Application of Principal Component Analysis to High-Resolution Infrared Measurement Compression and Retrieval

HUNG-LUNG HUANG

Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin—Madison, Madison, Wisconsin

PAOLO ANTONELLI

Department of Atmospheric and Oceanic Sciences, University of Wisconsin—Madison, Madison, Wisconsin

ABSTRACT

A simulation study is used to demonstrate the application of principal component analysis to both the compression of, and meteorological parameter retrieval from, high-resolution infrared spectra. The study discusses the fundamental aspects of spectral correlation, distributions, and noise; the correlation between principal components (PCs) and atmospheric-level temperature and water vapor; and how an optimal subset of PCs is selected so a good compression ratio and high retrieval accuracy are obtained.

Principal component analysis, principal component compression, and principal component regression under certain conditions are shown to provide 1) nearly full spectral information with little degradation, 2) noise reduction, 3) data compression with a compression ratio of approximately 15, and 4) tolerable loss of accuracy in temperature and water vapor retrieval. The techniques will therefore be valuable tools for data compression and the accurate retrieval of meteorological parameters from new-generation satellite instruments.

1. Introduction

In the last decade, using advanced technologies, new instruments have been built and proposed to improve observations of atmospheric temperature, water vapor, and winds. These instruments, mostly interferometers,[1] represent the realization of a new measurement concept based on high density, high spectral resolution infrared measurements. They will provide measurements with spectral resolutions finer than 1 cm^{-1} and have quasi-continuous spectral coverage in the 3–18-μm spectral region. The high-resolution measurements will greatly improve the vertical resolution of retrieved atmospheric variables (Huang et al. 1992). However, the volume of data generated at the satellite will increase by at least 2 orders of magnitude, exceeding the capacity of the current downlink technology and requiring expensive

ground data processing systems. Under these circumstances the development of an efficient and accurate data compression procedure becomes crucial for the full realization of the new measurement concept. The need for efficiency is imposed by the fact that the onboard computing power is limited, but accuracy is required to maximize the preserved information about the state of the atmosphere embedded in the raw measurements.

In this paper, we investigate the applicability of principal component analysis (PCA) to the problem of data compression and inversion (Wark and Fleming 1966; Smith and Woolf 1976; Smith et al. 1995) by formalizing quantitatively and qualitatively the trade-off between its efficiency and its accuracy. PCA is currently being used for the developmental retrieval algorithms for both the NAST-I and Atmospheric Infrared Sounder projects.

PCA has proven to be a valuable tool not only for compressing spectral data (in the brightness temperature space) but also for smoothing out part of the instrument noise that is contained in the measurements (in this paper, noise will normally refer to the Gaussian part of the instrument noise, unless otherwise noted). The key concept behind the evaluation of PCA as a data compression tool [i.e., principal component compression, (PCC)] is how the *best compression* achievable is defined. This concept in remote sensing is not uniquely

---

[1] That is, Interferometeric Monitor for Greenhouse Gases, Cross Track Infrared Sounder, Michelson Interferometer for Passive Atmospheric Sounding (Fischer and Oelhaf 1996), Infrared Atmospheric Sounding Interferometer, and National Polar-Orbiting Operational Environmental Satellite System Aircraft Sounder Testbed–Interferometer (NAST-I) (Cousins and Gazarick 1999).

Corresponding author address: Dr. Allen Huang, CIMSS, University of Wisconsin—Madison, 1225 West Dayton St., Madison, WI 53706.
E-mail: allen.huang@ssec.wisc.edu

defined; at least two definitions are well suited and widely accepted for interferometric data compression.

According to the first definition (hereinafter definition 1), the best compression is the one that preserves most of the information contained in the raw data, that is, the one that minimizes the loss of information. This definition represents a clear criterion for the evaluation of PCC, but its applicability depends on the nature of the information loss. The original signal, in fact, contains information on both the real state of the atmosphere and the instrument noise. A loss of the first component would lead to a loss of accuracy in the retrieved variables, but a loss of the second component would generally lead to an enhancement of the retrieval accuracy. Therefore, definition 1 is not completely applicable unless it is possible to determine the nature of the information loss due to the compression.

The second definition (hereinafter definition 2) is: the best compression is the one that guarantees the best retrieval performances. This definition is independent of the nature of the information loss but depends on the properties of the retrieval algorithm used (robustness to noise, sensitivity to nonlinear features, etc.). Different retrieval algorithms may not only lead to different evaluations of PCC as a compression technique but also to differences in the best compression ratio achievable. In this paper, PCC is evaluated with respect to both definitions 1 and 2. We will show that, given a retrieval algorithm based on linear regression, the best compression ratio achieved by PCC according to definition 1 does not strictly lead to the best retrieval results. In any case, a consistent conclusion can be drawn about the *optimal* compression ratio achievable by PCC according to both definitions. In fact, the loss in the accuracy of the retrieved variables is shown to be almost constant, and negligible, about its minimum for a wide range of values of the compression ratio. Therefore, even if the best compression ratio can not be uniquely determined, an optimal estimation of it can be found.

To assess PCC qualities according to both definitions and for different noises, only simulated high-resolution infrared data are used. Linear regression applied to compressed data has been chosen to retrieve temperature and water vapor profiles from the simulated observations.

In this study, it is shown that when the noise is Gaussian distributed with zero mean and known standard deviation (std), a compression ratio of about 15 can be achieved, with temperature retrieval degradation of less than 0.05 K and a water vapor retrieval degradation of less than 0.05 g kg^{-1}, with respect to the best linear regression retrieval achievable either for compressed or uncompressed data. In addition, the performance of PCC has been evaluated, according to definition 2, for different stds and statistical distributions of the instrument noise. For those cases, the conclusions drawn about the compression ratio still hold.

Section 1 describes in detail the characteristics of the datasets and instrument measurements used in the study and the procedure followed in generating the noise. Section 2 introduces PCA from a mathematical point of view and describes some of its most relevant properties. Section 3 focuses on PCC, that is, PCA applied to data compression; it describes separately the results obtained by applying PCC to noise-free and noisy data and discusses both the negative (atmospheric information loss) and the positive (noise reduction) effects of PCC on the synthetic data. In this section, we describe also some of the statistical properties of the compressed data. Section 4 discusses the inversion of the compressed radiance data using linear regression [this technique is hereinafter referred to as principal component regression (PCR)]. In the last part of the paper, the conclusion about the achievable compression ratio will be discussed, and some ideas for future work will be proposed.

Although most satellite infrared measurements are likely to be cloud contaminated, a realistic simulation of cloud effects is beyond the goal of this paper. The results of the current study are therefore applicable only to clear-sky observations, although cloud detection techniques are developing (Smith et al 1998; Cuomo et al. 1999) and, in the future, will be incorporated into this research.

2. The dataset and the instrument measurements

To apply PCA to high-resolution interferometric data, both a training set (12 000 profiles) and a testing set (2000 profiles) of temperature, water vapor, and ozone profiles were selected at random from a 1996 global radiosonde dataset. The profiles have been arbitrarily clustered in three main classes: 1) subpolar profiles [located at latitudes greater than 50°N (S)], 2) midlatitude profiles [located at latitudes between 25° and 50°N (S)], and 3) tropical and subtropical profiles (located at latitudes between 25°N and 25°S). Figure 1 shows the spatial distribution of the observations and the percentage of observations belonging to each class with respect to the entire ensemble.

The geographical distributions for the testing and training datasets are shown to be consistent, but the relative lack of observations over the oceans and in the Tropics might affect the representativeness of the datasets. For each profile, a noise-free, clear-sky spectrum of $N_c = 3888$ channels of brightness temperatures was generated through a radiative transfer model (Strow et al. 1998), with a spectral resolution of 0.6 cm^{-1} and a spectral coverage from 550 to 2750 cm^{-1} (Smith et al. 1990). This radiative transfer model allows both the atmospheric temperature and the concentration of absorption gases such as water vapor, ozone, methane, and carbon dioxide (CO_2) to vary. In the current implementation, the variation of CO_2 amount is fixed and all other minor gases (including methane and carbon mon-
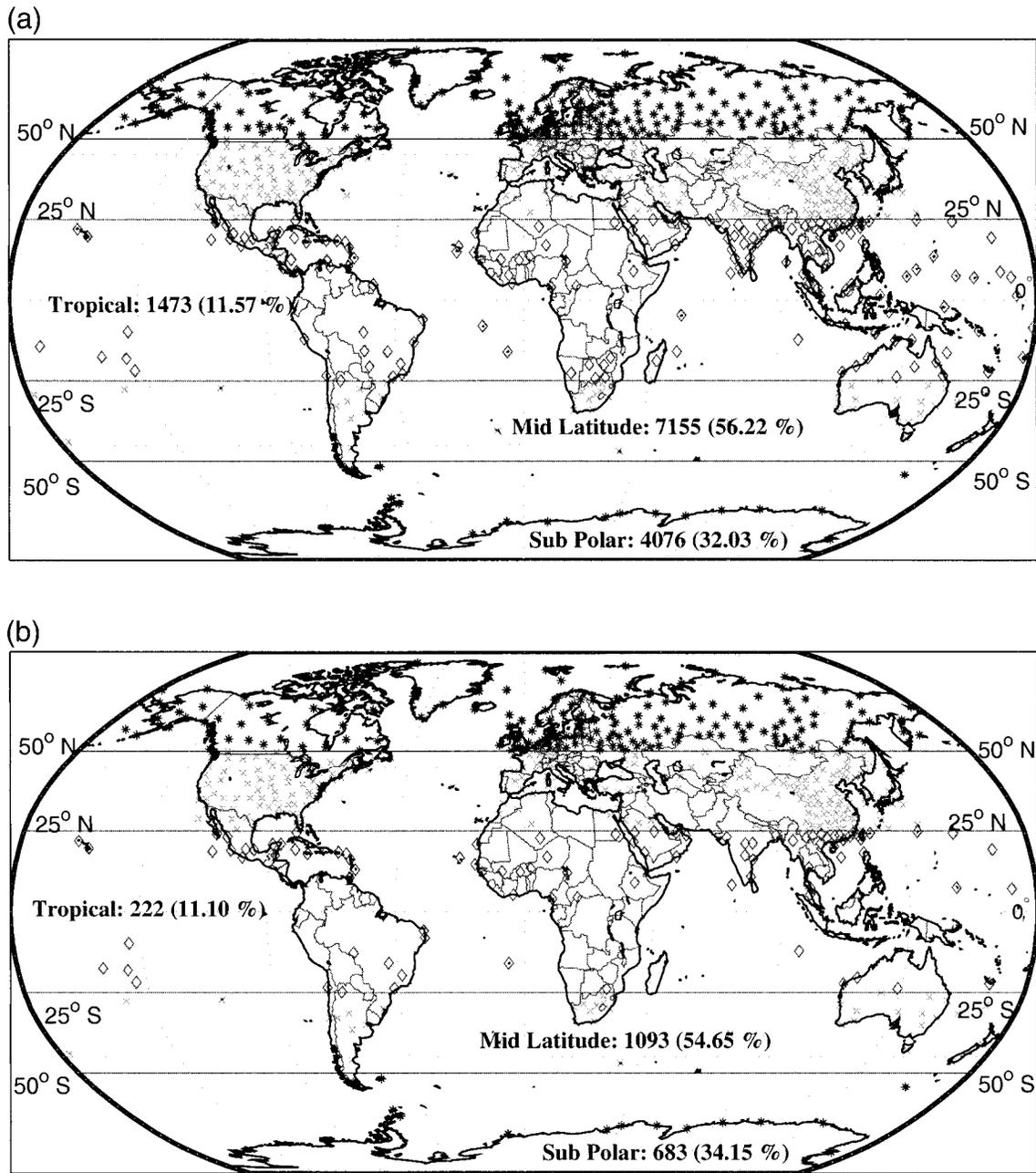
(a)

(b)

FIG. 1. Observation distributions for (a) training and (b) testing datasets.

oxide) are parameterized and varied as per climatological data.

