Predictability from a Forecast Provider's Perspective

Ken Mylne

Met Office, Bracknell RG12 2SZ, UK.
email: ken.mylne@metoffice.com

1. Introduction

Predictability is not a new issue for forecasters or forecast providers. Forecasters have always dealt with uncertainty, usually describing it subjectively with terms such as "mainly in the NW" and "up to an inch in places". Many forecasters' daily jobs involve providing bespoke services to individual customers, and by understanding those customers' businesses the forecasters are able to provide them with information on some of the risks and uncertainties which will impinge on their activities and affect their decision-making. What has been changing in recent years is the ability of forecast provider organisations, such as the Met Office, to assess the uncertainties more quantitatively. Forecast services are increasingly provided automatically in order to minimise costs and delays and allow flexible production of forecasts for many sites. It is therefore necessary to find ways of expressing uncertainty automatically and in a way which is meaningful and useful to customers. By being quantitative, usually as probabilities, we can also offer significantly better services to our customers, as long as the figures are meaningful. We also need to work with the end users to help them understand what the numbers mean and how to make use of them in their decision-making.

This paper will discuss the use of probabilities in providing forecast services to customers, and describe some of the ways that ensembles are used in the Met Office to support and improve our services.

2. Predictability - what can we predict?

For a forecast provider like the Met Office, predictability is about balancing customer desires for certainty, with what we can and cannot predict. Customers would like certainty to ease decision-making but this is frequently impossible due to chaos and processes we cannot resolve. So what can we predict? A good starting point is usually climatology:

- Past statistics tell us the climatological probability of an event

  e.g. Snow falls on 17 out of every 100 January days ⇒ daily prob of snow in January = 17%.

Assuming the climatology is static and representative, this provides a perfectly reliable probability forecast and to be useful, any forecast system must improve on this. Conversely, where we don't know better we should issue climatology as the best available guidance to a customer. An example of where this might be done is a long-range forecast, issued ahead of the time when we believe we have predictive skill. For example, an insurance company providing cover against weather disruption will assess their risks and set premiums based on climatology.

As well as setting a baseline for probability forecasts, climatology is also useful in interpreting probabilities. A common criticism of probability forecasts is that forecasters are simply "covering themselves" or "don't know", particularly when a mid-range probability such as 50% is issued. However a forecast of 50% can be extremely informative - consider the following forecast issued in November: "There's a 50% prob of snow in London tomorrow." While not impossible, climatology tells us that snow in London is rare in November, so a 50%

probability for the next day is indicating a very high risk compared to normal. This forecast therefore contains a strong signal, even though the forecaster could quite honestly say "I do not know if it will snow in London tomorrow", and it is important to present that signal clearly.

Thus climatology provides a baseline for predictability, and any forecast should be an attempt to refine the probability to give more information. Figure 1 illustrates forecasts of a parameter *x* which has a climatological distribution shown in green. There are several standard ways to generate forecasts of *x* based on numerical weather prediction (NWP). A deterministic forecast is the outcome of a single run of an NWP model, and gives a single solution as shown, but which will normally be in error. Over a number of previous forecasts it is possible to generate statistics of the errors of deterministic forecasts, and from these the deterministic forecast may be supplemented by an error distribution function as shown in red, providing a simple estimate of the forecast probability density function (PDF). In this case this is illustrated with a Gaussian distribution which may be generated from knowing just the standard deviation of the errors, but the method may also be applied using different forms. For example a Gamma distribution may be more appropriate for parameters such as rainfall which typically have a skewed climatological distribution (Wilks, 1995). Whatever distribution is used, the form of it is fixed and does not depend on either the meteorological situation or the value of the deterministic forecast to which it is applied - hence on occasions where the deterministic forecast is extreme (and therefore of particular interest) the PDF from the error distribution is least likely to be representative of the true forecast probabilities.

