Determination of eof's from a large sample

Vesa Karhila and Juhani Rinne

University of Helsinki, Department of Meteorology

DETERMINATION OF EOF'S FROM A LARGE SAMPLE

Vesa Karhila and Juhani Rinne

1. INTRODUCTION

EOF's have been determined from 11876 500 mb analyses given on a round grid of 1404 gridpoints.

Normally EOF's are found as eigenvectors of a covariance matrix computed from an analysis sample. From covariance matrix with a dimension of 1404 we can find 1404 eigenvectors (EOF's). From these 1404 EOF's we would normally use at most some 200. Instead of computing a huge covariance matrix for determination of all 1404 EOF's (which is above our computer re ources) we have computed sets of EOF's from smaller covariance matrices for smaller areas and joined these sets with a special technique which we discuss below.

The loss of variance during different joining operations and the accuracy of the EOF's determined are also discussed. The convergence of EOF-series on the 6th of Sep. 1968 is shown as an example.

2. THE SAMPLE

Our data come from Fleet Numerical Weather Center (FNWC) Monterey, California. The data consists of 14733 500 mb analyses which cover the period from Nov. 1945 to Dec. 1970. Analyses are given in the 63 * 63 square FNWC-grid which has the grid interval of 381km at 60^{o}N.

This material was divided into dependent and independent samples. EOF's were determined from the dependent sample which consists of the years 1946-1952, 1956-1959 and 1962-1970, altogether 11876 analyses. The remaining 2857 analyses were left as test samples.

Only 1404 gridpoints forming a round area in the middle of the FNWC-grid were used (out of 3969 possible). Because of map distortion this round grid covers 56% of the original area although only 35% of the original gridpoints are used.

3. DETERMINATION OF THE EOF'S

The round grid was divided into four subgrids, 351 gridpoints each, by dividing each fourth gridpoint of the one dimensional representation of the round grid into one subgrid. A covariance matrix was computed for each subgrid using all 11876 analyses and 351 EOF's were determined for each subgrid. These EOF's were transformed to time dependent coefficients C (see (A.2) in Appendix). With 351 components in series expansions all the original variance 25175m^2 is explained.

Fig. 1 is a schematic illustration of the joining procedure performed. The variances given in fig. 1 are partial: the total variance is obtained by summing the variances of the subgrids.

By the nature of EOF-series in explaining redundant information most of the variance is explained by first components. On the other hand high indexed tail components describe mainly noise and errors of the original analyses. Thus we can and we must truncate our EOF-series and instead of using all 351 components per subgrid we use only 175. In these truncation about 14 m^2 of variance is lost from each subgrid, altogether some 55 m^2 is lost.

In two truncated sets of coefficients C we have the main information of these both subgrids. However there is still a lot of redundancy in the two sets of coefficients and they are not orthogonal, they are not easy to use. We can remove this redundancy by orthogonalizing the two sets of coefficients. We can compute covariances between coefficient time series. As eigenvectors of such matrices joining functions J are found. With these functions J new coefficients can be determined. These new coefficients are orthogonal and valid over joined grid area. Functions f can also be joined with the aid of J's.
(A detailed description of this method will be published later).

By repeating this method the final set of coefficients and functions are found. In our case this method was applied three times.

The first two times were at the level of quarter subgrids. After the truncations covariance matrices were computed and functions J^1 and J^2 were determined. With these functions two sets of coefficients C were found, these new coefficients being orthogonal in their own subgrids both covering half of the round grid. Only 175 coefficients per analysis and per subgrid were determined out of 350 possible. In these truncations about 20 m^2 of variance was lost from both subgrids, altogether about 40 m^2.

The method was applied a third time on these new truncated sets of coefficients on the level of half area subgrids. Joining functions J^3 were determined and final coefficients valid over the original round area were computed. Again only 175 coefficients out of 350 were computed. In this final truncation 32 m^2 were lost. Thus in these seven truncations done during the three joining steps 126 m^2 of the original variance was lost, which is 0.50%. This much must be sacrified if the degrees of freedom are to be reduced from 1404 to only 175.

4. PROPERTIES OF EOF'S DETERMINED

4.1 Variance reduction

EOF's were also tested in independent (in respect to synoptic changes) samples. Variance reduction in different samples are shown in table 1.

	sample	cases	original variance	explained variance	%
dependent	1946-1970 exl.53-55 60-61	11876	25175	25048	99.50
independent	1953-1955	1370	25878	25777	99.61
	1960-1961	1426	25745	25670	99.71

TABLE 1. Variance explained by 175 EOF's.

Higher values in independent samples, which seem paradoxical,may be due to smoother analyses which are more easily explained by eof's.

4.2 Example of two single EOF's

As an example two EOF-components are shown, namely functions f_{14} and f_8. In the component f_{14} (fig 2) there is a big similarity with trigonometric functions. Wavenumber 5 is clearly seen although there is quite a lot of asymmetry included. Component f_8 (fig 3) is an example which is far from any single spherical harmonic. This component is concentrated over Atlantic where it tends to build a high (or a low). A lot of mathematical components are needed to achieve the same variance reduction as with this single 'local' component in EOF-series.

4.3 Example of EOF-series expansion

Figs 4-14 show how EOF's work in representing the 500 mb analysis on Sep. 6th, 1968. This case was chosen to see how the blocking high over Scandinavia is being built up by EOF's. The formula (A.1) in this case was

$$Z = Zm + 143 * f_1 + 1 * f_2 + 16 * f_3 + 15 * f_4 + \dots - 0.4 * f_{170} .$$

Fig 4 is the constant term zm, which is the mean field computed from 11876 analyses.

In figs 5-13 the left hand pictures (5a-13a) are the effect of single terms $C_\mu f_\mu$ and the right hand pictures (5b-13b) show the cumulative sum up to μ (μ= 1,2,...,6, 20, 50 and 100). The blocking is being built up mainly by components 5 and 6 (figs 9 and 10).

Fig 14 shows the EOF-series expansion after 170 components have been summed (degrees of freedom 170). Fig 15 shows the original grid analysis for comparison (degrees of freedom 1404).

APPENDIX

BASIC FORMULAE AND NOTATIONS

A variable z, here the 500 mb height field, can be represented with the aid of EOF's

$$\text{(A.1)} \quad z(x,t) = zm(x) + \sum_{\mu=1}^{N} C_\mu(t)\, f_\mu(x) + \text{residual}\,(x,t,N),$$

where zm is a long time mean field, C's are time dependent coefficients, f's are space functions, x refers to a gridpoint or a space coordinate, t refers to time and N is the truncation point of the EOF-series.

