TECHNICAL MEMORANDUM

262

Skill and relative value of the ECMWF Ensemble Prediction System

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

Skill and relative economic value of the ECMWF Ensemble Prediction System

By **D. S. Richardson**

European Centre for Medium-Range Weather Forecasts

Submitted to *Q. J. R. Meteorol. Soc.*, October 1998.

Summary

The economic value of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational Ensemble Prediction System (EPS) is assessed relative to the value of a perfect deterministic forecast. The EPS has substantial relative value throughout the medium range. Probability forecasts derived from the EPS are of greater benefit than a deterministic forecast produced by the same model. In fact, for many users the probability forecasts have more value than a shorter-range deterministic forecast. Based on the measures used here, the additional information in the EPS (reflecting the uncertainty in the initial conditions) provides a benefit to users equivalent to many years development of the forecast model and assimilation system.

The impact of ensemble size on forecast value is considered. The difference in performance between 10 and 50 members may appear relatively small, based on standard skill measures, yet the larger ensembles have substantial benefit to a range of users. Further increases in ensemble size will provide additional value.

1. Introduction

An ensemble prediction system (EPS) has been in regular production at the European Centre for Medium-Range Weather Forecasts (ECMWF) since December 1992 (Palmer et al., 1993; Molteni et al., 1996). The current operational EPS comprises a control forecast initialised from the operational analysis, plus 50 additional integrations initialised from perturbations to the control analysis (Buizza et al., 1998). All forecasts are made with the operational ECMWF model, but run at lower horizontal resolution (T_L159) than the single high resolution deterministic forecast (T_L319). The EPS complements the deterministic forecast by the provision of information about the probability distribution of future weather, based on uncertainty in the initial analysis.

The performance of the EPS is routinely monitored using a range of verification measures. These assessments demonstrate that the EPS is a skilful prediction system. They also demonstrate the improvement of the enhanced EPS introduced in December 1996 (Buizza et al., 1998). However, these measures do not explicitly address the question which is perhaps of most concern to potential users, namely "Is the EPS worth paying for?".

Providing an answer to such a question is not straightforward. To obtain economic benefit from a forecast, a potential user must have alternative courses of action available, the consequences of which will depend on the weather that occurs. If, by using forecasts, the user decides on actions which he would not otherwise take, and benefits economically from these alternative actions, then the forecasts have been of value to the user (Murphy, 1994). Thus, a proper evaluation of the benefits of a forecast system to a particular user will involve not only the intrinsic skill of the forecasts, but also detailed knowledge of the exact weather-sensitivity and decision-making process of the user.

A number of studies have investigated the relationship between skill and economic value of weather forecasts, either using specific examples (Roebber and Bosart, 1996), or in general terms (Murphy and Ehrendorfer, 1987). Katz and Murphy (1997a) provide a comprehensive account of research in this area, and an extensive collection of references. As these studies indicate, the relationship between skill and value is complex, yet the potential value of ensemble forecasts has not so far been considered. Reliance on skill measures alone may give a misleading impression of forecast value. Decisions on system configuration (for example, choice of optimal ensemble size), may be different if the effect on economic value is considered rather than just the overall skill. There are thus two reasons to study the economic value of the EPS: to evaluate the potential benefit to users and to assess the impact of possible system changes in a context relevant to users.

In this paper we use a simple decision-analytic model to examine the economic value of the EPS relative to that of a (hypothetical) perfect deterministic forecast. As examples of the operational EPS products we consider predicted probabilities of 850 hPa temperature anomaly exceeding certain thresholds over Europe. The decision model (introduced in Section 2) is applied to the deterministic control forecast (Section 3) and to the EPS probability forecasts (Section 4) for these predictions for January and February 1998. The benefit of the EPS over the deterministic forecast is assessed in Section 5. In Section 6, the effect of ensemble size on forecast value is studied using EPS precipitation forecast data for winter 1996-97. Conclusions are drawn in Section 7.

The decision-analytic model discussed in this paper has also been applied to seasonal forecast ensembles. Results are discussed in Palmer et al., 1998.

2. The cost-loss ratio decision model

The economic value of weather forecasts is often discussed in terms of so-called decision-analytic models (Murphy, 1977; Katz and Murphy, 1997b). A decision maker has a number of alternative courses of action available and the choice of action is to some extent influenced by the forecast. Each action has an associated cost and leads to an economic benefit or loss depending on the actual weather that occurs. The task of the decision maker is to minimise the expected loss by choosing the appropriate action. To emphasize the main points of the decision framework, we consider here the simplest decision model, known as the static cost-loss model.

Consider a decision maker who has just two alternatives, to take action or to do nothing, the choice depending exclusively on his belief that a given weather event E will occur or not. Taking action incurs a cost C irrespective of the outcome. If the event does occur and no action has been taken then the decision maker incurs a loss L. For example, the weather event could be the occurrence of ice on roads and the action "to grit the roads"; C would be the cost of the gritting procedure while L would be the economic loss due to traffic delays and accidents on icy roads. The expense associated with each combination of action and occurrence of E is shown in Table 1 (the expense matrix).

Table 1. Cost and loss for different outcomes

		Occurs	
		No	Yes
take action	No	0	L
	Yes	C	C

The decision maker wishes to pursue the strategy which will minimise his expense over a large number of cases. If only climatological information is available there are just two options: either always take protective action or never protect. Always taking action incurs a cost C on each occasion (irrespective of whether the event occurs or not), while if action is never taken the loss L occurs only on that proportion $\bar{o}$ of occasions when the event occurs, hence the average expense is $\bar{o}L$. Thus in the absence of information other than climatology, the optimal course of action is always act if $C < \bar{o}L$ and never act otherwise.

It is convenient to consider the expense of the various courses of action in terms of the "cost-loss" ratio C/L. Given only climatological information, the expected mean expense per unit loss (ME) is then

$$ME(\text{climate}) = min\left(\frac{C}{L}, \bar{o}\right). \tag{1}$$

If the cost of protection is greater than the potential loss there is no benefit to be obtained from taking any protective action. Thus C/L can effectively be considered to be in the range 0 to 1. ME can be plotted as a function of C/L on an expense diagram (Fig. 1). The minimum ME given climate information (Eq. (1)) is shown by the solid curve. The dashed curve shows the minimum expense which could be obtained given perfect knowledge of the future weather - the decision maker would only need to take action when the event was going to occur and would never incur a loss, so the expected expense would be

$$ME(\text{perfect}) = \bar{o}\frac{C}{L}. \tag{2}$$

Of course, given the chaotic nature of the atmosphere and our inevitably uncertain knowledge of the exact initial state of the atmosphere, such perfect forecasts are not likely to be achieved in practice. However, there is clearly potential for reduction of the ME from that of climatological information towards the perfect-forecast limit.

The provision of additional information in the form of forecasts may allow the decision-maker to revise his strategy and reduce his expected expense. The extent by which the expense is reduced is a measure of the value of the forecasts to the decision maker.

We define the relative value V of a forecast system as the reduction in ME as a proportion of that which would be achieved by a perfect forecast:

$$V = \frac{ME(\text{climate}) - ME(\text{forecast})}{ME(\text{climate}) - ME(\text{perfect})}. \tag{3}$$

Thus maximum relative value V=1 will be obtained from a perfect forecast system, while V=0 for a climate forecast. If V>0 then the user will benefit from the system. This definition gives an absolute upper bound to V and is a convenient reference level for the user: if a perfect knowledge of the future weather will save the user an amount S (over the use of purely climatological information) then the EPS with relative value V will save the user 100V% of S.

3. Skill and relative value for a deterministic forecast system

Consider first a deterministic forecast system, that is each forecast is a simple statement either that a weather event E will occur or that it will not occur. The performance of the system over a period of time can be summarised in a contingency table which shows the proportion of correct and incorrect forecasts of a weather event occurring or not occurring (Table 2).

TABLE 2. CONTINGENCY TABLE FOR FORECAST AND OCCURRENCE OF BINARY EVENT

		observed		
		No	Yes	
forecast	no	a	b	a+b
	yes	c	d	c+d
		a+c	b+d	

The hit rate (HR) is defined as the proportion of occurrences of the event which were correctly forecast, while the false alarm rate (FAR) is the proportion of non-occurrences for which the event was (incorrectly) forecast. Note that both HR and FAR are expressed in terms of the observed relative frequency of the event $\bar{o}$; it is assumed that $\bar{o} > 0$, i.e. that the event does occur in the sample.

$$HR = \frac{d}{b+d} = \frac{d}{\bar{o}} \tag{4}$$

$$FAR = \frac{c}{a+c} = \frac{c}{(1-\bar{o})} \tag{5}$$

From Table 2 and the expense matrix (Table 1), the expected mean expense (ME) for the forecast system is:

$$ME = \frac{bL+(c+d)C}{L} = b+(c+d)\frac{C}{L} \tag{6}$$

This can be written in terms of HR and FAR using Eqs. (4) and (5) as

$$ME = \mathrm{FAR}\frac{C}{L}(1-\bar{o}) - \mathrm{HR}\bar{o}\left(1-\frac{C}{L}\right)+\bar{o} \tag{7}$$

Substituting from Eqs. (1), (2), and (7) into Eq. (3), the relative value of the forecast system is:

$$V=\frac{min\left(\frac{C}{L},\bar{o}\right)-\mathrm{FAR}\frac{C}{L}(1-\bar{o})+HR\bar{o}\left(1-\frac{C}{L}\right)-\bar{o}}{min\left(\frac{C}{L},\bar{o}\right)-\bar{o}\frac{C}{L}} \tag{8}$$

So the relative value of a particular forecast system depends on the external (to the system) parameters C/L and, and the model-dependent parameters HR and FAR.

Various measures of forecast skill may be derived from Table 2 (Wilks, 1995; Stanski et al., 1989). Here we use the Kuipers score (KS) which has the desirable characteristic of equitability (Gandin and Murphy, 1992), in that random or constant forecasts will score zero and perfect forecasts will have a score of 1. In the notation of Table 2, this can be written as

$$\mathrm{KS} = \frac{ad-bc}{(a+c)(b+d)} = HR-FAR \tag{9}$$

To illustrate the performance of a deterministic forecast system, we consider the control integration (the forecast from the unperturbed analysis). Results for the prediction of day 6 850 hPa temperature anomalies exceeding certain thresholds are shown in Table 3 for January and February 1998 over Europe. The control forecast has substantial skill for all thresholds.

