



National Aeronautics and  
Space Administration

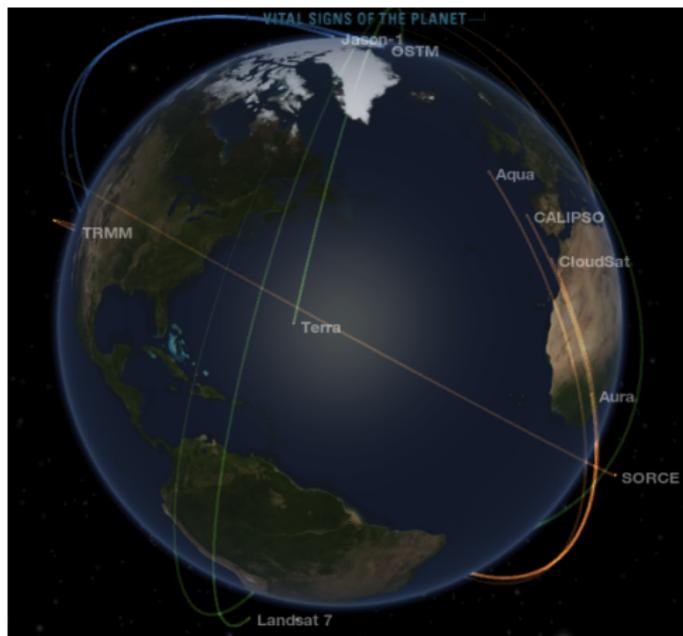
**Jet Propulsion Laboratory**  
California Institute of Technology  
Pasadena, California

# Massive Data Set Analysis for NASA's Atmospheric Infrared Sounder

Amy Braverman   Eric Fetzer   Brian Kahn  
Evan Manning   Bob Oliphant   Joao Teixeira

Jet Propulsion Laboratory  
California Institute of Technology

February 13, 2013



## NASA Eyes on the Earth

- ▶ Introduction
- ▶ AIRS mission and observations
- ▶ Approach
- ▶ Data reduction
- ▶ Algorithms and operational considerations
- ▶ Statistical perspective
- ▶ Data product
- ▶ Example analysis
- ▶ Conclusions



- ▶ NASA's Earth Observing System satellites return massive quantities of multivariate, spatio-temporal data about Earth's atmosphere, ocean, land, cryosphere, etc.
- ▶ These data are typically collected in chunks (called granules) as spacecraft orbits Earth, and downlinked and processed granule by granule.
- ▶ Scientific inference on these data is difficult because of their volume and because they are stored in small (granule) subsets, sometimes in different locations.
- ▶ How can we efficiently discover what's in these data sets so that we can design an appropriate strategy for making inferences from them?



- ▶ We need a data “product” to facilitate this type of exploration by the science community: a reduced data set that captures important statistical characteristics.
- ▶ If we knew what analyses users intended to carry out, we could design a reduced data “product” optimized for those analyses. But we don’t.
- ▶ The product must be created in a way that is doable within NASA’s data processing pipeline: granules can’t be staged all at once, so data reduction must proceed on many small subsets simultaneously with intermediate results combined at the end.
- ▶ Perform exploratory analysis on the set representatives instead of the original data.



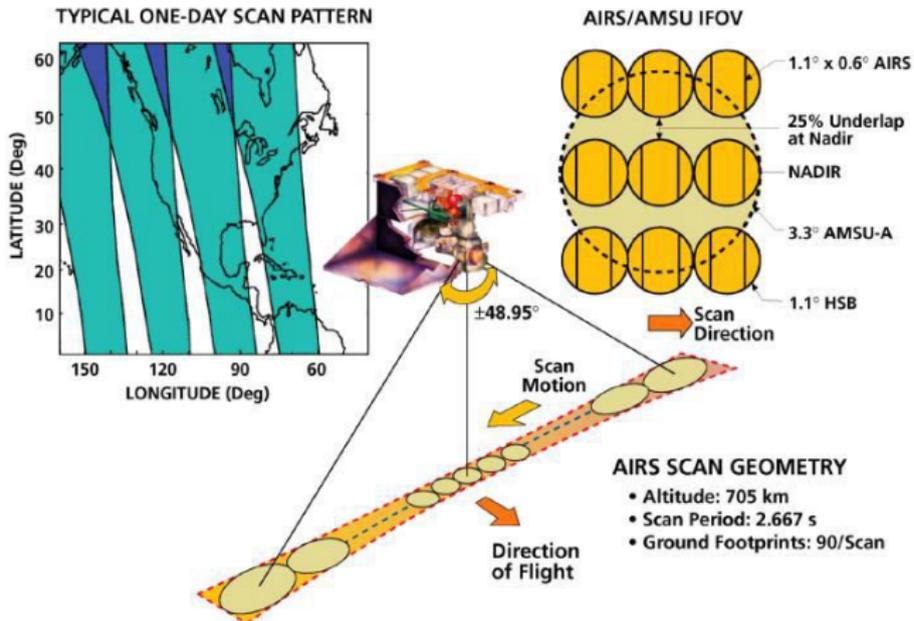
## In this talk:

- ▶ Set forth an approach to data reduction adopted by NASA's Atmospheric Infrared Sounder mission.
- ▶ Discuss how it was implemented operationally.
- ▶ Provide an example analysis that shows what can be learned from it.



