Visualizing Distributions from Multi-Return Lidar Data to Understand Forest Structure

David Kao[1], Marc Kramer[2], Alison Luo[3], Jennifer Dungan[2] and Alex Pang[3]
[1]NASA Advanced Supercomputing Divison and [2]Earth Science Division, NASA Ames Research Center
[2]Computer Science Department, UCSC
David.L.Kao@nasa.gov, kramerm@fsl.orst.edu, {alison,pang}@soe.ucsc.edu, Jennifer.L.Dungan@nasa.gov

***Abstract*— Spatially distributed probability density functions (pdfs) are becoming relevant to the Earth scientists and ecologists because of stochastic models and new sensors that provide numerous realizations or data points per unit area. One source of these data is from multi-return airborne lidar, a type of laser that records multiple returns for each pulse of light sent towards the ground. Data from multi-return lidar is a vital tool in helping us understand the structure of forest canopies over large extents. This paper presents several new visualization tools that allow scientists to rapidly explore, interpret and discover characteristic distributions within the entire spatial field. The major contribution from this work is a paradigm shift which allows ecologists to think of and analyze their data in terms of the distribution. This provides a way to reveal information on the modality and shape of the distribution previously not possible. The tools allow the scientists to depart from traditional parametric statistical analyses and to associate multimodal distribution characteristics to forest structures. Examples are given using data from High Island, southeast Alaska.**

I. INTRODUCTION

Historically, scientists have relied on the use of statistical descriptors such as the mean, median, standard deviation, skewness and kurtosis to describe and compare the populations they are sampling. While these statistical measures work well for describing and comparing unimodal distributions, they often fail to capture the nature and dynamics of multimodal distributions. Yet multimodal distributions commonly occur in populations found in the natural environment. Multimodal distributions create a challenge for visualization because they are less easy to summarize and thereby render with a few variables or parameters. In a geospatial context, multimodal distributions provide an even greater problem. In this context, a distribution may exist at many locations in a spatial field, such as at every cell in a grid. If even some of the distributions are multimodal and cannot be summarized easily, the whole field cannot be rendered using conventional approaches.

In this paper, we develop and apply visualization techniques and tools in a new way: to visualize, query and compare distributions of earth science data on a grid. The data we have chosen to use is of a new type of increasing interest to ecologists. These data come from lidar (Light Detection And Ranging) instruments, airborne remote sensors that measure vegetated surfaces such as forests. Such measurements are collected to gain a detailed understanding of the canopy structure across an entire study area, rather than at a few select plots. Distribution data on the height of the vegetation canopy results from the collection of multiple observations over a fixed area. The distribution data in this study were derived from raw multi-return lidar data of forest and provide information on forest structure, tree size and density. Forest plots recovering from natural disturbance tend to have unimodal distributions of stem sizes and canopy heights with low standard deviations, whereas older, less disturbed forest plots tend to have multimodal distributions [5].

To better understand the forest canopy distributions derived from lidar data, we have developed some new visualization tools. These tools were selected because they allow scientists to query their data in new ways in order to better understand the distributions of ecological phenomena, both at single locations and across the spatial domain. Such visualization tools allow for exploration, interpretation and discovery and extraction of characteristic distributions. These tools may also be applied to visualize other ecological phenomena for which detailed spatially explicit distribution data exist or can be derived. The scientists' job is to associate ecological meaning to these distributions. Such meaning can be derived from field reconnaissance, expert knowledge or ancillary information. Among the science questions these tools can help answer are:

1) What new scientific insights can be gained from working with distribution data?
2) How does exploring, probing and performing query through visualization tools enhance the understanding of the distribution data?

A key contribution of this paper is the development of new visualization tools that allow scientists to continually perceive and explore their data in terms of the distribution, rather than through coarse statistical descriptors, or clustering algorithms, as was previously the case. Thus new and important information about the nature and dynamics of the distribution can be captured both visually and quantitatively.

II. BACKGROUND

The challenge to visualizing spatially explicit, multimodal distributions is the four dimensional nature of the problem. To consider probability density functions (pdfs) over space, two dimensions are the orthogonal spatial dimensions, a third is the variable scale (in this case the height scale given by lidar) and the fourth is the frequency scale. Previously, we have reported on techniques for visualizing 4D spatial distribution data sets [3] using parametric statistics. That is, the pdf at

every cell is characterized by a few statistical parameters such as mean, standard deviation, skewness, etc. and visualized. When some of the pdfs, particularly in mixed forest areas, have multimodal distributions, statistical summaries are not sufficient. To address this, we have also developed shape-based descriptors for distributions [4]. The basic idea here is to describe the shape of a distribution using information about the number of modes, the location of the modes, the width and height of each mode, etc. This descriptive information is then mapped to visual parameters. We demonstrate how that approach can be brought to bear on the lidar data in Section V.

