A Comparison of Gamma and Lognormal Distributions for Characterizing Satellite Rain Rates from the Tropical Rainfall Measuring Mission

HYE-KYUNG CHO, KENNETH P. BOWMAN, AND GERALD R. NORTH

Department of Atmospheric Sciences, Texas A&M University, College Station, Texas

(Manuscript received 4 February 2003, in final form 6 June 2004)

ABSTRACT

This study investigates the spatial characteristics of nonzero rain rates to develop a probability density function (PDF) model of precipitation using rainfall data from the Tropical Rainfall Measuring Mission (TRMM) satellite. The minimum χ^2 method is used to find a good estimator for the rain-rate distribution between the gamma and lognormal distributions, which are popularly used in the simulation of the rain-rate PDF. Results are sensitive to the choice of dynamic range, but both the gamma and lognormal distributions match well with the PDF of rainfall data. Comparison with sample means shows that the parametric mean from the lognormal distribution overestimates the sample mean, whereas the gamma distribution underestimates it. These differences are caused by the inflated tail in the lognormal distribution and the small shape parameter in the gamma distribution. If shape constraint is given, the difference between the sample mean and the parametric mean from the fitted gamma distribution decreases significantly, although the resulting χ^2 values slightly increase. Of interest is that a consistent regional preference between two test functions is found. The gamma fits outperform the lognormal fits in wet regions, whereas the lognormal fits are better than the gamma fits for dry regions. Results can be improved with a specific model assumption depending on mean rain rates, but the results presented in this study can be easily applied to develop the rainfall retrieval algorithm and to find the proper statistics in the rainfall data.

1. Introduction

a. Background

Tropical rainfall is an extremely important climate variable because three-fourths of the energy that drives the atmospheric wind circulation comes from the latent heat released by tropical precipitation (Kummerow et al. 2000). Although tropical rain plays an essential role in the global hydrological cycle and the atmospheric energy budget, it is one of the most difficult meteorological parameters to study because of the lack of reliable data and the large variations of rainfall in space and time. The ocean covers 75% of the Tropics. Much of the land area of the Tropics is difficult terrain, such as rainforests and deserts, in which to install rainfall measurement instruments. The only way to provide global-scale rainfall data over the Tropics is by means of spaceborne satellite sensors (North 1987). The Tropical Rainfall Measuring Mission (TRMM) satellite is designed to measure tropical and subtropical rainfall by using microwave, visible, and infrared sensors. TRMM carries the first spaceborne precipitation radar, and its orbital inclination allows TRMM to observe each position on earth's surface between 35°N and 35°S. The orbit precesses with respect to the diurnal cycle with a period of about 6 weeks. The TRMM satellite provides a consistent database of instantaneous rain-rate measurements throughout the Tropics, and rainfall information has been successfully collected since its launch on 27 November 1997 (Kummerow et al. 2000).

To improve modeling of rainfall processes and their influence on the global circulation, and ultimately to predict rainfall, many researchers have been searching for the physical and statistical properties of rainfall using observational data. The statistical characterization of rain is useful in understanding the large-scale space and time variability of the process and is helpful in assessing the accuracy of the rainfall retrievals by imposing constraints that must be satisfied by the spatial and temporal averages of the high-resolution estimates, from the perspective of spaceborne sensors (Meneghini et al. 2001). One topic of interest is the probability density function (PDF) of rain rates. From a meteorological and climatological point of view, the PDF of rain rate is indispensable and particularly so when it comes to the estimation of total rainfall from space by satellite (Kedem et al. 1994). The parameters of the particular distribution are used for representing very different climatological conditions (Wilks 1995) because

Corresponding author address: Dr. Hye-Kyung Cho, Department of Atmospheric Sciences, Texas A&M University, College Station, TX 77845-3150.
E-mail: hkc@ariel.met.tamu.edu

the PDF of rain rate over large areas is fairly stable over time and space (Meneghini and Jones 1993). Operationally, the studies of the PDF of rain rate are used for improving the present forecast format of the precipitation (Briggs and Wilks 1996) and calculating the various moments or other statistics from the dataset (Martin 1989). In previous papers, various functional forms have been proposed to represent the probability distribution for rainfall, but general agreement has not been reached on which best represents the PDF for rain rates (Wilks 1990; Kedem et al. 1990). This may result because data are collected under different physical conditions, and the shape of rain-rate PDFs is complicated and sensitive to the model assumptions. Considering these present situations, the TRMM is a great challenge to the field of the rainfall statistics because it provides the long-term, spatially extensive dataset for the tropical rainfall, and its purpose is also the indirect estimation of mean rain rate over a large area throughout a certain period of time (Kedem et al. 1997).

In this study, the properties of tropical precipitation are investigated by characterizing the PDF of rain rates using the TRMM dataset. With a minimum of model assumptions, we fit observed TRMM rain rates at each location with the gamma and lognormal distributions and compare the relative performance between two test functions in representing the PDF of rain rates. The sensitivity to model assumptions and the spatial characteristics of the rain-rate histograms are also investigated.

These results will be helpful for improving the present TRMM rainfall retrieval algorithm and the estimation procedure for mean rain rate and other statistical variables. Clearly, there is a great need for better knowledge of the variability of rainfall statistics in space and time and of how they depend on scale (Bell et al. 1990). The results of this study can also be applied to find the spatial dependency structure of rain rates and the general circulation model (GCM) simulations of tropical rainfall.

b. Rain-rate distribution

The extremely variable nature of rain makes it difficult to estimate area or time averages and higher moments of the rainfall amounts directly from the observational data (Martin 1989). Rain-rate PDFs can be fitted by many theoretical distributions, and the mean and variance can be calculated from the parameters of such empirical PDFs. One set of parameters from certain hypothetical distributions, however, can give a satisfying PDF shape at one location but not at others. In addition, the combination of small sample sizes and the inherent variability of precipitation can lead to large uncertainties in the estimated PDFs. Thus, finding generalized PDFs is difficult (Revfeim 1982).

Because of a strong peak at $r = 0$ (i.e., no-rain case), the PDFs of rain rates are easily considered as mixed distributions; that is, a combination of a discrete function at $r = 0$ and a continuous component for $r > 0$ (Kedem et al. 1990). To represent the cumulative PDF for all rain rates (including $r = 0$), the discontinuous part of the cumulative PDF is given by a step function at $r = 0$ of strength $(1 - p)$, in which p is the ratio of rainy data to total data. The continuous portion of the PDF is to be estimated from the data and in principle can be assumed to be any arbitrary theoretical distribution (Kedem et al. 1990).

Experience shows that the PDFs for $r > 0$ are highly asymmetrical and skewed toward larger rain rates. Therefore, a Gaussian PDF is not useful in this case. There are many PDFs that are bounded on the left by zero and positively skewed. Among these distributions, the gamma distribution is widely used to model rain rates. The gamma distribution is defined by

$$f(x)dx = \frac{x^{\alpha-1}\exp(-x/\beta)}{\Gamma(\alpha)\beta^{\alpha}}dx, \qquad x,\ \alpha,\ \beta > 0, \quad (1)$$

where α and β are the shape and scale parameters, and Γ is the usual gamma function. The mean and variance of this distribution are given by $\alpha\beta$ and $\alpha\beta^2$, respectively. By changing α and β, the gamma distribution can represent many shapes (Fig. 1a). The versatile shape of the gamma distribution makes it an attractive candidate for representing precipitation data. It is more difficult to work with than the Gaussian (or lognormal) distribution, however, because the two parameters of the gamma distribution do not correspond exactly to the moments of sample data, as is the case of the Gaussian distribution (Wilks 1995).