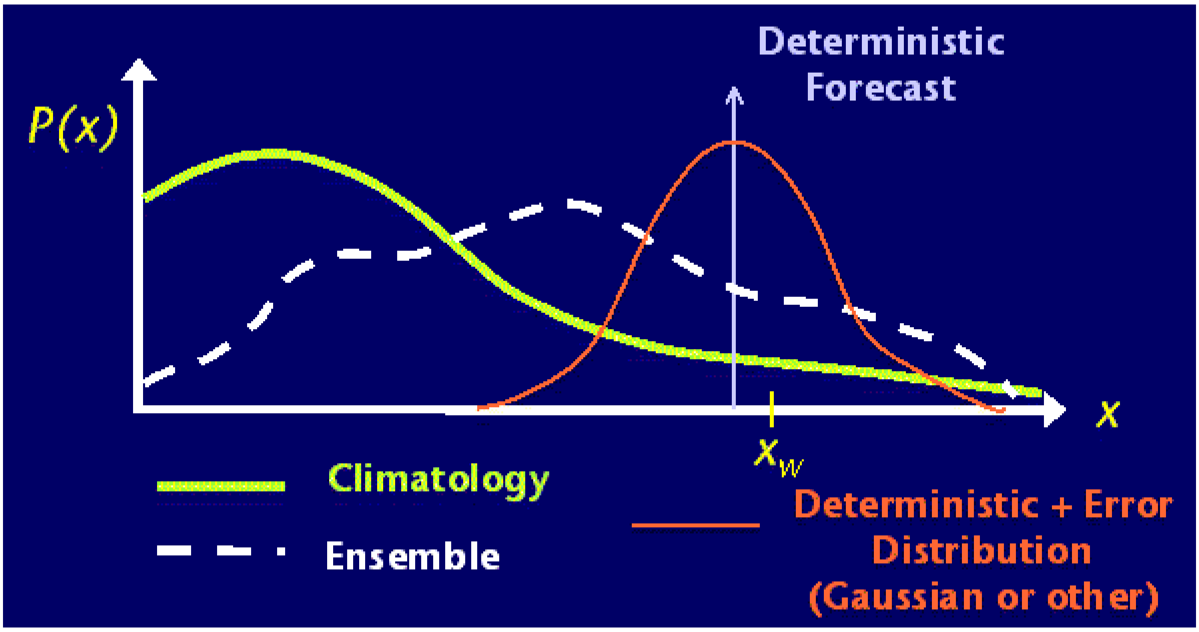

Figure 1: Illustrating how forecast information can refine the climatological distribution function. The deterministic forecast provides a single solution. The simplest way to provide probability information is to add an error distribution to a deterministic forecast which may be a gaussian as shown, or may follow some other fitted or observed distribution. An ensemble forecast provides a case-dependent probability distribution taking account of the meteorological information available to the ensemble. x_w shows the value of a hypothetical warning threshold of x.

An ensemble forecast attempts to sample the forecast PDF taking account of the current meteorological situation, and thus the current predictability state of the atmosphere, and may generate a complex form as illustrated, including distinctly different values of *x,* all generated from plausible meteorological forecasts. Each of these methods provides increasingly sophisticated refinements of the climatological distribution, from which forecasts of x may be expressed as probabilities which are different from that provided by climatology alone. For example, the distributions in figure 1 could be used to generate probabilities of exceeding the warning threshold x_w. The probability of exceeding x_w is given by the area under a PDF to the right of x_w. Hence, climatology gives a low probability in the tail of the distribution; the deterministic forecast alone would give a probability of zero, but when enhanced by the gaussian it suggests a probability around 40%; the ensemble suggests a probability distinctly greater than climatology but still within the tail of the distribution. While the ensemble method is clearly the most expensive way to generate probability forecasts, its flow-dependent nature means that for most applications it is likely to give better results than other methods. The error distribution method can provide some useful guidance, but since it is not flow dependent it may be seriously in error on some occasions, perhaps the most important ones.

3. Quality Measures of Probability Forecasts

Describing uncertainty quantitatively is only beneficial if the numbers can be shown to be meaningful. What does 30% probability mean? If the forecast probability of exceeding x_w is 30%, and x_w is indeed exceeded, this neither makes the forecast right nor wrong. But out of 100 independent forecasts of 30%, x_w should be exceeded 30 times. If this is the case the forecasts are said to be perfectly reliable. A reliability diagram plots the frequency of occurrence of an event against the forecast probability. Verification must be done over many forecasts. Reliability is not sufficient on its own for a useful probability forecast. Climatology provides perfectly reliable probabilities, but contains no occasion-specific forecast information. Useful forecasts also need resolution, which measures how much the forecasts deviate from climatology, and they need discrimination which indicates the ability of the forecast system to distinguish between occasions when an event does occur from ones when it does not. Discrimination is shown by the slope on a reliability diagram - if the graph is horizontal, the probability of the event occurring is independent of the forecast probability, so the forecasts are useless.

As well as being numerically meaningful, it is also important that forecasts are unambiguous and are relevant to the user application - both provider and customer must be clear exactly what the probability refers to. For example, if a forecast states there is "a 30% probability of rain in England", does this mean 30% at any one place, or 30% “somewhere in England”? Is this a 30% risk of a trace being recorded, or of a downpour? It must be clearly stated exactly what is being predicted.

4. Use of Ensemble Forecasts at the Met Office

4.1 Long-range forecasting

In long-range (monthly and seasonal) forecasts, predictability is inherently low and forecast systems simply aim to skew the climatological distribution slightly in the right direction. A 9-member ensemble is forced by current and expected sea-surface temperature anomalies to estimate the expected effect on mean behaviour of the atmosphere compared to normal. When run over extended periods the climate of the model(s) may differ significantly from the climate of the real atmosphere. Model climatology is determined by running the model for many past seasons. Forecasts are then expressed as anomalies relative to climatology, and may be interpreted or calibrated by reference to real climatology. Probabilities issued are frequently close to climatological probabilities, showing only small amounts of resolution due to low predictability, but verification shows that there is some discrimination and therefore the forecasts have value to some users who can adjust their actions in response to small changes in probability. The Met Office issues seasonal forecasts on its web-site and monthly forecasts are provided commercially to a number of customers. Long-range forecasting is discussed in more detail by other speakers in the Seminar.

