Numerical Weather Prediction

First-Guess Early Warnings project Report on the Verification System

Forecasting Research Technical Report No. 364

Tim Legg and Ken Mylne

email: nwp_publications@metoffice.com

First-Guess Early Warnings project
Report on the Verification System
November 2001

Tim Legg and Ken Mylne

Abstract

The First-Guess Early Warning (FGEW) project is aimed at giving NMC forecasters timely probabilistic guidance on the occurrence of severe weather events, in support of National Severe Weather Warning Service early warnings. FGEW output is based on forecast data from the ECMWF Ensemble Prediction System (EPS). Model event thresholds have been adjusted based on data from winter 2000/01 to optimise the system.

This report gives skill assessments of warnings generated by the FGEW software, before and after optimisation, compared to warnings issued by NMC forecasters, using a variety of verification measures. The EPS data have greatest skill at 4 days ahead, and have only a very limited ability at 1-3 days ahead to discriminate between high and low probabilities. NMC forecasters' warnings are skilful up to 2 days ahead, but the FGEW system is capable of providing forecasters with useful guidance 4 days ahead. This should encourage the issue of Early Warnings longer in advance of severe events, thus fulfilling one of the aims of the project.

1. Introduction

The principal aim of the First-Guess Early Warnings (FGEW) project is to give NMC forecasters timely guidance, in probabilistic form, on the occurrence of certain defined severe weather events. This is done by looking at and interpreting forecast output from the ECMWF Ensemble Prediction System (EPS). The purpose of this is to encourage the issue of Early Warnings at longer lead-times, as part of the Met Office National Severe Weather Warning Service (NSWWS). To establish the extent to which this aim is met, forecasts are assessed against whether Flash Warnings are subsequently issued for the same event. This assessment is performed in the manner set out by Mylne (2000a).

This report focusses on the following events, defined for use with the NSWWS: (i) severe gales (gusts of at least 70mph), (ii) heavy snowfall (at least 4cm depth accumulating within 2 hours), and (iii) heavy rainfall (15mm or more precipitation falling within a 3-hour period). These are not hard-and-fast definitions; rather, warnings are issued when weather conditions are expected to endanger or seriously inconvenience human activity.

Automatic software has been running since the summer of 2000, and a trial began with NMC forecasters on 11 September 2000. Estimated probabilities are generated for the defined severe weather events on a regional basis within the UK, based on EPS forecast values at grid-points. However, it was recognised at the outset of this project that it might be unusual for the EPS to give high forecast probabilities of these severe events. During the trial, forecasters were able to use these probabilities when considering the issue of an Early Warning message. Forecasters are free to issue these Early Warnings independently of when the automatic software suggests issue of a warning, and have recourse to all other forecast data from different models etc. which may influence their decision as to whether or not a warning is issued.

The aim of this report is to demonstrate whether the warnings generated by the FGEW software have better skill than those issued by NMC forecasters. The skill of probability forecasts based on each of these is compared with the strategy of always issuing 'null' forecasts, which are a convenient and useful reference standard to use for such rare events. Verification tools used include Reliability Diagrams, Brier Skill Scores, Relative Operating Characteristics, and Correct-Alarm Ratios / Miss Rates. For verification purposes, an event is deemed to occur if a Flash Warning (a very-short-range warning of severe weather, issued with high confidence) is issued – see Mylne (2000a). Proposals for altering the FGEW system to use the EPS data to maximise the forecast skill are also made in this report.

Care has to be exercised to ensure that this verification is done correctly and in a meaningful way. For a given day and for a given possible Severe Weather event, basically four outcomes are possible: an Early Warning was/was not issued, and the event did/did not occur. From this information, contingency tables are drawn up and used for further analysis, on an event-by-event basis. We have strived to avoid any inappropriate 'double-counting' of events in our verifications.

Since the ECMWF EPS was upgraded to run at T_L255 resolution as from 21 November 2000, only 200 days of verification data had become available by early-May 2001. (This includes five weeks of data which were back-run for October/November to provide an overlap period for testing and comparison with T_L159 ensemble data as was formerly used.) During this time there were 8 episodes of severe gales, 12 of heavy snowfall, and 20 of heavy rainfall (these episodes are defined as periods during which Early and/or Flash Warnings were issued, and sometimes extend over two or more days). (Note that included within this period were noteworthy severe events of gale damage and flooding caused by weather systems which crossed the UK area on 30th-31st October and again on 4th-6th November 2000.) Thus, the sample-size from which results are drawn is limited, and our

conclusions are likely not to be statistically significant. Over the course of time, this situation should improve, but shortage of data will always be a problem for probabilistic verification of severe events.

Furthermore, a bug in the EPS, which led to the ensemble having less spread than it ought to have had, was only diagnosed in January and corrected as from 6 February 2001. This EPS bug caused the initial perturbations to be too small, so the ensemble will have had insufficient spread during this period, especially at shorter time-ranges. The effects of this on our work are difficult to quantify but will be mentioned again below. One likely effect would be over-confidence, because many events would be (unrealistically) forecast with very low or very high probabilities, the ensemble values all falling within too narrow a range. It will be interesting to see whether the distributions of probabilities of events are more evenly spread in future. The severe storms of October and November 2000, included in the period affected by the spread bug, were predicted with very high probabilities, which may have helped give favourable results in terms of verification. However, when these dates were re-run as test cases after correction of the bug, probabilities of severe conditions were still quite high although slightly reduced.

The bug means that the verification results presented here may not be fully representative as they cannot truly reflect the outcome as it would have been if the EPS had been running correctly throughout the period studied. Unfortunately we will have to wait until the next autumn/winter season to be able to assess the performance of the system using correctly formulated ensemble data, as warnings during the spring and summer will be few, and hence more fine-tuning of the sensitivity of the system will be required after that (i.e. in mid-2002). We must also wait until then to assess the sensitivity of the system following the revisions recommended below, against independent data.

Another factor which we have borne in mind throughout this project is the limitations of model output. Output is obtained only at a network of grid-points (roughly 80km apart), and it is impossible for the model to fully resolve processes occurring on smaller scales. Sub-grid-scale processes can be important, especially for localised precipitation events. Hence we always expected to set the event thresholds when looking at EPS model output at values below those defined in the NSWWS requirements. Also, because 6-hourly EPS output is used, we have to estimate the occurrence of events whose NSWWS definitions cover shorter periods (heavy snowfall and heavy rainfall) as best we can. These limitations are discussed in more detail by Legg and Mylne (2000).

Detailed results are presented in this report for warnings of severe gales, heavy snowfall, and heavy rainfall events. Few warnings of blizzards were made, some of these coinciding with snowfall warnings, and hence results for blizzards would be similar. The rarity of all of these events means that as yet we have an insufficient sample of data to give truly representative results, and indeed we may never have sufficient. Nevertheless, the results to date show that the FGEW forecasts do have some limited potential skill, especially on the third and fourth days of the forecast period. It must be stressed, though, that these results give an approximate upper limit to the level of skill that the system is likely to achieve, as we are assessing using the same data as were used to obtain the event thresholds. Only when the next winter season is complete will we be able meaningfully to assess the system using truly independent data.

The FGEW forecasts are being used by NMC forecasters as an aid to providing Early Warnings further in advance of severe weather events, which was one of the stated aims of the FGEW project. An Early Warning of heavy rain and severe gales was issued on 2 November 2000 extending to five days ahead, which is the furthest ahead that an Early Warning has ever been issued by the Met Office. Based on feedback received from NMC forecasters, and our detailed study of the numerical data over this past winter, changes will be made to the event thresholds and the use of time-windowing which will improve the quality and skill of the warnings generated by the system, and recommendations

are presented below for what event thresholds and time-windowing should be used in the operational system during Winter 2001/2002.

In the following sections of this Report, the phrase "operational FGEW system" will be used to refer to the system as it was during the winter 2000/01 (operational trial) period, and the phrase "optimised FGEW system" will refer to the system as we propose to operate it for the 2001/02 winter period. Section 2 gives a general introduction to contingency tables, which form the basis of most of the assessment work. Sections 3-7 cover assessment of the operational FGEW system during the past winter season. Section 8 then explores possible ways of improving the skill of the system, and gives recommendations for an improved 'optimised' version of the system, including supporting verification information. Section 9 concludes.

2. Contingency Tables of events

Contingency tables can be drawn up based on whether or not an event was predicted to occur, and whether or not it did occur (e.g. Table 1). The four possible contingencies as shown in Table 1 can be thought of as 'hits', *H* (event correctly forecast), 'misses', *M* (event occurred but not forecast), 'false alarms', *F* (event was forecast but did not occur), and 'correct rejections', *C* (event neither forecast nor occurred). In the context of this work, 'forecast' means that an Early Warning was issued, and 'observed' means that a Flash Warning was issued, as explained below.

Table 1. Two-by-two contingency table of events

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

Such a contingency table can be obtained for each of a series of probability levels. The forecast is assigned as a 'yes' if the probability was equal to or greater than the probability level. Due to the nature of this project, which is concerned with low-probability severe events, a non-standard set of probability thresholds has been used throughout: 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. This can be done for any individual area of the UK or for probabilities of an event occurring anywhere within the UK, for any defined Severe Weather event. Contingency tables are made up from sampling warning events as described in the verification design specification (Mylne, 2000a). Thus our results are based on a case-by-case treatment, and any day during which no warnings of any kind (Early or Flash) are valid is treated as a 'Correct Rejection'. An unbiased system, i.e. one which predicts an event the correct proportion of the time, will have a mean forecast probability equal to the sample mean frequency of occurrence. A perfect system would produce no Misses or False Alarms at all.

In the verification plan (Mylne, 2000a) it was suggested that, to allow for timing errors, an Early Warning should be deemed successful even if its period of validity did not quite match the period during which the event actually occurred (i.e. the period of validity of the Flash Warning), so long as there was an overlap in the geographical area covered by the warning and it was for the same kind of event. The allowed mismatch in timing is greater at longer forecast ranges (Table 2, in which the

'forecast time' is measured from the time of issue of the Early Warning to the beginning of the validity period of the Flash). Verification has been repeated both with and without this allowance.

Table 2. Permitted interval between Early and Flash Warnings

Forecast time	Permitted interval
< 24h	0
≥ 24h, < 48h	6 hours
≥ 48h, < 72h	12 hours
≥ 72h	18 hours

The actual occurrence of a severe event is difficult to determine objectively, so an event is deemed to have occurred in any period and geographical area for which a Flash Warning is valid; this is possible because Flash Warnings are very-short-range warnings issued when any such event is imminent or already occurring, making them suitable for use as a proxy in this way. This eliminates any need to use model analyses, which have severe limitations especially for small-scale events, for verification purposes. 'On-screen analyses' were not used, as the solution of using Flash Warnings was simpler but more effective. Use of radar products to assess precipitation warnings was discounted, because the effort required to set this up was not considered worthwhile. (Flash Warnings themselves are being separately verified in another project, the Automated Verifications of Warnings project.) Note that it is possible that a warning subjectively assessed as correct, because disruption occurred, could be regarded as incorrect by the FGEW objective assessment system if for example heavy rain occurred but the original warning was for snowfall.

A further complication in the assessment is that we must avoid inappropriate 'double-counting' of particular events. This may happen if, for example, an Early Warning is valid at the end of one day and a Flash Warning is issued early on the following day. In this case, these warnings might cover the same weather event but give rise to both a Miss and a False-Alarm, which is clearly inappropriate. Such events are ascribed to the earliest appropriate day (in this example, the day of validity of the Early Warning) and counted only once. We must avoid different outcomes arising if, on separate occasions, similar combinations of warnings are given but at different times of day. On the other hand, if an Early Warning is valid for a period of two days but a Flash is issued for only one of these days, then we would count a Hit and a False-Alarm so as to penalise the overly-long Early Warning. If an Early Warning justifiably covers a two-day period then it would be rewarded with two Hit counts.

3. Relative Operating Characteristic

Relative Operating Characteristic (ROC) is an assessment in terms of Hit Rate and False-Alarm Rate. Referring back to Table 1, the Hit Rate is $H/(H+M)$ and the False-Alarm Rate is $F/(F+R)$. (Note that this definition of False-Alarm Rate gives the number of false alarms as a proportion of all non-occurrences of the event, i.e. we are stratifying by observation. This is the standard definition for ROC, and differs from the more commonly-used False-Alarm Rate which is stratified by forecast.) The Hit Rate indicates the proportion of occurrences of an event that are successfully forecast. Ideally, we would get a high Hit Rate and a low False-Alarm Rate. These quantities, which are both stratified by observation, can be calculated for any forecast probability threshold (from 0.0 to 1.0 inclusive) by assuming that the event is predicted if its forecast probability exceeds that threshold. Hence a set of

values of Hit Rate and False-Alarm Rate can be obtained and plotted on a graph, to obtain the ROC curve. For a skilful system, the ROC curve is bowed towards the upper-left part of the graph. A useful measure of skill is the area under a ROC curve, which would be equal to 0.5 for a skill-less system (in which Hit Rate and False-Alarm Rate would be equal for any given probability threshold, as the system has no ability to tell us when the event will/will not occur) and 1.0 for a perfect set of forecasts. The 'ROC area' is indicated on each of the graphs here. ROC measures the ability of forecasts to discriminate between when events do and do not occur, which is useful for decision-making applications. As noted above, we have used a non-standard set of probability thresholds in our verification, because these severe events are frequently forecast to have low probabilities, and hence there are more points towards the top-right of each ROC curve than would be obtained using standard probability thresholds at intervals of 0.1 throughout. This leads to a smoother curve, and a larger area being measured under the ROC curve, thus making the FGEW system seem more skilful than it otherwise would, but helps to show the ability of the system to discriminate between occurrences and non-occurrences of an event at low probabilities.

