MISR Global Data Products: A New Approach

Amy Braverman, *Member, IEEE*, and Larry Di Girolamo

***Abstract*—This paper describes a new type of global, gridded product being created by the Multi-angle Imaging SpectroRadiometer (MISR) team. The product is a compressed version, or summary, of MISR geophysical data products on a 1° monthly global grid. Data belonging to each grid cell are summarized by a multivariate histogram. The numbers, sizes, and shapes of the histogram bins vary among cells, and they adapt to the shape of the data in high-dimensional space. Also, bin representatives are means rather than midpoints. These modifications allow data to be summarized parsimoniously and with lower error than is possible using customary, simple, descriptive statistics. The method is demonstrated by compressing test MISR aerosol data, and performance is assessed by comparing computations using compressed data with those using the original.**

***Index Terms*—Clustering algorithms, data compression, entropy-constrained vector quantization, Level 3 products, massive datasets.**

I. Introduction

THE Multi-angle Imaging SpectroRadiometer (MISR) was launched into earth orbit aboard NASA's Terra satellite on December 18, 1999. Along with Terra's four other instruments, MISR has been collecting—and will continue to collect—massive quantities of data [1]. MISR alone is currently producing derived geophysical products at the rate of about 3.5 TB per year. One goal of the Terra mission is to provide the research community with long-term datasets for global climate studies, but even the most well-equipped users can expect global, exploratory analyses on this scale to be difficult. Recognizing this, many instrument teams, including the MISR team, have resolved to produce lower-volume, lower-resolution summaries of their geophysical data products. These are so-called Level 3 products.

Level 3 products are typically constructed by partitioning data collected in a month on a $1^\circ \times 1^\circ$ latitude–longitude spatial grid and then summarizing the data belonging to each grid cell with a set of simple, descriptive statistics such as means, standard deviations, and counts. While such summaries have the advantage of being well-understood and easy to compute, they discard most of the information in the data. For instance, the mean and standard deviation fully describe a data distribution only if the data are normally distributed. If not, these statistics characterize typical value and spread, but say nothing about skewness, number of modes, outliers, uniformity, clustering, or any other data features potentially important for science analysis. In addition, means and standard deviations do not describe relationships among geophysical parameters, and they may in fact obscure them. Covariances can be reported, but they say nothing about nonlinear relationships or interactions among three or more parameters.

Here we introduce a new type of Level 3 product, called the MISR Level 3 Joint Global Climate (L3JGC) product, designed to preserve more of the multivariate data structure present in MISR's geophysical products. Rather than providing a single mean and standard deviation for each parameter, L3JGC provides a set of representative vectors and associated weights. The vectors have as many components as there are parameters to be summarized. Vector weight is the number of data points the vector represents. In other words, if there are d parameters to be summarized for a given grid cell, the traditional Level 3 product provides d means, d standard deviations, and possibly some of the $d(d-1)/2$ covariances. The new product provides K d-dimensional (d-D) representatives and K counts, where the sum of the counts equals the number original data points in the cell for which measurements exist on all d parameters. K may vary among grid cells, depending on how many representatives are needed to adequately characterize their data. This will be done by balancing fidelity to the data against increased complexity incurred when greater numbers of representatives are used. We call the set of representatives and weights a summary, or compressed version, of the original $1^\circ \times 1^\circ$ cell data.

Several aspects of L3JGC distinguish it from traditional Level 3 products. First, it summarizes d parameters jointly, i.e., it treats N measurements of d parameters taken at the same time and place as N points in d-D space. Distances in this high-dimensional space are used to form groups from which representatives and weights are determined. Second, L3JGC is a family of nonparametric data distribution estimates, one for each grid cell. Each can be thought of as a high-dimensional histogram in which sizes, shapes, and numbers of bins adapt to the shape of the data in high-dimensional data space. Traditional histograms use a geometric partitioning to create rectangular bins. Here, inherent clustering in data space influences the shapes of the bins and allows the data to be represented with less error. Moreover, L3JGC is parsimonious: the number of representatives in a cell is only as large as necessary to adequately represent data belonging to it.

This paper describes the method used to create L3JGC and demonstrates how L3JGC can be used in data analysis. Section II describes the algorithm. Section III demonstrates its use on a test dataset constructed from some preliminary MISR aerosol retrievals. Section IV uses the resulting L3JGC-like dataset for a simple data analysis and assesses quality of the results by comparing them to those obtained when the

Manuscript received September 4, 2001; revised March 22, 2002. This work was conducted at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

A. Braverman is with the Jet Propulsion Laboratory, Pasadena, CA 91109-8099 USA (e-mail: Amy.Braverman@jpl.nasa.gov).

L. Di Girolamo is with the Department of Atmospheric Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801-3070 USA (e-mail: larry@atmos.uiuc.edu).

Publisher Item Identifier 10.1109/TGRS.2002.801159.

same analysis is performed on the original test data. Finally, Section V contains an assessment of the exercise and discusses some issues regarding product use. The statistical basis for the ideas presented here is discussed in [2].

II. Method

A. Summarizing Data

In this section, we introduce the method underlying creation of L3JGC. First, however, we introduce some notation, define a data summary, and discuss measures of quality for it.

Each month, each $1^\circ \times 1^\circ$ grid cell ("L3 cell") has an associated set of geophysical measurements. For the sake of this discussion assume those measurements are all at the same spatial resolution, say, 1.1 km^2, and focus on a single L3 cell; d parameters for the same 1.1-km^2 region can be concatenated to form a d-D column vector y, and the collection of N such vectors representing regions whose centers fall within the L3 cell are denoted by $\{y_n\}_{n=1}^N$; y may also be called an observation or a data point.

A summary of the data belonging to an L3 cell is a set of representative vectors and their associated weights. The y_n's in an L3 cell are partitioned into groups, also called clusters, and the mean vector of each cluster serves as its representative. The weights are the numbers of y_n's belonging to the clusters, also called counts. We let $\beta(k)$ denote the d-D mean vector of cluster k, and $N(k)$ denotes the number of y_n's it contains. Sometimes, we will also include a within-cluster error measure $\Delta(k)$, which is the average squared distance between data points in the kth cluster and their representative. If K is the number of clusters, the cell summary can be written compactly as $\{\beta(k), N(k)\}_{k=1}^K$ or $\{\beta(k), N(k), \Delta(k)\}_{k=1}^K$.

Two measures of summary quality are its distortion and entropy. Distortion is the average squared distance between data points and their representatives. This is just the weighted average of the within-cluster errors defined above, where the weights are given by the proportions of data points in each cluster. More formally, y_n's membership is recorded by an assignment function $\alpha_K(y_n)$ providing an integer identifying to which of the K clusters y_n belongs. With this notation, $\beta[\alpha_K(y_n)]$ is y_n's representative, and $N[\alpha_K(y_n)]$ is the corresponding count. Then distortion is

$$\Delta = \sum_{k=1}^{K} \frac{N(k)}{N} \Delta(k) = \frac{1}{N} \sum_{n=1}^{N} \|y_n - \beta[\alpha_K(y_n)]\|^2.$$

Entropy measures the average number of bits necessary to specify cluster membership of a data point, and it can be interpreted as a measure of summary descriptive complexity [3]. It is calculated from the probability distribution defined by the proportions $N(k)/N$:

$$h = -\sum_{k=1}^{K} \frac{N(k)}{N} \log \frac{N(k)}{N}.$$

Note that α_K fully determines the representatives, counts, errors, distortion, and entropy, since they all depend on cluster assignment. Low summary distortion generally comes at the cost of high descriptive complexity, and our goal is to find a set of assignments that achieves an optimal balance between the two.

