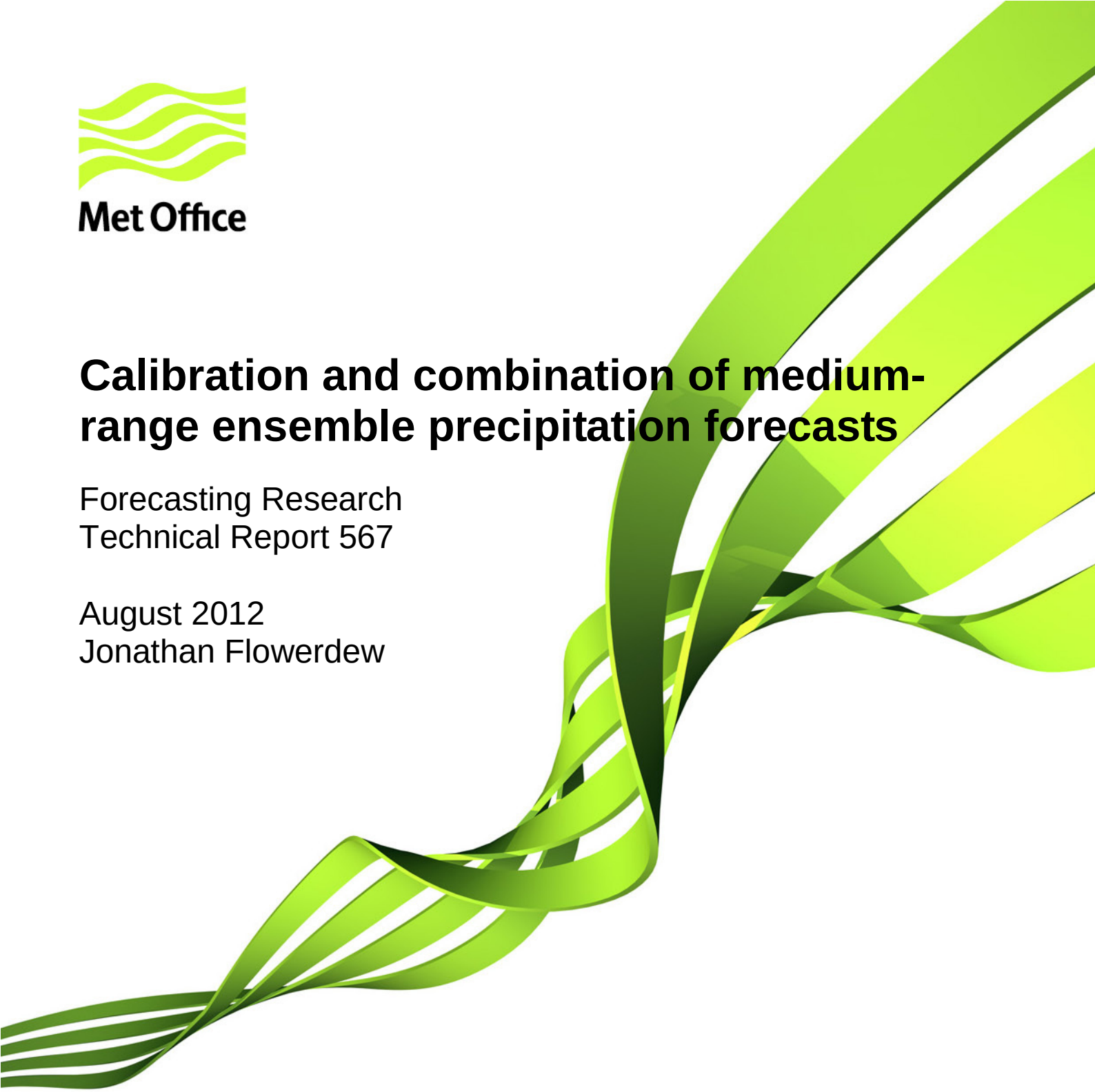

Calibration and combination of medium-range ensemble precipitation forecasts

Forecasting Research
Technical Report 567

August 2012

Jonathan Flowerdew

Contents

Abstract

Ensemble forecasts aim to improve decision-making by predicting the distribution of possible outcomes. Raw forecasts from a single system can be subject to deficiencies in climatology, spread, or the ability to resolve events, due to limitations in the underlying model, data assimilation and ensemble perturbations. Combining forecasts from multiple ensembles can provide extra samples and average over errors specific to individual systems. Statistical post-processing can use suitable verification data to re-shape the raw forecasts to improve their accuracy and reliability. This report investigates the impact of these inter-related techniques on ensemble forecasts of precipitation, which is both a key variable for forecasts users and a challenging variable for forecast models and statistical post-processing schemes. For 12-hour accumulations on a 1˚ grid over the UK, it is found that errors in climatology are relatively small, whilst errors in reliability and spread as a function of forecast magnitude are more significant. A calibration method motivated by these diagnostic results uses reliability tables to re-shape the forecast distribution, targeting both climatology and spread in one coherent operation. This produces a set of ensemble members retaining spatial, temporal and inter-variable structure from the raw forecasts, which should be beneficial to downstream applications such as hydrological models. The method virtually eliminates unreliability when verified against fixed thresholds, although some issues remain against more demanding verification using thresholds that vary with local climatology.

1. Introduction

Ensemble weather forecasts aim to improve decision-making by predicting the distribution of possible outcomes. The most useful forecasts will have a distribution which is as sharp as possible, whilst remaining statistically reliable, and include relevant probabilities of extreme or dangerous events.

Ensemble forecasting systems seek to achieve these aims by using weather models which are as accurate as computationally feasible, whilst also sampling the impact of relevant sources of uncertainty such as initial state and model formulation. All forecasting systems are imperfect: they may perform better in some situations than others, may suffer from systematic errors in both climatology and spread, and provide only a small sample of possible outcomes.

There are two broad post-processing techniques which have been used to mitigate the impact of these deficiencies on downstream products. Past observations can be used to quantify systematic errors such as bias (eg Johnson and Swinbank, 2009) or statistical reliability (Primo *et al.*, 2009) and adjust future forecasts based on these results. This process is known as forecast calibration, and a wide variety of methods have been proposed (Atger, 2001; Stensrud and Yussouf, 2007; Coelho *et al.*, 2006; Fraley *et al.*, 2010; Hagedorn *et al.*, 2008; Hamill *et al.*, 2008; see also Applequist *et al.*, 2002 for deterministic input). On the other hand, forecasts from multiple ensemble systems can be combined into a multi-model ensemble (Park *et al.*, 2008; Johnson and Swinbank, 2009; Fraley *et al.*, 2010). This increases the member count, samples over structural uncertainty and models that may do better or worse in different situations, and creates the potential for cancellation of systematic errors, all without necessarily requiring large volumes of past training data. There has been some debate in the literature over whether multimodel ensembles or calibration of the best single model ensemble provide the optimum practical forecasting system (Park *et al.*, 2008; Fraley *et al.*, 2010; Hagedorn *et al.*, 2012; Hamill, 2012). At the same time, one might hope for extra benefit by applying both techniques together.

Previous work at the Met Office examined the benefit of multimodel combination, bias correction, and spread adjustment for variables such as mean sea-level pressure and surface temperature (Johnson and Swinbank, 2009). This report focuses on precipitation, which is a key variable for most forecast users, including the general public, flood protection, and water management. It is also a particularly awkward variable, challenging to both the underlying meteorological models and statistical post-processing systems. The statistical difficulties include small spatial scale, strong spatial variations in climatology, the importance of timing errors, the tendency of error to scale with forecast magnitude, a non-Gaussian distribution, and the 'nugget' of probability at zero precipitation. These factors have a profound influence on which calibration methods will be effective: for instance, a simple additive bias correction does not make much sense given the special meaning of zero precipitation. Whilst the immediate focus of this report is on medium-range precipitation alone, the techniques presented are sufficiently general that it is hoped they may also be useful for other variables, and indeed for higher-resolution short-range forecasts.

The rest of this report is laid out as follows. Section 2 describes the sources of forecast and observation data used in this study. Section 3 provides basic verification of the raw forecasts, and a simple multimodel combination. The next two sections investigate the nature of the systematic errors which a calibration scheme might attempt to correct, examining the forecast climatology in section 4 and the spread-error relationship in section 5. Attempts to directly calibrate these attributes met with limited success, but section 6 describes a method based on reliability tables which is generally beneficial. Particular aims of this method include simultaneous calibration of climatology and forecast uncertainty, a lack of assumptions about the shape of the underlying distributions, and output as a set of ensemble members retaining spatial structure from the raw ensemble. Conclusions are given in section 7.