Space functions f are to be determined from a sample (the dependent sample) so that they will be orthogonal and normalized to one

$$\text{(A.2)} \quad \sum_{x \in A} f_\mu(x)\, f_\nu(x)\, dA(x) = \delta_\mu^\nu ,$$

where δ_μ^ν is Kronecker's delta, A is the area covered by the grid and the dA's are elemental areas associated with gridpoints and normalized by

$$\text{(A.3)} \quad \sum_{x \in A} dA(x) = 1.$$

Time dependent coefficients C will be orthogonal in the dependent sample and they will be normalized to the corresponding eigenvalues λ

$$\text{(A.4)} \quad \overline{C_\mu C_\nu} = \delta_\mu^\nu \lambda_\mu ,$$

where the bar $\overline{\quad}$ refers to time average. These coefficients C for any 500 mb height field, both in the dependent and independent samples, can be determined by

$$\text{(A.5)} \quad C_\mu(t) = \sum_{x \in A} f_\mu(x) \left[z(x,t) - zm(x) \right] dA(x).$$

If only coefficients C of the dependent sample are known the space functions f can be determined from the dependent sample by

$$(A.6) \qquad f_\mu(x) = \frac{1}{\lambda_\mu} \overline{C_\mu(t)\left[z(x,t) - zm(x)\right]}.$$

The variance of a single analysis explained by N EOF's is simply

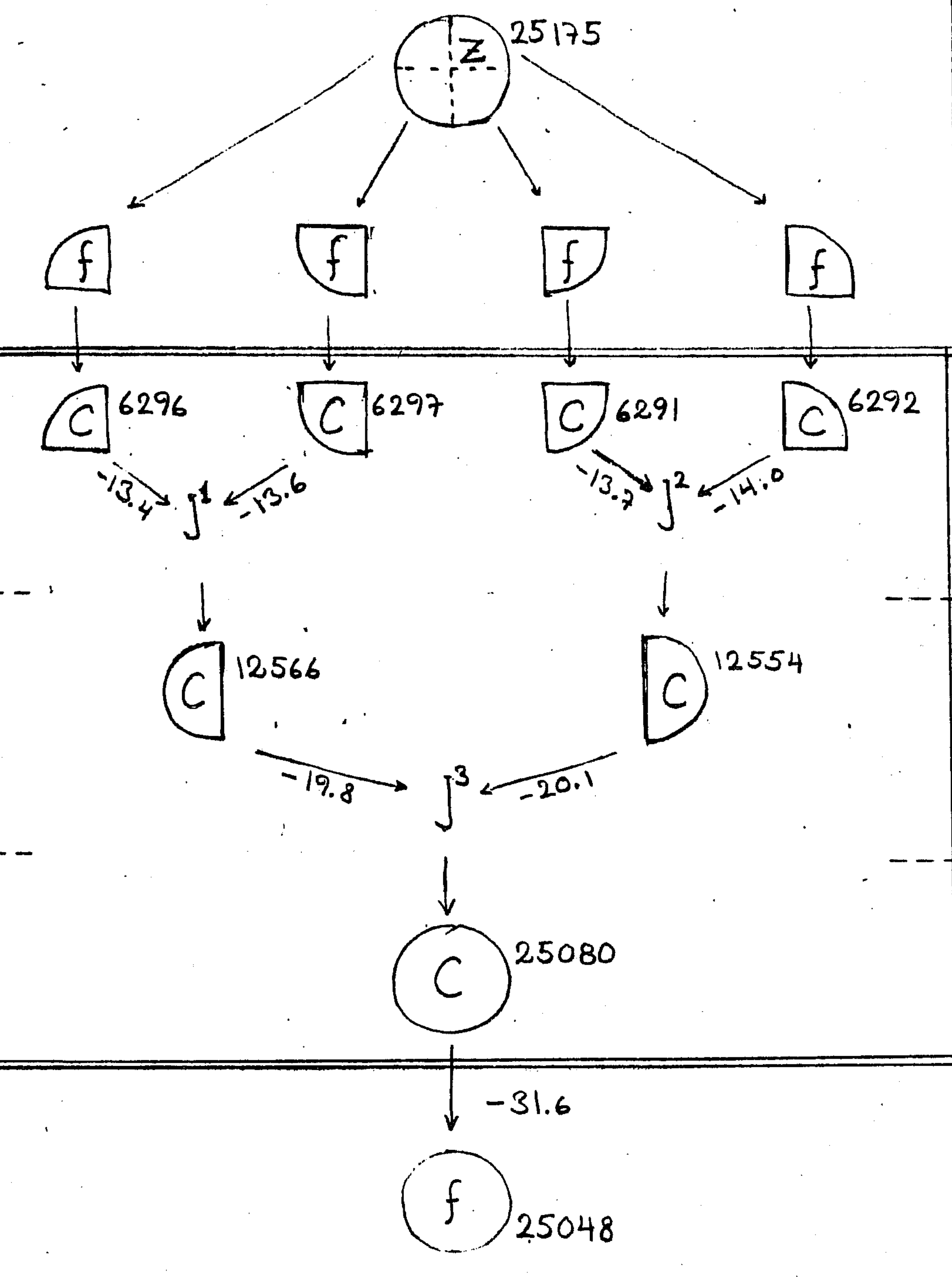
$$(A.7) \qquad \sigma_z^2(t) = \sum_{\mu=1}^{N} C_\mu^2(t)$$

and the variance of a whole sample is ($\sigma^2 = \overline{\sigma^2 (t)}$)

$$(A.8) \qquad \sigma_z^2 = \sum_{\mu=1}^{N} \overline{C_\mu^2}$$

In the dependent sample, using (A.4), (A.8) will be

$$(A.9) \qquad \sigma_z^2 = \sum_{\mu=1}^{N} \lambda_\mu .$$



Variance	degrees of freedom
25175	1404
25175	1404
-55	-704
25120	700
-40	-350
25080	350
-32	- 175
25048	175

Fig 1.

