JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 95, NO. D2, PAGES 1965–1972, FEBRUARY 20, 1990

Estimation of Mean Rain Rate: Application to Satellite Observations

BENJAMIN KEDEM[1]

Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park

LONG S. CHIU[2]

Applied Research Corporation, Landover, Maryland

GERALD R. NORTH[3]

NASA Goddard Space Flight Center, Greenbelt, Maryland

A method for the estimation of the mean area average rain rate from dependent data is developed and applied to the GARP Atlantic Tropical Experiment GATE data. The method consists of fitting a mixed distribution, containing an atom at zero, by minimum chi-square in combination with certain time-space sampling designs. In modeling the continuous component of the mixed distribution it is shown that the lognormal distribution provides a very close fit for the nonzero area average rainrates. A comparison with the gamma distribution shows that the lognormal distribution is a better choice as expressed by the minimum chi-square criterion. Some of the time-space sampling designs correspond to satellite sampling. The results indicate that a satellite visiting an area of about 350×350 km^2 in the tropics approximately every 10 hours over a period can provide a rather close estimate for the mean area average rain rate.

1. INTRODUCTION

1.1. *The Problem and Approach*

Global measurements of precipitation are important for prognostic and diagnostic studies of the atmospheric circulation. Because of the huge extent of the tropical oceans and the inevitable errors associated with in situ measurements from ships, satellite observation is probably the ultimate mode by which global precipitation measurements can be made [*Austin and Geotis*, 1980; *Atlas and Thiele*, 1981]. Satellite observation gives complete spatial coverage for each passage at distinct revisits separated hours apart. An important problem then is to estimate the total amount of rainfall volume from this type of data. When the measuring device on board the satellite is a radar or a radiometer that is capable of measuring instantaneous rain rate, this problem is equivalent to the estimation of the mean of the distribution of rain rate when the satellite observation is confined to a given period over a certain area. The feasibility of this method and the associated statistical considerations are examined in this paper in some detail.

One way to approach this estimation problem, is to model the rain field as a random field [*Eagleson*, 1967; *Rodriguez-Iturbe and Mejía*, 1974; *Bras and Colon*, 1978] an approach that usually presupposes stationarity or homogeneity of the random field. Our approach, on the other hand, is through statistical inference directed at the mixed distribution of rain rate observed in time and space and avoids stationarity and/or homogeneity assumptions. This approach sheds light on the tail behavior of the distribution of rain rate. As a by-product of this approach we obtain a useful linear relationship between the expected rain rate and the probability that it exceeds a given threshold.

1.2. *Goals and Preliminary Remarks*

The purpose of this paper is threefold. First, we demonstrate that a mixed lognormal distribution provides a very close fit for area average rain rate. Specifically we fit a mixed lognormal distribution $\Delta(1 - p, \mu, \alpha^2)$ [see *Aitchison and Brown*, 1963, p. 95] to rain rate averaged over 4×4 km^2 pixels obtained from the GARP (Global Atmospheric Research Program) Atlantic Tropical Experiment (GATE) and estimate the mean of the distribution and its standard error. Actually, this is done twice corresponding to each of two different phases of GATE, where the information about the continuous component describing the distribution of the positive values is truncated at 1 mm/h. Second, we examine the effect of different sampling designs in time and space on the fitted model and its parameters. As we shall see, these designs provide valuable insight into the time-space dependence of rain rate. Third, we interpret these designs from the point of view of remote sensing via satellite.

It is a fact, often ignored, that the distributions associated with rain characteristics are supported on a continuum except for an atom at zero. This means that the random characteristic in question has a positive probability of being equal to zero, but otherwise its distribution is continuous. Such a distribution, which is made of a discrete component and a continuous component, is called a mixed distribution. Clearly, there is no problem in modeling the discrete component, for it is nothing but a spike at zero whose magnitude is equal to the probability of admitting the value zero. The

[1]Also at Applied Research Corporation, Landover, Maryland.
[2]Also at NASA Goddard Space Flight Center, Greenbelt, Maryland.
[3]Also at Department of Meteorology, Texas A & M University, College Station.

Paper number 89JD02762.
0148-0227/90/89JD-02762$05.00

continuous component, on the other hand, can be modeled in more than one way unless physical or mathematical evidence points to a particular distribution. The case we have in mind, of course, is that of rain rate.

Rain rate has a distribution of the mixed type because there is a positive probability that it does not rain at all. The continuous distribution component of rain rate can be modeled in several ways, and there does not seem to be a general agreement as to its precise nature. This may be due to the fact that different data sets collected under different physical conditions do not point to any specific distribution but rather point to a whole range of probability distributions whose tail behavior differs greatly. Thus *Neyman and Scott* [1967] suggest the gamma distribution, while *Lovejoy and Schertzer* [1985] suggest a fat tailed hyperbolic distribution. *Houze and Cheng* [1977] argue that a lognormal fit is reasonable, and in practice it is even quite common to fit a truncated normal distribution. In addition to empirical studies, attempts have been made to describe the distribution of rain rate by resorting to stochastic models. Such an attempt has also been made by *Kedem and Chiu* [1987], who employed a stochastic regression scheme to model time series of area average rain rate. Some conditions on the stochastic regression model parameters necessary for asymptotic (in time) lognormality were satisfied rather closely by quite a few time series from GATE, a fact that encourages the entertainment of the lognormal distribution as a possible model. In this paper we continue to investigate the appropriateness of the lognormal distribution as a model for positive rain rate averaged over an area.

