Principal Components Analysis for AIRS and IASI

L. Zhou, Z. Cheng, M. Goldberg
W. Wolf, T. King, H. Sun, and C. Barnet…

Outline

- Principal Components Analysis (PCA)
- The use of PCA in processing AIRS data
- The use of PCA in processing IASI data
- Data Compression
- Summary and future work

What is PCA?

- Principal Components Analysis (PCA) is a classical approach to extract independent information

- It is commonly used to reduce the dimensionality of a data set with a large number of interdependent variables

- For data with Gaussian distribution, PCA extracts statistically independent features from the data

Why PCA?

- AIRS Granules contain 6 minutes of data
- 240 Granules per day
- 35GB per day

9 FOV

2378 channels

4 bytes/word

- IASI Granules contain ~3 minutes of data
- 491 Granules per day
- 30GB per day

4 FOV

8461 channels

2 bytes/word

Why PCA?

- Data size is very large compared with current sounders (one orbit ~ 2GB vs. 8 MB)
- Information is not independent. PCA can reduce the dimension of the representation and find primary independent pieces of information
- Principal Component Scores (PCS) can be provided instead of the individual channels
- Individual channels can be reconstructed with minimal signal loss

Methodology

- $\mathbf{R} = r_1.\mathbf{i}_1 + r_2.\mathbf{i}_2 + r_3.\mathbf{i}_3 + \ldots + r_n.\mathbf{i}_n$

  where $\mathbf{i}_1 = (1,0,0,0,0\ldots0_n)$; $\mathbf{i}_2 = (0,1,0,0,0\ldots0_n)$

- Data are rotated onto a new set of axes, such that the first few axes have the most explained variance.

- $\mathbf{R} = p_1.\mathbf{E}_1 + p_2.\mathbf{E}_2 + p_3.\mathbf{E}_3 + \ldots + p_n.\mathbf{E}_n$
  where $\mathbf{E}$ are eigenvectors and $p_1 = \mathbf{R}\,\mathbf{E}_1$

- So instead of **R** vectors of length n, we can have a truncated **P** vectors of length m, where m << n

Generating eigenvectors

- Compute the covariance matrix S, from a global ensemble of AIRS normalized spectra (radiances are normalized by expected instrumental noise) .

- An eigenvector decomposition is performed on S to generate the eigenvectors E and eigenvalues Λ

$$S = E \Lambda E^T$$

- E = matrix of orthonormal eigenvectors

  Λ = vector of eigenvalues (diagonal matrix of eigenvalues, $\lambda 1 > \lambda 2 > \ldots > \lambda n$)

Applying eigenvectors

- Compute principal component scores:

$$P = E^T R \quad ; \quad \text{elements of } R = (r_i - r_i\,(\text{mean}))\,/\text{nedr}_i$$

- Invert equation and compute reconstructed radiances R*

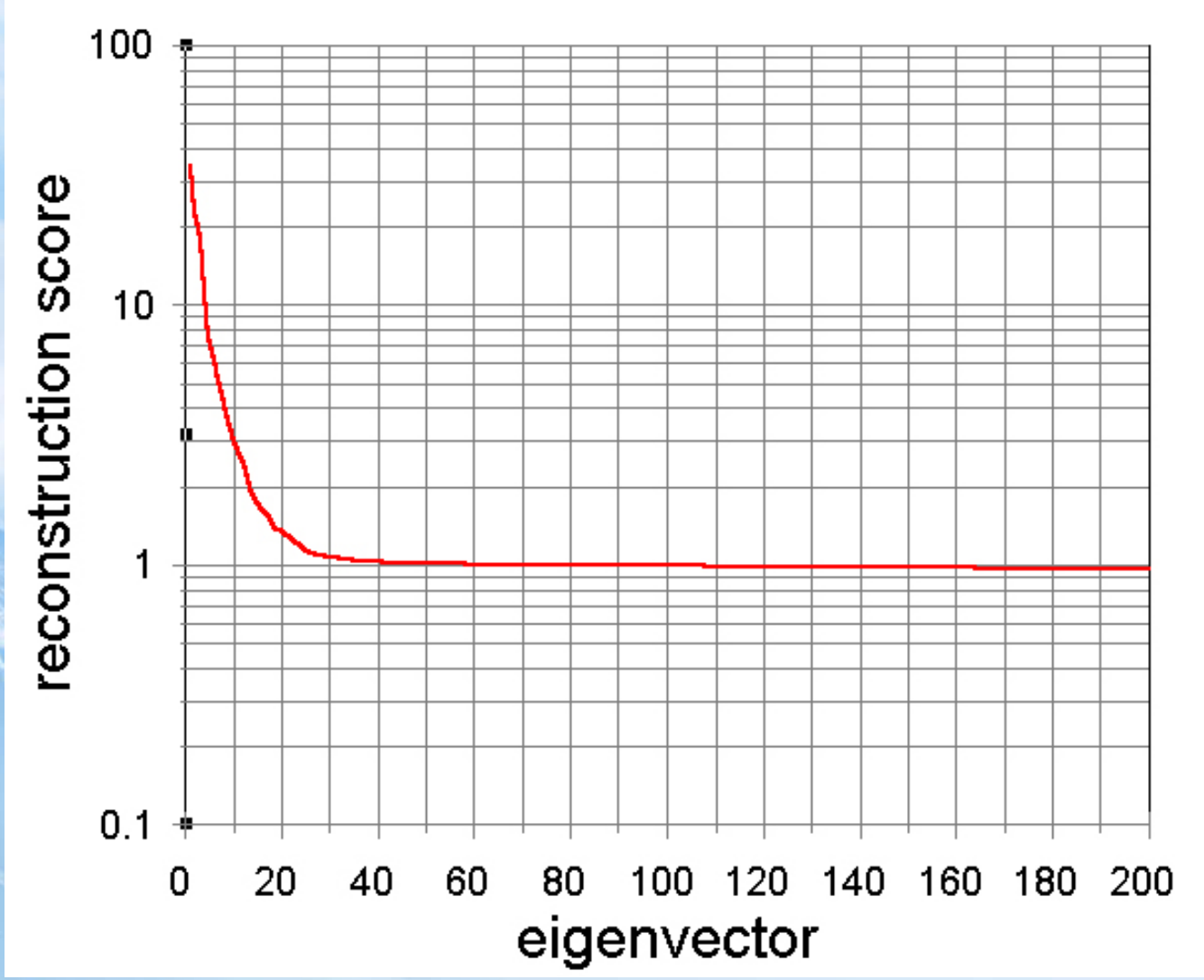
- $R^* = E^* P^*$

- <u>Reconstruction score</u> $= [\ 1/N\ \Sigma(R^*_i - R_i)^2\ ]^{1/2}$

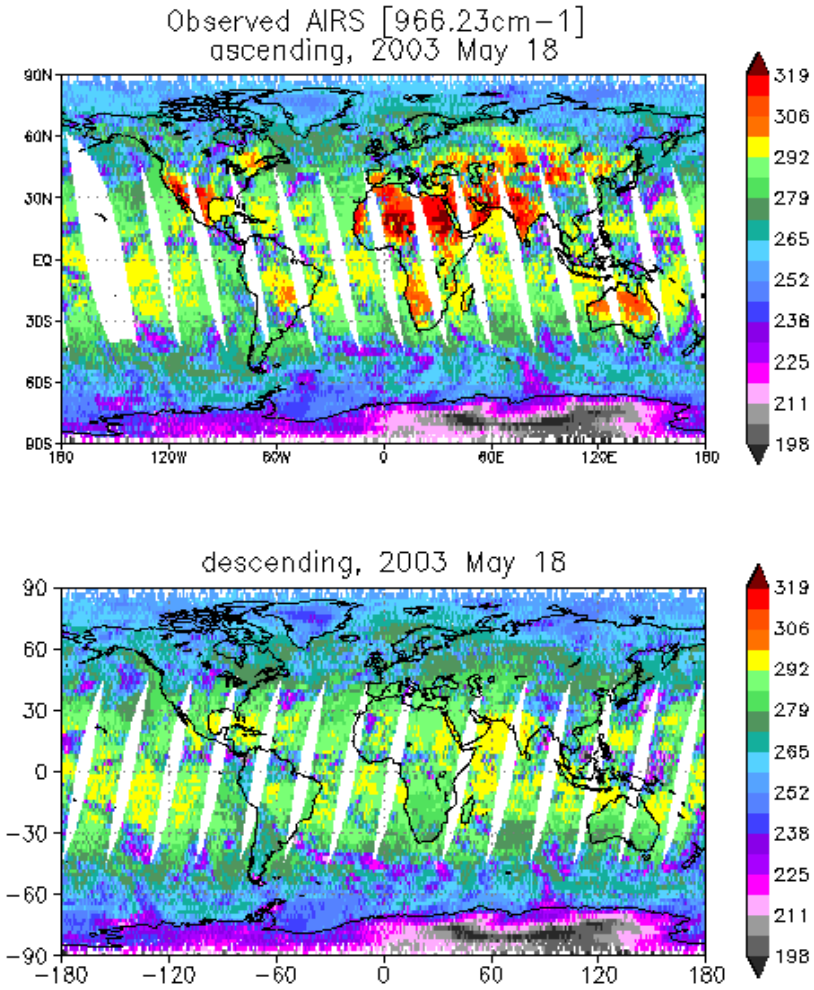
  $i = 1 \ldots N$ channels

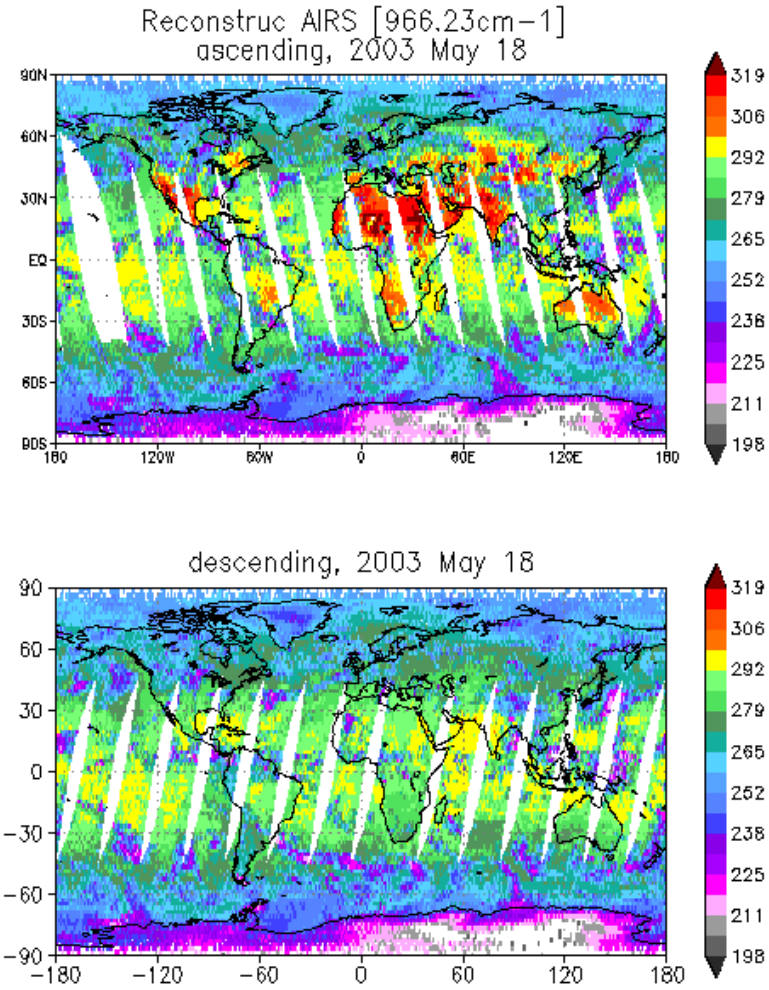
It's a measurement of how well the radiance can be reconstructed. If it is less than 1, it means that the radiance can be reconstructed within the noise level