By using long-term rain-rate data for selected locations, Ison et al. (1971) and Swift and Schreuder (1981) concluded that daily precipitation amounts follow the gamma distribution. Wilks and Eggleston (1992) characterized the precipitation climatology by using the shape parameters of gamma distributions. The gamma distribution is also used in reporting monthly and seasonal precipitation anomalies in the *Journal of Climate* seasonal climate summary of Wilks and Eggleston (1992).

Another distribution that is commonly used to represent rain rates is the lognormal distribution. The lognormal distribution is similar in appearance to the gamma distribution (Fig. 1b). The lognormal distribution assumes that the logarithms of the data are normally distributed. The lognormal distribution is given by

$$f(x)dx = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{[\ln(x/x_0)]^2}{2\sigma^2}\right\}dx,$$
$$x_0 = e^{\mu}, \qquad x > 0, \quad (2)$$

where μ and σ are the mean and standard deviation of the logarithmically transformed variables, respectively. Modeling with the lognormal distribution allows the use of normal-theory statistics on a logarithmic scale, and

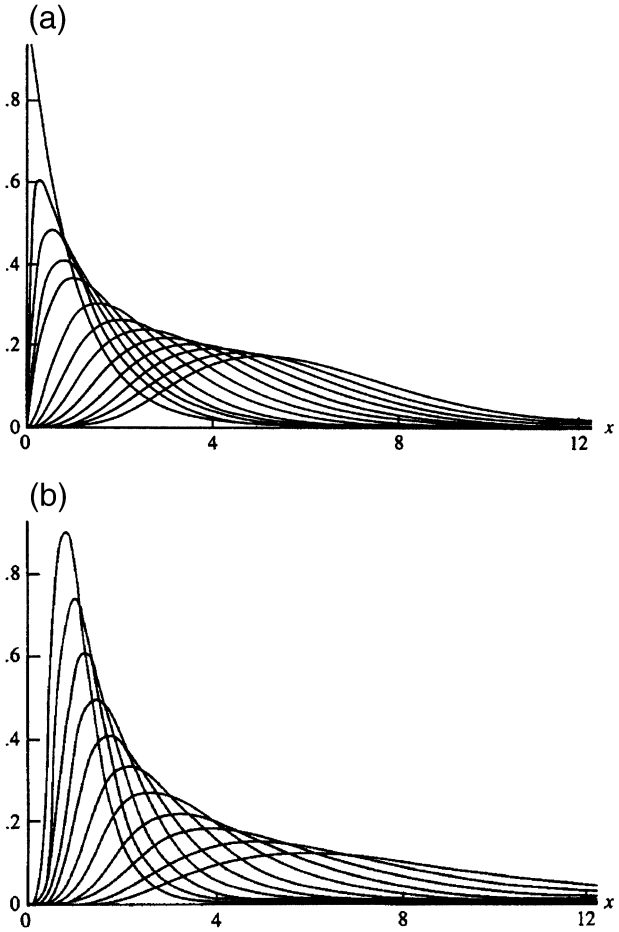

FIG. 1. (a) Example gamma distributions and (b) example lognormal distributions. The gamma and lognormal distributions have similar shapes. These distributions are useful when the variable of interest is skewed to the right.

parameter fitting is simple and straightforward (Wilks 1995).

Using the Global Atmospheric Research Program (GARP) Tropical Atlantic Experiment (GATE) data, Kedem et al. (1990, 1994) fitted the rain rates with a lognormal distribution and calculated mean rain rates based on the parameters of the distribution. In practice, the TRMM precipitation radar (PR) algorithm 3A26 estimates space–time rain-rate statistics by assuming a lognormal distribution for the rain rates and minimizing the rms difference between the hypothetical distribution and a discrete approximation to the distribution obtained directly from the measurements (TSDIS/TRMM 1999). Much of the observational justification for the lognormal distribution is based upon radar estimates of rainfall calculated using a radar reflectivity factor. Jameson and Kostinski (1999) found, however, that the distribution of rain rates derived using direct rain gauge and disdrometer observations in the Tropics shows clear departures from the lognormal distribution.

Many previous studies have tried to find which of the two distributions works better to represent the variation of the precipitation (Swift and Schreuder 1981; Meneghini and Jones 1993; Kedem et al. 1994; Jameson and Kostinski 1999). General agreement, however, is not reached because these results are not free of model assumptions and data selection. Moreover, to construct a clean statistical test between two test functions, enormous amounts of data are needed to assess the behavior of the tail of the PDFs (Bell 1987). Thus, it is a great challenge to investigate the ability of two distributions and compare their relative performances in fitting observational precipitation data using TRMM data.

2. Data and methods

a. TRMM 3G68 combined data

In this study, the TRMM 3G68 combined rain-rate product is used. The data were obtained from the TRMM Science Data and Information System (TSDIS). The TRMM ''combined'' algorithm merges information from the two TRMM sensors, the PR, and the 10-GHz channel of the TRMM microwave imager (TMI) into a single retrieval to produce a ''best'' rain estimate for TRMM (Haddad et al. 1997). The TRMM 3G68 combined product uses the high-quality retrievals done for the narrow swath in PR data and uses the wide swath retrievals generated in TMI.

The 3G68 data are an hourly gridded product. The dataset includes 24 hourly grids in a single daily file. The 3G68 data are spatially averaged in each $0.5^\circ \times 0.5^\circ$ latitude–longitude box that has a TRMM overpass, recording the time of observation of the first pixel in that grid box, the average rainfall, and the necessary statistics from the instantaneous observations within the grid box. A single hourly observation contains only one TRMM overpass (TSDIS/TRMM 1999). With this fine spatial resolution, each grid box contains between zero and five area-averaged observations per day depending on location. Approximately 11% of grid boxes have no data on a given day.

b. Averaging data

To study the effects of spatial averaging in the data, area-averaged rain rates with different resolutions, such as $5^\circ \times 5^\circ$ from the $0.5^\circ \times 0.5^\circ$ dataset are calculated. For each day, all available data within an averaging box are used to compute the daily mean. Daily area-averaged rain rate values are calculated from

$$r = \sum_{i=1}^{N} W_i R_i \quad \text{and} \tag{3}$$

$$W_i = \frac{A_i}{\sum_{k=1}^{N} A_k}, \tag{4}$$

where N is the number of $0.5^\circ \times 0.5^\circ$ boxes, included

in the averaging box and R_i is the ''instantaneous'' rain rate. No-rain data (i.e., $R_i = 0$) are included to avoid overestimating mean rain rates. Here, W_i is the area-weighting function for R_i, and A_i is the area of the ith $0.5° \times 0.5°$ grid box. The alternative way to get area-averaged rain rates is to use the total pixel counts instead of A_i in Eq. (4), but the results from this method are not much different from those used in this study. One of the purposes of this study is to find the nature of the change of rainfall statistics as the averaging area increases, namely, whether the statistical characteristics within a $0.5° \times 0.5°$ grid box would be retained in an averaged grid box. Rainfall, however, exhibits extensive variability on a wide range of spatial and temporal scales, and the data correlation in space and time is unknown; it is not a simple question to answer. Thus, in section 3d, the fitted distributions at the different grid resolutions are compared. Data from 8 December 1997 to 30 June 2001 (1301 days) were used. Rain rates are given in millimeters per hour.

c. Estimation methods

Three standard techniques for fitting a theoretical PDF to data are the method of moments, the maximum likelihood method, and the minimum χ^2 method. The method of moments uses the sample mean and variance as the mean and variance of the fitted distribution. Although it estimates the parameters of the distribution directly, it is inefficient, especially when a histogram is highly skewed (Wilks 1995).

