

# Technical note: Analytical estimation of the optimal parameters for the EOF retrievals of the IASI Level 2 Product Processing Facility and its application using AIRS and ECMWF data

X. Calbet and P. Schlüssel

EUMETSAT, Am Kavalleriesand 31, 64 295 Darmstadt, Germany

Received: 21 June 2005 – Published in Atmos. Chem. Phys. Discuss.: 10 October 2005

Revised: 22 December 2005 – Accepted: 13 January 2006 – Published: 16 March 2006

**Abstract.** The Empirical Orthogonal Function (EOF) retrieval technique consists of calculating the eigenvectors of the spectra to later perform a linear regression between these and the atmospheric states, this first step is known as training. At a later stage, known as performing the retrievals, atmospheric profiles are derived from measured atmospheric radiances.

When EOF retrievals are trained with a statistically different data set than the one used for retrievals two basic problems arise: significant biases appear in the retrievals and differences between the covariances of the training data set and the measured data set degrade them.

The retrieved profiles will show a bias with respect to the real profiles which comes from the combined effect of the mean difference between the training and the real spectra projected into the atmospheric state space and the mean difference between the training and the atmospheric profiles.

The standard deviations of the difference between the retrieved profiles and the real ones show different behavior depending on whether the covariance of the training spectra is bigger, equal or smaller than the covariance of the measured spectra with which the retrievals are performed.

The procedure to correct for these effects is shown both analytically and with a measured example. It consists of first calculating the average and standard deviation of the difference between real observed spectra and the calculated spectra obtained from the real atmospheric state and the radiative transfer model used to create the training spectra. In a later step, measured spectra must be bias corrected with this average before performing the retrievals and the linear regression of the training must be performed adding noise to the spectra corresponding to the aforementioned calculated stan-

dard deviation. This procedure is optimal in the sense that to improve the retrievals one must resort to using a different training data set or a different algorithm.

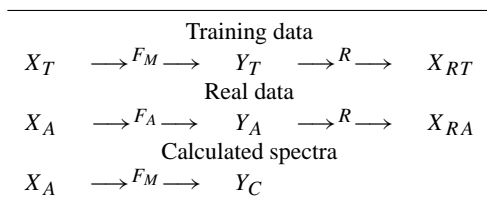
## 1 Introduction

Temperature and water vapour are basic meteorological parameters of high importance for weather forecasting as well as atmospheric chemistry studies. Observations from high-spectral-resolution infrared sounding instruments on board of satellites can provide unprecedented accuracy and vertical resolution of temperature and water vapour profiles. It is, however, not trivial to retrieve the full information content from radiation measurements. Accordingly, improved retrieval algorithms are desirable to achieve optimal performance of existing and future instrumentation, such as ground-based Fourier Transform InfraRed (FTIR) spectrometers (Schneider et al., 2005) or the satellite-based Advanced Microwave Sounding Unit (AMSU) (Houshangpour et al., 2005).

A series of European satellites, known as Metop, will be launched in the frame of the EUMETSAT Polar System (EPS) in low Earth orbits. The first launch of the Metop satellites is planned for 2006 and will carry the Infrared Atmospheric Sounding Interferometer (IASI). IASI is a high-spectral-resolution infrared sounding instrument developed by the Centre National d'Etudes Spatiales (CNES) and based on a Fourier transform spectrometer. IASI spectra are represented by 8461 spectral samples, between 3.62 and 15.5  $\mu\text{m}$ , with a spectral resolution of 0.5  $\text{cm}^{-1}$  after apodisation. Its spatial resolution is 25 km at nadir with an IFOV (Instantaneous Field of View) size of 12 km at a satellite altitude of 819 km. As part of EPS, EUMETSAT is developing the

Correspondence to: X. Calbet  
(xavier.calbet@eumetsat.int)

**Table 1.** Variable synopsis.  $X_T$ : modified “Sampled database of 60-level atmospheric profiles from the ECMWF analyses” (Chevallier, 2002).  $F_M$ : radiative transfer model RTTOV8 (Saunders, 2004).  $Y_T$ : spectra derived from the modified ECMWF sample (Chevallier, 2002) and RTTOV8. R: EOF retrieval.  $X_A$ : ECMWF analyses (ERA40).  $F_A$ : radiative transfer of the real atmosphere and instrument.  $Y_A$ : AIRS measured spectra.  $Y_C$ : calculated spectra from ECMWF analyses and RTTOV8.



operational IASI Level 2 Product Processing Facility (IASI L2 PPF), which will generate atmospheric state retrievals from the IASI radiance spectra (Schlüssel et al., 2005).

One of the retrieval techniques available in the IASI L2 PPF is based on Empirical Orthogonal Functions (EOF), which is a valuable and very computer efficient method. It consists in performing a linear regression of the principal components or EOF of the measured brightness temperature spectra and the atmospheric state parameters. In this paper, the particular EOF retrieval method developed for the IASI L2 PPF will be reviewed analytically and tested with real data available from the AIRS instrument.

AIRS is a high-spectral-resolution infrared sounder launched in May 2001 on board the NASA Aqua satellite (Aumann et al., 2003). It has a spectral coverage from 3.7 to 15.4  $\mu\text{m}$  with a spectral resolution of 1200 ( $\lambda/\Delta\lambda$ ) and a total of 2378 channels. Its spatial resolution is about 28 km at nadir with an IFOV size of 14 km.

The EOF retrieval method has been studied before with synthetically generated data (e.g. Huang and Antonelli, 2001), but further problems arise when used with real data as is acknowledged by Zhou et al. (2002). Namely, the existence of a significant bias between the measured and modeled derived radiance and the dominant influence of the radiative transfer model errors on the observational error analysis.

To make this paper more readable, the real world example data is presented throughout the analytical demonstrations, but conceptually this paper could be divided in two separate parts. The first one (Sects. 2 to 5) deals with the analytical derivation of the best parameters to be used in EOF retrievals. The demonstration is general enough to account for different types of EOF retrievals using the same algorithm as shown in this paper. It can be applied whether radiances or brightness temperature measurements are used. The method can also be applied whether it is calibrated and validated using numerical model analyses or using radiosonde data. The first condition to apply the analytical results is that it is only cal-

ibrated and validated with one set of atmospheric profiles, that is, either radiosondes or numerical model analyses, but not both at once. The second condition is that the “total” noise of the measurements has gaussian statistics. By “total” noise it is meant the observed minus “calculated” measurement standard deviation as shown in Fig. 12. This “total” noise includes the instrumental noise, the forward radiative transfer model errors and the representativeness of the data used as the real atmospheric profiles. Once these two conditions are met, the analytical results show which bias corrections and noise figures are the optimal ones in the EOF retrievals.

The second part of the paper (Sect. 6 and throughout Sects. 2 to 5) verifies the analytical results with a real world example, the EOF retrievals of the IASI L2 PPF using real AIRS spectra. In this particular example, AIRS brightness temperatures are the measured quantities and the atmospheric profiles are calibrated and validated against ECMWF analyses. It has been verified (not shown in this paper) that the noise of the observed minus calculated brightness temperatures do show gaussian statistics, and hence the analytical optimal bias and standard deviation corrections can be applied.

## 2 EOF retrievals

The IASI L2 PPF EOF retrieval consists of two distinct parts. The first one of them is the “training” process in which the retrieval parameters are determined. The second one consists in performing retrievals with the available data using these parameters, validating the theoretical approach. These parts will be explained briefly in the next two subsections. Table 1 summarizes all the main variables used in this paper.

### 2.1 Training EOF retrievals

The EOF retrievals can be trained with synthetically generated data derived from a representative sample of atmospheric states. In the IASI L2 PPF case, the profiles used for training are a modification of the “Sampled database of 60-level atmospheric profiles from the ECMWF analyses” (Chevallier, 2002), and will be denoted by  $X_{T,ki}$ . The corresponding AIRS spectra,  $Y_{T,ji}$ , are calculated from these profiles using the RTTOV-8 (Saunders, 2004) radiative transfer model,  $F_M$ ,

$$Y_{T,ji} = F_M(X_{T,ki}), \quad (1)$$

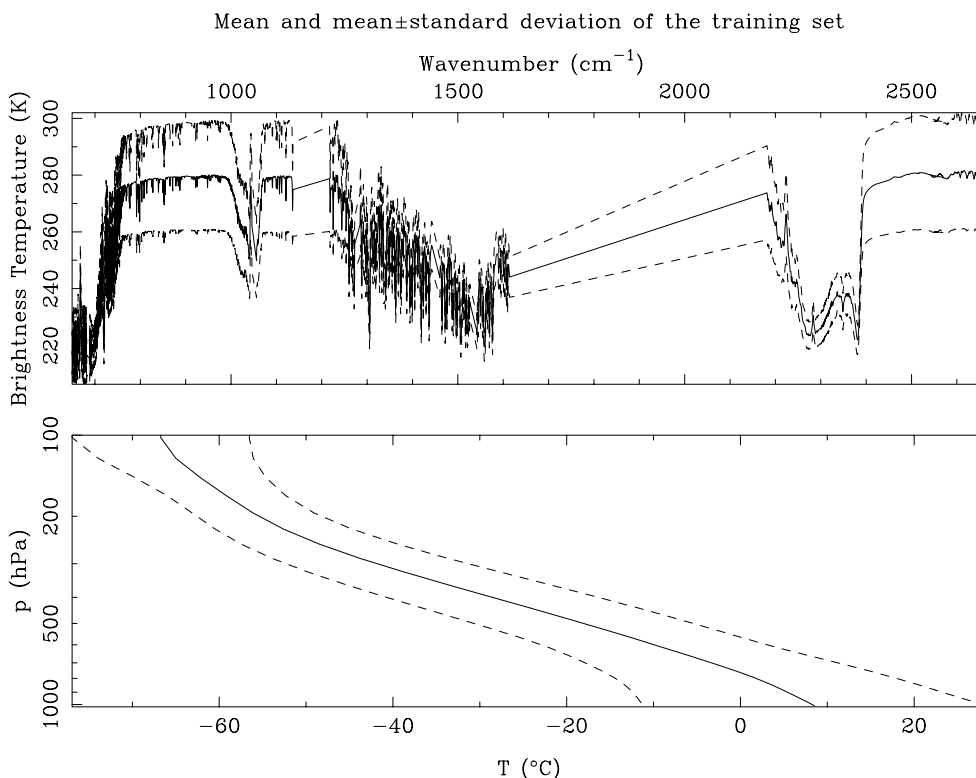
where,

$$i = 1 \dots n_T \text{ (Item number),}$$

$$j = 1 \dots m \text{ (Channel number),}$$

$$k = 1 \dots q \text{ (Atmospheric state parameter number),}$$

the subindex  $M$  stands for “model”, the subindex  $T$  for “training” data,  $n_T$  is the number of items or training sample



