Long-Range Forecasting and Climate Research

Extended range prediction experiments

using an 11-level GCM

by

J. M. Murphy and A. Dickinson

Revised January 1988

R-C 18

April 1988

LONG RANGE FORECASTING AND CLIMATE
RESEARCH MEMORANDUM NO. 18

EXTENDED-RANGE PREDICTION EXPERIMENTS USING AN 11-LEVEL GCM

by

J M MURPHY AND A DICKINSON

MET O 13 (SYNOPTIC CLIMATOLOGY)
METEOROLOGICAL OFFICE
LONDON ROAD
BRACKNELL
BERKSHIRE RG12 2SZ

APRIL 1988

NOTE. This paper has been accepted for publication in Meteorology and Atmospheric Physics.

LONDON, METEOROLOGICAL OFFICE.

Long-Range Forecasting and Climate Research
Memorandum No.18
Extended-range prediction experiments using an
11-level GCM.

02660688 FH1B

Summary

Recent research in dynamical extended-range prediction at the UK Meteorological Office (UKMO), based on 40-day integrations of a global 11-level general circulation model, is described. The forecast anomaly correlation scores, calculated with respect to some set of background atmospheric normals, contain a significant contribution due to differences between the normals and the true atmospheric climate for the years containing the experimental initialisation dates. This contribution varies according to the choice of normals. The best set to use in practice are those which minimise the measured scores, since they are closest to the true climate. The model's own climatology is sufficiently realistic for it to be suitable for long-range forecasting. However, significant climate drift still occurs in all seasons, and empirical correction for this increases the model's skill substantially, although the use of dependent data exaggerates the improvement somewhat. On average, the skill of winter and spring forecasts exceeds that of summer and autumn cases for days 1-15 and 6-20, for the domain 30-90°N. Although the mean skill remains well above zero throughout the forecast period, there are few cases of high skill, on the hemispheric scale, at extended-range. However, study of local skill, over a region centred on the UK, shows that the model's ability to forecast surface pressure anomalies compares favourably with that of the experimental long-range forecasts produced at UKMO using statistical forecasting techniques and medium-range dynamical predictions. The major improvement is for days 6-15. Based on the anomaly correlation score, encouraging results are obtained concerning the frequency with which the degree of local skill reaches a potentially useful level, and the prospects for predicting this skill in advance. However, further analysis using alternative skill scores is required to confirm these results.

1. Introduction

It is well known (eg Lorenz, 1982), that there is no prospect of making skilful predictions of instantaneous weather patterns beyond a range of two weeks, due to the growth with time of small analysis errors. However on certain occasions (Mansfield, 1986; Miyakoda et al, 1983), skilful forecasts of temporally or spatially filtered fields are possible on the monthly time scale using general circulation models (GCMs). Such filtering methods isolate the larger, more predictable scales of motion (Shukla, 1981). Since the extent of low frequency variability depends on geographical position (Blackmon et al, 1977), extended-range predictability may be higher in some regions than others.

The influence of tropical sea surface temperature (SST) anomalies on the tropical and extratropical atmospheric circulation has been demonstrated (eg Horel and Wallace, 1981). Significant improvements in forecast skill in tropical regions have been obtained through the inclusion of observed SSTs (Miyakoda et al, 1986a; Owen and Palmer, 1987), although for extratropical regions results are less clear cut (Mansfield, 1986; Cubasch and Wiin-Nielsen, 1986; Miyakoda et al, 1986a).

In this paper only mid-latitude predictability is considered. The integrations reported all use climatological SST, so any forecast skill arises purely from the initial conditions. The technique of ensemble forecasting involves running a number of forecasts from slightly different initial states, each compatible with the errors associated with the latest available analysis. This enables any skilful signal to be separated from unpredictable noise arising from imprecise knowledge of the initial state. Results from 'perfect model' experiments show that improved extended-range skill, relative to that of individual forecasts, can be attained through ensemble-averaging (Seidman, 1981; Hoffman and Kalnay, 1983; Murphy, 1988). Recent 'real data' experiments suggest that some degree of increased skill can also be realised in practical predictions (Miyakoda et al, 1986b; Molteni et al, 1986; Murphy, 1988). In addition the last two studies, and also Kalnay and Dalcher (1987), show that, at least locally, ensemble spread may provide a useful a priori prediction of forecast skill.

The members of an ensemble forecast may be assumed to represent equally probable evolutions of the atmosphere. An important question in this context is whether the atmosphere obeys a unimodal, near normal distribution of states, or resides mainly in one or other of a limited number of quasi-stable flow regimes, with intermittent rapid transitions between them. Reinhold (1987) gives a discussion of relevant work. Evidence of bimodality in the space-integrated large-scale wave amplitude has been found by Benzi et al (1986), although the distribution of zonal wind values was unimodal. Benzi et al show that modification of the Charney and de Vore (1979) resonance theory of multiple equilibria, through the inclusion of some non-linear effects, produces results consistent with these observations. If a view of the general circulation based on flow regimes is appropriate, ensemble forecast distributions may sometimes exhibit clustering (Murphy and Palmer, 1986), each cluster representing a different regime, rather than taking a Gaussian form.

Since the skill of medium and extended-range forecasts varies considerably from case to case (Bengtsson and Simmons, 1983; Mansfield, 1986; Molteni et al, 1986), the ability to predict skill in advance is of paramount importance. The use of ensemble spread is one possible approach. In addition, relationships between the skill of medium-range forecasts, and diagnostics such as the amplitude and persistence of the large-scale flow, measures of local synoptic activity, or certain characteristic patterns such as Wallace and Gutzler's (1981) PNA teleconnection pattern, have been examined by Hollingsworth et al (1985), Branstator (1986), and Palmer and Tibaldi (1986), with promising results.

Major experimental extended-range forecast projects are now underway at several centres (eg Tracton, 1986), encouraged by the research outlined above. (See Hollingsworth et al (1987) for a comprehensive review of topics in extended-range mid-latitude prediction.) In this paper results are presented from recent experiments conducted at the UK Meteorological Office (UKMO) using a global 11-level GCM. A large sample of individual 40-day integrations is used to determine the model's mean forecast skill, and the variability thereof, for different seasons of the year (section 4). The GCM's climate drift is discussed briefly in section 3, and its effect on skill is subsequently assessed. Some consideration is given to the general problem of forecast verification, relating to the contribution to the measured skill arising from differences between the true atmospheric climate, and the background normals used in practical assessment. Finally, in section 5 the model's skill in the UK area is compared with that of the experimental monthly long-range forecasts produced at UKMO, to assess the likely impact on current extended-range forecast procedures of regular input from monthly GCM integrations.

Study of these individual forecasts also provides a suitable background for the future assessment of the practical impact of ensemble forecasts, which is also a central part of the UKMO project.

2. Model and experiments

The model is a global, 11-level grid-point GCM with a latitude/longitude grid of resolution 2.5 x 3.75°. Described in detail in Slingo (1985), it is primarily designed for long climate integrations, and is therefore considered suitable for extended-range forecast experiments, although it cannot match the short-range skill of a high resolution numerical weather prediction model. Particular features include an interactive radiation scheme (incorporating the seasonal cycle) with predicted clouds, and a parametrization of gravity wave drag (Palmer et al, 1986), put in to ease the problem of excessive winter-time westerly flow in northern mid-latitudes (see Slingo and Pearson, 1987).

Table 1 shows the initialisation dates of each of the sixty-four 40-day integrations considered in this paper. The same version of the model was used in all experiments. Each forecast used climatological SSTs, updated every 5 days during the integration. Initial data was taken from the years 1982-7, with 16 forecasts for each season. The dates were centred around the appropriate solstice or equinox with a maximum of four dates per season per year, any pair of dates in the same year being at

least 10 days apart (with one exception where the gap was 9 days). No attempt was made to pre-select dates likely to yield forecasts of anomalously high skill. The integrations used initial data interpolated from UKMO operational analyses, apart from a few cases (asterisked in Table 1), where the data was obtained from the European Centre for Medium Range Weather Forecasts (ECMWF).

Most of the results presented subsequently, including the calculation of model climate drift, are based on the 48 forecasts marked A in Table 1. The remaining runs (marked B) are used to provide an independent test of the reliability of the climate drift estimates.

In addition to these experiments, a set of ten 7-member extended-range ensemble forecasts have been produced along the lines of the case discussed by Murphy and Palmer (1986), using the lagged-average forecast technique of Hoffman and Kalnay (1983), with initialisation dates chosen to correspond with parallel work at other centres. It is intended to continue producing these ensemble forecasts at a rate of one every three months, and results will appear in a future paper.

3. Model climate drift

(a) Systematic error

In general systematic errors (SE) develop during a GCM integration initialised from real data. The model climate, determined from a large number of runs started from independent initial conditions, 'drifts' away from the corresponding observed climate towards its own internal statistical equilibrium. The SE in the model forecasts must be corrected to maximise their usefulness. The best way, of course, is to improve the GCMs to remove the shortcomings which cause the development of the SE in the first place, and much work is already being done in this area (see Hollingsworth et al, 1987). However, until such errors can be totally eradicated, schemes for mitigating the effects of the SE will be required, based on empirical statistical corrections. Recent work by Miyakoda et al (1986b) and Molteni et al (1986) has shown that improvements in skill can be obtained, by subtracting estimates of the SE in the mean flow from the forecasts prior to verification, and some relevant results are given in the following section. Firstly, a brief illustration of the model's SE is given using 500 mb geopotential height (H500) maps.

In Figures 1a-d the mean forecast flow for the A integrations, averaged over days 16-30, is shown along with the corresponding mean observed pattern, and the difference between them, for each season. Heights are lower than observed in most areas, although the mean latitudinal gradient is quite realistic in all seasons. However areas of large difference are apparent in each case, caused by errors in the positioning of the jet stream, or by troughs and ridges being wrongly positioned or of incorrect amplitude. For example, in common with other GCMs, the amplitude of the winter-time Rockies ridge is underestimated, resulting in a large area of negative difference centred to the north of Alaska. Two major areas of difference in the spring patterns lie in the eastern Atlantic, where the jet stream is too far south, with exaggerated diffluence to the north, and over the east coast of Russia, where a deep trough and

downstream ridge are positioned too far east. In the summer the sharp trough at 75°E has no counterpart in the observed flow, and the area of strongest flow over North America and the Atlantic lies too far south. Errors of similar magnitude occur in the autumn pattern.

Since they are based on a finite number of cases, the observed differences may be partially attributable to random forecast errors, rather than genuine climate drift. Accordingly, Figure 1 also shows maps of the statistical significance of the differences, determined in each case by performing a t-test at each grid point on the two samples of twelve forecast and observed fields. The stippled areas show regions within which the difference is everywhere significant at the 5% level, assuming 12 combined degrees of freedom. This value is used, rather than 22, because integrations from the same year for a given season are not totally independent of one another (Table 1). The choice is made on the basis that runs started 30 days apart are independent, and those from intervening dates provide a small degree of extra information.

Ironically, the results are then liable to underestimate the significance for a different reason. The t-test assumes that the two samples are selected independently of each other. However, in practice they are identical at forecast day zero, and can only be taken as totally independent once all forecast skill has disappeared. In fact, the total significant area is always well above the chance level of 5% in any case, averaging 27% over the four seasons. Therefore, although the inherent spatial correlations make it difficult to assess overall pattern significance from a field of univariate tests (Livezey, 1985) it appears reasonable to regard the difference maps as reliable estimates of the model SE.

Corresponding maps for days 1-15 (not shown) give broadly similar difference patterns, although the intensity of the features is generally lower, reflecting the development of the SE between the two periods.

(b) Low frequency variability

Another important aspect of the model's climate is the extent to which it reproduces the low frequency variability of the atmosphere, since the prediction of such variations is the crux of the extended-range forecasting problem. Figures 2a and b show the standard deviation, over the A integrations, of the 15 day mean forecast H500 field for days 16-30, along with that of the verifying observed fields, for winter and summer. Differences occur in the patterns of variability for both seasons. For example, in winter the areas of high variability situated near Iceland, and off the east coast of North America, are not captured by the model. Point-by-point variance ratio tests, assuming 6 degrees of freedom for each variance, suggest that the differences are not significant, subject to the aforementioned caveat. Therefore, although the model underestimates variability in these experiments over the blocking-prone areas (Blackmon et al, 1977), it is not certain that this is a genuine feature of the model climate. Nevertheless, Table 2 shows clearly that the spatially-averaged variance is underestimated, particularly in summer. The same is true of spring and autumn.

Even if the atmospheric patterns in Figure 2 are successfully reproduced, there are other aspects of low frequency behaviour which must be modelled correctly, such as the occurrence of regime behaviour and transitions between circulation types (see introduction). Whilst it is not intended here to discuss such questions in detail, it is worth noting a simple indicator of the model's ability to produce changes in its large-scale circulation during a forecast. Figure 3 shows, as a measure of persistence, the anomaly correlation between the 15 day mean of forecast days 1-15 and subsequent overlapping 15 day means, averaged over the A integrations. Anomalies were calculated relative to normals based on the years 1951-80, with the estimate of the appropriate seasonal model SE, calculated as in Figure 1 as a function of forecast time, subtracted from each integration. The mean model persistence exceeds the corresponding atmospheric persistence, calculated in the same way from the verifying observations. Thus the model tends to retain the long-wave pattern generated over the medium-range part of the forecast to an excessive degree.

No model can predict, with complete certainty, a fundamental shift in the general circulation which occurs at extended-range. Nevertheless, the probability distribution of possible outcomes may in principle be correctly predicted by ensemble methods given, say, a GCM capable of modelling regime behaviour realistically. Figure 3 suggests that the present model may not entirely fulfil this requirement. If high frequency variability is crucial in triggering regime transitions (Reinhold, 1987), the model's deficiency in this respect (Slingo and Pearson, 1987), may be an important drawback.

4. Assessment of forecast skill

In this section forecast verification results are given for 15 day mean H500 and mean sea level pressure (PMSL) fields for 30-90°N, using the anomaly correlation score to measure skill.

(a) Choice of climatology

Ideally, forecast and observed anomalies should be formed relative to the true atmospheric climate appropriate to the experiment year. Unfortunately the true climate, at any point in time, is unknown. Its precise determination would require an infinite ensemble of Earths, each of whose atmospheres were subject to identical external forcing conditions. In practice anomalies are calculated relative to a set of atmospheric normals, constructed by averaging observed data from a number of years prior to the experiment year. In general these differ from the true climate for two reasons. Firstly, the true climate may itself vary with time (ie systematic climate change), and secondly the normals are subject to sampling error, being formed from a finite body of data. These errors (the systematic element is so termed for the sake of convenience), appear in both the forecast and observed anomalies, and thus increase the measured correlations.

A simple calculation illustrates this point. Let f_i and a_i represent, for some experiment, the forecast and corresponding observed anomaly relative to the true climate, at some grid point i. The model SE is

assumed to have been removed, although the basic argument does not depend on this. The anomaly correlation coefficient, c_{true}, relative to the true climate, is given by

$$c_{true} = \overline{a_i f_i}/(w_f w_a),^{1/2}$$

where $\overline{\ }$ denotes an area-weighted spatial average over the domain of interest, $w_f = \overline{f_i^2}$ and $w_a = \overline{a_i^2}$. However, the correlation that is measured in practice, c_{meas}, is

$$c_{meas} = \overline{(a_i+e_i)(f_i+e_i)}/[\overline{(a_i+e_i)^2}\ \overline{(f_i+e_i)^2}]^{1/2},$$

where e_i is the total (systematic + random) error incurred in estimating the true climate using a given set of normals. In general e_i will vary with position, but its mean size is characterised by $w_e = \overline{e_i^2}$. The mean contribution to c_{meas} from this source is

$$\langle c_e \rangle = w_e/[(\langle w_a \rangle + w_e)(\langle w_f \rangle + w_e)]^{1/2},$$

using $\langle\ \rangle$ to denote an average over many independent forecasts. $\langle c_e \rangle$ may be recognised as the mean level of correlation which would be obtained by verifying each forecast against observed data chosen at random from the distribution of states appropriate to the true climate.

If the assumption is made that $\langle w_f \rangle = \langle w_a \rangle$, it follows easily that

$$\langle c_{true} \rangle = (\langle c_{meas} \rangle - \langle c_e \rangle)/(1 - \langle c_e \rangle) \quad \ldots \quad (1)$$

In fact this result is not very sensitive to the ratio $\langle w_f \rangle/\langle w_a \rangle$. Thus, whilst the (closely-related) variances discussed in section 3 suggest that $\langle w_f \rangle/\langle w_a \rangle$ varies between about 0.7 and unity for the fields of interest, depending on season, the error incurred in the estimation of $\langle c_{true} \rangle$ by assuming a value of unity is entirely negligible.

Equation (1) quantifies the effect of the errors e_i in increasing $\langle c_{meas} \rangle$ relative to $\langle c_{true} \rangle$. Also, the change in $\langle c_{meas} \rangle$ arising from the use of a different set of normals may be derived from equation (1), giving

$$\langle \delta c_{meas} \rangle = [(1 - \langle c_{meas} \rangle)/(1 - \langle c_e \rangle)]\ \langle \delta c_e \rangle \ldots . \quad (2)$$

In Figure 4 the time variation of $\langle c_{meas} \rangle$, obtained by averaging results from the 48 A integrations, is given for the H500 and PMSL fields. Alternative scores are shown corresponding to three different sets of atmospheric normals, based on the years 1972-81, 1951-70 and 1951-80. The model SE is removed as described in section 3, which is the case for all results in this and the following sub-section. Also plotted for each set of normals are values of $\langle c_e \rangle$, estimated by verifying each forecast against the observed data corresponding to one of the remaining eleven forecasts from the same season, initialised on the same day and month in a different year. It is assumed that the true climate is stationary over the years

containing the forecast initialisation dates. Note that the years of the experiments are not included in the normals, since to do so would decrease the measured skill unfairly.

Significant positive values of $\langle c_e \rangle$ are observed, which are above the level which would occur if the e_i were due purely to sampling error. (Assuming $\langle w_f \rangle = \langle w_a \rangle$, the expected value of $\langle c_e \rangle$ would be $(1+N)^{-1}$ for N-year normals.) Values of $\langle c_{meas} \rangle$ and $\langle c_e \rangle$ are a little higher for H500 than for PMSL. 1951-80 normals give the lowest scores, the values for 1951-70 being slightly higher, and those for 1972-81 a little higher still. Values of $\langle c_{meas} \rangle$ and $\langle c_e \rangle$ for 1977-81 normals (not shown) are typically 0.05 higher than those for 1972-81. The correspondence between the variations in $\langle c_{meas} \rangle$ and $\langle c_e \rangle$ is well represented by equation (2).

Figure 4 also shows the time variation of $\langle c_{true} \rangle$, derived using equation (1) from the observed values of $\langle c_{meas} \rangle$ and $\langle c_e \rangle$. (The derived values are identical, whichever pair of curves (a), (b), (c) or (d) are substituted into equation (1).) Even for 1951-80 normals, the difference between $\langle c_{meas} \rangle$ and $\langle c_{true} \rangle$ is considerable, so that whilst $\langle c_{meas} \rangle$ remains substantially positive throughout the forecast period, $\langle c_{true} \rangle$ is small beyond about two weeks.

Thus the unambiguous definition of forecast skill is problematical. The point at which the difference between the forecast and observed states reaches saturation level corresponds to $\langle c_{true} \rangle = 0$, or $\langle c_{meas} \rangle = \langle c_e \rangle$, since the forecast is then no better than one made from an initial state selected at random from the true climate distribution. If this is taken as the absolute limit of predictability, and skill is to be measured relative to this point, then $\langle c_{true} \rangle$ is the score to use. However the true climate is an abstract concept, and in practice forecasts can only be verified as anomalies from observed normals. The increased correlation obtained in so doing is due to the model's ability, quantified by $\langle c_e \rangle$, to reproduce the true climate relative to that defined by the normals. (By correcting for the model SE using dependent data, some degree of 'cheating' has occurred in this respect (see later), but the principle is unaltered.) If this ability may be legitimately regarded as part of the forecast skill, then $\langle c_{meas} \rangle$ is the appropriate statistic. This is a moot point, since the contribution arising from $\langle c_e \rangle$ could equally well be obtained from, say, last year's forecast, or indeed last year's observations.

However, $\langle c_{meas} \rangle$ is clearly the better yardstick of practical forecast utility from the viewpoint of potential long-range forecast customers, since it measures skill relative to the perceived climate, upon which weather-sensitive decisions would be based in the absence of a suitable prediction method. The choice of normals for this purpose should be those which minimise $\langle c_e \rangle$, and hence $\langle c_{meas} \rangle$, since they are, by this criterion, the best available estimate of the true climate. The mean scores indicate that 1951-80 is the appropriate choice, and that the use of normals based on 10 years or less leads to an overestimation of forecast skill.

In principle, the dependence of $\langle c_{meas} \rangle$ on the choice of normals may itself vary with season. The difference between the scores for 1951-80 and 1951-70 normals is quite small in all seasons (Figure 5). However, the difference between 1951-80 and 1972-81 is somewhat greater in autumn and

particularly spring, but is reversed in winter, the scores for 1972-81 normals being a little lower. Nevertheless, the experimental results presented in the following sections all refer to 1951-80 normals, regardless of season.

(b) Variability of skill

Figure 6 shows, for PMSL, the seasonal variation in $\langle c_{true} \rangle$ compared with the corresponding variation in $\langle c_{meas} \rangle$. Compared with summer and autumn, winter and spring show higher values of $\langle c_{true} \rangle$ up to days 16-30, the differences being especially large for days 1-15 and 6-20. For these two periods values of $\langle c_{meas} \rangle$ are also higher, although the differences are smaller due to opposite seasonal variations in $\langle c_e \rangle$.

The overall variability among the 48 A integrations is shown in Figure 7, which gives histograms of c_{meas} for 15 day mean PMSL fields of days 1-15 and 16-30. The top category ($c_{meas} > 0.6$) represents forecasts showing sufficient skill to be classified as 'useful'. The second category ($0.2 < c_{meas} \leq 0.6$) is intended to encapsulate forecasts retaining a recognisable element of skill, which is insufficient to render them useful over the entire domain. The remaining cases are considered to have lost predictability. The criterion of 0.6 is taken from Hollingsworth et al (1980), who applied it to daily fields. Such a value is perhaps harsh for time averaged fields, since the same degree of fine detail may not be required. Hence the definition of useful skill, although valuable as a means of classifying the forecasts, should not be interpreted too literally.

For days 1-15 all the forecasts fall into the top two categories, with just under two thirds achieving the higher level. This includes 21 of the 24 winter and spring cases, whereas only nine of the remaining runs qualify. In the second half month only one forecast achieves this level, although the majority (66%) still attain the second category. For these runs the prospects for extracting useful information probably depend on the ability to identify which regions of the forecast domain are likely to be skilful in any particular case.

The time variation of skill for selected individual forecasts, relative to the appropriate seasonal mean, is given in Figure 8. The criterion for selection, that the skill should remain either above or below the seasonal mean over the whole forecast period, was fulfilled in only nine cases. Thus only occasionally can a forecast be classified unambiguously as good or bad throughout. However, seven of the nine cases are in the good category, suggesting that on around 15% of occasions unusually skilful hemispheric 40-day predictions are possible.

(c) Effect of climate drift

If the model SE is not removed prior to verification, the values of $\langle c_{meas} \rangle$, $\langle c_e \rangle$ and $\langle c_{true} \rangle$ are consistently smaller (Figure 9), since the SE increases the value of w_f, the forecast anomaly intensity, without on average affecting the covariance between the forecast and observed anomalies. The extent of the difference, similar for H500 and PMSL, is substantial throughout the forecast period. It is sufficient to decrease

the $\langle c_{meas} \rangle$ value by over 50% at days 16-30 and beyond in the case of PMSL. The seasonal breakdown (Figure 10), shows that the difference is greatest for summer and autumn cases, thus the gap in skill between winter/spring and summer/autumn is accentuated. The difference in winter is somewhat smaller than that reported by Miyakoda et al (1986b), whose model did not include gravity wave drag. Also, the number of forecasts achieving the top category of skill for days 1-15, and the second category for days 16-30, is significantly reduced (Figure 11, cf Figure 9), showing the importance of removing the SE to maximise forecast usefulness.