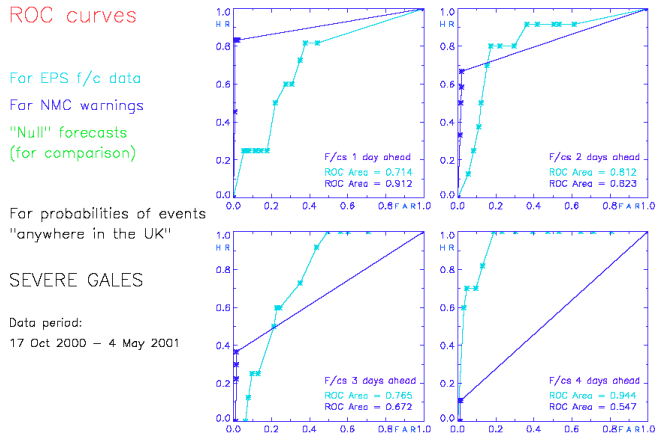
ROC for the operational FGEW system

The ROC curves shown in Figs 1-3 are for probabilities in individual UK areas, for Severe Gale, Heavy Snowfall and Heavy Rainfall events respectively, for forecasts one to four days ahead (D+1 to D+4). All except D+1 have ROC areas well in excess of 0.5. This indicates that the system has some ability in terms of resolution, and produces useful probabilistic information that can be used for decision-making, though the lack of smoothness in the curves, notable especially for probabilities anywhere in UK, is a testament to the inadequate sample sizes. The degree of resolution demonstrated here increases from 1 to 4 days ahead, and it is clear that the system performs best at D+4. Performance tails off again beyond D+4 (not shown). This skill is due to the system's performance within the low end of the probability range (represented by the points near the top-right of each ROC curve, which represent probability thresholds of, in order, 0.0, 0.01, 0.03, 0.05, 0.09, 0.13, etc.). Performance is poor at D+1 (forecasts have almost no apparent resolution, due mainly to insufficient ensemble spread), especially for probabilities of events in individual areas.

The superior performance of the FGEW system at D+4, compared to other days, is worthy of note. This appears strange at first sight, but we think we can explain why. The perturbations used in the ensemble are designed to maximise the error growth rate over the first 48 hours. Due to this rapid growth rate, the initial perturbations are very small in order to give the correct ensemble spread at 48 hours, resulting in deficient spread at less than 48 hours ahead. Because the spread is maximised at 48 hours, the ensemble cannot be a random sample of the possible atmospheric states at that time, and therefore cannot be expected to give reliable probabilities of events. It is only when the effects of non-linearity have had an opportunity to take more effect beyond 48 hours that we can expect a quasi-random sampling of the distribution, and hence reliable probabilities. The 'spread bug' may have exacerbated these effects also.

In the assessments presented in this report, Early Warnings have been deemed successful even if the period of validity did not quite match the period during which the event actually occurred, with a permitted mismatch of up to 18 hours (Table 2). However, one point of note is that ROC curves for assessments re-calculated with no such mismatch permitted were poorer, especially for forecasts 3 days ahead, with a corresponding reduction in ROC areas for all cases (e.g. Fig. 4 for Heavy Rainfall warnings). No such difference was found for any of the other verification measures, reported later. All results presented hereafter were obtained with the mismatch of up to 18 hours allowed.

Fig. 1 (a)



(b)

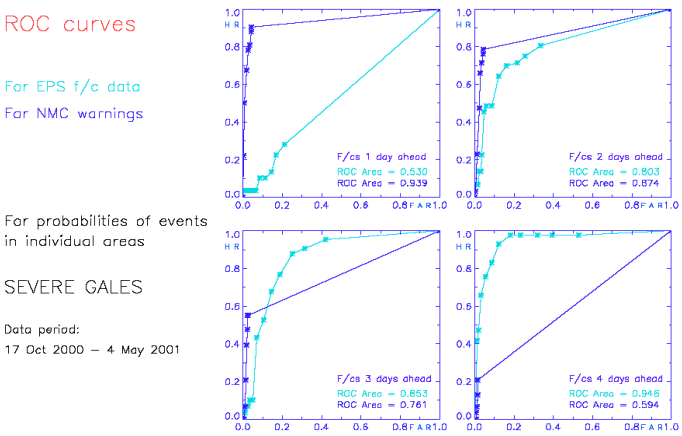
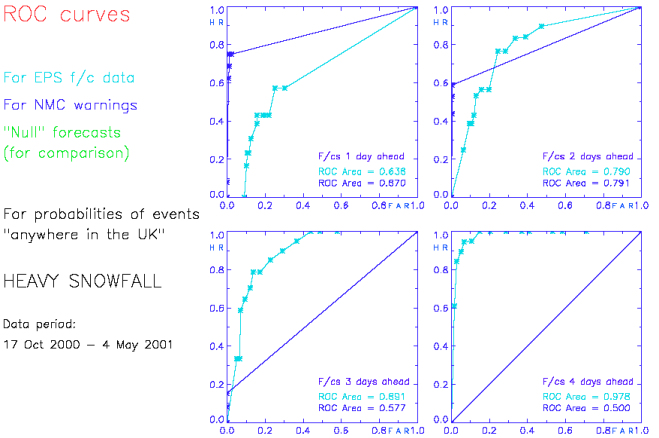


Fig. 2 (a)



(b)

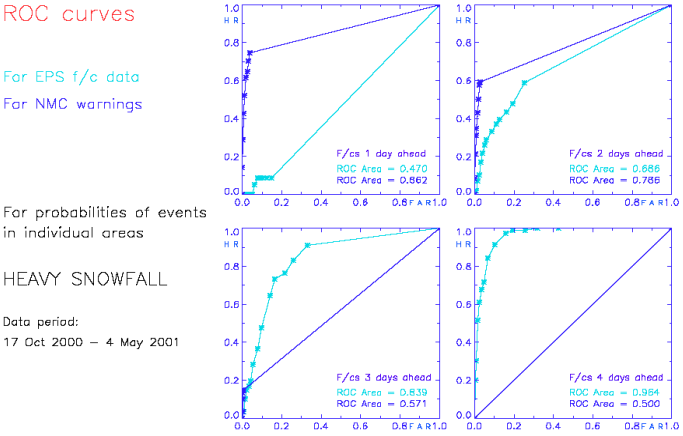
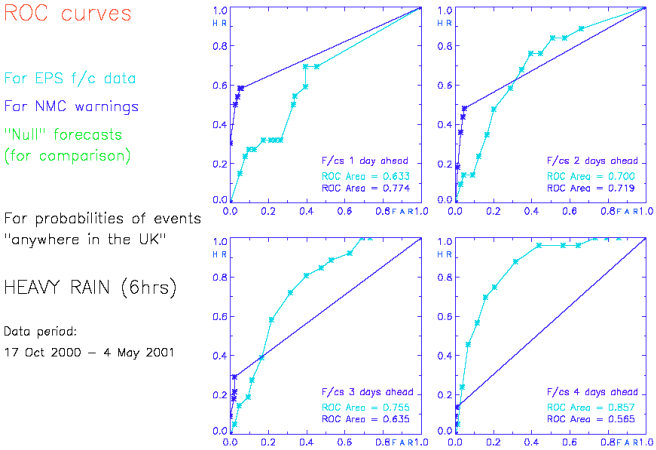
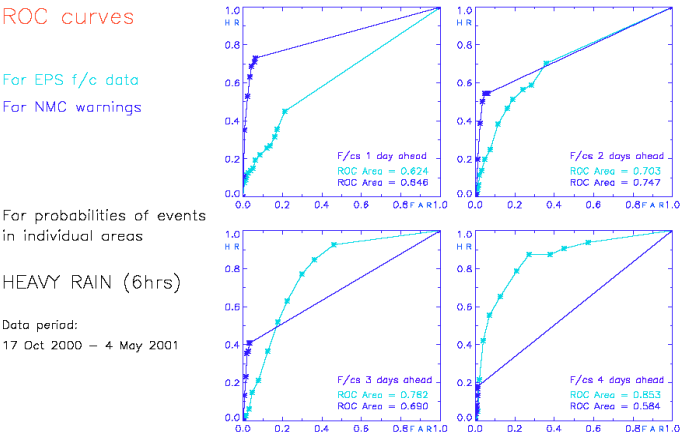


Fig. 3 (a)



(b)



Figs 1-3 Relative Operating Characteristic curves, for 1, 2, 3 and 4 days ahead, for Severe Gale warnings (Fig. 1), Heavy Snowfall warnings (Fig. 2), and Heavy Rainfall warnings (Fig. 3), from the operational FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring (a) anywhere in the UK, and (b) in individual areas.

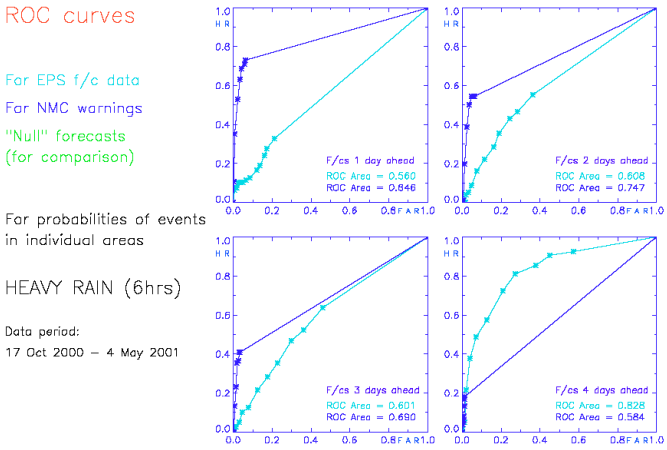


Fig. 4 Relative Operating Characteristic curves, for 1, 2, 3 and 4 days ahead, for Heavy Rainfall warnings, from the operational FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring in individual areas, but with no mismatch allowed between Early Warning time and verifying Flash Warning time. (Cf. Fig. 3b)

4. Reliability Diagrams

For an ideal probabilistic forecasting system, out of all occasions when a probability of *x*% is assigned to an event, that event will occur on *x*% of occasions. By determining for each value of *x* (binned into a series of finite ranges) the proportion of occurrences (the number of times the event was forecast and did occur divided by the total number of times the event was forecast, i.e. *H/(H+F)* in Table 1), a "reliability diagram" can be generated. The ideal reliability curve takes the form of a straight line along the diagonal *y=x*. Reliability curves can be useful in highlighting certain shortcomings – for example, graphs which stray below the *y=x* line are symptomatic of a system which overestimates forecast probabilities. The resolution capabilities of a forecasting system are indicated by the slope of a reliability diagram, for example a system with no resolution would produce a roughly horizontal reliability curve. It is also useful on such a diagram to include an indication of how often each probability 'bin' was forecast ('sharpness'), as has been done on the diagrams which follow.

Note once again that we have used a non-standard set of probability thresholds in this work, with a greater number of 'bins' for low probabilities. This enables us to explore the low end of the probability range more closely, and avoids over-populating the probability 'bins' for p=0.0 and p=0.1.

The reliability of the operational FGEW system

Reliability curves are shown for forecasts at 1, 2, 3 and 4 days ahead, based on probabilities of events occurring anywhere within the UK, for severe gales (Fig. 5), for heavy snowfall events (Fig. 6), and for heavy rainfall events (Fig. 7). Similar graphs for probabilities of events occurring in individual areas are shown in Figs 8-10. These pairs of graphs are designed to compare the performance of the FGEW automatic scanning system (based on EPS output) (on the right in each pair) with that of Early Warnings issued by NMC forecasters (left). We shall discuss the FGEW warnings first, and then consider the NMC warnings.

The reliability curves derived from data used in the operational FGEW system are clearly very noisy (especially for the whole-UK results in Figs 5-7), and at first glance the forecasts appear poor. However there are some positive results. Bearing in mind that ROC results were far better at D+4 than at shorter times, consider first the D+4 results in part (d) of the figures. In these cases there is a clear slope with higher forecast probabilities correctly indicating an increased frequency of occurrence, and the event

never occurs when the lowest probabilities are issued. Thus we have good resolution of whether there is any risk, and higher probabilities when the risk is higher. However, it is also noticeable that the system is seriously over-forecasting severe weather since the curves all fall well below the ideal dotted diagonal line; this is also indicated by the figures under the graphs which show that the average forecast probability is very much higher than the sample climatology. Looking at D+2 and D+3 forecasts the results are much less encouraging. There is no significant slope from low to high probabilities, and the most we can say is that in most cases the events are rare when the probability is zero. Where the event does sometimes occur with zero forecast probability, its frequency is well below the sample climatological frequency given by the horizonal dotted line. Correspondingly for the higher probabilities the frequency of occurrence is usually well above the climatological frequency. (This is more apparent in Figs 8-10 for the individual areas than for the whole-UK results in Figs 5-7.) Thus the system is usually able to discriminate between a zero risk and some risk, but no more. For D+1 forecasts even this is not true, but this is perhaps not surprising given that the perturbations should not be expected to have grown large enough at this stage of the forecast. Beyond D+4 (not shown), the skill of the system declines as might be expected, though not as fast as the increase in the first four days. The over-estimation of EPS probabilities becomes more severe at

