KONINKLIJK NEDERLANDS METEOROLOGISCH INSTITUUT

WETENSCHAPPELIJK RAPPORT
SCIENTIFIC REPORT
W.R. 78-7

J. van Maanen

Some experiments with the ECMWF's analysis program.

De Bilt, 1978

Publikatienummer K. N. M. I. W. R. 78-7 (M. O.)

U. D. C.: 551. 509. 38

Table of contents

SOME EXPERIMENTS WITH THE ECMWF's ANALYSIS PROGRAM

by Jan van Maanen

1. Introduction

This report gives a description of work done during a visit to the European Centre for Medium Range Weather Forecasting from 1977 October 17th to November 11th. The main purpose of the visit was to study the analysis scheme used at the European Centre. With that aim in mind, I worked on two small projects involving the analysis program. The first experiment was a satellite impact study. Several analyses of the same date were made to study the influence of satellite data on the analysis. Another experiment studied a certain aspect of the analysis scheme, the division of the atmosphere into boxes. The main part of this report describes the second experiment. The first section describes the satellite experiment.

2. The impact of satellite data

A few experiments were taken to study the impact of satellite information on the analysis. This was done by omitting part of the data that are normally available to the analysis program. It would have been interesting to run forecasts on the different analyses; this was however impossible, because the nesessary programs were not available. Three analyses are shown in figures 1, 2 and 3. In the first one all available data are used, in the second one all satellite information is omitted. The differences are small. The largest differences can be found over the Pacific Ocean, for instance near 40^{o}N, 180^{o}E. Over the southern hemisphere (not shown) the differences are much larger, of course, and are in some places over 20 gpdam.

It is also interesting to compare figures 1 and 3. In the analysis of fig. 3 all information except radiosondes is used. Especially over Europe and Asia many small differences occur,

for instance near 50°N, 40°E, were a trough changed into a cut-off low. A trough near the coast off California became less pronounced and a trough near the eastern United States disappeared over land.

From maps of the analysis error (not shown) it can be seen that when radiosonde reports are omitted the standard deviation of the analysis error has a sharp discontinuity near the coasts of Europe and Asia. In this case the error over land is much larger than over the ocean.

3. Description of the main features of the analysis program

In this section a description is given of the most important features of the analysis scheme only. More details can be found in [1].

It is assumed that there is always a guess field available. This guess field can be climatology, persistence or a forecast of a numerical model. Furthermore, it is assumed that the deviation of the analyzed value from the guess field at a certain grid point can be determined as a sum of weighted deviations of the observed values from the guess values. If A denotes any meteorological variable, A_m^o the observed value at station m, A_m^p the guess field (predicted) value at the location of station m, A_k^p the guess field value at grid point k and A_k^i the analyzed (interpolated) value at grid point k, then the basic analysis equation is

$$A_k^i = A_k^p + \sum_{m=1}^{N} C_{km} \left(A_m^o - A_m^p \right) \qquad (1)$$

C_{km} is the weight given to information of station m when analyzing point k. If we write all quantities as deviations from the true value (superscript t), one obtains

$$a_k^i = a_k^p + \sum_{m=1}^{N} C_{km} \left(a_m^o - a_m^p \right)$$

in which the following definitions are used:

$$a_k^i = A_k^i - A_k^t \;,\; a_k^p = A_k^p - A_k^t \;,\; a_m^o = A_m^o - A_m^t \;,\; a_m^p = A_m^p - A_m^t$$

The analysis problem now is to determine the weights in an optimal way, i.e. to minimize the variance of a_k^i when a large number of analyses is considered. For the variance of a_k^i the following equation can be derived:

$$E_k = \langle a_k^{i\,2} \rangle = \langle a_k^{f\,2} \rangle + 2 \sum_m C_{km} \left[\langle a_k^f a_m^o \rangle - \langle a_k^f a_m^f \rangle \right] + \sum_m \sum_n \left[\langle a_m^o a_n^o \rangle + \langle a_m^f a_n^f \rangle - \langle a_m^o a_n^f \rangle - \langle a_m^f a_n^o \right] C_{km} C_{kn}$$

With brackets $\langle\ \rangle$ we denote the average of a large number of cases. In order to find the values of C_{ki}, i=1,...,N, that minimize E_k, differentiate with respect to C_{ki}, and put the derivative equal to zero.

$$\frac{\partial E_k}{\partial C_{ki}} = -2\left[\langle a_k^f a_i^f \rangle - \langle a_k^f a_i^o \rangle\right] + 2 \sum_j C_{kj} \left[\langle a_i^o a_j^o \rangle + \langle a_i^f a_j^f \rangle - \langle a_i^o a_j^f \rangle - \langle a_i^f a_j^o \rangle\right] = 0 \qquad (2)$$

In this way one obtains a set of N linear equations. The solution of this set yields the weights C_{ki}. We assume that the observation and prediction errors are not correlated with each other, so

$$\langle a_i^o a_j^f \rangle = \langle a_k^f a_i^o \rangle = 0 \qquad (3)$$

In matrix notation the set (2) can be written as

$$(\underset{\sim}{B} + \underset{\sim}{P})\,\underset{\sim}{c} = \underset{\sim}{M}\,\underset{\sim}{c} = \underset{\sim}{d}$$

where the elements of $\underset{\sim}{B}$, $\underset{\sim}{P}$, $\underset{\sim}{c}$ an $\underset{\sim}{d}$ are given by

$$B_{ij} = \langle a_i^o a_j^o \rangle, \quad P_{ij} = \langle a_i^f a_j^f \rangle, \quad c_i = C_{ki}, \quad d_i = \langle a_i^f a_k^f \rangle$$

Once the weights C_{ki} are determined, it is possible to find an explicit expression for E_k, the squared error in the analysis. By summing the normal equations (2) over i, substituting into the expression for E_k and using (3) one obtains

$$E_k = \langle a_k^{f\,2} \rangle - \sum_m C_{km} \langle a_k^f a_m^f \rangle$$

It is not necessary that all A_m refer to the same quantity. For example, A_1^p and A_1^o can refer to the 500 mbar height, A_2^p and A_2^o to the North-South component of the wind of the same radiosonde, A_4^p and A_4^o to a satellite thickness, and so on.

For the determination of the matrix $\underset{\sim}{P}$ information is needed about quantities $< a_i^p \, a_j^p >$. We discuss only the case that both a_i^p and a_j^p refer to the height of an isobaric level. (With the aid of the geostrophic relations one can derive the value of $< a_i^p \, a_j^p >$ in other cases.) It is too complicated, and also unnecessary, to compute $< a_i^p \, a_j^p >$ for every possible pair of stations. Instead, one assumes the correlation

$$\frac{< a_i^p \, a_j^p >}{[< a_i^{p\,2} >< a_j^{p\,2} >]^{1/2}}$$

to be a function of r, the distance between the stations i and j. This function has the form e^{-br^2}. This implies that the direction of i with respect to j is not important, and also the exact location of station i is not important. The variance $< a_i^p \, a_i^p >$ is specified as a function of latitude only. The quantity b depends on the first-guess field used, which can be climatology, persistence or a forecast from a numerical model.

The vector $\underset{\sim}{q}$ describes, in the same way, correlations between the observation points and the analysis point.

The matrix $\underset{\sim}{B}$ specifies statistical information about the errors of the observation system.

Most types of observations have independent errors, so many elements of the matrix $\underset{\sim}{B}$ are 0. The numbers on the main diagonal specify the variance of the observation errors. If $i \neq j$, B_{ij} vanishes unless the errors of observations i and j are correlated. For instance, the errors of satellite observations of different places are correlated, and so are the errors of the height of different levels observed by one radiosonde.