2. Datasets

2.1. Forecasts

Forecast data for this project have been taken from the THORPEX Interactive Grand Global Ensemble (TIGGE; Bougeault *et al.*, 2010) archive, http://tigge.ecmwf.int/. This provides access to data from a variety of global medium-range ensemble forecasting systems. Like Johnson and Swinbank (2009), this study only considers forecasts from the European Centre for Medium Range Weather Forecasts (ECMWF), UK Met Office (UKMO) and United States National Centers for Environmental Prediction (NCEP). This triad provides significant model diversity whilst keeping data volumes manageable. They are the three models which could be most easily obtained in real-time by any future UK operational multimodel system. They are also amongst the best performing models in the archive, helping to illustrate the best performance which might be obtained, and also reducing the importance of issues such as weighting of low-skill models in multimodel combination.

For simplicity, the results presented here consider only perturbed forecast members, without the unperturbed control forecasts. This makes each ensemble a homogenous unit, and creates a statistically purer target for initial investigation. In the case of the Met Office ensemble, including the control forecast with the perturbed members would be contrary to the principles underlying both the Ensemble Transform Kalman Filter (ETKF; Wang and Bishop, 2003) which generates the initial perturbations, and the online spread calibration system (Flowerdew and Bowler, 2011). Control forecasts do provide extra information, with lower root mean square (rms) error than perturbed members, so an optimal multimodel system would probably want to make use of them. However, this raises further questions such as how to optimally weight the control forecast, and

whether this weight should vary with lead time. In any case, early experiments suggested that the inclusion or exclusion of control members makes little difference to the verification scores; they are after all a small fraction of the total member count.

The main results of this report consider data covering a period of almost two years (September 2007 to July 2009). This provides a reasonable sample for both training and verification, and was the longest period for which all relevant data was conveniently available. To limit the data volume, only 00 UTC forecasts have been considered, evaluated in successive 12-hour intervals from 0 to 15 days. For convenience, and to avoid downloading the full global fields, data were interpolated on the ECMWF computer system to a common 1-degree grid over Europe. This is fairly close to the native grid-scale of the forecasting systems during the period considered. Unfortunately, the interpolation routine used by the archiving system is known to have problems deriving accumulations across the reduction in grid resolution of the ECMWF ensemble at T+10 days. Many of the ECMWF and multimodel results below show glitches at this lead time: to the extent that they affect this lead time alone, they may well reflect this unavoidable limitation of the archiving system rather than the actual response of the ECMWF ensemble to the change in grid.

2.2. Observations

All of the results shown in this report use observations taken from the Met Office 'ukpp' analyses. These combine gauge-adjusted radar, satellite-derived precipitation, and high-resolution short-range forecasts on a 2 km grid over a domain slightly larger than the British Isles. The model forecasts are compared to the average of all ukpp pixels whose centre lies within each model gridbox. In the time dimension, the observations are similarly integrated from rain rates at hourly intervals to 12-hour accumulations. This helps to create a 'fair' evaluation, in which the model predictions of grid-box average precipitation are compared to observations that are genuinely representative of the whole gridbox. Whilst many of the techniques considered in this report could be applied to the problem of mapping gridbox-average predictions to individual stations, this is not considered here, and would likely result in lower predictive skill.

In order to reduce statistical noise and test a wider range of regimes, it was originally hoped to use an equivalent 'europp' dataset covering a larger European domain, making use of the European radar network. However, initial studies indicated that the observations were not comparable to ukpp and led to poorer verification scores over the ukpp domain before 2009. The main results therefore focus on the UK domain alone. Some additional verification was performed against short-range forecasts in order to use the full European domain, but these results will not be presented in detail.

3. Verification of raw forecasts

Figure 1 measures the overall performance of raw ensemble forecasts via the implied probabilities to exceed chosen rainfall thresholds. The plots show the decomposition of the Brier Skill Score (BSS; Wilks, 2006) into its 'resolution' and 'reliability penalty' components: the latter measures the difference between each forecast probability and the observed event frequency when that probability is forecast, whilst the former measures the ability to discriminate between situations in which the event is very likely to occur and situations in which it is very unlikely. The reliability penalty can in principle be eliminated by remapping the forecast probability values given sufficient representative training data. The resolution score is more tied to the quality of the underlying forecast system, although it can be affected by corrections to the bias or spread of the underlying ensemble, since these will change the distribution of members with respect to thresholds. The overall BSS is the resolution score minus the reliability penalty. A perfect forecast would have a resolution score of one and a reliability penalty of zero, whilst BSS

less than zero indicates a forecast no better than the climatological probability of the event.

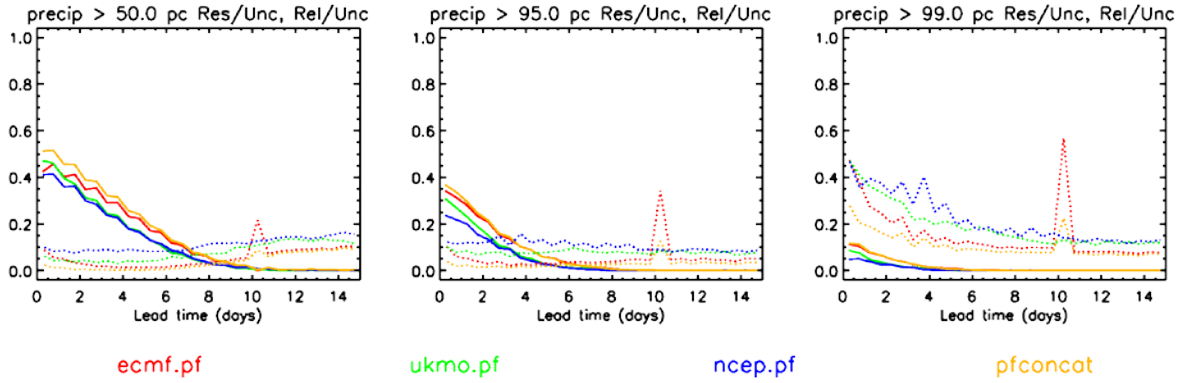

Figure 1. Resolution (solid) and reliability (dotted) components of the Brier Skill Score for selected quantiles of local in-sample climatology. Forecasts use the perturbed members from the ECMWF (red), Met Office (green) and NCEP (blue) 15-day ensemble systems, and the simple aggregation of all these members (yellow). The spike in ECMWF reliability penalty at T+10 days is associated with a coarsening of the model grid, and may be an artefact of the TIGGE archive interpolation code, as discussed in section 2.1.

The verification thresholds have been specified indirectly, as quantiles of the relevant month's observations for the same time of day within a 5x5-gridbox domain centred on each gridpoint. This reduces 'false skill', where forecasts achieve positive BSS simply by knowing the climatological variation of the event probability with season, location, or time of day (Hamill and Juras, 2006). It defines the event by whether it is 'normal', 'unusual' or 'rare' for each location. The score is then uniformly representative of all locations, and statistical noise is minimised. When interpreting the quantile values, it should be noted that they refer to the climatology of the observations within each target month, rather than overall climatology. Thus, the 99th percentile measures the ability to identify the top 1% of local events within each month, whereas a typical month might not contain any events reaching the top 1% of the overall climate. This approach focuses on forecast performance for the low and high precipitation events within each month, rather than climatological extremes which will arise in only a few months and would be poorly and unevenly sampled even within a two-year dataset. It also avoids the need for a long archive of ukpp observations from which to construct an accurate overall climatology, which would be of dubious value anyway given that the ukpp system has not been designed for long-term stability.

Both the reliability and resolution scores rank the single-model ensembles in the order expected from the quality of the underlying models and the fineness of the mesh on which they are run. The resolution scores decay towards zero at long lead times, as the forecasts lose their ability to discriminate between situations in which the event is more or less likely. The behaviour of reliability with lead time depends on the threshold. The underlying reliability diagrams (not shown) all decline to a situation in which most forecast probabilities just give the climatological event frequency. However, the range of forecast probabilities narrows more sharply with lead time for the higher percentiles, so that the overall reliability score improves with lead time rather than declining as for the 50th percentile. The combination of reliability and resolution makes the forecasts worse than climatology (BSS < 0) at long lead times for the lower quantiles, and all lead times for the 99th percentile. The high values of reliability penalty suggest significant scope for calibration of forecast probabilities, something that is much less apparent in verification against fixed thresholds (not shown).

3.1. Multimodel combination

The yellow curves in Figure 1 consider the aggregation of all three ensembles, where each individual member is given equal weight (the alternative of giving each ensemble equal weight is discussed below). The results indicate that this simple post-processing strategy does indeed produce precipitation forecasts which are equivalent or superior to the best single-model ensemble. The advantage in resolution score is greatest for lower thresholds and shorter lead times. The reliability advantage is also largest at short lead times, but remains nonzero at long lead times and is pronounced for both high and moderate thresholds.

