Title

Probabilistic assessment of cloud fraction using Bayesian blending of independent datasets: Feasibility study of a new method

Permalink
https://escholarship.org/uc/item/7qv2w8ch

Journal
Journal of Geophysical Research: Atmospheres, 118(10)

ISSN
2169-897X

Authors
Shen, Samuel SP
Velado, Max
Somerville, Richard CJ
et al.

Publication Date
2013-05-27

DOI
10.1002/jgrd.50408

Peer reviewed

Probabilistic assessment of cloud fraction using Bayesian blending of independent datasets: Feasibility study of a new method

Samuel S.P. Shen,[1] Max Velado,[1] Richard C. J. Somerville,[2] and Gabriel J. Kooperman[2]

Received 16 October 2012; revised 8 April 2013; accepted 10 April 2013; published 28 May 2013.

[1] We describe and evaluate a novel method to blend two observed cloud fraction (CF) datasets through Bayesian posterior estimation. The research reported here is a feasibility study designed to explore the method. In this proof-of-concept study, we illustrate the approach using specific observational datasets from the U. S. Department of Energy Atmospheric Radiation Measurement Program's Southern Great Plains site in the central United States, but the method is quite general and is readily applicable to other datasets. The total sky image (TSI) camera observations are used to determine the prior distribution. A regression model and the active remote sensing of clouds (ARSCL) radar/lidar observations are used to determine the likelihood function. The posterior estimate is a probability density function (pdf) of the CF whose mean is taken to be the optimal blend of the two observations. The data at hourly, daily, 5-day, monthly, and annual time scales are considered. Some physical and probabilistic properties of the CFs are explored from radar/lidar, camera, and satellite observations and from simulations using the Community Atmosphere Model (CAM5). Our results imply that (a) the Beta distribution is a reasonable model for CF for both short- and long-time means, the 5-day data are skewed right, and the annual data are almost normally distributed, and (b) the Bayesian method developed successfully yields a pdf of CF, rather than a deterministic CF value, and it is feasible to blend the TSI and ARSCL data with a capability for bias correction.

Citation: Shen, S. S. P., M. Velado, R. C. J. Somerville, and G. J. Kooperman (2013), Probabilistic assessment of cloud fraction using Bayesian blending of independent datasets: Feasibility study of a new method, *J. Geophys. Res. Atmos.*, *118*, 4644–4656, doi:10.1002/jgrd.50408.

1. Introduction

[2] The U. S. Department of Energy's Atmospheric Radiation Measurement (ARM) program was established to improve our understanding of cloud processes and the atmospheric radiation budget for climate change assessment and prediction [*Stokes and Schwartz*, 1994; *Ackerman and Stokes*, 2003]. Cloud fraction (CF) is a critical parameter in climate models. It affects the balance of the solar energy input to the climate system and the long wave radiation emitted by the Earth [*Bass et al.*, 2010; *Ramanathan et al.*, 1989; *Trenberth et al.*, 2009]. Both gridded numerical climate model data and satellite remote sensing data need to be compared with station-based in situ observations [*Xi et al.*, 2010]. The representativeness of a point observation to a grid volume, the accuracy of the grid-volume or grid-point modeling output, and the reliability of the satellite pixel data for moving and broken clouds all require careful scrutiny [*Bar-Or et al.*, 2010]. Instrument and retrieval algorithm improvements, multi-instrumental observations, numerical model simulations, and optimal blending of different datasets are all appropriate ways to improve reliable estimation of cloud parameters for climate modeling and other applications. A suite of mathematical methods needs to be developed to carry out the needed optimal data analysis. The purpose of this paper is to introduce a Bayesian blending method that combines two ground-based observations to form a probabilistic CF measure at the ARM Southern Great Plains (SGP) site.

[3] A geometric definition of CF for climate models is the percentage of sky area (from the local nadir view) which is covered by clouds [*Qian et al.*, 2012; *Kassianov et al.*, 2005]. This snapshot geometric definition is difficult to implement using an in situ observation over a large region [*Xie et al.*, 2010; *Xi et al.*, 2010]. A physically more useful definition of CF is the ratio of the space-time cloud volume to the entire space-time atmospheric volume for a given area and period of time [*Xi et al.*, 2010]. The complexity of cloud patches comes from their characteristics: irregular in spatial geometry, discontinuous in time, and varying in both location and time. Furthermore, the space-time cloud volume may not be a mathematically simply connected domain, because clouds can have irregular spatial geometry for different types of clouds, discontinuities in both space and time

[1] Department of Mathematics and Statistics, San Diego State University, San Diego, California, USA.

[2] Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California, USA.

Corresponding author: S. S. P. Shen, San Diego State University, 5500 Campanile Drive, GMCS 415, San Diego, CA, 92182–7720, USA. (shen@math.sdsu.edu)

2169-897X/13/10.1002/jgrd.50408

Table 1. Mean, Variance, Skewness, and Kurtosis of the 5-Day and Annual Cloud Fraction Data Over SGP Based on the TSI CF Data and BPE Results

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
					5-Day					
μ(BPE)	0.3575	0.4128	0.4699	0.4633	0.4443	0.4381	0.3820	0.4651	0.4222	0.4226
μ (TSI)	0.3490	0.4012	0.4831	0.4808	0.4443	0.4309	0.3840	0.4475	0.4111	0.4264
σ(BPE)	0.2583	0.2530	0.2386	0.2379	0.2134	0.2358	0.2357	0.2805	0.2088	0.2171
σ (TSI)	0.2355	0.2244	0.2263	0.2163	0.1985	0.2266	0.2105	0.2563	0.1992	0.2210
γ_3(BPE)	0.5831	0.2635	0.2078	0.3628	0.0182	0.1119	0.2894	0.2044	0.1593	0.1647
γ_3(TSI)	0.7289	0.4110	0.1720	0.3884	0.2107	0.3422	0.4696	0.2458	0.2437	0.3705
γ_4(BPE)	−0.5865	−1.1695	−1.1171	−0.7128	−0.7527	−0.8706	−0.9044	−1.2105	−0.6534	−0.7202
γ_4 (TSI)	−0.2520	−0.7598	−0.7134	−0.5180	−0.2223	−0.5297	−0.6387	−0.9652	−0.4186	−0.3385
					Annual					
$\boldsymbol{\mu}$(BPE)	0.3147	0.3924	0.4534	0.4774	0.4391	0.4627	0.4013	0.4659	0.4266	0.4434
$\boldsymbol{\sigma}$(BPE)	0.0344	0.0348	0.0350	0.0353	0.0352	0.0354	0.0356	0.0352	0.0353	.0352
$\boldsymbol{\gamma_3}$(BPE)	0.0825	0.0446	0.0230	0.0053	0.0229	0.0084	0.0331	0.0122	0.0259	.0204
$\boldsymbol{\gamma_4}$(BPE)	−0.0158	−0.0181	−0.0185	−0.0192	−0.0190	0.0194	−0.0197	−0.0191	−0.0192	−0.0191

for broken clouds, and vague spatial and temporal boundaries, especially for thin clouds. It is well known that clouds are extremely difficult to observe, because they have both spatial and temporal complexity [*Bar-Or et al.*, 2010]. Many outstanding issues exist concerning observations of CFs by different instruments and comparisons between observational data and model simulations [*Qian et al.*, 2012; *Xi et al.*, 2010; *Xie et al.*, 2010; *Kassianov et al.*, 2005]. These issues and our incomplete understanding of cloud processes limit our ability to represent clouds and their climate effects realistically in global climate models (GCMs).

[4] All the available CF datasets have strengths and weaknesses. In this paper, we have chosen to employ two specific datasets. However, we are well aware of the shortcomings of these data [*Kassianov et al.*, 2011], and we refer readers to that paper and references in it as an introduction to the large amount of literature on that topic. Our objective here is not to assess the strengths and weaknesses of these datasets. In the research reported in this paper, we emphasize that we have instead elected to use these two datasets simply for convenience in this proof-of-concept research, in order to illustrate the method that we have devised and to explore its feasibility and properties. This paper focuses on developing a probabilistic measure of CF by using a Bayesian blending of the data from two ground instruments: Total Sky Image (TSI) camera and Active Remote Sensing of Clouds (ARSCL) radar/lidar. We will justify our point observation results with grid box observations and modeling by using the gridded data from NOAA Geostationary Operational Environmental Satellite-8 (GOES8) and from NCAR Community Atmosphere Model Version 5 (CAM5). Our base data are the daytime hourly data of TSI and ARSCL. In our Bayesian blending procedure, the prior distribution is constructed from the TSI data, and the ARSCL data help build the likelihood function via a linear regression procedure. The posterior estimate is a probability density function (pdf) which blends the two observations. The GOES8 and CAM5 data are used for the results comparison and analysis.

[5] The paper is organized as follows: section 2 describes the data, section 3 describes the method, our results and conclusions are presented in section 4, and some discussion and concluding remarks are offered in section 5.