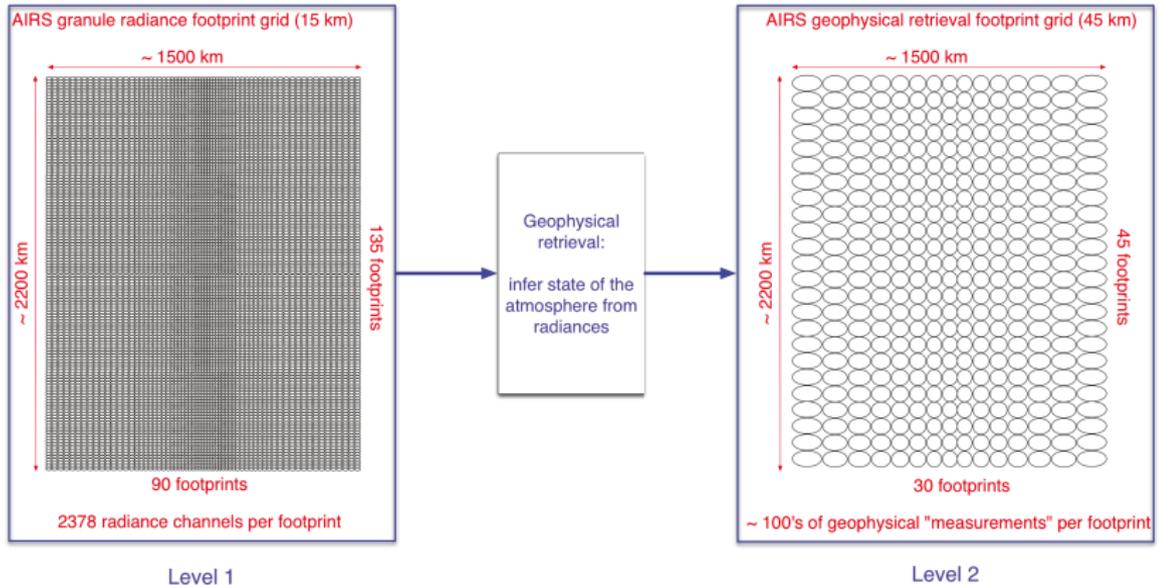
# AIRS Mission and Observations

## AIRS data collection:



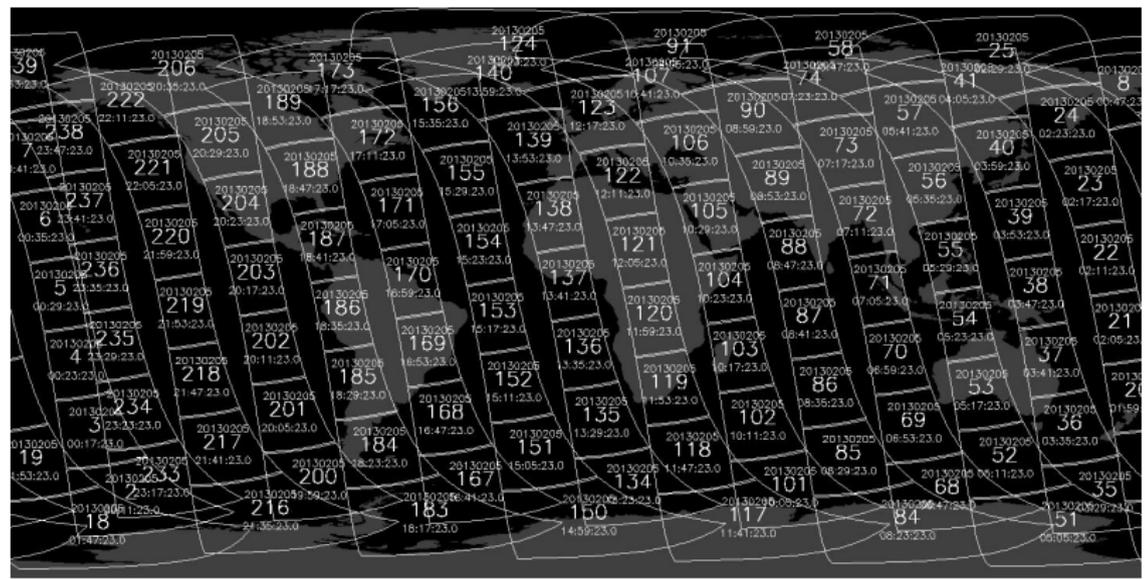


## AIRS data collection:





## AIRS data collection:



Ascending orbits, February 6, 2013.





## Approach:

- ▶ Stratify data on a spatio-temporal grid (e.g., monthly,  $5^\circ \times 5^\circ$  latitude-longitude).
- ▶ Reduce data in each cell in a way that preserves statistical characteristics within and between cells.
- ▶ Build up to monthly data set by reducing five days at a time, then combine these “pentads” to form the monthly.



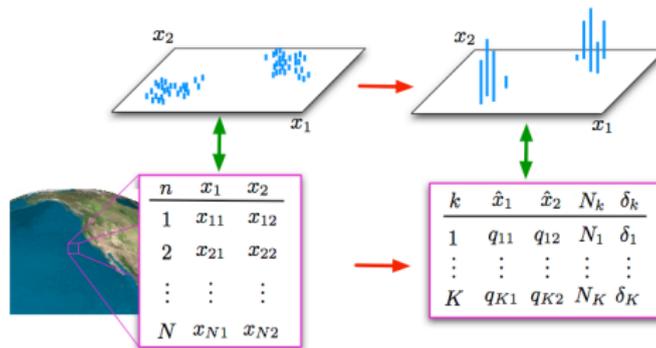
- ▶ For our purposes, each Level 2 footprint represented by a vector of length 35:

<i>Indices</i>	<i>Field</i>
1-11	atmospheric temperature (tair) at 11 levels
12-22	water vapor (h2ommr) at 11 levels
23-32	cloud fraction (cldfrc) at 10 levels; excludes surface
33	land/water indicator
34	quality indicator
35	day/night indicator

- ▶ 240 files per day, each  $45 \times 30$  footprints, since September 2002.
- ▶ Monthly data volume:  $240 \times 45 \times 30 \times 35 \times 8 \times 30 = 2.72 \text{ GB/month}$ .

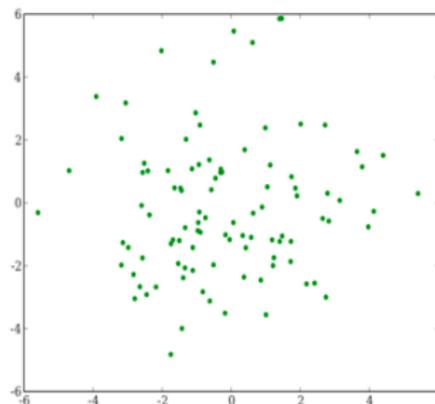
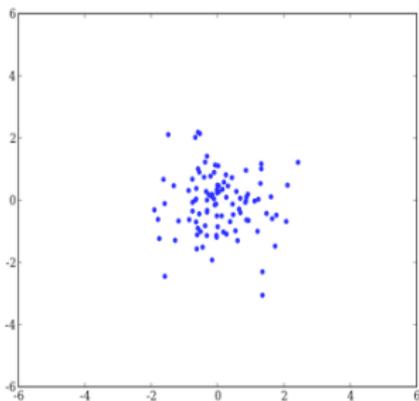


- ▶ How to reduce data in a way that preserves information with minimum data volume?