The paper is organized as follows. After a description of the GATE data and a precise formulation of the random variable that represents the data (section 2), the main thrust of a statistical analysis based on time-space sampling designs is discussed in section 3. Section 4 deals with a comparison between the fits provided by the lognormal and gamma distributions to the GATE data, and in section 5 we examine the implication of the results of the time-space sampling designs on satellite sampling. Finally, conclusions and a summary are presented in section 6.

2. The Data and Some Basic Statistical Issues

In this section we address basic issues concerning our approach to modeling a time-space characteristic whose probability distribution contains an atom at the origin. It is perhaps best to start with a description of the data set that triggered the present investigation.

2.1. *The GATE Data*

GATE is an observational program conducted in the summer of 1974. During three consecutive periods each of about 3 weeks in duration, detailed rainfall measurements from rain gauges and radars on an array of research vessels were made over an area called the B scale. The center of the B scale area is located at 8.5°N, 23.5°W in the Atlantic Ocean and encompasses an area of about 400 km in diameter. *Arkell and Hudlow* [1977] composited the radar measurements from ships and presented an atlas of radar echoes at 15-min intervals. *Patterson et al.* [1979] converted the radar measurements to instantaneous rain rates averaged over 4×4 km^2 pixels. Our data consist of these average rain rates. Each of the three periods is referred to as a phase. In this work we use the data from the first two phases and refer to them in the sequel as GATE 1 and GATE 2, respectively.

2.2. *The Population of Area Average Rain Rate*

To carry on the statistical analysis of rain rate in a sound and clear manner, we must identify precisely the random variable in question. For this purpose, concentrate for a moment on the GATE 1 data collected during a period of nearly 3 weeks. We wish to study the distribution of the instantaneous area averages of rain rate confined to this period only, over the B scale area.

It is important to note that although the instantaneous average rain rate over a given pixel was recorded every 15 min, in principle we could do so every minute or every second or indeed continuously in time. It is helpful to think of a given pixel visited at a certain instant of time as an "instantaneous pixel" from which we obtain the reading of an instantaneous average rain rate.

The GATE 1 data consist of rain rates averaged over 4×4 km^2 (nonoverlapping) pixels which cover the B-scale area. A closer look reveals that we really have a finite collection of values of averaged rain rates obtained from the infinite parent population that consists of all possible instantaneous average rain-rate values that in principle could have been observed over the 4×4 km^2 pixels that cover the B-scale area in the given period of about 3 weeks. In other words, we have an infinite population of values of a random variable that assigns to every "instantaneous 4×4 pixel" the corresponding average rain rate, and we identify the distribution of the instantaneous average rain rates with the distribution of the random variable. Having thus defined the random variable whose values are the instantaneous average rain rates that could in principle be observed over the 4×4 (nonoverlapping) pixels of the B scale area during the first phase of GATE, we can now study its distribution in a meaningful way. Denote this random variable by R. The same formulation holds for the other two phases of GATE.

2.3. *Mixed Distributions*

Most probability distributions encountered in practice are "regular." That is, they are either discrete, supported on a countable number of values, or continuous, in which case the distribution is supported on a continuum. However, there are many situations in which the cumulative distribution function contains jumps at some points but is otherwise continuous. Such a distribution is neither discrete nor continuous but rather a combination of a discrete component and a continuous component, and is referred to as a mixed distribution. The case of rain rate presents an example of a mixed distribution, for the event $\{R = 0\}$ has a positive probability $1 - p$, say, but otherwise $P(R = r) = 0$, $r > 0$. More precisely, let G be the cumulative distribution function of R, $G(r) = P(R \le r)$. Then it can be represented as a convex combination of two increasing functions H, F

$$G(r) = (1 - p)H(r) + pF(r) \tag{1}$$

where

$$H(r) = 0 \qquad r < 0$$

$$H(r) = 1 \qquad r \geq 0$$

and F is a continuous distribution function such that $F(r) = 0$, $r \leq 0$, with a density $f(r) = F'(r)$, $r > 0$. It follows that the generalized density $g(r)$ corresponding to $G(r)$ takes the form

$$\begin{aligned} g(r) &= 0 \qquad r < 0 \\ g(r) &= 1 - p \qquad r = 0 \\ g(r) &= pf(r) \qquad r > 0 \end{aligned} \tag{2}$$

where $f(r)$ is the density of R conditional on $R > 0$ (see also *Aitchison and Brown* [1963, p. 95] and *Feuerverger* [1979]). The kth moment of this distribution is given by

$$E(R^k) = \int_{-\infty}^{\infty} r^k \, dG(r) = p \int_0^{\infty} r^k f(r) \, dr \tag{3}$$

and in particular, the mean and variance are given by

$$E(R) = p \int_0^{\infty} rf(r) \, dr \tag{4}$$

$$\text{Var } (R) = p\left\{ \int_0^{\infty} r^2 f(r) \, dr - p\left[\int_0^{\infty} rf(r) \, dr \right]^2 \right\} \tag{5}$$

In general, the kth central moment is given by

$$E(R - E(R))^k = (-ER)^k(1 - p) + p \int_0^{\infty} (r - E(R))^k f(r) \, dr \tag{6}$$

