WORLD METEOROLOGICAL ORGANIZATION

CBS-DPFS/EPS/Doc. 7(2)

(31.I.2006)

Item: 7

ENGLISH ONLY

COMMISSION FOR BASIC SYSTEMS
OPAG DPFS

EXPERT TEAM ON ENSEMBLE PREDICTION SYSTEMS

EXETER, UNITED KINGDOM
6-10 FEBRUARY 2006

REPORT ON APPLICATIONS OF EPS FOR SEVERE WEATHER FORECASTING

(Submitted by Mr. Ken Mylne)

ACTION PROPOSED

The meeting is invited to review the document and consider input to its conclusions and recommendations as appropriate.

Early Warnings of Severe Weather from Ensemble Forecast Information

T. P. LEGG AND K. R. MYLNE

Met Office, Exeter, United Kingdom

(Manuscript received 6 May 2003, in final form 25 March 2004)

ABSTRACT

A system has been developed to give probabilistic warnings of severe-weather events for the United Kingdom (UK) on a regional and national basis, based on forecast output from the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS). The First-Guess Early Warnings (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 optimize the probabilistic early warnings over an initial training period of one winter season, and the resulting warnings were then verified over a 16-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 3–4 days ahead 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 that 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 behavior is found to also occur in equivalent deterministic forecasts and so is not attributable to the ensemble perturbation strategy. Nevertheless it is suggested that while the EPS perturbations work well for the medium range, alternative perturbation strategies may be required for successful short-range ensemble prediction.

1. Introduction

Over recent years numerical weather prediction (NWP) models have improved to the extent that developments in weather prediction can now be focused increasingly on severe or hazardous weather. The development of severe weather usually involves strong nonlinear 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. This process is similar to that which frequently leads to synoptic-scale uncertainty in medium-range prediction, except that the rapid evolution typically associated with severe weather developments can often lead to large uncertainty occurring at shorter lead times than normally expected (Lorenz 1969). Ensemble prediction systems (EPSs; Mureau et al. 1993; Molteni et al. 1996; Toth and Kalnay 1997) were introduced in the 1990s to attempt to quantify the uncertainty in medium-range forecasts due to synoptic-scale baroclinic instabilities. 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 sample the probability distribution of the forecast, and is therefore normally used to generate probability forecasts. Operational medium-range ensembles have been run by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) since 1992, but to date most applications and verifications have focused on relatively common and less severe events. This paper describes a first attempt to apply the ECMWF EPS for practical probabilistic prediction of severe weather in the United Kingdom (UK), and provides 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 (Hymas 1993). 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 1–2 days ahead. The First-Guess Early Warn-

Corresponding author address: Tim Legg, Forecasting Research, Met Office, FitzRoy Road, Exeter, Devon EX1 3PB, United Kingdom. E-mail: tim.legg@metoffice.com

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

Event	Definition
Severe gale	Gusts of 70 mph or more
Heavy snowfall	At least 4-cm depth of fresh snow falling within a 2-h period
Blizzard	Moderate or heavy snowfall, with mean wind speeds of at least 30 mph
Heavy rainfall	At least 15 mm of rainfall occurring within a 3-h period
Prolonged heavy rainfall	At least 25 mm of rainfall occurring within a 24-h period

* In practice, forecasters issue warnings when weather conditions are expected to endanger or seriously inconvenience human activity, so absolute values of event thresholds may be lower in heavily populated regions. We will discuss how this is handled in the FGEW system in section 4b.

ings (FGEW) project was established to estimate probabilities of severe weather objectively from the ECMWF EPS in 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 (Hymas 1993), 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. 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. Section 6 presents verification results. Results will be discussed in section 7, and conclusions summarized in section 8.

2. Ensemble prediction for severe weather

Operational EPSs 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 Center (Houtekamer et al. 1996) in 1994. Since then, the ECMWF EPS 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 80 km in midlatitudes 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 linearized adjoint of the full NWP model with dry physics at T42L40 resolution, but identify a good approximation to the dynamic modes with the fastest linear growth over the first 48 h of the forecast. Initial perturbations also include evolved SVs, calculated 48 h 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, the ECMWF also run 5 additional ''multianalysis'' (MA) members. These are the ''control'' members from the larger MA EPS described by Richardson (2001) and use the same model as the rest of the EPS, but are started from different analyses. Four use the operational analyses of different NWP centers [Met Office, Météo-France, Deutsche Wetterdienst (DWD) and NCEP]; the fifth is a ''consensus'' analysis calculated as the mean of these four 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 MA members compared to 50 SV-perturbed members, they are used in this paper with double weighting in calculation of probabilities. While there is no direct evidence that the MA members should be twice as likely to be ''correct'' as the SV members, it is expected that they may sample slightly different areas of uncertainty than the SV members, and this double weighting ensures that they can have a non-negligible impact on the forecast probabilities. Several different weightings were verified experimentally over an initial trial period. Differences in performance were quite small but suggested that a weighting of four provided the best verification. However, with relatively little evidence to support such a strong weighting, the more conservative weighting of two is used operationally.

Despite these numerous developments from the initial version of the EPS, most of the verification reported in the literature to date 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 practice, 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; they showed that calibration substantially improved the quality of ensemble probabilities for nonextreme 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 h ahead.

3. Predictability of severe weather

The defined requirement for the issue of early warnings in the NSWWS is a 60% probability of sufficiently severe weather conditions to cause danger, or significant disruption to normal life, somewhere in the United Kingdom. It is interesting to speculate on how often severe weather is likely to be predictable at this level more than about 24 h in advance. As noted in the introduction, the development of severe weather normally involves the nonlinear interaction of quite small-scale flow anomalies in the atmosphere. One or more of these anomalies may individually be climatologically extreme and therefore difficult to predict, making their interactions more difficult to predict. Small differences in the position, intensity, or timing of such anomalies in the model can lead to large differences in forecast evolution. In an ensemble, if we are successful in perturbing the features to which the forecast is sensitive, then most members of the ensemble will therefore produce interactions different from the single realization of the atmosphere, and the forecast probability of the severe event will inevitably be low. Evidence from the December 1999 storms over France and Germany showed that only a minority of ensemble members (or of deterministic forecasts from different centers) succeeded in predicting severe storms, even at ~24 h ahead (Palmer 2002). Of those which did predict storms, there was variability in both intensity and location of severe conditions, so the local probability of severe weather at any particular location and time was lower than the probability of severe weather occurring somewhere within a larger region over a period of time. Thus we should not expect an ensemble to generate high probabilities of severe weather, especially not locally, except on rare occasions when the atmosphere is in an exceptionally predictable state.

There is good reason to expect that this sensitivity to the interactions of flow anomalies is characteristic of the real atmosphere as much as of the NWP model. Just as the model may be in an unstable state where its evolution is sensitive, so may the real atmosphere, so the forecast uncertainty is an inherent characteristic of the atmosphere and not just a weakness of NWP models. There will be exceptions to this, occasions when there is much severe weather potential and a high probability of realizing it (although we can never know the true probability in the real atmosphere). However, even on these occasions there is likely to be uncertainty about exactly where and when severe conditions will develop. Therefore, although the forecast probability of severe weather occurring somewhere within a large region may be high, local probabilities will still be low. The European storms of December 1999 were a good example of this. An exceptionally strong jet stream in the upper troposphere provided the potential and was itself quite predictable, but the fine details and resulting cyclogenesis were much less certain.