More generally, we proposed an operator-based approach to visualizing spatial distribution data sets [7]. The main idea here is to treat distributions as first-class objects with their own set of methods and operations. This allows one to compare, add, subtract, etc. two distributions. More complex operations can also be constructed from these simpler operations. The benefit of this approach is that standard visualization techniques such as contour lines, isosurfaces, streamlines, etc. can be extended to support the new data type.

Previous efforts to visualize lidar data [2], [9] presented ways in which a user can navigate through forest lidar data sets within a virtual environment. This is essentially the creation of a digital elevation model of the canopy top. Unlike this approach, our approach looks at aggregated multiple lidar returns. Therefore the data at each cell location is actually a collection of height values. In this study, we visualized distributions from 0.1 hectare cells, the size of field plots for which forest stand measures exist. The techniques developed in [3], [4], [7] is brought to bear upon this problem.

III. DATA

Forest canopy height distribution data were collected using a multi-return lidar system. The system, the digital airborne topographic imaging system (DATIS-2; 3-Di Technology, MD, USA) is a small footprint lidar. The sensor is capable of retrieving multiple (up to 5) returns of elevation and intensity for every shot as it passes through a forested canopy. Over wooded terrain, the first return measures forest canopy height, while the last return measures ground elevation. The laser fires at a rate exceeding 4000 pulses per second and scans across the aircraft flight path (see Figure 1). Since the speed of light is known, the reflection time of the laser light back to the aircraft is measured, allowing the distance to the terrain surface to be calculated. To locate the elevation points, the latitude and longitude of the aircraft are recorded with a high accuracy (< 1 cm) Global Positioning System (GPS). The accuracy of point locations is further increased by compensating for the the aircraft's attitude (pitch, roll and yaw), measured using an inertial measurement unit. DATIS-2 was flown in a Cessna 206 in May 2001. The data were initially collected at a density exceeding 2 shots per m^2. Raw data were processed into 81 measures of maximum forest canopy height for each 0.1 hectare cell across the island, resulting in 1800 0.1 grid cells with distribution data for each cell.

The data were collected above High Island (approximately 500 hectares), which is located in the middle of the Alexander Archipelago (Figure 2). The maximum elevation on the island is 150 m. The parent material is fractured basalt. Average annual precipitation is 1.9 m, with the wettest months during fall and winter. Extreme temperature fluctuations are infrequent due to the maritime influence. Cloud cover, precipitation, cool ambient air temperatures (4-10 °C), and high relatively humidity (80%) are characteristic throughout the year. The island is dominated by productive western hemlock (*Tsuga heterophylla* (Raf.) Sarg.) with scattered Sitka spruce (*Picea sitchensis* (Bong.) Carr.).

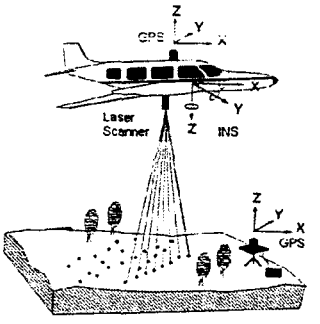

Fig. 1. Airborne lidar data acquisition

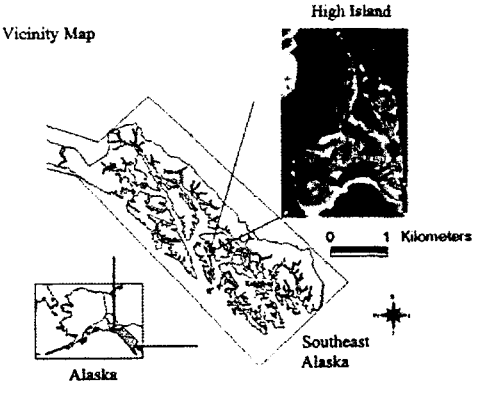

Fig. 2. Vicinity map of High Island in Alaska where the lidar data were collected

IV. ALGORITHMS

Prior to visualization, algorithms are applied to the raw data to estimate and characterize their distributions. In particular, density estimation is used to generate a probability density function from the 81 heights at each grid cell, a peak hunting algorithm is used to find all the modes in the pdfs, and an operator is selected to allow distribution matching.

A. Density Estimation

For each grid cell in the field, there are multiple lidar returns, each with an associated height. These represent a sample of the full set of heights of all elements in the canopy. We use each sample to make an estimate of the "true" density, that is, the distribution of the full set of heights. One

common density estimator, the histogram, does not produce a mathematically valid pdf and is very sensitive to the bin width used. There are many other estimators possible depending on the nature of the data [8]. In this application with lidar data, we selected a kernel estimator because it provides robust density estimation and is widely used.