This goal is similar to that of quantization in signal processing [4]. There, data are random signals to be sent over channels of limited capacity. Signals are assigned to one of K classes, and only class indicators are sent. At destination, indicators are replaced by representative values for classes as estimates of raw signals. The quantization problem is to design a system that produces good estimates within constraints on the number of bits necessary to distinguish classes (channel capacity) and in view of the statistical character of the signals. This is a constrained optimization problem: find α_K to minimize average error subject to a limit on the average number of bits per transmission.

Here, the data are the observations we wish to summarize, and for fixed K we seek α_K such that the distortion between the summary and the observations is small. The bit restriction is analogous to a requirement that summary entropy be small as well. However, unlike the signal processing situation, here we have no hard upper limit on entropy analogous to channel capacity. Instead, we formulate the optimization problem as follows: find the assignment of data points to K clusters to minimize a Lagrangian objective function

$$L_{\lambda,K} = \frac{1}{N} \sum_{n=1}^{N} \left[\|y_n - \beta[\alpha_K(y_n)]\|^2 + \lambda \left(-\log \frac{N(\alpha_K(y_n))}{N} \right) \right] \quad (1)$$

where K and λ are fixed constants.

The first term on the right-hand side of (1) is the average squared distance between data points and their representatives under α_K. The second term represents the average log-proportion of points assigned to the clusters. This term can be thought of as a penalty that is small when α_K assigns data points to clusters to produce a low-entropy configuration. The parameter λ translates the penalty into units of squared distance compatible with the first term. If $\lambda = 0$, $L_{\lambda,K}$ is minimized when α_K assigns each y_n to the cluster with the nearest squared Euclidean distance representative, and all K clusters receive at least one data point. If $\lambda > 0$, the α that minimizes $L_{\lambda,K}$ may assign some data points to clusters with representatives that are not nearest to them because more massive clusters exist further away. In fact, some clusters may receive no data points at all, and the number of nonempty clusters may be fewer than K.

To find the set of assignments that minimizes $L_{\lambda,K}$, we use an iterative, randomized algorithm described in Section II-B and given fully in the Appendix. Algorithm parameters λ and K must be set in advance, and a procedure for determining their values is discussed in Section II-C.

B. Algorithm

The algorithm to find optimal assignment functions is based on the Entropy-Constrained Vector Quantization algorithm (ECVQ) [5]. ECVQ is an iterative descent algorithm for minimizing $L_{\lambda,K}$ by choice of α. Fig. 1 is a diagram of ECVQ.

Assuming K and λ are fixed, ECVQ begins by assigning each data point randomly to one of K clusters and computing cluster

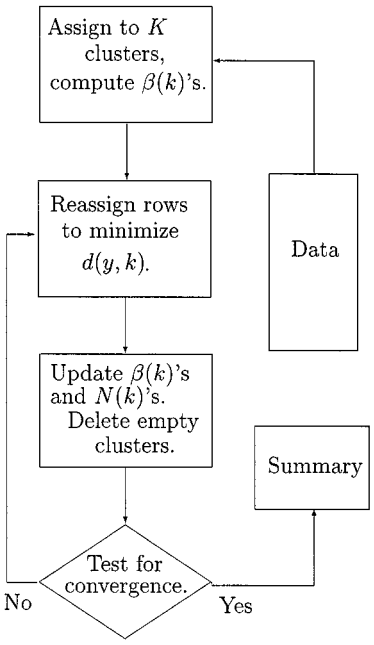

Fig. 1. ECVQ algorithm finds assignments of data points to clusters when both distortion and entropy of the distribution resulting from the assignments are taken into account.

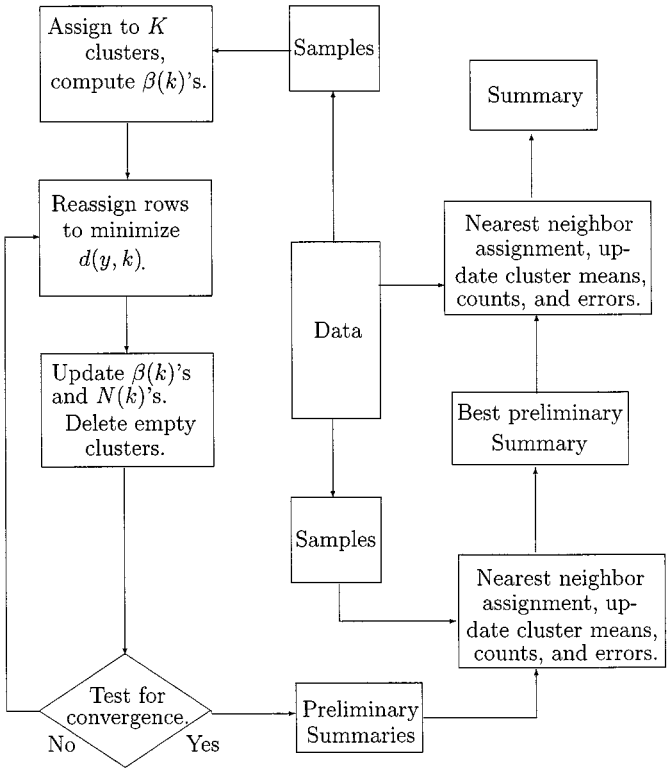

Fig. 2. Modified ECVQ algorithm. The ECVQ algorithm is repeated multiple times with different random samples on each trial to create multiple preliminary summaries. The distortions that would be incurred if the entire dataset were summarized using these summaries are also estimated from the samples. The preliminary summary with smallest estimated distortion is used to summarize the full dataset.

means. Next, each data point is reassigned to the cluster with the nearest mean vector, and then cluster means and counts are updated. The following steps are then iterated until convergence: 1) each data point is reassigned to the cluster for which the penalized distance

$$d(y,k) = \|y - \beta(k)\|^2 + \lambda[-\log N(k)/N]$$

is minimized, where $\beta(k)$ and $N(k)$ are the cluster means and counts, respectively; 2) $\beta(k)$ and $N(k)$ are updated, and empty clusters deleted. The final values of $\beta(k)$ and $N(k)$ constitute the basic summary, and sometimes the within-cluster distortion will be included as well. The ECVQ algorithm is guaranteed to converge in a finite number of iterations [5]. Though not guaranteed to converge to a global minimum of $L_{\lambda,K}$, the ECVQ solution will always be an improvement over the starting assignments, and the resulting summaries provide a sensible, lower-volume version of the original data.

ECVQ in this form is not practical for MISR data for several reasons. First, ECVQ is iterative and too computationally intensive for large volumes of data. Second, solutions are subject to sampling variation because the initial assignment of data points to clusters is random. Third, ECVQ solutions are not nearest-neighbor: the final assignments do not minimize squared Euclidean distances between data points and their cluster representatives. These problems are mitigated by modifying ECVQ, as shown in Fig. 2.

First, we run ECVQ separately on a number of independent random samples taken with replacement from the data in the cell being summarized. This produces a number of different sets of representatives and counts, here called preliminary summaries. Each preliminary summary is derived from one sample, called its design sample, and we call the others its test samples. For instance, if we use 50 samples, the first one is the design sample for the first preliminary summary, and the other 49 are the first preliminary summary's test samples. The second sample is the design sample for the second preliminary summary, and samples $1, 3, \ldots, 50$ are its test samples, and so forth. Second, each preliminary set of representatives is then used to summarize its test samples by i) assigning each set of test sample points to their nearest squared Euclidean distance representatives, ii) recomputing the cluster means based on these assignments, and iii) obtaining the distortion between test sample and the preliminary summary. In the example, this yields 49 distortions for each of the 50 preliminary summaries. Distortions for the same preliminary summary are averaged to produce an estimate of the distortion that would be incurred had that summary been used to summarize the entire cell dataset rather than just the test samples. Third, the preliminary summary with the lowest estimated distortion is designated the best preliminary summary. The best preliminary summary is used to summarize the entire cell dataset by assigning every data point in the cell to the cluster with the closest squared Euclidean distance representative. New cluster means, counts, and distortions are computed, and these are reported as the final summary for the cell. We designate the final summary by $\{\tilde{\beta}(k), \tilde{N}(k)\}_{k=1}^{\tilde{K}}$, or $\{\tilde{\beta}(k), \tilde{N}(k), \tilde{\Delta}(k)\}_{k=1}^{\tilde{K}}$, where the tilde notation indicates all data in the cell are represented. With this scheme all data are summarized, but only a random sample is used to design the summary.