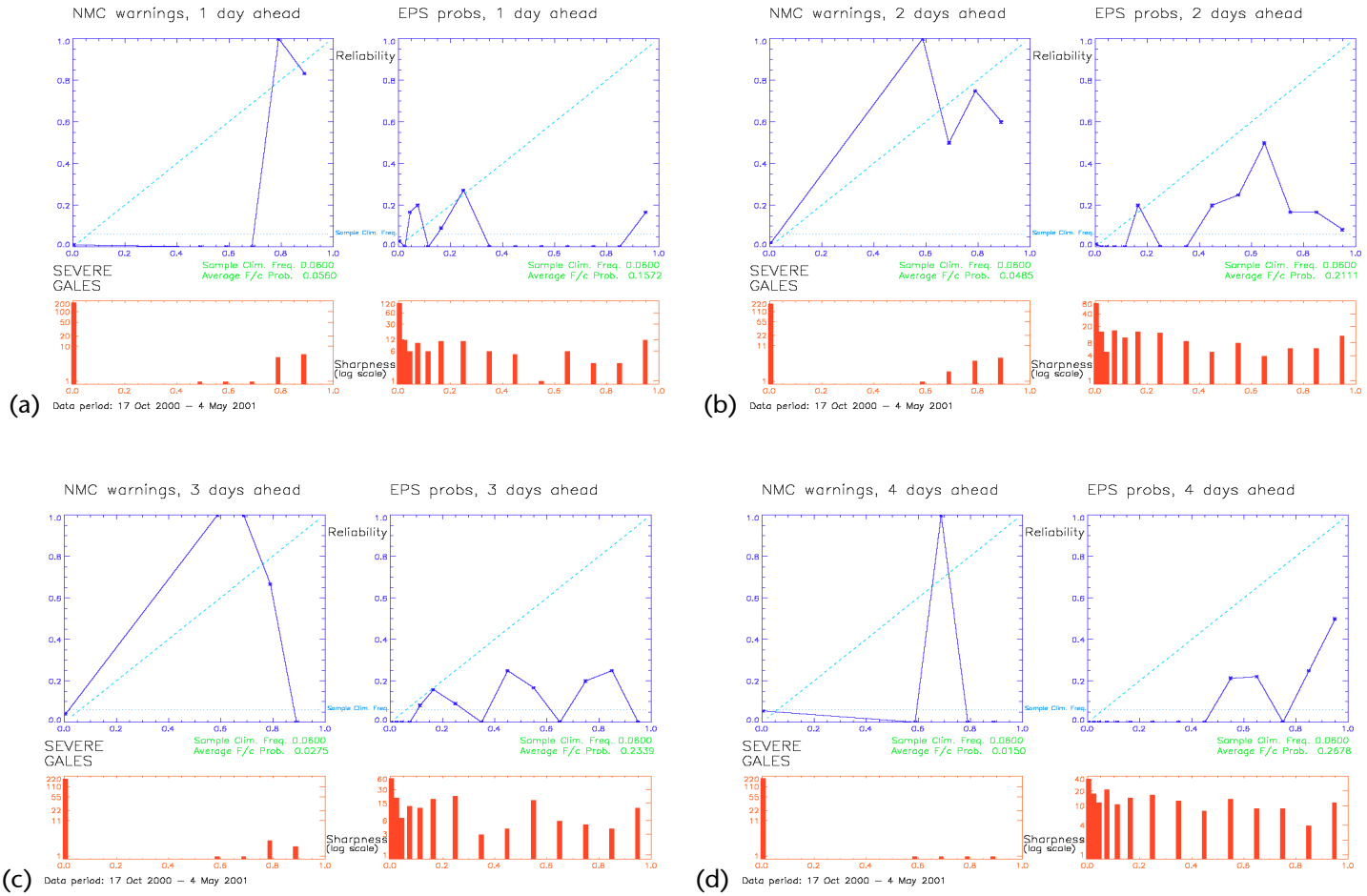


Fig. 5 Reliability curves for Severe Gale warnings, from (left of each panel) NMC warning probabilities of events and (right of each panel) probabilities from the operational FGEW system, of events occurring anywhere in UK, at (a) 1, (b) 2, (c) 3, and (d) 4 days ahead. Sharpness diagrams are also included, at the bottom of each panel, with logarithmic y-axis scaling.

longer forecast-times, and this is because of the time-windowing which becomes wider with time. The system has a clear tendency to show over-forecasting in almost all instances; in other words, mean forecast probabilities are higher than the corresponding sample climatological frequencies (i.e. incidence of issued Flash Warnings).

Probabilities of events in each of the 12 individual UK areas must always be lower than or equal to the probabilities anywhere within the UK. Issued warning probabilities for individual areas can be as low as 20% on an Early Warning message, so long as the probability 'anywhere in UK' is 60% or more. The sharpness graphs in Figs 8-10 are concentrated more towards the left (lower probabilities) than those for the whole UK. There are twelve times as many data-points contributing to these graphs, and hence the data sample size is much larger and more representative, although these data cannot all be statistically independent. The reliability graphs of events occurring in individual areas are smoother than those for probabilities anywhere in the UK, partly due to a greater number of data points being available, but otherwise the results are very similar. Note once again that the results for forecasts at D+4 are better than those for other days.

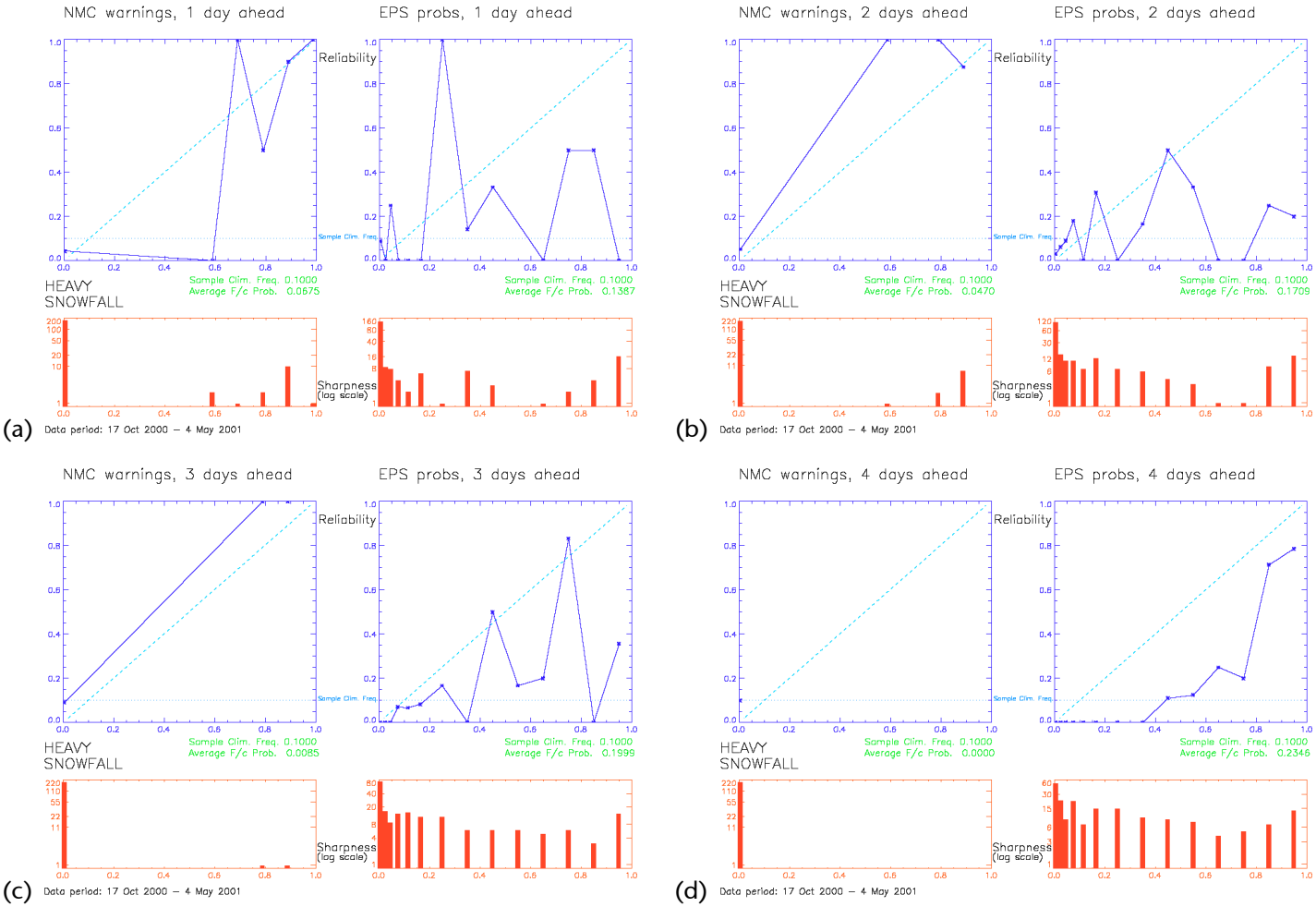


Fig. 6 As Fig. 5, but for Heavy Snowfall warnings.

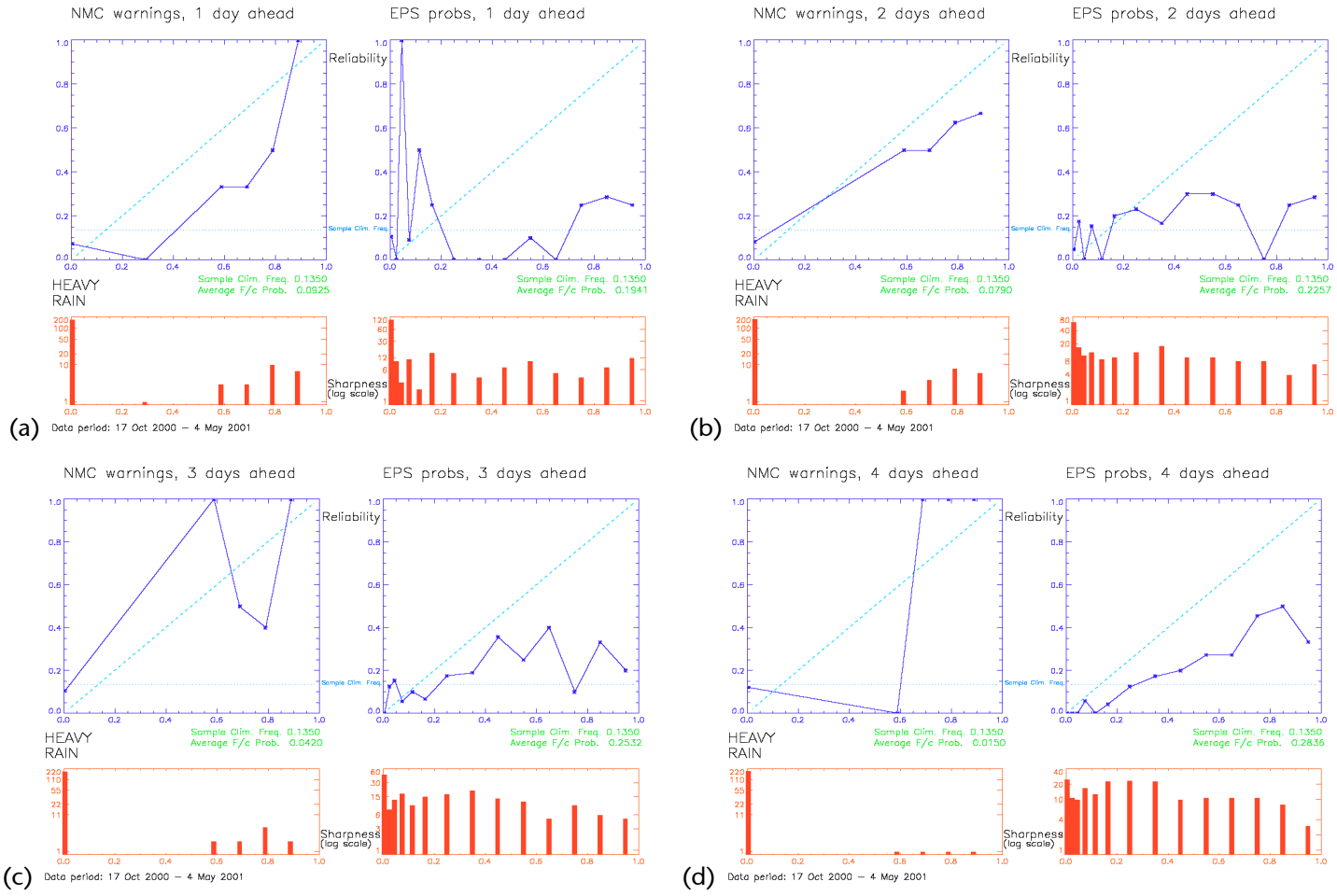


Fig. 7 As Fig. 5, but for Heavy Rainfall warnings.

The deficient ensemble spread during the first part of the period studied will undoubtedly have had some effect on these results, and on the calibration of the system. We have already noted a tendency for our system, tuned as it was for the operational trial, to produce event probabilities which were too high on average (over-forecasting). There has also been a tendency towards over-confidence, which is a feature one would expect of any ensemble system having insufficient spread i.e. the range of values covered by the ensemble is too narrow. Any attempt made to alter the system to compensate for this might be found, in future when verified on independent data, to have over-compensated, and the ensemble system might then therefore appear to be under-confident. (Tuning of event thresholds can only really be used to cure over/under-forecasting. Over-confidence can only be corrected by calibration of probabilities, but this can only be done if the forecast probabilities have a sufficient degree of resolution, and if we have a large enough sample size, so we have not attempted to correct for this.) It is noteworthy that the best results are for D+4. With the corrected ensemble spread, it is speculated that forecasts at D+3 may have sufficiently greater spread that over-confidence might no longer occur, and we suspect that the calibration may then be found to have over-compensated at D+4. We stress again at this point that, in the light of verification for the coming (2001/02) winter, further adjustments to the calibration of the system are likely to be made, ready for the winter after that.

