Nonlinear Processes in Geophysics (2003) 10: 469–475

Nonlinear Processes in Geophysics

Can an ensemble give anything more than Gaussian probabilities?

J. C. W. Denholm-Price

The Met Office, UK

School of Mathematics, Kingston University, Kingston upon Thames, KT1 2EE, UK

Received: 30 August 2002 – Revised: 24 February 2003 – Accepted: 24 March 2003

Abstract. Can a relatively small numerical weather prediction ensemble produce any more forecast information than can be reproduced by a Gaussian probability density function (PDF)? This question is examined using site-specific probability forecasts from the UK Met Office. These forecasts are based on the 51-member Ensemble Prediction System of the European Centre for Medium-range Weather Forecasts. Verification using Brier skill scores suggests that there can be statistically-significant skill in the ensemble forecast PDF compared with a Gaussian fit to the ensemble. The most significant increases in skill were achieved from bias-corrected, calibrated forecasts and for probability forecasts of thresholds that are located well inside the climatological limits at the examined sites. Forecast probabilities for more climatologically-extreme thresholds, where the verification more often lies within the tails or outside of the PDF, showed little difference in skill between the forecast PDF and the Gaussian forecast.

1 Introduction

Ensemble weather forecasting techniques are used to account for uncertainty in initial conditions (Molteni et al., 1996; Toth and Kalnay, 1993; Houtekamer et al., 1996) and increasingly also for model errors in numerical weather prediction (Buizza et al., 1999). The small sample of the forecast probability density function (PDF) made by present-day ensembles, such as the Ensemble Prediction System (EPS) (Molteni et al., 1996) from the European Centre for Medium-range Weather Forecasts (ECMWF), which currently consists of 50 perturbed members plus an unperturbed control, means that the ensemble forecast PDF cannot resolve the full complexity of the atmospheric PDF (Palmer, 2000).

Probabilistic forecasts at specific sites attempt to resolve a PDF with much smaller dimension than a global model as it is conditioned by external influences, such as synoptic conditions, and local influences, such as orography and land surface features. For a number of years the Met Office has used the ECMWF EPS to obtain site-specific probability forecasts from a system called 'Previn' (Legg et al., 2002). The data analysed in this paper were taken from the EPS system that was operational over the 2001/2002 winter. Its spectral resolution of $T_L 255$ gives a horizontal resolution of $\approx$80 km over the UK.

Improvements to Previn made around this time include a site-specific Kalman filter to provide downscaling from the large-scale EPS fields to local weather parameters, and statistical post-processing to improve the probabilistic reliability. These are discussed briefly below, see Denholm-Price and Mylne (2002) for full details.

Ensemble-generated PDFs are frequently non-Gaussian, whether they are site-specific or forecasts on a model grid. Precipitation and wind speed PDFs are often highly skewed. Temperature PDFs are often nearly Gaussian but on occasion may not be, for example when the PDF has multiple modes. It may be that such PDFs are due to meteorological effects and represent real forecast information. For example, a bimodal temperature PDF could be caused by distinct clusters of solutions, placing a site under different wind directions. However, there is little evidence in the literature to support the idea that this represents useful (e.g. as measured by forecast value, Richardson (2000)) or even probabilistically reliable forecast information (e.g. as measured by a rank histogram, (Hamill and Colucci, 1997)). It has been suggested that an EPS gives no more information than a skilful mean forecast and an idea of the unpredictability in terms of the ensemble spread (Atger, 1999), which is essentially Gaussian 'information'.

In this paper, verification results from Previn forecasts of 2 m temperature (T) and 10 m wind speeds (WS) are used to demonstrate the existence of information beyond just the mean and spread within site-specific PDFs. To do this, PDFs from various configurations of the Previn system are compared with reduced-information Gaussian PDFs. If there

Correspondence to: J. C. W. Denholm-Price
(j.denholm-price@kingston.ac.uk)

were no more than Gaussian information in the Previn PDFs then there would be no difference in probabilistic skill between the full PDF and an appropriately-determined Gaussian. Hamill and Colucci (1998) performed similar tests on precipitation PDFs using a smaller short range ensemble, by fitting a gamma distribution to the ensemble PDF after calibration and bias-correction. Their results were essentially neutral, indicating no significant difference in skill between forecasts from the calibrated ensemble and fitted gamma distributions. Here a Gaussian fit is used for both temperature and wind speed forecasts. Even though it is unlikely to model accurately a skewed wind speed PDF many Previn wind speed PDFs are not skewed. The utility of a Gaussian fit for this purpose is discussed further in Sect. 3.2.

