THE USE AND VERIFICATION OF EMCWF FORECAST RESULTS, A CASE STUDY

BY
K. ARPE
EUROPEAN CENTRE FOR MEDIUM RANGE WEATHER FORECASTS

1. Introduction

The purpose of this lecture is to discuss the question of the verification and utilisation of medium range forecast products. This question is rather new and almost unexplored. The Centre has made about 30 experimental ten-day forecasts to date, giving 12 different cases. To compare these forecasts with reality, different presentations, verification and diagnostic tools have been developed at the Centre and by applying them on a single case (6 February 1976), I will illustrate their values and limitations by considering two forecasts based on the same initial data, but with two different models. The forecasts will be notated X27 and X33 in the following. Both models are identical in their adiabatic formulation (see Burridge and Haseler,1977) and have the same resolution : horizontally it is a global regular latitude-longitude grid with 1.825° grid distance (N48), and vertically they have 15 sigma levels with higher resolution in the boundary layer and stratosphere. They differ, however, with regard to parameterization of the physical processes : X27 has the GFDL formulation (see Miyakoda et al 1970) and X33 has the ECMWF formulation (see Tiedtke et al 1978). These forecasts were made as part of a larger experiment, the result of which will be published in the near future as an ECMWF Technical Report. The forecast results will be compared with the operational analyses made by NMC, Washington.

We shall concentrate our study on daily maps and other presentations of the 1000 mb and 500 mb height fields and of the 850 mb temperature field. We shall compare the subjective impression gained by these maps with objective skill scores.

2. 1000 mb height fields

The first three days of the forecasts were very successful and only minor differences between the two forecasts occur. Therefore, we shall disregard the very first days and concentrate the comparison and the evaluation on day 4 and day 5 (fig.1). The maps at the bottom show the observed fields and on top the two forecasts. For both days I shall point out the differences between forecasts and observation and give some general remarks:

Day 4 :

In general the main features like the big low pressure area with two centres over the Norwegian sea and north-eastern Canada as well as the large belt of anticyclones stretching from the eastern Atlantic over southern Europe towards Siberia is quite well forecast and consequently we can regard this as a useful forecast. However, a few other important meteorological features are not particularly well predicted as for instance the following :

The low at 40°E, 37°N got quite a different position in both forecasts, but this problem becomes less important on the next day, since a development is actually predicted in this area, especially in X33. The cyclone at 90°E, 60°N is forecast 10° too far south-south-west from where it should be, also this problem will become less important since this low weakens substantially both in the forecasts and in reality on the next day. The cyclone at 60°W, 40°N is forecast too far south and too weak, especially by X33. This problem becomes even worse on the next day at X33. Over the Pacific the individual cyclones are not always predicted correctly in position, but the area of cyclone movement and genesis is correctly predicted and therefore this problem might not be too serious.

Day 5 :

The main problem on this day is that both forecasts failed to predict the breaking of the anticyclone over southern Europe. Furthermore, the change over the central United States from cyclonic to anticyclonic pattern was not forecast. At 140^{O}E, 50^{O}N a cyclone intensification was forecast while the observed low weakened.
On the whole these major failures reduce the usefulness considerably,especially compared to the 4th day.

To obtain an overview over the developments of the whole period of forecasts fig. 2 shows Hovmöller's trough-ridge diagrams. For the observed field in the lowest panel also the axes of troughs and ridges are automatically evaluated and then copied to the other panels , showing the forecasts, to make a comparison easier. Up to day 4 the forecasts are quite accurate but after that the errors grow rapidly.

In figs. 3 and 4 we want to see if the most commonly used skill scores, i.e. RMS error and anomaly correlation coefficients, do reflect the subjective impression gained from figs. 1 and 2. For convenience the values of a persistence forecast are included in both figures and the climatological variance (here called NORM) is marked in fig. 3. The RMS error between day 4 and day 6 stays almost constant, especially for X33, although there was a marked change in forecast quality from a subjective point of view. This is a clear warning to be careful when using this skill score. On the other hand, the RMS error just reaches the NORM after day 4 and this point is often used as an indication for the end of predictability which would agree with the subjective judgement. The correlation coefficients more clearly indicate the expected change in skill after day 4.

We have already listed above the errors of the forecasts on day 5 based on fig. 1. If we take a more general view, we will find that these errors are not as severe as fig. 1 indicated. The splitting of the anticyclone over southern Europe was in fact forecast by X27 one day later and by X33 less intense two days later. Also, the cyclone at 140°E, 50°N fades away one day later. We could also give a long list of areas where the forecasts are quite good beyond day 4 and shall now try to find ways of separating the useful information from the useless information of the forecasts. As the models are no longer able to predict the exact positions of moving cyclones, it may be that time means over several days are still correclty forecast. Fig. 5 shows an example of a 3 day mean 1000 mb height field covering the period day 4½ to 7½. Here we find forecast failures and successes similar to those in the daily maps.
The cyclone at 20°W, 70°N was forecast too far east, the anticyclone at 45°E, 53°N is in the forecast too far south, the forecast predicts a cyclone at 140°E, 55°N which is not observed. The splitting of the anticyclone over the Mediterranean Sea is not obvious. We obtained correct forecast features like the anticyclone at 30°E, 40°N, the cyclone over the northern Atlantic, pattern over Siberia etc. which could already be seen in fig. 1. On the whole this method of extracting information out of the forecast products seems to give no better skill than using daily maps.

This can also be seen by using Hovmöller's trough-ridge diagrams of the time smoothed data (fig. 6). Comparing this with fig. 2 it becomes obvious that time smoothing does not at all improve the similarity between forecasts and analyses.

Thus the present models do not seem to be good in predicting the quasi-stationary part, therefore we want to find out if they are better in predicting the transient part of motion (fig.7 and fig.8). Fig.7 shows the cyclone tracks for the same period as before. The areas of cyclonic activity seem to be forecast reasonably well and it may be useful to have this information beyond the point of normal predictability, here beyond day 4. But displaying the centres of cyclones, as done here, has the disadvantage that cyclones of different strength get the same weight.

To circumvent this difficulty we took from each map between day 4½ and 7½ the largest closed contours surrounding individual lows. These seven charts were then overlaid and the envelopes of the overlapping contours are shown in fig. 8 for both the observation and for X33. It shows a fairly good similarity. Such information is probably very useful and a skill score can be easily defined. The problem is to define the areas of cyclonic activity. In fig. 8 areas of closed isobaric lines around a cyclone were taken subjectively, and to get an objective measure, the following properties could be used :

a) maximum of vorticity
b) maximum or minimum of 1000 mb height field tendency
c) areas of closed isobaric lines around a cyclone
d) minimum of height field
e) maximum or minimum of temperature tendency
f) maximum of horizontal gradient of temperature.

This list can be easily extended and also combinations of these parameters can be used. In these examples three day means or maxima were taken but this is not necessarily an optimum number of days,and further investigations of this question have to be carried out.

From the last few figures we can draw the following conclusion : beyond day 4 there is still useful information in the transient part of motions, and tools to extract this information from the forecast products are available. There was also useful information in the quasi-stationary part of motion, but we do not yet have the tool to separate it from the useless information. How far these results are common for other forecasts has still to be investigated.

3. 500 mb height fields

From our experience with short-range forecasts we may expect that the 500 mb height fields have a larger predictability than the 1000 mb height fields. Fig. 9 shows the 500 mb height field maps for day 4 and 5. On day 4 the similarity between forecasts and observation is quite good, only the low at 25°W, 70°N, the trough at 120°W and the ridge at 155°E, 55°N are too weak.

On the 5th day the errors start to grow for both models but both models show different failures :

X33 : The Atlantic ridge, the low at 20°W, and the trough at 120°W are too weak.

X27 : The low at 90°W, 65°N is much too strong, the trough at 60°W is too weak, the low at 20°W, 70°N got a wrong position, and the ridge at 150°E, 55°N is too weak.

Both models did forecast the low over the Black Sea as a trough only. Although this is already a long list of detailed deficiencies, such a forecast map may still be of value because the main features are predicted reasonably well. Furthermore, the listed deficiencies were only errors of intensity or slight differences in position which will not greatly affect the usefulness. Although both forecasts look different, it is hard to judge which of both is the better one.

Further 500 mb height field maps are not shown here,but it can be seen from Hovmöller's trough-ridge diagram in fig. 10 that the skill on day 6 is worse than on day 5. It is also quite obvious from this figure that the forecasts are quite accurate up to day 4.

In agreement with what we expected,we obtained with some restrictions one more day of useful information when taking the 500 mb height field compared to the 1000 mb height field.
When working out the end of predictability, only those areas where the forecast failed were taken into account. However, in fig. 10 we can still find areas of good forecast beyond day 5, e.g. the trough starting at 50^{O}E and the area between 270^{O}E and 360^{O}E are correctly forecast for all 10 days. To establish whether we can extract this valuable information left in the forecasts, we shall try some data manipulations : Fig. 11 shows Hovmöller's trough-ridge diagrams of the same fields like before, but smoothed by using a 60-hour mean. Obviously this procedure does not improve the similarity between forecasts and observations at all, it may be even worse. It confirms our experience with the 1000 mb level height fields. The present version of the model is not better in predicting the quasi-stationary part of the circulation than in predicting the total fields, at least in this example.

We shall next investigate the effect of applying a space filter. Fig. 12 shows the trough-ridge diagrams like before but excluding zonal wave numbers higher than 3. As could be expected, this does not help either, as these long waves tend to be quasi-stationary. Both methods of filtering should, therefore, lead to similar results.