4. Division into boxes

The most straightforward way of solving the normal equations (2) is to set up a new matrix $\underset{\sim}{M}$ for every grid point. $\underset{\sim}{M}$ depends only on which observations are used for the analysis of grid point k. However, at the ECMWF one uses a different approach.

The atmosphere is divided into boxes, having horizontal dimensions of about 700 kilometres. The same set of observations is used for the analysis of all grid points within the box. Of course, normally this set of observations is taken from a larger region than the analysis box itself. In the matrix equation (3) this implies that for all grid points the same matrix is used, only the right-hand side vector $\underset{\sim}{d}$ is different for every grid point. Therefore, only one matrix inversion per box is needed instead of one inversion for every grid point as is usually done. The consequence of this subdivision is the risk of discontinuities at the boundaries. The last sections of this report describe some experiments to investigate how serious this problem is.

5. Discontinuities at the box boundaries

As a first experiment a small-scale analysis was made. The grid points were situated along the 40°N parallel, from the Greenwich meridian to 40°W with a spacing of one degree. In the first experiments the first-guess was formed by climatology, while in all experiments the observations of 1976 Feb. 10th, 00 GMT were used. These observations were taken from the Data Systems Test magnetic tapes, made for FGGE purposes.

In the expression for the correlation function of the height prediction error, e^{-br^2}, a value of $0.61 \times 10^{-12} m^{-2}$ was used for the quantity b. The distance between two stations, expressed in metres, is r. This value for b implies a high correlation between observations (or better, between deviations of the first-guess field from the true field). Further details about the correlations and the measure of the various observation error statistics can be found in [2]. The analysis program computes the normalized increments, i.e. corrections to the first-guess height field, divided

by the standard deviation of the prediction error estimate. This quantity is shown in some of the following figures. In the notation of section 2 the normalized increment is given by

$$\frac{a_k^i - a_k^p}{\langle a_k^r a_k^p \rangle^{\frac{1}{2}}} = \frac{\Sigma_i C_{ki}(a_i^o - a_i^p)}{\langle a_k^r a_k^r \rangle^{\frac{1}{2}}}$$

In figure 4 results of the first experiment are given. The weight given to a specific observation is plotted as a function of the place of the grid point that is analyzed, while the edges of the boxes are indicated as well. It is evident there are jumps in the weight of this observation, especially at the boundary between two boxes near 2^{o}W.

In figure 5 the normalized increment is plotted. In box nr. 207 the gradient is smaller than in the neighbouring box. nr. 208, while there is a jump in height between box. nr. 206 and 207. By some coincidence the increment in the region 40^{o}W to 20^{o}W is always very near to +2.0.

The conclusion that can be drawn is that in these experiments there are unacceptable jumps in the analysis at the box boundaries. The discontinuities are caused by data that are included in the matrix $\underline{M}$ for one box and not included in the matrix of the neighbouring box. To prevent jumps in the analysis an observation should be included for the analysis of a box whenever its correlation with any analysis point within the box is significantly different from zero. However, due to computer limitations, the order of the matrix is limited to a maximum of 151, so not more than 151 data values per box can be taken into account. The smaller the correlation of an observation with the analysis point, the smaller is the effect of neglecting the observation. It can be expected that the discontinuities are smaller if a smaller correlation function is chosen. Therefore, another set of experiments was taken with a different and more realistic correlation function. The operational analysis will use a forecast as a first-guess instead of climatology. In this case the correlation function for the prediction error has a narrower shape, and this will decrease the influence of stations, especially if they are far from the analysis point. For the correlation function of the error of the forecast field a value of b was used of $2.00 \times 10^{-12} m^{-2}$.

A number of small-scale analyses of an area west of Europe were made to check whether a realistic correlation function based on a forecast as first-guess field would overcome the difficulties of jumps. Depending on the number of radiosondes available, the analysis program has two modes of operation. If there are many radiosonde data, the analysis proceeds level by level, i.e. for each level the normal equations are solved. All non-radiosonde data are on principle available for selection in the matrix $\underset{\sim}{M}$, but from the radiosonde data only the data from the level that is analyzed at the moment are used. In areas with few radiosondes all information of all observations is used, with the restriction that not more than 151 data values can be included. A data value is, for instance, one satellite thickness observation, or one wind component of one level of a radiosonde. As will be shown later, the restriction to 151 data values is a severe one.

The clearest way to show discontinuities is using data only from the box that is analyzed. In this way two boxes never have any observation in common. The result of such an analysis is displayed in figure 6. The boxboundaries clearly show up in this figure. Of course, such an analysis is unacceptable.

In figures 7 and 8 is shown the result of an analysis under more realistic conditions. One cause of jumps is eliminated, namely the different way of analyzing boxes with many or few radiosondes. The boxboundaries remain visible. The jumps are caused by the fact that two neighbouring boxes are not analyzed with the same information. Therefore, a closer look at the data selection was taken. It appeared that rather much of the space in the matrix was taken by satellite data, possibly excluding more important information from stations situated farther away. For this reason one experiment was taken with all satellite information omitted. It can be seen in figure 9 that this did not help very much. A boxboundary over the United Kingdom and along 40°N is easily identified.

The last example (figure 10) shows an analysis in which only radiosonde data are used. It is the only example of an analysis that appears smooth at the boxboundaries.

6. Requirements for data selection

As already mentioned before, the data used for the analysis should be selected in a more careful way. We shall first describe the procedure that was used in all experiments shown (except for the experiment of figure 6). A neighbouring box is defined as a box that has a boundary in common with the box under consideration, or is less than half a boxsize away. A boxsize is 660 km. The order of selection of data is as follows: first, take all data from the central box, next all radiosonde data in the neighbouring boxes, and finally all other data in the neighbouring boxes. The neighbouring boxes are selected in the order shown in figure 11. As soon as 151 data values are reached, the rest of the information is omitted. If there are less than 151 data values in the central box and the neighbouring boxes, information of neighbours of neighbours is used, and so on. A closer look at the data used for the analysis of a specific box can be found in figure 12. It can be seen that the order of selection of neighbouring boxes is not optimal. For instance, when analyzing box nr. 163, non-radiosonde observations of box 208 are used in preference to those of box 123. Also boxes 164 and 165 have, with the exception of radiosondes, almost no observations in common.

In this way it is not surprising to find discontinuities at the boxboundaries. The discontinuities at the surface level are probably larger.

The consequence is that too many observations are not used, even though they can have a considerable influence on analysis points near the boundary.

In order to make the analysis smoother at the edges, a better selection of data is necessary. Consider the box being analyzed. One should make sure that a certain amount of data of the neighbouring boxes are used. In data-dense areas it is impossible to use all information of all neighbouring boxes. Typically, there are about 6 neighbouring boxes (including the so-called overflow box, that contains extra information from the central box), each of which can cantain information of, say, 5 radiosondes, 10 surface observations, and 3 satellite soundings for a total of (5 x 3 + 10 x 3 + 3 x 10) x 6 boxes = 450 data values. Even if the number

of data values in a satellite sounding is reduced, the number of values (and hence the size of the matrix to be inverted) can be very large.

7. Order of selection of information

To prevent the jumps of the analysis at the boxboundaries, one should subdivide the neighbouring boxes. The easiest way to do this is a subdivision of the neighbouring boxes into four subboxes. When analyzing a certain box, one should make sure that at least all data that are not more than 350 km away are included. This means that data are used from an area at least $4\frac{1}{2}$ times as large as the area of one box.