Throughout this paper the skill of the ensemble PDF with respect to the Gaussian fit is measured using the Brier (skill) score. The outline is as follows: In Sect. 2 the experimental design is discussed, the results are presented in Sect. 3 and some conclusions are drawn from the results in Sect. 4, together with some discussion of future work.

2 Experimental setup

Previn forecasts are derived from EPS data with four incremental stages of post-processing. After each stage a potentially different ensemble is generated from which probability forecasts are made. First the EPS is interpolated (bilinearly) from the model grid to the locations of 30 sites distributed over the UK. In this paper, forecasts derived from these data are denoted by 'RAW'. The following notation is used to identify subsequent stages of processing throughout the paper:

1. KFMOS: Interpolated EPS fields are downscaled and bias-corrected using a Kalman filter which implements exponentially-weighted, multivariate regression from a 60 day training set. Regression models were determined experimentally during 2000 to give the best reduction in bias over a subset of the UK stations. The system uses a 'perfect-prog' MOS-style approach, utilising data from the first forecasts that are available at a given time of day to correct all forecasts at that local time throughout the 10 day forecast. Corrections for 00, 06, 12 and 18 h local time are therefore calculated using forecast data from T+12, T+18, T+0 and T+6 h, respectively (the EPS data time is midday UTC). It is assumed that errors at these early forecast times are dominated by representativity error rather than model error (as suggested by Orrell et al. (2001)).

2. CAL: Probabilistic reliability of the KFMOS forecasts is corrected by rank histogram calibration (Denholm-Price and Mylne, 2002; Hamill and Colucci, 1997). The rank histograms define probabilistic weights that are used when calculating probabilities from the ensemble. They are formed by measuring the relative frequency with which observations from three months of past verification data fall into the 52 bins whose edges are defined by the 51 ranked ensemble members.

3. CAL+W: PDF tails from CAL forecasts are augmented by fitted Weibull distributions (Bauer, 1996). This is done to model more accurately the edges of the distribution where the calibrated weights are often large. The smallest and largest members of the ensemble define the right or left edge of the bin that is called the lower or upper outlier bin, respectively. A Weibull distribution is fitted to the verifying observations which are found in these bins using six months of verification data. The fitted distribution replaces the extrapolation over a somewhat arbitrary distance of one unit that is used at the edges of the CAL forecast PDFs. The fitted Weibull tends to broaden the PDF and gives a more realistic tail to the distribution as it is derived from the verification data.

The probabilistic weights are updated at the beginning of every month using the last three months' data taken from 30 stations across the UK. The fitted Weibull distributions are updated similarly but they also use three months of data from the same season of the previous year. These periods were chosen to balance the need for adequate amounts of verification statistics against the likelihood of future model changes. See Denholm-Price and Mylne (2002) for details.

After each stage of the post-processing, the ensemble estimate of the PDF is available. This gives three different forecast systems to test (KFMOS, CAL and CAL+W), where each stage of post-processing potentially adds more detail to the PDF. The Brier skill score BSS (Wilks, 1995) is used to compare the probabilistic skill of the various Previn forecasts with their Gaussian counterparts.

$$BSS = \frac{BS_{\text{Gauss}} - BS_{\text{Previn}}}{BS_{\text{Gauss}}} \tag{1}$$

where the Brier score (BS) for each forecast (Previn and Gauss) is given by

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

where y_k are the forecast probabilities for a dichotomous event (as yet undefined) and o_k is determined from the observations: $o_k = 1$ if the event occurred and $o_k = 0$ otherwise – see Wilks (1995, pp. 259–260) for further details.

Since $BS \geq 0$ and $BS = 0$ only for a 'perfect' forecast, BSS takes positive values for improved (Brier) skill in Previn (BS_{Previn}) with respect to the chosen Gaussian (BS_{Gauss}), $BSS = 0$ if there is no difference in skill and $BSS < 0$ means that Previn is less skilful than the Gaussian. An overall value for BSS is calculated from the full 3 month data set at each forecast time. In the following, section a method for estimating confidence intervals for this overall value is outlined.

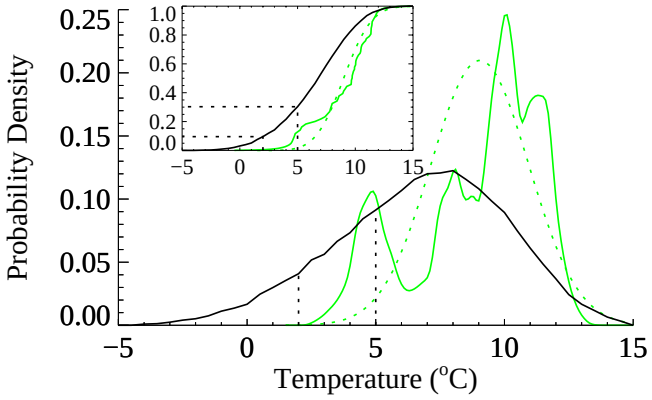

Fig. 1. An example of a KFMOS temperature forecast distribution (solid green), fitted Gaussian (dashed green) and climatological distribution (solid black lines) from 16 January 2002 at T+120 h, together with two forecast thresholds (black dashed lines) illustrating how forecasts of different events respond to the 'detail' of the forecast PDFs. The main plot contains the PDFs and the inset plot the cumulative distributions. The Gaussian was fitted to the middle of the CDF with $a = 0.68$ (see Sect. 2.3).