However the scores for the corrected forecasts may be too optimistic, since the SE was not calculated from independent data. This was tested by verifying the 16 independent B integrations, both with and without the SE, as deduced from the A integrations, removed. Figure 12 compares the scores for the B and A integrations (copied from Figure 9). The skill of the uncorrected B integrations is somewhat lower than that of the uncorrected A integrations. The mean skill of the B integrations is increased by removing the SE, but the average increase over all time levels, compared with that obtained using dependent data, is only 40% for PMSL, and 57% for H500, suggesting that the results for the corrected A integrations are indeed artificially high. However, in section 3 it was argued that the estimates of the model SE are statistically reliable, ie relatively uncontaminated by residual random forecast errors.

A possible means of reconciling these results is to postulate that the model SE is a function of flow regime. Then the mean SE, whilst accurate for the set of integrations from which it is deduced, may not be representative of the climate drift observed in an independent set of forecasts. It is beyond the scope of this paper to investigate this idea in detail, but some supporting evidence is given in Figure 13, which shows the seasonal variation, for the B integrations, of the effect of removing the SE. In winter the PMSL scores are actually made slightly worse, whereas in autumn a substantial improvement is observed. If, for example, the flow regimes occurring in the winter of 1985-86 were not represented in the winters of the A integrations, with the converse true for autumn 1986, these results could be understood in the light of the above hypothesis.

Thus the scores for the corrected A integrations may be taken as an upper skill limit, appropriate if the variation of the model SE with flow regime is well documented. However, if only the mean model SE is known, the maximum skill attainable in practice may lie about halfway between the scores given for corrected and uncorrected A integrations.

5. Skill for UK area

This section compares the model's skill in a localised area centred on the UK (see Figure 14), to that of the experimental long-range forecasts for one month ahead, performed every half-month at UKMO, hereafter referred to as 'the issued forecasts'. The regular inputs to these forecasts (Folland and Woodcock, 1986), consist of the latest available operational 5- and 10-day forecasts from UKMO and ECMWF respectively (usually day 2 of these runs corresponds to day 1 of the forecast period), and certain statistical techniques. The dominant statistical method is a multivariate forecasting technique (MVA), described by Maryon and Storey (1985). This

uses linear discriminant equations, determined from historical data, to predict mean PMSL patterns over the North Atlantic and Europe for the two half-months of the forecast. The predictors are the strengths of a set of northern hemisphere covariance eigenvector patterns of recently observed half-monthly mean PMSL and 1000-500 mb thickness, and regionally-averaged SST anomalies over a number of ocean areas for the month preceding the forecast. The predictands are the probabilities of each of a set of six pre-defined, season-dependent, 'cluster' patterns of PMSL. The patterns of the most likely clusters can be adjusted in a second, regression stage of the method.

Once the final forecast PMSL patterns have been determined from the above inputs, temperature and rainfall predictions for various UK districts are derived, using objective regression relationships, analogue techniques and SST information.

Although extended-range dynamical predictions are occasionally available (Murphy and Palmer, 1986), they have not made a significant contribution to date, so the issued forecasts represent an appropriate baseline, relative to which the impact of regular operational extended-range GCM forecasts may be judged.

The skill of the issued forecasts for temperature, rainfall and PMSL is reported in detail by Folland et al (1986), hereafter referred to as F. Here only PMSL is considered, comparing the A integrations to the 96 issued forecasts produced during the period February 1983-January 1987. The A integrations are taken from the period April 1982-October 1985.

(a) Comparison of mean skill

Figure 15 shows the average anomaly correlation of the A integrations and the issued forecasts, measured relative to 1951-80 normals, for the area shown in Figure 14, calculated from pressure values at the 25 marked points. Scores are also given for a smaller area framing the UK itself, based on the six points marked x. The averaging periods of days 1-5, 6-15 and 16-30 represent the three ranges into which the issued forecasts are normally split, and results are also shown for the first half-month and the monthly mean. Of the two regions, the 6 point area gives the fairer comparison, since the prime concern in producing the issued forecasts is to maximise PMSL forecast skill over the UK itself, as the final temperature and rainfall forecasts are restricted to this area. However in terms of the mean anomaly correlation scores for the issued forecasts, there is actually little difference between the two areas for days 1-5, 6-15 or 16-30 (Figure 15).

The average skill of the uncorrected model forecasts exceeds that of the issued forecasts for all periods considered, except for the monthly mean over the 6 point area. Here the equality of the issued forecasts seems fortuitous, since the model scores are higher in both the constituent half-months. The superiority of the model is considerably enhanced when the SE is removed, although the scores for the corrected forecasts may be somewhat optimistic (section 4). Also, the model and issued forecast initialisation dates are not identical, which may prejudice the observed differences. Nevertheless beyond days 1-5, dominated by input from the

medium-range dynamical forecasts, the mean skill of the issued forecasts is very low (although F shows that for rainfall the skill is somewhat higher). The model forecasts apparently offer a clear improvement, particularly over days 6-15. For this period, over the 25 point area, the mean difference in skill between the model forecasts (with or without the SE removed) and the issued forecasts is significant at the 5% level, according to a t-test performed after transforming the correlations using Fisher's (1958) z-statistic, which renders their distributions more nearly normal (Branstator, 1986). The calculation assumes 23 and 47 degrees of freedom for the model and issued forecasts respectively, on the basis discussed in section 3. The differences are not significant for days 16-30, due to the large spread in the scores (see sub-section (b) below). The increased skill afforded by the model is also apparent from the mean r.m.s. error scores. For the corrected (uncorrected) model forecasts over the 6 point area, these are 4.02 (4.27), 6.87 (7.27) and 7.45 (7.81) mb for days 1-5, 6-15 and 16-30. The issued forecast scores are 4.90, and 8.45 and 7.83 mb respectively.

The production of the issued forecast for days 6-15 often involves the forecaster in subjective combination of the outlook from the medium-range dynamical forecasts, and the statistical predictions. Since their skill is so much greater (Figure 15), the availability of extended-range GCM forecasts would largely remove the need for such activity at this range. However, objective combination of ensembles of extended-range model integrations may lead to further improvements in skill (see introduction).

For days 16-30 the issued forecasts are usually based entirely on the statistical techniques, although the forecaster's interpretation is still necessary, for example, to produce a final best-estimate prediction from the probability forecast output of MVA. Since the model's skill is itself low at this range, the need for subjective or objective combination of dynamical and statistical predictions is likely to survive the introduction of regular monthly dynamical forecasts.

Measuring skill from the correspondence between spatial patterns, as above, may not be entirely appropriate for the user, who is probably more interested in the time series of weather at a point, or over a small area. Figure 16 shows time series of forecast and observed PMSL anomalies, averaged over the 6 point area for days 1-15 and 16-30, created by arranging the issued forecasts in chronological order along the abscissa. The graphs depict the ability of the operational system, based on forecasts made at half-monthly intervals, to predict the sequence of half-monthly mean anomalous weather at a range of one and two half-months ahead. Figure 17 shows analogous graphs for a pseudo time series of the model experiments, created by grouping them season by season, with the forecasts in each season in chronological order. Note the model's skill at days 1-15 in predicting large departures from normal, compared with the issued forecasts. Loss of signal is apparent at days 16-30 for both prediction systems. Figure 18 gives the correlation between the time series of forecast and observed anomalies for each of the graphs shown in Figures 16 and 17. Also shown are corresponding correlations for analogous sequences based on days 1-5, 6-15 and 1-30. The model score exceeds the issued forecast score for all periods, the difference being greatest for days 6-15.

The time series correlation score, for both the corrected model and issued forecasts, exceeds the corresponding mean spatial anomaly correlation score for days 1-5, for both the 6 and 25 point areas. For days 6-15 and 16-30 the difference between the two scores is generally small. However, for the uncorrected model forecasts the time series score is the larger for all periods. Surprisingly, the difference in skill between the uncorrected and corrected forecasts, apparent for spatial anomaly correla-tion, is almost completely eroded. A possible explanation follows from the fact that large anomaly cases contribute disproportionately to a time series correlation value. In most cases of large observed anomaly (Figure 17), the anomaly is negative. Since the mean model SE is also negative, removing it actually results in worse predictions in such cases, which may offset the positive effect of removing the SE in the remaining cases.

Time series correlations may also be calculated for point anomalies, as for the area-averaged anomalies discussed above. Figure 19 shows maps of such point correlations for days 1-15 and 16-30 of the corrected model forecasts. Some variation over the 25 point area is apparent, particularly for days 16-30, the area centred west of Ireland showing lower skill than elsewhere. The pattern of skill variation is not strongly related to the variation of the r.m.s. forecast or observed anomaly. The mean point-by-point correlation for the 25 point area exceeds the correlation for area-averaged anomalies for days 1-15, 16-30 and 1-30, the scores being 0.71 cf 0.68, 0.30 cf 0.19 and 0.56 cf 0.43 respectively. This result, which reflects the skill in predicting the anomaly pattern over the area, is encouraging, since the point-by-point score may be more relevant from the user's viewpoint.

(b) Variability of skill for 25 point area

Returning to the spatial anomaly correlation c_{meas}, Figure 20 shows the distribution of values for days 1-15 and 16-30 of the issued forecasts and the corrected and uncorrected A integrations. A wide spread of scores is apparent in each case, especially at days 16-30. The incidence of $c_{meas} > 0.6$ is greatest for the corrected model forecasts, with 40% of cases qualifying even at days 16-30, compared with only 15% for the issued forecasts. Although use of the 25 point area may penalise the issued forecasts unfairly (sub-section (a)), any such effect is probably small, given that the mean correlation for the 6 point area is similar for the periods considered. The larger area is preferred to make the distributions more meaningful in terms of forecast value — see below.

The model results suggest that useful local skill may exist on a substantial minority of occasions at extended-range, even if hemispheric or global skill is low. However, the anomaly correlation score is sensitive to anomaly magnitude and to phase differences (Arpe et al, 1985). Circumstances can easily be envisaged under which its value for a limited area is unrepresentative of a forecast's true worth. Indeed, the percentage of cases with $c_{meas} > 0.6$ for days 16-30 is almost as large when each forecast is verified against unrelated observed data, as described in section 4. Thus, some of the cases in the top category in Figure 20 may not be as useful as the score indicates. Equally, some forecasts with low

anomaly correlation may have greater value than the score suggests. Whilst the results remain encouraging, a more sophisticated verification scheme is required to establish accurately the frequency of occurrence of useful skill over limited areas. A possible candidate, not considered here, is a score based on the distance between the forecast and observed states in a probability space defined by the climatological distribution of observed values (see F, for example).

(c) Prediction of skill

The issued forecasts each carry a confidence level C (highest), D or E (lowest), for the month as a whole. This is determined subjectively, the major factor being the degree of consistency between the different forecast techniques, analogous to the use of ensemble spread in dynamical predictions. Figure 21 shows that the skill of those forecasts assigned confidence C (21% of the total), is somewhat higher than average, especially for the 6 point area, the better yardstick of performance for the issued forecasts. The largest improvement occurrs at days 6-15. Therefore, despite the low level of average skill, some degree of skill prediction is possible for the issued forecasts (see F).

Such an ability increases the utility of the forecasting system, accordingly an attempt was made to pick out skilful model forecasts. The criterion used was that the magnitude of the spatially averaged forecast anomaly, over the 25 point area, should be greater than 4.5 mb for days 1-15. This is a crude way of examining the possibility that cases showing stable anomalous features over the medium-range period of the forecast are more predictable. Eleven of the 48 cases qualify, and the average anomaly correlation is substantially higher in these forecasts at extended-range (Figure 22), the difference being particularly large for days 16-30. For days 1-15, ten of the eleven forecasts show $c_{meas} > 0.6$, dropping to six for days 16-30, (ie 55% cf 40% in Figure 20b). Also, Figure 23 shows that the skill of these forecasts compares favourably with that of the C confidence issued forecasts.

To some extent this result is built in by the dependence of anomaly correlation on anomaly magnitude. Certainly the average r.m.s. forecast error for the large anomaly cases is little different from that for all 48 cases. For days 1-15 and 16-30 the scores (large anomaly cases first), are 4.36 cf 4.67 mb, and 7.12 cf 6.98 mb respectively. However, if the presence of errors of a given size is less detrimental to the forecast quality in large anomaly cases (Branstator, 1986), then the increased anomaly correlation may be a reasonable reflection of greater forecast usefulness. Repetition of this analysis using a skill score based on probability distances, as suggested above, would prove illuminating in this regard.

The average score for a persistence forecast, based on the 15 day period immediately preceding the forecast, is very low in the large anomaly cases (Figure 22). Relative to this measure of persistence, the increase in anomaly correlation is therefore even more pronounced. It remains possible that the extra skill is due to persistence of patterns developed during the early stages of the forecast itself. However, even relative to for forecast days 1-15, persistence is generally lower in the large anomaly

cases, both in the model and the verifying observations, as shown by Figure 24. Thus the extra skill cannot be explained in these terms. Perhaps the model is better able to predict changes in the circulation at extended-range when its medium-range forecast is unusually accurate, as in the large anomaly cases.

Although a definitive statement requires a more detailed analysis, these results provide some encouragement concerning the prospects for local skill prediction in a GCM-based extended-range forecasting system. A vital element of such a system would be the ability to recover predictions of temperature and rainfall of equal skill to those of the circulation patterns considered here.

6. Conclusions

A large sample of 40-day integrations of a global 11-level GCM, using climatological boundary conditions, has been used to assess the prospects for dynamically-based extended-range forecasting.

Seasonal estimates of the model's systematic error (SE), each based on 12 integrations, show that significant climate drift occurs in all seasons, which should be corrected empirically to obtain maximum forecast skill. Low frequency variability is also underestimated, although there are insufficient experiments to determine precisely the patterns of model variability. The degree of persistence apparent in a model forecast, relative to its first 15 days, is on average greater than in the verifying observations, suggesting that the model has difficulty in generating regime transitions, possibly due to its lack of sufficient high frequency variability.

Atmospheric normals, based on data from past years, differ in general from the true climate for the years of the experiments. Since the difference appears in both the forecast and verifying anomaly fields, an extra positive contribution to the observed forecast anomaly correlation scores results, the magnitude of which depends on the choice of normals. The mean score which would be obtained relative to the true (but unknown) climate can be estimated from the experimental results, and is found to be small beyond a range of two weeks. However, the mean observed correlation is 0.3 or greater at days 16-30 and beyond. This residual may be attributed to the model's capacity to reproduce the true climate relative to the normals in use.

In practice skill must be measured relative to the best available climate, namely the set of normals which minimises the mean observed correlation. For these experiments, normals based on the years 1951-80 give lower scores than a number of alternative choices based on subsets of this period.

Seasonal variation in skill is observed, winter and spring giving the best results. For days 1-15, 60% of all forecasts exceed the 0.6 level of anomaly correlation, however for days 16-30 most of the scores lie between 0.2 and 0.6. At this range the identification of locally skilful areas is a particularly high priority, since only a few forecasts give anomalously high hemispheric skill throughout the forecast period.

The above results were all calculated with the model SE removed. Omission of this correction degrades the scores significantly. Results from independent cases show smaller differences between the corrected and uncorrected forecasts, suggesting that the above scores are somewhat optimistic. However, the reduced differences in the independent cases may be partially attributable to dependence of the SE on flow regime.

The model's average skill in predicting surface pressure one month ahead, over a limited area centred on the UK, exceeds that of the experimental long-range forecasts issued at UKMO, which are based on medium-range dynamical forecasts and statistical prediction techniques. The major improvement occurs at days 6-15, and is enhanced considerably by removal of the model SE (subject to the above remarks). The model's average skill is low at days 16-30, however its superiority at days 6-15 is in principle sufficient to render other methods largely obsolete at this range.

The model's skill also exceeds that of the issued forecasts, in terms of the ability to reproduce observed sequences of time-averaged anomalies, obtained by arranging forecasts in chronological order. In this respect the model's score generally exceeds its corresponding mean spatial anomaly correlation score, particularly when the time series refer to point anomalies, rather than area-averaged values.

Over the limited area, 40% of the (corrected) model forecasts show anomaly correlations in excess of 0.6 at days 16-30, suggesting that useful local skill is possible on a relatively high proportion of occasions at extended-range. Also, the anomaly magnitude is a useful predictor of forecast anomaly correlation. However, both results are influenced by the sensitivity of this skill score, which is not ideally suited to limited area verification. Further investigation involving alternative measures of skill is required, before firm conclusions can be drawn on these topics.

References

Arpe, K., Hollingsworth, A., Tracton, M.S., Lorenc, A.C., Uppala, S., Kallberg, P. 1985 The response of numerical weather prediction systems to FGGE Level IIb data. Part II: Forecast verifications and implications for predictability. Quart. J. Roy. Met. Soc., 111, 67-102.

Bengtsson, L., Simmons, A.J. 1983 Medium-range weather prediction — operational experience at ECMWF. In Hoskins, B.J. and Pearce, R.P. (eds); Large-scale dynamical processes in the atmosphere. London, Academic Press.

Benzi, R., Malguzzi, P., Speranza, A., Sutera, A. 1986 The statistical properties of general atmospheric circulation: observational evidence and a minimal theory of bimodality. Quart. J. Roy. Met. Soc., 112, 661-674.

Blackmon, M.L., Wallace, J.M., Lau, N.C., Mullen, S.L. 1977 An observational study of the northern hemisphere wintertime circulation. J. Atmos. Sci., 34, 1040-1053.

Branstator, G., 1986 The variability in skill of 72-hour global-scale NMC forecasts. Mon. Wea. Rev., 114, 2628-2639.

Charney, J.G. deVore, J.G. 1979 Multiple flow equilibria in the atmosphere and blocking. J. Atmos. Sci., 36, 1205-1216.

Cubasch, U., Wiin-Nielsen, A.C. 1986 Predictability studies with the ECMWF spectral model for the extended range: The impact of horizontal resolution and sea surface temperature. Tellus, 38A, 25-41.

Fisher, R.A. 1958 Statistical methods for research workers. Hafner Publishing Co.

Folland, C.K., Woodcock, A. 1986 Experimental monthly long-range forecasts for the United Kingdom. Part I. Description of the forecasting system. Meteorol. Mag., 115, 301-318.

Folland, C. K., Woodcock, A., Varah, L.D. 1986 Experimental monthly long-range forecasts for the United Kingdom.
Part III. Skill of the monthly forecasts. Meteorol. Mag., 115, 377-395.

Hoffman, R.N., Kalnay, E. 1983 Lagged-average forecasting, an alternative to Monte Carlo forecasting. Tellus, 35, 100-118.

Hollingsworth, A., Arpe, K., Tiedtke, M., Capaldo, M., Savijarvi, H. 1980 The performance of a medium-range forecast model in winter — impact of physical parametrization. Mon. Wea. Rev., 108, 1736-1773.

Hollingsworth, A., Lorenc, A.C., Tracton, M.S., Arpe, K., Cats, G., Uppala, S., Kallberg, P. 1985 The response of numerical weather prediction systems to FGGE Level IIb data. Part I: Analysis. Quart. J. Roy. Met. Soc., 111, 1-66.

Hollingsworth, A., Cubasch, U., Tibaldi, S., Brankovic, C., Palmer, T. N., Campbell, L. 1987 Mid-latitude atmospheric prediction on time scales of 10-30 days. In 'Variability in the atmosphere and oceans', Roy. Met. Soc. Monograph.

Horel, J. D., Wallace, J. M., 1981 Planetary-scale atmospheric phenomena associated with the Southern Oscillation. Mon. Wea. Rev., 109, 813-829.

Kalnay, E., Dalcher, A., 1987 Forecasting forecast skill. Mon. Wea. Rev., 115, 349-356.

Livezey, R. E., 1985 Statistical analysis of general circulation model climate simulation: sensitivity and prediction experiments. J. Atmos. Sci., 42, 1139-1149.

Lorenz, E.N., 1982 Atmospheric predictability experiments with a large numerical model. Tellus, 34, 505-513.

Mansfield, D.A., 1986 The skill of dynamical long-range forecasts, including the effect of sea surface temperature anomalies. Quart. J. Roy. Met. Sec., 112, 1145-1176.

Maryon, R.H., Storey, A.M. 1985 A multivariate statistical model for forecasting anomalies of half-monthly mean surface pressure. J. Climatol., 5, 561-578.

Miyakoda, K.,Gordon, T.,Caverly, R., Stern, W., Sirutis, J., Bourke, W. 1983 Simulation of a blocking event in January 1977. Mon. Wea. Rev., 111, 846-869.

Miyakoda, K.,Sirutis, J.,Knutson, T. 1986a Experimental 30-day forecasting at GFDL. In ECMWF Workshop on predictability in the medium and extended range, 17-19 March 1986. ECMWF, Shinfield Park, Reading, UK.

Miyakoda, K.,Sirutis, J.,Ploshay, J. 1986b One-month forecast experiments — without anomaly boundary forcings. Mon. Wea. Rev., 114, 2363-2401.

Molteni, F., Cubasch, U., Tibaldi, S. 1986 30- and 60-day forecast experiments with the ECMWF spectral models. In ECMWF Workshop on predictability in the medium and extended range, 17-19 March 1986. ECMWF, Shinfield Park, Reading, UK.

Murphy, J.M. 1988 The impact of ensemble forecasts on predictability. To appear in Quart. J. Roy. Met. Soc.

Murphy, J.M., Palmer, T.N., 1986 Experimental monthly long-range forecasts for the United Kingdom.
Part II. A real-time long-range forecast by an ensemble of numerical integrations. Meteorol. Mag., 115, 337-349.

Palmer, T.N., Shutts, G.J., Swinbank, R. 1986 Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parametrization. Quart. J. Roy. Met. Soc., 112, 1001-1039.

Palmer, T.N., Tibaldi, S. 1986 Forecast skill and predictability. ECMWF Technical Memorandum No. 127, ECMWF, Shinfield Park, Reading, UK.

Owen, J.A., Palmer, T.N.	1987	The impact of El Nino on an ensemble of extended-range forecasts. Mon. Wea. Rev., 115, 2103-2117.
Reinhold, B.	1987	Weather regimes: the challenge in extended-range forecasting. Science, 235, 437-441.
Seidman, A.N.	1981	Averaging techniques in long-range weather forecasting. Mon. Wea. Rev., 109, 1367-1379.
Shukla, J.	1981	Dynamical predictability of monthly means. J. Atmos. Sci., 38, 2547-2572.
Slingo, A. (ed)	1985	Handbook of the Meteorological Office 11-layer atmospheric general circulation model. Volume 1: model description. Dynamical Climatology Technical Note No. 29, Meteorological Office, Bracknell, England.
Slingo, A., Pearson, D.W.	1987	A comparison of the impact of an envelope orography and of a parametrization of orographic gravity-wave drag on model simulations. Quart. J. Roy. Met. Soc., 113, 847-870.
Tracton, M.S.	1986	Activities and plans in dynamic extended range forecasting (DERF) at NMC. In ECMWF workshop on predictability in the medium and extended range, 17-19 March 1986. ECMWF, Shinfield Park, Reading, UK.
Wallace, J. M., Gutzler, D.S.	1981	Teleconnections in the geopotential height field during the northern hemisphere winter. Mon. Wea. Rev., 109, 784-812.

Table 1 Initialisation dates and grouping of model forecasts. Asterisked cases were initialised from 12Z ECMWF data, the remainder used midnight UKMO analyses.

