Numerical Weather Prediction

The MOGREPS short-range ensemble prediction system: Verification report

Trial Performance of MOGREPS January 2006 - March 2007

Forecasting Research Technical Report No. 503

Neill Bowler, Marie Dando, Sarah Beare & Ken Mylne

email: nwp_publications@metoffice.gov.uk

The MOGREPS short-range ensemble prediction system: Verification report

Trial Performance of MOGREPS January 2006 - March 2007

Neill Bowler, Marie Dando, Sarah Beare & Ken Mylne

1. Introduction

MOGREPS (Met Office Global and Regional Ensemble Prediction System) is an ensemble prediction system (EPS) designed for short-range forecasting to provide a capability for assessment of uncertainty and the generation of probability forecast products over the time-scale of 1-2 days. It is designed to complement the ECMWF EPS which provides the same service for medium-range forecasts over lead-times of 3-15 days. MOGREPS grew out of an initial feasibility study in 2002-03 which considered the requirements for a short-range EPS and designed the framework for a LAMEPS (Limited Area Model EPS). The aim of the project was to provide a near-mesoscale resolution ensemble to address the uncertainties which are important to short-range forecasting, with a particular emphasis on more severe or extreme events which have a high impact on many Met Office customers and which often require the higher resolution of a regional model to be adequately represented. The North Atlantic and European (NAE) version of the Unified Model (UM) was chosen for the regional ensemble to provide coverage of the main region for development of weather systems affecting the UK and Europe.

After a period of research on perturbation methods, the LAMEPS Implementation Project was initiated in 2004 to implement the developing system in the Met Office operational suite in time for a full trial beginning in September 2005. The purpose of the trial was to assess the quality of the NAE ensemble for real-time forecasting, and in particular to assess whether it was able to offer improved uncertainty and probabilistic information for the short-range than was already available from the ECMWF EPS. It was considered essential to run the trial over a substantial period of at least a year to produce valid statistical samples for probabilistic forecasts. The year-long trial was formally completed in September 2006 and a preliminary verification report was completed and presented as MOSAC Paper 11.5 (2006). After the trial MOGREPS continued to run and this full verification report, which is a Key Deliverable of the Met R&D Programme, draws on data up to March 2007.

As the name implies, MOGREPS provides both a global and a regional ensemble capability, but the main interest in MOGREPS is in performance of the higher resolution NAE ensemble for high impact events. The global ensemble exists mainly to provide the lateral boundary conditions for the regional ensemble, so the emphasis in this report is mostly on the regional NAE ensemble. Some verification of the global ensemble was provided in the preliminary report in MOSAC Paper 11.5 (2006), and showed that the ensemble is performing well. Aside from its original purpose of supporting the regional ensemble, global MOGREPS is also now run routinely out to 15 days as part of the Met Office contribution to the WMO research programme, THORPEX. Under the Met Office THORPEX project, Watkin *et al* (2007) reported that the 15-day MOGREPS was competitive with the ECMWF EPS

although overall slightly less skilful. Since the ECMWF EPS is a much more mature system this is a notable achievement and offers encouragement that the global ensemble should provide good support for the regional ensemble.

As noted above the high impact events of interest to customers are often also severe or extreme weather events. In practise probabilistic verification of extreme events is not possible because sample sizes are so small, but emphasis in the report will be given wherever possible to performance for extreme events.

The verification results in this report are split into three sections, divided by the systems which have been used to analyse the performance. This choice is motivated by the fact that each of the verification systems is attempting to provide different information. Section 2 of the report provides a brief description of MOGREPS, while section 3 introduces the main verification diagnostic tools which are used in subsequent sections. Section 4 uses the station-based verification system and looks at the categorical verification of the ensemble forecasts. This assessment can be made on the ECMWF ensemble in addition to the global and NAE MOGREPS ensembles, allowing a comparison with pre-existing capability. Section 5 uses the station-based verification system, but looks at the ability of the spread of the NAE ensemble to predict the skill of the ensemble mean forecast. Section 6 uses the area-based verification system, and looks at the overall performance of the NAE ensemble in more detail, using the maximum possible number of verifying observations to focus as much as possible on more extreme (and hence generally higher impact) events. Section 7 presents some preliminary results for verification of tropical cyclones in the 15-day global version of MOGREPS. Some conclusions are drawn in section 8.

2. Description of MOGREPS

MOGREPS consists of two ensembles, one global and one regional using a higher-resolution LAM covering the North Atlantic and Europe (the NAE ensemble). The NAE domain is shown in figure 2.1. The resolution of both ensemble systems has been chosen to be approximately half the resolution

Figure 2.1: Map showing the domain of the Met Office's models. The NAE model covers much of the north Atlantic and Europe, and is shown in a darker shading here. (Also shown is the UK meso-scale model area in a lighter colour.)

of the corresponding deterministic models in the Met Office operational suite. The global ensemble is run at N144 resolution (approximately 90km in the mid-latitudes) with 38 vertical levels, which compares with N320 resolution and 50 levels for the deterministic global model. The NAE ensemble is run at 24km resolution and 38 vertical levels, which compares with 12km resolution and 38 vertical levels for the deterministic model. Both ensembles are run with 24 members (unperturbed control plus 23 perturbed members).

The run times of the ensembles are offset by 6 hours to distribute the computing burden more evenly through the day. The global ensemble runs at 00 and 12UTC and the NAE ensemble at 06 and 18UTC. Thus the NAE ensemble takes its LBCs and initial condition perturbations from a 6-hour forecast of the global ensemble. For a short-range ensemble to be useful in an operational framework it is critical that the forecasts are available as early as possible after data-time, and this arrangement allows the NAE ensemble to be run immediately that the new NAE analysis becomes available (rather than having to wait for the global ensemble to run first). Global MOGREPS forecasts are run to 72 hours ahead and NAE MOGREPS to 36 hours for results presented in this report, although the latter has more recently been extended to 54 hours to fully encompass Day 2.

Initial condition perturbations for MOGREPS are generated using an Ensemble Transform Kalman Filter (ETKF) (Bishop *et al.,* 2001). This may be thought of as a natural generalisation of the error breeding method (Toth and Kalnay, 1993), in which the perturbations determined for each cycle are a linear combination of the forecast perturbations from the previous cycle. This mixing allows the perturbations to be orthogonalised, and has been seen to lead to improved performance over error breeding (Wang and Bishop, 2003). The ETKF calculates the new set of perturbations from the forecast perturbations using a transform matrix. The perturbations are rescaled to ensure they are consistent with observation errors in 4D-Var using a variable inflation factor. The ETKF provides a set of perturbations which are added to the Met Office 4D-Var analysis to provide the initial states for ensemble members.

In the current implementation of MOGREPS all initial condition perturbations are calculated in the global ensemble. The regional ensemble takes the 6h forecast perturbations from the global ensemble and adds them to the latest NAE analysis to provide the initial conditions. It is planned to introduce initial perturbations calculated within the NAE ensemble in the near future, but this will not affect any of the results presented in this report.

Uncertainty due to model error is addressed in MOGREPS through stochastic perturbations to the model, mainly to the parameterised model physics. Three schemes are implemented in the global ensemble, but only the Random Parameters (RP) scheme is currently employed in the NAE ensemble. The RP scheme targets uncertainty due to the choice of tuneable parameters in a number of parameterisation schemes in the UM. Parameter values are based on empirical results but are subject to uncertainty. In the standard UM implementation parameter values are held constant at a chosen value, but under the RP scheme they are allowed to vary smoothly from time-step to time-step in a random fashion within the error bounds of the empirical

estimation of the parameter values. The global ensemble also employs a Stochastic Convective Vorticity (SCV) scheme and a Stochastic Kinetic Energy Backscatter (SKEB) scheme. The SCV scheme addresses uncertainty due to the impact of organised convection and is most appropriate in the tropics and for lower resolution models, so is not employed in the NAE. The current SKEB scheme was found to be ineffective in the NAE model, so was not implemented. A new SKEB2 scheme is under development and is expected to be implemented in the NAE in the future.

It is important to emphasise that uncertainties in initial condition and model are not independent, as the impact of stochastic physics perturbations is propagated into the initial condition perturbations of the next cycle through the ETKF.

MOGREPS is supported by a comprehensive display system which allows forecasters to view a wide range of forecast fields and products such as probabilities. MOGREPS forecasts are also stored in the FSSSI database for a large set of sites around the UK and Europe, and a few elsewhere around the world from the global ensemble only.

2.1 Upgrades to MOGREPS

MOGREPS was initially implemented in August 2005. Since then there have been a small number of significant upgrades which could have had some impact on the verification performance. Upgrades affecting the NAE ensemble are as follows:

Date	**Description of Change**	**Expected impact on NAE Ensemble**
Oct 2005	Upgrade global EPS to UM6.1	Minor
June 2006	Local ETKF	Significant reduction in ensemble spread over mid-latitudes – improvement.
“	SKEB introduced to global EPS	Minor

The introduction of the local ETKF in June 2006, which is believed to have significantly reduced the over-spread of the NAE ensemble over Europe, is the only change believed to have had a major impact on performance during the verification period. Many of the verification results presented are based only on the later part of the trial period after this change (including the period after the formal end of the trial in September 2006) so the results are expected to largely reflect the performance of the current system at the end of March 2007. However note that the results from the Area-based Verification system presented in section 6 include data from before this change in order to capture as large a sample as possible for rare events, which may have some impact on the results reported in that section.

3. Description of verification measures

A number of verification methods are used in this report, and a brief outline of these will be given here. A more detailed description of almost all of the verification methods used in the world today can be found on Beth Ebert's web pages at
http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

3.1 Brier skill score

The Brier skill score (BSS), reliability score and resolution score are used to assess forecast quality. These scores derive from the well-known Brier score (Wilks, 1995, p262)

$$BS = \frac{1}{N}\sum_{i=1}^{I} N_i \left(p_i - \overline{o}_i \right)^2 - \frac{1}{N}\sum_{i=1}^{I} N_i \left(\overline{o}_i - \overline{o} \right)^2 + \overline{o}(1-\overline{o}) \qquad (3.1)$$

$\overline{o}$ is the sample climatological value of the probability of the event, that is

$$\overline{o} = \frac{1}{N}\sum_{k=1}^{N} o_k \qquad (3.2)$$

N is the total number of forecasts, and o_k is the probability that the event was observed (either zero or one). $\overline{o}_i$ is the frequency with which the event was observed, when the forecast probability fell into category i. N_i is the number of forecasts of the event in the same category, I is the number of categories and p_i is the forecast probability. The BSS is derived from the Brier score as

$$B_{skill} = 1 - \frac{BS}{BS_{ref}} \qquad (3.3)$$

The in-sample climatology is usually taken as the reference forecast, which gives

$$B_{skill} = B_{resolution} + B_{reliability} \qquad (3.4)$$

where

$$B_{resolution} = \frac{1}{\overline{o}(1-\overline{o})N}\sum_{i=1}^{I} N_i \left(\overline{o}_i - \overline{o} \right)^2 \qquad (3.5)$$

measures the propensity of the forecast to give high or low values of the probability (as opposed to forecasting climatological values) and

$$B_{reliability} = \frac{1}{\overline{o}(1-\overline{o})N}\sum_{i=1}^{I} N_i \left(p_i - \overline{o}_i \right)^2 \qquad (3.6)$$

is a penalty function for departure of the forecast from perfect reliability. The Brier skill score is a useful measure of the skill of a forecast, since it cannot be hedged, which is one the reasons it is the skill score used to define the probability of precipitation KPT (Met Office Key Performance Target).

3.2 Attributes (reliability) diagrams

A natural companion to the Brier skill score is an attributes diagram. This consists of the standard reliability diagram and also includes reference lines related to the algebraic decomposition of the Brier score and Brier Skill score (Wilks, 1995).

The reliability curve shows the frequency with which an event was observed to occur, plotted against the frequency with which it was forecast to occur. The curve for a perfectly reliable ensemble will lie along the diagonal line with gradient 1:1. The histogram is known as the sharpness or refinement distribution, which provides information regarding the frequency of use of each probability value.

The horizontal dashed line marked 'no resolution' in section 6 is related to the resolution term in the decomposition of the Brier score (see above). See figure 6.1 on page 33 for an example. The no resolution line represents the sample climatology of the observations - points falling on this line indicate forecasts that are unable to resolve occasions where the event is more or less likely than the overall climatological probability. The larger the distance between the reliability curve and the no resolution line the greater the resolution of the forecast. The resolution term of the Brier score consists of the sum of the weighted average of the square of the vertical distance between the points on the reliability curve and the no resolution line. The vertical dashed line in section 6 is located at the intersection of the perfect reliability line and the no-resolution line and marks the climatological forecast probability of the event. By definition such a forecast has no resolution and perfect reliability. The dashed diagonal line that sits midway between the perfect reliability line and the no resolution line marks the line of 'no skill'. For points along this line $B_{resolution}$ is equal to $B_{reliability}$, and the forecast has no skill.

3.3 Relative operating characteristic (ROC)

Plots of the relative operating characteristic (Mason, 1999) show the hit rate (H) against the false alarm rate (F) for different confidence levels (such as probability of precipitation greater than 50%). The hit rate and false alarm rate are defined as follows

$$H = a \big/ a+b$$
$$F = c \big/ c+d \qquad (3.7)$$

where a, b, c and d are the standard contingency table values shown in table 3.1. Values at different probability thresholds define a series of points, which are often joined by straight lines. A measure of the skill of a probability forecast is the area under the ROC curve (from (0,0) to (1,1)).

	Event forecast	Event not forecast
Event observed	a (hit)	b (miss)
Event not observed	c (false alarm)	d (correct rejection)

Table 3.1. Contingency table for a categorical forecast.

It has been proposed (Wilson, 2000) that rather than calculate the area under a ROC curve based on a series of straight-line segments that a parametric curve be fitted to the data, and the ROC area be estimated as the area under this curve. This estimate should provide the limiting skill of the ensemble as its size tends towards an infinite number of members. The parametric fitting is taken from a straight line fit when the empirical hit and false alarm rates have been transformed to standard normal deviates. It is this parametric fit which is used in this report.

3.4 Spread and skill

One of the key aims of an ensemble from a forecaster's perspective is to predict the skill of the deterministic forecast. Ideally, when the spread of the ensemble is small then the forecaster can have confidence that the deterministic forecast will be reliable, whereas when the spread is large it is more important to express uncertainty and make allowance for errors. There are two aspects to spread and skill. Firstly the spread of the ensemble should match the error in the ensemble mean forecast on average. This is a common assessment, based on finding the average spread and error of the ensemble forecast for a particular lead time. This type of verification is touched upon in section 6. The second type of assessment looks at whether the spread of the ensemble provides an accurate prediction of the error in the ensemble mean on any given instant. This is assessed by looking at the correlation between the spread and error, and is considered in section 5.

4. Station-based verification – categorical verification and comparison with ECMWF

Verification was performed using data from the station-based verification system. This system takes the ensemble forecast data from the site-specific forecast database, FSSSI, (which has been interpolated to the station location) and verifies this against the observations at each site. Quality control of the observation data is based on the probability of gross error for an observation, as determined by the data-assimilation system.

Data from FSSSI includes MOGREPS global and regional forecasts as well as forecasts from the ECMWF ensemble. The station-based verification is performed against observations at a set of 79 sites in the UK and Europe. For a number of quantities the ECMWF forecasts are verified against UK stations only, and in order to perform a comparison, the MOGREPS data is restricted in the same way. Table 1 lists the quantities that are verified, and for each of these quantities whether the verification is performed against UK only, or UK and European stations. Any results presented for which European stations are available will have a small influence from European stations, but still be mainly based on UK stations. All results presented are for the validity times of 0Z and 12Z combined.

UK stations only (56 sites)			UK and European stations (79 sites)		
Variable	Validity time	Threshold	Variable	Validity time	Threshold
Wind Speed	0Z	Force 5, 6, 7, 8 & 9			
Wind Speed	12Z	Force 6, 8 & 9	Wind speed	12Z	Force 5 & 7
Temperature	0Z	> 5, 10, 15, 20 < 2, -2	Temperature	0Z	< 0, -5
Temperature	12Z	> 20, 25 < 5, 2, 0, -2	Temperature	12Z	> 10, 15
12h accum precip	0Z	> 0.1, 0.5, 1, 5, 10, 20			
12h accum precip	12Z	> 0.1, 1, 20	12h accum precip	12Z	> 0.5, 5, 10

Table 4.1. Quantities for which only UK stations, or UK and European stations are used in the verification.

Reliability tables have been calculated for each day of the verification period, and these were re-sampled to provide confidence intervals. The confidence intervals shown are the 5% and 95%. The forecast probabilities are binned into 10% bins, which allows comparison of ensembles with different numbers of members. Although binning the forecast into 10% bins aids comparison between the ensemble systems, some advantage for the system with more ensemble members (ECWMF in this case) will persist. If, on any given occasion forecasts are not available from any one of the forecasting systems, then the verification is not calculated for any system. This ensures that exactly the same data-set has been used for each of the systems.

4.1 Precipitation performance

Figures 4.1 to 4.4 show the Brier skill score, reliability and resolution for the three ensemble forecasts. The results are for probability forecasts of 12h accumulated precipitation greater than 0.5, 1, 5 and 20 mm respectively. For accumulations of 0.5 mm (figure 4.1) the NAE ensemble has a similar resolution to the ECMWF ensemble, which is better than the global ensemble, though not significantly. The NAE ensemble is significantly more reliable than the global ensemble, which is significantly more reliable than the ECMWF ensemble. Overall, this means that the NAE ensemble is significantly more skilful than the global and ECMWF ensembles, which have approximately equal skill.

The results for higher precipitation thresholds (figures 4.2-4.4) are similar to those for 0.5 mm accumulation. As the threshold moves to higher precipitation values the ECMWF ensemble becomes more reliable. At a threshold of 5mm (figure 4.3) the NAE and ECMWF ensembles perform similarly and both are significantly more skilful than the global ensemble. At a threshold of 20 mm (figure 4.4) the NAE ensemble is less reliable than the other models, and is significantly less skilful than the ECMWF ensemble. This is due to the NAE ensemble substantially over-forecasting the occurrence of heavy rain.

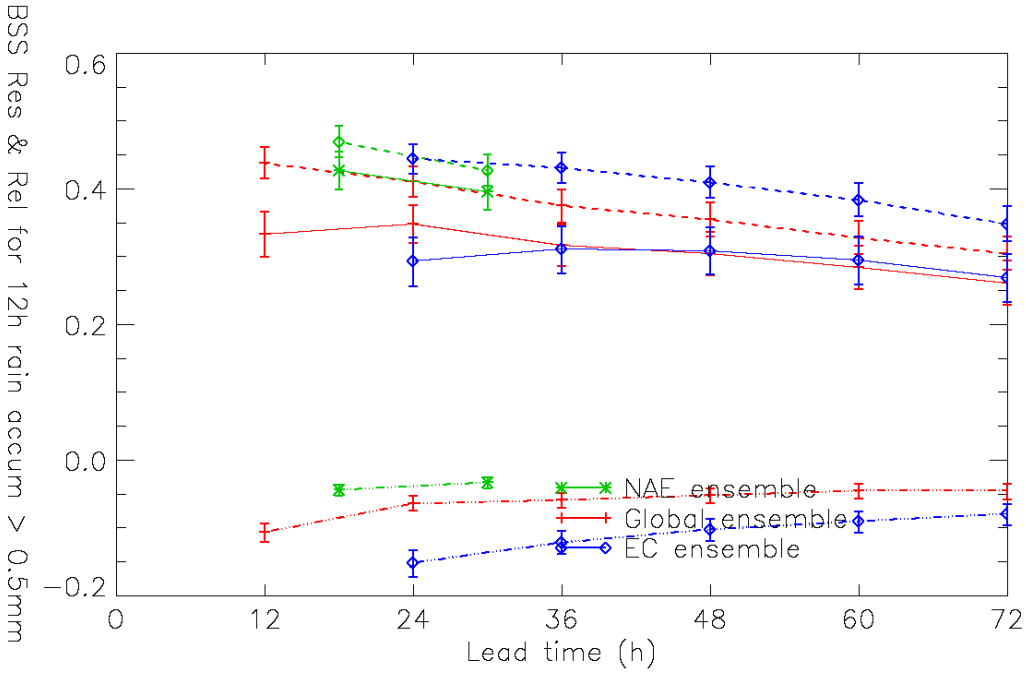


Figure 4.1. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components as defined in equations 3.5 and 3.6 for the NAE, global and ECMWF ensembles for forecasts of 12h accumulated precipitation greater than 0.5 mm. The verification period is 1 July 2006 to 31 March 2007.