2. Data

[6] The first CF dataset is the observations by the TSI camera, which has a 352 × 288 pixel resolution and can measure the CF during daylight [*Morris*, 2005]. It retrieves CF as the ratio of the number of cloud cover pixels to the 101,376 total field-of-view (FOV) pixels, i.e., 101,376 = 352 × 288, when the local solar elevation angle (i.e., the angle between the sun direction and the horizon) is greater than or equal to 10 degrees. Thus, the TSI measures the daytime CF. The daytime length varies according to seasons. The camera sampling rate is one image per 30 s. Here, we use the 2000–2009 hourly climate model best estimate (CMBE) dataset [*Xie et al.*, 2010]. This hourly dataset was aggregated from the 30-s data. This value added product is the CF of the plane viewed from nadir direction [*Xie et al.*, 2010; *Kassianov et al.*, 2005]. The plane is thus tangent to the TSI hemispheric dome FOV at the highest point from the TSI's Earth surface location. The nadir view CF is a correction to the CF directly obtained from the hemispheric dome FOV of TSI [*Kassianov et al.*, 2005].

[7] The second observational dataset is the 1996–2009 hourly daytime ARSCL radar/lidar data, which are derived from the observations of a millimeter wave cloud radar and a micropulse lidar [*Clothiaux et al.*, 2001; *Xie et al.*, 2010]. Here, the daytime is synchronized with the TSI daytime, i.e., the ARSCL data are used for analysis only during the time when TSI data exist. Due to the nature of radar and lidar instruments, the ARSCL data are along a very narrow FOV of well less than 1^{o} around zenith, compared to the TSI hemispheric dome FOV extending 160^{o} also around zenith. Thus, unlike TSI, the ARSCL CF has to be approximated by temporal cloud averages. The original temporal grid is at 10-s resolution (Table 1 of *Xie et al.* [2010]). The hourly data are an aggregation of these 10-s data. The ARSCL CF is defined as the ratio of the number of cloud covering temporal intervals to the total temporal intervals in a given time period [*Xi et al.*, 2010], which is 1 h for the data we use here. Thus, the ARSCL CF definition using the temporal ratio represents the frequency of cloud occurrence [*Qian et al.*, 2012] and is hence different from the CF defined by the TSI instrument. The latter uses the spatial ratio of cloud covered area divided by the total area of the FOV (projected on the plane from the nadir view) [*Kassianov et al.*, 2005]. Because radar radio waves and lidar

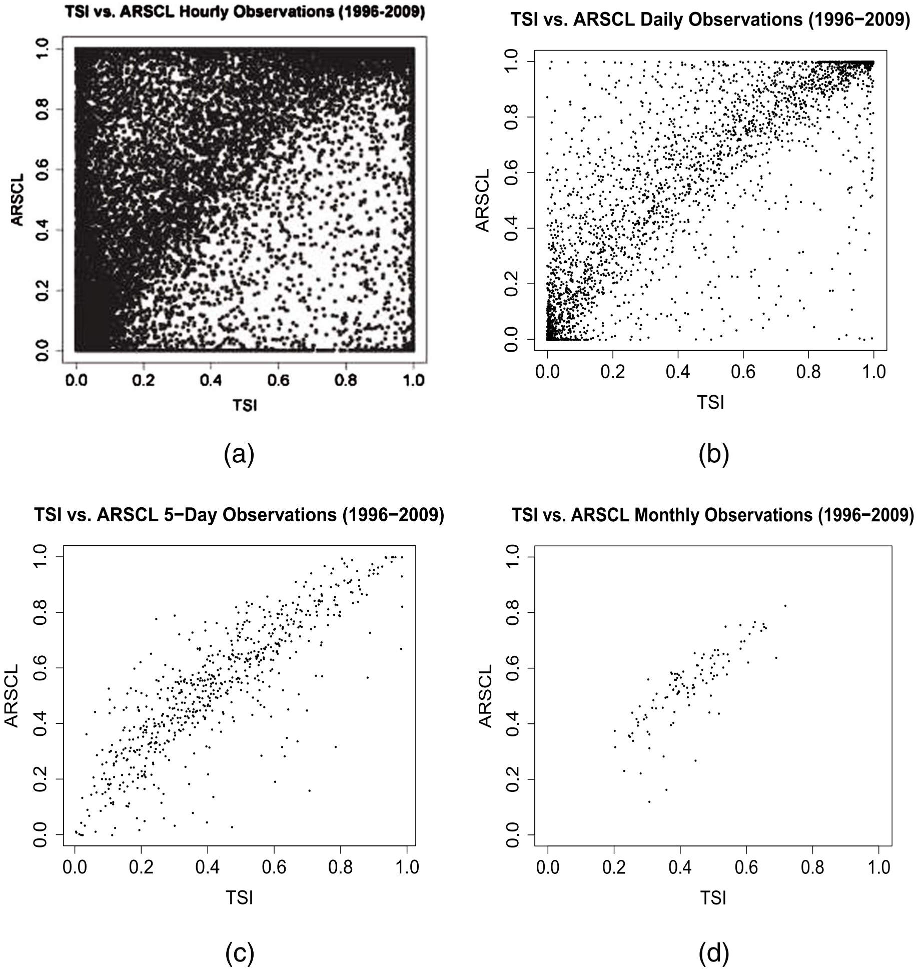
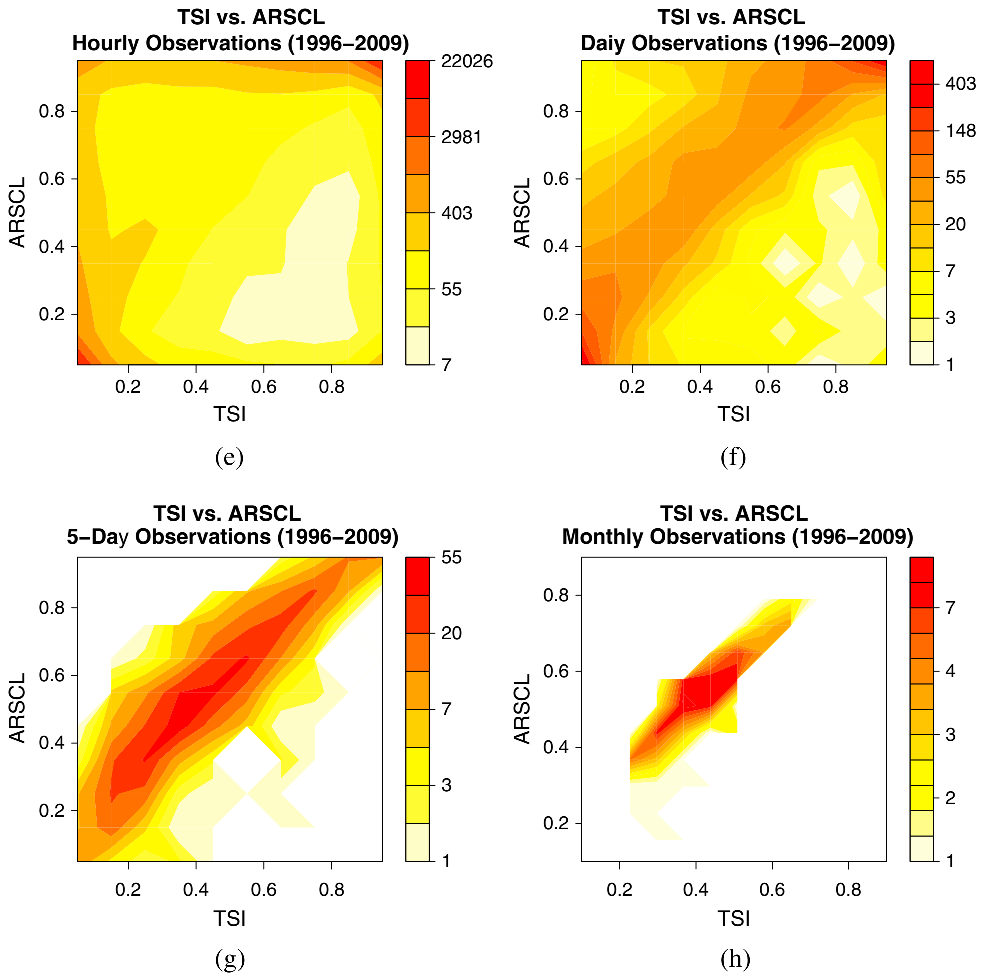


Figure 1. Scatter plots of TSI camera and ARSCL radar daytime cloud fraction data at different time scales: (a) hourly, (b) daily, (c) 5-day, and (d) monthly. Two-dimensional frequency plots of TSI camera and ARSCL radar daytime cloud fraction data at different time scales: (e) hourly, (f) daily, (g) 5-day, and (h) monthly. The two-dimensional resolution is 0.1 × 0.1. The legend indicates the number of points in each 0.1 × 0.1 grid box.

laser beams can reach high altitudes, the ARSCL CF data for SGP are a function of both height and time, while the TSI CF data are not a function of the cloud height. In addition, the ARSCL radar/lidar products also include additional cloud parameters: cloud top height, cloud bottom height, liquid water content, and others [*Xi et al.*, 2010]. Here, we focus on examining the vertically integrated CF, comparing these vertically integrated ARSCL data with the TSI data, and developing a probabilistic method to blend the two datasets together with the final product data described by CF pdfs rather than fixed CF values. For comparison purposes, only the ARSCL daytime data are used.