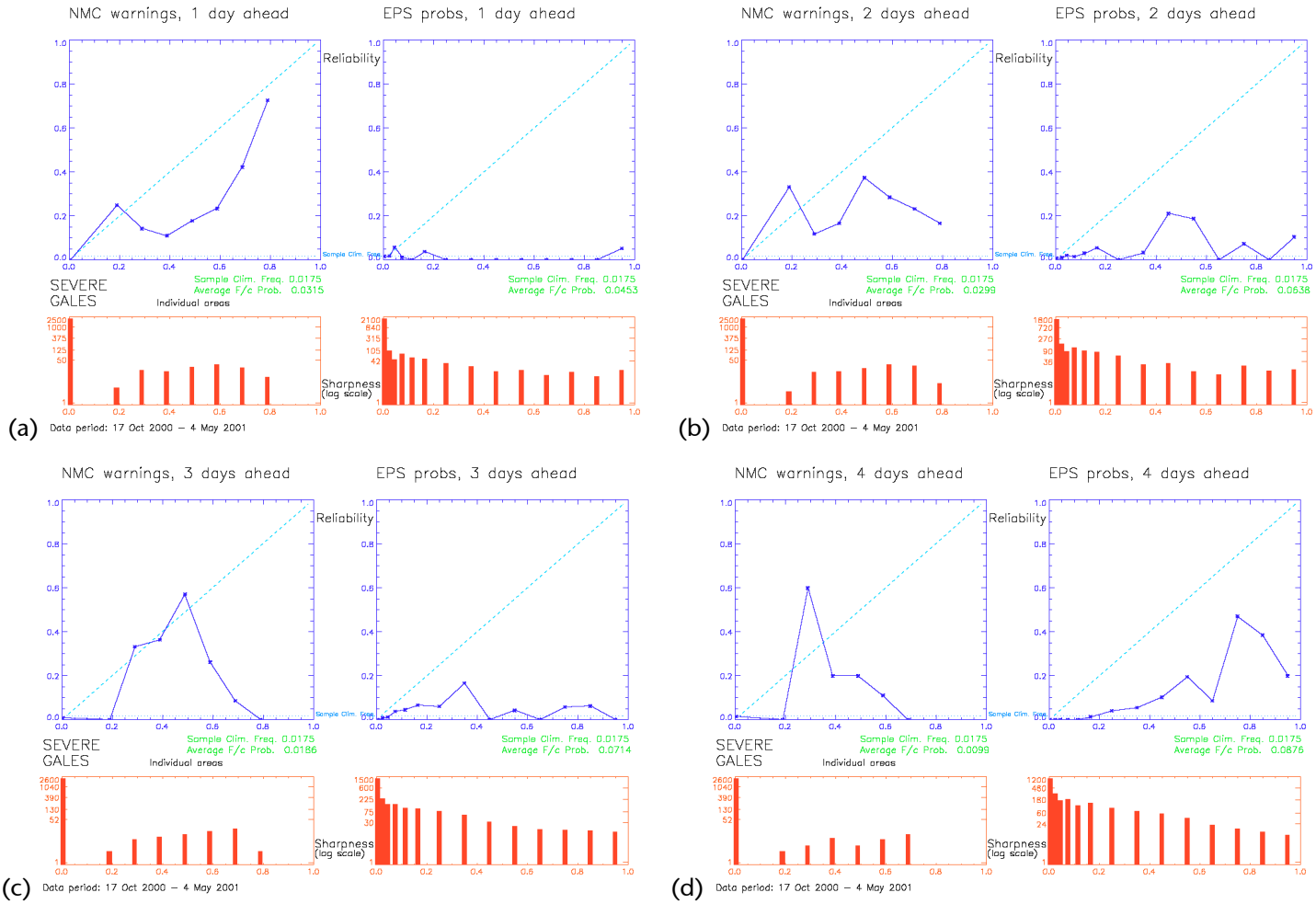


Fig. 8 Reliability curves for Severe Gale warnings, from (left of each panel) NMC warning probabilities of events and (right of each panel) probabilities from the operational FGEW system, of events occurring in individual areas, at (a) 1, (b) 2, (c) 3, and (d) 4 days ahead. Sharpness diagrams are also included, at the bottom of each panel, with logarithmic y-axis scaling.

Reliability of NMC issued Early Warnings

Early Warnings are only issued by NMC when the probability somewhere in the UK is at least 60%, therefore the corresponding graphs contain no points between 0% and 60% forecast probability (except for a small number of cases when existing warnings were re-issued in downgraded form).

The tendency of the FGEW reliability curves to show over-forecasting in many instances, as described above, is shared to some degree by the corresponding results from NMC warnings. Average probabilities of NMC warnings for anywhere in the UK are close to or a little below the sample climatological frequency of each of the events at D+1 and D+2, and fall lower at longer ranges because fewer warnings are issued; indeed very few warnings have been issued by NMC more than three days ahead. This is clearly seen on the reliability diagrams, which show that many events occurred even when no Early Warning was issued. The proportion of events occurring given zero probability is close to zero for Severe Gales at 1 and 2 days ahead, but higher in other instances, and approaches the sample climatological frequency at days 3-4 as very few warnings were given and most occurrences of the event were missed. There is some room for improvement here, although for a probabilistic system we should always expect at least a small proportion of events to be missed.

Over-forecasting in the NMC warnings is apparent for individual area probabilities, suggesting that, while the number of Early Warnings overall is about right or slightly too low, the number of areas included in many of these warnings is too many or else the probabilities for individual areas are too high. On average, for a weather event during which Flash Warnings are issued, these warnings cover only approximately 4 of the 12 UK areas. This suggests that individual-area probabilities need to be lower even though the anywhere-in-UK probabilities should be little changed. Lower probabilities over several individual areas could also be used to reflect forecasters' uncertainty regarding the areal extent of an impending severe weather event.

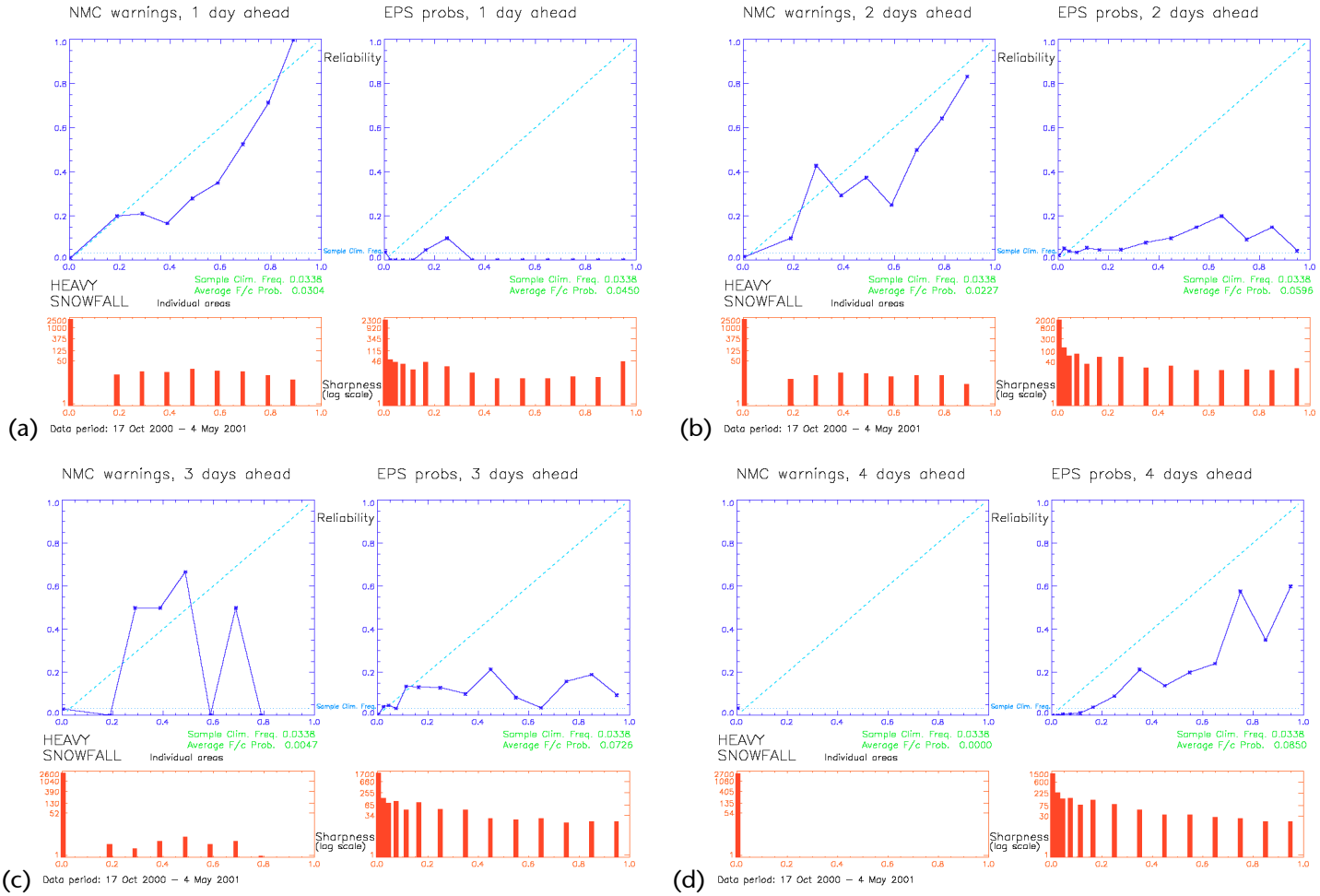


Fig. 9 As Fig. 8, but for Heavy Snowfall warnings.

Discussion

One use of a reliability curve is to calibrate the system in a way that corrects for any apparent bias in the forecast probabilities. For example, if, when the system gives a probability of 70%, the event only verifies on 50% of occasions, the forecast probability itself would be modified to 50%. This approach could be used to correct for any over-confidence, which would result from an ensemble which has insufficient spread. (A symptom of such over-confidence is a reliability curve whose slope is too shallow, with too many forecasts having probabilities close to 0 or 1, and many occurrences of events which were forecast with low probability and non-occurrences of events which were given high probability.) But we could only use them for calibration if the reliability curve is monotonic, and the results we have shown here include many which are not, partly due to the data sampling problem, which itself means that we have too few data for such a method of calibration to be considered robust. For these reasons, this approach has not been followed further. However, the mean forecast probabilities have been used to determine by how much the event thresholds should be altered so as to reduce the bias of the probability forecasts, i.e. we can correct for over-forecasting, but not for any over-confidence in the forecast probabilities. However, we suggest that a slight degree of over-forecasting be allowed, as we wish not to have too many 'missed events', while being careful to ensure that forecasters aren't presented with too many false alarms recommending them to issue Early Warnings.

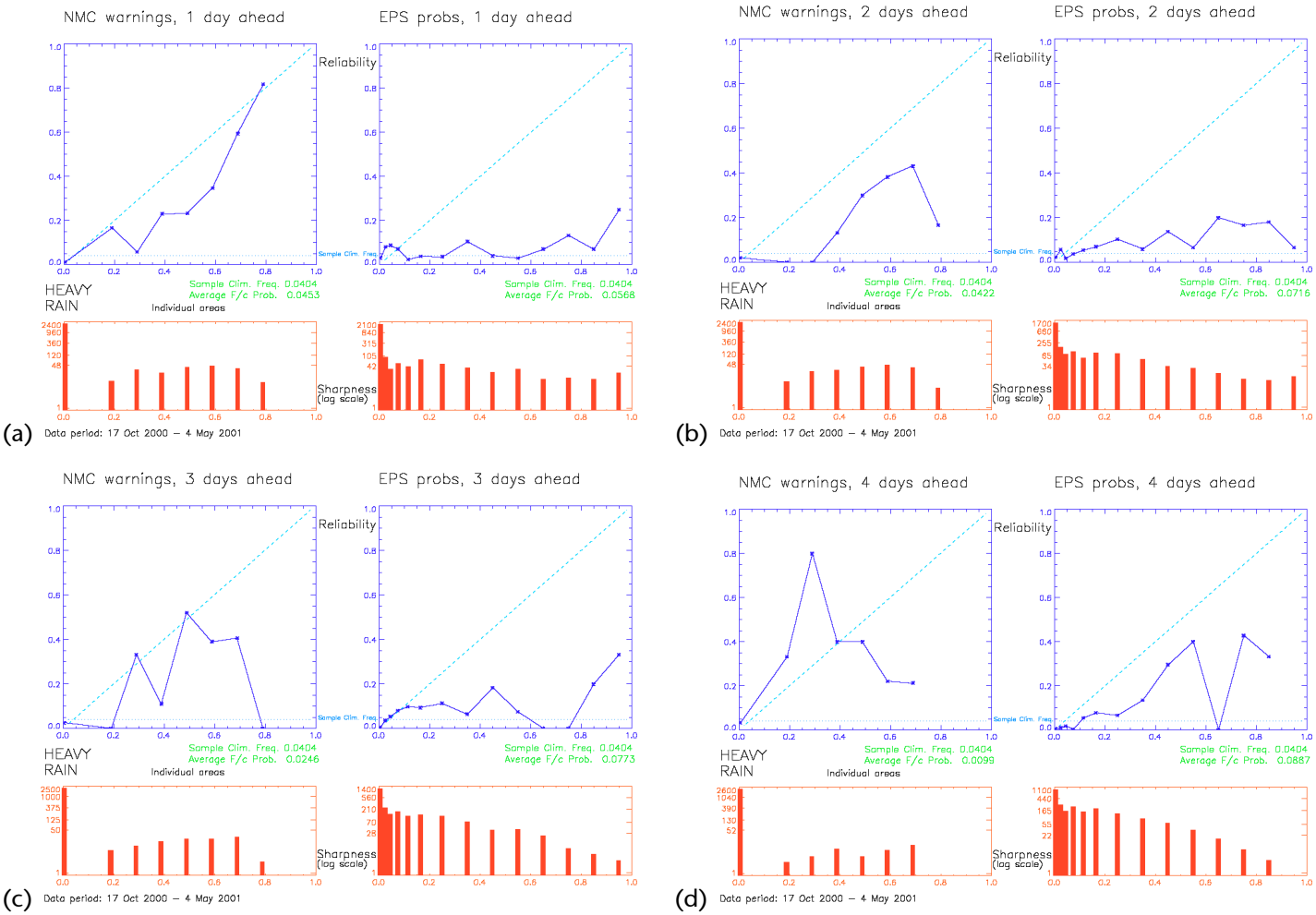


Fig. 10 As Fig. 8, but for Heavy Rainfall warnings.

5. Brier Scores and Brier Skill Scores

The Brier Score *BS* is a measure of mean square error in a probability forecast. It is defined as the average value of the square of the difference between the forecast probability and the observation (Equation 1). Note that the observation can only be 1 or 0, depending on whether the defined event occurred or not. Brier Score is bounded by the values 0.0 and 1.0; a lower value represents a better forecast. Denoting the total number of forecasts by *N*, and the forecast probability and observation by p_f and p_o, the Brier Score *BS* is calculated as

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

However, comparing Brier scores for different events is not meaningful if their climatological probabilities are different. This fact can be seen by considering two events, one of which occurs on average half of the time, and the other of which has a mean occurrence rate of 0.1, say. If the forecast is randomly distributed between 'yes' and 'no', while preserving the sample climatological probability over many cases, then the Brier score for the more common event will be larger than that for the rarer event. The apparent conclusion that, even for a random and hence unskilled system, we are worse at forecasting the more common event, is due solely to the climatological probabilities being different.

To avoid this problem, we calculate the Brier Skill Score (BSS). This is obtained by comparing the Brier Score of our forecasting system with that obtained by some reference forecast (such as climatology, persistence, or a strategy of always forecasting zero probability) (Equation 2). As many of the events in FGEW are rare events, the strategy we have used is to compare issued forecasts with null-probability forecasts, to determine whether the forecasts are better than a hypothetical 'fall-back' strategy of never issuing warnings. Brier Skill Score *BSS* is calculated from the Brier score of our forecasts BS_{fc} and that of the reference forecast system BS_{ref} as

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

Brier Scores and Brier Skill Scores have been calculated for forecasts up to 6 days ahead, again comparing automatic forecasts based on EPS output with Early Warnings issued by NMC. The Skill Scores are calculated relative to 'null' (zero probability) forecasts.