A schematic picture of the joining method used. The round grid is divided into four subgrids which are joined step by step. Letters indicate the form of information at each step: Z refers to the original analyses, f to the set of EOF's, C to the set of time dependent coefficients. J's refer to the sets of functions used in joining operations. In the picture and in the left number column positive numbers are the variances explained, negative numbers are the variances lost in joinings. In the righmost column numbers indicate the maximum degrees of freedom. 25175 m^2 is the original variance given in 1404 gridpoints, 25048 m^2 is the variance explained by 175 EOF's.

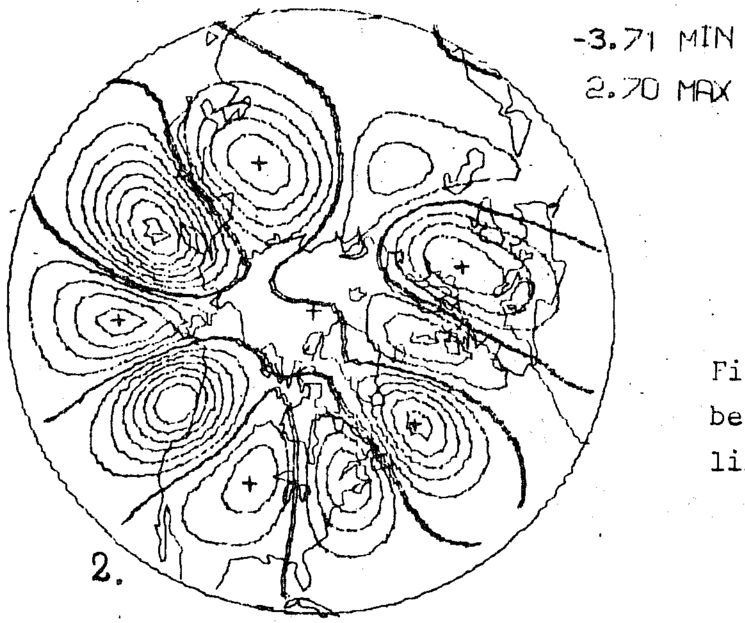


Fig 2. Component f_{14}, spacing between isolines is 0.5, thick line shows the zero isoline.

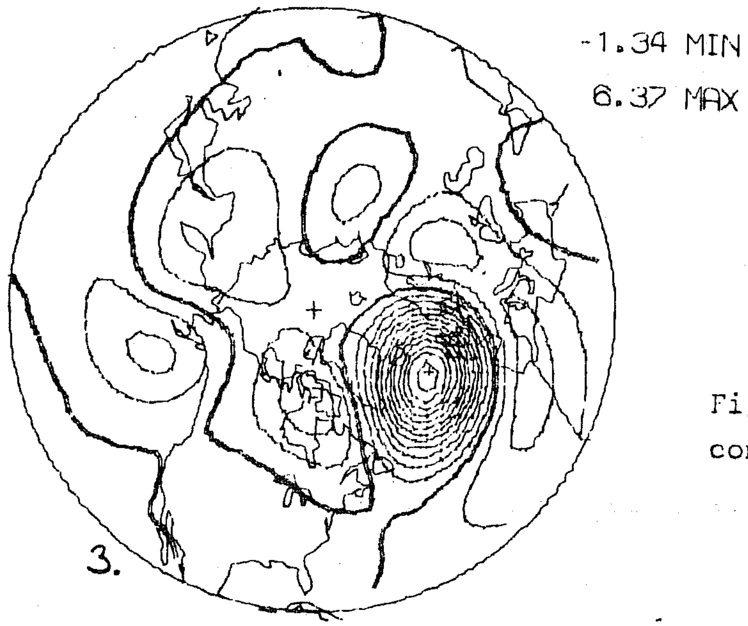


Fig 3. As fig 2 but for component f_8.

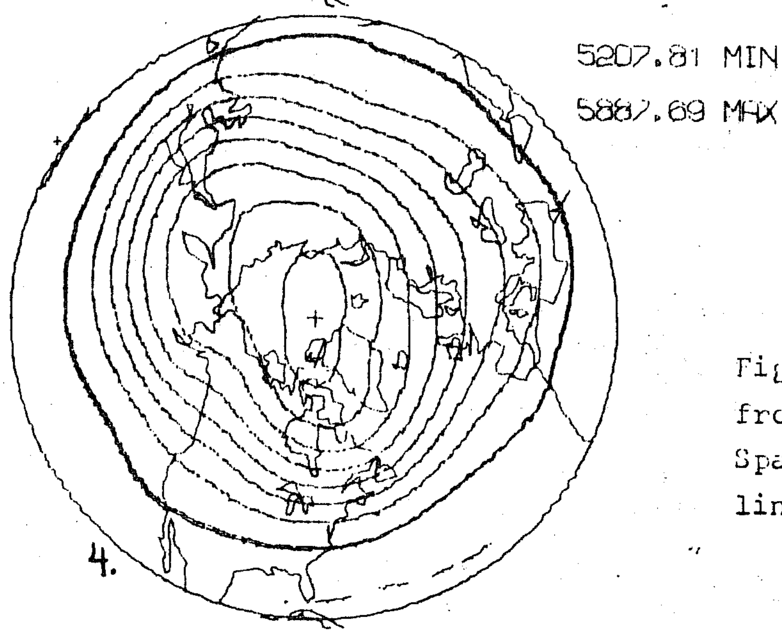
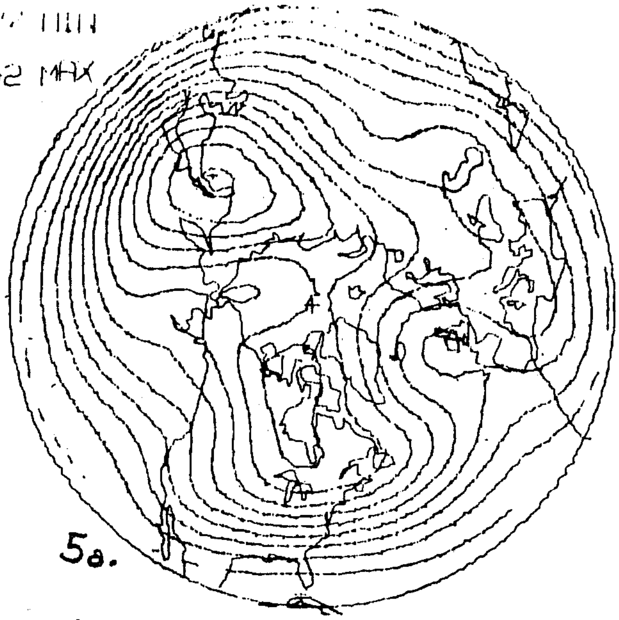


Fig 4. Mean field zm computed from twenty years of data. Spacing is 80m and the thick line stands for 5800m.



