ON THE ECONOMIC VALUE OF ENSEMBLE BASED WEATHER FORECASTS

Yuejian Zhu[1], Zoltan Toth[1], Richard Wobus[1], David Richardson[2], and Kenneth Mylne[3]

National Centers for Environmental Prediction

Submitted to the

Bulletin of American Meteorological Society

Second revised version

September 7, 2001

[1]SAIC at NCEP

[2]European Centre for Medium Range Weather Forecasts, Reading, UK

[3]Met Office, Bracknell, UK

Corresponding author's address:

Yuejian Zhu, NCEP, Environmental Modeling Center, 5200 Auth Rd., Room 207, Camp Springs, MD 20746
e–mail: Yuejian.Zhu@noaa.gov

ABSTRACT

The potential economic benefit associated with the use of an ensemble of forecasts vs. an equivalent or higher resolution control forecast is discussed. Neither forecast systems are postprocessed, except a simple calibration that is applied to make them reliable. A simple decision making model is used where all potential users of weather forecasts are characterized by the ratio between the cost of their action to prevent weather related damages, and the loss that they incur in case they do not protect their operations. It is shown that the ensemble forecast system can be used by a much wider range of users. Furthermore, for many, and beyond 4 days lead time for all users the ensemble provides greater potential economic benefit than a control forecast, even if the latter is run at a higher horizontal resolution. It is argued that the added benefits derive from (1) the fact that the ensemble provides a more detailed forecast probability distribution, allowing the users to tailor their weather forecast related actions to their particular cost/loss situation, and (2) the ensemble's ability to differentiate between high and low predictability cases. While single forecasts can statistically be supplemented by more detailed probability distributions, it is not clear whether with more sophisticated postprocessing they can identify more and less predictable forecast cases as successfully as ensembles do.

1. Introduction

During the past decade, due to increased computer resources, the development of more realistic atmospheric models, and the recognition of the importance of atmospheric predictability in general, ensemble forecasting became a major component of Numerical Weather Prediction (NWP). NWP centers around the globe (European Center for Medium Range Weather Forecasts, Molteni et al. 1996; the National Centers for Environmental Prediction, Toth and Kalnay 1993; the Canadian Meteorological Center, Houtekamer et al. 1996; the Fleet Numerical Oceanographic and Meteorlogical Center, Rennick 1995; the Japan Meteorological Agency, Kobayashi et al. 1996; and the South African Weather Bureau, Tennant 1998, personal communication) began producing operational ensemble forecasts, where the models are integrated a number of times, started from slightly perturbed initial conditions, in addition to generating the traditional ”control” forecast that starts from the best available atmospheric analysis. Through the ensemble approach one can generate probabilistic forecasts for assessing the case dependent forecast uncertainty related to small errors in the initial conditions and the models used.

When new forecast techniques emerge, some questions naturally arise: Does the new method provide guidance that is of higher quality or more use than existing methods? Is the potential benefit from running a new technique cost effective? Is the new method sufficient with respect to old methods (Ehrendorfer and Murphy, 1988), i. e., is using the old technique redundant, given the new guidance? These are questions that should be addressed with respect to using the relatively new ensemble technique, as compared to relying on the use of a traditional single control forecast.

In earlier studies we presented a detailed analysis of the quality of probabilistic forecasts generated based on the NCEP ensemble forecasting system (Toth and Kalnay, 1997). The performance of the NCEP ensemble forecasts was also compared to that of the ECMWF ensemble prediction system (Zhu et al., 1996), and a single higher resolution

MRF control forecast (Toth et al., 1998). These earlier studies gave valuable insight into the behavior of the different forecast systems, thus providing feedback to the developers. Nevertheless, the ultimate measure of the utility of weather forecasts is arguably the economic and other benefits associated with their actual use in the daily decision making process of individuals and different organizations.

Simplistically, users of weather forecasts either do, or do not take action (e. g., introduce protective action to prevent/reduce weather–related loss), depending on whether a particular weather event is forecast or not. Cost–loss analysis of different complexity can be applied to evaluate the economic impact of the use of weather forecasts on the users (Murphy, 1985; Katz and Murphy, 1997). Studies of the economic value of weather forecasts can either be descriptive, assessing the value of forecasts used, often suboptimally, by existing customers; or prescriptive, identifying the potential value of forecasts, assuming they are used in an optimum manner (Stewart, 1997).

In this paper we evaluate the potential economic value associated with the use of an ensemble of forecasts, vs. an equivalent, and a higher resolution control forecast, after minimal postprocessing. A relatively simple cost–loss model that was discussed previously by Richardson (2000a) and Mylne (1999) and is similar to that of Wilks (2001) will be used. We note that cost–loss analysis is only one type of model that can be applied to investigate the potential value of weather forecasts, as described in Katz and Murphy (1997). The cost–loss analysis approach followed in this study obviously has its limitations. For example, not all values can be expressed in terms of dollar amounts; the loss of life is one such example. Nevertheless the economic analysis used offers a framework that, after some simplifications, can generally be applied in most cases.

