Pattern Recognition in the Satellite Temperature Retrieval Problem

OWEN E. THOMPSON, MITCHELL D. GOLDBERG[1] AND DONALD A. DAZLICH[2]

Department of Meteorology, University of Maryland, College Park, MD 20742

(Manuscript received 6 February 1984, in final form 30 August 1984)

ABSTRACT

Two pattern recognition procedures are developed to provide improvements to first-guess fields for satellite temperature retrievals. The first is a technique whereby a radiometer measurement may be used to select one or more historical radiosonde temperature profiles as *analog* estimates of ambient thermal structure. The vertical scales of the analogs are those of radiosondes—the vertical resolving power of the satellite radiometer being relevant only to a decision process. The analog selection process is shown to be much more effective if implemented in an orthogonalized space of measurement information. The second procedure is one which partitions *a priori* dependent data into shape-coherent pattern libraries using structure informaton inherent in the data itself. This is an alternative to traditional partitioning schemes whereby proxy classifiers such as season, location and surface type are used.

These pattern recognition techniques are shown to be capable of reducing first-guess profile errors by nearly 50%, in an independent test of about 800 diverse retrievals. The impact of pattern recognition on temperature retrieval error is assessed using regression and physical-iterative retrieval algorithms. The influence of improved first-guess fields is markedly different on these two types of algorithms. Pattern recognition is shown to have a strong, positive impact on the physical-iterative method but little significant impact on regression when evaluated in an overall batch sense. A case study suggests that a small number of very poor retrievals may particularly mask the potential benefits of pattern recognition on both methods.

1. Introduction

This study deals with the application of pattern recognition techniques to the satellite sounding problem. While the focus in these papers is satellite temperature soundings, the concepts to be discussed might apply equally to other problems in which one is to infer one set, or vector, of geophysical variables from a different set of fundamental measurements.

For temperature sounding, our interest in pattern recognition follows a concern about the influence of *a priori* information on satellite retrievals. Retrieval algorithms require a first-guess profile and the influence of this first guess on the results has been a subject of study. Smith *et al.* (1972) adapted the minimum information method of Foster (1961) to a numerical weather prediction first guess field, arguing that the forecast profile was a better *a priori* estimate of the thermal field than climatology. Using a similar solution algorithm, Thompson and Wolski (1976) showed that a bias in the first-guess location of the tropopause is pretty much transferred through the algorithm into the retrieval results. The inability of physical retrieval algorithms to efficiently correct shape errors in the first guess field is due fundamentally to the comparatively low vertical resolving power of satellite sounders (see Conrath, 1972; Gautier and Revah, 1975; Thompson, 1982). Highly statistical methods, such as the regression approach of Smith *et al.* (1970), and the empirical orthogonal function approach of Smith and Woolf (1976) use *a priori* correlation and covariance statistics in addition to a statistical first guess field. Mixed physical-statistical algorithms such as minimum rms (Foster, 1961; Strand and Westwater, 1968; Rodgers, 1970) also use *a priori* covariance information and statistical mean first guess field. As shown by Crosby and Weinreb (1974) and Spänküch *et al.* (1977), poor results may be obtained with such methods if unrepresentative statistics are used in the sounding.

There have been investigations directed toward improving the first-guess representation. Westwater and Grody (1980), Gage and Green (1982a,b), Westwater, Decker, Zachs and Gage (1983) have demonstrated improvements that could be made with ground-based remote sensing devices. Fritz (1977) has demonstrated potential improvements in soundings by conditioning the retrieval problem with RAOBs nearby in space and time. We propose here that pattern recognition techniques can be used to advantage to form improved first guess fields and to select representative *a priori* data sets for sounding.

The techniques to be tested involve the following basic concept. Spectral radiances measured by a satellite sounder can be thought of as forming a vector in some n-dimensional space. In this conceptual discussion, it is convenient to think of this vector as the radiance or brightness temperature vector itself,

[1] Current affiliation: Systems and Applied Sciences Corp., Hyattsville, MD.

[2] Current affiliation: MACOM/Sigma Data, NASA/GSFC, Greenbelt, MD 20771.

although it will be shown that other vector representations are more effective in some applications. For measuring frequencies chosen in the sense of a satellite temperature sounder, the *orientation* of this vector in its space is determined by the *shape* of the profile of temperature versus height which acts, through radiative transfer physics, to produce the radiances. Thinking inversely, one may be able to infer the shape of a temperature profile by *determining* the orientation of its associated radiance representation vector, this process done without explicitly solving the inverse problem. Moreover, one should be able to determine how similar are the shapes of two different temperature profiles by measuring how closely oriented are their corresponding radiance representation vectors. If the radiance representation vectors are scaled to unit magnitude, then this degree of colinearity can be easily and simply measured by a scalar inner product of the two normalized vectors. This concept can be turned into useful shape or pattern recognition algorithms used to preprocess satellite radiometer data with the goal of producing first-guess fields with improved vertical structure.

There are two pattern recognition procedures to be developed in this paper. The first is an analog retrieval method whereby one examines the orientation of a current radiance measurement representation vector, in relation to the orientations of the members of some historical library of such vectors, to choose the one (or more) from the library which is nearly colinear with the current measurement vector. The corresponding library temperature sounding (or the mean of several) is regarded as the *analog retrieval* of the ambient temperature profile. We will show that analog retrievals are quite useful in forming good first guess fields having more representative vertical structure. The second pattern recognition procedure to be developed involves partitioning *a priori* data sets into pattern groups. Such pattern groups are alternatives to "traditional" partitioning involving latitude, season, underlying surface, etc. Any retrieval algorithm can be conditioned by pattern groups, as easily as by traditional groups, and an analog selection procedure can be implemented to select from pattern groups, if desired. The combination of using pattern recognition for forming a first guess field for a physical retrieval method will be shown to be competitive with traditionally implemented statistical methods, such as regression.

2. Pattern recognition I: Analog retrieval

a. The analog concept

The search for historical analogs to current situations is a strategy that pervades many fields of inquiry, including many nonscientific ones. The oldest and most familiar meteorological application is the so-called analog weather forecasting method which can, in one form or another, be traced back to the late 1800s. More recently, scientists have tried different pattern recognition techniques to improve weather forecasting; e.g., Lorenz (1969), Karpeyev (1969), Sonechkin (1969), Radinovic (1975), Woodcock (1980); to predict hurricane movement; e.g., Hope and Neumann (1970); to classify weather patterns; e.g., Vapnik and Romanov (1978), or short-term climate fluctuations (Barnett and Preisendorfer (1978), or again to classify cloud cover; e.g., Parikh and Rosenfeld (1978) Parikh and Ball (1980).

Thompson *et al.* (1983) presented preliminary results on an analog temperature retrieval method for satellite sounder, and Chedin *et al.* (1984) have independently developed a similar technique for improving TOVS soundings. Lipton and Vonder Haar (1983) have reported on the use of pattern recognition methods for water vapor profile retrievals. In addition, the use of structure functions [e.g., empirical orthogonal functions (EOFs)] to represent ensembles of data is now a common form of pattern sensitive analysis; it has been used explicitly in the temperature retrieval problem by Alishouse *et al.* (1967) and Smith and Woolf (1976). Jalickee and Ropelewski (1979) have used transformed (rotated) EOFs as a pattern recognition procedure to sort out low-level potential temperature profiles. Wark (personal communications, 1983) and co-workers at NOAA/NESDIS (see preceding paper) are adapting this technique to partition *a priori* data for satellite temperature sounding algorithms. Moreover, even the traditional partitioning of data by latitude, season, surface type, etc., is a structure identification strategy implicit to virtually all satellite sounding methods.

An explicit analog temperature retrieval algorithm may be viewed in the following way. Suppose a temperature profile $T(z)$ is represented in some vector space by vector $\mathbf{V} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + \cdots + v_n\mathbf{e}_n$, where the $\mathbf{e}_i$ are unit basis vectors. This representation could take several forms for the satellite temperature retrieval problem. For example, $\mathbf{V}$ could represent spectral radiances or equivalent brightness temperatures in the frequency domain of the radiometric instrument, or $\mathbf{V}$ could be a decomposition of those physical quantities into a space of theoretical or empirical orthogonal functions, or again $\mathbf{V}$ could be some representation of $T(z)$ in another convenient or useful basis. For reasons to become apparent later, it is useful to scale these vectors by a transformation involving some parameters a_i and b_i:

$$U_i = (v_i - a_i)/b_i$$

and then to normalize the result,

$$\mathbf{r} = \frac{\mathbf{U}}{|\mathbf{U}|}.$$

The vector $\mathbf{r}$ is now a unit vector and, if a_i and b_i are chosen properly, may as likely fall into any sector of the n-dimensional vector space. Given two observation vectors, $\mathbf{V}_1$ and $\mathbf{V}_2$, their colinearity may be

simply measured by the inner product of their associated "pattern vectors,"

$$(\mathbf{r}_1 \cdot \mathbf{r}_2) = \cos(\theta); \quad \theta = \text{angle between } \mathbf{r}_1 \text{ and } \mathbf{r}_2.$$

An analog temperature retrieval method can be constructed as follows: Given an historical library of temperature profiles $\{T_k(z)\}$ and associated pattern vectors, $\{\mathbf{r}_k\}$ the analog retrieval of a given object profile $\tilde{T}(z)$ whose pattern vector is $\tilde{\mathbf{r}}$ is that member of the library $T_M(z)$, whose associated pattern vector yields maximum inner product $(\tilde{\mathbf{r}} \cdot \mathbf{r}_M)$.

b. *Satellite measurement representation vector space*

The vector space in which temperature profiles are to be represented is subject to choice. For the satellite temperature sounding problem to be analyzed here, three different choices were tested: RADV will denote methods involving the spectral radiance vector itself; BRTV will denote methods involving radiance equivalent brightness temperature vector; and REOF denotes methods in which radiance vectors are projected onto empirical orthogonal functions (EOFs) of the radiance library, and pattern vectors are constructed from the resulting expansion coefficients of each radiance vector. For RADV and BRTV methods, the dimensionality of the vector space is the number of sounding channels, while for REOF the space may be collapsed to lower dimension by using only the most significant EOFs.

c. *Scaling strategy*

For RADV and BRTV, the scaling parameters, a_i and b_i, are chosen to be library means and standard deviations of the elements of the appropriate representation vectors. This choice of scaling is intended to assure, in this initial study, that each radiometric channel has an equal influence on the selection procedure.[1] For the REOF methods, some scaling[1] is implicit in the decomposition of radiance data into EOF expansion coefficients, and these coefficients are further scaled by the square root of the corresponding eigenvalue.

d. *Analog selection strategy*

The inverse satellite temperature sounding problem is ill-conditioned; i.e., two or more quite different temperature profiles might occur which produce satellite radiance measurements that are indistinguishable within the noise level of the radiometer. In an analog retrieval procedure, the larger the historical library, the greater the probability that a good analog can be found and that a bad analog can be found as well. This is the subject of a separate study to be reported elsewhere. For the present study, some strategy must be adopted to avoid the selection of bad analogs—the analog method counterpart of regularization of inverse solution methods.

