Variability of Space–Time Mean Rain Rate

B. KEDEM AND R. PFEIFFER

Mathematics Department and Institute for Systems Research, University of Maryland at College Park, College Park, Maryland

D. A. SHORT

Laboratory for Atmospheres, NASA/GSFC, Greenbelt, Maryland

(Manuscript received 1 February 1996, in final form 25 September 1996)

ABSTRACT

A mixed lognormal distribution is fit to rain-rate data for the purpose of estimating the space–time mean. Using Fisher information, the large sample variance is obtained for grouped and ungrouped data estimates. The asymptotic variance results are used in deriving the efficiency of the grouped data estimator as a function of the mixed lognormal parameters. The method is applied to data from the Tropical Ocean Global Atmosphere Coupled Ocean–Atmosphere Response Experiment binned into 2 km × 2 km and 4 km × 4 km pixels. The variance of estimators is smaller for the 4 km × 4 km case, indicating that the lognormal model is more appropriate for the lower-resolution data.

1. Introduction

The Tropical Rainfall Measuring Mission (TRMM) presents a great challenge to the field of statistics. Its purpose is the indirect estimation of mean rain rate over an area throughout a period of time by means of an array of spaceborne instruments. In dealing with this problem, the objective of this paper is to discuss an estimation procedure for mean rain rate outlined in Kedem et al. (1990) and to quantify the variance of the resulting estimate using large-sample methodology.

To do so, the first step is to identify precisely the random variable in question. The data dealt with are a finite collection of averaged rain rates over 4 km × 4 km nonoverlapping pixels, as in the Global Atmospheric Research Program (GARP) Atlantic Tropical Experiment (GATE), or 2 km × 2 km pixels, as in Tropical Oceans Global Atmosphere Coupled Ocean–Atmosphere Response Experiment (TOGA COARE) 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 given area in the given period of time. In other words, it is an infinite population of values represented by a random variable that assigns to every ''instantaneous pixel'' the corresponding average rain rate, and the distribution of the instantaneous average rain rates is identified with the distribution of the random variable. Having thus defined the random variable, denoted by X, its distribution can now be studied in a meaningful way.

The variable X is of a mixed type. This means that X takes the value 0 (no rain) with positive probability $1 - p$ and $0 < p < 1$; otherwise, conditional on rain, X has a positive continuous distribution with distribution function $F(x, \theta)$. The distribution function of X is therefore

$$P(X \le x) = G(x) = (1 - p)H(x) + pF(x, \theta),\ x \ge 0,$$

where $H(x)$ stands for the step function, $H(x) = 0$ for $x < 0$, and $H(x) = 1$ for $x \ge 0$.

The generalized density function of X is

$$g(x) = (1 - p)^{1 - I[x>0]}\,[pf(x, \theta)]^{I[x>0]},\ x \ge 0,$$

where $f(x, \theta)$ denotes the density function of X conditional on rain and $I(A)$ is the indicator function of the event A.

Meteorological experiments performed at several regions on the globe have provided many datasets of rain-rate measurements during different seasons. From statistical analysis it can be concluded that the probability distribution of rain rate, when it rains, is unimodal and highly skewed to the right. A model that seems reasonable for modeling rain rate, conditional on rain, is the two-parameter lognormal distribution $\Lambda(x \mid \mu, \sigma)$.

The mean rain rate over a fixed area is $E(X) = pE(X \mid X > 0)$. If the continuous part is described by the lognormal distribution $\Lambda(x \mid \mu, \sigma)$, then the mean rain rate satisfies $E(X) = pe^{\mu + \frac{1}{2}\sigma^2}$.

Previous efforts at the estimation of mean rain rate usually focused on estimating the mean of the contin-

Corresponding author address: David A. Short, Laboratory for Atmospheres, NASA/Goddard Space Flight Center, Greenbelt, MD 20771.
E-mail: short@trmm.gsfc.nasa.gov

uous part of the distribution $E(X \mid X > 0)$ and used the proportion of zero observations in the dataset to estimate $1 - p$. But herein lies one of the difficulties of the problem. The instruments involved only measure rain rate or related quantities within a certain range. A spaceborne weather radar is usually limited at low rain rates by small signal-to-noise ratios and at high rain rates by attenuation effects. The data analyzed here were observed by 5-cm surface-based radars, which are also affected by the same problems. Especially in the case of rain rates close to zero it is usually impossible to tell whether one is dealing with an observation that falls outside the instrument range, a noisy observation, or a case ''no rain.'' Therefore, including a contaminated proportion of zeros among the data in estimating the probability of rain in a given area over a given time period may give erroneous results.

One way to overcome the obstacle of inaccurate data is to group the observed values into categories. Following Kedem et al. (1990), we use minimum chi-square estimation for the parameters, but in the present work the whole range of data—including zero—is divided into categories, not just the data points greater than a given positive threshold.

A similar approach was taken by Meneghini and Jones (1993). Within the dynamic range of measurable rain rates of an instrument, the idea of ''multiple thresholds'' was used to estimate the distribution function of rain rate over a given area. A model cumulative distribution function (cdf) was chosen with unknown parameters adjusted so that the mean-square error between the model cdf and the empirical distribution function was minimized. Once the unknown parameters are obtained and the model distribution function is fully specified, first-order statistics of the distribution can be derived. Using random sampling, this method can provide the variance of the resulting estimates by appealing to nonlinear least squares theory.