In principle, the multimodel ensemble has two kinds of potential advantage compared to the single model ensemble: it samples over structural uncertainty in the modelling system formulation, and it simply has more samples to fill out the predicted distribution. To examine the relative importance of these effects, similar verification was performed using only 20 members from each configuration, randomly selected for each data time. For the larger ensembles (ECMWF and multimodel), this harms resolution at all lead times, and reliability at long lead times (not shown). The multimodel advantage over ECMWF in both resolution and reliability remains at short lead times, suggesting that it comes predominantly from the diversity provided by multiple distinct systems. At longer lead times, the multimodel advantage in reliability is not just lost but reversed for higher thresholds. The detriment to both systems fits the fact that high member count is needed to represent low probabilities. The smaller degradation of ECMWF perhaps reflects the fact that the restricted multimodel is trading good ECMWF members for poorer members taken from the other systems.

Theoretically, one might debate whether it is more appropriate to give equal weight to each member or each underlying ensemble system. Model-weighted combination emphasises the systematic differences between the different systems, helping cancellation of systematic errors, and reducing the dominance of ECMWF data that otherwise arises from its large number of members. A member-weighted combination, as used here, assumes each member is an equally valuable sample of the possible outcomes. In practice, early tests (not shown) demonstrated that these two alternative choices make very little difference to the verification results, with perhaps a slight preference for member-based weighting. It is possible that ensemble-based weighting may be more advantageous in situations where forecast bias is more importance (see section 4 below) or where the ensemble with the greatest number of members does not also happen to be the one with the best deterministic skill.

The most obvious theoretical justification for simple aggregation as a method of ensemble combination would be if there was one universal distribution of possible outcomes for any given forecasting situation, and all members from all ensembles are samples of this distribution. However, this is clearly not the case. Some models are more accurate than others, so their distribution of possible outcomes should be narrower. Different models may be better at handling different forecasting situations, and thus able to exclude outcomes that other models cannot. In the extreme case of two forecasts which are statistically independent, the uncertainty given both forecasts is related to the product, not the average, of the two PDFs, producing a distribution which narrows as more and more independent forecasts are added (as is familiar in the case of averaging over many independent measurements). However, early experiments (not shown) demonstrated that the errors of the mean precipitation forecasts from different ensemble systems are in fact highly correlated, and that simple product-like combination does not perform well. The general solution to the ensemble combination problem will be given by Bayes' Theorem, where correlated errors are expressed through conditional probabilities of one forecast on the other. Without working through the maths in detail, it seems reasonable that the convolution created by error correlation will produce a result more

like averaging the component PDFs as correlation increases (although its standard deviation should not exceed the smallest of the input standard deviations if the weights are optimal).

4. Climatology comparison

The first type of forecast deficiency which many post-processing systems attempt to correct is any systematic difference between forecast and observed values. This is to be contrasted with systematic errors in the prediction of forecast uncertainty, which are discussed in subsequent sections. For instance, a simple post-processing scheme for surface temperature forecasts might seek to identify and remove the mean difference (bias) between forecast and observations (eg Johnson and Swinbank, 2009). This could be particularly important in a downscaling context where gridbox average forecast values are to be mapped to a single station which may have an atypical elevation and meteorological context.

The statistical features of precipitation argue against the use of simple bias corrections. Overall additive terms would incorrectly affect all forecasts of zero precipitation. Corrections which are multiplicative or only affect nonzero forecasts do not help to adjust the frequency of zero precipitation forecasts. Here, a more general approach is taken, comparing quantiles of the forecast and observed climatologies. This provides a natural way to compare distributions of any shape, and identify potentially different biases for heavy as opposed to light precipitation. (A similar differentiated comparison could be achieved using the bias conditioned on narrow ranges of forecast values. However, this has the disadvantage of damping forecast variability by drawing towards the climatological mean as forecasts become less accurate).

Here, as in the rest of this report, observation error is assumed negligible in comparison to forecast error. If observation error were non-negligible, the forecasts should be compared not to observed climatology, but to a narrower climatology of 'truth', derived so that it reproduces observed climatology when dressed with observation error.

Figure 2 illustrates the variation of forecast and observed climatology with calendar time (top) and lead time (bottom). The calculation starts by finding the 95th percentile of values within a 5x5-gridbox domain centred on each gridpoint, for each data source and three-month block of data. The top plot shows results for each three-month block at a single lead time, whilst the bottom plot takes the mean at each gridpoint over all three-month blocks, to summarise the overall behaviour whilst giving equal weight to all seasons. Thus far, the calculation produces separate values for each gridpoint. To summarise these spatial distributions, the plots show the 10th, 50th and 90th percentiles over space of their respective gridpoint results. Thus, the 95th percentile in a 'typical' gridbox follows the solid lines, whereas the dotted lines illustrate the gridboxes with unusually low or high 95th percentiles in each particular case. The following discussion focuses mostly on the solid lines.

The top plot shows that the forecasts and observations exhibit broadly similar annual cycles. The forecasts tend to underpredict this quantile, but there is significant variability about this average result. This has two implications for calibration. First, a large sample is required to separate signal from noise. Recalculating the lead time plot for different data volumes suggests about a year of data is needed to stabilise the calibration signal. Second, even once this signal is stabilised, it can only correct the overall mean error, not the full difference between forecast and observations for any one three-month period.

The bottom plot shows that the calibration signal is fairly stable with lead time, although there is some drift at short lead times, and the bias of the ECMWF model changes following the reduction in grid resolution at T+10 days. The fact that the dotted lines show less systematic error than the solid lines suggests there is an important spatial

component to the bias signal. Overall, similar plots suggest that the forecasts tend to overpredict the 15–35th percentile (~0.1mm), underpredict the 60–95th percentile (0.8–7mm), and overpredict the 99th percentile (~12mm).

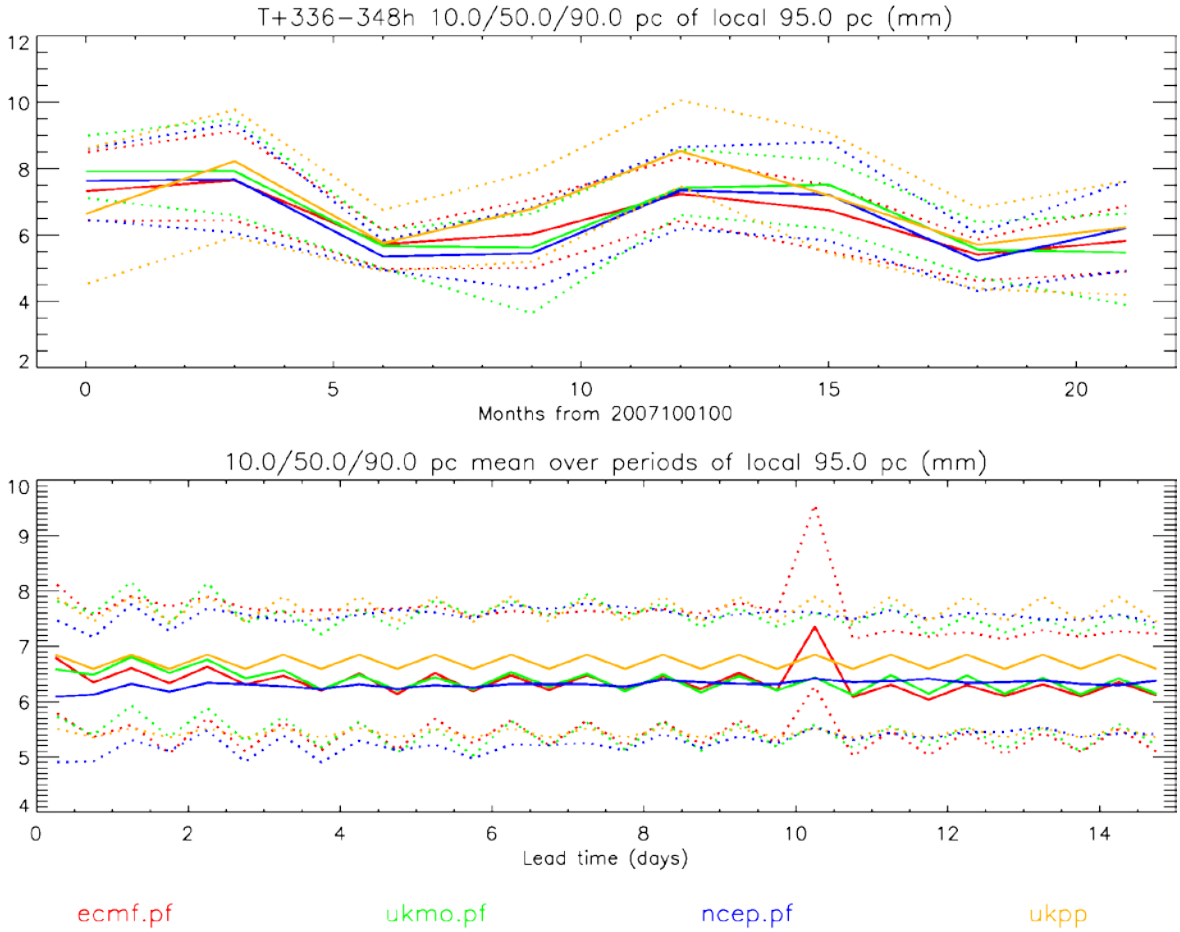

Figure 2. 10th (lower dotted), 50th (solid) and 90th (upper dotted) percentiles over space of the 95th percentile of three-month 5x5-gridbox climatologies. The top plot shows results at T+14.0-14.5 days as a function of the central month. The bottom plot shows results as a function of lead time, where the mean of the 95th percentiles from all three-month periods has been calculated for each gridpoint before taking quantiles over space. Red, green and blue identify forecast sources as in Figure 1, whilst yellow indicates the ukpp observations.

