Numerical Weather Prediction

Severe weather early warnings from ensemble forecast information

Forecasting Research Technical Report No. 415

Tim Legg and Ken Mylne

email: nwp_publications@metoffice.com

Severe Weather Early Warnings from ensemble forecast information

T P Legg and K R Mylne
Met Office, Bracknell, United Kingdom

16 April 2003

Lead author: Tim Legg, Forecasting Research, Met Office, London Road, Bracknell, Berkshire RG12 2SZ, United Kingdom

Abstract
A system has been developed to give probabilistic warnings of severe-weather events for the UK on a regional and national basis, based on forecast output from the ECMWF Ensemble Prediction System (EPS). The First-Guess Early Warning (FGEW) project aims to give guidance to operational forecasters, to help them give earlier warning of severe weather in support of the UK National Severe Weather Warning Service (NSWWS).

Calibration was applied to the EPS model output to optimise the probabilistic Early Warnings over an initial training period of one winter season, and the resulting warnings then verified over a 15-month period spanning two winter seasons. The skill of warnings from several versions of FGEW is assessed using a range of probabilistic skill scores, and is also compared with that of warnings issued by forecasters. Results show that the system is capable of providing useful warnings with some probabilistic skill. Most of the skill is attributable to warnings issued at low probabilities, but when higher probabilities do occur this provides a valuable signal which has been used by forecasters on a number of occasions to issue warnings earlier than was done previously.

Maximum skill of the FGEW warnings is found at a lead-time of 4 days, with virtually no skill at shorter lead-times of 1 or 2 days. This is believed to be due to the ensemble perturbation strategy applied at ECMWF which is optimised for medium-range forecasting. Different perturbation strategies will be required for future development of short-range ensemble prediction systems.

1. Introduction

Over recent years NWP models have improved to the extent that developments in weather prediction can now be focussed increasingly on severe or hazardous weather. The development of severe weather frequently involves strong non-linear interactions, often between quite small-scale features in the atmosphere. Such interactions are inherently difficult to predict since even small errors in the analysis or timing of such features can lead to large differences in the forecast evolution due to the non-linearity of the processes involved. This uncertainty is similar to that which leads to synoptic-scale uncertainty in medium-range prediction, except that the shorter length-scales and rapid evolution can often lead to greater than normal uncertainty at short lead-times. Medium-range forecasting has improved greatly since the introduction of ensemble prediction systems (EPS) (Mureau *et al*., 1993; Molteni *et al*., 1996; Toth and Kalnay, 1997) in the 1990s. An EPS uses a number of runs of an NWP model, differing by small perturbations to the initial conditions and perhaps the model physics, to estimate the probability distribution of the forecast – it is normally therefore used to generate probability forecasts. The inherent uncertainty of severe weather developments means that ensemble prediction is likely to be beneficial. Severe-weather events are also well-suited to the issue of probability forecasts – forecast producers are often concerned that end users and the general public do not understand probabilities, but interpretation is often considered to be easier when thinking of extreme and rare events. Operational medium-range ensembles have been run by ECMWF (European Centre for Medium-Range Weather Forecasts) and NCEP (US National Centers for Environmental Prediction) since 1992, but to date most applications and verifications have focussed on relatively common and less severe events. This paper describes a first attempt to apply the ECMWF EPS for prediction of severe weather in the UK, and provides a probabilistic verification over an extended run of forecasts.

The Met Office provides a National Severe Weather Warning Service (NSWWS) as part of its Public Meteorological Service responsibilities to national and local government authorities. The NSWWS includes Early Warnings which are given in probabilistic form and may be issued up to 5 days ahead, and Flash Warnings which are issued when severe weather is expected within the next few hours with a high degree of certainty. Historically the probabilities for the Early Warnings have been assessed subjectively by forecasters, and in practice warnings have rarely been issued more than about 1-2 days ahead. The First-Guess Early Warnings (FGEW) project was established to estimate probabilities of severe weather objectively from the ECMWF EPS in exactly the form required for the NSWWS, with the aim of giving forecasters more confidence to issue warnings more frequently and earlier. Since Flash Warnings are issued for the same weather events as Early Warnings, and have been shown to have a very high accuracy when verified against observations (K. Hymas, personal communication), they provide convenient proxy observations for the verification of FGEW warnings. NSWWS definitions of severe weather events used by the FGEW system are given in Table 1.

Table 1. Definitions of some of the severe-weather events in the NSWWS

Event	Definition

Severe gale	Gusts of 70m.p.h. or more
Heavy snowfall	At least 4cm depth of fresh snow falling within a 2-hour period
Blizzard	Moderate or heavy snowfall, with mean wind-speeds of at least 30m.p.h.
Heavy rainfall	At least 15mm of rainfall occurring within a 3-hour period
Prolonged heavy rainfall	At least 25mm of rainfall occurring within a 24-hour perioc
These are not precise definitions; rather, warnings are issued when weather condition are expected to endanger or seriously inconvenience human activity, and hence absolute values of event thresholds may be lower in heavily-populated regions.	

Section 2 of this paper will briefly review the development and application of ensembles, and consider what we might expect of ensembles in prediction of severe or extreme events. Section 3 will describe the predictability of severe weather; Section 4 covers the calculation of probabilities in the form required for the NSWWS, and introduces several variants of the FGEW system which have been tested. Section 5 introduces the probabilistic verification methods used, and explains how these were used to calibrate the system. Sections 6-9 present verification results, comparing the different variants of FGEW, and Section 10 describes the 'cost-loss' approach. Results will be discussed in section 11, with suggestions for future developments; conclusions are summarised in Section 12.

2. Ensemble Prediction for Severe Weather

Operational ensemble prediction systems (EPS) are now well-established for medium-range forecasting, having started at both ECMWF (Molteni *et al.*, 1996) and NCEP (Toth and Kalnay, 1997) in 1992, and also at the Canadian Meteorological Centre (Houtekamer *et al*., 1996) in 1994. Since then the ECMWF EPS, used in this study, has been upgraded (Buizza *et al*., 2000) to run with 51 members (an unperturbed control plus 25 pairs of perturbed members generated by adding and subtracting each perturbation to the analysis) at a resolution of $T_L255L40$ (approximately 80km in mid-latitudes with 40 vertical levels). Initial condition perturbations are calculated as linear combinations of singular vectors (SVs) (Molteni and Palmer, 1993; Mureau *et al*., 1993). These SVs are calculated using a linearised adjoint of the full NWP model with dry physics at T42L40 resolution, but identify a good approximation to the dynamic modes which grow most rapidly over the first 48 hours of the forecast. Initial perturbations also include evolved SVs, calculated 48 hours previously and evolved over that period to provide a better estimate of uncertainty in the early part of the forecast (Buizza *et al*., 2000; Barkmeijer *et al*., 1999). Stochastic physics has been incorporated in an attempt to take some account of model errors as well as initial condition errors (Buizza *et al*., 1999). With these developments the EPS has matured to become the principal tool for medium-range forecasting in most European National Meteorological Services; applications of the EPS at the Met Office are described by Legg *et al*. (2002) and its use by Met Office medium-range forecasters is described by Young and Carroll (2002).

In addition to the standard 51 members of the EPS, ECMWF also run 5 additional "multi-analysis" (MA) members. These use the identical version of the ECMWF model

to the rest of the EPS, but are started from different analyses. Four use the operational analyses of different NWP centres (Met Office, Meteo-France, DWD (Deutsche Wetterdienst) and NCEP (US National Centers for Environmental Prediction)); the fifth is a "consensus" analysis calculated as the mean of these 4 and the ECMWF analysis. These MA members are currently experimental, but may be used to provide some additional uncertainty information not contained in the SV perturbations. Since there are only 5 of them compared to 50 SV-perturbed members, they are used in this paper with double-weighting in calculation of probabilities – several different weightings were used experimentally and some results suggested that a higher weighting of 4 provided the optimal results, but the lower weighting is currently used as a more conservative approach until stronger evidence is available.

Despite these numerous developments from the initial version of the EPS, most verification has been based on moderate-severity events and broad-scale parameters, notably 500 hPa geopotential height. Legg *et al*. (2002) presented some verification of site-specific probability forecasts of surface weather parameters, for which skill was rather limited, especially for more extreme event thresholds. An ensemble can only attempt to estimate the probability distribution of forecast states, and in practise ensembles normally show insufficient spread to cover the full uncertainty in the forecast. Mylne *et al*. (2002) corrected this spread to provide calibrated site-specific probability forecasts from the EPS; calibration substantially improved the quality of ensemble probabilities for non-extreme events, but actually degraded the skill for extreme events. The initial aim of the FGEW project, to attempt to predict real severe weather events from the EPS, was thus ambitious. However it was considered important to test the ability of the EPS to help with severe weather prediction in order to make full use of the EPS for applications of real concern to Met Office customers. The NSWWS Early Warnings, issued up to 5 days ahead, are well-suited for prediction with the EPS which is designed for optimal performance in forecasting more than 48 hours ahead.

3. Predictability of Severe Weather

Before considering the implementation of the FGEW system, it is worth considering the predictability of severe weather, and what we should expect from ensemble prediction.

The defined requirement for the issue of Early Warnings in the NSWWS is a probability of 60%. It is interesting to speculate on how often severe weather is likely to be predictable at this level more than about 24h ahead. Evidence from the December 1999 storms over France and Germany showed that only a minority of ensemble members (or of deterministic forecasts from different centres) succeeded in predicting severe storms, even at ~24h ahead. Fig. 1 illustrates schematically that, in a synoptic situation when severe weather is possible, as shown by the control forecast C which evolves to exceed a warning threshold in severity, once a forecast moves into the chaotic non-linear regime, most ensemble members (C±p) are likely to be drawn towards the model's climatology. The result of this is that the forecast pdf (probability density function) is always likely to be skewed away from severe weather. Thus, 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.

(Although the diagram illustrates this idea with the central control forecast C predicting severe weather and perturbed analyses (C±p) 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 EPS in its current formulation is designed on an assumption that the large-scale evolution of the atmosphere is normally quasi-linear over approximately the first 48 hours, which would suggest that higher probabilities might be obtained within this time-range. However, Smith and Gilmour (1999) have found that typically there is important non-linearity in forecasts even within the first 48 hours. Since the development of severe weather is likely to involve strongly non-linear processes on small scales, this is particularly likely to be true when the state of the atmosphere is such that severe weather is possible. For severe-weather situations the quasi-linear limit is likely to be much less than 24h, and there may be little chance of predicting high probabilities of extreme conditions.

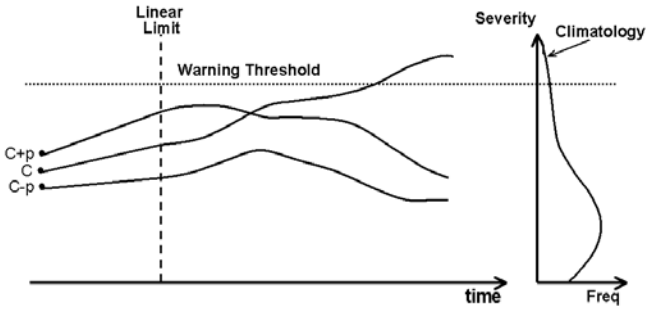

Figure 1. Schematic illustration of the effect of non-linearity on an ensemble forecast consisting of control forecast starting from initial condition C and perturbed forecasts with initial conditions C±p where p is a perturbation. Vertical axis represents some measure of severity, with climatology as shown on the right. In the early stages of the 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.

