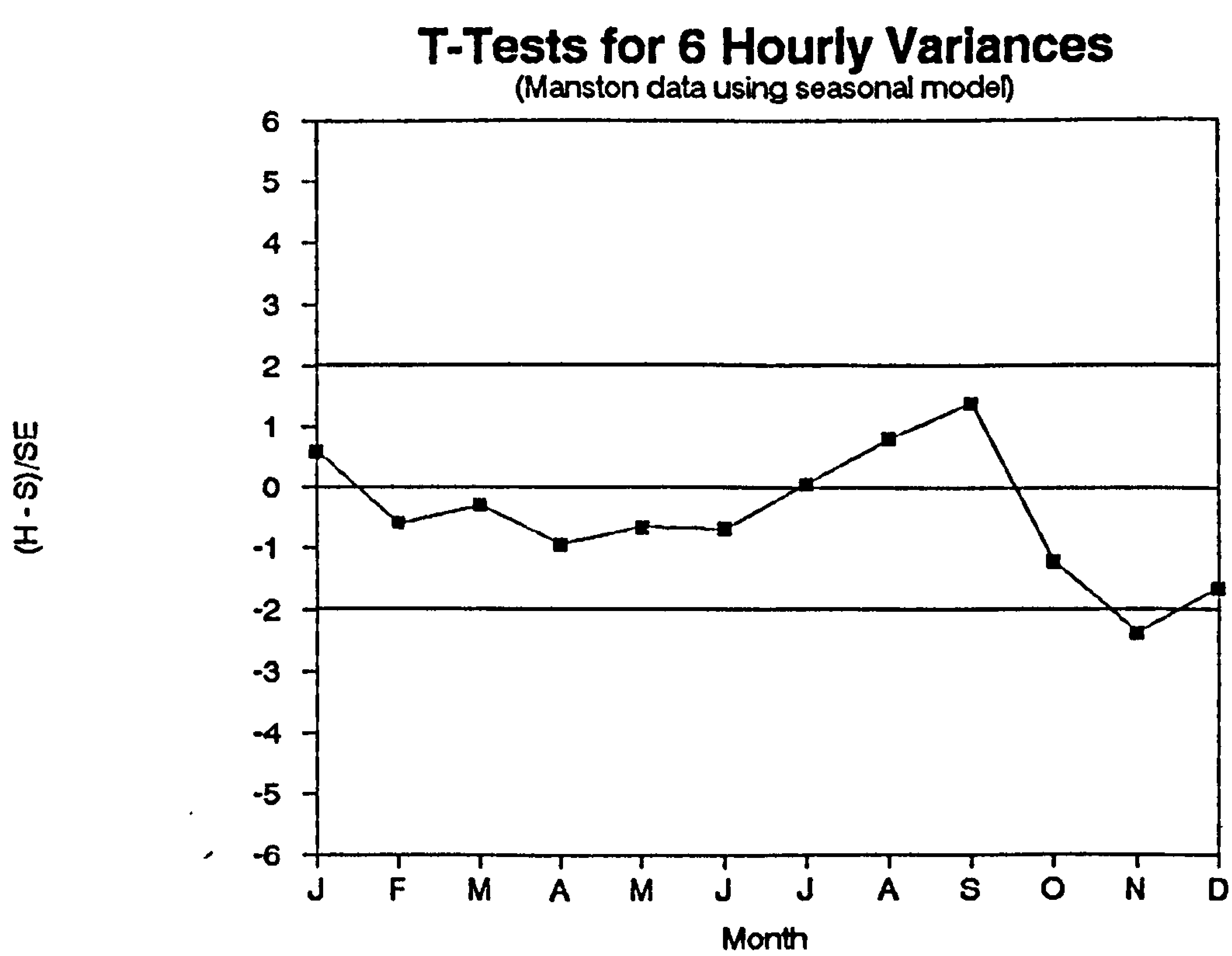


Figure D.8

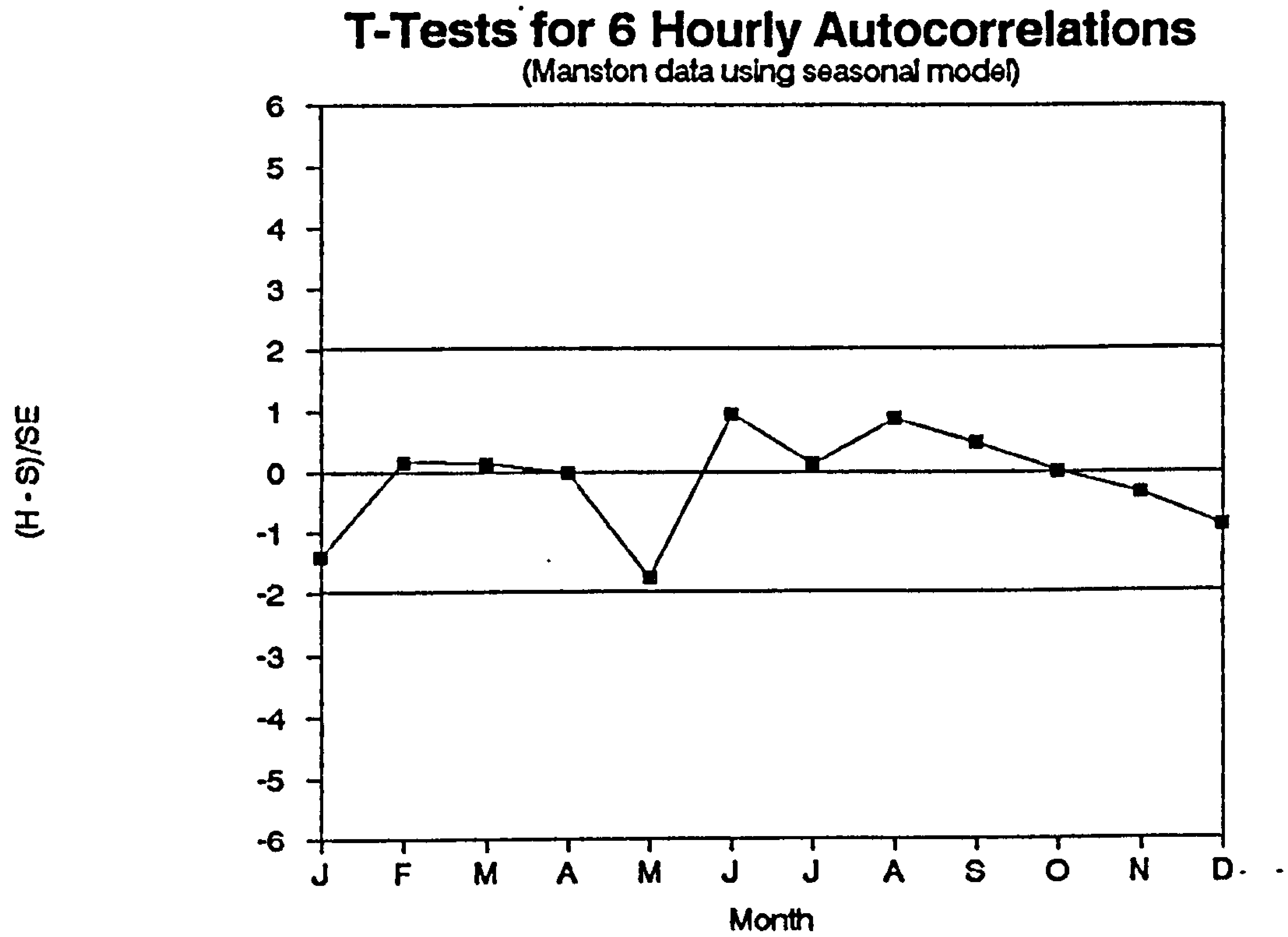


Figure D.9

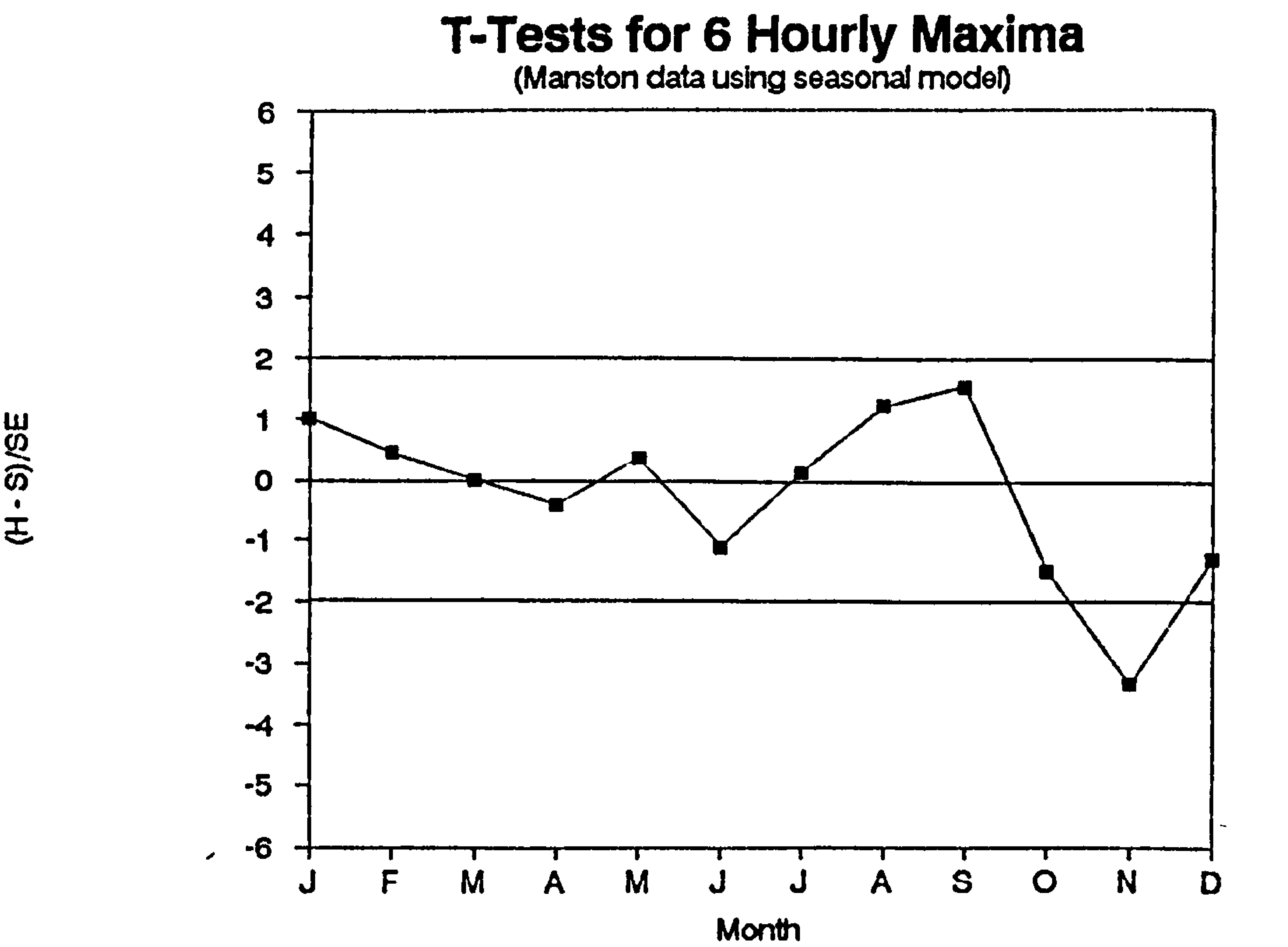


Figure D.10

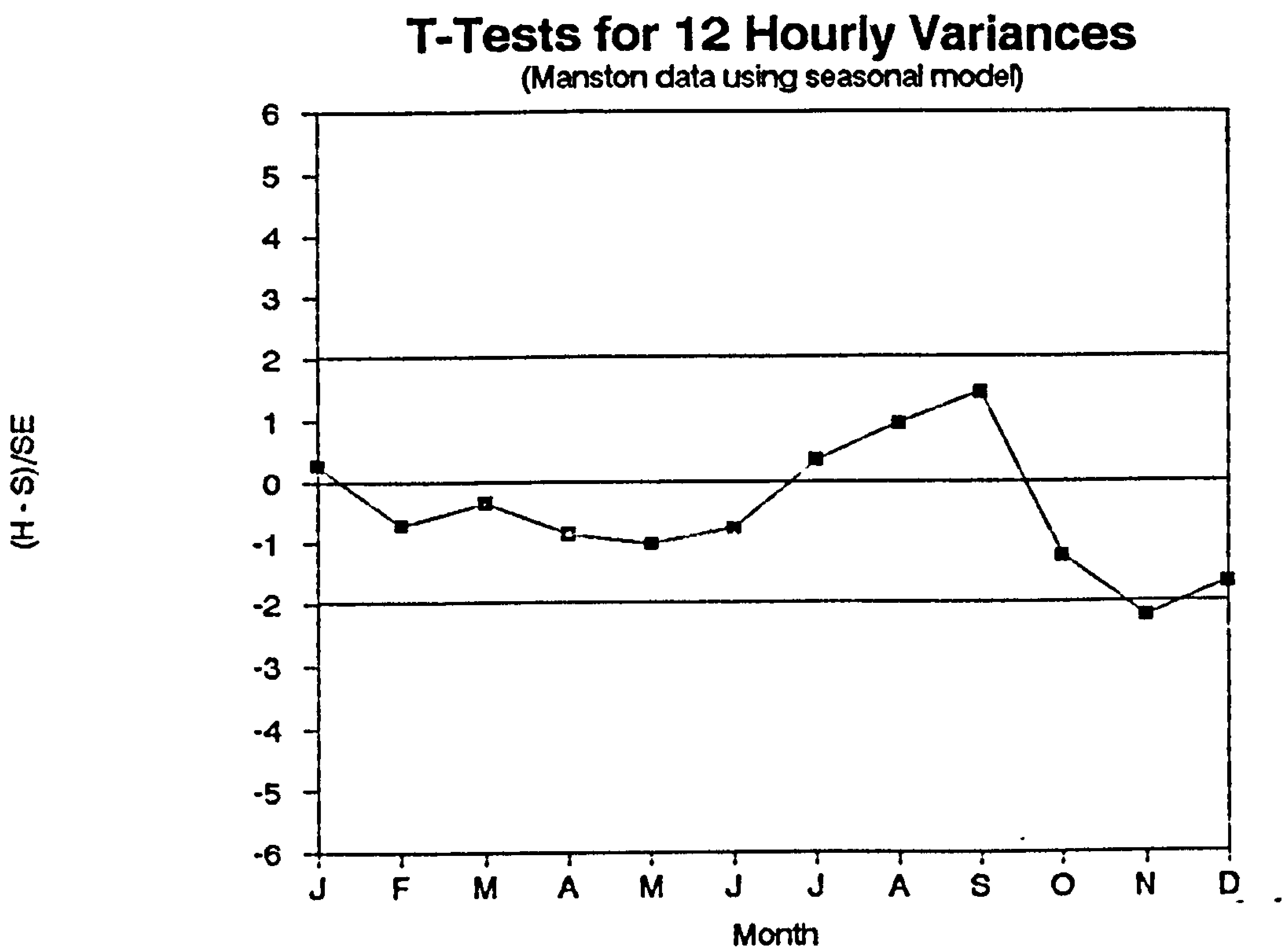


Figure D.11

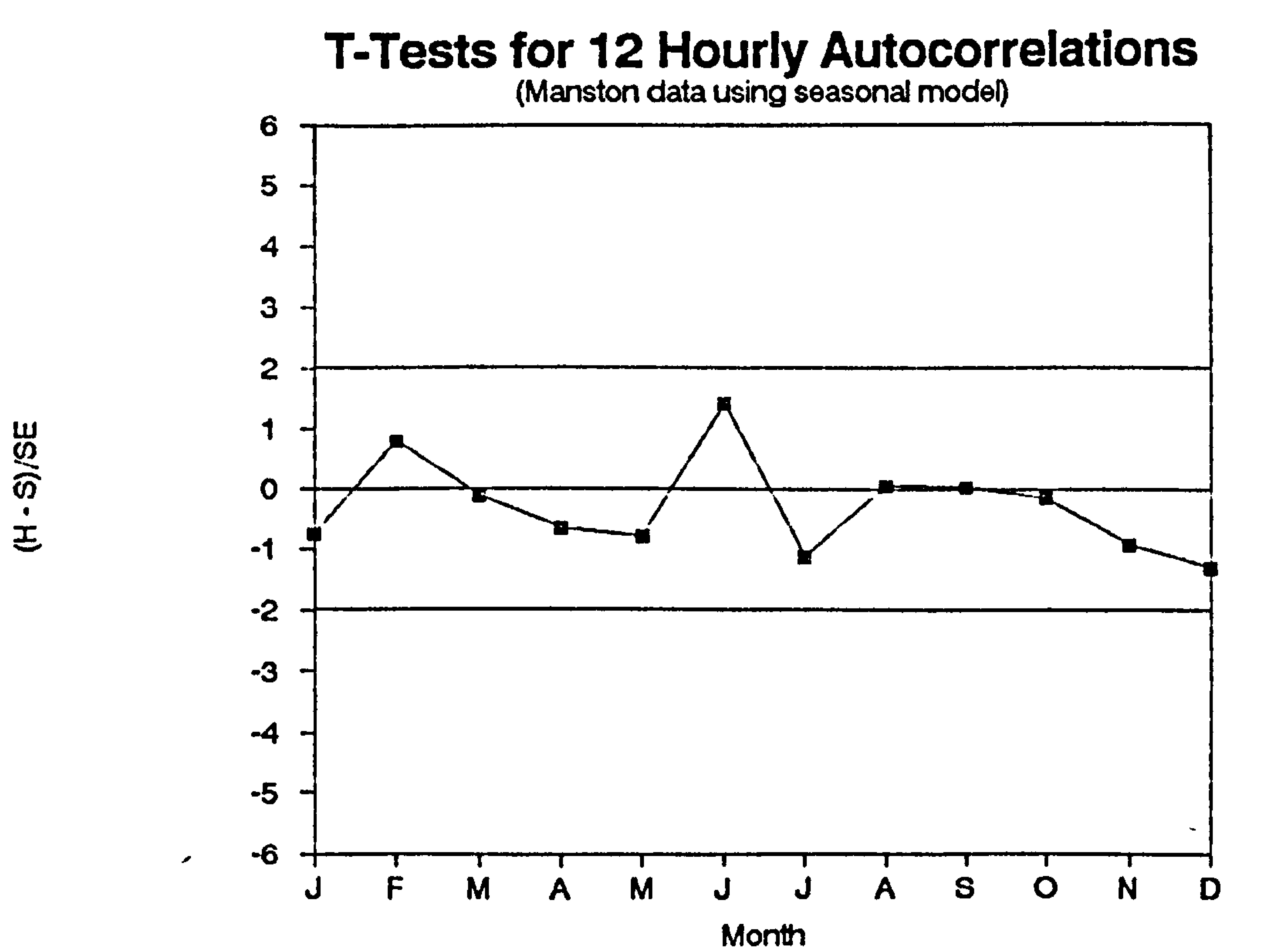


Figure D.12

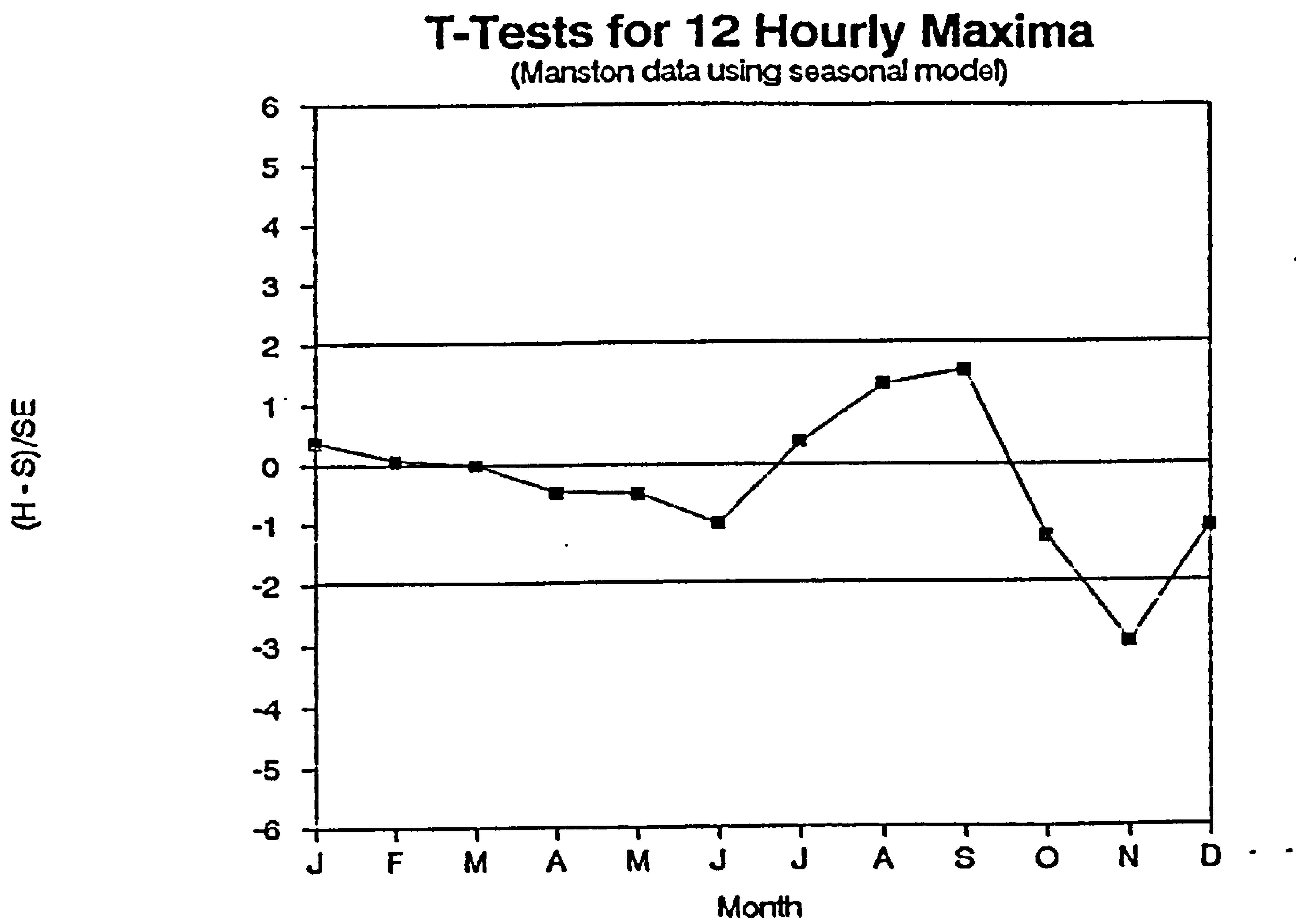


Figure D.13

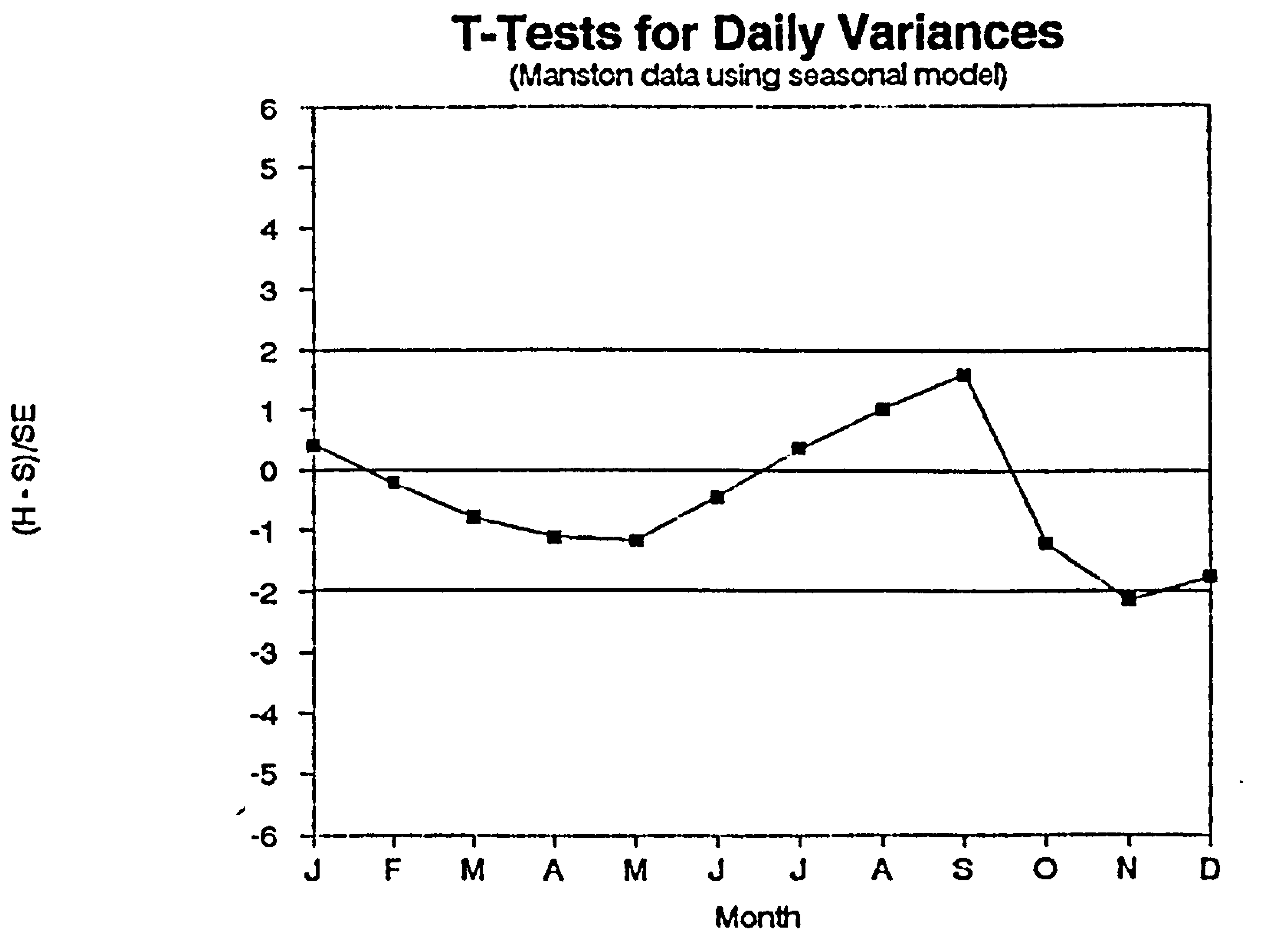


Figure D.14

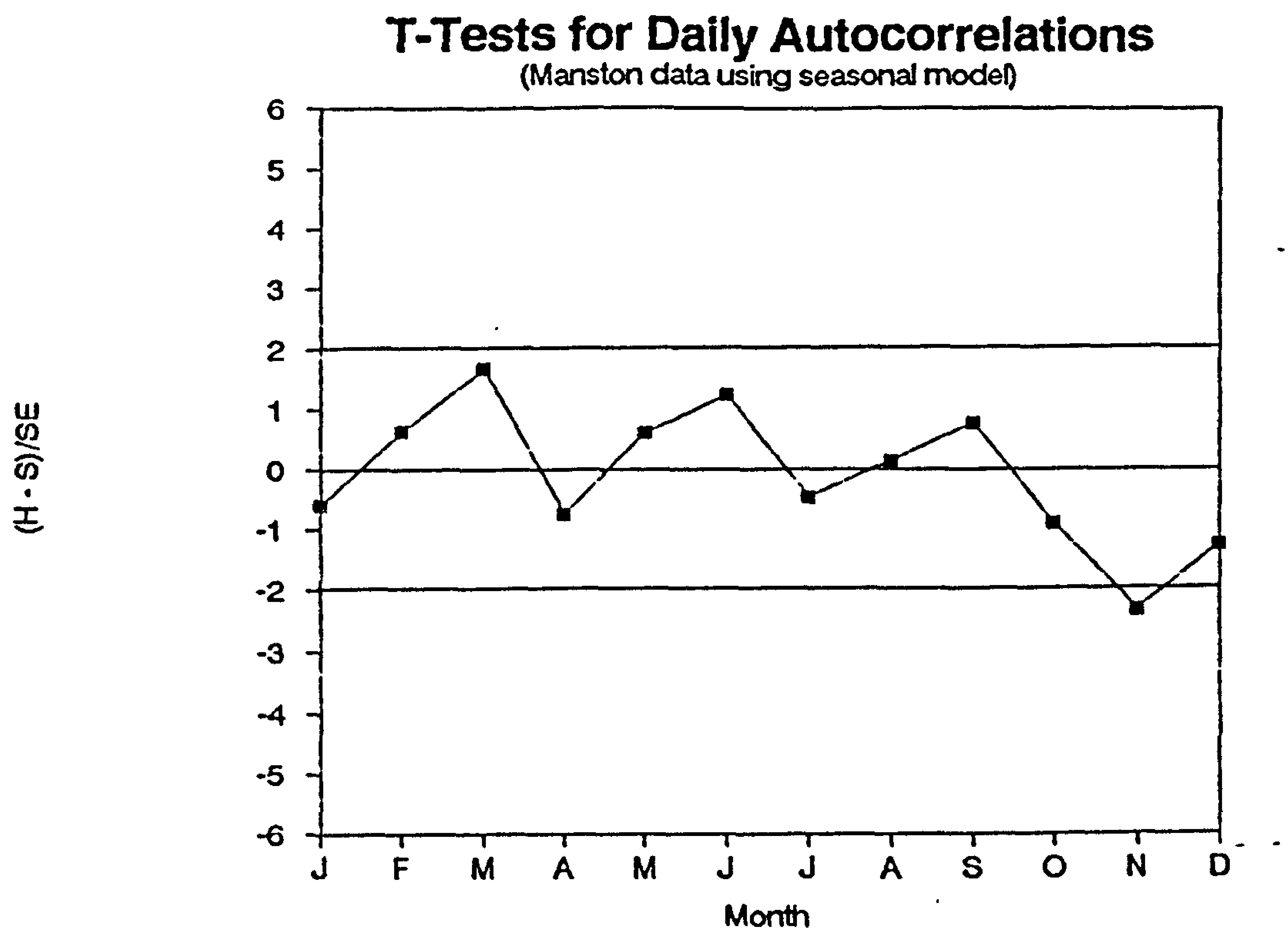


Figure D.15

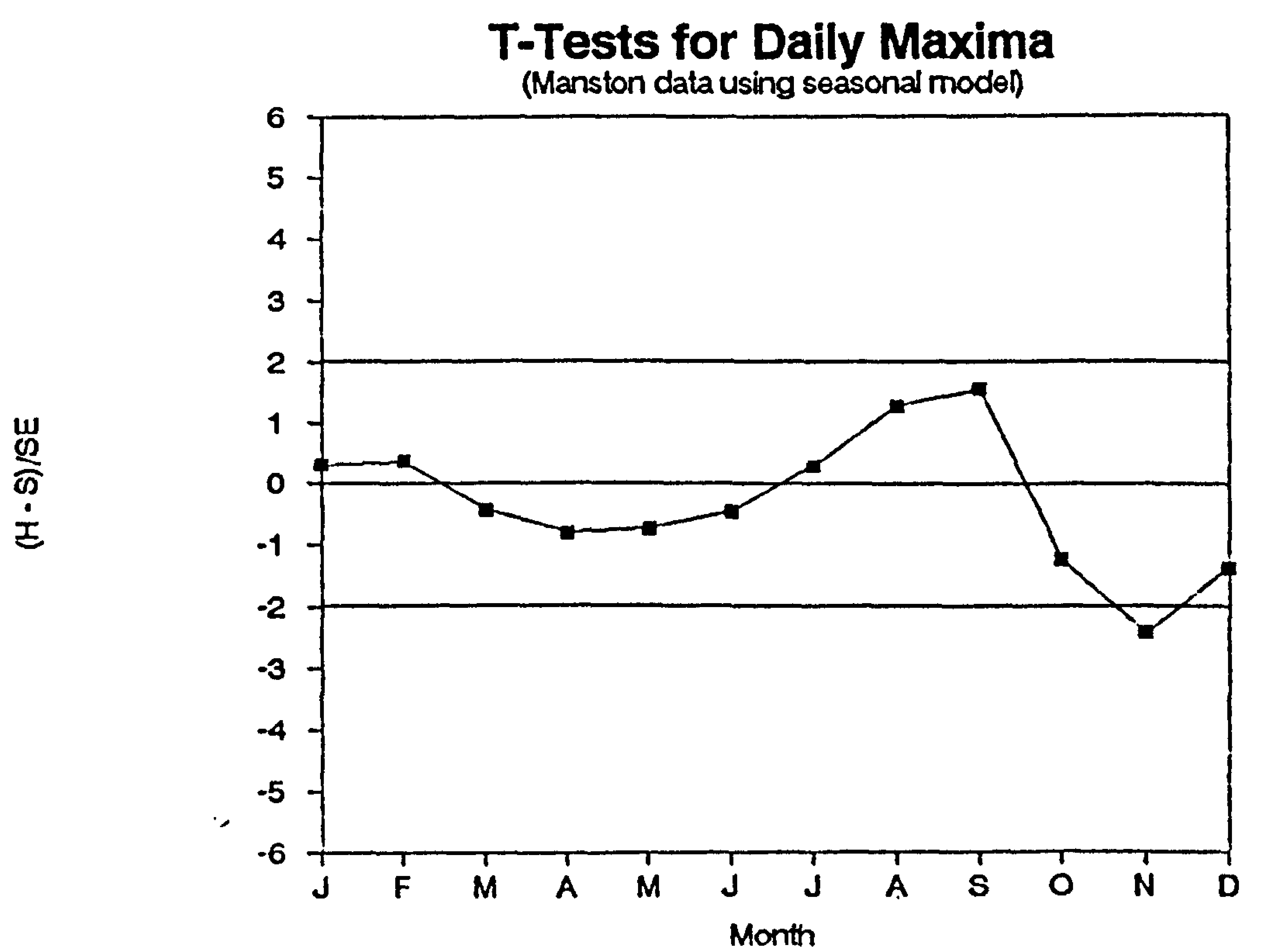


Figure D.16

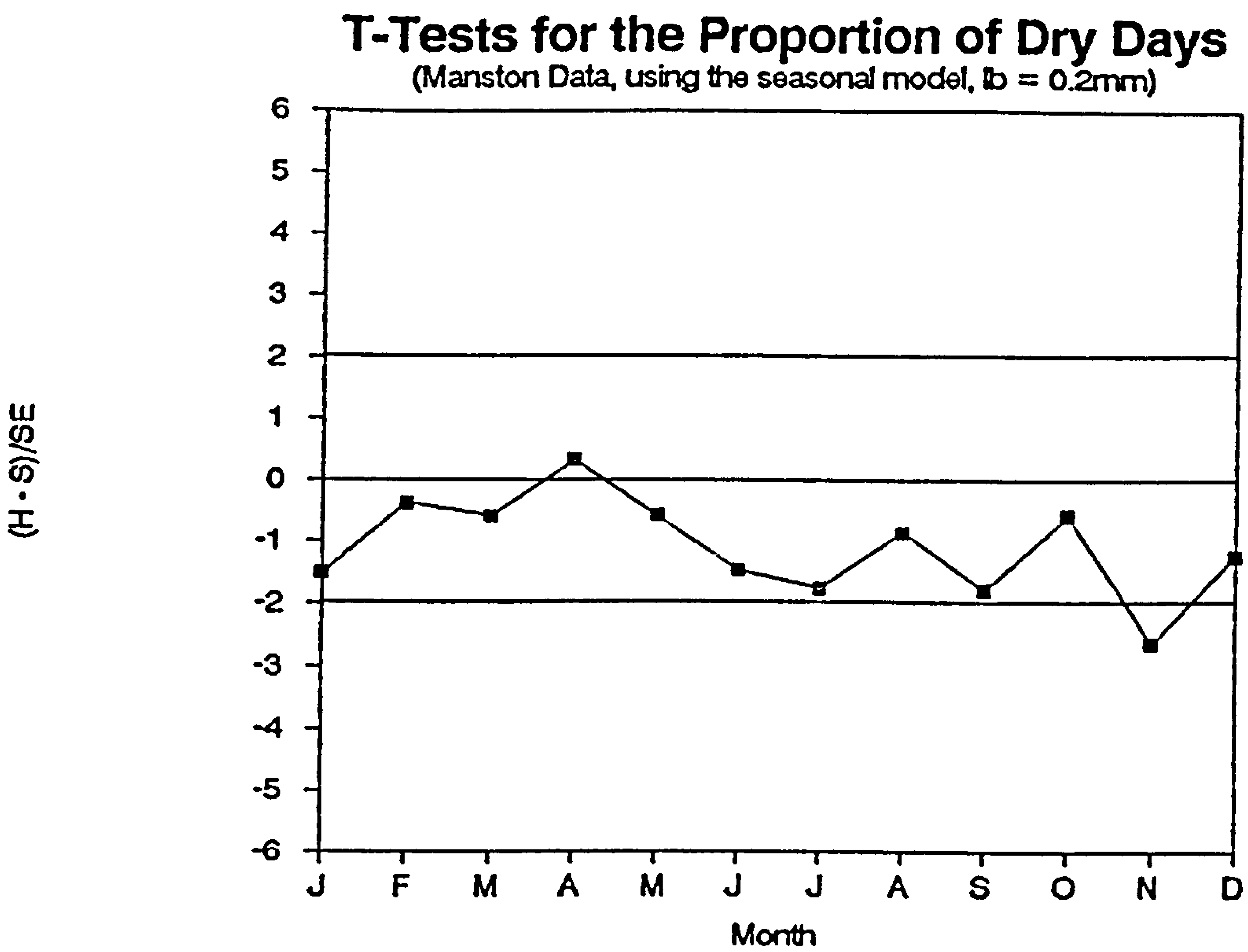


Figure D.17

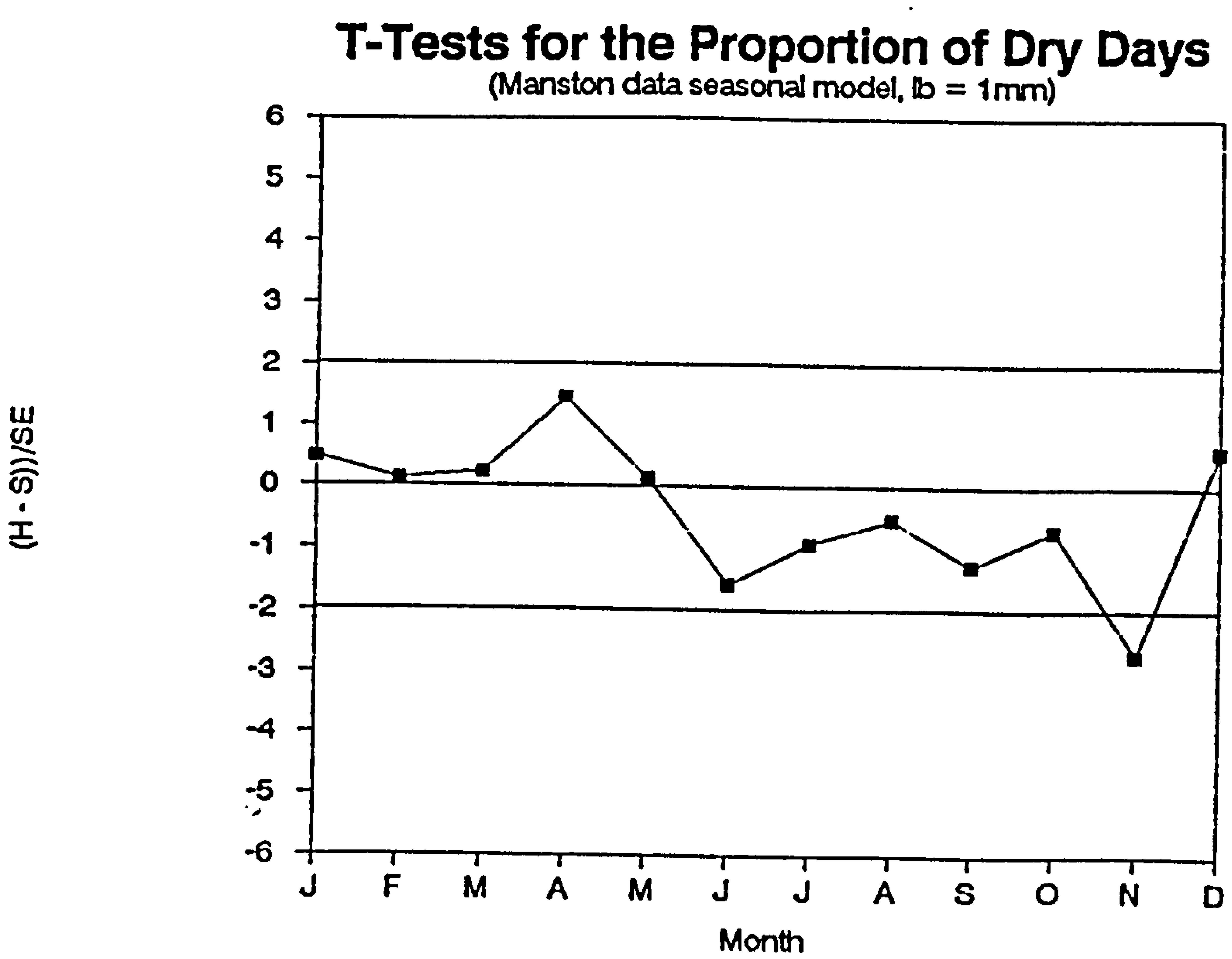


Figure D.18

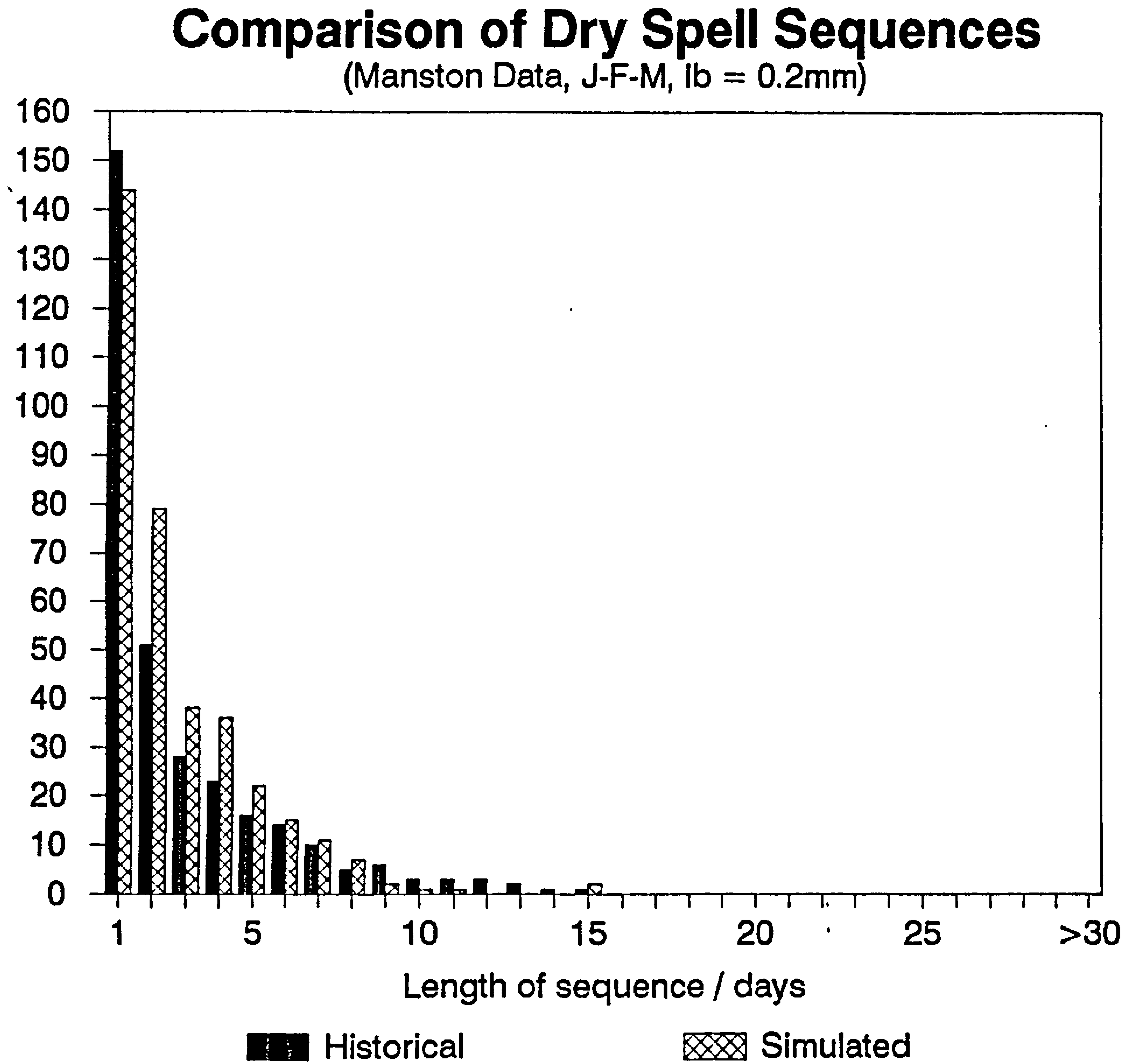

Comparison of Dry Spell Sequences
(Manston Data, J-F-M, lb = 0.2mm)
Frequency (over a 20 year period)
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of sequence / days
Historical
Simulated

Figure D.19

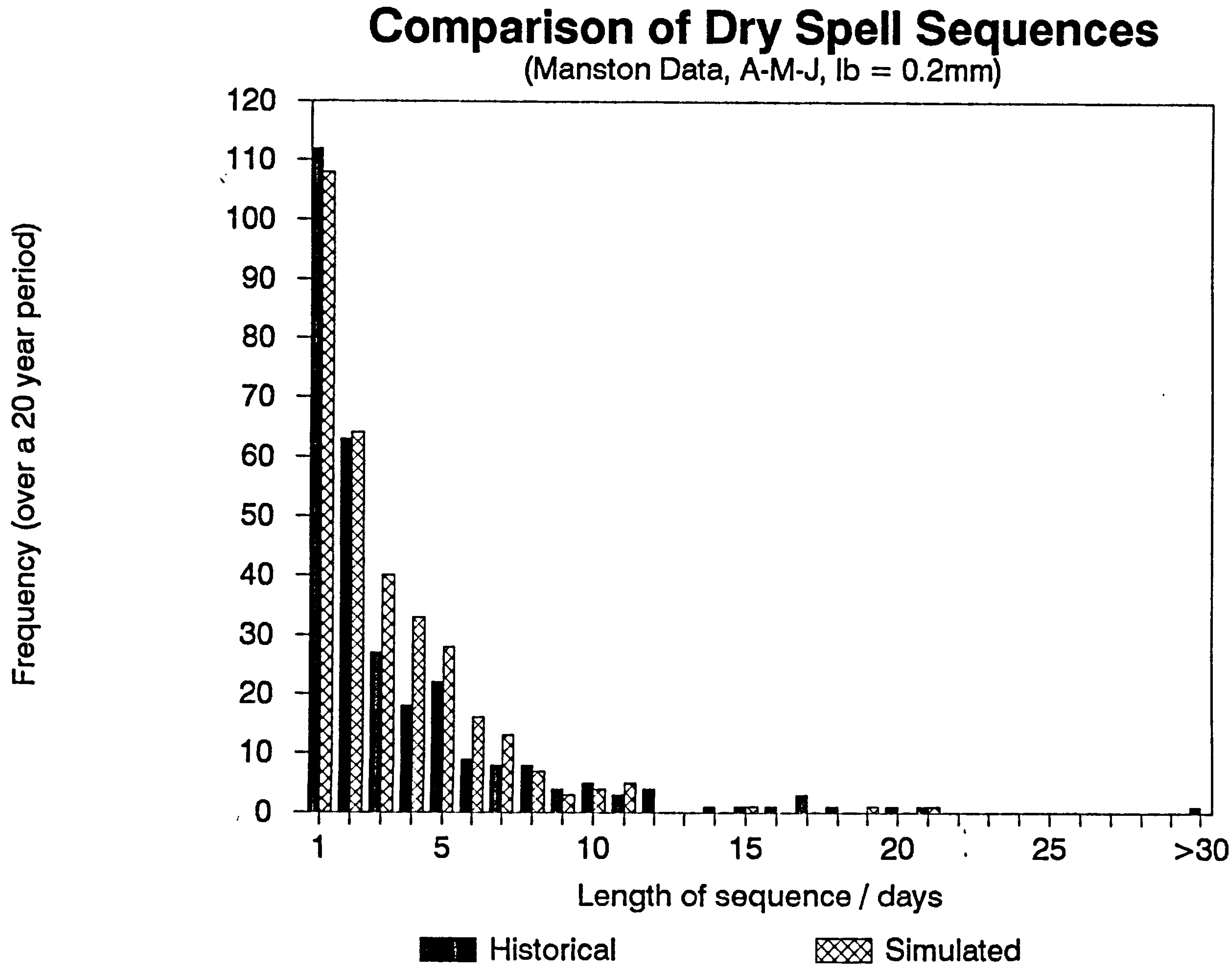

Comparison of Dry Spell Sequences
(Manston Data, A-M-J, lb = 0.2mm)
Frequency (over a 20 year period)
120
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of sequence / days
Historical
Simulated

Figure D.20

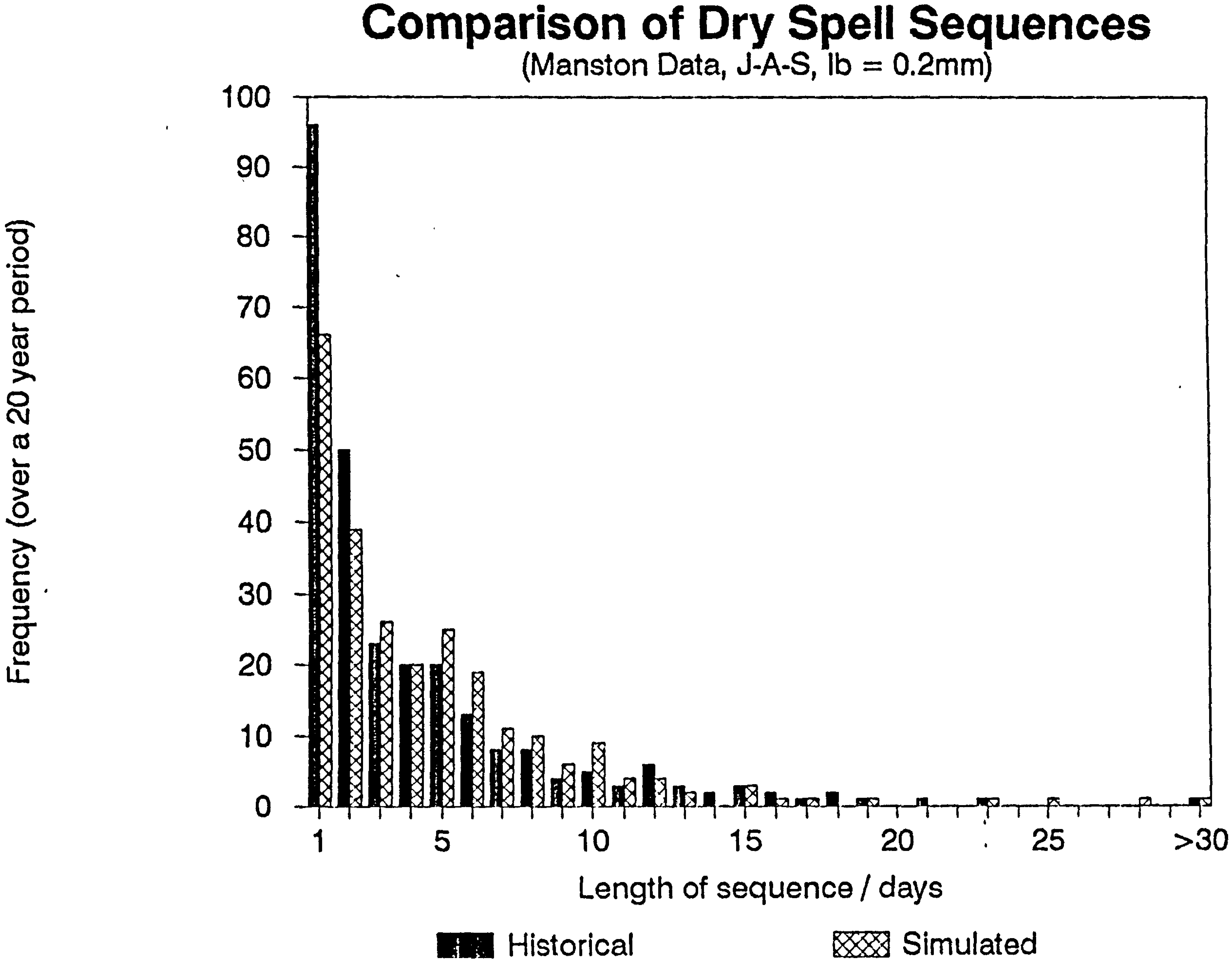

Comparison of Dry Spell Sequences
(Manston Data, J-A-S, lb = 0.2mm)
Frequency (over a 20 year period)
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of sequence / days
Historical
Simulated

Figure D.21

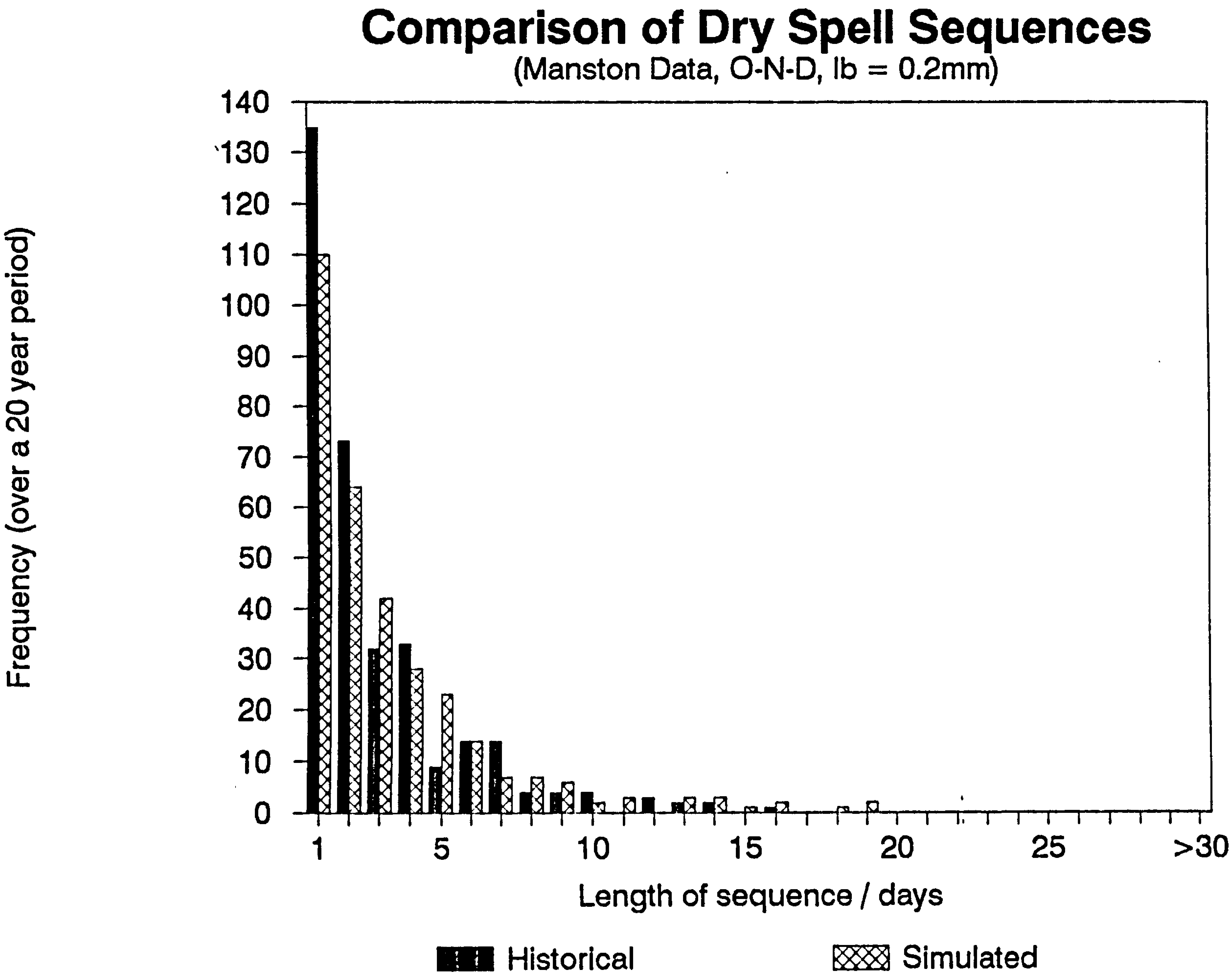

Comparison of Dry Spell Sequences
(Manston Data, O-N-D, lb = 0.2mm)
Frequency (over a 20 year period)
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of sequence / days
Historical
Simulated

Figure D.22

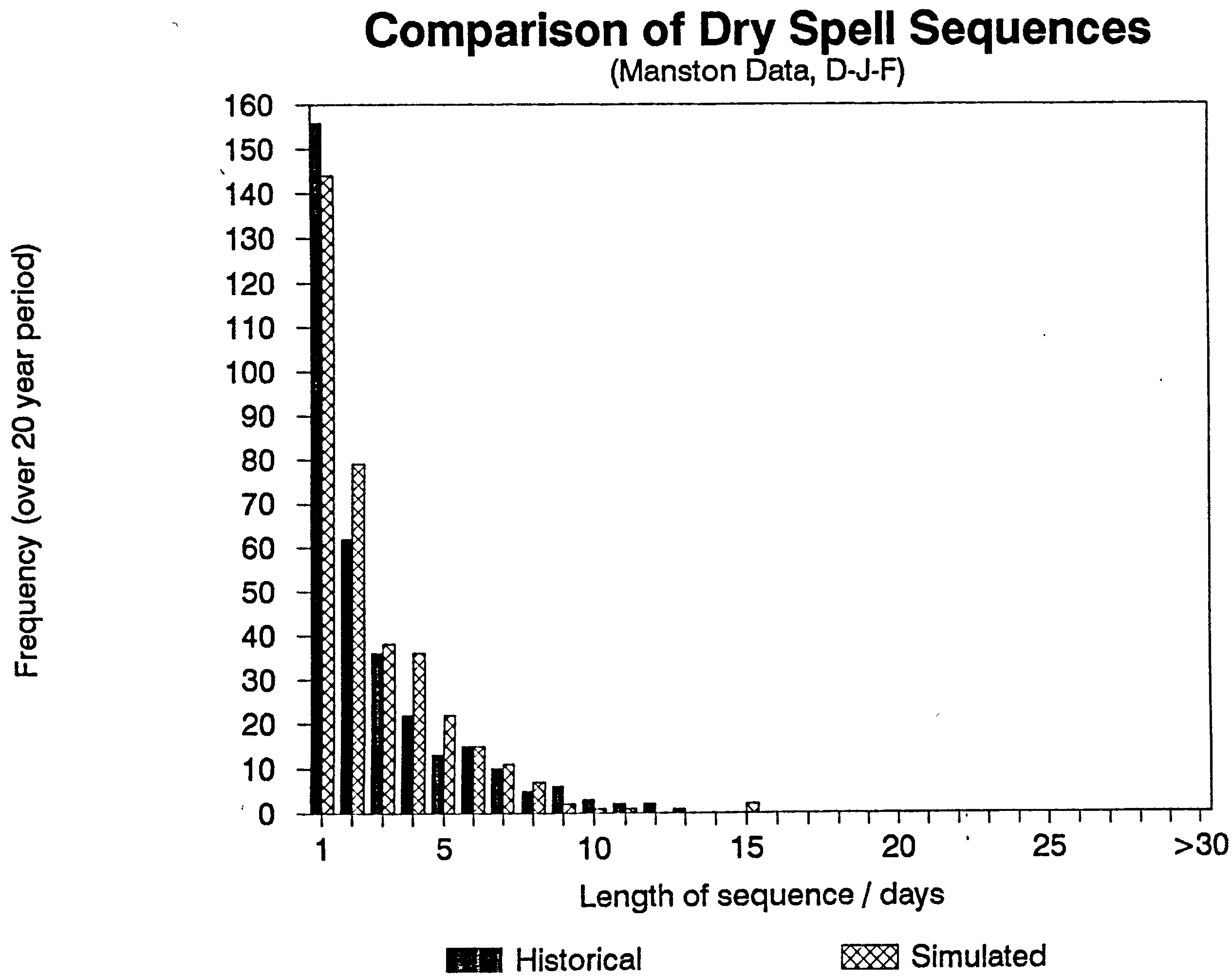

Comparison of Dry Spell Sequences
(Manston Data, D-J-F)
Frequency (over 20 year period)
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of sequence / days
Historical
Simulated

Figure D.23

Comparison of Dry Spell Sequences

(Manston Data, M-A-M)

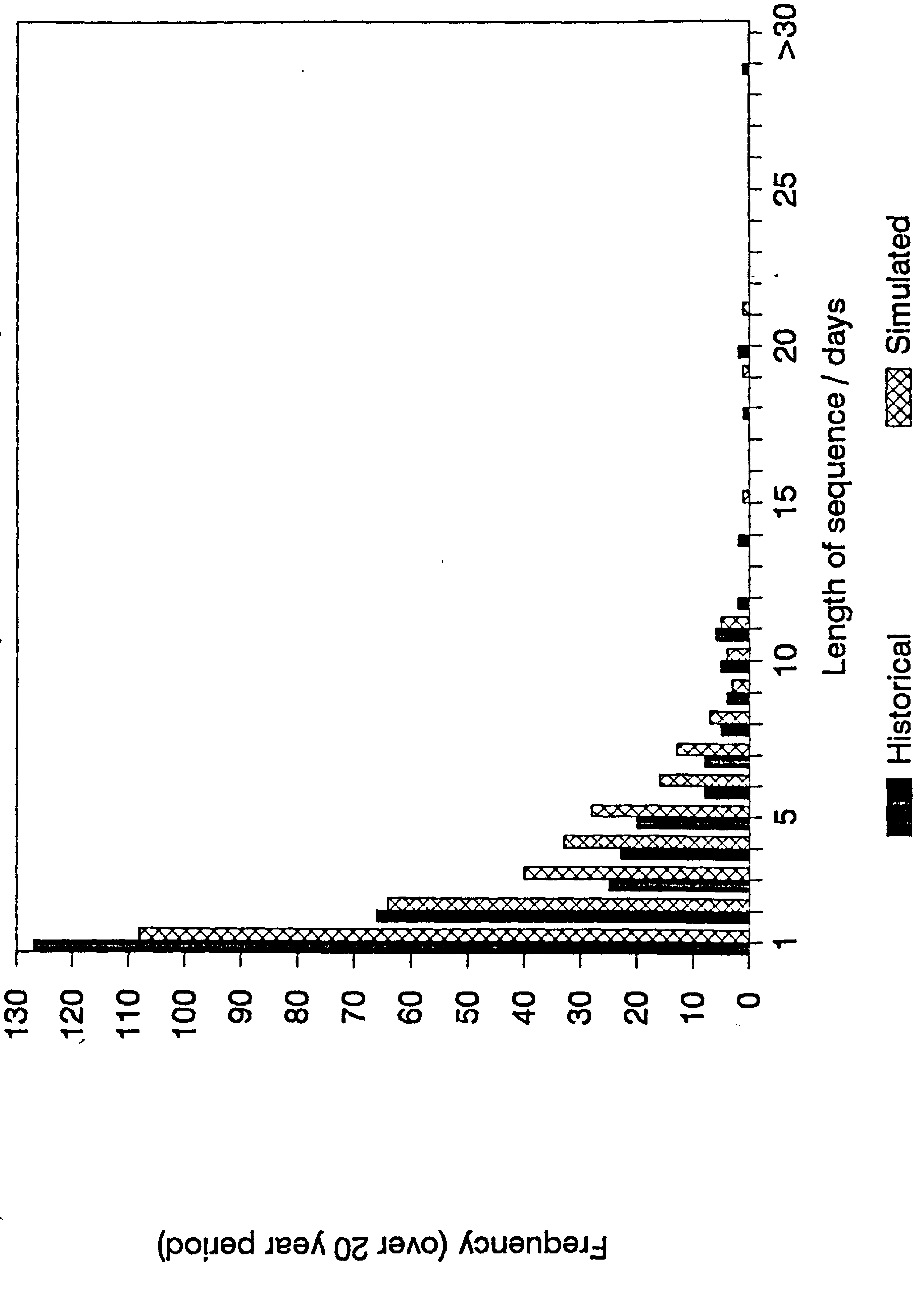


Figure D.24

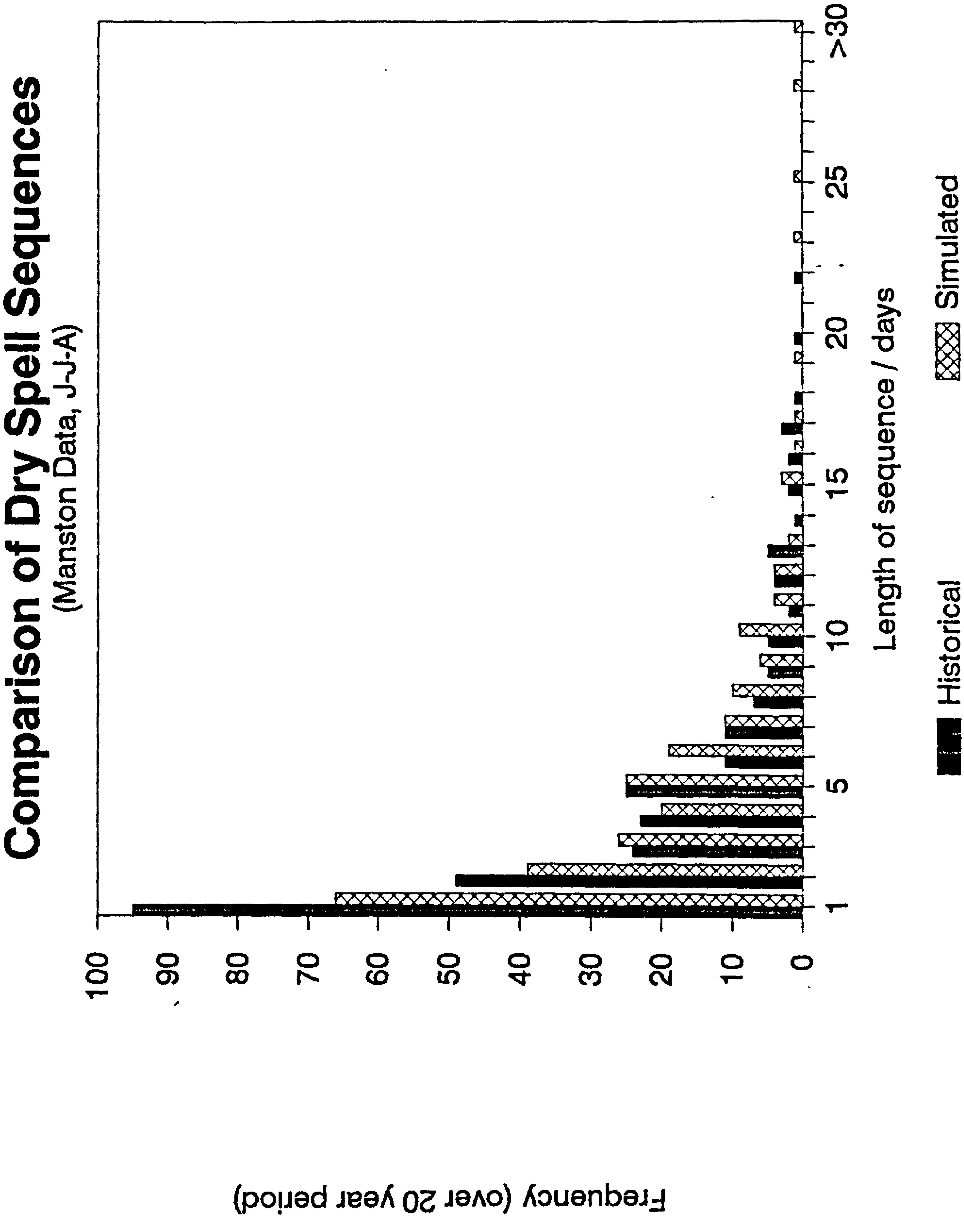
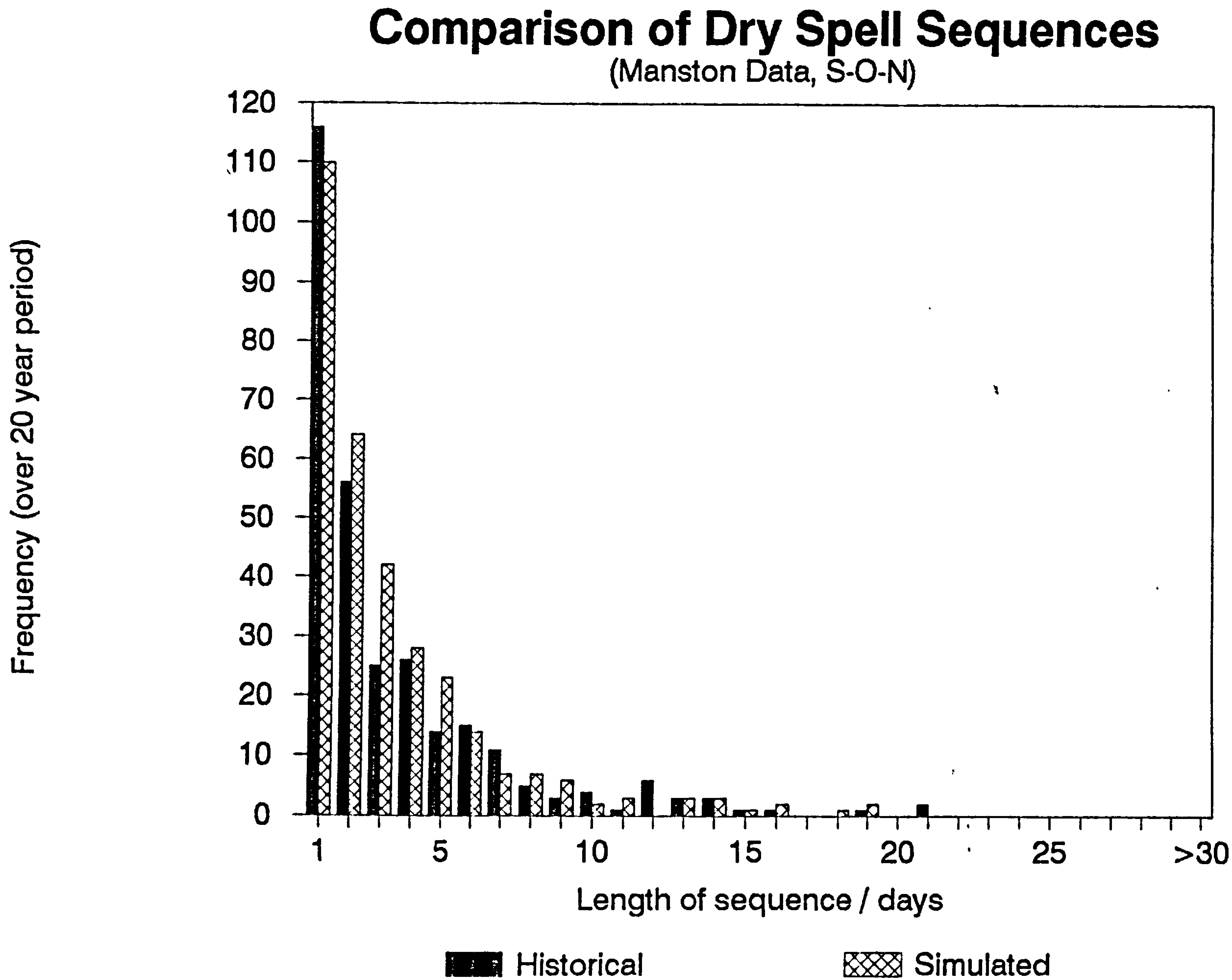

Comparison of Dry Spell Sequences
(Manston Data, J-J-A)
Frequency (over 20 year period)
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
>30
Length of sequence / days
Historical
Simulated

Figure D.25



Comparison of Dry Spell Sequences
(Manston Data, S-O-N)
Frequency (over 20 year period)
0
10
20
30
40
50
60
70
80
90
100
110
120
1
5
10
15
20
25
>30
Length of sequence / days
Historical
Simulated