Single and double truncation, together with maximum likelihood estimation, was also used by Hong (1994) to fit a mixed lognormal distribution to rain-rate data. But aside from the truncation there was no further grouping of the data.

Our approach is somewhat different and is within the statistical framework of categorical data. We use minimum chi-square estimation to get an estimate for mean rain rate and use asymptotic theory to derive the variability and confidence intervals of the estimate. We also quantify the amount of information that gets lost by grouping the measurements.

2. Minimum chi-square estimation for mean rain rate

The key idea of the method described herein—introduced by Pearson (1900)—is grouping of the data. As mentioned before, the measurements dealt with are very coarse, and grouping is to some degree a way to overcome this obstacle. In practice, all continuous data are subject to the limits of measurement accuracy, but the inherent grouping may be so fine as to have negligible effect. With large amounts of data, as is the case for satellite observations, grouping may also facilitate display and handling of the data.

The procedure runs as follows (e.g., see Rao 1973, his chapter 5). Given a random sample, the real line is partitioned into k categories and the number of data points that fall into each category is counted, giving us the set of observed frequencies $n_1, \ldots, n_k$. The hypothetical probabilities $\pi_1(\boldsymbol{\theta}), \ldots, \pi_k(\boldsymbol{\theta})$, where $\pi_j(\boldsymbol{\theta}) = P$(observation falls into the jth category), are functions of the unknown parameters. A measure of discrepancy is defined between the observed frequencies $n_1, \ldots, n_k$ and the hypothetical expectations $n\pi_1(\boldsymbol{\theta}), \ldots, n\pi_k(\boldsymbol{\theta})$, where $n = \Sigma n_i$. The estimates for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ are obtained by minimizing such a measure with respect to $\boldsymbol{\theta}$.

The specific measure used in the paper is the minimum chi-square statistic

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{[n_i - n\pi_i(\boldsymbol{\theta})]^2}{n\pi_i(\boldsymbol{\theta})}. \tag{1}$$

The minimum chi-square estimator is therefore the solution of

$$\sum_{i=1}^{k} \left[\frac{n_i}{\pi_i(\boldsymbol{\theta})}\right]^2 \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_k} = 0, \qquad k = 1, \ldots, m.$$

This estimator is asymptotically equivalent to the grouped data MLE (maximum-likelihood estimate)—that is, the solution of the equations

$$\sum_{i=1}^{k} \frac{n_i}{\pi_i(\theta)} \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_k} = 0, \qquad k = 1, \ldots, m.$$

This fact is used later to obtain the asymptotic variance of the minimum chi-squared estimator (MCE). Fisher showed that the approximate distribution of $\chi^2(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ denotes the estimate that minimizes (1), is $\chi^2(k - m - 1)$; when any other method of estimation is used, this conclusion may not be true. Large values of this statistic—that is, values in the upper tail of the $\chi^2(k - m - 1)$ distribution—therefore are evidence for lack of fit.

As stated earlier, the distribution function of instantaneous rain rate X is assumed to be

$$P(X \le x) = G(x) = (1 - p)H(x) + pF(x) \qquad x \ge 0, \tag{2}$$

where $H(x)$ stands for the step function, $H(x) = 0$ for $x < 0$, and $H(x) = 1$ for $x \ge 0$. Here, $F(x)$ has a density function $f(x)$ that depends on some unknown parameters. If the continuous part is chosen to be a lognormal distribution with parameters μ and σ, then $\boldsymbol{\theta} = (p, \mu, \sigma)'$. If the continuous part is gamma, then $\boldsymbol{\theta} = (p, \alpha, \beta)'$, where α is the shape parameter and β the scale parameter.

For a distribution of the form (2) and for cell boundaries $0, x_1, \ldots, x_{k-1}, x_k$, where $x_k = \infty$, to cover the whole range of the random variable, we have the cell probabilities

$$\pi_1(\boldsymbol{\theta}) = 1 - p + pF(x_1),\ \pi_{i+1}(\boldsymbol{\theta}) = p[F(x_{i+1}) - F(x_i)],\ i = 1, \ldots, k-1.$$

Note that $F(x_k) = F(\infty) = 1$.

The MCE estimate obtained by minimizing (1) for the given cell probabilities is denoted by $\hat{\boldsymbol{\theta}}$. The estimate for mean rain rate is obtained by plugging the parameter estimates in the formula for the mean—that is, $\hat{E}(X) = \hat{p}e^{\hat{\mu}+\frac{1}{2}\hat{\sigma}^2}$ if the continuous part of the distribution is lognormal, and $\hat{E}(X) = \hat{p}\hat{\alpha}\hat{\beta}$ if the continuous part is gamma.

An important question is the price paid for grouping the data. As the distribution function underlying the data is continuous except for a jump at 0, considering only the cell frequencies n_i does not fully use the information available in the observations.

To get more insight into this problem, the asymptotic variance of the estimate for mean rain rate derived from grouped data is compared with the asymptotic variance obtained from the original, ungrouped observations.

The next section deals with the calculation of the asymptotic variance of mean rain rate and also with the loss of information that occurs by grouping.