4. The FGEW system

a. Calculation of representative probabilities

A common concern is that end users do not understand what probability forecasts mean. To avoid this it is vital that probabilities are given for clearly defined events. For example, a warning of heavy rain in England must specify exactly how ''heavy rain'' is defined; also whether the probability refers to ''somewhere in England'' getting heavy rain, or ''any specific location'' within that area—two very different probabilities. Fortunately, the NSWWS specifies events quite clearly: an event may be observed anywhere within the area stated for the forecast to verify. Table 1 gives the definitions of severe weather events used in the NSWWS. In early warnings, probabilities are given for these events occurring ''anywhere in the United Kingdom'' and also within each of 12 local areas of the United Kingdom. In either case, the weather need only occur somewhere within the area.

Conventionally, ensemble probabilities are often shown as contoured charts of grid-point values at specific times. Severe events tend to occur quite locally over small areas at any fixed time, thus affecting only a few grid points in each ensemble member. Given the spread of the ensemble, those EPS members that generate severe weather are likely to have it in different locations, and perhaps with differing timing, especially in forecasts several days ahead. Hence, point probabilities of severe events are almost invariably low. However, for NSWWS early warnings the defined weather threshold need only be exceeded at one grid point in a region for an EPS member to count towards the required probability. Similarly, early warnings are issued to cover a stated time period, typically between 12 and 36 h in length. The precise timing of an event is not critical for a warning several days ahead, so the threshold need only

TABLE 2. Proxy events used to represent NSWWS severe-weather definitions in the EPS model.

NSWWS severe-event definitions		Proxy events in EPS model			
		Northwest UK	Eastern UK	West and southwest UK	Central and southern England
Severe gales	Gusts of 70 mph	77 mph	68 mph	69 mph	67 mph
Heavy snow	4 cm snow within 2 h	2.5 cm in 6 h	2.5 cm in 6 h	2.5 cm in 6 h	2.5 cm in 6 h
Heavy rain	15 mm rain within 3 h	22.0 mm in 6 h	12.0 mm in 6 h	11.5 mm in 6 h	11.0 mm in 6 h

be exceeded at a single time within a 12-h window for an EPS member to count toward the probability. 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. Calculating probabilities for regions and for 12-h 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 required for the NSWWS warnings.

b. Definition of severe-weather thresholds

In calculating probabilities for FGEW it was necessary to specify the severe weather events carefully from the EPS fields. 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 ''15 mm in 3 h.'' EPS fields are only output every 6 h so it is immediately necessary to define a ''proxy event'' that can be identified from 6-h 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 parameterized 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 mismatch between model output fields and the real-world warning definitions, an NWP model with 80-km grid length cannot resolve the locally observed extremes in a severe-weather event. Thus the thresholds defined to represent severe-weather events in the model are expected to be less extreme than the real-world events in Table 1. It was noted in Table 1 that, in practice, slightly different warning thresholds are used in different parts of the United Kingdom, depending on the sensitivity of the region. Hence, for FGEW purposes, the United Kingdom was divided into four regions and the thresholds are calibrated separately for each.

The initial specification of the proxy events was necessarily somewhat arbitrary, but the precise thresholds used were subsequently calibrated to optimize performance over an initial training period from 17 October 2000 to 4 May 2001. For an unbiased probability forecast system the mean forecast probability should equal the sample climatology. Thresholds were calibrated to minimize the bias in event probabilities over this training period. Because the FGEW proxy events were calibrated against forecaster-issued flash warnings (see verification details in section 5), this process automatically made allowance for any variations in the weather sensitivity of the four regions. Since the forecasters issue flash warnings according to the severe weather that actually occurs, it also made allowance for any difference in frequency of severe weather in the training period from normal climatology. Calibrated proxy event thresholds used in the operational FGEW system are presented in Table 2.

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 the model climatology. We use an approximate climatology of the EPS model generated by Lalaurette (2003). To calculate probabilities of severe weather as required for FGEW, 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 above, this method avoids the need for tuning over a long training period as the relationship between the model and site observations can be established from pre-existing statistical data. This approach could thus be used to calibrate the system for any warning threshold required, and any location for which site climatology is available.

Warning probabilities are based on forecasts 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. 1. First, 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. The same percentile point on the model climate distribution for the representative grid

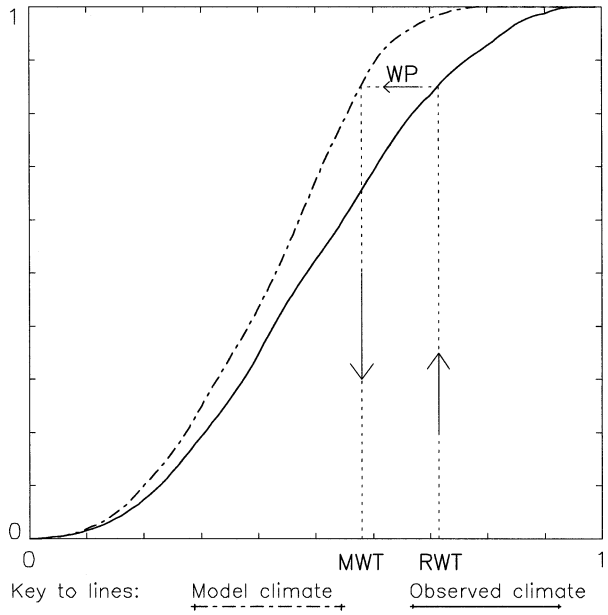

FIG. 1. An illustration of how model and site climatologies may be used to determine event probabilities. The horizontal axis is labeled in arbitrary units of weather severity. The vertical scale represents cumulative probability, so the probability of exceeding a cumulative value c is $1 - c$. Model Warning Threshold (MWT), Real Warning Threshold (RWT), Warning Percentile (WP).

point then defines the warning threshold for the model forecasts. For a single point this threshold can then be used to determine a forecast probability $P_{f/c}$ directly from the EPS forecast distribution as illustrated in Fig. 1. For FGEW, the warning thresholds defined in this way are used to calculate area probabilities in the same way as in the standard FGEW system, using several sites to represent each area. This method is 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 initialized at 1200 UTC each day.
- *102-member EPS (version C).* 102-member EPS obtained using the 1200 UTC operational 51-member EPS plus a second 51-member EPS run from 0000 UTC each day.
- *Climatology method (version D).* As version A, but calibrated using the climatology-based method described in section 4c.

TABLE 3. Verification statistics from postevent analysis of NSWWS Flash Warnings issued between Apr 1994 and Mar 1998 (Hymas 1995, 1996, 1997, 1998).

Year	Events	Flashes issued	Missed events	False alarms	Incorrect
1994/95	92	90	2	2	4
1995/96	111	109	2	4	6
1996/97	69	68	1	2	3
1997/98	116	116	0	2	2
Total	388	383	5	10	15

Also, NSWWS ''issued'' warnings will be referred to as ''N'' in some of the comparisons that follow. These versions are labeled on all of the figures and referred to throughout the text.