Figs. 13,14,and 15 are now provided to compare the subjective evaluation with objective skill scores. This time a separation into contributions of bands of zonal wave numbers is provided. The RMS-error of the total field, fig. 13 top, gives a predictability of 6 days if the intersection of NORM and RMS error is taken; this does not agree with the 5 day limit found by synoptic evaluation. The figure also gives a clear advantage for X33 and this comes from the very long waves with wave numbers 1 to 3 in the bottom panel. This wavenumber band was presented in fig. 12 and from that no difference in quality between both forecasts can be found, neither does fig. 13 shows the expected decrease in skill between day 4 and day 6. This skill score, therefore, fails to reflect the subjective impression.

The short waves with zonal wavenumber 10 to 20 already reach an RMS-error level equal to the NORM and to persistence forecast on day 2. This is partly due to a very short predictability of these waves, but it can be due also to uncertainties of analyses in this wavenumber band. In fact comparing data from two different analyses schemes may cause similar RMS-differences. This wavenumber group should, therefore, not be used for verification purposes.

More information about differences in analyses schemes will be published soon.

The correlation coefficients in fig. 14 are better in showing the difference of forecast quality between day 4 and day 6. They also indicate quite different skill for both experiments with a clear advantage to X33 due to the very long waves. This is hardly borne out by fig. 12. But from fig. 12 it can be seen that at just about day 5 the amplitudes of the long waves are smaller than at the beginning or end of the forecast period. With relatively small amplitudes a subjective judgement would assume a good forecast even if the waves are not in phase, while the correlation coefficient might indicate the opposite. This may be the reason why the agreement between subjective judgement and correlation coefficient was good for the 1000 mb level and less good for the 500 mb level at about day 5.

To get rid of the normalizing effect of the correlation coefficient we use an idea of Williamson (1978).
For all wavenumbers the forecast data are changed so that they keep the phase of the original forecast data and take the amplitude of the analysis fields. Then the RMS-error is computed. It is called RMS phase error. Fig. 15 gives this score for our example for the 500 mb height field. It shows the sudden change of growth rate of errors after day 5, especially for the long waves. It does not give an advantage to either of the forecasts on day 5 for the total error. Both facts do agree with our subjective judgement.

4. 850 mb temperature field

An important potential use of the ECMWF forecast products will be a prediction of temperature. Therefore this parameter too should be used for verification. As can be seen from the lecture by Klein (page 221 - 272 of these proceedings), 850 mb and 700 mb temperature forecasts are useful predictors for the surface temperature.

To illustrate the quality of 850 mb temperature forecasts fig. 16 shows some Hovmöller trough-ridge diagrams for this parameter. This time a latitudinal mean over 20° instead of a single latitude is taken (the unaveraged data leads to similar conclusions). It is obvious that the length of predictability is much greater in this temperature field than in both height fields. One could say that the similarity between forecast and analysis is very good up to day 4 and good up to day 7. After day 7, especially X27,gets considerable phase errors between 130°E and 210°E. But as the amplitudes are quite small within this region and as the dominant features are reasonably well predicted, the forecast even after day 7 may be useful. To further illustrate the quality of forecast, fig. 17 shows the anomalies of temperature for day 6. The main areas of anomalies are fairly well predicted.

Figs. 18, 19 and 20 show the skill scores for the 850 mb temperature field. From day 4 onwards the RMS errors are always close to the NORM. This would indicate that the useful prediction ends with day 4 which is contrary to our subjective judgement. In the same period the correlation coefficients decrease steadily with time but keep quite a high level of 0.55 at day 6 and 0.2 at day 10. This is contrary to the RMS phase error, which keeps an almost constant level of 4 to 5 K over the same period. This disagreement between the different skill scores makes their use quite difficult. All skill scores agree in one point only, i.e. X33 is better than X27 after day 7 and this also agrees with the subjective judgement.

5. Conclusions

Our experience from short-range weather forecasts that 500 mb height fields have a longer period of predictability than 1000 mb height fields was upheld also for one medium-range forecast. It was also found that some more days of useful forecast skill would be gained by using the 850 mb temperature field. But also from 1000 mb height fields useful information could be gained beyond that period of predictability by taking 3-day cyclone tracks or areas of cyclonic activity

No single best way of verifying the forecast could be found in this study The subjective judgement was reflected best by the RMS phase error, and by the correlation coefficients and worst by the RMS error.

In conclusion we re-iterate that the purpose of this lecture has been the discussion of methods of forecast evaluation and verification. No general validity can be assigned to the estimates of useful predictability as we have discussed only one forecast from models that are still under development. However, the results in this case will give a guidance for evaluating a larger sample of forecasts.

References

Burridge, D. and J. Haseler, 1977 A Model for Medium Range Weather Forecasting - Adiabatic Formulation ECMWF Technical Report No. 4

Miyakoda, K., Hembree, G D., Strickler, R.F. and J. Shulman 1972 Cumulative Results of Extended Forecast Experiments. I. Model Performance for Winter Cases. Monthly Weather Review, 100, 836-855.

Williamson, D.L., 1978 The Relative Importance of Resolution, Accuracy, and Diffusion in Short-Range Forecasts with the NCAR Global Circulation Model. Monthly Weather Review, 106, 69 - 88.

Tiedtke, M., Geleyn, J.-F., Hollingsworth, A. and J.-F. Louis 1978 ECMWF Model - Parameterization of Sub-Grid Processes. ECMWF Technical Report No. 10.

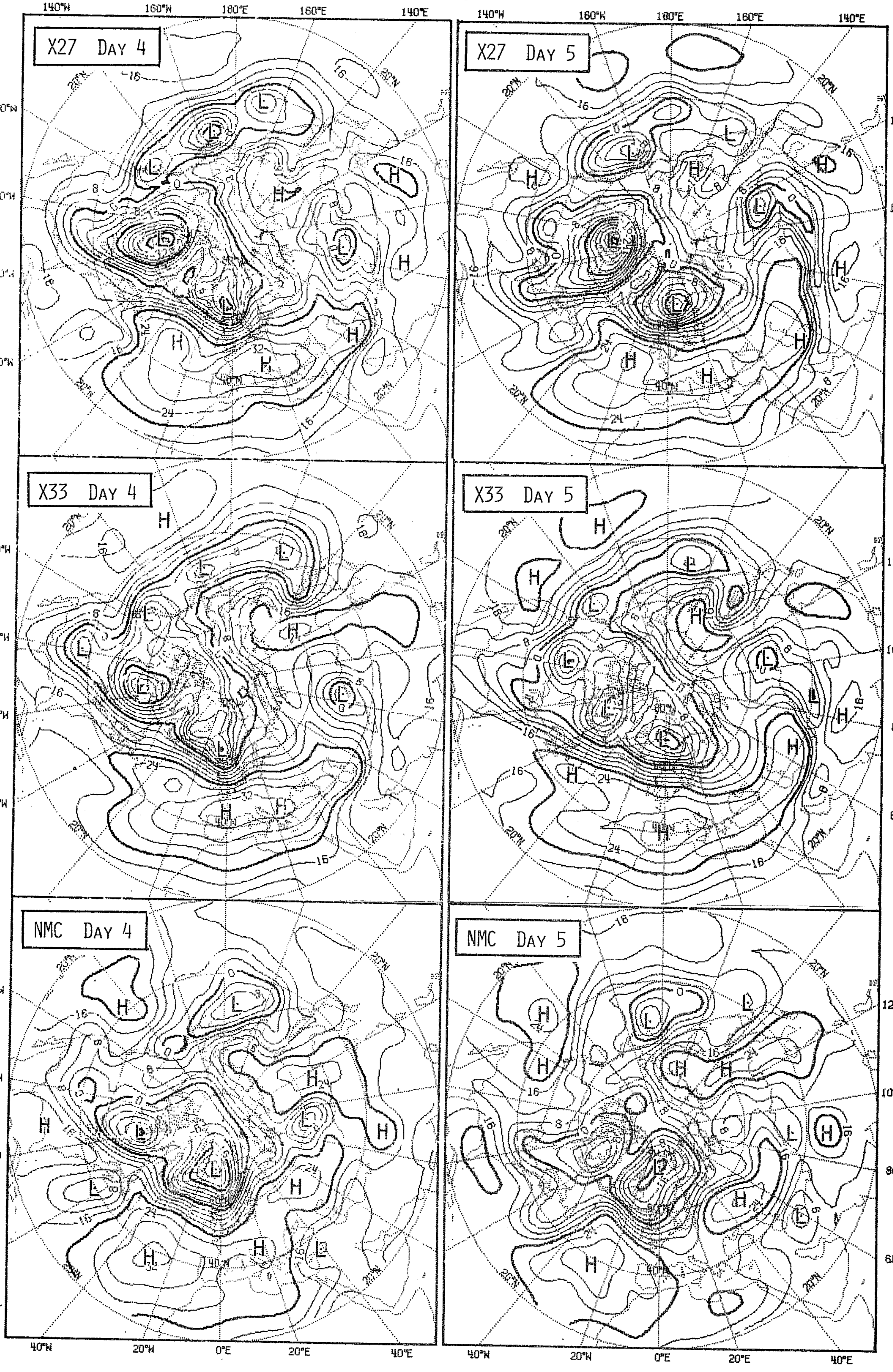

Fig. 1 1000 mb height fields for day 4 (left) and day 5 (right).
Contour interval : 4 decametres. Lowest panels