An inverse satellite retrieval solution is quite different than that of a radiosonde observation (RAOB). The vertical resolving power of the satellite retrieval system is much less than that of a RAOB, and a satellite temperature retrieval is a smoothed representation of ambient thermal structure. In an analog retrieval procedure, a RAOB from a historical library is retrieved as an estimate of the existing temperature structure. Such a RAOB is rich in vertical structure consisting of some long scales which influence the radiance vector and also some shorter scales which are simply not represented in the satellite measurement. Thus, some strategy should be developed to preserve relevant vertical scales in the analog retrieval but filter out irrelevant smaller scale variation.

In this study, the strategy used to address both of these problems is to select several near-analogs from the historical library and then to construct a profile average of these. Procedurally, this means that for a given object pattern vector $\mathbf{r}$, one selects all library profiles whose inner product, $(\mathbf{r} \cdot \mathbf{r}_k)$ exceeds some critical value, called the *averaging limit,* and then averages the selected RAOBs.

e. *Properties of analog retrievals*

An ensemble of 1600 RAOBs over land or sea, during summer or winter, and between latitudes 30°S and 60°N served as the data base for this study. This data set, assembled by N. Phillips of NOAA/NMC, was interpolated to 65 levels between the surface and 1 mb. Radiances and brightness temperatures were synthesized for 11 CO_2 sounding channels of a HIRS-like sounder using transmittances as provided in the NASA/GLAS retrieval system. The channels and noise levels are the same as those used by Thompson (1982) and shown in Table 1. The data were first divided into dependent and independent sets of 800 and 799 profiles respectively; one very bad profile was discarded at the outset. Scaled and normalized RADV and BRTV pattern vectors were calculated for both sets using the component means and standard deviations of measurement vectors of the dependent data for scaling. Empirical orthogonal functions of the dependent data were used to construct REOF pattern vectors for the independent data, the pattern vectors having components of expansion coefficient

[1] The reader should not regard these choices of scaling as proven optimum choices for all applications. In a library search procedure, this scaling has an important influence on the final selection. For example, if high altitude radiometric channels were scaled up and low altitude channels scaled down, then the analog selection process would yield a profile that is good at high levels but possibly quite poor at low levels. Similarly, one can enhance the influence of certain structure functions in the REOF procedure by choosing different scaling. Scaling strategy should be developed consistent with the final goal of the search.

TABLE 1. Characteristics of eleven simulated HIRS sounder channels.

Channel index	Center (cm^{-1})	Noise ($mW\ m^{-2}\ sr^{-1}\ cm^{-1}$)
1	668.60	0.82
2	679.05	0.15
3	689.70	0.11
4	703.80	0.08
5	716.70	0.05
6	731.85	0.06
7	2192.50	0.0011
8	2211.65	0.0012
9	2237.35	0.0009
10	2271.20	0.0007
11	2506.60	0.0005

divided by the square root of corresponding eigenvalue; i.e., the standard deviations of the ensembles of expansion coefficients for the dependent set.

For each type of pattern vector—RADV, BRTV, REOF with nine functions included—all possible inner products ($\mathbf{r}_k \cdot \tilde{\mathbf{r}}_l$) and corresponding 10–1000 mb rms temperature profile differences were calculated for each member [$\tilde{\mathbf{r}}_l$, $\tilde{T}_l(z)$] of the independent ensemble using the dependent ensemble of pairs [$\mathbf{r}_k$, $T_k(z)$] as historical library. This is some 639 200 pairs of values for each type of pattern vector. Figure 1a shows a scatter diagram of these two quantities for RADV results. The contours refer to the 639 200 pairs and represent the density of data points in the diagram, scaled against uniform density. Thus, for example, a contour value of N means that points are N-times more dense along that contour than if data points were uniformly distributed over the diagram. Each of the points shown on these diagrams represents a *direct mode analog,* which is a directly selected, individual RAOB from the dependent data which has the smallest 10–1000 mb rms deviation from a given independent RAOB.

While this direct analog selection cannot be implemented operationally, it serves as a base line against which to compare the satellite selected analogs. Over large regions in the lower left and upper right, the normalized point density is <1, meaning that far fewer cases occur in these regions than if there were nothing more than a random relationship between the two measures. The orientation of the pattern from lower-right to upper-left suggests a desirable correlation such that the larger the radiance pattern vector inner product the more likely it is that the corresponding rms RAOB temperature difference will be small. On a less promising note, the large majority of points occur for fairly large inner products, suggesting that the RADV approach is not efficiently separating the two measures. Also, the mean value of the temperature profile errors is not small by the standards of temperature retrievals. For the RADV results, the scatter of data along the temperature axis is fairly broad, indicating that the radiance measure may not be able to efficiently select the best analog.

Figure 1b shows results for the BRTV analog approach. The plotting technique is identical to Fig. 1a and, further, the results are very similar to RADV results. This indicates that the scaling inherent in

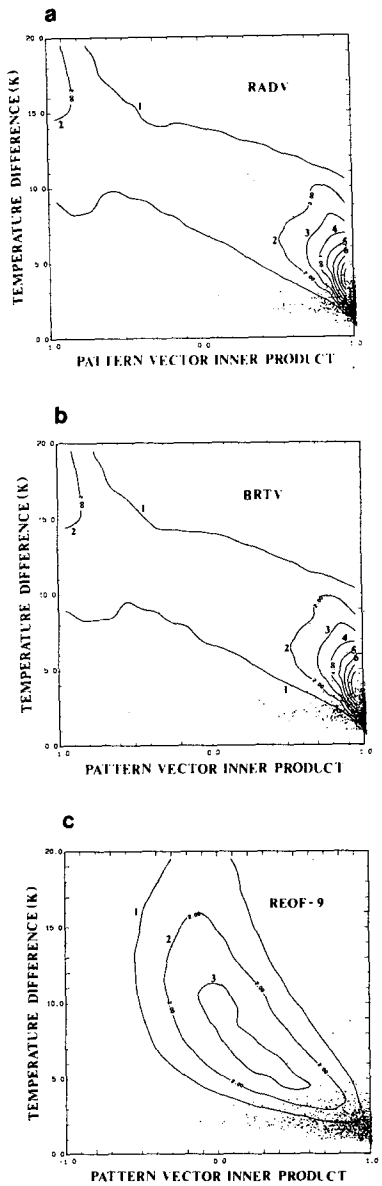

FIG. 1. Normalized point density scatter of 10–1000 mb rms differences between pairs of RAOB profiles and the inner product of pattern vectors constructed from corresponding pairs of satellite measurements. Represented are 639 200 separate pairings from an ensemble of observations between 30°S and 60°N, summer or winter, over ocean or land. Dots are best analog pairs (see text). (a) Spectral radiance pattern vector, RADV; (b) equivalent brightness temperature pattern vector, BRTV; (c) empirical orthogonal function pattern vector using 9 functions, REOF-9.

conversion of radiance to brightness temperature does not have much impact on this method.

Figure 1c shows results for the REOF analog method. The graph here is remarkably different from those of Figs. 1a, b. Most notably, the maximum normalized point density occurs nearer to the center of the diagram for the REOF approach and the spread of the pattern is broader. The difference between these patterns can be interpreted theoretically, leading to useful insight into optimum analysis of data. A theoretical analysis, given in the Appendix, is important to readers who would like a deeper insight into analog methods, but not essential to readers wishing only to appreciate results.

Due to the scatter of rms temperature difference values for the given values of inner product shown in Fig. 1, it seems clear that choosing a single analog would be a risky procedure since it may be either very good or very bad. Thus, experiments were performed to decide how one might select a small batch of near-analogs for averaging, as discussed earlier. Analog methods, REOF and RADV, were applied to a subensemble of winter cases over ocean between latitudes 30 and 60°N. Analog retrievals were performed using various averaging limits (i.e., minimum value of $(\mathbf{r}_k \cdot \tilde{\mathbf{r}}_l)$ for screening). The REOF method was implemented using projections onto 3, 4, · · ·, eleven EOFs for constructing pattern vectors. Figure 2 shows the 10–1000 mb rms analog retrieval

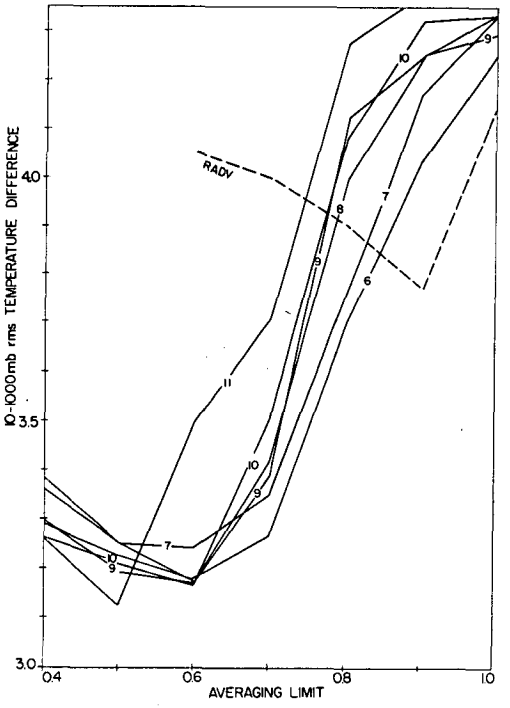

FIG. 2. Mean analog retrieval errors, 10–1000 mb, for RADV and REOF methods vs averaging limit. Winter cases between 30°N and 60°N over oceans.

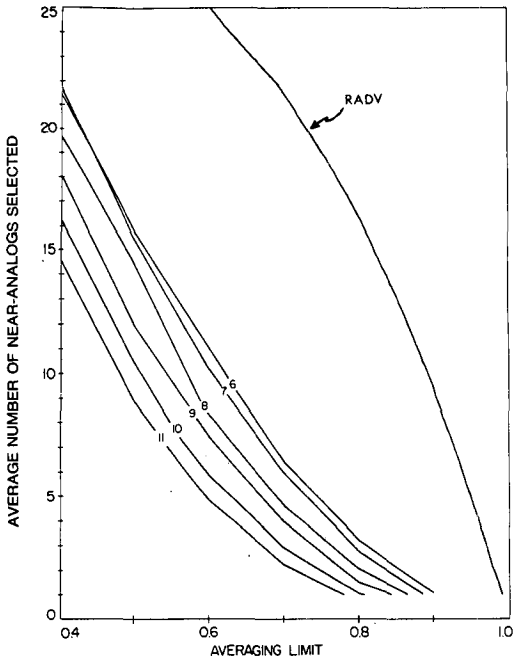

FIG. 3. Number of near-analogs average together to produce ad REOF analog retrieval vs averaging limit; test sample as in Fig. 2.

error as a function of averaging limit and number of EOFs for REOF. This should be examined in conjunction with Fig. 3, which shows the mean number of near-analogs averaged together to produce the results in Fig. 2. It is immediately clear that the REOF analog procedure is superior provided the averaging limit is set so that at least a half-dozen near-analogs are averaged together. Using radiances or brightness temperatures themselves is simply not a good approach since the channel interdependence degrades the pattern recognition power. In Figs. 2 and 3 note that truncation of the EOF projection is not a serious factor in the selection strategy. The vertical resolution of a given near-analog is that of the selected RAOB, and the vertical resolution of the final (averaged) analog retrieval is controlled by the averaging process. The truncation of EOFs affects only the analog selection process itself and, evidently, the higher-order EOFs are not very important in this selection.