**Fig. 1.** Mean (solid line) and mean  $\pm$  one standard deviation (dashed line) of the temperature profile of the modified “Sampled database of 60-level atmospheric profiles from the ECMWF analyses” (Chevallier, 2002) (bottom) and their corresponding spectra statistics calculated using RTTOV-8 (top).

size,  $m$  is the number of channels in the infrared spectrum and  $q$  is the number of atmospheric state parameters.

Figure 1 illustrates the mean and mean  $\pm$  one standard deviation of the temperature profiles of these sample analyses and of their corresponding spectra obtained using RTTOV-8. Figure 2 shows a particular example of this dataset. The whole EOF retrieval process has been applied from surface pressure up to the highest RTTOV-8 level, 0.1 hPa. Since we are interested mainly in tropospheric retrievals only the data below 100 hPa is shown.

To obtain the EOF, the covariance matrix of the spectra must be calculated,

$$C_{T,jl} = \sum_{i=1}^{n_T} (Y_{T,ji} - \overline{Y_{T,j}})(Y_{T,li} - \overline{Y_{T,l}}), \quad (2)$$

where  $\overline{Y_{T,j}}$  is the average of the brightness temperature for all samples,  $n_T$ .

The covariance matrix can be diagonalized in the form,

$$\sum_{j=1}^m C_{T,ij} e_{jk} = \sigma_{T,k}^2 e_{ik}, \quad (3)$$

where  $e_{ik}$  are the eigenvectors and the eigenvalues are defined as  $\sigma_{T,k}^2$  for convenience. The eigenvalues  $\sigma_{T,k}^2$  will be ordered from higher to lower values as the  $k$  index increases.

The principal components or EOF scores of the spectra can now be calculated with,

$$Z_{T,ik} = \sum_{j=1}^m e_{ji} (Y_{T,jk} - \overline{Y_{T,j}}), \quad (4)$$

where,

$$k = 1 \dots n \text{ (Item number)},$$

$$i = 1 \dots p \text{ (Eigenvector number)},$$

and the value  $p$  is the number of eigenvectors used, which can run from 1 to the total number of channels,  $m$ .

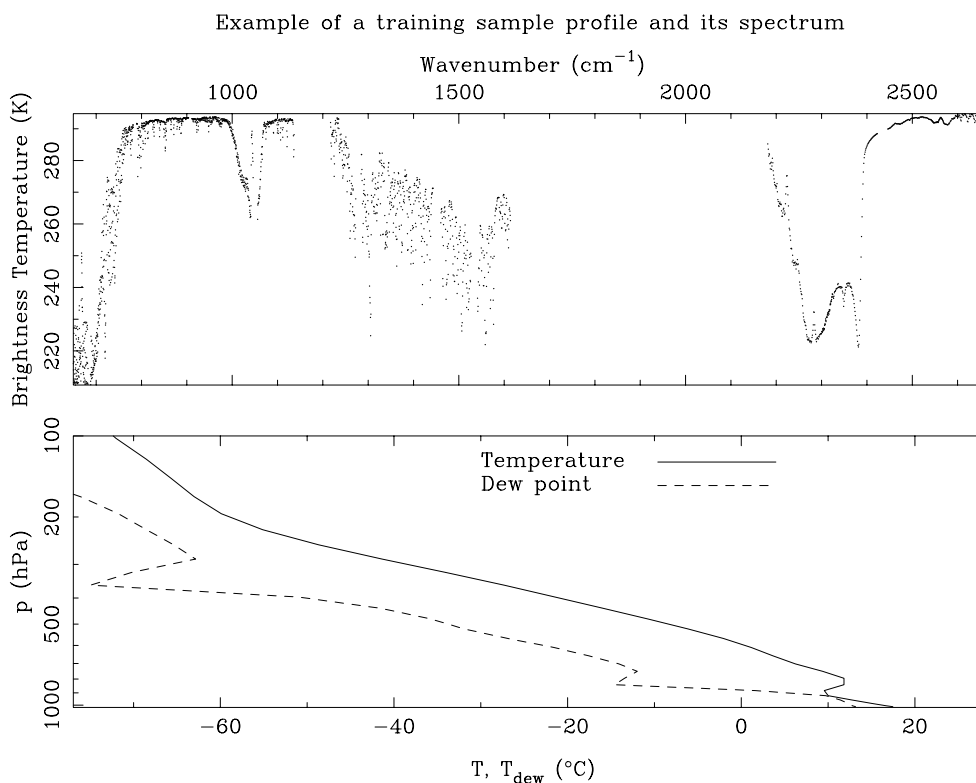
Finally, to be able to perform the retrievals, a linear regression with the atmospheric states is done,

$$X_{T,ki} = \sum_{j=1}^p \beta_{kj} Z_{T,ji} + \overline{X_{T,k}}, \quad (5)$$

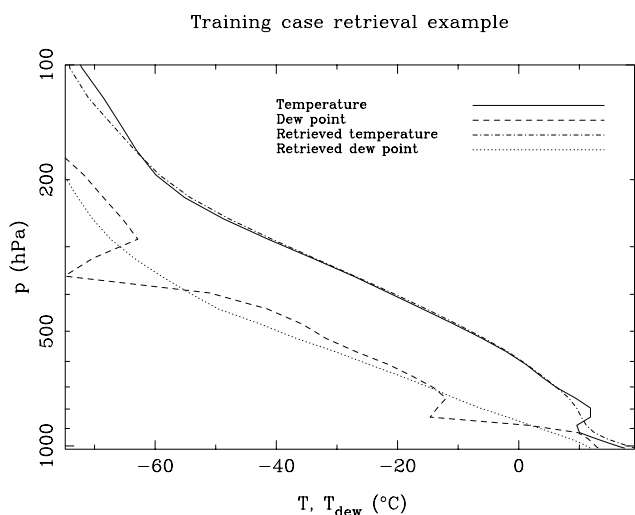
where  $\overline{X_{T,k}}$  is the atmospheric state average of all samples,  $n_T$ .

The linear regression coefficients can be calculated by least square minimization,

$$\beta_{kj} = \frac{1}{\sigma_{T,j}^2} \sum_{i=1}^{n_T} (X_{T,ki} - \overline{X_{T,k}}) Z_{T,ji}. \quad (6)$$



**Fig. 2.** Example of a particular training sample (bottom). The temperature (solid line) and dew point temperature (dashed line) are shown as well as its corresponding brightness temperature spectrum calculated using RTTOV-8 (top).



**Fig. 3.** Retrieval profile of the particular training example in Fig. 2 using 200 eigenvectors. The temperature (solid line) and dew point temperature (dashed line) of the original training profile are shown, as well as the retrieved temperature (dash-dotted line) and dew point temperature (dotted line).

## 2.2 Performing EOF retrievals

The retrieval method can be tested, for comparison purposes, with the same training cases. They will be defined as,

$$X_{RT,ki} = \sum_{j=1}^p \beta_{kj} \sum_{l=1}^m e_{lj} (Y_{T,li} - \overline{Y_{T,l}}) + \overline{X_{T,k}}, \quad (7)$$

where the subindex  $RT$  stands for “retrieval of the training” cases. A training profile retrieval, using the data from the example in Fig. 2, is shown in Fig. 3.

