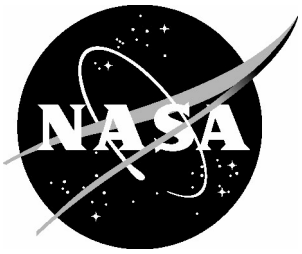
NASA/TP-2005-213274



Comparison of Histograms for Use in Cloud Observation and Modeling

Lisa Green
Middle Tennessee State University, Murfreesboro, Tennessee

Kuan-Man Xu
Langley Research Center, Hampton, Virginia

May 2005

The NASA STI Program Office . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results ... even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at ***http://www.sti.nasa.gov***

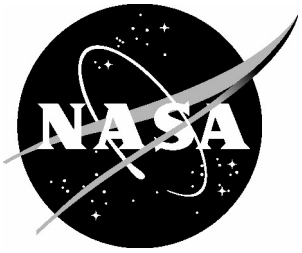
- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA STI Help Desk at (301) 621-0134

- Phone the NASA STI Help Desk at (301) 621-0390

- Write to:
  NASA STI Help Desk
  NASA Center for AeroSpace Information
  7121 Standard Drive
  Hanover, MD 21076-1320

NASA/ TP-2005-213274



Comparison of Histograms for Use in Cloud Observation and Modeling

Lisa Green
Middle Tennessee State University, Murfreesboro, Tennessee

Kuan-Man Xu
Langley Research Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

May 2005

Acknowledgments

This research has been supported by NASA EOS interdisciplinary study program. The authors would like to thank Drs. Zach Eitzen, Don Garber, Takmeng Wong, and Mr. Gary Gibson for reading the preliminary draft of this report.

Available from:

NASA Center for AeroSpace Information (CASI)
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161-2171
(703) 605-6000

Abstract

Cloud observation and cloud modeling data can be presented in histograms for each characteristic to be measured. Combining information from single-cloud histograms yields a summary histogram. Summary histograms can be compared to each other to reach conclusions about the behavior of an ensemble of clouds in different places at different times or about the accuracy of a particular cloud model.

As in any scientific comparison, it is necessary to decide whether any apparent differences are statistically significant. The usual methods of deciding statistical significance when comparing histograms do not apply in this case because they assume independent data. Thus, a new method is necessary. The proposed method uses the Euclidean distance metric and bootstrapping to calculate the significance level.

Introduction

Cloud observation and cloud modeling create large amounts of data. The information from observation and modeling data can be presented as a histogram for each characteristic or parameter of a given cloud. Combining information from several different clouds, whether real or modeled, requires combining these histograms into a summary histogram. Summary histograms, which represent signals of a large data sample, can then be compared to each other to reach conclusions about the behavior of an ensemble of clouds in different places at different times or about the accuracy of a particular cloud model.

As in any scientific comparison, it is necessary to decide whether any apparent differences are statistically significant. The usual methods of deciding statistical significance when comparing histograms do not apply in this case because they assume the data are independent; thus, a new method is necessary. In this study, the proposed method is to choose a distance metric and use bootstrapping to calculate the significance level. Details of this method are described in this report.

Satellite Data

Observations of a cloud, either in a computer model or through satellite remote sensing, consist of measurements made in a grid, the points of which are referred to as "footprints" in satellite remote sensing and grid boxes in cloud modeling. Several different quantities are measured or inferred at each footprint of every cloud: solar insolation, short-wave reflected radiation, albedo, cloud optical depth, ice water path, cloud ice diameter, liquid water path, cloud droplet radius, outgoing long-wave radiation, emissivity, cloud top temperature, cloud top height, cloud top pressure and sea surface temperature. Because the sizes of the satellite-observed clouds vary greatly, the total number of footprints measured in a cloud could be either fairly small or quite large. The data set used here, which consists of measurements made by the Clouds and the Earth's Radiant Energy System (CERES; Wielicki et al., 1996) instrument on the Tropical Rainfall Measuring Mission (TRMM) satellite during March 1998 contained 352 clouds varying in size from 74 footprints to 6883 footprints, with a mean of 545 footprints. Each footprint is typically 10–15 km in diameter.

The CERES instrument observes the various quantities at each footprint, and histograms (recorded in a separate file for each cloud) summarize these values. The coded file names contain information about the date and location of the cloud. Because these "single-cloud" histograms contain data that were measured in the same cloud system, the data may not be independent. For example, if one footprint measures the cloud height at 15 km, it is likely that nearby footprints also have similar values for this measurement. Thus, a histogram reporting cloud top height that has a large number of observations in the bin centered at 15.25 km is more likely to have observations in the neighboring bins centered at 14.75 km and 15.75 km, provided that the chosen bin size is 0.5 km.

These single-cloud histograms are typically not studied alone but are combined with histograms from other clouds with certain common attributes. It is these summary histograms that are compared with each other. For example, 100 clouds from one geographic region could be combined into a histogram, which would be compared to a similar histogram summarizing 150 clouds from another geographic region. Model simulations of clouds could be compared to satellite observations of the same clouds, or clouds from one month/year could be compared to clouds from another month/year in the same geographic region.

While the satellite footprint observations within each cloud are dependent, this study assumes that different clouds are independent of each other. An argument could be made that the clouds are dependent based on similar weather dynamics over a large geographic region where many clouds are developed and maintained, for example, but any dependency involved would be difficult to quantify. It should be noted that if the bootstrapping procedure described subsequently is used to analyze data in which the different clouds are clearly dependent, it will fail to yield meaningful results.

Measuring Differences in Histograms

Several methods for evaluating the differences between two summary histograms were examined. They included standard goodness of fit tests from the statistical literature as well as several distance measures. The goodness of fit tests required independent assumptions and, as mentioned previously, the individual cloud histograms lack such independence. It is possible that a Chi-squared goodness of fit test could be modified to apply, but the exact nature of the modifications is a topic for future study. Thus, it was decided to use a measure of distance between histograms as a statistic.

Several distance measures were examined. They were chosen because they had been used with good results in other applications that require comparison of histograms, for example, image retrieval, remote sensing, or object tracking. The characteristics of each measure were examined by comparing the behavior for simple test histograms similar to those in figure 1 below. Through this process, either the L_2 measure or the Jeffries-Matusita (JM) distance (also called the Hellinger distance) was used.

The L_2 measure is the typical Euclidean distance between two vectors, which is defined by

$$L_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

where a_i and b_i are the proportion of values in the ith bin of the respective histograms. If the histogram is reported as an approximate probability distribution function, in which the area contained by the histogram is one, this proportion (either a_i or b_i) is found by multiplying the bin height by the bin width. It is possible to define a_i and b_i as the bin heights without multiplying by the width. Such a change

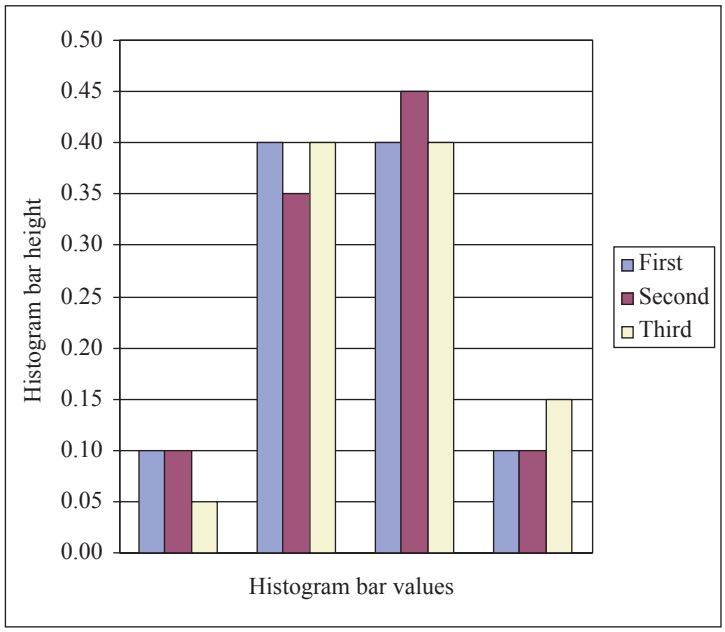
would affect the value of L_2 but would not change the relative scale of smaller L_2 values to larger values. However, if the area contained by the histogram is one, the maximum value L_2 or JM, defined subsequently, can achieve is $\sqrt{2}$. If desired, these values can be divided by $\sqrt{2}$ to ensure that the maximum value is one.