TABLE 3. SKILL OF THE EPS CONTROL FORECASTS OF TEMPERATURE OVER EUROPE

event	Jan-Feb 98			
	FAR	HR	KS	$\bar{o}$
T < -8K	0.039	0.445	0.406	0.058
T < -4K	0.144	0.611	0.468	0.228
T >+4K	0.091	0.548	0.457	0.179
T >+8K	0.027	0.393	0.367	0.043

Forecasts for the smaller thresholds are more skilful than forecasts for the more extreme events, and positive anomalies appear more difficult to predict than negative anomalies. However, the question for potential users is: "How does this skill translate to economic value of a forecast?".

For a given weather event and forecast system, $\bar{o}$, HR and FAR are given and the relative economic value V of the forecast system depends only on the cost-loss ratio. V is shown in Fig. 2 as a function of C/L for the forecasts of the four events. Although the model is skilful according to the scores in Table 3, it is clear that the usefulness to a decision maker depends greatly on his particular cost-loss ratio. For C/L greater than about 0.6 none of the event forecasts are useful; for C/L

between 0.1 and 0.5 forecasts of the ±$4K$ events are useful, while for C/L less than 0.1 it is only the forecasts of larger anomalies which have value.

Maximum value always occurs for $C/L = \bar{o}$; at this point the expense of taking either climatological option (always or never protect) is the same: climatology does not help the decision maker and the forecast has the greatest benefit. As the cost approaches the limits of 0 and 1, the climatological options become harder to beat - high expense resulting from even occasional incorrect forecasts outweighs the low expenditure of the default action.

The maximum value itself is given by

$$V_{\max} = \mathrm{HR} - \mathrm{FAR} \tag{10}$$

Thus skill (Eq. (8)) is related to the usefulness of the forecasts: KS is the maximum relative value that can be obtained from the system. Whether this potential maximum value will be achieved depends on the cost-loss ratio of the user; the closer C/L is to $\bar{o}$ the higher will be the value. Note that this maximum value is independent of C/L and $\bar{o}$; if two systems predicting different events (with quite different $\bar{o}$) have the same KS then the potential maximum value will be the same, but it will occur for different values of C/L (equal to the respective observed frequencies).

4. Probability forecasts

If forecasts are supplied to the decision maker as probabilities then the question facing the user is at what probability threshold should action be taken. Should the user take action if the event is forecast with a probability of, say, 50% or should he wait until the forecast is more certain (perhaps 80%)? Is there an optimum probability above which action should be taken?

In effect, this choice of a threshold probability p* converts the probability forecast to a deterministic one - consider those forecasts with higher probability for the event as forecasts that the event will occur and those with lower probability as forecasts that the event will not occur. For a given p*, the value of the system can then be determined in the same way as for a deterministic system. By varying p* from 0 to 1 a sequence of values for HR and FAR and hence of V can be derived; the user can then choose that value of p* which results in the largest V. Note that since V also depends on $\bar{o}$ and C/L, the appropriate value of p* will be different for different users and different weather events.

Probability forecasts of the temperature events considered in the previous section are produced using the EPS. The relative operating characteristic (ROC; Mason, 1982; Harvey et al., 1992) is a plot of HR against FAR for a set of threshold probabilities p* between 0 and 1 (figure 3). The endpoints of the ROC (1,1 and 0,0) result from the baseline actions of always forecasting or never forecasting the event respectively. A perfect forecast system, with HR=1 and FAR=0, would give a point at the top left corner of the graph, so the closer the ROC is to the top left corner the better. If a forecast system had no ability to discriminate occurrence of an event from non-occurrence, then HR and FAR will always be equal and the ROC for the system would lie along the diagonal line HR=FAR. The area (A) under the ROC is used as an index of the accuracy of the forecast system (Mason, 1982; Buizza et al., 1998; Buizza and Hollingsworth, 1998). A perfect system would have A=1.0, while the no-skill system (HR=FAR) would have A=0.5. The areas under the ROCs of Fig. 3 are shown in the legend. The area can be converted to a skill score relative to climatology and chance in the usual way (Stanski et al., 1989) as

$$ASS = \frac{A(\text{forecast}) - A(\text{climate})}{A(\text{perfect}) - A(\text{climate})} = 2A - 1\,. \tag{11}$$

For each probability threshold p*, the corresponding HR and FAR can be used to generate a value curve, just as in the deterministic case. The set of curves for the T>+4K event are shown in Fig. 4. The EPS forecasts have value for most users, although the benefit varies substantially for users with different cost-loss ratios. The most important feature of Fig. 4 is that the relative value depends crucially on the appropriate choice of threshold probability p*. Users with small cost-loss ratios, i.e. relatively large potential losses, will benefit by taking action even when the forecast probability is low, while for users with high cost, value is obtained by taking action only if there is high forecast probability for the event. An inappropriate choice of p* can result in substantial reduction in forecast value. For example, a decision maker with a cost-loss ratio of 0.1 will receive over 40% relative value by acting when the EPS probability is 10% or more, but would gain no value at all from the EPS if no action is taken until the forecast probability is greater than 50%.

This example illustrates the importance of providing probability information to users: the value of the EPS forecasts depends significantly on the choice of probability threshold p* and on the user's cost-loss ratio. There is no single threshold for which the EPS has value for all users - different users must use different thresholds to benefit from the forecasts. If the EPS forecast is reduced to a single deterministic forecast for all users, for instance by using the ensemble mean or by choosing an arbitrary threshold, the value to some users will be reduced and may even be eliminated completely.

A probability forecast system is said to be reliable if the event occurs on a proportion p of those occasions where the forecast probability is p, i.e. the forecasts may be taken at face value (Stanski et al., 1989; Murphy, 1993). For a reliable forecast system, maximum value for a particular user will be obtained by choosing the threshold probability p* $= C/L$. More generally, an appropriate p* can be chosen to compensate for the forecast bias, e.g. from Fig. 4. While reliability may be an intrinsically desirable objective, the value to a user is unaffected, so long as he is aware of any unreliability and chooses p* accordingly.

To provide maximum value to users with small cost-loss ratios, the EPS will need to resolve equally small probability thresholds: an ensemble of at least N $= L/C$ members will be needed to realise the maximum potential benefit. In practice, much larger ensembles may be needed to provide reliable probability estimates. The variation of value with p*, and hence ensemble size is discussed in more detail in Section 6.

Maximum value over all C/L occurs at $C/L = \bar{o}$, as for the deterministic case, and is given by

$$V_{\max} = Max_{p*}(HR - FAR) = Max_{p*}(KS) \equiv KS_{max} \tag{12}$$

Thus for reliable forecasts, maximum relative value is $V_{\max} = KS(p^* = \bar{o})$.

We have defined V as value relative to the value of a perfect deterministic forecast. In practice, the inherent uncertainty in our estimates of the current state of the atmosphere mean that perfect deterministic forecasts, and hence perfect relative value (V=1), are unattainable. An estimate of the potential relative value of the EPS, given current analysis uncertainty, can be made by considering the ensemble as a "perfect EPS", eliminating model errors. The "perfect EPS" can be evaluated by using one ensemble member as the verification rather than the actual analysis (Buizza and Palmer, 1998). Alternatively, for thc point probability forecasts studied here, scores for the perfect EPS may be derived directly from the forecast probabilities. Let the function g(p) denote the frequency with which the EPS forecasts a given event with probability p. Let $p'(p)$ be the frequency with which the event occurs when the forecast probability is p. The performance of the EPS in forecasting the event is completely specified by these two functions. For example, hit rate can be written as

$$\mathrm{HR(p^*)} = \left(\left[\int_{p^*}^{1} p'(p)g(p)dp\right] / \left[\int_{0}^{1} p'(p)g(p)dp\right]\right) \tag{13}$$

For a "perfect EPS", $p'(p) = p$ for all p, and the skill and value of the ensemble depends only on the distribution of forecast probabilities g(p). This method is statistically more robust than evaluating against a randomly chosen ensemble member (for large enough ensemble and sample size the two methods will give the same result).

The relative value of the perfect EPS (V_{perf}) is shown in Fig. 4. The relative value of the actual EPS forecasts is about 0.1 below that of the perfect EPS for almost all C/L, with greater potential gains for small C/L. While the difference between the relative value V of the EPS and the "perfect EPS" relative value V_{perf} gives an indication of the potential for improvement by elimination of model errors, the difference between V_{perf} and the perfect *deterministic* limit V=1 indicates the importance of initial condition uncertainty in limiting the value which can be expected in practice. In the perfect EPS context, V_{perf} can only be improved by reducing the initial uncertainty, i.e. by reducing analysis errors. Future developments in data assimilation are designed to lead to better analyses and lower initial uncertainties - this will lead to higher relative value.

5. Comparison of deterministic and probabilistic forecasts

One of the benefits of the ROC is that it allows direct comparison of deterministic and probabilistic forecast systems. The HR and FAR for the control forecast (table 2) are plotted together with the EPS probability ROCs in figure 3. For this event the points for the control lie below the EPS ROC. The control forecasts are less useful than the EPS forecasts for this event, since for the same FAR a higher hit rate is obtained using the EPS probabilities. This will be true fore all users irrespective of $\bar{o}$ and C/L - if $\bar{o}$, C/L and FAR are fixed, V increases with HR (Eq. (8)).

Comparison of the value curves for the EPS probability forecasts and the control deterministic forecast highlights the advantage of the probability forecasts (Fig. 5). The flexibility of being able to choose the threshold probability greatly increases the range of users who will benefit from the EPS forecasts. Even though the deterministic forecasts appear close to the EPS curves on the ROCs, the extra value of the probability forecasts can be substantial. The relative value of a "perfect EPS" is also shown; the improvement over the actual EPS is similar for each event, increasing V by 0.1-0.2 for most users.

