APPLICATION OF CLUSTER ANALYSES OF RADIANCE DATA MEASURED
BY SATELLITE AND COMPUTED FROM FORECAST PROFILES

Dongsoo Kim
Cooperative Institute for Research in Environmental Sciences (CIREs)
University of Colorado/NOAA, Forecast Systems Laboratory, Boulder, Colorado U.S.A.

1. INTRODUCTION

The Forecast Systems Laboratory (FSL) of the National Oceanic and Atmospheric Administration (NOAA) runs a real-time 3-hour data assimilation cycle, the Mesoscale Analysis and Prediction System (MAPS). With an Optimum Interpolation (OI) analysis scheme, MAPS successfully integrates asymptotic wind information (Benjamin et al., 1991a), but asymptotic mass information from TIROS Operational Vertical Sounder (TOVS) data is not yet incorporated. The use of radiance information instead of retrieved temperatures in numerical weather prediction (NWP) was advocated by Eyre and Lorenc (1989), and recent impact studies with a global assimilation system have had favorable results (see, for example, Eyre et al., 1992). The approach discussed here involves the computation of radiiances from the forecast profiles, the so-called "forward computation". The computed radiances undoubtedly contain many sources of error including observation errors (in radiance), MAPS model errors (in the predicted profiles) and errors in the computation of radiances. The success of assimilation hinges on accurate specification of these errors, such as biases and covariances.

An OI scheme has been designed to include TOVS radiance information. Though available, global error statistics of computed radiances should not be used in a mesoscale assimilation system, primarily because a mesoscale NWP model with a rapid update cycle will be better than a global model in forecasting airmasses, whose characteristics largely determine the outgoing radiance. Thus, MAPS output and supporting observations were collected in order to obtain more representative bias and covariance error statistics within the MAPS domain - the lower 48 United States and a buffer zone on all four sides.

If all collected data were lumped together, there is concern that the statistical sample might contain members from different populations. OI requires homogeneity of errors (constant bias and covariance). In such a diverse sample, the error distributions might also depart from Gaussian. Because of these concerns the data base was partitioned using cluster analysis to obtain more representative error statistics. This idea is borrowed from the French Laboratoire de Meteorologie Dynamique (LMD) and adopted for a regional data assimilation system in which both model forecast and TOVS data are available in real time (Scott et al., 1991).
KIM, D. APPLICATION OF CLUSTER ANALYSES...

Section 2 discusses cluster analyses of the collected data set. Section 3 discusses the application to the cluster allocation for a TIROS pass over the midwestern United States, and application to the retrieval problem. Results are discussed in Section 4, and suggestions and plans for future work are given in Section 5.

2. CLUSTER ANALYSES

2.1 Data Collection

During March of 1991 and 1992, FSL collected NOAA-10 satellite data near the synoptic times of 0000 and 1200 UTC. A collocated sample comprises 1) measured radiances, 2) radiances computed from a MAPS 3-hour forecast, 3) radiances computed from a rawinsonde sounding (RAOB), 4) MAPS forecast profile on mandatory levels, and 5) RAOB profile on mandatory levels.

The time window for data collection was ± 2 hours from the synoptic times, and the spatial window was 20 km from the RAOB site. The closest HIRS field of view was selected regardless of cloud condition. The first three and the last three elements of each scan line are not saved.

The computation of radiances from the MAPS forecast profiles was reported in the proceedings of an earlier ITSC meeting (Kim, 1991). During the preprocessing of TOVS data, no limb correction is applied, either to HIRS or MSU channel radiances. Also the estimation of cloud parameters (cloud top pressure and cloud fraction) within the HIRS field of view is part of the forward computation; no “cloud clearing” efforts are made. The cloudy radiance is considered to be the measurement. Furthermore, the tuning factors in the transmittance model were set to default values, namely, δ was set to zero, and γ was set to one. These tuning parameters may help correct biases in the radiances, but they also make it difficult to obtain good error covariances.