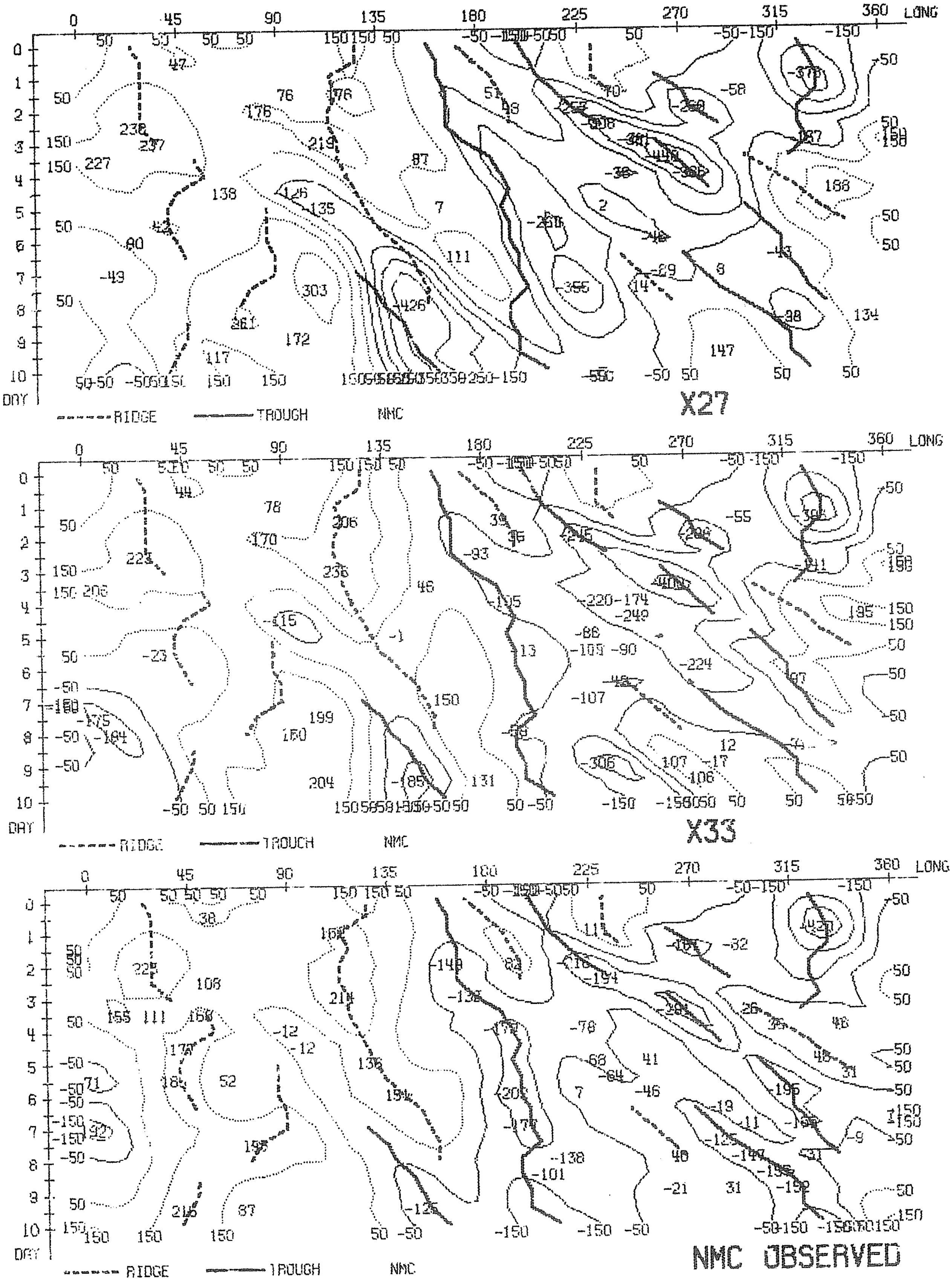

Fig. 2 Hovmöller's trough ridge diagrams for the 1000 mb height field at 55^{o}N. Contour interval : 100 m. Solid lines : Negative values, dotted lines: positive values. Thick solid lines : Axes of observed troughs. Thick dashed lines: Axes of observed ridges

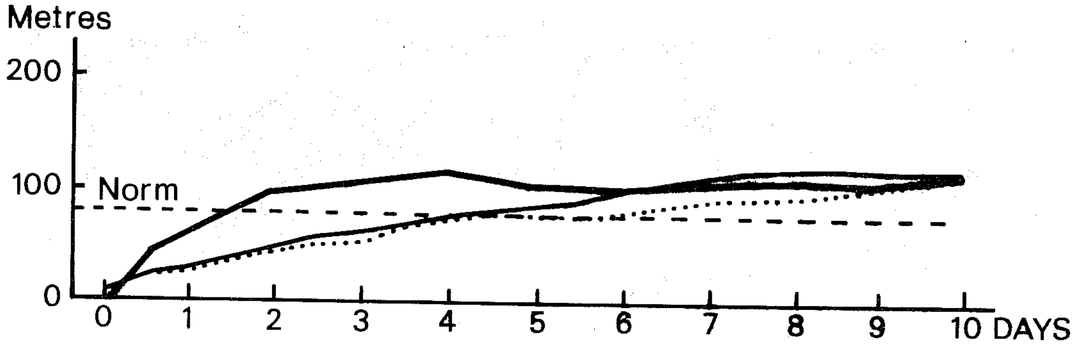

Fig. 3 RMS error of the 1000 mb height field. Mean between 20^{O}N and 82.5^{O}N.

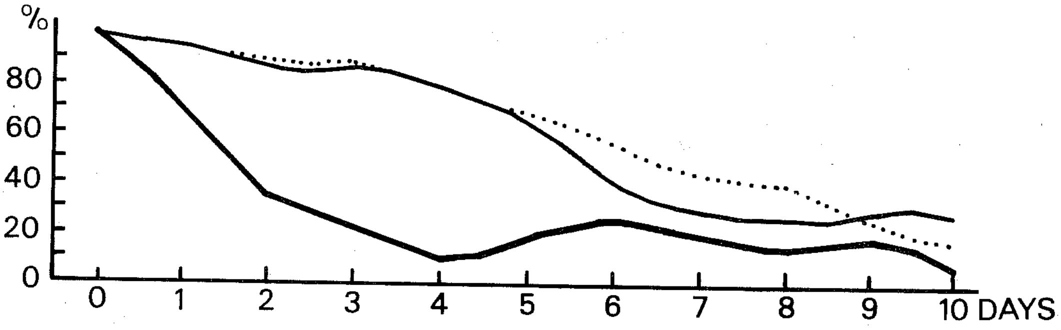

Fig. 4 Anomaly correlation of the 1000 mb height field. Mean between 20^{O}N and 82.5^{O}N.

▬ PERSISTENCE ── X 27 ······ X 33

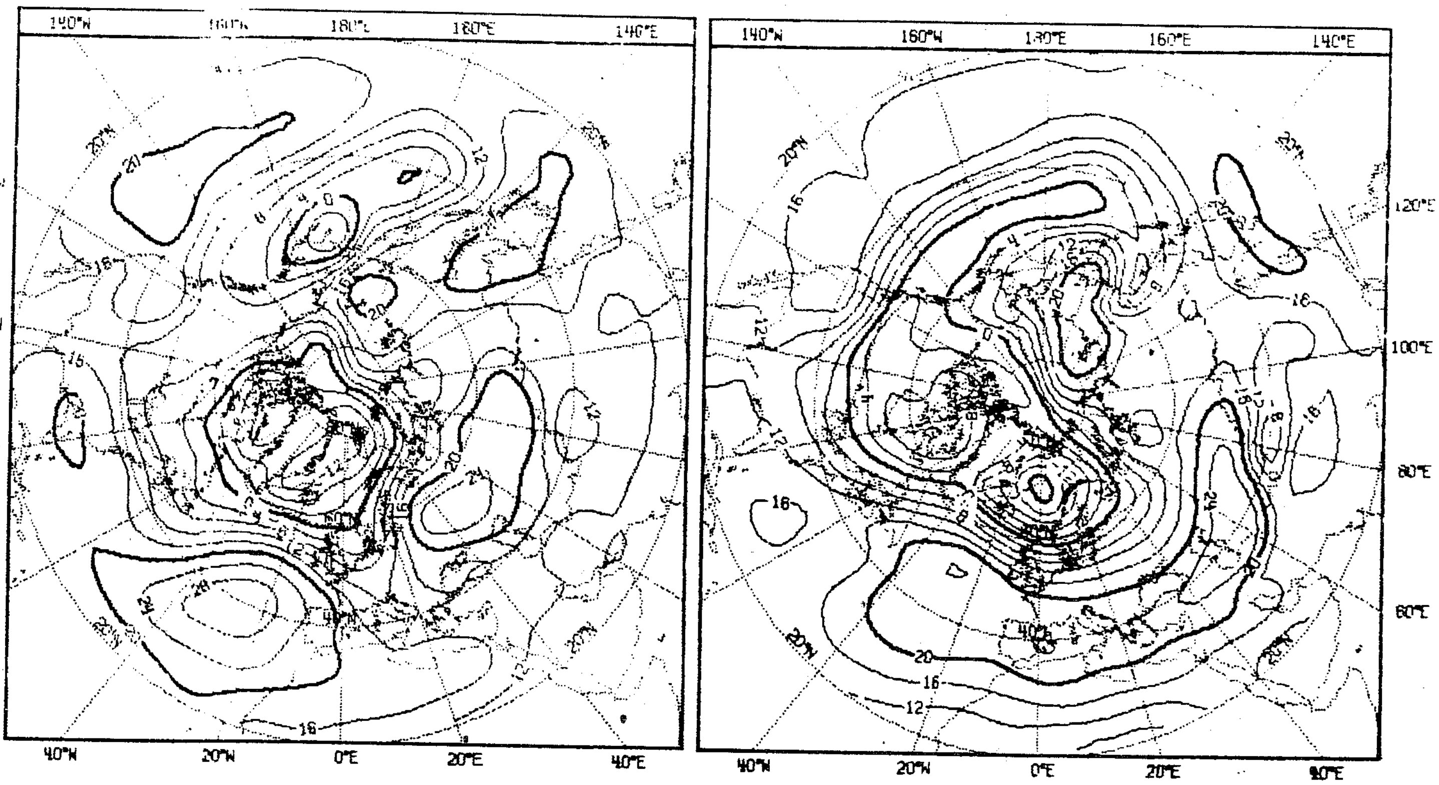

Fig. 5 1000 mb height field, mean between day 4.5 and 7.5. Contour interval : 4 decametres. Right : X33 Left : Verifying analysis.

