# Compression of AIRS data with principal components

**Jonathan A. Smith**

*ECMWF, Shinfield Park, Reading RG2 9AX, U.K*
*email: jonathan.smith@ecmwf.int*

## 1. Introduction

The Atmospheric InfraRed Sounder (AIRS) is the first of a series of high spectral resolution infrared sounders that will become available for use in operational numerical weather prediction (NWP). Currently only 324 AIRS channels out of the total 2,378 are transmitted in near-real time to NWP centres. This restriction of channels, along with a spatial thinning to only one in nine spectra, reduces the volume of AIRS data but ensures the timely arrival of these data at NWP centres before the NWP assimilation begins. To use more channels from AIRS, the data can be compressed by expressing each spectra as coefficients to a set of principal components (PC) rather than as brightness temperature (BT) per channel. Data from all channels used in the creation of the PC set can then be reconstructed from the coefficients. The reconstructed data can however only contain information that is present in the PC set used in the reconstruction. Information not present in the training set used to create the PC set will be lost. On the other hand, not all PCs in the original set need be used in the reconstruction, especially useful if it is considered that some PCs just contain noise.

The work here has set out to investigate PC encoded AIRS data in the context of the ECMWF Integrated Forecast System (IFS). A preliminary evaluation has been carried out on AIRS PC data routinely available from NESDIS (Goldberg *et al.*, 2003). This focused on the changes seen when the original data were replaced with these NESDIS PC reconstructed spectra; first, compared to model background and then, in a short assimilation trial.

This work also considers the optimum way to create the PC sets, in particular whether creating the PC set from only clear spectra gives an advantage when reconstructing spectra from clear observations.

The results here will be used in future trials of AIRS reconstructed data and also in setting plans for the treatment of the next high spectral resolution data for NWP, from the Infrared Atmospheric Sounding Interferometer (IASI) due for launch in 2005.

The following section describes the use of PC sets to encode spectra and how the results are then evaluated. Tests with NESDIS PC sets then follow in Sections 3 and 4. The spectra used to create different PC sets are then described in Section 5 and the relative performance of these sets investigated in Section 6.

## 2. Encoding data with principal components

A spectral observation, *obs*, over $n$ channels can be encoded into a vector of coefficients, $c$, using a matrix, $\mathbf{U}$, made of $m$ PC by :

$$c = \mathbf{U}^T obs \tag{1}$$

where the superscript $T$ denotes the transpose of the matrix as the PC are arranged as columns of $\mathbf{U}$. The PC are spectral eigenvectors of length $n$. The length of $c$ equals $m$ and so $\mathbf{U}$ is an $n$ by $m$ matrix. With $m < n$, *obs* now expressed as $c$ has been compressed. The spectrum can be decoded from $c$ by :

$$recon = \mathbf{U}\, c \tag{2}$$

where *recon* is the reconstructed spectrum of *n* channels. Any spectral features not represented in **U** cannot be present in *recon*. This loss of information from each spectra can be represented as a reconstruction error, *RE*, and following Goldberg *et al.* (2003) this has been calculated here as :

$$RE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left(obs_i - recon_i\right)^2} \tag{3}$$

the root mean square of the difference between *obs* and *recon* across the *n* channels in the spectrum.

There can be several components contributing to *RE*. One that can be desirable is noise, which if removed leads to "de-noising" of channels. When channels are reconstructed they are done so using the spectral patterns over several PC. If only a few channels are then passed on for processing they benefit from the noise averaging from the other channels contributing to the PC spectral patterns.

Any true spectral features not in **U** will also contribute to *RE*. These could be due to an unusual atmospheric state not included in the data used to create **U**. It would be desirable to include any such states into **U**, by recalculation of **U**, or the addition of an eigenvector to **U**. Other spectral features may be those of data not represented in the model - such as aerosol not included in the fast radiative transfer model used in assimilation.

Note that if *obs* and *recon* are in noise normalised radiance units then $RE < 1{\cdot}0$ signifies that, on average, *obs* has been reconstructed to within the noise level over all the channels. Here $RE_R$ denotes when *RE* has been calculated from noise normalised radiance data and $RE_{BT}$ when the calculation has been from BT data.

A spectral reconstruction error, *SRE*, has also been calculated per channel. For the $i^{th}$ channel this is given by:

$$SRE_i = obs_i - recon_i \tag{4}$$

The mean and standard deviation of $SRE_i$ have been calculated over several spectra.

The $SRE_i$ for one spectrum can be gathered together into a *residual*, a vector *n* channels long. This can be compressed and sent with *c* to allow complete reconstruction of *obs*. To maintain compression when sending both *c* and *residual* relies on little information being left in *residual*, which then allows it to be highly compressed - see, for example, Schlüssel (2004).

