Mining Massive Earth Science Data Sets for Large Scale Structure

Amy Braverman and Eric Fetzer
Jet Propulsion Laboratory,
California Institute of Technology
Mail Stop 126-347
4800 Oak Grove Drive
Pasadena, CA 91109-8099
email: Amy.Braverman@jpl.nasa.gov

***Abstract*—The traditional way to look for large scale structure in very large observational or model generated data sets is to examine maps of means and standard deviations of parameters of interest on a coarse spatio-temporal grid. This approach is popular because it is easy to implement and understand, but unfortunately it throws away almost all of the distributional information in the data. Moreover, maps are computed for individual parameters of interest, and therefore do not retain information about relationships among two or more parameters. In this work, we use a modified data compression algorithm to produce multivariate distribution estimates for each grid cell. The algorithms optimally mediates between data reduction and fidelity loss using information-theoretic principles. Changes in these distribution estimates over time, space and resolution reflect large scale data structure. This is the basis for a data mining algorithm that characterizes those changes using a pseudo-metric for the distance between distributions. We demonstrate using data from the Atmospheric Infrared Sounder (AIRS) on board NASA's Aqua satellite.**

***Index Terms*—Massive data sets, data compression, probability distributions, AIRS.**

I. Introduction

THE motivation for this work is the need to facilitate exploratory data analysis of very large data sets produced by NASA Earth Observing System (EOS) satellites. EOS intends these data to be used by the greater research community in the study of Earth's climate system. However, for many researchers these data are too voluminous for the type of interactive, exploratory data analysis needed to formulate hypotheses and point the way to more detailed investigations. To ameliorate this problem, NASA instrument teams produce low volume, lower resolution, summary data sets typically comprised of means, standard deviations and other simple statistics for certain variables over an appropriate time period, and at coarse spatial resolution. For example, typical summary products might be aggregated daily or monthly over half, one, or five degree latitude-longitude spatial grid cells. Data processing constraints make this strategy attractive because means and standard deviations can be calculated in a "streaming" mode: one simply accumulates totals for each grid cell over the summary time period, and performs the appropriate division. Unfortunately, almost all distributional information is lost in this process, and this may be the information of greatest interest. For instance, subtle changes with time and space in the number of modes or occurrences of outliers may reflect important physical changes.

In [1] and [2] we proposed practical algorithms for constructing non-parametric, multivariate distribution estimates as replacements for simple means and standard deviations alone. The algorithms are based on entropy-constrained vector quantization (ECVQ) [3], an algorithm originally designed to estimate the rate-distortion function of a stochastic information source. We use a modified version of ECVQ designed specifically to output a set of representative, mutlivariate vectors along with their respective weights, for each grid cell to be summarized. These vectors and their weights define a discrete probability distribution that can be thought of as a coarsened version of the empirical distribution of the raw, unsummarized data. We place these distribution estimates in each cell of a monthly, $5^\circ \times 5^\circ$ coarse resolution grid, and mine the data for large scale structure by comparing distributions across time and space.

This paper discusses some initial investigations using Atmospheric Infrared Sounder (AIRS) data for January 2003.

II. Atmospheric Infrared Sounder Data and Science

Data used in this exercise are from the Atmospheric Infrared Sounder (AIRS) experiment, consisting of the AIRS instrument and the companion Advanced Microwave Sounding Unit (AMSU). AIRS is an instrument on-board NASA's EOS-Aqua satellite, launched on May 4, 2002. Aqua is in sun synchronous, polar orbit 705 kilometers above Earth, and crosses the equator during the ascending, or northward, part of its orbit at 1:30 pm local time. AIRS successively scans across a 1500 kilometer field of view taking data in 90 circular footprints as shown in Figure 1. As the spacecraft advances, the sensor resets and

This research is performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

obtains another scan line. 135 scans are completed in six minutes, and this 90 by 135 footprint spatial array constitutes a‘granule of data. AMSU observations are obtained at 1/9 the rate of those from AIRS, with each AMSU footprint colocated with nine AIRS footprints as shown in Figure 1. The instrument successively scans 240 granules per day, 120 on the day-time, ascending portions of orbits and 120 on night-time, descending portions. Granule ground footprints precess so granule 1 on a given day is not coincident with granule 1 on the next day. The descending granule map for July 20, 2002 is shown in Figure 2 as an example.