The mixed distribution (equation (1)) leads to an interesting relationship between $E(R)$and $P(R > \tau)$. First, note that for fixed $\tau \geq 0$,

$$p = \frac{E(R)}{E(R|R > 0)} = \frac{P(R > \tau)}{P(R > \tau|R > 0)} \tag{7}$$

Thus solving for $E(R)$, we obtain the linear relationship for any fixed $\tau \geq 0$,

$$E(R) = \beta_\tau P(R > \tau) \tag{8}$$

where β_τ is given by

$$\beta_\tau = \frac{E(R|R > 0)}{P(R > \tau|R > 0)}$$

When $\tau = 0$, (8) reduces to (4). When $\tau > 0$, the Markov inequality ($\tau P(R > \tau) \leq E(R)$) yields the lower bound

$$\tau \leq \beta_\tau$$

A useful interpretation of (8) is that for sufficiently large area, and assuming spatial and temporal statistical homogeneity (i.e., constant β_τ),

$$\langle R \rangle \simeq \beta_\tau P_\tau \tag{9}$$

where $\langle R \rangle$ is the area average of rain rate, and P_τ is the fraction of the area covered with rain rate above the threshold τ. The relationship (9) holds to a great degree of accuracy, a fact that has been verified for the GATE data by *Chiu* [1988] and for other data sets by *Atlas et al.* [1988] and *Rosenfeld et al.* [1988]. It should be noted, however, that (8) is true for any mixed distribution with an atom at zero and can be viewed as a theorem, while (9) is a manifestation of the law of large numbers corresponding to (8).

2.4. *The Mixed Lognormal Distribution*

When f comes from a parametric family indexed by a vector of parameters θ, we replace the notation $f(r)$ by $f(r, \theta)$. In this case, $g(r)$ is replaced by $g(r, p, \theta)$. When a random sample $R_1, R_2, \cdots, R_n$ containing $n - m$ zeroes is given, the likelihood of p, θ corresponding to the random sample is given by

$$L(p, \theta) = p^m(1 - p)^{n-m} f(r_1, \theta) \cdots f(r_m, \theta) \tag{10}$$

and the maximum likelihood estimator of p is readily given by $\hat{p} = m/n$. Note that when the data are dependent, (10) does not represent the true likelihood. To use the maximum likelihood estimation procedure, one must know the precise dependence, and when the latter is not known, other methods are needed such as the one described in section 3.

Let $\theta = (\mu, \sigma^2)$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and assume that $f(r, \theta)$ is given by the lognormal density

$$\begin{aligned} f(r, \theta) &= \frac{1}{r\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2} (\log r - \mu)^2 \right] \qquad r > 0 \\ f(r, \theta) &= 0 \qquad r \leq 0 \end{aligned} \tag{11}$$

In this case we shall say that R has a mixed lognormal distribution with parameters p, μ, σ^2. *Aitchison and Brown* [1963, p. 95] denote this distribution by $\Delta(\delta, \mu, \sigma^2)$, where $\delta \equiv 1 - p$. Then (4) and (5) become

$$E(R) = p \exp(\mu + \sigma^2/2) \tag{12}$$

$$\text{Var } (R) = p \exp(2\mu + \sigma^2)[\exp(\sigma^2) - p] \tag{13}$$

Throughout the paper the measurement unit of R is millimeters per hour. This means, from (12) and (13), that the unit of μ is log (mm/h), but σ^2 is a pure number.

3. Analysis of the GATE Data

We would like to demonstrate, using minimum chi-square estimation, that the mixed lognormal distribution provides a close fit for the GATE data. Since no modeling is required for the discrete component of the distribution, it suffices to consider the density $f(r, \theta)$ and show that it fits the nonzero average rain rates rather closely. At the same time we will combine the estimation problem with the study of the dependence structure of the data in time and space, by sampling the data according to certain designs. This novel approach has certain advantages which simple random sampling lacks.

3.1. *Minimum Chi-Square Estimation*

In the face of dependent data, a useful procedure to estimate θ, and at the same time obtain a good idea of the goodness of fit and the dependence structure, is to use the procedure of minimum chi-square estimation together with certain time-space sampling designs. This indeed amounts to

using less than the full data set, but the lessons learned from this exercise provide valuable insight into the time-space dependence in the data as well as an answer to the problem of how often a rain field should be sampled via satellite in order to estimate the average amount of precipitation with reasonable precision.

Suppose the nonzero average rain rates can be grouped into k categories with o_i observations falling in category i while $e_i(\theta)$ is the corresponding expected value (dependent on θ!) obtained under $f(r, \theta)$. Consider the quadratic form,

$$\chi^2(\theta) \equiv \sum_{i=1}^{k} \frac{(o_i - e_i(\theta))^2}{e_i(\theta)} \tag{14}$$

The parameter $\hat{\theta}$ which minimizes $\chi^2(\theta)$ is called the minimum chi-square estimate of θ. The advantages of using minimum chi-square estimation are as follows: (1) $\chi^2(\theta)$ can be evaluated regardless of dependence; (2) the minimum chi-square approach adapts itself easily to the situation when rain rates are known only within a certain range, as is the case with most remote-sensing techniques, including radar; (3) estimation and goodness of fit are embodied in a single quantity; (4) the procedure can be easily interpreted, for we simply want $\chi^2(\theta)$ to be as close to zero as possible; (5) $\chi^2(\theta)$ can be easily evaluated even for very large samples; (6) $\chi^2(\theta)$ can provide valuable information about the dependence structure of the data when different sampling designs are used; (7) $\chi^2(\theta)$ can be used in comparing different models for $f(r, \theta)$; and (8) other theoretical advantages (enumerated by *Berkson* [1980]) when the data are independent.

