TECHNICAL MEMORANDUM

265

A probability and decision model analysis of PROVOST seasonal multi model ensemble integrations

T.N. Palmer, Č. Branković
and D.S. Richardson

Research Department

November 1998

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Abstract

A probabilistic analysis is made of seasonal ensemble integrations, with emphasis on the Brier score and related Murphy decomposition, and the relative operating characteristic. Results from the analysis of relative operating characteristic are input to a simple decision model. The decision model analysis is used to define a user-specific objective measure of the economic value of seasonal forecasts. The analysis is made for two simple meteorological forecast conditions or 'events' E based on 850hPa temperature. The ensemble integrations result from integrating four different models over the period 1979-1993. For each model a set of 9-member ensembles are generated by running from consecutive analyses.

Results from the Brier skill score analysis taken over all Northern Hemisphere grid points indicate that whilst the skill of individual models ensembles is only marginally higher than a probabilistic forecast of climatological frequencies, the multi-model ensemble is substantially more skilful than climatology. Both reliability and resolution are better for the multi-model ensemble, than for the individual-model ensembles. This improvement arises both from the use of different models in the ensemble, and from the enhanced ensemble size obtained by combining individual-model ensembles; the latter reason was found to be the more important. Brier skill scores are higher for years in which there were moderate or strong El Niño events. Restricting to Europe, only the multi-model ensembles showed skill over climatology. Similar conclusions are found from analysis of the relative operating characteristic.

Results from the decision-model analysis show that the economic value of seasonal forecasts is strongly dependent on the cost C to the user of taking precautionary action against E, in relation to the potential loss L if precautionary action is not taken and E occurs. However, based on the multi-model ensemble data, the economic value can be as much as 50% of the value of a hypothetical perfect deterministic forecast. For the hemisphere as a whole, value is enhanced by restricting to El Niño years. It is shown that there is economic value in seasonal forecasts for European users. However, the impact of El Ninño on economic value over Europe is mixed; value is enhanced by El Niño only for some potential users with specific C/L.

The techniques developed are applicable to complex events E for arbitrary regions. Hence these techniques are proposed as the basis of an objective probabilistic and decision-model evaluation of operational seasonal ensemble forecasts.

1 Introduction

This paper discusses the meteorological skill and potential economic value of a set of seasonal ensemble integrations of atmospheric global circulation models (GCMs). These integrations were made as part of the PROVOST project (**PR**ediction **O**f climate **V**ariations **O**n **S**easonal and interannual **T**imescales), supported by the European Union. Within the terms of this project, ensembles of 120-day integrations were made by four different modelling groups. The ensembles were made with prescribed observed sea surface temperature (SST) and initial dates spanned the period of the ECMWF 15-year reanalysis (1979-1993). More details concerning the models and the ensemble construction are given in section 2.

The primary purpose of this paper is to discuss the skill and value of seasonal ensemble forecasts within an explicitly probabilistic framework. Consider a meteorological condition or 'event' E defined over a particular season and particular spatial location. Examples of E might be: the seasonal- mean temperature is higher than normal, the seasonal-mean rainfall is one standard deviation lower than normal, there are at least 10 days where daily-average wind speed exceeds 20m/s etc. For each member of a forecast ensemble, E is predicted to occur, or not to occur. The ensemble as a whole provides a forecast probability $p(E)$ for E, based on the fraction of ensemble members in which E occurs.

There are a number of techniques available with which to assess the skill of the forecast probabilities $p(E)$; here we focus on two. The first is the reliability diagram and related Brier score discussed in section 3. In analysing the Brier score, we make use of Murphy's (1973) decomposition into reliability and resolution. The second technique is the relative operating characteristic (ROC; Stanski et al, 1989) discussed in section 4. This measures the success and false-alarm rates of the ensemble, made by assuming E will occur if it is forecast with a probability exceeding some specified probability threshold p_t.

Although objective measures of skill can be obtained from the Brier and ROC statistics, it is impossible to say what constitutes a level of useful skill for seasonal forecasts. The reason is simply that on the seasonal timescale (arguably on all timescales), it is not possible to define a level of useful skill that is independent of the needs of the forecast user. However, the ROC statistics provide the required input to a simple decision model which allows one to assess objectively the potential economic value of the ensemble seasonal forecasts from an (idealised) user perspective. The user has available the probabilistic forecast information to help decide whether to take (at some cost) some precautionary action to mitigate possible future weather-related loss or damage. The decision model applied to the PROVOST ensemble data, is discussed in section 5. It can be noted that readers only interested in the decision-model analysis can skip section 3 (but not 4) without substantial loss.

As mentioned above, the ensembles discussed here have been run with four different models. For each model, an ensemble comprises 9 integrations initiated from consecutive 24- hour analyses. The combination of ensembles from the four individual models provides a 36-member

multi-model ensemble. In some sense, this multi- model ensemble spans uncertainties in both initial conditions and model formulation. We assess, on the basis of the probabilistic forecast techniques discussed in this paper, whether results from the multi-model ensemble provide a more skilful and more valuable set of data than that obtained from any one individual model.

2 Experimental Details and Systematic Error

The four model formulations used in this study are as follows: a) ECMWF IFS cycle 13r4, T63L31 resolution with semi-Lagrangian timestepping; b) Météo-France Arpège cycle 12, T42L31 with Eulerian timestepping; c) the UKMO Unified Model HADAM3, 2.5x3.75 degree L19 resolution with Eulerian timestepping; d) as b) but with T63L31 resolution, the integrations being run by Eléctricité de France.

For each of the GCMs a)-c), 9-member ensembles were run over all seasons for the period 1979-93 (coinciding with the period of the ECMWF 15-year reanalysis, ERA-15; Gibson et al, 1997). In addition, a set of 9-member ensembles was created using model d) for the winter seasons. In this paper, attention is focussed on results for these winter seasons. For technical reasons, only 14 winter seasons from 1979/80-1992/93 from the 15 year reanalysis were studied.

For each model, the ensemble members were initiated from consecutive 12Z ERA-15 analyses, from 1 to 9 days preceding the beginning of the season (here taken as 1 December for the winter season). The length of each integration was 4 months plus 1 to 9 days, depending on the intial date. The integrations were run with prescribed observed SSTs, also taken from ERA-15, and updated daily in the integrations. Model output was archived every 24 hours at 12Z and stored in common format at ECMWF. Data from ERA-15 was used for verification purposes.

Monthly-mean SST anomalies were averaged over the area 7N-7S, 160E-80W in order to define an El Niño/Southern Oscillation (ENSO) index. Within the reanalysis period, two strong ENSO events occurred, a warm event in 1982/83 and a cold event in 1988/89. In addition there were moderate warm events in 1986/87 and 1991/92, and a prolonged cold event from 1984/1986. As in Brankovic and Palmer (1998), these are used in the analysis below to define a subset of ENSO years.

Before discussing the probabilistic skill scores associated with these ensembles, Fig 1 shows the 500-hPa systematic error from the 4 models (for January to March, JFM) The patterns of systematic error in the ECWMF and both Arpège models are quite similar; only the UKMO model has a distinct pattern of systematic error. Comparing the two Arpège models, it is clear from this study that seasonal-mean systematic error is not particularly sensitive to horizontal resolution in the range T42 to T63. Although it appears that both the ECMWF and Arpège models may have a common deficiency, it is difficult to draw any definite conclusions about the causes of the systematic error from these results. The ECMWF and Arpège models do

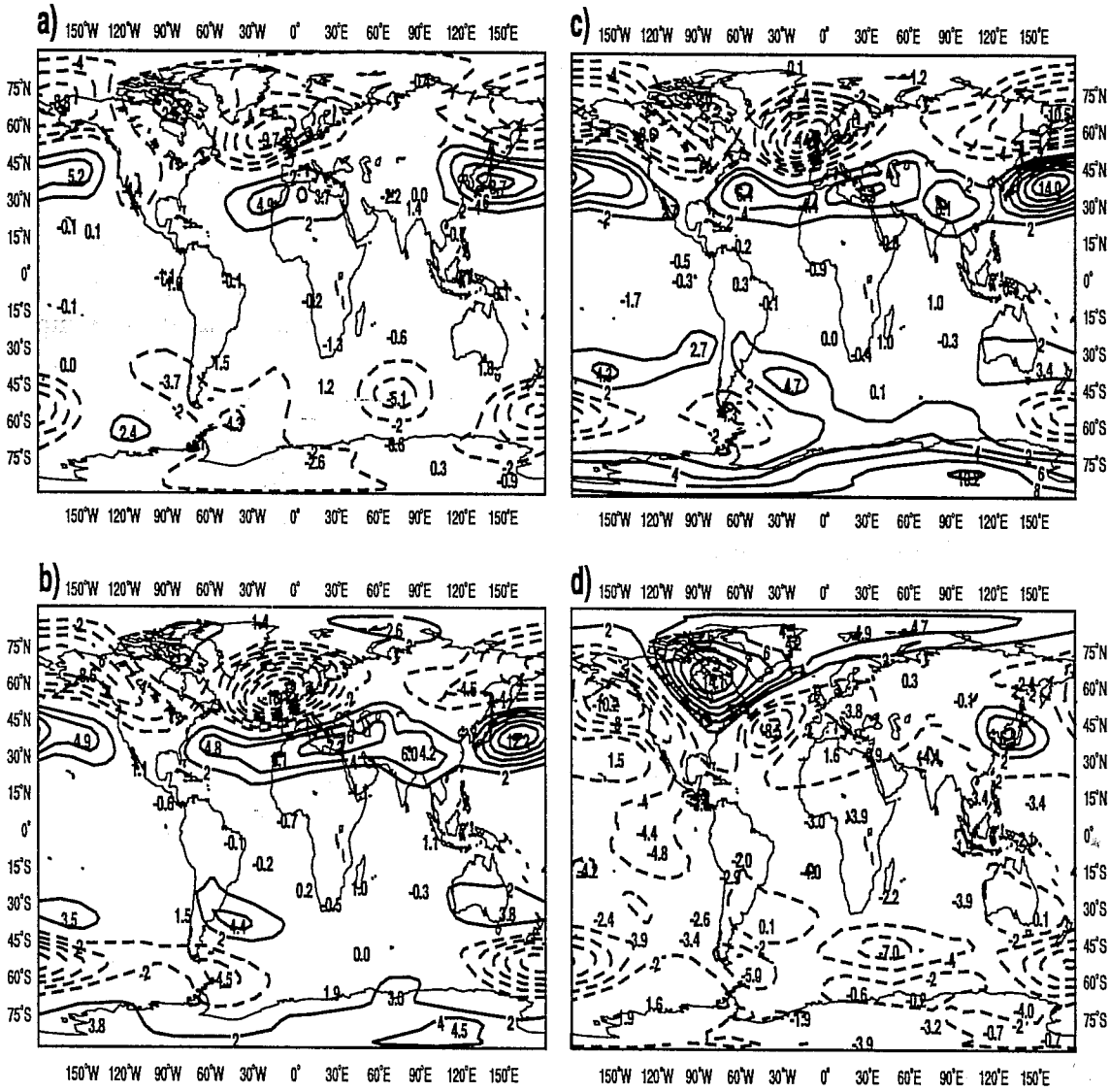

Figure 1: 500hPa height systematic error for January-February for a) ECMWF model, b) Arpège T42 model, c) Arpège T63 model, d) UKMO model.