2. Cost–loss analysis

A decision maker becomes a user of weather forecasts if he/she alters his/her actions based on forecast information. Whether a user is expected to benefit from the use

of a forecast system in the long run can be assessed based on 2x2 matrices (Table 1). If the user does not take action and the event does not occur (correct rejection), there is no cost to the user (N=0). If the event does occur and the user is not protected (miss), he/she will suffer a loss L. If a user takes preventive action to guard against this potential loss, the user will incur a cost (C<L). If the event does not occur (false alarm), C is the total cost on the user's side; if the event occurs (hit), in addition to C, the user may also incur some reduced, unprotectable loss L_u. Note that the sum of C and L_u is usually called mitigated loss (M), and typically $C \leq M < L$. The expenses associated with each combination of action and outcome are shown in Table 1, where the total loss is expressed as the sum of the loss which can be

		FORECAST/ACTION	
		YES	*NO*
OBSERVATION	YES	Hit (h) ***Mitigated Loss ($C+L_u$)***	Miss (m) ***Loss ($L=L_p+L_u$)***
	NO	False Alarm (f) ***Cost (C)***	Correct Rejection (c) ***No Cost (N)***

Table 1. Contingency table indicating the costs/losses accrued by the use of weather forecasts, depending on forecast and observed values.

protected against (L_p), and the remaining unprotectable loss (L_u).

a. Expected expense

We assume that the user takes action depending on whether the event is forecast or not. If the relative frequency of the four different outcomes in Table 1 is known and marked by *h, f, c,* and *m*, one can assess, in a statistical sense, the expected expense of a user of a forecast system:

$$E_{forecast} = h(C+L_u) + fC + m(L_p+L_u). \qquad (1)$$

Since we assume that a correct rejection is associated with no cost (N) on the part of the user of weather forecasts, this term is omitted from Eq. 1. Furthermore, one can determine

the expected expense associated with using climatological information only:

$$E_{climate}=Min[o(L_p+L_u),\ C+oL_u] = oL_u + Min[oL_p,\ C], \qquad (2)$$

where *o* is the climatological frequency of the event. Based on the climatological frequency of the event, and on the user's associated costs and losses, the user will either *always* or *never* take protective action. A decision maker will choose to use a forecast system if his/her expected expense associated with the forecast system will be lower than that associated with using only climatological information.

The minimum expense for a user, given a perfect forecast system that provides accurate predictions for the occurence and non–occurence of a particular event, can be written as:

$$E_{perfect}=o(C+L_u). \qquad (3)$$

In this ideal situation, the user takes protective action if and only if a harmful event actually occurs.

b. Economic value

Using Eqs. (1–3) the definition of the relative economic value *(V*) of a forecast system can be given as

$$V = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}}. \qquad (4)$$

Using a forecast system that is perfect will result in an economic value of 1 (maximum value), while a forecast system associated with the expected expense equal to (larger than) that attainable using climatological information only will have zero (negative) economic value. Economic value is unbounded from the negative side. Negative values indicate that following a forecast system will actually cost the user more than following the best climatological option. The best way to avoid such a misuse of forecasts is to use the concept of "value" as discussed here and elsewhere to optimize decision making.

Noting that h+m=o, we can rewrite Eq. 1 as:

$$E_{forecast}=oL_u+(h+f)C+mL_p. \qquad (5)$$

It is now clear that the term oL_u is common to each expected expense and that this unavoidable part of the total loss will not appear in the expression for *V*. Substituting Eqs. 2, 3, and 5 into Eq. 4 gives:

$$V = \frac{min[oL_p, C] - (h + f)C - mL_p}{\cdot min[oL_p, C] - oC}. \qquad (6)$$

Dividing each term on the right hand side of Eq. 7 by L_p and recognizing that the ratio *(r)* of the cost of protection to the amount of potential loss which can be protected is $r=C/L_p$, we arrive at:

$$V = \frac{min[o, r] - (h + f)r - m}{\cdot min[o, r] - or}. \qquad (7)$$

Note that the economic value of forecasts *(V)* depends only on two forecast performance parameters (*h* and *m* in Eq. 7), which can also be expressed by the hit rate (*HR)* and false alarm rate *(FAR)* used in the definition of the relative operating characteristics (ROC), indicating the close relationship discussed further in section 4 between economic value and ROC characteristics. Beyond the parameters describing the forecast system, *V* also depends on *o*, the climatological frequency of the event, and on $r=C/L_p$, the cost–loss ratio characterizing the users of a forecast system. The fact that all users can be characterized in this framework by a single variable, C/L_p, offers a convenient way to evaluate the potential economic value of any forecast system for all users on a two–dimensional, *V* vs. C/L_p plot.

3. Experimental setup

In the following section we compare the economic value of the MRF T62 and T126 resolution control forecasts to that of a 14–member set of the T62 horizontal resolution NCEP ensemble for the April – June 1999 period. Note that the computational cost of generating either a higher, T126 resolution control forecast, or a 14–member T62

resolution ensemble is an order of magnitude higher than that of running a T62 resolution control forecast only. In the example below, weather events are defined as the 500 hPa geopotential height at gridpoints over the Northern Hemisphere extratropics (as routinely defined for verification purposes at NCEP, 20N – 77.5N) being in any of 10 climatologically equally likely bins. The reason for the choice of 500 hPa geopotential height is that climatological information in the above format was readily available from an earlier study (Toth et al. 2001). Climatological decisions (see Eqs. 2 and 4) are also based on this 15–year climatology.

The user of an ensemble of n forecasts has n options for use as decision criteria with respect to his/her weather related action. He/she can choose to take action only if all n forecasts predict the adverse weather, act if at least $n–1$, $n–2$, ..., or even if at least 1 member predicts the adverse weather. Each of these decision criteria corresponds with a different economic value. Based on their C/L_p ratio, users can choose the decision criterion that offers the most value to them. In fact, it can be shown that the best decision level p, corresponding to the predicted probability of the weather event, is equal to C/L_p (assuming perfectly reliable forecasts, Murphy 1977). The higher the cost of the protective action relative to the potential loss, the more certainty the user requires about the forecast before he/she takes action. One of the potential advantages of using an ensemble forecast system is that it naturally provides a multitude of such decision criteria. Different users can then tailor their use of the forecast information to their particular application, characterized by their cost–loss ratio.