4.2 Medium-range forecasting

Medium-range forecasting (3-10 days) has been transformed over the past 10 years by the availability of the ECMWF EPS (Ensemble Prediction System). Prior to that forecasts were deterministic, based heavily on the Met Office's global model, with statements about confidence based on the agreement or otherwise of a few other models. Today the main products are still largely deterministic, including isobaric charts with frontal systems, as these products are extremely popular with customers, but they are now based on what is perceived to be the most probable solution from the EPS. Forecasters have access to a wide range of tools for visualisation of the EPS, and use field modification software (Carroll, 1997) to produce meteorologically consistent fields representing the most probable outcome. Figure 2 shows results from subjective verification, assessing the quality of forecast charts, which shows that the modified forecast fields (MOD) perform better than any of the individual models available to the forecasters.

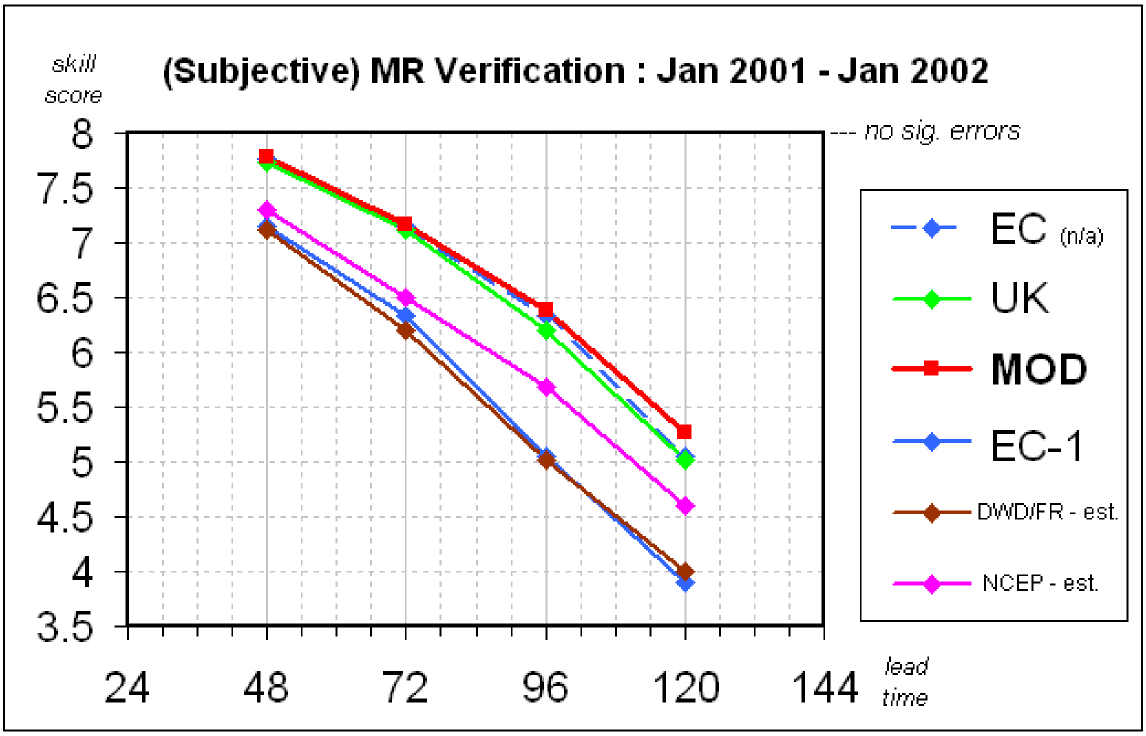

Figure 2: Results of subjective verification performed between January 2001 and January 2002 comparing the quality of charts from various models with the modified charts (MOD) generated by forecasters in the National Meteorological Centre (NMC). Note that the latest ECMWF model output (EC) is not available early enough to be used in producing the MOD charts.

In addition to generating the most-probable chart solution, medium-range guidance forecasters are also able to generate alternative solutions where the EPS suggests a different solution with a probability of occurrence of more than about 20%. Such alternatives are normally based on clustering of ensemble members, and probabilities from the numbers of members in different clusters. All chart products are further supplemented with detailed discussion of the confidence and risks indicated by the ensemble.

4.3 Forecasting Severe Weather

Much emphasis is now being put on improving predictions of severe weather. Since the development of severe weather is frequently highly non-linear, this is an appropriate application of ensembles; at the same time it is a particularly demanding application, and is also difficult to verify since severe weather occurs relatively rarely so data samples are small. For long-range forecasting we noted that model climatology is often significantly different from real climatology - the same is true in the medium or short range when considering severe weather, since many severe-weather developments depend on quite small-scale processes which are not fully resolved. It is therefore often necessary again to calibrate forecasts relative to model climatology rather than interpreting model output directly - Francois Lalaurette will discuss this more in his lecture.

