J. R. Statist. Soc. A (1984),
147, *Part* 1, *pp.* 1–34

A Model fitting Analysis of Daily Rainfall Data

By R. D. STERN and R. COE

University of Reading, UK

[*Read before the* Royal Statistical Society on *Wednesday, October 26th, 1983, the President*, Professor P. Armitage, *in the Chair*]

SUMMARY

This paper discusses the fitting and use of models for daily rainfall observations. Non-stationary Markov chains are fitted to the occurrence of rain, and gamma distributions, with parameters which vary with the time of year, are fitted to the rainfall amounts. Numerical methods are used to derive results from these models that are important in agricultural planning. Examples include the distributions of soil water content and lengths of dry spells. The process of fitting and using these models provides a straightforward and flexible analysis for rainfall records.

Keywords: AGRICULTURAL PLANNING; CHAIN DEPENDENT PROCESS; DRY SPELLS; FOURIER SERIES; GAMMA DISTRIBUTION; GENERALIZED LINEAR MODEL; MARKOV CHAINS; RAINFALL; WATER BALANCE

1. INTRODUCTION

Throughout the world considerable effort is devoted to the collection of rainfall data. Many long records exist and most countries now have a reasonably dense network of rainfall stations. The work on data collection is not, at present, matched by a corresponding effort on analysis. This paper demonstrates a method of analysis of daily rainfall records and shows how the models derived can be used to give results that are of direct use in agricultural planning. In many parts of the world the climate is the major constraint on agriculture and agricultural planning decisions must be made with this in mind. Although the various climatic variables interact with the crop in complex ways, rainfall is the limiting factor in most parts of the tropics.

In some studies (e.g. Woodhead *et al.*, 1970; Panabokke and Walgama, 1974) the only characteristics of rainfall that are considered are percentage points of the total rainfall in successive 7- or 10-day periods. While these can be a useful guide to which crops are viable, there are many other aspects of the rainfall pattern that are also important. For example, in many areas of the seasonally arid tropics, crops must be planted early and the date of the start of the growing season may coincide with the first heavy rainfall. Davy *et al.* (1976) observed that millet was often planted in Nigeria after the first occurrence of at least 20 mm of rain over a 2-day period. The distribution of the date of this event is therefore of interest. Crops will be at risk from dry spells occurring during the growing season. The level of the risk can sometimes be assessed by evaluating the probability that a long dry spell occurs when the plant is particularly sensitive, such as just after germination, or at flowering. A calculation of the chance of long dry spells through the season is therefore often useful.

Another agriculturally important feature of the rainfall pattern is the timing of the end of the wet season. If this occurs too soon the crop may not have sufficient water to reach maturity. However, excessive wet weather may prevent ripening or harvesting. In general, crops will use stored soil moisture for growth beyond the end of the rains, and so the end of the growing season is the date when the soil profile is too dry for growth to continue. This date can be evaluated by

Present address: Department of Applied Statistics, University of Reading, Whiteknights, Reading RG6 2AN.

 9254–0238/84/147001 $2.00

considering a water balance model with rainfall as input to the soil and evaporation (plus possibly runoff and drainage) as output. The evaporation used to evaluate this water balance may be calculated as a function of various climatic variables (Doorenbos and Pruitt, 1977), illustrating the importance for agriculture of variables other than rainfall. Unfortunately, at many sites, reliable records for other variables are available for only a few years. Hence many simple water-balance models (e.g. Cocheme and Franquin, 1967; Hills and Morgan, 1981) incorporate actual rainfall data in different years together with average evaporation figures. The probability distribution of characteristics of the water balance is then a reflection of the year to year variability of the rainfall.

Most studies of the characteristics of rainfall mentioned above have consisted simply of a summary of the observed data, with perhaps the subsequent assumption that the characteristic studied is normally distributed (e.g. Benoit, 1977; Archer, 1981). This type of analysis is discussed in detail by Stern *et al.* (1982). The main advantages of the method are simplicity and the lack of assumptions that are made. However, long records are required because, for each characteristic studied, each year of data provides just one observation. It is also difficult to study important conditional questions (such as the distribution of the end of the growing season given that the soil profile is full at the beginning of September) without either splitting the record into subsets or introducing some new assumption into the analysis.

Our primary objective in this paper is to demonstrate the viability of an alternative approach which, for a given site, consists of modelling the pattern of rainfall on a daily basis, followed by a calculation of the required results from the model. The type of model used consists of a non-stationary Markov chain to describe the occurrence of rain, and gamma distributions to describe the rainfall amounts. The process of fitting the models together with a brief review of some of the considerable literature is given in Section 2 of the paper. There is a much smaller literature on how such models can be used. Their use to answer some of the questions posed above is discussed in Section 3.

Data from a 53-year record for Morogoro (Tanzania—mean annual rainfall 900 mm) and a 37-year record for Irbid (Jordan—mean annual rainfall 430 mm) are used to illustrate the analyses.

2. FITTING MODELS TO THE DAILY RAINFALL DATA

The modelling of rainfall data has a large literature, reviewed by Waymire and Gupta (1981). Some of the models proposed (Le Cam, 1961; Kavvas and Delleur, 1981) are based on cluster processes in an attempt to incorporate some meteorological ideas. However, the majority of models are derived empirically, and meteorological significance is possibly attached to the fitted parameters. Early work concentrated on describing the distribution of wet or dry spell lengths (Lawrence, 1954; Cooke, 1953; Williams, 1952). The distributions fitted have been extended to give "good fits" at different sites (Green, 1970; Singh *et al.*, 1981) and built into comprehensive alternating renewal type models (Buishand, 1977). Markov chains were an obvious candidate to model the sequence of wet and dry days. Gabriel and Neumann (1962) used a first-order stationary Markov chain. The models have since been extended to allow for non-stationarity, both by fitting separate chains to different periods of the year (Caskey, 1963; Dumont and Boyce, 1974; Heerman *et al.*, 1968; Jackson, 1981) and by fitting continuous curves to the transition probabilities (Feyerherm and Bark, 1965; Woolhiser and Pegram, 1979). The order of Markov chain required has been discussed extensively (Lowry and Guthrie, 1968; Gates and Tong, 1976; Chin, 1977), the obvious conclusion being that different sites require different orders. It is this flexibility of the Markov chain models, as well as the ease with which parameters are estimated, that leads us to use them. Another advantage which Markov chains have over other models of rainfall occurrence is the ease with which results can be obtained from the fitted model without resorting to simulation.

Some authors have attempted to describe rainfall amounts by fitting Markov chains with many states each representing a range of amounts (Khanal and Hamrick, 1974; Haan *et al.*, 1976). One unsatisfactory element of these models has been the large number of parameters to be estimated.

The alternative is to model rainfall amounts on wet days separately. The distribution of these amounts is extremely skew, and a gamma distribution, or some modification of it, has often been used (Buishand, 1977; Katz, 1977).

The remainder of this Section describes the fitting of Markov chain models to the wet/dry sequence and gamma distributions to the amounts of rain on wet days.

2.1. *Fitting Models to the Occurrence of Rain*

We restrict attention for the present to two state Markov chains. The states are labelled "dry" and "rain", but the boundary point between the two may be above zero for some applications, or to avoid difficulties arising from the inconsistent recording of very small rainfall amounts. The models may be fitted to the whole year but, at some sites, there is little point in modelling the occurrence of rain during the dry season because the probability of rain is very small. The model is in general fitted to the T days of the year from day t_1 to t_T.

Let

$$J(t) = \begin{cases} 0 & \text{if day } t \text{ is dry,} \quad t = t_1, \ldots, t_T, \\ 1 & \text{if day } t \text{ has rain.} \end{cases}$$

Attention is restricted to first- and second-order Markov chains but everything follows easily for higher order chains. The assumption that $J(t)$ forms a second-order Markov chain is the assumption that

$$P[J(t) = 1 \mid J(t-1), J(t-2), J(t-3) \ldots] = P[J(t) = 1 \mid J(t-1), J(t-2)], \quad t = t_1, \ldots, t_T$$

and fitting the Markov chain model involves estimating the $4T$ parameters

$$p_{hi}(t) = P[J(t) = 1 \mid J(t-1) = i, J(t-2) = h], \quad t = t_1, \ldots, t_T.$$

The numbers of transitions are sufficient statistics for $p_{hi}(t)$ so the data may be reduced to the $2 \times 2 \times 2 \times T$ table with entries

$$n_{hij}(t) = \text{Number of days with } J(t) = j, \quad J(t-1) = i, \quad J(t-2) = h, \quad t = t_1, \ldots, t_T.$$

Day 60 (February 29th) has data only in leap years so day 59 precedes day 61 in non-leap years and is used to calculate $n_{hij}(61)$; similarly, day 366 of the previous year precedes day 1. The obvious estimates of $p_{hi}(t)$ are the observed proportions

$$r_{hi}(t) = n_{hi1}(t)/n_{hi+}(t), \quad h, i = 0, 1, \quad t = t_1, \ldots, t_T,$$

where + indicates summation over the subscript.

