KW

MET O 11 TECHNICAL NOTE NO. 156

THE GMDH: A METHOD FOR PREDICTING ERRORS IN THE FORECAST MODEL

by

F. Kates

Meteorological Office
Forecasting Research Branch
London Road
Bracknell
Berkshire
United Kingdom

March 1982

1. Introduction

In the absence of a perfect model, any forecast will contain errors. These errors can be broken into two parts. The first is random and is by definition unpredictable. The second part is systematic, and it is this which we are trying to predict.

The method used to predict the errors is the GMDH - Group Method of Data Handling. This is a statistical prediction algorithm and was first proposed by A G Ivakhnenko (1968) and is explained below. In Met O 11 it has been applied to predicting errors in the Octagon 72 hr Forecast 500 mb height field, but the algorithm has been used extensively in other areas. For example in Japan it has been used to predict air pollution levels (Morita et al (1977), Tamura and Kendo (1977)), and river flow (Ikeda et al (1976)). In America it has been used to model the US economy (Scott and Hutchinson (1976)) and to forecast annual crop yeild from meteorological data (Mehra (1977)). In the USSR it has been used to model inflation processes in the British economy (Parks et al (1974)), and to predict average monthly sums of effective temperatures in Central Asia (Vysotskii and Yunusov (1977)).

2. The data

For every grid point in the Octagon field (which consists of 3037 points) there is a time series of forecast height values and a series of verifying initializations. The time series consists of seven weeks of data collected from the 0Z and 12Z forecast runs. Each point in a time series is known as an epoch. Thus the first forecast in the series is epoch 1, the second epoch 2, and so on.

The error in the forecast, called the 'actual' error to distinguish it from the estimated error, is defined to be the difference between the forecast and the verifying initialization. Thus:

Actual Error = Forecast value - Verifying Initialization Value

In most of the work done, the actual error is taken from the Octagon 72 hr forecast. The values of the 72 hr forecast height are then used as predictors for this error. Other values have been used as predictors, for example the values of the 48 hr forecast, but those results are not presented here.

For ease of identification, each gridpoint is numbered, from 1 to 3037. See Fig 1.

3. The Algorithm

The basic algorithm is described by Dixon and Purvis (1980) and consists of 5 steps.

(1) Divide the data into a training sequence and a testing sequence.

(2) Select N predictors (X_1, X_2, .. X_N)

(3) Combine the predictors two at a time to form the $N(N-1)/2$ quadratic polynomials:

$$Z_k = a_{k0} + a_{k1} X_i + a_{k2} X_j + a_{k3} X_i^2 + a_{k4} X_i X_j + a_{k5} X_j^2$$

$$i, j = 1, 2, \cdots N, \; i \neq j$$

$$k = 1, 2, \cdots N(N-1)/2$$

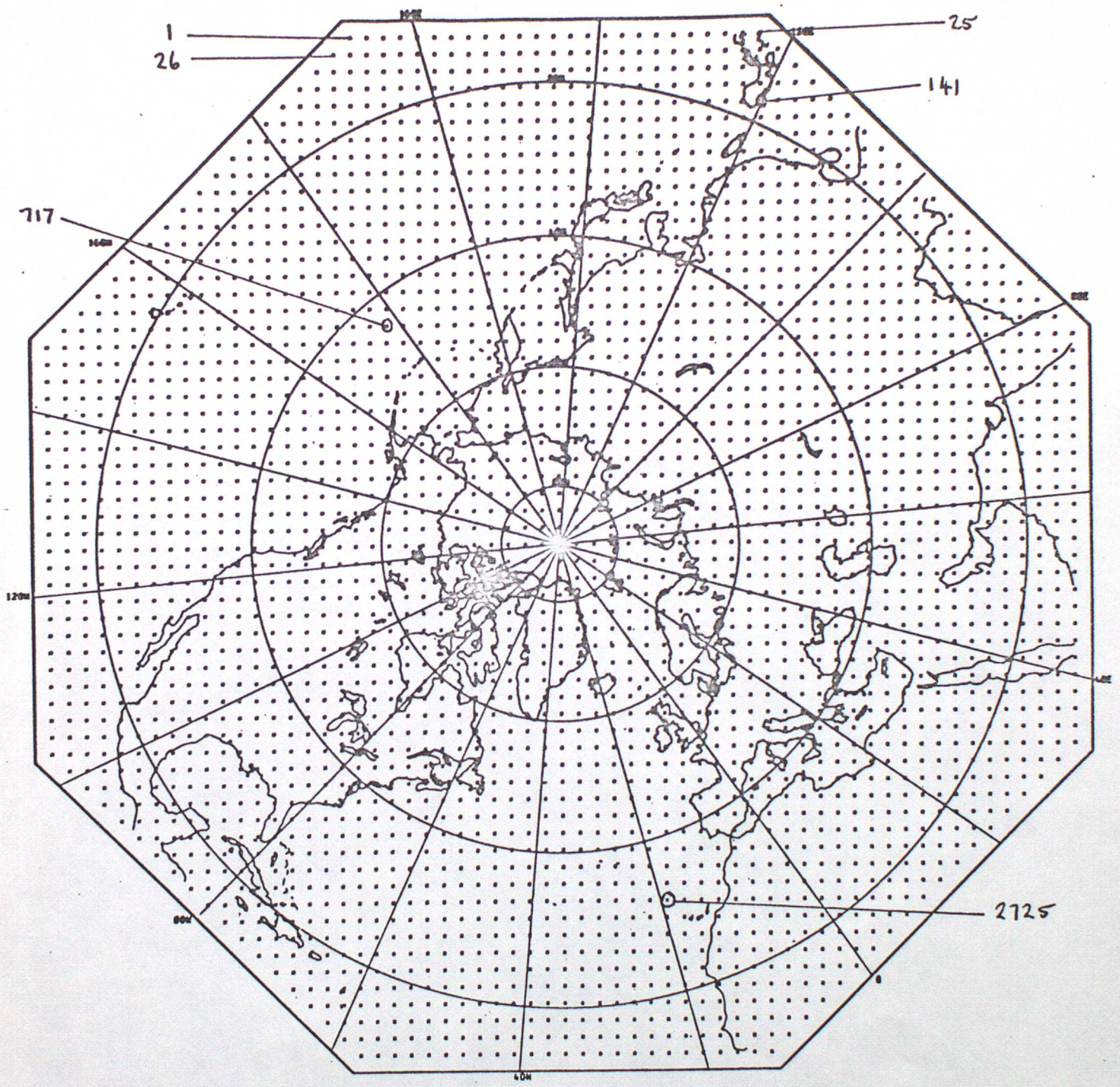


FIG 1 : THE OCTAGON FIELD AND THE GRID POINT NUMBERING SYSTEM.

and in each case determine the coefficients by least squares fitting of the training set data.

(4) Evaluate these polynomials over the testing sequence, and select the m polynomials which give the m smallest mean square errors. These are the new predictors.

(5) Go back to (3) and cycle through (3), (4) and (5) using the new predictors. Keep on looping round until the best polynomial found at (4) gives a mean square error larger than that given by the best polynomial at the previous passage through (4). The best polynomial from the previous stage is then taken as the optimum predictor.

Note that Z_k evaluated at any epoch gives an estimate for the error at that epoch. This represents the main difference between the GMDH and the more traditional least squares approach, in that the GMDH uses the prediction process to select the model whereas the least squares approach does not. Most of the least squares theory is concerned with getting a polynomial which fits the data well, but no attempt at actually predicting anything is made until the very end of the process. The prediction is almost incidental to the whole method. In the GMDH however, the prediction process forms a fundamental part of selecting the model.