Relative frequency values based on counting how many ensemble members predict a certain event usually provide probabilistic forecasts that are not reliable in a sense that they do not necessarily match corresponding observed frequency values. This is because of deficiencies in model and ensemble formulation. For example, when half of the ensemble members predict a weather event, that event may, over a long verification

period, verify only 40% of the time. Such biases in ensemble–based probabilities are generally consistent in time and can be easily eliminated (see, e. g., Zhu et al., 1996). The calibrated forecast that would be issued based on the past verification statistics in the above case, where half of the ensemble members predict an event, for example, is 40%. The April – June 1999 ensemble based probabilistic forecasts evaluated in this paper have been calibrated using independent data from February 1999 verification statistics.

For each cost–loss ratio shown in Figs. 1–4 the decision criterion for the ensemble is based on the calibrated probability forecasts. In particular, it is assumed that a user will take protective action if the calibrated probability forecast value is greater than or equal to the cost–loss ratio $(p \geq C/L_p)$. For the extremely high (and low) probablity values where the finite ensemble cannot provide optimum guidance, the best available guidance was used, i. e., the highest (lowest) probability values associated with all (at least one) members predicting the weather event.

The above decision making algorithm, based on the users' cost–loss ratio and probabilistic forecasts calibrated based on independent verification data, represents an operationally feasible strategy for the use of ensemble guidance. In some earlier studies (Mylne 1999; 2001; Richardson 2000a) a slightly different approach was used, where the optimum decision level for a particular user was identified directly by evaluating the economic value associated with the use of different decision criteria. This process, in some sense, is equivalent to calibration. Note, however, that in these earlier studies calibration was performed on the forecast sample that was evaluated (dependent data), assuming that the forecasts can be perfectly calibrated. In contrast, no such assumption is made in the present study. By using calibrated probabilities that are adjusted based on prior and independent verification statistics, the economic value of the ensemble forecast system is evaluated in a more realistic setting, accounting for the information loss that inevitably occurs in the calibration process.

In contrast to the ensemble system that naturally offers multiple decision levels, deterministic guidance from a single forecast, unless its form is changed via postprocessing, can only be interpreted by a user in one way. If a particular adverse weather event is forecast, the user can take protective action, and do nothing otherwise. The yes–no forecast of a deterministic system, based on past verification statistics, can be converted to dichotomous probabilistic forecasts just as the ensemble–based probabilistic forecasts can be calibrated, see, e. g., Murphy (1986), and Toth et al. (1998). Since there is only one decision level envolved, we assume probabilistic forecasts based on a control forecast can be perfectly calibrated, and for calibration we use verification statistics based on the sample period. Given that the ensemble forecasts are not assumed to be perfectly calibrated and that their actual calibration is done with a simple algorithm the economic value results shown in the next section represent a conservative estimate with respect to the benefits of using an ensemble as compared to a single control forecast.

4. Results

a. Economic value

In Fig. 1 we show the economic value of the two control forecasts vs. an ensemble of forecasts at 24–hour lead time, as a function of the C/L_o ratio, as discussed above. The economic value comparison results indicate that most potential users, except those with cost–loss ratios in a relatively narrow band between 0.2 and 0.5, can realize more economic value when using the ensemble forecasts. At and beyond (72–)120 hours lead time (Figs. 2–4) (virtually) all users are better off using the ensemble system than the control forecasts. Furthermore, the range of cost–loss ratios for which the forecasts exhibit value, compared to using climatological information only, is substantially widened, indicating that a much larger group of users can benefit from the ensemble forecasts as compared to the control forecasts.

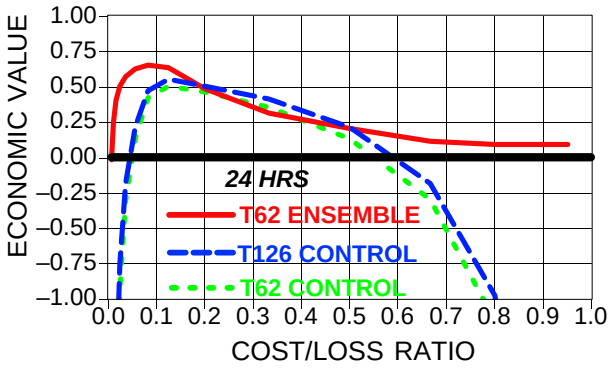

Fig. 1. Economic value of 24–hour MRF T126 (dashed) and T62 (dotted) control forecasts, and 14–member T62 ensemble forecasts (solid) in predicting events defined in terms of 10 climatologically equally likely bins for 500 hPa height over the NH extratropics, for April–June 1999, for users characterized by different cost/loss ratios (horizontal axis). For the ensemble, the optimum decision strategy evaluated here is based on the probabilistic forecasts, calibrated using February 1999 verification statistics, being greater or equal to C/L_p ($p>C/L_p$). Values below –1 are not plotted.

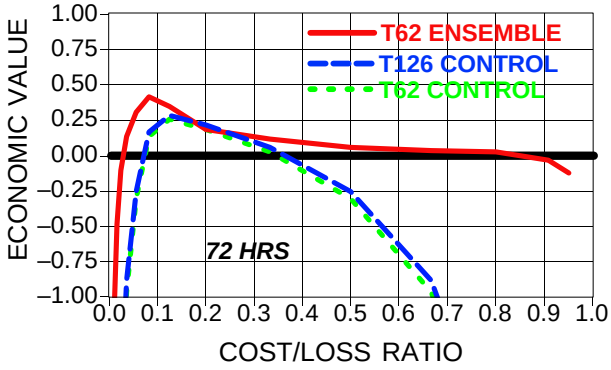

Fig. 2. Same as Fig. 1, except for 72–hour forecast lead time.

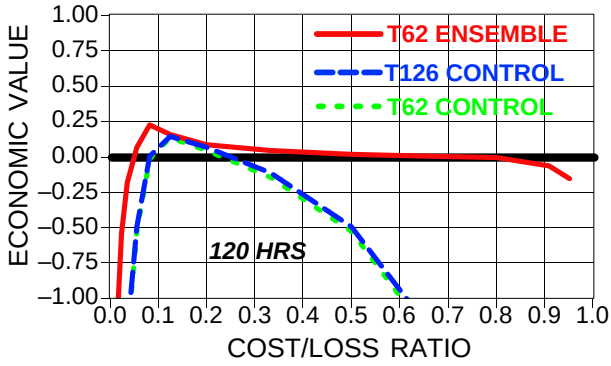

Fig. 3. Same as Fig. 1, except for 120–hour forecast lead time.