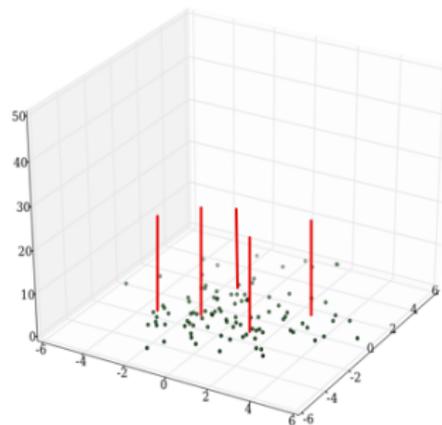
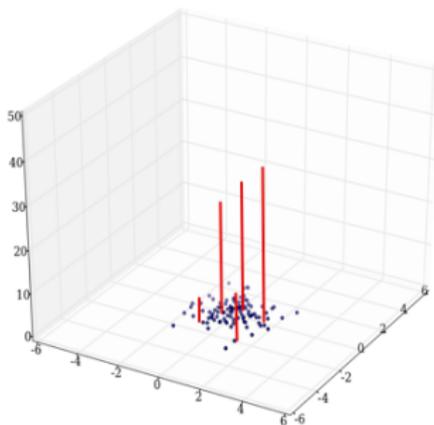


- ▶ How to reduce data in a way that preserves information with minimum data volume?





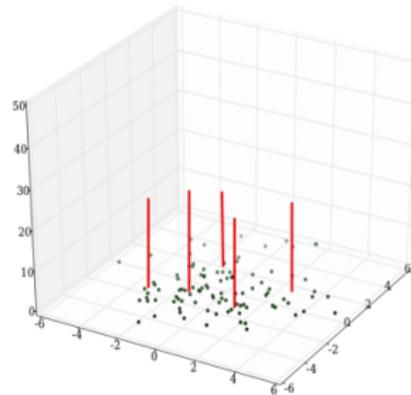
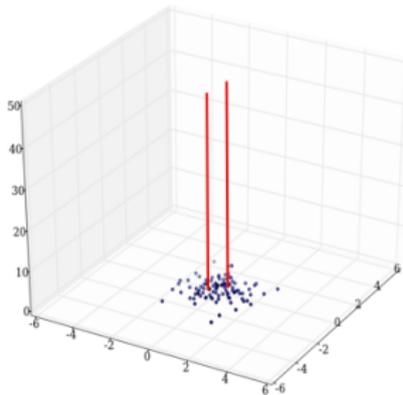
- ▶ How to reduce data in a way that preserves information with minimum data volume?



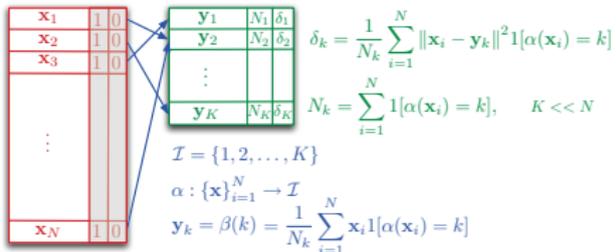
- ▶ Unequal qualities of representation.



- ▶ How to reduce data in a way that preserves information with minimum data volume?



- ▶ Equalize the qualities of representation.





$x_1$	1	0
$x_2$	1	0
$x_3$	1	0
$\vdots$		
$x_N$	1	0

$y_1$	$N_1$	$\delta_1$
$y_2$	$N_2$	$\delta_2$
$\vdots$		
$y_K$	$N_K$	$\delta_K$

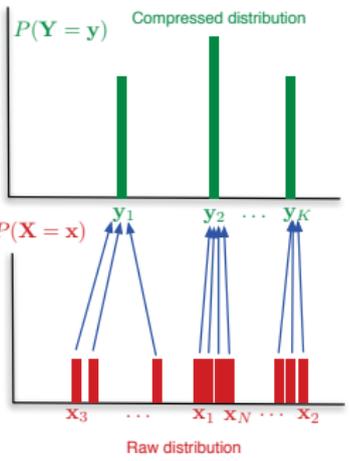
$$\delta_k = \frac{1}{N_k} \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{y}_k\|^2 1[\alpha(\mathbf{x}_i) = k]$$

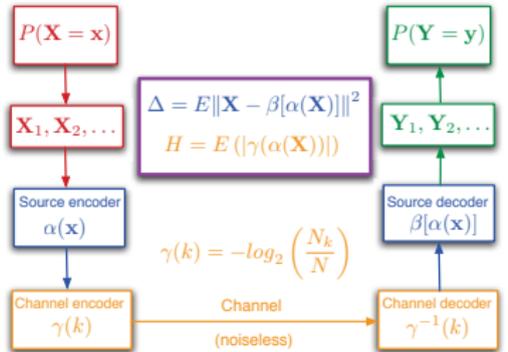
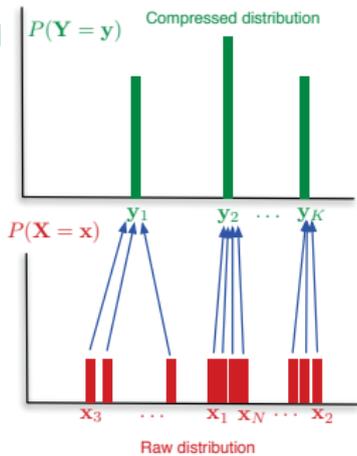
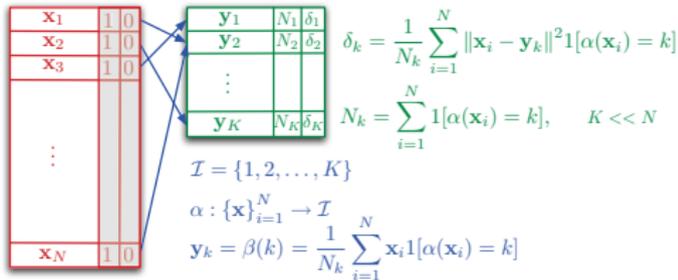
$$N_k = \sum_{i=1}^N 1[\alpha(\mathbf{x}_i) = k], \quad K \ll N$$