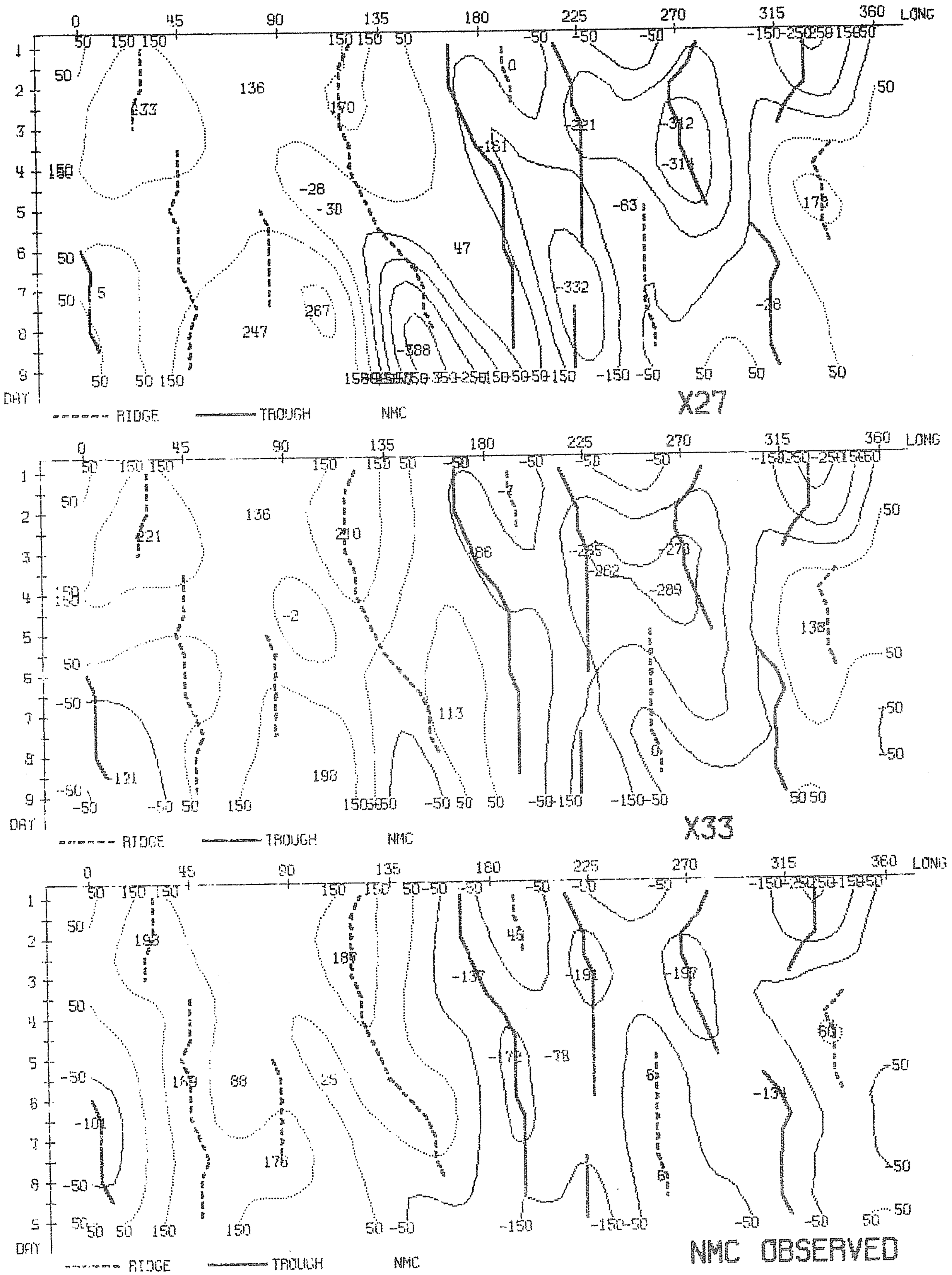

Fig. 6 Hovmöller's trough ridge diagrams for the 1000 mb height field at 55^{o}N. Data are smoothed by a 60 hour running mean. Contour interval and line shape as before.

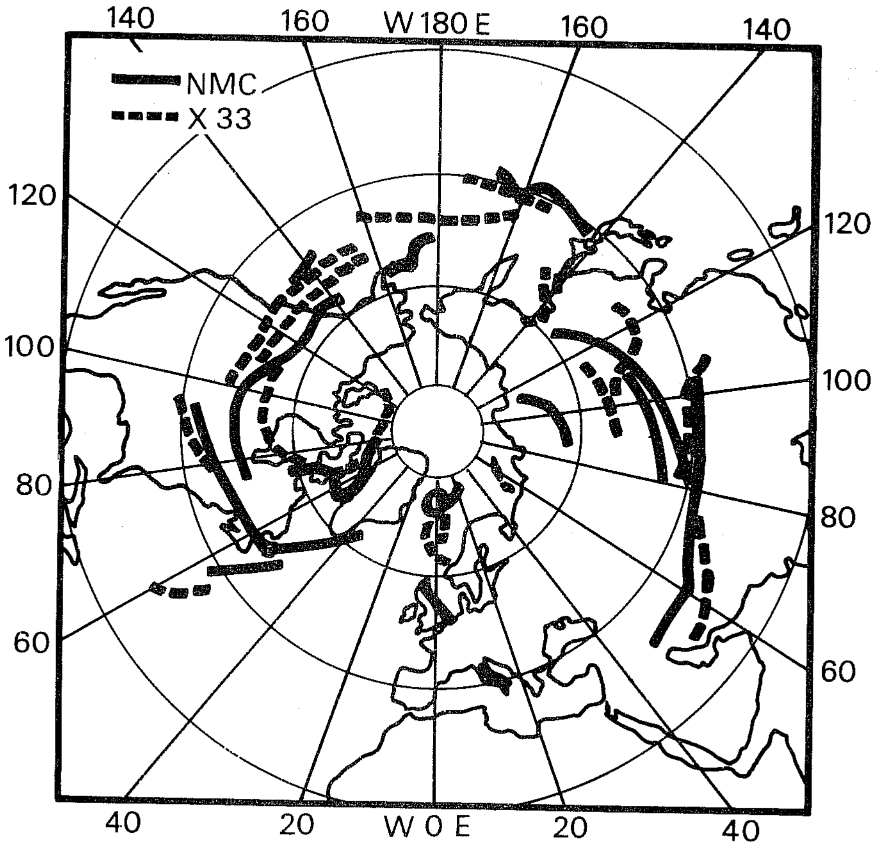

Fig. 7 Cyclone tracks between day 4½ and 7½ for X33 (dashed lines) and observation (solid lines).

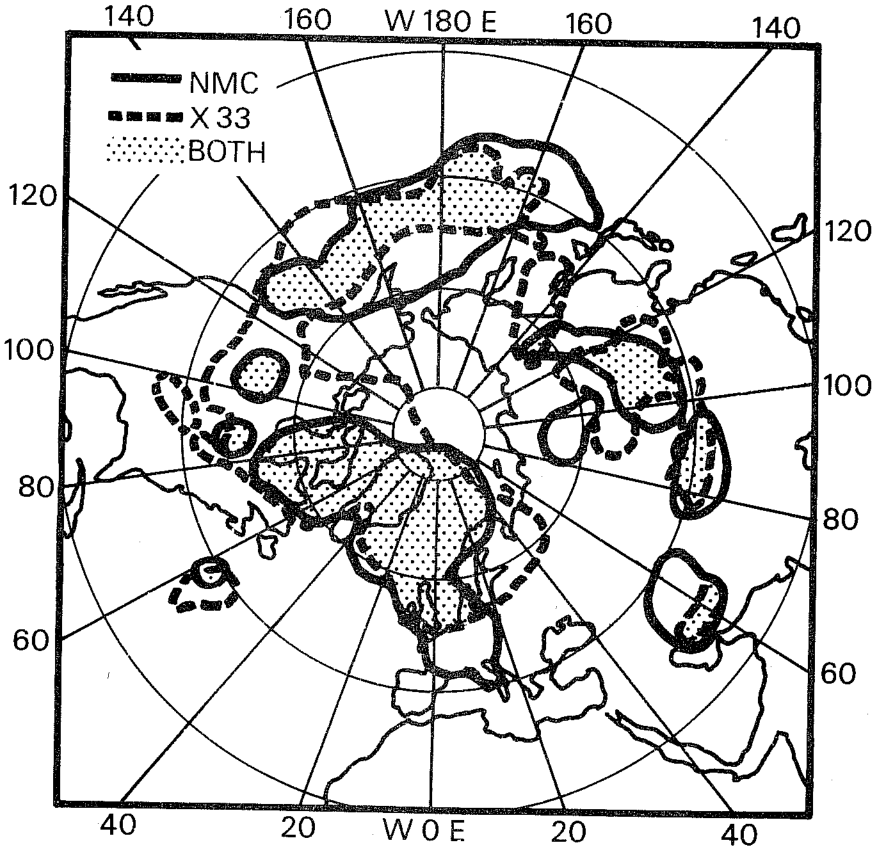

Fig. 8 Areas of cyclonic activity between day 4½ and day 7½ for X33 (dashed lines) and observation (solid lines). Correct forecast areas are dotted.

