# APPLICATION OF A CLUSTERING METHOD FOR CLOUD COVER ANALYSIS OVER TROPICAL REGIONS

Michel Desbois, Geneviève Sèze
Laboratoire de Météorologie Dynamique du CNRS
Ecole Polytechnique, 91128 Palaiseau Cedex, France

Abstract : The latest developments of a cloud classification technique applied on METEOSAT images are presented. This technique is applied to an African sector, from 11 to 14 June, 1979. The cloud cover results are presented at the scale of the ECMWF model grid.

## 1. INTRODUCTION

Several cloud classification techniques from satellite pictures have been proposed in the last few years as an accurate cloud cover measurement is needed for radiation budget studies, initialization and diagnostic studies of General Circulation Models, and for climatological purposes : Reynolds and Vonder Haar (1977), ESA (1980), Coakley and Bretherton (1982), Desbois et al. (1982), Minnis and Harrison (1982).

Some of these methods, like the one developed by the authors, are based on histogram analysis and partitioning by objective techniques. It was used in several studies, specially during the ISCCP algorithm intercomparison. The basic steps of this method are described below. The influence of space sampling for obtaining a representative classification at large scales is stressed. Some remarks are made concerning the usefulness of time sampling, both for ground reference properties and representativeness of the cloud types. The importance of the shape of the histograms is also outlined.
We examine then the capability for the method to separate high semi-transparent clouds from other ones (specially middle level clouds and

thick high clouds, generally convective). A way to take into account some spatial properties of the image is proposed and tested. Finally, we present some classifications obtained over Africa for the period 11, 12, 13 and 14 June 1979.

## 2. THE BASIC STEPS OF THE LMD CLASSIFICATION METHOD

The principles of the application of the so-called "dynamic cluster method" for the classification of meteorological satellite images have been described in Desbois et al. (1982). Since then, the method has been intensively used and tested in numerous experiments (for example ISCCP), on several satellite instruments, and with several adaptations. All these applications follow the same basical steps :

a) A "learning set" is chosen. The learning set is composed of the radiometric counts corresponding to a representative number of pixels (in several channels). The statistical requirements of the method make it necessary to have several thousand pixels in the learning set, but this number cannot be increased very much for computer time saving considerations. Note that the learning set can be different from the images on which the classification will be applied.

b) One, two, three or more dimensional histograms are constructed from this learning set. This operation requires a good alignment of the different image channels, as the position of each pixel in the histogram will determine its class.

c) The dynamic cluster method is applied to the histograms either from randomly chosen kernels, or from pre-determined kernels. In all the applications described here, we will use randomly chosen kernels to avoid the influence of a non-objective pre-determination. The results of this fundamental step of the method are :

. the number of classes which have been separated,

. the center of gravity and the variances corresponding to each class of the learning set,

. the percentage of points in each class of the learning set.

d) Each point of the studied image (or images) is attributed to one of the classes, according to a definition of the distance to a class which is the sum of the Euclidian distance to the center of gravity of the class and of the variance of the kernel of the class : more points are then attracted to homogeneous classes than to dispersed classes. The result of this operation then gives the total number of points of the studied image in each class, and afterwards the percentage of coverage by this class. Classified images can also be constructed.

e) One further step, which is not mandatory in the classification procedure itself, is the identification of the clouds, or a temperature and height attribution. This step can include a subjective identification of the cloud class by comparison of the classified images and the original images in the different channels and an objective determination of the cloud top temperature and the optical thickness, taking into account atmospheric corrections, ground characteristics, etc...

As in other objective classification techniques, the basic method has some drawbacks which were shown by the intercomparisons done for ISCCP :

. When the histogram does not present well defined peaks, but a regular decrease from a single maximum, the method still separates classes. The boundaries of these classes are not very significative, as they can vary from one initial random choice of the kernels to another.

. In many cases, specially the ones described above, the choice of the initial parameters of the method (number of initial classes, number of

elements in a kernel) can affect the classification (greater the kernel size, smaller the number of final classes).

These drawbacks must always be recalled when applying this method. A proper choice of its initial parameters has to be made. However, when real classes do exist, they are correctly separated. In the applications we did, the problem was that in general we found too many classes rather than too few.

## 3. SPACE SAMPLING OVER LARGE AREAS

One of the problems of cloud classifying by automated techniques in individual squares is that the classes found are not always equivalent for adjacent squares or the same squares taken at different times, depending on the cloud content of the square. This is useful when studying details of the cloud distribution on the mesoscale, but does not make it easy to investigate large-scale space distributions or time variations of the great cloud types.

Moreover, in the case of the present statistical method, better results are obtained with learning sets sufficiently large, containing several thousand pixels.

A solution to this problem is to take sufficiently large learning sets from squares at least of the size of the minimum required to get significative statistical results and to have a representative cloud population in the area. Far larger squares, from which a sample of the required size is extracted, can be taken in order to get continuous results in a large region. The method used here allows us to go back from the classes found at these large scales (step c) to classifications of the points of image segments of any size by applying step d) of the method to these segments.

The effects of this kind of approach have been observed in three independent experiments :

- studies to find the ISCCP algorithm intercomparisons (GOES satellites),
- METEOSAT studies over tropical regions on the meshes of a general circulation model,
- METEOSAT studies over Europe.

We present here some conclusions from these studies, illustrated by the

African cases. One experiment was conducted on METEOSAT data over tropical regions to study the effect of the spatial content of the learning set. The grid used corresponds in this case to the grid of a General Circulation Model used in Laboratoire de Météorologie Dynamique.