Brier Scores and Brier Skill Scores of the operational FGEW system

Figs 11-13 show Brier scores and Brier Skill Scores at forecast times up to 6 days ahead for the operational FGEW system and for NMC-issued Early Warnings, for probabilities of events occurring anywhere in the UK. NMC Warnings have rarely been issued more than three days ahead, which means that Brier Scores are equal to those of 'null forecasts' and Brier Skill Scores are zero, though these NMC Warnings do have skill at shorter ranges. The scores for FGEW are heavily influenced by the tendency for forecast probabilities to be too high, and hence many of the Skill Score values are strongly negative, especially at 2 and 3 days ahead.

Fig. 11

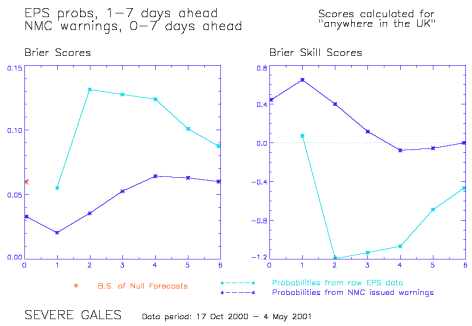


Fig. 12

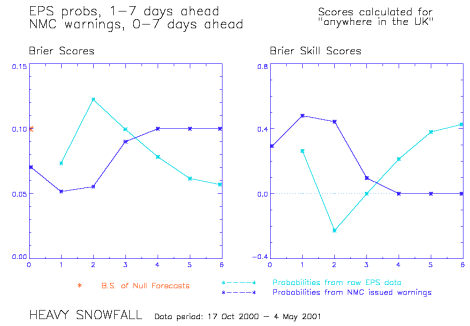
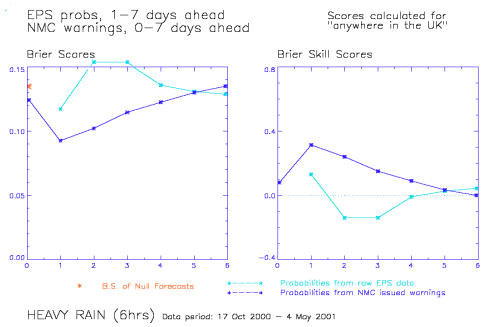


Fig. 13



Figs 11-13 Brier Scores (left of each Figure) and Brier Skill Scores (right of each Figure), for Severe Gale warnings (Fig. 11), Heavy Snowfall warnings (Fig. 12), and Heavy Rainfall warnings (Fig. 13), from the operational FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring anywhere in UK.

Brier Scores for individual-area probabilities are lower (Figs 14-16), because the forecast probabilities are lower and a greater proportion of 'non-events' occurred, but cannot be directly compared with those for probabilities anywhere in the UK for the reason explained above. Brier Skill Scores can, however, be compared in this way, but they remain disappointing for FGEW forecasts. There is a small degree of positive skill at D+1 (though skill is much lower than for NMC warnings), but beyond this the probability bias gets worse and the skill scores are negative.

Fig. 14

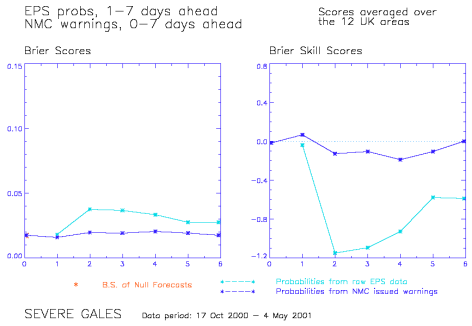


Fig. 15

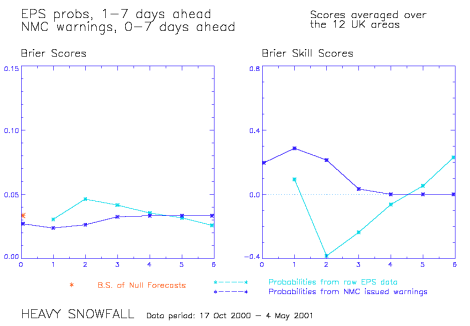
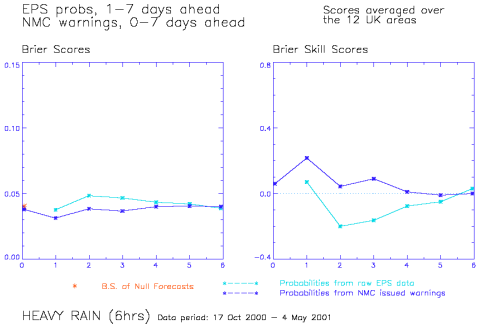


Fig. 16



Figs 14-16 Brier Scores (left of each Figure) and Brier Skill Scores (right of each Figure), for Severe Gale warnings (Fig. 14), Heavy Snowfall warnings (Fig. 15), and Heavy Rainfall warnings (Fig. 16), from the operational FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring in individual areas.

6. Correct Alarm Ratio and Miss Rate

The Correct Alarm Ratio is the proportion, out of all warnings issued, that actually prove correct. Referring to Table 1 again, this is equal to $H/(H+F)$. It is bounded by the values 0.0 and 1.0, and a large value is desirable.

The Miss Rate is the proportion, out of all occasions when no warning was given, that the event does occur. Using the notation of Table 1 once more, Miss Rate = $M/(M+R)$. Again bounded by the values 0.0 and 1.0, a small value of Miss Rate is ideal. It is convenient to plot these two quantities on the same graph.

However, because Early Warnings are required to be issued whenever the probability of an event is 60% or greater, we do expect some false-alarms, and also some events must be expected to occur when the forecast probability is below 60%, for a reliable and well-calibrated system. Hence the Correct Alarm Ratio cannot be expected to reach 1.0, and the incidence of misses ensures that the Miss Rate will not be zero. As the requirement for warning issue is a probability of at least 60%, we hope to obtain a Correct Alarm Ratio greater than 0.6. A useful forecasting system will produce a Miss Rate lower than the sample climatological frequency of the event. For completely random forecasts, note that the Correct Alarm Ratio and the Miss Rate for any event would both be equal to the event's sample climatological frequency; also, if a warning is never issued, the Miss Rate and the sample climatological frequency will also be the same. Other numerical ratios can be inferred from contingency tables, but the merits of these have not been explored here.

Correct Alarm Ratios and Miss Rates for the operational FGEW system

These results are shown in Figs 17-19, for probabilities of events occurring anywhere in the UK, for forecasts 1 to 4 days ahead. Note that the axis scaling for Correct-Alarm Ratio, on the left-hand side of the graphs, is not the same as that for Miss Rate, on the right-hand side. The x-axis of these graphs is a forecast probability threshold, so for example the points plotted at 0.6 are for the event being treated as forecast if the probability is 60% or more. Again these graphs permit comparison of the performance of the FGEW system with that of NMC issued warnings. These forecasts all perform better than random forecasts, except at very high probabilities (for which the sample sizes are

Fig. 17

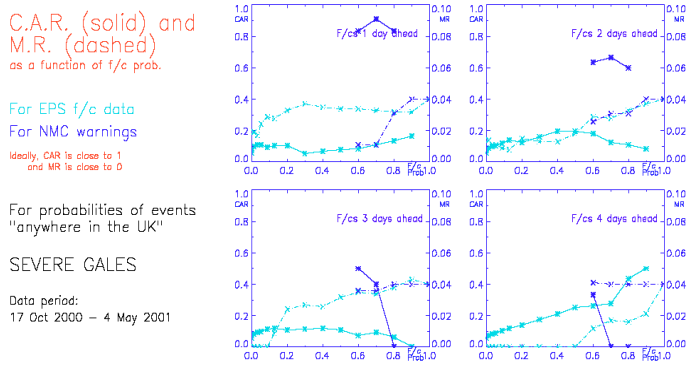


Fig. 18

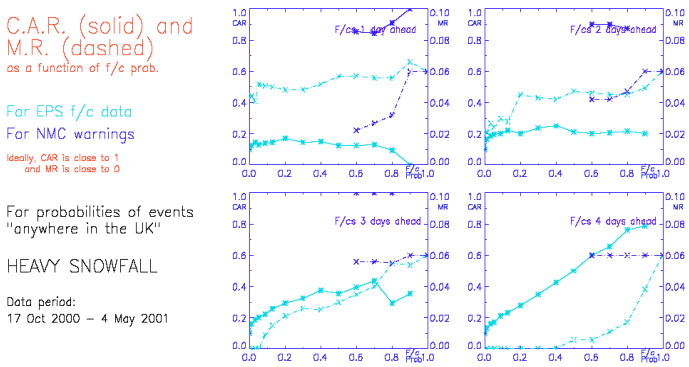
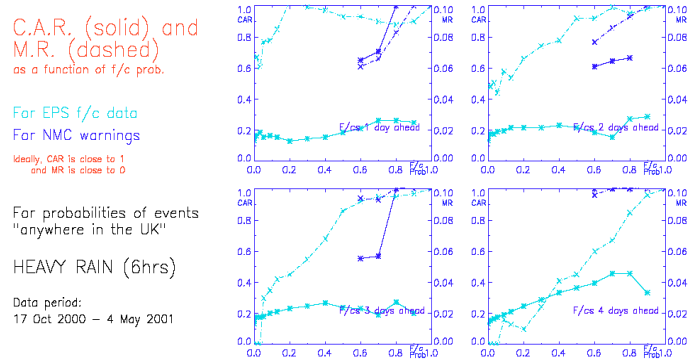


Fig. 19



Figs 17-19 Correct Alarm Ratios (solid lines) and Miss Rates (dashed lines), for 1, 2, 3 and 4 days ahead, for Severe Gale warnings (Fig. 17), Heavy Snowfall warnings (Fig. 18), and Heavy Rainfall warnings (Fig. 19), from the operational FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring anywhere in UK.

especially small). The Miss Rates we have obtained are always equal to or less than the sample climatological frequency (shown by the Miss Rate for a probability threshold of 1.0 where the curve cuts the right-hand axis), which is encouraging. It is difficult to compare Miss Rates for FGEW with those for NMC as they are calculated as a proportion of the occasions when no warning was given, and FGEW issued far more warnings than NMC. Correct Alarm Ratios for FGEW are poor (low) at up to 3 days ahead; they are much better at D+4, but, except for heavy snowfall, still below the 60% which we would ideally achieve. Ideally we would see larger values of Correct Alarm Ratios for higher probability thresholds, and this is indeed seen on most of these graphs.

7. Proportion of events that were not forecast

Another way of defining the miss rate is to consider, out of all occurrences of an event, what proportion were not forecast. This is equal to *(1-HR)*, where *HR* is the hit-rate defined above, and can be calculated at any forecast probability threshold. As before, there are no points on the curves for NMC probabilities below p=0.6 for events occurring 'anywhere in UK'.

Graphs showing miss rates calculated by this alternative definition appear in Figs 20-22. Miss rates of EPS forecasts are rather poor at D+1 for all three weather events, because at this range the system has little discriminatory ability; on many occasions at D+1 when the system does forecast an event, the event does not occur. Results are better at longer lead-times, most especially D+4. For NMC forecasts,

Fig. 20

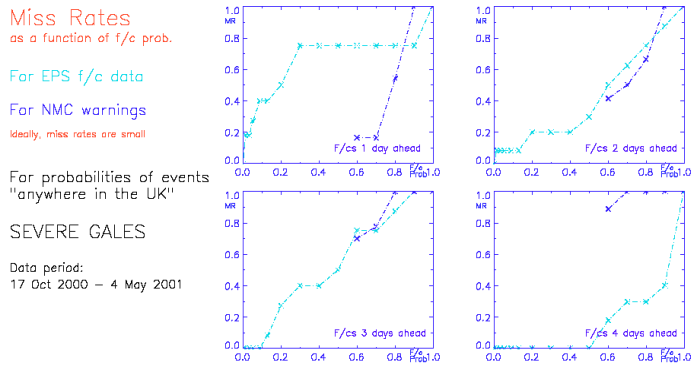


Fig. 21

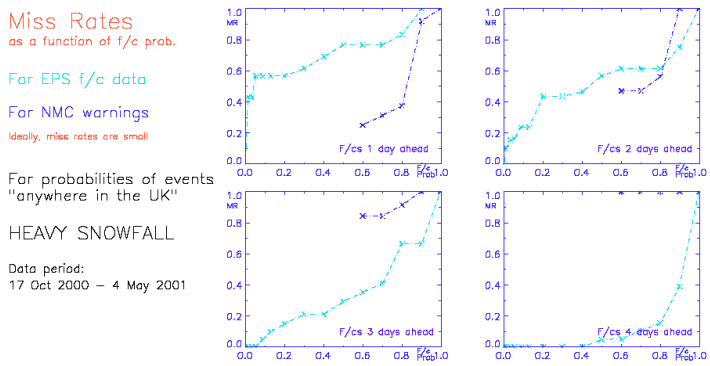
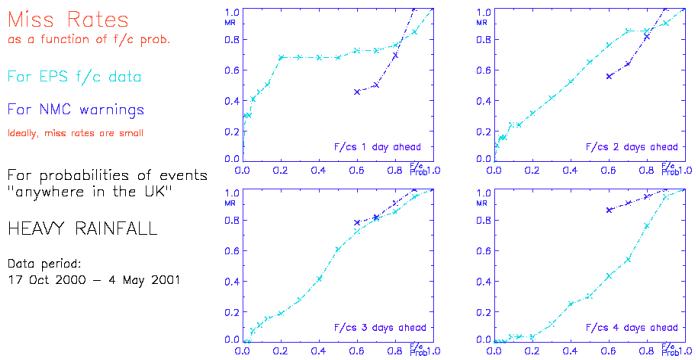


Fig. 22



Figs 20-22 Miss Rates (alternative definition – see text), for 1, 2, 3 and 4 days ahead, for Severe Gale warnings (Fig. 20), Heavy Snowfall warnings (Fig. 21), and Heavy Rainfall warnings (Fig. 22), from the operational FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring anywhere in UK.

miss rates at D+1 and D+2 are relatively low, but so few events are forecast at 3 and 4 days ahead that miss rates are high. This shows that the EPS forecasts are better in this respect at 3 and 4 days ahead. In some cases where the EPS appears to do better than NMC forecasters, this is, however, at the expense of a higher false-alarm rate.

