Principal Component Analysis (PCA) of AIRS Data

Mitchell D. Goldberg[1], Lihang Zhou[2], Walter Wolf[2] and Chris Barnet[1]

NOAA/NESDIS/Office of Research and Applications, Camp Springs, MD[1]
QSS Group Inc.[2]

1. Introduction

The Atmospheric InfraRed Sounder (AIRS) (Aumann et al. 2003) is the first of a new generation of high spectral resolution infrared sounder with 2378 channels measuring outgoing radiance between 650 cm^{-1} and 2675 cm-^{1}. High spectral resolution in the infrared region allows for the derivation of atmospheric soundings of temperature, moisture, ozone and trace gases with higher accuracy and higher vertical resolution. Many of the AIRS channels are highly correlated, which is easily observable by inspecting an AIRS correlation matrix for a given global ensemble. Hence there are not 2378 independent pieces of information, instead we will show that there are far fewer (less than 100). We use principal component analysis (PCA), also call eigenvector decomposition, to extract independent orthogonal structure functions (eigenvectors). PCA is often used to approximate data vectors having many elements with a new set of data vectors having fewer elements, while retaining most of the variability and information of the original data. The new data vectors are called principal component score vectors (or coefficients of eigenvectors), and because they consist of the components of the original data vector in an orthogonal coordinate system, the elements of a given principal component score vector are independent of each other (unlike the original spectrum).

For AIRS, PCA is used for a) data compression, b) reconstructing radiances with the properties of reduced noise, c) independent instrument noise estimation, d) quality control, and e) deriving geophysical parameters.

2. Principal Component Analysis

Principal component analysis for high spectral resolution sounders is described by Huang and Antonelli (2001) and Goldberg et al. (2003). Elements of a principal component score vector are projections of the spectrum onto each of the orthogonal basis vectors, which are the eigenvectors (principal components) of the radiance covariance matrix. The total number, *n*, of eigenvectors is equal to the total number of channels. However, it can be shown that a much smaller set of *k* eigenvectors (< 100), ordered from largest to smallest eigenvalues, is sufficient to explain most of the variance in the original spectra. The covariance matrix is derived from an ensemble of AIRS normalized spectra, i.e. radiance divided by the instrument noise. The matrix of eigenvectors, **E**, is related to the covariance matrix, **S**, by:

$$\mathbf{S} = \mathbf{E}\,\boldsymbol{\lambda}\mathbf{E}^{T} \tag{1}$$

where **S** , **E** and **λ** are all dimensioned *n* x *n*, and **λ** is a diagonal matrix of eigenvalues. The principal component scores vector **p** is computed from:

$$\mathbf{p} = \mathbf{E}^{T}\,\mathbf{r} \tag{2}$$

where **r** is the vector of centered (departure from the mean) normalized radiances. The next equation is used to reconstruct the radiances from a truncated set of *k* eigenvectors **E*** and a vector of principal component scores **p***. (The symbol * indicated that the matrix or the result of a matrix operation is due to truncated set of vectors).

$$\mathbf{r}^* = \mathbf{E}^*\mathbf{p}^* \tag{3}$$

The normalized reconstructed radiance vector is **r***, **E*** has dimension *n* x *k*, and the vector **p*** has length *k*. To obtain the un-scaled radiance, one must add the ensemble mean normalized radiance used in generating the covariance matrix and multiply the sum by the noise used in constructing the normalized radiances

The number of principal components needed to reproduce the signal in the original radiances is determined by examining the magnitude of the eigenvalues and examining the spatial correlation of the principal component scores. Since we are using normalized radiances, the square root of the eigenvalues can be interpreted as signal to noise. Principal component scores (PCS) can be thought of as superchannels since each one is a linear combination of all channel, as defined the associated eigenvector. The first score contains the largest signal to noise ratio, which as shown in Table 1 is very large. When the eigenvalues fall below unity, the noise contribution is larger than the signal. Based on Table 1, this transition occurs near the 60th eigenvalue. However when we examined the PCS spatial correlations, additional principal components are needed. Ideally the spatial correlation should be near zero for the last eigenvector of the truncated set, otherwise the PCs are not capturing all of the signal. Fig. 1a-b compares global fields of the 60th and 100th PCS. Note the spatial coherent signal remaining in fig. 1a, where as little if any spatial coherence can be seen at the 100th PCS in fig 1b.