We call the final summary's distortion the *a posteriori* distortion. We also report the average of the 50 preliminary summary distortions and call it the *a priori* distortion, since it approximates the average distortion incurred before specifying a particular preliminary summary for use on the whole dataset. *A posteriori* distortion can be thought of as a goodness-of-fit measure for a specific summary. It is the relevant error to propagate through transformations of summarized data in order to gauge how closely such quantities approximate true values of those transformations computed using the original dataset. *A priori* distortion can be thought of as a process performance measure incorporating both goodness-of-fit and algorithm stability over different design samples. It is the relevant error measure for making decisions about future applications of the modified ECVQ algorithm to the same data. For instance, *a priori* distortion is the relevant error for selecting an appropriate value of λ in Section II-C.

Problems of computational intensity are mitigated by using samples for the computationally intense, iterative part of the procedure. Moreover, using multiple random samples on multiple trials and averaging estimated summary errors yield *a priori* distortion estimates that account for sampling variation. Also, this strategy constitutes a randomized algorithm that improves chances of finding a true global minimum of $L_{\lambda,K}$ if it exists. Finally, since data are assigned to their nearest cluster means prior to updating for the last time, final assignments are approximately nearest neighbor and, therefore, error minimizing. The modified procedure is called modified ECVQ.

C. Setting K and λ

The algorithm described in Section II-B requires the parameters K and λ be set in advance. The parameter K establishes the general level of fidelity. The larger K is, the lower will be the distortion of the theoretically best summary, the summary produced by the assignment function that is the global minimizer of $L_{\lambda,K}$ when $\lambda = 0$. Even though our algorithm may use $\lambda > 0$ and is not guaranteed to find the global minimum, larger values of K tend to produce lower distortions. All other things being equal, one would therefore like to set K as large as possible. On the other hand, the total number of clusters for all L3 cells must not exceed some value determined by the file size allocated for their storage. We set K to the largest integer not exceeding this value divided by the number of cells being summarized.

The parameter λ translates the penalty in $L_{\lambda,K}$ to its equivalent in units of squared distance, but it is not obvious what λ should be. If one were compressing a single cell dataset in isolation, the most accurate summary, for given K, is obtained by setting $\lambda = 0$. Then, $L_{\lambda,K}$ is minimized by assigning each data point to the nearest of the K representatives regardless of consequences for complexity. Now consider compressing two datasets, say two neighboring L3 cells, using no more than K clusters each. Suppose cell A contains data distributed over some hyperspherical region in d-space, and suppose also that cell B contains the same number of points tightly clustered in a much smaller region. The situation is illustrated for $d = 2$ by the scatterplots in the top panels of Fig. 3. The modified ECVQ solutions obtained with $\lambda = 0$ are shown in the middle two panels of Fig. 3 for $K = 5$. Both summaries have K representatives, but the quality is much better for cell B, since the average squared distance between data points and their representatives is smaller.

K clusters are more than the number necessary to achieve cell A's quality level in cell B, or, put another way, cell B's summary is more descriptively complex than necessary. In this sense, cell B's data are simpler than cell A's because a summary with lower entropy can achieve the same distortion level. In the interest of parsimony, B's summary should be simpler, and we say cell B is undercompressed at $\lambda = 0$.

Alternatively, if λ is very large ($\lambda = \infty$) the complexity penalty in $L_{\lambda,K}$ outweighs the error term. In each cell, observations are assigned to a single cluster for which the representative is the cell mean vector. Entropies for both cells' summaries are $\log 1 = 0$, and summary errors are just the sums of the variances of the data vector elements. This is an overcompressed condition in which summaries do not reflect differences in data variation. Those differences are subsumed into summary errors rather than being manifest by differences in descriptive complexity.

Between these extremes is some value of λ that neither undernor overcompresses the two cells' data relative to one another. Such a solution is shown in Fig. 3(e) and (f), and it corresponds to $\lambda = 0.2$. Summary error levels are relatively similar compared to solutions at $\lambda = 0$ or $\lambda = \infty$. Condition $\lambda = 0.2$ produces summaries of similar quality, so differences between them reflect data distribution differences rather than discrepancies in goodness-of-fit of the summaries to their data. If this exercise were repeated using $K = 10$ instead of $K = 5$, the same type of result would be obtained except that error levels would generally be lower.

Parameter K controls overall fidelity of a collection of summaries to their data. Parameter λ is used to tune the summaries to a common distortion level in that regime so that summaries are of comparable quality. In practice, K is set by practical considerations described earlier; then, various values of λ are tested by running the modified ECVQ algorithm on a subset of L3 cells. For instance, in the example in Section III, we use all L3 cells for which latitude and longitude are even multiples of 5° and test $\lambda = 0, 0.1, 0.2, \ldots, 1.0$. We start with this range because $\lambda = 0$ is its minimum possible value, and $\lambda = 1.0$ makes one unit of distortion equal to one bit of descriptive complexity in the objective function. We look at the variance across L3 cells of *a priori* distortion and select λ to minimize this quantity. *A priori* distortion is relevant because this is a pretest. Final cell summaries will not be created here; a future application of the algorithm will be required, and sampling variation must therefore be taken into account.

If this procedure suggests λ should be an endpoint of the initial test range, the range is refined, so we can be reasonably sure we are finding a minimum. If the distortion-minimizing λ is zero, we retest using $0, 0.01, 0.02, \ldots, 0.09$. If it is one, we retest using $1.1, 1.2, \ldots, 2.0$, and in principle continue the process until the λ that minimizes the variance of *a priori* distortion is in the interior of a test range. In practice, we stop after a reasonable number ranges, say five, have been tested.