Table 1. Observed climatological probabilities of various events from winter 2001/2. The wind speed thresholds correspond with winds of at least Beaufort Force 7 and 9

Climatological probability	Midday (T+0)	Midnight (T+12)
$P_C\left(T \geq 2°\mathrm{C}\right)$	0.90	0.78
$P_C\left(T \geq 5°\mathrm{C}\right)$	0.70	0.52
$P_C\left(WS \geq 13.9\,\mathrm{ms}^{-1}\right)$	0.38	0.27
$P_C\left(WS \geq 20.8\,\mathrm{ms}^{-1}\right)$	0.14	0.08

2.1 Bootstrap confidence intervals

A bootstrapped resampling procedure is used to derive useful confidence limits for the Brier skill scores in order to allow meaningful statistical conclusions to be drawn from these comparisons. For this study, three months of forecasts and verifying observations were archived from 30 sites in the UK. The bootstrapping was performed by taking 10 000 random samples (with replacement) from this period, each sample being equivalent to 30 days' data. A distribution of BSS is then obtained by calculating the skill from each subsample.

The full skill score calculated from 3 months' data represents the average skill from the chosen season and the bootstrap resampling attempts to account for the variability between seasons. Although it is more common to take samples (with replacement) of the same length as the whole data set, this was found to produce what seemed to be unreasonably narrow confidence intervals. For example, the error bars in the figures that are discussed below would be 20–50% narrower if 90 day samples were used. It was felt that the use of 90 day samples would lead to over-confident conclusions regarding the level of skill in the full ensemble compared with the fitted Gaussian and so 30 day samples were adopted.

The error bars applied in the analysis below are 95% confidence intervals. They are derived from the bootstrapped distributions of BSS using the 'bias-correction and acceleration' (BCa) technique as described by Efron and Tibshirani (1993). The BCa intervals are corrections to the standard percentile intervals. For a 95% confidence interval the interval would be (0.025,0.975). The BCa technique adjusts this interval so that the mean of the bootstrap distribution matches the original estimate of the BSS (from the full data set) and the width gives a more precise estimate of the chosen confidence interval – see Efron and Tibshirani (1993) for details. The differences between the BCa intervals and the original were, in most cases, quite small (less than 1%) and do not affect the conclusions of this paper in any way. According to Efron and Tibshirani the BCa intervals are accurate to $\mathcal{O}(1/n)$, where n is the number of bootstrap samples, whilst the original percentile estimates are accurate to $\mathcal{O}(1/\sqrt{n})$. Therefore, the small differences between the BCa and original intervals may be due to the relatively large number of bootstrap samples used ($n = 10^4$).

2.2 Climatology of the events

We are interested here in the detail within the PDF. The event thresholds are chosen either to lie inside the bulk of the climatological distribution or in its tails, in order to test the skill of the unsmoothed PDF against its smooth Gaussian counterpart. A climatologically extreme event is one that is found in the tails of the climatological distribution or outside the distribution altogether. It frequently lies outside the forecast distribution as it usually has a low forecast probability. Forecasts of such an event do not rely on the detail of the forecast PDF and so probabilities from the smooth and detailed PDFs should be similar. Conversely, a climatologically frequent event should usually be found within the forecast PDF of a skilful forecast system. In this case, if the forecast PDF contains non-Gaussian detail then comparing the skill of the full PDF with a smooth Gaussian fit measures the improvement in forecast skill due to that detail.

This is illustrated in Fig. 1 where the observed climatological distribution of temperature is shown in black compared with a single KFMOS forecast distribution in green, together with the two event thresholds $T = 2°\mathrm{C}$ and $T = 5°\mathrm{C}$ (black dotted lines), whose climatological probabilities are listed in Table 1. The inset plot shows the equivalent cumulative distributions. In the main part of Fig. 1 it is evident that the $T = 2°\mathrm{C}$ threshold lies outside the forecast PDF in this instance, as will normally be the case for such a relatively extreme event. Thus, the bulk of the probability density lies to the right (in this case) of the threshold, which can also be seen from the inset cumulative plot, and the detailed shape of the PDF is irrelevant to the forecast of $P(T \geq 2°\mathrm{C})$.