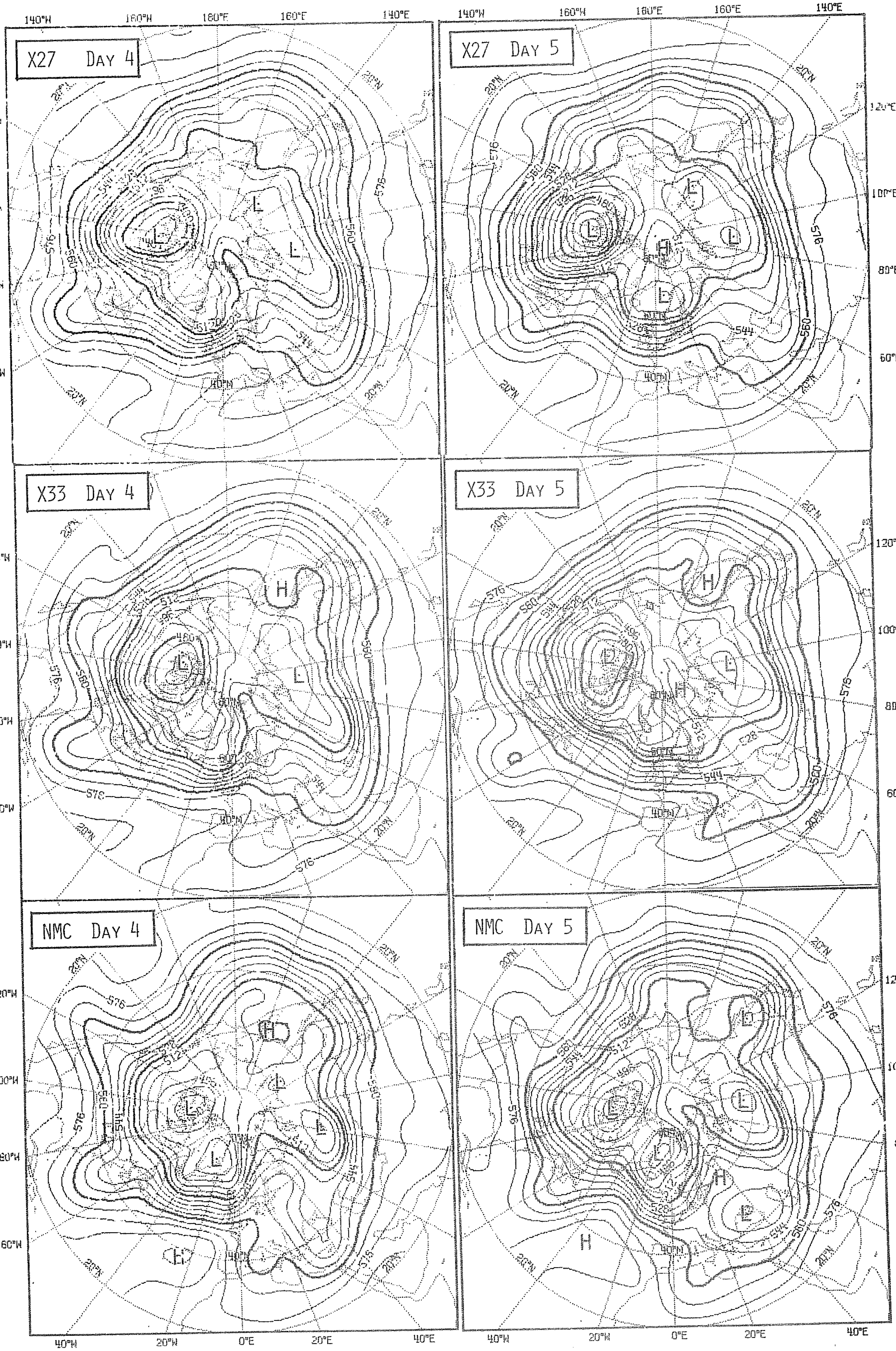

Fig. 9 500 mb height fields for day 4 (left) and day 5 (right).
Lowest panels : verifying analyses. Contour interval : 8 decametres.

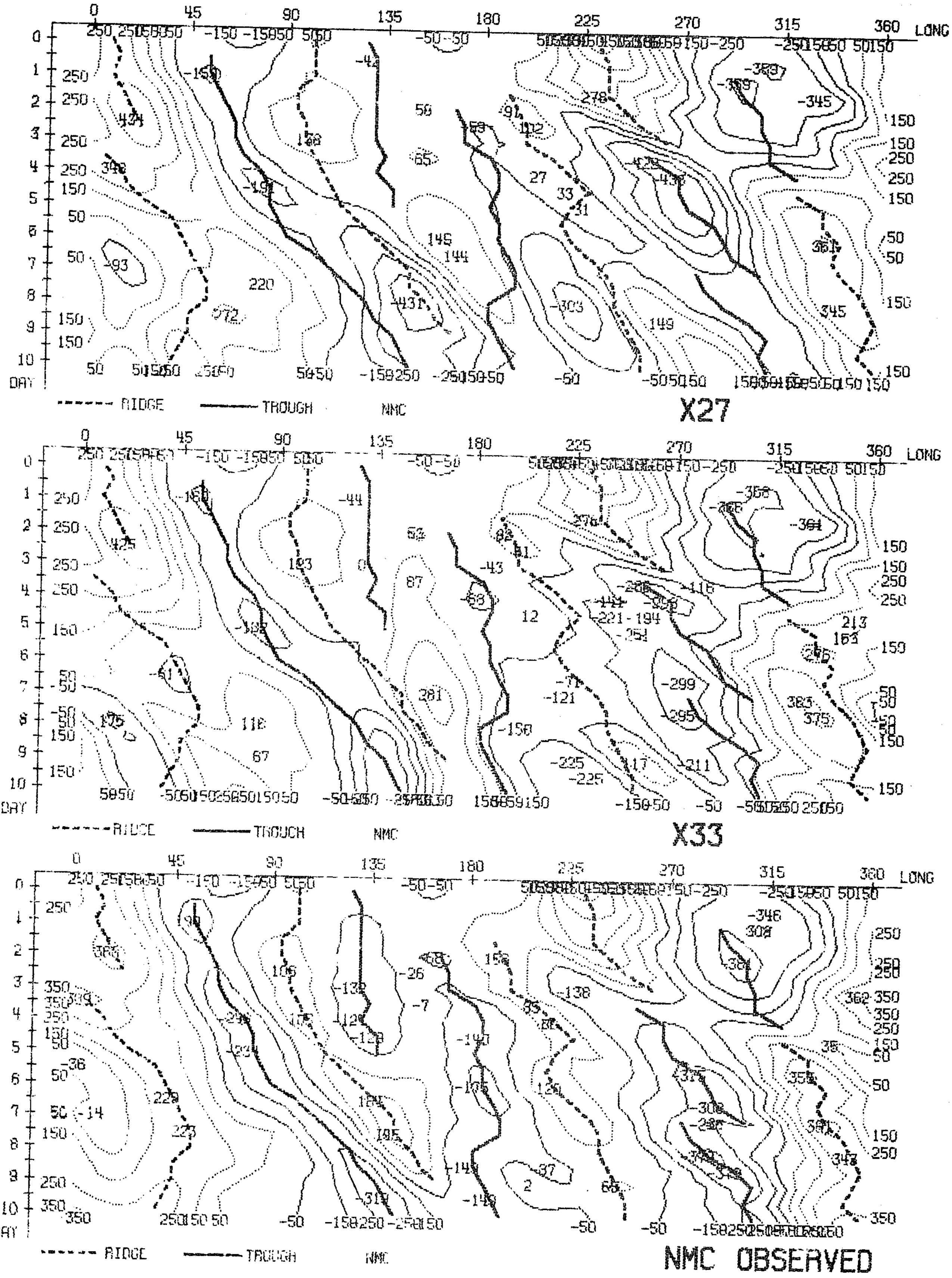

Fig.10 Hovmöller's trough ridge diagrams for the 500 mb height field at 55°N. Contour interval and line shape as before.

