# Data Mining for Climate Model Improvement

Amy Braverman
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

Robert Pincus and Crispian Batstone
Climate Diagnostics Center
NOAA Earth System Research Laboratory
325 S. Broadway, R/PSD1
Boulder CO 80305

***Abstract*—Given very large volumes of remote sensing data and climate model output, one would like to be able to compare them in order to understand where, when and why model data do not agree with observations. Due to the large volumes, and to incongruities between instrument observation techniques and models, the traditional approach is to reduce both data sources by averaging important parameters up to coarse, common resolution. This destroys information about high-resolution dependencies among parameters, which are often important sources of model-data discrepancies. Instead, we replace parameter means with full, multivariate distribution estimates of multiple quantities of interest. We then perform formal statistical hypothesis tests to determine whether distributions produced from model output agree with those for the same coarse grid cell obtained from observations. If differences exist, we can isolate them with another suite of hypothesis tests that identify the distributional characteristics causing the problems. In this talk, we report on work to assess and diagnose the Geophysical Fluid Dynamics Laboratory's AM2 atmospheric model.**

***Index Terms*—Massive data sets, data compression, probability distributions, climate model diagnosis.**

## I. Introduction

THIS paper reports the current status of our work to diagnose and evaluate climate models by comparing their output to trusted observations. Our approach is to summarize the two data sets by estimating their multivariate probability distributions for selected variables, here vertical profiles of equivalent potential temperature and saturation equivalent potential temperature at 35 vertical levels in the atmosphere. We compare and examine the distributions rather than individual observations. This has two advantages. First, it requires us only to look at these reduced volume, reduced complexity distributional summaries rather than the more unwieldy raw data or information destroying mean values. Second, it allows us to bring to bear the tools of probability theory to analyze the results. This in turn makes it possible for us to conduct formal hypothesis tests that quantify our findings.

The rest of this paper describes the data we use, the method of estimating multivariate distributions, and develops some probability based methods for analyzing them.

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

## II. Model Output and Observations

For this study we used observational data from the Atmospheric Radiation Measurement (ARM) Program's Southern Great Plains (SGP) site. A host of instruments measure atmospheric profiles of many different variables every 30 minutes, and ARM data are often taken as ground truth because of their accuracy and long-term consistency. We used a three year record (1999-2001) of 35-level profiles of equivalent potential temperature, $\theta_e$, and equivalent saturation potential temperature,$\theta_{es}$, and compared them to data for the same location and time period produced by the Geophysical Fluid Dynamics Laboratory's (GFDL) AM2 model. AM2 produces values every 20 minutes for variables including $\theta_e$ and $\theta_{es}$ at the same pressure levels as found in the ARM data. The levels are, in order of decreasing altitude, (in millibars): 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 525, 550, 575, 600, 625, 650, 675, 700, 725, 750, 775, 800, 825, 850, 875, 900, 925, and 950.

The two data sets can be thought of as two multivariate time series. ARM has a data point every 30 minutes and GFDL has data point every 20 minutes. Both series begin on January 1, 1999 and end on December 31, 2001. Each data point is a column vector of length 70, with the first 35 components occupied by a profile $\theta_e$, and the second 35 occupied by the coincident profile of $\theta_{es}$. For notational convenience, we refer to the ARM data point at time $t$ as $\mathbf{x}_{t,A}$ and the GFDL data point at time $t$ as $\mathbf{x}_{t,G}$.

A natural way to compare the series $\mathbf{x}_{t,A}$ and $\mathbf{x}_{t,G}$ would be compute the euclidian distances between their values at those $t$ which they have in common. Since ARM collects measurements very 30 minutes and GFDL outputs a prediction every 20 minutes, the only time points in common are on the hour. This strategy would therefore amount to decimating both data sets, and throwing away information. One could come up with an averaging scheme to align the time points, but averaging too destroys information. More importantly, while it is possible to perform such computations and visually inspect the results in the case of data from one site, it is less likely to be possible if observational time series were available at many sites. We hope that this will in fact be the case when satellite derived observations make their way into routine climate model diagnosis and validation activities. Finally, even

with data from just one site, it's difficult to imagine how one would visualize and understand relationships between all pair-wise combinations of the 70 quantities, let alone higher order interactions.