Conversely, the $T = 5°C$ threshold lies within the bulk of the climatological distribution shown in Fig. 1 and so this threshold lies within the forecast distribution more often than $T = 2°C$. Probability forecasts using the $T = 5°C$ threshold are therefore sensitive to details in the full PDF. For example, the small spike in the full KFMOS PDF near 5°C in Fig. 1 is not reproduced by the fitted Gaussian (the dotted green line). In this case the probability forecasts from the full PDF and the fitted Gaussian differ significantly. This can be seen in the inset cumulative plots in Fig. 1 where it is clear that the full PDF (solid green line) and fitted Gaussian (dashed green line) are not coincident between 5°C and 7°C.

Table 1 shows the chosen events and their observed climatological probabilities (denoted by P_C) based on data from winter 2001/2 at midday and midnight local time. The temperature distribution is approximately Gaussian so the two thresholds lie to the left of and near to the mean, at either midday or midnight. The distribution of wind speed is shaped more like a gamma distribution, strongly peaked towards the origin with a long tail at high wind speeds. The two wind speed thresholds are chosen to lie within the 'bulk' of the distribution (13.9 ms^{-1}) and within the tail (20.8 ms^{-1}). This avoids the potentially inaccurate low wind speed end where the observations may have relatively large errors.

2.3 The Gaussian fit

In order to look for skilful information in the full PDF compared to a Gaussian fit, a good Gaussian representation of the PDF is needed (a poor Gaussian would be easy to beat and give misleadingly good test results). Since the Previn PDFs are often non-Gaussian in shape, especially beyond the approximately linear growth phase within the EPS 48 h singular-vector optimisation time, finding the Gaussian fit by maximum-likelihood (Wilks, 1995) may be unsatisfactory. Instead $\mathcal{G}(x)$, the Gaussian cumulative distribution function (CDF), is fitted to two points x_1 and x_2 on the Previn CDF, such that $\mathcal{G}(x_{1,2}) = \frac{1}{2}(1 \pm a)$, where a is the fraction of the unit area in either PDF between the two points and equal parts of the area lie in each tail. Varying a changes the position where the two CDFs match. Here $a = 0.68$ is used to fit the Gaussian near the middle of the Previn PDF so that x_1 and x_2 are approximately one standard deviation either side of the mean. $a = 0.95$ selects the tails of the forecast distributions so that x_1 and x_2 are approximately two standard deviations from the mean. This changes the position and width of the fitted Gaussian, as is illustrated in Fig. 2 where a Previn CAL+W PDF is shown (solid line) with two fitted Gaussian PDFs (broken lines). Using $a = 0.68$ fits the Gaussian near to the centre of the PDF, which is the left-most Gaussian in Fig. 2. The Gaussian fit with $a = 0.95$ matches the fitted CDF to the 2.5% and 97.5% points of the Previn PDF and thus, tends to select the tails rather than the middle of the PDF. In Fig. 2 the CAL+W PDF is skew and has a longer tail to the right, so the $a = 0.95$ fitted Gaussian is wider and shifted to the right than that obtained with $a = 0.68$.

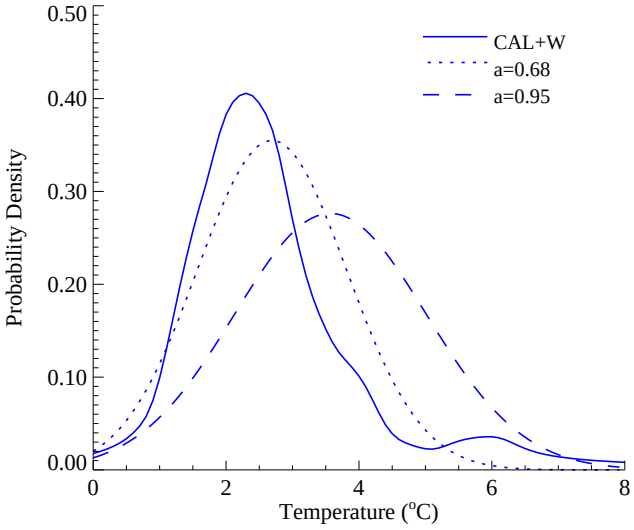

Fig. 2. An example CAL+W PDF (solid line) from 9 January 2002 at T+120 h. The broken lines are Gaussian fits using the method described in Sect. 2.3.

3 Comparing the skill of the Previn PDF to a fitted Gaussian

In this section a number of Previn forecasts are compared with different Gaussian forecasts. In each case data from the same three months are used from winter 2001/2002 and pooled over 30 UK stations. Ensemble data are available every six hours, from the analysis at midday on forecast day 1 (indicated by T+0) to midday on day 10 (T+240).