8. Optimisation of the FGEW system

Using the system as originally set up for the operational trial which began in September 2000, the verification results presented above are relatively poor, though with D+4 forecasts showing some skill. In particular, a strong tendency to over-estimate event probabilities was noted, which is seen most clearly in the reliability curves but also has a large effect on the Brier Scores. The two things that could be altered in an attempt to improve this were (a) the event thresholds and (b) the use of 'time-windowing' (designed to cater for uncertainties in the timing of particular weather events). Event probabilities can be reduced by raising the thresholds that must be reached to trigger a warning, or by reducing the width of the time-windowing. The amount of time-windowing allowed is somewhat subjective, and was allowed to increase with forecast time (up to 24 hours either side at 5-6 days ahead). Event thresholds are more tangible, and increasing these too much would be unrealistic, so it was thought best to try reducing (or even eliminating) the time-windowing while leaving the event thresholds at or below the values defined in the NSWWS event definitions.

The original event thresholds were determined using objective techniques, as described in the Report on the Scanning System (Legg and Mylne, 2000), but were based on a small data sample. They were also set before the EPS upgrade in November 2000 (though, following the upgrade and a preliminary comparison exercise, and in the light of comments received so far from the forecasters who were using the system, thresholds were increased somewhat), so it was always anticipated that they would need to be adjusted in the light of verification. In making these adjustments, we have aimed to achieve forecasts whose mean probability is close to the sample climatological frequency for any given event. Thus we have attempted to optimise the system as best we can using the available data covering the past winter season. This should ensure that the system will have a useful degree of skill in terms of probabilistic prediction of severe-weather events. Assessments of the probability forecasts obtained from the system before and after this optimisation process was performed are described below, and form the basis of proposals for changes to the system (in terms of time-windowing and numerical values of event thresholds) to be introduced ready for the next winter season.

The main tool that was used as an aid to the re-calibration was the Reliability Diagrams, together with mean forecast probabilities which were compared with the sample mean frequencies of occurrence for each event. For this reason, reliability curves are discussed first in this section, and recommendations for how to optimally tune the system are made, with greatest weight given to the performance of forecasts at D+4. The corresponding results for ROC, Brier Scores/Brier Skill Scores, and Correct Alarm Ratios/Miss Rates are then described.

(i) Reliability Diagrams

Trials of improved experimental versions of the FGEW system

In an attempt to improve the verifications of the system, various changes were made to the scanning program, and the resulting forecast probabilities re-assessed. First of all, the time-windowing allowed in the scanning program was halved, which led to a noticeable improvement in the reliability curves, because the mean forecast probabilities were reduced slightly, giving fewer instances of 'non-events' that had high forecast probabilities. Removing the scanning-program time-windowing completely gave some further improvement to the reliability curves, though some apparent over-forecasting

remained in some instances. (Indeed, with the exception of D+4, the system clearly does not have resolution across the whole range of probabilities, and is merely capable of identifying occasions when an event definitely will not occur.)

Also, the mean forecast probabilities for all forecast days now matched the sample climatological frequencies better, although severe gale probabilities were still over-estimated on average. So then the assessments were repeated just for severe gales, with the event thresholds altered by a common factor. The best multiplication factor was found to be *1.1. This gave a close match between mean forecast probability and sample climatological frequency, though admittedly the resolution of the forecasts was still poor for probabilities above 0.2. Despite the poor nature of the reliability curves overall, even with revised event thresholds and reduced time-windowing, it was seen once again that those for D+4 forecasts indicate much better performance.

With time-windowing removed from the scanning program, we have the desirable finding that average event probabilities given by FGEW software are now roughly independent of forecast time, rather than showing a steady increase, which we believe was an artificial effect caused by the time-windowing. However, for heavy-rainfall events, there is a slow but steady decrease of mean event probability with time; there is no obvious reason why this should happen, though there may be a real tendency towards model under-activity for rainfall.

Although the removal of time-windowing makes for a close match between average forecast probability and sample observed frequency, we feel it is necessary to retain some degree of time-windowing to allow for increasing uncertainty with time as to when an event will occur (some ensemble members will develop a given synoptic system earlier than others, for example). However, it is felt that the original time-windowing, allowing differences of up to a whole day on either side at D+5, was undesirable because it gave an artificial inflation of forecast event probabilities.

It was also noticed during the operational trial of the system that probabilities of severe gale events and heavy rainfall events were often highest in southern areas of the UK. Closer investigation revealed that the average bias in the probability forecasts was indeed greater in the south. So the assessments were repeated with the event thresholds raised slightly more in the south than in the north, and some further small improvements were noted though not for all forecast-days. We suggest increasing the event thresholds in the scanning system for these events slightly more in southern areas in order to reduce the mean probability bias here.

Proposed revisions to the scanning system: Skill of the optimised FGEW system

The proposed event thresholds to be used in the FGEW scanning system for the coming winter are listed in Table 3 at the end of this Report. The time-windowing will extend to 6 hours either side of the forecast time throughout. Event thresholds will be increased by factors of between 1.1 and 1.2, and the increases for severe gales and for heavy rainfall will be slightly greater in southern areas of the UK.

Reliability curves for the optimised FGEW system are illustrated in Figs 23-25 (for probabilities of events occurring anywhere in the UK) and 26-28 (for individual area probabilities); these can be compared with the performance of the operational FGEW system and the warnings issued by NMC (Figs 5-10).

There is now a close match between the mean forecast probabilities and the sample mean frequency of occurrence. The reliability curves are now quite good at D+4. Although they are noisy due to the small data samples, it is clear that the mean slope of the curves for all three weather types is quite close to the ideal. Occasions when high probabilities can be predicted are, however, rare, as was expected by Mylne (2000b) in a review of the 60% probability threshold used for Early Warnings. For days 2 and 3, however, results are still poor and at best there is only some useful discrimination of whether or not there is any risk of severe weather. FGEW could be used to generate alerts, i.e. low-

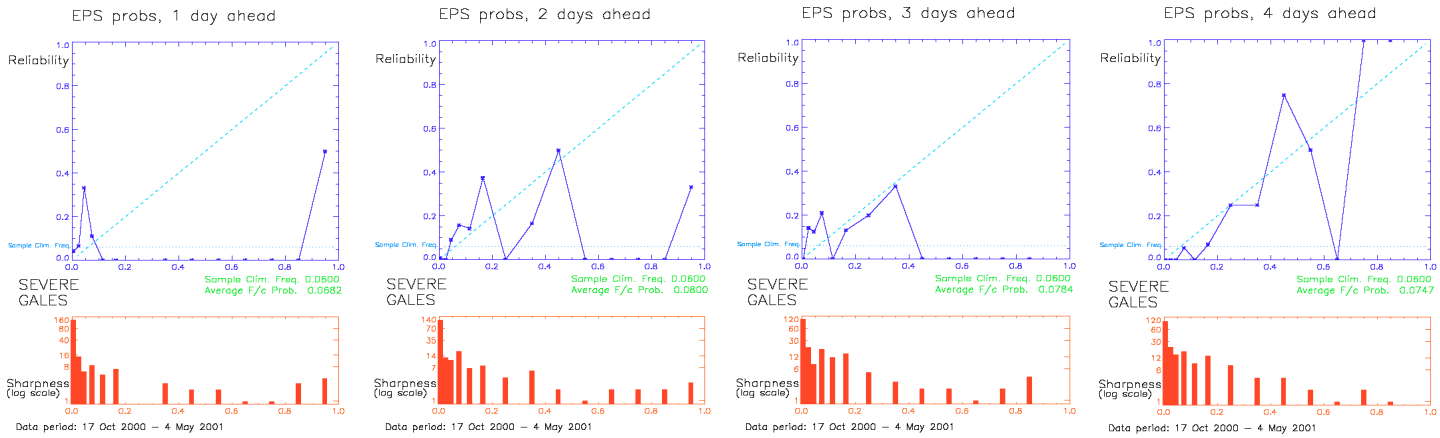


Fig. 23 Reliability Curves for Severe Gale warnings, from the optimised FGEW system, for probabilities of events anywhere in UK, at (left to right) 1, 2, 3 and 4 days ahead. Sharpness diagrams are also included, with logarithmic y-axis scaling.

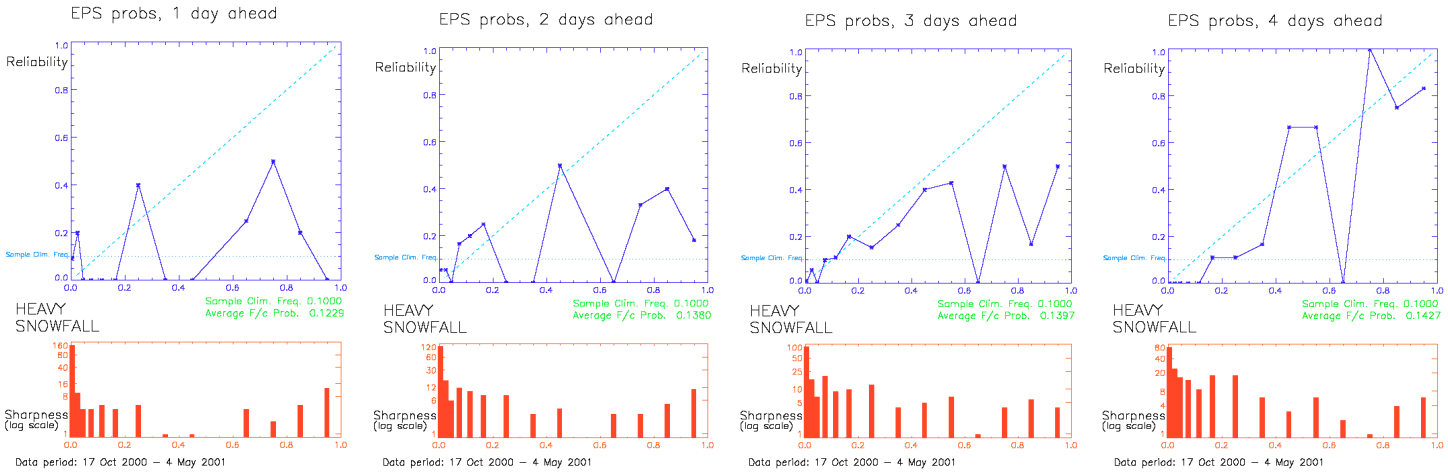


Fig. 24 As Fig. 23, but for Heavy Snowfall warnings.

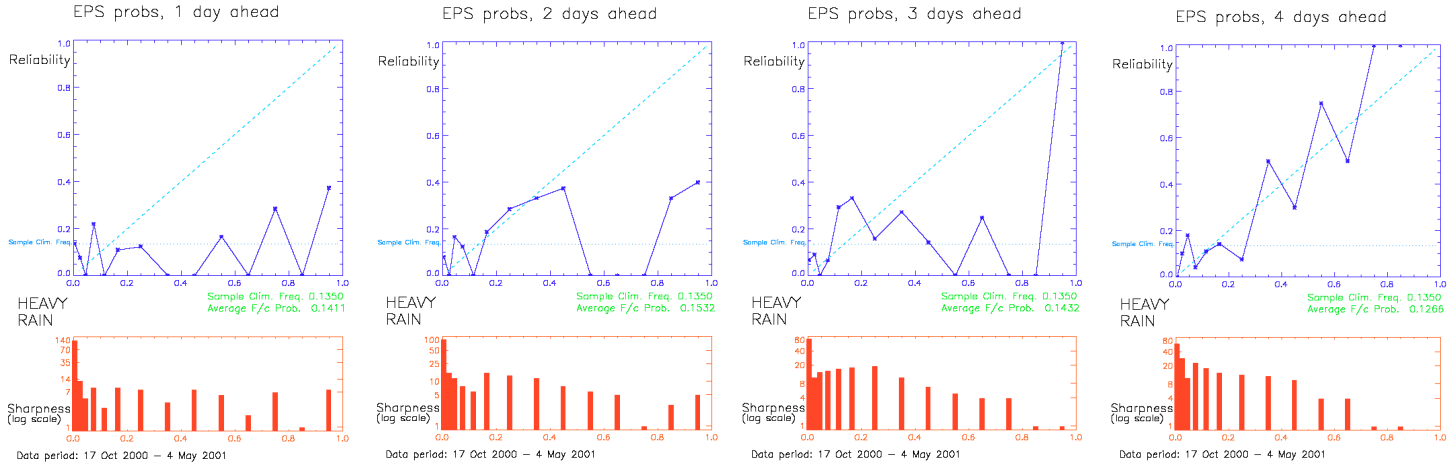


Fig. 25 As Fig. 23, but for Heavy Rainfall warnings.

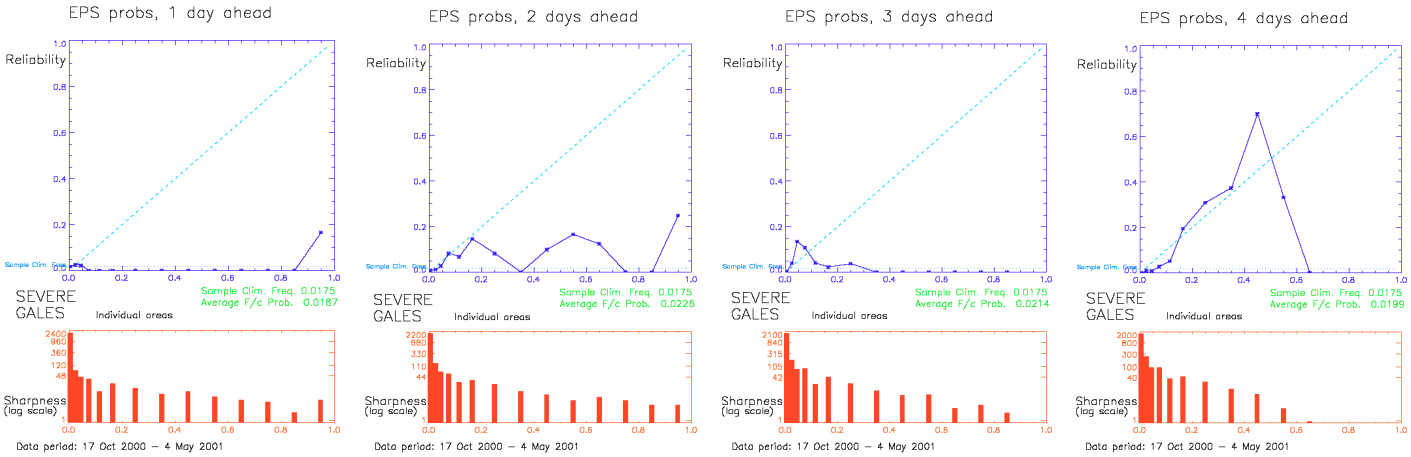


Fig. 26 Reliability Curves for Severe Gale warnings, from the optimised FGEW system, for probabilities of events occurring in individual areas, at (left to right) 1, 2, 3 and 4 days ahead. Sharpness diagrams are also included, with logarithmic y-axis scaling.