AIRS observes Earth and its atmosphere in 2378 infrared spectral channels. Roughly speaking, the channels sense the surface or different altitudes in the atmosphere. The instrument counts photons at the different wavenumbers, or inverse wavelengths. These counts are converted to brightness temperatures ranging from zero to about 340 degrees Kelvin. Certain atmospheric characteristics are related to photon emission, and these characteristics can be retrieved by solving complex sets of equations. The AIRS retrieval is performed on each set of nine AIRS footprints and a single, overlapping AMSU footprint, giving 1350 retrievals per granule.

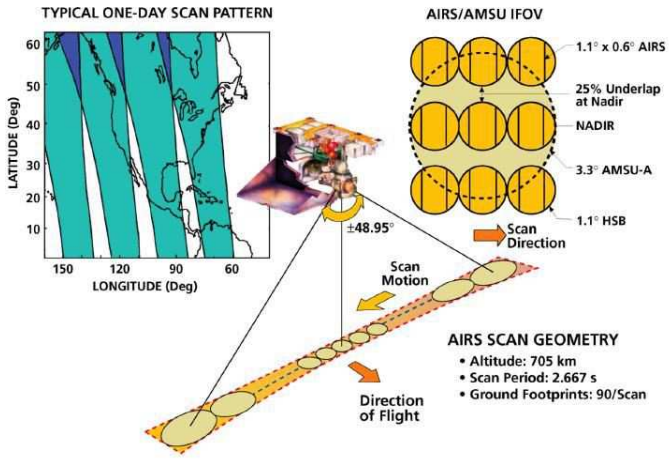

Fig. 1. AIRS scan geometry. Left: Ground-tracks for ascending portion of orbit. Center: AIRS instrument and ground footprints in one scan line. Right: AIRS, AMSU and HSB views of one footprint. AMSU and HSB are two other instruments on Aqua.

Among the retrieved geophysical variables are vertical profiles of temperature and water vapor at a subset of atmospheric levels scrutinized by AIRS, cloud fraction at two vertical levels, and several other diagnostic quantities. The variables of interest to us here are a subset of the temperature and water vapor levels in the lower part of the atmosphere, and the cloud fractions at those same levels. The retrievals are most accurate in this range, and much of the information about atmospheric state is contained therein. Variables are shown in Table I. Levels are numbered in order of increasing altitude, as shown in Table II Thus, each AIRS data point is a vector of the 35 variables shown in Table I. There are 1350 such data points in a granule,

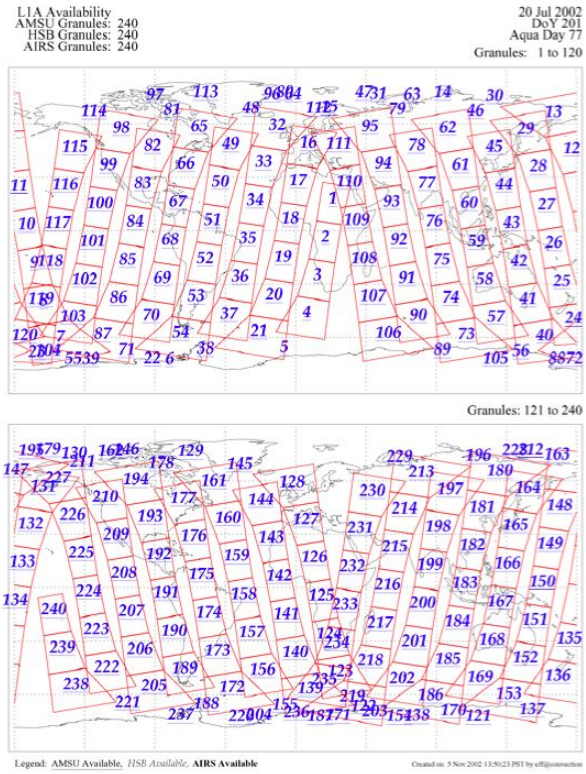

Fig. 2. 240 AIRS granule for a single day.