In this study we take a different approach. Instead of reducing the data for a given time window to point statistics for each variable individually and analyzing their evolution over time, we reduce the data by estimating their multivariate distributions for a given time window. In this paper that time window is a single, three-year snapshot. In the future, we will apply the same set of techniques to higher-resolution time windows, and look at their evolution over time.

## III. Estimating Multivariate Distributions

Given 70-dimensional observational data for the ARM site from ARM itself and from the GFDL model, we illustrate by constructing grand summaries for each data source for the entire three year period. We use a modified version of the ECVQ algorithm ([1], [2]) to partition the entire series $\{\mathbf{x}_{t,A}\}$ for all $t$ (every 30 minutes from January 1, 1999 to December 31, 2001) into a set of disjoint groups called clusters. 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 [1].

ECVQ can be seen in at least three different ways, shown in Figure 1. First, it is a penalized clustering algorithm. It partitions a collection of multidimensional data points into disjoint groups, 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 a 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 1. Raw data points are $C$-dimensional ($C = 70$ here) observations, $\mathbf{x}$ of which there are many: $N$. Representative vectors are also $C$-dimensional, and denoted $\mathbf{y}$. The cluster analysis assigns each $\mathbf{x}$ 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 the numbers of raw data points assigned to cluster $k$, $M_k$'s, and within-cluster mean squared errors are $\delta_k$'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 histogram in which every data point has weight $1/N$. Data points are grouped to form a new distribution. Here again, the $\alpha$'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 $\alpha$ 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, $\mathbf{y} = \beta[\alpha(\mathbf{x})]$. $\beta$ is called the decoder, and in this application is always the group or cluster centroid. 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 $p_k = M_k / \sum_{k=1}^{K} M_k$.

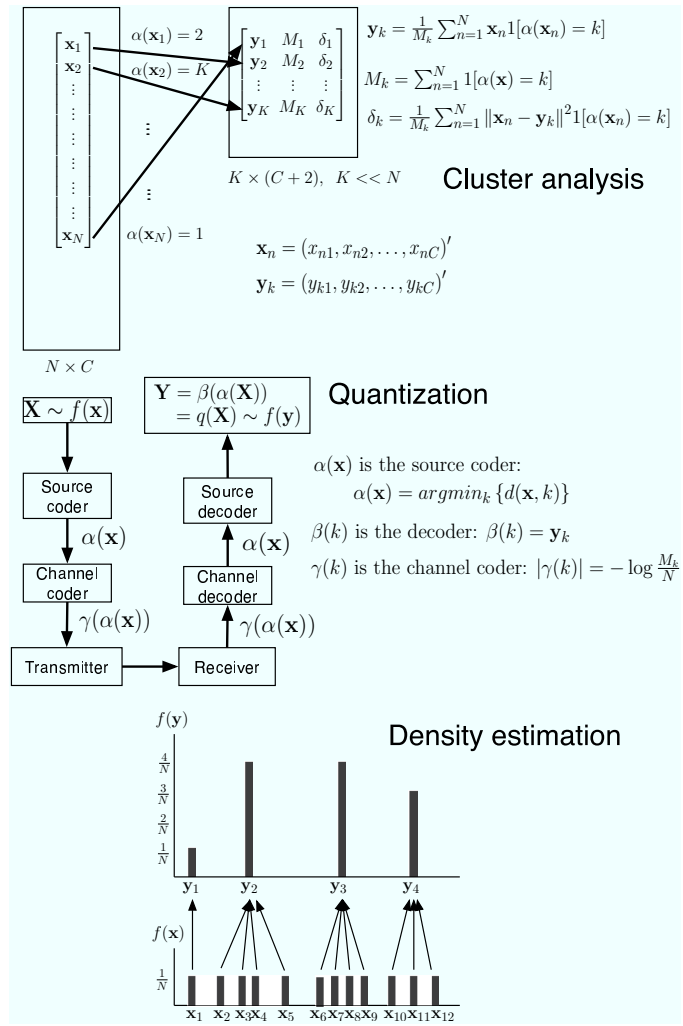


Fig. 1. Three interpretations of the ECVQ algorithm: as a clustering procedure (top), as a quantization algorithm (middle), and as a density estimation method (bottom). $\alpha(\mathbf{x})$ is the source encoding function, returning the index of the cluster to which $\mathbf{x}$ is assigned. $\beta(k)$ is the source decoding function, 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 quantization view reveals something the other two do not: the problem is more complex than simply finding

the optimal assignment $N$ of data points to $K$ clusters, 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.