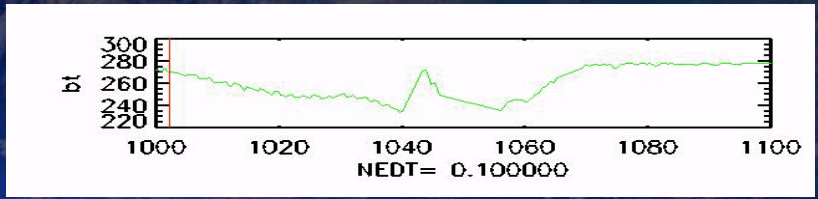
Reconstruction score

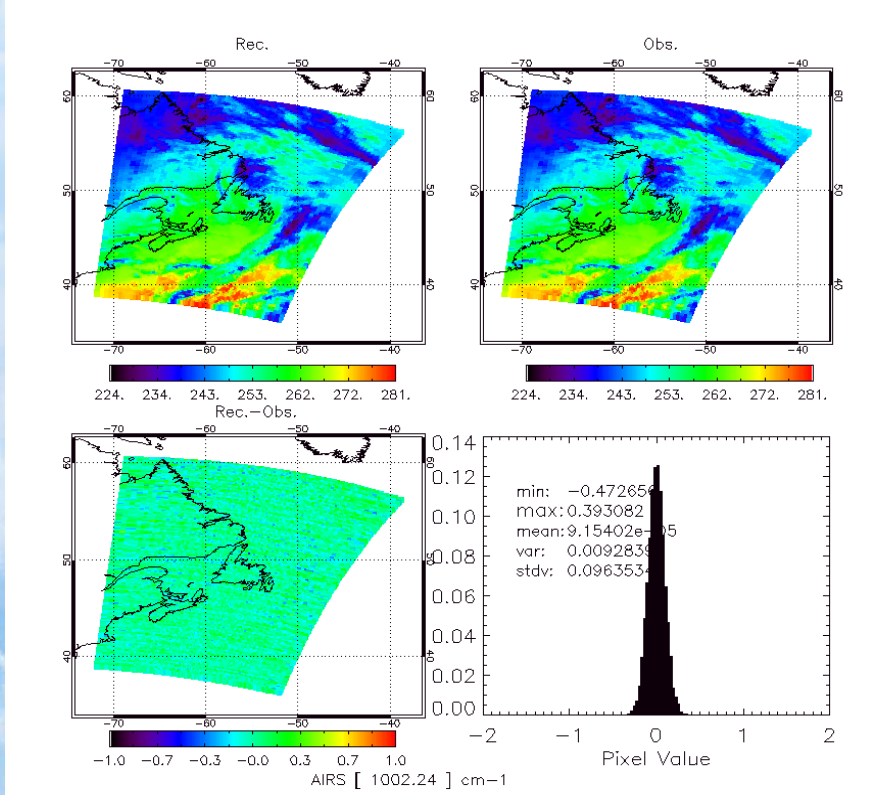


Application 1: Reconstruct Radiance









Ozone Band

Reconstructed; Observed; Difference

Application 2: Regression Retrieval:

•Use PC scores in least squares regression to derive geophysical retrievals (temperature, moisture, ozone profiles)

- the inverse of the predictor matrix is more stable
- the regression solution is computationally fast

In matrix notation the form of the regression coefficients C, dimensioned *m* number of parameters by the *k* number of principal component scores, is

$$\mathbf{C} = \mathbf{X}\mathbf{P}^{*T}(\mathbf{P}^{*}\mathbf{P}^{*T})^{-1}$$

where **X** is a training dependent predictand ensemble matrix, of dimension *m* by sample size *s*. **P*,** the training predictor ensemble matrix, is dimensioned *k* by *S*. On independent data the *m*-dimensioned solution vector is obtained from the matrix multiplication of **C p*** , where **p*** is the independent vector of principal component scores of length *k*.

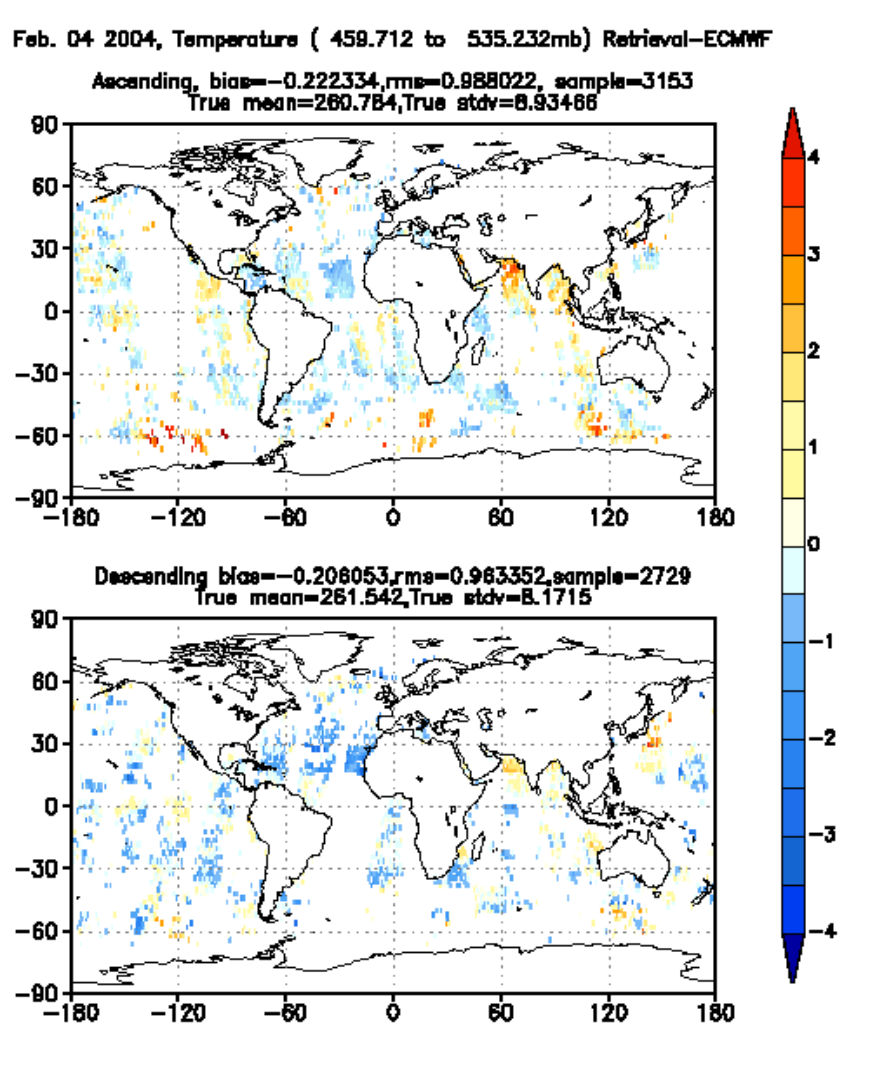
Regression Retrieval(2):

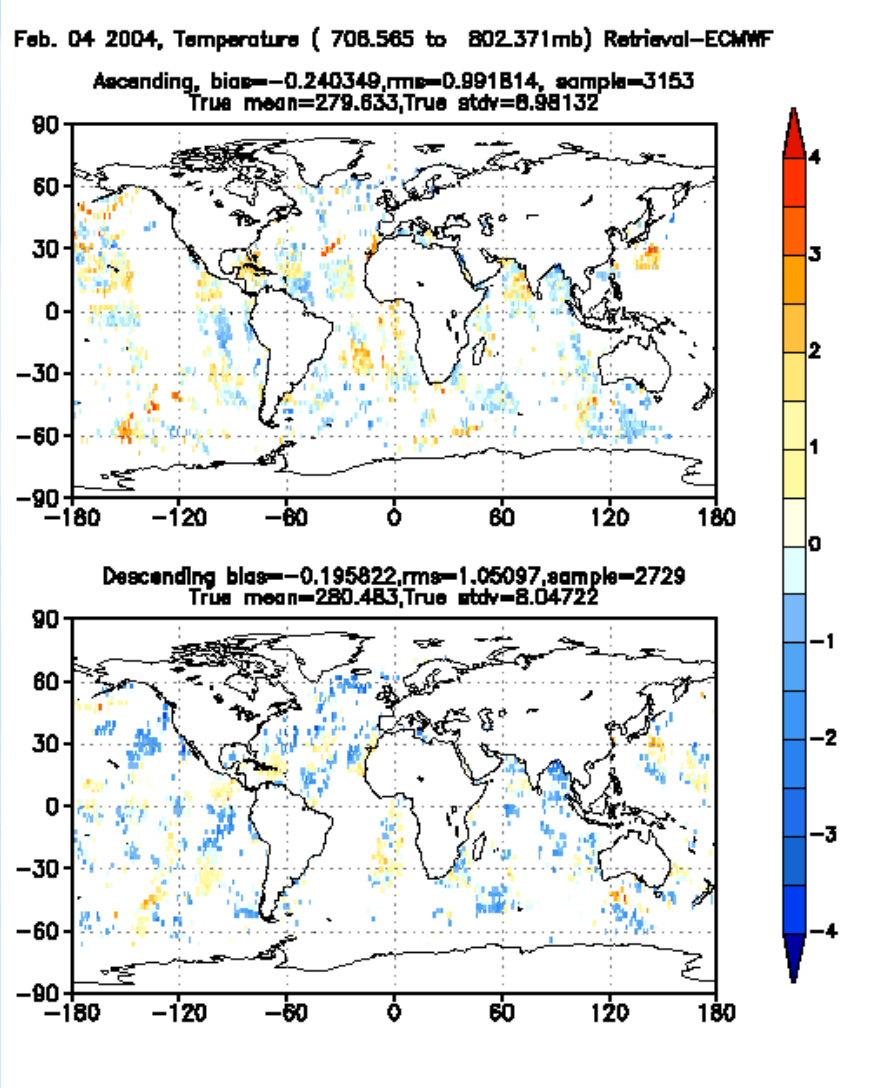
- For regression we used the cloud cleared radiance and ECMWF from three level-2 focus days: Sept. 6, 2002; Jan. 25, 2003; and June 08, 2003.