not have the same set of physical parametrizations, and have somewhat different numerics (eg Lagrangian vs Eulerian timestepping), yet their mean systematic errors are similar. On the other hand the ECMWF and UKMO models have both different numerics and different physics.

The main purpose of showing the systematic error is to note that, in many parts of the extratropics, the systematic error of four state-of-the-art prediction models is comparable in magnitude with the standard deviation of atmospheric interannual variability (this is shown explicity for the ECMWF model in Brankovic and Palmer; 1998). The fact that the model error is as large as the signal to be predicted, lends some support to the use of a multi-model ensemble. On the other hand, the fact that three of the four models have similar systematic errors suggests that this particular choice of multi-model ensemble has an inherent bias, and may not be optimal.

For each model, the simulated climatology is defined as the mean of all ensemble members for that model, taken over the whole 14-winter period. A simulated anomaly field is defined with respect to the appropriate model climatology. In this way, a linear correction is applied to take account of some of the model error discussed above. When multi-model ensembles are constructed, we combine the anomaly fields computed in this way. The verifying analysis anomalies are computed using the ERA-15 climatology

3 The Brier Score and its Decomposition

As discussed in the introduction, we consider an event E which, for a particular ensemble forecast, occurs a fraction p of times within the ensemble. If, when the verifying analyses are available, E actually occurred then let $v = 1$. Otherwise $v = 0$. Repeat this over a sample of N different ensemble forecasts, so that p_i is the probability of E in the ith ensemble forecast and $v_i = 1$ or $v_i = 0$, depending on whether E occurred or not in the ith verification ($i = 1, 2...N$). In practice, the N ensemble forecasts could be taken not only from different years, but also from different gridpoints.

The Brier score (Brier, 1950) (used routinely to evaluate the ECMWF medium-range ensemble-forecast system; Molteni et al, 1996, Palmer et al, 1996, Talagrand et al, 1998) is defined by

$$b = \frac{1}{N}\sum_{i=1}^{N}(p_i - v_i)^2,\ 0 \leq p_i \leq 1,\ v_i \epsilon \{0, 1\} \qquad (1)$$

Like the conventional rms score, the Brier score is positive, equalling zero only in the ideal limit of a perfect deterministic forecast. The worse the forecast, the higher the Brier score (up to a maximum of 1 for a consistently incorrect deterministic forecast).

For a large enough sample, the Brier score can be written as

$$b = \int_0^1 [p-1]^2 o(p)g(p)dp + \int_0^1 p^2[1-o(p)]g(p)dp \tag{2}$$

where $g(p)$ is a probability density function (pdf) such that $g(p)dp$ is the relative frequency that E was forecast with probability between p and $p+dp$, and $o(p)g(p)dp$ gives the relative frequency of cases when E was forecast with probability p and $p+dp$, and E actually occurred. To see the relationship between (1) and (2) note that $\int_0^1 [p-1]^2 o(p)g(p)dp$ is the Brier score for ensembles where E actually occurred, and $\int_0^1 [p-0]^2(1-o(p))g(p)dp$ is the Brier score for ensembles where E did not occur.

Simple algebra on (2) gives

$$b = \int_0^1 [p-o(p)]^2 g(p)dp - \int_0^1 [\bar{o}-o(p)]^2 g(p)dp + \bar{o}[1-\bar{o}] \tag{3}$$

where

$$\bar{o} = \int_0^1 o(p)g(p)dp \tag{4}$$

is the (sample) climatological frequency of E. This is Murphy's (1973) decomposition of the Brier score.

There are three terms on the right-hand side of (3). The first is the 'reliability'

$$b_{rel} = \int_0^1 [p-o(p)]^2 g(p)dp. \tag{5}$$

A reliability diagram (Wilks, 1995) is one in which $o(p)g(p)\delta p$ is plotted against $pg(p)\delta p$ for some finite binning of width δp. In a perfectly reliable system $o(p) = p$ and the graph is a straight line oriented at deg 45 to the axes, and $b_{rel} = 0$. Reliability measures the mean square distance of the graph of $o(p)$ to the diagonal line.

The second term on the right-hand side of (3) is 'resolution'. A simple but reliable probability forecast would be to always predict the climatological probability $\bar{o}$ of E. However, such a forecast would have no power of discrimination between years because the same forecast would be made every year). Resolution is defined as

$$b_{res} = \int_0^1 [\bar{o}-o(p)]^2 g(p)dp \tag{6}$$

Note that the negative of b_{res} enters the Brier score decomposition. Hence the more skilful the system, the larger is b_{res}. A system with relatively high b_{res} is one where the dispersion of $o(p)$ about $\bar{o}$ is as large as possible. Conversely, a forecast system has no resolution when, for all forecast probabilities, the event verifies a fraction $o(p) = \bar{o}$ times. Resolution measures the mean square distance of the graph of $o(p)$ to the sample climate horizontal line.

The third term on the right-hand side of (3) is the 'uncertainty'

$$b_{unc} = \bar{o}[1 - \bar{o}] \tag{7}$$

and ranges from 0 to 0.25. If E was either so common, or so rare, that it either always occurred or never occurred within the sample of years studied, then $b_{unc} = 0$; if E occurred 50% of the time within the sample, then $b_{unc} = .25$. Uncertainty is a function of the climatological frequency of E, and is not dependent on the forecasting system itself. It can be shown that the resolution of a perfect deterministic system is equal to the uncertainty.

When assessing the skill of a forecast system, it is often desirable to compare it with the skill of a forecast where the climatological probability $\bar{o}$ is always predicted. The Brier score of such a climatological forecast is $b_{cli} = b_{unc}$ (using the sample climate), since, for such a climatological forecast $b_{rel} = b_{res} = 0$. In terms of this, the Brier skill score, B, of a given forecast system is defined by

$$B = 1 - b/b_{cli}. \tag{8}$$

$B \leq 0$ for a forecast no better than climatology, and $B = 1$ for a perfect deterministic forecast.

Skill-score definitions can similarly be given for reliability and resolution, ie

$$B_{rel} = 1 - b_{rel}/b_{cli} \tag{9}$$
$$B_{res} = b_{res}/b_{unc} \tag{10}$$

the latter recognising, as noted above, that the largest possible value of b_{res} is equal to b_{unc}. Hence, for a perfect deterministic forecast system, $B_{rel} = B_{res} = 1$.

We illustrate these skill measures in Figs 2-3, which are reliability diagrams for the ECMWF-model ensemble, and for the multi-model ensemble (using all 4 models), respectively. For Fig 2a and 3a, the event $E_{<0}$ is: December-February (DJF) seasonal mean 850hPa temperature anomaly is below normal. For Fig 2b and 3b, the event $E_{<-1}$ is: DJF 850hPa temperature anomaly is less than -1K. All points in the extratropical Northern Hemisphere (NH), and all years in the PROVOST data set are used. For each event, the horizontal axis gives the forecast probability p, the vertical axis gives the observed frequency $o(p)$, based on a binning of data in probability bins of width 0.1 (or 0.05 in the case of the extreme probabilities). The number of occurrences within these bins are given next to the data points. The frequency of occurrences are also plotted as histograms next to the reliability diagrams; these give the pdf $g(p)$. The sample-mean climatological probability of occurrence $\bar{o}$ for E is shown as the horizontal dashed line on the reliability diagram. For $E_{<0}$ (Fig 2a and 3a), $\bar{o} = 0.5$; for $E_{<-1}$ (Fig 2b and 3b), $\bar{o} \sim 0.2$. For each reliability diagram B, B_{rel} and B_{res} are also shown in the top left-hand corner.