Square Root of Eigenvalues (first 72)

1	7497.60	19	14.68	37	3.38	55	1.25
2	1670.40	20	13.49	38	3.11	56	1.19
3	945.52	21	12.28	39	2.82	57	1.16
4	496.01	22	11.32	40	2.53	58	1.15
5	284.01	23	10.70	41	2.41	59	1.09
6	266.30	24	9.08	42	2.39	60	1.05
7	156.95	25	8.24	43	2.34	61	1.02
8	139.67	26	7.85	44	2.24	62	0.98
9	88.27	27	6.77	45	2.03	63	0.90
10	72.83	28	5.98	46	1.86	64	0.86
11	60.03	29	5.83	47	1.78	65	0.81
12	53.42	30	5.39	48	1.71	66	0.80
13	45.01	31	5.34	49	1.65	67	0.78
14	39.72	32	4.98	50	1.61	68	0.77
15	34.54	33	4.34	51	1.54	69	0.73
16	26.57	34	4.09	52	1.52	70	0.72
17	22.62	35	3.62	53	1.35	71	0.70
18	17.60	36	3.48	54	1.34	72	0.66

Table 1 Square root of the first 72 eigenvalues

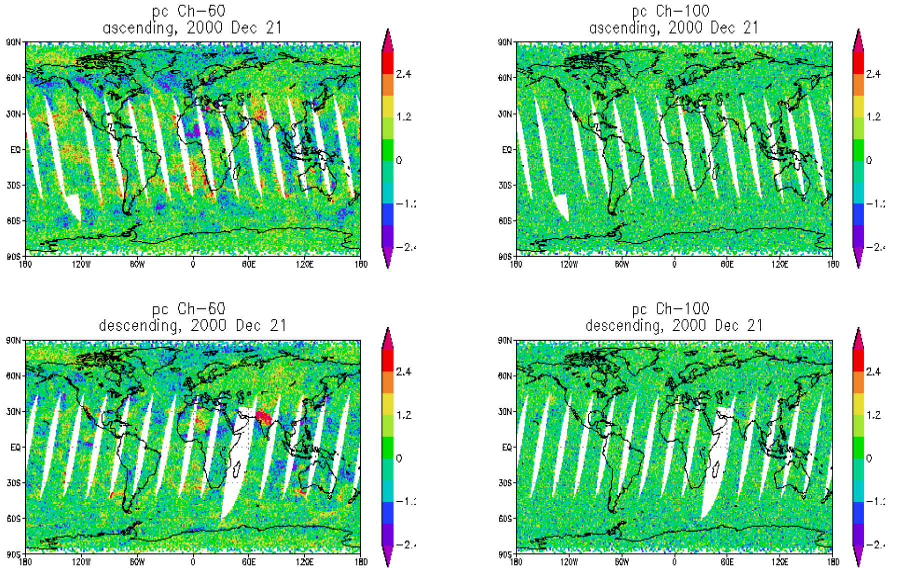

Figure 1a-b Global filed of the 60th and 100th Principal Component Scores (Note date of Dec 21, 2000, these plots are based on simulated AIRS data).

