A Mixed Exponential Distribution Model for Retrieving Ground Flash Fraction from Satellite Lightning Imager Data

W. J. KOSHAK

NASA Marshall Space Flight Center, Huntsville, Alabama

(Manuscript received 6 January 2010, in final form 22 June 2010)

ABSTRACT

A Bayesian inversion method is introduced for retrieving the fraction of ground flashes in a set of flashes observed from a (low earth orbiting or geostationary) satellite lightning imager. The method employs a constrained mixed exponential distribution model to describe the lightning optical measurements. Because the method also retrieves certain population statistics of ground and cloud flash optical properties, the method can be applied to an arbitrary geographical region, including those regions where the lightning optical statistics either are not known or are difficult to obtain. The approach is tested by performing simulated retrievals, and retrieval error statistics are provided. A first-attempt retrieval of the global geographical distribution of ground flash fraction is obtained using the 5-yr Optical Transient Detector (OTD) dataset; the spatially averaged ground flash fraction over the global-scale domain studied was 0.151 with a standard deviation of 0.081. The ability to retrieve ground flash fraction has important benefits to the atmospheric chemistry community. For example, using the method to partition the existing OTD/Lightning Imaging Sensor (LIS) satellite global lightning climatology into separate ground and cloud flash climatologies would improve estimates of regional and global lightning nitrogen oxides (NO_x) production; this, in turn, would improve both regional air quality and global chemistry/climate model predictions.

1. Introduction

The study by Koshak (2010) showed that the distributions of ground and cloud flash optical characteristics, as seen from the Optical Transient Detector (OTD), overlap appreciably. Therefore, space-based flash-type discrimination (on a flash-by-flash basis) is fundamentally difficult. However, Koshak (2010) also indicated that the mean values of the optical characteristics for ground and cloud flashes are distinct, so that an analysis of a sample of N flashes could possibly provide information about the fraction of ground flashes (i.e., the ground flash fraction) within the sample.

The follow-on study by Koshak and Solakiewicz (2011) confirmed this hypothesis. They established a mathematical formalism for analyzing the mean optical data and provided a straightforward approach for estimating the ground flash fraction in a sample of flashes. In numerical tests, they retrieved the ground flash fraction for 52 regions that were widely distributed across the conterminous United States (CONUS). The retrieval errors were under 11.1%, and as low as 6.1%. However, their method is based on employing CONUS-based estimates of the population means of the optical characteristics. This implies that the ground flash fraction retrieval accuracy could degrade if the method is applied outside the CONUS where the CONUS mean estimates might no longer be accurate. In other words, if the optical characteristics that are used [i.e., the maximum number of events in a group (MNEG), or the maximum group area (MGA)] vary greatly from their respective CONUS-mean values, then retrieval error will increase to unacceptable levels.

Though Koshak and Solakiewicz (2011) demonstrated that the ground and cloud flash values of MNEG and MGA did not vary greatly across CONUS (a region of highly variable lightning, highly variable thunderstorm structures, and highly variable cloud morphologies with diverse cloud-scattering properties), it is still uncertain how variable MNEG and MGA are across the globe. To determine the global variability, satellite-based total lightning observations would need to be compared with independent ground flash observations so that the satellite observations could be separated into ground and

Corresponding author address: Dr. William Koshak, 320 Sparkman Drive, Earth Science Office, VP61, Robert Cramer Research Hall, NASA Marshall Space Flight Center, Huntsville, AL 35805. E-mail: william.koshak@nasa.gov

DOI: 10.1175/2010JTECHA1438.1

cloud flashes. Statistics of the spatial variation of MNEG and MGA for both the ground and cloud flashes could then be computed. However, because of the lack of ground flash detection systems across the globe (with particularly poor ground flash detection over the oceans), the present ability to make detailed assessments of the global variability of MNEG and MGA is somewhat limited; and no studies have yet been conducted to assess the variability given the existing coverage. Fortunately, this will not always be the case because improvements to global ground flash detection/coverage are in progress [e.g., the Vaisala Corporation has introduced a new Global Lightning Dataset (GLD360) product with plans for continued network expansion (F. DeFina, Vaisala Corporation, 2009, personal communication)].

Because of the lack of knowledge of the global variability of MNEG and MGA, and because relying on CONUS-based estimates of these optical characteristics is fundamentally restrictive, it is highly desirable to devise a retrieval algorithm that not only retrieves the unknown ground flash fraction for a particular geographical region, but also the population means of the unknown optical characteristics (either MNEG or MGA) for that region. Such an algorithm would obviously be more robust than that provided in Koshak and Solakiewicz (2011) because it would not depend on using CONUS-based estimates of lightning optical characteristics.

Other motivations for this study are of a general nature and have been previously discussed. The ground flash fraction (or the closely related parameter called the "Z ratio," given by the ratio of the number of cloud flashes to ground flashes) is thought to be particularly useful in a number of areas, for example, severe weather warning, lightning–convection relationships, lightning nitrogen oxide (NO_x) production, the contribution of lightning to the global electric circuit, and cross-sensor validation (Koshak 2010; Boccippio et al. 2001). In addition, a satellite-based ground flash fraction retrieval product would obviously be valuable for cross comparison/validation with global ground flash measurements, such as the GLD360 dataset mentioned above.

In this study, a method is introduced for retrieving the ground flash fraction of a sample of N flashes observed from space [e.g., from the OTD, the Lightning Imaging Sensor (LIS), or the future Geostationary Operational Environmental Satellite (GOES)-R Geostationary Lightning Mapper (GLM)]. The method is more sophisticated and more general than that provided in Koshak and Solakiewicz (2011) because retrievals of the unknown population means of optical characteristics (i.e., of MGA) are obtained. Hence, the method is useful for global application.

The study begins by introducing a general form for the mixture of two optical distributions (section 2). The mixture consists of a linear combination of two distributions—one distribution represents the distribution of a ground flash optical characteristic and the other distribution represents the distribution of the associated cloud flash optical characteristic. The mean and variance of the mixture distribution are provided. A Bayesian method for retrieving the ground flash fraction and attributes of the ground and cloud flash optical distributions is provided in section 3. Next, section 4 introduces a special case of the mixture distribution, called the "mixed exponential distribution model," which is the primary focus of this study. Section 5 details the "label switching" ambiguity associated with mixed exponential distribution models, and section 6 introduces a useful approach for initializing the numerical search associated with the Bayesian retrieval process. Section 7 shows how population statistics can be used as constraints to the retrieval process. Finally, numerical tests of the Bayesian retrieval method (applied to the mixed exponential distribution model) are provided in section 8; the method is also applied in section 9 as a first attempt to retrieve the ground flash fraction on a global scale using OTD data. Concluding remarks are provided in section 10.

2. The distribution of a mixture

Consider a set of $i = 1, \ldots, N$ flashes that are observed over a time period Δt by a satellite lightning imager (e.g., a low-earth-orbiting sensor like the LIS, or a geostationary sensor like the future GLM). For each of the N flashes, the sensor measures a particular optical characteristic x. For example, this characteristic could represent flash radiance, flash area, flash duration, the number of optical groups in the flash, or the number of optical events in the flash. One could consider several other optical characteristics, such as the maximum number of events in a 2-ms sensor frame time for a given flash, radiance of the first event in the flash, radiance of the brightest group, MNEG, and MGA. Note that the common nomenclature for describing the OTD/LIS data is used; that is, a flash is composed of optical groups, and each optical group is composed of optical events (see Mach et al. 2007).

In general, the distribution of the optical characteristic x will be different for ground and cloud flashes. See examples of the differences (for various optical characteristics) in Koshak (2010). The distribution of x for ground flashes is denoted by $p_g(x)$, and the distribution of x for cloud flashes is denoted by $p_c(x)$. Hence, the distribution of a mixture of ground and cloud flashes, called the mixture distribution $p(x)$, is given by

$$p(x) = \alpha p_g(x) + (1 - \alpha)p_c(x), \quad (1)$$

where α is the ground flash fraction, as discussed in Koshak and Solakiewicz (2011). For the set of N flashes, αN of them are ground flashes and $(1 - \alpha)N$ are cloud flashes. Formally, note that $p(x)$, as well as $p_g(x)$ and $p_c(x)$, are probability density functions (pdfs); the terms ''distribution'' and ''density'' are used interchangeably. Hence, the population mean and variance of the mixture distribution are

$$\begin{aligned}\mu &\equiv \int_{-\infty}^{\infty} xp(x)\,dx = \alpha\mu_g + (1 - \alpha)\mu_c,\\ \sigma^2 &\equiv \int_{-\infty}^{\infty} (x - \mu)^2 p(x)\,dx,\\ &= \alpha\sigma_g^2 + \alpha(1 - \alpha)(\mu_g - \mu_c)^2 + (1 - \alpha)\sigma_c^2. \quad (2)\end{aligned}$$

Here, the population mean and variance of the ground and cloud flashes are, respectively,

$$\begin{aligned}\mu_g &\equiv \int_{-\infty}^{\infty} xp_g(x)\,dx, \quad \sigma_g^2 \equiv \int_{-\infty}^{\infty} (x - \mu_g)^2 p_g(x)\,dx,\\ \mu_c &\equiv \int_{-\infty}^{\infty} xp_c(x)\,dx, \quad \sigma_c^2 \equiv \int_{-\infty}^{\infty} (x - \mu_c)^2 p_c(x)\,dx. \quad (3)\end{aligned}$$

Obtaining the results in (2) are straightforward, and the second result in (2) requires a little algebra.

3. Bayesian inference of model parameters

a. The MAP solution

As discussed above, the satellite lightning imager will observe N flashes over a particular region during some time interval Δt. Considering the optical characteristic x, this means that one has the N vector of observations $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ from which to retrieve the ground flash fraction α. In general, however, a particular retrieval algorithm might not use every piece of information in $\mathbf{x}$ to retrieve α. To include all classes of Bayesian-type retrievals, one can consider a vector $\mathbf{d}$ that is derived from $\mathbf{x}$. For example, the components of $\mathbf{d}$ could be the first several moments of the sample observations in $\mathbf{x}$; that is, the first component of $\mathbf{d}$ could be the sample mean $\overline{x}$ of the N observations, the second component could be the sample variance s^2 of the N observations, and so on. The entire vector $\mathbf{d}$ might even degenerate into a single scalar, for example, $\mathbf{d} \rightarrow \overline{x}$. The components of $\mathbf{d}$ could even represent various sample survival values (complementary cumulative distribution values), that is, the first component of $\mathbf{d}$ could be the fraction of the elements in $\mathbf{x}$ that exceed some value X, and the second component of $\mathbf{d}$ could be the fraction that exceed some value Y, and so on. Finally, if all of the raw data are used directly, then one simply has $\mathbf{d} = \mathbf{x}$. This is the preferred approach because one intuitively expects more accurate retrievals when all of the observational information is used. In the case where $\mathbf{x}$ might be a very large vector (i.e., N very large), practical limitations might warrant using the other forms of $\mathbf{d}$ described so that the resulting dimensions of $\mathbf{d}$ are less than N.