Also noted are NWP model upgrades (Benjamin et al., 1991b) in the March 1992 data set, such as grid resolution change from 80 km to 60 km, vertical resolution change from 18 levels to 25, improved precipitation (stable) physics, and rougher terrain.

2.2 Quality Control

If any measurement of HIRS or MSU is either missing or out of range (150 K to 350 K in brightness temperature), a set of collocated samples is not saved. If HIRS channel 8 measured radiance (converted to brightness temperature) is different from MAPS radiance by more than 2 K, then this set of collocated samples is omitted from the data base because of gross error in the forward computation. After these restrictions were imposed, the March 1991 data base contained 444
samples, and the March 1992 data base contained 253 samples. The number of samples in March 1992 is smaller than in March 1991 because of an unstable ingest system.

HIRS channels 7 and 8 are not considered in the statistical analysis because they were already used in estimating cloud parameters and subsequently in computing radiances from profiles (Eyre and Menzel, 1989). The cloud parameters are available as a byproduct of the forward computation, but they do not play any role in "cloud clearing". Rather, they are used as diagnostic variables to check the results of cluster analysis.

2.3 Cluster Analysis

Cluster analysis aims to allocate a set of "objects" to a set of mutually exclusive and exhaustive groups such that individuals in different groups are dissimilar (Chatfield and Collins, 1980). The object in this study is defined to be a vector whose 11 elements are the radiance increments (measured minus MAPS-computed) of HIRS channels 4, 5, 6, 10, 11, 12, 13, 14 and 15, and MSU channels 2 and 3. So the number of variables is eleven.

The dissimilarity is represented as the square root of the sum of squares of each pair (called Euclidean distance) after the 11 variables in each object are standardized so that mean is zero and variance is one. The Euclidean distance (ED) function is defined by

\[ ED_{ij} = \left[ \sum_{k=1}^{11} (R_{ik} - R_{jk})^2 \right]^{1/2}, \quad i, j = 1, \ldots, n \]  

where \( R \) is the radiance increment and \( n \) is the number of objects.

Equation (1) can be modified based on prior knowledge of the variables. For example, if variables HIRS 10, 11, and 12 are not standardized, then \( ED \) values are basically controlled by these three variables because their variances are greater than those of other variables.

Whether two objects are merged into a cluster depends upon their \( ED \) value. With \( n \) objects, there are \( n - 1 \) mergers and \( 2^{n-1} \) possible orderings for the objects in the cluster tree. The method of merging each object is "complete-linkage," in which the distance between groups is determined from the most remote pair of objects (see Chatfield and Collins, 1980). There are strategies for subdividing larger clusters and merging smaller clusters that are controlled by threshold values, but the threshold values determine the structure we want to explore. For this reason, the whole tree structure (a dendrogram) is laid out and the number of clusters is empirically restricted to nine as discussed in the following section. We retain clusters with more than 30 objects and regroup smaller clusters into an unrepresentative cluster.
2.4 Hierarchical Trees

The clustering algorithm described in the previous section is applied to plot the dendrogram in Fig. 1 for 1991 (top/left) and 1992 (top/right). The abscissa of the dendrogram is the ED value. The dashed lines are cutting heights to generate nine clusters. Three clusters in the 1991 data set and two clusters in the 1992 data set contain more than 30 objects as seen in the histograms at the bottom of Fig. 1.

2.5 Statistical Characteristics of Clusters

The means and standard deviations of each cluster are shown in Table 1-A for the 1991 data set and in Table 1-B for the 1992 data set. The 402 objects are in the three clusters in Table 1-A, and 216 objects are in the two clusters in Table 1-B, so that the numbers of objects in unrepresentative clusters are 42 and 37 respectively. The standard deviation of each cluster is smaller than that of the total sample, which shows the benefits of clustering in getting internally homogeneous subgroups.

The clusters determined from the radiance differences show the model error (RAOB minus MAPS) structures as seen in Fig. 2 and Fig. 3. Figure 2 shows cluster means of model error in temperature (top), specific humidity (middle), and measured minus MAPS-computed brightness temperature (bottom). Left panels are for the 1991 data set and right panels are for the 1992 data set.