This region is more homogeneous than the temperate regions studied elsewhere, presenting mainly latitudinal variations of ground and atmosphere characteristics. Comparisons between the local mesh classification (about 6000 pixels per mesh) and the total area classification (50 meshes - about 350 000 pixels) have been made for 16 meshes, on 12/06/79. Note that, in this case, the 3 channels (VIS, IR, WV) of METEOSAT were used.

The standard deviation between the results obtained by local classification and classification on 8 meshes is 10 % of cloudiness. It increases slowly up to 13.5 % when the classification is done on the 50 meshes. These results are better than in the case of the temperate regions : this is due partly to a better homogeneity of the region studied, partly to a better processing : use of a third channel, which allows a better discrimination for high clouds (water vapour channel) ; grouping of the classes in 4 levels taking into account ground and atmospheric variations. However, the same problems of stability of the partitions and significance of the boundaries between classes remain.

## 4. TIME SAMPLING

There is another way to get representative samples for a given region : it is to take the learning set for the classification from a time series of pictures for several successive days, at the same hour. This method was tested for 2 experiments over Europe (Desbois & Seze, 1984). The cumulative histograms obtained (Fig. 1) present distributions representative of all kinds of clouds

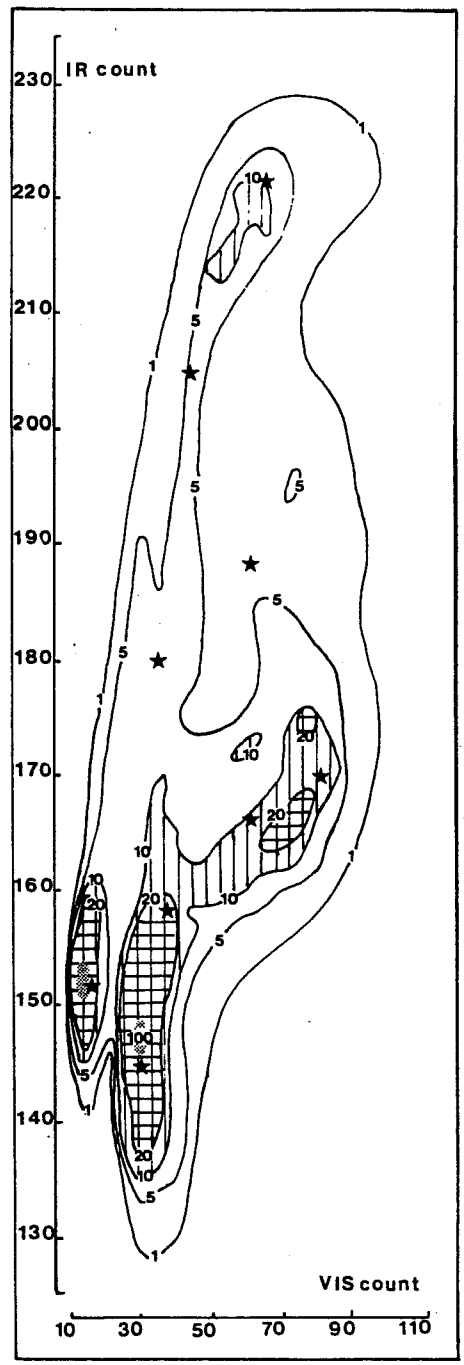


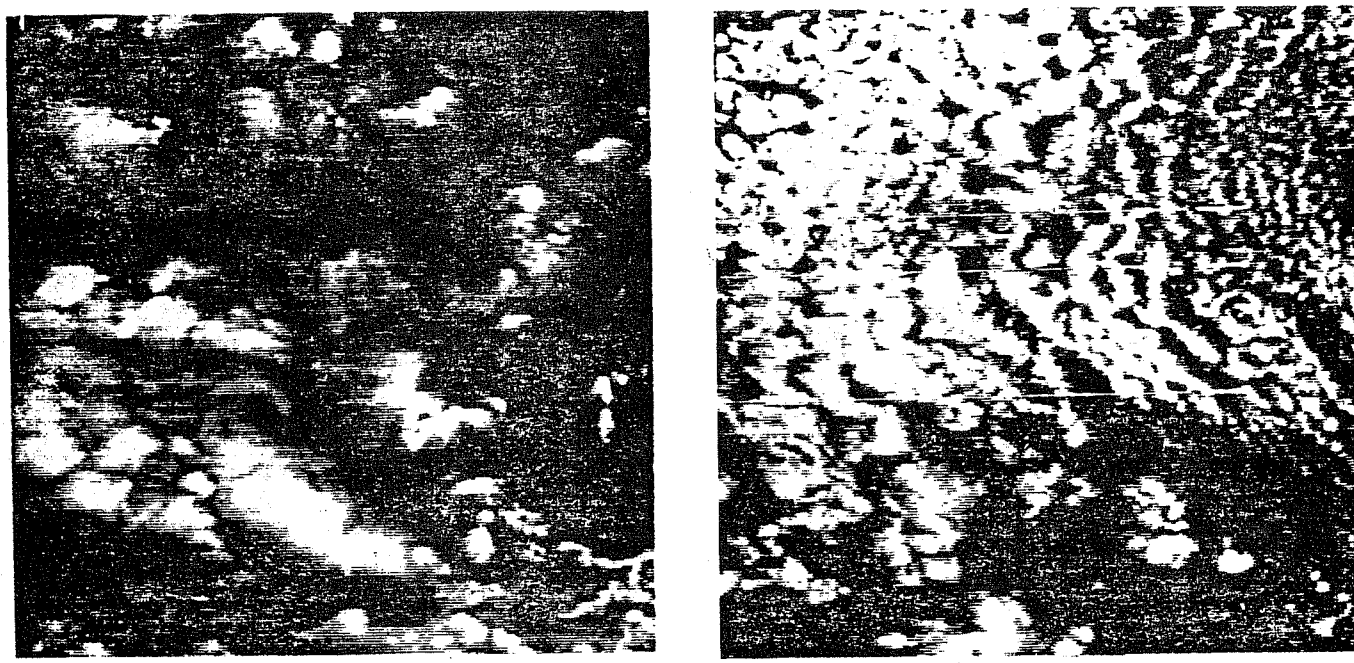

Figure 1 : Time cumulative VIS-IR histogram obtained over an European region for 6 consecutive days of February 1982. The stars represent the centers of gravity of the classes found.