The Jeffries-Matusita distance is defined by

$$JM = \sqrt{\sum_i \left(\sqrt{a_i} - \sqrt{b_i}\right)^2}$$

It has been used in applications, such as image retrieval, in which it is necessary to find small differences in data. Note that both these formulas assume that the histograms have bins with equal widths.

Figure 1 demonstrates some of the differences in the behavior of these two distances. According to the L_2 distance, the second and third histograms are equally distant from the first histogram. However, when one uses the Jeffries-Matusita distance, the third histogram is farther away from the first than the second histogram is. This extra distance demonstrates that in the Jeffries-Matusita distance, differences in small bins are more important than differences in large bins, while in the L_2 distance, it is simply the magnitude of the difference that matters.



Distance measure	Distance between first and second	Distance between first and third
L_2	0.071	0.071
Jeffries-Matusita (Hellinger)	0.056	0.117

Figure 1. Behavior of L_2 distance versus Jeffries-Matusita distance for idealized histograms.

Bootstrapping Procedure

The formulas discussed previously define distances between histograms, but they do not address the question of statistical significance; that is, does the difference in the summary histograms imply that they came from two different populations? Bootstrapping will address this question. A short introduction to bootstrapping is included in appendix A, and a copy of a computer program that implements the algorithm discussed in this section is given in appendix B. It is important to note that it is possible for

bootstrapping to yield incorrect results. For example, it is possible to conclude that there is a difference in the underlying populations when it is not actually true. However, this fact is true of any statistical procedure.

The CERES/TRMM data from March 1998 were saved in 352 files. Each file contains histograms of data about a particular cloud and is named in a manner that describes the date and location of that cloud. The various clouds were each classified as being in the Eastern, Western, or Central Tropical Pacific regions. In the algorithm described subsequently, Region 1 denotes one of the three regions, while Region 2 denotes one of the remaining regions. The process for comparing Region 1 to Region 2 clouds is as follows:

1. Make lists of file names for Region 1 and 2 clouds. In this case, there are 88 clouds classified as being in Region 1 and 135 classified as in Region 2.

2. Merge the lists of file names for Regions 1 and 2. There are 223 clouds on the new list. Under the null hypothesis, these clouds come from the same population; thus, the choice of which clouds were in Region 1 and which were in Region 2 was equivalent to a random choice from the merged list under the null hypothesis.

3. Choose 88 file names to represent the "Random 1" contingent of clouds by randomly sampling with a replacement from the list of 223 file names. Do the same to choose 135 clouds for the "Random 2" contingent of clouds.

4. Create summary histograms for the new sets of "Random 1" and "Random 2" clouds and calculate the values of the distance measures.

5. Compare the bootstrapped distance value between the "Random 1" and "Random 2" clouds to the distance value calculated for the true arrangement of the clouds. If the new value is larger than the true value, add one to a counter. Repeat steps 3–5 for a total of 5000 iterations.

6. Divide the value of the counter by 5000. If this proportion is small, less than 5 percent for a 95 percent confidence level, we have evidence that there is a difference between the cloud populations. If desired, the bootstrapped values can be stored and graphed to allow visualization of the true value location compared to the bootstrapped values.

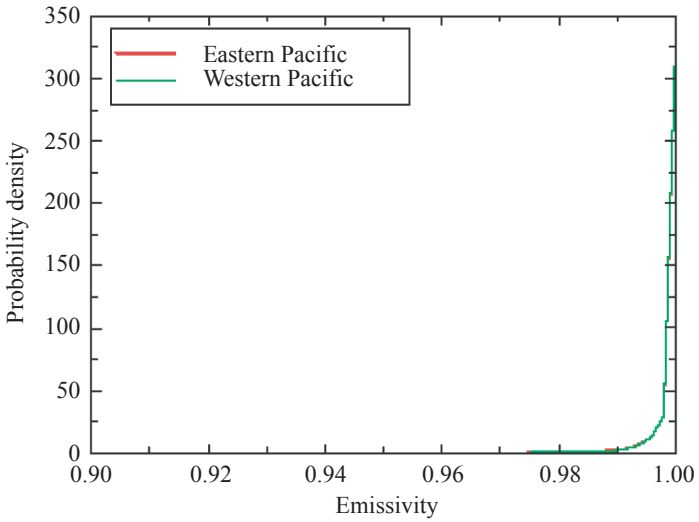
This algorithm tests the null hypothesis that all the clouds in the files are from the same population. If this hypothesis is true, then the clouds in Region 1 and Region 2 are essentially equivalent. Therefore, the distance between the histograms for the "true" ordering is essentially a random number picked from the sampling distribution of the bootstrapped distances.

In the algorithm, the proportion of bootstrapped distances that are greater than the "true" distance is calculated. If the true distance is a random choice, as implied by the null hypothesis, this proportion could be any value between zero and one. A very small value for this proportion is evidence that the true distance was not a random choice from the sampling distribution. The proportion will be called the approximated significance level (ASL), and a value less than 0.05 will be evidence against the null hypothesis.

Results

The algorithm previously described was implemented by using the March 1998 CERES/TRMM data. Details of this data set are given at <http://cloud-object.larc.nasa.gov>. They are not described here because this report focuses on the methodology for comparing the differences between summary histograms and presenting statistically significant tests. Results presented in this section compare clouds in the Eastern Pacific to clouds in the Western Pacific. Similar comparisons between the Western and Central Pacific clouds and between the Eastern and Central Pacific clouds are not presented.

Figure 2 shows the histograms (left panel), measured distances, and ASLs (right panel) of emissivity in the Eastern and Western Pacific. Emissivity is a measure of how strongly a body radiates energy and has a value between zero and one (Wallace and Hobbs, 1977). The histograms for the two locations overlap to such an extent that it is difficult even to see the line representing the Eastern Pacific. The measured distances agree, yielding very small values of L_2 and Jeffries-Matusita. The corresponding ASLs are quite large, an example of a situation in which we can clearly conclude that there is no difference between the populations represented for the distribution of emissivity.