3.2. *Time-Space Sampling Designs*

Recall that we think of the B scale area as covered by 4 × 4 km^2 pixels and that the GATE data consist of instantaneous snapshots over this area every 15 min. The time-space sampling designs we used in this work are characterized by the triples (n, k, l). The first index n denotes sampling frequency in time and is always a multiplier of 15 min; k and l refer to spatial sampling in east-west and north-south directions, respectively. They are always multipliers of 4 km. For example, the design (1, 10, 10) denotes sampling "instantaneous pixels" every 15 min in time but separated by 40-km intervals in both east-west and north-south directions. This is similar to "visiting" rain gauges located on a grid in space and separated by 40 km every 15 min. The design (48, 1, 1) samples every possible pixel every 12 hours (48 × 15 min). This is similar to sampling by a densely scanning sensor on board a polar-orbiting satellite that passes over the same location every 12 hours.

3.3. *Truncation at 1 mm/h*

In the GATE data it is difficult to distinguish zero observations from those that are positive but very close to zero. This may be due, among other sources, to noise in radar reflectivity. To overcome this technical difficulty, the data were truncated at 1 mm/h, as was done in an earlier study by *Austin and Geotis* [1978]. Thus instead of fitting $f(r, \theta)$ to the positive average rain rates, we fitted the truncated density

$$\frac{f(r, \theta)}{\int_1^\infty f(r, \theta)\, dr} \tag{15}$$

TABLE 1. Results From Two Different Designs

	GATE 1: (8, 8, 8)		GATE 2: (10, 5, 5)	
Class	o_i	$e_i(\hat{\theta})$	o_i	$e_i(\hat{\theta})$
1–2	453	450	697	704
2–4	590	598	864	855
4–6	325	324	459	448
6–8	207	188	261	259
8–10	116	116	156	162
10–12	60	76	94	107
12–16	82	88	126	128
16–20	52	46	70	69
>20	80	79	135	130
Total N	1965	1965	2862	2862

The minimum chi-square estimate is $\hat{\theta}$. For GATE 1, $\hat{\theta} = (\hat{\mu} = 1.140, \hat{\sigma}^2 = 1.047)$, $\hat{\alpha} = 5.28$, $\hat{\beta}^2 = 51.5$, and $\chi^2(\hat{\theta}) = 6.74$. For GATE 2, $\hat{\theta} = (\hat{\mu} = 1.043, \hat{\sigma}^2 = 1.205)$; $\hat{\alpha} = 5.184$, $\hat{\beta}^2 = 62.788$, and $\chi^2(\hat{\theta}) = 2.589$.

to the R_i, $R_i > 1$, where $f(r, \theta)$ is given by (11) and $\theta = (\mu, \sigma^2)$. The corresponding $e_i(\theta)$ can be easily computed under lognormality.

The truncated data R_i, $R_i > 1$, were grouped in nine categories 1–2, 2–4, 4–6, 6–8, 8–10, 10–12, 12–16, 16–20, and >20 mm/h. The number of categories and their sizes were chosen so that the expected number of observations in each bin was at least 10 across all the sampling designs used. This rule guarded against sampling designs that were too sparse. Let N denote the number of R_i greater than 1 mm/h. Then (15) and lognormality give

$$e_1(\theta) = N \frac{\Phi((\log 2 - \mu)/\sigma) - \Phi(-\mu/\sigma)}{1 - \Phi(-\mu/\sigma)}$$

$$e_2(\theta) = N \frac{\Phi((\log 4 - \mu)/\sigma) - \Phi((\log 2 - \mu)/\sigma)}{1 - \Phi(-\mu/\sigma)} \tag{16}$$

and so on for $e_3(\theta), \cdots, e_9(\theta)$, where Φ denotes the cumulative distribution function of the standard normal distribution. Similar expressions can be obtained, provided that we specify $f(r, \theta)$.

3.4. *Results From Different Designs*

The use of (14), (15), and (16) is first illustrated in Table 1 with the design (8, 8, 8) from GATE 1 and the design (10, 5, 5) from GATE 2. In the first case, $\chi^2(\theta)$ is minimized for $\hat{\theta} = (1.14, 1.047)$, and in the second case $\chi^2(\theta)$ is minimized for $\hat{\theta} = (1.043, 1.205)$. The corresponding minimum chi-square values are 6.74 and 2.589. For the moment, accept these numbers as reference values. The table also gives the estimated mean ($\hat{\alpha}$) and variance ($\hat{\beta}^2$) of the estimated or fitted lognormal distribution. The results from many other designs are summarized in Table 2. Note that $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\alpha}$, and $\hat{\beta}^2$ are the estimates of the parameters in the original nontruncated distribution, but they were obtained from the truncated data. Figure 1 shows the histogram corresponding to the design (8, 8, 8) and the fitted lognormal density.

There is a way to estimate p from the truncated data, since n, the total number of observations zero or nonzero, is known as well as N, the total number of average rain rates