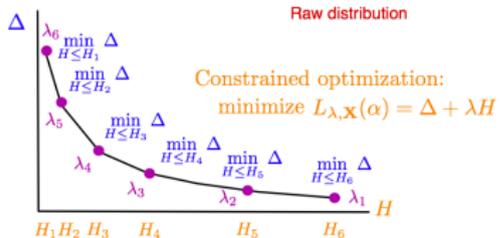
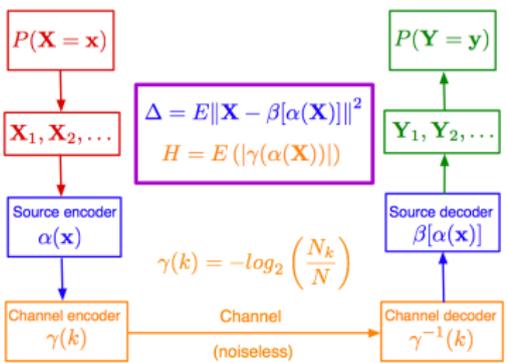
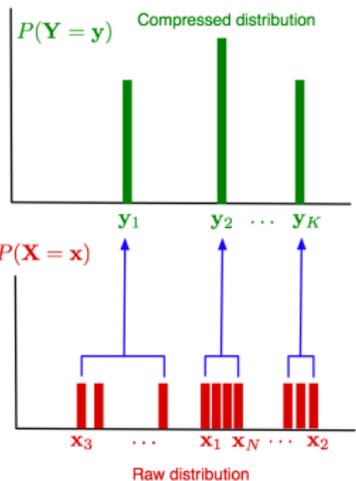
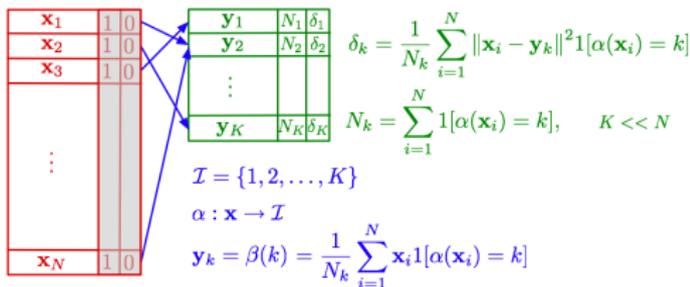
$$\mathcal{I} = \{1, 2, \dots, K\}$$

$$\alpha : \{\mathbf{x}\}_{i=1}^N \rightarrow \mathcal{I}$$

$$\mathbf{y}_k = \beta(k) = \frac{1}{N_k} \sum_{i=1}^N \mathbf{x}_i 1[\alpha(\mathbf{x}_i) = k]$$

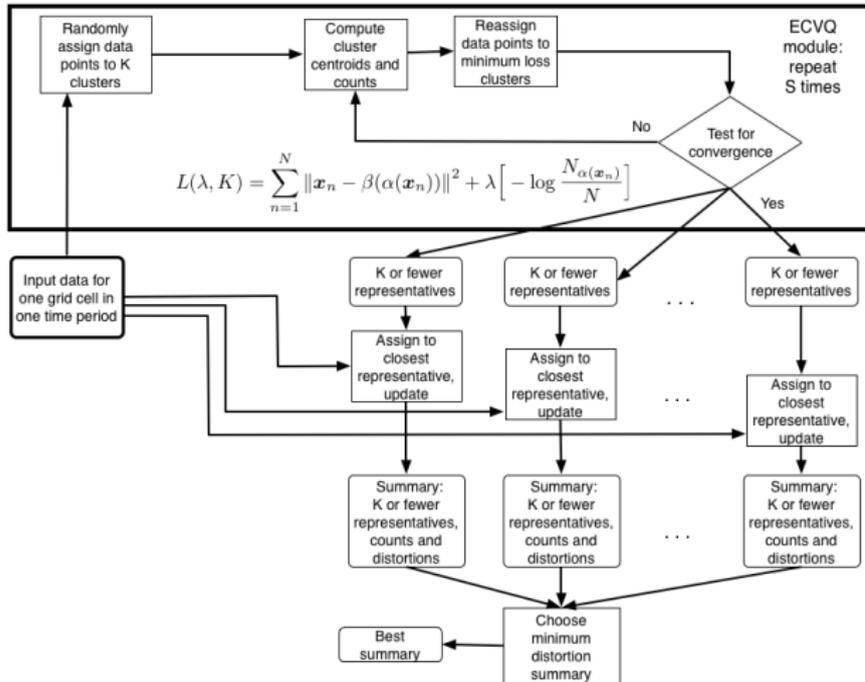






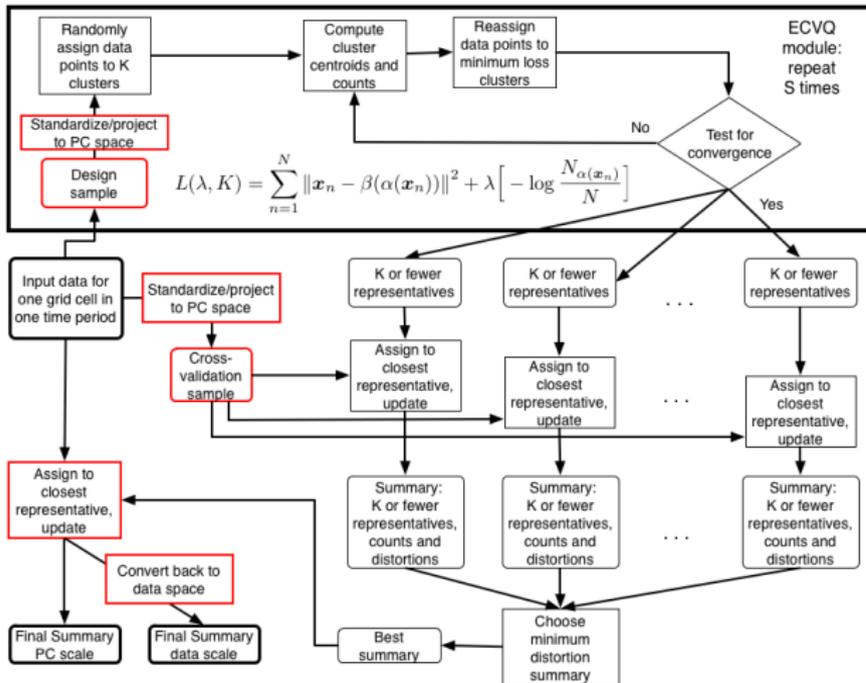


# Basic algorithm



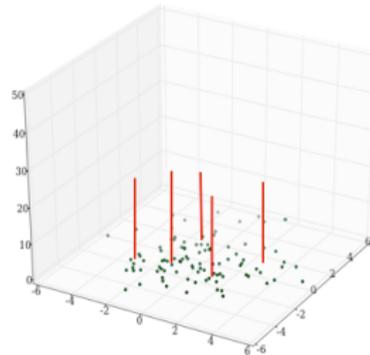
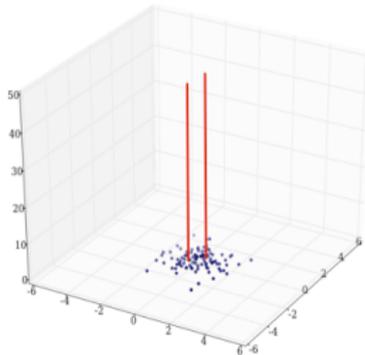


