Remote sensing from the infrared atmospheric sounding interferometer instrument
1. Compression, denoising, and first-guess retrieval algorithms

F. Aires
Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, USA
NASA Goddard Institute for Space Studies, New York, New York, USA

W. B. Rossow
NASA Goddard Institute for Space Studies, New York, New York, USA

N. A. Scott and A. Chédin
Laboratoire de Météorology Dynamique, École Polytechnique, Palaiseau, France

Received 19 June 2001; revised 4 February 2002; accepted 5 February 2002; published 19 November 2002.

[1] A principal component analysis (PCA) scheme is developed for treatment of observations from the high spectral resolution Infrared Atmospheric Interferometer (IASI) spaceborne instrument. Compression and denoising of IASI observations are performed using this PCA. This preprocessing methodology also allows for a fast pattern recognition to obtain a first guess from a climatological data set. The performance of the compression, denoising, and multivariate first-guess retrieval are evaluated with a large diversified data set of radiosondes atmospheres including rare events. Overall, the instrumental noise in the overall observed IASI spectrum goes from 0.9 to 0.2 K after denoising. This analysis procedure will be used by *Aires et al.* [2002c] to retrieve simultaneously temperature, water vapor and ozone atmospheric profiles. *INDEX TERMS:* 0399 Atmospheric Composition and Structure: General or miscellaneous; 3260 Mathematical Geophysics: Inverse theory; 3337 Meteorology and Atmospheric Dynamics: Numerical modeling and data assimilation; 3360 Meteorology and Atmospheric Dynamics: Remote sensing; *KEYWORDS:* infrared interferometer, principal component analysis, channel selection

Citation: Aires, F., W. B. Rossow, N. A. Scott, and A. Chédin, Remote sensing from the infrared atmospheric sounding interferometer instrument, 1, Compression, denoising, first-guess retrieval inversion algorithms, *J. Geophys. Res.*, *107*(D22), 4619, doi:10.1029/2001JD000955, 2002.

1. Introduction

[2] The Infrared Atmospheric Sounding Interferometer (IASI), is a high resolution (0.25 cm^{-1}) Fourier transform spectrometer scheduled for flight in 2005 on the European polar Meteorological Operational Platform satellite funded by the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) and the European Space Agency member states. This instrument is intended to replace the High-Resolution Infrared Radiation Sounder (HIRS) as the operational infrared sounder and is expected to reach accuracies of 1 K in temperature and 10% in water vapor with vertical resolutions of 1 km and 2 km respectively (cloud-free). IASI, jointly developed by the Centre National d'Études Spatiales (and EUMETSAT, covers the wavelength range from 3.5 μm to 15.5 μm at considerably higher spectral resolution than HIRS and, together with the Advanced Microwave Sounding Unit (AMSU) and the Microwave Humidity Sounder, is expected to lead to dramatic improvements in the accuracy and vertical resolution of remotely sensed temperature and humidity profiles and ozone amount.

[3] The dimension (number of measurements per field-of-view) of IASI observations is much higher than for previous instruments: 8461 channels compared to 19 for HIRS on TIROS-N Operational Vertical Sounding (TOVS). This is a major problem in the definition of retrieval algorithms. Classical retrieval algorithms are often unable to deal with this amount of information. Iterative methods require a fast direct model with its Jacobians (i.e., the first derivative of the observation with respect to retrieved variables), such as the radiative transfer model for IASI (RTIASI) model [*Matricardi and Saunders*, 1999]. Variational assimilation techniques also need a fast forward model with the Jacobians or the tangent linear operator. These approaches are unlikely to use the full raw IASI information because of the large dimension of the observations. Neural network techniques also have difficulties handling such an amount of information [*Aires et al.*, 1998]. To deal with this high-dimension problem, various techniques to select channels in the IASI spectrum have been developed by *Rabier et al.* [2002] or by *Aires et al.* [2002a].

0148-0227/02/2001JD000955

[4] Noise is also a major concern since IASI has high levels of instrument noise in some spectral regions. In this context, the full use of the IASI channels can help reduce instrument noise effect by exploiting information redundancy in channels.

[5] We propose here to use, as did *Huang and Antonelli* [2001], a feature extraction approach based on a principal component analysis (PCA) representation of the IASI observations. This idea has been used in the past, but the study of *Huang and Antonelli* [2001] has shown that it is useful for high resolution, noisy infrared observations. We apply this method to IASI using the instrumental characteristics specified for this instrument (in terms of spectral resolution and instrumental noise levels) with an IASI-dedicated radiative transfer model. We also show that this PCA representation of the IASI observations can be used to design a fast pattern recognition for the retrieval of a first guess from a climatological data set (which is needed in a retrieval scheme). Even if this study concerns specifically the IASI instrument, the algorithms that we develop can easily be adapted to other instruments, in particular such as the atmospheric infrared sounder on board the Aqua spacecraft.

[6] We have tested all these approaches (compression, denoising and first-guess retrieval) over a large number of real atmospheric situations as measured by radiosondes, taken from the Thermodynamic Initial Guess Retrieval (TIGR) database [*Chédin et al.*, 1985; *Achard*, 1991; *Escobar*, 1993; *Chevallier et al.*, 1998].

[7] This paper is organized as follows: the description of the IASI instrument is presented in section 2. Section 3 describes the PCA and its application to IASI spectra. Section 4 present compression, denoising and first-guess retrieval results based on PCA of IASI spectra. Section 5 concludes this study with some perspectives on this work.

2. IASI

2.1. Characteristics of IASI

[8] The two major advances of the IASI instrument are (1) the dramatically increased number of spectral channels: For each field of view, 8461 measures are available covering the spectral range from 645 to 2760 cm^{-1} with a resolution (unapodized) of 0.25 cm^{-1}; and (2) the increased resolution power: With IASI the resolution power is two order of magnitude higher with such instruments as the TOVS HIRS radiometer. So, it is expected that the vertical resolution and the accuracy of retrievals will substantially increase; the IASI mission specifications are a mean error of 1 K in atmospheric temperature and 10% in relative humidity profiles with, respectively, 1 Km and 2 Km vertical resolution. Table 1 represents some of the most important spectral regions and their associated absorbing constituent.

Table 1. IASI Spectral Information

Spectral Region, cm^{-1}	Variable
650 to 770	CO_2 temperature sounding
770 to 980	surfaces cloud properties
1000 to 1070	O_3 sounding
1080 to 1150	surfaces cloud properties
1210 to 1650	water vapor sounding; N_2O and CH_4 column amounts
2100 to 2150	CO column amount
2150 to 2250	CO_2 temperature sounding; N_2O column amount
2350 to 2420	CO_2 temperature sounding
2420 to 2700	surfaces cloud properties
2700 to 2760	CH_4 column amount

Table 2. *NE*Δ*T* Noise Characteristics of IASI at 280 K

ν, cm^{-1}	*NE*Δ*T*, K
650	0.419
700	0.157
750	0.145
800	0.145
850	0.150
900	0.150
950	0.165
1000	0.165
1050	0.176
1100	0.200
1150	0.200
1200	0.095
1250	0.096
1300	0.098
1350	0.100
1400	0.105
1450	0.105
1500	0.111
1550	0.116
1600	0.125
1650	0.137
1700	0.160
1750	0.170
1800	0.200
1850	0.224
1900	0.250
1950	0.240
2000	0.130
2050	0.135
2100	0.141
2150	0.151
2200	0.172
2250	0.200
2300	0.239
2350	0.287
2400	0.351
2450	0.400
2500	0.700
2550	0.900
2600	1.100
2650	1.300
2700	1.600
2750	1.935

[9] The IASI noise is presently simulated by a white Gaussian noise without error correlations on observations, this is a realistic assumption for an interferometer. However, radiances are assumed to be apodized, which introduces some correlation between adjacent channels. The *NE*Δ*T* at 280 K noise characteristics are given in Table 2 [*Cayla et al.*, 1995] (more recent results are from F. Cayla et al., personal communication, 1998). The *NE*Δ*T* at 280 K represents the standard deviation $st_{280}(\nu)$ of the Gaussian noise for a given wave number ν. At any other scene brightness temperature, Tb', the standard deviation, $st_{Tb'}(\nu)$, of the Gaussian noise is computed by:

$$st_{Tb'}(\nu) = \frac{\frac{\partial B(Tb=280,\nu)}{\partial Tb}}{\frac{\partial B(Tb=Tb',\nu)}{\partial Tb}} \cdot st_{280}(\nu), \qquad (1)$$

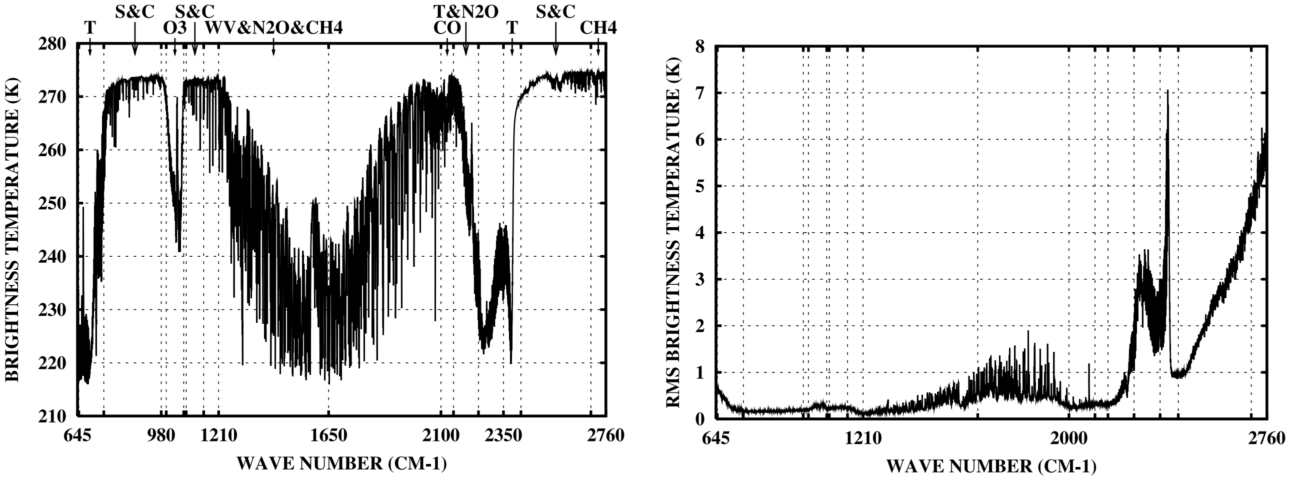


Figure 1. (left) Mean infrared atmospheric sounding interferometer instrument (IASI) spectrum and (right) corresponding standard deviation of IASI instrumental noise. Principal spectral absorption regions are indicated, as in Table 1.

which shows that the noise level increases when Tb' decreases. Figure 1 shows the IASI spectrum averaged over the whole TIGR data set (a data set of climatological situations that will be described in section 3.2) with the corresponding noise standard deviation spectrum. Figure 1 shows that some spectral regions could have a noise standard deviation higher than 2 K for a standard atmospheric situation.

