5.1

OPERATIONAL CALIBRATED PROBABILITY FORECASTS FROM THE ECMWF ENSEMBLE PREDICTION SYSTEM: IMPLEMENTATION AND VERIFICATION.

Kenneth R. Mylne*, Caroline Woolcock, James C. W. Denholm-Price, and Robert J. Darvell.
Met Office, Bracknell, United Kingdom.

1. INTRODUCTION

The ECMWF EPS (Molteni *et al*, 1996) has been used operationally at the UK Met Office in the *Previn*** system since 1999. This paper discusses the recent upgrade to the system to improve site-specific probability forecasts. In §2 a Kalman filter system is described which corrects biases in the 10 day site-specific forecasts of 2m temperature (T_{2m}), 10m wind speed (*WS*) and precipitation (PPT12 and PPT24 – 12 and 24 hour accumulations), which depend on the site and synoptic situation. It also enables the calculation of site-specific minimum and maximum temperatures (T_{min} and T_{max}).

The 51 Kalman Filtered ensemble members (including T_{min} and T_{max}) are then used to generate probabilistic forecasts, and their calibration is described in §3. The operational implementation of the system and its verification are considered in §4 and §5, respectively, before a summary of these developments in §6.

The original Previn system produced probabilistic information for 41 sites in the UK and their location is illustrated in figure 1. Subsets of these sites were used in this study, with 15 sites used to test the Kalman filter and 30 used in the calibration. The upgraded system has since been expanded to include many more UK, European and global sites

2. KALMAN FILTER

Currently (Autumn 2001) the EPS runs at T_L255 resolution (~80km over the UK). The gridded fields are interpolated to specific sites but this leaves a significant site-specific and synoptically-dependent bias in the forecasts. The Kalman filter was implemented as an exponentially-weighted least-squares regression filter and was (unsurprisingly) found to perform better than a simple running-mean bias-correction at these sites for all weather parameters except precipitation.

Figure 1: Map of UK stations.

Corrections are made to every EPS forecast range at 6 hourly intervals, up to 10 days. In principle the Kalman filter could be applied at each forecast time, that is at T+0, 6, 12, …, 240 hours. In practice such a correction incorrectly decreases the ensemble spread. This is illustrated in figure 2 for a single station (WMO code 03026, namely Stornoway in Scotland) where the ensemble spread is plotted relative to the unfiltered spread (at each forecast range). The reduction in spread occurs because statistics based on past forecasts include forecast errors as well as site-specific errors. Correction based on these statistics therefore pulls each ensemble member back towards climatology, to correct for forecast errors, thus reducing the spread.

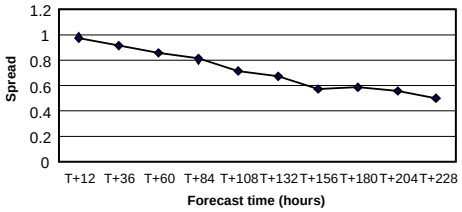

Figure 2: EPS spread after Kalman-filtering at each forecast period, relative to the unfiltered spread.

* *Corresponding author address:* K R Mylne, Met Office, London Road, Bracknell, RG12 2SZ, United Kingdom; e-mail: ken.mylne@metoffice.com

** Standing for 'Predictability Visualisation'.

Instead a regression model is updated using the control EPS member for each station and validity time. That is, corrections for 0Z, 6Z, 12Z and 18Z are derived from forecasts at T+12,T+18,T+24 and T+30. The appropriate model is applied to each forecast time to correct the forecast (*e.g* 12Z model applied to T+240). For T_{min} and T_{max} the validity time depends on the location of the station. In addition it was found that the full Kalman filter regression was inappropriate for precipitation and instead a simple bias-correction is applied to PPT12.