Variable(s)	*Levels*	*Units*
Temperature	1 - 11	$^\circ K$
Water vapor	1 - 11	gm/kg (dry air)
Cloud fraction	2 - 11	None
Land fraction	NA	None
Retrieval type	NA	0=Good/1=not good)
Scannode type	NA	0=Ascending/1=descending

TABLE I
AIRS VARIABLES INCLUDED IN OUR TEST DATA.

and 240 granules per day. In what follows, we these data acquired over one month, January 2003, to illustrate how distribution estimates are constructed and used.

III. Estimating Distributions

Given the 35-dimensional AIRS data for January 2003 we construct a monthly summary by partitioning the data points into subsets according to their membership in cells of a $5^\circ \times 5^\circ$ spatial grid. We then apply the modified ECVQ algorithm to each grid cell. In this section we briefly describe the modified ECVQ algorithm to aid understanding of what the distribution estimates represent. More detail can be found in [2].

ECVQ can be seen in at least three different ways. First, it is a penalized clustering algorithm. It partitions a collection of multidimensional data points into disjoint groups,

Level	*Pressure (mb)*
1	1000
2	925
3	850
4	700
5	600
6	500
7	400
8	300
9	250
10	200
11	150

TABLE II
AIRS PRESSURE LEVELS.

called clusters, and reports the centroid of each cluster as the cluster's representative. Second, it is a density estimation algorithm. The set of cluster representatives and their associated numbers of member data points define a discrete probability distribution, which is coarsened version of the original, empirical distribution of the data. Third, it is a quantization algorithm that finds the optimal encoder for a stream of stochastic signals that must be sent over a channel with limited capacity. These three interpretations are depicted schematically in Figure 3. Raw data points are C-dimensional observations, $\mathbf{x}$ of which there are many: N. Representative vectors are also C-dimensional, and denoted $\mathbf{y}$. The cluster analysis assigns each $\mathbf{s}$ to a group, indexed by k, via the encoding function, $\alpha(\mathbf{x})$. Cluster representatives are the mean vectors of all data points assigned to clusters. Cluster weights are M_k's, and within-cluster mean squared errors are δ_k's.

The same definitions apply to the density estimation view, except that the cluser weights are normalized to proportions. Here, the original distribution is represented by a (red) histogram in which every data point has weight $1/N$. Data points are grouped to form a new distribution (green). Here again, the α's provide the assignments. In the quantization view, a signal $\mathbf{X}$ from a stochastic information source, f, must be sent over a channel with finite capacity. Therefore, $\mathbf{X}$ can not be transmitted with perfect accuracy. A source encoder α, assigns every possible realization of $\mathbf{X}$ to one of K groups, and only the group index, $\alpha(\mathbf{x})$ is sent (in binary: $\gamma[\alpha(\mathbf{x})]$. At the receiver, the process is reversed to recover the group index, which is then replace by the group representative, $\mathbf{y}\beta[\alpha(\mathbf{x})]$. β is called the decoder, and in this application is always the group or cluster centroid determined by the encoder. An optimal code minimizes the estimation error, $E||\mathbf{x}-\mathbf{y}||^2$ ($E(\cdot)$ is the statistical expectation operator) subject to the constraint imposed by the channel capacity, H_{max}:

$$H(\mathbf{y}) = -\sum_{k=1}^{K} p_k \log p_k \leq H_{max},$$

where $H(\mathbf{y})$ is the entropy of the quantizer's output, $\mathbf{y}$, k is the number of groups, and $p_k = M_k / \sum_{k=1}^{K} M_k$. Thus, the quantization view reveals something the other two do not:

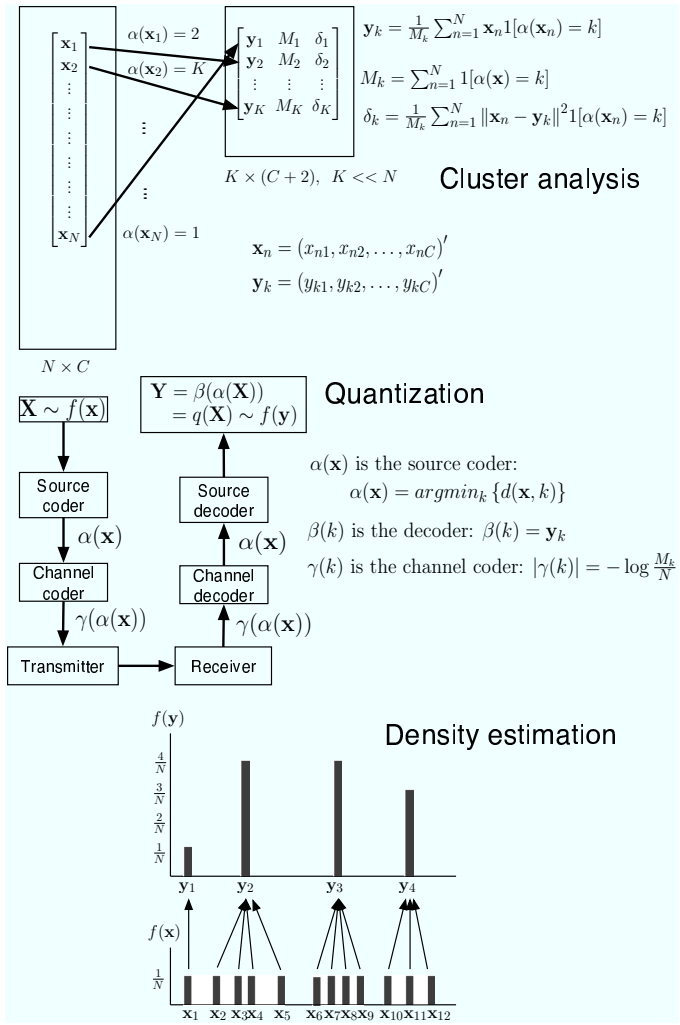

Fig. 3. Three interpretations of the ECVQ algorithm: as a clustering procedure (top), as a quantization algorithm (middle), and as a density estimation method (bottom). $\alpha(\mathbf{x})$ is the source encoding function returning the index of the cluster to which $\mathbf{x}$ is assigned. $\beta(k)$ is the source decoding function, in this returning the centroid of cluster k. $\gamma(k)$ is the channel coder which returns the binary representation of cluster index k. $\mathbf{X}$ and $\mathbf{Y}$ are random variables with possible realizations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k$ respectively. Note that $\mathbf{Y} = \beta[\alpha(\mathbf{X})]$ and is therefore a deterministic function of $\mathbf{X}$.

the problem is more complex than simply finding the optimal assignment N data points to K clusters or groups, otherwise unconstrained. The best assignment will balance mean squared error against complexity, H, and find the minimum mean squared error encoding function subject to a constraint on entropy. The practical implication is illustrated in Figure 4. The left panels of Figure 4 show two, two dimensional data sets as scatterplots, one with large variance and one with small variance. The center panels show the results of clustering the data with the K-means algorithm using five clusters. Note that the average squared distance from data points in the top data set to their nearest cluster representatives is larger than the average squared distance from data points in the bottom data set to their nearest cluster represenatives. In other words, accuracy of the cluster in the top data set representatives

of the raw data is greater than in bottom data set. A quantitative measure of this accuracy is the mean squared error, $E||\mathbf{x}-\mathbf{y}||^2$, also called the distortion in the signal processing literature. The right panel of Figure 4 the clustering obtained by the ECVQ algorithm. Only two clusters are formed for the bottom data set because only two are needed to achieve accuracy similar to that of the set of five clusters found for the top data set.