A kernel estimator is given by

$$f(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{t-z_i}{h}), \tag{1}$$

where $f(t)$ is the pdf estimated from the data values $\{z_i, i = 1, \ldots n\}$, h is a smoothing parameter, and the kernel function $K(t)$ satisfies the following property:

$$\int_{-\infty}^{\infty} K(t)\,dt = 1 \tag{2}$$

If the kernel function $K(t)$ is C^1 continuous, then kernel estimators are also C^1 functions; this is in contrast to histograms which are C^0 continuous. For this application, we used the Gaussian kernel function:

$$K(t) = \frac{1}{\sqrt{2\pi}}e^{-(1/2)t^2} \tag{3}$$

Kernel estimators are also influenced by h which controls the overall smoothness of the estimate. As h decreases, the kernel estimator becomes more sensitive to slight variations in the distribution; as h increases, contributions from more neighboring points are coalesced to form a smoother estimate. If h is not chosen appropriately, the shape of the estimate can vary significantly and even change the modality of the pdf (e.g. from unimodal to bimodal). So, rather than letting the user specify the value of h, a data-dependent h can be derived [8]:

$$h = 0.9 \times \ min(std.\ dev., interquartile\ range/1.34)\ n^{-\frac{1}{5}}. \tag{4}$$

For the lidar data, we compute the kernel density estimate $f_{ij}(t)$ for each grid point (i, j) using the data-dependent smoothing factor given in Equation 4. Thus, the smoothing factor h would vary across the field depending on the data values at each grid point.

B. Mode Finding

We used the following algorithm which we proposed in [4]. Given a pdf obtained from a density estimator, our goal is to determine the number of peaks in the distribution and their respective positions. The distribution may be very bumpy, in other words it has many local maxima. First, we compute all the local maxima in the distribution. By our definition, a local maximum is an interval [a,b] such that the density estimate is concave over [a,b], but not over any larger interval [8]. We refer to the local maximum as a *basic peak*. The height of a peak is defined as the vertical distance between the local maximum and one of the two ends of the interval with higher density. A peak with height higher than a user-defined threshold is considered significant. A concatenated peak, as its name implies, includes at least two basic peaks. We further

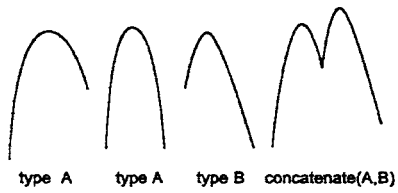

Fig. 3. Basic and concatenated type A and type B peaks.

classify both kinds of peaks into two types, type A and type B (See Figure 3). Type A peaks have the minimum density at the start of the interval and type B peaks have the minimum density at the end of the interval. If the start and end of the interval have the same density, then the peaks are classified as type A. We use the following rules for concatenating the peaks.

These two types of basic peaks may be concatenated to form combined peaks according to the following rules:

$$concatenate(A, A) = A \tag{5}$$

$$concatenate(A, B) = \begin{cases} A & \text{if the start of A} \leq \text{the end of B} \\ B & \text{otherwise} \end{cases} \tag{6}$$

$$concatenate(B, B) = B \tag{7}$$

The concatenate(B, A) produces no new peaks. These operations apply to both basic and concatenated peaks. The peaks to be concatenated must be adjacent to each other and concatenation only takes place if at least one of the two adjacent peaks is not significant. Then we iteratively loop through all the basic peaks and determine if the adjacent peaks can be concatenated into a concatenated peak. The iterations stop when there is no concatenation possible, that is, all pairs of adjacent peaks are considered as significant peaks. Finally, we count and record the locations and heights of these significant peaks.

C. Distribution matching

We used an operator that compares distributions in order to find ones that have shapes that match a distribution of interest. The operator returns a scalar value that indicates how similar the two distributions are. The operator applied for this purpose is the Kullback-Leibler (KL) operator [1], [6]. Let P and Q be two distributions of variable z to be compared. The Kullback-Leibler distance is defined as follows:

$$KL(P, Q) = \int_{-\infty}^{+\infty} P(z)\log\frac{P(z)}{Q(z)}dz \tag{8}$$

The greater the KL distance is, the less similar the two distributions are. We set a threshold to control how similar we want the search results to be to a distribution of interest or target distribution. All the distributions with a distance less than the threshold to the target distribution will be accepted

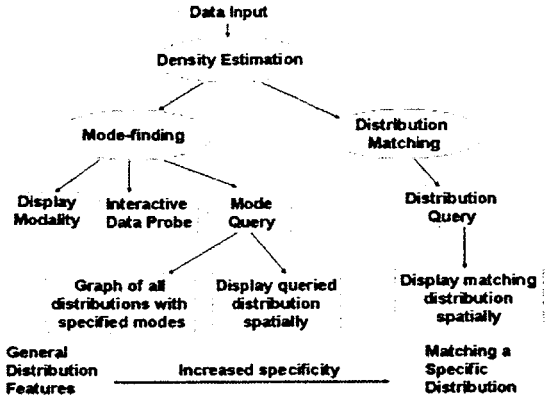

Fig. 4. Flow chart of the visual analysis of distribution data made possible by the techniques described in this paper. Ovals denote algorithms. Rectangles denote the techniques proposed in this paper. The axis from left to right shown near the bottom of the diagram shows how the techniques showcase general to increasingly specific features of the distributions.

as similar distributions to the target distribution. By relaxing the threshold, a larger number of similar distributions will be identified.

