PROBABILISTIC MONTHLY MEAN PREDICTIONS

USING FORECAST ENSEMBLES

M. Déqué

Centre National de Recherche Météorologique

Toulouse ,France

ABSTRACT

Two series of monthly mean forecasts for January 1981 to January 1985 have been carried out with a global version of the numerical model used at the French weather service .

In the first one, each forecast consists of 9 integrations with a T21 version. The forecast is expressed in a 3-category probabilistic way . The results show that lagged average forecasting is systematically superior to Monte Carlo forecasting and that the gaussian hypothesis for the forecast distribution is reasonable .

In the second one, each forecast consists of 5 lagged average integrations with a T42 version. The 3-category forecasts are generalized to density forecasts, under the gaussian assumption. The results show that the intra-forecast dispersion is not the best criterion to determine the standard deviation of the probabilistic prediction. With a climatological standard deviation and a model-predicted average, the probabilistic forecast of 850 hPa temperature is shown to be better than the simple climatological probabilistic forecast .

1. INTRODUCTION

The recent attempts to produce Dynamical Extended Range Forecasts by

several meteorologists (Miyakoda et al.,1986 ;Cubasch and Wiin Nielsen, 1986 ;Mansfield,1986; to quote but a few) have shown that a slight predictability remains in the large scale part of the forecast up to a monthly range, with a wide skill variability among the cases. The low , but significant, skill of such forecasts requires to complete a forecast by an indication on its skill This can be achieved by providing a forecast of a skill index (e.g. anomaly correlation). Such an approach has been developed for short and medium range forecasts by using the spread of a forecast ensemble (Kalnay and Dalcher, 1987) or by achieving a linear regression of the skill index by several predictors (Palmer and Tibaldi, 1986).

Another way to express our confidence in a prediction, instead of predicting a value and an expected error for this value, is to predict a probability density for this value. As it is presently impossible to develop a numerical model which directly produces probability densities ,as Epstein (1969) did with a simplified model, the numerical weather forecast models are used to provide several forecasts for the same variable, and then an empirical distribution is calculated (Hoffman and Kalnay, 1983). Such an ensemble forecast can be obtained by perturbations in the initial conditions which simulate the analysis errors (Monte Carlo Forecast) or by lagged initial conditions (Lagged Average Forecast). Probabilistic forecasts can also be obtained from a single numerical forecast by parameterizing the density (e.g. a gaussian distribution with the forecast value as mean and a given standard deviation) .

A probabilistic forecast is more difficult to display and to verify than a deterministic one. In the first part of the paper we shall describe simple techniques for mapping and checking a forecast, using 9-

case samples of forecasts with a low resolution model. In the second part we shall try to improve the forecasts by using 5-case samples of forecasts with a higher resolution model, and by optimizing the density distribution .

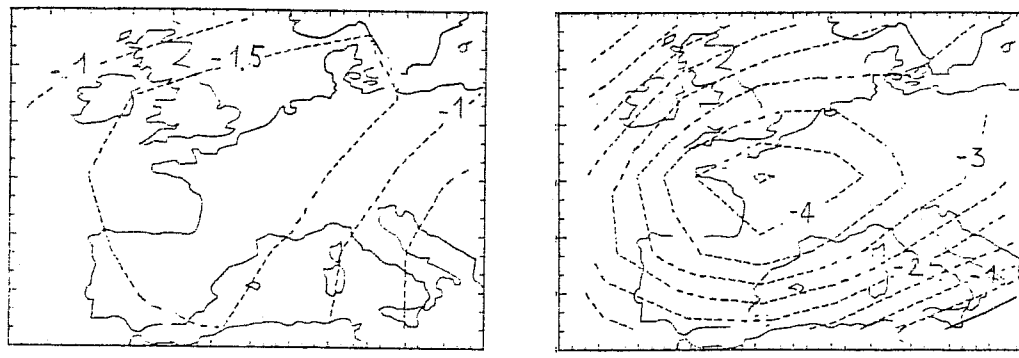
2. FORMULATION OF PROBABILISTIC FORECAST

2.1 Description of the first experiment

We have carried out 45 50-day integrations with a global T21-15L version of the operational model of the French weather service. The 45 integrations concern 5 January months (1981 to 1985) and each forecast consists of 9 integrations obtained as follows :3 initial situations are considered (14 ,15 and 16 December at 00Z) and each initial situation provides 3 different starting situations by small perturbations on the variables. Then 45 averages are computed for the 1 January-31 January period. A forecast for a given year consists thus of 9 predicted values and a reference value obtained by averaging the 36 (=45-9) remaining values. A deterministic prediction can be achieved by providing the difference between the 9-sample mean value and the reference value. This procedure allows to remove the model systematic error which is important at such lags and for such a resolution. The predicted anomaly is then compared with the observed anomaly which has been calculated by the same way from ECMWF daily analyses (e.g. the January 1981 observed anomaly is the difference between the January 1981 mean and the January 1982 to 1985 mean) .