The basic algorithm is shown in flow diagram form in Fig 2.. To understand this in more detail, consider the following example. To simplify the problem, choose 4 input predictors (the real problem has 9). This gives 6 possible quadratic polynomials at each stage:

$$z_1 = a_{11} + a_{12}x_1 + a_{13}x_2 + a_{14}x_1^2 + a_{15}x_1x_2 + a_{16}x_2^2 \quad (1)$$

$$z_2 = a_{21} + a_{22}x_1 + a_{23}x_3 + a_{24}x_1^2 + a_{25}x_1x_3 + a_{26}x_3^2 \quad (2)$$

$$z_3 = a_{31} + a_{32}x_1 + a_{33}x_4 + a_{34}x_1^2 + a_{35}x_1x_4 + a_{36}x_4^2 \quad (3)$$

$$z_4 = a_{41} + a_{42}x_2 + a_{43}x_3 + a_{44}x_2^2 + a_{45}x_2x_3 + a_{46}x_3^2 \quad (4)$$

$$z_5 = a_{51} + a_{52}x_2 + a_{53}x_4 + a_{54}x_2^2 + a_{55}x_2x_4 + a_{56}x_4^2 \quad (5)$$

$$z_6 = a_{61} + a_{62}x_3 + a_{63}x_4 + a_{64}x_3^2 + a_{65}x_3x_4 + a_{66}x_4^2 \quad (6)$$

The polynomial coefficients a_{ij} are calculated over the training sequence. The basic GMDH algorithm does this using the least squares process, but the version now used in Met O 11 uses a 'quick' formulation developed by Mr R Dixon. In addition to being quicker, the Dixon formulation avoids ill-conditioning which is a recurring problem in more traditional approaches. For example in the Box-Jenkins process (Box and Jenkins (1977)) a 1% failure rate is regarded as acceptable. If applied to the Octagon, this would mean that 30 points were corrupted, which is completely unacceptable. Using the Dixon formulation, there is no risk of ill-conditioning.

Having found the coefficients in the 6 polynomials, z_1 to z_6 are calculated using values of x_i taken from the testing sequence. For any particular epoch in the testing sequence, z_i is an estimate of the actual error. By comparing the estimates with the actual errors for every epoch in the training sequence, the best 4 prediction polynomials can be selected. Suppose for simplicity that the best 4 are z_1, z_2, z_3, z_4. Then we form the 6 quadratic polynomials in the same manner as before, using the z_i as predictors:

$$y_1 = b_{11} + b_{12}z_1 + b_{13}z_2 + b_{14}z_1^2 + b_{15}z_1z_2 + b_{16}z_2^2 \quad (7)$$

$$y_2 = b_{21} + b_{22}z_1 + b_{23}z_3 + b_{24}z_1^2 + b_{25}z_1z_3 + b_{26}z_3^2 \quad (8)$$

$$y_3 = b_{31} + b_{32}z_1 + b_{33}z_4 + b_{34}z_1^2 + b_{35}z_1z_4 + b_{36}z_4^2 \quad (9)$$

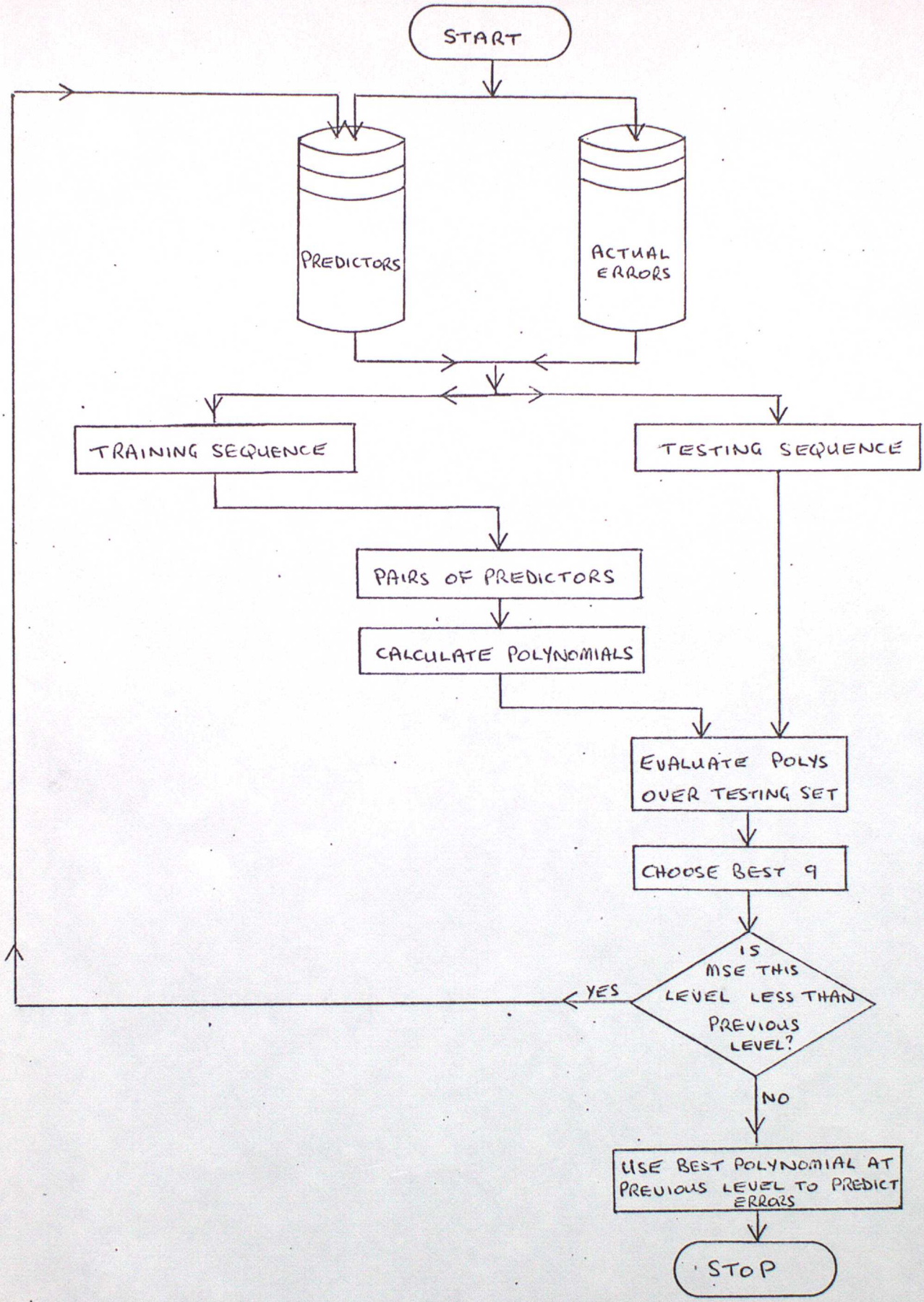


Fig 2. Flowchart for GMDH

$$Y_4 = b_{41} + b_{42} z_2 + b_{43} z_3 + b_{44} z_2^2 + b_{45} z_2 z_3 + b_{46} z_3^2 \qquad (10)$$

$$Y_5 = b_{51} + b_{52} z_2 + b_{53} z_4 + b_{54} z_2^2 + b_{55} z_2 z_4 + b_{56} z_4^2 \qquad (11)$$

$$Y_6 = b_{61} + b_{62} z_3 + b_{63} z_4 + b_{64} z_3^2 + b_{65} z_3 z_4 + b_{66} z_4^2 \qquad (12)$$