With all of these possibilities in mind for the choices of $\mathbf{d}$, one can write Bayes' law as

$$P(\mathbf{v}|\mathbf{d}) = \frac{P(\mathbf{d}|\mathbf{v})P(\mathbf{v})}{P(\mathbf{d})}, \quad (4)$$

where $\mathbf{v} = (v_1, \ldots, v_\eta) = (\alpha, v_2, \ldots, v_\eta)$. The $\eta - 1$ components $(v_2, \ldots, v_\eta)$ are population statistics of $p_g(x)$ and $p_c(x)$. The attributes $\mathbf{v}$ define the η parameters of a model that, in turn, are used to described the mixture distribution $p(x)$ in (1). The sample data vector $\mathbf{d}$ is used to infer the model parameters $\mathbf{v}$. Note that (4) provides an expression for determining the probability of getting the model parameters $\mathbf{v}$ given the sample data $\mathbf{d}$. The notation in (4) follows that provided in Rodgers (2000), that is, $P(\mathbf{v}|\mathbf{d})$ is the posterior pdf, $P(\mathbf{v})$ is the prior pdf, $P(\mathbf{d}|\mathbf{v})$ is the measurement pdf, and $P(\mathbf{d})$ is a pdf that serves as a normalization constant. Essentially, prior to any measurement one has the prior knowledge of the state $\mathbf{v}$ given by $P(\mathbf{v})$. After specific measurements are made, this prior knowledge is updated by $P(\mathbf{d}|\mathbf{v})$, which describes the knowledge of $\mathbf{d}$ if the state were $\mathbf{v}$; that is, $P(\mathbf{d}|\mathbf{v})$ describes the forward problem. The forward problem describes how one maps the unknown model parameters $\mathbf{v}$ into measurement space $\mathbf{d}$. Carrying out the algebra in Bayes' law given in (4), one obtains $P(\mathbf{v}|\mathbf{d})$, that is, the probability density of the state $\mathbf{v}$ when the measurement $\mathbf{d}$ is given.

The approach here is to determine the most probable $\mathbf{v}$ given the sample data $\mathbf{d}$. Because $\mathbf{d}$ is given, it can be treated as a constant vector. The maximum a posteriori (MAP) solution described in Rodgers (2000) is the value of $\mathbf{v}$ that maximizes $P(\mathbf{v}|\mathbf{d})$. Formally, the maximization is expressed by obtaining the critical point solution to the following set of equations:

$$\begin{aligned}&\frac{\partial P(\mathbf{v}|\mathbf{d})}{\partial \mathbf{v}} = \mathbf{0}\\ &\quad \Rightarrow \frac{\partial}{\partial \mathbf{v}}\left[\frac{P(\mathbf{d}|\mathbf{v})P(\mathbf{v})}{P(\mathbf{d})}\right] = \frac{1}{P(\mathbf{d})}\frac{\partial}{\partial \mathbf{v}}[P(\mathbf{d}|\mathbf{v})P(\mathbf{v})] = \mathbf{0}\\ &\quad \Rightarrow \frac{\partial}{\partial \mathbf{v}}[P(\mathbf{d}|\mathbf{v})P(\mathbf{v})] = \mathbf{0}. \quad (5)\end{aligned}$$

b. Model parameter independence

The attributes (model parameters) are commonly assumed to be independent. For example, one does not expect the ground flash fraction itself to depend on the population mean optical statistics, and the population statistics of ground and cloud flashes are expected to be reasonably independent. The condition for independence can be expressed as

$$P(\mathbf{v}) = P(v_1, v_2, \ldots, v_\eta) = P(v_1) \times P(v_2) \times \cdots \times P(v_\eta). \tag{6}$$

As a simple illustration of independence, consider the case of three model parameters. In general, one has $P(v_1, v_2, v_3) = P(v_1|v_2, v_3) \times P(v_2, v_3) = P(v_1|v_2, v_3) \times P(v_2|v_3) \times P(v_3)$. However, if v_1 is independent of v_2 and v_3, then $P(v_1|v_2, v_3) = P(v_1)$; and if v_2 is independent of v_3, then $P(v_2|v_3) = P(v_2)$. Hence, independence implies that $P(v_1, v_2, v_3) = P(v_1) \times P(v_2) \times P(v_3)$.

c. Uniform and normal priors

When there is little or no knowledge about a model parameter and/or one does not want to bias solution retrieval in any way, the *ignorance prior* is often assumed. The ignorance prior for the jth model parameter is that of a uniform distribution given by

$$P(v_j) = 1/(b_j - a_j) = c_j, \tag{7}$$

where the distribution is defined over the interval $I[a_j, b_j]$, and c_j is a jth constant.

In some case, however, one might have data or other information that suggests a particular model parameter is normally (or approximately normally) distributed. A normally distributed prior takes on the form

$$P(v_j) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{-1/2[(v_j-\mu_j)/\sigma_j]^2}, \tag{8}$$

where μ_j and σ_j are the population mean and standard deviation, respectively.

d. The MAP solution assuming independence and mixed uniform and normal priors

Suppose that model parameter independence is assumed. Also suppose that the first $j = 1, \ldots, \eta_u$ model parameters are assumed to follow a uniform distribution, and the remaining $\eta - \eta_u$ model parameters are assumed to follow a normal distribution. In this case, (6) reduces to

$$P(\mathbf{v}) = \left[\prod_{j=1}^{\eta_u} c_j\right]\left[\prod_{k=\eta_u+1}^{\eta} \frac{1}{\sigma_k\sqrt{2\pi}} e^{-1/2[(v_k-\mu_k)/\sigma_k]^2}\right]. \tag{9}$$

The probability density of drawing the vector **d** given the model parameters **v** is simply the product of the probability densities associated with drawing the ith component of **d** given **v**; that is,

$$P(\mathbf{d}|\mathbf{v}) = \prod_{i=1}^{m} p(d_i|\mathbf{v}). \tag{10}$$

Note that $p(d_i|\mathbf{v})$ is just a way of representing the mixture distribution $p(d)$ when the conditionality on **v** is explicitly given; see section 2 for more on the mixture distribution.

Substituting (9) and (10) into the last equation set given in (5) gives a nonlinear system of η equations in η model parameter unknowns, where **d** is regarded as a constant m vector, with $m \leq N$. It is normally either difficult or impossible to solve such a system analytically, so the MAP solution is typically found by using numerical optimization methods such as those provided in Press et al. (1992, chapter 10). Because the product of small numbers (probabilities) is quite small and often results in computational underflow errors, it is customary to take the natural logarithm of the function to be maximized. The natural logarithm removes the underflow errors by converting very small fractions into "regular range" negative numbers. Hence, in practice, the MAP solution can be found by maximizing the function S given by

$$\begin{aligned} S(\mathbf{v}) &\equiv \ln[P(\mathbf{d}|\mathbf{v})P(\mathbf{v})] = \ln P(\mathbf{d}|\mathbf{v}) + \ln P(\mathbf{v}) \\ &= \ln\prod_{i=1}^{m} p(d_i|\mathbf{v}) + \ln\left[\prod_{j=1}^{\eta_u} c_j \prod_{k=\eta_u+1}^{\eta} \frac{1}{\sigma_k\sqrt{2\pi}} e^{-1/2[(v_k-\mu_k)/\sigma_k]^2}\right] \\ \Rightarrow S(\mathbf{v}) &= \sum_{i=1}^{m} \ln[p(d_i|\mathbf{v})] + \sum_{j=1}^{\eta_u} \ln c_j + \sum_{k=\eta_u+1}^{\eta} \ln\left(\frac{1}{\sigma_k\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{k=\eta_u+1}^{\eta}\left(\frac{v_k-\mu_k}{\sigma_k}\right)^2. \end{aligned} \tag{11}$$

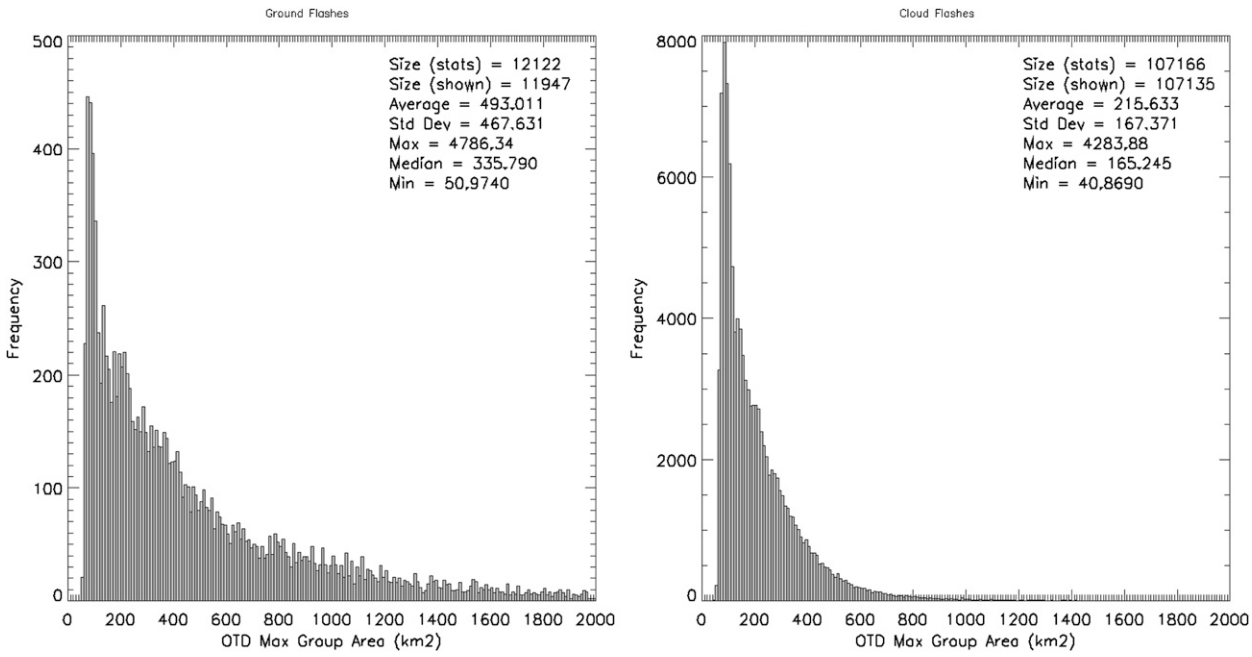


FIG. 1. Distribution of MGA for (left) ground flashes and (right) cloud flashes as found in Koshak (2010).

4. The mixed exponential distribution model

The preceding section provided a general framework for obtaining the MAP solution when model parameter independence is assumed, and when the model parameter prior distributions can be assumed to be either uniform distributions, normal distributions, or a combination of both. In this section, a particular form for the model mixture distribution $p(d_i|\mathbf{v})$ is chosen, and the prior distributions are specified.

The primary approach taken in this study is to thoroughly examine the known observations composing $p_g(x)$ and $p_c(x)$, and then build a physically meaningful model based on these observations. By virtue of the mathematical form of the model invoked, physical constraints are implicitly added to the retrieval process. These ''model constraints'' ultimately help remove solution ambiguity. [In addition, one should note that it is possible to avoid model constraints altogether by invoking the Central Limit Theorem of statistics (see the appendix for additional details).]