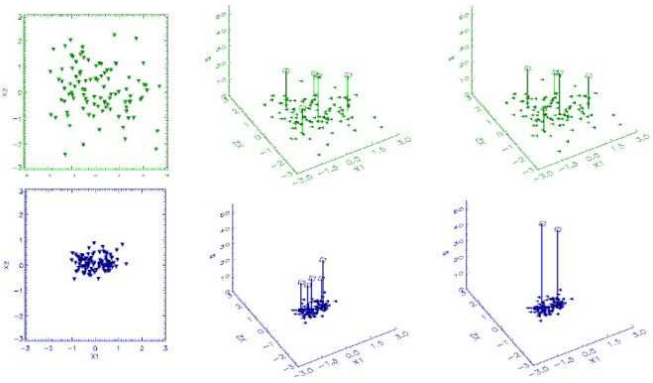

Fig. 4. Illustration of ECVQ clustering vs. K-means clustering. The left panel shows scatterplots of two, bivariate data sets, one more homogeneous than the other. The center panel projects the scatterplots onto the floor of a three dimensional space in which the vertical axis is number of cluster members. The locations of the pins show where K-means would locate five cluster representatives in each case. Pin heights show how many of the original data points on the floor are assigned to each cluster. Note that the average squared distance from points to the nearest pin is smaller in the bottom data set. The right panel shows the results for ECVQ. Only two pins are used to represent the homogeneous data set, and the average squared distance from points to nearest pin is more nearly equal for the two. The interpretation is that if one is satisfied a priori with the accuracy of the pins as a proxy for the heterogeneous data, then one needs fewer pins to be equally satisfied in the case of the homogeneous data.

The upshot is that while there is no a priori right number of clusters for representing a single data set in isolation, there are better clusterings and worse clusterings when applying the algorithm to more than one data set. That is case in our application where we apply clustering to the data belonging to each cell of a monthly, coarse resolution spatial grid. Our objective is to compare those data sets to one another through their clustered representations, and we want those representations to have similar accuracy. This is so that differences in the clusterings reflect real differences in the underlying data, and not differences quality. ECVQ achieves this, and the result is that entropy of the clustered representations is modulated across grid cell data sets. Those containing more complex data sets receive a more information-rich representation. Here, information is used in the rigorous sense of information theory. Often, this means more clusters are allocated to these data sets, but in truth the real distinction is that the clusterings of these data yield discrete probability distributions with higher entropy, $H(\mathbf{y})$.

We applied the ECVQ algorithm to data on the 35 variables described in Section II for the month of January 2003. Algorithmic details of the procedure can be found in [2]. The original data volumes follow from 240 granules per day for 31 days, and 4.9 MB per file. The ECVQ output is one, 25 MB file containing latitude and longitude indices for each $5^\circ \times 5^\circ$ grid cell, cluster index within grid cell, and for each cluster, a 35-dimensional representative vector, total cluster count, and cluster mean squared error. In addition, we broke down the total cluster count into contributions from each of six, five-day periods in the month. (Actually, the last period has six days.) The purpose is to retain some sub-monthly time information, the use of which will be apparent in Section.

IV. Data Mining

The result of the procedure described in Section III is one set of clusters for every $5^\circ \times 5^\circ$ grid cell. The number of clusters can vary from cell to cell depending on the number needed to adequately capture information-theoretic data complexity. There are no more than 50 clusters anywhere. We divide the cluster counts by the total numbers of raw data points in each grid cell to create discrete probability mass functions (pmf's). Our aim here is to quantify the difference between any two distributions in a way that allows us to systematically look for patterns of similarity related to geographic location. In the future, the same will be done for temporal relationships, i.e. look for patterns in time series of pmf's.

To define a difference measure for pmf's, let P_1 and P_2 be the pmf's belonging to two different grid cells. $P_1 = P(\mathbf{Q}_1 = \mathbf{q}_1$ and $P_2 = P(\mathbf{Q}_2 = \mathbf{q}_2$, where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are random vectors for which the possible realizations are the cluster representatives in the grid cells and the probabilities with which they are obtained are given by the normalized cluster counts. We define the distance between P_1 and P_2 as

$$\Delta(P_1, P_2) = \min_{p(\mathbf{q}_1, \mathbf{q}_2)} \sum_{ij} ||\mathbf{q}_{1_i} - \mathbf{q}_{2_j}||^2 p(\mathbf{q}_{1_i}, \mathbf{q}_{2_j}),$$