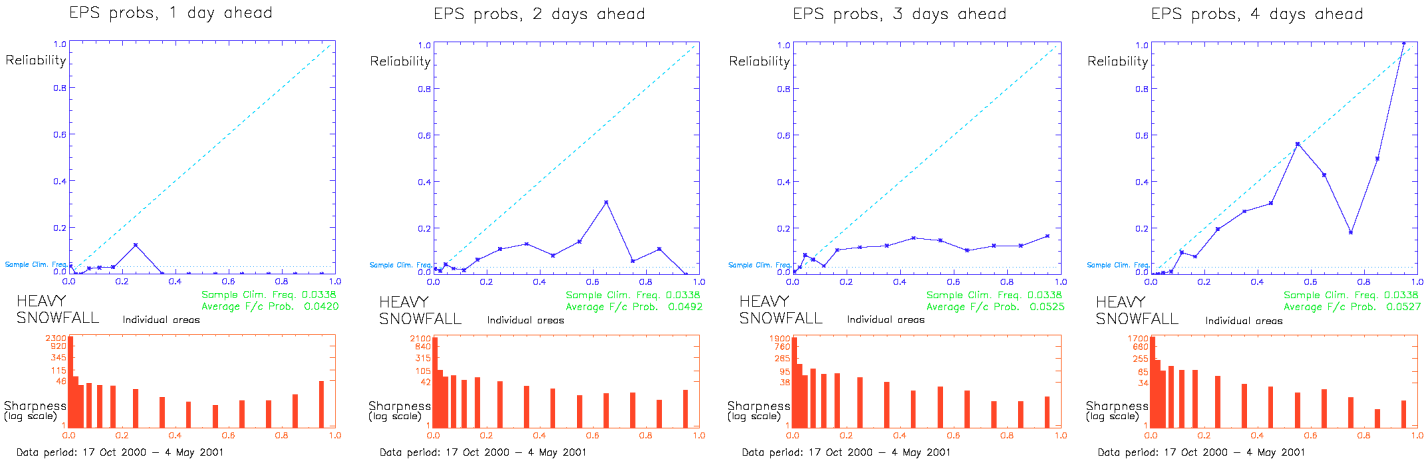


Fig. 27 As Fig. 26, but for Heavy Snowfall warnings.

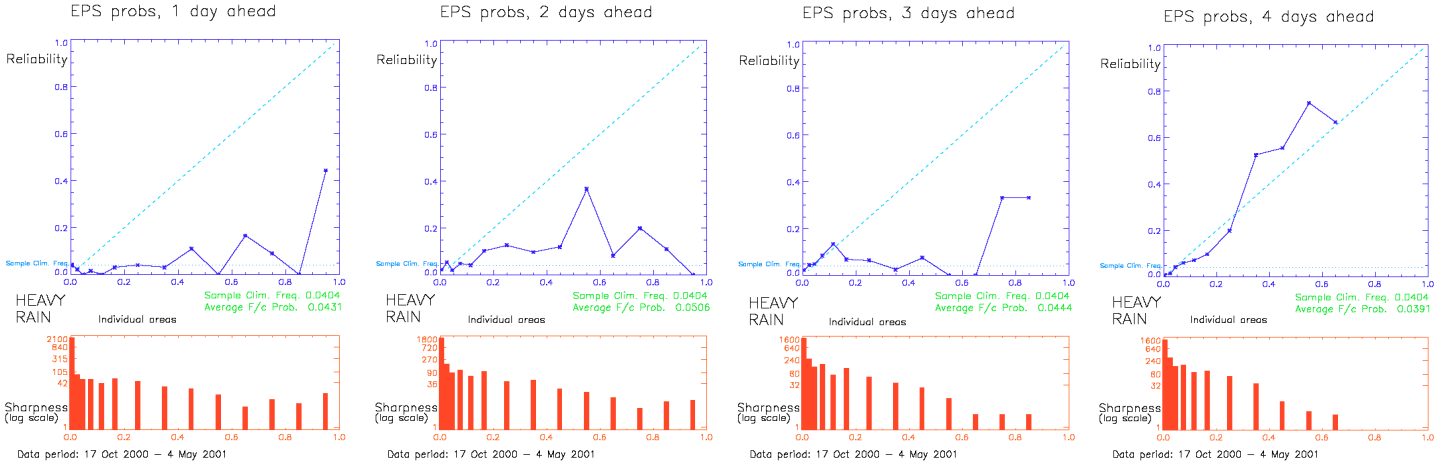


Fig. 28 As Fig. 26, but for Heavy Rainfall warnings.

probability warnings of severe events, but there is little skill beyond that except at 4 days ahead. For D+1 there is no useful skill.

(ii) Relative Operating Characteristic

ROC curves, and the magnitudes of the areas under them, do not improve when the time-windowing used is reduced or removed (see Figs 29-31 for results from the optimised system, shown just for D+2 and D+4, for individual-area probabilities). This suggests that, although the probability biases have been removed, the forecasts achieve no better resolution than those from the operational FGEW system. Individual points on the ROC curves tend to move downwards and towards the left, as both Hit Rate and False Alarm Ratios are lower for any given probability threshold, but this has little effect on the shape of the curves or the area under them. It was thought that wider time-windows might smooth out the forecast probabilities too much, leading to a loss in resolution, thus the optimised system may have produced improved ROCs, but our findings here refute this.

Fig. 29

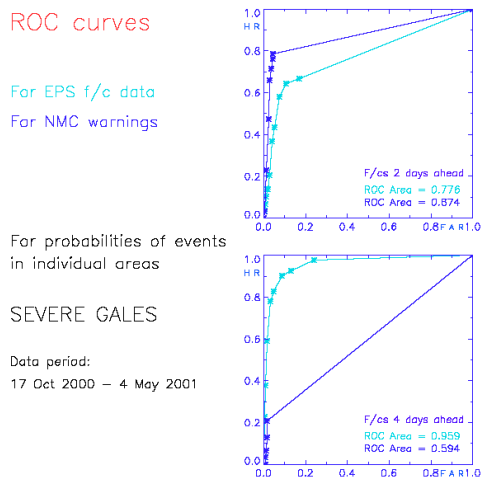


Fig. 30

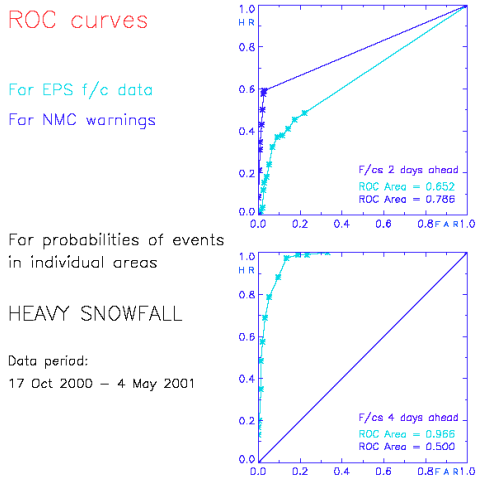
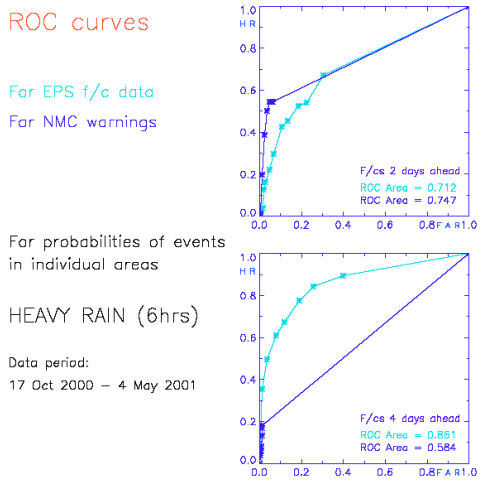


Fig. 31



Figs 29-31 Relative Operating Characteristic curves, for 2 and 4 days ahead, for Severe Gale warnings (Fig. 29), Heavy Snowfall warnings (Fig. 30), and Heavy Rainfall warnings (Fig. 31), from the optimised FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring in individual areas.

(iii) Brier Scores and Brier Skill Scores

Brier Scores and Brier Skill Scores of the optimised FGEW system show much better performance than for the operational version of the system, because the probability bias has been vastly reduced and there are few remaining high-probability forecasts of events which did not occur. We now see substantially positive Brier Skill Scores throughout, particularly beyond D+3 (Figs 32-37, on which scores for NMC warnings remain unchanged). Tests with different sets of changes to the system suggested that the optimised FGEW system as proposed in this report performs as well as any other set of changes. Probabilities obtained from the optimised system do offer useful probabilistic information at 3 or more days ahead, but this is not so at 1 and 2 days ahead, which is when most NMC Early Warnings (which are more skilful at this range) are issued. The reasons for the shortcoming at 1 and 2 days ahead have been discussed already, and include the lack of spread in the ensemble, which means that the probabilities are often too high or too low; it remains to be seen whether this

improves when we assess using an independent sample of data which is free of the 'spread bug' described earlier. However, the effects of non-random sampling in the ensemble will not change.

Fig. 32

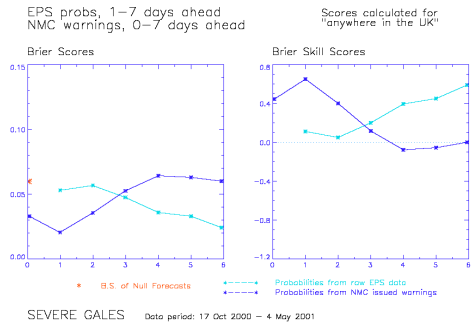


Fig. 33

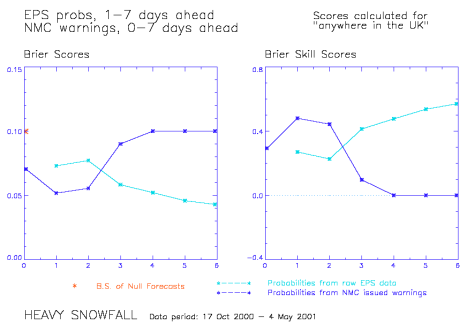
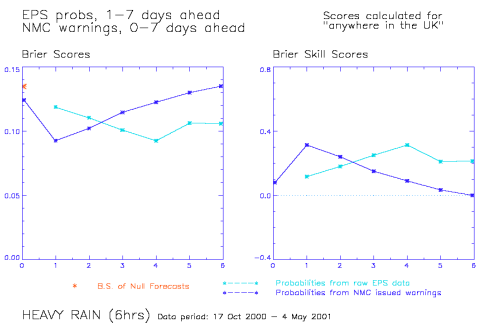


Fig. 34



Figs 32-34 Brier Scores (left of each Figure) and Brier Skill Scores (right of each Figure), for Severe Gale warnings (Fig. 32), Heavy Snowfall warnings (Fig. 33), and Heavy Rainfall warnings (Fig. 34), from the optimised FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring anywhere in UK.

Fig. 35

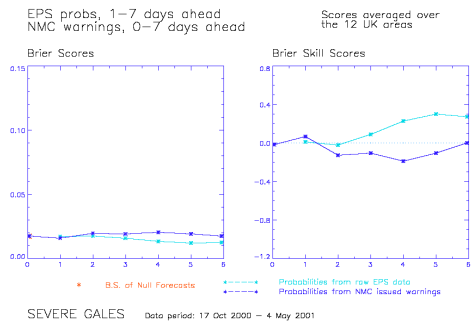


Fig. 36

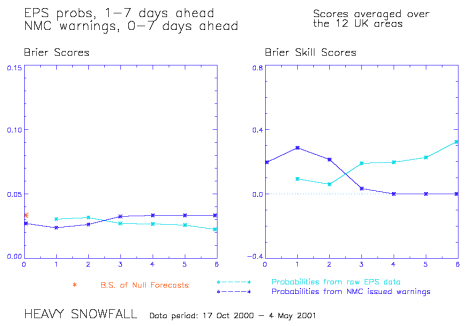
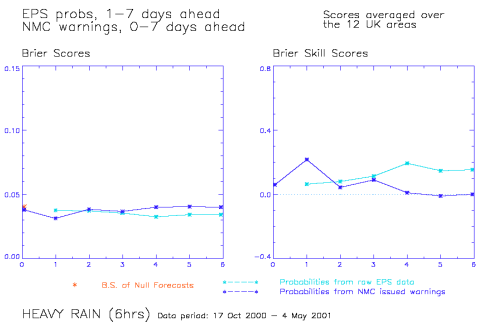


Fig. 37



Figs 35-37 Brier Scores (left of each Figure) and Brier Skill Scores (right of each Figure), for Severe Gale warnings (Fig. 35), Heavy Snowfall warnings (Fig. 36), and Heavy Rainfall warnings (Fig. 37), from the optimised FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring in individual areas.

(iv) Correct Alarm Ratio and Miss Rate

Re-running the system with our proposed event thresholds and time-windowing, results once more show an improvement (Figs 38-40). However, the Miss Rates do generally increase. The reason for this is that we are now forecasting events less often (for any given probability threshold); the increase in Miss Rate is accompanied by a substantial and beneficial drop in the False-Alarm Rate. Note once again that Correct Alarm Ratios remain much better at D+4 than for other forecast-days, and the biggest improvements here are indeed seen at D+4. Correct Alarm Ratios now exceed 0.6 for the higher probability thresholds at D+4 for all events, and also at D+3 for heavy rainfall.