Again the coefficients are calculated using the training sequence and the Y_i's evaluated over the testing sequence. The Y_i's are then compared with the actual errors, the four best are found .... and so on. If we write out Y_1 explicitly, we get:

$$\begin{aligned} Y_1 = {} & b_{11} + b_{12}\left(a_{11} + a_{12}X_1 + a_{13}X_2 + a_{14}X_1^2 + a_{15}X_1X_2 + a_{16}X_2^2\right) \\ & + b_{13}\left(a_{21} + a_{22}X_1 + a_{23}X_3 + a_{24}X_1^2 + a_{25}X_1X_3 + a_{26}X_3^2\right) \\ & + b_{14}\left(a_{11} + a_{12}X_1 + a_{13}X_2 + a_{14}X_1^2 + a_{15}X_1X_2 + a_{16}X_2^2\right)^2 \\ & + b_{15}\Big(\left(a_{11} + a_{12}X_1 + a_{13}X_2 + a_{14}X_1^2 + a_{15}X_1X_2 + a_{16}X_2^2\right) \\ & \qquad \cdot\left(a_{21} + a_{22}X_1 + a_{23}X_3 + a_{24}X_1^2 + a_{25}X_1X_3 + a_{26}X_3^2\right)\Big) \\ & + b_{16}\left(a_{21} + a_{22}X_1 + a_{23}X_3 + a_{24}X_1^2 + a_{25}X_1X_3 + a_{26}X_3^2\right)^2 \end{aligned} \qquad (13)$$

Clearly this is getting much too cumbersome to program directly, and the complexity increases if further cycles are considered. The Japanese and Russians get around this problem to some extent by beginning with a very simple polynomial, say:

$$z_1 = a_1 + a_2 X_1$$

and then building up the degree of the polynomial at each stage, (see A G Ivakhnenko (1978)) However this approach is not feasible on a problem the size of the envisaged Met O 11 application.

The solution to the problem is found in the algebraic structure of the polynomial for Y_1 (equ 13). See Fig 3.

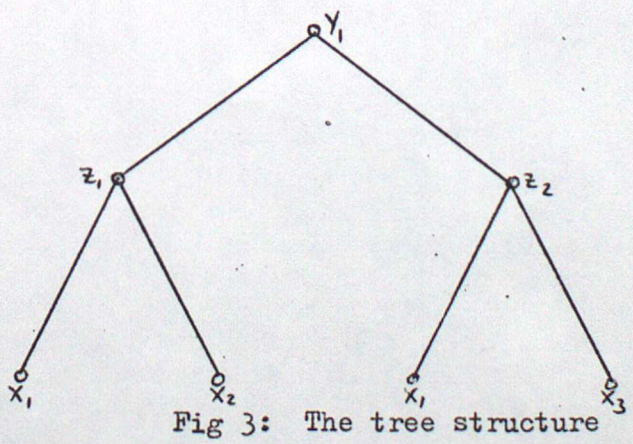


Fig 3: The tree structure

This can be split into three parts, see Fig 4.

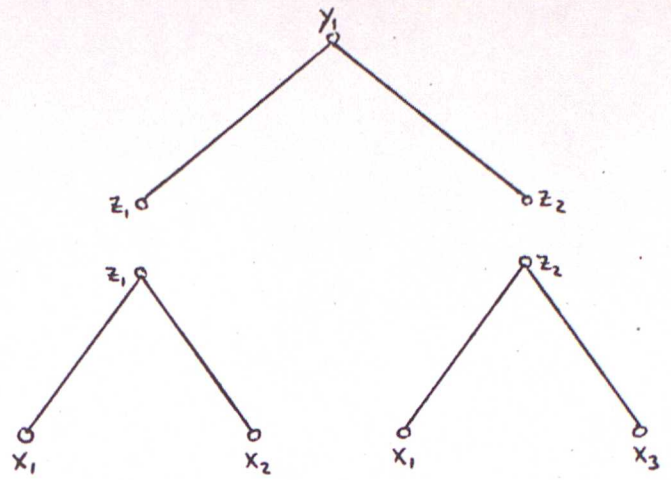


Fig 4: The tree structure split into identical parts.

The three parts are structurally identical. Proceeding in this manner, the problem becomes much simpler. The value of z_1 can be calculated given the coefficient vector $(a_{11}, a_{12}, a_{13}, a_{14}, a_{15}, a_{16})$ so that at any stage the algebraically complex z_i can be reduced to a single number.

The z_i in the above example are known as tree level 1 results. Because they are used to calculate y_i they are also known as 'intermediate' results. y_i are tree level 2 results.

The usual analogy given in the literature (see Scott and Hutchinson (1976)) is the process used by a horticulturist in selectively breeding a species of plant to obtain a hybrid variety having certain desired properties. The horticulturist takes a group of flowers and cross-fertilizes each with every other flower in the group. He collects the resulting seeds and plants them. When these have grown, he examines the new generation of flowers and discards those which are not an improvement in the species. This is the equivalent of one layer in the GMDH algorithm. This process is repeated over and over. Provided he has been discarding plants properly at each generation, the horticulturist should see the resulting generations tending to have more and more of the properties he requires. Finally a single flower results, which is the best he can obtain with the number of generations he has grown.

In the Met O 11 application, the 'flowers' are the predictors used at each level. These are 'cross-fertilized' or combined in pairs to form 'seeds' or polynomials. Those which have the desired property of giving good estimates of the error are then used as the new 'flowers' or predictors for the next generation.

4. Preselection

In the paper by Ikeda et al (1976) the idea of preselecting the predictors was suggested. In the Met O 11 application, for every grid point and every epoch there are 117 possible predictors. This is because the predictors are constrained to be taken from no more than 12 epochs (144 hours) in the past and can be taken from the eight surrounding points as well as the target point. See fig 5. From these 117, we have to select 9, and this can be done in two ways. The simplest way is to make an a priori choice of predictors. Thus predictors may be chosen to be taken from 4 epochs (48 hours) back. This would be fixed for every point in the Octagon field.

The more complicated method of selecting the 9 predictors allows a different 9 to be specified for each point. For a given point, the 117 predictors are used individually to make an estimate of the error. On the basis of this estimate, the 9 best predictors are chosen. It is assumed that if a predictor gives accurate estimates when used individually, then it will give accurate estimates when used in combinations with other predictors, though this may not always be

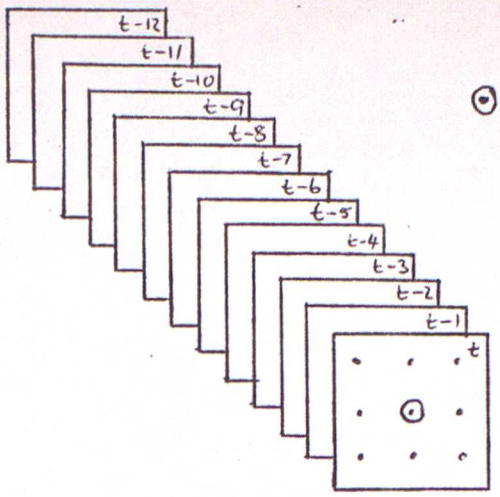


Fig 5: Configuration of the 117 possible predictors.

true. In addition to giving good estimates, it is also possible to select predictors on the basis of their correlation with each other, so that the 9 eventually selected are not too highly correlated.

5. Results

The programs as written allow up to tree level 6 results to be calculated. However, most of the recent work has been limited to tree level 1, as it was found that once a bad combination of predictors has been selected at one level, the tree structure rapidly becomes contaminated through subsequent levels. All of the results given in this paper are tree level 1 results.

The results from a typical run are given in Figs 7,8,9 and 10 for epoch 87. This epoch is significant in that it is the first epoch for which a forecast was issued based on data not included in the training or testing sequences. See Fig 6. Epoch 80 is the last epoch in the testing sequence and is the last epoch for which the error would be known in an operational run.