[8] Cloud movement, generation, and dissipation can result in a dramatic variation of CF over the SGP region within an hour. Thus, strictly speaking, the hourly CF for a given ground area should be defined as the ratio of the space-time cloud volume to the total space-time volume [*Xi et al.*, 2010], rather than a single snapshot of the spatial cloud cover fraction. The radar and camera observations of CF are approximations to this space-time cloud ratio. The hourly camera data is the average CF from the 120 snapshots within the TSI FOV. The hourly ARSCL CF data are defined as the ratio of the total number of 10-s temporal intervals with cloud presence between ground level and 14 km height to the 360 total 10-s temporal intervals making up 1 h [*Xi et al.*, 2010; *Qian et al.*, 2012]. However, because of the different CF definitions and the different FOVs between radar and camera instruments, the CFs of the camera and radar are not the same, although the instruments are colocated at the same SGP site, namely Lamont, Oklahoma, USA (36° 36’ 18.0" N, 97° 29’ 6.0" W). Instrumental sensitivities or errors also contribute to the differences. The differences become smaller when considering a long-time mean. For example, the 5-day or monthly ARSCL and TSI data are better approximations to a representative mean space-time CF than the hourly data. Compared with hourly or daily data, the 5-day and monthly ARSCL and TSI data are less scattered with a significant correlation of 0.84. Figure 1 shows scatter plots and two-dimensional frequency plots of ARSCL vs. TSI data for the SGP site at hourly, daily, 5-day, and monthly time scales for TSI and ARSCL during the common observational period: 2000–2009. For hourly data, Figure 1a shows many more ones (i.e., overcast sky) for ARSCL than for TSI, because for the short time



(e)

(f)

(g)

(h)

Figure 1. (continued)

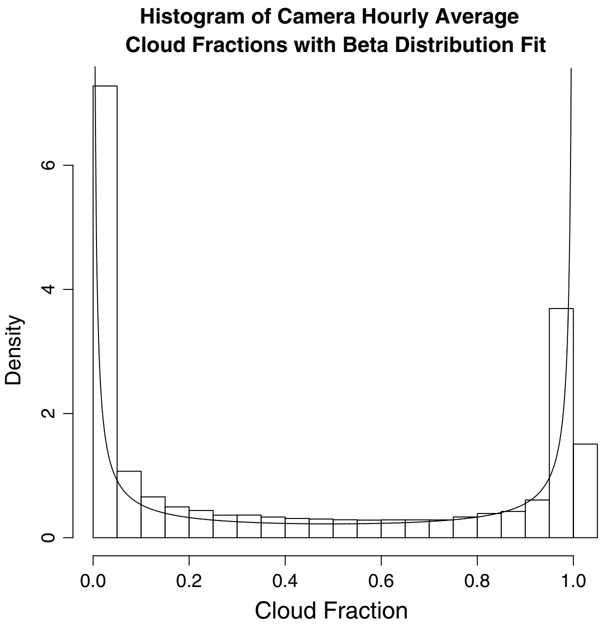


Figure 2. Histogram of the hourly TSI cloud fraction data and its Beta distribution fitting.

average, the pencil radar observation (i.e., 0° FOV) is likely to render either 0 or 1, while the TSI camera has a 160° FOV and has a larger spatial coverage than ARSCL radar. Even for TSI, most of the observations are either 0 or 1, because the edges of cloud clusters are often not in the TSI's FOV (see Figure 2). Figure 1a reveals that a large number of points are near TSI zeros and ARSCL ones. In order to clearly show the number of points in these domains, we have included two-dimensional frequency plots as Figures 1e–h, corresponding to the time scales in Figures 1a–d. Figure 1e shows that most data points are in these domains. This high frequency of occurrence of either 0 or 1 for the hourly data implies a possible good fit for a Beta distribution *Beta* $(a,\beta)(x)$, which has a bimodal shape with peaks at x = 0 and x = 1 when the shape parameters α and β are in the interval (0,1). The mathematical expression of the pdf of a Beta distribution is

$$Beta\ (\alpha,\beta)(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1} \qquad (1)$$

where Γ is a Gamma function of shape parameters α and β and is independent of random variable x. The mean and standard deviation of the random variable x are $\mu_B = \alpha/(\alpha+\beta)$

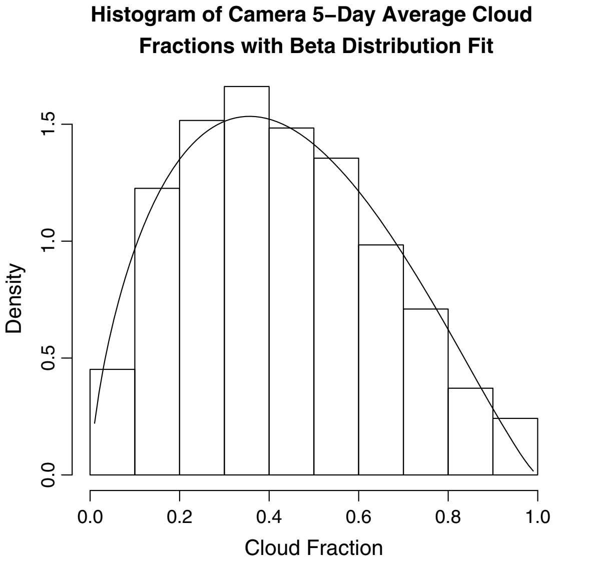


Figure 3. Histogram of the 5-day TSI cloud fraction data and its Beta distribution fitting.

and $\sigma_B = \left\{\frac{\alpha\beta}{\left[(\alpha+\beta)^2(\alpha+\beta+1)\right]}\right\}^{1/2}$. The positive values of the shape parameters depend on the time scales of the observations. When the time scale is short, say one hour, the parameters are less than 1 and *Beta* $(\alpha,\beta)(x)$ have large values near the end points of (0,1) because of the negative exponent in $x^{\alpha-1}$ and $(1-x)^{\beta-1}$. These peaks near the end points of (0,1) are shown in Figure 2 for the histogram and *Beta* $(\alpha,\beta)(x)$ fitting of the hourly TSI data. As the time scale becomes longer, the shape parameters become larger than 1. Consequently, the Beta distribution approaches a normal distribution with mean μ_B and standard deviation σ_B. See Figure 3 for the histogram and *Beta* (1.69,2.25)(x) fitting of the 5-day TSI data. Although the 5-day distribution is not yet normal, its peak is close to the midpoint of (0,1) and not near the end points.

[9] For daily data (Figure 1b), many scatter points are near the point (1,1), indicating that both TSI and ARSCL are effective in observing the case of a whole day of overcast sky. For many nearly clear-sky conditions detected by TSI (i.e., TSI CF =0), many nonzero ARSCL values exist, ranging from 0.1 to 0.6. For many nearly overcast sky conditions detected by ARSCL (i.e., ARSCL CF = 1), the TSI CF values vary in the range (0.8, 1.0). Figure 2, the SGP CF scatter plot of ARSCL vs. TSI, of *Qian et al.* [2012] also demonstrates the same properties of ARSCL overestimation relative to TSI. Examining the 5-day data (Figure 1c), TSI and ARSCL data are highly correlated, but once again, ARSCL still overestimates CF, particularly for nearly clear skies. The behavior of the scatter plot and 2D frequency plot for a longer time average is similar to that of the 5-day plot. As an example, Figures 1d and 1h show the scatter plot and 2D frequency plot, respectively, for the monthly data.

[10] Although the scatterplots are less scattered for longer time scales, the correlation coefficients between the TSI and ARSCL data change very little over the range of time scales from hourly to daily, 5-day, and monthly. The numerical values are 0.79 (hourly), 0.85 (daily), 0.84 (5-day), and 0.81 (monthly). Despite this small variation in correlation, the reduced number of points for longer time scales indicates significant correlations between the two datasets at longer time scales. Therefore, Figure 1 implies that at a longer time scale, both TSI and ARSCL are able to measure the strictly defined space-time CF. This asymptotic behavior as time scale increases was also explored by *Xi et al.* [2010] who compared GOES8 satellite observations with 35 Hz Millimeter Wavelength Radar results.

3. Method

[11] We will now describe the Bayesian method to blend the TSI and ARSCL data at the time scale at which the two approximately measure the same parameter: space-time CF. The result will be a Bayesian blended probabilistic measure of CFs.