This experiment and its results are described in Déqué (1988). The deterministic prediction is shown to be better than the climatological one (i.e. zero anomaly) and the persistence one (mean 15 Nov.-15 Dec. anomaly) for 850 hPa temperature over Europe. We shall focus our

discussion on this last field in the following since its prediction on a monthly time scale could be helpful to human activities, particularly in winter (e.g. for fuel storage). Here Europe corresponds to the area 35°N-60°N and 15°W-20°E. The Root Mean Square (RMS) error is 2.1°C for the model, 2.5°C for the climatology, and 3.6°C for the persistence (the model systematic error is -2.6°C). For each January except 1984 , the model is the best predictor .



Predicted Observed

Fig.1 :predicted and observed 850 hPa temperature anomaly (°C) for January 1985 .

Fig.1 shows the forecast for the best predicted January, that is January 1985. The sign of the anomaly is well captured, but not its amplitude .

2.2 Probabilistic forecast

The above mentioned model skill is interesting since we are not very exacting for long range forecasts, but it expresses a large uncertainty on the results. A mathematical way to translate our uncertainty is the probability theory. Two approaches may be considered. The first one , the parametrical method, consists of choosing a kind of density (for example a gaussian distribution) and of adjusting a small number of parameters that define accurately the distribution. In this case the prediction is carried out by providing the set of parameters. In the

second one, the non-parametrical method, the density is discretized in a small number of categories, and the prediction consists of giving the set of probabilities (one for each category). One can also consider a combined approach which adjusts a given kind of density and uses this density to calculate probabilities for a few categories. In this section we shall make the assumption that the 9 individual forecasts have equal probabilities and that our uncertainty is only due to the dispersion of these values. We shall choose a non-parametrical approach since with 9 values one can expect to represent phenomena different from the classical gaussian shape (e.g. bimodality). We have thus to choose the number and the boundaries of the categories. We have decided to take categories with equal probabilities when the distribution is the climatological one. With two categories, the threshold is located near the average value (which is 0 since we consider anomalies) and the quality of the predictions is rather bad because you find often a small predicted value and a small observed one with the opposite sign. We shall consider here 3 categories. The thresholds (for each grid point) are given by the terces of the 5-January observed anomaly distribution. For 850 hPa temperature over Europe, they correspond nearly to -1°C and 1°C ,and the 3 categories will be referred to as cold, medium, and warm. The forecast at each grid point consists thus of 3 positive values p_1 , p_2 ,p_3 the sum of which is 1. They are computed by dividing by 9 the number of individual forecasts which belong to each category .

2.3 Verification

The problem at this stage is to know what a good forecast is.The distribution must be sharp and the likeliest category must correspond to the observed one. .Set o_i equal to 1 when the ith category is observed, equal to 0 otherwise (i=1 ,2 ,3). The best forecast is $p_i=o_i$.

It seems natural to measure the skill by the Brier Score (Brier, 1950):

$$BS=(p_1-o_1)^2+(p_2-o_2)^2+(p_3-o_3)^2 \qquad (1)$$

This criterion lies between 0 and 2, 0 corresponding to the best prediction, 2 to the worst one. Let us suppose that we observe a cold anomaly ($o_1=1$); if we had predicted a warm ($p_3=1$) or a medium ($p_2=1$) one, the BS is 2, though the error seems more important in the former case. Murphy (1971) proposed a Ranked Probability Score based on the distance between the cumulative probabilities :