Figure D.26

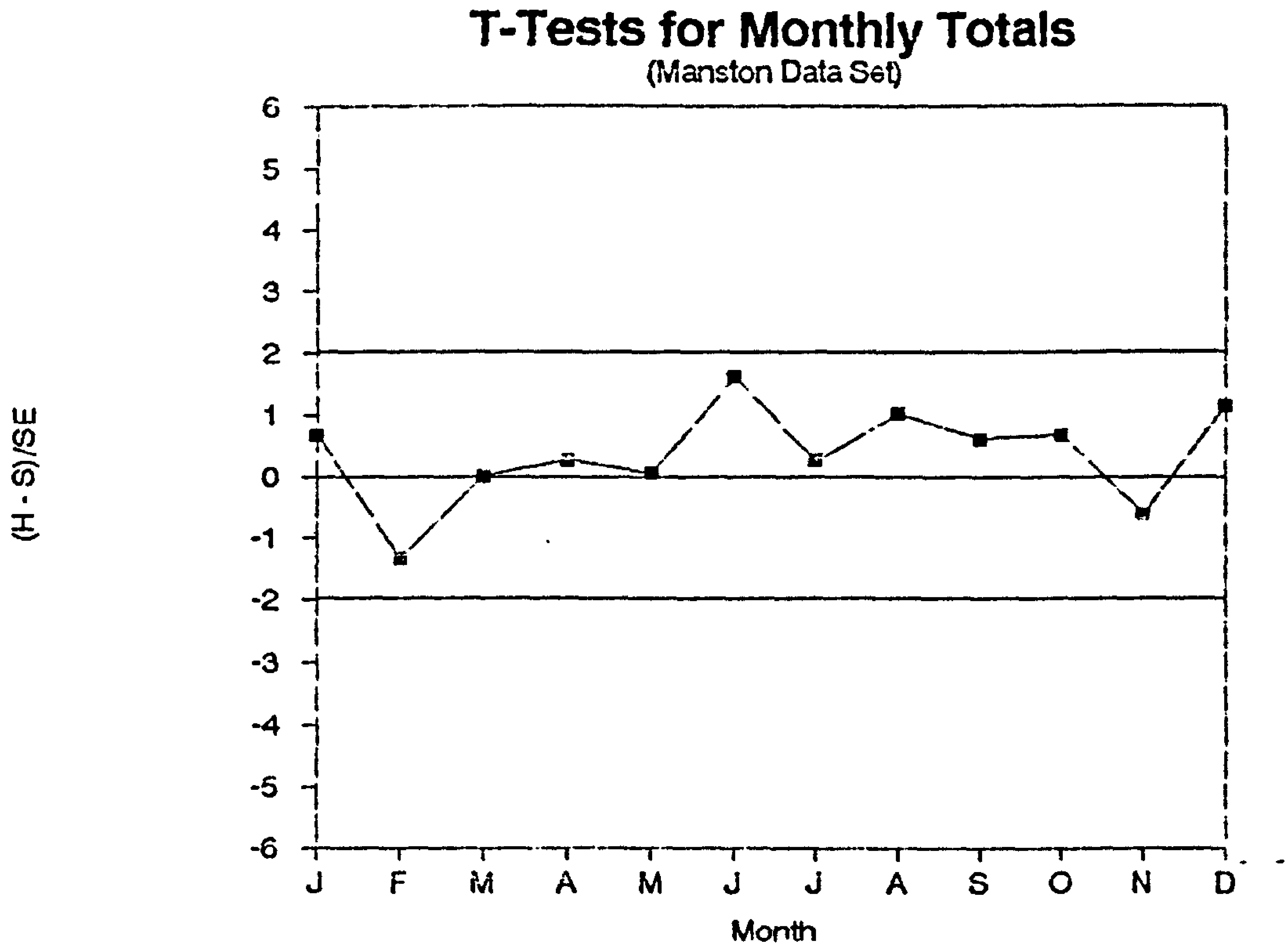


Figure E.1

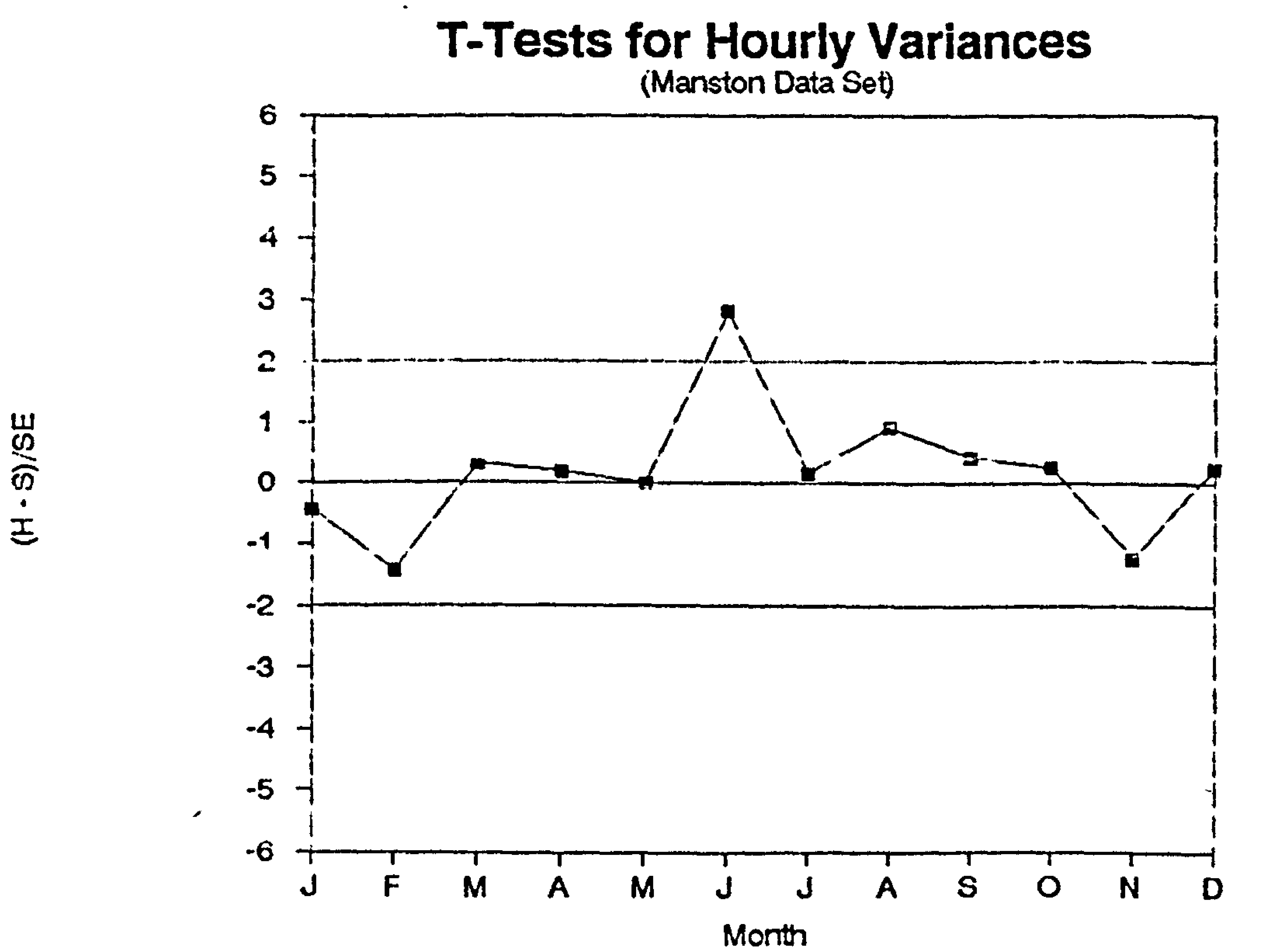


Figure E.2

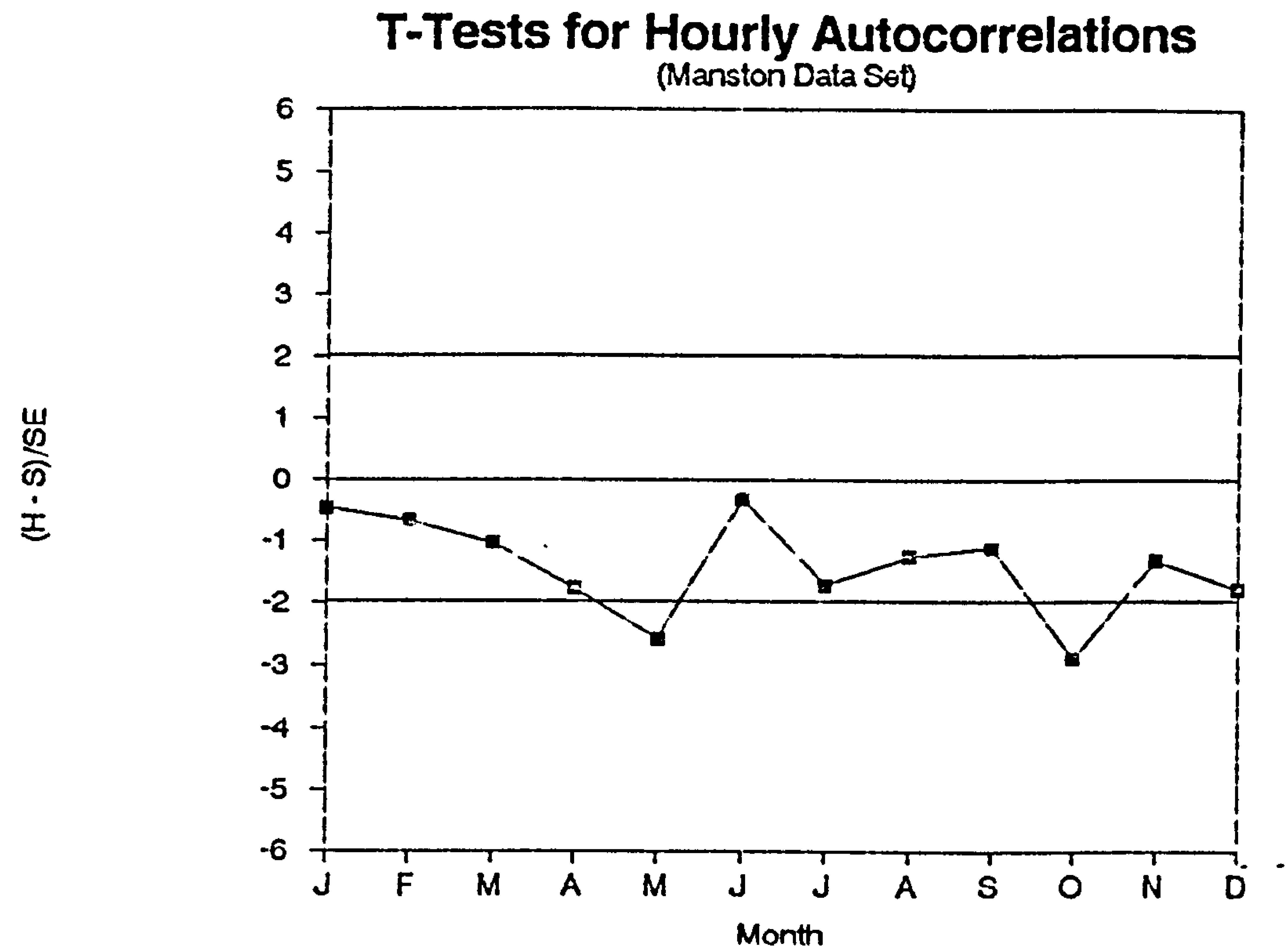


Figure E.3

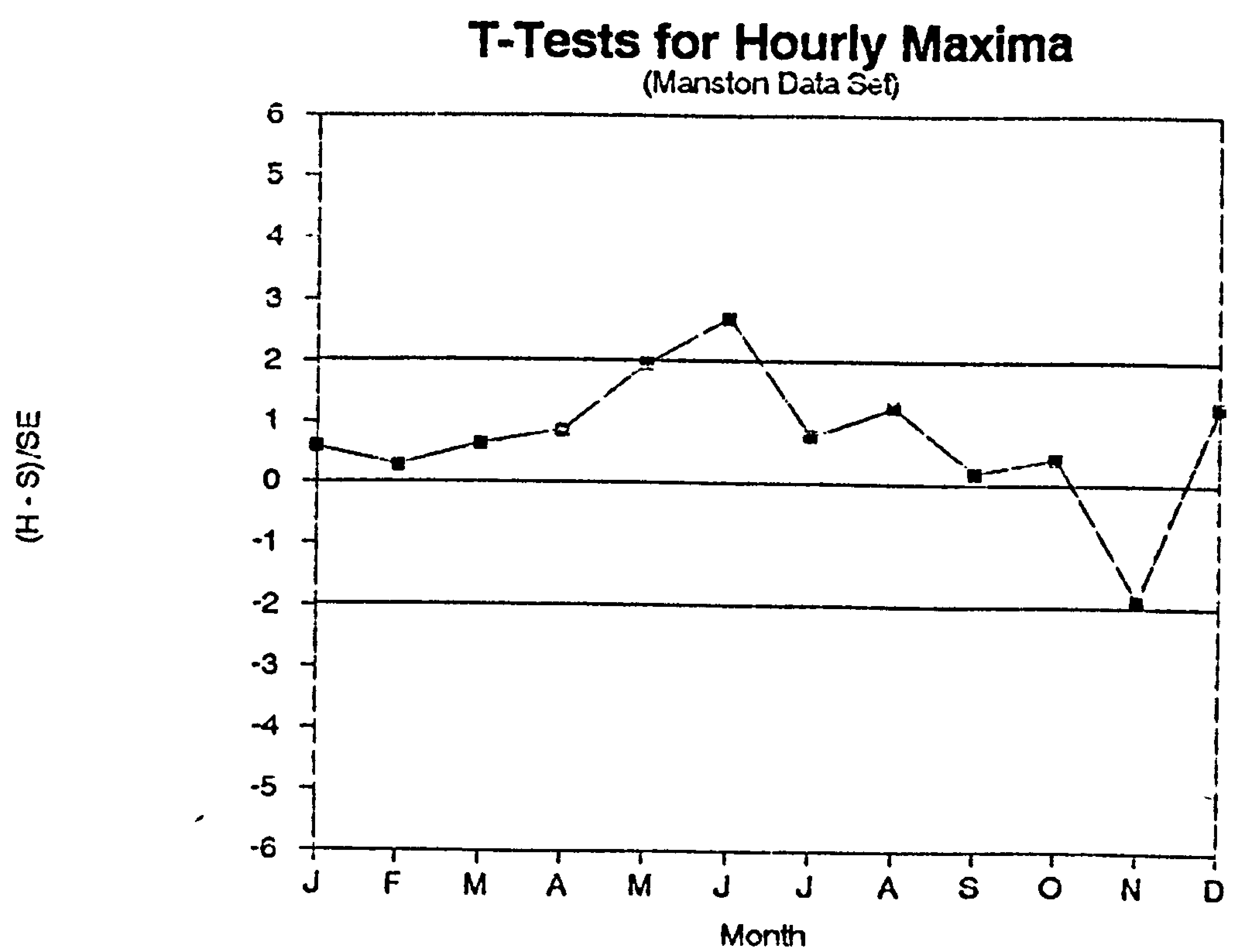


Figure E.4

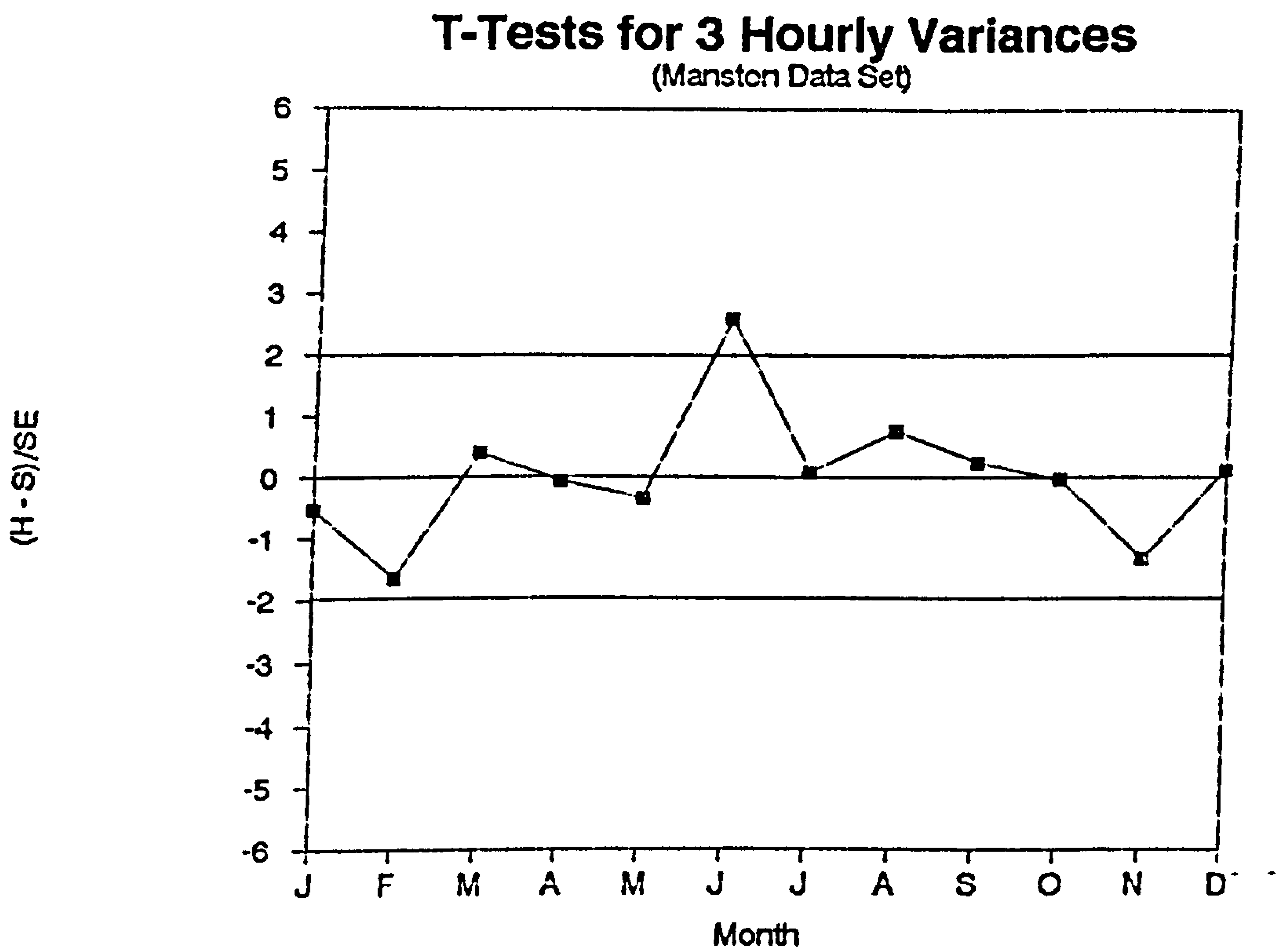


Figure E.5

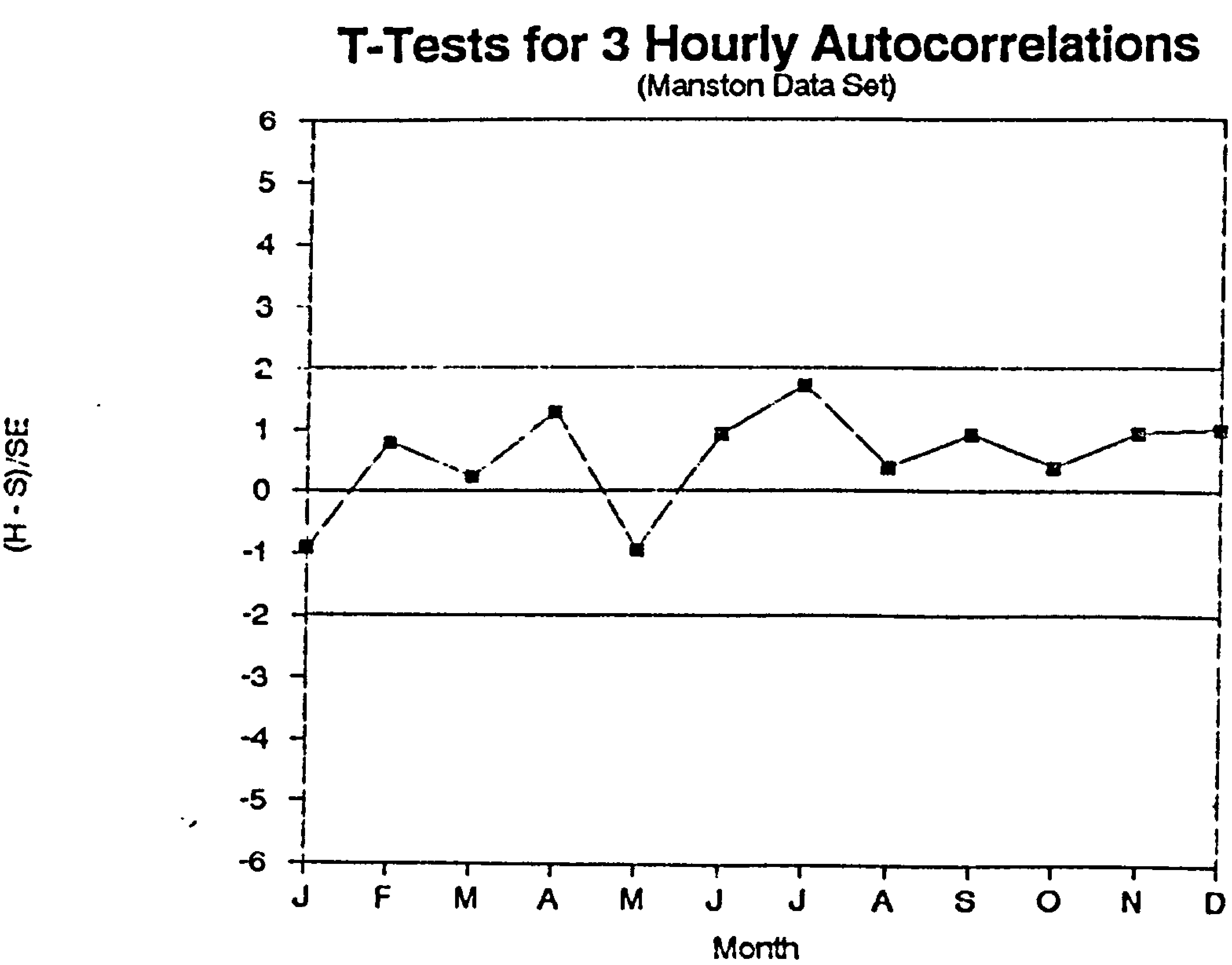


Figure E.6

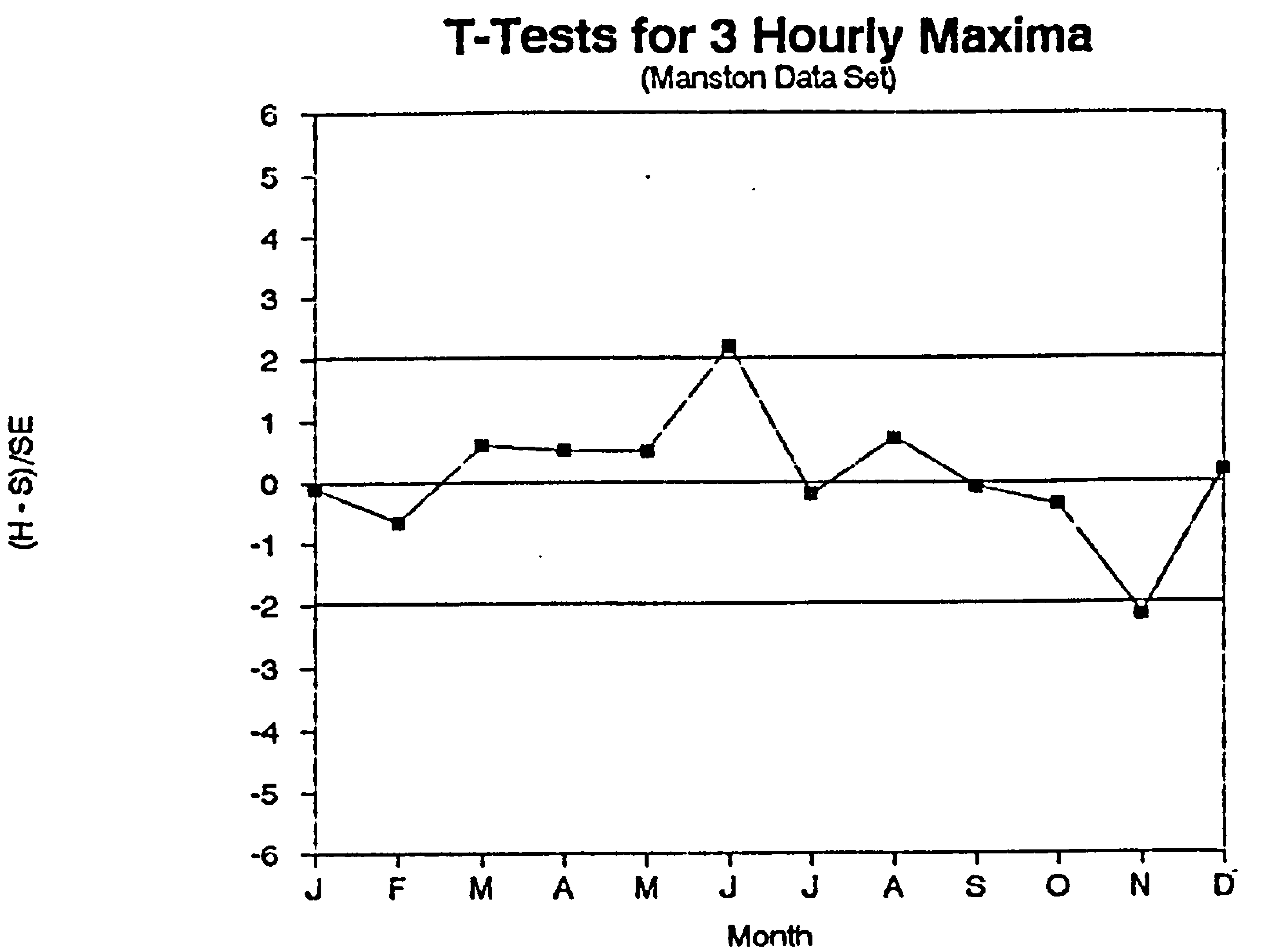


Figure E.7

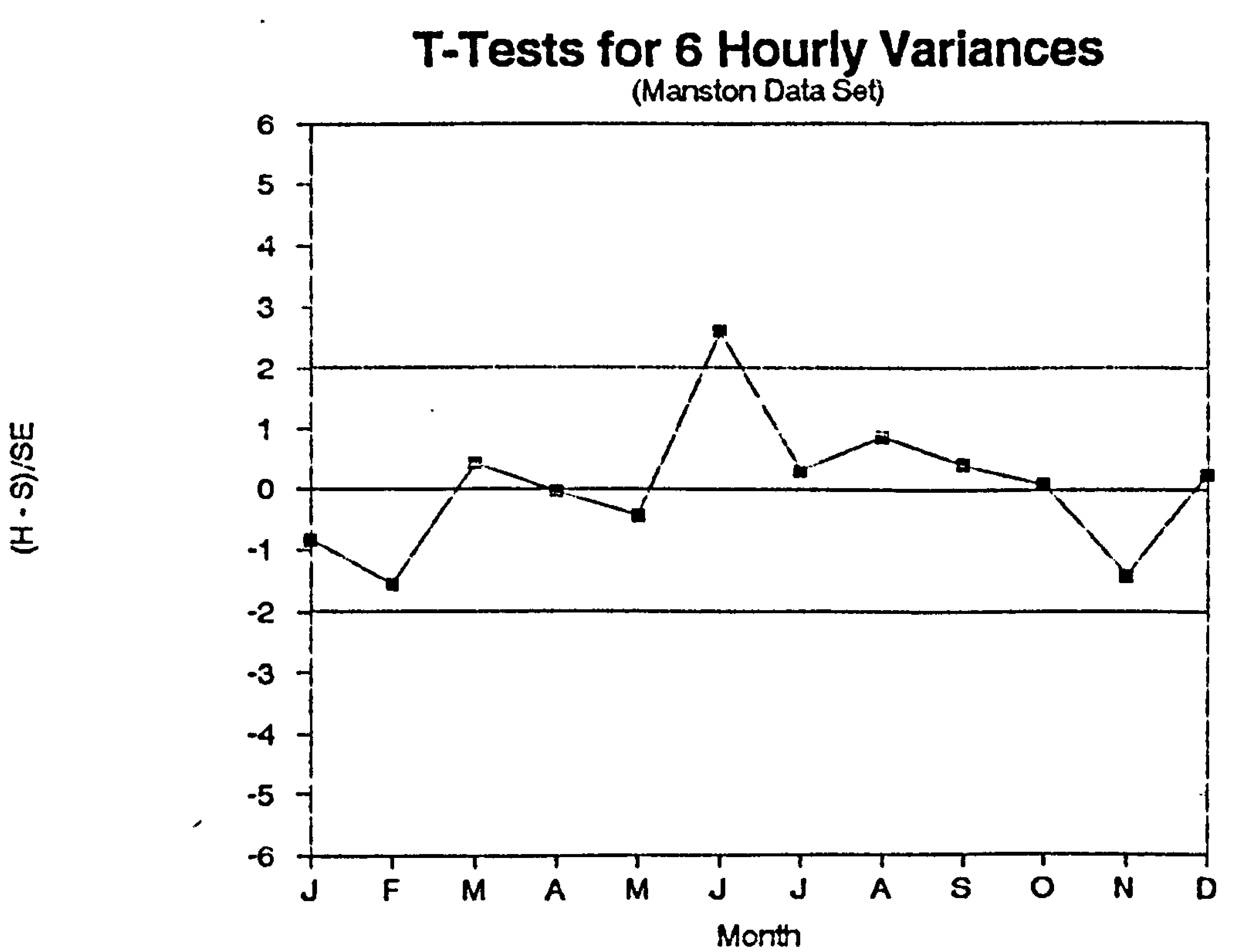


Figure E.8

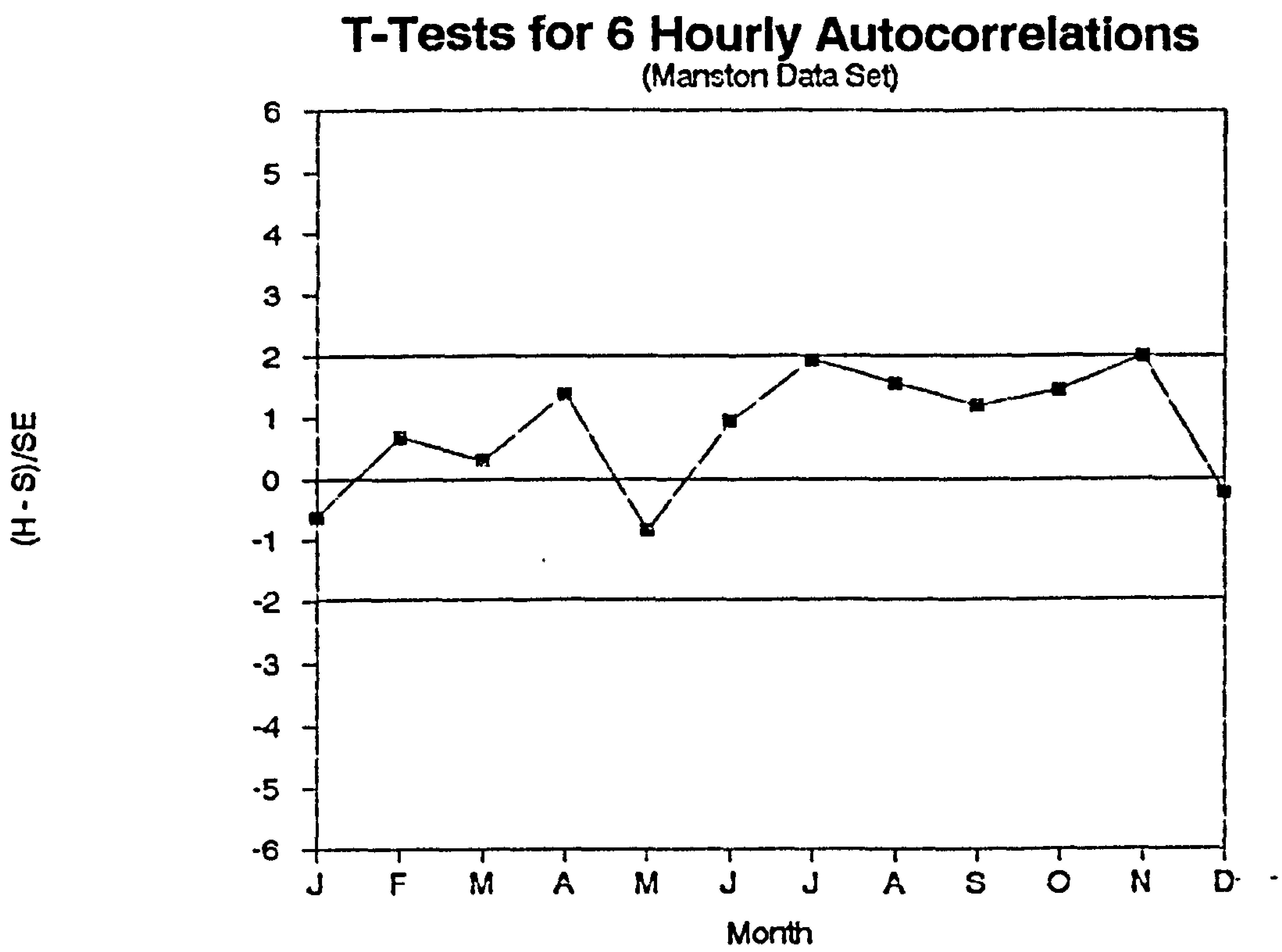


Figure E.9

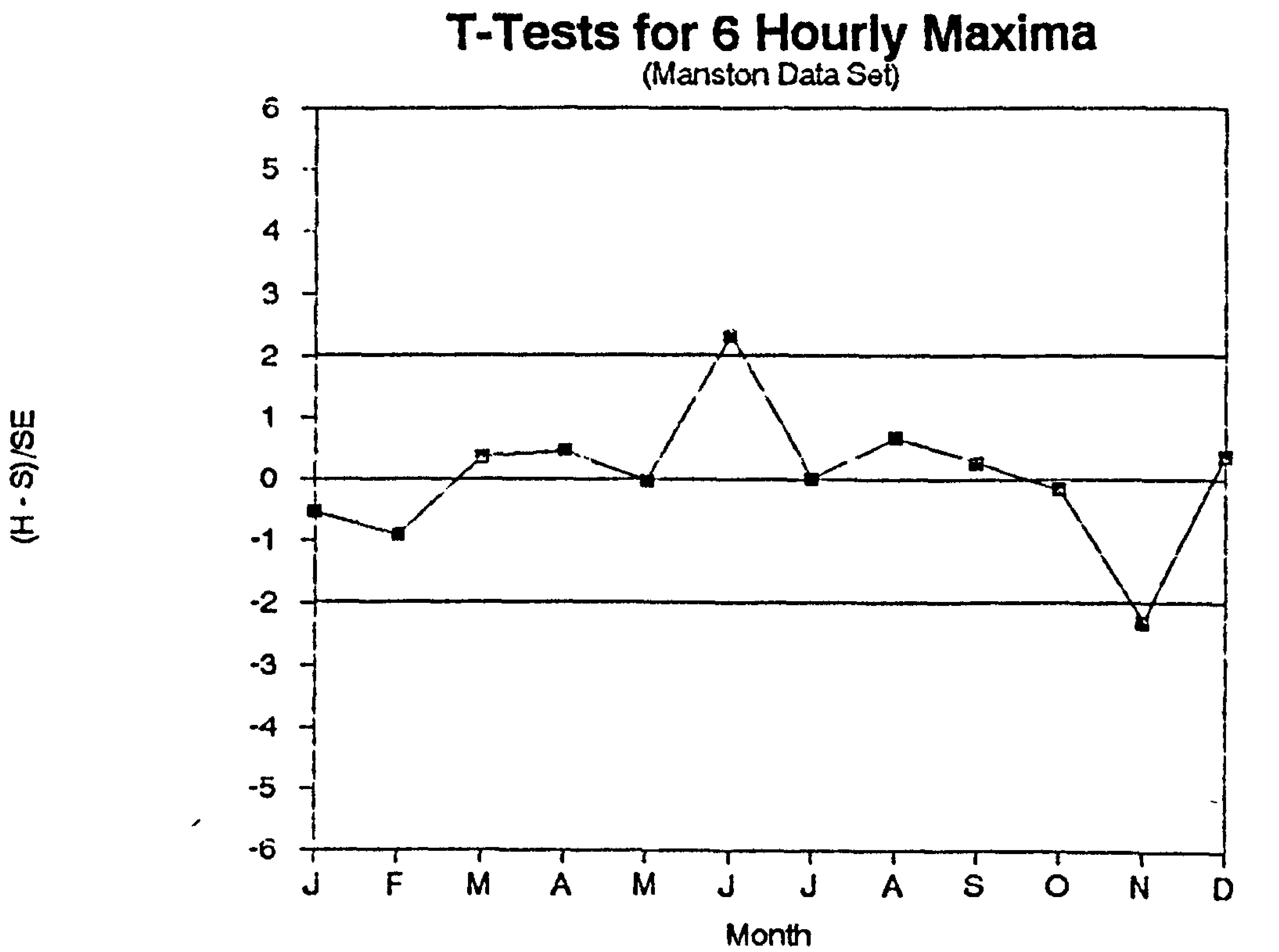


Figure E.10

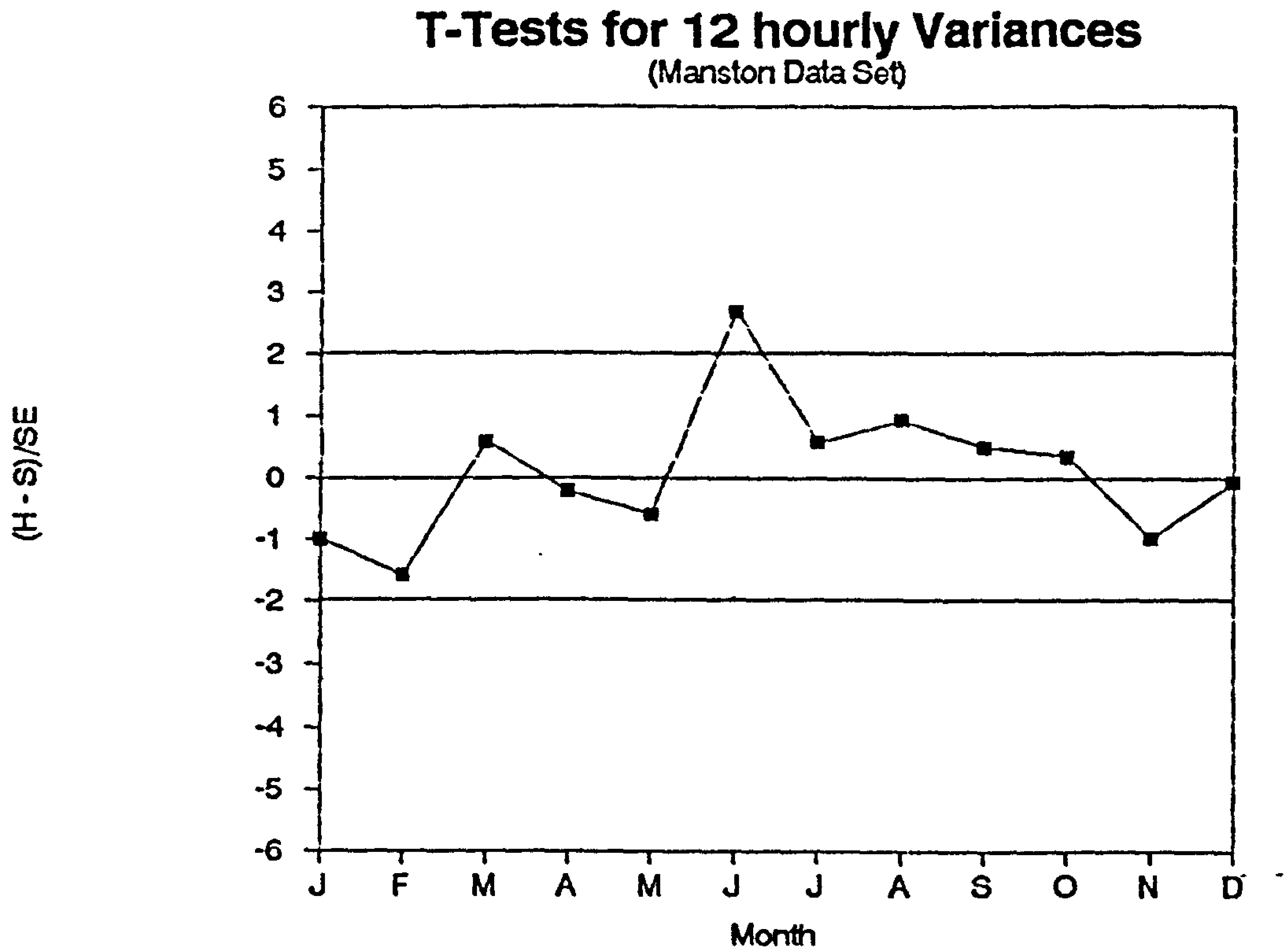


Figure E.11

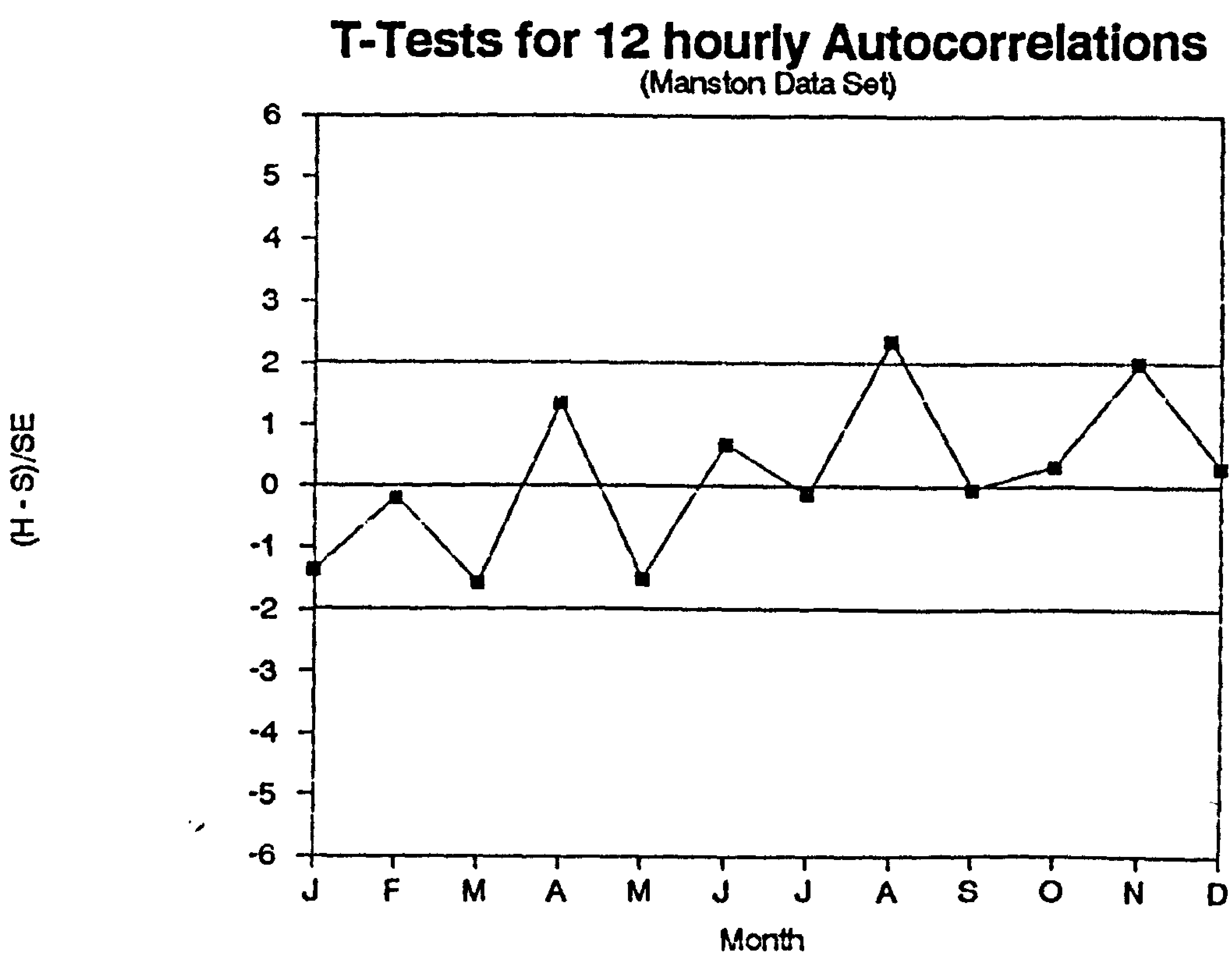


Figure E.12

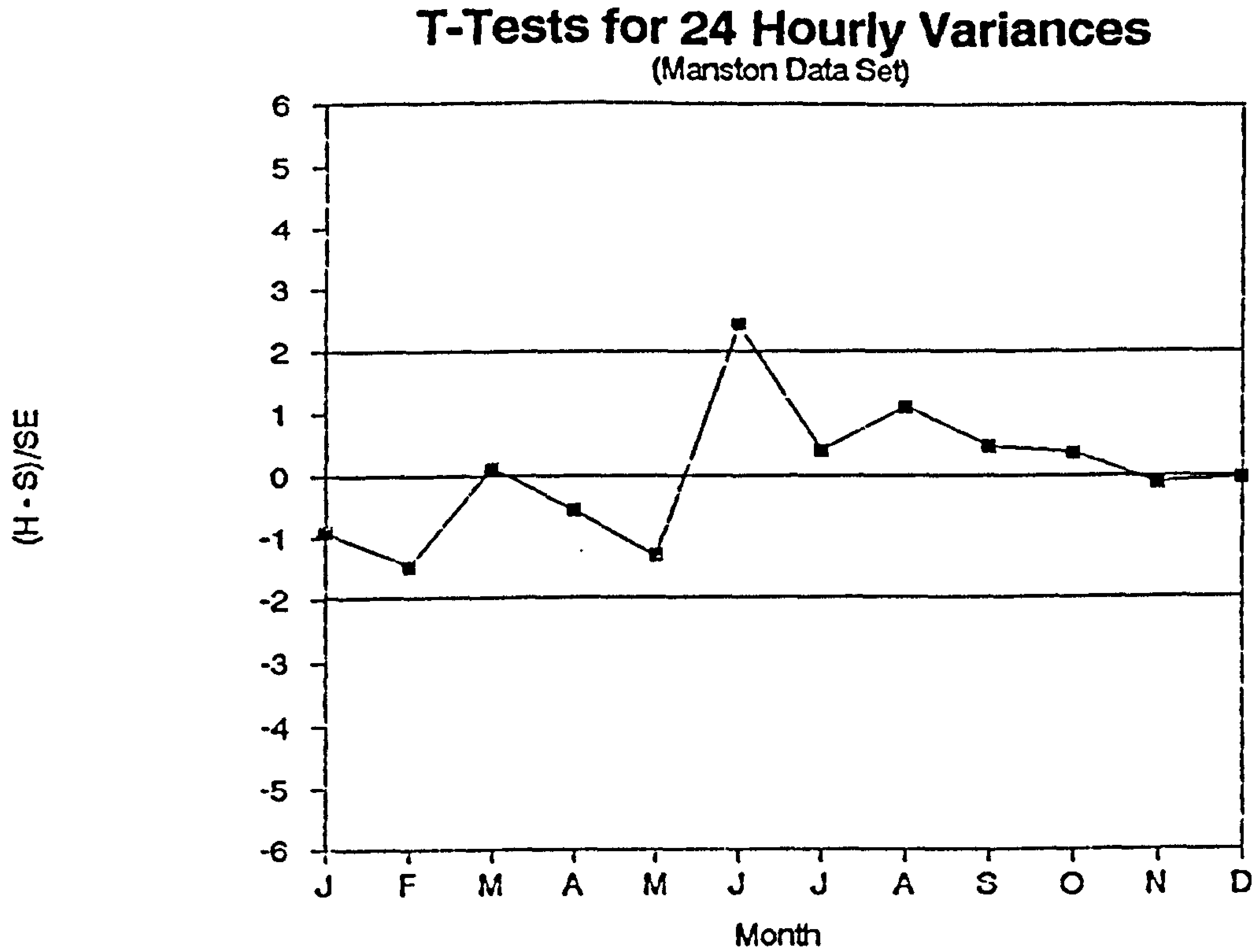


Figure E.13

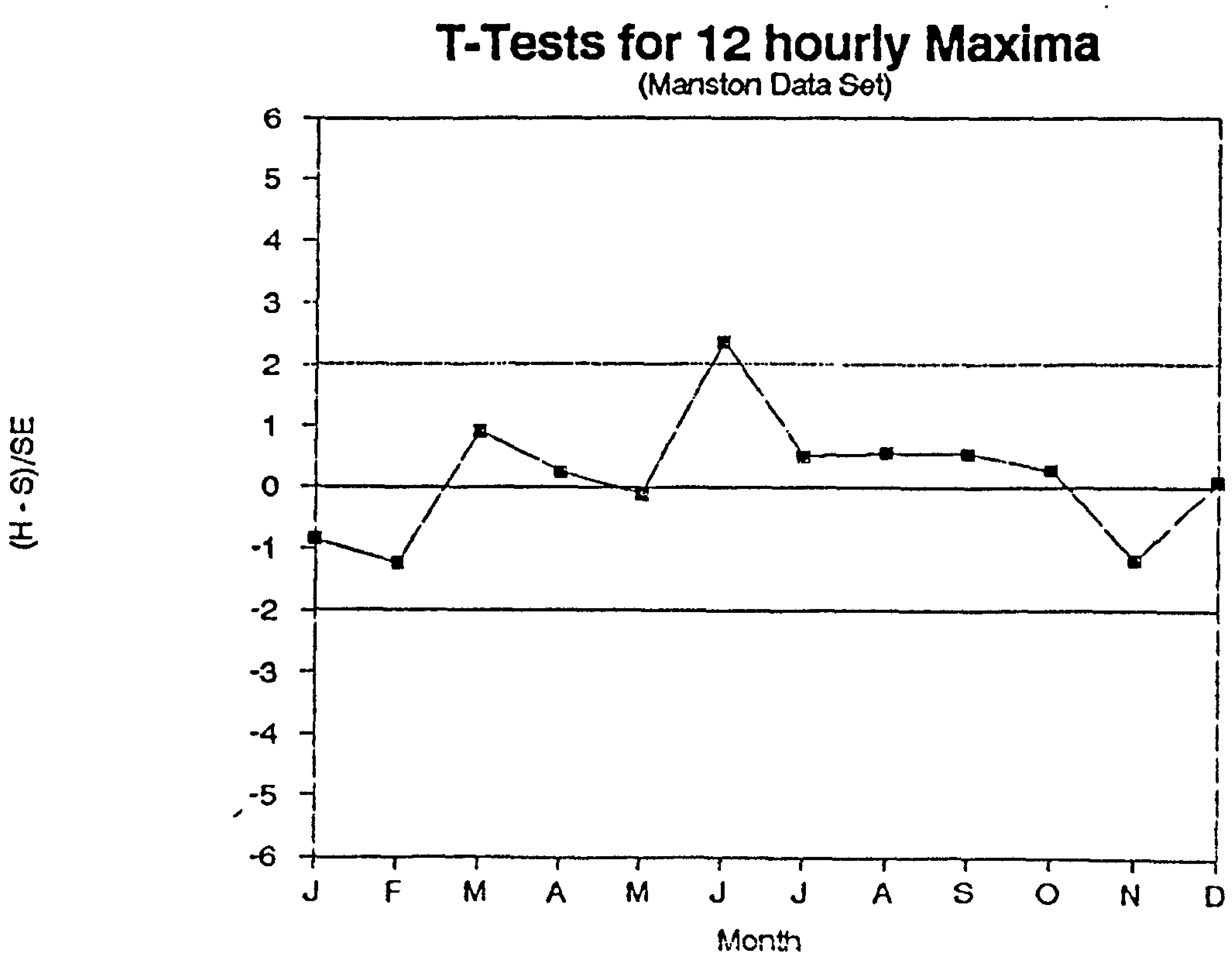


Figure E.14

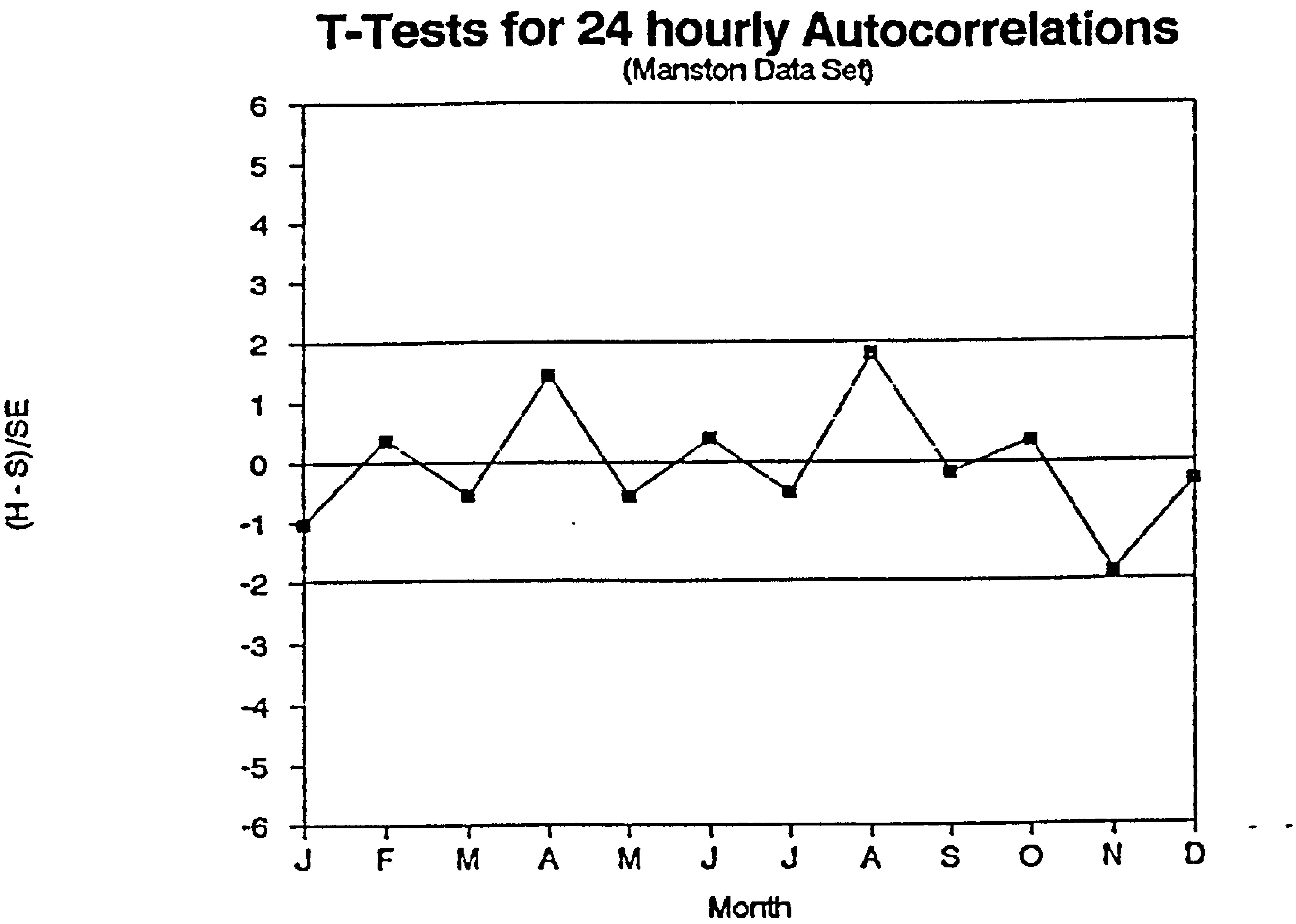


Figure E.15

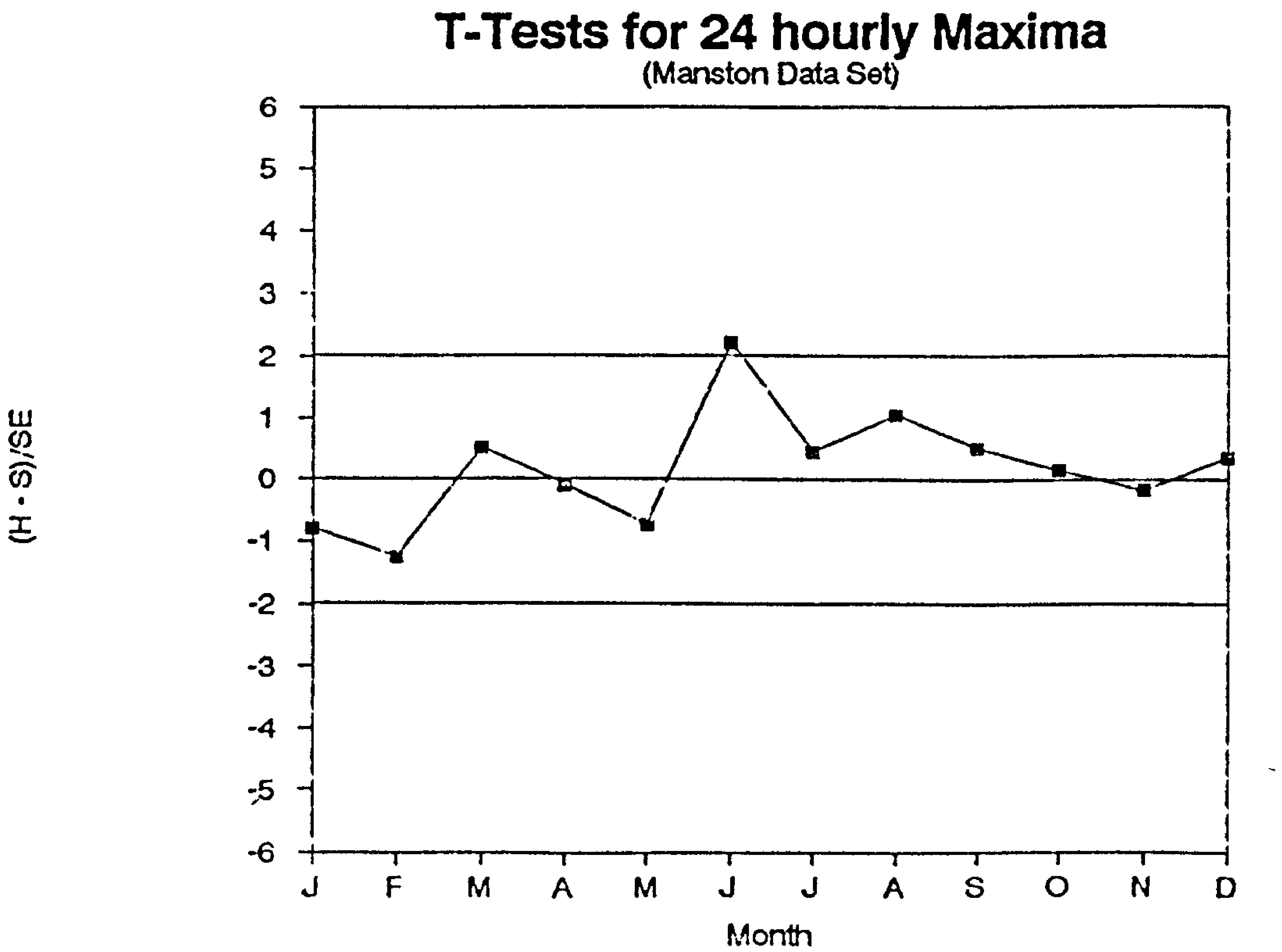


Figure E.16

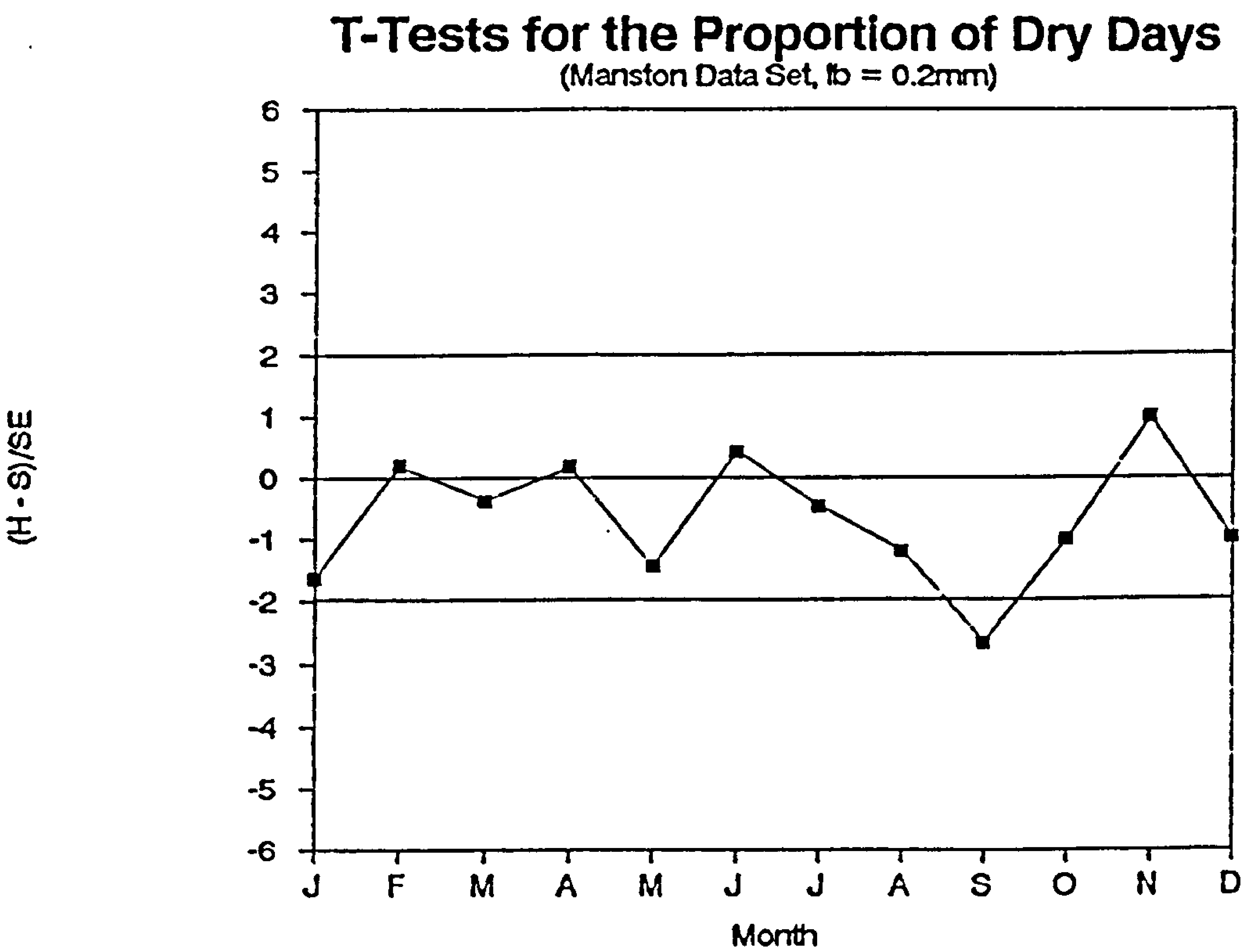


Figure E.17

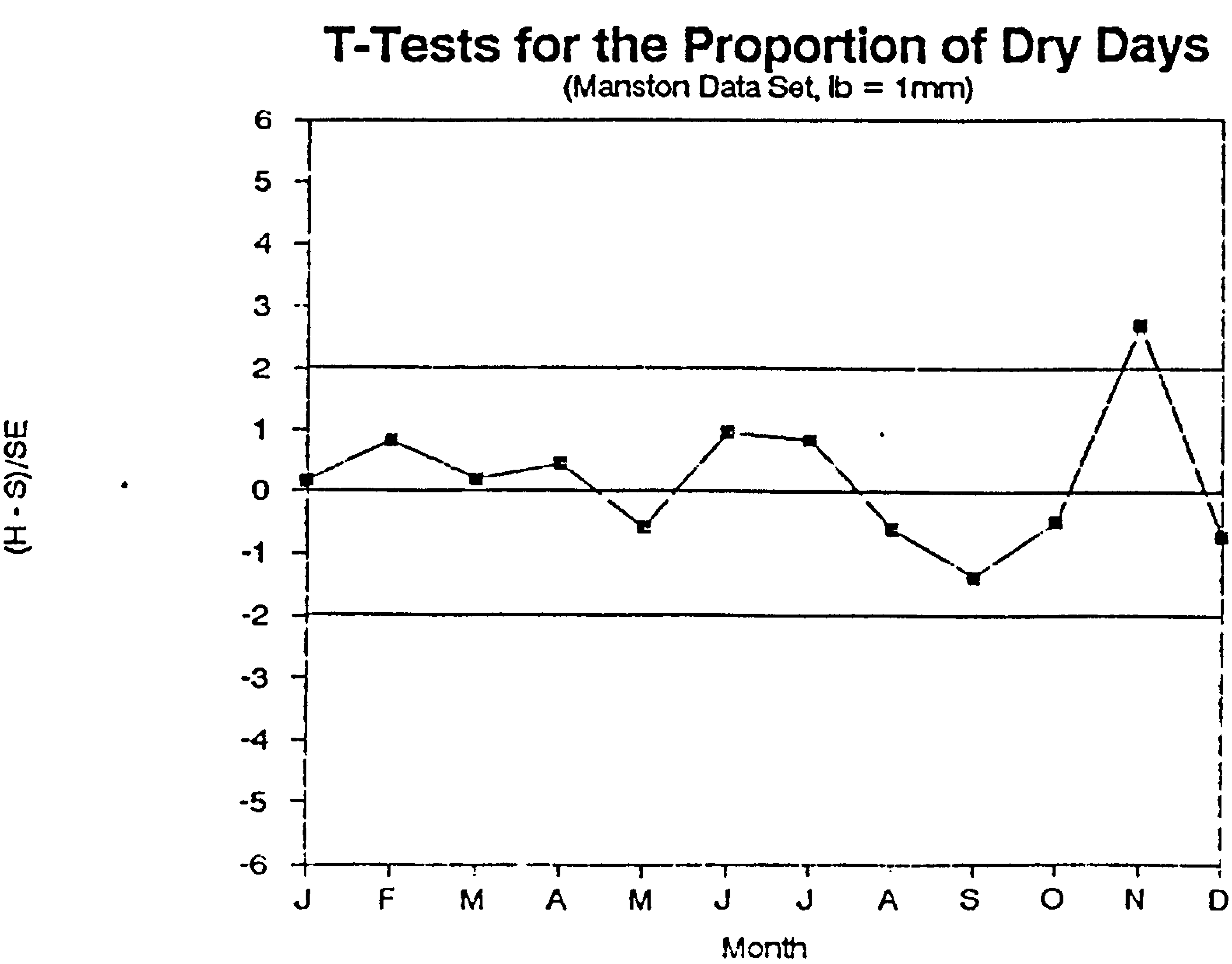


Figure E.18

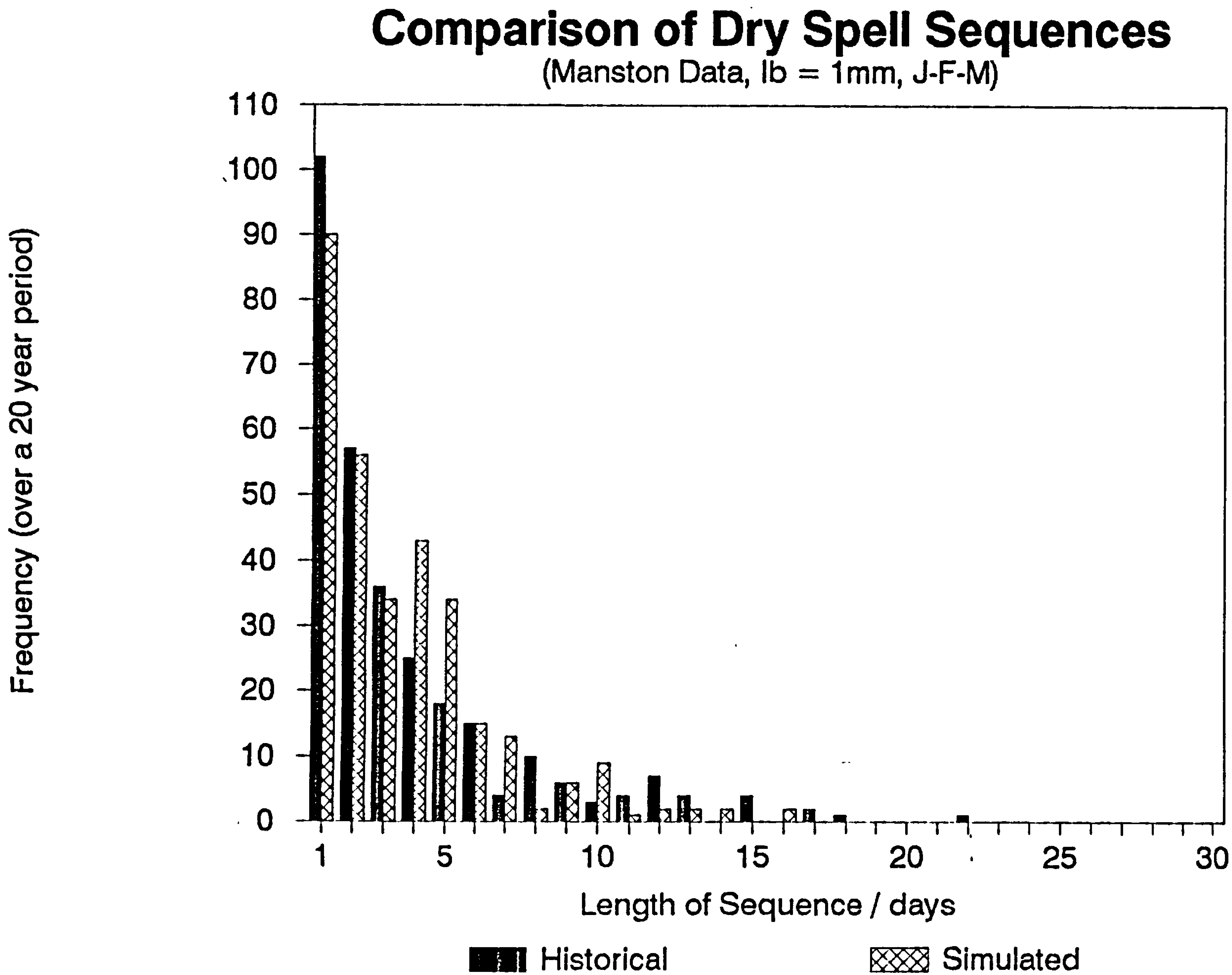

Comparison of Dry Spell Sequences
(Manston Data, lb = 1mm, J-F-M)
Frequency (over a 20 year period)
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of Sequence / days
Historical
Simulated

Figure E.19

Comparison of Dry Spell Sequences

(Manston Data, lb = 1mm, A-M-J)

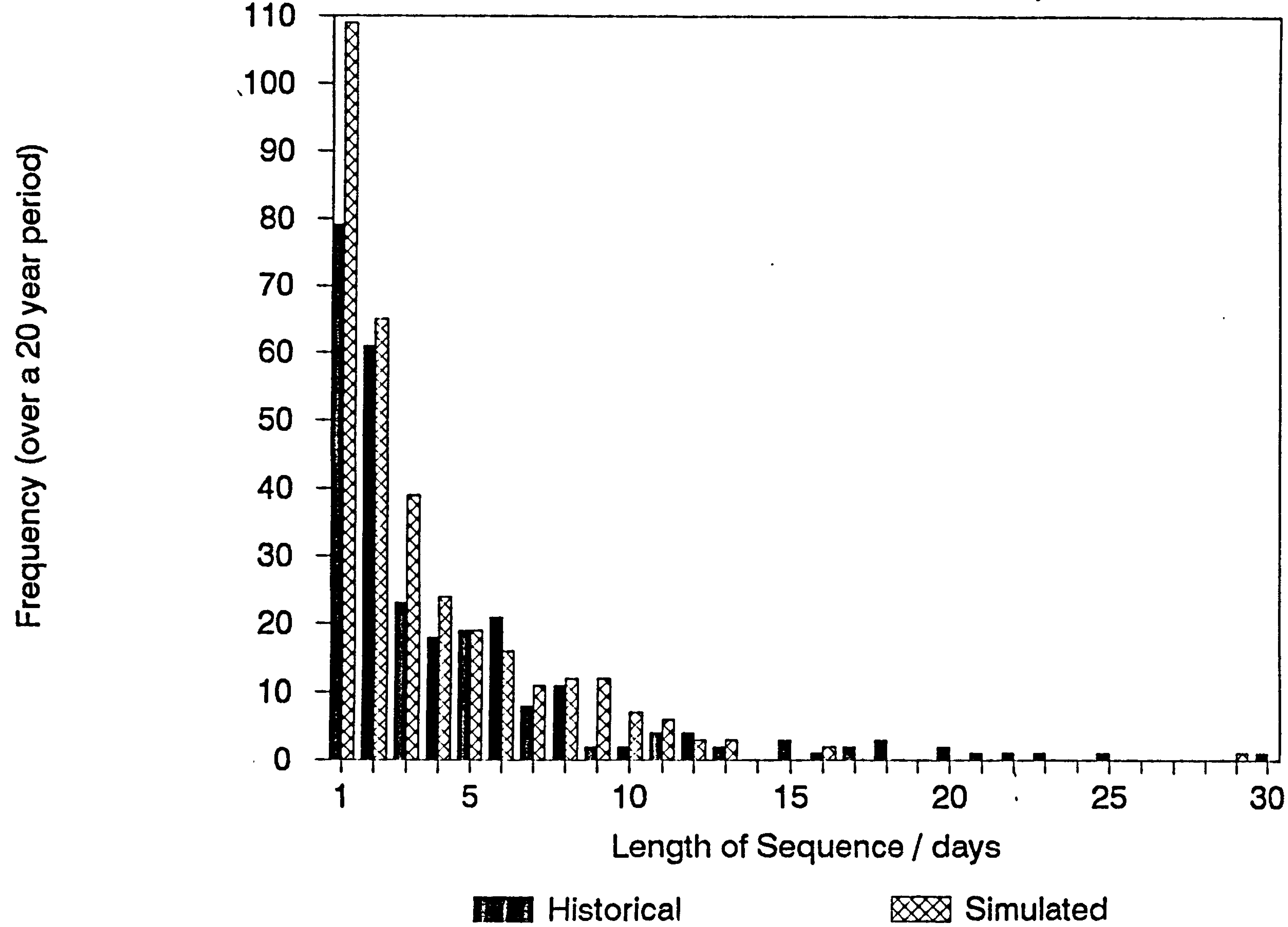
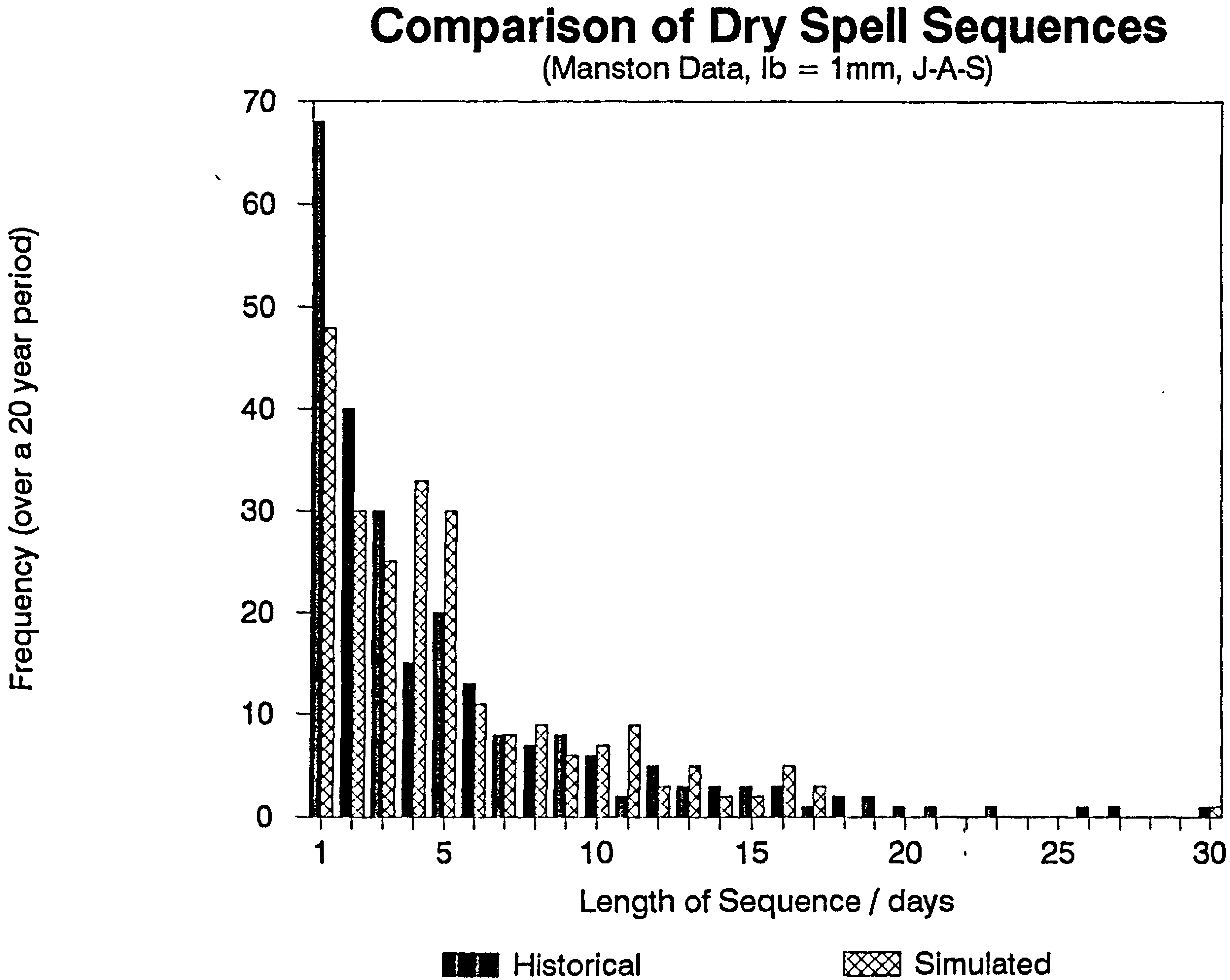


Figure E.20



Comparison of Dry Spell Sequences
(Manston Data, lb = 1mm, J-A-S)
Frequency (over a 20 year period)
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of Sequence / days
Historical
Simulated

Figure E.21

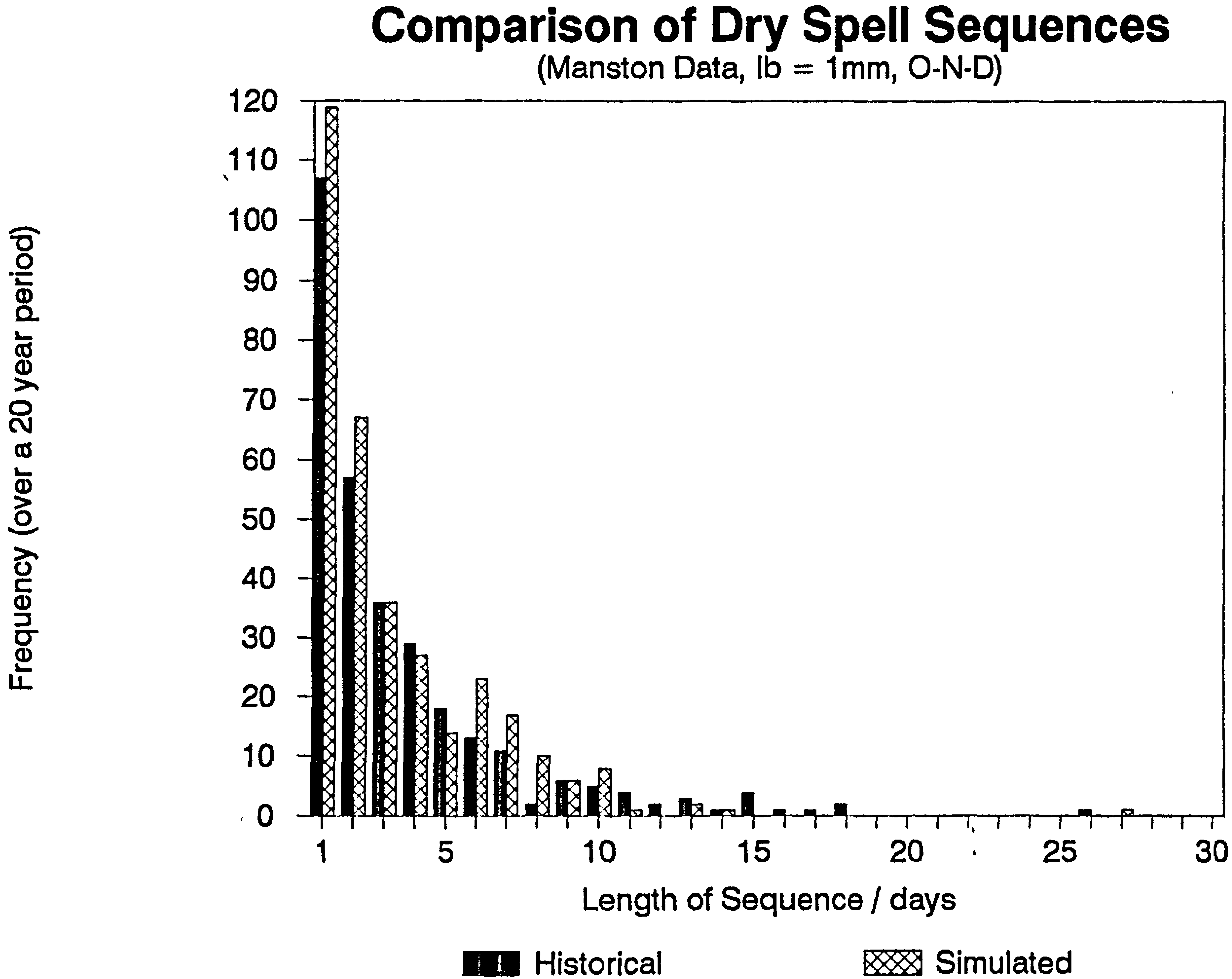
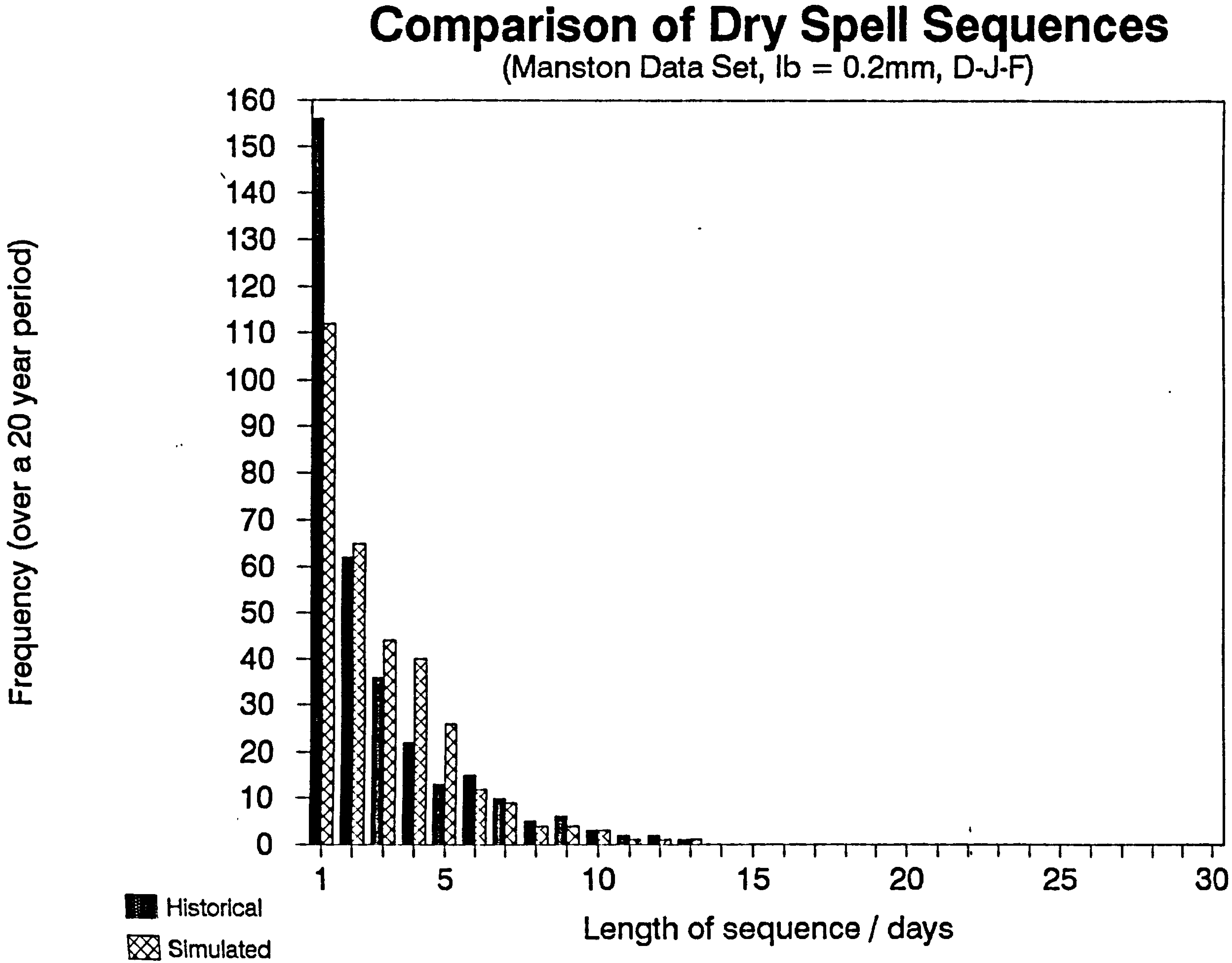

Comparison of Dry Spell Sequences
(Manston Data, lb = 1mm, O-N-D)
Frequency (over a 20 year period)
120
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of Sequence / days
Historical
Simulated

Figure E.22



Comparison of Dry Spell Sequences
(Manston Data Set, lb = 0.2mm, D-J-F)
Frequency (over a 20 year period)
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of sequence / days
Historical
Simulated

Figure E.23

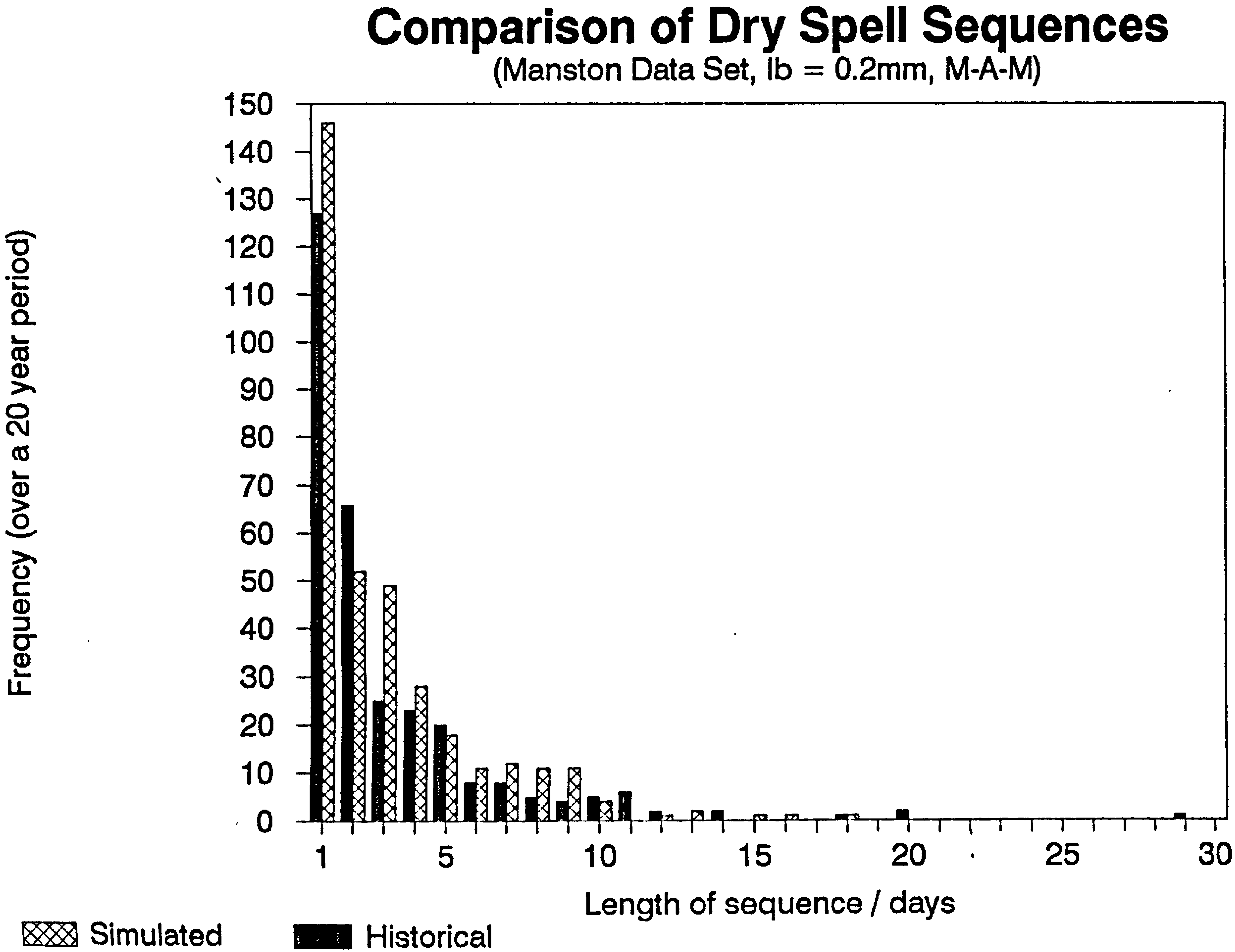

Comparison of Dry Spell Sequences
(Manston Data Set, lb = 0.2mm, M-A-M)
Frequency (over a 20 year period)
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of sequence / days
Simulated
Historical

Figure E.24

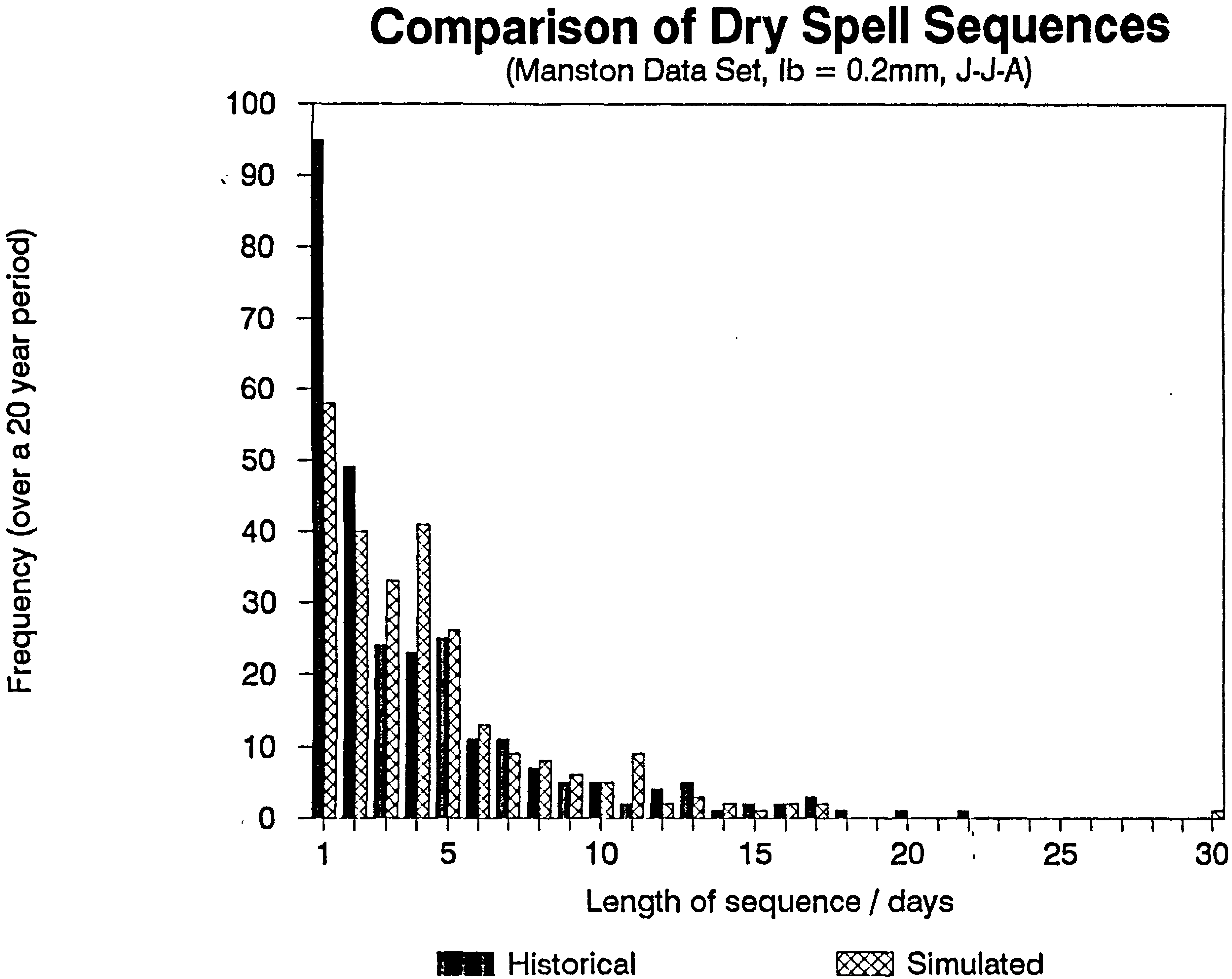

Comparison of Dry Spell Sequences
(Manston Data Set, lb = 0.2mm, J-J-A)
Frequency (over a 20 year period)
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of sequence / days
Historical
Simulated

Figure E.25

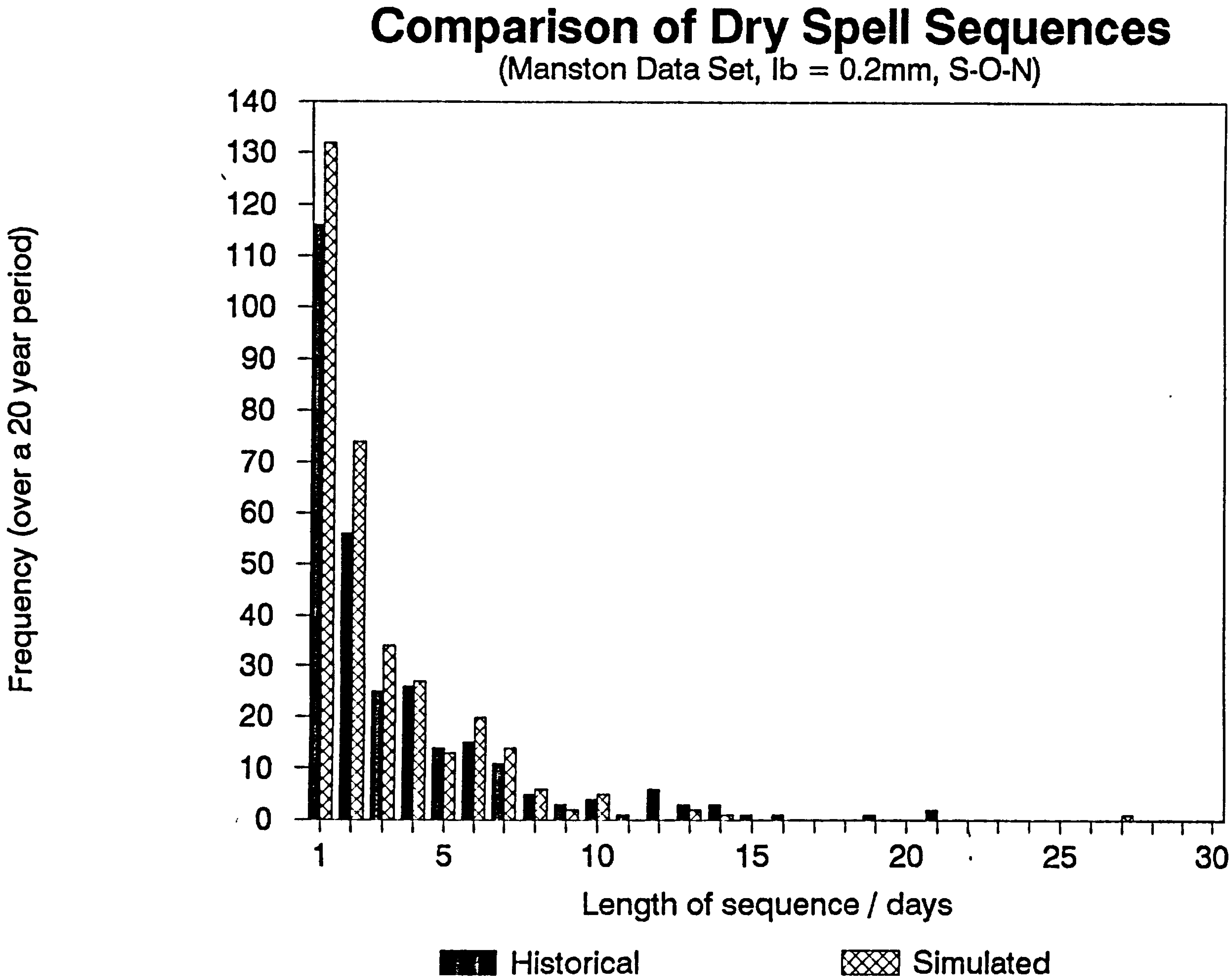

Comparison of Dry Spell Sequences
(Manston Data Set, lb = 0.2mm, S-O-N)
Frequency (over a 20 year period)
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
5
10
15
20
25
30
Length of sequence / days
Historical
Simulated