Figures 2a and 2c show the mean temperature and water vapor profiles plus and minus the std, and Fig. 2e shows the mean, noise-free, simulated spectrum from the selected 1996 database. Figure 2b shows the temperature profiles characterized by the maximum and minimum mean temperature (over the 42 vertical pressure levels ranging from 0.1 to 1050 hPa), and Figs. 2d and 2f show the water vapor profiles and the noise-free spectra associated with the temperature profiles in Fig. 2b. Note that, because the profiles are defined only at 42 pressure levels, the simulated spectra differ from the real spectra not only because of the forward model's intrinsic limitations but also because the representation of the atmospheric (column) state has a finite number of degrees of freedom. The next section describes what PCA is and under which conditions it guarantees the best compression according to definitions 1 and 2.
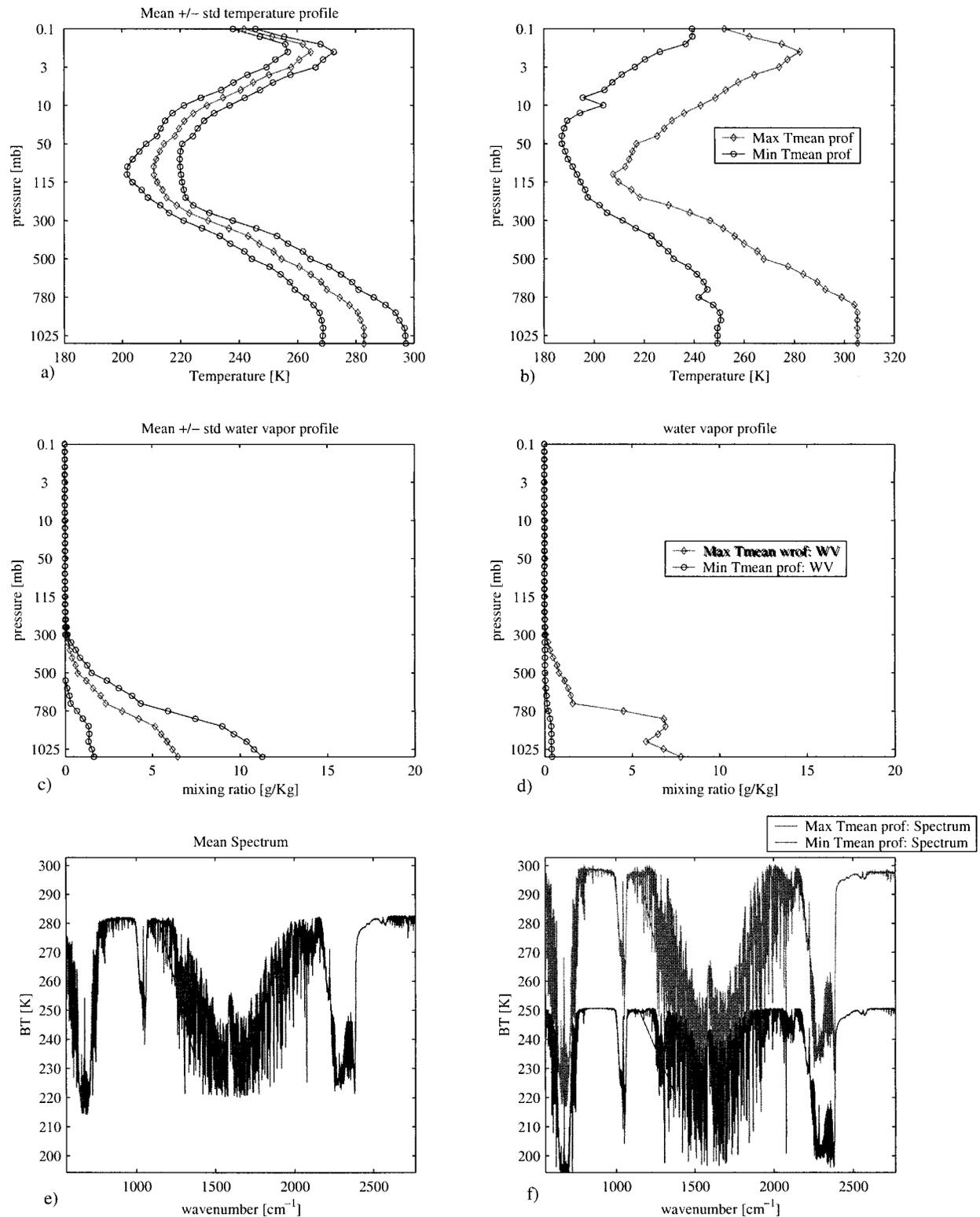
FIG. 2. (a) Training mean temperature profile and mean profile plus and minus standard deviations of training dataset. (b) Temperature profiles characterized by maximum and minimum vertical mean over the pressure levels. (c) Training mean water vapor profile and mean profile plus and minus standard deviations of training dataset. (d) Water vapor profiles associated with the temperature profiles in (b). (e) Brightness temperature spectrum calculated for mean profile condition of training dataset. (f) Brightness temperature spectra computed from the profiles in (b) and (d).
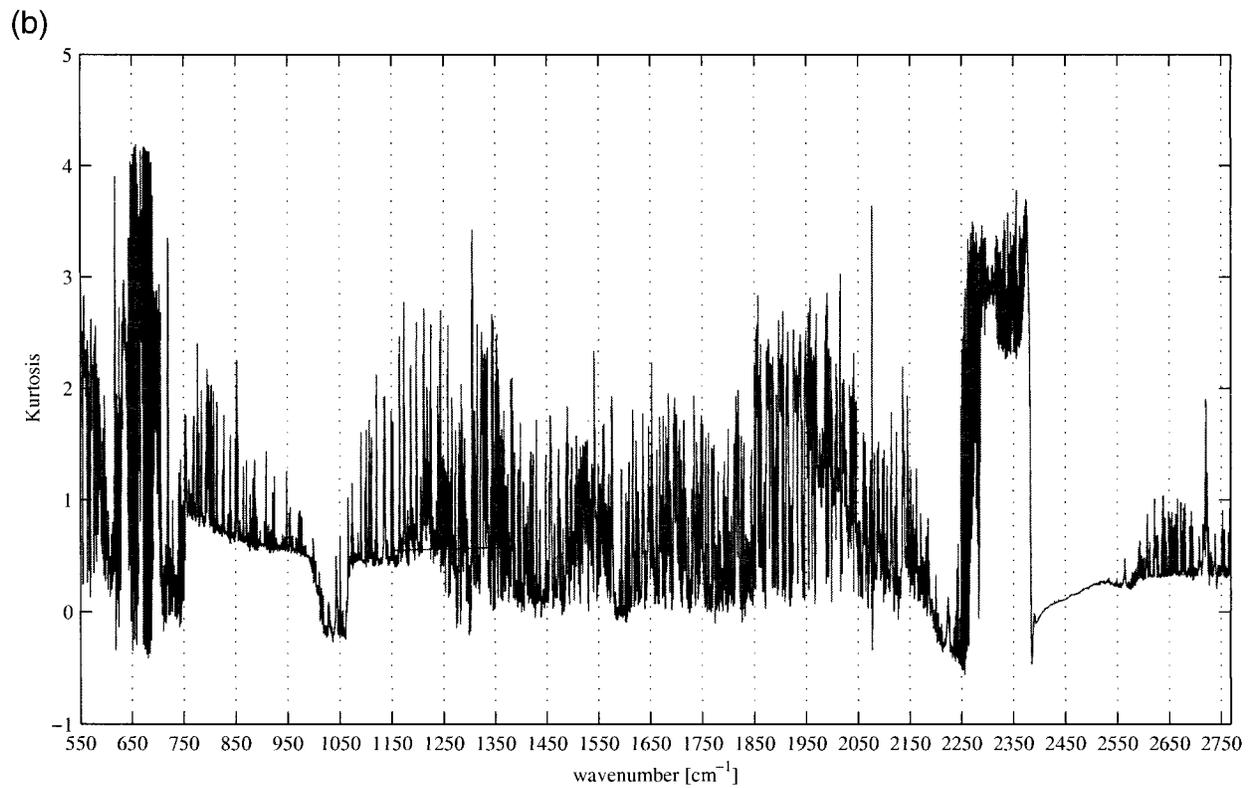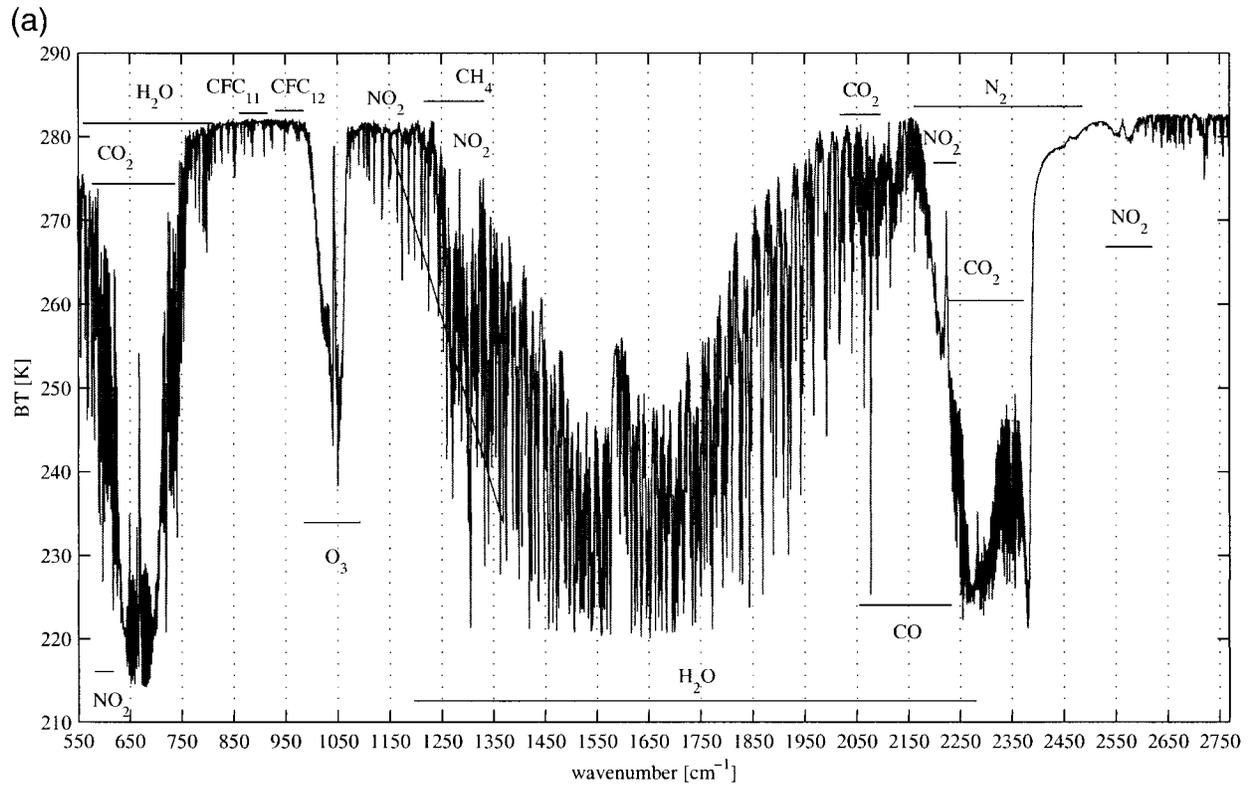
(a)

(b)

FIG. 3. (a) Training mean brightness temperature spectrum. (b) Non-Gaussian index (kurtosis) spectrum.
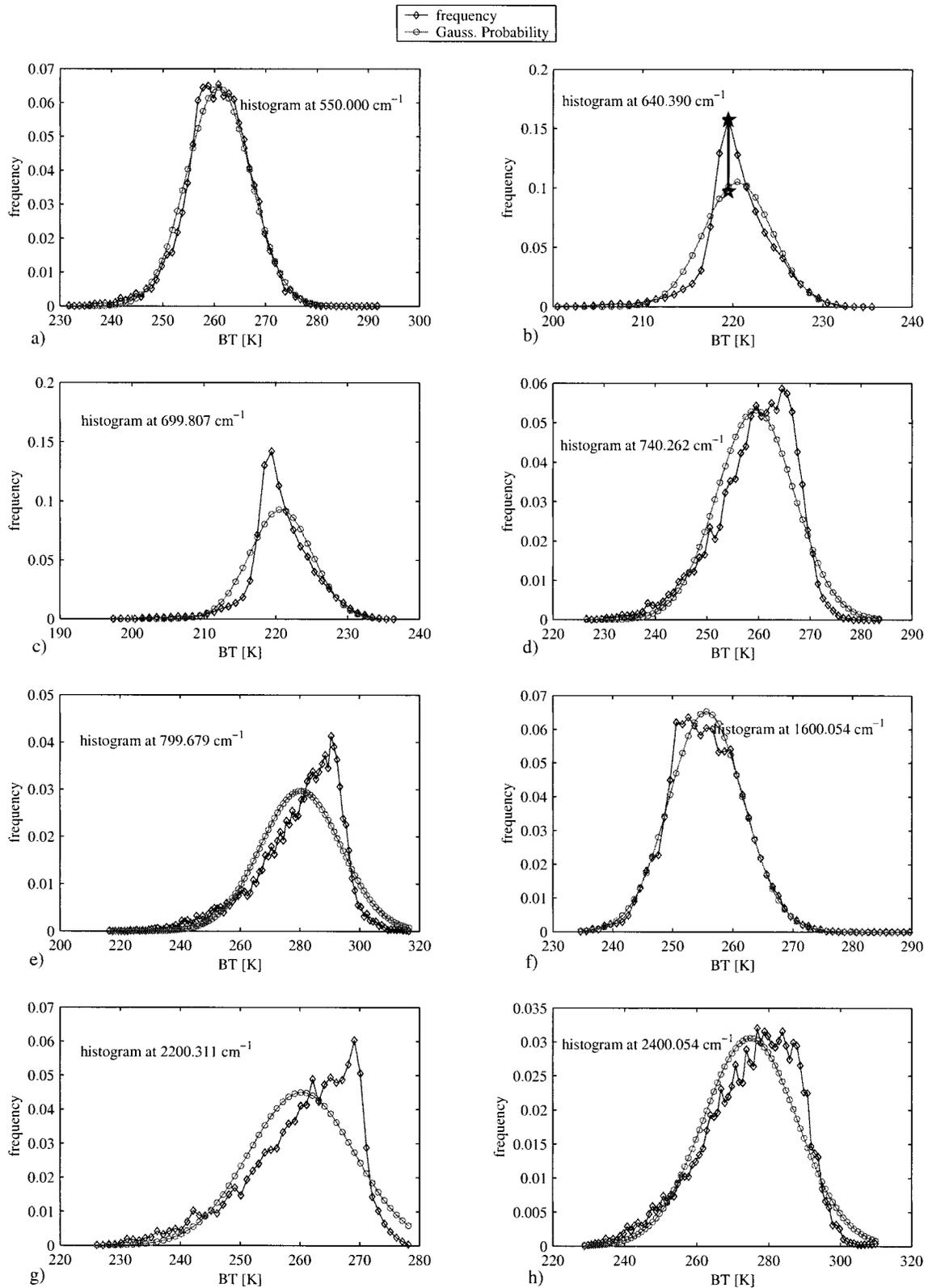
FIG. 4. (a), (b) Examples of the differences between Gaussian channel distributions and the Gaussian distributions with same mean and same std of the actual measurements in that specific channel. (c), (d) Examples of the differences between super-Gaussian channel distributions and the Gaussian distributions with same mean and same std of the actual measurements in that specific channel. (e)–(h) Examples of the differences between sub-Gaussian channel distributions and the Gaussian distributions with same mean and same std of the actual measurements in that specific channel.