## 3. Comparison with ECMWF model background "first guess"

The theory and details of the creation of the NESDIS PC coefficients are given in Goldberg *et al.* (2003) and in Goldberg(2004). A 1524 channel PC set was available for the comparison with the IFS background ("first guess") for every 18$^{th}$ AIRS spectra (files tagged "U1" containing the central spectrum from every alternate "golfball" - the 3 x 3 AIRS spectra within an AMSU-A field of view). The data within **U** for this NESDIS PC set were taken on one day - the 20$^{th}$ December 2002, but thinned spatially by only using 3 lines from every AIRS data granule (a granule contains 6 minutes of observations). Calculations are carried out using noise normalised spectra - using the measured instrument noise. The first two hundred eigenvectors ($m = 200$) are used to create the coefficients, *c*, which are stored in BUFR (Binary Universal Form for the Representation of meteorological data) files which are available in near-real time. Used with two hundred eigenvectors in **U** provided by NESDIS, *c* was used to calculate *recon* locally. The operational 324 channel AIRS data were replaced by 263 channels from *recon* (61 channels of the 324 were not included in the 1524 channels used in the PC set). The replacement data were then compared with the IFS first guess fields over a 12 hour period.

The results are shown in Figure 1, in the lower plot as the difference in BT (ΔBT) per channel and in the upper plot as the standard deviation of departure from the model first guess. Both are plotted against AIRS channel number for clarity so, as an approximate guide to wavenumber, the bands noted in Table 1 are marked in light blue on the figure. Mean ΔBT (blue circles) shows little change but the standard deviation (magenta circles) shows increases for *recon*. These increases are in three groups;

i. a group for higher peaking temperature sounding channels below channel number 400,

ii. another for channel numbers from 1000 to 1300 that are mainly sensitive to ozone, and

iii. a group above channel number 2100 - at the solar end of the spectrum.

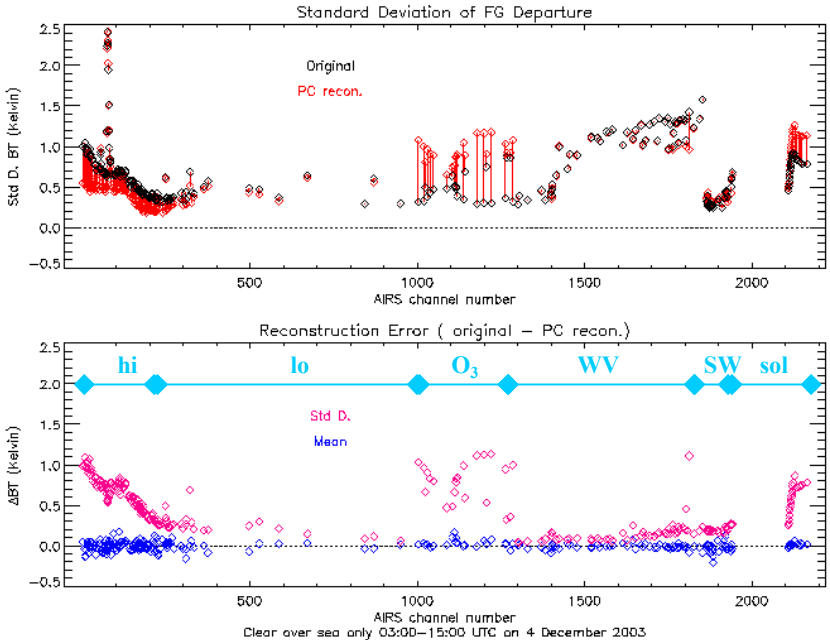


Figure 1. Spectral comparisons of the original data and data reconstructed using NESDIS transmitted coefficients. Lower plot shows differences between the original data and reconstructed data. (Markers for the spectral bands in Table 1 are in light blue along the top of this plot) Upper plot shows standard deviation from the IFS first guess.

Each group has a corresponding change in the upper plot of standard deviation in departure from the first guess. For the last two groups, (ii) and (iii), the standard deviation has increased. The change from the original observations has taken the reconstructed data further away from the model first guess. However for the first group, (i), the reconstructed data have a decreased standard deviation, they vary less from the model than the original observations do. The channels in group (i) appear to have been "de-noised" - random noise not being reconstructed as it was not included in the *m* eigenvectors within **U**.

| plotted symbol | | **hi** | **lo** | **$O_3$** | **WV** | **SW** | **sol** |
|---|---|---|---|---|---|---|---|
| main spectral properties | | high peak $CO_2$ | lower peak $CO_2$ | ozone | water vapour | $CO_2$ short wave | solar |
| break point | AIRS channel | 201 | 1003 | 1285 | 1843 | 1937 | |
| | wavenumber ($cm^{-1}$) | 707·0 | 1001·4 | 1228·2 | 1598·5 | 2248·6 | |

*Table 1. Break points between the spectral bands overlaid along channel number axis.*

## 4. Initial assimilation trial with ECMWF model

As well as *c* from the 1524 channel PC set, reconstructed radiances, *recon*, created by NESDIS from a 2047 channel PC set were available in BUFR format. As these data were presented as 322 channels in identical form to the operational 324 channel data, they were ideal for use in an initial assimilation trial. (there are two channels missing, AIRS channel numbers 1138 & 2357, but operationally these are blacklisted at ECMWF). The 2047 channel PC set was created in the same way as the 1524 channel PC set and 200 eigenvectors are again used but this time to create *recon* rather than just *c*. The assimilation trial was carried out by Tony McNally (personal communication) with no other changes to the IFS as at Cycle 26R3 beyond replacing the AIRS BT data with BT data from *recon* ( which unavoidably reduced the data volume by half as reconstructed data were only at alternate golfballs, where as operationally the central spectrum from every golfball is available).