Error bars in the estimates of skill are the 95% BCa confidence intervals, giving a probability that the actual skill would lie outside the indicated range of 5%. In the ensuing discussion the word 'significant' is used to mean 'significant according to the 95% BCa confidence interval'.

3.1 Temperature forecasts

Figures 3 and 5 compare the skill of the full Previn PDF to the Gaussian that is fitted to the middle of the PDF with $a = 0.68$ for two temperature thresholds. In Fig. 3 (P ($T \geq 2°C$)) the 95% confidence error bars usually cross the zero-skill line, indicating there is no significant difference between the full PDF and its smoothed Gaussian counterpart. This is because the forecast threshold ($T = 2°C$) often lies outside the highest density parts of the forecast PDF (P_C ($T \geq 2°C$) is close to one in Table 1). This means that the detailed shape of the PDF is less important than the size of the lower tail and so the Brier skill of the smooth Gaussian is competitive with the full PDF.

One exception to this occurs in the first day of the forecast. The CAL forecasts (in red) are significantly less skilful than their Gaussian counterparts. This is because the EPS, with its singular vector optimisation time of 48 h, has insufficient spread at this time. This leads to a relatively large weight

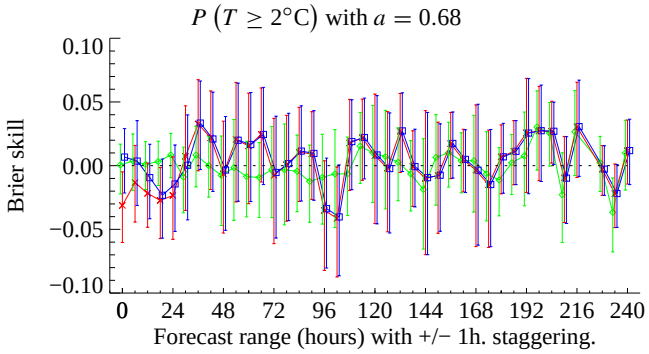

Fig. 3. Brier skill scores from KFMOS (green), CAL (red) and CAL+W (blue) Previn forecasts of 2 m temperature, comparing the full PDF against a Gaussian fit with $a = 0.68$. Note that for clarity the KFMOS and CAL+W curves are displaced by $\pm$1h, respectively.

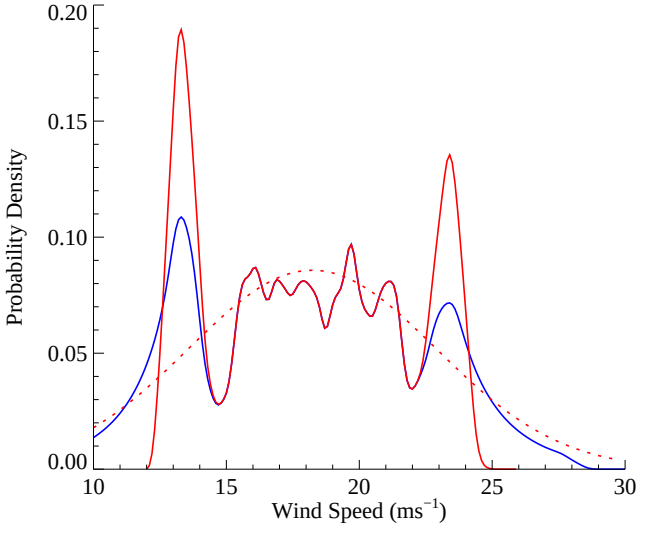

Fig. 4. An example of Previn temperature forecast PDFs from 4 January 2002 at T+12 h, illustrating the presence of spikes due to large outlier weights in the CAL forecasts (solid red). For comparison the fitted Gaussian is shown (dotted red line) as well as the CAL+W forecast PDF (solid blue line, overlapping the CAL between 15 and 22 ms^{-1}). The Gaussian was fitted to the middle of the CAL CDF with $a = 0.68$ (see Sect. 2.3).

being given to the outlier bins of the CAL ensemble (at T+24 the weight is 10 times larger than the ideal weight of $1/52$). Probabilities from the tails of the CAL PDF are found simply by extrapolating linearly this weight from the maximum or minimum of the ensemble over a width of 1 unit (either 1°C or 1 ms^{-1} as appropriate). The large outlier weight leads to spikes at the edges of the PDF and this is illustrated by the CAL PDF in Fig. 4 (the solid red line). The Gaussian is more skilful than the CAL PDF in Fig. 3 (hence the negative skill score) as it smoothes out the spikes introduced into the PDF by the large outlier weights, as illustrated in Fig. 4 by