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. Verification of flash warnings has shown a very high correspondence with actual severe weather. Hymas (1993) reported only two misleading messages out of 83 flash warnings issued in the first 2 yr of the service. In subsequent years the numbers of flash warnings issued by forecasters increased. Table 3 summarizes verification figures for flash warnings issued between April 1994 and March 1998 from postevent analysis published in reports to Met Office customers (Hymas 1995, 1996, 1997, 1998). Out of 388 identified events, there were 5 missed events for which no flash warning was issued (1.3%) and 10 false alarms (2.6%), so warnings were assessed as incorrect on less than 4% of occasions. Given this high level of accuracy, flash warnings make convenient proxy observations for verification of early warnings, avoiding the need for complex analysis of several types of real observations.

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 4). For probability forecasts, contingency tables are determined for each of a range of probability thresholds, with the event deemed to be forecast

TABLE 4. Two-by-two contingency table of events for forecast verification.

	Forecast	Not forecast	Totals
Observed	H	M	$H + M$
Not observed	F	R	$F + R$
Totals	$H + F$	$M + R$	$H + M + F + R$

if the probability exceeds the threshold. The event is defined to be observed when a flash warning is issued.

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 nonevent for a correct rejection. The basic unit used is a calendar day, but rules were devised to avoid double counting where a warning or event spans 2 days.

- An early warning is judged to be a hit if any part of its validity period overlaps the period of a verifying flash warning. However, where an early warning spans more than 24 h and only one day is validated, the second will record a false alarm to penalize excessively long warnings.
- Where either an early warning or the verifying flash warning spans 2 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 h *and* flash warnings are valid on both days, in which case two hits are recorded.

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 an irregular set of probability thresholds is used for the contingency tables, with extra thresholds at low probabilities: 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.

In the NSWWS early warning probabilities are generated both for the United Kingdom 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 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 1200 UTC data reach forecasters around 0600 UTC the following day, so a 3-day ($D + 3$) FGEW warning can only be used two days ahead.

A major problem with verifying severe weather warnings is 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. Verification results

a. 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 4) together. Because both HR and FAR are stratified by the observations (there are $H + M$ events and $F + R$ nonevents), ROC measures the forecast system's ability to discriminate between occurrences and nonoccurrences of an event. Hit rate 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. For lower probability thresholds the event is forecast more frequently, resulting in higher values of both HR and FAR; corresponding points on the graph therefore appear toward the top right. A perfect forecast would have HR $= 1$ and FAR $= 0$, so for a skillful system the curve is bowed toward 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.

ROC curves for heavy-rainfall probabilities over the whole United Kingdom are shown in Fig. 2. 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 forecasting systems skill is greatest 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 $D + 4$. There is no significant difference between the different versions of FGEW. Also included in Fig. 2 are ROC points showing the HR and FAR for deterministic FGEW forecasts based on (i) the EPS control and (ii) the $T_L511L60$ high-resolution deterministic model. (The latter forecast is generated at twice the horizontal resolution of the EPS and with extra vertical levels; no separate FGEW calibration has been conducted for this model and the higher resolution allows it to resolve

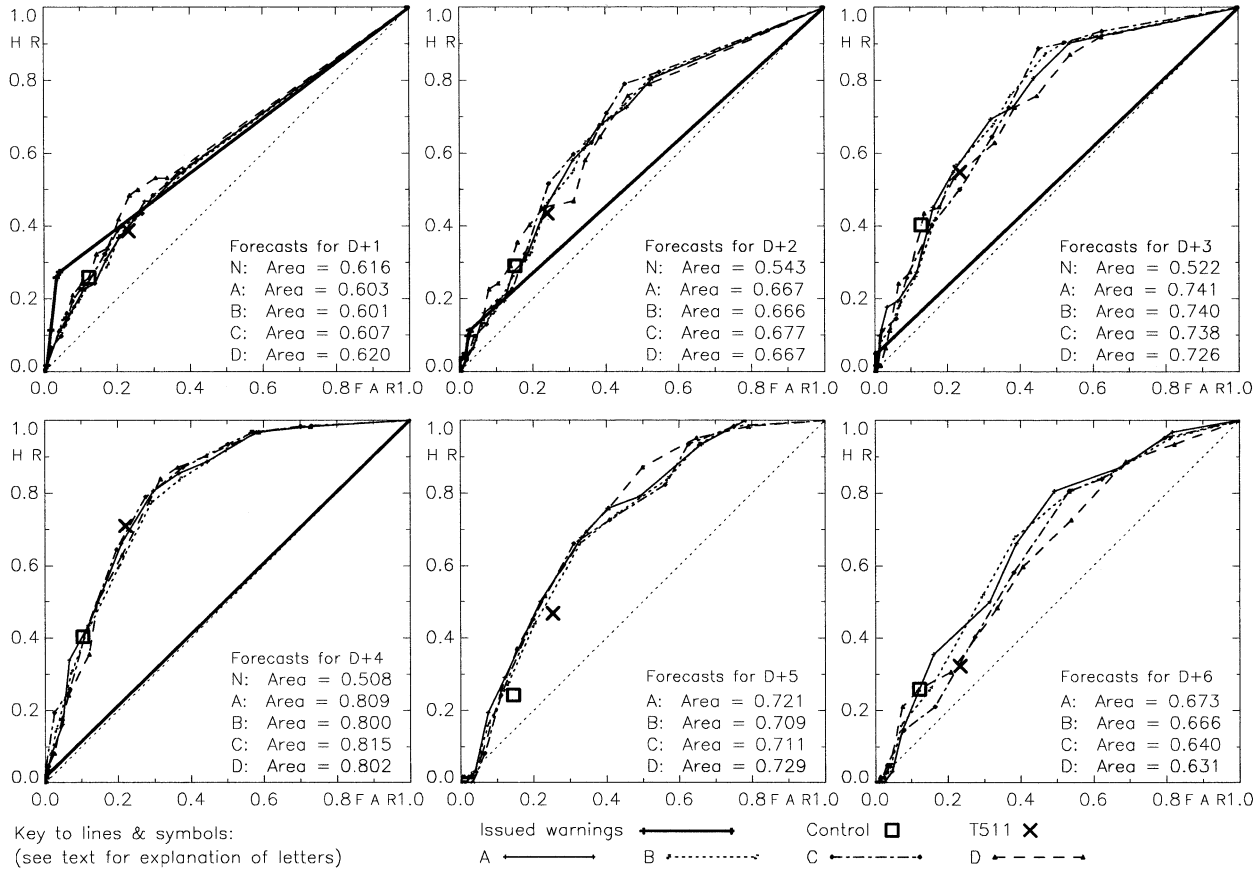

FIG. 2. ROC curves for probabilities of heavy-rainfall events occurring anywhere in the United Kingdom (plotted for all probabilities). Forecasts for 1–6 days ahead. Data period: 1 Oct 2001–12 Feb 2003. Letter codes denote versions of the FGEW system as defined in section 4d.

severe weather more frequently, with the result that both HR and FAR are higher for the $T_L511L60$ than for the EPS control.) It can be seen that these deterministic FGEW forecasts also display the same behavior, that the ROC point is closer to the top-left corner of the graph at $D + 4$ than at $D + 2$ or $D + 1$. This behavior is therefore associated with the ability of the models to forecast this type of severe weather, and is not specific to the ensemble methodology. This effect and its implications will be discussed in section 7.