Over recent years the Met Office has attempted to use the EPS to generate early warnings of severe weather in support of the UK National Severe Weather Warning Service (NSWWS). Early warnings can be issued up to 5 days in advance when the probability of an event occurring "somewhere in the UK" is 60% or more. In addition to an overall UK probability, probabilities are also given for 12 local regions. In practice forecasters only rarely issue warnings more than 36h in advance, so the EPS First-Guess Early Warning (FGEW) system was designed to provide forecasters with consistent and verified objective probabilities in order to encourage earlier issue and reduce the overall Miss Rate.

FGEW warnings provide a good example of the need to calculate the relevant probability for the application. At 3-5 days ahead the precise timing of severe weather does not matter so the probability calculation looks inside a time-window. Similarly an ensemble member counts towards the probability if it generates severe weather at any grid-point in the UK (or a sub-region). Thus the probabilities are much higher than those seen at fixed times at individual grid-points.

The 60% probability threshold defined for the issue of Early Warnings reflects customer desire for high confidence, but in practice this is rarely attained. Figure 3 illustrates schematically that in a synoptic situation when severe weather is possible, once a forecast moves into the chaotic non-linear regime, most ensemble members are likely to be drawn towards the model's climatology. (Although the diagram illustrates this idea with the central control forecast predicting severe weather and perturbed analyses leading to less severe conditions, this argument is just as true when it is one or more perturbed ensemble members which predict severe conditions.) The result of this is that the forecast PDF is always likely to be skewed away from severe weather, and although the ensemble can be expected to include members with severe events, it would be unusual for it to predict high probabilities of severe weather. This indeed turns out to be the case in practice.

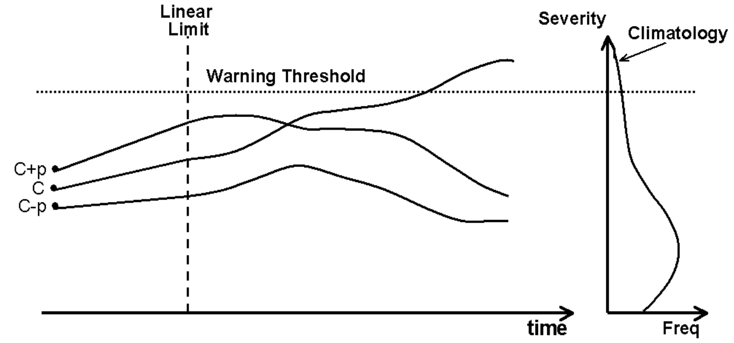

Figure 3: Schematic illustration of the effect of non-linearity on an ensemble forecast. In the early stage of a forecast, ensemble members diverge quasi-linearly. In later stages, even when one member predicts severe weather, most members can be expected to be drawn towards model climatology.

Figures 4 and 5 show examples of verification of 4-day forecasts from the FGEW system which illustrate that most of the forecast information is contained in low-probability forecasts. Figure 4 shows reliability diagrams for early warnings of heavy rain from several different test versions of the system, shown in different colours, with corresponding histograms of the number of times each forecast probability was issued. Apart from the pale blue version, which was poorly calibrated and over-predicting severe weather, the histograms show that forecasts were rarely issued with higher probabilities, particularly above 40%. (Note that because the events are rare the probability bins used have been concentrated towards the low probability end, and that most forecasts give probabilities below 10%.) As a result of the small samples, the reliability diagrams are very noisy at higher probabilities, but they do show the right general trend with severe weather increasingly more likely to occur when higher probabilities are issued. For the green, yellow and orange curves, which use the operational system calibration from different runs of the EPS (12 UTC, 00 UTC and combined respectively) the overall bias is small with mean forecast probabilities close to the overall sample frequency of 0.032. It was noted above that for severe weather it is important to calibrate the forecasts relative to model climatology - this was done for the FGEW system by optimising this overall bias and the forecast reliability over the winter 2000/01; verification results shown here are taken from the following winter 2001/02 and confirm that the calibration was quite successful.