The three points (0,0) (HR_{control}, FAR_{control}) and (1,1) may be considered to define a ROC curve for the deterministic forecast. The accuracy of this forecast can then be measured in the same way as for the probability forecast by the area A under the curve (i.e. the area of the quadrilateral defined by these points and (1,0)). It is straightforward to show that for the deterministic forecast the area skill ASS (Eq. (11)) is equal to the standard KS skill score:

$$KS = 2A - 1 \equiv ASS. \tag{14}$$

The skill of the deterministic control and of the EPS probability forecasts of the +4K anomaly are compared for forecast days 3 to 10 (D3 to D10) in Fig. 6. For the EPS, ASS is substantially higher than KS_{max}, giving the impression of greater skill and a larger improvement over the control forecast. A deterministic forecast with skill equal to KS_{max} will still have less value than the probability forecasts for a range of C/L, because of the flexibility of varying p* for different users. The difference between ASS and KS_{max} for the ensemble reflects this additional benefit, although ASS is difficult to interpret quantitatively. In the following we use KS_{max} rather than ASS as an index of EPS skill because of its simple interpretation as the maximum value of the ensemble skill. Use of ASS instead would show a greater skill advantage for the EPS in the following analysis.

Comparison of the KS curves for EPS and control show a substantial advantage of the probability forecasts at all forecast lead times. In fact the D10 probability forecast is as skilful (has the same maximum value) as the D6 control. This gives a measure of the benefit of the EPS - for the single deterministic forecast to have the same skill as the current EPS at day 10, the forecast model would need to be improved so that deterministic D10 forecasts are as skilful as the present D6 forecasts. This skill advantage is shown for all temperature events in Fig. 7 (missing points indicate where the EPS is more skilful than the D3 control; scores were not available for D1 and D2 forecasts). For most lead times and events studied here, the EPS has an advantage (D_{adv}) of 2-4 days over the control. This is a considerable difference which highlights the importance of the EPS - improvements in deterministic skill of such magnitude are generally achieved over many years of development of all aspects of model and data assimilation formulation (Simmons et al., 1989).

The analysis of the previous paragraph demonstrated the advantage of the EPS in terms of skill or, equivalently, maximum relative value. For different users, the relative value of EPS and the control forecasts varies substantially. The change in value with lead time is shown for all users in Fig. 8 for the event T<-4K, the situation where the EPS has the smallest skill advantage (Fig. 7). Although D_{adv} is less than one day at D6, the D5 control forecast is only beneficial for a limited range of C/L. One forecast system is said to be sufficient for another if it provides greater value for all users (Ehrendorfer and Murphy, 1988). It is clear from Fig. 8 that even the D3 control forecast is not sufficient for the D6 EPS. The sufficiency of the D3 control for the D10 EPS is considered in Table 4 for the four temperature thresholds. For each event the D3 control skill (maximum value) is, as expected. substantially greater than the D10 EPS skill. However for every event there are some users for whom the D10 EPS has considerably more value than the D3 control. The day 3 control forecast is not sufficient for the D10 EPS for any of the temperature events.

TABLE 4. SKILL AND MAXIMUM INCREASE IN VALUE OF EPS D10 OVER CONTROL D3 FORECASTS OF TEMPERATURE OVER EUROPE

event	Jan-Feb 98		
	KS(D3 control)	KS_{max}(D10 EPS)	Max(V(D10 EPS)-V(D3 control))
T < -8K	0.688	0.404	0.194
T < -4K	0.728	0.343	0.070
T >+4K	0.708	0.454	0.110
T >+8K	0.648	0.514	0.328

The greatest benefit of the EPS over the shorter range control forecast is for users with low C/L (Fig. 8). This has particular significance when the potential cost savings of more timely warning are taken into account. If a user has more time to prepare for protective action he may be able to reduce the cost of taking action, reducing C/L towards the region where the EPS has the greatest advantage.

6. THE EFFECT OF ENSEMBLE SIZE ON FORECAST VALUE

It was shown in Section 4 that for a given user, maximum relative value is gained when the threshold probability is equal to the user's cost-loss ratio. It is therefore important to consider the variation of value with threshold probability. If small changes in p* lead to significant differences in value then the EPS resolution will need to be sufficient to resolve the required probability thresholds.

The relationship between ensemble size, threshold probabilities and value is illustrated using EPS forecasts of precipitation over Europe in the 12 hour period T+108 to T+120 (day 4.5 to day 5) exceeding a given threshold. Three events are considered, representative of different climatological frequency: total precipitation exceeding 1mm, 5mm, and 10mm. The EPS forecasts are verified against the T+12 to T+24 accumulated precipitation forecasts from the operational deterministic high resolution (T213) ECMWF model; results are for 92 cases from winter 1996-97 (Buizza and Hollingsworth, 1998).

In addition to ensemble size, an important factor in the determination of an optimal ensemble configuration is the resolution of the forecast model. For a given allocation of computing power, a compromise must be reached between ensemble size and model resolution (Buizza et al., 1998). Precipitation forecasts are perhaps particularly sensitive to resolution. For this reason, the operational high resolution (T213) deterministic forecast is assessed in addition to the EPS control forecast to give an indication of the effect of the reduced EPS resolution.

ROCs for the three events are shown in Fig. 9; the area under each curve and the maximum relative value of the system are given in Table 5. The EPS has skill for all events and the forecasts all have substantial relative value. For each precipitation amount, the ROC points for both deterministic forecasts lie below the corresponding EPS probability ROC and the maximum value (or skill, KS) is substantially lower than that of the EPS. The skill of the high resolution operational forecast is, however, greater than that of the lower resolution EPS control, indicating the potential for improvement to the EPS from using a higher resolution forecast model. While the overall value of the EPS is greater than

TABLE 5. EPS PERFORMANCE FOR PRECIPITATION EVENTS

precipitation threshold	$\bar{o}$	ROC area	V_{max}	KS(control)	KS(operational)
1 mm	0.289	0.826	0.50	0.338	0.397
5 mm	0.0552	0.831	0.56	0.218	0.239
10 mm	0.0135	0.794	0.57	0.140	0.190

the single high resolution forecast, there are circumstances where the lower resolution EPS fails to capture an important extreme event, whereas an integration run at enhanced resolution provides a successful forecast (A. Hollingsworth, personal communication). In such circumstances, while a single high resolution forecast will not always be correct, it can provide valuable additional information to forecasters on the potential effect of resolution.

For the EPS, V_{max} is higher for the 5mm and 10 mm events than for the 1mm threshold, but skill (as determined by ROC area) is lower for the 10mm event than for the other precipitation thresholds. It is apparent from Fig. 9 that while points on the ROC are reasonably evenly spaced for precipitation over 1mm, data for the ROCs for 5mm and 10mm are increasingly restricted towards the lower left portion of the graph. If attention is focused on just those parts of the ROC curves where data is available for all three precipitation amounts, it is clear that the EPS performance is better for the higher precipitation events (for a given false alarm rate, the highest hit rate is achieved for the 10 mm event, lowest hit rate for the 1mm event).

The reason for the restricted coverage of the larger event ROCs is related to the lower observed frequency of the higher rainfall events (Table 5) and the threshold probabilities used to calculate the HR and FAR. If the observed frequency of the event is small relative to the forecast probability threshold p*, and the forecast system is reasonably reliable (i.e. the rare events are forecast rarely), the great majority of points in the contingency table (Table 2) for this p* will be in the No/No cell ("a" in Table 2); hence the FAR will be low and the HR will be limited accordingly. To extend the ROC data points towards the top right of the diagram, additional lower probability thresholds must be considered. In practice, the resolution of probability thresholds is limited by the size of the ensemble. For an N-member ensemble, all possible hit and false alarm rates can be calculated by taking the threshold probabilities at intervals $dp^* = 1/N$. The ROCs in Fig. 9 were calculated using all threshold probabilities available from the 51-member EPS, i.e. p* intervals of approximately 0.02. While this resolution is sufficient for the relatively frequent 1mm precipitation event, a larger ensemble would be needed to specify the full ROC for the rarer 10mm event.

The ROC is often generated using a fixed set of probability thresholds at 10% intervals (Stanski et al., 1989; Buizza and Palmer, 1998). This use of a limited set of probability thresholds may give a misleading impression of the potential performance of the forecast system. The effect is illustrated in Fig. 10 and Table 6. in which the ROCs and scores for each precipitation event have been recalculated using dp*=0.1, rather than the full set of probability thresholds used for Fig. 9. There is little effect on the ROC for the relatively frequent 1mm event; the area decreases slightly but V_{max} is unchanged. However, the scores for the larger precipitation amounts are substantially degraded and give the misleading impression that the 10mm event is considerably less skilful than the 1mm event.

TABLE 6. EPS PERFORMANCE FOR PRECIPITATION EVENTS (DP*=0.1)

precipitation threshold	ROC area	V_{max}
1 mm	0.819	0.50
5 mm	0.764	0.51
10 mm	0.656	0.31

The variation of relative value with C/L is shown for the two sets of probability thresholds in Fig. 11. For all precipitation events there is little effect from changing dp* for $C/L > 0.1$. However, large differences are apparent for smaller C/L; The curves are plotted with a logarithmic x-axis so that this can be seen clearly. Although $V_{\max}$ does not increase for dp*=0.02 for the 1mm event there are still substantial gains for users with low cost-loss ratios, use of the smaller dp* giving values of up to 20% for C/L where the dp*=0.1 probability thresholds gave no value. The differences in value are notably larger for the more extreme events. Also shown in Fig. 11 are estimates of the relative value which could be obtained with a sufficiently large ensemble. These estimates were derived using a parametrised model of the ROCs of Fig. 9 (Mason, 1982; Harvey et al., 1992); the method is summarised briefly in the appendix. There is potential for substantial improvement for users with low cost-loss ratios, particularly for the more extreme events (as $\bar{o}$ decreases, so too does the probability threshold p* for maximum value).

For cost-loss ratios larger than about 0.1, there seems to be little benefit to be obtained solely from increasing the resolution of the probability threshold. However, increasing ensemble size may be of value in providing more reliable estimates of forecast probability. The effect of ensemble size is examined by using just 10 of the 51 ensemble members to calculate HR and FARs for the original p* thresholds (dp*=0.1). ROC area and maximum value are both reduced (Table 7, c.f. Table 5), while relative value is lower for all users than for 50 members (figure 12).