Figure E.26

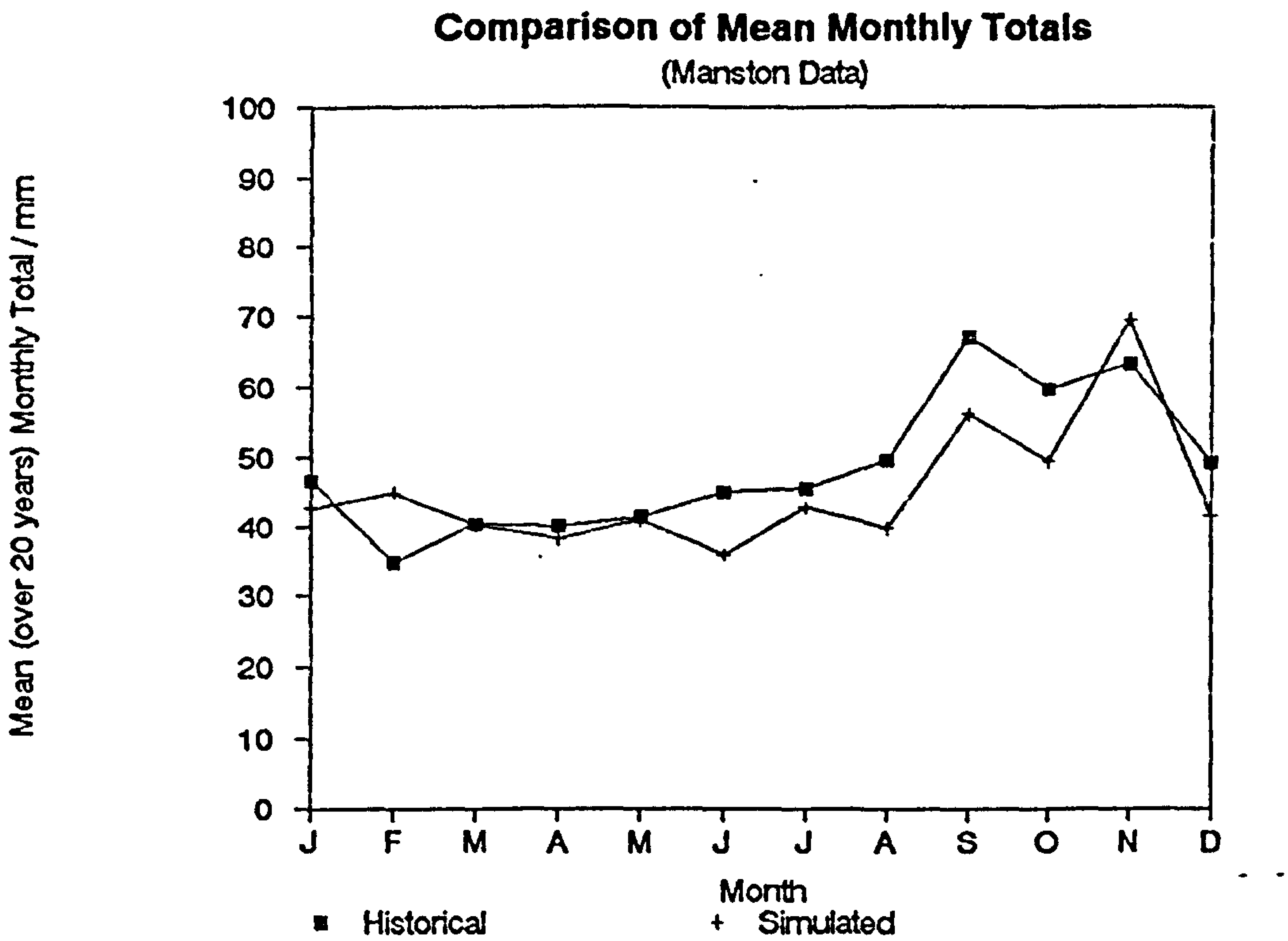


Figure E.27

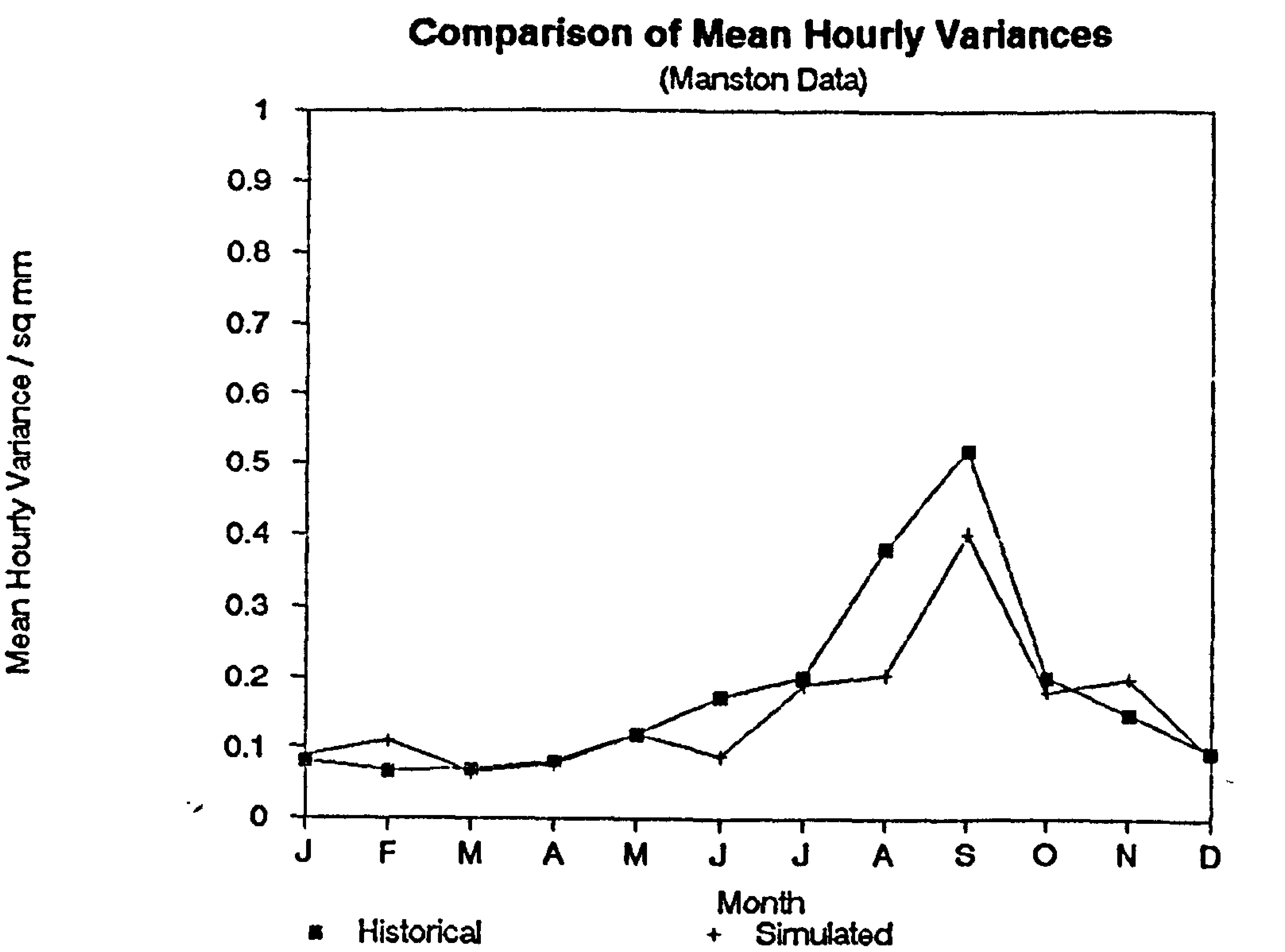


Figure E.28

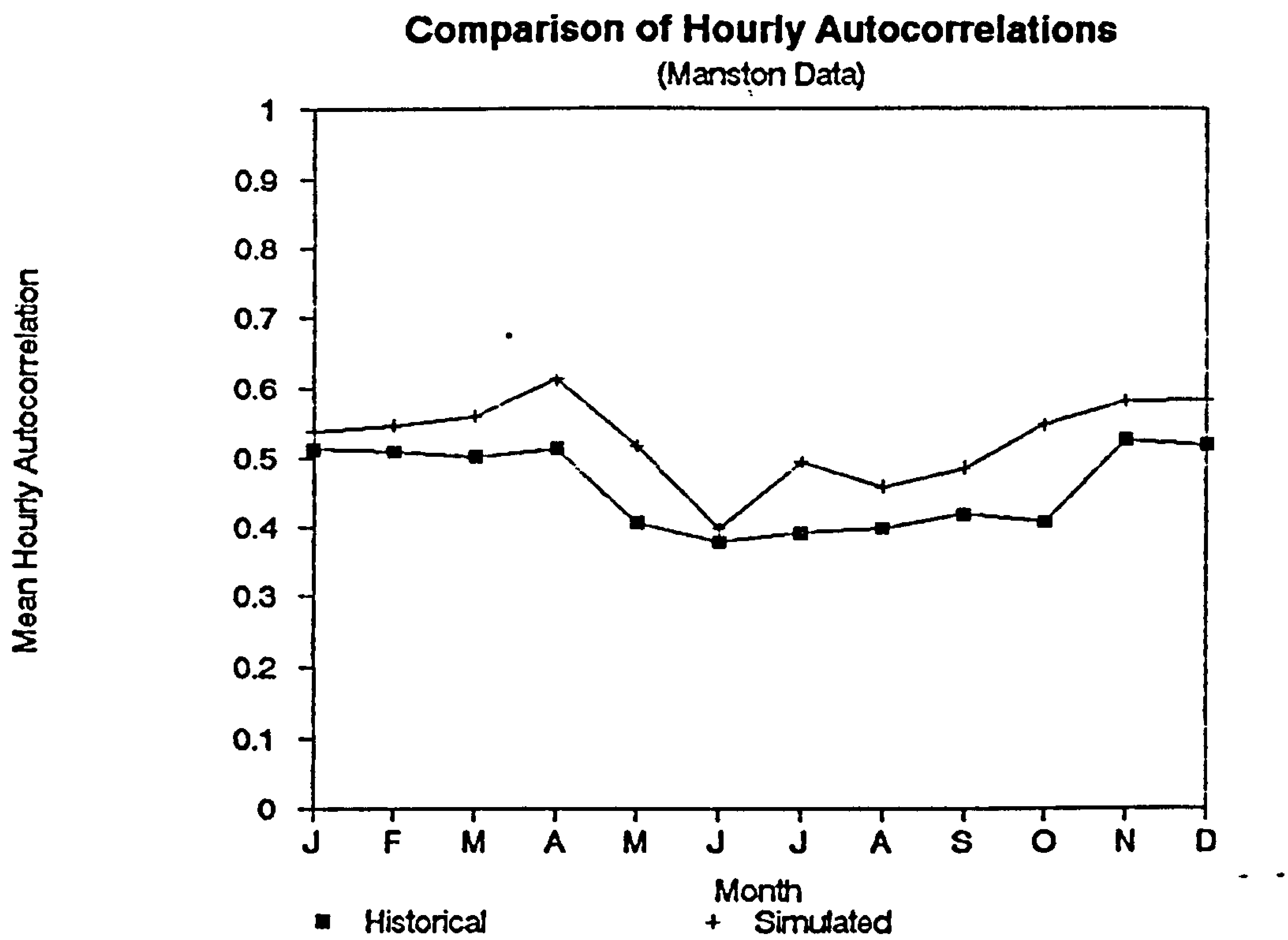


Figure E.29

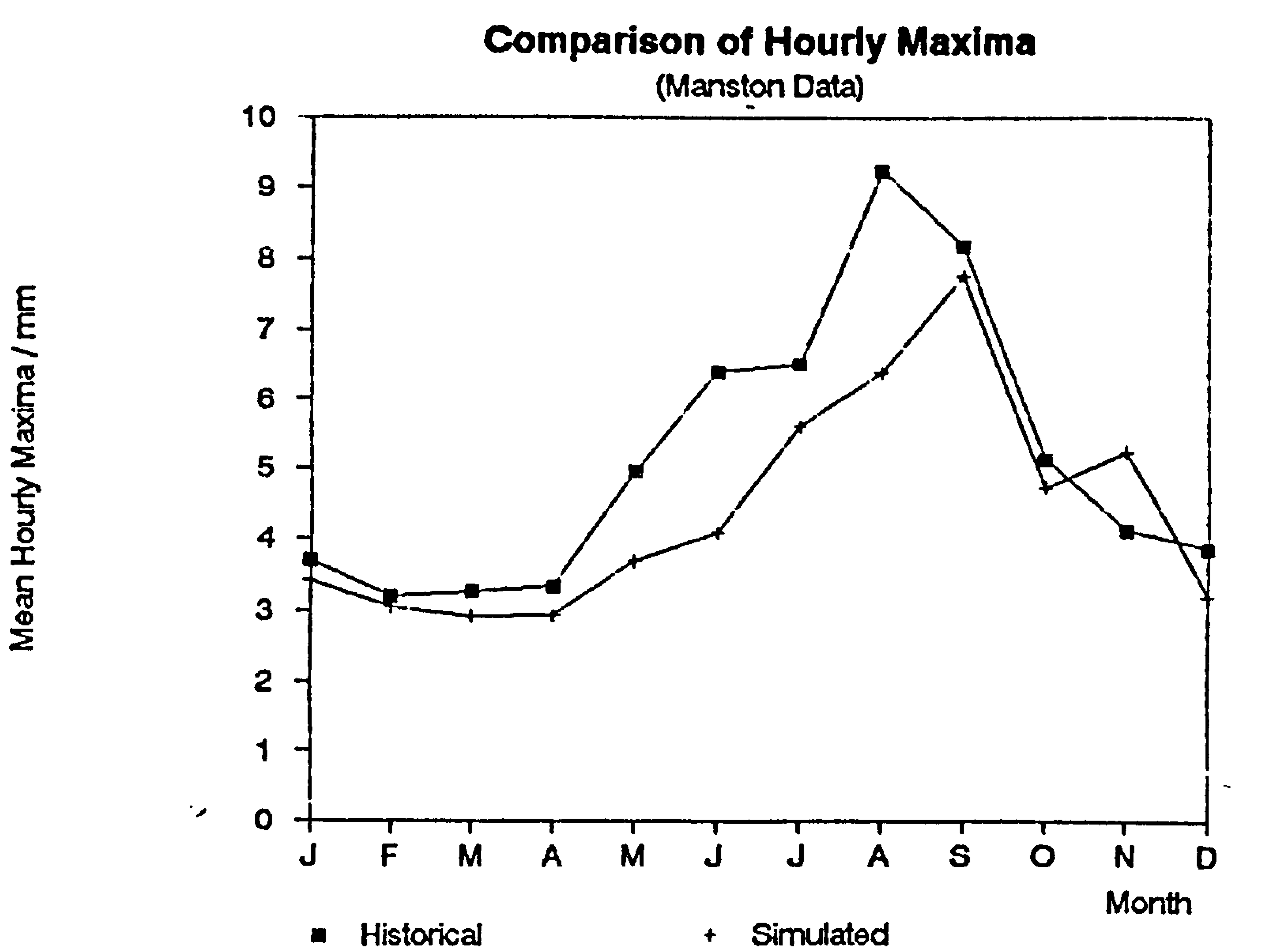


Figure E.30

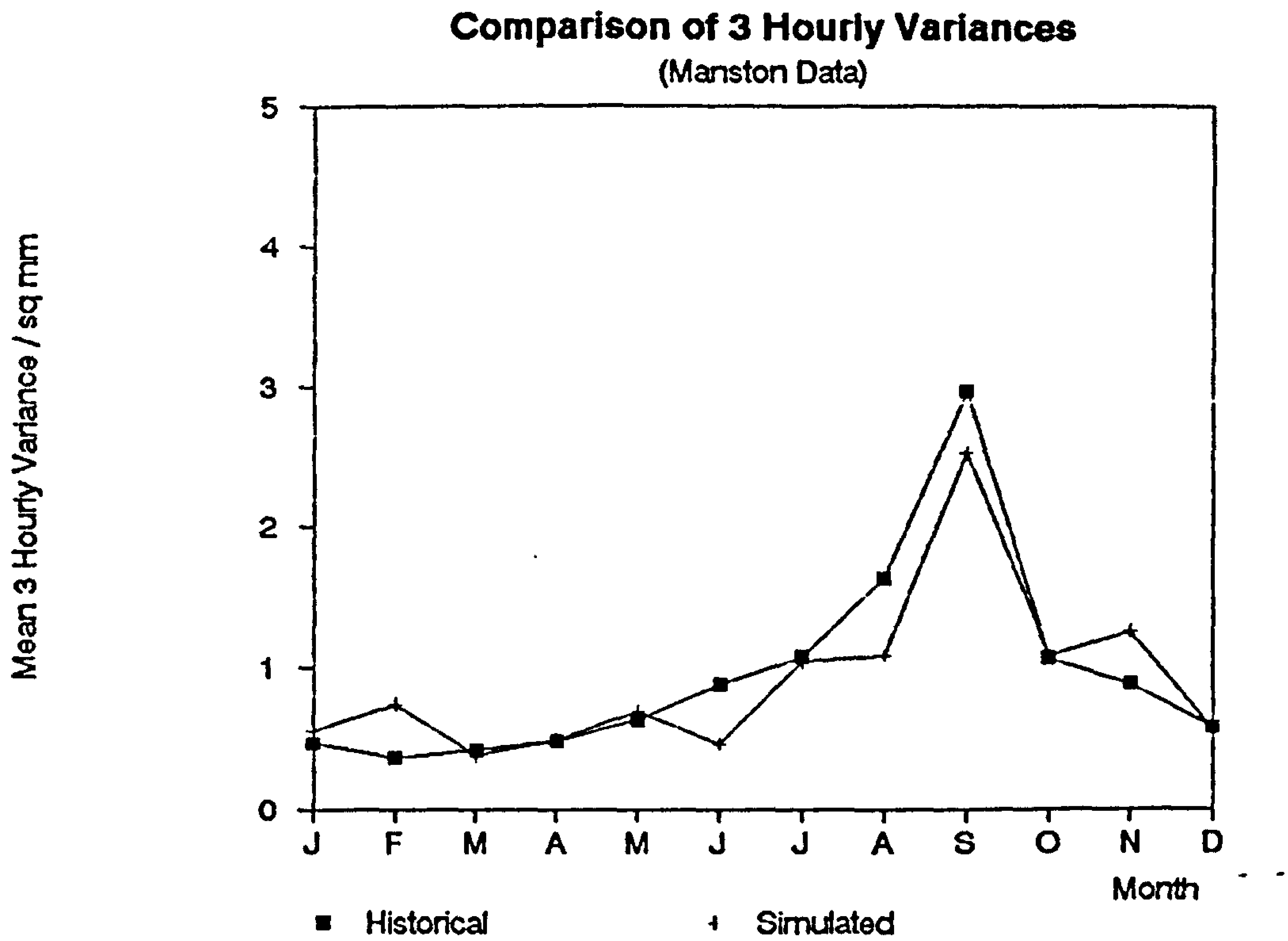


Figure E.31

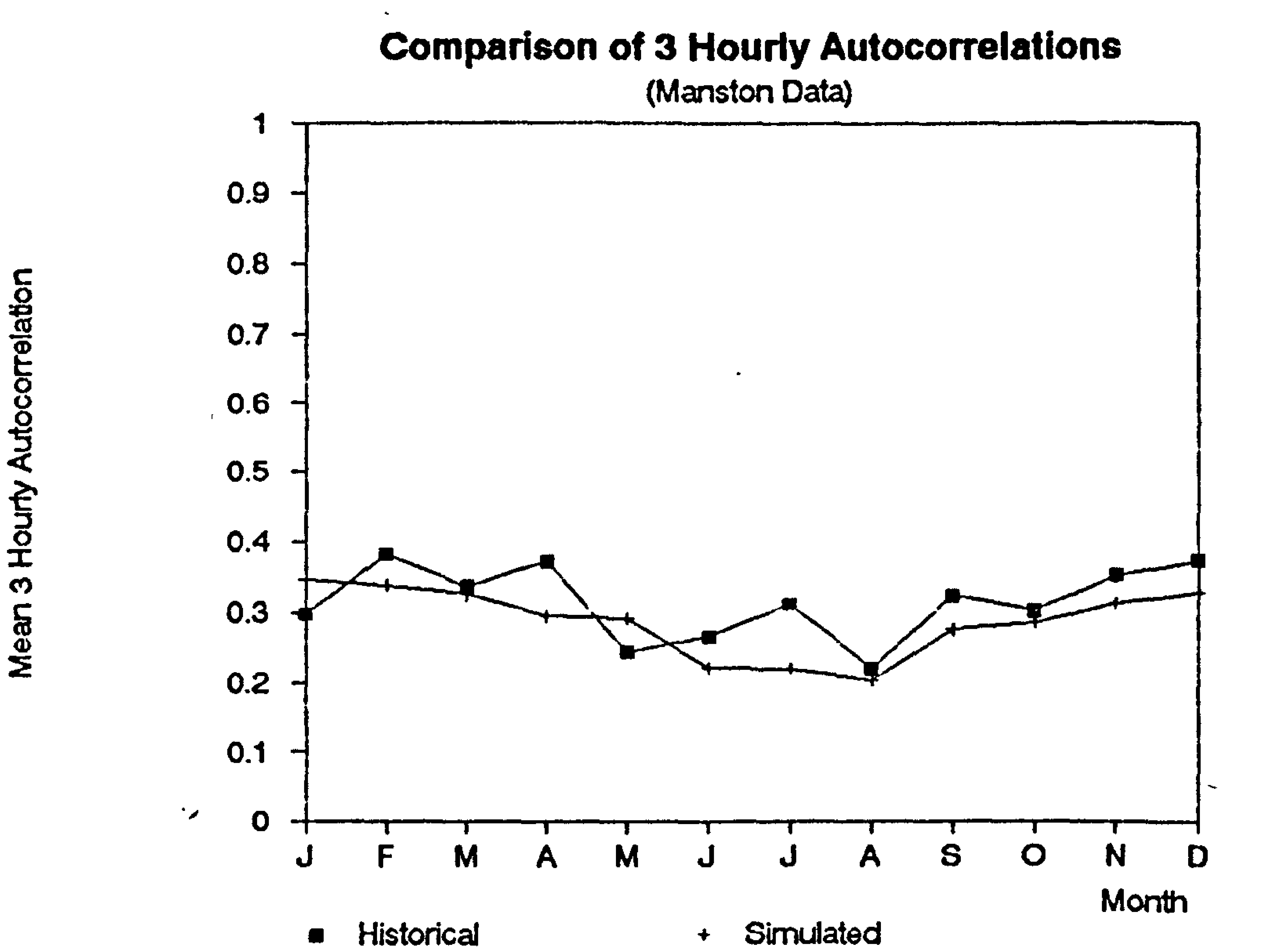


Figure E.32

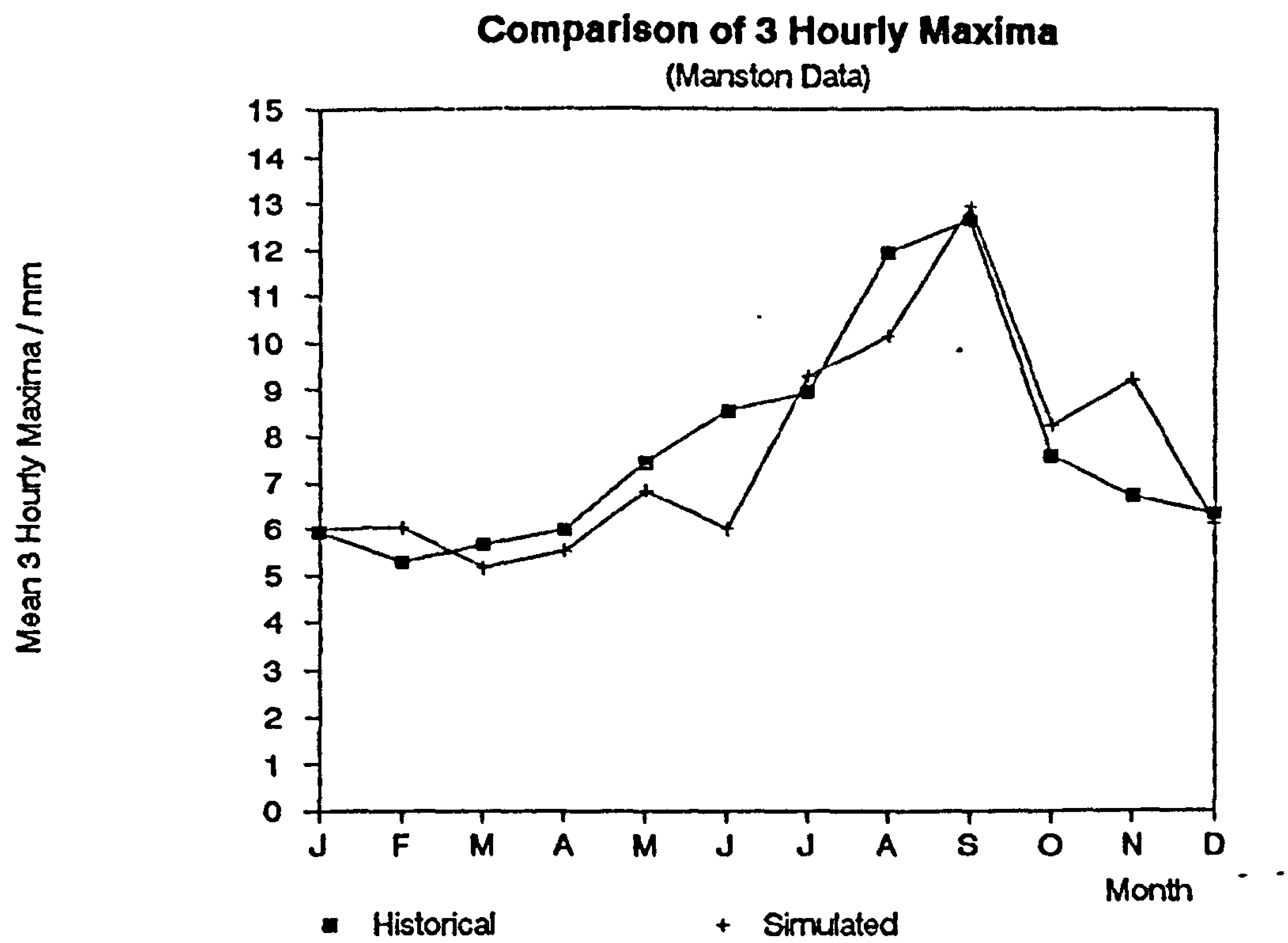


Figure E.33

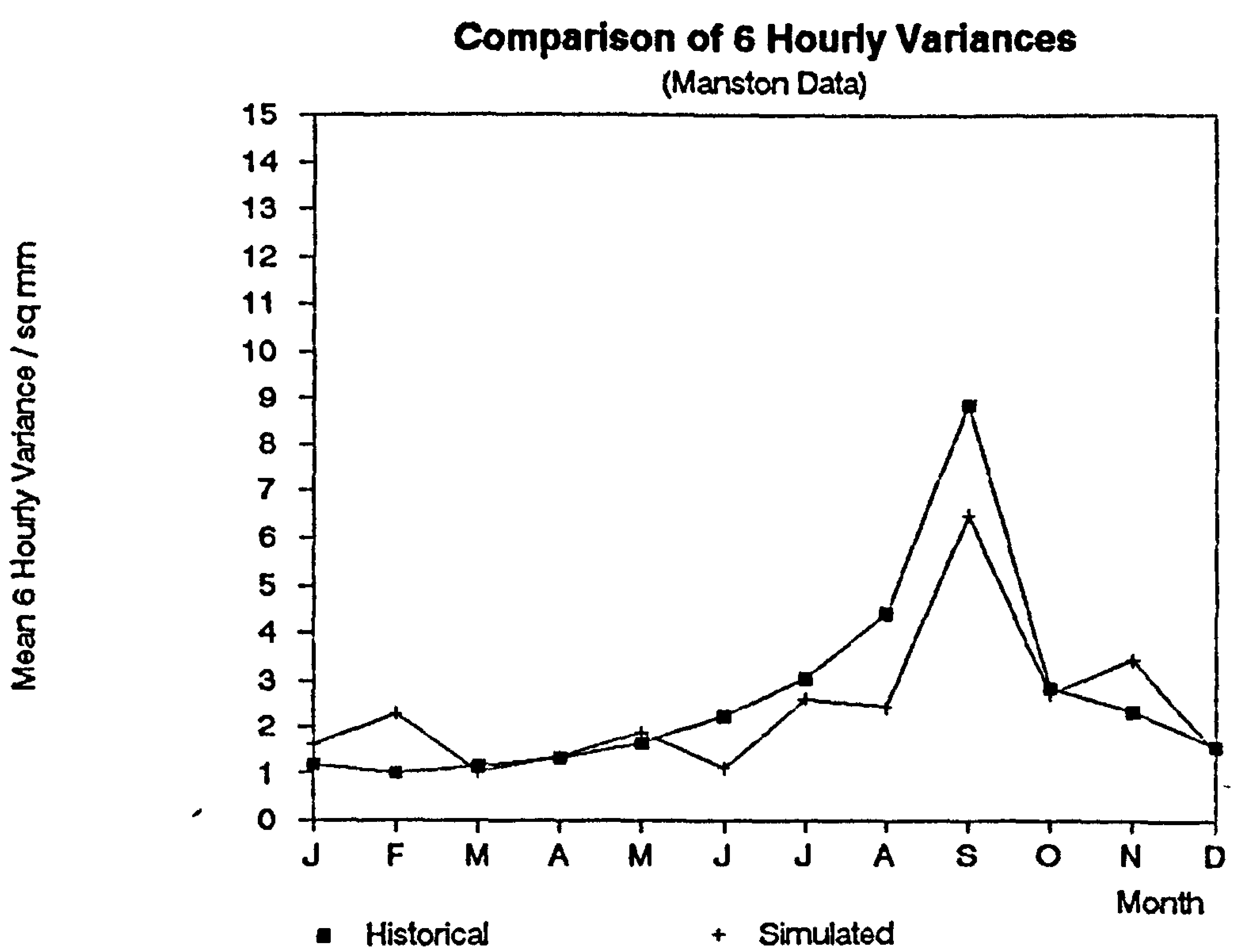


Figure E.34

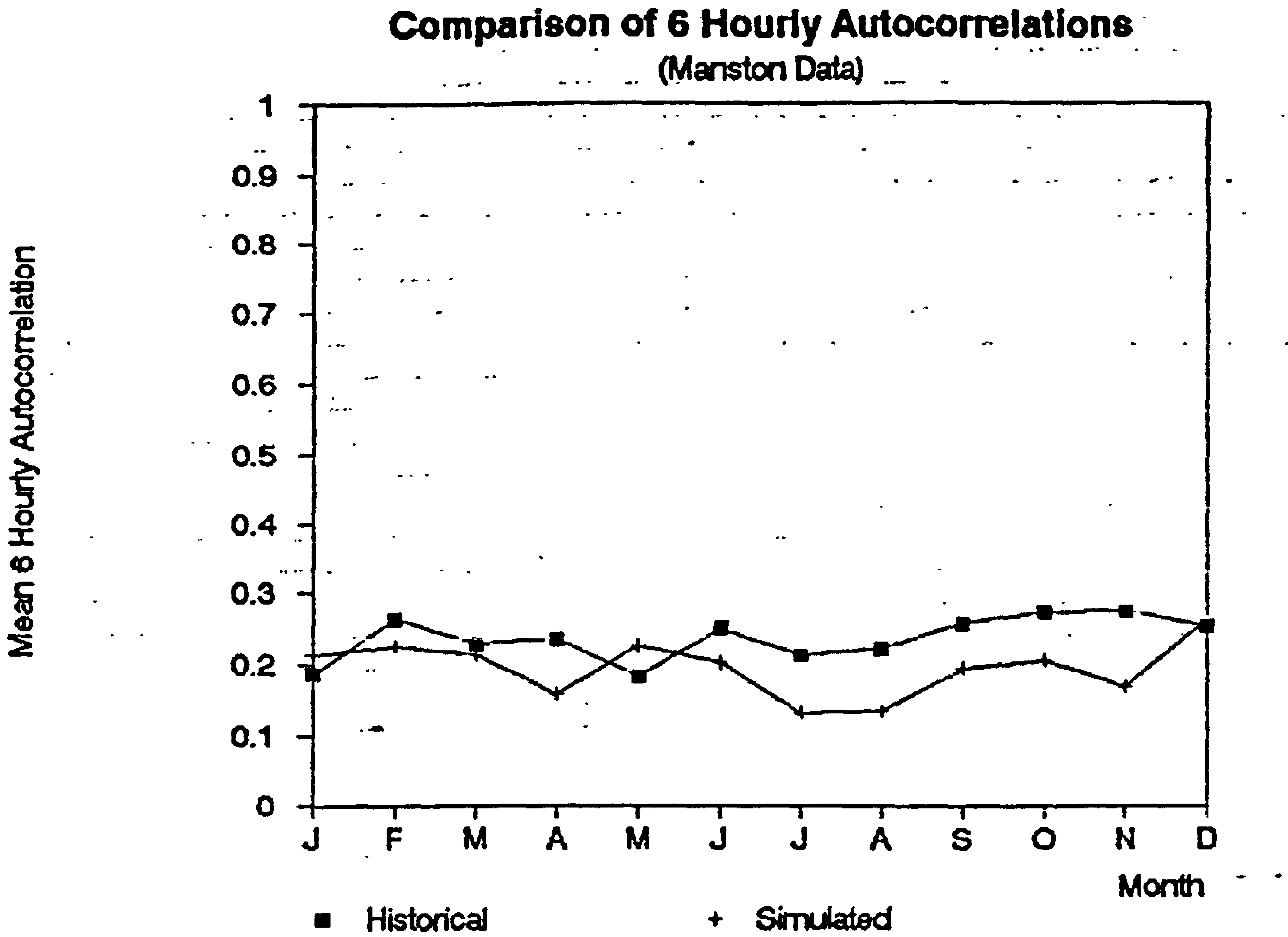


Figure E.35

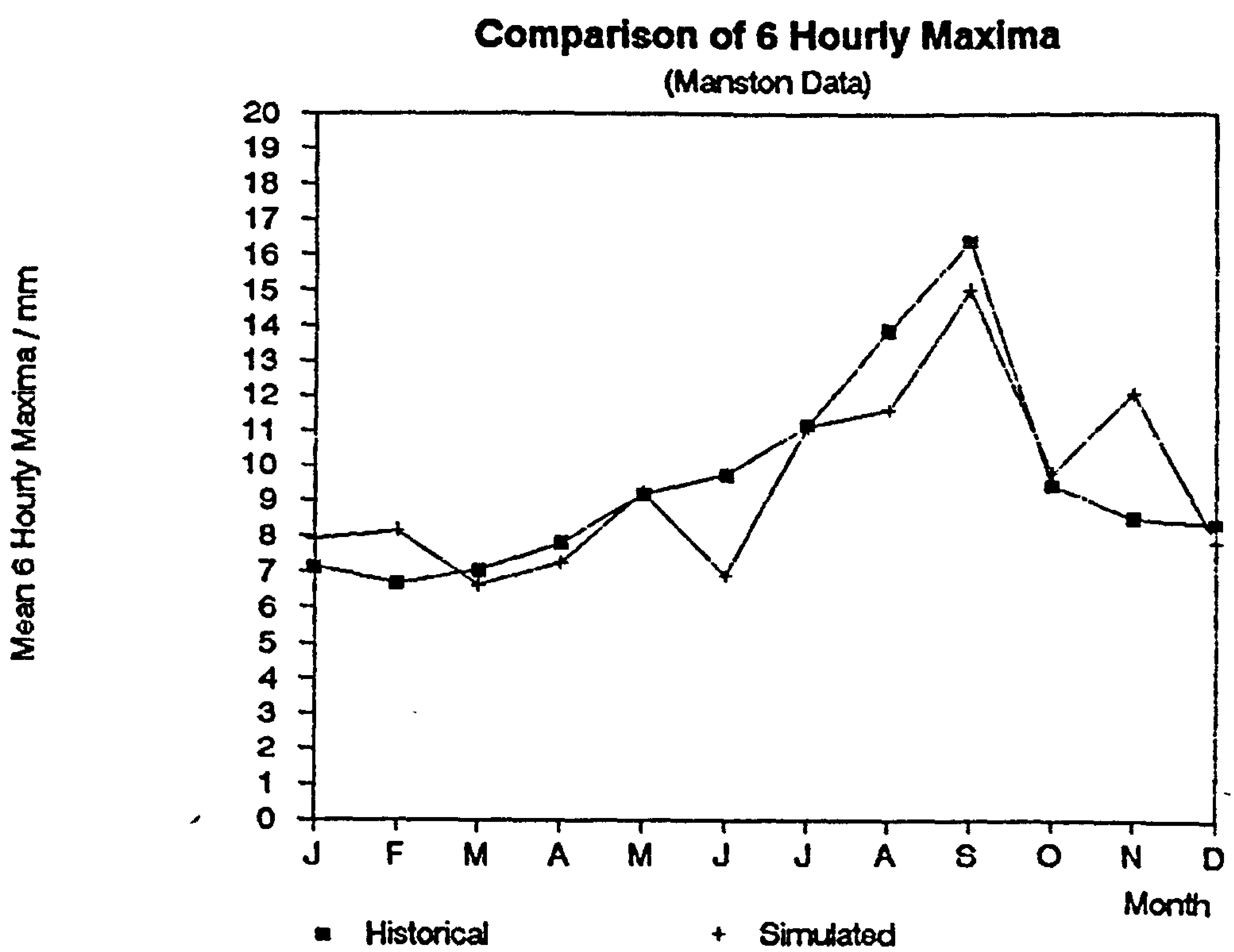


Figure E.36

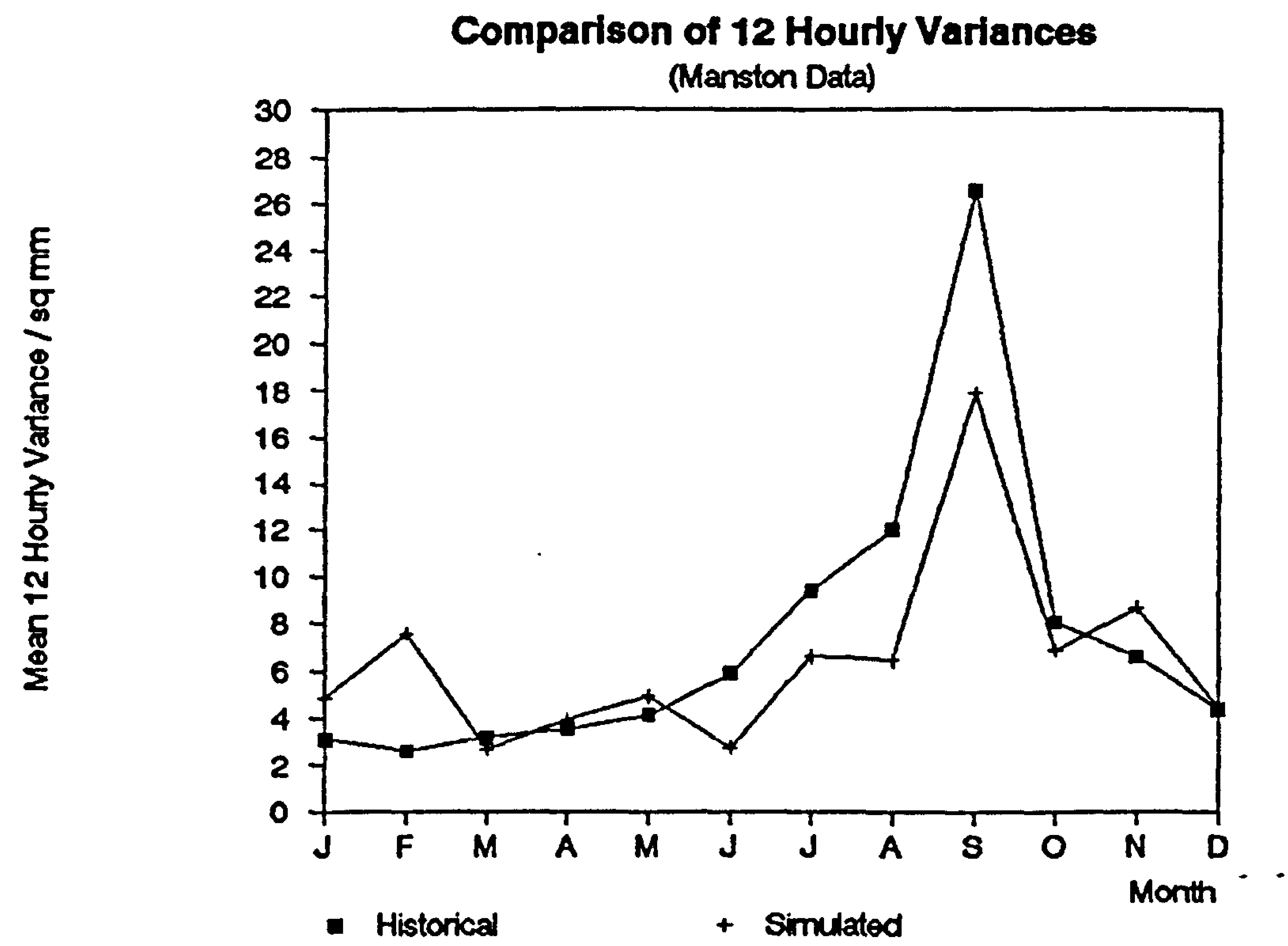


Figure E.37

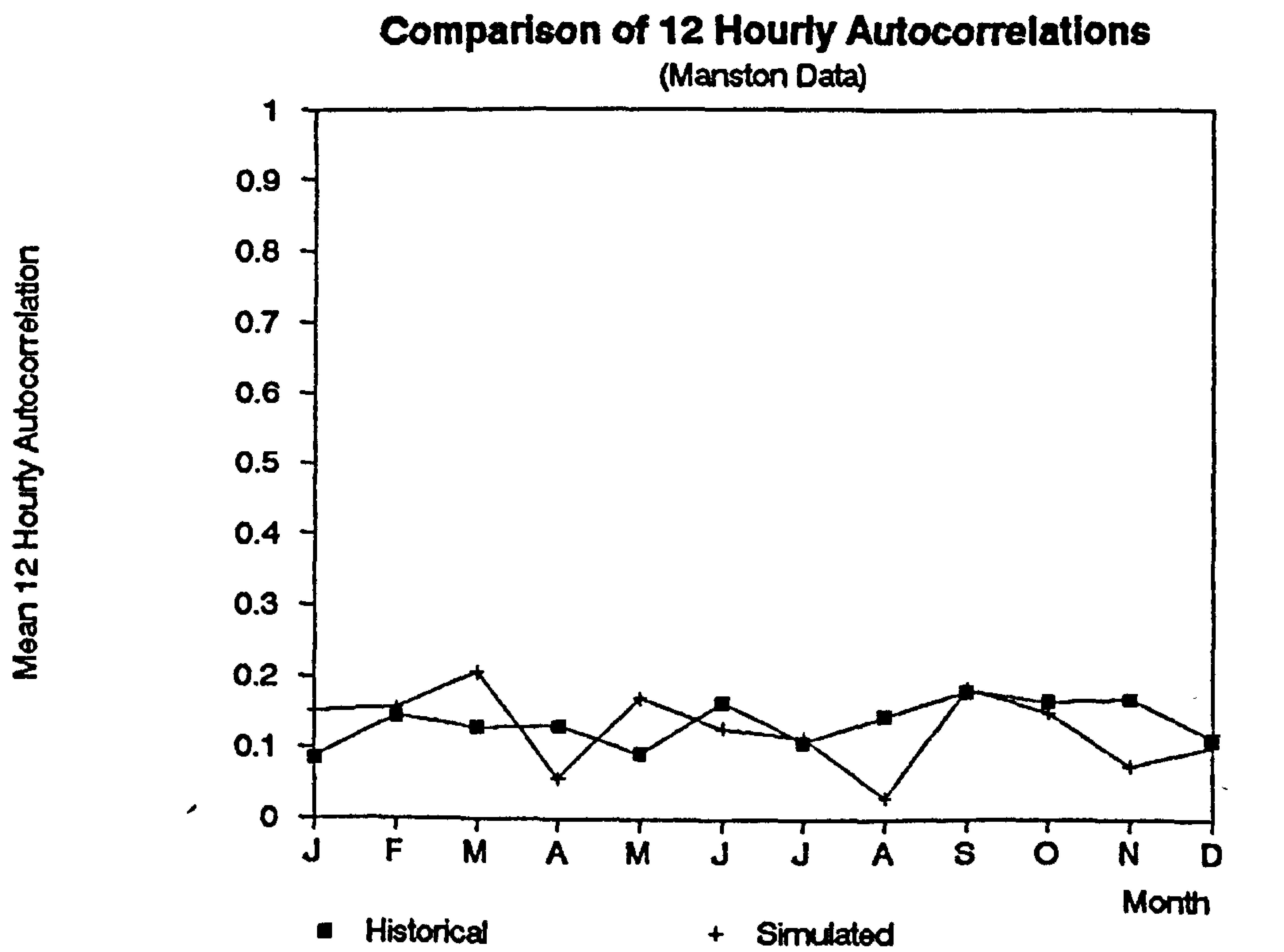


Figure E.38

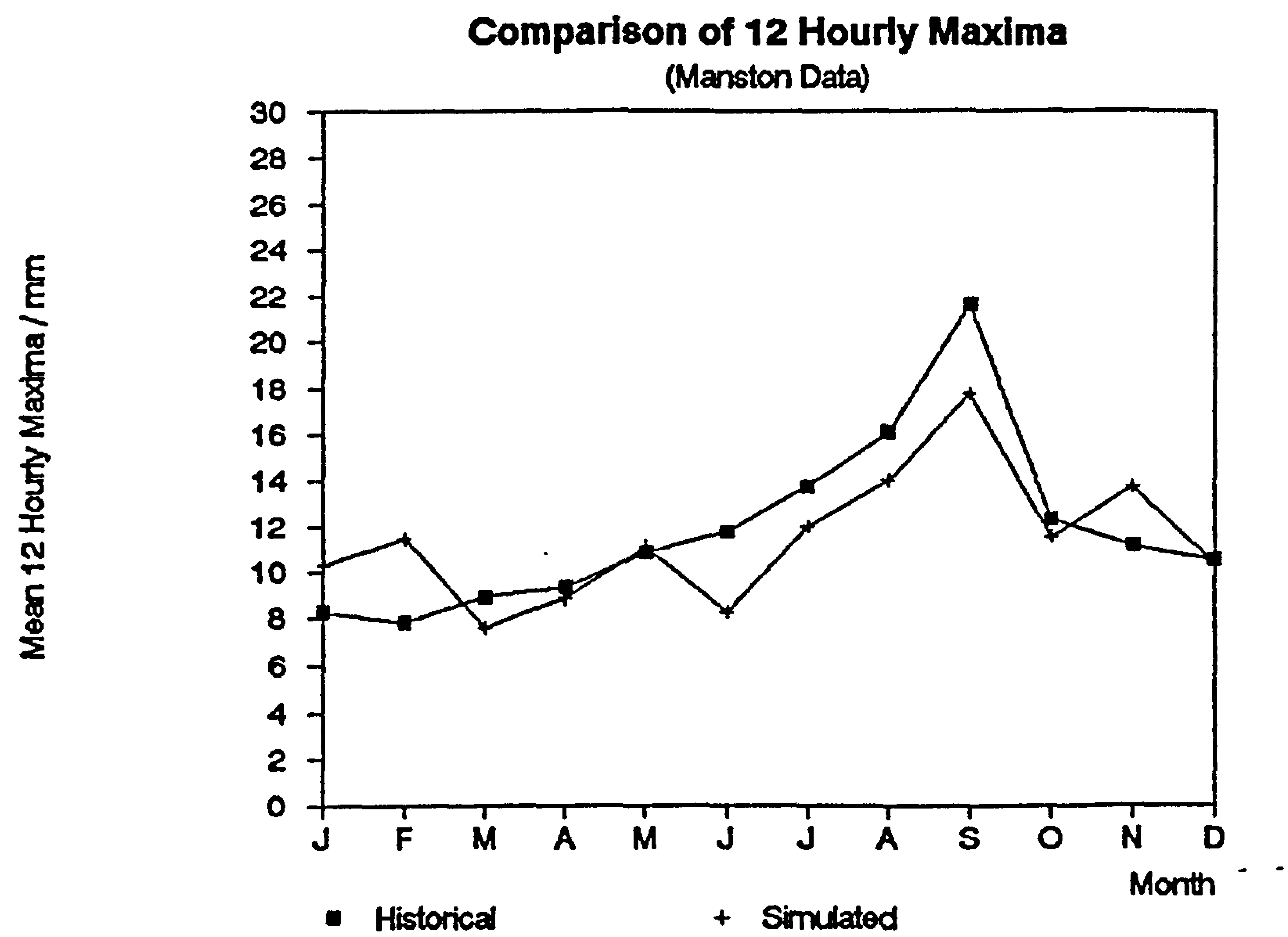


Figure E.39

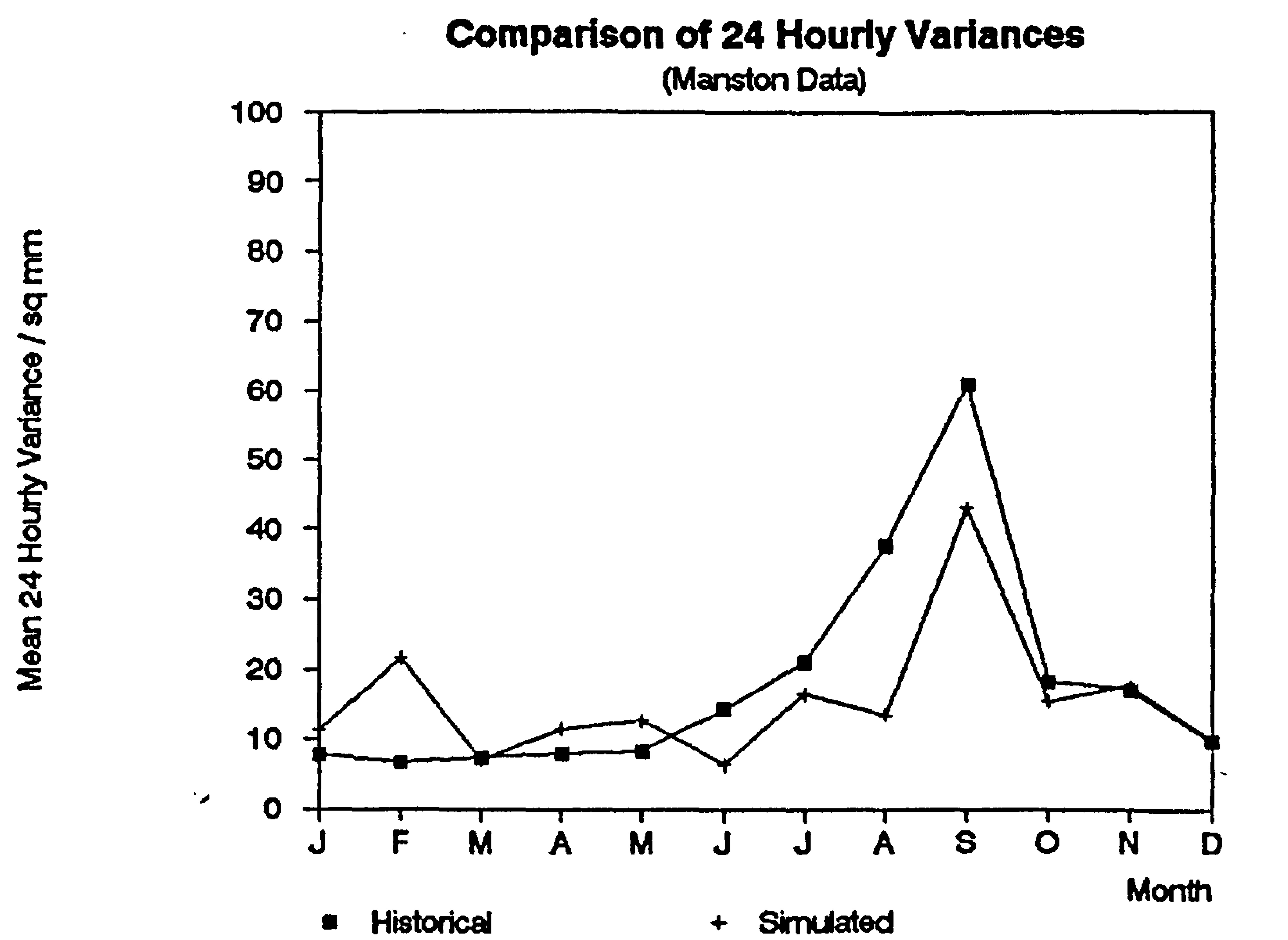


Figure E.40

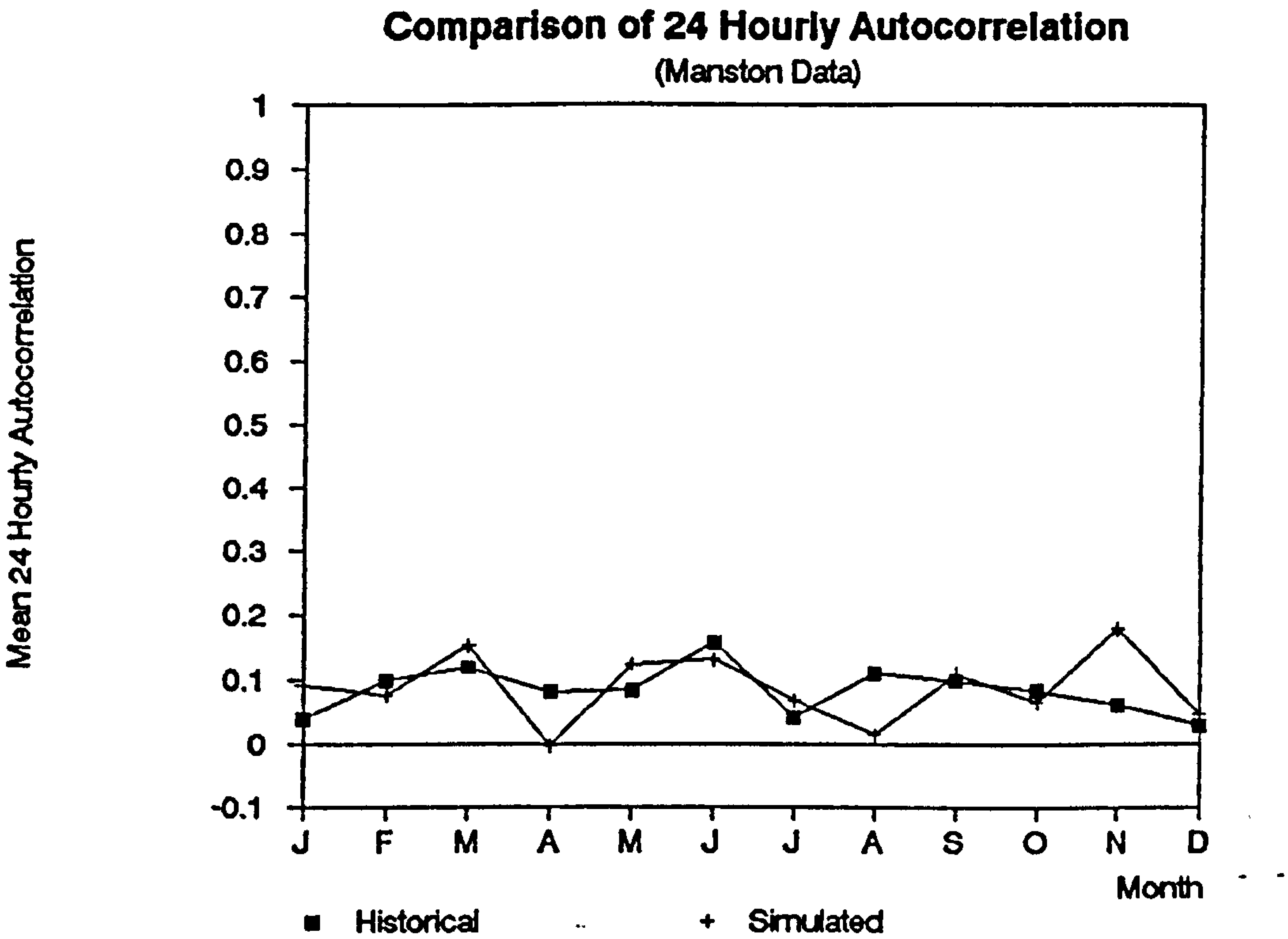


Figure E.41

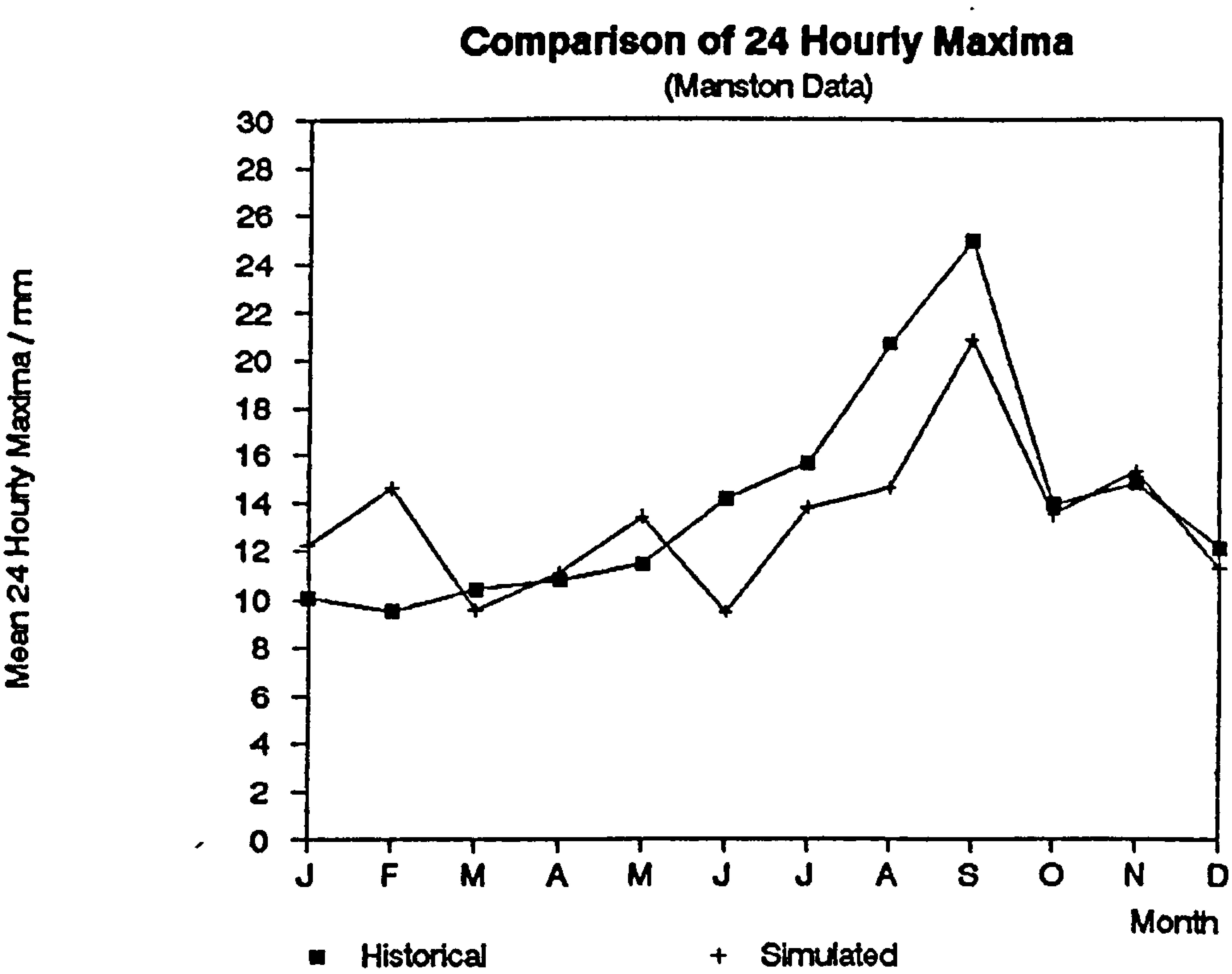


Figure E.42

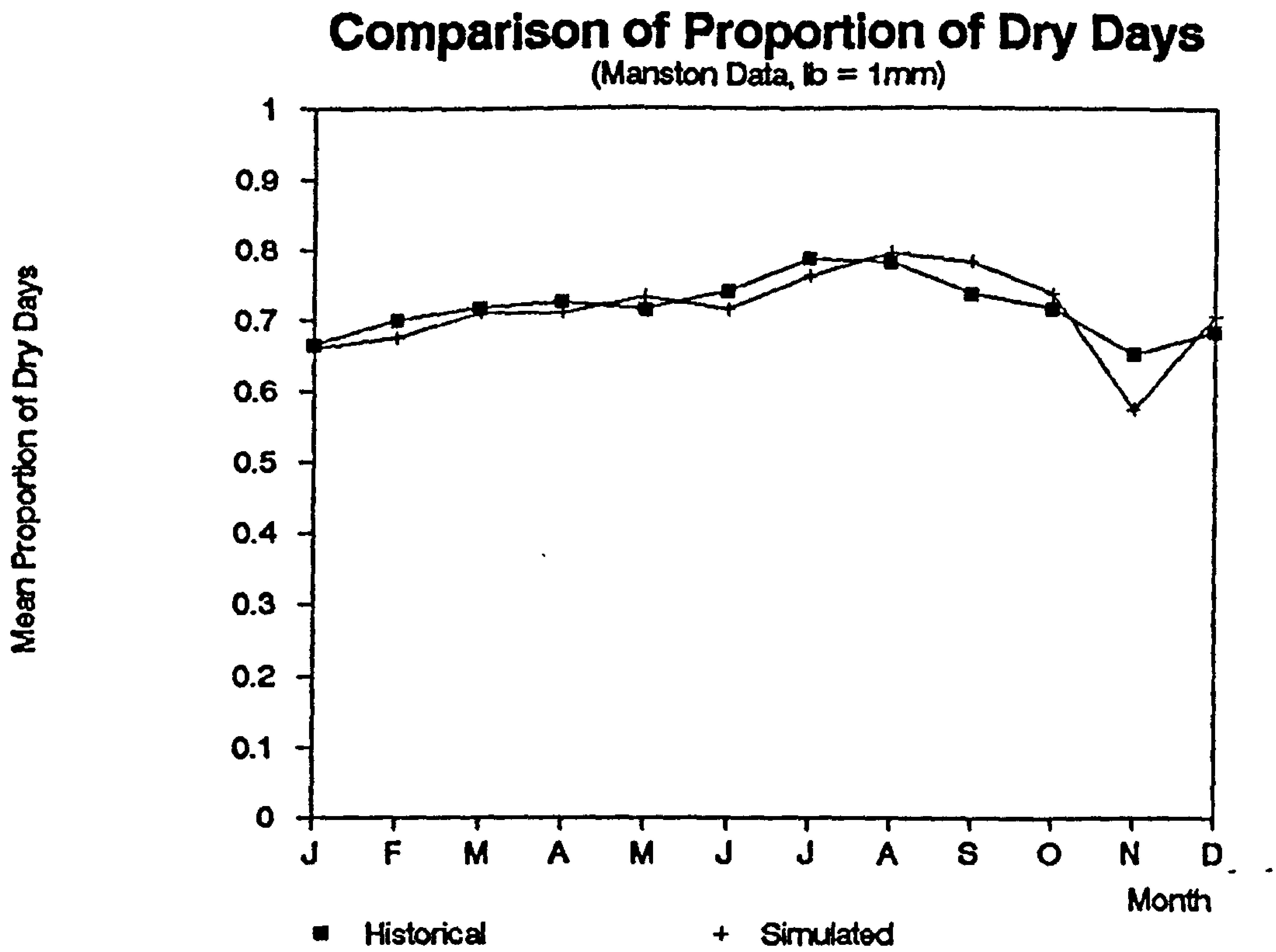


Figure E.43

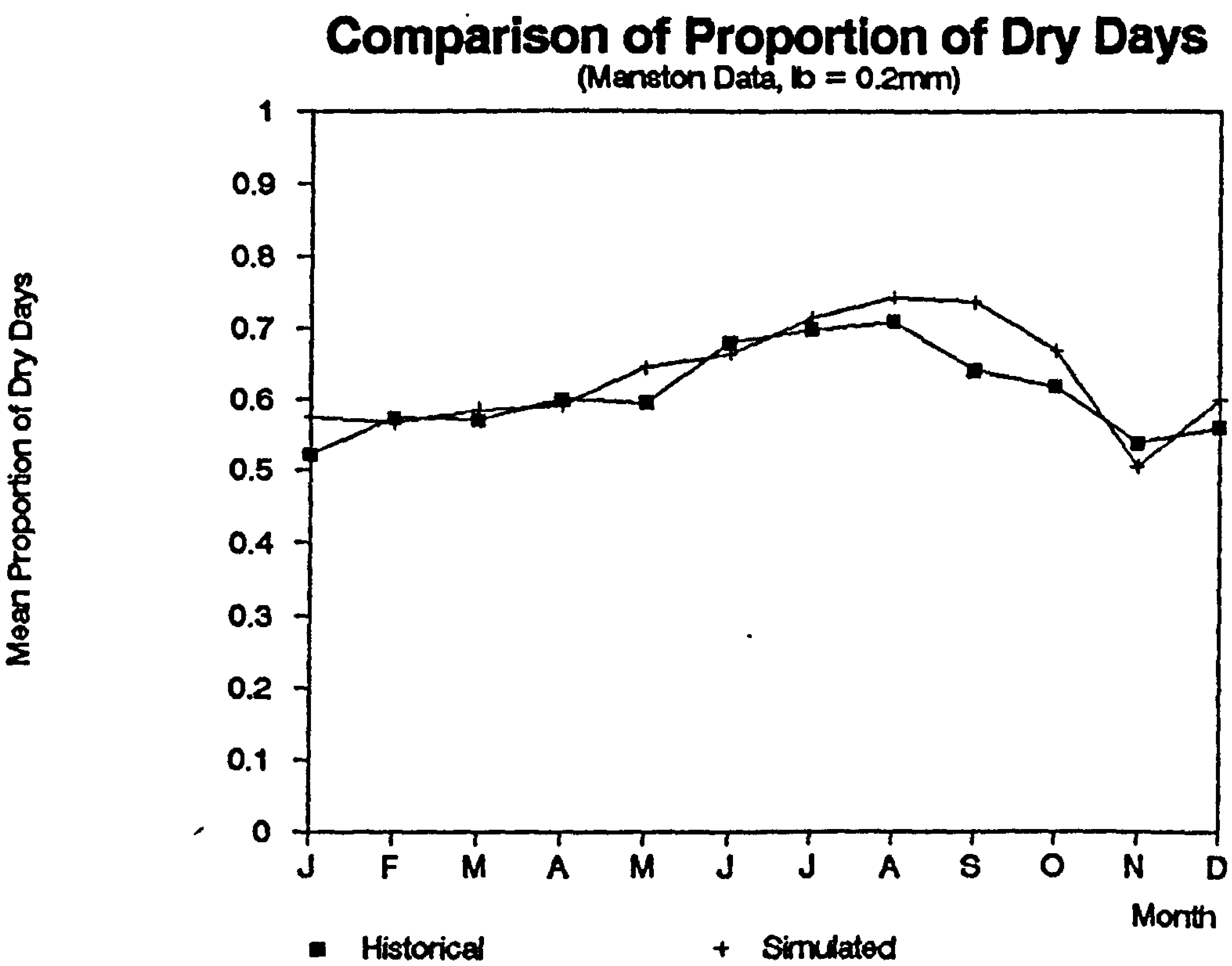


Figure E.44

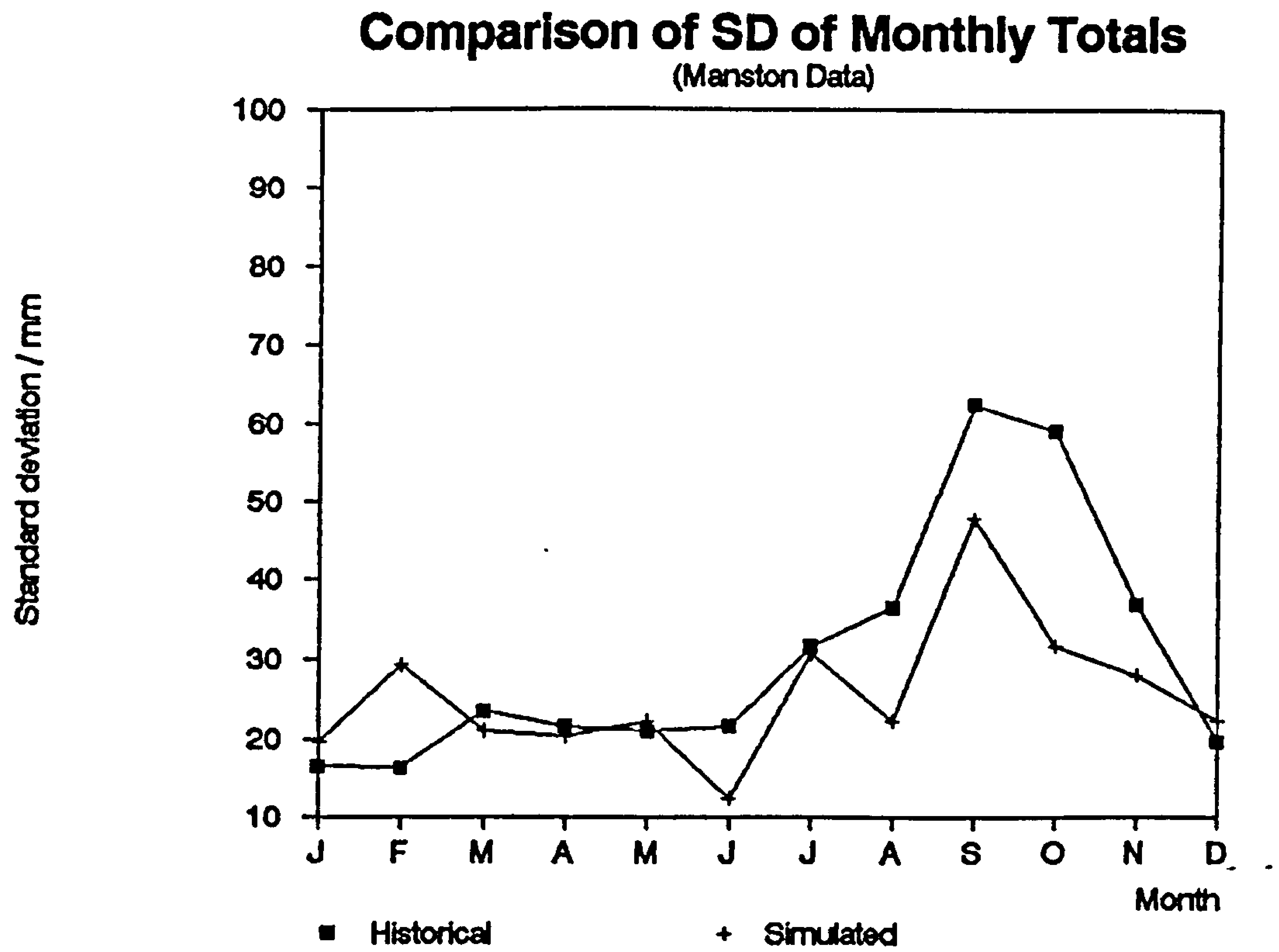


Figure E.45

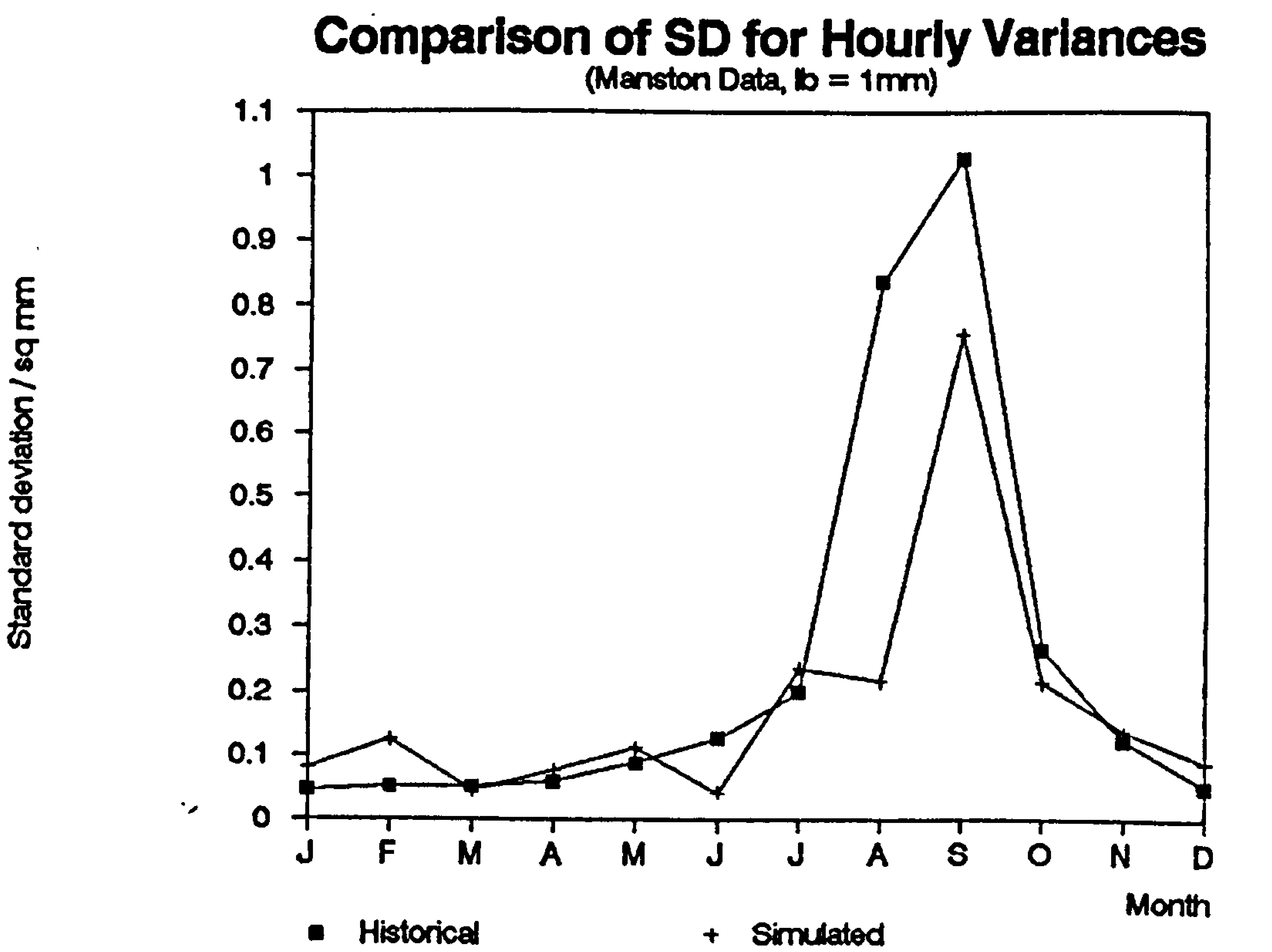


Figure E.46

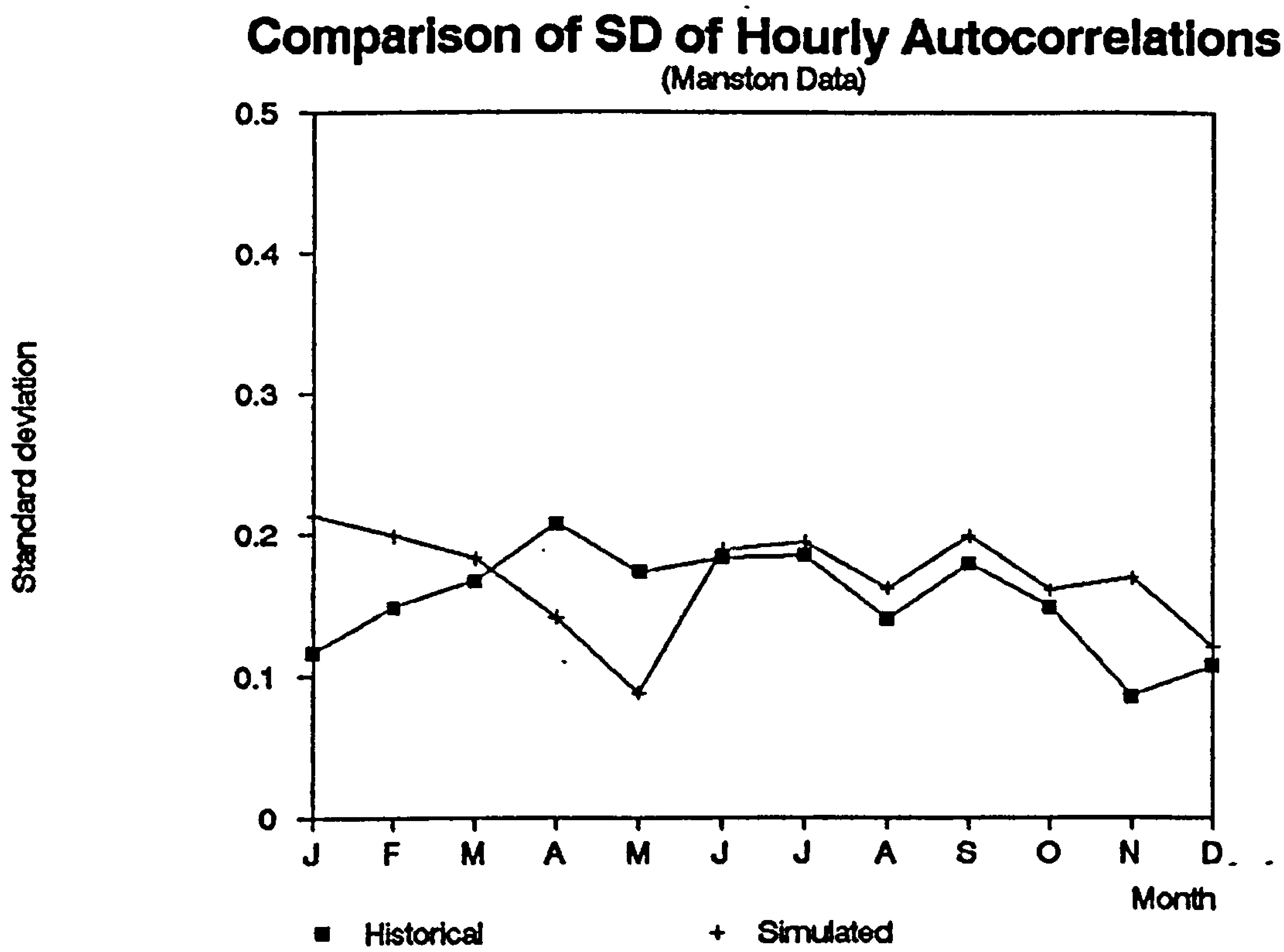


Figure E.47

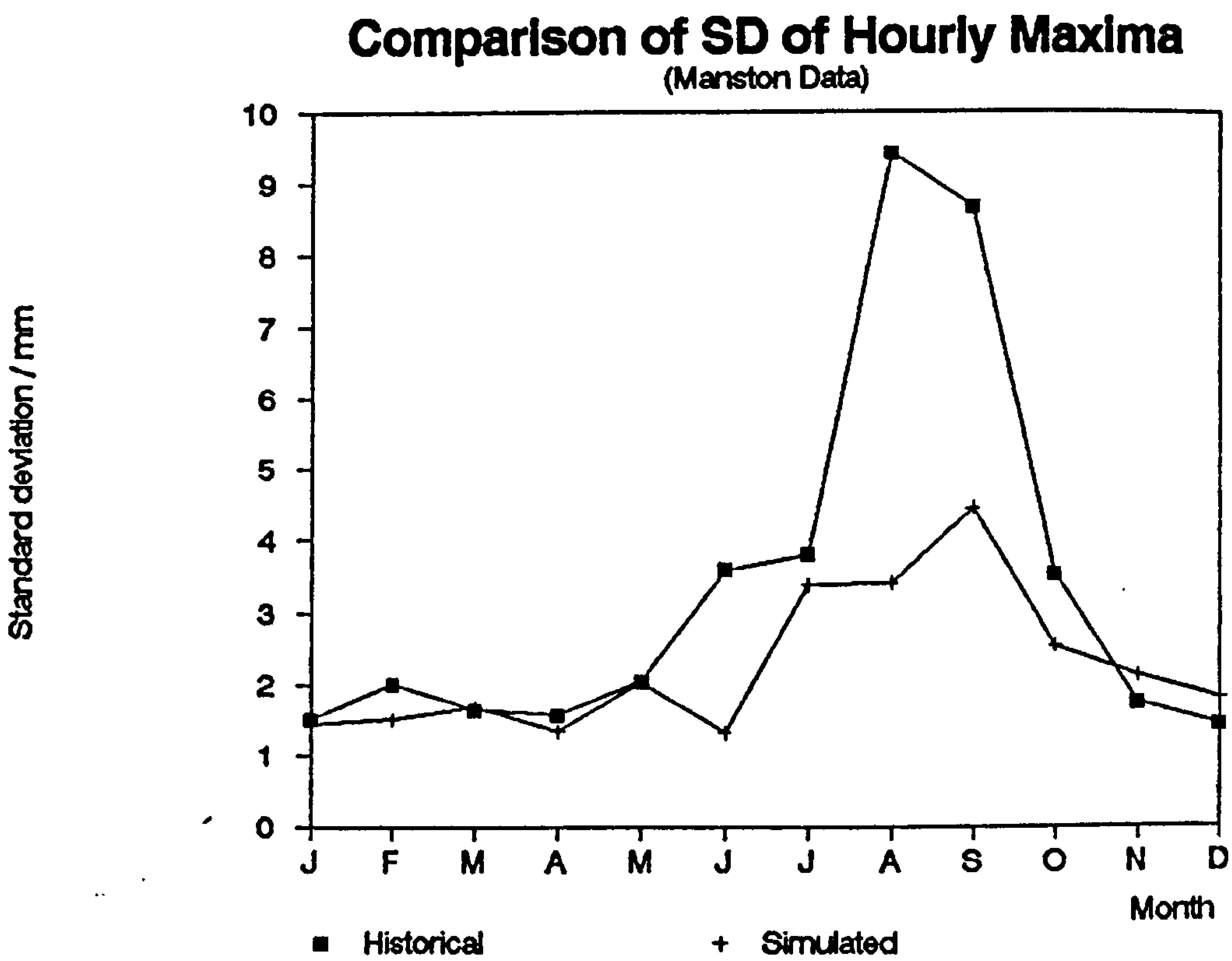


Figure E.48

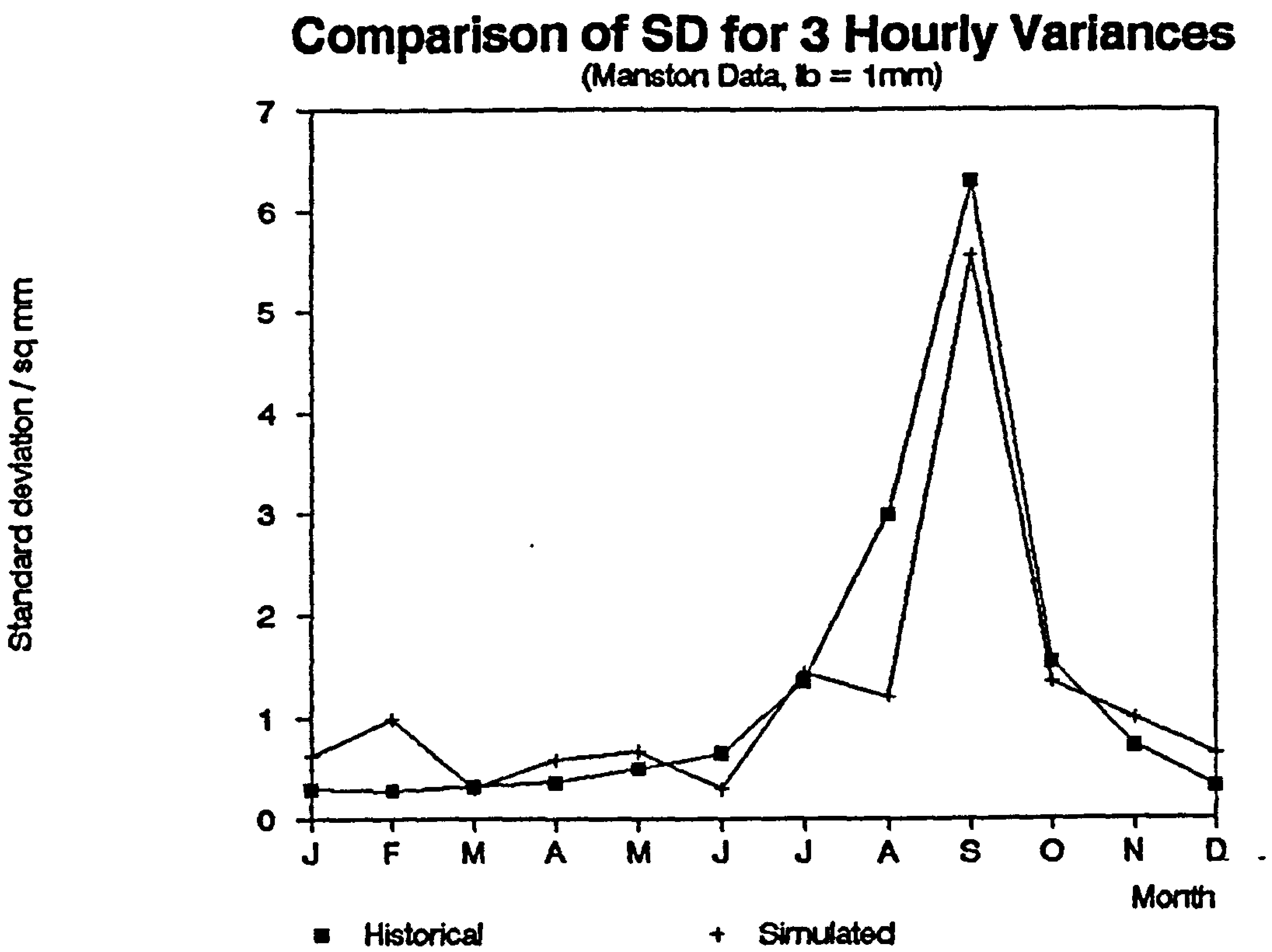


Figure E.49

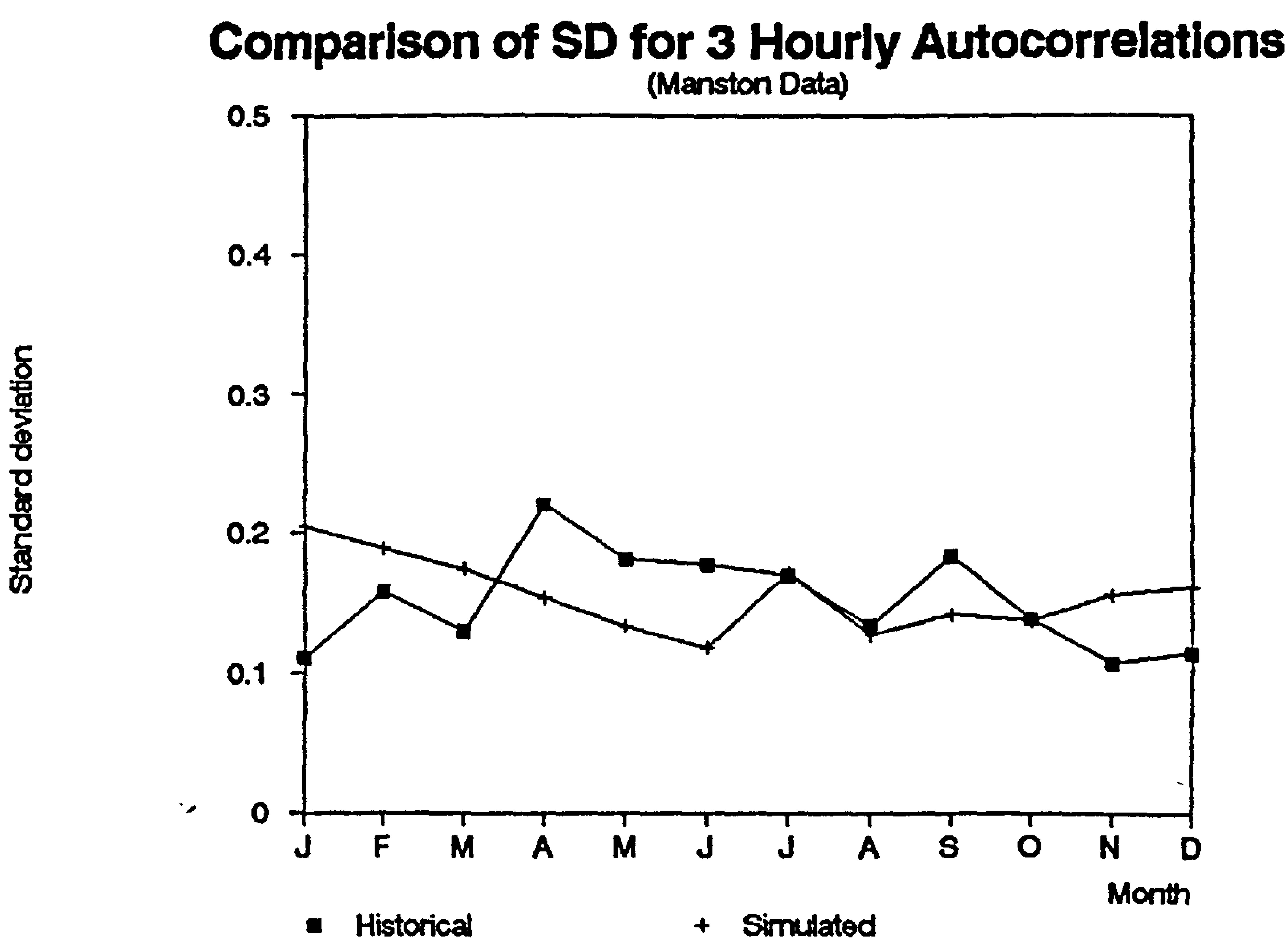


Figure E.50

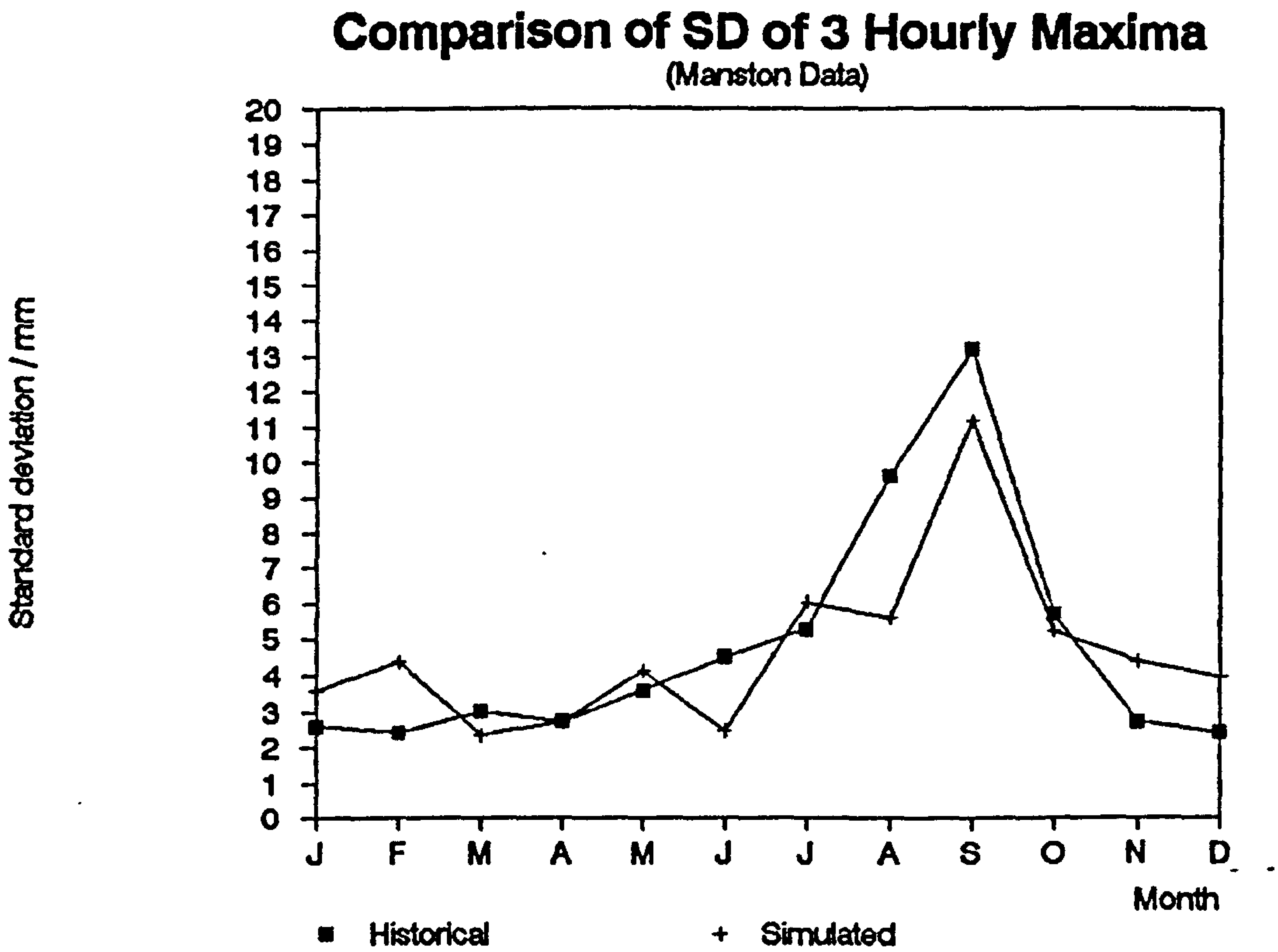


Figure E.51

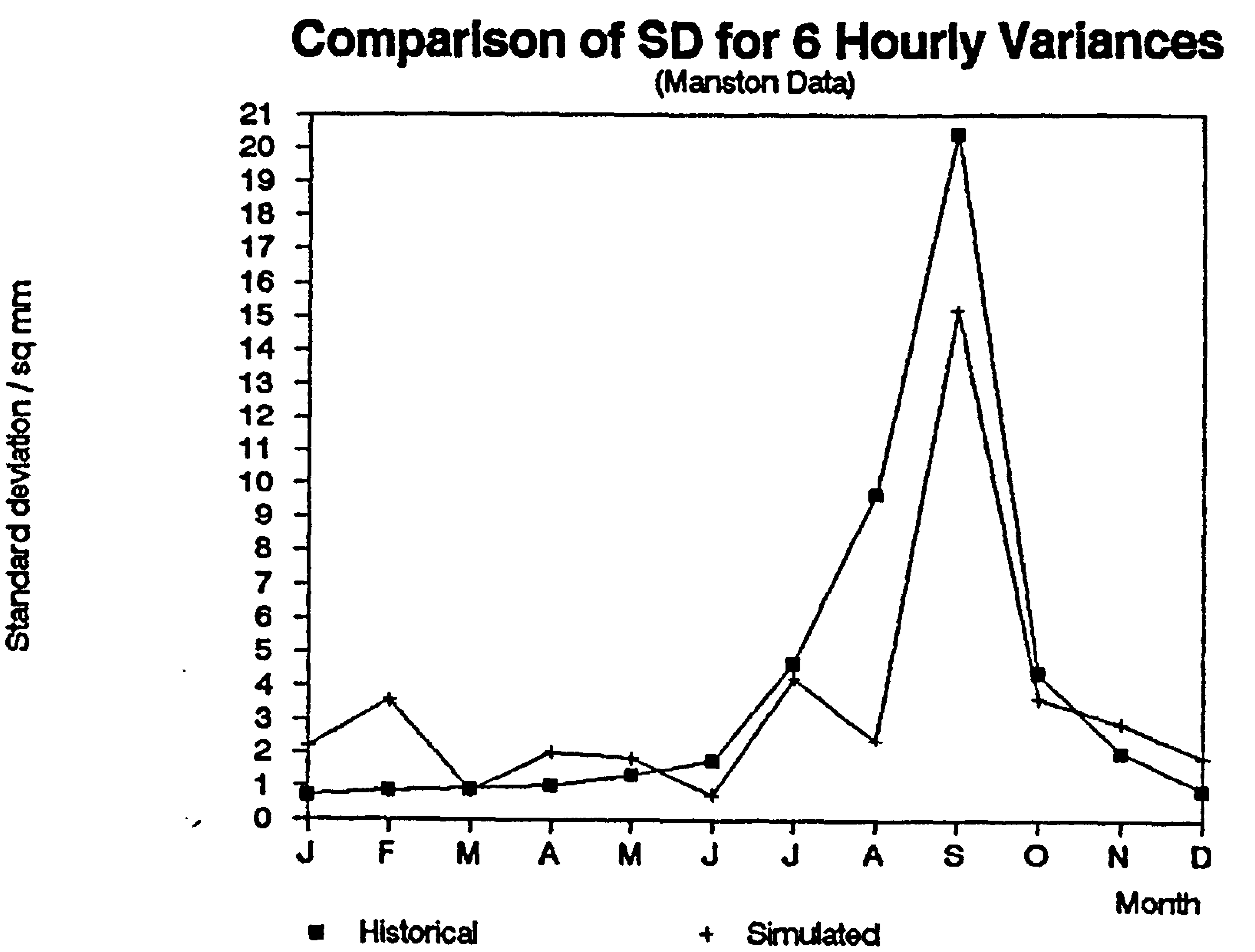


Figure E.52

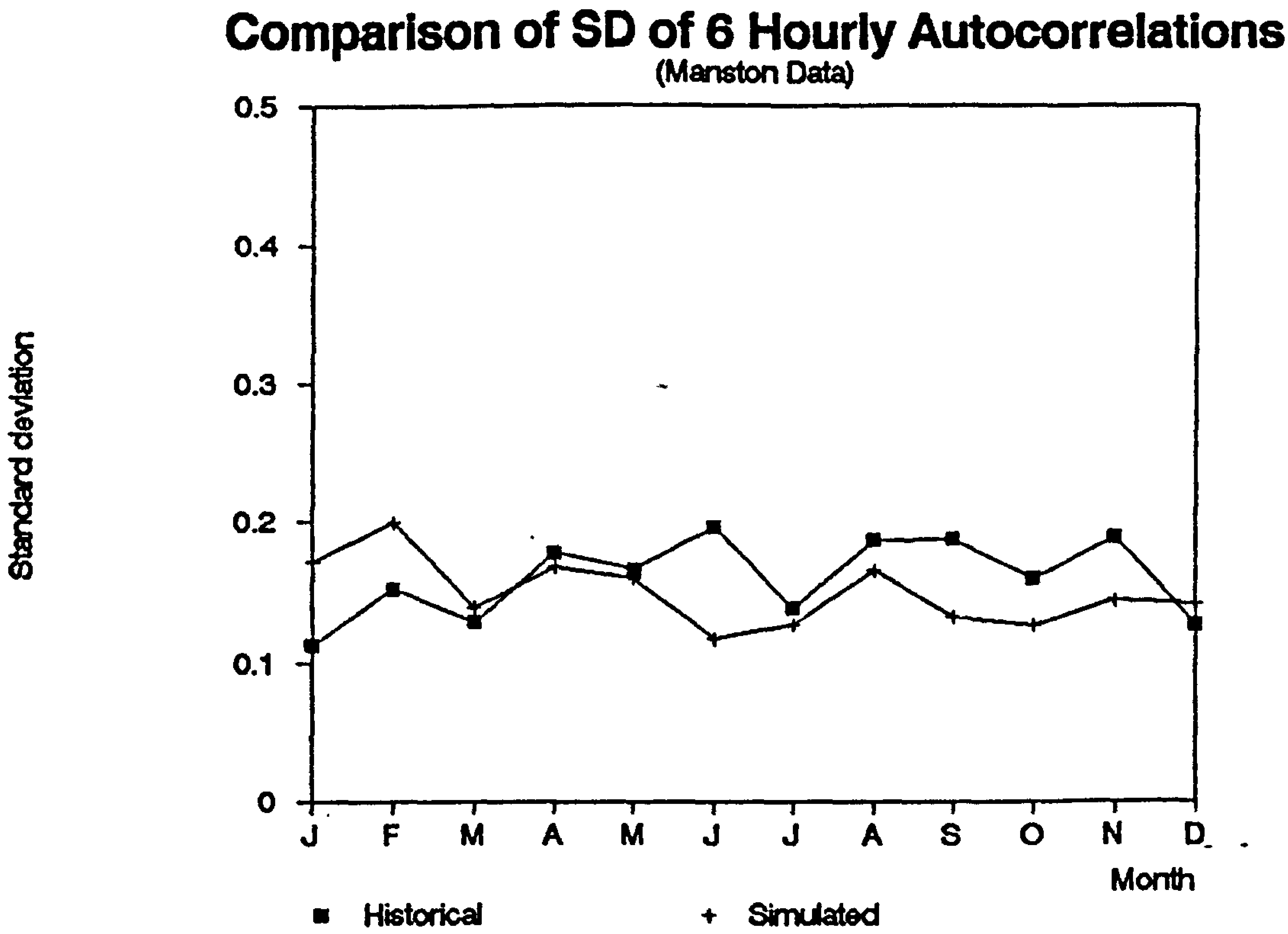


Figure E.53

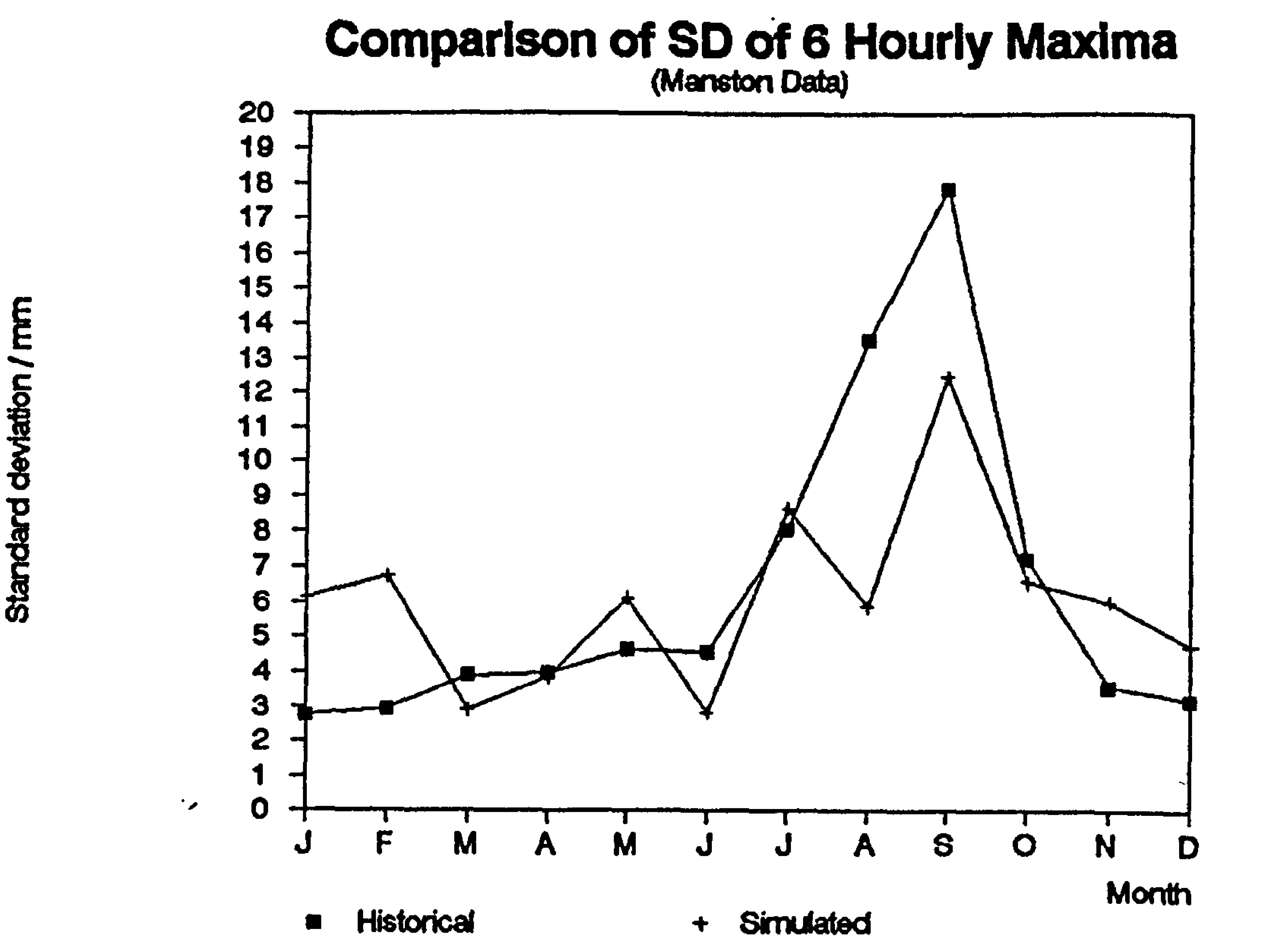


Figure E.54

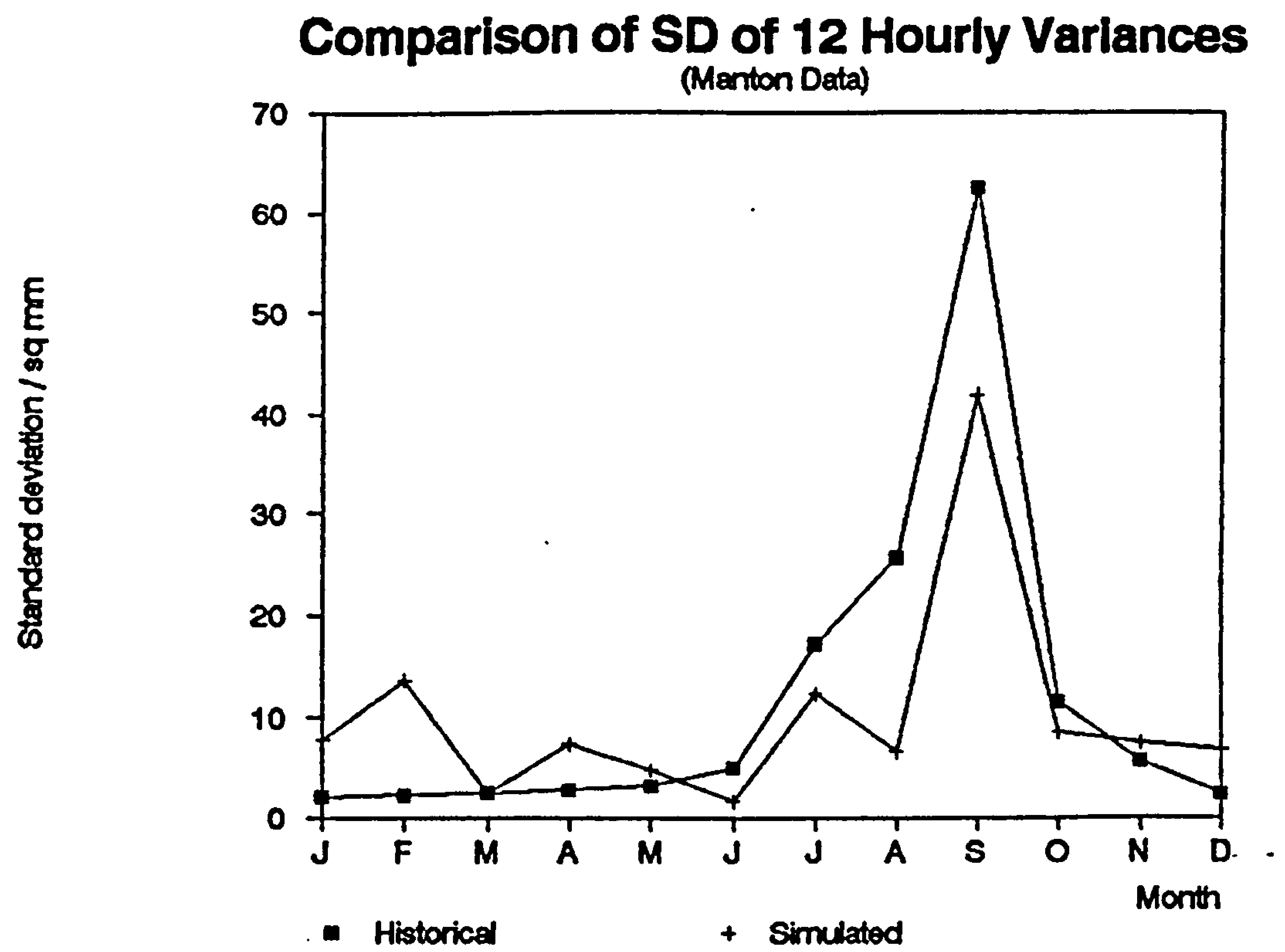


Figure E.55

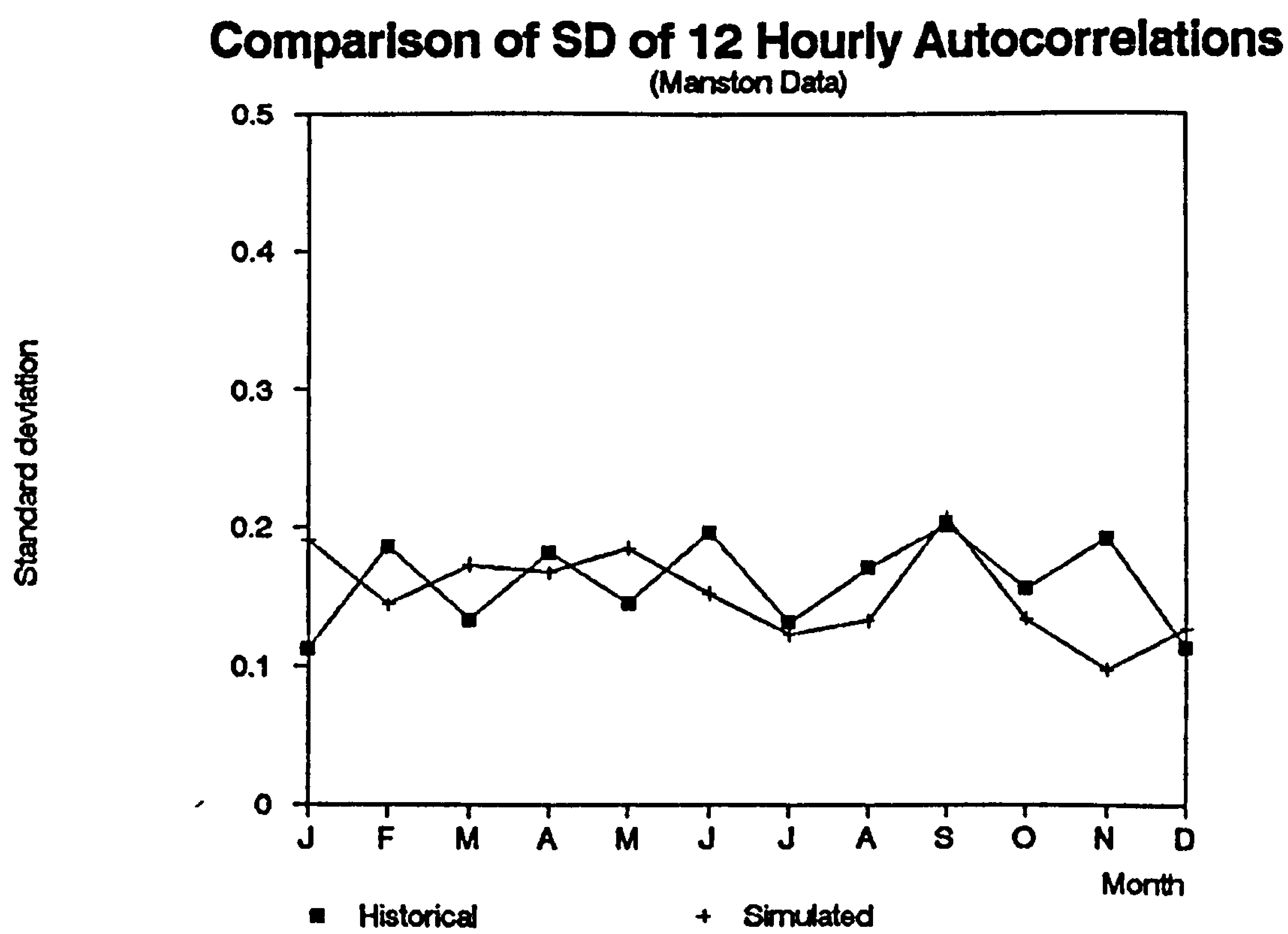


Figure E.56

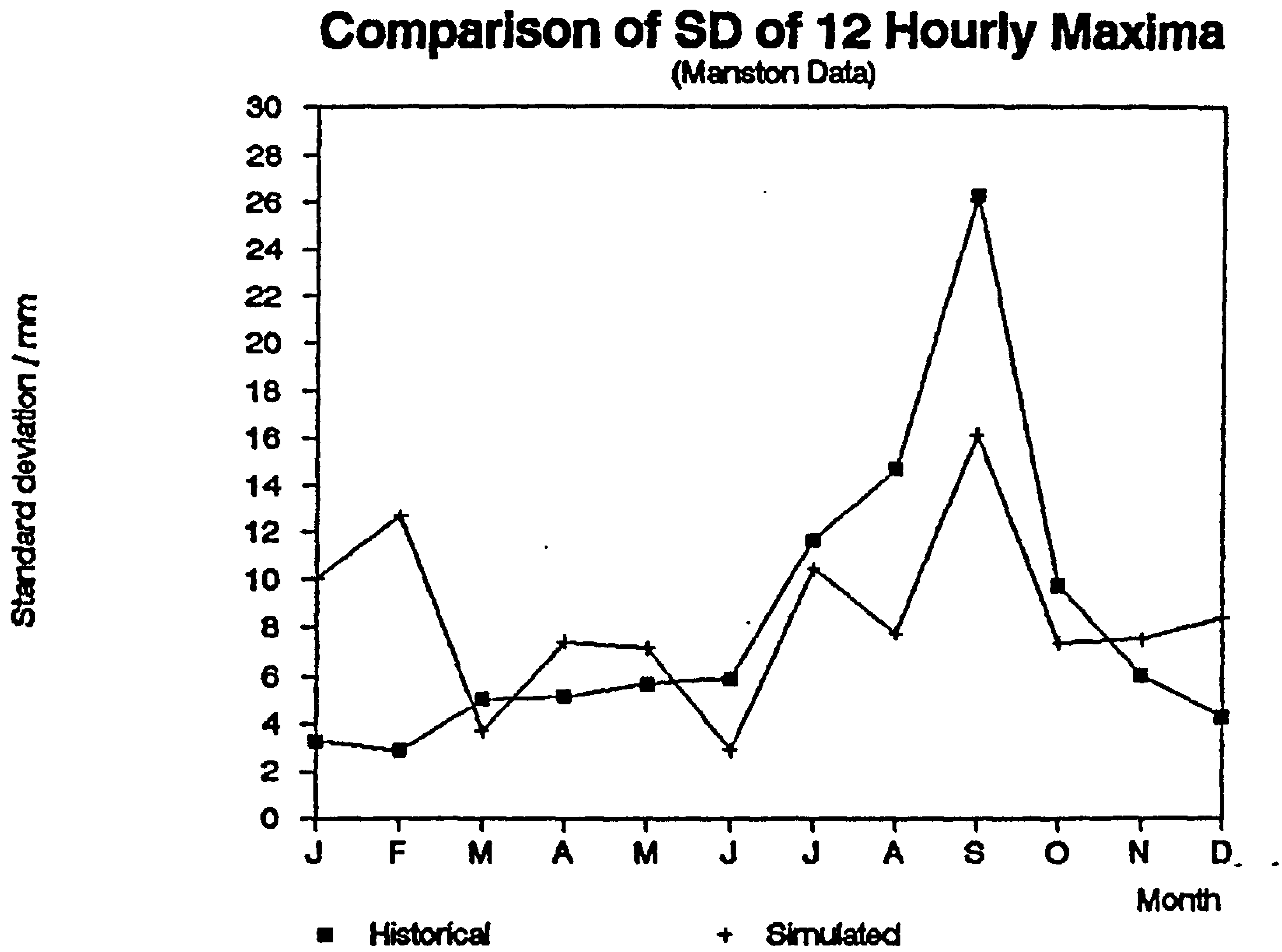


Figure E.57

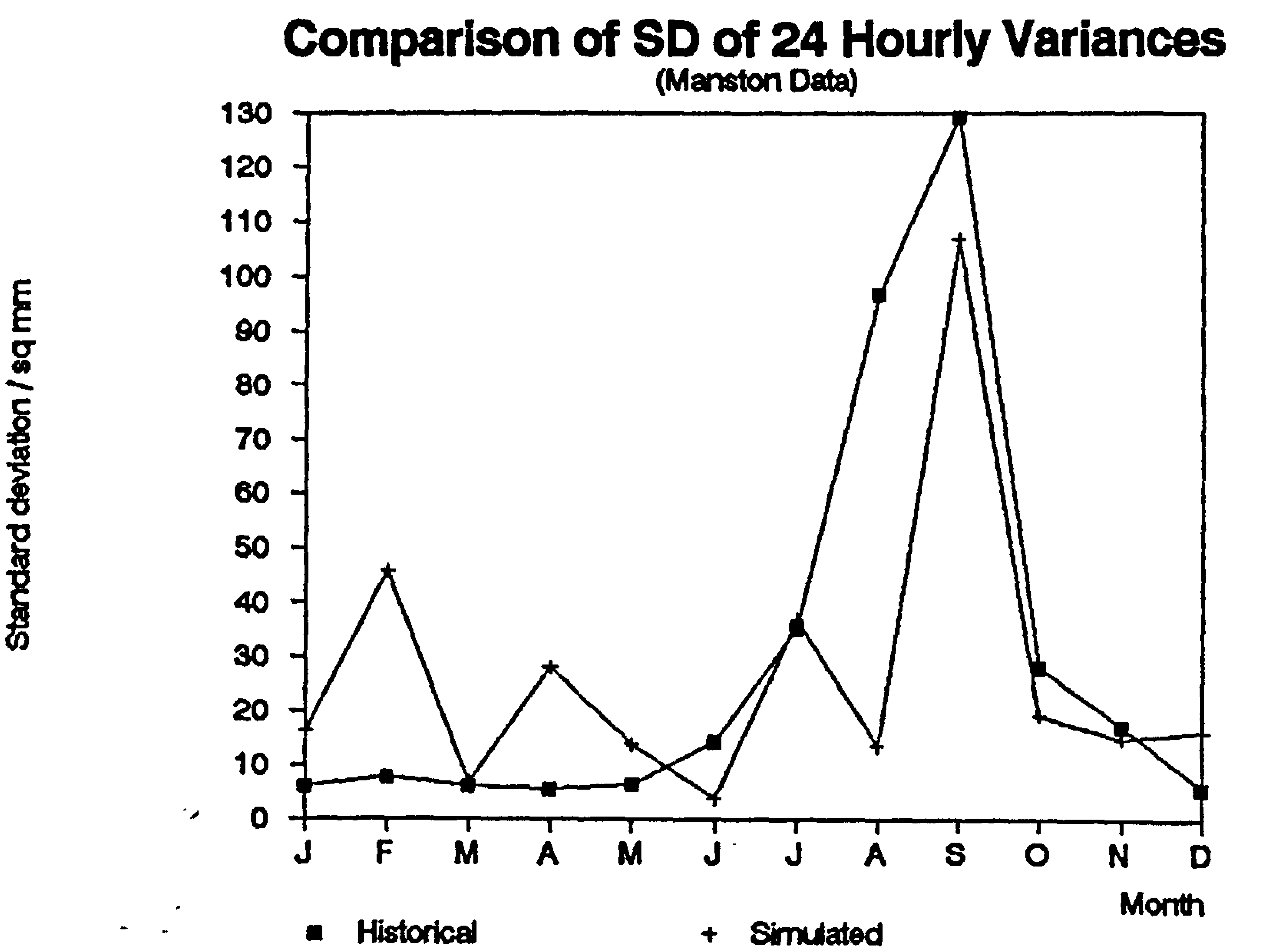


Figure E.58

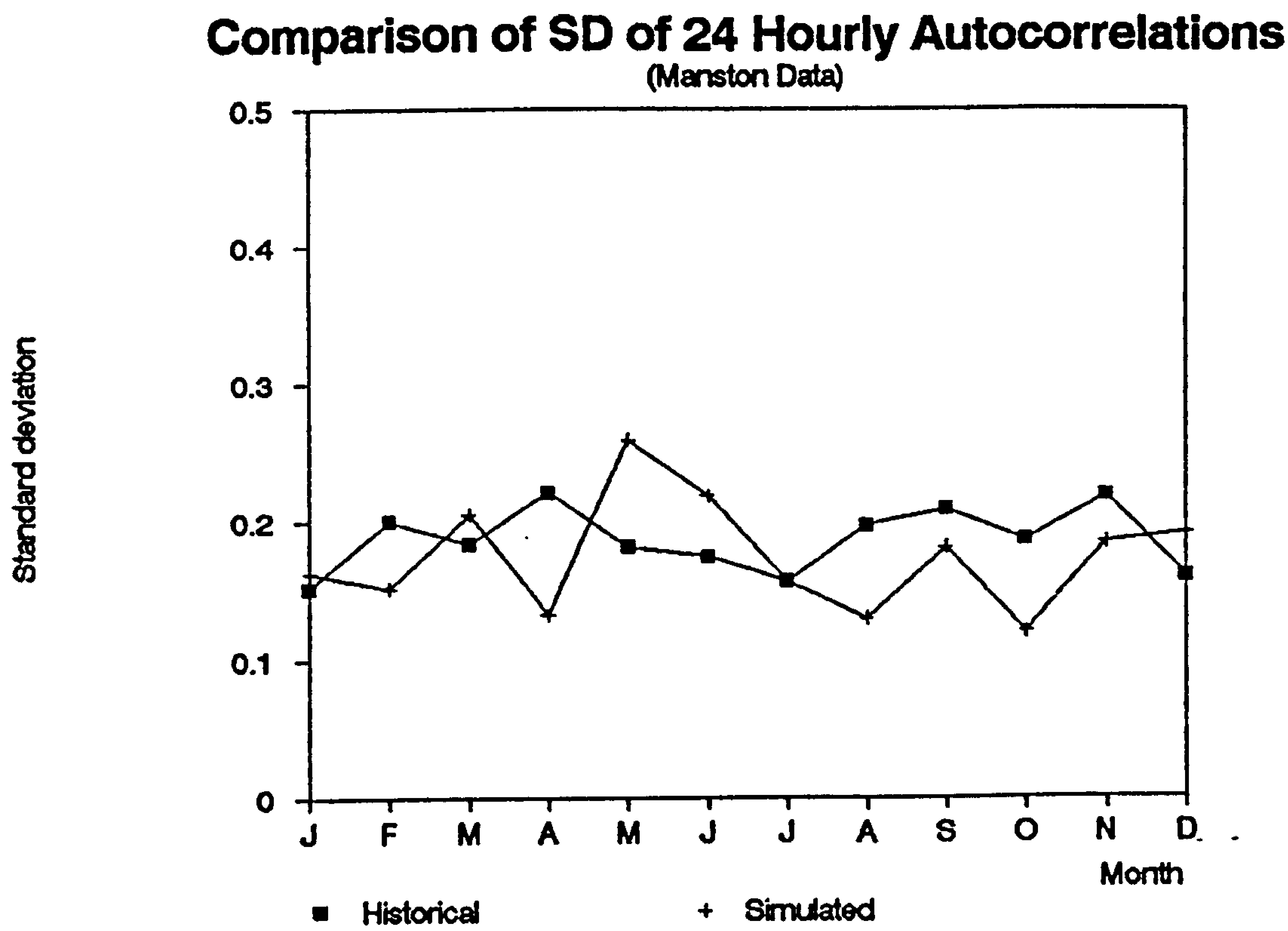


Figure E.59

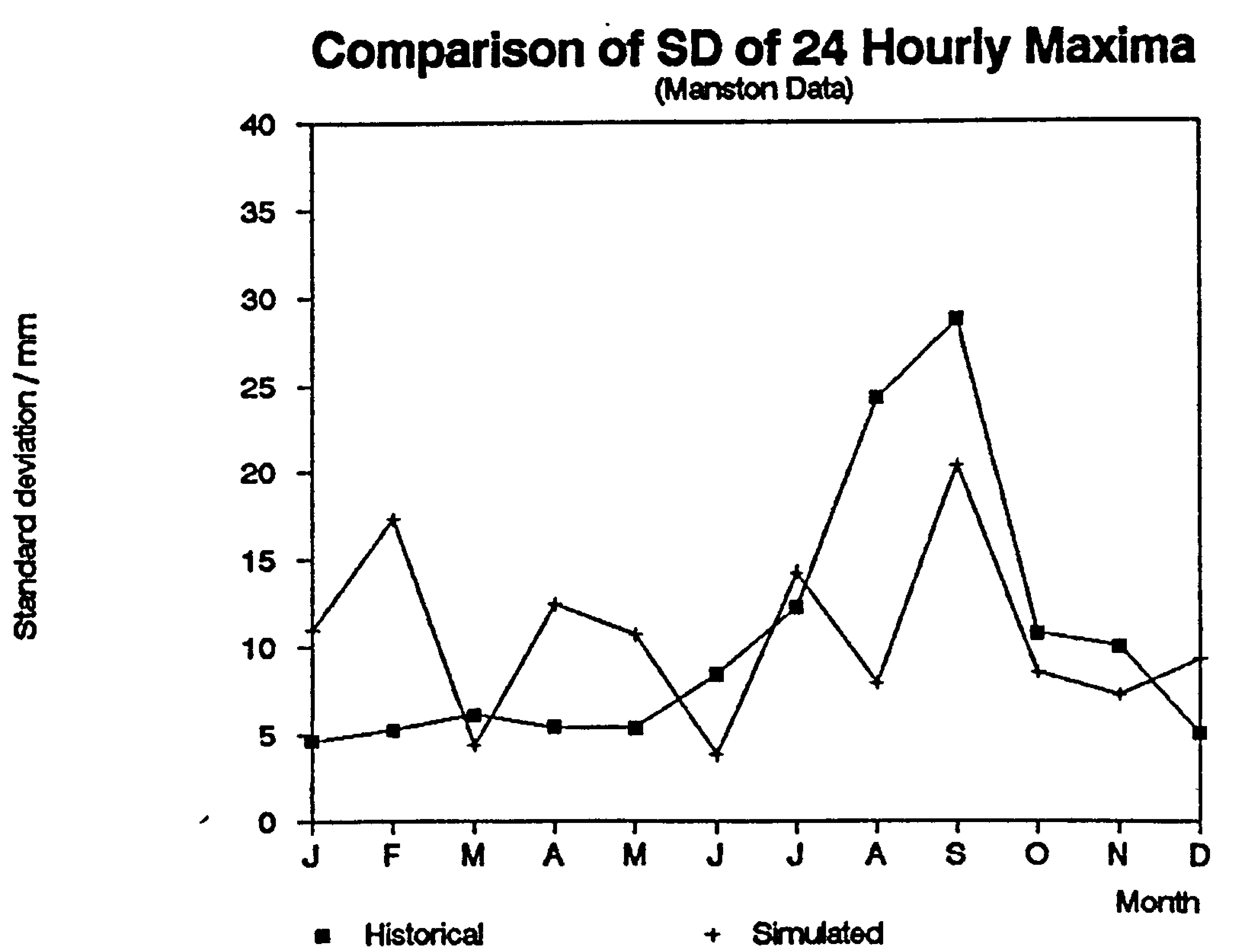


Figure E.60

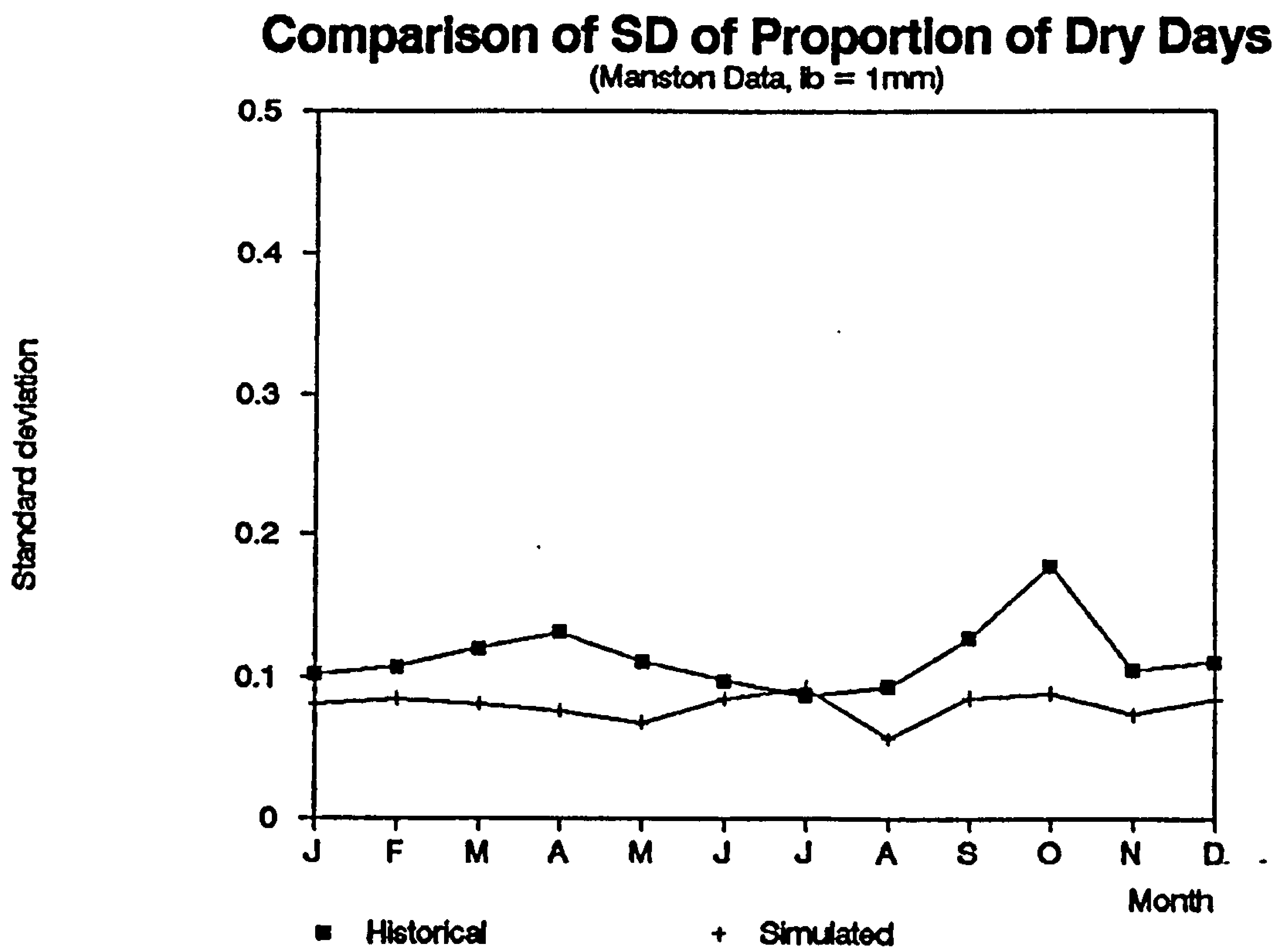


Figure E.61

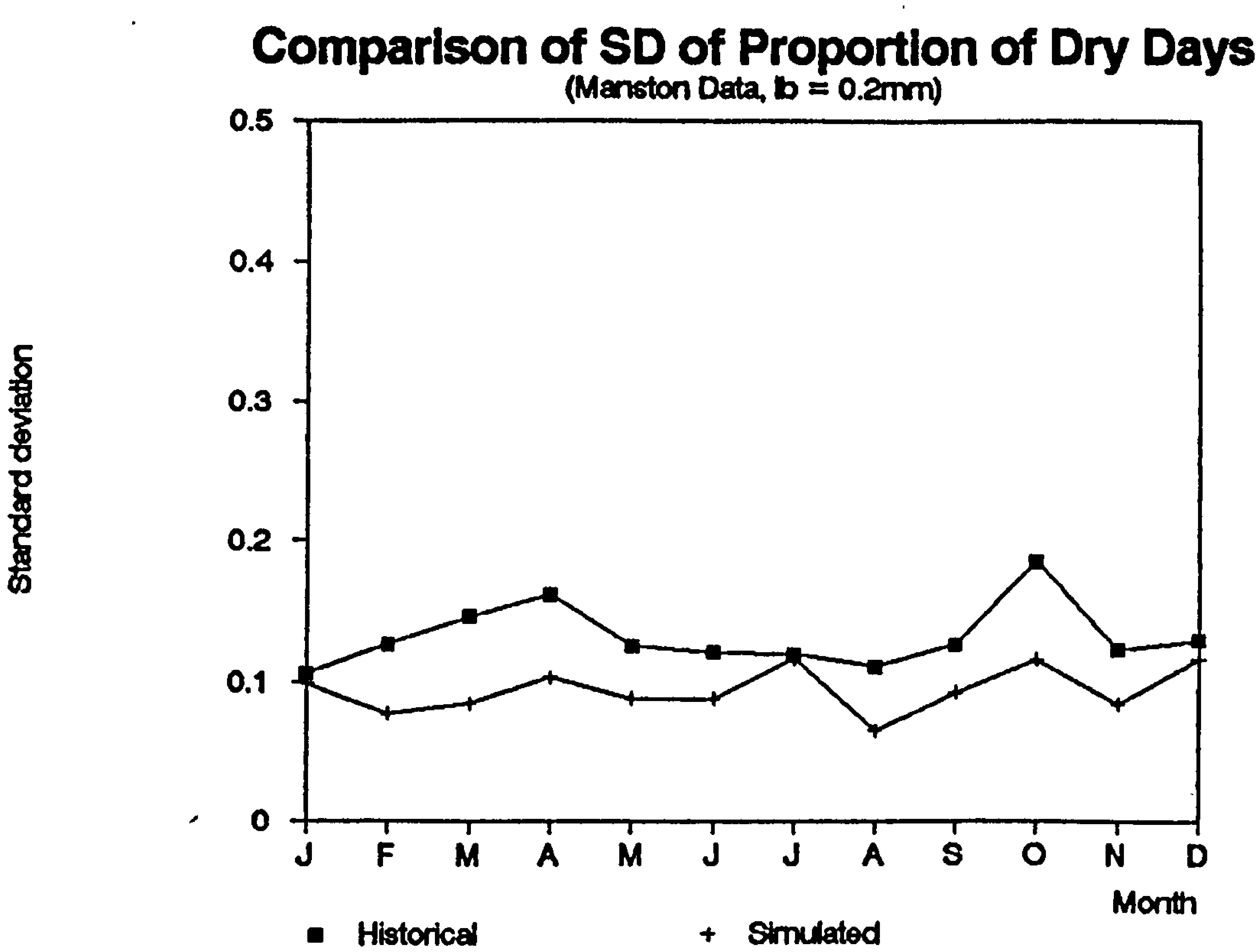


Figure E.62

APPENDIX F: A CLOSER LOOK AT THE 3 OBSERVATIONS EXCLUDED FROM THE MULTIVARIATE (REGIONAL) REGRESSION ANALYSIS

Table F.1 gives the range of parameter estimates for the observations used in the multivariate regression analysis of Chapter 6, and Table F.2 gives the parameter estimates of the 3 observations that were excluded from the analysis. Comparing these Tables it is clear that the excluded observations have parameter estimates well outside the range of the parameter estimates used in the analyses. Table F.2 also includes the predicted parameter estimates obtained using the multivariate regression model, and as might be expected some of the predicted parameter estimates are substantially different from the actual estimates. To see whether these differences constitute an overall difference in rainfall, some key expressions were evaluated and are shown in Table F.3 [Note that this Table also contains the value of the statistic extracted from the rainfall data].