In Koshak (2010), two closely related optical characteristics were suggested as being valuable in the retrieval of α; one was the MNEG and the other was the MGA. Because the MNEG distribution contains only integer numbers that lead to less smooth distributions, the MGA optical characteristic is used here. Hence, in all that follows, note that x = MGA.

Figure 1 shows the distribution of x (i.e., MGA) for ground and cloud flashes, as provided in Koshak (2010). These data were derived from an analysis of 5 yr of OTD data. Note that the distribution of x ramps up quickly to a peak frequency value, and then gradually decays for larger values of x. The low end of the distribution gets complicated/truncated by the OTD pixel resolution, which varies across the OTD field of view (and is about 8 km at nadir). Rather than attempting to accommodate the low-end peculiarities using a complicated multiparameter model, it was better to simply focus on all $x \geq 64$ km^2. That is, a new random variable $y \equiv x - 64$ km^2 is used.

Figure 2 shows plots of the distribution of y for the OTD ground and cloud flashes. The red analytic curve in each plot is an exponential distribution of the form $(1/\overline{y})e^{-y/\overline{y}}$, where $\overline{y}$ takes on the data average shown in the upper-right corner of the plot. For a mixture of ground and cloud flashes, one can consider a superposition of exponential distributions. This gives, in general, the following mixed exponential distribution model:

$$p(y) = \begin{cases} 0, & y < 0 \\ \alpha p_g(y) + (1-\alpha)p_c(y) = \dfrac{\alpha}{\mu_g} e^{-y/\mu_g} + \dfrac{(1-\alpha)}{\mu_c} e^{-y/\mu_c}, & y \geq 0, \end{cases} \tag{12}$$

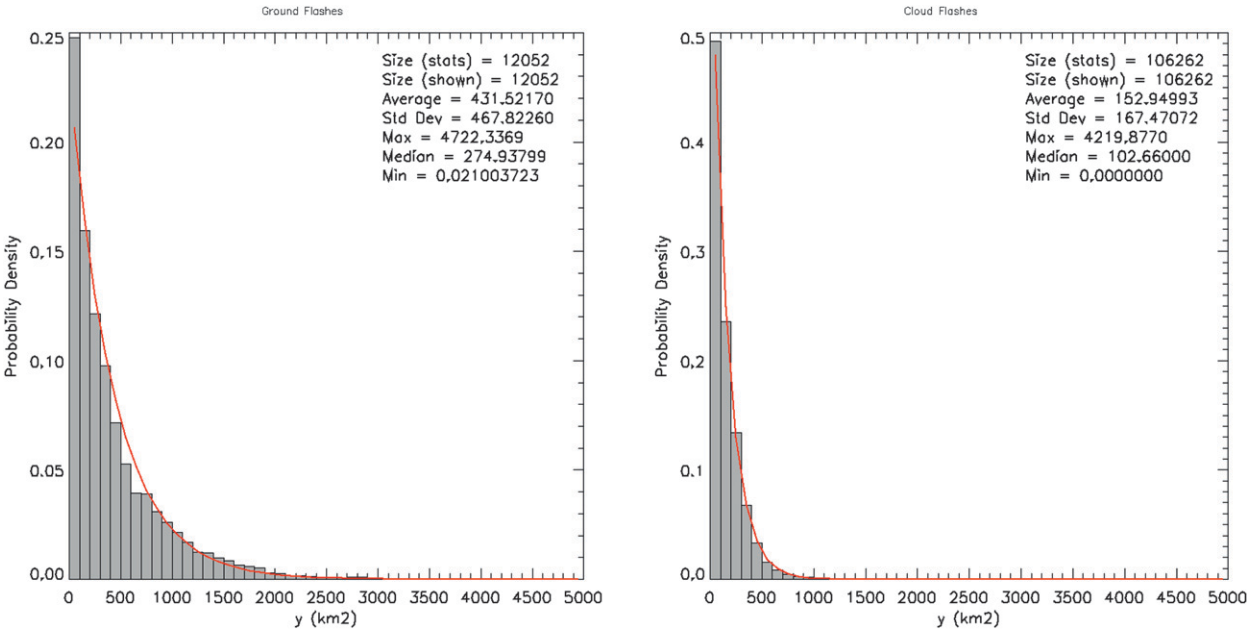


FIG. 2. The distribution of $y \equiv x - 64$ km^2 for the OTD (left) ground flashes and (right) cloud flashes. In each plot is shown an exponential distribution (red curve) having as a mean the OTD CONUS average of y, given in the upper-right hand corner of the plot. Note how well the exponential model fits the two datasets. The statistics shown are based only on values of $y \geq 0$.

where the population means are given by

$$\mu_g \equiv \int_0^\infty y p_g(y)\, dy, \quad \mu_c \equiv \int_0^\infty y p_c(y)\, dy. \tag{13}$$

The population mean and variance of the mixture distribution $p(y)$ are

$$\begin{aligned} \mu &\equiv \int_{-\infty}^{\infty} y p(y)\, dy = \alpha\mu_g + (1-\alpha)\mu_c, \\ \sigma^2 &\equiv \int_{-\infty}^{\infty} (y-\mu)^2 p(y)\, dy, \\ &= \alpha(2-\alpha)\mu_g^2 - 2\alpha(1-\alpha)\mu_g\mu_c + (1-\alpha^2)\mu_c^2. \end{aligned} \tag{14}$$

Mixed exponential distribution models are commonly used in a variety of disciplines [e.g., such models are employed in the medical literature to analyze patient "length of stay" in the hospital (Keatinge 1999)].

Note that the mixed exponential model in (12) has three model parameters, that is, $\mathbf{v} = (\alpha, \mu_g, \mu_c)$, and $p(y)$ could be rewritten in the more explicit fashion as that of $p(y|\mathbf{v})$. In addition, $\mathbf{d} = \mathbf{y}$; that is, $\mathbf{y}$ is derived from $\mathbf{x}$ and consists of all of the elements of $\mathbf{x}$ that are greater than or equal to 64 km^2 (i.e., the instrument nadir pixel ground footprint area). Next, assumptions are made about the forms of the model parameter prior distributions.

The prior distribution $P(\alpha)$ is assumed to be uniform. Because $0 \leq \alpha \leq 1$, this implies that $P(\alpha) = P(v_1) = 1/(b_1 - a_1) = 1/(1-0) = 1$. One could argue that lower values of α are more likely (i.e., cloud flashes generally out number ground flashes), but because the main interest of this study is to retrieve α, extra caution has been taken to avoid prebiasing retrieval results.

For $P(\mu_g)$ and $P(\mu_c)$ it is useful to consider the data in Table 3 of Koshak and Solakiewicz (2011), where the statistics of the spatial variability in MGA across CONUS are provided. The mean (and standard deviation) of the ground flash MGA over 52 widely distributed locations across CONUS was 488.7 km^2 (39.5 km^2); the cloud flash MGA values were 213.7 km^2 (16.8 km^2). These are only rough estimates because the statistics depend on how many flashes are analyzed in the vicinity of each of the 52 locations; nonetheless, the standard deviation values give a "ballpark" idea of the spatial variability. In addition, given that CONUS is a region of highly variable lightning, highly variable thunderstorm structures, and highly variable cloud morphologies with diverse cloud scattering properties, the spatial variability of MGA across the globe is probably not too much larger than the CONUS standard deviation values indicated here. In addition, the mean y values shown in Fig. 2 are likely to be reasonably close in value to the respective (μ_g, μ_c) values for a particular geographic region over the globe, plus or minus some random variability. Taking all of this into account, it is assumed that $P(\mu_g)$ is normally distributed with a mean given by the CONUS mean (i.e., 431.521 70 km^2 from the left-hand distribution in Fig. 2) and with a standard deviation given by 50 km^2

(i.e., something on the order of, or a bit larger than, the spatial variability of y across CONUS given by the values 39.5 km^2 and 16.8 km^2 mentioned above). Similarly, it is assumed that $P(\mu_c)$ is normally distributed with a mean given by the CONUS mean (i.e., 152.949 93 km^2 from the right hand distribution in Fig. 2) and again a standard deviation of 50 km^2. [Note that the standard deviation of MGA across CONUS is identical to the standard deviation of y across CONUS since y is just shifted in magnitude; i.e., y = MGA – 64 km^2.]

With these assumptions imposed on the prior distributions, and utilizing the mixed exponential model in (12), the function S in (11) simplifies to

$$S(\mathbf{v}) = \sum_{i=1}^{m} \ln\left[\frac{\alpha}{\mu_g} e^{-y_i/\mu_g} + \frac{(1-\alpha)}{\mu_c} e^{-y_i/\mu_c}\right] + \ln\left(\frac{1}{\sigma_2\sqrt{2\pi}}\right) + \ln\left(\frac{1}{\sigma_3\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{\mu_g - \mu_2}{\sigma_2}\right)^2 - \frac{1}{2}\left(\frac{\mu_c - \mu_3}{\sigma_3}\right)^2, \quad (15)$$

where $\mu_2 \pm \sigma_2 = 431.521\ 70\ \text{km}^2 \pm 50\ \text{km}^2$, and $\mu_3 \pm \sigma_3 = 152.949\ 93\ \text{km}^2 \pm 50\ \text{km}^2$.

To obtain the MAP solution, that is, the value of $\mathbf{v}$ that maximizes S in (15), the Broyden–Fletcher–Goldfarb–Shannon variant of the Davidon–Fletcher–Powell (DFP) algorithm is used to find the minimum of $-S$. (see chapter 10 of Press et al. 1992 for details on numerical optimization).

5. Label switching

Note that if all of the model parameter prior distributions are defined as uniform distributions, then $S(\mathbf{v})$ depends only on the first sum in (15) plus a constant. This makes the retrieval process highly susceptible to solution nonuniqueness because of the so-called "label switching" problem (Redner and Walker 1984; Diebolt and Robert 1994; Richardson and Green 1997; Celeux 1998; Stephens 2000; Jasra et al. 2005). Label switching can be clearly defined by considering the following transformation:

$$\tilde{\alpha} = 1 - \alpha, \quad \tilde{\mu}_g = \mu_c, \quad \tilde{\mu}_c = \mu_g. \quad (16)$$

For $y > 0$, the mixture distribution is

$$\begin{aligned}\tilde{p}(y) &= \frac{\tilde{\alpha}}{\tilde{\mu}_g} e^{-y/\tilde{\mu}_g} + \frac{(1-\tilde{\alpha})}{\tilde{\mu}_c} e^{-y/\tilde{\mu}_c}, \\ &= \frac{(1-\alpha)}{\mu_c} e^{-y/\mu_c} + \frac{[1-(1-\alpha)]}{\mu_g} e^{-y/\mu_g}, \\ &= \frac{\alpha}{\mu_g} e^{-y/\mu_g} + \frac{(1-\alpha)}{\mu_c} e^{-y/\mu_c} \equiv p(y). \end{aligned} \quad (17)$$

In other words, the mixture distribution is invariant under the transformation given in (16). This means that a mixture with the properties $\mathbf{v} = (\alpha, \mu_g, \mu_c)$ is indistinguishable from a mixture with the properties $\mathbf{v} = (1-\alpha, \mu_c, \mu_g)$. This is a fundamental ambiguity related to the fact that one can always suppose that the ground flashes are really all cloud flashes, and that the cloud flashes are really all ground flashes; hence, the population means interchange and the ground flash fraction α converts to its complement $1-\alpha$.