When the form of the distribution itself can be hypothesized, asymptotically efficient estimators may be obtained by the maximum likelihood method. The maximum likelihood method seeks to find values of the distribution parameters that maximize the likelihood function, which is a function of the unknown parameters for fixed values of the already observed data (Wilks 1995). If data are not independent, however, and the correlation is not well known, the maximum likelihood method needs modification to obtain reliable estimates (Casella and Berger 1990; Kedem et al. 1990). Because of its computational efficiency, approximate solutions of the maximum likelihood equation have been developed (Öztürk 1981; Wilks 1995).

The minimum χ^2 method estimates the parameters of the fitted distribution by minimizing the difference between the hypothesized and observed distributions. If the data are grouped into k categories ($i = 1, 2, 3, \ldots, k$), the observed frequency in each class is denoted as O, and the expected probability from the hypothesized distribution is π, then the χ^2 value can be calculated from

$$\chi^2(\theta) = \sum_{i=1}^{k} \frac{[O_i - n\pi_i(\theta)]^2}{n\pi_i(\theta)} \quad \text{and} \tag{5}$$

$$n = \sum_{i=1}^{k} O_i. \tag{6}$$

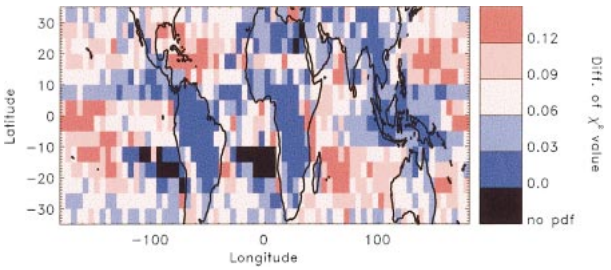

FIG. 2. Difference of the χ^2 values between the method of moments and the minimum χ^2 method. A positive value means a better fit by the minimum χ^2 method when compared with the method of moments.

The parameters (θ) of the hypothetical distribution π are adjusted so as to minimize χ^2. The χ^2 value can be evaluated regardless of dependence between data and gives information about the goodness of fit. Moreover, the procedure is simple because it merely finds the χ^2 value closest to zero (Kedem et al. 1990).

The χ^2 values can also be used to compare results from different PDFs. When one judges the relative performance among the differently fitted distributions at a given location, the distribution with a smaller χ^2 value is directly closer to the observational distribution. However, if one wants to compare the characteristics between different locations, the χ^2 value does not give an objective criterion because the χ^2 value itself is dependent on the sample size (n) at each location. For example, the distribution fitted in the dry region usually has a smaller χ^2 value than that in the wet region, even though its performance may be poorer. Thus, the χ^2 value is not useful in the comparison of the performances between the different locations. To remedy this problem, an alternative form of the χ^2 value, the ϕ value, is calculated from

$$\phi = \sum_{i=1}^{k} \frac{[O_i/n - \pi_i(\theta)]^2}{\pi_i(\theta)} = \frac{\chi^2(\theta)}{n}. \tag{7}$$

Although the overall spatial distribution of the ϕ values is not much different from that of the χ^2 values, the ϕ values are more useful for comparing results from different locations than is the case for χ^2, alone. Even if a quantitative method for evaluating the closeness of the fitted distribution to the data (such as χ^2) may indicate an inadequate fit, it will not indicate which part of the histogram fails to match the hypothesized distribution. Therefore, we also check the fit by superposing the fitted theoretical distribution on the data at each location. Graphical comparisons of the data histograms and the fitted distribution allow diagnosis of where and how the theoretical representations are inadequate (Wilks 1995).

To show the efficiency of the minimum χ^2 method, Fig. 2 illustrates the difference between the χ^2 values from the method of moments and the minimum χ^2 method for the gamma distribution. The average improve-

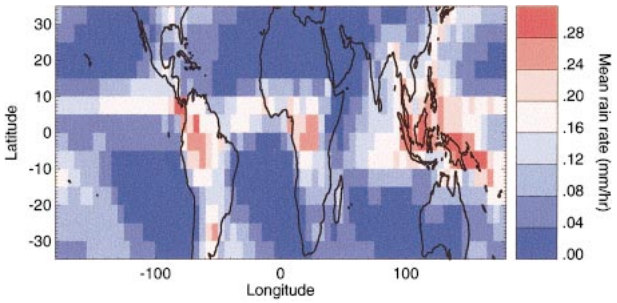

FIG. 3. Mean rain rate from TRMM 3G68 data (including zero rain rates) for the period 8 Dec 1997–30 Jun 2001. Each grid box is 5° × 5°.

ment by the minimum χ^2 method is 151%, and the median is 128%. In rainy regions such as tropical South America, tropical Africa, or the western Pacific, there is only modest improvement, but in dry regions, where the mean rain rates are less than 0.5 mm h^{-1}, the minimum χ^2 method results are better than those computed by the method of the moments. The maximum-likelihood method may be a versatile and important alternative, but the minimum χ^2 estimation provides a convenient tool to compare between the gamma and lognormal distribution, and so only results from the minimum-χ^2 method are presented here.

For efficiency, the initial guesses of parameters are, however, based on the sample median and variance values. In this study, the median is used instead of the mean because the median is often preferred as a measure of central tendency for heavily skewed frequency distributions (Swift and Schreuder 1981). A sufficient number of iterations is applied to all test functions in order to achieve the convergence, and the χ^2 values from the gamma and lognormal distributions are compared for evaluating the closeness of fitted distributions to the underlying rain-rate data.

d. Model

To fit the theoretical PDFs to the data, the data must first be binned. The actual number of observations in each bin is compared with the expected number from the theoretical PDF, and the parameters of the distribution are calculated using the minimum χ^2 method. There is no general method to find the optimal choice for the bin width and number of groups, but some useful guidance is given in Croarkin and Tobias (1999). They suggest using $2n^{2/5}$ as a good starting point for choosing the number of bins, where n is the total number of observations available. Another common practice requires every group to have at least five data points (Croarkin and Tobias 1999; Kedem et al. 1990).

Here, we have tried to find bin sizes that work satisfactorily for a range of locations and mean rain rates. After extensive experimentation, we chose to use 20 equal-width bins between 0.001 and 2.5 mm h^{-1}. This selection includes 99.7% of all nonzero rain rates. In this study, a unified dynamic range and bin density are applied for easier comparison between two test functions. To check the appropriateness of this choice of bins, the results from different bin numbers and dynamic ranges are investigated in section 3b.