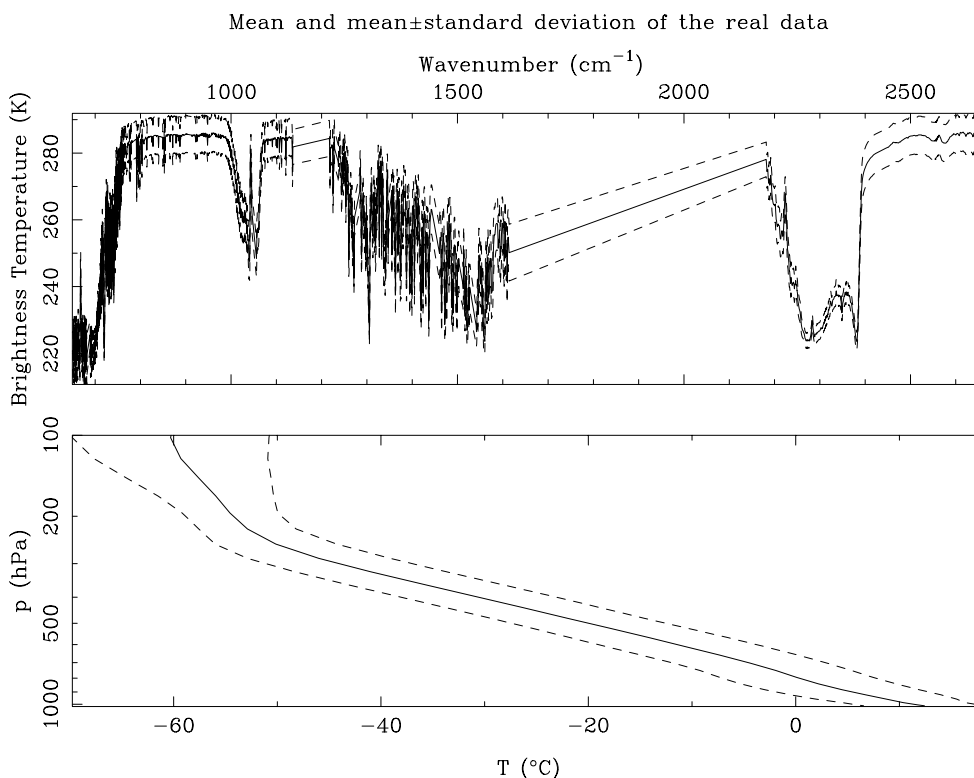
The real spectra can be derived from the atmospheric states by measuring them in a real atmosphere,

$$Y_{A,ji} = F_A(X_{A,ki}), \quad (8)$$

where,

- $i = 1 \dots n_A$  (Item number),
- $j = 1 \dots m$  (Channel number),
- $k = 1 \dots q$  (Atmospheric state parameter number),

the subindex  $A$  stands for “atmospheric” real cases,  $n_A$  is the number of measurements,  $m$  is the channel number in the infrared spectrum,  $q$  is the total number of atmospheric parameters and  $F_A$  represents the whole real system including the atmosphere and the measuring instrument.



**Fig. 4.** AIRS measured brightness temperature mean (solid line) and mean  $\pm$  one standard deviation (dashed line) of 8650 clear sky measurements during nighttime over ocean of the day 6 October 2003 (top). Also shown are the statistics of the corresponding ECMWF (ERA40) temperature analyses to those measurements (bottom).

In this paper, the real atmospheric measurements,  $Y_A$ , are the 8650 clear sky spectra from AIRS taken during 24 h of nighttime over ocean of a randomly chosen day, namely 6 October 2003. For the detection of clear-sky situations a number of threshold tests are applied as proposed by Lutz (2002) and Lutz et al. (2003), which are summarized in Table 2. The tests are very restrictive to assure that the amount of undetected cloud contamination remains negligible. Further restrictions consist of (Table 2):

- Nighttime measurements to avoid solar contamination of the spectra.
- Latitudes equatorward of  $50^\circ$  to avoid cold surfaces where cloud detection is difficult.
- Small scan angles ( $<15^\circ$ ).

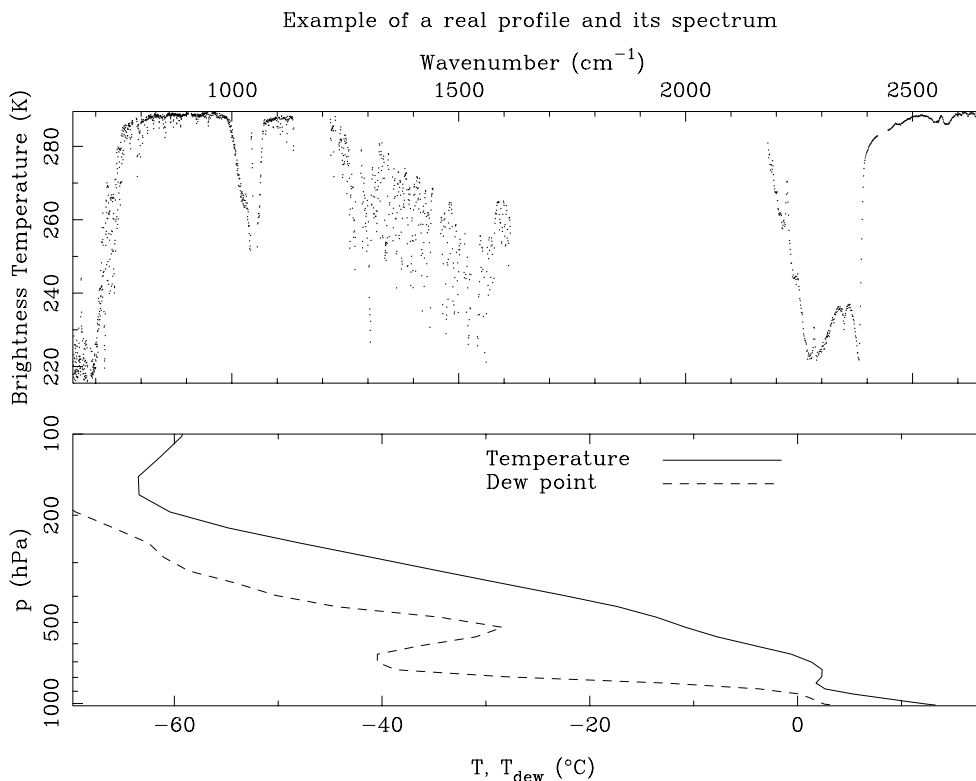
The closest, in space and time, ECMWF analyses of each one of the spectra is assumed to be the “real” atmospheric state,  $X_A$ . These analyses have been extracted from the ECMWF 40-year re-analysis project (ERA40). Figure 4 illustrates the mean and mean  $\pm$  one standard deviation of the AIRS spectra dataset and of their corresponding ECMWF temperature analyses. Figure 5 shows one particular example of the real atmospheric dataset.

**Table 2.** Scene selection.  $T(10.8 \mu\text{m})$ , for example, is the brightness temperature of an AIRS channel that lies in that wavelength ( $10.8 \mu\text{m}$ ). SST is the sea surface temperature derived from ECMWF analysis.

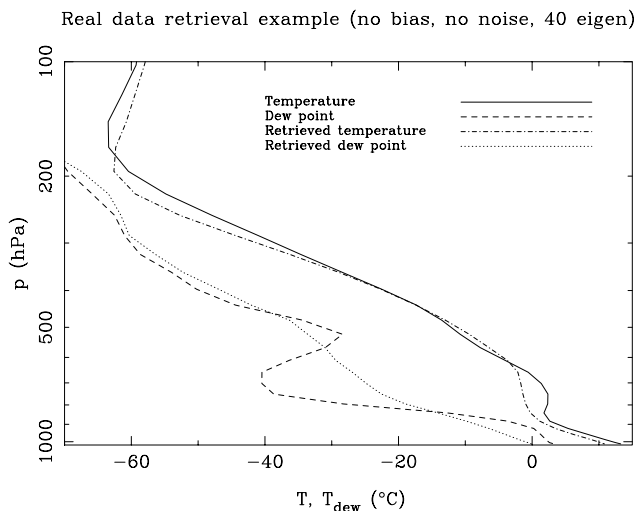
Cloud detection
$-1 \text{ K} < T(3.9 \mu\text{m}) - T(10.8 \mu\text{m}) < 3 \text{ K}$
$T(10.8 \mu\text{m}) > 276 \text{ K}$
$T(11.0 \mu\text{m}) > \text{SST} - 2.2 \text{ K}$
$T(4.0 \mu\text{m}) - T(11.0 \mu\text{m}) > 12 \text{ K}$
$T(9.3 \mu\text{m}) - T(11.0 \mu\text{m}) < 0 \text{ K}$
$T(11.0 \mu\text{m}) - T(12.0 \mu\text{m}) < 1 \text{ K}$
$T(11.0 \mu\text{m}) - T(13.6 \mu\text{m}) > 18 \text{ K}$
Others
$ \text{Solar zenith angle}  > 100^\circ$
$ \text{Latitude}  < 50^\circ$
$ \text{Scan angle}  < 15^\circ$

The retrievals of the real atmospheric states,  $X_{RA,ki}$ , can now be performed by using the linear regression as before,

$$X_{RA,ki} = \sum_{j=1}^p \beta_{kj} \sum_{l=1}^m e_{lj} (Y_{A,li} - \overline{Y_{T,l}}) + \overline{X_{T,k}}. \quad (9)$$



**Fig. 5.** Example of a particular real data sample. The measured AIRS spectra is shown (top), as well as the closest in space and time ECMWF temperature (solid line) and dew point temperature (dashed line) analysis (bottom).



**Fig. 6.** Retrieval profile of the particular real AIRS data example in Fig. 5 using 40 eigenvectors. The temperature (solid line) and dew point temperature (dashed line) of the ECMWF analysis are shown, as well as the retrieved temperature (dash-dotted line) and dew point temperature (dotted line). No bias correction or noise added to the training data set has been used in this case.

An example of a retrieval performed from the AIRS spectrum example shown in Fig. 5 is illustrated in Fig. 6.

### 3 Statistics of the retrievals

To determine the performance of the retrievals a comparison with some known truth must be made. In the case of the retrievals performed on the same training cases the obvious choice for comparison are the original profiles. In the case of the measured AIRS spectra, the retrievals will be compared with ECMWF analyses ( $X_A$ ). For most retrieved parameters, it is usually the case that the difference between the retrieved profiles and the original or real ones has a Gaussian distribution. Because of this, a good choice to characterize the statistics of the retrievals is to calculate the mean and standard deviation of this difference.