- Data was selected by screen out where it is too cloudy (using an AMSU predicted airs test), or the cloud cleared radiance can not be reconstructed well.

- Use the first 85 PCS as predictors, generate regression retrievals of atmospheric temperature, moisture, ozone, and skin temperature

Regression Retrieval(3):



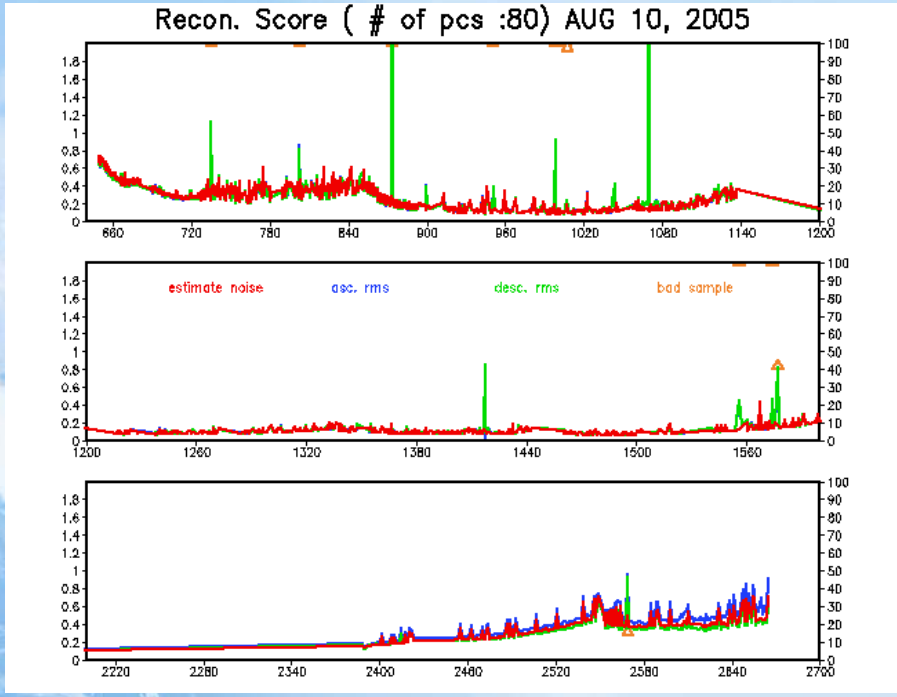


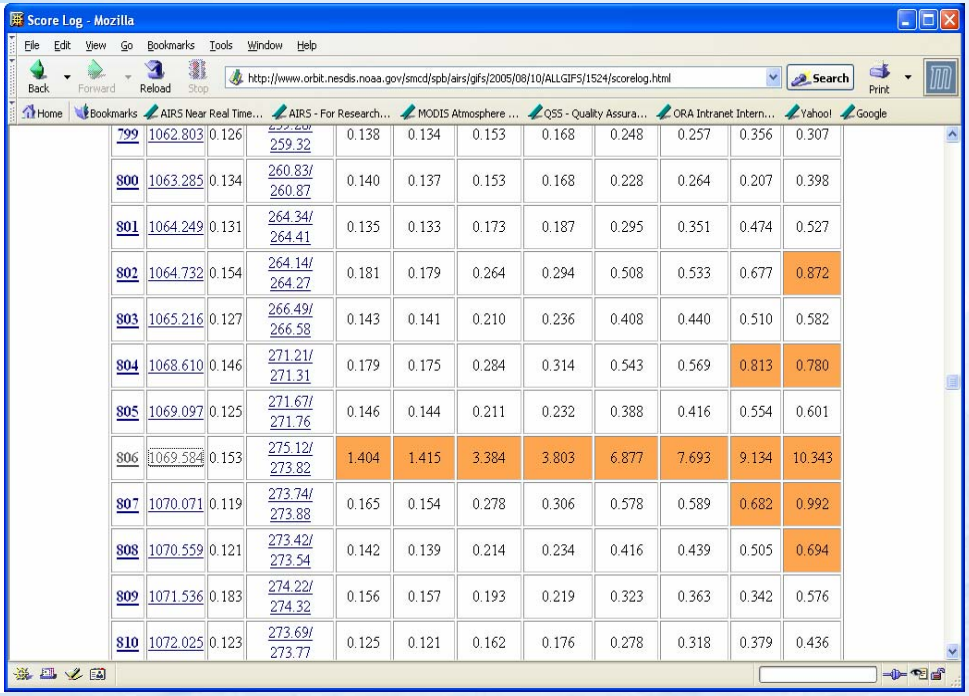
Application 3: Detector Monitoring/Quality Control

- When a large difference between reconstructed and the original radiances is found, it is often due to a problem in the original radiances

- In AIRS operation we use the neighboring channels to compute the pcs, and then reconstruct the erroneous channel.

Detector Monitoring/Quality Control





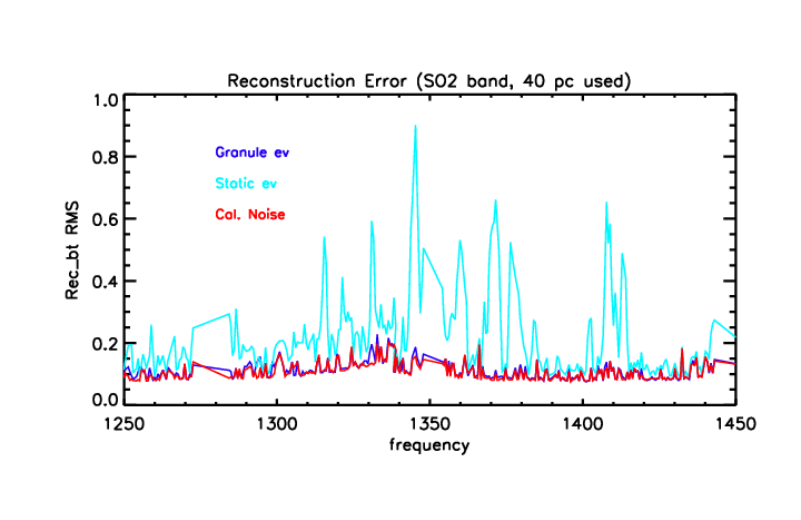
799	1062.803	0.126	259.26/ 259.32	0.138	0.134	0.153	0.168	0.248	0.257	0.356	0.307
800	1063.285	0.134	260.83/ 260.87	0.140	0.137	0.153	0.168	0.228	0.264	0.207	0.398
801	1064.249	0.131	264.34/ 264.41	0.135	0.133	0.173	0.187	0.295	0.351	0.474	0.527
802	1064.732	0.154	264.14/ 264.27	0.181	0.179	0.264	0.294	0.508	0.533	0.677	0.872
803	1065.216	0.127	266.49/ 266.58	0.143	0.141	0.210	0.236	0.408	0.440	0.510	0.582
804	1068.610	0.146	271.21/ 271.31	0.179	0.175	0.284	0.314	0.543	0.569	0.813	0.780
805	1069.097	0.125	271.67/ 271.76	0.146	0.144	0.211	0.232	0.388	0.416	0.554	0.601
806	1069.584	0.153	275.12/ 273.82	1.404	1.415	3.384	3.803	6.877	7.693	9.134	10.343
807	1070.071	0.119	273.74/ 273.88	0.165	0.154	0.278	0.306	0.578	0.589	0.682	0.992
808	1070.559	0.121	273.42/ 273.54	0.142	0.139	0.214	0.234	0.416	0.439	0.505	0.694
809	1071.536	0.183	274.22/ 274.32	0.156	0.157	0.193	0.219	0.323	0.363	0.342	0.576
810	1072.025	0.123	273.69/ 273.77	0.125	0.121	0.162	0.176	0.278	0.318	0.379	0.436

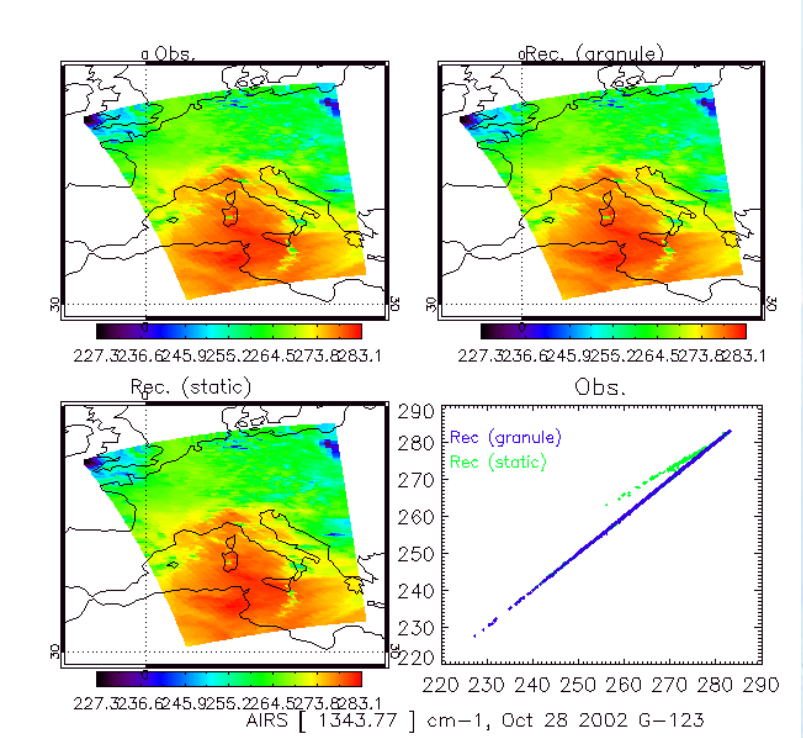
Large reconstructed error can be use to identify suspicious data

Application 4: Diagnose Climate Anomaly Signals – A Case Study

- Static PCS are very stable over time
- It is useful to monitoring the changes in the global environment.
- When an anomaly event occurs, we may see the signature of it in reconstruction scores.
- A case study has been done for Mt Etna volcano eruption.

