J1.5

ASSESSMENT OF A MULTI-CENTRE "POOR MAN'S" ENSEMBLE PREDICTION SYSTEM FOR SHORT-RANGE USE

Kenneth R. Mylne* and Kelvyn B. Robertson
Met Office, Bracknell, United Kingdom.

1. Introduction

Current operational Ensemble Prediction Systems (EPSs), such as those operated by NCEP and ECMWF, are mostly designed for medium-range use, typically 3-10 days ahead. Forecasters are also interested in ensemble information for shorter range, particularly to help identify risks of severe weather development. Following the major cyclonic storms which struck W. Europe in December 1999, which were mostly poorly forecast by deterministic NWP models, the Met Office started a project to investigate the skill of a Poor Man's Ensemble Prediction System (PEPS) as a potentially cheap and effective tool for short-range probability forecasting over 24 - 72 hours. A PEPS is formed by combining the operational forecasts from a number of different NWP centres' systems and treating the combination as an EPS.

There are several reasons to think that a PEPS may provide an effective, as well as cheap, approach to short-range ensemble prediction. Experiments with short-range EPS in the USA (e.g. Hou *et al*, 2001; Stensrud *et al*, 1999; Wandishin *et al*, 2001) have clearly shown that it is important to sample uncertainties due to errors in both the initial conditions and the model evolution to provide effective ensemble dispersion in the short-range. One effective way to sample model error uncertainties is the use of a multi-model ensemble approach (eg Evans *et al* 2000; Mylne *et al*, 2001.) The PEPS is inherently a multi-model ensemble, both in the forecasts but also in the data-assimilation systems where the different models are used to produce background fields. In order to estimate probabilities in the forecast, the analysis errors should ideally be sampled randomly. For medium-range ensembles, however, it is necessary to maximise the dispersion between members over the early part of the forecast by using initial condition perturbations selected to sample the dynamically growing modes which are most likely to lead to medium-range forecast errors. To achieve this ECMWF uses so-called singular vector perturbations (eg Molteni e*t al*, 1996) while NCEP uses error breeding (Toth and Kalnay, 1993). While essential to sample the medium-range uncertainty effectively, these approaches give a highly selective, non-random sampling of the analysis errors. It is believed that medium-range ensembles can still provide useful probabilistic information in the medium-range due to the effects of non-linearity which effectively randomise the sample after the first 1-3 days of the forecast. In the short range this is exactly the period required, so it is more doubtful whether reliable estimates of probability can be expected, especially from the singular vector approach which maximises linear growth over the first 48 hours. By contrast the PEPS provides essentially a random sampling of the initial condition errors, since each model is started from an independent analysis generated from different sub-sets of the observational data, processed by independent assimilation systems. Thus, by randomly sampling both the initial condition and model evolution errors, there is some reason to believe that the PEPS may give better estimates of probabilities in the short-range than can be obtained from the medium-range EPSs.

Other indications that a Poor Man's approach may give good results have been given by Ziehmann (2000) and Ebert (2001). However these studies were based on small ensembles and a limited number of output fields. The Met Office project aims to use a much larger ensemble, and in the longer term to look at a wider range of meteorological fields including some surface parameters. This paper reports on the first stage of the project, using a pilot system based on data available on the MARS archive at ECMWF - this is essentially the forecasts which are exchanged freely under WMO agreements. Probabilistic skill of the PEPS is compared to the ECMWF EPS which, although designed for the medium-range, provides a useful reference. Work is also now in progress to collect higher resolution data from NWP centres around the world for the second stage of the project.