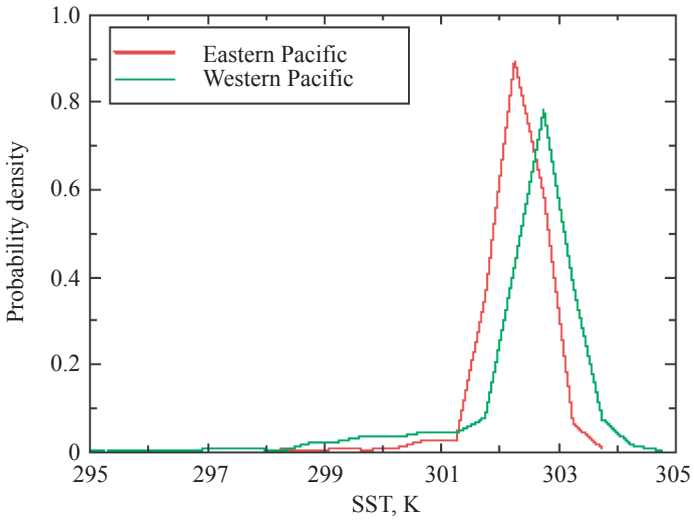


L_2	0.008
ASL for L_2	0.684
JM	0.027
ASL for JM	0.740

Figure 2. Histograms (left panel), measured distances, and the approximated significance levels (ASLs) (right panel) of emissivity for Eastern and Western Pacific clouds during the March 1998 period.

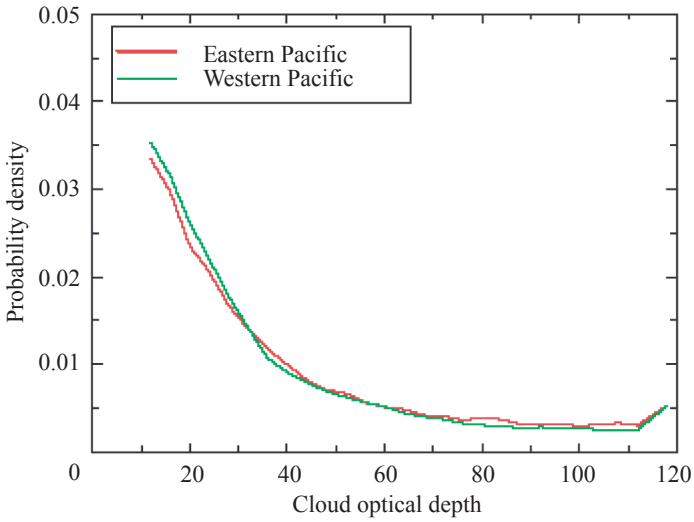
Figure 3 shows the histograms of sea surface temperature (SST) associated with the clouds in the eastern and western Pacific in March 1998. The histograms have similar shapes and are very concentrated about their modes; however, the mode for the Eastern Pacific clouds is 302.25 K, while the mode for the Western Pacific clouds is 302.75 K. Also, the range of SSTs for the Eastern Pacific clouds is 298.25–303.75 K, while the range for the Western Pacific clouds is much larger, from 292.25–304.75 K. This difference is reflected in fairly large values for L_2 and Jeffries-Matusita but is shown much more starkly in the extremely low ASL values. Thus, we can conclude that there is a difference in SSTs between the clouds in the two regions.

Figures 4 and 5 show the histograms and numeric information for cloud optical depth and cloud ice diameter, respectively. Optical depth measures the depletion of a beam of radiation as a result of passing through a cloud layer (Wallace and Hobbs, 1977). The L_2 values for these graphs are both 0.027.



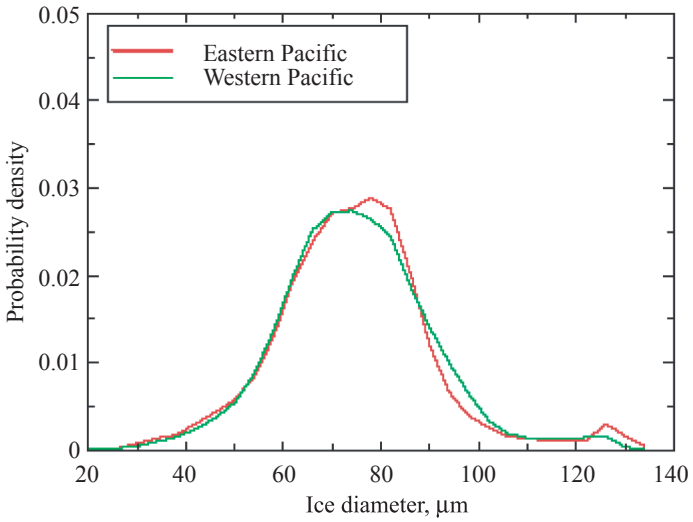
L_2	0.340
ASL for L_2	0.001
JM	0.488
ASL for JM	0.0002

Figure 3. Histograms (left panel), measured distances, and ASLs (right panel) of sea surface temperature (SST) for Eastern and Western Pacific clouds during the March 1998 period.



L_2	0.027
ASL for L_2	0.196
JM	0.052
ASL for JM	0.224

Figure 4. Histograms (left panel), measured distances, and ASLs (right panel) of cloud optical depth for Eastern and Western Pacific clouds during the March 1998 period.



L_2	0.027
ASL for L_2	0.700
JM	0.087
ASL for JM	0.596

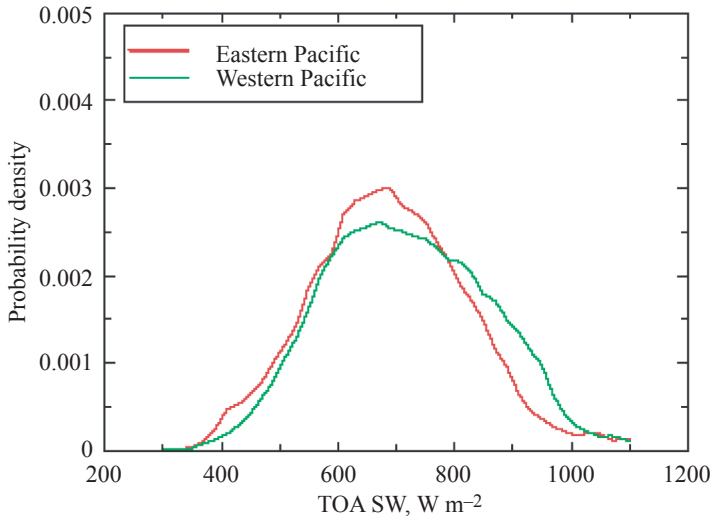
Figure 5. Histograms (left panel), measured distances, and ASLs (right panel) of cloud ice diameter for Eastern and Western Pacific clouds during the March 1998 period.

However, the corresponding ASLs are not similar to each other. The ASL for cloud optical depth is 0.196, while the ASL for ice diameter is 0.700. This difference is a result of differences in the amount of variation in the original data and demonstrates that the significance level does not depend on the numeric value of the corresponding statistic, but on the relative size of the statistic in relation to the bootstrapped values. The ASL for the JM distance is also much higher in the ice diameter (>100 percent) than in cloud optical depth, although the JM distances are only different by less than 50 percent between these two quantities, further supporting the assertion that the ASL value is not directly related to the statistic.

Figures 6–10 demonstrate further examples of histograms with their corresponding L_2 and Jeffries-Matusita values and the ASLs yielded by the bootstrapping algorithm. In each of these cases, the statistical evidence does not imply a difference in the underlying populations, although visual inspections, which may misidentify the areas covered by the two curves as the distances, may suggest otherwise in some of them. There are two explanations for this apparent disagreement. One is that the data samples are not large enough. The data sample sizes are extremely small in liquid water path and cloud droplet radius because the cloud top heights of most of the cloud footprints are too high to contain the liquid phase clouds. Another reason is that the definitions of the L_2 and JM distances are *not* equivalent to the areas contained between the two PDFs, which are observed by visual inspections.

