JP1.24

EARLY WARNINGS OF SEVERE WEATHER FROM THE ECMWF ENSEMBLE PREDICTION SYSTEM

Kenneth R. Mylne* and Timothy P. Legg
Met Office, Bracknell, United Kingdom.

1. Introduction

One of the aims of ensemble prediction is to improve the forecasting of severe weather. To the extent that the development of severe weather is frequently highly non-linear and therefore sensitive to forecast errors, this is an appropriate application of ensembles; at the same time it is a particularly demanding application, and is also difficult to verify since severe weather occurs relatively rarely so data samples are small.

This paper describes a project in the Met Office to attempt to use the Ensemble Prediction System (EPS) (Molteni *et al,* 1996*)* run by ECMWF (European Centre for Medium Range Forecasts) to generate early warnings of severe weather in support of the UK National Severe Weather Warning Service (NSWWS). The NSWWS provides several tiers of warning to the public and the emergency services. The type of warning considered here are Early Warnings. These can be issued up to 5 days in advance when the probability of an event occurring "somewhere in the UK" is 60% or more. In addition to an overall UK probability, probabilities are also given for 12 local regions. Since the warnings are probabilistic by definition, they are well suited to an ensemble approach. In practice warnings have only rarely been issued more than 36h in advance and around 24h is much commoner - one of the aims of the project was to provide forecasters with more information to help give them the confidence to issue warnings further in advance.

2. Predictability of Severe Weather

The defined requirement for the issue of Early Warnings in the NSWWS is a probability of 60%. However it is interesting to speculate on how often this is likely to be predictable for severe weather at more than about 24h ahead. Evidence from the December 1999 storms over France and Germany showed that only a small proportion of ensemble members (or of deterministic forecasts from different centres) succeeded in predicting severe storms, even at ~24h ahead. Figure 1 illustrates schematically that in a synoptic situation when severe weather is possible, once a forecast moves into the chaotic non-linear regime, most ensemble members are likely to be drawn towards the model's climatology. (Although the diagram illustrates this idea with the central control forecast predicting severe weather and perturbed analyses leading to less severe conditions, this argument is just as true when it is one or more perturbed ensemble members which predict severe conditions.) The result of this is that the forecast pdf (probability density function) is always likely to be skewed away from severe weather. Thus, although the ensemble can be expected to include members with severe events, it would be unusual for it to predict high probabilities of severe weather. The EPS in its current formulation is designed on an assumption that the evolution of the atmosphere is normally quasi-linear over approximately the first 48h, which would suggest that higher probabilities might be obtained within this time-range. However Smith and Gilmour (1999) have found that typically there is important non-linearity in forecasts at <48h. Since the development of severe weather is likely to involve non-linear processes, this is particularly likely to be true when the state of the atmosphere is such that severe weather is possible. For severe weather situations the quasi-linear limit may be much less than 24h, and there may be little chance of predicting high probabilities of extreme conditions.

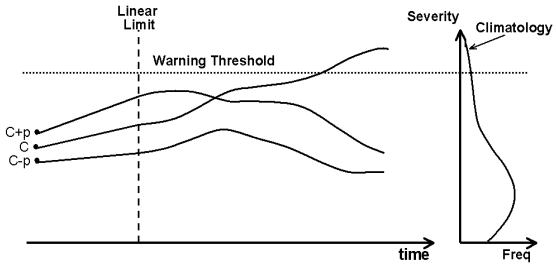


Fig. 1 Schematic illustration of the effect of non-linearity on an ensemble forecast. In the early stage of a forecast, ensemble members diverge quasi-linearly. In later stages, even when one member predicts severe weather, most members can be expected to be drawn towards model climatology.

Given this analysis, it is unlikely that an Early Warnings system based on the EPS is going to be able to capture the majority of severe weather events several days in advance with a 60% probability. Indeed, since the above analysis applies equally well to the real atmosphere as to a model, it can be argued that the occurrence of severe weather is fundamentally a low probability event in the atmosphere, and thus that on most occasions it should only be appropriate to

* *Corresponding author address*: Ken Mylne, Met Office, Bracknell, Berks RG12 2SZ, UK.
email: Ken.Mylne@metoffice.com

issue warnings at low probabilities. This suggests that the 60% threshold will results in many events being missed. Nevertheless it was still considered valuable to develop an ensemble-based system in the hope that it might provide useful alerts to forecasters. Also, a new tier of warnings was recently introduced allowing issue of warnings at lower probabilities on rare occasions when there is a risk of exceptionally severe conditions.