3. Fisher information and estimation of the variance of mean rain rate

a. Fisher information for the mixed distribution

The Fisher information matrix is a measure of the value of a statistical experiment designed to investigate a parametric model and is also crucial in the study of asymptotic properties of maximum likelihood type estimators. Under regularity conditions its inverse is the asymptotic covariance matrix of the MLE distribution and it is defined elementwise by

$$I_{ij} = E\left[-\frac{\partial^2 \log f(x, \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right],$$

where $f(x,\boldsymbol{\theta})$ denotes the density function of the distribution in question and $\boldsymbol{\theta}$ is the parameter vector of interest. In the case of a single parameter, the information matrix reduces to a scalar value $I = E[-\partial^2/\partial\theta^2 \log f(x,\theta)]$. See Rao (1973, chapter 5), and DeGroot et al. (1983). In the rain-rate estimation problem $\boldsymbol{\theta} = (p, \mu, \sigma)'$—that is, the Fisher information matrix is a 3×3 matrix.

All the calculations presented in this section were performed using the lognormal distribution $\Lambda(x \mid \mu, \sigma)$ for the continuous part. The mean rain rate is therefore $E(X) = pe^{\mu+\frac{1}{2}\sigma^2}$.

For ungrouped observations, the Fisher information matrix for the mixed distribution is

$$\mathbf{I}_f = \begin{bmatrix} \dfrac{1}{p(1-p)} & 0 & 0 \\ 0 & \dfrac{p}{\sigma^2} & 0 \\ 0 & 0 & \dfrac{2p}{\sigma^2} \end{bmatrix}. \tag{3}$$

For the grouped data,

$$I_{ij} = \sum_{l=1}^{k} \frac{1}{\pi_l(\boldsymbol{\theta})} \frac{\partial\pi_l(\boldsymbol{\theta})}{\partial\theta_i} \frac{\partial\pi_l(\boldsymbol{\theta})}{\partial\theta_j}.$$

Thus, the Fisher information matrix for the grouped data is

$$\mathbf{I}_g = \begin{bmatrix} i_{11} & i_{12} & i_{13} \\ i_{12} & i_{22} & i_{23} \\ i_{13} & i_{23} & i_{33} \end{bmatrix},$$

where $F_i \equiv F(x_i)$,

$$\begin{aligned}
i_{11} &= \frac{1 - F_1}{p(1 - p + pF_1)}, \\
i_{12} &= \frac{\phi_1}{\sigma(1 - p + pF_1)}, \\
i_{13} &= \frac{\phi_1 u_1}{\sigma(1 - p + pF_1)}, \\
i_{22} &= \frac{p}{\sigma^2}\left(\sum_{j=2}^{k} \frac{(\phi_j - \phi_{j-1})^2}{F_j - F_{j-1}} + \frac{p\phi_1^2}{1 - p + pF_1}\right), \\
i_{23} &= \frac{p}{\sigma^2}\left(\sum_{j=2}^{k} \frac{(\phi_j - \phi_{j-1})(\phi_j u_j - \phi_{j-1}u_{j-1})}{F_j - F_{j-1}} + \frac{pu_1\phi_1^2}{1 - p + pF_1}\right), \\
i_{33} &= \frac{p}{\sigma^2}\left(\sum_{j=2}^{k} \frac{(\phi_j u_j - \phi_{j-1}u_{j-1})^2}{F_j - F_{j-1}} + \frac{pu_1^2\phi_1^2}{1 - p + pF_1}\right).
\end{aligned} \tag{4}$$

Here, ϕ_i denotes the density function of the standard normal distribution evaluated at $u_i = [\log(x_i) - \mu]/\sigma$, where $x_i = 0.5, 1.5, 2.5, 4, 6, 8, 10, 12, 16, 20$, and ∞. In the summation, therefore, $k = 11$; note the fact that $F_{11} = 1$ and $\phi_{11} = 0$.

For estimates $(p^*, \mu^*, \sigma^*)'$ derived from maximum likelihood with fully observed data, it follows that

$$\sqrt{n}[(p^*, \mu^*, \sigma^*)' - (p, \mu, \sigma)'] \to N(\mathbf{0}, \mathbf{I}_f^{-1}),$$

where $\mathbf{0} = (0, 0, 0)'$. Likewise, for estimates $(\hat{p}, \hat{\mu}, \hat{\sigma})'$ derived from maximum likelihood for grouped data,

$$\sqrt{n}[(\hat{p}, \hat{\mu}, \hat{\sigma})' - (p, \mu, \sigma)'] \to N(\mathbf{0}, \mathbf{I}_g^{-1}).$$

b. Estimation of the variance of mean area rain rate using the delta method

The mean rain rate is estimated through the parameters μ, σ, and p. The Fisher information matrix is therefore calculated for those three parameters and not for mean rain rate. To get the variability of the quantity of interest, apply a well-known statistical tool, the delta method, which is basically given by the following result (e.g., see Rao 1973).

Suppose $\mathbf{T}_n = (T_{n1}, \ldots, T_{nm})'$ is asymptotically multivariate normal with mean $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ and covariance matrix $\boldsymbol{\Sigma}/n$. Suppose the scalar function $g(t_1, \ldots, t_m)$ has a nonzero differential $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)'$ at $\boldsymbol{\theta}$, where