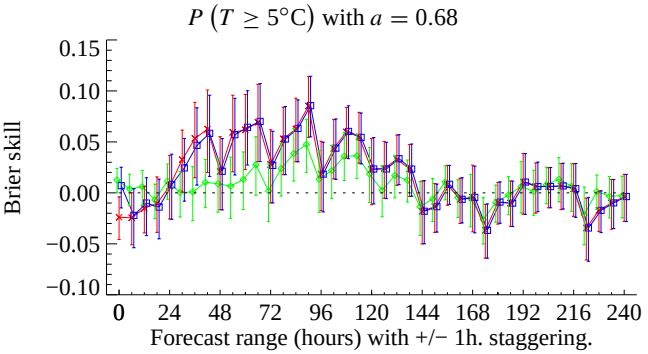

Fig. 5. Brier skill scores from KFMOS (green), CAL (red) and CAL+W (blue) Previn forecasts of 2 m temperature, comparing the full PDF against a Gaussian fit with $a = 0.68$. Note that for clarity the KFMOS and CAL+W curves are displaced by $\pm$1 h, respectively.

the red dashed line. The CAL+W forecasts (the blue line in Fig. 4) improve on the CAL by replacing the 1 unit extrapolation with fitted Weibull tails (see Denholm-Price and Mylne (2002) for a full discussion) and there are no spikes in the unweighted KFMOS PDFs, so the skill for these forecasts before T+24 h in Fig. 3 remains broadly neutral.

Figure 5 tells a different story. In this case the threshold $T = 5°\text{C}$ lies in the middle of the climatological distribution and so these forecasts test the detail within the bulk of the PDF. Unlike Fig. 3 there are now times when the full PDF is significantly more skilful than the Gaussian. This is most evident in the calibrated forecasts (CAL and CAL+W) between T+36 and T+114 but occasionally includes the KFMOS. The midday forecasts in that period (T+48, T+72 and T+96) are exceptions as the 95% confidence limit crosses the zero skill line. At midnight the climatological value, P_C ($T \geq 5°\text{C}$), is close to 0.5 which maximises the impact of any detail in the forecast PDFs compared to the smooth Gaussian. At midday the climatological probability is larger so the threshold is shifted from the mean of the distribution. It follows that there are more forecasts where the threshold lies away from the mean of the forecast PDF, the forecast skill is less affected by any detail in the PDF and therefore the skill of the full PDF over the Gaussian is reduced. Note that this argument only holds as long as the observed and forecast climates are similar, which is the case for temperature at all values but only for wind speed beyond 12 ms^{-1}.

3.2 Wind speed forecasts

In Fig. 6a there is no evidence of significantly positive skill in the forecasts of $P\left(WS \geq 20.8\,\text{ms}^{-1}\right)$ over the Gaussian fit for what is, according to Table 1, a relatively rare event. Like Fig. 3, the negative skill in the early forecast range is due to spikes in the PDF introduced by the calibration (CAL) being smoothed by the Gaussian. However this problem also extends at one forecast time to the CAL+W, indicating that

the Weibull tails are perhaps insufficiently broad in this case, although it is unlikely that forecasts at this range would be useful as the forecast data are not available until after T+12.

Beyond T+12 there is no significant difference in skill between the fitted Gaussian and the full PDF (except at T+60 and T+72 where the calibrated skill is slightly positive). This is despite the fact that the fitted Gaussian cannot model any skewness in the wind speed PDFs. Since $WS \geq 20.8\,\mathrm{ms}^{-1}$ is a fairly extreme event, as shown in Table 1, the threshold will usually lie to the right of the forecast PDF. Therefore, it is unsurprising that the skill in Fig. 6a is mostly not significantly different from zero. The use of a Gaussian fit has little bearing on the conclusions in this case as the threshold is usually outside the bulk of either PDF.

In Fig. 6b forecasts of wind speed for the lower threshold $WS \geq 13.9\,\mathrm{ms}^{-1}$ are considered. As before, there is negative skill in the CAL and CAL+W forecasts before T+12. In this case, however, there is some significantly positive skill between T+48 and T+120, although it is smaller in magnitude than for the temperature forecasts in Fig. 5 and only occurs with the calibrated forecasts (CAL and CAL+W). This indicates that the fitted Gaussian is less skilful than the full PDF on occasion for CAL and CAL+W. For all KFMOS forecasts and for the calibrated forecasts beyond T+120 (other than T+150) there is no significant difference in skill between the Gaussian fit and the full PDF. This is despite the fact that the Gaussian is not the most appropriate PDF to fit to a bounded parameter that has a skewed climatological distribution like wind speed. Comparing the full PDF to a better fit (e.g. a Gamma distribution) is unlikely to change the conclusion that generally there is little skill in the full PDF over the fitted PDF, although a better fit might remove the small positive skill found in Fig. 6b.