3. PCA

PCA is a multivariate analysis technique that was first introduced by Pearson in 1901 and developed independently by Hotelling in 1933. It is commonly used to reduce the dimensionality of a dataset with a large number of interdependent variables. This reduction is achieved by finding a set of N_t orthogonal vectors in the input space of dimension N_c, with $N_t < N_c$, which accounts for as much as possible of the data variance. Hence, the problem of dimensionality reduction is reduced to finding a linear transformation from the N_c-dimensional input space to an N_t-dimensional subspace spanned by N_t orthogonal vectors defined above and hereinafter referred to as principal components (PCs). The first PC is defined as the direction along which the variance of the input data has its maximum. The second PC is the vector in the orthogonal $(N_t - 1)$-dimensional subspace complementary to the first PC, which explains most of the remaining variance and so on until the last PC is the direction of minimum variance.

Let x be an N_c-dimensional array of observables (in our case, brightness temperatures) with probability density function $p(x)$, and define C_{ij} as the ith, jth element of the covariance matrix **C** of a set of observations of x:

$$
C_{i,j} = \frac{1}{M_{\mathrm{tr}}} \sum_{p=1}^{M_{\mathrm{tr}}} \left(x_{i,p} - \frac{1}{M_{\mathrm{tr}}} \sum_{l=1}^{M_{\mathrm{tr}}} x_{l,p} \right)
$$
$$
\times \left(x_{j,p} - \frac{1}{M_{\mathrm{tr}}} \sum_{l=1}^{M_{\mathrm{tr}}} x_{l,p} \right), \qquad (1)
$$

where M_{tr} represents the total sample number of x over the training dataset. Then "The k-th principal component of the input vector x is the normalized eigenvector $\boldsymbol{\nu}_k$ corresponding to the eigenvalue λ_k of the covariance matrix **C**, where the eigenvalues are ordered $\lambda_1 > \lambda_2 > \cdots > \lambda_{N_c}$" [proof in Deco and Obradovic (1996)].

It can be shown easily that for the case $N_t = N_c$ PCA linearly decorrelates the output, that is, diagonalizes the covariance matrix of the output components. Hence, PCA essentially performs a singular value decomposition of the covariance matrix. However, the diagonalization of the covariance matrix (i.e., decorrelation) does not necessarily yield *statistical independence*. Statistical independence implies that the probability distribution is factorizable. Decorrelation in the general case yields to linear independence. It yields to statistical independence only under the assumption that the input variables are Gaussian-distributed. This assumption holds only for some of the channels. The implications of this problem will not be addressed in this paper, but given that different algorithms, better suited for problems having different input distributions, could result in better compression performances, we will discuss the probability density function of the measurements in some detail.

To describe how far the actual distributions are from being Gaussian, we have used the kurtosis, here defined as

$$
\mathrm{kurt} = \frac{E\{z^4\}}{(E\{z^2\})^2} - 3, \qquad (2)
$$

where $E\{\ \}$ represents the expected values, and z is a random variable. The values of the kurtosis, which can be either positive or negative, are plotted for each single channel in Fig. 3b. The channels that have kurtosis close to zero are generally Gaussian channels (Figs. 4a,b), those with positive kurtosis are called super-Gaussian channels (Figs. 4c,d), and the channels that have negative kurtosis are called sub-Gaussian channels (Figs. 4e,f). Super-Gaussian channels have typically a "spiky" PDF with heavy tails; sub-Gaussian channels, on the other hand, tend to have a "flat" PDF. In Fig. 4, the frequency distributions for six different channels are compared with the Gaussian distribution with the same mean and variance of the actual distribution; for each plot, the value of the kurtosis is indicated. From Figs. 3 and 4, it is clear that the distribution for many of the channels is not Gaussian, especially in some of the most important absorption bands. Therefore, in our case, after the application of PCA, the compressed data are decorrelated in the brightness temperature space, but they might not be statistically independent. In other words, the fact that they are not Gaussian-distributed may affect the optimal value achievable for the compression ratio.

Among the different mathematical and statistical properties of PCA, a relevant one for the present discussion is the compression error. The compression error (least squares error) is defined as LSE $= \overline{\|x - x_r\|}$, where x_r is the data array, which has been compressed and then reconstructed according to the procedure described in the next section. The $\|\ \|$ represents the Euclidean norm.

It can be shown that the LSE is minimized by any linear transformation whose rows span the same space spanned by the PC. The proof is based on the optimal reconstruction theorem: "The reconstruction error LSE is minimal when the rows of the linear transformation from the N_c-dimensional input space to the N_t-dimensional PC space are vectors spanning the same subspace as the N_t eigenvectors of the input covariance matrix corresponding to the N_t largest eigenvalues" [proof in Deco and Obradovic (1996)].

Note also that the transformation that leads to the optimal reconstruction is not the same as that obtained with the application of PCA and does not guarantee the decorrelation of the output components, even though they span the same principal eigenspace. Deco and Obradovic (1996) showed that in the reconstruction process, only when each element of the input array x is Gaussian-distributed is the "principal subspace projection method" (i.e., reconstruction after PCA compression) equivalent to "the principle of minimum information loss." As already mentioned, under the assumption that the input data (measurements) are Gauss-

ian-distributed, because of its reconstruction and decorrelation properties, PCA should be a natural candidate for data compression.

4. PCC

The following notation will be used: $S^{(tr)}$: noise-free training spectrum, $S^{(ts)}$: noise-free testing spectrum, $\tilde{S}^{(tr)}$: noisy training spectrum, $\tilde{S}^{(ts)}$: noisy testing spectrum, $R^{(ts)}$: reconstructed noise-free testing spectrum, and $\tilde{R}^{(ts)}$: reconstructed noisy testing spectrum.

a. PCC applied to noise-free data

According to this method, the compression of noise-free data is achieved through the following steps. First Eq. (1) for **C** is rewritten as as

$$C_{i,j} = \frac{1}{M_{tr}} \sum_{p=1}^{M_{tr}} \left[S_{p,i}^{(tr)} - \frac{1}{M_{tr}} \sum_{l=1}^{M_{tr}} S_{l,i}^{(tr)} \right]$$
$$\times \left[S_{p,j}^{(tr)} - \frac{1}{M_{tr}} \sum_{l=1}^{M_{tr}} S_{l,j}^{(tr)} \right], \quad (3)$$

where $i = 1, \ldots, N_c$ and $j = 1, \ldots, N_c$. The PCs are evaluated by diagonalizing **C**, that is, by finding the matrices **U** and **D** such that

$$\mathbf{C}_{N_c \times N_c} = \mathbf{U}_{N_c \times N_c} \mathbf{D}_{N_c \times N_c} \mathbf{U}^T_{N_c \times N_c}, \quad (4)$$

where **D** is a diagonal matrix whose elements are the eigenvalues of **C** and **U** is a matrix whose columns are the eigenvectors of **C**. The differences between the noise-free testing spectra and the mean training spectrum are projected onto the first N_t PCs (i.e., the N_t PCs associated with the N_t largest eigenvalues of **C**):

$$\Omega_{i,j} = \sum_{k=1}^{N_c} \left[S_{k,i}^{(ts)} - \frac{1}{M_{tr}} \sum_{l=1}^{M_{tr}} S_{k,l}^{(tr)} \right] U_{k,j}, \quad (5)$$

where $i = 1, \ldots, M_{tr}$ and $j = 1, \ldots, N_t$. Using those projection coefficients, the spectra are reconstructed as follows:

$$R_{i,j}^{(ts)} = \sum_{k=1}^{N_t} \Omega_{i,k} U_{j,k} + \frac{1}{M_{tr}} \sum_{l=1}^{M_{tr}} S_{l,j}^{(tr)}, \quad (6)$$

where $i = 1, \ldots, M_{tr}$; $j = 1, \ldots, N_c$; Ω_{ik} is the projection coefficient of spectrum k onto the ith PC; $(1/M_{tr}) \sum_{l=1}^{M_{tr}} S_{k,l}^{(tr)}$ is the kth component (channel) of the mean spectrum (over the training dataset from which $\Omega_{i,k}$ is also derived; and $R_{i,j}^{(ts)}$ is the ith reconstructed channel of the jth testing spectrum.

The reconstruction root-mean-square (rms) error is defined as

$$\text{rms} = \left\{ \frac{\sum_{j=1}^{N_c} \sum_{k=1}^{M_{ts}} [R_{j,k}^{(ts)} - S_{j,k}^{(ts)}]^2}{M_{ts} N_c} \right\}^{1/2}. \quad (7)$$
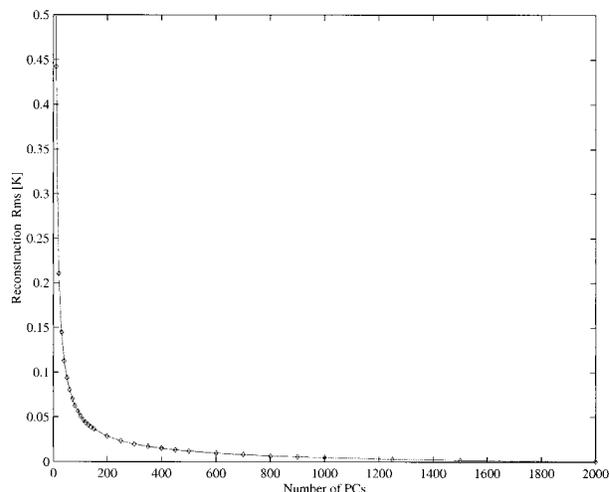

FIG. 5. Brightness temperature information loss in the noise-free data compression.

The reconstruction rms is plotted in Fig. 5 versus the number of PCs used, and it represents purely the loss of information due to compression. Figure 5 shows that, in the noise-free case, PCC is lossy, in the sense that it does not preserve the total information contained in the raw data, and that the rate of loss tends to zero as the number of PCs tends to the number of channels.

Before continuing further with the description of PCC, it is worth interpreting, with the help of linear correlation, the physical meaning of some of the PCs. As shown in the appendix, the linear correlation between the jth channel and the ith PC can be written as

$$\chi_{i,j} = \frac{\sigma_i}{\sigma_j} U_{i,j}, \quad (8)$$

where σ_i is the std of the projection coefficients on the ith PC, σ_j is the std of the ith component of $S^{(ts)}$, and $(U^T)_{ij}$ is the ith component of the jth PC.

Using this property of the PCs, it is possible to demonstrate that the first PC has high linear correlation with the window channels and low linear correlation with the CO_2 and water (H_2O) absorption band channels. The second and third PCs show the opposite tendency—they are highly correlated with the absorption channels and weakly correlated with the window channels. Figures 6a–c, show the correlation between the first three PCs and all the channels. Figure 6d shows the dependence of the mean rms correlation, defined as

$$\text{rms}_k = \left[\frac{\sum_{j=1}^{N} (\chi_{k,j})^2}{N_c} \right]^{1/2}, \quad (9)$$