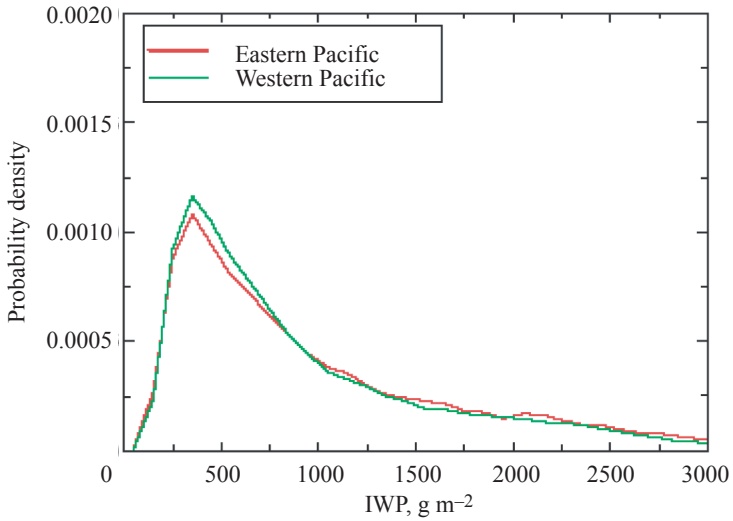
The histograms of top-of-the-atmosphere (TOA) solar insolation shown in figure 11 are an example of disagreement between the L_2 statistic and the Jeffries-Matusita statistic. In this case, the ASL generated with the L_2 statistic implies that there is no difference between the underlying populations. However, the ASL generated with the Jeffries-Matusita statistic does imply a significant difference that is probably related to the large JM value (0.519) and the nearly single-point PDFs with rather different modes of TOA solar insolation for individual clouds. The randomized populations may produce small JM values when PDFs with similar modes are combined, which results in fewer smaller bins than in the true distribution of cloud populations.

Figures 12–14 also demonstrate disagreement between the L_2 and JM distances. However, in these cases it is the significance level generated by the L_2 statistic that suggests that there is a difference in the



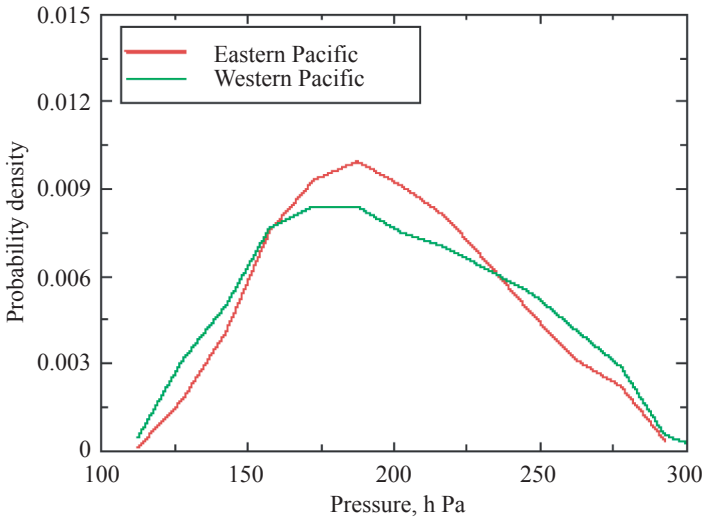
L_2	0.036
ASL for L_2	0.298
JM	0.144
ASL for JM	0.302

Figure 6. Histograms (left panel), measured distances, and ASLs (right panel) of shortwave reflected radiation flux for Eastern and Western Pacific clouds during the March 1998 period.



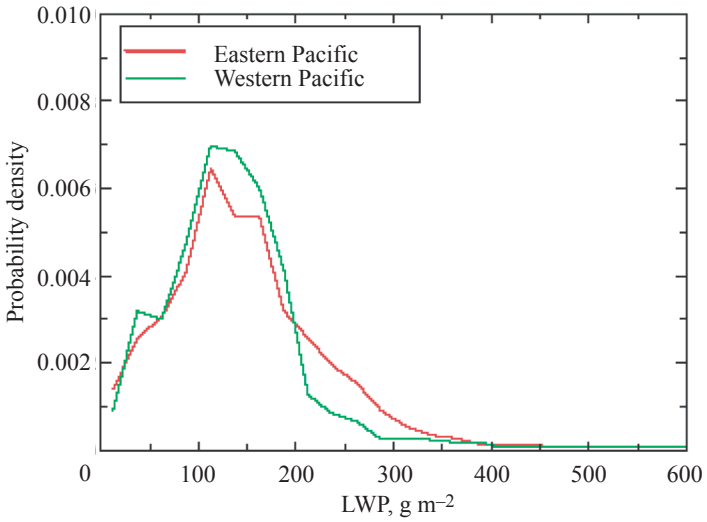
L_2	0.024
ASL for L_2	0.276
JM	0.056
ASL for JM	0.358

Figure 7. Histograms (left panel), measured distances, and ASLs (right panel) of ice water path (IWP) for Eastern and Western Pacific clouds during the March 1998 periods.



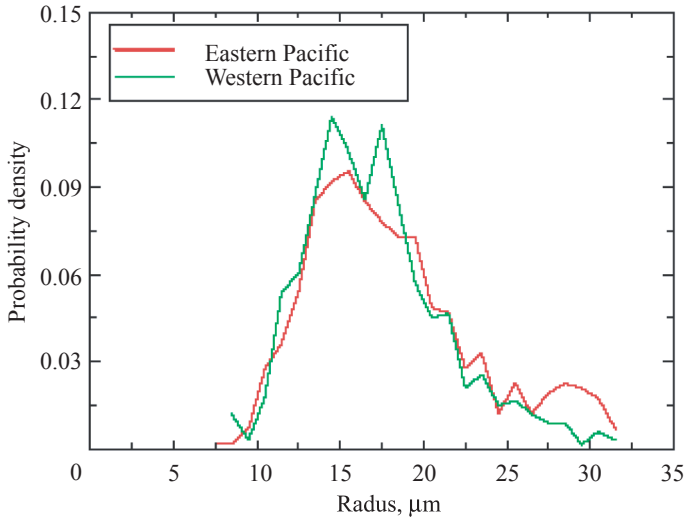
L_2	0.054
ASL for L_2	0.072
JM	0.106
ASL for JM	0.156

Figure 8. Histograms (left panel), measured distances, and ASLs (right panel) of cloud top pressure for Eastern and Western Pacific clouds during the March 1998 period.