It is possible that at this stage the matrix is already too large to be handled. The criterion for this depends on the maximum matrix size the computer can handle. If the matrix is too large, there are two possibilities: either more observations should be combined to form one superobservation, or the central box itself must be subdivided. This subdivision makes it possible to do a separate analysis of the four subboxes.

In datasparse areas it is possible (and necessary) to use observations farther away. The selection of the order in which data are included must be as simple as possible, and new data should be selected first in directions where there are not many observations yet. This implies that we need two scans through the data. A possible way of doing this will be outlined now.

First, we need a count of the observations in every box. With this count we can select the direction in which the least amount of information is available.

The reduction in error variance at a certain analysis point in the central box will of course depend on the presence of radiosonde data in the subbox. However, two radiosonde observations have a high correlation with each other, and therefore the exact number of radiosondes seems less important. It might be sufficient to make only a determination of the <u>presence</u> of any radiosondes in the subbox, and not of the <u>total number</u> of the observations. Other observation types, like satellite observations, can be treated in the same

way. For each observation type i we get a number c_i that is zero, if the type is absent in the subbox, and one, if present in the subbox. The total information content of the subbox, I, is equal to the sum over all information types:

$$I = \sum_i w_i c_i$$

w_i is a number that depends on the observation accuracy of observation type i and is not specified here.

If the information content of all subboxes is determined, then it is known in which direction the least information is available.

The next step is adding to the matrix all information of the subbox or subboxes not used yet, that are in this direction of the least information.

Acknowledgements

This report is the result of a visit to the European Centre for Medium Range Weather Forecasting. I am indebted to the Centre for the opportunity to study and use their programs. Special thanks are due to Dr. L. Bengtsson, G. Larssen, A. Lorenc, C. Little and C. Clarke of the analysis section for stimulating discussions and help.

References

1 A. Lorenc, I. Rutherford and G. Larsen.
The ECMWF Analysis and Data-Assimilation Scheme: Analysis of Mass and Wind Fields. Technical Report ECMWF Nr. 6.

2 G. Larsen, C. Little, A. Lorenc, I. Rutherford.
Analysis Error Calculations for the FGGE. Internal Report 11 of ECMWF.

-o-o-o-

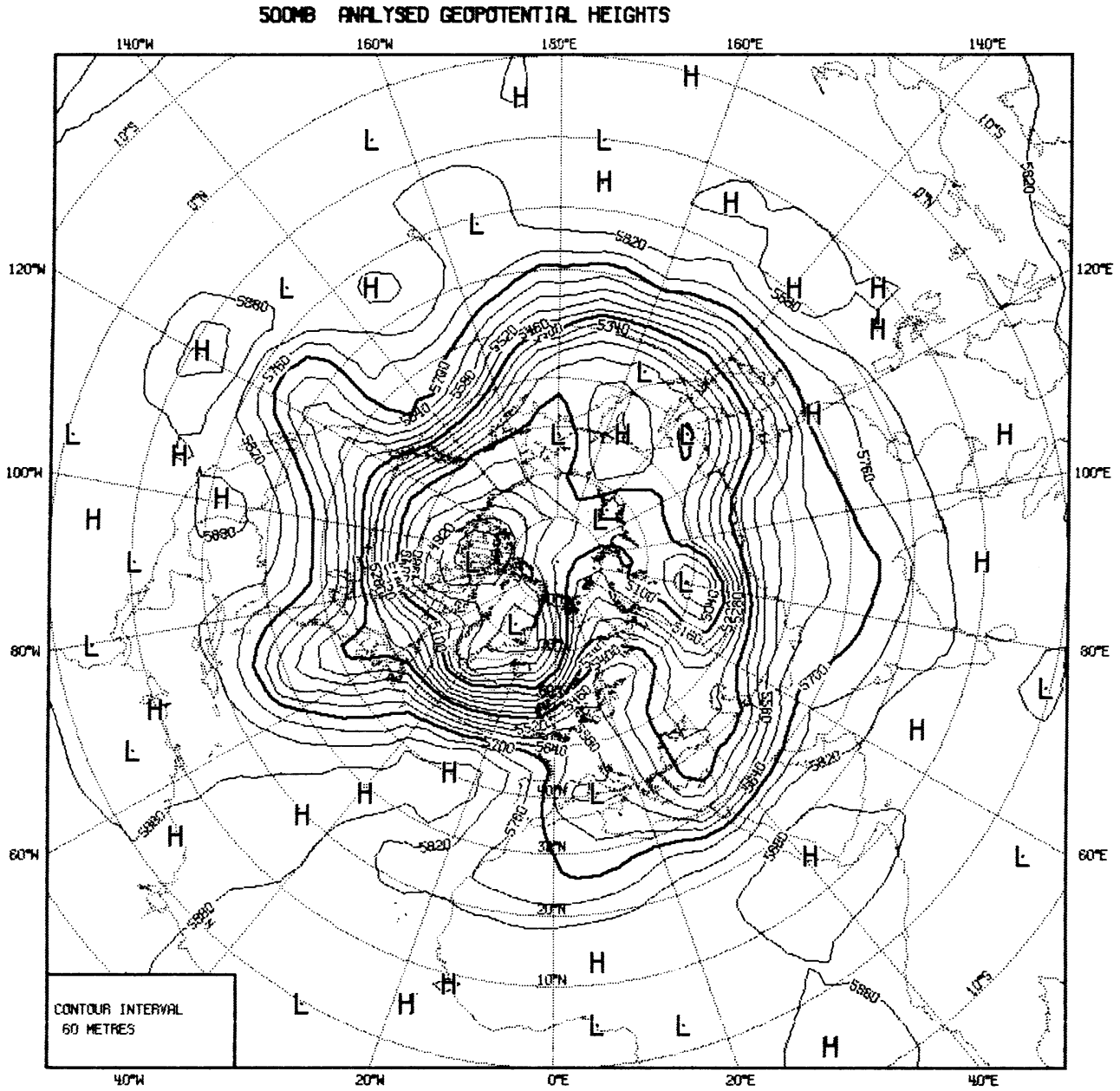

Fig. 1 Analysis of 1976 Feb. 10th, 00 GMT. All available observations are used.