Figure 3 summarises the mapping from forecast to observed climatology based on 11 months of data for one particular lead time. A line is shown for each 5x5 group of gridpoints, testing the forecasts' ability to track spatial variations in climatology rather than just its overall statistics. The series of bias trends listed in the previous paragraph are just about discernable, but the plot emphasises that the UK-mean biases are small in comparison to the corresponding rainfall magnitudes. Against ukpp data averaged up to the same 12-hour accumulations on a 1˚ grid, it appears the forecast climatology is really quite good. The scatter of lines shows larger biases for particular gridpoints, suggesting some potential benefit from a spatially-localised climatology calibration.

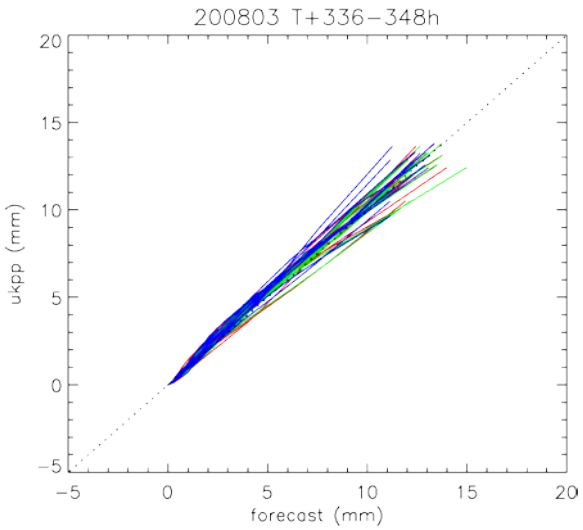

Figure 3. Mapping between corresponding quantiles of forecast and ukpp data at T+14.0-14.5 days in the eleven-month 5x5-gridbox climatology covering 200710-200808. A line is shown joining 15,20,…90,95,97,99% quantiles for every fifth gridpoint in each direction (quantiles below 15% were omitted since they are generally zero). Results are shown for the ECMWF (red), Met Office (green) and NCEP (blue) perturbed-member ensembles.

4.1. Climatology calibration

Having compared quantiles of forecast and observed climatology, it is a small step to ask whether this mapping can be used to improve the forecasts. In principle, this provides a very general way to correct the mean, width or shape of the forecast climatology. The Local Quantile-Quantile Transform used by Bremnes (2007) is based on the same idea. A simple test was constructed as follows. The available two-year period was divided into three-month blocks, where the four 'preceding' blocks (in a cyclic sense) are used to calibrate each target date. Within each block, the 15,20,…,90,95,97,99th percentiles of the 5x5-gridbox domain around each gridpoint were identified, separately for each data source and lead time. The calibration uses the average of these quantiles over the four blocks. An average of three-month quantiles was chosen over twelve-month quantiles to make the result equally applicable to all seasons and avoid the sampling noise that might otherwise result from results being dominated by the most extreme seasons. The forecast values are calibrated by linear interpolation/extrapolation between the matching percentiles of forecast and observed climatology, as illustrated in Figure 3. The percentiles were chosen to explore the resolved shape of the climatology mapping, whilst hopefully limiting noise sufficiently that extrapolation at the extremes remains plausible. The detailed treatment of both zero and very large precipitation amounts will be discussed in more detail in section 6.1.

Results (not shown) demonstrate that this approach is fairly effective at correcting the overall magnitude of the sampled quantiles, their diurnal cycle, and the ECMWF spike and change at T+10 days. Some straightening and tightening of the residual calibration curves can be seen in plots equivalent to Figure 3. For the components of the Brier Skill Score, there is a small positive impact on resolution, but some degradations to reliability in cases where bias is removed without correcting overspread. These results broadly support the idea that bias, even generalised to a reshaping of climatology, is not the most important calibration issue for the forecasts considered here.

5. Spread-error comparison

Unlike single deterministic forecasts, ensemble systems aim to provide an estimate of case-specific uncertainty by predicting the complete distribution of possible outcomes. This section focuses on comparing the width of the predicted and actual error distributions: mathematically, whether the mean square deviation of the ensemble members from the ensemble mean (plus the mean square observation error, generally assumed small) matches the mean square deviation of truth from the ensemble mean.

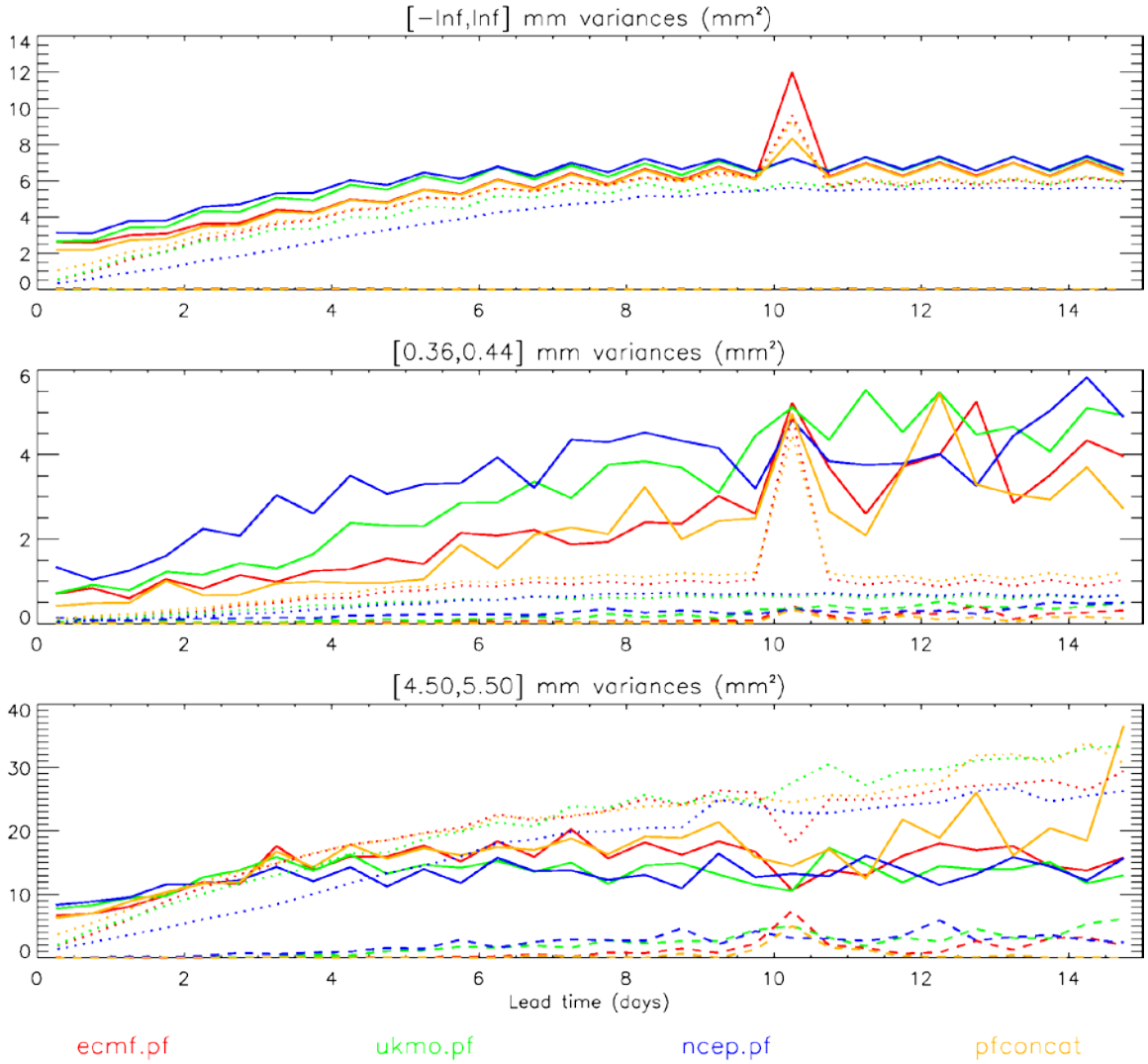

Figure 4. Mean square spread (dotted), bias (dashed) and the error variance of the ensemble mean (solid), averaged over all cases (top), and those with ensemble mean forecasts around 0.4 mm (middle) and 5.0 mm (bottom).