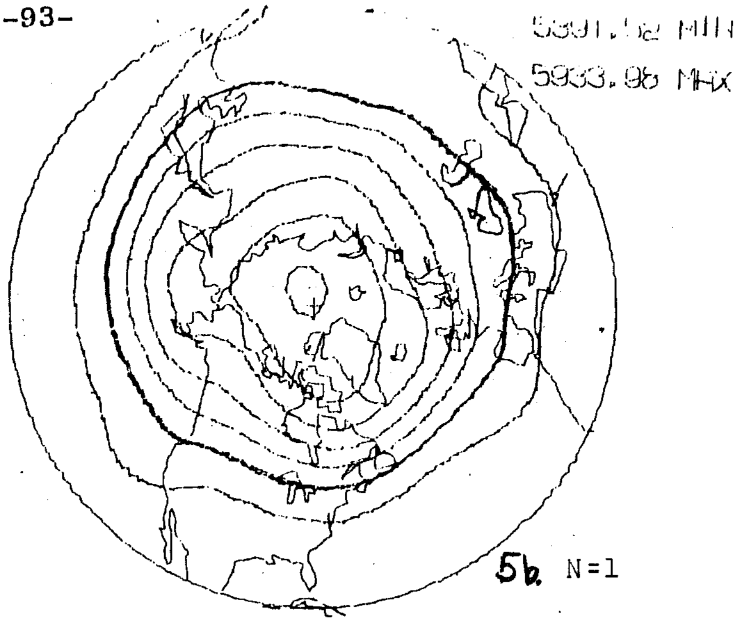
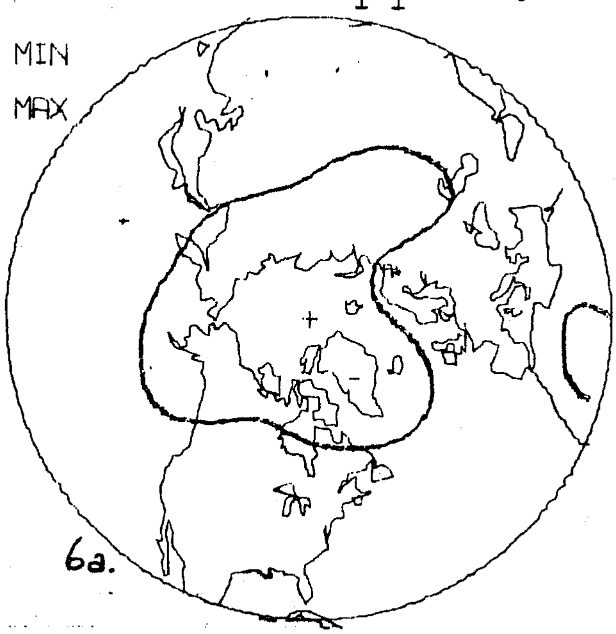


Fig 5a. Effect of component f_1 in (A.1) on Sep. 6th, 1968, weighted by C_1=143m. Spacing 20m, thic lines 0m.
Fig 5b. Sum $zm+C_1 f_1$ on Sep. 6th, 1968. Spacing 80m, thick 5800m.



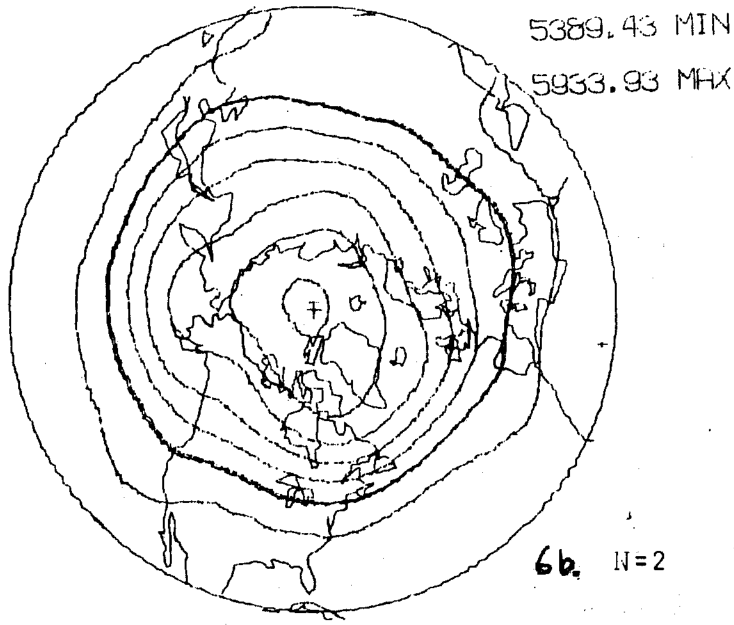
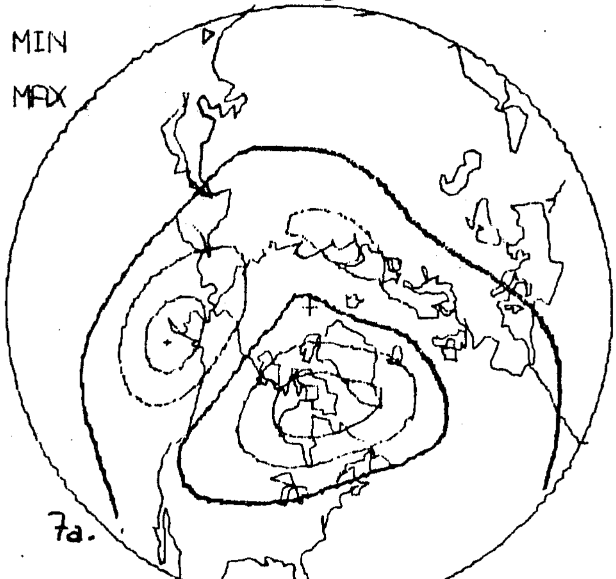


Fig 6a. As fig 5a but for f_2, C_2=1m.
Fig 6b. As fig 5b but for N=2 of sum $zm+\sum_{\mu=1}^{N} C_\mu f_\mu$.