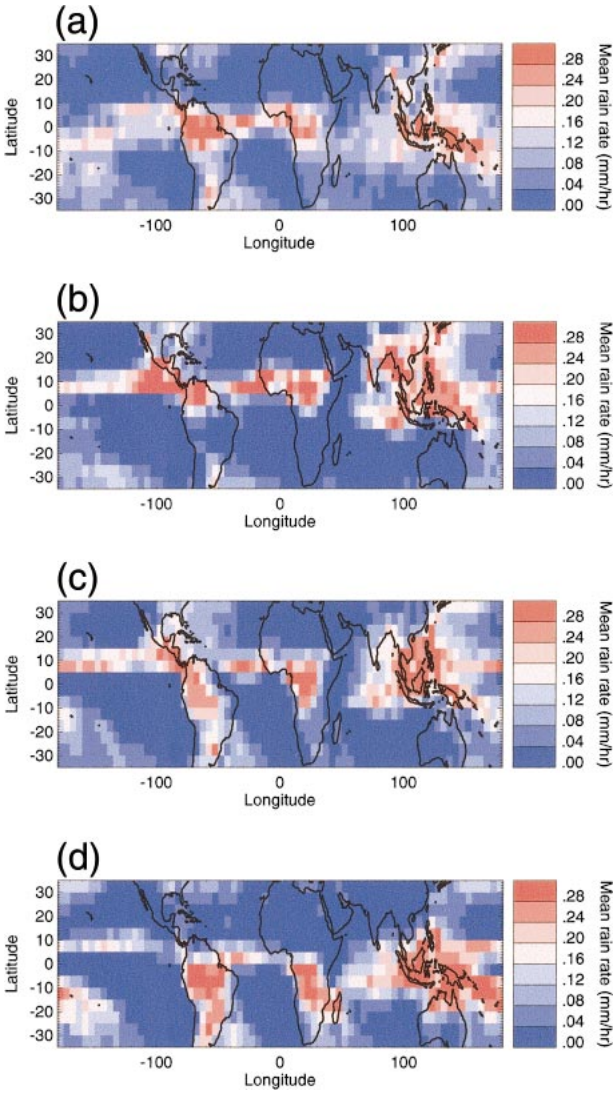

FIG. 4. Seasonal mean rain rates for (a) MAM, (b) JJA, (c) SON, and (d) DJF.

3. Results

a. Basic statistics of the TRMM data

In this section, the basic statistical and climatic characteristics of the TRMM data are investigated. Figure 3 shows the mean rain rates for the 5° × 5° grid for the period from 8 December 1997 to 30 June 2001. The mean rain-rate values are not sensitive to the grid resolution. Rainfall is concentrated over tropical land masses and over the ocean in the intertropical convergence

TABLE 1. The mean and median ϕ values from different bin sizes and dynamic ranges with the gamma (G) and lognormal (L) distributions and the difference between the two distributions.

Bin sizes (mm h^{-1})	Mean ϕ			Median ϕ			Log better fit (%)
	Gamma (G)	Lognormal (L)	Difference ($G-L$)	Gamma (G)	Lognormal (L)	Difference ($G-L$)	
Dynamic range 0.001–2.5 (includes 99.7% data)							
0.06	0.106	0.093	0.012	0.069	0.070	−0.002	44.94
0.12	0.069	0.048	0.020	0.043	0.037	0.004	55.31
Dynamic range 0.001–1.5 (includes 98.8% data)							
0.06	0.116	0.110	0.006	0.074	0.075	−0.002	41.18
0.12	0.076	0.062	0.013	0.044	0.042	−0.001	46.56
Dynamic range 0.01–2.5 (includes 85.0% data)							
0.06	0.119	0.106	0.012	0.079	0.078	−0.004	41.74
0.12	0.070	0.054	0.016	0.043	0.041	0.000	49.35
Dynamic range 0.01–1.5 (includes 83.3% data)							
0.06	0.138	0.128	0.009	0.087	0.089	−0.003	38.58
0.12	0.080	0.069	0.011	0.049	0.047	−0.002	43.57
Dynamic range 0.05–2.5 (includes 52.0% data)							
0.06	0.161	0.162	0.001	0.106	0.109	−0.003	36.64
0.12	0.088	0.088	−0.001	0.055	0.060	−0.004	34.73
Dynamic range 0.05–1.5 (includes 51.7% data)							
0.06	0.196	0.193	0.003	0.128	0.128	−0.002	37.21
0.12	0.106	0.105	0.001	0.066	0.068	−0.003	32.54

zone (ITCZ). The highest rainfall occurs over the Maritime Continent between the Indian and Pacific Oceans. The fraction of measurements with nonzero rain rate, p, is consistent with the mean rain-rate distribution, but p itself is sensitive to the grid resolution. As the grid size increases, the probability of rain occurring somewhere in the box also increases. The mean p values for $2.5° \times 2.5°$, $5° \times 5°$, and $10° \times 5°$ boxes are 20.35%, 37.20%, and 49.14%, respectively. The seasonal mean rain rates are shown in Fig. 4. The seasonal p values have similar geographical patterns to the seasonal mean rain-rate distributions. Over land, the seasonal shift of rainfall into the summer hemisphere is apparent. Seasonal variations in the oceanic ITCZs can also be clearly seen.

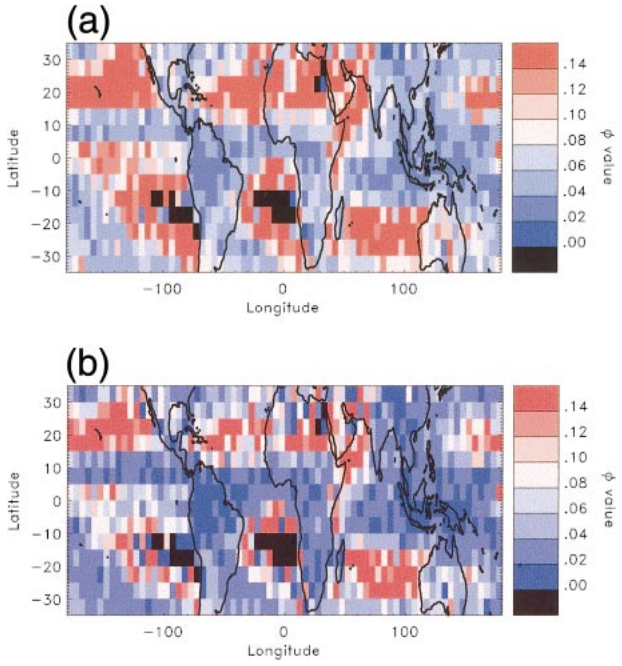

FIG. 5. The ϕ values for (a) 40 bins and (b) 20 bins.

b. *The sensitivity of model parameters*

1) SENSITIVITY TO BIN SIZE

Histograms were tested with different constant bin widths to investigate the sensitivity of the fitted PDFs to bin size. Table 1 shows results from two different bin widths, 0.06 and 0.12 mm h^{-1}. As the bin width increases, the χ^2 value decreases. The major reason for this is that there are more events per bin, and the error of estimating the number within each bin decreases. Although histograms with wider bins have smaller χ^2 values, wider bins do not represent the histogram as well. For most of the results presented here, the histograms are constructed using 20 equal-sized bins between 0.001 and 2.5 mm h^{-1} (bin width is 0.06 mm h^{-1}). This covers the overwhelming majority of observed rain rates. The overall spatial distribution of the χ^2 values is not sensitive to the choice of bin width (Fig. 5).