Season	Initialisation date	Class of integration
Spring	5. 4.82 *	
	6. 3.83 *	
	16. 3.83 *	
	26. 3.83 *	
	5. 4.83 *	
	6. 3.84	A
	16. 3.84	
	26. 3.84	
	5. 4.84	
	6. 3.85	
	16. 3.85	
	26. 3.85	
	6. 3.86	
	16. 3.86	B
	26. 3.86	
	5. 4.86	
Summer	6. 6.83	
	16. 6.83	
	26. 6.83	
	6. 7.83	
	6. 6.84	
	16. 6.84	A
	26. 6.84	
	6. 7.84	
	6. 6.85	
	16. 6.85	
	26. 6.85	
	6. 7.85	
	5. 6.86	
	16. 6.86	B
	26. 6.86	
	6. 7.86	
Autumn	6.10.82 *	
	6. 9.83	
	16. 9.83	
	26. 9.83	
	6.10.83	
	6. 9.84	A
	16. 9.84	
	26. 9.84	
	5.10.84	
	6. 9.85	
	16. 9.85	
	26. 9.85	

	6. 9.86	
	16. 9.86	B
	26. 9.86	
	6.10.86	
	6.12.82 *	
	16.12.82 *	
	26.12.82 *	
	5. 1.83 *	
	6.12.83	
	16.12.83	A
	26.12.83	
Winter	5. 1.84	
	6.12.84	
	16.12.84	
	26.12.84	
	5. 1.85	
	6.12.85	
	16.12.85	B
	26.12.85	
	5. 1.86	

Table 2 Zonal mean values for winter and summer of 500 mb geopotential height variance (dam^2), over 12 model forecasts and verifying observations.

	Winter		Summer	
Latitude (°N)	Forecast variance	Observed variance	Forecast variance	Observed variance
80	63.9	77.7	9.3	46.2
70	76.0	93.0	13.6	24.2
60	70.1	66.7	11.2	18.9
50	51.1	55.0	11.1	20.4
40	32.1	47.2	6.4	10.1
30	13.2	14.0	2.7	3.3

Figure captions

Figure 1a — 15-day mean 500 mb geopotential height (dam) for days 16-30, averaged over 12 winter forecasts, with corresponding observed pattern and difference map. Point-by-point significance levels (%) of the difference are shown, as determined from a t-test assuming 6 degrees of freedom in each sample, with areas significant at the 5% level stippled.

Figure 1b — As Figure 1a for spring forecasts.

Figure 1c — As Figure 1a for summer forecasts.

Figure 1d — As Figure 1a for autumn forecasts.

Figure 2a — Standard deviation of 15-day mean forecast and observed 500 mb geopotential height (dam), among 12 winter cases, for days 16-30.

Figure 2b — As Figure 2a for summer forecasts.

Figure 3 — 48 case average persistence of 15-day mean forecast 500 mb geopotential height anomalies for 30-90°N, with respect to forecast days 1-15 (-·-·). Model systematic error has been removed. (····) shows the corresponding persistence in the verifying observations.

Figure 4a — 48 case average forecast anomaly correlation for 15-day mean 500 mb geopotential height, 30-90°N (——), and score for 'random' forecast (----), measured relative to normals from: (a) 1972-81; (b) 1951-70; (c) 1951-80; with model systematic error removed. The highest dashed curve corresponds to (a) and the lowest to (c). (····) shows derived mean score relative to the true climate.

Figure 4b — As Figure 4a for mean sea level pressure.

Figure 5a — Forecast anomaly correlation for 15-day mean 500 mb geopotential height, 30-90°N, averaged over 12 winter cases with model systematic error removed, and measured relative to normals from: (a) 1972-81; (b) 1951-70; (c) 1951-80.

Figure 5b — As Figure 5a for spring forecasts.

Figure 5c — As Figure 5a for summer forecasts.

Figure 5d — As Figure 5a for autumn forecasts.

Figure 6a — 12 case average forecast anomaly correlation for 15-day average mean sea level pressure, 30-90°N, with model systematic error removed.
(——) winter; (····) spring; (----) summer; (-·-·) autumn.

Figure 6b As Figure 6a for derived mean anomaly correlation relative to true climate.

Figure 7 Histograms of forecast anomaly correlation for 15-day average mean sea level pressure, 30-90°N, with model systematic error removed, for (a) days 1-15, (b) days 16-30, based on 48 forecasts. Each bin covers a correlation range of 0.4, centred about the value shown.

Figure 8 Forecast anomaly correlation for 15-day average mean sea level pressure, 30-90°N, with model systematic error removed, for selected individual forecasts, expressed as the departure from the relevant seasonal mean score. Forecast initialisation dates were:
(a) 00Z 5.1.84; (b) 12Z 6.3.83; (c) 00Z 26.6.83;
(d) 12Z 26.3.83; (e) 00Z 16.3.84; (f) 00Z 26.9.85;
(g) 00Z 16.6.84; (h) 00Z 6.7.85; (i) 00Z 5.4.84.

Figure 9a 48 case average forecast anomaly correlation for 15-day mean 500 mb geopotential height, 30-90°N (———), and score for 'random' forecast (----). (••••) shows derived score relative to true climate. In each case the upper curve shows the skill when model systematic error is removed, and the lower curve the skill of the uncorrected forecasts.

Figure 9b As Figure 9a for mean sea level pressure.

Figure 10a 12 case average increase in forecast anomaly correlation, for 15-day mean 500 mb geopotential height, 30-90°N, achieved when model systematic error is removed. (———) winter; (••••) spring;
(----) summer; (-•-•) autumn.

Figure 10b As Figure 10a for mean sea level pressure.

Figure 11 As Figure 7 for uncorrected forecasts.

Figure 12a Forecast anomaly correlation for 15-day mean 500 mb geopotential height, 30-90°N. Upper and lower curves show scores with and without model systematic error removed, averaged over: 48 class A integrations (———); 16 independent class B integrations (----). Systematic errors were calculated from the A integrations.

Figure 12b As Figure 12a for mean sea level pressure.

Figure 13 Average increase in forecast anomaly correlation for mean sea level pressure, 30-90°N, achieved when model systematic error is removed, for four independent cases. (———) winter; (••••) spring;
(----) summer; (-•-•) autumn.

Figure 14 Map showing the region, defined by the 25 marked points, used for limited area verification. Some results are also given for the inset region based on the six points marked x.

Figure 15 Average forecast anomaly correlation for mean sea level pressure over 6 and 25 point UK areas. Scores for 48 model forecasts, with and without systematic error removed, are compared with those for 96 issued forecasts, for various time-averaging periods.

Figure 16 Sequences of 15-day mean forecast (———) and verifying (----) mean sea level pressure anomaly for the issued forecasts, created by arranging them in chronological order along the abscissa. Anomalies are averaged over the 6 point UK area for: (a) forecast days 1-15; (b) forecast days 16-30.

Figure 17 As Figure 16 for a pseudo time series of the model forecasts (with systematic error removed), created by grouping them seasonally in the order winter, spring, summer, autumn, with forecasts in each season arranged in chronological order.

Figure 18 As Figure 15, except that skill is measured by the correlation between sequences of spatially-averaged forecast and verifying anomalies, created by arranging (model) issued forecasts in a (pseudo) time series, as in Figures 16 and 17.

Figure 19 Maps of point-by-point correlation, over the 25 point UK area, between time series of model forecast and observed anomalies, created as in Figure 17, for: (a) forecast days 1-15; (b) forecast days 16-30.

Figure 20a As Figure 7 for skill over 25 point UK area for forecast days 1-15, including: (a) 96 issued forecasts; (b) 48 model forecasts; (c) model forecasts with systematic error removed.

Figure 20b As Figure 20a for forecast days 16-30.

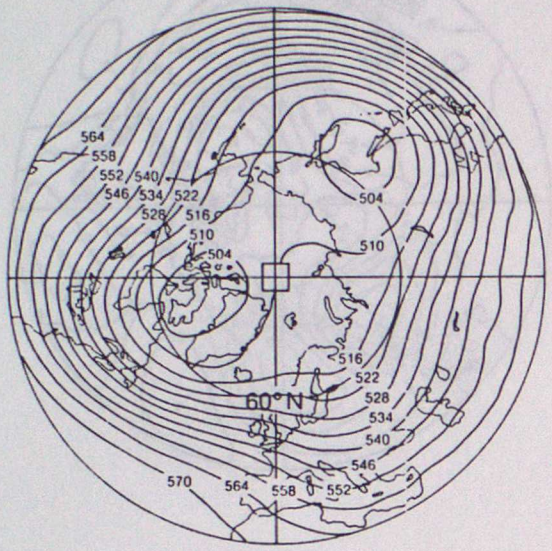
Figure 21 As Figure 15, but comparing average forecast anomaly correlation for all 96 issued forecasts with that for a subset of 20 cases assigned a confidence level of C.

Figure 22 Forecast anomaly correlation for 15-day average mean sea level pressure over 25 point UK area, averaged over 11 'large anomaly' model forecasts (———), compared with remaining 37 forecasts (••••). (----) shows the persistence score for the 15-day period immediately preceding and including the forecast initialisation date, averaged over the large anomaly cases, compared with the remaining 37 cases (-•-•).

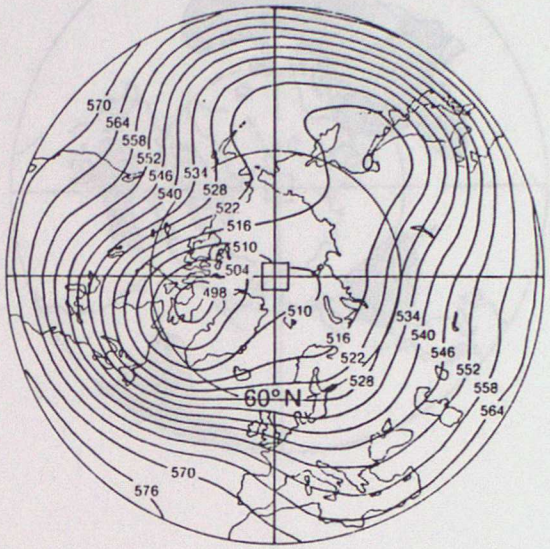
Figure 23 As Figure 15, but comparing average forecast anomaly correlation for 48 model forecasts, with systematic error removed, with that for the subset of 11 large anomaly cases.

Figure 24 Persistence of 15-day average mean sea level pressure anomalies over 25 point UK area with respect to forecast days 1-15. Scores given are for model forecasts (with systematic error removed) and verifying observations, averaged over: (a) 11 large anomaly cases, (———) and (----); (b) remaining 37 cases, (••••) and (-•-•).