V. VISUALIZATION TOOLS

In the following subsections, we describe several visualization techniques designed to provide capabilities ranging from synoptic, general views of the full data set to more specific, localized and detailed query and display. The techniques allow extensions well beyond summary descriptors such as the quadratic mean, robust mean and chi-square or other "non-parametric" summaries or clustering algorithms. Briefly, they are a map display of the number of modes for each distribution, an interactive data probe, mode exploration (finding locations that have distributions with modes specified by the user, finding distributions that have a specified number of modes and graphing distributions that have specific modes) and distribution exploration (matching complete distribution shapes). For each technique, we describe how it can assist the scientist in exploring distributions. We apply these tools to distributions of canopy height derived from the lidar data, focusing on three characteristic distributions that are of particular interest to the scientist. Collectively, these tools can be used effectively to analyze distribution data. Figure 4 depicts relationships among the techniques, starting from the number of modes (which many different distributions may have in common), to querying a specific mode in a distribution (which a smaller subset will share), and finally to matching a distribution, which another, possibly even smaller subset of distributions will share. So, the queries work with increasing specificity.

It is envisioned that these visualization tools will provide scientists with new understanding of their data, previously not possible. The extraction features provided with many of these tools will allow creation of new data sets from specific distributions, or portions (e.g. modes) of distributions, that are of interest.

A. Displaying the modality of the distribution data

The first look at the spatially explicit distribution data shows the number of modes for each cell, calculated using the algorithm described in Section IV-B (Figure 5). This synoptic descriptor of all the data across the sample space provides the scientist with the first glimpse of new information related to the distribution. The modality of each cell gives some indication of patterns of multimodality. This is the first new information about the distribution that is not available through coarse statistical descriptors. This display helps answer questions such as what proportion of the data is unimodal or multimodal? Are the number of modes spatially clustered or concentrated in any one subregion of the field?

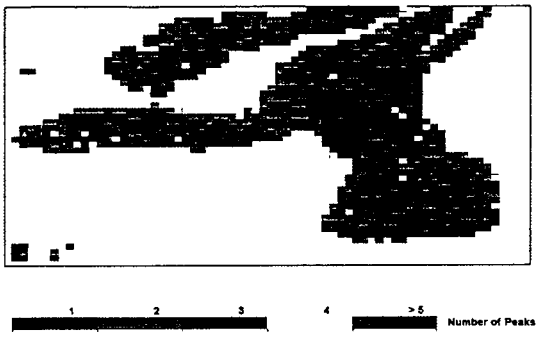

Fig. 5. The spatial locations of unimodal and complex multimodal distributions in the High Island lidar data.

B. Interactive Data Probe

We have implemented an interactive data probe that allows the user to view the distribution of an individual cell at the current probe position set by the user. The interactive data probe is straightforward and useful for visualizing the pdf at any location in the field. It provides a per point basis query and shows the modality of the distribution. Only one density estimate is displayed at a given time. To begin gaining familiarity with the data the scientist can probe the forest data at different locations in order to have a good overall feel for the distributions in the study area. In addition, when viewing the distribution of a given cell, adjacent cells can be selected (through an up, down, left and right keyboard feature), thus allowing the scientist to visually traverse portions of the forest. This feature allows the scientist to view and relate distributions of particular forest regions of interest that s/he might already be familiar with through field reconnaissance or other ancillary data.

Figure 6 shows the kernel density estimate and the histogram at the probe position shown on the right image of the figure. The image is colored by the mean canopy height.