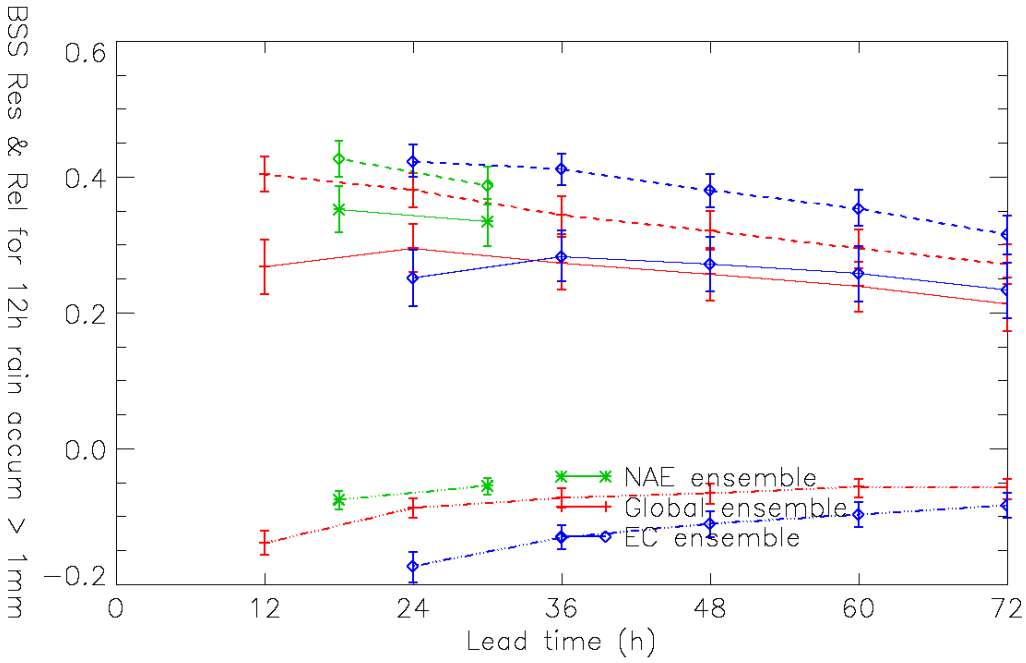


Figure 4.2. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 12h accumulated precipitation greater than 1 mm. The verification period is 1 July 2006 to 31 March 2007.

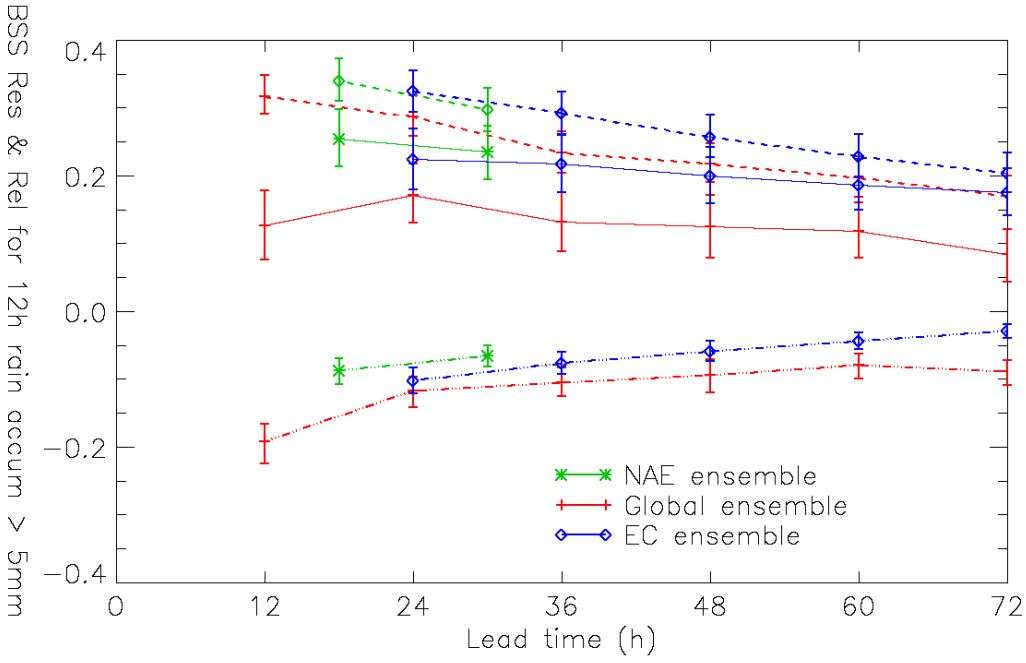


Figure 4.3. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 12h accumulated precipitation greater than 5 mm. The verification period is 1 July 2006 to 31 March 2007.

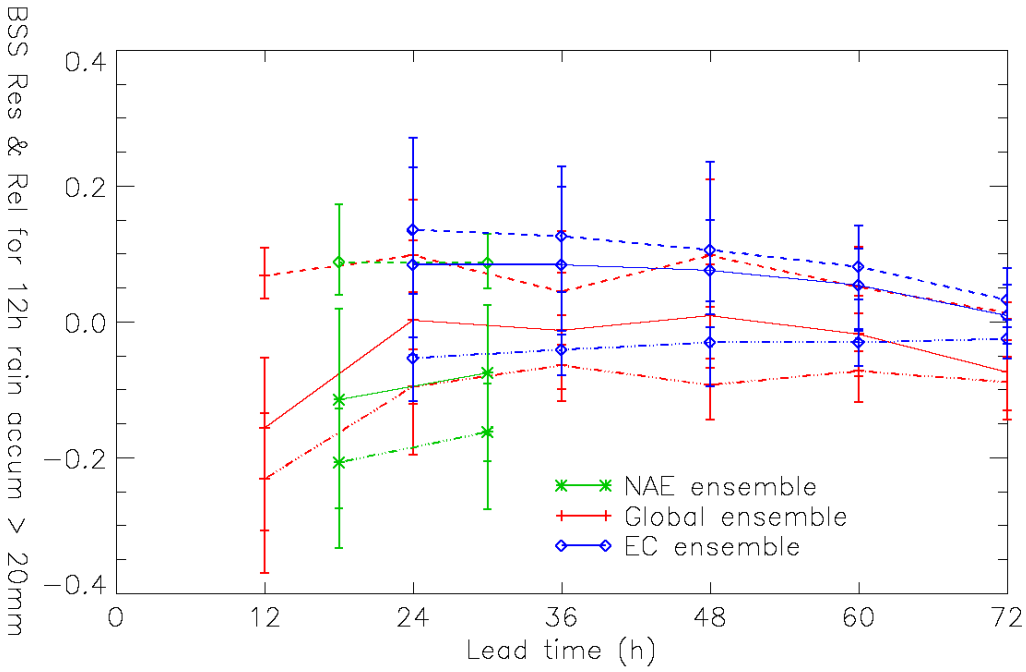


Figure 4.4. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 12h accumulated precipitation greater than 20 mm. The verification period is 1 July 2006 to 31 March 2007.

The reliability and sharpness diagrams for 12h accumulated precipitation greater than 0.5 mm for forecast lead times of T+30 (NAE) and T+36 (global and ECMWF) are shown in figure 4.5. All three ensemble systems are over-forecasting the occurrence of light precipitation – the NAE and global ensembles having similar levels of bias, and the ECMWF having a worse bias. Since the verification is performed against a series of stations, some over-forecasting of light rain may be expected (since the precipitation would need to be down-scaled to a specific site). Thus, the bias seen in the global ensemble would be expected to be greater than for the NAE ensemble (since it is lower resolution) and the ECMWF ensemble would be worst affected (since it is transferred to the Met Office at 1.5 degree resolution). In fact, results from the area-based verification system against Nimrod analyses (see figure 6.1) indicate that the NAE ensemble is not over-forecasting light rain, when the observations are averaged over grid-boxes of similar size to the model grid. The resolution is unlikely to be the whole story, with some of the poor reliability of the ECMWF ensemble most likely due to modelling problems.

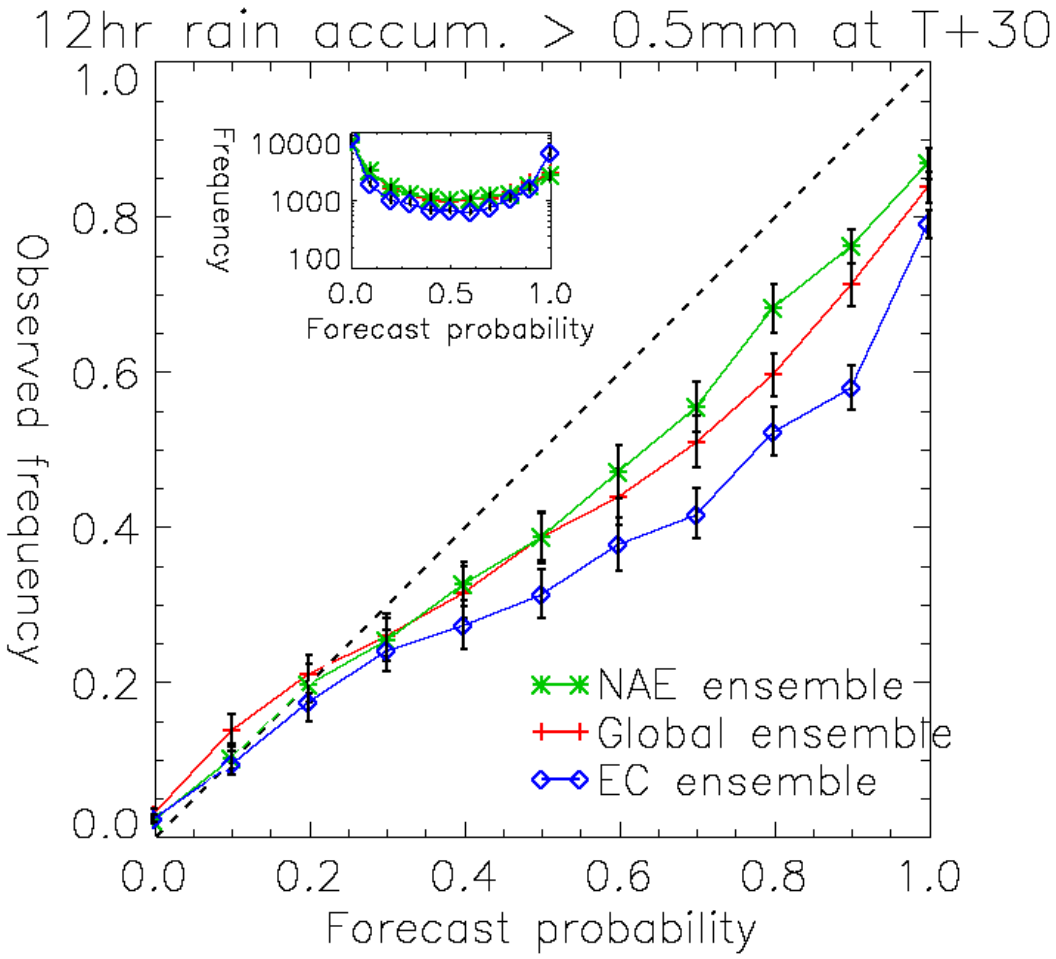


Figure 4.5. Reliability and sharpness diagrams for forecasts of 12h accumulated precipitation greater than 0.5 mm over the months of July 2006 to March 2007. The lead time of the forecasts is T+36 (global and ECMWF) and T+30 (NAE).

4.2 Precipitation results month-by-month.

The Brier skill score for the three ensembles are shown in figure 4.6, month-by-month for forecasts of 12h accumulated precipitation greater than 0.5 mm. This shows that the lack of reliability of the ECMWF ensemble for low rain accumulations is most prevalent during the summer months. Figure 4.7 shows the Brier skill score (and its components) for forecasts made between 6 November 2006 and 31 March 2007. For this period, the performance of the ECMWF ensemble is much improved, relative to the other models, with performance better than the global ensemble (though not significantly). The NAE ensemble still performs better than the ECMWF ensemble, though not significantly. The relative performance of the models changes less with the changing season for higher precipitation thresholds.

Thus, we may conclude that the MOGREPS ensembles are not biased in their forecasting of light rain, but that the ECMWF ensemble is affected by a bias in the Summer.

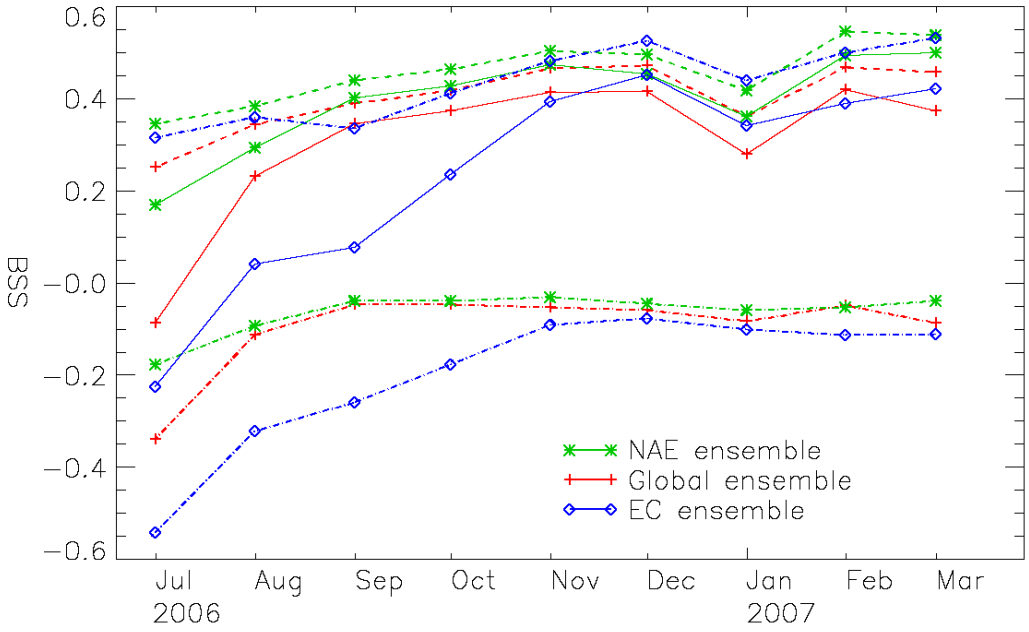


Figure 4.6. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 12h accumulated precipitation greater than 0.5 mm. The verification period is for each month from July 2006 to March 2007.

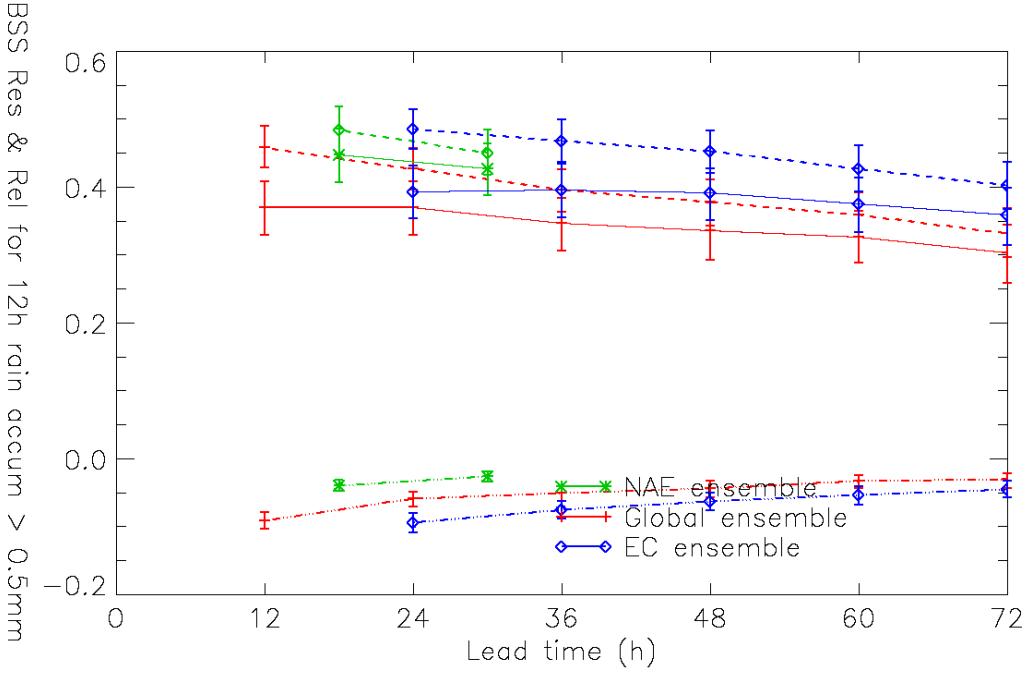


Figure 4.7. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 12h accumulated precipitation greater than 0.5 mm. The verification period is 6 November 2006 to 31 March 2007.

4.3 ROC of precipitation forecasts

Another way to examine the verification results is to look at the relative operating characteristic (ROC). This is essentially a measure of the resolution of the forecast. Figure 4.8 shows the ROC curve for forecasts of 12h accumulated precipitation greater than 0.5 mm for forecast lead times of T+30 (NAE) and T+36 (global and ECMWF). The curves shown are best-fit curves to the data points from the forecasts (Wilson, 2000) with the dashed lines giving confidence intervals. Figure 4.9 shows the area under the ROC curve calculated at various lead times. The ECMWF ensemble has the largest area, follows by the NAE ensemble and then the global ensemble. The ECMWF ensemble is significantly more skilful than the global ensemble, though not than the NAE ensemble.

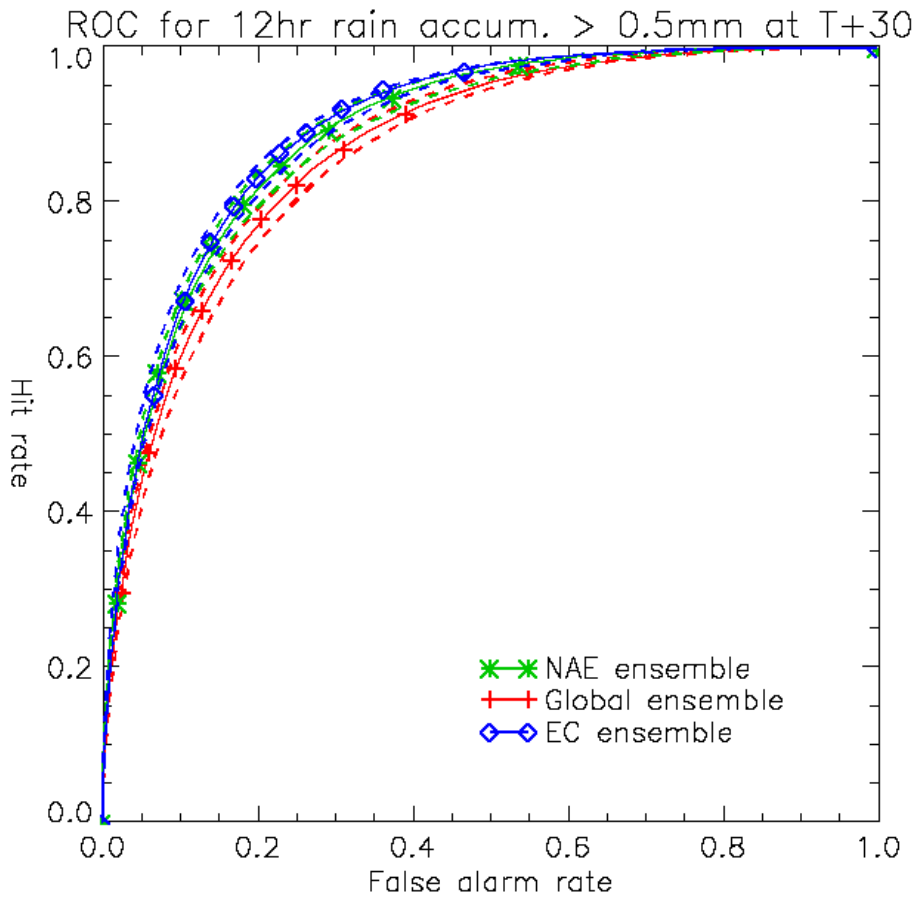


Figure 4.8. ROC for the NAE, global and ECMWF ensembles for forecasts of 12h accumulated precipitation greater than 0.5 mm. The verification period is 1 July 2006 to 31 March 2007.