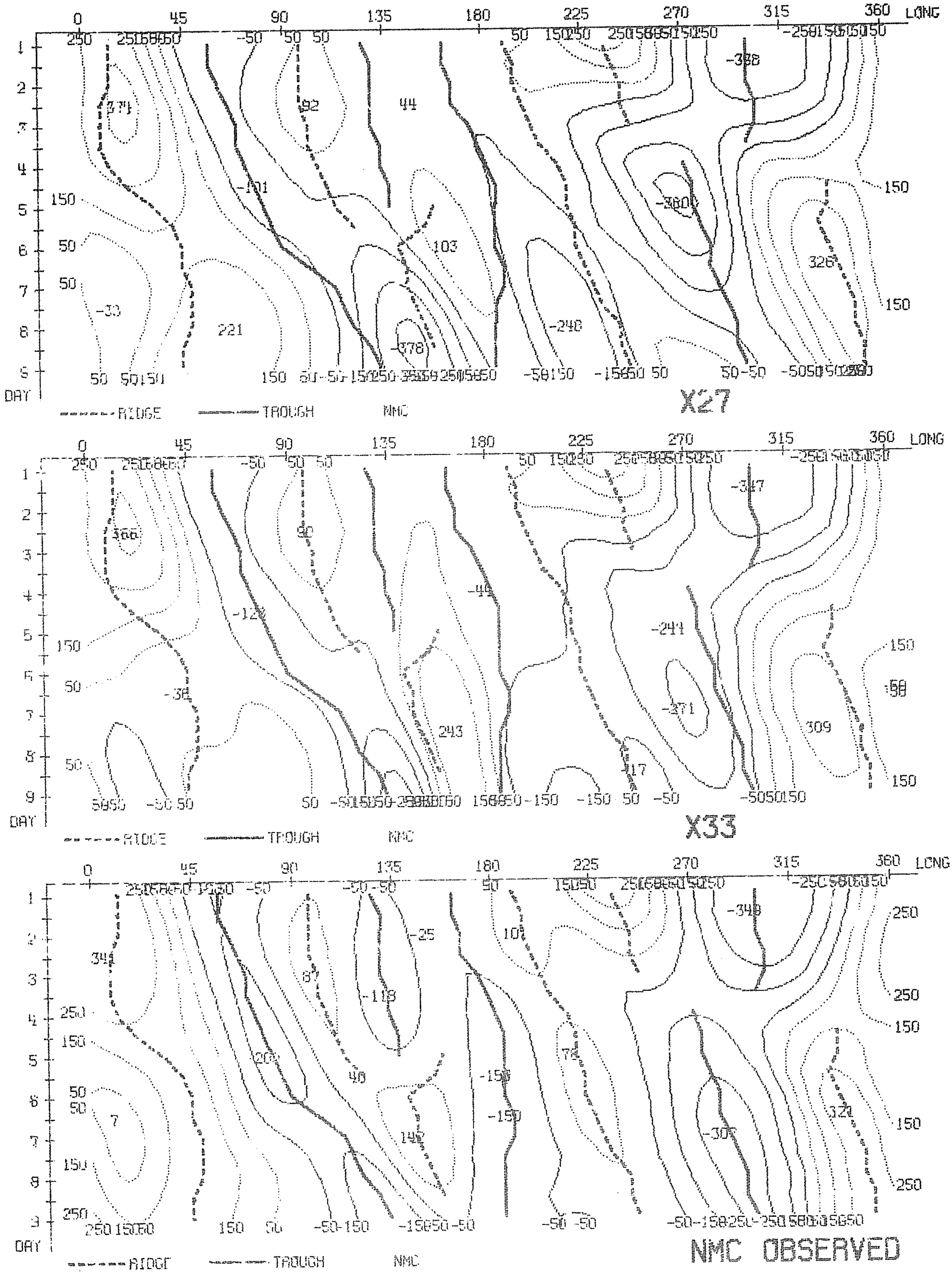

Fig.11 Hovmöller's trough ridge diagrams for the 500 mb height field at 55^{O}N. Data are smoothed by a 60 hour running mean. Contour interval and line shape as before.

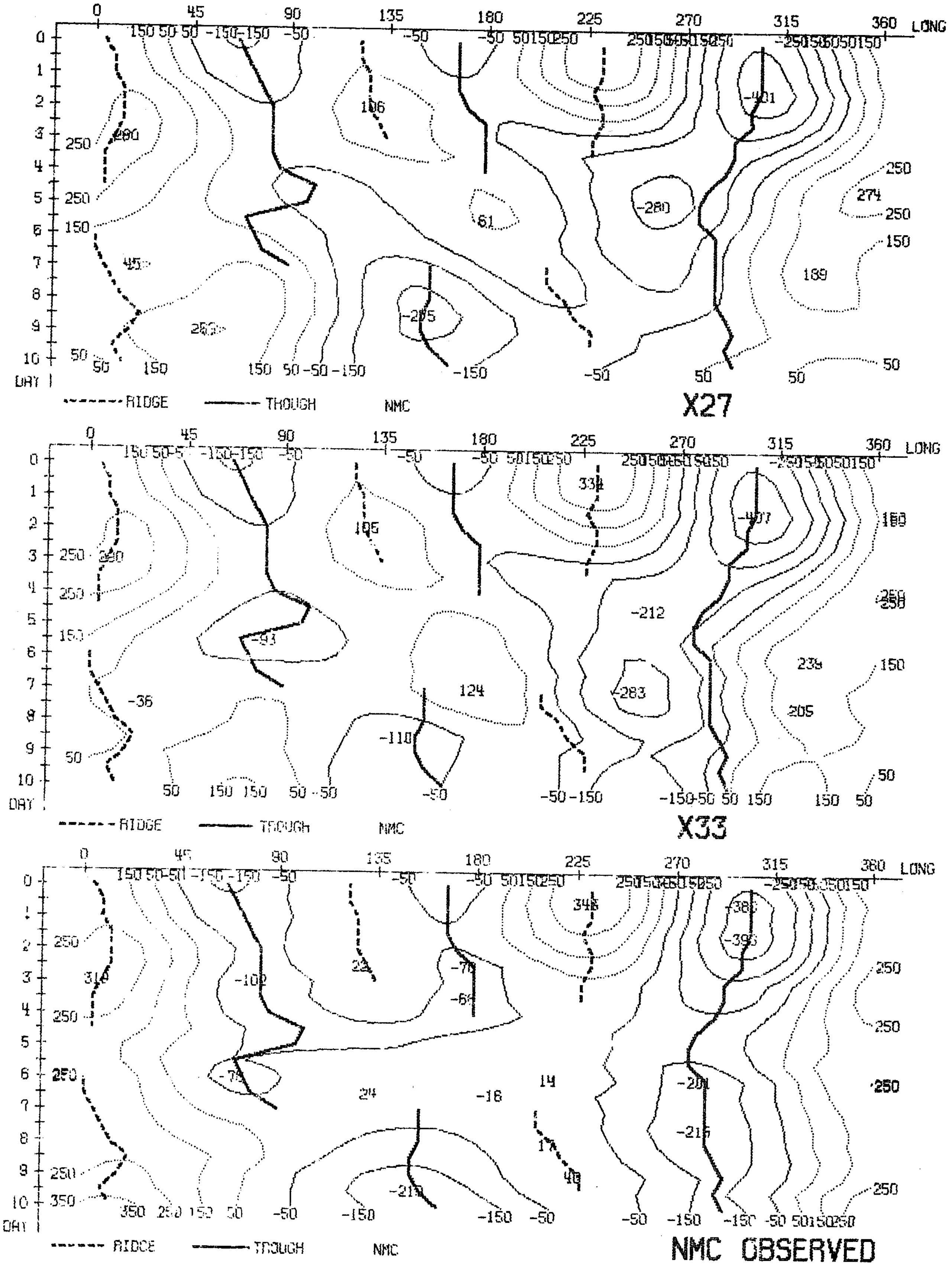

Fig.12 Hovmöller's trough ridge diagram for the 500 mb height field at 55°N of the long waves only. Zonal wave numbers higher than 3 are excluded. Contour interval and line shapes as before.

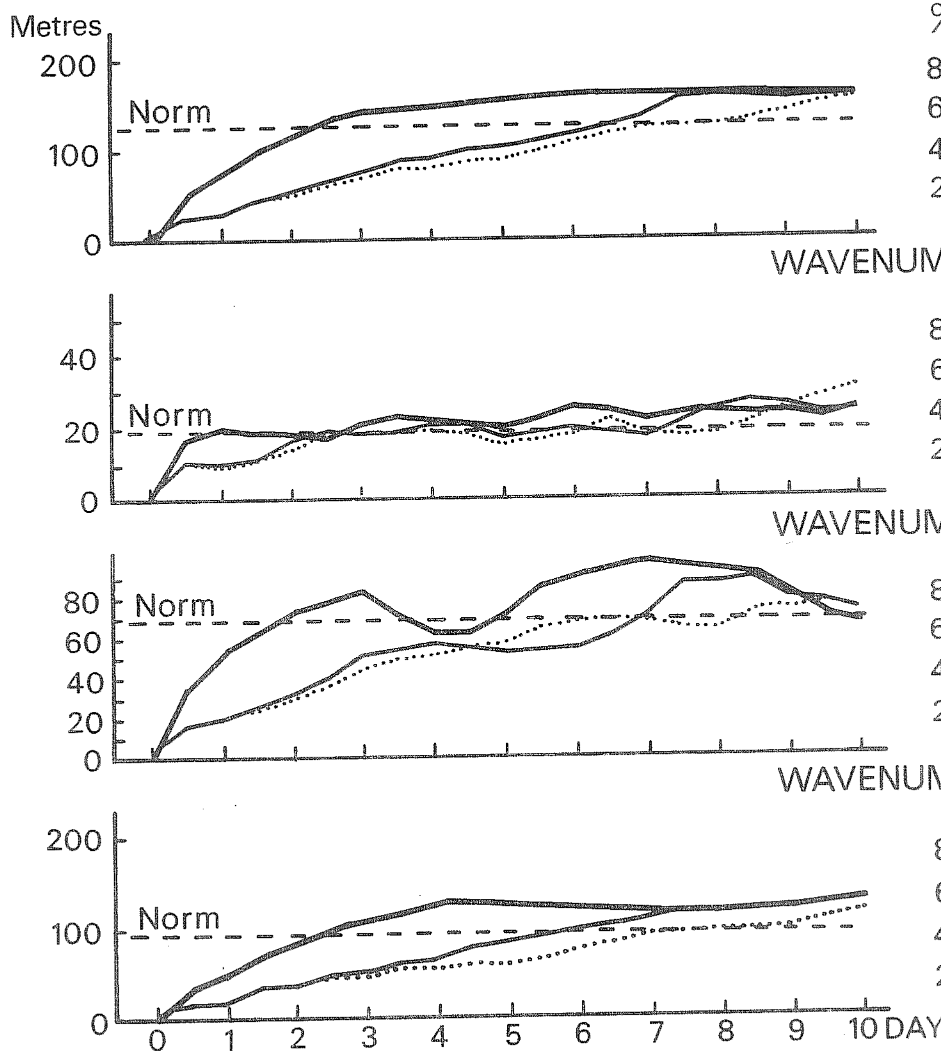

Fig.13 RMS error of the 500 mb height field for the total and three wave number groups. Mean between 20°N and 82.5°N.

