JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 112, D20301, doi:10.1029/2007JD008391, 2007

Sampling techniques in high-dimensional spaces for the development of satellite remote sensing database

Filipe Aires[1] and Catherine Prigent[2]

Received 3 January 2007; revised 10 May 2007; accepted 17 July 2007; published 17 October 2007.

[1] This study presents various strategies to sample databases from large atmospheric data sets in high-dimensional spaces for satellite remote sensing applications. In particular, two sampling algorithms are examined: the traditional uniform sampling that lists all possible situations and the clustering sampling (K-means) that respects the natural variability probability distribution functions. In order to assess the quality of both sampling methods, the extracted databases are used to extract first guesses for satellite remote sensing schemes. They are also employed as training databases for the calibration of statistical retrieval algorithms. The analysis of these sampling algorithms is illustrated by constructing both a first guess (FG) extraction and a retrieval databases of temperature and water vapor profiles over sea for the Atmospheric Microwave Sounding Unit (AMSU) instrument. The advantages and problems of each sampling approach are thoroughly examined and sensitivity studies are conducted to analyze the impact on the FG extraction and retrieval of various algorithmic parameters such as the distance being used, the size of the databases, or the instrumental noise sensitivity. The K-means clustering algorithm, not yet used for this kind of problems, is very efficient compared to the more traditional uniform sampling approach. It is also shown that it is important to have quasi-automatic and flexible tools that can be used to generate problem-specific databases.

Citation: Aires, F., and C. Prigent (2007), Sampling techniques in high-dimensional spaces for the development of satellite remote sensing database, *J. Geophys. Res.*, *112*, D20301, doi:10.1029/2007JD008391.

1. Introduction

[2] Two main sources of information are used in satellite remote sensing of geophysical parameters: First and foremost, the satellite measurements themselves, second, a priori information that helps constrain the inverse problem. A schematic representation of the data and method modules in a general retrieval algorithm is presented in Figure 1. First guess (FG) information is often used in the initial step of a retrieval process. The FG information can come from a priori knowledge of the situation (e.g., FG provided by a Numerical Weather Prediction (NWP) model) but it can also depend upon the satellite observations (i.e., FG derived from a climatological database using the satellite observations). Databases are developed to optimize their use in retrieval scheme.

[3] In order to derive statistical remote sensing algorithms that retrieve geophysical parameters from satellite observations, it is necessary to develop ''training'' (or learning) database. The training database is used to calibrate the statistical model in charge of the retrieval: Estimating the coefficients of a linear regression retrieval scheme, learning a Neural Network (NN) model [*Aires et al.*, 2001]. or performing Bayesian retrievals [*Kummerow et al.*, 2001]. This database can come from colocalized satellite measurements and in situ observations (i.e., empirical inversion) or it can result from the simulation of the satellite measurements using a radiative transfer model (i.e., physical inversion). This training database can be chosen to describe as well as possible the natural variability probability distributions or it can be chosen to emphasize some data characteristic of particular interest for the application.

[4] FG information can be important for the retrieval scheme since the inversion is sometimes a minimization procedure that modifies iteratively an initial solution. The iterative process is highly sensitive to the FG solution (starting point of the iteration) as the optimization scheme can be trapped in a local minimum. The number of iterations, hence the computer time, is also sensitive to the quality of the FG. Furthermore, this additional information helps eliminate ambiguities when multiple solutions are possible (i.e., nonuniqueness of the inverse problem) [*Tarantola*, 1987]. In the framework of variational assimilation in operational centers, the FG is provided by the NWP (a priori information). This paper will focus on FG that depend essentially upon the satellite observations so that the resulting satellite retrievals can more easily be used for model validation. These FG databases are defined both on the geophysical variables and on the satellite measure-

[1]Laboratoire de Météorologie Dynamique, Institut Pierre-Simon Laplace/Centre National de la Recherche Scientifique, Université de Paris VI/Jussieu, Paris, France.

[2]Laboratoire d'Etudes du Rayonnement et de la Matière en Astrophysique, Centre National de la Recherche Scientifique, Observatoire de Paris, Paris, France.

0148-0227/07/2007JD008391$09.00

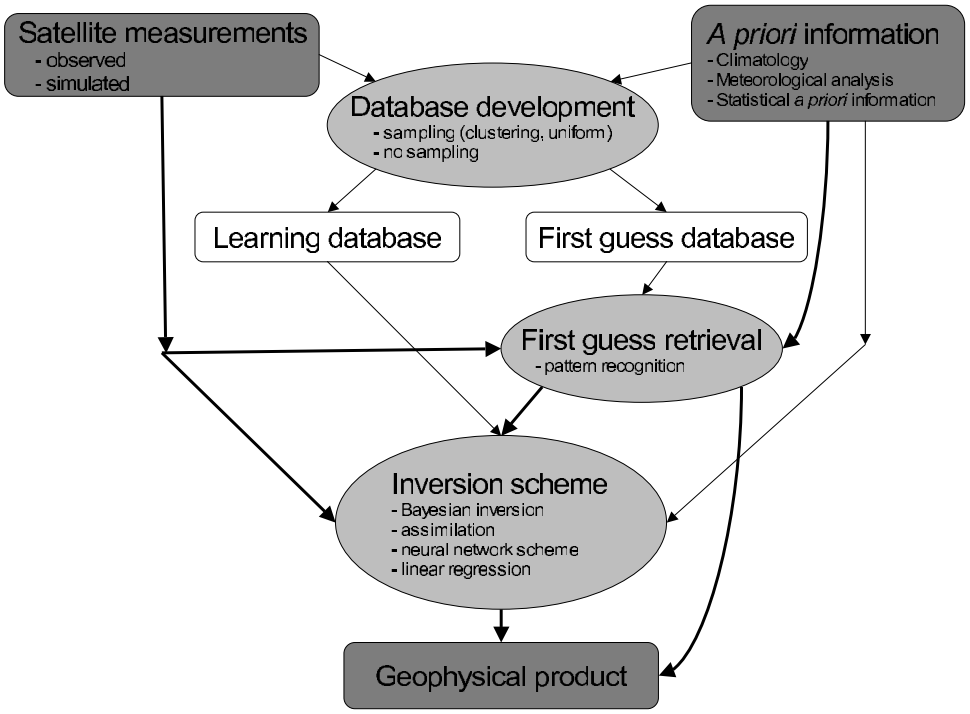

Figure 1. Schematic representation of the data and method modules in a general retrieval algorithm. Thinner arrows are for the preliminary development stage, and bold arrows describe the data flow in the operational mode.

ment spaces so that a pattern recognition procedure can link the satellite measurements to the geophysical FG.

[5] In the meteorology/climatology field, there is generally a profusion of data. One can use the full original data set as the training database. For example, *Aires et al.* [2001] used a full year of global observations to train a NN retrieval scheme. However, for other applications, it might be difficult to perform radiative transfer simulations for millions of samples and for multiple conditions such as various scanning angles or surface properties, especially with the new generation of instruments that can include thousands of channels (e.g., the "Improved Atmospheric Sounding in the Infrared" is an ESA instrument with 8461 channels that has been launched in October 2006 on board the METOP mission). Furthermore, for some specific applications, a careful monitoring of the data set to be used can be necessary (e.g., some regimes of particular interest might be emphasized). In these cases, an adequate sampling of the initial database will select numerous and diverse samples representing most of the variability of the original large data set, but limiting the number of samples to a manageable size. For many remote sensing applications, the problem is often to perform a pertinent sampling in high-dimensional spaces, to generate adequate FG or training databases. The optimal sampling algorithm is highly dependent on the application.

[6] The Thermodynamical Initial Guess Retrieval (TIGR) database [*Chédin et al.*, 1985; *Escobar*, 1993; *Chevallier et al.*, 1998] is a good example of such data sets used for FG retrieval or training databases. TIGR contains 2311 atmospheres sampled from a large ensemble of radiosondes. For recent applications, TIGR has been improved to account for specific issues: For surface temperature variability [*Aires et al.*, 2002a] or for a database composed of reanalysis from the ECMWF (European Centre for Medium range Weather Forecast) [*Chevallier et al.*, 2000]. For recent instruments, and complex applications using their synergy, there is a need for easy-to-use, flexible, and efficient tools to construct specifically designed databases. An example of such technical developments is given by *Vrac et al.* [2005].

[7] Several technical issues will be discussed in this paper. Which distance should be used? Which sampling algorithm should be chosen: Should the FG or training databases be statistically representative of the original data set variability or should the samples be uniformly distributed? How to deal with the correlations among variables? On which space should the sampling be done: In the space of the geophysical variables or in the space of the satellite observations? With a good understanding of the answers to these questions, it becomes possible to define semiautomatic sampling algorithms that can easily evolve to different time/space frames, to specific conditions (i.e., adapted to a physical process or a set of geophysical variables), or to particular instruments. No direct comparisons with existing FG or training databases will be intended, the focus being a better understanding of the algorithmic aspects so that specific and adequate databases can be designed for each new application.

[8] The original data set on which the sampling algorithms are tested is presented in section 2. Section 3 introduces the various technical concepts, in particular the two main sampling algorithms used in this study. Applica-

tion to the generation of a FG database is discussed in section 4, together with various sensitivity studies. Section 5 focuses on the development of training databases. Conclusions are drawn in section 6.

2. Original Data Set

2.1. ERA40 Geophysical Data Set

[9] A large data set of radiosondes can be used to build the original data set. However, first these radiosondes are not uniformly distributed over the globe, second they can require significant quality control and interpolation/extrapolation work to be used as inputs to radiative transfer simulations. In this study, reanalysis from NWP models are adopted. In a variational assimilation context, the horizontal and vertical resolution appropriateness and the compatibility of the atmospheric situations with the model are essential.

[10] In the following experiments, we use a data set composed of the 1999 ECMWF ERA40 reanalyses [*Simmons and Gibson*, 2000] over the globe. Only data over ocean under cloud-free conditions are considered. Data with relative humidity less than 1% or higher than 100% are discarded. Over the 1-a period, this results in an original database of more than 4 million points. We randomly select more than 140.000 points from the original data set in order to keep our multiple experiments manageable. However, other independent data sets used for validation purpose are also extracted from the same initial database.

[11] In order to run radiative transfer simulations, the necessary information is kept for each point in the data set but only the temperature (in Kelvin), water vapor (relative humidity in percentage) and ozone (in kg/kg) profiles are used for the sampling experiments. We define the "*GEO*-space" to be the space of these three geophysical profiles (dim $n = 3$ variables $\times$ 23 levels = 69). We will focus on the temperature and the water vapor profiles since these are classical and interesting geophysical variables monitored by the considered satellite instrument.

2.2. AMSU Observations

[12] The AMSU-A and AMSU-B on board the latest generation of the National Oceanic and Atmospheric Administration (NOAA) polar orbiting satellites measure the outgoing radiances from the Earth atmosphere and the surface. AMSU-A is designed to retrieve the atmospheric temperature from about 3 hPa ($\simeq$45 km) down to the Earth's surface. It has 12 channels located close to the oxygen absorption lines below 60 GHz and four window channels at 23.8, 31.4, 50.3, and 89 GHz. AMSU-B is used for atmospheric water vapor sounding and makes measurements in three channels in the vicinity of the strong water vapor absorption line at 183 GHz and in two window channels at 89 and 150 GHz. The two instruments have instantaneous fields of view of 3.3° and 1.1° and sample 30 and 90 Earth views respectively. The AMSU observation scan angle varies from $-48°$ to $+48°$ with the corresponding local zenith angle reaching 58°. A detailed description of the AMSU sounders is reported by *Goodrum et al.* [2000].

[13] The use of AMSU measurements in operational Numerical Weather Prediction (NWP) models can provide accurate monitoring of both air temperature and moisture profiles with good temporal and spatial sampling. Compared to infrared sounding measurements, AMSU observations are less sensitive to high thin and nonprecipitating clouds. Several retrieval techniques have been developed for temperature and/or humidity sounding with AMSU-A/AMSU-B and other microwave radiometer measurements. *Rosenkrantz* [2001] used surface and atmosphere modeling to retrieve temperature-moisture profiles from AMSU-A/AMSU-B data. A NN technique has been used by *Shi* [2001] to estimate air temperature profiles from AMSU-A; a similar technique has been utilized by *Franquet* [2003] for the retrieval of water vapor. Over land however, the AMSU measurements are not fully exploited. Efforts to assimilate AMSU radiances over land are underway in several NWP centers.

[14] We use the RTTOV radiative transfer code to simulate AMSU-A and AMSU-B observations based on the ERA40 description of each atmosphere in the data set presented in section 2.1. The 23 pressure levels of the ERA40 reanalyses are actually interpolated to the 43 levels used by RTTOV. RTTOV is a fast radiative transfer model originally developed at ECMWF [*Eyre*, 1991] and that is now supported by the EUMETSAT NWP-SAF (Satellite Application Facility) [*Saunders et al.*, 1999; *Matricardi et al.*, 2001]. The model allows for rapid simulations of radiances for satellite infrared and microwave radiometers given an atmospheric profile of temperature, variable gas concentrations, cloud and surface properties, referred to as the state vector. Numerous platforms and sensors are supported.

[15] The "*TB* space" (TB is for brightness temperature) is defined as the space with 20 coordinates corresponding to the 15 channels from AMSU-A and the five channels from AMSU-B (see section 2.2). This "*TB* space" is the "dual" of the *GEO* space defined in section 2.1.

3. Sampling Techniques

3.1. Principal Component Analysis

[16] Principal Component Analysis (PCA) of both geophysical variables (*GEO*) and brightness temperatures (*TB*) are used here to perform a preliminary analysis of variables and observations. Furthermore, the PCA will be useful for the Mahalanobis distance introduced in section 3.2.