For each specific forecast parameter the optimum statistical model must be determined. For example it is reasonable to expect both wind speed and direction to influence local temperatures (especially near water or orographic features). Tests to determine which statistical model was the best for each forecast parameter (temperature, wind speed and precipitation) were performed by comparing the 'skill' of forecasts produced from various statistical models against the unfiltered EPS. An example is shown in figure 3 where the skill of a simple bias correction and seven different statistical models are compared. The best model was determined for each parameter and these are summarised below in table 1.

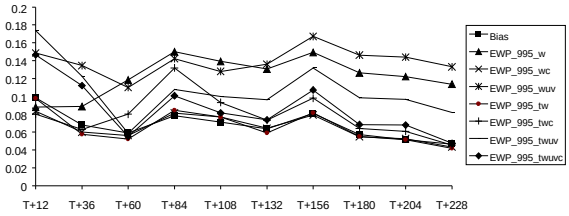

Figure 3: Skill scores for 10m wind speed from various Kalman filter models.

A similar procedure was undertaken for comparing the effect of varying the Kalman filter's 'memory' (*i.e.* the amount of past data used in the statistical models) and the optimal 'memory' was found to be approximately 60 days.

Forecast parameter	Regression model
T_{max}	*T, w, u-v, c*
T_{min}	*T, w, c*
T_{2m}	*T, w, c*
WS_{10m}	*T, w, u-v, c*

Table 1: Kalman filter parameters.

Regression model parameers: *T*=2m temperature, *w*=10m wind speed, *c*=constant (bias), *u-v*=10m wind direction (cosine).

3. CALIBRATION

The Kalman filter corrects site-specific biases effectively but the filtered PDF is still not always *reliable*, in that forecast probabilities do not always verify at the forecast frequency, resulting in deficient reliability diagrams (Wilks, 1995). Verification rank histograms from (site-specific) Previn forecasts often exhibit overpopulated extreme ranks (outliers) which may indicate insufficient spread in the ensemble (or other effects – see Hamill (2001) for a discussion). Calibrating the forecast probabilities overcomes these deficiencies and the chosen method follows Hamill and Colucci (1997) in using the verification rank histogram to predict the correct probabilistic weights used in the PDF.

3.1 Probabilistic weights

Previn PDFs are recalibrated using weights derived from verification rank histograms from the previous 3 months, averaged over a group of stations (initially these were a 30 station subset from figure 1). Verification rank histograms count the number of verifying observations falling within bins delimited by pairs of *ranked* (in numerical order) ensemble members. The EPS has 51 members so there are 50 bins delimited by two members and two *outlier bins* at either end. The outlier bins contains all observations that fall outside the ensemble, with the lower outlier holding the observations less than the lowest ensemble member, and *vice versa* for the upper outlier bin. Ideally, observations would fall in each bin with equal probability and therefore 1/52 of the observations should fall into each bin (*including* the outliers). In reality the ensemble is not ideal and departures from the ideal uniform shape are found (*e.g.* see Hamill, 2001).

The ensemble members come from integrations of the same numerical model, starting from analyses that are perturbed with combinations of singular vectors (Molteni and Palmer, 1993). The singular vectors are optimised to give maximal growth over the first 48 hours of forecast time. The ensemble spread in the early stages of the EPS forecast is low for all parameters, giving overpopulated rank histogram outliers. In addition the outliers may be overpopulated at any forecast range due to the difficulty of downscaling the T_L255-resolution forecast to specific sites.

Examples of the weights are shown below in figures 4 and 5 for temperature at T+54 and precipitation at T+156, respectively (note both have logarithmic vertical scales and the horizontal line denotes the 'ideal' weight of 1/52 for a 51-member ensemble). Temperature is under-spreading with many EPS forecasts being too cold (LHS bin) or too warm (RHS bin) which the weights correct by broadening the calibrated PDF, emphasising the tails.