$$RPS=(p_1-o_1)^2+(p_1+p_2-o_1-o_2)^2+(p_1+p_2+p_3-o_1-o_2-o_3)^2$$

$$=(p_1-o_1)^2+(p_3-o_3)^2 \qquad (2)$$

This criterion has the same properties as the BS but takes into account the fact that the categories are ranked (the highest value is obtained when one predicts an extreme category and one observes the opposite one). If (p_1,p_2,p_3) is a random vector independent of (o_1,o_2,o_3), that is if we do not have any information to produce the forecast, the lowest expected value for the RPS is obtained when p_i is constant and equal to the expectation of o_i. This kind of prediction is the climatological probabilistic prediction. Since we have chosen the categories with equal probabilities, the climatological forecast is $p_i=1/3$ and the expected RPS is 4/9. An informative prediction must provide an averaged RPS below 0.44. In the case of the forecast of 850 hPa temperature over Europe, the averaged (over the 5 Januaries and over the area) RPS is 0.29. If we consider the different Januaries, the model is always better than the climatology except in 1981. The model provides valuables probabilistic predictions. The persistence predictions have not been considered since they have a lower skill than the climatological ones in the deterministic approach, and since it is difficult and artificial to introduce a probability for this kind of

prediction .

2.4 Mapping

We dispose thus of 3 values (p_1,p_2,p_3) to display instead of 1 when the forecast is deterministic. One can exhibit 3 maps of iso-probability , or only two since the sum is 1. One can also superimpose the isolines on a single map, but the results are often difficult to interpret. In fact we have two degrees of freedom and the most appropriate way of mapping is to draw an arrow field. Epstein and Murphy (1965) have proposed to use barycentric coordinates. Let $O_1O_2O_3$ be a triangle (see Fig.2) . A set of probabilities (p_1,p_2,p_3) may be represented by a point P with barycentric coordinates (p_1,p_2,p_3):

$$p_1\underline{PO_1}+p_2\underline{PO_2}+p_3\underline{PO_3} = \underline{0} \qquad (3)$$

Since the p_is are positive ,the point lies inside the triangle. When a category is deterministically predicted, $p_i=1$, the corresponding point is the vertex O_i. The climatological forecast $p_i=1/3$ is represented by the barycenter (or center of mass) G. The vector $\underline{GP}$ indicates the degree of confidence by its length (a short arrow corresponding to a weak confidence) and the preferred category by its direction (towards the left, cold; upwards, medium; towards the right, warm). We can choose any kind of triangle. Epstein and Murphy proposed to take an equilateral triangle. In this case the distance between two points is proportional to the BS. We have taken in Fig.2 a right isosceles triangle, so that the distance is proportional to the RPS .

Fig.3 exhibits the forecast for 1985 as Fig.1 but in a probabilistic way. Most arrows are directed to the left, indicating a cold anomaly . Over Great Britain, the arrows are shorter, indicating a smaller probability for the cold category .

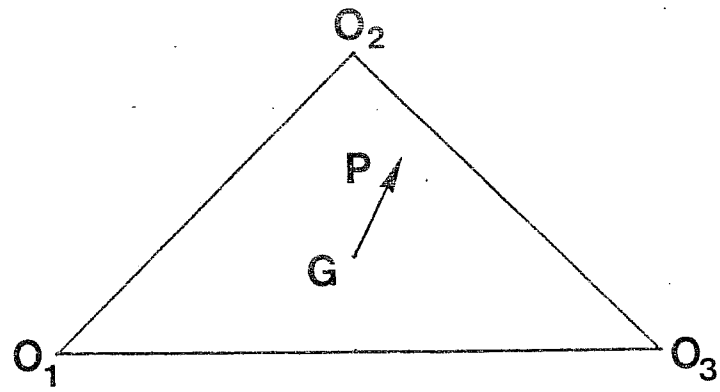


Fig.2:vector representation of a probabilistic forecast .

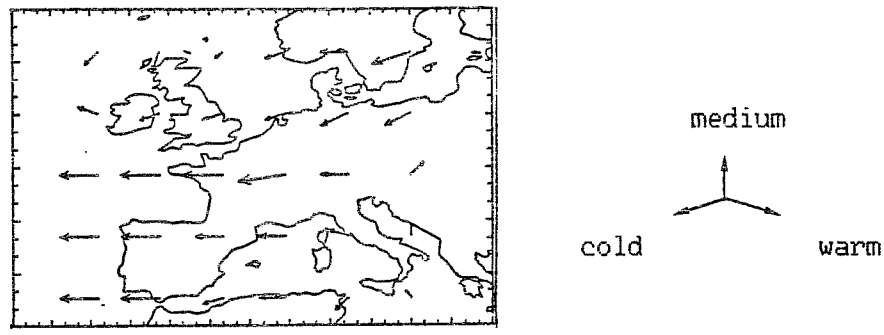


Fig.3 Probabilistic forecast for 850 hPa temperature in January 1985.