This constrained optimization problem is solved by formulating a lagrangian objective function of the following form:

$$L_\lambda = \frac{1}{N}\sum_{n=1}^{N}\left[\|\mathbf{x}_n - \beta[\alpha(\mathbf{x})]\|^2 + \lambda\left(-\log\frac{N([\alpha(\mathbf{x})])}{N}\right)\right].$$

$N[\alpha(\mathbf{x})]$ is the number of data points assigned to the cluster $\alpha(\mathbf{x})$. For a given value of the parameter $\lambda$, we find the assignments of raw data points to clusters, $\{\alpha(\mathbf{x}_n)\}_{n=1}^N$ to minimize $L_\lambda$. Our algorithm begins with *random* assignments, and iteratively updates $\{\alpha(\mathbf{x}_n)\}_{n=1}^N$ and the cluster centroids until the algorithm converges. Because the algorithm begins with random assignments, the resulting clustering is also random, so we repeat this process $S = 1000$ times using different random initial assignments each time. Of the thousand resulting clusterings, we choose the one which minimizes distortion, $N^{-1}\sum_{n=1}^N \|\mathbf{x}_n - \beta[\alpha(\mathbf{x}_n)]\|^2$. So, while we iterate to obtain the set of assignments that minimize a combination of distortion and complexity, our final selection criterion is based on distortion alone in order to achieve the best representation of the original data.

The parameter $\lambda$ is the rate-distortion parameter. One can see from the form of $L_\lambda$ that different values of lambda put different amounts of weight on the complexity penalty term. Higher values of $\lambda$ result in more compression. This begs the question of which $\lambda$ to use. The answer is that there is no single correct value: the way in which the data set collapses as $\lambda$ increases is itself and important characteristic of the data being compressed. At $\lambda = 0$, ECVQ reduces to the $K$-means algorithm [3], assigning data points to $K$ clusters in order to minimize the first term in $L_\lambda$. This represents a baseline in terms of distortion and entropy for the chosen value of $K$. In our experiment, we chose $K = 100$ because we feel that 100 is an upper limit on the number of true atmospheric states possible at the ARM site. We then tested $\lambda = 1, 2, 3, 4, 5, 6, 7, 8, 10$. Experience has shown that when the raw data are normalized and projected in a lower dimensional data space capturing at least 95 percent of the variation in the data before clustering is applied, this range of $\lambda$ values spans an appropriate range of possible outcomes. By that we mean that at $\lambda = 0$ we obtain $K$ clusters with baseline levels of distortion and complexity (entropy), and at $\lambda = 10$ we generally obtain very few clusters and an output distribution with low entropy and high distortion.

We do in fact perform the clustering on the normalized, projected data rather than on the original data for two reasons. First, we normalize by subtracting the overall mean and dividing by the overall standard deviation of vector components in order to put all vector components on the same footing. Second, we project the normalized data points into the space spanned by the first eight principal components of the raw data set in order to reduce the dimension of the space in which clustering is performed. We then use the cluster assignments to calculate the cluster centroids in the original 70-dimensional data space so the resulting distribution estimate is easily interpretable from a physical standpoint. We report both versions: the normalized-projected cluster centroids and the physical cluster centroids along with their counts and within-cluster distortions.

We applied the algorithm described in this section to both the ARM and GFDL three-year data sets. In the next section we describe our analysis of the results.

## IV. Distributional Analysis of ARM and GFDL Data

In this section we examine the results of the procedure described in Section III.

### A. Visual Comparisons

The results are one multivariate distribution estimate for each $\lambda$ for each ARM and GFDL. Figure 2 shows the rate-distortion plots that describe the increase in distortion and reduction in complexity as $\lambda$ increases from zero to 10.