TABLE 7. EPS PERFORMANCE FOR PRECIPITATION EVENTS 10 MEMBERS

precipitation threshold	ROC area	$V_{\max}$
1 mm	0.800	0.47
5 mm	0.711	0.40
10 mm	0.638	0.28

This result indicates that increasing ensemble size will benefit both by allowing finer resolution of probability thresholds and by giving improved estimates of the forecast probability distribution. The potential value estimates of Fig. 11 which only consider the effect of increasing resolution of dp* should therefore be considered as lower bounds to the improvement obtainable with larger ensembles.

7. CONCLUSIONS

The EPS is an important component of the operational forecasting capability at ECMWF, complementing the operational deterministic forecast with probabilistic information reflecting the uncertainty in the analysed atmospheric state. A simple decision-analytic model has been used to study the potential economic value of EPS forecasts of temperature and precipitation.

The EPS has considerable value throughout the medium range, although different users will benefit to a greater or lesser extent, depending on their specific economic costs. Comparison of the EPS probability forecasts with deterministic predictions from the control integration demonstrates the advantage of the probabilistic approach. Probability forecasts are generally more useful than deterministic forecasts of comparable quality because of the facility for the user to select a probability threshold appropriate to his needs. A forecaster's arbitrary prescription of such a threshold without knowledge of a particular user's requirements can severely reduce the value of the system to that user.

The value of the operational EPS was compared to that of a "perfect EPS" in which model errors are eliminated. The potential improvement in relative value is 0.1-0.2 for most users for the four temperature events considered. The substantial difference between the relative value of the perfect EPS and the perfect deterministic limit (V=1) is an indication of the effect of the initial uncertainty in limiting the value which can be expected in practice. If the initial uncertainty can be reduced by improvements to the forecast model and data assimilation system, this will lead to improvements in the potential value of the EPS.

In our point verification, the EPS probability forecasts have several days advantage over the control forecast. In other words, the additional information in the EPS, reflecting the uncertainty in the initial conditions, provides a benefit to users equivalent to many years development of the deterministic forecast model and assimilation system. In fact, for some users, EPS probability forecasts for 10 days ahead have more value than day 3 deterministic forecasts.

The analysis presented in this paper demonstrates the overall benefit of the EPS probability forecasts over a single deterministic forecast. For precipitation forecasts, which may be particularly sensitive to model resolution, the EPS was shown to be superior to the higher resolution T213 deterministic forecast. However, the improvement of the T213 forecast over the T_L159 EPS control indicates the potential benefit of increased resolution. Despite the overall benefit of the EPS, it was noted that there are circumstances where the EPS resolution is not great enough to capture an important event. While a single high resolution forecast will not always be correct, it can provide valuable additional information on the sensitivity of a particular situation to model resolution; it is for this reason that a high resolution deterministic forecast is run operationally at ECMWF alongside the lower resolution EPS.

Maximum relative value occurs when the cost-loss ratio, C/L, is equal to the observed frequency of the event. For a particular user, the greatest value of a reliable probability forecast is found for threshold probability p* equal to the user's C/L. Thus users with small C/L will only receive maximum benefit if equally small probability thresholds can be resolved by the EPS. Recent studies of ensemble size (Buizza and Palmer, 1998; Buizza et al., 1998, Buizza and Hollingsworth, 1998) have followed the common practice of using a standard set of probability thresholds, at 10% intervals. While some benefit was found for all users by using all 50 ensemble members rather than just 10 to estimate these thresholds, a far greater increase in value for small C/L was found by using the full ensemble to discriminate smaller probability thresholds. Further increases in ensemble size will give significant additional improvement for low C/L, especially for more extreme events.

Ideally, the weather sensitivity, decision processes and relevant costs of all potential users would be known. It would then be straightforward to calculate the overall value of the EPS. Proposed changes to the forecast system could be evaluated in terms of net benefit, and the effect on any given user could be determined. In practice, little is known about the decision-making processes of many weather -sensitive activities (users themselves are often unclear about this information). One possible approach (Roebber and Bosart, 1996) is to study overall value as a function of various hypothetical distributions for C/L, although the absolute loss, L, of different users and the relative importance of different weather events will also have a significant impact on total value.

However, until more is known about the spectrum of potential users, it is important to be aware of the effect of the EPS on the full range of C/L. As Roebber and Bosart (1996) point out, high cost-loss ratios are difficult for a business to sustain; also, competition between businesses is likely to act to reduce C/L. The provision of forecasts with longer lead time may also serve to reduce costs (see Section 5). Finally, for extreme events, potential losses may greatly exceed the cost of protection in many cases. The advantage of the EPS and potential benefits of increased ensemble size for low C/L may thus be of considerable significance.

Although it may be difficult to determine the cost and losses for a particular user, the value study presented here does present forecast verification in a form relevant to the user's needs. The decision framework, although simple, emphasises the fact that users will not all feel the same benefits of the EPS, nor will they be affected equally by changes to the forecasting system.

Acknowledgements

I thank Tim Palmer, Adrian Simmons and Tony Hollingsworth for their helpful comments on an earlier version of this paper.

APPENDIX A THE GAUSSIAN MODEL FOR THE ROC

The ROC curves presented in this paper have all been produced empirically as plots of hit rate and false alarm rate derived from forecast verification data. Mason (1982)and Harvey et al. (1992) demonstrate that a parametrised model of the ROC, derived from signal detection theory (SDT) provides a good fit to such empirical forecast data.

The model assumes that information about the occurrence or not of an event can be represented by a single one-dimensional variable X. Uncertainty about the outcome is reflected in case to case variations in X, which are given by two fixed probability distributions, one for the distribution of X given the event occurs $f_s(X)$ (the signal distribution in SDT) and another for the distribution of X given the event does not occur $f_n(X)$ (the noise distribution).

For any given value X_c (the decision criterion), the hit rate and false alarm rate are then simply the areas to the right of X_c under the respective pdfs

$$HR = P(X > X_c | \text{event occurs}) = \int_{X_c}^{\infty} f_s(x)dx \tag{A.1}$$

$$FAR = P(X > X_c | \text{event does not occur}) = \int_{X_c}^{\infty} f_n(x)dx \tag{A.2}$$

So as the decision criterion X_c varies, a sequence of HR and FAR are produced which trace out the ROC for the system. The model and the ROC are thus defined by the distributions $f_n(X)$ and $f_s(X)$. The simple case of both distributions being Gaussian is generally found to produce good results (Mason, 1982; Harvey et al., 1992).

If both distributions are Gaussian (means μ_s, μ_n; standard deviation σ_s, σ_n) then HR and FAR can be expressed as areas under the standard Gaussian distribution

$$HR = \int_{X_c}^{\infty} f_s(x)dx = \int_{z_s}^{\infty} f(z)dz \tag{A.3}$$

where

$$z_s = \frac{X_c - \mu_s}{\sigma_s} \tag{A.4}$$

with a similar expression for FAR. z_s and z_n are related by

$$z_s = \frac{\sigma_n}{\sigma_s} z_n + \frac{\mu_n - \mu_s}{\sigma_s} \tag{A.5}$$

Thus, a set of HR and FAR may be transformed to the corresponding standardized Gaussian deviates z_s and z_n (Eq. (A.3)); the strength of the linear relationship between z_s and z_n (Eq. (A.5)) is a measure of the validity of the Gaussian model for the original data. Correlation between the transformed variables is greater than 0.99 for all the events discussed in this paper (statistically significant beyond the 1% level), confirming that the Gaussian model is applicable to the EPS data.

The Gaussian model is completely specified by two parameters, usually chosen as the distance between the means of the two Gaussian distributions (normalised by the standard deviation of f_n) and the ratio of the two standard deviations.

These may be estimated from a set of empirical ROC data. The model may then be used to construct a parametrised version of the original ROC, but with any desired resolution in probability threshold. Thus the model may be used to provide an estimate of the potential benefit of having finer resolution in dp*.

REFERENCES

Buizza, R. and Hollingsworth, A., 1998. Probability precipitation prediction using the ECMWF Ensemble Prediction System. *Weather and Forecasting.*, submitted.

Buizza, R. and Palmer, T. N., 1998. Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, to appear.

Buizza, R., Petroliagis, T., Palmer, T. N., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A., and Wedi, N., 1998. Impact of model resolution and ensemble size on the performance of and ensemble prediction system. *Q. J. R. Meteorol. Soc.*. To appear.

Ehrendorfer, M., and Murphy, A. H., 1988. Comparative evaluation of weather forecasting systems: sufficiency, quality and accuracy, *Mon. Wea. Rev.*, **116**, 1757-1770.

Gandin, L. S., and Murphy, 1992. Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, 120, 361-370.

Harvey, L. O. Jr., Hammond, K. R., Lusk, C. M., Mross, E. F., 1992. The application of signal detection theory to weather forecasting behaviour. *Mon. Wea. Rev.*, **120**, 863-883.

Katz, R. W., and Murphy, A. H., Eds., 1997a. Economic value of weather and climate forecasts. *Cambridge University Press*, 222 pp.

Katz, R. W., and Murphy, A. H., 1997b. 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.

Mason, I., 1982. A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.

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

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

Murphy, A., H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281-293

Murphy, A., H., 1994. Assessing the economic value of weather forecasts: an overview of methods, results and issues. *Met. Apps.*, **1**, 69-73

Murphy, A., H., and Ehrendorfer, M., 1987. On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Weather and Forecasting*, **2**, 243-251.

Palmer, T. N., Brankovic, C., and Richardson, D. S., 1998. A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. To be submitted.

Palmer, T. N., Molteni, F., Mureau, R., and Buizza, R., 1993. Ensemble prediction. ECMWF Seminar Proceedings *'Validation of models over Europe: Vol. 1'*, ECMWF, Shinfield Park, Reading, RG2 9AX, UK.

Roebber, P. J. and Bosart, L. F., 1996. The complex relationship between forecast skill and forecast value: a real-world analysis. *Weather and Forecasting*, **11**, 544-559.

Simmons, A. J., Mureau, R., Petroliagis, T., 1995. Error growth and estimates of predictability for the ECMWF forecasting system. *Q. J. R. Meteorol. Soc.*, **121**, 1739-1771.

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

Wilks, D. S., 1995. Statistical methods in the atmospheric sciences. *Academic Press*, 464pp.