2) SENSITIVITY TO DYNAMIC RANGE

The minimum and maximum of values included in the histogram (the dynamic range) also affect fits of the

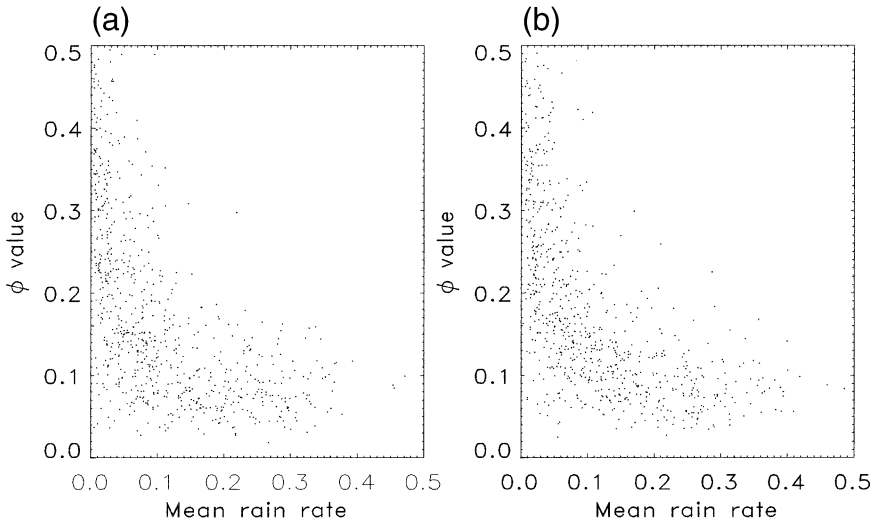

FIG. 6. Scatterplots of seasonal mean rain rate and the ϕ value. The ϕ values indicate the difference between the annual rain-rate histogram and the histogram for (a) JJA and (b) DJF. When the seasonal mean rain rate is large, the shape of the rain-rate histogram for that season is similar to the annual rain-rate histogram.

theoretical PDFs to the rain-rate data. The data in both tails (which are larger than the maximum value or smaller than the minimum value of the dynamic range) are not included, and the truncations in each tail do not have an effect on the overall shape of the rain-rate histogram. Because of the large geographical variation in rainfall, it is difficult to find a single dynamic range suitable for all areas of interest.

Several approaches were tried to test the sensitivity of the results to the choice of dynamic range. One approach is to select the dynamic range locally based on the sample mean and variance. The maximum value is set to the sample mean plus 6 times *s*, where *s* is the sample standard deviation, and bin width is given by 0.3 times *s*. The minimum value is set to 0.001. Although some rainy regions exhibit well-matched distributions with the flexible dynamic range, the overall results are degraded in comparison with choosing a fixed dynamic range. Over dry regions, the rain-rate distributions deviate noticeably from normal, and the sample mean and variance do not represent data well.

Six different sets of fixed dynamic ranges were tested. The mean and median χ^2 value for each dynamic range along with the difference between the gamma and lognormal distribution are shown in Table 1. The last column of Table 1 shows the fractional ratios (%) where the lognormal distribution has smaller χ^2 values than the gamma distribution.

In general, as the dynamic range increases, the average χ^2 values decrease. The ratio of the number of rejected data points to the total number of rainy data points and the shape of the histogram turn out to be sensitive to the minimum value rather than the maximum value. Changing the minimum value from 0.01 to 0.001 gives smaller χ^2 values at most locations. A minimum value of 0.05 fits slightly decreases the χ^2 values in most rainy regions when compared with a minimum value of 0.01.

In most rainy regions and over the continents, a maximum value of 2.5 gives the smallest χ^2 values. On the contrary, results over the ocean, where mean rain rates are roughly in the range from 0.04 to less than 0.15 mm h^{-1}, are poorer when using a large maximum value.

In this study, the primary dynamic range is chosen between 0.001 and 2.5 mm h^{-1}. With this dynamic range, the majority of the observed values from the rainy regions are covered, but dry regions have some zero frequency bins in heavy rain rates.

3) SEASONAL VARIATION

Seasonal variations of the rain-rate histograms are investigated by calculating the ϕ values between the rain-rate histogram of each season and the overall histogram including all seasonal data. Relations between each seasonal histogram and the overall histogram are dependent on the seasonal mean rain rate. Figure 6 shows scatterplots between seasonal mean rain rates and χ^2 values calculated from the overall histogram and the histograms for June–July–August (JJA; hereinafter 3-month periods are denoted by the first letter of each respective month) and DJF. If the seasonal mean rain rate is high, the rain-rate histogram of that season is similar to the overall rain-rate histogram. In regions with lower mean rain rates, the shape of seasonal histograms can be quite different from the annual histogram.

c. Gamma and lognormal distributions

The two test functions (gamma and lognormal distributions) are fitted by the minimum χ^2 method. Figure

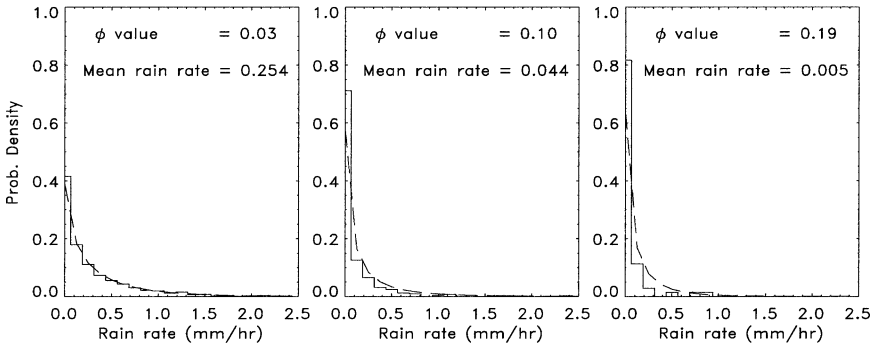

FIG. 7. Examples of fitting the rain-rate histogram with a gamma distribution using the minimum χ^2 method.

7 shows some examples of fitting the gamma distribution to data at three locations with different rainfall characteristics, and it also illustrates the relationship between the ϕ value and mean rain rates. The ϕ values range from 0.03 to 0.19 in these examples. All examples use the same bin size and dynamic range.

A comparison between the gamma and lognormal distributions shows that both PDFs give similar results. Fits in rainy regions generally have smaller χ^2 values. Both distributions fail to find a fit when the mean rain rate is less than approximately 0.01 mm h^{-1}. The gamma distribution generally has a smaller χ^2 value over more regions, but the lognormal distribution also has a comparable χ^2 value with that of the gamma distribution.

In general, the gamma distribution tends to underestimate heavy rain rates and very light rain rates as compared with the PDF fitted by the lognormal distribution. The lognormal distribution, on the other hand, tends to underestimate intermediate rain rates as compared with the gamma distribution. Figures 8a and 8b illustrate the differences between fits by the gamma and lognormal distributions. The two locations shown have the largest difference of the χ^2 values between two PDFs.

Figure 9 shows maps of the ϕ value for the two distributions and the difference between them. The gamma distribution generally outperforms the lognormal distribution in rainy regions, where the mean rain rates are over 0.18 mm h^{-1}. These regions have a wide range of rain rates and relatively high frequencies of high rain rates. In contrast, the lognormal distributions have smaller ϕ values where the mean rain rate is less than 0.15 mm h^{-1} and the very light rain rates have the highest frequencies. The spatial distribution of the goodness of fit of the two distributions is consistent regardless of bin numbers and dynamic ranges.

To test the accuracy of estimating the probability of the theoretical PDFs, the mean and variance computed by the parameters from the fitted distributions were compared with sample mean and variance from the original, ungrouped rain-rate data with values in the range from 0.001 to 2.5 mm h^{-1}.