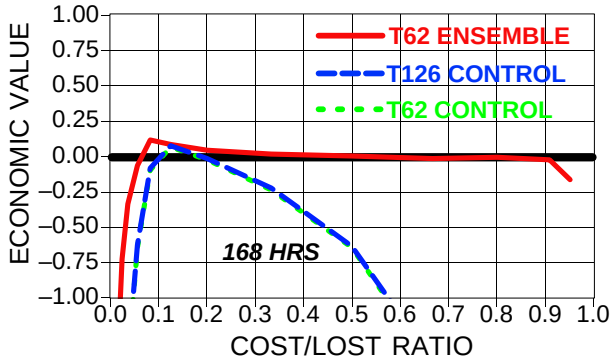

Fig. 4. Same as Fig. 1, except for 168–hour forecast lead time.

Note that on each of the figures the largest economic benefit is, as expected theoretically (see, e. g., Richardson, 2000a), attained by users whose C/L_p ratio is approximately equal to *o*, the climatological frequency of the weather event, which in our case is 0.1. Note also that with increasing lead time the economic value, as compared to using perfect forecasts, just as the forecast information content (see Fig. 8 of Toth et al., 1998), is reduced. Finally, the low negative values for the ensemble at very low cost/loss ratios arise due to the small size of the ensemble. As described in section 3, the lowest calibrated probability level available from the ensemble (p_l) corresponds to the outcome of at least one ensemble member predicting the event in question. For lack of a better choice this criterion is used for all cost/loss ratios below p_l, leading to poor performance in that range.

b. Summary measures

Beyond ecnomic value defined in Eq. 7 and evaluated in Figs. 1–4 there exist a number of measures that attempt to summarize the general value of different forecast systems. Some of these summary measures are based on an assumption about the distribution of properties protected by all (or a group of) users with various C/L ratios. These summary measures, therefore, can be considered as overall economic value estimates given their assumption about the distribution of protected values along different C/L ratios. Unfortunately little if any information is available on most users' cost–loss ratio. As an

alternative, Wilks (2001) considers several artificial distributions of C/L among users in his cost–loss analysis.

Relative Operating Characteristics–area (ROC–area, see, e. g., Mason, 1982) is one common summary measure of ensemble forecast performance based on signal detection theory. Using the notation of Table 1, an ROC diagram plots the hit rate *HR=h/(h+m)* of a forecast system against its false alarm rate *FAR=f/(f+c).* The overall performance of a forecast system is measured by the ROC–area defined by the points (0,0), (1,1), and the point(s) representing the forecast system (see, eg., Stanski et al., 1989). The closer a curve is to the upper left hand corner, the more ability the studied forecast system has in delineating between conditions under which a certain event (in this study, one of 10 climatologically equally likely bins) does or does not occur. A perfect forecast system would have a ROC–area of 1 while a system with no ability to distinguish in advance between different weather events would have a score of 0.5 (i. e., points lying on the diagonal defined by 0,0 and 1,1). As shown by Mylne (1999) and Richardson (2000b, 2001) the ROC–area is closely related to the economic value of a forecast system. Note that ROC disregards forecast reliability (or lack of it) altogether, effectively assuming that before their use the forecasts can be perfectly calibrated.

Another summary score is the Brier Skill Score (BSS), which is a measure related to ROC–area for systems that produce reliable forecasts (i. e., forecast probabilities that exactly match observed frequencies, Talagrand et al. 1998). BSS measures the overall economic value associated with a particular forecast system, assuming that when all users are considered, the same amount of property is at stake at each cost–loss ratio value (Murphy 1966). For the comparison of a single control forecast with ensemble based probabilistic forecasts BSS gives an even larger benefit to the ensemble than ROC–area, reflecting the wider range of users who gain positive value from the ensemble. Another alternative would be to use the economic value associated with the C/L_p yielding the

maximum value (Richardson 2000a). Compared with the ROC–area, this would give somewhat smaller differences (but still up to 50% in Figs. 1–4). The maximum value recognizes the fact that the ensemble forecast system is better for users with C/L=o, but does not reflect the additional benefits for the majority of other users that exist due to the provision of multiple decision levels in the form of multiple level probabilistic forecasts (Richardson 2000a).

As an example of a summary measure fig. 5 shows an ROC–area based skill score

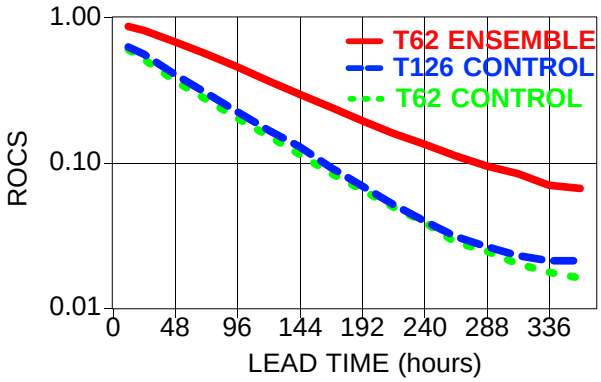

Fig. 5. ROC (Relative Operating Characteristics) area skill score for the T126 (dashed) and T62 (dotted) control forecasts, and the 14–member T62 ensemble forecasts (solid) for the 500 hPa height, NH extratropics, for April–June 1999. Scale on vertical axis is logarithmic.