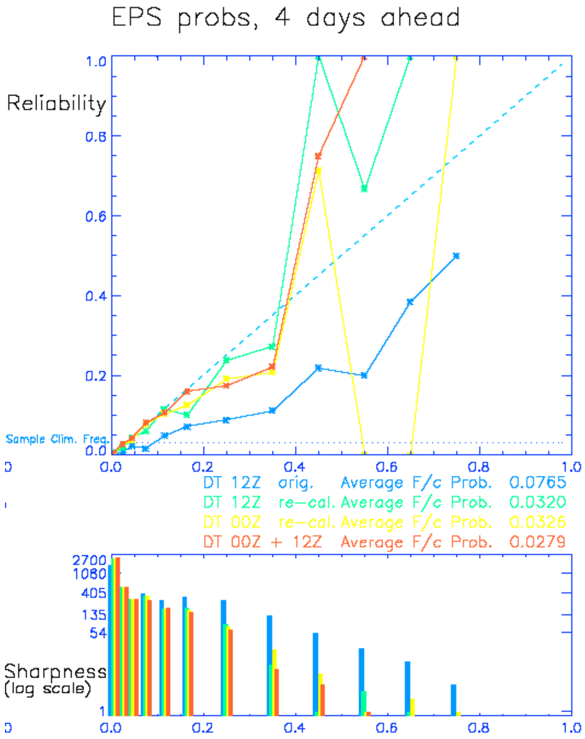

Figure 4: Reliability diagrams for 4-day forecasts of heavy rain from different test versions of the First-Guess Early Warnings system (different colours) verified between July 2001 and May 2002. The Sharpness diagram underneath indicates the number of times each forecast probability was issued.

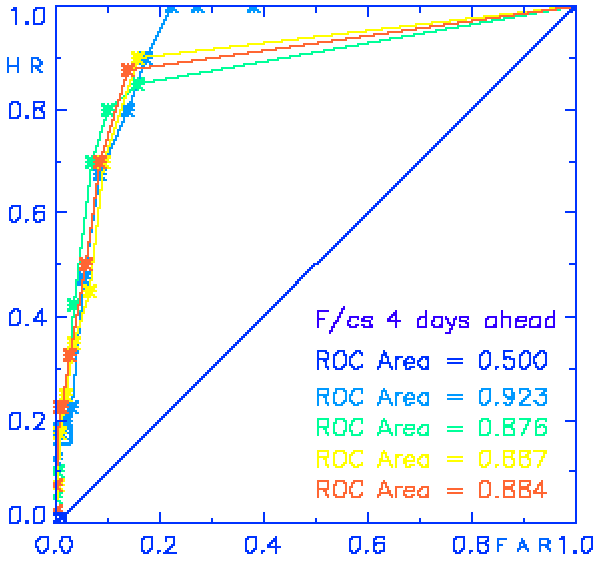

Figure 5: ROC curves, and areas under the curves, for 4-day forecasts of severe gales from different test versions of the First-Guess Early Warnings system (different colours) verified between July 2001 and May 2002.

The ROC (Relative Operating Characteristics - see Stanski *et al,* 1989) curves shown in figure 5 for 4-day severe gale warnings show that the system has considerable ability to discriminate occasions when gales are more likely to occur. (ROC points lying on the main diagonal represent "no-skill"; points above the line show discrimination ability.) However, in a ROC curve the points nearest the top right of the graph give hit rates and false alarm rates corresponding to the lowest probability thresholds, so most of the discrimination ability is due to low probability forecasts. Note that although quite high hit rates (vertical axis) can be achieved for these low probability forecasts, this is at the cost of large numbers of false alarms - although the false alarm rates plotted (horizontal axis) look quite small, because they are expressed as a fraction of all non-events they can still represent quite large numbers of occasions. Thus the First-Guess system is able to provide some reliable probabilistic information on the likelihood of severe weather, but only on rare occasions is it able to provide the high probabilities that most customers would require before taking any protective action.

One notable result from the FGEW system is that the EPS forecasts at 4 days ahead were better than those at shorter range. For day 2 and 3 forecasts (not shown) the ROC curves were less bowed towards the top left corner, and the reliability diagrams showed little increase in probability of occurrence when forecast probabilities were high. This is believed to be due to the nature of the singular vector (SV) perturbations used in the EPS which are optimised for maximum growth over the first 2 days of the forecast. Over this period the perturbations are attempting to identify the maximum ensemble growth and so might be expected to successfully identify any possible extreme weather developments, but the results show that the resulting probabilities of severe weather are very poor. To generate reliable probabilities requires a random sampling of the PDF, and it may be that the SV strategy, by focussing on maximum growth, is not sufficiently random in the early stages of the forecast. Beyond the 48h optimisation period, after the perturbations have undergone significant non-linear growth, the sampling may be more effectively random, thus resulting in more reliable probability forecasts. Brier skill scores (Wilks, 1995) (not shown) show that the overall probabilistic skill of the EPS early warnings increases with increasing lead-time out to 4 days, and with some configurations of the system out to 6 days.

The FGEW system illustrates the difficulty of meeting customer requirements for high-confidence forecasts where predictability is low. While the system has some considerable skill in identifying the possibility of severe weather, and some ability to produce unbiased, reliable probabilities, the fundamental low predictability of the severe weather illustrated in figure 3 means that on most occasions warning can only be given at low probabilities. Users need to learn how to make use of such warnings, and I will return to this later.