2.2. Dimension Reduction

[10] Using directly (i.e., without preprocessing) all 8461 IASI channels in a retrieval algorithm is a simplistic strategy that would give poor results for practical and theoretical reasons. High-dimension data have to be reduced to limit the "curse of dimensionality" [*Bishop*, 1999]. The curse of dimensionality stipulates that, as the dimension M of the data space increases, the difficulty of a statistical regression procedure (representing the inverse radiative transfer equation), increases significantly and the number, E, of examples required for the regression increases exponentially with the dimension M. This is a general problem in regression analyses and is not particular to neural networks. The curse of dimensionality, however, may remain tractable because the intrinsic complexity of the function to be estimated, which is really the factor controlling the number of examples required, does not increase exponentially with the dimension.

[11] However, practical problems also occur. For example, the number of parameters in the regression model increases with M. Excess degrees of freedom in the regression, combined with the introduction of noninformative data (i.e., variability not linked to the desired output, like the instrumental noise or an inadequate vertical resolution), may distort the regression process: The quality criterion used to parameterized the inverse radiative transfer model becomes more complex so the global minimum is harder to estimate. Furthermore, computations are longer with such a large number of parameters.

[12] Thus the goal of dimension reduction is to present to the regression model the most relevant information from the initial rough data. One way to reduce dimension is feature selection: only a part of the observation is selected for the regression [*Jain and Zongker*, 1997]. An example of such an approach is channel selection [*Rodgers*, 1990]. For example Jacobian-based channel selection algorithms use the Jacobians of the radiative transfer model (RTM) to investigate the information content of the instrument channels in order to select the more pertinent ones [*Rabier et al.*, 2002; *Aires et al.*, 2002a]. This approach is particularly valuable when one is interested in some specific channels, for example in the case of trace gases retrieval.

[13] But the channel selection approach limits exploitation of channel redundancy for noise reduction. This is an important drawback for IASI since noise can be important in some spectral region, especially in the third spectral band of IASI (Figure 1). Another way to reduce the dimension of the data is by feature extraction, i.e., an operator acts on the entire observed IASI spectrum to extract its more pertinent characteristics. PCA is often used for this purpose: the dimension reduction is obtained by combining mutual information among the measured brightness temperatures. As explained in next section, a compromise needs to be made between reducing the data dimension and preserving the redundant information in the rough data.

3. Principal Component Analysis of IASI Spectra

[14] Although widely used for statistical analysis, the PCA technique is also very efficient for compression purposes [*Jolliffe*, 2002]. It is used here to compress and denoise the IASI observations. In the following, all IASI spectra are produced by a RTM (we use the RTIASI model here [*Matricardi and Saunders*, 1999]) computation applied to the TIGR data set since IASI does not exist yet.

3.1. Principal Component Analysis

[15] Let $\mathcal{D} = \{y^e; e = 1, \ldots, E\}$ be a database of E spectra, y, of dimension $M = 8461$. Let Σ be the $M \times M$ covariance matrix of the $\mathcal{D}$ database. Let V be the $M \times M$ matrix with columns equal to the eigenvectors of Σ and let L be the diagonal $M \times M$ matrix with the M associated eigenvalues in decreasing order (by definition $\Sigma \cdot V = V \cdot L$).

Table 3. Temperature Levels, Water Vapor and Ozone Layers for the IASI Retrieval Scheme

Layer or Level Number	Temperature Levels, hPa	Water Vapor Layers, hPa	Ozone Layers, hPa
1	0.1	0.1 to 167.9	0.1 to 0.6
2	0.2	167.9 to 253.7	0.6 to 2.6
3	0.6	253.7 to 358.2	2.6 to 20.4
4	1.4	358.2 to 478.5	20.4 to 45.2
5	2.6	478.5 to 610.6	45.2 to 69.9
6	4.4	610.6 to 795.0	69.9 to 102.0
7	6.9	795.0 to 1013.2	102.0 to 1013.2
8	10.3	0.1 to 1013.2	0.1 to 1013.2
9	14.8	...	...
10	20.4	...	...
11	27.2	...	...
12	35.5	...	...
13	45.2	...	...
14	56.7	...	...
15	69.9	...	...
16	85.1	...	...
17	102.0	...	...
18	122.0	...	...
19	143.8	...	...
20	167.9	...	...
21	194.3	...	...
22	222.9	...	...
23	253.7	...	...
24	286.6	...	...
25	321.5	...	...
26	358.2	...	...
27	396.8	...	...
28	436.9	...	...
29	478.5	...	...
30	521.4	...	...
31	565.5	...	...
32	610.6	...	...
33	656.4	...	...
34	702.7	...	...
35	749.1	...	...
36	(795.0+839.9)/2.	...	...
37	(882.8+922.4)/2.	...	...
38	(957.4+985.8)/2.	...	...
39	(1005.4+1013.2)/2.	...	...

[16] We define the $M \times M$ filter matrix $F = L^{-1/2} \cdot V^t$. The matrix F is used to project IASI spectra, y, onto a new orthonormal base composed by the columns of F: $\{F_{\star i}; i = 1, \ldots, M\}$:

$$\begin{cases} h = F \cdot y = F_{1\star} \cdot y_1 + \ldots + F_{M\star} \cdot y_M \\ y = F^{-1} \cdot h = F^t \cdot h = h_1 \cdot F_{\star 1} + \ldots + h_M \cdot F_{\star M}, \end{cases} \quad (2)$$

where ^{t} is the transpose operator. The vectors $\{F_{i\star}; i = 1, \ldots, M\}$, i.e., rows of F, are called the filters and the normalized eigenvectors $\{F_{\star\ i}; i = 1, \ldots, M\}$ are called the PCA basis functions. Because these eigenvectors are an orthogonal basis set for representing the IASI spectra, y, we will refer to them as eigenspectra. By definition, the coordinates of the new data h are uncorrelated since:

$$\langle h \cdot h^t \rangle = \langle F \cdot y \cdot y^t \cdot F^t \rangle = \langle F \cdot \Sigma \cdot F^t \rangle = I_{M \times M}, \quad (3)$$

where $\langle \cdot \rangle$ represents the mathematical expectation.

[17] Practically, the first step in a PCA approach is to compute the 8461 × 8461 covariance matrix $\Sigma = \langle (y - \langle y \rangle) \cdot (y - \langle y \rangle)^t \rangle$ of the database, where y is an IASI spectrum composed of the 8461 wavelengths. The eigenvalue matrix L and the corresponding eigenvectors V of this covariance matrix Σ are then computed using a Cholesky or a singular value decomposition (SVD).

3.2. Data Set

[18] Our data set is composed of a large number of real atmospheric situations measured by radiosondes, taken from the TIGR database [*Chédin et al.*, 1985; *Achard*, 1991; *Escobar*, 1993; *Chevallier et al.*, 1998, 2000]. We use the TIGR3 database composed of 2311 atmospheres: 872 in

Table 4. Cumulated Explained Variance Percentage of IASI Spectra With Respect to the Number of PCA Components

Number of PCA Components Used	Cumulated Explained Variance, %
1	91.97
2	95.08
3	97.74
4	98.44
5	99.03
6	99.42
7	99.61
8	99.75
9	99.82
10	99.85

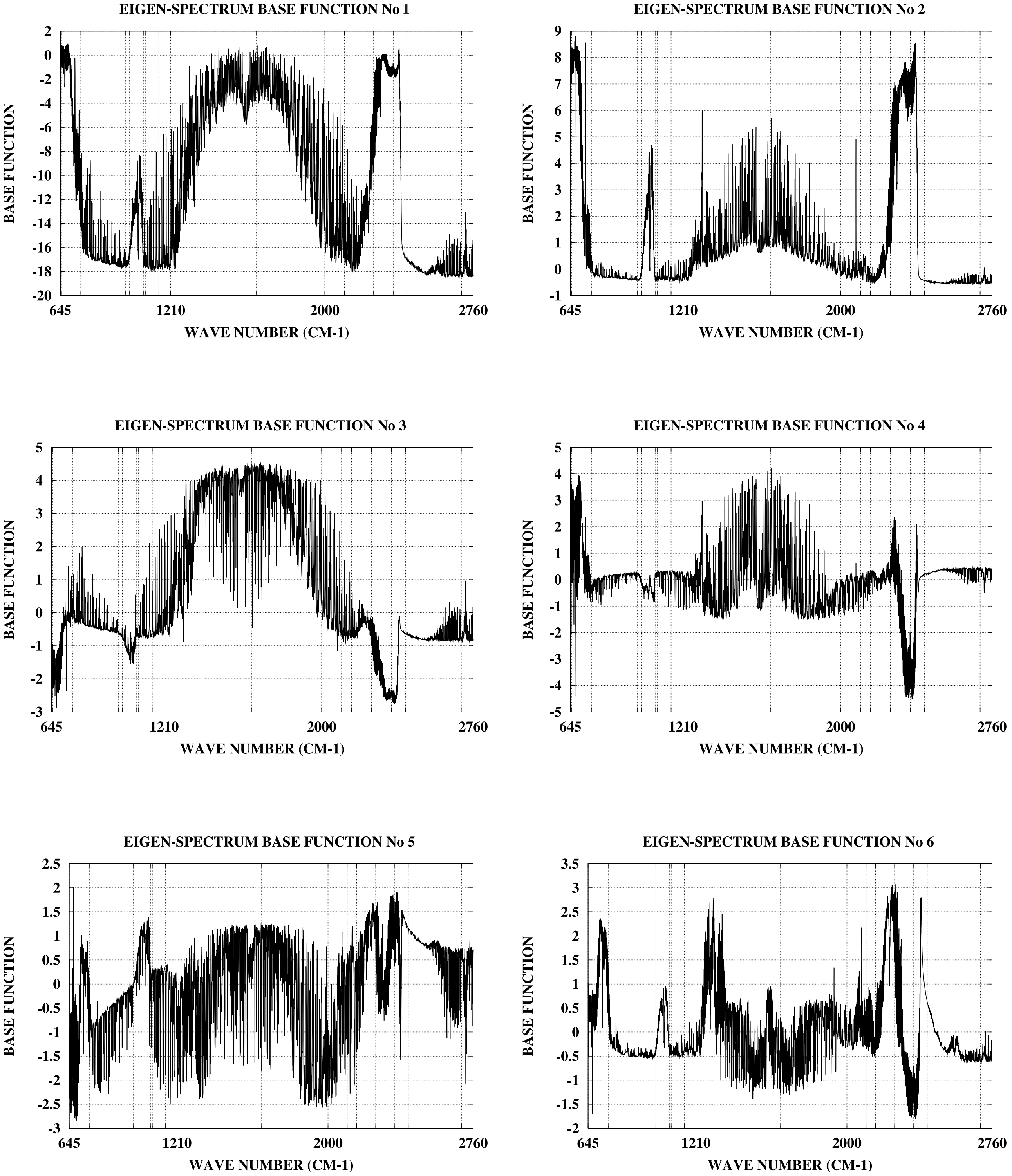


Figure 2. First 10 infrared atmospheric sounding interferometer instrument eigenspectrum basis functions.