The differences seen in the spectra are illustrated in Figure 2. The two plots are from the cloud detection scheme (which is detailed in McNally & Watts 2003 and also in McNally 2004). They are both for a band of longwave channels from one spectra; one plot using the original data and the other plot the reconstructed data from NESDIS. The band covers channels from "hi" and "lo" in Table 1. These channels are ranked according to the altitude that cloud would have to have for a significant impact on the radiance in that channel. Here channels affected at the top of the atmosphere are plotted with low rank and those affected near the surface with high rank (to the right hand side of each plot). As in Figure 1 these are plotted as BT departures from the model first guess in blue. The yellow line is a smoothed fit to the data. Limits used in the cloud detection scheme are marked by dotted, horizontal lines. The height of the tropopause in the model first guess is marked by a red triangle. Blue stars are plotted over the yellow smoothed line where the cloud detection scheme has diagnosed cloud. In each case diagnosed cloud follows the cold tail up from the lower atmosphere until the slope levels off within the ±0·5 Kelvin departure limits. The reconstructed data have a small effect in the level cloud is diagnosed at. The more striking feature is the departure from channel to channel - much less in the reconstructed spectrum. In this case the reduction in channel departure is seen above and below the tropopause, whereas the data in Figure 1 suggest that over a longer time period it is the stratospheric channels that are de-noised the most.

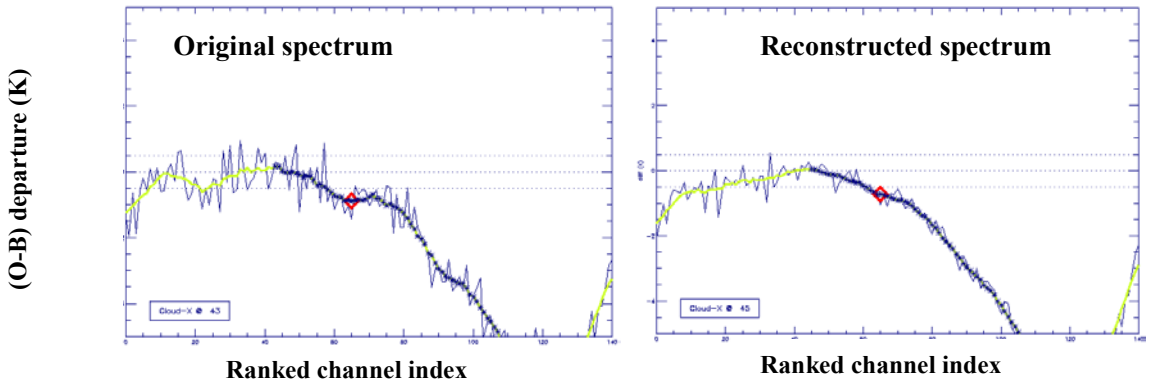


Figure 2. Plots of ECMWF AIRS cloud detection scheme for longwave band using the original data, left, and reconstructed radiances from NESDIS, right.

*recon* data were assimilated over 7 days and compared with a control run using the original, *obs*. Forecast quality and the fit to other data were the same but changes were seen in the analysis increments at high altitudes. Figure 3 shows the difference between increments for *recon* minus *obs* runs. Green shading shows little or no difference, with yellow and orange shading where the *recon* increments are greater - towards high latitudes and altitudes.

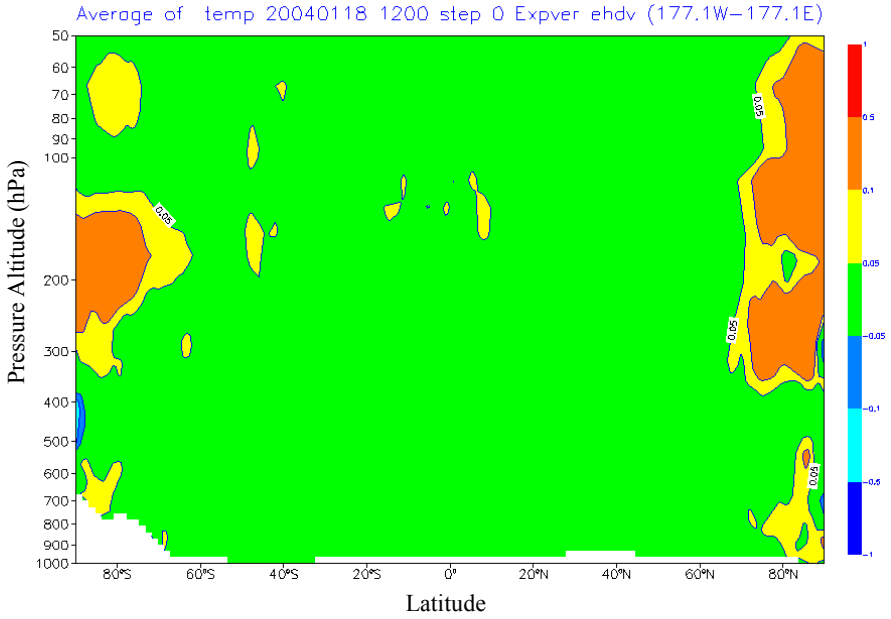


Figure 3. Zonal mean difference (NESDIS PC *recon* minus original *obs*) of RMS analysis increments from a 7 day ECMWF IFS assimilation trial.

