PNAS
Proceedings of the National Academy of Sciences of the United States of America www.pnas.org

On the Lognormality of Rain Rate

Benjamin Kedem, and Long. S. Chiu

PNAS 1987;84;901-905
doi:10.1073/pnas.84.4.901

Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 901–905, February 1987
Applied Mathematical Sciences

On the lognormality of rain rate

(Galton–Watson process/stochastic regression/martingale difference/global atmospheric research program, Atlantic tropical experiment data)

BENJAMIN KEDEM[†‡] AND LONG. S. CHIU[§]

[†]Department of Mathematics, University of Maryland, College Park, MD 20742; and [§]Applied Research Corporation, Landover, MD 20785

Communicated by Robert Zwanzig, October 22, 1986 (received for review July 28, 1986)

ABSTRACT **A stochastic regression model is used in modeling rain rate. Under some conditions on the model parameters, it is shown that rain rate is asymptotically lognormal. An application of the model to the GATE (global atmospheric research program, Atlantic tropical experiment) data shows a remarkable agreement between the assumed and estimated model parameters for rain rate averaged over sufficiently large area and a sampling interval of 15 min.**

There is ample evidence based on observations that rain characteristics tend to be approximately lognormally distributed. This observation is shared by quite a few research workers who considered different data sets. These pertain to the amount and duration of rainfall and to horizontal and vertical cloud extent in tropical and extratropical regions under a wide variety of convective conditions (1–4). Even more intriguing is the fact that area averages of rain rate tend to follow a lognormal rather than the "expected" normal distribution (4). The questions are then, What makes the lognormal distribution so prevalent when it comes to rain systems, and is there any theoretical basis for these observational findings? On practical grounds, we may ask whether it makes sense at all to fit a lognormal distribution to rain characteristics and under what conditions. This is the subject of the present note. We will focus on the lognormality of rain rate.

Many authors believe that the lognormal distribution is a natural outcome of the so-called law of proportionate effect (5). Accordingly, $\{X_j\}$ satisfies the law of proportionate effect if

$$X_j - X_{j-1} = \varepsilon_j X_{j-1},$$

where the ε_js are mutually independent and are also independent of the X_js. While the law of proportionate effect is of fundamental importance in motivating the lognormal distribution, the independence assumption on the ε_j is quite restrictive and can in fact be relaxed. It is sufficient that the ε_js obey conditions that guarantee the asymptotic normality of sums in terms of these variates. For this to hold, they need not be independent and may even be dependent on the X_js.

In the present note we discuss a certain type of dynamic regression model that, with less restrictive conditions, helps to explain the observed lognormality of rain rate. The model has a strong intuitive appeal and is quite flexible in that it requires only a few parameters that can be easily estimated from data. Using a specific estimation procedure, the model is fitted to the GATE (GARP—global atmospheric research program—Atlantic tropical experiment) data. It is shown that some requirements for asymptotic lognormality are satisfied by the data. Furthermore, realizations produced by the model appear to be very similar to those produced by real rain-rate data.

It should be emphasized that our result is model-based and that by itself does not constitute a proof that rain rate is precisely lognormally distributed. We merely provide reasonable conditions that lead to lognormality, and indeed some of our conditions are well supported by the GATE data. It seems to us that the present approach is an improvement over the approach that solely relies on the law of proportionate effect.

A Stochastic Model for Rain Rate

To unravel the lognormal mystery, we begin with a rather naive notion of a rain element. Conditional on rain, we conceive of a rain element as a volume in space containing small droplets of water that have the following dynamics. Let time be discrete. At the $n - 1$ time step, some droplets give rise to a new generation of droplets through a complicated physical process, some droplets leave the volume while new ones, called immigrants, arrive to join the droplets of the new generation. It is really a process of replacement and immigration where the replacement refers to droplets already in the volume. The droplets are being replaced by a non-negative number of droplets where zero could mean complete departure or emigration. Thus at time n, the number of droplets in the volume in space is the sum of the replacement droplets and the immigrants. Let X_{n-1} stand for the (random) number of droplets in the volume at time $n - 1$ and suppose the ith droplet there is replaced by $Y_{n,i}$ fresh droplets while I_n denotes the number of immigrants. Then at time n, the rain element contains