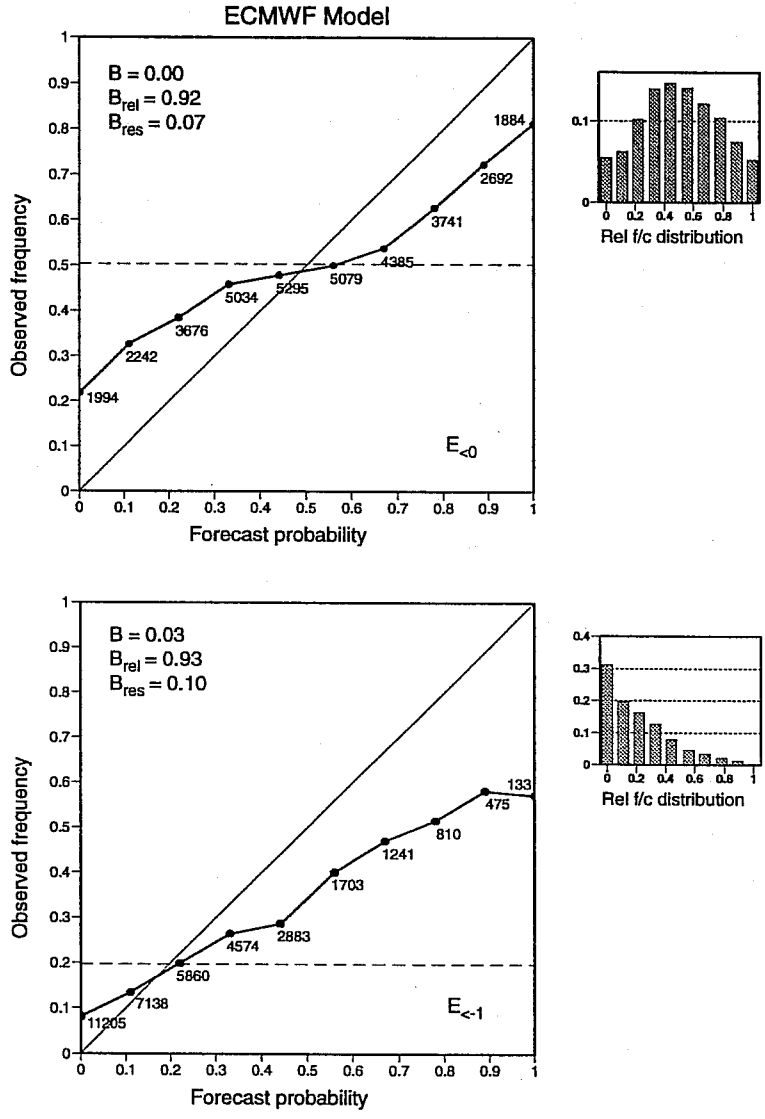

Figure 2: Reliability diagram, Brier skill score, and Murphy decomposition, for $E_{<0}$ (top) and $E_{<-1}$ (bottom) for ECMWF-model ensemble. All extratropical NH points. The values on the reliability graph give the number of occurences where E was forecast within each of the probability category bins. These values are also plotted as a frequency distribution to the right of the reliability diagram.

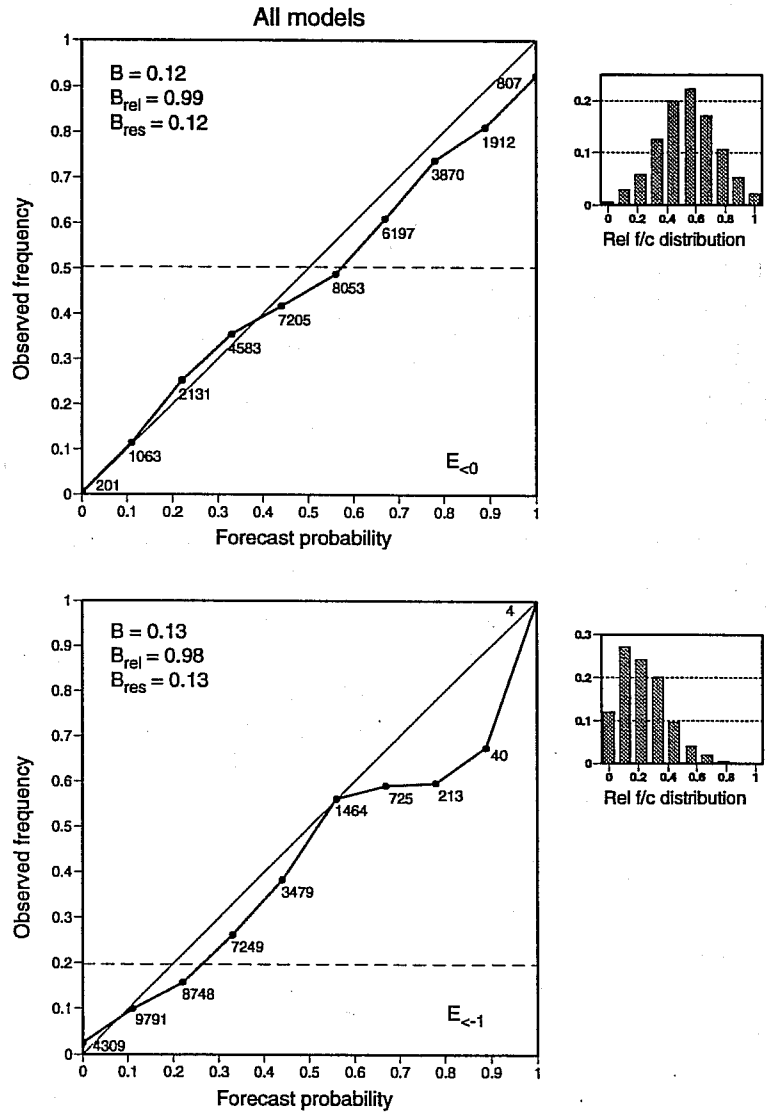

Figure 3: As for Fig 2 but multi-model ensemble.

For the ECMWF model (Fig 2), the Brier skill score B for $E_{<0}$ is equal to zero; ie, for this event, the skill of probability forecasts is no better than the skill of a climatological probability forecast. Looking at the Murphy decomposition, the ensemble probability forecasts are certainly reliable ($B_{rel} = 0.92$), and graph of $o(p)$ has positive slope, and parallels the diagonal in places. On the other hand, the probability forecasts are 'over confident'. For example, for occasions where all ensemble members predict $E_{<0}$, the event only verifies 80% of the time. Similarly, for occasions where no members forecast the event, it occurs about 20% of the time. Since such cases of forecast unanimity are relatively rare, this over-confidence does not contribute substantially to B_{rel} (which could be taken as a shortcoming of the score). In contrast to reliability, the ensembles have poor resolution $B_{res} = 0.07$. The corresponding pdf is strongly weighted towards the climatological frequency where $o(p)$ is tilted towards the climatological frequency line.

For $E_{<-1}$, the Brier skill score for the ECMWF model is slightly positive $B = 0.03$ and therefore indicates somewhat higher skill than that of a climatological forecast. Looking at the decomposition, it can be seen that the reliability is only slightly better than for $E_{<0}$. At first sight this may appear surprising given that the slope is consistently less than that of the diagonal. However, the slope in the vicinity of the pdf maximum is similar to that for $E_{<0}$. On the other hand, the resolution is clearly better ($B_{res} = 0.10$) than for $E_{<0}$.

By contrast, the Brier skill score for the multi-model ensemble (Fig 3) is better than climatology, and better than the ECMWF model, for both events ($B = 0.12$ for $E_{<0}$, $B = 0.13$ for $E_{<-1}$). Visually, it can be seen that the reliability is considerably improved for both events, and for $E_{<0}$, the graph of $o(p)$ is very close to the diagonal (where $B_{rel} = 0.99$). In addition, B_{res} is improved using the multi-model ensemble. In some respects the increase in resolution might appear surprising, since the pdf of the multi-model ensemble is more sharply peaked towards the climatological frequency of the event in question. However, the tilting of the graph of $o(p)$ away from the horizontal in the multi-model ensemble, more than offsets for the sharpening of the pdf.

Tables 1-2 give a summary of Brier skill scores for all extratropical NH gridpoints and for the subset of all European gridpoints respectively. In each table the top sub-table is for $E_{<0}$, the bottom is for $E_{<-1}$. For both events, and for both regions, it can be seen that the multi-model Brier skill score is more skilful than the equivalent score from any of the single model ensembles. Note that in the case of the European region, the Brier skill score is either negative or identically zero for all individual models, whilst the multi- model scores are (just!) positive.

The improvement in multi-model Brier skill score over the ECMWF skill score can arise both from the use of multiple models, and from the fact that the multi-model ensemble is four times larger than the individual-model ensembles. In order to assess which of these factors was most important, Brier skill scores have been estimated by constructing a 9- member ensemble, taking the first three members from the ECMWF, T63-Arpege, and UKMO ensembles. For the extratropical NH, this 9-member multi-model ensemble has $B = 0.04$ for $E_{<0}$ and $B = 0.06$ for $E_{<-1}$. For both events, this corresponds to about one third of the improvement over the

ECMWF ensemble when using the full 36-member multi-model ensemble. On this basis, the improvement in the multi-model ensemble over the individual model ensemble arises partially from the use of more than one model, but, perhaps more importantly, from the larger ensemble size implied by combining individual-model ensembles.

As discussed in Brankovic and Palmer (1998), the ensemble-mean skill is improved if one restricts onself to ENSO years within the PROVOST integration period. Tables 1-2 also show the Brier skill scores calculated over the subset of 5 ENSO winters. For the NH, there is an improvement in skill using this subset of years, for each individual-model ensemble, and for the multi-model ensemble. For Europe, on the other hand, Brier skill scores are actually worse, restricting to ENSO years. This impact of ENSO over Europe, between different measures of skill, will be further discussed in the sections below.

As discussed above, the theoretical maximum value for B is unity, corresponding to a perfect deterministic forecast. This theoretical maximum is not a reasonable least upper bound on what can be achieved, since, even with a perfect model, inevitable uncertainties in initial conditions lead to chaotic variability within the ensemble. One way of estimating a more realistic upper bound on skill is to verify the ensemble against one of its members (chosen at random). Equivalently, one can calculate a 'perfect- model' Brier skill score, B_{per}, by taking the $g(p)$ as given by the ensemble, and putting $o(p) = p$ (perfect reliability) in (3) and (4). In this way, B_{per} is defined only from the first and second moments of p with respect to $g(p)$. Tables 1-2 show B_{per} using each of the four model ensembles. Note that the perfect-model estimates are model dependent; this is not surprising given that the degree of internal model variability will be model dependent. (For models with relatively small internal variability, the perfect-model estimate of Brier skill score will be larger.) For the NH, the average perfect-model estimates are $\bar{B}_{per} = 0.25$ for both $E_{<0}$ and $E_{<-1}$. For Europe, the average Brier skill score is a little lower $\bar{B}_{per} = 0.17$ for both $E_{<0}$ and $E_{<-1}$, but is certainly above zero. hence there is potential seasonal predictability over Europe. It can be noted that B_{per} has deliberately not been estimated for the multi-model ensemble. This is because there are clear differences between model climatologies associated with systematic rather than random uncertainties between model formulation. In this sense, the multi-model ensemble cannot be considered as a representative of a perfect-model ensemble.