This argument is presented in terms of the expected behaviour of an ensemble forecast for clarity. However this does not mean that low predictability is a failure of the ensemble method. The sensitivity to small perturbations illustrated in Fig. 1 is characteristic of the real atmosphere as much as of the NWP model. Just as the model may be in a state where there is potential for severe weather but most ensemble members will not develop it, so may the real atmosphere, so severe weather development is often a fundamentally low probability event, and we should not expect to be able to forecast high probabilities often. There will be exceptions to this, occasions when there is much potential and a high probability of releasing it, but even on these occasions there is likely to be uncertainty about exactly where the severe weather will develop – on these occasions the large-scale probability of severe development will be high, but local probabilities will still be low. A good example of this was the storms of December 1999 in Europe. These storms developed as a result of an exceptionally strong jet-stream in the upper troposphere. The strong jet was a large-scale feature and was very predictable, but the resulting cyclogenesis was much less certain. Palmer (2002) showed that only a minority of EPS members developed a deep cyclone, and among those that did there was considerable variation in depth and location. The highest local probabilities of severe wind gusts were around 30%.

4. The FGEW System

a. Calculation of Representative Probabilities

There is often concern that end users do not understand what probability forecasts mean. To avoid this it is vital that the weather events for which probabilities are given are clearly defined. For example, a warning of heavy rain in southern England must specify exactly how "heavy rain" is defined; also whether the probability refers to "somewhere in southern England" getting heavy rain, or "any specific location" within that region experiencing the heavy rain – two very different probabilities. Fortunately for this project, the NSWWS specifies events quite clearly. Table 1 gives the definitions of severe weather events used in this project; probabilities are given for these events occurring "anywhere in the UK" and also within each of 12 sub-regions of the UK. In either case, the weather need only occur somewhere in the region, and not everywhere.

Conventionally, ensemble probabilities are often shown as contoured charts of grid-point values for specific times. Severe events tend to occur quite locally over small horizontal areas at any fixed time, thus affecting only a few grid-points in each ensemble member. The uncertainty represented by the ensemble spread means that even among those EPS members which generate severe weather they are likely to have it in different locations, and perhaps with differing timing, especially in forecasts several days ahead. As a result, the contoured grid-point probabilities of severe events are almost invariably low. However for the NSWWS Early Warnings we require the probability of severe weather occurring anywhere in a region – for the overall probability which determines whether a warning is issued or not, this region is the whole of the UK. Thus, to calculate the probability required, each ensemble member need only exceed the defined weather threshold at one grid-point in the whole region to count towards the required probability. Similarly, Early Warnings are issued to cover a defined period of time, typically between 12 and 36 hours – the precise timing of an event is not essential when issuing a warning several days ahead. Thus in calculating the warning probabilities the threshold need only be exceeded at a single time within a 12-hour time-window for an EPS member to count towards the probability. (Note that in the early development of FGEW it seemed sensible to allow this time-window to expand for forecasts further ahead, as timing uncertainty increases with lead-time. However, this resulted in the probability bias in the forecasts changing with lead-time which made consistent tuning of the weather thresholds impossible, so the time-window was fixed at 12 hours for all lead-times.) Calculating probabilities for entire regions and for time-windows in this way results in much higher probabilities of severe weather than are seen at individual grid-points at fixed times, and also provides the best estimate of the probabilities actually required for the NSWWS warnings.

b. Definition of Severe Weather Thresholds

In calculating probabilities for FGEW it was necessary to specify the events carefully from the EPS fields. As described above, for any probability forecast, it is essential to define both the weather events and the probabilities. Considering the weather events in table 1, it is clear that these cannot be identified directly from the EPS output fields. For example the heavy rainfall definition is "15mm in 3 hours". EPS fields are only output every 6 hours so it is immediately necessary to define a "proxy event" which

can be identified from 6-hour rainfall accumulations. The definition of severe gales is given in terms of gusts, but the standard EPS product is mean wind speed, so empirically-based 'gust factors', differing over land and sea grid-points, were used to estimate gusts from mean speeds. (Note: ECMWF does now also offer a parametrized gust product from the EPS. FGEW experiments have been conducted using this product, but no overall benefit was found and it is not used in the current implementation of FGEW.) As well as the basic mis-match between model output fields and the real-world warning definitions, an NWP model with 80km resolution also cannot be expected to resolve the locally-observed extremes in a severe-weather event. Thus the thresholds defined to represent real severe weather events in the model are expected to be less extreme in real terms than the real-world events in Table 1.

The initial specification of the proxy events was necessarily somewhat arbitrary, but the precise thresholds used have subsequently been calibrated. For an unbiased probability forecast system the mean forecast probability should equal the sample climatology. Thresholds were calibrated to minimise the bias in event probabilities using the first winter season's verification data (17 October 2000 to 4 May 2001), and are thus tuned to provide optimal performance.

c. An Alternative Approach – Climatology-based Severe Weather Thresholds

One way around the problem of having to calibrate the model output in terms of sensitivity is to look for extreme forecasts relative to a model climatology. Lalaurette (2003) has generated an approximate climatology of the EPS based on a limited period (up to 3 years) of EPS forecasts, exploiting the fact that 51 ensemble members running out to 10 days evolve into many more realisations than the real atmosphere does over the same period. Lalaurette uses the climatology to define an Extreme Forecast Index (EFI) which measures how extreme a forecast distribution is relative to the model's climate distribution. This approach provides alerts of the risk of extreme weather, but does not offer explicit probabilities of extreme weather as required for FGEW. Instead, here we relate the model climatology to the real climatologies at observing sites to obtain an objective calibration of warning thresholds. Compared to the standard FGEW method described in 4b. above, this method has the advantage that it could be used to calibrate the system for any new warning threshold required, and anywhere in the world, without the need for a prolonged training and tuning period.

Warning probabilities are estimated for approximately 50 UK observing sites, using corresponding model grid-points (compared with around 200 grid-points used in the standard FGEW approach). For each observing site, the warning threshold is calibrated using the process illustrated in Fig. 2. Firstly, the real NSWWS warning threshold (table 1) is compared with the observed site climatology (for the time of year) to determine the percentile point it represents on the climatological distribution. Then from the model climate distribution for the representative grid-point, the value is determined which has the same percentile, or frequency of occurrence, and this sets the warning threshold for the model forecasts. This threshold is then used with the ensemble forecast distribution to determine the forecast probability for that site. FGEW probabilities for NSWWS regions are then calculated in the same way as those using the standard FGEW thresholds (section 4b.), with regions represented by the

sites within them rather than the model grid-points, and using the same time-windowing (section 4a.). This method is currently used only for wind and heavy rain warnings, as suitable climatologies for snowfall are not available.

d. FGEW System Versions

Verification results will be presented from the current operational version of FGEW, plus three experimental versions, as follows. Version letters will be used to identify them in the remainder of the paper:

- Operational FGEW system (Version A)

The operational system uses the 51-member operational EPS, plus the five MA members, doubly weighted compared to the regular EPS members (section 2).

- 51-member EPS (Version B)

The standard operational 51-member EPS initialised at 12 UTC each day.

- 102-member EPS (Version C)

102-member EPS obtained using the 12 UTC operational 51-member EPS plus an experimental 51-member EPS run from 00 UTC each day.

- Climatology method (Version D)

As Version A but calibrated using the climatology-based method described in section 4c.

Also, NSWWS 'Issued' warnings will be referred to as "N" in some of the comparisons which follow.

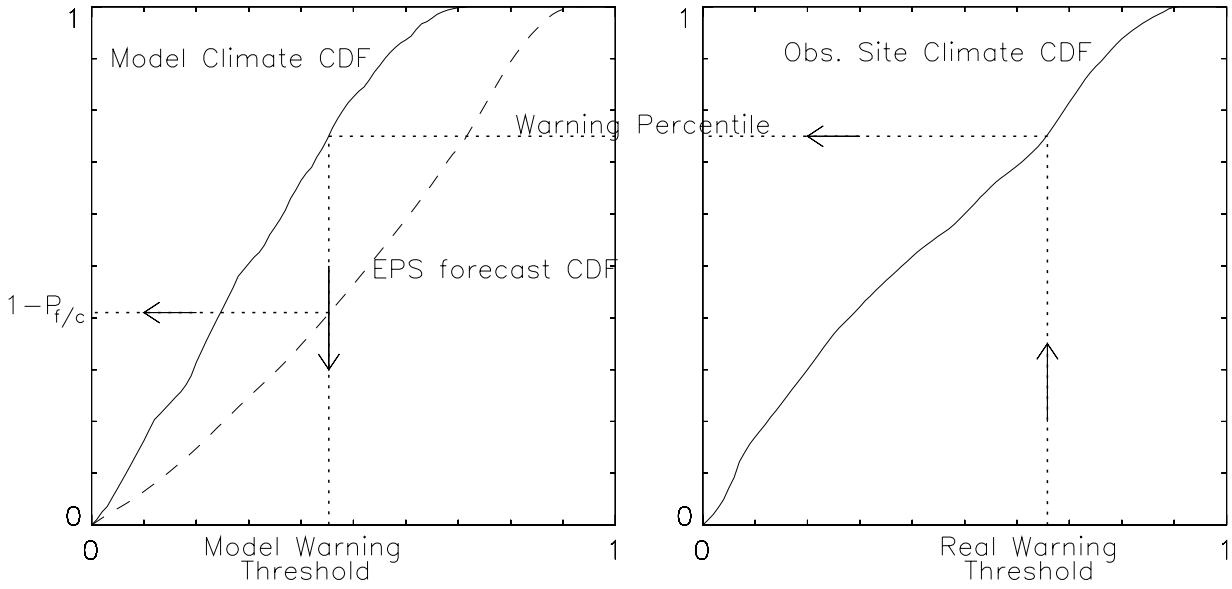

Figure 2. An illustration of how model and site climatologies may be used to determine event probabilities. Horizontal axes are labelled in arbitrary units of weather severity; vertical scales represent cumulative event probability.

5. Verification methodology

As stated above, Early Warnings are verified against the issue of Flash Warnings. Flash Warnings are issued for the same severe events as Early Warnings but within a very few hours of the event when confidence is high. Thus they make convenient proxy observations for severe weather, avoiding the need for complex analysis of real

observations. Verification of Flash Warnings has shown a very high correspondence with actual severe weather, with very few missed events or false alarms (K. Hymas, personal communication).

Verifications presented herein mostly use standard probabilistic verification scores, including Relative Operating Characteristic (ROC; Stanski *et al.*, 1989), Reliability and Brier Score (Wilks, 1995) and Relative Economic Value (Richardson, 2000). Much of this verification is event-based, using contingency tables of 'Hits' *H*, 'Misses' *M*, 'False-Alarms' *F*, and 'Correct Rejections' *R* (Table 2). For probability forecasts contingency tables are determined for a range of probability thresholds, with forecasts assigned as 'yes' if the forecast probability exceeds the threshold.