In view of the fact that the minimum analog retrieval errors shown in Fig. 2 exceed 3 K, it should be recalled that the analog procedure is intended for the formation of first guess fields for a physical retrieval algorithm. The techniques provide for shape coherence but not absolute closeness of analog and object profiles. As will be shown later, the vertical shapes of these analog first-guess profiles are better than mean profiles. A physical retrieval algorithm can easily reduce the error by removing biases that are allowed to exist in the analog procedure.

3. Pattern recognition II: Pattern group partitioning

a. *Pattern partitioning as an alternative to traditional partitioning*

When retrieving information on the atmospheric structure of temperature and moisture from satellite sounder data, one finds it customary first to identify the location, season and underlying surface type of the measurement. This is done so that some appropriate batch of historical data may be used to regularize or otherwise condition the result. This traditional partitioning of data leads to more accurate soundings than no partitioning at all. Crosby and Weinreb (1974), Spänküch *et al.* (1977) and others have shown, however, that optimum results may not be obtained if the object profile to be inferred is not well represented by the historical ensemble.

In a case-study sense, one can easily imagine situations such as cold polar outbreaks, intrusions of maritime tropical air over midlatitude continental regions, episodes of "Indian summer," etc., for which traditional preliminary identification of a case by latitude, season and surface may not be optimum. In a given synoptic system, one expects the thermal structure characteristics of prefrontal and postfrontal air masses to be significantly different even if they occur in the same latitude band over the same surface type. Furthermore, temperature retrieval errors are often not random over the globe but, instead, tend to correlate with meteorological patterns.

Satellite measured spectral radiance observations (SATOBs) by themselves possess information about the vertical structure of the thermal field. An ensemble of SATOBs can be partitioned, using pattern recognition techniques, into distinguishable pattern libraries having coherent vertical structure without explicit reference to latitude, season or surface type.

The partitioning procedure to be developed here has two distinct steps. First, a dependent ensemble of corresponding RAOB–SATOB pairs will be partitioned into pattern libraries. This involves an analysis of the dependent set of RAOBs with the goal of defining distinct patterns of RAOB vertical structure. Second, an objective pattern recognition technique using SATOB data which provide that a given (independent) SATOB can be associated with one of the dependent pattern libraries will be formulated. We will develop and test this partitioning procedure using the dependent and independent ensembles of the Phillips data before proceeding with a full set of retrieval experiments combining pattern recognition with retrieval algorithms.

b. *Partitioning a priori RAOB data*

Suppose the matrix of covariance about the mean of a large heterogeneous set of temperature profiles was subjected to singular value decomposition, and the individual profiles $T_k(z)$ were projected onto the resulting empirical orthogonal eigenfunctions $\psi_l(z)$:

$$T_k(z) = \bar{T}(z) + \sum_l A_{kl}\psi_l(z). \quad (1)$$

If the data were structurally homogeneous, the statistical distribution of expansion coefficients A_{kl} would have zero mode for each l so that the mean profile $\bar{T}(z)$ is a reasonable first approximation to a large number of ensemble members. If the data set were structurally heterogeneous, then coefficient distributions might have, for each l, one or more modes significantly different from zero, implying that pattern vectors formed from expansion coefficients may be found in clusters in phase space rather than uniformly distributed about a single representative mean. If such clusters are distinct, they may be easily identified as a distinct pattern group and the vector mean for each cluster is a good approximation to members of that cluster. However, if clusters are "fuzzy" or otherwise not distinct, then partitioning becomes much more difficult.

Seven hundred ninety-eight members of the dependent data set were decomposed in the manner described here. Each RAOB member was represented by a vector with components (A_{k1}/λ_1), $(A_{k2}/\lambda_2), \cdots,$ (A_{kL}/λ_L) where λ_l^2 is the lth eigenvalue of the dependent RAOB covariance matrix. In our coordinate system, $L = 65$, although one need not consider more than two or three EOFs in any partitioning step. Figure 4 shows a vector diagram representing the first two components of these *representation vectors.* The circle has unit radius and represents the standard deviations of the scaled components in each direction. Figure 5 shows profiles of the mean and rms difference from the mean for these data. Figure 6 shows the frequency distributions of the first two expansion coefficients, scaled by the appropriate eigenvalue. Also shown here is a Gaussian distribution function scaled to the same number of profiles. Figure 7 shows profiles of the first two eigenfunctions of this data. Now, it is clear from Fig. 6 that the expansion coefficients are not randomly distributed about zero mean and mode. Thus, many profiles in the dependent data have vertical structure much different than the mean profile in Fig. 5. Figure 4 shows that the representation vectors are not uniformly distributed in direction. Further, there are more cases outside the unit circle than one would expect in a Gaussian distribution. The variance profile in Fig. 5 exhibits rather large values, and the ratio of the second eigenvalue to the first eigenvalue is roughly 1:2, which means that both structure functions in Fig. 7 are quite important in determining the structure of the set. The first EOF for this set has the apparent principal role of adjusting overall temperature to warmer or colder values, and the second EOF, along with the first, has the apparent role of making ad-

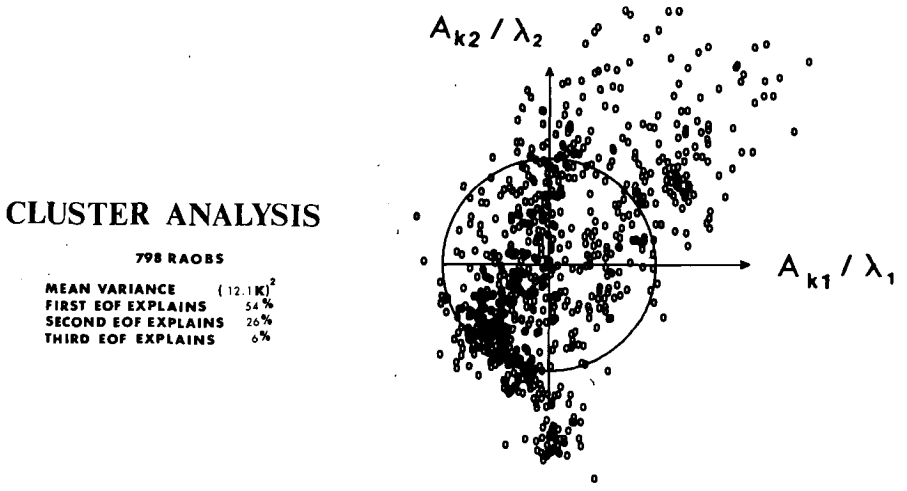

FIG. 4. Vector scatter diagram of the first two components of a temperature profile representation vector involving the projection of RAOB profiles onto empirical orthogonal expansion functions (EOFs). Sample is 798 RAOBs during summer or winter, between 30°S and 60°N, over ocean or land.

justments of tropopause height. Clearly, this batch of data contains many diverse shapes, and the mean profile is not very representative of the members.

Now, it is desirable to divide this dependent data into libraries which have more coherent structure and smaller amplitude, better behaved variation about the mean. Jalickee and Ropelewski (1979) showed a method for rotating the eigenfunctions so that, for example, two new axes could be constructed in a display such as Fig. 4 which pass "closer" to the data points than do the axes shown. In their data, however, some traditional partitioning had already been done and the resulting subensembles exhibited much sharper clustering than is apparent in our data in Fig. 4. For reasons explained earlier, we are specifically avoiding any prepartitioning by traditional classifiers so as to test pattern recognition in its most fundamental mode. Thus, we proceed, somewhat subjectively, as follows. The frequency distribution of A_{k1}/λ_1 is skewed to negative values with a hint of a secondary peak around $A_{k1}/\lambda_1 \sim 1 \cdot 1$. The dependent data were partitioned into two subensembles using the critical value $A_{k1}/\lambda_1 = 0.3$ for the separation, a value somewhere between the two local maxima. *Each resulting subensemble was then reanalyzed separately in the manner of Figs. 4–7.* This means computing new means, covariance matrices, EOFs, etc. The lower-order EOFs of these new subsets contain variance information which was embedded in the higher-order functions of the ensemble before partitioning. This is why only two or three EOFs need be considered at each partitioning step. Figures 8a–d and 9a–d show the resulting characteristics. The subensemble of Fig. 8 shows very nice properties. The total variance is much smaller than the original dependent data ensemble, and the ratio of successive eigenvalues has been reduced. The frequency distribution of the new expansion coefficients A_{k1} is much

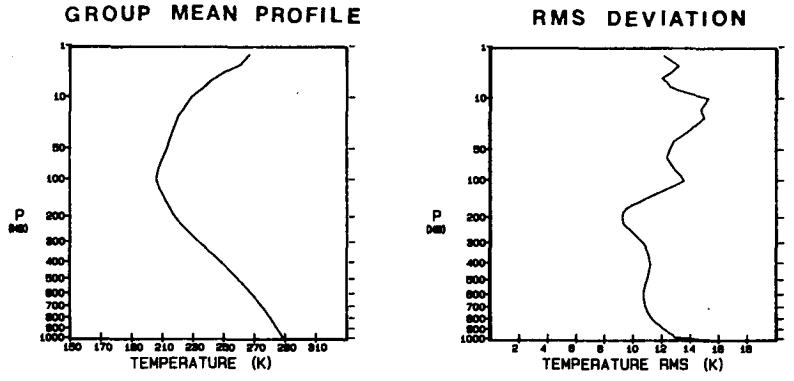

FIG. 5. Ensemble mean temperature profile and rms deviation about the mean for the data of Fig. 4.