However, because the MGA distributions for the conterminous United States provided in Koshak (2010) have the property $\mu_g > \mu_c$, it is reasonable to consider only those solutions where $\mu_g > \mu_c$ holds. This approach for helping to mitigate the effects of label switching is called an *identifiability constraint* (Stephens 2000).

Even after invoking the identifiability constraint, label switching is still a problem when the true value of α is near (or at) zero or unity. For example, when $\alpha = 0$, μ_g is undefined so that the retrieved value of μ_g is meaningless and therefore not comparable with μ_c. When $\alpha = 1$, the converse is true; that is, μ_c is undefined so that the retrieved value of μ_c is meaningless and therefore not comparable with μ_g.

Nonetheless, one could still try comparing the meaningful μ_g value (case $\alpha = 1$) or the meaningful μ_c value (case $\alpha = 0$) with the CONUS means in Fig. 2. For example, if one assumes that the case $\alpha = 0$ is correct, then one would expect the retrieved μ_c value to be reasonably close to the cloud flash CONUS mean of 152.949 93 km^2 given in Fig. 2. If, instead, μ_c was closer to the ground flash CONUS mean of 431.521 70 km^2, then one would suspect that the case $\alpha = 1$ is actually the truth. Similarly, if one assumes that the case $\alpha = 1$ is correct, then one would expect the retrieved μ_g to be reasonably close to the ground flash CONUS mean of 431.521 70 km^2 given in Fig. 2. If, instead, μ_g was closer to the cloud flash CONUS mean of 152.949 93 km^2, then one would suspect that the case $\alpha = 0$ is actually the truth. However, the author has found that such comparisons still are not sufficiently effective in removing solution ambiguity.

Fortunately, assuming that the prior distributions $P(\mu_g)$ and $P(\mu_c)$ are normally distributed helps to mitigate the label-switching problem because the symmetry in S that leads to the label-switching problem is broken [see section 8b, which quantifies the benefits of invoking the normal priors]. Note that the means of these normal distributions (i.e., $\mu_2 = 431.521\ 70$ km^2, and $\mu_3 = 152.949\ 93$ km^2) given in the previous section are consistent with the identifiability constraint because $\mu_2 > \mu_3$.

6. Initializing the search using constraints based on population statistics

Recall that the DFP minimization algorithm mentioned in section 4 is used to find the minimum of $-S$, where $S = S(\mathbf{v}) = S(\alpha, \mu_g, \mu_c)$ is given in (15). The DFP algorithm requires an initialization (i.e., ''first guess'') of $\mathbf{v}$. A good first guess improves the chance of getting successful convergence to the global minimum of the function. In this section, constraints that are based on population statistics are introduced, and are shown to be very useful in obtaining good initializations of $\mathbf{v}$. These constraints also clarify what is (mathematically) involved with picking solutions that have $\mu_g > \mu_c$.

For m large, the sample mean and sample variance provide good estimates of the associated populations statistics in (14); that is,

$$\begin{aligned} \overline{y} &\cong \mu = \alpha\mu_g + (1-\alpha)\mu_c, \\ s^2 &\cong \sigma^2 = \alpha(2-\alpha)\mu_g^2 - 2\alpha(1-\alpha)\mu_g\mu_c + (1-\alpha^2)\mu_c^2, \end{aligned} \tag{18}$$

where the sample mean and variance are given by

$$\overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i, \quad s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \overline{y})^2. \tag{19}$$

For a fixed (α, μ, σ^2), and viewing (μ_g, μ_c) as coordinate axes, the first equation in (18) is an equation of a line and the second equation in (18) is the equation of an ellipse that is not in standard form; that is, it is a rotated ellipse. The equation of the line allows one to eliminate μ_c as follows:

$$\mu_c = A + B\mu_g, \quad A = \overline{y}/(1-\alpha), \quad B = \alpha/(\alpha - 1). \tag{20}$$

Substituting this expression for μ_c into the variance equation in (18) gives a result that is quadratic in μ_g with solution

$$\begin{aligned} \mu_g &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \\ a &= c_1 + c_2 B + c_3 B^2, \\ b &= c_2 A + 2c_3 AB, \quad c = c_3 A^2 - s^2, \\ c_1 &= \alpha(2-\alpha), \quad c_2 = -2\alpha(1-\alpha), \quad c_3 = 1 - \alpha^2. \end{aligned} \tag{21}$$

Because $a = a(\alpha)$, $b = b(\alpha; \mu)$, and $c = c(\alpha; \mu, \sigma^2)$, note that one obtains the following functional dependencies: $\mu_g = \mu_g(\alpha; \mu, \sigma^2)$ and $\mu_c = \mu_c(\alpha; \mu, \sigma^2)$. Simplifying the positive root of (21) and substituting it into the first equation of (20) gives the positive root solutions as

$$\begin{aligned} \mu_g &= \mu + \sqrt{\frac{1}{2}\left(\frac{1-\alpha}{\alpha}\right)(\sigma^2 - \mu^2)}, \\ \mu_c &= \mu - \sqrt{\frac{1}{2}\left(\frac{\alpha}{1-\alpha}\right)(\sigma^2 - \mu^2)}. \end{aligned} \tag{22}$$

By similar manipulations, the negative root solutions are

$$\begin{aligned} \mu_g &= \mu - \sqrt{\frac{1}{2}\left(\frac{1-\alpha}{\alpha}\right)(\sigma^2 - \mu^2)}, \\ \mu_c &= \mu + \sqrt{\frac{1}{2}\left(\frac{\alpha}{1-\alpha}\right)(\sigma^2 - \mu^2)}. \end{aligned} \tag{23}$$

In addition, the results in (18) imply that

$$\sigma^2 - \mu^2 = 2\alpha(1-\alpha)(\mu_g - \mu_c)^2 \geq 0. \tag{24}$$

Hence, the expressions under the radicals in (22) and (23) are nonnegative. This means that the positive root solution in (22) necessarily results in the condition $\mu_g \geq \mu_c$, whereas the negative root solution results in the condition $\mu_g \leq \mu_c$. In other words, because this study only picks solutions with $\mu_g > \mu_c$, this means that only the positive root solution is considered; that is, this study uses (22), not (23). This is not an arbitrary strategy because all of the analyses of CONUS data provided by Koshak (2010) have in fact suggested that $\mu_g > \mu_c$. (As discussed in section 5, one is initially pressed with the issue of picking solutions either of type $\mu_g > \mu_c$ or $\mu_g < \mu_c$ in order to address the label-switching issue in a consistent way.)

Moreover, the result in (22) is used in this study to help initialize $\mathbf{v}$. Using the estimates in (18), and making the initialization $\alpha_{\text{initial}} = 0.5$, (22) gives the initializations

$$\begin{aligned} (\mu_g)_{\text{initial}} &= \overline{y} + \sqrt{\frac{1}{2}(s^2 - \overline{y}^2)}, \\ (\mu_c)_{\text{initial}} &= \overline{y} - \sqrt{\frac{1}{2}(s^2 - \overline{y}^2)}. \end{aligned} \tag{25}$$

Note that this is usually an excellent initialization because the estimates in (18) are excellent when the sample size N (i.e., the number of flashes observed) is large.

7. Aggressive application of population statistic constraints (reduced search spaces)

For completeness, it is worth mentioning that it is possible to more aggressively apply the population statistics provided in (18) when obtaining a solution. For example, one can substitute (22) into (15) to eliminate (μ_g, μ_c) altogether so that the function S is converted

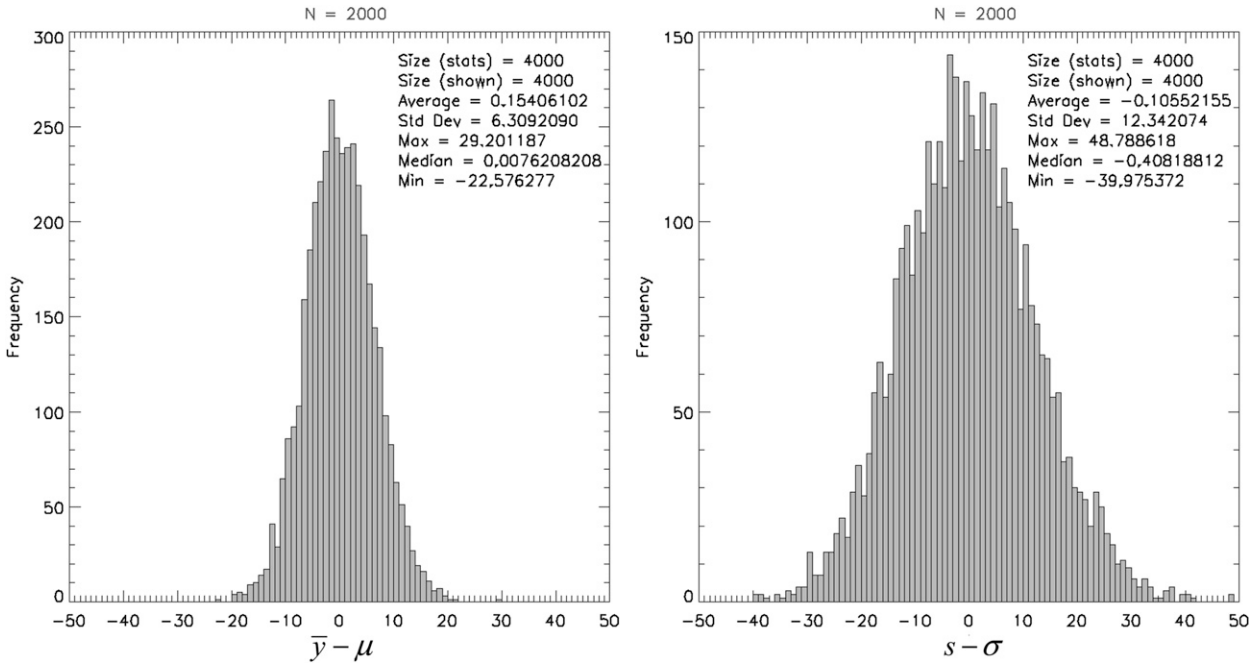


FIG. 3. The distribution of (left) $\bar{y} - \mu$ and (right) $s - \sigma$ for the case $N = 2000$. Because the difference between the sample and population value is reasonably small, the initialization methodology introduced in (22) is highly useful.

into a function of just one variable, that is, $S(\mathbf{v}) \rightarrow S(\alpha)$. In this case, $(s^2 - \bar{y}^2)$ is used to estimate $(\sigma^2 - \mu^2)$ with some estimation error ε. This means that one would just need to scan through the values of α between 0 and 1 to find the maximum of $S(\alpha)$; that is, the optimum ground flash fraction would just be the value of α at which the $S(\alpha)$ curve has its maximum. This in fact does work for values of α between about 0.2 and 0.8. However, note in (22) that when one scans for values of α near zero the value of the factor $(1 - \alpha)/\alpha$ approaches infinity, and when one scans for values of α near unity the factor $\alpha/(1 - \alpha)$ approaches infinity. This results in unreasonable error magnification of the estimation error ε, and hence a poor retrieval results when the true ground flash fraction is near zero or unity.