Epoch	77	78	79	80	81	82	83	84	85	86	87
72 hr Fcst							→	→	→	→	→
Init	→	→	→	→							
Error											

Fig 6: The significance of epoch 87

Fig 7 gives the 500 mb height field as forecast by the Octagon for 0Z 4/2/78, and Fig 8 gives the verifying initialization for the same time. Fig 9 gives the error, that is the difference between the forecast and the initialization. Points to note are that the forecast over the USA was somewhat in error, the position and intensity of the low near Alaska is incorrect and the ridge west of the UK is not forecast sharp enough. Also, the low south of Greenland is not forecast at all.

Fig 10 gives the forecast after the GMDH estimated error has been removed, and Fig 11 gives the residual errors that has not been predicted. The ridge over the USA has been corrected to a large degree but no obvious improvement has been made to the low near Alaska. The ridge west of the UK has not been altered very much. There is a suggestion of a trough at 45°W which is possibly an attempt at producing the low which should have been south of Greenland. Overall, the error has been reduced by 9.4%, and on a diagram of this scale it is therefore not surprising that no great improvements can be seen.

Figures 12, 13 and 14 give the results for 3 single points, the locations of which are shown in Fig 1. In each, the actual error is plotted for epochs 39 to 90, together with the estimated error for epochs 80 to 89. Below each graph, three means are given. The first, labelled "mean (actual) for whole period" is the average value of the actual error over epochs 39 to 90. The second is the mean of the actual error taken over epochs 80 to 89, and the third is the mean of the estimated error taken over epochs 80 to 89. Note that the vertical scale is different in each case.

For point 141 (Fig 12), the actual error seems to have a very organized pattern, and consequently it is not surprising that the GMDH estimate accounts for 75% of the error. However, the actual error for point 717 (Fig 13) shows no such organization and yet the GMDH estimate accounts for 79.3% of the error.

Fig 14 shows the actual and estimated errors for point 2725. The actual error, while not varying as wildly as at point 717, does not show any systematic pattern. The GMDH actually increased the error by 225.1%.

6. Conclusions

The results obtained show overall a very modest reduction in the error of the Octagon 72 hr 500 mb height field. This is probably due to the fact that there is no large systematic error inherent in the forecast model. At certain points where there does seem to be a quasi-periodic error, the GMDH predicts it exceptionally well.

The results, coupled with the fact that the GMDH only uses a very short time series, would indicate that this is an ideal technique for predicting any systematic event, such as the diurnal change in temperature at any single point. Indeed, users of least squares methods in general should give this method very serious consideration.

7. Acknowledgements

The work presented in this paper was largely carried out by Mr R Dixon and Mr G W Purvis, sine qua non.

8. References

1. A.G. Ivakhnenko: "The group method of Data Handling, a Rival of the method of Stochastic Approximation". Soviet Automatic Control, Vol 4, No 3, 1968.

2. T. Morita, M. Konishi, K. Nose, Y. Asada and I. Yamamoto: "Use of GMDH for Estimation of Regional Air Quality" Environmental Systems Planning and Control Part 1, Kyoto, Japan, 1-5 Aug 1977.

3. H. Tamura and T. Kondo: "Revised GMDH Algorithm Using Self-Selection of Optimal Partial Polynomials and its Application to Large-Spatial Air Pollution Pattern Identification". Soc. Instrun. & Control Eng. (Japan), Vol. 13, No 4, Aug 1977.

4. S. Ikeda, S. Fujishige and Y. Sawaragi: "Non-linear Prediction Model of River Flow by Self-Organization Method". International Journal of Systems Science, Vol. 7, No. 2, 1976.

5. D. E. Scott and C. E. Hutchinson: "The GMDH Algorithm - A technique for Economic Modelling". Modelling and Simulation, Vol. 7, No. 2, April 1976.

6. R. K. Mehra: "Group Method of Data Handling (GMDH): Review and Experience". IEEE Conf on Decision and Control, 1977.

7. P. Parks, A. G. Ivakhnenko, L.M. Boychuk and V. K. Svetalskii: "Self-Organization of a Model of the British Economy by Balance-Of-Variables Criterion for control with Prediction Optimization". Automatika (USSR), Vol 7, No 6, 1974.

8. V. N. Vysotskiy and N. I. Yunusov: "Improving the Noise Immunity of a GMDH Algorithm Used for Finding a Harmonic Trend with Nonmultiple Frequencies". Soviet Automatic Control, Vol 10, No 5, 1977.

9. R. Dixon and G. W. Purvis: "Modification of Numerical Forecast Height Errors by the GMDH". WMO Collect. Pap. Present. WMO Symp. Probab. Stat. Meth. Weath. Forecast. Nice 8-12 Sept 1980.

10. G. E. P. Box and G. M. Jenkins: "Time Series Analysis. Forecasting and Control". San Francisco (Holden-Day), 1970.

11. A. G. Ivakhnento: "The Group Method of Data Handling in Long-Range Forecasting". Technological Forecasting and Social change, Vol 12, 1978.

A large number of papers on the GMDH by A. G. Ivakhnenko and his associates are to be found in Soviet Automatic Control (USA).

Appendix: A GMDH Glossary

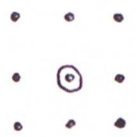
Actual Error: Forecast height minus the verifying initialization height.

Corrected Field: Octagon forecast minus GMDH estimated error.

Epoch: A 12 hour time step. Forecasts are issued every 12 hours. The first forecast in a time series is epoch 1, the second epoch 2 and so on.

Estimated Error: The GMDH prediction of the actual error.

Geographical Predictors: Predictors taken not just from the time series of a particular point but also from the 8 surrounding points.



GMDH: Group Method of Data Handling.

Intermediate Result: Any estimated error from a tree level other than the final one.

Preditand: The actual error. What the GMDH is trying to predict.

Predictor: Anything used to predict the actual error.

Preselection: The selection of 9 predictors for each point by an objective rather than a subjective method.

Residual: Actual error minus estimated error. This should be smaller than the actual error.

Testing sequence: That half of the time series which is used to evaluate the prediction polynomial. This is also known as the checking set in some of the literature.

Training sequence: That half of the time series which is used to calculate the coefficients for the prediction polynomial.

Tree Level: One cycle through the algorithm is known as a tree level because of the algebraic structure.

FIELD DERIVED FROM TIME SERIES EXTENDING FROM 23/12/77 TO 5 /2 /78.(EPOCHS 1 TO 90)