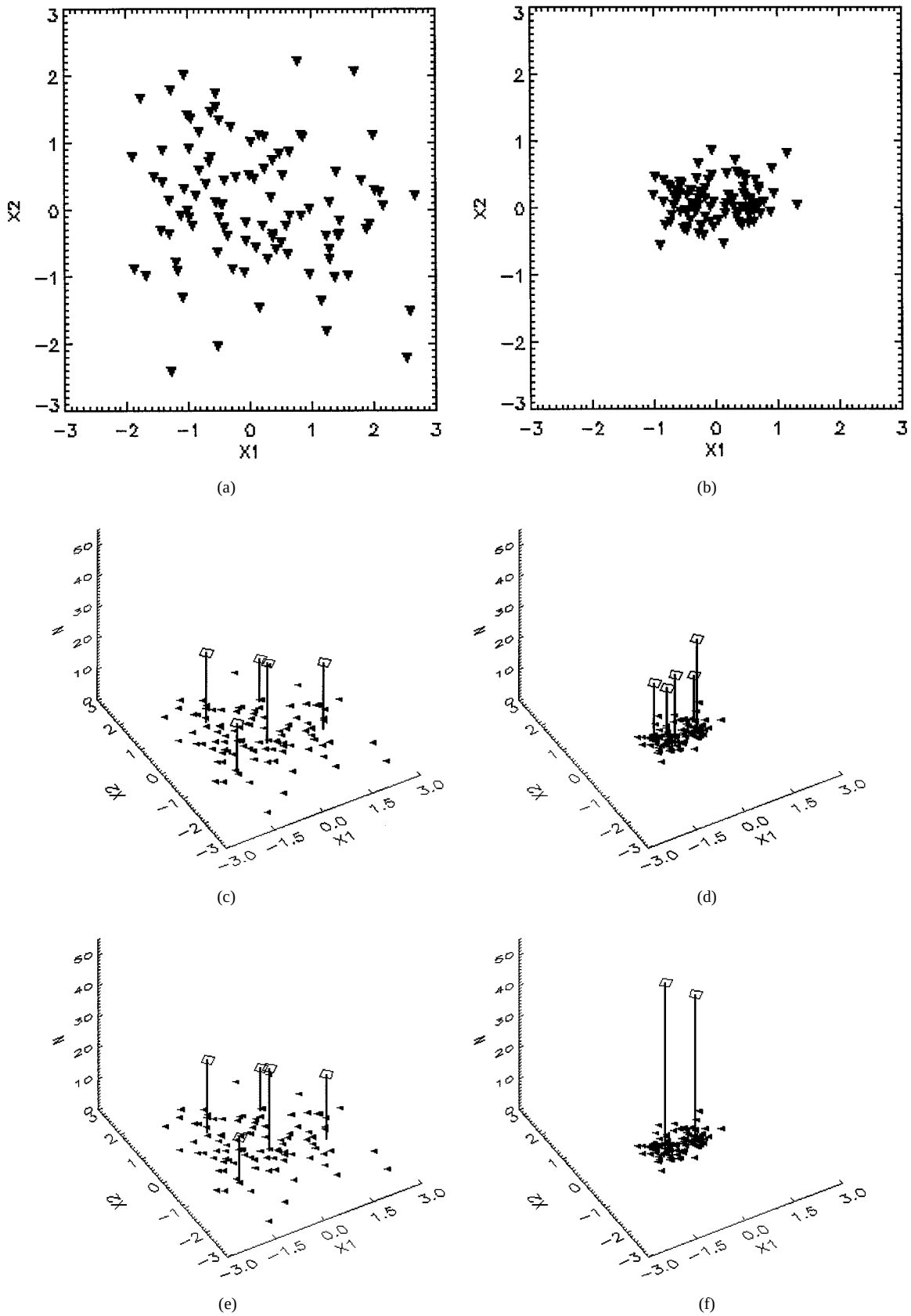

Fig. 3. (Top) Scatterplots of two datasets: (a) one heterogeneous (cell A in the text) and (b) one homogeneous (cell B in the text). (Middle) Summaries of the two datasets using five clusters in both cases. These summaries are produced by the ECVQ algorithm with $K = 5$ and $\lambda = 0$. Scatterplots are projected onto plot floors. Locations of spikes show locations of cluster representatives. Spike heights indicate cluster counts. Region of high data density in the homogeneous dataset is represented by five clusters in close proximity. (Bottom) Summaries of the two datasets produced by the ECVQ algorithm with $K = 5$ and $\lambda = 0.2$. The region of high data density in the homogeneous dataset is represented by fewer, more massive clusters than was the case when $\lambda = 0$. Note that the average (squared) distance between data points and their nearest representatives is more equal when $\lambda = 0.2$ than when $\lambda = 0$.

III. Example: Compressing MISR Aerosol Data

To demonstrate the use of the modified ECVQ algorithm, we create a compressed version of some MISR test data. We use test data extracted from preliminary MISR aerosol retrievals over southern Africa (latitude -40° to 0°, longitude 0° to 55° over land only) during a six-week period beginning in August 2000. As of that time, MISR aerosol retrievals over land provide an optical depth estimate τ and five goodness-of-fit measures, denoted χ_1^2, χ_2^2, χ_3^2, χ_4^2, and χ_5^2, describing how well MISR's observed radiances match sets of radiances predicted by five different aerosol models. For our example, optical depth and the five χ^2's are appended to form a six-dimensional (6-D) data vector $y = (\tau, \chi_1^2, \chi_2^2, \chi_3^2, \chi_4^2, \chi_5^2)'$. There is one such data vector for each 17.6-km^2 region.

Since aerosol retrievals are only valid in the absence of clouds, MISR's stereo-derived cloud mask (SDCM) is used to screen out cloudy subregions. Each 17.6-km^2 region is made up of a 16×16 array of 1.1-km^2 subregions, and the SDCM designates each of these as cloudy or clear with high or low confidence. Test data were created from the regional retrievals by copying each data vector N_r times, where N_r is the number of high-confidence, clear subregions in region r. Each subregional data vector bears the latitude and longitude of the corresponding subregion, and these 1.1-km^2 data points are the observations to be summarized in this exercise. Of course, this creates a certain amount of artificial clustering. In creating L3JGC, it will be necessary to reconcile spatial resolutions in this way to combine parameters derived at different resolutions. The larger the disparity in resolution, the more pronounced will be this artifact. Under such circumstances, counts reported in the summaries should be interpreted as weights rather than actual numbers of observations. That being said, we ignore the issue for the remainder of this example.

The modified ECVQ algorithm is applied to all 743 cells independently with $K = 40$, 50 random samples of size 500 drawn with replacement, and $\lambda = 0.1$. The paramters K and λ were set according to the criteria described in Section II-C. The condition $K = 40$ assures there can be no more than a total of $743 \times 40 = 29\,720 \approx 30\,000$ representatives output. To select a value of λ, we tested values $0, 0.1, 0.2, \ldots, 1.0$ using 29 L3 cells: those with latitudes and longitudes evenly divisible by 5°. Condition $\lambda = 0.1$ minimized the variance of the *a priori* distortions across those cells. The number of samples and sample size was chosen to balance the benefits of large samples against the cost in terms of computational speed. Seven hundred forty-three summaries are created

$$\left\{ \left\{ \tilde{\beta}_{u,v}(k), \tilde{N}_{u,v}(k), \tilde{\Delta}_{u,v}(k) \right\}_{k=1}^{\tilde{K}_{u,v}} \right\}_{(u,v)}$$

$u = 0, \ldots, -40$, $v = 0, \ldots, 55$, where u and v index cell latitude and longitude. A typical cell having about 15 000 data points was compressed in about 12 min on a 400-MHz RISC 12000 processor.

Figs. 4–6 show some diagnostics. Fig. 4 displays the numbers of test data vectors present by grid cell $N_{u,v}$. There are 743 nonempty cells. The color bar in Fig. 4 is truncated at 20 000 to ensure distinguishability at the low end of the scale. At the high end, the cell with the largest number of data points has 31 137 observations. The total size of the test data for all cells is 6 304 861 6-D data points.

Fig. 5 shows the number of clusters $\tilde{K}_{u,v}$ allocated to each grid cell. The largest number of clusters allocated to any cell is 31. Note that cells with relatively high $\tilde{K}_{u,v}$ are not necessarily the same ones with high $N_{u,v}$. More clusters are allocated to cells with more complex data, not necessarily to ones with more data points. The total number of clusters for all cells combined is $\sum_{u,v} \tilde{K}_{u,v} = 9322$.

Fig. 6 shows the square root of summary *a posteriori* distortion in each cell relative to cell average data point norm

$$\sqrt{\tilde{\Delta}}_{u,v}^{\text{rel}} = \frac{\sqrt{N_{u,v}^{-1} \sum_{k=1}^{\tilde{K}_{u,v}} \tilde{\Delta}_{u,v}(k) \tilde{N}_{u,v}(k)}}{N_{u,v}^{-1} \sum_{n=1}^{N_{u,v}} \|y_{n,u,v}\|}.$$

This is a measure of how well the reported summary represents an average data point. Of the 743 summaries, 738 have errors which are less than 5% as large as the average data point size. The color bar in Fig. 6 is truncated at 5%, since this is sufficient for almost all the values shown, and provides distinguishability at the low end of the scale. Of the five cells with errors greater than 5%, the largest relative error is about 28%. There is no apparent geographic pattern in the errors. The five high-distortion cells are those with a small numbers of data points where the algorithm may be less stable.