The usual analysis of Markov chains assumes stationarity, that is

$$p_{hi}(t) = p_{hi}, \quad t = t_1, \ldots, t_T,$$

and little is said about how to proceed if the process is not stationary. For many parts of the world the assumption of stationarity is not appropriate even for periods as short as 1 month. For example, Fig. 1 shows $r_{hi}(t)$, pooled over 5 days for clarity, plotted against t for the data from Morogoro. It is clear that, even during the part of the year when the probability of rain is high (January to May), the process is not stationary. Seeing the data in this form immediately suggests fitting curves to model the $p_{hi}(t)$ through the year.

The $r_{hi}(t)$ are proportions, so one approach to curve fitting would be to transform to approximate normality and use standard regression methods (Feyerherm and Bark, 1965). An exact method is available, however. The log-likelihood is

$$l = \sum_{t=t_1}^{t_T} \sum_{h,i} [n_{hi1}(t) \log (p_{hi}(t)) + n_{hi0}(t) \log (1 - p_{hi}(t))].$$

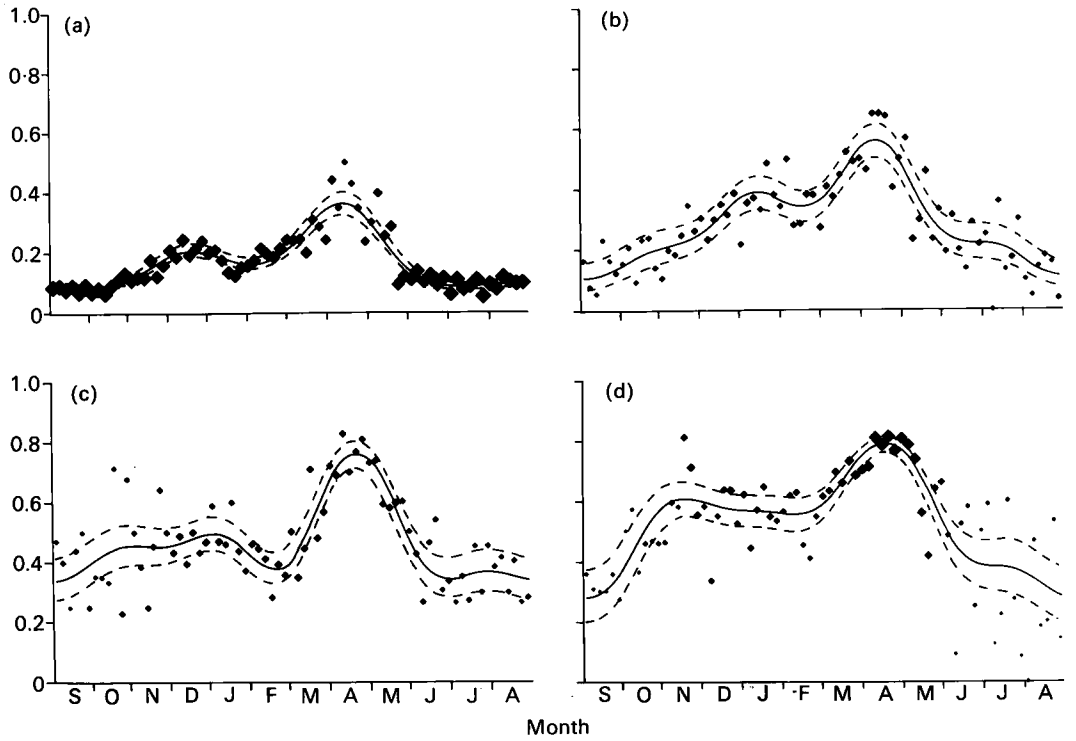

Fig. 1. Morogoro, Tanzania (data from 1928 to 1981): observed ($r_{hi}(t)$) and fitted ($\hat{p}_{hi}(t)$) proportions of rainy days, second-order Markov chain, with approximate 95 per cent confidence limits for $p_{hi}(t)$. (a) $(h, i) = (0, 0)$. (b) $(h, i) = (1, 0)$. (c) $(h, i) = (0, 1)$. (d) $(h, i) = (1, 1)$. Observed points pooled over 5 days, area proportional to number of observations.

This is the same as the kernel of the log-likelihood for a multinomial model for a $2 \times 2 \times 2 \times T$ contingency table or for $4T$ independent binomial responses, where $n_{hi1}(t) \sim B(n_{hi+}(t), p_{hi}(t))$. The binomial approach is more efficient, because the number of data points is half that of the other approach, as rain days alone are considered as the response.

The model of time response used is

$$p_{hi}(t) = h(g_{hi}(t)),$$

where h is a known link function connecting the probabilities, $p_{hi}(t)$, to the function $g_{hi}(t)$ which is linear in unknown parameters. Since the binomial distribution is a member of the exponential family fo distributions, the model is a generalized linear model (Nelder and Wedderburn, 1972) and hence the maximum likelihood estimates can be obtained easily. The package GLIM (Baker and Nelder, 1978) was used for fitting all the models.

The link function h is taken to be the logit,

$$p_{hi}(t) = \exp(g_{hi}(t))/[1 + \exp(g_{hi}(t))],$$

which ensures that the estimates of p lie between 0 and 1. An alternative approach is to use an identity link function and find maximum likelihood estimates of the parameters in g while constraining p to lie in the required range (Woolhiser and Pegram, 1979). This considerably increases the computational complexity and prevents standard methods from being used.

The function $g_{hi}(t)$ may take many different forms and as a routine we use Fourier series,

$$g_{hi}(t) = a_{hi0} + \sum_{k=1}^{m} [a_{hik} \sin(kt') + b_{hik} \cos(kt')], \quad h, i = 0, 1, \tag{2.1}$$

where $t' = 2\pi t/366$. These have the desirable properties of modelling complex bimodal rainfall patterns with few parameters and, where the whole year is modelled ($T = 366$), of being continuous from day 366 to day 1.

The required number of harmonics, m, may be decided using multiple regression techniques in which the explanatory variables enter in a fixed order. Models with increasing values of m are fitted successively until no improvement in fit is gained by including additional terms. Maximum likelihood estimation is used, so likelihood ratio tests may be used to assess the increase in goodness of fit. Thus for each model the deviance, G^2_m (Baker and Nelder, 1978) is calculated and the difference in deviances, $G^2_m - G^2_{m+1}$, measures the effect of including the $(m+1)$th harmonic.

TABLE 1

Deviances and degrees of freedom (df) for curves with m harmonics fitted to data from Morogoro calculated on a daily and 5-day basis

		Daily *h, i*					*Five day* *h, i*			
m	d.f.	0, 0	1, 0*	0, 1	1, 1*	d.f.	0, 0	1, 0	0, 1	1, 1
0	365	725.1	598.1	586.5	680.0	72	456.5	242.8	220.7	327.0
1	363	447.0	457.5	540.3	555.3	70	180.5	103.3	172.8	205.1
2	361	404.7	445.8	485.3	470.7	68	138.6	91.6	114.9	120.8
3	359	336.9	436.2	457.6	462.1	66	72.3	82.1	88.2	112.2
4	357	336.1	427.1	447.6	453.9	64	71.7	72.8	77.5	104.8
5	355	331.9	425.9	446.9	451.8	62	68.1	71.0	77.1	101.9

* Degrees of freedom for these two curves are reduced by 2.

Fourier series have been fitted to the data from Morogoro; the deviances obtained are shown in Table 1. The reductions in deviance may be compared with appropriate χ^2 distributions; the results indicate that three harmonics are required for $(h, i) = (0, 0)$ and four harmonics for the other three curves.