[17] An important point concerns the normalization of data before performing the PCA. This normalization should be applied very carefully. If two original coordinates are as important but have a different variability range, no normalization means that the PCA will emphasize the coordinate with the larger variability to the detriment of the other one. On the contrary, normalization can be dangerous if one data coordinate is not informative (e.g., measurement noise exceeding the information) and has a small variability range: Normalization would put weight onto this variable and less onto the valuable ones, which would deteriorate the PCA results. Furthermore, it should be pointed out here that using a unit for a geophysical variable is already a normalization choice.

[18] In this application, the variables in the *GEO* space are the temperature, water vapor, and ozone atmospheric profiles in the 23 discrete vertical levels. These pressure levels correspond to the vertical levels used by the radiative

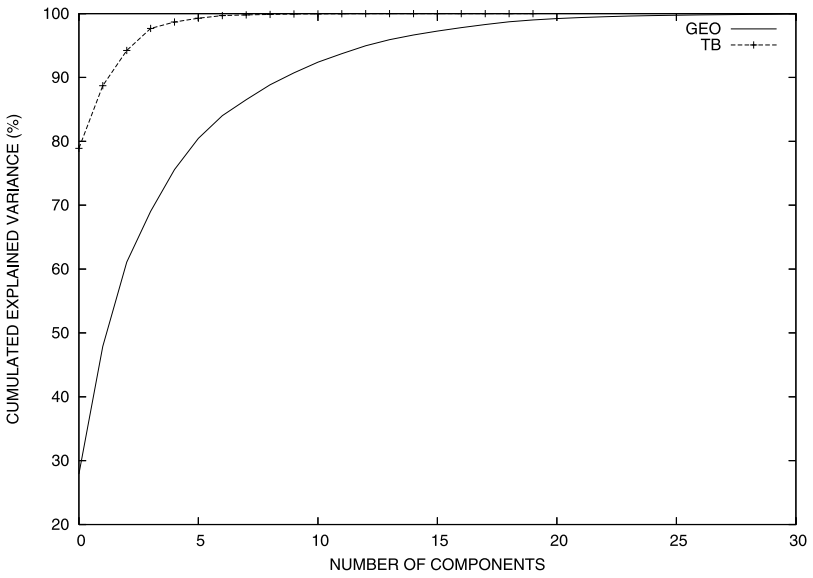

Figure 2. Percentage of variance explained by the first PCA components in the *GEO* (solid line) and the *TB* spaces (dashed lines).

transfer code we will be using. It is decided not to normalize these variables. Ozone with its small variability range (in kg/kg) will have a very limited impact on the PCA, meaning that the ozone profile will be well represented by the PCA only if it is highly correlated with the temperature or water vapor profiles. Since temperature and water vapor have variability ranges of the same order of magnitude (temperature in Kelvin and relative humidity in percentage), they have a comparable impact on the PCA results. Not normalizing here limits the impact of high-altitude water vapor which is not the priority in our application.

[19] The cumulative percentage of variance explained by the first components is presented in Figure 2 for the PCA of geophysical variables (n = 69) describing the atmosphere and for the associated brightness temperatures (n = 20). The results of the PCA in the *GEO* space are dependent on the resolution of the data (from the ERA40 reanalyses in our case). ''Resolution'' refers here to the intrinsic variability of the data, not just the number of discretization of the vertical. With the data sets presented in section 2, there are more degrees of freedom in the *GEO* space than in the *TB* space, meaning that the satellite observations are insufficient to characterize completely the *GEO* space, even when no observation noise is added. This characterizes an underdetermined (i.e., underconstrained) inverse problem. Nonetheless, if seven PCA components can be extracted from the satellite observations and can ''perfectly'' be related to the first seven PCA components of the GEO variables, the retrieval would be able to represent about 90% of the *GEO* space variability [*Tarantola*, 1987; *Rodgers*, 2000].

[20] The errors made by the PCA representation are shown in Figure 3 for various compression levels (i.e., number of PCA components kept). For $n' = n = 69$, all the components remain, which means that there is no compression and the PCA representation is perfect. The error increases when components are suppressed in the PCA representation, but it can be seen that a representation with 40 components is sufficient for the temperature profiles. About 30 components can describe the water vapor profiles correctly. This means that the temperature and the water vapor profile have about 40 and 30 degrees of freedom respectively. 40 components are enough to represent simultaneously the temperature and water vapor profiles. These numbers are consistent with Figure 2. It is obviously impossible to retrieve these 40 components from the limited number of components in the *TB*s. This simple PCA analysis gives directly an assessment of the upper limit information content of the satellite measurements represented by the best case scenario because the first *TB* PCs can be associated to more *GEO* PCs or to *GEO* PCs of higher rank. For a complete information content analysis, the reader is referred to *Thépaut and Moll* [1990].

3.2. Distances

[21] In order to sample a data set, a good distance measure is required to measure similarities and differences between two points in this database. To extract relevant samples, an appropriate distance has to be selected.

[22] The Euclidean distance is widely used in statistics or in physics:

$$D_E(y^0, y) = \left[(y^0 - y)^t \cdot (y^0 - y)\right]^{\frac{1}{2}}.$$

It gives the same weight to each y's coordinate so that coordinates with higher variability ranges will drive the distance. The weighted Euclidean distance is given by:

$$D_W(y^0, y) = \left[(y^0 - y)^t \cdot W \cdot (y^0 - y)\right]^{\frac{1}{2}}.$$

where W is a diagonal matrix with the elements W_{ii} giving the relative importance of the ith coordinate. The weighted Euclidean distance allows to define a priori the weight of each of y's coordinates.

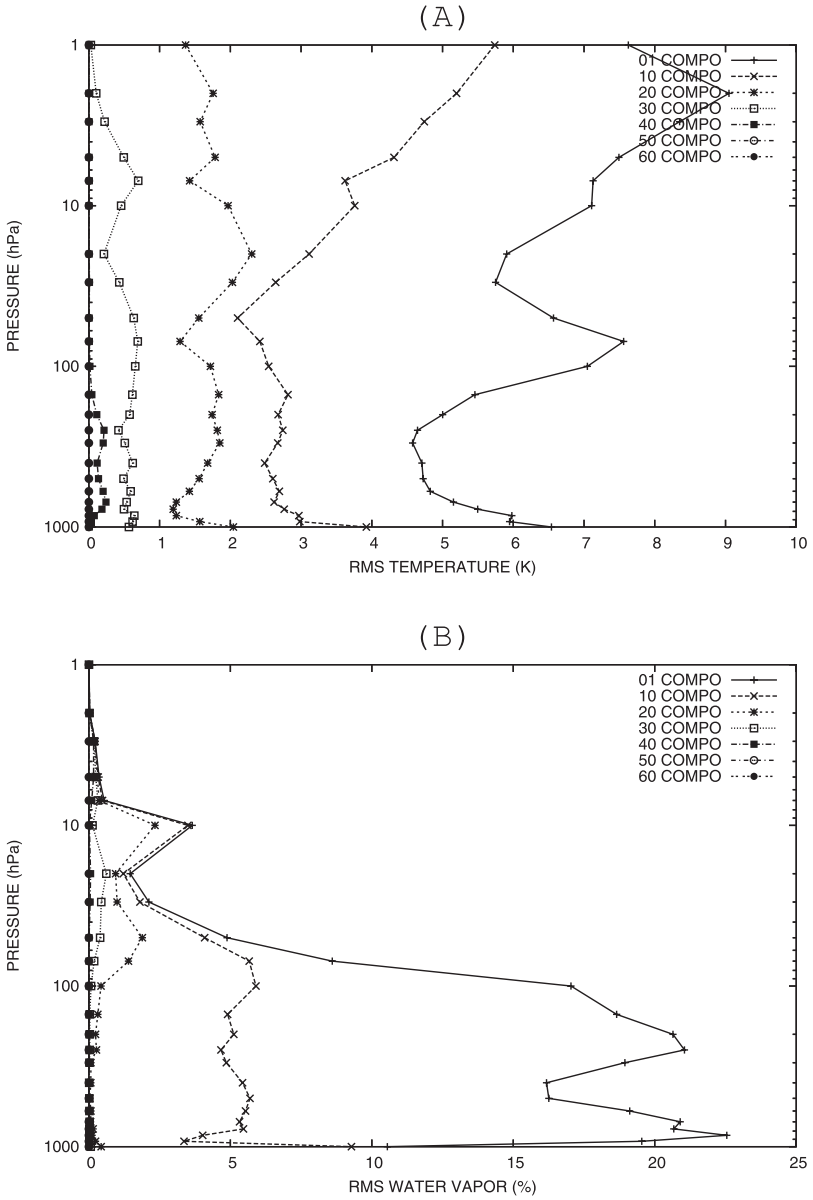

Figure 3. Error of the PCA representation for (a) the temperature and (b) the water vapor atmospheric profiles for an increasing number of PCA components from 1 to 60.

[23] Another important distance is the Mahalanobis distance, e.g., [*Crone and Crosby*, 1995]:

$$D_M(y^0, y) = \left[(y^0 - y)^t \cdot \Sigma_y{}^{-1} \cdot (y^0 - y) \right]^{\frac{1}{2}}.$$

By exploiting the correlations among variables described in the covariance matrix Σ_y, the Mahalanobis distance gives less weight to variables with high variance and groups of variables highly correlated in order to optimize the use of each available independent piece of information. The link with the weighted Euclidean distance is clear: If y's coordinates are independent $W = \Sigma_y^{-1}$ and the weights $W_{ii} = 1/\sigma_i^2$. We propose here to use an Euclidean distance based on the first n' PCA components, h (see section 3.1). This Truncated-PCA Euclidean (TPCAE) distance would be equivalent to the Mahalanobis distance if we used all the PCA components ($n' = n$) [*Jolliffe*, 2002] (i.e., in that case, the covariance matrix Σ_y becomes the identity matrix). Using fewer components removes irrelevant information and produces a faster pattern recognition step. This TPCAE distance was used, for example, by *Aires et al.* [2002b] for a remote sensing retrieval problem with first guess. This TPCAE distance benefits from the correlation structure to optimally weight the data coordinates. However, using this TPCAE distance can be computationally expensive (i.e., not only the FG or learning database needs to be in the PCA space, but all new satellite observations need to be projected). If only a pattern recognition is needed and the optimal weighting of the data coordinates is not essential in the application of interest, using the TPCAE distance can be superfluous and inefficient. For more complex and inhomogeneous data, the TPCAE distance is quite interesting. As long as enough components are used, the number N

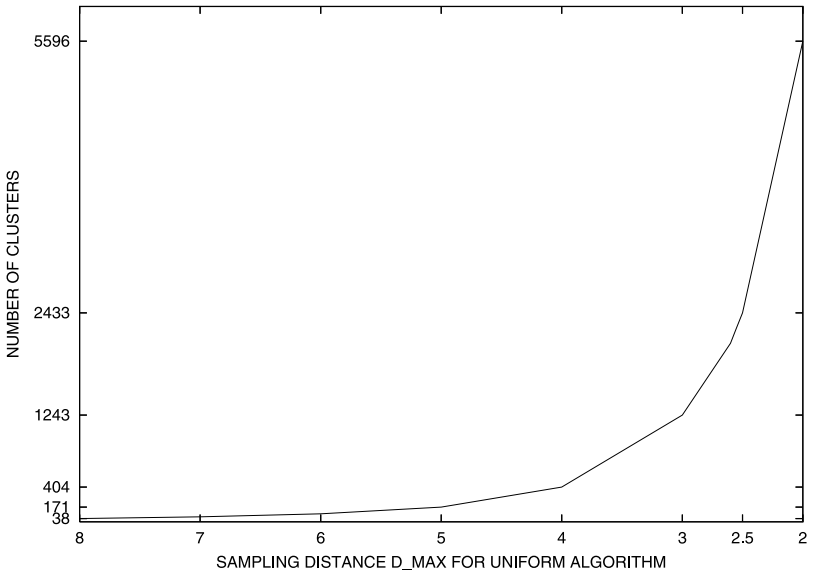

Figure 4. Number of prototypes extracted by the uniform sampling algorithm in the *uniGEO* for decreasing D_{MAX} (i.e., increasing precision).

of components kept in the distance is not too critical: Since the TPCAE distance takes into account the importance of each components, adding higher-order components has no impact on the distance, except to slow down the computation. In practice, it is recommended to be conservative with n' as long as the computations can be dealt with, especially if the rare events are of interest. In the following, the PCAE distance is used for each sampling methods with 40 components for the *GEO* space and 15 for the *TB* space.

[24] Rare events are not the purpose of this paper, but it is important to mention the Minkowski distance [*Huber*, 1981] that can emphasize the extraction of rare event at a low algorithmic cost.

3.3. Data Space

[25] On which data space should the sampling applied, the *GEO* or *TB* space? There are advantages to work on the *GEO* space. First, it is easier to extract geophysical distributions close to the natural ones. Second, since inverse problems suffer from the nonuniqueness of the solution, if there are ambiguities (i.e., two different geophysical situations associated to a *TB* measurement), sampling in the *GEO* space will help represent this nonuniqueness in the extracted databases. Third, controlling the importance of a specific geophysical variable in the sampling process is easier when sampling directly in the *GEO* space.

[26] On the other hand, sampling in the *TB* space can also be attractive. First, *TB* measurements are more or less homogeneous so that a normalization among them is often unnecessary and if required, it is usually easier to perform. Second, the sampling is applied directly on the space used to perform the FG extraction, the *TB* space, and in the information theory sense, the sampling is optimal in this *TB* space. Sampling of geophysical prototypes that cannot be discriminated in the *TB* space is irrelevant since the FG extraction works directly on the *TB*s. Sampling in the *TB* space ensures that we sample optimally the available information for the FG extraction or the training of the retrieval algorithm. Again, sampling in the *GEO* or the *TB* space is a choice that depends upon the use of the extracted database.