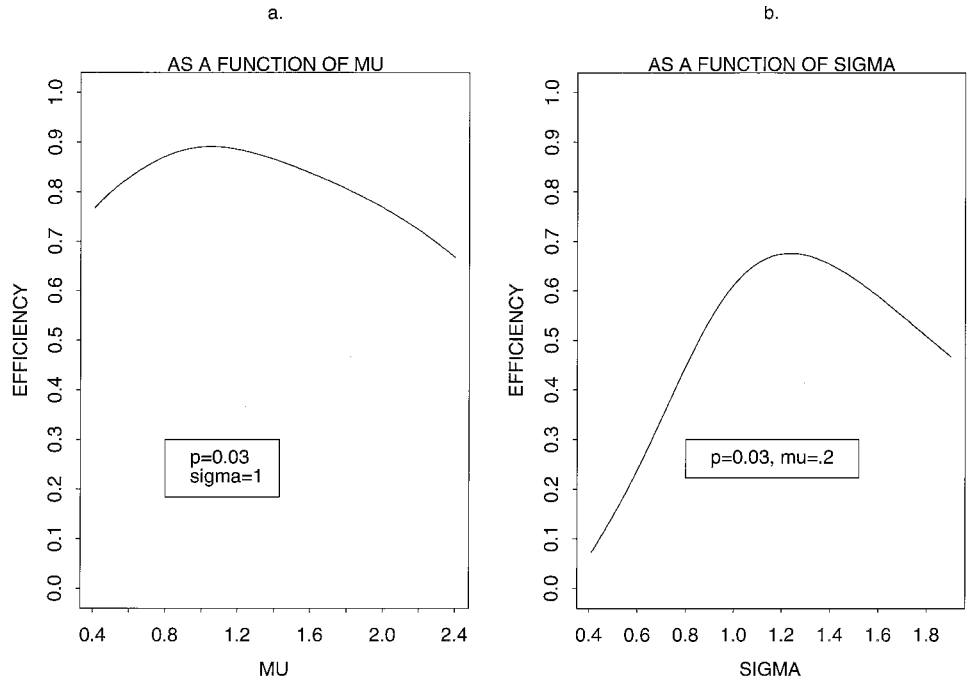

FIG. 1. Efficiency, defined as the ratio of the variance of maximum-likelihood estimators (for mixed lognormal distributions) from full data to that from grouped data, (a) versus μ, the mean value of the log of rain rate, for $p = 0.03$ (probability of rain) and $\sigma = 1$ (the variance of the log of rain rate), and (b) versus σ for $p = 0.03$ and $\mu = 0.2$.

$$\gamma_i = \left.\frac{\partial g}{\partial t_i}\right|_{t=\theta}.$$

Then, $\sqrt{n}[g(\mathbf{T}_n) - g(\boldsymbol{\theta})] \to N(\mathbf{0}, \boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma})$.

In our case, $\boldsymbol{\theta}' = (p, \mu, \sigma)$. Due to the asymptotic equivalency of the minimum chi square and the grouped data MLE, the minimum chi-square estimator $\hat{\boldsymbol{\theta}}$ is asymptotically normal, with the inverse of the Fisher information matrix as the covariance matrix.

For $g(\boldsymbol{\theta}) = E(X) = pe^{\mu+\sigma^2/2}$ the asymptotic variance for mean rain rate is obtained by applying the above result. It can easily be verified that the differentiability condition is satisfied for the parameter range of interest.

For the MLEs from ungrouped observations,

$$\boldsymbol{\Sigma} = \mathbf{I}_f^{-1} = \begin{bmatrix} p(1-p) & 0 & 0 \\ 0 & \dfrac{\sigma^2}{p} & 0 \\ 0 & 0 & \dfrac{\sigma^2}{2p} \end{bmatrix},$$

and the asymptotic variance becomes

$$\operatorname{var}[g(\hat{\boldsymbol{\theta}})] = \operatorname{var}[\hat{E}(X)] \doteq \frac{1}{n}e^{2\mu+\sigma^2}\left[p(1-p) + p\sigma^2 + \frac{1}{2}p\sigma^4\right].$$

The expression for the asymptotic variance of the grouped data estimate is more complicated, as $\mathbf{I}_g$ is no longer diagonal, and therefore the inverse cannot be given in a closed form. It is obtained numerically.

Figure 1 shows a plot of the efficiency—that is, the ratio of the full data variance over the grouped data variance—to get an idea of how much information and performance is lost due to grouping. For $p = 0.03$ and $\sigma = 1$, the maximal loss in efficiency for μ between 0.4 and 2.0 is around 30%. For μ close to 1.0, the efficiency is 90% and the loss in performance due to grouping is small. The effect of grouping is much more visible in the behavior of the variance of the estimators when σ is varied. For $p = 0.03$ and $\mu = 0.2$, efficiency can be as low as 10%. For σ around 1.2 though, the efficiency is 70%. Also, notice the strong sensitivity of the performance of our method with respect to σ when the estimation results for different data resolutions are compared (see next section). The parameter p, as can be seen from the tables, hardly influences the efficiency of the estimator.

Corresponding to Fig. 1, Tables 1 and 2 give some numerical values for the efficiency. In Table 1 it is looked at as a function of σ for fixed μ, and in Table 2 it is treated as a function of μ.

4. Analysis of the TOGA COARE dataset

a. The TOGA COARE data

In this section, part of the TOGA COARE database is analyzed. This database contains rainfall maps as

TABLE 1. Efficiency for the variance as a function of σ for $\mu = 0.2$.

$p = 0.03$											
σ	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
Efficiency	0.147	0.239	0.342	0.447	0.549	0.611	0.655	0.674	0.670	0.653	0.624
$p = 0.5$											
σ	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
Efficiency	0.097	0.172	0.267	0.374	0.477	0.560	0.615	0.641	0.643	0.627	0.604

measured by two shipboard Doppler radars. The rainfall maps are binary images composed of a 556-byte header followed by a 278 byte $\times$ 278 byte array of data with a spatial resolution of a 2 km $\times$ 2 km pixel.