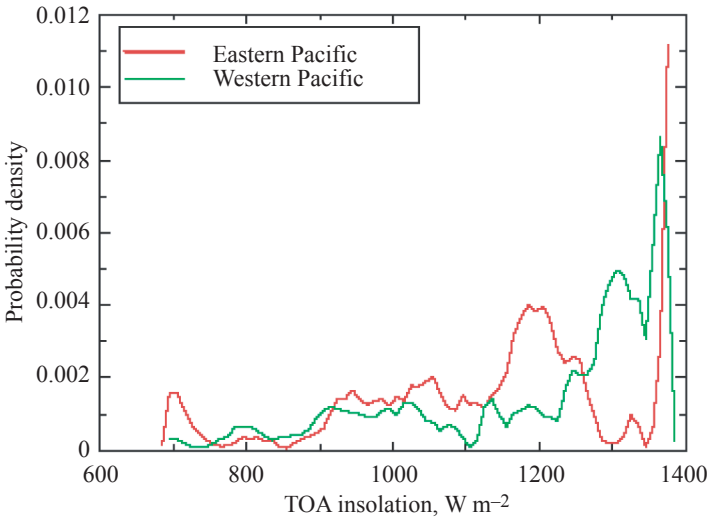
L_2	0.081
ASL for L_2	0.543
JM	0.203
ASL for JM	0.554

Figure 9. Histogram (left panel), measured distances, and ASLs (right panel) of liquid water path for Eastern and Western Pacific clouds during the March 1998 period.



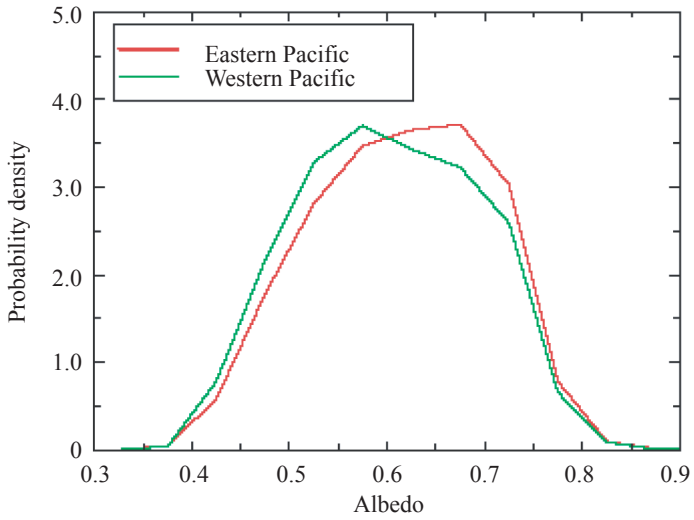
L_2	0.093
ASL for L_2	0.4344
JM	0.245
ASL for JM	0.339

Figure 10. Histograms (left panel), measured distances, and ASLs (right panel) of cloud droplet radius for Eastern and Western Pacific clouds during the March 1998 period.



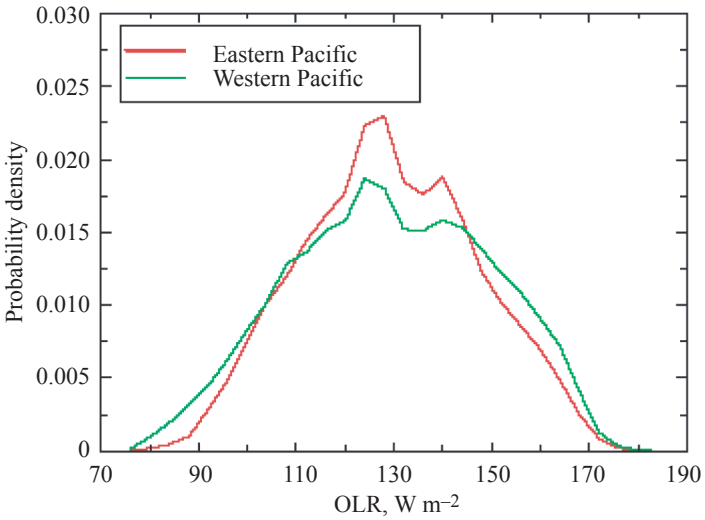
L_2	0.146
ASL for L_2	0.125
JM	0.519
ASL for JM	0.018

Figure 11. Histograms (left panel), measured distances, and ASLs (right panel) of top-of-the-atmosphere (TOA) solar insolation for Eastern and Western Pacific clouds during the March 1998 period.



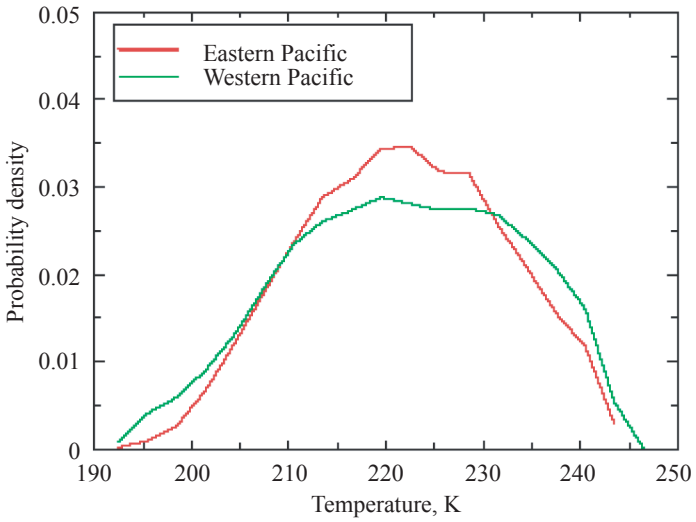
L_2	0.060
ASL for L_2	0.024
JM	0.085
ASL for JM	0.070

Figure 12. Histograms (left panel), measured distances, and ASLs (right panel) of TOA albedo for Eastern and Western Pacific clouds during the March 1998 period.



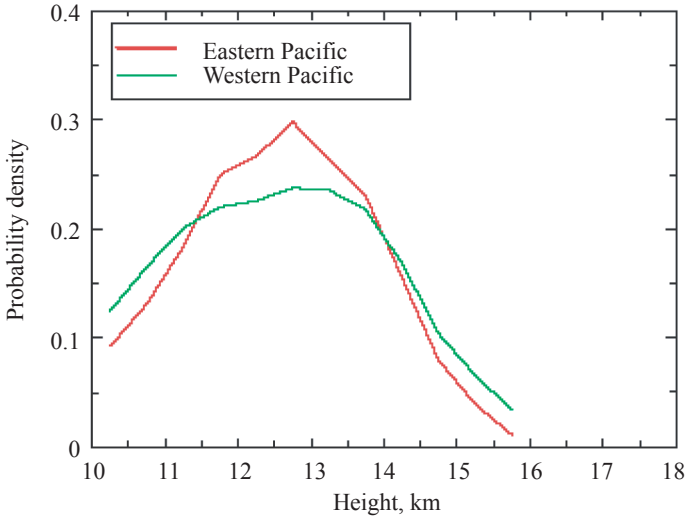
L_2	0.045
ASL for L_2	0.033
JM	0.120
ASL for JM	0.107

Figure 13. Histogram (left panel), measured distances, and ASLs (right panel) of outgoing longwave radiation for Eastern and Western Pacific clouds during the March 1998 period.



L_2	0.049
ASL for L_2	0.041
JM	0.115
ASL for JM	0.121

Figure 14. Histograms (left panel), measured distances, and ASLs (right panel) of cloud top temperature for Eastern and Western Pacific clouds during the March 1998 period.



L_2	0.057
ASL for L_2	0.050
JM	0.107
ASL for JM	0.141

Figure 15. Histograms (left panel), measured distances, and ASLs (right panel) of cloud top height for Eastern and Western Pacific clouds during the March 1998 period.

populations. The significance levels generated by the Jeffries-Matusita statistic do not support the conclusion that the populations are different. Figure 15 is included in this group because the approximated significance level for the L_2 statistic in this case is only slightly larger than 0.05. Technically, an ASL that is larger than 0.05 does not support the hypothesis that two populations are different.