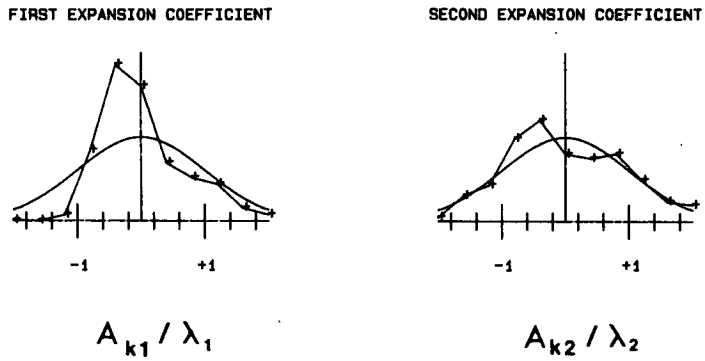

FIG. 6. Frequency distribution of the first two scaled EOF expansion coefficients of the data in Fig. 4; also shown are Gaussian curves scaled to the total number of cases.

closer to the Gaussian shape, and the vectors in the cluster diagram appear much more uniformly distributed in direction and magnitude than the total ensemble of Fig. 4. The mean profile is much different from that of Fig. 5 and seems to have a "polar flavor." Although the distribution of A_{k2}/λ_2 in Fig. 8c is imperfect, this subensemble will be set aside as one of the final "pattern libraries."

Turning to Fig. 9, we observe that the frequency distribution of expansion coefficients still show strongly non-Gaussian behavior. Even though the total variance of this subensemble is only slightly larger than that for the data in Fig. 8, further partitioning is desirable. In the vector diagram of Fig. 9, one sees two clusters of points—in the lower left quadrant and lower right quadrant—that seem to stand out from the remaining data. These two clusters were separated out of the batch and reanalyzed, and the remaining elements also reanalyzed as a separate set. There are now four separate subensembles of the original data in Figs. 4–7.

This process of data partitioning using EOF analysis of each subensemble was continued until the original dependent set was partitioned into nine separate subensembles. The details of this process are not shown here, but the idea is established. These subensembles are candidates for a pattern recognition scheme for classification of each member of the independent data into a shape library.

c. *Pattern recognition of independent satellite data*

The partitioning of the dependent RAOB data set into pattern libraries is for the purpose of classifying each independent SATOB as represented by one of the historical pattern library data sets. The pattern library data sets can then be used to derive required *a priori* data for satellite sounding. To classify radiometer measurements, one calculates a brightness representation vector for each library mean temperature profile and also for the measurement to be classified. Brightness representation vectors have components of the form

$$r_i = \frac{(T_B(\nu_i) - \bar{T}_B)}{\left[\sum_i (T_B(\nu_i) - \bar{T}_B)^2\right]^{1/2}}$$

where $T_B(\nu_i)$ represents equivalent brightness temperature for frequency ν_i, and $\bar{T}_B$ is arbitrarily chosen to

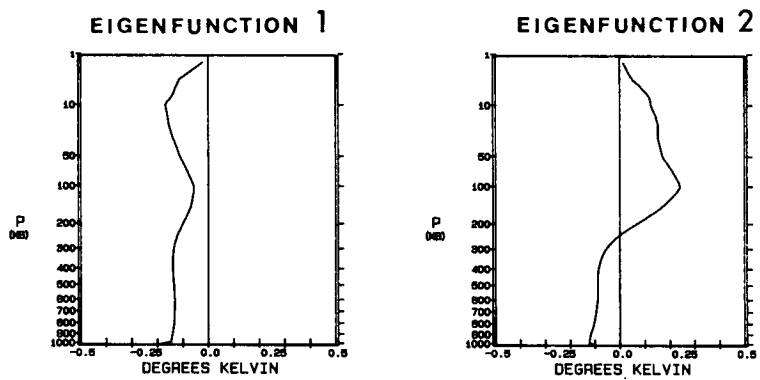

FIG. 7. First two EOFs of the data in Fig. 4.

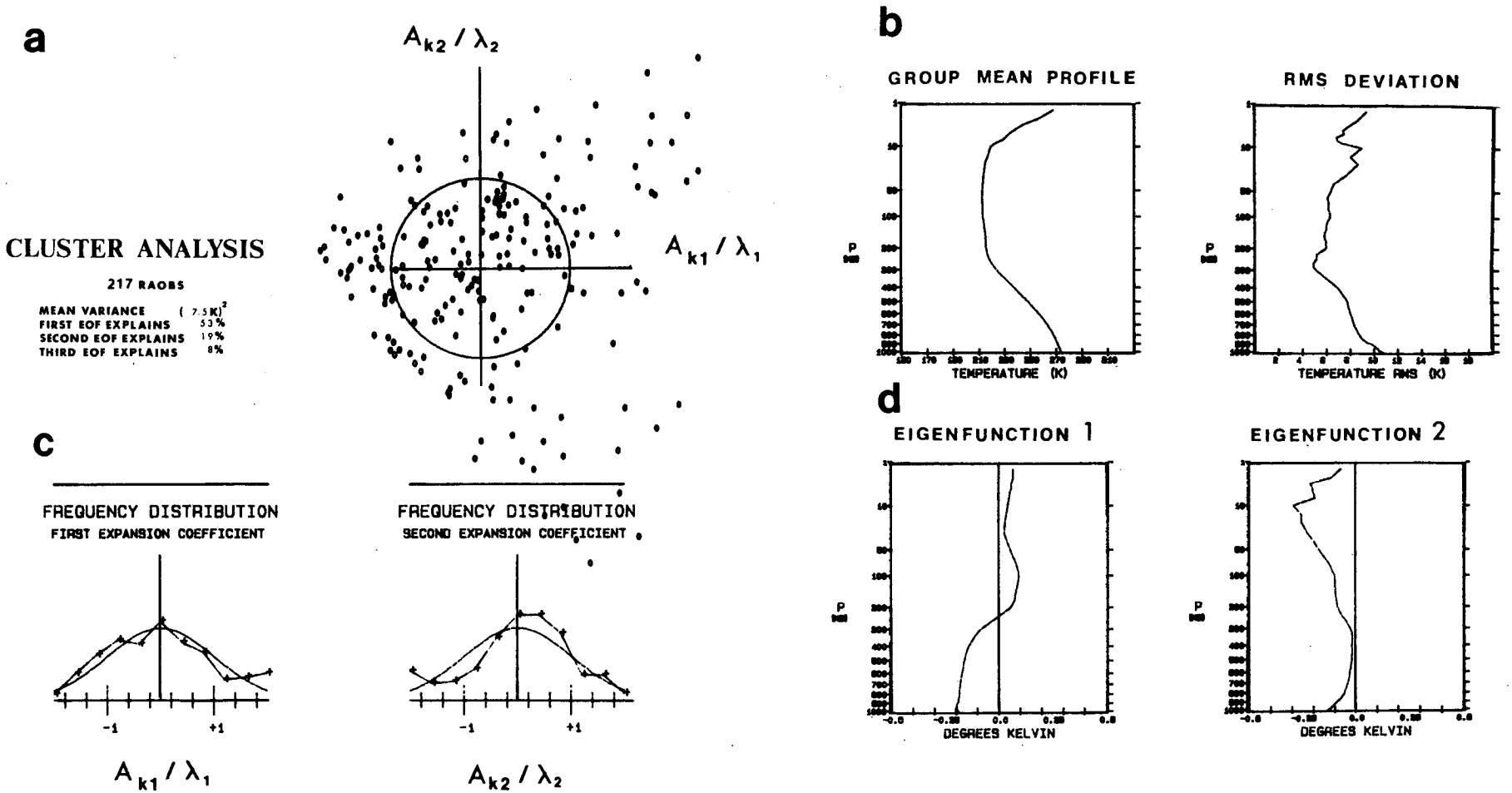

FIG. 8. As in Figs. 4–7 except for a subset of the data comprising Figs. 4–7. (See text for definition of subset composition.)

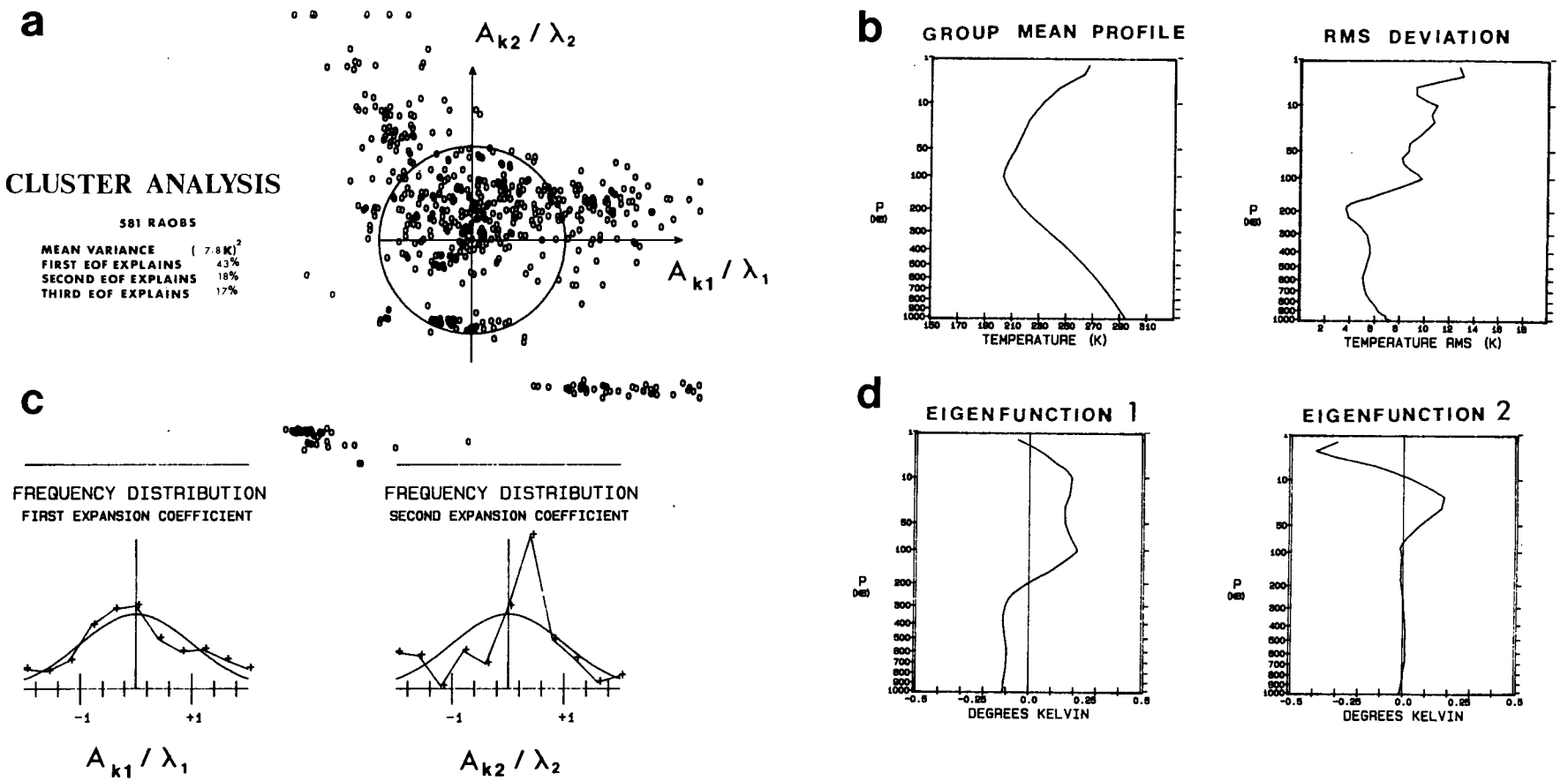

FIG. 9. As in Figs. 4–7 except for a residual subset of the data comprising Figs. 4–7 obtained by extracting the data of Fig. 8 from the data of Fig. 4.