$$X_n = \sum_{i=1}^{X_{n-1}} Y_{n,i} + I_n \qquad (n = 1, 2 \ldots) \qquad [1]$$

droplets with the convention that $\Sigma_1^0 \equiv 0$. For Eq. **1** to cover dry periods and shifts from dry (wet) to wet (dry) periods, the following interpretation is adopted. Most of the time when it is not raining, the rain element is dry and both X_n and I_n vanish. The rain element becomes active as soon as I_n admits a positive value. This sets the X_n, and hence the $Y_{n,i}$, in motion until the X_n vanish. The process restarts when I_n admits again a positive value. I_n can be thought of as the part of the process responsible for the occurrence of rain storms while $\Sigma Y_{n,i}$ pertains to the duration and amount of rain.

The most important parameters associated with the dynamic model (Eq. **1**) are

$$EY_{n,i} = m, \ EI_n = \lambda \qquad (n, i = 1, 2 \ldots).$$

No further assumption is needed for the present use of the model except for *Assumptions 1* and *2* below.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: GATE, global atmospheric research program, Atlantic tropical experiment.
[‡]To whom reprint requests should be addressed.

When the occurrences of rain are not too frequent, we expect λ to be small and close to zero. When it does rain, it usually persists for a while before it stops. This means that m should be close to 1 but still strictly less than 1. If m is greater than or equal to 1, the duration and amount can be explosive. Thus an indication of goodness of fit of Eq. **1** to rain-rate data is small λ and m close to but smaller than unity. It is interesting to apply the model to real data to see if these conditions are met.

When $\{Y_{n,i}\}$, $\{I_n\}$ are families of mutually independent, non-negative-integer-valued random variables, the process $\{X_n\}$ is called a Galton–Watson process with immigration (6). This type of process was introduced as early as 1915 by Smoluchowski (7) whose work is reported by Chandrasekhar (8). Smoluchowski (7) used the model to study the fluctuations in the number of particles contained in a small volume that exhibit random motion. However, we do not necessarily require the *Y*s and *I*s to be independent.

There is a well-known device that transforms Eq. **1** into a more convenient regression equation that takes into account past values of X_n (9, 10). Let $\mathcal{F}_n$ be the σ-field generated by the random variables $(X_0, X_1, \ldots, X_n)$, and note that

$$E(X_n|\mathcal{F}_{n-1}) = mX_{n-1} + \lambda.$$

Define ε_n by the difference

$$\varepsilon_n = X_n - E(X_n|\mathcal{F}_{n-1}),$$

and write Eq. **1** as

$$X_n = mX_{n-1} + \lambda + \varepsilon_n. \quad [2]$$

Then $\{X_n\}$ is seen to be a stochastic difference equation where ε_n is a martingale difference (11); i.e., ε_n is $\mathcal{F}_n$-measureable and $E(\varepsilon_n|\mathcal{F}_{n-1}) = 0$ for every n. An important example is the case of independent ε_n with mean 0, which is not required here. Other than its formal importance as expressed in Eq. **2**, martingale differences follow the central limit theorem under quite general conditions.

Since X_n refers to the density of droplets in the rain element, it is related to the rain rate. But multiplication of Eq. **2** by a constant leaves the model intact, and we can actually think of X_n as representing rain rate. We, therefore, model rain-rate dynamics by Eq. **2** where X_n admits only non-negative values.

Continuity Assumption

In its present form, Eq. **2** is a fairly general model that could represent a wide range of physical and statistical processes. To ensure the lognormality of X_n, some more assumptions are needed.

Let $\{X_n\}$, where $n = 0, 1, \ldots$, be the stochastic process (Eq. **2**) that stands for the rain-rate process at a given rain element. Assume that the $X_0, X_1, X_2, \ldots$ are readings at time $0, T, 2T, \ldots$, where the sampling interval T is small. The main assumption we shall adhere to is that of continuity: when the sampling interval T is sufficiently small, we require that, conditional on rain, X_n and X_{n-1} be close to each other as is the case with many continuous phenomena in nature. This assumption is reasonable when X_n represents the average rain rate over a sufficiently large area sampled at short time intervals. For normality we also require the sum of squares of the ε_n to explode. More precisely, conditional on rain (i.e., positive X_ns) we assume the following.