3.4. Uniform Sampling

[27] The uniform (or topological) sampling algorithm can sample a number of prototypes in a high-dimensional data set. The different steps are as follows:

[28] 1. Choose randomly a sample in the original data set. This sample is the first prototype.

[29] 2. Take the following sample in the data set. If this sample is at a distance larger than an a priori chosen threshold D_{MAX} from the set of previous prototypes, it becomes a new prototype. Otherwise, the sample is discarded.

[30] 3. Repeat step 2 until all the samples in the original data set have been processed.

[31] This algorithm produces a set of prototypes. Each sample from the original data set can be associated to its closest prototype that defines clusters of points. This algorithm is used for example by *Achard* [1991], *Escobar* [1993], or *Chevallier et al.* [2000].

[32] The maximum distance, D_{MAX}, between each sample of the original data set and any of the prototypes indicates the ''precision'' of the prototype representation: The lower D_{MAX} and the closer a prototype to any sample in the data set. The number of extracted prototypes depends on D_{MAX} but it cannot be determined a priori: Several experiment needs to be conducted to tune D_{MAX} to provide a number of prototypes close to what is desired. In Figure 4, the number of prototypes extracted by the uniform sampling algorithm using the Euclidean distance on the *GEO* space is represented against the D_{MAX} threshold. The number of required prototypes grows exponentially with decreasing D_{MAX} threshold (i.e., higher precision). As a consequence, it becomes rapidly impossible to exceed a certain level of precision, given the fact that the number of prototypes cannot be indefinitely increased. This is a direct consequence of the curse of dimensionality discussed in section 7.

[33] The uniform sampling is sensitive to the order of presentation of the samples in the original data set when the number K of extracted prototypes is relatively small. This is less a problem for large K. A random reordering in the original data set is nonetheless encouraged to perform a more robust sampling and to obtain better uniformity in the variables not included in the clustering process (such as the date of the situation).

[34] This algorithm can be computationally demanding when extracting a large number of prototypes. The curse of dimensionality is a problem for the number of prototypes and it also affects the computation time to estimate them. The number of distances to be computed during the sampling process is about $K \times (E + 1)/2$. It grows linearly with K, the number of prototypes, and with E, the number of samples in the original data set. Since the number of prototypes K increases exponentially with an increasing precision (i.e., decrease of D_{MAX}, see Figure 4), the number of computations increases also exponentially with precision. *Chevallier et al.* [2007] use a ''preliminary filtering phase'' to solve this problem.

[35] It can also be noted that the final solution (i.e., the ensemble of extracted prototypes) is constructed piece by piece, one prototype after another. It is known in optimization theory that such ''local'' solutions are less optimal than global solutions where the whole solution is estimated at once. This is particularly true for small K, but is not a concern anymore for a large number of extracted components.

3.5. Sampling Using Clustering

[36] The K-means algorithm [*Lloyd*, 1992] is an example of clustering method and it is selected here to sample a large and high-dimensional data set (any other clustering algorithm could be used instead). This clustering method has been used in the atmospheric science disciplines for example in the work by *Desbois et al.* [1982] for the classification of clouds. In the works by *Jakob et al.* [2005] and *Gordon et al.* [2005], K-means were used to relate radiative, cloud, and thermodynamic properties and to validate models with observations. *Omar et al.* [2005] use clustering to classify aerosol measurements and *Cordisco et al.* [2006] classify satellite observations in snow types. The K-means algorithm steps are simple:

[37] 1. First, K prototypes are selected in the data space to be clustered (*GEO* or *TB* spaces in our case). They are generally randomly chosen uniformly in the original data set.

[38] 2. Assign each sample of the data set to its closest prototype by using the data distance. This cluster allocation determines K clusters of points.

[39] 3. When all samples have been assigned, calculate the mean of the K clusters. These cluster centers become the new prototypes. To add stability, a ''learning rate'' can be adopted so that the new mean is a linear combination of the previous mean and its new estimate [*Moody and Darken*, 1989].

[40] 4. Repeat steps 2 and 3 until the convergence is reached. A criterion checks this convergence: In our case, the training phase is stopped when the relative change in the prototypes is small. This is done by monitoring the relative change curves.

[41] The K-means clustering presents several problems. First, the results depend on the initial values of the prototypes, and suboptimal partitions can be found. The standard solution is to start with few different starting points and check the robustness of the outputs. Second, a cluster can happen to be empty during the sampling process (i.e., no samples in the original data set are closer to its associated prototype), so that the prototype cannot be updated. This is a problem that must be overcome. Third, results depend upon the metric being used. We provided in section 3.2 few possible distances, each one with its own advantages and inconveniences. Lastly, the results also depend upon K, the number of extracted prototypes.

[42] A major difference between the uniform and the clustering sampling is that the later one obtains a database with distributions closed to the original data set distributions. This is not the case for the former one that obtains by definition more ''uniform'' distributions. A popular sampling algorithm that also respect the original data set distributions is the ''random'' sampling used for example in Monte Carlo techniques. Quasi-random sampling is also often used [*Press et al.*, 2002], it correspond to a sampling where the sample points are maximally avoiding each other for a better sampling efficiency and clustering sampling uses a similar idea. Obtaining realistic distribution can be an advantage: for example to perform Bayesian statistics, it might be preferable to use distributions that respect the natural variability [*Gelman et al.*, 2003]. Uniform distributions can also be positive: the statistical weight of rare events is increased in this case and this can beneficial for the calibration of an inversion scheme since rare events can be more informative than more frequent situations redundant in the training database.

[43] Another advantage of clustering is that the number of clusters K is defined a priori, contrarily to the uniform sampling approach. This is a disadvantage for specific applications but not for sampling problems. For a discussion on the number of prototypes, see *Milligan and Cooper* [1985], *Mimmack et al.* [2001] or *Rendell and Whitehead* 2003]. The problem of ordering the samples does not happen in the clustering sampling approach since initial prototypes are moved during the clustering toward the most populated parts of the space.

[44] If aberrant data are present in the original data set, the uniform sampling will extract them as prototypes since they are different from the other data. These unrealistic situations being most of the time ''isolated'' in the original data set, they will not be extracted as prototypes by clustering algorithm. Outliers in a FG data set are not a key issue: They will slow down the FG extraction but will be rarely selected as FG. On the other hand, they can be detrimental in a training database since they will impede the training of the retrieval scheme, even if some special learning algorithms have been designed to deal with outliers, i.e., robust statistics [*Moore and McCabe*, 2006]. Since we use ERA40 data for the application in this paper, we limit the possibility of such aberrant data.

[45] To evaluate the clustering algorithm, we use again the 1-a ERA40 data set described in section 2.1. The number of prototypes K is chosen equal to nine. The number nine was chosen as a good compromise between enough classes to test the method, but keeping the number

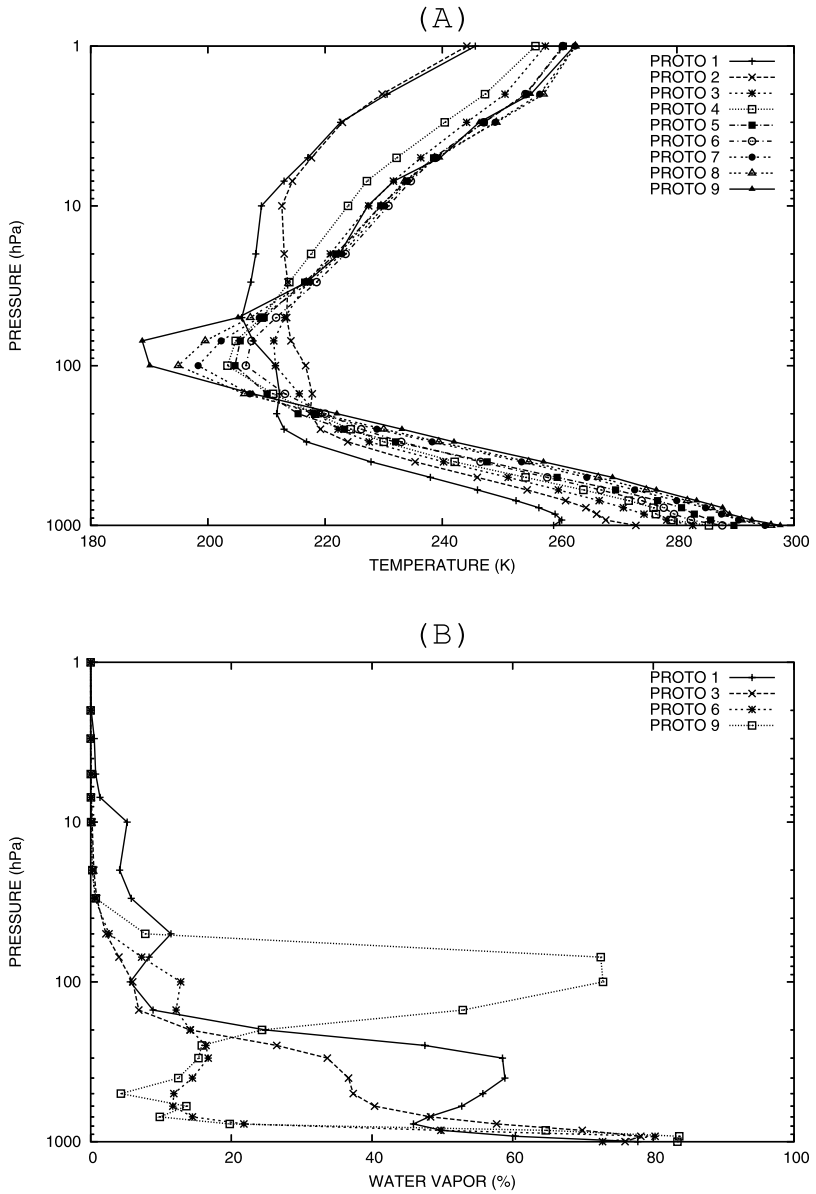

Figure 5. (a) Temperature and (b) water vapor atmospheric prototype profiles extracted in the *clustGEO* experiment for $K = 9$ clusters. For clarity purposes, only four prototypes are represented for water vapor in Figure 5b.

low enough to be able to interpret the extracted prototypes. The tools developed here can be used to conduct a climatological study [*Jakob et al.*, 2005]. Such classification with a small number of clusters can also be used to develop specific retrieval schemes for each cluster (see *Chédin et al.* [1985] for such an approach in the 3I algorithm in which each class is an air mass).

[46] After convergence, it is possible to represent the $K = 9$ extracted prototypes (Figure 5). Since the clustering has been applied to the *GEO* space of the temperature, water vapor and ozone profiles, the dimension of the prototypes is $3 \times 23 = 69$. However, the prototypes are shown only for the temperature and water vapor profiles that are of interest in our particular application. Prototypes 1 and 2 are polar situations, with low surface temperatures and low tropopause levels around 250 hPa. Following clusters from 3 to 9 have a tropopause close to 100 hPa and become warmer at the surface with a maximum around 298°K. Only 4 water vapor profile prototypes are presented in Figure 5 for clarity. Their humidity peak varies: The polar prototype 1 reaches 60% between 200 and 300 hPa, prototypes 3 and 6 have water vapor mostly in the surface layers, and the tropical prototype 9 has a 70% water vapor content at a higher altitude, around 100 hPa.

[47] No ordering is imposed in the clusters by the classification algorithm. However, for the clarity of the presentation we reorganized the clusters with an increasing mean Total Column Water Vapor (TCWV) on the successive extracted clusters. This cluster reordering was performed after the clustering is finished but it could also be imposed during the actual clustering. See Kohonen map algorithm for a clustering with imposed ordering of the

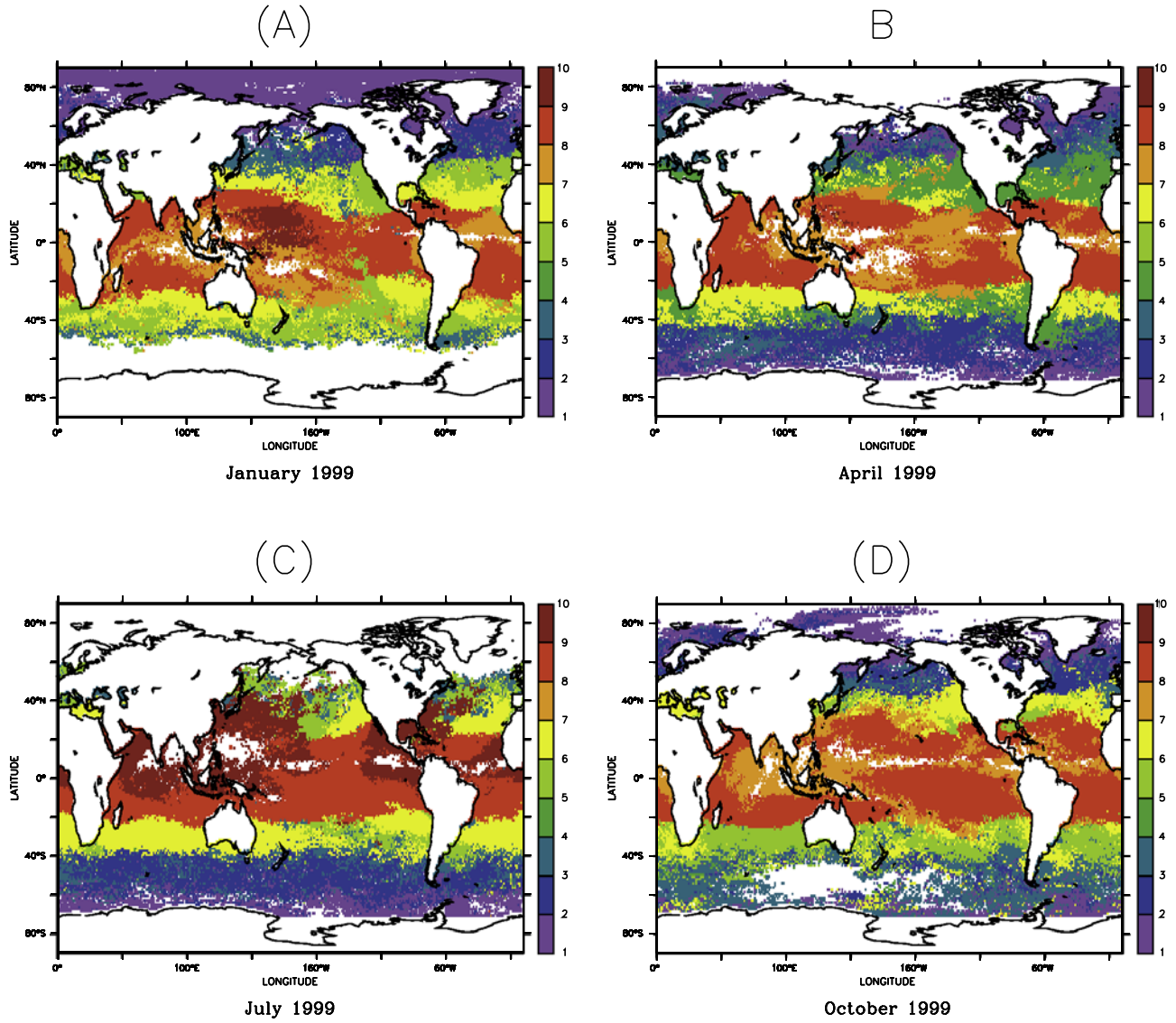

Figure 6. Maps of the most frequent prototypes in (a) January, (b) April, (c) July, and (d) October 1999 for the *clustGEO* experiment with $K = 9$ clusters.