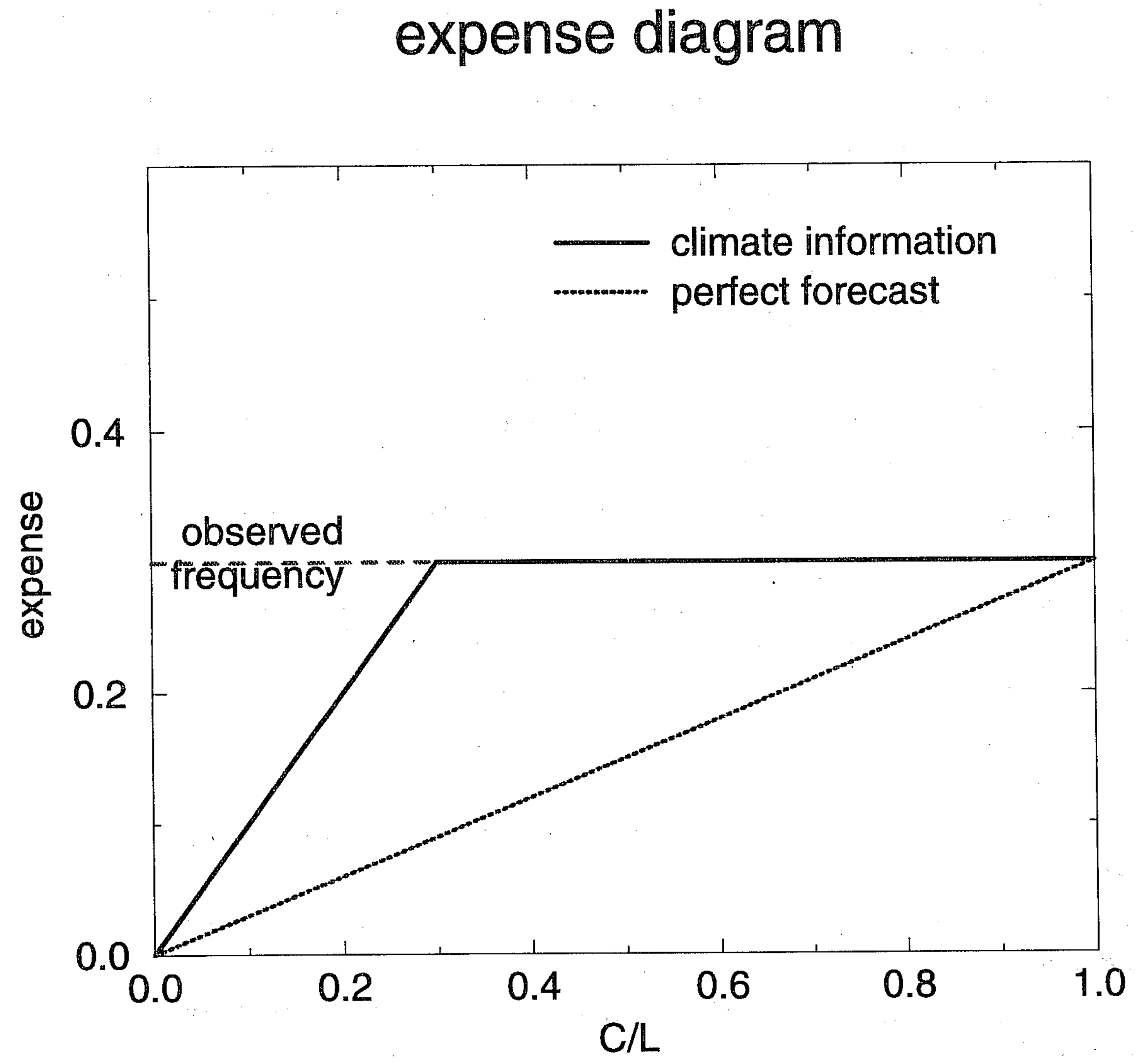

Figure 1: The expense diagram. In this example the event under consideration occurs with a relative frequency of 0.3. Expense per unit loss, given only climatological information, is shown as a function of the cost-loss ratio C/L (solid line). Given perfect knowledge of future occurrences of the event, all users could reduce their expense to the level shown by the dotted line. Information from an imperfect forecast system may allow reduction in expense from the climatological level towards (but not below) this perfect forecast limit.

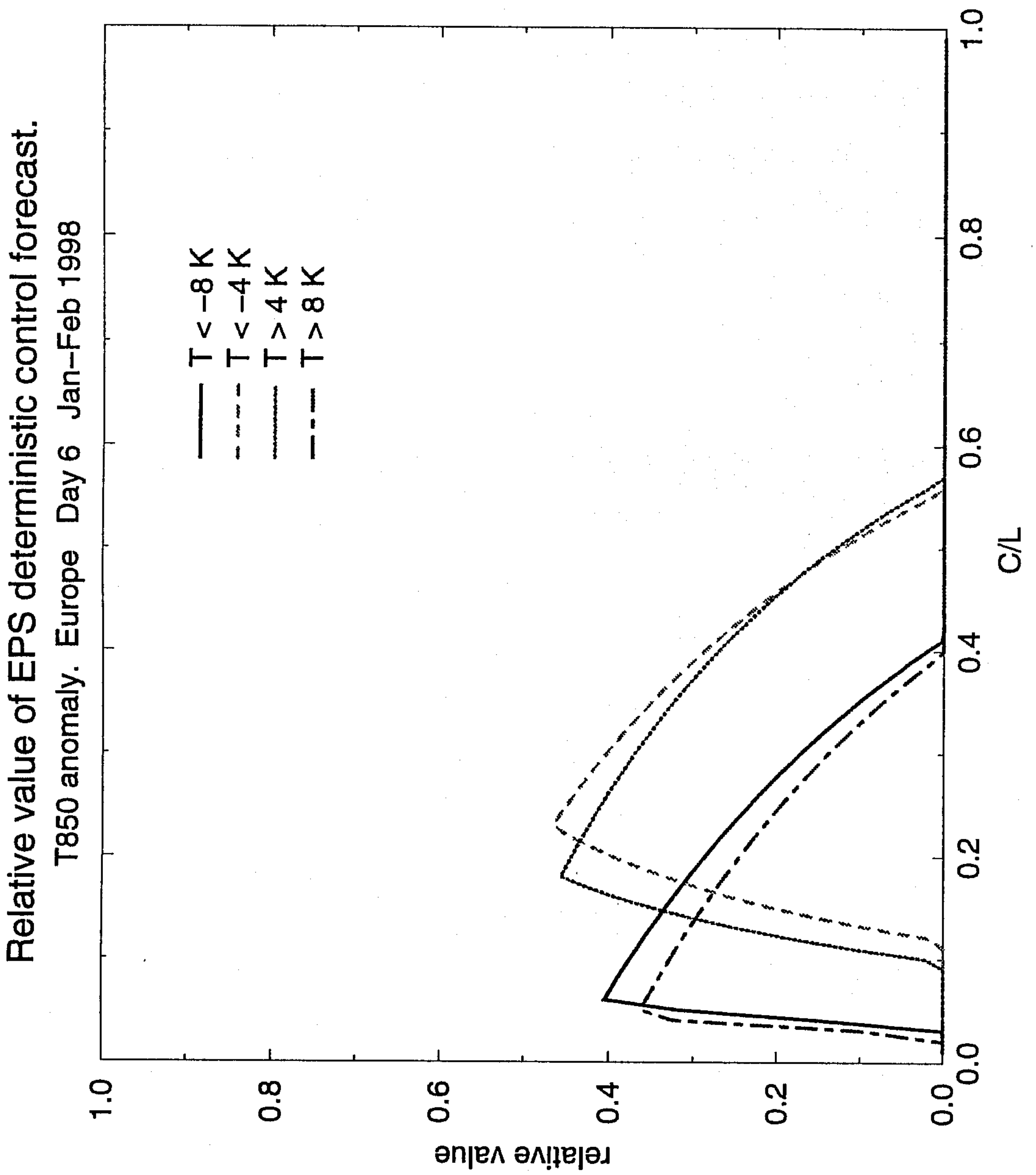

Figure 2: Relative value V of EPS control deterministic forecasts of 850 hPa temperature anomalies T exceeding 4 different thresholds over Europe at day 6 for January and February 1998. V plotted for events T<-8K (solid line), T<-4K (solid line),T>4K (dotted line),T>8K (dash-dotted line).