ASSUMPTION 1. $|X_i - X_{i-1}| << X_{i-1}$.

ASSUMPTION 2. $\frac{1}{n}\sum_{i=1}^{n} E[(\varepsilon_i/X_{i-1})^2|\mathcal{F}_{i-1}] \rightarrow c^2 > 0$, *a.s. (almost surely)* $(n \rightarrow \infty)$.

Since $E(\varepsilon_i/X_{i-1}|\mathcal{F}_{i-1}) = 0$, and since by *Assumption 1*, ε_i/X_{i-1} is essentially bounded as $m \rightarrow 1$ and $\lambda \rightarrow 0$, it follows that (12, 13)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i/X_{i-1} \xrightarrow{d} N(0, c^2) \qquad (n \rightarrow \infty).$$

Asymptotic Lognormality of Rain Rate

Let $\chi[A]$ be the indicator of the event A, and define δ_n by

$$\delta_n = (X_n - X_{n-1})/(X_{n-1} + \chi[X_{n-1} = 0]).$$

Then Eq. **2** can be written as

$$X_n = (1 + \delta_n)X_{n-1} + (\lambda + \varepsilon_n)\,\chi[X_{n-1} = 0]. \quad [3]$$

Thus, conditional on rain (i.e., $X_0, X_1, X_2, \ldots$, all positive), it follows that

$$X_n = (1 + \delta_n)(1 + \delta_{n-1}) \ldots (1 + \delta_1)\,X_0, \quad [4]$$

from which we obtain by *Assumption 1* that

$$\log(X_n/X_0) \simeq \sum_{i=1}^{n} \delta_i, \quad [5]$$

or

$$\log(X_n/X_0) + \sum_{i=1}^{n} [(1 - m) - \lambda/X_{i-1}] \simeq \sum_{i=1}^{n} \varepsilon_i/X_{i-1}. \quad [6]$$

Therefore, for m sufficiently close to 1 and λ close to 0, *Assumptions 1* and *2* imply that for large n

$$\left(\frac{X_n}{X_0}\right)^{1/\sqrt{n}} \left[1 + (1 - m) - \frac{\lambda}{n}\sum_{i=1}^{n}\frac{1}{X_{i-1}}\right]^{\sqrt{n}} \sim \Lambda\,(0, c^2), \quad [7]$$

where $\Lambda(0, c^2)$ denotes the lognormal distribution with parameters 0 and c^2 (5). When $m \rightarrow 1$ and $\lambda \rightarrow 0$, we obtain the useful approximation

$$\left(\frac{X_n}{X_0}\right)^{1/\sqrt{n}} \sim \Lambda\,(0, c^2). \quad [8]$$

The 0 parameter is expected if we assume that X_n for large n is independent of X_0 and that the two are identically distributed. Under these conditions both $X_n^{1/\sqrt{n}}$ and $X_0^{1/\sqrt{n}}$ are asymptotically $(\mu, c^2/2)$ for some μ (5).

Statistical Estimation of *m* and λ

A great deal of the foregoing discussion depends on m being close to but strictly smaller than 1, and λ being positive but close to 0. To verify these conditions, the parameters should be estimated as precisely as possible. Fortunately, this estimation problem is a special case of a general problem investigated in detail by Lai and Wei (11) who give conditions under which the least squares estimates converge almost surely to the respective parameters. Winnicki (10) has suggested that m and λ should be estimated from the weighted model

$$X_n/(X_{n-1}+1)^{1/2} = mX_{n-1}/(X_{n-1}+1)^{1/2} + \lambda/(X_{n-1}+1)^{1/2} + \varepsilon_n^*, \quad [9]$$

where $\varepsilon_n^* \equiv \varepsilon_n/(X_{n-1}+1)^{1/2}$, by minimizing the sum of squares of the ε_n^*. The estimates obtained in this way are called weighted least squares and are shown, under some conditions, to be superior to the ordinary least squares when m is close to 1. Now, the Lai and Wei theory (11) can be applied to the stochastic regression model (Eq. **9**), since ε_n^* in Eq. **9** is still a martingale difference. This is done next.