# Massive data set algorithm

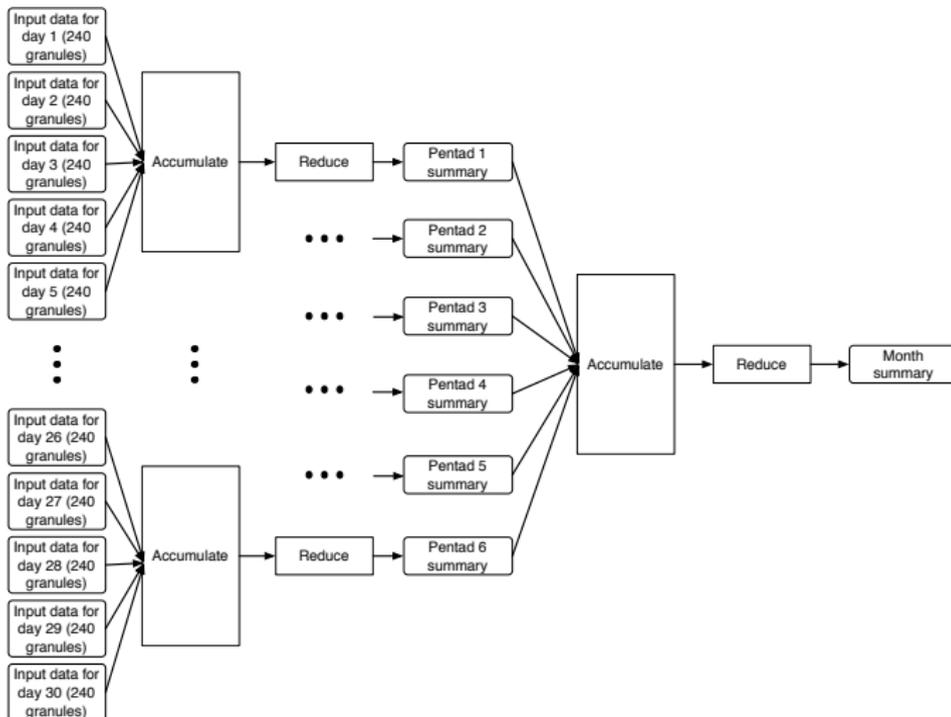


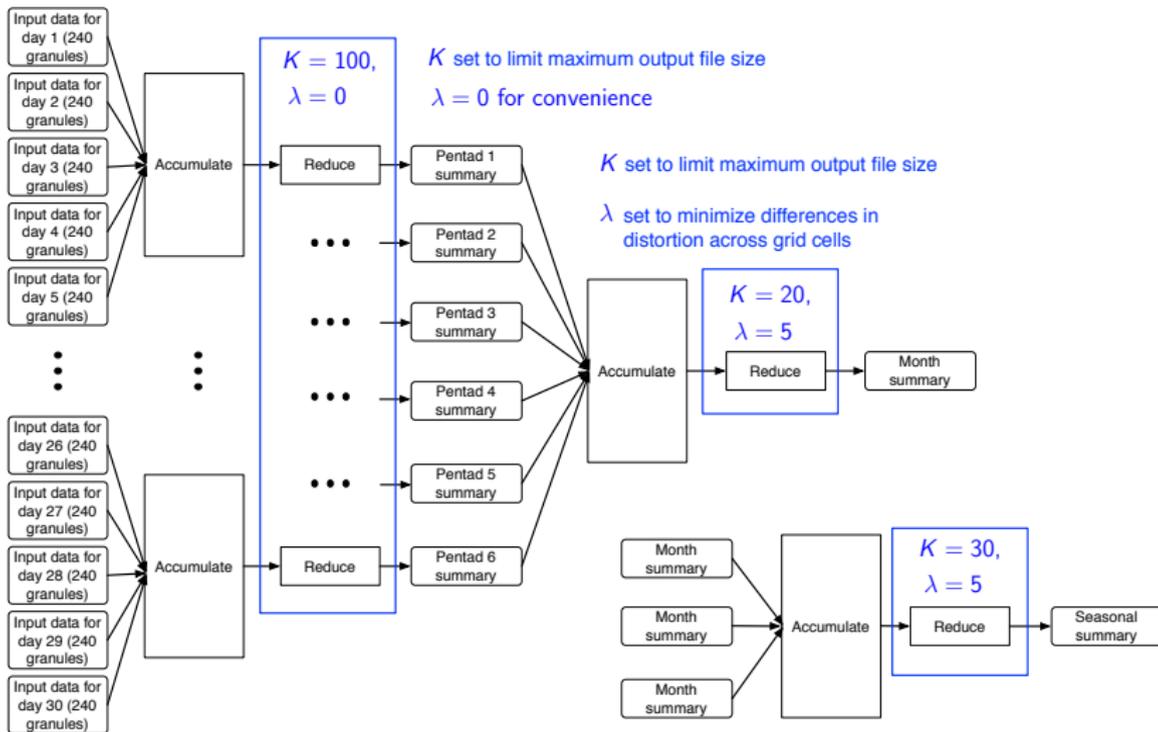


## Setting algorithm parameters $\lambda$ and $K$ :



- ▶  $K$  sets absolute maximum number of representatives. Which regime do you like better?
- ▶  $\lambda$  controls assignments (distribution) within a given regime.







## A statistical perspective

- ▶ Let  $\mathbf{X}$  be a  $(p \times 1, p = 35 \text{ here})$  vector representing a raw observations. The support of  $\mathbf{X}$  is the set of raw observations,  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ .  
 $P(\mathbf{X} = \mathbf{x}_i) = 1/N$  for all  $i = 1, \dots, N$ .
- ▶ Let  $\mathbf{Y} = q_1(\mathbf{X})$  be a function that maps  $\mathbf{X}$  to its representative,  $\mathbf{Y}$ . The support of  $\mathbf{Y}$  is the collection of representatives,  $\{\mathbf{y}_1, \dots, \mathbf{y}_{K_1^*}\}$ , where  $K_1^*$  is the number of representatives.  $P(\mathbf{Y} = \mathbf{y}_j) = N_j/N$ , where  $N_j$  is the number of raw observations assigned to representative  $\mathbf{y}_j$ .
- ▶ By construction,  $\mathbf{Y} = E(\mathbf{X}|\mathbf{Y})$ . We say that  $\mathbf{Y}$  is self-consistent for  $\mathbf{X}$  (Tarpey and Flury, 1996).