Figure 4 shows spread and error statistics as a function of lead time. Mean square quantities are shown rather than standard deviations so that bias and variance add linearly, and any independent observation error just appears as a constant offset. The top plot shows results averaged over all cases. This suggests reasonable correspondence between spread and error, consistent with a small observation error, and some underspread at the shortest lead times. All but the NCEP ensemble reproduce the diurnal cycle of error at reduced magnitude. The best ensembles have both the smallest error and the largest spread, and therefore the best match of spread to error. The multimodel ensemble has some advantage over the best single-model ensemble, although the gap is slight after the first few days.

The ensemble predicts a distribution of possible outcomes for each specific forecasting case, which the observations should match if that situation could somehow be repeated many independent times. Thus, the mean square spread and error variance should match not just when averaged over all cases, but over any subset of cases which can be identified from the forecasts alone. The remaining plots of Figure 4 consider forecasts where the ensemble mean falls within a relatively narrow range of values. For ensemble mean precipitation around 0.4 mm (middle panel), the mean square spreads are too small and grow too slowly. It is admittedly hard to achieve variances of 2 mm^2 or more on means of 0.4 mm with a variable bounded at zero: this suggests the forecasts may be too Gaussian in this regime, with insufficiently long tails. For mean precipitation around 5 mm (bottom plot), the ensembles become overspread after the first few days. Further analysis (not shown) explored the behaviour of error as a joint function of spread and the ensemble mean forecast. This showed that the gradient of error with respect to spread tends to be less than unity against observations, it declines with lead time, and tends to be closer to ideal for large as opposed to small values of the ensemble mean forecast.

Overall, these results indicate that there are significant deficiencies in the relationship between ensemble spread and forecast error, which only emerge when that relationship is decomposed by variables such as lead time, location, spread, and the ensemble mean forecast. Some of the deficiencies may arise from sampling error rather than systematic overspread or underspread in particular situations. At longer lead times, where the forecasts have little skill, they should essentially predict climatology, and the ensemble mean should always equal the climatological mean. On some occasions, a random draw of a few tens of members from climatology will produce a sample mean significantly above (or below) the true climatological mean. Since precipitation is bounded at zero, such a sample will tend to have a smaller (or, respectively, larger) standard deviation than the true climatology, which is the pattern observed here at long lead times. Kolczynski *et al.* (2011) consider the impact of sampling error on the spread-skill relationship in an idealised context. Whatever the origin, it remains true that these forecasts have deficiencies in mean and spread which could be improved by post-processing: at the extreme of no skill, one would be better replacing a noisy sample by a fixed climatology.

5.1. Spread calibration

The above results suggest there may be benefit to scaling the ensemble spread to better represent forecast errors, provided the calibration is decomposed using the key predictor variables identified in the previous subsection. In principle, such a differentiated spread calibration has the potential to improve both reliability and resolution, since different forecasts can be scaled by different amounts. The basic principle of variance scaling has proved useful for more mainstream atmospheric variables (Flowerdew and Bowler, 2011), and was also considered by Johnson and Swinbank (2009).

A basic calibration scheme was implemented to test this idea. As before, it uses binned data and linear interpolation to avoid more detailed assumptions about the underlying distributions. It again uses twelve months of training data cyclically preceding the target forecast. Mean square spread and error are accumulated in bins defined by the ensemble mean and spread. The number of bins is kept small to limit statistical noise, and the boundaries are chosen based on sample counts and relationship turning points from the diagnostic results (0.4, 3.0 and 7.0 mm for ensemble mean, 2 and 20 mm^2 for squared spread). To further control noise, the final statistics to be used for each gridpoint and bin are aggregated over a rectangular domain of surrounding gridpoints as required to reach at least 200 samples (which would give a 10% error on the variance of independent Gaussian variables). This means that common situations are trained on the most locally-relevant data, whilst rare situations draw data from a wider area to reduce

statistical noise. For each gridpoint and lead time, the desired ensemble spread is linearly interpolated from the training data based on the raw ensemble mean and spread. The perturbations are then scaled by the factor required to produce this spread. This has the advantage of preserving the spatial structure of each ensemble member, and thus the relationships between different locations, variables, or lead times.

The results (not shown) are mixed. The match between spread and error as a function of lead time and ensemble mean forecast is greatly improved, and there is also some improvement in error as a function of spread. Probabilistic scores show a small positive impact on resolution for higher thresholds. There is a large positive impact on reliability at short lead times for all but the highest threshold, but neutral or negative impact at longer lead times. Rank histograms (Hamill and Colucci, 1997) suggest that simple scaling does not result in the correct distribution shape. In addition, the calibration slightly degrades the forecast climatology, because it changes the ensemble spread without compensating changes to the variability of the ensemble mean (Johnson and Bowler, 2009). Finally, there is the question of how to handle zero precipitation with a simple scaling approach, especially when reducing the spread for means above zero, or increasing the spread for means of zero. Ultimately, the framework proposed in the following section proved more conceptually and practically successful.

6. Reliability-based calibration

Both of the calibration methods described above had limited, mixed impact and in particular left significant unreliability with respect to local climatological thresholds. It is often asserted than unreliability can in principle be fixed by relabeling forecast probabilities, given sufficient representative training data, whilst resolution is harder to improve. This section presents a calibration method based on this idea, providing a baseline against which to compare other approaches.

6.1. Design aims

The reliability-based calibration scheme was designed around the following principles:

- The primary measures of forecast performance are the probabilistic scores against climatological thresholds, since these consider the complete forecast PDF and relate to the probabilistic way in which ensemble forecasts should be used.
- The previous calibration schemes attempted to improve probabilistic scores indirectly, by improving the climatology or spread of the underlying member forecasts. The new scheme targets the reliability directly. Within the limits of stationarity and statistical noise, the result should be reliable by construction, since each calibrated probability is based on a past observed event frequency.
- It was noted above that separate spread calibration harms the forecast climatology. Reliability calibration aims to produce a reliable PDF conditioned on the forecast. The sum of all such conditional PDFs is the climatology, so reliability calibration can in principle simultaneously correct probabilities, spread, and climatology.
- One of the motivations for the climatology and spread calibration schemes was that, by adjusting the underlying ensemble members, they preserve spatial, temporal and inter-variable structures. These are important so that the forecasts both look realistic when plotted and integrate realistically when used to drive downstream systems such as hydrological models. At first sight, a calibration of pointwise reliability produces probabilities not ensemble members, and thus loses these important correlations. However, the method includes a step which

reconstructs a set of ensemble members retaining the spatial structure of the raw ensemble, the effectiveness of which is considered in section 6.4.

- The scheme makes very few assumptions about the shape of the forecast or observed distributions. This is particularly useful for precipitation, given the statistical features noted in the Introduction. Whilst assumed distribution shapes can help to manage statistical noise, and may be required for more extreme events than those considered here, there is always the risk that the assumed shape will be unreliable in some situations. The scheme tests the limits of the data-driven approach, and could ultimately be compared to methods based on assumed distributions.
- The scheme tries to avoid making adjustments based on noise rather than signal by aggregating or discarding data as required to achieve a specified minimum number of samples. The aim is to make the calibration as specific and local as possible, consistent with the underlying data volume. Ultimately, the scheme degrades to climatological probabilities when it is unable to be more specific.
- Calibrating rare events requires training data that at least gets close to including them. Section 4 noted that about a year of training data was needed to stabilise the climatology calibration signal. The scheme is thus designed to use a relatively long set of equally-weighted training data (here one year). This is in contrast to schemes such as running bias correction (eg Johnson and Swinbank, 2009), which by tackling a simpler problem are able to use a shorter training period which can potentially capture biases specific to the current weather regime. It is hoped that the reliability-based scheme can achieve some regime-specificity by its more differentiated use of the training data.
- The scheme could in principle be supplied with any training data, including reforecasts (eg Hamill *et al.*, 2008). However, the interest here is in the impact of calibration on multiple ensemble systems and their multimodel combination. Reforecasts are not available for all these systems. In addition, there is no long archive of ukpp observations and, even if there were, changes in processing over the years might limit its usefulness as training data. By contrast, a year of recent historical forecasts and observations might practically be obtained for most forecasting systems. The results suggest the inevitable inhomogeneities are not too harmful, though comparison with calibration based on reforecasts would be required to be sure.
- The scheme aims to make maximum use of the underlying ensemble forecast. In particular, it uses as predictors the full forecast probabilities to exceed a range of thresholds. This is in contrast to many calibration schemes, particularly those based on reforecasts, where predictors are typically just the ensemble mean and possibly the spread (eg Hagedorn *et al.*, 2008; Hamill *et al.*, 2008). The hope is that the full raw probabilities make better use of the detailed atmospheric dynamics and physics included within each ensemble scenario, reducing the amount of work the statistical scheme has to do. Ultimately, the correctness or not of this idea would have to be demonstrated by comparison to calibration schemes based on alternative compromises, such as reforecasts. Simple tests on the reliability-based calibration scheme where training uses a randomly-chosen subset of members do show degraded performance.
- The initial scheme presented here is very much a proof of concept, designed to demonstrate potentially important features whilst remaining reasonably simple and transparent. Various possible extensions and improvements will be mentioned below.