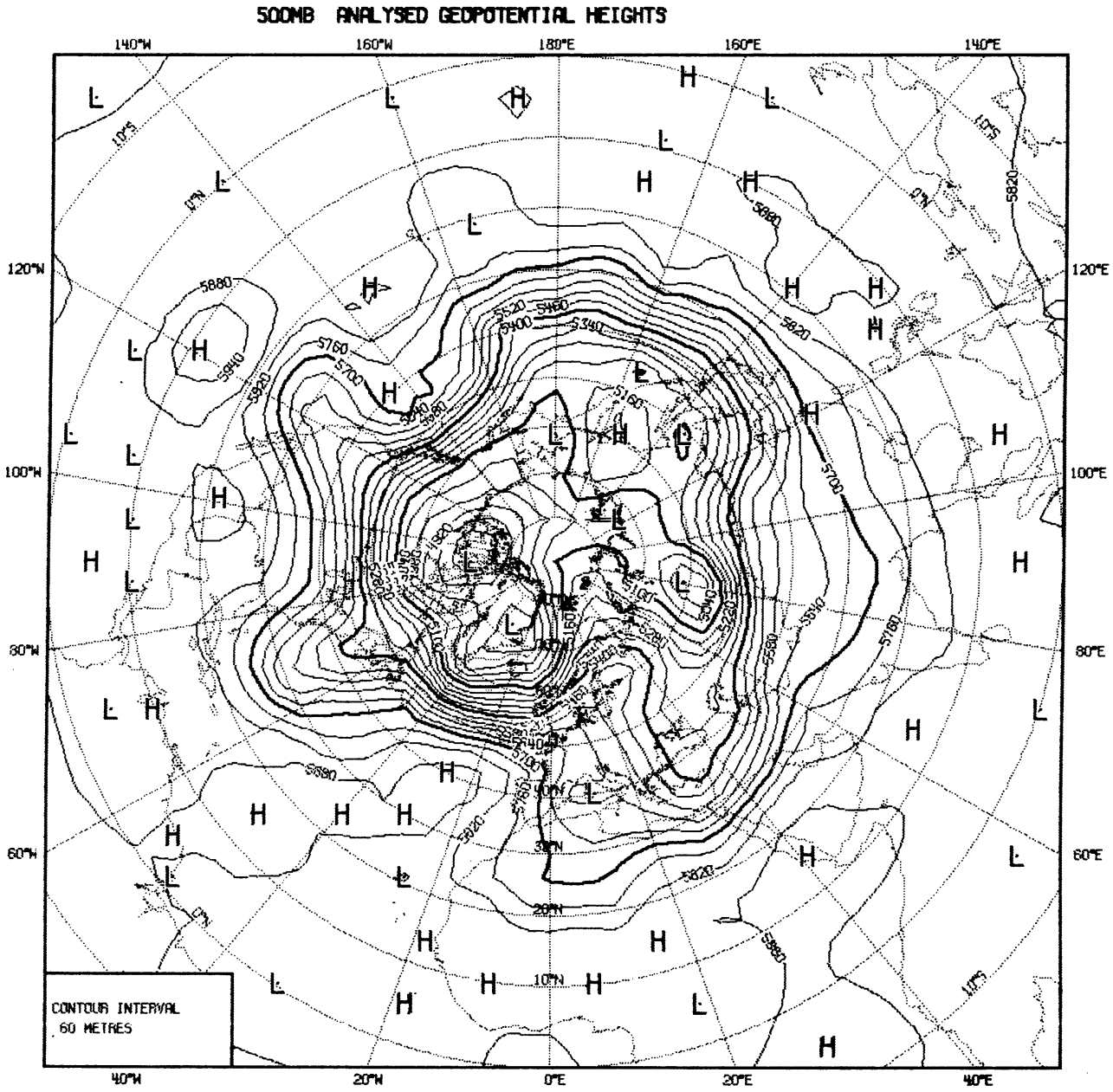

Fig. 2 Analysis of 1976 Feb. 10th, 00 GMT. All available observations are used, exeept satellites.

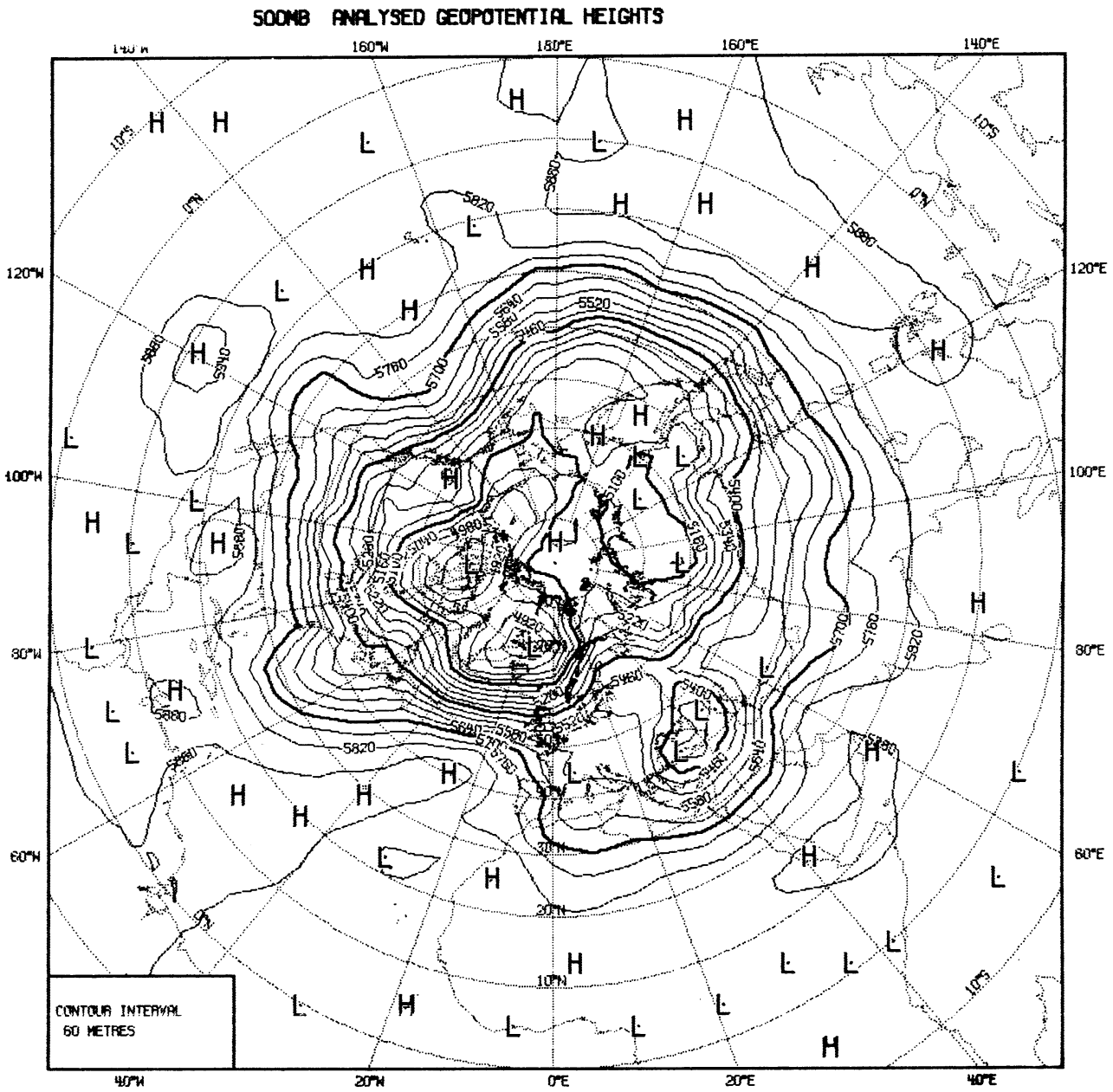

Fig. 3 Analysis of 1976 Feb. 10th, 00 GMT. All available observations are used, except radiosondes.