A Case Study: Mt. Etna Eruption(2)





The use of PCA in processing AIRS data

- AIRS is currently being processed and distributed in near-real time in NESDIS

- There are many important applications of PCA to AIRS processing, such as:

  – Detector Monitoring and Noise Estimate
  – Radiance Reconstruction
  – Regression Retrieval of Physical Parameters
  – Climate Anomaly Signal Diagnostic
  – Data Compression…

The use of PCA in processing IASI data

- Simulated IASI data

  each granule:

  * 2 scan lines
  * 4 IASI FOV
  * 8461 channels

- For easy computation, divided the 8461 channels into three bands

  band 1: 2261 channels $645cm^{-1} \sim 1210cm^{-1}$

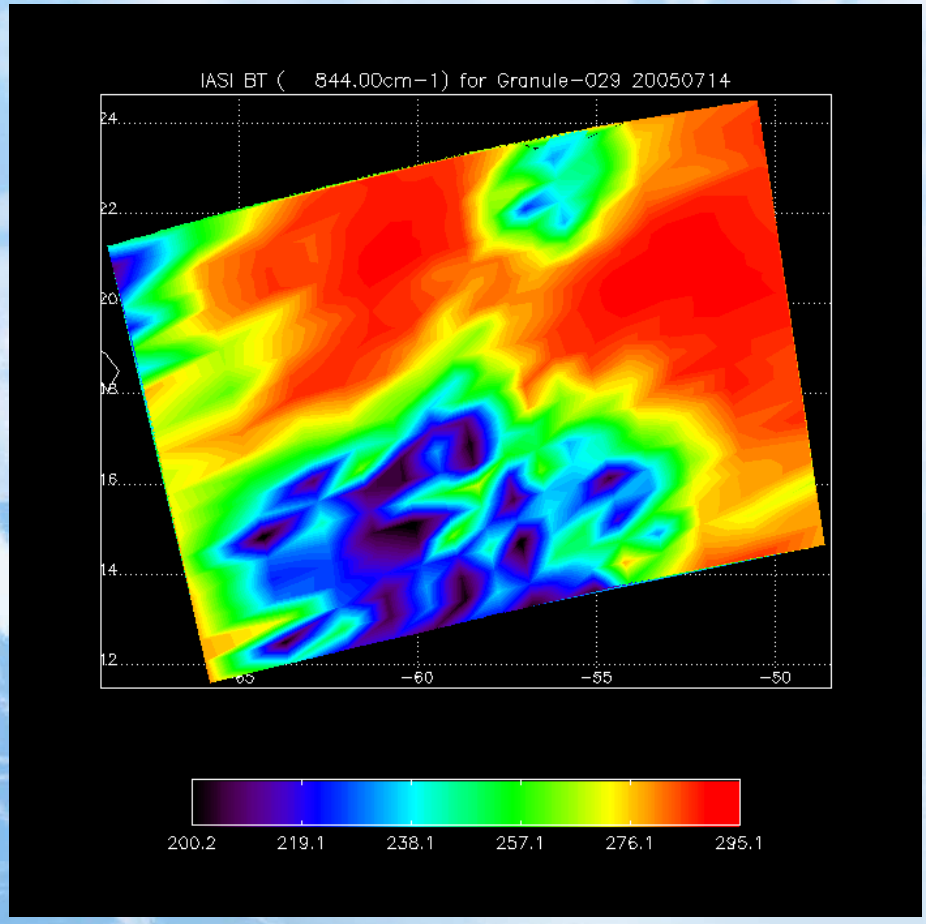
  band 2: 3160 channels $1210.25cm^{-1} \sim 2000cm^{-1}$

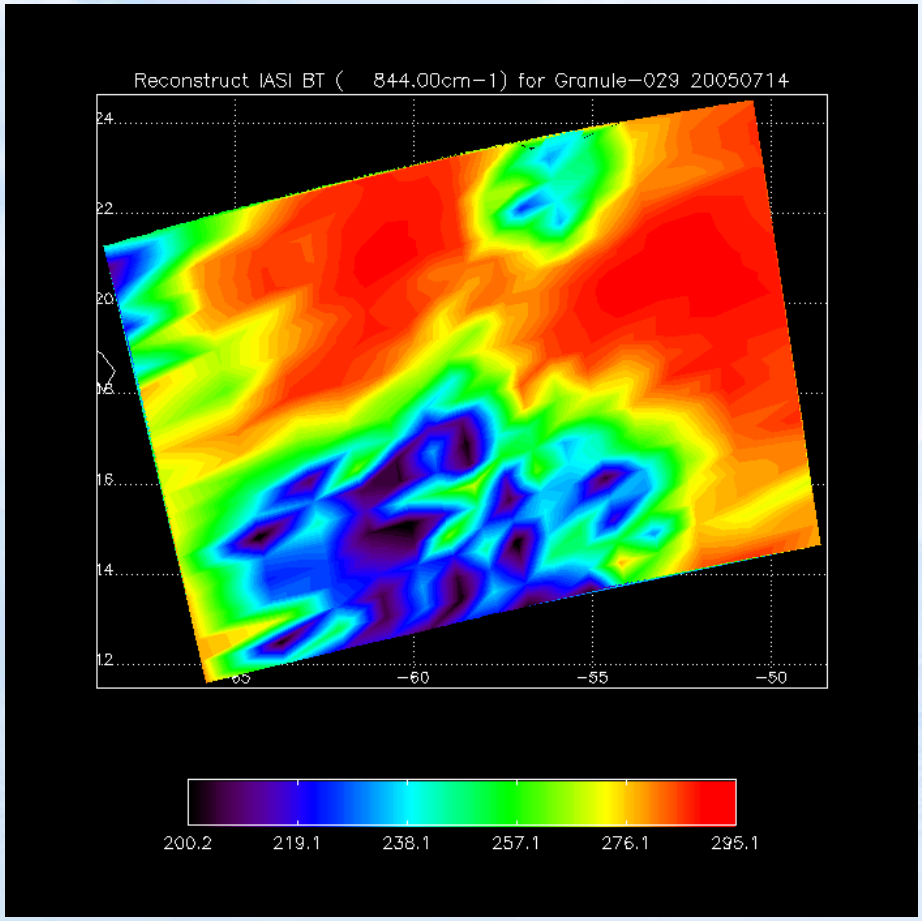
  band 3: 3040 channels $2000.25cm^{-1} \sim 2760cm^{-1}$

The use of PCA in processing IASI data

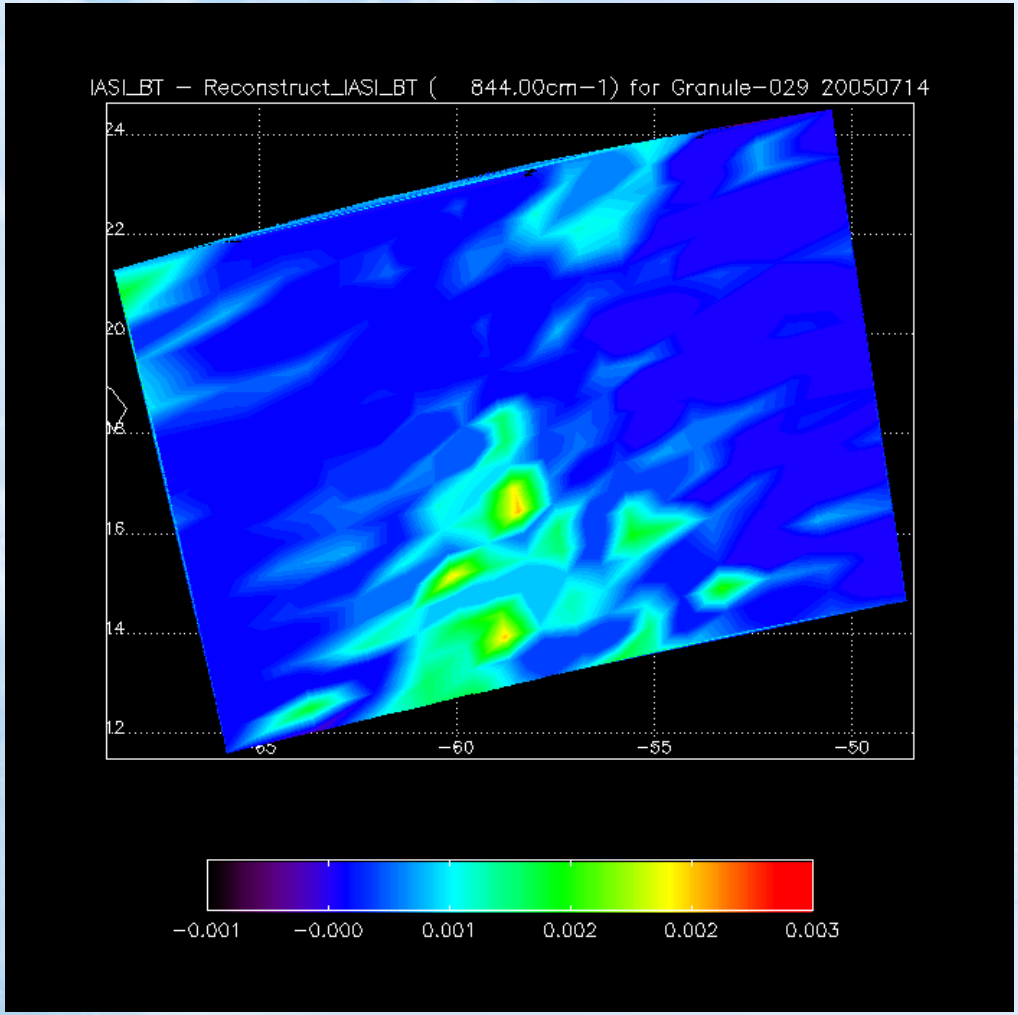
- The algorithm to generate and apply IASI eigenvectors is same as AIRS
- Computed 200 principal components for each band
- Reconstruct the radiance by using principal components
- Compute the PC scores