## 5. Creation of PC sets: all-sky vs clear

To further investigate the properties of PC sets and their application in NWP it was decided to create PC sets with different properties. Given that, at the moment, only channels deemed clear of cloud are assimilated within the IFS, could better scores be achieved for clear spectra by having a PC set created from clear spectra alone ? A "clear only" PC set can also be used as the basis of a cloud detection scheme - by adding eigenvectors containing cloud information to the **U** of clear eigenvectors. For details of this method see, for example, Smith & Taylor (2004) or Schlüssel (2004).

Spectra used to create **U** were restricted to the 2107 AIRS channels marked as valid in July 2003 for AIRS Level 2 (L2) processing. Data at full spatial and spectral resolution were downloaded offline. One set of spectra used to create a PC set used all views, at all scan angles, over 1 day (15th July 2003) but thinned by a ninth by using only the central spectrum from a golfball. This gave around 324 000 spectra. For another set, spectra were selected from the following day, but only if they were over the sea and passed tests as clear. The tests were thresholds on data included in the AIRS L2 Standard Atmospheric/Surface Product. A spectrum was treated as clear and over the sea if it ; ***passed*** one of the following three tests :

1. no layers of cloud were retrieved, L2 field *"numCloud"* = 0, or
2. only one layer was retrieved but with cloud cover < 2 %, L2 fields *"CldFrcStd"* + *"CldFrcStdErr"* < 0·02, or
3. observed SST was spatially homogeneous and within 3 Kelvin of an NWP derived SST, L2 field *"clear_flag_11um"* set to 1

***and*** that the following was true :

- land fraction < 1 % L2 fields *"landFrac"* + *"landFrac_Err"* < 0·01, and

***and*** for spectra that relied upon tests on a retrieved quantity, tests 1. & 2. above, that also :

- retrieval quality flags were all OK, L2 field name *"RetQAflag"* = 0

Just under 3% of the spectra from this day passed these tests. The majority of these were within 40º latitude of the Equator (Tony Lee, *personal communication*) as L2 retrievals are not yet validated beyond 40º

latitude. To increase the representation from higher latitudes more spectra were added from the previous day (15th July). These were spectra that passed the same SST test (3. in the list above) at latitudes beyond 40º from the Equator. When added this gave 90 000 clear spectra in total.

The two sets of spectra, "All" and "Clears", were noise normalised and spectral covariance matrices calculated. The variances over each set are shown in Figure 4. The differences are as expected - taking in data from all scenes the "All" channels show greater variance over the tropospheric channels.

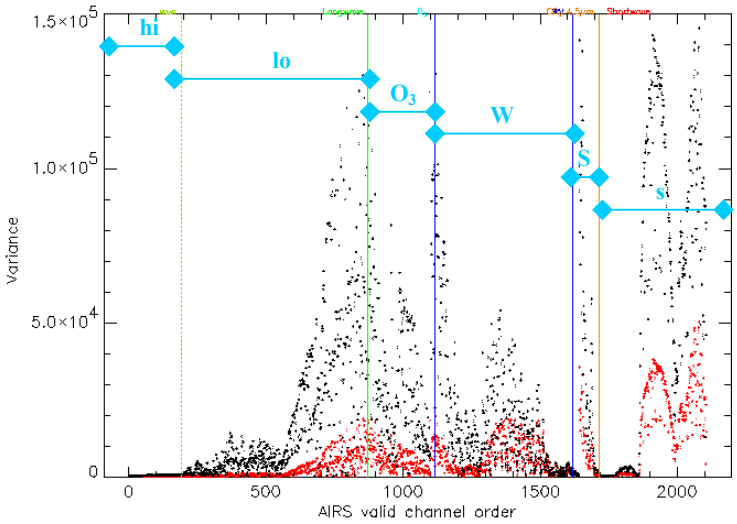


Figure 4. Variance from "All" (in black) and "Clears" (in red) spectra from July 2003 used to create the PC sets.
Light blue markers illustrate the spectral bands noted in Table 1.

Eigenvectors were computed from the covariance matrices and $m = 200$, so only the first two hundred eigenvectors were used to create **U** for each set.

## 6. Test of reconstruction error

The two PC sets "All" and "Clears" were tested with data distanced in time from the data used for U. The 14th October 2003 was chosen, which was before the shutdown of AIRS . Every ninth spectra from granules 1 to 188 on the 14th October were encoded and RER computed for these two sets and the NESDIS 1524 channel set as well. The results have been analysed geographically and spectrally.