4 Relative Operating Characteristic

The relative operating characteristic (ROC; Stanski et al, 1989) is another test of the performance of a probabilistic forecast. It is based on the notion that a prediction of E is assumed, providing E is forecast by at least a fraction $p = p_t$ of ensemble members, where the threshold p_t is defined a priori. ROC is used routinely to evaluate the performance of the operational ECMWF medium-range ensemble forecasts (Buizza et al, 1998).

a)	ECMWF	EDF	Météo France	UKMO	All Models
All years	0.00	0.08	0.08	-0.01	0.12
ENSO years	0.07	0.19	0.12	0.10	0.20
Perfect Model	0.29	0.22	0.23	0.24	
b)	ECMWF	EDF	Météo France	UKMO	All Models
All years	0.03	0.08	0.04	0.01	0.13
ENSO years	0.08	0.16	0.10	0.11	0.20
Perfect Model	0.30	0.24	0.23	0.22	

Table 1: Brier Skill Score. NH grid points. a) $E_{<0}$ b) $E_{<-1}$

a)	ECMWF	EDF	Météo France	UKMO	All Models
All years	-0.05	-0.05	-0.05	-0.01	0.05
ENSO years	-0.25	-0.11	-0.07	-0.05	-0.3
Perfect Model	0.21	0.14	0.19	0.14	
b)	ECMWF	EDF	Météo France	UKMO	All Models
All years	-0.03	0.00	-0.12	0.02	0.05
ENSO years	-0.07	-0.08	-0.15	-0.03	-0.01
Perfect Model	0.20	0.15	0.16	0.15	

Table 2: Brier Skill Score. European grid points. a) $E_{<0}$ b) $E_{<-1}$

Consider first a deterministic forecast of E (either that it will occur, or that it will not occur). Over a sufficiently large sample of independent forecasts, we can form the forecast-model contingency matrix giving the frequency that E occurred or not, and whether it was forecast or not, ie

		Occurs	
		No	Yes
Forecast	No	α	β
	Yes	γ	δ

Based on these values, the so-called 'hit rate' (H) and 'false-alarm rate' (F) for E are given by

$$\begin{aligned} H &= \delta/(\beta+\delta) \\ F &= \gamma/(\alpha+\gamma). \end{aligned} \tag{11}$$

Hit and false alarm rates for an ensemble forecast can be defined as follows. Suppose it is assumed that E will if the forecast probability $p > p_t$ (and will not occur if $p < p_t$). By varying p_t between 0 and 1 we can define $H = H(p_t)$, $F = F(p_t)$. In terms of the pdf $g(p)$

$$\begin{aligned} H(p_t) &= \int_{p_t}^{1} o(p)g(p)dp/\bar{o} \\ F(p_t) &= \int_{p_t}^{1} (1-o(p))g(p)dp/(1-\bar{o}) \end{aligned} \tag{12}$$

The ROC curve is a plot of $H(p_t)$ against $F(p_t)$. A measure of skill is given by the area under the ROC curve (A_{ROC}). A perfect deterministic forecast will have $A_{ROC} = 1$, whilst a no-skill forecast for which the hit and false alarm rates are equal, will have $A_{ROC} = 0.5$. As discussed in the next section, the ROC curve values are of direct use in assessing the user- specific value of a probability forecast, based on decision-model analysis.

Figs 4-5 are the same as Figs 2-3 but for ROC rather than reliability. (The values $H(p_t)$ and $F(p_t)$ have been estimated in terms of bins of width 0.1). The A_{ROC} value is shown in each figure. Results are broadly in agreement with those from the previous section. For both the ECMWF-model ensemble, and the multi-model ensemble, $E_{<-1}$ is predicted more skilfully than $E_{<0}$. In all cases, A_{ROC} is greater than the no-skill value of 0.5. For both events, the multi-model ensemble is more skilful than the ECMWF-model ensemble.

Tables 3-4 are the same as Tables 1-2 but for A_{ROC} rather than B. For both events and for the NH and Europe, it can be seen that the multi-model A_{ROC} is higher than that for any of the

a)	ECMWF	EDF	Météo France	UKMO	All Models
All years	0.648	0.680	0.683	0.626	0.693
ENSO years	0.707	0.756	0.715	0.712	0.753
Perfect Model	0.807	0.768	0.773	0.777	
b)	ECMWF	EDF	Météo France	UKMO	All Models
All years	0.706	0.715	0.702	0.656	0.736
ENSO years	0.752	0.780	0.750	0.741	0.800
Perfect Model	0.852	0.832	0.827	0.828	

Table 3: Area under ROC curve. NH grid points. a) $E_{<0}$ b) $E_{<-1}$

a)	ECMWF	EDF	Météo France	UKMO	All Models
All years	0.593	0.562	0.596	0.596	0.625
ENSO years	0.509	0.575	0.617	0.559	0.590
Perfect Model	0.762	0.709	0.748	0.711	
b)	ECMWF	EDF	Météo France	UKMO	All Models
All years	0.626	0.626	0.520	0.617	0.628
ENSO years	0.621	0.576	0.555	0.632	0.614
Perfect Model	0.792	0.775	0.763	0.760	

Table 4: Area under ROC curve. European grid points. a) $E_{<0}$ b) $E_{<-1}$

individual model ensembles. For the NH, for both events, the single-model and the multi-model A_{ROC} is higher for the set of ENSO years than for the whole dataset. For Europe, there is no consistent improvement or degradation by restricting to ENSO years; for example, for $E_{<0}$ both Arpège models show an improvement in A_{ROC} by restricting to ENSO years, whilst ECMWF and UKMO (and the multi-model ensemble) show a degredation.

Perfect-model A_{ROC} values are also shown in Tables 3-4. As for the Brier skill score, perfect-model hit and false alarm rates were estimated from the ensemble pdf $g(p)$, putting $o(p) = p$, ie assuming perfect reliability. Hence, for example, for the perfect-model ensemble

$$H_{per}(p_t) = \int_{p_t}^{1} pg(p)dp / \int_{0}^{1} pg(p)dp \tag{13}$$

As with perfect-model Brier skill score, there is some variation in A_{ROC} between the different models. Hemispheric values are larger than European values, though the latter are substantially higher than the no-skill value of 0.5, again demonstrating potential seasonal predictability over Europe.

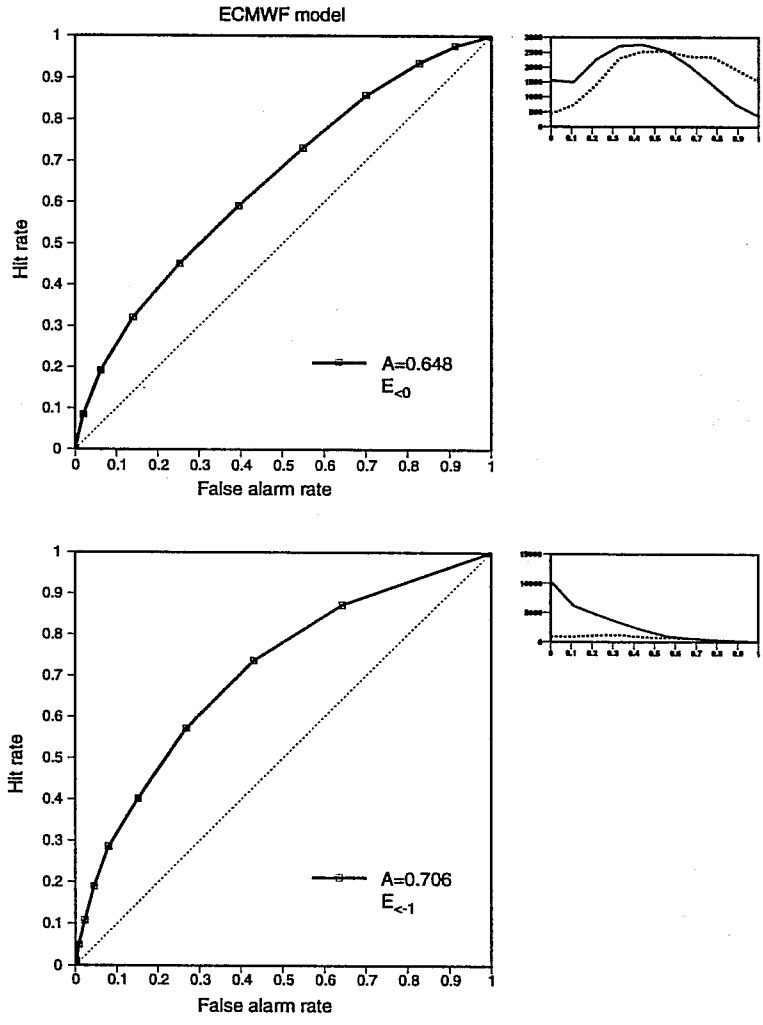

Figure 4: As Fig 2 but for ROC curve.