C. Mode Exploration

The modes of a distribution can be explored in a variety of other ways using mode exploration tools. For all of these

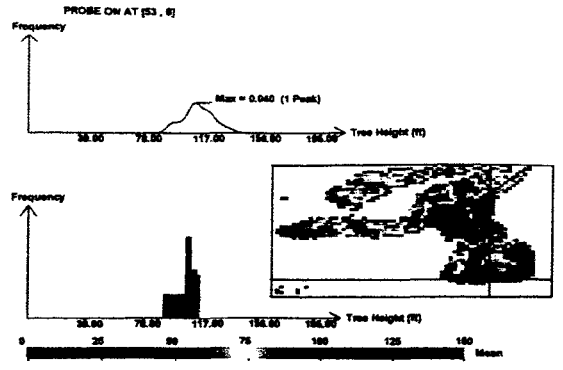

Fig. 6. The image on the right shows the mean canopy height of the lidar distribution data at each grid cell location. The bottom plot is a histogram of the data values making up the distribution at the cross-hair. At the probe position, the tree height is approximately 117 ft, which is denoted by the black vertical line in the histogram. The top plot is a kernel estimate of those values.

tools, the mode is computed using the peak hunting algorithm as described in Section IV-B. Meaningful mode exploration depends on using the proper density estimator, so that shapes in the distribution are real and not an artifact of low sampling density in the data. Conversely, it is important that all real information in the distribution be retained, so the smoothing function must minimize the loss of real information contained in the distribution. Our mode exploration tools comprise of the following processes: (1) mode query, (2) visualizing the distributions from the results of a mode query, and (3) visualizing the distributions and the spatial locations from the results of a mode query. Each of these processes is described in more detail in the following sections.

1) Mode query: Mode query allows the scientist to specify queries that show: (1) the abundance, (2) the spatial location (possibly corresponding to ancillary information about that area or prior knowledge), (3) the spatial extent and (4) the spatial structure (e.g. random vs clustered) of distributions with a specific mode. This feature allows all the distributions that contain a particular mode of interest to be identified and displayed. For example, after using the interactive data probe and relating the distributions observed with field observations, the scientist became interested in finding all unimodal distributions with a mode between 117 and 194 feet. The mode query tool now allows for visual identification of all those distributions that match the query. Figure 7 shows the locations of the forest that have unimodal distributions with a canopy height mode between 117 to 194 feet and with the minimum density value $M = 0.05$. These locations are grid cells denoted with black squares. It is not surprising that the mean field of the distribution data at these grid cells are relatively high as indicated (in the red and magenta color range in Figure 7). For reference, the topography of the data are shown in Figure 8. Our tool allows the user to interactively change the range of values where the mode should lie within and the minimum density value M, along with the number of modes for the entire distribution. When any one of these parameters changes, a new query is generated and the tool updates the screen by highlighting grid cells that satisfy the query.

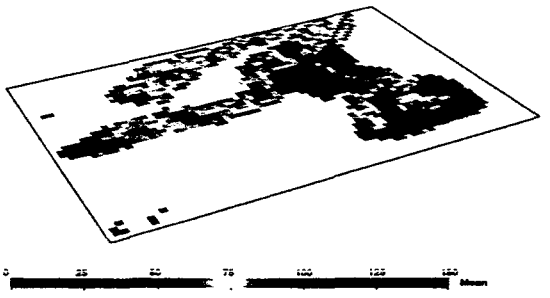

Fig. 7. The image plane shows the mean canopy height for each cell in the field colored by six classes according to the key. The cells marked by black squares denote those locations found by the following query: (1) the distribution has a mode (peak) between the heights of 117 and 194 feet, and (2) the frequency of the distribution is above 0.05 (or 5%) of the canopy height measurements.

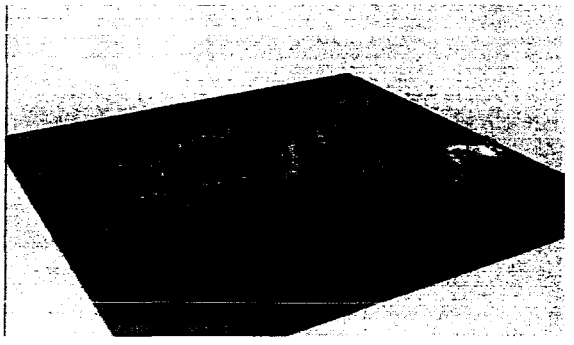

Fig. 8. A graphical model of the High Island forest from the lidar data. Individual trees in the forest are represented by the tree-like icons.

2) Visualizing the distributions from a mode query: From a mode query, there may be tens or hundreds of grid cells that match a specified mode criterion. The visualization challenge here is how to display all of these density estimates effectively, so that the scientists can begin to explore the shapes and diversity of the distributions identified through the query. Furthermore, these similarities and differences need to be highlighted for analysis e.g. for those pdfs that are "very" different, the information about which grid cells these pdfs represent should be shown/highlighted. Similarly, the grid cells of those conforming pdfs (pdfs that are similar) should be clustered or colored in the same group. If the pdfs are very different, then the scientist would like to know how they differ and where they differ.

The most common approach to view several pdfs is to simply plot them side by side for visual comparison. This can be done by plotting a set of pdfs, or as many pdfs that can possibly fit on the screen in multiple windows. If the query only found a few pdfs, then this method is ideal and effective for comparing these pdfs. However, if there were tens

or hundreds of pdfs found by the query, the user would need to view so many graphs as to make this method impractical.