The geographical results are summarised in Figure 5, where $RE_R$ values are mapped out and colour coded. In Fig. 5(a) "All" shows largest $RE_R$ over Antarctica where some feature in the Antarctic spring cannot now be recreated from **U** calculated in July at the time of the Antarctic winter. Over land in Australia, Africa, the Middle East and South America, $RE_R$ is slightly higher. To compare with the "Clears" PC set only spectra over the sea and flagged as clear at L2 were considered. This amounted to 6245 spectra, or 2·5 % of the total number of spectra. Of these clear spectra, 63 % had a better, lower, $RE_R$ with "Clears". These 63 % have been plotted in Fig. 5(b) to show their location, mainly at low latitudes. They are all within ±41º latitude of the Equator, whereas clear spectra spread further to 59º either side of the Equator. There is a maximum peak in the distribution in the Indian Ocean. For all clear spectra this day, the "Clears" set had a mean $RE_R$ of 1·085 and standard deviation of 0·081. For "All" it was slightly less at 1·091 and standard deviation 0·084. This small difference is statistically significant. Standard Error, *SE*, was calculated approximately using :

$$SE = \frac{SD}{\sqrt{p}} \tag{5}$$

where *SD* is the standard deviation and *p* is the number of points the standard deviation was calculated over.

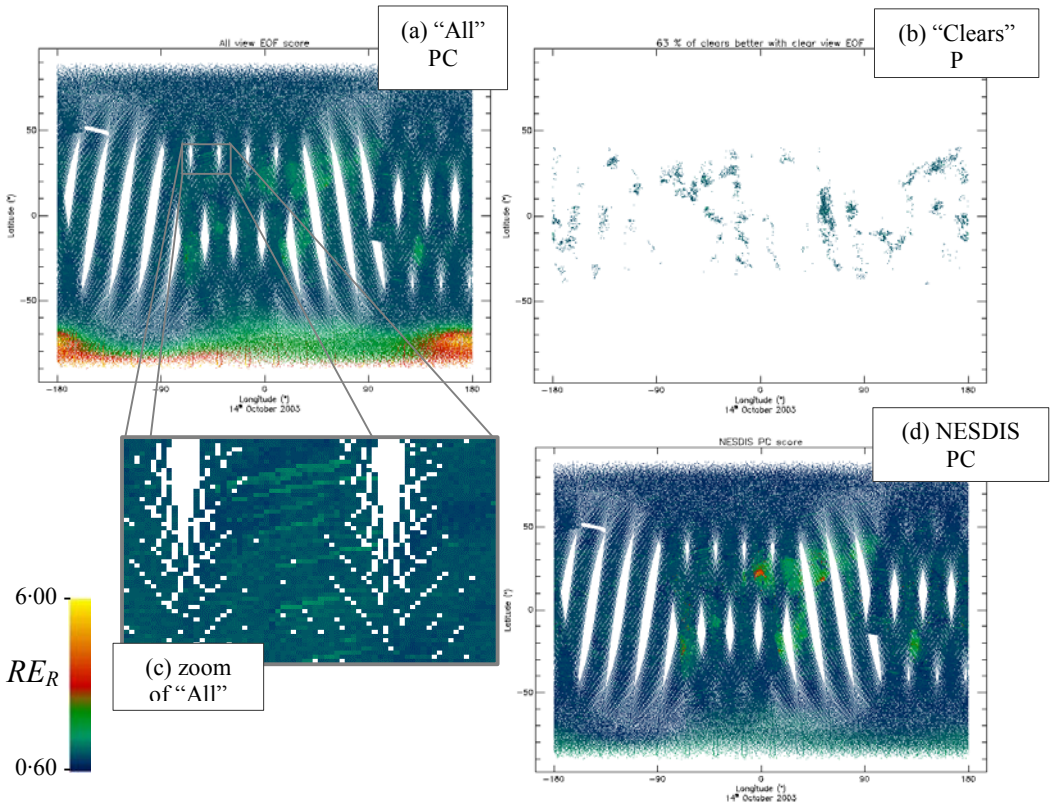


Figure 5. Three maps of $RE_R$ for Granules 1 to 188 on 14th October 2003. Alternative sets of PC, "All" in (a) and "Clears" in (b), and the NESDIS 1524 channel PC set are at (d), in the lower right, for comparison. For "Clears" map, spectra are only plotted where $RE_R$ is lower than for "All". Zoom image in (d) is a magnification on a section with "All" map to illustrate lines in $RE_R$ across the swath.

Then the standard error of the difference between means ($SE_{diff}$) is derived using

$$SE_{diff} = \sqrt{SE_j^2 + SE_k^2} \tag{6}$$

where $SE_j$ is $SE$ for the $j^{th}$ data series. For "Clears" the mean $RE_R$ is just over 4 times $SE_{diff}$ removed from that for "All". For the NESDIS 1524 PC set the mean was less at 0·981 but standard deviation was up at 0·109. The difference is much larger, at over 60 times, in terms of $SE_{diff}$. Fig. 5(c) maps the NESDIS 1524 PC set $RE_R$. The Antarctica is better represented, but the same areas of land show up again as for "All". There is a "hot spot" of increased $RE_R$ over south west Sahara. This could be due to the elevated temperatures (Mitch Goldberg, personal communication), were the difference in BT is small but when calculated in radiance, even when noise normalised, is large. Fig. 5(d) shows part of Fig. 5(a) with increased detail. Stripes can be seen across the swath as $RE_R$ crosses ove a colour shading boundary. These features could be due to "striping", an artefact of the way the AIRS data are calibrated by using space views at the end of each line. As the space view data have the same noise characteristics as other data, an increase in noise here due to the expected variability will affect the whole line.