tropical air-mass, 388 in midlatitude type 1, 354 in midlatitude type 2, 104 in polar type 1 and 593 in polar type 2. These atmospheres are described by their temperature and gas concentration profiles. For the retrieval scheme, we have used the discretization described in Table 3. The discretization in temperature is the same as the one used by RTIASI, except for the 3 near-surface levels (37, 38, and 39) that are each the combination of two European Centre for Medium-Range Weather Forecasts (ECMWF) levels which we consider too thin for IASI (the first ECMWF level is at 60 m). For water vapor, we take layers of about 2 km, which follows the recommendation for IASI. The ozone discretization is not regular; it emphasizes the layers near 30 hPa where the ozone abundance is a maximum. The water vapor and ozone

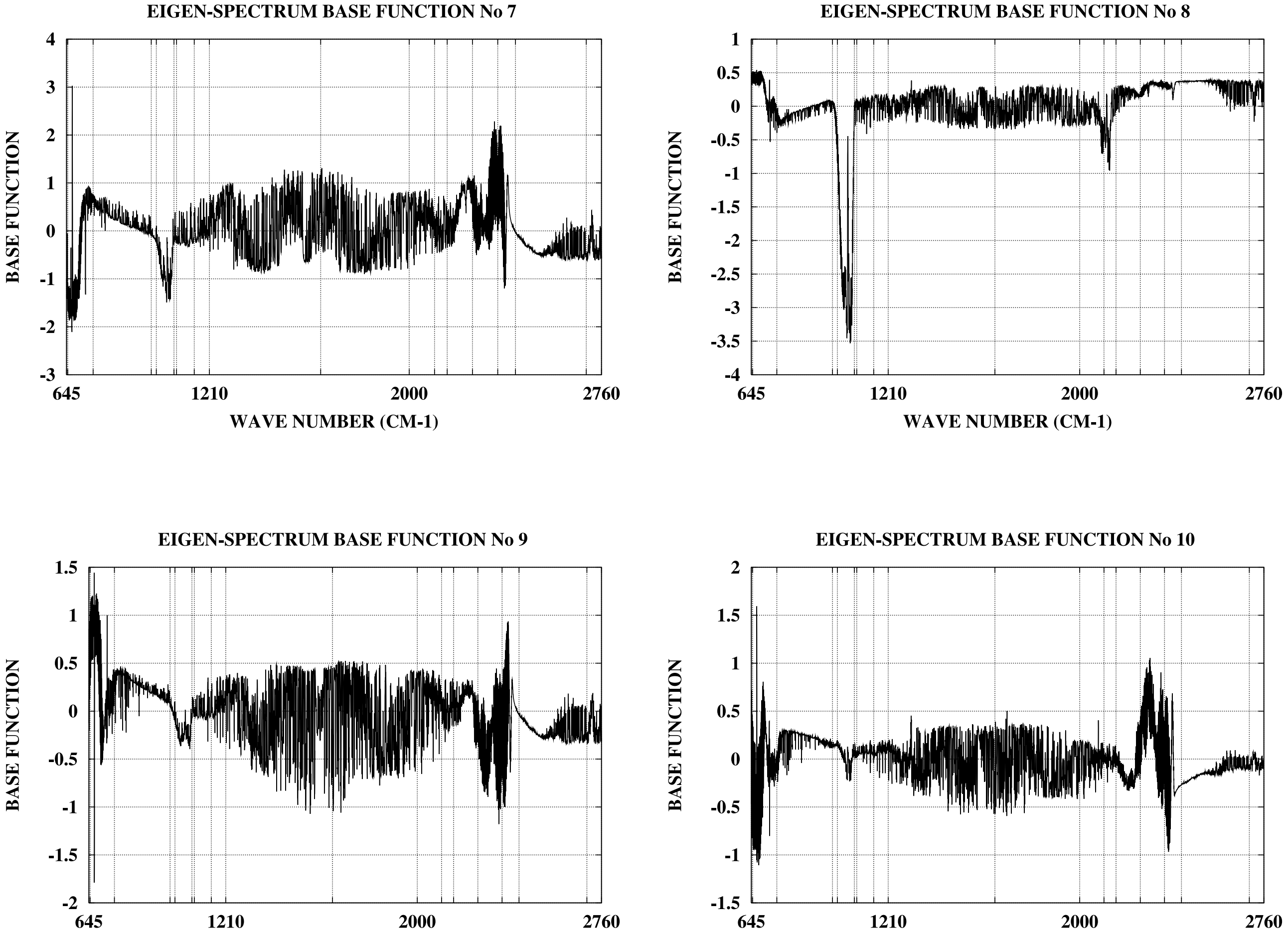


Figure 2. (continued)

discretizations are kept as subdiscretizations of the ECMWF scheme.

[19] The TIGR atmospheres have been selected from a collection of more than 150,000 radiosonde measurements; they are very irregular profiles because they come from radiosonde measurements. The data set represents, as much as possible, all kinds of possible atmospheric situations. This database has been designed for a pattern recognition (i.e., a climatological first-guess retrieval) process, so the same emphasis is uniformly put on rare and frequent atmospheric situations [*Chevallier et al.*, 1998]. So not only is the range of variability occasionally extreme (see *Aires et al.* [2002a] for a description of the variability envelop of TIGR), but also the occurrence and strength of inversions in the profiles introduces complicated structures that are very challenging to any retrieval method. This complexity represents, by design, more extreme conditions, and less smooth behavior than that encountered under operational conditions where model output is used as the first guess, so our estimate of the retrieval errors could be an overestimate. However, the use of a large and complex climatological data set allows the inversion model to be calibrated globally including rare events.

[20] The ozone variability representation is not sufficient in this version of TIGR. So, it is expected that in our results, the retrieval error for ozone will probably be an underestimate of the correct error level for IASI. A new data set is presently being developed to improve the ozone representation.

[21] The RTIASI forward radiative transfer algorithm developed at ECMWF [*Matricardi and Saunders*, 1999] has been used to compute the IASI brightness temperatures associated with the TIGR atmospheres for clear conditions over the sea. RTIASI provides also the Jacobians of the RTE [*Sherlock*, 2000] but we don't use this information in this study. For each of the 2311 atmospheres of TIGR, we have simulated 5 noise realizations using the specifications for the instrument, Table 2 and equation (1). Our data set is then composed of 11,555 examples.

3.3. Analysis of the Eigenspectrum Basis Functions

[22] PCA is done on the previously described data set in noise-free condition. In that way, the structures (i.e., the eigenspectra) found by the analysis represent actual physical variability, not perturbated by the noise.

[23] In Table 4, the cumulated percentage of explained variance is indicated as a function of the number of components. We see that the 99% level is attained with only 10 components. PCA uses optimally the redundant information existing in the IASI channels by adaptatively determining the principal components h_i as a weighted sum of partially redundant channels: $h_i = \sum_{j=1}^{M} F_{ij} \cdot y_j, \forall i = 1, \ldots, N$. The terms h_i can be seen as "meta channels" that have been adaptatively