Denote the weighted least squares estimators by $\hat{m}$ and $\hat{\lambda}$, and the design matrix by $\mathbf{X}_n$. Then

$$X_n = \begin{pmatrix} X_1/(X_1+1)^{1/2} & 1/(X_1+1)^{1/2} \\ X_2/(X_2+1)^{1/2} & 1/(X_2+1)^{1/2} \\ \vdots & \vdots \\ X_n/(X_n+1)^{1/2} & 1/(X_n+1)^{1/2} \end{pmatrix}$$

Define a 2 × 2 matrix **A** by, $\mathbf{A} \equiv \mathbf{X}_n' \mathbf{X}_n$ and let $\lambda_{\min}(n)$ and $\lambda_{\max}(n)$ be, respectively, the smaller and larger eigenvalues of **A**. Then the relevant result of Lai and Wei (11) can be stated as follows, assuming model Eq. **9**. Assume

(i) $\sup_n E(|\varepsilon_n^*|^\alpha|\mathcal{F}_{n-1}) < \infty$ a.s. for some $\alpha > 2$,

and that

(ii) $\lambda_{\min}(n) \to \infty$, such that as $n \to \infty$,

$\log \lambda_{\max}(n) = o[\lambda_{\min}(n)]$ a.s.

Then

$$(\hat{m}, \hat{\lambda}) \to (m, \lambda) \quad \text{a.s.}$$

Thus, when conditions i and ii are satisfied, the result guarantees a strong sense of convergence of the weighted least squares estimates. The estimates themselves are given in ref. 10 as

$$\hat{m} = \frac{\sum_{i=1}^n X_i \sum_{i=1}^n 1/(X_{i-1}+1) - n\sum_{i=1}^n X_i/(X_{i-1}+1)}{\sum_{i=1}^n (X_{i-1}+1)\sum_{i=1}^n 1/(X_{i-1}+1) - n^2}, \quad [10]$$

$$\hat{\lambda} = \frac{\sum_{i=1}^n X_{i-1}\sum_{i=1}^n X_i/(X_{i-1}+1) - \sum_{i=1}^n X_i \sum_{i=1}^n X_{i-1}/(X_{i-1}+1)}{\sum_{i=1}^n (X_{i-1}+1)\sum_{i=1}^n 1/(X_{i-1}+1) - n^2}, \quad [11]$$

where n is the series size.

Since observed rain rate is finite, condition i is automatically satisfied. To verify condition ii analytically is difficult in general, but it can be verified from data. The rain-rate data we have in mind are described in the next section. For rain-rate averages obtained from squares of 32 × 32 km² at 15-min intervals, the results from a typical time series are given in Table 1. The series size ranges from n = 100 to n = 1700, and it is seen that condition ii is satisfied since $\lambda_{\min}(n)$ tends to infinity faster than $\log[\lambda_{\max}(n)]$. Similar results were obtained for other time series and so, for all practical purposes, the door is now open to the actual estimation of m, λ using these data.

Table 1. An example where condition ii is satisfied

n	$\lambda_{\min}(n)$	$\lambda_{\max}(n)$	$[\log \lambda_{\max}(n)]/\lambda_{\min}(n)$
100	6.368	97.212	0.719
200	53.972	185.318	0.097
400	108.249	351.806	0.054
600	142.773	533.428	0.044
800	151.526	722.730	0.043
1000	421.242	901.590	0.016
1200	438.366	1084.106	0.016
1500	475.987	1359.853	0.015
1700	514.737	1540.269	0.014

The rain-rate series are sampled every 15 min over a square of 32 × 32 km².

Application To GATE Data

We applied the model to rainfall data collected during GATE. GATE was conducted in the summer of 1974. During roughly three triweekly periods, detailed rainfall measurements from rain gauges and radars on an array of research vessels were made over an area called the B-scale. The B-scale encompasses an area of about 400 km in diameter. Arkell and Hudlow (14) composited the radar ship data and presented 15-min radar reflectivity scan data. The radar reflectivity data are converted to rain rates that are binned into 4 × 4 km² pixels in ref. 15. This data set is probably as yet one of the most extensive rainfall measurements made over the oceans.

Time series of rain rate for individual pixels (4 × 4 km² resolution) and for area averages 10 × 10 pixels (or 40 × 40 km²) have been extracted from the first triweekly period in GATE (called phase 1). The parameters of the model are estimated by the method of weighted, least squares described above. Table 2 gives the estimated m and λ for 10 × 10 pixel arrays and for individual pixels situated at the center of the GATE area.