2. Data and Analysis Methods

Results in this paper are based on data taken daily from the MARS archive at ECMWF. In addition to ECMWF forecasts, fields are available from several other NWP centres which distribute them freely on the Global Telecommunications System (GTS). Many of these fields are only distributed at low resolution (5x5° lat/long), and only selected parameters are available. On this basis it was decided to conduct studies on as

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

many centres' data as possible on the common grid of 5x5 degrees and using the parameters H500 (500hPa geopotential height) and PMSL (mean sea-level pressure). Data from the following NWP centres were taken: ECMWF, Met Office (UK), Meteo-France, DWD (Germany), NCEP (USA), JMA (Japan) and BoM (Australia). In the case of ECMWF, both the High-resolution deterministic model and EPS Control runs were used; in addition to the operational Met Office forecasts, two additional runs using a lower resolution version of the model and started from Met Office and ECMWF analyses respectively were also used. Finally six perturbed members of the ECMWF EPS were included to assess what benefits are available by incorporating some singular vector perturbations.

Different configurations of the PEPS were formed by 15 different combinations of these models, to assess the benefits of incorporating, for example, some singular vector perturbations (ECMWF EPS members) or versions of the same model at different resolutions or run from different analyses. In forming PEPS combinations it was felt important to try and set up systems to mimic an operational environment, and only use data that would be available in those circumstances, as in practice this would be one of the limiting factors in the skill of a Poor Man's ensemble. Different models were available from different analysis times (00 or 12 UTC) so it was important to combine outputs with different lead-times for the verification time of interest (e.g. T+36 from 00 UTC from some models with T+24 from 12 UTC from other models). Similarly, ECMWF forecasts, for example, run only once a day (12UTC) and use a late data cut-off time so only become available about 12 hours later. To create a PEPS which would be realistic for short-range operational use it was therefore necessary to use ECMWF forecasts based on data up to 24h older than some of the other models, since that would be the most up-to-date forecasts available operationally.

The 15 configurations used were:

A. Full PEPS consisting of 9 forecasts from different models and their analyses, including ECMWF at two resolutions, and Met Office model at two resolutions plus one run started from the ECMWF analysis.
B. 9 configurations formed by removing one of the models from A above.
C. 5-member ensemble formed from Met Office and ECMWF model runs only.
D. 5-member ensemble formed from the operational runs of 5 centres excluding ECMWF.
E. 15 members incorporating A above plus 6 perturbed forecasts from the EPS.
F. 11 members incorporating C above plus 6 perturbed forecasts from the EPS.
G. An 8-member subset of the EPS consisting of 4 pairs of perturbed members.

As the Poor Man's ensemble has the smallest advantage over the EPS at about 06UTC when any contribution from ECMWF would include forecasts based on analyses only 12h previously (as opposed to 24h), it was decided to assess the skill of the combination of models available at this time of day. (For most models this is a data time of 00UTC.). To limit the amount of data only T+24, T+48, T+72, T+96, T+120 and T+144 were assessed.

In addition to the different PEPS configurations, some preliminary experiments were conducted to test the benefit of including time lagged members (forecasts based on analyses with earlier data times). It was found to be beneficial to have the 12 hour ECMWF lagged forecasts included, but further additions of lagged forecasts from other centres (and ECMWF with more than 12 hours lagging) mostly tended to degrade the ensemble. The first signs of benefit came at T+96 where one of the four assessed regions indicated a small degree of improvement by including 24 hour lagged forecasts. It was only at T+120 that inclusion of time-lagged members became more clearly useful, with lagging of 24 hours slightly more beneficial than lagging of 36 hours. Even at these later times the improvement in Brier Skill (defined below) was small, and since the primary interest was in the use of PEPS for the short-range (up to T+72), this report will only present results using ensembles with a nominal data time of 00 UTC and some models lagged by 12 hours.

Data were collected over a period of 126 days from 7th February to 12th June 2001. Over this period there were only a few occasions when some or all of the 15 ensembles didn't include their full quota of members. For those occasions, the software was designed to calculate probabilities from however many members were available - another capability which would be essential in any operational PEPS, which would depend on the supply of data from numerous centres around the world.

PEPS forecasts were verified against the ECMWF operational analysis and results compared with the ECMWF EPS. Most models verify better against their own analyses than those from other models, particularly for short-range forecasts, due to the role of the model in the assimilation cycle which generates the analyses. Thus the use of the ECMWF analysis gives any advantage in comparative verification to the ECMWF EPS, avoids any risk that apparently better performance by the PEPS might be due to the choice of verifying analysis.