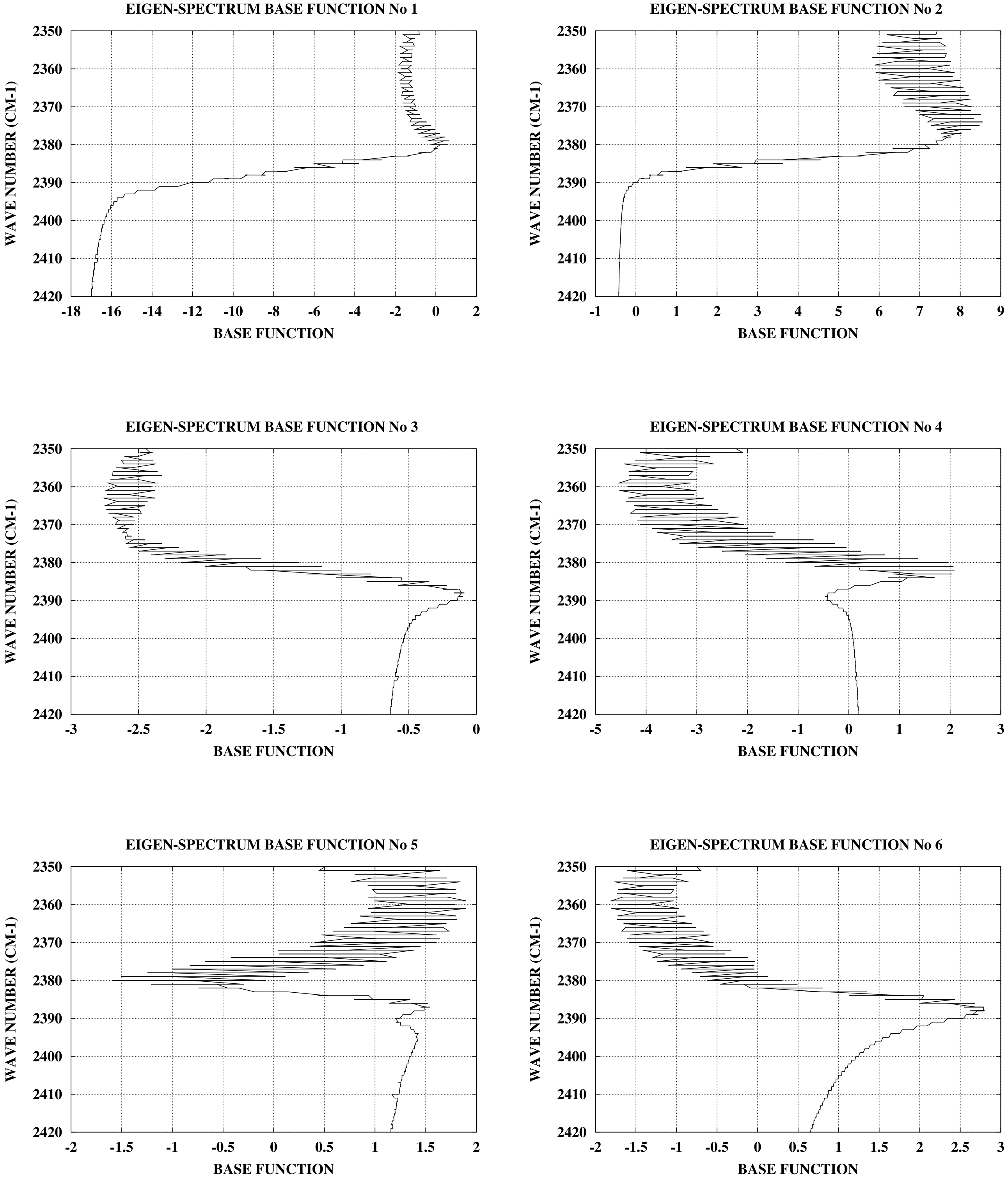


Figure 3. Interpretation of infrared atmospheric sounding interferometer instrument eigenspectra for temperature, in the 2350–2420 cm^{-1} spectral region.

(in the statistical sense) determined using the $\mathcal{D}$ data set of examples.

[24] The first ten eigenspectra of Σ, i.e., the basis functions $F_{\star i}$ ($F = L^{-1/2} \cdot V^t$), are shown in Figure 2. Each one gives particular information on the statistical dependence among the selected channels. For example, the eigenspectrum basis function 1 describes the average deviation from the mean spectrum. Its shape is the same as the mean spectrum in Figure 1 but inverted. We recognize all the absorption features described in Table 1: temperature, water vapor, ozone, *CO*, etc.

[25] The eigenspectrum basis function number 2 is related more to temperature (650–770 and 2150–2420 cm^{-1}) and ozone (1000 to 1070 cm^{-1}), less to water vapor.

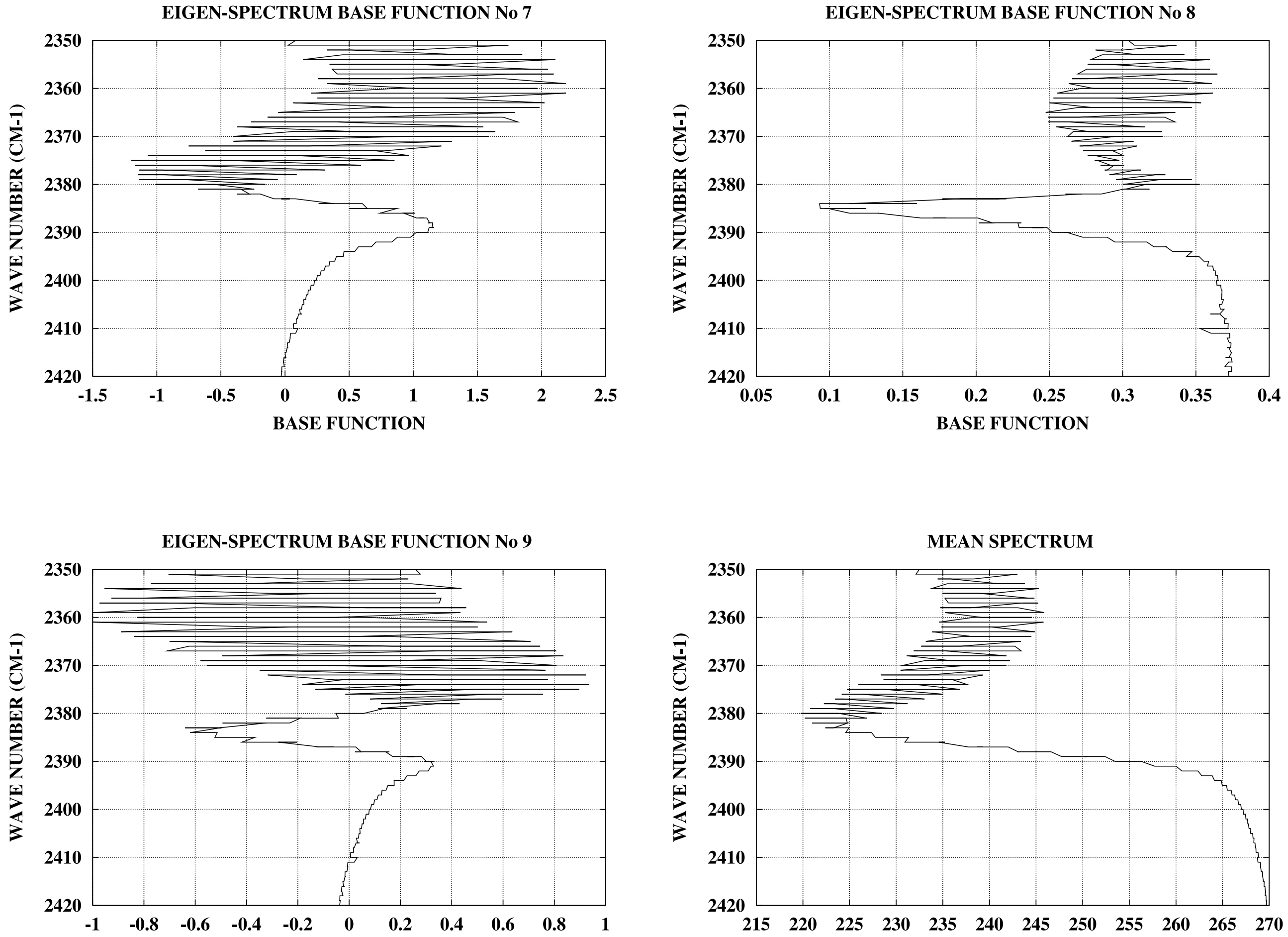


Figure 3. (continued)

The eigenspectrum basis function 3 is the opposite: it gives an information related more to the 1210–2000 cm^{-1} spectral region of the water vapor. Some of higher number eigenspectra look more localized in wavelength. For example, eigenspectrum basis function number 8 isolates 1000 to 1070 cm^{-1} (ozone) and 2100 to 2150 cm^{-1} (*CO*).

[26] The 2350–2420 cm^{-1} spectral region is usually dedicated to CO_2 temperature sounding. We can use the relatively "direct" link between atmospheric temperature and brightness temperature in this region to understand the behavior of the eigenspectra. In Figure 3, the pieces of the first 9 IASI eigenspectra in that spectral region are presented: the basis function value is on the abscissa, and the wave number is on the ordinate. The mean spectrum of TIGR in this region is also shown (bottom right). Smaller wave numbers sound upper-atmosphere temperature, and larger wave number sound the near-surface temperature. The lower-order eigenspectra are smoother than the following ones and have a regular monotonic profile shape. For example, we see that the first eigenspectrum is similar (but inverted) to the mean spectrum: it is then a good basis function to represent the mean spectrum, i.e., regular smooth information. The higher order eigenspectra have more pronounced inversion(s) at different "altitudes". These basis functions are used by the PCA to express the different atmospheric profiles sharpes as deviations from the average shape, with an increasing amount of detail as the number of components used increases.

[27] The interested reader can get a more detailed analysis of the eigenspectra from *Huang and Antonelli* [2001], where the correlation between PCA-compressed infrared high resolution spectra and geophysical parameters, like the temperature, have been analyzed to characterize the significance of the PCA components.

4. Use of IASI Eigenspectra

4.1. Compression of IASI Spectra

[28] Let $\bar{F}$ be the $N \times M$ truncated matrix of F. The PCA decomposition uses this truncated matrix to project IASI spectra, y, of dimension M = 8461 into a space of lower dimension N (with $N \leq M$):

[29] $h = \bar{F} \cdot y$ and $\hat{y} = \bar{F}^{-1} \cdot h$ (same as equation (2) but with N instead of M and where $\bar{F}^{-1}$ is a generalized inverse matrix since $\bar{F}$ is not square). The compression error $\|y - \hat{y}\|$ is given by $\|h_{N+1} \cdot F_{\star N+1} + \ldots + h_M \cdot F_{\star M}\|$, where $\| \cdot \|$ is the Euclidean norm. PCA is optimum for the least squares errors criterion $\frac{1}{E}\sum_{e=1}^{E} \|y^e - \hat{y}^e\|^2$ [*Jolliffe*, 2002].

[30] For the compression, we only retain the first N filters, but a compromise needs to be found between a good compression level and a small compression error. The more components used for compression, the lower the compres-

sion error is. With only N = 50 (the 50 first principal components), the RMS compression error of the IASI spectra averaged over the whole TIGR data set is close to 0.05 K, which is much lower than the average IASI noise which is close to 1 K.

[31] Figure 4 shows the spectral distribution of the compression errors. The more eigenspectra used for the compression, the lower the compression error. Taking 10 components is not enough, but with 50 components, the level of error becomes very reasonable.

[32] A global PCA uses the same covariance (or dependency) structure, whatever the air mass, but this structure can also be made to vary with the air-mass. However, the compression level that we obtain is robust: using atmospheres from different air masses gives equivalent compression errors. Even if the shape of components are variable in the different air masses, the general eigenspectra decomposition seems to be able to catch the variabilities describing temperature, water vapor or ozone. So a specialized PCA could be more precise (less components needed for compressing the IASI spectra) but this is not the real issue at this stage of the study. We will investigate this point in the future.

[33] Another possibility is to perform the PCA across more than one instrument. The resulting eigenspectra can then represent the same information (like the temperature profile) in the different instruments. The same geophysical variability can then be described in parallel by different behavior in the different wavelengths of instruments channels. This point will be studied in the future for the use in parallel of IASI (infrared) and AMSU (microwave).