[12] The Bayesian method is developed based on Bayes' theorem, sometimes called Bayes' law for conditional probability, which is commonly included in textbooks for elementary undergraduate statistics courses [e.g., *Johnson*, 2010]. With today's continuing advances in computer power, the Bayesian method is undergoing rapid development. It has already become a powerful data analysis tool and has been applied to a wide range of problems. For an introduction to the Bayesian method, see [*Albert*, 2009]. The Bayesian method for climate data analysis has already been used by many researchers [e.g., *McFarlane et al.*, 2002; *Coelho et al.*, 2004; *Chiu and Petty*, 2006]. Although every Bayesian approach is based on the same Bayes' theorem, there are numerous ways to carry out the calculation steps for particular types of application problems. The novelties of a new Bayesian method are typically in the procedures for constructing prior distributions and likelihood functions, as well as the development of creative numerical techniques for massive computations. The novelty of the method presented in this paper lies in our specific ways of constructing prior distribution and likelihood functions for TSI and ARSCL. Next, we describe our procedures of constructing the prior distribution, likelihood function, and posterior distribution for CFs from TSI and ARSCL.

[13] For the construction of the prior distribution, different kinds of prior models have been considered in the past. For example, *Chiu and Petty* [2006] fitted a lognormal model, while *McFarlane et al.* [2002] used the Gaussian distribution. Here, we will fit TSI CF data to a Beta distribution using the moment method. The TSI statistical moments computed for 5-day data for each year are shown in Table 1.

[14] Let us consider the 5-day CF as an example. Let θ be a random variable, representing CF from TSI. We estimate a Beta pdf for θ using the method of moments (see Chapter 9 of *Wackerly et al.* [2008]). The first four statistical moments of each year for a few selected time scales are shown in Table 1.

[15] For the entire 10 years from 2000 to 2009, Figure 3 shows a histogram of TSI 5-day data and our fitted Beta density function, which is the prior distribution $\pi(\theta) = Beta$ (1.69,2.25)(θ), a formula explicitly described in equation (1). Here, the parameters $\alpha = 1.69$ and $\beta = 2.25$ are computed from the sample mean and standard deviation of 730 five-day data from 1 January 2000 to 31 December 2009 excluding 29 February. The mean is 0.4282, and the standard deviation is 0.2204.

Table 2. The Regression Coefficients (β_0, β_1) and Standard Deviations σ of the Regression for the 5-day Averages and Monthly Average CF[a]

	5-Day			Monthly		
Year	β_0	β_1	σ	β_0	β_1	σ
2000	0.1265	1.0108	0.1035	0.1034	0.9853	0.0722
2001	0.0567	0.8504	0.1595	−0.1223	1.4523	0.1282
2002	0.1313	0.7805	0.1502	0.1871	0.8575	0.0501
2003	0.2499	0.7132	0.1070	0.2057	0.7730	0.0535
2004	0.1431	0.9005	0.1394	0.1046	0.9893	0.1034
2005	0.2582	0.7454	0.1228	0.2313	0.8064	0.0344
2006	0.1843	0.9340	0.0971	0.1594	1.0006	0.0494
2007	0.1440	0.8788	0.1135	0.0730	1.0268	0.0469
2008	0.1255	0.9818	0.0818	0.1692	0.8710	0.0530
2009	0.1103	0.9799	0.0611	0.1332	0.9306	0.0284

[a] The values of β_0, β_1 and σ for the yearly average CFs from 2000 to 2009 are 0.2538, 0.6429 and 0.0376, respectively.

[16] For a specific 5-day TSI datum, say, days 1 to 5 of 2003, the parameters α and β are determined by the 5-day TSI datum as the mean and the TSI standard deviation of the year 2003 shown in Table 1. Thus, each 5-day TSI CF corresponds to a prior distribution with the mean equal to the 5-day CF datum and the standard deviation and higher moments estimated from the 73 five-day data in that year.

[17] One may notice a dramatic difference between the two histograms shown in Figure 2 for hourly data and Figure 3 for 5-day data. The hourly data have a bimodal distribution with peaks at 0 or 1, while the 5-day data follow a uni-modal distribution due to the long-time average. According to the central limit theorem (see Chapter 7 of *Wackerly et al.* [2008]), the mean of large samples approaches a normal distribution. The 5-day mean CF is much closer to a normal distribution than the 1-h mean. Still, the 5-day data are skewed right with more frequent occurrence of less than 50% CFs. We judge that this is probably the reality for the SGP cloud characteristics: more frequent occurrence of clear moments or partially cloudy times than of overcast or very cloudy times.

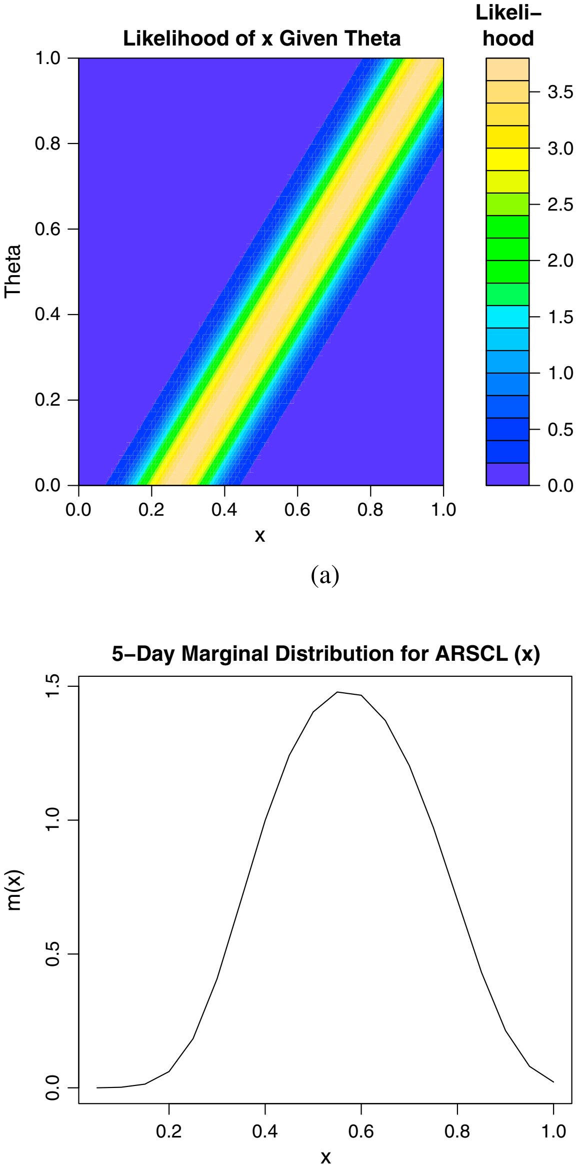
[18] In summary, our prior distribution is modeled by a Beta distribution. The position(s) of the peak(s) move from the end points toward the midpoint as time scales become longer.

[19] For the construction of the likelihood function, we use both ARSCL and TSI CF data and a regression procedure. The likelihood function is often referred to as the forward model in the community of satellite data retrieval [*McFarlane et al.*, 2002; *Chiu and Petty*, 2006]. The likelihood function may be modeled by incorporating some physical understanding via mathematical or statistical models determined by physical properties or data. Here, we choose to use ARSCL and TSI data to build a regression model, whose residual is normally distributed. This normal distribution acts as the forward model, whose variance is fixed for a given year and whose mean is the 5-day ARSCL datum. Each year thus has 73 likelihood functions corresponding to the 73 five-day data values. The computational details of this construction are explained below.

[20] Let X represent the second observation given the prior CF distribution. Regression is used to model X for a given CF θ [*Graybill and Iyer*, 1994]:

$$x|\theta = \beta_0 + \beta_1\theta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \tag{2}$$

where ε is the regression model error with a standard deviation equal to σ. This equation describes the radar CF given the TSI camera observation. The regression coefficients (β_0, β_1) were calculated by fitting a linear regression model between TSI and ARSCL data for each year, e.g., the year 2000 had its own regression coefficients for the 5-day averages. The same procedure was done for the monthly averages. Thus, for the 5-day averages, each year has 73 data points for calculating the regression coefficients; for the monthly averages, each year has 12 data points for regression. For the annual average, however, we have only 10 data points from 2000 to 2009, so we have only one pair of coefficients for the annual time scale. Table 2 shows these regression coefficients (β_0, β_1) and the standard deviation



(a)

(b)

Figure 4. (a) Likelihood function for different values of x and θ. (b) The normalization factor $m(x)$.