Creation of contingency tables for the types of warnings considered here is complicated by the fact that both warnings and events may span any time period. It is thus necessary to define a set of rules by which a forecast may be considered sufficiently accurate to be a Hit, and how to define a non-event for a Correct Rejection. The basic unit used was a calendar day, but rules were devised to avoid double-counting where a warning or event spanned two days:

- An Early Warning is judged to be a Hit if any part of its validity period overlaps the period of a verifying Flash.
- Where either an Early Warning or the verifying Flash spans two calendar days it will be counted on the first day of validity only (regardless of which day they actually overlap). Both warnings are then ignored for the second day. The only exception is where the Early Warning spans more than 24 hours *and* Flashes are valid on both days, in which case two Hits are recorded.
- Where an Early Warning spans more than 24 hours and only one day is validated, the second will record a False Alarm to penalise over-long warnings.

Since the forecast probabilities are expected to be low for warnings of rare events, it is important to design the verification such that it resolves information about low-probability warnings. To achieve this a non-standard set of probability thresholds is used both for the contingency tables and for the probability bins in reliability diagrams – 0.01, 0.03, 0.05, 0.09, 0.13, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

Table 2. Two-by-two contingency table of events for forecast verification

	Forecast	Not forecast	
Observed	*H*	*M*	*H+M*
Not observed	*F*	*R*	*F+R*
	H+F	*M+R*	*H+M+F+R*

In the NSWWS Early Warnings probabilities are generated both for the UK as a whole, and for each of 12 individual areas. Verification results were produced for both, and a mixture are presented below. As expected from the discussion above, issued probabilities are generally lower for the individual areas, but the sample sizes are larger (although it must be noted that these samples are not fully independent).

One aim of the FGEW project was to encourage earlier issue of warnings. Results will concentrate on the performance of the FGEW system at 2-5 days ahead compared to NSWWS warnings issued one day ahead, the only range at which sufficient NSWWS warnings are issued for meaningful verification. Note that operationally FGEW warnings based on 12 UTC data reach forecasters around 06 UTC the following day, so a 3-day (D+3) FGEW warning can only really be used two days ahead.

A major problem with verifying severe weather warnings is the small sample sizes due to the rarity of events. This problem is particularly acute when verifying probability forecasts. Results are presented here for a verification period from 1 October 2001 to 12 February 2003. (Tuning of the system was performed on a prior verification period from 17 October 2000 to 4 May 2001.) During the verification period there were Flash Warnings issued on only 16 days for severe gales, 17 days for heavy snowfall, and 62 days for heavy rainfall. Correspondingly, most of the results presented here will be for heavy rainfall, with fewer for severe gales and heavy snowfall. Inevitably, the effects of small samples will be apparent in the results presented. Nevertheless we believe that there are a number of consistent messages in the results, and that some useful conclusions can be drawn.

6. Relative Operating Characteristic

Relative Operating Characteristic (ROC) (Stanski *et al.,* 1989) explores the Hit Rate $HR=H/(H+M)$ and False-Alarm Rate $FAR=F/(F+R)$ (using the notation in Table 1) together. Because both *HR* and *FAR* are stratified by the observations (there are $H+M$ events and $F+R$ non-events), ROC measures the forecast system's ability to discriminate between occurrences and non-occurrences of an event. *HR* and *FAR* are evaluated for a range of probability thresholds from the contingency tables described above, and plotted in a graph of *HR* against *FAR*, giving the ROC curve. A perfect forecast would have $HR=1$ and $FAR=0$, so for a skilful system the curve is bowed towards the upper-left part of the graph, indicating useful probabilistic information that can be applied to decision-making. Forecasts with no discriminating power have $HR=FAR$. A useful summary measure of skill is the area under the ROC curve, which is 0.5 for a skill-less system and 1.0 for perfect forecasts. Note that the probability thresholds used here, focussing on lower probabilities, will tend to increase the ROC area compared to using thresholds every 10%. ROC curves using the full range of probability thresholds represent the discrimination skill of the forecasts for users who have the flexibility to tune their decision-making according to their cost-loss ratio associated with protecting against severe weather (Mylne, 2002). NSWWS Early Warnings are only issued when the UK-wide probability is at least 60%, so, for a direct comparison of FGEW and NSWWS warnings, some ROC curves are also produced omitting forecast probabilities below 60%.

ROC results

ROC curves for heavy-rainfall probabilities over the whole UK are shown in Fig. 3. FGEW probabilities show clear evidence of skill, with the greatest ROC area at D+4, showing that the ensemble is best able to discriminate heavy rain events at this range. This result is remarkable, given that for most forecast systems the skill is maximum at short range and decreases at increasing range. However it is worth noting that this result was very robust, and did not, for example, depend on the calibration thresholds used. Altering the calibration affected the area under the ROC curves at all ranges,

but did not alter the fact that it was maximised at D+4. Possible reasons for this effect, and implications, will be discussed in section 11. There is no significant difference between the different versions of FGEW.

NSWWS warnings issued by forecasters (black) clearly have skill at 1 day ahead, and there is a small amount of skill at 2 days; beyond this, there have been too few warnings issued to allow meaningful results. At first sight D+4 skill of FGEW warnings appears better than the D+1 issued warnings, but in fact much of this skill comes from the low end of the probability range, represented by points closer to the top-right part of the graphs. Since the NSWWS warnings are only issued when the UK probability is 60% or more, a fair comparison can only be made by excluding points corresponding to FGEW probabilities below 60%. This is shown in Fig. 4, comparing FGEW 2- and 4-day forecasts with D+1 issued warnings, and it can be seen that at both 2 and 4 days FGEW is able to discriminate a small number of events at the 60% level, but less than the issued warnings achieve at D+1. Version D of FGEW, using the climatology calibration method, appears to perform slightly better than the other versions at this 60% threshold. Overall skill at the 60% threshold required for the NSWWS is very limited, but there may be some scope for issuing a limited number of warnings earlier than has been done in the past, based on FGEW D+4 products.

ROC curves for 2- and 4-day FGEW forecasts of heavy rain in the 12 individual UK regions are shown in Fig. 5, and are similar to those for the whole UK (Fig. 3). Again FGEW performance is best at D+4.

ROC curves for Severe Gales are shown in Fig. 6, for 2 and 4 days ahead for the whole UK, and at 4 days for the individual regions. D+1 issued warnings are overlaid as before. Again FGEW skill is maximum at D+4, but with most of the skill coming from the low probability thresholds. Results for heavy snow warnings are not shown, but are very similar.

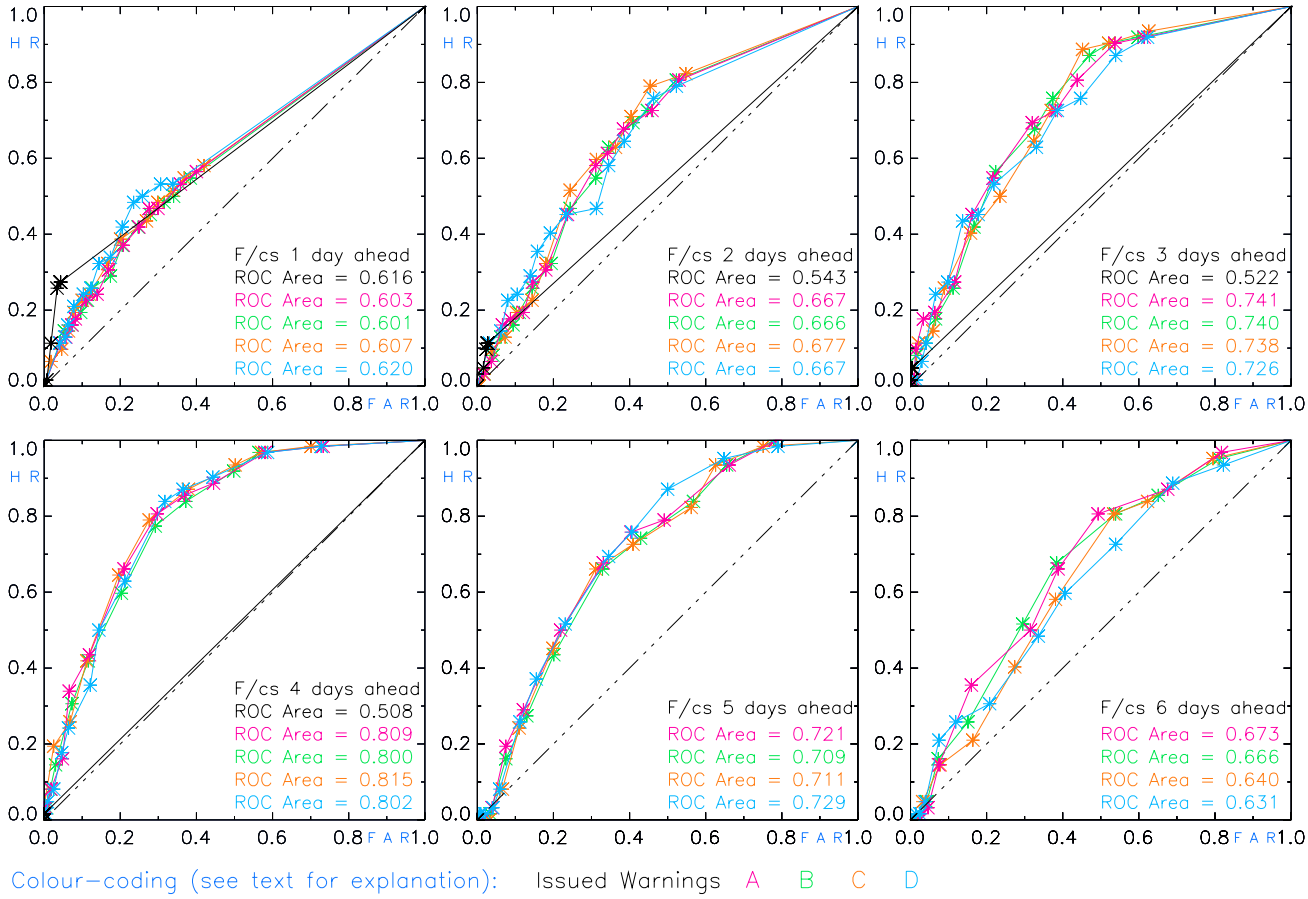

Figure 3. ROC curves for probabilities of Heavy Rainfall events occurring anywhere in UK (plotted for all probabilities). Forecasts for 1 to 6 days ahead. Data period: 1 Oct 2001 – 12 Feb 2003.