The data array is a 5° latitude by 5° longitude area (556 km $\times$ 556 km) within the intensive flux array. The origin is the northwestern corner of a world map, with coordinates 0°, 153°E. Rainfall-rate maps exist for data collected between 10 November 1992 and 23 February 1993. The time between two snapshots is 10 min. In our analysis, the data sequences cover the period 1–18 January 1993. For more information concerning the data, see Short et al. (1995).

b. Time–space sampling designs

For the chi-square estimation of the parameters (p, μ, σ) of the mixed lognormal distribution, different time–space sampling designs characterized by the triples (k, l, l) are used. The first index k denotes sampling frequency in time and is a multiple of 15 min; l refers to spatial sampling and is a multiple of 4 km; missing data points were not taken into account. These designs are chosen to conform with the designs used for the GATE I dataset. For example, the sampling design (4, 8, 8) therefore corresponds to pixels sampled every hour in time and separated by 32-km intervals. The total number of sampled pixels for any specific design is denoted by n.

NAG-FORTRAN library subroutines were used for the minimization of the chi-square statistic.

c. Results from different designs, 4 km $\times$ 4 km resolution

The main part of the TOGA analysis was carried out for a pixel size of 4 km $\times$ 4 km. This resolution of the data was obtained by simply averaging four neighboring pixels, if none of them was a missing data point. The results are shown in the tables below.

The results obtained by analyzing the data in the original 2 km $\times$ 2 km resolution show that lognormality was not the best choice of a model, as the variability of the estimates was high, and the estimate for mean area average rain rate seemed to be too large. The fit to the 4 km $\times$ 4 km data was much better. This is due to a reduced variability in the data, as high and low rain rates get averaged out. For example, for the snapshot taken at 0000 1 January 1993, the maximal rain rate observed for the 2 km $\times$ 2 km resolution is 64.5 mm h^{-1}; for the 4 km $\times$ 4 km resolution, the maximal rain rate is 27.8 mm h^{-1}. Table 3 shows that the results for a coarser resolution are quite uniform. The categories used here are [0, 0.5), [0.5, 1.5), [1.5, 2.5), [2.5, 4), [4, 6), [6, 8), [8, 10), [10, 12), [12, 16), [16, 20), and [20, ∞). Here, 0.5 mm h^{-1} was chosen, as it is within the dynamic range of the TOGA COARE radars.

From Table 3 we see that the mean rain-rate estimates range from 0.2238 to 0.3228 mm h^{-1} and the values of the chi-square statistic from 5.108 to 32.584. The largest values of the chi-square statistic correspond to close spatial sampling designs, which indicates dependence in the data. The same phenomenon was observed in Kedem et al. (1990)—namely, close designs resulted in large chi-square values. The estimates for σ vary around 1.9, and the values for μ range from -0.0288 to 0.1016. The method of moments gave somewhat lower estimates for μ and σ.

Table 4 contains the estimates of the variability of mean rain rate for the TOGA COARE data. It shows that the effect of grouping the data on the variance is small. Once the variance of the estimates is obtained, confidence intervals for mean rain rate can be computed. The average 95% confidence interval for ungrouped data with a resolution of 4 km $\times$ 4 km is (0.2178 mm h^{-1}, 0.3354 mm h^{-1}), and the width of the interval is 0.1176; the average 95% confidence interval for grouped data is (0.1979 mm h^{-1}, 0.3554 mm h^{-1}), with an average width of 0.1574, as expected due to loss in efficiency.

TABLE 2. Efficiency for the variance as a function of μ for $\sigma = 1.0$.

$p = 0.03$											
μ	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Efficiency	0.890	0.890	0.884	0.875	0.864	0.852	0.837	0.836	0.820	0.803	0.785
$p = 0.5$											
μ	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Efficiency	0.868	0.868	0.862	0.852	0.838	0.823	0.806	0.789	0.770	0.749	0.723

TABLE 3. Chi-square estimation for mixed lognormal distribution with 4 km × 4 km resolution.

Design	n	$\hat{p}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{E}(X \mid X > 0)$ (mm h^{-1})	$\hat{E}(X)$ (mm h^{-1})	Chi-square
2,4,4	263 895	0.0473	−0.0199	1.954	6.618	0.3133	32.584
2,8,8	66 825	0.0468	−0.0357	1.982	6.888	0.3228	7.536
2,16,16	16 048	0.0425	−0.0213	1.965	6.754	0.2871	5.108
4,4,4	132 803	0.0468	−0.0050	1.895	5.996	0.2811	17.885
4,8,8	33 728	0.0455	0.0085	1.899	6.117	0.2788	8.851
4,16,16	8014	0.0432	−0.0071	1.973	6.965	0.3012	6.758
8,4,4	65 369	0.0457	−0.0288	1.925	6.198	0.2834	16.922
8,8,8	16 556	0.0437	−0.0331	1.817	5.044	0.2238	9.448
16,4,4	32 420	0.0386	0.1016	1.823	5.835	0.2252	7.756
16,8,8	8009	0.0419	0.0047	1.987	7.238	0.3040	13.660