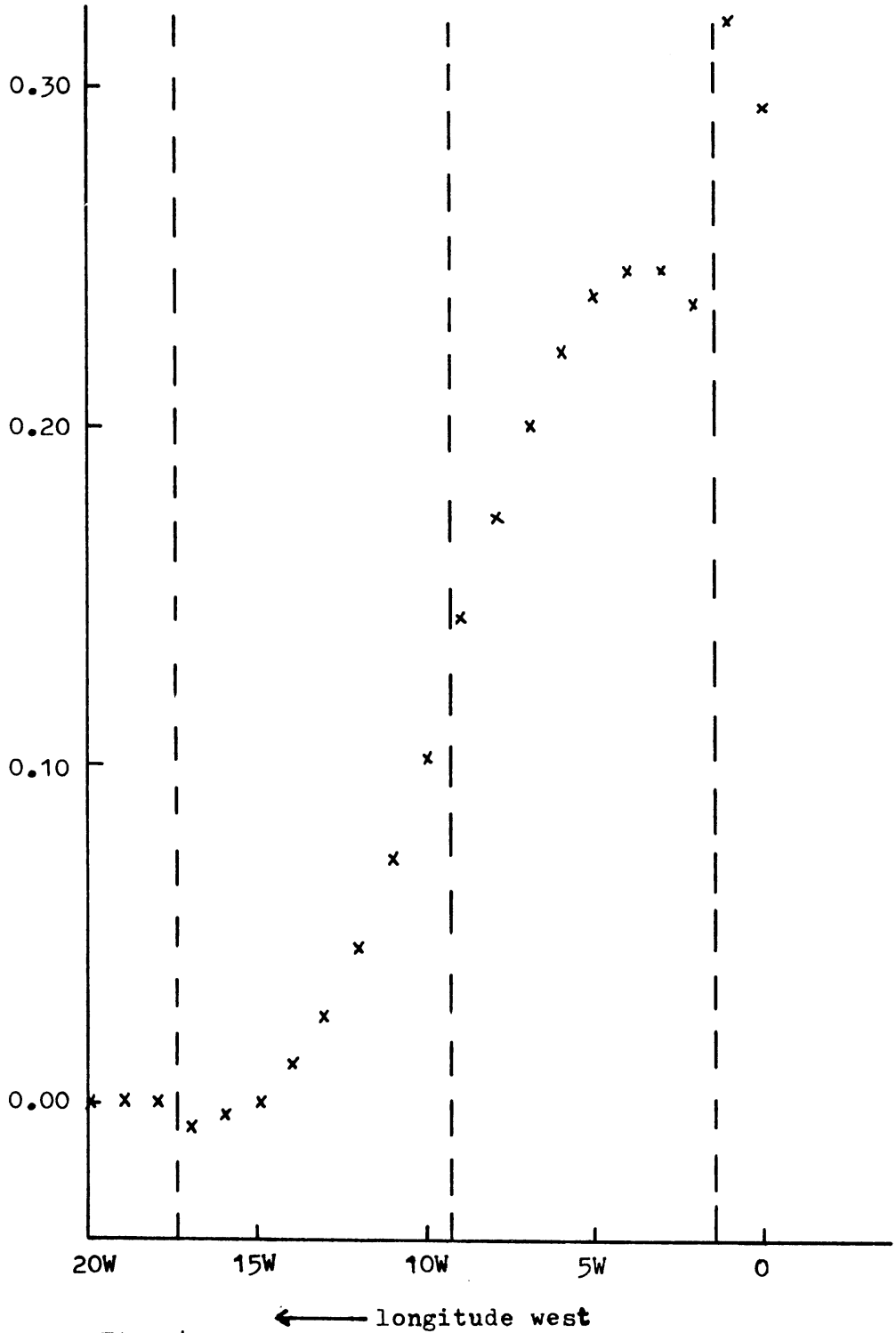

Fig. 4 Weight given to the height observation of the 500 mbar level of a radiosonde at 40°N, 3.6°W (Lisbon).

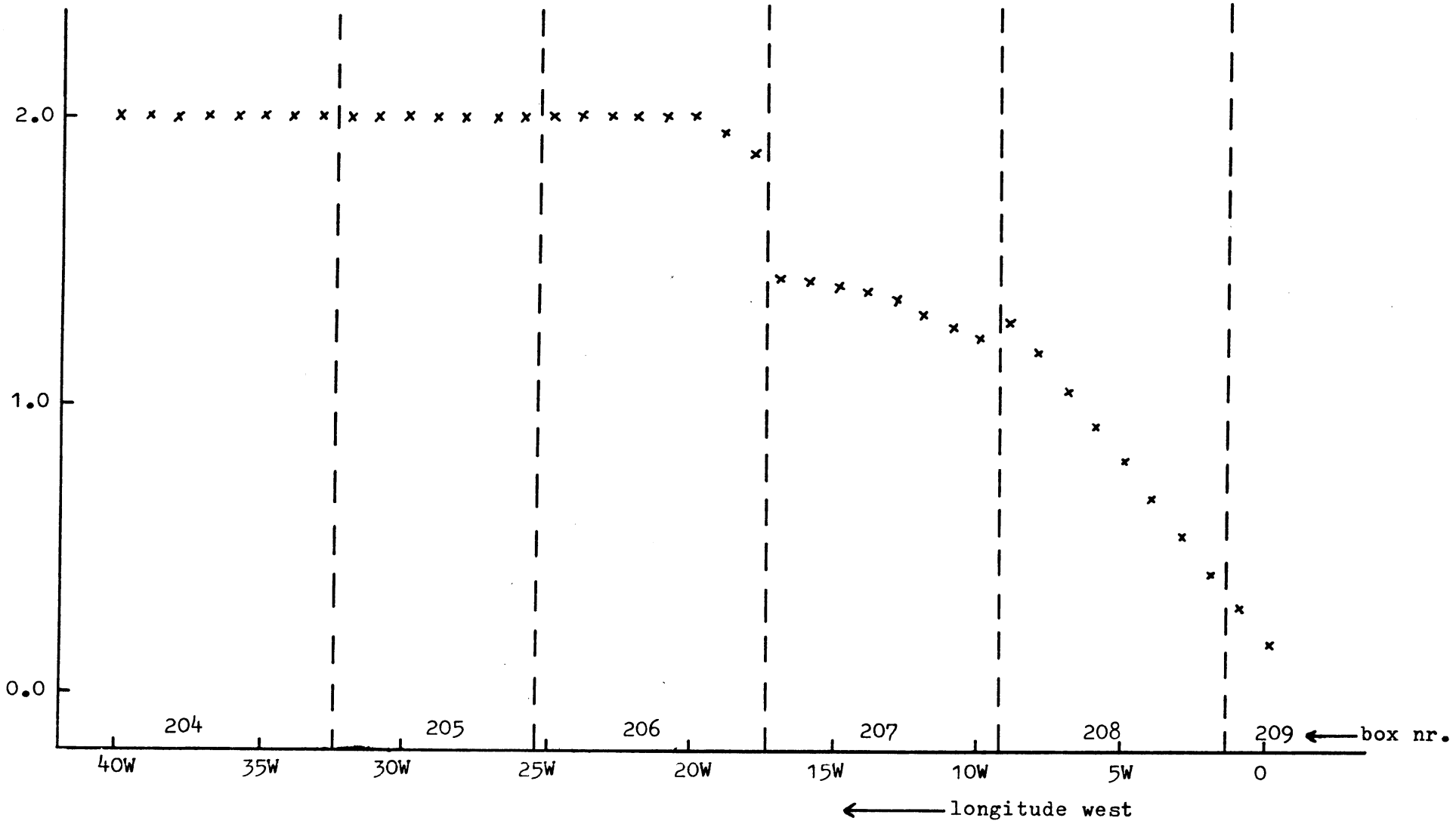

Fig. 5 Normalized increments of every grid point between 40°W and 0°W at 40°N.