From the plots of Appendix G (Figures G.26 - G.41) it is clear that the percentage errors obtained for observations 2 and 3 are fairly typical of other percentage errors for observations included in the regression analysis. However, observation 1 does have large percentage errors in predicted values, so that one might expect this observation to come from a short rainfall record. However, this was not the case as this observation came from 25 years of daily rainfall data taken from Inverness, suggesting that the regionalised model would be inadequate for this station-month. This inadequacy could be due to the station lying in a region that is influenced by micro-climates. Under such circumstances it would be better to fit the stochastic model to daily data taken from the site under investigation.

Table F.1

Range of Values for Parameter Estimates

(Excluding 'outlying' observations)

Parameter:	λ	β	η	ν	ξ
	hr^{-1}	hr^{-1}	hr^{-1}	cells/ storm	hr/mm
Maximum	0.0320	0.5511	16.03	19.79	3.213
Minimum	0.0057	0.0500	0.2794	1.415	0.1008

Table F.2

Parameter Estimates for Excluded Observations

Observation	λ	β	η	ν	ξ	Source
1	0.0125	0.249	506.	0.00693	11.9	Data†
1	0.0160	0.141	0.802	1.44	5.12	Predicted*
2	0.0175	395.	0.458	0.442	1.00	Data†
2	0.0106	0.118	1.00	0.585	4.67	Predicted*
3	0.0152	410.	0.489	0.463	1.00	Data†
3	0.0105	0.113	0.987	0.578	3.87	Predicted*

* = predicted using the multivariate (regional) regression model.

† = parameter estimates obtained by fitting the model to the data.

Table F.3

Evaluated Key Expressions

Observation	$\mu(24)$	$\gamma(24)$	$\phi(24)$	Source
1	1.02	4.03	0.641	Data†
1	1.71 (68%)	10.8 (168%)	0.576 (10%)	Predicted*
1	1.02	4.32	0.651	Historical
2	2.07	37.2	0.658	Data†
2	2.03 (2%)	23.5 (37%)	0.684 (4%)	Predicted*
2	2.07	52.7	0.649	Historical
3	1.62	26.2	0.694	Data†
3	1.71 (6%)	18.5 (29%)	0.702 (1%)	Predicted*
3	1.61	37.6	0.681	Historical

(Absolute percentage errors are shown in brackets, $\mu(24)$ = mean of 24 hourly time series, $\gamma(24)$ = variance of 24 hourly time series, $\phi(24)$ = proportion of dry days (μ, γ, and ϕ are functions of the model parameters)).

* = predicted using the multivariate (regional) regression model.

† = parameter estimates obtained by fitting the model to the data.

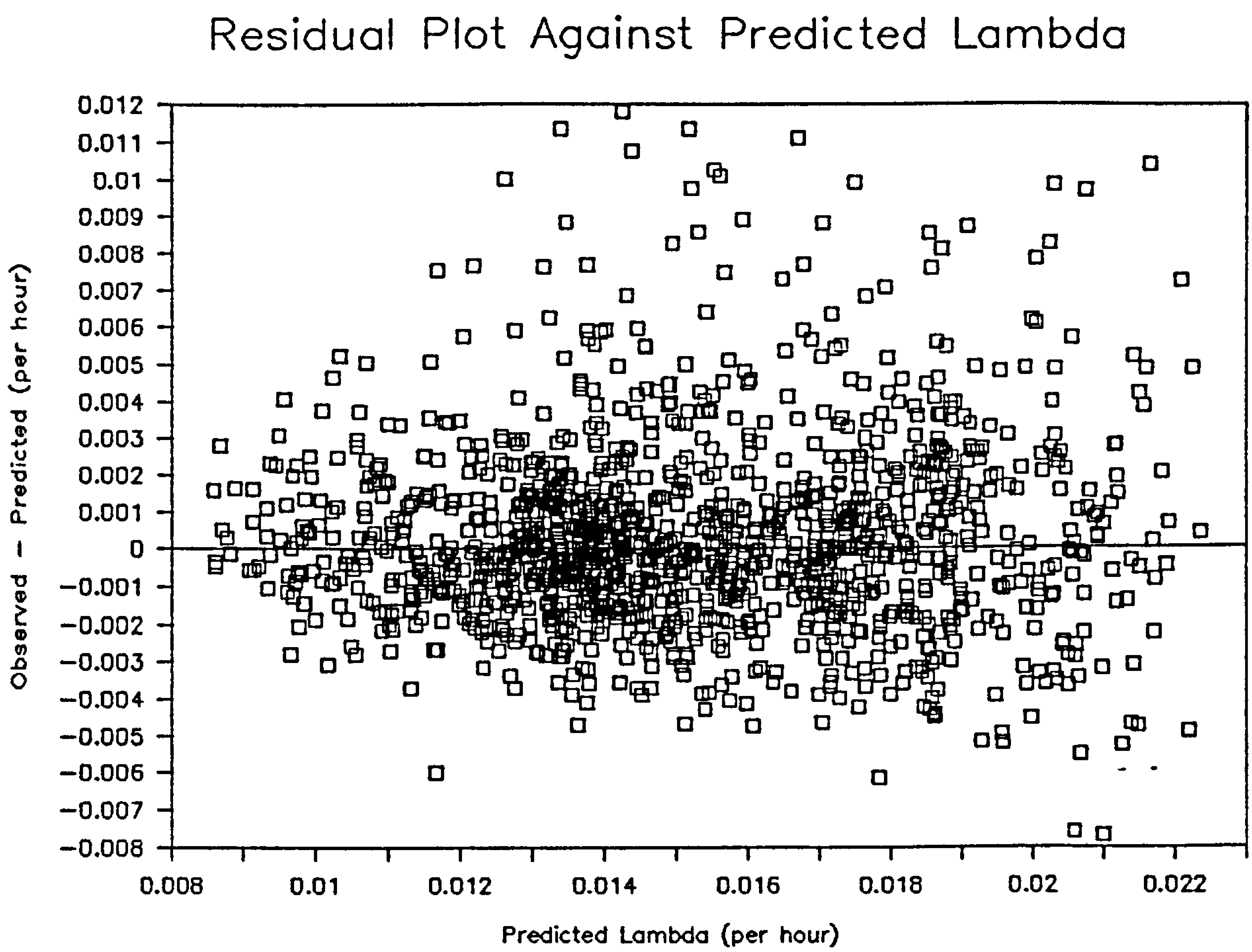


Figure G.1

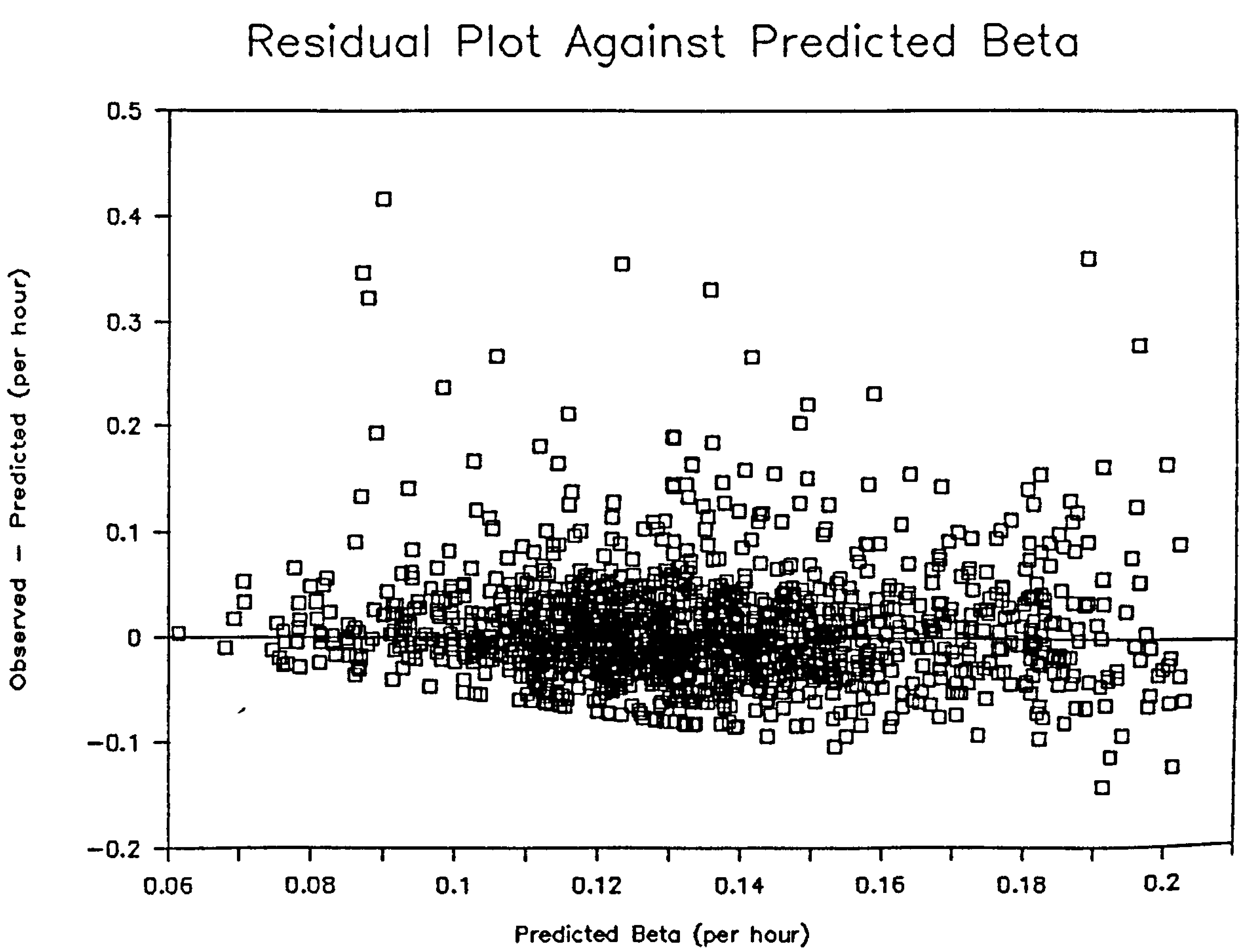


Figure G.2

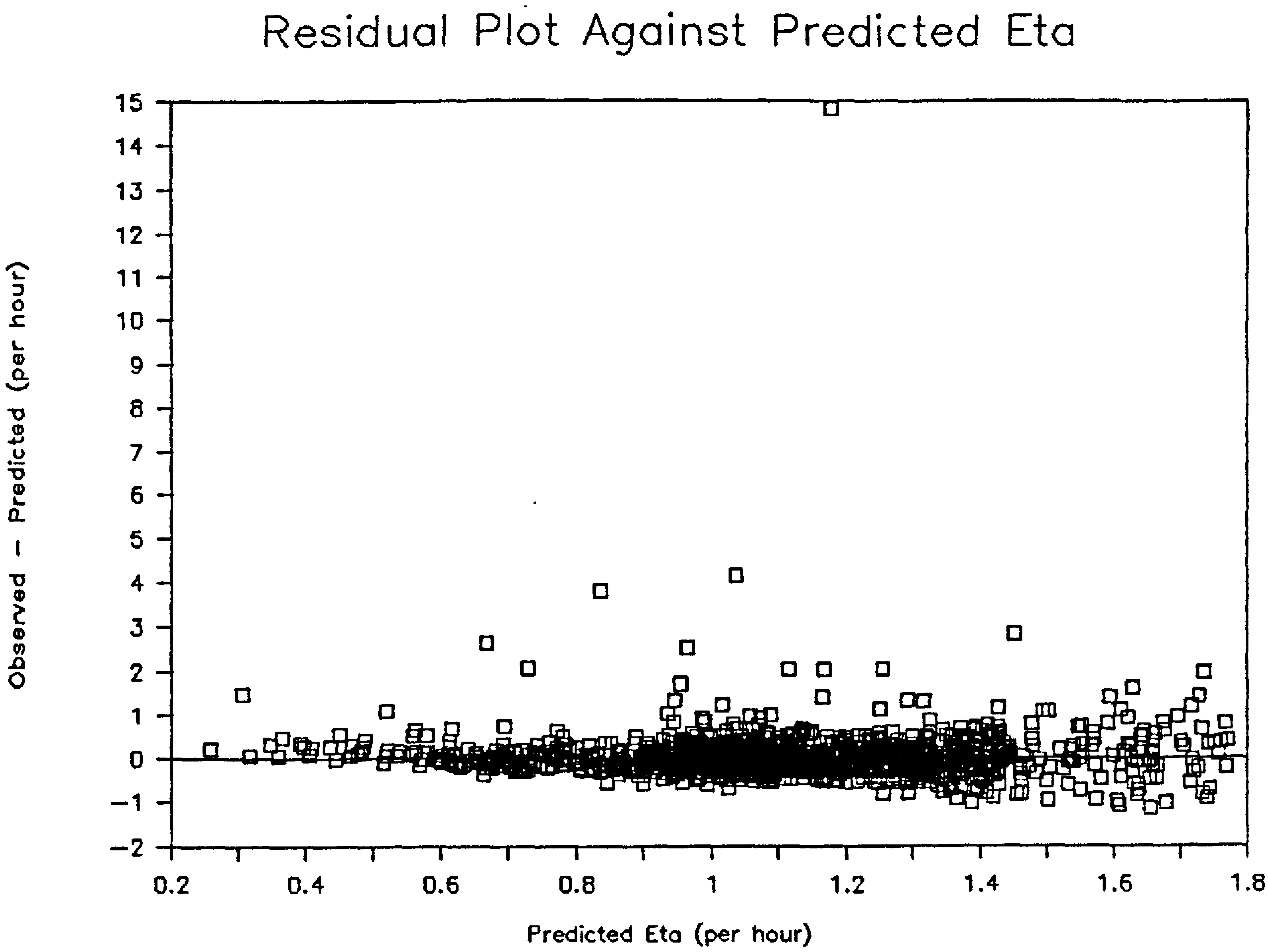


Figure G.3(a)

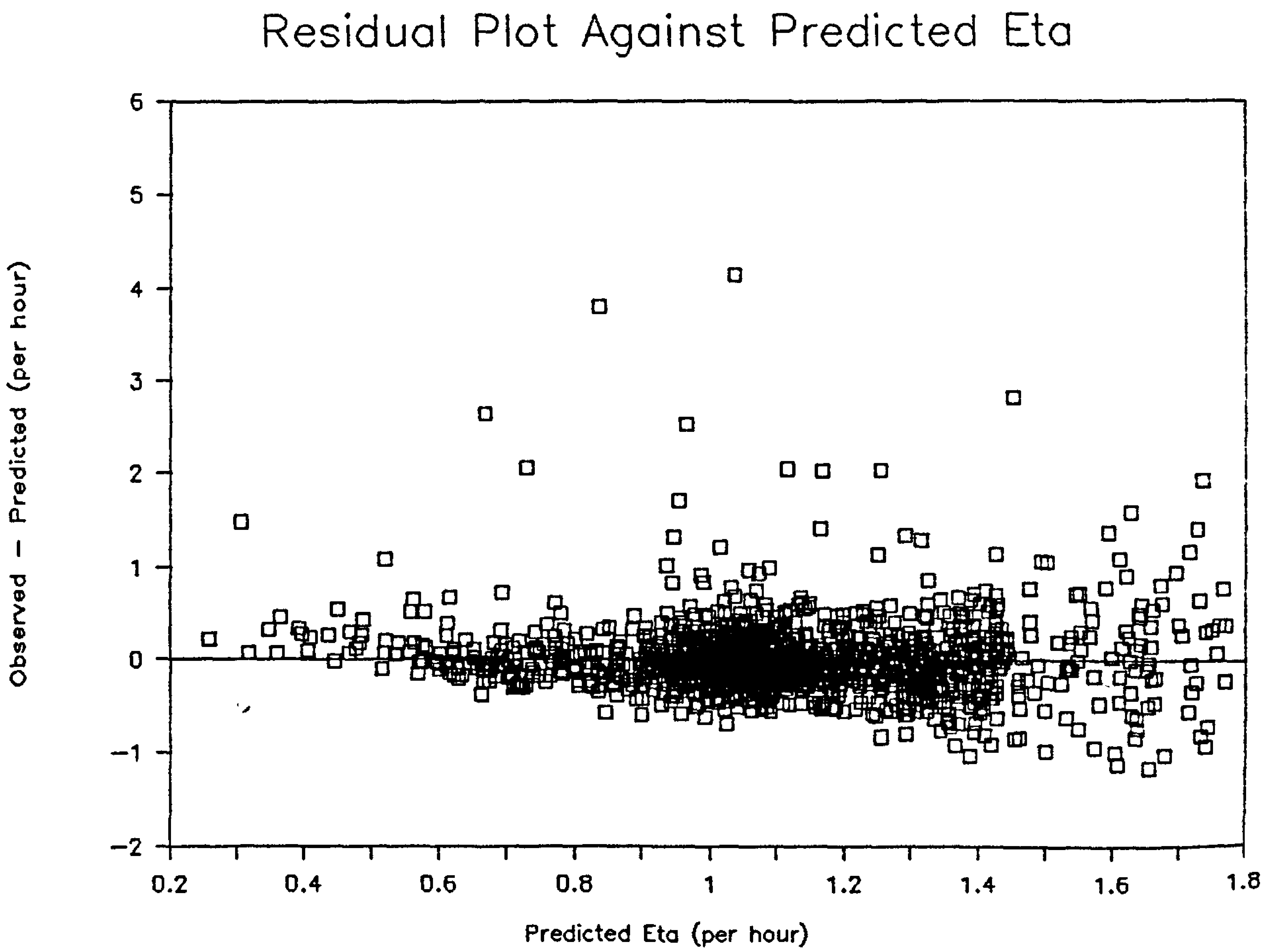


Figure G.3(b)