The mean of the difference or biases of the training cases is,

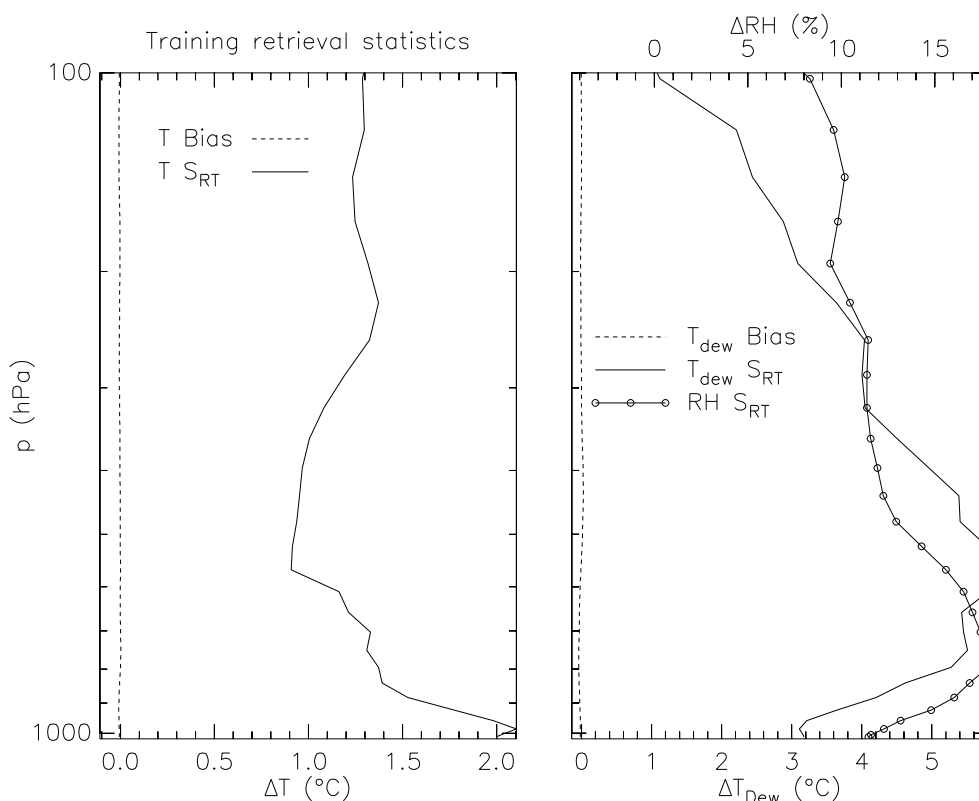
$$\overline{X_{RT,k} - X_{T,k}}. \quad (10)$$

With the real cases, the bias is,

$$\overline{X_{RA,k} - X_{A,k}}. \quad (11)$$

The square of the standard deviation or the covariance of the retrieved versus the original profiles is,

$$S_{RT,k}^2 \equiv \frac{1}{n_T} \sum_{i=1}^{n_T} (X_{RT,ki} - X_{T,ki})^2, \quad (12)$$



**Fig. 7.** Bias (dotted line) and standard deviation (solid line) of the retrievals performed on the training profiles using 200 eigenvectors for temperature (left) and dew point temperature and relative humidity (right).

for the training cases and,

$$S_{RA,k}^2 \equiv \frac{1}{n_A} \sum_{i=1}^{n_A} (X_{RA,ki} - X_{A,ki})^2, \quad (13)$$

for the real measured ones.

Figure 7 shows the computed statistics for the training cases. A zero bias is shown and a standard deviation between 2 K for the lower levels and 1 K for the upper ones.

Figure 8 shows the same statistics for the real data, that is, AIRS EOF retrievals compared with ECMWF analyses. The most significant feature in this graph is the large bias shown in the retrievals, which degrades their performance considerably. The standard deviation is within reasonable limits and is similar to the training cases of Fig. 7.

#### 4 Analytical derivation of the statistics of the retrievals

To understand the large bias observed in Fig. 8, an analytical derivation of the bias and standard deviation will be shown in this section. The bias of the training cases can be readily calculated obtaining the result,

$$\overline{X_{RT,k} - X_{T,k}} = 0, \quad (14)$$

with,

$$k = 1 \dots q \quad (\text{Atmospheric state number}). \quad (15)$$

In the case of the real cases, the bias result is,

$$\overline{X_{RA,k} - X_{A,k}} = \sum_{j=1}^p \beta_{kj} \sum_{l=1}^m e_{kj} (\overline{Y_{A,l}} - \overline{Y_{T,l}}) + (\overline{X_{T,k}} - \overline{X_{A,k}}), \quad (16)$$

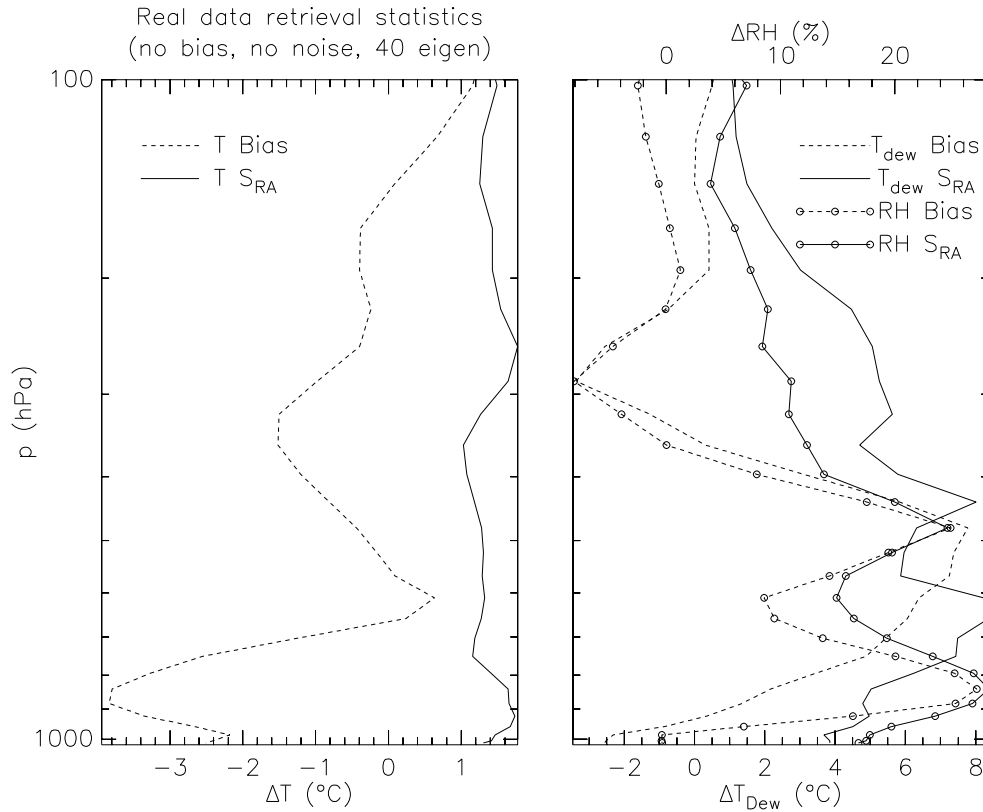
with,

$$k = 1 \dots q \quad (\text{Atmospheric state number}) \quad (17)$$

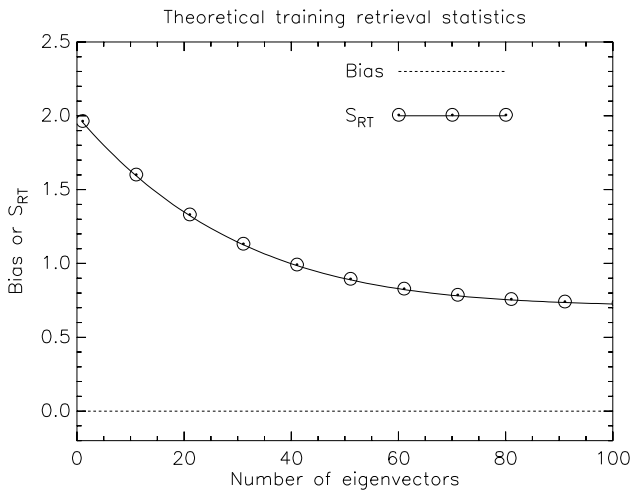
This important result shows that the biases of the retrievals are a sum of two terms. One is the second parenthesis in the right hand side of Eq. (16), which is the bias of the difference between the modeled atmospheric profiles used for training and the real atmospheric profiles. Another one is the first parenthesis in the right hand side of Eq. (16), which is the bias of the difference between the real atmospheric spectra and the modeled one used for training, transferred to the atmospheric profile space by the inversion process.

The training standard deviation can be resolved to give,

$$S_{RT,k}^2 = \frac{1}{n_T} \left[ \sum_{i=1}^{n_T} (X_{T,ki} - \overline{X_{T,k}})^2 - \right.$$



**Fig. 8.** Bias (dotted line) and standard deviation (solid line) of the retrievals performed on the real AIRS spectra when compared to collocated ECMWF analyses (ERA40) using 40 eigenvectors for temperature (left) and dew point temperature and relative humidity (right). The AIRS data consists in 8650 clear sky spectra taken during 24 h on 6 October 2003 during nighttime over ocean. No bias correction or noise added in the linear regression has been used in this case.



**Fig. 9.** Analytically derived curved for the bias (dotted line) and standard deviation (solid line with circles) of the retrievals of the training case as a function of the number of eigenvectors.

$$\sum_{j=1}^p \frac{1}{\sigma_{T,j}^2} \left[ \sum_{i=1}^{n_T} (X_{T,ki} - \overline{X_{T,k}}) Z_{T,ji} \right]^2, \quad (18)$$

with,

$$k = 1 \dots q \quad (\text{Atmospheric state number}). \quad (19)$$

From this equation we immediately see, as is shown in Fig. 9, that as we increase the number of eigenvectors, the standard deviation of the retrieval error will decrease monotonically. Its minimum value, which should be greater than zero, is reached when we use the maximum number of eigenvectors,  $p=m$ .