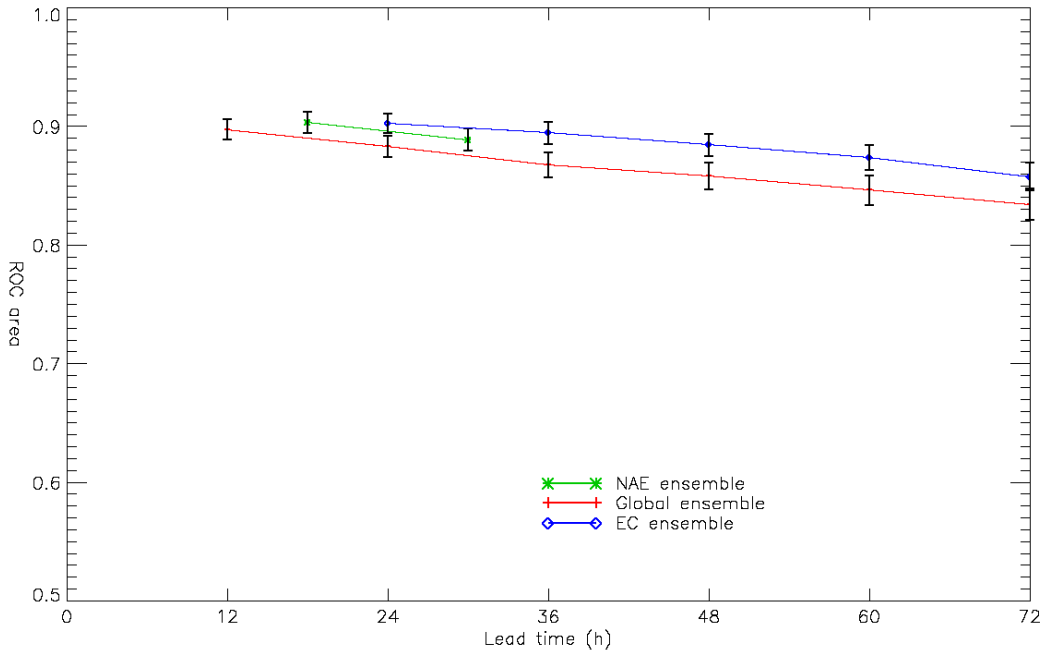


Figure 4.9. Area under ROC curve, calculated from best-fit curve, for the NAE, global and ECMWF ensembles for forecasts of 12h accumulated precipitation greater than 0.5 mm. The verification period is 1 July 2006 to 31 March 2007.

Thus, we conclude that the NAE ensemble performs better, overall, than the other models for forecasts of 12h accumulated precipitation. For higher accumulation thresholds, and at the lower thresholds during winter the ECMWF ensemble is competitive with the NAE ensemble. The global ensemble often performs less well than the other ensembles, though it is competitive over the whole period for low precipitation thresholds. The MOGREPS ensembles have lower ROC values than the ECMWF ensemble, although it should be remembered that ROC does not reflect the better reliability of MOGREPS.

4.4 Temperature verification

There have been a number of problems with the verification of temperature forecasts. Normally, the values of soil moisture used in the forecast model should be derived from the latest analysis. However, before November 2006, the MOGREPS ensemble was using the climatological values for the soil moisture. This problem was corrected on 5 November 2006. The error was particularly noticeable during the summer months when the soil was therefore too moist, resulting in 2m temperature forecasts which were too low. This can be seen in the month-by-month verification chart of the BSS for forecasts of 2m temperature greater than 15^{o}C (see figure 4.10). In addition to this, the FSSSI system moved from receiving the forecast data in degrees Celsius to Kelvin on the 2nd March 2007. Unfortunately, the forecast from the 6Z NAE

ensemble run on this day was processed incorrectly, and temperatures around 270°C were forecast! This was passed through the system, and corrupted the post-processing (KFMOS) system for the NAE ensemble for the rest of the month of March. Therefore, the subsequent verification is performed for the period 6 November 2006 – 28 February 2007. Finally, no results for forecasts of low temperatures will be presented, because there seems to be a problem with the MOGREPS results for lead times greater than one day. It is not known what is causing this problem.

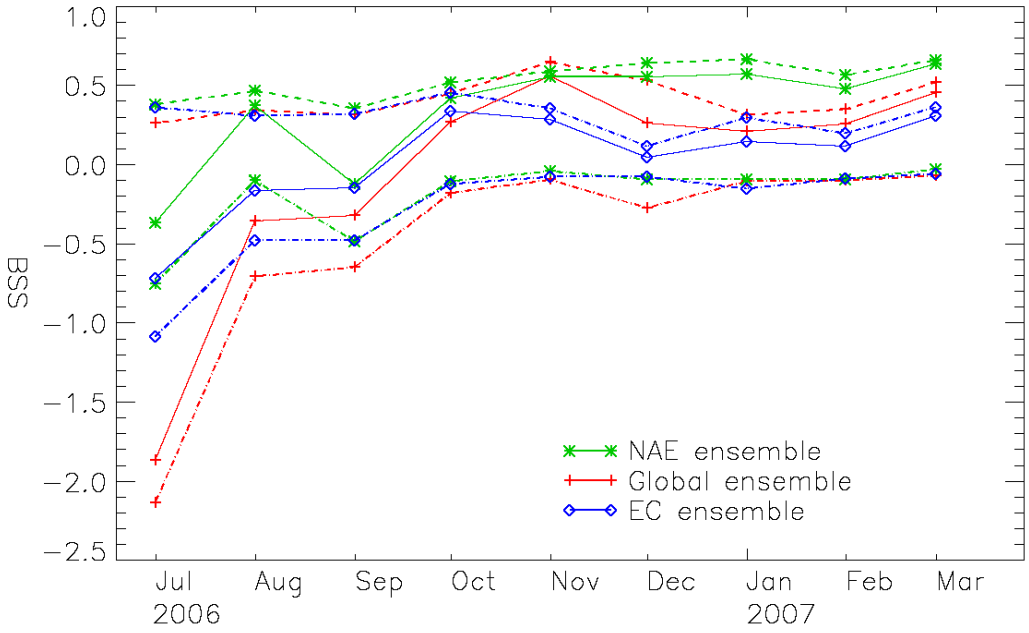


Figure 4.10. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 2m temperature greater than 15°C. The verification period is for each month from July 2006 to March 2007.

The performance of the raw forecasts are shown in figures 4.11-4.12. This shows the verification for forecasts of 2m temperature greater than 10°C and 15°C, respectively. The forecasts of low temperatures would be more appropriate to show for this winter period. For both temperature thresholds the NAE ensemble is the most skilful, and the global ensemble the next most skilful. The differences are significant at the 95% level for all distinctions, except for the difference between the global and ECMWF ensembles for the 10 degree threshold. The reliability and sharpness diagrams for forecasts of 2m temperature greater than 10°C are shown in figure 4.13. The NAE ensemble is closer to the diagonal (perfect reliability) than the other two ensembles, with the ECMWF ensemble possessing a clear bias to under-forecast the temperature.

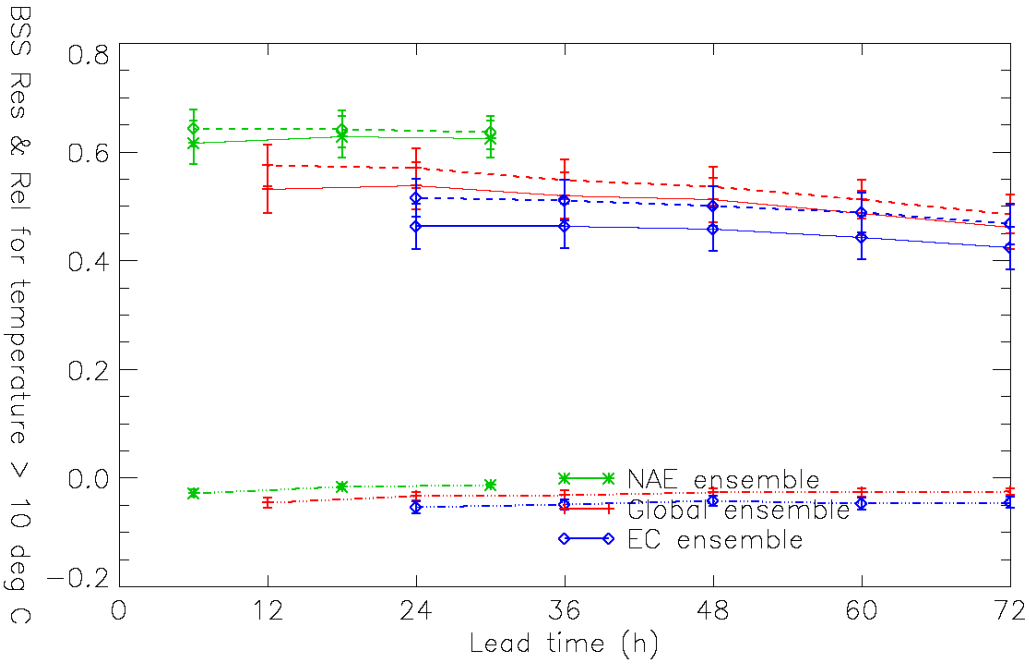


Figure 4.11. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 2m temperature greater than 10°C. The verification period is from 6 November 2006 to 28 February 2007.

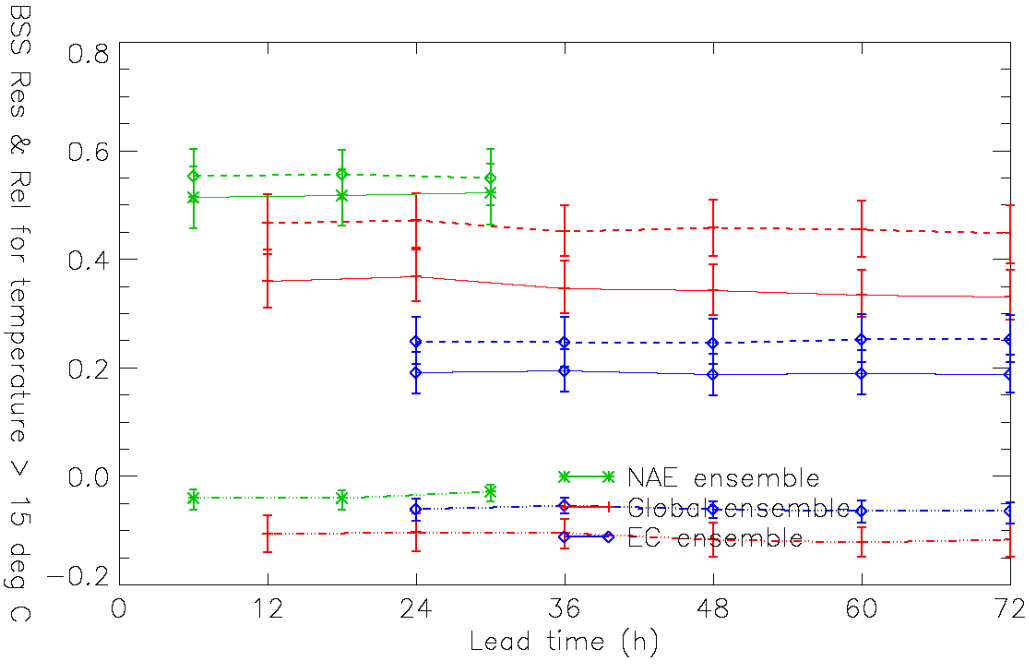


Figure 4.12. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 2m temperature greater than 15°C. The verification period is from 6 November 2006 to 28 February 2007.

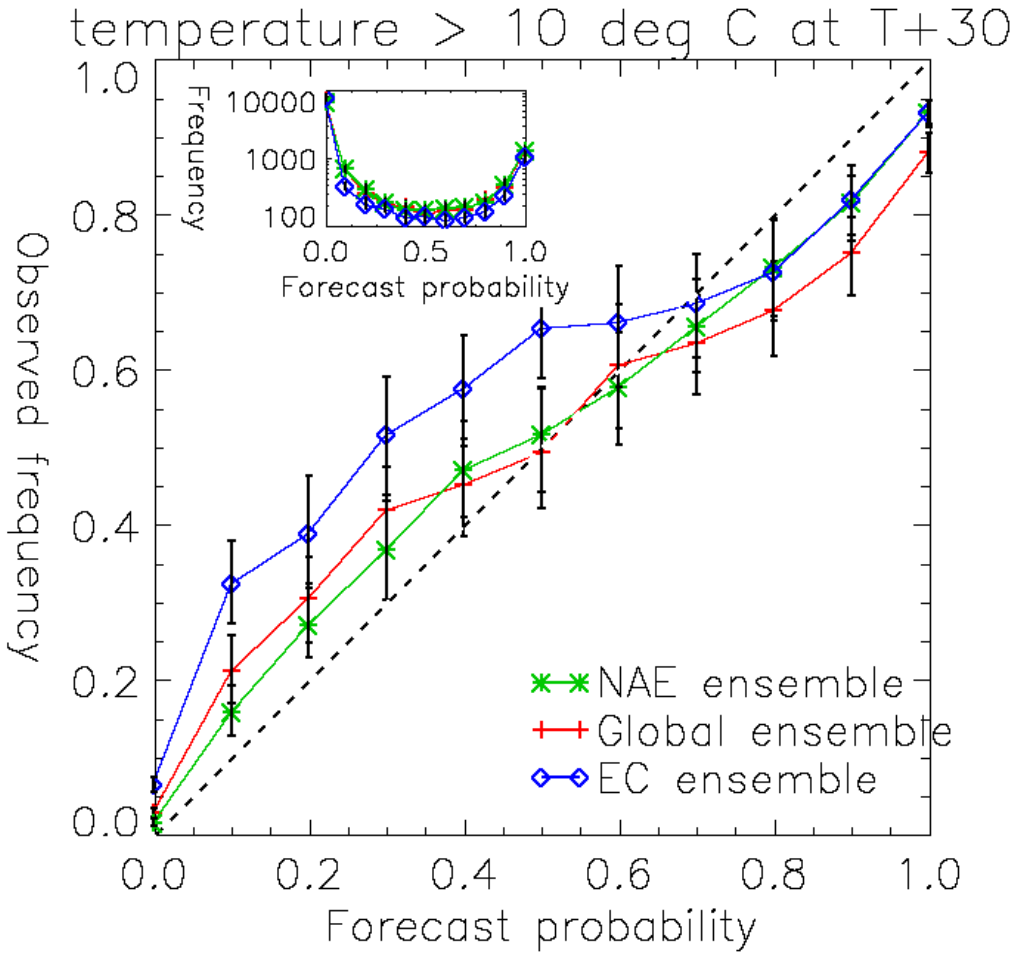


Figure 4.13. Reliability and sharpness diagrams for forecasts of 2m temperature greater than 10°C. The verification period is from 6 November 2006 to 28 February 2007. The lead time of the forecasts is T+36 (global and ECMWF) and T+30 (NAE).

After post-processing via KFMOS, the forecasts from all three ensembles are improved (see figures 4.14 and 4.15). The ECMWF ensemble benefits most from the post-processing, as would be expected. None of the differences between the ensemble systems are significant in this case.

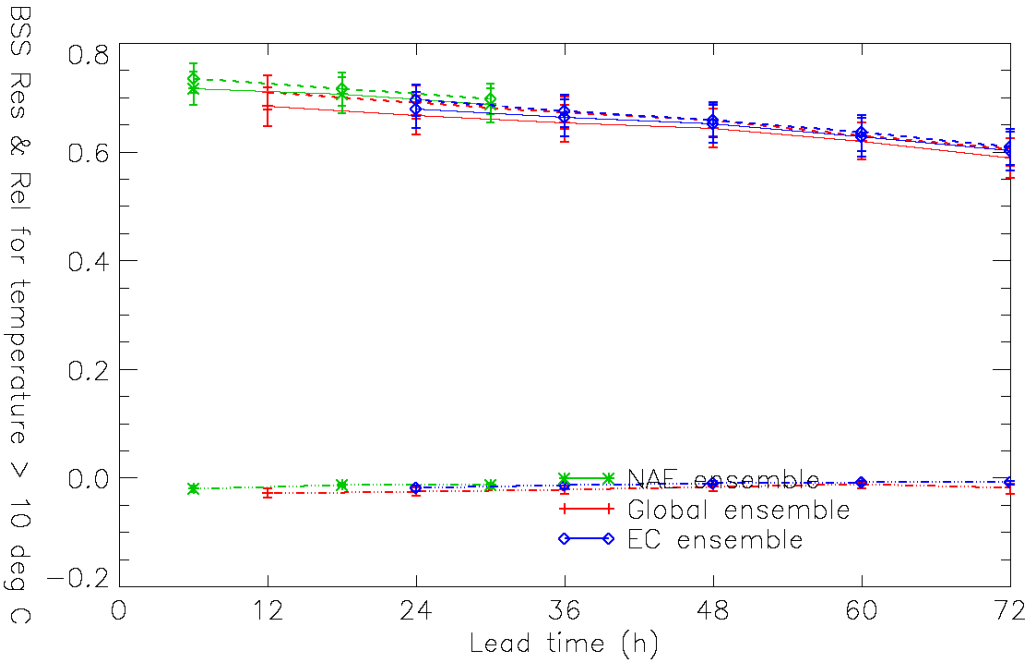


Figure 4.14. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 2m temperature greater than 10°C. The verification period is from 6 November 2006 to 28 February 2007. All the forecasts have been post-processed using the KFMOS bias correction.

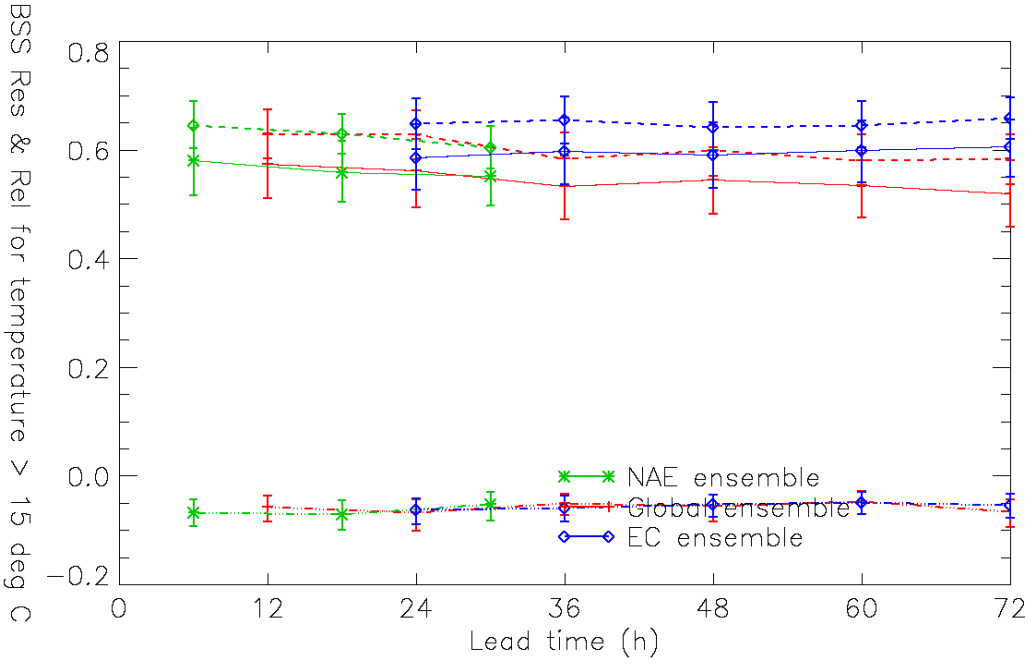


Figure 4.15. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of 2m temperature greater than 15°C. The verification period is from 6 November 2006 to 28 February 2007. All the forecasts have been post-processed using the KFMOS bias correction.

4.5 Wind speed verification

For forecasts of wind speed figure 4.16 shows the variation is BSS for forecasts of wind speed of at least force 5. There is a clear variation in skill with season, which is common to all the ensemble systems. Thus, it would be appropriate to compare their performance over the whole period. However, the KFMOS bias correction for MOGREPS wind speed contained an error that was corrected on 12th October 2006. Thus, we use the period 6 November 2006 to 31 March 2007 for wind speed verification.