500MB
FIELD OF 72 HR FCST

THIS EPOCH NO=87

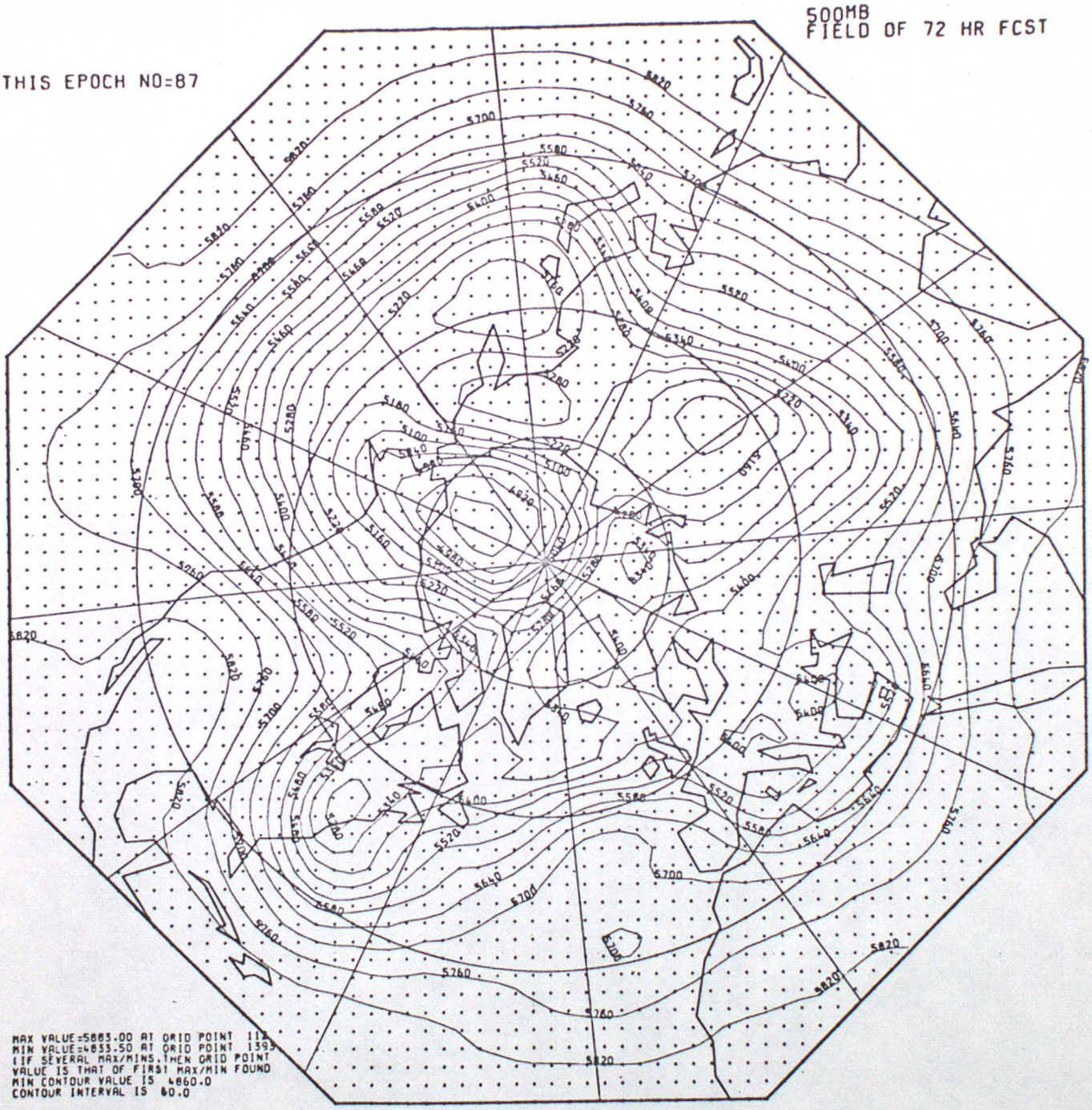


FIG. 7

FIELD DERIVED FROM TIME SERIES EXTENDING FROM 23/12/77 TO 5 /2 /78.(EPOCHS 1 TO 90)

500MB
FIELD OF 72HR VERIFYING INITS

THIS EPOCH NO=87

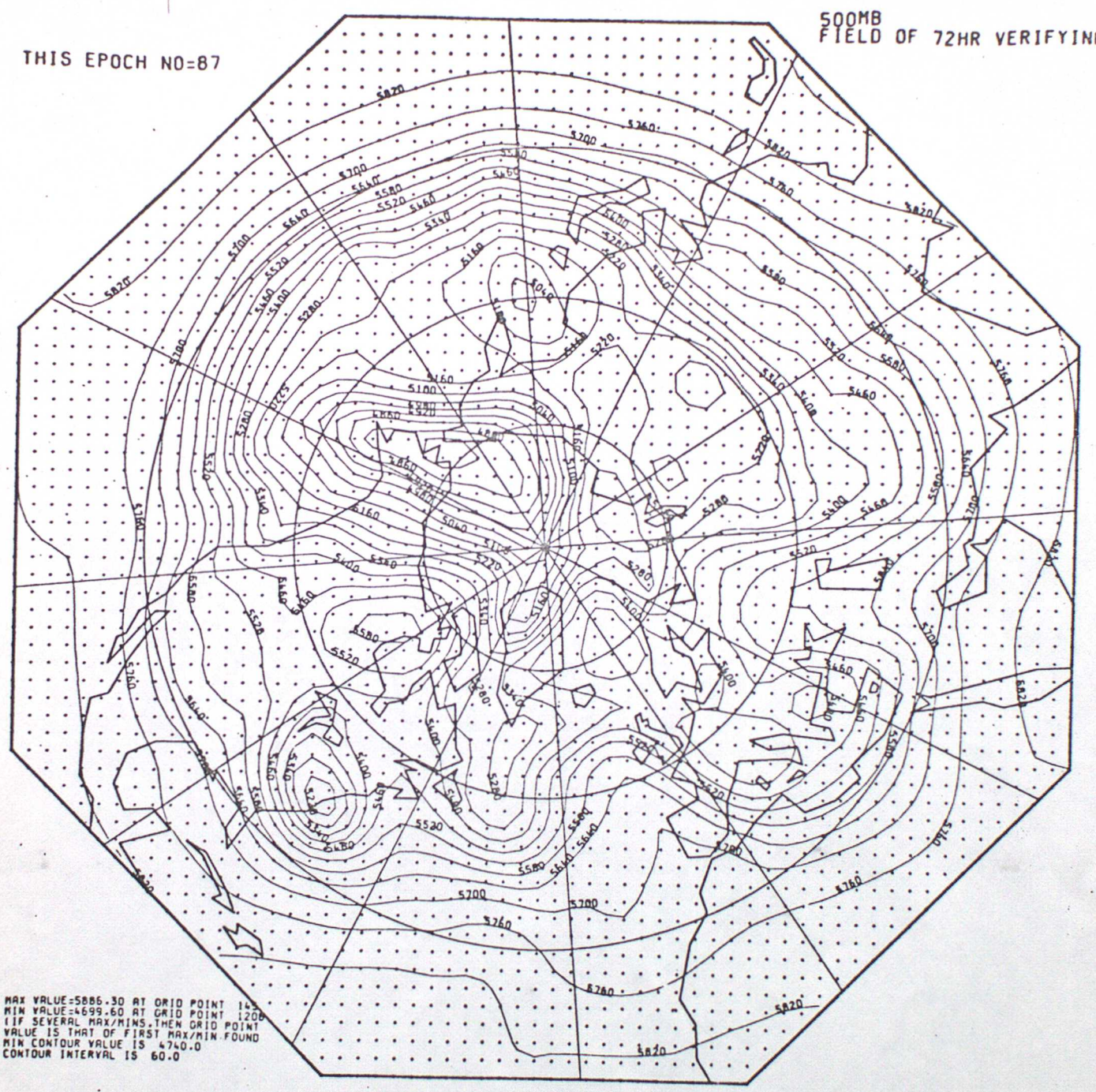


FIG. 8

FIELD DERIVED FROM TIME SERIES EXTENDING FROM 23/12/77 TO 5 /2 /78.(EPOCHS 1 TO 90)