3. Results

3.1 Brier Skill Score

The Brier Score (BS) (e.g. Wilks, 1995) is the mean-squared error of probability forecasts, defined as:

$$BS = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2$$

where:
n = total number of observations
y_k = forecast probability of an event occurring (i.e. proportion of ensemble members that are predicting the event)
o_k = 0 or 1; 0 if the event did not occur, 1 if the event did occur.

BS varies between 0 and 1, with lower values indicating better probability forecasts. To compare the skill of the PEPS with the EPS, the Brier Skill Score (SS) is defined as:

$$SS = \frac{(BS_e - BS_p)}{BS_e}$$

where BS_e is the Brier score for the EPS, and BS_p is the Brier score for the PEPS. A *SS* above 0 indicates the system being assessed (PEPS) has positive skill relative to the reference forecast (EPS); perfect deterministic forecasts would score *SS=1*.

In figure 1 Brier skill score is plotted for probabilities of PMSL being below a range of event thresholds from 970 to 1030 hPa over the Northern Hemisphere at T+24, T+48 and T+72h. Each graph has a number of different lines which correspond to the different PEPS configurations. Here it is unfortunately necessary to present the graphs in monochrome, which prevents the individual configurations being distinguished. However, it is clear that most configurations have considerable skill compared to the EPS across all the probability thresholds and all three lead-times shown. The variation in skill between most configurations is much less than the improvement in skill relative to the EPS. Only one configuration is consistently poorer than the EPS, which is (predictably) the 9-member subset of the EPS (G in the list above). The other versions which do less well than the majority are the 5-member ensembles (C and D) and the one excluding other centres but including 6 EPS members (F). The better performance comes consistently from those combinations which include contributions from as wide a range of different models and analyses as available (A, B and E). Inclusion of the 6 EPS members (configuration E) does not add significantly to the performance of the PEPS, although this configuration was found to perform best over longer range forecasts of 96, 120 and 144h (not shown). Over these longer time-scales the PEPS performed similarly to EPS, and was poorer by T+144 except for the configuration E which remained marginally (but probably not significantly) better. This indicates that the singular vector perturbations are, as predicted in the introduction, effective for the medium range but