The results for 20 time series obtained from large area averages of 10 × 10 pixels are shown in Table 2. For each time series m and λ are estimated using Eqs. **10** and **11**. The estimated m are very close to but less than 1 while $\hat{\lambda}$ is quite small. We see that, for large area averages sampled (really visited!) at T = 15-min intervals, the results are satisfactory and so a lognormal fit makes good sense.

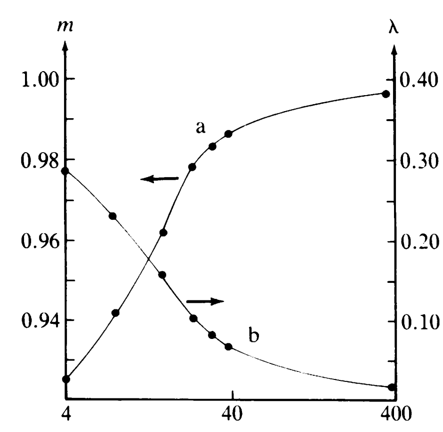

FIG. 1. Curve a, the monotone increase in $\hat{m}$, and curve b, the monotone decrease in $\hat{\lambda}$ as functions of the pixel size (square root of the area).

Table 2. Pairs of estimates ($\hat{m}$, $\hat{\lambda}$) for 40 × 40 km² and 4 × 4 km² pixels obtained from 20 time series in the center of the GATE area

40 × 40 km² pixel					4 × 4 km² pixel				
0.98, 0.05	0.97, 0.05	0.92, 0.09	0.97, 0.05	0.96, 0.07	0.93, 0.40	0.94, 0.38	0.94, 0.41	0.93, 0.50	0.90, 0.68
0.98, 0.07	0.99, 0.05	0.98, 0.05	0.98, 0.07	0.98, 0.08	0.85, 0.51	0.91, 0.40	0.93, 0.37	0.94, 0.38	0.94, 0.42
0.98, 0.08	0.99, 0.06	0.99, 0.06	0.99, 0.06	0.99, 0.07	0.88, 0.34	0.92, 0.32	0.95, 0.22	0.94, 0.34	0.90, 0.61
0.98, 0.08	0.98, 0.08	0.99, 0.06	0.99, 0.08	0.99, 0.07	0.88, 0.38	0.89, 0.39	0.94, 0.21	0.91, 0.37	0.92, 0.50

For individual pixels (4 × 4 km²) m is still fairly large although not as close to 1 as in the 10 × 10 pixel array case, but λ is relatively large as seen from Table 2. The reason for this can be attributed to the 15-min sampling interval: for smaller pixels we need to sample more often than 15 min to achieve results similar to those for large pixels. This suggests that the lognormal limit is approached for large aggregates at the 15-min sampling rate, and more generally, that there exists a time scale that corresponds to a spatial scale. This dependence of the model parameters on the averaging area can be seen very clearly from Fig. 1 where $\hat{m}$ and $\hat{\lambda}$ are given as a function of the pixel size (i.e., the averaging area) while the sampling interval is fixed at T = 15 min. The pixel sizes examined are 4 × 4, 8 × 8, 16 × 16, 24 × 24, 32 × 32, 40 × 40, and 352 × 352 km². Thus our theoretical considerations suggest that lognormality of positive rain rate can already be observed fairly closely by averaging over pixels whose area is roughly as small as 40 × 40 km² where the sampling frequency is 15 min. This finding is enhanced by a histogram plot in Fig. 2 derived from 53,600 GATE pixels of 40 × 40 km². The figure displays the distribution of the area averages of positive rain rate on a logarithmic scale. The distribution appears to be fairly symmetric in support of the above discussion. A corresponding histogram on a linear scale together with a matched theoretical histogram from the lognormal distribution $\wedge$ (−1.438, 3.614) are shown in Fig. 3. Unfortunately, we cannot attach to this fit the usual statistical measures of goodness of fit due to the high degree of dependence in the data.