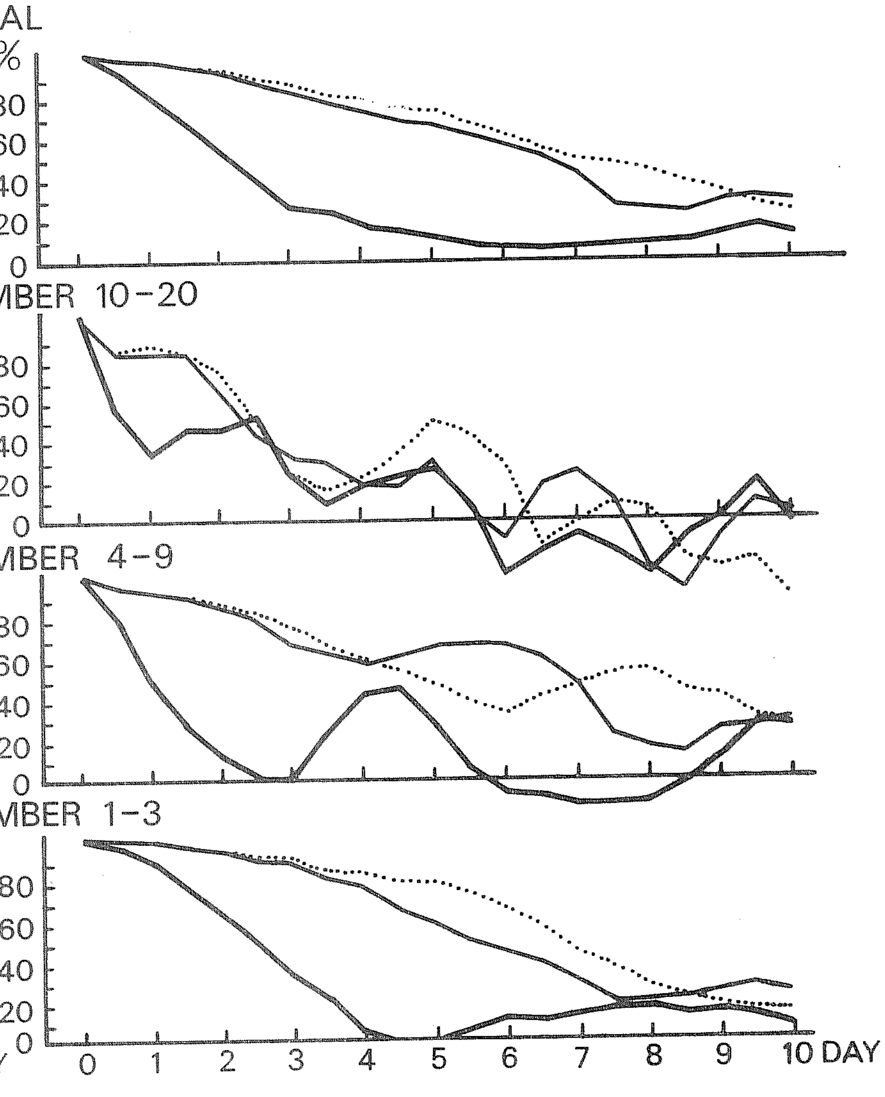

Fig.14 Anomaly correlation of the 500 mb height field for the total and three wave number groups. Mean between 20°N and 82.5°N.