A formal goodness-of-fit test of the final selected model is available, as G^2 itself has an approximate χ^2 distribution if the model is correct. Except for the (0, 0) curve, the values of G^2 for the selected models are all "significantly" large. However, the distributional properties of G^2 are known only for large samples. It is not clear what this means in practice, but the asymptotic distribution of G^2 is known to underestimate the actual distribution when the expected frequency is less than 1 in some cells (Fienberg, 1980), as in the present example. The Pearson X^2 statistic has the same asymptotic distribution as G^2 but is less sensitive to this problem. The values of X^2 for the selected models (Table 2) are all lower than G^2 and indicate adequate fits.

TABLE 2

Deviance (G^2), Pearson's X^2 and degrees of freedom (d.f.) for curves with m harmonics fitted to data from Morogoro

h, i	0, 0	1, 0	0, 1	1, 1
m	3	4	4	4
G^2	336.9	427.1	447.6	453.9
X^2	330.4	385.2	390.1	390.6
d.f.	359	355	357	355

The problem of finding suitable measures of goodness-of-fit when there are small observations may be removed by pooling the data. Although this is not recommended for many contingency tables it is useful in the present example. If data from days t_k to t_l are pooled, this assumes $p_{hi}(t)$ is constant for $t = t_k, \ldots, t_l$. Since the probabilities usually vary slowly through the year, this is not unreasonable if $(l-k)$ is "small". We find it convenient to pool the data over 5-day periods throughout the year but there is no reason why the pooling should not be into groups of variable size. Thus towards the ends of the rainy season, when $n_{11+}(t)$ tends to be small, larger groups would be used for $p_{11}(t)$ than at the height of the rainy season. The deviances have been calculated for the Morogoro data pooled over 5 days (Table 1). The reductions in deviance and the fitted values are very close to those obtained from the original data and this is the case in all examples we have looked at. The pooling has another advantage: the computation time for fitting is less as the number of data points is reduced.

The four fitted curves, $\hat{p}_{hi}(t)$, are plotted in Fig. 1. Approximate 95 per cent confidence limits for $p_{hi}(t)$ are also plotted, based on the asymptotic standard error of $\hat{g}(t)$.

As in any regression problem, the goodness-of-fit of the model may be assessed by comparing the observed and fitted values. However, simply superimposing the fitted curves on a plot of $r_{hi}(t)$ is not sufficient. The $n_{hi}(t)$ are not constant, so the observed values $r_{hi}(t)$ have different weights when curves are fitted; this must be borne in mind when looking at the plots. A simple method is to plot $r_{hi}(t)$ with the area of the symbols proportional to $n_{hi}(t)$ (Fig. 1). This shows that although there is considerable scatter of points around the fitted lines when $h = 1$, the points far from the line have small values of $n_{hi+}(t)$. For example, for the 5-day group around August 2nd, $r_{11} = 0$ but n_{11+} is only 7. At times of the year when the values $n_{11+}(t)$ are small and hence the estimate of $p_{11}(t)$ has a large standard error, the overall probability of rain is small and so $p_{00}(t)$, which is estimated well, will dominate most calculations using the fitted model.

A more objective approach to assessing the goodness-of-fit is to calculate standardized residuals,

$$\sqrt{n_{hi+}(t)} \cdot [r_{hi}(t) - \hat{p}_{hi}(t)] / \sqrt{[\hat{p}_{hi}(t)\,(1 - \hat{p}_{hi}(t))]}$$

which asymptotically have a standard normal distribution. However we do not find these particularly useful here. Trends in these residuals, indicating some bias in the model, can be spotted equally easily in plots such as Fig. 1. It is difficult to interpret "significantly" large residuals as it is not clear whether the degree of lack of fit indicated is important in terms of the fitted model. Genuine outliers will be very rare as they will be present only if an error in the data or freak rain event occurs on the same day of most years.

In the discussion above a second-order Markov chain was fitted, but it is of interest to determine whether a Markov chain of lower order is adequate. A first-order Markov chain may be considered as a second-order chain in which

$$p_{0,i}(t) = p_{1,i}(t) = p_i(t), \quad i = 0, 1, \quad t = t_1, \ldots, t_T.$$

A test of this hypothesis against the alternative $p_{0,i}(t) \neq p_{1,i}(t)$ is just a test of whether two curves fit the data as well as four curves—a standard comparison of regressions problem. As before, likelihood ratio tests are used. For Morogoro the analysis confirmed that a second-order chain was necessary.

An important aspect of the analysis is the wide range of models that can be fitted and the ease with which restrictions can be made on the transition probabilities in order to minimize the number of parameters being estimated. Coe and Stern (1982) gave an example of fitting parallel curves, on the logit scale, to data for a second-order chain. Two further examples are given in Fig. 2. At some sites second-order chains are required. However, models with three rather than four curves may prove adequate with the probability of rain depending only on the state of the previous day if it was rainy. This idea may be extended to models with order greater than two. At Nagpur (India), Fig. 2a, the model finally adopted was a fifth-order chain, which required only the six curves $p_1(t)$, $p_{10}(t), \ldots, p_{10000}(t)$ and $p_{00000}(t)$ where, for example, $p_{100}(t)$ = probability of rain following exactly two dry days. The further constraint that all the curves

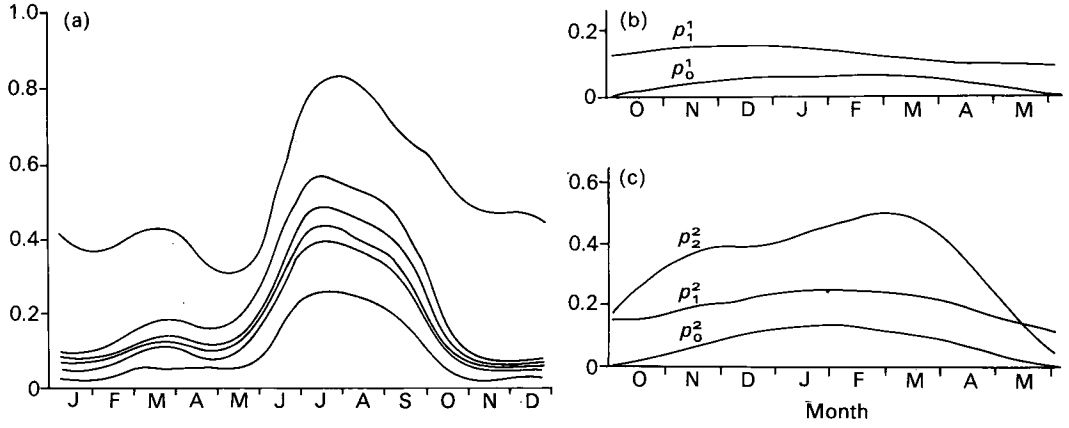

Fig. 2. Examples of other Markov chain models. (a) Nagpur, India (data from 1891 to 1955): fitted proportions of rainy days, restricted fifth-order Markov chain. From top down: $\hat{p}_1$, $\hat{p}_{10}$, $\hat{p}_{100}$, $\hat{p}_{1000}$, $\hat{p}_{10000}$, $\hat{p}_{00000}$. (b) Irbid, Jordan (data from 1937 to 1974): fitted probabilities of trace, three-state, first-order Markov chain. (c) Irbid, Jordan: fitted probabilities of rain, three-state, first-order Markov chain.

except $p_1(t)$ were parallel on the logit scale also proved to be reasonable. With these restrictions the model plotted in Fig. 2a required only 22 parameters to be estimated. In Fig. 2b and 2c the model for Irbid (Jordan) is given that will be used later in the paper. Here a first-order chain was adequate but 3 states were required, and a third state, "trace", was taken as all rainfall amounts less than 2.5mm, including those values actually recorded as "trace" in the data. The model adopted was of the three curves denoted $p_2^2(t)$, $p_1^2(t)$ and $p_0^2(t)$ for the probabilities of rain together with two curves $p_2^1(t) = p_1^1(t)$ and $p_0^1(t)$ for the probabilities of "trace", where the indices 0, 1, 2 refer to dry, trace and rain respectively.

2.2. *Modelling Rainfall Amounts*

The model for rainfall amounts must describe the distribution of rainfall on days when rain occurs. This distribution may depend on the time of year and also on what has occurred on previous days. Simple models will be treated initially with possible extensions considered later. Let $X(t)$ be the amount of rain on day t when $J(t) = 1$. $X(t)$ is undefined when $J(t) = 0$. The distribution of $X(t)$ is highly skewed and gamma distributions have been found to fit well. In practice a shifted gamma distribution is fitted as amounts less than some small value (e.g. 0.1 mm) are never recorded. If a lower limit greater than this has been used in the definition of $J(t)$ then it is assumed that the same limit is used here.