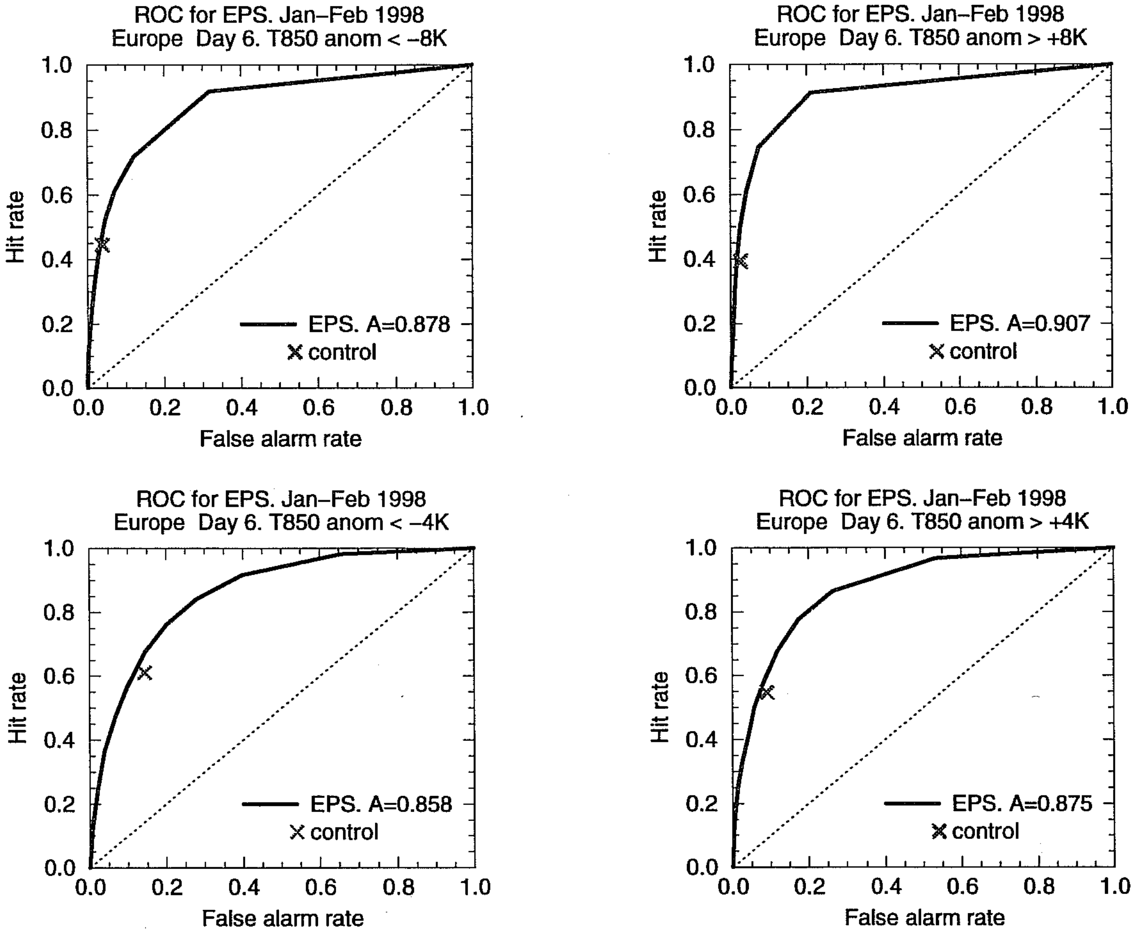

Figure 3: ROCs for EPS forecasts of 850 hPa temperature anomalies T exceeding 4 different thresholds over Europe at day 6 for January and February 1998. Curves shows ROC for EPS probability forecast; the hit rate and false alarm rate for the deterministic control forecast is also shown (x).

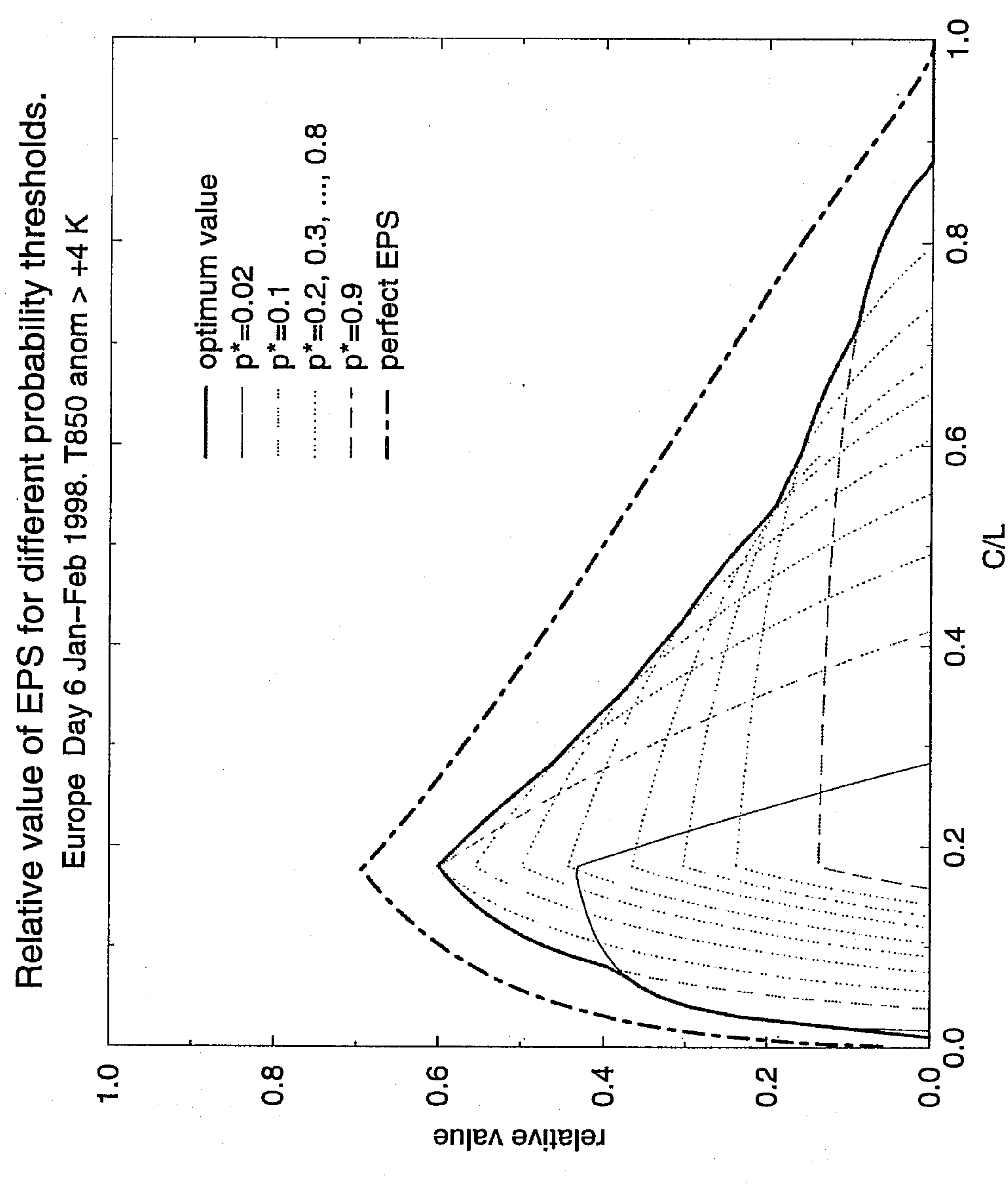

Figure 4: Relative value for EPS forecasts of 850 hPa temperature anomalies T>+4K over Europe at day 6 for January and February 1998. Thin curves show V for various probability thresholds p*; the envelope of these curves (heavy solid line) indicates the overall relative value of the EPS, obtained by choosing the optimal p* for each C/L. The relative value of a "perfect EPS" is also shown (dash-dotted line).

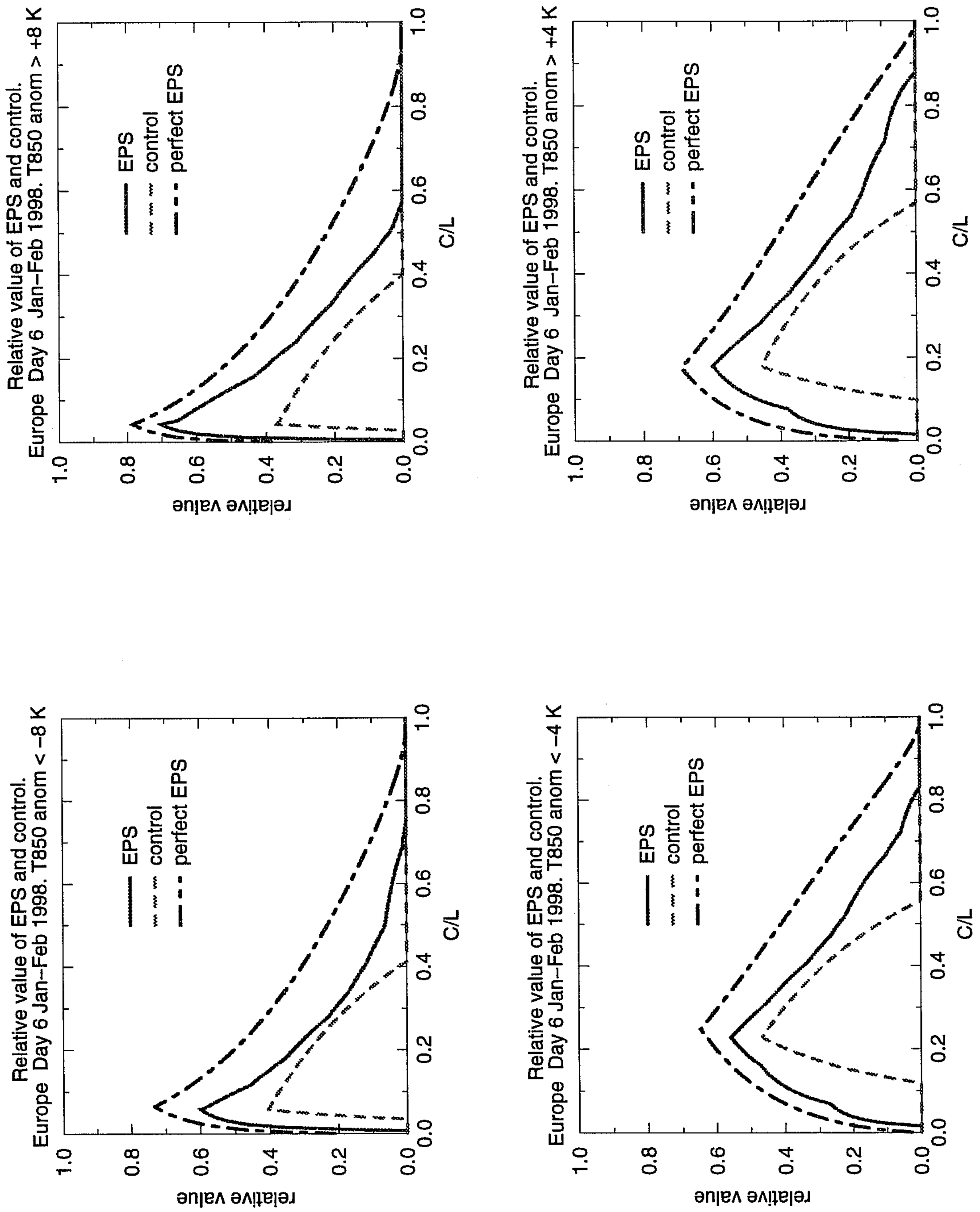

Figure 5: Relative Value for the deterministic control forecast (dashed line), EPS probability forecasts (solid line) and a "perfect EPS" (dash-dotted line) for the four temperature events over Europe at day 6 for January and February 1998.