4.4 Site-specific Forecasts

Most weather forecast customers require site-specific forecasts for their particular locations, so the Met Office has invested significant effort in extracting site-specific weather parameters from each EPS member to allow the generation of probability forecasts. Around 400 sites are now available in a database for product generation, and several graphical tools are available to display forecasts for customers. Figure 6 shows an example of a Stacked Probability Chart generated from the marine wave model within the EPS, showing the risks of exceeding various significant wave-height thresholds at a site in the North Sea - this chart, generated routinely for use by offshore oil industry customers, is designed for risk assessment and is ideal for the identification of "weather windows" in which work can be carried out.

Interpolating weather parameters directly from NWP model grids to specific locations is subject to large errors as the NWP model cannot resolve the sub-grid-scale features which are important in generating the micro-climate of the real site. The NWP model can only attempt to represent weather parameters on some sort of grid-box average, and the true resolution of a model is around 4-5 grid-lengths. Consequently the Met Office applies a multi-variate Kalman filter to relate interpolated model field values to observed weather parameters statistically. Parameters such as temperature are related to model temperatures, but also wind direction which is particularly important in coastal locations, for example, so the Kalman filter provides more than just a simple bias correction. Use of the Kalman filter also allows the derivation of parameters which may not be available directly from the model, but which are available from site observations and which are required by customers, such as maximum and minimum temperatures.

Verification of probability forecasts from ensembles normally shows that the ensemble does not spread sufficiently to cover the full uncertainty, and this is particularly true when considering site-specific forecasts. By removing most of the site-specific biases, the Kalman filter greatly improves the coverage of the uncertainty at local sites, but spread is still not sufficient. A second post-processing is therefore also applied to increase the spread of the site-specific forecasts based on past verification (see Mylne *et al,* 2002). Routine verification of operational forecasts is important to demonstrate the skills of forecast production systems to customers, and the site-specific probability forecasts are verified continuously. Figure 7 shows an example of a reliability diagram for forecasts of wind exceeding Beaufort Force 6 at 3 days ahead. The green line is for winds interpolated directly from model fields, and shows that the wind is significantly over-forecast as forecast probabilities are consistently too high. This bias is very largely corrected by the Kalman filter (red) but the forecasts are still over-confident, shown by the fact that the slope of the graph is less than the ideal 45 degrees. The two blue curves show two versions of the final calibration, and lie very close to the ideal diagonal, indicating that the calibration is successfully improving the reliability of the forecasts.

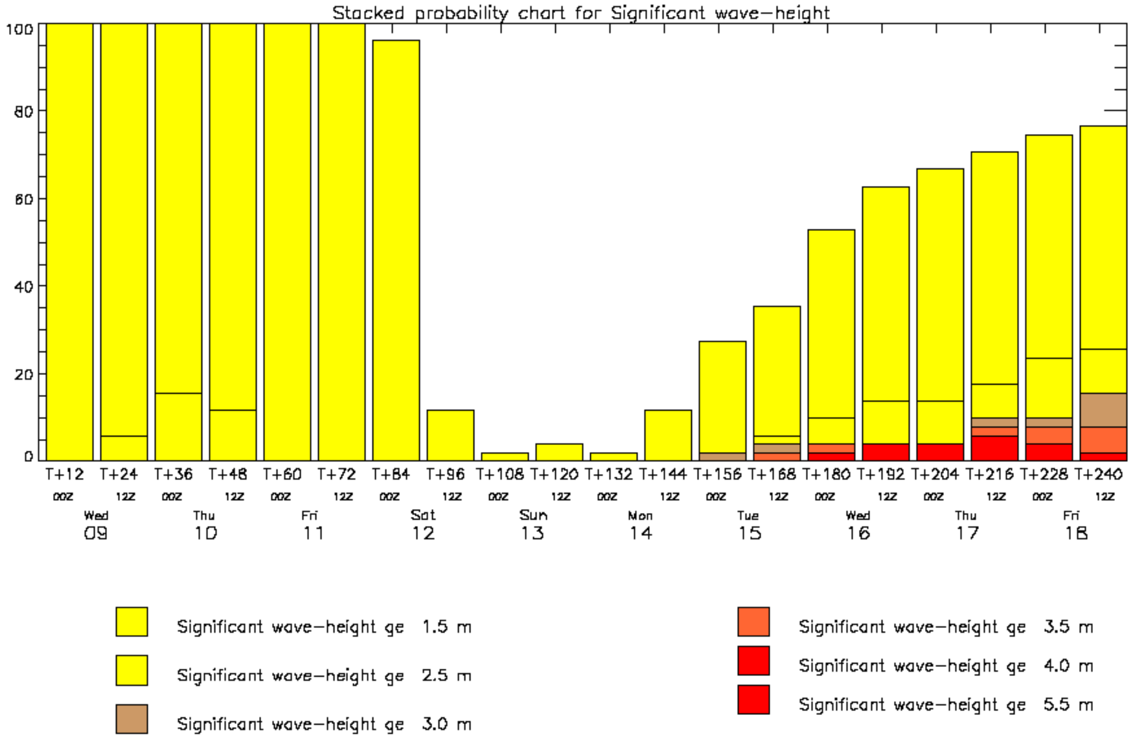

Figure 6: Stacked Probability chart showing probabilities of exceeding various wave-height thresholds.