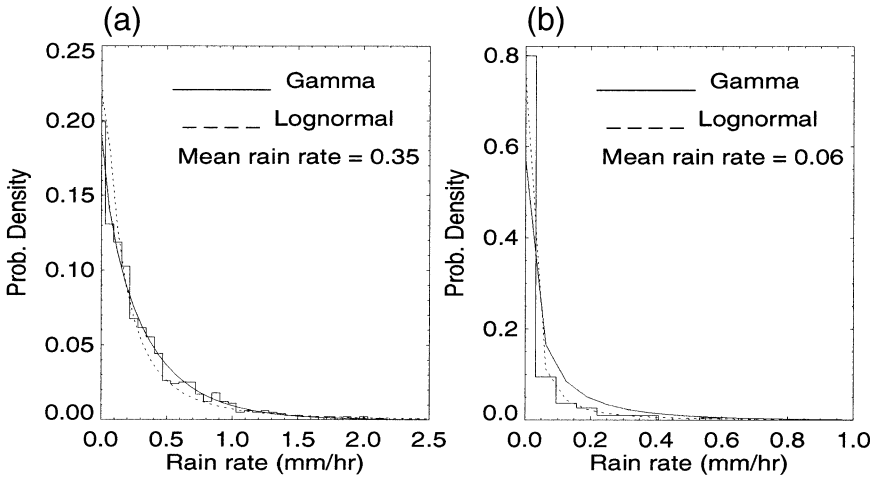

FIG. 8. Comparison of the gamma and lognormal distribution fits to the rain-rate histograms. (a) The gamma fit is better than the lognormal fit, and (b) the lognormal fit is better than the gamma fit. The locations selected have the largest differences of the ϕ value between the gamma and lognormal distribution.

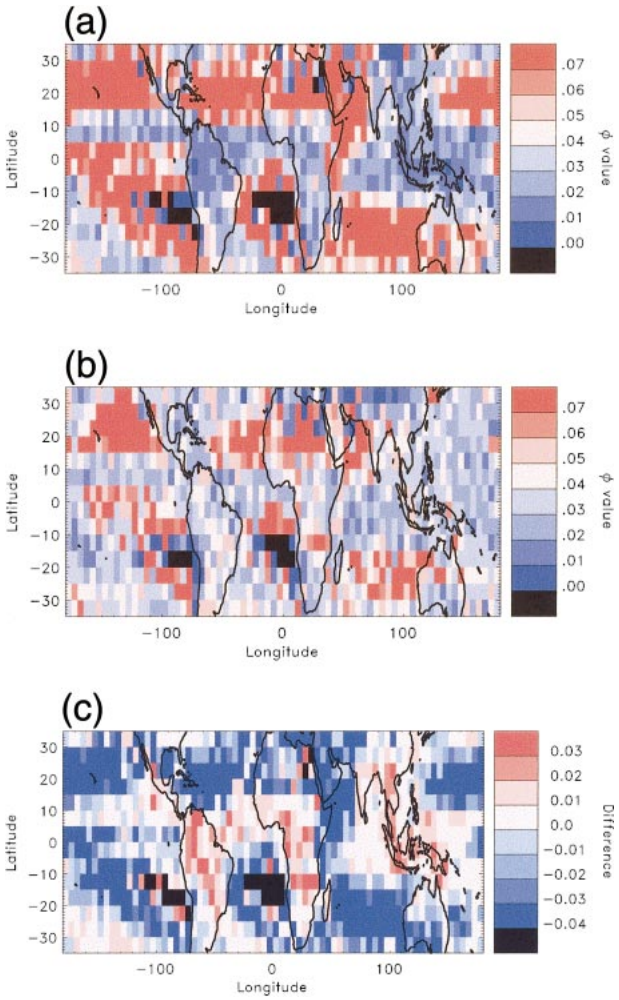

FIG. 9. The ϕ values from (a) the gamma distribution and (b) the lognormal distribution. (c) The difference of the ϕ values. The positive values in the (c) indicate regions where the gamma fit is better than the lognormal fit.

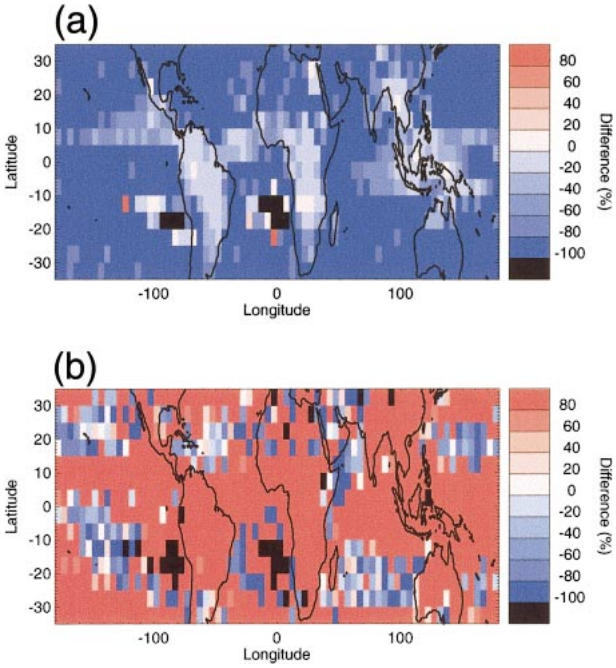

FIG. 10. Comparison between sample mean and parametric mean from (a) the gamma distribution and (b) the lognormal distribution. Positive values mean that the parametric mean is larger than the sample mean at a given location.

It is conventional to express the precision of parameter estimation in terms of the variance of the sampling distribution and, in particular, to compare competing theoretical distributions in terms of the variances of the resulting sampling distributions (Wilks 1990). Both test functions generally give biased estimates of the observed mean and variance. The geographical pattern of the means estimated from the gamma distribution is similar to that of the sample means, but the gamma distribution underestimates the mean and variance in all areas (average underestimate is 73%). The lognormal distribution, in contrast, overestimates the mean and variance in 53% of the boxes, and the mean estimated from the lognormal distribution is more than 2 times the sample mean. Figure 10 shows maps of the difference between the sample mean and the means from the two fitted theoretical distributions. As compared with Fig. 9c, as one distribution yields a closer fit to the data, the mean estimated from that distribution approaches the sample mean.

To explain why the gamma parametric mean underestimates the sample mean in most regions, the relationship between these two variables was tested using a Monte Carlo method. First, random numbers distributed according to the gamma function are generated at a given shape parameter; then, the minimum χ^2 method is applied to find α and β. The minimum χ^2 method successfully finds the given α and β for all cases, but the mean estimated from the gamma fits underestimates the sample mean when α is less than 0.5, independent of sample size. Considering that most shape parameters in Fig. 9a are much less than 1 because the shape of the rain histogram is highly skewed to the right, this result explains why the gamma parametric means underestimate the sample mean. To remedy the underestimates, a minimum value for the shape parameter α is set to 0.3. Figure 11 shows the difference between the sample means and the means from the gamma distribution fits with minimum α. Figure 12a shows the differences of the ϕ values between the gamma distribution with and without minimum α. Figure 12b shows the differences with the lognormal distribution shown in Fig. 9b. One immediately notices that the difference between parametric mean and sample mean decreases and ϕ values increase when the minimum value of α is set. The difference of the ϕ value caused by the constraint of minimum values of α is larger in the dry regions because the dry regions have smaller α than the

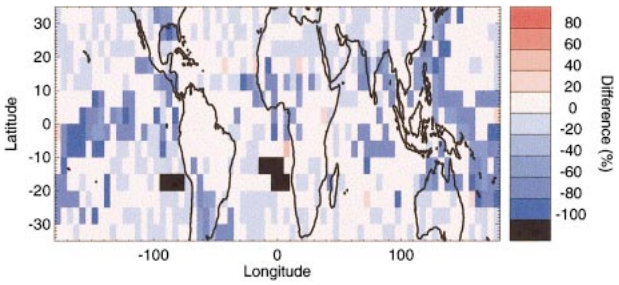

FIG. 11. Comparison between sample mean and parametric mean from the gamma distribution using a minimum value of the shape parameter α.

rainy regions. When compared with the lognormal distribution, the gamma distribution still provides a closer fit over the rainy regions, although the regions where the lognormal distribution has smaller ϕ values than gamma increases.