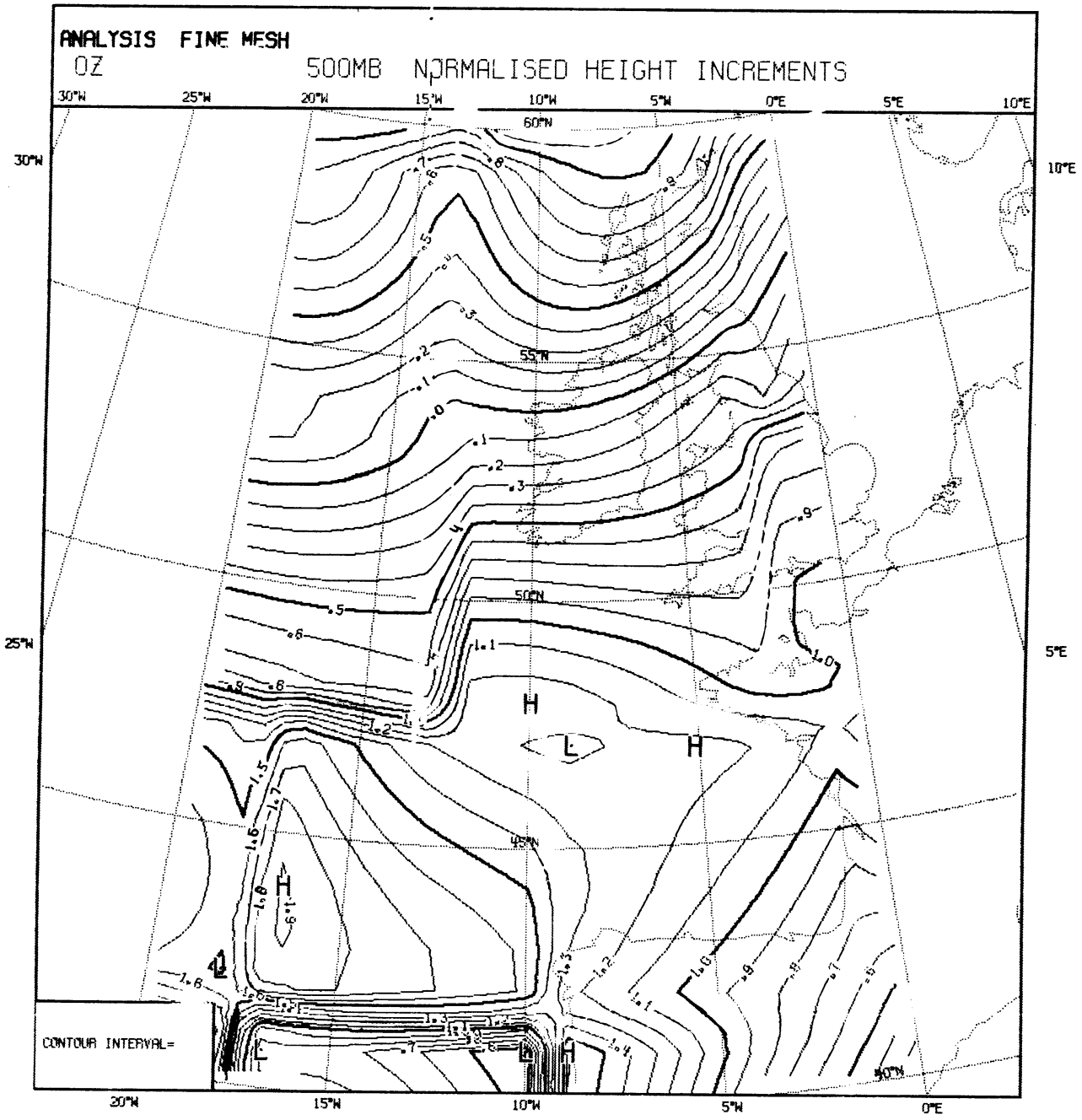

Fig. 6 Analysis, in which for every box only information was used from the same box and not from surrounding boxes.

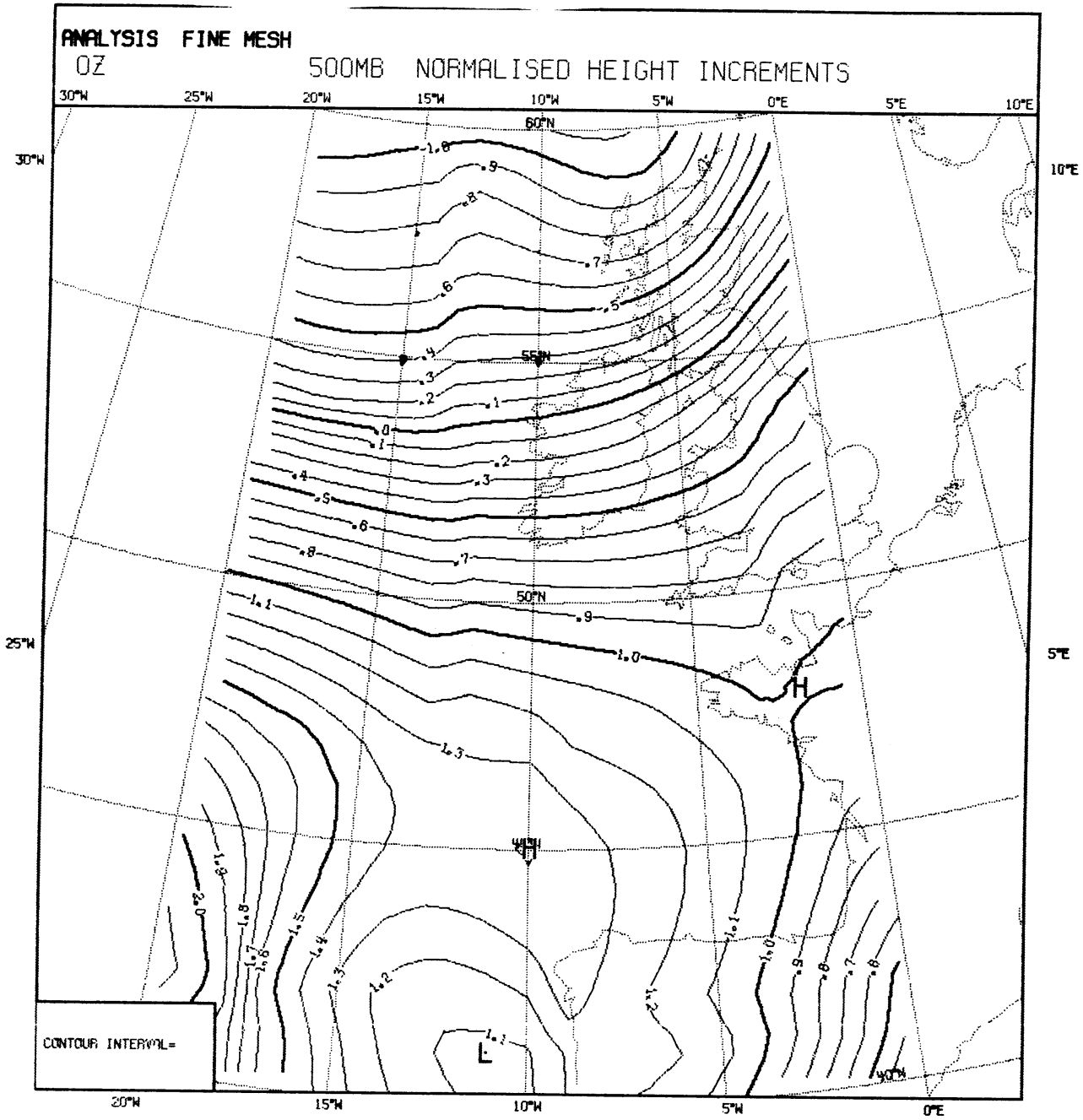

Fig. 7 Example of an analysis with grid points spaced by one degree. All information is used.