3. Applications

3.1. Noise Filtering

Because reconstructed radiances are derived from the principal components containing most of the signal as opposed to noise, the reconstructed radiances are nearly noise-free. The reconstructed radiances can be used in the retrieval process or directly assimilated. Fig. 2 shows the AIRS instrument noise at scene brightness temperature, and the root mean square (rms) difference between reconstructed brightness temperatures, from 60 principal component scores, and noise-free simulated brightness temperatures. To compute these results, we simulated brightness temperatures from a global ensemble with and without expected instrument noise. The reconstructed brightness temperatures are computed from the instrument noise-contaminated data. The original noise curve in Fig. 2 is simply the rms error of the two datasets (noise and noise-free). The rms difference between the reconstructed brightness temperatures and the noise-free simulated brightness temperatures is extremely small in comparison to the instrument noise. The reconstructed data are more similar to noise-free observations. Since the reconstructed rms error is very small, we can use the reconstructed data to estimate the noise. This is done by simply computing the rms difference between the reconstructed brightness temperatures and the original noisy data. The difference between the original noise curve in Fig. 2 and the noise estimate using PCA is shown in Fig. 3. The difference is extremely small and we use this technique as an independent approach for estimating instrumental noise. Furthermore, when we find an occasional large difference between reconstructed and the original radiances it is often due to a problem in the original radiances. So PCA is also used for quality control. Another advantage of using reconstructed radiances is that a reduced channel set can be used in a retrieval algorithm or in radiance assimilation with the benefits of using information from the entire spectrum, since each reconstructed radiance is a linear combination of all channels.

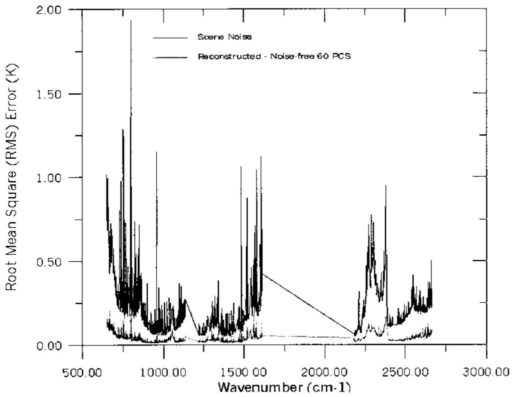

Fig. 2 Root mean square (rms) error of noise minus noise-free brightness temperatures (scene noise) and rms of reconstructed brightness temperatures from 60 principal component scores minus noise free brightness temperatures

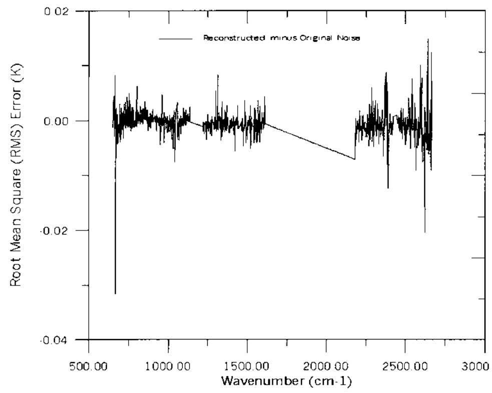

Fig. 3 Difference between the rms of reconstructed brightness temperatures from 60 principal component scores minus noise contaminated brightness temperature and the scene noise (Fig. 2).

3.2. Data compression

Instead of distributing 2378 AIRS channels, a data producer can distribute 100 or so PCS. Thereby, reducing the amount of data to be distributed or archived by a factor of 20 or more. The user can reconstruct all or a subset of the channels. We have demonstrated that the eigenvectors can be based on a historical (independent) set or based on a dependent ensemble. Our historical eigenvectors are based on a single day of observations. The dependent eigenvectors are generated from the granule itself. Fig. 4 shows the original brightness temperature field for a single channel (1001.81 cm^{-1}), along with reconstructed brightness temperature fields derived from three sets of eigenvectors: a) historical set from January 2003, b) a historical set from May 2004 (two days prior to the data shown in Fig. 4), and c) from the granule. Visually, the four images are indistinguishable. Fig. 5 shows the histograms of the difference between the original and reconstructed radiances based on three sets of eigenvectors. Shown in Fig. 5 are the minimum, maximum, mean, variance and standard derivation of the differences between reconstructed and observed brightness temperatures. Note that the corresponding values from the three plots are very similar. The reconstructed brightness temperatures for all three sets used 85 eigenvectors, even though far fewer are actually needed for granule eigenvectors because the total variability within a granule is much lower than a global ensemble.