Another approach is to use the first equation in (18) to eliminate α; that is, the expression $\alpha \cong (\bar{y} - \mu_c)/(\mu_g - \mu_c)$ can be substituted into (15) in order to convert $S(\mathbf{v}) \rightarrow S(\mu_g, \mu_c)$. This conversion reduces down the dimensionality of the problem, which has practical benefits (smaller search space, less CPU time to obtain a solution, and a more constrained solution), but this approximation of α is nonoptimum if the approximation $\bar{y} \cong \mu$ happens to be poor.

The author has experimented with the two above retrieval methods, and finds that, although they have some beneficial features, a numerical maximization of the full three-dimensional function $S(\mathbf{v})$ given in (15) is preferred. The main reason for this decision is that the basic objective of finding the $\mathbf{v}$ that maximizes $P(\mathbf{v}|\mathbf{y})$ is not compromised in any way.

8. Numerical tests

a. Dependence of retrieval error on sample size

To begin, a simple test is performed that illustrates the effect of finite sampling on retrieval errors. A known mixed exponential distribution is assumed; that is, the population means are fixed at $\mu_g = 400$ km^2 and $\mu_c = 170$ km^2, and the ground flash fraction is fixed at $\alpha = 0.3$. The three parameters $\mathbf{v} = (\alpha, \mu_g, \mu_c)$ fully define the mixture distribution $p(y)$. Note that the values of μ_g and μ_c chosen here are physically reasonable; that is, the CONUS y means based on the 5-yr OTD data analysis of Koshak (2010) are $\mu_g = 431.521\ 70$ km^2 and $\mu_c = 152.949\ 93$ km^2. Next, the known mixture distribution is used to randomly generate a set of $m = N$ simulated measurements. These simulated measurements are inverted using the Bayesian inference methodology, and the retrieval error in each of the three parameters (α, μ_g, μ_c) is computed and stored. Another set of N measurements are randomly generated from the mixture distribution, and again the retrieval errors are computed and stored. A total of 4000 simulated retrievals were performed in this way.

To appreciate the advantages of employing (22) in the initialization of $\mathbf{v}$, Fig. 3 shows the distribution of the

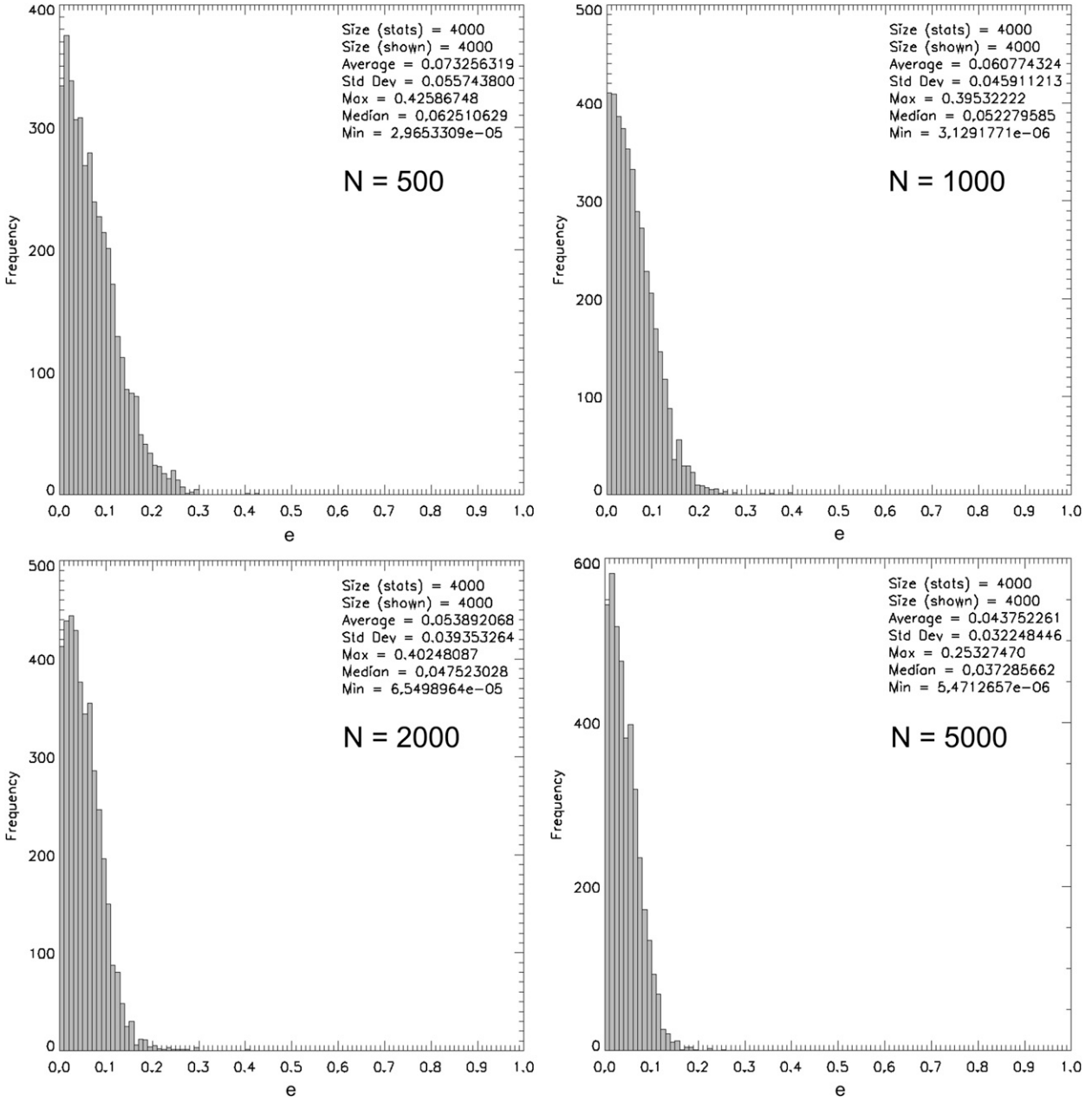


FIG. 4. The distribution of ground flash fraction retrieval errors $e = |\alpha_r - \alpha|$ as a function of the number of flashes N in the mixture. Retrieval errors decrease as N increases.

deviation of the sample mean from the population mean, that is, $\bar{y} - \mu$ and the distribution of the deviation between the sample standard deviation and the population standard deviation $s - \sigma$. Here, the sample size $N = 2000$. As expected, the values of $\bar{y}$ are close to μ and the values of s are close to σ. Hence, the approximations in (18) are quite reasonable for the initialization provided in (22).

Figure 4 shows the statistics of the retrieval errors for the cases with $N = 500$, 1000, 2000, and 5000 measurements. The retrieval error is given by $e = |\alpha_r - \alpha|$, where α_r is the retrieved ground flash fraction and α is the true ground flash fraction (i.e., 0.3). As expected, the retrieval errors decrease as N increases because the estimates in (22) improve, and also because more information is used in the Bayesian retrieval method. Tests of this type were also performed for the case $\alpha = 0.7$, and similar results were found.

Average errors in μ_g were (25.5, 24.3, 23.6, and 20.8 km^2) for the respective values ($N = 500, 1000, 2000,$

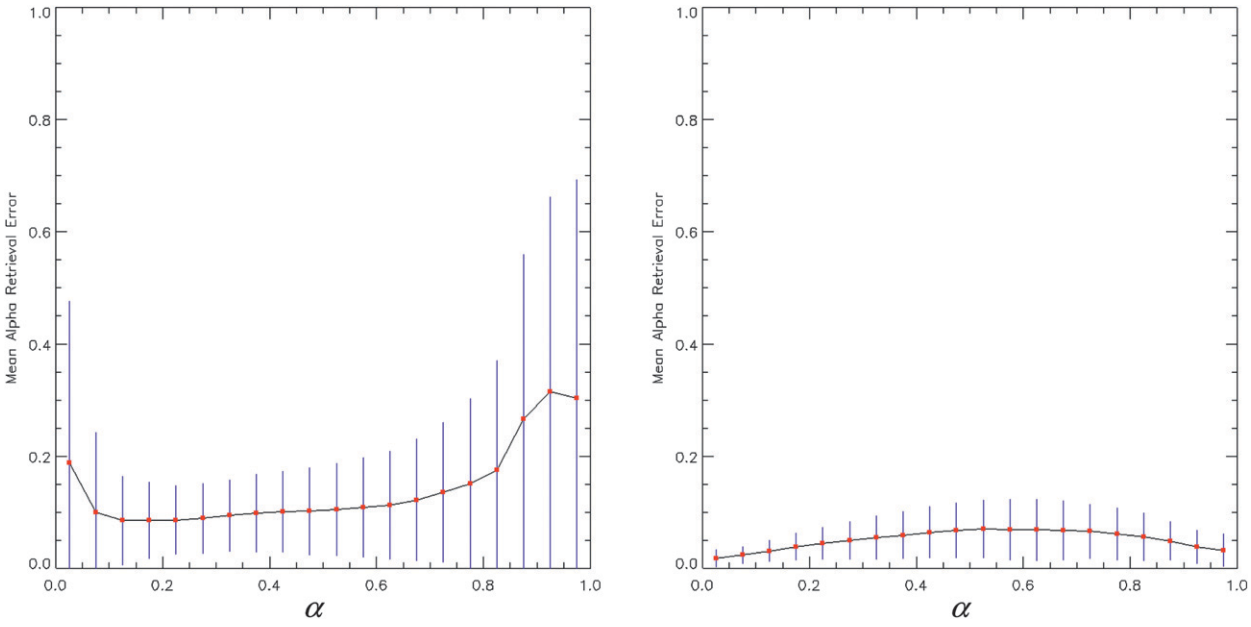


FIG. 5. Average ground flash fraction retrieval error as a function of the known ground flash fraction when (left) uniform priors are assumed for $P(\mu_g)$ and $P(\mu_c)$, and (right) normal priors are assumed for $P(\mu_g)$ and $P(\mu_c)$. The vertical lines indicate the standard deviation about the average errors shown. Use of the normal priors substantially reduces errors.

and 5000), and average errors in μ_c were (14.7, 11.7, 9.5, and 7.0 km^2). All of these errors are a reasonably small percentage of the known values (μ_g = 400 km^2 and μ_c = 170 km^2).

b. *Dependence of retrieval error on α and the prior distributions*

Next, it is illustrated how the mean retrieval error in α varies as a function of α, and on the type of prior distributions used. With the same fixed values (μ_g = 400 km^2, μ_c = 170 km^2), 100 simulated retrievals were performed in each α bin of width 0.05. That is, 100 known values of α were randomly chosen from the first bin (α, 0.0–0.05), 100 known values from the second bin (α, 0.05–0.1), and so on up to the last bin (α, 0.95–1.0). These tests provide the mean retrieval errors provided in Fig. 5. The vertical lines indicate the standard deviations about the mean error values. As always, a uniform distribution is used for the prior $P(\alpha)$. In the left-hand plot of Fig. 5, uniform distributions are used for the priors $P(\mu_g)$ and $P(\mu_c)$, but in the right-hand plot of Fig. 5 the usual normal priors are used. The large errors that occur for values of α near zero or unity in the left-hand plot (here called "tail errors") are a consequence of the label-switching ambiguity discussed in section 5. Note that the use of the normal priors nicely removes these large errors. Similar improvements have been found for simulations that employed other known values of (μ_g, μ_c).

c. *Full error analysis*

In addition to the sample size N, retrieval errors will depend on the true values of all three variables (α, μ_g, μ_c) because these define the specific form of the mixture distribution $p(y)$. Therefore, thorough simulated retrievals were run across the full range of each of these three variables. Based on the typical variability in μ_g and μ_c found across CONUS in Koshak and Solakiewicz (2011, their Table 3), the following ranges were chosen: μ_g = 256–608 km^2 and μ_c = 98–208 km^2. The range on α was of course 0–1. As before, random draws from the known mixture distribution $p(y)$ are used to generate N simulated measurements **y**.