5. Short-Range Predictability

So far we have been discussing medium and long-range prediction, but most forecast customers are primarily interested in the short-range (1-2 days). At this range NWP is deterministic, and forecasts have improved steadily due to increased resolution, improved model formulation and data assimilation, and better use of observations, but there are still many uncertainties in the forecasts issued. Large synoptic-scale errors are rare but typically involve rapid cyclogenesis and are therefore critically important. Much more common are errors in sub-synoptic details such as frontal waves, QPF (Quantitative Precipitation Forecasting), convection and more detailed weather parameters of importance to customers like cloud height and visibility. Uncertainty in the short-range detail is still assessed subjectively by forecasters with few objective tools to help them, but research is now starting into whether these issues can be addressed with ensembles. A few centres such as

NCEP are already experimenting with short-range ensembles; the Met Office is currently developing plans for an ensemble to be based on a Limited Area Model (LAM) covering the Atlantic and Europe with a horizontal resolution of around 20km. This resolution will only start to address some of the uncertainty issues, but ultimately it is hoped to run ensembles over the UK at very high resolution to enable probabilistic prediction of detailed weather parameters such as thunderstorms. Adequate computing resources for this will not be available before 2008 at the earliest.

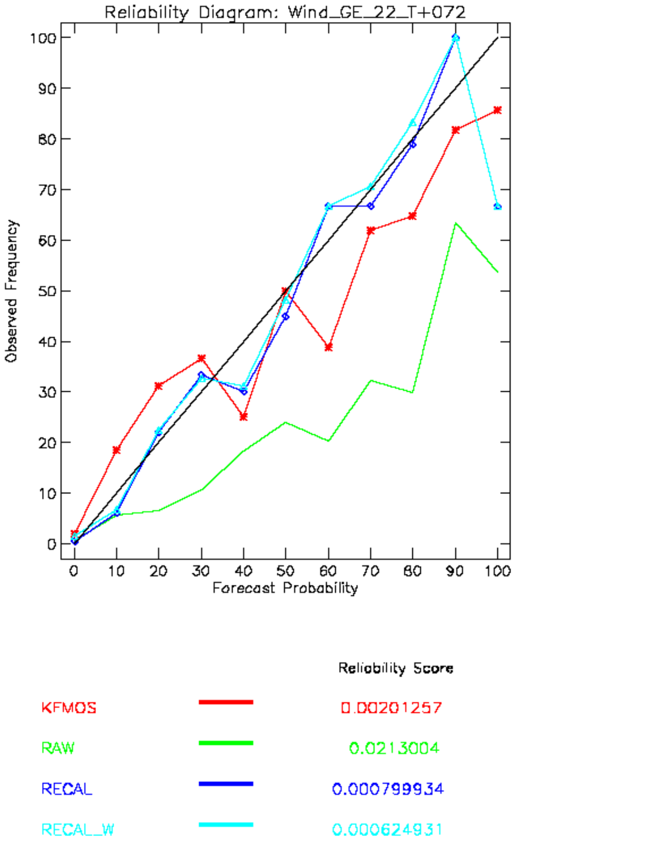

Figure 7: Reliability diagram for forecasts of wind speed exceeding 22kt (Beaufort Force 6) at T+72 (3 days) at sites in the UK over winter 2000/01. The different coloured lines are explained in the text. Figures under the graph give the reliability score for each curve (see Wilks, 1995).