Let $X'(t) = X(t) - \delta$, with observations $x_j(t)$, $j = 1, \ldots, n(t)$, where δ is some lower limit and $n(t)$ is the number of years in which day t had rain. The distribution of $X'(t)$ is then taken as the gamma with density function

$$f(x) = (\kappa/\mu(t))^{\kappa} x^{\kappa-1} \exp\left[-\kappa x/\mu(t)\right]/\Gamma(\kappa).$$

$E(X'(t)) = \mu(t)$ and the time dependence is taken to be of the form $\log(\mu(t)) = g(t)$. If $g(t)$ is linear in unknown parameters then this model is again a generalized linear model. Fourier series (2.1) are used as a routine. The methods of estimating parameters and assessing the goodness-of-fit are essentially the same as when modelling the probability of rain but are complicated by the second parameter, κ, the shape parameter of the gamma distribution.

It is helpful to compare the model considered here with regression models fitted to normally distributed data. The variance in such models is usually assumed to be constant. This assumption may be tested and the analysis modified if necessary. The analogy in the gamma model is the assumption that the coefficient of variation $(1/\sqrt{\kappa})$ is constant for all values of t. The $n(t)$

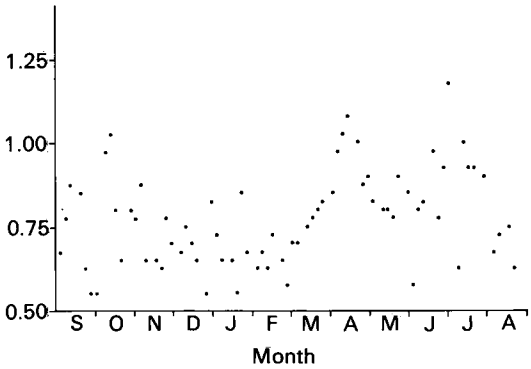

Fig. 3. Morogoro, Tanzania: estimated value of κ for each 5-day group.

repeated observations on each day (or 5-day group) mean that an estimate of κ is available for each value of t. In Fig. 3 the values have been plotted for Morogoro. The increased variance in the estimates outside the main wet season is due to the smaller values of $n(t)$ at these times. There is however also an indication that κ may be slightly larger during April, May and June than in the rest of the year. This complication, which is not a feature of many of the sites we have analysed, is ignored intially.

The primary aspect of interest when fitting Fourier series to μ is how many harmonics, m, are required. Models are again compared by considering reductions in deviance; null distributions are approximately multiples of χ^2 with $\kappa \cdot \text{deviance} \sim \chi^2$. The parameter κ is unknown so, by analogy with the analysis of variance, "mean deviances" are calculated as the deviance/d.f. The ratio of two mean deviances is then taken as having an approximate F distribution (Baker and Nelder, 1978). The denominator should refer to a model known to fit well so that the "pure error" deviance is used, calculated as the deviance that would be obtained by fitting a separate mean to each value of t. Table 3 gives the analysis of deviance for assessing the value of m for the data from Morogoro. The total number of observations, $N = \Sigma\, n(t)$, is often large, 5476 in this example. In practice therefore the data are conveniently summarised by the sufficient statistics $\Sigma\, x_j(t) = n(t)\bar{x}(t)$ and $\Sigma \log x_j(t)$. The function $g(t)$ is fitted to the $\bar{x}(t)$ by treating them as gamma variates with weights $n(t)$. In Table 3 the data have also been pooled over 5-day periods. The interpretation is the same as for ordinary ANOVA tables and two harmonics are required to model the mean rain per rainy day. Fitted values together with approximate 95 per cent confidence limits are given in Fig. 4.

TABLE 3

Analysis of deviance for testing the effect of increasing the number of harmonics fitted to the mean rain per rainy day at Morogoro

Source	*d.f.*	*Deviance*	*Mean deviance*	*F*
Between day	72	913.5		
First harmonic	2	675.0	337.5	214.3
Second harmonic	2	108.0	54.0	34.3
Third harmonic	2	3.4	1.7	1.08
Fourth harmonic	2	5.1	2.6	1.65
Residual	64	122.0	1.906	1.210
Within day	5403	8508.0	1.575	
Total	5475	9421.5		

The dependence of $\mu(t)$ on $J(t-1)$, $J(t-2)$ or earlier values may be introduced into the model. Let $\mu_{hi}(t) = E[X'(t) \mid J(t-1)=i, J(t-2)=h]$ and let $\bar{x}_{hi}(t)$ be the corresponding observed mean rainfalls. Four separate curves may be fitted to $\bar{x}_{hi}(t)$. If $\mu_{0i}(t) = \mu_{1i}(t)$ then two curves will fit as well as four and if $\mu_{hi}(t) = \mu(t)$ then just one curve is required. Thus once again standard comparison of regression techniques may be used. Table 4 shows such an analysis for the data from Morogoro, with the conclusion that one curve is sufficient, that is, the mean rain per rainy day seems not to depend on the state of previous days. This result is not typical. At many sites $\mu_{h1}(t) > \mu_{h0}(t)$ for most of the year.

TABLE 4

Analysis of deviance for testing dependence of $\mu(t)$ on $J(t-1)$ and $J(t-2)$ at Morogoro. Each curve is a two-harmonic Fourier series

Source	*d.f.*	*Deviance*	*Mean deviance*	*F*
Between day	289	1344.0		
One curve	4	782.5	195.6	125.6
Two curves	5	12.5	2.5	1.6
Four curves	10	16.5	1.7	1.1
Residual	270	532.5	1.972	1.27
Within day	5185	8075.8	1.558	
Total	5474	9382.3		

As with the probability of rain, there are three related approaches to assessing goodness of fit. A graphical comparison of the observed and fitted values (Fig. 4) is probably the most useful. The values $n(t)$ are very much larger towards the middle of the rainy season, with relatively few observations to estimate the shape of the curve at the beginning and end. This pattern of replication is unfortunate as accurate modelling of rainfall amounts at the beginning and end of the season is particularly important. Standardized residuals may be defined as $\sqrt{n(t)}\,[\bar{x}(t) - \hat{\mu}(t)]/\hat{\mu}(t)$ which have an asymptotic $N(0, 1/\kappa)$ distribution. A plot of these could also be used to check for the constancy of κ. However, this plot must be less sensitive than the method used above, being based on between-day rather than within-day information. The third approach is to compare the residual between-day deviance with the "pure error" within-day deviance. The F-value calculated is often small but "significant" when compared with the appropriate F distribution (as in Table 3). This is not surprising considering the large number of degrees of freedom. We are also assuming the F distribution is a good approximation to the actual

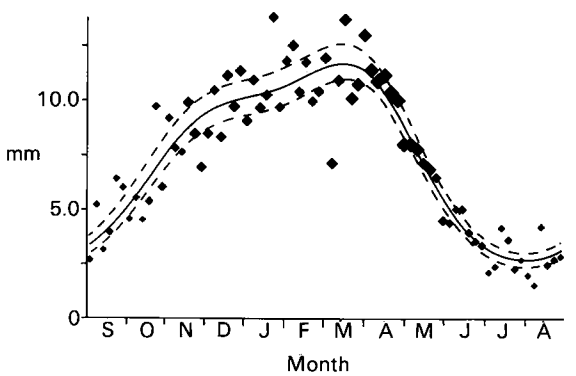

Fig. 4. Morogoro, Tanzania: observed and fitted mean rain per rainy day, with estimated 95 per cent confidence limits for the fitted mean. Observed values plotted with area proportion to $n(t)$.

distribution of this statistic. A third reason for this F value being large is inaccuracies in the data. The within-day deviance is calculated as

$$D^2 = 2 \Sigma n(t) [\log \bar{x}(t) - \overline{\log x}(t)] \quad \text{where } \overline{\log x}(t) = [\Sigma \log x_j(t)]/n(t).$$

In practice small rainfall values are recorded with a large relative error. These inaccuracies usually result in $\overline{\log x}(t)$ usually being too large and hence D^2 being too small. Thus the F ratios with D^2 as denominator will often be inflated.