2.5 Further investigations

The Lagged Average Forecasts are shown to be significantly better than the Monte Carlo Forecasts in this experiment (Déqué,1988) and it is shown that the skill increase by ensemble averaging is larger from a single to 3 integrations than from 3 to 9 integrations. It seems that less than 9 should be sufficient. Anyway this number will allow us to test the hypothesis of a gaussian distribution for the predicted anomalies. If we calculate (p_1,p_2,p_3) using a gaussian distribution with mean and standard deviation computed from the 9 values, the mean RPS is the same (i.e. 0.29) and for the individual Januaries, the values differ slightly from the previous ones. The distribution of the 9 temperature anomalies is therefore near to a gaussian distribution .

3. OPTIMIZATION OF PROBABILISTIC FORECASTS

3.1 Description of the second experiment

In order to improve our results, we have doubled the horizontal resolution of the model (T42 truncation) and reduced the time step from 45 min to 30 min. The cost of a single 50-day integration is multiplied by a factor 5. Considering the results of the previous experiment, we have only performed 5 integrations for each forecast. They are obtained by the LAF technique, taking 14 Dec. 00Z ,14 Dec. 12Z ,15 Dec. 00Z ,15 Dec. 12Z, and 16 Dec. 00Z as initial conditions. The forecast periods are the same (1981 to 1985) and the predicted monthly anomalies are interpolated on the same grid as in section 2 by spectral truncation . When compared with observed anomalies, the forecasts show a general skill improvement, at least in the lower atmosphere. The RMS error of 850 hPa temperature over Europe is 1.8°C .

3.2 Probabilistic forecast

Our samples are now shorter (5 instead of 9) and it seems bold to calculate probabilities for 3 categories using the empirical distribution. As the gaussian hypothesis does not decrease the skill with a 9-sized sample, it should be more appropriate to calculate the mean and standard deviation of the 5-sized sample and to make use of a gaussian density .

Moreover, the choice of 3 categories is somewhat arbitrary in that case. For instance, if the 3 categories are $t<0°C$, $0°C<t<2°C$,and $2°C<t$ and if we predict a large probability for the first category, people who are concerned with blossom freezing will find the information helpful , but those who ask about fuel oil freezing cannot be satisfied. Then the choice of the categories should be left to the users and the adequate meteorological information is the whole probability density. Under the

gaussian assumption, we need to calculate the mean m and the standard deviation s. The forecast maps involve thus two isoline network. The forecast verifications need to generalize the RPS formula to an infinite set of categories which have the same probability with respect to the climatological distribution. Set $F_c(t)$, $F_f(t)$, and $F_o(t)$ the cumulative densities for the climatology, the forecast, and the observed valute t_o respectively. Then ,

$$F_o(t)=H(t-t_o) \qquad (4)$$

where H(t) is the Heavyside function H(t)=0 when t<0 and H(t)=1 when t>0. A generalized RPS may be given by :

$$GRPS=2\int_0^1 (F_f(F_c^{-1}(u))-F_o(F_c^{-1}(u)))^2 du \qquad (5)$$

The factor 2 makes the GRPS vary between 0 and 2 like the RPS. The small categories are $(F_c^{-1}(u)<t<F_c^{-1}(u+du))$ and have the probability du. In the non parametrical approach, the forecast is given by n values t_1, $t_2,\ldots,t_n$ and the climatology by N values $T_1,T_2,\ldots,T_N$ .Then

$$GRPS= 2/N \sum_{k=1}^{N} (1/n \sum_{i=1}^{n} H(T_k-t_i) -H(T_k-t_o))^2 \qquad (6)$$

If we substitute a gaussian distribution of parameters m and s for the set (t_i), (5) becomes:

$$GRPS= 2/N \sum_{k=1}^{N} (G((T_k-m)/s) -H(T_k-t_o))^2 \qquad (7)$$

where G is the cumulative density of the normalized gaussian law. If we use a gaussian distribution for the climatology, (5) cannot be expressed analytically. Moreover the gaussian distribution is not as well verified for the temperatures from different years as for the temperatures from different integrations of the same January. Therefore we shall make use of (7) to calculate the GRPS. For each year the climatological set (T_k) will consist of the 4 observed anomalies of the remaining years. The mean GRPS for the climatological forecast is 0.44 if we use (6) and 0.41 if we use (7), irrespective of the kind of field

and of the area of verification (in the first case, one can demonstrate that the mean GRPS is $(1+(N+1)/N^2)/3$). In the case of 850 hPa temperature over Europe, the mean GRPS of the T42 forecasts is 0.34 , which shows that the model forecasts are superior to the climatological ones .