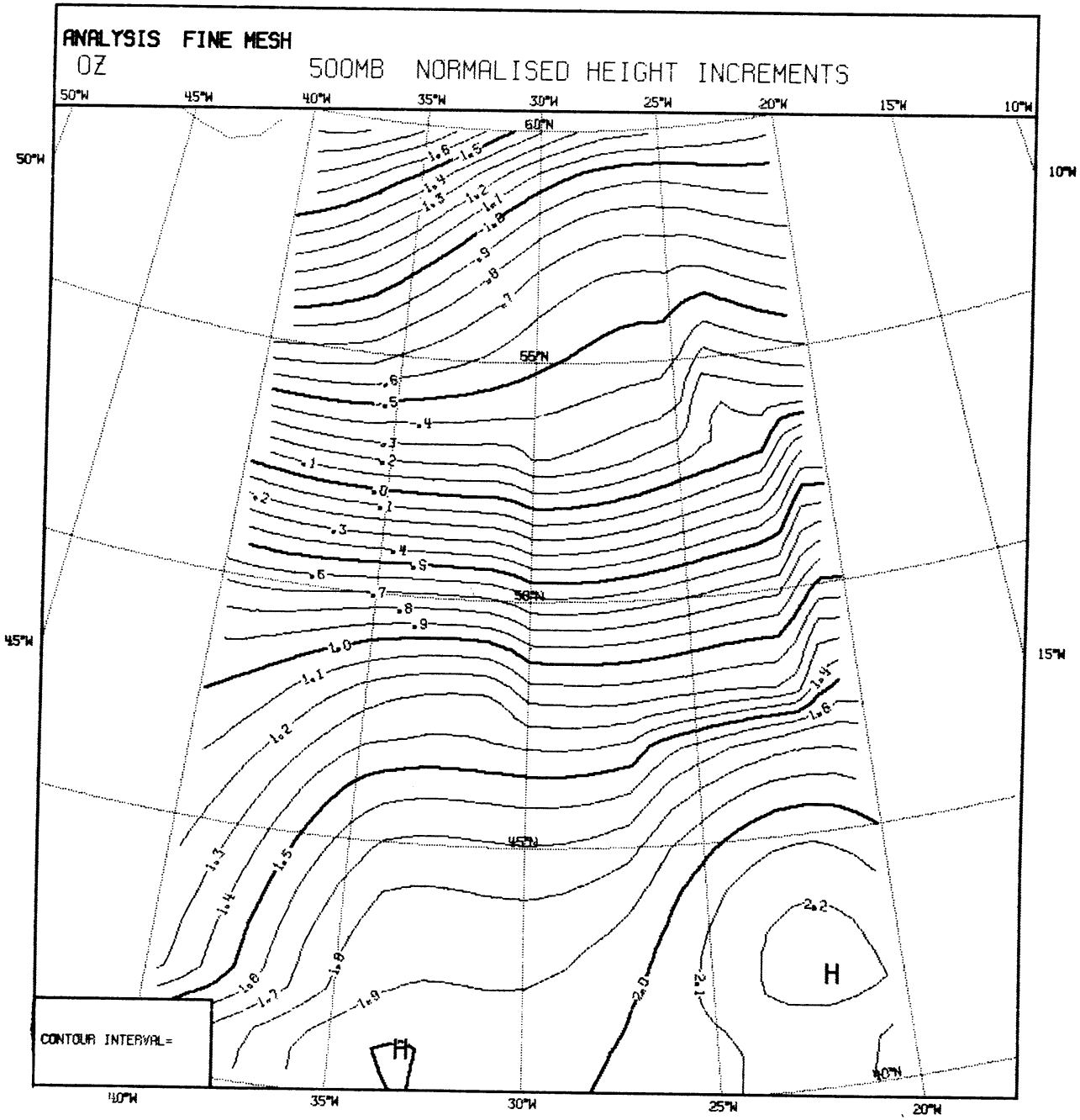
ANALYSIS FINE MESH
OZ
500MB NORMALISED HEIGHT INCREMENTS
50°W
45°W
40°W
35°W
30°W
25°W
20°W
15°W
10°W
60°N
55°N
50°N
45°N
40°N
H
CONTOUR INTERVAL=

Fig. 8 As Fig. 7.

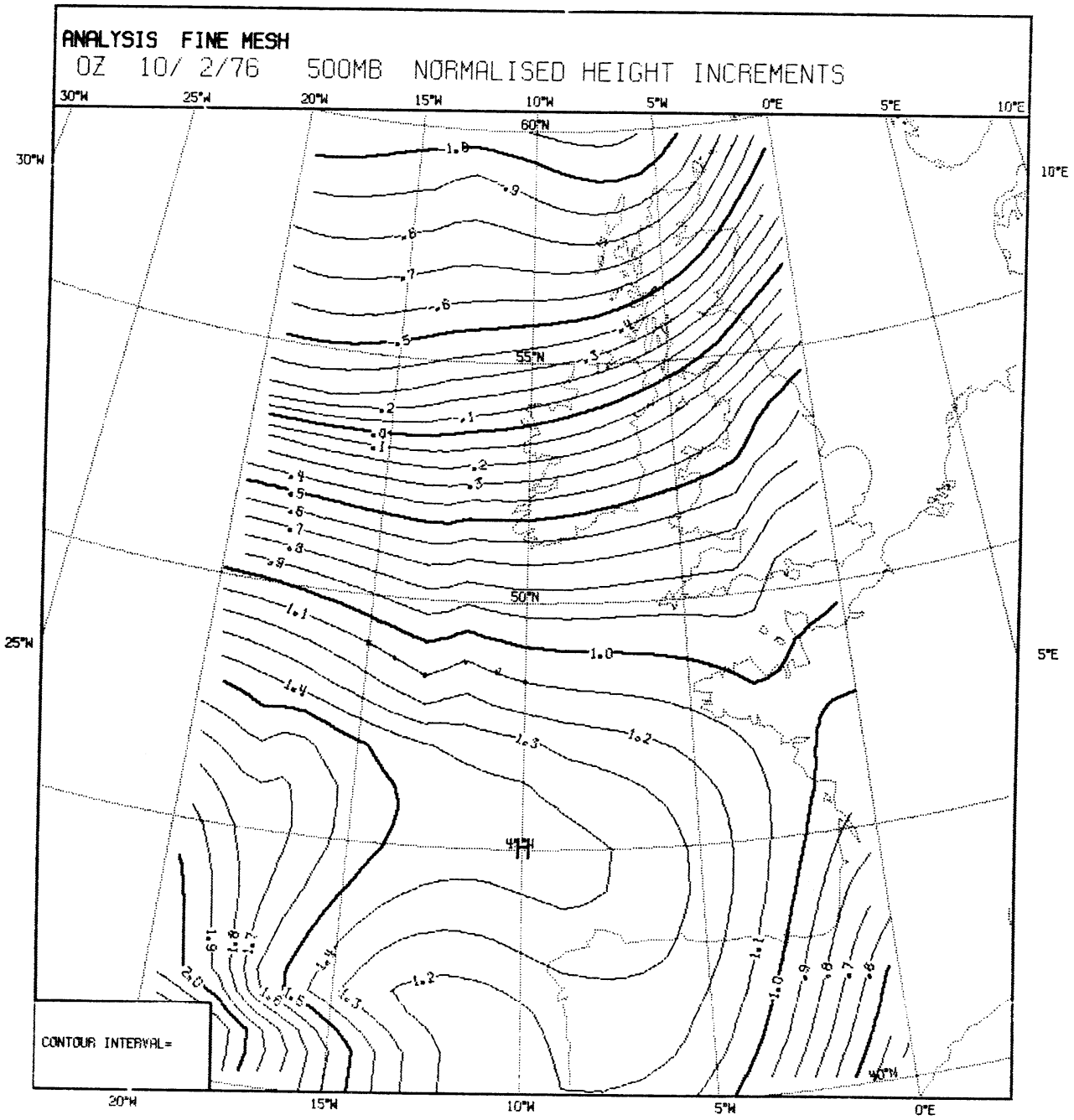

Fig. 9 As Fig. 7, but no satellite information is used.