An estimate of κ is required to complete the model and maximum likelihood estimation is used. If the problem of the inaccurate recording of small amounts can be ignored then the repeated observations for each value of t allow a "pure error" estimate of κ to be calculated. This is the solution of the equation

$$\log \kappa - \psi(\kappa) = D^2/2n, \tag{2.2}$$

where $n = \Sigma n(t)$ and $\psi(\cdot)$ is the digamma function (Abramowitz and Stegun, 1968). Tables giving the solution of this are available or a rational approximation (Greenwood and Durand, 1960) may be used. The mean within-day deviance, $D^2/n - T$, may be taken as a crude estimate of the scale parameter, $1/\kappa$ (Baker and Nelder, 1978) on the grounds that for a well-fitting model $\kappa \cdot D^2 \sim \chi_{n-T}$. For normally distributed data this estimate of the scale parameter is the maximum likelihood estimate of σ^2 adjusted for bias, the true maximum likelihood estimate being D^2/n. Here the maximum likelihood estimate of κ is also known to be biased. The bias may be estimated if $\mu(t)$ is constant (Shenton and Bowman, 1977) but is not known otherwise. Thus it seems reasonable to modify the right-hand side of (2.2) to $D^2/2(n-T)$. Small simulation studies support this but more work is needed. Secondly, taking $1/\kappa = D^2/n$ is equivalent to approximating the left-hand side of (2.2) by $1/2\kappa$. The approximation $\psi(\kappa) = \log \kappa - 1/2\kappa$ is poor for the small values of κ that we are dealing with (Abramowitz and Stegun, 1968).

The under-estimation of D^2 due to the inaccuracies in recording small amounts results in an over-estimation of κ. This effect may be removed in one of two ways. An estimate of κ based on between-day information should be robust to these inaccuracies. The maximum likelihood estimate for the between-day estimate is

$$\Sigma n(t) [\log (n(t)\kappa) - \psi(n(t)\kappa)] = D_b^2/2,$$

where D_b^2 is the residual between-day deviance. If $n(t)\kappa$ is large (> 10) then this may be approximated as $\kappa^* = T/D_b^2$. This is reasonable at Morogoro as the smallest value of $n(t) = 25$. A bias-reduced estimate is thus d/D_b^2, where d is the residual between-day degrees of freedom. This is just the intuitive estimate of the scale parameter of any generalized linear model. Any between-day estimate of κ would, however, be influenced by lack of fit in the model for $\mu(t)$. Censoring may be used to obtain a robust estimate of κ based on within-day information. The nunber of observations less than some censoring point ϵ is used for the estimation but not their actual values. The value of ϵ chosen will depend on the data, and 1.95 mm has been used for Morogoro.

The estimates of κ obtained for Morogoro (Table 5) show that the between-day and censored values are similar and smaller than the maximum likelihood estimate. The between-day estimate is easiest to obtain and is used as a routine. All the estimates are considerably less than 1, so the exponential distribution, used by several authors for its algebraic simplicity (Todorovic and Woolhiser, 1975), would not be suitable. This is so for data at most of the places that we have analysed.

TABLE 5
Estimates of κ for Morogoro

Estimation method	*Estimate*
Maximum likelihood	0.768
Bias-reduced between day	0.521
Censored	0.596

3. USE OF THE MODELS

The model for a given site usually involves between 20 and 50 parameters which define the equation of the curves for the probabilities of rain, and the parameters of the distribution of amounts of rainfall. These parameters, or summary statistics derived from them, may then be used to assess the pattern of rainfall at the site. For example, in the analysis of a transect of 11 stations in West Africa, Garbutt *et al.* (1980) defined the "average length of the rains" as the period during which the fitted curves for the overall probability of rain was greater than 0.2. With this simple definition the length ranged from 45 days at stations close to the Sahara, to 237 days at stations on the coast.

In many situations, however, the models should be seen as an intermediate step and we consider here the aspects of the pattern of rainfall that were discussed earlier in relation to agricultural planning. It is straightforward to derive any required results from the models by first simulating many "years" of data (Stern and Coe, 1982) and similar models to those described here have been used in a number of simulation studies of water balance or crop growth in which rainfall is one of the inputs (e.g. Jones *et al.*, 1970). This section demonstrates that many important characteristics can be studied without resorting to simulation.

3.1. *The Recurrence Relations*

The pair $(J(t), X(t))$ is a chain dependent process. Various properties of these processes have been investigated when they are stationary (e.g. O'Brien, 1974) but little can be done in general. In the context of rainfall, Gabriel and Neumann (1962) gave a formula for the distribution of the number of rainy days in a period of a given length for a first-order, two-state stationary Markov chain. Their formula is not simple even for this straightforward model but Katz (1974) shows how their result can easily be derived using simple recurrence relations. The recurrence relations for the longest dry spell in a period were given by Sen (1980).

Similar recurrence relations may be derived for calculating the distribution of other characteristics. In most cases it is convenient to consider a bivariate Markov chain $(Y(t), J(t))$, where $J(t)$ is defined as in Section 2. The states of $Y(t)$ depend on the rainfall characteristic being examined; they are ordered with the first and last states being barriers, which may be absorbing of reflecting. Attention is restricted to where $J(t)$ is a first-order, two-state Markov chain. Extension to higher order or more states is straightforward. Lloyd (1979) showed how to convert a bivariate chain to a univariate chain to enable recurrence relations for the joint distribution to be obtained. Computationally it is often convenient to keep the bivariate form. Then

$$P[J(t)=j,\ Y(t)=l] = \sum_k \sum_i s_{ijkl}\, P[J(t-1)=i,\ Y(t-1)=k],$$

where $s_{ijkl} = P[J(t)=j,\ Y(t)=l \mid J(t-1)=i,\ Y(t-1)=k]$. Transitions between states of J depend only on the transition probabilities $p_i(t) = P[J(t)=1 \mid J(t-1)=i]$ so

$$s_{i0kl}(t) = (1-p_i(t)) \cdot q_{i0kl}(t) \quad \text{and} \quad s_{i1kl}(t) = p_i(t) \cdot q_{i1kl}(t)$$

where $q_{ijkl}(t) = P[Y(t)=l \mid Y(t-1)=k, J(t)=j, J(t-1)=i]$. The $q_{ijkl}(t)$ will depend on the structure of Y. Some examples are given below.

A simple example is the distribution of the total number of rainy days in a given period. If the period is the m days from $t_0 + 1$ to t_m then the states of $Y(t)$ are $0, 1, \ldots, m$. The required distribution will be the distribution of $Y(t_m)$. The only non-zero values of $q_{ijkl}(t)$ are

$$q_{i0kk}(t) = 1, \quad i = 0, 1 \quad \text{and} \quad q_{i1k(k+1)}(t) = 1, \quad i = 0, 1.$$

The initial conditions are usually given by $Y(t_0) = 0$ and the marginal $P(J(t_0) = i)$, $i = 0, 1$. If the quantities $p_{ij}(t)$ do not depend on t or on i then we have a very complex way of generating the binomial distribution.

As a second example consider the chance of a dry spell of K or more days within the period. The states of $Y(t)$ are $1, 2, \ldots, K$ and the state K is an absorbing barrier. The non zero terms of q_{ijkl} are

$$\begin{aligned} q_{00k(k+1)}(t) &= 1, \quad k < K, \\ q_{01k1}(t) &= 1, \quad k < K, \\ q_{1j11}(t) &= 1, \quad j = 0, 1, \\ q_{ijKK}(t) &= 1, \quad i, j = 0, 1 \end{aligned}$$

and the required probability is given by $P[Y(t_m) = K]$. The initial conditions are given by using the same recurrence relations prior to day t_0 but with a reflecting rather than an absorbing barrier at K, i.e. $q_{01k1} = 1$ replaces $q_{01KK} = 1$. Alternatively, the probability of a K day dry spell in the period conditional on day t_0 being rainy can be evaluated by setting $P[J(t_0) = 1, Y(t_0) = 1 = 1$.