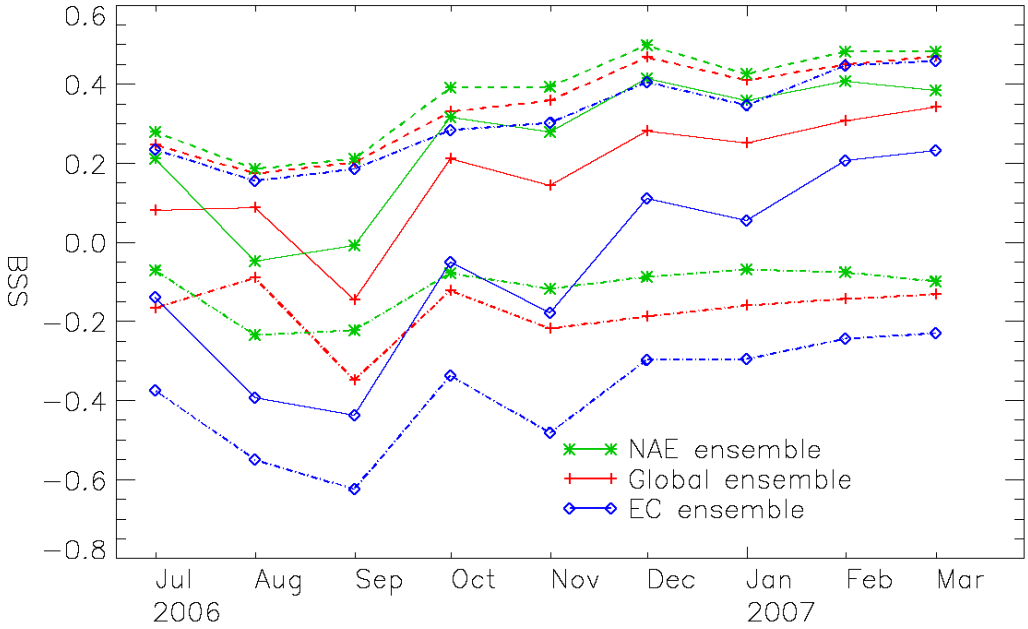


Figure 4.16. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of wind speed at least force 5. The verification period is for each month from July 2006 to March 2007.

The performance of the ensembles for forecasts of wind speed of at least force 5 and force 7 are shown in figures 4.17 and 4.18, respectively. These show that both the NAE and global ensembles are significantly more skilful than the ECMWF ensemble, with the NAE ensemble significantly more skilful than the global ensemble at the lower threshold. The reliability penalty is the main difference between the three ensembles. Figure 4.19 shows the reliability diagram for forecasts of wind speed of at least force 5 for forecast lead times of T+30 (for the NAE) and T+36 (for the global and ECMWF ensembles). All the ensembles appear to have a bias – they forecast the occurrence of this event too often, with the bias being most severe for the ECMWF ensemble.

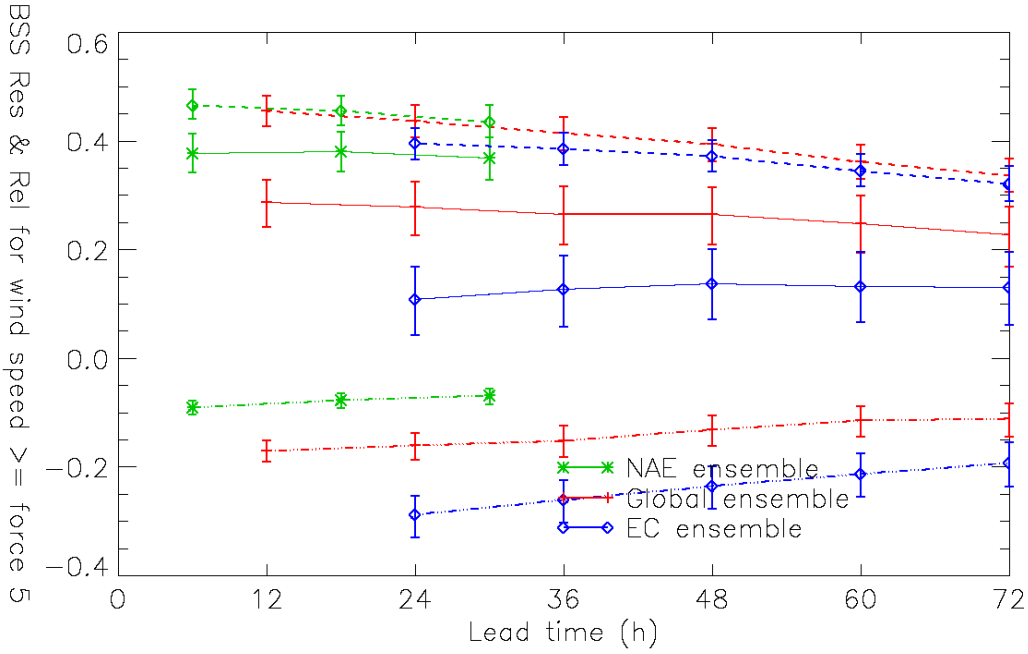


Figure 4.17. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of wind speed at least force 5. The verification period is from 6 November 2006 to 31 March 2007.

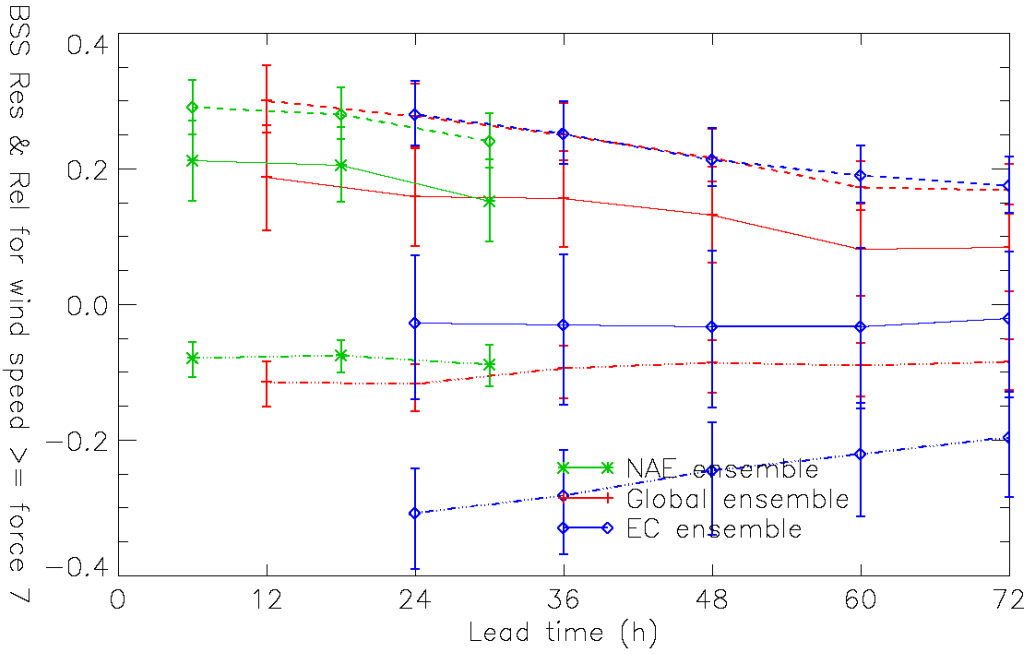


Figure 4.18. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of wind speed at least force 7. The verification period is from 6 November 2006 to 31 March 2007.

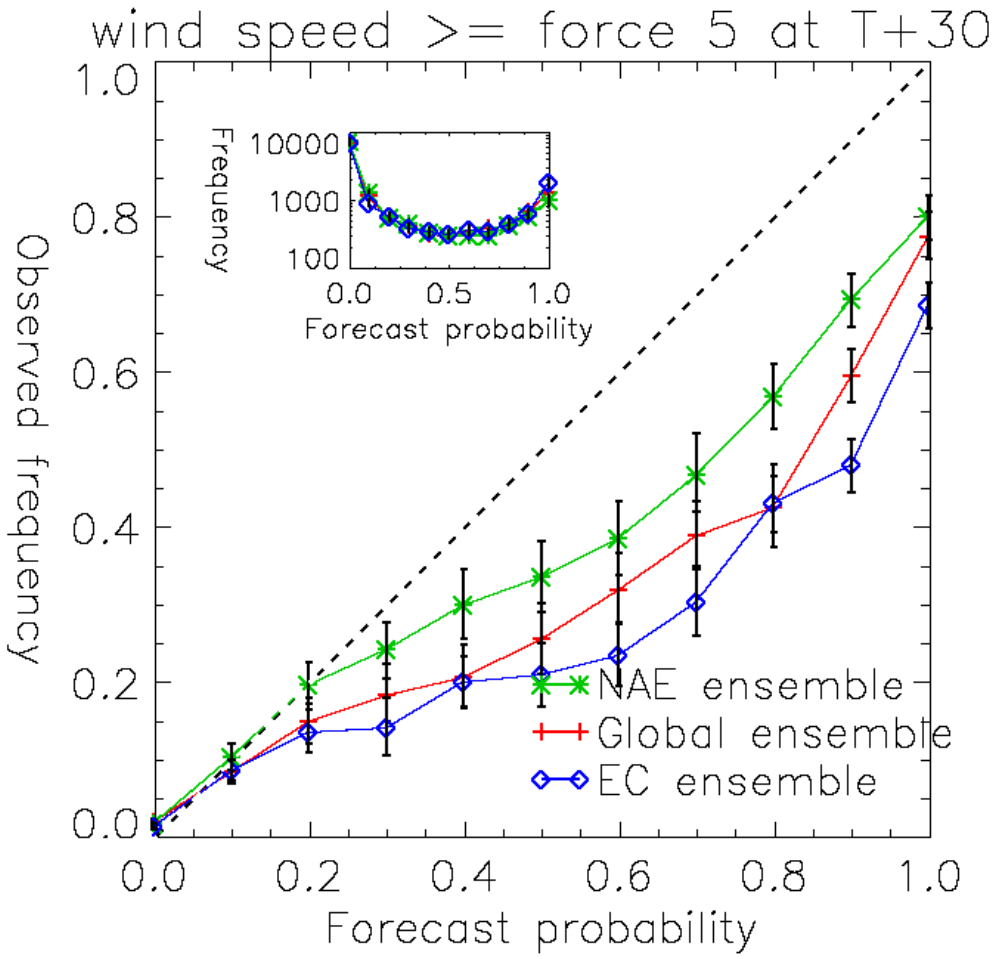


Figure 4.19. Reliability and sharpness for the NAE, global and ECMWF ensembles for forecasts of wind speed at least force 5. The verification period is from 6 November 2006 to 31 March 2007. The lead time of the forecasts is T+36 (global and ECMWF) and T+30 (NAE).

Equivalent graphs to figures 4.17 and 4.18, but for forecasts which have been post-processed using the KFMOS scheme are shown in figures 4.20 and 4.21. The difference between the three ensembles is much less. For forecasts of wind speed of at least force 5 the NAE ensemble performs marginally better than the ECMWF ensemble and significantly better than the global ensemble. The ECMWF ensemble is significantly more skilful than the global ensemble at longer lead times. For forecasts of wind speed of at least force 7, the ECMWF ensemble performs best after post-processing, with the NAE ensemble next most skilful. In this case, the ECMWF ensemble is significantly more skilful than the global ensemble at all lead times.

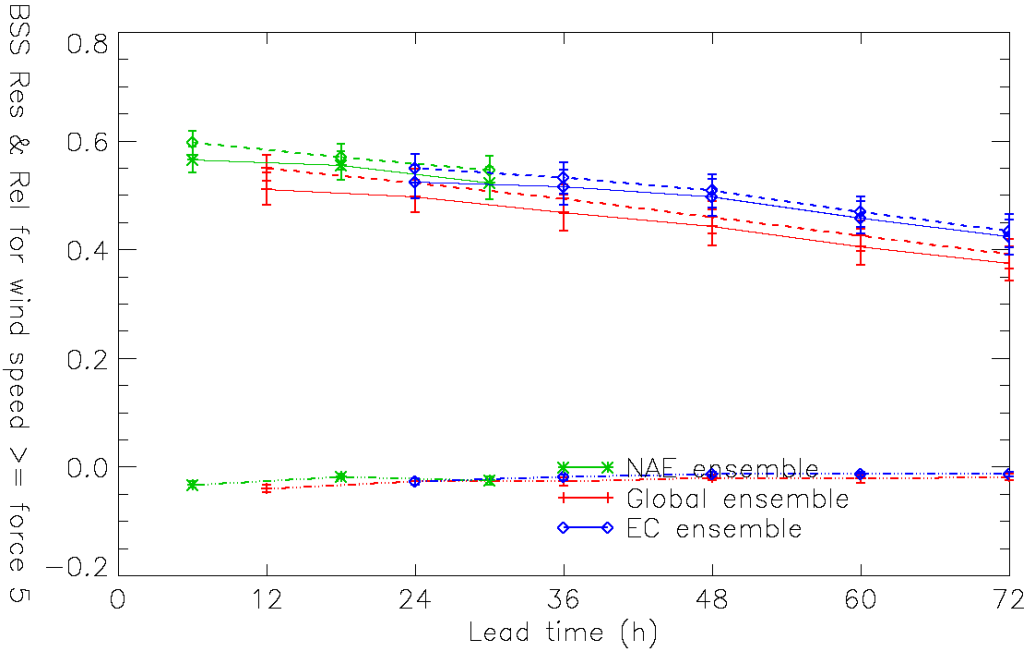


Figure 4.20. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of wind speed at least force 5. The verification period is from 6 November 2006 to 31 March 2007. All the forecasts have been post-processed using the KFMOS bias correction.

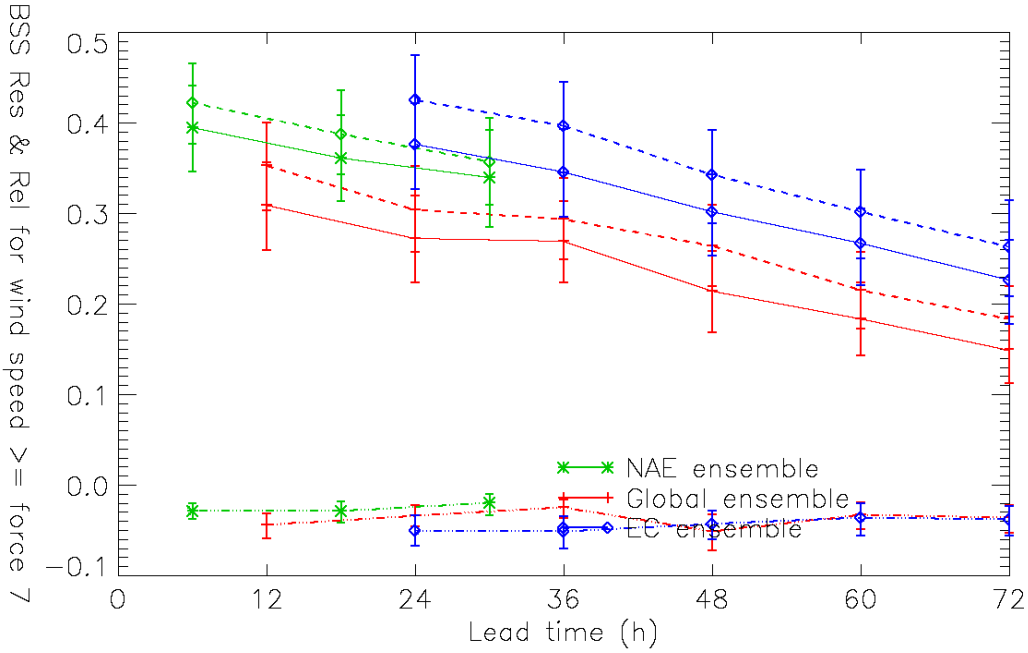


Figure 4.21. Brier skill score (solid), and reliability (dash-dot) and resolution (dashed) components for the NAE, global and ECMWF ensembles for forecasts of wind speed at least force 7. The verification period is from 6 November 2006 to 31 March 2007. All the forecasts have been post-processed using the KFMOS bias correction.

4.6 Conclusions

For precipitation the NAE ensemble is generally the most skilful, the only exception to this is at the threshold of 12h accumulated precipitation of more than 20 mm. Much of the benefit from the NAE compared to the ECMWF EPS comes from better reliability. The resolution score of the two ensembles are often very similar, with ECMWF EPS having slightly higher values of ROC area for the 0.5mm threshold. The improvement for the NAE ensemble system is particularly good over summer, when the ECMWF forecast performs poorly for light rain.

For temperature forecasts the NAE ensemble performs better than the other two ensembles before post-processing. After post-processing, the NAE ensemble performs best for forecasts of 2m temperature greater than 10^{o}C, and the ECMWF ensemble performs best for forecasts of 2m temperature greater than 15^{o}C, although there are few events in this category for the verification period studied.

For forecasts of wind speed, the NAE ensemble performs better than the other two systems before post-processing. After post-processing, the NAE ensemble performs best for forecasts of wind speed of at least force 5. For forecasts of at least force 7, the ECMWF ensemble is most skilful after post-processing.

Overall, therefore, it is clear to say that the NAE ensemble is the most skilful of the three ensemble systems studied in most situations. Where the KFMOS post-processing has been applied to reduce site-specific biases it reduces the differences in performance between the ensembles such that benefits of the NAE compared to the EPS are mainly not statistically significant. However it should be borne in mind that for many applications customers' needs cannot be fully met by univariate post-processed output. For customers who need the correlations between variables the performance of raw ensemble output provides the best indication of useable skill.

5. Station-based verification – Spread Skill Results

5.1 Introduction

In a perfect ensemble system, which has infinite ensemble members and all sources of uncertainty accounted for, the true state of the atmosphere should always be contained within the forecast distribution. In such a system the spread of the ensemble forecasts could be used to represent the forecast uncertainty. For example, where there is a large spread in the ensemble forecasts the uncertainty of the forecast would be high whereas a small spread would indicate low forecast uncertainty. Such a system would therefore be able to produce reliable probability forecasts. The aim of this study is to evaluate the spread-skill relationship of MOGREPS and to

determine if the spread of the ensemble can be used to estimate the uncertainty of the forecast.

5.2 Data

To investigate the spread-skill relationship of MOGREPS, output was used from the NAE domain and station-based verification was performed. Two periods were investigated, summer 2006 (JJA) and winter 2006/7 (DJF). The results for the winter and summer seasons were calculated separately so that the spread-skill relationship could be compared for different meteorological regimes. For both of the three month periods 30 hour forecasts of 1.5m temperature and 10m wind speed were considered. Throughout these results only one lead time is considered because there is an inherent relationship between the forecast lead time and ensemble spread. Incorporating a mix of lead times would therefore provide misleading results. 30 hour forecasts were selected for investigation because the longer length forecasts allow greater development of the small initial differences in the ensemble members and therefore the potential for greater spread.

The method used to determine the spread–skill relationship of MOGREPS will now be described using the 10m wind speed in DJF as an example. The same method was used for the summer period and these results are shown at the end. On each day during the 3 month period MOGREPS produced a new set of forecasts for the NAE domain. These daily forecasts were verified using observations made at 55 UK stations. Raw ensemble data was used and the ensemble mean at each station was calculated. This was then plotted against the observation to determine if there was a bias in the ensemble mean. Figure 5.1 shows the wind speed measurements plotted against the ensemble mean forecast for each station, on each day, during the 3 month period (DJF), producing about 5,000 data points. The blue trend line superimposed on the data shows there is a bias in the ensemble mean wind speed with the ensemble over-forecasting low wind speeds and under-forecasting high wind speeds. The black line represents the diagonal where the data trend line should be located when there is no bias present. A bias correction was therefore applied to the ensemble forecasts to ensure that the trend line fits the diagonal.

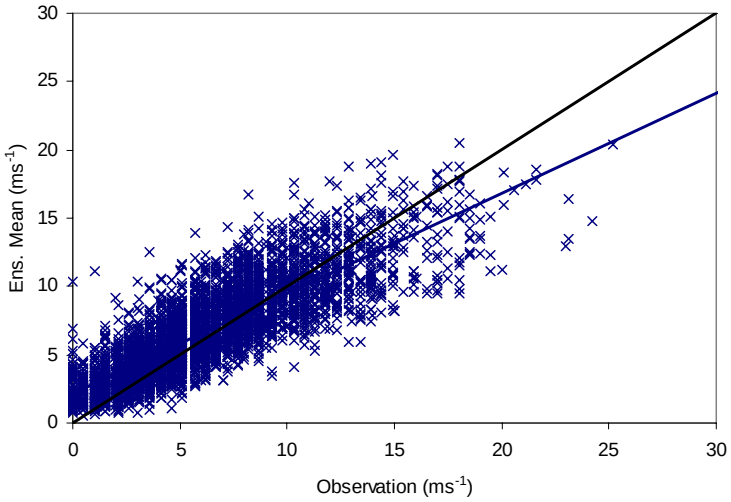


Figure 5.1: Measured wind speed in DJF plotted against the ensemble mean forecast (ms^{-1}).