4.2. Denoising of IASI Spectra

[34] There is a possibility to suppress part of the noise during the compression process. It is assumed that the lower-order principal components ($h_1,\ldots,h_N$) of a PCA decomposition describe the real variability of the observations, or the signal, (here the IASI spectra) and that the remaining principal components ($h_{N+1},\ldots,h_M$) describe higher frequency variabilities in the IASI spectrum. These higher frequencies are more likely to be related to the white Gaussian noise of the instrument, or to be related to the variability of minor atmospheric constituents. Because we are interested in only major constituents like temperature, water vapor and ozone, we can consider the higher order components to describe noise (instrumental plus noninformative information). So, the PCA representation of the spectra could advantageously be used for denoising.

[35] In practice, a PCA is first performed on no-noise spectra in order that the resulting eigenspectra contain only signal information and are not used to describe noise. In the operational stage, observed spectra, y, are projected into the regular subspace of the first components, describing the real variability of IASI spectra (we will comment on how to choose N in the following). In the resulting compression h, the variability attributed to the instrumental noise is then partially suppressed. The compression h can be directly used in a retrieval scheme, or it can be uncompressed to obtain $\hat{y}$, the spectrum partially denoised.

[36] In Figure 5, the denoising error (compressed and then uncompressed noisy spectrum minus no-noise spectrum) is shown with respect to the number of PCA components used for the compression. After a decrease of the error with increasing PCA number due to a better compression, the denoising error increases. This increase of the denoising error for an increase of number of components results from a more accurate representation of the noise. Asymptotically, the compression error should converge to zero (perfect representation of the noisy IASI spectra), but the denoising error converges to the instrument noise (perfect reconstruction of the noisy spectra). The

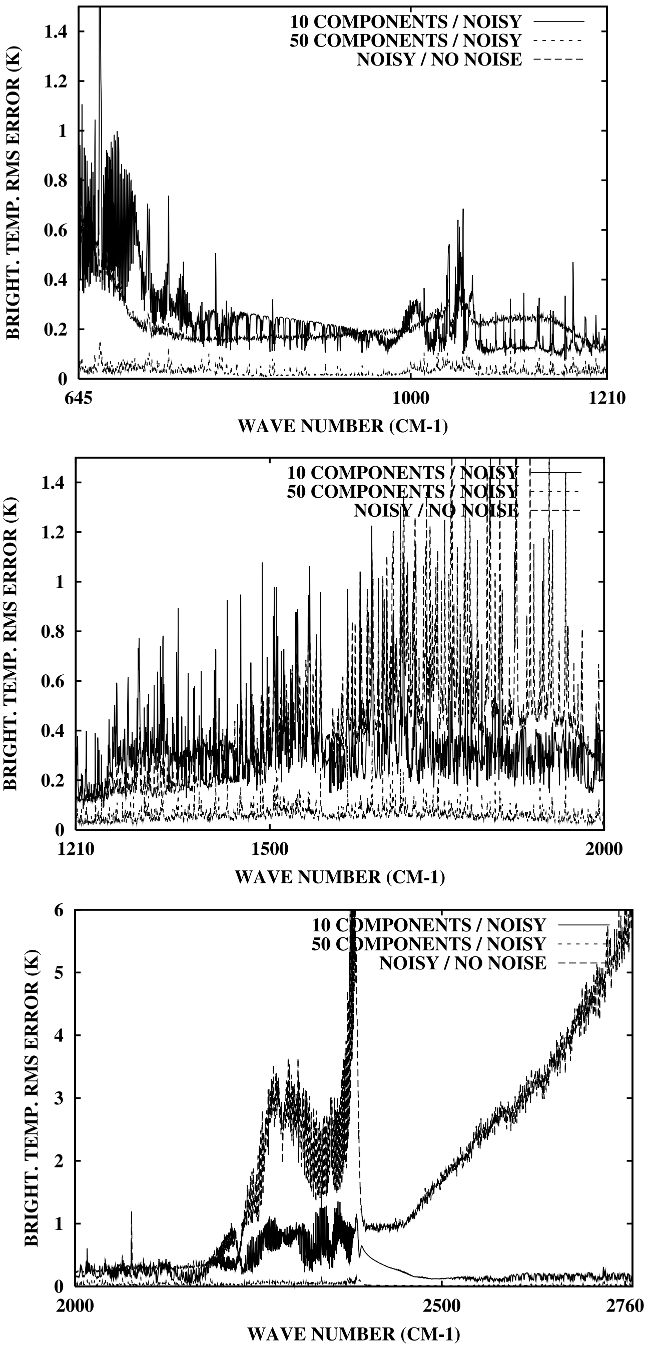


Figure 4. Statistics of compression in the 3 spectral bands of infrared atmospheric sounding interferometer instrument for 10 components (top line) and 50 components (bottom line), instrumental noise standard deviation is shaded for comparison purpose. See color version of this figure at back of this issue.

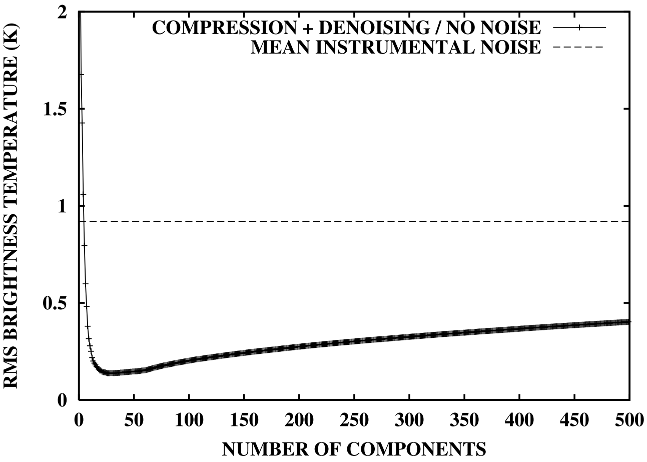


Figure 5. Denoising error (solid line) and overall instrumental noise (dashed line), with respect to the number of principal component analysis components used in the compression.

optimum is given for N = 30 components. This number depends not only on the spectral characteristics of the IASI observations, but also on the noise level, and on the data set (here the TIGR database) used to perform the PCA and the resulting statistics.

[37] Figure 6 shows the spectral statistics of the denoising errors on the TIGR atmospheric database. Using only 10 components is not enough. In the first and second spectral bands, the denoising error is still often larger than the instrumental noise. However, it is shown that the third band is already considerably denoised (0.2 K of RMS instead of more than 2 K!). The use of 30 components for the compression/denoising has excellent statistics: Denoising statistics for this compromise is the lowest point of the curve in Figure 5. An error analysis (not shown) indicates that 30 components is the best compromise between the compression error, requiring a large number of components, and a denoising error, requiring the limitation of the number of components used so as to avoid representing the noise.

[38] This compromise is good, of course, only in a statistical sense. Actually, it is interesting to note that with 200 components, some spectral regions are represented with an equivalent, or even better, denoising level than with 30-components (see for example 1500−1800 cm^{-1}). But on average over the whole spectra, the denoising errors are higher because more noise has been represented by the additional components (see first spectral band). If a spectral region is of particular interest (because of a particular constituent absorption), it is important to note that the denoising of the entire spectrum is not necessarily the optimal solution. The particular spectral region of interest may be neglected in a statistical point of view with respect to the other channels: the compression/denoising scheme will not well represent this information. Control of errors for each spectral region is crucial if such spectral regions are of particular interest. Then, even if 30 components seems to be the perfect compromise for compression/denoising of the whole spectrum, it might be useful to use higher order components for other purpose. Particular cases could use a combination of approaches. This is especially true when the regression scheme used is able to extract nonlinear information from the components. If we are interested in very localized channels, that display complex behavior (nonlinear with respect to the amount of the absorbing constituent, unstable, etc.), a PCA, even with a high number of components, will not be ideal: it probably will use too many components to describe this complex behavior. An alternative would be to use, in that particular case, the raw

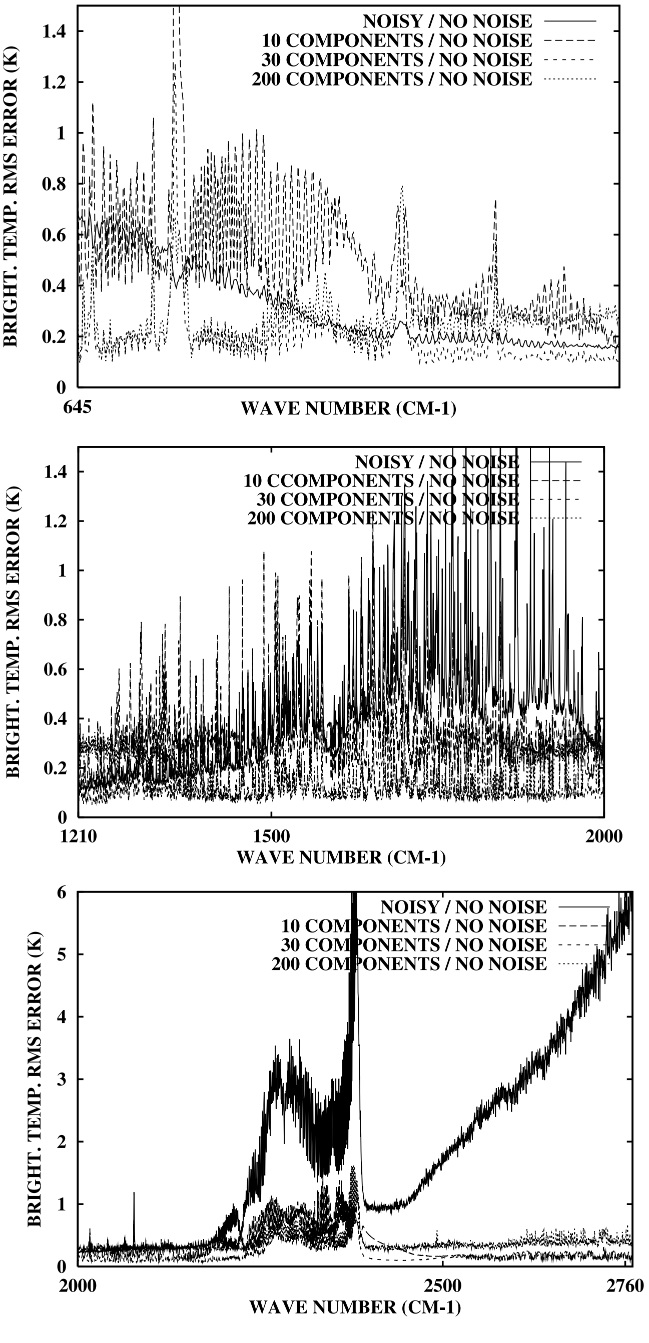


Figure 6. Statistics of denoising errors in the three spectral bands of infrared atmospheric sounding interferometer instrument using 10, 30 and 200 principal component analysis components, for the Thermodynamic Initial Guess Retrieval database situations, instrument noise (red line) is shown for comparison purpose. See color version of this figure at back of this issue.