where $\chi_{k,j}$ represents the correlation between PC_k and channel j. The mean rms correlation tends to reach a minimum and to fluctuate as the order approaches the number of channels.
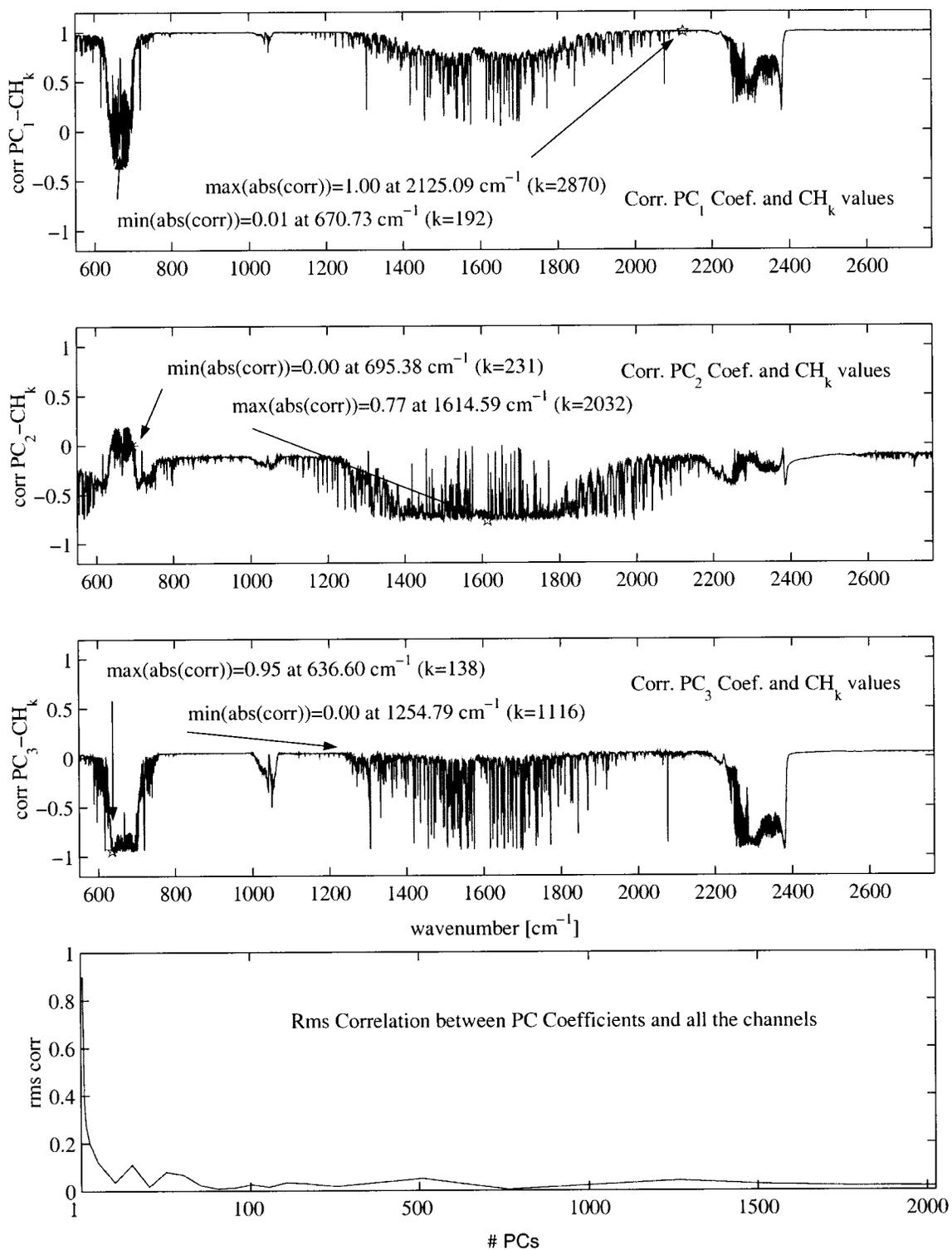
FIG. 6. Correlations between the (a) first, (b) second, and (c) third PC and all the channels. (d) Dependence of the mean rms correlation on PC number.
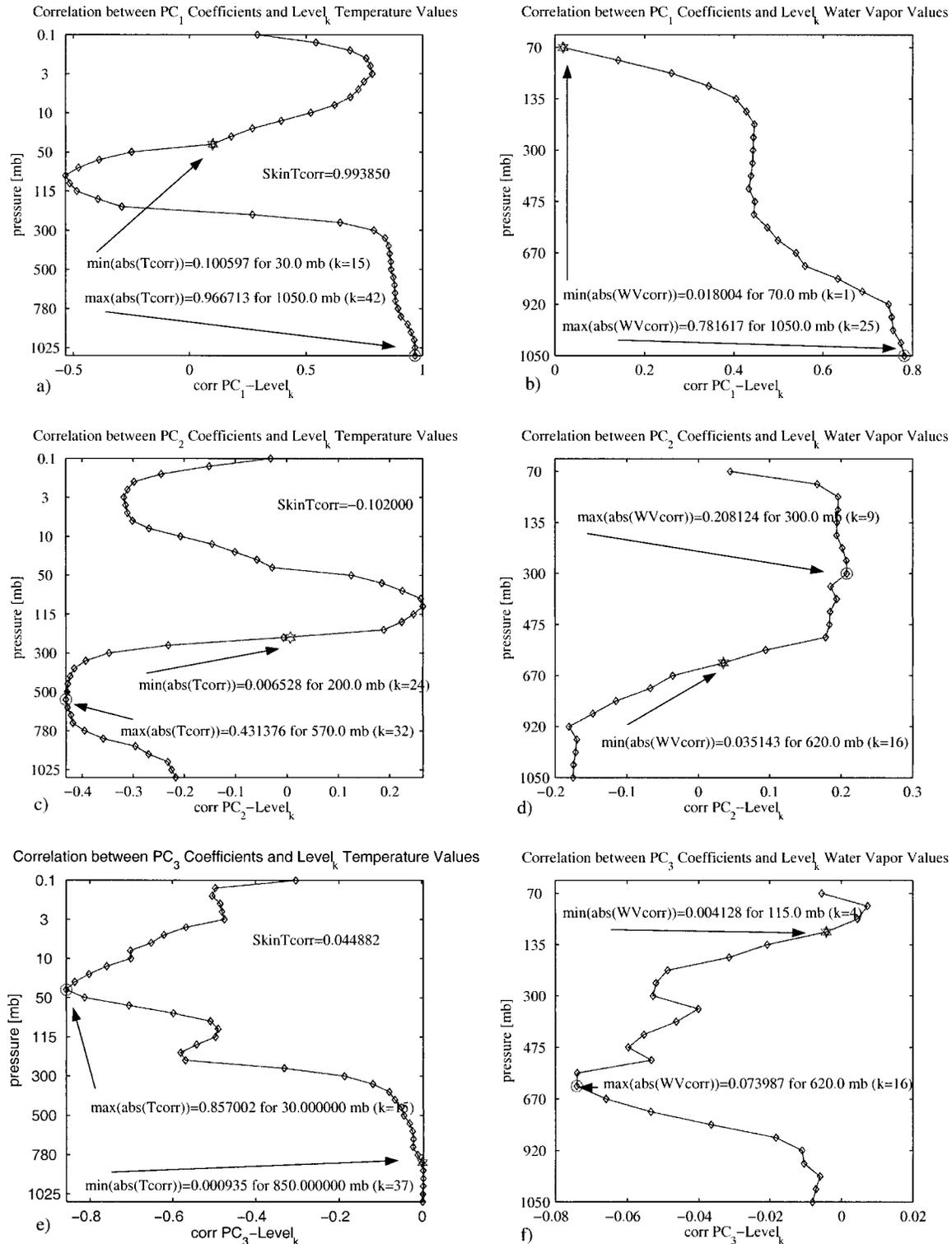
FIG. 7. Correlation between first PC and (a) temperature profile and (b) water vapor profile. Correlation between second PC and (c) temperature profile and (d) water vapor profile. Correlation between third PC and (e) temperature profile and (f) water vapor profile.

The linear correlation represents a useful tool to interpret the *physical meaning* of the different PCs, but it does not represent the higher-order correlation terms that may be important, especially for the absorption band channels. Using linear correlation it is also possible to estimate, to a first approximation, how much information each PC carries about a single temperature or water vapor level. Figures 7a,b show the correlation between the first PCs and all the different temperature and water vapor levels. As might be expected from the correlation between the PCs and the channels, the first PC is highly correlated with the lowest atmospheric level of temperature and water vapor. Less obvious is the correlation for the second and the third PCs and the atmospheric variables at different levels (Figs. 7c,d and 7e,f). The second PC (highly correlated with H_2O absorption band channels) has its maximum and minimum correlation with temperature at 500 and 300 hPa, respectively, and has its maximum and minimum correlation with water vapor at 300 and 670 hPa, respectively. The third PC has zero correlation with the lowest levels of temperature and maximum correlation with the temperature at 50 hPa and has its correlation maximum with temperature out of phase with the correlation maximum with water vapor.

b. PCC applied to noisy data

The introduction of noise substantially alters the conclusions made in the noise-free case. First, we will define how noise is simulated. The instrument noise for an interferometer has two main components. The first component, known as detector noise, is Gaussian-distributed with zero mean and std dependent on the measured radiance B (see section 1) through the following expression:

$$\sigma = \text{NeDT}\left(\frac{\partial B}{\partial B_T} \middle/ \frac{\partial B}{\partial B_T}\bigg|_{250 \text{ K}}\right), \tag{10}$$

where NeDT is 0.25 K at a scene temperature of 250 K and B_T represents the value of the brightness temperature. The detector noise is not correlated in the time domain (white noise), and it is the dominant term. The second term is associated with the dynamics of the mechanical components of the interferometer. It is generally correlated in time, and it may be modeled simply by a small percentage of the measured signal plus a small percentage of the first time derivative of the signal itself. With the introduction of noise, the signal to be compressed becomes

$$\tilde{S} = S + \eta, \tag{11}$$

where $\tilde{S}$ is the noisy signal and η is the noise. Using the linearity of PCC, it is possible to write

$$\text{PCC}[\tilde{S}^{(ts)}] = \text{PCC}[S^{(ts)}] + \text{PCC}(\eta). \tag{12}$$

Because we have already discussed the first term on the right side of Eq. (12), the following section will evaluate the PCC effects on the second term and then will discuss them jointly.

Figures 8a–c compare the reconstructed noise using three different numbers of PCs to the original noise for a single spectrum. The first few PCs carry little information about the noise, and the higher-order PCs carry more and more noise information. The rms of reconstructed noise has also been plotted in Fig. 9 versus the number of PCs used in the compression.

Let us define two new variables that represent a measure of the differences between the noisy signal reconstructed after compression and the original uncompressed noise-free signal (rmsdiff) and the differences between the signal reconstructed after compression and the original noisy signal (nrmsdiff), respectively:

$$\text{rmsdiff} = \left\langle \frac{\left\{\sum_{i=1}^{N_c} [S_i^{(ts)} - \tilde{R}_i^{(ts)}]\right\}^2}{N_c} \right\rangle^{1/2}, \quad \text{and}$$

$$\text{nrmsdiff} = \left\langle \frac{\left\{\sum_{i=1}^{N_c} [\tilde{S}_i^{(ts)} - \tilde{R}_i^{(ts)}]\right\}^2}{N_c} \right\rangle^{1/2}. \tag{13}$$

Figure 16a (described later; obtained by comparing Figs. 5 and 9) shows that PCC has two different effects when applied to noisy spectra: using too few PCs results in part of the signal information being lost; using too many PCs results in more and more noise being reconstructed.

Figure 10a shows a minimum in the rms reconstructed differences rmsdiff at approximately 150 PCs. With real data, the original noise-free signal is not available, and the evaluation of the effects of PCC on the data has to be based on the differences between the original noisy signal and the reconstructed noisy signal. Figure 10b shows how the evaluation of PCC based on the differences between the original noisy signal and the reconstructed one could be misleading; the noise filtering effect would be interpreted as signal information loss.

According to definition 1, an objective criteria for the evaluation of PCC could be based on the minimization of rmsdiff. That would ensure the minimum loss of information, but, as we will show, it does not guarantee the best performance for the retrieval process. For example, although a compression obtained using the first 150 PCs minimizes the reconstruction rms differences, it does not yield the best temperature and water vapor retrieval. Therefore, definition 2, based directly on the retrieval performance, is also needed to provide a different perspective on the definition for the optimal compression of spectral data. According to this definition, the best compression ratio value is the minimum among all the compression ratio values for which the retrieval rms error is the smallest at every level.

5. PCR

In multiple linear regression, one of the major difficulties with the least squares estimators is multicollinearity. Multicollinearity occurs when there are near-constant linear functions of two or more predictor variables (in this case, the brightness temperature values) and is often, but not always, indicated by large correlations between subsets of the variables. Figure 11 shows the mean absolute value of the correlation between each channel and all the others defined as

$$\overline{\chi} = \frac{1}{N_c - 1} \sum_{\substack{i=1 \\ i \neq j}}^{N_c} |\chi_{i,j}|, \tag{14}$$

where χ_{ij} represents the correlation between channel i and channel j. The channels with the highest mean correlation are the windows. Figure 3a shows the mean spectrum and the absorption bands. The absorption line channels (except for the ozone band) are characterized by smaller mean correlation. If multicollinearities exist, the variances of some of the estimated regression coefficients can become very large, leading to unstable and often misleading estimates of the regression equation (Jolliffe 1986).

PCR alleviates the existing multicollinearities among the high-spectral-resolution brightness temperature measurements. It simply uses the projections of the predictor variables (brightness temperature) onto a subset of PCs in place of the predictor variables themselves. Because the PCs are linearly uncorrelated, there are no multicollinearities between the projection coefficients. If all the PCs were included in the regression, then the resulting regression would be equivalent to that obtained by least squares regression, and the large variances caused by multicollinearities would not be reduced. However, if some of the PCs related to the smaller eigenvalues are deleted from the regression equation, then the large variances for the regression coefficients are greatly reduced, and reliable regression estimates can be achieved.

According to Eq. (5), the brightness temperature measurements can be transformed into expansion coefficients Ω_{ij} by using a subset of N_t PCs, where $i = 1$ to N_t and $j = 1$ to M_{tr}. Equation (5) represents the mapping of the whole spectrum of information of N_c channels onto the PC space of N_t coefficients. This mapping significantly reduces the dimensionality of the matrix needed to be inverted later, because N_t is usually much smaller than N. The linear regression relationship between expansion coefficients of infrared measurements and temperature or water vapor profiles is defined by

$$\rho_{l,i} = \phi_{l,j}(\Omega^{T})_{j,i} + [\overline{\rho_l} - \phi_{l,j}(\overline{\Omega^{T}})_j], \tag{15}$$

where l is the number of vertical levels representing the temperature or water vapor profile ρ; the overbar indicates a mean over i; and i and j are the same as in Eq. (5). The variable $\mathbf{\Phi}$ becomes the regression retrieval coefficient matrix that needs to be derived from a very large training ensemble dataset $S^{(tr)}$ to perform regression retrieval analysis on other independent infrared measurements $\tilde{S}^{(ts)}$. The least squares solution for $\mathbf{\Phi}$ in Eq. (15) can be derived as

$$\phi_{l,j} = (\rho_{l,i} - \overline{\rho_l})(\Omega_{i,j}) \frac{[(\Omega^{T})_{j,i}\Omega_{i,j}]^{-1}}{M_{tr}}, \tag{16}$$

where superscripts $(\ )^{-1}$ and $(\ )^{T}$ stand for the matrix inverse and transpose, respectively. The matrix to be inverted has the size of N_t by N_t, where N_t is the number of PCs in the selected subset. In the case of using channel measurements directly in this regression analysis, the size of this matrix is significantly increased to N_c by N_c, where N_c for typical current and future high-spectral-resolution infrared spectra is on the order of thousands.