The final choice of which statistic to use should be based on each of their behaviors in distinguishing histograms. Figures 11–15 will help in this endeavor because they are examples in which the two statistics yield different conclusions. The choice of which is more accurate should depend on the nature of the properties being measured. The difference between the Eastern and Western Pacific solar insolation data was determined to be significant by the Jeffries-Matusita distance, but not by the L_2 distance. Solar

insolation measures the amount of energy coming from the Sun. It changes based on time of year and latitude but does not change based on longitude (Wallace and Hobbs, 1977). Conversely, the difference between the Eastern and Western Pacific data for albedo, outgoing longwave radiation, cloud top temperature, and cloud top height was determined significant by the L_2 distance, but not by the Jeffries-Matusita distance. It would be reasonable for these quantities to differ based on location because of the climatological contrast between the two regions. During the El Niño period, the occurrence of cloud systems shifted from the Western Pacific to the Central and Eastern Pacific (Cess et al., 2001), but certain properties can still be different in the two regions. Based on these observations, it is recommended that the L_2 distance be used rather than the Jeffries-Matusita distance.

Concluding Remarks

Information about individual clouds is stored in histograms that can be combined to create summary histograms. These summary histograms can then be compared to summary histograms from other places, other times, other observational methods, or model simulations. It is necessary, therefore, to design a method to determine whether any differences in these histograms are statistically significant.

One can determine statistical significance by choosing a statistic and then calculating the approximated significance level by using a bootstrapping procedure. It is important to note that statistical significance is determined by comparing histograms to other histograms that are generated by the same type of data. That is, comparing the distances between histograms for emissivity to the distances between histograms for solar insolation does not yield any important information. Instead, the approximated significance levels can be compared, if desired.

Comparing the behavior of two distance measures under the bootstrapping procedure leads to the recommendation that the L_2 statistic (the typical Euclidean distance between two vectors) be used rather than the Jeffries-Matusita distance. The graphs in which the two measures differ were examined, and the Jeffries-Matusita distance suggested that differences in solar insolation between Eastern Pacific and Western Pacific clouds were significant. The L_2 measure did not agree with this conclusion. The properties of solar insolation argue against any difference in longitude having an effect on the value. Other graphs in which the L_2 measure suggested significance and the Jeffries-Matusita measure did not were examined. In these cases, it was reasonable to assume that longitude could affect the values of the quantities being measured.

Therefore, it is recommended that the L_2 statistic be combined with a bootstrapping procedure to compare summary histograms to each other. Once the bootstrapping procedure has been applied, the approximated significance level that is generated can be compared to a desired significance level, for example 0.05, to reach a conclusion about whether the underlying populations differ.

References

Efron, B. and Tibshirani, R.: *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.

Hogg, R. and Craig, A.: *Introduction to Mathematical Statistics, Fifth Edition*, Prentice-Hall, New Jersey, 1995.

Wallace, J. and Hobbs, P.: *Atmospheric Science: An Introductory Survey*, Academic Press, New York, 1977.

Wielicki, B. A.; Barkstrom, B. R.; Harrison, E. F.; Lee, R. B. III; Smith, G. L.; and Cooper, J. E.: 1986: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment. *Bull. Amer. Meteor. Soc.,* **77**, 853–868.

Cess, R. D.; Zhang, M.; Wang, P.-H.; and Wielicki, B. A.: 2001: Cloud Structure Anomalies Over the Tropical Pacific During the 1997/98 El Niño. *Geophys. Res. Lett.*, **28**, 4547–4550.

Appendix A

An Introduction to Bootstrapping

Most statistical actions, such as finding confidence intervals or performing hypothesis tests, rely on knowing the distribution of the test statistic. For example, if we have collected numerical data that are normally distributed, we can show that the sample mean is normally distributed and the sample standard deviation, properly normalized, is distributed like a Chi-squared random variable. These facts are responsible for several standard statistics formulas.

What if the distribution of the statistic is unknown? This lack of information could result if the distribution of the data is unknown or if the formula for calculating the statistic is complicated. Often, a distribution for the data is assumed anyway to simplify further work. This assumption, however, invites criticism of the final results and does not address complicated calculations.

If we know the distribution of the data, we can approximate the distribution of the statistic by simulating the experiment many times and recording the value of the statistic for each trial, and the set of values for the statistic can be used to estimate any required quantity. This method simplifies the analysis for complicated calculations but does not address the question of whether we know the distribution of the data.

Bootstrapping is based on the fact that the empirical distribution function—that is, a function that puts equal weight on each of the sample data points—forms an estimate of the distribution function of the data. To approximate the distribution of the test statistic, we will sample randomly with replacement *from the data*, calculate the value of the statistic, and repeat.

An example may clarify the process. Suppose the following 15 numbers were gathered during an experiment:

0.726	0.712	0.401	0.892	0.621
0.902	0.556	0.020	0.612	0.346
0.819	0.539	0.330	0.950	0.687

The sample mean calculated from these data is $\bar{X} = 0.6075$. What is a 95-percent confidence interval for the mean? If we can assume that the data are normally distributed, we can use the standard formula of $\bar{X} \pm 1.96\, s/\sqrt{n}$. This formula arises because the standard error—the standard deviation of the statistic—is $s/\sqrt{n}$, and 1.96 standard deviations to the left and right of the center encloses 95 percent of the area of a normal curve. However, we have no way of verifying the normality assumption, and graphing these data in a histogram yields a graph that does not look like the familiar bell-shaped curve. Instead, we will approximate the distribution of $\bar{X}$ and take the numbers which enclose 95 percent of the area.

To approximate the distribution of $\bar{X}$, first sample with a replacement from the previous 15 numbers. For example, the following numbers could result:

0.902	0.950	0.539	0.612	0.621
0.621	0.712	0.330	0.712	0.950
0.621	0.401	0.020	0.556	0.687

Notice that while some of the numbers are repeated, there are no numbers that were not in the original sample. The new sample mean is $\bar{X}_1 = 0.6156$. Resampling 1000 times yields a set of sample means, which are plotted in the following histogram (see fig. A1):

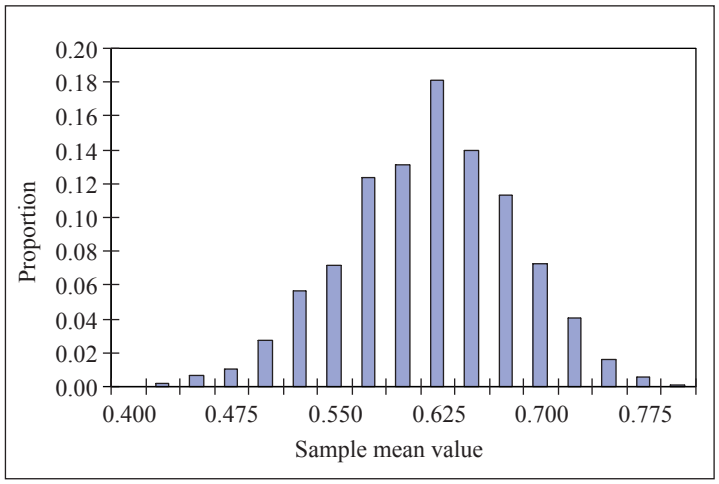


Figure A1. Histogram of sample means.