(ROCS), defined by Richardson (2000a) as:

$$ROCS=2(ROCA - 0.5), \qquad (8)$$

that is an indicator for the overall utility of a forecast system, as a function of lead time. As can be seen from Figs. 1–4, the relative benefit of the ensemble (compared to the controls) will tend to be greater than this for users with C/L<o, but will be less for some users with C/L>o.

The ensemble forecast system is found to outperform the control forecasts at all lead times. For example, at day 2 (6) lead time the use of the ensemble forecast system

provides close to 70% (130%) improvement over the control forecasts; to put it in another way, a 4–day (10–day) ensemble forecast has a score as high as a 2–day (6–day) single control forecast. These results are in good agreement with Figs. 1–4, and at later lead time with those of Mylne (1999) and Richardson (2000a).

It is interesting to note how much more the users can benefit from the ensemble–based multiple decision level forecast system than from a simple increase in the horizontal resolution of the control forecast from T62 to T126 – each of which requires approximately an order of magnitude more resources than running a low resolution T62 control only. The high ensemble scores make the difference between the two different resolution model versions look rather small. However, in terms of NWP advancements, the T126 resolution model represents a rather significant improvement over the T62 version of the NCEP MRF model.

The economic value results in Figs. 1–4 and the related ROC–area results in Fig. 5 refer to the value of direct output from the NWP systems investigated, after minimal postprocessing that makes both the control and ensemble systems reliable. When evaluating these results, however, we must note that a single control forecast, without statistical postprocessing, offers only one threshold for decision makers while an ensemble of *n* members offer *n*, depending on how many members indicate the occurence of a critical weather event. The large difference between the ensemble and the control curves in Figs. 1–5 highlights the importance of using detailed probabilistic forecast information in economic decision making processes.

The fact that direct output from a control forecast offers only a single decision level is clearly reflected on ROC curves like that shown in Fig. 6 for 5–day lead time forecasts where both control forecasts are represented by one point only. For a single forecast that, without postprocessing, offers only one decision criterion, ROC–area is defined by the triangle given by (FAR,HR) for the single point and the (0,0) and (1,1) points (see example

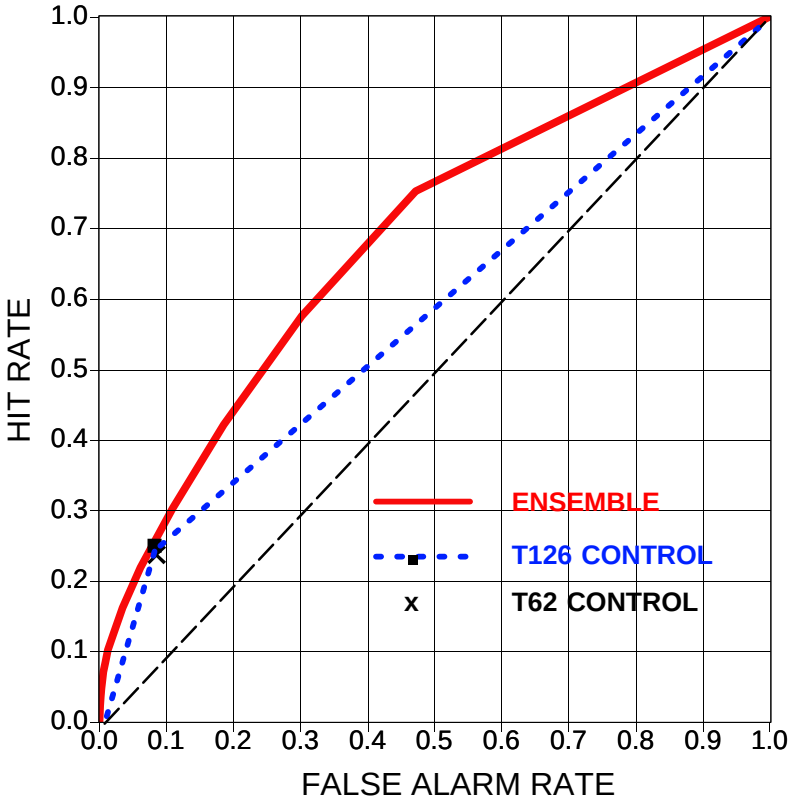

Fig. 6. ROC (Relative Operating Characteristics) curve for a 5–day lead time 14–member T62 ensemble of forecasts and for the T126 and T62 control forecasts predicting events defined in terms of 10 climatologically equally likely bins for the 500 hPa height, NH extratropics, April–June 1999.

in Fig. 6). Note in the example of Fig. 6 that the control point is only slightly below the ensemble curve. The main difference in ROC–area between the control and ensemble forecast systems comes from the larger number of thresholds used to define the ensemble ROC curve. This greater number of thresholds is directly related to the wider range of users for which the ensemble has positive value compared to the control forecast (cf. Figs 1–4). The substantial difference in ROC–area between ensemble and control forecasts emphasizes the importance of the flexibility this range of decision thresholds offer to the users.

We note that if we parametrized the ROC curves for both the ensemble and the control forecasts (Mason, 1982; Richardson 2000a) the difference in ROC–area would be substantially reduced. However this would indicate potential (and not actual) differences in value, achievable only if a sufficiently wide range of useful forecast thresholds for the single control forecast system could somehow be made (Harvey et al. 1992). One possible way to achieve this would be to predict 50 mm precipitation, for example, not only when the control forecast exceeds that amount, but with less probability, also when the forecast reaches 20 or even only 10 mm (see, e. g., Atger 2001).