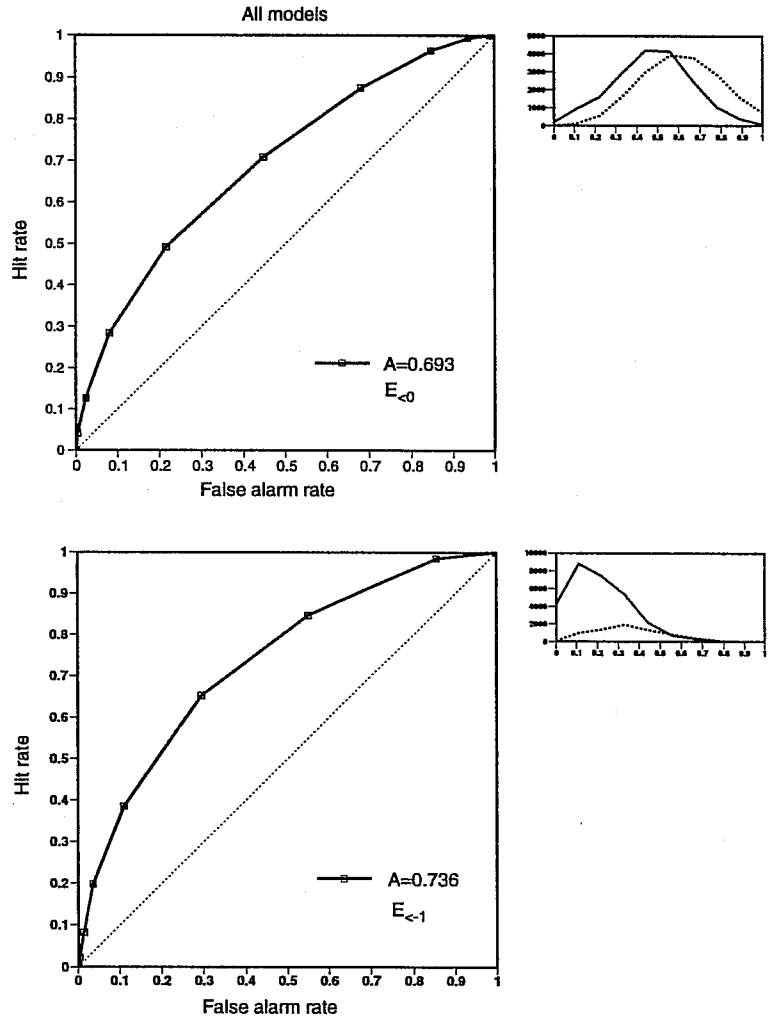

Figure 5: As Fig 4 but for multi-model ensemble.

5 Decision-Model Analysis

Although B and A_{ROC} provide objective measures of skill for ensemble forecasts, it is difficult to say what constitutes a threshold of useful skill for seasonal forecasts. This is not surprising since 'usefulness' is a user-specific concept. In an attempt to define 'usefulness' objectively, we consider here a simple decision model (Murphy, 1977; Katz and Murphy, 1997) whose inputs are the the hit and false-alarm rate estimates $H(p_t)$ and $F(p_t)$. This decision model has been used by Richardson (1998) to define the economic value of ECMWF medium-range ensemble forecasts.

Consider a potential forecast user who can take some specific precautionary action depending on the likelyhood that E will occur. Taking precautionary action incurs a cost C irrespective of whether or not E occurs. However, if E occurs and no action has been taken, then a loss L is incurred. The expense associated with each combination of action/innaction and occurrence/non-occurrence of E is given in the decision-model contingency matrix

		Occurs	
		No	Yes
Take Action	No	0	L
	Yes	C	C

The decision maker wishes to pursue the strategy which will minimise expenses over a large number of cases.

If only climatological information on the frequency $\bar{o}$ of E is available, there are two basic options: either always or never take precautionary action. Always taking action incurs a cost C on each occasion, whilst never taking action inurs a loss L only on the proportion $\bar{o}$ of occasions when E occurs, giving an expense $\bar{o}L$.

The purpose of this section is to analyse whether the PROVOST forecast data, if it were used by the hypothetical decision maker, would reduce expenses beyond what could be achieved using climatological information alone. Consider first a deterministic forecast system with characteristics described by the forecast-model contingency matrix in section 4. Then, using the forecast and decision contingency values, the user's expected mean expense M (per unit loss) is

$$M = \frac{\beta L + (\gamma + \delta)C}{L} \tag{14}$$

This can be written in terms of the hit-rate H and the false-alarm F using (11), so that

$$M = F\frac{C}{L}(1-\bar{o}) - H\bar{o}(1-\frac{C}{L}) + \bar{o} \tag{15}$$

For a perfect deterministic forecast $H = 1$, $F = 0$, hence

$$M_{per} = \bar{o}\frac{C}{L} \tag{16}$$

To calculate the mean expense per unit loss knowing only climatology, suppose first the decision maker always protects, then $M = C/L$ (equivalent to using a forecast system where the event is always predicted and for which $H = 1$ and $F = 1$). Conversely, if the decision maker never protects then $M = \bar{o}$ (equivalent to using a forecast system where the event is never predicted and for which $H = 0$ and $F = 0$). Hence if the decision maker knows only the climatological frequency $\bar{o}$, M can be minimised by either always or never taking precautionary action, depending on whether $C/L < \bar{o}$, or $C/L > \bar{o}$ respectively. Hence, the mean expense per unit loss associated with a knowledge of climatology only, is

$$M_{cli} = min(\frac{C}{L}, \bar{o}). \tag{17}$$

We define the value V of forecast information to be a measure of the reduction in M over M_{cli}, normalised by the maximum possible reduction associated with a perfect deterministic forecast, ie

$$V = \frac{M_{cli} - M}{M_{cli} - M_{per}} \tag{18}$$

For a forecast system which is no better than climate, $V = 0$; for a perfect deterministic forecast system $V = 1$.

As discussed in Section 4, an ensemble forecast gives hit and false-alarm rates $H = H(p_t)$, $F = F(p_t)$, as a function of probability thresholds p_t (see (12)). Hence V is defined for each p_t, ie $V = V(p_t)$. Using (15), (16) and (17)

$$V(p_t) = \frac{min(\frac{C}{L}, \bar{o}) - F(p_t)\frac{C}{L}(1-\bar{o}) + H(p_t)\bar{o}(1-\frac{C}{L}) - \bar{o}}{min(\frac{C}{L}, \bar{o}) - \bar{o}\frac{C}{L}}. \tag{19}$$

For given C/L and event E, the optimal value is

$$V_{opt} = \max_{p_t} V(p_t). \tag{20}$$

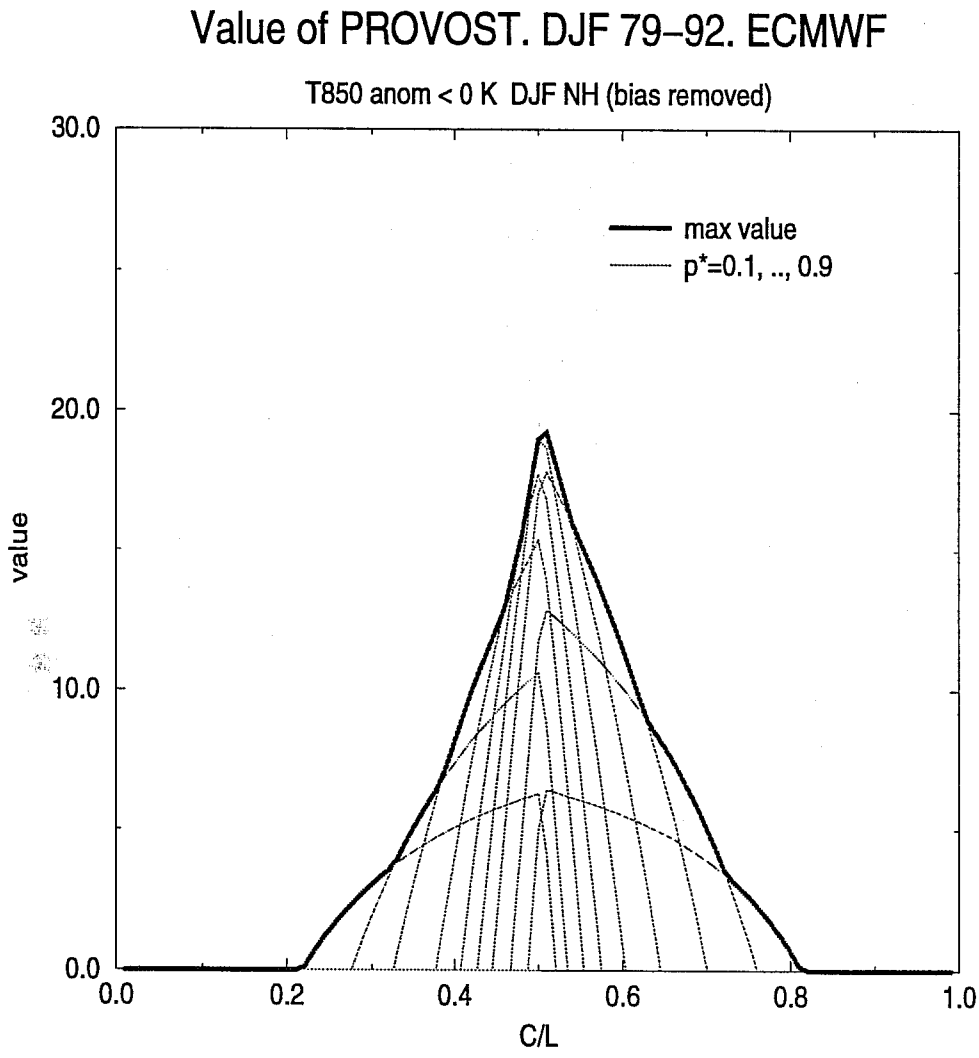

Figure 6: $V(p_t)$ for $p_t = 0.1, \ldots, 0.9$, together with the optimal envelope value V_{opt}, for $E_{<0}$. Extratropical NH grid points.