TABLE 2. Minimum χ^2 Estimates From Different Designs

Design	$\chi^2(\hat{\theta})$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{p}$
		GATE 1		
(2, 4, 4)	11.125	1.137	1.089	0.083
(2, 8, 8)	4.199	1.157	1.062	0.082
(4, 4, 4)	3.733	1.129	1.098	0.083
(4, 8, 8)	3.849	1.140	1.079	0.084
(5, 20, 20)	0.803	1.185	1.067	0.077
(6,6, 6)	2.916	1.152	1.056	0.081
(6, 8, 8)	3.814	1.169	1.085	0.082
(8, 4, 4)	2.728	1.126	1.093	0.083
(8, 6, 6)	4.931	1.182	1.061	0.080
(10, 8, 8)	5.151	1.162	1.059	0.083
(10, 10, 10)	3.386	1.085	1.176	0.081
(20, 10, 10)	7.757	1.095	1.071	0.075
(24, 1, 1)	27.443	1.159	1.056	0.083
(48, 1, 1)	48.077	1.255	1.019	0.088
		GATE 2		
(4, 4, 4)	50.358	1.065	1.211	0.069
(3, 10, 10)	14.874	1.032	1.211	0.071
(5, 3, 3)	40.581	1.099	1.160	0.072
(5, 5, 5)	12.241	1.056	1.191	0.069
(5, 10, 10)	4.084	1.031	1.186	0.073
(8, 8, 8)	17.391	1.046	1.108	0.070
(10, 5, 5)	2.590	1.043	1.205	0.070
(10, 10, 10)	5.405	0.960	1.348	0.072
(20, 3, 3)	12.622	1.050	1.205	0.068
(20, 5, 5)	2.741	0.998	1.257	0.071
(20, 10, 10)	2.449	0.918	1.392	0.074
(30, 5, 5)	7.469	0.976	1.261	0.075
(30, 10, 10)	10.498	0.982	1.474	0.077
(48, 1, 1)	77.339	1.041	1.056	0.0663

greater than 1. In this case, by appealing to (7) we estimate p by

$$\hat{p} = \frac{N}{n \int_1^\infty f(r, \hat{\theta})\, dr} \tag{17}$$

where the integral offsets the truncation. Table 2 also gives $\hat{p}$.

Table 2 shows that in most designs the fit as measured by $\chi^2(\hat{\theta})$ is quite remarkable taking into account the large number of observations that ranged from several hundreds to several thousands to tens of thousands. For example the designs (10, 10, 10), (4, 4, 4), and (24, 1, 1) correspond to sample sizes 723, 15,467, and 42,237, respectively. As the pixels become more separated in time and space the $\chi^2(\hat{\theta})$ values tend to behave very much as if the data were independent. For example, when every pixel is sampled in space, the χ^2 value becomes large, as can be seen from (48, 1, 1) in both GATE 1 and GATE 2. Also, the design (4, 4, 4) in GATE 2 shows that the χ^2 is inflated because of high dependence. But as we introduce further separation in time and space the χ^2 values deflate appreciably, as can be seen, for example, from the design (10, 10, 10) in both GATE 1 and GATE 2. Thus although we do not know the distribution of $\chi^2(\hat{\theta})$ because of the dependence in the data, the values in the table can help in forming an idea of what large and small $\chi^2(\hat{\theta})$ values are. It is interesting to note that the ninety-fifth percentile of the chi-square distribution with 6 degrees of freedom is equal to 12.6. It seems that with sufficient separation in time and space it is not unreasonable to judge the goodness of fit by comparing the $\chi^2(\hat{\theta})$ with this value. In fact, simulation results with independent data support this suggestion well by giving rise to minimum chi-square values of the same order of magnitude.

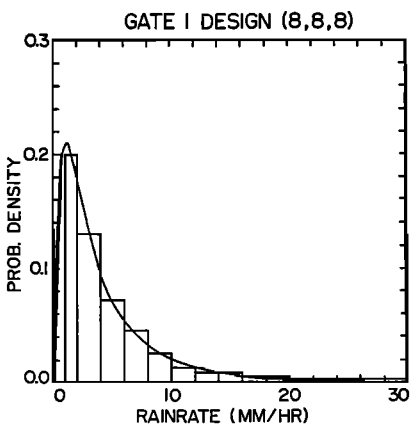

Fig. 1. The scaled histogram from the values of the (8, 8, 8) design and the fitted lognormal density with parameter θ = (1.140, 1.047).

However, it should be made clear that $\chi^2(\hat{\theta})$ values that behave as if the data were independent do not constitute a proof of independence. Such behavior is merely evidence in favor of independence.

Another fact manifested in Table 2 is that regardless of $\chi^2(\hat{\theta})$ value, the values of $\hat{\mu}$, $\hat{\sigma}^2$ obtained from the different designs are very close. For example, the GATE 1 the design (6, 6, 6) gives $\hat{\theta}$ = (1.152, 1.056), while the design (24, 1, 1) gives $\hat{\theta}$ = (1.159, 1.056), but the respective $\chi^2(\hat{\theta})$ values are 2.916 and 27.443. Similarly, in GATE 2 the designs (4, 4, 4) and (10, 5, 5) yield fairly close estimates, while the respective $\chi^2(\hat{\theta})$ values are 50.358 and 2.59. Can a "tight" design in time and space yield an excellent fit and still display a large $\chi^2(\hat{\theta})$? The answer to this question is affirmative, as the following argument shows.