The solution to the real cases is a more complicated expression,

$$S_{RA,k}^2 = \frac{1}{n_A} \sum_{i=1}^{n_A} (X_{A,ki} - \overline{X_{T,k}})^2 + \frac{1}{n_A} \sum_{j=1}^p \beta_{kj} \sum_{l=1}^p \beta_{kl} \sum_{i=1}^{n_A} Z_{A,ji} Z_{A,li} - \frac{2}{n_A} \sum_{j=1}^p \beta_{kj} \sum_{i=1}^{n_A} (X_{A,ki} - \overline{X_{T,k}}) Z_{A,ji}, \quad (20)$$

with,

$$k = 1 \dots q \quad (\text{Atmospheric state number}). \quad (21)$$

To get a grasp of this equation, some simplifications must be made. Assuming that the covariance matrix of the EOF scores of the real cases is also diagonal,

$$\sum_{i=1}^{n_A} Z_{A,ji} Z_{A,ki} = \sigma_{A,j}^2 \delta_{jk}, \quad (22)$$

and that the cross-covariance matrix of the measured spectra and the modeled spectra is the same,

$$\sum_{i=1}^{n_A} (X_{A,ki} - \overline{X_{T,k}}) Z_{A,ji} = \sum_{i=1}^{n_T} (X_{T,ki} - \overline{X_{T,k}}) Z_{T,ji}, \quad (23)$$

the following result is obtained,

$$S_{RA,k}^2 = \frac{1}{n_A} \sum_{i=1}^{n_A} (X_{A,ki} - \overline{X_{T,k}})^2 - \sum_{j=1}^p \frac{\left( \sum_{i=1}^{n_T} (X_{T,ki} - \overline{X_{T,k}}) Z_{T,ji} \right)^2}{\sigma_{T,j}^2} \left[ 1 - \frac{\sigma_{A,j}^2 - \sigma_{T,j}^2}{\sigma_{T,j}^2} \right] \quad (24)$$

with,

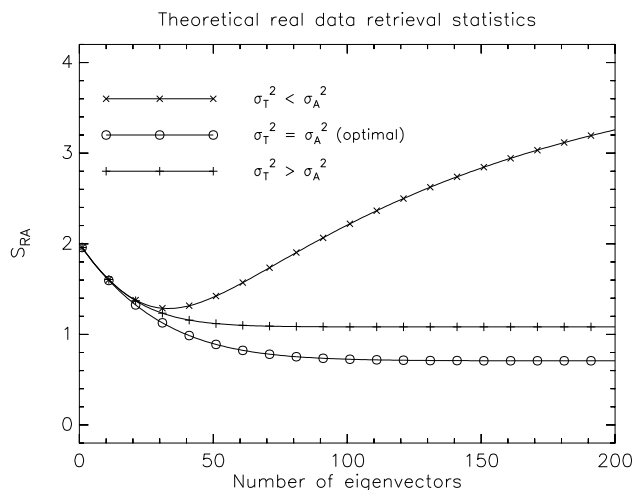
$$k = 1 \dots q \quad (\text{Atmospheric state number}). \quad (25)$$

The behavior of this covariance as a function of the number of eigenvectors is shown in Fig. 10. Three different cases can be distinguished,

1. **Exact match of modeled covariance and measured covariance**,  $\sigma_{T,j}^2 = \sigma_{A,j}^2$ . This case has the same solution as in the purely training case, Eq. (18). The results are shown in Fig. 10. The retrieved errors tend to decrease as the number of eigenvectors increase.
2. **Modeled covariance bigger than real covariance**,  $\sigma_{T,j}^2 > \sigma_{A,j}^2$ . In this case the retrieved errors also tend to decrease as the number of eigenvectors increase, as is shown in Fig. 10, but the overall errors are bigger than in the previous case.
3. **Modeled covariance smaller than real covariance**,  $\sigma_{T,j}^2 < \sigma_{A,j}^2$ . The behavior of this case, Fig. 10, is seen by assuming that  $\sigma_{A,j}^2 - \sigma_{T,j}^2$  is approximately constant as a function of the eigenvalue index  $j$ , on the basis that this difference will effectively be a residual noise of the measurements,  $Y$ , and recalling that the eigenvalues decrease with increasing index  $j$ . In this case, the errors in the retrievals decrease as the number of eigenvectors increases and then shows a minimum at the eigenvalue index  $k$  such that,

$$\sigma_{T,k}^2 = \sigma_{A,k}^2 - \sigma_{T,k}^2 \quad (26)$$

before increasing afterwards.



**Fig. 10.** Analytically derived curves for the standard deviation of the retrievals as a function of the number of eigenvectors of the real atmospheric cases. Three cases are shown: when the standard deviation of the real atmospheric states,  $\sigma_A^2$ , is bigger (x signs), the same (circles) or smaller (plus signs) than the standard deviation of the training cases,  $\sigma_T^2$ .

To calculate the optimal retrievals in the general case, Eq. (20), the smallest possible standard deviation of the differences between the retrieved and observed profiles should be obtained. This can be done by finding its minimum,

$$\frac{\partial S_{RA,k}^2}{\partial \beta_{kr}} = 0, \quad (27)$$

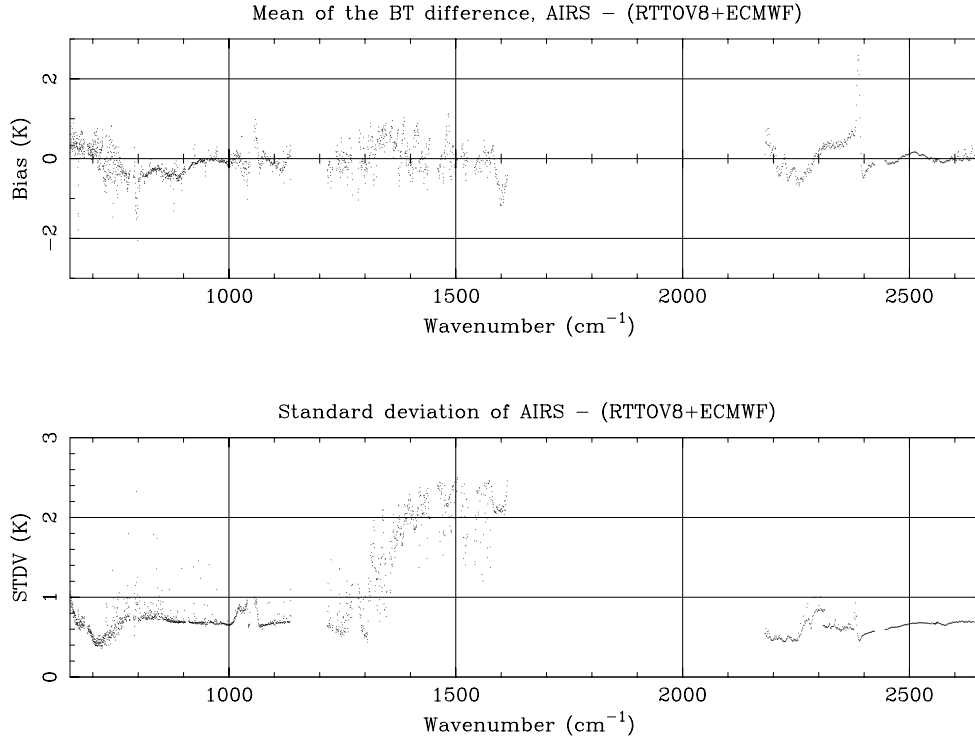
which gives as a result,

$$\begin{aligned} \sum_{i=1}^{n_T} (X_{T,ji} - \overline{X_{T,j}})(Y_{T,ki} - \overline{Y_{T,k}}) = \\ \sum_{i=1}^{n_A} (X_{A,ji} - \overline{X_{T,j}})(Y_{A,ki} - \overline{Y_{T,k}}), \\ \sum_{i=1}^{n_T} (Y_{T,ji} - \overline{Y_{T,j}})(Y_{T,ki} - \overline{Y_{T,k}}) = \\ \sum_{i=1}^{n_A} (Y_{A,ji} - \overline{Y_{T,j}})(Y_{A,ki} - \overline{Y_{T,k}}). \end{aligned} \quad (28)$$

This result for the general case confirms what was previously obtained in the particular case of Eq. (24), Fig. 10, when  $\sigma_{T,j}^2 = \sigma_{A,j}^2$ .

## 5 Estimation of the optimal parameters

The result from optimal parameters of Eq. (28) provides what is the ideal situation when performing retrievals. In real cases this is not normally the case and there is usually a significant



**Fig. 11.** Bias (top) and standard deviation (bottom) of the difference between the measured AIRS brightness temperature and the “calculated” ones with the ECMWF analyses and RTTOV8.

difference between the modeled and the real covariance matrices caused by both instrumental noise and model noise. In which way can we estimate the retrieval parameters so that we get the best possible retrievals with a given set of radiative transfer model and observations?

### 5.1 Estimation of the biases and covariance matrix corrections

A good estimation of the bias and covariance matrix correction to the training and measured cases can be obtained by calculating the mean and covariance of the difference between the measured spectra and the “calculated” one, denoted by  $Y_{C,ki}$ . Given a set of measurements  $X_{A,ji}$  and  $Y_{A,ki}$ , the calculated spectra can be derived from the set of atmospheric profiles and the radiative transfer model used by,

$$Y_{C,ki} = F_M(X_{A,ji}). \quad (29)$$

The bias of the difference between the real measured spectra and the calculated one can now be obtained by,

$$\overline{Y_{A,k} - Y_{C,k}}, \quad (30)$$

and the standard deviation by,

$$\frac{1}{n_A} \sum_{i=1}^{n_A} [Y_{A,ki} - Y_{C,ki} - (\overline{Y_{A,k} - Y_{C,k}})]^2. \quad (31)$$

Both statistics are shown in Fig. 11. In Fig. 12 the instrumental noise is compared with the standard deviation of Eq. (31).