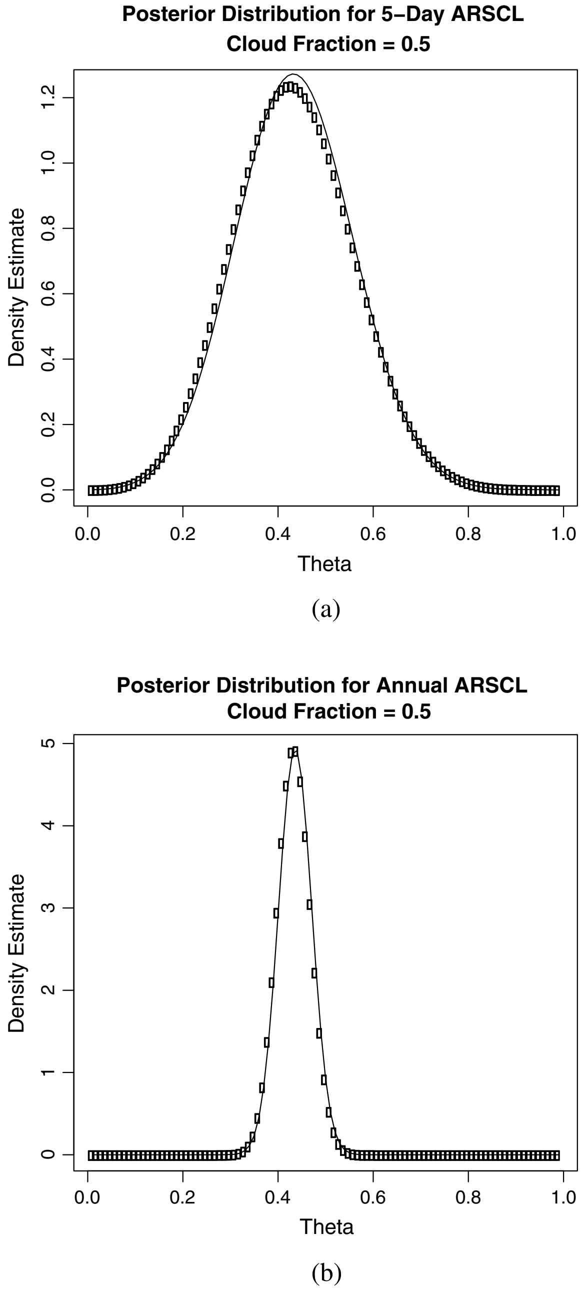


Figure 5. Posterior distribution of the cloud fraction (a) when the ARSCL datum is 0.5 for 5-day CF, and (b) when the ARSCL datum is 0.5 for annual CF.

σ of the regression residual for the two time scales: 5-day and monthly. Thus, the explicit expression of a likelihood function $f(x|\theta) \sim N(\beta_0 + \beta_1\theta, \sigma^2)(x)$ is

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\, exp\left[-\frac{(x-\beta_0-\beta_1\theta)^2}{2\sigma^2}\right] \tag{3}$$

[21] Figure 4a shows the likelihood function for the 5-day ARSCL and TSI data for the year 2003. For each fixed θ, x is normally distributed, and the mean is $\beta_0 + \beta_1\theta = 0.2499 + 0.7132\theta$, while the standard deviation is $\sigma = 0.1070$. Figure 4a's likelihood function was plotted based on a 100×100 grid on the x - θ plane.

[22] If TSI and ARSCL have statistically identical observations, then $\beta_0 = 0$ and $\beta_1 = 1$, implying the likelihood function equal to $N(\theta,\sigma^2)(x)$. This kind of likelihood function is often found in the literature (e.g., *Chiu and Petty* [2006], *McFarlane et al.* [2002]). In this case, the posterior distribution will be independent of the choice of which data are used for prior and which data for likelihood. However, Table 2 shows that β_1 often deviates far away from 1. The smallest $\beta_1 = 0.7454$ is from the 2005 five-day data, and the largest $\beta_1 = 1.0108$ is from the 2001 monthly data. β_0 is always positive except in the 2001 monthly data ($\beta_0 = -0.1223$ for 2001). A nonzero β_0 value implies the existence of bias in either TSI data or ARSCL data or both. According to Figure 1c, there are two bias possibilities: (a) ARSCL yields an overestimate of CF for the nearly clear sky and an underestimate for the heavily covered sky, and (b) TSI yields an overestimate for the heavily covered sky and an underestimate for the nearly clear sky. We agnostically chose TSI data for constructing the prior distribution and ARSCL data for the likelihood in this study. The accuracy and error assessment of TSI and ARSCL data are still under investigation for various applications, including climate model validation and parameterization [*Qian et al.*, 2012; *Xie et al.*, 2010]. We thus do not assert here which prior distribution or which likelihood function is superior. Instead, we attempt to describe a methodology to derive a pdf for every pair of TSI and ARSCL data.

[23] The likelihood function developed this way is applied to 5-day, monthly, and annual data. Further investigation would be needed to determine whether this model is applicable to the hourly data, because there is strong nonstationarity at this time scale, for which model (2) may become invalid. A moving time window method may be adopted for an approximation of piecewise stationarity [*Shen et al.*, 2012].

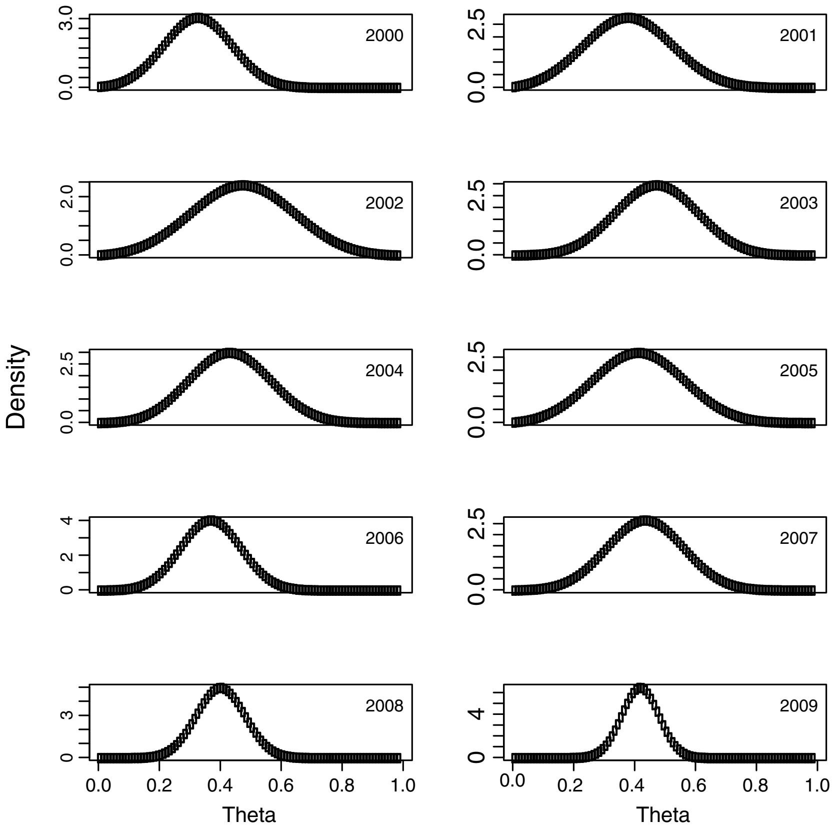
[24] The third step is to apply Bayes' theorem of conditional probability. For a given TSI CF datum, the prior distribution $\pi(\theta)$ is determined. For the corresponding ARSCL CF datum for the same time and location, the likelihood function is determined. The Bayes' formula yields the Bayesian posterior estimate (BPE)

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{m(x)} \tag{4}$$

where $m(x)$is a normalization factor equal to

$$m(x) = \int_0^1 \pi(\theta)f(x|\theta)d\theta. \tag{5}$$

[25] This normalization factor is also called the marginal distribution and can be numerically computed. See Figure 4b for the shape of this function. The posterior density function $\pi(\theta \mid x)$ given in equation (4) is the probabilistic estimate of the CF based on two observations, an assumption of the prior distribution, and a regression analysis. Thus, the CF is no longer expressed in terms of a single fixed value. Instead, it is described by a pdf given by equation (4). See Figures 5 and 6 for some examples of pdf outputs.



(a)

Figure 6. The posterior distribution of the cloud fraction of each year from 2000 to 2009 at the SGP site (a) for 5-day CF as the mean of the 73 five-day CF values, and (b) for annual CF.