Figure 6 shows the mean retrieval error in α across these full ranges; the sample size in this simulation is N = 2000. In each cell shown, 100 retrievals were performed; that is, 100 different simulated measurement vectors **y** were mathematically inverted using the Bayesian methodology for each cell. For example, in Fig. 6a, a particular cell is defined by the ranges α = 0–0.1, μ_g = 320–352 km^2, and μ_c = 108–118 km^2. This means that the known values of (α, μ_g, μ_c) were randomly selected from these respective ranges and the resulting mean retrieval error (and standard deviation) in α was 0.022 (0.015), as given in the cell.

To help interpret the plan view of the errors, each cell is color coded according to its mean error in α; low-level errors are shades of green, midlevel errors are shades of

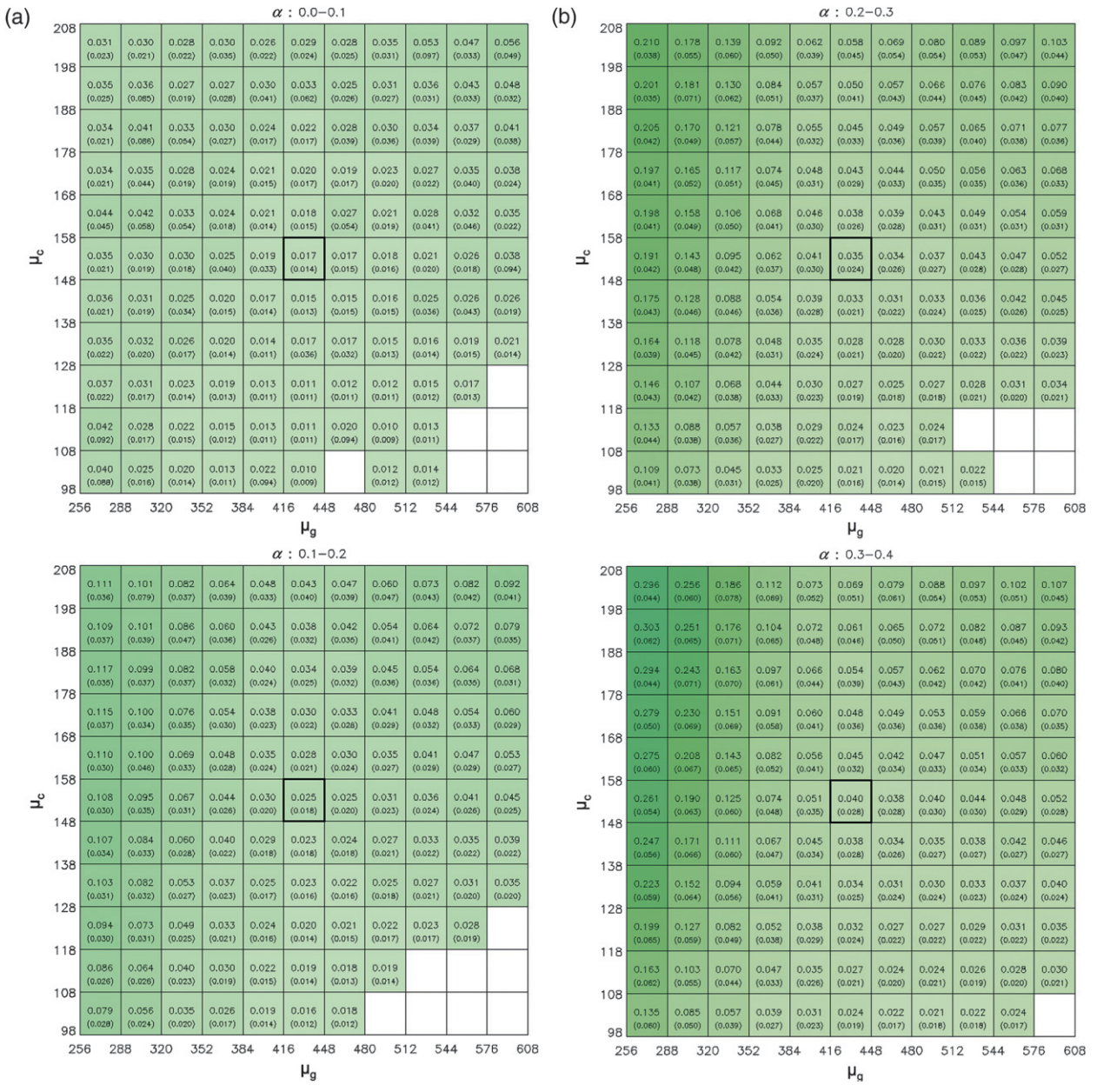


FIG. 6. The mean retrieval error (standard deviation) in α as a function of the indicated ranges of the model parameters; these plots cover the range (a) $\alpha = 0-0.2$, (b) $\alpha = 0.2$–0.4, $\alpha = 0.4$–0.6, (d) $\alpha = 0.6$–0.8, and (e) $\alpha = 0.8$–1. Note: low, midlevel, and high retrieval errors are in shades of green, blue, and red, respectively.

blue, and high-level errors are shades of red. Blank (white) cells represent regions where a mean value could not be computed over all 100 retrievals because one or more retrievals of the 100 had nonconvergent solutions; that is, the Bayesian method did not provide a physically meaningful answer because of poor convergence to the minimum of $-S$. However, these cases were rare and the mean errors for the remaining good retrievals in the affected cell were generally similar to the mean errors in adjacent cells.

Because the actual value of α over the globe is expected to be on the low end ($0 \leq \alpha \leq 0.4$), it is encouraging to see that the Bayesian retrieval method performs well in this range. Excessive retrieval errors are only found for large values of α when μ_g is under about 320 km^2. The central cell, which has a square box drawn around it for emphasis, contains the CONUS mean values (μ_g = 431.521 70 km^2 and μ_c = 152.949 93 km^2). Because the normal priors employed use these CONUS values as population means, retrieval errors are small in this cell as expected.

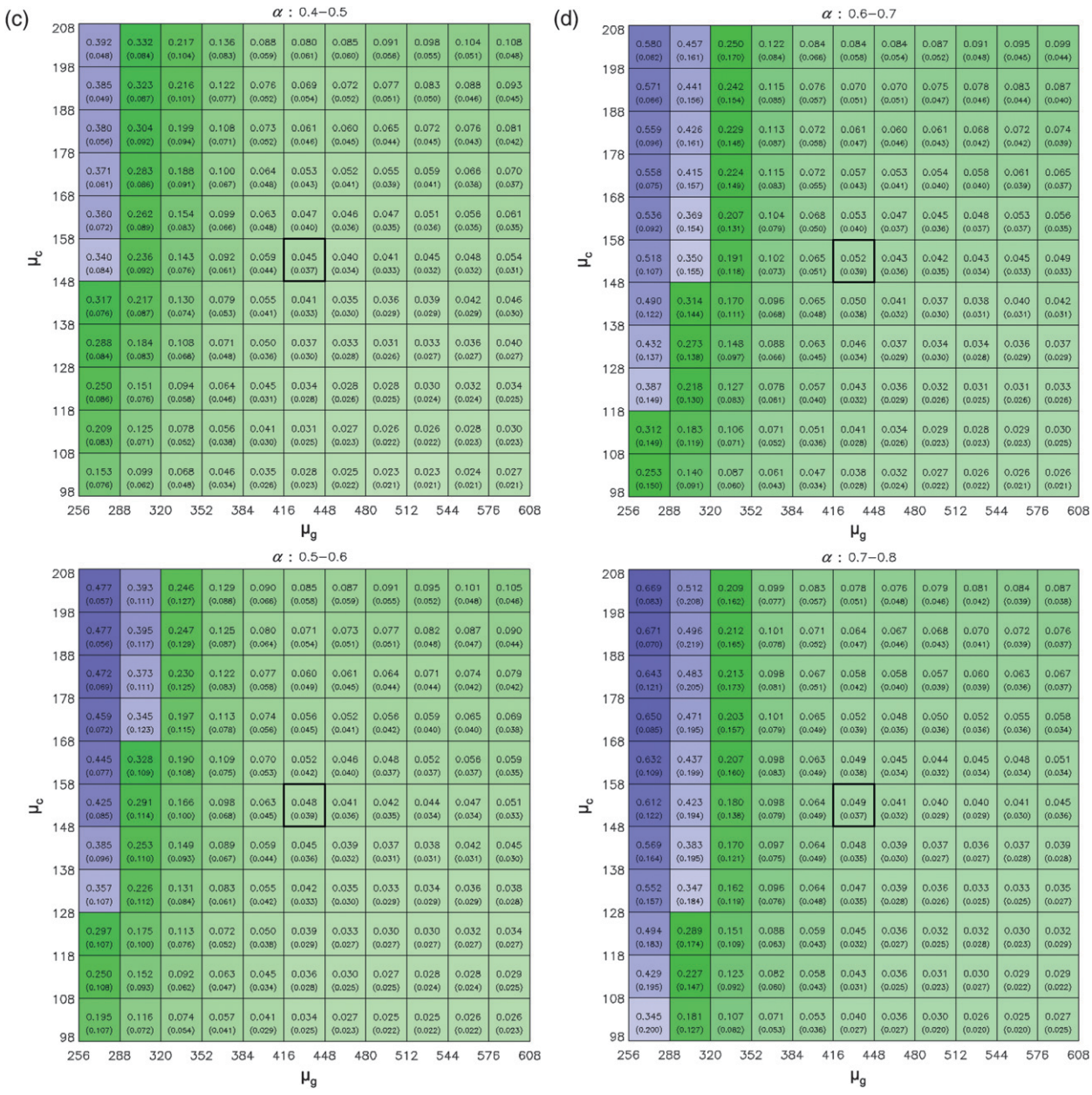


FIG. 6. (*Continued*)

It is important to note, as emphasized previously, that retrieval errors tend to decrease as the sample size of flashes N increases. Hence, the same analysis was run again, but with $N = 5000$. As expected, the retrieval errors generally decreased. For example, the central cell had an average reduction in ground flash retrieval error of 28.6%.