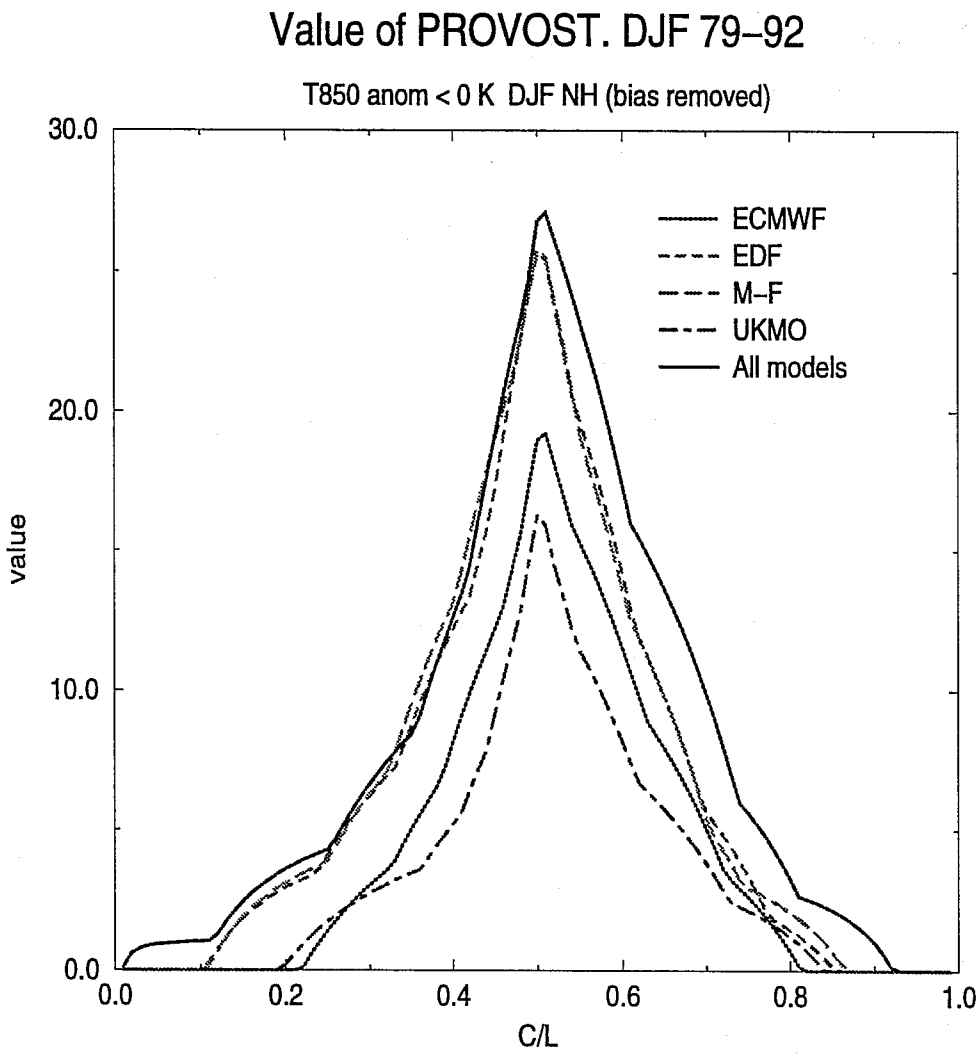

Figure 7: V_{opt} for $E_{<0}$. Individual-model ensembles and multi-model ensembles. Extratropical NH grid points.

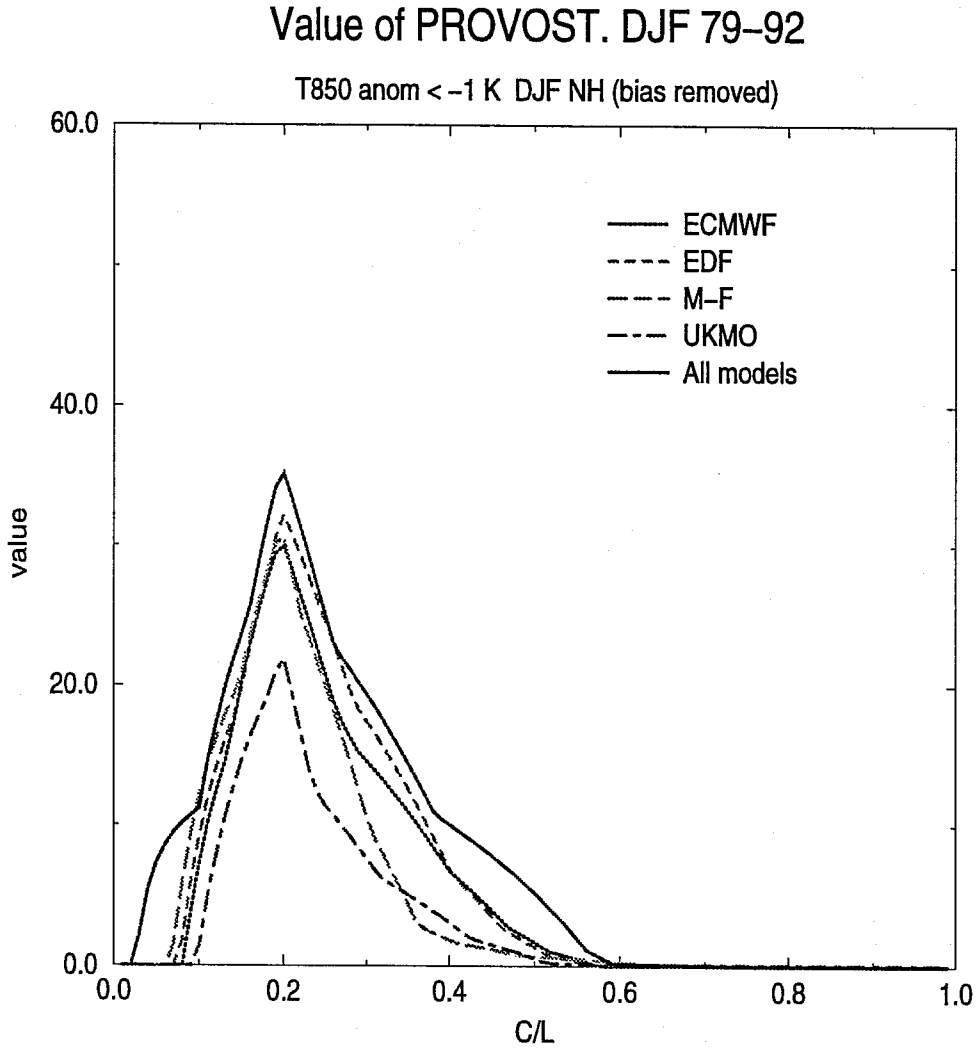

Figure 8: As Fig 7 but for $E_{<-1}$

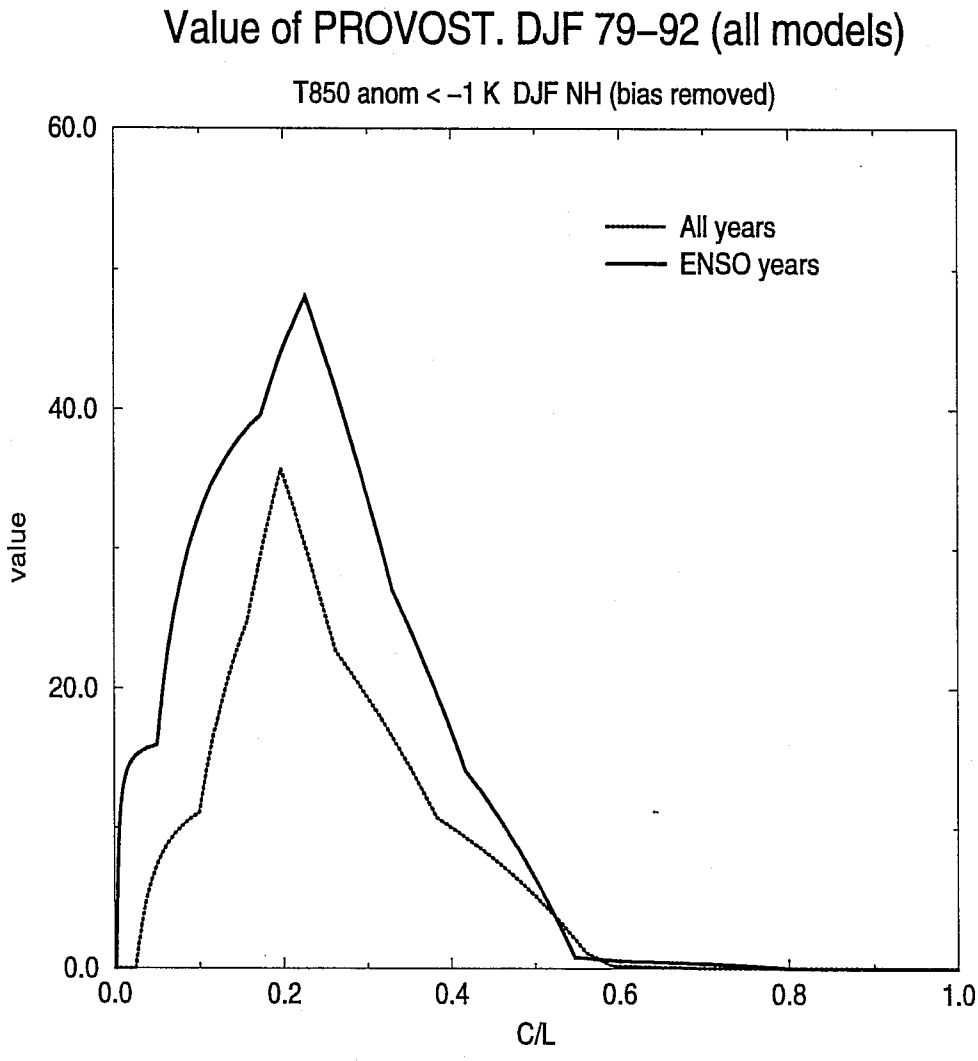

Figure 9: V_{opt} for $E_{<-1}$. Multi-model ensemble. Dashed: all years. Solid: ENSO years only. Extratropical NH grid points.