In Figure 6 $SRE_i$ has been plotted for all three PC sets but for clear spectra only (as would be expected the "Clears" set shows large standard deviation when used to encode spectra from all scenes). Only channels common with the NESDIS 1524 channel selection have been plotted (giving 1523 channels). All three show little spectral variation in mean $SRE_i$, a slight increase in the spread of the mean from channel to channel in the longwave end of the water vapour band and in the "SW" 4·5 μm $CO_2$ temperature sounding band. There are more spectral features in the standard deviation of $SRE_i$. There are several step changes, which could be instrumental or spectral features - the steps occur both at breaks in AIRS spectral coverage and at the crossover of detector modules. The highest standard deviation is at the solar end. However there are sections were the variation in these channels have been captured within the "All" and "Clears" PC sets as the standard

deviation drops close to zero. The effect is greater for the "Clears" set. This would be expected to contain less atmospheric information but has been tested here with the same number of eigenvectors as the other sets ($m$ = 200). This may allow other signals, such as noise in a band, to be reconstructed. The NESDIS PC set shows a lower overall standard deviation in $SRE_i$, most noticeable at the longwave end in comparison with the dotted guideline arbitrarily drawn as a guide at a value of 1.

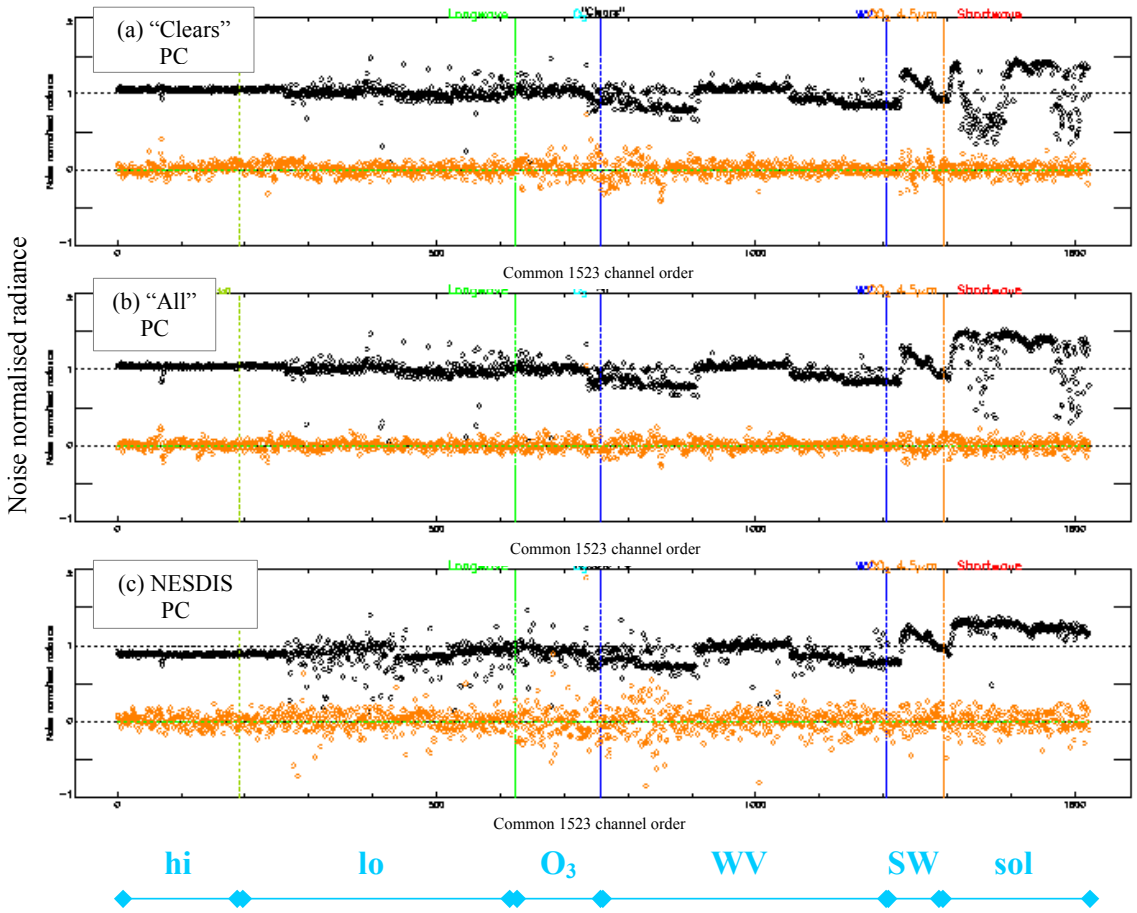


Figure 6. Spectral structure of reconstruction error, $SRE_i$, for three PC sets. Data reconstructed are clear spectra from granules 1 to 188 on 14th October 2003. Orange diamonds plot the mean and black diamonds the standard deviation of $SRE_i$ per channel. Markers for the spectral bands in Table 1 are in light blue along the bottom of the figure.