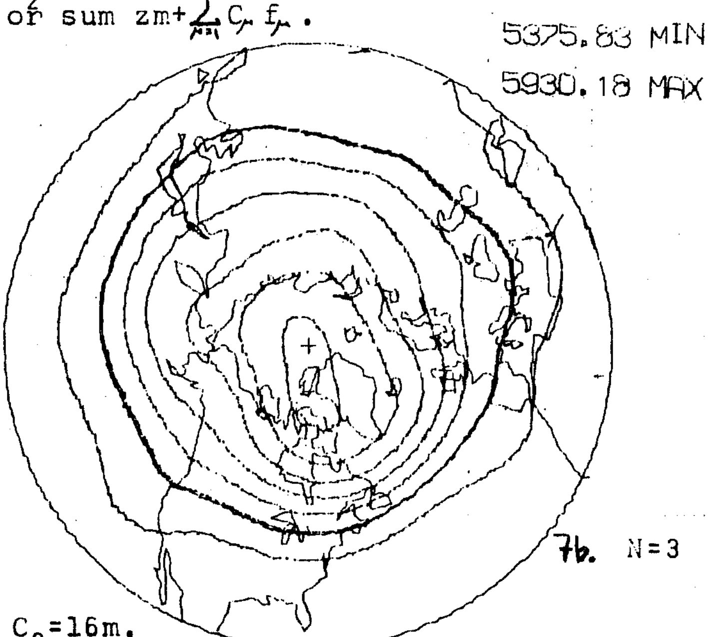
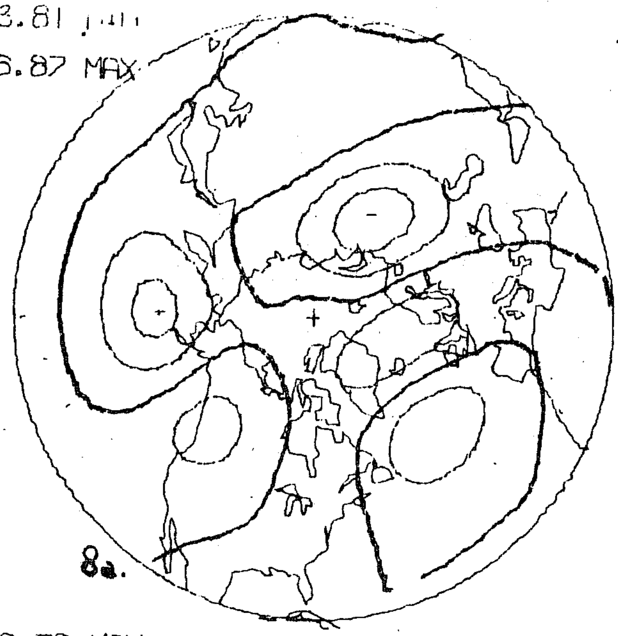
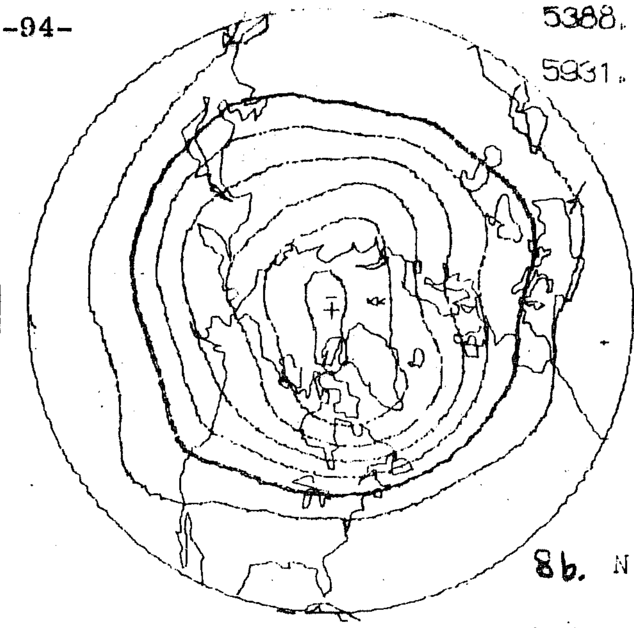
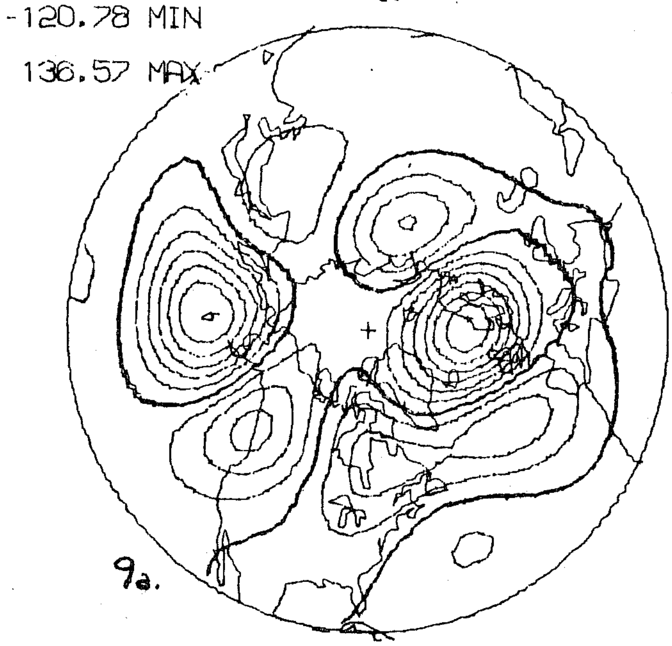
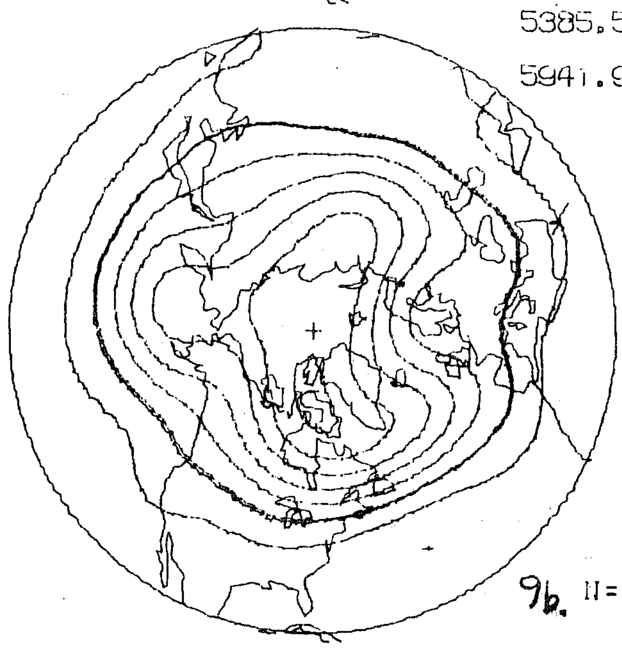