With the data bias-corrected, the standard deviation of the ensemble members for each location on each day, were calculated. A strong correlation would not necessarily be expected when comparing the standard deviation of each individual event against the value of the ensemble mean error and it can be misleading, discussed in Houtekamer (1993). Therefore the events were grouped together according to the value of their forecast standard deviation. Each standard deviation bin contained an equal number of events so the spread of the standard deviation contained in each bin varied. This avoided the bins containing the largest standard deviation values being determined by a small number of events. The average value of the standard deviation in each bin was then compared to the root mean square error (RMSE) for the events in the bin providing a more robust measure of the spread-skill relationship.

To help evaluate the spread-skill relationship of MOGREPS, these results were compared to results from two other, artificially generated ensembles. The first was an ensemble which had no spread-skill relationship and is referred to as the 'no skill' ensemble. The second was an ensemble which had 'perfect spread' which means the observation was always contained within the ensemble distribution. Comparing the MOGREPS results with the results from these other two ensembles illustrates the quality of its spread-skill relationship.

The ensemble with no spread-skill relationship was generated by randomly associating the standard deviation value from each event with an ensemble mean error value from a different event. These new, 'no skill' events are then grouped again according to the value of the standard deviation. The value of the average standard deviation within each bin was then plotted against the RMSE value for each of the 15 standard deviation bins, shown by the pink data set in figure 5.2. The blue data set is the MOGREPS results.

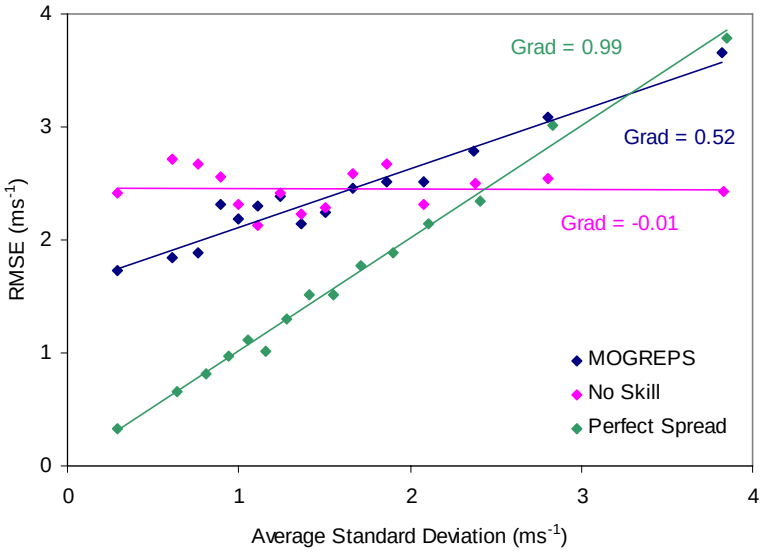


Figure 5.2: The average standard deviation in each bin plotted against the RMSE in each bin. Values for wind speed (ms^{-1}) in DJF at T+30. The gradient of each trend line is also included.

The final data set plotted on figure 5.2, the green data set, is that generated by an ensemble with perfect spread. These results are produced by replacing the observation for each event with a randomly selected ensemble member forecast. These pseudo-observations are therefore by definition always within the distribution of the ensemble, creating an ensemble with perfect spread. The results in figure 5.2 and the results for the other variables are discussed in the next section.

5.3 Results

The results in figure 5.2 show a strong spread-skill relationship for wind speed in DJF. The gradient of the MOGREPS trend line (0.52) lies between the 'perfect spread' and 'no skill' lines. As expected MOGREPS does not match the 'perfect spread' line, but it does show a spread-skill relationship. The results for wind speed in JJA and temperature in JJA and DJF are shown in figures 5.3 to 5.5.

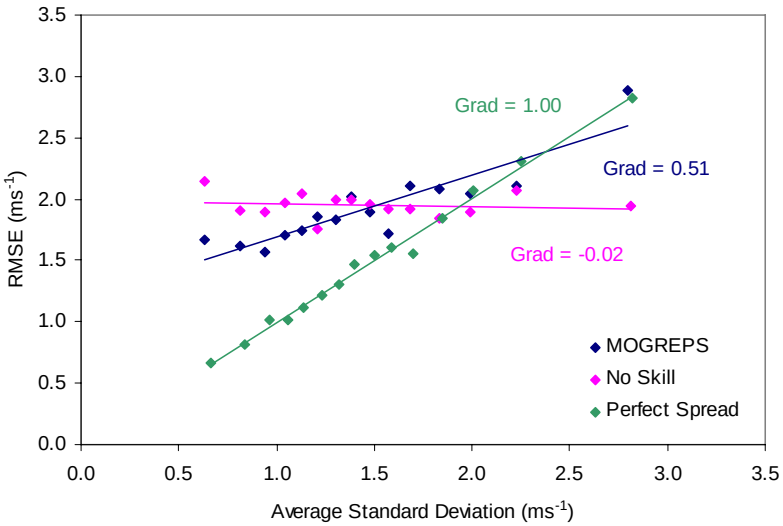


Figure 5.3. The average standard deviation in each bin plotted against the RMSE in each bin. Values for wind speed (ms^{-1}) in JJA at T+30.

A bias correction was performed for the JJA wind speed forecasts to correct for the ensemble over-forecasting at low wind speeds. The bias corrected data was then binned according to standard deviation and the results are shown in figure 5.3. The gradient of the MOGREPS trend line (0.51) indicates a strong spread-skill relationship and lies approximately mid-way between the 'no skill' and 'perfect spread' ensembles.

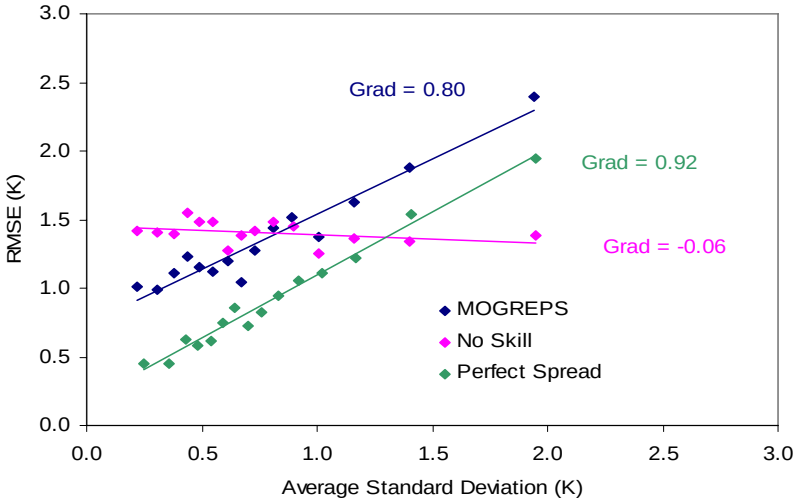


Figure 5.4. The average standard deviation in each bin plotted against the RMSE in each bin. Values for temperature (K) in DJF at T+30.

For the ensemble mean temperatures in DJF no bias correction was required. Figure 5.4 shows that in DJF there is a strong spread-skill relationship for this variable with MOGREPS closely following the perfect spread line with a gradient of 0.80. In JJA, however, shown in figure 5.5, there is very little evidence of a spread-skill relationship with very little difference between MOGREPS and the 'No Skill' line. Both lines have a very shallow gradient, with MOGREPS having a gradient of 0.12 and the 'No Skill' line having a gradient of 0.05.

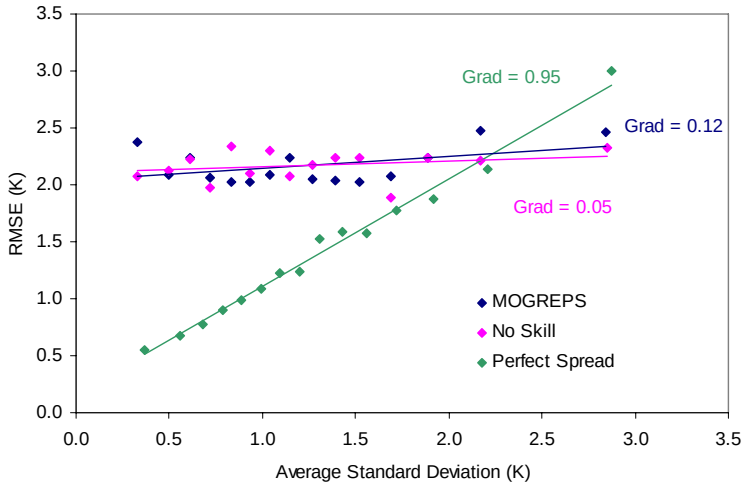


Figure 5.5. The average standard deviation in each bin plotted against the RMSE in each bin. Values for temperature (K) in JJA at T+30.

Another factor taken into consideration is the impact of the observation error on the performance of MOGREPS. In the 'perfect spread' ensemble observation error is not a component of the RMSE because the pseudo-observation is the forecast from one of the ensemble members. Therefore, to allow a fair comparison the observation error was estimated for temperature and wind speed.

The estimated observation errors in the station list for surface observations were used as a guide to the maximum value of temperature and wind speed errors. The estimate in the station list for the temperature observation errors was 1.1K and the estimate for wind speed error was 2.0ms^{-1}. In three of the four cases the estimated errors in the station list appeared to be larger than the errors on the actual observations. The errors were therefore estimated to give values which are lower than the lowest MOGREPS points in figures 5.2 to 5.5, providing error estimates of ~1.3ms^{-1} for wind speed and 0.8K for temperature in DJF. The RMSE value in each bin was then corrected (RMSE$_c$) by removing the observation error contribution:

$$\text{RMSE}_c = \sqrt{RMSE^2 - ObsErr^2}$$

To remove the effect of MOGREPS being under or over-spread the RMS of the standard deviations (RMSS) of all the events was calculated. This was then compared to the RMSE$_c$ for all the events. For a correctly spread ensemble RMSS should be equivalent to RMSE$_c$, so using this information a correction factor was applied to the standard deviation of each event. This correction factor was applied to all three 'ensembles' and the estimated observation error was removed from MOGREPS and the 'no skill' ensemble. The results are shown in figures 5.6 to 5.9.

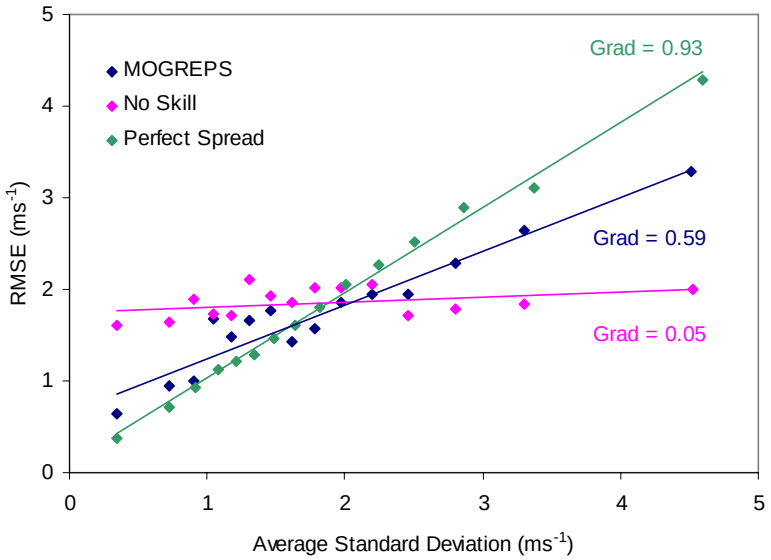


Figure 5.6. The average standard deviation in each bin plotted against the RMSE in each bin, corrected for observation error. Values for wind speed (ms^{-1}) in DJF at T+30.

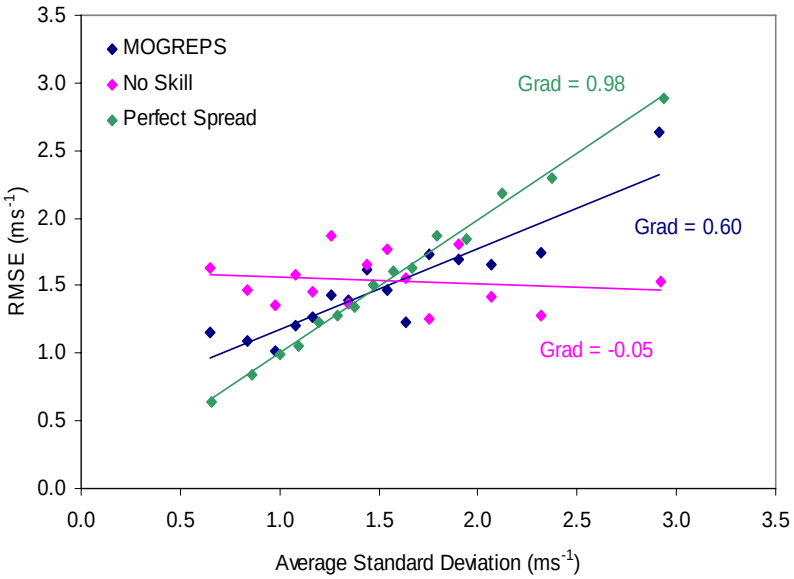


Figure 5.7. The average standard deviation in each bin plotted against the RMSE in each bin, corrected for observation error. Values for wind speed (ms^{-1}) in JJA at T+30.

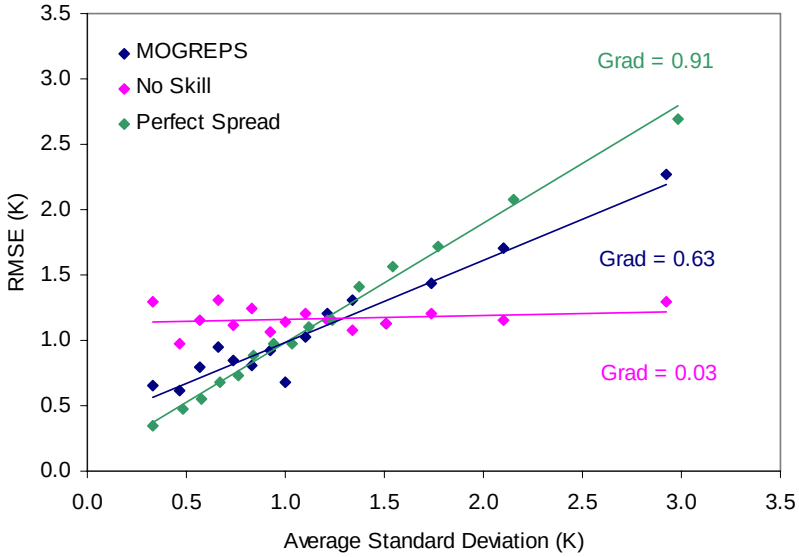


Figure 5.8. The average standard deviation in each bin plotted against the RMSE in each bin, corrected for observation error. Values for temperature (K) in DJF at T+30.

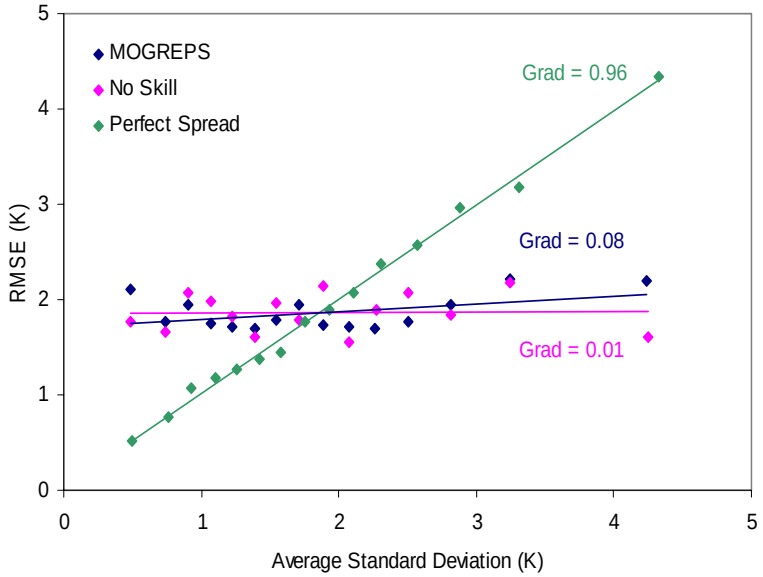


Figure 5.9. The average standard deviation in each bin plotted against the RMSE in each bin, corrected for observation error. Values for temperature (K) in JJA at T+30.

Table 5.1 summarises the results from figures 5.6 to 5.9. It contains the gradients of the trend lines for the 3 ensembles for each variable, in each season. The results for the 'no skill' ensemble show that there is no spread-skill relationship with the gradient ranging between ±0.05. MOGREPS however, displays gradients between 0.59 and 0.63 for 3 out of the 4 cases. The only case where it does not perform well is for temperature in JJA where the gradient is only 0.08, indicating that in this case there is virtually no

spread-skill relationship. The perfect spread ensemble has a gradient close to one in all cases with values ranging from 0.91 to 0.98.

	PS	Real	NS	Skill score
Wind Speed DJF	0.93	0.59	0.05	0.61
Wind Speed JJA	0.98	0.60	-0.05	0.63
Temperature DJF	0.91	0.63	0.03	0.68
Temperature JJA	0.96	0.08	0.01	0.07

Table 5.1. Trend line gradients for the 'perfect spread' data (PS), MOGREPS (Real) and the 'no skill' (NS) data sets.

5.4 Conclusions

The results shown are for 30h forecasts of 1.5m temperature and 10m wind speed at specific sites over 2 different seasons. The results are therefore subject to the limitations of a small sample but should still be indicative of the general underlying trends.

When taking into account the impact of observation errors on the RMSE the results show that MOGREPS performs substantially better than an artificially generated 'no skill' ensemble. In 3 of the 4 cases considered here the skill score associated with the gradient of the trend line, representing the spread-skill relationship, is between 0.6 and 0.7. The differences between MOGREPS and the 'perfect spread' ensemble are influenced by the limited number of ensemble members in MOGREPS and the perturbation strategies employed by the ensemble.

6.1 Verification of Probabilistic products

6.1.1 Precipitation Verification

Nimrod analyses blend rainfall accumulations derived from radar and surface gauge measurements to provide a spatially coherent estimate of the observed precipitation field. Verification against Nimrod data is performed at one degree resolution within the ABV (approximately four and a half times the grid length of the forecast model and equivalent to the effective resolution of the forecast). Currently Nimrod data is only available within the ABV over the UK region. Here we focus on results for 6 and 24 hour rainfall accumulations.

Presented in figure 6.1 and 6.2 are attributes diagrams for the T+36 6hr precipitation forecast greater than or equal to 0.3 mm and the T+36 24hr precipitation forecast greater than 0.5mm for the period 1st January 2006 to 28th February 2007. It can be seen from figures 6.1 and 6.2 that the reliability curve lies very close to the diagonal indicating that the ensemble probabilities exhibit very good reliability at the low precipitation thresholds. However there is evidence that the ensemble is marginally over-confident, slightly over-forecasting high probabilities and under-forecasting the low probabilities. Figure 6.3 shows that the near perfect reliability for the 6hr accumulation

forecast observed in figure 6.1 is also evident at shorter forecast lead times, whilst the resolution of the forecasts decrease with increasing forecast lead time. This is also evident at the forecasts of 24hr precipitation accumulations (not shown). It should be noted that the reference forecast used for calculation of the Brier skill score for all the results presented here is a climatological probability derived from the observations within the verification sample. This is not an ideal practice because this reference forecast is not available *a-priori* and it effectively gives the climatological forecast an advantage over the ensemble forecast and hence decreases the apparent skill of the ensemble forecast. However, no suitable *a-priori* climatology is currently available for the set of observations used in the ABV.

The reliability curves for the higher thresholds as shown in figures 6.4 to 6.7 are less smooth than the reliability curve shown in figures 6.1 and 6.2 due to the smaller number of events. These reliability curves lie below the line of perfect reliability indicating that the ensemble is over forecasting the probability of the larger rain amounts and appears to possess a wet bias (as can be seen in figure 6.8). Bias in this context is the total number of forecasts of the event divided by the total number of events that occurred. An unbiased ensemble would have a bias of one, values less (greater) than one indicate an under (over) forecasting bias. However, it can be seen that the ensemble forecasts do possess resolution and that in figure 6.4 and 6.5 they contribute towards a positive Brier Skill score.