It is conceivable that the performance of such a multiple decision level system based on a control forecast can reach or possibly even exceed the performance level of an ensemble–based system in case the unpostprocessed control forecast point lies above the ensemble curve on a ROC chart. This is the case for the low (T62) and high resolution (T126) controls for lead times up to 12 and 96 hours respectively (Fig. 7). This indicates that for some users the control forecasts are more valuable at short lead times. Beyond 4 days lead time both control forecast points, however, lie below the ensemble curve (Figs. 5 and 7), inidicating that the ensemble is a better forecast system for all users. At these lead times

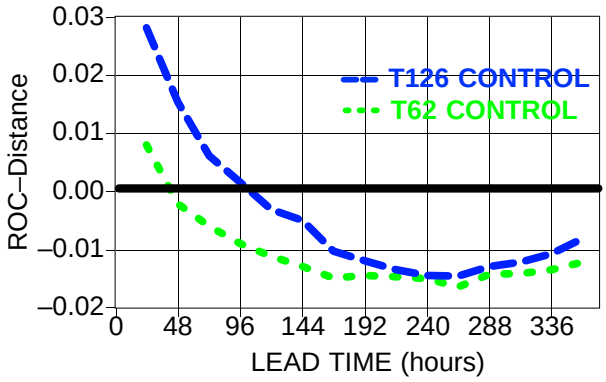

Fig. 7. Same as Fig. 5 except for ROC–distance, defined (on linear vertical axis) as the distance between a control point and the closest point on the ensemble polygon. Positive (negative) values indicate the control point is above (below) the ensemble curve.

there is no reason to assume that if more sophisticated statistical postprocessing is applied

to both the high resolution control and the low resolution ensemble forecast systems the one based on the control forecast only would perform better. These results indicate that using the same computational resources, potentially more economic benefit can be gained from generating an ensemble of forecasts than from increasing the horizontal resolution of the control forecast, at least for lead times beyond 4 days.

5. Discussion

a. Why the ensemble approach is successful

As discussed earlier, an ensemble of forecasts naturally offers a multitude of decision levels compared to a single yes–no decision based on a control forecast, providing detailed probability distributions instead of only two levels of probabilities. Toth et al. (1998) showed that the use of detailed ensemble based probability distributions (as compared to the use of only two probability levels) substantially improves forecast performance in terms of ROC, Ranked Probability Skill Score, and information content. As discussed earlier, multiple–value probability forecasts can of course be constructed based on a single deterministic forecast, using past verification statistics. Such a system can produce statistically postprocessed, bias–free probabilistic forecasts. Atger (1999) found that at least beyond 3 days lead time 500 hPa height forecasts the ECMWF ensemble prediction system outperformed such a system based on postprocessed control forecasts. In their compex economic value analysis addressing hypothetical applications in the electricity sector Smith et al. (2001) came to similar conclusions. Talagrand and Candille (1999, personal communication) reported similar results. The ensemble's better performance in these comparisons, since the control forecast was supplemented by a detailed probability distribution, must be due to some genuine information contained in the ensemble but not in the control based distributions.

The ensemble based distributions can surpass their control based couterparts in two ways. First, due to nonlinear effects (Toth and Kalnay 1997) the ensemble distribution

may be centered closer to truth than the distribution based on a single forecast only. While at short lead times a higher resolution control may have an advantage due to its increased accuracy (see, e. g., Fig. 3 of Toth et al., 1998), at longer lead times the ensemble has an advantage, due to its nonlinear error filtering capability. While these differences may have a substantial contribution, comparing Figs. 3 and 7 of Toth et al. (1998) suggests that they play only a secondary role. A more important contribution of the ensemble may be its ability to capture day–to–day variations in the expected uncertainty of the forecasts (Toth et al. 2001; Ziehmann 2001). The ensemble can distinguish between forecasts with higher and lower than average uncertainty at the time the forecasts are issued. As Toth et al. (1998) showed, the ensemble provides important extra information to the users through its case dependent uncertainty estimates. While statistical postprocessing of some sophistication applied on a control forecast system may be able to capture part of the day to day variations in predictability, it is not likely that all information that affects predictability (i. e., case dependent initial errors and their evolution in the forecast) could be captured through statistical approaches.

b. Limitations and open questions

All the results presented in this study pertain to forecasts of the 500 hPa height over the Northern Hemisphere extratropics, made at T62 and T126 model resolution. Similar results were obtained by Richardson (2000a) and Mylne (1999) using sensible weather elements. Further studies, however, are necessary to analyze the economic value related to the use of ensembles vs. higher resolution control forecasts with respect to other weather elements in higher resolution forecast models at various lead time ranges, with the use of more complex decision making tools. Can some general guidelines, such as the presence of large forecast uncertainty, and/or large and predictable variations in it, be established under which it is more advantageous to spend resources on running an ensemble, instead of increasing the spatial resolution of the model used in NWP

forecasting? Under what general conditions may an ensemble forecast system be sufficient (Ehrendorfer and Murphy 1988) for the high resolution control forecast, making the control forecast redundant (and its generation unnecessary)?

Ensemble forecasts for sensible weather elements should preferably be statistically postprocessed to eliminate possible systematic errors or model biases before they are used in weather forecasting. Calibration is an important issue for practical applications since, as Wilks (2001) showed, uncalibrated forecasts can suffer a great reduction in their expected economic value. Probabilistic forecasts based on the NCEP ensemble (500 hPa height, 10 climatologically equally likely events) were successfully calibrated by a simple method, used also in the present study, by Zhu et al. (1996) and Toth et al. (1998). The success of such a calibration depends on the relative stationarity of the NWP analysis and forecast system on one hand, and the natural climate system on the other. The calibration of forecasts for intermittant or less frequent events is a more problematic task that calls for further investigation.