In the case of the lognormal distribution, the main reason for the overestimate is that the fitted distribution has an inflated tail as compared with the observed histogram. The overestimate by the lognormal distribution was also reported in Meneghini and Jones (1993).

d. Effects of averaging

One of the purpose of the TRMM research is to find the nature of the change of rainfall statistics as the averaging area increases. In this section, we examine the effects of the grid resolution on fitting the theoretical distributions to the data to find the characteristics of the rain-rate distribution. As mentioned in section 2, the 3G68 TRMM product is an average rain rate from all instantaneous observations within $0.5° \times 0.5°$ grid boxes. Because the data are stratified by hour, the observations that are averaged within a grid box are nearly simultaneous. We created area-averaged datasets with resolutions of $2.5° \times 2.5°$, $5° \times 5°$, and $10° \times 5°$ latitude–longitude. Figure 13 shows the spatial distribution of the ϕ values for different grid resolutions for the gamma distribution. As the grid box size decreases, the ϕ values generally increase; they fail to achieve a fit at a larger number of locations. The spatial pattern of ϕ values, however, does not depend strongly on resolution, and the regional preference between two test functions is pertained to all grid resolutions. The gamma distribution still fits better in rainy regions, while the lognormal distribution gives better results in arid areas.

We compared the histograms from each $0.5° \times 0.5°$ grid box in the 3G68 TRMM data, with the histogram from the data combined into a $5° \times 5°$ latitude–longitude grid box. Although the accuracy of estimates degrades significantly as area decreases, the relative performance of different-sized boxes is not much different. Figure 14a is the histogram from a $5° \times 5°$ grid box through the whole data period. Figure 14b shows the histogram

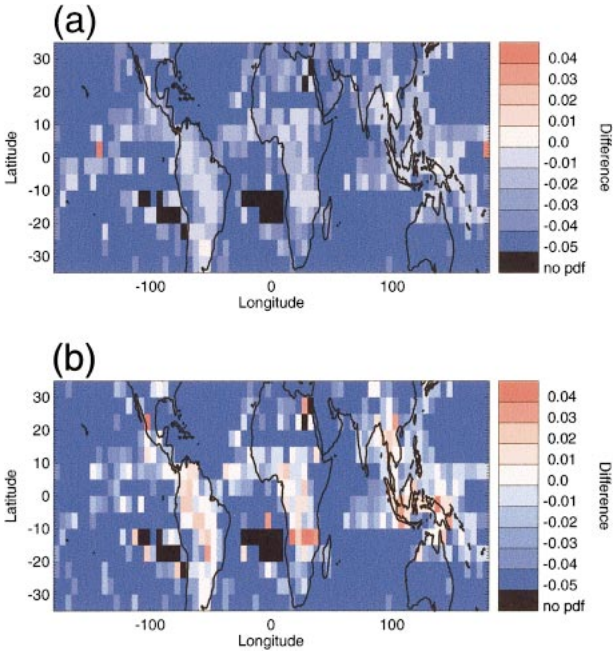

FIG. 12. (a) Difference of the ϕ values between the gamma distribution with and without a minimum value of α. (b) As in (a), but between the lognormal distribution and gamma distribution with a minimum value of α. A positive value indicates regions where the gamma fit with a minimum value of α is closer to the shape of the rain-rate histogram than the other fit.

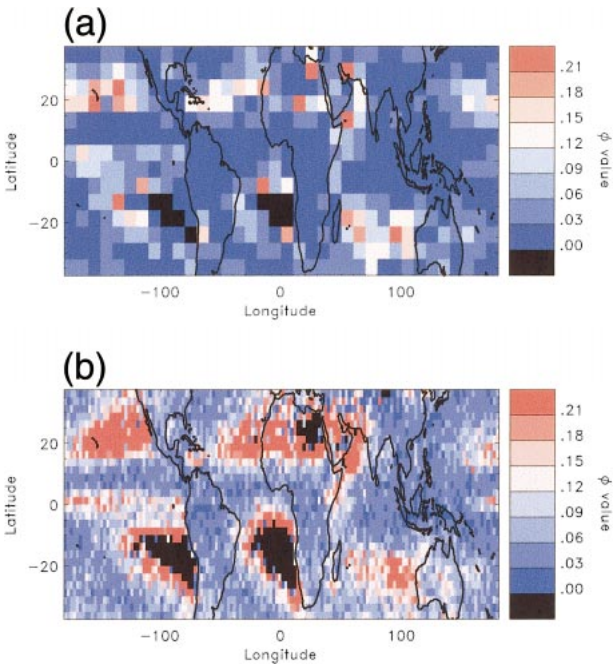

FIG. 13. The ϕ values for two different grid resolutions. The gamma distribution is fitted using the minimum χ^2 method. The size of the grid boxes is (a) $10° \times 5°$ and (b) $2.5° \times 2.5°$.

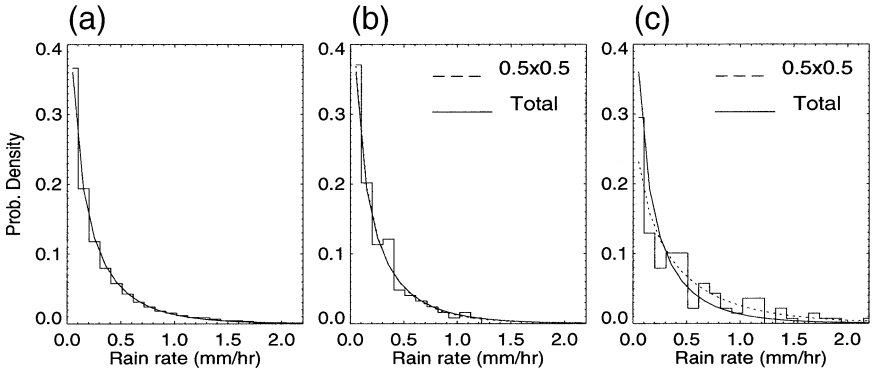

FIG. 14. Comparison of the histogram and fitted gamma distributions for different grid resolutions. The histogram for (a) the 5° × 5° grid box and (b), (c) the two 0.5° × 0.5° subboxes within (a) that have the smallest and largest difference from (a).

that is the least different from Fig. 14a, while Fig. 14c is the most different from Fig. 14a.

4. Summary and concluding discussion

We have investigated the spatial characteristics of nonzero rain rates to develop a PDF model of precipitation using the large rainfall dataset provided by the TRMM satellite. As a first step in using the TRMM data for the study of the PDFs of rain rates, the gamma and lognormal distributions were used to fit the data. The parameters of the fitted distributions are found by minimizing the χ^2 value between the data and the theoretical distributions. The sensitivity of model parameters to bin width and dynamic range was investigated. The results were not sensitive to the choice of bin width, but the choice of dynamic range had some effects on the results. The shape of the rain-rate histograms is sensitive to the choice of minimum value of the dynamic range everywhere, and the choice of maximum value of the dynamic range can affect the results, depending on mean rain rate and continentality. The result are generally improved for broader dynamic ranges.