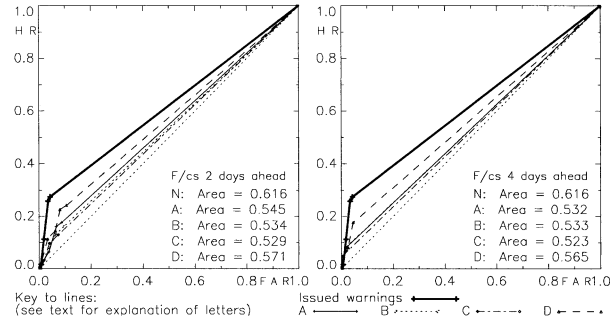

FIG. 3. ROC curves for probabilities of heavy-rainfall events occurring anywhere in the United Kingdom (plotted for probabilities ≥0.6 only). FGEW probabilities for (left) 2 and (right) 4 days ahead are plotted, with issued probabilities 1 day ahead overlaid for comparison. Data period: 1 Oct 2001–12 Feb 2003.

NSWWS warnings issued by forecasters (solid line) 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. 3, comparing FGEW 2- and 4-day forecasts with $D + 1$ issued warnings; 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 the $D + 1$ issued warnings are more skilful. 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

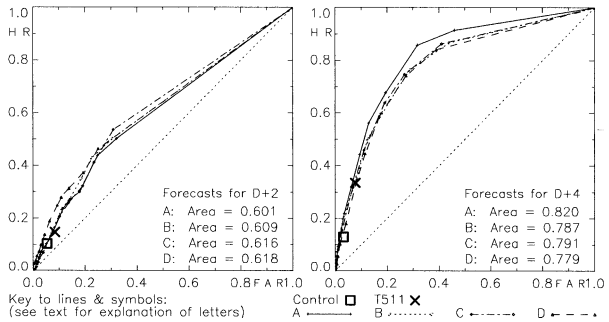

FIG. 4. ROC curves for FGEW probabilities of heavy-rainfall events occurring in individual areas (plotted for all probabilities) at (left) 2 and (right) 4 days ahead. Data period: 1 Oct 2001–12 Feb 2003.

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 areas are shown in Fig. 4, and are similar to those for the whole United Kingdom (Fig. 2). Results for severe gales are shown in Fig. 5, for 2 and 4 days ahead for the whole United Kingdom, and at 4 days for the individual areas; $D + 1$ issued warnings are overlaid. In each case the probabilistic FGEW skill is greatest at $D + 4$, but with most of the skill coming from the low probability thresholds. Results for the EPS control and $T_L511L60$ deterministic forecasts are also included. For heavy rain in individual areas (Fig. 4), the discriminating skill of the forecasts is again clearly greatest at $D + 4$. In the case of severe gales this is less clear. ROC points are close to the lower left of the diagram and there is no consistent pattern. This is probably due to the smaller data samples available for severe gale warnings, as the ROC point for a deterministic forecast is sensitive to the model behavior in individual forecasts when the sample size is small. Results for heavy snow warnings are not shown, but are very similar with the best performance at $D + 4$ for both probabilistic and deterministic forecasts.

b. 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 that strays below the $y = x$ line indicates overestimation of forecast probabilities. A near-horizontal curve would indicate a lack of event resolution in the forecasts, related to the ability to discriminate whether an event will or will not happen. Sharpness diagrams, histograms that indicate how often each probability bin was forecast, are normally plotted alongside to aid interpretation. In the reliability and sharpness diagrams presented here we have used the same set of thresholds to separate the bins as used for the ROC contingency tables above, focusing on low probabilities, except that we have merged the bins for highest forecast probabilities (above 60%) to reduce the effects of small sample sizes.

Figures 6–9 present a selection of reliability and sharpness diagrams for heavy rain and severe gale warnings. 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. All the reliability diagrams show 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 indi-

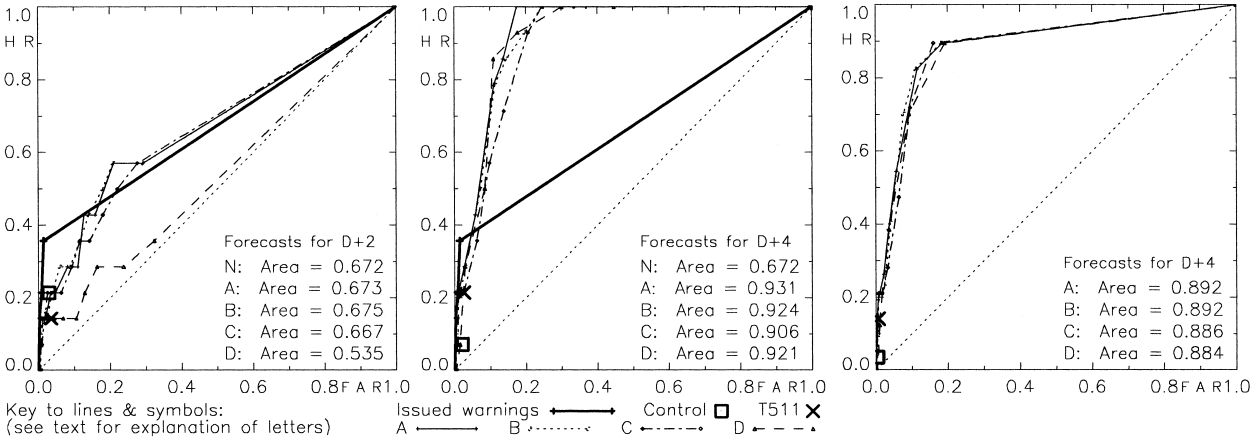

FIG. 5. ROC curves for probabilities of severe-gale events (plotted for all probabilities). (left) Probabilities for events occurring anywhere in the United Kingdom; FGEW 2 days ahead and issued 1 day ahead. (Center) Probabilities for events occurring anywhere in the United Kingdom; FGEW 4 days ahead 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.

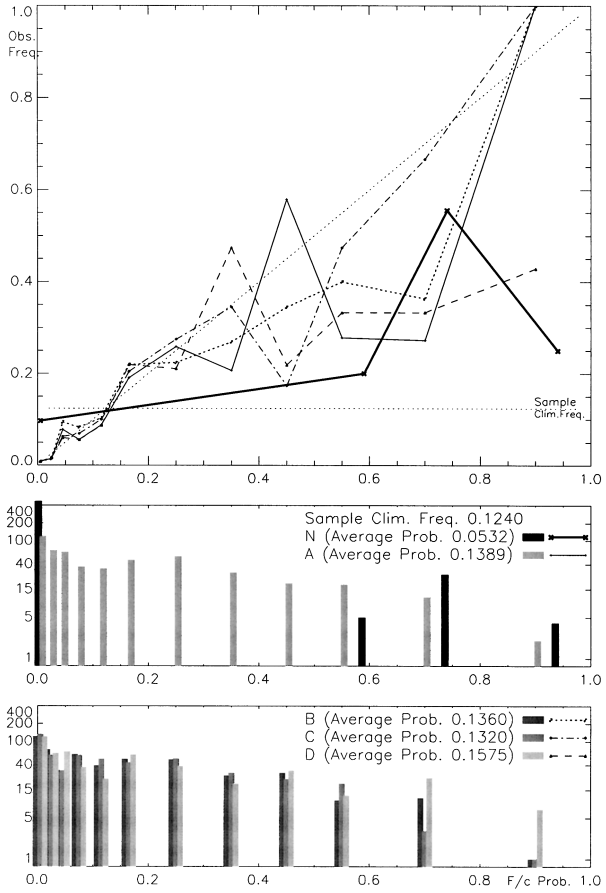

FIG. 6. (top) Reliability and (middle) sharpness for versions N and A; (lower) for versions B, C and D; note logarithmic vertical scales), for probabilities of heavy–rainfall events anywhere in the United Kingdom for FGEW warnings 4 days ahead and issued warnings 1 day ahead. Data period: 1 Oct 2001–12 Feb 2003. The reliability diagram plots the observed frequency of occurrence of severe weather as a function of forecast probability. Sharpness diagrams show the corresponding sample sizes for each category of forecast probability. Letter codes denote versions of the FGEW system as defined in section 4d.