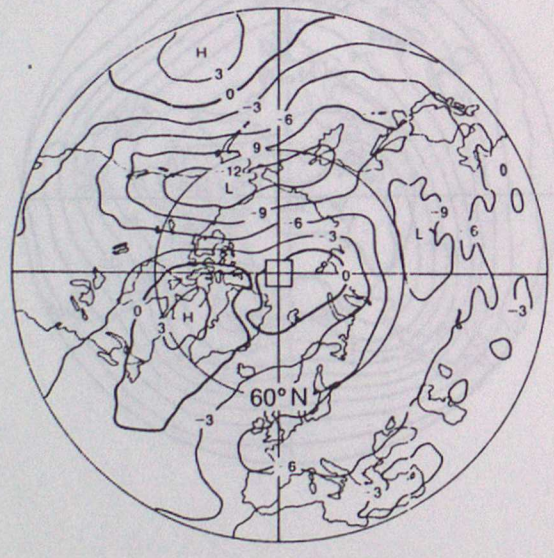
FIG 1a



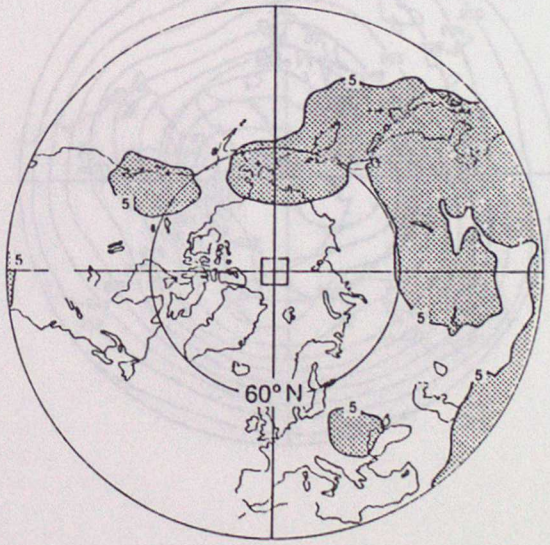
Forecast



Observed

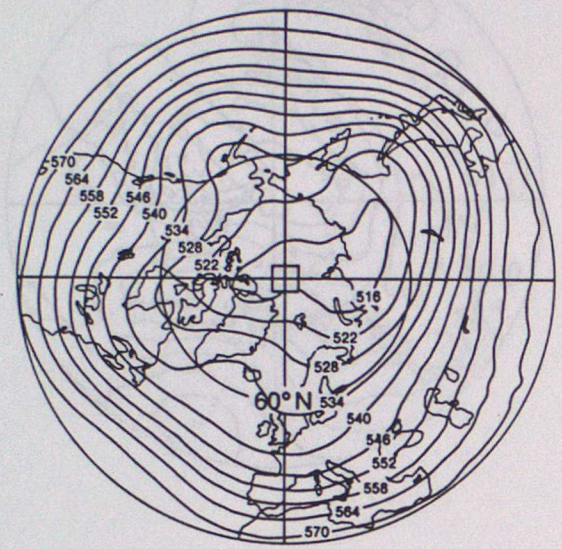


Difference

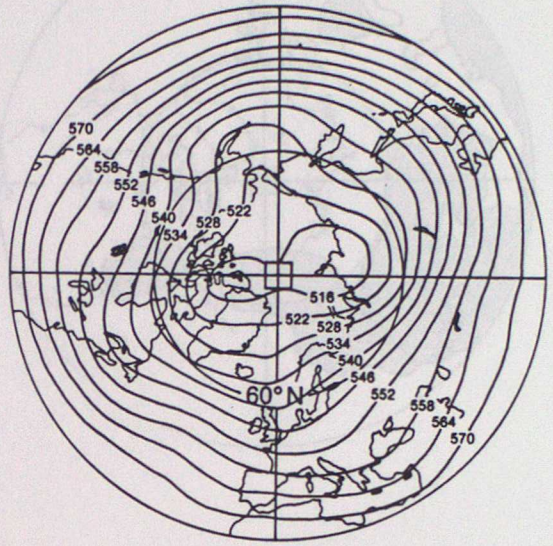


Significance level

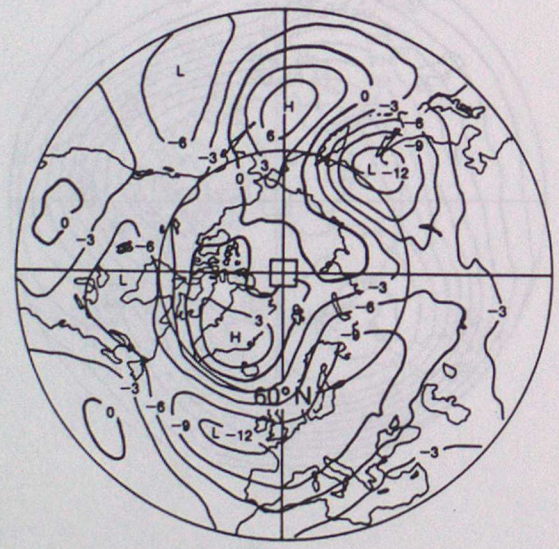
FIG 16



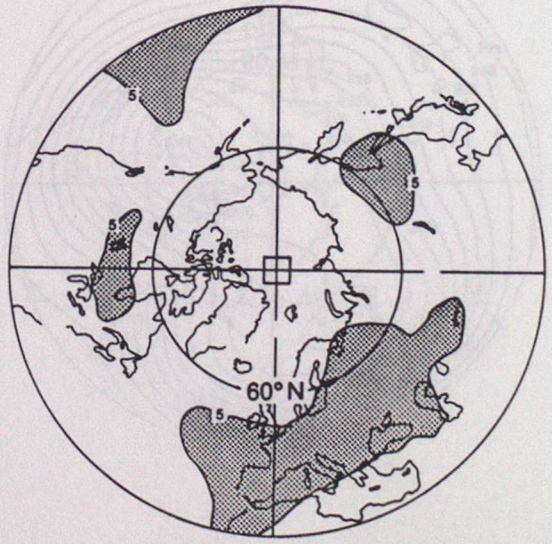
~~Forecast~~



~~Observed~~

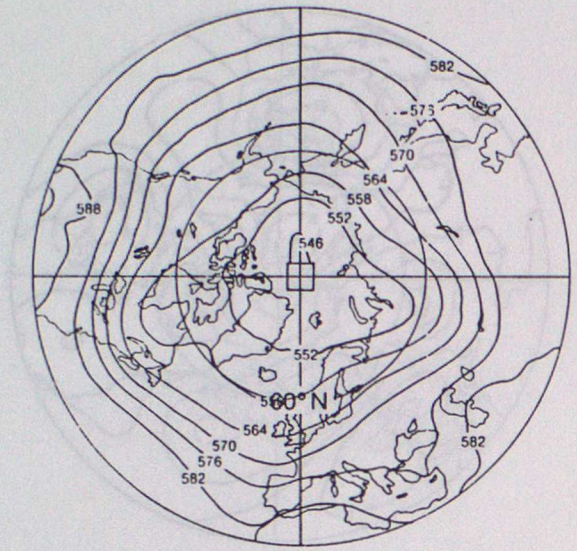


Difference

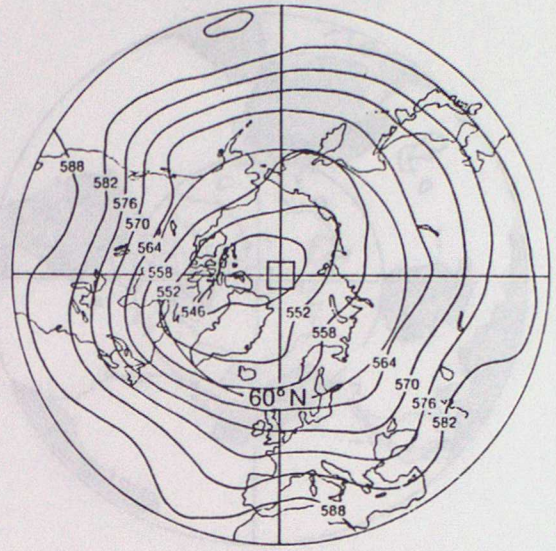


Significance level

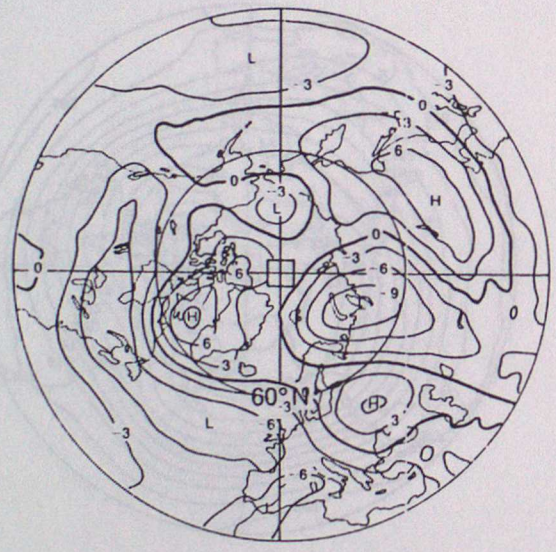
FIG 1c



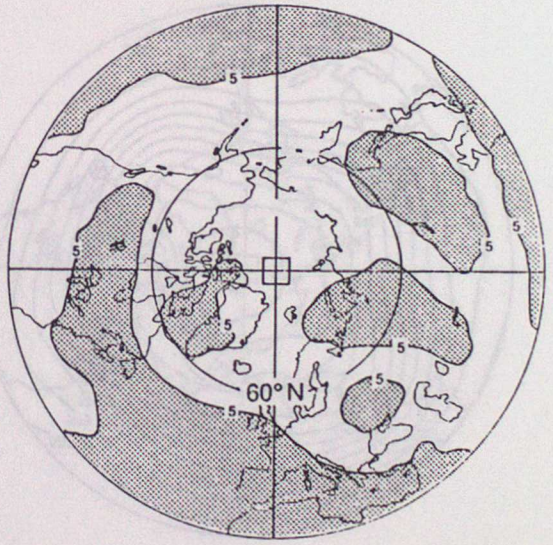
Forecast



Observed

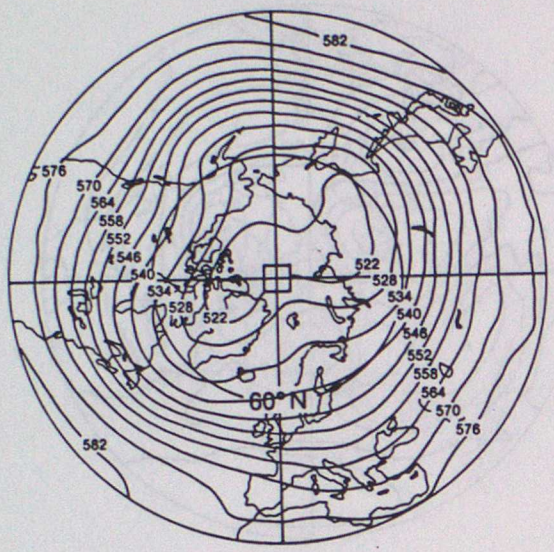


Difference

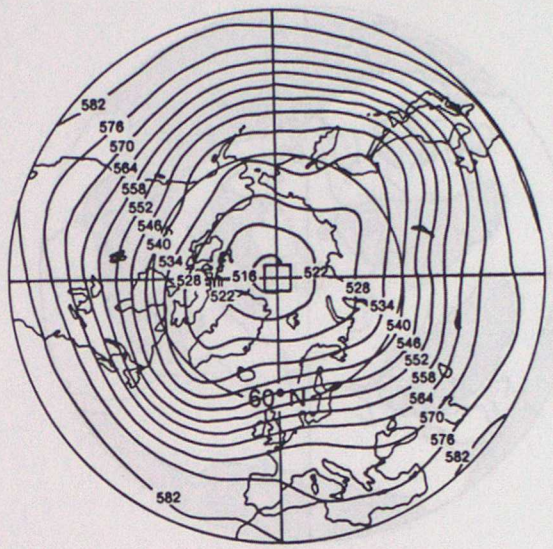


Significance level

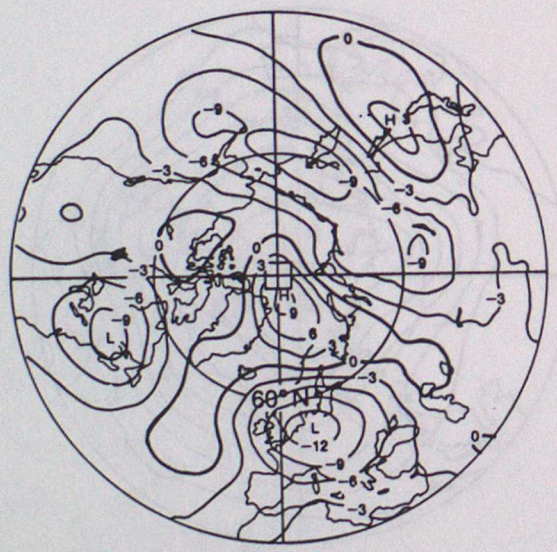
FIG 1d



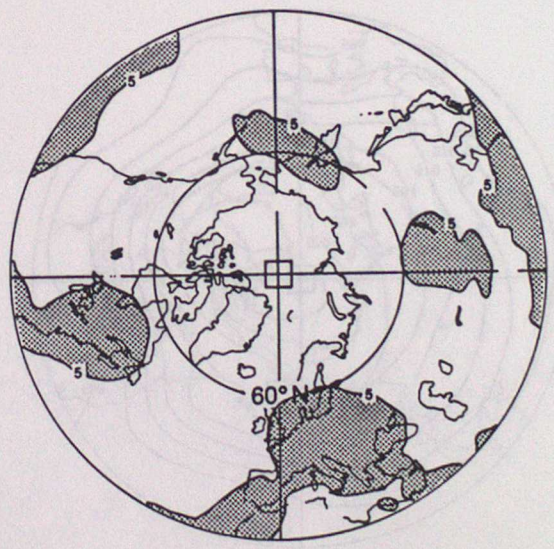
Forecast



Observed

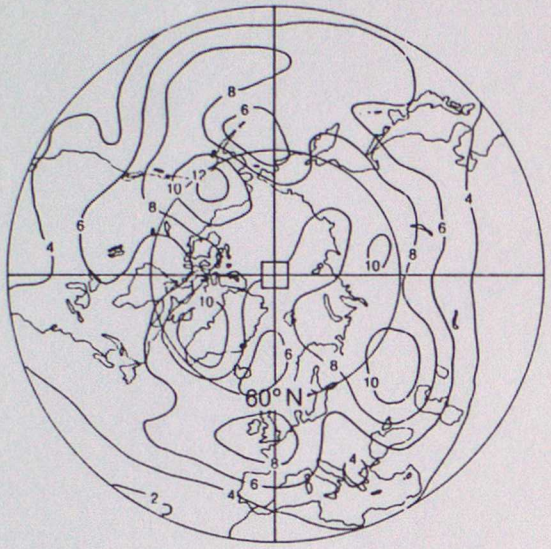


Difference

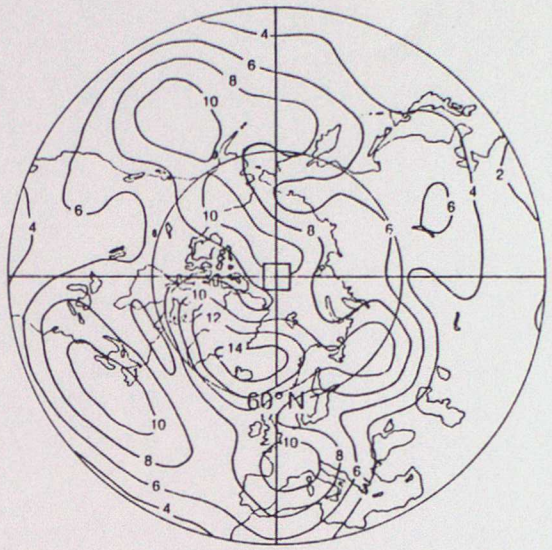


Significance level

FIG 2a

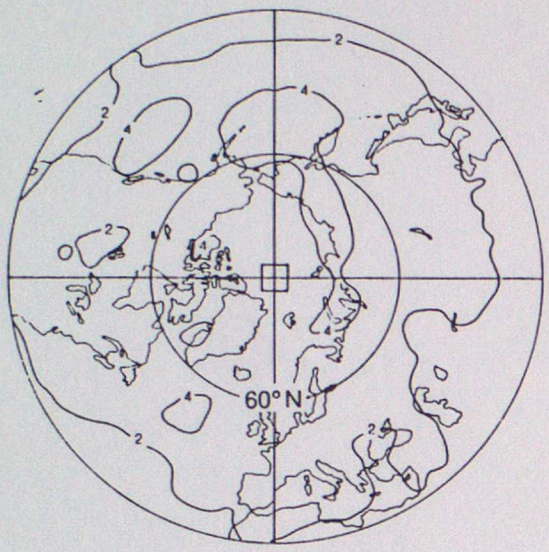


Forecast

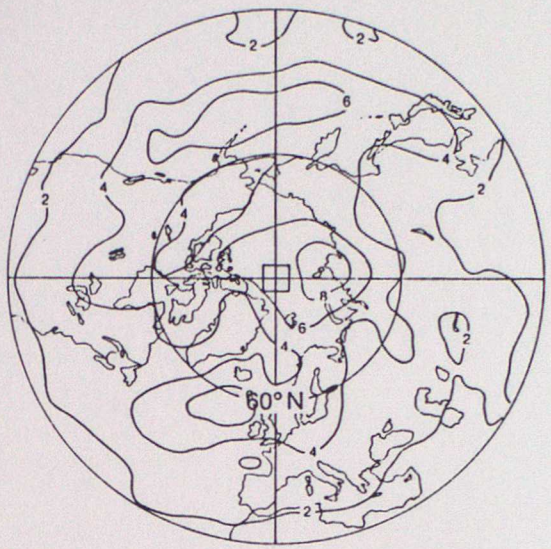


Observed

FIG 26



Forecast



Observed

FIG 3

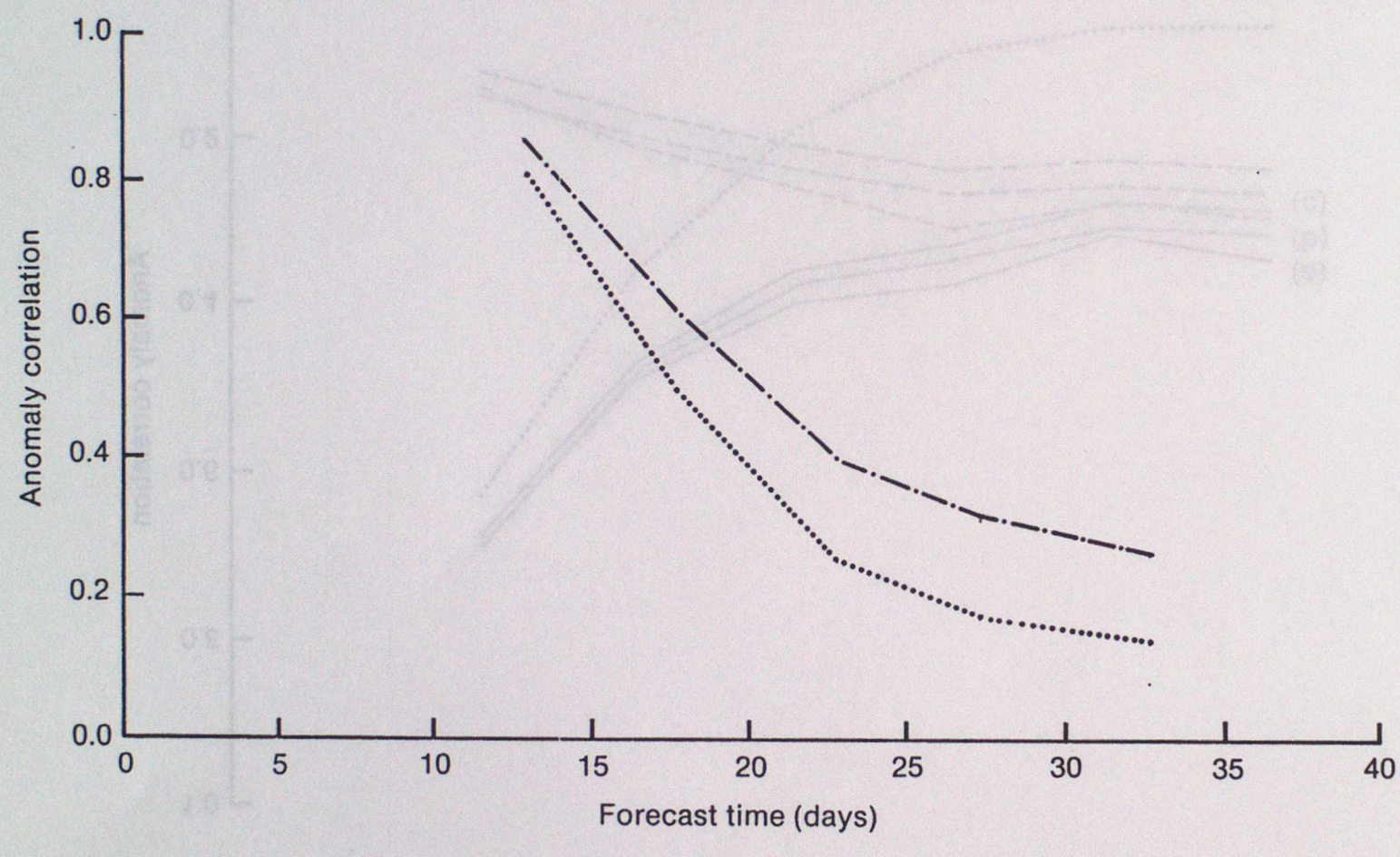

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 4.a

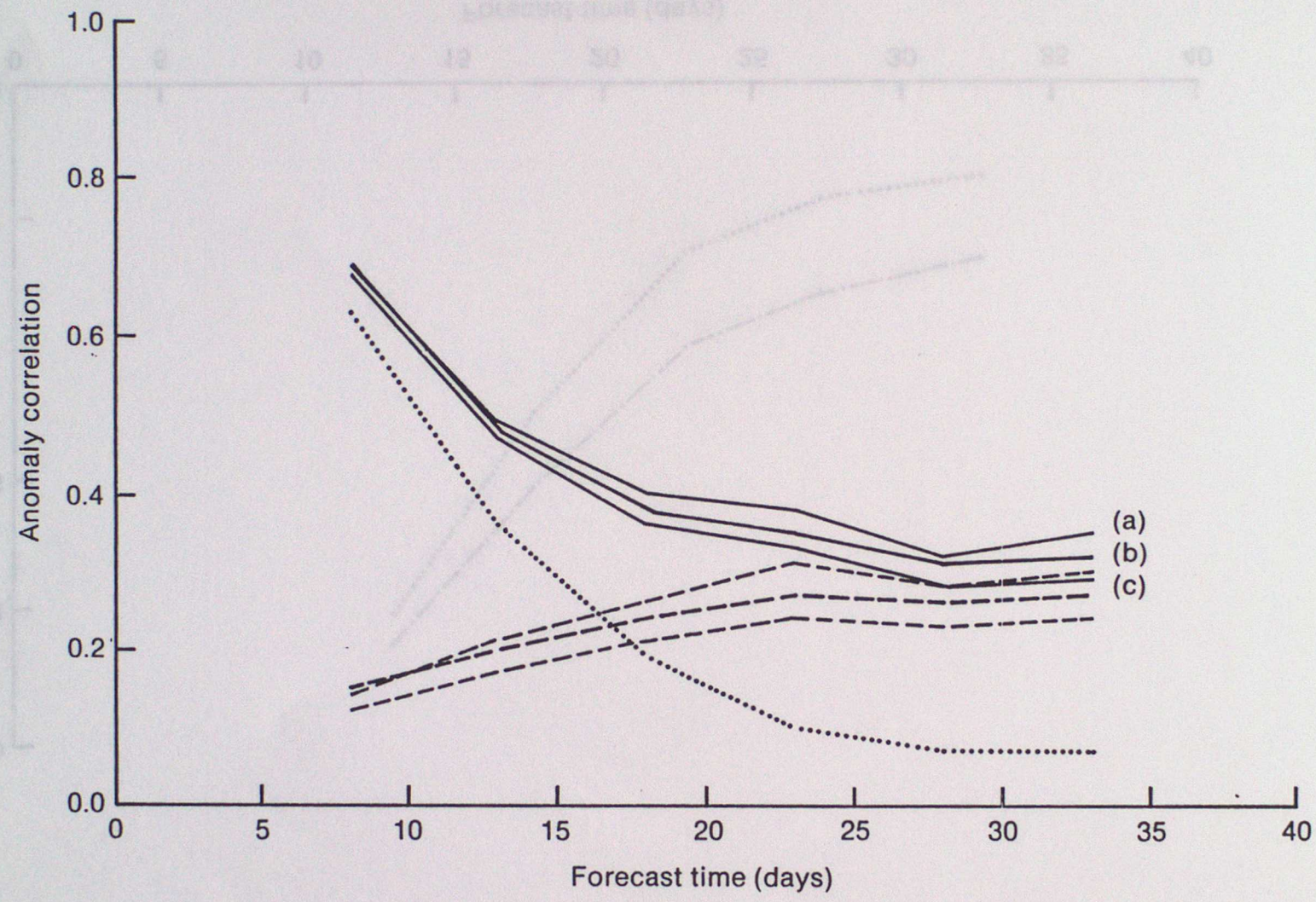

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)
(a)
(b)
(c)

FIG 46

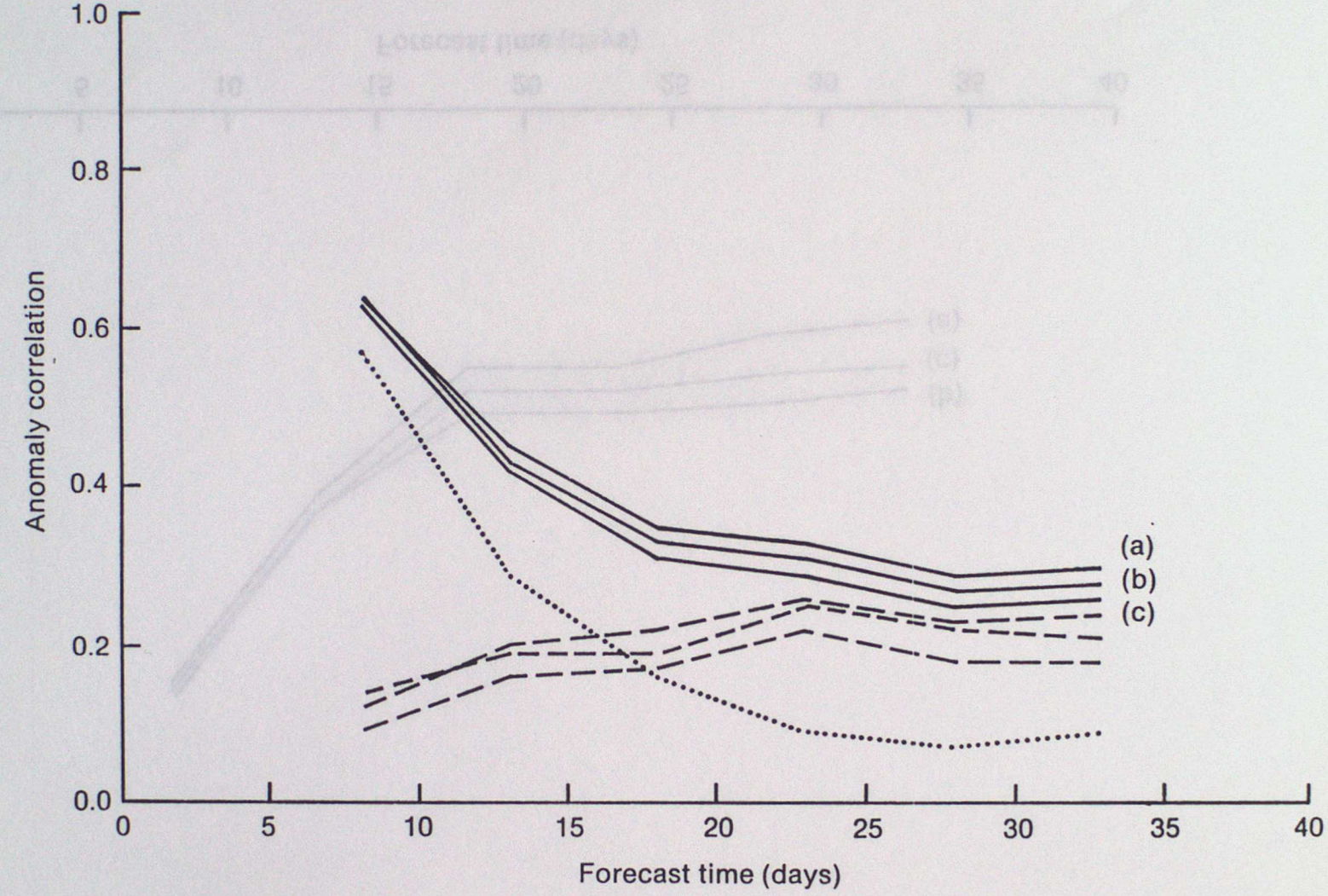

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)
(a)
(b)
(c)

FIG 5a

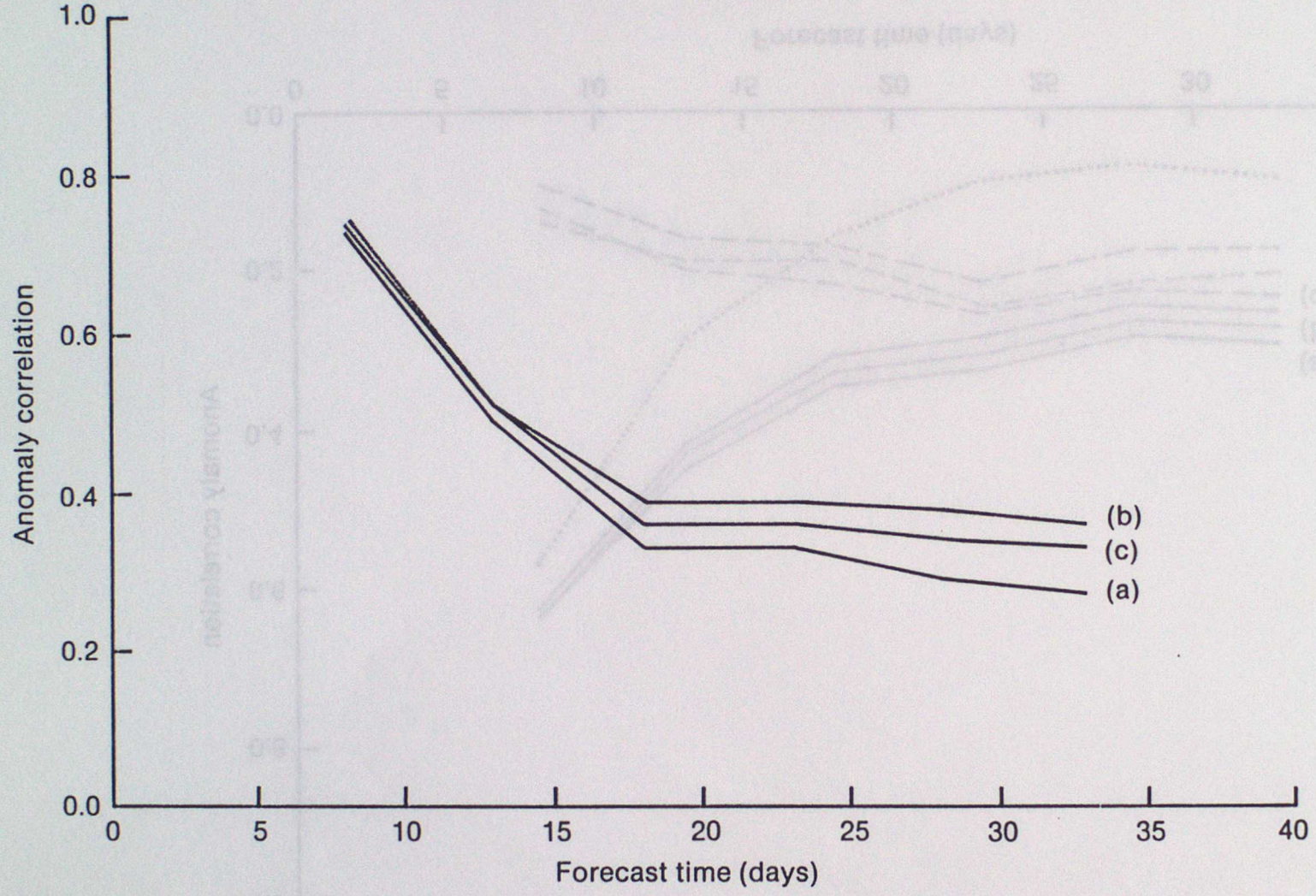

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)
(b)
(c)
(a)

FIG 56

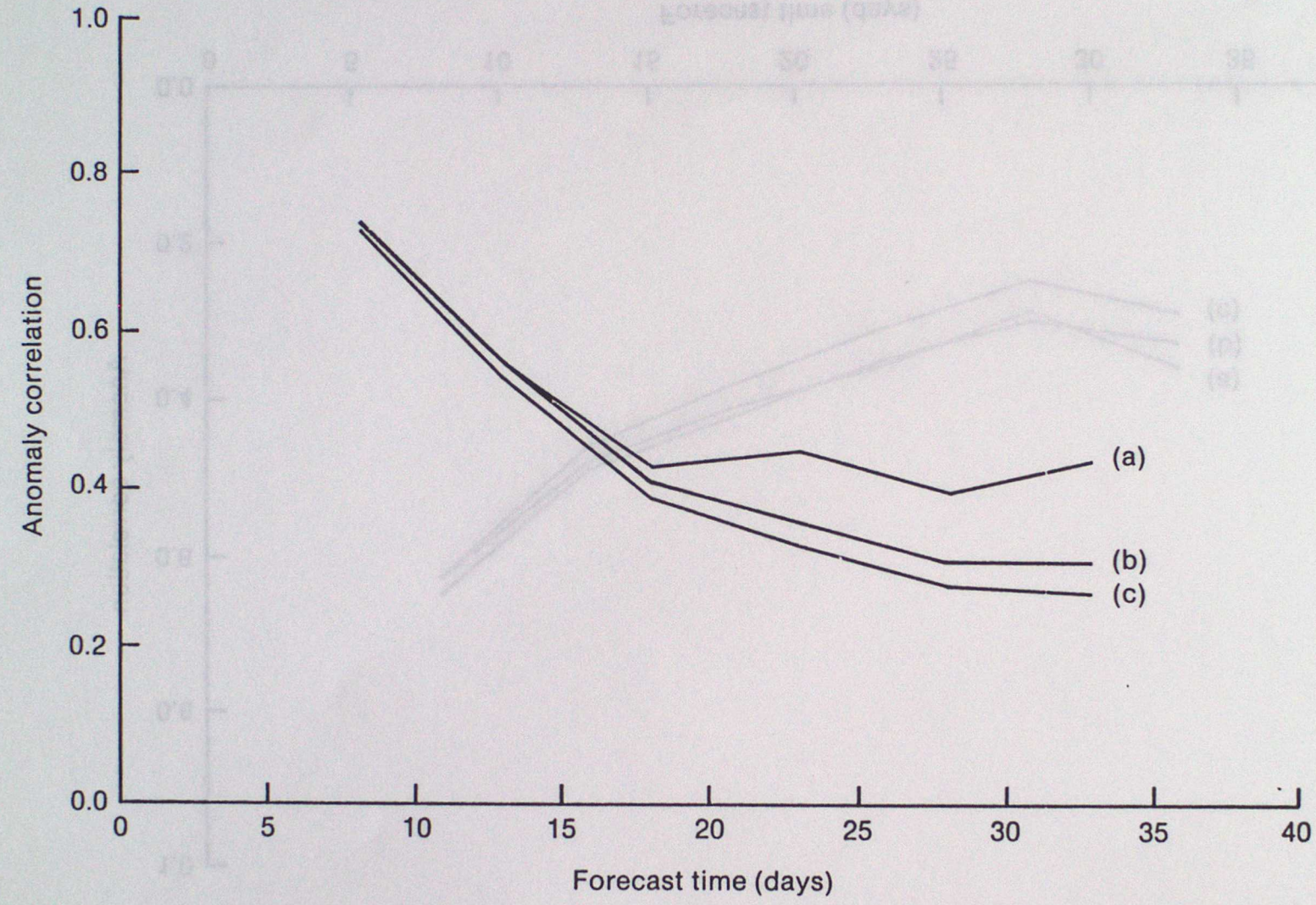

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)
(a)
(b)
(c)