classes [*Kohonen*, 1984]. After reordering, clusters have the following mean TCWV (in %): 4.2, 7.0, 16.8, 17.7, 18.2, 18.4, 28.4, 31.2, 38.8. The mean surface temperatures can also be used to order the clusters, but a strong link exists anyway with the TCWV content. The clusters are organized from the drier, colder atmospheres (i.e., polar atmospheres) to the hotter and more humid atmospheres (i.e., tropical cases). Figure 6 illustrates the spatial distribution of the clusters. In Figures 6a, 6b, 6c and 6d, the more frequent clusters are presented in each pixel during January, respectively April, July, and October 1999. White pixels correspond to locations with no clear situation during the month. The clusters are organized in latitudinal bands, from polar, temperate, to tropical situations which confirms the air mass description of the prototypes in the previous paragraph. The prototypes are well spread in latitude meaning that the sampling process used efficiently the atmospheric situations from the poles to the tropics. The climatological features are coherent, in particular, the Intertropical Convergence Zone pattern is correctly characterized. This experiment with a low number of extracted prototypes shows that this technique can be used to perform climatological studies but, most important, it proves that the technical choices are valid since the sampling technique is able to extract pertinent physical features. This is confirmed by the literature on the use of K-means in geosciences [*Desbois et al.*, 1982; *Prigent et al.*, 2001; *Jakob et al.*, 2005].

4. First Guess Database Generation

4.1. Sampling Configurations

[48] We now use the techniques described in the previous section to create a first guess (FG) database of more than 2000 prototypes (we will see that this is a reasonable compromise between a high number of prototypes for good retrieval results, and a manageable size for computation time). We have only an indirect control on the prototype's number (through D_{MAX}) for the uniform sampling. The goal

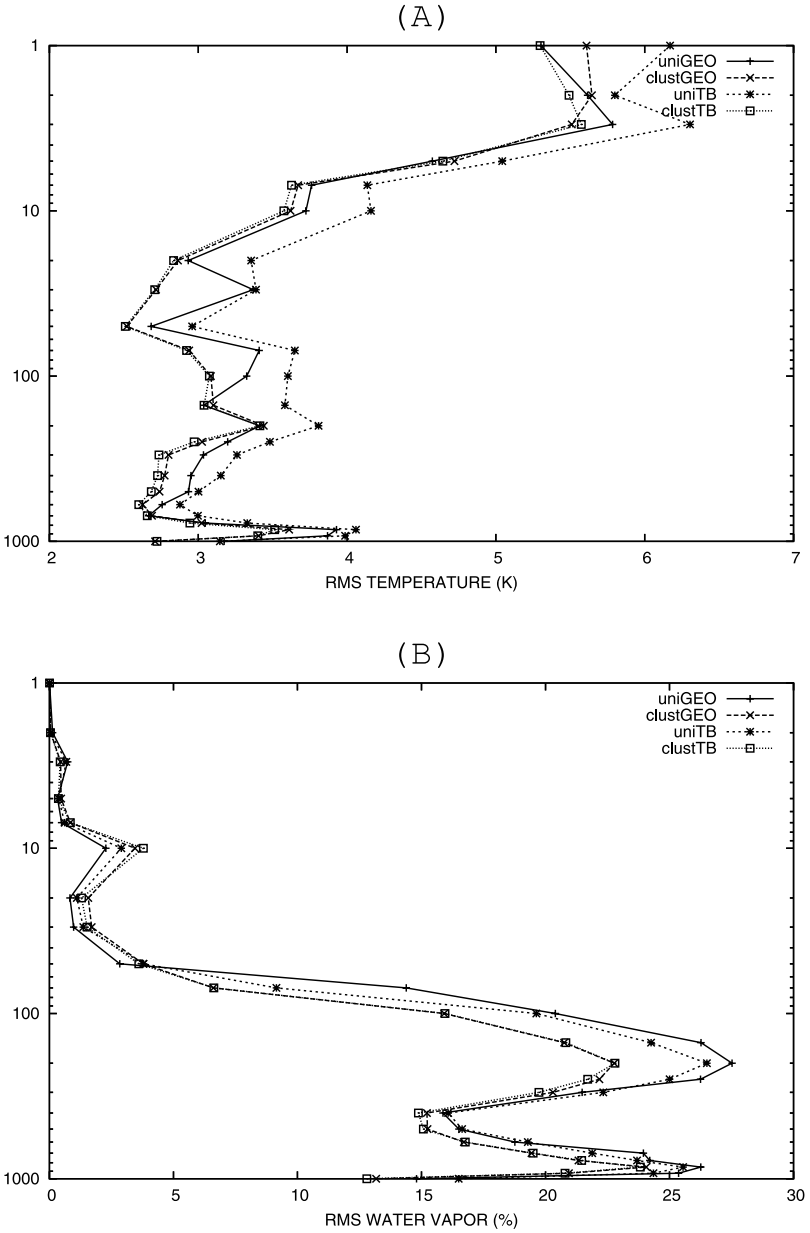

Figure 7. RMS errors of the FG extraction for (a) the temperature and (b) the water vapor atmospheric profiles in the *uniGEO*, *clustGEO*, *uniTB* and *clustTB* experiments.

being to compare results with the uniform and the clustering sampling algorithms, we first run the uniform sampling with a reasonable $D_{\max}$ that determines the number of prototypes for the uniform sampling (e.g., 2258 prototypes), then we use the same number of prototypes for the clustering.

[49] These prototypes need to represent as well as possible the variability in the 1-a ERA40 data set and to take into account the information content of the AMSU measurements. In order to test the various technical choices presented in section 3, four experimental configurations are defined: (1) uniform sampling on the *GEO* space, "*uniGEO*"; (2) clustering sampling on the *GEO* space, "*clustGEO*"; (3) uniform sampling on the *TB* space, "*uniTB*"; and (4) clustering sampling on the *TB* space, "*clustTB*."

[50] These four experimental configurations are compared by analyzing their respective FG extraction results. For each experiment, the prototypes are extracted in the *GEO* and the *TB* spaces ("dual" spaces). We note $uniGEO_{GEO}$ the GEOphysical prototypes of the *uniGEO* experiment, and $uniGEO_{TB}$ the corresponding prototypes in the *TB* space. Same notations hold for experiments *clustGEO*, *uniTB*, and *clustTB*.

[51] To estimate the FG extraction root mean square (RMS) errors in an testing data set $BASE = (BASE_{GEO}, BASE_{TB})$ independent of the original data set that was sampled to constitute the FG database, we proceed as follows, in the *uniGEO* case: (1) We first take the *ith TB* sample, tb^i, from $BASE_{TB}$ and its dual, GEO^i, in $BASE_{GEO}$. (2) Its closest prototype in $uniGEO_{TB}$ is found using a pertinent distance (see section 3.2): $\widehat{tb}_{uniGEO}(tb^i)$. (3) The *GEO* dual of $\widehat{tb}_{uniGEO}(tb^i)$ is extracted from $uniGEO_{GEO}$: $\widehat{GEO}_{uniGEO}(tb^i)$. (4) The RMS difference, RMS_{uniGEO}, for all

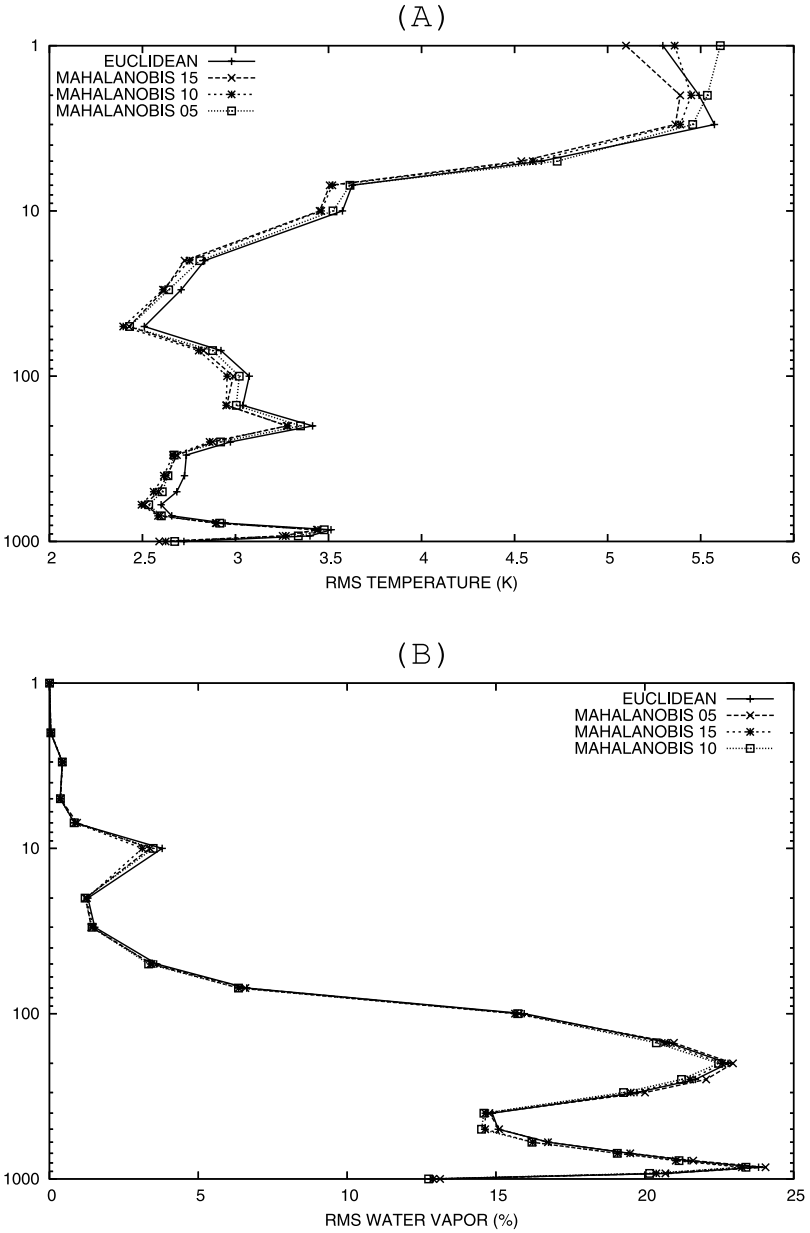

Figure 8. RMS errors of the FG extraction for (a) the temperature and (b) the water vapor atmospheric profiles in the *clustTB* experiment when using D_E, the Euclidean distance, and D_M, the Mahalanobis distance, with 15, 10 and 5 PCA components.

the samples i in *BASE* is the RMS difference between the pairs (GEO^i, $\widehat{GEO}_{uniGEO}(tb^i)$).

[52] The same approach is used for experiments *clustGEO*, *uniTB*, and *clustTB*.

[53] The four configurations are tested using the Euclidean distance in the independent data set *BASE*. The FG RMS errors for temperature are represented in Figure 7a. Clustering configurations, both for the *GEO* and *TB* spaces, outperform the uniform algorithm. Furthermore, the *clustTB*'s results are very similar to the *clustGEO* statistics except for a slight decrease in temperature errors on the higher atmospheric levels.

[54] The FG RMS errors for water vapor are shown in Figure 7b: The clustering configurations are very close to each other and they again do better than the uniform algorithm, except for levels higher than 40 hPa (note that the ERA40 water vapor at these levels needs to be considered with caution).

[55] In conclusion, the clustering is almost always better than the uniform algorithm. This was expected by design since the RMS errors criterion favors the clustering approach: the uniform sampling emphasizes rare events that do not influence the RMS, where the clustering focusses on the most common situations that will drive the RMS statistics. In other words, the clustering sampling extract the natural distribution that is used to estimate the RMS error, the uniform sampling extract a different distribution.