Another simple approach is to plot all of the pdfs in one single graph, giving the scientist a visual comparison of these pdfs. This method is only useful, however, if the scientist is interested in determining whether there are any pdfs that differ significantly from others. The scientist would be able to see the overall shapes of these pdfs using this approach. However, for more detailed comparisons of pdfs, this method would not be suitable since it is most likely that many pdfs would overlap in the graph which makes it difficult to distinguish the details. Figure 9 shows a graph of the pdfs that matched the query used in Figure 7. By displaying all of these pdfs in the same graph, the overall shape of these distributions can be seen to be very close.

Visualizing the distributions identified through a mode query not only allows the scientist to inspect them, it also allows the scientist to look for the following features: (1) outliers (how are outliers shaped, how many modes do they have, etc.), (2) trends (how are most of the distributions shaped, how many modes do they have), (3) diversity (how different are they from one another), (4) homogeneity (how similar), and (5) modality of the distributions (1, 2, 3 or more peaks).

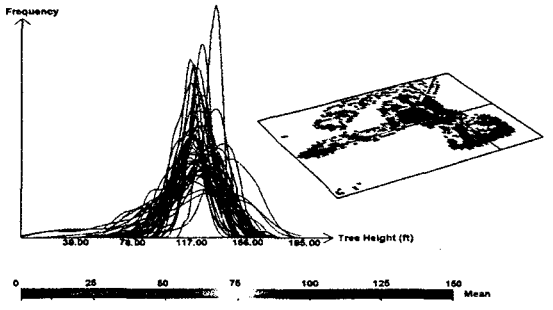

Fig. 9. This graph show the distributions of those pdfs that matched the same query as shown in Figure 7.

3) Visualizing the distributions and their spatial locations from a mode query: At this point, a scientist using the mode query tool has an idea of the locations and a graph of all the distributions that match the query. One source of information that is missing from the previous approach, shown in Figure 9, is that we do not know which grid cell the pdfs correspond to. In Figure 10, the same pdfs from Figure 9 are plotted right above their grid cells. We found this technique to be effective also for revealing the pdfs found by the query. By plotting the pdfs right above the corresponding grid cells, we can easily see the spatial locations of the matching pdfs. Note that the pdfs are drawn such that the density estimates are plotted along the axis perpendicular to the image plane. We construct a pdf curve for each grid cell found by the mode query. A pdf curve is created by horizontally displacing points along a vertical line by the magnitude of the density estimate. The height of the pdf curve is determined by the number of evaluation points of the density estimate given in Equation 1. In our example, 150 evaluation points are used. The color of each pdf curve presents the mean tree height of the distribution data at the corresponding grid cell. There are some parameters that the user can set for this approach. For example, the user can select another arbitrary direction to project the pdf curves and a different statistical descriptor to color the pdf curves. Though at first glance, the "drop down" pdfs shown in Figure 10 may appear to be some how cluttered as those shown in Figure 9. However, our tool allows the scientists to interactively rotating the graph shown in Figure 10 so that he/she may be able to view the "drop down" pdfs from different angles and thus allow for a better comparison of these pdfs and the shapes of these pdfs.

Visualizing the distributions and the spatial locations identified through a mode query allows a scientist to inspect them and find new spatial/distribution trends or clusters of "like" distributions. Conversely, the scientist can also find out where "different" distributions are located and if there are any random versus clumped patterns. Lastly, it also allows visual traversal of the distributions in a manageable way.

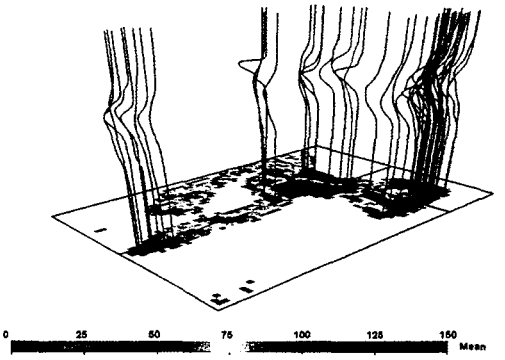

Fig. 10. Same query results as shown in Figure 9 except that the pdfs are plotted directly above the corresponding grid cells. These pdf curves provide another visual cue of the distributions found by a mode query.

D. Distribution Exploration

Distribution exploration is performed by distribution matching and visualizing similar distribution shapes. This allows scientist to identify all distributions that are similar in their entirety rather than in just a mode as described in Section V-C. Our tool allows the user to be more restrictive or more relaxed in the specificity of finding "like" distributions and allows all distributions to be ranked in terms of their similarity to the specified pattern. For example, matching could be restricted to certain data range, or only when the frequency is above a certain threshold. Likewise, matching could be relaxed by lowering acceptance threshold or using more liberal similarity metrics. Since density estimates vary in their quality, the ability to relax or restrict the definitions of similarity with the query tool allows user flexibility in identifying a range of like distributions and their spatial locations.