Taken together, Figs. 4–6 show how much data volume reduction is achieved and at what cost. One measure of compression is the proportional reduction in the number of records, here 99.85%. There are two additional fields in the summaries: cluster count and within-cluster distortion. It is also possible to offer an information-theoretic measure of data reduction based on entropy reduction, but this seems less relevant than simple reduction in file length for data-handling concerns addressed here. The low relative *a posteriori* errors suggest we can in general expect calculations based on these summaries to well approximate the same calculations using original data. However, this depends as much on the nature of the calculation as on the level of fidelity. Section IV contains an example involving a nonlinear transformation and demonstrates why using the summaries is better than using grid cell means.

IV. Sample Data Analysis

In this section, we investigate relationships between optical depth and χ^2 measures using compressed MISR aerosol test data, and we compare results with the same set of calculations performed on raw test data. We emphasize that the standard by which the modified ECVQ algorithm should be judged is how closely results of the two sets of calculations match, not by their substantive content. At the time of this analysis, the retrieval algorithm which created these test data was preliminary. Therefore, no conclusions should be drawn regarding quality of MISR data based on this analysis.

We examine correlation between optical depth, τ, and the variance, W, of the five χ^2's about their mean. W is a measure

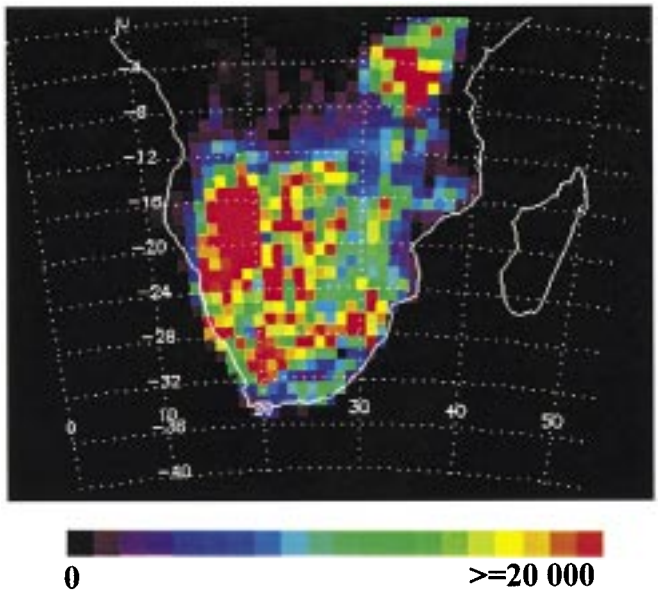

Fig. 4. Number of original data points belonging to L3 grid cells. $N_{u,v}$ is the value for the cell at latitude u, longitude v. The total number of data points represented is $\sum_{u,v} N_{u,v} = 6\,304\,861$.

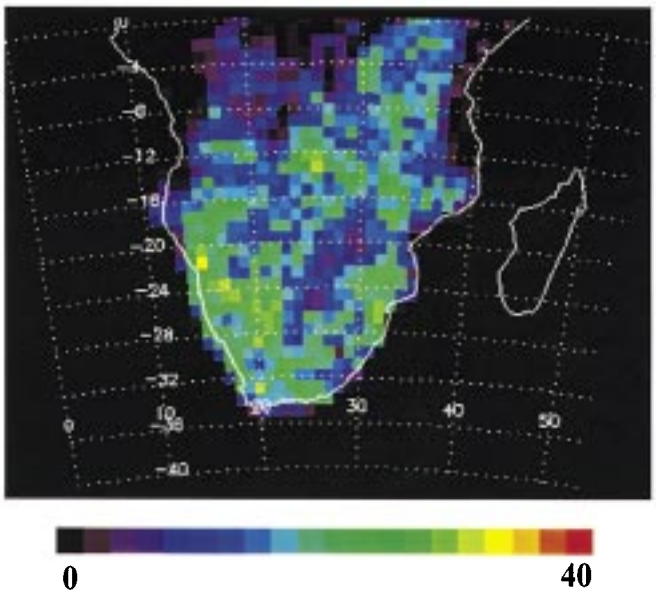

Fig. 5. Number of clusters allocated to L3 grid cells. $\bar{K}_{u,v}$ is the value for the cell at latitude u, longitude v. The total number of clusters output is $\sum_{u,v} \bar{K}_{u,v} = 9322$.

of homogeneity of the χ^2's. Low values indicate the five models explain observed radiances nearly equally well, and high values indicate some degree of differentiation. We use the correlation $\rho(\tau, W)$ for this demonstration, not for its scientific content, but because it is a nonlinear function of all six components of y_n. While linear functions of raw data are exactly reproduced by the same linear functions of compressed data, nonlinear functions are not preserved [2]. Quality of nonlinear transformation estimates depends on how closely summaries match raw data and on the transformation. This example illustrates that when summaries match their data well, good estimates of typical nonlinear functions such as correlation can be obtained.

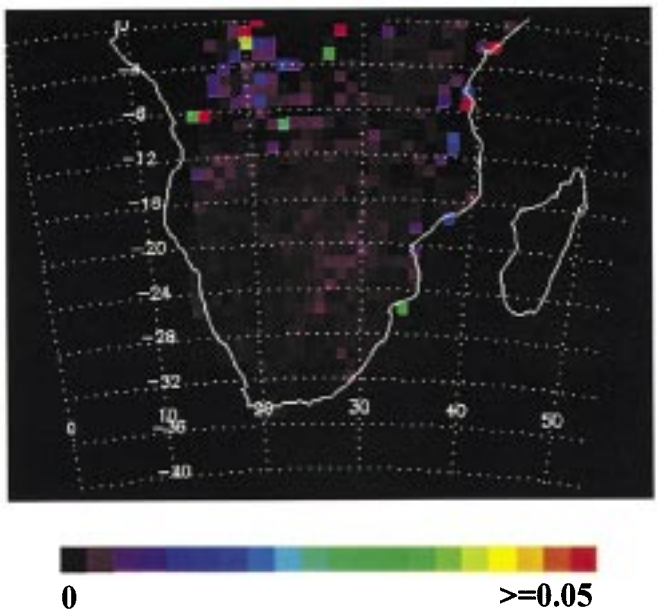

Fig. 6. Square root of *a posteriori* error, relative to average data point norm, by grid cell: $\sqrt{\bar{\Delta}}^{\mathrm{rel}}_{u,v} = \sqrt{\bar{\Delta}}_{u,v} / \sum_{n=1}^{N_{u,v}} \|y_{n,u,v}\|$, where $y_{n,u,v}$ is the nth data point in the grid cell at latitude u, longitude v.

W can be calculated for each data point

$$W_{n,u,v} = \frac{1}{5}\sum_{i=1}^{5}\left(\chi^2_{n,u,v,i} - \bar{\chi}^2_{n,u,v}\right)^2$$
$$\bar{\chi}^2_{n,u,v} = \frac{1}{5}\sum_{i=1}^{5}\chi^2_{n,u,v,i} \tag{2}$$

where i indexes aerosol model. We would like ρ, the true correlation between τ and W, for all 743 $1^\circ \times 1^\circ$ grid cells:

$$\rho_{u,v} = \frac{N_{u,v}^{-1}\sum_{n=1}^{N_{u,v}}(\tau_{n,u,v} - \bar{\tau}_{u,v})(W_{n,u,v} - \bar{W}_{u,v})}{\sqrt{\sigma^2_{\tau,u,v}}\sqrt{\sigma^2_{W,u,v}}}$$

where $\bar{\tau}_{u,v}$ and $\bar{W}_{u,v}$ are the mean values of τ and W in the L3 cell with latitude u and longitude v, and $\sigma^2_{\tau,u,v}$ and $\sigma^2_{W,u,v}$ are the variances:

$$\bar{\tau}_{u,v} = N_{u,v}^{-1}\sum_{n=1}^{N_{u,v}}\tau_{n,u,v}$$
$$\bar{W}_{u,v} = N_{u,v}^{-1}\sum_{n=1}^{N_{u,v}}W_{n,u,v}$$
$$\sigma^2_{\tau,u,v} = N_{u,v}^{-1}\sum_{n=1}^{N_{u,v}}(\tau_{n,u,v} - \bar{\tau}_{u,v})^2$$
$$\sigma^2_{W,u,v} = N_{u,v}^{-1}\sum_{n=1}^{N_{u,v}}(W_{n,u,v} - \bar{W}_{u,v})^2.$$

Here, it is possible to calculate $\rho_{u,v}$ because the test data are small. Fig. 7 shows the true values of $\rho_{u,v}$ by grid cell.