After one successfully obtains the representative regression coefficient matrix $\mathbf{\Phi}$ from the ensemble training dataset (i.e., 1996 radiosonde dataset), the regression retrieval of the independent high-spectral-resolution sounding-instrument measurements with specific noise, spectral coverage, and resolutions for temperature or water vapor profile retrievals is obtained using Eq. (17),

$$\rho_l^{ind} = \phi_{l,j}\Omega_{j,i}^{T\,ind} + \overline{\rho_l^{dep}} - \phi_{l,j}\overline{(\Omega^{T})_j^{dep}}, \tag{17}$$

where Ω^{ind} is computed from Eq. (5) using $\tilde{S}^{(ts)}$ from independent infrared measurements, and $\mathbf{\Phi}$ is simply obtained from Eq. (16), which is derived from the training dataset described in the previous section. The regression results p^{ind} are of temperature or water vapor profiles that can be compared with the true profiles used to simulate the independently measured $\tilde{S}^{(ts)}$.

6. PCR retrieval optimization

The optimal number of PCs, N_t in Eq. (15), can be determined by balancing two partially conflicting effects. To eliminate large variances of regression coefficients due to multicollinearities, it is essential to remove all those components whose variances are very small without deleting components that have large correlations with the dependent variable ρ. One approach is to use different values of N_t and to determine which N_t yields the smallest temperature or water vapor profile retrieval errors over an ensemble of independent simulated measurements. The temperature and water vapor retrieval rms error for the testing spectra have been evaluated at different pressure levels. The asterisk marks in Figures 12, 13, 14, and 15 represent the values of N_t that minimize the rms error for that specific level.

To evaluate the dependence of N_t on the noise characteristics, the optimal N_t has been determined for five different cases characterized by five different assumptions on prior information available about the real noise properties, as follows.

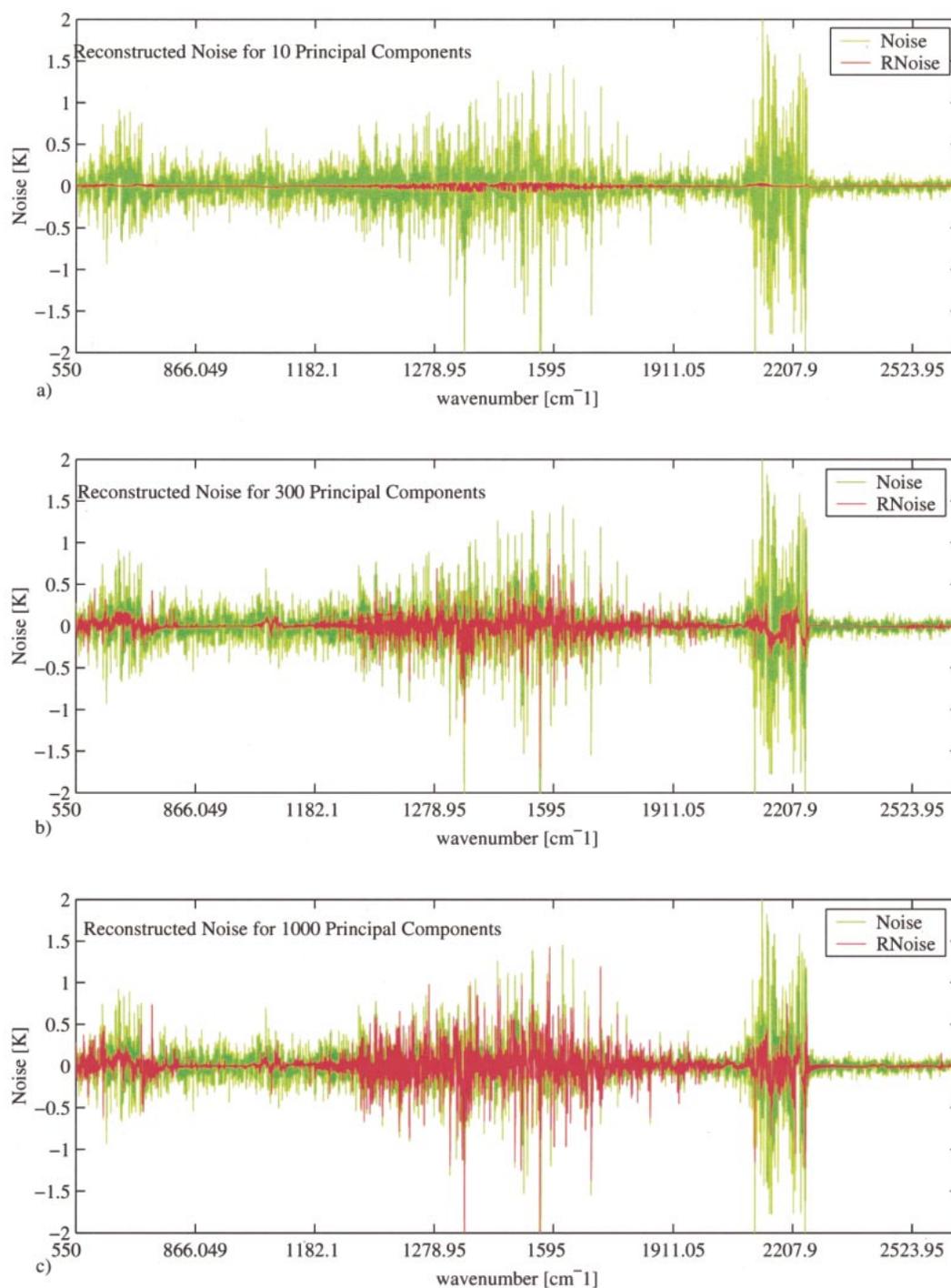1) No information available: the regression coefficients

FIG. 8. Noise compression: (a) 10-PCC, (b) 300-PCC, and (c) 1000-PCC compressed noise as compared with uncompressed noise for a single-spectrum case.

Φ are evaluated using noise-free training data (diamonds in Figs. 12 and 13).
2) Perfect knowledge: the regression coefficients are evaluated using noisy training data, where the noise has the same statistical properties of the testing data (circles in Figs. 12 and 13).

3) Supernoise (partial knowledge with high noise): the noise added to the testing spectra has (for each channel) an std that is 1.4 times the one used for the noise added to the training spectra used for the evaluation of the regression coefficients (dashed curves in Figs. 14 and 15).
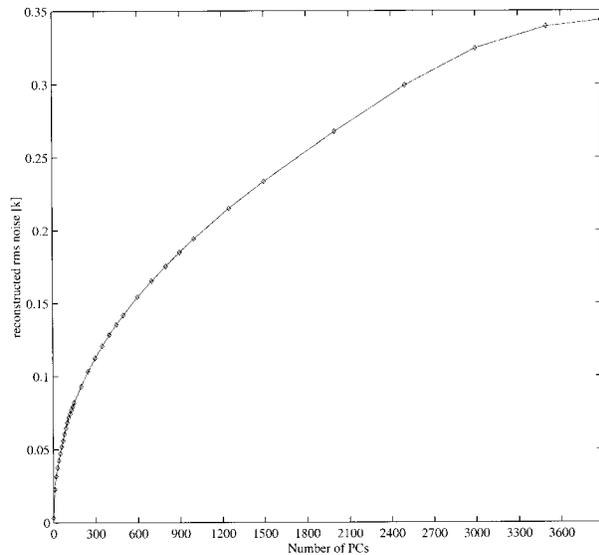
FIG. 9. Noise reconstruction rms error using different numbers of PCs.

4) Subnoise (partial knowledge with low noise): the noise added to the testing spectra has (for each channel) an std that is 0.6 times the one used for the noise added to the training spectra used for the evaluation of the regression coefficients (solid curves in Figs. 14 and 15).

5) Correlated noise (partial knowledge with correlated noise): the noise added to the testing spectra has the same characteristics as the noise added to the training spectra used for the evaluation of the regression coefficients but with the addition of a small component that is correlated in the B_T space; the testing spectrum noise is obtained by adding to the uncorrelated noise 0.1% of the signal (dotted curves in Figs. 14 and 15).

Figures 12 and 13 show, for every atmospheric level, that in the absence of noise information, the rms error minima are larger than the ones obtained in Case 2 (perfect knowledge). The lack of prior information on the noise causes a very fast degradation of the retrieval performances when N_t increases (i.e., PCR in case 1 is much more unstable than in case 2). The optimal N_t in case 2 is dependent on the level of interest. For temperature (Figs. 12a–h), the levels in the middle of the atmosphere, 100–850 hPa, have an optimal N_t that is smaller than the optimal N_t of the levels near the surface (1000 hPa). The minimum is reached at 2000 PCs, which is about one order of magnitude greater than the best compression ratio estimated according to the minimum information loss. The water vapor requires much fewer PCs to obtain the optimal retrieval accuracy (Figs. 13a–e). In addition, for case 2, the regression retrieval is proven to be very stable with respect to the number of PCs used. Figures 14 and 15 show that the presence of a small component of correlated noise does not affect

significantly the retrieval results. They also show that the over- or underestimation of the noise in generating the regression coefficients $\mathbf{\Phi}$ affects the PCR performances in the same way as cases 1 and 2, but with milder effects. Cases 3 (supernoise) and 4 (subnoise) simply represent intermediate cases between the extreme cases of case 1 and 2.

These results show that PCA allows a good compression ratio using only a subset of PCs that characterizes the high-spectral-resolution infrared measurement with much smaller dimension or data volume. In addition, PCA clearly guarantees small degradations of the temperature retrievals.

Another important characteristic of PCR is that the temperature retrieval performance obtained for a range of PCs between 200 and 1000 is almost constant. The rate of retrieval degradation is very small, as shown in Fig. 16b. These results demonstrate the advantage of PCR over traditional least squares techniques; least squares retrieval, not demonstrated here, tends to become unstable when correlated measurements are used.

Note that the best compression ratio, according to definition 1, is approximately 20, but, according to definition 2, the best compression ratio is about 2. Although these results seem to be very different, they are consistent with each other. In fact it has been proven that the retrieval degradation for a compression ratio up to 20 is very small. The reasons for this difference might be explained simply by applying PCR to the compressed spectra to retrieve the original noise-free spectra. Figure 17 shows the reconstruction errors for some of the channels. The circles represent the differences between the reconstructed signal and the real noise-free signal as a function of the number of PCs used in the reconstruction. The diamonds represent the differences between the regressed signal obtained by applying PCR to retrieve the brightness temperature values instead of temperature or water vapor values and the original noise-free signal. The regressed spectra differences, after reaching the minimum, are nearly constant, while the reconstructed spectra differences, after reaching the minimum, tend to increase. This indicates that PCR is very robust with respect to the Gaussian component of the instrument noise. In the retrieval process, increasing the number of PCs, even if high-order PCs introduce more and more noise, does not negatively affect PCR performance as long as the PCs keep carrying some information about the real signal. After 2000 PCs, the ratio between the noise introduced and the real signal information added is so large that PCR performances start degrading. Therefore, we should conclude that the evaluation of PCR according to Definition 2 depends on the robustness of the retrieval algorithm with respect to noise. The value of the compression ratio corresponding to the best retrieval performances is not really indicative as long as the retrieval performances are nearly steady for a wide range of compression ratio values.
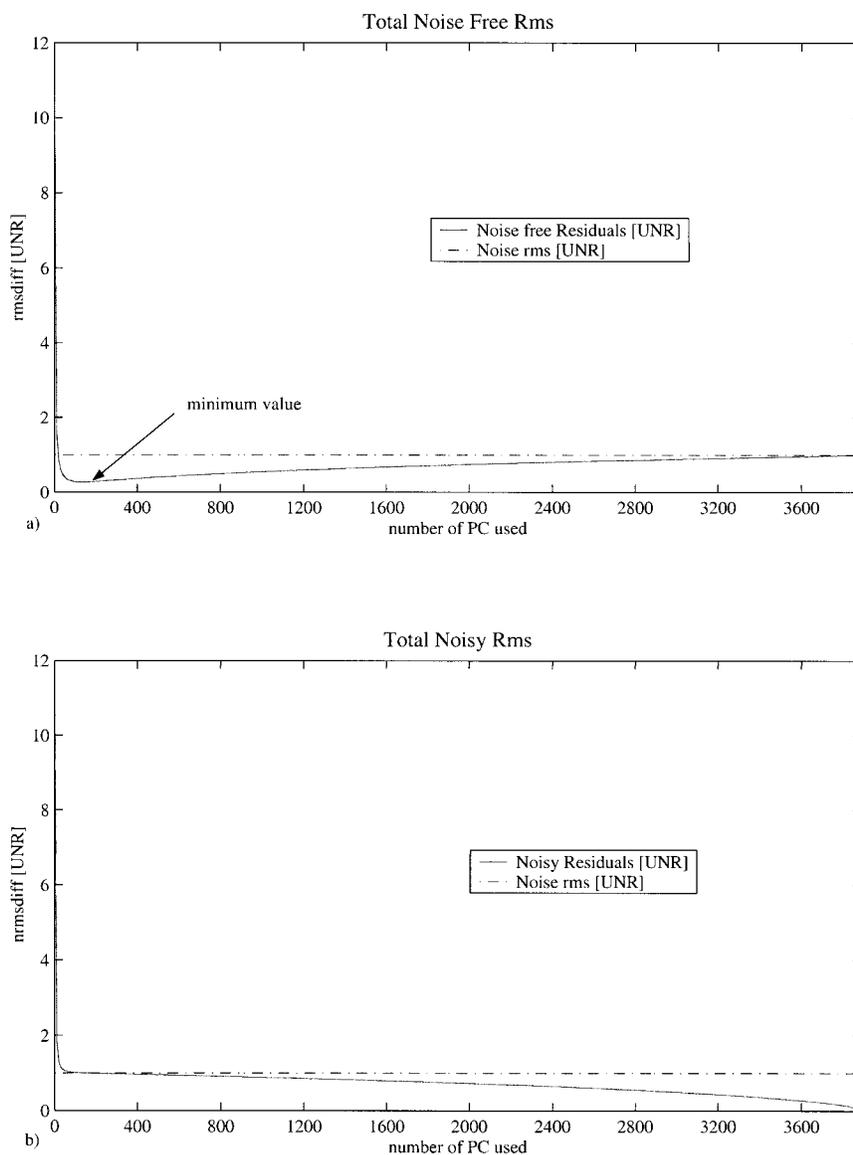
Total Noise Free Rms

Total Noisy Rms

FIG. 10. (a) Noise-free and (b) noisy-data PCC reconstruction residual rms error.