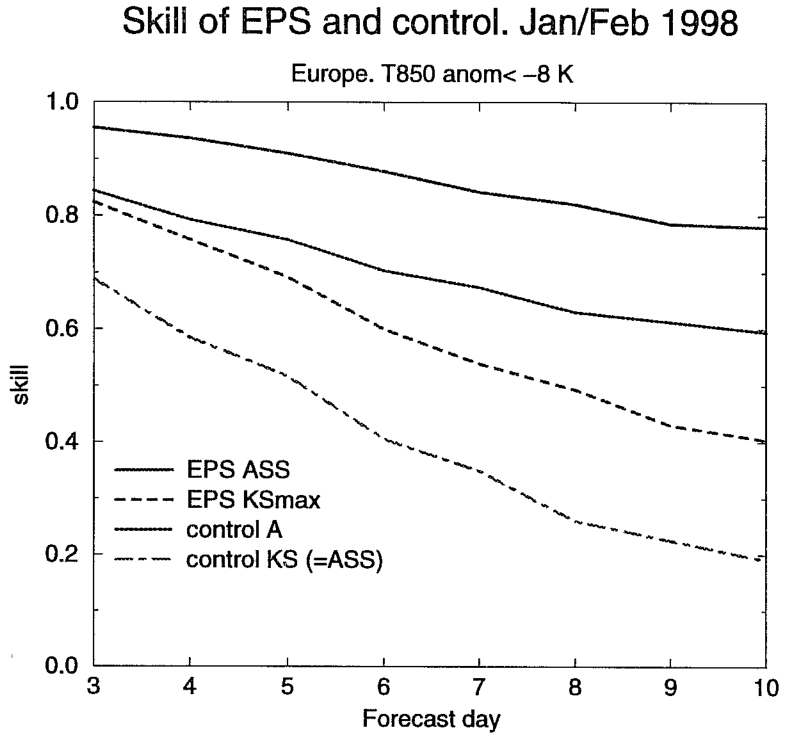

Figure 6: Skill of control forecast and EPS probability forecasts for 850 hPa temperature anomalies T>+4K over Europe for forecast days 1-10, for January and February 1998. Curves are for ROC area skill score ASS (solid line) and maximum Kuipers score KS_{max} (dashed line) for the EPS probability forecasts, and KS (=ASS, dash-dotted line) for the deterministic control forecast.

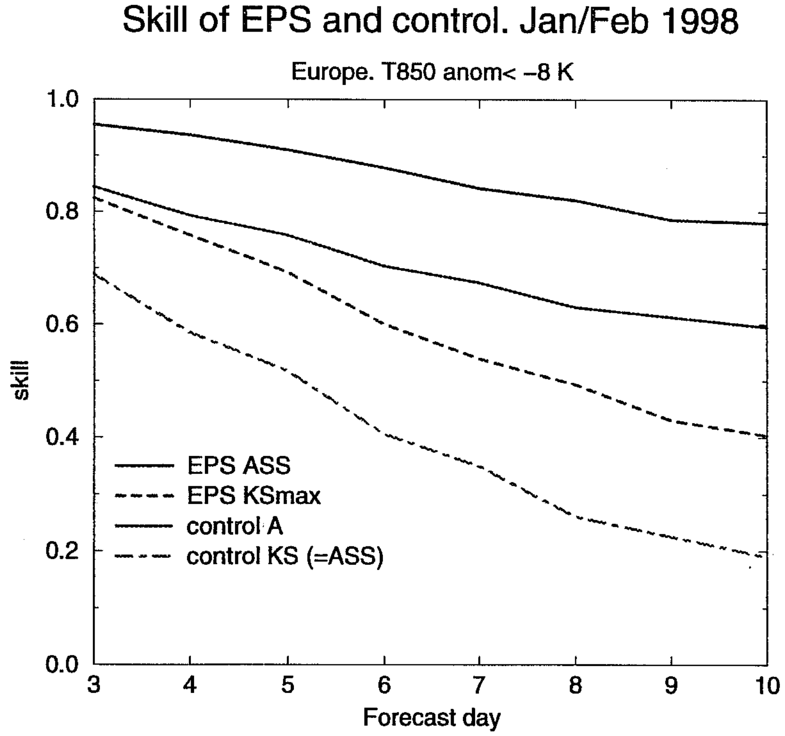

Figure 7: Skill advantage of EPS probability forecasts over the deterministic control forecasts for four temperature events over Europe, for January and February 1998. See text for details.

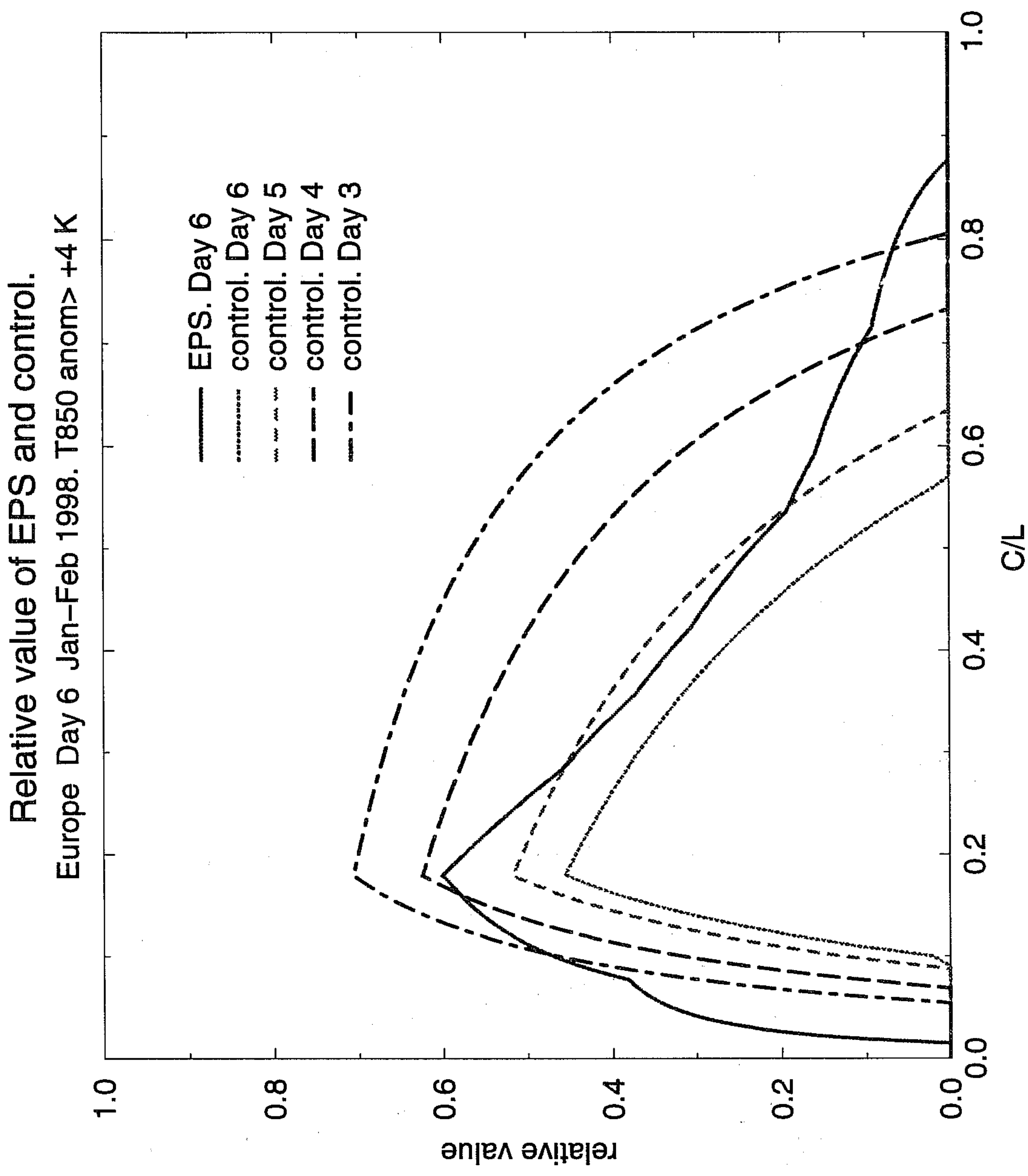

Figure 8: Relative value of EPS probability forecasts at day 6 (solid line) and deterministic control forecasts for different lead times (day 6, dotted; day 5, dashed; day 4, long dashed; day 3, dash-dotted) for 850 hPa temperature anomalies T>+4K over Europe at day 6 for January and February 1998.

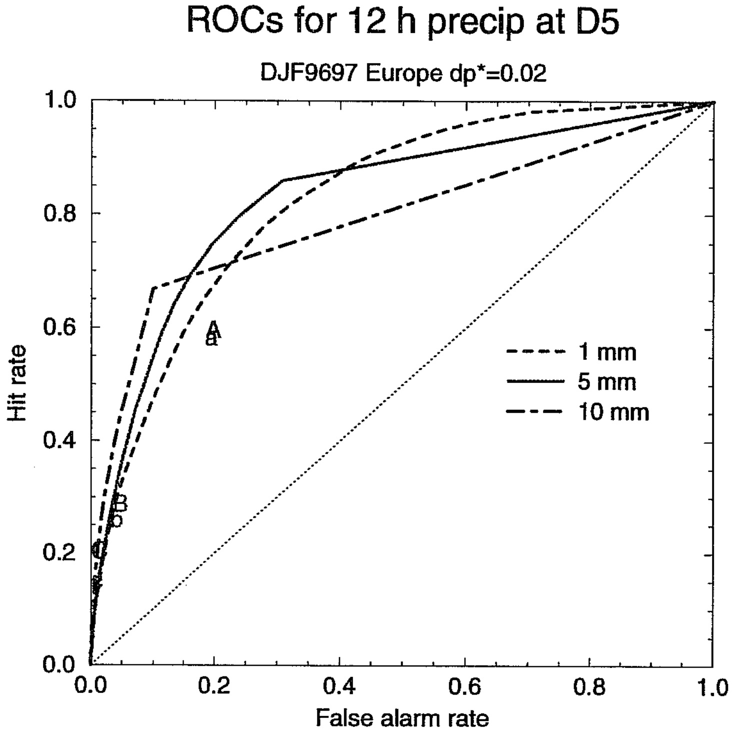

Figure 9: ROCs for EPS forecasts of 12 hour total precipitation exceeding 1, 5, and 10 mm over Europe at day 5 for winter 1996/97. Curves are for EPS probability forecasts calculated using all possible probability thresholds (approximately at 2% intervals). Hit rates and false alarm rates for the deterministic EPS control forecast (lower case letters) and high resolution operational forecast (upper case) are also shown (a,A, 1mm; b,B, 5mm; c,C, 10mm).