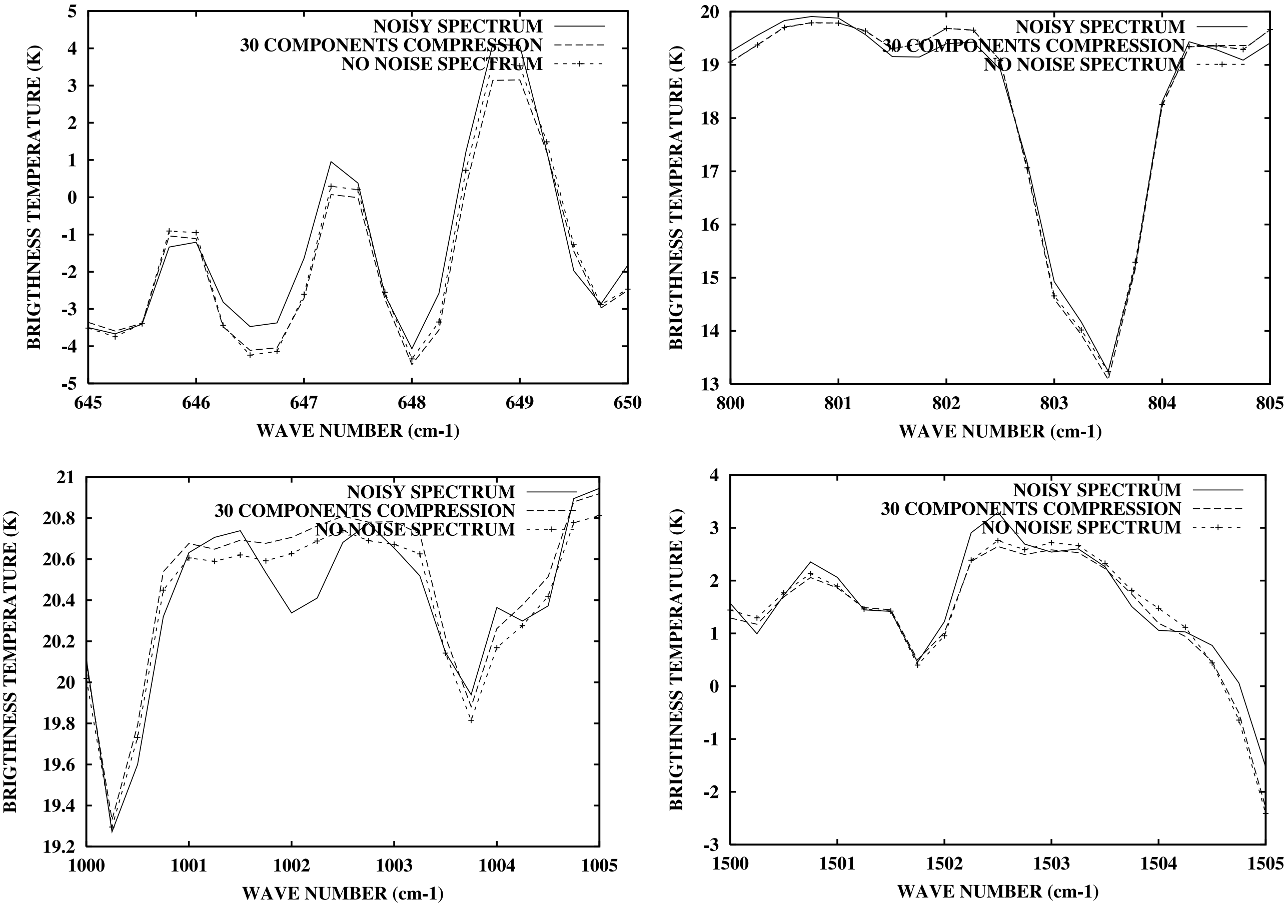


Figure 7. Comparison of one noise-free spectrum (dotted line with points), the same spectrum with noise (solid line), and the corresponding denoised spectrum using 30 principal component analysis components (dashed line).

specific channels. This would be the case for example for trace gas retrievals.

[39] In the study of *Huang and Antonelli* [2001], 150 PCA components is found to be the best compromise for the denoising process of high-resolution infrared measurements. The fact that we find that only 30 components are sufficient in the IASI case can be explained by the differences in the specifications of the instruments in the two studies. The spectral resolution is not the same (0.6 cm^{-1} compared to 0.25 cm^{-1} for IASI), the radiative transfer model is different, the climatological data sets to estimate the statistics are also different, and the number of channels is only 3888 (8461 for IASI). But the fact that IASI has a high instrument noise level in some spectral regions explains most of the difference in the two optimum compression levels: with a higher noise level, using more channels would only serve to better represent the noise variability. This is what we observe in practice, if we use more than 30 components, the denoising error starts to increase. It is possible that a PCA performed only on some less noisy spectral regions of a IASI spectrum would be better denoised using more than 30 components. This localized PCA on smaller IASI spectral regions will be the subject of a future work.

[40] In Figure 7, some specific spectral regions of interest are represented to illustrate the compression and denoising properties for one atmosphere. We see how our scheme is able to retrieve the signal part (i.e., no-noise spectrum) in a noisy observation (see for example the 645–650 cm^{-1} spectral region). This is particularly true for high noise-level spectral regions like 2495–2500 cm^{-1} where the scheme has used the information of flat spectrum to avoid the oscillations due to the instrument noise.

4.3. PCA-Based Pattern Recognition

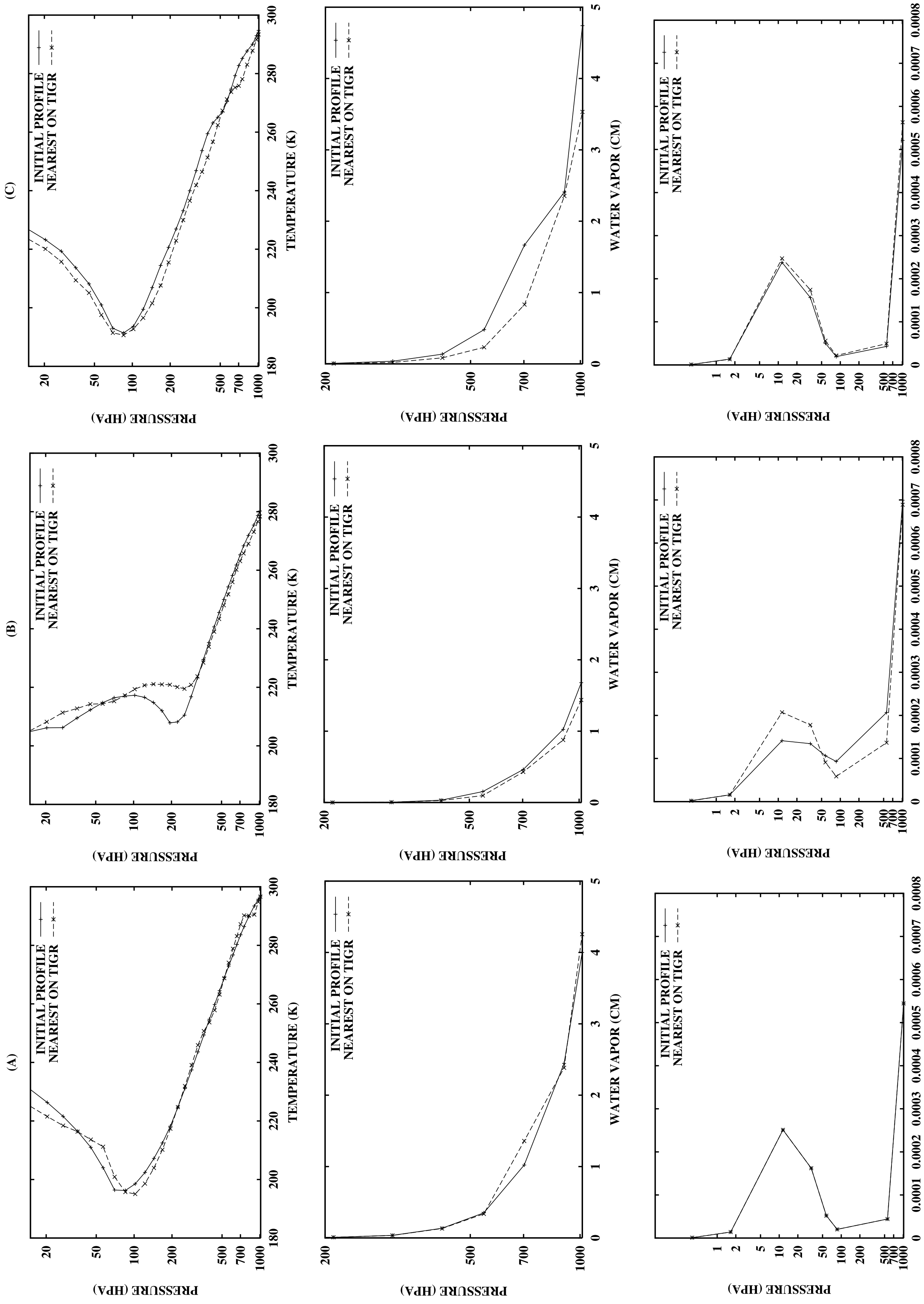
[41] For many ill-conditioned problems, the use of a first-guess estimate is very important to regularize the inversion process. In the Improved Initialization Inversion (3I) retrieval scheme [*Chédin et al.*, 1985; *Scott and Chédin*, 1981], the initial guess is found in the TIGR climate database. In a variational assimilation context, more focused on meteorology than on climatology, the first-guess solution is the 6-hour forecast; see *Prunet et al.* [1998] for an example in the IASI context.

[42] To retrieve such a first-guess efficiently from such a data set, often the Euclidean distance between observations y^0 and a spectrum from the data set, y, is minimized:

$$D_E(y^0, y) = (y^0 - y)^t \cdot (y^0 - y).$$

Another possibility is to use the Mahalanobis distance:

$$D_M(y^0, y) = (y^0 - y)^t \Sigma^{-1} (y^0 - y).$$


(A)
(B)
(C)
INITIAL PROFILE
NEAREST ON TIGR
PRESSURE (HPA)
TEMPERATURE (K)
180
200
220
240
260
280
300
20
50
100
200
500
700
1000
WATER VAPOR (CM)
0
1
2
3
4
5
OZONE (CM)
0
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
1
2
5
10
20

The Euclidean distance treats all variables in the same way where the Mahalanobis distance gives less weight to variables with high variance and groups highly correlated variables.