where $\mathbf{q}_{1_i}$ and $\mathbf{q}_{2_j}$ are the ith and jth possible realizations of $\mathbf{q}_1$ and $\mathbf{q}_2$ respectively. In other words $\Delta(P_1, P_2)$ is the expected squared distance between $\mathbf{Q}_1$ and $\mathbf{Q}_2$ under the joint distribution $p(\mathbf{q}_1, \mathbf{q}_2)$ that minimizes this distance subject to the constraints that $p(\mathbf{q}_1, \mathbf{q}_2)$ is consistent with the marginal distributions P_2 and P_2:

$$P(\mathbf{Q}_1 = \mathbf{q}_{1_i}) = \sum_j p(\mathbf{q}_{1_i}, \mathbf{q}_{2_j}),$$
$$P(\mathbf{Q}_2 = \mathbf{q}_{2_j}) = \sum_i p(\mathbf{q}_{1_i}, \mathbf{q}_{2_j}).$$

Since

$$\begin{aligned}\Delta(P_1, P_2) &= E||\mathbf{Q}_1 - \mathbf{Q}_2||^2 \\ &= E(\mathbf{Q}_1 - \mathbf{Q}_2)'(\mathbf{Q}_1 - \mathbf{Q}_2) \\ &= E\mathbf{Q}_1'\mathbf{Q}_1 - 2E\mathbf{Q}_1'\mathbf{Q}_2 + E\mathbf{Q}_2'\mathbf{Q}_2,\end{aligned}$$

and the first and last terms on the right are fixed because P_1 and P_2 are fixed, this amounts to maximizing the co-

variance of $\mathbf{Q}_1$ and $\mathbf{Q}_2$:

$$E\mathbf{Q}_1{}'\mathbf{Q}_2 = Cov(\mathbf{Q}_1, \mathbf{Q}_2) + [E\mathbf{Q}_1]'\,[E\mathbf{Q}_2]\,.$$

In other words, the joint pmf $p(\mathbf{q}_1, \mathbf{q}_2)$ that minimizes mean squared error infers joint probabilities that maximize the covariance, or equivalently, the correlation between $\mathbf{Q}_1$ and $\mathbf{Q}_2$. The interpretation is that distance between the pmf's of two grid cells is determined by assuming they are as correlated as possible while still satisfying the constraints imposed by their individual pmf's: we give the benefit of the doubt to the assumption that the cells are as related as possible.

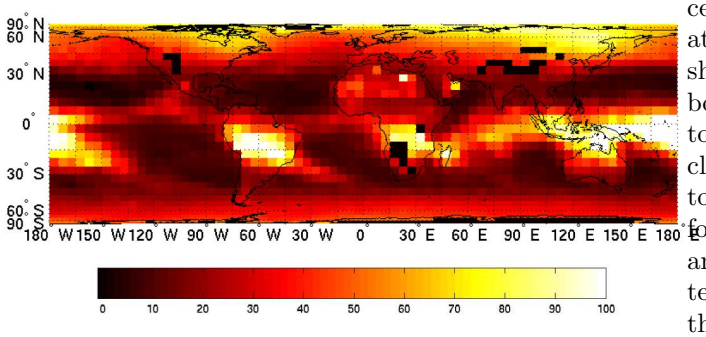

Fig. 5. Map of dissimilarity of grid cell pmf's relative to the grid cell containing Hawaii. Units are squared distance between 35-dimensional data points. Note that there is no data for a few grid cells in North America, southern Africa, and central Asia.

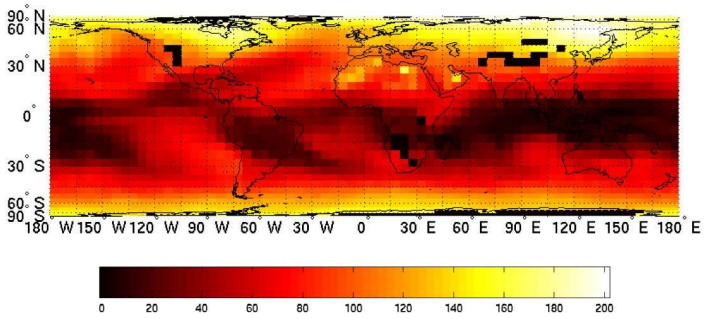

Fig. 6. Map of dissimilarity of grid cell pmf's relative to the grid cell containing Nauru. Units are squared distance between 35-dimensional data points. Note that there is no data for a few grid cells in North America, southern Africa, and central Asia.