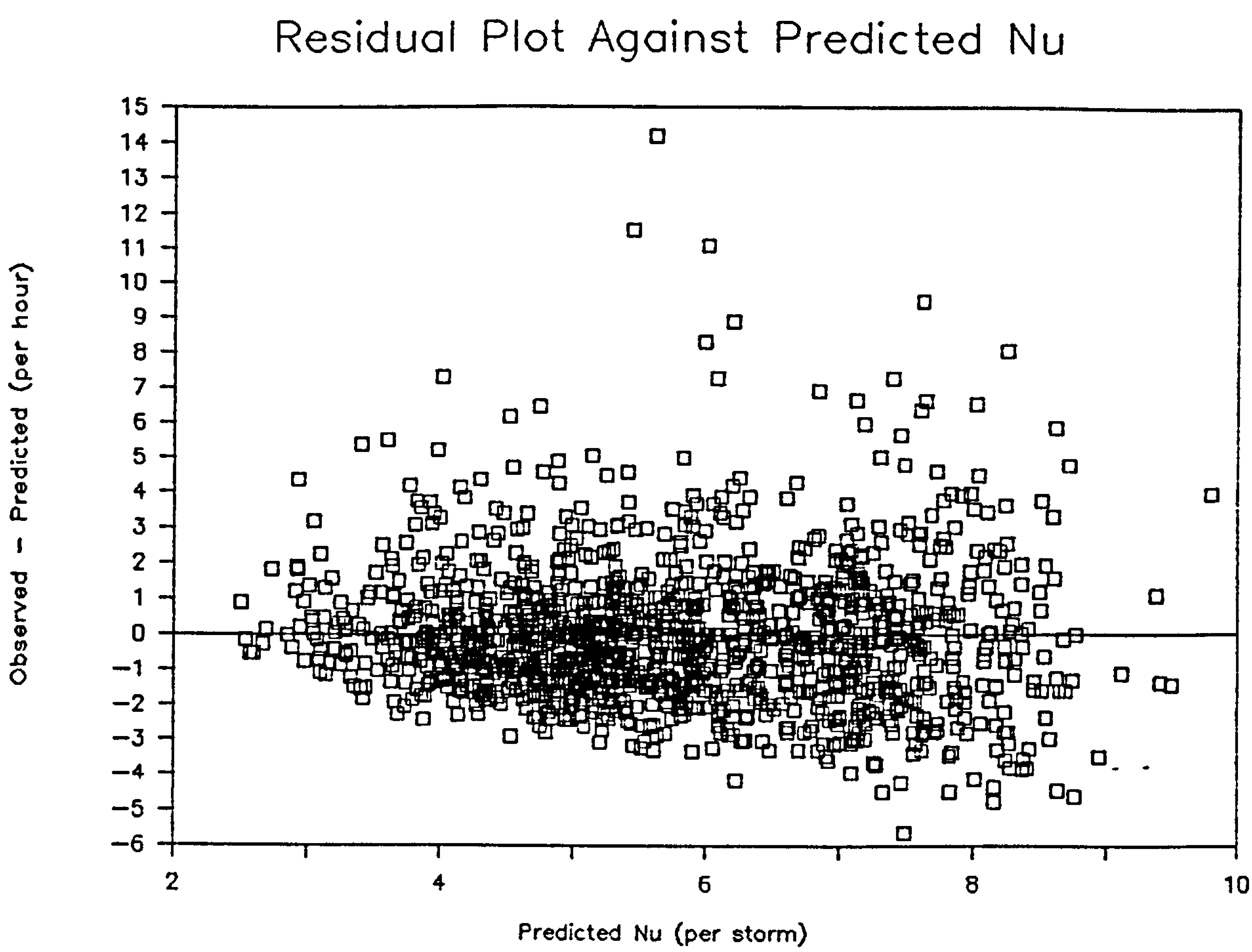


Figure G.4

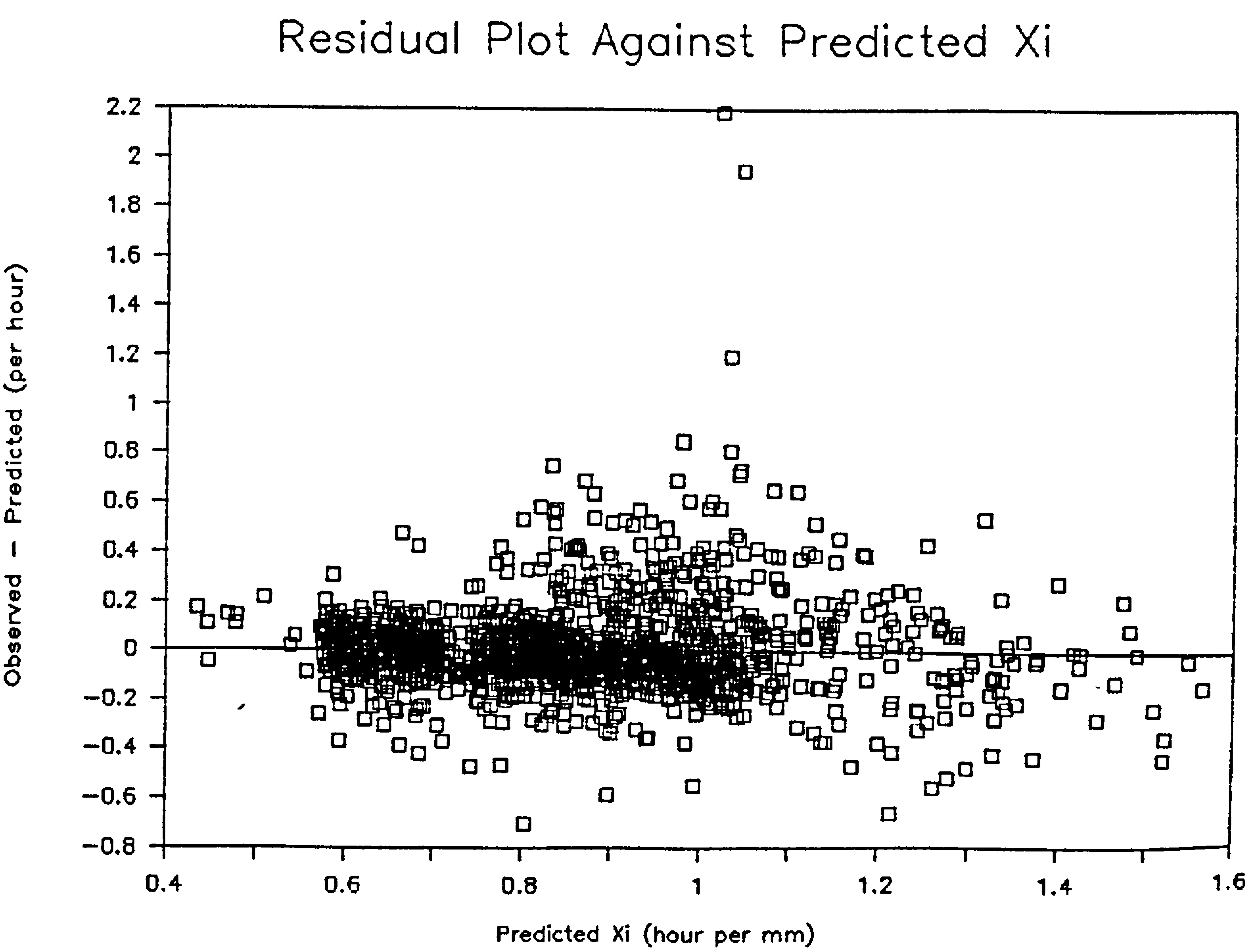


Figure G.5

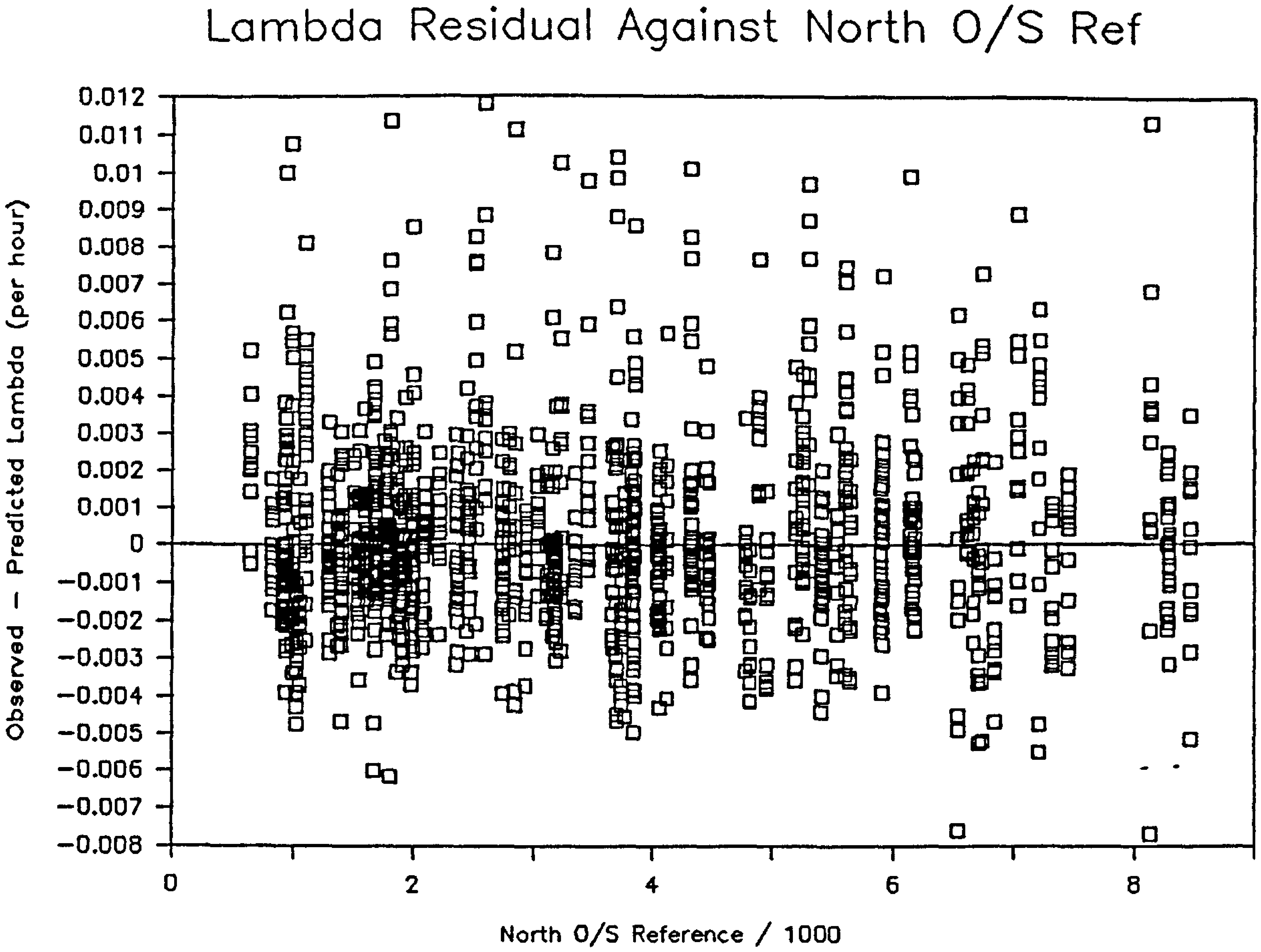


Figure G.6

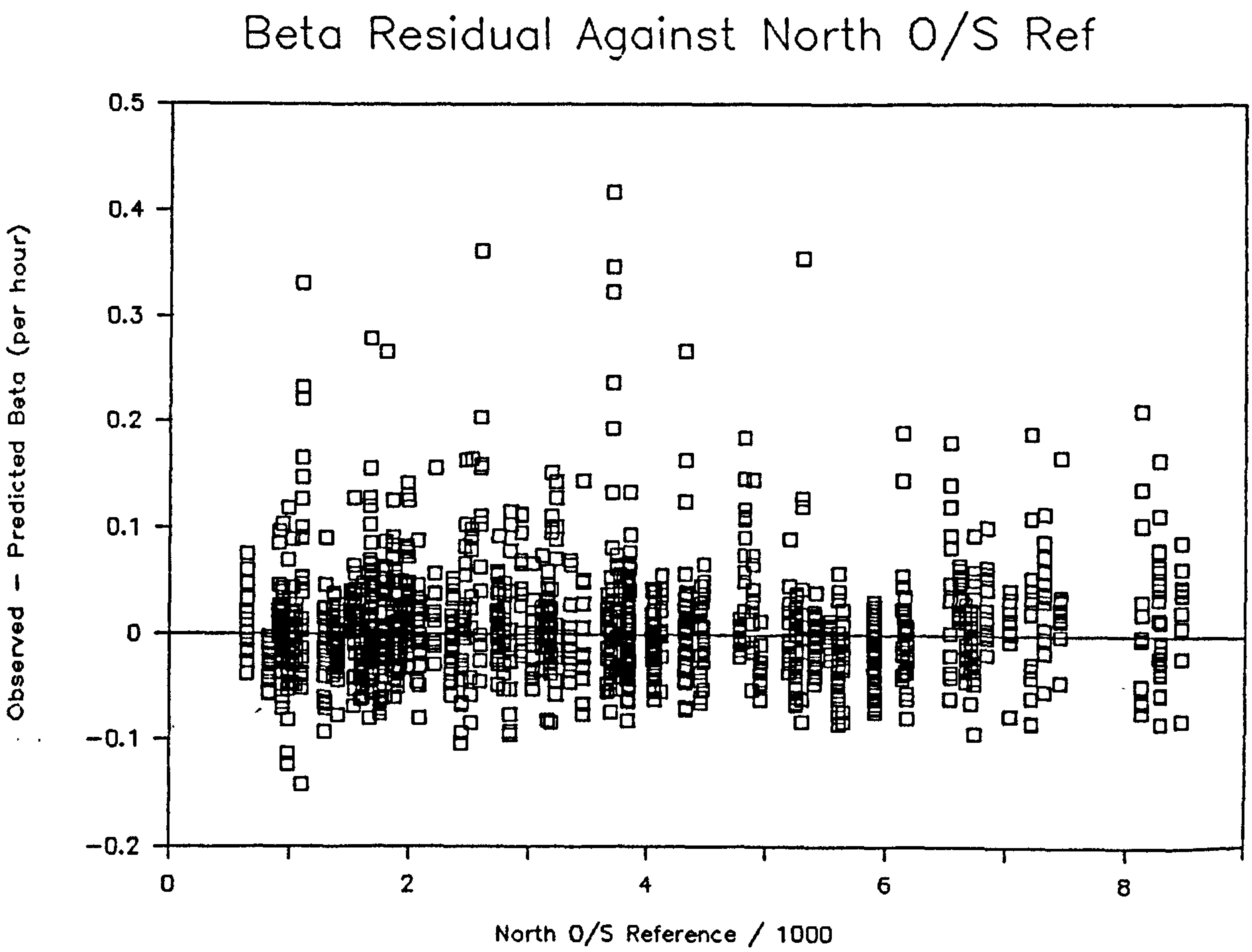


Figure G.7

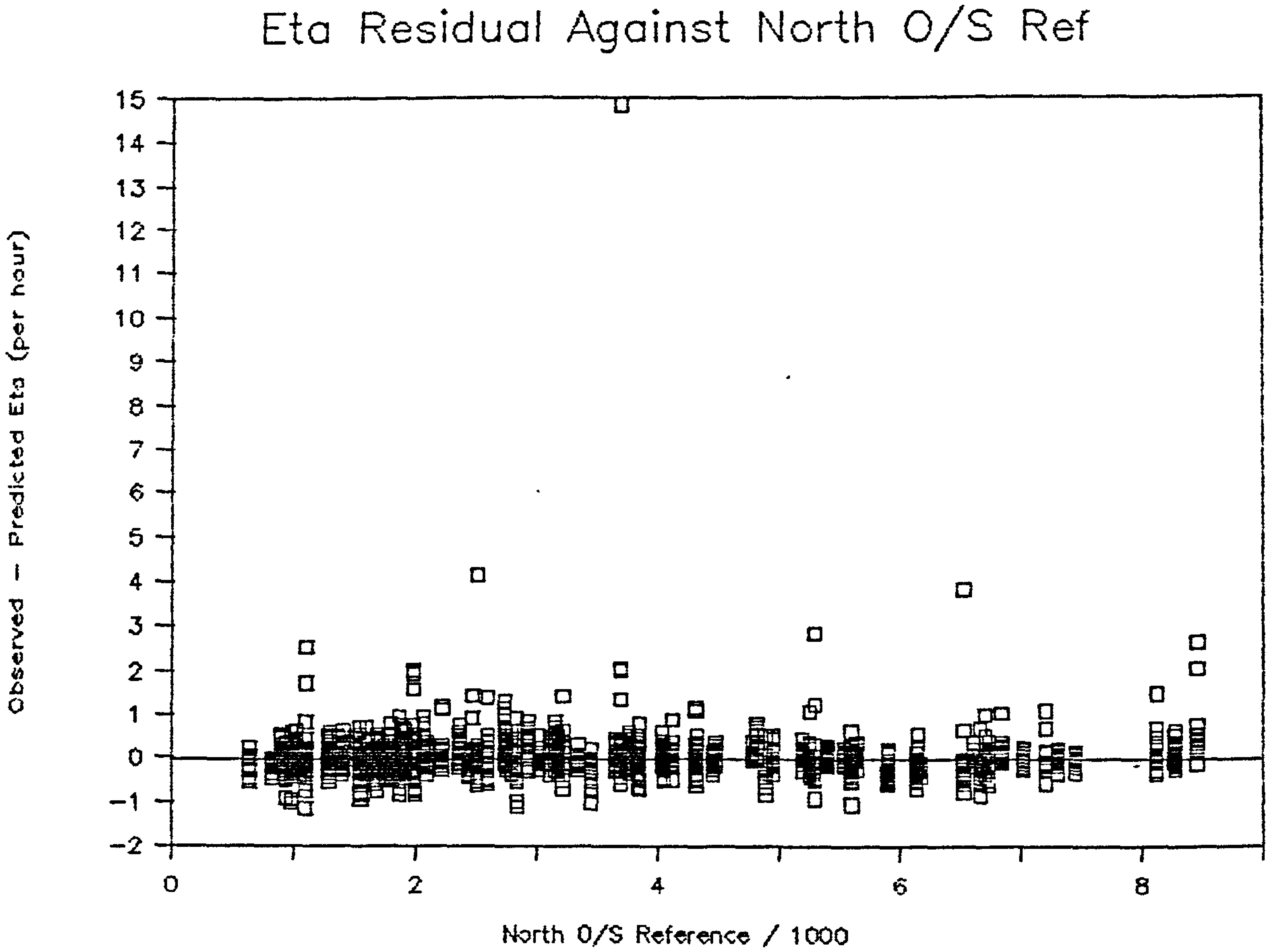


Figure G.8

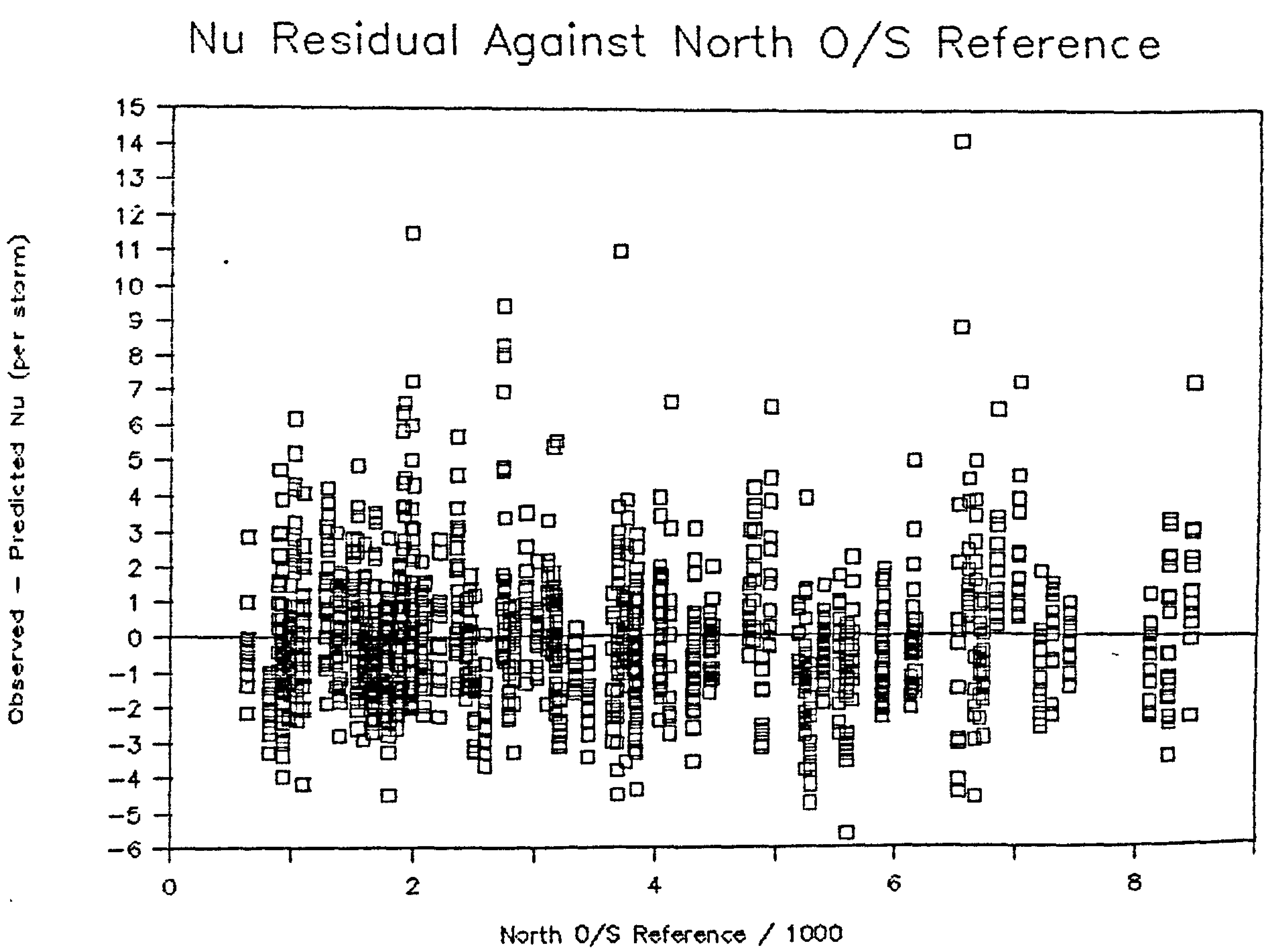


Figure G.9

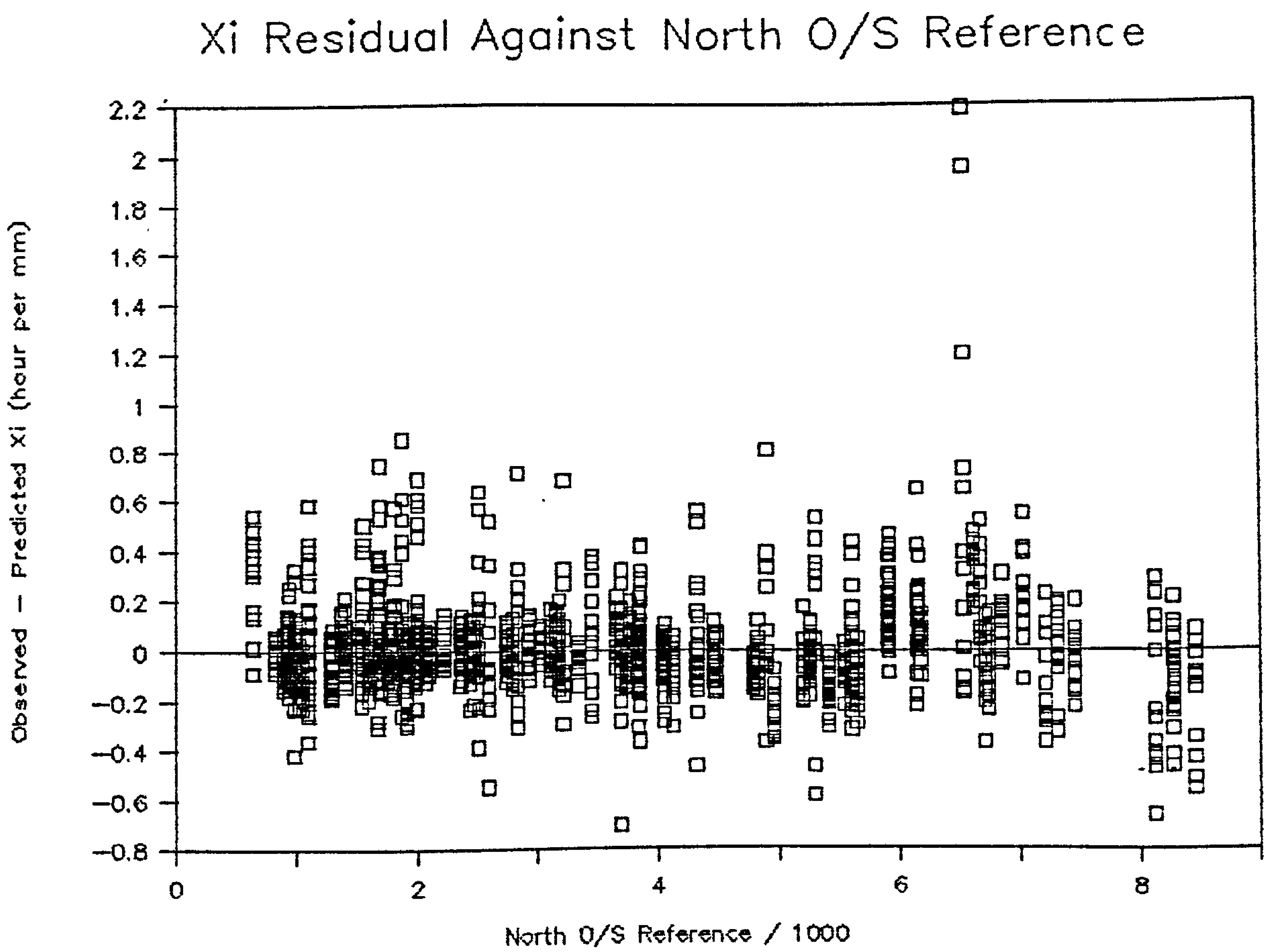


Figure G.10

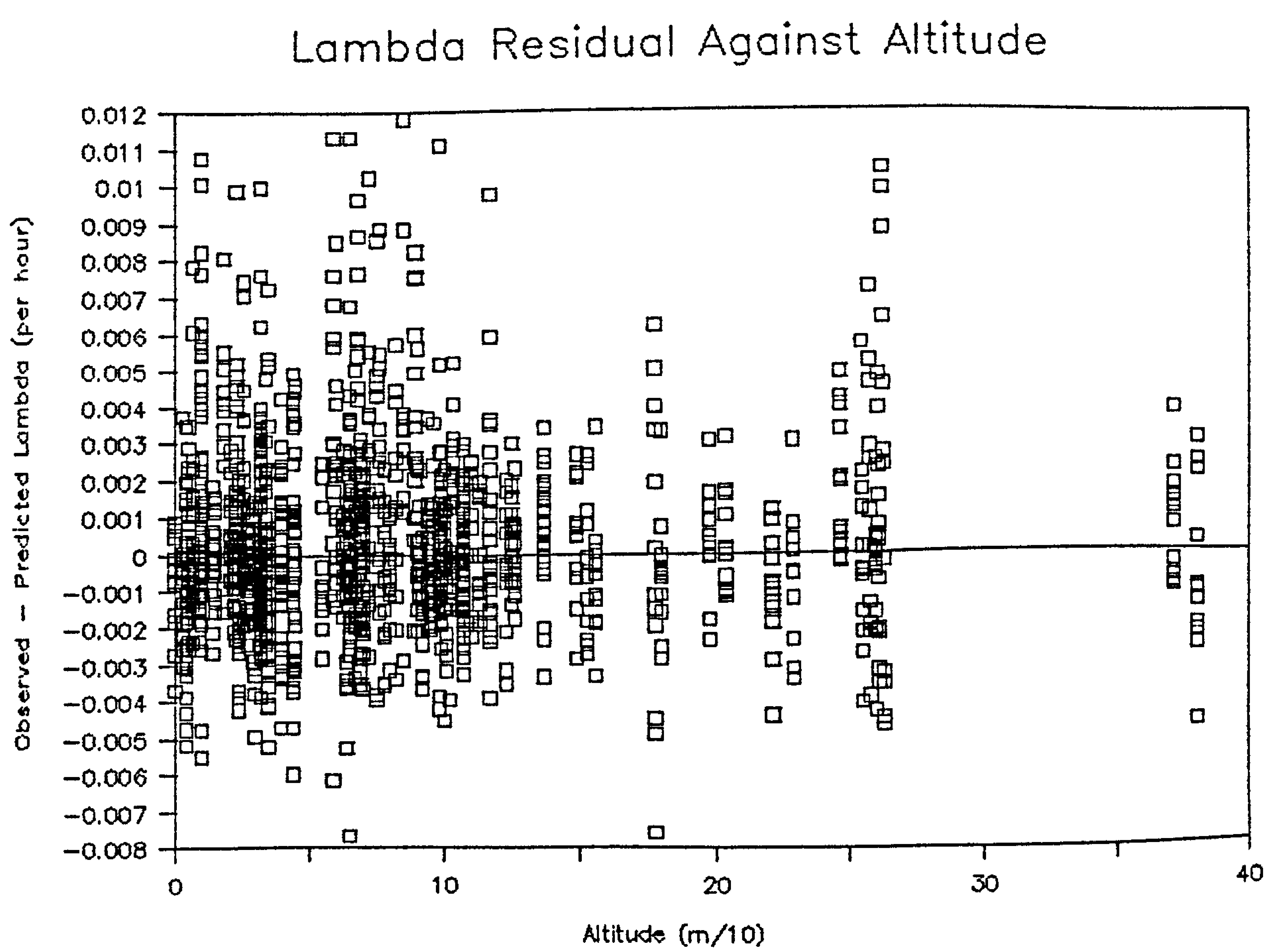


Figure G.11

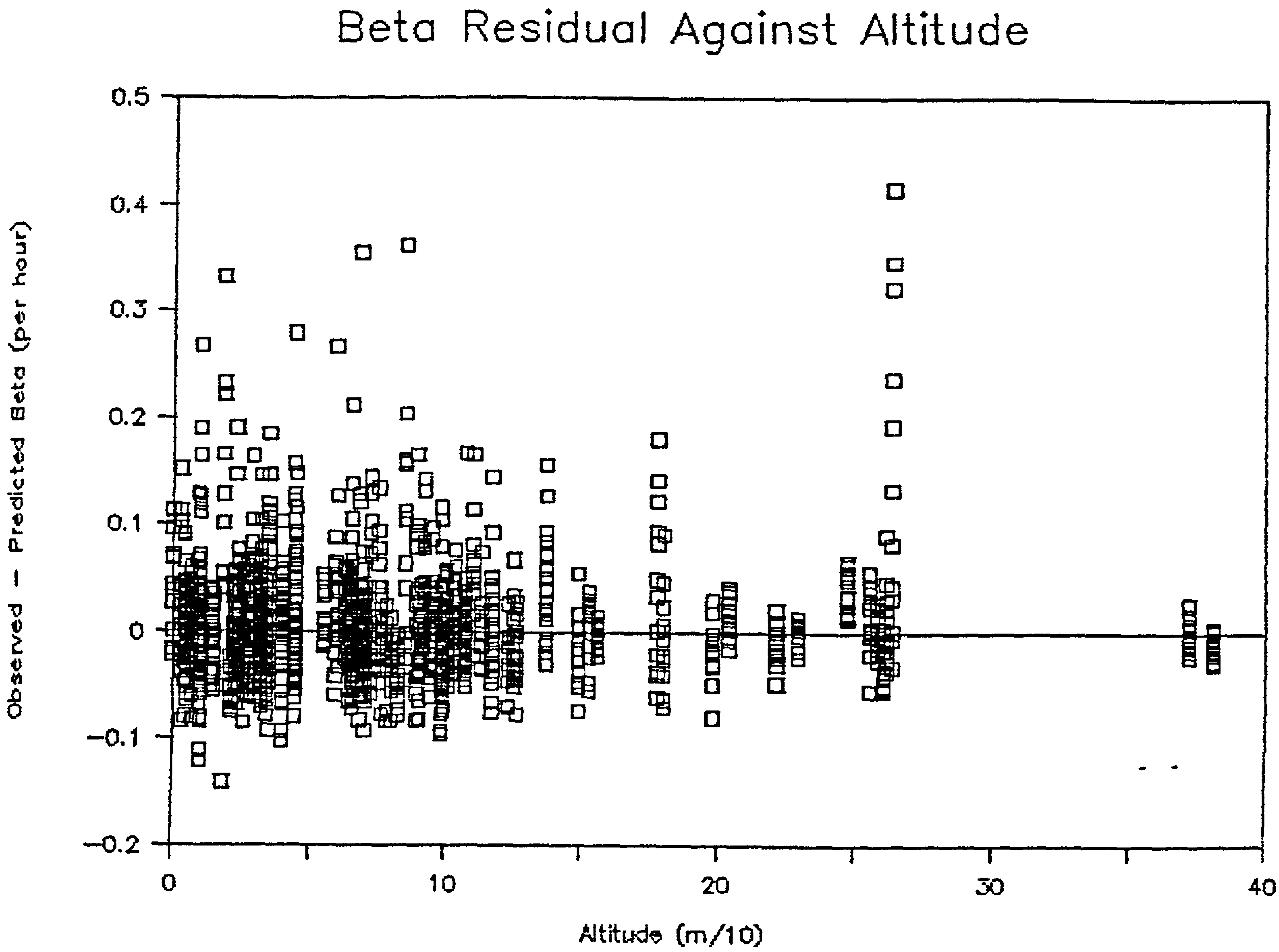


Figure G.12

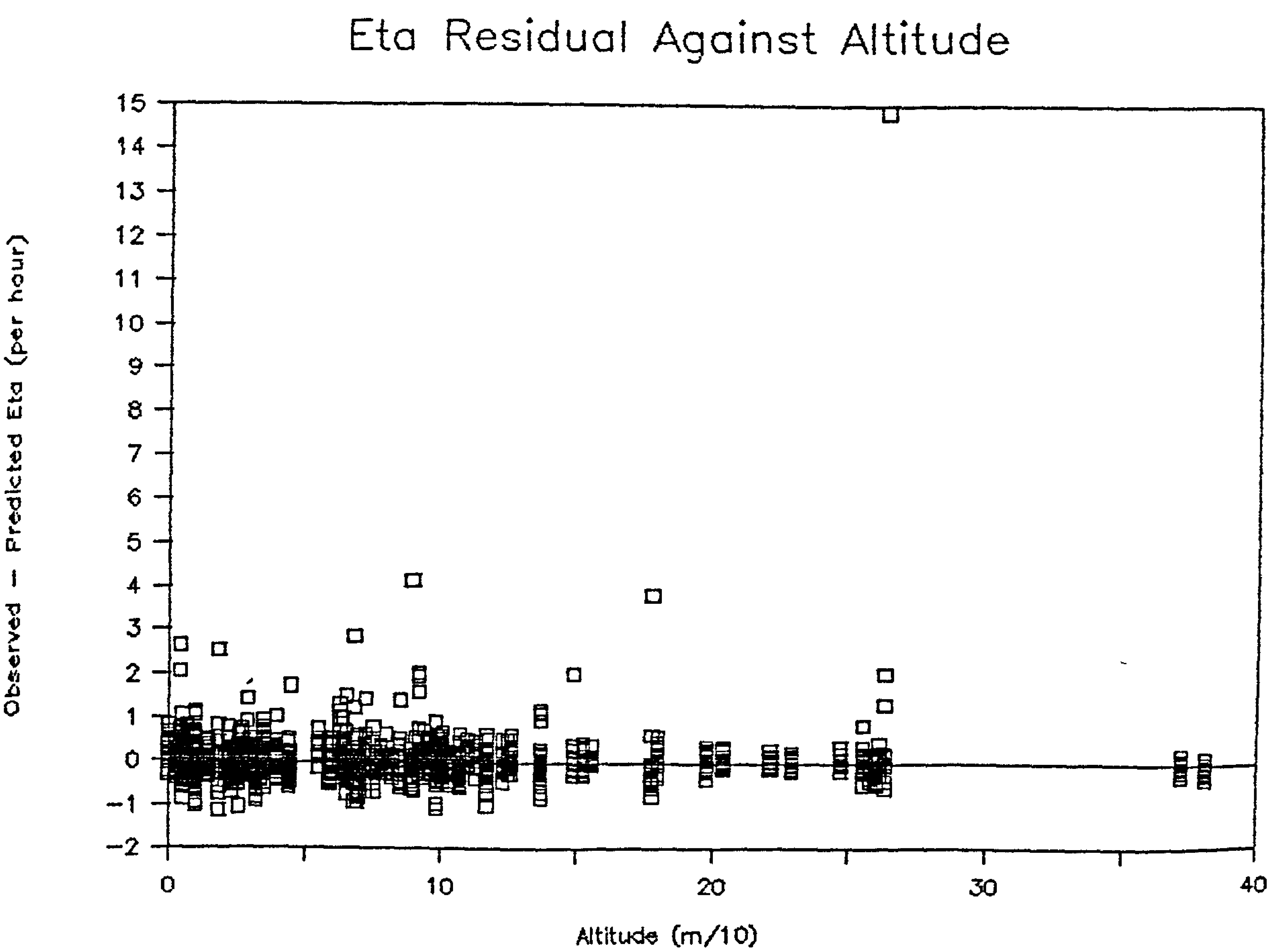


Figure G.13

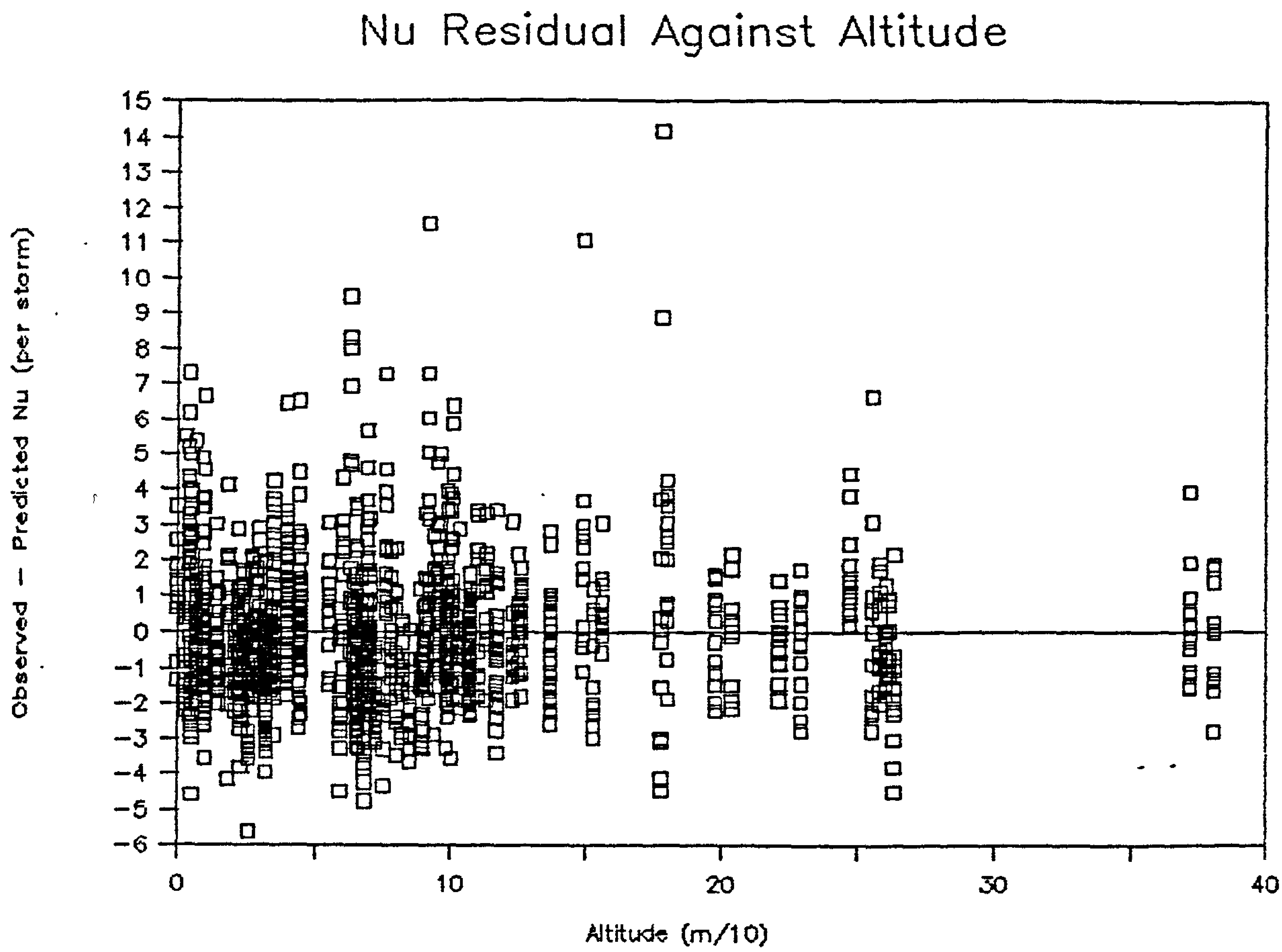


Figure G.14

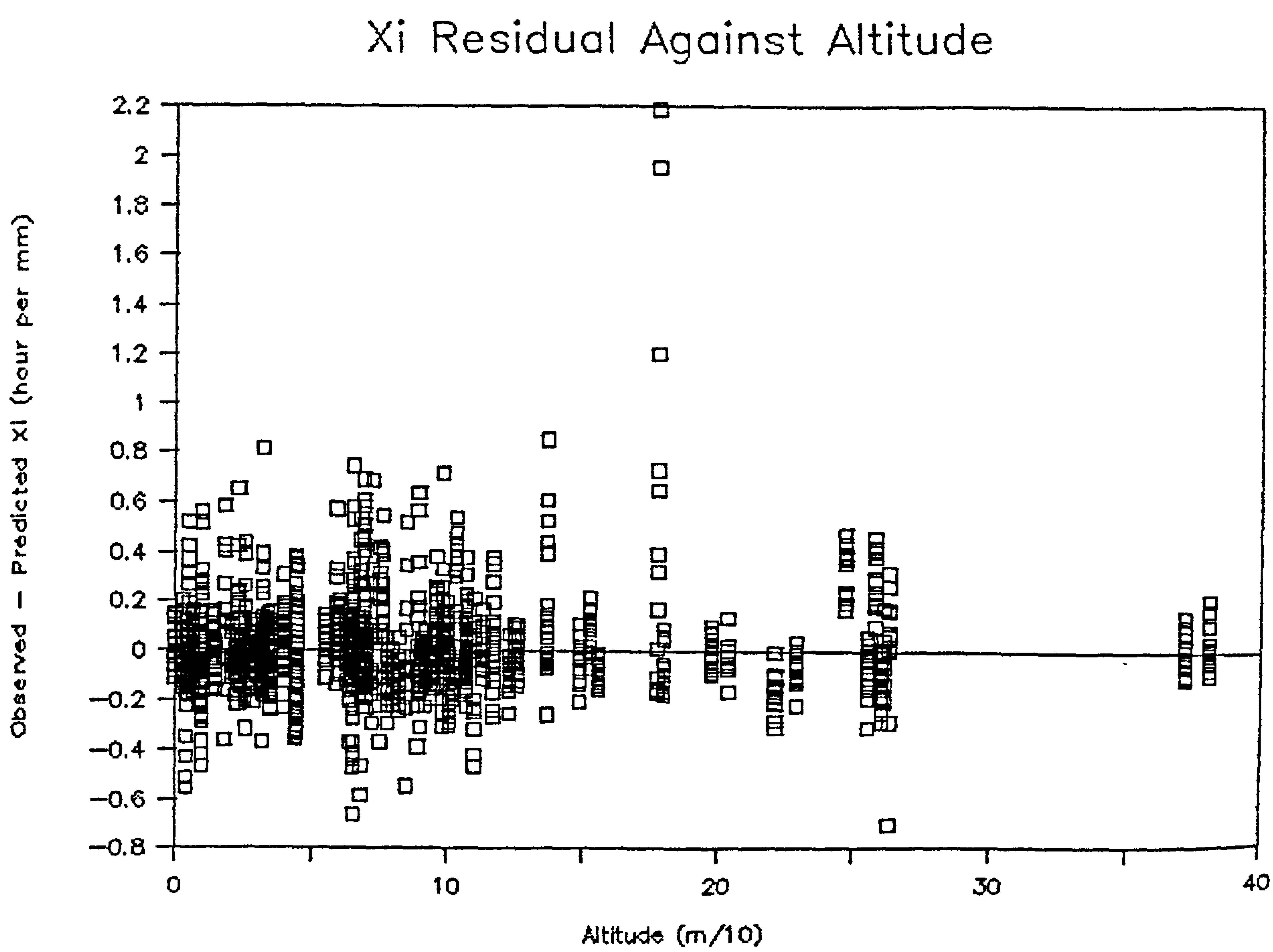


Figure G.15

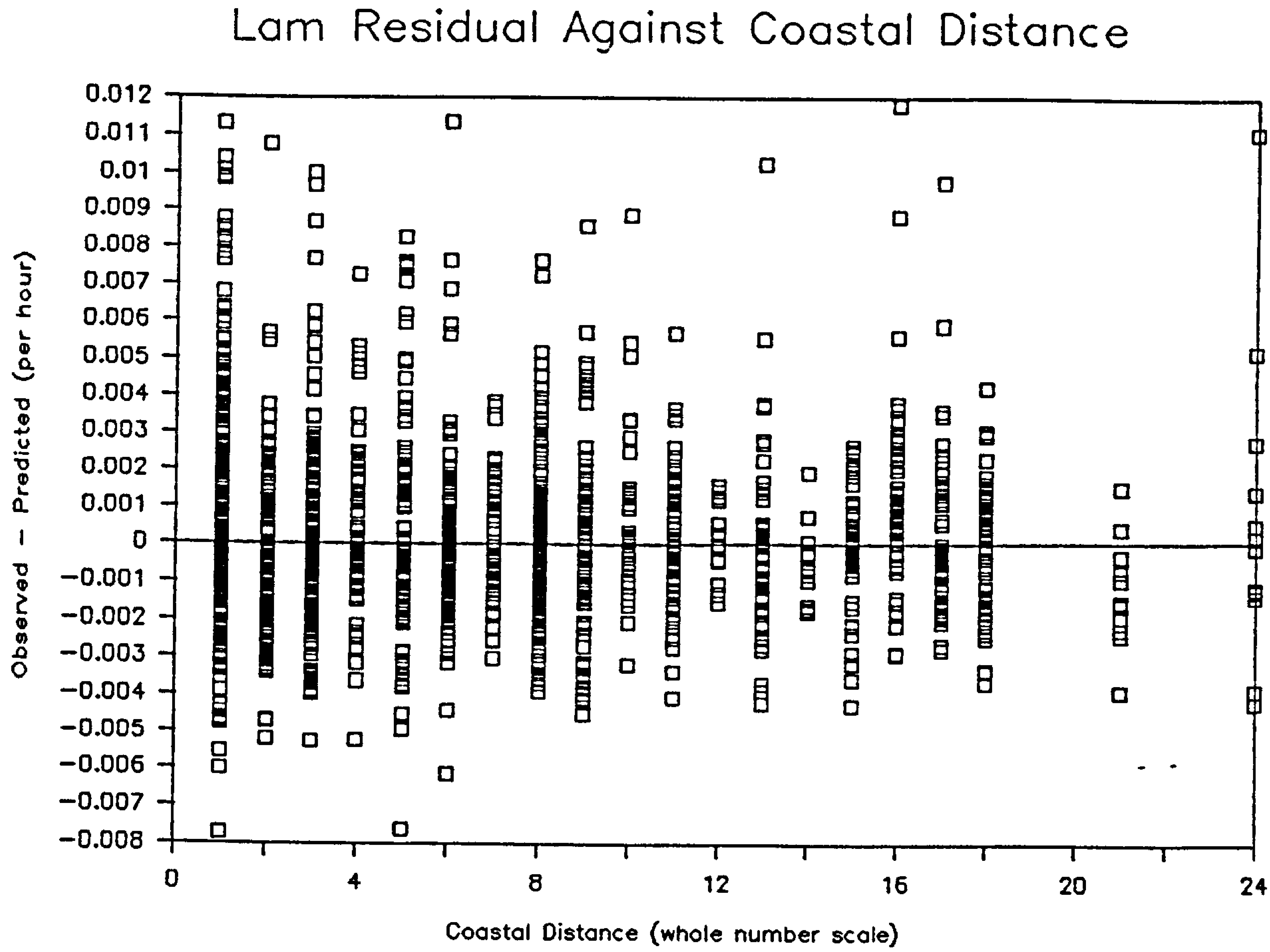


Figure G.16

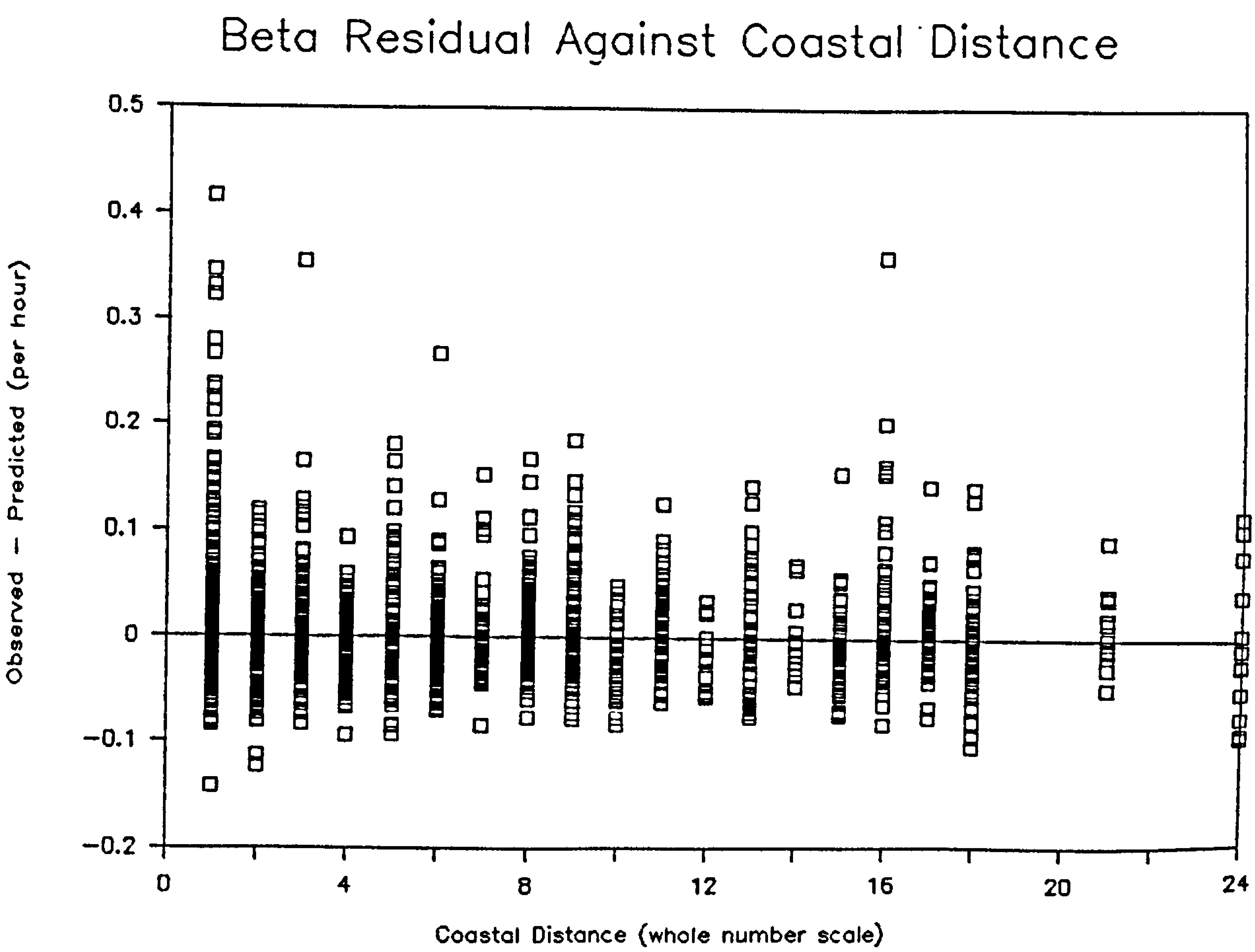


Figure G.17

Eta Residual Against Coastal Distance

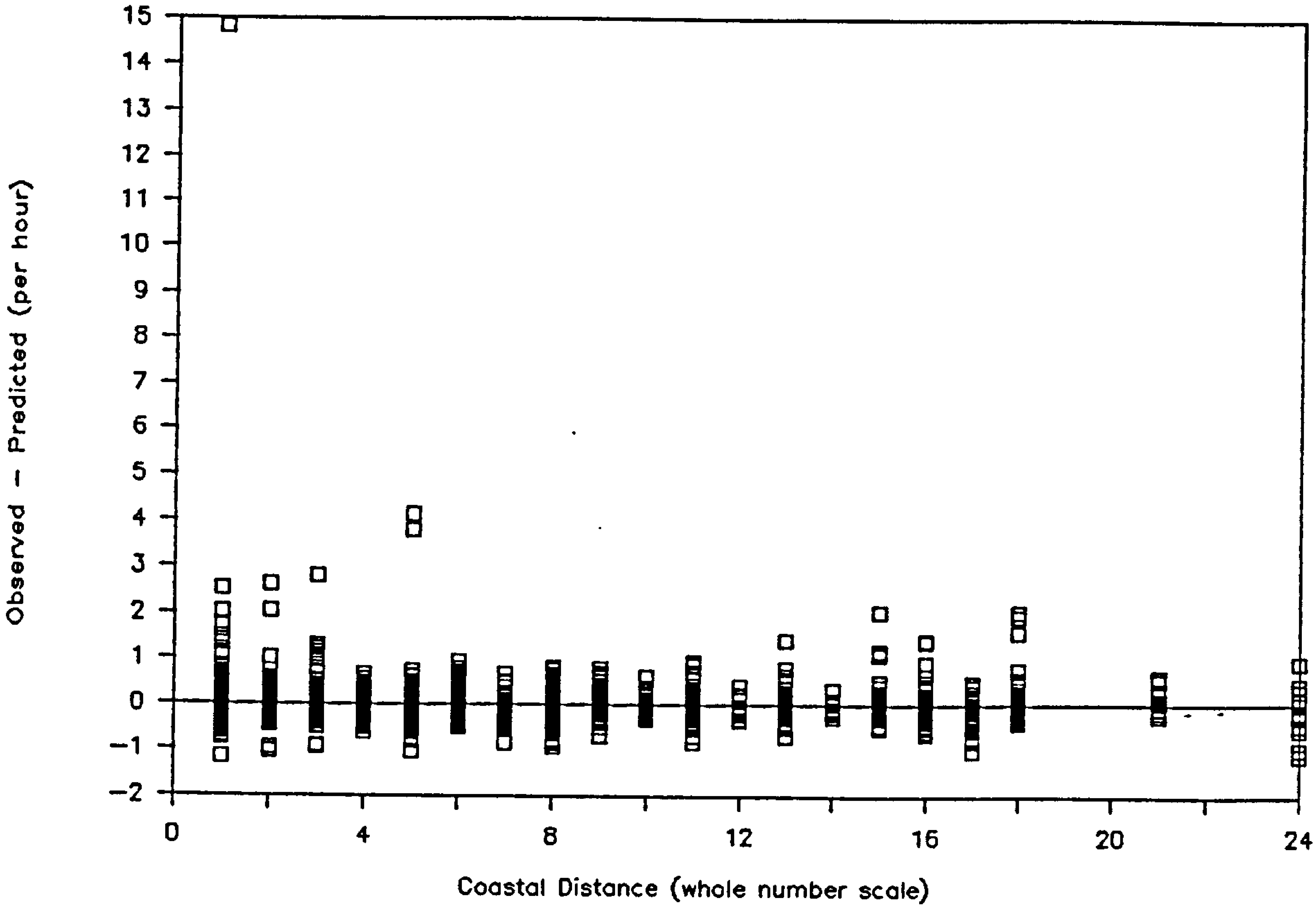


Figure G.18

Nu Residual Against Coastal Distance

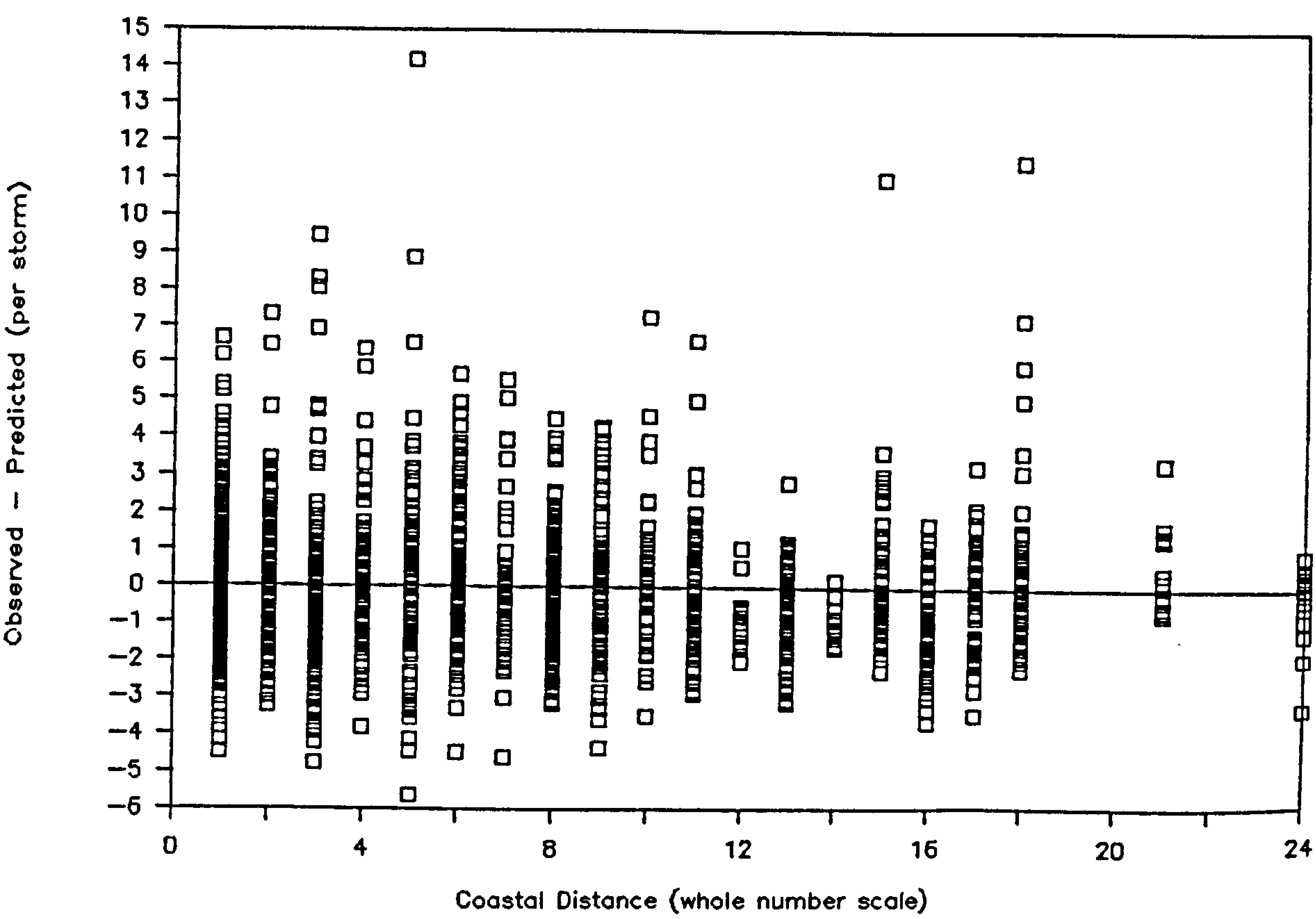


Figure G.19

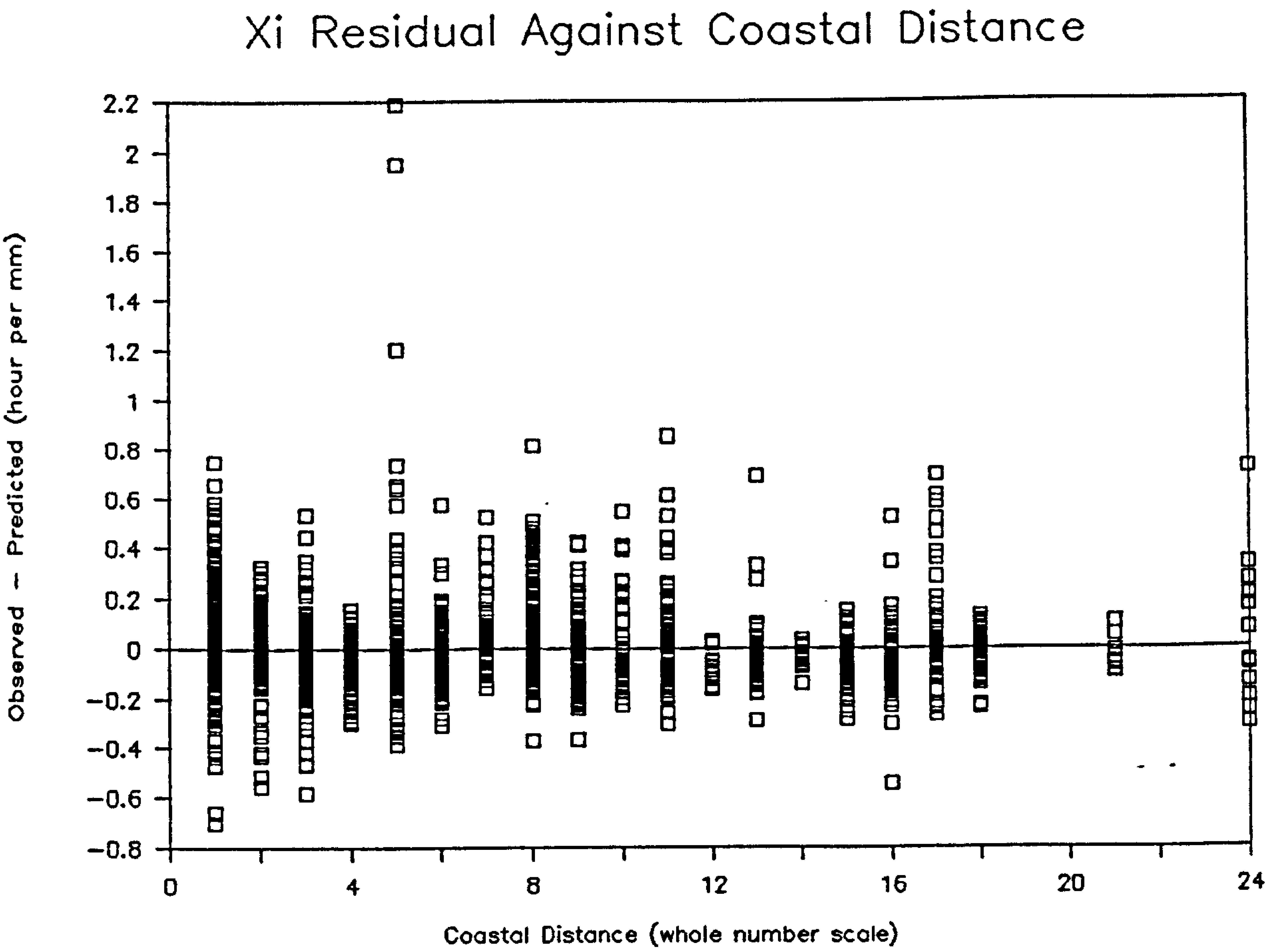


Figure G.20

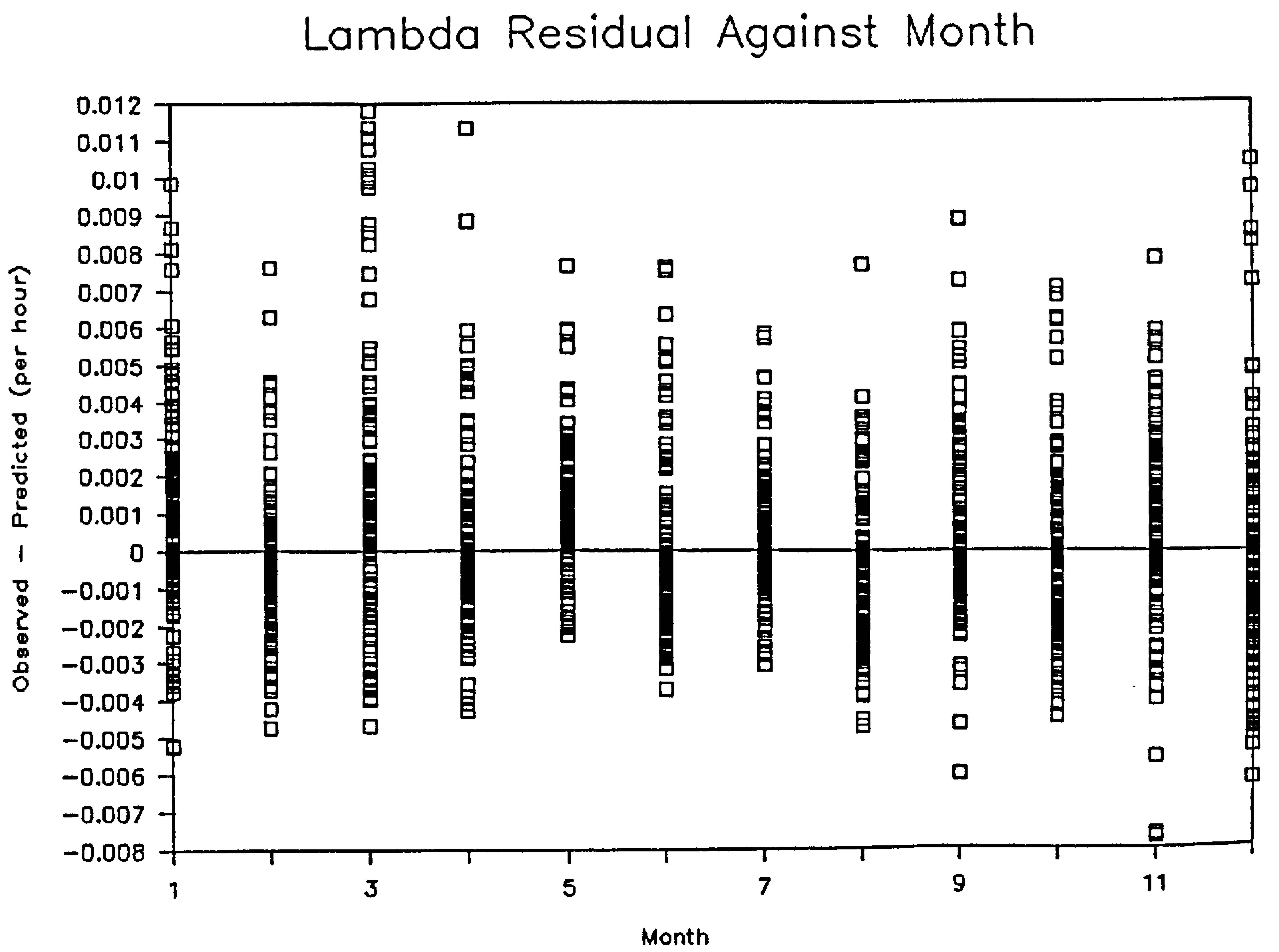


Figure G.21

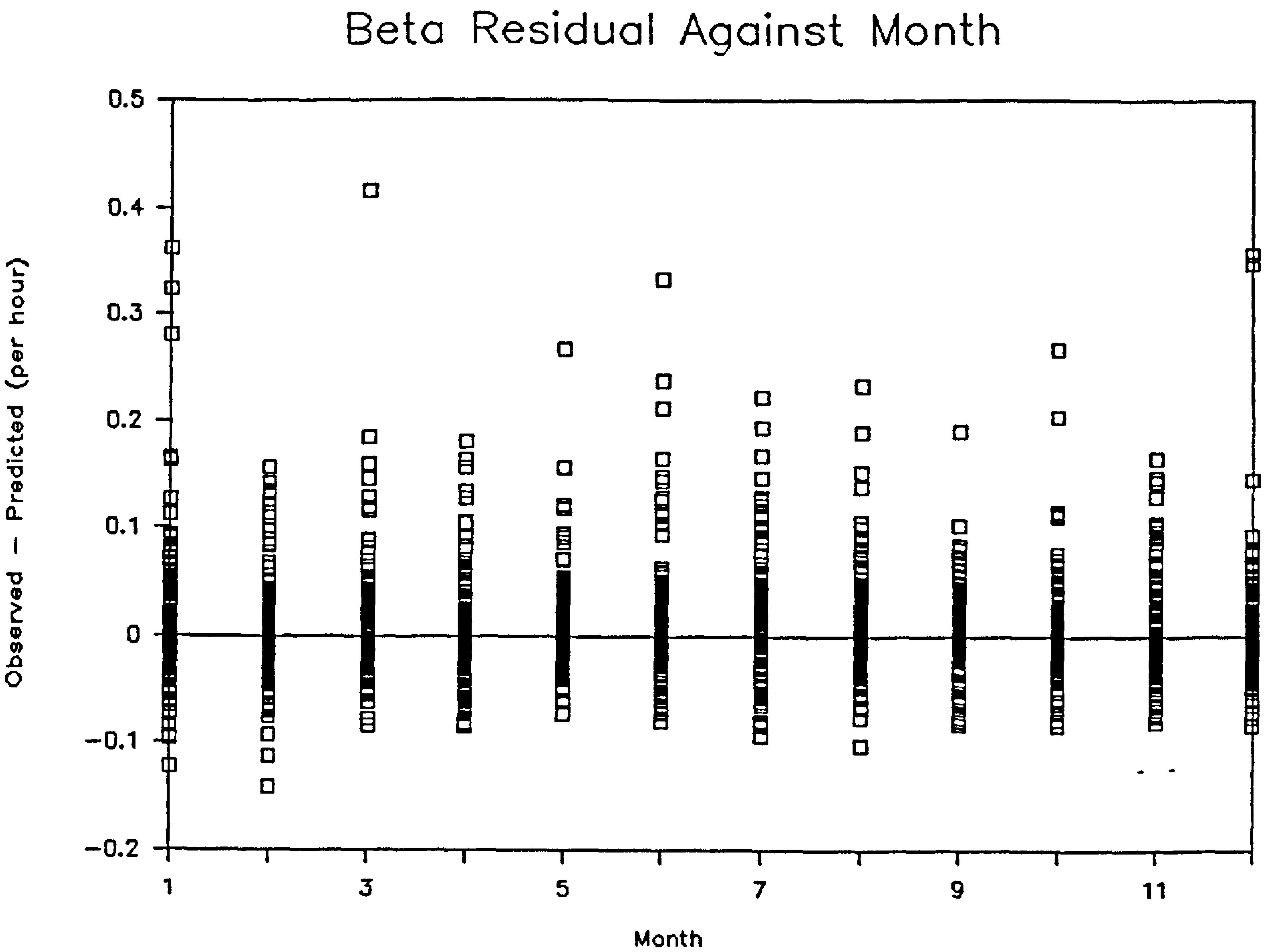


Figure G.22

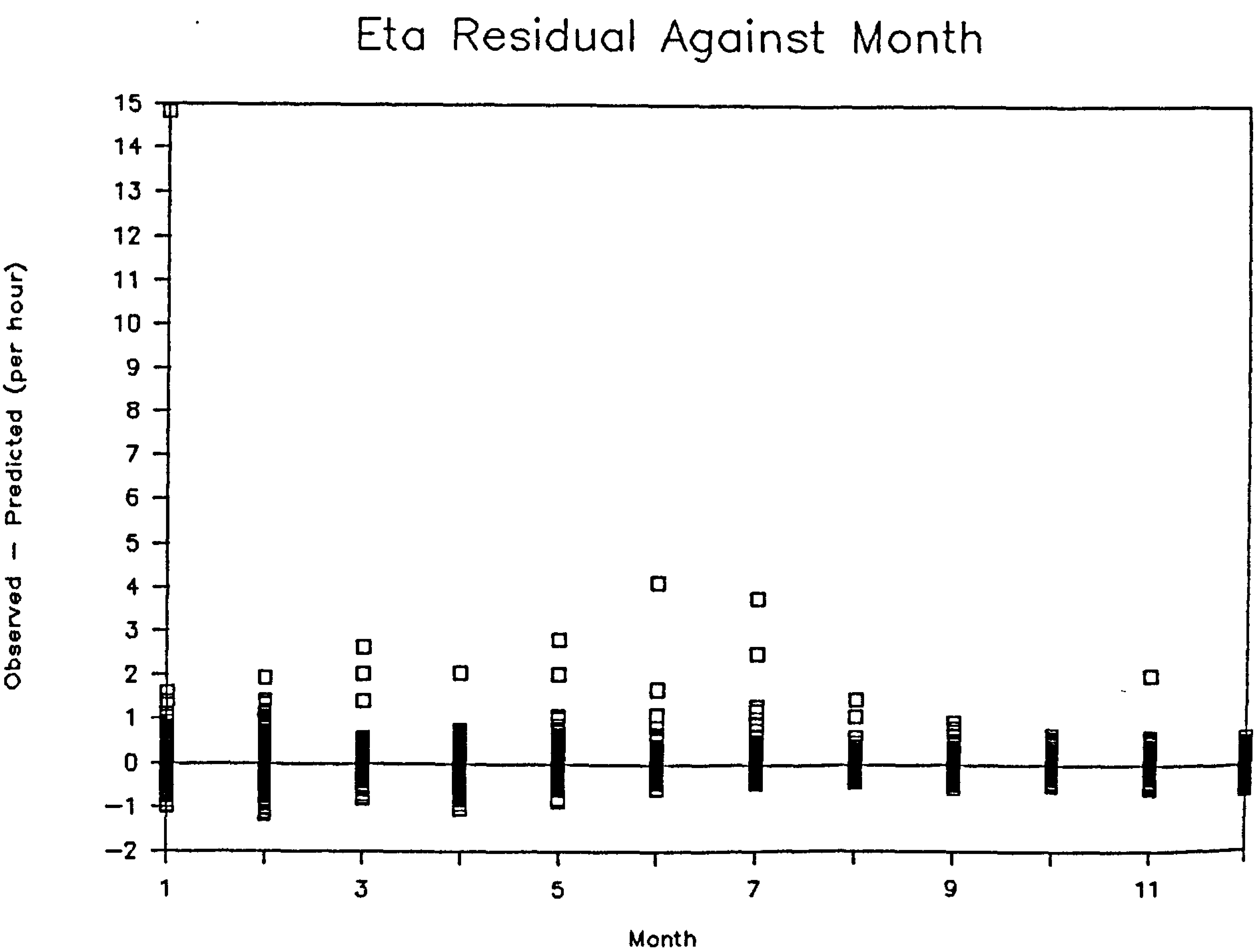


Figure G.23

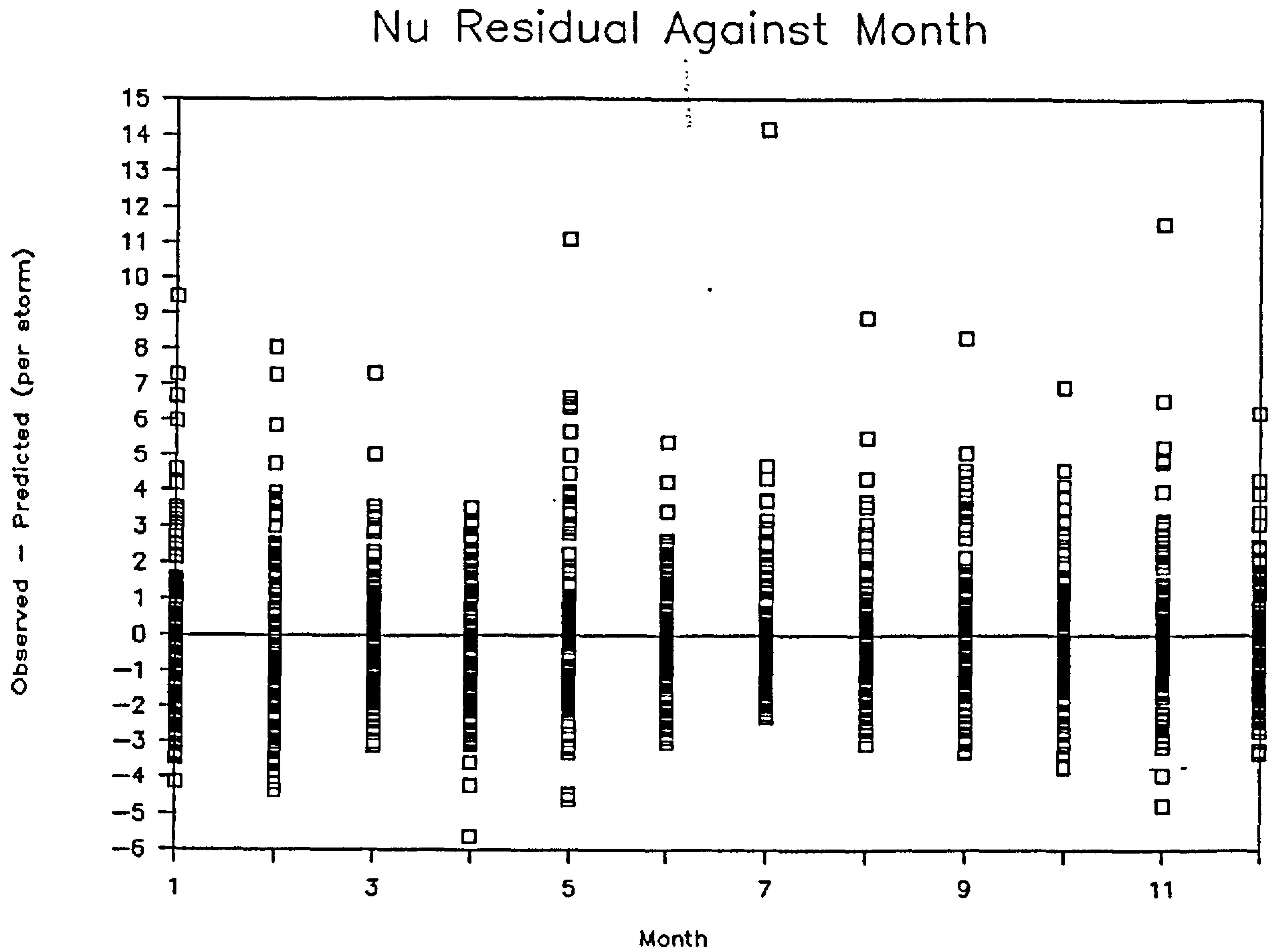


Figure G.24

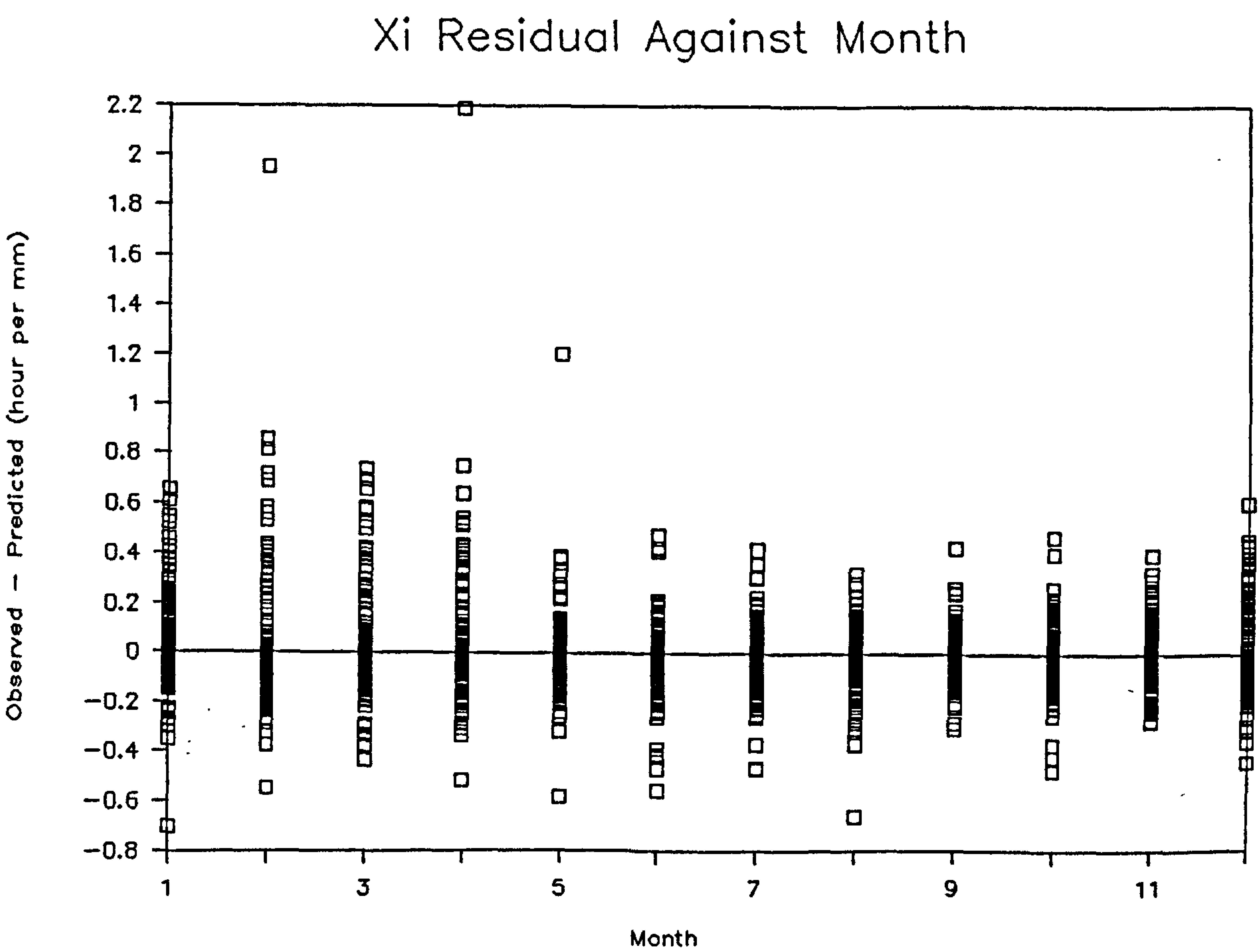


Figure G.25

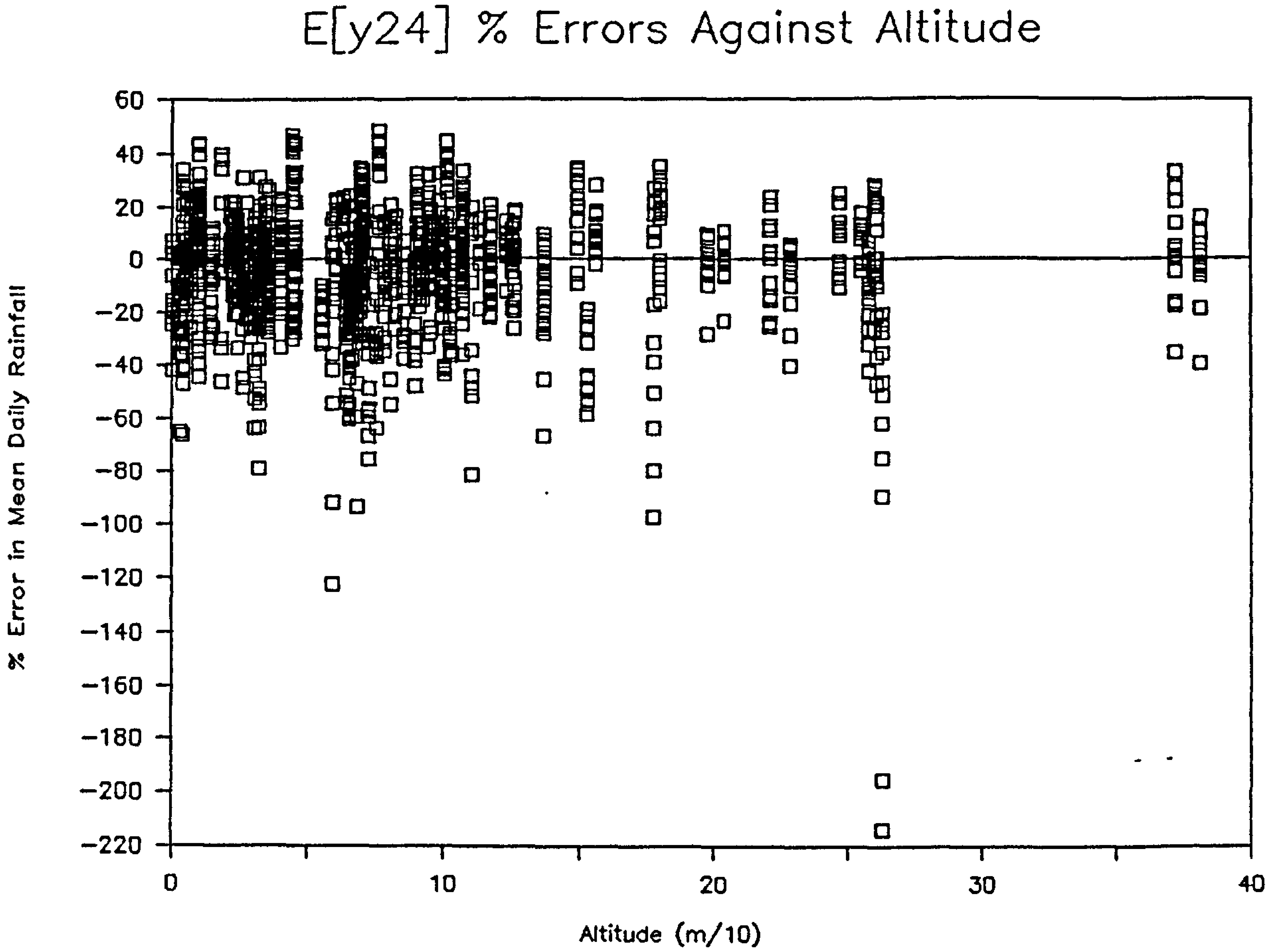


Figure G.26

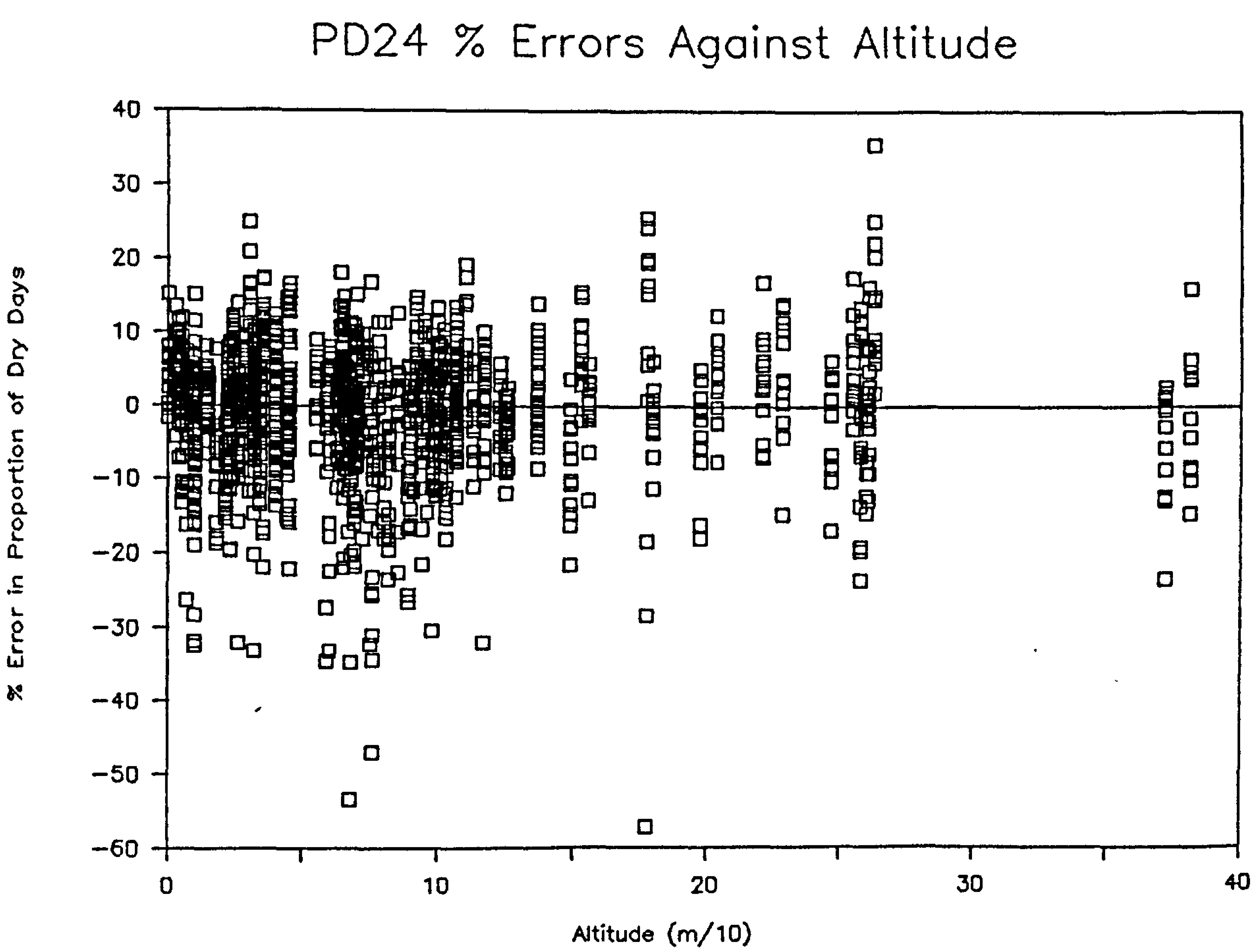


Figure G.27

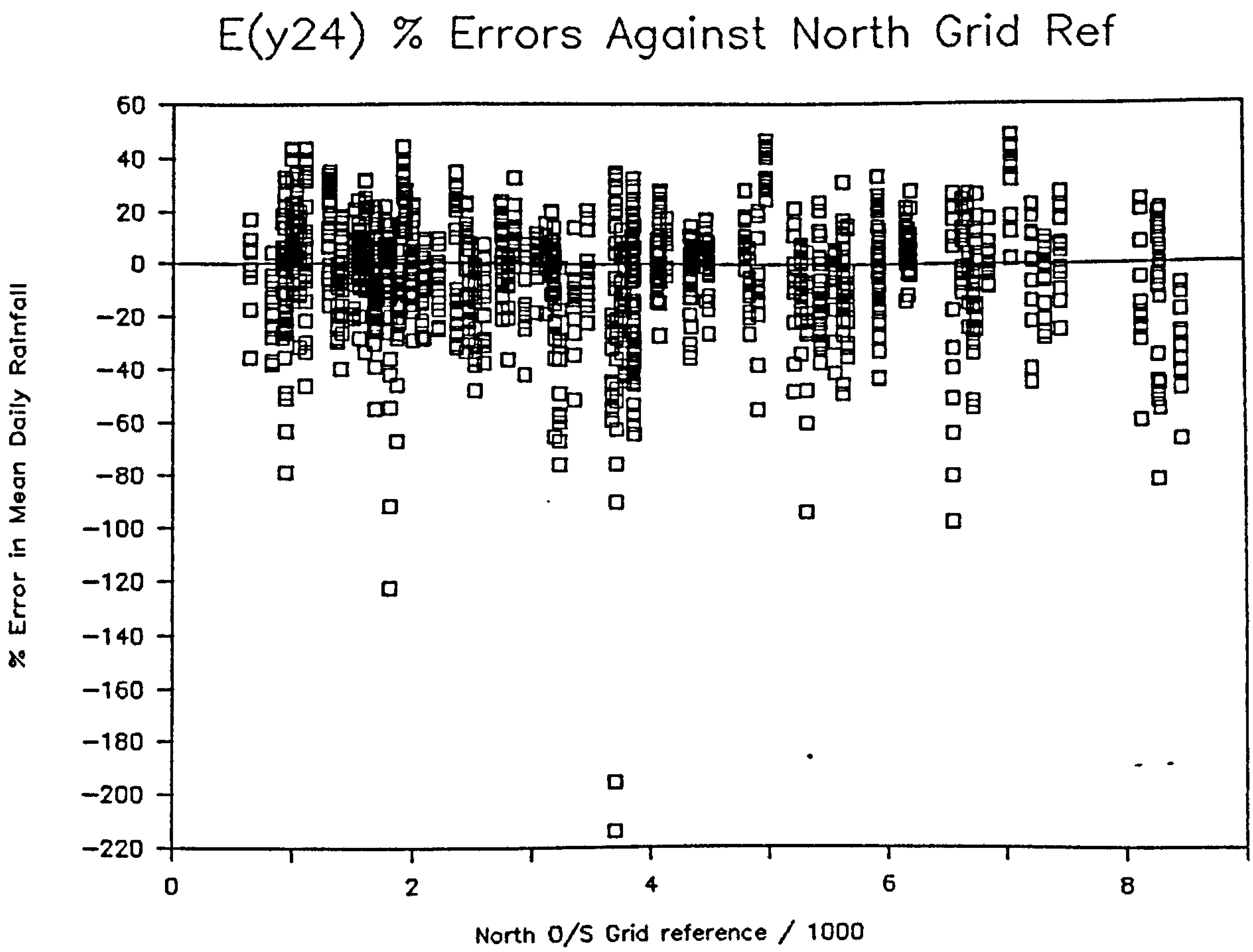


Figure G.28

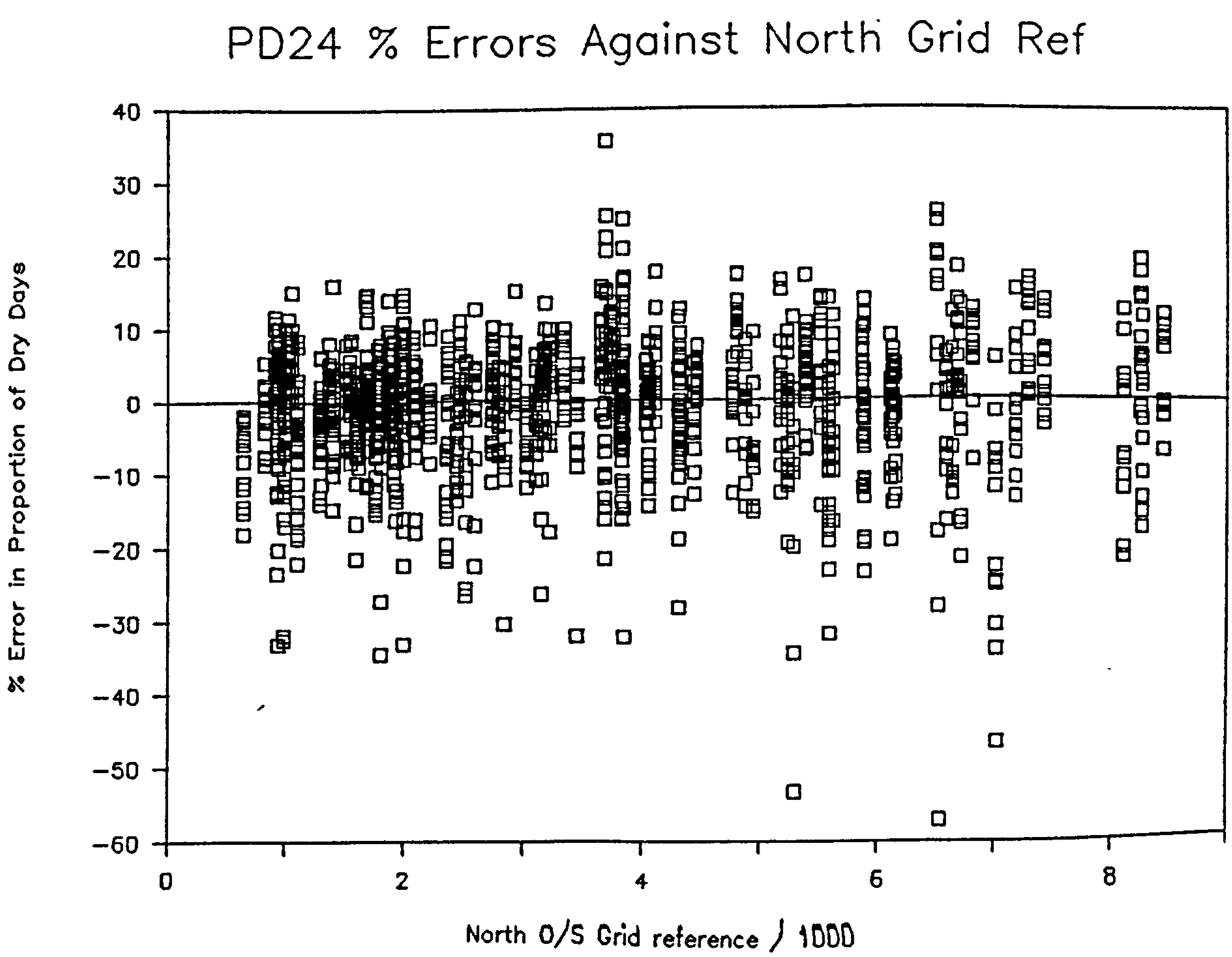


Figure G.29

E(y24) % Errors Against Coastal Dist

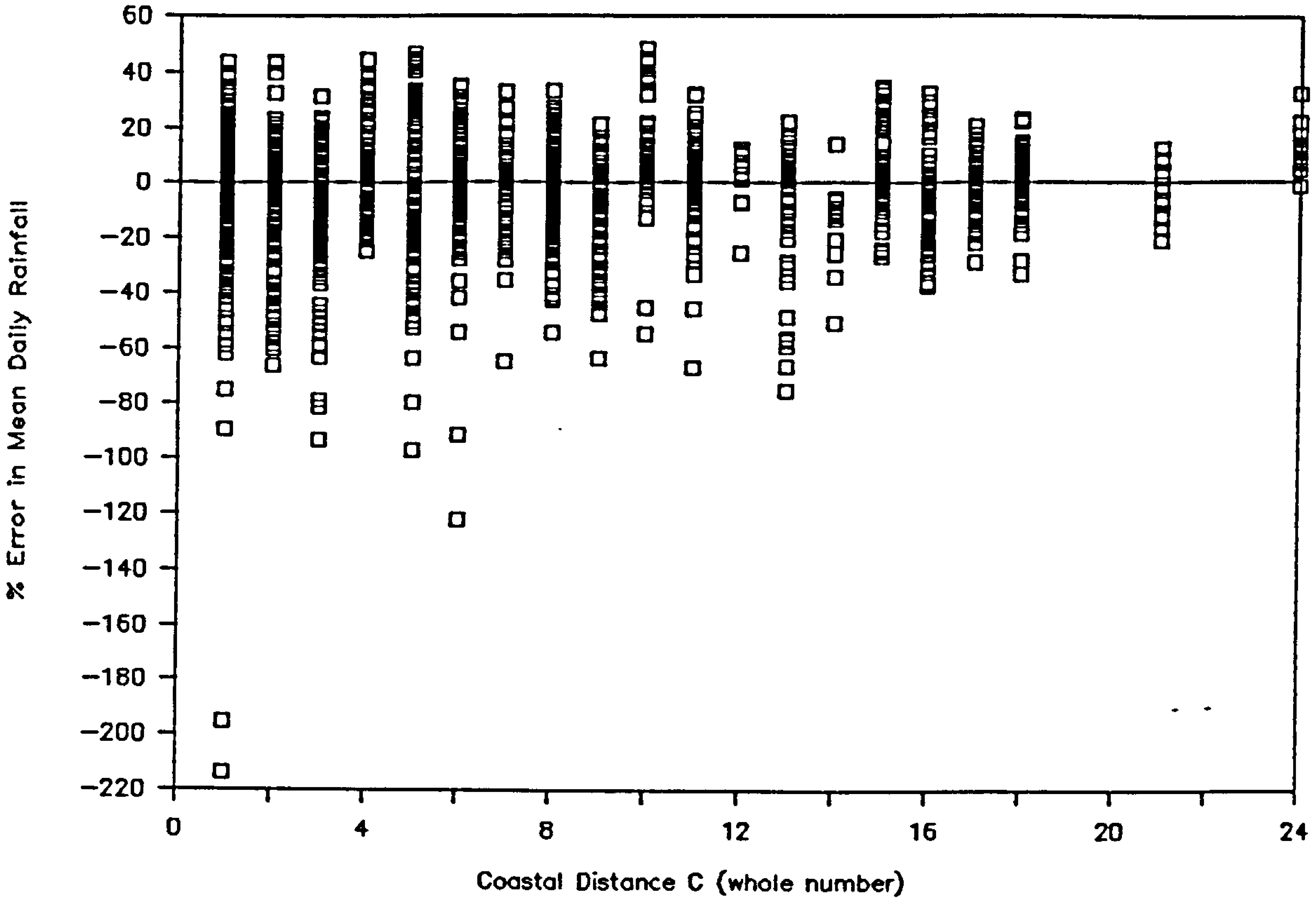


Figure G.30

PD24 % Errors Against Coastal Distance

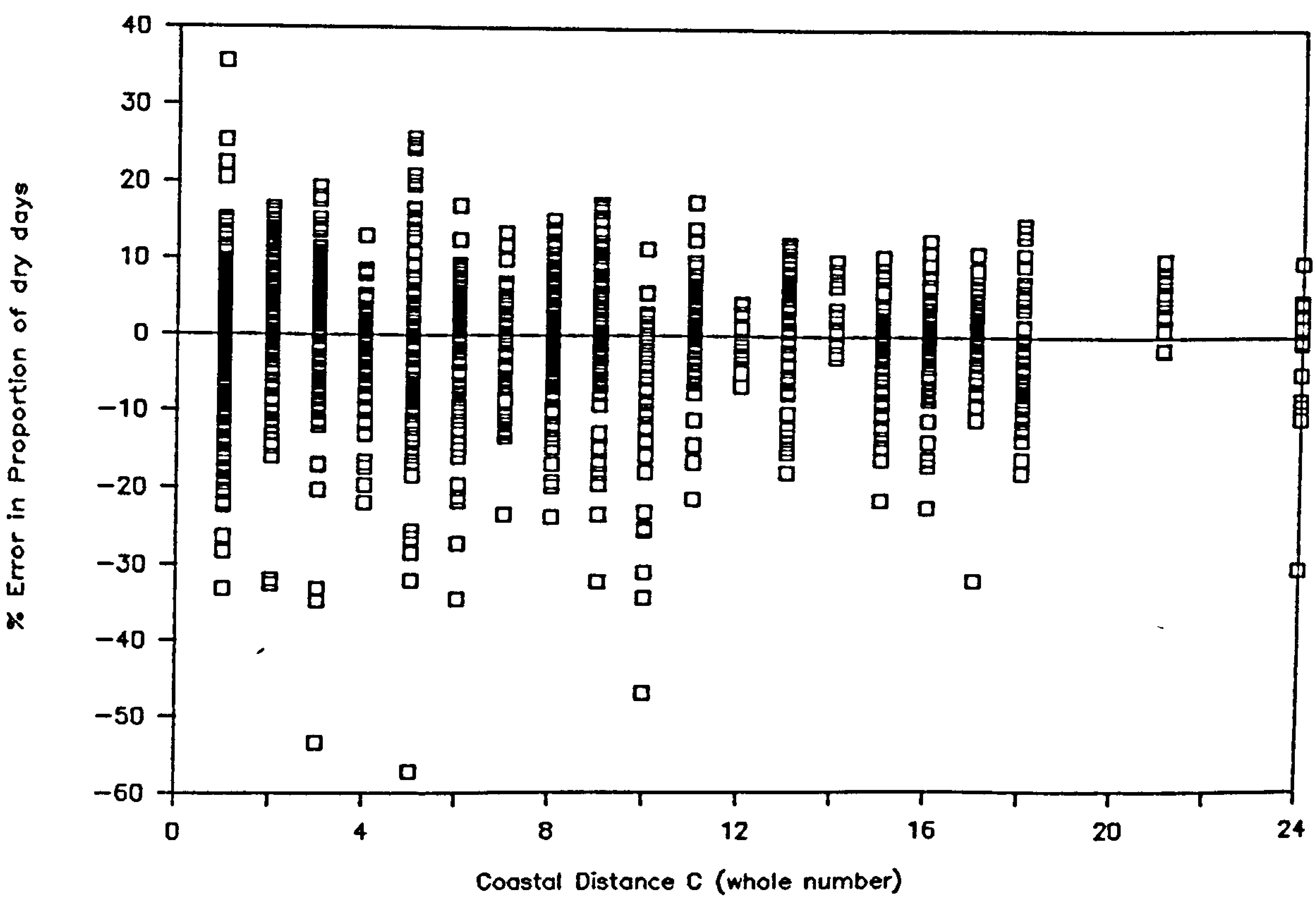


Figure G.31

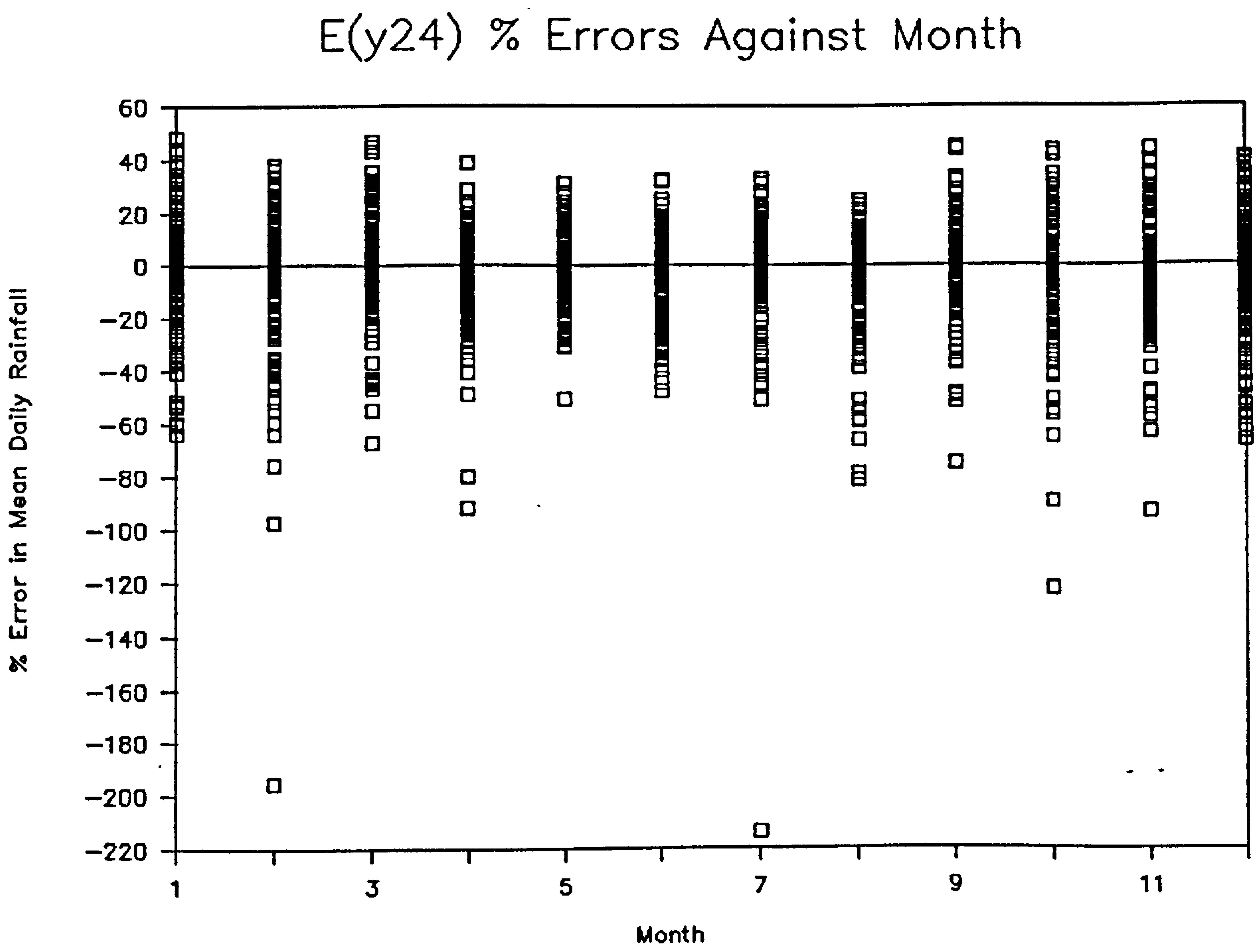


Figure G.32

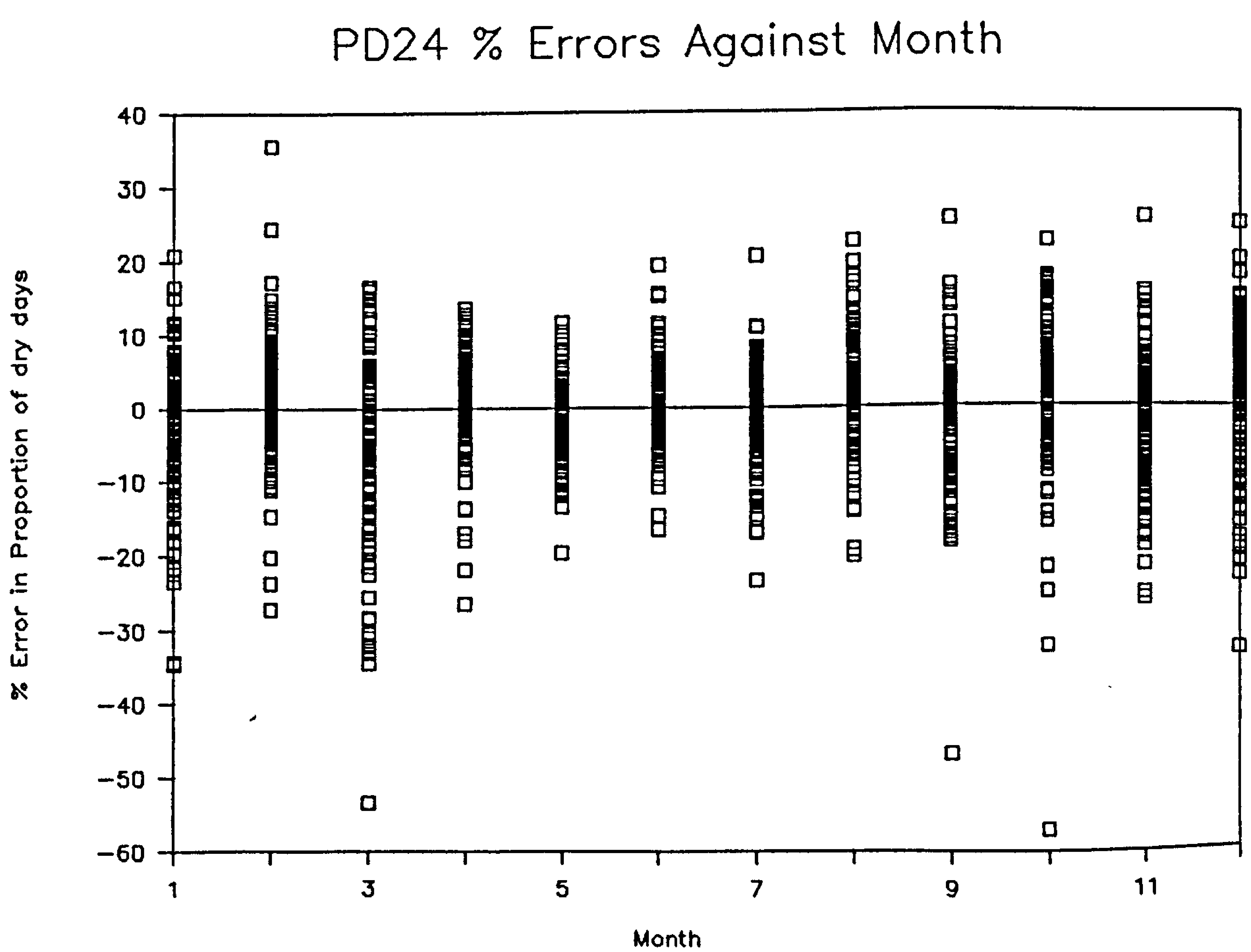


Figure G.33

Var(y24) % Errors Against Altitude

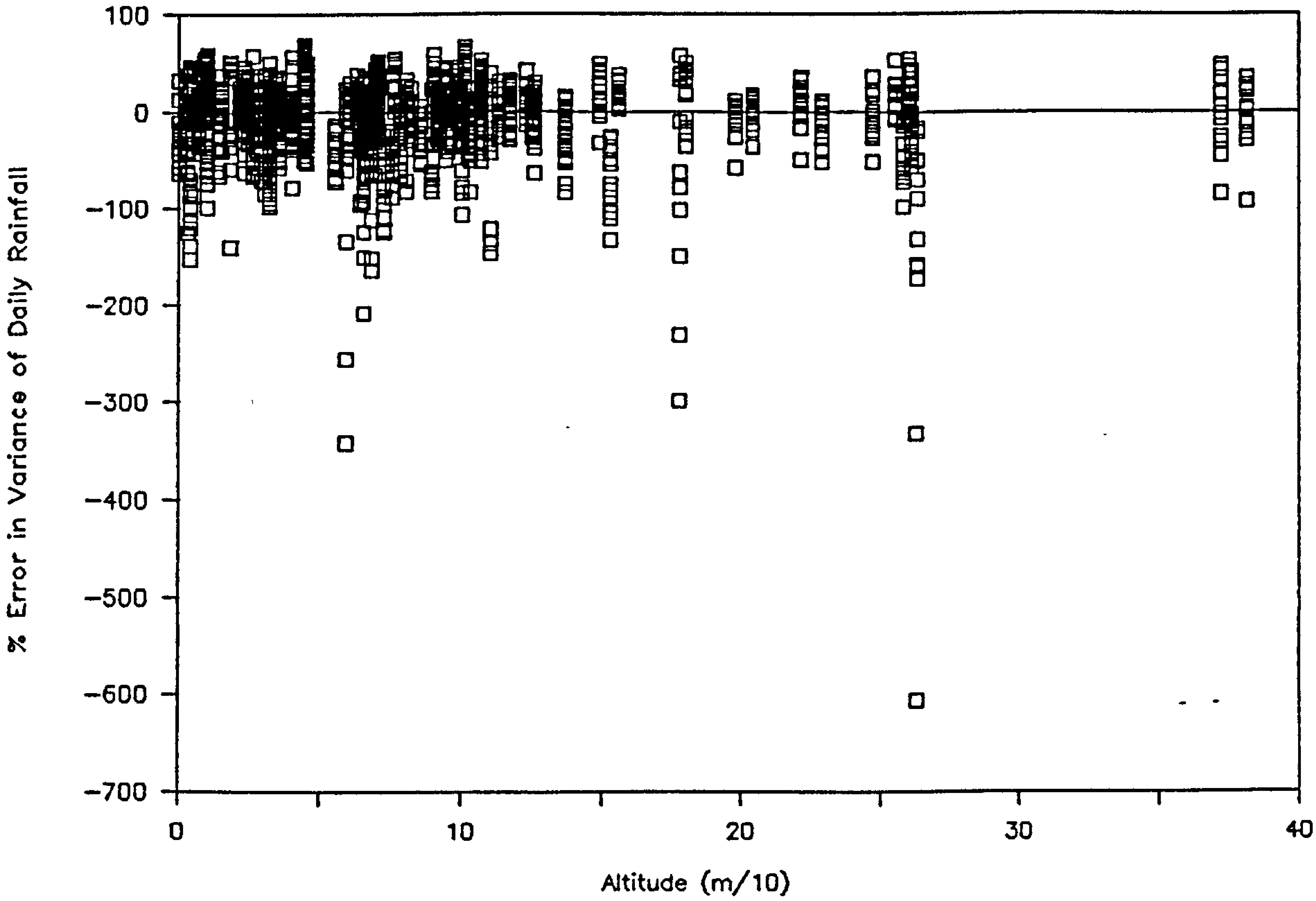


Figure G.34

Var(y24) % Errors Against Altitude

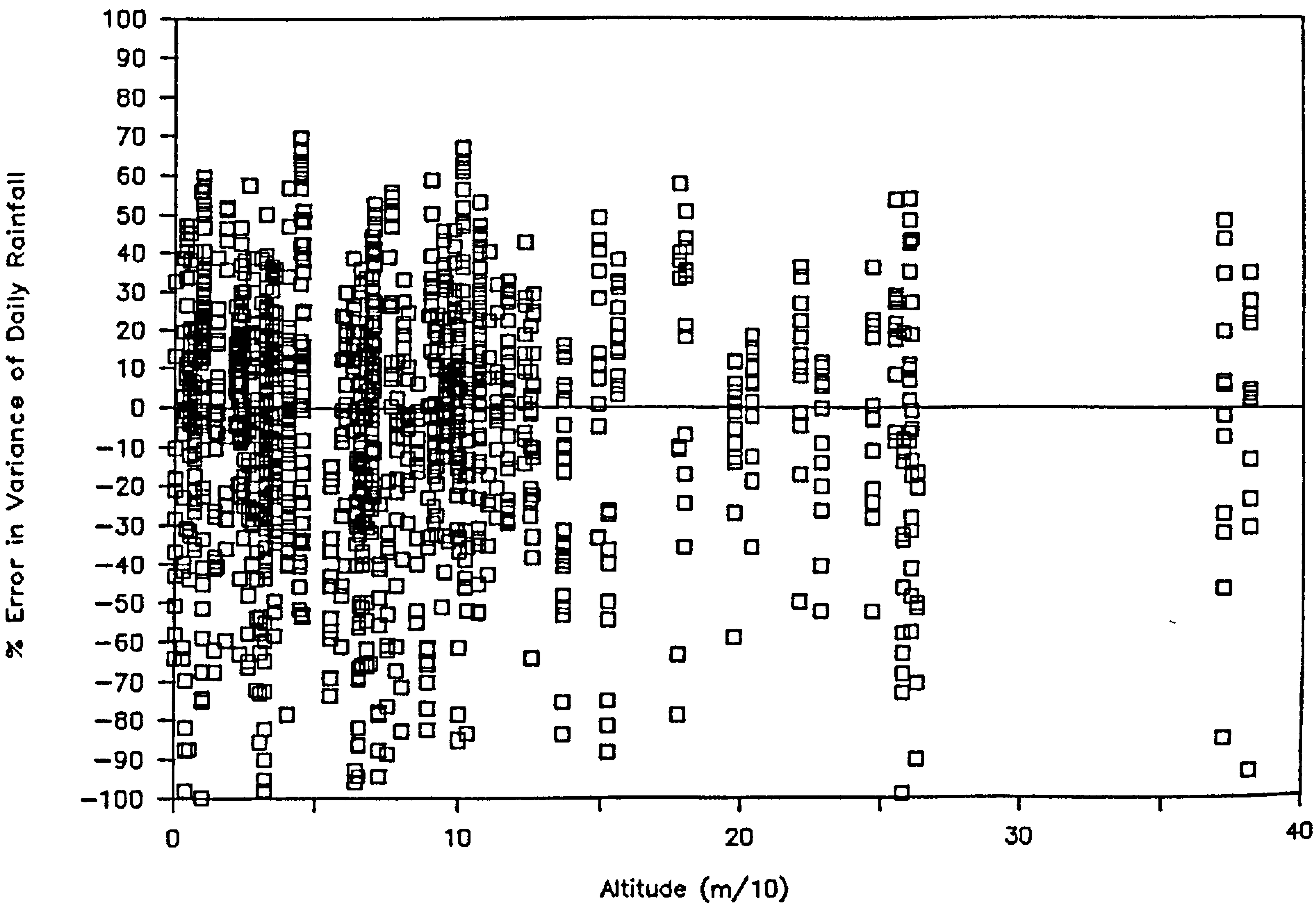


Figure G.35

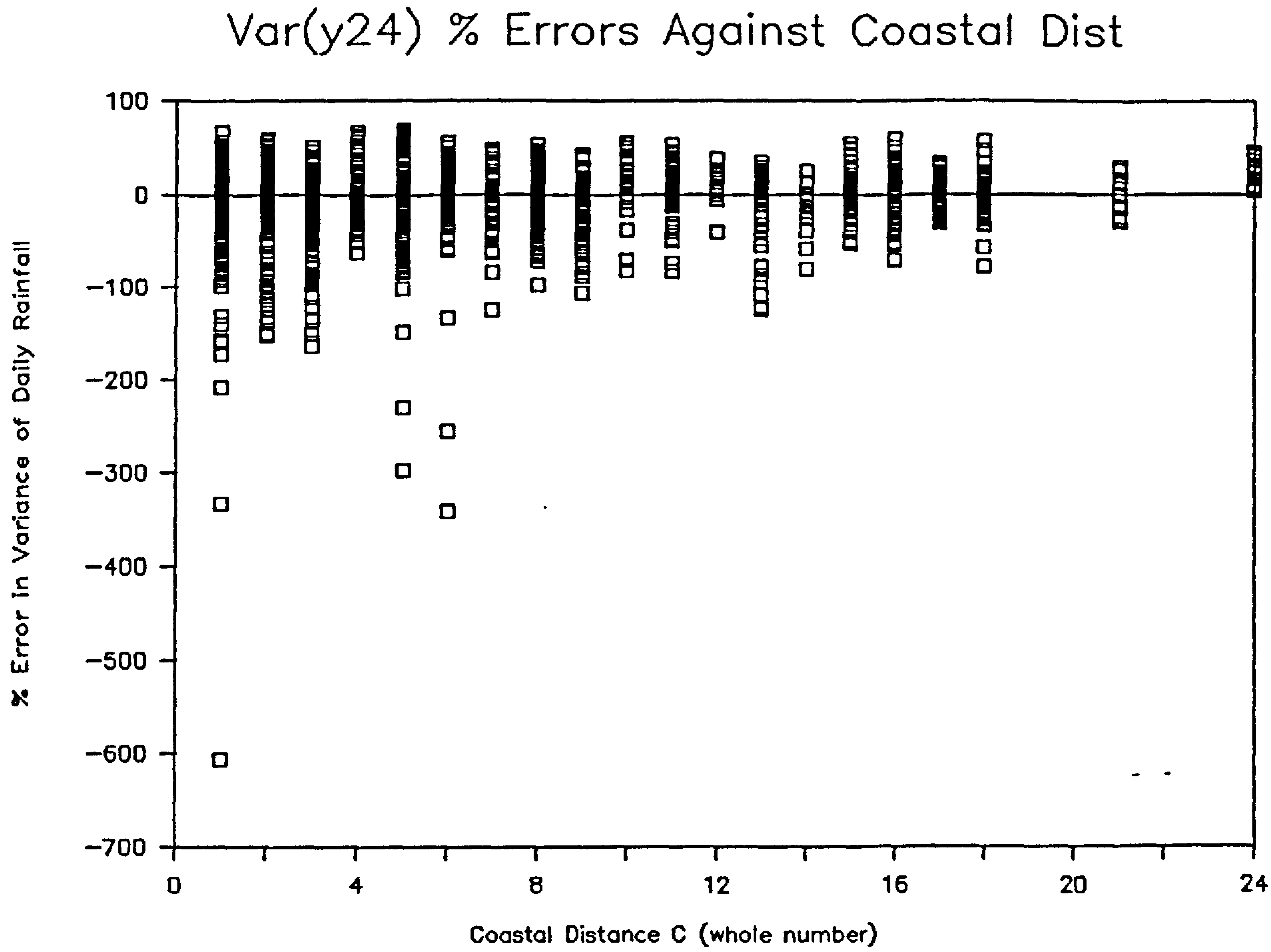


Figure G.36

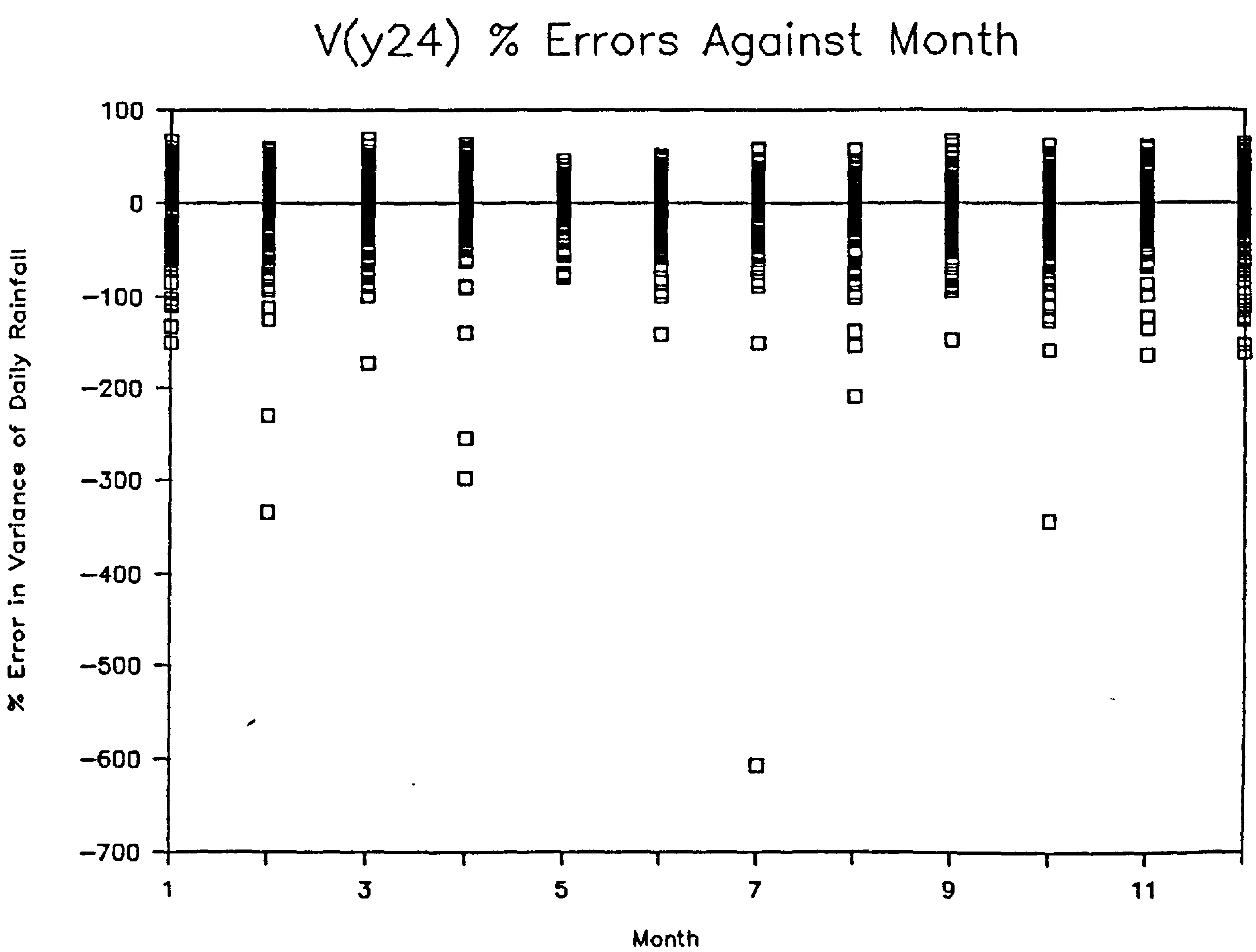
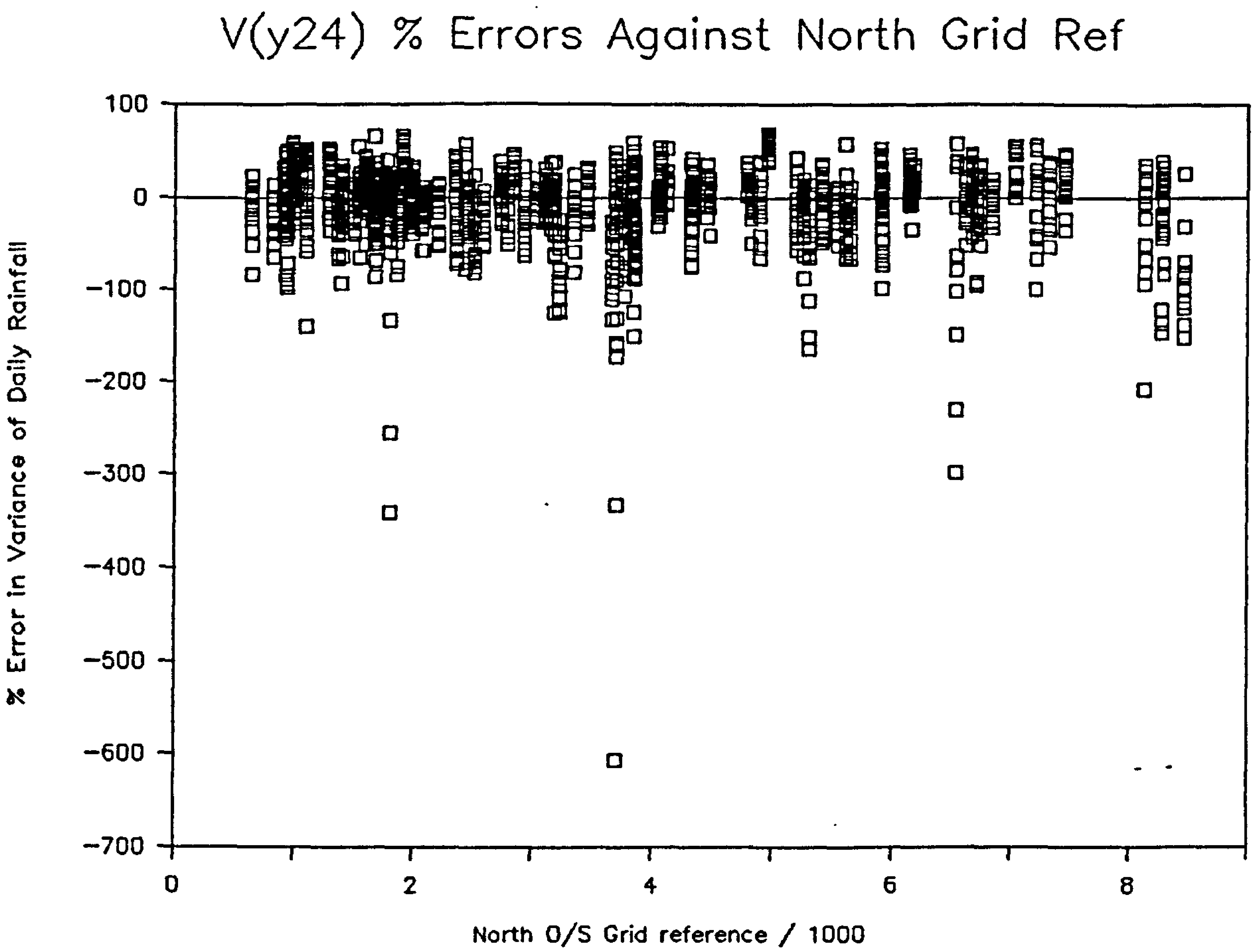
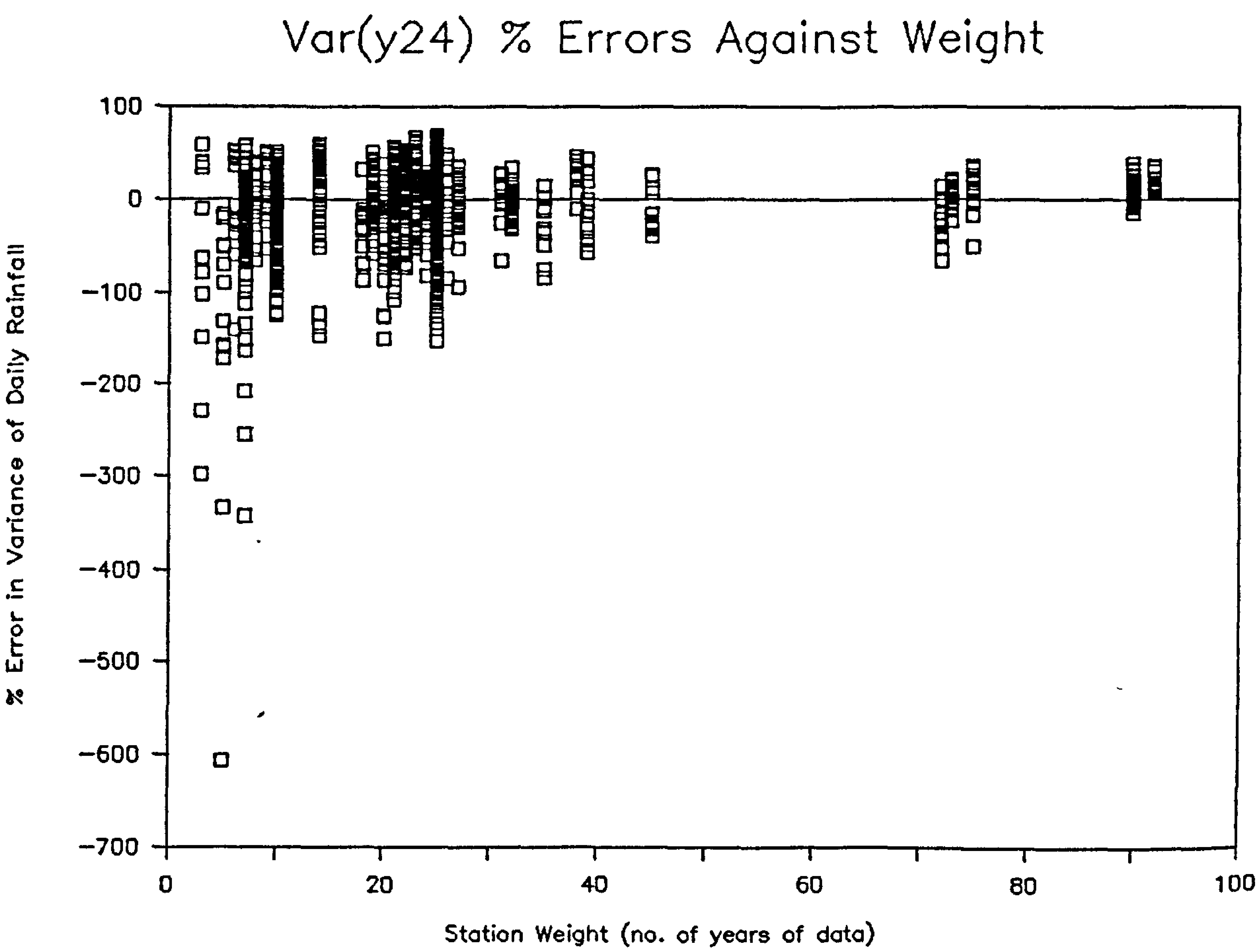