According to the bottom/left panel of Fig. 2, the relative radiance differences for cluster 1 against those of cluster 2 are positive in most channels. The pattern is the same for temperature model error as seen in the top/left panel of Fig. 2; the relative temperature model errors of cluster 1 against cluster 2 are positive throughout all levels except at 200 hPa. Similar argument also holds for specific humidity errors except at 500 hPa. Note that the radiometer will sense warmer temperatures when the layer is drier in HIRS channels 10, 11 and 12. It is difficult to apply the same argument to cluster 3 which has 58 objects.

Figure 3 is the same as Fig. 2 except for standard deviations (s.d.). Cluster 3 shows very large s.d. in the model temperature error above 300 hPa, but shows smaller errors than those in clusters 1 and 2 for both the specific humidity and most channel brightness temperatures. In general, the 1992 data set has smaller s.d. than the 1991 data set both in the model errors and radiance increments except HIRS 15 and MSU 2.

3. APPLICATIONS

3.1 Cluster Allocation
The allocation of a data vector with a particular cluster is different from determining the composition of the original cluster. To see how well the cluster identified, the Maximum Likelihood ($ML$) Distance function (Christensen, 1991) is applied to the March 1991 data set. The $ML$ function is defined by

$$ML_i = \log|\Sigma_i| + (y - \mu_i)^T \Sigma_i^{-1} (y - \mu_i),$$

(2)

where $\Sigma_i$ stands for covariance matrix of the $i^{th}$ cluster, $\mu_i$ is the vector of the $i^{th}$ cluster mean, $|\cdot|$ is the determinant and $y$ is the vector of the measured minus MAPS-computed brightness temperature. A given vector $y$ is tested against each cluster using Eq. (2). If the $\min(ML_i)$ is greater than some threshold (say 50), the vector $y$ is assigned to the unrepresentative cluster. If the $\min(ML_i)$ is less than the threshold, the vector is assigned to the cluster that yielded the minimum value. About 85% were correctly assigned.

The similar test was done with Mahalanobis Distance defined by

$$MH_i = (y - \mu_i)^T \Sigma_i^{-1} (y - \mu_i).$$

(3)

This function scored less satisfactorily than $ML$ distance. The $MH$ distance has been used for air mass classification in LMD's new 3I method (Scott, 1991), but a direct comparison cannot be made due to different cluster statistics. The better allocation by the $ML$ function in FSL's database indicates that the distribution of objects in each cluster is rather homogeneous.

To carry out the clustering procedure in Section 2, the TOVS-measured radiances and MAPS forecasts are needed; both are available in real time at FSL. With the statistical information, means and covariances obtained from a month-long data set, a satellite pass on March 25 00 UTC, 1991 over the U.S. Midwest is subjected to cluster identification at each HIRS field of view.

The Fig. 4 shows the cluster ID distribution marked in every third element of every third scan line (top), the MAPS 3-h 400-hPa temperature forecast (middle), and the MAPS forecast minus MAPS analysis at the same level (bottom). One is struck by the spatial coherence of the clusters. Cluster 1 corresponds to a cold bias in the forecast (see top/left panel of Fig. 2). Over the states of Arkansas and Mississippi cluster 1 corresponds to a thermal trough in the MAPS forecast. Cluster 2 is distributed where model errors have a warm bias. The symbol 0 does not represent any cluster due to the limit $\min(ML) < 50$.

3.2 Temperature Retrieval

Another application of clustering is the use of a retrieval operator (i.e., cross-covariance) to compute increments of retrieved profile from increments of brightness temperatures in the context of
McMillin (1991). We use constrained least squares estimation;

$$dt = b(y) + S_{yx}[\lambda E + S_{xx}]^{-1} + [x - b(x)],$$

(4)

where $dt$ is the increment of retrieved profile, $x$ is an observation increment in brightness temperature, $y$ is the profile increment, $E$ is the covariance of RAOB-computed minus MAPS-computed brightness temperature, $S_{xx}$ is $\text{cov}(x, x)$, $S_{yx}$ is $\text{cov}(y, x)$, $b(\cdot)$ is the sample mean, and $\lambda$ weights $E$ relative to $S_{xx}$. Since the matrix $E$ is the covariance between RAOB-computed radiance and MAPS-computed radiance, it represents errors contained in the forward model. This includes errors from the cloud parameter estimation. The matrices $E$ and $S_{xx}$ serve as constraints to prevent the appearance of unreasonable coefficients in the retrieval operator $S_{yx}$. The scalar value $\lambda$ is set depending on the physical characteristics of clusters.