3. Scanning the Ensemble

The EPS-based system attempts to support the following NSWWS Early Warnings events:

- *Severe Gales* - gusts of 70 mph or more
- *Heavy Rain* - at least 15mm within a 3-hour period
- *Heavy Snow* - 2cm/hour or more for at least two hours

These events are very demanding for an NWP model, and proxy events had to be defined to represent these in the model output. For example, for *gusts>70mph* we define a gust factor, based on empirical rules, to estimate gusts from the 10m mean wind speed. To define when the proxy elements reach a sufficiently severe level to trigger warnings, thresholds were initially calibrated using analysis fields from past cases when Flash Warnings had been issued. (Flash Warnings are another tier of warning in the NSWWS, issued for the same events as Early Warnings, but at very short range when there is a high degree of certainty. To avoid the need for complex identification of real events for calibration and verification, Flash Warnings were used to define "observed" severe weather events.) In practice this approach produced thresholds which were too low, and led to an excessive over-forecasting bias for most events. Thresholds were subsequently re-tuned using verification data from the winter season 2000/2001 to correct these biases.

In order to estimate probabilities of events, all ensemble members are scanned over grid-points covering the UK. Probabilities are derived for an event occurring “somewhere in the UK” (used for the 60% issue threshold) and also in 12 sub-regions. Probabilities are defined simply by the proportion of ensemble members predicting severe weather within the defined regions (specified for the NSWWS). In order to define the probabilities of an “event”, it is important to allow for uncertainties in both where and when it may occur, since what is essentially the same severe weather event may develop at slightly different locations or times in different ensemble members. For the purposes of an Early Warning the requirement is only to know that something may occur in some part of the UK on a particular day - extra detail of time or place can be added nearer the time. Thus an ensemble member is counted if it exceeds the severe weather threshold at *any* grid-point within a region, and within a time-window to allow for timing differences between ensemble members. Initially longer time-windows were used for longer-range forecasts, on the basis that timing errors would increase the further ahead the forecast. However this resulted in a greater over-forecasting bias for longer-range forecasts, so the time-window was later fixed at ±6h for all lead-times.

Alerts are issued to forecasters when UK forecast probabilities exceed 20%, and recommendations to issue warnings at over 60%.

4. Verification

As noted above, Early Warnings are verified against Flash Warnings of the same weather events. This can cause problems when more than one type of severe weather occurs together, but on most occasions provides a good measure of real events.

The greatest problem of verifying a severe weather prediction system is that actual events are rare and hence data samples are small. This problem is particularly acute for probabilistic forecast systems, as probabilities can only be assessed over large samples. Results are given below for the period from 17th October 2000 to 4th May 2001, and the effects of small data samples are clearly apparent so it is important to interpret results with care. To date this is the only data available for tuning the system, so it is important to remember results shown here represent the best *potential* skill of the system - warnings will be verified using independent data over the 2001/02 winter to give a better assessment of the true skill of the system.

4.1 ROC

ROC (Relative Operating Characteristic) (see Stanski *et al,* 1989*)* measures the decision-making ability of a forecast system in terms of hit rates (HR) and false alarm rates (FAR). ROC for a probability forecast is plotted as a graph of HR against FAR, with points defined for a range of probability thresholds. (For any threshold p_t, the event is deemed to have been forecast if the forecast probability exceeds p_t.) A skilful system will give a ROC curve bowed towards the upper left corner (HR>FAR) while a system with no skill will have HR=FAR. A summary score of the ROC skill is given by the area under the curve, with values greater than 0.5 representing useful skill, and 1.0 representing perfect deterministic forecasts. There has been considerable discussion of the optimal way to calculate the area under the ROC curve. Wilson (2000) argued that a curve should be fitted to the data, but it can be argued that this gives a theoretical limit for an infinitely large ensemble. It is common instead to plot a set of straight lines joining points for (FAR,HR) generated at standard values of p_t., often at 10% intervals. In this study ROC curves are plotted with points at probability thresholds of 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. Extra points from low probability thresholds are included to allow assessment of skill at predicting events with low probability. Low probabilities are of no direct interest for NSWWS Early Warnings, but may nevertheless be useful for alerting forecasters.