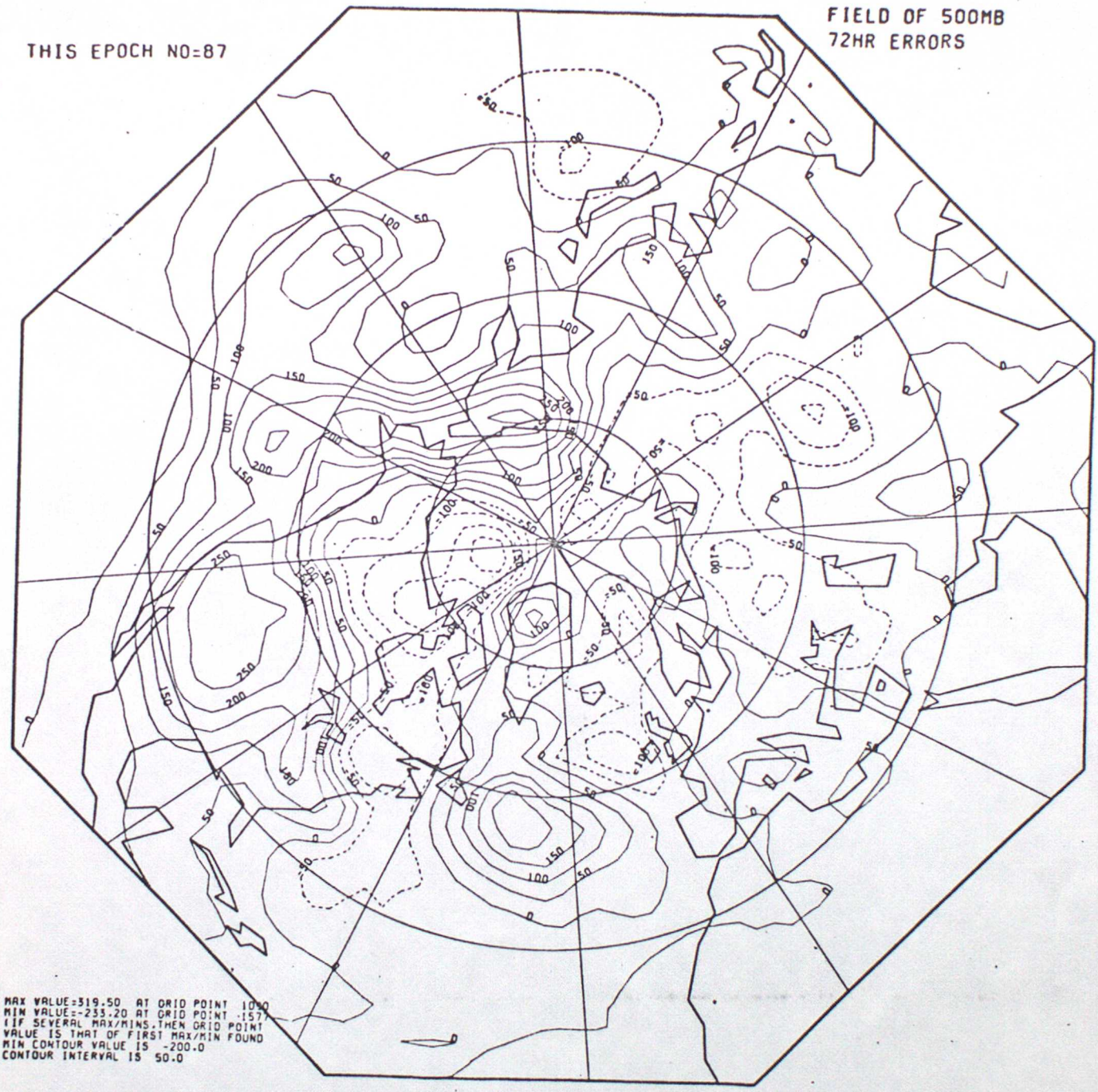


FIG. 9

FIELD DERIVED FROM TIME SERIES EXTENDING FROM 23/12/77 TO 5 /2 /78.(EPOCHS 1 TO 90)

500MB
FIELD OF 72 HR PREDICTIONS

THIS EPOCH NO=87

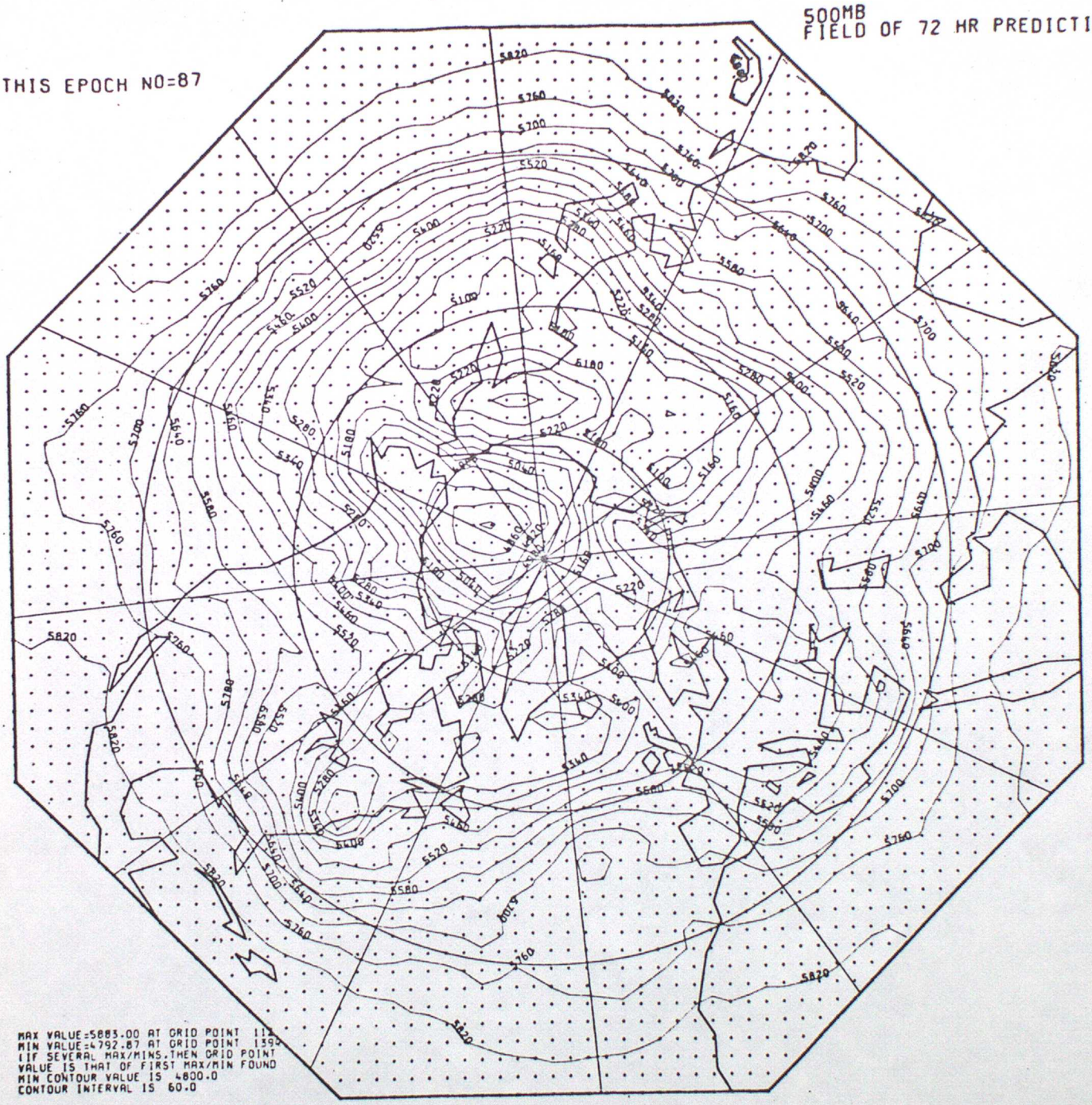


FIG. 10

FIELD DERIVED FROM TIME SERIES EXTENDING FROM 23/12/77 TO 5 /2 /78.(EPOCHS 1 TO 90)