5. Comparison with the method used by Meneghini and Jones

Meneghini and Jones (1993) estimate the parameters of the mixed distribution by minimizing

$$\Sigma_{i=1}^{k} \{F_n[R_T(i)] - F[R_T(i), \boldsymbol{\theta}]\}^2 \tag{5}$$

over the dynamic range $[R_T(1), R_T(k)]$ of the instrument, where $R_T(i)$, $i = 2, \ldots, k-1$ denote thresholds within the range and F_n is the empirical distribution function based on the n observed data points. This method relates to the minimum chi-square estimation in the following way. It is possible to rewrite (5) using the following notation. Choose the class boundaries for the minimum chi-square statistic to be $R_T(1), \ldots, R_T(k)$, then

$$F[R_T(j), \boldsymbol{\theta}] = \sum_{i=1}^{j} \pi_i(\boldsymbol{\theta}),$$

$$F_n[R_T(j)] = \frac{1}{n} \sum_{i=1}^{j} n_i,$$

where n_i denotes the observed frequency in the ith category. Therefore,

$$\sum_{i=1}^{k} \{F_n[R_T(i)] - F[R_T(i), \boldsymbol{\theta}]\}^2 = \frac{1}{n^2} \sum_{i=1}^{k} \left\{ \sum_{j=1}^{i} [n_j - n\pi_j(\boldsymbol{\theta})] \right\}^2.$$

To minimize the above expression is equivalent to minimizing the Cramer–von Mises statistic $W_d^2 = n \Sigma_{i=1}^{k} \{\Sigma_{j=1}^{i} [n_j - n\pi_j(\boldsymbol{\theta})]\}^2$ for discrete data on the finite instrument range.

Table 5 compares the estimates from minimum chi-square estimation and least squares estimation for a mixed lognormal distribution for several designs of the GATE I dataset using grouped data over the whole data range—that is, $R_T(1) = 0$ and $R_T(k) = \infty$ for $k = 10$, and the boundaries $x_i = 1, 2, 4, 6, 8, 10, 12, 16$, and 20. The first entry in each column is the minimum chi-square estimate, and the second entry is the least squares estimate. The two methods give very similar results.

Note that since for random samples the covariance matrix of $\{F_n[R_T(1)], \ldots, F_n[R_T(k)]\}$ is known, the variability of the least squares estimates can be obtained using nonlinear least squares theory.

It is useful to mention that another way to calculate the variance of mean rain rate is through a reparameterization of the density function in terms of p, $M = E(X)$, and $\beta^2 = \text{var}(X \mid X > 0)$. By calculating the Fisher information matrix for the parameterization $\boldsymbol{\theta} = (p, M, \beta^2)$, the asymptotic variance for mean rain rate is obtained directly as the inverse of that matrix. This obviates the need for the delta method.

Table 6 shows the exact variabilities for the GATE I data and the corresponding 95% confidence intervals, calculated using the reparameterization, the variances, and confidence intervals obtained by delta method for

TABLE 4. Estimates for mean rain rate, the sample variance and an approximate 95% confidence interval for mean rain rate based on ungrouped data with 4 km × 4 km resolution.

Design	n	$\hat{E}(X)$ grouped (mm h^{-1})	$[\hat{\text{var}}(X)]^{1/2}$ ungrouped (mm h^{-1})	95% CI ungrouped (mm h^{-1})	$[\hat{\text{var}}(X)]^{1/2}$ grouped (mm h^{-1})	95% CI grouped (mm h^{-1})
2,8,8	66 825	0.313	0.02043	0.2810, 0.3627	0.02784	(0.2662, 0.3775)
4,4,4	132 803	0.281	0.01179	0.256, 0.304	0.01560	(0.2492, 0.3116)
4,8,8	33 728	0.278	0.02364	0.2311, 0.3257	0.03132	(0.2158, 0.3411)
4,16,16	8014	0.301	0.05700	0.1868, 0.4150	0.07752	(0.1459, 0.4560)
8,4,4	65 369	0.283	0.01757	0.2480, 0.3183	0.02350	(0.2361, 0.3302)
8,8,8	16 556	0.223	0.02551	0.1692, 0.2713	0.03287	(0.1545, 0.2860)
16,4,4	32 420	0.225	0.01992	0.1853, 0.2651	0.02573	(0.1737, 0.2766)
16,8,8	8009	0.304	0.05899	0.1852, 0.4213	0.08057	(0.1421, 0.4644)

TABLE 5. Minimum chi-square estimation versus least squares for mixed lognormal distribution for GATE I. The least squares estimates are given in the second column.

Design	n	$\hat{p}$		$\hat{\mu}$		$\hat{\sigma}$		$\hat{E}(X)$	
30,10,10	4814	0.085	0.085	0.999	0.968	1.156	1.102	0.449	0.441
20,10,10	7138	0.075	0.075	1.095	1.089	1.071	1.032	0.399	0.378
6,6,6	65 208	0.081	0.083	1.151	1.111	1.056	1.067	0.447	0.444
10,8,8	21 844	0.083	0.087	1.161	1.084	1.059	1.094	0.466	0.468
2,4,4	434 148	0.083	0.085	1.137	1.084	1.089	1.089	0.467	0.456
24,1,1	585 216	0.083	0.086	1.158	1.092	1.056	1.090	0.462	0.465
48,1,1	292 608	0.088	0.095	1.255	1.105	1.019	1.130	0.519	0.545
4,4,4	217 074	0.083	0.085	1.129	1.080	1.098	1.089	0.469	0.454
8,8,8	27 305	0.083	0.085	1.129	1.104	1.047	1.057	0.449	0.446
10,10,10	14 276	0.081	0.082	1.085	1.065	1.176	1.093	0.479	0.431
5,10,10	28 552	0.083	0.083	1.058	1.063	1.208	1.091	0.498	0.435
5,20,20	6880	0.077	0.079	1.184	1.129	1.067	1.085	0.443	0.440

grouped data. As can be seen, the resulting estimates are close.