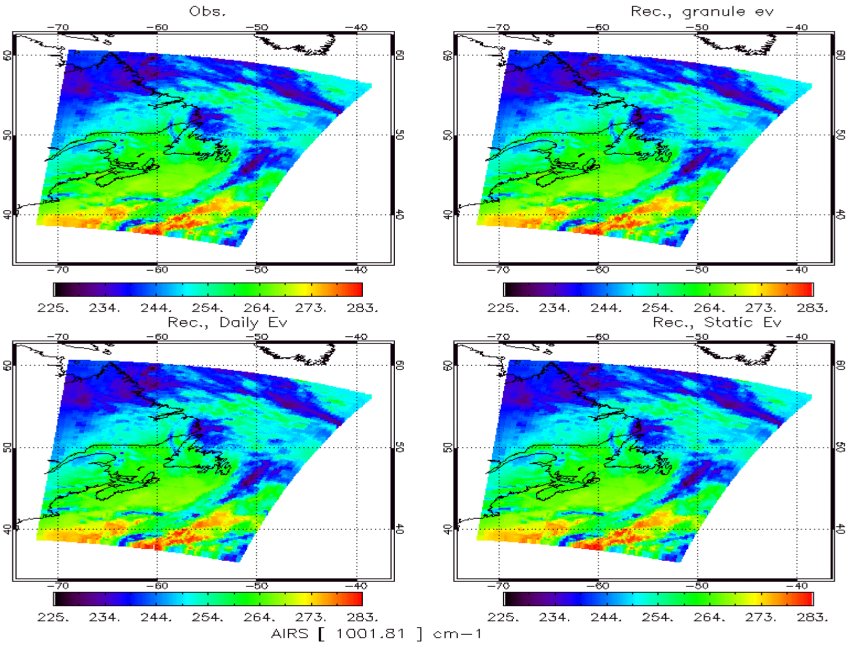

Fig. 4 Observed brightness temperatures (upper left), reconstructed from granule eigenvectors (upper right) , reconstructed from May 2004 eigenvectors (lower left) and reconstructed from Jan 2003 eigenvectors (lower right).

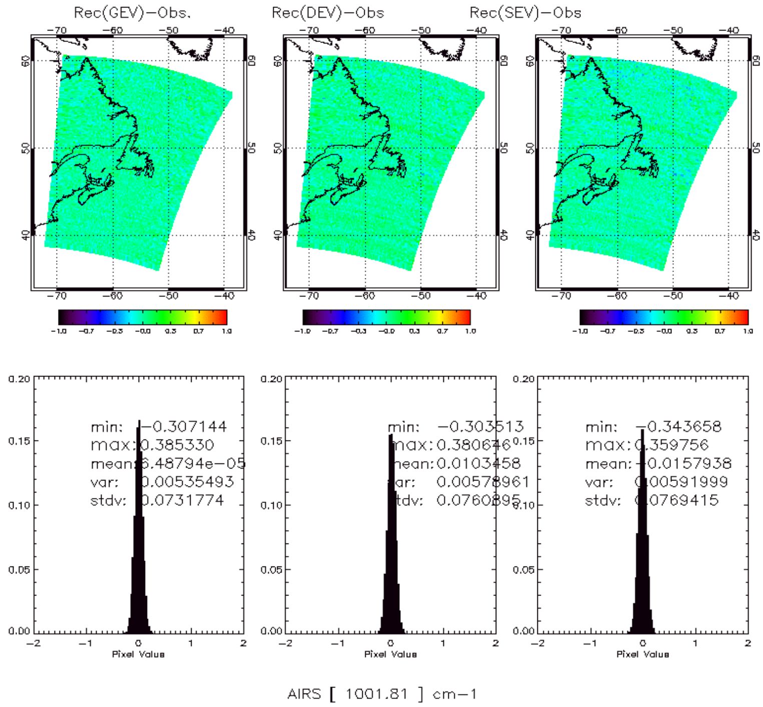

Fig. 5 Observed brightness temperatures minus reconstructed from granule eigenvectors (left) , minus reconstructed from May 2004 eigenvectors (center) and minus reconstructed from Jan 2003 eigenvectors (right).