[56] It is equivalent in this particular application to use the *clustGEO* or the *clustTB* with a slight advantage to the *TB* space. Furthermore, working in the *TB* space seems

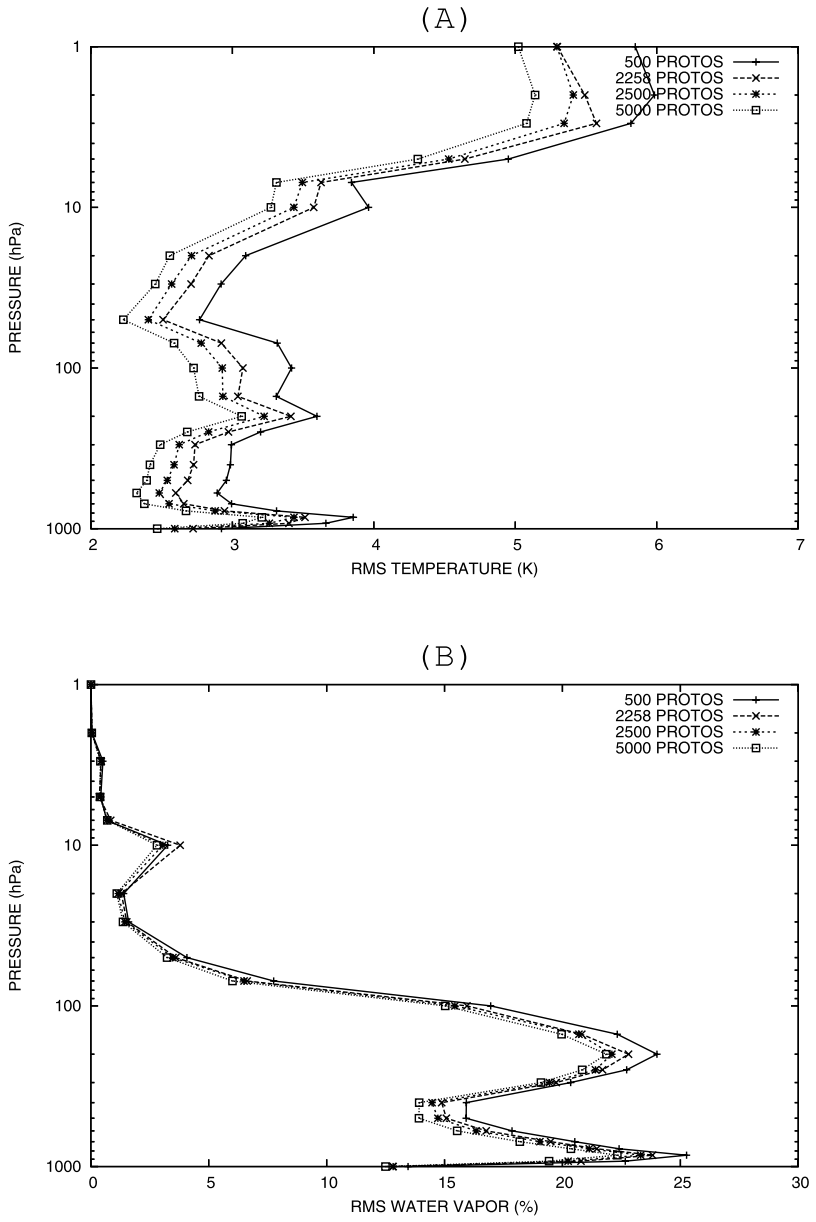

Figure 9. RMS errors of the FG extraction for (a) the temperature and (b) the water vapor atmospheric profiles in the *clustTB* experiment with 500, 2258, 2500, and 5000 extracted prototypes.

more convenient than in the *GEO* space for normalization issues: the natural variability of the atmosphere gets translated into the variability in the observations, so the normalization in the *TB* space is pertinent and is easier since observations have a similar range of variability. As a consequence, we will choose the *clustTB* version for the remaining experiments in the next sections.

4.2. Sensitivity Studies

[57] In this section, we use the *clustTB* configuration to measure the sensitivity of FG extraction results to various parameters of the sampling method, namely the distance, the number of extracted clusters, and the level of instrument noise.

[58] Figure 8 shows the dependence of the FG extraction of temperature and water vapor atmospheric profiles on the distance used in the sampling algorithm: the Euclidean distance D_E or the Mahalanobis distance D_M (with the number of PCA components $n' = 5$, 10 and 15) (see section 3.2). The Mahalanobis distance performs slightly better than the Euclidean one, even when using only $n' = 5$ components but the impact is not significant enough to definitely prefer one over the other. In this particular application, the choice of the distance does not seem to be determinant. This can be surprising knowing the importance of distances in general discriminant analysis. However, as already discussed in section 3.2, one distance can be favored over the others for various reasons, rapidity of computations, importance of the rare events, complexity of the normalization of data coordinates, and it is difficult to establish a definitive conclusion based on this application only. The robustness of the results with respect to the

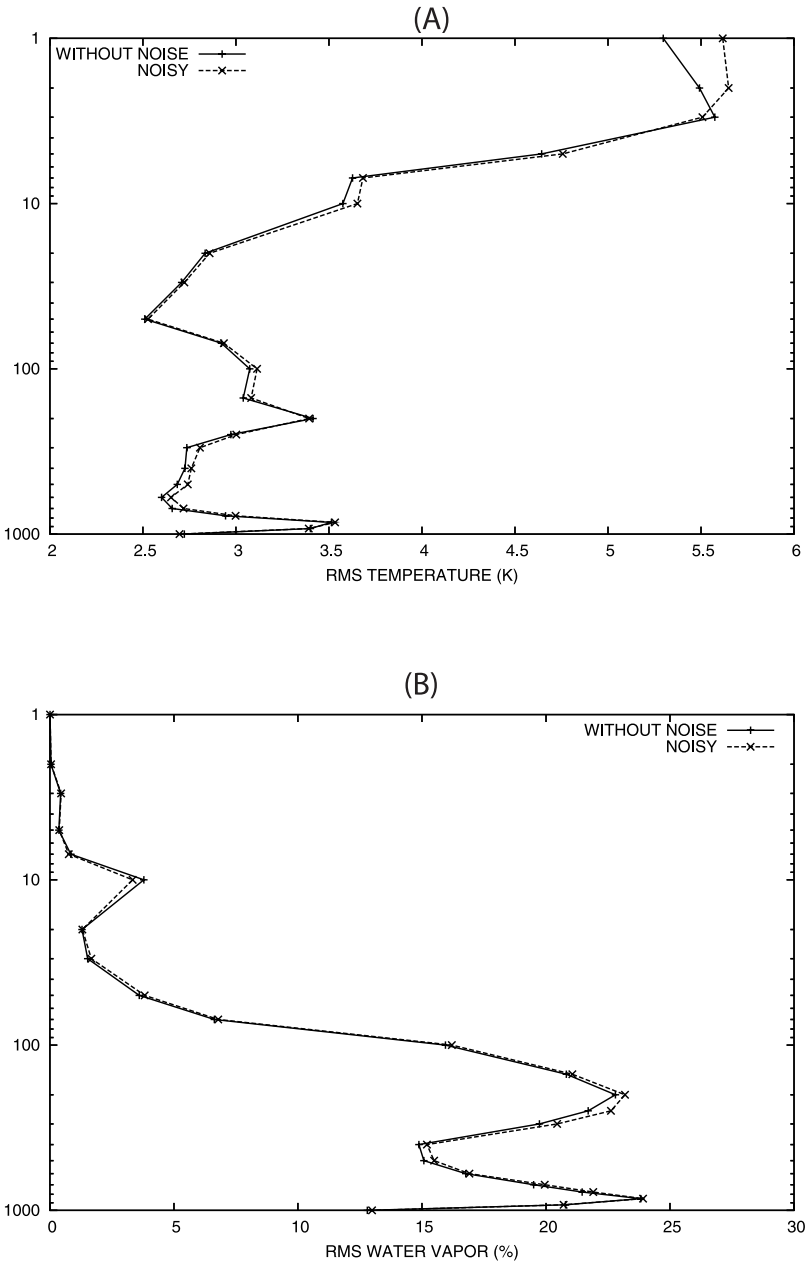

Figure 10. Sensitivity of the FG extraction to the level of instrument noise in the satellite observations for (a) atmospheric temperature and (b) water vapor profiles in the *clustTB* experiment.

distance confirms that we can use the Euclidean distance for the forthcoming experiments.

[59] The dependence of the FG extraction on the number of prototypes in the FG database is illustrated in Figure 9. FG errors are represented for temperature (Figure 9a), and for water vapor (Figure 9b), when using, from left to right, 2500, 2258, 2000, 1500, 1000 and 500 prototypes in the FG database. As expected, the RMS errors for both temperature and water vapor decrease when increasing the number of prototypes in the FG database. Increasing the number of samples in the FG database improves the FG extraction statistics but the required number of samples increases exponentially with the desired precision (because of the curse of dimensionality), setting a practical limitation on the quality of possible FG extraction error statistics.

[60] The FG is retrieved from the AMSU measurements which are subject to instrument noise. How sensitive is the FG extraction to the noise level? The 20 AMSU channels have a specified noise of, respectively, 0.20, 0.27, 0.22, 0.15, 0.15, 0.13, 0.14, 0.14, 0.20, 0.22, 0.24, 0.35, 0.47, 0.78, 0.11, 0.37, 0.84, 1.06, 0.70, and 0.60K [*Rosenkrantz*, 2001]. Figure 10 represents the FG extraction errors in the *clustTB* configuration for the temperature (Figure 10a) and the water vapor (Figure 10b) when a simulated instrument noise is added to the AMSU measurements. This noise is supposed to be Gaussian-distributed and independent from one channel to another. As expected, the FG extraction errors increase with the noise level but the sensitivity to noise is quite low, less than 0.1K in temperature and 1% in water vapor. This low sensitivity to noise demonstrates that

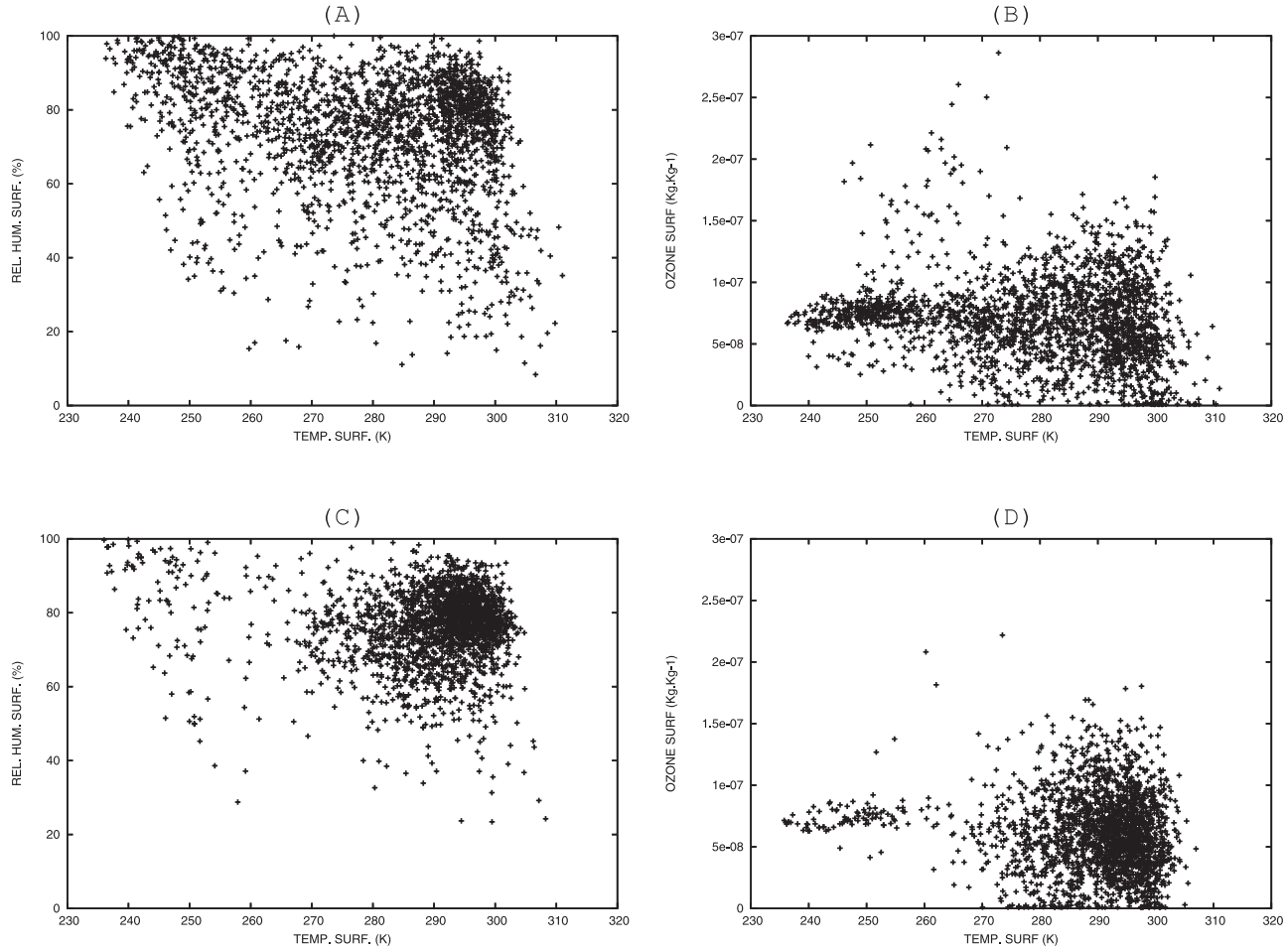

Figure 11. Scatterplots of the temperature with respect to (a) the relative humidity and (b) the ozone concentration at the surface layer for the *uniGEO*. (c and d) Similar but for the *clustTB* experiment.

the FG extraction scheme being a multivariate scheme, it benefits from the correlations among the AMSU channels so that instrument noises without channel correlation have a limited impact. This is an additional advantage of multivariate sampling techniques over methods that samples one variable at a time.

4.3. Further Analysis

[61] The impact of various algorithmic parameters on the FG extractions has been analyzed in the previous sections. Beyond these quantitative results, it is important to understand in depth how the algorithms work. The goal of this paper is not to choose one approach over another one but rather to better know how each method works and what are their respective advantages and inconveniences.