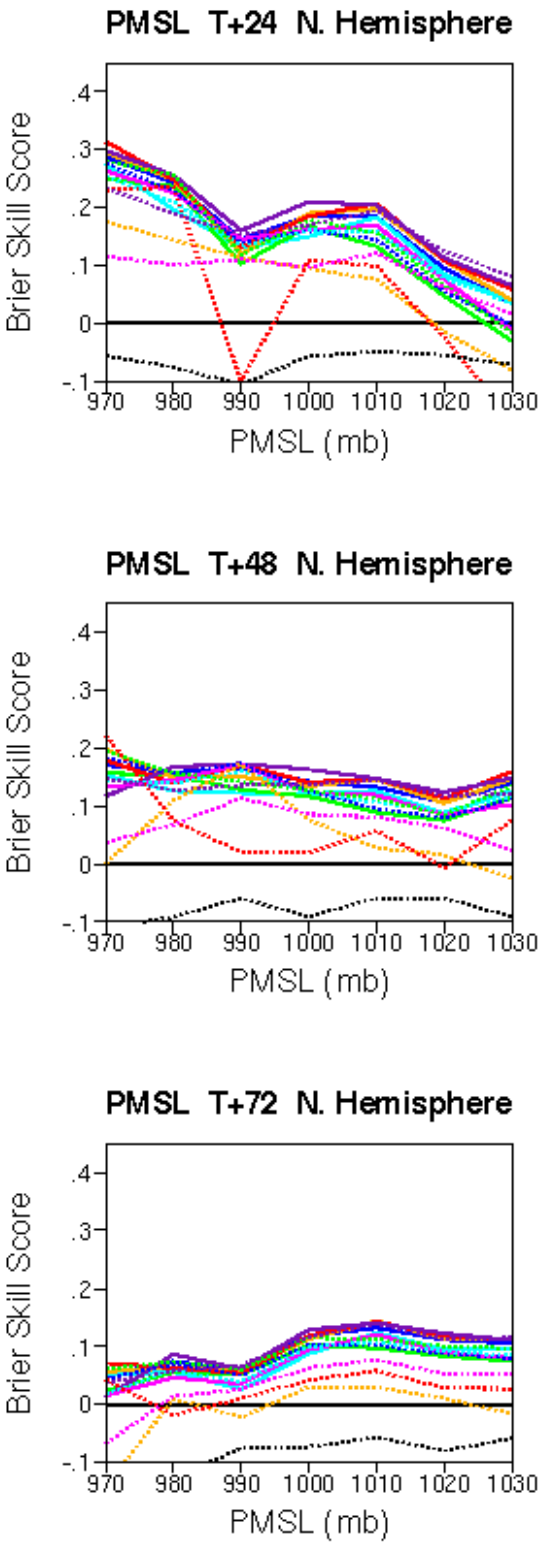


Figure 1: Brier Skill Scores relative to the ECMWF EPS for forecasts of PMSL below various thresholds at T+24 (top), 48 (middle) and 72h (bottom) over the Northern Hemisphere.

much less so for short-range prediction where the more random sampling of the PEPS results in better probabilities. Similar results were found over smaller regions covering Europe and N.America. Results were much poorer when verified over the whole globe, suggesting that the PEPS performed less well than the EPS in the southern hemisphere. The difference may be because some of the models perform less well in the southern hemisphere, where analyses are often poorer due to the large data-sparse areas of the oceans. It may also be due to the fact that we are verifying against the ECMWF analysis - as mentioned earlier, this may give an advantage to the ECMWF models, and this would be expected to be particularly true in such data-sparse areas where the ECMWF analysis is heavily weighted towards the ECMWF short-range forecasts.

A notable feature of the PEPS performance over the global area is that it is better than the EPS for more extreme events (such as PMSL<1020hPa or PMSL<970hPa), but is poor for the more common events such as PMSL <1000hPa. It is also notable in figure 1 that the greatest benefit at T+24 comes for the most extreme low pressure, 979hPa.

3.2 Reliability Diagrams

A reliability diagram is a plot of the observed relative frequency against the forecast probability. A diagonal straight line angled at 45 degrees represents perfect reliability, which means that if the event was forecast as 80%, then the event would happen on 80% of such occasions.

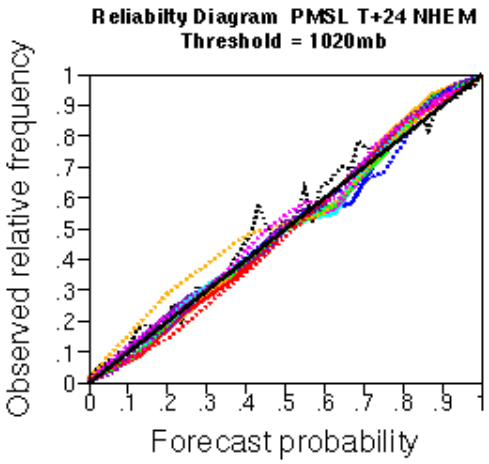
Reliability diagrams are shown in figure 2 for events of PMSL less than (a) 1020mb and (b) 980mb over the northern hemisphere. Again the different combinations of PEPS are shown by different lines, although they cannot be distinguished in monochrome. The EPS is clearly distinguished as the rather jagged dotted line. In order to generate a reliability diagram, forecasts and observations are binned according to the forecast probability. Because we are here comparing ensembles of very different sizes, use of a standard set of bins would have treated different configurations differently, and could have given misleading results for the smaller ensembles in particular. To avoid this, each ensemble was binned according to the full range of possible forecast probabilities available from the members of that ensemble. For this reason the EPS with 51 members (52 possible probabilities) gives a more jagged reliability diagram than the PEPS configurations. This is particularly noticeable with the rarer event PMSL<980hPa in figure 2(b), where some of the high probability bins are sparsely populated.