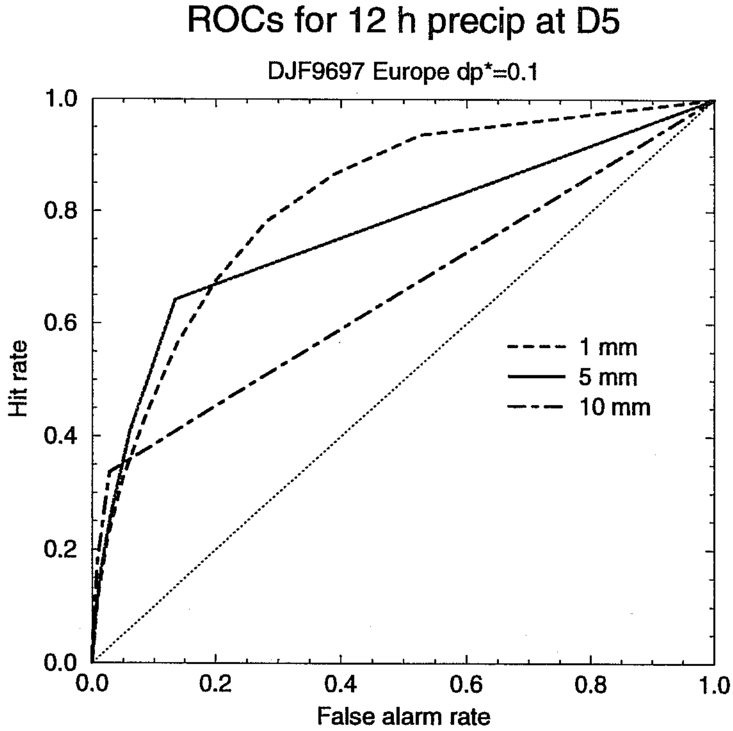

Figure 10: As Fig. 9, but for ROC curves calculated using probability thresholds p* at 10% intervals.

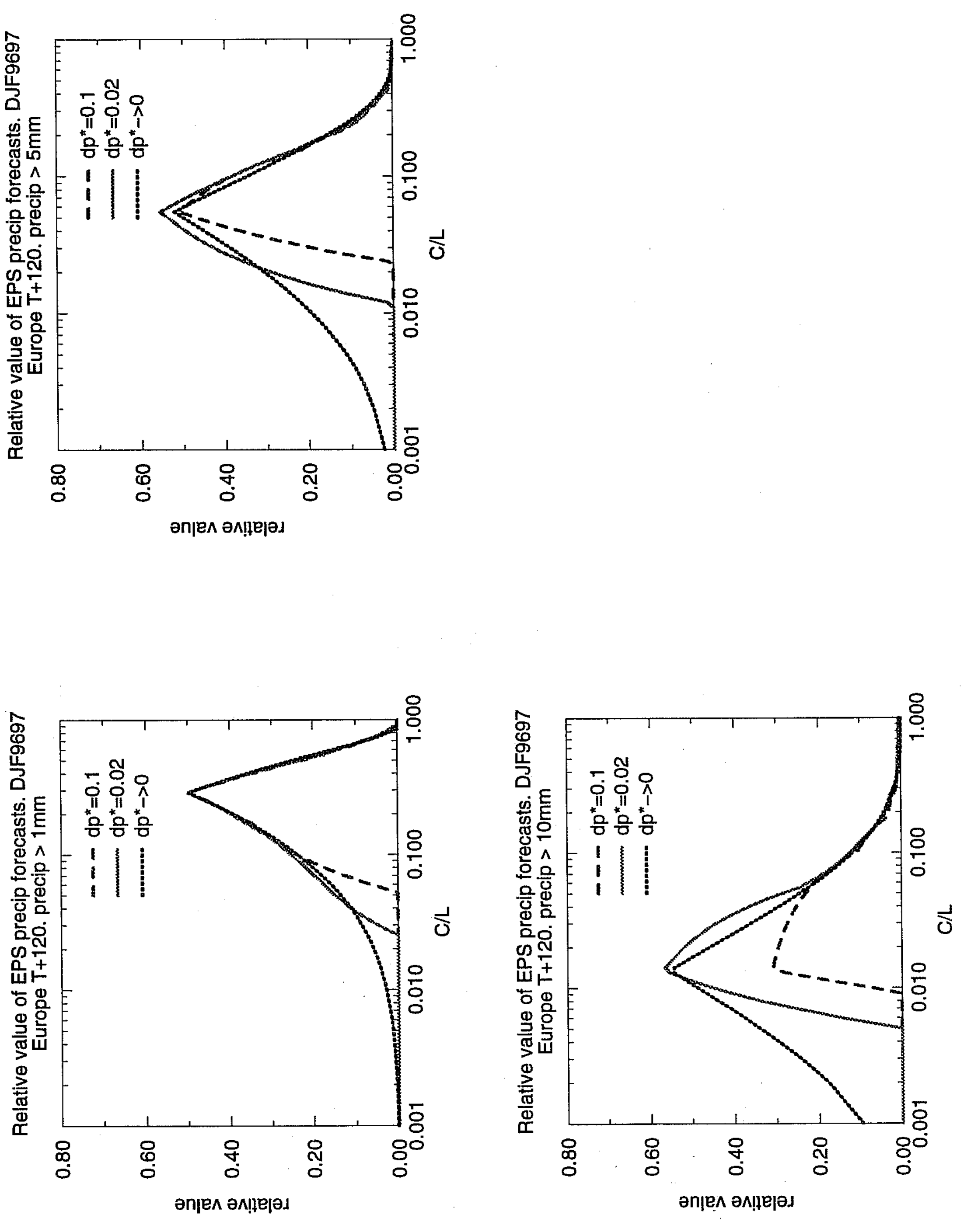

Figure 11: Relative value of EPS probability forecasts of 12 hour total precipitation exceeding 1, 5, and 10 mm over Europe at day 5 for winter 1996/97. Relative value is shown for different resolutions of probability threshold p*: standard 10% intervals, dp*=0.1 (dashed line); all possible thresholds, dp*=0.02 (solid line); estimate of potential relative value as dp* → 0 (dotted line).

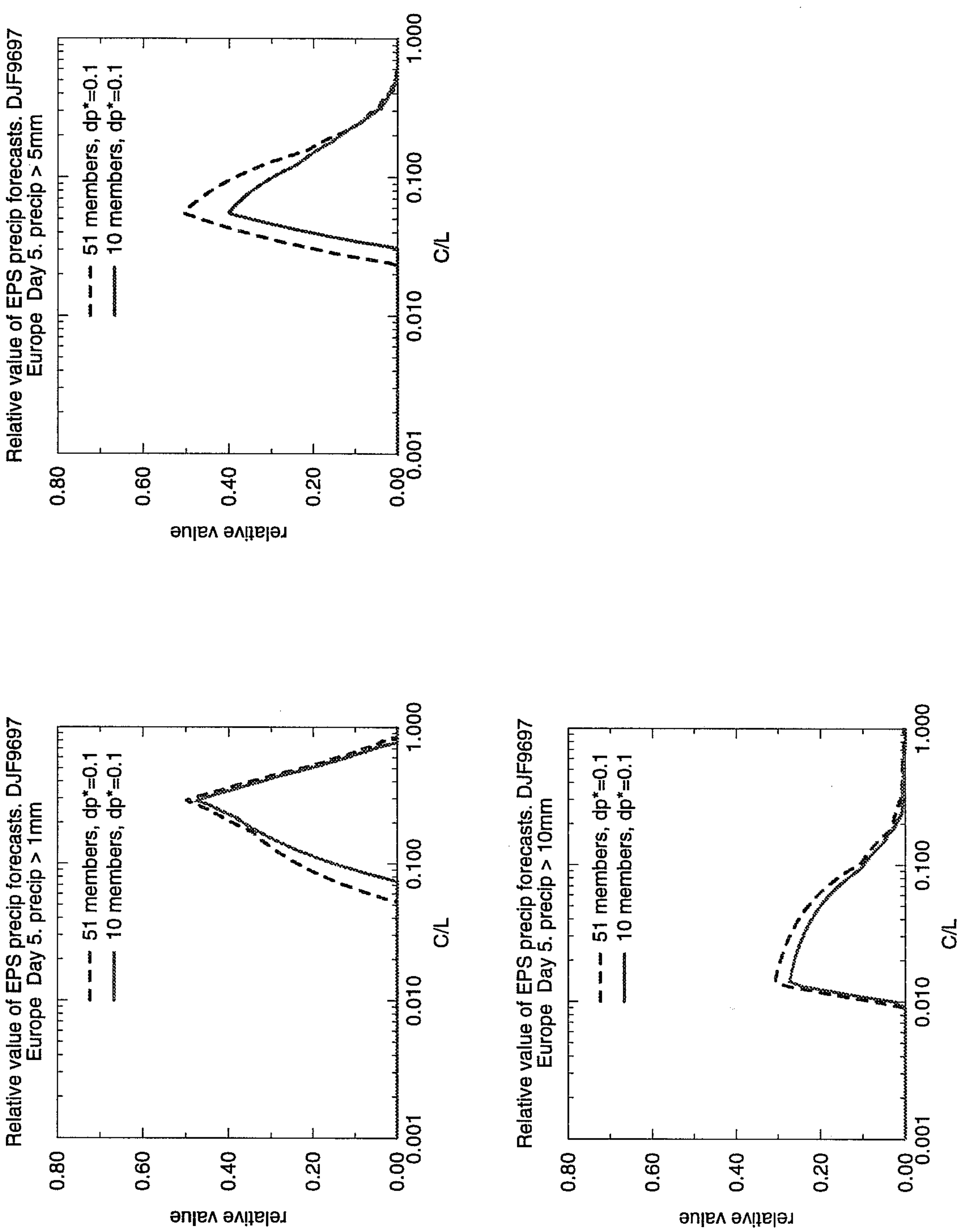

Figure 12: Relative value of 10-member and 51-member EPS probability forecasts of 12 hour total precipitation exceeding 1, 5, and 10 mm over Europe at day 5 for winter 1996/97. Probability thresholds at 10% intervals for both full 51-member EPS (solid line), and reduced 10-member EPS (dashed line).