(and surface) present on the series of images. One can deduce from these diagrams preferential levels of cloud tops as well as surface statistical properties. But they also stress the very large number of pixels which do not belong to a well defined cloud type, but to intermediate regions between these cloud or surface types : partial coverage of the pixels, semi-transparency of the clouds and multiple layers are responsible of this spreading of the cloud signatures.

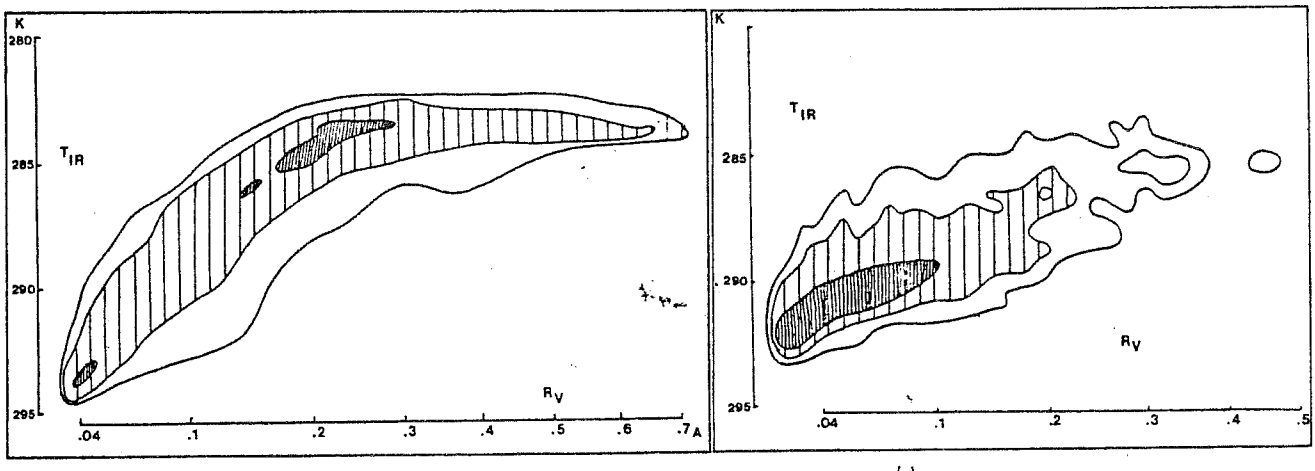
Once again, it appears difficult to get significative separations of the classes using a cluster separation on 2-D histograms only. This is well illustrated by the study of two particular kind of clouds : stratocumulus and cirrus.

## 5. STRATOCUMULUS AND CIRRUS CLOUDS

Stratocumulus clouds, specially oceanic ones, form well defined layers easily recognizable on satellite pictures (Fig. 2). However, they present large variations of optical thicknesses and produce many partially covered pixels, which generally result in 2-D histograms with no definite peaks, but with a very characteristic shape. This shape can be explained by radiative transfer models, as the one used by Arking (Fig. 3). In some cases, it is possible to interpret the 2-D histograms from this kind of model (Fig. 2). Areas with mostly partial coverage can be separated from areas with mostly variable optical thickness. Nevertheless, that does not solve the problem of getting "pure" cloud and sea classes which would allow an interpretation of intermediate points according to Arking's proposal (1984).
Another approach is the one of Coakley and Bretherton (1982), also called the spatial coherence method, which takes into account the spatial environment of each pixel by computing local IR varainces which are plotted on diagrams versus the IR values measured at the same point. The arches obtained, which are particularly clear for stratocumulus over sea, are interpreted



a) b)




a) b)

Figure 2 : Examples of VIS-IR histograms obtained over 2 kinds of oceanic stratocumulus.

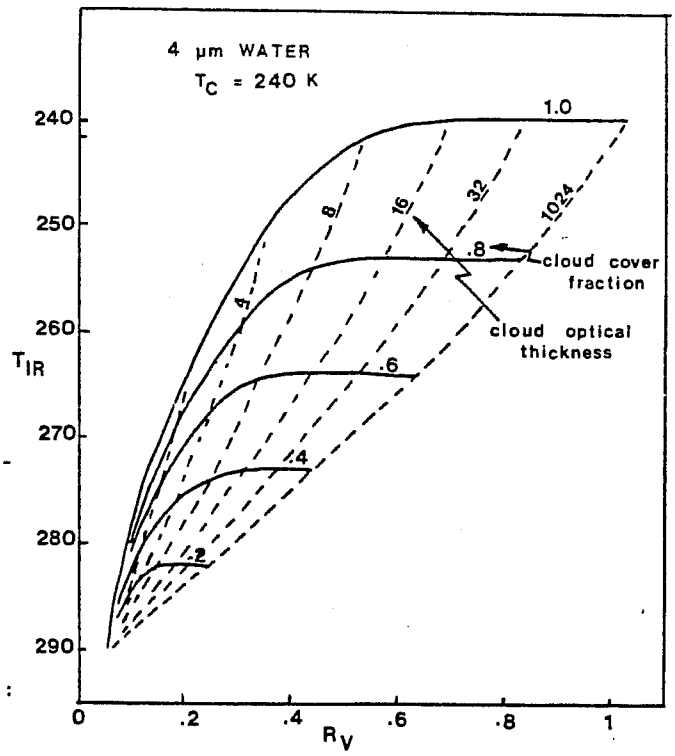



Figure 3 : Model histogram obtained by Arking and Childs for varying optical thicknesses and partial pixel cloud coverages, from radiative transfer calculations.