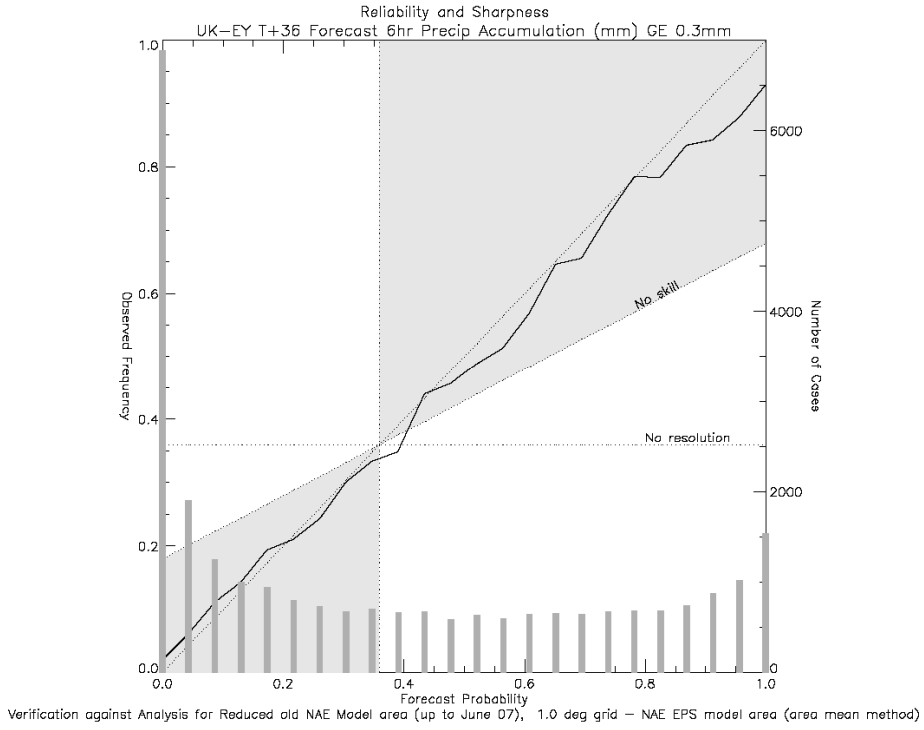


Figure 6.1. Attributes diagram T+36 forecast of 6hr accumulation of precipitation greater than or equal to 0.3mm.

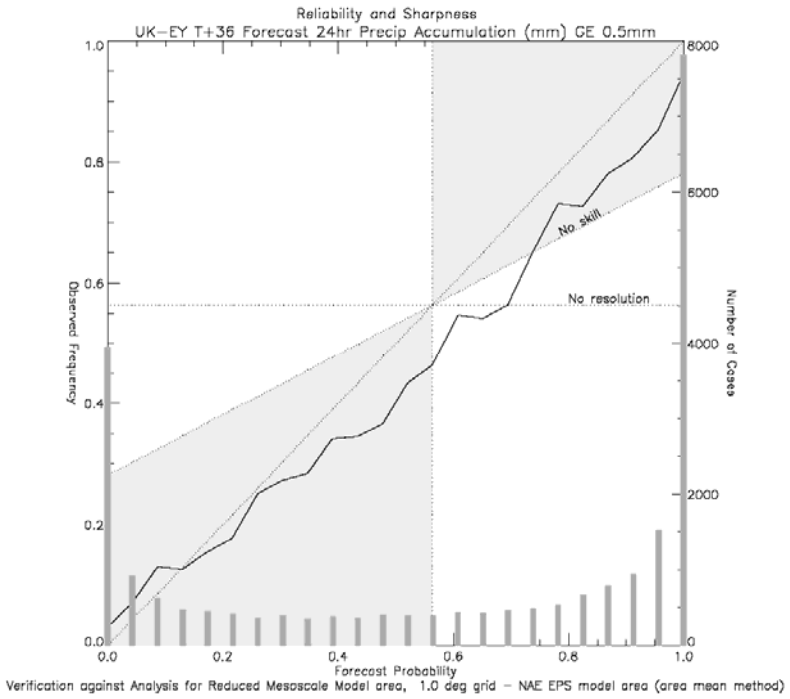


Figure 6.2. Attributes diagram T+36 forecast of 24hr accumulation of precipitation greater than 0.5mm.

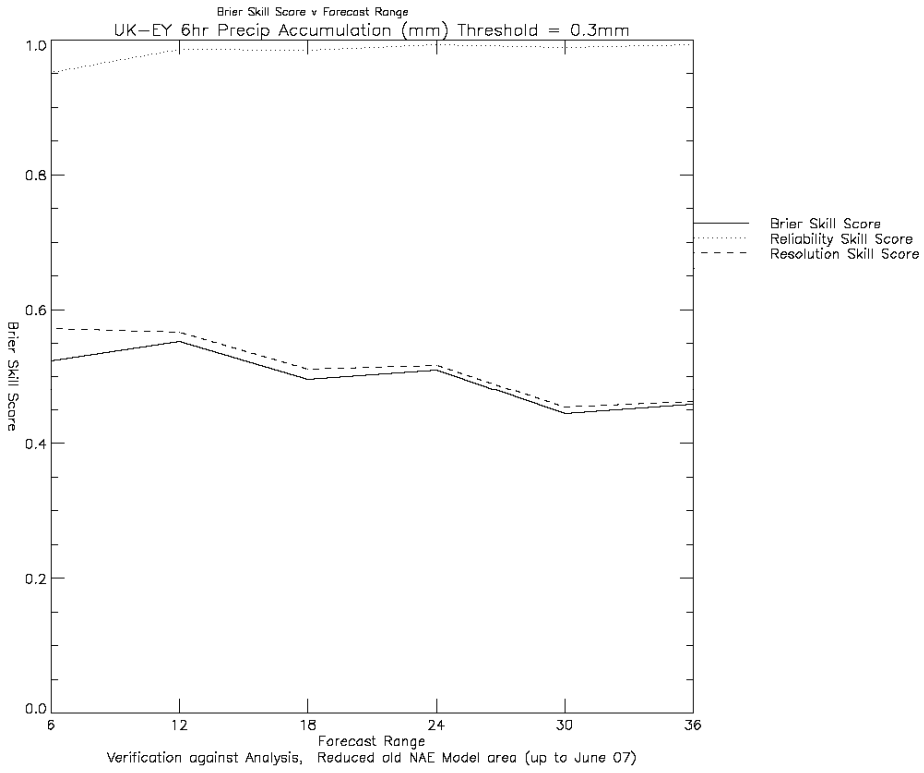


Figure 6.3. Brier skill score v forecast range for 6hr accumulation of precipitation greater than 0.3 mm.

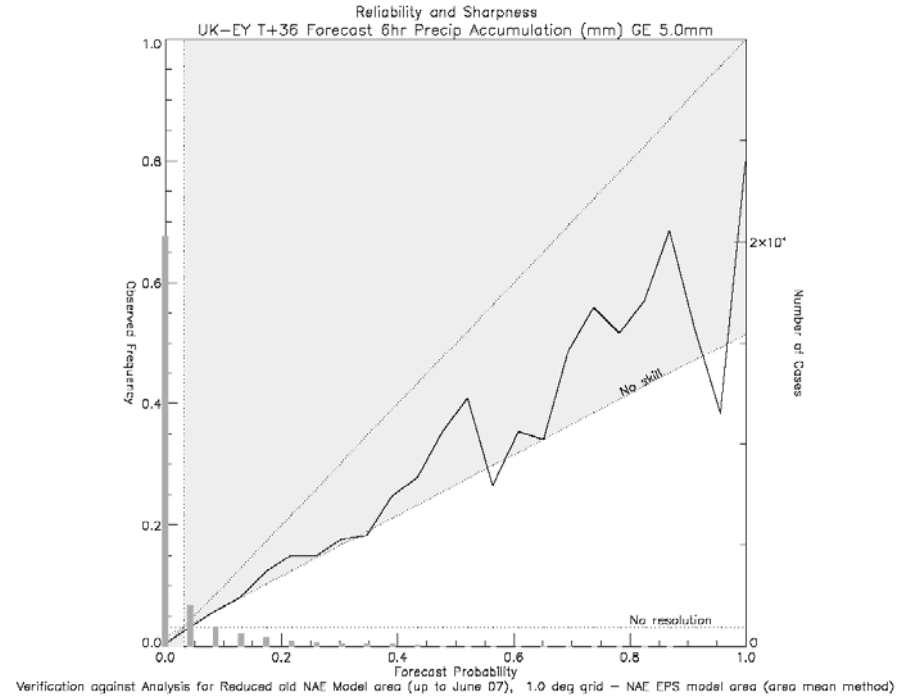


Figure 6.4. Attributes diagram T+36 forecast of 6hr accumulation of precipitation greater than or equal to 5.0 mm.

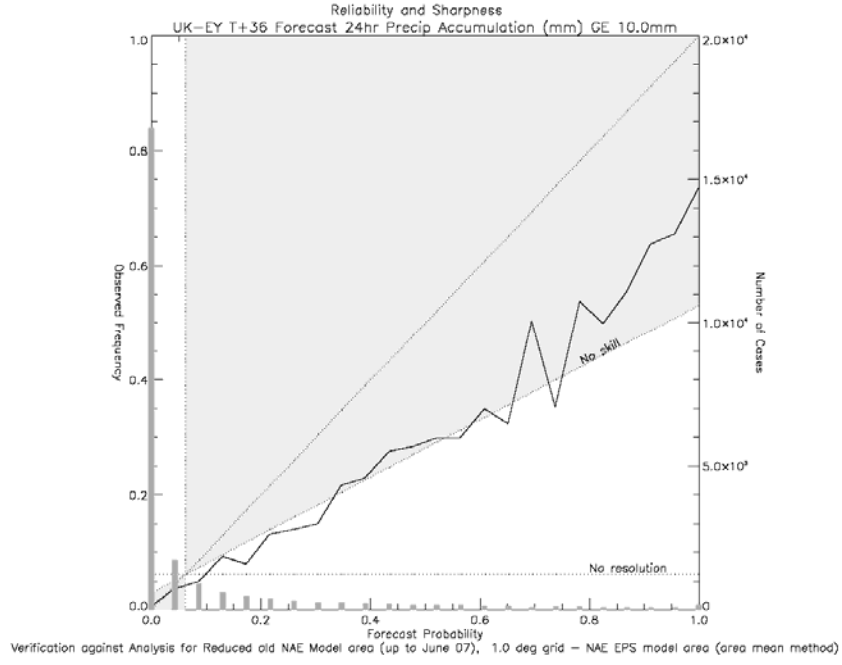


Figure 6.5. Attributes diagram T+36 forecast of 24hr accumulation of precipitation greater than 10.0 mm.

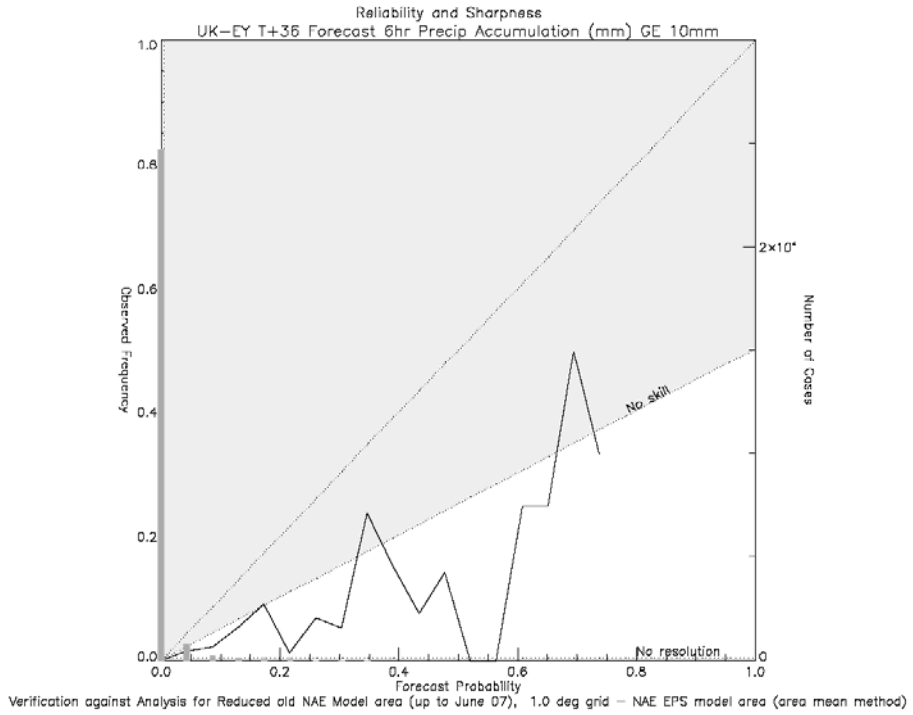


Figure 6.6. Attributes diagram T+36 forecast of 6hr accumulation of precipitation greater than 10.0 mm.

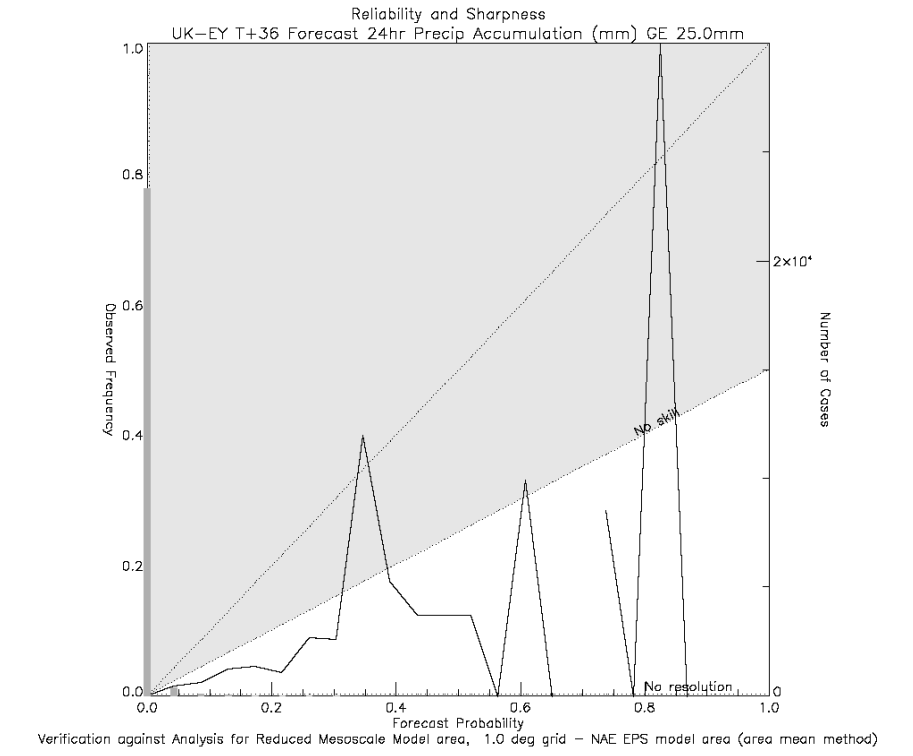


Figure 6.7. Attributes diagram T+36 forecast of 24hr accumulation of precipitation greater than 25.0 mm.

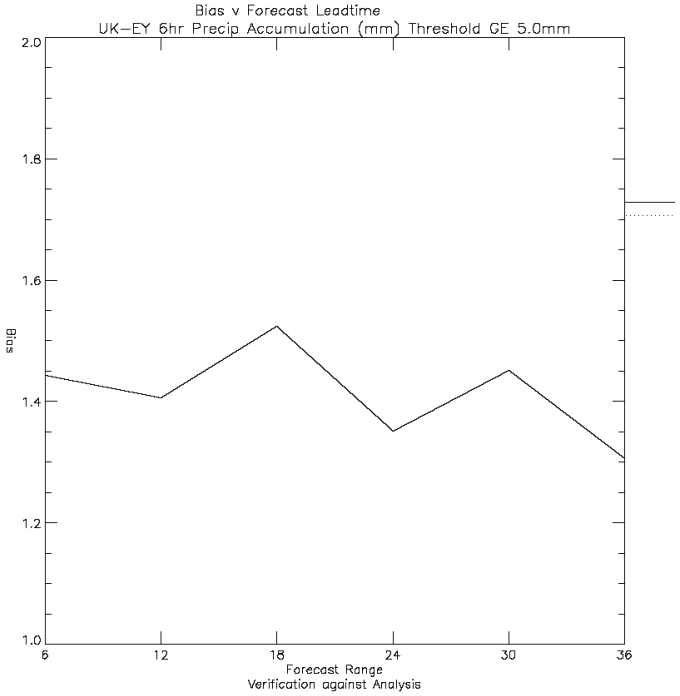


Figure 6.8. Bias in ensemble forecast v forecast lead time for 6hr precipitation greater than or equal to 5.0mm. A bias of one indicates a perfect forecast.

6.1.2 Wind Speed Verification

Attributes diagrams for wind speeds greater than force 8, 9 and 10, verified against surface observations for the period 1 January 2006 to 28th February 2007 are shown in figure 6.9 to figure 6.11. The verification has been performed over the reduced NAE model domain in order to increase the sample size as much as possible for these more-extreme wind speed thresholds. Due to the limited number of observations at these thresholds the reliability diagrams are increasingly noisy. Nonetheless, there are positive slopes to the reliability curves which indicate that, even at storm force 10, MOGREPS-R has some ability to provide information for extreme events. This result is particularly encouraging because MOGREPS-R wind products are fed into the EURORISK Windstorms project.

Figure 6.12, presents the bias in the ensemble forecasts for wind speed greater than Beaufort force 8 against forecast lead time for the three different geographical regions. It is interesting to note that the bias in figure 6.12 varies with geographical region, the larger the verification region the larger the under forecasting bias. This is consistent with the model under forecasting the 10m wind speed over land; the reduced NAE model domain contains many inland observation sites when compared to the UK Index list which contains many coastal sites.

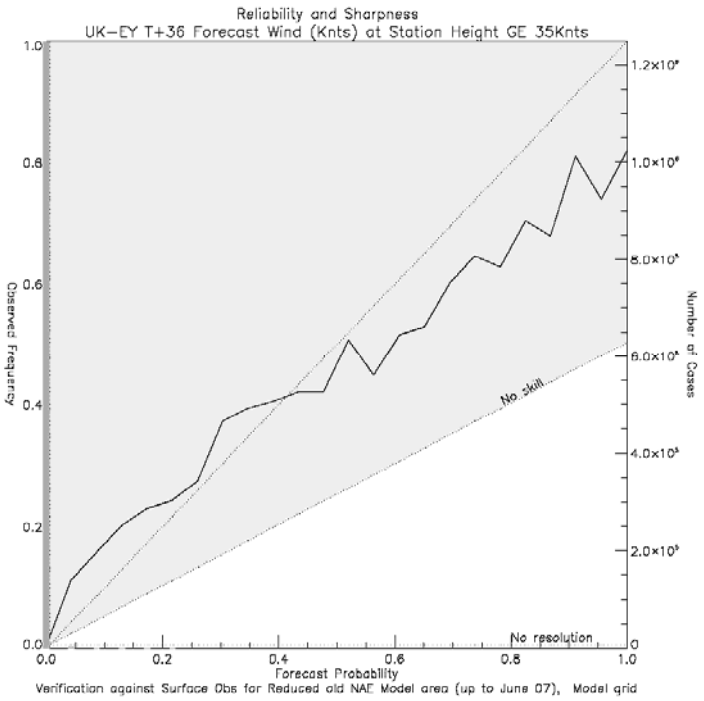


Figure 6.9, Attributes diagram T+36 forecast of 10m wind speed greater than or equal to 34 knots or Beaufort Force 8.

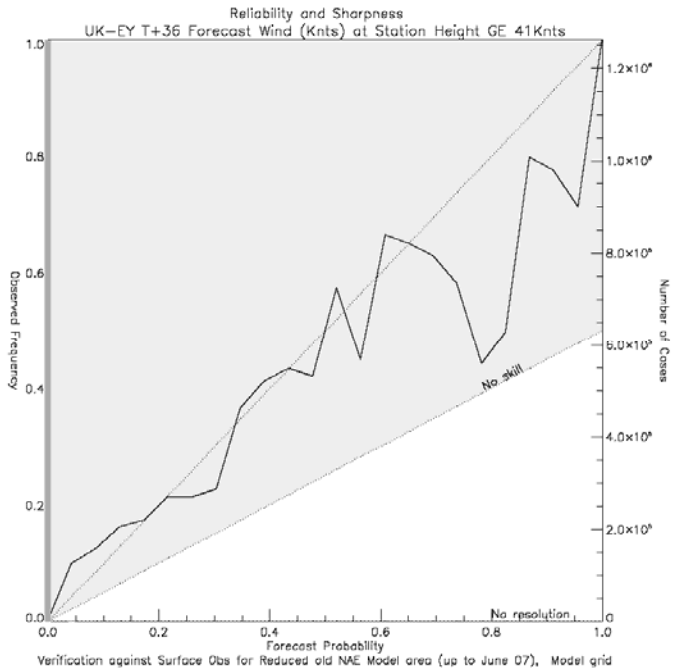


Figure 6.10, Attributes diagram T+36 forecast of 10m wind speed greater than or equal to 41 knots or Beaufort force 9.