Figure G.37



<u>**Figure G.38**</u>



<u>**Figure G.39**</u>

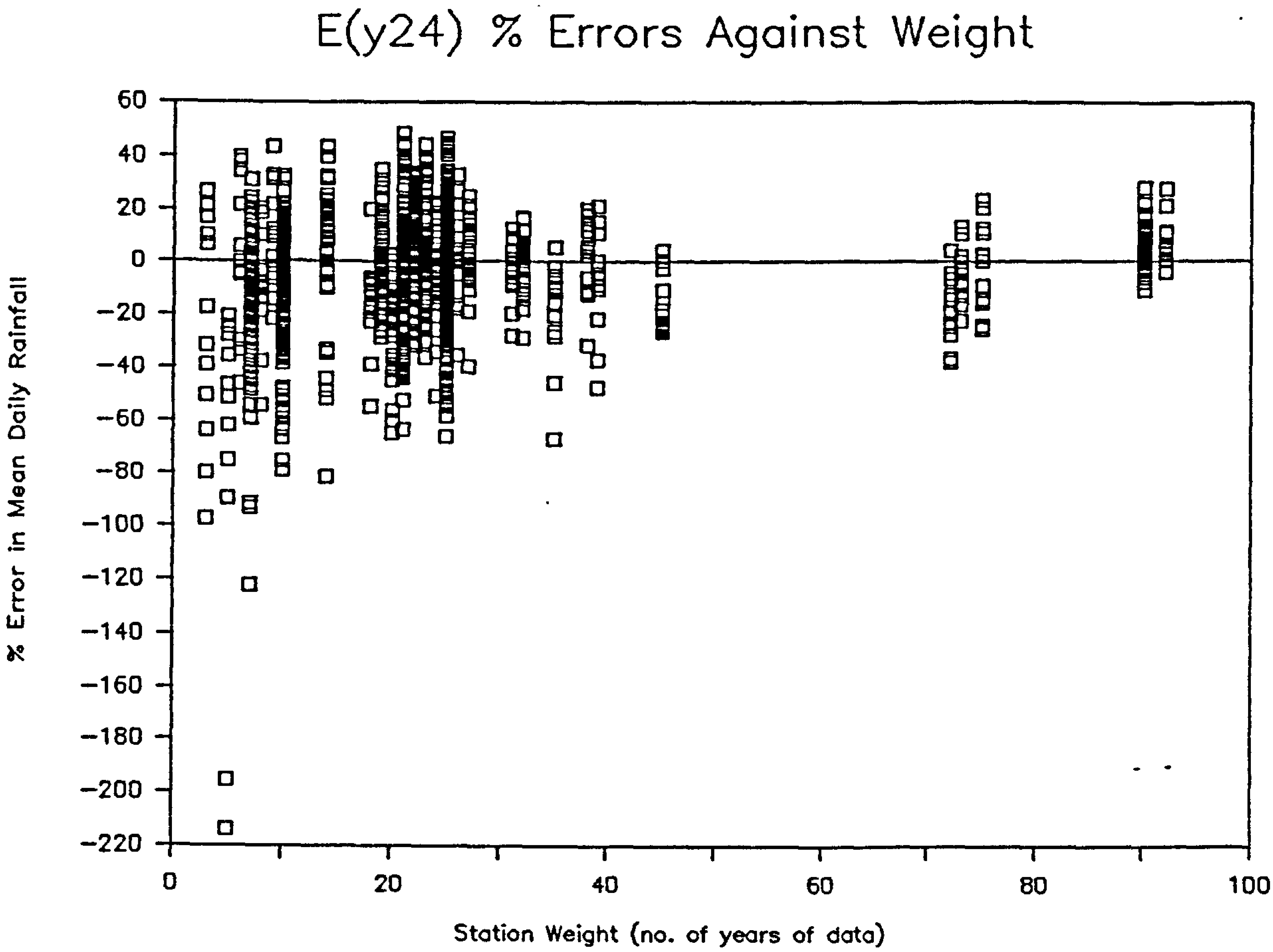


Figure G.40

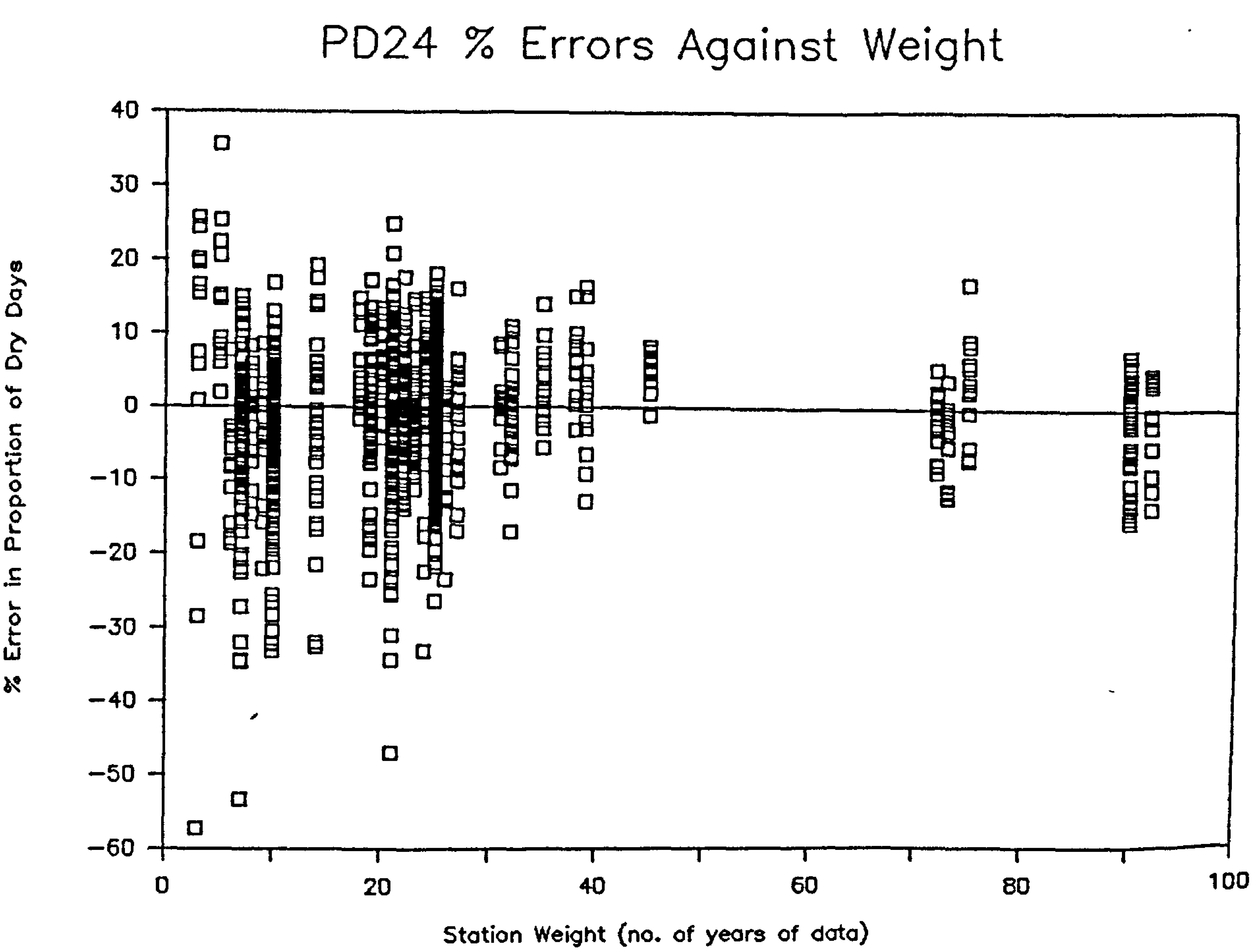
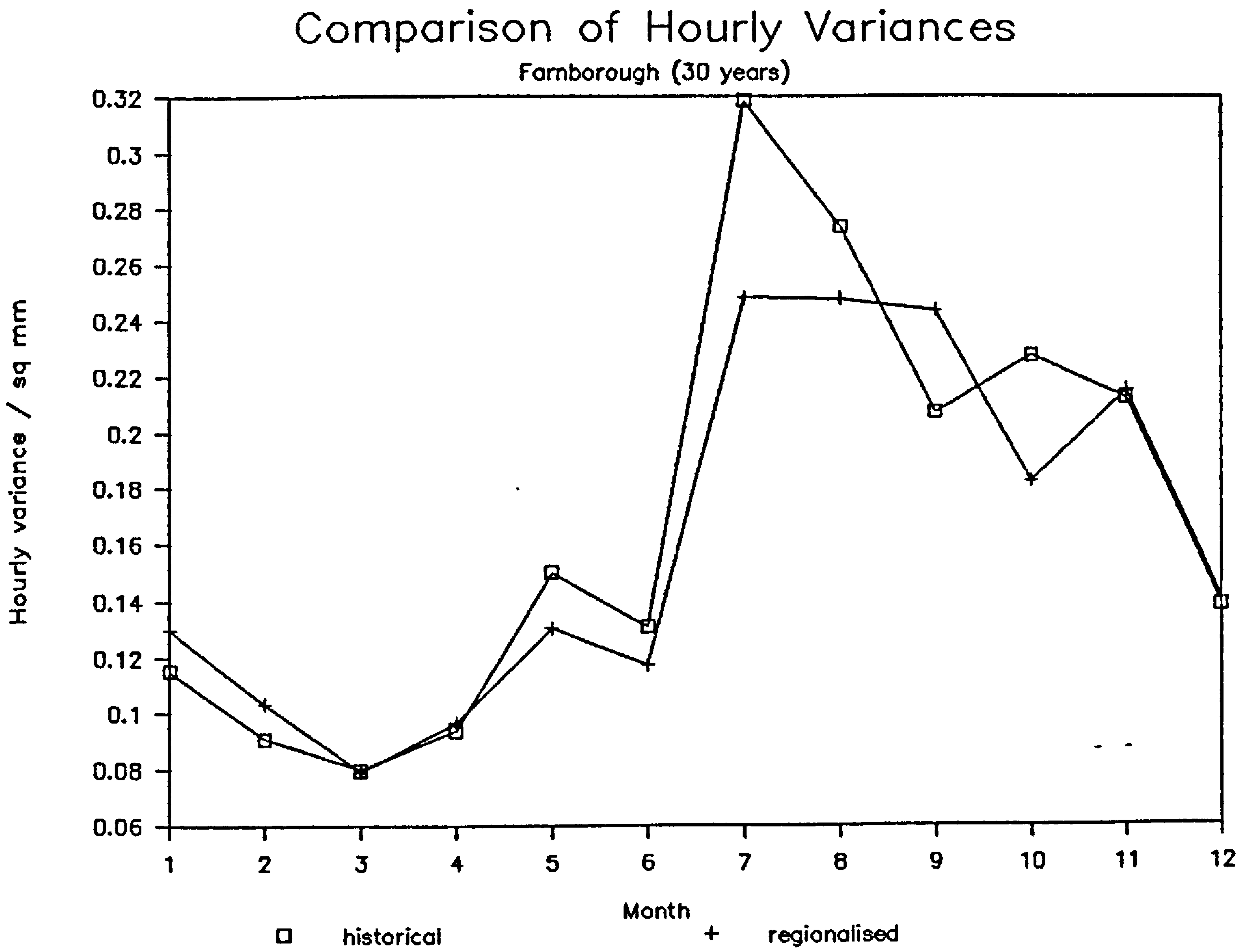
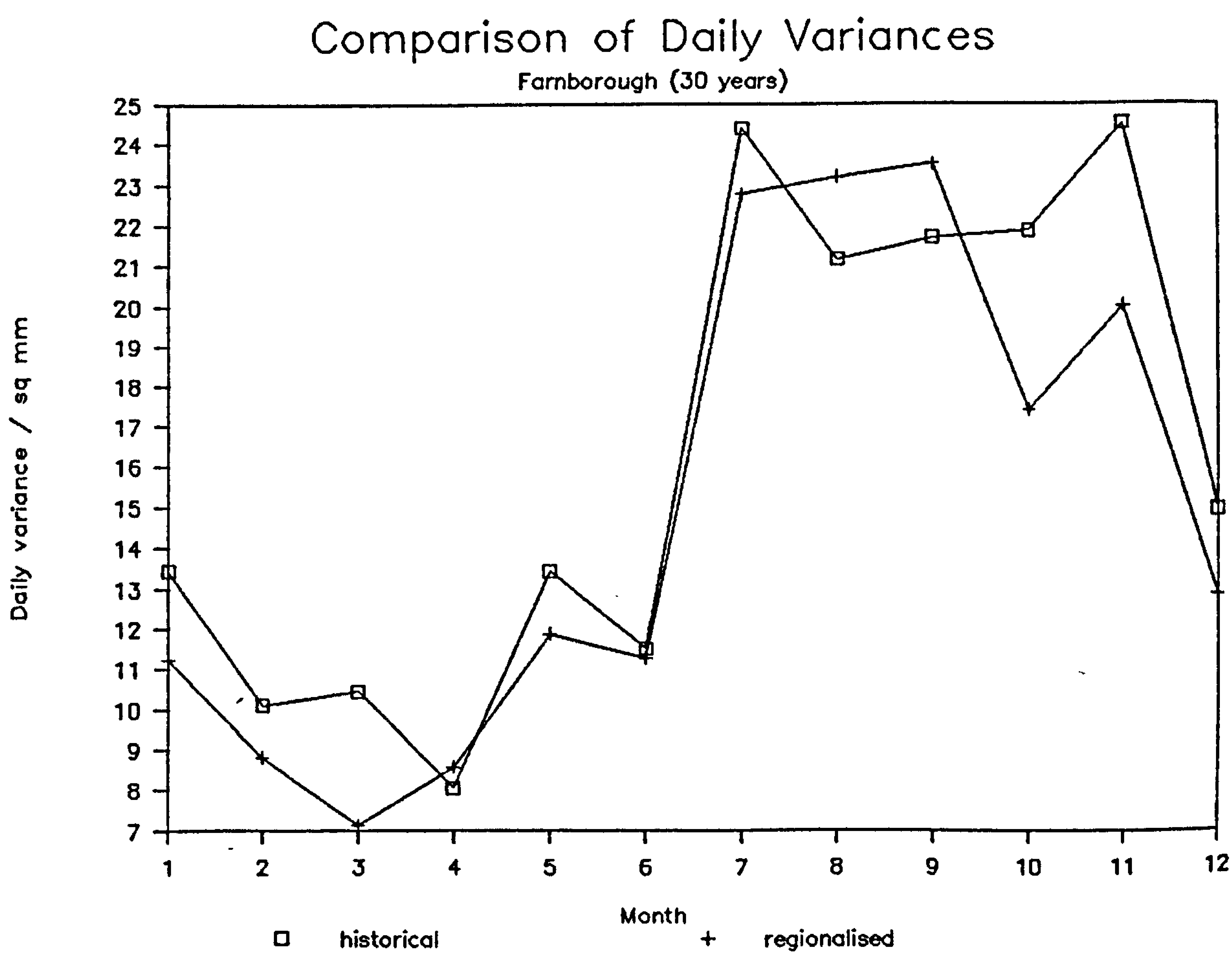


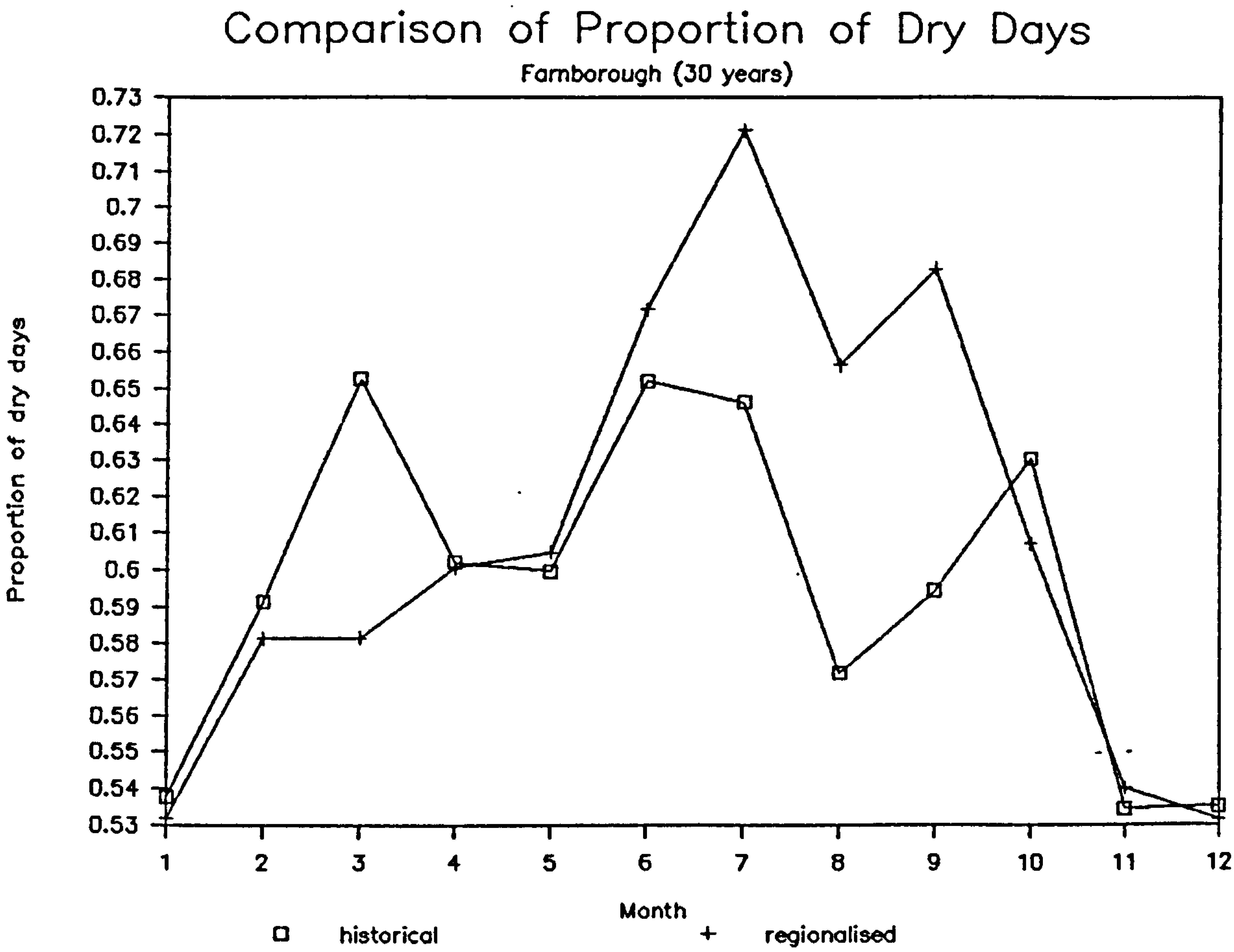
Figure G.41


Comparison of Hourly Variances
Farnborough (30 years)
Hourly variance / sq mm
0.32
0.3
0.28
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

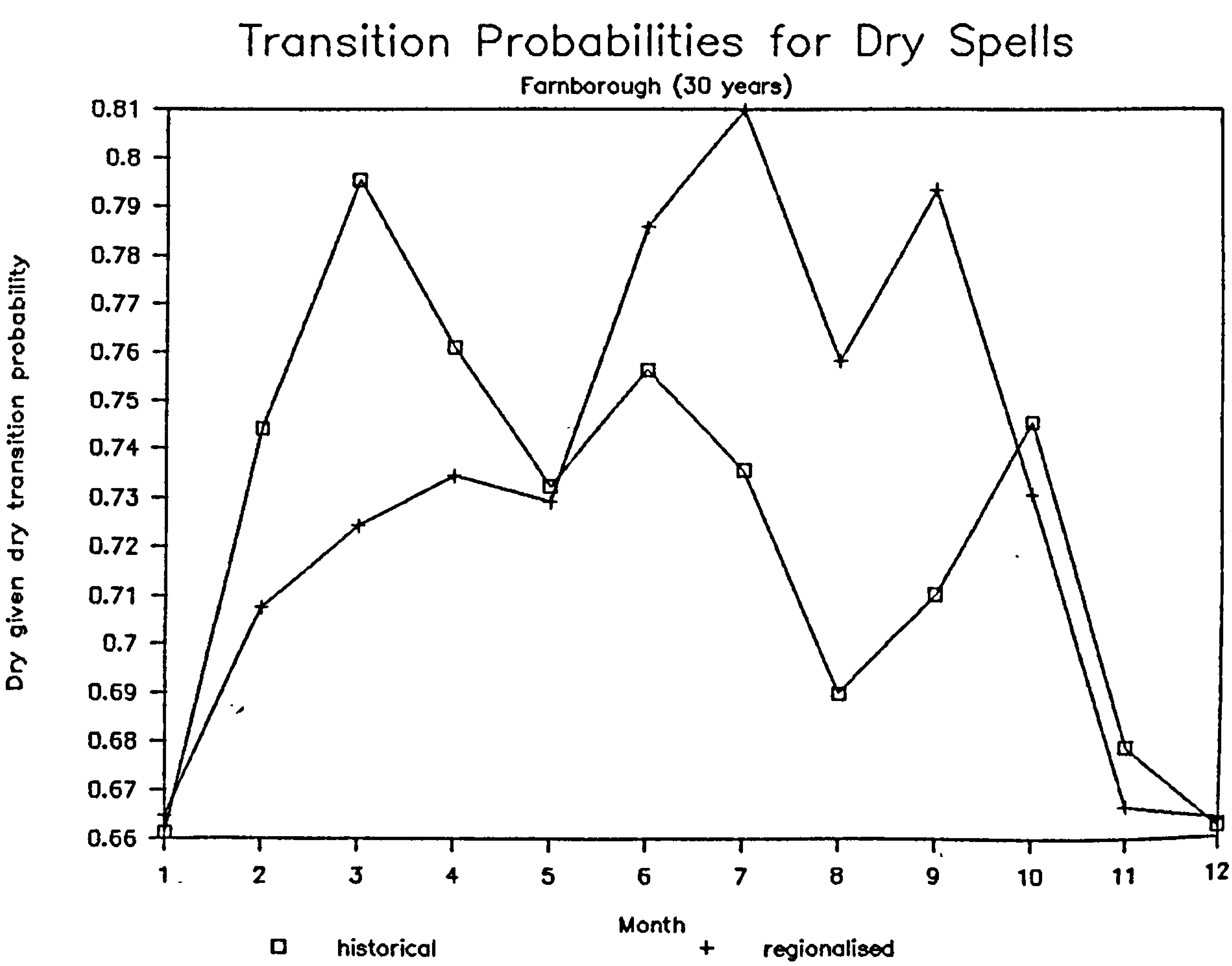
(G.42)


Comparison of Daily Variances
Farnborough (30 years)
Daily variance / sq mm
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

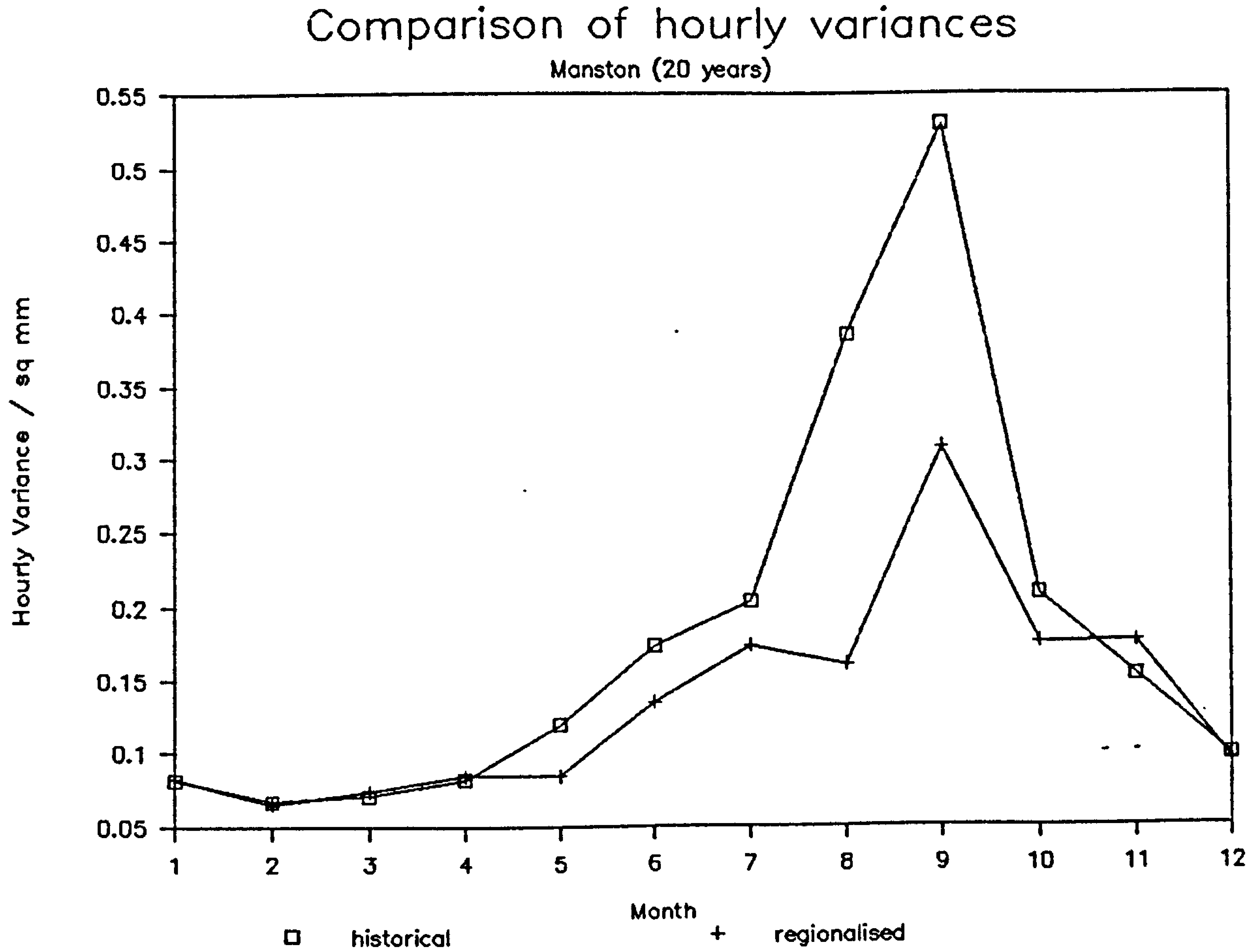
(G.43)


Comparison of Proportion of Dry Days
Farnborough (30 years)
Proportion of dry days
0.73
0.72
0.71
0.7
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0.61
0.6
0.59
0.58
0.57
0.56
0.55
0.54
0.53
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

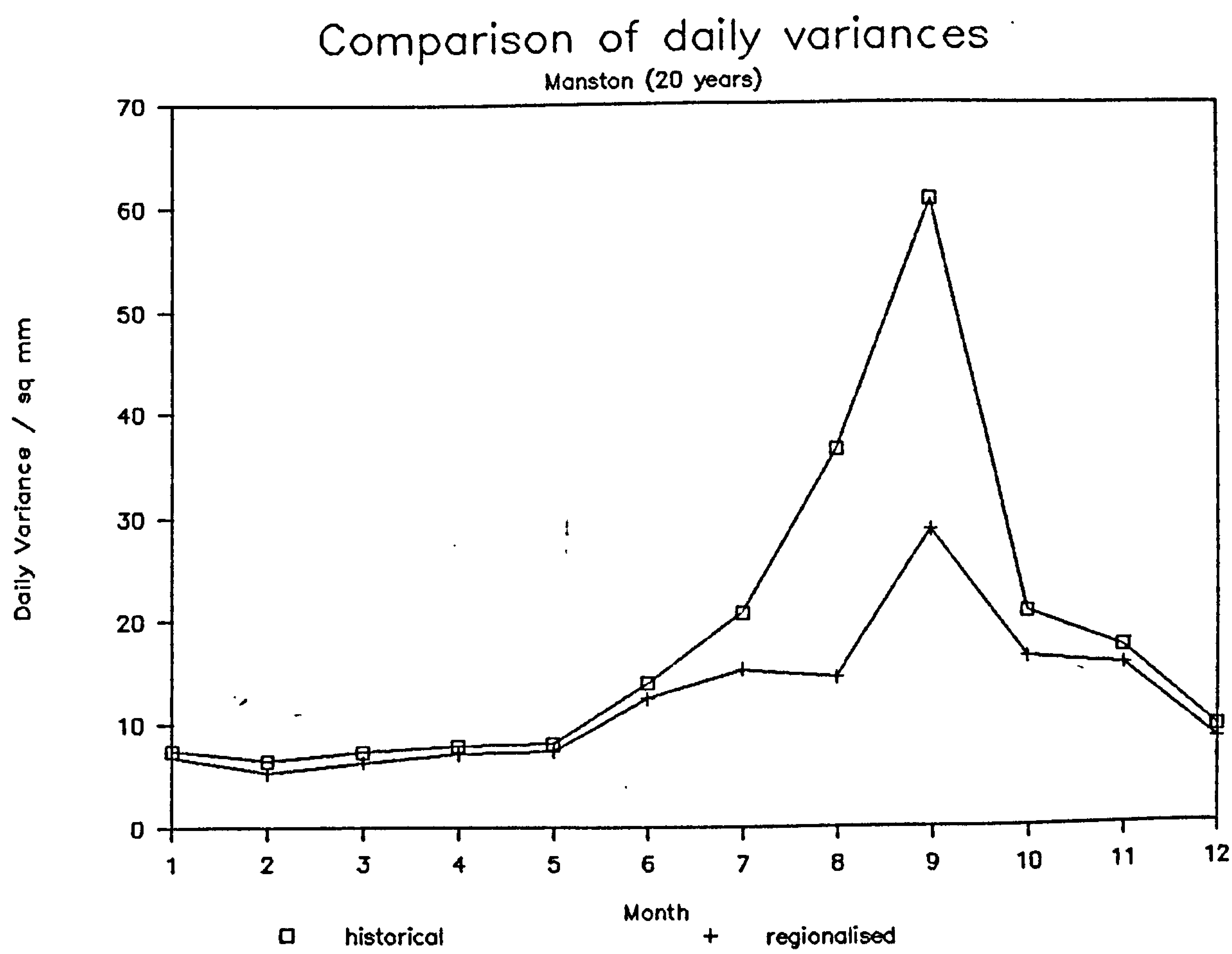
(G.44)


Transition Probabilities for Dry Spells
Farnborough (30 years)
Dry given dry transition probability
0.81
0.8
0.79
0.78
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.7
0.69
0.68
0.67
0.66
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

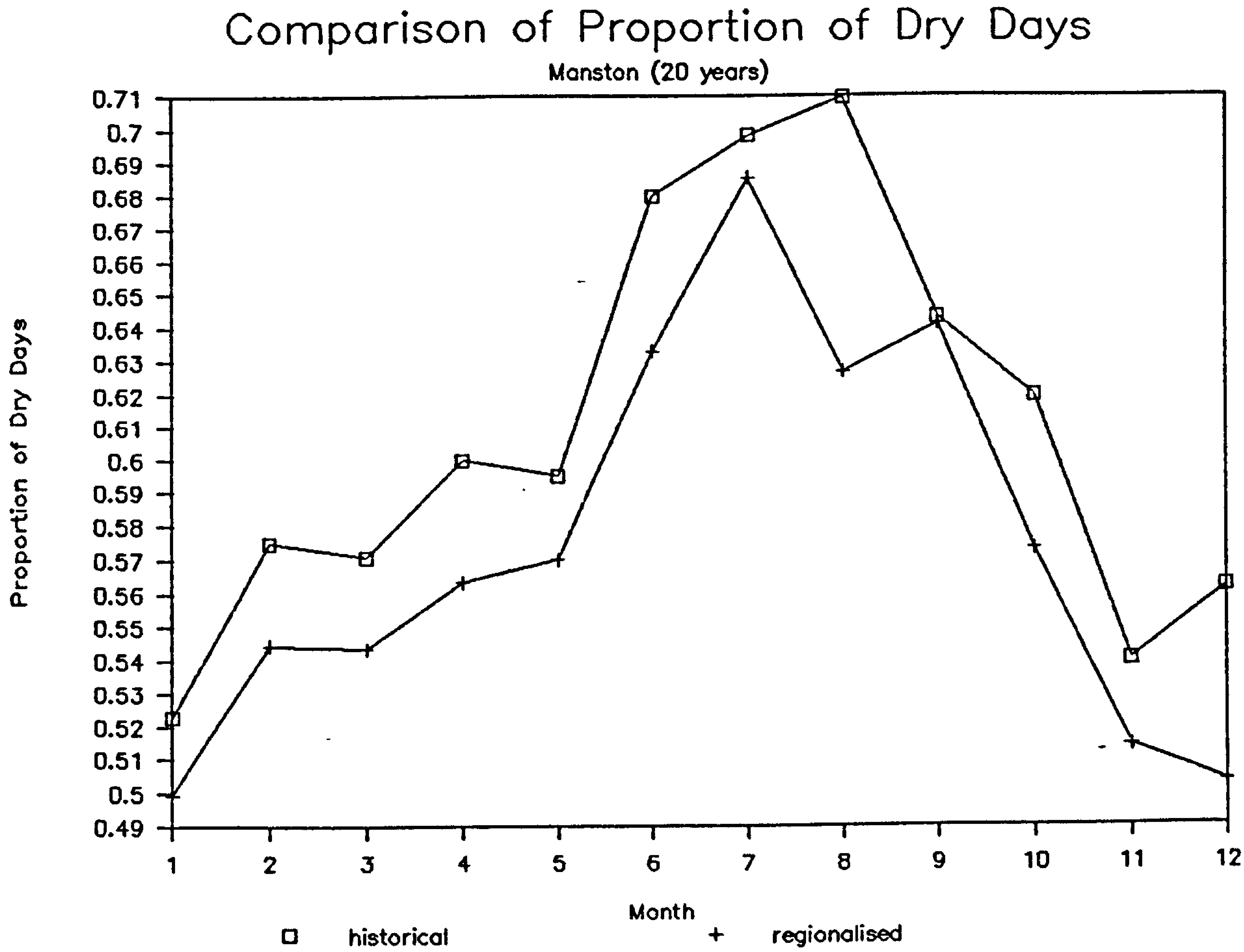
(G.45)


Comparison of hourly variances
Manston (20 years)
Hourly Variance / sq mm
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

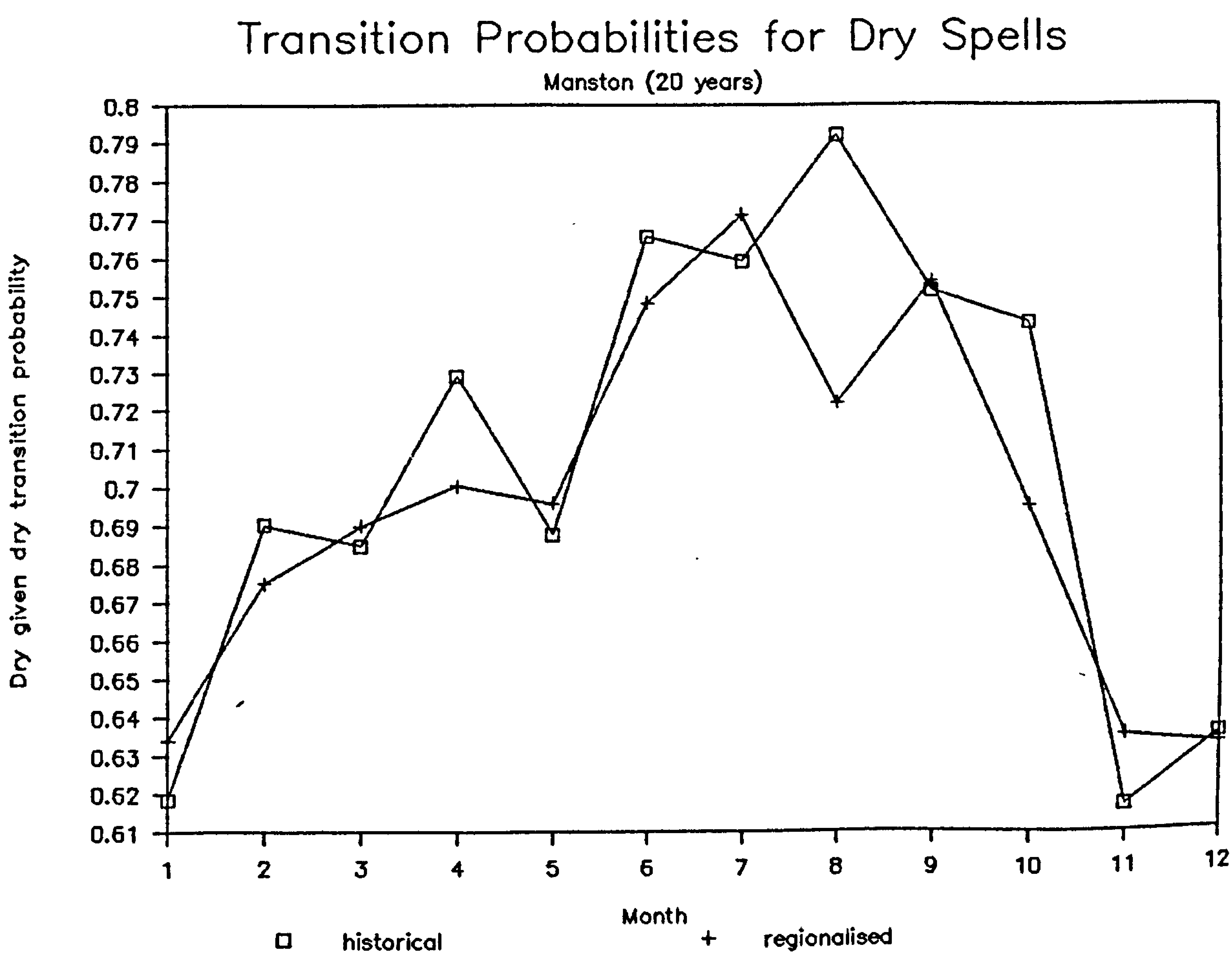
(G.46)


Comparison of daily variances
Manston (20 years)
Daily Variance / sq mm
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

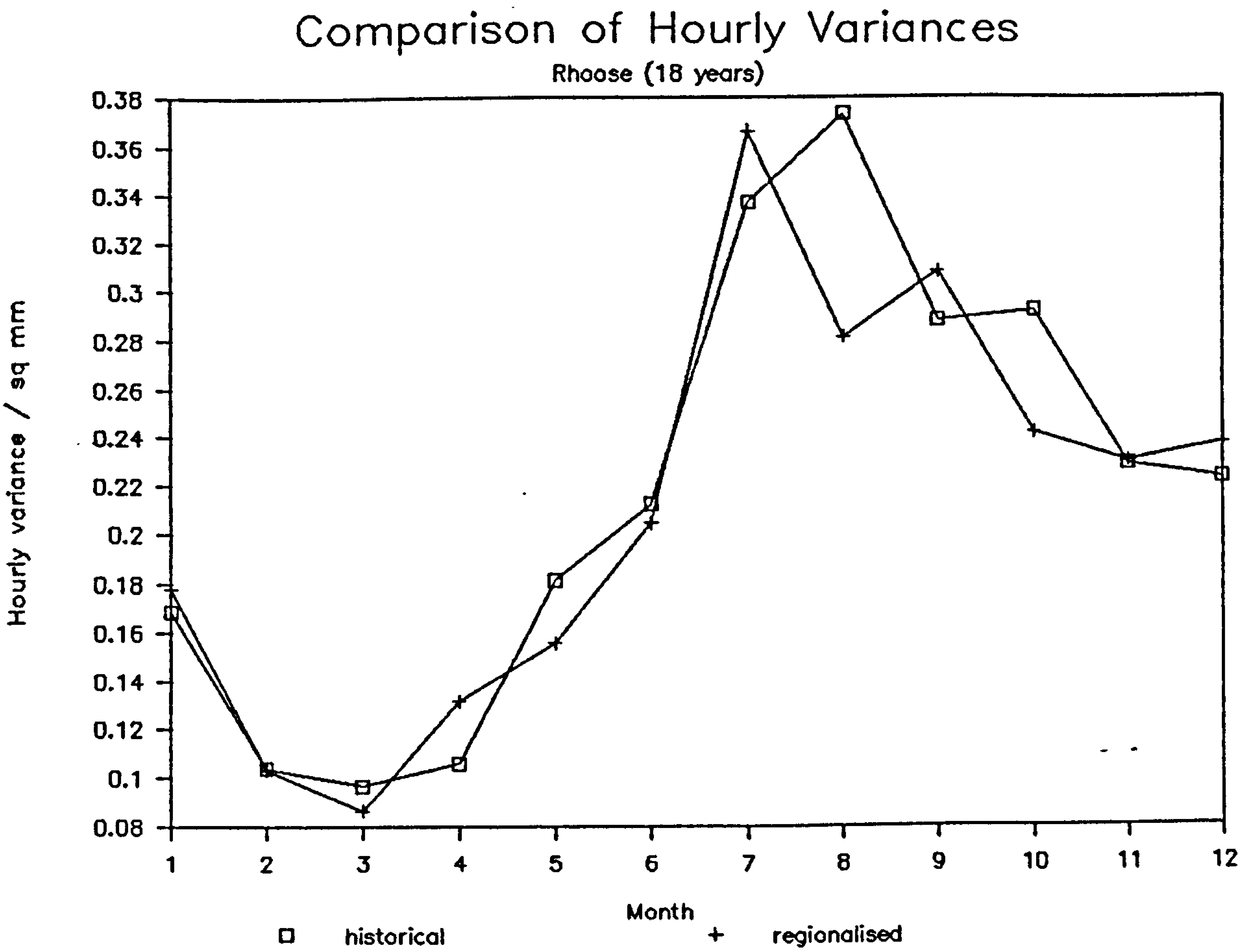
(G.47)


Comparison of Proportion of Dry Days
Manston (20 years)
Proportion of Dry Days
0.71
0.7
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0.61
0.6
0.59
0.58
0.57
0.56
0.55
0.54
0.53
0.52
0.51
0.5
0.49
1 2 3 4 5 6 7 8 9 10 11 12
Month
historical
regionalised

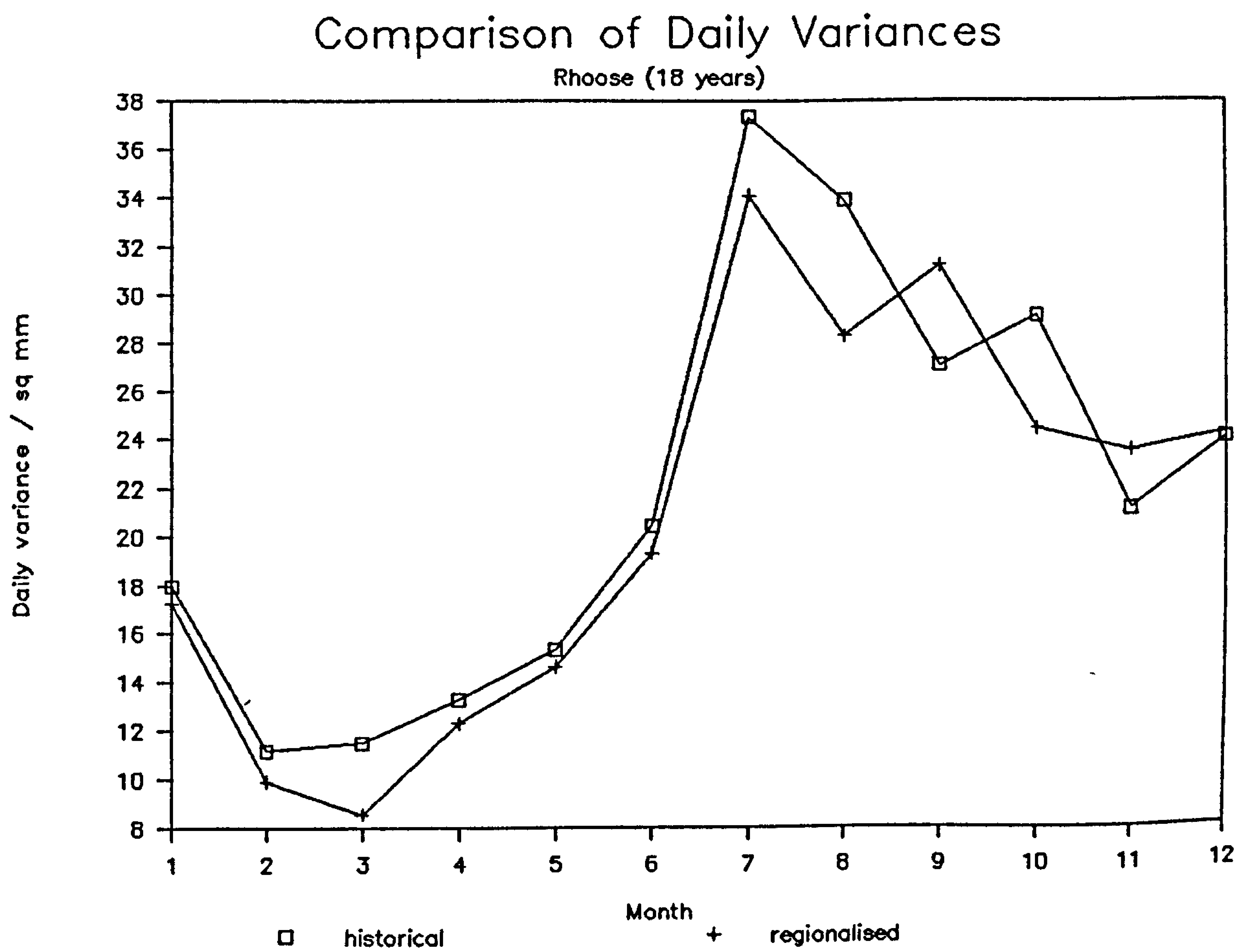
(G.48)


Transition Probabilities for Dry Spells
Manston (20 years)
Dry given dry transition probability
0.8
0.79
0.78
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.7
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0.61
1 2 3 4 5 6 7 8 9 10 11 12
Month
historical
regionalised

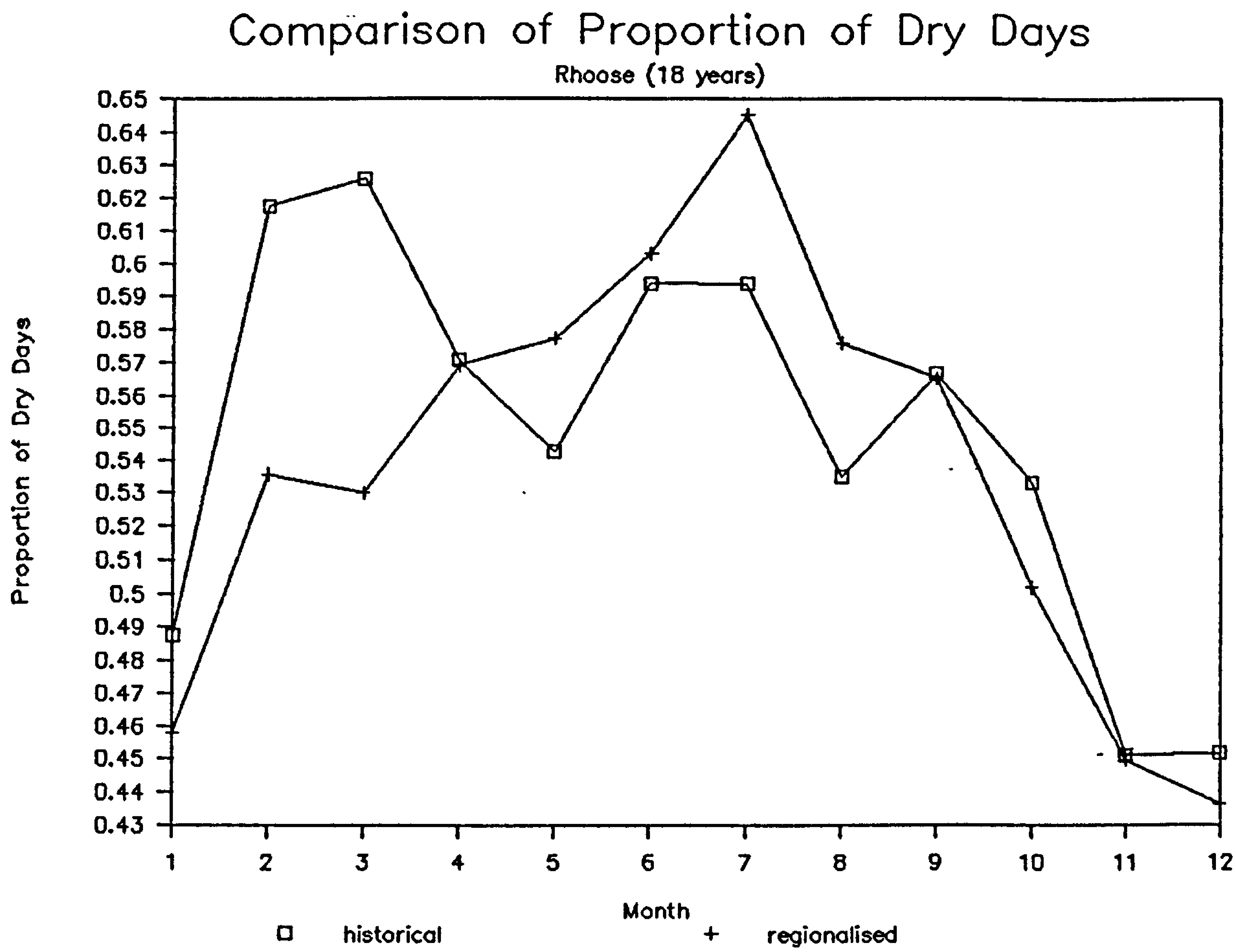
(G.49)


Comparison of Hourly Variances
Rhoose (18 years)
Hourly variance / sq mm
0.38
0.36
0.34
0.32
0.3
0.28
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
0.08
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

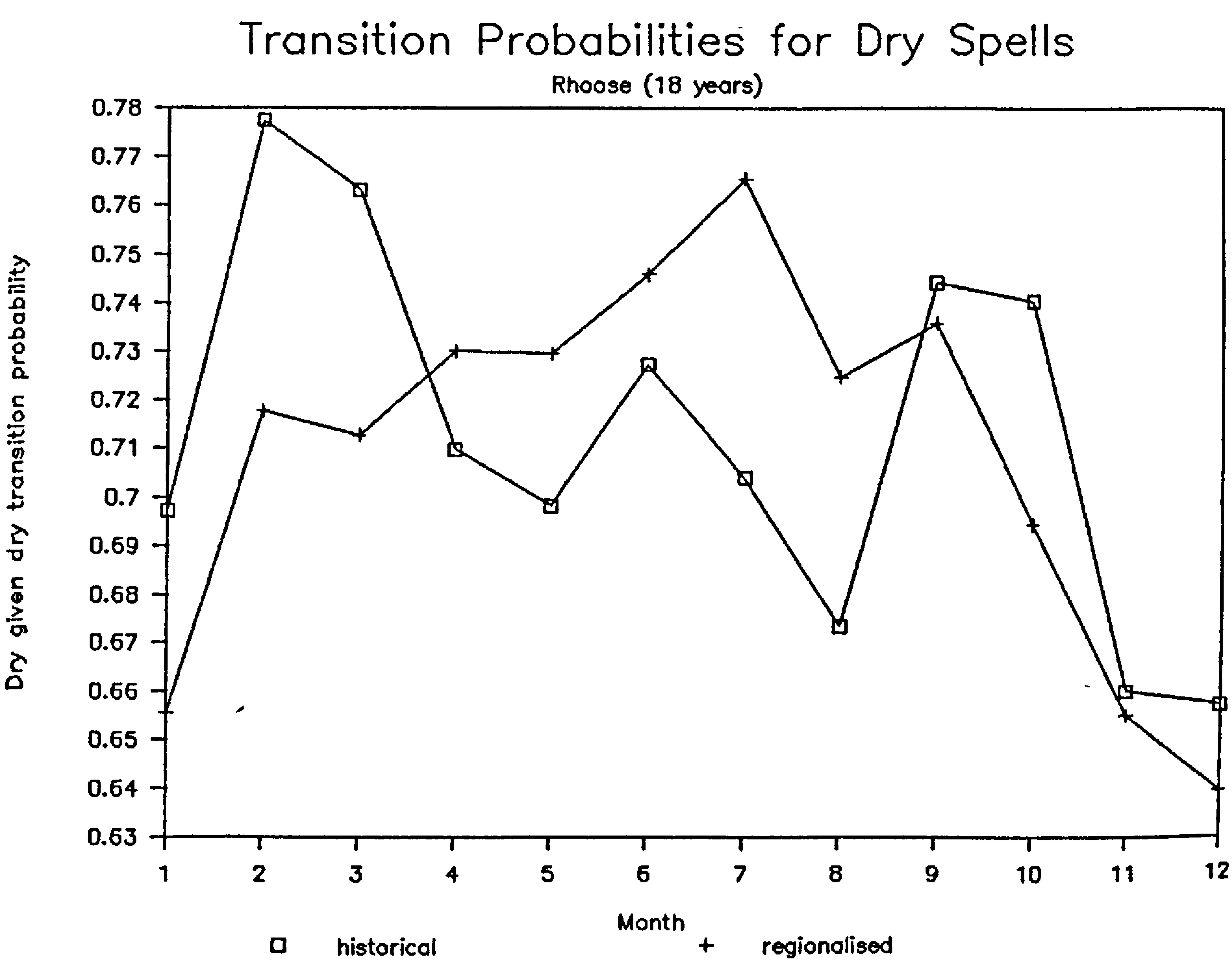
(G.50)


Comparison of Daily Variances
Rhoose (18 years)
Daily variance / sq mm
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

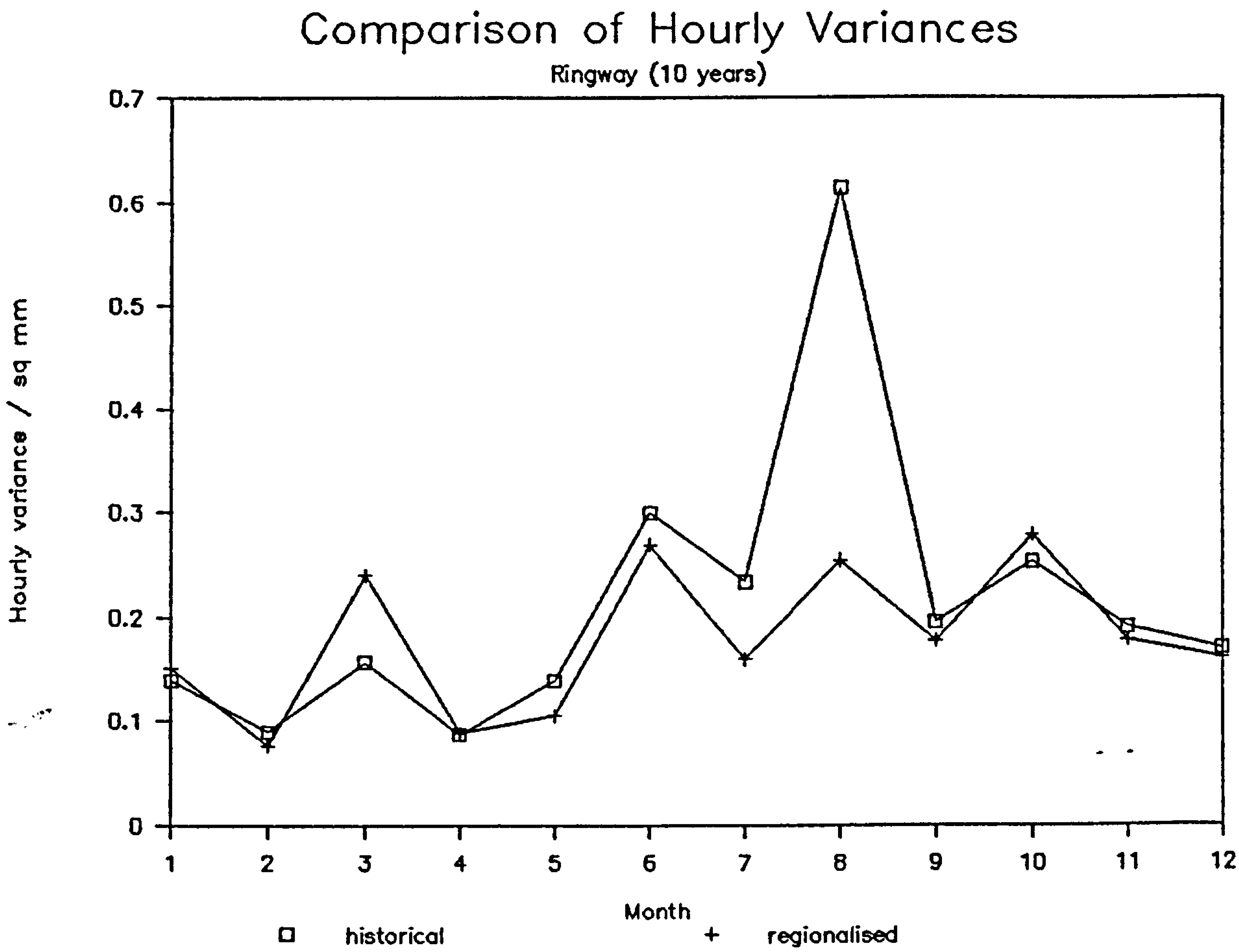
(G.51)


Comparison of Proportion of Dry Days
Rhoose (18 years)
Proportion of Dry Days
0.65
0.64
0.63
0.62
0.61
0.6
0.59
0.58
0.57
0.56
0.55
0.54
0.53
0.52
0.51
0.5
0.49
0.48
0.47
0.46
0.45
0.44
0.43
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

(G.52)

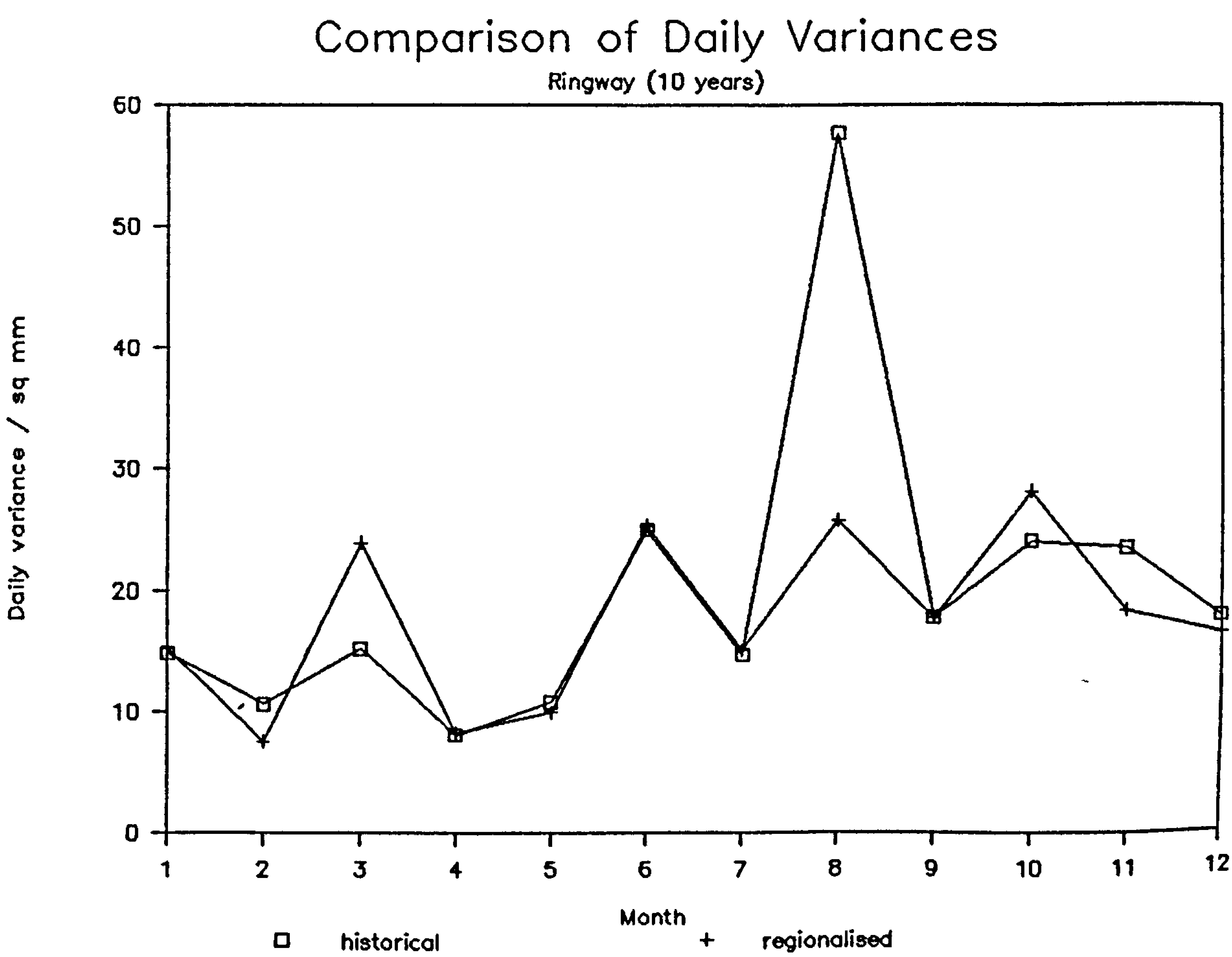

Transition Probabilities for Dry Spells
Rhoose (18 years)
Dry given dry transition probability
0.78
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.7
0.69
0.68
0.67
0.66
0.65
0.64
0.63
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

(G.53)



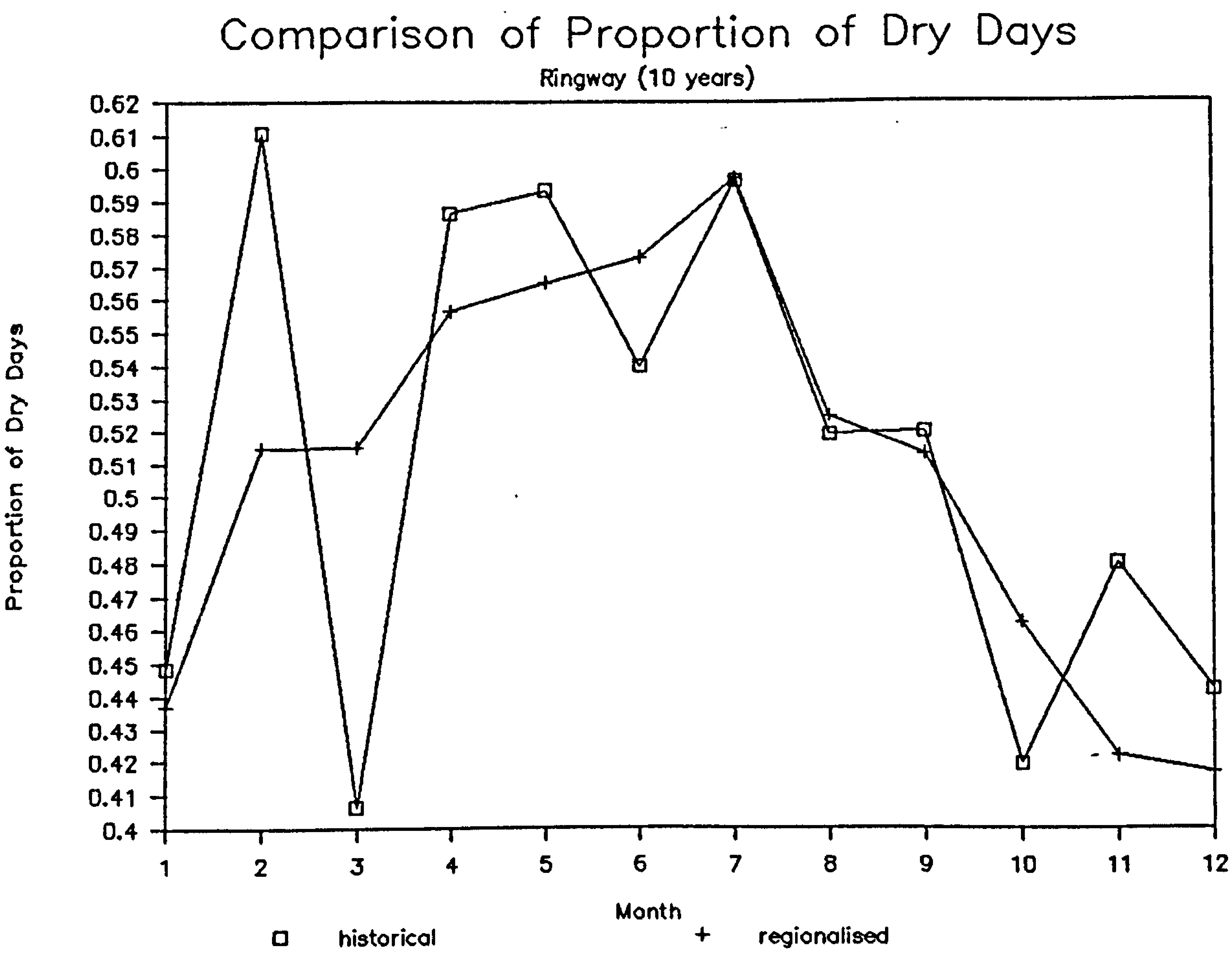
Comparison of Hourly Variances
Ringway (10 years)
Hourly variance / sq mm
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

(G.54)



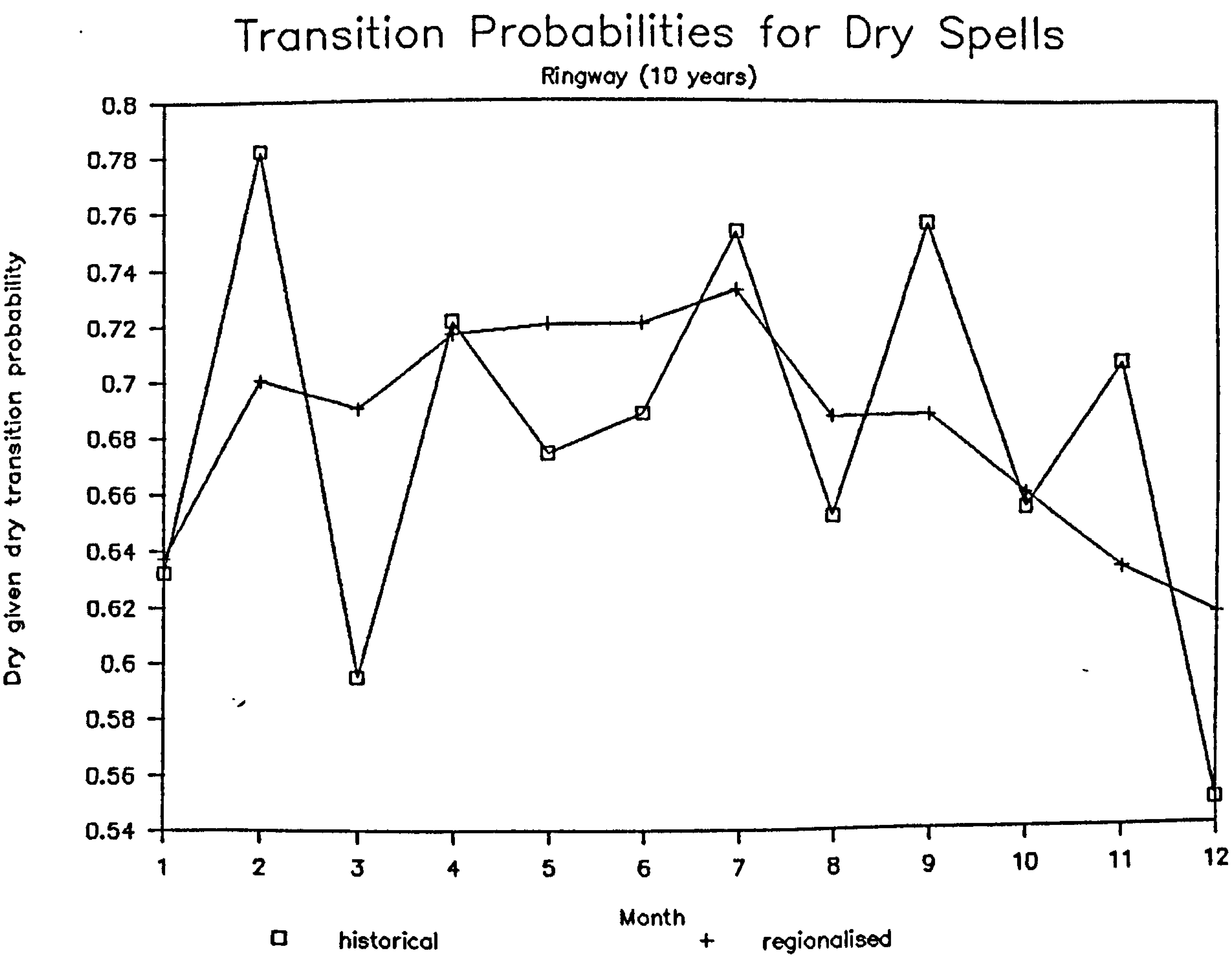
Comparison of Daily Variances
Ringway (10 years)
Daily variance / sq mm
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

(G.55)



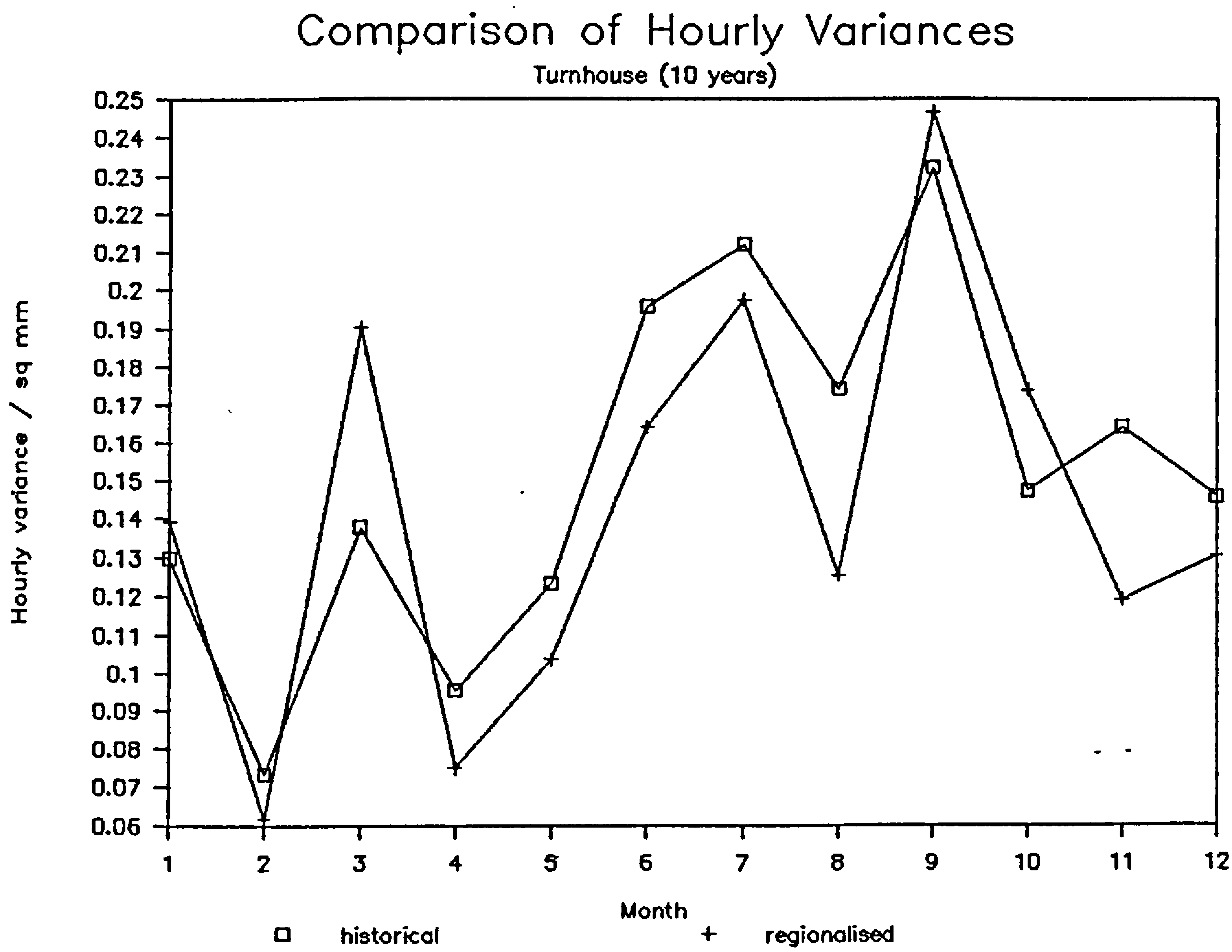
Comparison of Proportion of Dry Days
Ringway (10 years)
Proportion of Dry Days
0.62
0.61
0.6
0.59
0.58
0.57
0.56
0.55
0.54
0.53
0.52
0.51
0.5
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0.42
0.41
0.4
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

(G.56)

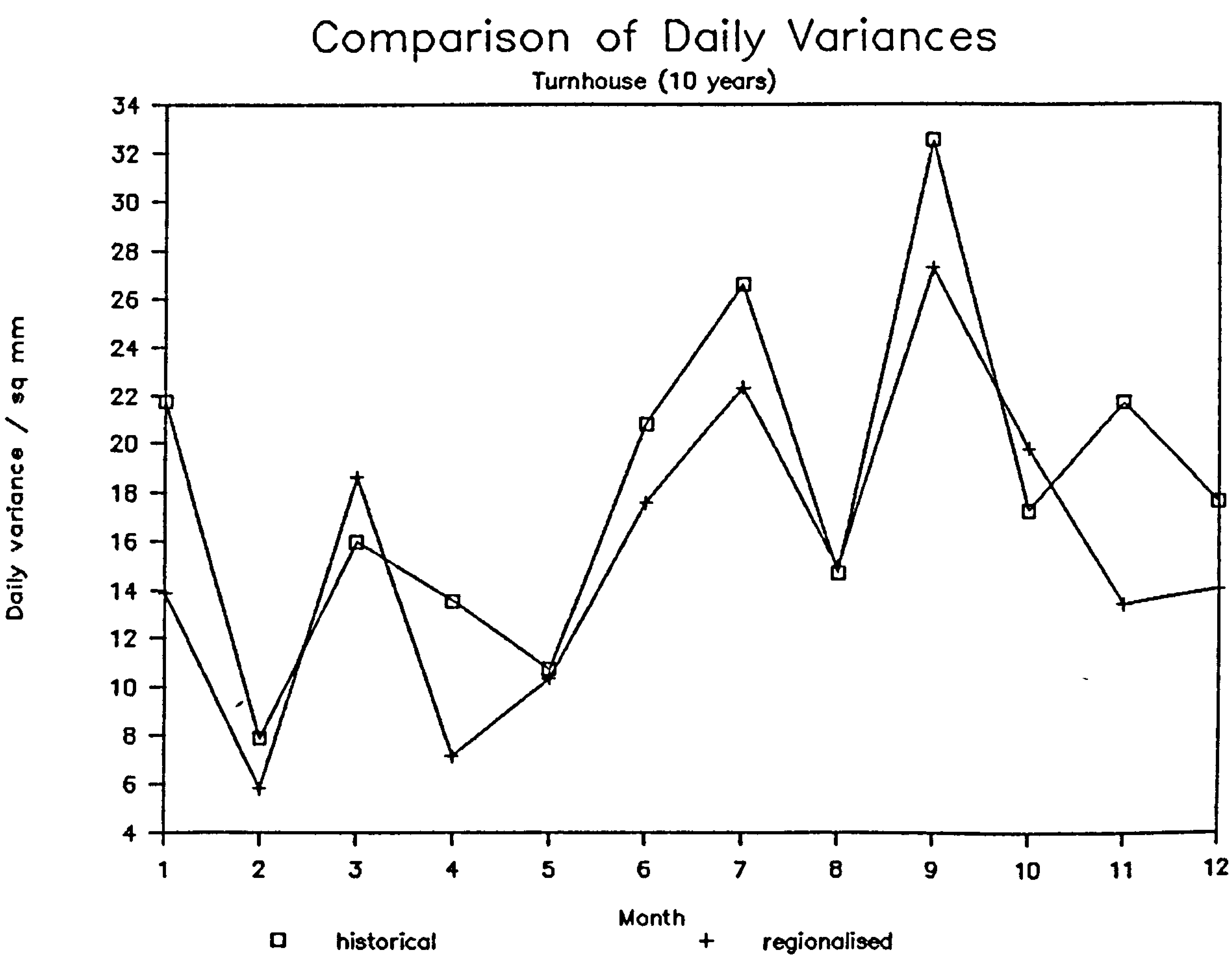


Transition Probabilities for Dry Spells
Ringway (10 years)
Dry given dry transition probability
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
0.56
0.54
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

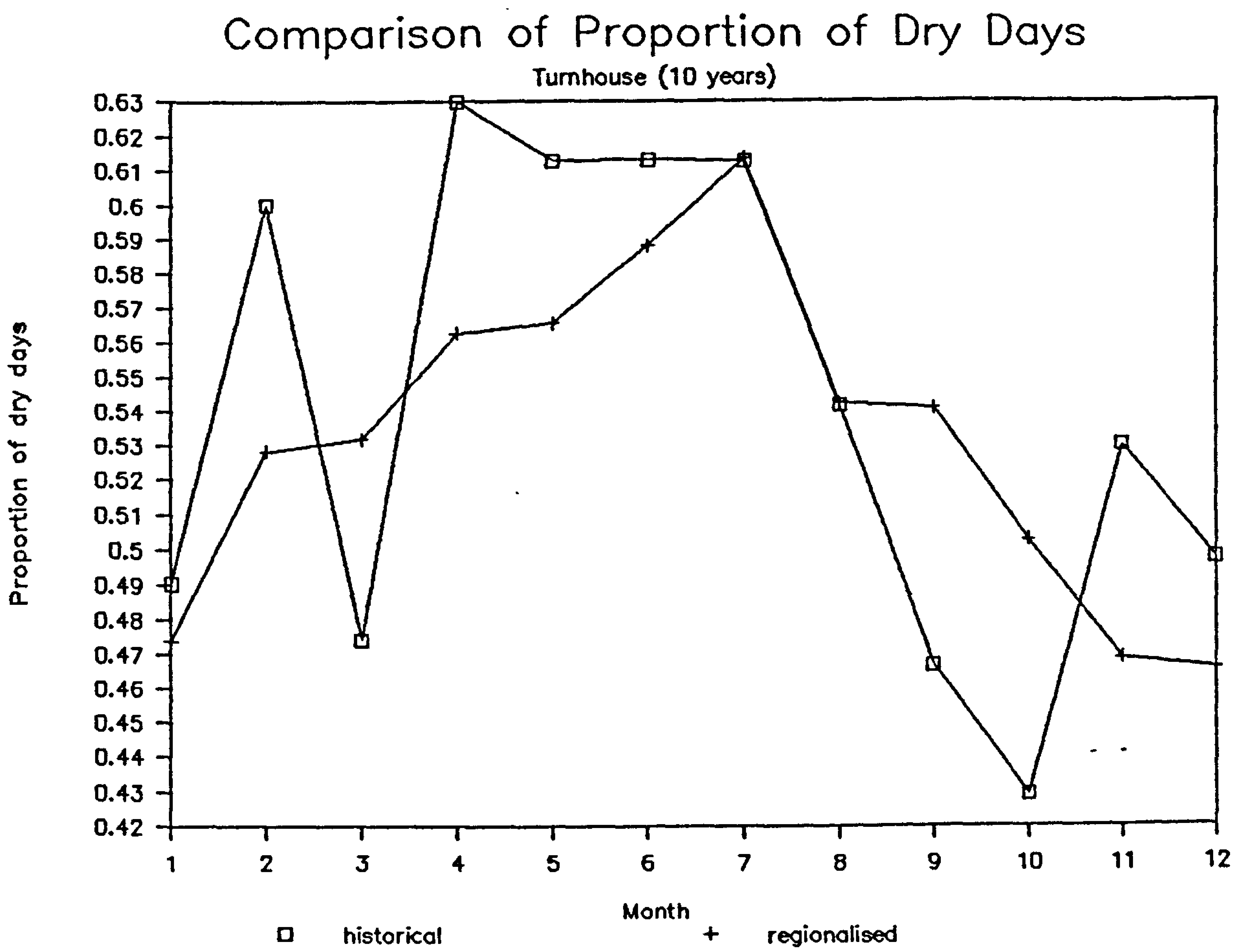
(G.57)


Comparison of Hourly Variances
Turnhouse (10 years)
Hourly variance / sq mm
Month
historical
regionalised

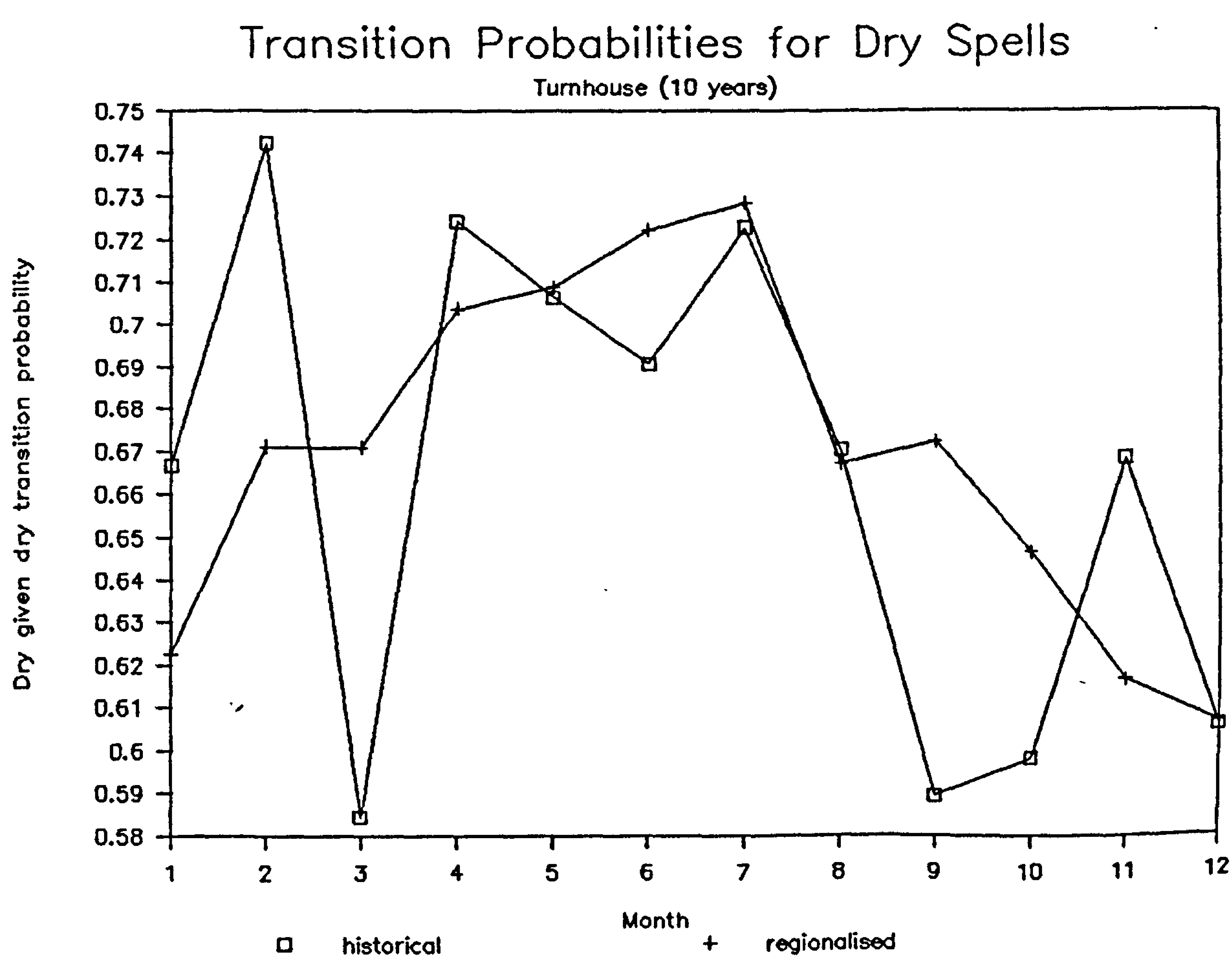
(G.58)


Comparison of Daily Variances
Turnhouse (10 years)
Daily variance / sq mm
Month
historical
regionalised

(G.59)


Comparison of Proportion of Dry Days
Turnhouse (10 years)
Proportion of dry days
0.63
0.62
0.61
0.6
0.59
0.58
0.57
0.56
0.55
0.54
0.53
0.52
0.51
0.5
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0.42
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

(G.60)


Transition Probabilities for Dry Spells
Turnhouse (10 years)
Dry given dry transition probability
0.75
0.74
0.73
0.72
0.71
0.7
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0.61
0.6
0.59
0.58
1
2
3
4
5
6
7
8
9
10
11
12
Month
historical
regionalised

(G.61)

APPENDIX H: TREATMENT OF MISSING VALUES

Rainfall data almost always contains missing values. As the data set was large (a total of 120 rainfall stations) and many of the computer programs written for the project had to be used on all the stations, it was found convenient to keep the number of years of data for a station the same for each month.