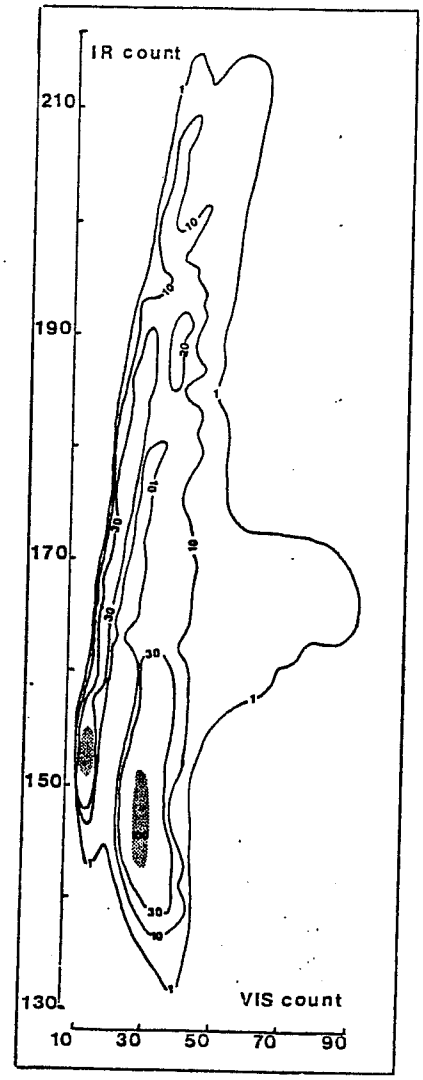



Figure 4 : typical shape of a VIS-IR histogram in the presence of cirrus.

in terms of partial coverage of the pixels. This method gives of course a better definition of homogeneous classes than the other ones, but the variations of the optical thickness of the cloud cannot be described. A combination of both kind of information seems necessary.

A similar presentation of the case of cirrus cloud could be done, but their variations of both optical thickness and emissivity, together with the problem of a non-uniform background below them, makes the problem still more intricate. Characteristics of the VIS IR histograms for cirrus is given on Fig. 4, going from very thin to thick clouds. The dispersion along the IR axis is very large, resulting also in large IR local variances, but smaller VIS variances. Once again a combination of spectral and spatial information seems appropriate for a good separation of cirrus.

Fortunately, the use of the Water Vapor channel in the case of METEOSAT can help in the separation of these clouds. It has been shown, specially by Szejwach (1982), that semi-transparent clouds have a particular position on the IR-WVP histograms, relative to the curve which represents the blackbody temperature of opaque clouds. This is illustrated by Fig. 5 which shows the semi-transparent area on an IR-WV histogram, and the corresponding points in the corresponding VIS-IR histogram. These areas remain well defined on both histograms, showing that a clustering technique applied to a 3-D histogram could find a characteristic position for these semi-transparent clouds.

## 6. TECHNIQUE USED FOR CLASSIFICATION AT ECMWF GRID SCALE

Accounting for the observations done in the previous sections of this paper, a good classification scheme using a clustering technique has to take into account :