6.2. Implementation

The reliability-based calibration scheme is illustrated in Figure 5. As before, the scheme is trained on a year of data cyclically preceding the target forecast. The training accumulates sample counts, forecast probabilities and observed frequencies for each gridpoint, lead time, threshold, and forecast probability bin. The main implementation uses ten thresholds spaced in powers of two from 0.1 to 51.2 mm. These aim to provide good resolution of low precipitation amounts whilst reducing statistical noise on high precipitation amounts. To further reduce statistical noise and memory usage, only five probability bins are used: three across the main probability range and one each for cases where zero or all members exceed the threshold. (This was motivated by short-lead-time reliability diagrams such as those shown in Figure 5 where the behaviour of these outermost forecasts is discontinuous with the intermediate probabilities, presumably due to underspread; tests showed a small benefit compared to five equally-spaced bins).

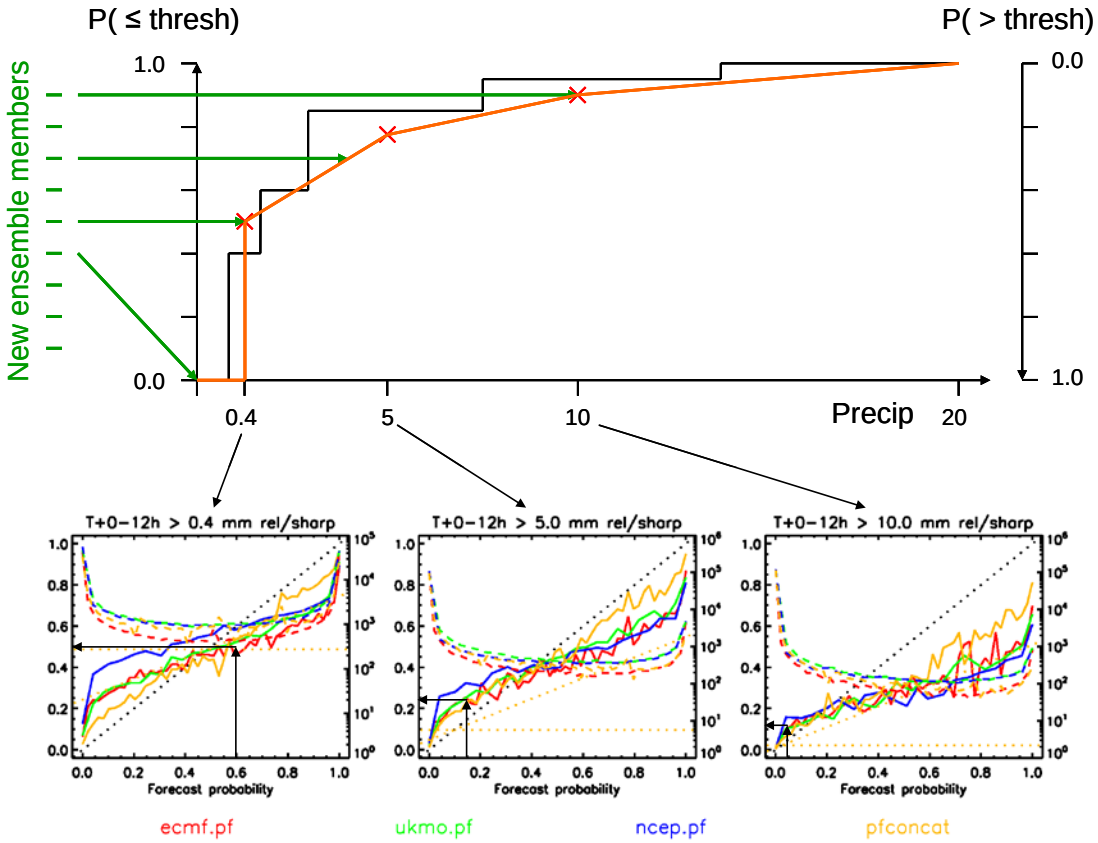

Figure 5. Illustration of the reliability-based calibration method. The raw ensemble members imply a Cumulative Density Function (CDF; black). The training (reliability diagrams) implies a calibrated CDF (orange). New members (green) are assigned to equally divide the probability range, in the same order as the raw ensemble members. Note the opposite sense in which reliability diagrams (probability to exceed a threshold) and CDFs (probability to be less than or equal to a threshold) are traditionally defined.

Before the training data are used, they undergo a spatial aggregation process similar to that described in section 5.1. This is performed separately for each bin, so that common events are calibrated locally whilst rare events pull data from a wider area. Bins with fewer than the minimum number of samples over the domain as a whole are discarded (an improved scheme might combine them with neighbouring bins).

For each gridpoint and lead time, the raw probabilities are established by counting the number of members which exceed each threshold. These are calibrated by linear interpolation/extrapolation from the relevant reliability diagram, or replaced by the observed (approximately climatological) frequency if only one bin exceeded the minimum sample count. If the process stopped at this point, it could produce maps of calibrated probabilities to exceed the predefined thresholds, but not individual forecast values or spatial relationships.

The rest of the process regards these calibrated probabilities as providing a calibrated CDF. Since each threshold is calibrated with a different set of predictors, it is possible for the calibrated probabilities to be non-monotonic as a function of threshold. In practice, the scheme appears to have sufficient control over statistical noise that this effect is small (with a mean probability decrease of about 0.015 across the approximately 5% of cases which are affected). The current implementation sorts the probabilities to force them to be monotonic, though the detailed treatment seems to have negligible impact on probabilistic scores.

The next step identifies a set of ensemble member values to represent the calibrated CDF. These are defined to divide the CDF into bins of equal probability, following the theory behind rank histograms (Hamill and Colucci, 1997). The corresponding precipitation values are obtained by linear interpolation between the calibrated thresholds. To provide a clean distinction between zero and non-zero precipitation, all results below the lowest threshold (0.1mm) are mapped to zero. To close the top of the distribution, the cumulative probability is set to one at twice the highest threshold. A more elaborate scheme might fit an extreme value distribution to close the top of the CDF (eg Ferro, 2007).

The key to preserving spatial, temporal and inter-variable structure is how this set of values is distributed between ensemble members. One can always construct ensemble members by sampling from the calibrated PDF, but this alone would produce spatially noisy fields lacking the correct correlations. Instead, the values are assigned to ensemble members in the same order as the values from the raw ensemble: the member with the locally highest rainfall remains locally highest, but with a calibrated rainfall magnitude. In this way, despite going via the intermediate formulation of probabilities to exceed thresholds, the overall calibration procedure amounts to a reshaping of the local CDF, preserving the order of the ensemble members. A similar ensemble reconstruction step was proposed by Bremnes (2007).

For the multimodel ensemble, the results presented here combine the underlying ensembles before calibration. In principle, one could calibrate the individual ensembles separately and then combine them: this might be desirable if they had significantly different biases, for instance. However, as demonstrated in section 4, conditional biases appear relatively unimportant to the 12-hour accumulations on a 1˚ grid considered in this report. As discussed in section 3.1, simply aggregating members from calibrated ensembles need not produce a correctly calibrated composite. Aggregating before calibration provides more samples to resolve both the raw and calibrated probabilities, and allows the calibration scheme to directly control the statistical properties of the final result.

6.3. Results at gridpoint scale

Against fixed thresholds (not shown), the reliability-based calibration virtually eliminates the reliability penalty without harming resolution, suggesting that the basic formulation is performing as intended and has sufficient control over statistical noise. The interpolation implied by the ensemble reconstruction scheme produces probabilities that are broadly competitive with those obtained when training directly includes the target threshold.

Indeed, in some cases, interpolation between thresholds actually improves resolution, perhaps because it allows members to slide over intermediate values. Whilst the calibration does improve the Brier Skill Score for precipitation to exceed 25mm, it does no better than climatology. Bin aggregation or fitting an extreme value distribution might help to improve this, but the essential problem is the rarity of such events. Experiments using calibration and verification against short-range forecasts, using the whole European domain to provide training data, demonstrate that the scheme can improve 25mm reliability without harming resolution in this case.