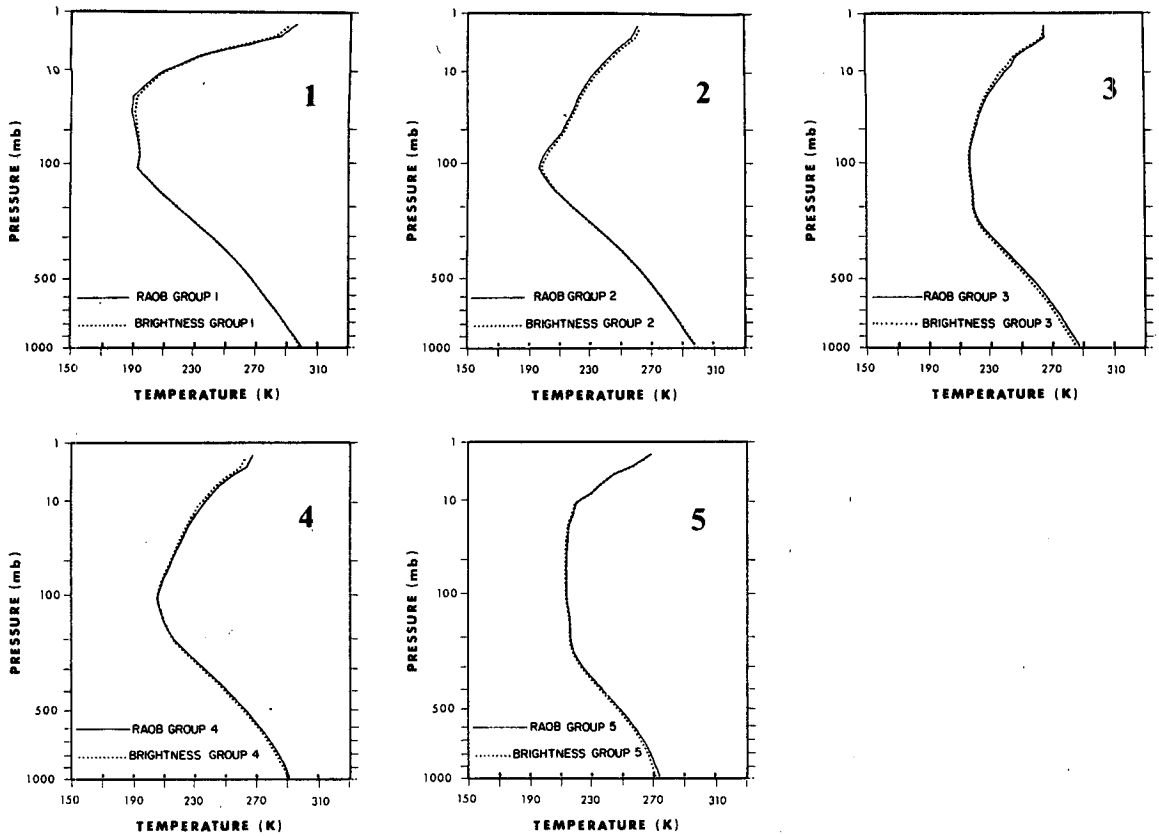

FIG. 10. Mean temperature profiles of five *a priori* data pattern libraries and five corresponding independent data pattern groups formed using pattern recognition methods. Pattern libraries are partitions of the data in Figs. 4–7.

be 270 K. The measurement is assigned to the library type whose representation vector produces the largest inner product with the measurement brightness representation vector. Once libraries are constructed, this classification of independent data can be done very economically since it involves only a small number of inner product calculations.

Operating in brightness temperature space, one finds it relevant to ask whether each of the nine subensembles mentioned previously are distinguishable in that domain. Because of limitations of their vertical resolving power, the eleven HIRS channels used in this study are not capable of clear distinction of all nine subensembles. (Clear distinction means not only that brightness differences are more than radiometer noise levels, but that inner products are not ambiguously close). Thus, the nine subensembles were reorganized into only five pattern libraries by combining certain pairs of subensembles with nearly similar mean shape. Profiles designated RAOB GROUP in Fig. 10 show the shape of the mean profile for these five libraries.

With five RAOB pattern libraries defined, the brightness pattern vector inner product partitioning algorithm was applied to the *same* data to determine how well these libraries could be separated using SATOBs. The first experiment utilized noise-free simulated measurements. Table 2 shows the number of original RAOB library cases missed and added by the brightness pattern vector partitioning for each library type. Note that while the total number of members in each brightness partitioned group is about the same as the corresponding RAOB partitioned group, there is considerable swapping in and out of groups. Libraries 2 and 4, for example, are somewhat difficult to distinguish by brightness temperatures. Of the 45 cases added to brightness group

TABLE 2. Cases missed or added when dependent data in Fig. 10 are partitioned using HIRS brightness temperature pattern vector inner products with RAOB library mean brightness pattern vectors.

	Library				
	1	2	3	4	5
RAOB partitioned	45	275	148	113	217
Cases missed	2	37	37	43	37
Cases added	8	45	39	61	3
Brightness partitioned members	51	283	150	131	183

2, 32 were from RAOB group 4, while 31 of the 61 cases added to brightness group 4 were from RAOB group 2. These two groups have similarities with group 3 also. Of the 45 additions to brightness group 2, 12 cases came from RAOB group 3. Of the 61 cases added to brightness group 4, 24 came from RAOB group 3. Further, patterns 3 and 5 swapped some members in this experiment: 30 of the 39 cases added to brightness group 3 came from RAOB group 5.

The equivalent rms deviations about the "mean" for the 798 member partitioned dependent data set were computed using either mean profiles of respective RAOB partitioned libraries or the mean profiles of respective brightness partitioned libraries. Figure 11 shows vertical profiles of those deviations integrated over the entire 798 member dependent data set. The RAOB partitioned data shows less variance above 300 mb but more below that level. Evidently, the greater number of lower atmospheric channels of the HIRS leads to a better separation of patterns at those lower levels, but suffers in its ability to partition the higher level structure. The overall 10–1000 mb mean rms deviation is 5.00 K for RAOB partitioning and 4.95 K for brightness partitioning. This same statistic computed for traditionally partitioned dependent data has value 5.63 K overall, broken down as follows: 7.10 K for winter–midlatitude, 5.16 K for summer–midlatitude, 5.10 K for winter–tropics, and 4.26 K for summer–tropics. This suggests that the swapping of library members discussed before may not be so crucial overall, but may degrade results at higher levels more than at lower levels. Presumably, the addition of the stratospheric sounding unit information would improve the brightness partitioning.

When the experiment was repeated with errors drawn from a random distribution with standard deviations given by noise values in Table 1, the brightness partitioned groups were very stable to the noise perturbations. Group 1 received only one additional member and missed none of those cases in the noiseless group. Group 2 missed one case and added none. Group 3 missed two noiseless cases and added five. Group 4 missed six noiseless cases and added one. Group 5 missed no noiseless cases and added two. Thus, while there is some swapping of the RAOB library members when partitioning is done using satellite measurements, the partitioning is relatively stable to radiance measurement errors. This is an important point, for it illustrates that the effect of radiance measurement errors on pattern partitioning is quite different from the effect of errors on inverse algorithms. As with the analog retrieval method, radiometer errors only affect a decision or selection process and do not "propagate" into the domain of the vertical temperature profile. The effect on the selection process depends simply on how distinct the libraries are (on a pattern vector inner product scale) and how much random errors impact a value of pattern vector inner product.

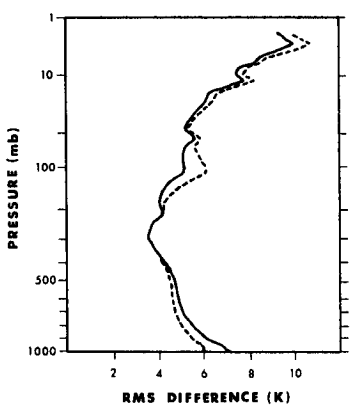

FIG. 11. Profiles of overall rms deviations of RAOBs about appropriate pattern library means for libraries partitioned using the RAOB structure (solid curve) and libraries partitioned using corresponding satellite equivalent brightness temperature structure (dashed curve).

The 800 *independent* RAOB profiles were sorted into five groups using the five pattern libraries and radiance pattern recognition method described above. The broken curves in Fig. 10 are profiles of the means of the five corresponding SATOB pattern recognized batches of independent data in comparision to the RAOB partitioned dependent data. The mean structure of the independent data groups (dotted curves), which have been sorted by satellite data pattern recognition, are virtually indistinguishable from the mean structures of the *a priori* dependent pattern libraries (solid curves), which were formed by the RAOB structure partitioning procedure. The brightness inner product pattern recognition method appears to be quite capable of partitioning the independent data into batches whose corresponding RAOBs are very similar to the dependent libraries.

It is interesting to compare the pattern recognition partitioning with the traditional partitioning from the point of view of how the two methods overlay. Table 3 shows the number of members of each of the five *a priori* pattern libraries that may be identified with different latitude bands, seasons and surface types. All library 1 members are winter, tropical profiles with about half over ocean and half over land. Library 2 profiles are mostly tropical, occurring in summer or winter over each surface type, but notice that nine of the total of 275 were drawn from middle latitudes. Library 3 is mostly middle latitude summer soundings but with 18 of 148 drawn from the tropics. Library 4 is diverse having been drawn from every traditional category except midlatitude winter over ocean. Library 5 is the most diverse with most drawn from midlatitude winter soundings but with cases from each traditional category.

TABLE 3. Comparison of members of traditional and pattern library partitions of 798 RAOB profiles. (Dependent data set)

	Library					
	1	2	3	4	5	Total
Tropical–winter–ocean	23	65	0	13	7	108
Tropical–winter–land	22	56	1	5	8	92
Tropical–summer–ocean	0	77	6	23	1	107
Tropical–summer–land	0	68	11	13	1	93
Midlatitude–winter–ocean	0	0	3	0	97	100
Midlatitude–winter–land	0	1	0	3	96	100
Midlatitude–summer–ocean	0	1	62	31	5	99
Midlatitude–summer–land	0	7	65	25	2	99
Total	45	275	148	217	113	798

Before we leave the subject of pattern recognition partitioning, attention is drawn to the vector diagram in Fig. 4, reprinted as Fig. 12, but with the five pattern libraries identified as well as can be done without the benefit of color. Even with "fuzziness" of the library clusters, the technique has produced a reasonably coherent separation of a messy array of data points.