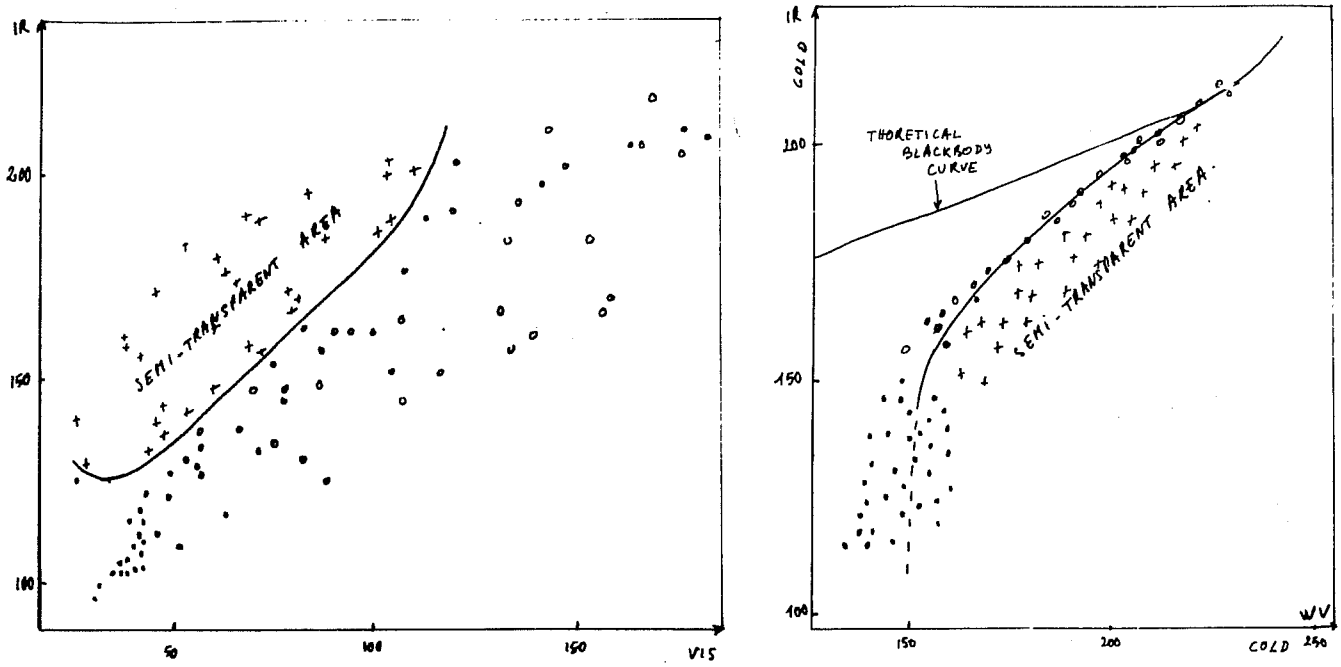



Figure 5 : Separation of semi-transparent clouds on VIS-IR and IR-WVP histograms.

. sampling on large areas or/and time sampling,

. local variances in IR and Visible,

. water vapor channel for high clouds (in the case of METEOSAT).

The period chosen to do comparisons with clouds produced in the ECMWF model is the FGGE period 11-15 June 1979. For this period, we used METEOSAT pictures from 11 June : 11:30, 17:30, 23:30 ; 12 June : 5:30, 11:30, 17:30, 23:30 ; 13 June : 5:30, 11:30 ; 14 June : 11:30 and 15 June : 11:30. The area processed is shown on Fig. 6, with the grid of ECMWF superimposed. This area was separated into four sub-areas on which the classification was applied. Due to the shortness of the period and to the large extent of the area of study, the choice has been made to have large areas for constituting the learning set rather than using time sampling. The sub-areas used have dimensions of the order of 4000 km x 1500 km, which is large enough to do representative clustering, but sufficiently restricted in latitude to avoid large variations of the surface properties. A water vapor channel was used in all the classifications, daytime and nightime. The use of visible and visible variance was restricted to the picture of 11:30, because the sun was too low in pictures of 5:30 and 17:30. In the 11:30 picture, VIS channel was corrected for the variation of the solar zenith angle. Five parameters were used for the classification at this hour, VIS, IR, WV, VARVIS, VARIR, instead of 3 at the other ones, IR, WV, VARIR.

## 7. RESULTS

Before giving the results of this classification, one has to verify their consistency : Are the same clusters found for the different areas and different times ? What is the effect of taking 3 parameters instead of 5 ? To what kind of clouds do the classes correspond ? To answer quickly the first question, we have always found between 6 and 11 classes in the classifica-

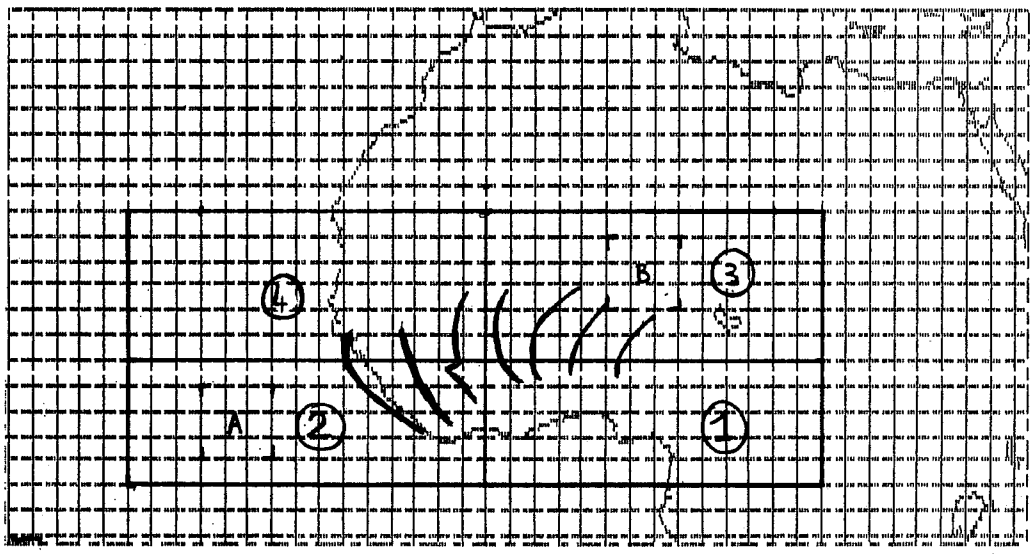



Figure 6 : Area analysed in the present study. The area is limited by the parallels 1°N and 33°N and the meridians 26°E and 33°W. Sub-areas used for the classification are numbered 1,2,3,4. The grid represents the ECMWF grid (1,875°). Successive positions of the squall line occuring during the period from 11 June at 17:30 to 13 June at 5:30 are also plotted.

tions with a larger probability for 7 and 8 classes. Centers of gravity found show generally the same characteristic main classes, but some classes can be split or some particular classes can appear, depending on the content of the square of analysis. For example, several classes of surface can generally be found in sub-area 3, which is mostly desert and has little cloudiness. In sub-area 1, the presence of haze is sometimes detected, whereas in sub-area 4, dust clouds can induce a particular clustering.