Intuitively, we may argue that in the presence of dependence when a rain rate value falls into a certain bin, its immediate neighbors in time and space will tend to follow suit and fall into the same bin as well. This gives rise to a distortion of the distribution of the observed frequencies as compared with the case under independence and hence gives rise to abnormal values of $\chi^2(\hat{\theta})$. This is only a heuristic argument that can help in shaping up our intuition. A more technical explanation is based on the fact that $\chi^2(\theta)$, θ being the true parameter, is a quadratic form whose distribution tends as $N \to \infty$, under some conditions, to the distribution of a linear combination of independent chi-square random variables each with a single degree of freedom. To see that define $\hat{p}_i = o_i/N$, $p_i = E(\hat{p}_i)$ and the 8×1 column vectors $\mathbf{p} = (p_1, \cdots, p_8)'$, $\hat{\mathbf{p}} = (\hat{p}_1, \cdots, \hat{p}_8)'$, $\mathbf{1} = (1, \cdots, 1)'$, and let $\mathbf{A}$ be an 8×8 matrix given by

$$\mathbf{A} = \text{Diag}\left(\frac{1}{p_1}, \cdots, \frac{1}{p_8}\right) + \mathbf{1}\,\mathbf{1}'\frac{1}{p_9}$$

Then we can express $\chi^2(\theta)$ in (14) as a quadratic form

$$\chi^2(\theta) = \sum_{i=1}^{9} \frac{(o_i - e_i)^2}{e_i} = N(\hat{\mathbf{p}} - \mathbf{p})'\mathbf{A}(\hat{\mathbf{p}} - \mathbf{p}) \tag{18}$$

Assume that under some fairly general conditions as $N \to \infty$,

TABLE 3. Mean Average Rain Rate and Its Standard Error Approximation

GATE 1			GATE 2		
Design	$\hat{E}(R)$	Standard Error	Design	$\hat{E}(R)$	Standard Error
(5, 10, 10)	0.438	0.015	(5, 10, 10)	0.368	0.015
(6, 8, 8)	0.453	0.013	(8, 8, 8)	0.348	0.014
(8, 8, 8)	0.438	0.015	(10, 10, 10)	0.369	0.022
(10, 10, 10)	0.431	0.021	(20, 3, 3)	0.359	0.009
(10, 5, 5)	0.442	0.011	(30, 5, 5)	0.374	0.018

$$N^{1/2}(\hat{\mathbf{p}} - \mathbf{p}) \rightarrow N(\mathbf{0}, \mathbf{V})$$

Then it can be shown [see *Bishop et al.*, 1975, pp. 472–473]

$$\chi^2(\theta) \rightarrow \sum_{i=1}^{8} \lambda_i Z_i^2 \tag{19}$$

where the Z_i^2 are independent $\chi^2_{(1)}$ random variables. If the sampled rain rates are indeed independent, then $\lambda_i \equiv 1$ for all i, and when θ is known, $\chi^2(\theta)$ has an asymptotic $\chi^2_{(8)}$. But if the sampled rain rates are dependent, the λ_i may be smaller or larger than 1, and consequently (19) may be inflated or deflated. A similar argument can be made about $\chi^2(\hat{\theta})$. This means in particular that when the distribution of a $\chi^2(\hat{\theta})$ from a "tight" (in time and space) design is such that the $\chi^2(\hat{\theta})$ values are relatively large, we must reject the hypothesis of goodness of fit for larger values than those obtained under independence or near independence. It is important to mention that *Moore* [1982] has observed a similar effect in stationary Gaussian processes with positive autocorrelation function. He notes that in this case the Pearson chi-square statistic for testing fit to a normal law is stochastically larger than in the case of independent and identically distributed observations. To sum up, dependence can lead to large values of $\chi^2(\hat{\theta})$ despite a possible nearly perfect fit. See also *Kedem and Slud* [1981], who discuss a similar quadratic form whose values are inflated owing to dependent data.

3.5. *Estimation of the Mean of the Area Average Rain Rate*

The expected value of R in the mixed lognormal distribution is given in (12) as $E(R) = p\alpha$ where $\alpha = \exp(\mu + \sigma^2/2)$. By substituting our estimates for these parameters, we obtain an estimate for $E(R)$ denoted by $\hat{E}(R)$. Ideally, under independence and the ability to separate the zero from nonzero observations, the variance of $\hat{E}(R)$ can be approximated under lognormality (compare *Aitchison and Brown* [1963, p. 99])

$$\text{Var}(\hat{E}(R)) \simeq \hat{p}^2 \frac{\hat{\alpha}^2}{n_1}\left(\hat{\sigma}^2 + \frac{\hat{\sigma}^4}{2}\right) + \hat{\alpha}^2 \frac{\hat{p}(1-\hat{p})}{n} \tag{20}$$

where n is the total number of observations and n_1 is the number of nonzero observations. We emphasize that (20) is only a rough approximation, since our data are dependent, and that it probably tends to underestimate the true variance. Table 3 gives the estimates of the mean and their standard errors approximated by (20) and obtained from different designs. The estimated mean average rain rate during the second phase of GATE was considerably lower than that of the first phase. The table indicates the consistency of the results regardless of the design. It can be argued that in general, (20) should be viewed as an underestimate of Var $(\hat{E}(R))$, but for sampling designs that give sufficient separation in time and space our empirical work suggests that (20) is a reasonable approximation. (See also the discussion leading to Figure 2 below.)

4. LOGNORMAL VERSUS GAMMA

As was mentioned in the introduction, there is no general agreement as to the form of $f(r, \theta)$. Some authors such as *Neyman and Scott* [1967] report very good fit of the gamma distribution to precipitation amounts, while *Kedem and Chiu* [1987] argue that for area averages of rain rate, the lognormal model is appropriate provided that certain conditions are met. It is therefore interesting to compare these two distributions by applying minimum $\chi^2(\theta)$ as a criterion. Recall that the gamma density is given by

$$f(r, \theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)} r^{\alpha-1} \exp(-\lambda r) \qquad r > 0$$

$$f(r, \theta) = 0 \qquad r \le 0$$

where $\theta = (\alpha, \lambda)$, $\alpha, \lambda > 0$.