Part of the simplicity of the recurrence relations considered above arises because the characteristics considered do not involve the distribution of rainfall amounts. As a third example we therefore consider the distribution of water balance through the season. The water in the soil profile on day t is $W(t) = W(t-1) + X(t) - e$, with $X(t)$ taken as zero when $J(t) = 0$, subject to the constraint that $W(t)$ cannot be negative or exceed some upper limit. The constant, e, represents evaporation or other loss of water. The $K + 1$ states of $Y(t)$ represent discrete increments in $W(t)$, with $Y(t) = 0$ corresponding to an empty profile and $Y(t) = K$ to a full one. In the same units implied by the number of states of $Y(t)$, let

$$G_{iy}(t) = P[Y(t) \geqslant y - 1/2 \mid J(t-1) = i].$$

Then if the evaporation is e units, the non-zero terms of q_{ijkl} are given for reflecting barriers by

$$\begin{aligned} q_{i0k(k-e)} &= 1, \quad i = 0, 1, \quad k - e > 0, \\ q_{i0k0} &= 1, \quad k - e \leqslant 0, \\ q_{i1k(k+q-e)} &= G_{i(q+1)} - G_{iq}, \quad i = 0, 1, \quad k + q - e > 0, \quad k + q - e < K, \\ q_{i1k0} &= 1 - G_{i(e-k)}, \quad k - e \leqslant 0, \\ q_{i1kK} &= G_{i(K-k-e)} \end{aligned}$$

A variety of initial conditions can be used. The results are of course approximate but can (at the expense of computer time) be made as accurate as is required by using a large value of K. Typically, if 100 mm of water corresponds to a full profile, we find that reasonable results can be achieved with $K = 100$. The exact result involves calculating the joint distribution function of $Y(t)$ and $J(t)$, $F(t, y, j) = P[Y(t) \leqslant y, J(t) = j]$ as

$$F(t, y, j) = \sum_i F(t-1, y+e, i)\,(1 - p_i(t)), \quad j = 0$$

$$= \int_0^y \sum_i F(t-1, y+e-w, i)\, p_i(t) \cdot g_i(w, t)\, dw, \quad j = l,$$

where $g_i(w, t)$ is the density of rainfall amounts on a day t given $J(t) = 1$ and $J(t-1) = i$. To calculate this, however, numerical integration must be used. This involves converting the integral to a sum by choosing a step size and discretizing F, a process that is equivalent to choosing K.

A possible definition of the end of the rains is the first date after day t_0 that the soil profile is empty. The distribution of this date is calculated by setting $q_{i100} = 1$ in the above equations and recording $P[Y(t) = 0]$ for $t > t_0$. The extension of the equations to allow for evaporation that varies through the year and differs on dry and rainy days corresponds merely to replacing the constant e in the above equation by $e_i(t)$. It is also straightforward to allow the evaporation to be a random variable instead of a fixed constant.

3.2. *Examples*

The data for Irbid are used to illustrate some of the results that can be derived on the risk of dry spells. Fig. 5 gives the estimated probability of a dry spell of 10, 15 and 20 days in successive 30-day periods where "dry" is defined as a day on which there was less than 2.5 mm rain. To compare different models, results have been derived from three alternative first-order Markov chain models for the occurrence of rain. They are as follows:

(i) A two-state chain with "rain" defined as a day with 0.1 mm or more.
(ii) A two-state chain with "rain" defined as a day with 2.5 mm or more.
(iii) A three-state chain with the states as "dry", "trace" = 0.1–2.49 mm and "rain" = 2.5 mm or more.

Fig. 2 shows the estimated probabilities of rain for the three-state chain, Model (iii). Models (ii) and (iii) include the same threshold for rain as is used here to define the dry spells; hence the only part of the model needed is that for the occurrence of rain. The simplest case is model (ii) which uses the equations given in Section 3.1 directly. Separate calculations are made for each dry spell length and for a sequence of starting dates which are arbitrarily taken 10 days apart. For each calculation the recurrence relations with reflecting barriers give the initial spell length distribution. The results plotted in Fig. 5 are the probability of absorption in the subsequent 30 days. They are easily extended for three-state chains to give the results for model (iii). For model (i) the part of the model that describes the distribution of rainfall amounts is also required. This is used to estimate the proportion of days with rainfall between 0.1 and 2.49 mm. These proportions are then used to construct the three-state chain for which the dry spell probabilities are then calculated.

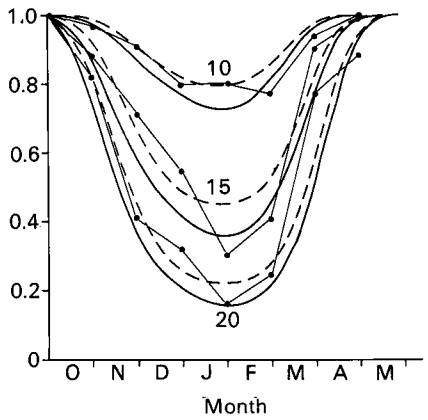

Fig. 5. Irbid, Jordan: observed and fitted proportion of years with a dry spell of at least 10, 15 and 20 days in the next 30 days. Observed (–•–•–), fitted using model (i) (———), fitted using models (ii) and (iii) (– – –).

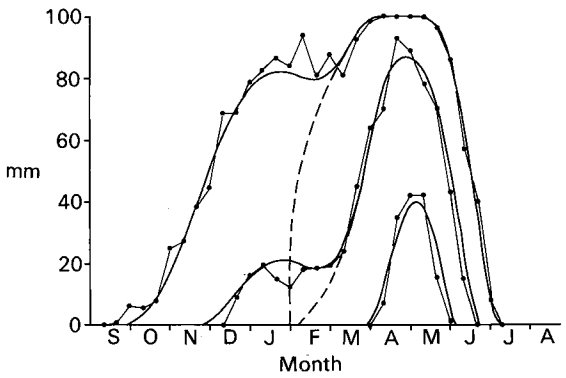

Fig. 6. Morogoro, Tanzania: 10, 50 and 90 per cent points of the distribution of water balance with maximum soil capacity of 100 mm and evaporation of 4 mm per day. Observed (–•–•–•–), fitted (———), fitted using initial conditions of a balance of 0 on January 28th (– – –).

The results from fitting the model indicate that the three-state chain model (iii) fits the data better than model (i) or (ii). Hence it is useful to compare the results for the 3 models and to see how they correspond to the observed data. In Fig. 5 the observed proportion of years with a dry spell of 10, 15 and 20 days is plotted together with estimates of the probabilities from models (i) and (iii). The estimates from model (ii) are indistinguishable from those for model (iii) and are close to the observed proportions. Model (i) gives estimated probabilities that are usually slightly less than the observed proportions. However, the differences are small and the results from this model may be adequate for some applications. More generally, however, this model illustrates that results can be derived for a range of thesholds other than those used in defining the model. Thus the process of fitting an appropriate model can be considered independently of the way in which the model will subsequently be used. This contrasts with the direct summary of the observed data where the actual data has to be re-examined for, in this case, each different threshold that is to be used.

The above example on dry spells illustrates the use of the model when $Y(t)$ has one absorbing and one reflecting barrier. If dry spells of length 10 are considered then $Y(t) = 10$ is an absorbing state and $Y(t) = 0$ is reflecting. The simple water balance described in Section 3.1 illustrates a variable with two reflecting barriers. If $Y(t)$ represents the water balance on day t and the distribution of this is required, then $Y(t) = 0$ and $Y(t) = K$ are both reflecting states. The model fitted in Section 3 to data from Morogoro has been used to calculate percentage points, plotted in Fig. 6 with $K = 100$, representing a maximum soil moisture capacity of 100 mm and $e = 4$ mm per day. The observed values have also been plotted at 10-day intervals and the agreement is very good. One would expect the 50 per cent point to be modelled adequately but the good fit of more extreme percentage points shows that the model also describes the year to year variability well.

The direct summary of the water balance shown in Fig. 6 uses a similar equation to that described by Hills and Morgan (1981). They also carried out an analysis to determine how sensitive the results were to differing values of evaporation and water-holding capacity. This analysis is quite tedious because the original daily records have to be re-summarized for each different set of assumptions. As an alternative, it is easy to use the recurrence relations for a range of values of the parameters. The fact that the Markov chain models used yield smooth curves for the percentage points also aids interpretation of the results.

At Morogoro the soil profile is empty in 30 per cent of the years on day 150 (January 28th) (Fig. 6). For such a year the recurrence relations may be used from day 150 with initial conditions $P[Y(150) = 0] = 1$ to give the distribution of water balance for the remainder of the season. The initial conditions for $J(t)$ must also be specified, and in Fig. 6 day 150 is assumed to be wet, that

is $P[J(150) = 1] = 1$. With the model used, these initial conditions affect the percentage points of the water balance for about 30 days. Overall, one of the strengths of the modelling approach together with the recurrence relations is the ease with which results can be derived for a range of initial conditions.