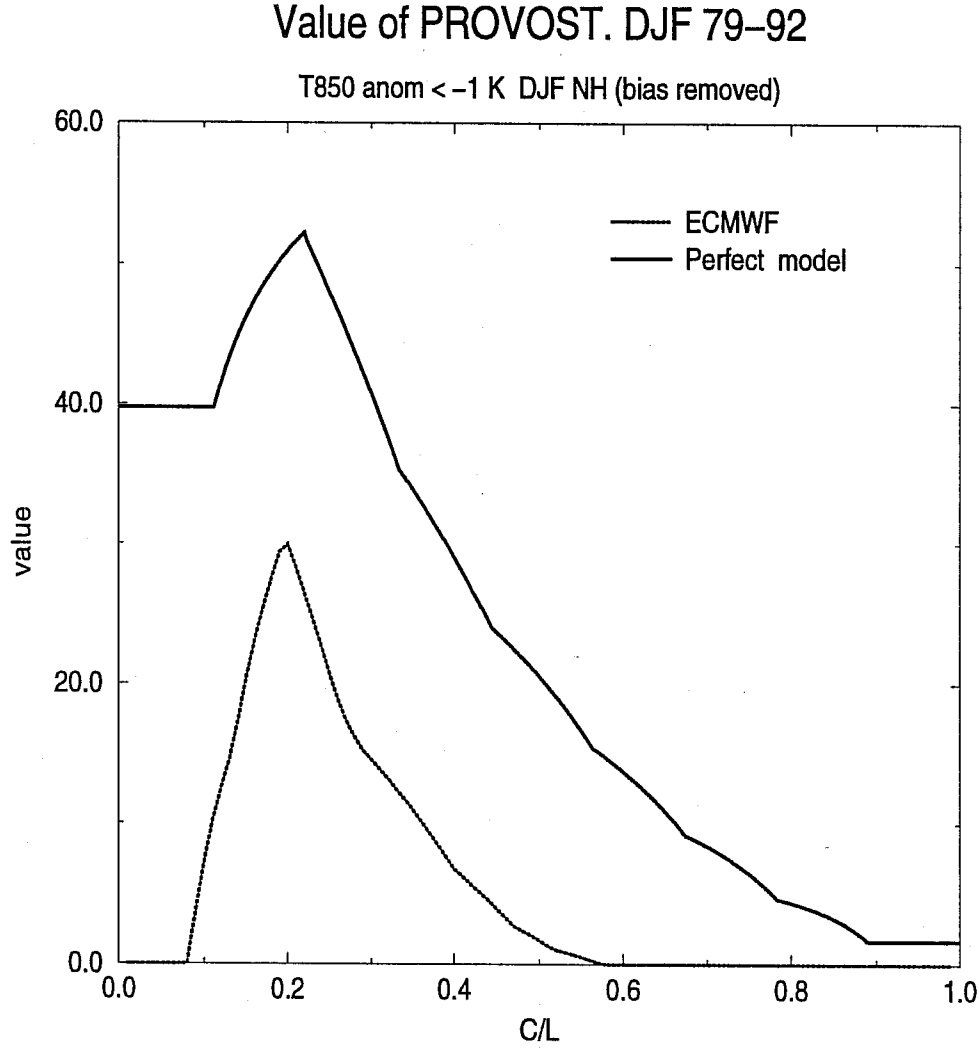

Figure 10: V_{opt} for $E_{<-1}$. Solid: Perfect-model ensemble (based on ECMWF model). Dotted - ECMWF-model ensemble (verified against real world). Extratropical NH grid points.

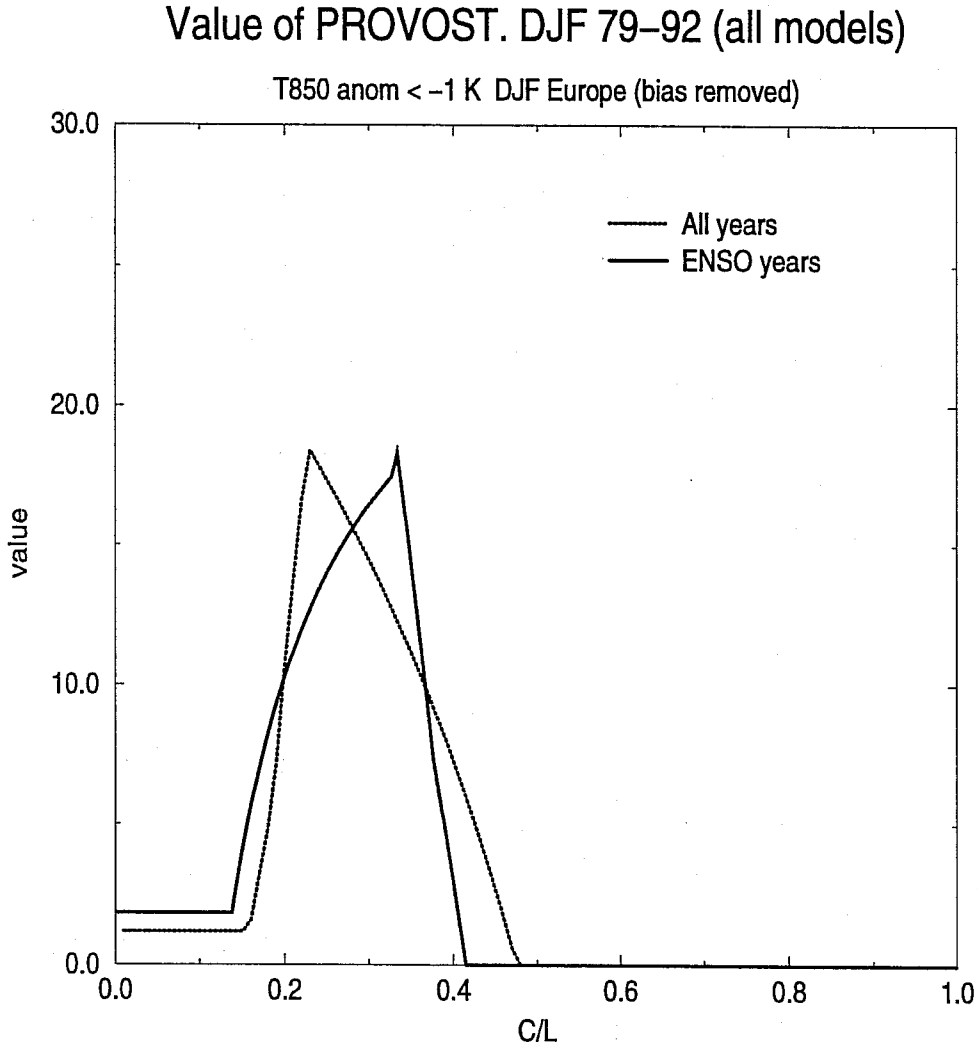

Figure 11: As Fig 9, but for European grid points only.

In Fig 6, $V(p_t)$ and V_{opt} are shown for $E_{<0}$, as a function of C/L, based on ECMWF ensemble data only, and taken over all extratropical NH grid points. The thin lines show the individual graphs $V(p_t)$ for $p_t = 0.1, 0.2, \ldots, 0.9$. For small p_t, the graph indicates value above climatology for users with C/L between about 0.2 and 0.5. For large p_t, $V(p_t)$ is positive for users with C/L between about 0.5 and 0.8. Consequently, the envelope function V_{opt} shows value for all users with C/L between about 0.2 and 0.8.This illustrates the benefit of probabilistic forecasts over deterministic forecasts. The value curve for a deterministic forecast would be no better than that of a single $V(p_t)$ curve, since a deterministic forecast has only one hit and false-alarm rate associated with it.

Fig 7 shows V_{opt} for $E_{<0}$ for all the individual-model ensembles and for the multi-model ensemble. For low C/L users, the value of the multi-model ensemble is about the same as the value obtained by two of the individual models. However, for users with $C/L > 0.5$, the multi-model ensemble provides the highest value.

For both Fig 6 and Fig 7, it can be noted that V_{opt} peaks for users with $C/L \sim 0.5$, which is the climatological frequency of $E_{<0}$. Climatological information is of no value to users whose cost-loss ratio is close to the climatological frequency of E. In this sense it is not surprising that the value of PROVOST data over climatology, is largest for users whose cost/loss ratio $C/L \sim \bar{o}$. (For these users, the cost of always or never taking precautionary action is about the same).

Fig 8 shows V_{opt} for $E_{<-1}$ for all the individual-model ensembles and for the multi-model ensemble. (Note that the scale of the ordinate axis is different in Figs 7 and 8.) For this event, maximum value occurs for users with $C/L \sim 0.2$, the climatological frequency of $E_{<-1}$. The value of the multi-model ensemble data, for this event, exceeds that of any of the individual models for most cost/loss ratios between 0 and .6. The value of forecasts of the $E_{<-1}$ event is significantly higher than that of the $E_{<0}$ event for $C/L \sim \bar{o}$, consistent with the fact (see sections 3 and 4) that both B and A_{ROC} were higher for $E_{<-1}$.

In earlier sections, we have noted that skill is improved if one restricts to ENSO years. Fig 9 shows V_{opt} for $E_{<-1}$ for the multi-model ensemble data restricted to ENSO years, for all extratropical NH points. For user with $C/L \sim \bar{o}$ the value of the forecast information is close to 1/2 of that obtainable from a hypothetical perfect deterministic forecast.

Again, as mentioned in earlier sections, the notion of a hypothetical perfect deterministic forecast is not a realistic one for seasonal forecasts. Fig 10 shows the value for a perfect-model ensemble made by using the hit and false alarm rates estimated under the perfect-model assumption (see Section 4). This is done for the ECMWF ensemble (for $E_{<-1}$). The actual value of the ECMWF ensemble is also shown for reference. On the basis of this curve, it could be said that seasonal forecasts of $E_{<-1}$ are potentially valuable to the entire range of users, though present levels of model error restrict usefulness to users whose C/L lies bewteen about 0,1 and 0.5. On the other hand, for users with high C/L, the potential value of forecast of $E_{<-1}$ is small compared with that from a hypothetical perfect deterministic forecast.

Finally, Fig 11 shows V_{opt} for $E_{<-1}$ from the multi-model ensemble over Europe. Comparing with Fig 7, it can be seen that the value of the PROVOST forecasts over Europe is not as great as that taken over all extratropical grid points (note different scaling of the ordinate axis in Fig 8 and 11). However, neither is it zero. For users with $C/L \sim \bar{o}$, the value is close to 20% of that obtainable with a hypothetical perfect deterministic forecast. It can be seen that restricting to ENSO years, improves value for some users, decreases value for other users. The fact that the impact of ENSO on value over Europe is user dependent, is consistent with the mixed signal obtained earlier, for the impact of ENSO on B and A_{ROC} over Europe.