The above procedure was applied to the gamma density and the results corresponding to various designs are given in Table 4. Except for two cases (designs) where the two densities essentially performed equally well, the lognormal fit is much better as judged by $\chi^2(\hat{\theta})$. Moreover, the parameters obtained for the gamma distribution greatly underestimate the mean positive rain rate in GATE.

5. APPLICATION TO SATELLITE SAMPLING

Our final goal in this paper is to address the important problem of interpreting the above sampling designs for the purpose of remote sensing by a polar-orbiting satellite. One design that mimics satellite sampling is (48, 1, 1). A satellite following this sampling design can provide instantaneous snapshots of a large area of roughly the GATE dimension of 350 × 350 km^2, every 12 hours. The Tropical Rainfall Measuring Mission [*Simpson et al.*, 1988] proposes such a satellite whose tasks include the estimation of the mean area average rain rate over a given period. It is of great interest therefore to see if such a design is indeed appropriate for this purpose.

A slightly "tighter" design that allows more frequent visits over an area, such as (40, 1, 1), can improve the estimate appreciably. To see this increase in precision, we

TABLE 4. Comparison of Minimum $\chi^2(\theta)$ Applied to the Lognormal and Gamma Densities

Design	N	Lognormal			Gamma $\chi^2(\hat{\theta})$		
		$\chi^2(\hat{\theta})$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\alpha}$	$\hat{\lambda}$	$\chi^2(\hat{\theta})$
(30, 10, 10)	333	6.04	1.00	1.16	0.29	0.12	5.57
(20, 10, 10)	456	7.76	1.10	1.07	0.37	0.13	12.14
(10, 10, 10)	972	3.39	1.09	1.18	0.30	0.10	13.73
(5, 10, 10)	1976	6.80	1.06	1.21	0.34	0.10	32.46
(10, 5, 5)	3936	8.77	1.12	1.12	0.35	0.12	44.24
(5, 5, 5)	7889	16.83	1.11	1.13	0.35	0.12	85.87
(10, 20, 20)	219	4.98	1.32	1.00	0.49	0.12	12.39
(5, 30, 30)	263	6.53	1.09	1.41	0.26	0.09	4.09
(5, 20, 20)	461	0.80	1.19	1.07	0.41	0.12	7.21

applied the (48, 1, 1) and (40, 1, 1) designs to the GATE data starting from the first, second, third, etc., snapshot in succession. For example, the design (48, 1, 1) when starting from the first snapshot samples the first, forty-ninth, ninety-seventh, etc., snapshots, and when starting from the second snapshot it samples the second, fiftieth, ninety-eighth, etc. snapshots, and so on. Altogether this procedure yields 48 and 40 distinct estimates corresponding to the designs (40, 1, 1) and (48, 1, 1), respectively. The histograms from these two sets of estimates are given in Figure 2. The third histogram is of estimates obtained from the design (1, 10, 10) translated in space and is displayed for the purpose of comparison. There is a reduction in the standard deviation of the mean estimates from the (40, 1, 1) design as compared with the (48, 1, 1) design. This has a significant implication in studying the diurnal cycle, for the (40, 1, 1) design will sample through the diurnal cycle.

It is important to note that the mean and standard deviation given in Figure 2 for various sampling designs were obtained from the direct distinct estimates of $E(R)$ without any distribution assumption. Yet the results indicated on the histograms agree fairly well with the results, obtained by the minimum chi-square method using the lognormal model and (20), given in Table 3. This fact shows that (20) is not unreasonable after all.

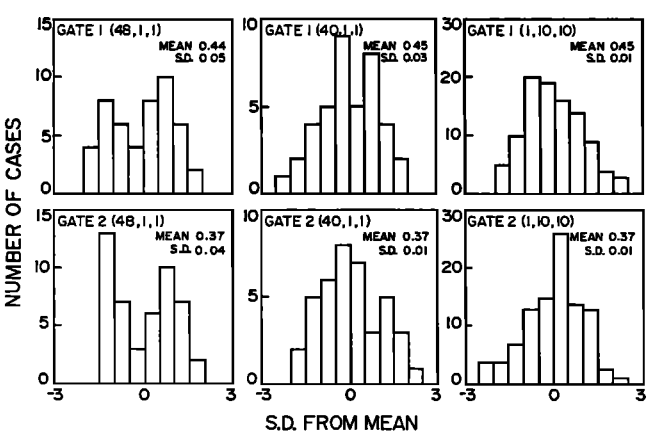

Fig. 2. Histogram of mean estimates from three designs applied repeatedly to GATE: (48, 1, 1) and (40, 1, 1) are translated in time, while (1, 10, 10) is translated in space. The results should be compared with those obtained from the minimum chi-square method given in Table 3.

6. Summary and Conclusions

We have shown that a truncated lognormal distribution provides an excellent fit for the area (4×4 km^2) average rain rate above 1 mm/h from GATE. A comparison with a truncated gamma, using area average rain rate above 1 mm/h, shows that for GATE the truncated lognormal gives a much better fit. Since the zero rain rates require no modeling, this suggests that a mixed lognormal distribution (not truncated) is very adequate for GATE. This finding is consistent with the development of *Kedem and Chiu* [1987]. However, it is very possible that other mixed distributions can provide adequate models as well.

We discussed the estimation of parameters in a mixed lognormal distribution amid dependent data by applying the method of minimum chi-square estimation combined with time-space sampling designs. By varying the designs we gained valuable insight into the dependence structure of the GATE data in time and space. The estimates from the various designs are (somewhat unexpectedly) quite close, a fact that points out the consistency of the minimum chi-square method despite of dependent data. The results obtained by the minimum chi-square method are in agreement with the direct estimates (Figure 2) obtained without recourse to an underlying distribution.