4. DISCUSSION

The surface problem remains critical in the forward computation. In the forward computation, the skin temperature is set equal to the air temperature. The predicted temperature on the lowest computational surface must be interpolated to the predicted surface pressure, which varies considerably over the mountaineous terrain.

Other than at the surface, we obtained comprehensive relationships between radiance increments for TOVS and MAPS, and profile increments for RAOBs and MAPS. The conventional measure of model errors, i.e., RAOB minus MAPS, was consistent with the radiance increment structure. The cluster analysis enhances this feature and highlights the utilities of TOVS radiance data.

Cluster 3 of the 1991 data set was found to have a large cloud fraction at upper levels (not shown); this certainly affected MAPS-computed radiances. Cluster 3 differs strongly from the other clusters, in the three humidity channels (HIRS 10, 11 and 12) and HIRS 15 (see bottom/left of Fig. 2). These channels have small s.d. (see bottom/left of Fig. 3), implying that an upper level overcast is the main source of bias in the computed radiances.

No cluster in the 1992 data resembles cluster 3 in the 1991 data. Even though NWP model improvement is a contributing factor, small sample sizes may have prevented generation of a cloud-sensitive cluster (the criterion is 30 samples).

5. CONCLUSION

Though this study confirms the utility of TOVS radiance data, their incorporation in a mesoscale data assimilation system remains a challenge. The approach discussed here was to refine the
KIM, D. APPLICATION OF CLUSTER ANALYSES...

data set via cluster analysis to obtain more representative statistics. As a first step, biases and
covariances from one month of data were obtained. Then statistics were used to classify satellite
radiance data vector into clusters. The spatial distribution of model errors are in general agreement
with the characteristics of each cluster as shown in the first example in Section 3. The two examples
in Section 3 should be merged into one suite within MAPS because the retrieval increment will be
the satellite information in the MAPS OI scheme.

Statistics from the cluster analysis are more representative than those developed from the total
sample. The cluster analysis should be allowed to evolve in time because forecast errors depend
on the weather regime as well as the sophistication of the NWP model itself, as shown in the 1992
March data analysis. Hence, it seems desirable to update statistics weekly or monthly (Dibben

6. ACKNOWLEDGMENTS

All statistical analyses are performed with S-PLUS by Statistical Sciences, Inc. The author thanks
Tom Schlatter and Nita Fullerton of FSL and Ed Westwater of WPL for their careful review.

7. REFERENCES

An isentropic three-hourly data assimilation system using ACARS aircraft observations. Mon.
Wea. Rev. 119, 888-906.

in the MAPS isentropic-sigma data assimilation system. Preprints, Ninth AMS Conference on

246 pp.

Christensen, R., 1991: Linear models for multivariate, time series, and spatial data. Springer-
Verlag, 318 pp.

Dibben, P. and D.E. Chapman, 1991: The monitoring of a local area sounding system. Sixth

Eyre, J., and A. Lorenc, 1989: Direct use of satellite sounding radiances in numerical weather

Eyre, J., and P. Menzel, 1989: Retrieval of cloud parameters from satellite sounder data: a simul-

information through one-dimensional variational analysis. Tech. Memo. No.187, ECMWF,
pp.31.

215


Table 1-A

Means and standard deviations of radiance increment (K) with MAPS-computed March 1991 data set for each cluster (NOAA-10). Number in parenthesis is the sample size.