To calculate analytically the covariances of this difference it should be noted that since the “calculated” profiles are derived using the radiative transfer model,  $F_M$ , it is reasonable to assume that their covariances are similar,

$$\overline{Y_C Y_C} \simeq \overline{Y_T Y_T}. \quad (32)$$

On the other hand, since the “calculated” profiles are derived from the real atmospheric states, their mean should be similar,

$$\overline{Y_{C,k}} \simeq \overline{Y_{A,k}}. \quad (33)$$

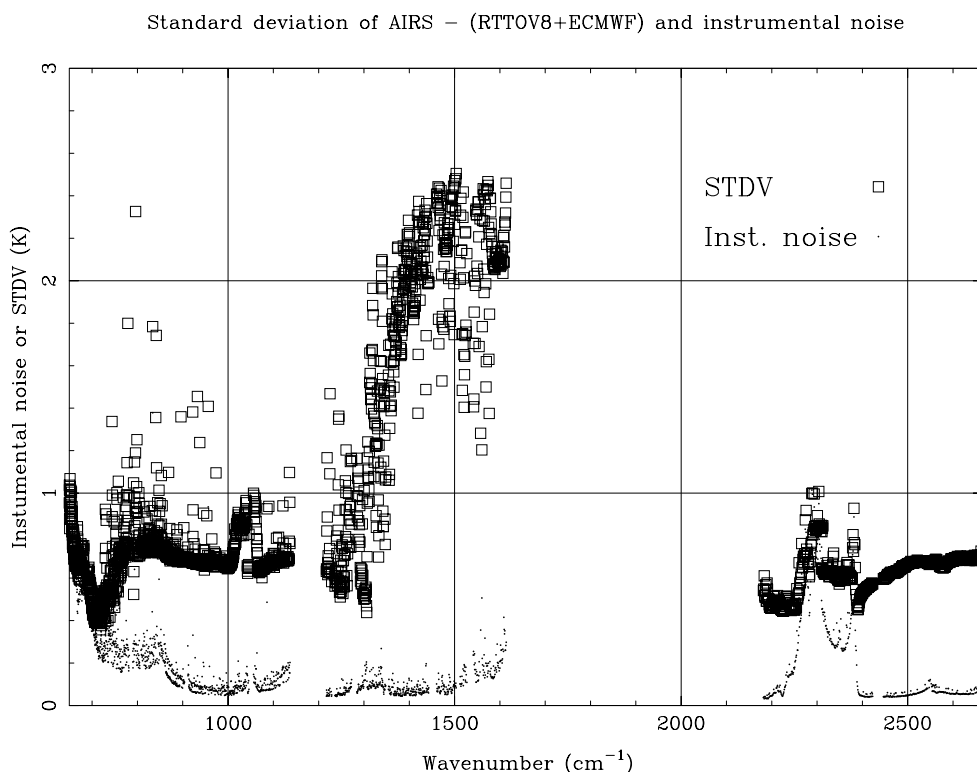
With this in mind one can assume that the measured radiances are equal to the calculated ones plus an added noise term,

$$Y_{A,ji} = Y_{C,ji} + \mu_{ji}, \quad (34)$$

such that the noise term is independent of the calculated value, in the sense that,

$$\sum_{i=1}^{n_A} (Y_{C,ji} - \overline{Y_{C,j}}) \mu_{ki} = \sum_{i=1}^{n_A} \mu_{ji} (Y_{C,ki} - \overline{Y_{C,j}}) = 0. \quad (35)$$

This assumption can hold if the added noise is random or it is systematic but “well behaved” in the sense that satisfies the above equation.



**Fig. 12.** “Total error”, equivalent to the standard deviation of the difference between the measured AIRS brightness temperature and the “calculated” ones with the ECMWF analyses and RTTOV8 (squares) and instrumental noise for the temperature profile of Fig. 5 (dots).

The final covariance of the differences can be calculated by using Eqs. (32) and (34),

$$\sum_{i=1}^{n_A} (Y_{A,ji} - Y_{C,ji} - \overline{(Y_{A,j} - Y_{C,j})}) \cdot (Y_{A,ki} - Y_{C,ki} - \overline{(Y_{A,k} - Y_{C,k})}) \simeq \sum_{i=1}^{n_A} Y_{A,ji} Y_{A,ki} - \sum_{i=1}^{n_M} Y_{T,ji} Y_{T,ki}. \quad (36)$$

## 5.2 Performing bias and covariance matrix corrections

It is now possible to correct the observations and the training sample to obtain the best possible EOF retrievals given the available data and radiative transfer model.

Modifying the measured radiances by subtracting the biases calculated using Eq. (30),

$$\hat{Y}_{A,ki} = Y_{A,ki} - \overline{Y_{A,k} - Y_{C,k}}, \quad (37)$$

the bias of the retrievals using these values,  $\hat{X}_{RA,ji}$ , can be obtained by,

$$\hat{X}_{RA,k} - X_{A,k} = \sum_{j=1}^p \beta_{kj} \sum_{l=1}^m e_{kj} (\overline{F_M(X_{A,k})} - \overline{F_M(X_{T,k})}) + (\overline{X_{T,k}} - \overline{X_{A,k}}), \quad (38)$$

and by assuming that the retrieval is nearly the inverse of the forward model,

$$\sum_{j=1}^p \beta_{kj} \sum_{l=1}^m e_{kj} F_M \simeq \text{Identity}. \quad (39)$$

The final resulting biases are nearly zero,

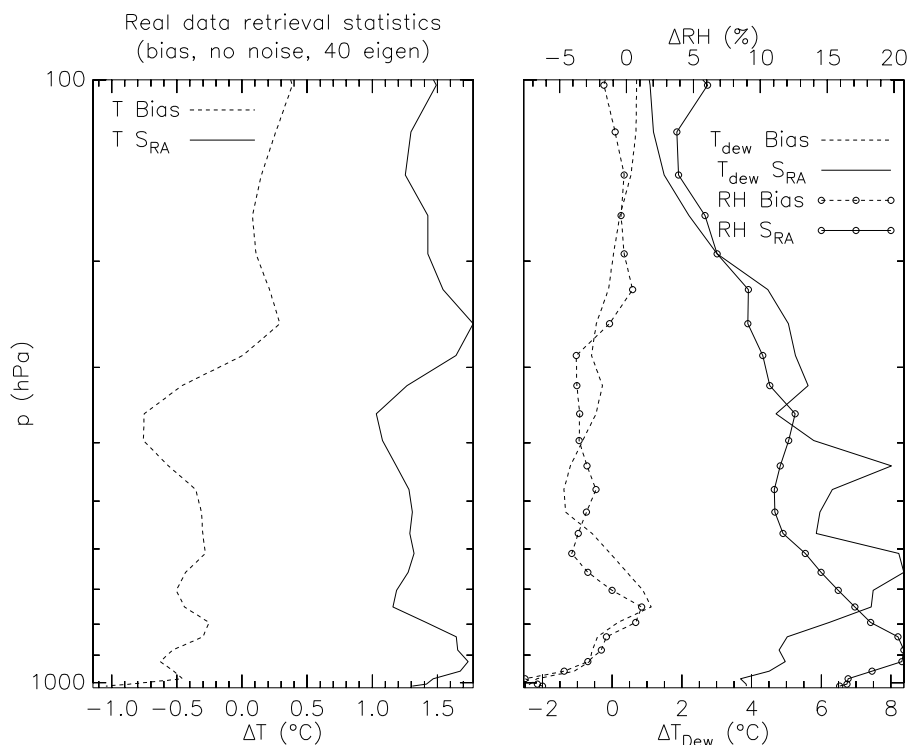
$$\overline{\hat{X}_{RA,k} - X_{A,k}} \simeq 0. \quad (40)$$

The covariance corrections will be applied on the spectra of the training cases, adding to them a random noise component,

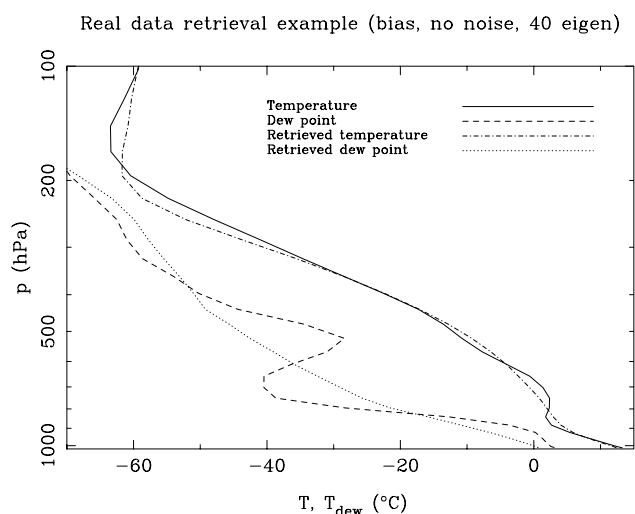
$$\hat{Y}_{T,ki} = Y_{T,ki} + \epsilon_{ki}, \quad (41)$$

where the covariance matrix of the added error,  $\epsilon_{ki}$ , is the same as the one in Eq. (36),

$$\sum_{i=1}^{n_M} \epsilon_{i,j} \epsilon_{i,k} = \sum_{i=1}^{n_A} (Y_{A,ji} - Y_{C,ji} - \overline{(Y_{A,j} - Y_{C,j})}) \cdot (Y_{A,ki} - Y_{C,ki} - \overline{(Y_{A,k} - Y_{C,k})}). \quad (42)$$



**Fig. 13.** Bias (dotted line) and standard deviation (solid line) of the retrievals performed on the real bias corrected AIRS spectra when compared to collocated ECMWF analyses using 40 eigenvectors for temperature (left) and dew point temperature and relative humidity (right).