[62] The first question concerns the sampling algorithms and their respective performances. The numerical experiments (section 4.1 and Figure 7) show without ambiguity that the clustering algorithm obtains better results than the uniform sampling for the FG extraction. Is the clustering approach always preferable to the uniform sampling? Scatterplots of temperature, water vapor, and ozone content in the surface layer are represented in Figure 11 for the propotypes extracted by the uniform algorithm in the *GEO* space (i.e., *uniGEO*) and the clustering algorithm in the *TB* space (i.e., *clustTB*). The uniform sampling algorithm selects more prototypes in the less populated domains of the natural variability in the *BASE* data set (not shown since this domain can well be guessed from the *uniTB* scatterplots) and, as a consequence, this sampling includes rare events. Furthermore, the prototypes are more regularly spaced than in the clustering approach which selects more prototypes in the densely populated regions of the *BASE* original data set. This is the behavior expected by design from the two algorithms.

[63] The spread of the uniform prototypes over a larger domain could be a strong advantage of the uniform algorithm over the clustering method. Nevertheless, Figure 12 shows that the distributions of the temperature (Figure 12c) and water vapor (Figure 12f) at the surface layer of the clustering sampling are closer to the natural variability of the *BASE* data set (Figures 12a and 12d), than those derived from the uniform sampling (Figures 12b and 12e). Uniform sampling by nature tends to increase the statistical weight of rare events, which can be of interest depending of the application. However, if spurious/aberrant data are present in the original data set, these situations can be chosen as prototypes by the uniform sampling (this might be a problem if we were using radiosondes instead of ERA40 reanalyses in our application). On the contrary, the cluster-

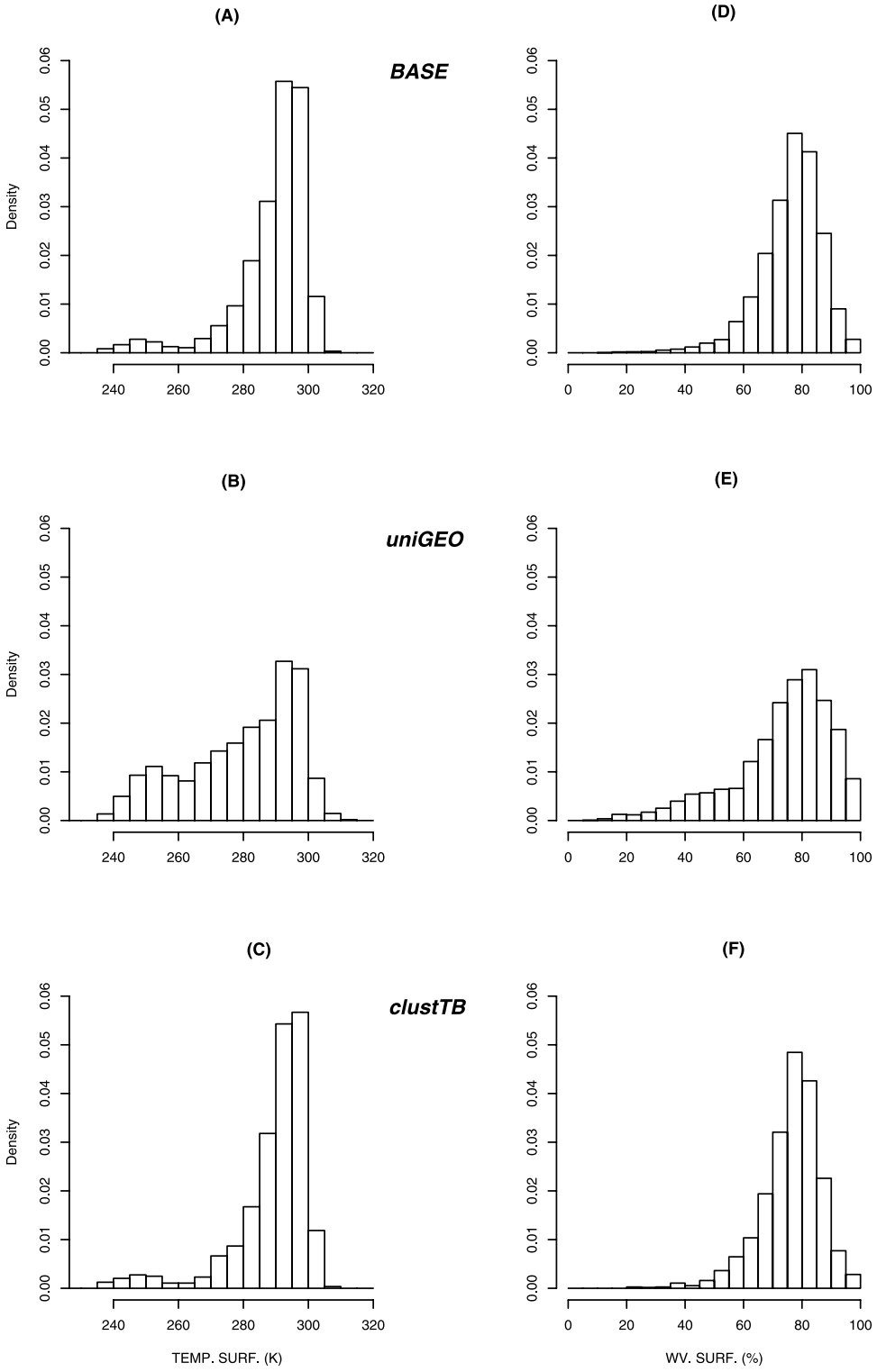

Figure 12. Histograms of temperature at the surface layer in (a) the *BASE* data set, (b) the *uniGEO* prototype database, and (c) the *clustTB* prototype database. (d–f) Similar but for the water vapor.

ing sampling provides sample distributions that respect the original variability which is a key property when using the extracted database to perform statistics. For the FG extraction, since the quality criterion was the RMS errors, the clustering sampling that mimics the *BASE* distributions and emphasizes the populated regions performs globally better than the uniform database. The uniform database FG extraction would nevertheless perform better than the clustering database for situations located at the edge of the natural variability.

[64] The histograms of the FG errors for both sampling techniques (Figure 13a for the *uniGEO* and Figure 13b for

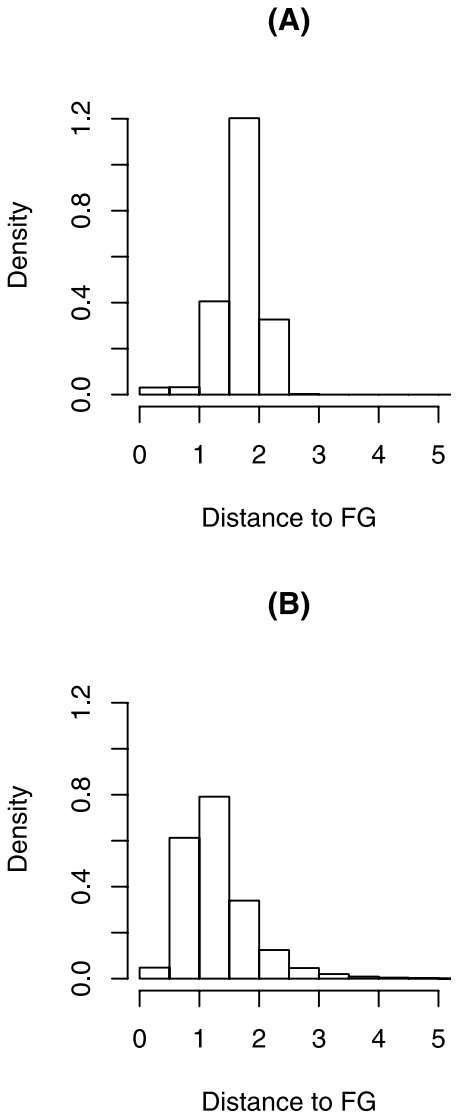

Figure 13. Histograms of the FG extraction errors for (a) the *uniGEO* and (b) the *clustTB* databases.

the *clustTB* configurations) provide another way to understand the differences between uniform and clustering sampling. For the uniform sampling method, the FG error is by design lower than a given threshold (i.e., $D_{MAX} = 2.6$ in our case). Bayesian techniques can favor this threshold on the maximum FG error because the error characteristics are simpler than for the cluster sampling. The clustering histogram has not a threshold on the maximum FG error but its average is still lower than for the uniform sampling algorithm. The clustering sampling gives more weight to the populated regions which improve its overall statistics whereas the uniform sampling can select a limited number of prototypes in the rare event regions that are statistically less important.

[65] The curse of dimensionality shows that the number of required prototypes in the FG database increases exponentially with the desired precision. A practical limit in the troposphere seems to be about 3K for the temperature and about 20% for the water vapor.

[66] FG extraction errors for the *clustGEO* and the *clustTB* configurations being quite similar (see Figure 7), it could be concluded that working in either space is equivalent but the arguments of section 3.3 must be considered to make a choice.

5. Training Database Generation

[67] In this section, the FG databases generated by the four experimental configurations of section 4 are used to train a NN retrieval scheme. The goal is to test the databases as training tools for a statistical retrieval scheme. No more information than the databases used in the FG extraction is introduced into the NN model: Any difference between FG and NN retrieval results has to be explained by the way the algorithms use the information. Any other statistical algorithms built from a training database could be used instead of the NN technique for this comparison purposes.

5.1. Retrieval Methodology

[68] Neural Network (NN) techniques have proved very successful in developing computationally efficient algorithms for remote sensing applications. A NN algorithm is applied to retrieve simultaneously the atmospheric temperature and humidity profiles at 23 fixed pressure levels from 1000 to 1hPa over sea using AMSU-A and AMSU-B observations. The Multilayered Perceptron (MLP) model is selected [*Rumelhart et al.*, 1986]: It is a nonlinear mapping model. Given an input X, it provides an output Y. In our case, X is composed of the AMSU observations and Y represents the temperature and the water vapor atmospheric profiles. The architecture has 20 neurons in the input layer (i.e., the AMSU *TB*s), 200 neurons in the hidden layer, and 46 neurons in the output layer (the two profiles).

[69] The NN is trained to reproduce the behavior described by a database of samples composed of inputs X^e (i.e., the *TB*s) and their associated outputs Y^e (i.e., *GEO*-physical variables), for $e = 1,\ldots, N$ with N the sample number in the training database. Provided that enough samples (X^e, Y^e) are available, any continuous relationship, as complex as it is, can be represented by a MLP [*Hornik et al.*, 1989; *Cybenko*, 1989]. The databases extracted from the algorithms defined in section 4 (*uniGEO*, *clustGEO*, *uniTB*, and *clustTB*) are used to train the NN models. Since this model is very general and efficient, the results obtained with these four databases can be considered a good measure of their respective quality as training databases.

5.2. Retrieval Results

[70] Figure 14 represents the NN RMS errors for the temperature (Figure 14a) and the water vapor atmospheric profile retrievals (Figure 14b). No bias is observed in the NN retrievals (not shown) so the RMS represents essentially the variance error. For the four configurations (*uniGEO*, *clustGEO*, *uniTB*, and *clustTB*), the clustering databases outperform the uniform sampling databases for both temperature and water vapor. This was expected since the RMS error criterion favors the clustering approach. Differences between the four configuration results can be significant with more than 1K in temperature and 5% in water vapor. Among the uniform sampling databases, the *uniGEO* is

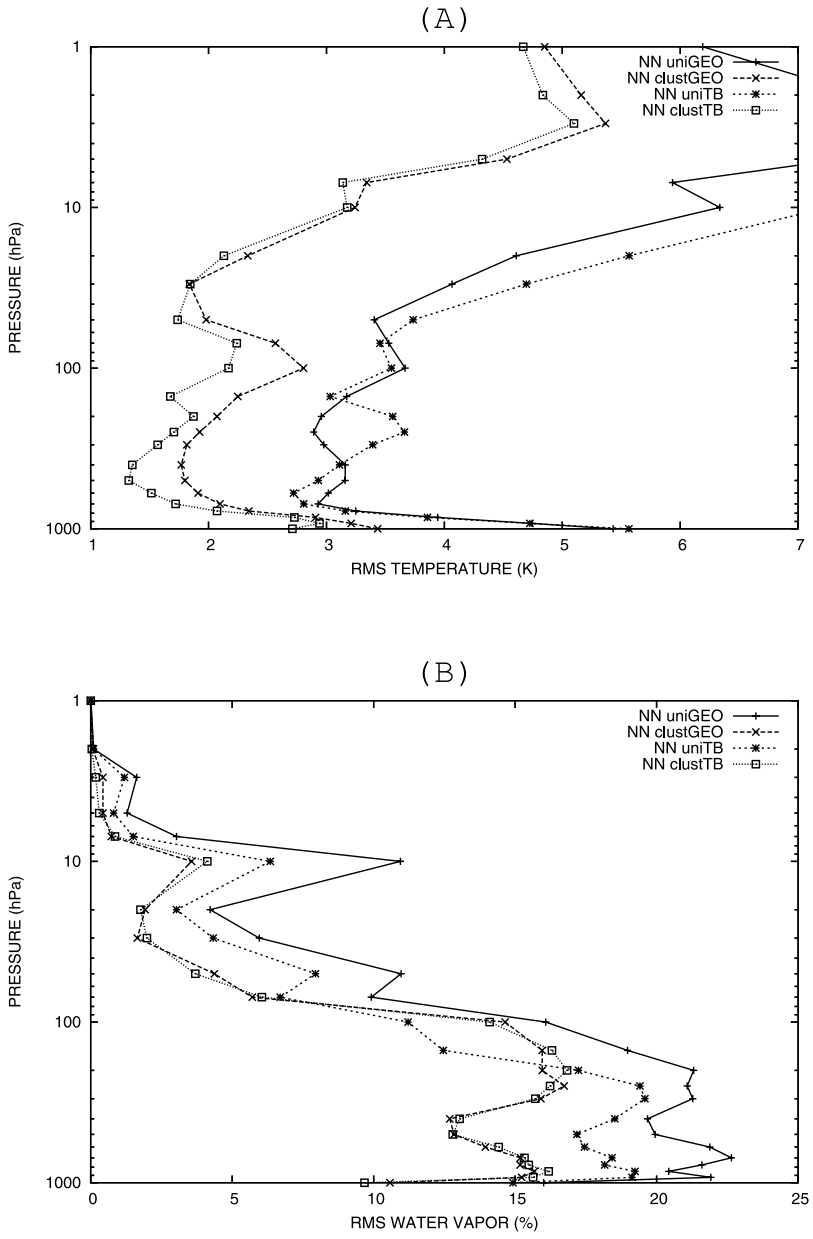

Figure 14. RMS errors in the (a) atmospheric temperature and (b) water vapor profiles for a NN retrieval trained with the databases from the *uniGEO*, *clustGEO*, *uniTB*, and *clustTB* experiments.

better than the *uniTB* but for the clustering sampling databases, results of the NN retrieval with *clustGEO* or *clustTB* are similar.