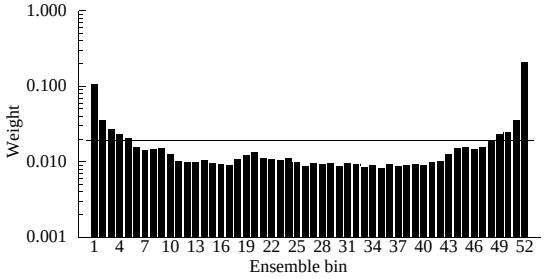

Figure 4: Weights for T_{2m} at T+54h.

Figure 5 illustrates the model's tendency to frequently forecast small amounts of rain over a grid box when no rain is observed at a specific location within the grid box, which causes the LHS bin to be extremely large. The weights in this case enhance the probability assigned to the ensemble member with the smallest amount of rain (often zero), shifting the PDF towards zero. This can have an adverse effect on forecasts of large amounts of rain. The difficulty of forecasting large amounts of rain at specific sites is a known problem with NWP models in general. More specific calibration techniques could be used to improve this situation.

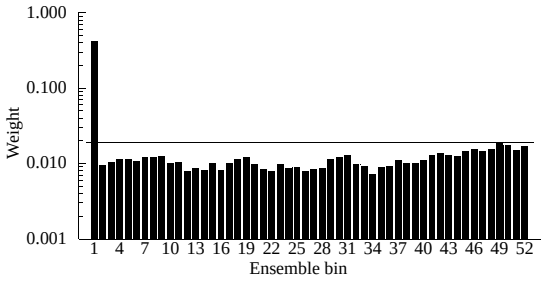

Figure 5: Weights for PPT12 at T+156h.

3.2 Fitted outlier tails

An overpopulated outlier bin causes a large amount of probability to be assigned to that bin in the forecast PDF, which is needed to make the PDF reliable. However there is no information from the EPS about the shape of the PDF in this bin. To provide such information, the Previn forecast PDFs are extended using Weibull distributions. These are fitted to the distribution of the *differences* between verifying observations and the relevant extreme EPS member. Data for these are taken from 6 months – 3 months prior to the current data and 3 months from the previous year starting from the current month (*e.g.* forecasts for August 2001 use data from May/June/July 2001 and August/July/June1999*). The Weibull fit to these data is improved by the Kalman filter which eliminates many extreme outliers that are strongly influenced by local effects.

Examples of the outlier distributions are shown in figures 6 and 7. In the former the Weibull fit is poor as the outlier distribution is clearly bimodal. This was found to be influenced strongly by coastal sites.

In figure 7 the fit is improved by Kalman filtering the data – indeed the spread of observations outlying the EPS is much reduced (the maximum difference decreases from more than 12°C in figure 6 to less than 4°C in figure 7).

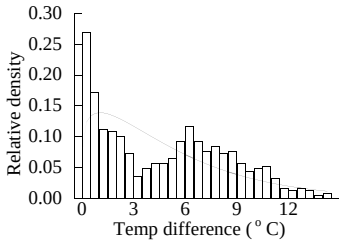

Figure 6: Upper outlier distributions for T_{2m}: T+240h., Winter 1998/9, bias correction only.

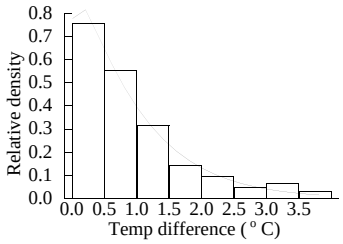

Figure 7: Upper outlier distributions for T_{2m}: T+240h., Winter 2001, Kalman filtered EPS.

Similar distributions are found for wind speed (benefiting similarly from the Kalman filter) but precipitation (PPT) exhibits different behaviour:

The lower PPT outlier bin contains many observations where no rain was observed at the station yet small amounts were forecast over the grid box. This results in a distribution of outliers that is dominated by a single bin at the origin (such as shown in figure 5) and is poorly modelled by the Weibull distribution. However since PPT is non-negative the dominant lower outlier bin is equivalent to a large weight applied to the lowest EPS member (usually at or near the origin) and thus a fitted tail is unnecessary.