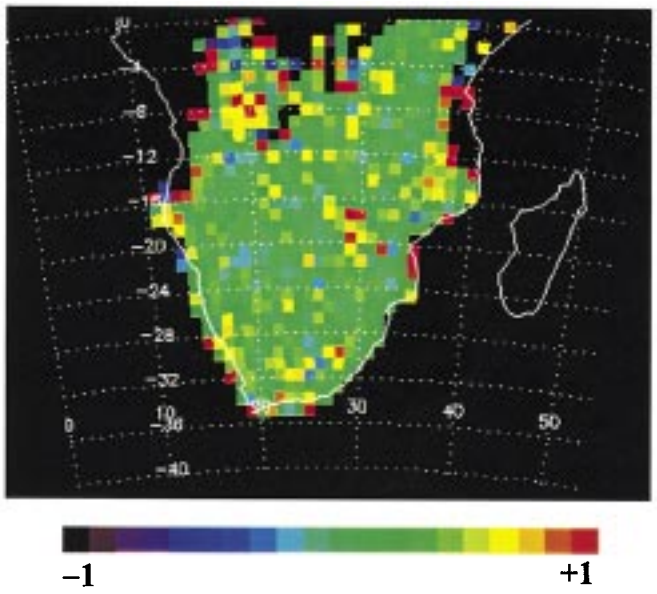

Fig. 7. True correlation, $\rho_{u,v}$, between optical depth, τ, and goodness-of-fit χ^2 variance, W, in the MISR aerosol test data.

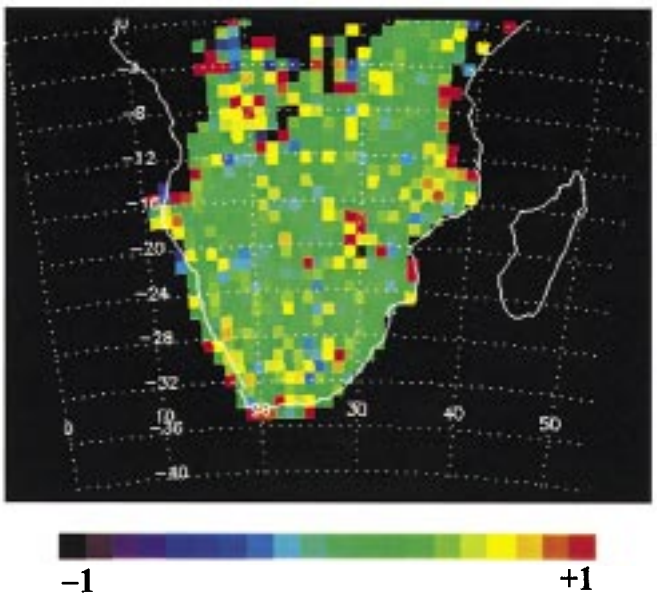

Fig. 8. Estimated correlation, $\hat{\rho}_{u,v}$, between optical depth, τ, and χ^2 variance, W, obtained from summarized MISR aerosol test data.

With large volumes of data like those we obtain from MISR over a month, it may not be possible to calculate quantities like $\rho_{u,v}$ directly. Instead, $\rho_{u,v}$ can be estimated by $\hat{\rho}_{u.v}$ using compressed data:

$$\hat{\rho}_{u,v} = \frac{N_{u,v}^{-1} \sum_{k=1}^{\tilde{K}_{u,v}} \tilde{N}_{u,v}(k)(\hat{\tau}_{k,u,v} - \hat{\bar{\tau}}_{u,v})(\hat{W}_{k,u,v} - \hat{\bar{W}}_{u,v})}{\sqrt{\hat{\sigma}^2_{\tau,u,v}}\sqrt{\hat{\sigma}^2_{W,u,v}}}$$

where

$$\begin{aligned}
\hat{\bar{\tau}}_{u,v} &= N_{u,v}^{-1} \sum_{k=1}^{\tilde{K}_{u,v}} \tilde{N}_{u,v}(k)\hat{\tau}_{k,u,v} \\
\hat{\bar{W}}_{u,v} &= N_{u,v}^{-1} \sum_{k=1}^{\tilde{K}_{u,v}} \tilde{N}_{u,v}(k)\hat{W}_{k,u,v} \\
\hat{\sigma}^2_{\tau,u,v} &= N_{u,v}^{-1} \sum_{k=1}^{\tilde{K}_{u,v}} \tilde{N}_{u,v}(k)(\hat{\tau}_{k,u,v} - \hat{\bar{\tau}}_{u,v})^2 \\
\hat{\sigma}^2_{W,u,v} &= N_{u,v}^{-1} \sum_{k=1}^{\tilde{K}_{u,v}} \tilde{N}_{u,v}(k)(\hat{W}_{k,u,v} - \hat{\bar{W}}_{u,v})^2 \\
N_{u,v} &= \sum_{k=1}^{\tilde{K}_{u,v}} \tilde{N}_{u,v}(k). \qquad (3)
\end{aligned}$$

$\hat{\tau}_{k,u,v}$ is the first component of $\tilde{\beta}_{u,v}(k)$, and $\hat{W}_{k,u,v}$ is computed from the remaining five components of $\tilde{\beta}_{u,v}(k)$ using a formula analogous to (2). Fig. 8 shows estimated correlations by grid cell. It matches Fig. 7 closely, and neither shows any particularly striking geographic features. Conclusions drawn from the two figures would be similar.

This analysis would be impossible using a data product which represents L3 cell data by its mean or any other single value. Correlation is a property of probability or data distributions, and cannot be estimated unless there is a distribution of data points from which to calculate it. Moreover, even quantities which can be estimated from the mean alone may be very poorly estimated. For example, consider estimating the mean value of W. True mean W in the cell at latitude u, longitude v is given by $\bar{W}_{u,v}$ and is shown for all cells in the upper panel of Fig. 9. The estimate of true mean W computed from compressed data is given by (3) and shown in the middle panel of Fig. 9. Given only the mean vector for any cell, $\bar{y}$, the only possible estimate of true mean W is $(5^{-1})\sum_{i=1}^{5}(\bar{y}_i - \bar{\bar{y}})^2$, where $\bar{\bar{y}} = (5^{-1})\sum_{i=1}^{5}\bar{y}_i$. This estimate is shown in the bottom panel of Fig. 9.

The estimate is obviously poor. The reason is that W is a nonlinear function of the data. The mean of a nonlinear function does not equal the nonlinear function applied to the mean. For these types of calculations, its important to have distributional information over and above means and standard deviations.

V. Discussion

This paper presents a new type of global data product being produced to summarize MISR geophysical data. The algorithm used to produce it is modified from a signal processing application and is demonstrated using test MISR aerosol data. Examples show that typical nonlinear functions of the original data can be estimated well from the new product even though the new product is much smaller than the data it summarizes. Results are contrasted with estimates derived from cell means.