Concerning the consistency between classifications with 5 or 3 parameters, which is very important for the consistency of daytime and nightime classifications, we give in Table 1 the results obtained at the same hour (11:30 on 12 June) using 5 and 3 parameters (area 2). 8 clusters were found in each case.

| Parameters / Cloud id. | 5 parameters | | | | | | 3 parameters | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | VIS | IR | VARVIS | VARIR | WV | % | IR | VARIR | WV | % |
| High clouds | 181 | 215 | 62 | 47 | 210 | 4 | 213 | 40 | 209 | 4 |
| Cirrus | 84 | 196 | 58 | 78 | 192 | 4 | 184 | 91 | 184 | 5 |
| | 72 | 163 | 74 | 109 | 167 | 5 | 143 | 61 | 157 | 4 |
| | 23 | 129 | 39 | 88 | 153 | 12 | 139 | 98 | 157 | 7 |
| Low | 99 | 123 | 78 | 60 | 140 | 9 | 117 | 64 | 142 | 13 |
| | 30 | 113 | 66 | 55 | 141 | 15 | 112 | 38 | 142 | 18 |
| Surface | 13 | 106 | 24 | 33 | 139 | 38 | 106 | 15 | 140 | 30 |
| | 56 | 105 | 41 | 40 | 112 | 12 | 104 | 25 | 110 | 14 |

Table 1 Centers of gravity of the classes found for the 12 of June at 11:30, using 5 parameters and 3 parameters classifications. The values are in METEOSAT counts (8 bits) going from dark to bright and from warm to cold for increasing counts.

These results show clearly a good consistency between the two classifications, specially for high clouds. It is of course more difficult for the

algorithm to separate low clouds from the surface when there is no visible, but the orders of magnitude of the percentages remain the sames. What is far easier with the 5 channels classification is the identification of the classes : VIS and VARVIS parameters are very helpful for this.

Therefore, in the subsequent processing, we have used the daytime classification for the identification of the classes, attaching the nightime classes to the closest values of IR, VARIR and WV. The continuity of the results obtained by this method is apparently good. That can be identified in Fig. 7 which represents the evolution of high clouds, cirrus and medium clouds for the whole period on the whole area.

An analysis of the classification associated with a comparison of classified images with the original ones allows a better identification of the classes found during the whole period. 13 classes were identified at first :

1. High, thick clouds.
2. Cirrus (thin to thick) over surface.
3. Cirrus over medium clouds.
4. Cirrus over low clouds.
5. Medium level clouds.
6. Low clouds.
7. Not well separated low and medium.
8. Very thin clouds (medium or high) - edges.
9. Sea.
10. Land.
11. Land or sea with haze.
12. Dust clouds.
13. Sea with partial coverage of the pixels.

Then, each class of any particular classification was attached to one of these typical numbers. An example of the time evolution of the cloud coverage so obtained is given on Table 2 (for region 2).