FIG 5c

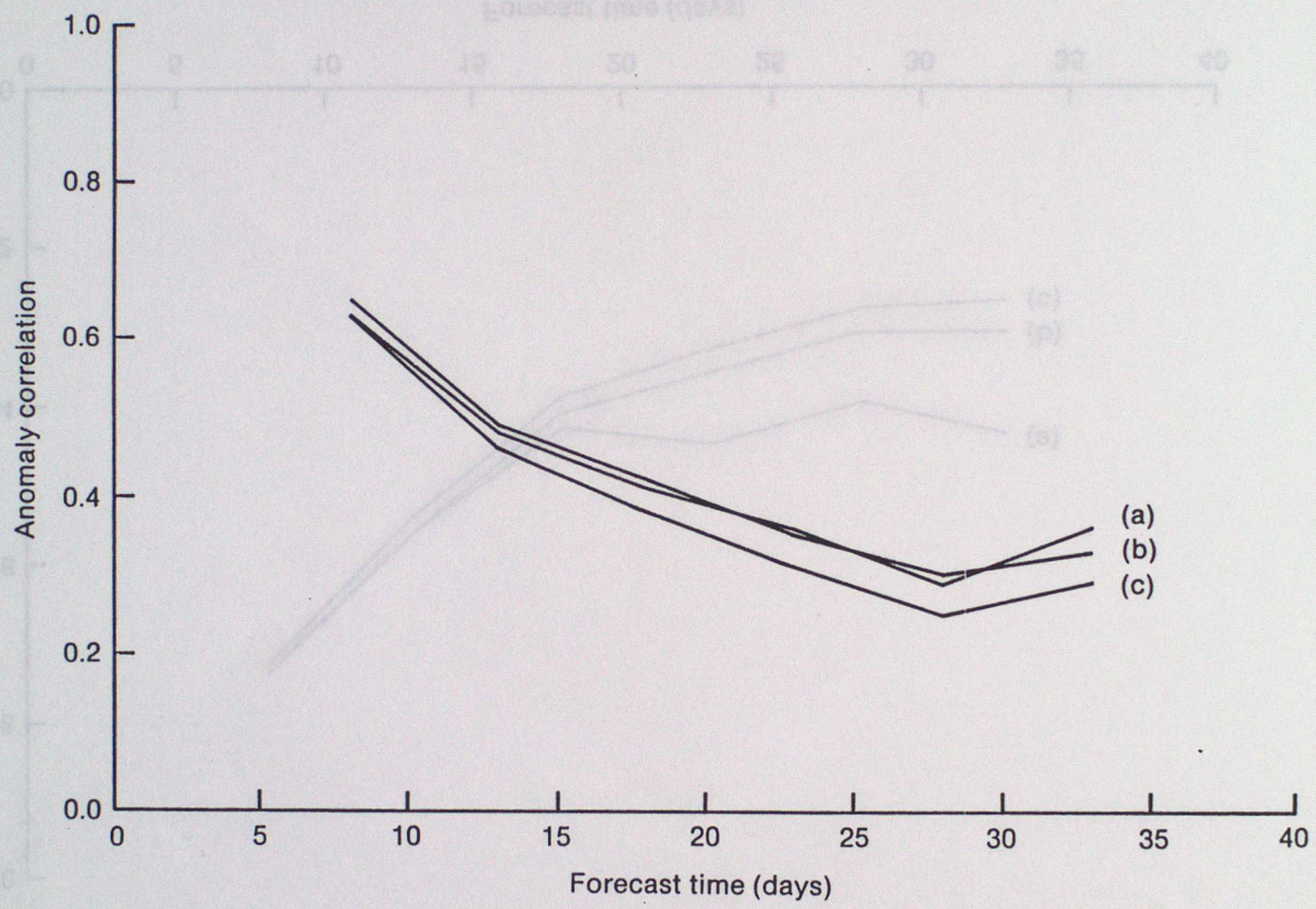

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)
(a)
(b)
(c)

FIG 5d

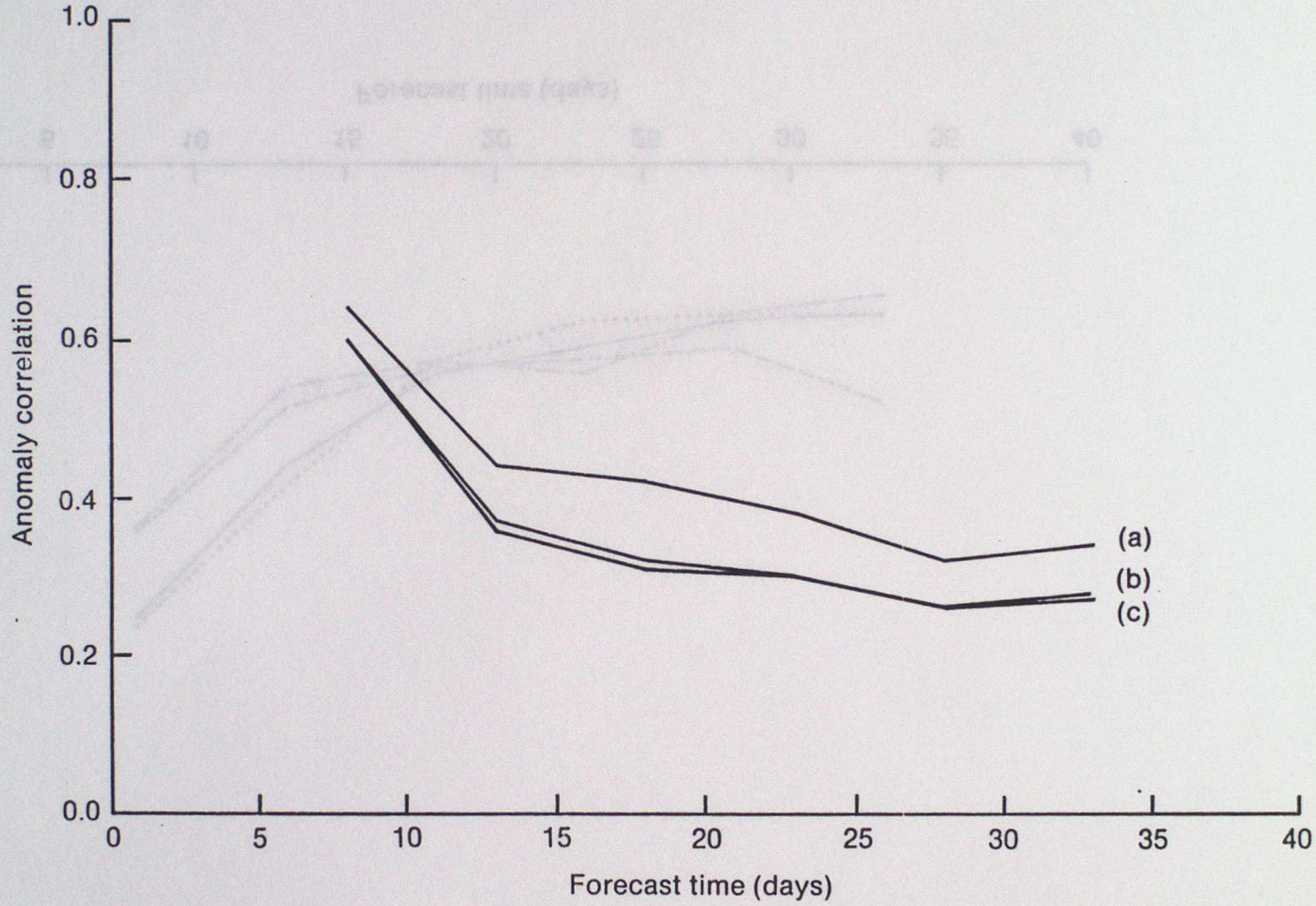

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)
(a)
(b)
(c)

FIG 6a

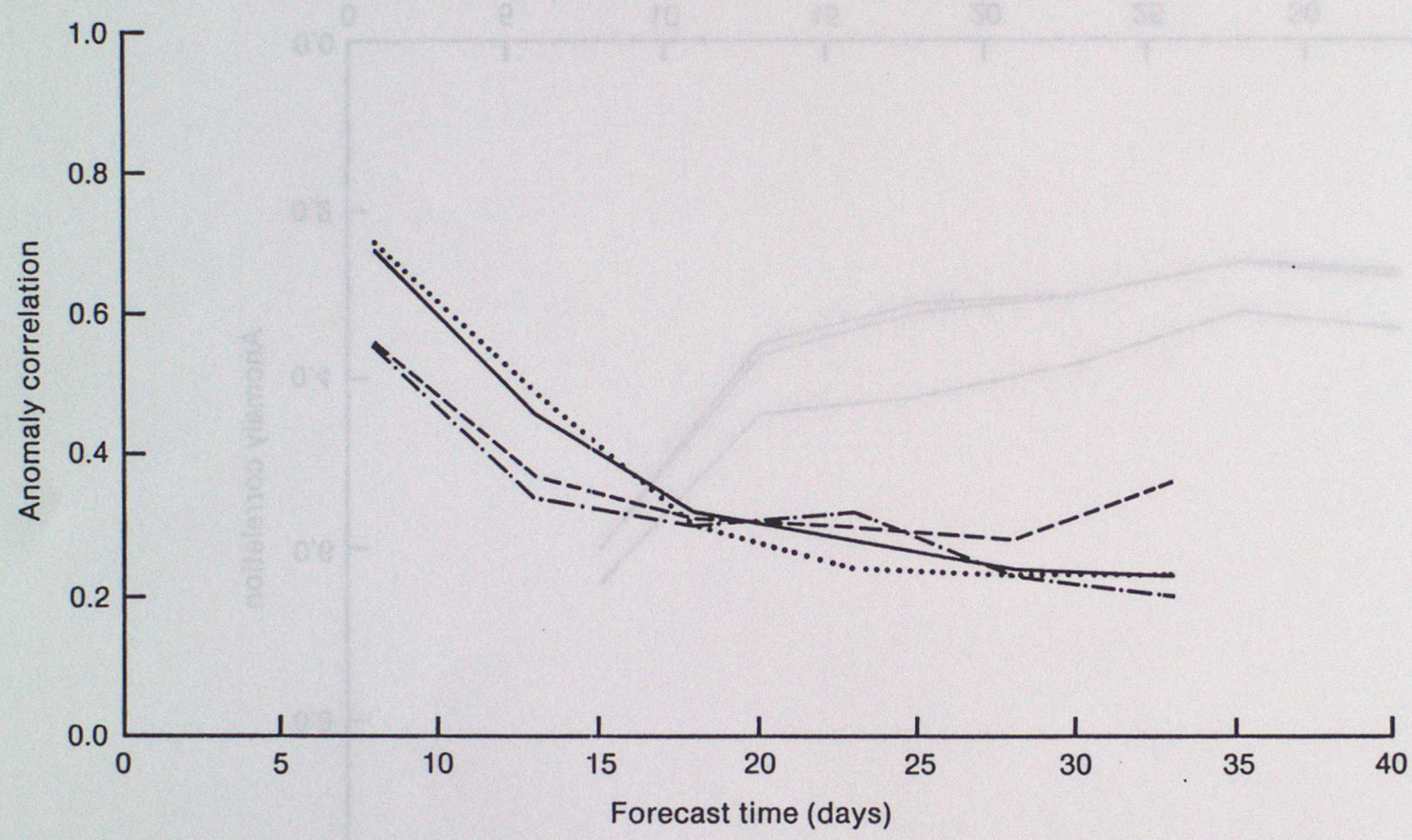

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 66

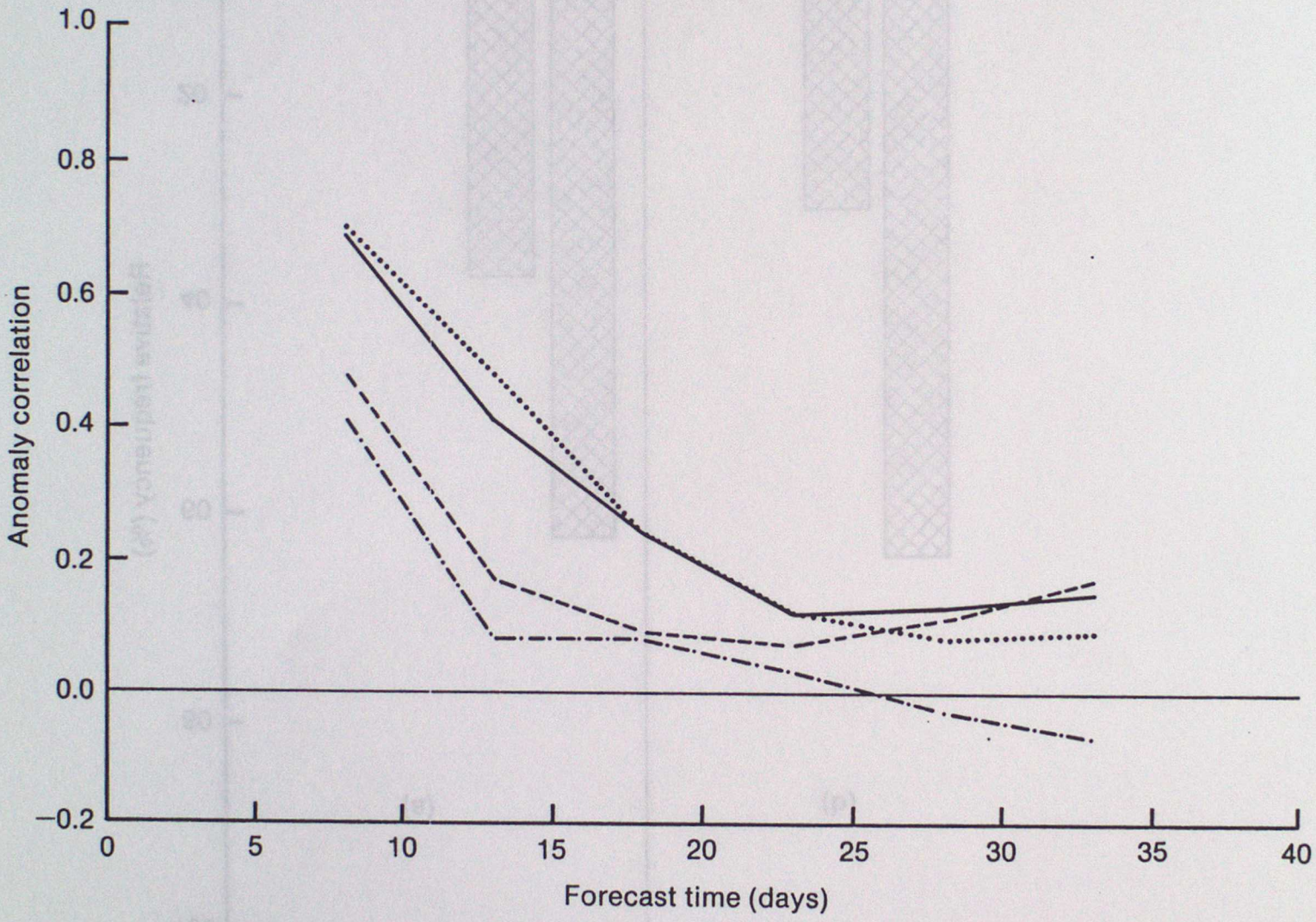

1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 7

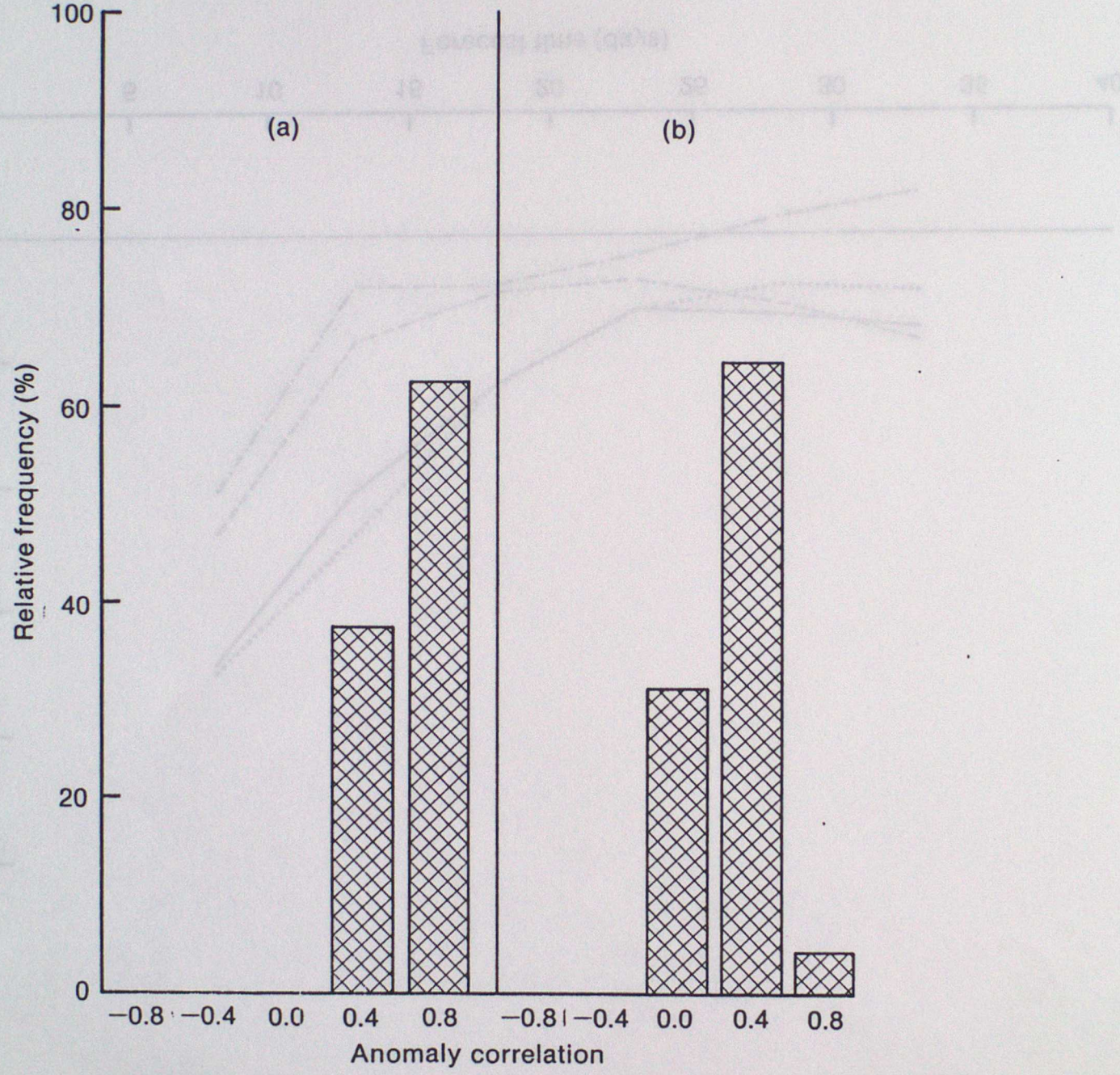

100
80
60
40
20
0
Relative frequency (%)
(a)
(b)
−0.8
−0.4
0.0
0.4
0.8
−0.8
−0.4
0.0
0.4
0.8
Anomaly correlation

FIG 8

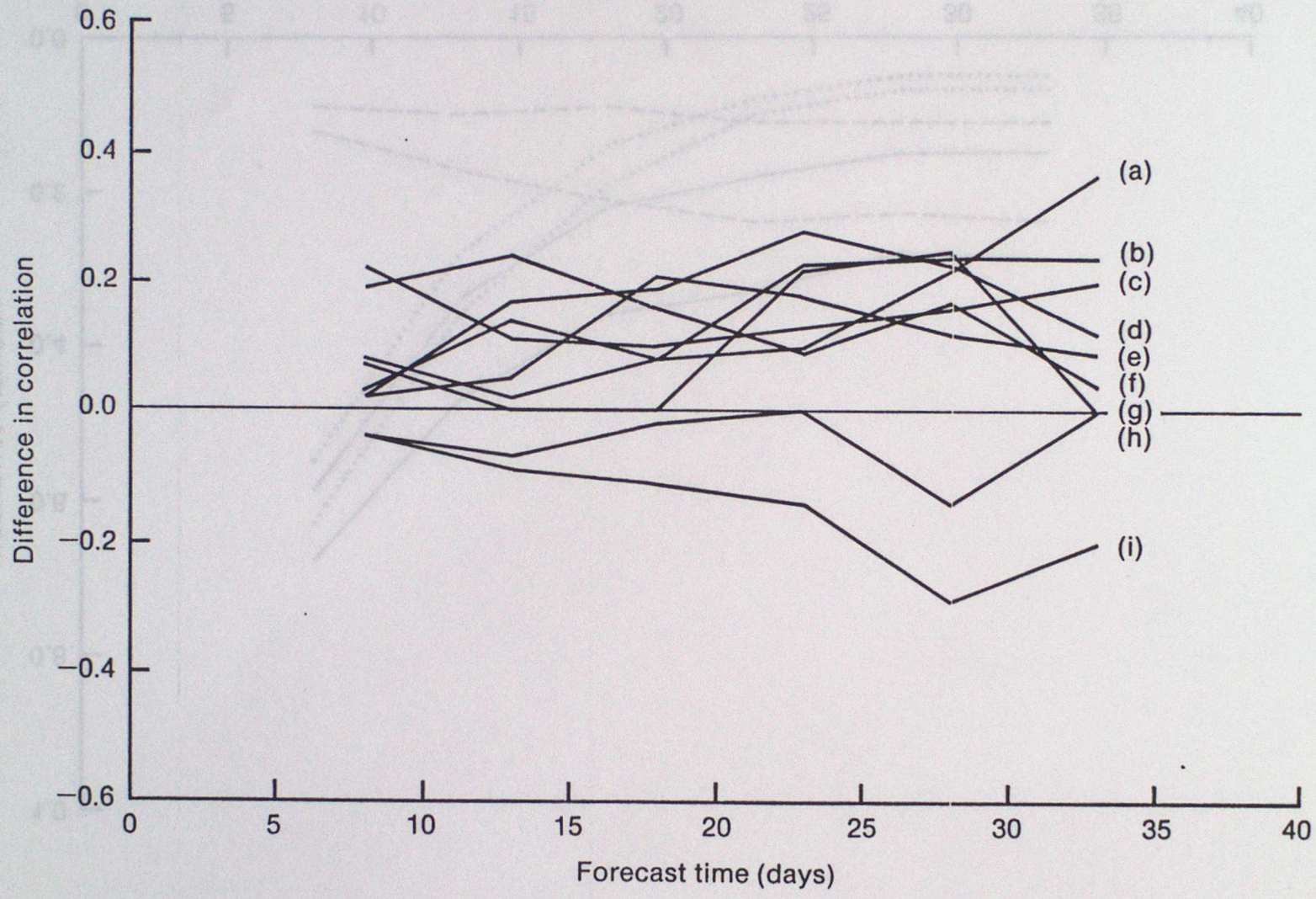

0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Difference in correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