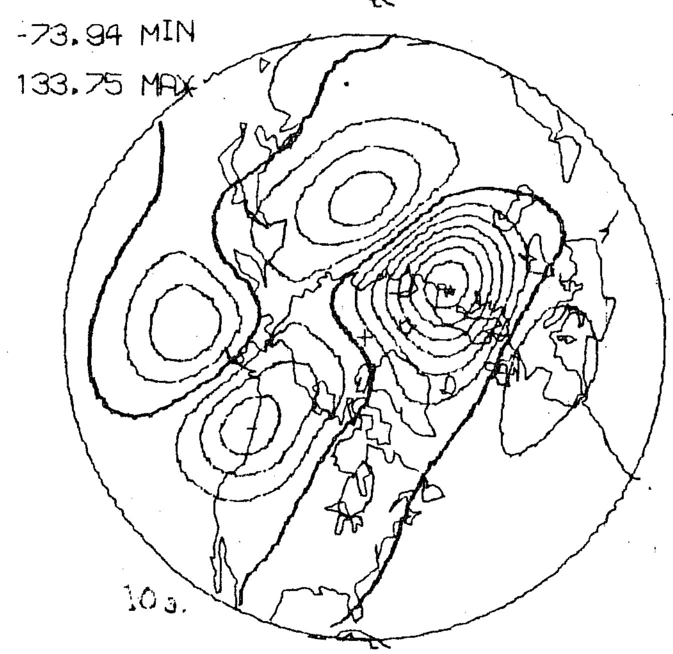
Fig 7a. As fig 5a but for f_3, C_3=16m.
Fig 7b. As fig 6b but for N=3.