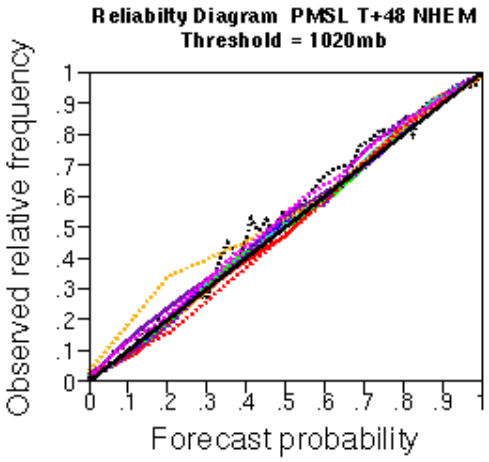
The main feature of figure 2(a) is that all the forecasts show a remarkably good reliability, with both EPS and all PEPS configurations lying close to the ideal diagonal. For the rarer event in figure 2(b) the reliability is not quite so good, with all the ensembles showing a small degree of over-confidence, indicated by the lines having a slightly lower slope than the ideal 45°, especially at the longer range T+72. The reliability of the EPS is a little more difficult to compare due to the noise caused by small sample sizes for higher probabilities of a relatively rare event. Overall there is again little difference between the reliability of the PEPS and EPS. Most configurations of the PEPS appears to be slightly better than EPS in the 30-60% probability range at T+24. Of the PEPS configurations, the 8-member sub-sample of the EPS (G) is similarly over-forecasting in this range, and is the most over-confident version over most of the range at T+48 and 72. While these differences between PEPS and EPS are quite small, they are consistent with the Brier Skill results suggesting that the PEPS gives better overall probability forecasts at the short-range.

3.3 Rank Histograms

Rank Histograms (see Hamill and Colucci, 1997) are used to measure the extent to which an ensemble is able to account for the full uncertainty in the forecast by its ability to encompass the observations. For example, in an ensemble of size n there will be n different forecast values. If these are ranked in ascending order this defines n+1 bins into which the observation could fall (including the 2 bins that are smaller or larger than any ensemble value). Over sufficient cases the rank histogram plots how the observations distribute themselves throughout the n+1 bins. Ideally each observation should be just as likely to fall between ranked members r and r+1 as between any others so that all bins are equally populated. In practice it is common to see the first and last bins being significantly over-populated, indicating that the ensemble has insufficient spread to cover the full uncertainty.

Comparison of rank histograms between ensembles of very different sizes, as done here, requires some care. For example a large ensemble not capturing 16 of the observations will have tails looking far more alarming than if the same thing happened with a smaller ensemble. Both ensembles have failed to capture 16 observations and therefore could be deemed equally useful, although it is likely that the smaller one, with the flatter rank histogram, has given more reliable probabilities. Where the comparison is useful, is to indicate how well the ensemble is performing compared with an ideal ensemble of the same size. If the ensemble performs well measured in this way (has small tails roughly the same size as the other bins), it is likely that an increase in ensemble size will be of significant benefit and considerably reduce the number of missed observations. If the ensemble doesn't perform well (large tails) this suggests that the perturbation strategy is failing to account for some of the important sources of error,