Simulation Versus Real Data

We end this note with a short graphical comparison between a time series from Eq. 1 and a typical time series from the GATE data. It should be noted that in the foregoing discussion we made no restrictions on the Ys and Is in Eq. 1 except for the requirements that they be non-negative integers. In fact Eq. 2 is a more general model since even this last restriction is removed. Thus, if Eq. 1 is capable of producing realizations that resemble real rain-rate data, this shows all the more the adequacy of Eq. 2 that is the model we used all along in the foregoing discussion.

Now, there are many ways to simulate Eq. 1. One simple and fast way is to take the Ys and Is as independent Poisson random variables with parameters m and λ, respectively. By this process we generated the time series in Fig. 4*B*. Fig. 4*A* shows a typical time series from GATE that constitutes 100 hr. The sudden bursts of rain storms, duration, intensity, decay, and inter-arrival times between storms in the real and simulated realizations are quite intriguingly similar.

Summary

The puzzling experimental fact that rain rate tends to follow a lognormal distribution was explained with the aid of a model. Accordingly, under some conditions, as a rain storm de-

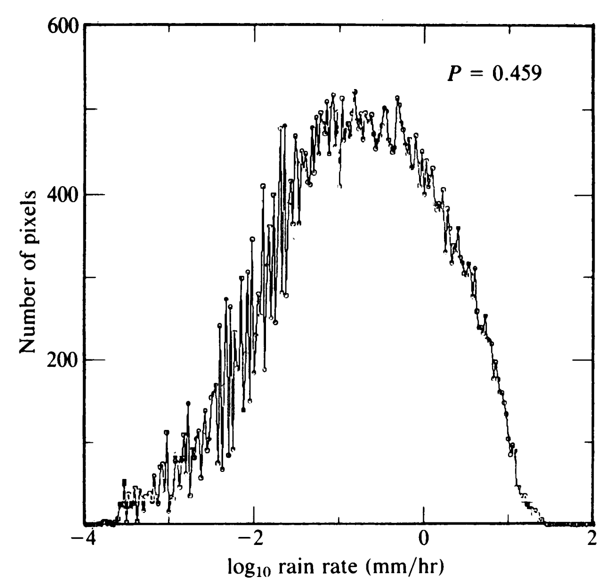

FIG. 2. A histogram of $\log_{10}$ of the rain rate obtained from a large number of 40 × 40 km² GATE pixels.

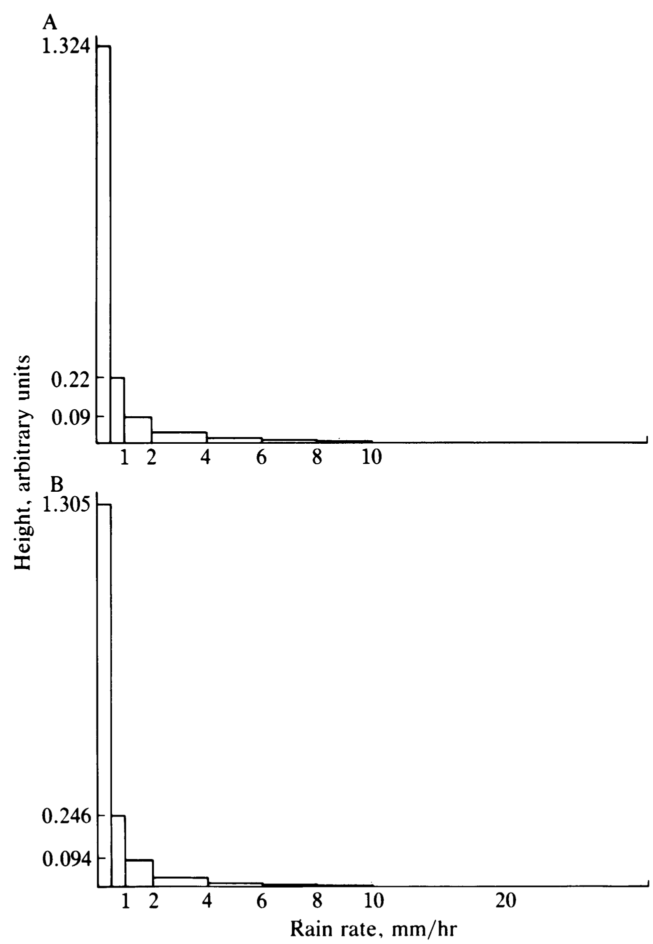

FIG. 3. A histogram (*A*) of area averages of rain rate obtained from a large number of 40 × 40 km² GATE pixels, and a corresponding (*B*) lognormal histogram from $\wedge$ (−1.438, 3.614).