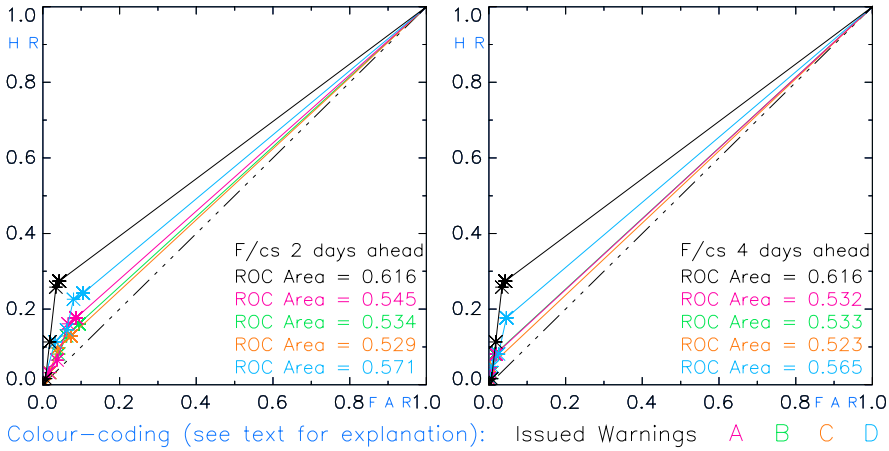

Figure 4. ROC curves for probabilities of Heavy Rainfall events occurring anywhere in UK (plotted for probabilities ≥ 0.6 only). FGEW probabilities for 2 and 4 days ahead are plotted, with Issued probabilities 1 day ahead overlaid for comparison. Data period: 1 Oct 2001 – 12 Feb 2003.

In terms of the ability to discriminate between occasions when severe events are or are not likely to happen, the graphs in this section (Figs 3 to 6) show that at 1-2 days ahead the Issued warnings do have some skill, which is not matched by the FGEW probabilistic warnings at this range. (Remember that, for the practical reasons

described above, it is in fact more relevant and meaningful to compare the Issued warnings at day 1 with the FGEW results at day 2, etc.) However, warning probabilities from FGEW for 4 days ahead – which in practice become available to the forecaster 3 days ahead of the verifying time – perform better than those at 1-2 days ahead. Thus the forecaster could gain extra information by heeding the FGEW output at this range, and should be able to issue warnings at least as skilful as those currently offered just 1-2 days ahead.

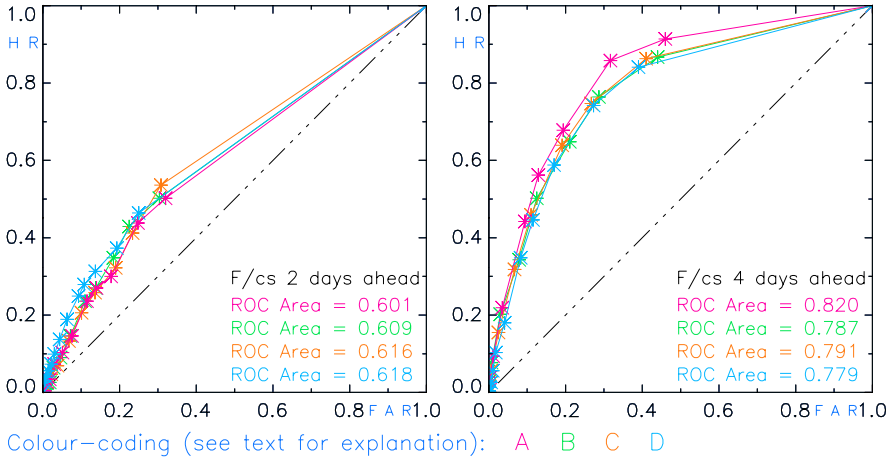

Figure 5. ROC curves for FGEW probabilities of Heavy Rainfall events occurring in individual areas (plotted for all probabilities) at 2 and 4 days ahead. Data period: 1 Oct 2001 – 12 Feb 2003.

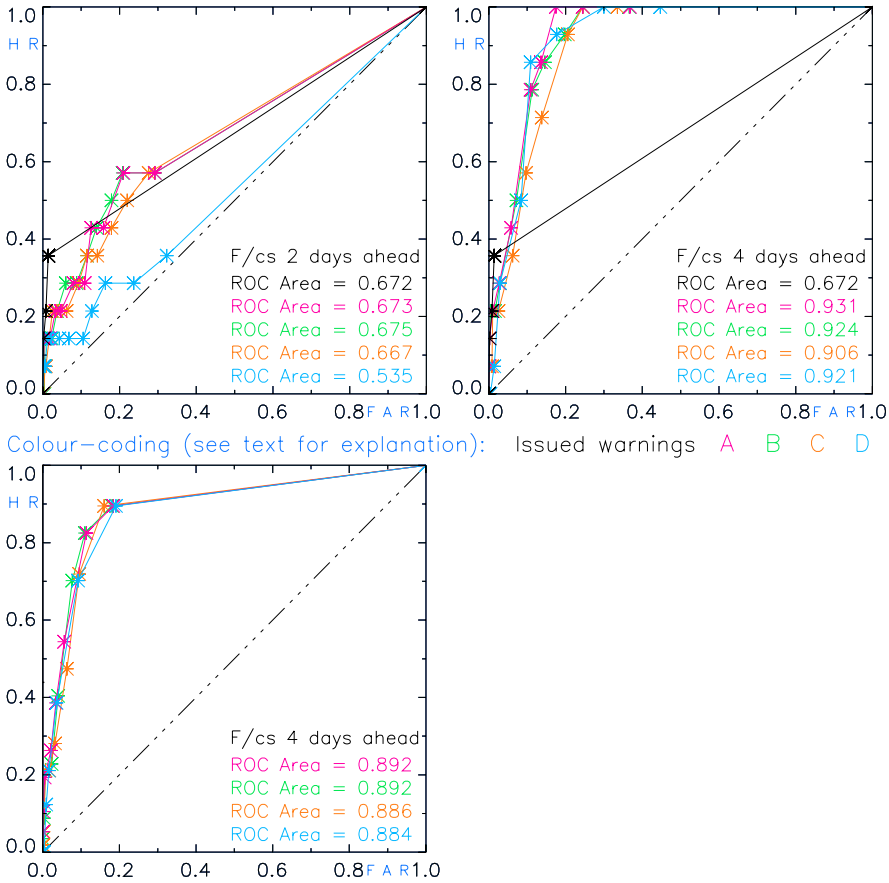

Figure 6. ROC curves for probabilities of Severe Gale events (plotted for all probabilities). (Left) probabilities for events occurring anywhere in UK; FGEW 2 days ahead and Issued 1 day ahead. (Centre) probabilities for events occurring anywhere in UK; FGEW 4 days ahead and Issued 1 day ahead. (Right) FGEW probabilities for events occurring in individual areas 4 days ahead. Data period: 1 Oct 2001 – 12 Feb 2003.

7. Reliability

For an ideal probabilistic forecasting system, of all occasions when a probability of *x*% is assigned to an event, that event will occur on *x*% of occasions. A reliability diagram, in which the frequency of occurrence of an event is plotted against the forecast probability (binned into a series of finite ranges), illustrates the extent to which this ideal is met (Wilks, 1995). An ideal forecasting system will produce a straight diagonal line along $y=x$. A reliability diagram which strays below the $y=x$ line indicates overestimation of forecast probabilities, while a near-horizontal curve would indicate a lack of resolution in the forecasts.

It is also useful to indicate how often each probability bin was forecast ('sharpness'), as on the diagrams that follow. Note that we have used the same set of thresholds as above, focussing on low probabilities, to separate the bins.

Reliability results

Figs 7-10 present a selection of reliability diagrams for heavy rain and severe gale warnings, along with accompanying sharpness diagrams. Each diagram includes the various versions of the FGEW system at 2 or 4 days ahead, overlaid with the Issued NSWWS forecasts at D+1 for comparison. A common feature of all the reliability diagrams is a high level of statistical noise (jagged graphs) for higher probabilities, above 20-40%. This is characteristic of small samples, and is unavoidable in forecasts of rare events, but does not prevent some useful conclusions being drawn.

Included in each diagram is a horizontal line indicating the sample climatological frequency, and mean probabilities are given for each forecast system for comparison. For an unbiased probability forecast system the mean forecast probability should equal the sample climatology. This requirement was used in the development of the FGEW system to calibrate the warning event thresholds, using a prior set of verification data from the previous year (17 October 2000 to 4 May 2001).

Fig. 7 shows reliability curves for Heavy Rainfall, for probabilities of events in the UK, comparing the performances of different versions of the FGEW system at D+4 with that of Issued warnings at D+1. The FGEW reliability curves for D+4 show excellent reliability for probabilities up to about 30%, where the sample sizes are quite large. At higher forecast probabilities the samples are noisy but there is a clear tendency to over-estimate the probabilities, indicated by the curves falling below the ideal diagonal. The occurrence of severe weather is substantially higher than the climatological frequency, so there is some useful forecast information in the higher probabilities, but there is little evidence of resolution between different forecast probabilities above about 40%, i.e. the proportion of events that occurred is largely independent of the forecast probability. Looking at the D+1 issued NSWWS warnings, a Flash Warning was clearly more likely to be issued after an Early Warning than when

no such warning had been issued. However, a large proportion of Flash Warnings were not preceded by Early Warnings, as shown by the point for issued probability of zero which only lies slightly below the sample frequency. This is largely due to the restriction which prevents forecasters from issuing warnings when the probability is less than 60%.

There is no clear difference in performance between the different FGEW versions. There is some indication that the 102-member version C performs better at the higher probabilities, though no statistical significance can be placed on this.

Fig. 8 shows similar curves but the FGEW forecasts are those for D+2. In this case there is virtually no resolution. The only positive feature is that in the lowest two probability bins (p<3%) the probability of occurrence is substantially below the sample frequency, while for all other forecast probabilities it is, on average, slightly above. This is consistent with the ROC results which indicated that there was much better discrimination, which is closely related to resolution, at D+4 than at D+2. Fig. 9 presents reliability diagrams for Heavy Rainfall in the individual areas of the UK at D+4 for FGEW and at D+1 for issued warnings. FGEW reliability curves show good resolution, with a strong positive slope, although this is slightly less than the ideal, indicating slight over-confidence (i.e. high probabilities are too high, low probabilities too low). The issued warnings have to be interpreted with care, remembering that they can only be issued on occasions when the UK probability is estimated to be 60% or more, but given this restriction the issued probabilities appear quite reliable. There is no clear distinction between most FGEW versions, except that the climatology-based version D is over-forecasting more severely at higher probabilities. This version appeared in Fig. 4 to offer better discrimination of events at high probability, but it can now be seen that this is at the cost of significant over-forecasting which indicates that the true resolution is at lower probabilities as with the other versions of the system. Results for other lead-times are not shown, but, as with the ROC assessments, the FGEW system consistently performs best at D+4.