ROC curves for warnings of Severe Gales issued by the EPS-based Early Warnings system at 1 to 6 days

ahead are shown in grey in figure 2. These results are based on probabilities calculated for the "whole UK", and so are relevant to the decision on whether to issue a warning. Warnings issued by forecasters are shown in black for comparison. (It should be noted that the EPS forecasts only become available for issue approximately 18 hours after data time, so for practical purposes the day 4 EPS forecasts should be compared with day 3 forecaster warnings, for example.) It is notable that the EPS forecasts at 4 days ahead were better than those at shorter range, as the ROC curve is bowed closer to the top left corner. This effect was even more marked for Heavy Rain and Heavy Snow warnings, since the 2-day ROC curve was not as good as here for Severe Gales. For 5 and 6 days ahead the ROC skill declines, as might be expected looking further ahead.

It is important to note that the points of the curve start with the lowest probability threshold nearest the top right, so much of the "area under ROC" above the no-skill line actually comes from low-probability forecasts.

For comparison, the forecasters had good skill at day 1, with progressively lower hit rates further ahead. Their false alarm rates were consistently very low, because they are known to be reluctant to issue warnings until they are very confident.

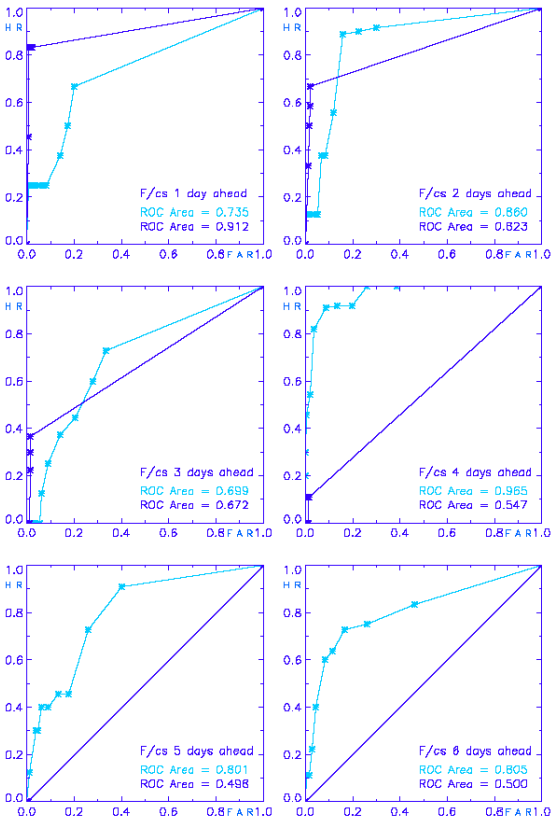


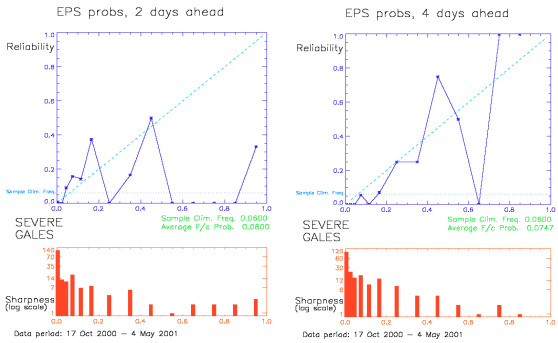
Fig. 2: ROC curves for warnings of Severe Gales "anywhere in the UK" issued by the EPS-based system (grey) and Met Office forecasters (black) at 1, 2, 3, 4, 5 and 6 days (see captions).

This result that the EPS-based system performs better at day 4 than at earlier days is remarkable. As noted earlier the results here are shown after tuning of the system to this set of results - however this result, that day 4 forecasts gave better ROC curves than shorter period forecasts, was also observed before re-tuning, and was in fact relatively insensitive to the tuning applied to the system. Possible reasons for this result will be discussed below.