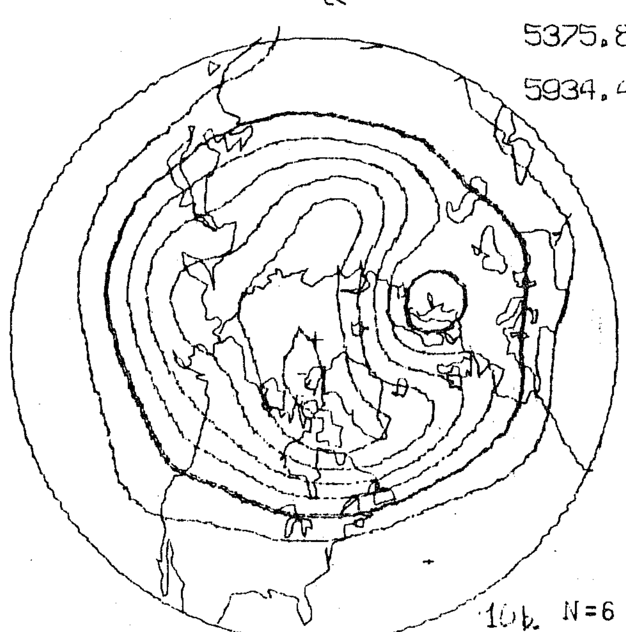






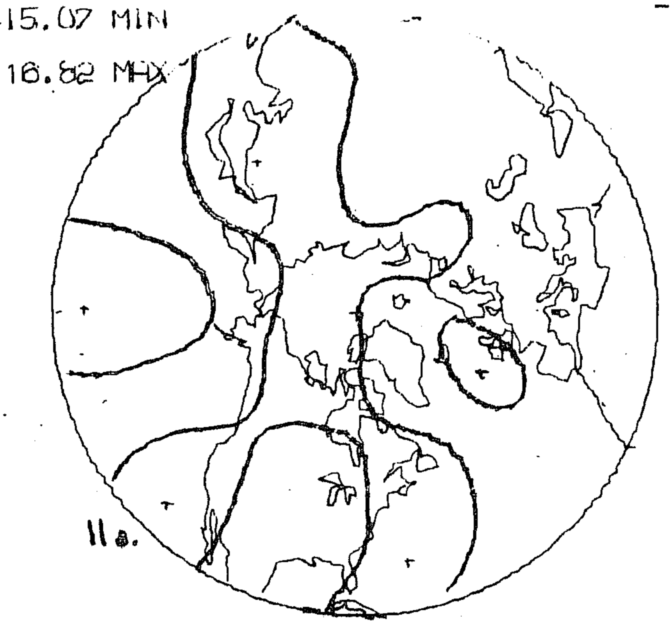


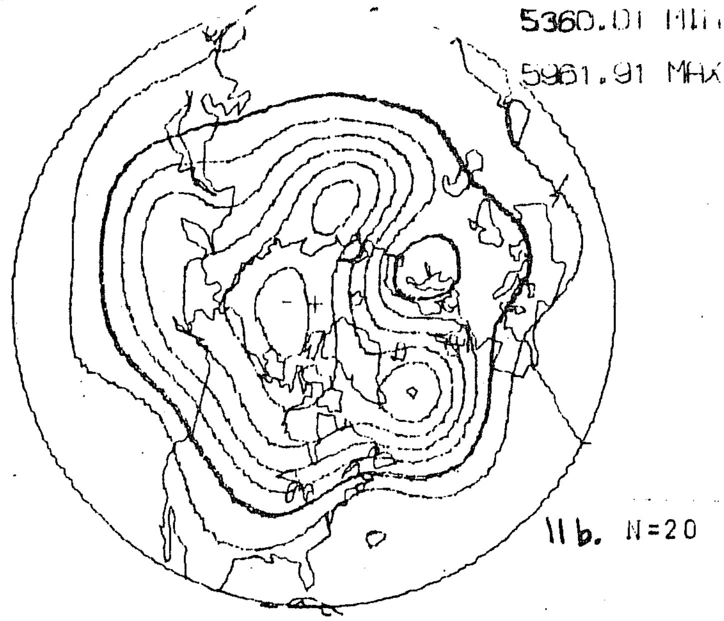


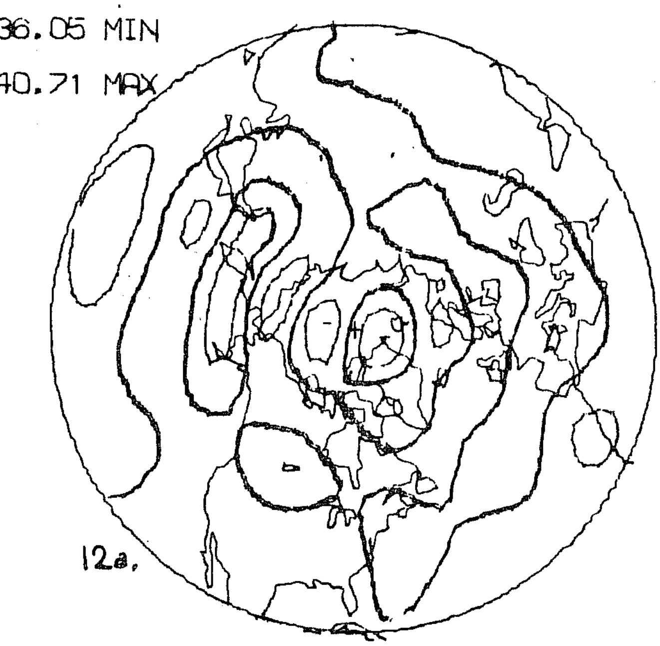


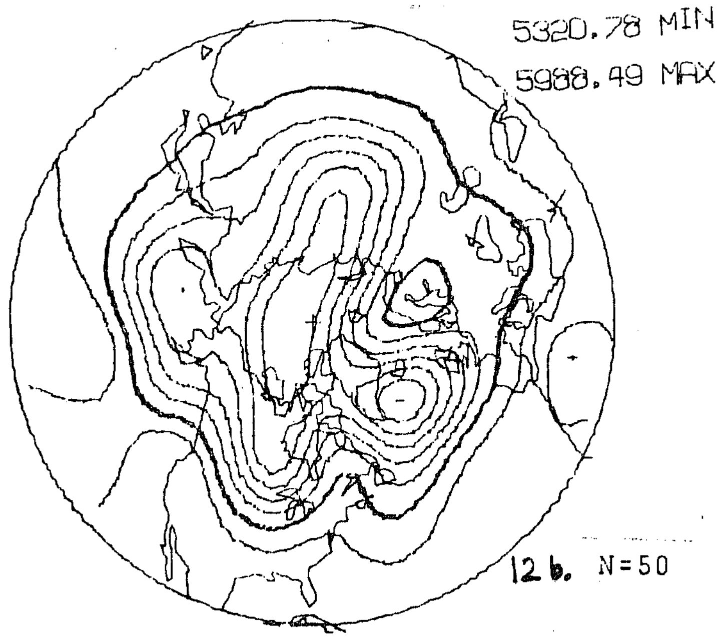
Figs 8a-10a. As fig 5a but for f_4 (C_4=15m), f_5 (C_5=34m) and f_6 (C_6=32m).

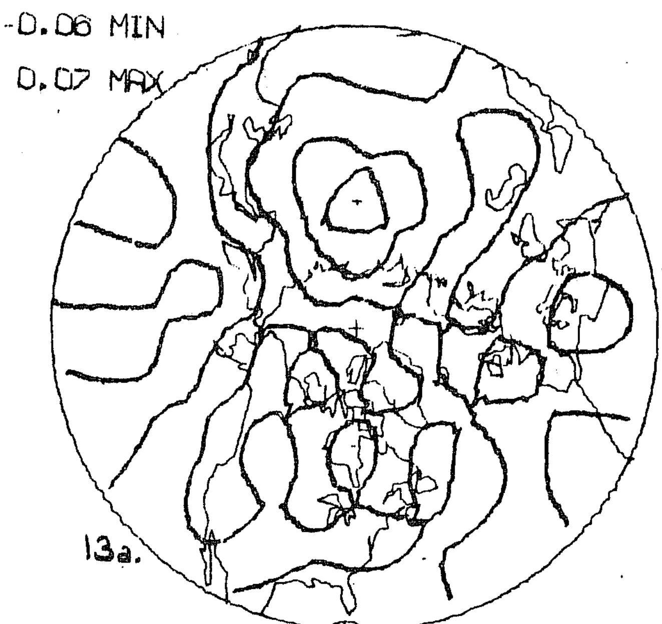
Figs 8b-10b. As fig 6b but for N = 4, 5 and 6.

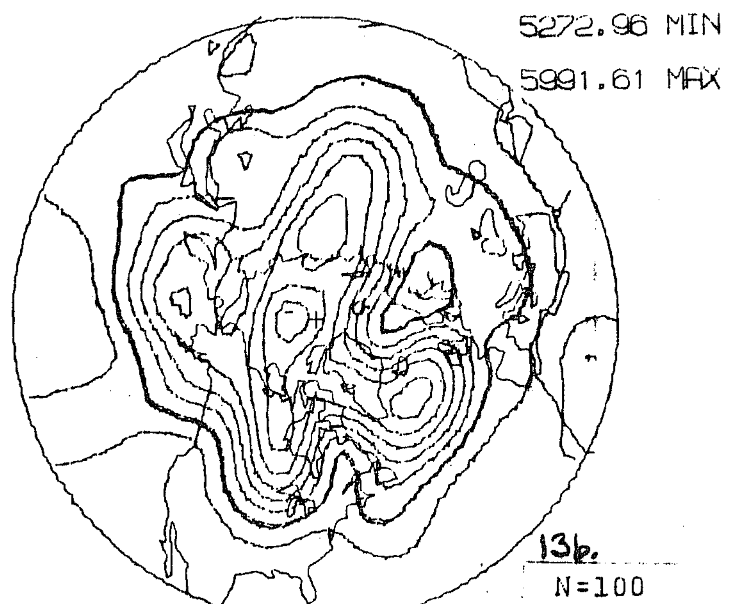












Figs 11a-13a. As fig 5a but for f_{20} (C_{20}=5m), f_{50} (C_{50}=-12m) and f_{100} (C_{100}=0.02).

Figs 11b-13b. As fig 6b but for N = 20, 50 and 100.