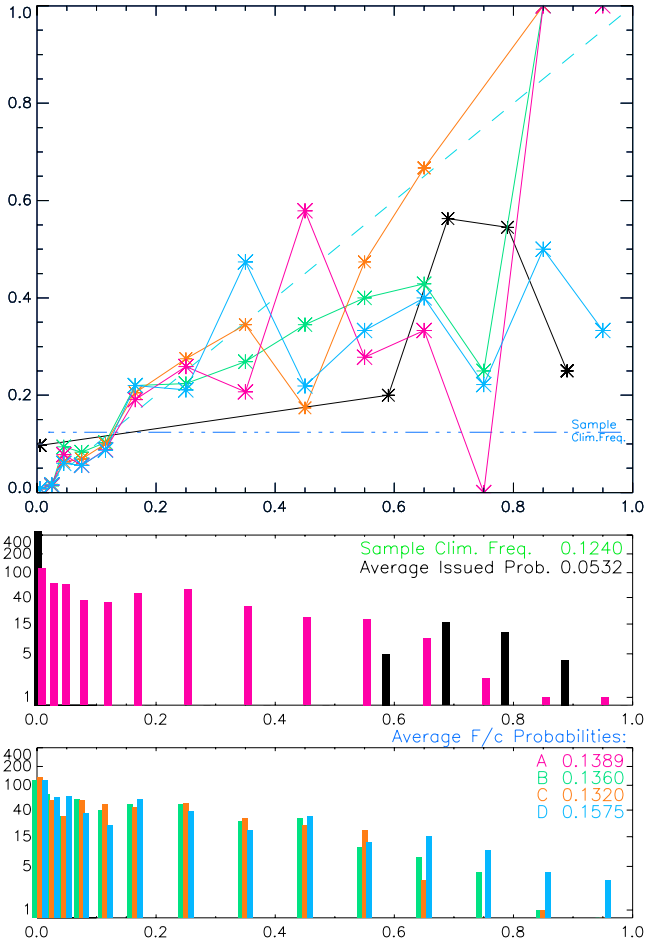

Figure 7. Reliability (top panel) and sharpness (second panel for versions N and A; third panel for versions B, C and D; note logarithmic scales), for probabilities of Heavy Rainfall events anywhere in UK, for FGEW warnings 4 days ahead and Issued warnings 1 day ahead (colour-coded). Data period: 1 Oct 2001 – 12 Feb 2003.

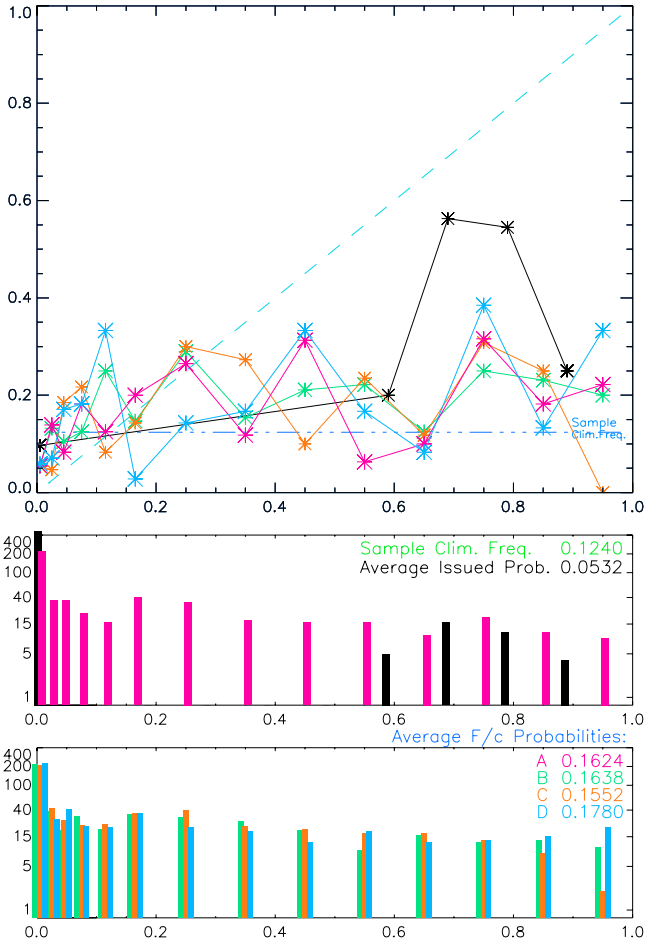

Figure 8. Reliability (top panel) and sharpness (second and third panels as Fig. 7; note logarithmic scales), for probabilities of Heavy Rainfall events anywhere in UK, for FGEW warnings 2 days ahead and Issued warnings 1 day ahead (colour-coded). Data period: 1 Oct 2001 – 12 Feb 2003.

Fig. 10 presents results for severe-gale events for the whole UK at D+4. These are broadly similar to the rainfall results, although subject to smaller sample sizes. FGEW warnings at D+4 show reasonable reliability at the lower probability thresholds. There were no occurrences of an event following any FGEW probability below 3% for D+4 forecasts, and few such cases at other forecast ranges, so the system shows good discrimination of a severe weather risk, at least at low probability. Issued warnings at D+1 again have a fairly high success rate when high probabilities are issued, but many events are missed due to the 60% threshold. Results for heavy-snowfall events (not shown) are similar with a reasonable degree of resolution at D+4, much less at D+3 and D+5 and no skill for other forecast days.

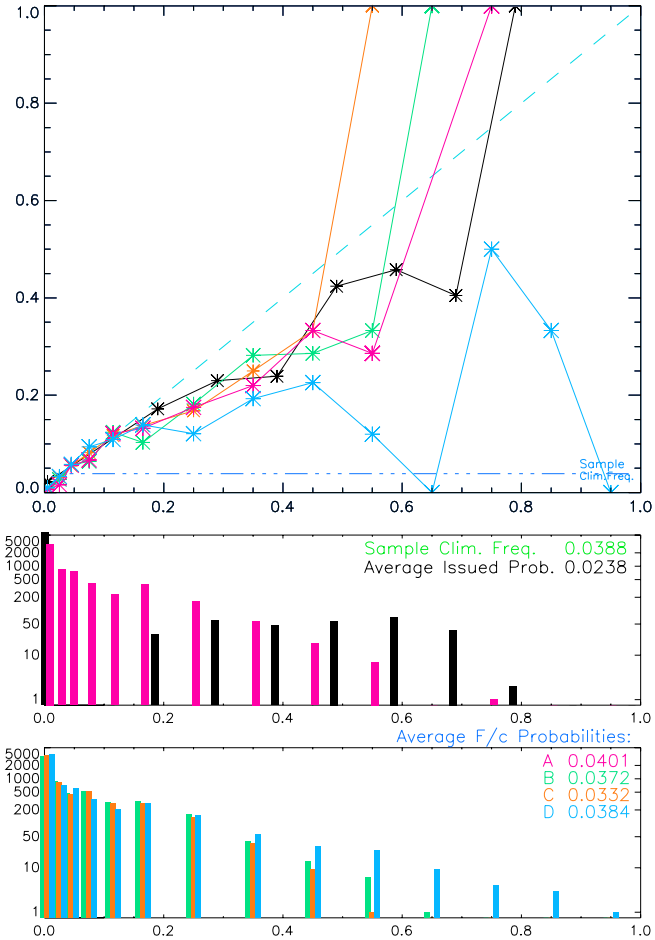

Figure 9. Reliability (top panel) and sharpness (second and third panels as Fig. 7; note logarithmic scales), for probabilities of Heavy Rainfall events in individual areas, for FGEW warnings 4 days ahead and Issued warnings 1 day ahead (colour-coded). Data period: 1 Oct 2001 – 12 Feb 2003.

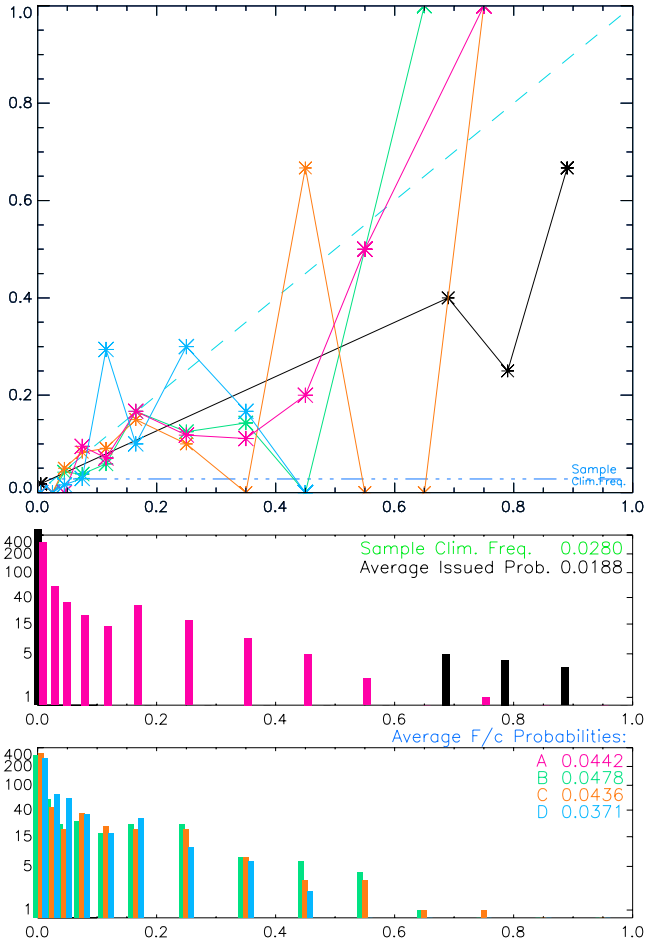

Figure 10. Reliability (top panel) and sharpness (second and third panels as Fig. 7; note logarithmic scales), for probabilities of Severe Gale events anywhere in UK, for FGEW warnings 4 days ahead and Issued warnings 1 day ahead (colour-coded). Data period: 1 Oct 2001 – 12 Feb 2003.

8. Brier Skill Scores

The Brier Score *BS* is a measure of mean square error for probability forecasts (Wilks, 1995):

$$BS = \frac{1}{N}\sum_{n=1}^{N}(p_f - p_o)^2 \qquad (1)$$

where p_f and p_o are the forecast and observed probabilities respectively, and *N* is the sample size; note that p_o can only be 1 (event occurred) or 0 (non-event). *BS* is bounded by the values 0.0 and 1.0; a lower value represents better forecasts.

Comparing Brier scores for different events can be misleading if their climatological probabilities are different, so it is more meaningful to calculate the Brier Skill Score *BSS*, obtained by comparing the Brier Score of the forecasting system BS_{fc} with that obtained by some reference forecast BS_{ref}:

$$BSS = 1 - \frac{BS_{fc}}{BS_{ref}} \qquad (2)$$

Typical reference forecasts used are climatology or persistence. For Early Warnings we did not know the prior climatological probability, and since they are rare events

we chose to use a null forecast (always forecasting the probability to be zero) for the reference. This shows whether the forecasts are better than the easy 'fall-back' option of never issuing warnings. After the initial training period this gave a crude estimate of the climatological probability for subsequent seasons, although based on only a single season. In Figs 11-13 *BSS* is plotted relative to null forecasts, but scores relative to this crude climatological forecast are also marked on the axes on the right side of the graphs.