Both the gamma and lognormal distributions were found to be good for representing the probability distributions from TRMM data. The gamma fits outperformed the lognormal fits in rainy regions, while the lognormal fits were better than the gamma fits in dry regions. Comparison between the two test functions showed that the gamma distribution underestimates high rain rates and light rain rates, while the lognormal distribution underestimated intermediate rain rates. As compared with the sample mean and variance from the original ungrouped rain-rate data, the parametric mean from the gamma distribution underestimated, while the lognormal distribution overestimated at all locations. This is in accord with previous results (Meneghini and Jones 1993). The difference in the estimation of high rain rates is thought to produce this result. Reducing the size of grid boxes reduces the sample size, which increases the uncertainty in the model fits, but the characteristics of the rain-rate distribution are not sensitive to the grid resolution.

The χ^2 values supported the proposed PDFs for describing variations in observed data. To compare the performance of the estimated PDFs at different locations, a statistical significance test was performed by the goodness-of-fit χ^2 test with the dynamic range from 0.001 to 1.5 (Fig. 15). Both test functions, in the area where rain is plentiful as compared with other areas, yielded a statistically reliable fit, while both of them failed to find a fit in very dry regions at the 5% significance level. Figure 15 confirms the regional preference between the two test functions dependent on the mean rain rates.

We tested the regional preference between the two test functions using the differently designed PDFs and found consistent results for all experiments. Based on our results, we suggest that the relative performance between the gamma and lognormal distribution for the rain-rate distribution is related to the mean rain rates at a given location.

The regional preference between the gamma and lognormal distributions could be useful for the calculation of the statistical properties of rain-rate fields and the simulation of precipitation in GCMs. This study suggests that the choice between the two test functions may

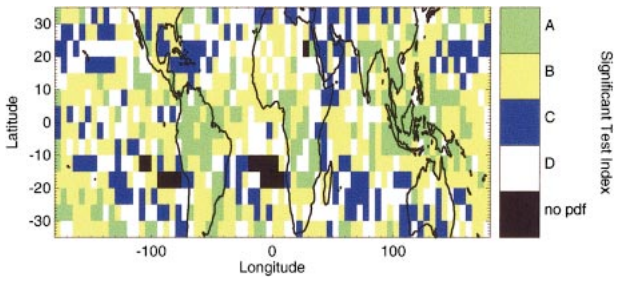

FIG. 15. Statistical significance of the fits at the 5% significance level: gamma only (A), both (B), lognormal only (C), and neither (D).

be helpful in increasing the accuracy of rainfall statistics and for accessing the accuracy of the retrieval algorithm of rain rates. This is also useful in determining the location dependence of rain-rate fields.

The maximum likelihood method might be an alternative to the minimum χ^2 method. When compared with the minimum χ^2 method, the calculated mean and variance using the maximum likelihood method are a little bit closer to the sample mean and variance in some regions. However, the maximum likelihood estimation does not provide any criterion for the comparison between two test functions such as the χ^2 value, and the early results from the maximum likelihood estimation in the representation of the observational distribution are even worse than the results from the minimum χ^2 method. Thus, to apply an approximate maximum likelihood estimation to the same dataset and the comparison between the two estimation methods poses a challenge for future work. Another interesting topic for future study is to represent the rain-rate PDFs with other functional PDF forms that have more flexibility than the two-parameter gamma and lognormal distributions used in this study.

Acknowledgments. The authors thank Thomas T. Wilheit and Benjamin Giese for their comments on an early version of the manuscript. The authors also thank the referees for insightful and helpful comments. The TRMM 3G68 data were obtained from the TRMM Science Data and Information System at the NASA Goddard Space Flight Center.

REFERENCES

Bell, T. L., 1987: A space–time stochastic model of rainfall for satellite remote sensing studies. *J. Geophys. Res.,* **92,** 9631–9643.

——, A. Abdullan, R. L. Martin, and G. R. North, 1990: Sampling errors for satellite-derived rainfall: Monte Carlo study using a space–time stochastic model. *J. Geophys. Res.,* **95,** 2195–2205.

Briggs, W. M., and D. S. Wilks, 1996: Estimating monthly and seasonal distributions of temperature and precipitation using the new CPC long-range forecasts. *J. Climate,* **9,** 818–826.

Casella, G., and R. L. Berger, 1990: *Statistical Inference.* Duxbury Press, 650 pp.

Croarkin, M. C., and P. Tobias, cited 1999: NIST/SEMATECH engineering statistics Internet handbook. [Available online at http://www.nist.gov/stat.handbook/.]

Haddad, Z. S., E. A. Smith, C. D. Kummerow, T. Iguchi, M. R. Farrar, S. L. Durden, M. Alves, and W. S. Olson, 1997: The TRMM ''Day-1'' radar/radiometer combined rain-profiling algorithm. *J. Meteor. Soc. Japan,* **75,** 799–809.

Ison, N. T., A. M. Feyerherm, and L. D. Bark, 1971: Wet period precipitation and the gamma distribution. *J. Appl. Meteor.,* **10,** 658–665.

Jameson, A. R., and A. B. Kostinski, 1999: Fluctuation properties of precipitation. Part V: Distribution of rain rates—Theory and observations in clustered rain. *J. Atmos. Sci.,* **56,** 3920–3932.

Kedem, B., L. S. Chiu, and G. R. North, 1990: Estimation of mean rain rate: Application to satellite observations. *J. Geophys. Res.,* **95** (D2), 1965–1972.

——, R. Pfeiffer, and D. A. Short, 1997: Variability of space–time mean rain rate. *J. Appl. Meteor.,* **36,** 443–451.

Kummerow, C., and Coauthors, 2000: The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit. *J. Appl. Meteor.,* **39,** 1965–1982.

Martin, R., 1989: A statistic useful for characterizing probability distributions, with application to rain rate data. *J. Appl. Meteor.,* **28,** 354–360.

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.

——, ——, T. Iguchi, K. Okamoto, and J. Kwiatkowski, 2001: Statistical methods of estimating average rainfall over large space–timescales using data from the TRMM precipitation radar. *J. Appl. Meteor.,* **40,** 568–585.

North, G. R., 1987: Sampling studies for satellite estimation of rain. Preprints, *10th Conf. on Probability and Statistics in Atmospheric Science,* Edmonton, AB, Canada, Amer. Meteor. Soc., 129–135.

Öztürk, A., 1981: On the study of a probability distribution for precipitation totals. *J. Appl. Meteor.,* **20,** 1449–1505.

Revfeim, K. J. A., 1982: Comments on ''On the study of a probability distribution for precipitation totals.'' *J. Appl. Meteor.,* **21,** 1942–1945.

Swift, J. L. W., and H. T. Schreuder, 1981: Fitting daily precipitation amounts using the Sb distribution. *Mon. Wea. Rev.,* **109,** 2535–2540.

TSDIS/TRMM, 1999: File specification for TRMM products—Level 2 and Level 3. Vol. 4, TSDIS-TSU interface control specification. 71 pp. [Available online at http://trmm.gsfc.nasa.gov/3a26.html.]

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Climate,* **3,** 1495–1501.

——, 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

——, and K. L. Eggleston, 1992: Estimating monthly and seasonal precipitation distributions using the 30- and 90-day outlooks. *J. Climate,* **5,** 252–259.