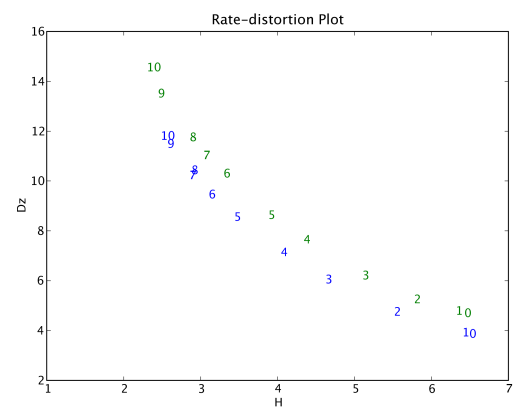


Fig. 2. Rate-distortion plots for ARM (blue) and GFDL (green) data sources. Each marker is positioned at the entropy of the best clustering along the x-axis and the distortion of that clustering along the y-axis. The marker's value corresponds to the value of $\lambda$ used.

Several features are worthy of note. First, both curves are convex and decreasing as predicted by rate-distortion theory [4]. In other words, lower distortion comes at the cost of higher complexity and the trade-off is not linear. Second, the ARM curve (blue) is interior to the GFDL curve (green). This means that achieving the same level of distortion requires greater complexity for GFDL than for ARM. The observations are simpler, or less noisy, than the corresponding model output. Third, the distortions of ARM and GFDL are most similar for $\lambda = 5$. That is, the vertical distance between green and blue markers of the same value is minimized at $\lambda = 5$, and this therefore represents a preferred value of $\lambda$ for comparing ARM to GFDL. The reasoning is that we want the representations of the two data sources to have similar accuracies so that

differences we observe in their distribution estimates reflect true distributional differences, and not differences in their qualities of representation.

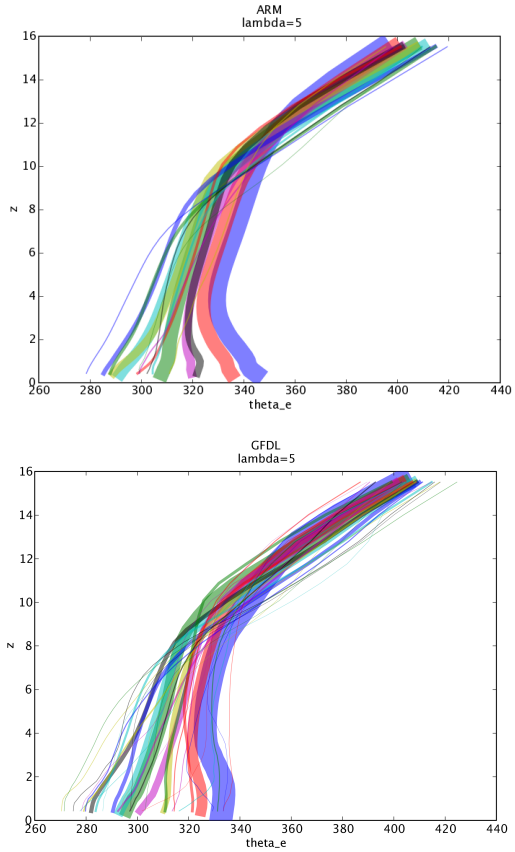


Fig. 3. Distribution of vertical profiles for ARM (top) and GFDL (bottom). This figure show values of $\theta_e$. Each profile is the first 35 components of the representative of one cluster. Clusters are colored to link the same cluster between the $\theta_e$ and $\theta_e s$ views. The widths of the profiles are proportional to the number, or equivalently the proportion, of raw data points represented.

At $\lambda = 5$, ARM has 16 clusters while GFDL has 33. This is consistent with the earlier conclusion that the GFDL data are more complex than ARM data. Distributional mass appears to be more evenly distributed among the largest ARM clusters relative to the largest GFDL clusters. Also, the two largest clusters (red and blue) exhibit quite different behavior in the lower part of the atmosphere (up to 4 km). Subsequent analysis shows that the states they represent occur during the summer. We went back to the original ARM time series and assigned each $\mathbf{x}_{t,A}$ to the nearest cluster representative in the ARM distribution. Then, we looked at the time series of assigned cluster labels and found that the most frequent ARM cluster (blue) accounts for most of the July observations in the record, and the second-most frequent ARM cluster (red) accounts for most of June and August in all three years.

These two clusters occur during the summer when thunderstorms are likely. In both data sources the surface temperature is quite high and the atmosphere unstable, and the most frequent cluster is the warmest at the surface.