7. Best compression ratio for PCC

We now address the compression effect as a result of the application of PCA in the measurement space. Figure 18 displays the reconstruction residual of noisy spectra of independent measurements in the spectral range from longwave infrared of 550 cm^{-1} to the shortwave infrared of 2740 cm^{-1} using 250 PCs. The rms of measurement noise is also shown for comparison. Using only 250 PCs, the complete spectral measurement of 3888 channels can be represented within measurement noise. The loss of measurement information caused by the reduced number of PCs, 250 versus 3888, is compensated by the noise-reduction effect inherent in PC reconstruction. Through the use of an optimal subset of PCs, most of the information about the atmospheric state can be retained, and the random component of measurement noise of each spectral channel is reduced. If one were to quantify the whole 3888 channel measurements with 250 PCs, the measurements would be compressed without significant loss of signal. The signal-to-noise ratio may also be enhanced at the same time because of noise reduction provided by the PC reconstruction process. Figures 12, 13, 14, and 15 further confirm that, for a wide range of compression ratio values, the retrieval performances are nearly constant. Setting a retrieval degradation threshold at 0.05 K for each individual profile level of temperature and from 0.05 to 0.002 g kg^{-1} for different levels of water vapor, a compression ratio of ~15 for temperature and ~20 for water vapor is feasible.
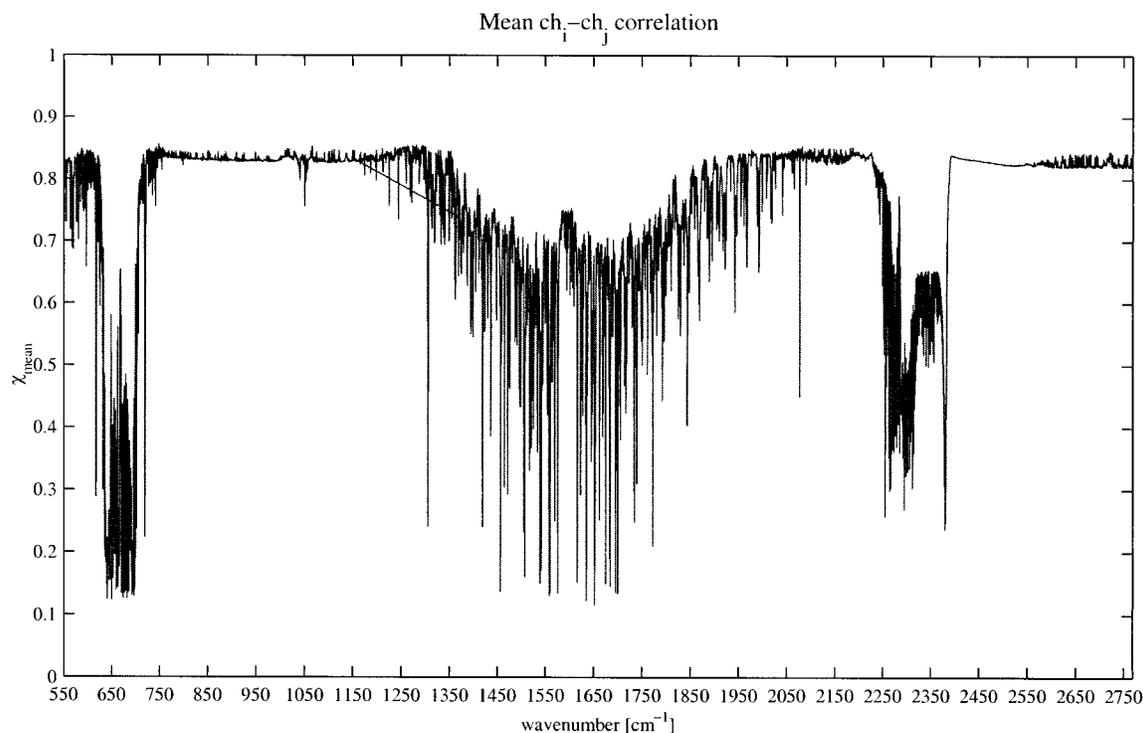
FIG. 11. Channel-to-channel mean correlation.

8. Summary and future work

These results demonstrate that PCA, PCC, and PCR methods will be effective in processing the high volume of data provided by the new generation of instruments and in accomplishing accurate sounding profile retrievals. The best compression ratio and the best retrieval can be simultaneously achieved in terms of the minimization of both the reconstruction residuals and regression retrieval error. PCA is demonstrated to be an efficient way to decorrelate the measurements, even when the input is not always Gaussian. Some of the most significant PCs are highly correlated with retrieval variables and therefore could be selectively used in linear regression, PCR, to retrieve temperature and water vapor profiles. An optimal subset of PCs has been shown to preserve the information, to reduce noise, and to compress efficiently the infrared measurements. Furthermore, an optimal subset of PCs ensures an improvement of the PCR retrieval computational and numerical efficiency by significantly reducing the dimensionality of the matrix that has to be inverted. It is shown that, regardless of different noise configurations, similar results of PCC and PCR are obtainable.

Although not discussed here, one significant source of error for this simulation study is the radiative transfer forward-model error, caused by the uncertainty of spectroscopic knowledge of the absorption characteristics of various abundant atmospheric gases. This error component will be included in future studies when real aircraft or satellite high-spectral-resolution infrared mea-

surements along with in situ data will be available to characterize this spectroscopic component.

The results obtained for PCA, PCC, and PCR are based on specific assumptions about the training and testing datasets, about the distribution of the infrared measurement noise, and about the linearity of PCA and PCR, assumptions that under certain measurement conditions may not be optimal. Clustered, regional, and seasonal datasets may further optimize the results of PCC and PCR; we will defer this study to future work.

The simulation of measurement noise is always subject to difficulties. In the follow-up research, we will simulate the noise in measurement space and interferogram space, and we will perform the same simulation experiments in this space. We will also investigate the effects of spectral resolution and sampling on PCC and PCR performance. The final assumption about the linearity of the transformation used to compress and retrieve the data (PCC and PCR) is a good first approximation for real problems in which the nonlinearity of the measurements is not easy to model and to implement. However, we do plan to proceed with a nonlinear PCA or independent component analysis in the near future. Linear PCR will be compared with a nonlinear algorithm, such as an artificial neural network–based algorithm.