For the hourly/minutely data; if there were many missing values for a month, the data for that month were deleted from the station record. Consequently the record length (in years) would be reduced by one for that station. If there were many missing values for another month, different from the previous month, then the data for that month were replaced by data from the same month of the deleted year. If only a few data were missing in a month, then these values were taken as zero. This approach was adopted (as opposed to taking an average) because the missing values tended to occur in sequences (rather than isolated values). The approach had no effect on the development of the fitting procedure because the Manston data set had no missing values (p66). Furthermore, the effect of this approach on the regionalisation procedure could be neglected because less than 1%[1] of the hourly data were replaced by zeros, and most (about 80%) of the data used in the regionalisation procedure came from daily rainfall stations.

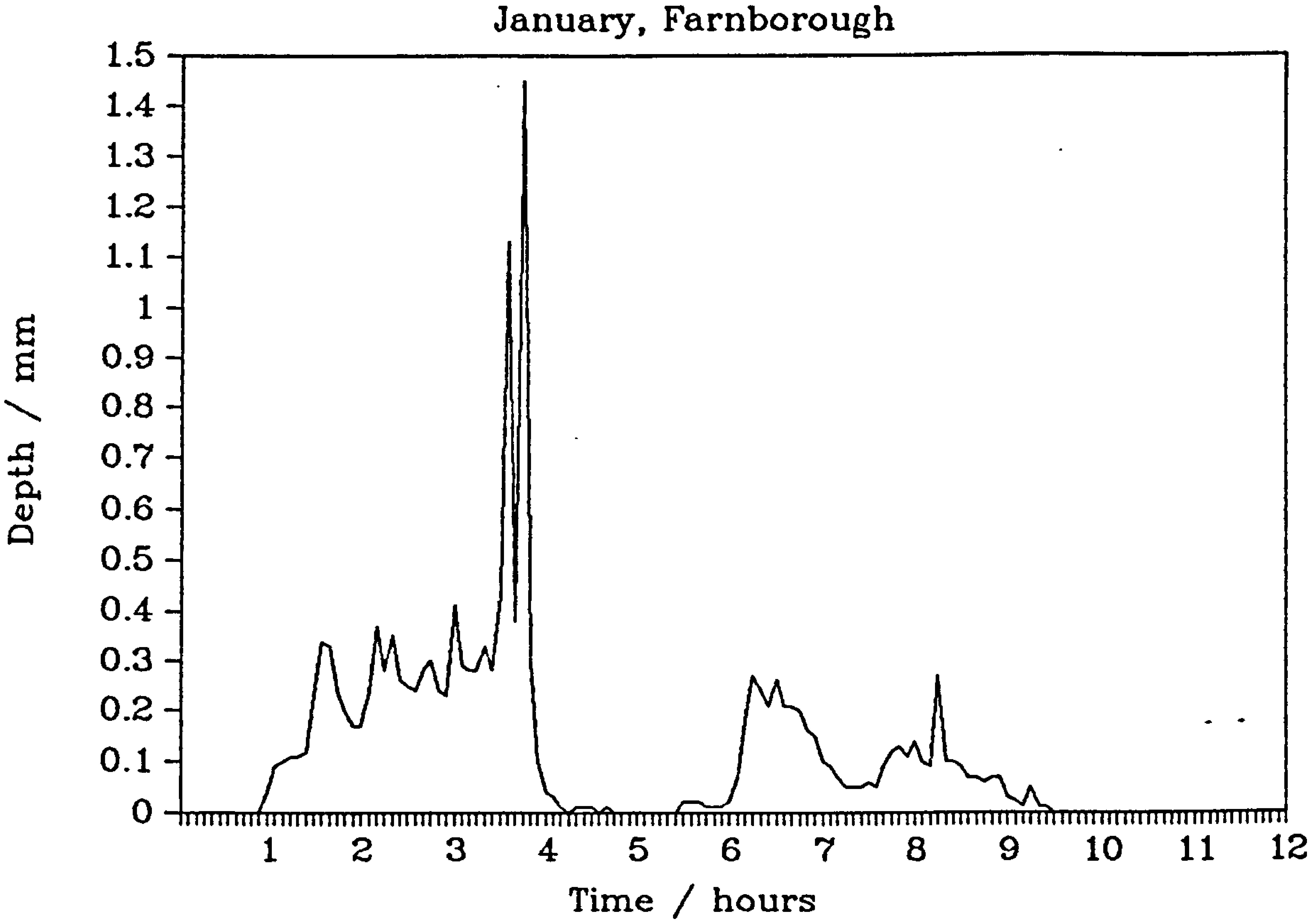
For the daily data; if, for one of the years, the data for a month were missing, the whole data set was scanned to find the same month (in a different year) with a monthly total close to the month with missing values, and the data from this month were then used to replace the missing data. The data for the year would be deleted (i.e. the record length reduced by one) if there were many months in the year with missing values. If only a few data were missing in a month then these values were taken as zero. In addition, traces were also taken as zero. N.B. less than 0.01%[2] of the missing daily values were replaced by zero, so that the effect of this approach on the regionalisation procedure could be neglected.

Finally, for programming convenience, the data for leap years were ignored (so that February always had 28 days).

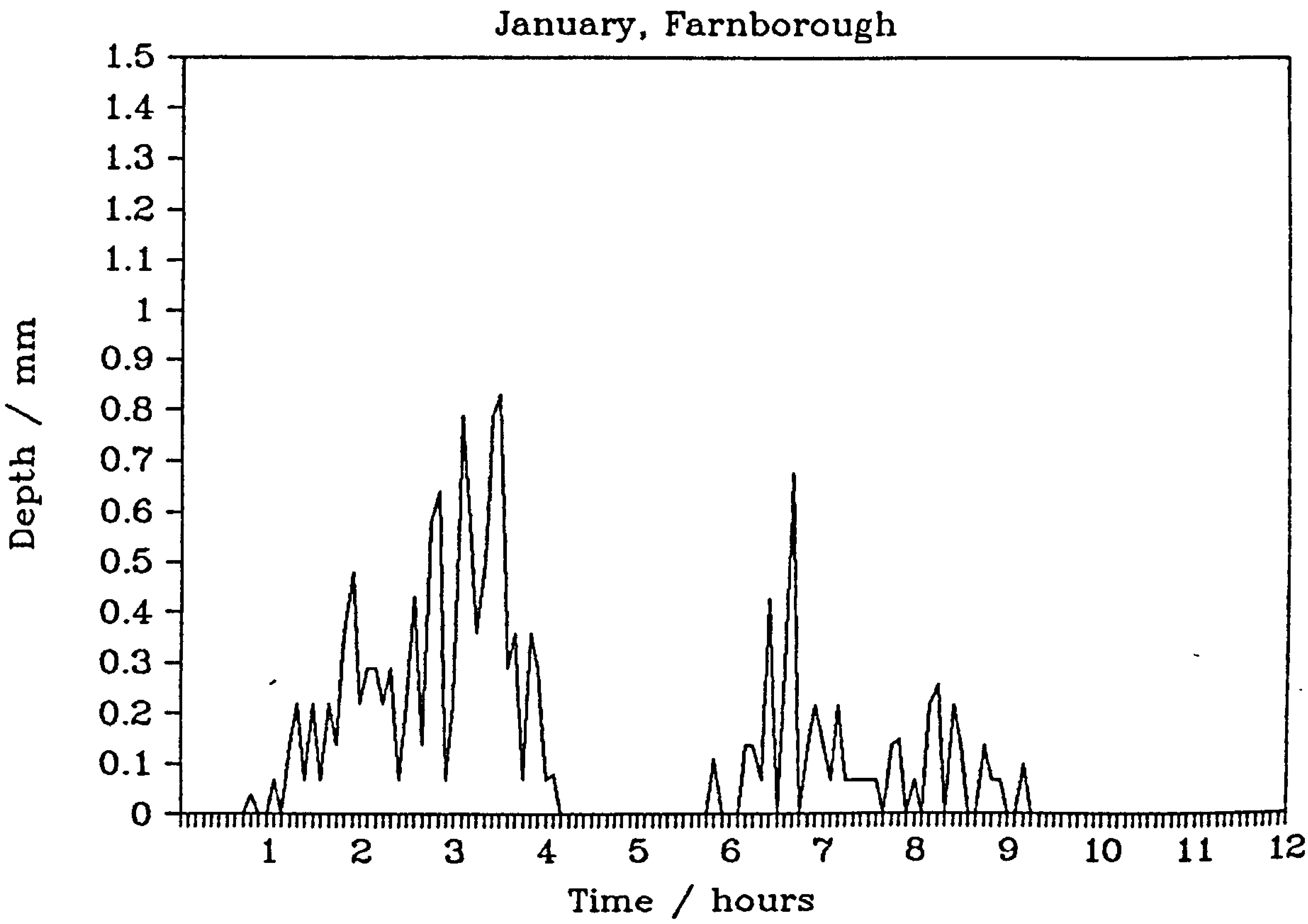
[1]*This estimate was obtained by selecting a station-month (Blackpool-January) that had a large number of missing values, when compared with other station-months.*

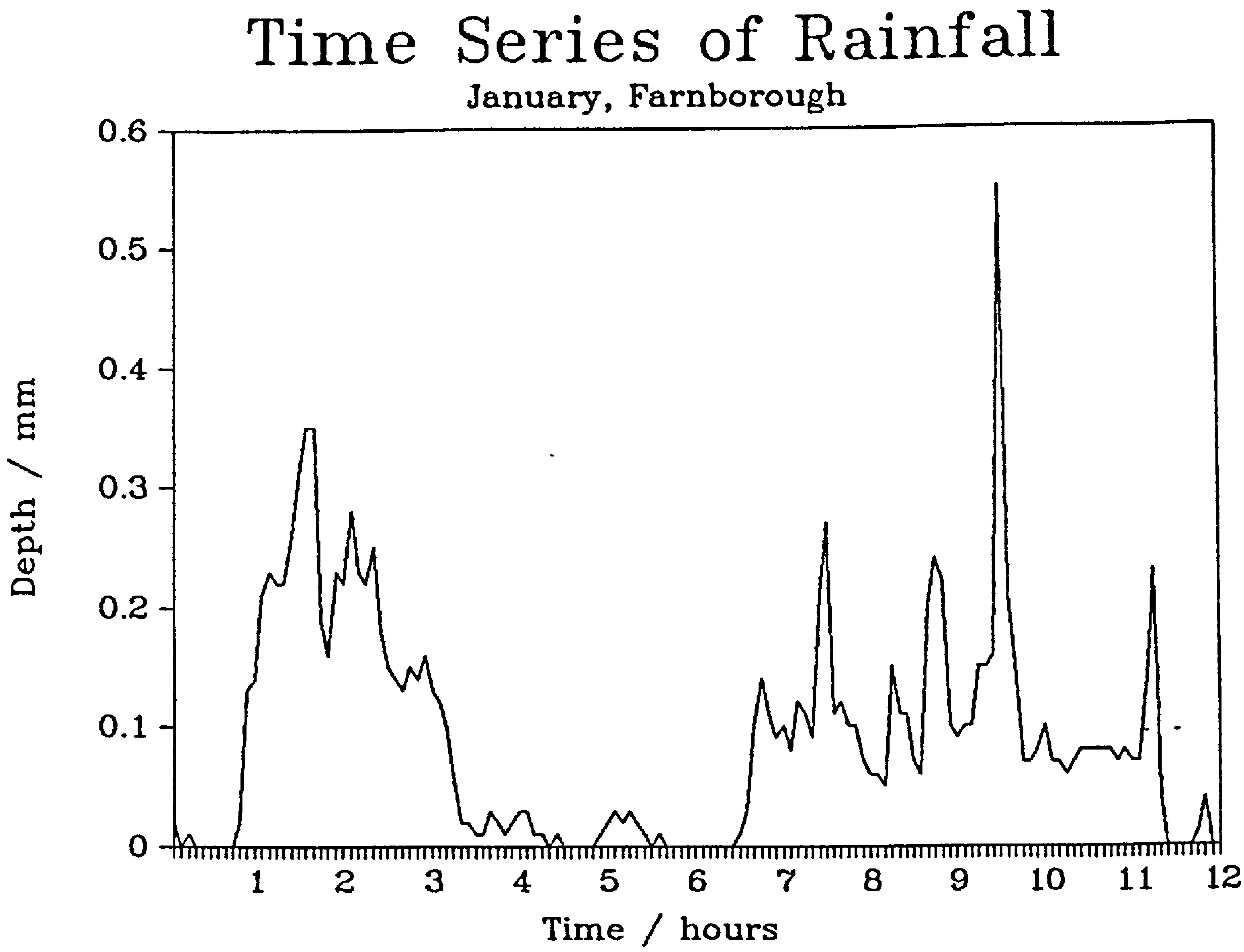
[2]*This estimate of 0.01% was obtained by taking a random sample of 20 daily stations and counting the number of times missing values were replaced by zero. This happened on 13 occasions, which gives a percentage of 100.13/25.20.365 = 0.007%, which was rounded to 0.01%.*

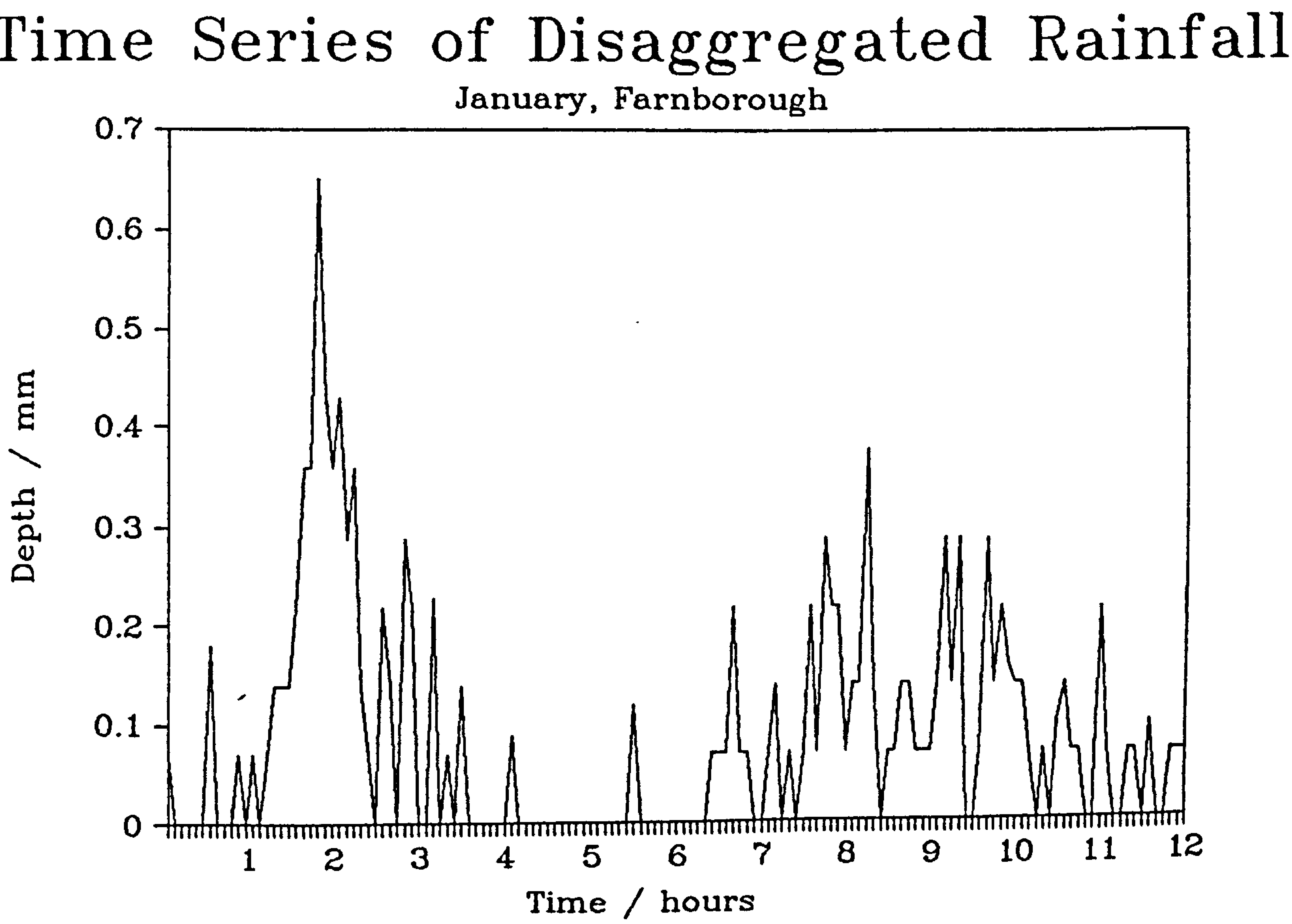
Time Series of Rainfall



Time Series of Disaggregated Rainfall

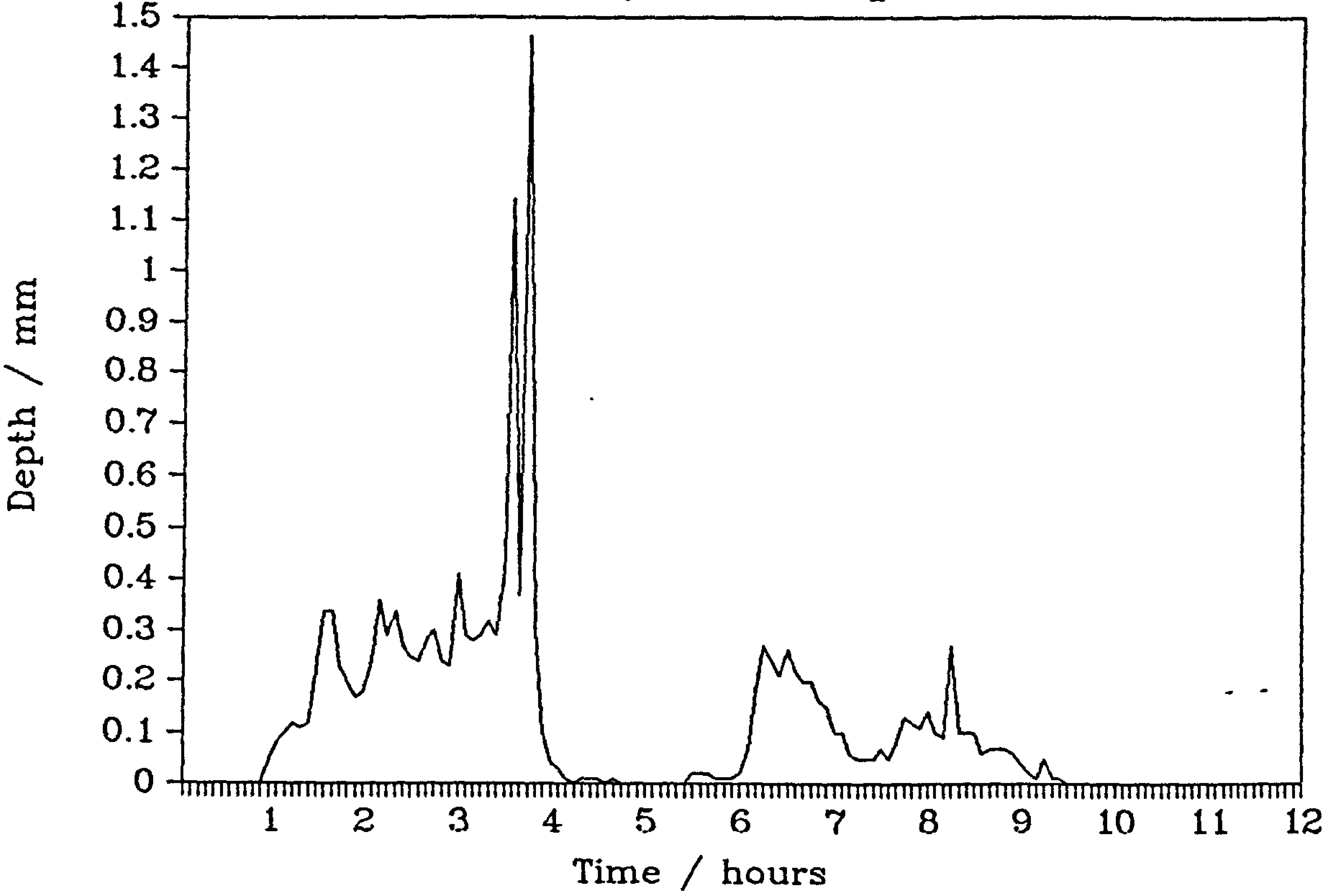



Time Series of Rainfall
January, Farnborough
Depth / mm
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
11
12
Time / hours


Time Series of Disaggregated Rainfall
January, Farnborough
Depth / mm
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
11
12
Time / hours

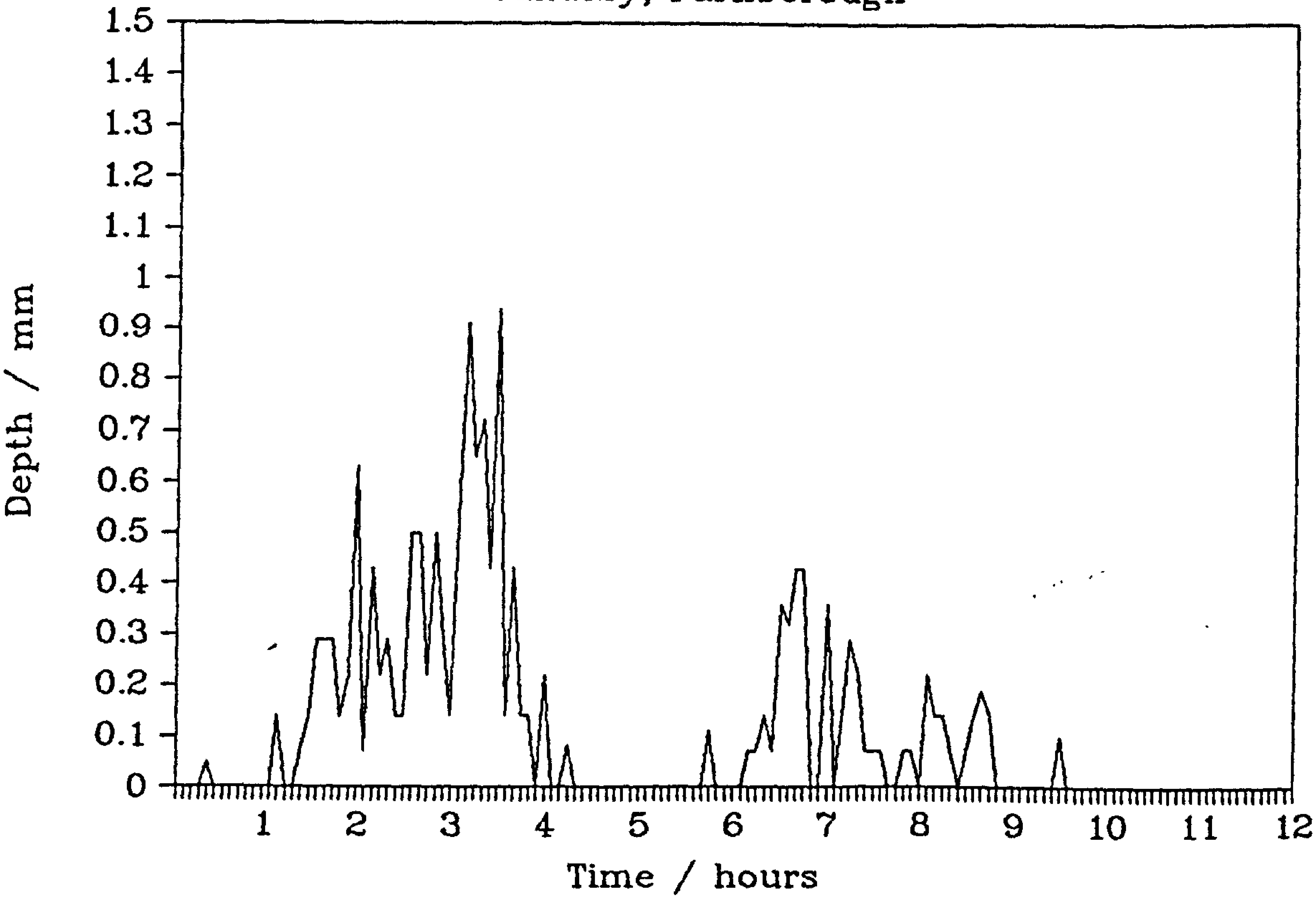
Time Series of Rainfall

January, Farnborough



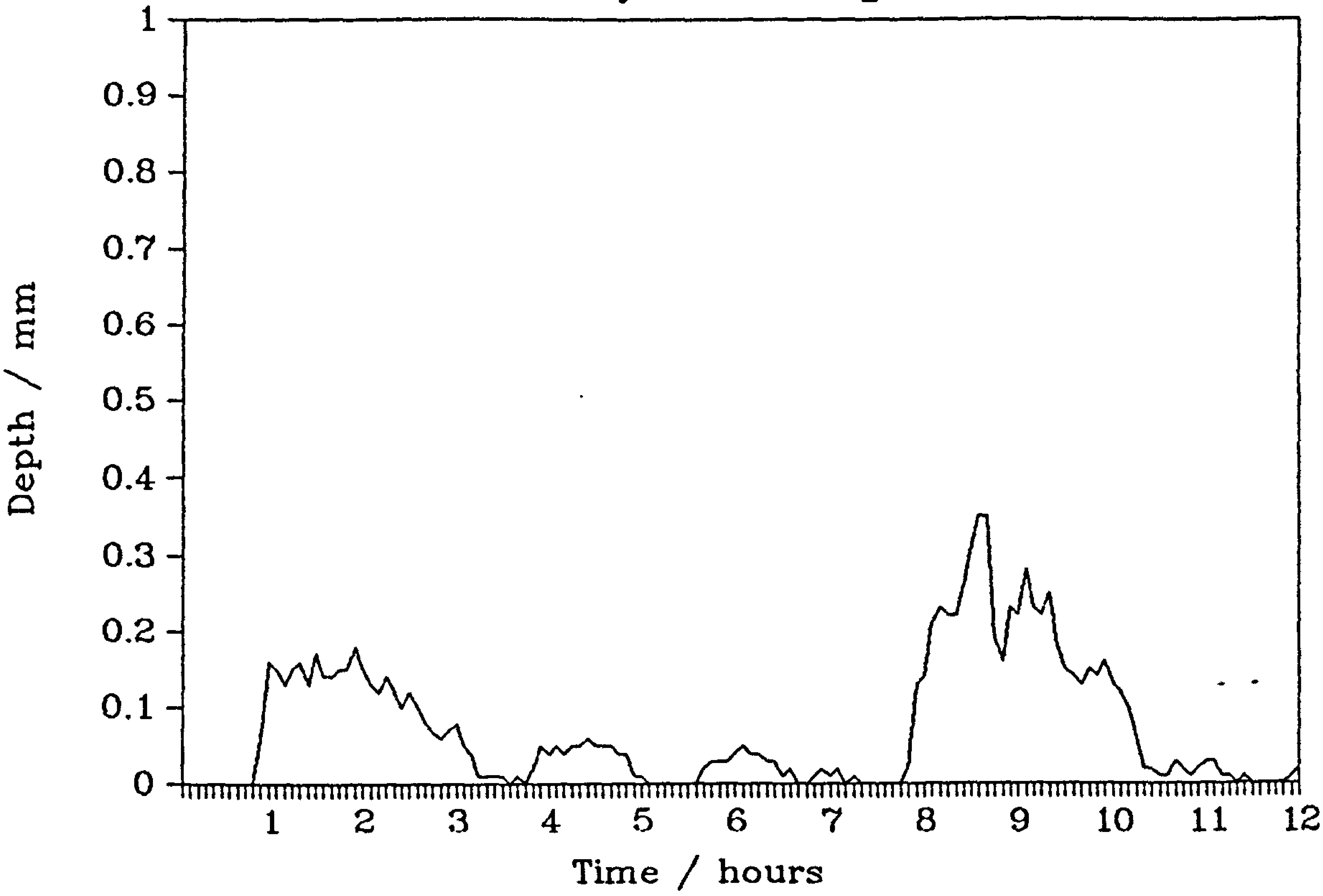
Time Series of Disaggregated Rainfall

January, Farnborough



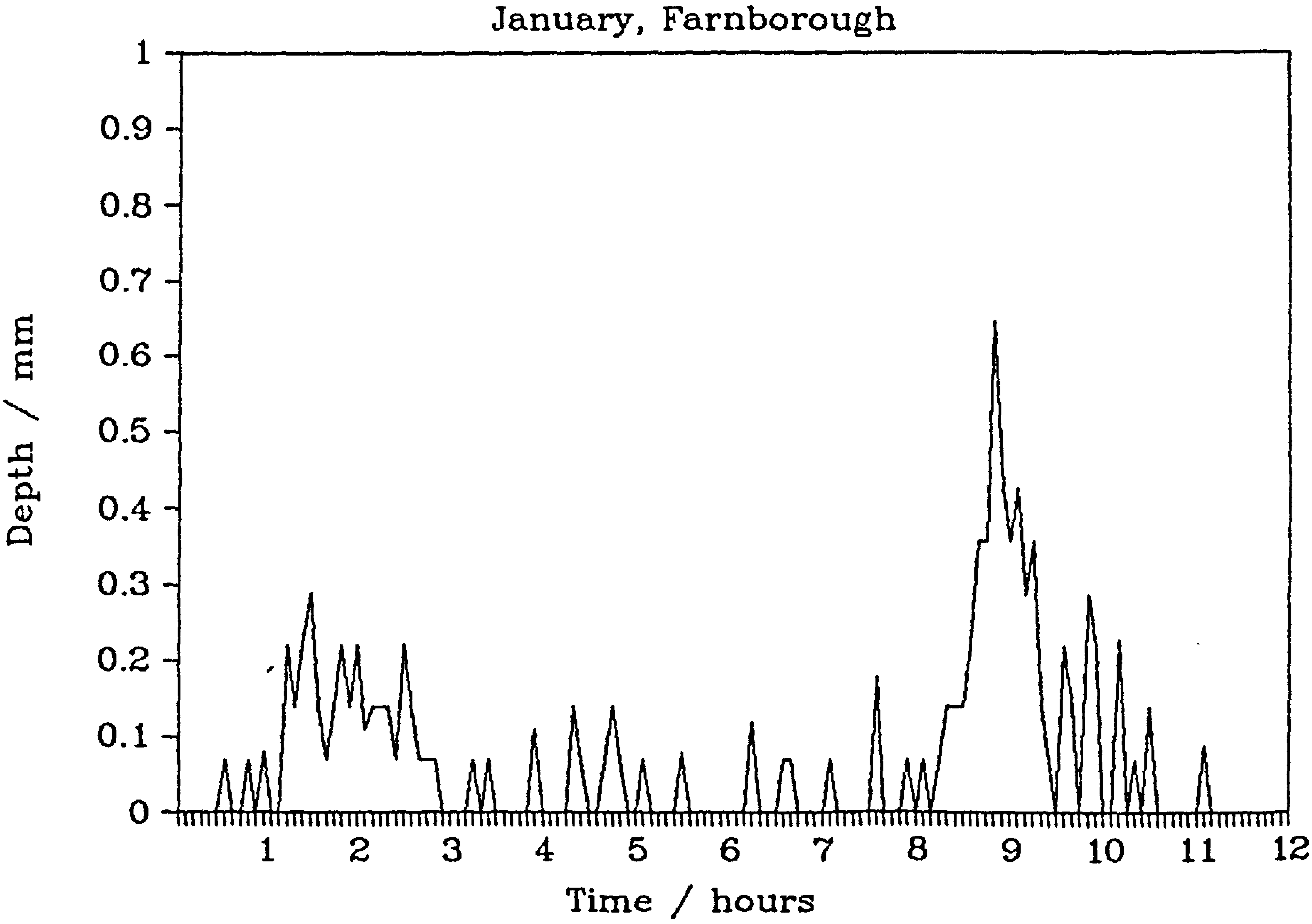
Time Series of Rainfall

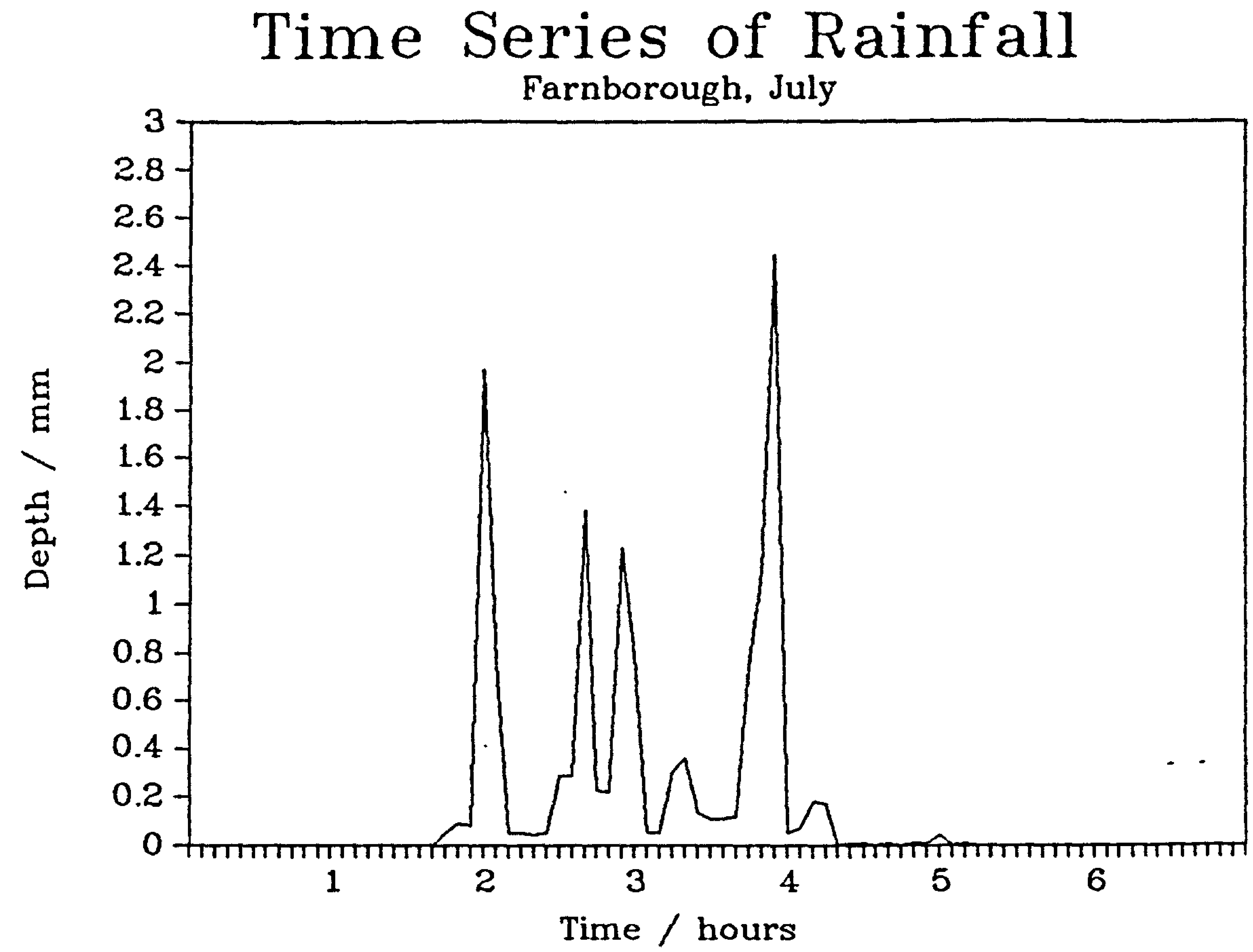
January, Farnborough

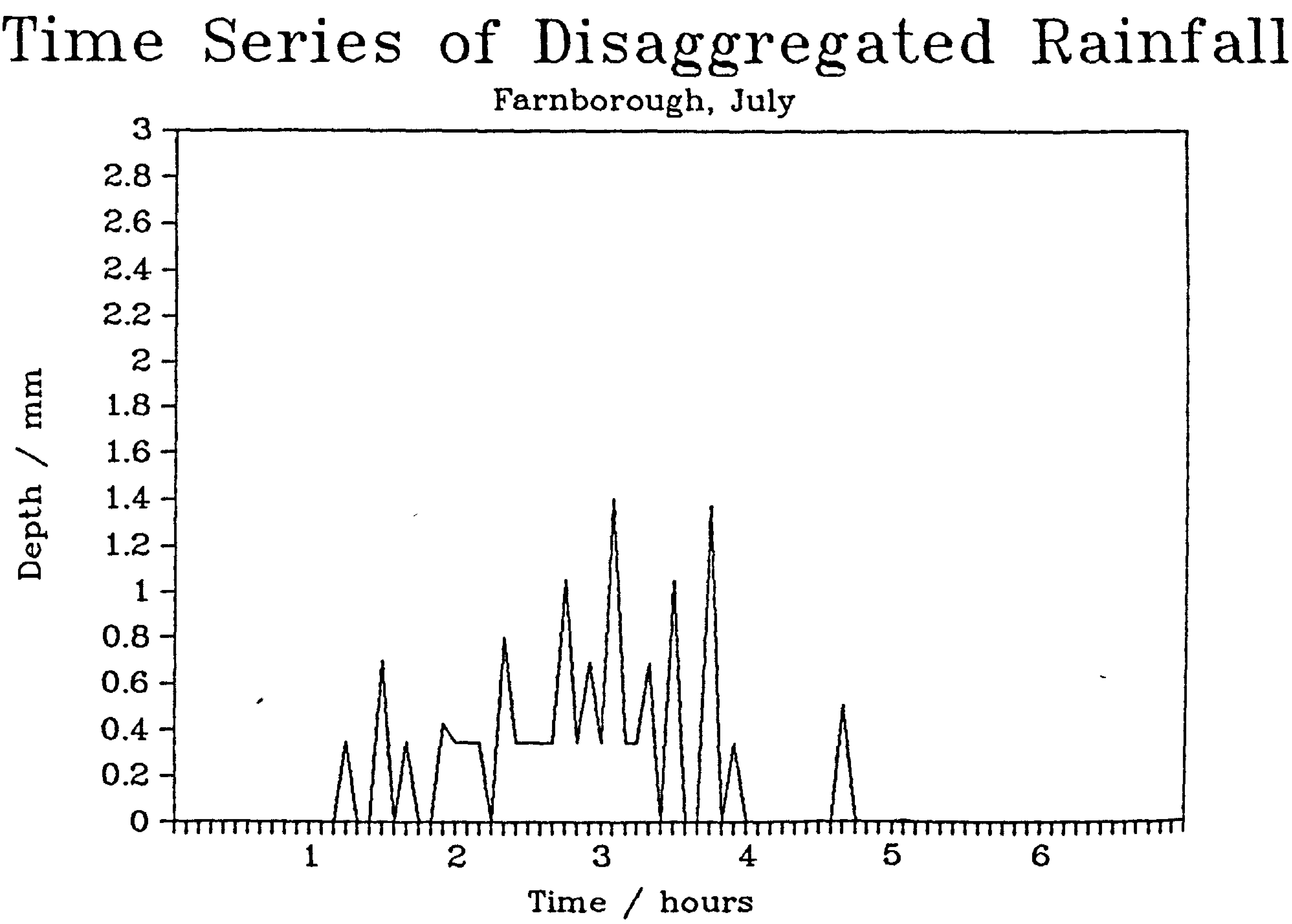


Time Series of Disaggregated Rainfall

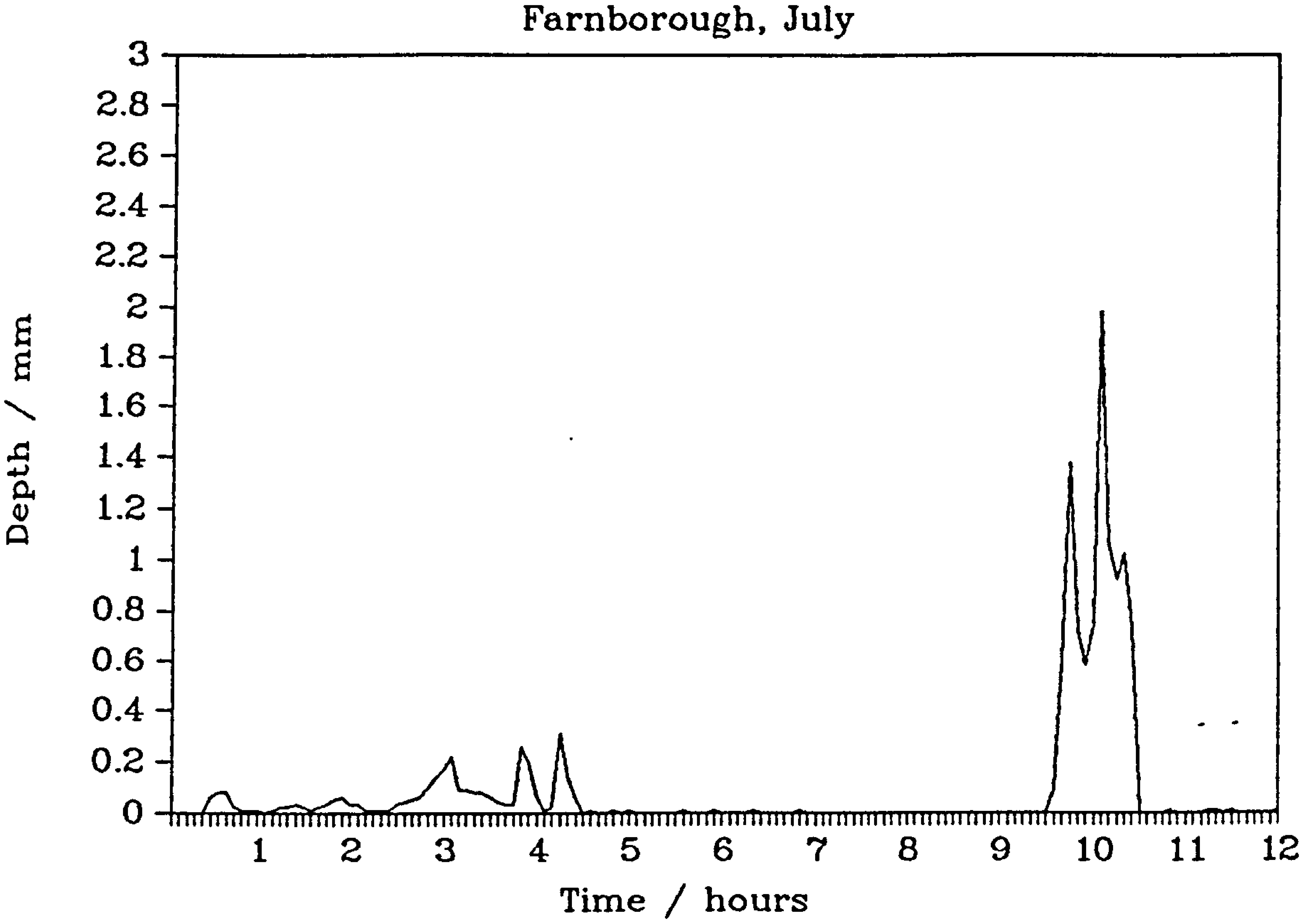
January, Farnborough



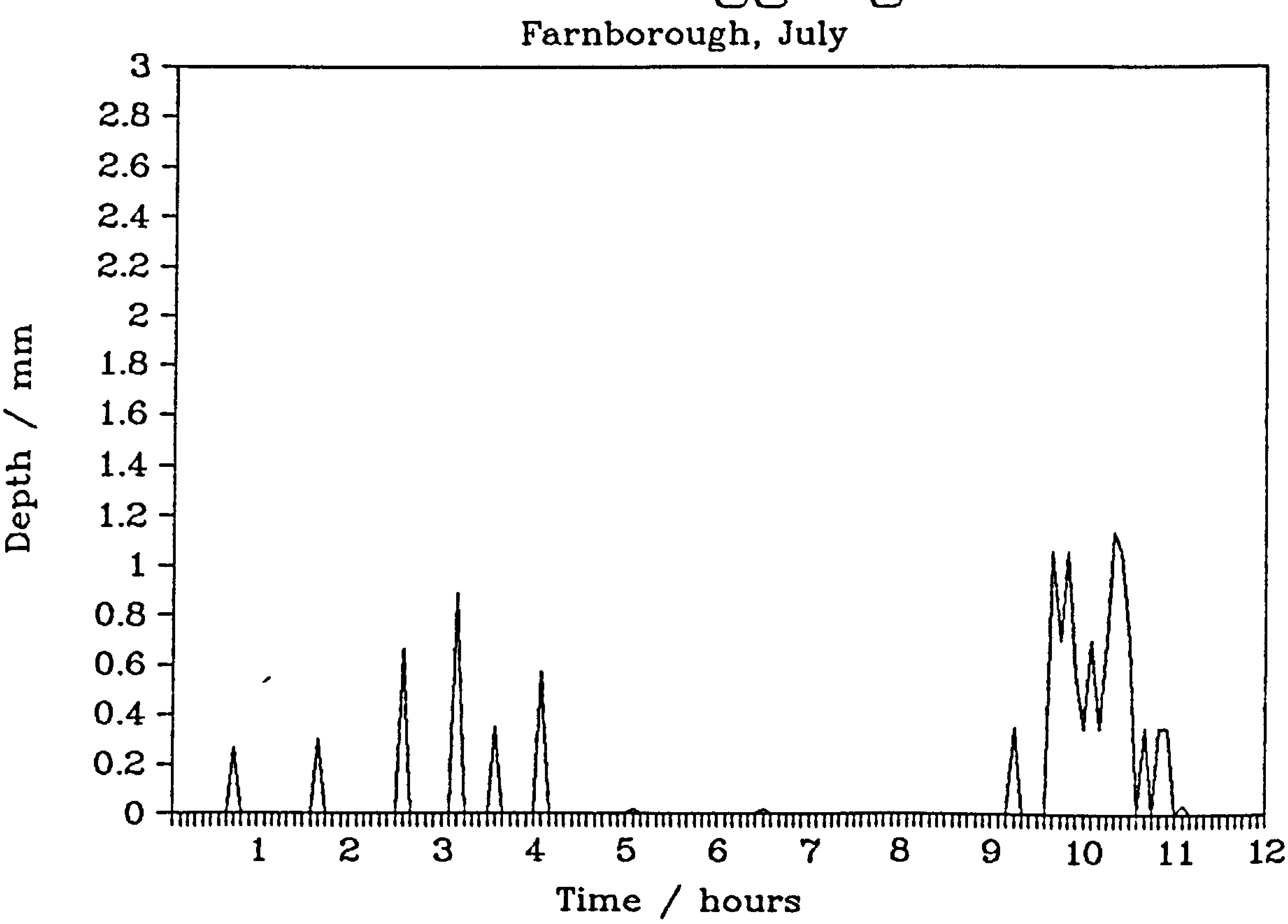

Time Series of Rainfall
Farnborough, July
Depth / mm
Time / hours
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
1
2
3
4
5
6


Time Series of Disaggregated Rainfall
Farnborough, July
Depth / mm
Time / hours
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
1
2
3
4
5
6

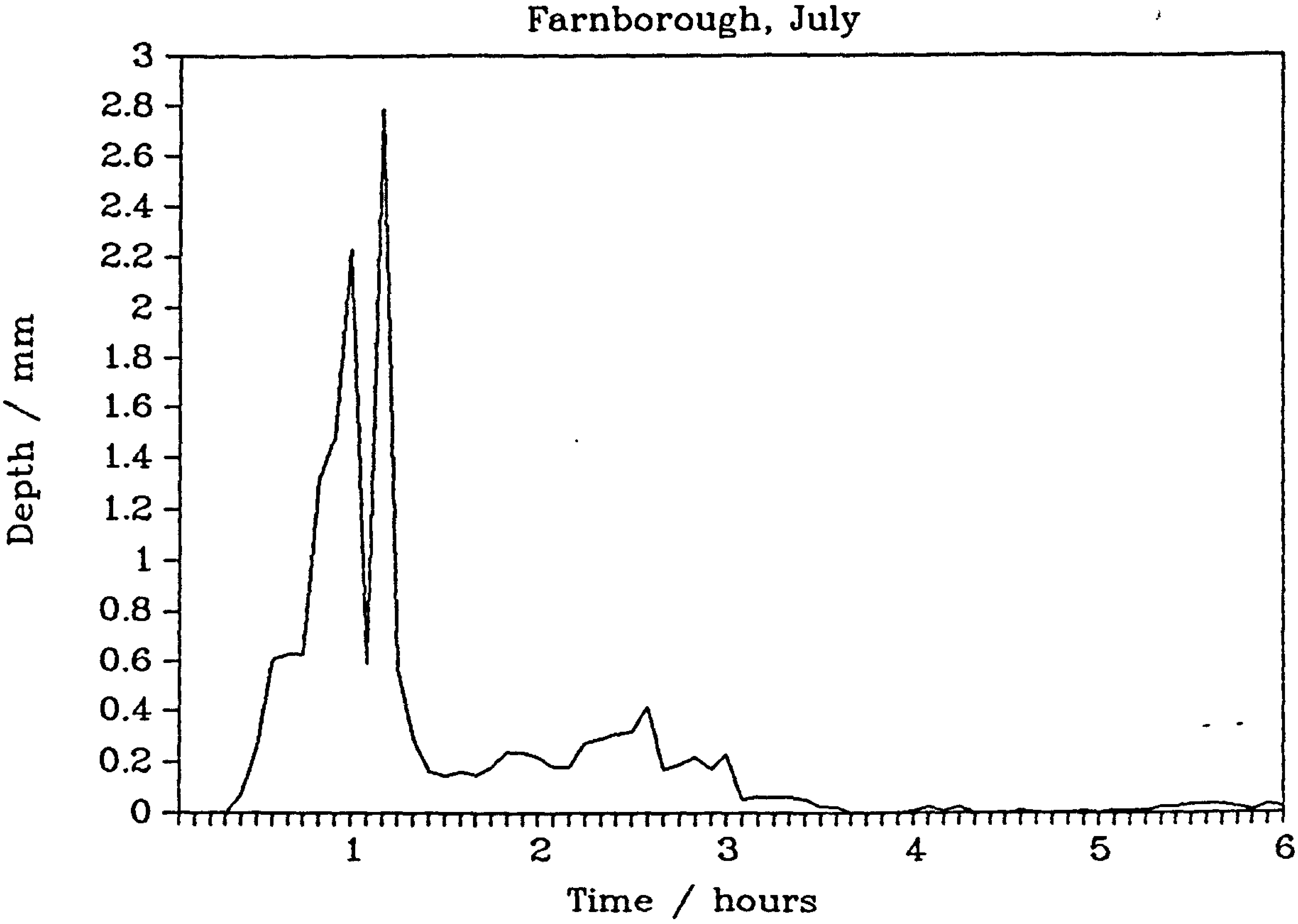
Time Series of Rainfall



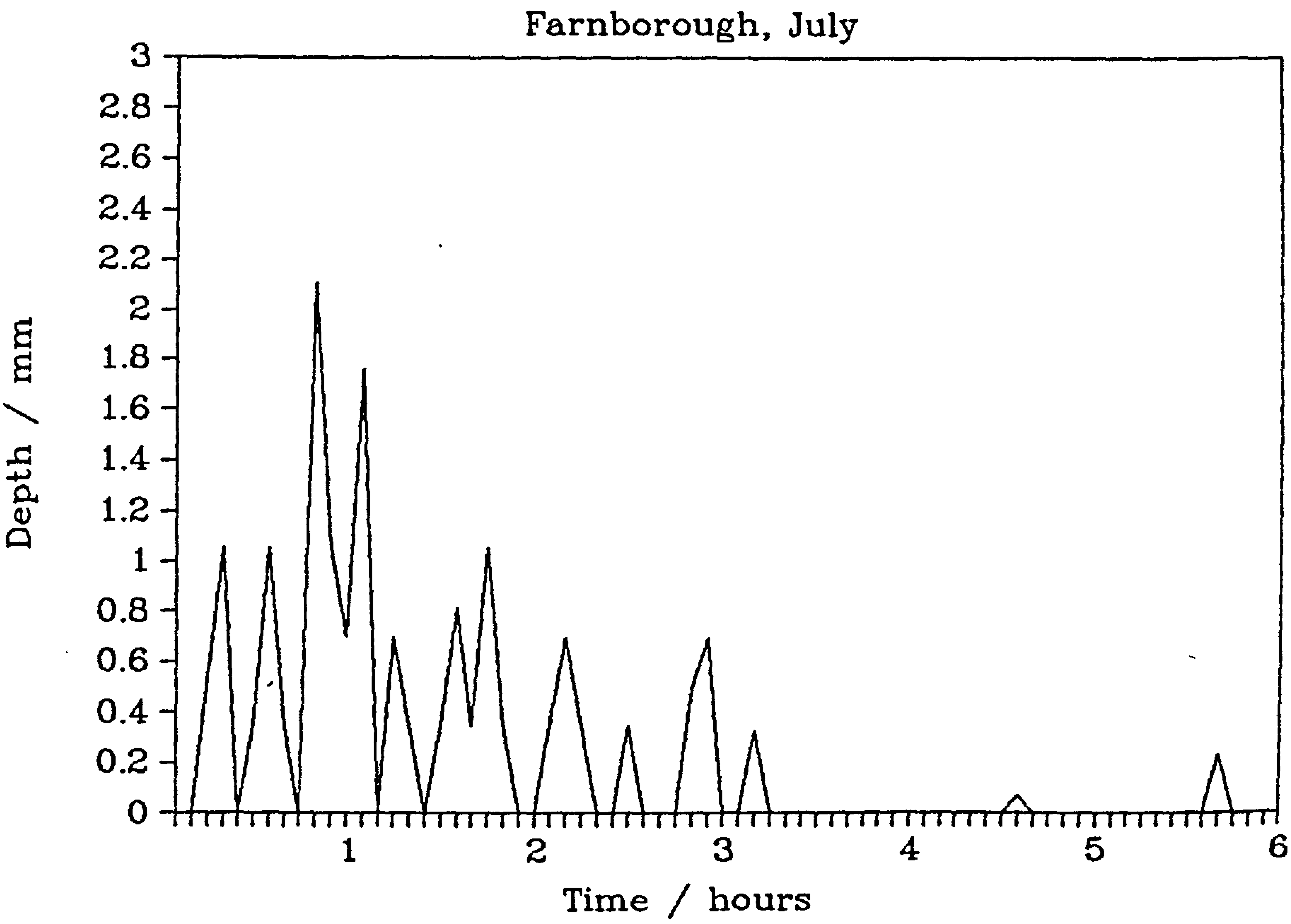
Time Series of Disaggregated Rainfall



Time Series of Rainfall

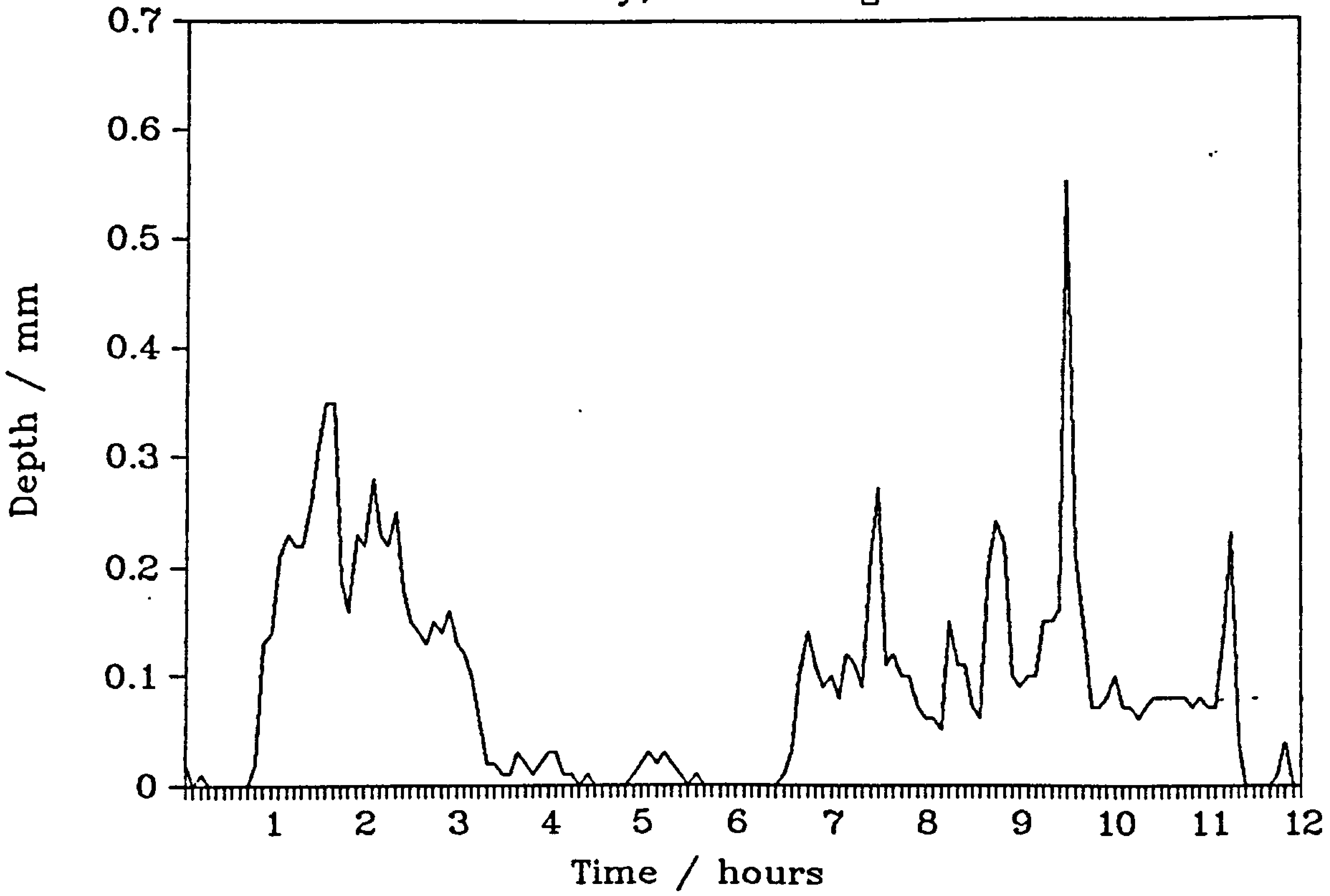


Time Series of Disaggregated Rainfall



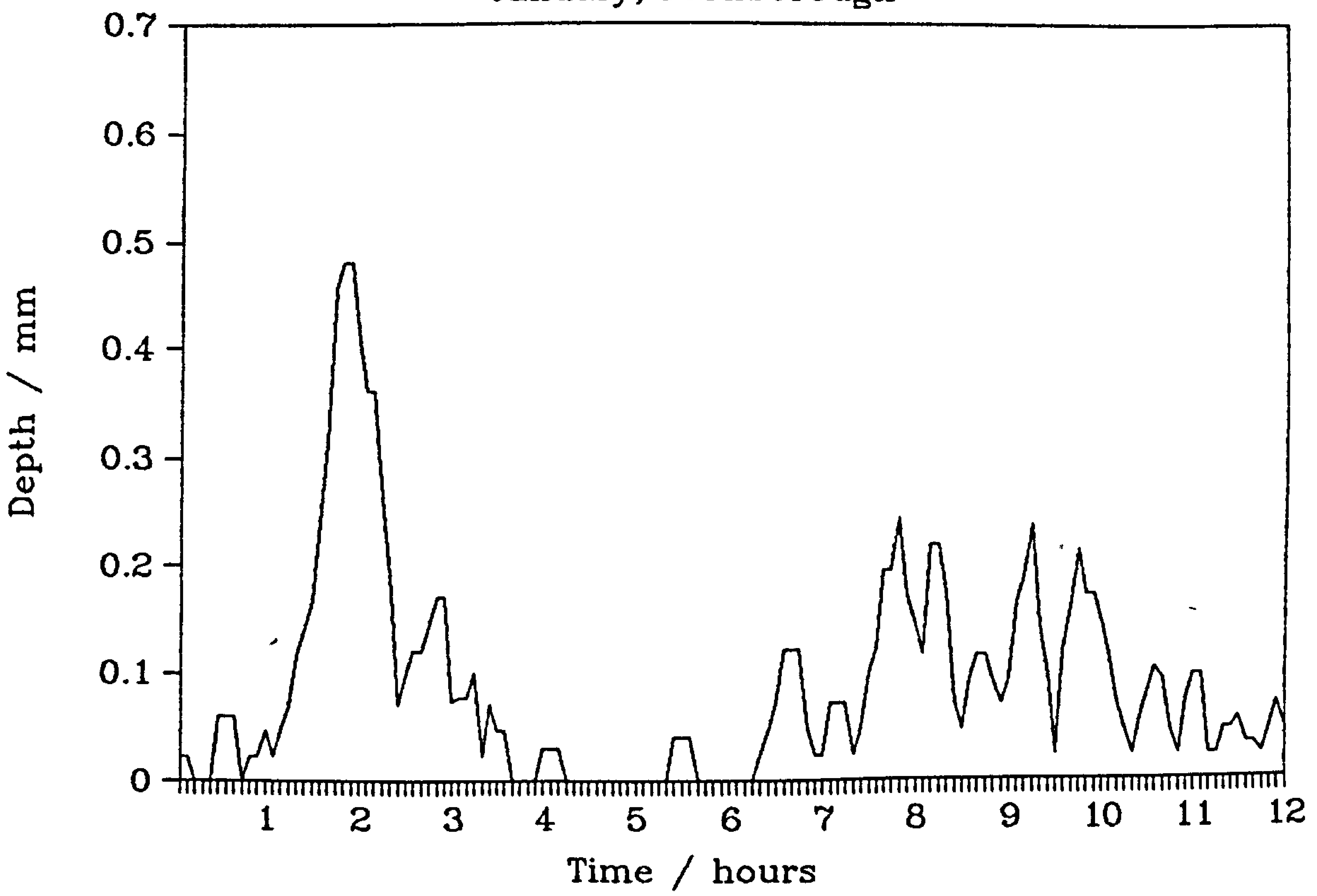
Time Series of Rainfall

January, Farnborough



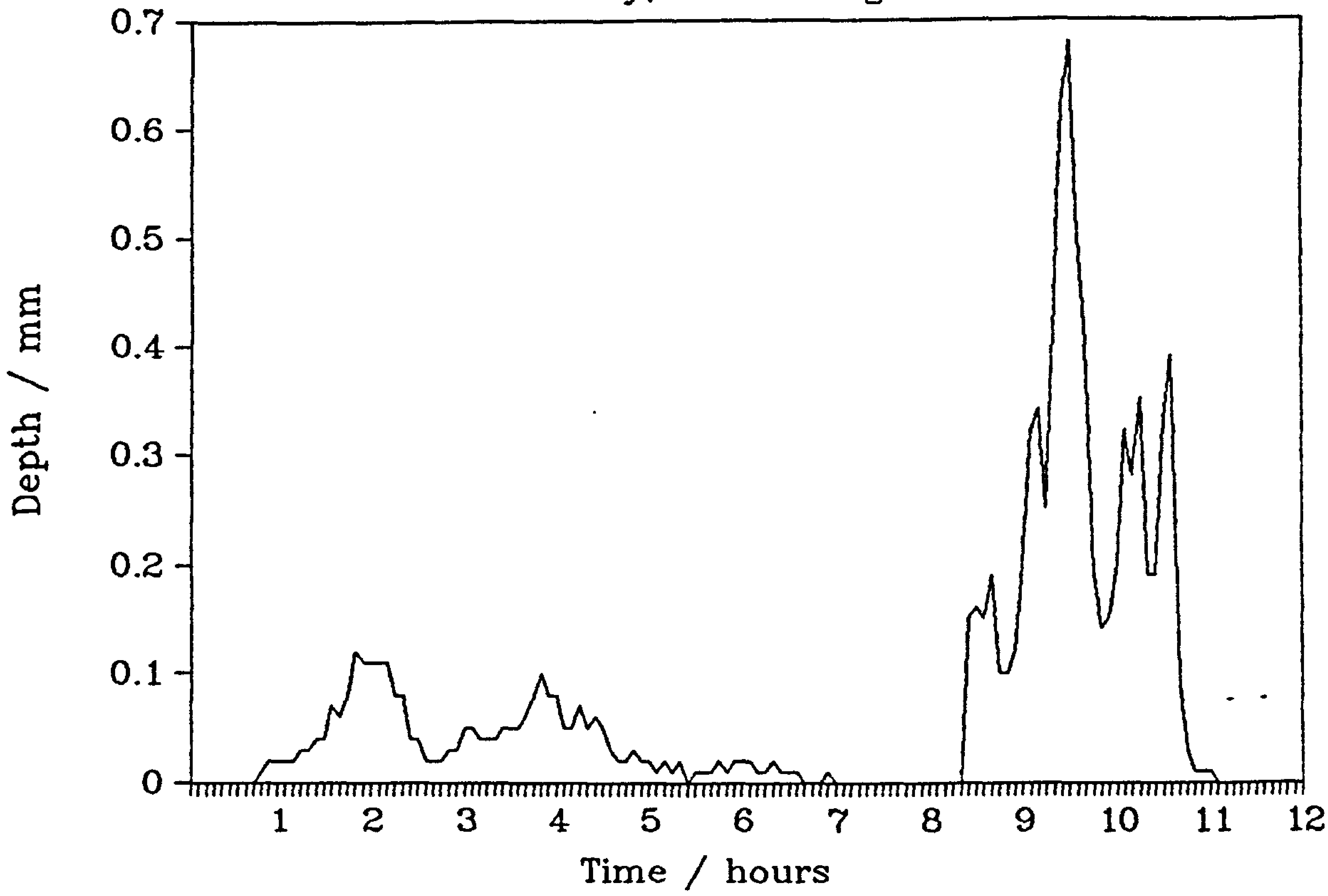
Time Series of 'Smoothed' Disaggregated Rainfall

January, Farnborough



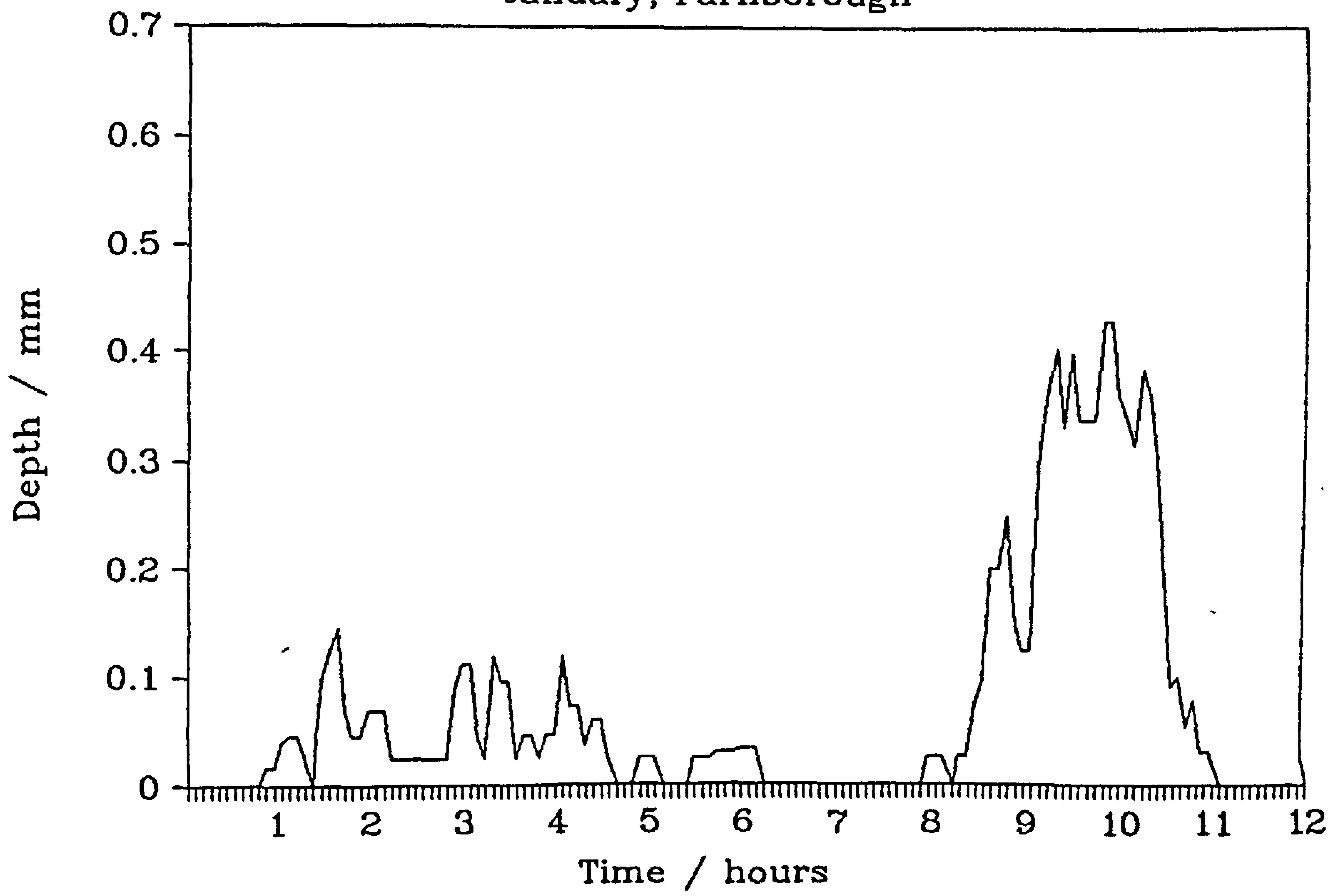
Time Series of Rainfall

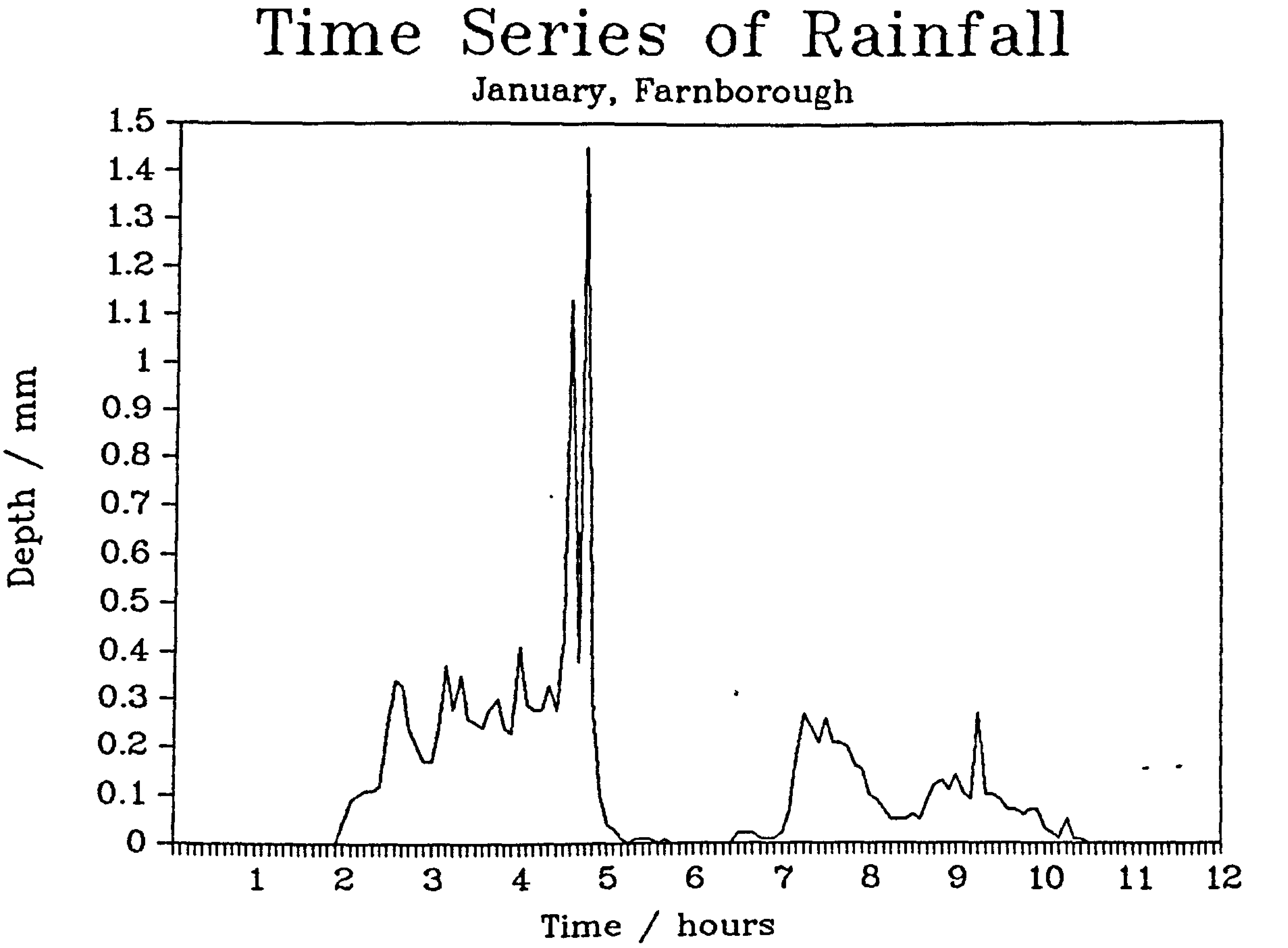
January, Farnborough



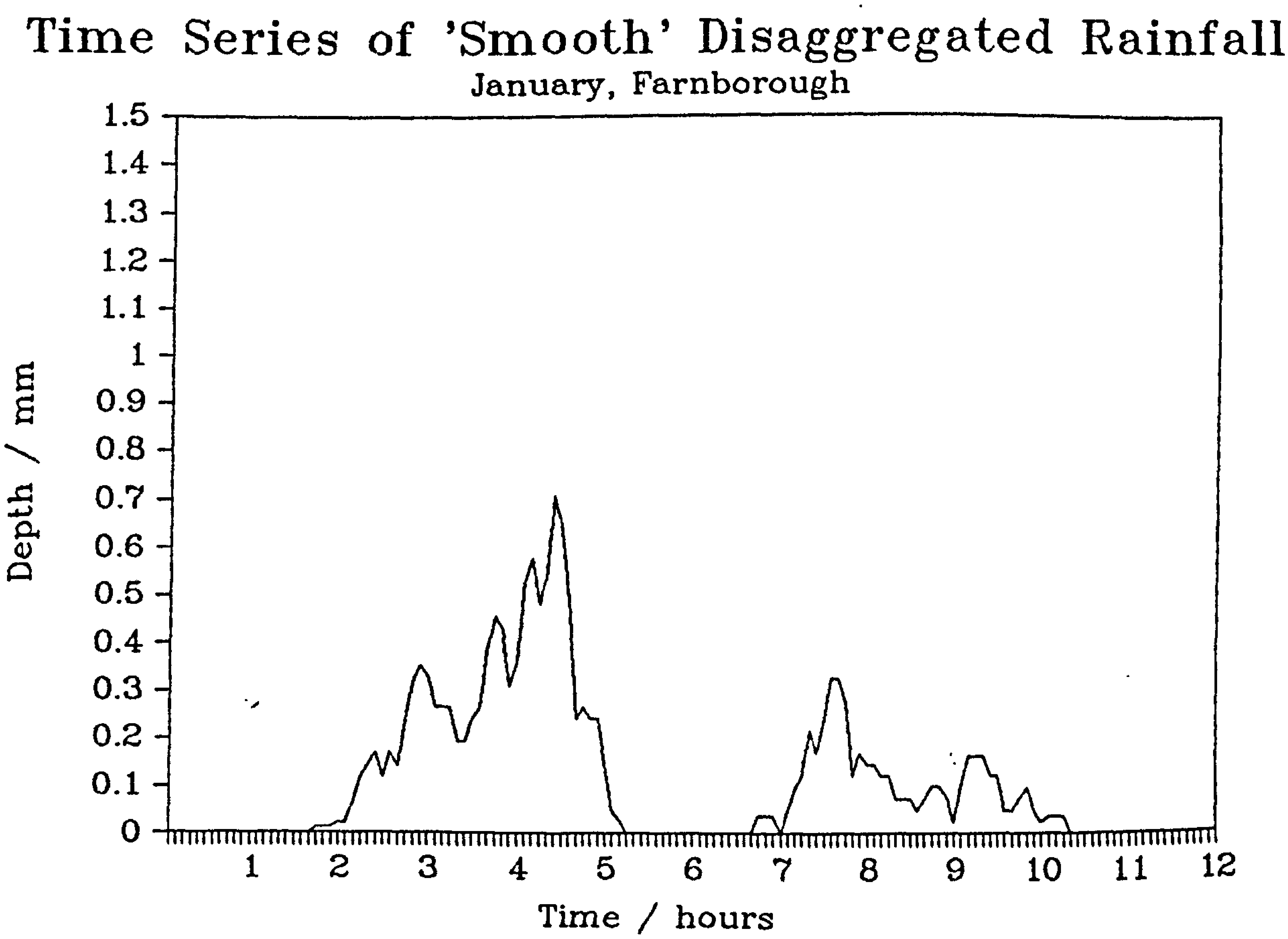
Time Series of 'Smooth' Disaggregated Rainfall

January, Farnborough





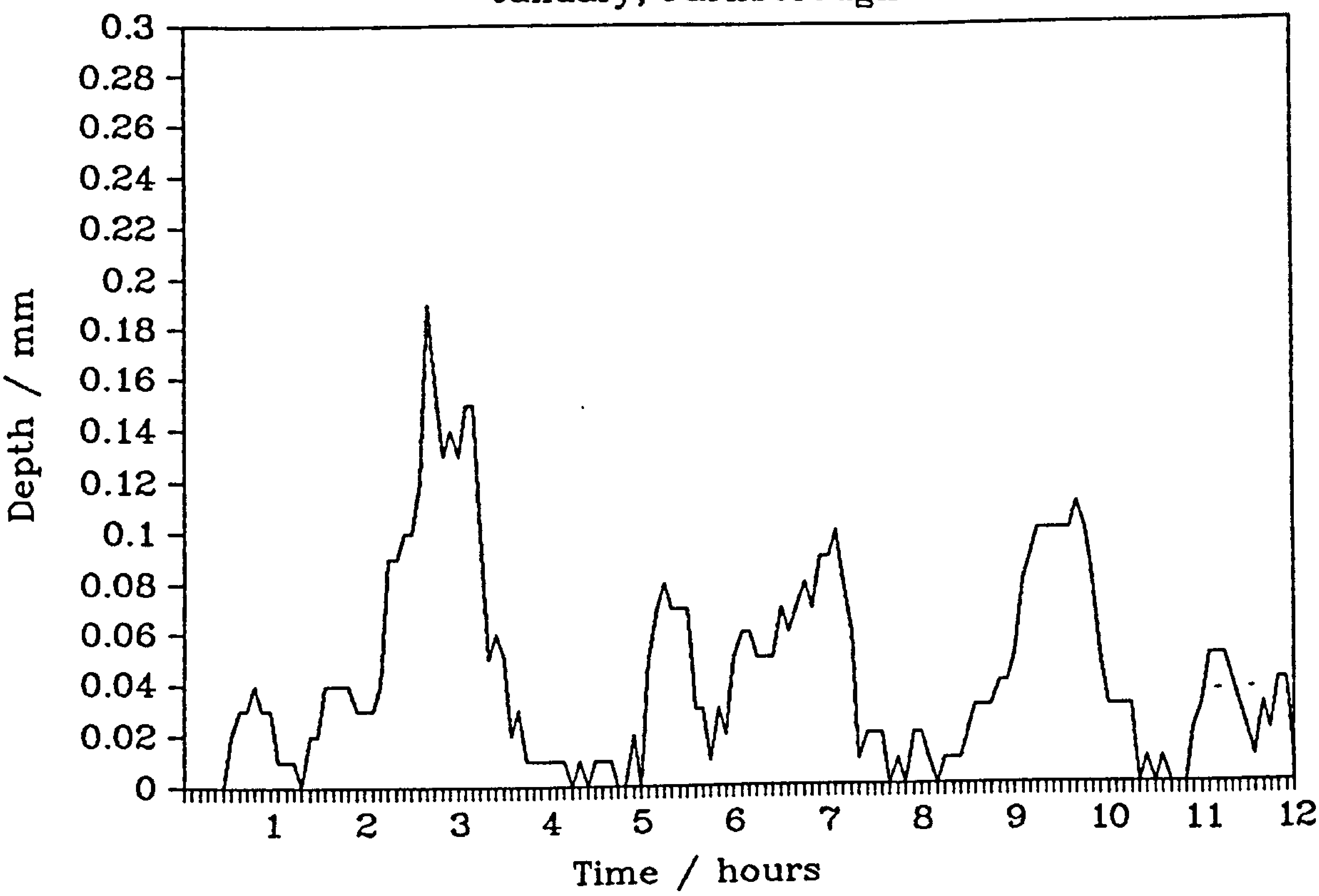
Time Series of Rainfall
January, Farnborough
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Depth / mm
1
2
3
4
5
6
7
8
9
10
11
12
Time / hours



Time Series of 'Smooth' Disaggregated Rainfall
January, Farnborough
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Depth / mm
1
2
3
4
5
6
7
8
9
10
11
12
Time / hours