When dry spells were considered, the probability of an event occurring within a period was calculated by defining one state of $Y(t)$ to be absorbing. This technique may also be used to obtain the distribution of the date of first occurrence of an event. This is simply $P(Y(t) = K)$ for $t = t_0, t_0 + 1, \ldots$, where K is the absorbing state. An example of such an event is the end of the growing season which may be of interest in its own right or may be used together with information on the start of the rains (Stern, 1982; Stern *et al.*, 1981) to look at the length of growing season. Stored water is important at the end of the season and so a sensible definition of the end would be the first day when the soil profile becomes empty. This definition may, however, lead to too many "false ends" (Dennett *et al.*, 1981) and, to prevent this, the additional condition of no rain occurring for at least 10 days after the profile becomes empty is used. Thus the event being modelled is complex but this need not complicate the recurrence relations; all that is required is 10 more states added to the "bottom" of $Y(t)$ and a definition of the appropriate transition probabilities. Various initial conditions may be used. Table 6 was calculated assuming the profile was full on April 27th and that day was rainy.

TABLE 6

Distribution of the date of the end of the growing season at Morogoro . (This is defined as the first occasion after April 27th on which the water balance equation reaches the bottom barrier and remains there for 10 consecutive days)

Date t	*May 27th*	*June 6th*	*June 16th*	*June 26th*	*July 6th*	*July 16th*	*July 26th*
Probability that the growing season has ended by date *t*	0	0.036	0.069	0.300	0.605	0.811	0.912

4. CONCLUDING REMARKS

This paper shows how a class of models may be fitted to rainfall data and then used to provide the type of information that agricultural planners commonly require from the data. The case for using these models is made in two stages. The first is the claim that a comprehensive analysis of rainfall data should use daily records and not be based on 7-, 10-day or monthly totals. There is then a choice between a direct analysis of the characteristics of interest and an analysis via a model of the pattern of rainfall on a daily basis. The case made here is that it is now practical to consider adopting the modelling approach as a routine.

An important feature of the direct analysis is its simplicity, making it attractive to non-statisticians. We have however found that the types of models fitted are also quite easy for non-statisticians to understand. The model is not difficult to explain in terms of the "chances of rain" and the "amount of rain when it occurs". Scientists are then most impressed to be told that what is plotted in, for example, Fig. 1 is a non-stationary second-order Markov chain! In principle the direct analysis can be done by hand whereas the modelling and use of the fitted model both require a computer. However, the volumes of raw data involved mean that few realistic direct analyses would be attempted without computing facilities, so lack of these would be a difficulty with either method of analysis.

One problem with the modelling approach is that it makes more assumptions about the structure of the data. One that is particularly important is whether there is any long-term dependence of the pattern of rainfall within the year. The literature gives conflicting results, and in many papers the statistical methods used are poor (Adedokin, 1979; Mooley, 1971; Winstanley, 1974).

For the stations that we have looked at in detail we have found no evidence of any important correlations between rainfall events at different periods within the year. (Dennett *et al.*, 1983.)

The major limitation of the direct analysis is inefficient use of the data, so that long records are required if we are to obtain estimates having reasonably small standard errors. In most countries there are few reliable long records, compared to the number of records of 10 or 15 years' duration, most of which have not been analysed. Initial studies indicate that the modelling approach gives sensible results for records of this length. In practice the few long records can therefore be used to validate the type of model adopted, by comparison with the direct analysis, and then to fit and use similar models for neighbouring shorter records.

One poor feature of almost all the direct analyses of rainfall data for agriculture has been the absence of any information on standard errors. This may partly be a problem of terminology because estimates of percentage points of rainfall totals have commonly been referred to as "confidence limits", (e.g. Manning, 1955; Kowal and Knabe, 1972). We have, so far, followed this tradition in our use of the modelling approach, but the assessment of approximate standard errors for the estimates of characteristics of interest is an important topic requiring further work.

We believe that the process of fitting and using the models and producing plots can be made into a routine that is usable by non-statisticians. The package GLIM permits a wide range of realistic models to be fitted; its macro facilities make the process easy to use, and allow a straightforward interface with programs that use the fitted models. We have analysed data from more than 15 countries and have had little difficulty in finding models that appear, from plots and analysis of deviance tables, to fit well. In using the models we have been pleasantly surprised by the extent to which results derived from the model agree with those from a direct analysis of the data. The results in Figs 5 and 6 are typical in this respect.

The type of model used is flexible and has many possible variations in addition to those discussed in Section 2 that may be useful at a particular site. For example, the order of the Markov chain could vary through the year. This model is easy to fit and does not complicate the recurrence relations. One use of this modification is for sites where, because of the lack of rainfall, data are insufficient to estimate the parameters of a high order Markov chain for the whole year. The variations in the model illustrated in Fig. 2 for the occurrence of rain could also be applied to the mean rain per rainy day, $\mu(t)$. One assumption made in all the models considered is that the distribution of the rainfall amount $X(t)$ may depend on $J(t-1)$ but does not depend on $X(t-1)$. Such dependence would make both fitting and using the appropriate models more difficult. The dependence has been investigated by Buishand (1977) and Katz (1977). For the sites considered we have found no important correlations between the rainfall amounts on successive rainy days.

In using the models there are some characteristics of the rainfall for which the recurrence relations become complicated even for the simple models used in this paper. Examples include questions on rainfall amount, dry spell length or water balance in a period of the year when that period depends on another characteristic. For example, "What is the distribution of the amount of rain falling in the first 30 days of the wet season?" This introduces another level of summation into the recurrence relations and we have therefore used simulation to derive results of this type. Further examples where simulation is needed were given by Stern and Coe (1982).

We have not discussed here any spatial aspects of rainfall data. This is a large and complex field. However, the modelling approach should be as useful when analysing several related sites as when looking at just one. The models at different sites may be compared directly or used to derive distributions of interest at each site which are then compared.

ACKNOWLEDGEMENTS

This work was supported by the UK Overseas Development Administration. We thank our colleagues in the Tropical Agricultural Meteorology group and the Department of Applied Statistics for comments and assistance.

REFERENCES

Abramowitz, M. and Stegun, I A. (1968) *Handbook of Mathematical Functions*. New York: Dover.

Adedokun, J. A. (1979) Towards achieving an in-season forecast of West African precipitation. *Arch. Met. Geoph. and Biokl.* Ser A, **28**, 19–38.

Archer D. R. (1981) Rainfall sequences in Northern Malawi. *Weather*, **36**, 2–9.

Baker, R. J. and Nelder, J. A. (1978) *The GLIM System*. Oxford: NAG.

Benoit, P. (1977) The start of the growing season in Northern Nigeria. *Agric. Meteor.*, **18**, 91–99.

Buishand, T. A. (1977) *Stochastic Modelling of Daily Rainfall Sequences*. Mededlingen Landbouwhogeschool, Wageningen.

Caskey, J. E. (1963) A Markov chain model for the probability of precipitation occurrence in intervals of various lengths. *Mon. Weather Rev.*, **91**, 298–301.

Chin, E. H. (1977) Modelling daily precipitation process with Markov chain. *Water Resources Res.*, **13**, 949–956.

Cocheme, J. and Franquin, P. (1967) A study of the agroclimatology of the semi-arid area south of the Sahara. FAO/UNESCO/WMO Tech. Bull. 86.

Coe, R. and Stern, R. D. (1982) Fitting models to daily rainfall data. *J. Appl. Meteor.*, **21**, 1024–1031.

Cooke, D. S. (1953) The duration of wet and dry spells at Moncton, New Brunswick. *Quart. J. Roy. Met. Soc.*, **79**, 536–538.

Davy, E. G., Mattei, F. and Solomon, S. I. (1976) An evaluation of climate and water resources for development of agriculture in the Sudano-Sahelian zone of West Africa. WMO Special Environmental Report No. 9. WMO, Geneva, Switzerland.

Dennett, M. A., Rodgers, J. A. and Stern, R. D. (1981) Rainfall at Kampi-ya-Mawe and Katumani, Kenya. Report No. 3, Tropical Agric. Meteor. Group, University of Reading.

———(1983) Independence of rainfalls through the rainy season and the implications for the estimation of rainfall probabilities. *J. Climatol.*, to appear.