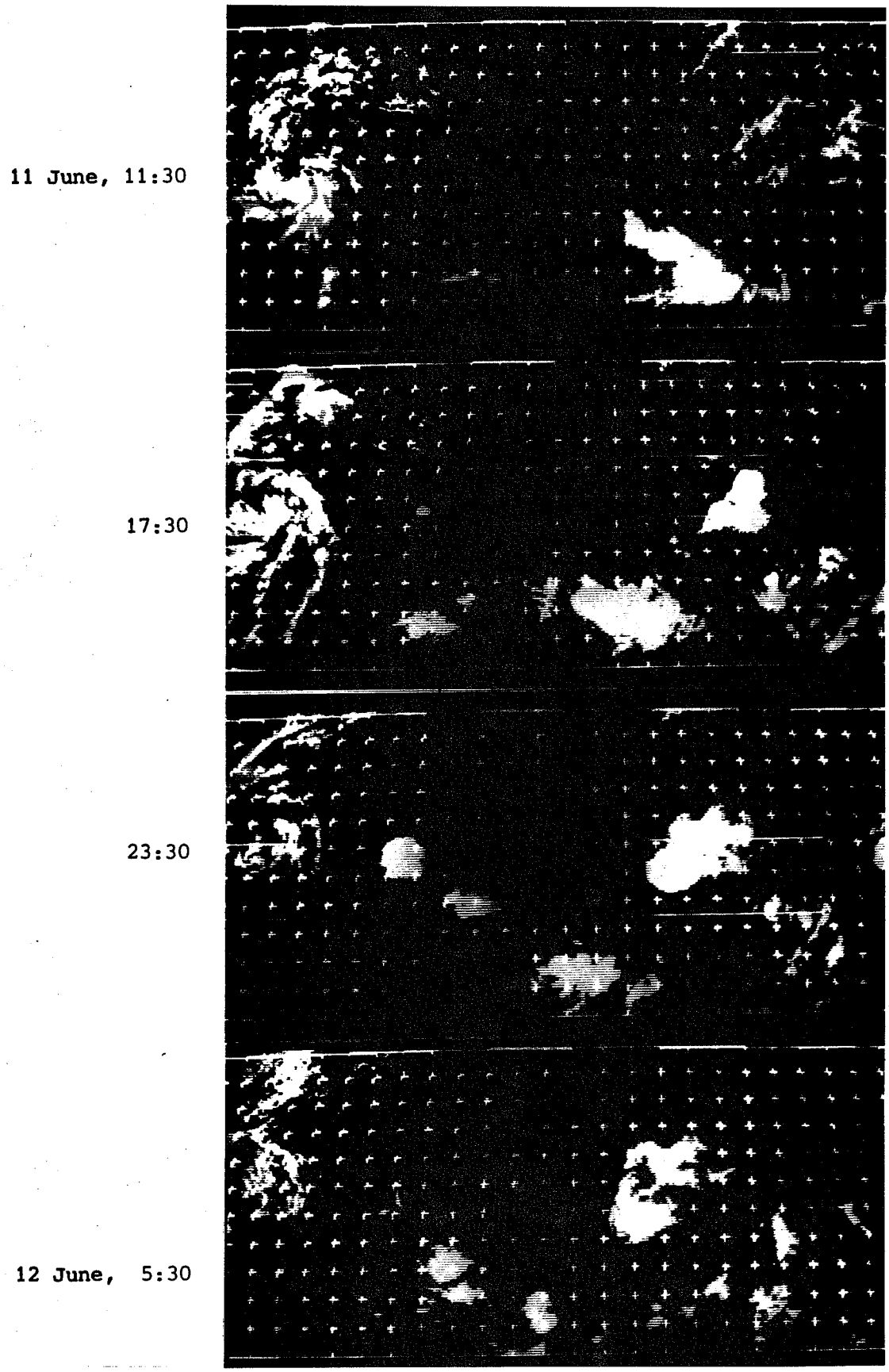



**Figure 7 : classified images from 11 June, 11:30 to 13 June, 5:30. Only deep high clouds (in white) and cirrus and medium clouds (in grey) have been plotted.**