---

* Year 2000 data were not archived for this purpose, thus avoiding a fault which adversely affected the EPS performance at that time.

At the upper end, the PPT outlier distribution exhibits a long tail where the observed PPT was larger than forecast. This kind of error is a well known problem when downscaling NWP PPT. The Weibull distribution does fit the majority of the upper outlier and it is hoped that it has some benefit whilst recognising that it is unlikely to compensate for the model's inability to correctly predict localised rainfall.

4. OPERATIONAL IMPLEMENTATION

The derivation of calibration statistics, either the probabilistic weights or fitted tails, requires large amounts of verification data. It is not possible to pool data from 0Z, 6Z, 12Z and 18Z verifying times at all forecast ranges since the weights and tails evolve significantly both with time of day and forecast range. Instead data are pooled from groups of stations. The Kalman filter reduces (but does not eliminate) the differing effects of orography between the sites and facilitates grouping by station. This enables the calibration to be performed using relatively recent data each month.

Initially calibration used 30 UK stations in 2001 (a significant number of the 41 stations in figure 1 have closed since the inception of Previn). Averaging data from these stations over 3 months is sufficient to produce statistically stable weights under most circumstances. Thus the weights used for the current month are based on the verification rank histogram from the past three months.

Ideally a larger sample is required in order to generate representative tails (*e.g.* figures 6 and 7 contain 500 and 127 points, respectively). Since the tails are observed to vary with the season a short averaging time is also needed. To accomplish this, data are pooled from the last 3 months in the current year together with 3 months from the previous year. In 2001 this was complicated slightly by the lack of archived data from January to November 2000 but will not affect operations from February 2002. For example, the fitted outlier tails from August 2001 use data from May, June and July 2001 together with August, July and June 1999. This forms a compromise between the need to avoid sampling 'between seasons' and the rate of upgrades to the EPS – ideally back statistics would be made available at each EPS upgrade, but this is quite infeasible given ECMWF's current resources.

In order to monitor the Weibull tails and provide information about their success in modelling the outlier distributions a resampling procedure is also adopted. The procedure followed is as follows: The pooled outlier data are made approximately independent by imposing a 24 hour separation in the data at each station (it being assumed that the outlier differences between stations are somewhat independent). A global Weibull fit is calculated from these data and then 1000 bootstrapped resamples (with replacement) are taken, usually of between 100 and 200 points from a global set of more than 500. The fit to each subsample is assumed to be a point from the 'climatology' of possible distributions and so the location of the global fit within the resampled distribution measures how representative the global fit is of the 'climate'. To be useful the global sample must be large so that the subsamples are reasonable realisations of the fitted tails – resampling from small amounts of data is meaningless, from a climatological perspective.

5. VERIFICATION

A probabilistic approach is used to verify the improvements to forecasts from Previn. Brier scores and reliability diagrams for specific forecasts are illustrated below.

The Brier score is analogous to a probabilistic mean-squared error – it is the mean-squared difference, over a sufficiently long average, between the forecast probability and the observed frequency. For forecast probabilities y_k and observations o_k (o_k=1 if the event occurs, 0 otherwise) the Brier score is:

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

BS=0 for a perfect forecast and is otherwise positive (small values are better). Brier scores are calculated for a specific probability forecast and hence a specific weather parameter threshold. Thus there are many cases that can be studied (and are generated operationally within the Previn system) so for brevity an illustrative example is included here.

In figure 8 the Brier scores for probability forecasts of midnight 2m temperatures greater than 10^{o}C are shown for all forecast ranges from summer (JJA) 2001. The upper line (with the largest score so the *least* skilful forecast) is from the uncorrected EPS forecast and the group of lines beneath are, in order of decreasing Brier score (increasing skill), the Kalman-filtered EPS, the calibrated EPS using the weights alone and finally the calibrated EPS with weights and fitted Weibull tails.