[71] The FG extraction can be considered as a legitimate retrieval/inversion algorithm, sometimes called ''analogue inversion''; this is represented by a direct link between the first guess retrieval and the geophysical products in the scheme of Figure 1. Classical look-up table inversion algorithm is such an example of simple FG inversion. It is interesting to compare the results of the NN retrievals of Figure 14 to those of the FG extraction (Figure 7). The NN is trained with the same database as the FG extractions so the NN retrieval results can directly be compared to the FG ones of Figure 7: The differences come from a better exploitation of the databases by the NN model. The improvement is considerable for the temperature, with a decrease of the RMS error between 1 and 1.5 K depending on the atmospheric level. The water vapor atmospheric profile is marginally improved in general, except a 5% decrease in RMS at the tropopause level. The considerable improvement in temperature can be explained by the fact that the FG extraction can only give a database prototype as answer, whereas the NN retrieval (or any other statistical retrieval scheme) is able to interpolate between the prototypes of the training database. This is an important point: A good retrieval scheme interpolates between the prototypes inside its training database and is able to overcome the curse of dimensionality of section 7 by increasing the number of samples in its training database through interpolation.

[72] We now focus on the *clustTB* configuration that provides the best retrieval results for this application. Two examples of retrievals are presented in Figure 15:

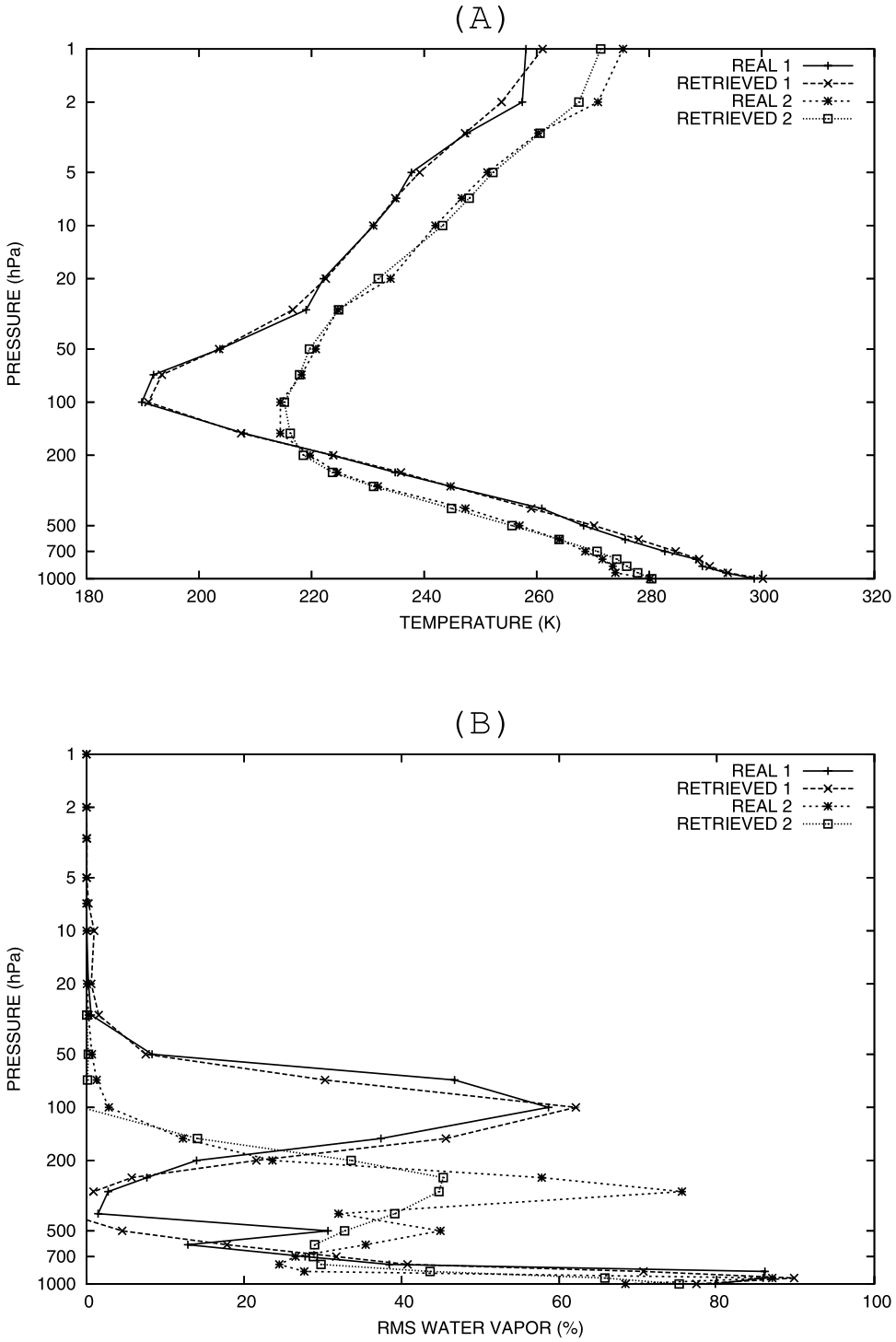

Figure 15. Two examples of NN retrievals for (a) the temperature and (b) the water vapor atmospheric profiles.

Figure 15a shows two temperature atmospheric profiles with their NN retrieval and Figure 15b is similar but for the water vapor. The temperature profiles are well estimated as measured by the global statistics in Figure 14a. The general profile structure is correctly retrieved, in particular at the tropopause level. Some oscillations can be found around the real temperature profile but this is much more visible for the water vapor profile. The presence of such oscillations means that there is no bias error in the estimations (i.e., no atmospheric contribution is systematically missed by the sounder). Variance errors are the consequence of the limited vertical resolution of the AMSU instrument. These errors are an illustration of the compensation problem: An underestimation in one atmospheric layer is compensated by an overestimation at a neighboring layer. This uncertainty cannot be suppressed because of the lack of

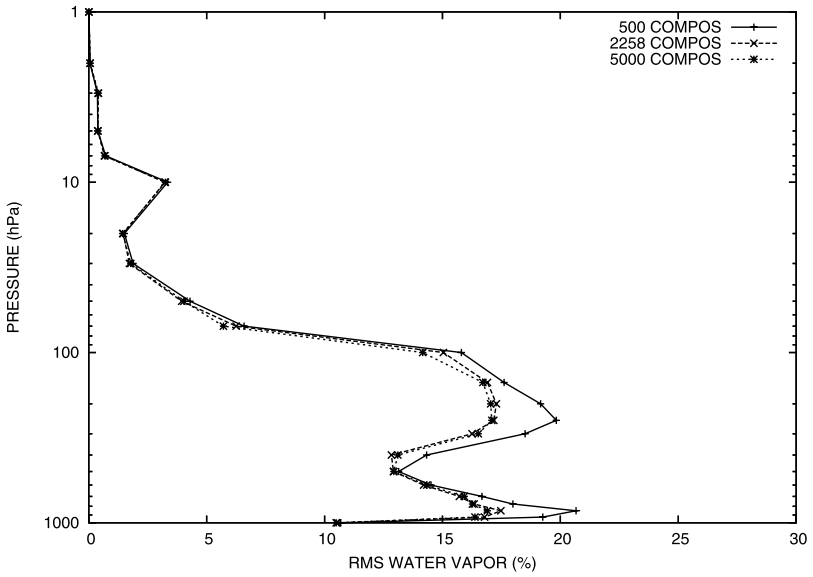

Figure 16. Atmospheric water vapor profile RMS errors for a NN trained with the databases in the *clustTB* experiment with 500, 2258, and 5000 extracted profiles.

information in the satellite observations. Such errors occur in underdetermined inverse problems [*Tarantola*, 1987]. Increasing the number of samples in the training database would not solve this problem since it is a direct limitation of the AMSU measurements.

5.3. Sensitivity Studies

[73] As in section 4.2, we study the dependency of the retrieval scheme on parameters of the sampling algorithms. Again, the *clustTB* configuration is chosen to perform these tests. Since the FG extraction can be considered as a legitimate inversion scheme, i.e., ''analogue inversion'' algorithm, most of the conclusions for the FG extractions in section 4.2 are valid for the NN retrieval algorithm as well. In particular, the distance used to sample the training data set does not impact much the retrieval results (not shown).

[74] The sensitivity to noise is also low: Experiments have been conducted to measure the impact of the instrument noise on the NN retrieval statistics. Same characteristics have been used as in section 4.2, with a magnitude of noise from one to two standard deviations. The impact of noise (not shown) is negligible showing clearly that the FG extraction scheme benefits from the correlations among the AMSU channels so that a channel-incoherent instrument noise has a limited effect. A similar behavior was already observed for the FG extraction.

[75] The temperature estimations appear less stable (not shown) than the water vapor to measure its sensitivity to the number of samples in the training database. The RMS errors of the water vapor atmospheric profiles are represented in Figure 16 when using the *clustTB* configuration with training databases that include 500, 2258, and 5000 prototypes. The RMS errors for the 500 prototypes are larger, as expected, but the 2258 and the 5000 prototypes experiment have similar statistics. This indicates that beyond a certain number of samples, the NN retrieval does not improve. Contrarily to the FG extraction statistics of Figure 9, this precision saturation is not the result of a limitation of the database: The limiting factor is the information content of the AMSU observations themselves. The NN retrieval scheme, by interpolating between the samples in the training database, succeeds in overpassing the limitation in sample number. The NN can still suffer from the curse of dimensionality but it is able to limit the number of samples to a manageable size. It still needs a database that explores the relevant regions of the data space, but needs a fewer sampling density in each of these regions. In other words, the NN retrieval scheme reduces the problem of curse of dimensionality.

6. Conclusion

[76] The objective of this study was to discuss the characteristics of sampling methods in the framework of remote sensing databases in high-dimensional spaces. It is based on a controllable and realistic experiment, the retrieval of a temperature and water vapor profile with AMSU-A and AMSU-B. It is dangerous to draw definite conclusions based only on one particular test but this experiment was carefully selected to be representative of complex atmospheric remote sensing problems. The RMS errors criterion that we use to compare solutions favors the clustering approach so for this application the clustering approach does better than the uniform sampling method. Each strategy possess its own advantages and inconveniences and the goal of this paper was to identify them. We can also state that the sampling in the space of geophysical variables is almost equivalent to the sampling in the satellite observation space which is more easy to work on in practice. The intrinsic quality of each method is not the only criterion to choose a sampling algorithm, the computational cost and the engineering constraints also have to be taken into account.

[77] It has been shown that the FG extraction is sensitive to the curse of dimensionality: The number of required

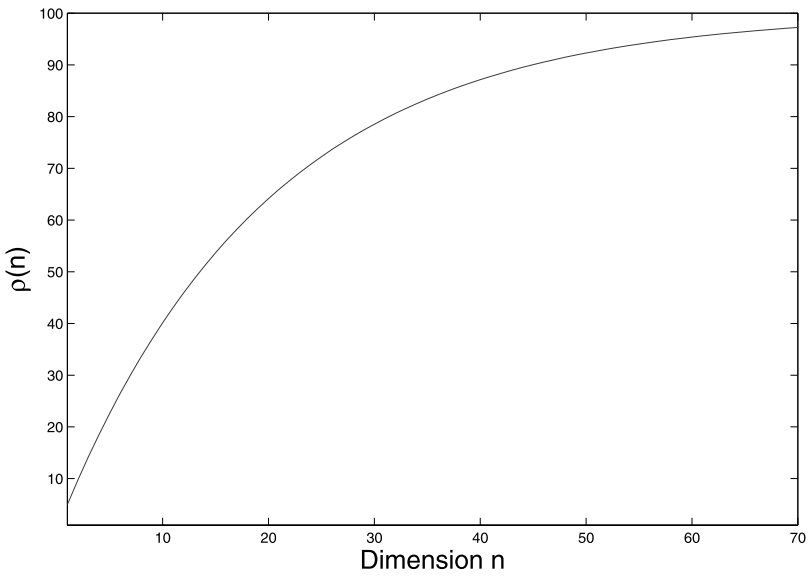

Figure A1. Percentage $\rho(n)$ of the "periphery" volume in the unit hypersphere with respect to the space dimension n.

samples increases exponentially with the desired precision. On the contrary, the NN retrievals are less dependent on the number of samples in the training data sets used to train them: The size of the database is only determined by the information content of the satellite observation because the training process interpolates correctly between the samples.

[78] The quasi-automatic tools developed for this study can be adapted for the generation of specialized FG and training databases for each particular application. These extracted databases will be used to develop remote sensing algorithms designed for specific conditions such as particular geographical domains, instrument characteristics (e.g., instrumental noise levels), and more generally any specific a priori information on the observing system.