Through using the interactive data probe, the scientist was able to "visit" portions of the forest he was already familiar

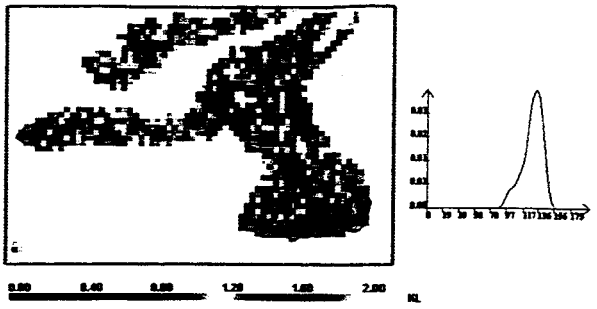

Fig. 11. Using the distribution matching tool, the scientist found all distributions that are similar to one (graph shown on the right) found to be recovering from a recent disturbance event.

Fig. 12. This figure shows a forest recovering from a catastrophic disturbance event which was found to have a unimodal distribution. Using the data probe tool, the scientist identified the characteristic distribution of this forest shown in Figure 11. Other forests recovering from more recent disturbance events were found to have this same characteristic distribution.

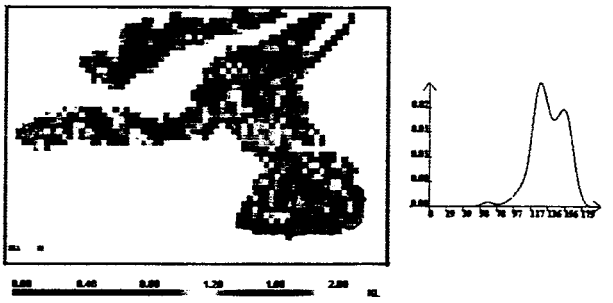

Fig. 13. The distribution on the right represents tree heights of a forest clump recovering from a recent disturbance event with residual (surviving) trees. The image on the left shows contour lines of where such distributions can be found on the island. The image is color mapped according to the similarity metric used in matching the characteristic distribution on the right with the distribution at each pixel location.

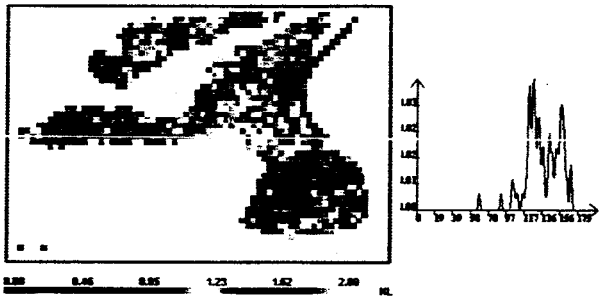

Fig. 14. Similar setup as Figure 13, but using a characteristic distribution that implies no evidence of any recent disturbances.

with (through field reconnaissance). He was then able to obtain three characteristic distributions of forest that are in various stages of recovery from major storm-driven disturbances. Using the distribution matching tool, all distributions that are similar in their entirety to those three distributions of interest were identified using contour lines as illustrated in Figures 11, 13, and 14. Identifying similar or matching distributions can be a powerful way to perform hypothesis testing, guide additional field work, and generate new data products of interest.

1) Visualizing matching distributions: Once the contours lines are generated from the results of the distribution matching tool, an additional visualization tool provides yet another way of studying subtle differences in the matching distributions. This tool constructs color mapped characteristic distribution surfaces to depict the variations of the pdfs along the contour lines. For each grid cell along a contour line, a vertical line is plotted right above the corresponding grid cell. Then, a surface mesh is formed by connecting vertical lines from the adjacent points along the contour line. The surface mesh is colored by the density estimates. Since there are usually several contour lines, our tool would generate several disjoint characteristic distribution surfaces. As with the pdf curves shown in Figure 10, the height of the surfaces is determined by the number of evaluation points of the density estimate. Figure 15 shows the characteristic distribution surfaces of the matching distributions. As with the visualization tools provided in the modality exploration, this visualization tool provides even further refinement of relative homogeneity, heterogeneity and associated possible spatial patterning of the characteristic distributions.