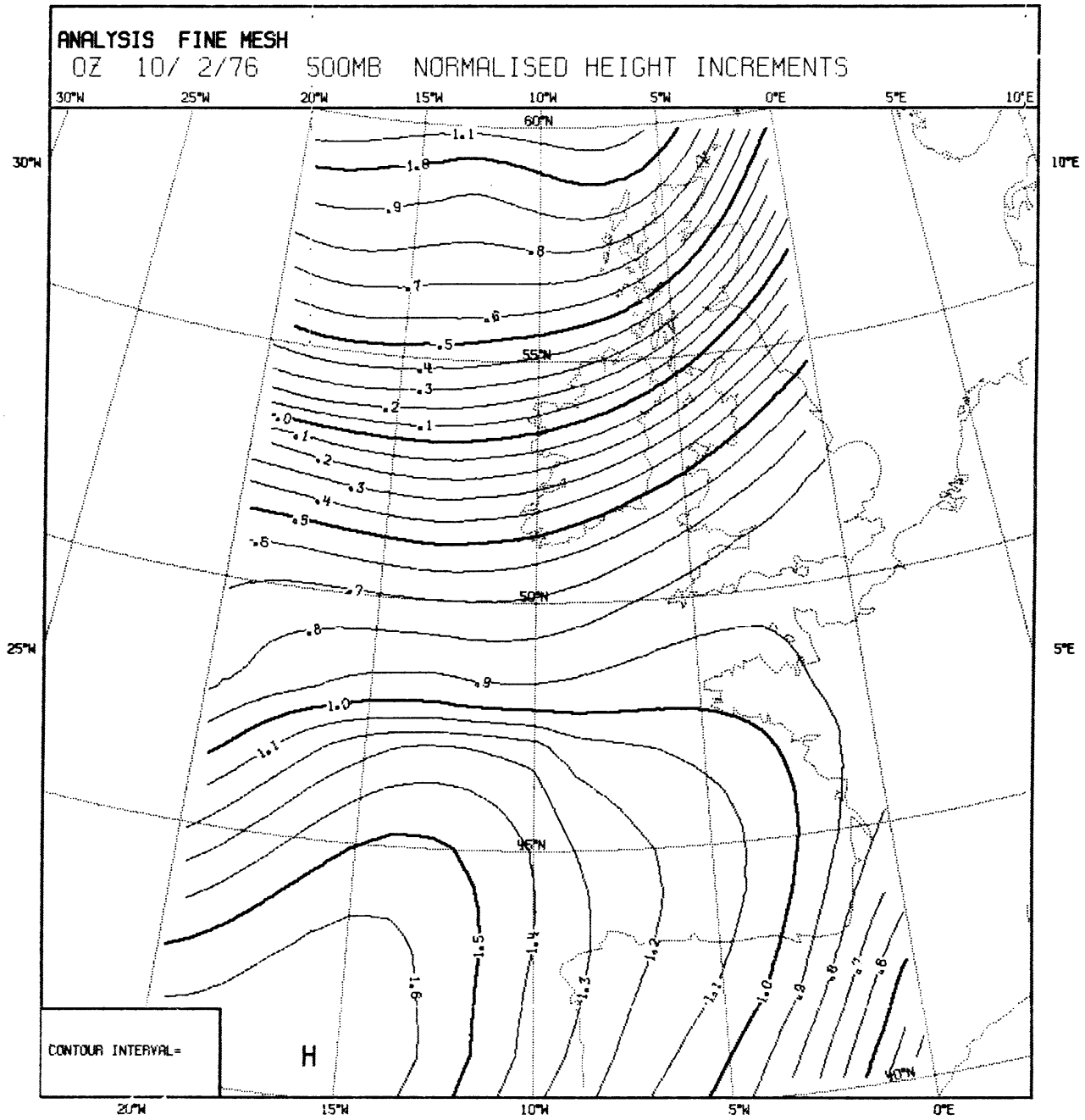

Fig. 10 As Fig. 7, but only radiosonde information is used.

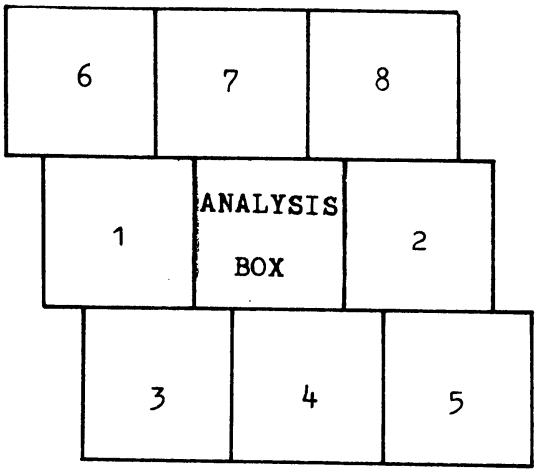

Fig. 11 In this figure the order in which information is selected for the analysis of the central box is given.

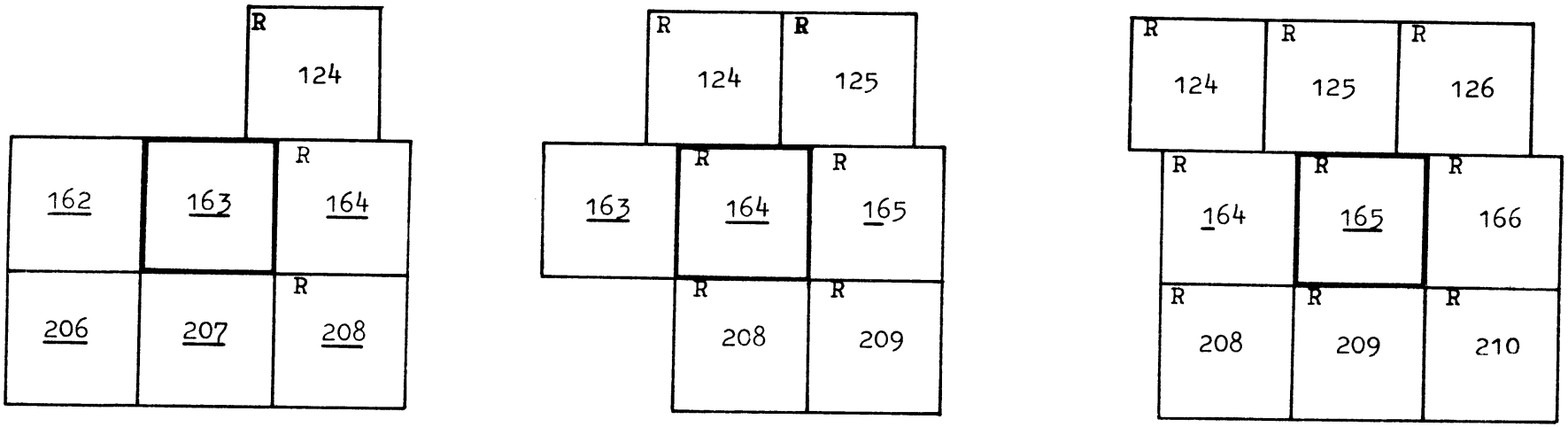

Fig. 12 Boxes which supplied information for the analysis of the central box are shown. The analysis box is indicated with a heavy line. The information supplied can be from radiosondes (these boxes have an R in the upperleft corner), or from other information (of these boxes the box number is underlined).