Fig. 38

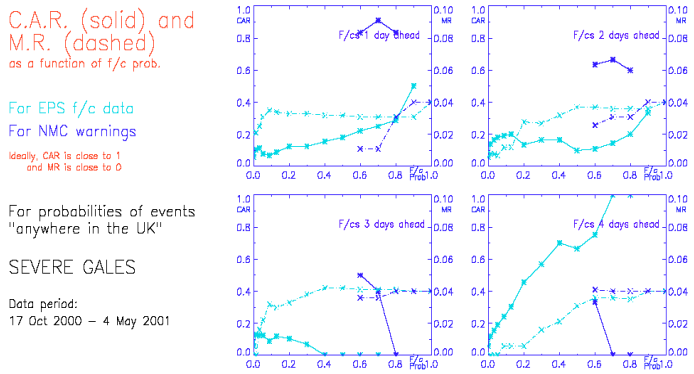


Fig. 39

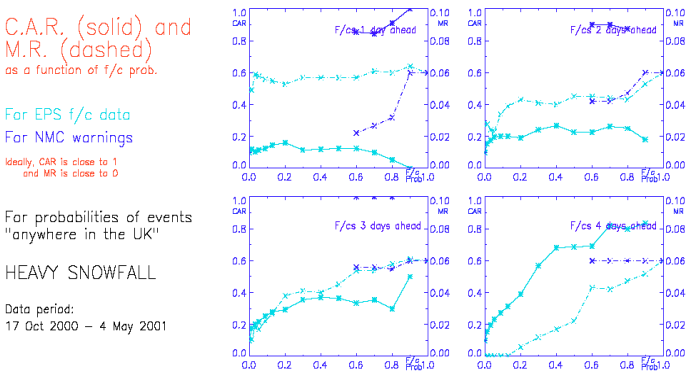
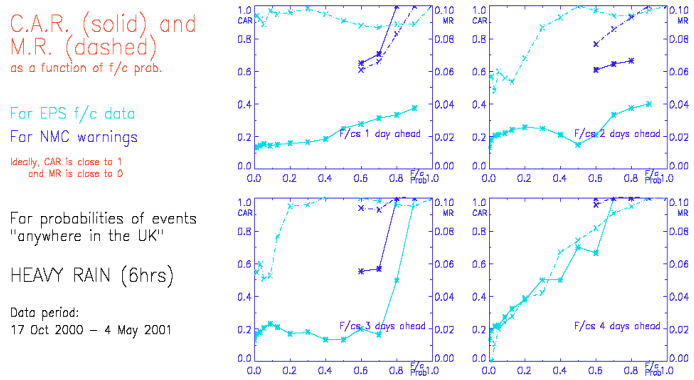


Fig. 40



Figs 38-40 Correct Alarm Ratios (solid lines) and Miss Rates (dashed lines), for 1, 2, 3 and 4 days ahead, for Severe Gale warnings (Fig. 38), Heavy Snowfall warnings (Fig. 39), and Heavy Rainfall warnings (Fig. 40), from the optimised FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring anywhere in UK.

(v) Proportion of events that were not forecast

With revised thresholds and time-windowing, biases in forecast probabilities are much reduced. This means that any given forecast probability threshold is exceeded less often, i.e. fewer events are forecast, giving an increase in miss rates. This is consistent with the predictions of Mylne (2000b) that we should not expect to be able to predict severe events with high probabilities - the operational system did succeed to some extent, but only at a cost of excessive false alarms. Thus, for severe gales and for heavy rainfall, the EPS miss rates at 3 and 4 days ahead are mostly no longer lower than those for NMC forecasts, although for heavy snowfall the EPS probabilities do still retain some advantage. However, what these graphs do not show *per se* is that the false-alarm rates are substantially improved (i.e. lower) following re-calibration of the system (as mentioned in sub-section (ii) above).

It is worth noting that if a lower probability threshold were used for the issue of warnings (say, 30% instead of 60%) these graphs show that the number of missed events could be substantially reduced. Of course this would be at the cost of more false alarms, but this may nevertheless be useful information to some users who stand to suffer large losses from severe events.

Fig. 41

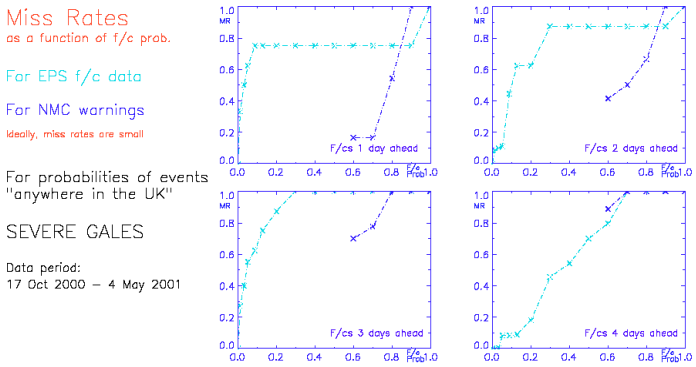


Fig. 42

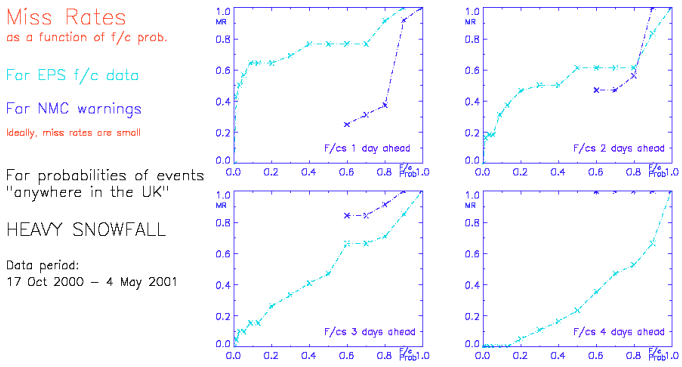
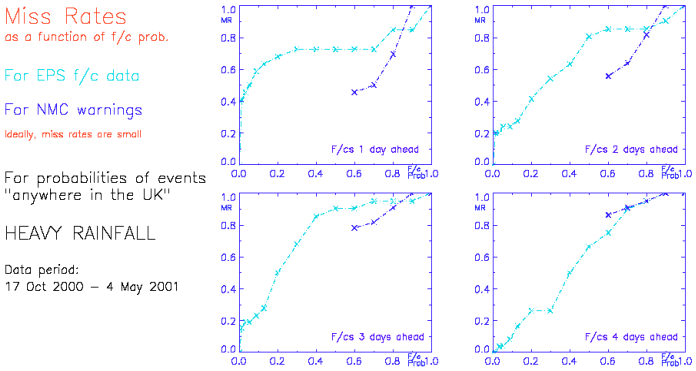


Fig. 43



Figs 41-43 Miss Rates (alternative definition – see text), for 1, 2, 3 and 4 days ahead, for Severe Gale warnings (Fig. 41), Heavy Snowfall warnings (Fig. 42), and Heavy Rainfall warnings (Fig. 43), from the optimised FGEW system (light blue curves) and from NMC warnings (dark blue curves), for probabilities of events occurring anywhere in UK.

9. Conclusions and Recommendations

Results and Conclusions

Assessment results to date for the FGEW scanning system, based on 6½ months of data since the latest upgrade to the ECMWF EPS, show that the operational system has good resolution in 4-day forecasts, although the thresholds used lead to quite severe over-forecasting. Using a process of 'calibration by assessment' (discussed in more detail below) it has been shown that this over-forecasting can be effectively eliminated, giving a potential for good probabilistic forecasts at D+4. Note, however, that this calibration has not been tested using independent data, due to the small data samples available for analysis, so results in the coming season are unlikely to be quite as good as shown here. Also it must be noted that these results may not be typical of other periods, because the assessments cover a period during a substantial part of which there was a bug which was subsequently discovered to have been affecting the spread of the EPS. Thus the true skill of the re-calibrated system can only be assessed over the coming winter season.

Results for shorter forecast periods of 1 to 3 days were less good. Indeed the system has no skill at D+1, and at 2 and 3 days has only a limited ability to discriminate occasions when there is no risk of severe weather from occasions when there is some risk. It may therefore be useful in issuing alerts to forecasters at this range, but not in assessing the actual probabilities of severe weather.

Comparison of forecast skill with the warnings issued by NMC forecasters was complicated by the fact that NMC rarely issues warnings earlier than two days ahead. Given the poor skill of FGEW at short range, it is not surprising that the forecasters did much better there. Equally FGEW did much better at 4 days because it performed well and NMC hardly ever issued forecasts at that range.

It should be noted here that because of the late data-cut-off used at ECMWF, NMC forecasters issuing a 3-day warning would actually be using D+4 EPS data. It is of no consequence that the D+1 FGEW data is no use, since it is not available early enough to be practically useful anyway. On the other hand, the skill of FGEW D+4 forecasts suggests that they could be used to issue useful 3-day warnings from NMC, and this could substantially improve the warning given to customers by issuing warnings more frequently. This should help achieve one of the aims of the project, which is to encourage earlier issue of warnings of severe events by NMC forecasters.

A notable feature of the D+4 results is that forecasts of high probabilities are quite rare. This is consistent with the predictions of Mylne (2000b). Assuming they are representative, the reliability diagrams of the optimised system in Figs 20-22 show that there is considerable potentially useful information at probabilities between 20 and 60% which will not be available to customers because of the rather arbitrary threshold of 60% at which warnings are issued. If these results are confirmed by independent data in future years, there would be a strong case for recommending to the NSWWS customers that Early Warnings should be issued at a lower probability threshold.

It is interesting to consider why the FGEW system performs so much better at day 4 than at earlier times. The EPS is purposely designed for medium-range use, and at D+1 the perturbations are still very small (although growing rapidly), so poor performance here is unsurprising. At D+2 and D+3 the perturbations should have completed their period of rapid growth and be representative of typical forecast errors, but the performance is still poor. The singular vector perturbations used are designed to look for maximum error growth over the first 48 hours of the forecast, so they represent far from a random sampling of the forecast pdf at that time. However, without a random sampling of the pdf, we should not expect to get reliable estimates of forecast probabilities. It is not until the effects of non-linearity are able to mix up the forecasts beyond about 48 hours that we can expect the ensemble to give us a quasi-random sampling of the forecast pdf, and it is believed that this is why the probability forecasts are much better at day 4.

Recommendations for Calibration for the 2001/02 Season

It has been shown that there is a need to refine the thresholds against which EPS forecast output is compared to determine event probabilities, in order to maximise the skill of the forecasting system and hence its usefulness in decision-making applications. A process of 'calibration by assessment' has been used, giving greatest weight to the performance of forecasts at 4 days ahead, in order to optimise the performance of the system.

By increasing some of the event thresholds by a modest amount, and reducing the amount of 'time-windowing' built into the event-scanning software to 6 hours either side of forecast-time throughout, skill scores and reliability curves based on the available data to date are substantially improved. It is recommended that the operational system should be altered accordingly as from August 2001. (Note: this change has already been implemented.) Once this has been done, there is the potential for the FGEW system to have considerable skill in alerting NMC forecasters when to issue Early Warnings, most especially at Day 4, which would be used by NMC for 3-day warnings; subsequent system tuning will be possible in the light of further experience. Further research is required to investigate how to improve warnings at Days 1-3.

Table 3 shows the event thresholds as used in the scanning system as it was last winter (in *italics*), and alongside are the recommended event thresholds for the revised version of the system (in **bold**). Instead of the time-windowing widening from 6 to 24 hours between 1 and 6 days ahead, we propose that it remains at 6 hours on either side throughout.

Table 3 Proposed event thresholds for the optimised FGEW scanning system, to be applied for the 2001/02 winter period

Basic event definition	FGEW Scanning System thresholds... *Original (italic)* / **Revised (bold)**			
	N. Scotland S.W. Scotland N. Ireland	S.E. Scotland N.E. England E.Angl./Lincs.	N.W. England Wales S.W. England	Midlands Cen. S. England S.E. England
Severe gales: Gusts to 70mph	*70mph* **77mph**	*62mph* **68mph**	*60mph* **69mph**	*58mph* **67mph**
Snowfall: 4cm accumulation in 2hrs	*1.5cm in 6hrs* **1.8cm in 6hrs**	*1.5cm in 6hrs* **1.8cm in 6hrs**	*1.5cm in 6hrs* **1.8cm in 6hrs**	*1.5cm in 6hrs* **1.8cm in 6hrs**
Blizzard: 1cm snow accum. and 30mph mean wind	*0.7cm, 30mph* **unchanged**	*0.7cm, 30mph* **unchanged**	*0.7cm, 30mph* **unchanged**	*0.7cm, 30mph* **unchanged**
Heavy rain: 15mm in 3hrs	*20.0mm in 6hrs* **22.0mm in 6hrs**	*11.0mm in 6hrs* **12.0mm in 6hrs**	*10.0mm in 6hrs* **11.5mm in 6hrs**	*9.5mm in 6hrs* **11.0mm in 6hrs**
Prolonged heavy rain: 25mm in 24hrs	*27.5mm in 24hrs* **unchanged**	*27.5mm in 24hrs* **unchanged**	*25.0mm in 24hrs* **unchanged**	*27.5mm in 24hrs* **unchanged**
Exceptionally severe gales: Gusts to 80mph	*80mph* **88mph**	*71mph* **78mph**	*69mph* **79mph**	*67mph* **77mph**
Exceptionally heavy snowfall: 10-15cm accum. in 3hrs	*4.5cm in 6hrs* **5.5cm in 6hrs**	*4.5cm in 6hrs* **5.5cm in 6hrs**	*4.5cm in 6hrs* **5.5cm in 6hrs**	*4.5cm in 6hrs* **5.5cm in 6hrs**

References

T.P. Legg and K.R. Mylne, 2000 – First-Guess Early Warning project: Report on the Scanning System. 23 June 2000.

K.R. Mylne, 2000a – First-Guess Early Warnings Project, Verification Plan and Design. June 2000.

K.R. Mylne, 2000b – Review of Probability Thresholds in Early Warnings. June 2000.