Statistical postprocessing has also been a critical element in the interpretation of traditional single control forecasts (e. g., Carter et al., 1989). Note that the purpose of statistical postprocessing of the ensemble forecasts is different from that of a single control forecast. MOS, for example, not only attempts to eliminate the bias from the forecasts on which it is applied but also hedges the forecasts (Murphy, 1978) toward climatology (the larger the expected forecast error, the more so). A single control forecast is normally used to provide a best estimate of the future state of the atmopshere, and hedging serves well this purpose. Ensemble forecasting, however, has a different goal, providing a detailed forecast probability distribution. In this case hedging, that brings all forecasts, originally intended to represent the inherent forecast uncertainty, closer to climatology is

counterproductive[1]. Additional open questions that remain to be investigated include whether and to what extent the application of more sophisticated statistical postprocessing algorithms applied on both the control and ensemble forecast systems can bring the performance of the control system closer to that of the ensemble.

c. Implications for weather forecasting

The role of forecasters is to provide all relevant information on future weather to the users. As the results and discussion above indicate, it is critical that the users have access to multiple–value probabilistic information that captures day–to–day variations in the expected uncertainty of the forecasts. A weather forecast is in fact not complete unless it is expressed in the form of probability distributions. And in the case of appreciable uncertainty, the goal of weather forecasting, including statistical postprocessing, such as MOS and other methods, should be the provision of a detailed case dependent probability distribution (Murphy 1977), and not only a best estimate of the state of the atmosphere. Such information facilitates the use, and increases the potential economic value of weather forecasts. It is not surprising that companies selling weather derivatives[2] are among the core users of ensemble forecasts.

The users in turn can take this information, along with other factors, into consideration when making their decisions related to operations that are sensitive to the weather (Pielke, 1999). Many of the users who could potentially benefit from ensemble forecasts may be unaware of this, because of their possible negative experience with weather guidance based on a single control forecast. This is well demonstrated by the relatively narrow C/L_p range in which users gain value from using the control forecasts

1. Note that if the mean of the ensemble is used as a best estimate of the future state of the atmosphere it can be further improved in an rms error sense by some additional smoothing (see Leith, 1974; Houtekamer and Derome, 1995).

2. Weather derivatives are insurance–type policies that pay the client if certain agreed weather events occur. The premium depends on the expected forecast uncertainty. Unlike normal insurance policies, derivitives can be sold on by the original client. The cost of the transaction may well differ from the initial premium as updated forecast information becomes available.

compared to using the ensemble (see Figs. 1–4), and also by the results of Smith et al. (2001). These potential users may not realize, until they are introduced to probabilistic forecasting, that the relatively low *average* hit rate of certain weather forecasts is not an obstacle to their usage, especially if reliable forecast probabilities show *variations* from case to case.

Initially, some users may feel uncomfortable with the notion of ”probabilities”, thinking they need to make decisions and for that they need a ”yes or no” forecast. The idea behind the cost–loss analysis discussed above is that if reliable probabilistic forecasts are available, each user can choose, depending on their estimated or known cost–loss ratio, a different criterion (probability level) for making their own ”yes–no” decision. After all, weather forecasters are for making weather forecasts, and decision makers are for making decisions (Murphy, 1978). If the forecaster conveys all available information, the weather forecast, for example, will be no longer ”yes, it will rain”, but rather, ”there is an 80% chance of rain”. Well trained users with cost–loss ratios 0.8 and below will interpret this forecast as ”yes”, while those with ratios above 0.8 as ”no”. We know that each weather forecast has an associated case dependent uncertainty, and that this uncertainty can generally be quantified by an ensemble of forecasts (Toth et al. 2001); it is in the users’ best interest to seek and utilize this information.

d. An example

As an example, let us consider the use of minimum temperature forecasts by two farmers in the same geographical area who grow different crops that are all sensitive to freezing temperature that climatologically occurs 20% of the time (o=0.2). Let us assume that the cost of protecting their crops is the same but their potential loss differs dramatically due to differences in the vulnerability and value of their crops. The farmer with less to lose (C/L_p=0.9, high cost–loss ratio) will only spend on protection if the frost is almost a certainty (p=0.9, or higher forecast probabilities), whereas the farmer who can suffer large losses

(C/L_p=0.05) will want to take protective action even if the forecast probability values are low (p=0.05, or higher). Note that in this example the farmers translate the probabilistic weather forecast into their ”protect – do not protect”, yes–no decision, using very different decision criteria (high vs. low probability values).

If a forecaster provides only his/her best estimate on whether the minimum temperature will be above or below freezing, this forecast will likely be useless for either farmer (cf. Fig. 2). Such a forecast, with an intermediate average hit rate of say 80% and missing rate (m/(m+c)) of 10%, will be useful only for users with intermediate cost–loss ratios. Neither the low, nor the high cost–loss ratio customer can benefit from such a product. Instead, they will tend to use climatological information and the former farmer will *always*, while the latter *never* protect his/her crop. To be of any use for them, the forecasts would need to be issued in the form of multiple probability values including their C/L_p values of 0.05 and 0.9. To achieve that, one needs information on case dependent uncertainty (instead of average uncertainty associated with all unclassified “yes” forecasts). As discussed earlier, such guidance can be readily derived from an ensemble of forecasts, which in turn can lead to substantial savings for the farmer with low (high) cost–loss ratio by identifying those cases when he/she can forgo (implement) protection.

6. Conclusions

An economic value analysis based on a simple cost–loss model was carried out on minimally postprocessed 500 hPa geopotential height model output from low resolution ensemble and low and high resolution control forecasts. The analysis revealed that a wider range of potential users can benefit from the ensemble than from the control forecasts, compared to relying simply on climatological information. Moreover, for most users the ensemble offers more economic value than the control forecasts. Similar results were obtained by Richardson (2000a) and Mylne (1999), who studied the economic benefit of

precipitation and temperature forecasts at ECMWF, and surface wind speed forecasts at the Met Office (Bracknell, UK) respectively.