Finally, the retrievals of μ_g and μ_c presented here represent an improvement over Koshak and Solakiewicz (2011) wherein CONUS estimates of these two parameters were used; that is, the Bayesian retrieval scheme provides a comprehensive methodology for obtaining the most likely parameters (α, μ_g, μ_c) for an arbitrary region of the globe and therefore does not depend on estimates derived from CONUS observations.

9. Preliminary application on a global scale

As a preliminary application of the Bayesian retrieval method using the mixed exponential distribution model, the entire 5-yr OTD dataset has been analyzed to estimate the ground flash fraction on a global scale. The

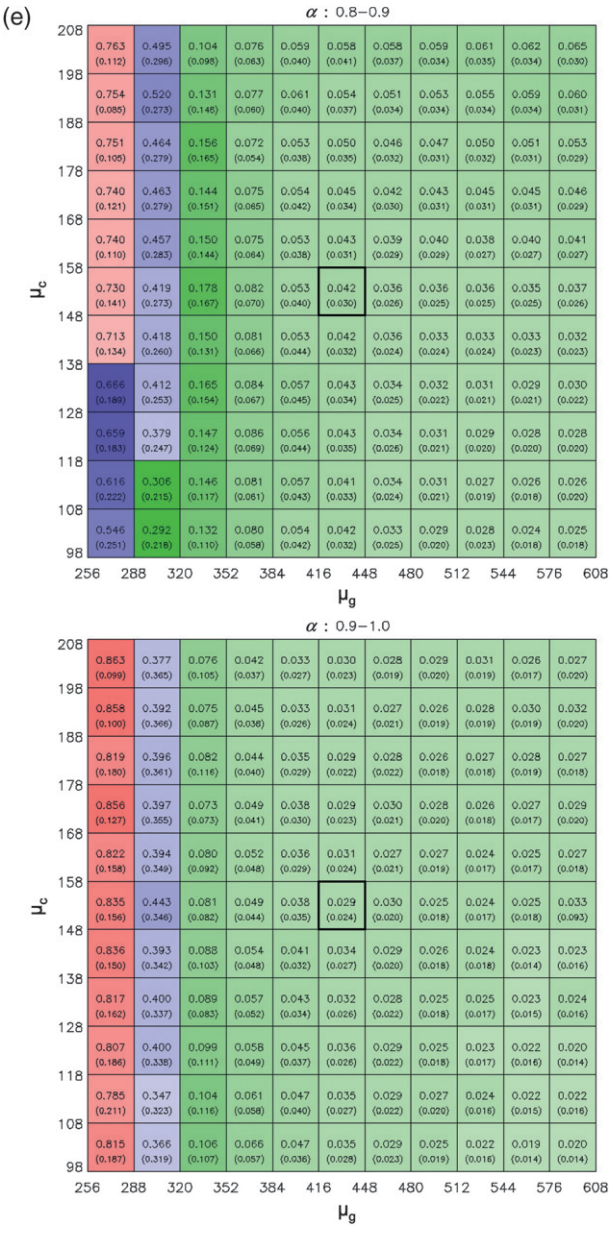


FIG. 6. (*Continued*)

globe is divided into 4° × 4° latitude–longitude bins, and no retrieval is performed if a bin contains less than 2000 flashes. The retrieval results are provided in Fig. 7.

Figure 7a provides the number of flashes in each latitude/longitude bin. Note that the number of flashes in central Africa is quite large (i.e., the global maximum number of flashes in a bin occurred in this region, and had a value of 32 173 flashes). Hence, based on the simulation results in section 8, which showed that retrieval errors decrease with increasing sample size N, ground flash fraction retrievals in central Africa are expected to be quite good.

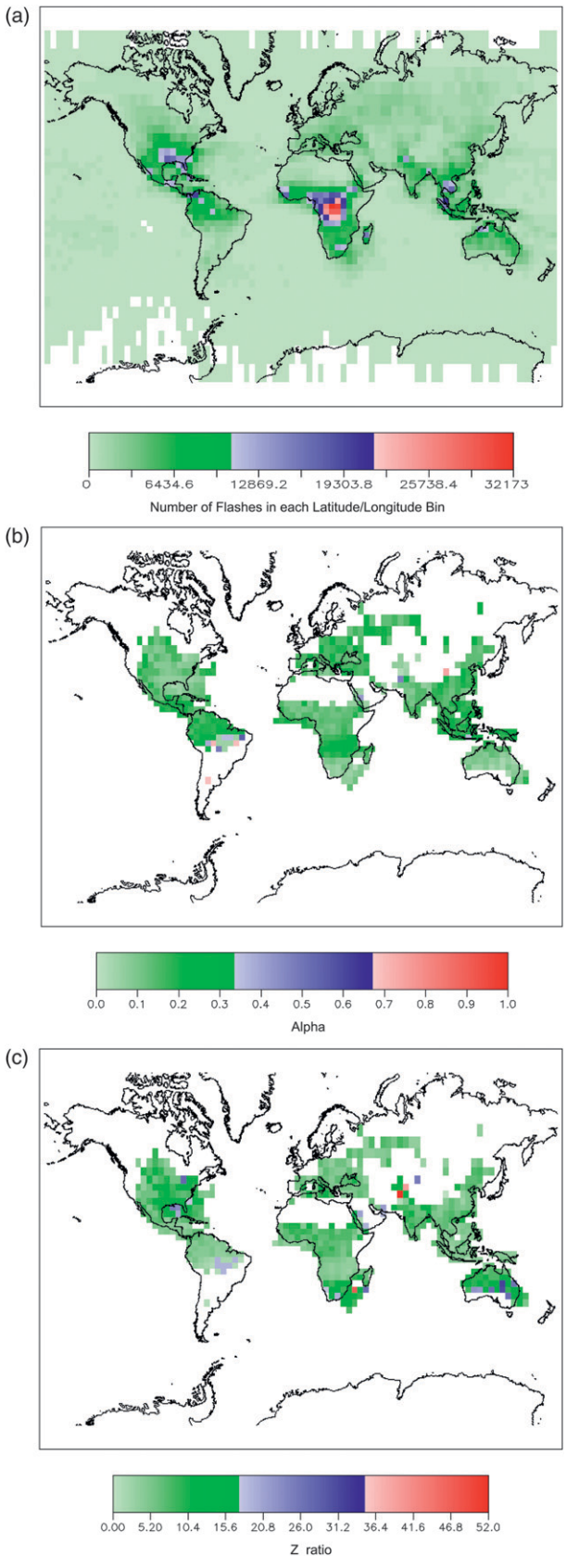
Figure 7b provides the spatial distribution of the retrieved ground flash fraction, and Fig. 7c provides the associated Z ratio [number of cloud flashes/number of ground flashes = $(1 - \alpha)/\alpha$]. One must be cautious in making comparisons of these values to others in the literature. For example, these results do not strictly account for the effects of the diurnal cycle; that is, the OTD data were not averaged over a 55-day cycle, as was done in Boccippio et al. (2001). The results shown are just a Bayesian analysis of all of the flashes that happen to occur in the particular latitude–longitude bin, no matter what time of day these flashes may have occurred. Also, the results shown are obviously not at the storm scale. However, as long as the flash sample size is maintained (i.e., roughly 2000 or more flashes, depending on the retrieval accuracy desired) the Bayesian retrieval method is fully applicable when both the space and time scales are made smaller and shorter, respectively.

For this global-scale analysis, the 4° bin size was used so that not too many bins across the globe would fall below the $N = 2000$ flash threshold. In fact, with the 4° bin size, there were a total of 562 bins across the globe that met this threshold. The spatial average (and standard deviation) of α across the domain analyzed was 0.151 (0.081) with a median of 0.140, and the range was 0.019–0.739. For Z, the spatial average (standard deviation) was 7.58 (5.45) with a median of 6.14, and a range of 0.353–51.6. Finally, note algebraically that $\overline{Z} \neq (1 - \overline{\alpha})/\overline{\alpha}$, in general.

10. Summary

This paper has introduced a Bayesian inversion technique (section 3) for retrieving the optimum ground flash fraction and other statistical parameters of a set of N flashes observed by a satellite lightning imager (such as OTD, LIS, or the future GLM). The method is based on describing the observed flashes as a mixture of two distributions of a particular flash optical characteristic (section 2); one distribution describes the ground flashes and the other distribution describes the cloud flashes.

The flash optical characteristic analyzed by the Bayesian retrieval algorithm is the maximum group area (MGA); selection of this variable was motivated by Koshak (2010), who identified MGA as a useful "return stroke" detector for a large sampling of flashes. Using 5 yr of CONUS OTD flashes, it was shown that when the MGA data for ground flashes is slightly shifted (i.e., decremented by the instrument nadir footprint, ~64 km^2 for OTD), the resulting "shifted MGA" variable is exponentially distributed. Similarly, the shifted MGA data for cloud flashes is also exponentially distributed. However, the sample means of each of these exponential distributions were



shown here to be significantly distinct (Fig. 2). The best estimates of the CONUS population means were found to be $\mu_g = 431.521\ 70\ \text{km}^2$ and $\mu_c = 152.949\ 93\ \text{km}^2$, respectively.

Because the shifted MGA data for ground and cloud flashes are each exponentially distributed, with statistically distinct means, a mixed exponential distribution model (section 4) has been introduced to describe the statistical distribution of an arbitrary mixture of ground and cloud flashes. In total, the mixed exponential distribution model contains the following three model parameters: 1) the ground flash fraction α, 2) the population mean μ_g of the shifted MGA distribution for ground flashes, and 3) the population mean μ_c of the shifted MGA distribution for cloud flashes.

Given a set of N shifted MGA observations from an arbitrary geographical region over the globe, the Bayesian retrieval algorithm finds the optimum values of (α, μ_g, μ_c) for that region. From exhaustive numerical simulations (section 8), it is shown that reasonable retrieval accuracy is obtained if N is about 2000 or larger, and that improved retrieval accuracy can be obtained by increasing N. [For GLM application, one could consider, for example, 10 thunderclouds in a region, with each producing a nominal flash rate of 10 flashes per minute. This means that one would only need to wait 20 min (=2000 flashes per 100 flashes per minute) before applying the Bayesian algorithm.]

The importance of applying normally distributed prior distributions within the Bayesian framework as a way of minimizing the deleterious effects of the infamous label-switching problem (section 5) is demonstrated; the normal priors effectively help to minimize retrieval errors when the true value of the ground flash fraction is either near zero or unity (Fig. 5). In addition, because the Bayesian algorithm relies on a numerical maximization of a log-likelihood function to obtain the optimum model parameters, it is important to obtain good initial estimates of the model parameters to improve the chances that the parameter search will converge to the global maximum. To serve this purpose, a highly useful initialization scheme that is based on fundamental population statistic constraints is introduced and applied (see section 6).

←

FIG. 7. (a) The number of flashes in each 4° × 4° latitude–longitude bin. White areas indicate zero flashes. (b) The retrieval of the ground flash fraction α on a global scale using the 5-yr OTD dataset. Only 4° × 4° bins with 2000 or more flashes are analyzed. (c) The Z ratio associated with the retrieved ground flash fraction shown in (b).