With the simulation of future instrument measurements, it is possible to analyze the behavior and the performance characteristics of principal component compression and linear regression retrieval in a way that
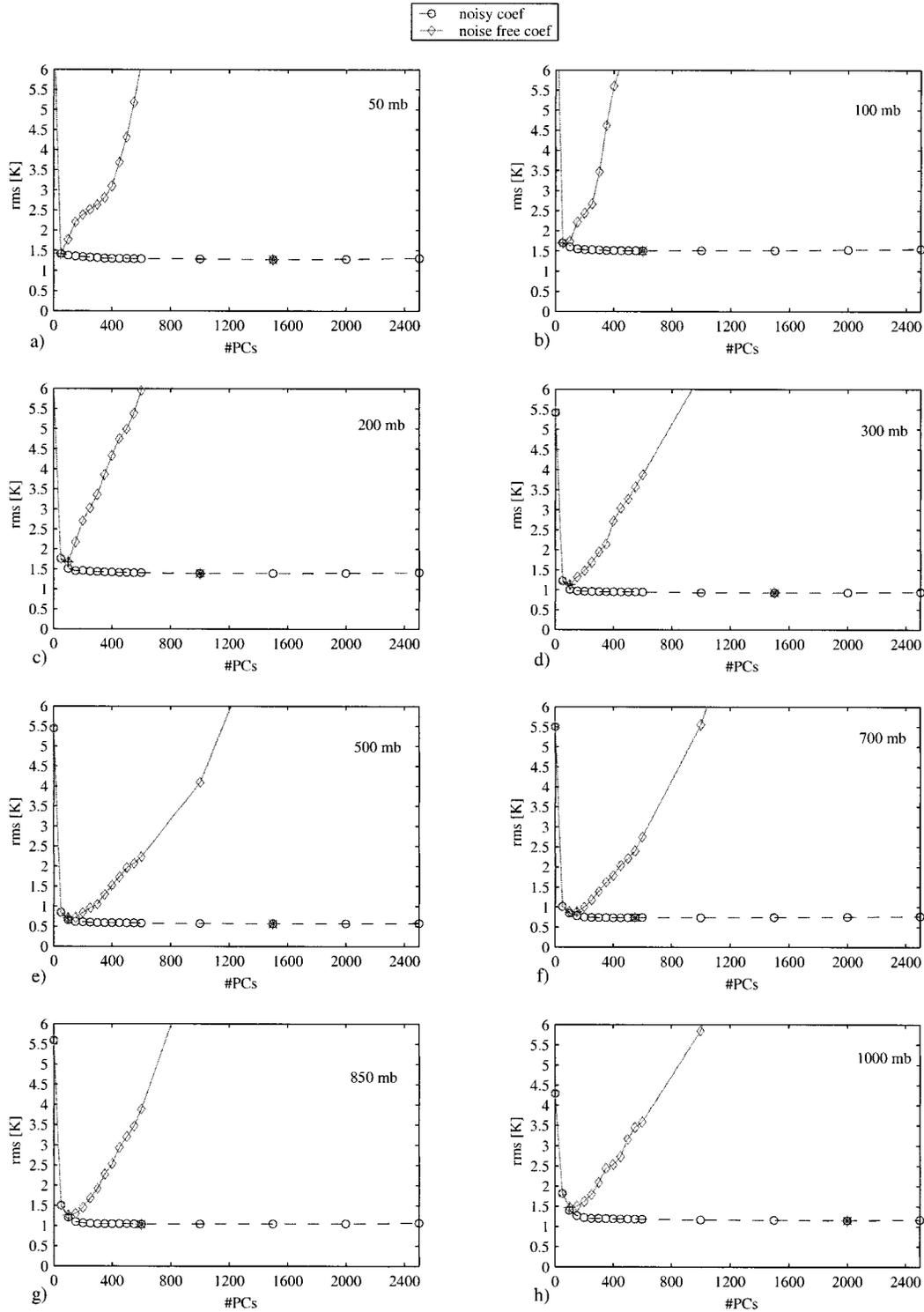
FIG. 12. Temperature retrieval degradation of noisy and noise-free PCR: (a) 50, (b) 100, (c) 200, (d) 300, (e) 500, (f) 700, (g) 850, and (h) 1000 hPa.
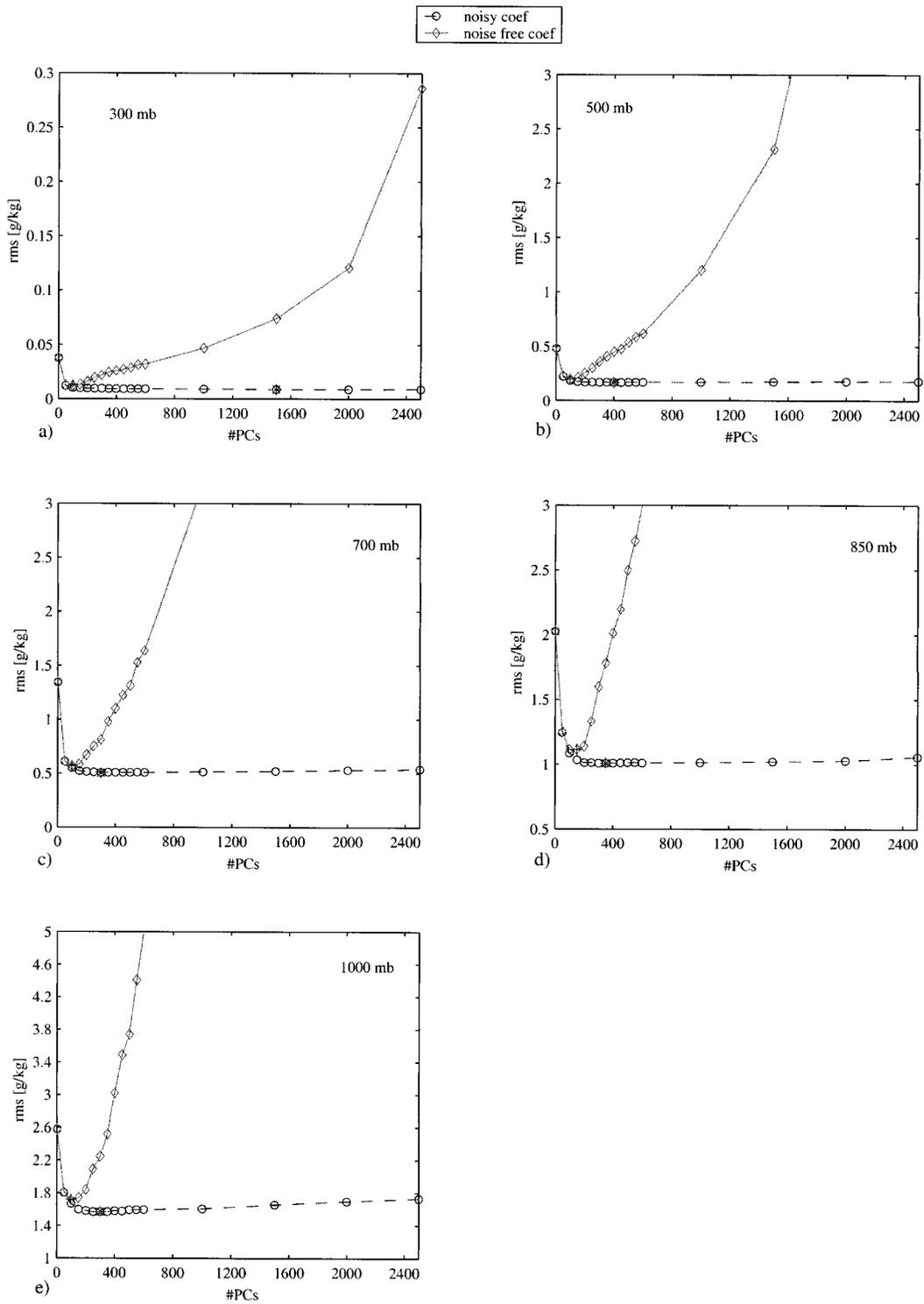
FIG. 13. Water vapor retrieval degradation of noisy and noise-free PCR: (a) 300, (b) 500, (c) 700, (d) 850, and (e) 1000 hPa.
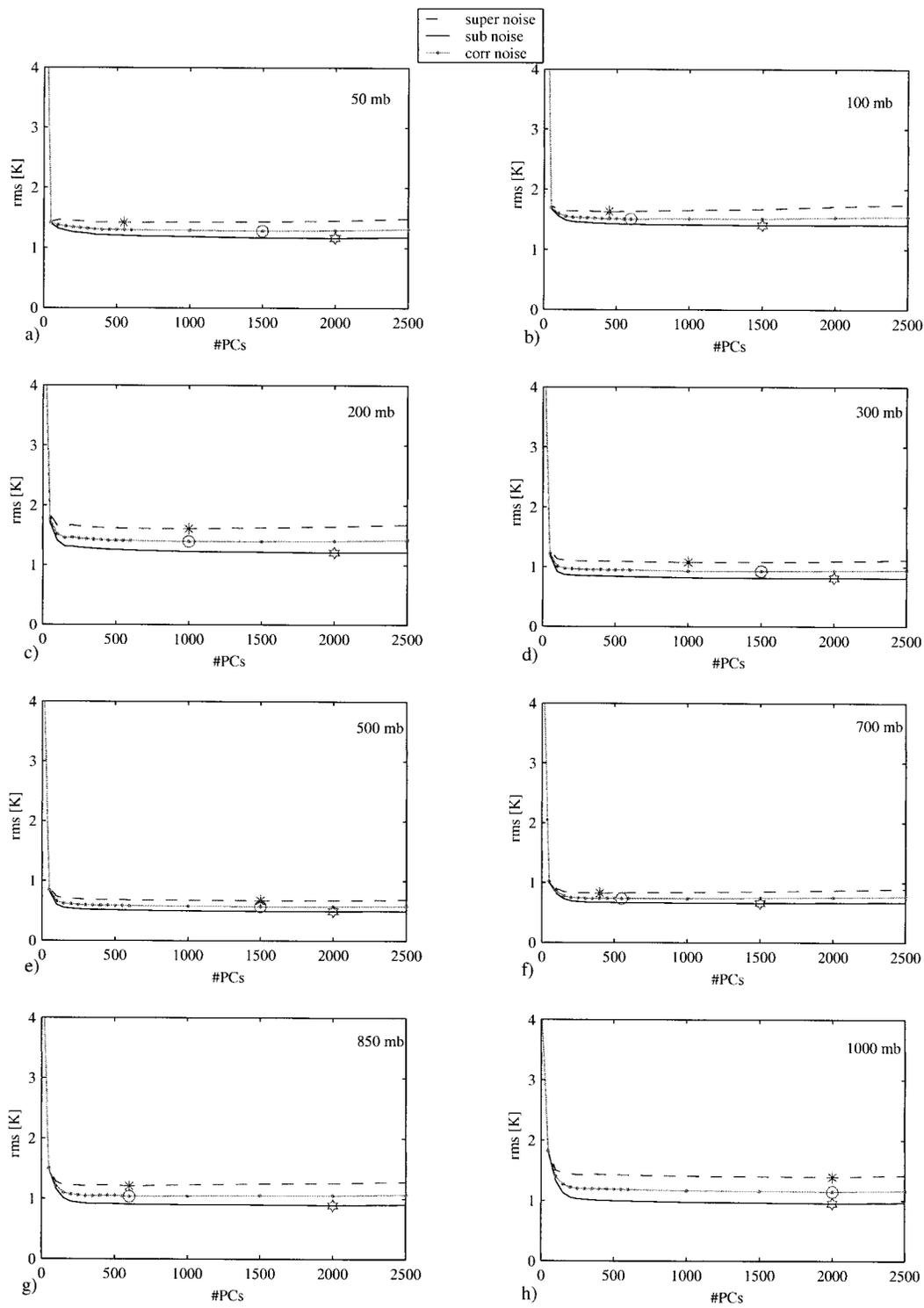
FIG. 14. Temperature retrieval rms error of super-, sub-, and correlated-noise PCR: (a) 50, (b) 100, (c) 200, (d) 300, (e) 500, (f) 700, (g) 850, and (h) 1000 hPa.
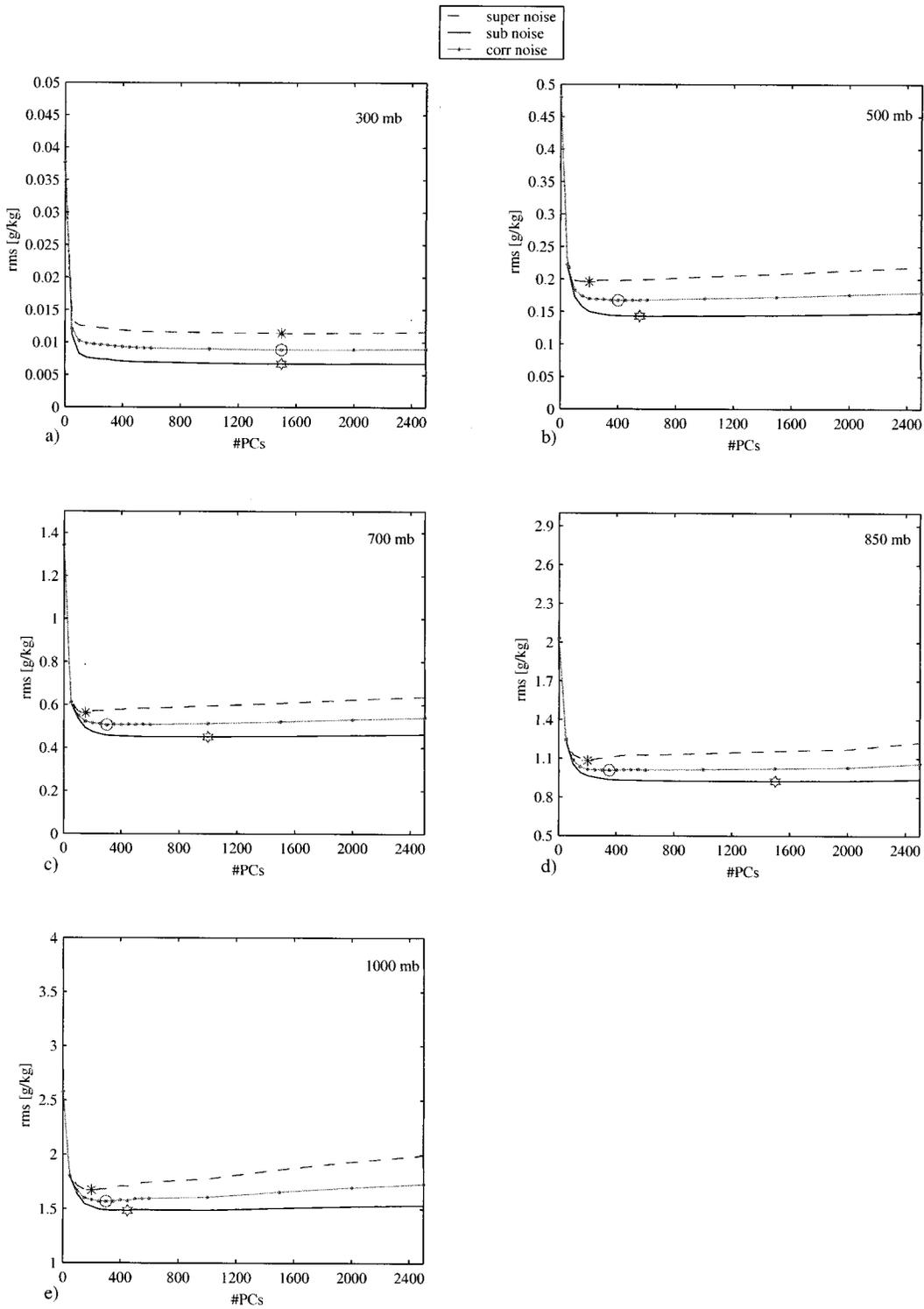
FIG. 15. Water vapor retrieval rms error of super-, sub-, and correlated-noise PCR: (a) 300, (b) 500, (c) 700, (d) 850, and (e) 1000 hPa.
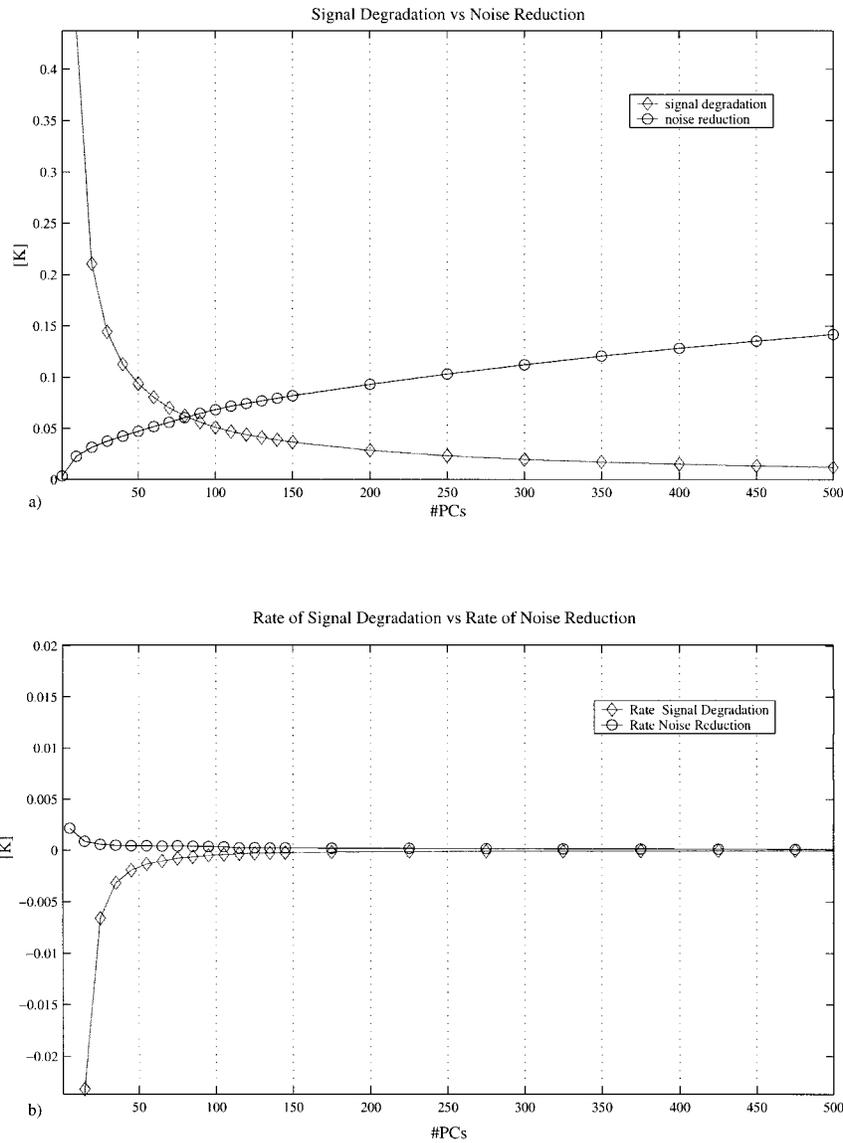
Signal Degradation vs Noise Reduction

Rate of Signal Degradation vs Rate of Noise Reduction

FIG. 16. (a) Brightness temperature signal degradation and noise reduction/reconstruction. (b) Rate of brightness temperature signal degradation and rate of noise reduction/reconstruction.

is almost impossible with real data. It is also possible to understand and to test the feasibility of the methods before applying them to any real observations.