[26] Various kinds of statistical properties can be calculated from these pdfs. The commonly used properties are the mean, standard deviation, skewness, and kurtosis, which are related to the first four central moments. These four statistics are calculated using the following formulas via numerical integrations:

$$\mu_1 = \int_0^1 \theta\pi(\theta|x)d\theta \qquad (mean) \tag{6}$$

$$s_2 = \left[\int_0^1 (\theta - \mu_1)^2 \pi(\theta|x)d\theta\right]^{1/2} \qquad (standard\ deviation) \tag{7}$$

$$\gamma_3 = \int_0^1 \left(\frac{\theta - \mu_1}{s_2}\right)^3 \pi(\theta|x)d\theta \qquad (skewness) \tag{8}$$

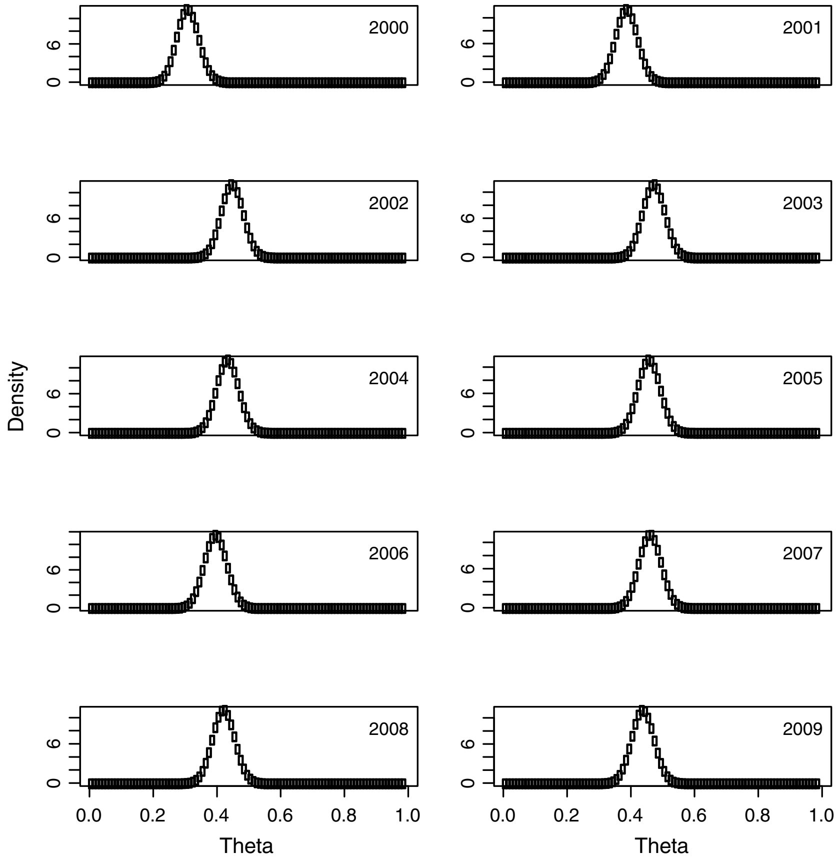
$$\gamma_4 = \int_0^1 \left(\frac{\theta - \mu_1}{s_2}\right)^4 \pi(\theta|x)d\theta - 3 \qquad (kurtosis) \tag{9}$$

[27] Table 1 shows numerical examples calculated from the above four formulas.

4. Results and Conclusions

[28] The BPE method yields a probabilistic output from two observations according to formula (4). Figure 5 shows a posterior distribution of θ for a given radar observation $x = 0.5$ for the year 2003 at a 5-day scale and annual scale, respectively. The posterior distribution is not symmetric for either scale. According to Table 1 calculated from posterior estimates, the mean, standard deviation, skewness, and kurtosis of CF in 2003 are 0.4633, 0.2379, 0.3628, and −0.7128 for the 5-day data, and 0.4774, 0.0353, 0.0053, and −0.0192 for the annual data. These are positively skewed, i.e., skewed right with preference to less cloudy skies.

[29] Figure 6 shows the CF pdf for the SGP site for the 10-year period of 2000–2009 obtained with the BPE approach for two time scales: 5-day and annual. Figure 6a shows the posterior distribution of the average of the 5-day data for each year. The annual 5-day data are the average of the 73 five-day data of each year. The first four moments of the 5-day TSI CF data are shown in Table 1. We also calculated the first four moments of ARSCL CF data (not shown in this paper). The BPE procedure produced the posterior distributions shown in Figure 6a. The distributions are centered around 0.3–0.5. For a single mode in the middle of an interval, the posterior distribution is close to being symmetric around the mean although it is not normal. Due to the short time scale, the mean varies from year to year, ranging from a minimum of



(b)

Figure 6. (continued)

0.3575 in 2000 to a maximum of 0.4699 in 2002 (see Table 1). The standard deviations have little change from year to year, remaining around 0.23. The distributions are almost symmetric and slightly positively skewed. The three most skewed years are 2000, 2003, and 2006. The least skewed year is 2004 whose distribution is almost symmetric.

[30] The kurtoses for both 5-day and annual data are negative, which implies a platykurtic distribution, i.e., one that is less peaked than a normal distribution. The 5-day data have larger absolute values of negative kurtosis, hence the pdf shapes are flatter, while the annual data's kurtosis values are closer to 0 and the pdf shapes' peakedness is close to that of normal (see Table 1 and Figure 6). The least peaked year for the 5-day data is 2007 with a kurtosis equal to -1.2105.

[31] At the annual time scale, we should not expect too much inter-annual variation of cloud statistical properties. It is thus not surprising that the CF pdfs in Figure 6b show similar shapes. The standard deviations are almost a constant at around 0.03 in the 10 years and are about 15% of those of the 5-day data. The skewness is also small since the distribution for the annual data is almost symmetric. The maximum is 0.0825 in 2000 and the minimum is 0.0053 in 2003. The positive skewness implies that the distributions are all slightly skewed right, indicating a less-than-0.5 CF preference.

[32] Besides estimating the pdf of the annual mean of the 5-day CF, we can also calculate the posterior distribution for any given 5-day, say, 181st–185th days (i.e., 37$^{\text{th}}$ 5-day data each year) of each year from 2000 to 2009 (see Figure 7). We use the 5-day datum of a given year as the mean and the standard deviation of the year to approximate the standard deviation of the sample 181st–185th days of the same year. The TSI and ARSCL data for the 181st–185$^{\text{th}}$ days for all the ten years are (2000: 0.3862, 0.5174), (2001: 0.3016, 0.4306), (2002: 0.7819, 0.7924), (2003: 0.1377, 0.1795), (2004: 0.4885, 0.5614), (2005: 0.3473, 0.4451), (2006: 0.2695, 0.3978), (2007: 0.7202, 0.8861), (2008: 0.2852, 0.3752), and (2009: 0.4376, 0.5273). The 37$^{\text{th}}$ 5-day data for 2005 and 2009 were missing from CMBE dataset and were replaced by the data of the 5-day nearest to 37$^{\text{th}}$: 32$^{\text{nd}}$ for 2005 and 32$^{\text{nd}}$ for 2009. With the mean and standard deviation, the Beta distribution's parameters α and β can be determined, and hence the prior distribution is determined. The parameters for the likelihood function model are the same for the entire year 2003 with $\beta_0 = 0.2499, \beta_1 = 0.7132, \sigma = 0.1070$ (see Table 2). When the ARSCL 5-day datum for 181st–185th days of each year is given (i.e., a given x value, denoted by x_2), the likelihood function $f(x_2|\theta)$ is uniquely determined, and the posterior distribution $\pi(\theta|x_2)$ can then be determined by equation (4) via a numerical integration for finding $m(x_2)$. The posterior distributions for the 181st–185$^{\text{th}}$ days are shown in Figure 7 for each year from 2000 to 2009. Some distributions are obviously asymmetric. The figures appear to imply that (a) when the posterior CF mean is large, then the posterior distribution is skewed left (e.g., years 2002, 2007), (b) when the posterior CF mean is small, then the posterior distribution