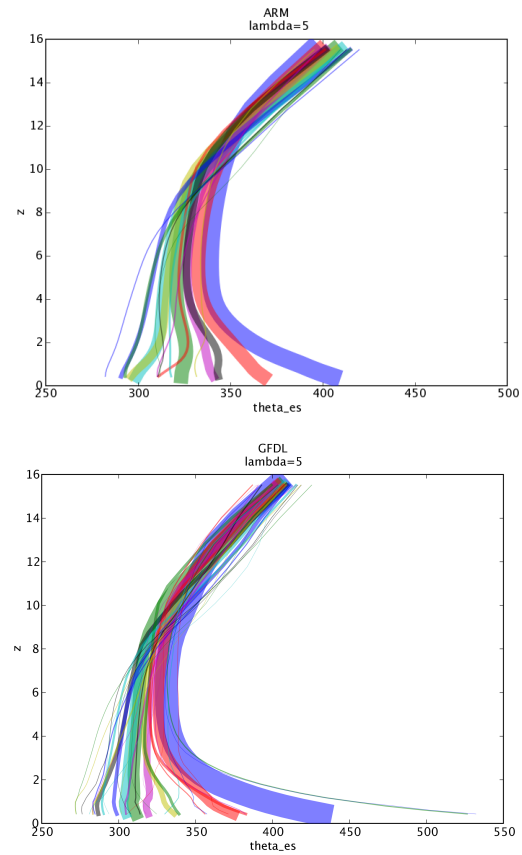


Fig. 4. Distribution of vertical profiles for ARM (top) and GFDL (bottom). This figure show values of $\theta_{es}$. Each profile is the last 35 components of the representative of one cluster. Clusters are colored to link the same cluster between the $\theta_e$ and $\theta_e s$ views. The widths of the profiles are proportional to number, or equivalently the proportion, of raw data points represented.

The clusters from the GFDL model, however, are substantially drier than those of ARM. This reflects a well-known bias in GCMs: summertime thunderstorms at the ARM site are frequently triggered by the eastward propagation of convection that initiates over the Rocky Mountains. This behavior is not usually reproduced by global models which, in the absence of convective rainfall, become too warm and dry over the plains.

### B. Hypothesis Testing

In this section we use the distributions arising from clustering to test various hypotheses about the similarity between the ARM and GFDL multivariate distributions. First, we define a distance between probability distributions as follows. Let $P_1$ and $P_2$ be two distributions, say those of the ARM and GFDL data discussed above. $P_1 = P(\mathbf{Q}_1 = \mathbf{q}_1)$ 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, and their probabilities are the corresponding 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 $i$th and $j$th 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 $P_1$ 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 covariance 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 pmfs is determined by assuming they are as correlated as possible while still satisfying the constraints imposed by their individual pmfs: we give the benefit of the doubt to the assumption that the phenomena described by the two distributions are as correlated as possible.

Our first hypothesis test is $H_0 : P_2 = P_1$. That is, the distribution of the GFDL data actually arose by from a population with the distribution of the ARM data. The test statistic is $\Delta(P_1, P_2)$. We simulate the null distribution (the distribution of the test statistic under the assumption that $H_0$ is true) as follows:

1. Draw a sample of size $N = 26194$ with replacement from $P_1$. The sample defines a distribution that puts mass $1/N$ on sampled value. Collect duplicates to form an empirical distribution, e.g. if cluster 1 is sampled $n_1$ times put mass $n_1/N$ on cluster 1. Call this sample-derived distribution $P^*_{1b}$.
2. Calculate $\Delta^*_b = \Delta(P_1, P^*_{1b})$.
3. Repeat the two previous steps for $b = 2, \ldots, 100$.

The null distribution of $\Delta$ for this test is shown in Figure 5. Now we compare the actual value of $\Delta(P_1, P_2)$ against the null distribution. The actual value is 2.81, which is off the graph to the right. In other words, if the null hypothesis were true, the chances of seeing $\Delta(P_1, P_2) \geq 2.81$ are very small– so small it is for all practical purposes zero. Therefore we reject the null hypothesis $H_0 : P_2 = P_1$. We reach the same conclusion if we switch the roles of ARM and GFDL.