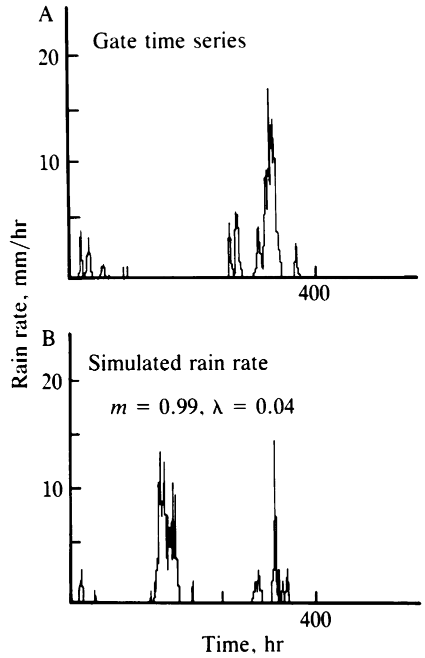

FIG. 4. The 400 observations from a typical GATE time series taken every 15 min where the pixel size is 32 × 32 km^2 (*A*), and 400 observations from Eq. 1 with $m = 0.99$ and $\lambda = 0.04$ (*B*).

velops, rain rate tends to follow a lognormal distribution. The conditions on the model parameters are shown to be satisfied fairly closely by the GATE data for time series that consist of rain-rate averages over sufficiently large pixels observed every 15 min. A variant special case of the model is capable of producing realizations that appear to be very similar to real rain-rate time series. Another fact is that the eigenvalue conditions needed for the almost sure convergence of the weighted, least squares estimates are well satisfied by the GATE data. In light of all these consistencies, it is hoped that the model (Eq. 2) can serve in settling other intriguing facts about rain.

The authors wish to express their gratitude to G. R. North and J. Winnicki for stimulating discussions. B.K. was partially supported by Office of Naval Research Grant N00014-86-K0007. L.S.C. was supported by National Aeronautics and Space Administration Grant NAS-29271.

1. Biondini, R. (1976) *J. Appl. Meteorol.* **15,** 205–224.
2. Lopez, R. E. (1977) *Mon. Weather Rev.* **105,** 865–872.
3. Houze, R. A. & Cheng, C. P. (1977) *Mon. Weather Rev.* **105,** 964–980.
4. Chiu, L. S., Kedem, B. & North, G. R. (1986) *Estimating the Mean Areal Average Rainfall: A Mixed Distribution Approach*, preprint.
5. Aitchison, J. & Brown, J. A. C. (1963) *The Lognormal Distribution* (Cambridge Univ. Press, London), pp. 12, 22.
6. Athreya, K. B. & Ney, P. E. (1972) *Branching Processes* (Springer, New York), p. 263.
7. Smoluchowski, M. (1916) *Physik. Zeits.* **17,** 557–585.
8. Chandrasekhar, S. (1943) *Rev. Mod. Phys.* **15,** 1–89.
9. Heyde, C. C. & Seneta, E. (1972) *J. Appl. Probab.* **9,** 235–256.
10. Winnicki, J. (1986) Dissertation (Univ. of Maryland, College Park).
11. Lai, T. L. & Wei, C. Z. (1982) *Ann. Stat.* **10,** 154–166.
12. McLeish, D. L. (1974) *Ann. Probab.* **2,** 620–628.
13. Basawa, I. V. & Prakasa Rao, L. S. (1980) *Statistical Inference for Stochastic Processes* (Academic, New York), p. 388.
14. Arkell, R. & Hudlow, M. (1977) *GATE International Meteorological Radar Atlas* (Government Printing Office, Washington, DC).
15. Patterson, V. L., Hudlow, M. D., Pytlowany, P. J., Richards, F. P. & Hoff, J. D. (1979) *GATE Radar Rainfall Processing System* (Natl. Oceanic and Atmospheric Admin., Washington, DC), NOAA Tech. Memo. EDIS 26.