500MB
FIELD OF 72 HR RESIDUALS

THIS EPOCH NO=87

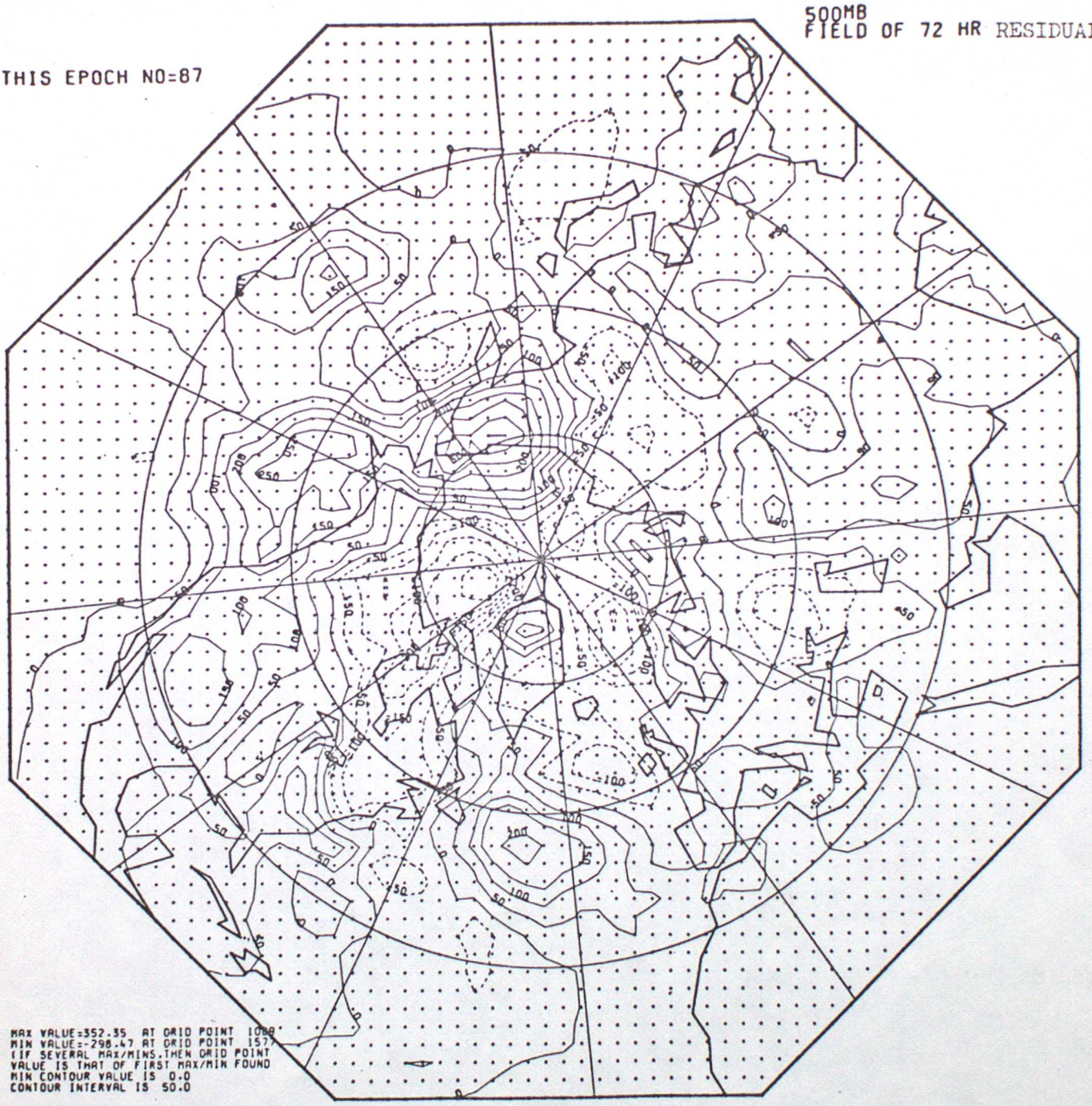
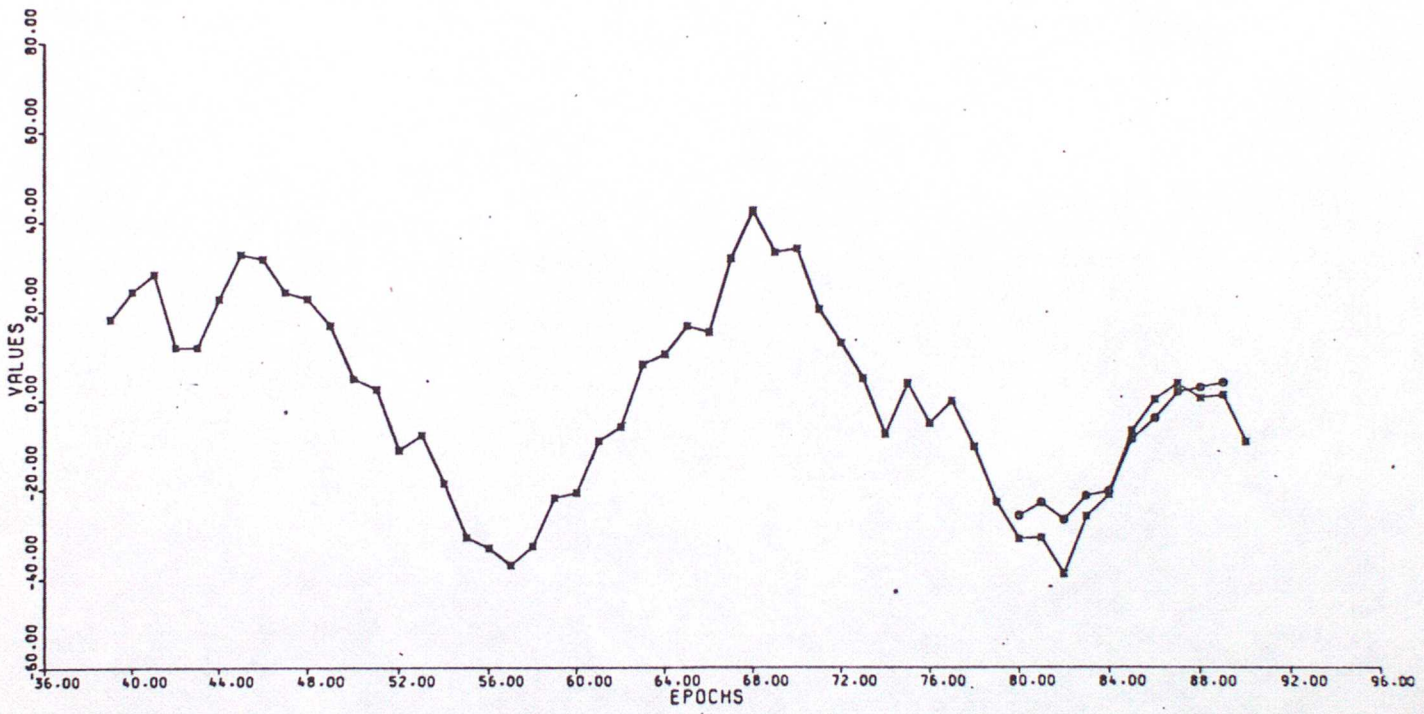


