Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

NEAREST NEIGHBOR SPACING DISTRIBUTIONS OF CUMULUS CLOUDS

Robert F. Cahalan
Laboratory for Atmospheres
NASA/Goddard Space Flight Center

1. INTRODUCTION

The statistical distribution of cumulus cloud spacings is important both in determining the resolution required for the remote sensing of cumulus clouds, and in the modeling of cumulus clouds and their associated radiation fields. As part of a program to develop statistical cloud models for Monte Carlo radiative computations, we are determining the nearest-neighbor spacing distributions of shallow (nonprecipitating) cumulus clouds from a number of digital images of the LANDSAT Multi-Spectral Scanner (MSS) and Thematic Mapper (TM) instruments, which have resolutions of 80 m and 30 m, respectively. Since typical shallow cumulus cloud spacings are a few hundred meters, they are not resolved by meteorological satellites, which have resolutions of 1 km or more. Our results are expressed in terms of a parameter δ which provides a measure of the tendency of clouds to be widely spaced (for $\delta > 0$) or to cluster together (for $\delta < 0$).

One may of course observe cloud clusters even in a completely random cloud field. The eighteenth century astronomer William Herschel worried about a similar problem--that perhaps the large number of "binary" star systems might be random line-of-sight coincidences. The Rev. John Mitchell, a contemporary statistician, showed that the observed number was significantly greater than that expected from random coincidences, and twenty years later Herschel found that many of them had moved in a way consistent with Kepler's laws (Zeilik, 1976).

A related question is: What is the nearest-neighbor spacing distribution of a set of points which have coordinates chosen independently from a uniform distribution? For points on a line, it is given by $\exp[-x]$, where x is the euclidean distance from each point to its nearest-neighbor. This decreases monotonically from its maximum value at zero separation. For points in a plane, the distribution is $2x\exp[-x^2]$, which vanishes at zero separation, rises to a maximum at $x^2 = 1/2$, and then drops rapidly to zero. When one generalizes to a space of arbitrary dimension d, one finds a Weibull distribution with shape parameter d (Onoyama et al., 1984). This has the property that large separations become more likely as the value of d increases.

In section 2 we further generalize this result by relaxing the requirement that the coordinates must be chosen independently. If we take the conditional probability $C(x)$ that there is a point at a distance x from some chosen point, given that there are none closer, to vary as x^δ, then we obtain a Weibull distribution with shape parameter or "effective dimension" $d_{eff} = d + \delta$. When $\delta > 0$ then C increases with separation (cloud "repulsion") and we have $d_{eff} > d$, so that large separations become more likely than they are in the case of independently distributed clouds. Conversely, when $\delta < 0$ then C decreases with x (cloud "attraction" or "clumping") and then $d_{eff} < d$, in which case large separations are suppressed. Section 2 describes a simple (and well-known) method for determining the shape parameter d_{eff}.

In section 3 we describe the LANDSAT data, and the procedures used to extract the cloud coordinates and their nearest-neighbor distances. Section 4 gives initial results for the cloud fields analyzed so far. Finally, in Section 5 we summarize our results and conclusions.

2. SPACING DISTRIBUTION OF RANDOM POINTS

For comparison with the empirical nearest-neighbor distributions of cumulus clouds described in Section 4, we have constructed a one-parameter family of distributions, as follows. First, we ignore the geometry of individual clouds by replacing each cloud by a single point at its center-of-mass. For convenience we refer below to these center-of-mass points simply as "clouds". We also ignore any angular dependence, and assume that all probabilities depend only on the distances separating the various cloud points. Thus in this model there is no preference for cloud streets, for example, though they may appear in a given realization.

A. Derivation of spacing distribution

Although in the LANDSAT data the cloud coordinates are projected onto two dimensions and discretized, it is convenient for this derivation to consider d-dimensional continuous variables. Consider a field of points in a d-dimensional euclidean space, and imagine a spherical shell of internal radius x, thickness dx, and volume dv centered on one of these points. The nearest-neighbor distribution is determined by the joint probability that: (1) there are no clouds within radius x of the given cloud; and (2) there is a cloud in the volume dv between radius x and x + dx. If we assume that this is proportional to the volume element dv, then we can express it as $P_{nn}(x)dv$, where $P_{nn}(x)$ is the probability per unit volume that the nearest neighbor is at x.