To Illustrate how these ideas are relevant to science analysis, we computed $\Delta(P_1, P_2)$ for all unique (Δ is symmetric in its arguments), pairwise combinations of $P_1 = P_{lat,lon}$ and $P_2 = P_{lat',lon'}$, $lat, lat' = 1, 2, \ldots, 36$, $lon, lon' = 1, 2, \ldots, 72$, where lat and lat^{prime} index latitude and lon and lon' index longitude in the $5° \times 5°$ spatial grid. Figures 5 and 6 show maps of $\Delta(P_1, P_2)$ using two different grid cells, one containing Hawaii which is located at 1°N, 167°E, and and one containing the island of Nauru in the tropical western Pacific at 20°N, 155°W, as P_1. Thus, the color scales show how dissimilar all the other cells are to these two. The next section provides a discussion of these results, and we note here that visual inspection of these maps is not what we mean by data mining. We mean using Δ as a basis for quantifying the relationships among grid cells by producing, for example, a tree diagram showing how similar cell distributions are across levels of resolution. This work is in progress.

V. Science Discussion

The maps in Figures 5 and 6 show how dissimilar grid cells are to the grid cell containing Hawaii (Figure 5) and Nauru (Figure 6). Both Hawaii and Nauru are outfitted with radiosondes and ground instruments so their characteristics are well understood. Figures 7 and 8 are visualizations of the eight most populous clusters in each of those cells. The three line plots show vertical profiles of temperature, water vapor, and cloud fraction. The fourth panel shows the numbers of members in the eight largest clusters, both in total on the left, and by pentad in the six columns to the right of the total. The colors associated with the clusters cycle as black, blue, red and green, and correspond to the black, blue, red and green solid lines in the line plots for the four most populous clusters, and black, blue, red and green dashed lines for the next most populous clusters. Admittedly, the differences can be difficult to see in the plots.

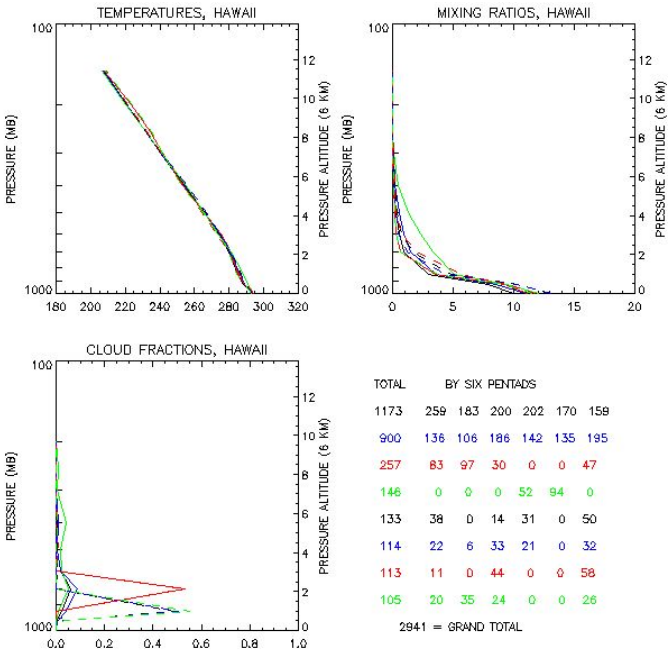

Fig. 7. Visualization of the eight most populous clusters representing the grid cell containing Hawaii. The three line plots show vertical profiles of temperature, water vapor, and cloud fraction. The fourth panel shows the numbers of members in the eight largest clusters, both in total on the left, and by pentad in the six columns to the right of the total. The colors associated with the clusters cycle as black, blue, red and green, and correspond to the black, blue, red and green solid lines in the line plots for the four most populous clusters, and black, blue, red and green dashed lines for the next most populous clusters.