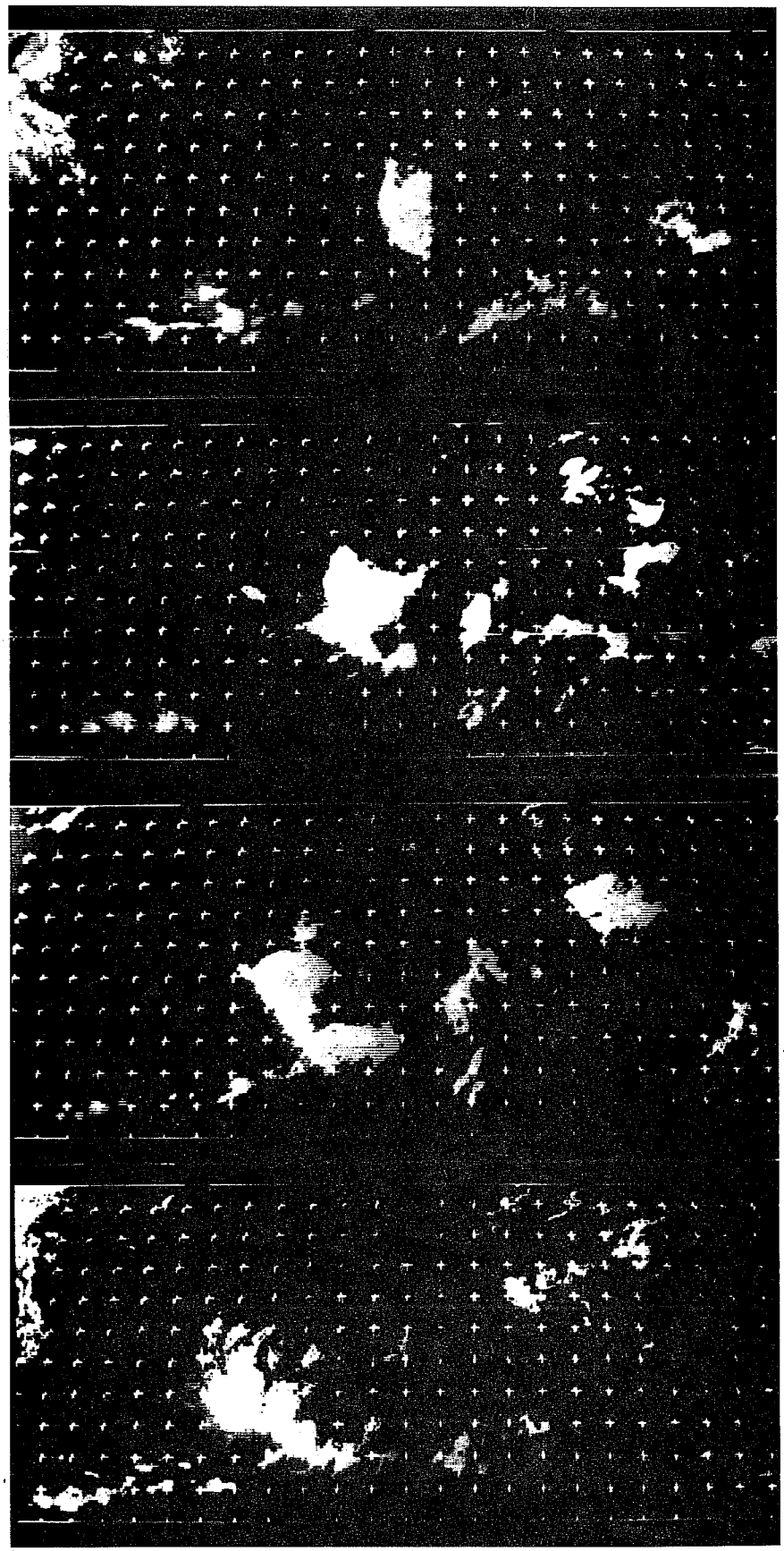



Figure 7: Continued

| Time ↓ | Cloud id. → | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11-6 | 11:30 | 1 | 3 | 18 | 3 | | 8 | 17 | | 22 | 25 | | | | |
| | 17:30 | 2 | 0 | 17 | 6 | | 11 | 19 | | 17 | 27 | | | | |
| | 23:30 | 3 | 2 | 11 | 8 | | | 23 | | 14 | 39 | | | | |
| 12-6 | 5:30 | 4 | 5 | 15 | 9 | | | 22 | | 8 | 32 | | | | |
| | 11:30 | 5 | 4 | 16 | 4 | | | 24 | | 12 | 30 | | 12 | | |
| | 17:30 | 6 | 11 | 11 | 0 | | | 22 | | 8 | 44 | | | | |
| | 23:30 | 7 | 8 | 10 | 5 | | | 22 | | 2 | 40 | | | | |
| 13-6 | 5:30 | 8 | 1 | 18 | 7 | | | 19 | | | 42 | | | | |
| | 11:30 | 9 | 9 | 25 | 0 | | 11 | 14 | | | 37 | | | | |
| 14-6 | 11:30 | 10 | 4 | 17 | 0 | 4 | | 21 | | | 34 | | | | 17 |
| 15-6 | 11:30 | 11 | 3 | 6 | 4 | | 5 | 34 | | 9 | 35 | | | | |

Table 2

Other examples of the time evolution of the cloud cover by the different classes are shown on Tables 3 and 4, for smaller areas representing 3x 3

| Time \ Cloud id. | 1 | 2 | 3 | 5 | 6 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 38 | 5 | 6 | 2 | 34 | 14 |
| 2 | | 30 | 9 | 6 | 16 | 19 | 20 |
| 3 | | 15 | 4 | 0 | 23 | 18 | 40 |
| 4 | | 11 | 8 | 0 | 22 | 18 | 42 |
| 5 | | 4 | 0 | 0 | 29 | 0 | 67 |
| 6 | 3 | 7 | 0 | 0 | 27 | 0 | 64 |
| 7 | 1 | 10 | 15 | 0 | 21 | 0 | 54 |
| 8 | 1 | 13 | 2 | 0 | 22 | 0 | 61 |
| 9 | 4 | 13 | 0 | 4 | 14 | 0 | 66 |
| 10 | 5 | 16 | 7 | 0 | 32 | 0 | 41 |
| 11 | 22 | 16 | 19 | 6 | 9 | 13 | 17 |

Table 3

meshes of the ECMWF grid. The first small area (A) is taken in Region 2, over sea and the second is taken in Region 3, over land (Fig. 6).

| | 1 | 2 | 3 | 5 | 6 | 8 | 10 |
|---|---|---|---|---|---|---|---|
| 1 | | 3 | 0 | 10 | 0 | 8 | 79 |
| 2 | | 21 | 0 | 25 | 0 | 0 | 53 |
| 3 | 1 | 16 | 0 | 44 | 0 | 19 | 20 |
| 4 | | 13 | 38 | 13 | 0 | 14 | 23 |
| 5 | | | 28 | 6 | 0 | 24 | 42 |
| 6 | 22 | 22 | 0 | 19 | 0 | 18 | 19 |
| 7 | 24 | 18 | 18 | 27 | 8 | 0 | 5 |
| 8 | 15 | 44 | 29 | 5 | 4 | 4 | 0 |
| 9 | | 43 | 10 | 2 | 0 | 13 | 31 |
| 10 | | 1 | 0 | 0 | 0 | 4 | 94 |
| 11 | 3 | 0 | 0 | 0 | 0 | 0 | 97 |

Table 4

## CONCLUSION

The cloud classifications obtained here do not pretend to give in a direct way the proportions of low, medium, high and cirrus clouds, with their usual meteorological definition and their top temperature and reflectivity. More physical work has to be done to interpret these results in that way.

However the method used in this paper produced coherent results between nightime and daytime images, specially for high and medium level clouds. Some problems remain with low-level clouds at night, but their proportion stays within reasonable limits. This satisfactory result is mainly due to the use of infrared variance together with infrared and water vapor radiances at night. The continuity of the classes obtained is sometimes broken

at the limits between the sub-areas where the classifications have been done ; that is due to the differences of surface properties between the sub-area but also to the very different relative proportions of cloud types inside the different sub-areas.

A non-negligible percentage of points are in intermediate classes which cannot be identified clearly as attached to a particular cloud type (class 8). These points belong to parts of the pluridimensional histogram which cannot be easily separated, associated with edges of different kind of clouds.

The summits of convective clouds may appear too extended in the results presented here ; but it would be easy to separate an additional class of very cold summits with small variances.

Although the classification obtained here can still be improved, it seems already sufficiently accurate, at least for high and medium level clouds, to do a preliminary comparison with the output of a General Circulation Model.

## REFERENCES


Arking, A. and J.D. Childs, 1984 : Extraction of cloud cover parameters from multispectral satellite measurements. J.Clim.Appl.Meteor. 23,

Coakley, J.A., and G. F.P. Bretherton, 1982 : Cloud cover from high resolution scanner data : detecting and allowing for partially filled fields of view. J.Geophys.Res. 87.

Desbois, M., G. Sèze and G. Szejwach, 1982 : Automatic classification of cloud on Meteosat imagery : applications to high level clouds.

Desbois, M., and G. Sèze, 1984 : Use of space and time sampling to produce representative cloud classification. Annales Geophysicae, 2-5, 599-605.

ESA, 1980 : Meteosat System guide. Vol. 5, Meteosat Management, ESOC Dept. Darmstadt. RFG.

Reynolds, D.W., and T.H. Vonder Haar, 1977 : A bispectral method for cloud parameter determination. Mon.Wea.Rev. 105, 446.

Szejwach, G., 1982 : Determination of semi-transparent cirrus cloud temperature from infrared radiances. Application to Meteosat. J.Appl.Meteor. 21, 384.