Sorting the values $\bar{X}_1,\ldots,\bar{X}_{1000}$ allows us to see that 95 percent of the values are in the interval (0.4773, 0.7227), which then serves as a confidence interval for μ. (The reader may notice that the distribution of $\bar{X}$ looks like a normal distribution. Figure A1 is a demonstration of the Central Limit Theorem and will not necessarily be the case for other test statistics.)

To demonstrate the use of bootstrapping in hypothesis testing, suppose we have another set of data, in addition to the first set, and the goal is to determine whether they came from the same distribution:

0.726	0.712	0.401	0.892	0.621
0.902	0.556	0.020	0.612	0.346
0.819	0.539	0.330	0.950	0.687

0.153	0.024	0.806	0.705	0.001
0.001	0.531	0.060	0.392	0.720
0.890	0.207	0.006	0.461	0.196

The null hypothesis is that these two sets of data are from the same distribution, and the alternate hypothesis will be that they are not. For the sake of argument, suppose we have chosen $\bar{X} - \bar{Y}$ as a test statistic. Then, for these data, $\bar{X} - \bar{Y} = 0.6075 - 0.3435 = 0.2640$. Does this example show a statistically significant difference?

Under the null hypothesis, the two samples shown in the previous charts are from the same distribution, so merging the samples yields an approximation of the underlying distribution. First, choose two sets of 15 numbers each by sampling with replacement from the entire set of 30 numbers. For example:

0.001	0.819	0.153	0.819	0.153
0.556	0.006	0.020	0.712	0.020
0.539	0.401	0.001	0.529	0.001

0.153	0.401	0.806	0.556	0.902
0.024	0.806	0.006	0.207	0.006
0.461	0.330	0.401	0.001	0.806

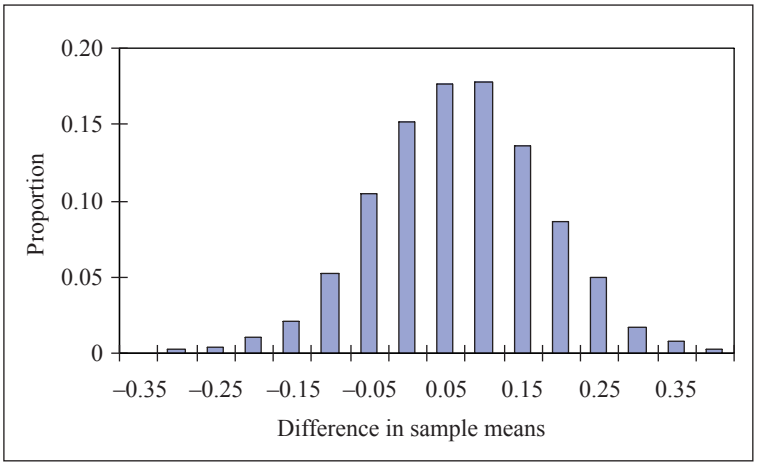


Figure A2. Histogram of differences.

The difference in sample means for this choice of numbers is $\bar{X}_1 - \bar{Y}_1 = -0.0750$. Repeating this exercise 1000 times yields values $\bar{X}_1 - \bar{Y}_1, \ldots, \bar{X}_{1000} - \bar{Y}_{1000}$, which are collected in figure A2.

There are 7 values greater than 0.2640 and 13 values less than -0.2640, so $p = 20/1000 = 0.02$. Assuming a significance level $\alpha = 0.05$, the fact that the p-value is smaller than α implies that the difference is statistically significant. The null hypothesis that the two sets of data came from the same distribution is rejected. In effect, if the two sets of data had come from the same population, the original number we obtained would likely be close to the center of this distribution. Since it is in one of the tails of the distribution, we have evidence that the assumption that led to the creation of this distribution must be incorrect.

Appendix B

Program Code

This appendix is a copy of a FORTRAN 77 program that implements the bootstrapping algorithm described in the text. The program assumes that it will be run in the same folder or directory as files containing formatted information about clouds. It also assumes that the names of these files contain a number as the 19th character, which describes whether the cloud is from the Eastern, Central, or Western Pacific. Finally, it assumes that a file named `'filenames.txt'` exists that lists the file names in the directory. Such a file can easily be created in DOS with the command `c:>dir S* > temp.txt`, or in UNIX with the command `>ls S* | cat temp.txt`.