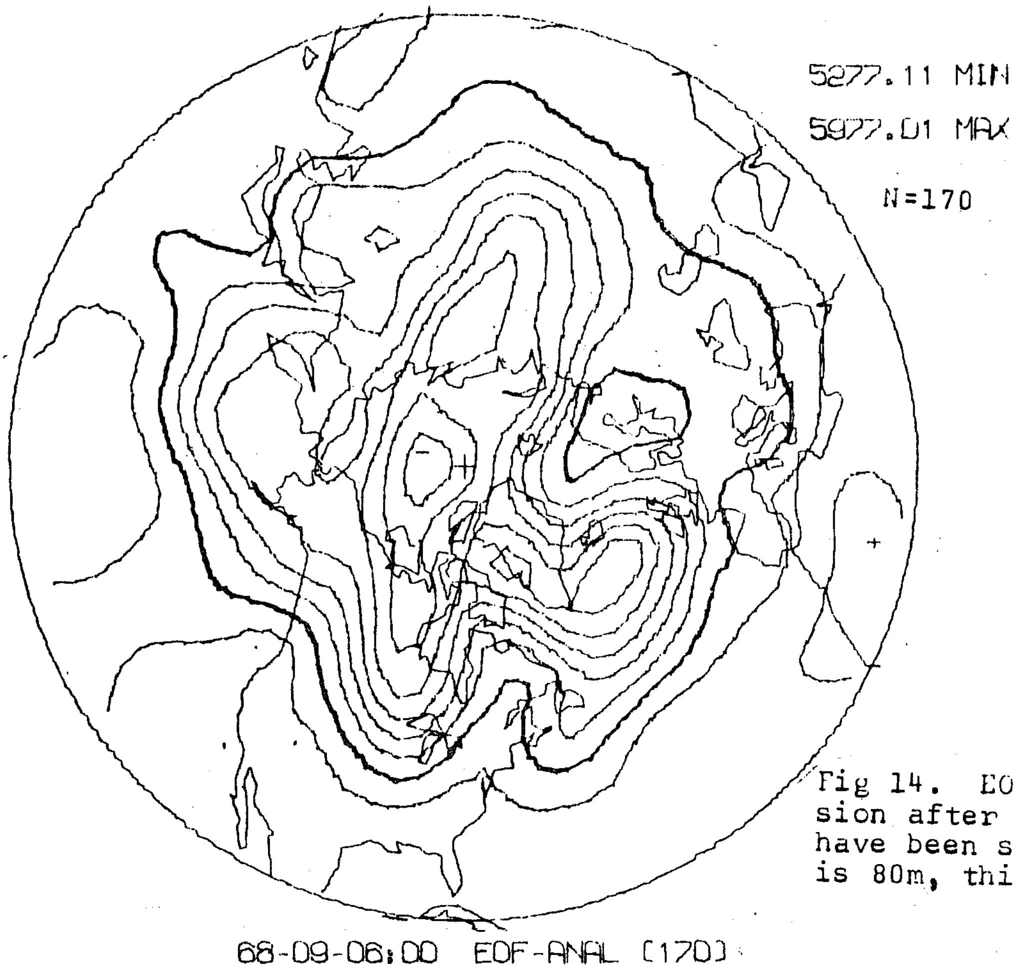


Fig 14. EOF-series expansion after 170 components have been summed. Spacing is 80m, thick line 5800m.

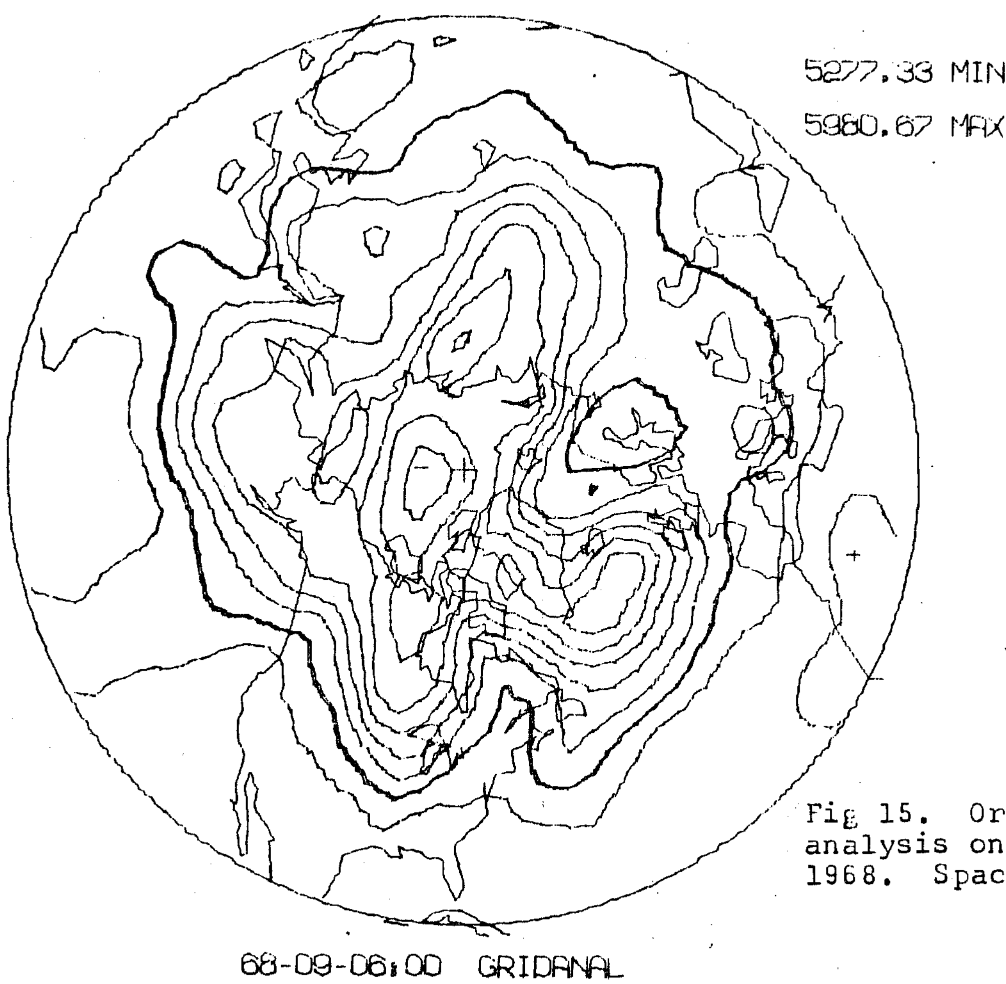


Fig 15. Original grid analysis on Sep. 6th, 1968. Spacing as in fig 14.