Figure 7 shows a problem seen when using data from after the AIRS solar storm shutdown. $SRE_i$ is shown from before and after, in (a) and (b) respectively. The spikes occur in both a 2107 channel set created at ECMWF in black (based on modelled data here, but the "All" and "Clears" sets gave similar results) and the NESDIS 1524 channel PC set plotted in magenta. Some detectors did not switch back on to the same state they were before - some improved, some were worse - but this meant that the PC sets were no longer valid. The channels creating the spikes in Fig. 7(b) are AIRS channel numbers : 756 (899·3 cm$^{-1}$), 765 (902·4 cm$^{-1}$), 957 (982·0 cm$^{-1}$) and 1802 (1569·3 cm$^{-1}$). None of these channels are included in the 324 channels used operationally. This illustrates a general problem of time consistency in reconstruction performance over the 2107 channels. Some of the earliest AIRS data, from October 2002, is used in the $SRE_i$ for Figure 8. Again a few badly reconstructed channels (the two main spikes are for AIRS channel 957, 982·0 cm$^{-1}$, and AIRS channel 1791, 1561·6 cm$^{-1}$) make useful comparisons using a spectrally averaged quantity such as $RE$ difficult.

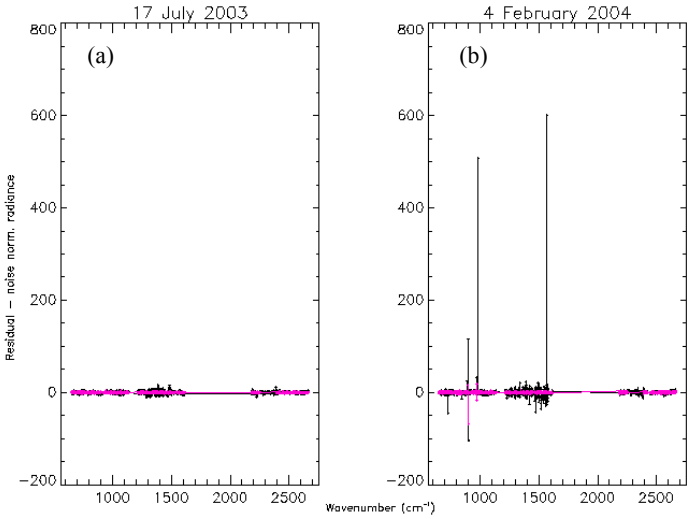


Figure 7. $SRE_i$ before, (a), and after, (b), the shutdown of AIRS. In each $SRE_i$ is plotted as small diamond symbols joined by a line in ; black for a 2107 channel PC set, and in magenta for the NESDIS 15245 channel PC set.

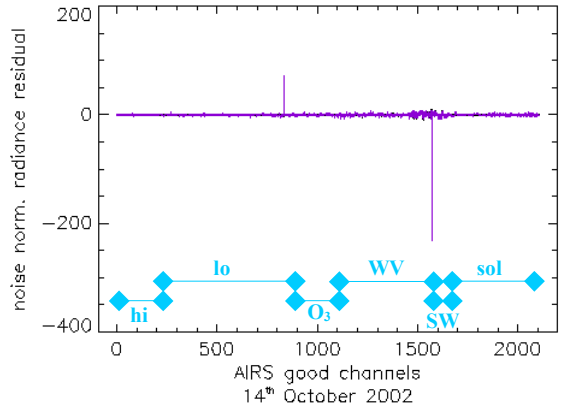


Figure 8. $SRE_i$ shown for AIRS data 9 months before the 2107 channel PC sets were created (July 2003). The purple line plots $SRE_i$ for the "Clear" PC set. Markers for the spectral bands in Table 1 are in light blue along the bottom of the plot.

Finally an example showing better, lower, values in clear air with a clear-only based PC set. Granule 180 on 15th October 2003 was over the Western Australian coast. Scene BT is shown in Figure 9(a) as a greyscale, with the land blanked off in purple as a location guide. Several clear areas, with broken and uniform cloud around, are present with a band of higher, unbroken, cloud extending from the bottom right corner. Fig. 9(b) shows $RE_{BT}$ for the "All" PC set with the higher cloud being least well reconstructed and showing up in red. With a PC set based only on clears the reconstruction of this high cloud is worse, as might be expected, and $RE_{BT}$ in Fig. 9(c) for this cloud shows as yellow. However, for the clear areas (either side of this band and, particularly, offshore in the top of the image) $RE_{BT}$ is often lower than for "All". When compared with the NESDIS 1524 channel PC set, 88 % of the clears had a lower $RE_{BT}$ (the percentage was even higher for "All). The PC set used for Fig. 9(c) was created accidentally, by mixing noise normalised and un-normalised data into a covariance matrix using the same 16th July spectra in the "Clears" PC set. As such this PC set used for Fig. 9(c) cannot be relied upon to contain all clear information, but for this granule it could better represent a large majority of the clear spectra.