PERSISTENCE

X 27

.... X 33

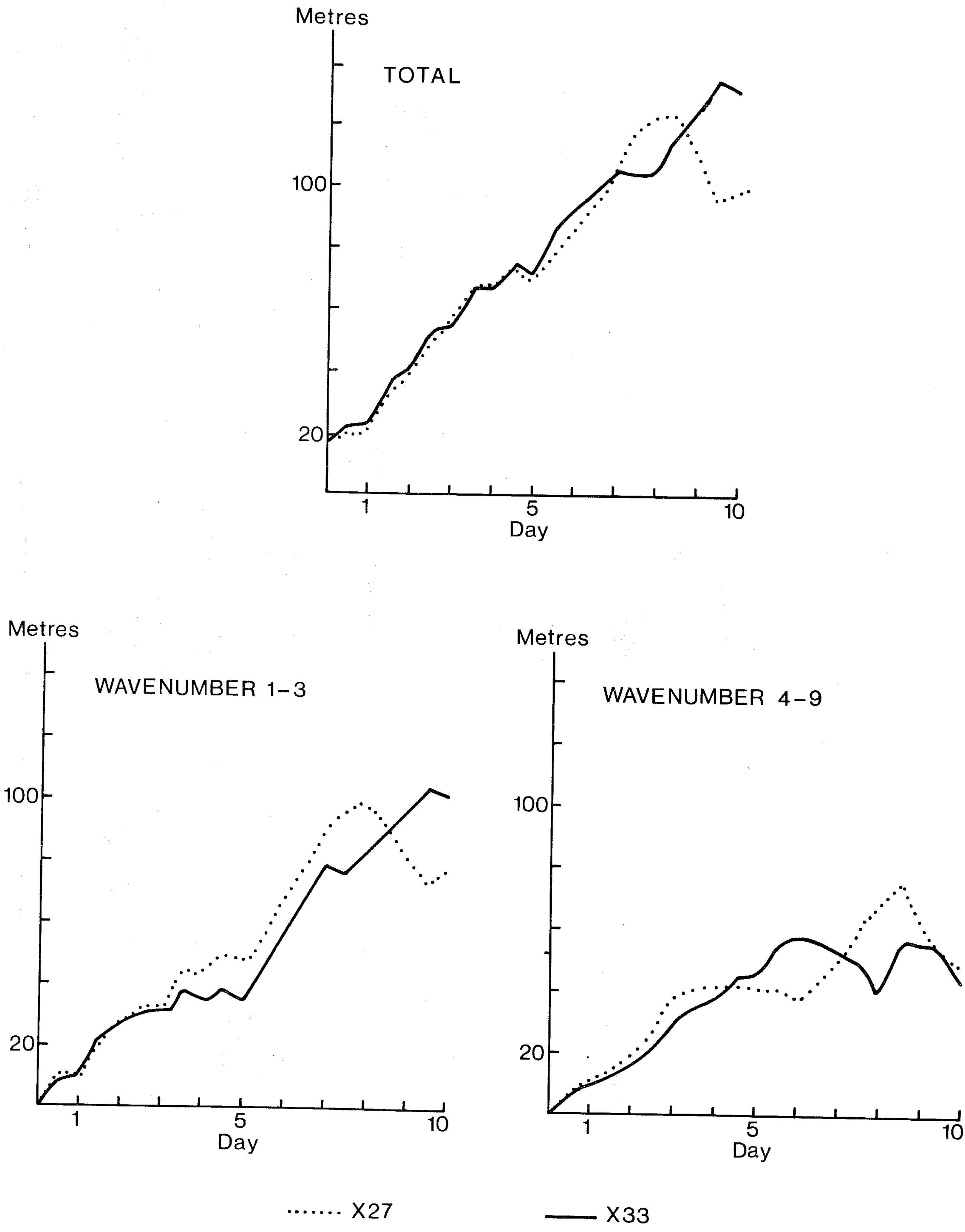

Fig. 15 RMS phase error for the 500 mb height field

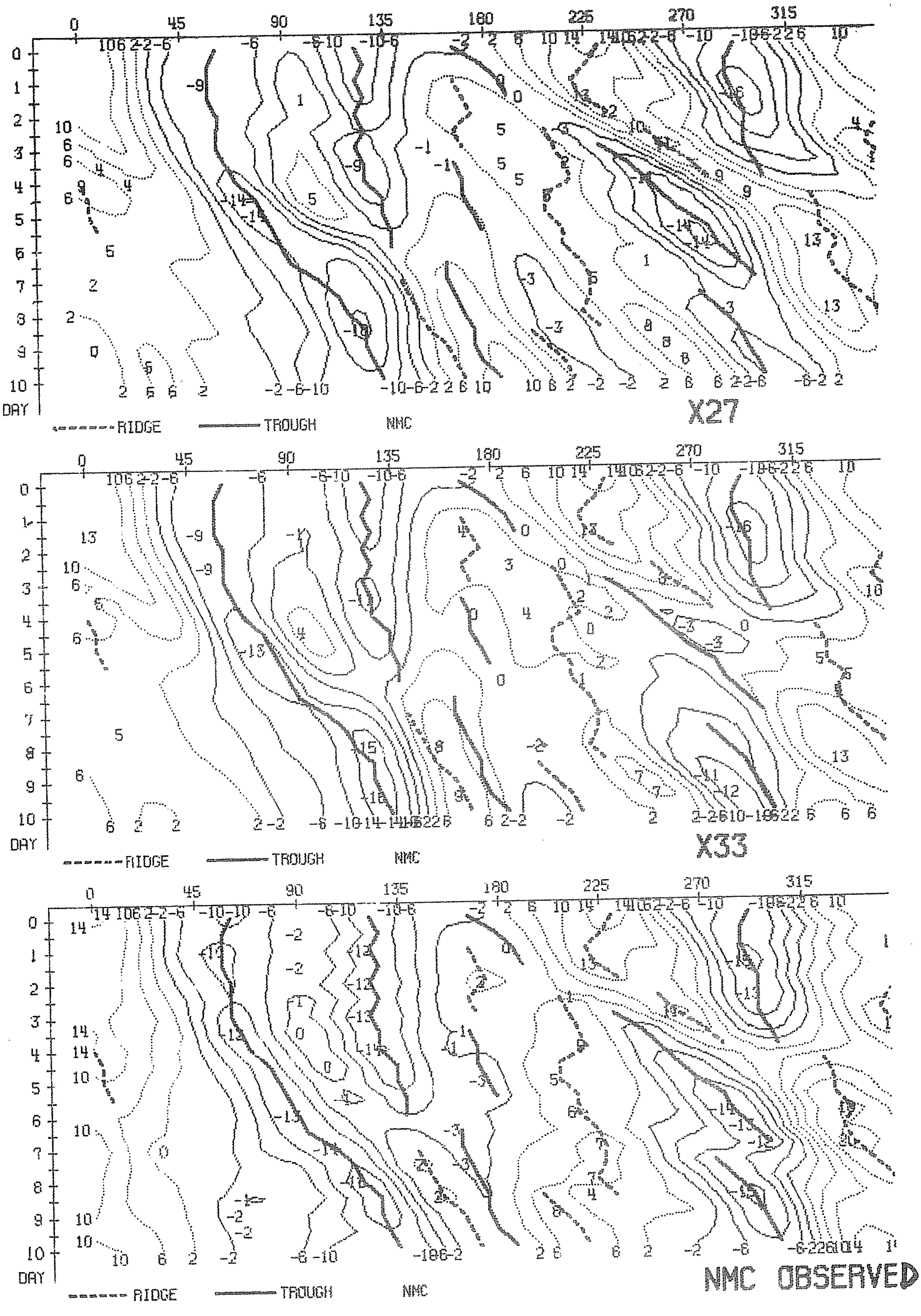

Fig.16 Hovmöller's trough ridge diagrams for the 850 mb temperature field. Latitudinal mean between 45°N and 65°N. Contour interval : 4K. Line shape as before.

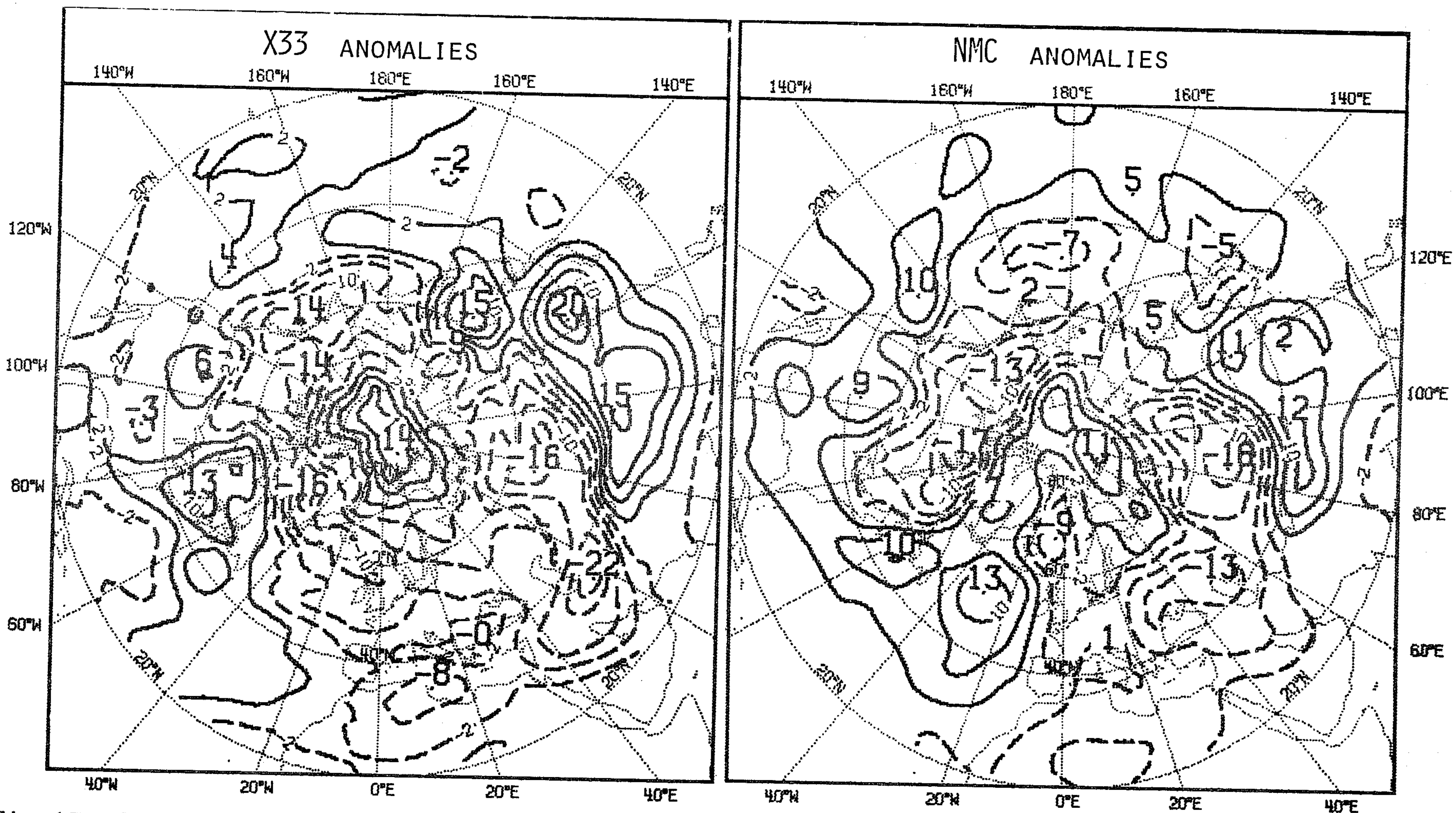

Fig.17 850 mb temperature anomalies for day 6. Contour interval 4K, negative lines are dashed and positive line solid.

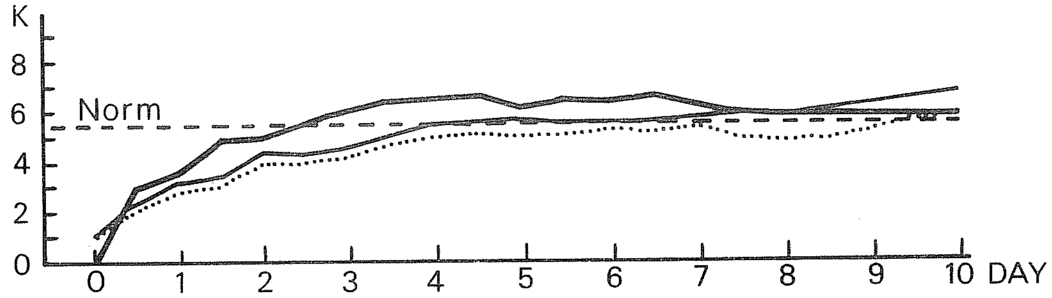

Fig. 18 RMS error of the 850 mb temperature field. Mean between 20^{o}N and 82.5^{o}N.

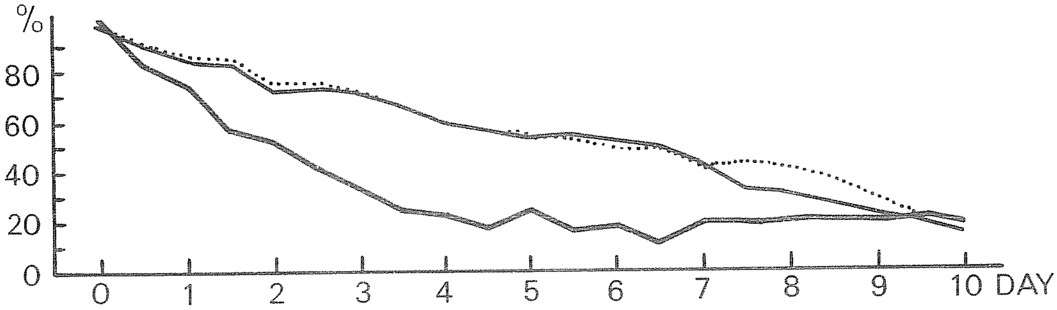

Fig. 19 Anomaly correlation of the 850 mb temperature field. Mean between 20^{o}N and 82.5^{o}N.

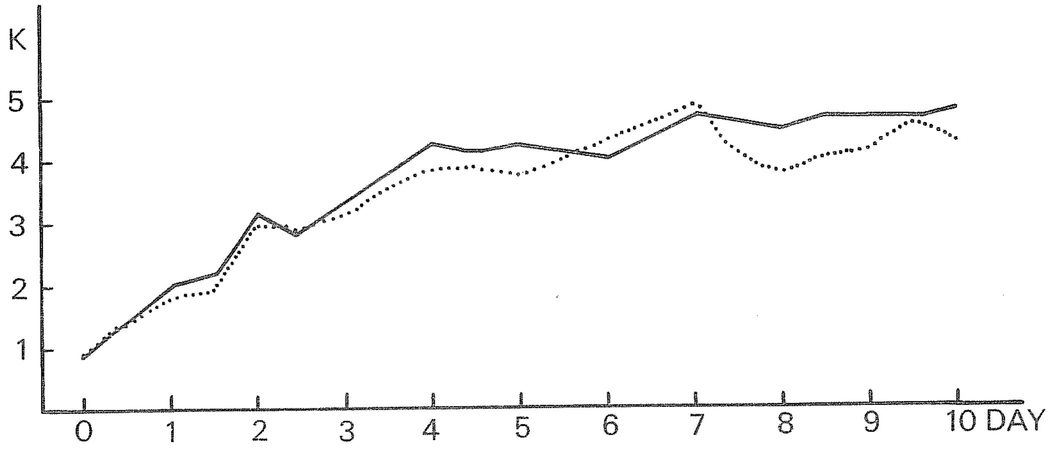

Fig. 20 RMS phase error of the 850 mb temperature field. Mean between 20^{o}N and 82.5^{o}N.

▬▬ PERSISTENCE ——— X 27 ······· X 33