[79] In addition to their practical application to the design of retrieval schemes, these routines can help analyze the information content of satellite observing systems. During the preparation of a new mission, these tools can be used to assess the sensitivity of the retrievals to parameters such as the spatial and spectral resolutions, the number and spectral locations of the channels, the instrument noise characteristics, or the impact of the first guess information. These techniques can quantify precisely the impact of each particular instrument characteristics on the retrieval.

[80] An atmospheric and surface data set that includes cloudy conditions is under construction for the Megha-Tropiques French-Indian mission that will be launched in 2009, and the tools described here will be used to develop the retrieval algorithm.

Appendix A: Curse of Dimensionality

[81] Handling of high-dimensional data ($\in \mathbb{R}^n$ with n large) in statistics or in function approximation introduces theoretical and practical problems. These difficulties are referred to as the "curse of dimensionality" in mathematics [*Bellman*, 1961] and have important consequences on multiple-integration, function approximation, or probability density function estimation.

[82] To obtain good statistics or function approximation, the original data space has to be densely sampled. When the data space has a high dimension, the number of required samples often grows exponentially. The sampling density of a limited n-dimensional space is proportional to $E^{1/n}$, with E the number of samples. This means that in order to obtain the same sampling density than a one-dimensional space with 100 samples ($E^{1/n} = 100$), 10^{20} samples are required in a ten-dimensional space.

[83] The understanding of the concept of distance can also be "distorted" in high-dimensional spaces. The volume of an hypersphere of radius R in a n-dimensional space n ($n \geq 2$) is:

$$V_R(n) = \frac{1}{n} \cdot R \cdot A_R(n).$$

where $A_R(n)$ is the area of the hypersphere (the area of a hypersphere with radius R in a n-dimensional space ($n \geq 2$) is given by $S_R(n) = \frac{2\sqrt{\pi^n}}{\Gamma\left(\frac{n}{2}\right)} R^{n-1}$, where $\Gamma(x) = \int_0^{+\infty} e^t t^{x-1}\, dt$) [*Hilbert and Cohn-Vossen*, 1952]. We call "periphery volume" the volume between the envelop of the hypersphere of radius $r = 1 - \varepsilon$ with $\varepsilon = 0.05$ and the hypersphere of unit radius ($R = 1$). Let $\rho(n)$ be the percentage of the periphery volume in the hypersphere of unit radius:

$$\rho(n) = 100 \frac{V_1(n) - V_r(n)}{V_1(n)}.$$

Figure A1 shows that when the dimension n increases, the percentage of the periphery volume inside the unit radius hypersphere increases rapidly toward 100%. As a consequence, for a uniform sampling of the unit radius hypersphere, the samples concentrate in the periphery when

dimension n increases. The behavior of distances can thus be puzzling in high-dimensional space.

[84] The exponential increase of required sample number and the distortion of distances does not imply that function approximation or statistical estimation cannot be successful in high-dimensional spaces. *Kolmogorov* [1957] showed that the dimension of the data space is not the leading factor in function approximation but that the intrinsic complexity of the function to be approximated or distribution to be estimated is the true constraining factor. Higher dimension gives the potential for more complexity (i.e., degrees of freedom), but this is not always the case. The ''intrinsic complexity'' (i.e., effective degrees of freedom) does not have to increase exponentially with the dimension so it does not suffer so severely from the curse of dimensionality. (This concept can be materialized in various ways, depending on the context (statistics, computer science, analysis, neural network theory). In statistics, the degrees of freedom, the entropy or other complexity measures can be used. The Vapnik-Chervonenkis dimension [*Vapnik and Chervonenkis*, 1971] is an attempt to measure the complexity of a function, but its practical use is limited.) The models to be used still need to characterize this intrinsic complexity. The error of a model is decomposed into two parts: the bias and the variance. A model with a lower complexity (not enough degrees of freedom) or a ''shape'' that is not adapted to the function to be approximated will have large bias error (i.e., inadequacy of the model). On contrary, a model with too high a complexity (too many degrees of freedom) will suffer from overfitting and the variance error will be high [*Geman et al.*, 1992]. By using good a priori information (e.g., a regularization term), it is possible to introduce additional constraints to limit the number of free parameters that would require too many samples in the training database.

[85] In addition to the higher number of samples required, the curse of dimensionality can also cause rapidity and computational problems, since the computation number might increase exponentially with the data dimension.

[86] **Acknowledgments.** We would like to thank Michel Desbois and Rémy Roca for interesting discussions linked to the Megha-Tropiques mission. We are grateful to Alain Chédin and Noëlle Scott for the first discussions on this topic during Filipe Aires's Ph.D. thesis. We would like to thank Frédéric Chevallier for sharing his experience on the generation of the TIGR database. Finally, Filipe Aires would like to thank for their kindness all the team of the public library in Amarante (Portugal) where most of this paper has been written.

References

Achard, V. (1991), Trois problèmes clés de l'analyse tridimentionelle 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.

Aires, F., C. Prigent, W. B. Rossow, and M. Rothstein (2001), A new neural network approach including first-guess for retrieval of atmospheric water vapor, cloud liquid water path, surface temperature and emissivities over land from satellite microwave observations, *J. Geophys. Res.*, *106*(D14), 14,887–14,907.

Aires, F., A. Chedin, N. Scott, and W. B. Rossow (2002a), A regularized neural network approach for retrieval of atmospheric and surface temperatures with the IASI instrument, *J. Appl. Meteorol.*, *41*(2), 144–159.

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

Bellman, R. (1961), *Adaptative Control Processes*, Princeton Univ. Press, Princeton, N. J.

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

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

Chevallier, F., A. Chédin, F. Chéruy, and J. J. Morcrette (2000), TIGR-like atmospheric profile database for accurate radiative flux computation, *Q. J. R. Meteorol. Soc.*, *126*, 777–785.

Chevallier, F., S. Di Michele, and A. P. McNally (2007), Diverse profile datasets from the ECMWF 91-level short-range forecasts, *Rep. NWPSAF-EC-TR-010*, Numer. Weather Predict. Satell. Appl. Facil., Met Off., Exeter, U. K.

Cordisco, E., C. Prigent, and F. Aires (2006), Snow characterization at a global scale with passive microwave satellite observations, *J. Geophys. Res.*, *111*, D19102, doi:10.1029/2005JD006773.

Crone, L., and D. Crosby (1995), Statistical applications of a metric on subspaces to satellite meteorology, *Technometrics*, *37*(3), 324–328.

Cybenko, G. (1989), Approximation by superpositions of a sigmoidal function, *Math. Control Signals Syst.*, *2*, 303–314.

Desbois, M., G. Seze, and G. Szejwach (1982), Automatic classification of clouds on METEOSAT imagery: Application to high-level clouds, *J. Appl. Meteorol.*, *21*(3), 401–412.

Escobar, J. (1993), 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.

Eyre, J. R. (1991), A fast radiative transfer model for satellite sounding systems, *ECMWF Res. Dep. Tech. Memo.*, *176*, Eur. Cent. for Med.-Range Weather Forecasts, Reading, U. K.

Franquet, S. (2003), Contribution à l'étude dy cycle hydrologique par radiométrie hyperfréquence: Algorithmes de restitution (réseaux de neurones) et validation pour la vapeur d'eau (instrument AMSU, SAPHIR) et les précipitations (AMSU, Radars au sol Baltrad), Ph.D. thesis, Univ. Paris-Diderot, Paris VII, Paris, 3 March.

Gelman, A., J. B. Carlin, D. B. Stern, and D. B. Rubin (2003), *Bayesian Data Analysis*, 2nd edition, 696 pp., CRC Press, Boca Raton, Fla.

Geman, S., E. Bienenstock, and R. Doursat (1992), Neural networks and the bias-variance dilema, *Neural Comput.*, *1*(4), 1–58.

Goodrum, G., K. B. Kidwell, and W. Winston (2000), *NOAA KLM User's Guide*, NOAA, Siver Spring, Md.

Gordon, N. D., J. R. Norris, C. P. Weaver, and S. A. Klein (2005), Cluster analysis of cloud regimes and characteristic dynamics of midlatitude synoptic systems in observations and a model, *J. Geophys. Res.*, *110*, D15S17, doi:10.1029/2004JD005027.

Hilbert, D., and S. Cohn-Vossen (1952), *Geometry and the Imagination*, 357 pp., Am. Math. Soc., Providence, R. I.

Hornik, K., M. Stinchcombe, and H. White (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, *2*, 359–366.

Huber, P. J. (1981), *Robust Statistics*, 320 pp., John Wiley, New York.

Jakob, C., G. Tselioudis, and T. Hume (2005), The radiative, cloud, and thermodynamic properties of the major tropical western Pacific cloud regimes, *J. Clim.*, *8*, 1203–1215, doi:10.1175/JCLI3326.1.

Jolliffe, I. T. (2002), *Principal Component Analysis*, 2nd ed., 487 pp., Springer, New York.

Kohonen, T. (1984), *Self-Organization and Associative Memory*, Springer, New York.

Kolmogorov, A. (1957), On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition, *Dokl. Akad. Nauk SSSR*, *114*, 953–956.

Kummerow, C., Y. Hong, W. S. Olson, S. Yang, R. F. Adler, J. McCollum, R. Ferraro, G. Petty, D.-B. Shin, and T. T. Wilheit (2001), The evolution of the Goddard Profiling Algorithm (GPROF) for rainfall estimation from passive microwave sensors, *J. Appl. Meteorol.*, *40*(11), 1801–1820.

Lloyd, S. P. (1992), Least squares quantization in PCM, *IEEE Trans. Inf. Theory*, *28*(2), 129–137.

Matricardi, M., F. Chevallier, and S. Tjemkes (2001), An improved general fast radiative transfer model for the assimilation of radiance observations, *ECMWF Res. Dep. Tech. Memo.*, *345*, Eur. Cent. for Med.-Range Weather Forecasts, Reading, U. K. (Available at http://www.ecmwf.int/publications)

Milligan, G. W., and M. C. Cooper (1985), An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, *50*(2), 159–179.

Mimmack, G. M., S. J. Mason, and J. S. Galpin (2001), Choice of distance matrices in cluster analysis: Defining regions, *J. Clim.*, *14*, 2790–2797.

Moody, J., and C. J. Darken (1989), Fast learning in networks of locally-tuned processing units, *Neural Comput.*, *1*(2), 281–294.

Moore, D. S., and G. P. McCabe (2006), *Introduction to the Practice of Statistics*, 5th ed., W. H. Freeman, New York.

Omar, A. H., J.-G. Won, D. M. Winker, S.-C. Yoon, O. Dubovik, and M. P. McCormick (2005), Development of global aerosol models using cluster analysis of Aerosol Robotic Network (AERONET) measurements, *J. Geophys. Res.*, *110*, D10S14, doi:10.1029/2004JD004874.

Press, W. H., B. P. Flannery, and S. A. Teukolsky (2002), Numerical Recipes in *C: The Art of Scientific Computing*, 2nd ed., 1032 pp., Cambridge Univ. Press, New York.

Prigent, C., F. Aires, W. B. Rossow, and E. Matthews (2001), Joint characterization of the vegetation by satellite observations from visible to microwavelengths: A sensitivity analysis, *J. Geophys. Res.*, *106*(D18), 20,665–20,685.

Rendell, L., H. Whitehead, and A. Coakes (2005), Do breeding male sperm whales show preferences among vocal clans of females?, *Mar. Mammal Sci.*, *21*(2), 317–322, doi:10.1111/j.1748-7692.2005.tb01231.x.

Rodgers, C. D. (2000), *Inverse Methods for Atmospheric Sounding—Theory and Practice*, World Sci., London.

Rosenkrantz, P. (2001), Retrieval of temperature and moisture profiles from AMSU-A and AMSU-B measurements, *IEEE Trans. Geosci. Remote Sens.*, *39*, 2429–2435.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986), Learning internal representations by error propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I, *Foundations*, edited by D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, pp. 318–362, MIT Press, Cambridge, Mass.

Saunders, R. W., M. Matricardi, and P. Brunel (1999), An improved fast radiative transfer model for assimilation of satellite radiance observations, *Q. J. R. Meteorol. Soc.*, *125*, 1407–1425.

Shi, L. (2001), Retrieval of atmospheric temperature profiles from AMSU-A measurements using a neural network approach, *J. Atmos. Oceanic Technol.*, *18*, 340–347.

Simmons, A. J., and J. K. Gibson (2000), The ERA40 project plan, Eur. Cent. for Med.-Range Weather Forecasts, Reading, U. K.

Tarantola, A. (1987), *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier, Amsterdam.

Thépaut, J.-N., and P. Moll (1990), Variational inversion of simulated TOVS radiancs using the adjoint technique, *Q. J. R. Meteorol. Soc.*, *116*(496), 1425–1448.

Vapnik, V. N., and A. Y. Chervonenkis (1971), On the uniform convergence of relative frequencies of events to their relative frequencies of events to their probabilities, *Theory Probab. Appl.*, *16*, 264–280.

Vrac, M., A. Chédin, and E. Diday (2005), Clustering a global field of atmospheric profiles by mixture decomposition of copulas, *J. Atmos. Oceanic Technol.*, *22*, 1445–1459.

F. Aires, Laboratoire de Météorologie Dynamique, Institut Pierre-Simon Laplace/Centre National de la Recherche Scientifique, Université Pierre et Marie Curie, case 99 4, place Jussieu, F-75252 Paris Cédex 05, France. (filipe.aires@lmd.jussieu.fr)

C. Prigent, Laboratoire d'Etudes du Rayonnement et de la Matière en Astrophysique, Centre National de la Recherche Scientifique, Observatoire de Paris, 61, av. de l'Observatoire, F-75014 Paris, France. (catherine.prigent@obspm.fr)