The Bayesian retrieval algorithm (using the mixed exponential distribution model) has been applied to obtain a first estimate of the global distribution of the ground flash fraction (Fig. 7b), as well as a closely related variable, the Z ratio (number cloud flashes/number of ground flashes). As expected, the spatially averaged ground flash fraction over the analyzed global-scale domain was small (0.151), reflecting the fact that cloud flashes normally outnumber ground flashes. Some retrieved values of the ground flash fraction are notably large in a few geographic locations, and this merits detailed follow-on studies to determine the specific cause(s). Owing to the coarse (4°) spatial resolution, the transient nature of OTD view times, and biases resulting from diurnal variability in lightning, care must be taken in intercomparing these global retrievals with results from other studies. In any case, the Bayesian technique is quite versatile, and the present limitations stem more from observational limitations rather than retrieval algorithm limitations.

Moreover, the global retrievals presented here represent an important first step in the task of partitioning the global OTD/LIS total lightning climatology into separate global ground and cloud flash distributions. Because the basic physical characteristics [peak current, multiplicity (number of strokes per flash), channel altitude, channel length, and presence of continuing currents] of ground and cloud flashes differ, they produce, in general, different amounts of nitrogen oxides ($NO_x = NO_2 + NO$). This has been explicitly shown in a recent detailed study by Koshak et al. (2010) that employed north Alabama Lightning Mapping Array (LMA) data and the National Aeronautics and Space Administration (NASA) Marshall Space Flight Center (MSFC) Lightning Nitrogen Oxides Model (LNOM). Therefore, obtaining separate ground and cloud flash distributions will allow one to make better estimates of global lightning NO_x production for chemistry/climate model predictions. Similarly, the partitioning of flashes at the regional level will directly assist in making better estimates of regional lightning NO_x production for regional air quality models.

Note that a recent study by Ott et al. (2010) suggested NO_x production from ground and cloud flashes might be similar; it also identified other studies that suggested the same. However, it is important to note that there is still healthy debate on the issue, and further studies are needed to clarify the situation. In any case, what is certain is that the physical characteristics of lightning mentioned above are highly variable. They are so variable, in fact, that the NO_x is certain to vary not only between ground and cloud flashes, but from one ground flash to the next, or from one cloud flash to the next. Hence, without question, partitioning total lightning distributions into separate ground and cloud flash distributions is highly desirable and informative.

Finally, this effort has demonstrated that ground flash fraction retrieval is a very fascinating problem from a theoretical point of view. There exists a multitude of different ways to approach the inverse problem; section 7 on ''reduced search spaces'' alluded to this fact. In addition, the appendix illustrates that one can consider removing all reference to the use of any specific model (such as a mixed exponential distribution model) by simply employing the Central Limit Theorem (CLT) of statistics. The benefit of such an approach is that one does not have to assume anything about the specific form of the shifted MGA distributions. However, a disadvantage is that one has to retrieve two additional population statistical parameters (making a total of five unknowns to be retrieved). Because the entire retrieval process is inherently less constrained, it will be highly susceptible to solution nonuniqueness issues. Therefore, actively imposing normal distribution priors (or other external constraints) will be required. In this sense, the ''no free lunch'' principle appears to be at work. Nonetheless, testing the CLT-based approach described in the appendix would make for some interesting future work.

Acknowledgments. This research has been supported by the NOAA/NESDIS/STAR GOES-R Risk Reduction Program under Space Act Agreement NA07AANEG0284 [Ms. Ingrid Guch, Chief, NOAA/NESDIS/STAR Cooperative Research Programs Division; Dr. Mark DeMaria, Chief, NOAA/NESDIS Regional and Mesoscale Meteorology Branch; and Dr. Steven J. Goodman, Senior (Chief) Scientist, GOES-R System Program], and by the Lightning Imaging Sensor (LIS) project (Program Manager, Ramesh Kakar, NASA Headquarters) as part of the NASA Earth Science Enterprise (ESE) Earth Observing System (EOS) project. The author also thanks Dr. Richard Solakiewicz for his comments on a draft version of this manuscript, and Dr. Harold Peterson (NASA Postdoctoral Program) for his clerical help in finding some of the references cited here.

APPENDIX

The Mean Distribution Model

In one respect, it is desirable to assume as little as possible about the specific form of the distributions $p_g(x)$ and $p_c(x)$ that define the mixture distribution $p(x)$ in (1). This is because the application of the ground flash

fraction retrieval algorithm could then be more flexibly applied across the globe; i.e., the mathematical forms of $p_g(x)$ and $p_c(x)$ derived from independent observations in a specific part of the world (e.g., CONUS) would not have to be applied to other geographic locations in the retrieval process.

One way to avoid dealing with the specific forms of $p_g(x)$ and $p_c(x)$ is to invoke the Central Limit Theorem of statistics [see Koshak (2010) for specific examples]. That is, one would derive mean optical characteristics from the measurement vector, **x**, and then deal with the distributions of the means, which are normal distributions. Details of this approach follow.

Suppose that one partitioned the set of N observations contained in **x** into m subsets where each subset contains n elements, i.e., $N = mn$. Next, suppose that the mean of each subset defines the m-vector **d**; i.e., $\mathbf{d} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_m)$. If n is larger than ~30, the Central Limit Theorem of statistics says that the subset means must be approximately normally distributed; i.e.

$$p(\overline{x}_i|\mathbf{v}) = \frac{1}{\sigma}\sqrt{\frac{n}{2\pi}}\exp\left[\frac{-n}{2}\left(\frac{\overline{x}_i - \mu}{\sigma}\right)^2\right], \quad i = 1, \ldots, m. \tag{A1}$$

Here, $\mathbf{v} = (\alpha, \mu_g, \mu_c, \sigma_g^2, \sigma_c^2)$ and (μ, σ^2) are given in (2) as

$$\begin{aligned} \mu &= \alpha\mu_g + (1-\alpha)\mu_c, \\ \sigma^2 &= \alpha\sigma_g^2 + \alpha(1-\alpha)(\mu_g - \mu_c)^2 + (1-\alpha)\sigma_c^2. \end{aligned} \tag{A2}$$

So the probability of drawing **d** given **v** is

$$P(\mathbf{d}|\mathbf{v}) = \prod_{i=1}^{m} p(\overline{x}_i|\mathbf{v}). \tag{A3}$$

But from the definition in (11) one has

$$S(\mathbf{v}) \equiv \ln[P(\mathbf{d}|\mathbf{v})P(\mathbf{v})] = \sum_{i=1}^{m} \ln p(\overline{x}_i|\mathbf{v}) + \ln P(\mathbf{v}). \tag{A4}$$

Assuming that the model parameters in **v** are independent, that the model parameter α is uniformly distributed, and that the other four model parameters are normally distributed with respective means and standard deviations given by (μ_k, σ_k); $k = 2, \ldots, 5$, then (A4) reduces to

$$S(\mathbf{v}) = \phi(\mu, \sigma) + \sum_{k=2}^{5} \ln\left(\frac{1}{\sigma_k\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{k=2}^{5}\left(\frac{v_k - \mu_k}{\sigma_k}\right)^2, \tag{A5}$$

where

$$\phi(\mu, \sigma) \equiv m\ln\left(\frac{1}{\sigma}\sqrt{\frac{n}{2\pi}}\right) - \frac{n}{2\sigma^2}\sum_{i=1}^{m}(\overline{x}_i - \mu)^2. \tag{A6}$$

Hence, one finds the value of **v** (by a suitable numerical method) that maximizes S.

However, if all 5 model parameter were assumed to have uniform priors, then (A5) would reduce to

$$S(\mathbf{v}) = \phi(\mu, \sigma) + \sum_{k=2}^{5} \ln c_k = \phi(\mu, \sigma) + \text{const}, \tag{A7}$$

where the c_k are constants as described in (7). But, since (A2) is two equations in the five unknowns (α, μ_g, μ_c, σ_g^2, σ_c^2), there are an infinite number of possible **v**'s that can satisfy maximization of (A7); i.e., more than one **v** can produce the exact same values of (μ, σ).

Hence, this approach, which does not require making any assumptions about the forms of $p_g(x)$ and $p_c(x)$ in (1), leads to solution ambiguity unless one constrains (at least) three of the five model parameters; for example, by invoking normal priors as in (A5).

REFERENCES

Boccippio, D. J., K. L. Cummins, H. J. Christian, and S. J. Goodman, 2001: Combined satellite- and surface-based estimation of the intracloud-cloud-to-ground lightning ratio over the continental United States. *Mon. Wea. Rev.,* **129,** 108–122.

Celeux, G., 1998: Bayesian inference for mixtures: The label switching problem. *COMPSTAT 1998—Proceedings in Computational Statistics,* R. Payne and P. J. Green, Eds., Physica-Verlag, 227–232.

Diebolt, J., and C. P. Robert, 1994: Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Stat. Soc.,* **56B,** 363–375.

Jasra, A., C. C. Holmes, and D. A. Stephens, 2005: Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat. Sci.,* **20,** 50–67.

Keatinge, C. L., 1999: Modeling losses with the mixed exponential distribution. *Proc. Casualty Actuarial Society,* Arlington, VA, Casualty Actuarial Society, 654–698. [Available online at http://www.casact.org/pubs/proceed/proceed99/99578.pdf.]

Koshak, W. J., 2010: Optical characteristics of OTD flashes and the implications for flash-type discrimination. *J. Atmos. Oceanic Technol.,* **27,** 1822–1838.

——, and R. J. Solakiewicz, 2011: Retrieving the fraction of ground flashes from satellite lightning imager data using CONUS-based optical statistics. *J. Atmos. Oceanic Technol.,* **28,** 459–473.

——, H. S. Peterson, E. W. McCaul, and A. Biazar, 2010: Estimates of the lightning NO_x profile in the vicinity of the North Alabama Lightning Mapping Array. *Proc. 21st Int. Conf. on Lightning Detection,* Orlando, FL, Vaisala. [Available online at http://www.vaisala.com/Vaisala%20Documents/Scientific%20papers/5.Koshak,%20Peterson.pdf.]

Mach, D. M., H. J. Christian, R. J. Blakeslee, D. J. Boccippio, S. J. Goodman, and W. L. Boeck, 2007: Performance assessment of the Optical Transient Detector and Lightning Imaging Sensor. *J. Geophys. Res.,* **112,** D09210, doi:10.1029/2006JD007787.

Ott, L. E., and Coauthors, 2010: Production of lightning NOx and its vertical distribution calculated from three-dimensional cloud-scale chemical transport model simulations. *J. Geophys. Res.,* **115,** D04301, doi:10.1029/2009JD011880.

Press, W. H., W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery, 1992: *Numerical Recipes in Fortran 77, The Art of Scientific Computing*. 2nd ed. Cambridge University Press, 933 pp.

Redner, R. A., and H. F. Walker, 1984: Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.,* **26,** 195–239.

Richardson, S., and P. J. Green, 1997: On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Stat. Soc.,* **59B,** 731–792.

Rodgers, C. D., 2000: *Inverse Methods for Atmospheric Sounding; Theory and Practice*. Series on Atmospheric, Oceanic and Planetary Physics, Vol. 2, World Scientific Publishing, 240 pp.

Stephens, M., 2000: Dealing with label switching in mixture models. *J. Roy. Stat. Soc.,* **62B,** 795–809.