IASI BT and Reconstruct BT

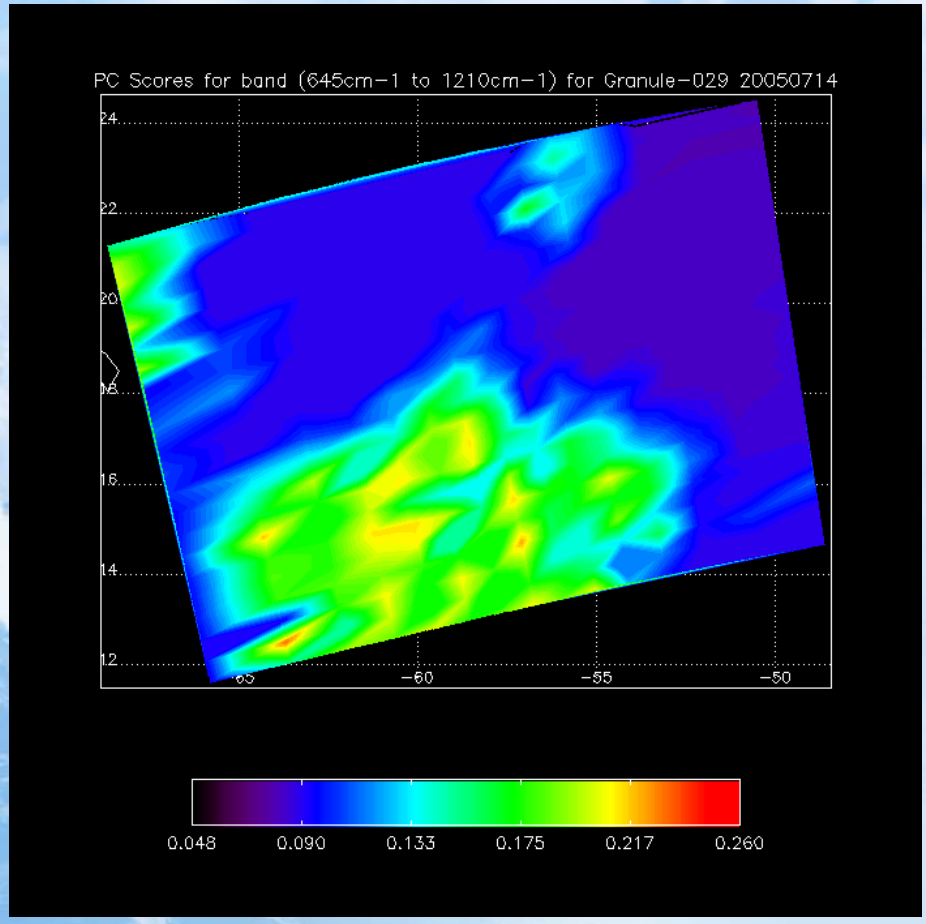


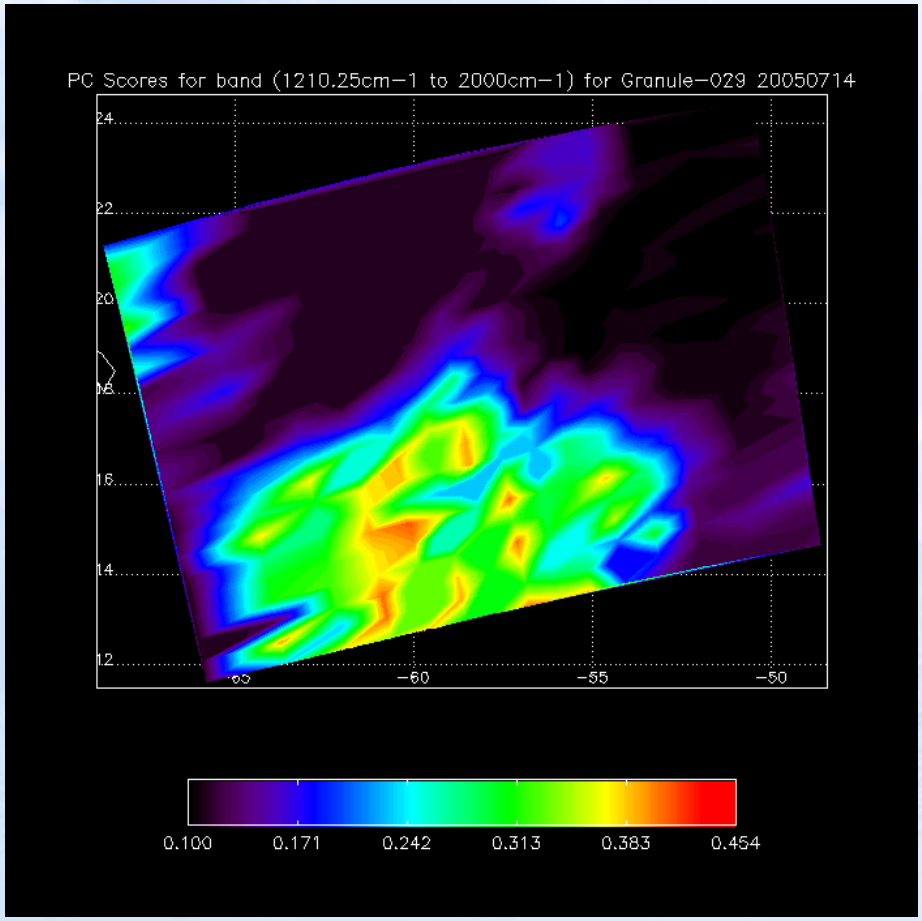


IASI BT - Reconstruct BT

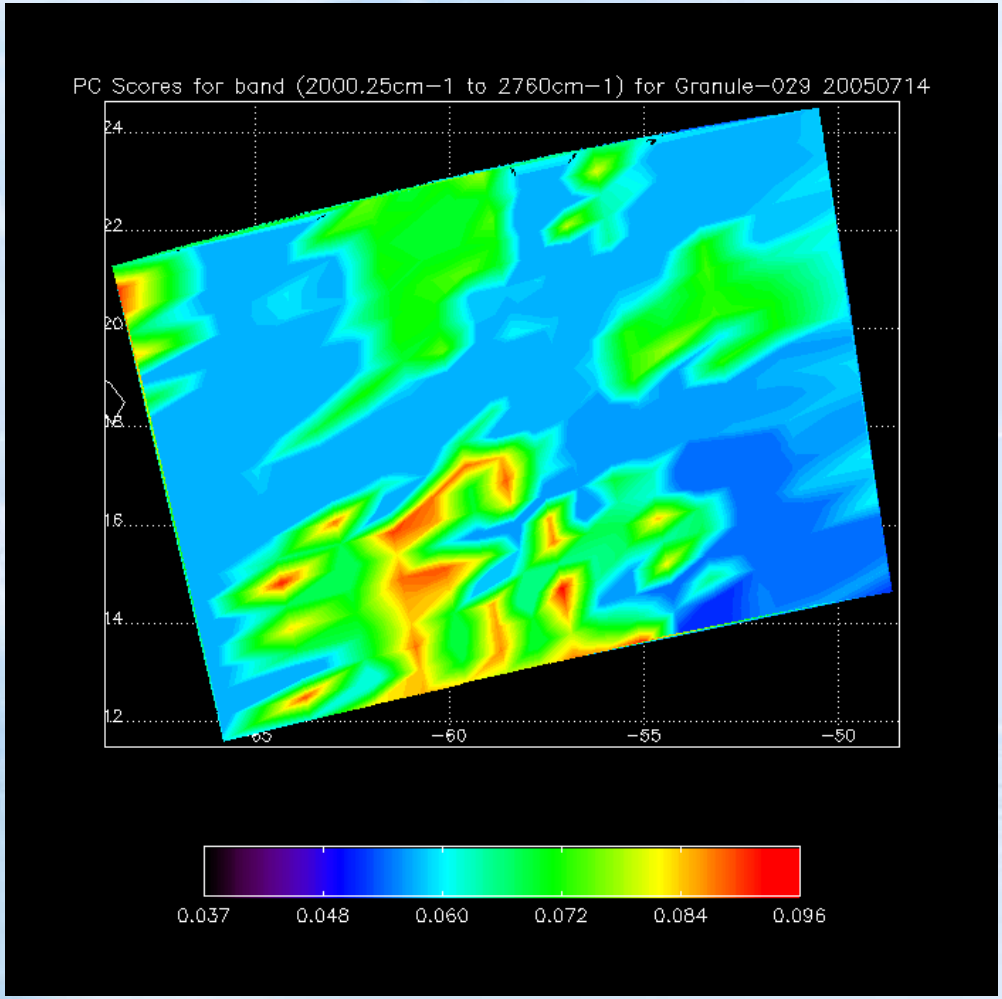


IASI Reconstruction Scores for Band 1 and Band2

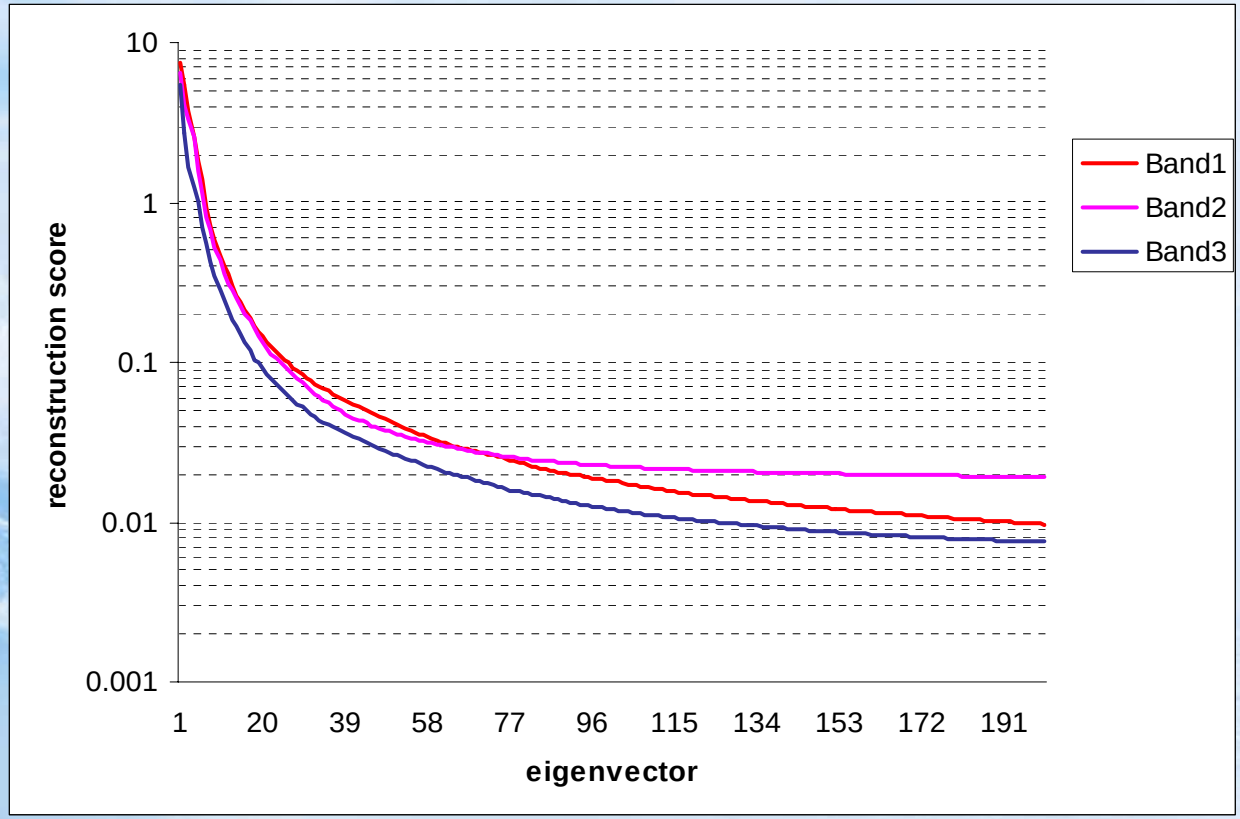




IASI Reconstruction Scores for Band 3

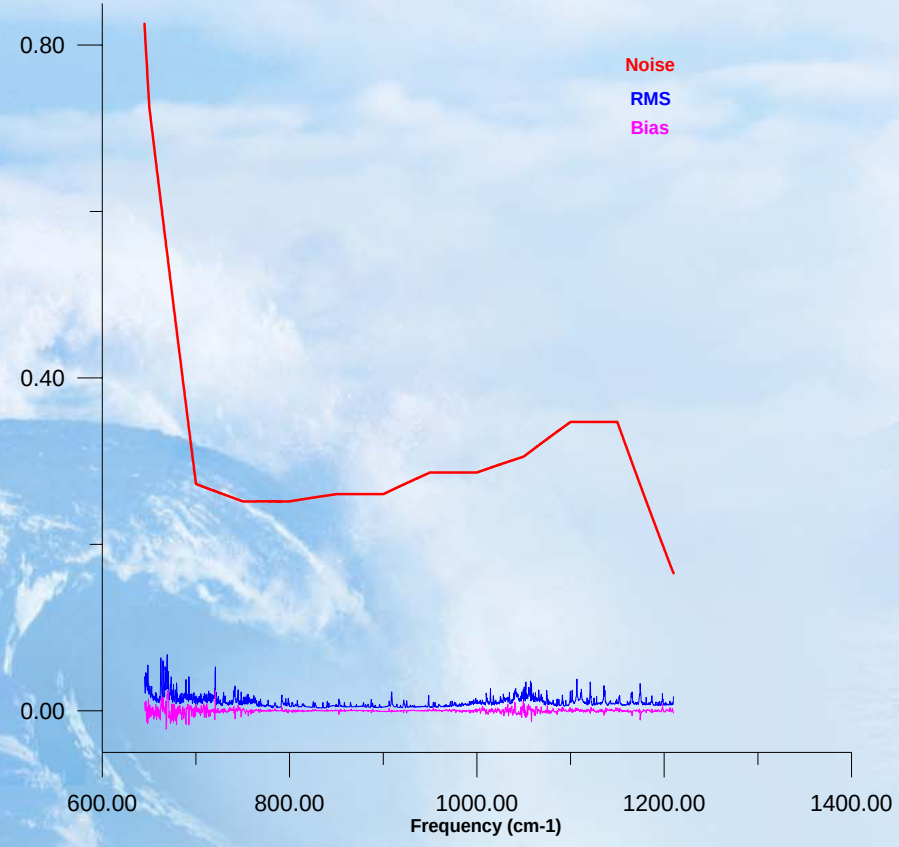


Reconstruction Score

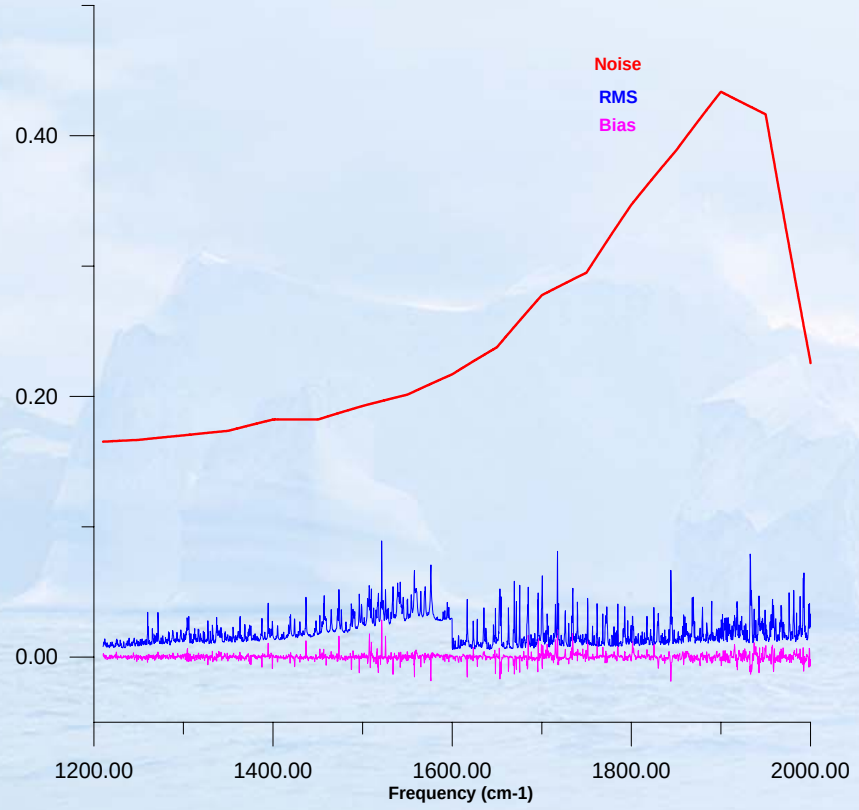


RMS and Bias

Band 1

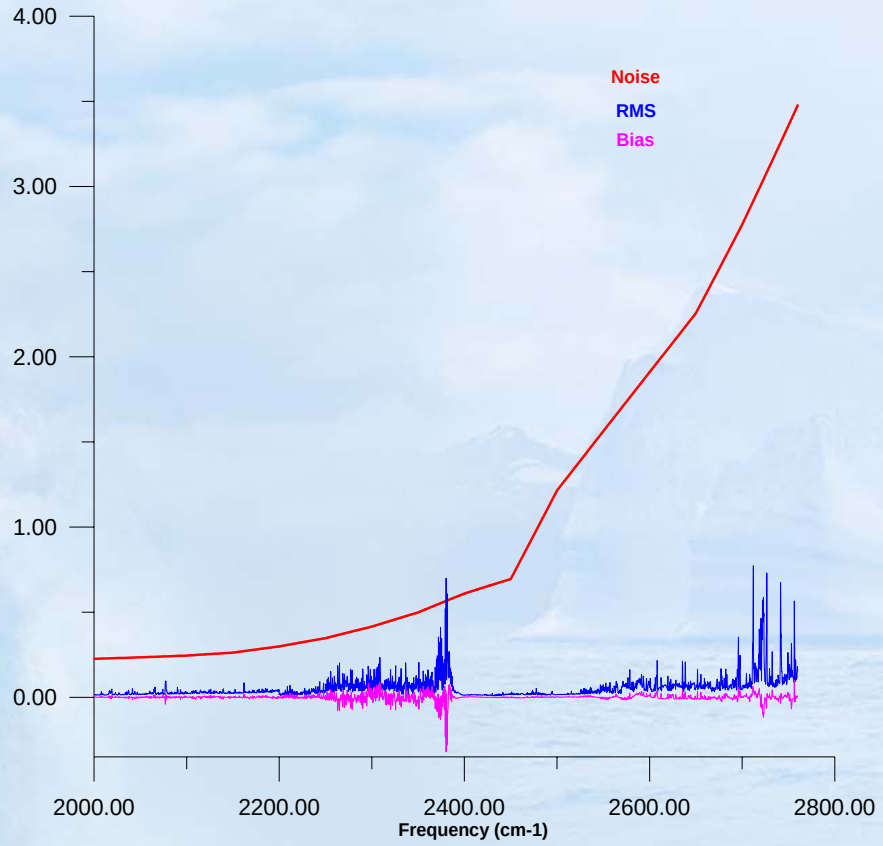


Band 2



RMS and Bias

Band 3



Apply PCA in Data Compression

- AIRS: 35GB per day
  IASI: 30GB per day

- Lossy Compression: Using PCA can compress the data by reducing the number of dimensions without much loss of information. For example, keep 200 PCs, the compressed data size is ~ 9.5%. The RMS for PCA compression (comparing with observed data) is similar to the noise level. For AIRS data, lots of PCA RMS are above 0.2K. Some values are above 1K.

Huffman Compression

- To improve the accuracy of the PCA compression, we can combine Huffman compression with PCA in our application

- Huffman compression is a lossless compression algorithm that is ideal for compressing text or small value data.

- Huffman compression uses a short bit code to replace the original data value. It assigns smaller codes to frequently used data and longer codes for data that is less frequently used.

Example

- For example consider a file consisting of the following data (31 characters)

AAAAAAAAAABBBBBBBBCCCCCDDDDDEE