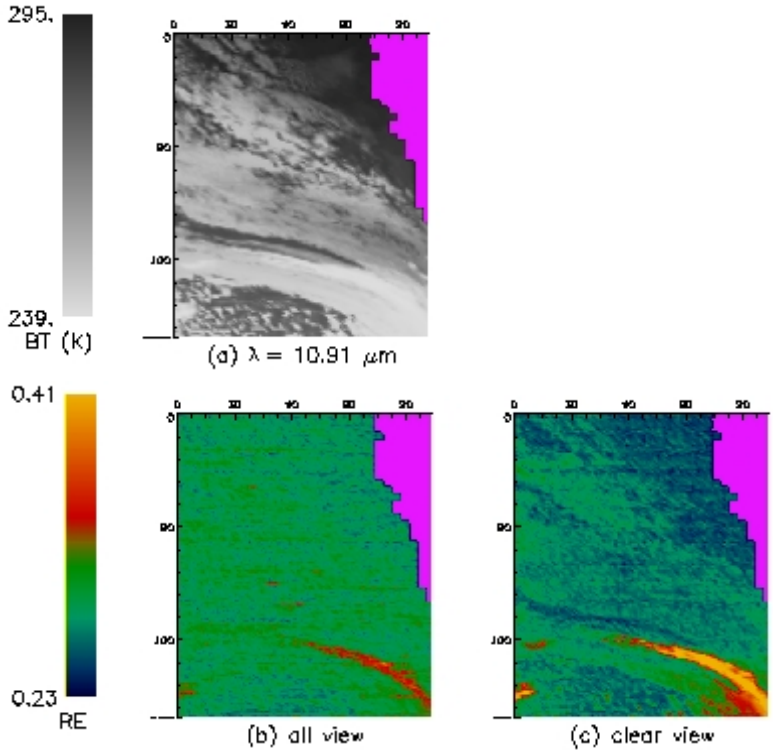


Figure 9. (a) shows scene BT in grey for Granule 180 on 15th October 2003 which was offshore from W. Australia. $RE_{BT}$ values from two different PC sets are colour shaded; in (b) for PC created from all view data, and in (c) using only clear views. (land is masked out with purple). Note the lower $RE_{BT}$ in (c) over several, cloud-free, areas.

## 7. Conclusion

In the NESDIS PC AIRS data de-noising - the removal of noise in the used channels - was seen for the longwave stratospheric $CO_2$ temperature sounding channels. Such channels used operationally in the IFS showed less variance to the background "first guess". Other channels, at shorter wavelengths, were the same as the original data or, at times, noisier. If a *residual* were transmitted along with *c* the channels exhibiting increased noise could be reconstructed exactly. The use of *RE* alone is not sufficient to asses the quality of the data for NWP, as de-noising will give an increase in *RE* similar to that for a poor reconstruction of desired spectral features. When assimilated into the IFS in a short trial, *recon* data from NESDIS PC showed an increase in stratospheric increments with no decrease in forecast quality or change in agreement with other data. This suggests that the de-noised channels were having a larger impact on the model, though more cases are required to confirm this.

The tests with the PC set made from "clear only" data showed there were some improvements in representing clear spectra. Though the differences were statistically valid, they were small and it would not be worthwhile using that particular "Clears" PC set for clear spectra in preference to the "All" PC set. The data have shown though that too many channels can lead to problems if some channels change. Increased noise at detector boundaries suggest that it may be possible to remove this instrument effect with PC. The striping that is apparent in *RE* maps shows the striping problem may be significant in NWP - with previous and current operational data thinned spatially, "stripes" were not readily seen though the data would still have been affected. A more stable set of channels for **U** would be necessary for the properties of clear only PC sets to be tested for time and location consistency. Further work should could also look at; varying *m* to allow a good reconstruction but keep any de-noising, to extend the clear spectra to higher latitudes and to add cloudy eigenvectors for cloud detection.

## Acknowledgements

We thank Mitch Goldberg and Walter Wolf at NESDIS for their help and access to their PC data and processed AIRS products. The results in Section 4 and Figures 2 & 3 were provided by Tony McNally. The full resolution AIRS data used in this study were archived and distributed by the NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC) Distributed Active Archive Center (DAAC).

## References

Goldberg, M.D., 2004: Principal component analysis of AIRS radiances. *in these proceedings*

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*, **41**, 379-389.

Lee, A. C. L., 2003: Strategy for dissemination and use of advanced sounder radiances. Met Office (RSI) Branch Working Paper, No. 205 [available from National Meteorological Library, FitzRoy Road, Exeter, EX1 3PB, United Kingdom]

McNally, A.P., 2004: The assimilation of AIRS radiances at ECMWF. *in these proceedings*

McNally, A.P. and P.D. Watts, 2003: A cloud detection algorithm for high-spectral-resolution infrared sounders. *Q. J. R. Meteorol. Soc.*, **129**, 3411-3423

Schlüssel, P., 2004: Compression of IASI Data and Representation of FRTM in EOF space. *in these proceedings*

Smith, J.A. and J.P. Taylor, 2004: Cloud detection using the EOF components of high spectral resolution infrared sounder data. *J. Appl. Meteorol.*, **43**, 196-210.