L3JGC is a global, descriptive summary of MISR data. It is created without physical or statistical modeling assumptions and is intended to facilitate global, exploratory data analysis. Presently, it is not intended to be used for inference. In this respect, it functions in the same capacity as traditional Level 3 products. For instance, simple means and standard

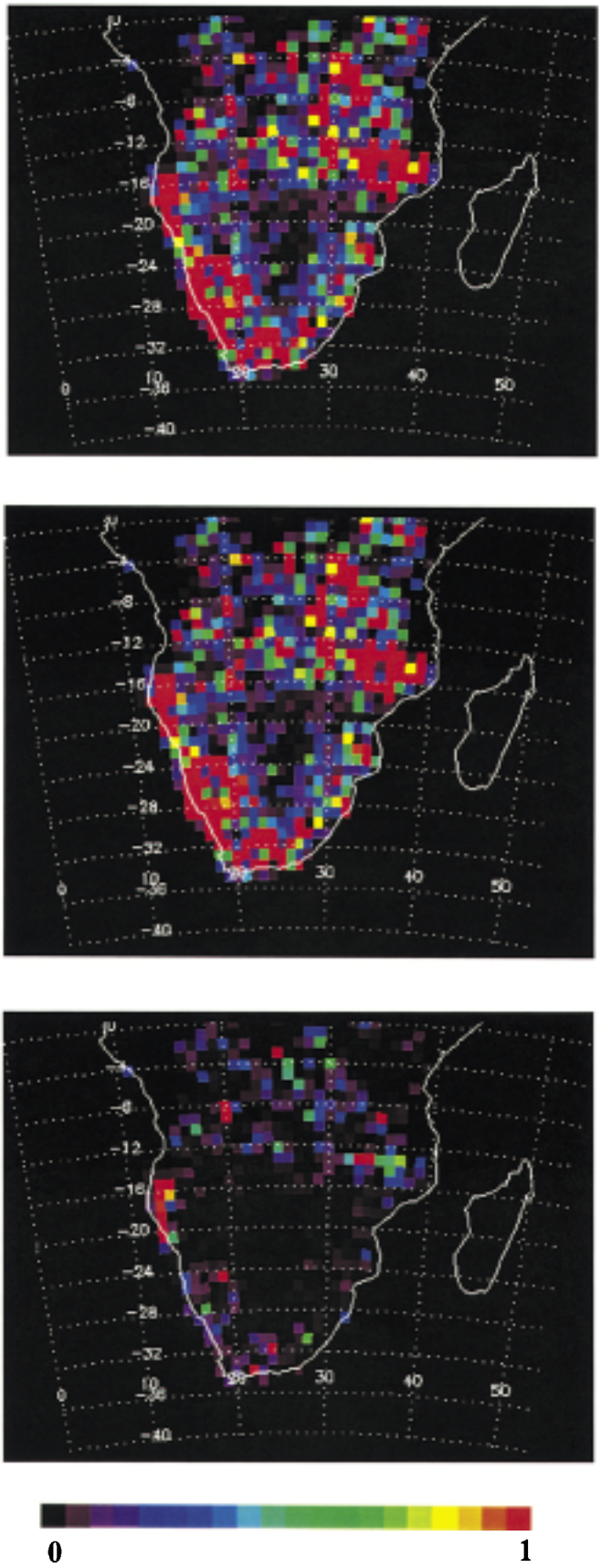

Fig. 9. (Top) True average χ^2 variance, $\bar{W}$, by L3 cell in the MISR aerosol test data. (Middle) Estimated average $\bar{W}$ computed from summarized data. (Bottom) Variance of χ^2's about their mean computed from cell mean vector.

deviations are appropriate for inference about underlying physical processes, assuming data generated by them are statistically independent. However, many researchers recognize that spatial and temporal dependences exist, and they will want to calculate their own statistics directly from high-resolution products to incorporate their own assumptions. Even so, global maps of means and standard deviations are useful in looking for expected and unexpected phenomena, patterns, and relationships in the data wholly apart from inference. The summaries described here contain more information about the multivariate distribution of the data than do simple means and standard deviations. They should, therefore, be useful in identifying phenomena, patterns, and relationships across space and time arising from higher order, multivariate features such as skewness, covariation, and other high-order interactions, multimodality, clustering, and outliers. Armed with this information, researchers can then make targeted requests for specific, manageable portions of MISR's high-resolution data products from which inferences can be made.

APPENDIX
ALGORITHM

A. Preprocessing

Given a dataset for one L3 cell, denote the N d-D observations therein by y_n, $n = 1, 2, \ldots, N$. Assume we know the global mean vector, μ, and covariance matrix, Σ, computed from all data in all cells. Select S samples of size M from the y_n. Denote the jth sample $\{x_{j1}, x_{j2}, \ldots, x_{jM}\}, j = 1, 2, \ldots, S$. Standardize all the sample points using the global mean and variances: let $z_{jm} = \Gamma^{-1/2}(x_{jm} - \mu)$, where $\Gamma = \mathrm{diag}(\Sigma)$.

B. Create Preliminary Summaries

Fix λ and K. Then, for each standardized design sample $j = 1, 2, \ldots, S$ do the following.

1) Set the iteration counter $t = 0$. Set the convergence criterion ϵ.
2) Set $K^{(0)} = K$, $\mathcal{I}^{(0)} = \{1, 2, \ldots, K^{(0)}\}$, $\alpha^{(0)}(z_{jm}) = m$, for $m = 1, 2, \ldots, (K^{(0)} - 1)$, $\alpha^{(0)}(z_{jm}) = K^{(0)}$, for $m = K^{(0)}, \ldots, M$. Here, the random initial assignment is accomplished by assigning the first $K^{(0)} - 1$ data points to the first $K^{(0)} - 1$ clusters and all remaining data points to the last cluster.
3) For $k = 1, 2, \ldots, K^{(0)}$ compute

$$N^{(0)}(k) = \sum_{m=1}^{M} 1[\alpha^{(0)}(z_{jm}) = k]$$

$$\beta^{(0)}(k) = \frac{1}{N^{(0)}(k)} \sum_{m=1}^{M} z_{jm} 1[\alpha^{(0)}(z_{jm}) = k]$$

$$\gamma^{(0)}(k) = -\log\left[\frac{N^{(0)}(k)}{M}\right]$$

where $1[\cdot] = 1$ if its argument is true and 0 otherwise.
4) For $m = 1, 2, \ldots, M$ set

$$\alpha^{(t+1)}(z_{jm}) = \mathrm{argmin}_k ||z_{jm} - \beta^{(t)}(k)||^2 + \lambda\gamma^{(t)}(k).$$

5) For $k = 1, 2, \ldots, K^{(t)}$ update:

$$N^{(t+1)}(k) = \sum_{m=1}^{M} 1[\alpha^{(t+1)}(z_{jm}) = k]$$

$$\beta^{(t+1)}(k) = \frac{1}{N^{(t+1)}(k)} \sum_{m=1}^{M} z_{jm} 1[\alpha^{(t+1)}(z_{jm}) = k]$$

$$\gamma^{(t+1)}(k) = -\log\left[\frac{N^{(t+1)}(k)}{M}\right].$$

6) For $k = 1, 2, \ldots, K^{(t)}$, remove k from $\mathcal{I}^{(t)}$ if $N^{(t+1)}(k) = 0$. Renumber so that elements of $\mathcal{I}^{(t)}$ are consecutive, and set $\mathcal{I}^{(t+1)} = \mathcal{I}^{(t)}$.

7) Set $t = t + 1$. Compute

$$L^{(t)} = \sum_{m=1}^{M} \left[\left\|z_{jm} - \beta^{(t)}[\alpha^{(t)}(z_{jm})]\right\|^2 + \lambda\gamma^{(t)}(\alpha^{(t)}(z_{jm}))\right].$$

If $[L^{(t-1)} - L^{(t)}]/L^{(t-1)} > \epsilon$, go to Step 4).