The joint probability $P_{nn}(x)dv$ can also be expressed as the product of: (1) the probability $P_o(x)$ that there are no clouds closer than x; and (2) the *conditional* probability that there is a cloud in dv, *given* that there are no clouds closer than x. If we assume that the conditional probability is also proportional to dv, then we have

$$P_{nn}(x)dv = P_o(x) * C(x)dv \qquad (1)$$

where C(x) is the probability per unit volume of finding a cloud at x, given no clouds closer. Now, the probability that there are no clouds closer than x equals the probability that the nearest neighbor lies between x and ∞, so that

$$P_o(x) = \int_x^\infty P_{nn}(x')\, dv' \qquad (2)$$

Taking the derivative of (2) and using (1) gives

$$dP_o/dv = -P_{nn} = -P_o{}^*C, \qquad (3)$$

from which

$$P_o(x) = K \exp[-\int_o^x C(x')dv'], \qquad (4)$$

and therefore we have

$$P_{nn}(x) = K\, C(x) \exp[-\int_o^x C(x')dv']. \qquad (5)$$

Finally, we use the fact that

$$dv = Bx^{d-1}dx, \qquad (6)$$

and

$$P_{nn}dv/dx = p_{nn}, \qquad (7)$$

where p_{nn} is the probability per unit *radius* that the nearest-neighbor is at radius x, so that (5) becomes

$$p_{nn}(x) = AC(x)\, x^{d-1} \exp[-B \int_o^x C(x')\;\; x'^{d-1}dx'], \qquad (8)$$

where A and B are constants determined by the normalization and first moments of p_{nn}.

If we allow the cloud coordinates to be chosen independently, the C(x) is independent of x, and p_{nn} reduces to the Weibull distribution with shape parameter d, a well-known result. (See for example Onoyama *et al*., 1984.) If instead we take

$$C(x) = k\, x^{\delta}, \qquad (9)$$

where δ may be thought of as a "repulsion" parameter measuring the tendency of clouds to prefer large spacings, then (8) takes the form

$$p_{nn}(x) = A\, x^{d_{eff}-1} \exp[-B\, x^{d_{eff}}], \qquad (10)$$

which is a Weibull distribution with shape parameter given by

$$d_{eff} = d + \delta . \qquad (11)$$

B. Determination of the shape parameter

Here we review a well-known method for estimating the shape parameter d_{eff} by a simple least-squares procedure. For a review of estimation procedures for the generalized (threshold-dependent) Weibull distribution, and the related Frechet, Gumbel, and Fisher-Tippett distributions, see Mann (1984) and also Sneyers (1984).

The cumulative distribution associated with (10) has the form

$$\Phi(x) = \exp[-B\, x^{d_{eff}}], \qquad (12)$$

so that

$$p_{nn} = -d\Phi/dx. \qquad (13)$$

Here $\Phi(x)$ is the probability of finding a nearest-neighbor distance greater than or equal to x. It can be estimated empirically by ordering the distances from largest to smallest, then assigning the rank k = 1 to the largest, k = 2 to the second-largest, etc, and finally setting

$$\Phi \sim F = k/N, \qquad (14)$$

where N is the total number of distances in the sample.

From (12) we have

$$\ln(-\ln\Phi) = d_{eff} \ln x + \ln B, \qquad (15)$$

so that if we put

$$Y = \ln(-\ln\Phi), \qquad (16a)$$

$$X = \ln x, \qquad (16b)$$

then (15) takes the form

$$Y = aX + b, \qquad (17)$$

where

$$a = d_{eff}, \qquad (18a)$$

$$b = \ln B. \qquad (18b)$$

According to (16) - (18), if the nearest-neighbor distances are distributed as in (10), then a scatter plot of the logarithm of a sample of N distances versus the double logarithm of the rank divided by N should lie close to a straight line with slope given by the shape parameter d_{eff}.