Brier Skill Score results

Fig. 11 shows *BSS* for FGEW warnings of heavy-rainfall events anywhere in the UK. *BSS* are positive relative to both null forecasts and the crude climatological forecasts throughout D+1 to D+6, except for version "D" whose skill declines with lead-time. The current operational version of the system ("A") appears to perform slightly better than version "B" excluding the MA members, but the 102-member ensemble ("C") performs slightly better still, although it is unlikely that the differences here are statistically significant. The skill of FGEW warning probabilities for events in individual areas (not shown) is lower than for anywhere-in-UK. *BSS* are not shown for NSWWS issued warnings as the 60% threshold makes the *BSS* rather meaningless.

BSS for severe-gale events (Fig. 12) are also positive for all lead-times out to D+6, surprisingly increasing at the longest lead-times. Reliability and sharpness diagrams for D+5 and D+6 (not shown) reveal that this skill comes entirely from low-probability warnings. The current operational version of FGEW performs at least as well as any other version. Except at D+4, version "D" again performs less well than other versions of FGEW.

For heavy-snowfall events (Fig. 13) *BSS* is highest at 2-3 days ahead, and declines at longer range. Note here that *BSS* w.r.t. climate forecasts is higher than that w.r.t. null forecasts, because the sample-mean frequency during the training period was significantly higher (0.100) than during the assessment period (0.034).

Overall, *BSS* results show that the FGEW system has positive skill relative both to climatology and to null forecasts. This is observed consistently across all three weather events verified. Details of the variation in behaviour with respect to forecast lead-time are not consistent, and the marked trends seen for severe-gales and heavy-snowfall are probably not reliable due to the small sample sizes available.

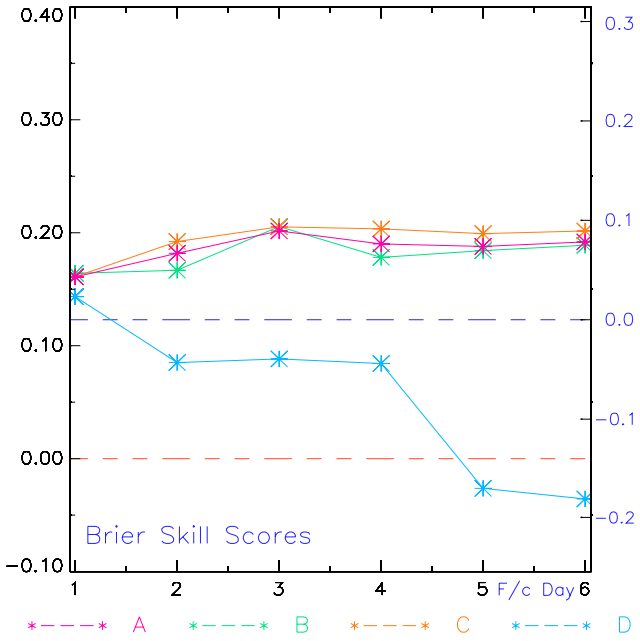

Figure 11. Brier Skill Scores (left-hand axis is skill w.r.t. null forecasts, right-hand axis is skill w.r.t. climate forecasts), for probabilities of Heavy Rainfall events occurring anywhere in UK, for FGEW and Issued warnings 1 to 6 days ahead. Data period: 1 Oct 2001 – 12 Feb 2003.

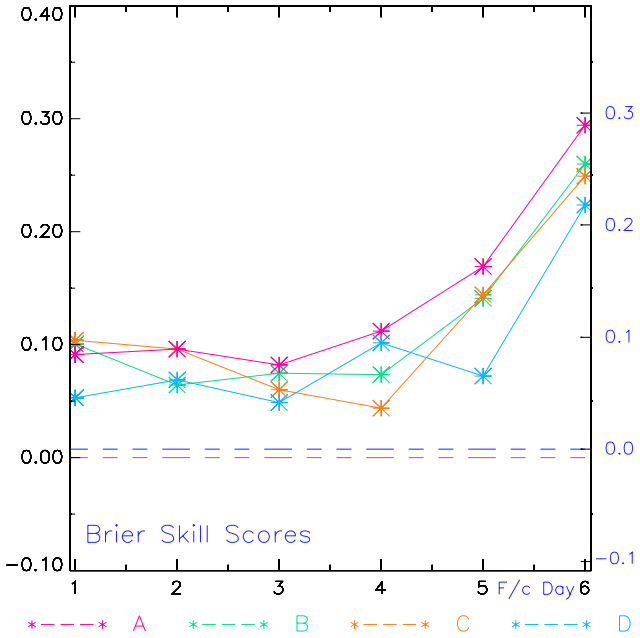

Figure 12. Brier Skill Scores (left-hand axis is skill w.r.t. null forecasts, right-hand axis is skill w.r.t. climate forecasts), for probabilities of Severe Gale events occurring anywhere in UK, for FGEW and Issued warnings 1 to 6 days ahead. Data period: 1 Oct 2001 – 12 Feb 2003.

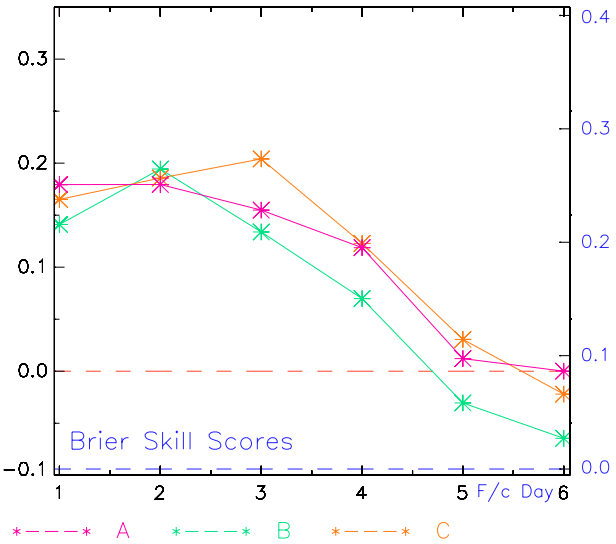

Figure 13. Brier Skill Scores (left-hand axis is skill w.r.t. null forecasts, right-hand axis is skill w.r.t. climate forecasts), for probabilities of Heavy Snowfall events occurring anywhere in UK, for FGEW and Issued warnings 1 to 6 days ahead. Data period: 1 Oct 2001 – 12 Feb 2003.

9. Proportion of Events Missed

An important consideration for users of warnings is the 'Proportion of events Not Forecast' (*PNF*). This is equal to *(1-HR)*, i.e. *M/(H+M)* where *HR* is the ROC Hit Rate and can be calculated at any forecast probability threshold. Clearly users want the value of *PNF* to be as small as possible (without incurring excessive false alarms).

Fig. 14 shows graphs of *PNF* at 1 to 4 days ahead, for heavy-rainfall probabilities anywhere in UK. The poor skill of the FGEW system at 1 day ahead is evidenced by the high *PNF* (approaching 0.5) for the lowest non-zero probability threshold. However, FGEW at Days 2-4 (at least for probabilities $p \geq 0.6$) performs almost as well as Issued warnings at Day 1, indicating that it should at least be possible to issue warnings around 2 days earlier. A point to note is that, at 2 or more days ahead for all probability thresholds above $p=0.3$, version D of the system achieves the lowest *PNF*, showing an apparent advantage. However this is consistent with the observation in section 7 that version D is consistently over-estimating event probabilities, and hence (at any probability threshold level) there are fewer times when an event is not forecast. Apart from this, it is the current operational version (A) of the FGEW system that performs best, though even for this version the majority of events are missed for probability thresholds above 0.3.

PNF increases monotonically from zero as the probability threshold is increased, but at D+4 especially this increase is relatively slow, meaning that very few events occurred when the forecast probability was very low.

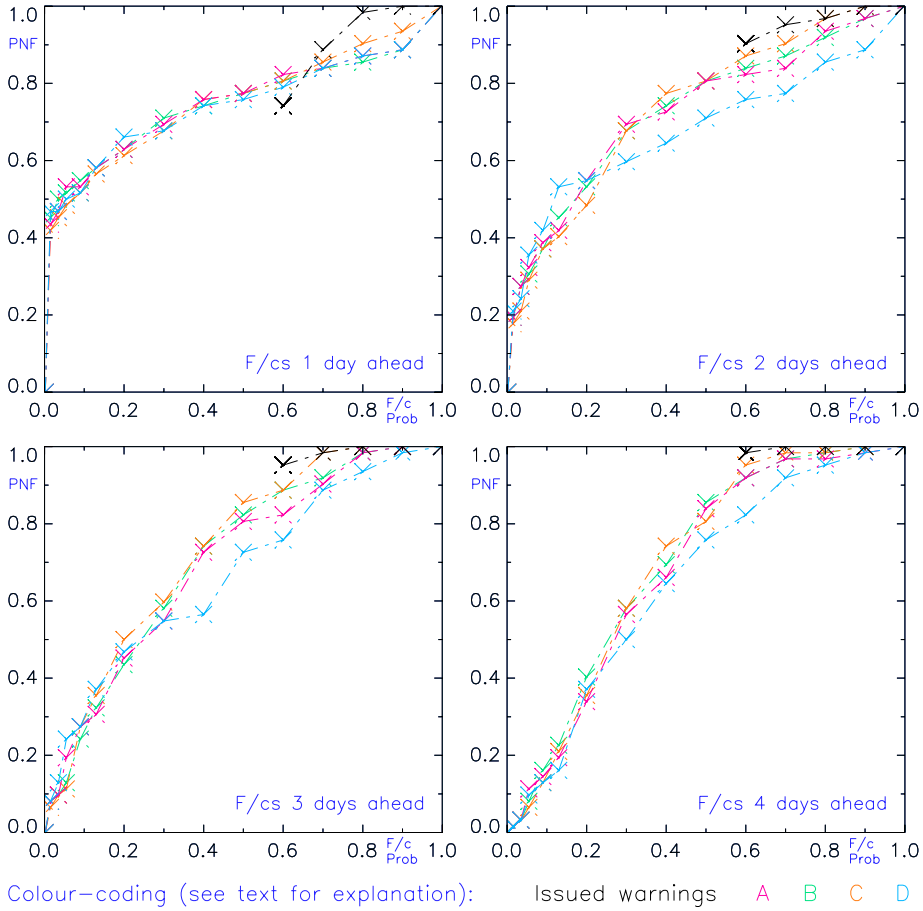

Figure 14. "Proportion of events Not Forecast", for probabilities of Heavy Rainfall events occurring anywhere in UK. Forecasts for 1 to 4 days ahead. Data period: 1 Oct 2001 – 12 Feb 2003.

For Severe Gale events, probabilities are on average lower, and, because there have been fewer events than for Heavy Rainfall, the graphs are noisier (not shown). FGEW performance is best at Days 3-4, but in this case issued warnings at Days 1-2 are more skilful than FGEW probabilities at any range.

Heavy Snowfall also suffers from a shortage of severe events (not shown), but FGEW is able to produce high forecast probabilities on some occasions when the event occurred. The skill of Issued warnings at day 1 is surpassed by FGEW at Days 3-5.