If each character is represented by using eight bits, then the number of bits require to store this file will be

$$31*8 = 248 \text{ Bits}$$

- To apply Huffman algorithm, we need to build a tree that replace the symbols by shorter bit sequences

Example

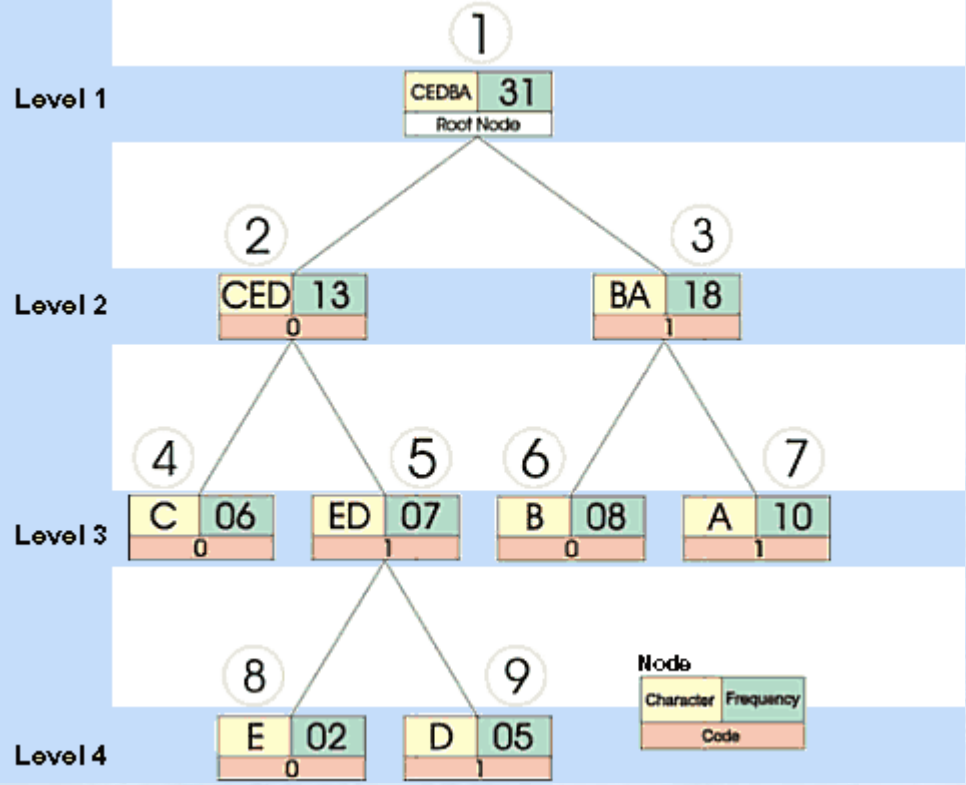
For each leg in this tree, suppose the left leg is 0 and the right leg is 1.We can get the code as:

A by the code 11 — 2bits
B by the code 10 — 2bits
C by the code 00 — 2bits
D by the code 011 — 3bits
E by the code 010 — 3bits

then the size of the file becomes

2*10 + 2*8 + 2*6 + 3*5 + 3*2 = 69 bits

This mean that 69 bits are used instead of 248, a compression ratio of 3.6 to 1 for this particular file



http://www.howtodothings.com/ViewArticle.aspx?article=392

Compress AIRS data by Combining PCA and Huffman compression

- We first compute 120 PCs and calculate the residual files. We need to store 120 PC scores and eigenvector files.
- The residual is between -10K to 10K.
- For near lossless compression, if the precision is 0.005K, then the residual data is between -9.99 and 9.99K. There are 2000 leaves in the tree. If the precision is 0.05K, the range of the residual data will become -9.9 to 9.9K. There are only 200 leaves in the tree.
- The more accurate compression means the more codes and bigger tree need to be built. And the code is also longer than that of less accurate compression data.
- The longer of the code, the bigger of the compressed file.
- The bigger of the tree, the slower to compress/decompress the file.