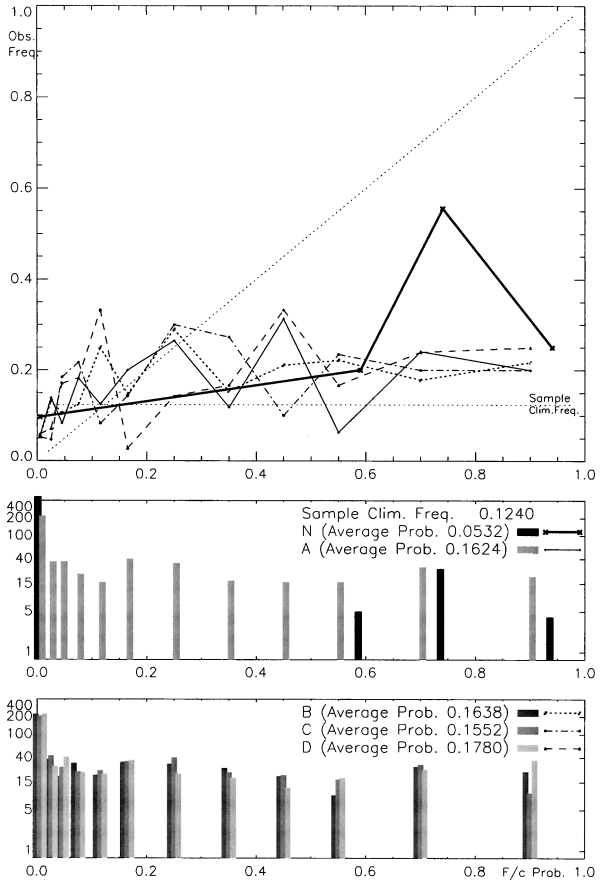

FIG. 7. As in Fig. 6, but for FGEW warnings 2 days ahead and issued warnings 1 day ahead.

cating 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 calibration of the FGEW warning event thresholds, using a prior set of verification data from the previous year (17 October 2000–4 May 2001).

Figure 6 shows reliability curves for heavy rainfall, for probabilities of events in the United Kingdom, 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. Occurrence of severe weather is substantially above the climatological frequency and higher than for lower forecast probabilities, providing useful forecast information, but there is a clear tendency to overestimate the probability, indicated by the curves falling below the ideal diagonal. 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 an issued probability of zero which lies only slightly below the sample frequency. This is partly 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.

Figure 7 shows results for $D + 2$ FGEW forecasts. 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

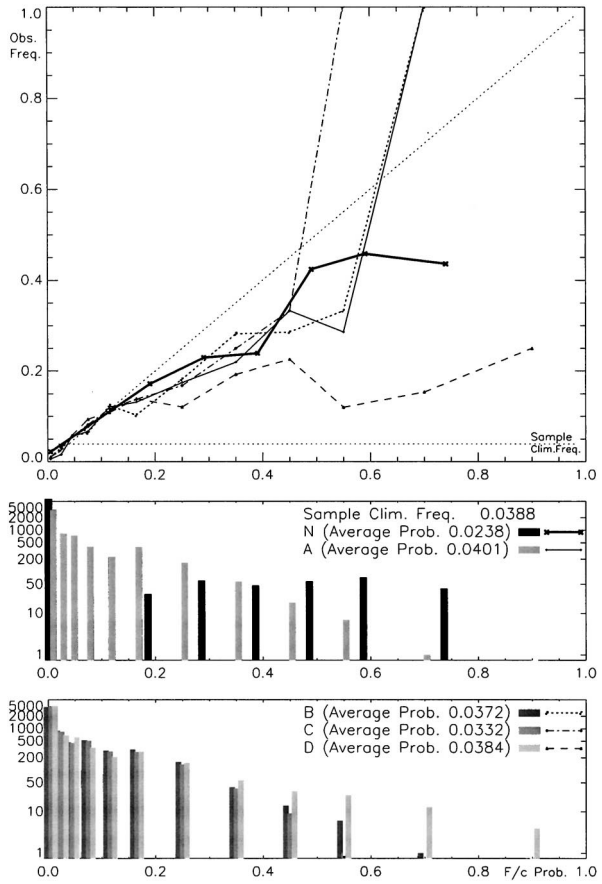

FIG. 8. As in Fig. 6, but for probabilities of heavy-rainfall events in individual areas, for FGEW warnings 4 days ahead and issued warnings 1 day ahead.

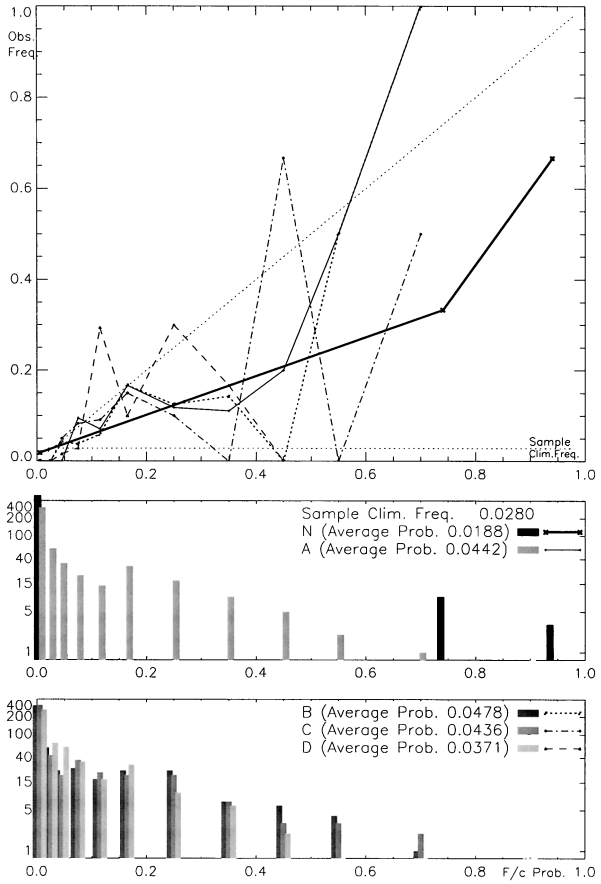

FIG. 9. As in Fig. 6, but for probabilities of severe-gale events anywhere in the United Kingdom, for FGEW warnings 4 days ahead and issued warnings 1 day ahead.

consistent with the ROC results that indicated that there was much better discrimination, which is closely related to resolution, at $D + 4$ than at $D + 2$. Comparing the sharpness diagrams of Figs. 6 and 7 reveals that the main difference is that at $D + 2$ the EPS forecasts the highest (and lowest) probabilities more frequently than at $D + 4$ but with no corresponding increase (decrease) in the occurrence of severe weather. This would suggest that poor performance of the ensemble at $D + 2$ could be related to lack of ensemble spread, although this cannot explain the better performance of the deterministic forecasts at $D + 4$ seen in the ROC results.