10. Cost-loss analysis

As noted earlier, to get the full benefit of probability forecasts, users need to make decisions at a probability threshold appropriate to their cost-loss ratio *C/L*. A user with a low *C/L* will be one who can take some protective action against severe weather at relatively low cost, but who stands to suffer a large loss in the event of severe weather occurring without protection. Such a user should thus take protective action when the probability of severe weather is quite low, whereas a user with a large *C/L* should only act when the probability is high (Mylne, 2002; Richardson, 2000). Using this simple decision model, ROC verification scores can easily be used to estimate the relative economic value of a forecast system for a range of user *C/L* (Richardson, 2000).

The relative value *V* of a forecast system is the reduction in mean expense, *E*, as a proportion of that which would be achieved using a perfect set of forecasts, against a benchmark of using only climatological information,

$$V = \frac{E(clim) - E(forecast)}{E(clim) - E(perfect)} \qquad (3)$$

For a perfect forecast system *V*=1, but if the user has no forecast information and simply takes the best action according to climatology (either always protecting or never protecting) then *V*=0. A forecast system from which the user can benefit will have *V*>0.

Fig. 15 shows the relative economic value of FGEW warnings of Heavy Rainfall over the whole UK at D+4, as a function of *C/L*. For a reliable forecasting system, *V* is greatest for *C/L* equal to the sample-mean frequency of the event, in this case around 0.13 consistent with Fig. 7. Also shown is the value curve for the issued warnings at D+1. The FGEW warnings clearly have much greater value to users with lower *C/L*, and this is simply because the restriction preventing the issue of warnings with probability below 60% prevents such users optimising their decision-making. It is also notable that the maximum value of D+4 FGEW forecasts is much greater than for the D+1 issued warnings. This is because the peak value, at *C/L*~0.13, is not well-matched to the 60% threshold limit.

One unusual feature of the cost-loss value curves in Fig. 15 is that some of them do not fall to zero at the extreme high and low *C/L* values, but reach a fixed non-zero value. This is an effect of the small sample sizes available in the verification data, and indicates that there is insufficient data to fully specify the ROC *HR* and *FAR* at every probability threshold.

Differences in the value of the different versions of FGEW are mostly quite small. The 102-member version C has marginally the greatest value towards the lowest *C/L* ratios, as the larger ensemble size improves the chance of capturing events at very low probabilities. The climatology-based version D has similar (but marginally lower) value over most of the *C/L* range, but is notably poorer for users with small *C/L*.

Fig. 16 shows the same cost-loss curves for Heavy Rainfall warnings over the whole UK given by the operational version A of FGEW at lead-times of 1 to 5 days, and also Issued warnings at D+1 and D+2. This supports the results from other diagnostics, indicating that the D+4 warnings from FGEW have the greatest user value, both in the absolute value of the forecasts and also in the range of different users, represented by different *C/L* ratios, who can benefit. The greatest-value issued warnings are those issued at D+1, but the sample size at D+2 is very small which makes comparison difficult. Peak value of the FGEW warnings is higher than the D+1 Issued warnings for all lead-times except D+1. The D+4 FGEW warnings have more user value than the D+1 Issued warnings for all except a very few users with *C/L* around 0.45.

Fig. 17 is similar to Fig. 15, but for Severe Gales. The effects of small sample sizes are more severe, so conclusions may not be reliable. The operational version A of FGEW appears to give the best performance. All versions are better than the D+1 issued

warnings for lower *C/L* values, but, unlike the Rainfall warnings, the issued warnings have more value for users with higher C/L.

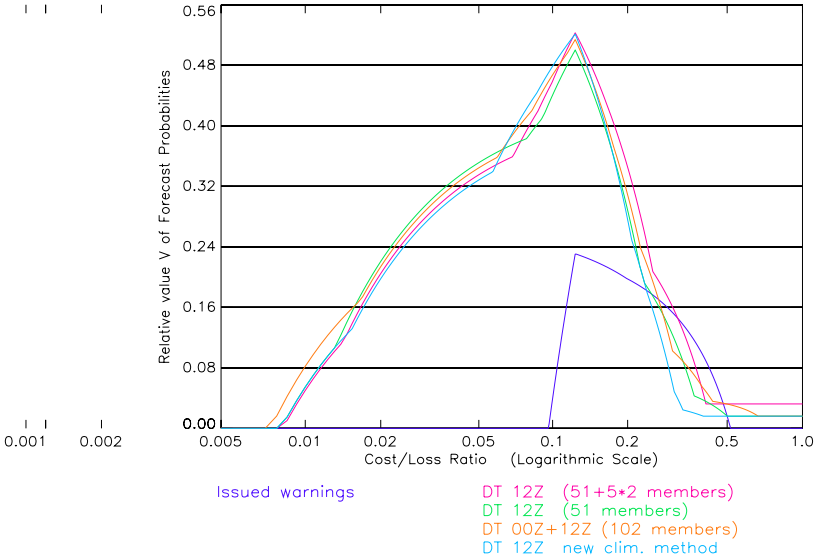

Figure 15. Cost/loss diagrams for Heavy Rainfall events, anywhere in UK. Issued Warnings for 1 day ahead, compared with FGEW probabilities for 4 days ahead. Colour-coded according to legend. Data period: 1 Oct 2001 – 12 Feb 2003.

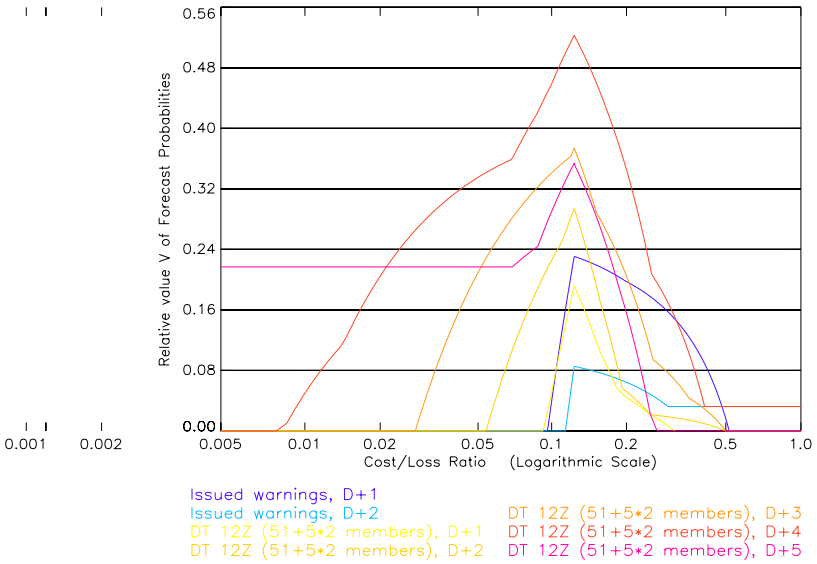

Figure 16. Cost/loss diagrams for Heavy Rainfall events, anywhere in UK. Issued Warnings for 1-2 days ahead, compared with Operational FGEW probabilities for 1-5 days ahead. Colour-coded according to legend. Data period: 1 Oct 2001 – 12 Feb 2003.

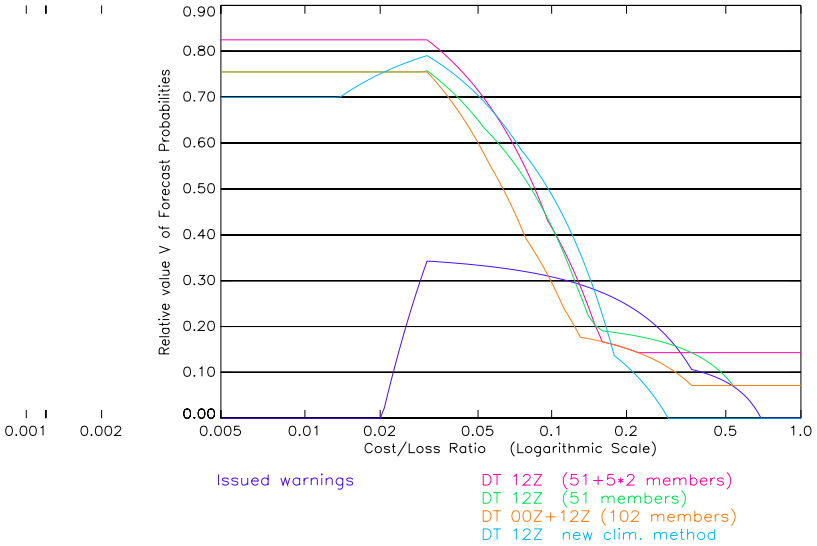

Figure 17. Cost/loss diagrams for Severe Gale events, anywhere in UK. Issued Warnings for 1 day ahead, compared with FGEW probabilities for 4 days ahead. Colour-coded according to legend. Data period: 1 Oct 2001 – 12 Feb 2003.

11. Discussion

The most remarkable result obtained from the experiments is that the FGEW system has its greatest skill for D+4 forecasts. The EPS is designed for medium-range use, and the SV perturbations are optimised for maximum ensemble growth over the first 48 hours of the forecast. It is to be expected that EPS performance may be poor at around 24 hours when the perturbations are still in their initial rapid-growth phase, but the normal expectation is that the performance will be best around the optimisation time of 48h, with a decline thereafter due to the normal limits of predictability. This expectation is supported by, for example, Barkmeijer *et al*. (1999) who showed ROC area for forecasts of 500hPa height anomalies decreasing steadily from D+2 onwards, and Buizza *et al*. (2000) who showed similar results for precipitation forecasts. Despite this, the FGEW result has proved extremely robust. For example, the result that ROC area is a maximum at D+4 is independent of the calibration of the warning thresholds described in section 4b. – calibration alters the absolute values of the ROC area, but not the fact that D+4 forecasts provide the largest area. As shown in Figs 7 and 8, once the calibration has been performed (by balancing the forecast bias such that the mean forecast probability is close to the climatology of the training sample), the D+4 forecasts also come closest to providing reliable probability forecasts. These results are consistent across all three of the weather types verified for FGEW: heavy rainfall, severe gales and heavy snowfall. One thing which is different about the verification reported here from most previously published EPS verification is the severity of the events being predicted, so the difference in results may reflect EPS performance for severe events. One of the few independent EPS verifications carried out for similarly extreme events was by Mullen and Buizza (2001, 2002) who verified rainfall forecasts over the USA. They reported a significant dip in performance at day 2 for the heavier rainfall thresholds (20 and 50mm per day), but not for lighter ones, although this was most apparent in the BSS rather than in ROC, specifically in the reliability term. It therefore appears that the

poor EPS performance around D+2, with better performance for longer lead-times, may be characteristic of forecasts of more extreme weather events.