Results obtained by considering a threshold closer to the left of the climatological distribution, e.g. $WS \geq 8\,\mathrm{ms}^{-1}$ (Beaufort force 5, whose climatological probability is 0.66 and 0.55 at midday and midnight, respectively) may be influenced by errors in the wind speed observations which are largest at low wind speeds. This may be why, when repeating the tests that lead to Figs. 6a and b for $P\left(WS \geq 8\,\mathrm{ms}^{-1}\right)$ a negative skill 'tail' is observed beyond T+192 (not shown) where the full PDF is less skilful than the fitted Gaussian. There the observed climate differs noticeably from the model climate, which may be due to observational errors.

3.3 Fitting the Gaussian to the distribution tails

One criticism of the Gaussian fits with $a = 0.68$ is that they may fail to capture the full width of the PDF. In the 51 member EPS, the outlier bins nominally contain $\frac{1}{52} \approx 1.9\%$ of the probabilistic weight. After calibration this is usually greater than $2\frac{1}{2}\%$ (see Denholm-Price and Mylne (2002) for details). Fitting the Gaussian to match the cumulative distribution at the points enclosing 95% of the area ($a = 0.95$) ensures that the $2\frac{1}{2}\%$ and $97\frac{1}{2}\%$ points of the two distributions match. This forces the Gaussian to match at least one point in the tails of the Previn PDFs.

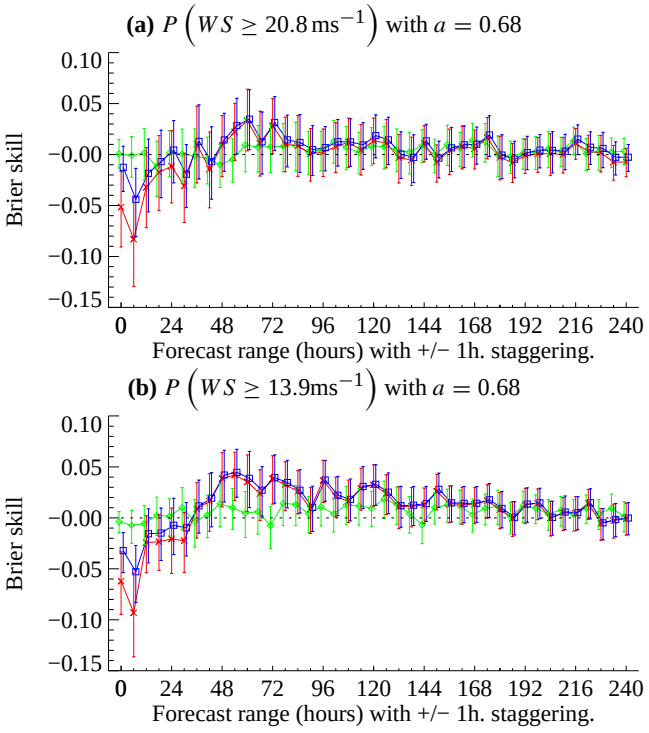

Fig. 6. Brier skill scores from KFMOS (green), CAL (red) and CAL+W (blue) Previn forecasts of 10 m wind speed comparing the full PDFs against Gaussians fitted with $a = 0.68$. Note that for clarity the KFMOS and CAL+W curves are displaced by ± 1 h, respectively.

Figure 7 shows the results of repeating the process used to derive Fig. 5, this time with $a = 0.95$. At first sight the skill appears improved (there is a wide region where all three processes have positive skill). In fact this demonstrates that the Gaussian fit with $a = 0.95$ is a worse forecast than the $a = 0.68$ fit: The Previn forecasts are unchanged yet their skill with respect to the $a = 0.95$ Gaussian is greater than their skill with respect to the $a = 0.68$ Gaussian. The apparently increased skill in Fig. 7 is due to the poor performance of the wider Gaussian.

4 Discussion and conclusions

Smoothing the full Previn PDFs (KFMOS, CAL and CAL+W) with a fitted Gaussian causes a reduction in forecast Brier skill in some cases (depending on the forecast range and threshold). This has been demonstrated with probability forecast thresholds in 2 m temperature and 10 m wind speed from site-specific forecasts derived from the ECMWF EPS. Bootstrapping was used to check the statistical significance of the results. The results indicate that there is more information in the site-specific temperature PDFs than can be represented by a single Gaussian PDF, no-matter how well fitted it is to the original PDF. The wind speed tests show little or no significant difference in skill.