Figures 2 and 3 show efficiency plots as functions of μ and σ using the reparameterization and an asymptotic formula (expression 9.58) from Aitchison and Brown (1963) to calculate the variances involved. It has to be mentioned that the Atchison and Brown approximation is only valid for large values of p and is therefore not applicable in the TOGA COARE setting, where p is around 0.04.

Table 7 contains the estimates for the parameters of the mixed lognormal distribution and mean rain rate for TOGA COARE obtained from least squares estimation. The procedure was applied to the whole data range and the same categories that were used in section 4 for some of the sampling designs. The residual sum of squares (RSS) serves as a measure of the goodness of fit in the same way the chi-square value did. As expected, the resulting parameter estimates are close to the MCS estimates, and the estimates for mean rain rate are within the confidence intervals of the MCS mean rain-rate estimates.

6. Summary

It has been shown how to obtain the asymptotic variance of the minimum chi-square estimator of the mean of a mixed lognormal distribution using categorized (grouped) data, where the first category includes the purely zero values. From the asymptotic variance, it is seen that for a certain range of the parameters the loss of efficiency of the MCS estimator is small, and for other parameter ranges the loss is appreciable and must be compensated by larger samples. A comparison between the minimum chi-square method and the least squares method by Meneghini and Jones (1993) shows that both methods give similar results.

From the application of the method to the TOGA COARE dataset it is seen that for a resolution of 4 km $\times$ 4 km a mixed lognormal model seems quite appropriate.

After obtaining the variance of the estimate, confidence intervals for mean rain rate can be computed.

Here, $E(X) = pe^{\mu+\frac{1}{2}\sigma^2}$ is equivalent to the space–time mean rain rate

$$\frac{1}{TA}\int_T\int_A r(x, t)\, dx\, dt,$$

where $r(x, t)$ is the space–time rain rate and the integration is over the particular space–time box of interest; T denotes the time interval and A the area in question. Hence, another estimator of mean rain rate is the linear estimator

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i(x, t),$$

where n is the total number of observations. This es-

TABLE 6. Variance estimates for GATE I.

Design	$\hat{E}(X)$ (mm h^{-1})	$[\hat{\text{var}}(X)]^{1/2}$ (mm h^{-1}) exact	95% CI exact	$[\hat{\text{var}}(X)]^{1/2}$ (mm h^{-1}) delta method	95% CI delta method
30,10,10	0.4483	0.04382	(0.3624, 0.5342)	0.04347	(0.3707, 0.5450)
20,10,10	0.3992	0.03069	(0.3391, 0.4593)	0.03051	(0.3382, 0.4602)
6,6,6	0.4465	0.01072	(0.4254, 0.4675)	0.01072	(0.4250, 0.4679)
10,8,8	0.4661	0.01918	(0.4284, 0.5036)	0.01913	(0.4277, 0.5043)
4,4,4	0.4685	0.00641	(0.4559, 0.4802)	0.00632	(0.4558, 0.4812)
8,8,8	0.4483	0.01629	(0.4163, 0.4802)	0.01627	(0.4158, 0.4809)
10,10,10	0.4788	0.02851	(0.4229, 0.5347)	0.02791	(0.4229, 0.5346)
5,10,10	0.4971	0.03420	(0.4549, 0.5393)	0.02096	(0.4552, 0.5390)
5,20,20	0.4429	0.03420	(0.3758, 0.5099)	0.03408	(0.3747, 0.5111)

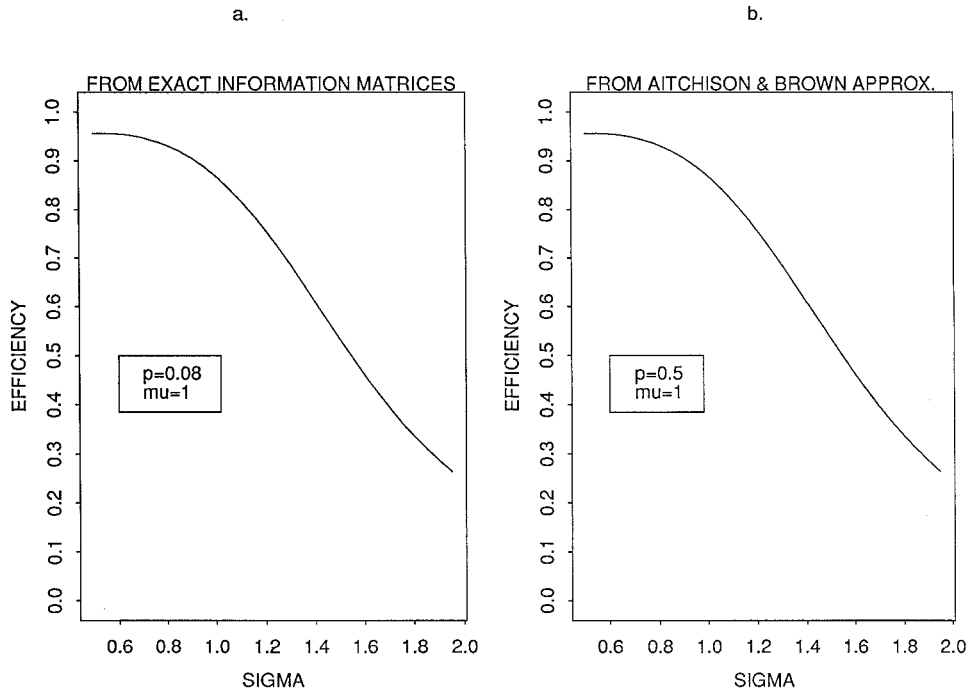

FIG. 2. Efficiency versus σ as determined from (a) exact information matrices with $p = 0.08$ and $\mu = 1$, and from (b) the approximation of Aitchison and Brown (1963), with $p = 0.5$ and $\mu = 1$.

timator can be used when the data are obtained from reliable instruments that can distinguish between zero and nonzero rain rate, and that do not suffer from limited dynamic range and/or bias problems. Our estimator is an approximation to the maximum likelihood—an efficient estimator—from grouped data, a device that mitigates these problems.