The economic value and ROC results presented in Figs. 1–5 clearly demonstrate the benefit of detailed probabilistic forecasts generated by an ensemble over categorical forecasts based on a single control integration, even if that single forecast is made at a resolution higher than that of the ensemble. For economic decision making it is imperative to use forecasts that provide multiple decision levels. Ensemble forecast systems naturally offer such guidance. The present paper has focused on the evaluation of direct output from the ensemble and from the control forecasts, without attempting to assess the potential benefits of advanced statistical postprocessing of either the ensemble–based probabilities or of the control forecast. Obviously, through statistical postprocessing control forecasts can be supplemented by detailed probability information. Atger (1999), Smith et al. (2001), and Talagrand and Candille (1999, personal communication) studied such forecast systems and found that ensemble forecasts, at least from 4 day lead time on, outperform them. These results suggest that the ensemble can provide some genuine and useful information, likely by the identification of day–to–day variations in forecast uncertainty, that will be difficult to reproduce by simply post–processing a control forecast.

A detailed analysis of ROC results (Figs. 5–7) indicate that beyond 4 days lead time the lower horizontal resolution T62 ensemble, generated at a similar computational cost, outperforms the high resolution T126 control forecast in every respect, suggesting that at least in this lead time range the ensemble forecast system is more cost effective. We conclude that the use of ensemble–based probabilistic forecasts has the potential to substantially increase the overall economic benefit weather predictions can deliver to society.

7. Acknowledgements

Roger Pielke of the University of Colorado, Peter Caplan of NCEP, Anthony Barnston (of IRI), John Jacobson (formerly with SAIC/GSC), and three anonymous reviewers provided helpful comments on earlier versions of this manuscript. We acknowledge the support and encouragement of Stephen Lord, Director of EMC.

8. References

Atger, F., 1999: The skill of ensemble prediction systems. MWR, 127, 1941–1953.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. Nonlinear Processes in Geophysics, under review.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. Wea. Forecasting, 4, 401–412.

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

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123,** 2181–2196.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124,** 1225–1242.

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

Kobayashi, C., K. Yoshimatsu, S. Maeda, and K. Takano, 1996: Dynamical one–month forecasting at JMA. Preprints of the 11th AMS Conference on Numerical Weather Prediction, Aug. 19–23, 1996, Norfolk, Virginia, 13–14.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Mason, I. B., 1982: A model for the assessment of weather forecasts. *Australian Meteorological Magazine,* **30**, 291–303.

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

Murphy, A. H., 1966: A note on the utility of probabilistic predictions and the probability score in the cost–loss ratio decision situation. *J. Appl. Meteor.*, **5**, 534–537.

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., 1978: Hedging and the mode of expression of weather forecasts. *Bull. Amer. Meteorol. Soc*., **59**, 371–373.

Murphy, A. H., 1985: Decision making and the value of forecasts in a generalized model of the cost–loss ratio situation. *Mon. Wea. Rev.*, **113**, 362–369.

Murphy, A. H., 1986: Comparative evaluation of categorical and probabilistic forecasts: Two alternatives to the traditional approach. *Mon. Wea. Rev.*, **114**, 245–249.

Mylne, K.R., 1999 The use of forecast value calculations for optimal decision making using probability forecasts. Preprints of the 17th AMS Conference on Weather Analysis and Forecasting, 13–17 September 1999, Denver, Colorado, 235–239.

Mylne K.R. 2001, ”Decision–Making from Probability Forecasts using Calculations of Forecast Value”. Forecasting Research Tech Note No 335, Met Office, Bracknell, UK (submitted to Met Apps, Jan 2001).

Pielke, Jr., R. A., 1999: Who Decides? Forecasts and Responsibilities in the 1997 Red River Floods. *Applied Behavioral Science Review,* **7,** 83–101.

Rennick, M. A., 1995: The ensemble forecast system (EFS). Models Department Technical Note 2–95, Fleet Numerical Meteorology and Oceanography Center. p. 19. [Available from: Models Department, FLENUMMETOCCEN, 7 Grace Hopper Ave.,Monterey, CA 93943.]

Richardson, D. S., 2000a: Skill and economic value of the ECMWF ensemble prediction system, Q. J. R.Meteorol. Soc., 126, 649–668.

Richardson, D. S., 2000b: Applications of cost–loss models. Proceedings of the Seventh ECMWF Workshop on Meteorological Operational Systems. November 15–19, 1999, Reading, England, 209–213.

Richardson, D. S, 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. Q. J. R. Meteorol. Soc., 127, in press.

Smith, L. A., M. S. Roulston, and J. Hardenberg, 2001: End to End Ensemble Forecasting: Towards Evaluating the Economic Value of the Ensemble Prediction System. ECMWF Technical Memorandum No. 336.

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

Stewart, T. R., 1997: Forecast value: descriptive decision studies. In: Economic value of weather and climate forecasts. Ed. by R. W. Katz and A. H. Murphy, Cambridge University Press, 147–181.

Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. Proceedings of ECMWF Workshop on Predictability, 20–22 October 1997, 1–25.

Toth, Z., and E. Kalnay, 1993: Ensemble Forecasting at the NMC: The generation of perturbations. Bull. Amer. Meteorol. Soc., **74**, 2317–2330.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev*, **125,** 3297–3319.

Toth, Z., Y. Zhu, T. Marchok, S. Tracton, and E. Kalnay, 1998: Verification of the NCEP global ensemble forecasts. Preprints of the 12th Conference on Numerical Weather Prediction, 11–16 January 1998, Phoenix, Arizona, 286–289.

Toth, Z., Y. Zhu, and T. Marchok, 2001: On the ability of ensembles to distinguish between forecasts with small and large uncertainty. *Weather and Forecasting*, **16**, 436–477.

Wilks, D.S., 2001: A skill score based on economic value for probability forecasts. *Meteorological Applications*, 8:209–219.

Zhu, Y, G. lyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints of the 15th AMS Conference on Weather Analysis and Forecasting, 19–23 August 1996, Norfolk, Virginia, p. J79–J82.

Ziehmann, C., 2001: Skill prediction of local weather forecasts based on the ECMWF ensemble. Nonlinear Processes in Geophysics, in print.