Doorenbos, J. and Pruitt, W. O. (1977) Crop water requirements. FAO Irrigation and Drainage Paper No. 24. FAO, Rome, Italy.

Dumont, A. G. and Boyce, D. S. (1974) The probabalistic simulation of weather variables. *J. Agric. Eng. Res.*, **19**, 131–145.

Feyerherm, A. M. and Bark, L. D. (1965) Statistical methods for persistent precipitation patterns. *J. Appl. Meteor.*, **4**, 320–328.

Fienberg, S. E. (1980) *The Analysis of Cross-classified Categorical Data*, 2nd ed. Cambridge, Mass.: MIT Press.

Gabriel, K. R. and Neumann, J. (1962) A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quart. J. Royal Met. Soc.*, **88**, 90–95.

Gates, P. and Tong, H. (1976) On Markov chain modelling to some weather data. *J. Appl. Meteor.*, **15**, 1145–1151.

Garbutt, D. J., Stern, R. D., Dennett, M. D. and Elston, J. (1981) A comparison of the rainfall climate of eleven places in West Africa using a two-part model for daily rainfall. *Arch. Met. Geoph. Biokl.* Ser B, **29**, 137–155.

Green, J. R. (1970) A generalised probability model for sequences of wet and dry days. *Monthly Weather Rev.*, **98**, 238–241.

Greenwood, J. A. and Durand, D. (1960) Aids for fitting gamma distribution by maximum likelihood. *Technometrics*, **2**, 55–65.

Haan, C. T., Allen, D. M. and Street, J. O. (1976) A Markov chain model of daily rainfall. *Water Resources Res.*, **12**, 443–449.

Heerman, D. F., Finkner, M. D. and Hiler, E. A. (1968) Probability of sequences of wet and dry days for eleven western states and Texas. *Colorado Agric. Exp. Station Tech. Bull.* 117.

Hills, R. C. and Morgan, J. H. T. (1981) An interactive approach to the analysis of rainfall records for agricultural purposes. *Expl. Agric.*, **17**, 1–16.

Jackson, I. J. (1981) Dependence of wet and dry days in the tropics. *Arch. Met. Geoph. and Biokl.* Ser B, **29**, 167–179.

Jones, J. W., Colwick, R. F. and Threadgill, E. D. (1970) A simulated environment model of temperature, evaporation, rainfall and soil temperature. ASAE Paper No. 70-404, American Soc. Agric. Engrs.

Katz, R. W. (1974) Computing probabilities associated with the Markov chain model for precipitation. *J. Appl. Meteor.*, **13**, 953–954.

———(1977) Precipitation as a chain dependent process. *J. Appl. Meteor.*, **16**, 671–676.

Kavvas, M. L. and Delleur, J. W. (1981) A stochastic cluster model of daily rainfall sequences. *Water Resources Res.*, **17**, 1151–1160.

Khanal, N. N. and Hamrick, R. L. (1974) Misc. Publ. No. 1275, U.S. Dept Agric., 197–210.

Kowal, J. M. and Knabe, D. T. (1972) *An Agroclimatological Atlas of the Northern States of Nigeria*. Zaria, Nigeria: Ahmadu Bello University Press.

Lawrence, E. N. (1954) Application of mathematical series to the frequency of weather spells. *Meteor. Mag.*, **83**, 195–200.

Le Cam, L. (1961) A stochastic description of precipitation. *Proc. 4th Berkeley Symp.*, pp. 165–186.

Lloyd, E. H. (1979) Stochastic storage problems. In *The Mathematics of Hydrology and Water Resources* (Lloyd E. H., O'Donnell, T. and Wilkinson, J. C. eds.) pp. 73–85. London: Academic Press.

Lowry, W. P. and Guthrie, D. (1968) Markov chains of order greater than one. *Monthly Weather Rev.*, **96**, 798–801.

Manning, H. L. (1955) Calculation of confidence limits of monthly rainfall. *J. Agric. Sci.*, **45**, 154–156.

Mooley, D. A. (1971) Independence of monthly and bimonthly rainfall over Southeast Asia during the summer monsoon season. *Mon. Wea. Rev.*, **99**, 532–536.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalised linear models. *J. R. Statist. Soc.* A, **135**, 370–384.

O'Brien, G. L. (1974) Limit theorems for sums of chain-dependent processes. *J. Appl. Prob.*, **11**, 583–587.

Panabokke, C. R. and Walgama, A. (1974) The application of rainfall confidence limits to crop water requirements in dry zone agriculture in Sri Lanka. *J. National Sci. Council Sri Lanka*, **2**, 95–113.

Sen, Z. (1980) Critical drought analysis of periodic-stochastic processes. *J. Hydrol.*, **46**., 251–263.

Shenton, L. R. and Bowman, K. O. (1977) *Maximum Likelihood Estimation in Small Samples*. London: Griffin.

Singh, S. V., Kripaluni, R. H., Shaka, P., Ismail, P. M. M. and Dahale, S. D. (1981) Persistence in daily and 5-day summer monsoon rainfall over India. *Arch. Met. Geoph. and Biokl.* Ser. A, **30**, 261–277.

Stern, R. D. (1982) Computing a probability distribution for the start of the rains from a Markov chain model for precipitation. *J. Appl. Meteor.*, **21**, 420–423.

Stern, R. D. and Coe, R. (1982) Use of rainfall models in agricultural planning. *Agric. Meteor.*, **26**, 35–50.

Stern, R. D., Dennett, M. D. and Dale, I. C. (1982) Analysis of daily rainfall measurements to give agronomically useful results. I. Direct methods. *Expl. Agric.*, **18**, 223–236.

Stern, R. D., Dennett, M. D. and Garbutt, D. N. (1981) The start of the rains in West Africa. *J. Climat.*, **1**, 59–68.

Todorovic, P. and Woolhiser, D. A. (1975) A stochastic model of *n*-day precipitation. *J. Appl. Meteor.*, **14**, 17–24.

Waymire, E. and Gupta, V. K. (1981) The mathematical structure of rainfall representation. 1. A review of the stochastic rainfall models. *Water Resources Res.*, **17**, 1261–1272.

Williams, C. R. (1952) Sequences of wet and dry days considered in relation to the logarithmic series. *Quart. J. Roy. Meteor. Soc.*, **78**, 91–96.

Winstanley, D. (1974) Seasonal rainfall forecasting in West Africa. *Nature*, **248**, 464.

Woodhead, T., Waweru, E. S. and Lawes, E. F. (1970) Expected rainfall and Kenya agriculture – confidence limits for large areas at minimum cost. *Expl. Agric.*, **6**, 87–97.

Woolhiser, D. A. and Pegram, G. G. S. (1979) Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. *J. Appl. Meteor.*, **18**, 34–42.

DISCUSSION OF THE PAPER BY DR STERN AND MR COE

Dr A. J. Lawrance (University of Birmingham): I would like to begin by congratulating the authors of tonight's paper for a readable, listenable and interesting contribution.

The objectives of the paper were clearly set, as being (in my paraphrasing) to develop into a viable strategy the modelling of daily rainfall data by Markov chains and gamma distributions, both of these having seasonal time varying parameters, and to use the models to investigate pertinent questions in agricultural planning. There is no doubt that they have achieved these objectives; this achievement should be judged against the background of earlier work in similar vein. The authors do not pretend that Markov chains, gamma distributions and Fourier series are new to the area, but earlier authors have viewed these ideas in much more *ad hoc* ways. In particular, the incorporation of seasonality directly into the transition probabilities via the logit link is to be welcomed; seasonality has been an intractable and unsatisfactory aspect of earlier work. The authors show how their Fourier seasonality fits elegantly into a Glim formulation, and the paper should do much to bring this British invention to the notice of more workers in the Geophysical sciences. I admire the way their formulation proceeds in order to take account of the non-standard aspects of the situation, and its emphasis on simplification of models. The introduction of seasonal time varying gamma distributions for the amount of daily rainfall is a sensible strategy; this sort of effect has been perceived necessary in earlier work, but the point to stress now is that the authors have found a successful formulation and have thoroughly implemented it. Their uses of recurrence relations to determine features, such as the total number of rainy days and the distribution of water balance, are valuable contributions.

However, it might be said that any model can be fitted to any set of data, and should we swallow totally the measured analysis we have heard presented tonight? When studying an applied paper such as this, the first thing I want to get is some feel of the data. For me this usually begins