6 Summary and Conclusions

Seasonal forecasting is inherently probabilistic (eg Palmer and Anderson, 1994). This has been recognised in the past through calculations of ensemble means and ensemble standard deviations - estimates of the first and second moments of the forecast probability distribution function. However, the approach in this paper is motivated by the need to develop a general methodology to evaluate the skill (and usefulness) of user-specific seasonal forecasts. Although in general terms, a user will want to know whether it will be rainy/dry, warm/cold, windy/settled, for more quantitative applications, forecasts should be tailored to specific user needs. For example, a user might want to know whether the seasonal-mean rainfall will be at least a standard deviation above average in a specific region of interest; another user may want to know whether there will be a period of ten or more consecutive days within the coming season when nightime minimum temperature will drop below freezing point. A third user might conceivably want to know whether the two events above will both occur. More generally, the analysis in this paper is based on the notion that the user wants to know whether some condition or 'event' E, defined in terms of the meteorological variables from the coming season (or seasons), will or will not occur. This analysis has been applied to the set of PROVOST uncoupled ensemble integrations. For each of four different atmospheric global circulation models, 9-member ensembles were run over the winter season for the period 1979-93 (coinciding with the period of the ECMWF 15-year reanalysis). The integrations were run with prescribed observed sea surface temperatures.

Given an ensemble forecast, one can evaluate the fraction of occasions that E is predicted within the ensemble. This is the ensemble probability forecast of E. In this paper we outline two objective measures to evaluate the skill of the ensemble system, considered as a tool to predict the probability of E.

The first measure is the Brier score and the associated Murphy decomposition (Murphy, 1973). In some sense, the Brier score is a generalisation of familiar rms scores for deterministic forecasts. The decomposition of the Brier score allows one to assess specific attributes of a probability forecast: reliability and resolution. For example, if every member of an ensemble predicted E, then the ensemble probabilty forecast of E would be 100%. In all such situations one would expect E to actually occur. On the other hand, taking all the situations where E was predicted

by the ensemble with, say, 60% probability, then we should only expect E to actually occur on 60% of the occasions. An ensemble system where E verifies a fraction p of occasions, when E is forecast with probability p, is said to be a reliable system. On the other hand, reliability is not enough for the ensemble system to have skill. A forecast where the same probability for E is given every time, based on the climatological frequency of E, will be reliable. However, such a system will be unable to distinguish between situations where E is relatively likey to occur, and situations where E is relatively unlikely to occur. A system which can make such a distinction is said to have resolution. The Brier score combines reliability and resolution into one single measure.

The second measure is the relative operating characteristic (ROC). This gives the forecast hit rate and false alarm rate for E, made by assuming that E will occur if it is forecast with a probability that exceeds some threshold p_t (and that E will not occur if it is forecast with a probability that does not exceed p_t). A plot of hit rate versus false-alarm rate for varying p_t is known as a ROC curve, and the area under the ROC curve, A_{ROC}, is a measure of the skill of the probabilistic forecast system.

Although the Brier skill score and A_{ROC} are objective measure of skill, it is impossible on the basis of these values alone, to say whether the forecast system has a useful level of skill. Usefulness is a user-specific concept; for some users, the ability to forecast probabilities that are only marginally different from climatology may be very valuable, for others, such probabilities would be almost worthless. The hit and false-alarm rate are fundamental parameters in one simple assessment of the user-specific value of the forecast system. The analysis is based on a simple and idealised decision model (Murphy, 1977; Katz and Murphy, 1997). We imagine a user who has to decide whether or not to take some form of precautionary action, at cost C, on the basis that if E occurs, a loss L will be incurred. For example, in the event of a mild winter, over-wintering crops might be damaged by aphid growth. A farmer might consider crop spraying as a precautionary action. Similarly, a farmer might consider selling some of his livestock in advance of a forecast drought. This precautionary action would imply a cost to the farmer in terms of reduced livestock price compared with that obtained if brought to market after a normal season. Knowing only the climatological frequency of E, and the cost/loss ratio (C/L), the user can decide to either always or never take precautionary action. The forecast system can be said to have value if the user's mean expense is less than this baseline expense. The decision model analysis shows in a succinct and effective way, the enhanced value of a probability forecast system over that of a deterministic forecast system.

In this paper, two simple events were defined $E_{<0}$: seasonal-mean 850 hPa temperature anomaly is less than zero, and $E_{<-1}$: seasonal-mean 850 hPa temperature anomaly is less than -1K. These events were evaluated taking either all grid points in the extratropical northern hemisphere, or all grid points over Europe.

It was found that the probabilistic skill of the multi-model ensemble was higher than that for any of the individual-model ensembles. Comparison of 9-member multi-model ensembles with 9-member individual-member ensembles suggested that about 1/3 of the Brier skill score

improvement of the full (36-member) multi-model ensemble arose from the use of different models, indicating that the majority of improvement arose from the larger ensemble size of the full 36-member multi-model ensemble. From a practical point of view, this suggests that optimal seasonal forecast performance could be obtained by a probabilistic synthesis of available operational forecasts, but that individual operational centres should strive to produce ensembles as large as practicable.

Skill and value over Europe was found to be small compared with similar measures over the whole extratropical Northern Hemisphere. For example, for $E_{<-1}$, the maximum value over the whole hemisphere was close to 50% of that associated with a perfect deterministic forecast. Over Europe, the maximum was close to 20%. However, this number is not negligible; if a major European company could double its profits given a hypothetical perfect deterministic seasonal forecast for Europe, then the results here suggest that it could increase them by up to 20% with a realistic ensemble forecast.

The PROVOST integrations were shown to suffer from considerable systematic error. In order to assess a realistic level of skill achievable in the situation where systematic error could be minimised, but where the intrinsic chaotic nature of the atmosphere is represented, the skill and value of a perfect-model ensemble were estimated. Essentially, these values correspond to the situation where one member of the ensemble is chosen, at random, to represent the verification. This approach could be justified on the basis that, if medium-range forecast experience is anything to go by, systematic error will be reduced significantly in the coming years. The level of skill and value estimates from the perfect- model ensembles were considerably higher than those obtained by comparing against the ECMWF reanalysis verification, both for the NH and European grid points.

Finally, it was found, that, over the extratropical NH, skill and value were enhanced by considering only those years where a moderate or strong ENSO event was in progress. Over Europe, the results were mixed. Brier skill scores were poorer during El Niño years, whilst A_{ROC} was higher for some models, lower for others. By considering the value diagnostic, some rationalisation of these mixed results could be made; by restricting to moderate or strong El Niño years, value over Europe was increased for some users, and decreased for other users.

The analysis performed in this paper has been far from comprehensive. We have considered only two rudimentary events, and have evaluated them over two rather extensive regions. As mentioned above, specific users may require more complex events involving different, or indeed multiple meteorological variables, and the user's domain of interest may well be much smaller than that considered. However, for all such events and domains the techniques outlined in this paper can be applied. In other words, given a user-defined event E and domain D, a reliability diagram with Brier skill score (and associated decomposition), together with a ROC diagram and value graph can be produced over D. Of couse, if the user requirement is too specific, then this will be reflected in a very noisy reliability diagram with low skill and value.

We conclude by suggesting that the tools developed in this paper be used routinely as part of an

assessment of operational seasonal forecasts, and note that this assessment might be performed most effectively through interaction between user and forecaster.

Acknowledgements

We acknowledge the help of Jean-Yves Canneill (Eléctricité de France), Michel Déqué (Météo-France) and Mike Harrison (United Kingdom Meteorological Office), for assistance in planning these ensemble experiments, and making data from their models available. We also acknowledge the work of Andreas Lanzinger (Zentralanstalt für Meteorologie und Geodynamik, Salzburg) and Francois Lalaurette (ECMWF), who developed much of the code for the reliability diagram computations, as part of the verification software for the ECMWF medium-range Ensemble Prediction System. Discussions with Mike Harrison and Olivier Talagrand are also gratefully acknowledged. This work was supported by the European Union Environment and Climate Programme under contract CT95-0109.

References

Brankovic, C. and T.N.Palmer, 1998: Estimates of seasonal predictability from ECMWF PROVOST ensemble integrations. Submitted.

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

Buizza, R., T. Petroliagis, T.N. Palmer, J. Barkmeijer, M.Hamrud, A.Hollingsworth, A.Simmons and N. Wedi., 1998: The Impact of Model Resolution and Ensemble Size on the Performance of an Ensemble Prediction System. Q.J.Roy.Meteorol.Soc., 124, 1935- 1960.

Gibson, J.K., P, Kallberg, S.Uppala, A.Hernandez, A.Nomura and E.Serano, 1997:ERA description. ECMWF Re-analysis project report series. ECMWF, Shinfield Park, Reading, RG2 9AX, UK.

Katz, R. W., and Murphy, A. H., 1997. Forecast value: prototype decision-making models. In Economic value of weather and climate forecasts, Katz, R. W., and Murphy, A. H., Eds.. Cambridge University Press, 222 pp

Molteni, F., R. Buizza and T.N. Palmer, 1996: The ECMWF ensemble prediction system: methodology and validation. Q.J.R.Met.Soc., 122, 73-119.

Murphy, A.H., 1973: A new vector partition of the probability score. J.Appl. Meteor., 12, 595-600.

Murphy, A., H., 1977. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. Mon. Wea. Rev., 105, 803-816.

Palmer, T.N. and D.L.T. Anderson, 1994: The prospects for seasonal forecasting. Q.J.R.Meteorol.Soc., 120, 755-793.

Palmer, T.N., R. Buizza and F. Lalaurette, 1998: Performance of the ECMWF Ensemble Prediction System. Proceedings of ECMWF Workshop on Predictability. ECMWF, Shinfield Park, Reading, RG2 9AX, UK. 199pp.

Richardson, D., 1998: Skill and economic value of the ECMWF ensemble prediction system. To be sumbitted.

Stanski, H.R. L.J.Wilson, W.R.Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Technical Report No. 8 (WMO/TD No. 358). World Meteorological Organization. Geneva. 114pp.

Talagrand, O., R.Vautard and B. Strauss, 1998: Evaluation of probabilistic prediction systems. Proceedings of ECMWF Workshop on Predictability. ECMWF, Shinfield Park, Reading, RG2 9AX, UK. 199pp.

Wilks, D.S., 1995: Statistical Methods in the Atmospheric Sciences. Academic Press. 467pp