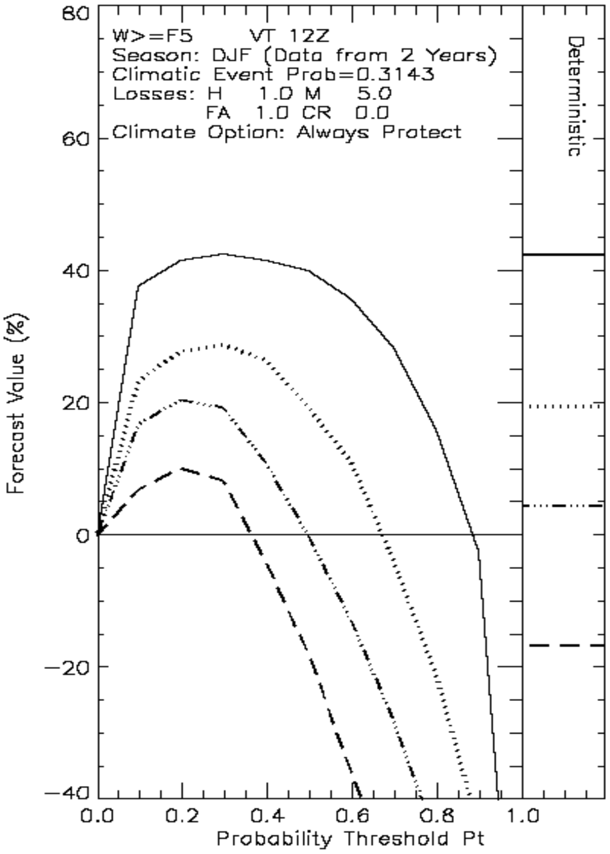

Figure 8: Economic value of uncalibrated probability forecasts plotted as a function of a probability decision threshold p_t *for a customer with a cost-loss ration* C/L=0.2. *(This example is based on forecasts of wind speed exceeding Beaufort Force 5 interpolated directly from the EPS with no post-processing.) Different lines are for different forecast lead-times: 48h (solid), 96h (dotted), 144h (dot-dash) and 192h (dashed).*

6. Real-World Customers and Probabilities

As mentioned earlier, customers tend to want near-deterministic forecasts, or at least high confidence on which to base decisions. Limited predictability of many parameters means that this is often not possible, and the most informative products we can provide are probabilities such as 2%, 50%, 80%. To balance the requirements and capabilities we need to ask what customers really need, and the answer is normally decisions. So how can we help them make decisions from probabilities? In another paper in this Seminar, David Richardson describes Economic Value for assessing forecasts, and the same approach can be used to guide decision-making with probability forecasts. By working with a customer to analyse their losses *L* associated with a weather event, and their costs *C* of protecting against that event, we can identify the cost-loss ratio *C/L*. Given this, the user's best strategy is to protect against the event whenever the probability is greater than *C/L* - averaged over many occasions, and provided the forecast probabilities are reliable, this strategy will maximise savings. Even if forecasts are not perfectly reliable, analysis of past forecast performance can allow us to identify the optimal decision threshold for a particular customer. Figure 8 presents an example of forecast value for a particular customer with *C/L=0.2* plotted against a decision threshold p_t. These curves are based on verification of uncalibrated probability forecasts interpolated directly from the EPS with no post-processing. For the user the best decision threshold is the p_t which maximises the value, so for 48h or 96h forecasts (solid and dotted lines) the user's best strategy is to protect against the weather when the probability exceeds 30%. At first sight one would expect the customer's optimum threshold to be 20%, because *C/L=0.2*, but because the forecasts are imperfectly calibrated they actually do better using 30%. (For further details see Mylne, 2001.)

In practice, real-world decision-making is usually much more complex, and more sophisticated decision tools are required, but methods such as this point the way to how forecast providers can work with their customers to maximise the benefit from forecasts where predictability is low. With increasing automation of forecast products, providers like the Met Office are increasingly working with customers to help them optimise decision-making, rather than simply providing a best-guess weather forecast. The use of probability products is a key part of this optimisation. After several years of introducing the ideas to customers we are now making significant progress in several sectors, notably severe weather warnings, the offshore oil industry (where potential losses are often massive) and weather derivatives traders who are very used to managing risk and basing decisions on small probabilities. Nevertheless, there are still very few customers who are prepared to take action on the basis of a probability as low as 10%.

7. Conclusions

Predictability is an issue for forecasters and customers on all time-scales, and ensembles are now well-established tools to aid assessment of predictability at long and medium ranges. Ensembles are used to improve the quality of deterministic forecasts by identifying the most probable solutions, and to supplement them with confidence information and alternative solutions. Forecasters are also provided with probabilistic guidance to help with risk assessment of severe weather. Many tools are now available to provide high-quality automatic probability forecasts to customers, who are starting to see the benefits in some sectors. Research is progressing to predictability issues in short-range forecasting where we are still largely dependent on the skills of experienced forecasters.

Acknowledgements

I am grateful to numerous colleagues at the Met Office who have performed much of the work described in this lecture. I am particularly indebted to Caroline Woolcock, Tim Legg, James Denholm-Price (now at Kingston University), Martin Young and Tim Hewson.

References

Carroll, E.B. 1997, A technique for consistent alteration of NWP output fields. *Meteorol Apps* 4, 171-178.

Mylne,K.R., 2001 Decision-Making from Probability Forecasts using Calculations of Forecast Value, Met Office Forecasting Research Tech Note No 335; to appear in *Meteorol Apps*.

Mylne, K.R., Woolcock C., Denholm-Price, J. C.W., and Darvell R.J. 2002: Operational calibrated probability forecasts from the ECMWF Ensemble Prediction System: implementation and verification. Preprints of

Symposium on Observations, Data Assimilation, and Probabilistic Prediction, AMS, 13-17 January 2002, Orlando, Florida, pp113-118.

Stanski,H.R., Wilson,L.J. and Burrows,W.R., 1989: Survey of Common Verification Methods in Meteorology, WMO WWW Tech. Report No 8, WMO TD No 358.

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