Result: AIRS data compression

- One granule of AIRS data: 86MB
- residual file: 86MB
- eigenvector file: 6MB
- compute 3 bands need store 120 PC (40pc for each band):

  $120*4*135*90 \approx 6MB$

  compute full band just need 40 PC

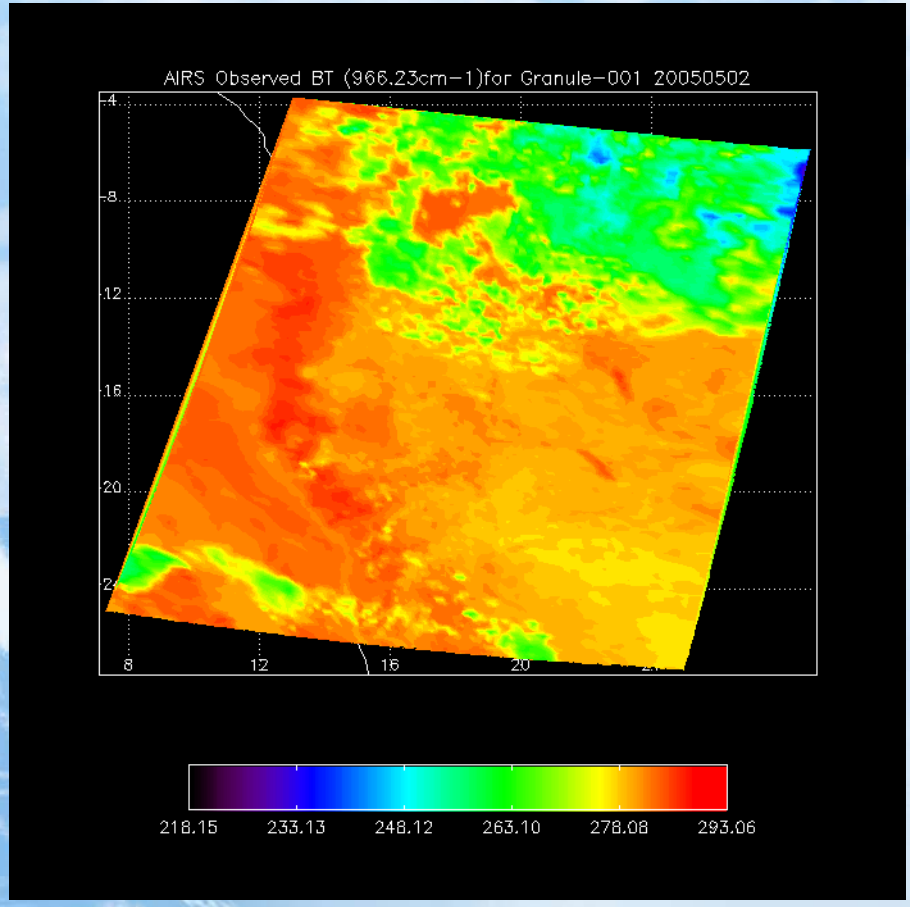
  $40*4*135*90 \approx 2MB$

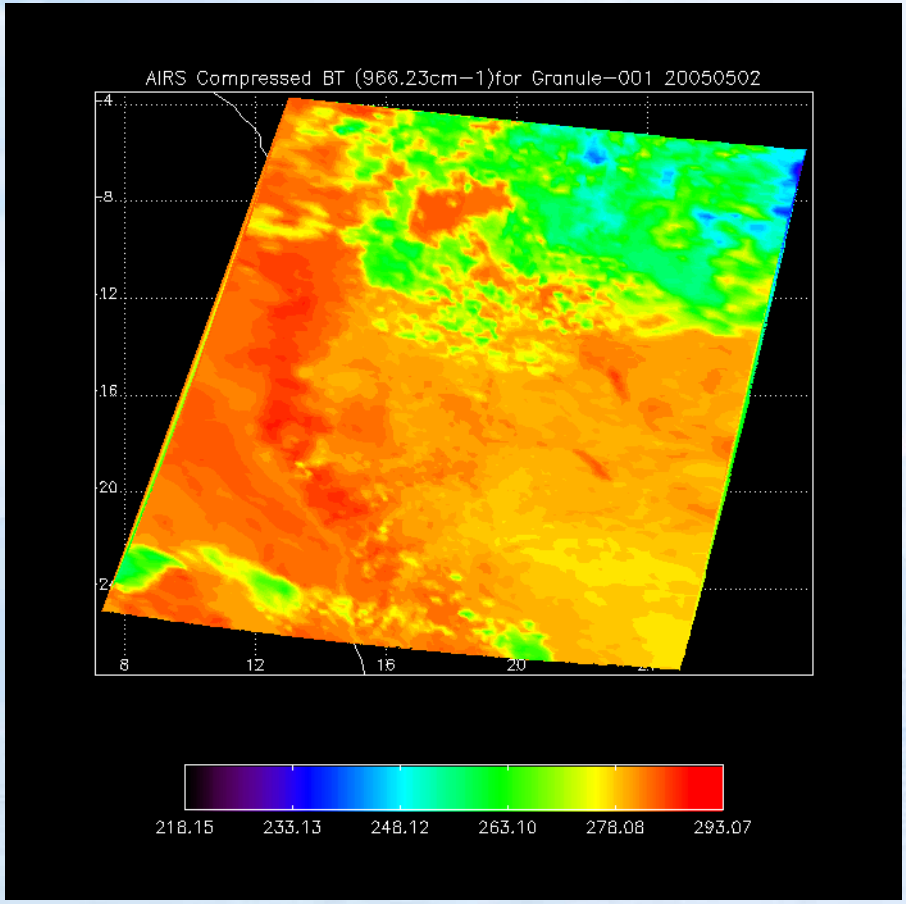
Result: AIRS data compression

- 0.005K accuracy (biggest bias is less than 0.005K):
  compressed residual file ~ 17MB
  compression ratio: 120 PC: 86/(6+6+17) ≈ 2.97
  40 PC : 86/(6+2+17) ≈ 3.44
  encoded time: ~ 17 seconds
  decoded time: ~ 25 minutes

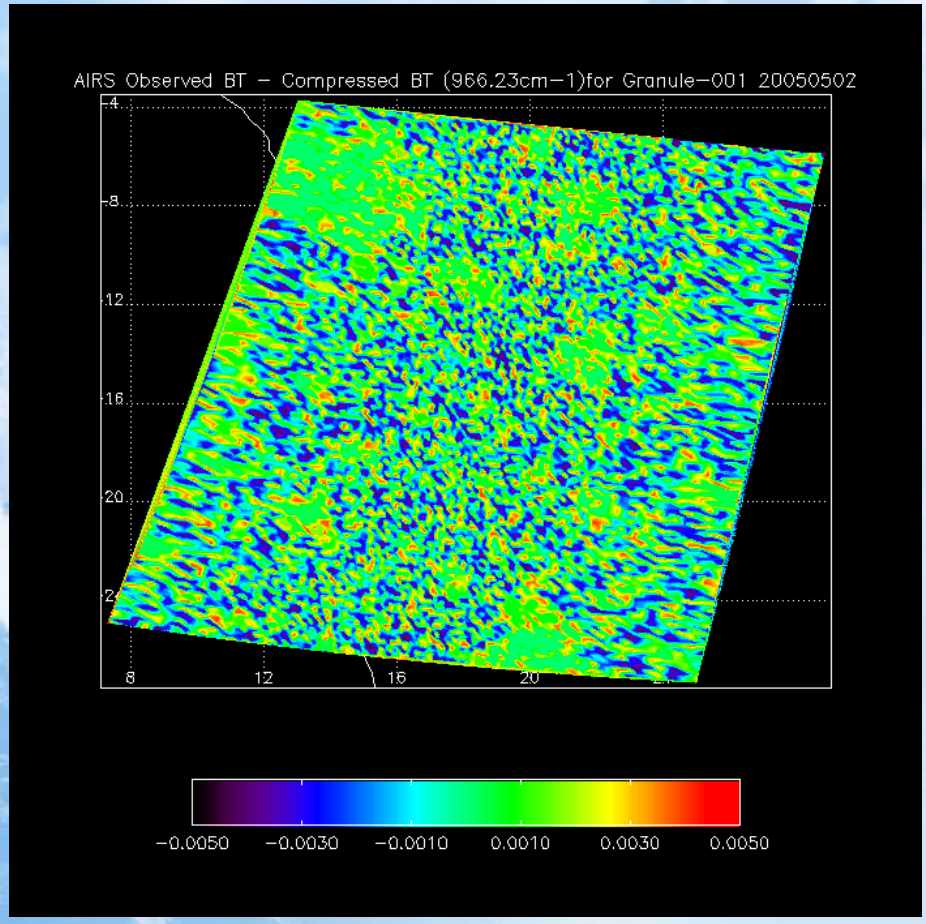
- 0.05K accuracy (biggest bias is less than 0.05K):
  compressed residual file ~ 9MB
  compression ratio: 120 PC: 86/(6+6+9) ≈ 4.1
  40 PC : 86/(6+2+9) ≈ 5.06
  encoded time: ~ 12 seconds
  decoded time: ~ 1 minutes 20 seconds