Figure 8 presents reliability diagrams for heavy rainfall in the individual areas of the United Kingdom 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 overconfidence (i.e., high probabilities are too high, low probabilities too low). The issued warnings have to be interpreted with care, remembering that the sample is selective because warnings for individual areas 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 overforecasting more severely at higher probabilities. This version appeared in Fig. 3 to offer better discrimination of events at high probability, but it can now be seen that this is at the cost of significant overforecasting, 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 9 presents results for severe-gale events for the whole United Kingdom 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

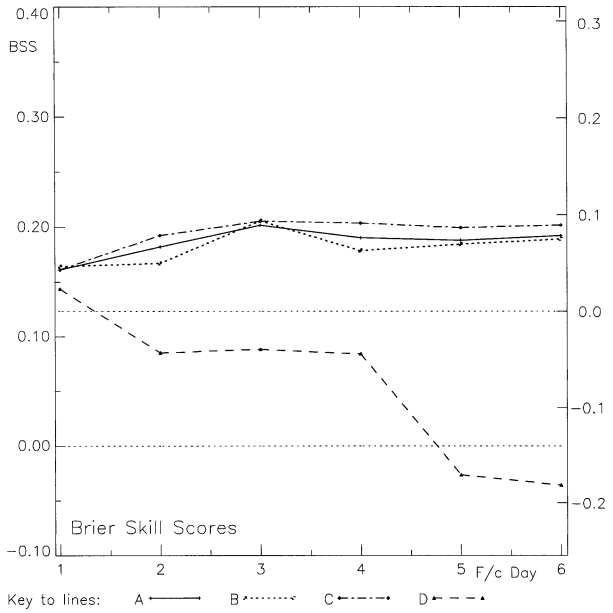

FIG. 10. BSS (left-hand axis is skill w.r.t. null forecasts, right-hand axis is skill w.r.t. climate forecasts) for probabilities of heavy-rainfall events occurring anywhere in the United Kingdom for FGEW warnings 1–6 days ahead. Data period: 1 Oct 2001–12 Feb 2003. Letter codes denote versions of the FGEW system as defined in section 4d.

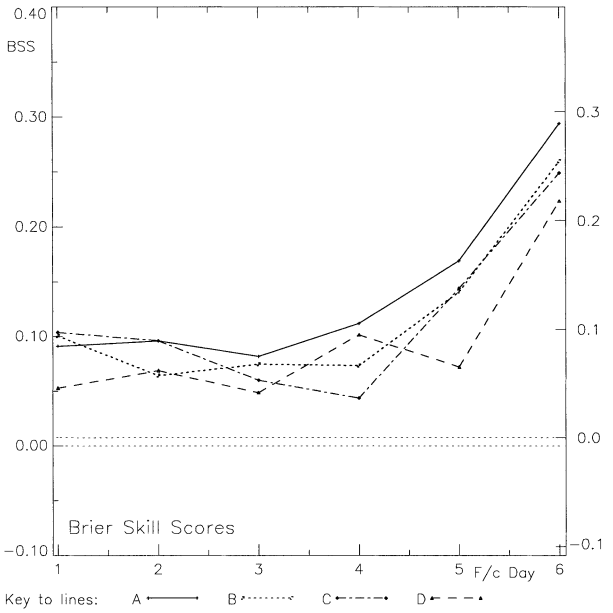

FIG. 11. As in Fig. 10, but for probabilities of severe-gale events anywhere in the United Kingdom.

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.

c. Brier skill scores

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

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

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 (nonevent). The Brier score 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 $\mathrm{BS}_{\mathrm{fc}}$ with that obtained by some reference forecast $\mathrm{BS}_{\mathrm{ref}}$:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{fc}}}{\mathrm{BS}_{\mathrm{ref}}}. \tag{2}$$

Typical reference forecasts used are climatology or persistence. Skilful forecasts have positive BSS; a negative BSS indicates a forecast worse than the reference forecasting system. For early warnings, no prior climatological probability was available, 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. In addition, a crude estimate of the climatological frequency of flash warnings was taken from the initial training period used for tuning. In Figs. 10–12, 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.

Figure 10 shows BSS for FGEW warnings of heavy-rainfall events anywhere in the United Kingdom. BSS are positive relative to both null forecasts and the crude climatological forecasts throughout $D + 1$ to $D + 6$, except for the climatology-based 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. The 102-member ensemble (C) performs slightly better still, although it is unlikely that the differences are statistically significant. The climatology-based system (D) is relatively poor, because, unlike the other methods, it has not been tuned explicitly to provide reliable probabilities. The skill of FGEW warning probabilities for events in individual areas (not shown) is lower than for anywhere in the United Kingdom. BSS are not shown for NSWWS issued warnings as the 60% threshold prevents valid calculation.

BSS for severe-gale events (Fig. 11) 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

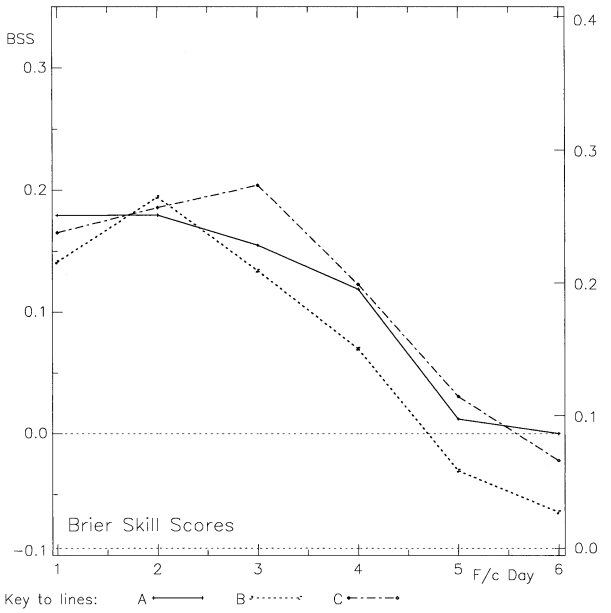

FIG. 12. As in Fig. 10, but for probabilities of heavy-snowfall events anywhere in the United Kingdom.

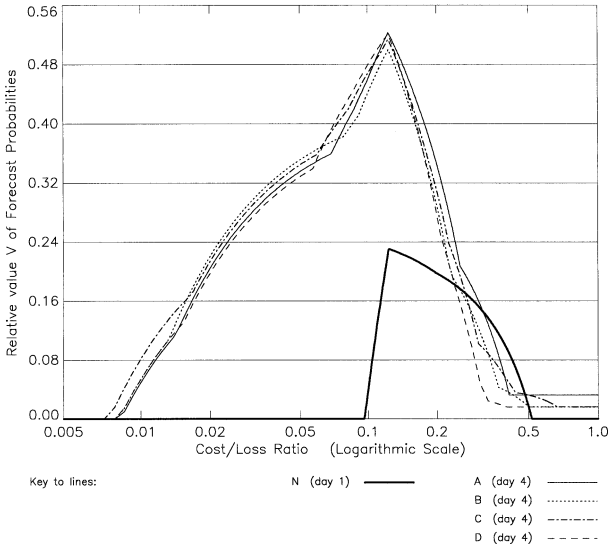

FIG. 13. Cost–loss diagrams for heavy-rainfall events anywhere in the United Kingdom. Issued warnings for 1 day ahead, compared with FGEW probabilities for 4 days ahead; line styles are used according to legend. Data period: 1 Oct 2001–12 Feb 2003. Letter codes denote versions of the FGEW system as defined in section 4d.