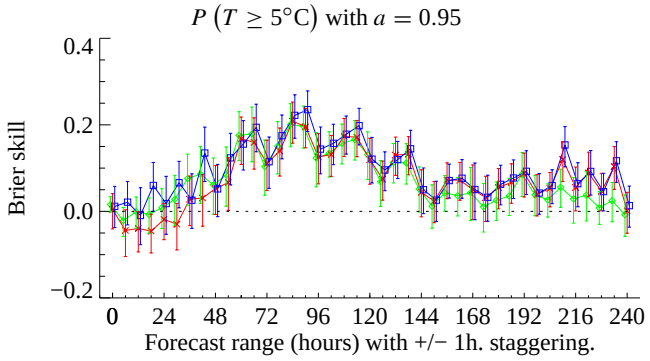

Fig. 7. Brier skill scores from KFMOS (green), CAL (red) and CAL+W (blue) Previn forecasts of 2 m temperature, comparing the full PDF against a Gaussian fit with $a = 0.95$. Note that for clarity the KFMOS and CAL+W curves are displaced by ± 1 h, respectively.

The results highlight, in a different way from the usual measures of skill, the importance of bias-correction and calibration when generating probabilistic forecasts for specific sites: The CAL and CAL+W forecasts give relatively more skilful detail than the bias-correction alone (KFMOS). In addition, the CAL+W forecast can improve on the CAL forecast by removing erroneous detail that might be inadvertently added to the PDF by a poor representation of the distribution tails.

In the future these tests should be expanded to include the decomposition of Brier skill into its components to assess whether the increased skill in temperature PDFs is due to improved reliability and/or resolution. It would also be useful to examine different measures of probabilistic forecast skill, such as the ranked probability skill score (RPSS) or forecast value (Richardson, 2000). It is to be hoped that the conclusions drawn from the temperature forecasts are sufficiently robust that they would also be supported by these different measures.

Since the PDFs at lower wind speeds tend to be skew, a fitted gamma distribution is likely to model the wind speed PDFs more closely than the Gaussian used here. It would therefore be a better smooth forecast against which to test the full PDF and as such might reduce the slightly positive skill in Fig. 6 to zero throughout. However, the conclusions drawn from the temperature PDFs would remain.

A measure of skill like the RPSS, which uses information from the whole PDF, might be more sensitive to the PDF shape than the Brier score used here. Similarly it would also be useful to investigate models with more than one Gaussian (or gamma) distribution. This would enable the examination of how much more information than a single Gaussian exists within the PDFs.

When presenting ensemble PDFs to users it is useful to smooth the forecast PDFs in order to remove inappropriate detail. Results from studies of this type may be used to determine how much smoothing is justified in order to retain the skill gained by calibrating the forecasts.

Acknowledgements. Thanks go to the Met Office for continued access to the Previn data since my move to Kingston University, to the EGS NP organising committee for soliciting the presentation of this work in 2001, to an EGS participant whose comments prompted further investigation into the use of the bootstrap technique and to the two anonymous reviewers whose detailed comments helped to improve the paper and suggested areas for future investigation.

References

Atger, F.: The skill of ensemble prediction systems, Mon. Wea. Rev., 127, 1941–1953, 1999.

Bauer, E.: Characteristic frequency distributions of remotely sensed in situ and modelled wind speeds, Int. J. Climatol., 16, 1087–1102, 1996.

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

Denholm-Price, J. C. W. and Mylne, K. R.: Report on calibrating probability forecasts from the ECMWF Ensemble Prediction System, Tech. Rep., The Met Office, forecasting Research Technical Report No. 386, 2002.

Efron, B. and Tibshirani, R. J.: An introduction to the Bootstrap., Monographs on Statistics and Applied Probability, vol. 57, Chapman and Hall, 1993.

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

Hamill, T. M. and Colucci, S. J.: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts, Mon. Wea. Rev., 126, 711–724, 1998.

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

Legg, T. P., Mylne, K. R., and Woolcock, C.: The use of medium-range ensembles at the Met Office. I: PREVIN – a system for the production of probabilistic forecast information from the ECMWF EPS, Meteorol. Apps., 9, 255–271, 2002.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: Methodology and validation, Quart. J. Roy. Meteorol. Soc., 122, 73–119, 1996.

Orrell, D., Smith, L., Barkmeijer, J., and Palmer, T.: Model error in weather forecasting, Non. Proc. Geophys., 8, 357–371, 2001.

Palmer, T. N.: Predicting uncertainty in forecasts of weather and climate, Reports on Progress in Physics, 63, 71–116, 2000.

Richardson, D. S.: Skill and relative economic value of the ECMWF ensemble prediction system, Quart. J. Roy. Meteorol. Soc., 126, 649–667, 2000.

Toth, Z. and Kalnay, E.: Ensemble forecasting at the NMC: The generation of perturbations, Bull. Amer. Meteorol. Soc., 74, 2317–2330, 1993.

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