Generally, only 20-40 eigenvectors are needed to reconstruct radiances at the noise level for eigenvectors generated from the granule itself. There are advantages and disadvantages of using historical or dependent eigenvectors. For global NWP applications, the advantages of historical eigenvectors are that only one set is needed for all cases, the eigenvectors are pre-computed, and the physical meaning of the eigenvector remain the same for all cases. For granule dependent eigenvectors, each granule would have a different set of eigenvectors, there is a much higher computational cost, and each eigenvector would have a different physical meaning. Granule dependent eigenvectors, however, will produce the best reconstructed radiances even in extreme cases, for example, volcanic eruptions. However, extreme cases are generally not assimilated in a model. Our recommendation is to use historical eigenvectors for global applications, and to use granule eigenvectors for archiving. We also recommend to archive images such as those shown in Fig. 4 and Fig. 5 as ancillary information so that users can see for themselves how well the reconstruction is performing. Our AIRS processing system at NESDIS produces datasets of PCS and a subset of 324 channels. Both are important and should be used, because statistics between reconstructed radiances and the 324 channels can be computed and monitored continuously in the same manner as currently employed by NWP centers to monitor the differences between measured versus computed radiances. It should also be noted that NESDIS produces a total reconstruction score contained in the PCS file, which is simply the square root of the sum of the differences between the observed noise-scaled radiances and the reconstructed. A reconstruction score of unity means that the overall difference between reconstructed radiances and the observed is at the instrument noise level. The reconstruction score is a very good quality assurance indicator, and we recommend only using reconstructed radiances with a score of less than 1.2.

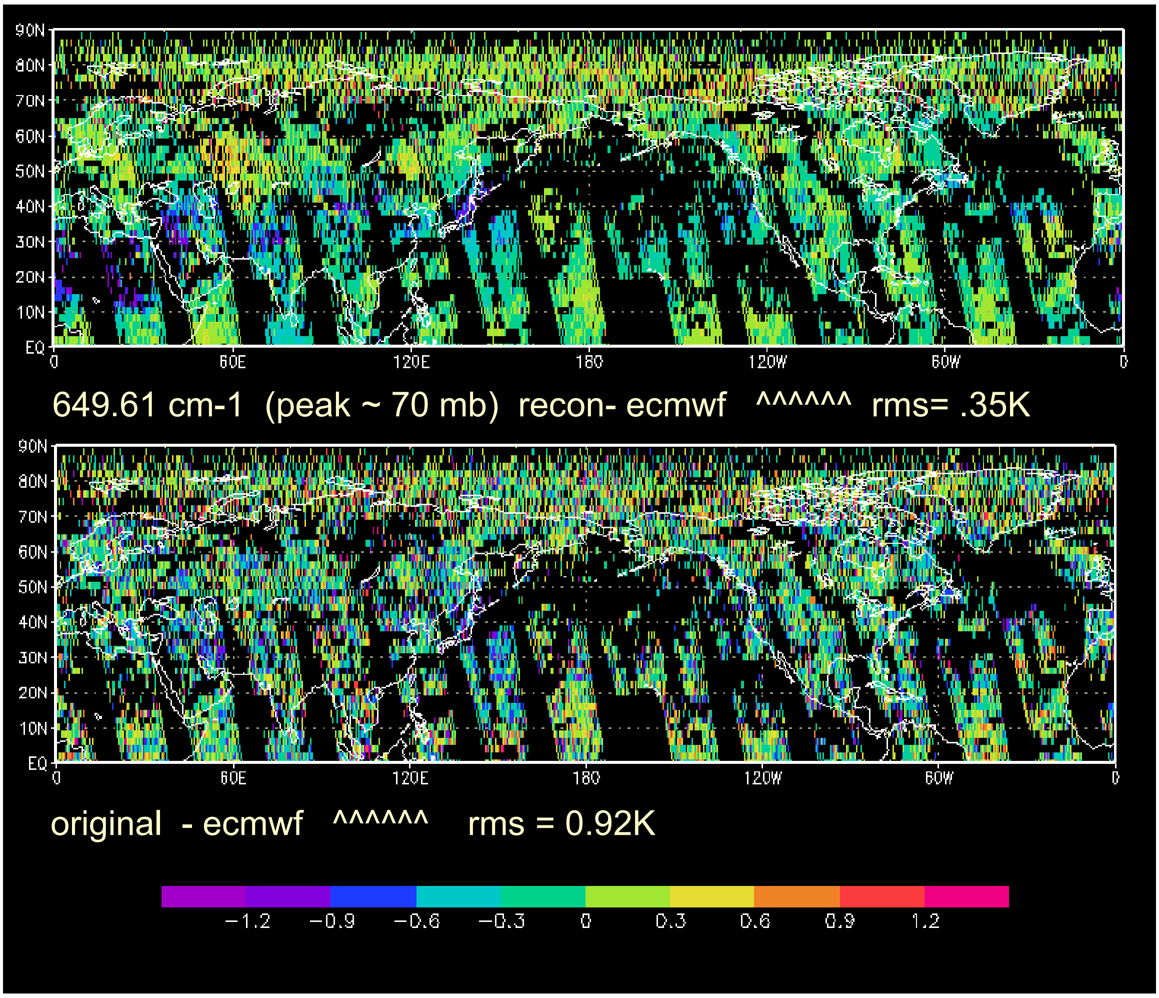
90N
80N
70N
60N
50N
40N
30N
20N
10N
EQ
0
60E
120E
180
120W
60W
0
649.61 cm-1 (peak ~ 70 mb) recon- ecmwf ^^^^^^ rms= .35K
original - ecmwf ^^^^^^ rms = 0.92K
-1.2
-0.9
-0.6
-0.3
0
0.3
0.6
0.9
1.2