VI. DISCUSSION

Several visualization techniques ranging from synoptic, general views of the full data set to more specific, localized and detailed query and display were described in this paper. The techniques allow extensions well beyond summary descriptors such as the quadratic mean, robust mean and chi-square or other "non-parametric" summaries or clustering algorithms. The utility of each technique fundamentally depends on the selection of an appropriate density estimator. The estimator determines how the data are smoothed and how modes are defined. Each estimator is different, and may be well-suited for one type of data but not another. The kernel estimator used in this application with lidar data, for example, may have smoothed possibly interesting features in the data. The

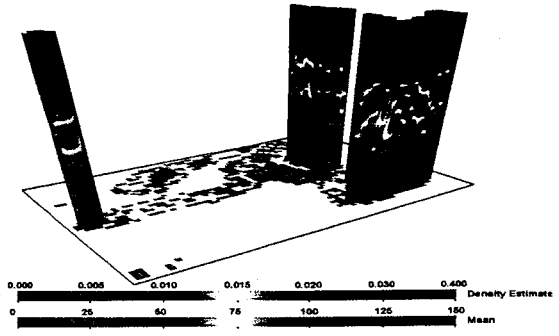

Fig. 15. This image shows the characteristic distribution surfaces for the contours lines shown in Figure 11. The surfaces are colored by the density estimates. The figure shows that the pdfs along the contours are mostly unimodal as indicated by the central magenta color band that runs across the middle section of the surfaces.

appropriateness of a given estimator depends partly on the number of raw data values per grid cell. In general, the larger the number of raw values, the more robust a given esimator will be. The precision of the data can also affect the size of the kernel used and the consequent smoothing of the distribution.

A key feature of these tools is their flexibility. Software that gives the scientist a choice of estimator and the ability to specify the parameters used in estimation will allow the accomodation of diverse data sets and exploratory data analysis. The kernel estimator we used in this study was selected because it is a robust, widely-used estimation technique but many other choices are possible.

Once an estimator is selected, the identification of modes is also not completely deterministic. Small bumps may be of little or no interest to the scientist, so what constitutes a mode in the display and query of modality can be user defined. Matching entire distributions is also a user-defined process. Success depends on increasing or decreasing the specificity of the distribution matching algorithm and having some meaningful criteria for doing so. Also, the distribution matching algorithm used is important. In this paper we used KL, but others are possible. Ultimately we envision a user-selection capability, so that various algorithms can be employed and their output assessed.

VII. CONCLUSION

Overall, our visualization tools provide new ways to query, visualize and compare distributions. The key contributions of this paper are as follows:

1) provide automated ways to analyze forest canopy distributions derived from lidar data,
2) make it easier for scientists to analyze distributions derived from lidar data, and
3) allow scientists to query distribution data for special features and then
   a) identify areas of the spatial field with similar distributions and
   b) discover potentially interesting distributions then find their locations.

Though the application described in this paper deals specifically with multi-return lidar data, our tools can be easily be used with distribution data sets from other applications. There are several open research problems in visualizing spatially varying distribution data sets, including the extension to distribution data that are sampled in a 3D domain and the extension to distribution data on more than one variable at a time.

VIII. ACKNOWLEDGEMENTS

This work is supported in part by the NASA Intelligent Systems Program Cooperative Agreement NCC2-1260 and NSF ACI-9908881. Additional support was provided to the second author through a National Research Council (NRC) postdoctoral research fellowship. We would like to thank Chris Hlavka and Michael Gerald-Yamasaki for their helpful comments and Anna Chen, Newton Der, Jose Renteria, Wei Shen, and Bing Zhang for help with programming and data preparation.

REFERENCES

[1] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[2] N.T. Eggleston, M. Watson, D.L. Evans, R.J. Moorhead, and J.W. McCombs II. Visualization of airborne multiple-return LIDAR imagery from a forested landscape. In *Proceedings of the Second International Conference of Geospatial Information in Agriculture and Forestry*, volume 1, pages 470–477, Jan 2000.

[3] D. Kao, J. Dungan, and A. Pang. Visualizing 2D probability distributions from EOS satellite image-derived data sets: A case study. In *Proceedings of Visualization '01*, pages 457–460, 2001.

[4] D. Kao, A. Luo, J. Dungan, and A. Pang. Visualizing spatially varying distribution data. In *Proceedings of the 6th International Conference on Information Visualization '02*, pages 219–225. IEEE Computer Society, 2002.

[5] M. G. Kramer, A.J. Hansen, M. Taper, and E. Kissinger. Abiotic controls on windthrow and forest dynamics in a coastal temperate rainforest, Kuiu Island, southeast Alaska. *Ecology*, 82:2749–2768, 2001.

[6] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 1951.

[7] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. In *Eurographics/IEEE TCVG Visualisation Symposium Proceedings*, May 2003. www.cse.ucsc.edu/research/avis/operator.html.

[8] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[9] M. Watson, N. Eggleston, D. Irby, R. Moorhead, and D. Evans. A virtual reality interface for analyzing remotely sensed forestry data. In Siggraph 2000 Conference Abstracts and Applications Catalog and CD-ROM, Sketches & Applications, 2000.