[43] We propose here to use an Euclidean distance based on the first *N* PCA components, *h*. This distance would be equivalent to the Mahalanobis distance if we used all the PCA components ($N = M$) [*Jolliffe*, 2002]. Using fewer components removes irrelevant information and produces a faster pattern recognition step (from a distance calculation with $M = 8461$ channels to a distance calculation with $N = 30$ components). This distance is then used to perform a pattern recognition in the climatological data set: for each observation y^0, the first guess is determined to be the atmospheric situation from the climatological data set *y* that has the minimum distance $D_E(h^0, h)$.

[44] Examples of pattern recognition of one TIGR atmosphere from the remaining TIGR atmospheres are presented in Figure 8, and RMS differences between the first-guess and real profiles are given for temperature, water vapor, and ozone in Figure 9. We note that the first guess for temperature is not very good (about 4/5 K of RMS error) but this can be explained by two factors: first, the pattern recognition for one TIGR situation is made into the TIGR data set; this is not optimal since TIGR has been designed such as a minimal distance exist between each situations of TIGR. This can reduce by a factor of two the sampling properties of the data set. Second, the pattern recognition is made for the whole spectrum, each constituent of the atmosphere is then taken into account, and the first guess has to be a compromise between each of the variables, temperature, water vapor, ozone, etc., instead of temperature only. It is normal for the first-guess error in temperature to increase with altitude since IASI has less and less information for high-level layers. A good first guess for water vapor is also difficult to obtain, the error is between 32 and 75%, but this can be due to the fact that IASI has little or no water vapor information for higher atmospheric layers [*Aires*, 1999]. The first-guess error of the total content for water vapor is about 32.5%. For ozone, the first guess is of good quality, between 10 and 25%, but this is due to an insufficient representation of the ozone in this version of the TIGR data set used in this study. This point is discussed in section 5. The first-guess error for ozone total content is about 10%.

[45] A new TIGR data set is being developed with, especially, an enrichment of the ozone variability. This should considerably improve the results of the first-guess retrieval. This first-guess retrieval however is already sufficient to improve the retrieval scheme [see *Aires et al.*, 2002c] because of the additional constraint it puts on the inverse problem.

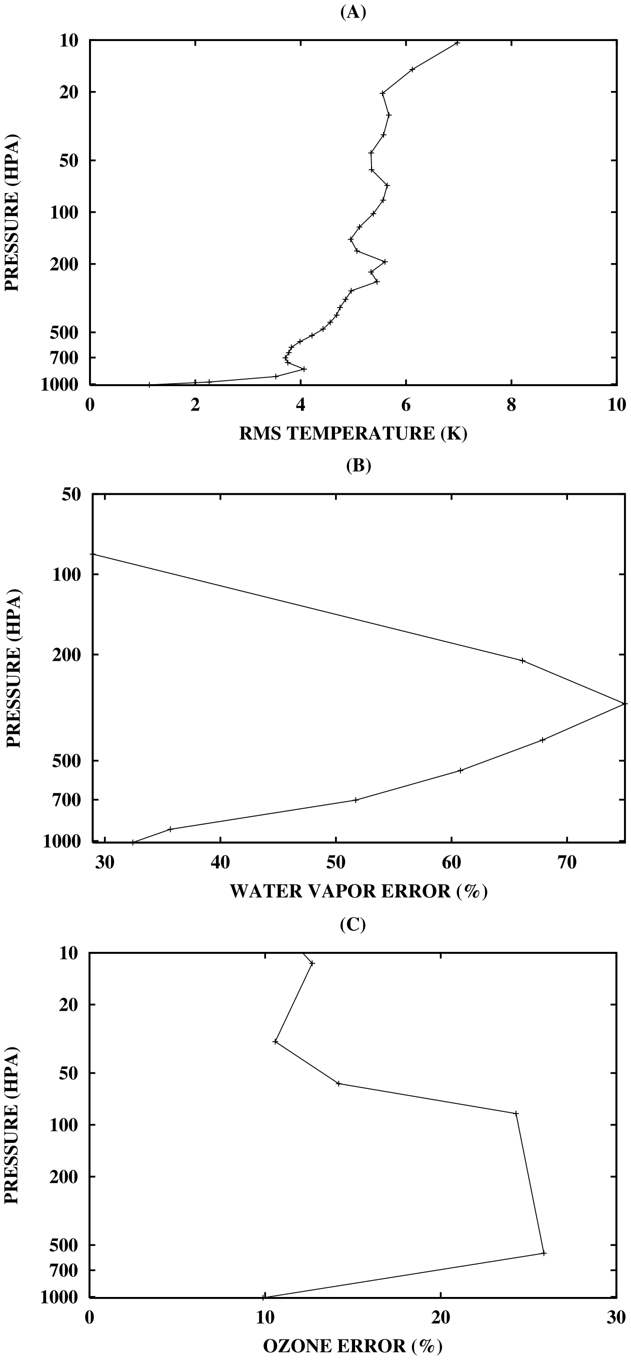


Figure 9. RMS error of the first guess for (a) temperature, (b) water vapor, and (c) ozone. Near-surface values for water vapor and ozone represent the total vertical content.

5. Conclusion and Perspectives

[46] We have developed a PCA-based method for compressing, denoising, and first-guess retrieval for the high-resolution interferometer IASI. Our approach allows for a more complete exploitation of the information in the IASI spectra. The compression step allows reduction of the dimension of the data used. The denoising process, using the redundancy information among channels, reduces considerably the instrument noise in IASI observations. The instrumental noise in the overall IASI observed spectrum

Figure 8. (opposite) First-guess retrieval examples: column A, tropical; column B, temperate; and column C, polar situations. Near-surface values for water vapor and ozone represent the total vertical content.

goes from 0.9K to 0.2K after denoising. The PCA representation of IASI spectra also allows for a fast and multivariate first-guess retrieval.

[47] These preprocessing steps (compression, denoising, first-guess retrieval) are an important first step for any retrieval algorithm because they provide a more strongly constrained inverse problem. For example, the simultaneous compression of many variables is a crucial point, since it allows us to exploit the complex inter-dependencies among the observations and between observations and variables.

[48] Our experiments were made with the TIGR database, i.e., a vast and complex set of real atmospheric situations (from radiosonde measurements which are much more irregular than model output) with rare events. This fact provides a global applicability of our method.

[49] This approach is completely general and does not depend on the retrieval method that is then employed. For example, our compression/denoising approach could be used in a variational assimilation scheme: the dimension of the data is smaller, noise is reduced, and the variables are decorrelated. This would simplify calculations and speed the scheme (a better approach would be to develop a numerical model in the space of the principal components). This PCA-based representation of IASI observations will be used by *Aires et al.* [2002c] to retrieve simultaneously the atmospheric temperature, water vapor and ozone profiles.

[50] A more optimal denoising approach would be to perform a PCA for each type of air mass. In effect, by using a global PCA, the same statistical structure of dependencies is used for each air-mass, which can be nonoptimal. A specialized PCA is expected to even better describe the natural variability of IASI observations. A new TIGR data set is under development where the ozone variability is improved.

[51] Use of more advanced statistical techniques for the compression and denoising of IASI spectra can be investigated. We are not particularly interested in nonlinear extraction techniques [*Karhunen and Joutsensalo*, 1994; *Monahan*, 2000] since the neural network retrieval method is already nonlinear. But the use of a technique such as the Independent Component Analysis (ICA) would be interesting: this new component extraction technique has been used in various fields including climatology [*Aires et al.*, 2000] and image processing [*Nadal et al.*, 2000]. *Aires et al.* [2002b] have shown that the ICA, which extracts statistically independent components instead of the decorrelated components of PCA (i.e., a weaker constraint), can provide a better way than PCA for extracting meaningful components. The PCA has the tendency of mixing many physical modes, and the use of the ICA to retrieve purer "physical" modes might improve any retrieval scheme.

Notation

ν wave number of an infrared atmospheric sounding interferometer (IASI) channel.
B Plank function.
TB brightness temperature.
st standard deviation of instrument noise for one channel.
y^o IASI brightness temperature spectrum observations.
M dimension of the IASI spectrum y.
E number of samples in the data set.
h principal component analysis (PCA) compression of the IASI spectrum y.
N dimension of the compression h $(N \leq M)$.
Σ $M \times M$ covariance matrix of spectra y.
V $M \times M$ matrix of eigenvectors of Σ.
L $m \times M$ diagonal matrix of eigenvalues of Σ.
F $M \times M$ filter matrix.
$\langle \cdot \rangle$ expectation operator.
$\mathcal{D}$ data set of examples for the PCA analysis.
D_E Euclidean distance.
D_M Mahalanobis distance.

[52] **Acknowledgments.** We would like to thank the European Centre for Medium-Range Weather Forecasts for providing the direct radiative transfer model RTIASI. This work was partly supported by special NASA Climate and Radiation Program funding provided by Robert J. Curran and now by Donald Anderson. The authors thank three anonymous referees for their fruitful comments and suggestions.

References

Achard, V., Trois problèmes clés de l'analyse tridimensionelle de la structure thermodynamique de l'atmosphère par satellite: Mesure du contenu en ozone, classification des masses d'air, modélisation hyper-rapide du transfert radiatif, Ph.D. thesis, Univ. Pierre et Marie Curie (Paris VI), Paris, 1991.

Aires, F., Problèmes inverses et réseaux de neurones: Application à l'interféromètre haute résolution IASI et à l'analyse de séries temporelles, Ph.D. thesis, Univ. Paris IX/Dauphine, Paris, 1999.

Aires, F., R. Armante, A. Chédin, and N. A. Scott, Surface and atmospheric temperature retrieval with the high resolution interferometer IASI, *Proc. Am. Meteorol. Soc.*, *98*, 181–186, 1998.

Aires, F., A. Chédin, and J.-P. Nadal, Independent component analysis of multivariate times series: Application to the tropical SST variability, *J. Geophys. Res.*, *105*(D13), 17,437–17,455, 2000.

Aires, F., A. Chédin, N. A. Scott, and W. B. Rossow, A regularized neural net approach for retrieval of atmospheric and surface temperatures with the IASI instrument, *J. Appl. Meteorol.*, *41*, 144–159, 2002a.