C. Monte Carlo simulations

In order to test our cloud analysis software, we have analzed some random points in a plane. Figure 1a shows a set of 512 points with coordinates chosen independently from a uniform distribution of integers from 1 to 512. Figure 1b shows each point connected by a straight line to its nearest-neighbor. Points which are closer to any of the four boundaries than to their nearest neighbor must be excluded from the sample. The remaining 456 distances were then sorted from largest to smallest, and a rank assigned. Figure 1c shows a scatter plot with ln(-ln(rank/456)) on the abscissa and ln(distance) on the ordinate. Since this is the inverse of (16), we obtain a least-squares fit line with slope approximately $1/d_{eff} = 0.5$. Distances of only a few pixels are excluded from the fit to minimize the effect of the discretization. The resulting slope is 0.50061. Figure 1d shows the distance histogram, along with the Weibull distribution for the estimated value of d_{eff}.

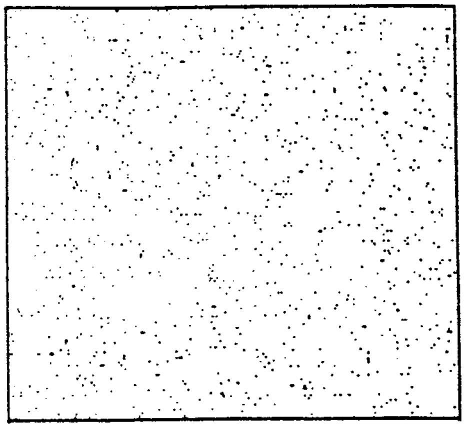

Figure 1a. Set of 512 points with coordinates chosen independently from a uniform distribution from 1 to 512.

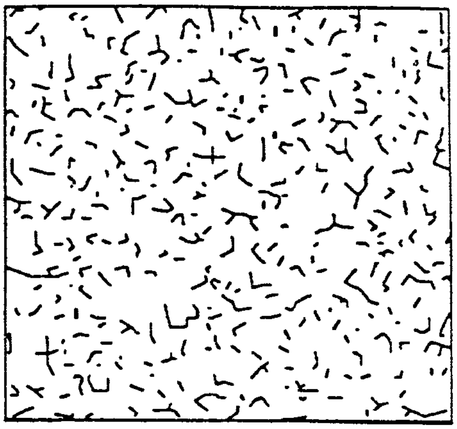

Figure 1b. Same as 1a, but with nearest-neighbors connected.

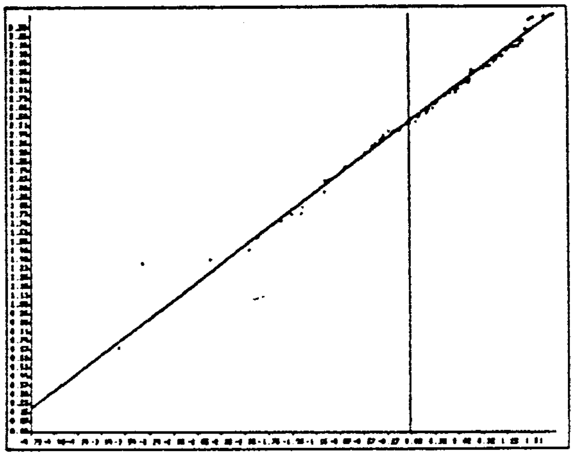

Figure 1c. Scatter plot of logarithm of distances versus double logarithm of rank over number.

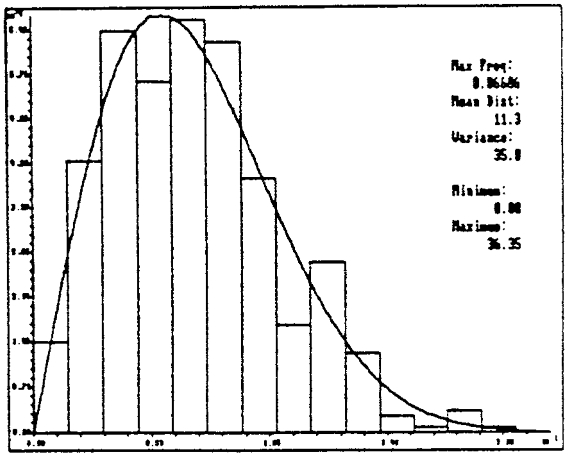

Figure 1d. Histogram of distances with Weibull distribution having shape parameter as in Figure 1c.