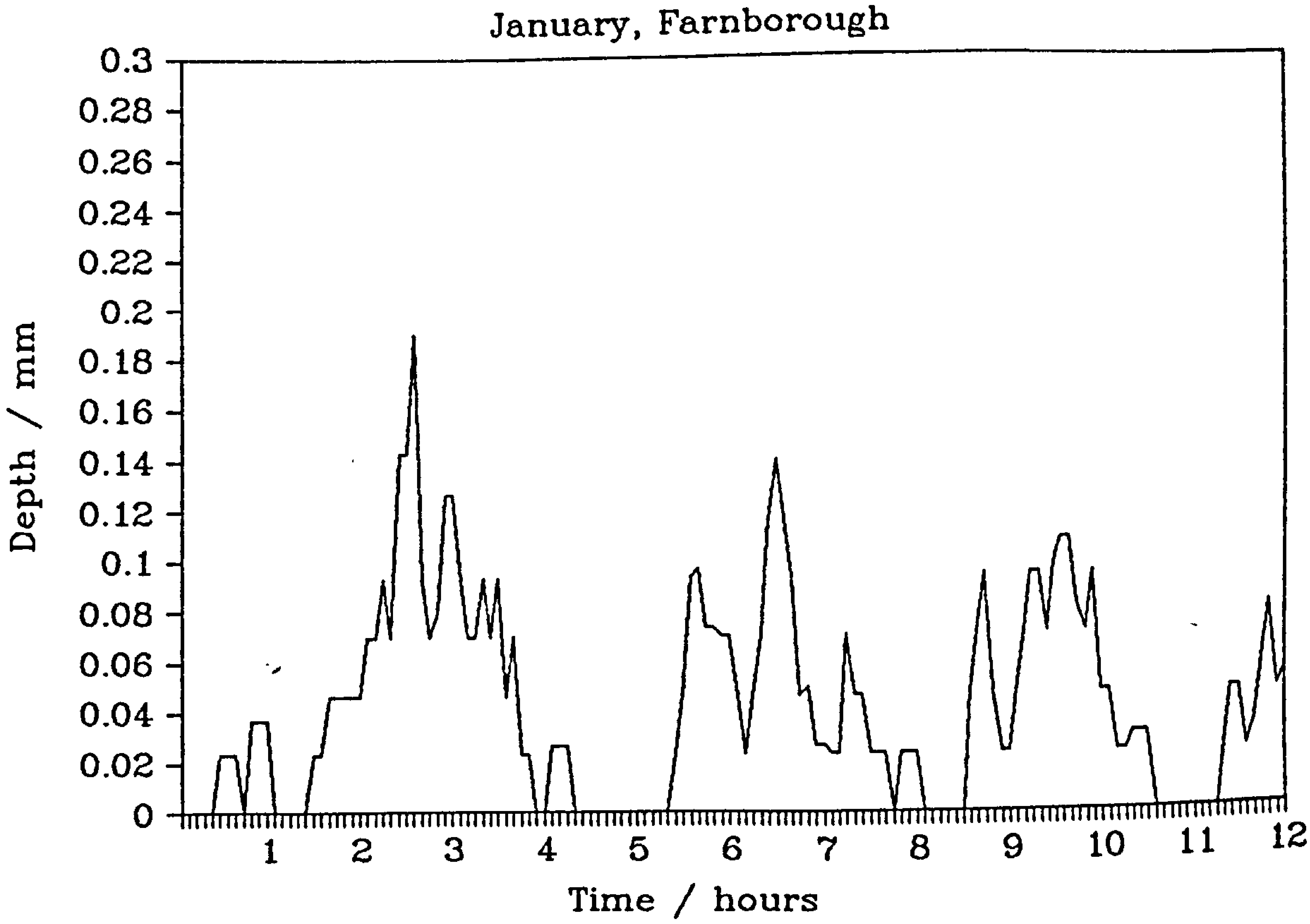
Time Series of Rainfall

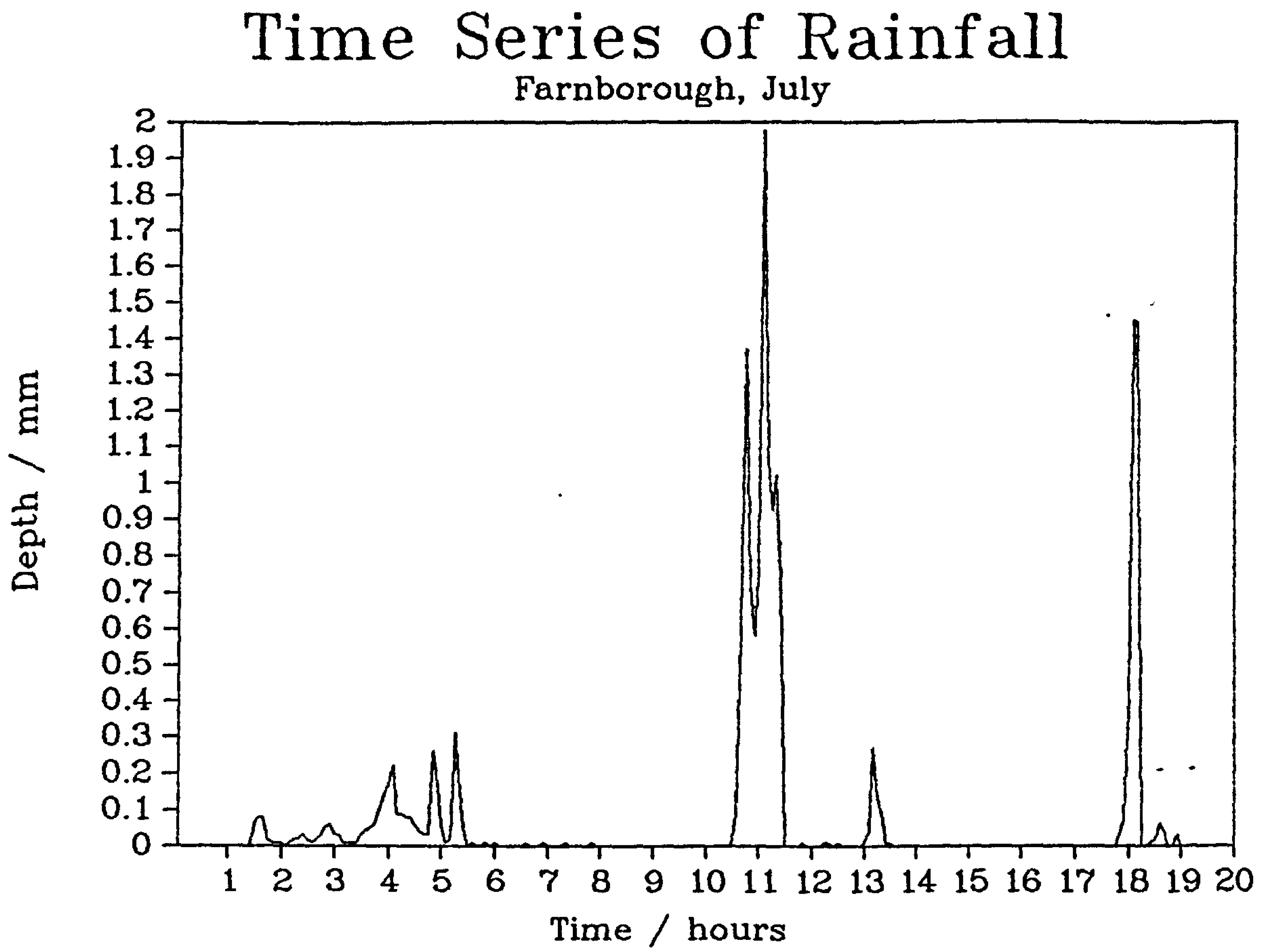
January, Farnborough

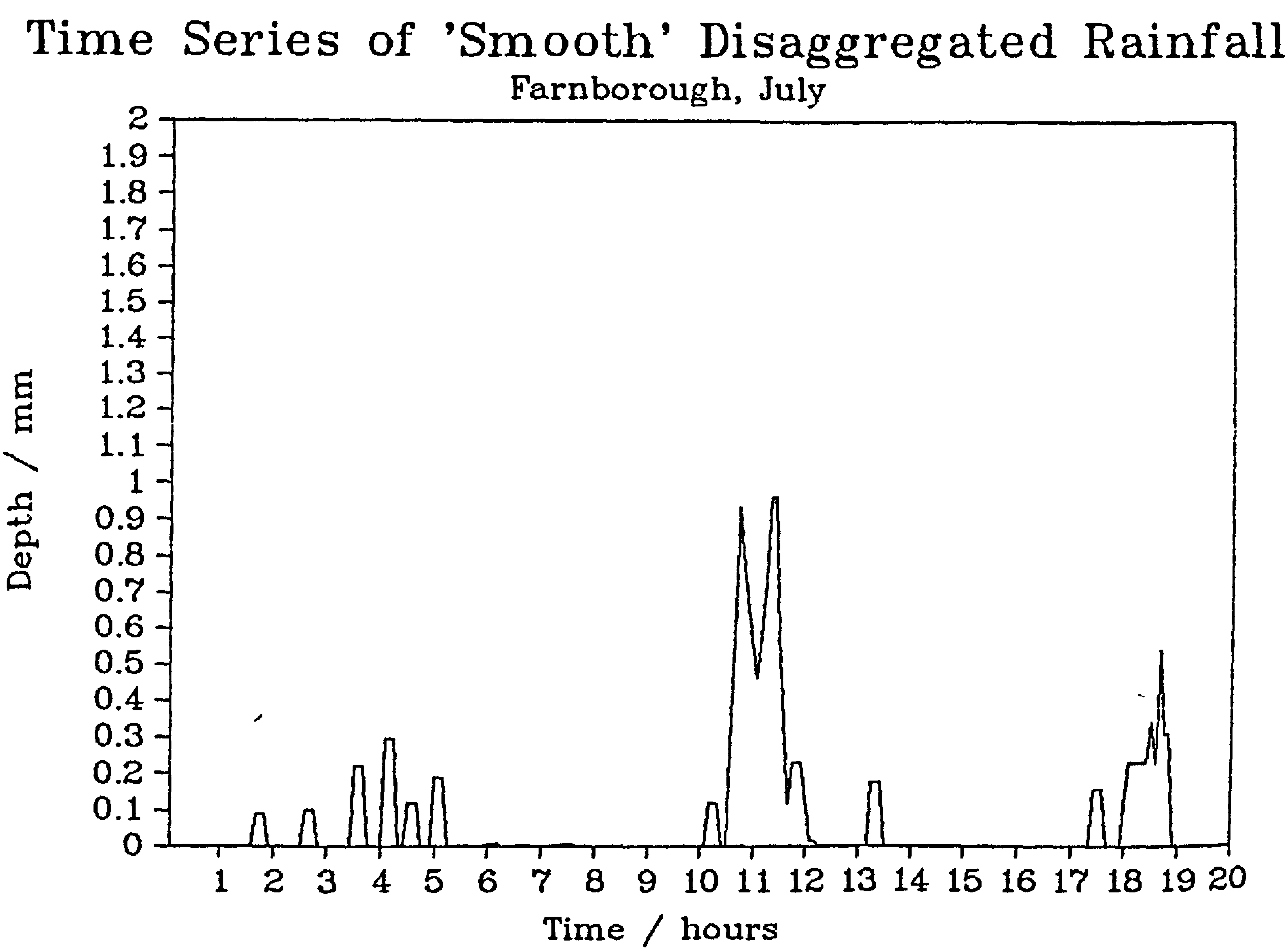


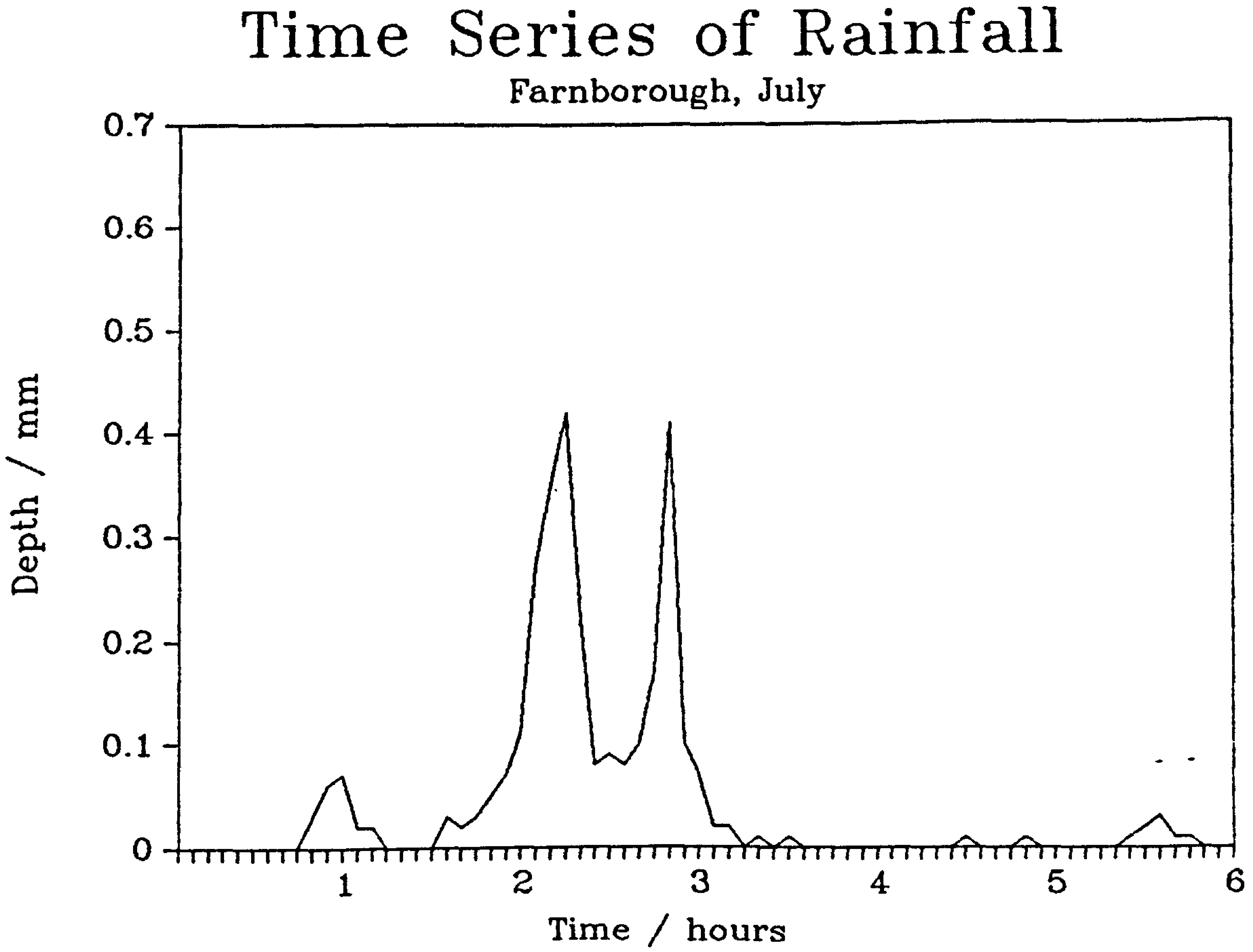
Time Series of 'Smooth' Disaggregated Rainfall

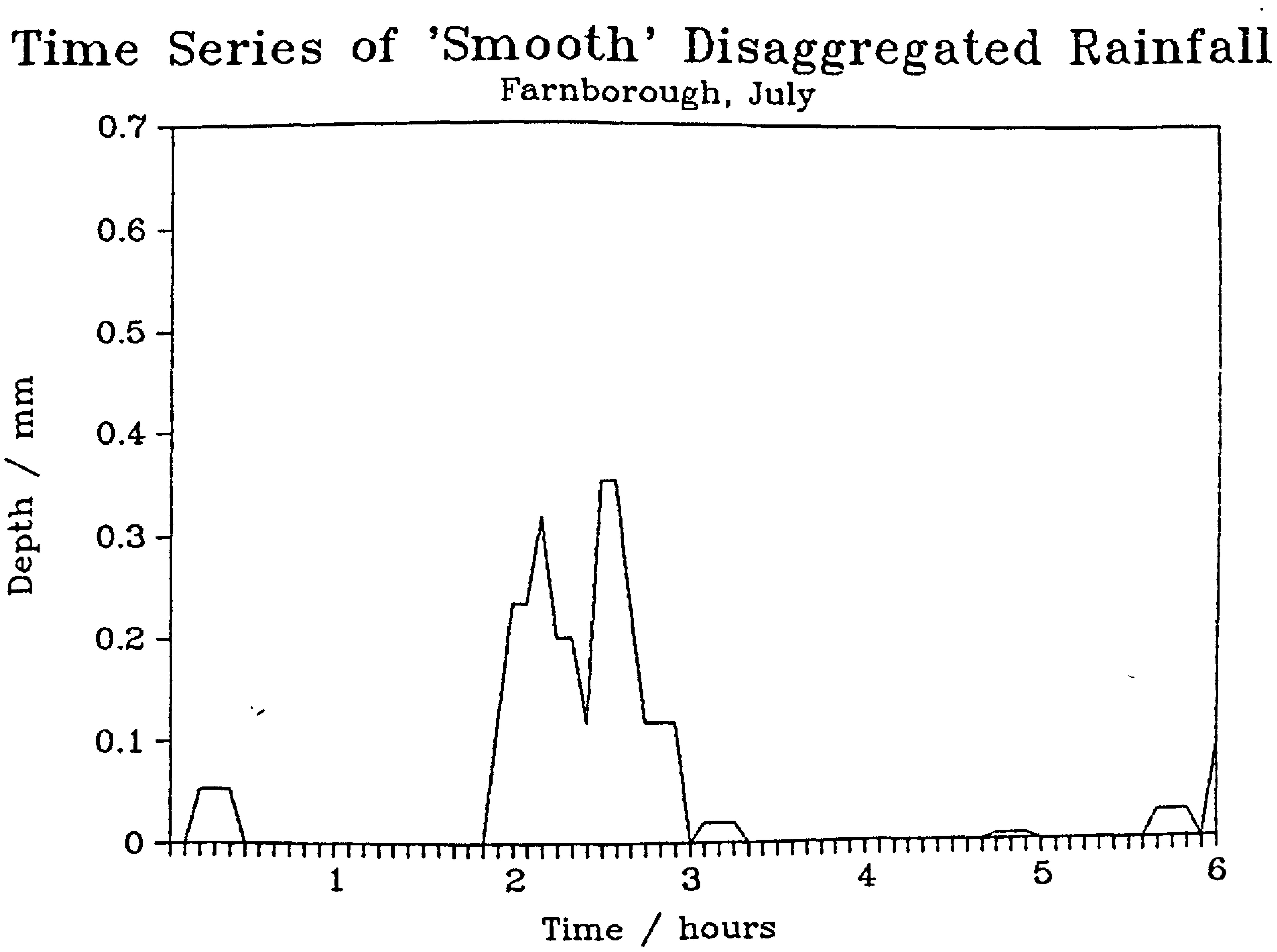
January, Farnborough

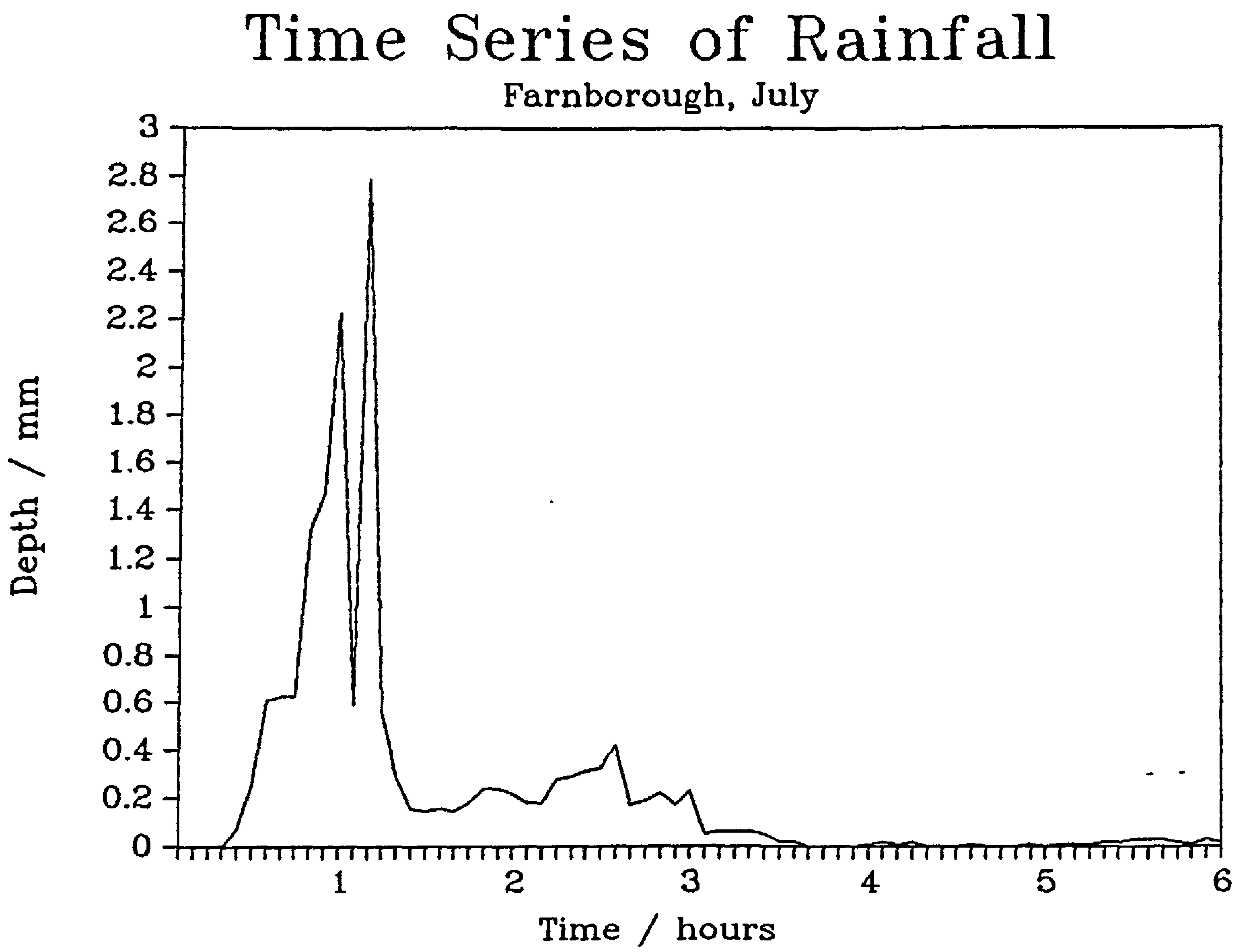



Time Series of Rainfall
Farnborough, July
Depth / mm
Time / hours
2 1.9 1.8 1.7 1.6 1.5 1.4 1.3 1.2 1.1 1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

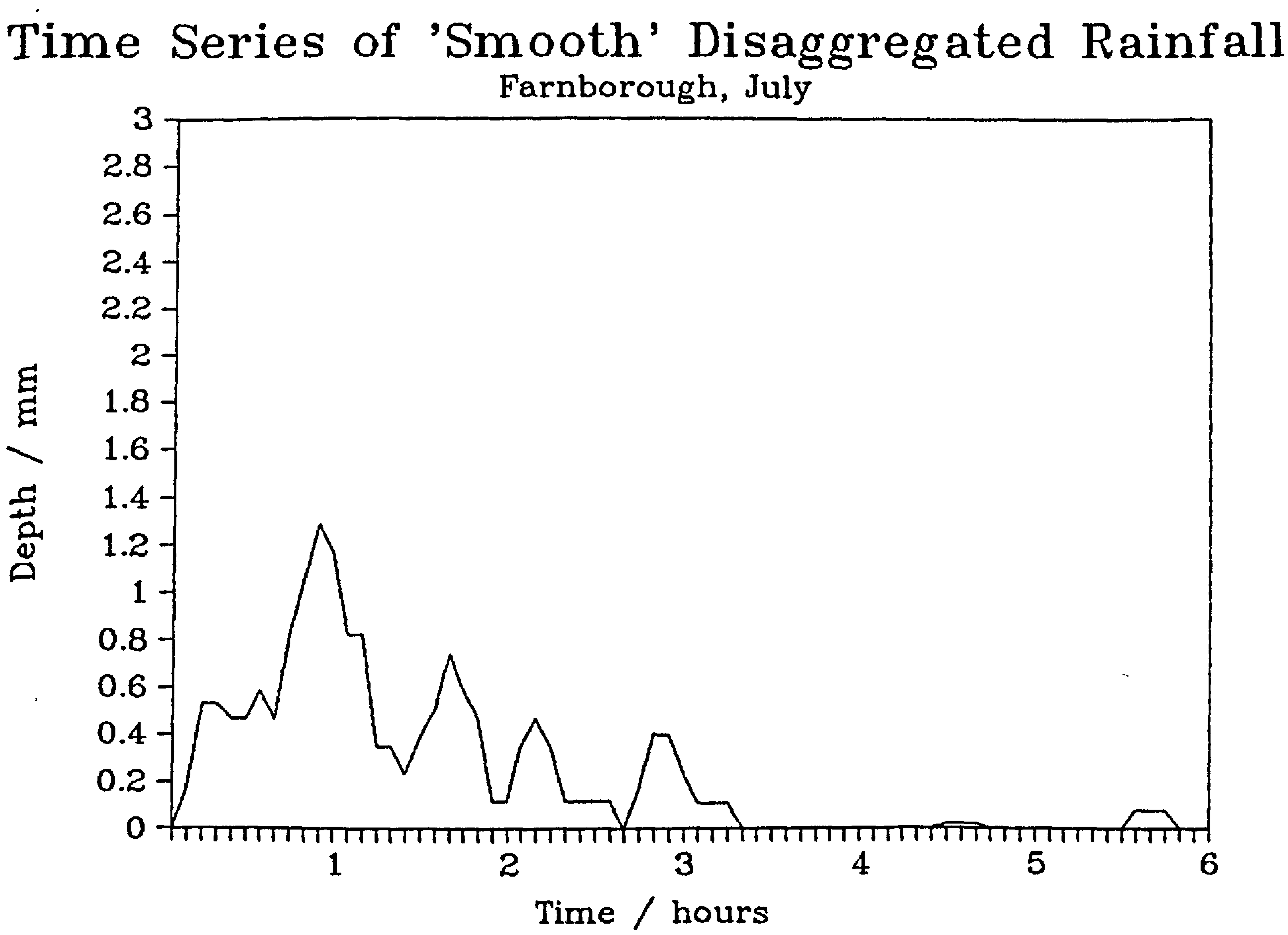

Time Series of 'Smooth' Disaggregated Rainfall
Farnborough, July
Depth / mm
Time / hours
2 1.9 1.8 1.7 1.6 1.5 1.4 1.3 1.2 1.1 1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20


Time Series of Rainfall
Farnborough, July
Depth / mm
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
Time / hours


Time Series of 'Smooth' Disaggregated Rainfall
Farnborough, July
Depth / mm
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
Time / hours



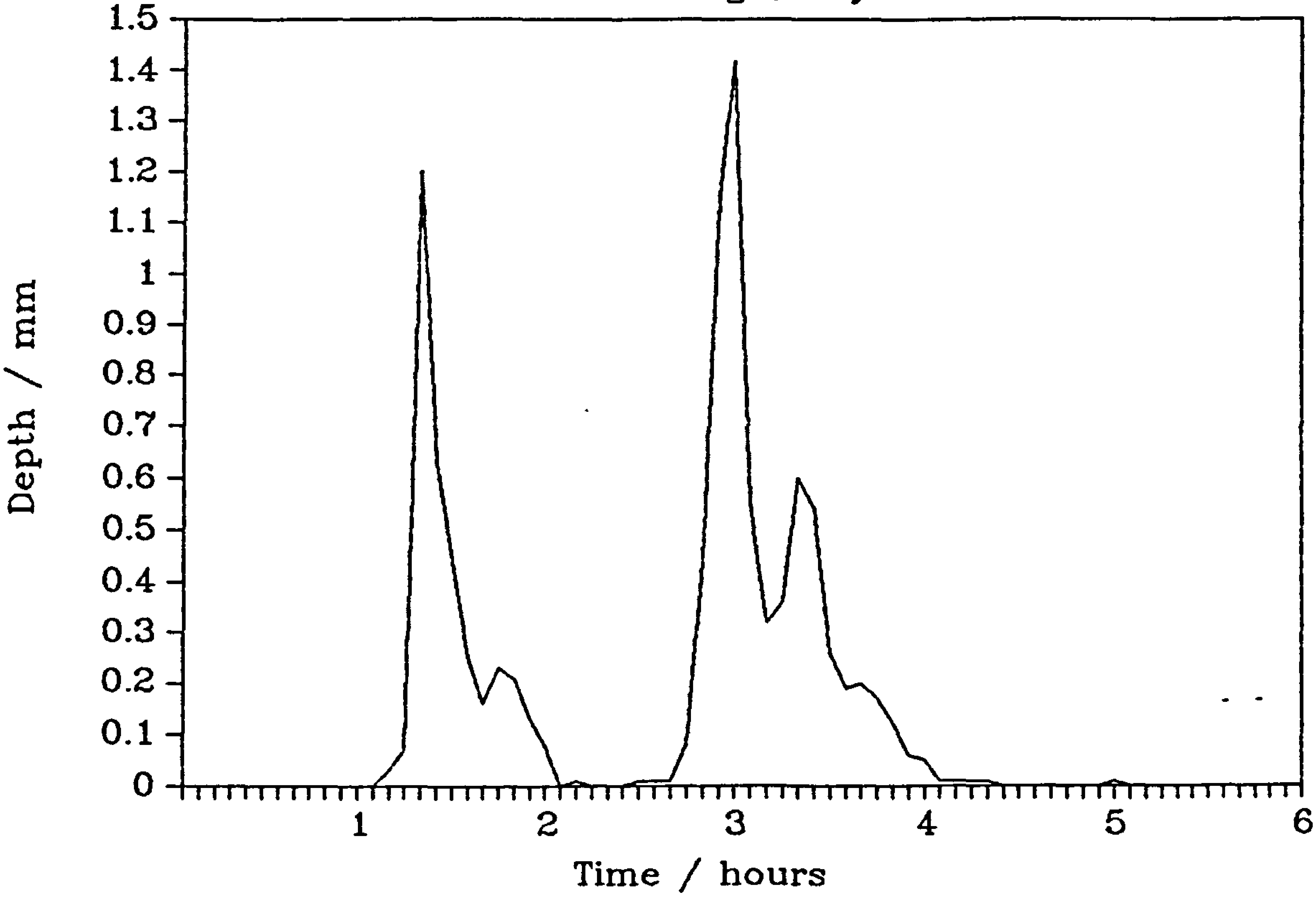
Time Series of Rainfall
Farnborough, July
Depth / mm
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Time / hours



Time Series of 'Smooth' Disaggregated Rainfall
Farnborough, July
Depth / mm
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Time / hours

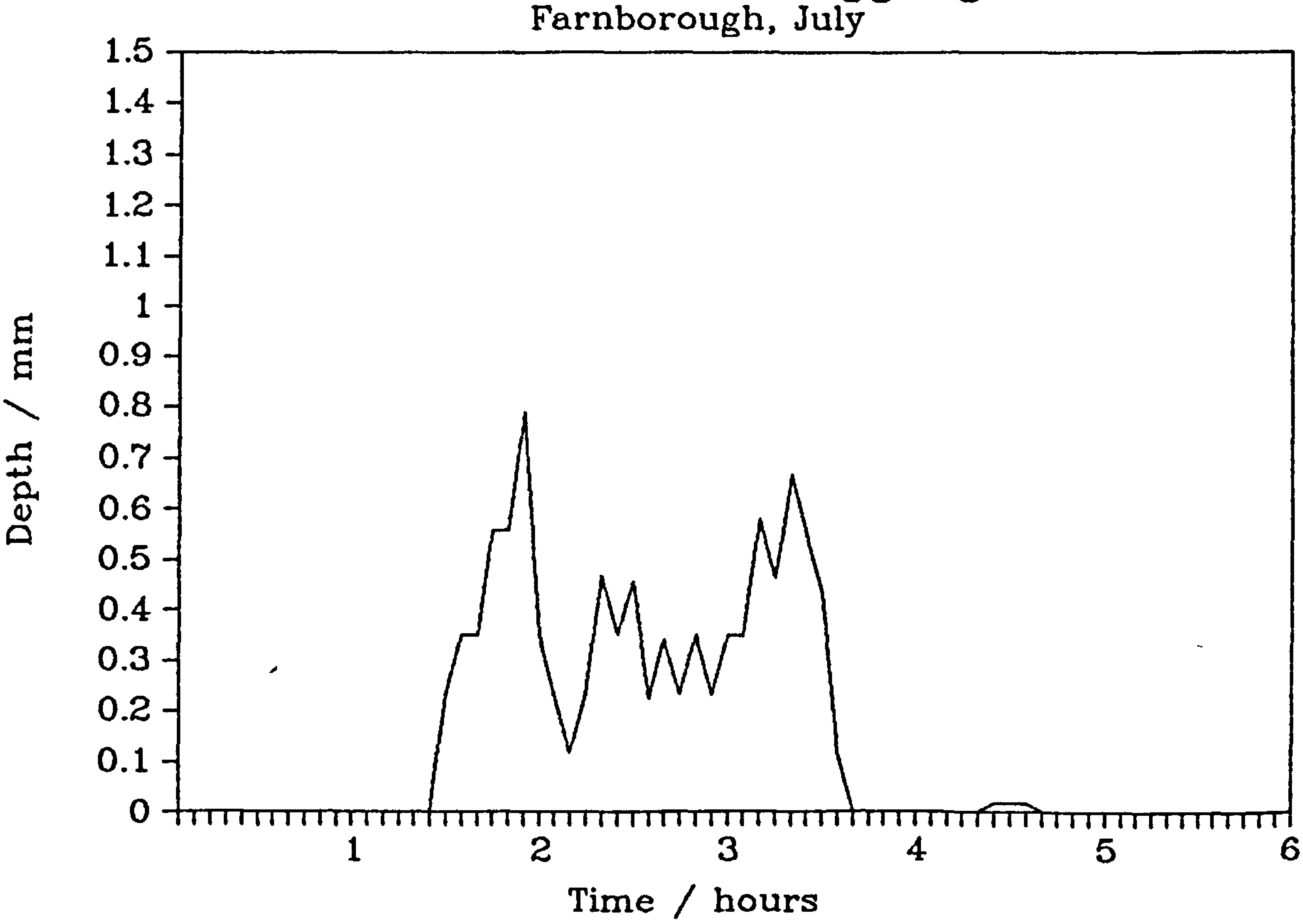
Time Series of Rainfall

Farnborough, July



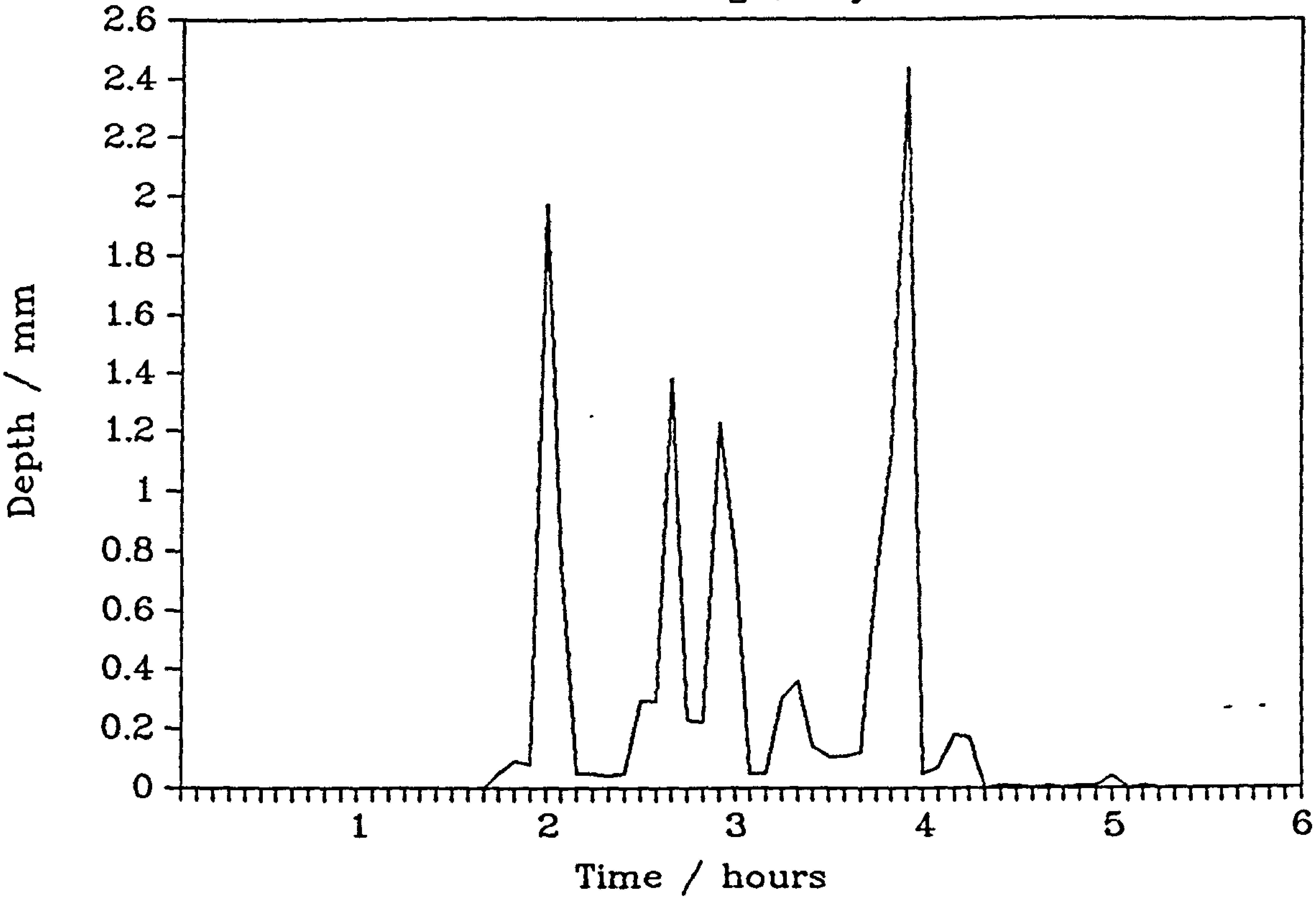
Time Series of 'Smooth' Disaggregated Rainfall

Farnborough, July



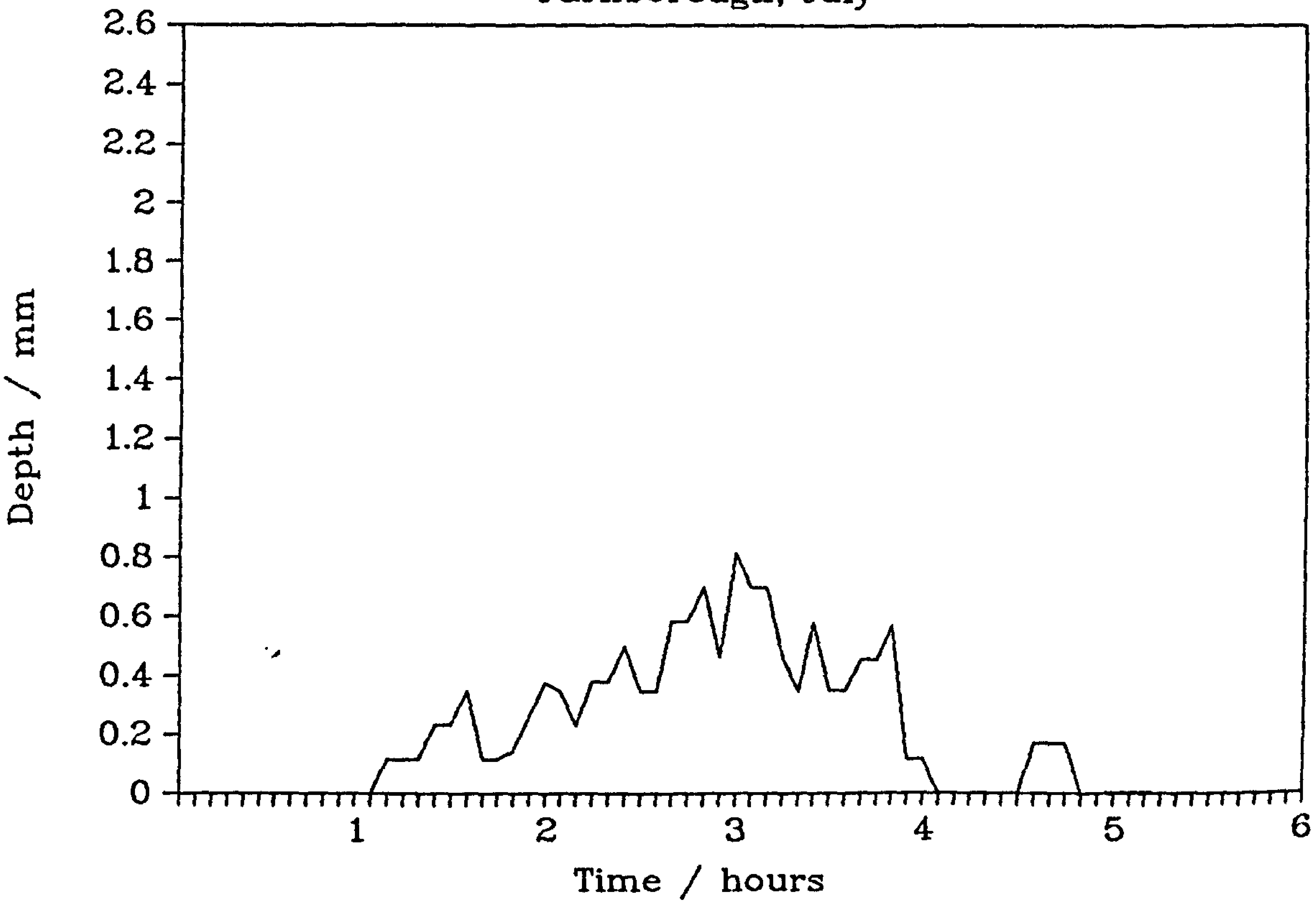
Time Series of Rainfall

Farnborough, July



Time Series of 'Smooth' Disaggregated Rainfall

Farnborough, July



APPENDIX J: CORRELOGRAMS OF MONTHLY MEAN DAILY RAINFALLS AND RESIDUAL SERIES

In the fitting procedure for the model, it has been assumed that the parameter estimates for a month do not need to be conditioned on previous monthly totals, i.e. it is assumed that the rainfall for one month does not affect the rainfall for other months. To investigate this, the mean daily rainfalls were found for each year-month of the 5 longest records of daily data. This produced 5 time series of monthly mean daily rainfalls. For each series, the overall mean daily rainfall for each month was subtracted from each of the monthly mean daily rainfalls and the correlogram for the residual series found. These correlograms are shown in Figures J1 - J5, from which it is clear that each residual series is uncorrelated. To make a comparison with one of the original series of monthly mean daily rainfalls, a correlogram is given for Poaka Beck (Figure J6), from which it can be seen that there is strong evidence of a seasonal autocorrelation pattern.

Mathematical summary: Consider one of the five stations, of record length 90 years, say. Let z_{ij} be the mean daily rainfall for year i, month j (where i = 1, ..., 90 and j = 1, ..., 12), i.e. let $\{z_{ij}\}$ be the time series of monthly mean daily rainfalls. The residual time series $\{z'_{ij}\}$ is given by:

$$z'_{ij} = z_{ij} - \sum_{i=1}^{90} z_{ij}/90$$

The correlograms (Figures J1 - J5) suggest that the residual series for each station are not significantly autocorrelated.

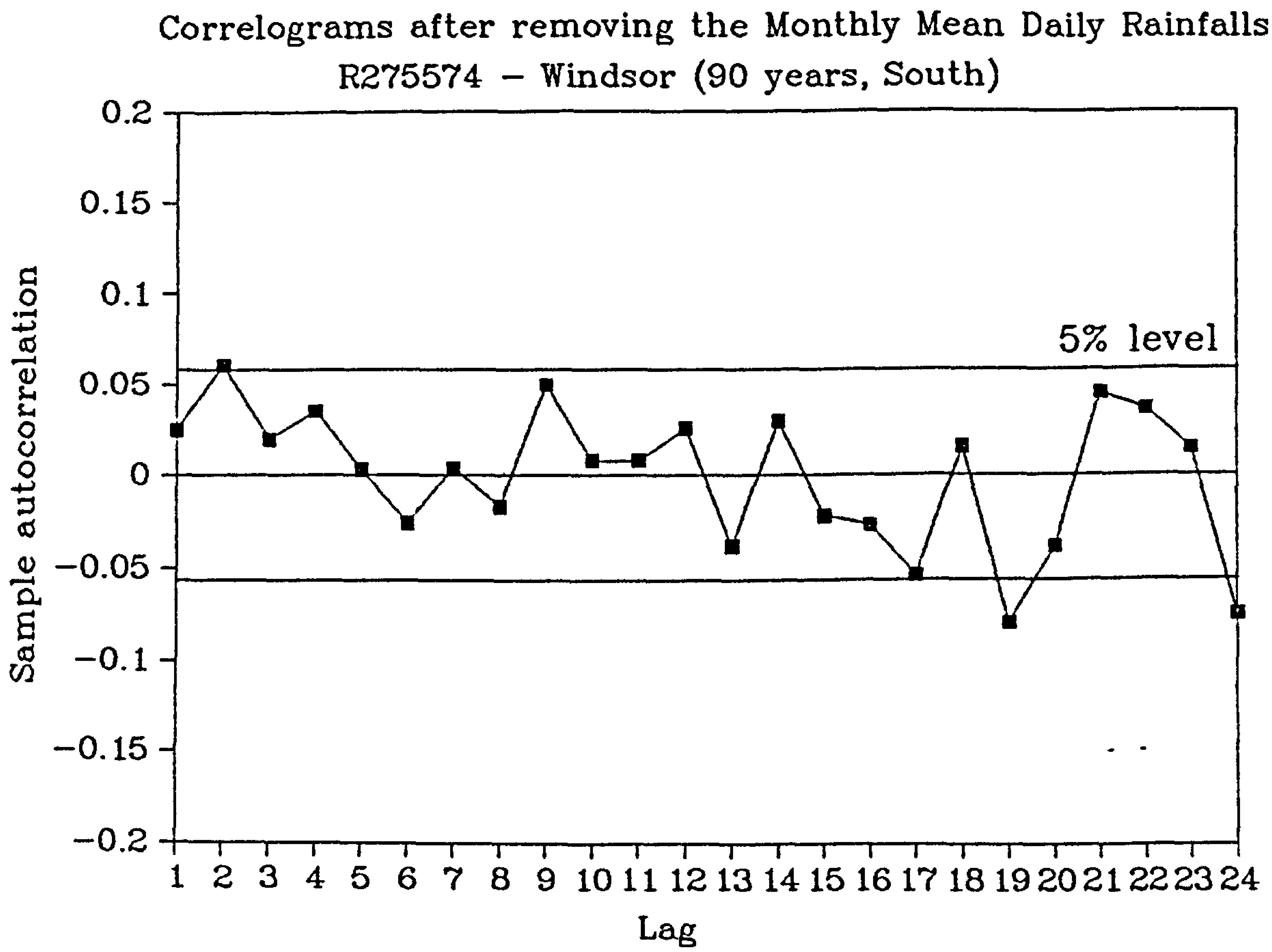


Figure J.1

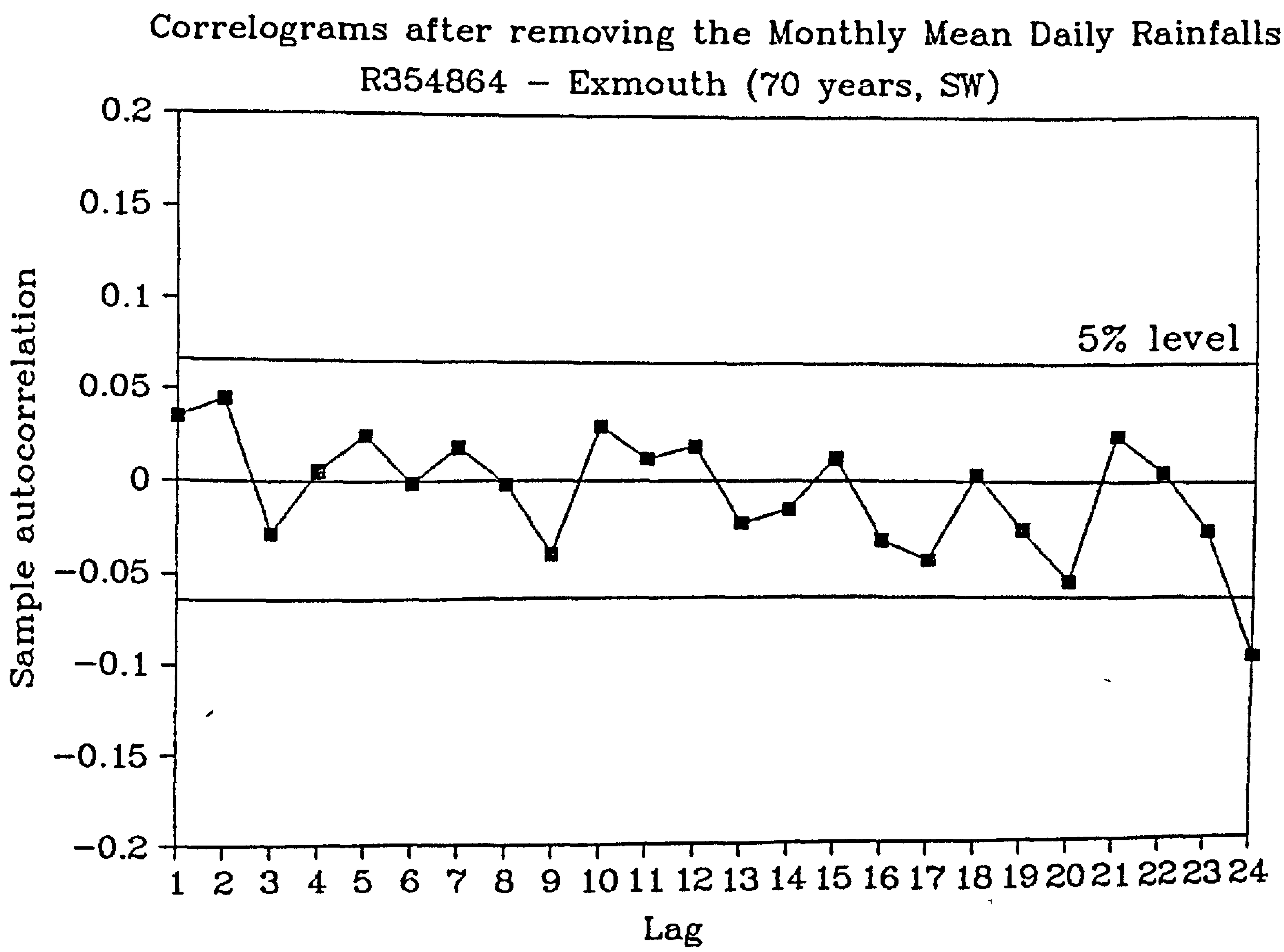


Figure J.2

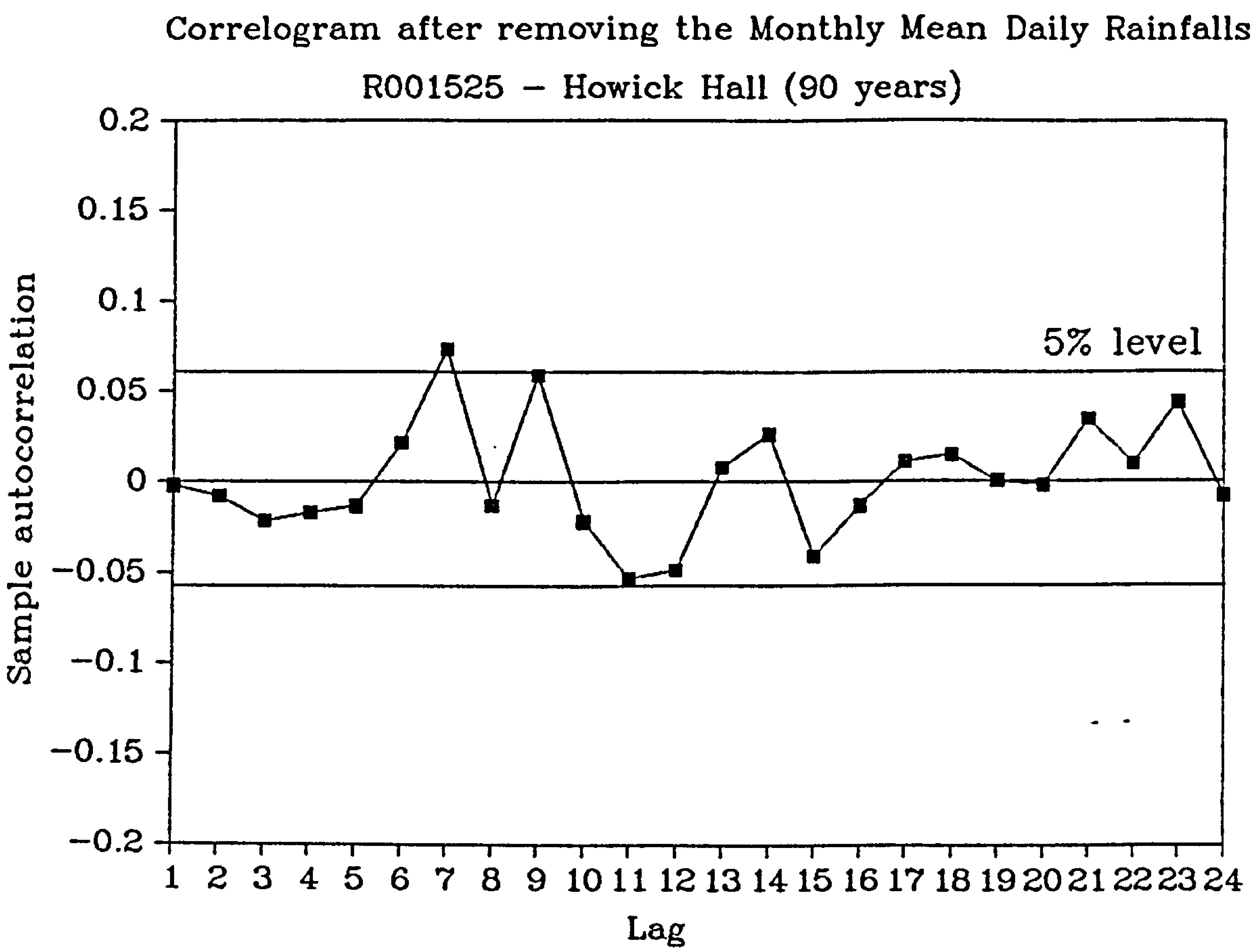


Figure J.3

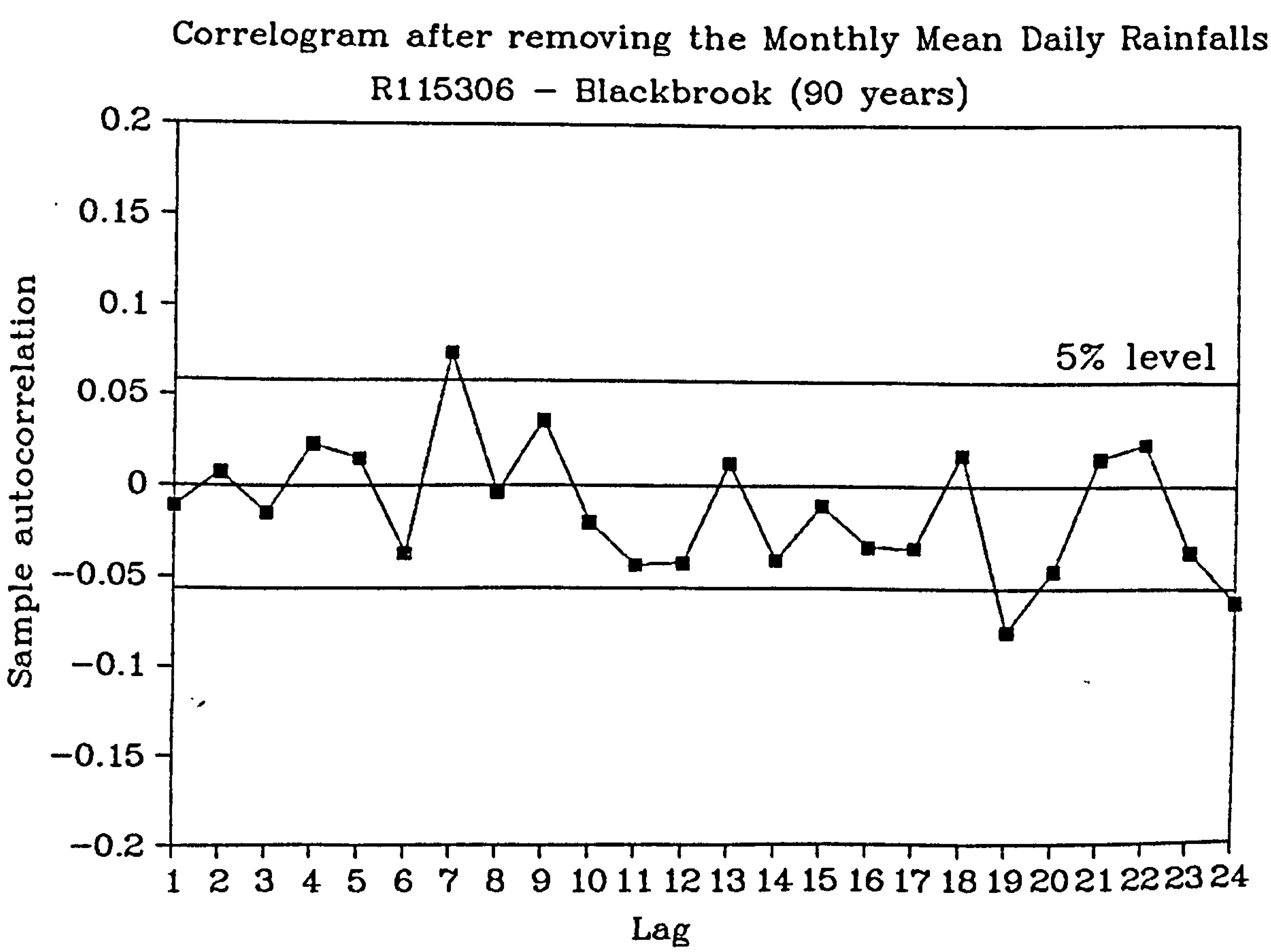


Figure J.4

Correlogram after removing Monthly Mean Daily Rainfalls

R588702 – Poaka Beck (90 years, NW)

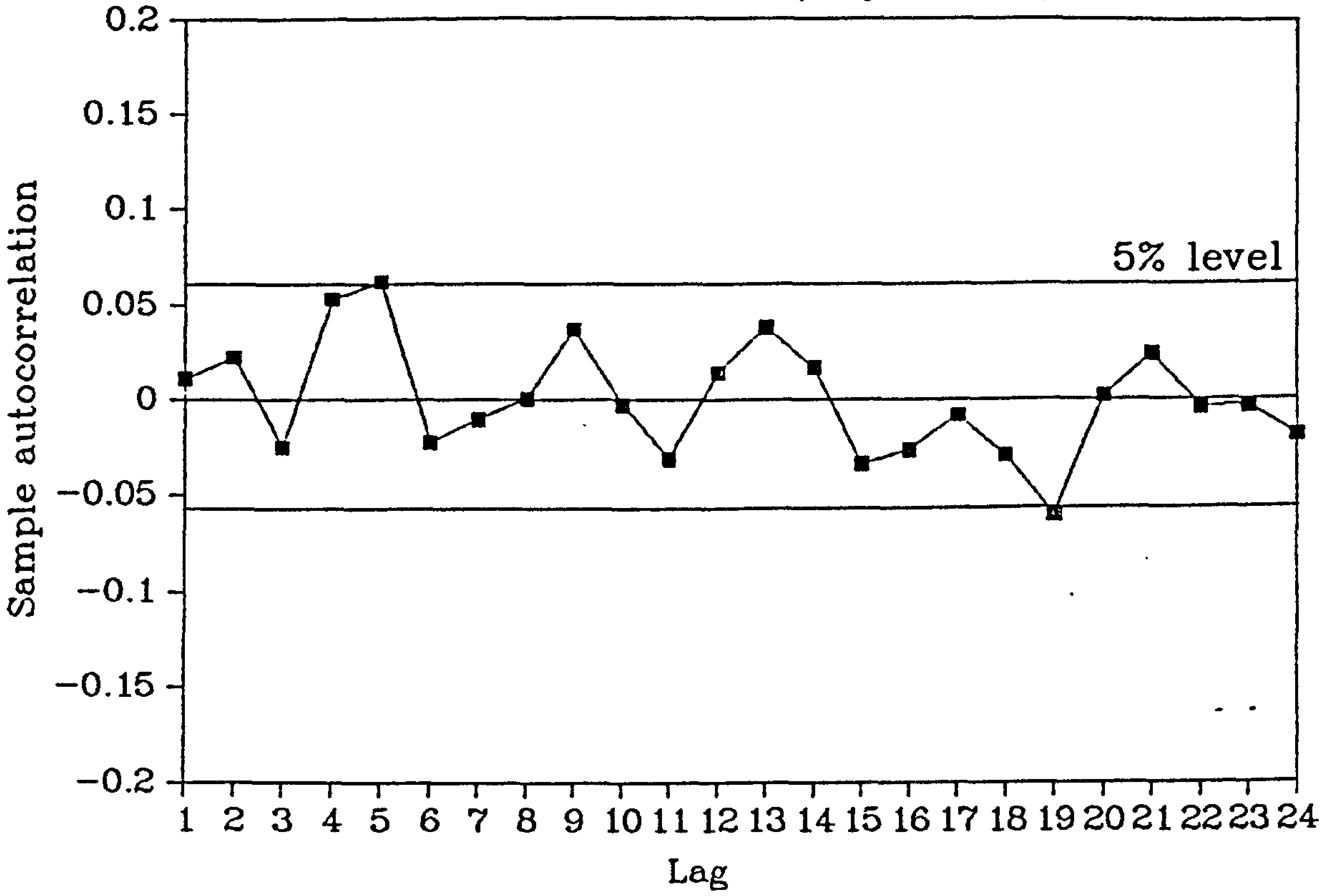


Figure J.5

Correlogram of Time Series of Monthly Mean Daily Rainfalls
(i.e. the result of NOT removing the Mean Daily Rainfalls)

R588702 – Poaka Beck (90 years, NW)

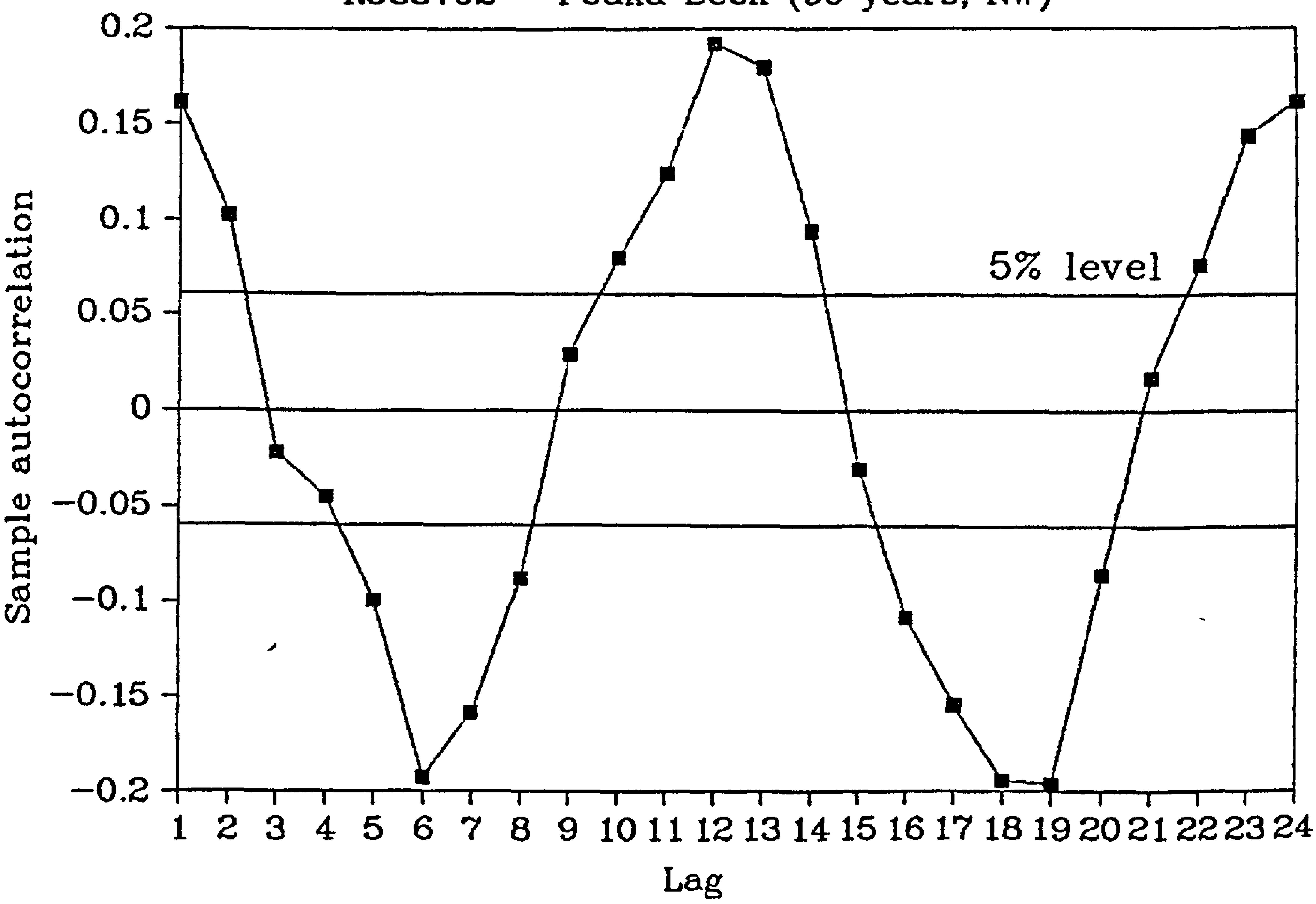


Figure J.6

APPENDIX K: THE EFFECT OF THE ERROR IN EXPRESSION (3.6)

After the thesis first submitted, an error was noticed in Chapter 3, Section 3.3. The error was in expression (3.6), which previously read:

$$p_t(h) = (1 - e^{-\beta t} + e^{-\beta(t+h)})(1 - \beta(e^{-\beta t} - e^{-\eta t})/(\eta-\beta)) \times \exp\left\{-\mu\beta(e^{-\beta t} - e^{-\eta t})/(\eta-\beta) - \mu e^{-\beta t} + \mu e^{-\beta(t+h)}\right\} \quad (3.6)$$

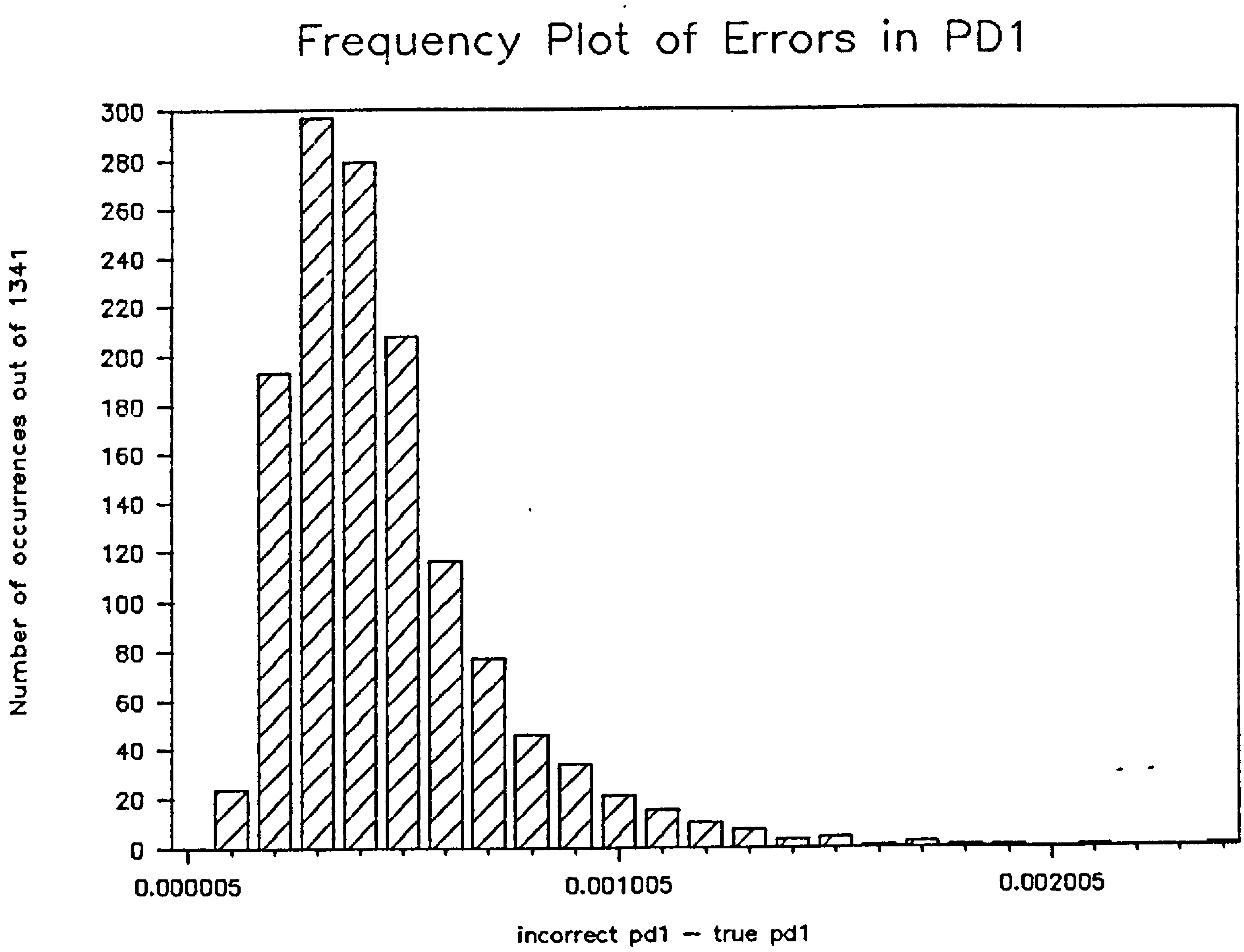
This has been replaced by:

$$p_t(h) = \left\{ e^{-\beta(t+h)} + 1 - (\eta e^{-\beta t} - \beta e^{-\eta t})/(\eta-\beta) \right\} \times \exp\left\{-\mu\beta(e^{-\beta t} - e^{-\eta t})/(\eta-\beta) - \mu e^{-\beta t} + \mu e^{-\beta(t+h)}\right\} \quad (3.6)$$

The incorrect expression above appeared in many of the programs used for the project. Therefore, comparisons were needed to see whether the consequences of this error would have any practical effect on the results presented in the thesis.

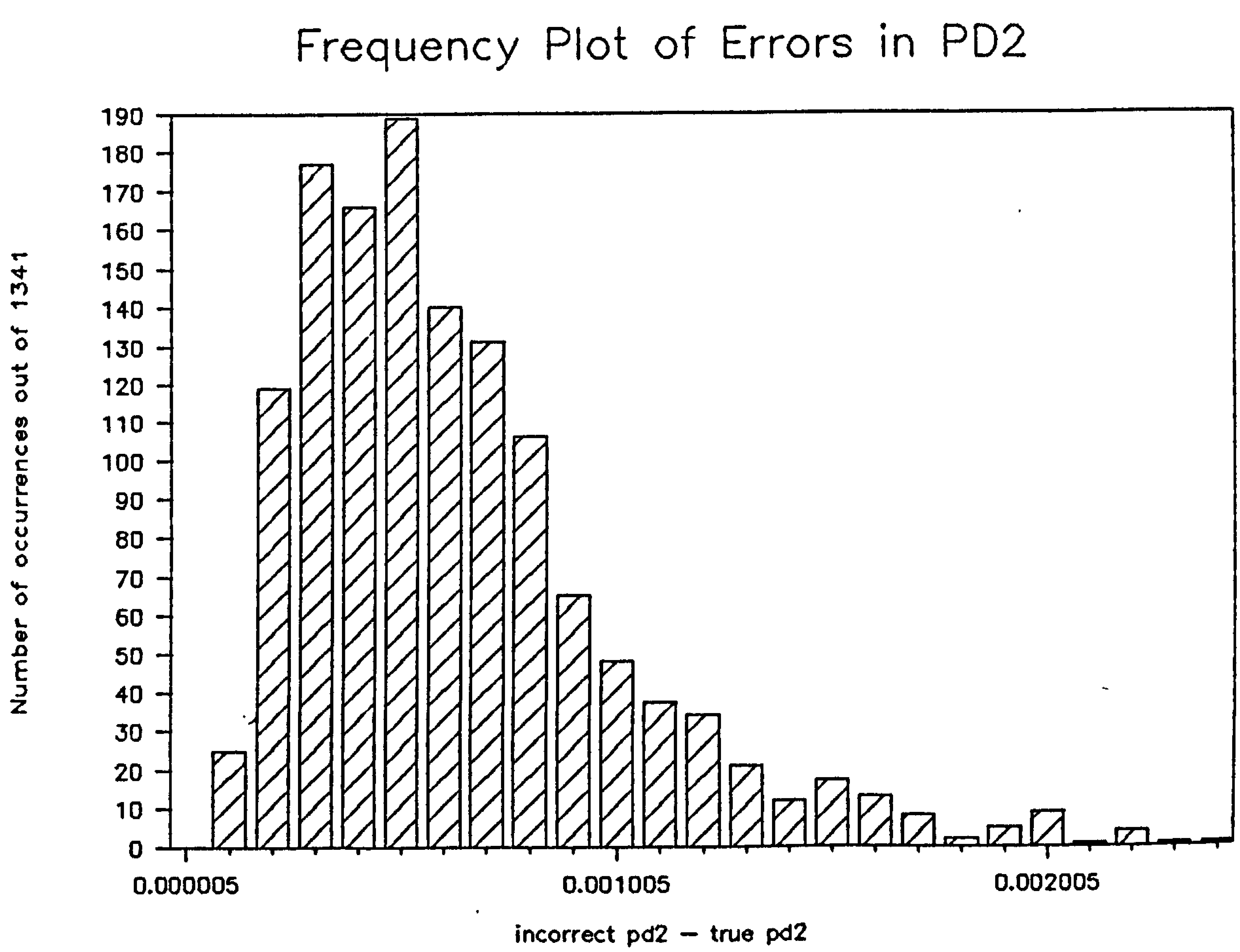
Expression (3.6) is needed in the expression for the probability that an arbitrary interval is dry (equations (3.9) and (4.4)) and is also needed in calculating the transition probabilities (equations (4.5) and (4.6)). In testing whether this error has any effect on the results described in this thesis, two levels of aggregation are considered: i) hourly and ii) daily. The effect of the error on the hourly level of aggregation can be assessed by evaluating $\phi(1)$ and $\phi(2)$ (equation (4.4)) using both the correct and incorrect version of (3.6), as these are used to calculate $\phi_{D|D}(1)$ and $\phi_{W|W}(1)$ (see equations (4.5) and (4.6)). Similarly, the effect of the error on the daily level of aggregation can be assessed by evaluating $\phi(24)$ and $\phi(48)$. These expressions were all evaluated for the 1341 parameter estimates used in the project, and the difference between the incorrect and correct value found. Figures K.1 - K.4 show frequency plots of these differences, where the notation PDh has been adopted for the proportion of dry intervals for a level of aggregation of h hours (evaluated using $\phi(h)$).

Looking through Figures K.1 - K.4 it can be seen that the effect of this error is at worst going to be in the third decimal place. Suppose we take 0.003 as the worst difference, and consider the smallest proportion of dry days in the Manston data set, which was 0.52 for January (see Table 4.2 of Chapter 4), then the worst percentage error is about 1% (when rounded up). Now the coefficient of variation for the proportion of dry days in a 10-year record is about 12% (see Table 4.1 of Chapter 4), which implies that the error is comparable to the sampling variability in a 10N-year record of rainfall data, where $1 = 12/\sqrt{N}$, i.e. a 1440-year record. The greatest station record length used in this project was 90 years. Hence, the effect of the error is regarded as small compared to the sampling variability of the rainfall data used in the project, and so can be neglected when interpreting the results in the thesis.



Frequency Plot of Errors in PD1
Number of occurrences out of 1341
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
0.000005
0.001005
0.002005
incorrect pd1 — true pd1

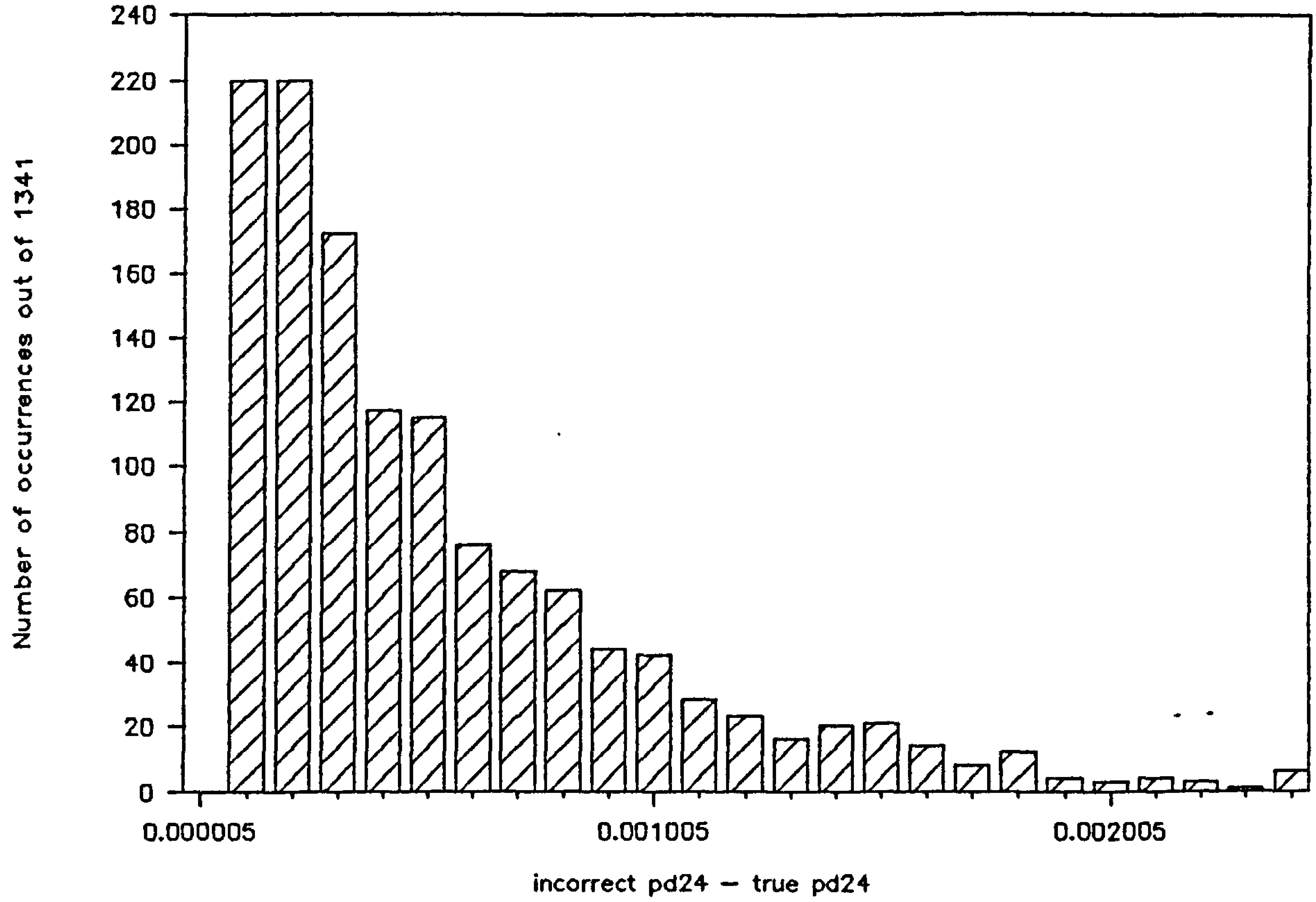
(K.1)



Frequency Plot of Errors in PD2
Number of occurrences out of 1341
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0.000005
0.001005
0.002005
incorrect pd2 — true pd2

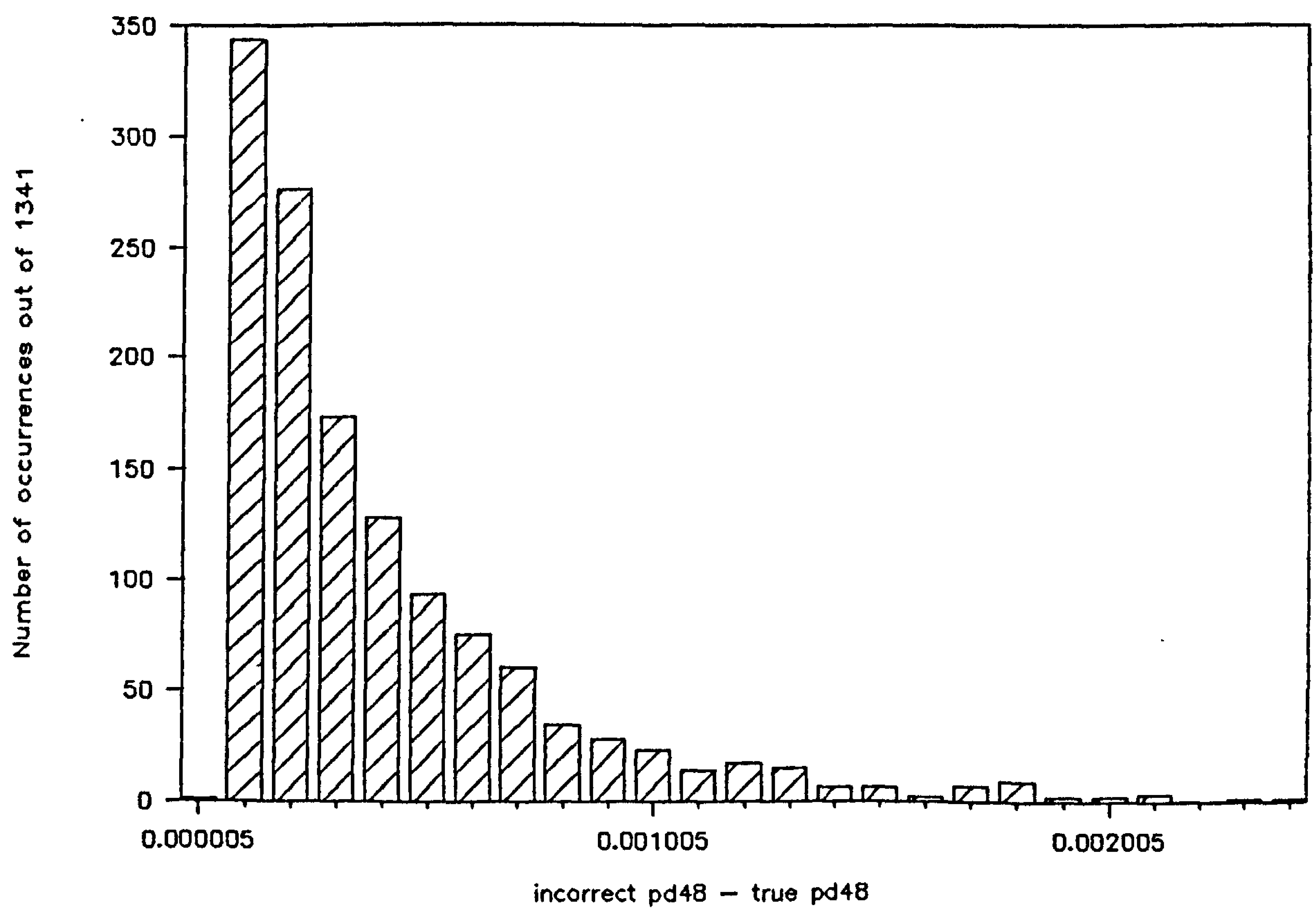
(K.2)

Frequency Plot of Errors in PD24



(K.3)

Frequency Plot of Errors in PD48



(K.4)

REFERENCES

Acreman M.C. (1990), A Simple Stochastic Model of Hourly Rainfall for Farnborough, England, Journal of Hydrological Sciences, 35, p119-148.

Amorocho J. and Wu B. (1977), Mathematical Models for the Simulation of Cyclonic Storm Sequences and Precipitation Fields, Journal of Hydrology, 32, p329-345.

Arnell V, Harremoes P, Jenson M, Johansen N.B. and Niemczynowicz J. (1984), Review of Rainfall Data Application for Design Analysis, Water Science and Technology, 16, Copenhagen, p1-45.

Aspin A.A. (1948), An Examination and Further Development of a Formula arising in the problem of comparing two mean values, Biometrika, 35, p88-96.

Breckling J. (1989), The Analysis of Directional Time Series: Applications to Wind Speed and Direction, Springer-Verlag, Berlin Heidelberg, Germany.

Calinski T. and Corsten L.C.A. (1985), Clustering Means in ANOVA by simultaneous testing, Biometrics, 41, p39-48.

Chatfield C. and Collins A.J. (1980), An Introduction to Multivariate Analysis, Chapman and Hall, London.

Cho H.R. (1985), Stochastic Dynamics of Precipitation: An Example, Water Resources Research, 21(8), p1225-1232.

Cho H.R. and Chan D.S.T. (1987), Mesoscale Atmospheric Dynamics and Modelling Rainfall Fields, Journal of Geophysical Research, 98(D8), p9687-9692.

Cowpertwait P.S.P. (1991), Further Developments of the Neyman-Scott Clustered Point Process for Modelling Rainfall, Water Resources Research, 27(7), p1431-1438.

Cox D.R. and Isham V. (1980), Point Processes, Chapman & Hall.

Cox D.R. and Spjotvoll E. (1982),On Partitioning Means into Groups, Scandinavian Journal of Sciences, 9, p147-152.

Cox T.F. and Jaber K.H. (1990), The Generalised Likelihood Ratio Test for the Behrens-Fisher Problem, Journal of Applied Statistics, 17(1), p63-71.

Crabtree R. (1988), Urban River Pollution in the UK: the WRc River Basin Management Programme, in: Hooke J.M. (Ed.), Geomorphology in Environmental Planning, p169-185 (John Wiley and Sons Ltd).

Dales M.Y. and Reed D.W. (1989), Regional Flood and Storm Hazard Assessment, Report No. 102, Institute of Hydrology, Wallingford, UK.

Eagleson P. (1978), Climate, Soil, and Vegetation, 2, The Distribution of Annual Precipitation Derived from Observed Storm Sequences, Water Resources Research, 14(5), p713-721.

Entekhabi D., Rodriguez-Iturbe I. and Eagleson P.S. (1989), Probabilistic Representation of the Temporal Rainfall Process by a Modified Neyman-Scott Rectangular Pulse Model: Parameter Estimation and Validation, Water Resources Research , 25, No. 2, p295-302.

Feller W. (1968) and (1971), An Introduction to Probability Theory and its Applications, John Wiley, New York, (Vol. 1 and Vol. 2).

Foufoula-Georgiou E. and Georgakakos K. (1988), Recent Advances in Space-Time Precipitation Modelling and Forcasting, Hato Asi on Recent Advances in the Modelling of Hydrologic Structures, Sintra, Portugal.

Foufoula-Georgiou E. and Guttorp P. (1987), Assessment of a Class of Neyman-Scott Models for Temporal Rainfall, Journal of Geophysical Research, 92(D8), p9679-9682.

Foufoula-Georgiou E. and Lettenmaier D.P. (1987), A Markov Renewal Model for Rainfall Occurrences, Water Resources Research, 23(5), p875-884.

Fisher R.A. (1935), The Fiducial Argument in Statistical Inference, Annals of Eugenices, 6, p391-398.

Gabriel K.R. (1959), The Distribution of the Number of Successes in a Sequence of Dependent Trials, Biometrika, 96, p454-460.

Gabriel K.R. and Neumann J. (1962), A Markov Chain Model for Daily Rainfall Occurrences at Tel-Aviv, Q.R.J. Meteorological Society, 88, p90-85.

Gradshteyn I.S. and Ryzhik I.M. (1965), Table of Integrals Series and Products, Academic Press.

Gregory S. (1975), On the Delimitation of Regional Patterns of Recent Climatic Fluctuations. Weather 30, 9, p276-287.

Gringorten I.I. (1963), A Plotting Rule for Extreme Probability Paper, Journal of Geophysical Research, 68, p813-814.

Hamlin M.J. (1983), The Significance of Rainfall in the Study of Hydrological Processes at Basin Scale, Journal of Hydrology, 65, p73-94.

Henderson R.J. (1989), Rainfall Time Series for Storm Overflow Assessment, Water Science Technology, 21, p1789-1791.

Henderson R.J. (1986), Rainfall Time Series for Sewer System Modelling, Report No. ER 195E, Water Research Centre, Swindon. UK.

Hershenhorn J. and Woolhiser D.A. (1987), Disaggregation of Daily Rainfall, Journal of Hydrology, 95, p299-322.

Holder R.L. (1985), Multiple Regression in Hydrology, Institute of Hydrology, Wallingford, Oxfordshire.

Jackson M.C. and Larke P.R. (1975), Frequencies of Specified Rainfall Amounts within Specified Elevation over England and Wales, Meteorological Office Hydrological Memorandum No. 41.

Jaber K.H. (1984), PhD thesis, University of Newcastle-Upon-Tyne.

James G.S. (1951), Tests of Linear Hypothesis in Univariate and Multivariate Analysis when the Population Variances are Unknown, Biometrika, 38, p19-43.

Kavvas M.L. and Delleur J.W. (1975), The Stochastic and Chronologic Structure of Rainfall Sequence: Application to Indiana, Water Resources Research Centre Technical Report, no. 57, Purdue University, West Layfayette, Indiana.

Kavvas M.L.,and Delleur J.W. (1981), A Stochastic Cluster Model of Daily Rainfall Sequences, Water Resources Research, 17, p1151-1160.

Krzanowski W.J. (1990), Principles of Multivariate Analysis - A User's Perspective, Oxford University Press.

Marien J.L. and Vandewiele G.L. (1986), A Point Rainfall Generator with Internal Storm Structure, Water Resources Research, Vol. 22, p475-482.

Mellor D. (1991), The MTB Model: Stochastic Spatial Rainfall Modelling, (unpublished paper submitted to the Journal of Hydrology).

Milly P.C.D. and Eagleson P.S. (1988), Effects of Storm Scale on Surface Runoff Volume, Water Resources Research, 24(4), pp620-624.

NAG (1977), Fortran Library Manual, NAG Executive, Oxford.

Natural Environmental Research Council (1975), Flood Studies Report, NERC.

Neyman J.E. (1939), On a New Class of Contagious Distributions Applicable in Entomology and Bacteriology, Ann. Math. Stat., 10, p35-57.

Neyman J. and Scott E.L. (1958), Statistical Approach to Problems of Cosmology, Journal of the Royal Statistical Society Series B, XX, No. 1, p1-43.

Nicks A.D. (1982), Space-Time Quantification of Rainfall Inputs for Hydrological Transport Models, Journal of Hydrology, 59, p249-260.

Ormsbee L.E. (1989), Rainfall Disaggregation Model for Continuous Hydrologic Modelling, Journal of Hydraulic Engineering, 115, p507-525.

Parzen E. (1967), Stochastic Processes, Holden-Day, San Francisco.

Quinn R. (1991), Extreme Value Analysis of Rainfall Time Series using the Neyman-Scott Model, MSc dissertation, Dept. Civil Engineering, University of Newcastle-Upon-Tyne.

Restrepo-Posada P.J. and Eagleson P.S. (1982), Identification of Independent Rainstorms, Journal of Hydrology, 55, p303-319.

Rodriguez-Iturbe I., Gupta V.K. and Waymire E. (1984), Scale Considerations in the Modeling of Temporal Rainfall, Water Resources Research, 20, p1611-1619.

Rodriguez-Iturbe I., Cox D.R. and Isham V. (1987a), Some Models for Rainfall based on Stochastic Point Processes, Proceeds of the Royal Society London (410), p269-288.

Rodriguez-Iturbe I., Febres De Power B., and Valdes J.B. (1987b), Rectangular Pulses Point Process Models for Rainfall: Analysis of Empirical Data, Journal of Geophysical Research, 92(D8), p9645-9656.

Scott A.J., and Knott M. (1974), A Cluster Analysis Method for Grouping Means in the Analysis of Variance, Biometrics, 30, p507-512.

Shaw S.R. (1983), An Investigation of the Cellular Structure of Storms Using Correlation Techniques, Journal of Hydrology, 62, p63-79.

Smith J.A. and Karr A.F. (1985), Statistical Inference for Point Process Models of Rainfall, Water Resources Research, 21, p73-79.

Smith E.R. and Schreiber H.A. (1973), Point Processes of Seasonal Thunderstorm Rainfall, Part 1, Distribution of Rainfall Events, Water Resources Research, 9(4), p871-884.

Stern R.D. and Coe R. (1984), A Model Fitting Analysis of Daily Rainfall Data, Journal of the Royal Statistical Society (Series A), 147, Part 1, p1-34.

Swartenbroekx P. (1987), The Use of the Point Generator (J.L. Marien, G.L. Vandewiele) on the C.D.C. Mainframe Computer, Iuphy-Document No. 15, University of Brussels.

Todorovic P. and Woolhiser D.A. (1976), Stochastic Structure of the local pattern of Precipitation, in Stochastic Approaches to Water Resources, Vol. 2, edited by H.W. Shen, Colorado State University, Fort Collins.

Todorovic P. and Yevjevich V. (1969), Stochastic Process of Precipitation, Hydrological Paper 35, Colorado State University, Fort Collins.

Valdes J.B., Rodriguez-Iturbe I. and Gupta V.K. (1985), Approximations of Temporal Rainfall From a Multidimensional Model, Water Resourses Research, 21, p1259-1270.

Water Research Centre/Water Authorities Association (1986), Sewerage Rehabilitation Manual, Second Edition.

Waymire, E. and Gupta V.K. (1981), The Mathematical Structure of Rainfall Representations: 1. A Review of Stochastic Rainfall Models, 2. A Review of the Theory of Point Processes, 3. Some Applications of the Point Process Theory to Rainfall Processes, Water Resources Research, 17(5), p1261-1294.

Waymire E., Gupta V.K. and Rodriguez-Iturbe I. (1984), A Spectral Theory of Rainfall Intensity at the Meso-β scale, Water Resources Research, 20(10), p1453-1465.

Welch B.L. (1947), Generalisation of Student's Problem when Several different Population Variances are involved, Biometrika, 34, p28-35.

Welch B.L. (1951), On the Comparison of Several Mean Values: An Alternative Approach, Biometrika, 38, p330-336.

Wigley T.M.L., Lough J.M. and Jones P.D. (1984), Spatial Patterns of Precipitation in England and Wales and a Revised Homogeneous England and Wales Precipitation Series, Journal of Climatology, 4, p1-25.

Wilson C.B, Valdes J.B. and Rodriguez-Iturbe I. (1979), On the Influence of the Spatial Distribution of Rainfall on Storm Runoff, Water Resources Research, 15(2), p321-328.

Yen B.C. and Chow V.T. (1980), Design Hyetographs for small drainage structures, Journal of Hydraulic Div, ASCE, 106(HY6), p1055-1076.