Aires, F., W. B. Rossow, and A. Chédin, Rotation of EOFs by the independent component analysis: Towards a solution of the mixing problem in the decomposition of geophysical time series, *J. Atmos. Sci.*, *59*(1), 111–123, 2002b.

Aires, F., W. B. Rossow, N. Scott, and A. Chédin, Remote sensing from the infrared atmospheric sounding interferometer instrument, 2, Simultaneous retrieval of temperature, water vapor, and ozone atmospheric profiles, *J. Geophys. Res.*, *107*(DX), 10.1029/2001JD0001591, in press, 2002c.

Bishop, C. M., *Neural Networks for Pattern Recognition*, 482 pp., Clarendon, Oxford, UK, 1999.

Cayla, F., B. Tournier, and P. Hebert, Performance budgets of IASI options, *Tech. Rep. IA-TN-0B-5476-CNE*, Cent. Natl. d'Etudes Spat., Toulouse, France, 1995.

Chédin, A., N. A. Scott, C. Wahiche, and P. Moulinier, The improved initialization inversion method: A high resolution physical method for temperature retrievals from Tiros-N series, *J. Clim. Appl. Meteorol.*, *24*, 128–143, 1985.

Chevallier, F., F. Chéruy, N. A. Scott, and A. Chédin, Neural network approach for a fast and accurate computation of the longwave radiation budget, *J. Appl. Meteorol.*, *37*, 1385–1397, 1998.

Chevallier, F., J.-J. Morcrette, F. Chéruy, and N. A. Scott, Use of a neural network-based longwave radiative transfer scheme in the ECMWF atmospheric model, *Q. J. R. Meteorol. Soc.*, *126*, 761–776, 2000.

Escobar, J., Base de données pour la restitution de paramètres atmosphériques à l'échelle globale; étude sur l'inversion par réseaux de neurones des données des sondeurs verticaux atmosphériques satellitaires présents et à venir, Ph.D. thesis, Univ. Denis Diderot (Paris VII), Paris, 1993.

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

Jain, A., and D. Zongker, Feature selection: Evaluation, application, and small sample performance, *IEEE Trans. Pattern Anal. Mach. Intell.*, *19*(2), 153–158, 1997.

Jolliffe, I. T., *Principal Component Analysis*, Springer Ser. Stat., Springer-Verlag, New York, 2002.

Karhunen, J., and J. Joutsensalo, Representation and separation of signals using nonlinear PCA type learning, *Neural Networks*, *7*, 113–127, 1994.

Matricardi, M., and R. Saunders, Fast radiative transfer model for simulation of infrared atmospheric sounding interferometer radiances, *Appl. Opt.*, *38*(27), 5679–5691, 1999.

Monahan, A. H., Nonlinear principal component analysis by neural networks: Theory and application to the Lorenz System, *J. Clim.*, *13*(4), 821–835, 2000.

Nadal, J.-P., E. Korutcheva, and F. Aires, Blind source separation in the presence of weak sources, *Neural Networks*, *13*, 589–596, 2000.

Prunet, P., J.-N. Thépaut, and V. Cassé, The information content of clear sky IASI radiances and their potential for numerical weather prediction, *Q. J. R. Meteorol. Soc.*, *124*, 211–241, 1998.

Rabier, F., N. Fourrié, D. Chafaï, and P. Prunet, Channel selection methods for infrared atmospheric sounding interferometer radiances, *Q. J. R. Meteorol. Soc.*, *128*, 1011–1027, 2002.

Rodgers, C. D., Characterization and error analysis of profiles retrieved from remote sounding measurements, *J. Geophys. Res.*, *95*(D5), 5587–5595, 1990.

Scott, N. A., and A. Chédin, A fast line-by-line method for atmospheric absorption computations: The automatized atmospheric absorption atlas, *J. Appl. Meteorol.*, *20*, 20,802–20,812, 1981.

Sherlock, V. J., Impact of RTIASI fast radiative transfer model error on IASI retrieval accuracy, *Tech. Rep. FR-319*, 34 pp., Met Office, Braknell, UK, 2000.

F. Aires and W. B. Rossow, NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY 10025, USA. (faires@giss.nasa.gov; wrossow@giss.nasa.gov)

A. Chédin and N. A. Scott, Laboratoire de Météorologie Dynamique, École Polytechnique, 91128 Palaiseau Cedex, France. (chedin@jungle.polytechnique.fr; scott@araf1.polytechnique.fr)

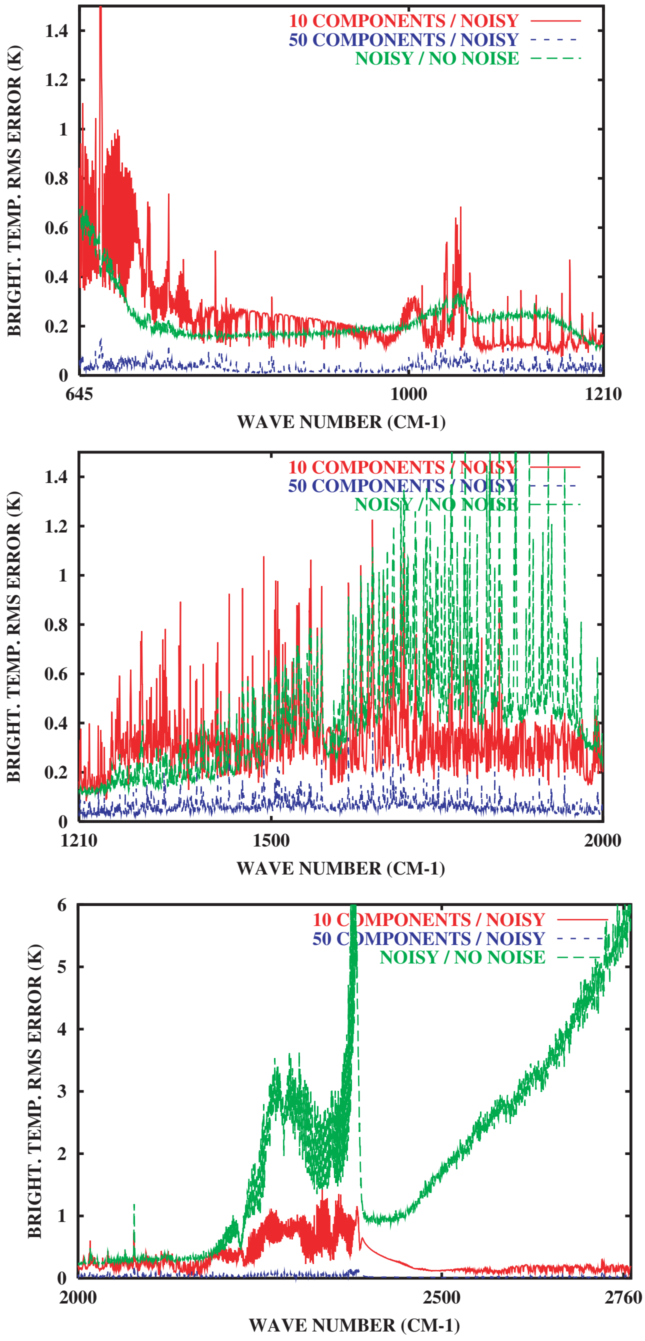


Figure 4. Statistics of compression in the 3 spectral bands of infrared atmospheric sounding interferometer instrument for 10 components (top line) and 50 components (bottom line), instrumental noise standard deviation is shaded for comparison purpose.

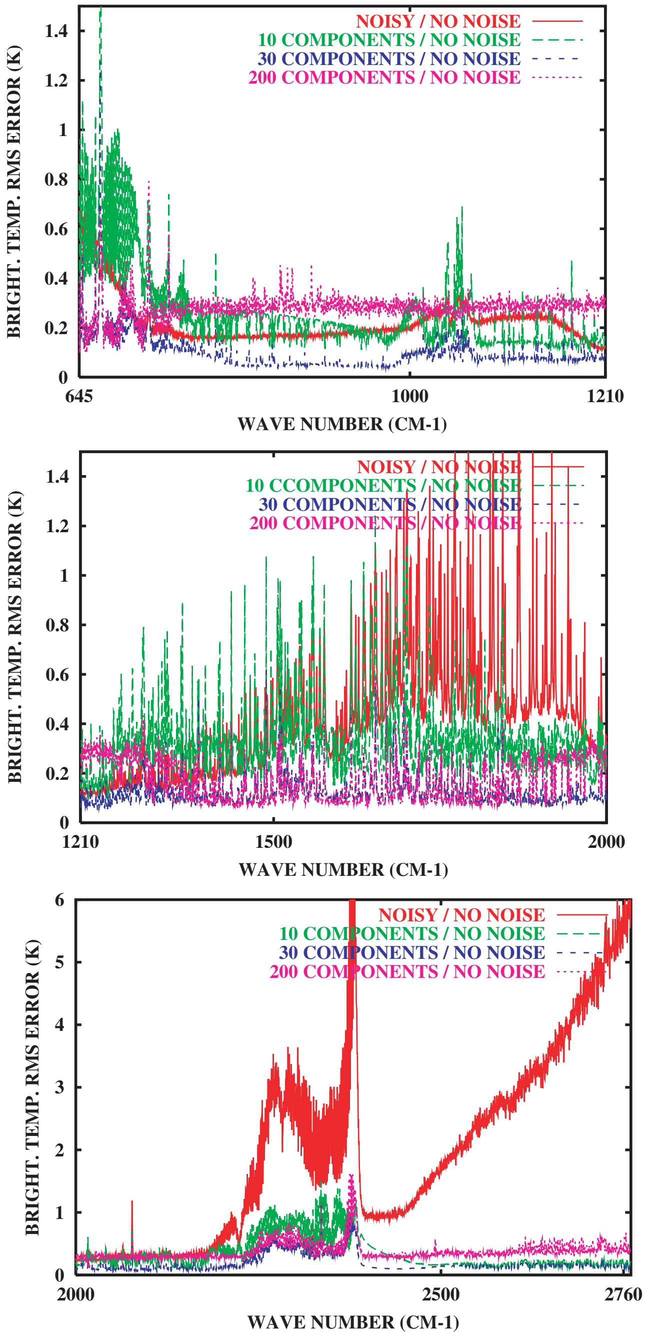


Figure 6. Statistics of denoising errors in the three spectral bands of infrared atmospheric sounding interferometer instrument using 10, 30 and 200 principal component analysis components, for the Thermodynamic Initial Guess Retrieval database situations, instrument noise (red line) is shown for comparison purpose.