FIG 9a

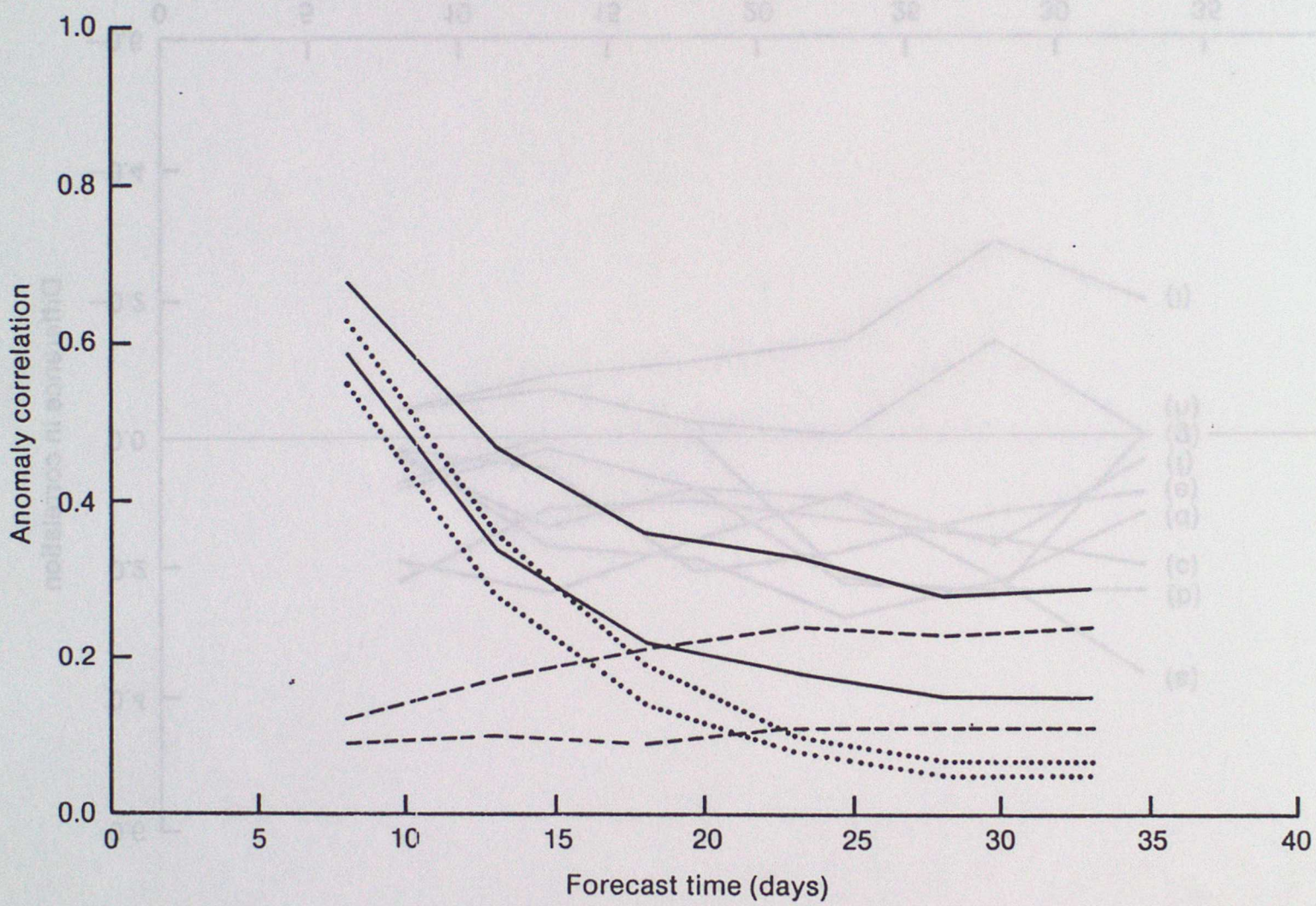

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 96

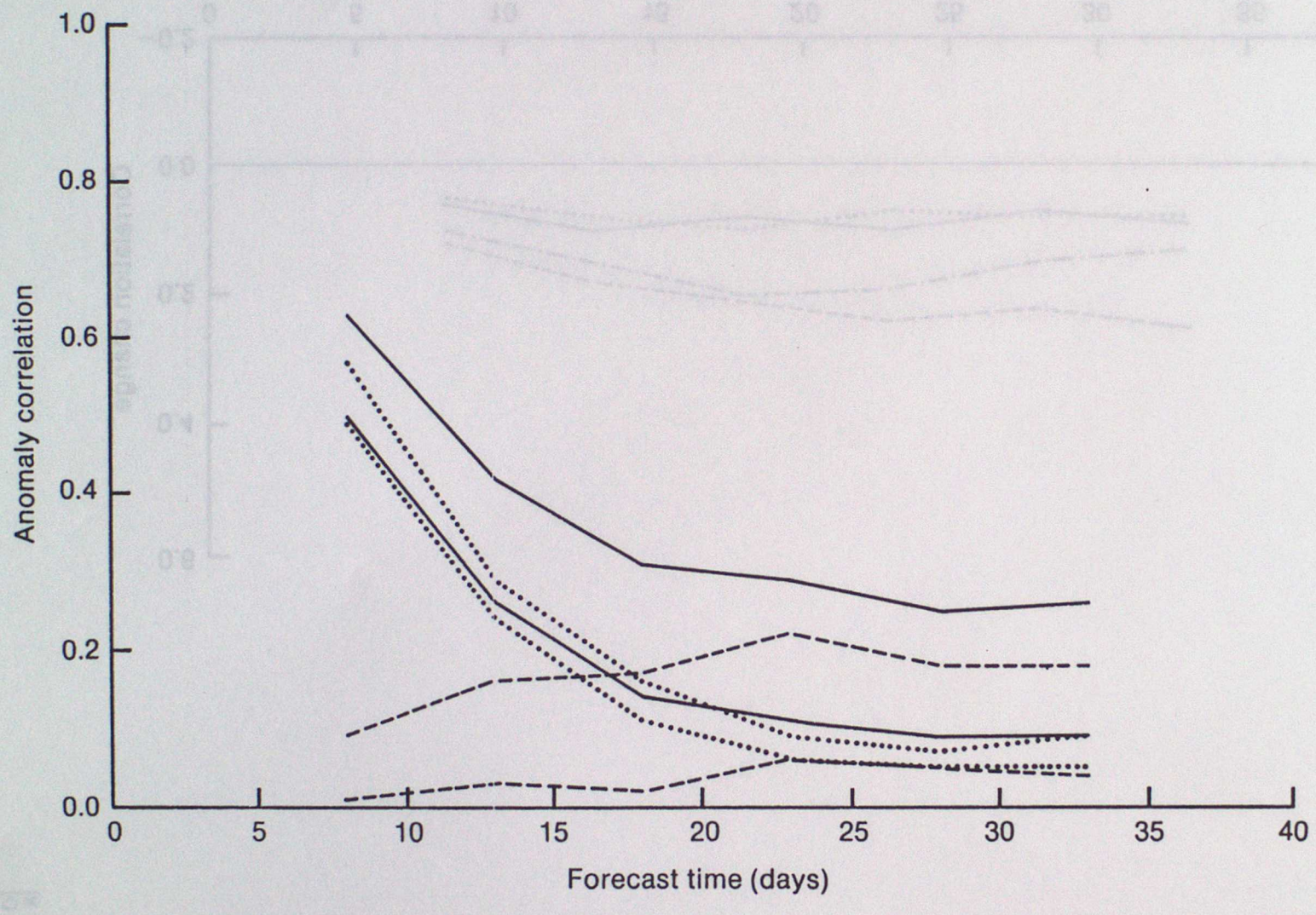

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 10a

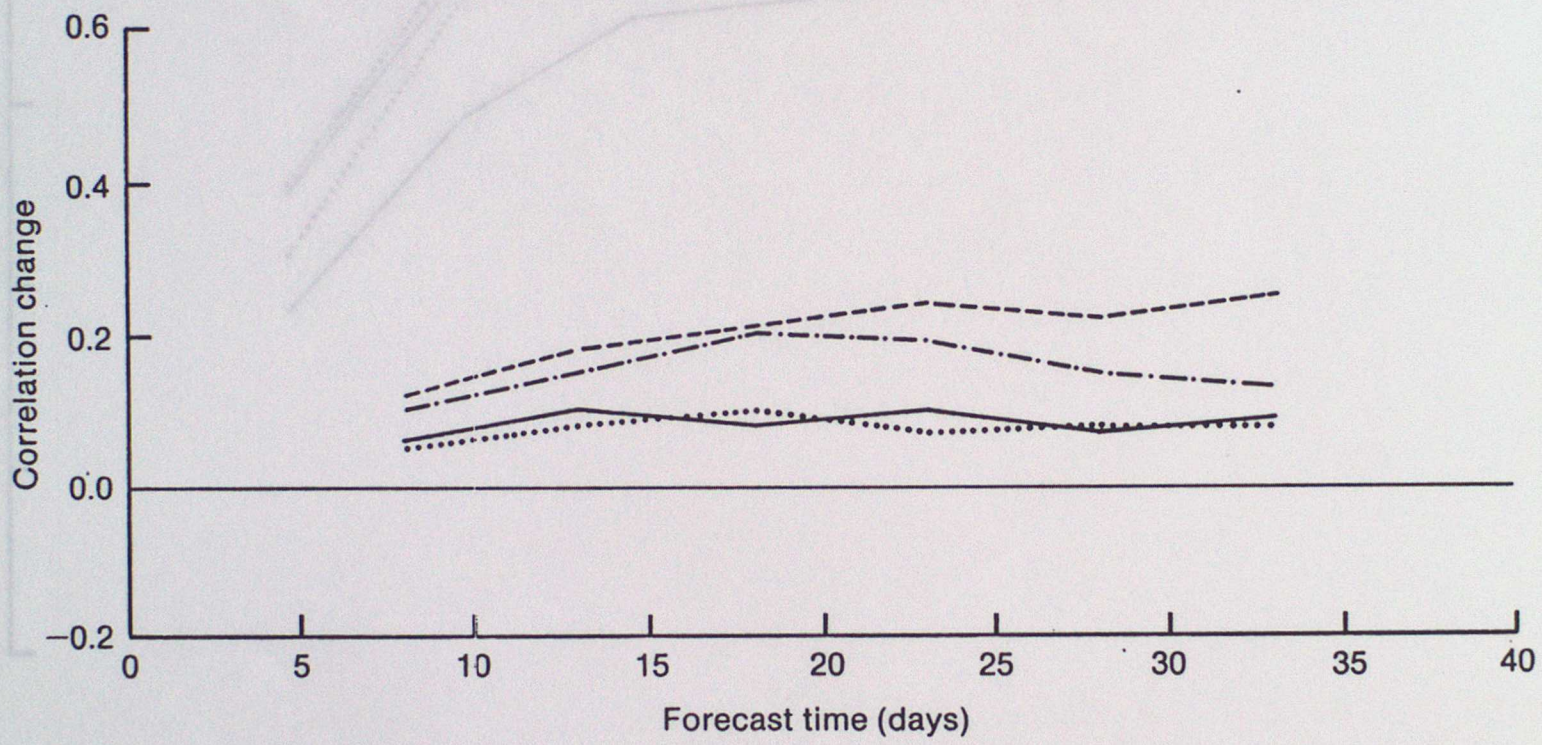

0.6
0.4
0.2
0.0
−0.2
Correlation change
0
5
10
15
20
25
30
35
40
Forecast time (days)

Fig 106

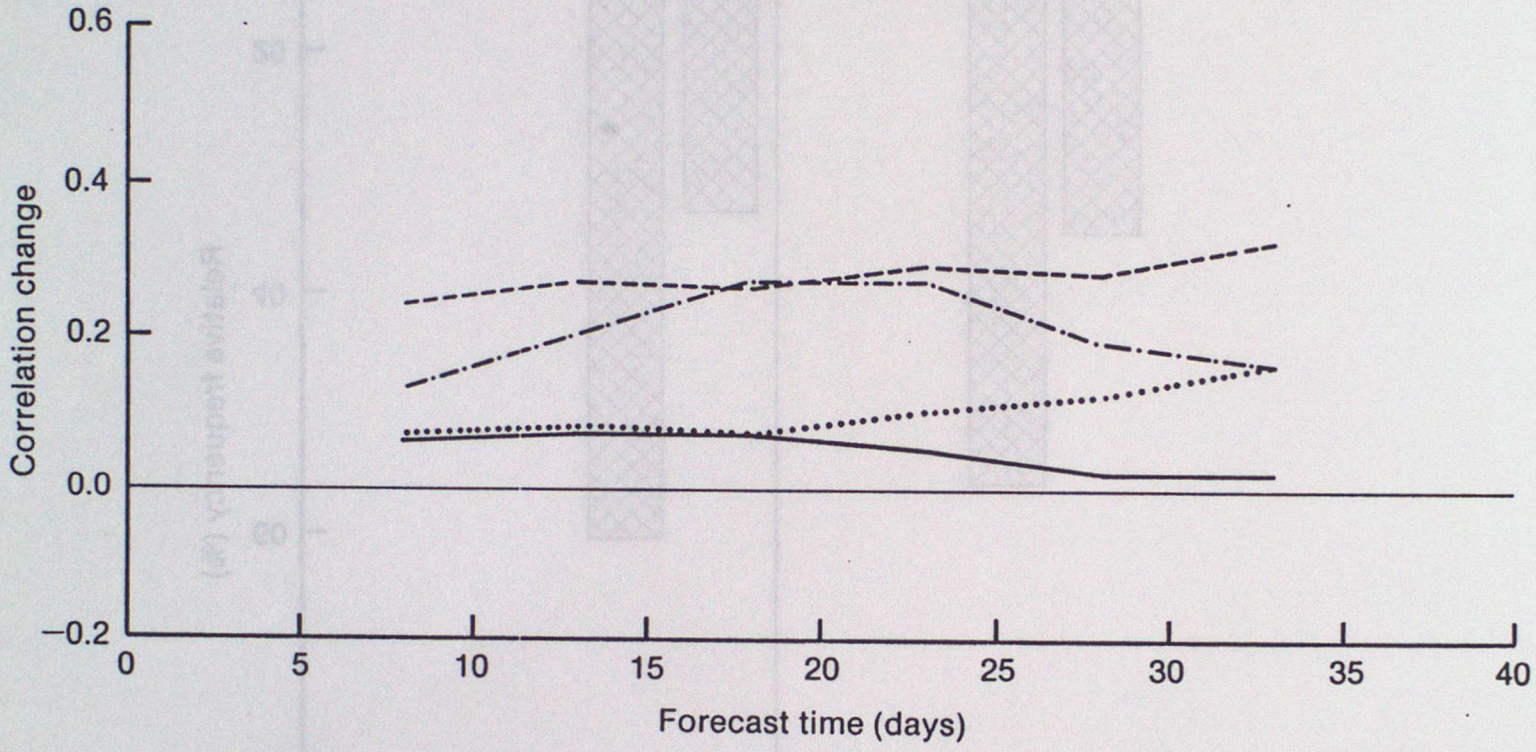

0.6
0.4
0.2
0.0
−0.2
Correlation change
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 11

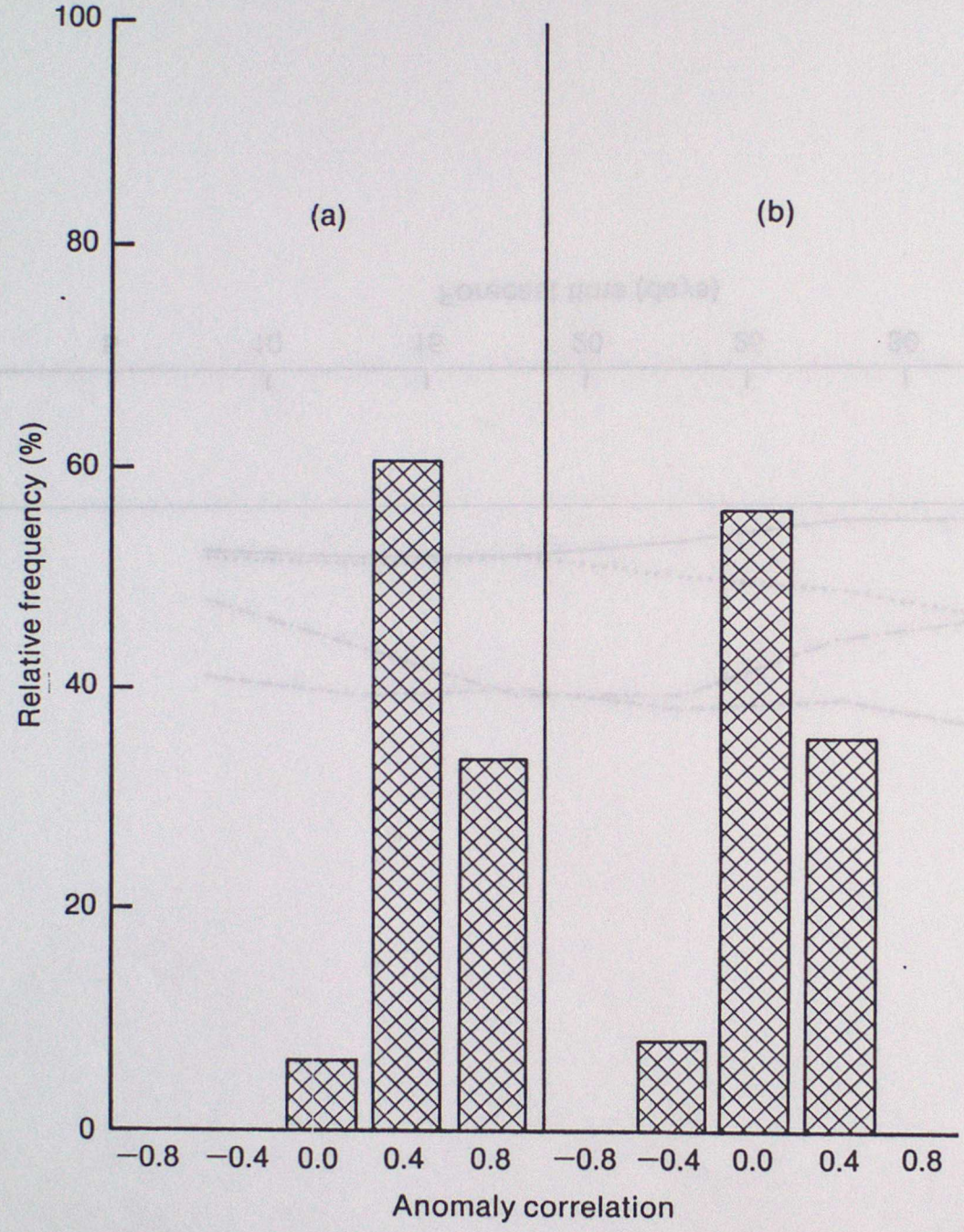

100
80
60
40
20
0
(a)
(b)
Relative frequency (%)
−0.8
−0.4
0.0
0.4
0.8
−0.8
−0.4
0.0
0.4
0.8
Anomaly correlation