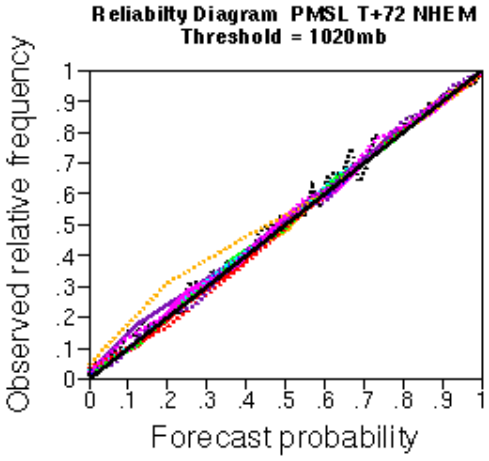
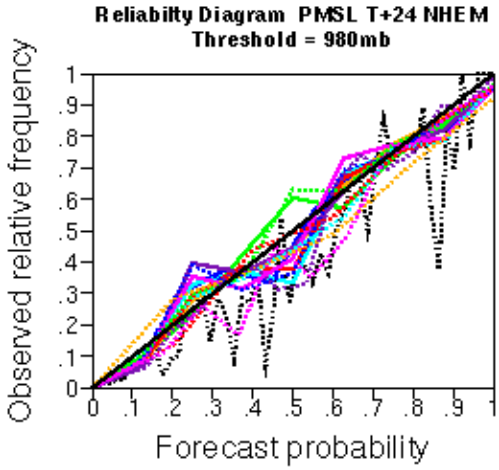
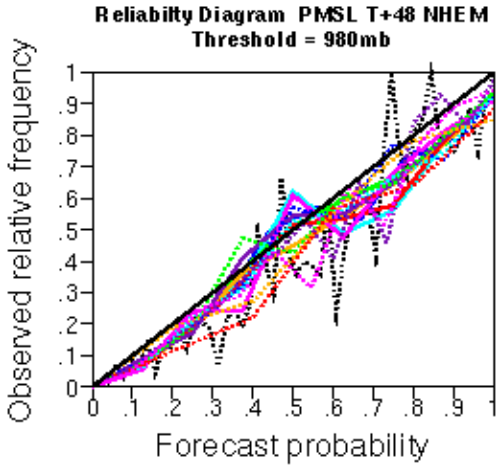


Figure 2a: Reliability diagrams for PEPS configurations for PMSL below 1020hPa at T+24 (top), 48 (middle) and 72h (bottom) over the Northern Hemisphere. The jagged dotted line is the EPS.





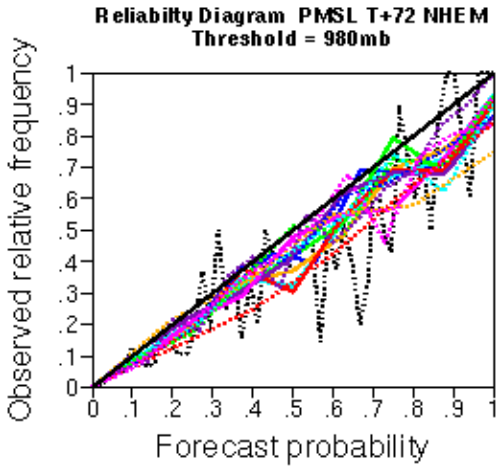


Figure 2b: As figure 2a, but for PMSL<980hPa.

and any advantage in increasing the ensemble size will be more limited.

Figure 3 shows rank histograms for PMSL over the northern hemisphere for the various PEPS configurations and the EPS. To compare ensembles of different sizes on one graph the bin populations have been normalised relative to their ideal populations (shown by the horizontal line) and the "histograms" plotted as lines. The EPS is clearly distinguished (as in the reliability diagrams) by being more jagged due to the large number of bins, and it is also immediately clear that the outlier bins at either end are more severely over-populated than for any of the PEPS configurations. This occurs at all lead-times shown, and indeed remains the case right out to T+144. The other line which is clearly distinguishable as a solid line with large peaks in the outlier bins and also a peak in the middle at a normalised rank of 0.5 is the 8-member subset of the EPS (configuration G). The peak in the middle of this rank histogram is believed to be caused by the fact that the EPS control, around which the pairs of perturbations are added, was not included in this configuration. The effect is strong at T+24 when the perturbations are still approximately distributed around the control, but disappears at later times in the forecast as the members become more randomly distributed.