Finally, fitting a model pdf is useful for incorporating values in the tail of the distribution—large values that cannot be observed due to instrument-limited dynamic range.

Acknowledgments. This work was supported by National Aeronautics and Space Administration Grant

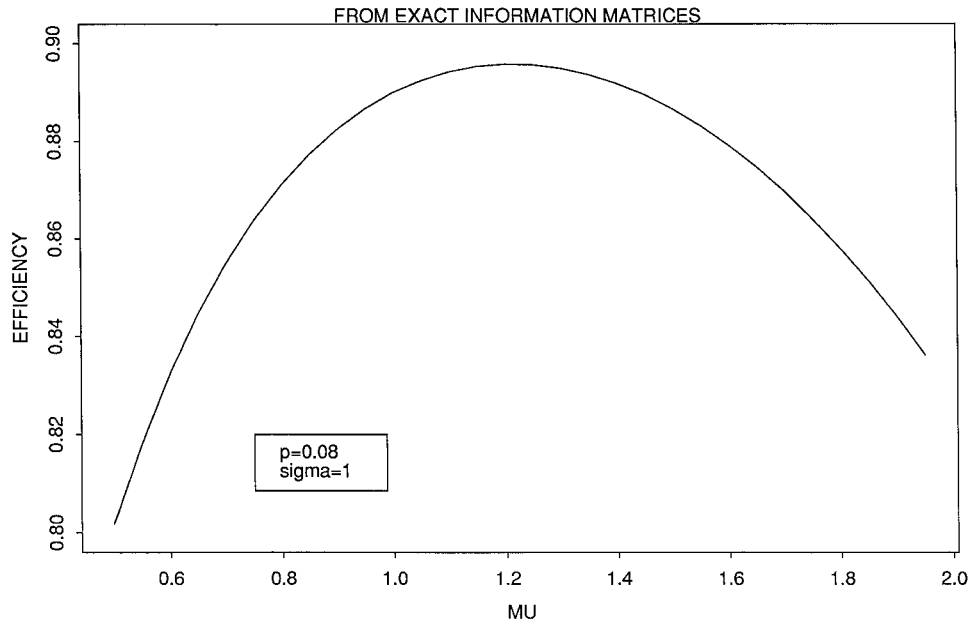

FIG. 3. Efficiency versus μ from exact information matrices for $p = 0.08$ and $\sigma = 1$.

TABLE 7. Least squares estimation for mixed lognormal distribution with 4 km × 4 km resolution.

Design	n	$\hat{p}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{E}(X\|X>0)$ (mm h^{-1})	$\hat{E}(X)$ (mm h^{-1})	RSS
4,8,8	33 728	0.0455	−0.0038	1.9054	6.119	0.2789	8.8D-8
4,16,16	8014	0.0429	−0.0281	1.9622	6.665	0.2860	5.9D-7
8,8,8	16 556	0.0446	−0.0801	1.883	5.442	0.2429	5.1D-7
16,4,4	32 420	0.0389	0.0575	1.915	6.627	0.2582	2.4D-7
16,8,8	8009	0.0418	−0.0740	1.962	6.374	0.2670	7.4D-7

NAG52783 and by National Science Foundation Grant EEC-940-2384. The authors thank the referees for useful comments.

REFERENCES

Aitchison, J., and J. A. C. Brown, 1963: *The Lognormal Distribution.* Cambridge University Press, 176 pp.

DeGroot, M. H., M. J. Bayarri, and P. K. Goel, 1983: Truncation, information and the coefficient of variation. Contributions to probability and statistics. *Essays in Honor of Ingram Olkin,* L. Gleser, M. Perlman, S. Press, and A. Sampson, Eds., Springer Verlag, 412–428.

Hong, Y., 1994: Retrieval of monthly rainfall over oceans from the Special Sensor Microwave/Imager (SSM/I). Ph.D. dissertation, Texas A&M University, 130 pp.

Kedem, B., L. S. Chiu, and G. R. North, 1990: Estimation of mean rain rate: Application to satellite observations. *J. Geophys. Res.*, **95,** 1965–1972.

Meneghini, R., and J. A. Jones, 1993: An approach to estimate the areal rain-rate distribution from spaceborne radar by the use of multiple thresholds. *J. Appl. Meteor.*, **32,** 386–398.

Rao, C. R., 1973: *Linear Statistical Inference and Its Applications.* Wiley and Sons, 625 pp.

Short, D. A., P. A. Kucera, B. S. Ferrier, and O. W. Thiele, 1995: COARE IOP rainfall from shipborne radars: 1. Rain mapping algorithms. Preprints, *Proc. 27th Conf. on Radar Meteorology,* Vail, CO, Amer. Meteor. Soc., 678–680.