FIG 12a

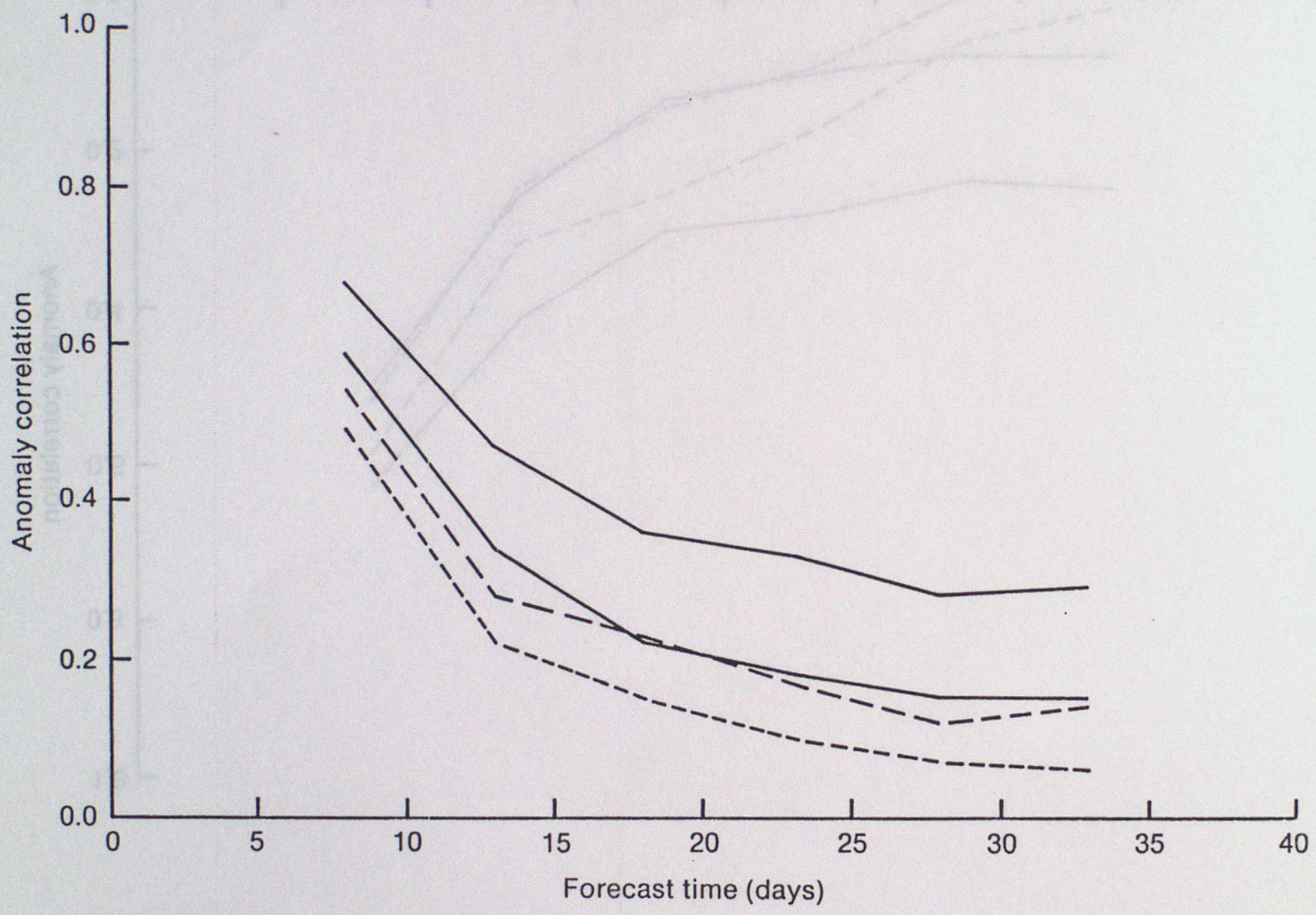

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 12 6

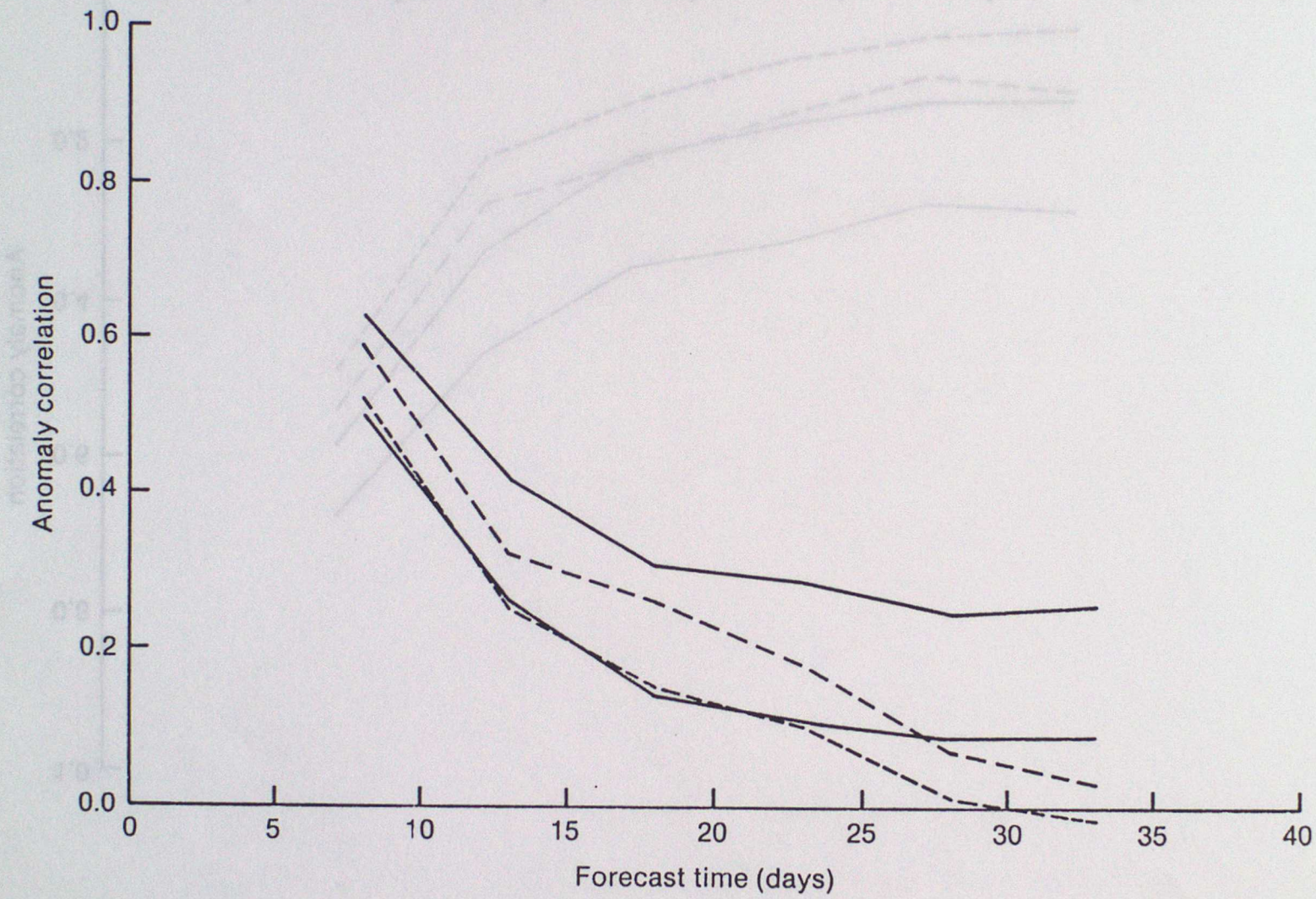

1.0
0.8
0.6
0.4
0.2
0.0
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 13

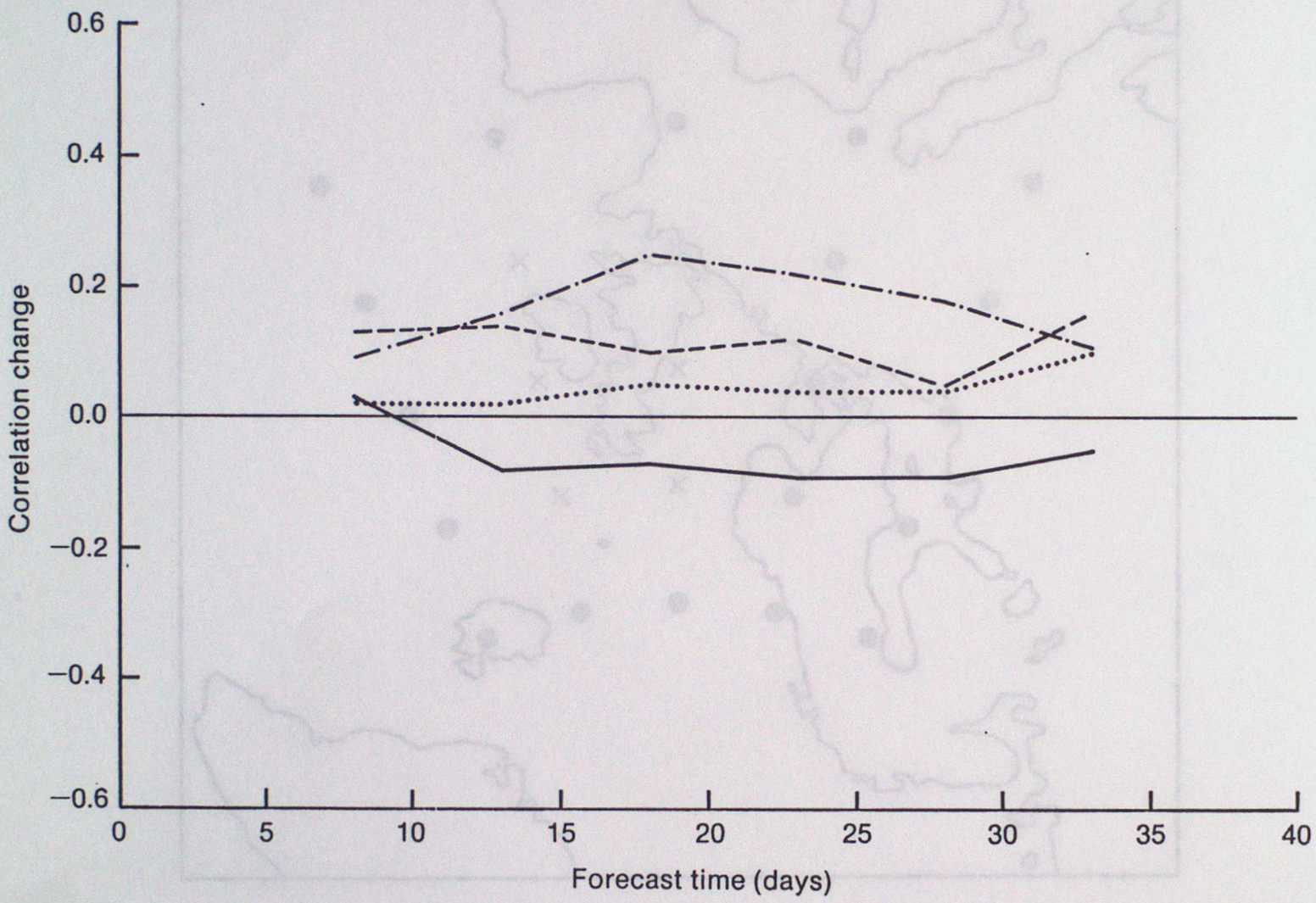

0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Correlation change
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 14

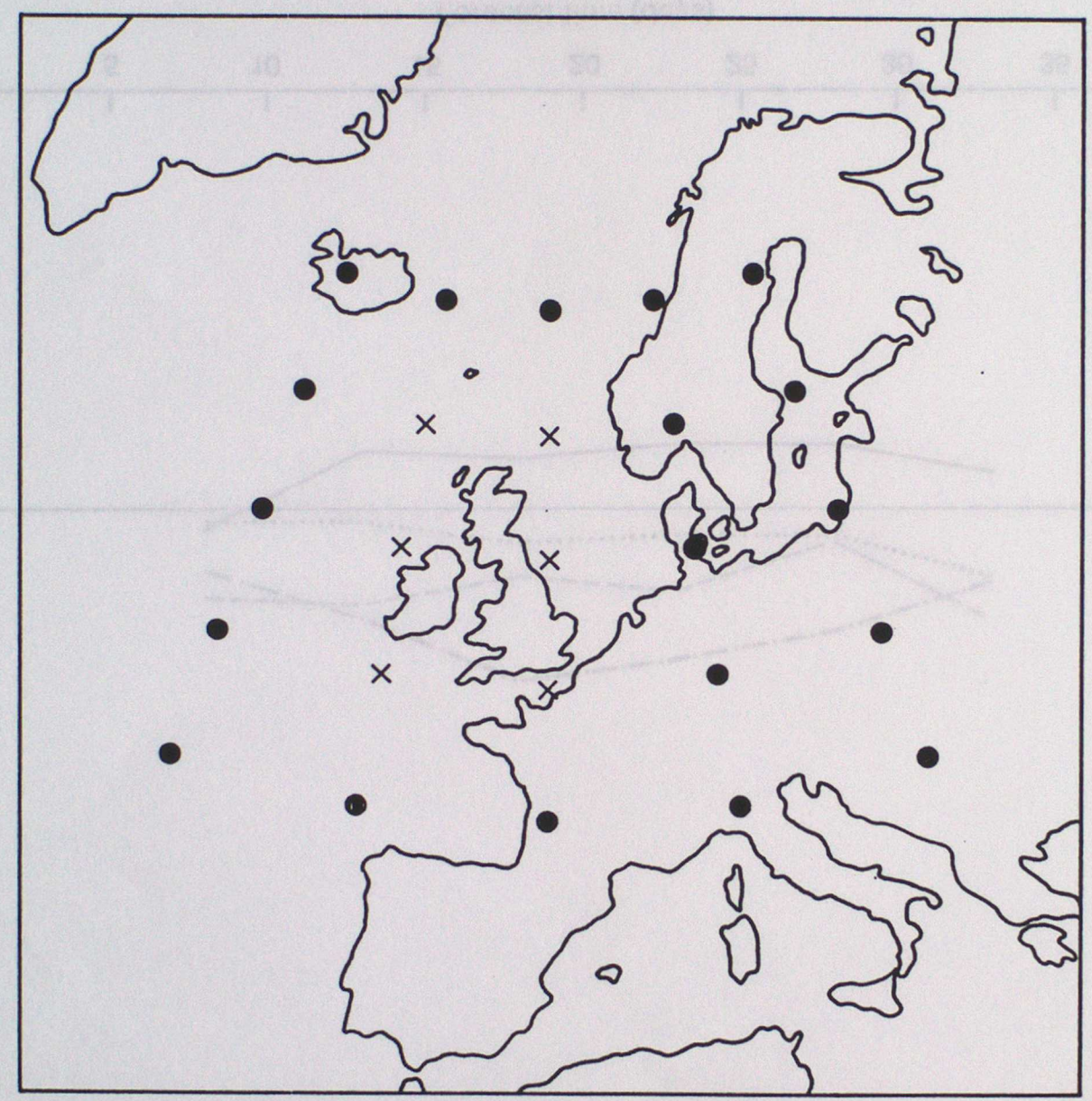


FIG 15

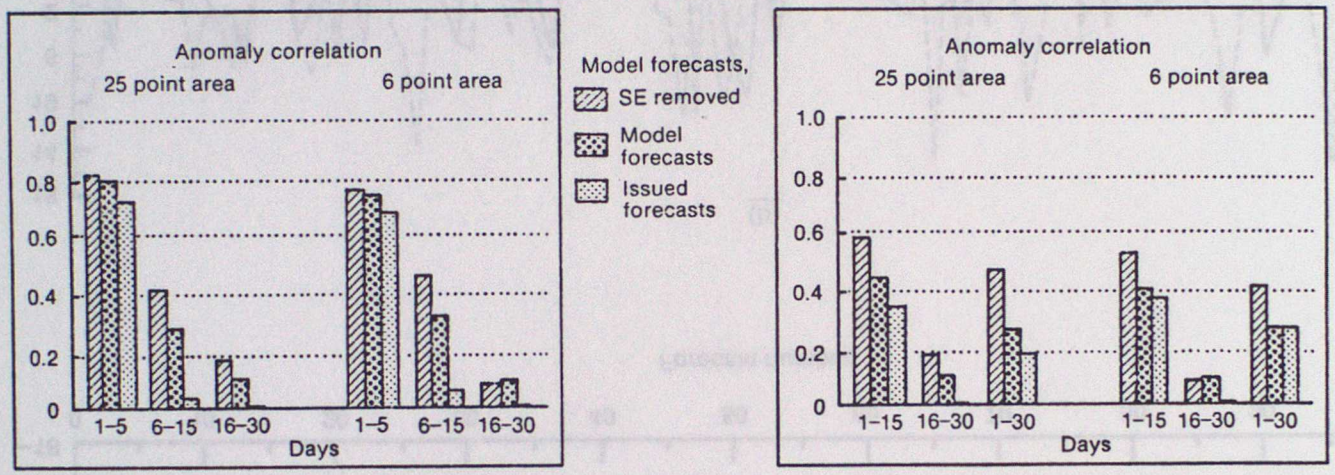

Anomaly correlation
25 point area
6 point area
1.0
0.8
0.6
0.4
0.2
0
1–5
6–15
16–30
1–5
6–15
16–30
Days
Model forecasts,
SE removed
Model forecasts
Issued forecasts
Anomaly correlation
25 point area
6 point area
1.0
0.8
0.6
0.4
0.2
0
1–15
16–30
1–30
1–15
16–30
1–30
Days

FIG 16

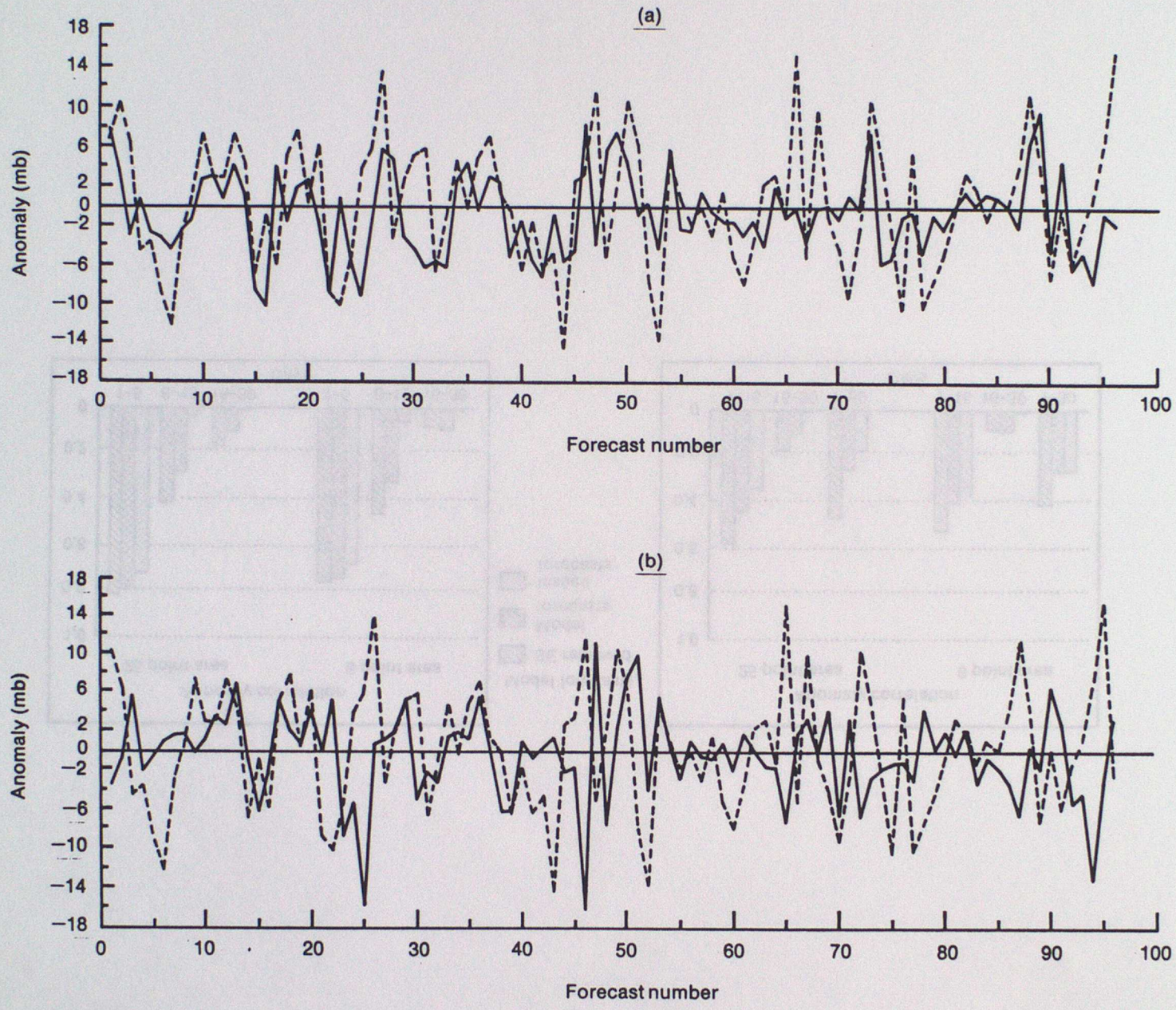
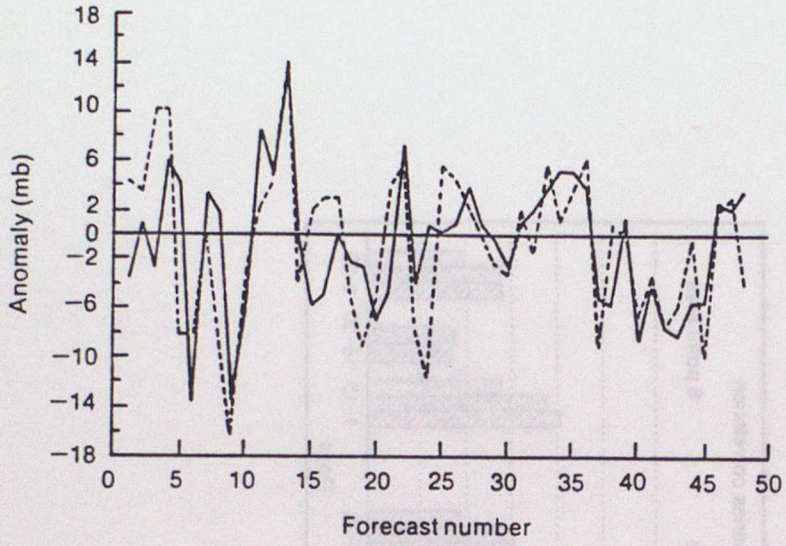

(a)
Anomaly (mb)
18
14
10
6
2
0
−2
−6
−10
−14
−18
0
10
20
30
40
50
60
70
80
90
100
Forecast number
(b)
Anomaly (mb)
18
14
10
6
2
0
−2
−6
−10
−14
−18
0
10
20
30
40
50
60
70
80
90
100
Forecast number

Fig 17

(a)


Anomaly (mb)
18
14
10
6
2
0
−2
−6
−10
−14
−18
0
5
10
15
20
25
30
35
40
45
50
Forecast number

(b)

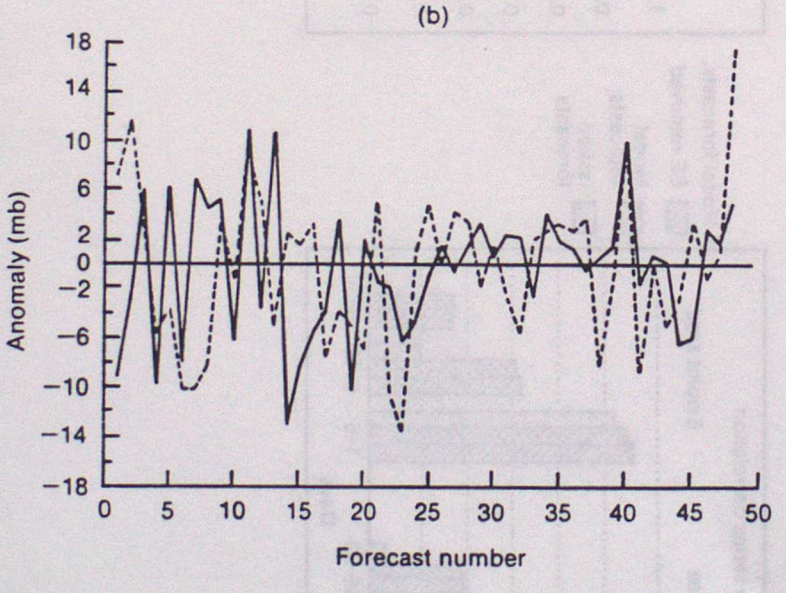

Anomaly (mb)
18
14
10
6
2
0
−2
−6
−10
−14
−18
0
5
10
15
20
25
30
35
40
45
50
Forecast number

FIG 18

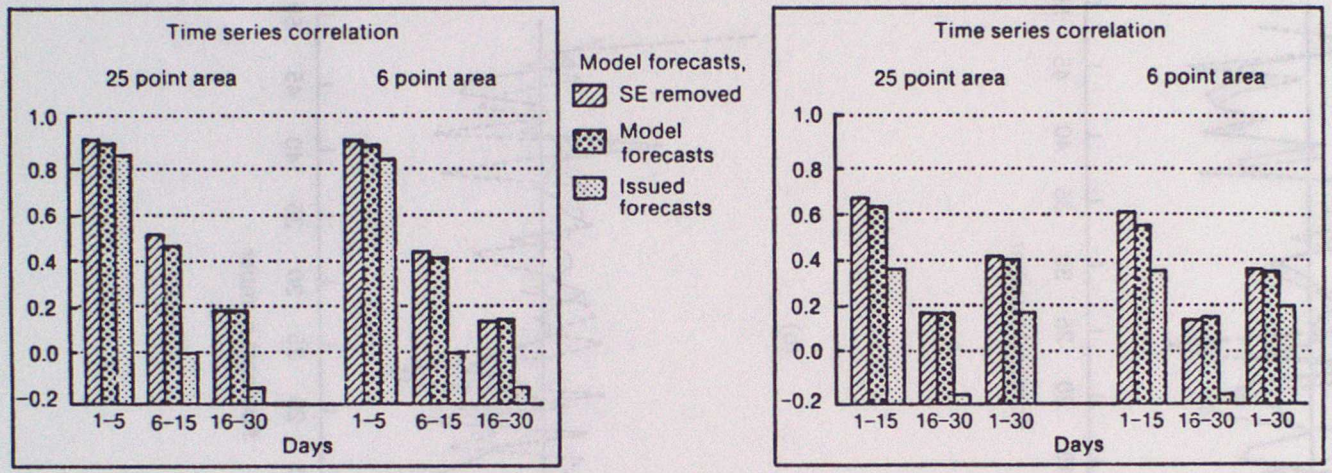

Time series correlation
25 point area
6 point area
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
1–5
6–15
16–30
1–5
6–15
16–30
Days
Model forecasts,
SE removed
Model forecasts
Issued forecasts
Time series correlation
25 point area
6 point area
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
1–15
16–30
1–30
1–15
16–30
1–30
Days