3.3 Question about the standard deviation

If we come back to the first experiment ,the mean GRPS is .30 .This result is surprising since the deterministic prediction is better with the T42 model than with the T21 version. Since the mean of the distribution is improved, the standard deviation must be inadequate. We have substituted, as a first step of checking , a mean standard deviation for the intra-forecast standard deviation: that is, for example, instead of the intra-forecast standard deviation of 1981, we used the average of the intra-forecast standard deviations of 1982 , 1983 ,1984, ,and 1984 to produce a probability prediction for 1981. The mean GRPS is found to be smaller (.28 instead of .34). The internal variability of the forecast ensemble is not a good estimate of the uncertainty of the forecast. Of course this result is verified whatever the field (temperature,wind,height), whatever the area (globe, intertropical,Europe) ,and for both T21 and T42 experiments. The problem is to choose a good estimate for s.We tried the observed and the forecast inter-month standard deviations. For most fields and areas the best choice seems to be the standard deviation between all the individual forecasts (i.e. the 45 forecasts for the T21 or the 25 forecasts for the T42). With such a choice, the GRPS for 850 hPa is .42 ,.42 , and .30 for the globe, northern hemisphere ,and Europe in the T21 experiment and .38, .40, and .28 in the T42 experiment (values to be compared with .41 for the climatological forecast). The shortcoming

of this kind of forecasts is that we do not predict our confidence in the results. We only predict the average of the gaussian law, the standard deviation being the same whatever the year .

3.4 Optimizing the standard deviation :first attempt

We can try to predict the standard deviation without using the intra-ensemble dispersion. The advantage of such a method is to provide a probability forecast with a single integration. We could use a method similar to that of Palmer and Tibaldi (1986) to calculate a standard deviation by linear regression with the mean forecast, but our sample is too short. We shall investigate however in this direction. Let us suppose that we are able to forecast the RMS error of 500 hPa height over the northern hemisphere (by regression on the first Principal Components of this field, for example; but here we shall take the actual value). Set z the standardized value of the prediction. In the T42 experiment we take for z 0. ,1.5 ,-1.5 ,0.6 ,-0.6 for January 1981 , 1982 ,1983 ,1984 ,1985 respectively. We shall use as standard deviation

$$s(j,i) = s^*(j)\ (1 + b\ z(i)) \qquad (8)$$

where $s^*(j)$ is the standard deviation between the forecast individual months (they are 25 for the T42) at each grid point j and b a parameter to optimize so that the GRPS is minimum. If b is positive, an expected good deterministic forecast for 500 hPa height over the northern hemisphere ($z(i)<0$) will increase our confidence in the prediction of 850 hPa temperature over Europe (which seems more natural than the contrary). Figure 4 (and other results not shown concerning other fields or other areas) shows that the results of section 3.3 ,that correspond to b=0, are not improved .

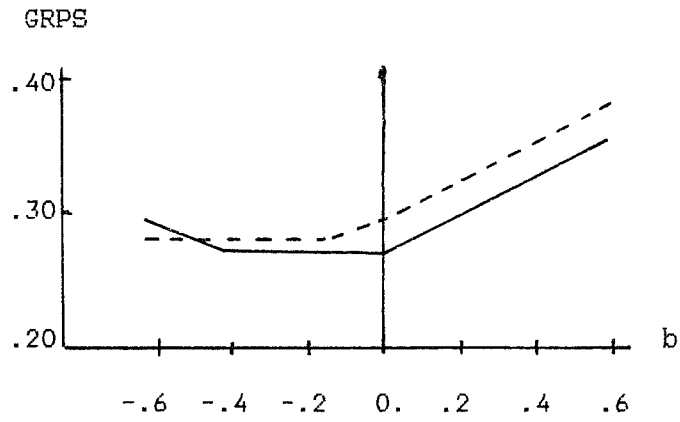


Fig. 4 :values of GRPS for 850 hPa temperature over Europe as a function of parameter b (see Eq. 8) ; T42 (solid line) and T21 (dashed line) experiment .