FIG. 11

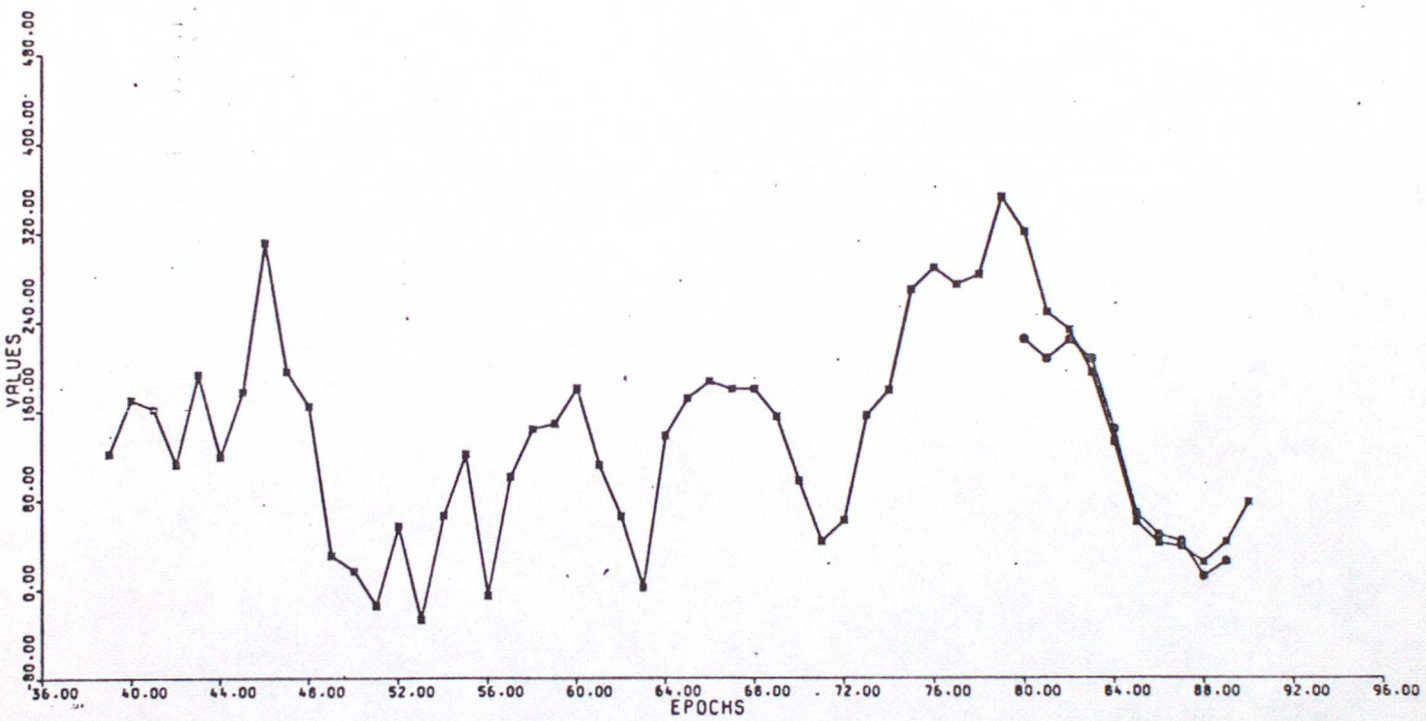
VALUES OF 500 MB 72HR ERROR (*) AND ESTIMATES (O)
FOR 23/12/77 TO 5 /2 /78
FOR POINT 141 FOR EPOCHS 39 TO 90



MEAN(ACTUAL) FOR WHOLE PERIOD =0.80
MEAN(ACTUAL) FOR EPOCHS 80-89 =-15.24
MEAN(ESTIMATE) =-12.49

FIG. 12

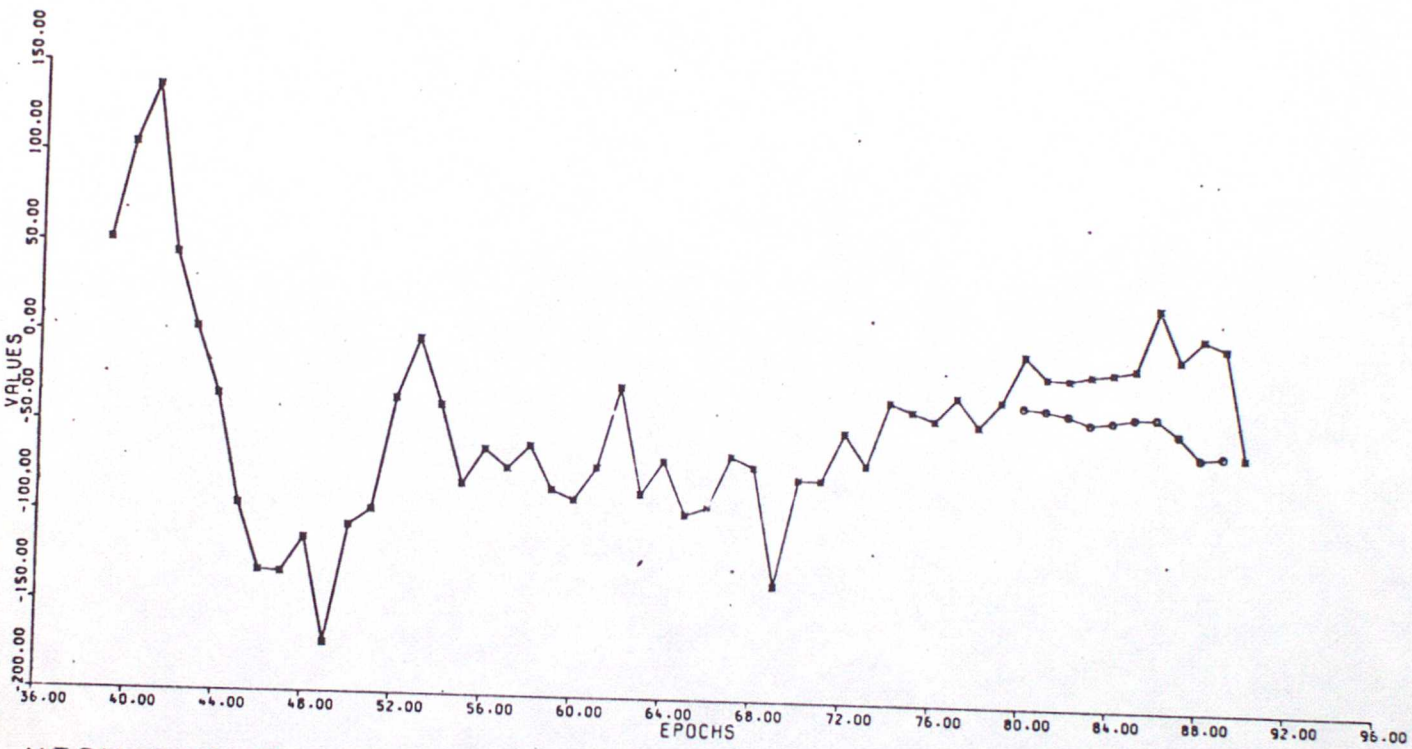
VALUES OF 500 MB 72HR ERROR (*) AND ESTIMATES (O)
FOR 23/12/77 TO 5 /2 /78
FOR POINT 717 FOR EPOCHS 39 TO 90



MEAN(ACTUAL) FOR WHOLE PERIOD =136.40
MEAN(ACTUAL) FOR EPOCHS 80-89 =132.11
MEAN(ESTIMATE) =118.82

FIG. 13

VALUES OF 500 MB 72HR ERROR (*) AND ESTIMATES (O)
FOR 23/12/77 TO 5 /2 /78
FOR POINT 2725 FOR EPOCHS 39 TO 90



MEAN(ACTUAL) FOR WHOLE PERIOD =-45.17
MEAN(ACTUAL) FOR EPOCHS 80-89 =-3.87
MEAN(ESTIMATE) =-41.45

FIG. 14