3. DATA AND ANALYSIS

Persistent large-scale regimes of shallow cumulus clouds are common in the subtropics over both the Atlantic and Pacific Oceans. (See for example Cahalan et al., 1981.) For our initial study of the spacing distribution, we have selected a LANDSAT TM scene of August 30, 1982 at 31°44'N, 78°54'W off the east coast of the United States. The TM instrument provides 3 reflected visible bands, 3 reflected near-infrared bands, and the water vapor window band of emitted infrared radiation. The reflected bands have a horizontal resolution of about 30 m, and a field-of-view about 185 km on a side. The reflectances and radiances are digitized to 8-bit values from 0 to 255. Since cloud reflectivity is often high enough to saturate the visible bands, we have chosen to study the near infrared band 7 with peak response at 2.05 microns.

The histogram of band 7 radiances shows a sharp peak of low reflectance values asociated with the dark ocean surface, followed by a long flat plateau of higher reflectance values due to clouds. We define the cloudy pixels by choosing a threshold just above the peak, so that the darker cloud edges are included. The digital image of 8 bit reflectances is then replaced by a binary image having 1 for cloudy pixels and 0 elsewhere. Contiguous areas are then identified, and each is given a unique integer identification number (ID). From the ID image, various cloud features such as area, perimeter, and center-of-mass coordinates are determined and stored in an ancillary file. We then compute the distance from each cloud to its nearest-neighbor, and to each of the four edges of the image. Clouds closer to an edge are excluded.

4. RESULTS

Figure 2a shows 294 center-of-mass cloud coordinates from a 1024 by 1024 pixel region in the LANDSAT scene described in the last section. This small sample clearly shows much more clustering than the random point set of Figure 1a. There are many small spacings within each cluster, but also large spacings associated with individual clouds not in a cluster. Figure 2b shows the nearest neighbors connected. Seventeen clouds are closer to an edge, and must be excluded. Figure 2c shows the scatter plot. Distances smaller than 5 pixels were excluded from the least-squares fit. The slope of 0.74611 gives a shape parameter of 1.34, and since $\delta = d_{eff} - 2$ we obtain

$$C(x) = x^{-2/3} \tag{19}$$

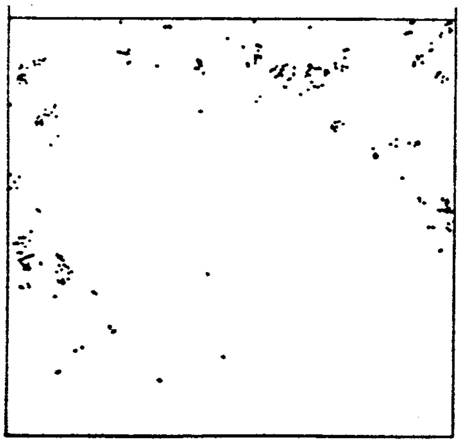

Figure 2a. Set of 293 points taken from LANDSAT scene described in section 3.

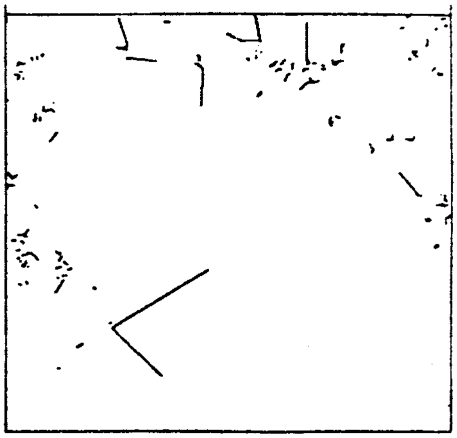

Figure 2b. Same as 2a, but with nearest-neighbors connected.