We have also tried to use the RMS error of 850 hPa temperature (instead of 500 hPa height).In this case we obtain lower values for the GRPS with b>0 than with b<0 but the results with b=0 are not improved .

3.5 Optimization of the standard deviation :second attempt

If we consider the intra-forecast dispersion, it would be a good estimate for the standard deviation if the model were perfect. Then the only source of error would be the increase of small analysis errors , due to the non-linearities in the model. We can consider that in fact there are two sources of error, model deficiencies and initial error growth, and we propose :

$$s(j,i) = (s_{intra}^2(j,i) + c\ e^2(j))^{1/2} \qquad (9)$$

where $s_{intra}(j,i)$ is the intra-ensemble standard deviation for year i , e(j) the total RMS error (including the systematic error) at grid point j, which has been calculated here with the 4 remaining years because of the shortness of the sample, and c a parameter to adjust. Fig.5 (and other results not shown concerning other fields or areas) shows that the best choice for c is about 0.2 for the T21 experiment and about 0.6 for the T42 one (which has smaller RMS errors). Anyway the values for

the GRPS found in 3.3 are not improved .

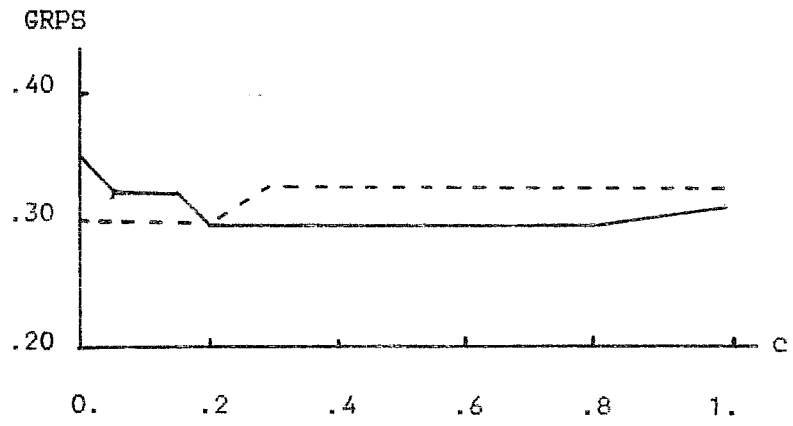


Fig.5 :as Fig.4 for parameter c (see Eq.9)

4. CONCLUDING REMARKS

When we want to visualize on a large area the results of a probability forecast, and to compare several cases, it is more useful to reduce the prediction to 3 categories and to represent the set of corresponding probabilities by a vector. When we want to provide the maximum information, the whole density is necessary. For practical reasons the forecast sample is often too short to use the empirical cumulative density. The gaussian hypothesis provides a satisfying approximation and involves only two parameters. If we believe in numerical models, we use as the average the ensemble mean (minus the model systematic error). Any other value should result of an artificial choice. But taking as the standard deviation the ensemble standard deviation yields clearly an underestimate since it comes to suppose the model is perfect .

The problem of the choice of the standard deviation is open, and a full optimization would require a large set of independent but homogeneous probability forecasts. With our five cases, it seems that a reasonable choice is something looking like the model natural variability. However it would be preferable that this standard deviation would vary from one

forecast to another, allowing for a discrimination between a priori good and a priori bad forecasts .

REFERENCES

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. Mon. Wea. Rev., 78, 1-3.

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

Déqué, M., 1988: The probabilistic formulation: a way to deal with ensemble forecasts. Annales Geophysicae, 6, in press.

Epstein, E.S. and A.H. Murphy, 1965: A note on the attributes of probabilistic predictions and the probability score. J. Appl. Meteor., 4, 297-299 .

Epstein, E.S.,1969: Stochastic dynamic prediction. Tellus, 21, 739-759.

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

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

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

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

Murphy, A.H., 1971: A note on the ranked probability score. J. Appl. Metor., 10, 155-156.

Palmer, T.N. and S. Tibaldi, 1986: Forecast skill and predictability .

ECMWF Technical Memorandum No. 127.