FIG 19

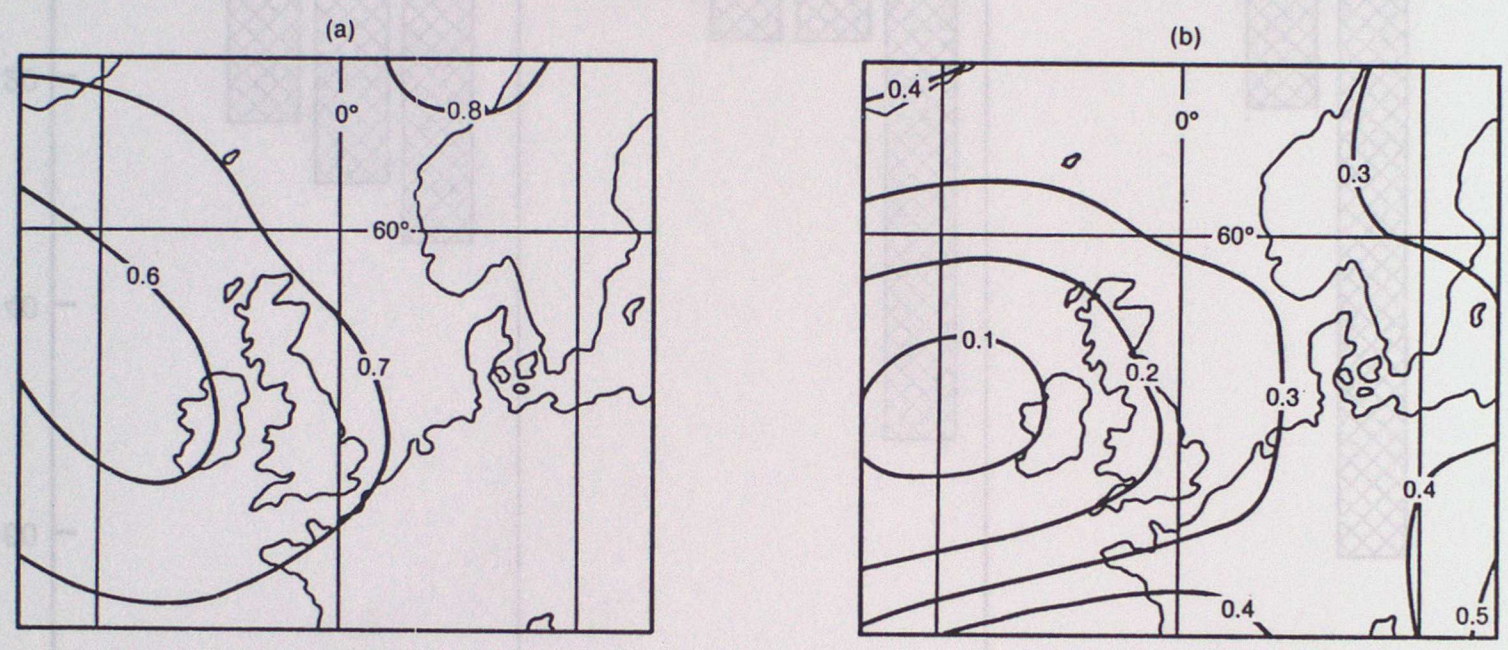


FIG 20a

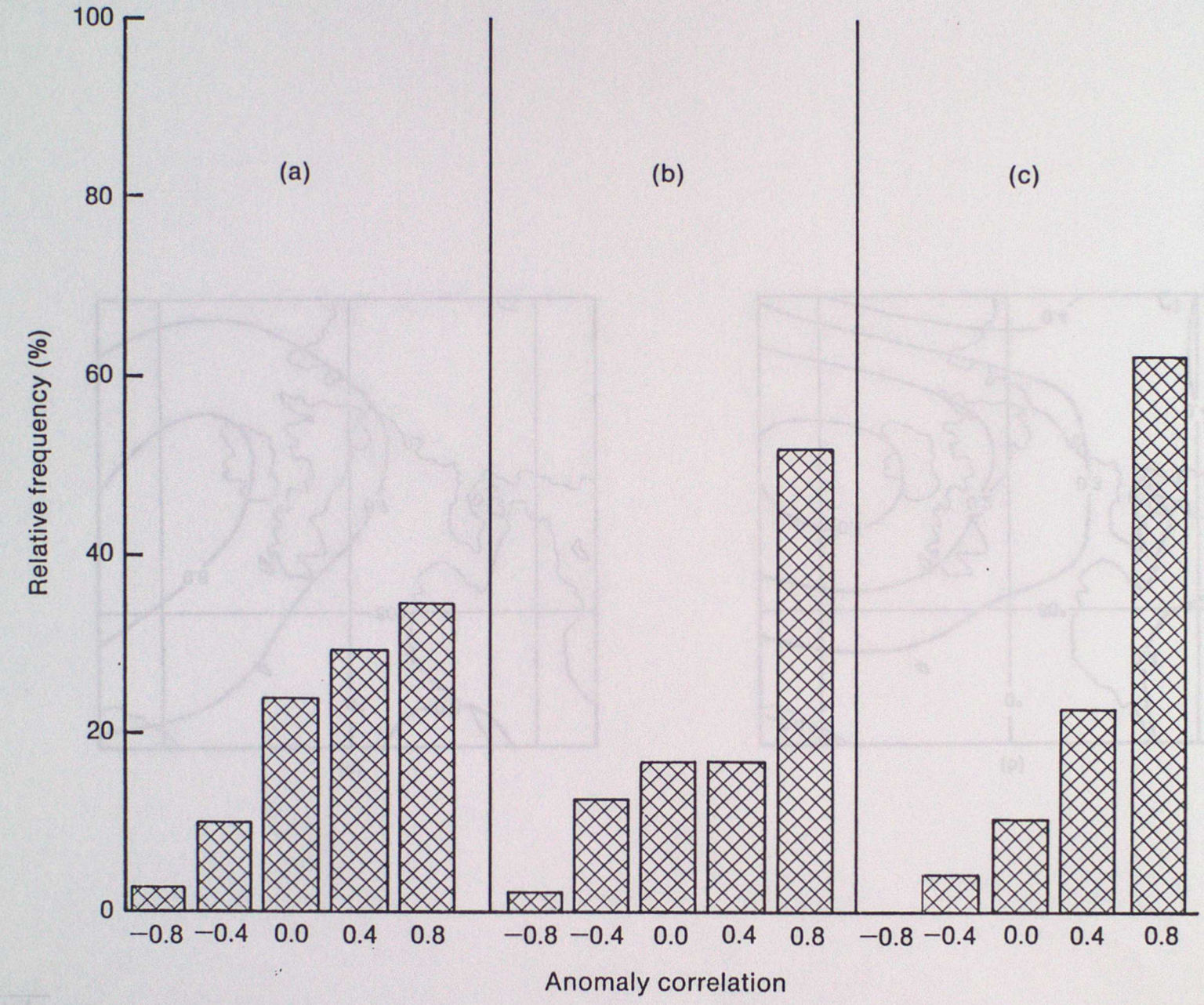

100
80
60
40
20
0
Relative frequency (%)
(a)
(b)
(c)
−0.8 −0.4 0.0 0.4 0.8
−0.8 −0.4 0.0 0.4 0.8
−0.8 −0.4 0.0 0.4 0.8
Anomaly correlation

FIG 206

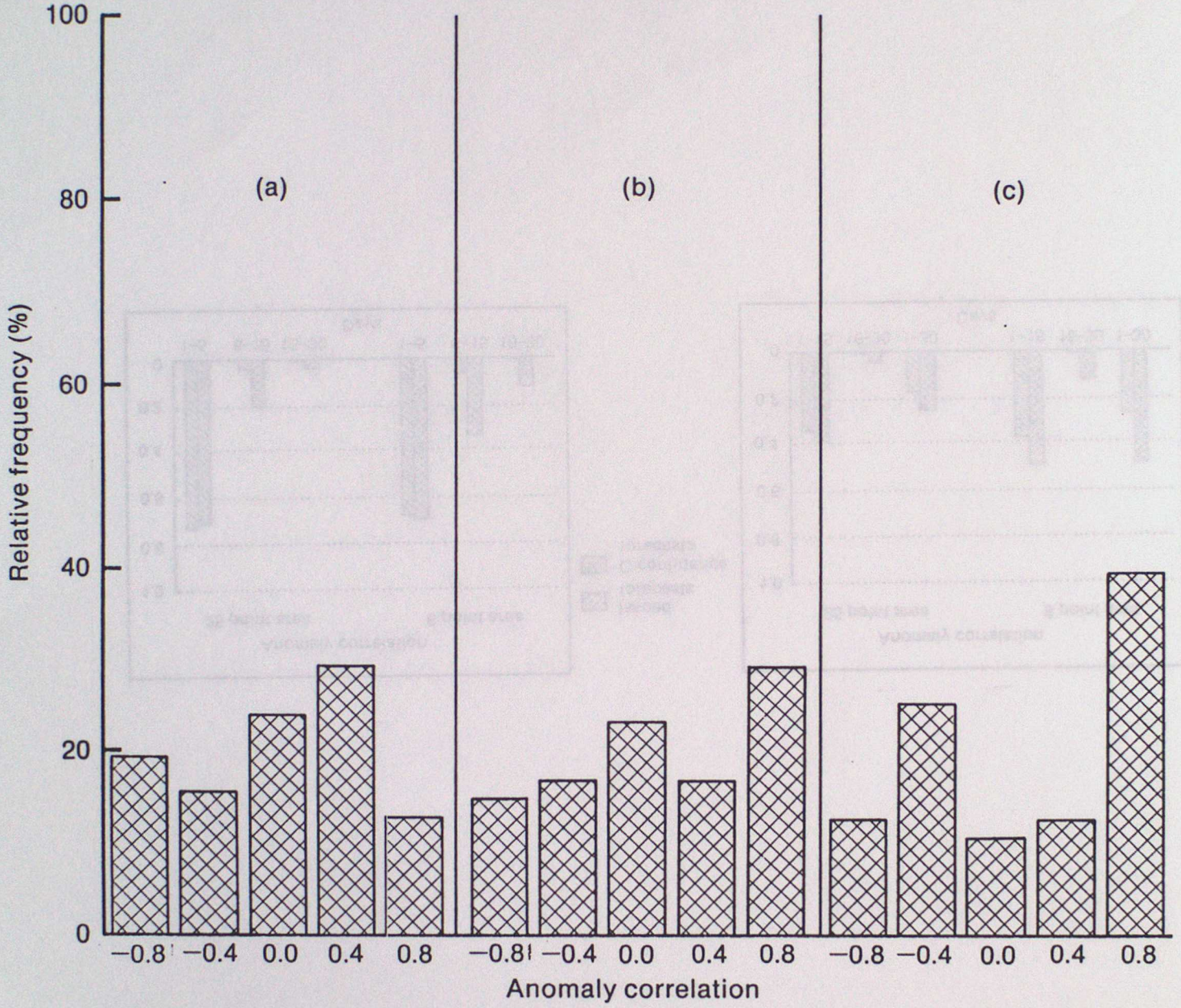

100
80
60
40
20
0
Relative frequency (%)
(a)
(b)
(c)
−0.8
−0.4
0.0
0.4
0.8
Anomaly correlation

FIG 21

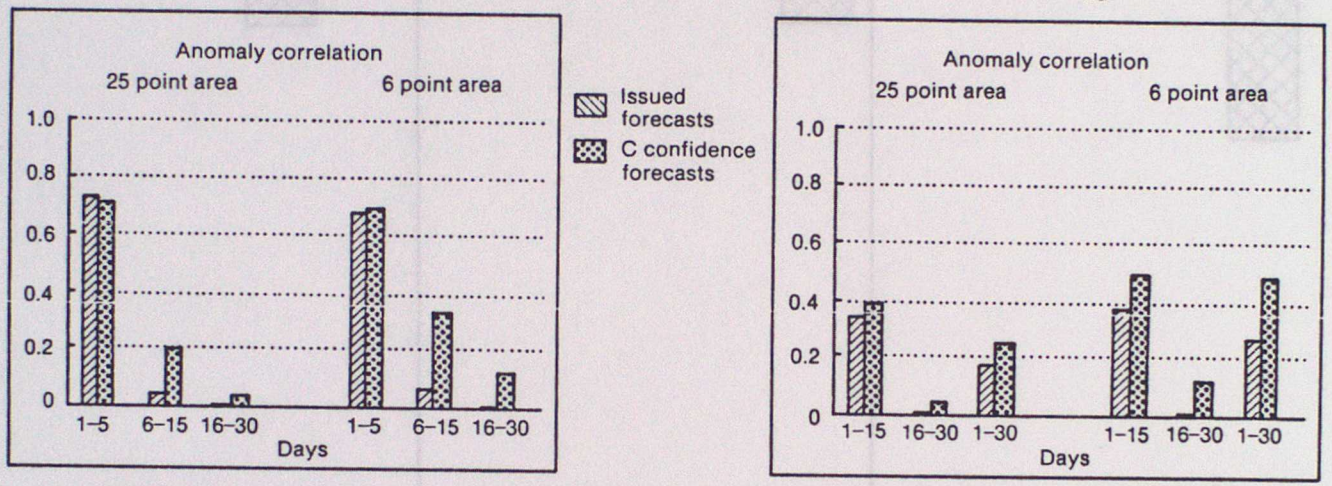

Anomaly correlation
25 point area
6 point area
1.0
0.8
0.6
0.4
0.2
0
1–5
6–15
16–30
1–5
6–15
16–30
Days
Issued forecasts
C confidence forecasts
Anomaly correlation
25 point area
6 point area
1.0
0.8
0.6
0.4
0.2
0
1–15
16–30
1–30
1–15
16–30
1–30
Days

FIG 22

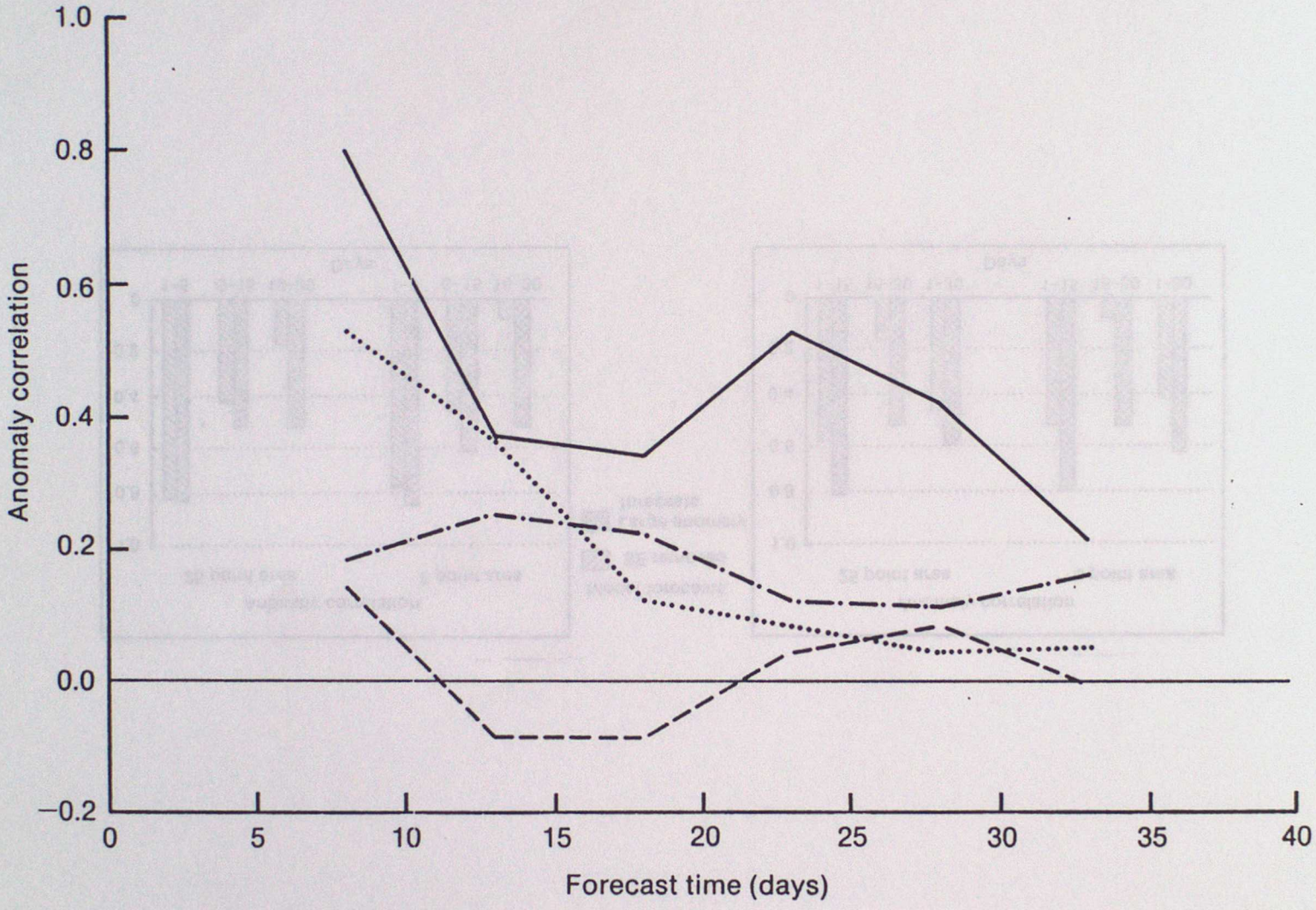

1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

FIG 23

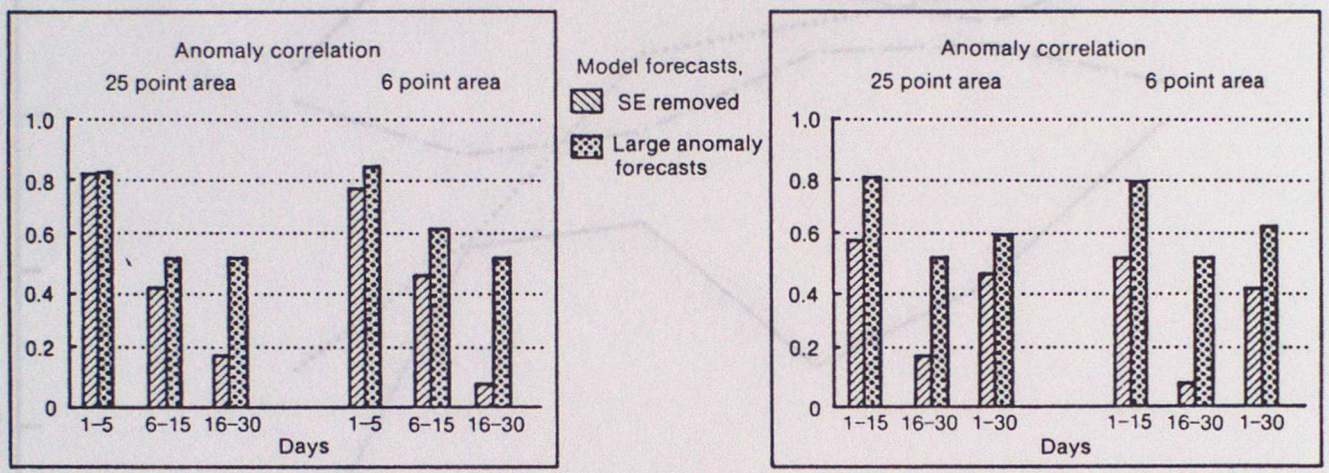
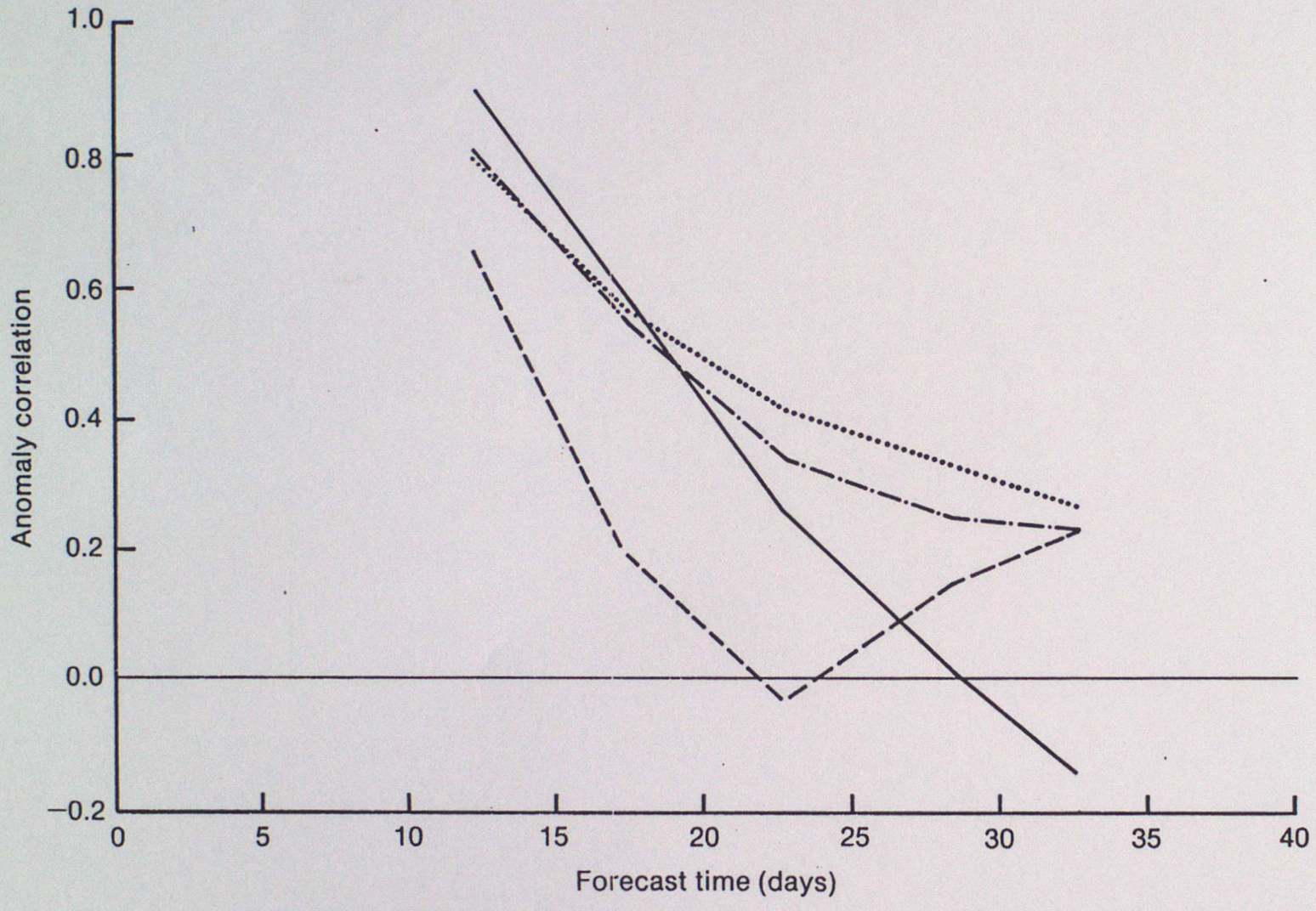

Anomaly correlation
25 point area
6 point area
1.0
0.8
0.6
0.4
0.2
0
1–5
6–15
16–30
1–5
6–15
16–30
Days
Model forecasts,
SE removed
Large anomaly forecasts
Anomaly correlation
25 point area
6 point area
1.0
0.8
0.6
0.4
0.2
0
1–15
16–30
1–30
1–15
16–30
1–30
Days

FIG 29


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Anomaly correlation
0
5
10
15
20
25
30
35
40
Forecast time (days)

INDEX TO LONG-RANGE FORECASTING AND CLIMATE RESEARCH SERIES

1) THE CLIMATE OF THE WORLD - Introduction and description of world climate.
by C K Folland (March 1986)

2) THE CLIMATE OF THE WORLD - Forcing and feedback processess.
by C K Folland (March 1986)

3) THE CLIMATE OF THE WORLD - El Nino/Southern Oscillation and the Quasi-biennial Oscillation.
by C K Folland (March 1986)

4) THE CLIMATE OF THE WORLD - Climate change: the ancient earth to the 'Little Ice Age'.
by C K Folland

5) THE CLIMATE OF THE WORLD - Climate change: the instumental period.
by C K Folland (March 1986)

6) THE CLIMATE OF THE WORLD - Carbon dioxide and climate (with appendix on simple climate models).
by C K Folland (March 1986)

7) Sahel rainfall, Northern Hempishere circulation anomalies and worldwide sea temperature changes, (To be published in the Proceedings of the "Pontifical Academy of Sciences Study Week", Vatican, 23-27 September 1986).
by C K Folland, D E Parker, M N Ward and A W Colman (September 1986)

8) Lagged-average forecast experiments with a 5-level general circulation model.
by J M Murphy (March 1986)

9) Statistical Aspects of Ensemble Forecasts.
by J M Murphy (July 1986)

10) The impact of El Nino on an Ensemble of Extended-Range Forecasts. (Submitted to Monthly Weather Review)
by J A Owen and T N Palmer (December 1986)

11) An experimental forecast of the 1987 rainfall in the Northern Nordeste region of Brazil.
by M N Ward, S Brooks and C K Folland (March 1987)

12) The sensitivity of Estimates of Trends of Global and Hemispheric Marine Temperature to Limitations in Geographical Coverage.
by D E Parker (April 1987)

13) General circulation model simulations using cloud distributions from the GAPOD satellite data archive and other sources.
by R Swinbank (May 1987)

14) Similation of the Madden and Julian Oscillation in GCM Experiments.
by R Swinbank (May 1987)

15) Numerical simulation of seasonal Sahel rainfall in four past years using observed sea surface temperatures.
by J A Owen, C K Folland and M Bottomley
(April 1988)

16) Not used

17) A note on the use of Voluntary Observing Fleet Data to estimate air-sea fluxes.
by D E Parker (April 1988)

18) Extended-range prediction experiments using an 11-level GCM
by J M Murphy and A Dickinson (April 1988)

19) Numerical models of the Raingauge Exposure problems - field experiments and an improved collector design.
by C K Folland (May 1988)

20) An interim analysis of the leading covariance eigenvectors of worldwide sea surface temperature anomalies for 1901-80.
by C K Folland and A Colman (April 1988)