- ▶ Note that  $E(\mathbf{Y}) = E[E(\mathbf{X}|\mathbf{Y})] = E(\mathbf{X})$ ; the original and reduced distributions have the same mean.
- ▶ The variance of  $\mathbf{X}$  can be written,

$$\begin{aligned} \text{Var}(\mathbf{X}) &= \text{Var}[E(\mathbf{X}|\mathbf{Y})] + E[\text{Var}(\mathbf{X}|\mathbf{Y})], \\ &= \text{Var}(\mathbf{Y}) + E[\text{Var}(\mathbf{X}|\mathbf{Y})], \end{aligned}$$

which shows that  $\text{Var}(\mathbf{Y}) \leq \text{Var}(\mathbf{X})$  by an amount equal to the average of the within-representative-group covariance matrices.



Self-consistency ensures that data reduction (replacing the distribution of  $\mathbf{X}$  with the distribution of  $\mathbf{Y}$ ) can be applied progressively (pentads  $\rightarrow$  months  $\rightarrow$  seasons) with each stage dependent only on the previous one for inputs, and means and variances equivalent to what would have been obtained had the reduction been done all at once.



- ▶ Let  $\mathbf{W} = q_2(\mathbf{Y})$  be a function that maps  $\mathbf{Y}$  to its representative,  $\mathbf{W}$ . The support of  $\mathbf{W}$  is the collection of representatives,  $\{\mathbf{w}_1, \dots, \mathbf{w}_{K_2^*}\}$ , where  $K_2^*$  is the number of representatives.  $P(\mathbf{W} = \mathbf{w}_k) = N_k/N$ , where  $N_k$  is the number of raw observations assigned to representative  $\mathbf{w}_k$ .
- ▶ By construction,  $\mathbf{W} = E(\mathbf{Y}|\mathbf{W})$  and  $\mathbf{W}$  is self-consistent for  $\mathbf{Y}$ , so  $E(\mathbf{W}) = E(\mathbf{Y}) = E(\mathbf{X})$ , and

$$E(\mathbf{Y}|\mathbf{W}) = E[E(\mathbf{X}|\mathbf{Y})|\mathbf{W}] = E[E(\mathbf{X}|\mathbf{Y}, \mathbf{W})] = E[E(\mathbf{X}|\mathbf{W})|\mathbf{Y}] = E(\mathbf{X}|\mathbf{W}).$$

- ▶ Finally, the variance of the original observations ( $\mathbf{X}$ ) can be reconstructed from the reduced representation,  $\mathbf{W}$ , and the intermediate and final within-representative-group covariance matrices:

$$\begin{aligned} \text{Var}(\mathbf{X}) &= \text{Var}[E(\mathbf{X}|\mathbf{Y})] + E[\text{Var}(\mathbf{X}|\mathbf{Y})] = \text{Var}[\mathbf{Y}] + E[\text{Var}(\mathbf{X}|\mathbf{Y})], \\ &= \text{Var}[\mathbf{W}] + E[\text{Var}(\mathbf{Y}|\mathbf{W})] + E[\text{Var}(\mathbf{X}|\mathbf{Y})]. \end{aligned}$$

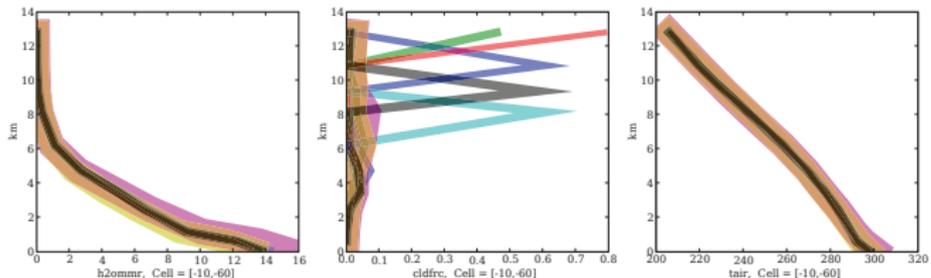


- ▶ Five-day and monthly reduced data product capturing profiles of temperature, water vapor, and cloud fraction at  $5^\circ \times 5^\circ$  spatial resolution.
- ▶ Run in two stages at the Goddard Distributed Active Archive Center (GDAAC).
- ▶ Raw monthly data volume: **~ 2.72 GB.**
- ▶ Reduced monthly data set volume: **~ 12.5 MB.**
- ▶ Processing time to create monthly data product: ~ 87 minutes (six pentads at ~ 51 minutes each, plus one monthly at ~ 36 minutes) on a single, 2.2 GHz AMD Opteron processor.
- ▶ Third stage of processing to create seasonal summaries for 2002 - 2005 run at JPL (~ 36 minutes on a single Mac 3.2 GHz processor). See below.



- ▶ Create summaries of 2002, 2003, 2004, and 2005 winter seasons at five degree resolution.
  - ▶ Winter 200x = Dec 200x, Jan 200(x+1), and Feb 200(x+1).
  - ▶ Constructed from constituent monthly summaries.

Winter 2002, grid cell [lat = -10, lon = -60] (Amazon)



- ▶ How to quantify the evolution of these multivariate distributions in time and space?



- ▶ Calculate the "distance" between two distributions (a measure of similarity).
- ▶  $36 \times 72 = 2592$ ,  $5^\circ \times 5^\circ$  degree grid cells, each containing a distribution.
- ▶ Form a  $2592 \times 2592$  symmetric distance matrix.
- ▶ Use multidimensional scaling (MDS) to analyze the distance matrix.



- Distance between two distributions,  $p$  and  $q$ :  $\Delta(p, q)$

$Y$  is a random draw  
from grid cell 2

$$q_j = P(Y = y_j)$$

$X, Y, p$ 's and  $q$ 's  
are given.