Thus at all forecast ranges this probability forecast is improved most by the Kalman filter and by a consistent small amount by the calibration. Typically the Kalman filter and calibration systems have a beneficial impact on Brier scores for temperature, wind speed and rainfall although the relative impact of each varies – the calibration usually has a stronger impact at early forecast times (where the weights are largest).

For extreme precipitation events (such as 20mm PPT falling in 12 hours) the calibration is actually

detrimental to the forecast. Less extreme events dominate the calibration statistics and in the case of precipitation the more extreme events are very poorly sampled. A different kind of calibration may prove to be more beneficial in such cases, with more explicit downscaling from the EPS grid box to the local sites.

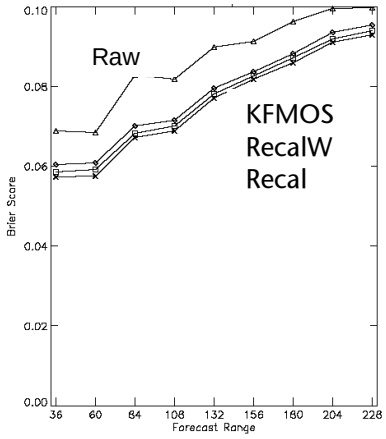

Figure 8: Brier score for $P(T_{2m} > 10°C)$ verifying at midnight during summer (JJA) 2001.

Symbols: Triangle=Raw EPS; Diamond=Kalman filter; Square=Calibration.

Calibration is intended to improve the probabilistic reliability of forecasts and this can be verified by comparing the forecast probability and frequency of occurrence of events. An example of such a 'reliability diagram' is shown in figure 9. The straight diagonal line indicates the ideal situation where forecast probabilities verify with the same frequency. The lower line corresponds to the uncorrected EPS and shows significant over-forecasting, predicting higher probabilities than verifying frequencies. The Kalman filter (diamonds) greatly reduces the over-forecasting by correcting the forecast bias, but the probabilities are still over-confident (indicated by a slope of less than 45^o on the diagram).

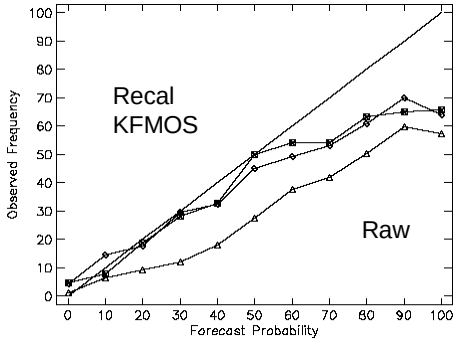

Figure 9 Reliability of $P(\mathrm{PPT}_{12} > 0.1\mathrm{mm})$ from forecasts at T+48 during summer 2001.

Symbols: Triangle=Raw EPS; Diamond=Kalman filter; Square=Calibration.

Over-confidence is typical behaviour where the ensemble spread is insufficient to cover the full uncertainty. The calibration (squares) reduces the over-confidence for forecast probabilities between 0% and 60%, but is unable to correct for higher probabilities.

Similarly, figure 10 shows a reliability diagram for the probability of winds exceeding 22 knots (Beaufort force 6, 11.3 ms^{-1}). In this case the raw EPS over-forecasts the wind probabilities. The Kalman filter corrects the bias quite effectively but again the forecasts remain over-confident. The calibration successfully corrects this to give fairly reliable probabilities throughout. The reliability diagram in figure 10 is somewhat 'noisy' making this assessment imprecise but the trend seems clear. The noise arises because of the small number of events in the higher-probability bins. This is illustrated by the so-called 'sharpness' diagram in figure 11. Evidently the highest probability bins contain less than 10 verifying events after calibration (note that the vertical scale is logarithmic).