Result (0.005K accuracy): AIRS data compression

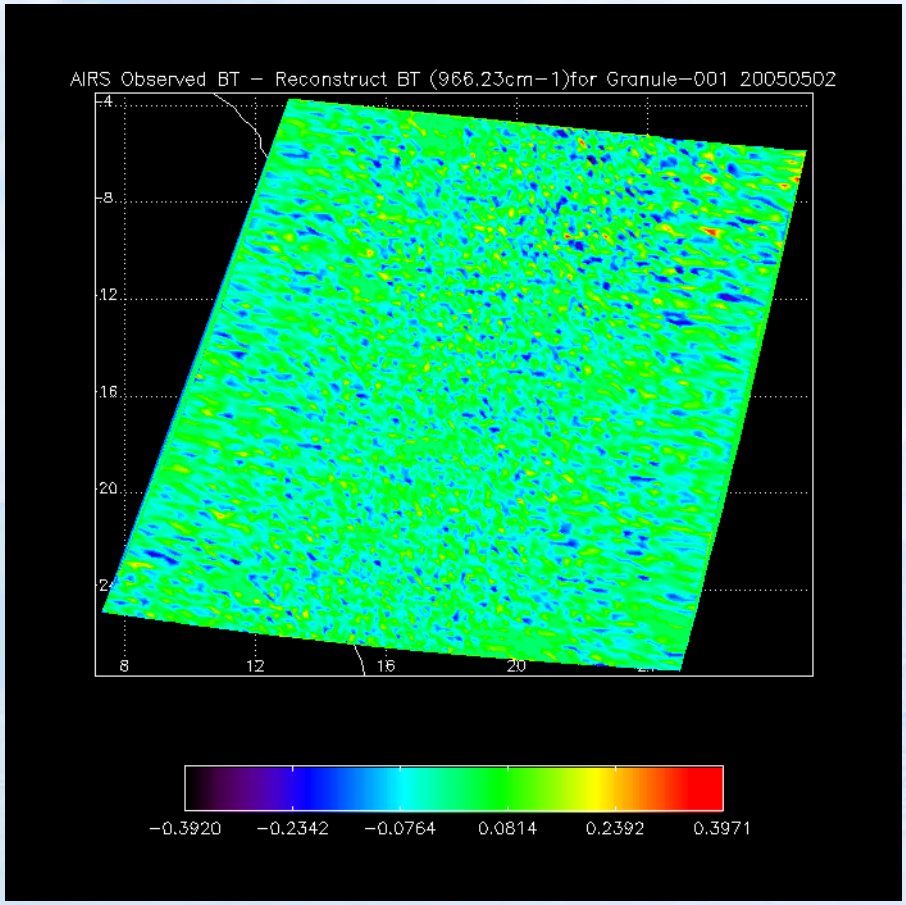




Result

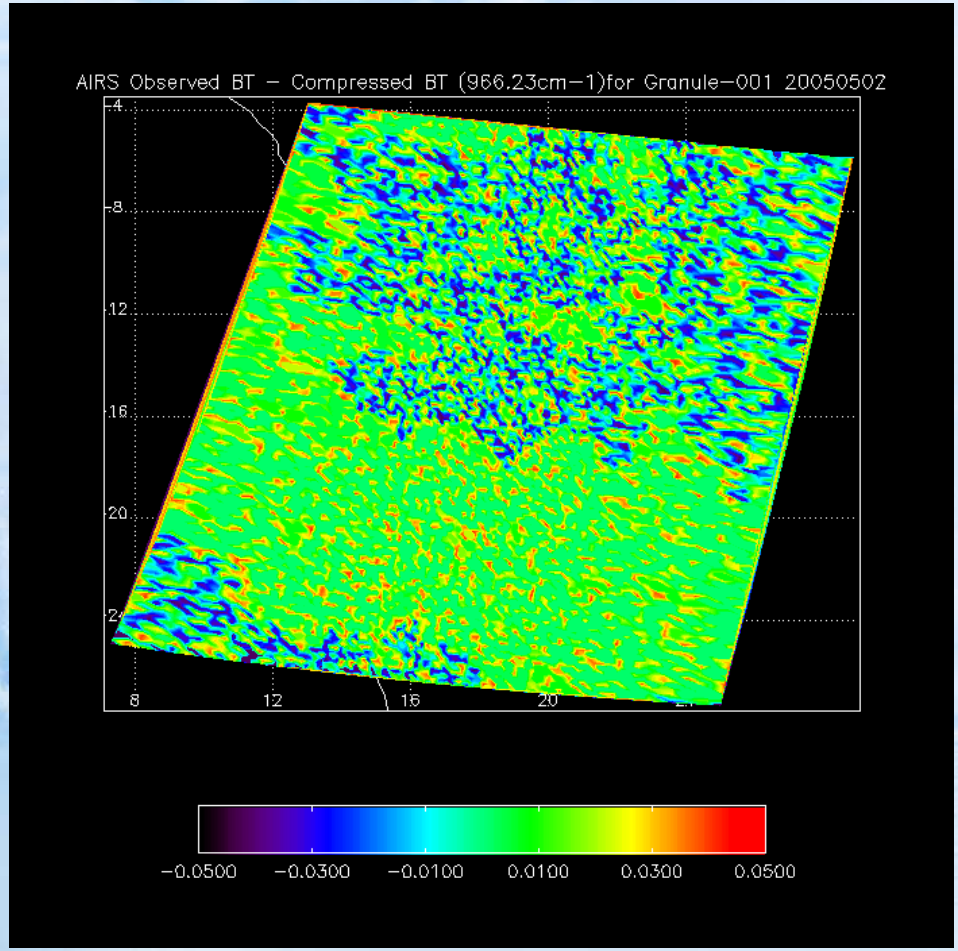


PCA+HUFFMAN



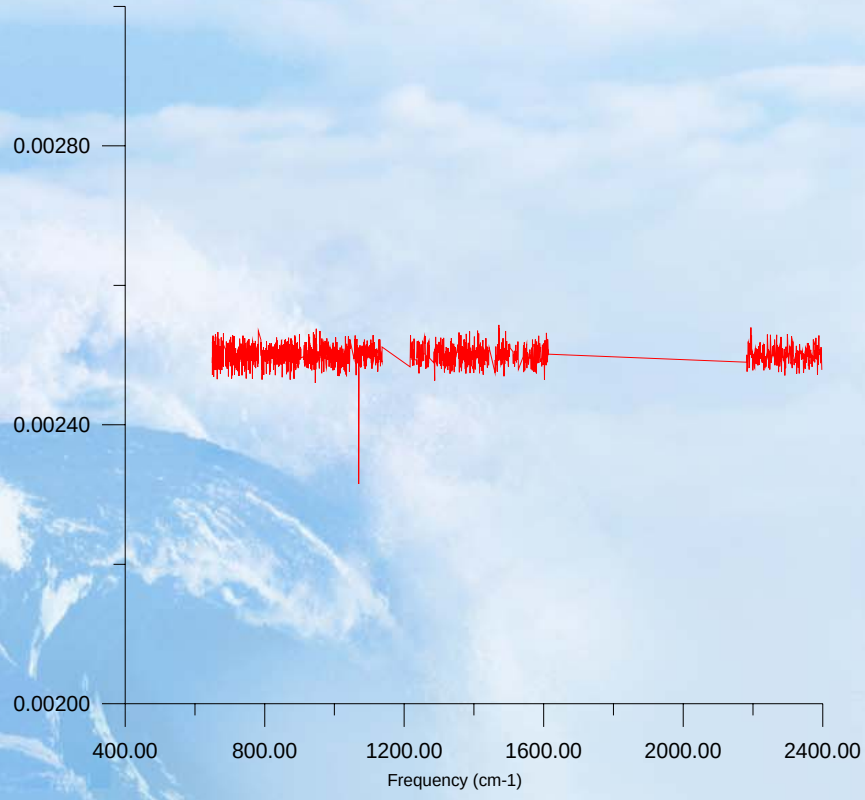
PCA Only

Result

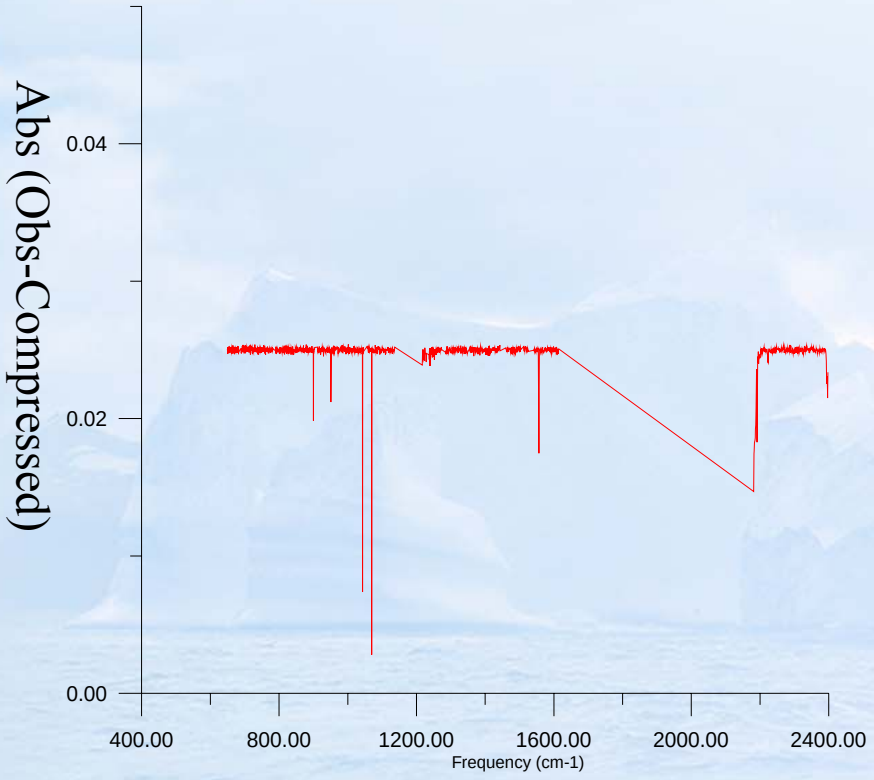

AIRS Observed BT – Compressed BT (966.23cm−1)for Granule−001 20050502
4
8
12
16
20
24
8
12
16
20
−0.0500
−0.0300
−0.0100
0.0100
0.0300
0.0500

Result

0.005K accuracy



0.05K accuracy



Summary and Future Work

- PCA has been proved to be very useful for AIRS processing.

- Will continue the work for IASI and other hyper-spectral instruments.

- FOR IASI, will work on:

  - Test on generating eigenvector by using different noise files
  - Compute the full band (all channels 8461) eigenvectors and compare the result with those of the separated bands

Summary and Future Work (2)

- Use PCA for quality control
- Data compression -Analyze accuracy vs. compression ratio
- Regression retrieval of physical parameters…
- Website development
- And more…

Summary and Future Work

THANK YOU!!!