Find  $\pi$ 's to maximize  
 $Corr(X, Y)$

$$p_i = P(X = x_i)$$

$X$  is a random draw  
from grid cell 1

$$\Delta(p, q) =$$

$$\min_{\substack{\pi: q_j = \sum_i \pi_{ij} \\ p_i = \sum_j \pi_{ij}}} E\|X - Y\|^2$$

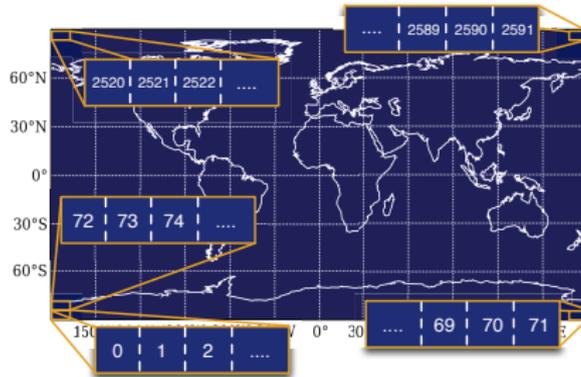
	$q_1$	$q_2$	$q_3$	$q_4$
$p_1$	$\pi_{11}$	$\pi_{12}$	$\pi_{13}$	$\pi_{14}$
$p_2$	$\pi_{21}$	$\pi_{22}$	$\pi_{23}$	$\pi_{24}$
$p_3$	$\pi_{31}$	$\pi_{32}$	$\pi_{33}$	$\pi_{34}$

$$p_i = \sum_j \pi_{ij}$$

$$q_j = \sum_i \pi_{ij}$$



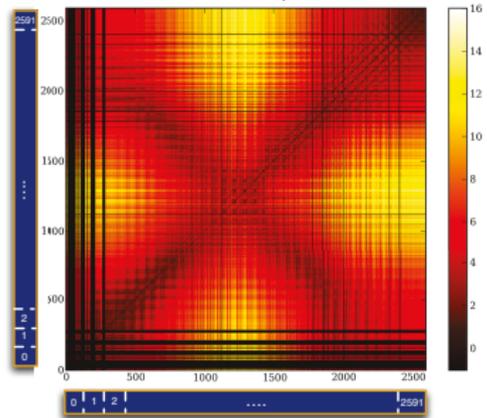
# Example analysis



2592  $5^\circ \times 5^\circ$  grid cells

$$\Delta(p, q)$$

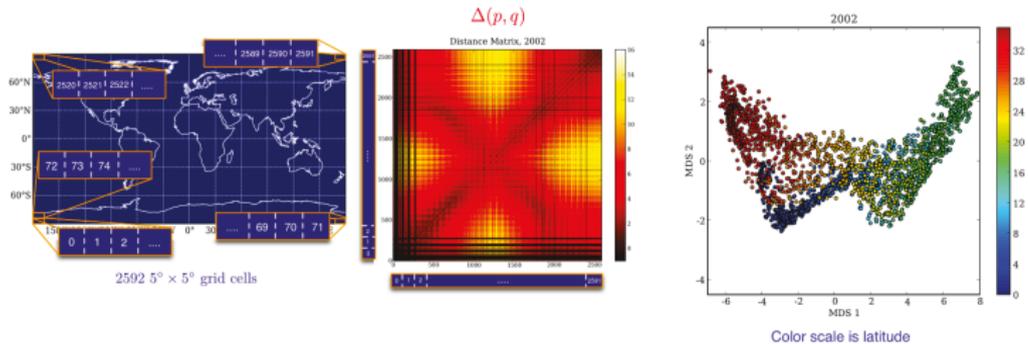
Distance Matrix, 2002





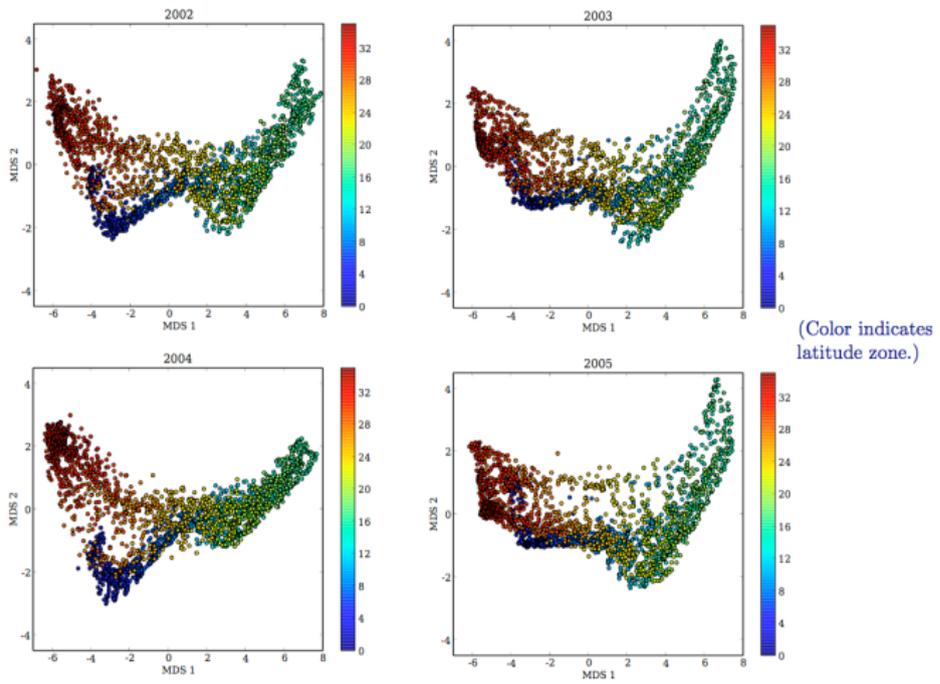
# Example analysis

- ▶ For each object (e.g. grid cell) in a distance matrix, represent it as a point in a low dimensional (e.g. 2) space.
- ▶ Situate the points in the low dimensional space so that relative inter-point distances approximate those in the distance matrix.