4. Temperature retrieval incorporating pattern recognition

a. Combining pattern recognition with retrieval algorithms

Up to this point, two pattern recognition procedures have been developed to improve first guess fields for temperature retrieval algorithms. We may now carry out retrieval experiments on the independent data to test the impact of the procedures on retrieval accuracy.

Six different retrieval systems will be tested on the 800 SATOBs and RAOBs comprising the independent data set. The six retrieval systems are based upon two familiar retrieval algorithms with traditional or pattern recognition preanalysis used to form first guess fields and statistics as necessary. The algorithms used are the regression retrieval algorithm of Smith *et al.* (1970) and the physical iterative algorithm of Smith (1970) as modified by Smith and Woolf (1981). This last algorithm is very similar to the iterative algorithm implemented in the NASA/GLAS temperature retrieval system (see Susskind *et al.*, 1982). These two algorithms were chosen since they represent widely different uses of *a priori* data and spectral transmittance information. These two retrieval algorithms can be implemented using different treatments of *a priori* data for derivation of first-guess fields and statistics. Thus, the six retrieval systems to be tested are denoted: REGR–TRAD–MEAN, REGR–PTRN–MEAN, SMITH–TRAD–MEAN, SMITH–PTRN–MEAN, SMITH–PTRN–ANLG, SMITH–PTRN–DRCT. The first item refers to either regression or Smith physical iterative algorithms. The second label refers to either traditional (TRAD) partitioning of dependent and independent data into tropical (30°S to 30°N) or midlatitude (30 to 60°N), summer or winter, land or oceanic categories, or to pattern (PTRN) partitioning of *a priori* data into the five pattern libraries of Fig. 10 and the use of brightness temperature vector inner products to associate each independent case with a pattern library. The third item in the label refers to first-guess profile and is either the use of an *a priori* ensemble mean (MEAN) profile as first guess, or an analog (ANLG) first guess formed using the REOF analog method truncated at nine functions and using an averaging limit of 0.6, or using the direct mode analog (DRCT). The REGR–TRAD–MEAN and SMITH–TRAD–MEAN are the standard retrieval systems against which the application of pattern recognition techniques are to be compared.

In implementing the Smith iterative algorithm, we used a test for convergence in order to terminate the iterations. If $\epsilon_i^n = (R_i^n - \tilde{R}_i)/S_i$ where R_i^n represents radiance at frequency ν_i produced by the nth iteration of the temperature profile, $T^n(z)$, $\tilde{R}_i$ the measured radiance at frequency ν_i, and S_i is noise level of the ν_i radiometer channel, then iterations are terminated whenever

$$\left[\frac{1}{I}\sum_{i=1}^{I}(\epsilon_i)^2\right]^{1/2}$$

reached the value 1 or did not change significantly with further iteration. In addition, since the RAOB profiles of the independent test set are known, a second test was applied at each step in which iteration was halted at that step if the 10–1000 mb rms

FIG. 12. As in Fig. 4 except that members of the five pattern partitioned libraries are shown.

temperature difference between the actual RAOB and the current iterative estimate increased. This second test was applied simply to avoid diverging iterations and to obtain the most accurate iterative retrieval, subject to the radiance convergence limit. For both retrieval algorithms, a quality control check was applied in which a retrieval was thrown out if the algorithm produced a retrieval for which

$$\left[\frac{1}{I}\sum_{i=1}^{I}(\epsilon_i)^2\right]^{1/2} \geq 5.$$

While this criterion is somewhat arbitrary, it eliminated a small percentage of cases from each experiment for which retrieval errors were unreasonably large, although it is sufficiently liberal to admit significant retrieval error.

Radiance measurements were simulated from RAOBs using the radiative transfer equation. The fairly mild effects of radiometer noise on the pattern recognition schemes to be implemented have already been discussed. Differences in noise effects on the regression and iterative algorithms are not the subject of the present study. Thus, the retrieval comparisons in this paper were carried out for noise-free simulated measurements so that the impact of pattern recognition on the retrieval algorithms could be assessed free of other practical complications. Surface temperatures were not retrieved from satellite radiances, but the surface contributions were subtracted exactly from simulated measurements so as to isolate the integral atmospheric contribution to the measurement. Clouds, water vapor, ozone and trace gas effects were ignored in the simulations. While all these simplifications compromise the immediate applicability of results to operational retrieval systems, this study was designed to evaluate fundamental concepts and to show guidance as to the potential advantage of pattern recognition.

For the retrieval experiments carrying labels PTRN and ANLG, the satellite data is subjected to preliminary pattern recognition steps before the retrieval algorithms are applied. For ANLG first guess, the REOF analog procedure was restricted to the appropriate *a priori* pattern group identified for that case.

b. Results

Figure 13 shows first-guess error profiles using several methods for first-guess production. The curve labeled TRAD–MEAN represents the overall batch rms profile of first-guess errors if the mean profile of the appropriate traditional *a priori* ensemble is used as first guess for each of the 800 independent cases. The 10–1000 mb rms value of this first-guess error profile is 5.75 K. In a similar manner, the curve labeled PTRN–MEAN represents first-guess errors when the mean profile of the appropriate *a priori* pattern library is used as first guess once each independent SATOB has been identified with a library by pattern recognition. The 10–1000 mb rms value of this first-guess error profile is 5.27 K. The curve labeled PTRN–ANLG represents first guess errors when REOF analog retrievals are drawn from the appropriate *a priori* pattern library once each independent SATOB has been pattern-recognized. The 10–1000 mb rms value of this first guess error is substantially reduced to 3 K, comparable to the winter, midlatitude, oceanic example of Fig. 2. The PTRN–DRCT curve is analogous to PTRN–ANLG except that a single, direct mode analog is drawn instead of forming an REOF analog. This experiment is included as a baseline test: direct mode analogs cannot be produced operationally.

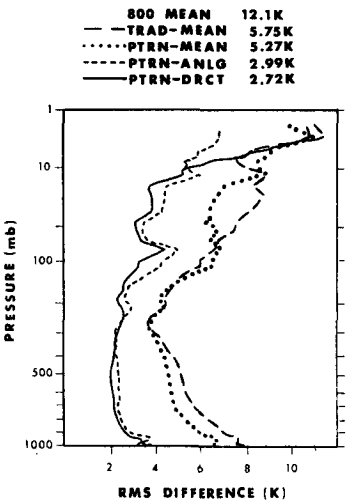

FIG. 13. First-guess profile errors for 800 independent cases using various possible first-guess selection methods.

The next step is to test the influence of these improvements in first-guess profiles on the retrieval algorithms. Figure 14 shows overall retrieval error statistics using REGR–TRAD–MEAN and REGR–PTRN–MEAN systems. Interestingly, even though pattern recognition has improved the first-guess field, the overall accuracy of the two methods is similar with the traditional regression approach showing a slight advantage both in accuracy and yield. The REGR–PTRN–MEAN retrievals were done using only five pattern groups instead of the traditional eight groups for REGR–TRAD–MEAN. The results clearly demonstrate that even a simple pattern recognition partitioning scheme is competitive with traditional partitioning. Furthermore, since the two partitioning procedures tend to overlay each other to some extent, large differences in overall retrieval results should not be expected unless pattern recognition partitioning were an utter failure. Nevertheless, it is significant that the regression algorithm is not improved by better first-guess properties as provided by the pattern partitioning implemented in this study.

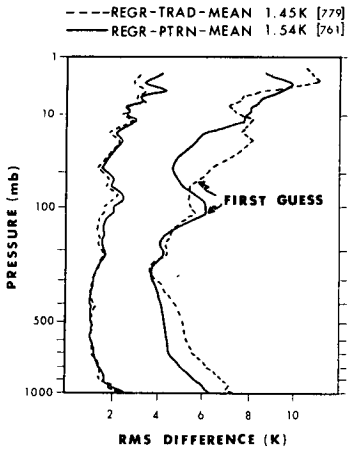

FIG. 14. Overall first-guess and temperature retrieval errors for the regression algorithm using traditional and pattern group partitioning of *a priori* and independent data.

There were 753 successful SMITH–TRAD–MEAN retrievals with overall 10–1000 mb rms error of 2.58 K. The SMITH–PTRN–MEAN experiment produced 752 retrievals with slightly improved overall error of 2.51 K. The influence of pattern recognition is best illustrated in Fig. 15, which compares SMITH–PTRN–ANLG with SMITH–TRAD–MEAN and REGR–TRAD–MEAN. A significant improvement to the SMITH algorithm occurs using both the library identification routine and the analog first guess retrieval. Overall 10–1000 mb rms retrieval error was reduced from 2.58 K for SMITH–TRAD–MEAN to 1.92 K for SMITH–PTRN–ANLG. The yield was greater for SMITH–PTRN–ANLG, with 31 more retrievals than SMITH–TRAD–MEAN, thus making the reduction in retrieval error even more significant. While the error level here is not as low as with regression, the performance gap between the physical method and statistical method has been substantially reduced by pattern recognition pre-processing of *a priori* data.

Although not shown in figure form, the overall 10–1000 mb rms retrieval errors for SMITH–PTRN–DRCT was 1.96 K for 771 cases. As mentioned earlier, the SMITH–PTRN–ANLG results exceed this baseline result mainly because averaging near analogs filters out small vertical scales which are not detectable in satellite data while preserving the significant shape structure in the first-guess field.

c. An interesting special case

Pattern recognition has shown a definite impact on the Smith physical retrieval algorithm, even when summarized in an overall batch sense on a large, heterogeneous independent data set. We expect that the impact may be even larger on particular subsets for which traditional first-guess fields are poor estimates of the target cases to be retrieved. While we wish to avoid selecting out a few "best cases" which may give a biased interpretation, we note that library 1 had the most unusual shape of all libraries, exhibiting high tropopause and deep, nearly isothermal lower stratosphere. Error statistics for traditional retrieval methods were computed for independent data group 1 and compared with pattern partitioned methods. Figure 16 shows individual first guess and retrieval errors (rms over height between 1000 and 10 mb) for the 39 cases of group 1. In 36 of 39 cases, the PTRN–MEAN first guess profile is significantly better than the traditional first guess. Moreover, in those three cases for which the PTRN–MEAN first guess fails badly, the traditional batch mean first guess is anomalously good. When the Smith algorithm is applied, the SMITH–PTRN–MEAN method dramatically outperforms SMITH–TRAD–MEAN in all cases except those for which PTRN–MEAN first guess is worse than the traditional first guess. When the regression algorithm is applied, the pattern approach is superior in 32 cases but very problematic in the remaining seven cases. Note particularly that the three for which PTRN–MEAN first guess was so poor are handled much better by the traditional regression. The reason is simple: the 45 member *a priori* pattern library 1 clearly did not have members representative of these three cases, whereas the two traditional libraries did; (108 tropical–winter–ocean members and 92 tropical–winter–land members). Since the regression algorithm is less sensitive to first guess, but crucially dependent on *a priori* statistics, it easily retrieved these cases in the traditional approach because they were "represented" in the *a priori* data, even as "outriders." The lesson appears to be that partitioning *a priori* data for a regression algorithm

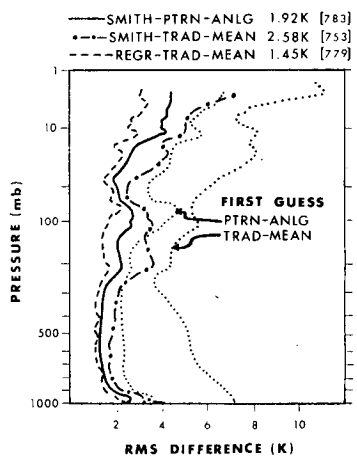

FIG. 15. Overall first-guess and temperature retrieval errors for the Smith algorithm using pattern group partitioning and REOF analog method for first guess selection. Results compared with the traditional regression and Smith iterative approach.