Fig. 6 Upper panel is the difference between reconstructed and ECMWF calculated brightness temperature, lower panel is the difference between observed and ECMWF calculated brightness temperatures.

Fig. 6 shows a good example for considering the assimilation of reconstructed radiances (brightness temperatures). The upper panel is the difference field between reconstructed brightness temperatures and brightness temperatures computed from the ECMWF model analysis, whereas the lower panel shows the difference field between observed and computed brightness temperature. Both are for a channel at 649.61 cm^{-1}, which peaks around 70 mb. The lower panel difference field appears to be very noisy, while the upper panel difference field is more coherent. Also note that the rms difference is significantly lower. The lower panel is noisy because this channel has a instrumental noise of about 0.9 K. If we assume that the reconstructed radiances are nearly noise-free, then the true background error of the model representing this channel is about 0.35K. Even though the reconstructed radiance for a given channel is a linear combination of all channels, the largest weight is from the channel itself and from nearby similar channels. The weights from the other channels tend to be much smaller.

3.3. Geophysical Retrievals using PC Regression

Another application is the use of PC scores in least squares regression to derive geophysical retrievals. For AIRS, we use 85 principal component scores for predictors and solve for atmospheric temperature, moisture, ozone profiles and surface temperature and surface emissivity. With 2000+ channels, many of the channels are similar to each other, making the covariance matrix nearly collinear. A significant advantage for using 85 principal component scores instead of all 2000+ channels is that the inverse of the predictor matrix is more stable and less collinear. Another advantage is that 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

$$C = XP^{*T}(P^{*}P^{*T})^{-1} \tag{5}$$

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*.

Retrieval rms errors (differences between the retrieval and collocated radiosondes) based on the AIRS PC regression are shown in Fig. 7. Also shown in this figure are retrieval errors from the NESDIS ATOVS system (Reale,. 2002). The AIRS retrieval errors (dashed curve), including the systematic bias are significantly lower than ATOVS. The larger errors in the lower tropospheric temperature are probably due to uncertainties arising from collocation temporal and spatial differences. However, the difference between the