It is immediately clear that the full 51 member EPS and the 8 member EPS subset both have a less even distribution than most of the PEPS configurations at all forecast times from T+24 out to T+72, and indeed this remains true out to T+144 (not shown). The 8-member EPS subset at first sight appears better than the full 51 member EPS, but it should be remembered that the flatter curve is what one would expect of a smaller ensemble, and gives a more useful comparison with the PEPS configurations.

The different configurations of PEPS are mostly similar to each other. There are two exceptions which, like configuration G, have noticeably over-populated outlier bins, and these are the two which do not include members from centres other than ECMWF and the Met Office, C and F. Like the BSS results, this confirms that the inclusion of a large number of independent analyses is highly beneficial in spanning the range of uncertainty in the short-range. However it should be noted that at T+24 the rank histograms of the other PEPS configurations, which do include the full range of analyses, are slightly curved up in the middle. This indicates that these versions are spanning rather too large a range of uncertainty, although they are still much closer to the ideal than the EPS or 8-member subset.

Another clear feature of most of the PEPS rank histograms at all lead-times is a bias towards under-population of the lower PMSL ranks and over-

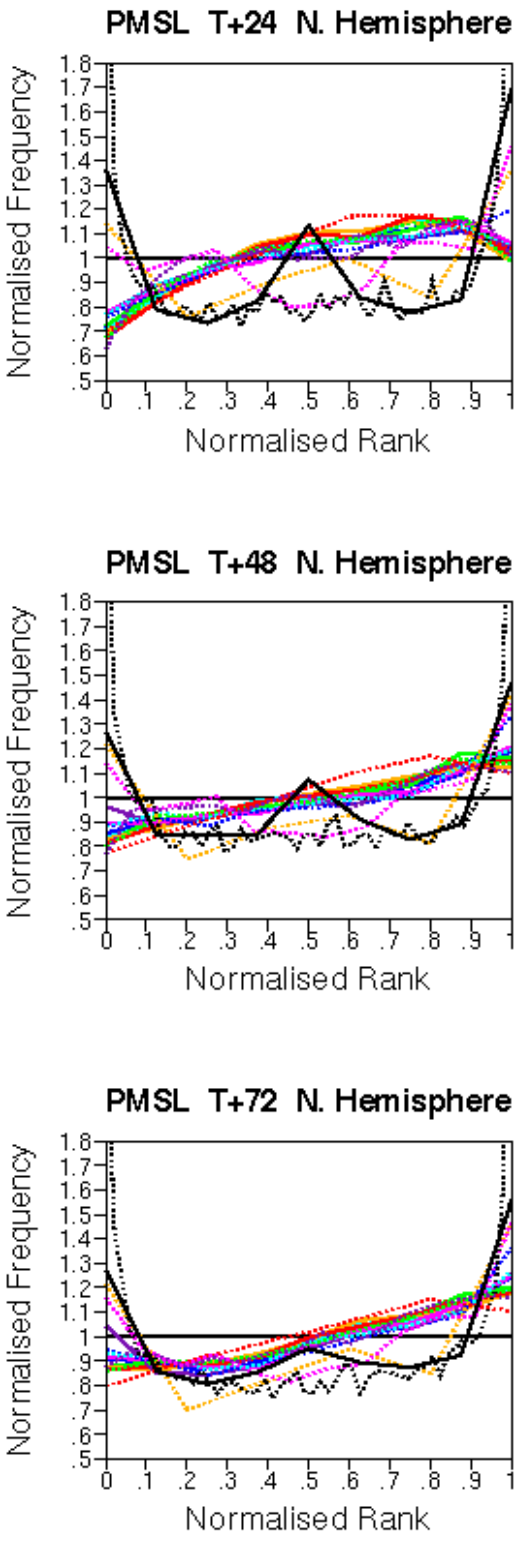


Figure 3: Normalised rank histograms for PMSL over the northern hemisphere for the various PEPS configurations. The EPS is the jagged dotted line.

population of the higher ranks. This suggests that several of the models have a bias towards forecasting low PMSL too frequently.