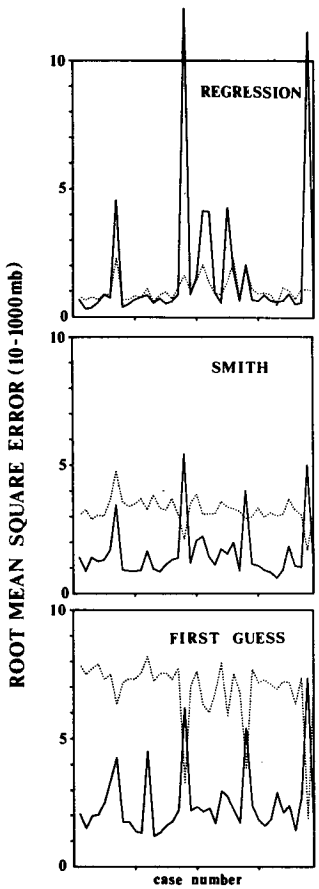

FIG. 16. Root-mean-square errors, between 10 and 1000 mb, for the 39 individual independent cases of pattern group 1: pattern group partitioning and corresponding pattern library mean profile as first guess (solid curves); traditional partitioning using either tropical winter or tropical summer *a priori* data ensembles as appropriate for each case (dotted curves).

should be a quite different process from partitioning for a physical algorithm owing to the different influences of first guess and variance about first guess. It will be interesting to study this matter further and to test pattern recognition procedures with a mixed physical–statistical algorithm such as the minimum rms method of Strand and Westwater (1968).

Table 4 shows a summary of retrieval experiments broken down by independent pattern groups. The REGR–PTRN–MEAN experiment showed best results for group 3 and worse results for groups 4 and 5. SMITH–PTRN–MEAN, SMITH–PTRN–ANLG and SMITH–PTRN–DRCT showed their lowest retrieval errors in pattern groups 1 and 2 and highest errors in group 5. The REGR–TRAD–MEAN and SMITH–TRAD–MEAN experiments were not summarized over pattern groups except for group 1 as discussed above. Note that for 37 accepted cases of group 1, the SMITH–PTRN–ANLG method produced a 1.6 K improvement over SMITH–TRAD–MEAN.

5. Conclusions and recommendations

In this study, we have demonstrated the impact of two satellite data pattern recognition preprocessing schemes on the accuracy of remote temperature soundings. The pattern recognition schemes are efficient in improving first-guess fields. Tested on an independent data ensemble of 800 soundings in winter or summer, between 30°S and 60°N, pattern recognition yielded an overall reduction of 10–1000 mb first guess errors from about 5.75 K to about 3 K. Physical (Smith iterative) and statistical (regression) retrieval algorithms were applied to the improved first-guess fields and compared with a traditional approach. The ability of these algorithms to capitalize on improvements in first guess is markedly different. The first-guess improvement has a significant impact on the physical retrieval method which uses only a first-guess field derived from *a priori* data. The impact on a purely statistical scheme is slightly negative in an overall batch sense but positive in a subset of cases for which traditionally derived *a priori* subsets are not as representative as pattern recognized *a priori* subsets. The regression algorithm provides good batch-error statistics if the *a priori* data subsets are sufficiently heterogeneous to be representative of a large number of cases to be retrieved, even if total variance of those subsets is larger than, say, a pattern recognized subset. A pattern recognition regression approach seems also to do well provided the object cases are well represented by pattern partitioned *a priori* data.

Neither of the retrieval algorithms used here are formally derived to optimize the influence of the first-guess field, although Fleming and Smith (1972) have shown that the Smith physical algorithm should have strong dependence on first-guess fields similar to the minimum information algorithm, as optimized by Smith *et al.* (1972). Due to the significantly different roles of *a priori* data in these methods, we are led to the following speculation. Given the connection between the Smith *et al.* optimized minimum information solution and the minimum rms solution of Foster (1961) and given that the improvement of the Smith algorithm shown in this study is far more dramatic than the slight degradation of the regression results, we postulate that the pattern recognition approach applied judiciously to a minimum rms algorithm should emphasize the positive aspects of pattern recognition and deemphasize the negative aspects.

Partitioning the *a priori* data by pattern recognition has shown promise but we are the first to admit that the method can be improved. The subjectivity involved in *a priori* RAOB data separation should be improved. That we were successful in our empirical approach should stimulate the development of a more theoretically based, objective pattern partitioning procedure. As discussed in this paper, the development of a pattern partitioning procedure should be linked

TABLE 4. Summary of first-guess errors (F), retrieval errors (R), and yield (Y) for all experiments.

Independent cases				Independent pattern groups 1 (39)	2 (291)	3 (164)	4 (139)	5 (167)	Total (800)
TRAD	MEAN	REGR	Y	37	—	—	—	—	779
			F	4.89	—	—	—	—	5.51
			R	1.07	—	—	—	—	1.45
		SMITH	Y	36	—	—	—	—	753
			F	4.91	—	—	—	—	5.52
			R	2.90	—	—	—	—	2.58
PTRN	MEAN	REGR	Y	37	271	156	133	164	761
			F	2.35	3.67	4.66	3.87	6.90	4.87
			R	1.41	1.44	1.26	1.68	1.67	1.54
		SMITH	Y	36	261	163	130	162	752
			F	2.36	3.55	4.95	3.90	7.02	4.96
			R	1.46	2.09	2.54	2.45	3.03	2.51
	ANLG	SMITH	Y	37	280	164	135	167	783
			F*	233	2.87	3.00	2.88	3.11	2.99
			R	1.29	1.77	1.82	2.01	2.11	1.92
	DRCT	SMITH	Y	36	279	162	131	163	771
			F*	1.66	2.81	2.58	2.62	2.82	2.72
			R	1.29	1.81	1.94	2.00	2.20	1.96

* ANLG and DRCT first-guess error statistics are shown for *all* members of each pattern group, even though some members are rejected by the retrieval scheme.

to a specific retrieval algorithm, for the question of optimality of *a priori* data subsets is very much a function of how the algorithm takes information from the *a priori* data. Next, the identification of object cases with some *a priori* pattern library does not seem to be so efficiently done by radiances or brightness temperatures. The dramatic differences in radiance, brightness temperature and EOF-based analog retrievals indicate that the vector space in which pattern recognition operates is crucial. We expect that pattern recognition of independent SATOBs would be much more effective if an orthogonal basis were used for pattern vectors. It is important to note that neither traditional partitioning nor our pattern partitioning here produces *a priori* pattern libraries whose means are orthogonal in any sense.

The analog retrieval method developed in this study shows very good promise for construction of appropriate first-guess fields. An important result is that an analog search procedure operating in an orthogonal basis is much more powerful than one operating in the obvious, but skewed and overspanned space of spectral radiances.

Finally, when different retrieval schemes are applied to large batches of cases, one often finds that some cases are handled better by one method than another. Pattern recognition is no different: in some cases, it works extremely well; in other cases, it may fail miserably. Therefore, a decision step should be implemented in which different methods are used for different cases. Pattern recognition schemes can be used to make this basic decision as well as to process those cases for which use of pattern recognition techniques is indicated.

Acknowledgments. The authors are grateful for the assistance of J. Susskind, NASA/GLAS, and L. McMillin, NOAA/NESDIS for providing information on HIRS transmittances and noise estimates. They also acknowledge the careful preparation of RAOB data by N. Phillips, NOAA/NMC which we borrowed for our tests. The careful and thorough review by three anonymous scientists is also gratefully acknowledged and appreciated. This research was sponsored by NASA through Grants NSG-5209 and NAG5-292.

APPENDIX

Frequency Distribution of Inner Products of Unit Vectors Uniformly Distributed in *n*-dimensional Space

Figures 1a, b, c show the distribution of radiance pattern vector inner products versus rms temperature profile deviations for nearly 640 000 comparisons of pairs of corresponding SATOBs and RAOBs in a particular finite data ensemble. To interpret the pattern of results, one must form some mental "model" of what results should look like for an infinite population. There are two issues: 1) what a random distribution should look like and 2) how a meteorologically organized natural distribution, such as Figs.

1a, b, c, compares. Fundamental to the concept of an analog selection method is the idea that some n-dimensional unit radiance pattern vector can be constructed which is a *random* vector ranging over the surface of an n-dimensional hypersphere. It is not clear what kind of pattern vector in the satellite radiance domain has such a random distribution. The radiance pattern vectors used in Figs. 1a, b, c are candidates, but the frequency distribution of pattern vector inner products is markedly different: very few inner products occur near zero in the RADV and BRTV experiments, whereas the REOF approach produces maximum frequency of occurrence at small, slightly positive values. Further, whether or not the concentration of points shown in the figures is to be expected requires a model of the frequency distribution of inner products of unit vectors which are uniformly distributed on an n-dimensional hypersphere.

Suppose a population of n-dimensional vectors is uniformly distributed over the surface of an n-dimensional hypersphere such that the density of endpoints on the surface is uniform. Choose a particular vector and consider the frequency distribution of inner products with all other vectors on the sphere. Denoting the direction of the particular vector as x, we observe that the number of vectors on the hypersphere producing inner product between $\cos(\theta_1)$ and $\cos(\theta_2)$ is proportional to the area of the "latitudinal band" on the hypersphere defined by θ_1 and θ_2. One can calculate this area by calculating the area of the "polar cap" defined by any angle θ and evaluating the difference for any two angles θ_1 and θ_2.