Having established that the GFDL and ARM distributions are not statistically similar, the next tests are aimed

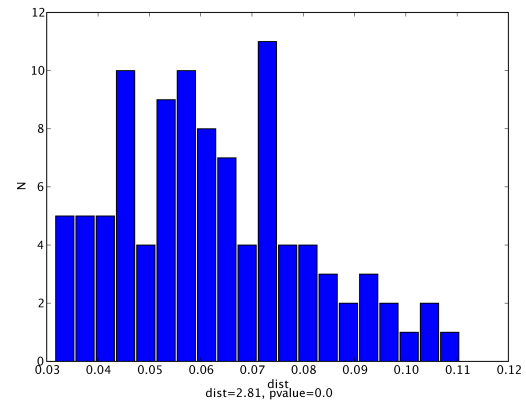


Fig. 5. Null distribution for the test of $H_0 : P_2 = P_1$, where $P_2$ is the GFDL distribution and $P_1$ is the ARM distribution.

at determining what gives rise to the discrepancy. We designate the ARM data as the source clusters and the GFDL data as the test clusters. We want to determine whether a given source cluster has large enough distortion that one could plausibly obtain the representative of a given test cluster by drawing randomly from the source. We assume the source cluster represents a distribution of data vectors with mean value equal to the source cluster's representative and with dispersion bounded by the cluster's distortion.

We make no assumptions about the form of the distributions involved. Instead we rely on Markov's Inequality [5] which states that for any positive random variable $Y$,

$$P(Y > a) \leq \frac{E(Y)}{a}.$$

We let $Y = \|\mathbf{X}_k - \mathbf{q}_k\|^2$ be the squared distance between a random draw from cluster $k$, $\mathbf{X}_k$, and the representative of cluster $k$, $\mathbf{q}_k = E(\mathbf{X}_k)$. Then:

$$P(\|\mathbf{X}_k - \mathbf{q}_k\|^2 > a) \leq \frac{E\|\mathbf{X}_k - \mathbf{q}_k\|^2}{a},$$

$$1 - P(\|\mathbf{X}_k - \mathbf{q}_k\|^2 \leq a) \leq \frac{\delta_k}{a},$$

and

$$1 - \frac{\delta_k}{a} \leq P(\|\mathbf{X}_k - \mathbf{q}_k\|^2 \leq a).$$

Recall that $\delta_k$ is the distortion of cluster $k$. For hypothesis testing at significance level $\alpha = 0.05$, we want the left hand side to be no less than 0.95, so $\delta_k/a = 0.05$, and $a = 20\delta_k$.

The upshot of these calculations is that the probability of obtaining, by random draw from cluster $k$, an observation more than 20 distortion units away from cluster $k$'s representative is less than 0.05:

$$P(\|\mathbf{X}_k - \mathbf{q}_k\|^2 > 20\delta_k) \leq 0.05.$$

We therefore conduct a series of tests of the following form: $H_0$ : the representative of test cluster $j$ could have been drawn at random from a distribution centered at source

cluster $k$'s representative, and with its associated distortion. We reject $H_0$ if test cluster $j$'s representative is more than $20\delta_k$ from cluster $k$'s representative.

Figure 6 shows the results of this family of tests. Cell $jk$ is red if $H_0$ is rejected, and blue otherwise. For example, source cluster 0 is consistent with test clusters 0 and 2, but not with test clusters 1 and 3. Note that the matrix is not symmetric because the numeric cluster labels simply distinguish one cluster from another within the source and test cluster sets. There is no a priori relationship between source and test clusters with the same label numbers.

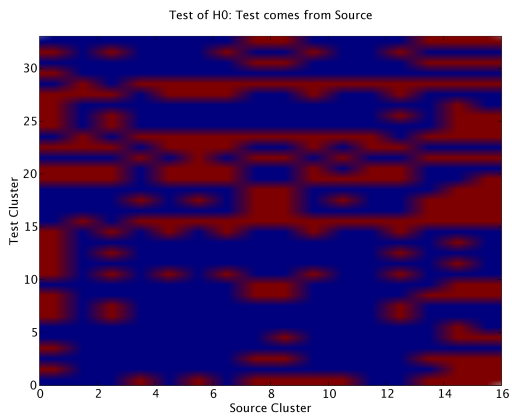


Fig. 6. Results of a family of hypothesis tests, $H_0$ : source cluster gives rise to test cluster . Red indicates the null hypothesis is rejected. Blue indicates it is not rejected