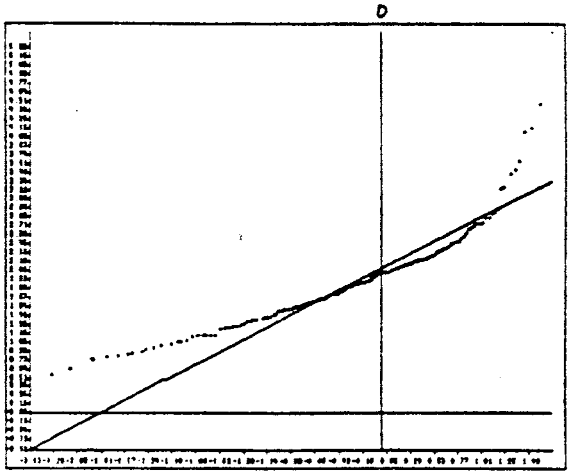

Figure 2c. Scatter plot of distances from Figure 2b versus double logarithm of rank over number.

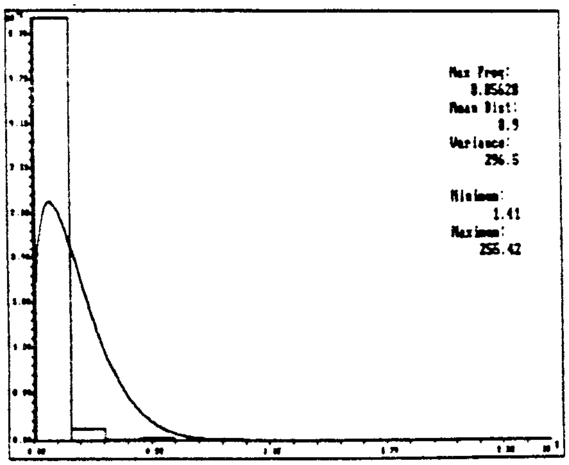

Figure 2d. Histogram of distances with Weibull distribution having shape parameter as in Figure 2c.

Because of cloud clustering, it is clearly necessary to perform this analysis on many samples with a large number of clouds, in order to obtain stable estimates of the large distances between clusters. Judging from Figure 2c, however, it may be that the monotonic form of C(x) assumed here will have to be generalized.

5. SUMMARY AND CONCLUSIONS

We have generalized the well-known result for the nearest-neighbor spacing of random points in d-dimensional euclidean space to allow for dependence between the points. The result (equation 8) is given in terms of a quantity C(x), which is the conditional probability per unit volume of finding a cloud at a distance x, given that there are no clouds closer. When $C \sim x^{\delta}$ then the nearest-neighbor distribution takes the form of a Weibull distribution with a shape parameter given by $d + \delta$ (equations 10 and 11).

The nearest-neighbor distribution of shallow cumulus clouds deviates from a Weibull distribution in situations when there is strong clustering. For such cases the simple power-law form of the conditional probability must be modified.

Randall and Huffmann (1980) have surveyed a number of possible physical explanations of cloud clustering. Their emphasis was on deep convection, and they studied a stochastic model in which cloud "seeds" were distributed independently. We may conjecture that the shallow cumulus cells studied here may be revealing a more appropriate way to distribute the seeds of deep convection.

Acknowledgments: I would like to thank Jon Robinson for developing the software described in sections 2C and 3, and Linda Pylant for typing the manuscript.

6. REFERENCES

Cahalan, R.F., D. A.. Short and G. R. North, 1981: Cloud Fluctuation Statistics, Mon. Wea. Rev., 110, 26-43.

Mann, N., 1984: Statistical Estimation of Parameters of the Weibull and Frechet Distributions, from Statistical Extremes and Applications, J. Tiago de Oliviera, ed., D. Reidel Publishing Co., c1984, pp. 81-89.

Onoyama, T., M. Sibuya, and H. Tanaka, 1984: Limit Distribution of the Minimum Distance Between Independent and Identically Distributed Random Variables, from Statistical Extremes and Applications, J. Tiago de Oliviera, ed., D. Reidel Publishing Co., c1984, pp. 549-562.

Randall, D. A. and G. J. Huffman, 1980: A Stochastic Model of Cumulus Clumping, J. Atmos. Sci., 37, 2068-2078.

Sneyers, R., 1984: Extremes in Meteorology, from Statistical Extremes and Applications, J. Tiago de Oliviera, ed., D. Reidel Publishing Co., c1984, pp. 235-252.

Zielik, M., 1976: Astronomy: The Evolving Universe, Harper and Row, Publishers, Inc., New York, 529pp.