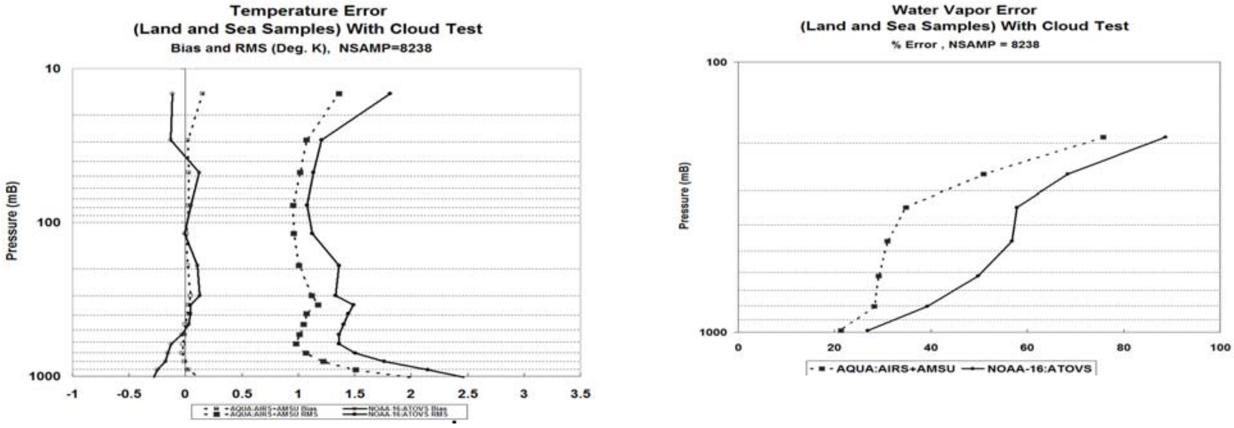

Fig. 7: Temperature (in K, top) and moisture (in %, bottom) RMS differences between the regression retrieval and the collocated radiosondes. The dashed curves are the AIRS errors, NOAA 16 errors are the solid curve.

ATOVS and AIRS retrieval remains large. Previous simulation studies have found that AIRS generally reduces the retrieval error by about 0.5K, and this appears to be holding for this radiosonde comparison. For moisture, the retrieval errors are significantly smaller than ATOVS. The large natural variability of water vapor combined with uncertainties in radiosonde-observed water vapor will prevent demonstrating the 10-15% accuracies often reported in simulated studies.

4. Recommendations for improved impact of AIRS data in NWP

Because of the demonstrated noise-filtering feature of using a truncated set of eigenvectors to reconstruct radiances, our recommendation to NWP centers is to assimilate reconstructed radiances instead of the original 324 channel radiances distributed by NESDIS. The reconstructed radiances are also available from NESDIS. This should be the first step. Second we strongly recommend the use of cloud-cleared radiances. Currently NWP centers are assimilating channel radiances that are not contaminated by clouds. The percentage of assimilated AIRS channel radiances therefore can range from 100% of the spectrally and spatially thinned data for channels peaking in the upper stratosphere, above the clouds, to 5% of the thinned data for channels peaking in the lower atmosphere. However, because the vertical resolving power of AIRS is concentrated in the lower atmosphere, the lower peaking and likely cloud contaminated AIRS channel radiances are the most important. The use of cloud-cleared radiances (Susskind et al., 2003) is therefore very important. Fig. 8 shows the global difference fields between an observed AIRS window channel brightness temperature at 1000 cm^{-1} and the clear AIRS channel simulated from ECMWF. Whereas Fig. 9 is the same