Apart from this bias and the slight over-spanning of the uncertainty at T+24, the rank histograms of most PEPS configurations at short-range are close to ideal. At longer range up to T+144 (not shown) the PEPS develops the normal U shape indicating that it is not spanning the full uncertainty, but it is noticeable that it remains closer to the ideal than either the EPS or the 8-member subset (G).

Conclusions

Preliminary experiments with the PEPS system using low resolution data have been compared to the ECMWF EPS, and results for short-range prediction in the northern hemisphere are encouraging. It should be noted that the EPS, with which results are compared, is optimised primarily for medium-range prediction, while the PEPS is under consideration as a short-range ensemble system. Results presented here have concentrated on PMSL as it is of more relevance to most forecast problems, but overall results for 500 hPa height are similar.

A number of different configurations of PEPS were compared. For most there was little difference in overall performance. However it was found to be important to include a good range of analyses (and models) from different NWP centres in order to span the uncertainty and give reliable probabilities in the short-range. The singular vector perturbations from ECMWF were less effective in spanning the uncertainty, although resulting probabilities from the EPS were nevertheless reliable. Brier Skill Scores indicated that the PEPS probabilities are considerably better than EPS, which may be due to the more random nature of the sampling of initial condition uncertainties.

One clear disadvantage of the PEPS compared to the EPS is ensemble size, and this is particularly relevant when considering severe weather. The EPS, with 51 members, will always have a greater chance of capturing severe weather developments with at least a low probability, and might therefore be expected to give greater decision-making value to certain users who can take action based on low probabilities. However, this may be partly countered by the evidence from figure 1 that the greatest improvement in probabilistic skill at T+24 actually occurs for the more extreme events.

Results were less good looking at the entire globe, suggesting poor performance in the southern hemisphere. Some of the models used may perform less well in the southern hemisphere, particularly if their analysis systems are not making such good use of satellite data as is now done by ECMWF. However this may also be partly caused by the use of the ECMWF analysis as the verifying truth - where there is little observational data the ECMWF analysis would be expected to agree well with ECMWF forecasts over the short-range. Southern hemisphere performance may need further investigation in the future.

Future Plans

The Met Office is now collecting forecast data directly from a much larger group of global NWP centres to assess the PEPS further. Data are being collected at 1.25° resolution for six fields (PMSL, H500, T850, 2m temp., 10m wind speed and precipitation) to allow a more comprehensive verification including more parameters of relevance to forecasters and forecast users. Contributions to this world-wide collaboration are welcome.

References

Ebert, E.E., 2001: Multi-model ensemble forecasts of heavy rain events in Australia *Symposium on Precipitation Extremes: Prediction, Impacts and Responses,* AMS, Albuquerque, 14-18 January 2001, pp332-335.

Evans, R.E., Harrison, M. S. J., Graham, R. J., Mylne, K. R., 2000: Joint medium-range ensembles from The Met. Office and ECMWF systems *Mon. Wea. Rev*., **128**, 3104-3127.

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

Hou,D., Kalnay,E. and Drogemeier,K. 2001: Objective verification of the SAMEX '98 ensemble forecasts *Mon.Wea.Rev*. **129**, 73-91.

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

Mylne,K.R., Evans,R.E., Clark,R.T., 2001 Multi-model multi-analysis ensembles in Quasi-Operational Medium Range Forecasting To appear in *Quart. J. Roy. Meteorol. Soc.*

Stensrud,D.J., Brooks,H.E., Du,J., Tracton,M.S. and Rogers,E. 1999: Using Ensembles for Short-Range Forecasting, *Mon Wea Rev* **127**, 433- 446.

Toth, Z. and Kalnay, E. (1993) Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Met. Soc.,* **74**, 2317-2330.

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

Wandishin, M.S., Mullen, S.L., Stensrud, D.J. and Brooks, H.E. 2001: Evaluation of a short-range multimodel ensemble system. *Mon Wea Rev*, **129**, 729-747.

Ziehmann,C.,2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus*, **52A**, 280-299.