We make the following observations. First, there is almost no variation in the temperature profiles at either loca-

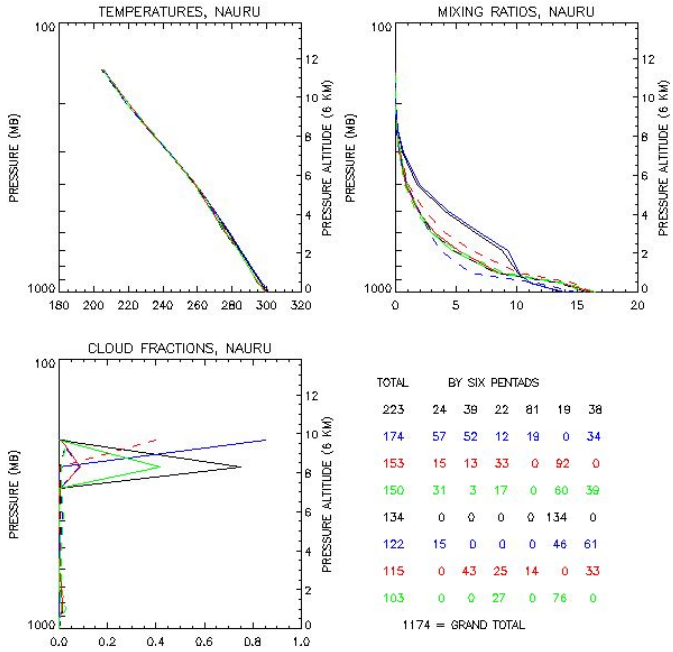

Fig. 8. Visualization of the eight most populous clusters representing the grid cell containing Nauru. See Figure 7 caption for explanation.

tion. Second, at Hawaii the three clusters with the highest total counts represent a shallow, humid layer with clouds at about 2 km. These are typical trade-wind cumulus clouds. Third, the fourth largest clusters is a moist layer at about 6 km. It represents storms moving through the area in late January. Note that all 146 original data points in this cluster come from the last two pentads in the month (January 20-25, and 26-31). Fourth, at Nauru, there are two dominant clusters with high water vapor amount up to several kilometers altitude. The primary difference between them is the height of the cloud layers. Fifth, the periodicity in cluster counts by pentad is consistent with the movement of a series of convective complexes through the area, which is typical. The less populous clusters represent clear, dry conditions between complexes.

The dark areas in Figures 5 and 6 are those we would expect to have conditions similar to Hawaii and Nauru, respectively. Hawaii and Nauru are moderately dissimilar to one another, and we now have a sense of what that means after looking at Figures 7 and 8. Looking at Figure 5, there were a few surprises. For example, climatology would lead us to believe that the Caribbean area would be quite similar to Hawaii, but this is less so than we expected. Other interesting features in the maps are the locations where there appear to be cells rather dramatically different than their neighbors.

VI. Conclusion

We have discussed and demonstrated the use of a probabilistic metric for determining how dissimilar two grid cells are based on estimates of the multivariate pmf's describing their data. Here we've provide just a toy example based on visual inspection of dissimilarity maps and plots of cluster representatives. The real goal is to use the metric to quantify the large scale ($5^\circ \times 5^\circ$ relationships among pmf's. This is work in progress. Even with such a quantification, the real benefit of this methodology can only be realized through careful interpretation by scientists. The grid cell pmf's are the data signatures of underlying physical processes, and the important work lies in connecting the two.

References

[1] A. Braverman, "Compressing Massive Geophysical Data Sets Using Vector Quantization," *Journal of Computational and Graphical Staistics*, vol. 11, no. 1, pp. 44–62, 2002.

[2] A. Braverman, E. Fetzer, A. Eldering, S. Nittel, and K. Leung, "Semi-streaming Quantization for Remote Sensing Data," *Journal of Computational and Graphical Staistics*, vol. 12, no. 4, pp. 759–780, 2003.

[3] P.A. Chou, T. Lookabaugh, and R.M. Gray, "Entropy-constrained Vector Quantization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 31–42, 1989.