The performance of the reliability-based calibration against the more demanding climatological thresholds is illustrated in Figure 6 (for comparison, the raw ensemble performance from Figure 1 is reproduced using pale lines). The calibration strongly reduces the reliability penalty in many cases, particularly for lead times where the forecasts have nonzero resolution and thus some prospect of skill. At longer lead times and the highest thresholds, the reliability appears to hit a nonzero floor, presumably limited by noise and the locality of the calibration data. The reliability penalty rises with lead time for the lower thresholds, and the forecasts still have negative overall BSS for the 99th percentile. One surprising feature is that the 'better' models end up with slightly larger 99th percentile reliability penalty at longer lead times. The cause of this is unclear, but is not seen against fixed thresholds or in other more idealised calibration experiments.

Individual reliability diagrams (not shown) demonstrate that the calibration successfully diagonalises all but the 99th percentile at short lead times, but is unable to sustain this at longer lead times. Against fixed thresholds, by contrast, the reliability diagrams are diagonalised for all lead times where the verification is not dominated by noise. This illustrates the sterner test imposed by the use of climatological thresholds, requiring the calibration to be locally appropriate. The reliability against climatological thresholds might be improved by greater locality of training data, such as might be provided by a longer training period or dynamic aggregation of training bins over raw probabilities in addition to space. Another approach might be to calibrate against climatological thresholds, although the implications for spatial aggregation would need to be carefully considered.

With the exception of 99th percentile reliability at long lead times, the multimodel ensemble remains competitive with or superior to the best single-model ensemble after all have been calibrated, particularly for resolution at short lead times and the 50th percentile. The calibration produces better reliability than the raw multimodel ensemble for the 99th percentile out to about T+5 days, and for the 50th percentile, particularly at long lead times. Otherwise, the much simpler raw multimodel ensemble is competitive with or superior to the best calibrated single model. The (resolution) advantage of the raw multimodel ensemble over the best calibrated single model for light precipitation is consistent with the result reported by Hamill (2012).

Tests comparing local calibration to uniform training from the whole ukpp domain show the former to be slightly beneficial for all scores except 99th percentile reliability. This includes small resolution improvements arising from improvements in local reliability. The 99th percentile reliability illustrates the fact that more samples are needed to accurately place more extreme quantiles, and performance might be improved by changing the aggregation criterion from a simple fixed sample count to a formula which recognised this effect. However, there is some evidence of a tradeoff between resolution and reliability: more aggregation tends to slightly improve the latter at the expense of the former.

Figure 6 does show some small declines in resolution at longer lead times, with similar impact on ROC area (not shown). This appears to arise when the calibration correctly

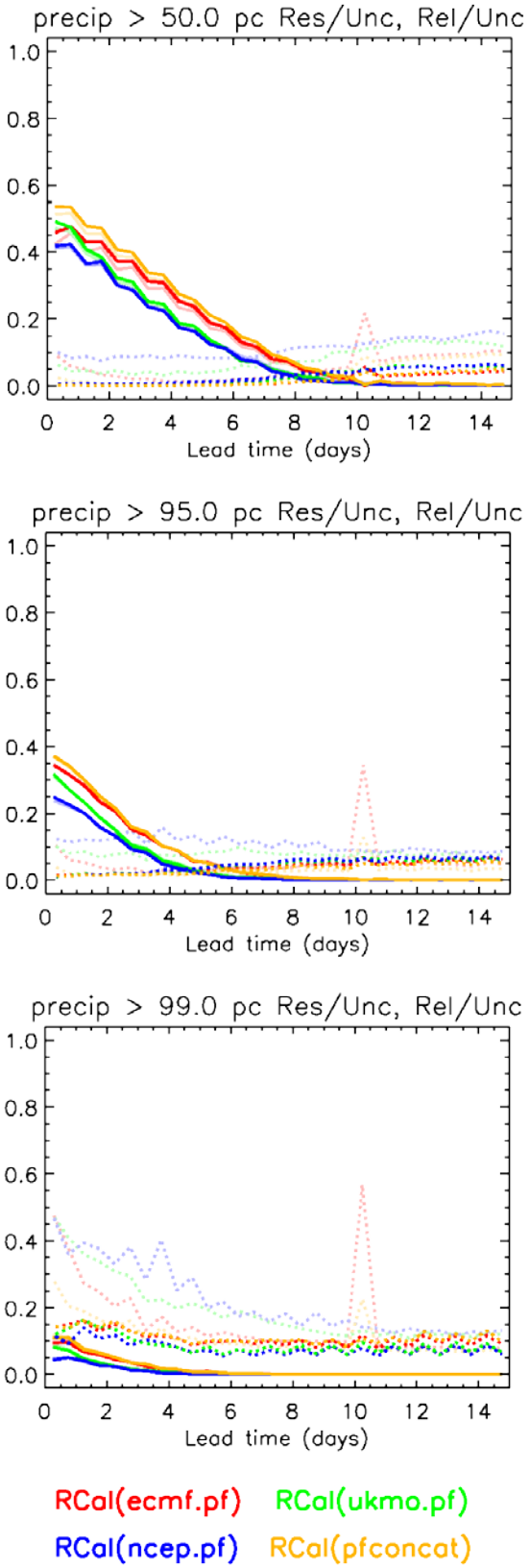

Figure 6. Performance of the single and multi-ensemble forecasts after reliability-based calibration. For comparison, the uncalibrated results from Figure 1 are reproduced using faint lines.

reduces forecast probabilities and then re-quantises these onto a smaller number of ensemble members. This is an unavoidable consequence of choosing to represent the calibrated probability with a finite number of ensemble members.

The reliability-based calibration has mixed impact on the climatology diagnostics introduced in section 4. The 50th percentile is improved to an extent competitive with the direct calibration of section 4.1. This may be related to the strong improvement in 50th percentile resolution for short lead times and the systems with more members, which is similar to that seen for direct climatology calibration. For higher quantiles, the draw towards observations is less convincing, and some results drift with lead time as the forecasts become more uncertain. One issue appears to be the ability to resolve more extreme probabilities with a finite number of members, particularly at long lead times where the ensemble effectively has to represent the whole climatology. This idea is reinforced by the fact that the systems with more members generally perform better, and this performance degrades when the forecasts (but not the training) are restricted to 20 members throughout. In principle, more correct climatology might be obtained by distributing each member randomly within its assigned quantile range, rather than always using the same fixed quantile values. However, such noise might harm more important skill measures. Combining rather than discarding bins that have too few samples may also help to improve the accuracy of climatological probabilities.

The diagnostics introduced in section 5 show the reliability-based calibration scheme drawing rms spread closer to rms error, although not as much as the direct spread calibration described in section 5.1. Reliability-based calibration also helps to make error as a function of spread more diagonal, perhaps even better than direct spread calibration. These results seem to support the idea that correct spread needs to be obtained as a consequence of a more complete calibration, rather than as the sole means to achieve that calibration.

Rank histograms (not shown) are much

flatter after reliability-based calibration than before, and demonstrate that the calibration successfully homogenises the multimodel ensemble at short lead times.

6.4. Impact of spatial structure

One of the aims of the calibration schemes considered in this report is to produce not just point probabilities to exceed predefined thresholds, but ensemble members that retain appropriate spatial, temporal and inter-variable structure. Whereas authors such as Berrocal *et al.* (2008) aim to model correlations statistically, the schemes considered in this report rely on the raw ensemble. No attempt is made to calibrate towards actual correlations, but equally the raw ensemble could provide case-specific correlations, including in the time dimension.

A simple test of this feature can be performed by calibrating at the grid scale but verifying averages over a larger scale (here the 3x3 region centred on each gridbox). The error variance of this average, for example, is the average of the 3x3 error covariance matrix, incorporating both the error variances of the individual gridboxes and the correlation between them. A similar verification technique is used by Berrocal *et al.* (2008).

The results are shown in Figure 7. Calibration at the grid scale (green) performs almost as well as direct calibration at 3x3 scale (yellow). Both have similar resolution and superior reliability to the raw forecasts (red). Assigning the calibrated quantiles to random ensemble members (blue) discards the spatial structure of the raw ensemble. This performs much worse than the other methods, with the exception of 99th percentile reliability at longer lead times. In this case, the extra randomness improves the reliability, perhaps indicating excessive correlations in the raw forecasts. Calibration at the grid scale with random quantile assignment would then provide more samples from which to construct the 3x3 average, reducing the floor on the reliability penalty. In practice, this is of limited benefit since all of the forecasts have negative skill for this threshold.