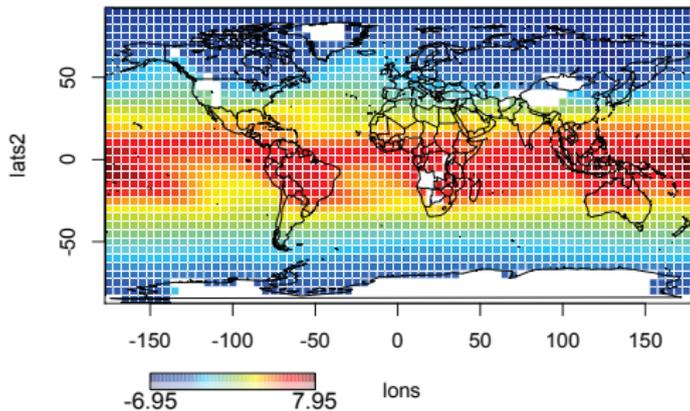


# Example analysis





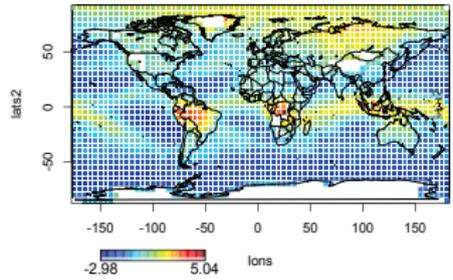
2002 MDS1



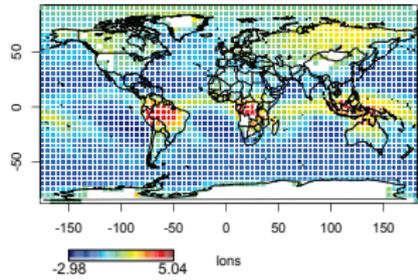


# Example analysis

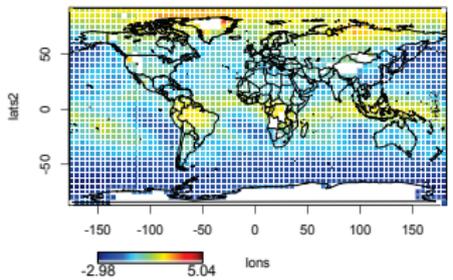
2002 MDS2



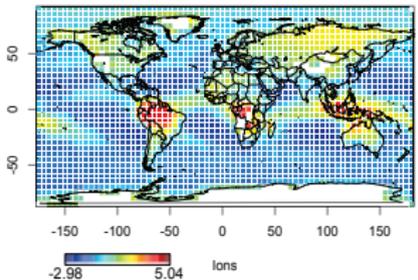
2003 MDS2



2004 MDS2



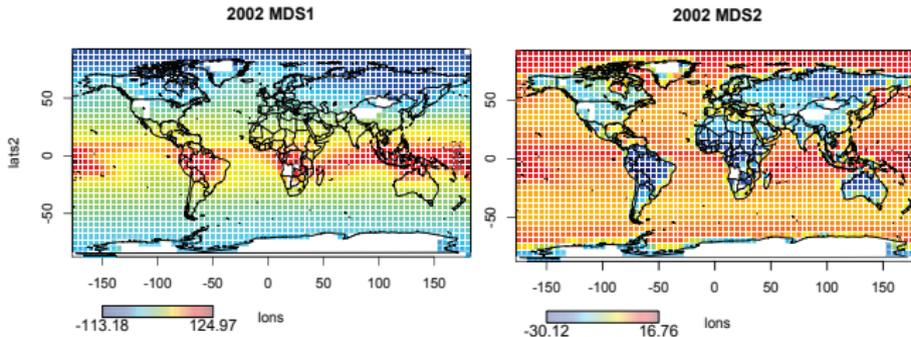
2005 MDS2





## Example analysis

- ▶ Left: MDS1 (2002) obtained from distribution distance matrix.
- ▶ Right: MDS1 (2002) obtained from grid cell mean distance matrix.

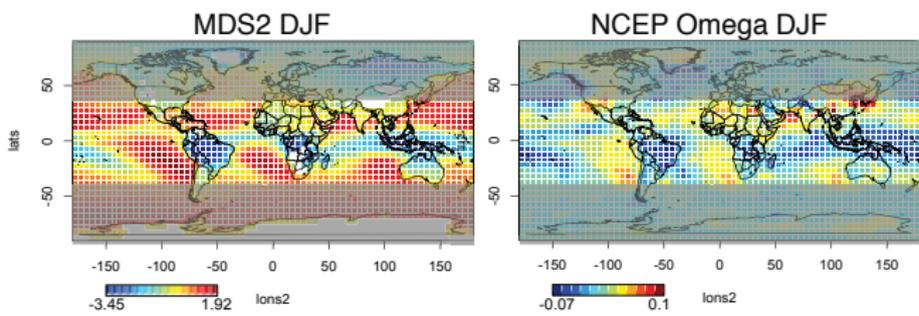


- ▶ Not the same, obviously.



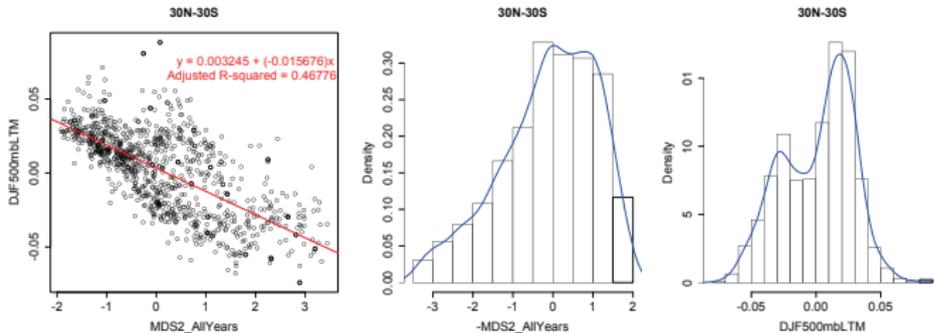
# Example analysis

- ▶ Left: MDS2 averaged over four winters. (Color scale reversed for comparison.)
- ▶ Right: NCEP reanalysis vertical velocity ("omega") averaged over all Decembers, Januarys and Februarys, 1968-1996, and spatially averaged to 5° resolution.





- ▶ Joint (left) and marginal distributions of MDS2 winter average (center) and NCEP reanalysis average “omega” (right) for the region 30°S – 30°N.



- ▶ Suggests that the distributional information in the AIRS data may allow us to tease out vertical velocity.
- ▶ Vertical velocity is a crucial climate variable, but no remote sensing instrument measures it directly.



- ▶ You can find this data product at  
<http://mirador.gsfc.nasa.gov/cgi-bin/mirador/homepageAlt.pl?keyword=AIRX3QPM>.
- ▶ Not enough “science” interest so far: people don’t understand what you can do with it.
- ▶ What can you do with it? *Simulate*:
  - ▶ draw from the distributions, calculate, repeat,
  - ▶ study sampling distributions of arbitrary statistics (with caution).



- ▶ This is a “data mining” product: a descriptive summary of the AIRS Level 2 data.
- ▶ All uncertainty accrues to the relationship between the reduced data and its unreduced parent. Does not account for uncertainty relative to the Earth’s processes.
- ▶ Need to think about how to simulate in a way that accounts for spatial dependence.
- ▶ Algorithm parameters set in a somewhat ad hoc way. (But hey...)
- ▶ In an operational mission (where good Statistics can make a huge difference), optimal is less relevant than do-able.



- ▶ More information in Braverman, A.J., Fetzer, E.J., Kahn, B.H., Manning, E.R., Oliphant, R.B., and Teixeira, J.A. (2012). Massive Data Set Analysis for NASA's Atmospheric Infrared Sounder, *Technometrics*, Volume. 54, Number 1, doi: 10.1080/00401706.2012.650504.
- ▶ Questions/comments? Reach me at [Amy.Braverman@jpl.nasa.gov](mailto:Amy.Braverman@jpl.nasa.gov).

*Copyright 2013, California Institute of Technology. Government sponsorship acknowledged.*