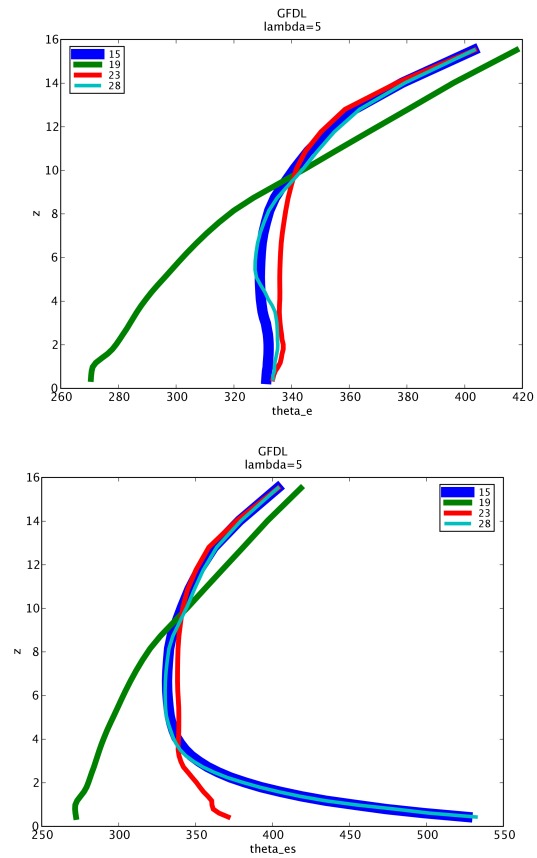


Fig. 7. Vertical profiles for GFDL clusters 15,19, 23, and 28.

It's a bit difficult to interpret Figure 6 at first glance. One can think of scoring the test clusters according to how much blue appears in the row corresponding to the cluster. Lots of blue means the test cluster is centrally positioned within the distribution of the source clusters. That is, a high-scoring test cluster is consistent with many source clusters by the rather forgiving standard used here. By this reasoning the likely culprits causing the hypothesis test $H_0 : P_2 = P_1$ to fail are test clusters 15, 19, 23 and 28. Figure 7 shows $\theta_e$ and $\theta_{es}$ vertical profiles for these clusters alone, without the clutter of the others.

These graphs suggest the following facts about the "problem" GFDL clusters. First, clusters 15 and 28 $\theta_{es}$ values are suspicious below about 2 km. They indicate that the model is much too hot and much too dry for the Oklahoma ARM site at any time of year. This may have to do with the failure of the model to precipitate properly, as we observed previously. Second, in cluster 19 $\theta_e$ and $\theta_{es}$ are nearly identical. Note that the scales on the two graphs are slightly different, making them appear more dissimilar than they really are. This suggests cloudy conditions that are unrealistically stable through the whole atmosphere. Finally, it's not immediately apparent from these graphs why cluster 23 is problematic.

This hypothesis testing framework, while still somewhat ad hoc, does identify problem clusters also identified by visual inspection of the profile plots of $\theta_e$ and $\theta_{es}$. It remains to show the relationship between this cluster-by-cluster hypothesis testing strategy, and the more comprehensive approach that tests $H_0 : P_2 = P_1$ directly.

## V. Summary and Conclusion

This paper reports on work in progress to use multivariate distribution estimates derived from ARM observations and GFDL model output to identify and diagnose discrepancies between models and observations. For proof of concept, we summarized and compared data from the two sources in a single, three year distribution for each. We conducted a formal hypothesis test to verify that the distributions are statistically different, and identified portions of the GFDL distribution likely to be responsible. We also characterized these suspicious model output points.

A single, three year analysis provides a convenient, tractable way to illustrate our approach, but to be of practical value more work is required. The analysis should be repeated on a seasonal or even monthly basis. For example, we need to know from what time period the suspiciously warm and dry GFDL cluster 15 arises. Is it the result of one set of anomalous predictions from one particular time period, or does it repeat periodically?

We also need to establish a theoretical link between the overall distribution-to-distribution hypothesis test, and the cluster-by-cluster tests. The two tests right now are based on different principles: a nonparametric resampling test and the application of Markov's Inequality (also nonpara-

metric). We want to unify them to ensure a consistent set of conclusions.

Our main conclusion from this effort is that the distributional paradigm holds significant promise for illuminating sources of discrepancies between model and observational data sets. However, we are at the early stages, and a great deal of work remains.

## 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] 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.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–296, 1967.

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

[5] S. Ross, *A First Course in Probability*, Prentice Hall, 2002.