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. 12) 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 null forecasts and to the crude estimate of climatology available. Details of the variation in behavior 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.

d. Cost–loss analysis

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 is 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 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 be used to estimate the relative economic value of a forecast system for a range of user C/L (Richardson 2000).

Figure 13 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 0.12 consistent with Fig. 6. 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, because the restriction preventing the issue of warnings with probability below 60% prevents such users from optimizing their decision-making. The maximum value of $D + 4$ FGEW forecasts is much greater than for the $D + 1$ issued warnings, and $D + 4$ FGEW warnings have more user value than the $D + 1$ issued warnings for all except a few users with C/L around 0.45. This is because the peak value, at $C/L \sim 0.12$, is not well-matched to the 60% threshold limit. (One unusual feature of the cost-loss value curves in Fig. 13 is that some of them do not fall to zero at the extreme high and low C/L values, but reach a fixed nonzero value. This is an effect of the small sample sizes available in the verification data, because 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 toward the lowest C/L ratios, as the larger ensemble size improves the chance of capturing events at low probabilities.

Figure 14 shows the same cost–loss curves for heavy-rainfall warnings over the whole United Kingdom given

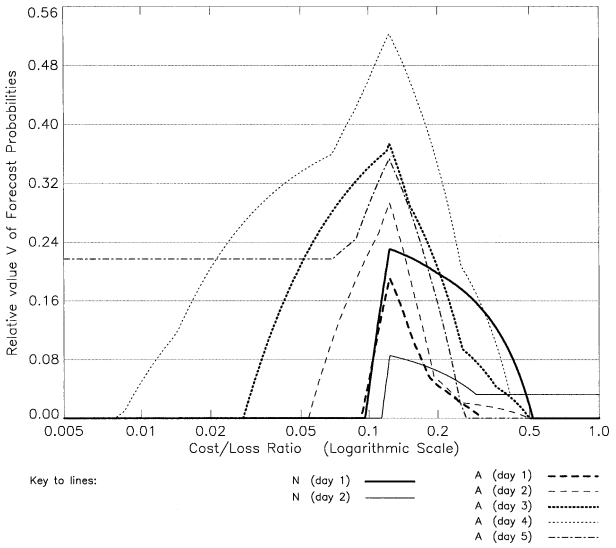

FIG. 14. As in Fig. 13, but for issued warnings for 1–2 days ahead, compared with operational FGEW probabilities for 1–5 days ahead.

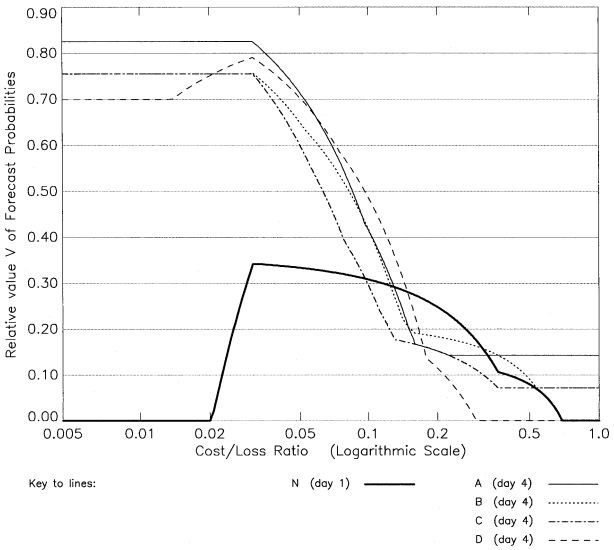

FIG. 15. As in Fig. 13, but for severe-gale events.

by the operational version A of FGEW at lead times of 1–5 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 (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 $D + 1$.

Figure 15 is similar to Fig. 13, 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.

7. Discussion

A consistent result obtained throughout the FGEW verification described above is that the forecasts have greatest skill at $D + 4$, with much less at $D + 2$. This result has proved extremely robust for all the weather types verified (heavy rainfall, severe gales, and heavy snowfall) and is not affected by the calibration described in section 4b. Calibration, performed by balancing the forecast bias so that the mean forecast probability is close to the climatology of the training sample, alters the absolute values of the area under the ROC curves but not the fact that $D + 4$ forecasts provide the largest ROC area. As shown in Figs. 6 and 7, once the calibration has been performed the $D + 4$ forecasts also come closest to providing reliable probability forecasts. While this result is most apparent in the probability forecasts from the EPS, it was shown in Figs. 2, 4, and 5 that it also applies to deterministic forecasts based on either the EPS control or the high-resolution $T_L511L60$ model. It therefore appears to be characteristic of the ECMWF NWP system for the type of extreme weather events being predicted by FGEW, and not caused by anything specific to the EPS, such as the perturbation methodology. T. Palmer (2003, personal communication) has noted that for some previous extreme weather events, notably the 16 October 1987 storm over southern England, numerical forecasts initiated several days ahead have been better than shorter-range forecasts. A possible explanation of the consistency between deterministic and probabilistic results is that when the control forecast is inclined to evolve in a particular direction, either toward or away from severe weather, the ensemble perturbations are not strong enough to divert it from this path in the early part of the forecast, and it is only by allowing the perturbations to evolve nonlinearly over a longer period of around 4 days that the ensemble becomes able to produce a more realistic sampling of the likelihood of severe weather. It is important to note that this should not be seen as a criticism of the EPS perturbations, which perform well at the medium-range for which they are designed.

By verifying forecasts of severe events over a sustained period of time, the FGEW verification may thus provide some quantification of a problem with short-range NWP which has not previously been documented. Most previous verifications have been for less extreme weather events and have not shown such behavior. For example Barkmeijer et al. (1999) showed that for EPS forecasts of 500-hPa height anomalies the ROC area decreased steadily onwards from $D + 2$, the optimization time of the SV perturbations. Buizza et al. (2000)

showed similar results for precipitation forecasts. One of the few independent EPS verifications carried out for similarly extreme events, and which provides some support for the current results, was by Mullen and Buizza (2001, 2002) who verified rainfall forecasts over the United States. They reported a significant dip in performance at day 2 for the heavier rainfall thresholds (20 and 50 mm per day) only, although this was most apparent in the BSS, specifically in the reliability term.

Although performance is better at $D + 4$ than $D + 2$ for deterministic as well as probabilistic forecasts, the extra verification information available from the ensemble forecasts may offer some help in identifying the cause. Mullen and Buizza (2001) showed that the EPS is underdispersive for precipitation forecasts, more so at $D + 2$ than at $D + 5$. Underdispersive ensemble forecasts typically result in overconfident probability forecasts, with reliability diagrams having a slope of less than 45°, as observed in the FGEW verification. Thus it is likely that the better probabilistic performance of FGEW at $D + 4$ than at $D + 2$ is due to an improved ensemble spread. Some evidence to support this is provided by the sharpness diagrams included alongside the reliability diagrams in Figs. 6 ($D + 4$) and 7 ($D + 2$). 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. The numbers of forecasts of zero probability are also reduced, from around 200 at $D + 2$ to around 130 at $D + 4$. Most of these zeroes become only low probabilities, but the effect is greatly to improve the reliability curve between 0 and 30% probabilities, almost eliminating events occurring when forecast probability is zero. Thus the spread of the forecasts at $D + 4$ appears to capture the uncertainty better than that at $D + 2$, which is consistent with the hypothesis above that the EPS perturbations are not strong enough to divert the model from its deterministic path in the early days of the forecast.