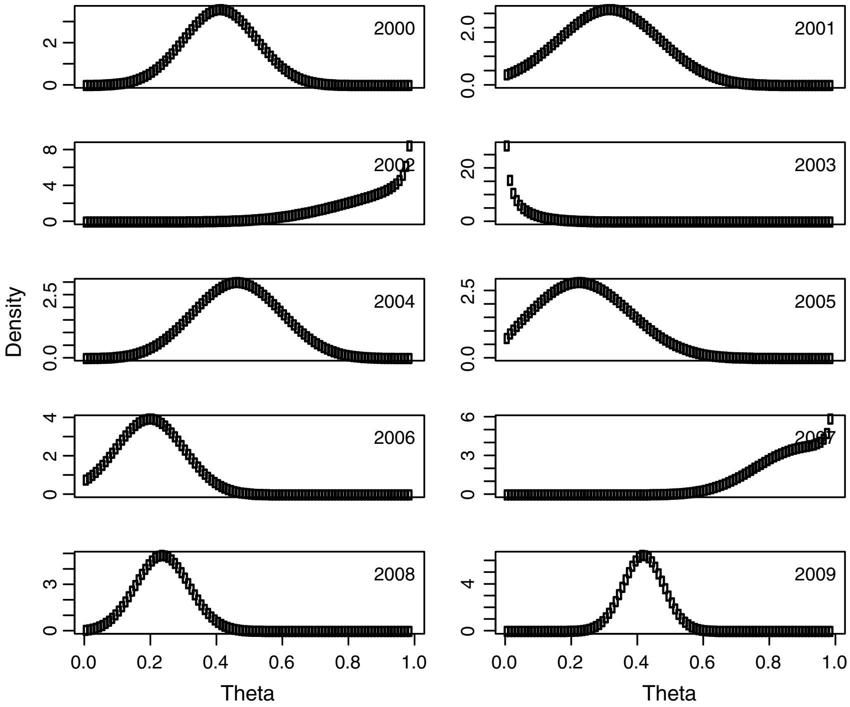
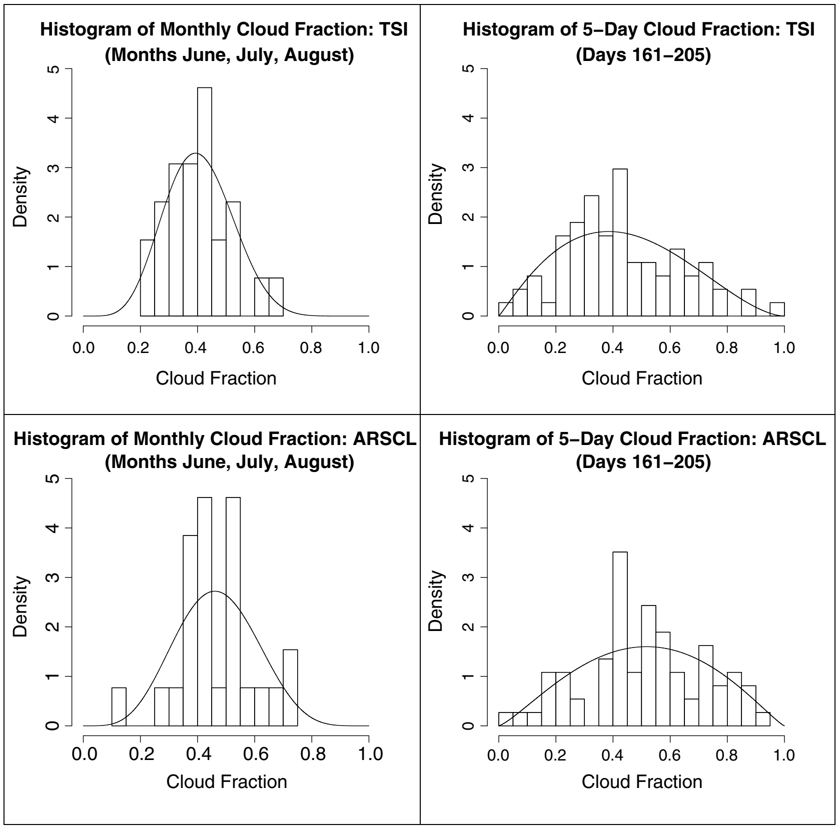
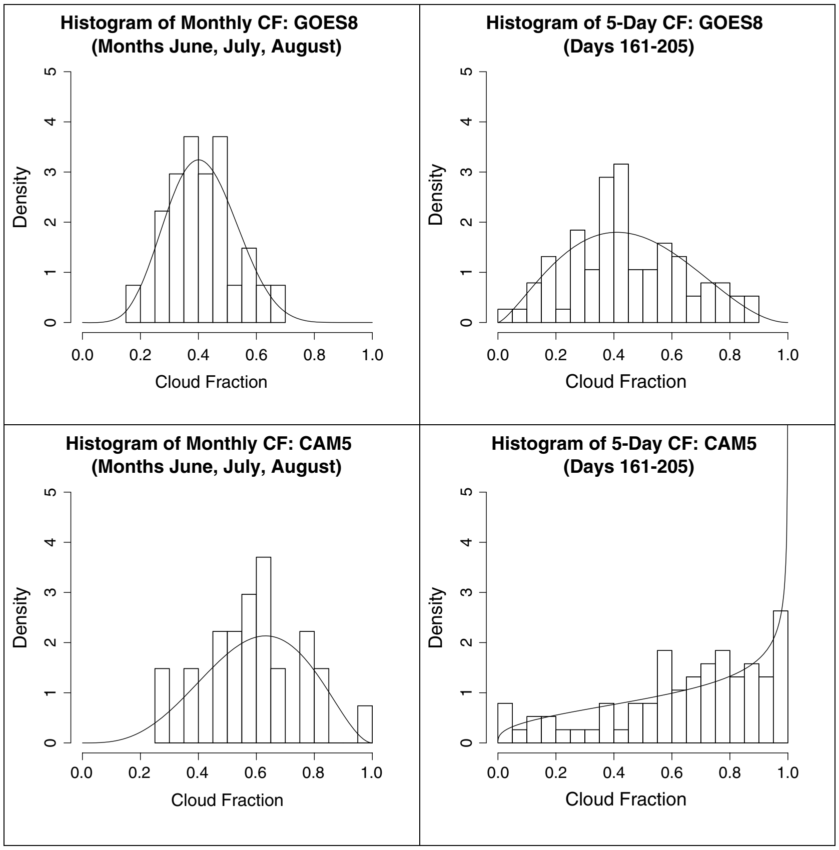


Figure 7. The posterior distribution of the cloud fraction of each year from 2000 to 2009 at the SGP site for the 181st – 185th days mean CF.



(a)

Figure 8. Histograms and their Beta distribution fitting of cloud fractions based on the monthly and 5-day time scales: (a) TSI and ARSCL data, and (b) GOES8 and CAM5 data.



(b)

Figure 8. (continued)

is skewed right (e.g., years 2001, 2003, 2005, 2006, 2008), and (c) when the posterior CF mean is near 0.5, then the posterior distribution is close to being normal (e.g., years 2000, 2004, 2009). These agree with our physical intuition: the CF has a high frequency of occurrence around the actually observed values.

[33] The results shown in Figures 6 and 7 and Table 1 demonstrate the advantage of the BPE approach over the traditional minimum mean square error approach [*North et al.*, 1991]. The traditional method is a least-squares procedure that can yield an optimal mean under a normal distribution assumption. The BPE approach thus yields much more information, including the pdf (which may be asymmetric as in the case of the 5-day CF) and other statistical properties of the estimated parameters. In the parameterization process of ensemble climate modeling, the pdf is critical for implementing the method of importance sampling. The peaked frequency of a parameter should be sampled more often than the tail distribution part of a parameter. Thus, our results here show a proof-of-concept example for stochastic parameterization for ensemble climate modeling procedures.

[34] According to *Xi et al.* [2010], simply making a direct real-time comparison between a point observation (ARSCL) and an areal observation (e.g., TSI and GOES8) or modeling result (e.g., CAM5) is inconsistent, since the point and area measurements in a very short time interval are observing different physical quantities. However, their work indicates that it is feasible to check the climatology and other long-term means from different observations or models for consistency validation. Further, probabilistic assessment by comparing pdfs will make sense. We explored the GOES8 and CAM5 CF data for the purposes of comparing them with the TSI and ARSCL data and with our BPE estimate. Our previously described results also provide evidence that it is useful to explore the probabilistic properties of the CFs from GOES8 and CAM5 data. For probability distribution, we have analyzed the summer months: June, July, and August in 2000–2009. Both ARSCL and GOES8 had 30 samples, where a sample is a monthly mean value, and we have used 3 months from each of the 10 years. TSI missed 3 months and had only 27 samples. CAM5 uses the grid point that is closest to the SGP site. Figure 8 compares the histograms and the fitted Beta distributions of GOES8, CAM5, TSI, and ARSCL summer CF data for 10 years (2000–2009). The pdfs were computed using the moment method. Two time scales are examined: monthly and 5-day.

[35] For the monthly data in June, July, and August, TSI, ARSCL, and GOES8 (see Figure 7's left column) have a mean of about 0.3–0.5 and are skewed right. The distributions of TSI and GOES8 appear very similar to each other, presumably because both of them have a large FOV, and the CFs from both instruments are defined in the same way as the areal fraction of clouds, although GOES8's FOV is much larger. One may notice that the monthly observed CFs are less than 0.8. This is expected since it is unlikely that almost overcast daytimes can persist for a month at SGP in the summer. It is quite

surprising that the distribution of monthly CAM5 data exhibits a certain degree of similarity to that of GOES and TSI, although the former is skewed left while the GOES and TSI are skewed right, and furthermore, CAM5 shows the existence of nearly overcast months with CF > 0.8. The CAM5 results have a higher mean of about 0.6–0.7 and are skewed left. The left column of Figure 8 and the above analysis indicate more frequent overcast or nearly overcast days in CAM5 results in the summer than in the observations.