4.2 Reliability Diagrams

Reliability diagrams (Wilks, 1995) plot the observed frequency of occurrence of an event against the forecast probability, so the ideal curve lies along the main diagonal of the graph. Reliability diagrams for the optimised system for severe gales and heavy rain warnings at 2 and 4 days ahead are shown in figure 3. Because the system has been tuned to this data these graphs can only represent the maximum potential skill of the system.

(a)



(b)

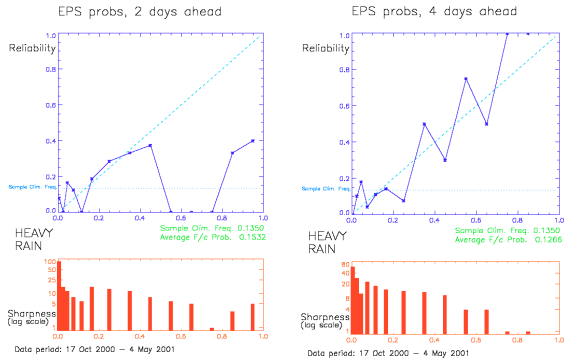


Fig. 3: Reliability (top) and Sharpness (bottom) diagrams for warnings of (a) Severe Gales and (b) Heavy Rain at 2 (left) and 4 (right) days.

Under each reliability diagram is plotted a sharpness diagram, a histogram showing the number of times each forecast probability has been issued. These show that sample sizes for high-probability warnings are very low (as predicted in section 2) and this results in the reliability diagrams being noisy. Considering this, the results at 4 days ahead are encouraging. Although

only a few higher probability forecasts were issued, when they were the severe weather was correspondingly more likely to occur.

By contrast the reliability diagrams for 2-day forecasts show there was virtually no resolution of whether the event was likely to occur - reliability curves are almost horizontal. The only positive feature is that when the forecast probability was zero, severe weather was very unlikely. Results for 3-day forecasts (not shown) were hardly any better than 2 days. These are consistent with the ROC results, since ROC is closely related to resolution.

5 Discussion

Results for the Early Warnings system, based on 6½ months of data since the latest upgrade to the ECMWF EPS, show that the system has some useful resolution in 4-day forecasts. This ability of the system to identify when the probability of severe weather is high was quite insensitive to any re-tuning of the system. Event thresholds used initially led to quite severe over-forecasting, but using a process of 'calibration by assessment' it has been shown that this over-forecasting can be effectively eliminated, giving a potential for reliable probabilistic forecasts at D+4. This calibration has, however, not yet been tested using independent data, due to the small data samples available for analysis. Results in the coming season are unlikely to be quite as good as shown here. It must also be noted that future performance may be somewhat different following correction of a fault in the ECMWF EPS, which was later found to have been affecting its performance during part of the period when these results were accumulated. The fault led to a much-reduced ensemble spread. 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 probabilities of severe weather.

It is interesting to consider why the system performed so much better at day 4 than at earlier times. The EPS is purposely designed for medium-range use, and at D+1 the dynamically active 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 was still poor. The most likely cause is the EPS fault described above, which meant the perturbations were smaller than they should have been, preventing effective ensemble growth in the early stage of the forecasts. It is therefore hoped that over the coming winter season, with the fault corrected, the day 3 forecasts, and perhaps also day 2, will be improved.

Overall, initial results from these experiments are encouraging but further verification from independent data is required before the true skill of the system can be evaluated. As anticipated in section 2, occasions on which severe weather can be predicted several days in advance with a high probability, say 60% as required by the NSWWS, are relatively rare although they do occur. Consequently, miss rates of warnings can be expected to remain relatively high as long as the 60% threshold for issue continues to be used. Nevertheless it is hoped that the good reliability of the 4-day warnings from the system will help to encourage forecasters to issue some warnings earlier than they have done in the past.

6. Future Work

Further verification will be carried out over the 2001/02 winter, and this may lead to further calibration of the event thresholds. Work is also planned to experiment with alternative proxies for the severe weather events, for example use of a wind gust parameter recently developed at ECMWF.

References

Molteni,F., Buizza,R., Palmer,T.N., Petroliagis,T., 1996: The ECMWF Ensemble Prediction System: Methodology and Validation *Q.J.R Met.Soc*. **122**, 73-119.

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

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

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

Wilson,L.J., 2000: Comments on "Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System", *Weather and Forecasting*, **15**, 361-364**.**