APPENDIX

Derivation of Eq. (8)

The standard deviation of Ω_i times the jth component of the ith PC divided by the standard deviation of the ith component of $S^{(ts)}$ represents the linear correlation between the jth channel and the ith PC. Let us define the correlation between the values of the projections of the spectra onto the ith PC and the values of the brightness temperature for the jth channel as

$$\chi_{i,j} = \frac{1}{M_{tr}\sigma_i\sigma_j} \sum_{k=1}^{M_{tr}} (\Omega_{k,i} - \overline{\Omega_{k,i}})(S_{k,j} - \overline{S_{k,j}}), \quad \text{(A1)}$$
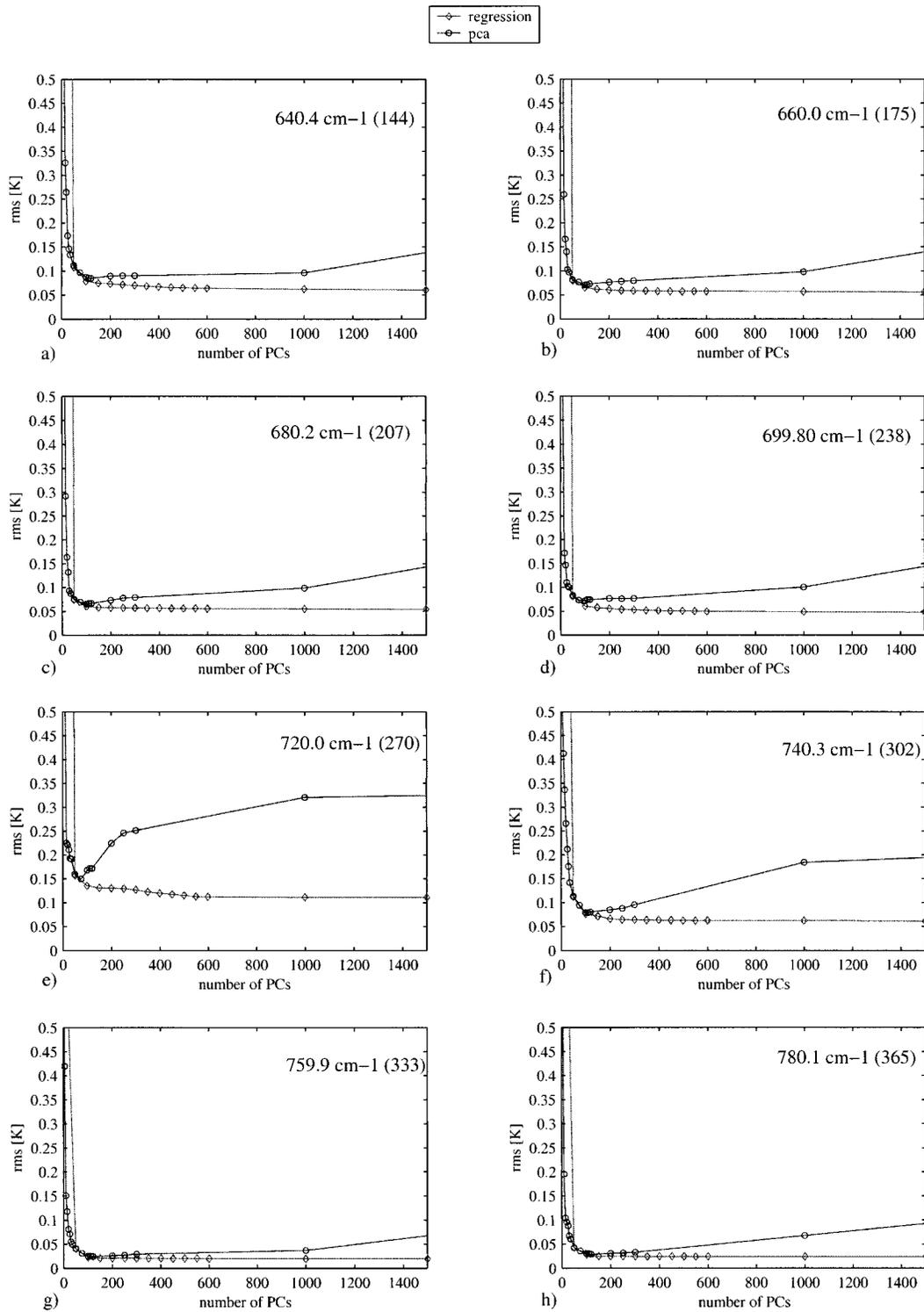
FIG. 17. Reconstruction rms error of linear regression and PCC methods: (a) 640.4 cm^{-1}, (b)–(h) other channels, as labeled.
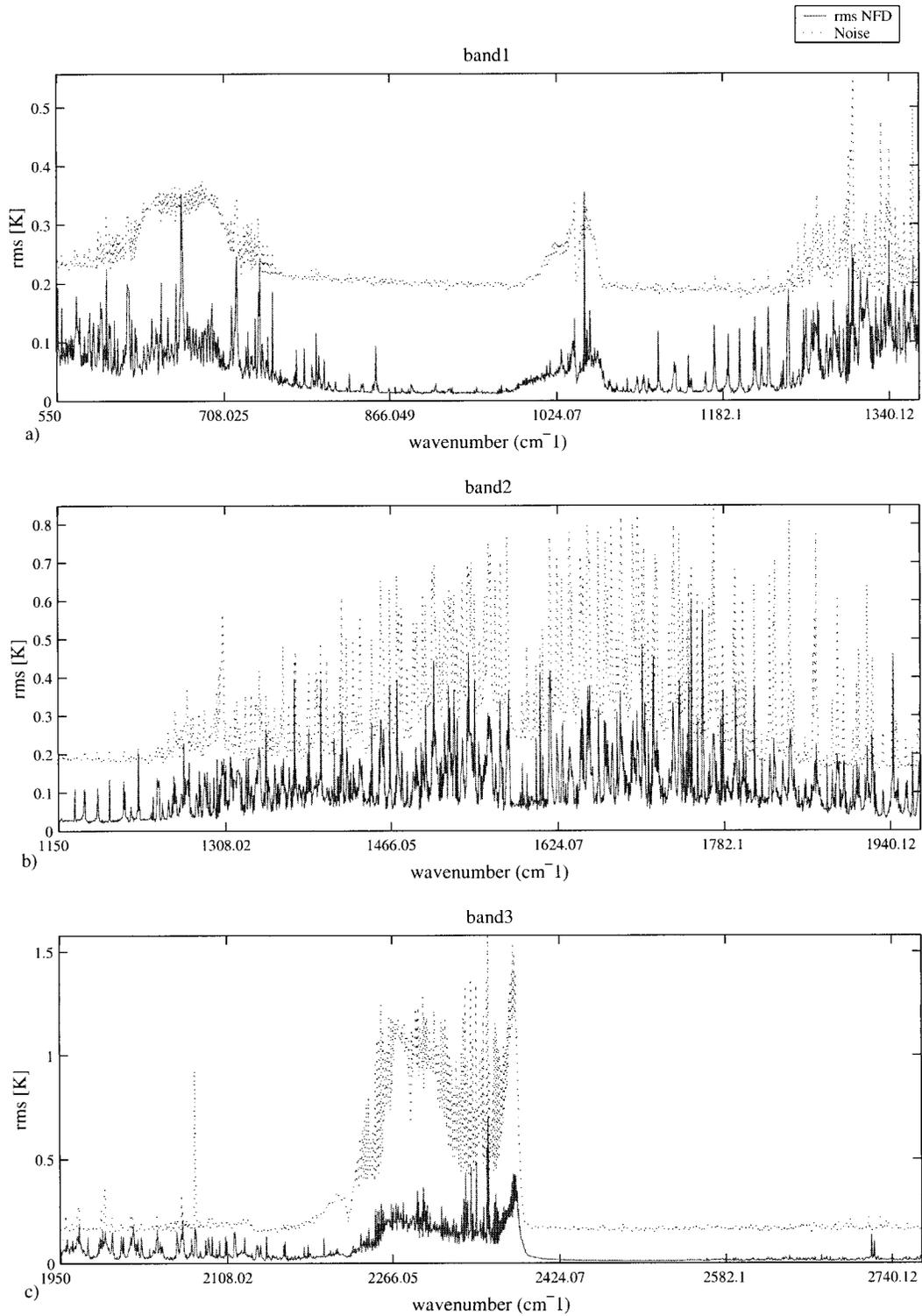
FIG. 18. PCC reconstruction residual rms error of brightness temperature using 250 PCs. Instrument noise rms is also shown: (a)–(c) bands 1 to 3, respectively.

where the overbar indicates the mean over the number of examples M_{tr}. Then, χ may be rewritten as

$$
\chi_{i,j} = \frac{1}{M_{tr}\sigma_i\sigma_j}
$$
$$
\times \sum_{k=1}^{M_{tr}} \left\{ \left[\sum_{l=1}^{N_c} (S_{k,l} - \overline{S_{k,l}})U_{l,i} \right] - \overline{\Omega_{i,k}} \right\}(S_{k,j} - \overline{S_{k,j}}),
$$
(A2)

and, taking into account that $\overline{\Omega_{i,k}} = 0$, the last expression becomes

$$
\chi_{i,j} = \frac{1}{M_{tr}\sigma_i\sigma_j}
$$
$$
\times \sum_{k=1}^{M_{tr}} \left\{ \left[\sum_{l=1}^{N_c} U_{l,i}(S_{k,l} - \overline{S_{k,l}}) \right](S_{k,j} - \overline{S_{k,j}}) \right\}.
$$
(A3)

Because the eigenvectors of **C** (PCs) are orthonormal, we may write

$$
\chi_{i,j} = \frac{1}{M_{tr}\sigma_i\sigma_j}
$$
$$
\times \sum_{k=1}^{M_{tr}} \sum_{p=1}^{N_c} \left\{ \left[\sum_{l=1}^{N_c} U_{l,i}(S_{k,l} - \overline{S_{k,l}}) \right](S_{k,p} - \overline{S_{k,p}})\delta_{p,j} \right\}
$$
$$
= \frac{1}{M_{tr}\sigma_i\sigma_j} \sum_{p=1}^{N_c} \sum_{l=1}^{N_c} U_{l,i}\delta_{p,j}
$$
$$
\times \sum_{k=1}^{M} (S_{k,l} - \overline{S_{k,l}})(S_{k,p} - \overline{S_{k,p}})
$$
$$
= \frac{1}{\sigma_i\sigma_j} \sum_{p=1}^{N_c} \sum_{l=1}^{N_c} U_{l,i}\delta_{p,j}C_{l,p}
$$
$$
= \sum_{p=1}^{N_c} \sum_{l=1}^{N_c} \frac{1}{\sigma_i\sigma_j}U_{l,i}\left(\sum_{\alpha=1}^{N} \lambda_\alpha U_{l,\alpha}U_{p,\alpha}^{T} \right)\delta_{p,j}
$$
$$
= \frac{\sigma_i}{\sigma_j}U_{i,j}.
$$
(A4)

REFERENCES

Cousins, D., and M. J. Gazarick, 1999: NAST interferometer design and characterization: Final report. MIT Lincoln Laboratory Project Report NOAA-26, 159 pp.

Cuomo, V., and Coauthors, 1999: A cloud-detection approach for IASI data. *Tech. Proc. 10th Int. ATOVS Study Conf.,* Boulder, CO, International TOVS Working Group, 595 pp.

Deco, G., and D. Obradovic, 1996: *An Information-Theoretic Approach to Neural Computing.* Springer, 261 pp.

Fischer, H., and H. Oelhaf, 1996: Remote sensing of vertical profiles of atmospheric trace constituents with MIPAS limb-emission spectrometers. *Appl. Opt.,* **35,** 2787–2796.

Huang, H.-L., W. L. Smith, and H. M. Woolf, 1992: Vertical resolution and accuracy of atmospheric infrared sounding spectrometers. *J. Appl. Meteor.,* **31,** 265–274.

Jolliffe, I. T., 1986: *Principal Component Analysis.* Springer-Verlag, 217 pp.

Smith, W. L., and H. M. Woolf, 1976: The use of eigenvectors of statistical covariance matrices for interpreting satellite sounding radiometer observations. *J. Atmos. Sci.,* **33,** 1127–1140.

——, and Coauthors, 1990: GHIS—The GOES High-Resolution Interferometer Sounder. *J. Appl. Meteor.,* **29,** 1189–1203.

——, H.-L. Huang, X. L. Ma, H. M. Woolf, and H. E. Revercomb, 1995: High Resolution Interferometer Sounder—An accurate method for profile retrieval without the use of contemporary "first guess" data. *Optical Remote Sensing of the Atmosphere Topical Meeting,* Salt Lake City, UT, Optical Society of America, 38–40.

——, S. Ackerman, H. Revercomb, H.-L. Huang, D. H. DeSlover, W. Feltz, L. Gumley, and A. Collard, 1998: Infrared spectral absorption of nearly invisible cirrus clouds. *Geophys. Res. Lett.,* **25,** 1137–1140.

Strow, L. L., H. E. Motteler, R. G. Benson, S. E. Hannon, and S. De Souza-Machado, 1998: Fast computation of monochromatic infrared atmospheric transmittances using compressed look-up tables. *J. Quant. Spectrosc. Radiat. Transfer,* **59,** 481–493.

Wark, D. Q., and H. E. Fleming, 1966: Indirect measurements of atmospheric temperature profiles from satellites: I. Introduction. *Mon. Wea. Rev.,* **94,** 351–362.