This program took about three hours to run. Most likely, the bulk of the time is used to open and close files. If so, the run time can be reduced by storing the values in the files at the beginning of the program run and then referring to the stored values rather than rereading the files.

```
      program eastwest
C     ***** variable declarations *****
      parameter (numfiles=352, maxbin=400, numiter=5000)
      character*24 tempname, east(numfiles),
     +       west(numfiles)
      integer eastindex, westindex,
     +       eastlength, westlength,
     +       easti(numfiles), westi(numfiles)
      real eastloc(maxbin, 14), aa1, bb1, binsize(14),
     +       eastcount(maxbin, 14), eastbintotal(14),
     +       westloc(maxbin, 14), westbintotal(14),
     +       westcount(maxbin, 14),
     +       L2(14), JM(14), L2boot(14),
     +       JMboot(14), pL(14), pJ(14)


      ivar=1
      eastindex=1
      westindex=1


C     *****read in list of filenames *****
C     *****requires file containing a list of all filenames*****
C     *****also requires the 19th character in the filename to *****
C     *****be a number, 1,2, or 3, which signifies east, west, or central****

      open (unit=20, file='filenames.txt',
     +     form='formatted',access='sequential')


50    read (20,*,end=100) tempname
```

```
      if (tempname(19:19).eq.'1') then
            east(eastindex)=tempname
            eastindex=eastindex+1
      elseif (tempname(19:19).eq.'2') then
            west(westindex)=tempname
            westindex=westindex+1
      elseif (tempname(19:19).eq.'3') then
C           central(centralindex)=tempname
C           centralindex=centralindex+1
      else
            print *, 'error!'
            go to 100
      endif


      ivar=ivar+1

      go to 50


100   continue

      close (20)

      length=ivar-1
      eastlength=eastindex-1
      westlength=westindex-1

C     print *, length, eastlength, westlength


C     *****EAST*****
C     print *, 'East'
C     *****Initialize variables*****

      do ivar=1,14
            do n=1,maxbin
                 eastcount(n, ivar)=0.0
            enddo
            eastbintotal(ivar)=0
      enddo


C     ***** Read in data from files *****

      do jvar=1,eastlength

       open(unit=21, file=east(jvar),
     +      form='formatted',access='sequential')
```

```
      nvar=1


125   read(21, *, end=150) nbin, aa1, bb1, lbin1

C     ***** Note, line below should be if (aa1 .ne. -999.00), but I got error
C     ***** messages for using not equal with a real variable.

      if (aa1 .le. -999.01 .or. aa1 .ge. -998.99) then

          eastloc (nbin, nvar) = aa1
          eastcount (nbin, nvar) = eastcount (nbin, nvar) +
     +                              float(lbin1)

      else
          eastbintotal (nvar) = eastbintotal (nvar) + lbin1
          binsize (nvar) = bb1

          nvar=nvar+1
      endif

      go to 125
150   continue

      close(21)

     enddo


C    *****WEST*****

C    print *, 'West'
C    *****Initialize variables*****

     do ivar=1,14
            do n=1,maxbin
                 westcount(n, ivar)=0.0
            enddo
            westbintotal(ivar)=0
     enddo

C    ***** Read in data from files *****

     do jvar=1,westlength

      open(unit=21, file=west(jvar),
    +        form='formatted',access='sequential')
```

```
      nvar=1

225   read(21, *, end=250) nbin, aa1, bb1, lbin1

C     ***** Note, line below should be if (aa1 .ne. -999.00), but I got error
C     ***** messages for using not equal with a real variable.

      if (aa1 .le. -999.01 .or. aa1 .ge. -998.99) then

           westloc (nbin, nvar) = aa1
           westcount (nbin, nvar) = westcount (nbin, nvar) +
     +                              float(lbin1)
      else
           westbintotal (nvar) = westbintotal (nvar) + lbin1
           binsize (nvar) = bb1

           nvar=nvar+1
      endif
      go to 225

250   continue

      close(21)

     enddo


C    *****Calculate metrics*****

     do ivar=1,14
      L2(ivar)=0.0
      JM(ivar)=0.0

      do n=1,maxbin
          L2(ivar)=L2(ivar) +
     +       (eastcount(n, ivar)/eastbintotal(ivar) -
     +        westcount(n, ivar)/westbintotal(ivar))**2
          JM(ivar)=JM(ivar) +
     +       (sqrt(eastcount(n, ivar)/eastbintotal(ivar)) -
     +        sqrt(westcount(n, ivar)/westbintotal(ivar)))**2

      enddo

      L2(ivar)=sqrt(L2(ivar))
      JM(ivar)=sqrt(JM(ivar))

C     write (6,*) ivar, ' L2:',L2(ivar), ' JM:', JM(ivar)
```

```
C         write (6,*) '-999.00', binsize(ivar)
      enddo

C     *****Now to do the random number set-up *****

      call date_time_seed@

      print *, 'random'

      do ivar=1,14
            pL(ivar)=0
            pJ(ivar)=0
      enddo
      do kvar=1,numiter
C      print *, kvar

       do jvar=1,eastlength
        easti(jvar)=nint(random()*(eastlength+westlength)+0.5)
C       write(6,*) easti(jvar)
       enddo

       do jvar=1,westlength
        westi(jvar)=nint(random()*(eastlength+westlength)+0.5)
C       write(6,*) westi(jvar)
       enddo

C      ***** "EAST" *****
C      print *, 'East'
C      *****Initialize variables*****

       do ivar=1,14
            do n=1,maxbin
                 eastcount(n, ivar)=0.0
            enddo
            eastbintotal(ivar)=0
       enddo

C      ***** Read in data from files *****

       do jvar=1,eastlength

C       write(6,*) easti(jvar), east(easti(jvar))

        if (easti(jvar) .le. eastlength) then
            open(unit=21, file=east(easti(jvar)),
    +        form='formatted',access='sequential')
        else
            open(unit=21, file=west(easti(jvar)-eastlength),
    +        form='formatted',access='sequential')
```

```
        endif

        nvar=1

425     read(21, *, end=450) nbin, aa1, bb1, lbin1

C       ***** Note, line below should be if (aa1 .ne. -999.00), but I got error
C       ***** messages for using not equal with a real variable.

         if (aa1 .le. -999.01 .or. aa1 .ge. -998.99) then

             eastloc (nbin, nvar) = aa1
             eastcount (nbin, nvar) = eastcount (nbin, nvar) +
    +                                 float(lbin1)

         else
            eastbintotal (nvar) = eastbintotal (nvar) + lbin1
            binsize (nvar) = bb1

            nvar=nvar+1
         endif

         go to 425

450      continue

         close(21)

        enddo


C       ***** "WEST" *****

C       print *, 'West'
C       *****Initialize variables*****

        do ivar=1,14
             do n=1,maxbin
                   westcount(n, ivar)=0.0
             enddo
             westbintotal(ivar)=0
        enddo

C       ***** Read in data from files *****

        do jvar=1,westlength

         if (westi(jvar) .le. eastlength) then
             open(unit=21, file=east(westi(jvar)),
```

```
     +         form='formatted',access='sequential')
          else
             open(unit=21, file=west(westi(jvar)-eastlength),
      +        form='formatted',access='sequential')
          endif

          nvar=1

525       read(21, *, end=550) nbin, aa1, bb1, lbin1

C        ***** Note, line below should be if (aa1 .ne. -999.00), but I got error
C        ***** messages for using not equal with a real variable.

          if (aa1 .le. -999.01 .or. aa1 .ge. -998.99) then

              westloc (nbin, nvar) = aa1
              westcount (nbin, nvar) = westcount (nbin, nvar) +
     +                                 float(lbin1)
          else
             westbintotal (nvar) = westbintotal (nvar) + lbin1
             binsize (nvar) = bb1

             nvar=nvar+1
          endif
          go to 525

550       continue

          close(21)

         enddo


C        *****Calculate metrics*****

         do ivar=1,14
          L2boot(ivar)=0.0
          JMboot(ivar)=0.0

          do n=1,maxbin
           L2boot(ivar)=L2boot(ivar) +
     +      (eastcount(n, ivar)/eastbintotal(ivar) -
     +       westcount(n, ivar)/westbintotal(ivar))**2
           JMboot(ivar)=JMboot(ivar) +
      +     (sqrt(eastcount(n, ivar)/eastbintotal(ivar)) -
      +      sqrt(westcount(n, ivar)/westbintotal(ivar)))**2

          enddo
```

```
          L2boot(ivar)=sqrt(L2boot(ivar))
          JMboot(ivar)=sqrt(JMboot(ivar))

          if (L2boot(ivar) .gt. L2(ivar)) then
              pL(ivar)=pL(ivar)+1
          else
          endif

          if (JMboot(ivar) .gt. JM(ivar)) then
              pJ(ivar)=pJ(ivar)+1
          else
          endif

         enddo

        enddo


C       ***** Output results *****

        print *, 'output'

        do ivar=1,14
            write(6,*) ivar, ' L2:', L2(ivar),
     +       ' pvalue:', pL(ivar)/numiter
            write(6,*) ivar, ' JM:', JM(ivar),
     +       ' pvalue:', pJ(ivar)/numiter
            write(6,*) ' '
        enddo


        end
```

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01- 05 - 2005	Technical Publication	

4. TITLE AND SUBTITLE

Comparison of Histograms for Use in Cloud Observation and Modeling

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Green, Lisa and Xu, Kuan-Man

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

229-01-02-10

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER

L-19051

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSOR/MONITOR'S ACRONYM(S)

NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

NASA/TP-2005-213274

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category 47
Availability: NASA CASI (301) 621-0390

13. SUPPLEMENTARY NOTES

Green, Middle Tennessee State Univ., Mufreesboro, TN. Xu, Langley Research Center, Hampton, VA.
An electronic version can be found at http://ntrs.nasa.gov

14. ABSTRACT

Cloud observation and cloud modeling data can be presented in histograms for each characteristic to be measured. Combining information from single-cloud histograms yields a summary histogram. Summary histograms can be compared to each other to reach conclusions about the behavior of an ensemble of clouds in different places or at different times or about the accuracy of a particular cloud model. As in any scientific comparison, it is necessary to decide whether any apparent differences are statistically significant. The usual methods of deciding statistical significance when comparing histograms do not apply in this case because they assume independent data. Thus, a new method is necessary. The proposed method uses the Euclidean distance metric and bootstrapping to calculate the significance level.

15. SUBJECT TERMS

Histogram; Cloud modeling; Satellite observation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Help Desk (email: help@sti.nasa.gov)
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	30	(301) 621-0390