This hypothesis can explain why the ensemble behaves in the same way as the deterministic forecast, but not why the deterministic forecast is poorer at short range. This may be related to the ability of the model to spin up severe weather developments. It was noted above that the improvement in the forecasts at $D + 4$ compared to $D + 2$ came from a reduction in the numbers of both missed events and false alarms. This also applies to the deterministic forecasts from the control and $T_L511L60$ model. Thus any spinup issue relates not just to an ability to develop severe weather (missed events), but also on occasions to develop less-severe weather (false alarms). Thus, this suggests that it is by evolving over a longer period of time, when nonlinear effects have more opportunity to significantly change the forecast evolution, that the model is better able to simulate the real probability of an extreme development. Why this should be is not clear and requires further research. One interesting question would be whether this behavior is specific to the ECMWF NWP system, or whether any other models behave similarly for extreme events.

The finding that ensemble forecasting may require a longer period of perturbation growth to provide reliable probabilities for more extreme events could have considerable implications for future developments of ensemble prediction. Most current operational ensembles, such as the ECMWF EPS, are designed for medium-range prediction, but current research is moving toward short-range ensembles, and a major aim of this work is to provide probabilities for severe-weather events. If the hypothesis above is correct, this would suggest that alternative perturbation strategies having more impact early in the forecast period may be required for successful short-range ensemble prediction. A simple amplification of the EPS perturbations would not be appropriate as this would result in overdispersion of the ensemble in medium-range predictions and possibly also at the short-range for nonsevere events. Various alternative strategies are under investigation, for example moist SVs optimized over a shorter period (Coutinho et al. 2004) or the Ensemble Transform Kalman Filter (Bishop et al. 2001; Wang and Bishop 2003), and some of these may prove more effective for short-range prediction of severe weather. Alternatively, it could be that to provide useful probabilities of severe weather requires a longer period of ensemble growth regardless of the perturbation strategy employed. If this is the case then it may severely limit the potential for use of ensemble prediction in the short range, at least until it is possible to run models at sufficiently high resolution to fully resolve the nonlinear processes, such as convection, which are most important at the short range.

8. 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 nonlinear 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 percent occurrence was considerably lower than forecast. Nevertheless the actual percent occurrence was substantially higher than climatology on these occasions, and this therefore still represented useful forecast information which could potentially 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 time scales. 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. Deterministic forecasts based on a single model forecast display this same characteristic, so this is not related to the EPS perturbation methodology. This behavior has not been observed for less extreme events where the best performance has been around 2 days or less. This clearly illustrates that we cannot assume that, because an EPS performs well for one forecast range or event threshold, it will do so at another. It is important to verify the performance at all ranges and thresholds of interest, and then only apply the results operationally where appropriate according to the verification results.

Since the difference in performance at $D + 4$ and $D + 2$ is observed in deterministic as well as probabilistic forecasts, the root cause clearly lies in the basic NWP system, either the data assimilation or the ability of the model to spin up severe weather correctly. This merits further investigation. While Mullen and Buizza (2001, 2002) observed some similar behavior with extreme rainfall thresholds, we are not aware of any other verification showing such a strong improvement in performance at $D + 4$ compared to $D + 2$. Similar independent verification of other forecast systems for extreme or rare events would be of great interest to confirm whether similar behavior is seen, or whether this is peculiar to the FGEW method of diagnosing and verifying severe weather.

Although the root cause of this behavior does not lie in the EPS perturbations, we would ideally hope that an ensemble would provide perturbed forecasts spanning the uncertainty better. Results suggest that the EPS perturbations may not be strong enough to sample the full uncertainty at $D + 2$. Alternative perturbation methods, providing more impact in the short range, may be required for successful short-range ensemble prediction. However we cannot discount the possibility that for extreme events, when we are sampling in the tail of the forecast distribution, a longer period of nonlinear evolution may always be required before the ensemble can provide a random sampling of the pdf.

Alongside the operational version of FGEW (A) we also tested a number of experimental versions. The differences between these systems were in fact relatively small, and the resulting differences in performance were also small and not statistically significant. 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 1200 UTC EPS run with the experimental 0000 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 the least skillful, particularly in terms of Brier score for severe gales. This is not surprising since the thresholds in the other versions have been tuned using a previous training set to optimize the probabilities, but the climatology method offers a useful approach for setting up an initial calibration that 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.

Acknowledgments. The authors would like to thank Tim Palmer for discussions and suggestions on this work, which led to significant improvements in our understanding of the results. Suggestions from David Richardson and three anonymous reviewers also greatly helped to improve this paper.

REFERENCES

Barkmeijer, J., R. Buizza, and T. N. Palmer, 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., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.,* **129,** 420–436.

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

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

Coutinho, M. M., B. J. Hoskins, and R. Buizza, 2004: The influence of physical processes on extratropical singular vectors. *J. Atmos. Sci.,* **61,** 195–209.

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

Hymas, K., 1993: The Meteorological Office National Severe Weather Warning Service (NSWWS). *Meteor. Mag.,* **122,** 53–61.

——, 1995: National Severe Weather Warning Service (NSWWS) April 1994 to March 1995. Annual Rep. to Met Office Customers, Met Office, Bracknell, United Kingdom, 11 pp.

——, 1996: National Severe Weather Warning Service (NSWWS) April 1995 to March 1996. Annual Rep. to Met Office Customers, Met Office, Bracknell, United Kingdom, 11 pp.

——, 1997: National Severe Weather Warning Service (NSWWS) April 1996 to March 1997. Annual Rep. to Met Office Customers, Met Office, Bracknell, United Kingdom, 11 pp.

——, 1998: National Severe Weather Warning Service (NSWWS) April 1997 to March 1998. Annual Rep. to Met Office Customers, Met Office, Bracknell, United Kingdom, 11 pp.

Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. J. Roy. Meteor. Soc.,* **129,** 3037–3057.

Legg, T. P., K. R. Mylne, and C. Woolcock, 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.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

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

——, R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

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

——, and ——, 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., F. Molteni, and T. N. Palmer, 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.

——, C. Woolcock, J. C. W. Denholm-Price, and R. J. Darvell, 2002: Operational calibrated probability forecasts from the ECMWF Ensemble Prediction System: Implementation and verification. Preprints, *Joint Session of 16th Conf. on Probability and Statistics in the Atmospheric Sciences and of Symposium on Observations, Data Assimilation, and Probabilistic Prediction,* Orlando, FL, Amer. Meteor. Soc., 113–118.

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.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667.

——, 2001: Ensembles using multiple models and analyses. *Quart. J. Roy. Meteor. Soc.,* **127,** 1847–1864.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: A survey of common verification methods in meteorology. WMO WWW Tech. Rep. 8, WMO TD No. 358, 114 pp.

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

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.,* **60,** 1140–1158.

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

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