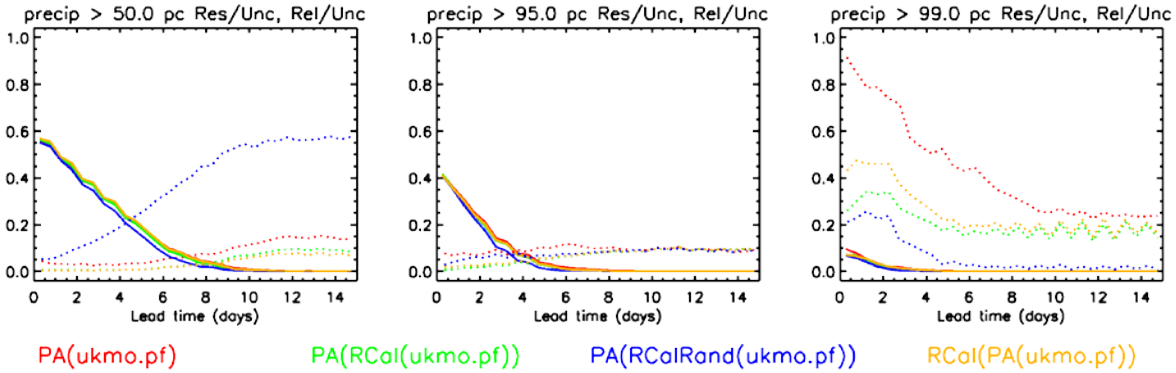

Figure 7. The impact on 3x3-gridbox averages of copying spatial structure from the raw ensemble. The plots show BSS reliability and resolution against climatological thresholds as in Figure 1, except that the climatology is now taken from the observed 3x3-gridbox averages. Colours show the raw ensemble (red), calibration at the grid scale (green), calibration at the grid-scale but with quantiles assigned randomly to ensemble members (blue), and direct calibration of 3x3 averages against the 3x3 average observations (yellow).

7. Discussion

All ensemble systems suffer from systematic errors and can thus potentially be improved by post-processing techniques. This report has considered two such techniques in the context of 12-hour precipitation accumulations on a 1˚ grid: combining forecasts from multiple ensemble systems, and calibrating forecasts based on historic verification data.

Forecast performance was assessed with a variety of techniques. Overall probabilistic performance was verified using reliability and resolution components of the Brier Skill Score against thresholds drawn from spatially- and temporally-local climatology. This approach has a number of advantages such as reducing 'false skill' and equalising the contribution from different regions. Forecast performance was best for the lowest thresholds. Statistical unreliability leads to negative overall skill at progressively earlier lead times the higher the threshold, so that the 99th percentile has negative skill at all lead times.

The forecast climatology was assessed by comparison against equal quantiles of the observed climatology. Some small systematic errors were found, but overall the forecast and observed climatologies were quite similar for the scales considered here.

The overall scale of predicted uncertainty was assessed by comparing rms spread to the rms error of the ensemble mean forecast. Overall results show good agreement, but important discrepancies emerge when the results are decomposed using other variables. Forecasts with ensemble mean around 0.4mm were systematically underspread, whilst long-range forecasts with ensemble mean around 5mm were systematically overspread. Further details emerged for error as a joint function of spread and ensemble mean.

In all of the verification, the simple aggregation of members from the ECMWF, Met Office and NCEP ensembles was competitive with or superior to the best single-model ensemble. Some of this advantage appears to come from system diversity, and some from having extra members. The greatest advantage was seen at short lead times. These advantages were generally retained after calibrating both the single- and multi-model ensembles. In situations of positive skill, the uncalibrated multimodel ensemble appears competitive with the best calibrated single-model ensemble.

Three calibration methods were tested. All of these were designed to take account of the statistical difficulties associated with precipitation, and produce complete ensemble members with spatial, temporal and inter-variable structure. Mapping quantiles from forecast to observed climatology was effective at correcting the climatology, but does not address the key problem of forecast uncertainty. Perturbation scaling was effective at matching ensemble spread to error, but harmed forecast climatology and failed to produce forecast PDFs with the correct shape.

The third scheme directly targets statistical reliability, simultaneously calibrating climatology, spread and probabilities. It virtually eliminates unreliability against fixed thresholds. Against climatological thresholds, the method also provides good improvements, but is unable to diagonalise reliability diagrams at long lead times and high thresholds, so that the overall 99th percentile BSS remains negative. A number of possible solutions were suggested, including more training data, more members to better resolve the climatology of rare events, improvements to the aggregation procedures, calibration against climatological thresholds, and fitting extreme value distributions.

The reliability-based calibration method includes a number of steps which could be performed in different ways. The calibrated probabilities used here were trained on the raw probabilities to exceed the same threshold, but the observed event frequency could be binned according to other variables such as forecasts at lower threshold. This might help to improve resolution scores, although a balance would need to be struck with the statistical noise implied by more finely-divided training data. The ensemble reconstruction approach could be applied to probabilities calibrated by alternative methods, such as logistic regression (Hamill *et al.*, 2008; Wilks, 2009).

Whilst the methods used in this report have been motivated by the statistical features of precipitation, they are generic and could be applied to other variables. Indeed, one of the attractions of the ensemble reconstruction approach is the suggestion that it might be able to produce self-consistent spatial, temporal, multi-variable forecasts. Future work at the Met Office will focus on extending the analysis to other variables and over a wider domain, which should improve the ability to calibrate and verify more rare events. It would also be interesting to see how similar methods performed on finer-scale forecasts.

Acknowledgements

The author thanks Stephen Moseley for information on the ukpp dataset, and Neill Bowler for some useful comments on an early version of this report.

References

Applequist S, Gahrs GE, Pfeffer RL. 2002. Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Weather and Forecasting* **17**: 783–799

Atger F. 2001. Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes in Geophysics* **8**: 401–417.

Berrocal VJ, Raftery AE, Gneiting T. 2008. Probabilistic quantitative precipitation forecasting using a two-stage spatial model. *Univ. Washington Dept. of Stats. Tech. Report* 532.

Bougeault P *et al.* 2010. The THORPEX Grand Global Ensemble (TIGGE). *Bull. Amer. Meteorol. Soc.* **91**: 1059–1072.

Bremnes JB. 2007. Improved calibration of precipitation forecasts using ensemble techniques. *met.no report* 04/02007.

Coelho CAS, Stephenson DB, Doblas-Reyes FJ, Balmaseda M, Guetter A, van Oldenborgh GJ. 2006. A Bayesian approach for multi-model downscaling: Seasonal forecasting of regional rainfall and river flows in South America. *Met. Apps.* **13**: 73–82.

Ferro CAT. 2007. A probability model for verifying deterministic forecasts of extreme events. *Weather and Forecasting* **22**: 1089–1100.

Flowerdew J, Bowler N. 2011. Improving the use of observations to calibrate ensemble spread. *Q. J. R. Meteorol. Soc.* **137**: 467–482.

Fraley C, Raftery AE, Gneiting T. 2010. Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian Model Averaging. *Mon. Weather Rev.* **138**: 190–202.

Hagedorn R, Buizza R, Hamill TM, Leutbecher M, Palmer TN. 2012. Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Q. J. R. Meteorol. Soc.* e-view.

Hagedorn R, Hamill TM, Whitaker JS. 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-metre temperatures. *Mon. Weather Rev.* **136**: 2608–2619.

Hamill TM, Colucci SJ. 1997. Verification of Eta-RSM short-range ensemble forecasts. *Mon. Weather Rev.* **125**: 1312–1327.

Hamill TM, Hagedorn R, Whitaker JS. 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Weather Rev.* **136**: 2620–2632.

Hamill TM, Juras J. 2006. Measuring forecast skill: is it real or is it the varying climatology? *Q. J. R. Meteorol. Soc.* **132**: 2905–2923.

Hamill TM. 2012. Verification of TIGGE multi-model and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Weather Rev.* **140**: 2232–2252.

Johnson C, Bowler N. 2009. On the reliability and calibration of ensemble forecasts. *Mon. Weather Rev.* **137**: 1717–1720.

Johnson C, Swinbank R. 2009. Medium-range multimodel ensemble combination and calibration. *Q. J. R. Meteorol. Soc.* **135**: 777–794.

Kolczynski WC Jr, Stauffer DR, Haupt SE, Altman NS, Deng A. 2011. Investigation of ensemble variance as a measure of true forecast variance. *Mon. Weather Rev.* **139**: 3954–3963.

Park Y-Y, Buizza R, Leutbecher M. 2008. TIGGE: Preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc.*, **134**: 2029–2050.

Primo C, Ferro CAT, Jolliffe IT, Stephenson DB. 2009. Calibration of probabilistic forecasts of binary events. *Mon. Weather Rev.* **137**: 1142–1149.

Stensrud DJ, Yussouf N. 2007. Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Weather and Forecasting* **22**: 3–17.

Wang X, Bishop CH. 2003. A comparison of breeding and ensemble transform Kalman Filter ensemble forecast systems. *J. Atmos. Sci.* **60**: 1140–1158.

Wilks DS. 2006. *Statistical methods in the atmospheric sciences. Second Edition.* Elsevier: Amsterdam.

Wilks DS. 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. *Met. Apps.* **16**: 361–368.

Met Office
FitzRoy Road, Exeter
Devon EX1 3PB
United Kingdom

Tel: 0870 900 0100
Fax: 0870 900 5050
enquiries@metoffice.gov.uk
www.metoffice.gov.uk