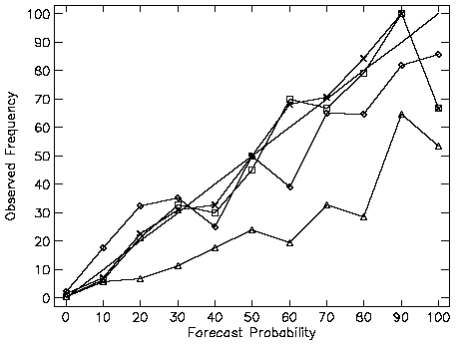

Figure 10: Reliability of $P(\mathrm{WS}_{10m} > 22\mathrm{kn})$ from forecasts at T+72 during winter 2001.

Symbols: Triangle=Raw EPS; Diamond=Kalman filter; Square=Calibration, Cross=Calibration with tails.

The verification data are taken from a single season (winter 2001 in the case of figures 10 and 11) and from 30 UK sites. It is likely that the dataset contains some poor forecasts which arise from synoptic systems affecting more than one site. Such a situation is responsible for the very unreliable final point on the calibrated reliability curve (the 100% point in figure 10). The ten or-so forecast/observation pairs contributing to this point all come from different stations on the same date. Evidently more data are required to remove the noise, which will become available from winter 2002.

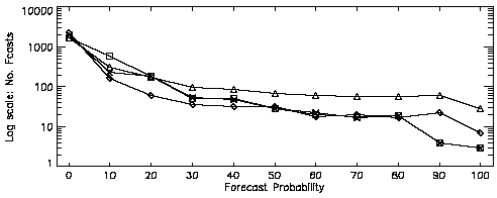

Figure 11: 'Sharpness' diagram showing the number of events in each bin from figure 10.

A further test of the capability of the calibrated forecasts is to examine forecasts of 95%-confidence temperatures. These are forecasts of the temperature range that the verification is expected to lie within 95% of the time and can easily be made from Previn. The verification should fall outside the forecast range 5% of the time. Verification of this is shown in figure 12 where the three lines depict the frequency of observations found outside the 95% range during spring (MAM) 2001. Clearly the calibration is beneficial at all forecast ranges (especially with the fitted Weibull tails), improving the reliability of the 95% limits considerably.

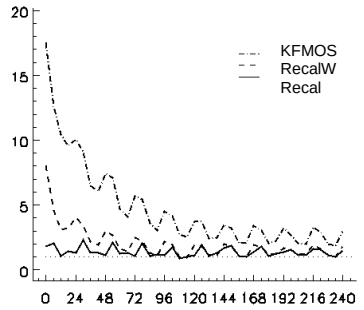

Figure 12: Frequency of observations outside 95% temperature intervals during spring 2001.

6. SUMMARY

Enhancements to the Met Office 'Previn' system for producing site-specific probability forecasts from the ECMWF EPS have been described. Fields from the EPS are interpolated for specific sites and corrections applied. The Kalman filter successfully corrects site-specific biases. Statistical calibration and augmentation of the resultant PDFs improves the reliability of probabilistic forecasts. Brier scores show that the full Kalman filter has a large beneficial effect on wind speed and temperature forecasts. Additional benefit comes from the calibration and is largest at early forecast times where the EPS spread is smallest. For precipitation a simple bias correction replaces the Kalman filter (which is less effective in terms of Brier scores) and the calibration is most beneficial for small precipitation amounts (less than 10mm in 12 hours). An alternative strategy is being considered for improving the probabilistic forecasts of more extreme events, such as larger precipitation thresholds. The calibration is not tuned for these rare events, with the statistics being dominated by more common events.

Reliability of the probability forecasts is usually improved by the Kalman filter and calibration, especially when considering 95% confidence temperature ranges. However there are applications (such as the aforementioned forecasts of large amounts of precipitation) for which the calibration is not suited. To enable users to identify these and make best use of the system, verification of probability forecasts with a large number of thresholds is performed routinely.

The improved Previn system is currently running quasi-operationally (Autumn 2001), producing corrected probability forecasts for many sites. It allows for the production of useful products for many potential Met Office customers, including the offshore oil and gas industry.

7. REFERENCES

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

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

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

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.

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