**Fig. 14.** Retrieval profile of the particular real AIRS data example in Fig. 5 using 40 eigenvectors. The temperature (solid line) and dew point temperature (dashed line) of the ECMWF analysis are shown, as well as the retrieved temperature (dash-dotted line) and dew point temperature (dotted line). Bias correction has been applied but no noise has been added to the training data in this case.

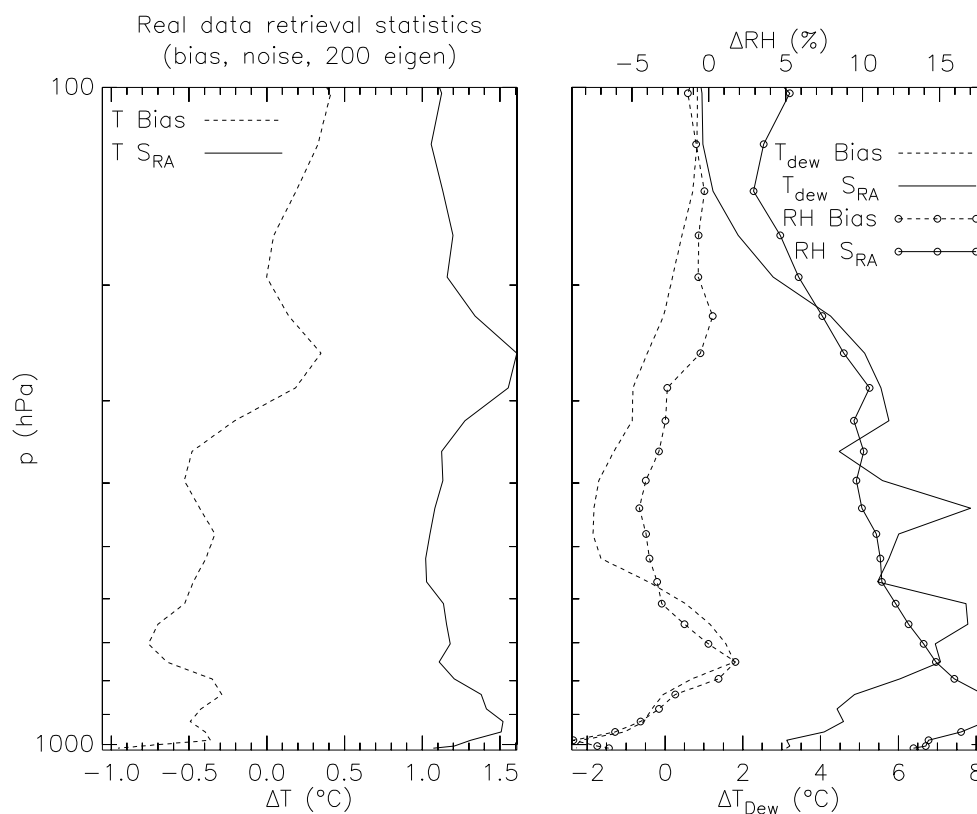
Calculating the covariance of the new training spectra by using Eq. (36) gives,

$$\sum_{i=1}^{n_T} (\hat{Y}_{T,ji} - \overline{\hat{Y}_{T,j}})(\hat{Y}_{T,ki} - \overline{\hat{Y}_{T,k}}) \approx \sum_{i=1}^{n_A} (Y_{A,ji} - \overline{Y_{T,j}})(Y_{A,ki} - \overline{Y_{T,k}}), \quad (43)$$

which is the same as the optimal covariance for the retrievals, Eq. (28). Assuming that the cross-covariances of atmospheric states and spectra are approximately equal for the modeled and measured case, that is, the first equation of the set of Eqs. (28) holds, then the conditions for an optimal EOF retrieval would have been reached.

## 6 Real case calculation of the statistics of the retrievals using the optimal parameters

It is now possible to calculate the statistics of the training and the real data retrievals with the corrected parameters and data to compare them with the theory. The biases and standard deviations calculated for the training cases have been shown in Fig. 7. In Fig. 18 the mean bias and standard deviation of the temperature profiles below 300 hPa versus the number of eigenvectors used is plotted. In this figure a zero bias and a



**Fig. 15.** Bias (dotted line) and standard deviation (solid line) of the retrievals performed on the real bias corrected AIRS spectra when compared to collocated ECMWF analyses using 200 eigenvectors for temperature (left) and dew point temperature and relative humidity (right). In this case the optimal noise has been added to the training profiles for the linear regression.

standard deviation that approaches a certain value asymptotically as the number of eigenvectors increases is shown. This result coincides with the analytical derivation of Eq. (18) and Fig. 9.

The biases and standard deviations of the real world uncorrected measurements has been shown in Fig. 8. In Fig. 17 the biases for the temperature profiles have been split in to the two sums of Eq. (16). Both terms, the bias between the modeled training atmospheric profiles and the real atmospheric profiles,  $\overline{X_{T,k} - X_{A,k}}$ , and the bias between the modeled training spectra and the real atmospheric spectra projected to the atmospheric profile space by the inversion,  $\sum_{j=1}^p \beta_{kj} \sum_{l=1}^m e_{kj} (\overline{Y_{A,l}} - \overline{Y_{T,l}})$ , show a significant contribution to the overall detected bias,  $\overline{X_{RA,k} - X_{A,k}}$ .

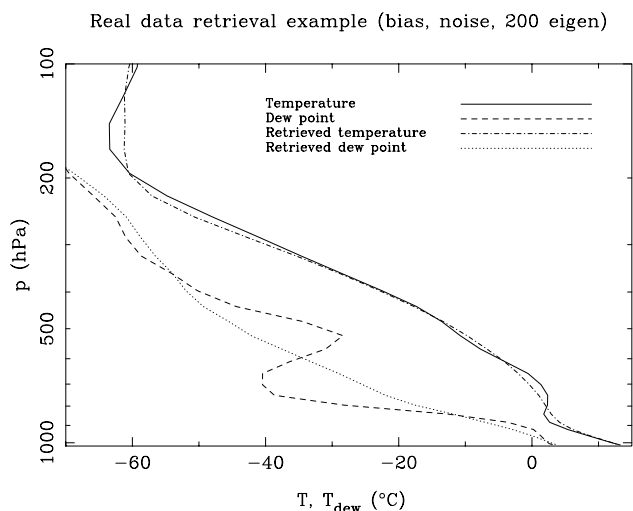
When the bias corrections of Eq. (37) are applied to the data, the final bias of the retrievals is reduced significantly as expected from Eq. (40). These results are shown in Fig. 13. The standard deviation lies between 1 and 1.7 K. A particular retrieval is shown in Fig. 14, which corresponds to the data example of Fig. 5.

When the optimal noise from Eq. (41) is added to the linear regression of the training, the retrievals are further improved as was expected. This is shown in Fig. 15. The standard

deviation has improved and now lies between 1 and 1.5 K. A particular retrieval is shown in Fig. 16, which corresponds to the data example of Fig. 5.

Figure 19 shows the mean bias and standard deviation of the temperature profiles below 300 hPa versus the number of eigenvectors used for the real atmospheric measurements. Results for four different noise types used in the linear regression of the training are shown. This figure shows a very similar behavior to the three cases of the theoretical curve of Fig. 10, i.e., when the covariance of the real atmospheric states is bigger, the same, or smaller than the covariance of the training cases, respectively.

In the end, the optimal standard deviation used for the retrievals is really showing the “total error” introduced in the retrieval, including instrument noise and calibration, radiative transfer model errors and errors in the measured atmospheric states (ECMWF analyses). In Fig. 12 this error is plotted together with the instrumental noise. It is shown that the “total error” in most wavenumbers is much larger than the instrumental noise.



**Fig. 16.** Retrieval profile of the particular real AIRS data example in Fig. 5 using 200 eigenvectors. The temperature (solid line) and dew point temperature (dashed line) of the ECMWF analysis are shown, as well as the retrieved temperature (dash-dotted line) and dew point temperature (dotted line). Bias correction and noise added to the training data set has been applied in this case.

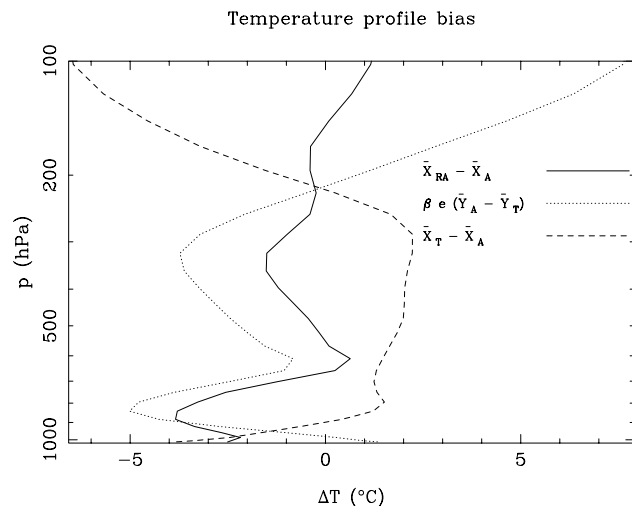
## 7 Conclusions

### 7.1 General

Given the specific algorithm shown in this paper, which consists of fitting a linear regression to the EOF components of synthetic spectral data, and a given atmospheric data set, it has been proven analytically that the optimal retrieval is obtained by performing the following steps:

1. Obtain from the real atmospheric profiles and the radiative transfer model (in our case ECMWF analyses and RTTOV-8) the “calculated” spectra. These spectra are then subtracted from the observed measured spectra (AIRS). Finally the mean of this difference and its standard deviation is calculated.
2. When performing the linear regression of the training data a Gaussian noise component should be added to the training spectra with a standard deviation that matches the one above, that is, the one obtained from the difference of observations minus “calculated” spectra.
3. When performing the retrievals, the measured spectra (AIRS) should be bias corrected with the aforementioned value, that is, the average of the difference between the observation minus the “calculated” spectra.