8) Set $\alpha_j^* = \alpha^{(t)}$, $K_j^* = K^{(t)}$, $I_j^* = I^{(t)}$. For $k = 1, 2, \ldots, K_j^*$ set $\beta_j^*(k) = \beta^{(t)}(k)$.

9) Report the preliminary cluster representatives on the standard scale $\{\beta_j^*(k)\}_{k=1}^{K_j^*}$.

C. Estimate Errors of Preliminary Summaries

For each preliminary summary $j = 1, 2, \ldots, S$, cluster the $S-1$ test samples corresponding to $i \neq j$. Let $m = 1, 2, \ldots, M$ index data points in the test samples. Assign each z_{im} to the cluster in $\{\beta_j^*(k)\}_{k=1}^{K_j^*}$ with the nearest Euclidean distance cluster representative, and calculate totals:

1) For $i = 1, 2, \ldots, S$ and $i \neq j$, do
   a) For $k = 1, 2, \ldots, K_j^*$ set $\tau(k) = \mathbf{0}$, $\tau_2(k) = \boldsymbol{O}$, where $\mathbf{0}$ is the d-D zero vector, and $\boldsymbol{O}$ is the $d \times d$ matrix of zeros. Set $\tau_N(k) = 0$.
   b) For $m = 1, 2, \ldots, M$ in the ith test sample, set

$$\begin{aligned}\alpha(z_{im}) &= \text{argmin}_{k \in \mathcal{I}_j^*} \|z_{im} - \beta_j^*(k)\|^2 \\ \tau(\alpha(z_{im})) &= \tau(\alpha(z_{im})) + z_{im} \\ \tau_2(\alpha(z_{im})) &= \tau_2(\alpha(z_{im})) + z_{im} z_{im}{}' \\ \tau_N(\alpha(z_{im})) &= \tau_N(\alpha(z_{im})) + 1.\end{aligned}$$

   c) For $k = 1, 2, \ldots, K_j^*$, remove cluster k if $\tau_N(k) = 0$. Renumber the remaining clusters consecutively, $1, 2, \ldots, K_j^{**}$.
   d) For $k = 1, 2, \ldots, K_j^{**}$ set

$$\begin{aligned}\beta_{ji}^{**}(k) &= \frac{\tau(k)}{\tau_N(k)} \\ N_{ji}^{**}(k) &= \tau_N(k) \\ \Delta_{ji}^{**}(k) &= \text{tr}\left[\frac{\tau_2(k)}{N_{ji}^{**}(k)} - \beta_{ji}^{**}(k)\beta_{ji}^{**}(k)'\right].\end{aligned}$$

   e) Set Δ_{ji}^{**}:

$$\Delta_{ji}^{**} = \sum_{k=1}^{K_{ji}^{**}} \frac{N_{ji}^{**}(k)}{N} \Delta_{ji}^{**}(k).$$

Δ_{ji}^{**} is the error incurred when the preliminary summary derived from sample j is used to summarize sample i.

2) Compute the estimate of the error incurred when the preliminary summary derived from sample j is used to summarize the full cell dataset

$$\Delta_j^{**} = \frac{1}{S-1} \sum_{i \neq j} \Delta_{ji}^{**}.$$

D. Identify the Best Preliminary Summary

1) Set $s^{\text{opt}} = \text{argmin}_j \{\Delta_j^{**}\}_{j=1}^{S}$.

E. Summarize the Full Dataset

1) For $k = 1, 2, \ldots, K_{s^{\text{opt}}}^{**}$ set $\tau(k) = \mathbf{0}$, $\tau_2(k) = \boldsymbol{O}$, and $\tau_N(k) = 0$.

2) For $n = 1, 2, \ldots, N$ set $z_n = \Gamma^{-1/2}(y_n - \mu)$.

$$\begin{aligned}\alpha(y_n) &= \text{argmin}_{k \in \mathcal{I}_{s^{\text{opt}}}^{**}} \|z_n - \beta_{s^{\text{opt}}}^{**}(k)\|^2 \\ \tau(\alpha(y_n)) &= \tau(\alpha(y_n)) + y_n \\ \tau_2(\alpha(y_n)) &= \tau_2(\alpha(y_n)) + y_n y_n{}' \\ \tau_N(\alpha(y_n)) &= \tau_N(\alpha(y_n)) + 1.\end{aligned}$$

Note that assignment is based on distances between the z_n and $\beta_{s^{\text{opt}}}^{**}(k)$ which are on the standard scale, but totals are computed using data on the original scale.

3) For $k = 1, 2, \ldots, K_{s^{\text{opt}}}^{**}$ remove cluster k if $\tau_N(k) = 0$. Renumber the remaining clusters consecutively, $1, 2, \ldots, \tilde{K}$.

4) For $k = 1, 2, \ldots, \tilde{K}$ set

$$\begin{aligned}\tilde{\beta}(k) &= \frac{\tau(k)}{\tau_N(k)} \\ \tilde{N}(k) &= \tau_N(k) \\ \tilde{\Delta}(k) &= \text{tr}\left[\frac{\tau_2(k)}{\tilde{N}(k)} - \tilde{\beta}(k)\tilde{\beta}(k)'\right].\end{aligned}$$

Report the summary of the full dataset

$$\left\{\tilde{\beta}(k), \tilde{N}(k), \tilde{\Delta}(k)\right\}_{k=1}^{\tilde{K}}.$$

The *a posteriori* error of this summary is $\sum_{k=1}^{\tilde{K}} (\tilde{N}(k)/N)\tilde{\Delta}(k)$. The *a priori* error for this L3 cell is $\delta = S^{-1} \sum_{j=1}^{S} \Delta_j^{**}$.

ACKNOWLEDGMENT

The authors would like to thank the referees and editors for their careful attention and helpful suggestions.

REFERENCES

[1] R. Kahn and A. Braverman, "What shall we do with the data we are expecting from upcoming earth observation satellites?," *J. Comput. Graph. Stat.*, vol. 8, no. 3, pp. 575–588, Sept. 1999.

[2] A. Braverman, "Compressing massive geophysical datasets using vector quantization," *J. Comput. Graph. Stat.*, vol. 11, no. 1, pp. 44–62, Mar. 2002.

[3] R. A. Ash, *Information Theory*. New York: Dover, 1965.

[4] A. Gerhso and R. Gray, *Vector Quantization and Signal Compression*. New York: Kluwer Academic, 1991.

[5] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 31–42, Jan. 1989.

Amy Braverman (M'01) received the Ph.D. degree in statistics from the University of California, Los Angeles, (UCLA), in 1999, the M.A. degree in mathematics from UCLA in 1992, and the B.A. degree in economics from Swarthmore College, Swarthmore, PA, in 1982.

She is currently a Statistician at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena. Her research interests are data reduction and analysis of massive datasets, data mining, and high-dimensional data visualization. She is also the Level 3 Scientist for MISR and the Atmospheric Infrared Sounder (AIRS).

Dr. Braverman is a member of the American Geophysical Union and the American Statistical Association, as well as being on the Board of Governors of the Interface Foundation of North America.

Larry Di Girolamo received the B.S. degree (with honors) in astrophysics from Queen's University, Kingston, Canada, in 1989 and the M.S. and Ph.D. degrees in atmospheric and oceanic sciences from McGill University, Montreal, QC, Canada, in 1992 and 1996, respectively.

He is currently an Assistant Professor in the Department of Atmospheric Sciences at the University of Illinois at Urbana–Champaign, Urbana. He teaches introductory courses on atmospheric science and advanced courses on satellite remote sensing. His research interests include the remote sensing of cloud properties, 3-D radiative transfer through heterogeneous cloud fields, and sampling strategies for remotely sensed data. He is Member of the MISR science team, as well as the AMS STAC committee on satellite meteorology and oceanography.