<table>
<thead>
<tr>
<th>Variable</th>
<th>All(444)</th>
<th>C1(199)</th>
<th>C2(145)</th>
<th>C3(58)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>Mean</td>
<td>S.D.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>S.D.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>S.D.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>S.D.</td>
</tr>
<tr>
<td>H4</td>
<td>-0.61</td>
<td>0.80</td>
<td>-0.46</td>
<td>0.74</td>
</tr>
<tr>
<td>H5</td>
<td>-0.01</td>
<td>0.68</td>
<td>0.11</td>
<td>0.52</td>
</tr>
<tr>
<td>H6</td>
<td>0.30</td>
<td>0.63</td>
<td>0.39</td>
<td>0.46</td>
</tr>
<tr>
<td>H10</td>
<td>1.47</td>
<td>1.89</td>
<td>1.15</td>
<td>1.35</td>
</tr>
<tr>
<td>H11</td>
<td>-1.31</td>
<td>2.22</td>
<td>-1.37</td>
<td>1.93</td>
</tr>
<tr>
<td>H12</td>
<td>-4.08</td>
<td>3.70</td>
<td>-4.40</td>
<td>3.60</td>
</tr>
<tr>
<td>H13</td>
<td>0.25</td>
<td>1.08</td>
<td>0.31</td>
<td>0.91</td>
</tr>
<tr>
<td>H14</td>
<td>-0.34</td>
<td>0.87</td>
<td>-0.28</td>
<td>0.72</td>
</tr>
<tr>
<td>H15</td>
<td>-2.32</td>
<td>1.27</td>
<td>-2.37</td>
<td>0.83</td>
</tr>
<tr>
<td>M2</td>
<td>-0.83</td>
<td>0.90</td>
<td>-0.54</td>
<td>0.78</td>
</tr>
<tr>
<td>M3</td>
<td>-1.32</td>
<td>0.96</td>
<td>-1.13</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 1-B

The same as Table 1-A except March 1992 data set.

<table>
<thead>
<tr>
<th>Variable</th>
<th>All(253)</th>
<th>C1(125)</th>
<th>C2(91)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>-0.68</td>
<td>0.78</td>
<td>-1.15</td>
</tr>
<tr>
<td>H5</td>
<td>-0.11</td>
<td>0.65</td>
<td>-0.43</td>
</tr>
<tr>
<td>H6</td>
<td>0.23</td>
<td>0.57</td>
<td>-0.01</td>
</tr>
<tr>
<td>H10</td>
<td>0.68</td>
<td>2.05</td>
<td>-0.28</td>
</tr>
<tr>
<td>H11</td>
<td>-1.64</td>
<td>2.59</td>
<td>-2.99</td>
</tr>
<tr>
<td>H12</td>
<td>-4.11</td>
<td>3.71</td>
<td>-6.01</td>
</tr>
<tr>
<td>H13</td>
<td>0.02</td>
<td>1.08</td>
<td>0.22</td>
</tr>
<tr>
<td>H14</td>
<td>-0.69</td>
<td>0.80</td>
<td>-0.80</td>
</tr>
<tr>
<td>H15</td>
<td>-2.73</td>
<td>1.38</td>
<td>-3.45</td>
</tr>
<tr>
<td>M2</td>
<td>-0.75</td>
<td>1.19</td>
<td>-1.28</td>
</tr>
<tr>
<td>M3</td>
<td>-1.05</td>
<td>0.77</td>
<td>-1.32</td>
</tr>
</tbody>
</table>
Fig. 1 Dendrograms (top) and histograms of cluster size (bottom) for 1991 (left) and 1992 (right) March data set. Dashed lines make 9 clusters. Cluster IDs are renamed according to their size, so that cluster 5 of 1991 data set is renamed as cluster 3 in the text.
Fig. 2 Cluster means of model error (RAOB-MAPS) in temperature (top), in humidity (middle), and measured minus MAPS-computed in brightness temperature (bottom) increments for 1991 (left) and 1992 (right) March data set.
Fig. 3 The same as Fig. 2 except those of standard deviation.
Fig. 4 Cluster IDs (top), MAPS forecast of temperature at 400 hPa (middle), and MAPS forecast error (bottom) for 1991 March 25 0035 UTC.