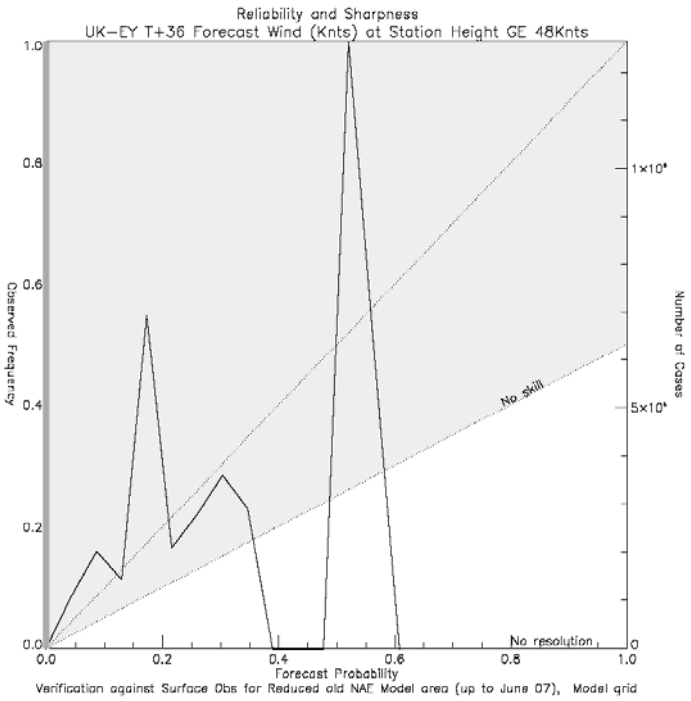


Figure 6.11, Attributes diagram T+36 forecast of 10m wind speed greater than or equal to 48 knots or Beaufort force 10.

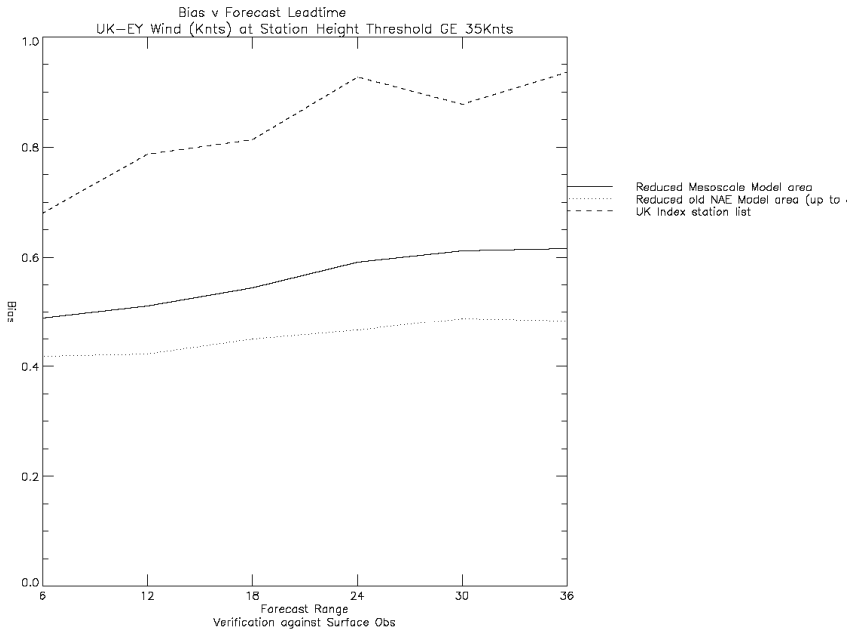


Figure 6.12, Bias v Forecast range for probability of wind speed greater than or equal to Gale force 8, for the UK Index Station list (dashed line), the reduced Mesoscale model area (solid line) and the reduced NAE model area (dotted line).

6.1.3 Visibility Verification

Prediction of visibility is very demanding for direct model output from any NWP model due to strong non-linear sensitivity of visibility to small errors in parameters such as humidity, and also to poorly simulated parameters such as aerosol. It is generally accepted that reliable NWP prediction of visibility requires much higher model resolution in both the horizontal and the vertical than the MOGREPS NAE model has. However visibility forecasts are very important to a number of Met Office customers, particularly in transport sectors, and subjective feedback from forecasters during the trial has consistently indicated that they have found useful guidance from the areas of high probability of fog or reduced visibility in MOGREPS NAE output. It is therefore of interest to examine whether any objective skill can be identified.

Visibility information is available from MOGREPS as an optional parameter on Meteograms and as probability charts of visibility less than 5000m, 1000m and 200m. Visibility has therefore been verified on a site-specific basis at the above thresholds.

Figures 6.13 to 6.15 presents attributes diagrams for visibility less than 200 m, 1000 m and 5000 m respectively. Whilst the reliability curves do exhibit some resolution, the reliability is poor and for figures 6.13 and 6.14 the reliability frequently falls below the no skill line. The curve in figure 6.15 is described as a conditional bias (Wilks, 1995) with the forecast probability being too high for the high probabilities, and too low for low probabilities. This behaviour, also referred to as over-confidence, is normally interpreted as the ensemble being under-spread. This suggests that there are some sources of uncertainty affecting this visibility threshold of 5000m that are not adequately represented in the MOGREPS ensemble. In this case, reduced visibility around 5000m is frequently caused by haze due to atmospheric aerosol, and this strong over-confidence may be related to the lack of any perturbation to aerosol concentration in MOGREPS.

Figure 6.16, shows the overall bias plotted against forecast range for a) 5000 m and b) 200 m. It can be seen that there is some variation in the levels of bias with forecast range, this could be related to the number of observations available at 06 and 18z (T+12,24,36) being less than at 00z and 12Z (T+6,18,30). In figure 6.16a, the bias is approximately 1 indicating that the ensemble is nearly unbiased overall.

Figure 6.16b also indicates that there is a very large over forecasting bias for fog (visibility less than 200m) and that this bias increases with increasing forecast range. The large bias is perhaps not surprising, given that fog is frequently a patchy phenomenon and while one observation site may have good visibility, another just down the road may have dense fog. In contrast visibility in the model is more consistent over large areas, due to model resolution and the lack of fine scale topographic features. Thus when the model predicts the conditions for fog formation, the diagnostic outputs are likely to suggest fog being much more widespread than it often is in reality, resulting in a tendency for the model to over predict fog amounts. As a result,

a forecast of fog from the NWP model should generally be interpreted as giving a probability of observing fog in the general area. It is notable that the bias changes greatly with forecast lead time. This indicates that the data assimilation system is attempting to remove this forecast bias. The bias observed in the ensemble results is likely to be dominated by the bias in the underlying forecast model which could perhaps be alleviated through bias correction prior to issuing the forecast.

It should also be noted that visibility has a highly skewed distribution of forecasts and observations which are dominated by high visibility events. In order to perform a meaningful assessment using continuous statistics (root mean square error and ensemble spread) it is perhaps helpful to consider performing the analysis on the logarithm of visibility, however, this has not been considered for this verification report.

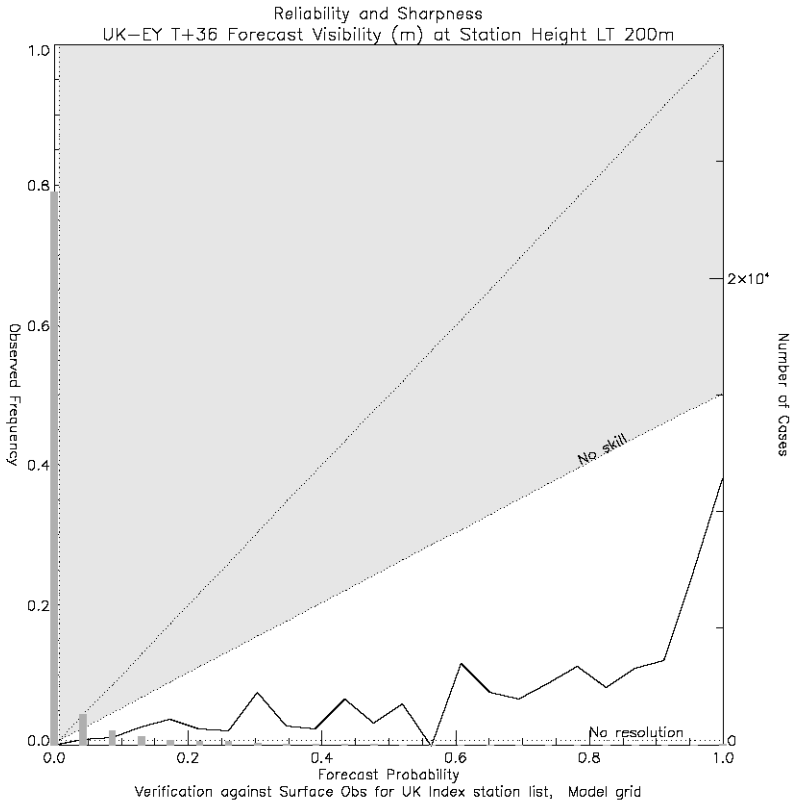


Figure 6.13, Attributes diagrams for T+36 forecast of Probability that visibility will be less than 200 m.

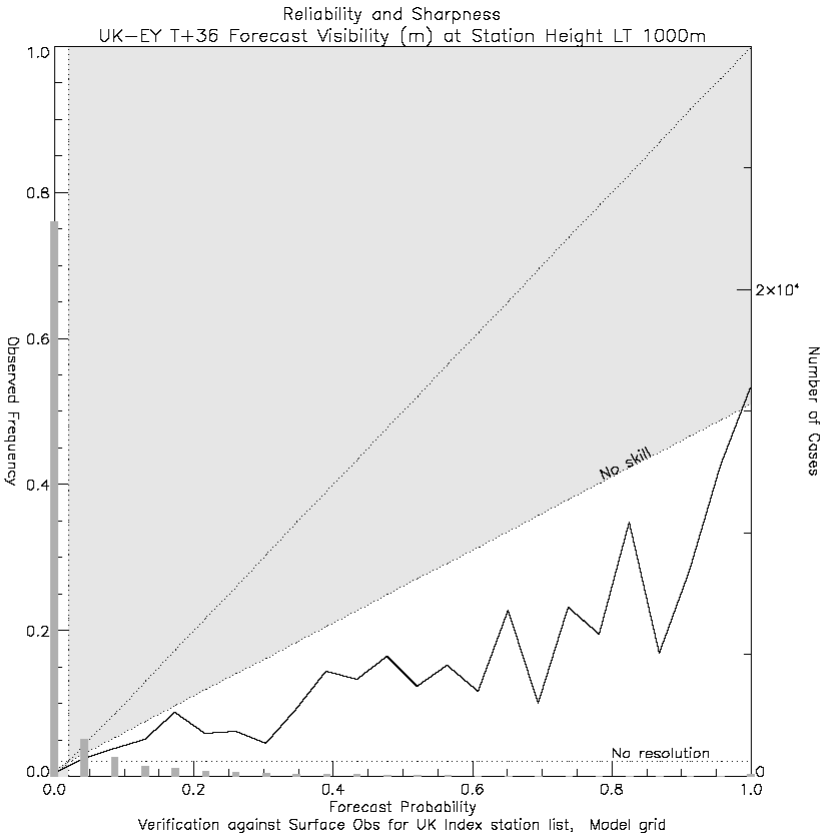


Figure 6.14, Attributes diagrams for T+36 forecast of Probability that visibility will be less than 1000 m.

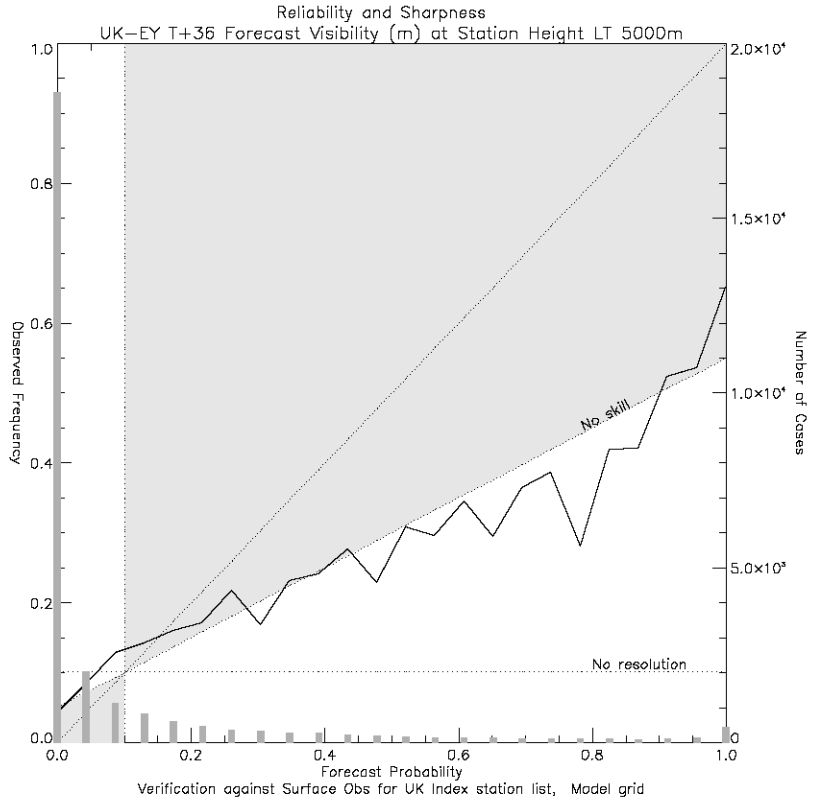


Figure 6.15, Attributes diagrams for T+36 forecast of Probability that visibility will be less than 5000 m.

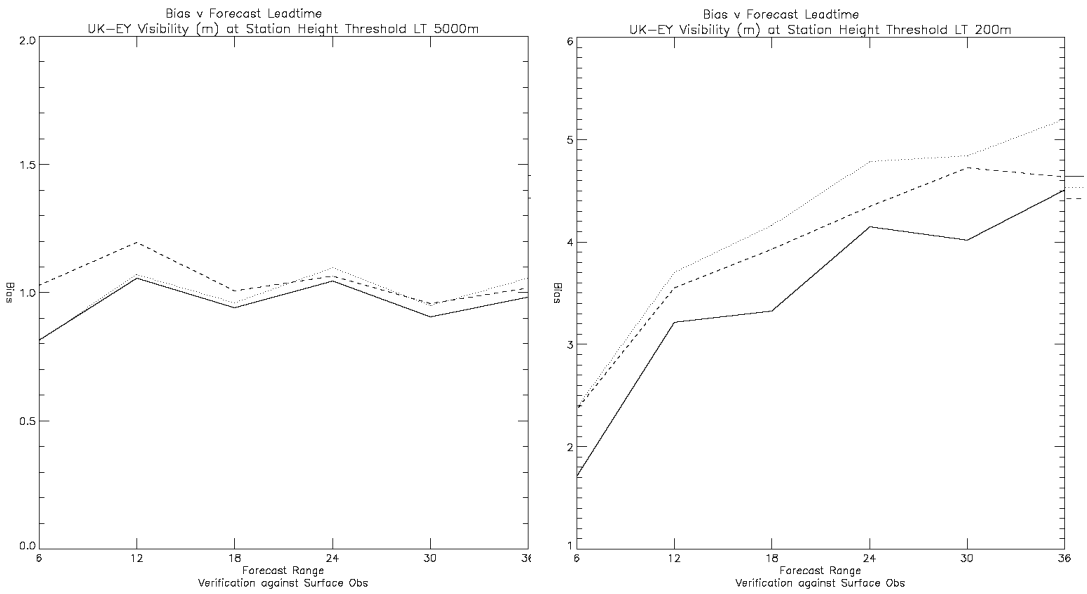


Figure 6.16, Bias v Forecast range for forecasts of visibility a) less than 5000 m and b) less than 200 m, for the UK Index Station list (dashed line), the reduced Mesoscale model area (solid line) and the reduced NAE model area (dotted line).

6.1.4 Cloud Base Height Verification

Forecasts of cloud base height given different levels of cloud cover are available from MOGREPS and are of particular relevance to aviation forecasts. It is therefore interesting to look at the performance for this joint probability, but as with visibility, low cloud base is poorly resolved by the model and therefore challenging for the ensemble. Figures 6.17 and 6.18, present attributes diagrams for cloud base height given 5/8th cover less than 500ft (152 m) and 1000 ft (304 m) respectively and figure 6.19 shows an attributes diagram for cloud base height given 3/8th cover less than 700ft (213m). Figures 6.17 to 6.19 show that the ensemble forecasts have no skill (in the sense of the Brier skill score, relative to sample climatology) and limited resolution. As with forecasts of poor visibility, we interpret this as meaning that the ensemble is under-spread and that some uncertainties in the forecast are unaccounted for. Figure 6.20, shows that the bias observed in forecasts of cloud base height less than 700ft given 3/8th cover is much larger over the reduced NAE model area than over UK Station list or over the reduced Mesoscale model area – suggesting that the forecast bias is worse over land.

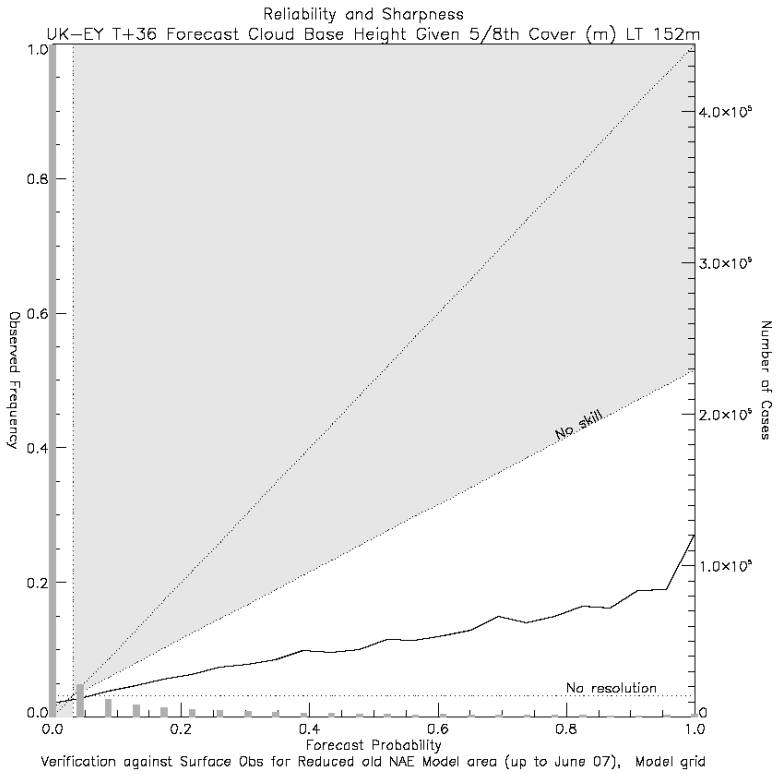


Figure 6.17, Attributes diagrams for T+36 forecast of Probability cloud base height will be less than 500 ft given 5/8 Cloud cover, verified over the reduced NAE model area.