If $\sigma(\theta)$ denotes the surface area of the polar cap between $\theta = 0$ and $\theta = \theta$ on an n-dimensional hypersphere of unit radius, then the fraction of total surface area is

$$\frac{\sigma(\theta)}{\sigma(\pi/2)} = \frac{\int_{r=0}^{\sin(\theta)} r^{n-2}dr/[1 - r^2]^{1/2}}{\int_{r=0}^{1} r^{n-2}dr/[1 - r^2]^{1/2}}.$$

With change of variables to $r = \sin(\beta)$, the integration submits to the reduction formula

$$\int \sin^m(\beta)d\beta = -(1/m)\sin^{m-1}(\beta)\cos(\beta) + [(m-1)/m]\int \sin^{m-2}(\beta)d\beta.$$

Thus, the frequency distribution of vector inner products can be determined by evaluating the integration for polar caps defined by $\theta_1, \theta_2, \cdots$ and subtracting results to define values for latitudinal strips, $\theta_2 - \theta_1$, $\theta_3 - \theta_2$, etc. A reviewer notes that the integral derivation is a standard result (cf Taylor, 1955), and that the derivative of $\sigma(\theta)/\sigma(\pi/2)$ may be explicitly written as $(2/\sqrt{\pi})\sin^{n-2}(\theta)/\{\Gamma[(n-1)/2]/\Gamma[n/2]\}$ for $n > 3$. Clearly, the resulting distribution is a function of dimension n.

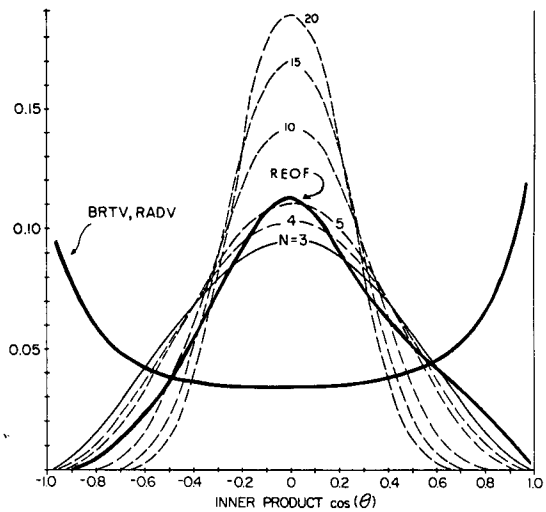

FIG. A1. Theoretical probability density function of inner products of random n-dimensional unit vectors. Also shown are empirical frequency distributions of a natural sample of RADV and REOF pattern vector inner products.

Figure A1 shows the theoretical frequency distribution of vector inner products for random n-dimensional unit vectors as a function of vector space dimension. Also shown are the empirical frequency distributions for results in Figs. 1a, b, c. That the RADV and BRTV pattern vector inner product distributions are far different from the theoretical curves is due to at least two reasons: first, the 11-dimensional frequency space of these vectors is not truly 11 orthogonal directions since the HIRS weighting functions are not linearly independent. The organization of information by sounding frequency means that one cannot even seek a lower-order orthogonal set simply by omitting channels; second, even though pattern vector components are scaled by subtracting mean values and dividing by standard deviation, this does not guarantee that the resulting components are random variables with mean and mode zero. For a collection of unit vectors to be random in the sense of the theoretical calculation, each component must be random about zero, and further, the components must not be correlated. The radiances or brightness temperatures do not have this property because the atmosphere exhibits particular modes of behavior.

The frequency distribution for the REOF pattern vectors is much more suggestive of the theoretical curves. Because the radiance measurement information is projected onto an efficient set of truly orthogonal basis functions, the frequency of inner-product values in this method achieves a maximum more toward the central zero value. The shift away from center and the asymmetry of the distribution show that the atmospheric thermal field, even in this rep-

resentation, is not random owing to the occurrence of particular meteorological modes of behavior. It is to be remembered that the data base for these calculations is an ensemble of cases between 30°S and 60°N during winter or summer, over land or ocean. Of the three approaches tested here, the REOF approach comes closest to having desirable properties for an analog retrieval method.

REFERENCES

Alishouse, J., L. Crone, H. Fleming, F. Van Cleef and D. Wark, 1967: A discussion of empirical orthogonal functions and their application to vertical temperature profiles. *Tellus,* **19,** 477–481.

Barnett, J. P., and R. W. Preisendorfer, 1978: Multifield and analog prediction of short-term climate fluctuations using a climate state vector. *J. Atmos. Sci.,* **35,** 1771–1787.

Chedin, A., N. A. Scott, C. Wahiche, P. Moulinier, N. Husson and I. Cohen-Hallaleh, 1984: The improved initialization inversion ("3I") method: A high resolution physical method for temperature retrievals from the Satellites of the TIROS-N Series. *Proc. Conf. Satellite/Remote Sensing and Applications,* Clearwater Beach, Amer. Meteor. Soc., 39–44.

Conrath, B. J., 1972: Vertical resolution of temperature profiles obtained from remote sensing measurements. *J. Atmos. Sci.,* **29,** 1262–1272.

Crosby, D. S., and M. P. Weinreb, 1974: Effect of incorrect atmospheric statistics on the accuracy of temperature profiles derived from satellite measurements. *J. Statist. Comput. Simul.,* **3,** 41–51.

Fleming, H. E., and W. L. Smith, 1972: Inversion techniques for remote sensing of atmospheric temperature profiles. *Proc. Fifth Symp. on Temperature,* Washington, D.C., Inst. Soc. Amer., 400 Stanwix St., Pittsburgh, PA, 2239–2250.

Foster, M., 1961: An application of the Weiner-Kolmogorov smoothing theory to matrix inversion. *J. SIAM,* **9,** 387–392.

Fritz, S., 1977: Temperature retrievals from satellite radiance measurements—An empirical approach. *J. Appl. Meteor.,* **16,** 172–176.

Gage, K. S., and J. L. Green, 1982a: A technique for determining the temperature profile from VHF radar observations. *J. Appl. Meteor.,* **21,** 1146–1149.

——, and ——, 1982b: An objective method for the determination of tropopause height from VHF radar observations. *J. Appl. Meteor.* **21,** 1150–1154.

Gautier, D., and I. Revah, 1975: Sounding of planetary atmospheres—A Fourier analysis of the radiative transfer equation. *J. Atmos. Sci.,* **32,** 881–892.

Hope, J. R., and C. J. Neumann, 1970: An operational technique for relating the movement of existing tropical cyclones to past tracks. *Mon. Wea. Rev.,* **98,** 925–933.

Jalickee, J. B., and C. F. Ropelewski, 1979: An objective analysis of the boundary layer thermodynamic structure during GATE. Part I: Method. *Mon. Wea. Rev.,* **107,** 68–76.

Karpeyev, G. A., 1969: Problem of similarity of meteorological fields and classification of atmospheric processes as one of the problems of pattern recognition theory. *Meteor. Gidrol.,* **12,** 35–40.

Lipton, A. E., and T. H. Vonder Haar, 1983: A classification–discrimination method for retrieving moisture profiles from satellite radiances. *Proc. Fifth Conf. on Atmospheric Radiation,* Baltimore, Amer. Meteor. Soc., 54–57.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturlly occurring analogues. *J. Atmos. Sci.,* **26,** 636–646.

Parikh, J. A., and A. Rosenfeld, 1978: Automatic segmentation and classification of infrared meteorological satellite data. *IEEE Trans. Syst. Man. Cybernetics SMC,* **8,** 736–743.

——, and J. T. Ball, 1980: Analysis of cloud type and cloud amount during GATE from SMS infrared data. *Remote Sens. Env.,* **9,** 225–245.

Radinovic, D., 1975: An analogue method for weather forecasting using the 300/1000 mb relative topography. *Mon. Wea. Rev.,* **103,** 639–649.

Rodgers, C., 1970: Remote sounding of the atmospheric temperature profile in the presence of cloud. *Quart. J. Roy. Meteor. Soc.,* **96,** 654–666.

Smith, W. L., 1970: Iterative solution of the radiative transfer equation for the temperature and absorbing gas profile of an atmosphere. *Appl. Opt.,* **9,** 1993–1999.

——, 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 ——, 1981: Algorithms used to retrieve surface-skin temperature and vertical temperature and moisture profiles from VISSR Atmospheric Sounder (VAS) radiance observations. *Proc. Fourth Conf. on Atmospheric Radiation,* Amer. Meteor. Soc., 13–17.

——, —— and W. J. Jacob, 1970: A regression method for obtaining real-time temperature and geopotential height profiles from satellite spectrometer measurements and its application to NIMBUS-III SIRS observations. *Mon. Wea. Rev.,* **98,** 582–603.

——, —— and H. E. Fleming, 1972: Retrieval of atmospheric temperature profiles from satellite measurements for dynamical forecasting. *J. Appl. Meteor.,* **11,** 112–122.

Sonechkin, D. M., 1969: Mathematical classification theory and its application to meteorology. *Meteor. Gidrol.,* **12,** 24–34.

Spankuch, D., Y. M. Timofeyev and J. Guldner, 1977: Comparison of different inversion method for the determination of vertical temperature profiles from simulated satellite measurements in narrow spectral intervals. *Z. Meteor.,* **27,** 234–242.

Strand, O. N., and E. R. Westwater, 1968: Minimum rms estimation of the numerical solution of a Fredholm integral equation of the first kind. *SIAM J. Num. Anal.,* **5,** 287–295.

Susskind, J., J. Rosenfield, D. C. Reuter and M. T. Chahine, 1982: The GLAS physical inversion method for analysis of HIRS/MSU sounding data. NASA Tech. Memo. 84936, NASA Goddard Space Flight Center, Greenbelt, MD 20771.

Taylor, A., 1955: *Advanced Calculus.* Ginn, Section 13.51.

Thompson, O. E., 1982: HIRS-ATMS satellite sounding system test—Theoretical and empirical vertical resolving power. *J. Appl. Meteor.,* **21,** 1550–1561.

——, and R. Wolski, 1976: Nonlinear functions of satellite measured spectral radiance as estimators of tropopause height. *J. Appl. Meteor.,* **16,** 281–289.

——, D. Dazlich and M. Goldberg, 1983: Pattern recognition and the satellite temperature retrieval problem. *Proc. Fifth Conf. Atmospheric Radiation,* Amer. Meteor. Soc., 31–34.

Vapnik, V. N., and L. N. Romanov, 1978: The recognition of functional dependences using meteorological data. *Iz. Acad. Sci. USSR Atmos. Ocean. Phys.,* **14,** 97–100.

Westwater, E. R., and N. C. Grody, 1980: Combined surface and satellite based microwave temperature profile retrieval. *J. Appl. Meteor.,* **19,** 1438–1444.

——, M. T. Decker, A. Zachs and K. S. Gage, 1983: Ground-based remote sensing of temperature profiles by a combination of microwave radiometry and radar. *J. Climate Appl. Meteor.,* **22,** 126–133.

Woodcock, F., 1980: On the use of analogues to improve regression forecasts. *Mon. Wea. Rev.,* **108,** 292–297.