The reason for the existence of a bias arise from the fact of using different sets of data for training and retrieval and from a divergence between observed and calculated radiative measurements with differing statistics. The origin of this “total” noise, and thus its bias and standard deviation, can



**Fig. 17.** Biases of the temperature profile (solid line is the total bias) for the real measurements. Both sums of Eq. (16) are shown: the bias between the modeled training atmospheric profiles and the real atmospheric profiles,  $\bar{X}_{T,k} - \bar{X}_{A,k}$  (dashed line), and the bias of the modeled training spectra and the real atmospheric spectra projected to the atmospheric profile space by the inversion,  $\sum_{j=1}^p \beta_{kj} \sum_{l=1}^m e_{kj} (\bar{Y}_{A,l} - \bar{Y}_{T,l})$  (dotted line).

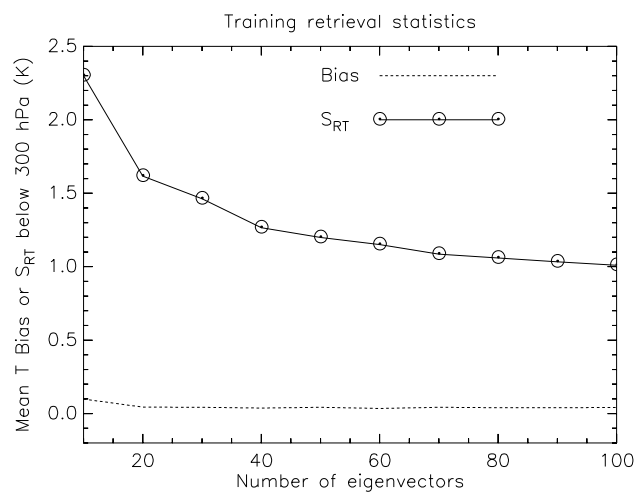
come from instrument noise, errors in the radiative transfer model and poor representativeness of the atmospheric states (ECMWF analyses in this case). It is very difficult, if not impossible, to discriminate between these three and to detect which source is the most significant one with the data used in this paper.

The bias correction is critical for the success of the EOF retrievals. If these bias corrections are not applied, significant biases appear in the retrievals degrading them significantly (compare Figs. 8 and 13).

Adding the optimal noise to the EOF retrievals is not critical and reasonable retrievals can be obtained without it (Fig. 13). Although its addition improves the retrieval by a noticeable amount (compare with Fig. 15). An added benefit to the use of the optimal noise is that the number of eigenvectors is not critical as long as it is high enough to reach the plateau observed in Fig. 19. This is not the case when a smaller than optimal noise is added and thus the optimal number of eigenvectors must be found (Eq. 26 and Fig. 10).

The optimal bias corrections and added noise that have been derived in this paper imply that to improve the EOF retrievals one must resort to either changing the overall algorithm or using other datasets, like for example, training the retrievals with latitude classified data or obtaining the real atmospheric profiles from another source such as radiosondes.

One drawback of this technique is that the retrievals will be fine tuned to whatever data we have used as real world atmospheric profiles (ECMWF in this case). The retrievals will try to resemble this real world data set.



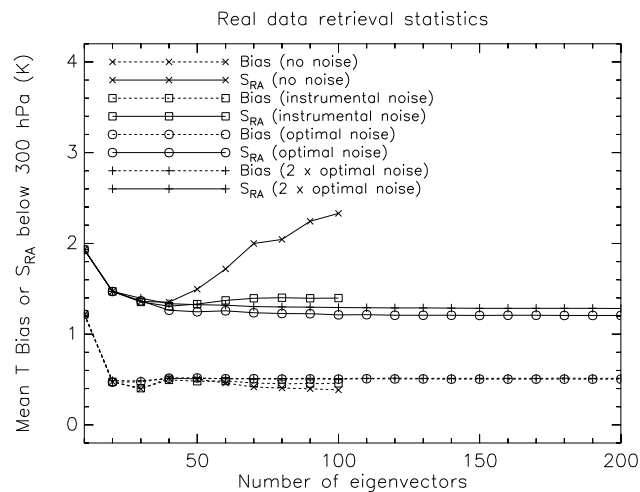
**Fig. 18.** Mean bias (dotted line) and standard deviation (solid line with circles) of the temperature profiles below 300 hPa versus the number of eigenvectors used when the training profiles are compared with its retrievals.

This technique has been tested with real data from 24 h of a randomly chosen data set (namely 8650 clear sky spectra on 6 October 2003 during nighttime and over ocean) and it has been optimized for this same data set. It is not exactly known how this technique can be extended to other dates, in the case that, for example, the biases change slightly with time. This effect could lead in the end to final biases when using the data for climatological purposes. This effect could be specially difficult to solve if the bias changes occur because of real atmospheric variability.

## 7.2 Other algorithms

To overcome the problem of the bias and noise corrections altogether an alternative EOF technique could be used, by using the same training data set as the one to be retrieved. The EOF retrieval could be trained with direct radiative measurements and radiosonde profiles for example. In this case the statistics of the training and retrieved data sets should be the same showing none of the problems studied in this paper. But if this ideal situation is not met and there is a statistical difference between the training dataset and the one used for retrievals a bias will show up (Eq. 16). In this case part of the theoretical analysis derived in this paper could be used. Biases corrections could be derived in a similar way as shown here (Eq. 37). If the standard deviations are also different, there will be a noise mismatch degrading the retrievals (Eq. 20). Standard deviation corrections could be applied by adding noise to one of the real measurements until both covariances are matched (Eq. 28).

Using real measurements for training is not exempt of drawbacks. The first one of them is that normally the set of satellite radiative data with collocated radiosondes mea-



**Fig. 19.** Mean bias (dotted line) and standard deviation (solid line) of the temperature profiles below 300 hPa of the real atmospheric retrievals when compared to the real profiles. Four different types of noise have been used in the linear regression of the training: no noise (x signs), instrumental noise (squares), optimal noise (circles) and twice the optimal noise (plus signs). Compare this figure with the theoretically derived one Fig. 10.

surements is usually scarce. This will give rise to probable differences between the statistics of the training data set and the retrieved one. Another one is that if the training data set is obtained in a specific region of the planet, it will not be global enough to perform universal retrievals, leading again to biases.

Edited by: U. Pöschl

## References

- Aumann, H. H., Chahine, M. T., Gautier, C., Goldberg, M. D., Kalnay, E., McMillin, L. M., Revercomb, H., Rosenkranz, P. W., Smith, W. L., Staelin, D. H., Strow, L. L., and Susskind, J.: AIRS/AMSU/HSB on the Aqua Mission: Design, Science Objectives, Data Products, and Processing Systems, *IEEE Trans. Geosci. Remote Sens.*, 41, 253–264, 2003.
- Chevallier, F.: Sampled database of 60-level atmospheric profiles from the ECMWF analyses, NWP SAF Technical Report No. 4, 2002.
- Houshangpour, A., John, V. O., and Buehler, S. A.: Retrieval of upper tropospheric water vapor and upper tropospheric humidity from AMSU radiances, *Atmos. Chem. Phys.*, 5, 2019–2028, 2005, **SRef-ID: 1680-7324/acp/2005-5-2019**.
- Huang, H. and Antonelli, P.: Application of Principal Component Analysis to High-Resolution Infrared Measurement Compression and Retrieval, *J. Appl. Meteorol.*, 40, 365–388, 2001.
- Lutz, H. J.: Scenes Analysis from MODIS and Meteosat Observations, Proceedings of the 2002 EUMETSAT Meteorological Satellite Data Users' Conference, pp. 8, 2002.

- Lutz, H. J., Inoue, T., and Schmetz, J.: Comparison of a Split-window and a Multi-spectral Cloud Classification for MODIS Observations, *J. Meteorol. Soc. Japan*, 81(3), 623–631, 2003.
- Saunders, R.: RTTOV-8 Users' Guide, NWP SAF, Met Office, <http://www.metoffice.com/research/interproj/nwpsaf/rtm/index.html>, 2004.
- Schlüssel, P., Hultberg, T. H., Phillips, P. L., August, T., and Calbet, X.: The operational IASI Level 2 Processor, *Adv. Space Res.*, 36, 982–988, 2005.
- Schneider, M., Hase, F., and Blumenstock, T.: Water vapour profiles by ground-based FTIR spectroscopy: study for an optimised retrieval and its validation, *Atmos. Chem. Phys. Discuss.*, 5, 9493–9545, 2005,  
**SRef-ID: 1680-7375/acpd/2005-5-9493.**
- Zhou, D. K., Smith, W. L., Li, J., Howell, H. B., Cantwell, G. W., Larar, A. M., Knuteson, R. O., Tobin, D. C., Revercomb, H. E., and Mango, S. A.: Thermodynamic product retrieval methodology and validation for NAST-I, *Appl. Opt.*, 4, 6957–6967, 2002.