So why might the EPS probability forecasts be better at D+4 than at D+2 or D+3 for extreme events? Ensemble forecasts frequently do not have enough dispersion amongst their members to fully represent the uncertainty in the forecasts. In order to provide reliable probability forecasts, the dispersion of ensemble members should be sufficient to generate uniform flat verification rank histograms (Hamill and Colucci, 1997). Mullen and Buizza (2001) present rank histograms for precipitation forecasts over the USA from the latest version of the EPS (as used in the FGEW experiments), verified using 24-hour accumulated rainfall data averaged onto a grid of 1.25°x1.25°. These show that the EPS is under-dispersive for precipitation forecasts, with excessive numbers of observations falling outside the ensemble spread at both ends of the distributions. Of particular relevance to forecasts of severe weather, the highest rank, representing observations with higher rainfall than all EPS members, is over-populated by a factor of about 2.5 (4.0) for D+2 forecasts in winter (summer). The over-population of the outlier ranks is reduced but not eliminated at D+5 (D+4 is not shown). The Flash warnings used in FGEW verification often take account of rainfall features occurring on shorter time and length scales than the averaged observations used by Mullen and Buizza (2001), so the excess of outliers is likely to be higher for FGEW forecasts. Under-dispersive ensemble forecasts typically result in over-confident probability forecasts, with reliability diagrams with a slope of less than the ideal 45°, as observed in the FGEW verification. Thus it is likely that the better performance of FGEW at D+4 than at D+2 is due to an improved ensemble spread – and perhaps more specifically a better ability to capture more extreme events, given that this behaviour is not seen when looking at less extreme events. Some evidence to support this is provided by the sharpness diagrams included alongside the reliability diagrams in Figs 7 (D+4) and 8 (D+2). It is notable that high probabilities are forecast on far more occasions at D+2 than at D+4. However the corresponding reliability diagrams show that these additional forecasts are almost entirely false alarms, as the high probabilities are not related to a higher occurrence of observations. More significantly, the numbers of forecasts of zero probability are approximately halved, from around 200 at D+2 to around 100 at D+4. Most of these forecasts are converted into low probabilities, but the effect is greatly to improve the reliability curve between 0 and 30% probabilities, almost eliminating missed events at zero probability. Thus the spread of the forecasts at D+4 appears to capture the true uncertainty much better than that at D+2.

It is not clear exactly why the ensemble spread should be more representative at D+4 than at D+2 for extreme events, but it is likely to be related to the perturbation strategy employed at ECMWF. Initial-condition perturbations are generated from SVs calculated to identify the fastest-growing modes of uncertainty as efficiently as possible (e.g. Buizza and Palmer, 1995; Molteni *et al*., 1996) by maximising the linear growth over an optimisation time of 48 hours. In addition, stochastic physics perturbations are applied to partially address uncertainties due to the model physics (Buizza *et al*., 1999). While this strategy provides an efficient sampling of the synoptic-scale uncertainty appropriate for medium-range forecasting, it may not be suitable for estimating probabilities of events close to the SV optimisation time of 48h.

Reliable estimation of probabilities requires a random sampling of the pdf, but around the optimisation time the sampling is far from random. However as the ensemble evolves beyond 48h one effect of non-linearity in the model is to mix up the solutions, effectively making the sampling more random. For most applications the strategy works satisfactorily because the probabilities being estimated are captured by the main body of the pdf which is relatively well-resolved, but for the current application we are sampling well into the tails of the pdf, where a non-random sampling is more likely to have an adverse impact on the probability estimates. In addition, as the effect of stochastic physics is cumulative through the forecast run, this will also help to improve the sampling for the estimation of probabilities over longer forecast ranges, and in particular this may help to compensate for the initial under-dispersion of the ensemble into the tails of the distribution – Mullen and Buizza (2001) showed that the introduction of stochastic physics, along with evolved SVs (Barkmeijer *et al*., 1999), greatly reduced the under-dispersion in precipitation forecasts. Thus these effects combined provide a plausible explanation for the improved performance at D+4 compared to D+2.

If our conclusions about the D+2 and D+4 results are correct, this will have considerable implications for future developments of ensemble prediction. Most current operational ensembles, such as the ECMWF EPS, are designed for medium-range prediction. Our results show that the EPS strategy is successfully able to provide useful forecast information on severe weather at this range, but demonstrates that the same methods cannot simply be applied also to the short range. Current research is rapidly moving towards short-range ensembles, and it is clear that different strategies will be required. If SV methods are to be applied they will need to be calculated at higher resolution over shorter optimisation times, and incorporating moist processes which are critical to the instabilities important at short-range. Alternatively it may be more appropriate to employ methods which focus on the modes growing most rapidly at or immediately before the analysis time, such as the Ensemble Transform Kalman Filter (Bishop *et al.,* 2001).

12. Conclusions

Probabilistic prediction of severe weather has for some time been seen as an ideal application of ensemble prediction, but few attempts have yet been made to apply ensembles operationally in this way. We have described a system built in support of the UK NSWWS, to aid forecasters in issuing warnings earlier and with greater confidence in probabilities. Severe weather thresholds for the ECMWF ensemble model were calibrated by tuning the probability bias over an initial training period; verification results were then obtained over a subsequent period which included two winter seasons. Despite this long verification period, sample sizes are small due to the rare nature of the severe weather events concerned, which is a limitation especially for probabilistic verification techniques. Nevertheless some useful conclusions can be drawn.

Predictability arguments suggested that we should not expect to be able to predict severe weather with high probabilities on many occasions, and this was confirmed in the results. On most occasions when severe weather occurred it was only possible to predict it at low probabilities. This is considered to be a predictability characteristic of

severe weather, and not just a limitation of the prediction method. Development of severe weather often requires the non-linear combination of several factors, and so the probability of occurrence has a fundamental low probability in the atmosphere as well as in a model. In fact, on those occasions when the FGEW system forecast higher probabilities of severe weather, above about 30%, the reliability diagrams showed that these forecasts were over-confident and the actual probability was considerably lower than forecast. Nevertheless the actual probability was substantially higher than climatology on these occasions, and this therefore still represented useful forecast information which could be calibrated before issue to end users.

The FGEW system has been shown to have a considerable capability in discriminating occasions when severe weather is possible or likely. This capability is mostly at low probabilities and therefore is of greatest benefit to users with lower cost-loss ratios for protective action. By contrast the warnings currently issued through the UK NSWWS are restricted to high probabilities (60% or more) which does not allow users who can make use of low-probability alerts to obtain all the benefit potentially available from the EPS. Ensemble prediction systems offer great opportunities for improved warnings of severe weather events, and these will improve further as ensembles are developed to focus more on severe weather on both medium and short timescales. However, to pass on the full benefit of the improved forecast information offered by ensembles to end users requires some fundamental changes in the way many forecast services are structured, interpreted and used.

The EPS provides the most reliable probabilities of severe weather at D+4, while forecasts at D+2 are virtually useless. This is attributed to the perturbation strategy used at ECMWF which is designed to sample the synoptic-scale uncertainties important for medium-range prediction. Alternative strategies will be required for future developments of short-range EPS.

Alongside the operational version of FGEW (A) we also tested a number of experimental versions. Overall the operational version performed as well as any of the test systems. For users with very small *C/L* ratios the 102-member ensemble (C), created by combining the standard 12 UTC EPS run with the experimental 00 UTC run which provides an additional 51 members, provides a little extra information by improving the chance of capturing events at very low probability. The version calibrated objectively using model and site climatologies (D) was notably less good, particularly in Brier Skill for severe gales. This is not surprising since the thresholds in the other systems have been tuned using a previous training set to optimise the probabilities, but the climatology method offers a useful approach for setting up an initial calibration which can subsequently be tuned in the light of experience.

Operationally, the FGEW system provides some useful extra information for Met Office forecasters in issuing NSWWS Early Warnings. Occasions when D+4 forecasts reach the 60% probability threshold required for issue of Early Warnings are rare, but when they do occur they provide a useful signal. The number of warnings issued around 3 days ahead (forecasters have access to D+4 FGEW warnings in time for issue of 3-day warnings) has increased significantly. However, to get the maximum value out of ensemble predictions of severe weather in the future will require changes in the way

warning services are structured, to provide warnings at lower probabilities so that users can make decisions appropriate to their own cost-loss ratios and exploit the ability of the EPS to predict low probabilities.

References

Barkmeijer, J., Buizza, R. and Palmer, T.N., 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF EPS. *Quart. J. Roy. Meteor. Soc.*, **125**, 2333-2351.

Bishop, C.H., Etherton, B.J. and Majumdar, S.J., 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Thoeretical aspects. *Mon. Wea. Rev.*, **129**, 420-436.

Buizza, R. and Palmer, T.N., 1995: The singular vector structure of the atmospheric general circulation. *J. Atmos. Sci.*, **52**, 1434-1456.

Buizza, R., Miller, M. and Palmer, T.N., 1999: Stochastic representation of model uncertainties in the ECMWF EPS. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887-2908.

Buizza, R., Barkmeijer, J., Palmer, T.N. and Richardson, D.S., 2000: Current status and future developments of the ECMWF Ensemble Prediction System. *Meteor. Appl.*, **7**, 163-175.

Hamill, T.M., and Colucci, S.J., 1997: Verification of Eta-RSM short-range ensemble forecasts, *Mon. Wea. Rev.*, **125**, 1312-1327.

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

Lalaurette, F., 2003: Early detection of abnormal weather using a probabilistic Extreme Forecast Index. *Accepted for publication in Quart. J. Roy. Meteor. Soc.*

Legg, T.P., Mylne, K.R. and Woolcock, C., 2002: Use of medium-range ensembles at the Met Office I: PREVIN – a system for the production of probabilistic forecast information from the ECMWF EPS. *Meteor. Appl.*, **9**, 255-271.

Molteni, F. and Palmer, T.N., 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. Roy. Meteor. Soc.*, **119**, 269-298.

Molteni, F., Buizza, R., Palmer, T.N. and Petroliagis, T., 1996: The ECMWF Ensemble Prediction System: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Mullen, S.L. and Buizza, R., 2001: Quantitative Precipitation Forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638-663.

Mullen, S.L. and Buizza, R., 2002: The Impact of Horizontal Resolution and Ensemble Size on Probabilistic Forecasts of Precipitation by the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **17**, 173-191.

Mureau, R., Molteni, F. and Palmer, T.N., 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart. J. Roy. Meteor. Soc.*, **119**, 299-323.

Mylne, K.R., 2002: Decision-making from probability forecasts based on forecast value. *Meteor. Appl.*, **9**, 307-315.

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 Joint Session of 16th Conference on Probability and Statistics in the Atmospheric Sciences and of Symposium on Observations, Data Assimilation, and Probabilistic Prediction, AMS, 13-17 January 2002, Orlando, Florida, pp 113-118.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649-667.

Palmer, T.N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128,** 747-774.

Smith, L. and Gilmour, I., 1999: Accountability and internal consistency in ensemble formation. In 'Proceedings of Workshop on Predictability, 20-22 October 1997'. ECMWF, 1999.

Stanski, H.R., Wilson, L.J., and Burrows, W.R., 1989: A Survey of Common Verification Methods in Meteorology. WMO WWW Tech. Report No. 8, WMO TD No. 358. 114pp.

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

Wilks, D.S., 1995: Statistical Methods in the Atmospheric Sciences - An Introduction. International Geophysics Series Vol. 59, Academic Press. 467pp.

Young, M.V. and Carroll, E.B., 2002: Use of medium-range ensembles at the Met Office II: Applications for medium-range forecasting. *Meteor. Appl.*, **9**, 273-288.