[36] For the 5-day data in late June and early July, we chose nine 5-day intervals (161^{st}–165^{th} days counting from January 1 with exclusion of 29 February, and 166^{th}–170^{th} days, 171^{st}–175^{th} days, 176^{th}–180^{th} days, 181^{st}–185^{th} days, 186^{th}–190^{th} days, 191^{st}–195^{th} days, 196^{th}–200^{th} days, and 201^{st}–205^{th} days). The full sample size is thus 90 for TSI, ARSCL, GOES8, and CAM5. The histograms on the right column were computed from the full samples. Again, the three distributions of observations from TSI, ARSCL, and GOES8 exhibit a certain degree of similarity (Figure 8, right panels). The peak frequencies of TSI, ARSCL, and GOES8 CFs are all near 0.4. TSI has the lowest CF mean value and is skewed right. The ARSCL distribution is the flattest. The GOES8 distribution is similar to TSI, but has fewer overcast days with a CF greater than 0.9 and also has fewer clear days with a CF less than 0.1. This result is not unexpected, because GOES8 has a large FOV and hence often reflects an average of overcast and clear areas. On the other hand, CAM5 CF exhibits a dramatically different distribution. It is strongly skewed left and has a larger mean CF value. A surprising feature is its peak frequency near 1.0, indicating many overcast times that persist for 5 days or longer.

[37] The above preliminary comparisons for similar and different features between CAM5 CFs and those from observations imply the necessity of detailed comparative studies on climate models' cloud parameterization and observed CFs. For a more comprehensive comparison among the CFs from TSI, ARSCL, and AR4 models, see *Qian et al.* [2012].

5. Discussion and Concluding Remarks

[38] We have introduced a BPE approach to optimally blend different CF datasets. To illustrate the approach, we have employed CF data obtained from the TSI camera and the ARSCL radar-lidar observations at the ARM SGP site. Both datasets have shortcomings, and we employ them only as sample datasets to illustrate the method and to evaluate its applicability, without making any claims as to the strengths and weaknesses of these datasets in comparison to products from other available observational technologies. This proof-of-concept study shows the feasibility of the BPE method. The method is applicable to non-Gaussian prior distributions and is different from the traditional least-squares approach, which assumes the normal distribution of both prior and likelihood functions. Since our likelihood function is constructed by a regression between two observations, our method thus helps correct systematic bias in either observation. When the correction is very large, the posterior estimate may become an extrapolation of the two observations. The analysis of the summer probabilistic distribution of the SGP CFs demonstrates the consistency between the CAM5 model CF and the observed CF from ARSCL, TSI, and GOES8 in the monthly scale, as well as the inconsistency in the 5-day scale. Our results imply the following: (a) CF is best defined as a space-time percentage of the cloud volume with respect to the total space-time volume, (b) the Beta distribution is a reasonable model for the CF for both short and long-time means, and (c) it is feasible to blend TSI and ARSCL data using our BPE procedure of constructing the prior CF using a Beta distribution and the likelihood function using a regression, in order to construct a CF pdf for various diagnostic and modeling applications.

[39] Given the above comparison of results from different observations and blending, it is still an open question as to how best to effectively compare "pencil" observations, such as ARSCL, with the grid box data of GOES8 and GCM. The randomization method of [*North et al.*, 1994] for accessing the ground truth errors of the TRMM satellite compared with the ground rain gauge observations might be an effective approach. The error will then be determined by the spectral properties of the covariance function of the CF. When a simple diffusive CF model is assumed, it will be possible to calculate the spectra analytically, which can lead to an analytic expression of the ground truth error.

[40] **Acknowledgments.** This research was supported in part by the U.S. Department of Energy (DOE)'s Atmospheric System Research, an Office of Science, Office of Biological and Environmental Research program, under Grants No. DE-FG02-09ER64764, No. DE-FG02-09ER647645 and DOE Grant No. DE-SC0002003, and by the National Science Foundation (NSF) under Grant No. AGS-1015926. This research was also supported in part by the Center for Multiscale Modeling of Atmospheric Processes (CMMAP), an NSF Science and Technology Center managed by Colorado State University under cooperative agreement No. ATM-0425247. The datasets of this research were acquired from the DOE Atmospheric Radiation Measurement Program. The authors thank Jeff Ledahl for his preliminary BPE formulation and Barbara Bailey for valuable discussions. We also thank the anonymous referees for their thorough reviews and constructive suggestions that significantly improved this paper.

References

Albert, J. (2009), *Bayesian Computation with R*, 2nd ed., 298 pp., Springer, New York.

Ackerman, T. P., and G. M. Stokes (2003), The Atmospheric Radiation Measurement Program - To predict reliably what increased greenhouse gases will do to global climate, we have to understand the crucial role of clouds, *Phys. Today*, *56*, 38–46.

Bass, L. P., O. V. Nikolaeva, V. S. Kuznetsov, and A. A. Kokhanovsky (2010), Radiation balance in a cloudy atmosphere with account for the 3D effects, *Atmos. Res.*, *98*, 1–8. doi:10.1016/j.atmosres.2009.12.001.

Bar-Or, R. Z., I. Koren, and O. Altaratz (2010), Estimating cloud field coverage using morphological analysis, *Environ. Res. Lett.*, *5*, 014022, doi:10.1088/1748-9326/5/1/014022.

Chiu, C., and G. W. Petty (2006), Bayesian retrieval of complete posterior PDFs of oceanic rain rate from microwave observations, *J. Applied Meteo. Climatology*, *45*, 1073–1095.

Clothiaux, E. E., et al. (2001), The ARM Millimeter Wave Cloud Radars (MMCRs) and the Active Remote Sensing of Clouds (ARSCL) Value Added Product (VAD), DOE Tech. Memo. ARM VAP-002.1, U.S. Department of Energy, Washington, D.C.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes, and D. B. Stephenson (2004), Forecast calibration and combination: A simple Bayesian approach for ENSO, *J. Climate*, *17*, 1504–1516.

Graybill, F. A., and H. K. Iyer (1994), *Regression Analysis: Concepts and Applications*, Duxbury Press, Belmont, California, pp. 699.

Johnson, R. A. (2010), *Statistics: Principles and Methods*, 6th ed., 696 pp., John Wiley, New York.

Kassianov, E., J. C. Barnard, L. K. Berg, C. Flynn, and C. N. Long (2011) Sky cover from MFRSR observations, *Atmos. Meas. Tech.*, *4*, 1463–1470.

Kassianov, E. I., C. N. Long, and M. Ovtchinnikov (2005), Cloud sky Cover versus cloud fraction: whole-Sky simulations and observations, *J. Applied Meteo.*, *44*, 86–98.

McFarlane, S. A., K. F. Evans, and A. S. Ackerman (2002), A Bayesian algorithm for the retrieval of liquid water cloud properties from microwave radiometer and millimeter radar data, *J. Geophys. Res.*, *107*(D16), 4317, doi:10.1029/2001JD001011.

Morris, V. R. (2005), Total Sky Imager Handbook, DOE Office of Science Technical Memo. DOE/SC-ARM/TR-017, doi:10.2172/1020716, U.S. Department of Energy, Washington, D.C.

North, G. R., S. S. P. Shen, and R. B. Upson (1991), Combining raingages with satellite measurements for optimal estimates of area-time averaged rainrates, *Water Resour. Res.*, *27*, 2785–2790.

North, G. R., J. B. Valdes, E. Ha, and S. S. P. Shen (1994), The ground truth problem for satellite estimates of rain rate, *J. Atmos. Ocean. Tech.*, *11*, 1035–1041.

Qian, Y., C. N. Long, H. Wang, J. M. Comstock, S. A. McFarlane, and S. Xie (2012), Evaluation of cloud fraction and its radiative effect simulated by IPCC AR4 global models against ARM surface observations, *Atmos. Chem. Phys.*, *12*, 1785–1810.

Ramanathan, V., R. D. Cess, E. F. Harrison, P. Minnis, B. R. Barkstrom, E. Ahmad, and D. Hartmann (1989), Cloud-radiative forcing and climate: results from the Earth Radiation Budget Experiment, *Science*, *243*, 57–63, doi:10.1126/science.243.4887.57.

Shen, S. S. P., C. K. Lee, and J. Lawrimore (2012), Uncertainties, trends, and hottest and coldest years of US surface air temperature since 1895: an update based on the USHCN V2 data, *J. Climate*, *25*, 4185–4203.

Stokes, G. M., and S. E. Schwartz (1994), The Atmospheric Radiation Measurement (ARM) Program: Programmatic background and design of the cloud and radiation test bed, *Bull. Amer. Meteo. Soc.*, *75*, 1201–1221.

Trenberth, K. E., J. T. Fasullo, and J. Kiehl (2009), Earth's global energy budget, *Bull. Amer. Meteo. Soc.*, *90*, 311–323.

Wackerly, D. D., W. Mendenhall III, and R. L. Scheaffer (2008), *Mathematical Statistics with Applications*, 7th ed., 912 pp., Thomson, Brooks/Cole, Belmont, California.

Xi, B., X. Dong, P. Minnis, M. M. Khaiyer (2010), A 10 year climatology of cloud fraction and vertical distribution derived from both surface and GOES observations over the DOE ARM SPG site, *J. Geophys. Res.*, *115*, D12124, doi:10.1029/2009JD012800, 2010.

Xie, S., et al. (2010), ARM Climate Modeling Best Estimate Data, *Bull. Amer. Meteor. Soc.*, *91*, 13–20.