It has been observed that pixels separated by roughly 40 km in space and by 10 hours in time give rise to chi-square values that would have been obtained had the data been independent. This observation is in good agreement with the work of *Laughlin* [1981], who used a correlation approach to study the time dependence in GATE. He showed that the time correlation in 4×4 km^2 pixels decays relatively fast, particularly in GATE 2.

Perhaps the most important conclusion that we can draw from this work is that to the extent that the GATE data are representative of oceanic rainfall in the tropics, revisiting an area of roughly the GATE dimension (350 by 350 km^2) at a repetition rate of about once every 10 hours provides an excellent estimate (standard error of about 0.01 to 0.03) for the area average 3-week mean rain rate for the region. This is within the capability of a single space platform with scanning sensors in a low-inclination (tropical) orbit.

Acknowledgment. The remarks of a referee improved greatly the clarity of the paper. We are grateful to him.

References

Aitchison, J., and J. A. C. Brown, *The Lognormal Distribution*, Cambridge University Press, New York, 1963.

Arkell, R., and M. Hudlow, *GATE International Meteorological Radar Atlas*, U.S. Government Printing Office, Washington, D. C., 1977.

Atlas, D., and O. Thiele, Precipitation measurements from space, workshop report, NASA Goddard Space Flight Center, Greenbelt, Md., Oct. 1981.

Atlas, D., D. Rosenfeld, and D. A. Short, The estimation of convective rainfall by area integrals, I, The theoretical and empirical basis, paper presented at Conference on Mesoscale Precipitation: Analysis, Simulation, and Forecasting, Mass. Inst. of Technol., Cambridge, Sept. 13–17, 1988.

Austin, P. M., and S. G. Geotis, Evaluation of the quality of precipitation data from a satellite-borne radiometer, final report, grant NSG 5024, 30 pp., Mass. Inst. of Technol., Cambridge, 1978.

Austin, P. M., and S. G. Geotis, Precipitation measurements over the oceans, in *Air Sea Interaction*, edited by F. Dobson, L. Hasse, and R. Davis, Plenum, New York, 1980.

Berkson, J., Minimum chi-square, not maximum likelihood, *Ann. Stat.*, *81*, 457–469, 1980.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis*, MIT Press, Cambridge, Mass., 1975.

Bras, R. L., and R. Colon, Time-averaged areal mean of precipitation: Estimation and network design, *Water Resour. Res.*, *14*(5), 878–888, 1978.

Chiu, L. S., Rain estimate from satellites; Areal rainfall-rain area relation, paper presented at the 3rd Conference on Satellite Meteorology and Oceanography, Am. Meteorol. Soc., Anaheim, Calif., Jan. 31–Feb. 4, 1988.

Eagleson, P. S., Optimal density of rainfall networks, *Water Resour. Res.*, *3*, 1021–1033, 1967.

Feuerverger, A., On some methods of analysis for weather experiments, *Biometrika*, *66*, 655–658, 1979.

Houze, R. A., and C.-P. Cheng, Radar characteristics of tropical convection observed during GATE: Mean properties and trends over the summer season, *Mon. Weather Rev.*, *105*, 964–980, 1977.

Kedem, B., and L. S. Chiu, On the lognormality of rain rate, *Proc. Natl. Acad. Sci. U.S.A.*, *84*, 901–905, 1987.

Kedem, B., and E. Slud, On goodness of fit of time series models: An application of higher order crossings", *Biometrika*, *68*, 551–556, 1981.

Laughlin, C., On the effect of temporal sampling on the observation of mean rainfall, in *Precipitation Measurements from Space*, edited by D. Atlas, and O. Thiele, NASA Goddard Space Flight Center, Greenbelt, Md., 1981.

Lovejoy, S., and Schertzer, D., Generalized scale invariance in the atmosphere and fractal models of rain, *Water Resour. Res.*, *21*, 1233–1250, 1985.

Moore, D. S., The effect of dependence on chi squared tests of fit, *Ann. Stat.*, *10*(4), 1163–1171, 1982.

Neyman, J., and E. L. Scott, Some outstanding problems relating to rain modification, *Proc. Berkeley Symp. Math. Stat. Probl.*, *5*, 293–326, 1967.

Patterson, V. L., M. D. Hudlow, P. J. Pytlowany, F. P. Richards, and J. D. Hoff, GATE radar rainfall processing system, *Tech. Memo., EDIS 26*, Natl. Oceanic and Atmos. Admin., Washington, D. C., 1979.

Rodriguez-Iturbe, I., and J. M. Mejía, The design of rainfall networks in time and space, *Water Resour. Res.*, *10*(4), 713–728, 1974.

Rosenfeld, D., D. Atlas, and D. A. Short, The estimation of convective rainfall by area average integrals, II., paper presented at Conference on Mesoscale Precipitation, Mass. Inst. of Technol., Cambridge, Sept. 13–16, 1988.

Simpson, J., R. F. Adler, and G. R. North, A proposal tropical rainfall measuring mission (TRMM) satellite, *Bull. Am. Meteorol. Soc.*, *69*, 278–295, 1988.

L. S. Chiu, Applied Research Corporation, Landover, MD 20785.

B. Kedem, Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742.

G. R. North, NASA Goddard Space Flight Center, Greenbelt, MD 20771.

(Received March 31, 1988;
revised July 20, 1989;
accepted August 15, 1989.)