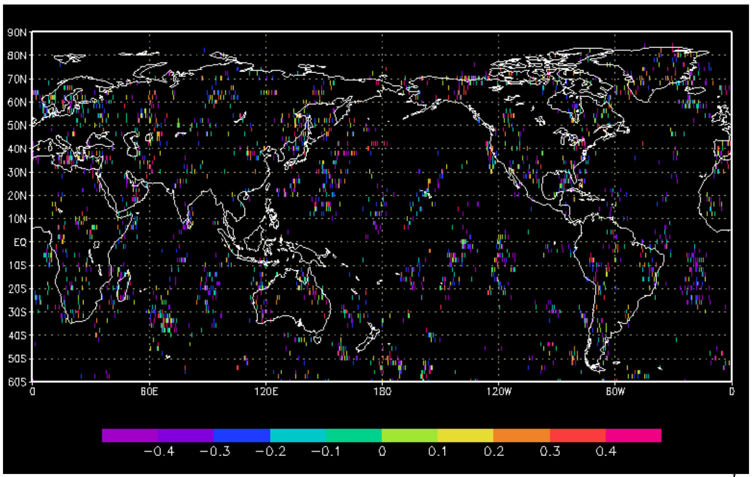

Fig 8. Observed minus clear ECMWF simulated brightness temperature for 965 cm^{-1} window channel. , all cases where the difference is +- 0.5 K. ;

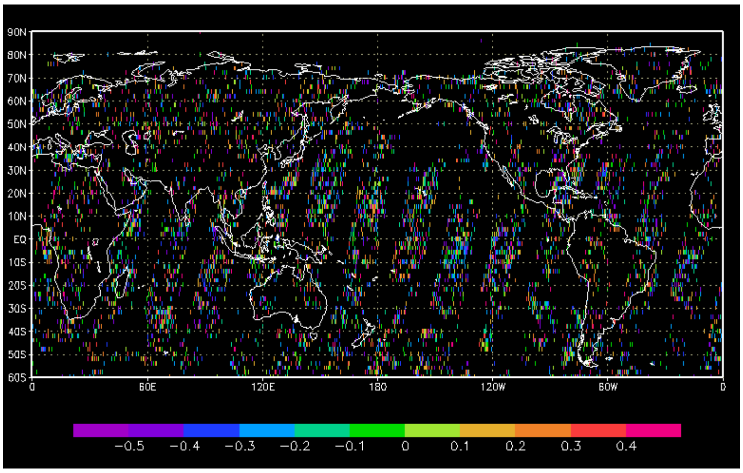

Fig 9. Cloud-Cleared minus clear ECMWF simulated brightness temperature for 965 cm^{-1} window channel. , all cases where the difference is +- 0.5 K.

as Fig. 8 with the exception that the observed AIRS is replaced with cloud-cleared radiances. Shown in both figures are difference values within +- 0.5 K. One can clearly notice the much larger coverage of data in Fig. 9.

5. Acknowledgements

This work was funded by the NESDIS Office of Research and Applications, the NASA AIRS Science Team, and the NESDIS Integrated Program Office. The views, opinions, and findings contained in this report are those of the author(s) and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. Government position, policy, or decision.

6. References

Aumann, H. M.T. Chahine, C. Gautier, M. Goldberg, E. Kalnay, L. McMillin, H. Revercomb, P.W. Rosenkranz, W.L. Smith, D. Staelin, L. Strow, and J. Susskind,"AIRS/AMSU/HSB on the AQUA mission: Design, science objectives, data products and processing systems," IEEE Transaction on Geoscience and Remote Sensing, *IEEE Trans. Geosci. Remote Sensing*, Vol. **41**, pp 253-264, Feb. 2003

Goldberg, M.D., Y. Qu, L.M. McMillin, W. Wolf, L. Zhou, and M. Divakarla, 2003: AIRS near-real-time products and algorithms in support of operational numerical weather prediction, *IEEE Trans. Geosci. Remote Sensing*, Vol. **41**, pp 379-389, Feb. 2003

Huang, H-L and P. Antonelli, Application of principal component analysis to high-resolution infrared measurement compression and retrieval. J. Appl. Meteor., 40, 365-388, 2001.

Reale, A.L. 2002. NOAA operational sounding products for advanced-TOVS. *NOAA Technical Report NESDIS 107*. U.S. Dept. of Commerce, Washington DC, 29 pp.

Susskind, J., C. Barnet and J. Blaisdell, 2003: Retrieval of atmospheric and surface parameters from AIRS/AMSU/HSB data under cloudy conditions, IEEE Transaction on Geoscience and Remote Sensing, *IEEE Trans. Geosci. Remote Sensing*, Vol. **41**, pp 390-400, Feb. 2003