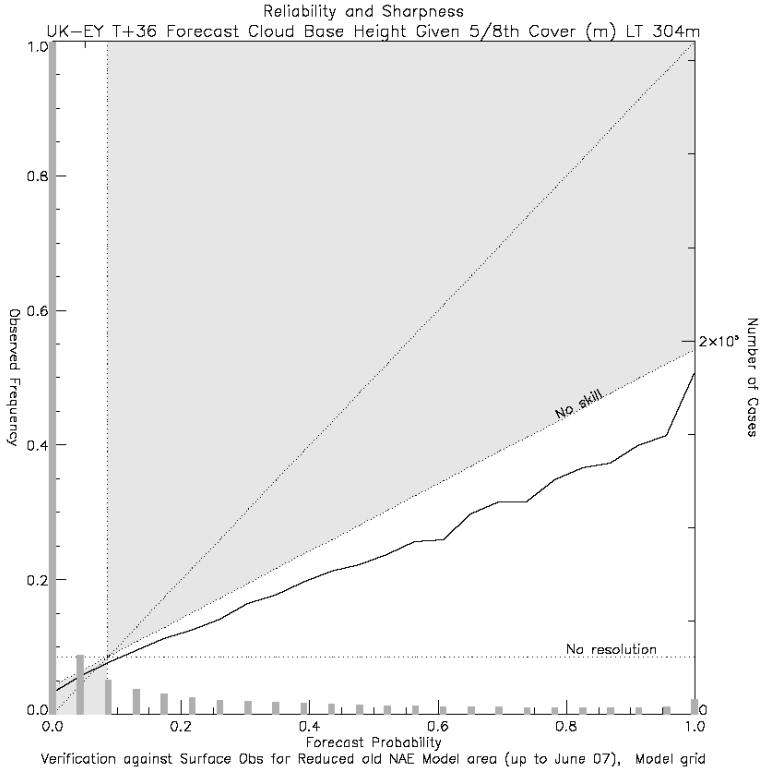


Figure 6.18, Attributes diagrams for T+36 forecast of Probability cloud base height will be less than a1000ft given 5/8 Cloud cover, verified over the reduced NAE model area.

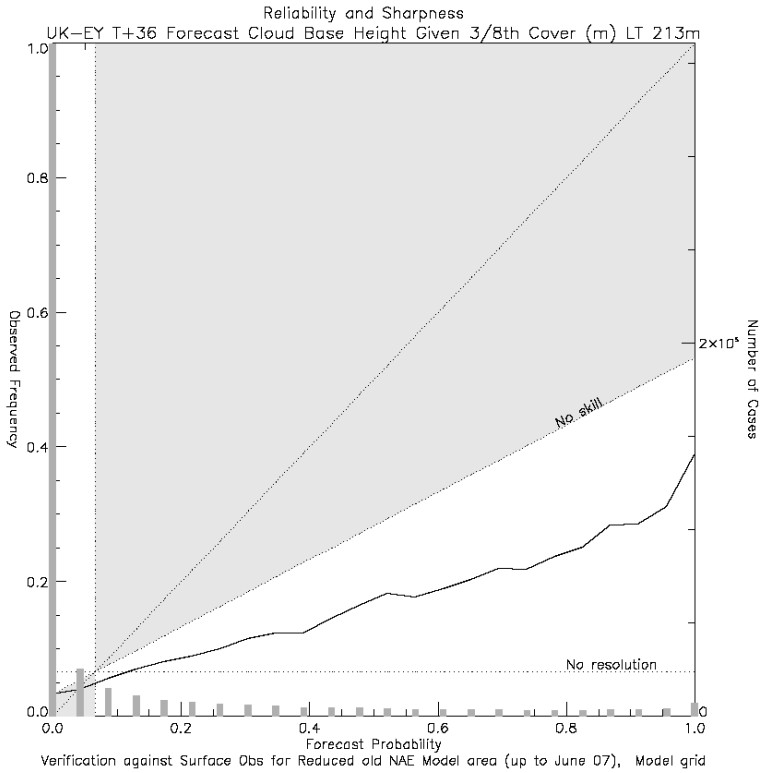


Figure 6.19, Attributes diagrams for T+36 forecast of Probability cloud base height will be less than 700 ft given 3/8 Cloud cover, verified over the reduced NAE model area.

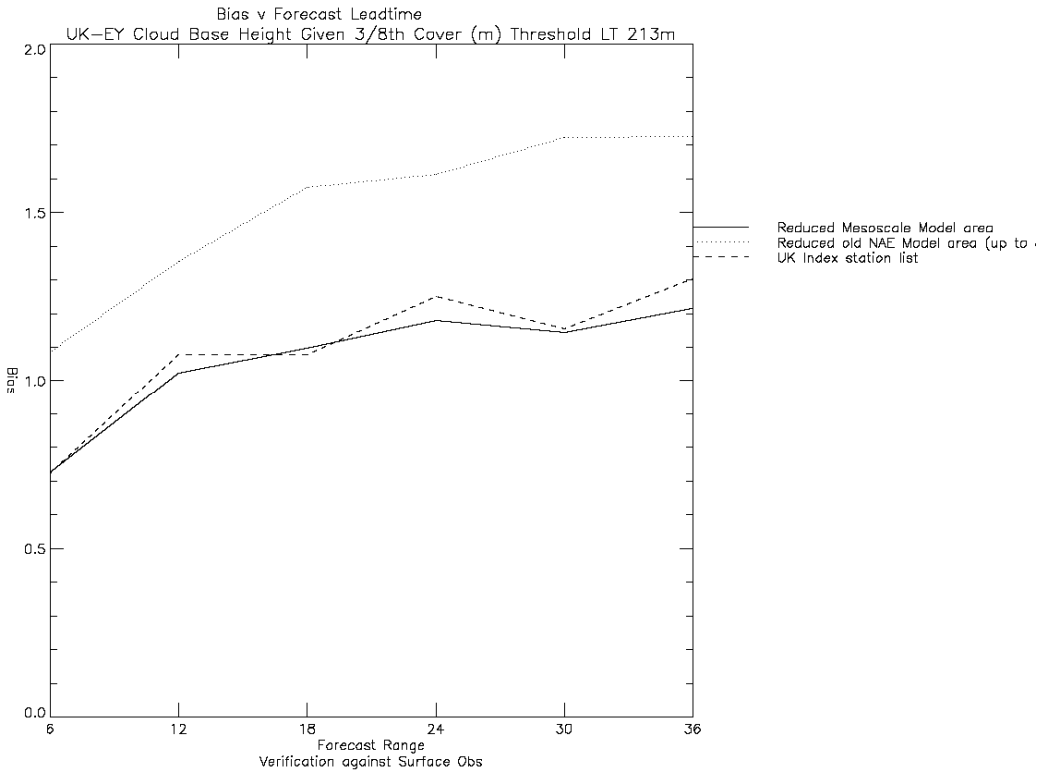


Figure 6.20, Bias against forecast lead time for forecast of cloud base height will be less than 700 ft given 3/8 Cloud cover.

6.1.5 Screen level Temperature Verification

It has been previously noted that verifying variables such as temperature over a large geographical area can introduce false skill. Therefore verification is presented for the UK station list only, acknowledging that even over this area there may be a degree of false skill attributed to the results. The SBV results have focussed on temperature thresholds greater than 10^{o}C and 15^{o}C and it should be noted that the under forecasting bias for the higher temperature thresholds associated with the climatological soil moisture is also apparent in results from the ABV (not shown).

Here we focus instead on results for temperature less than 5^{o}C and less than 0^{o}C for the period 1st January 2006 to 28th February 2007; attributes diagrams for these thresholds are presented in figures 6.21 and 6.22. Figure 6.21, indicates that the forecasts exhibit resolution but there is a conditional bias in the probability forecasts less than 0^{o}C, where there is an under forecasting of the lowest probabilities and an over forecasting of the higher probabilities. This is consistent with the ensemble being under-spread. Figure 6.22, presents a very good reliability curve at the less than 5^{o}C threshold, however, the ensemble is still slightly under-spread.

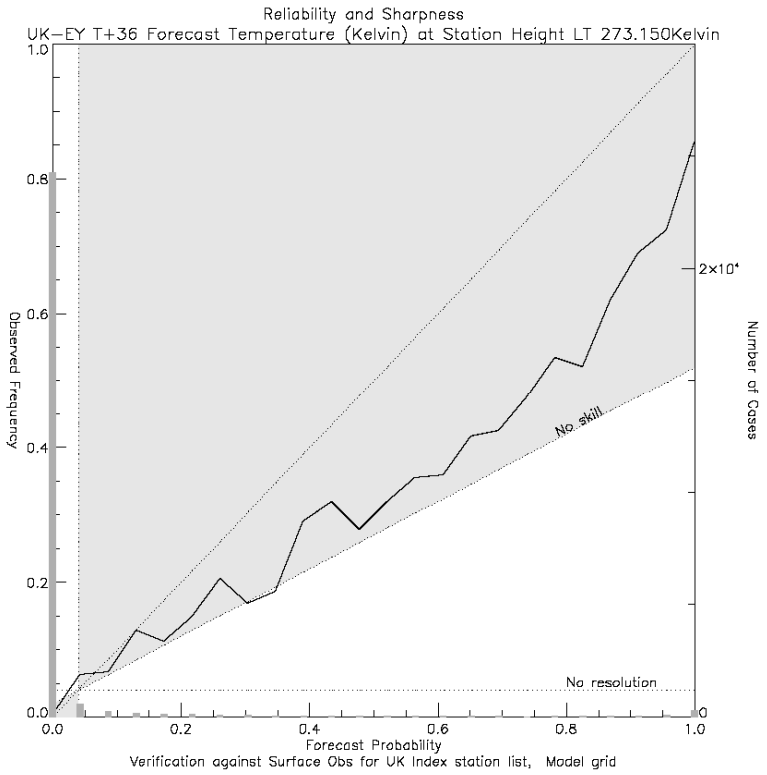


Figure 6.21, Attributes diagram for screen level temperature less than 0^{o}C verified against surface observations over the UK Index station list.

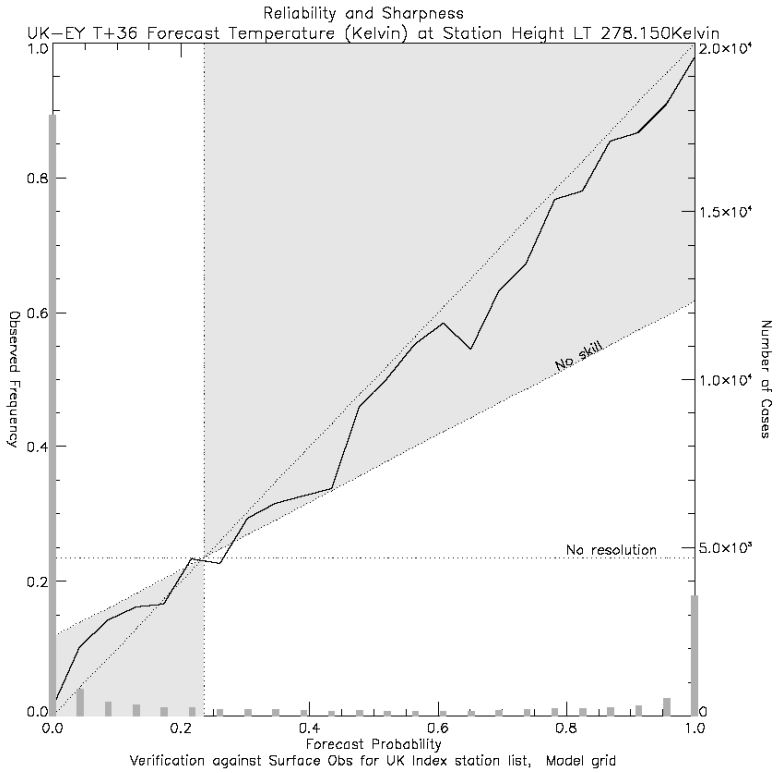


Figure 6.22, Attributes diagram for screen level temperature less than 5°C verified against surface observations over the UK Index station list.

6.2 Assessing ensemble performance using continuous statistics.

As well as considering the performance of probabilistic products issued from the MOGREPS suites it is also important to assess the performance of the ensemble as an ensemble system using measures such as the relationship between root mean square error (rmse) and the ensemble spread. RMSE can be calculated for the ensemble mean, in which case it is related to spread around the ensemble mean, or it can be calculated for the control and related to spread about the control. In an ideal system the spread of the ensemble about the ensemble mean will match the root mean square error of the ensemble mean forecast. In this section we consider the variables wind speed, geopotential height and temperature verified against radiosonde observations at 850 hPa, 500 hPa and 250 hPa pressure levels over the reduced NAE model domain. It is also important to consider the impact of observation errors on the verification results. The root mean square error of the forecast measured against the truth can be approximated by

$$rmse_{truth} = \sqrt{rmse_{forecast}^{\ 2} - rmse_{ob}^{\ 2}} \qquad (6.1)$$

Where $rmse_{forecast}$ is the root mean square error of the forecast measured against observations and $rmse_{ob}$ is the root mean square error associated with the observation. The root mean square error of the observations is estimated

using values specified within the OPS (Observation Processing System) in documentation on quality control (Ingleby,1998) and listed in Table 6.1.

Pressure Level	RMSE Wind (m/s)	RMSE Temperature (K)
250	2.85	1.2
500	1.85	0.8
850	1.60	0.8

Table 6.1, Observation errors for radio sonde observations taken from Ingleby (1998).

Accounting for observation errors can have a significant impact on the results. Whilst the figures presented here do not account for observation errors, we will consider them in the following discussion.

6.2.1 Geopotential Height

The root mean square error and spread in forecasts of geopotential height has traditionally been used as a summary measure of global ensemble performance. It is therefore appropriate that we first examine the performance of forecasts of geopotential height from the regional ensemble at the standard levels. Figures 6.23 to 6.25 show the root mean square spread and error against forecast lead time for forecasts of geopotential height at 250 hPa, 500 hPa and 850 hPa respectively. The root mean square error of the control and the ensemble mean are shown by the red and blue lines respectively, the spread of the ensemble about the control and ensemble mean forecasts are shown by the green and yellow lines respectively.

Simons and Hollingsworth (2002) estimated the root mean square error in the geopotential height at 500 hPa when measured by radio sondes to be of the order 10 m. Adjusting the root mean square error of the ensemble mean forecast for the effect of observation errors, the root mean square error becomes approximately 8 m at T+6 and 18 m at T+36 in figure 6.24. This clearly indicates that the ensemble is over spread in geopotential height at 500 hPa. There are no estimates available of the observation error in geopotential height at 250 hPa and 850hPa and it is therefore difficult to conclude what the effect of observation errors at these pressures will be. However, it is probably fair to conclude that the ensemble is over spread at all levels and that the effect is more severe at upper levels.

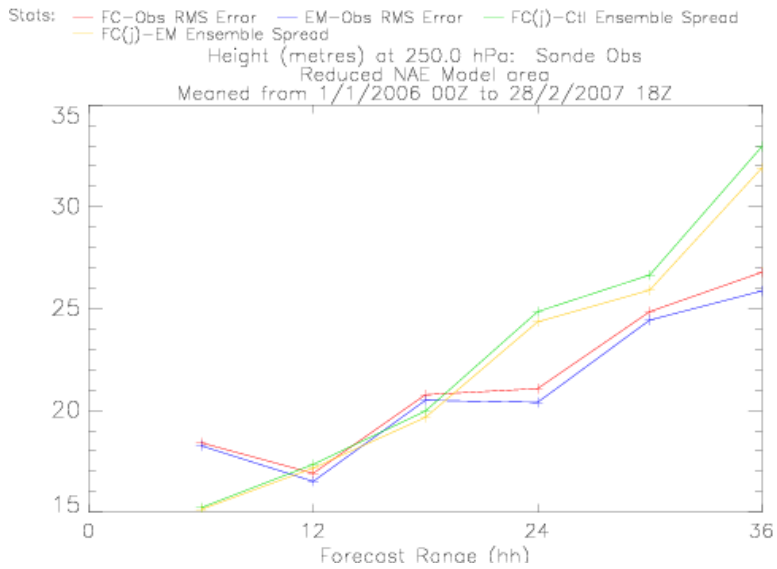


Figure 6.23, Root mean square spread and error against forecast lead time for forecasts of geopotential height 250 hPa. The root mean square error of the control and the ensemble mean are shown by the red and blue lines respectively, the spread of the ensemble about the control and ensemble mean forecasts are shown by the green and yellow lines respectively.

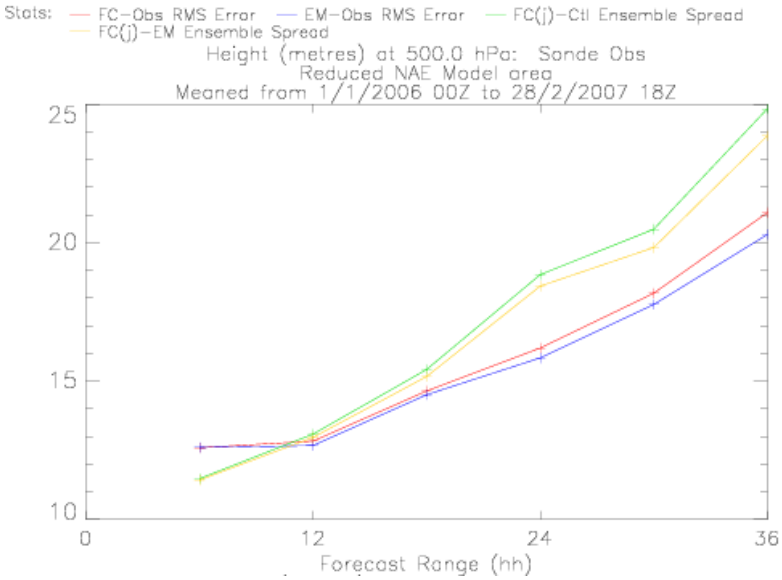


Figure 6.24, Root mean square spread and error against forecast lead time for forecasts of geopotential height at 500 hPa.

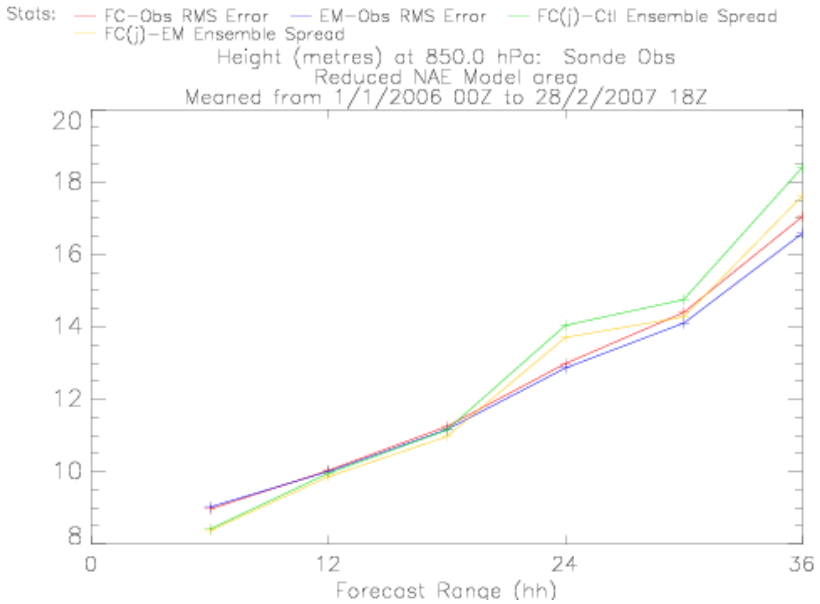


Figure 6.25, Root mean square spread and error against forecast lead time for forecasts of geopotential height at 850 hPa.

6.2.2 Temperature

Figures 6.26 to 6.28 present root mean square error and ensemble spread charts for temperature at 250 hPa, 500 hPa and 850 hPa respectively. At first glance it appears that the ensemble is under spread at 250hPa and 850hPa, with approximately the correct spread at 500hPa. When observation errors are considered we actually find the ensemble has approximately the correct spread at 850 hPa and that the ensemble is over spread at the upper levels.

6.2.3 Wind speed

Figures 6.29 to 6.32 present root mean square error and ensemble spread charts for wind speed at 250 hPa, 500 hPa, 850 hPa and station height respectively. At first glance it appears that the ensemble has the correct level of spread after T+24 in figure 6.29, however, when we account for observation errors in wind speed we find the ensemble is marginally over spread at 250hPa. Similarly at 500 hPa the ensemble appears to have approximately the correct level of spread after T+18. If observation errors are accounted for the ensemble has approximately the correct level of spread at T+6 and too much spread at T+36. Figures 6.31 and 6.32 appear to show that the ensemble is very under spread at all time ranges for wind speeds at 850hPa and at the surface respectively. We find that when observation errors are accounted for at 850hPa the ensemble has approximately the correct spread at T+36. The results in figure 6.32 are derived using surface observations and at station height the observation error is taken to be 1.7m/s (Ingleby, 1998). Correcting for observation errors using this value we find that the ensemble is still under spread at the surface at all time ranges.

To summarise the ensemble is under spread in wind speed at the surface, has approximately the correct spread at T+36 at 850hPa and too much spread at upper levels.

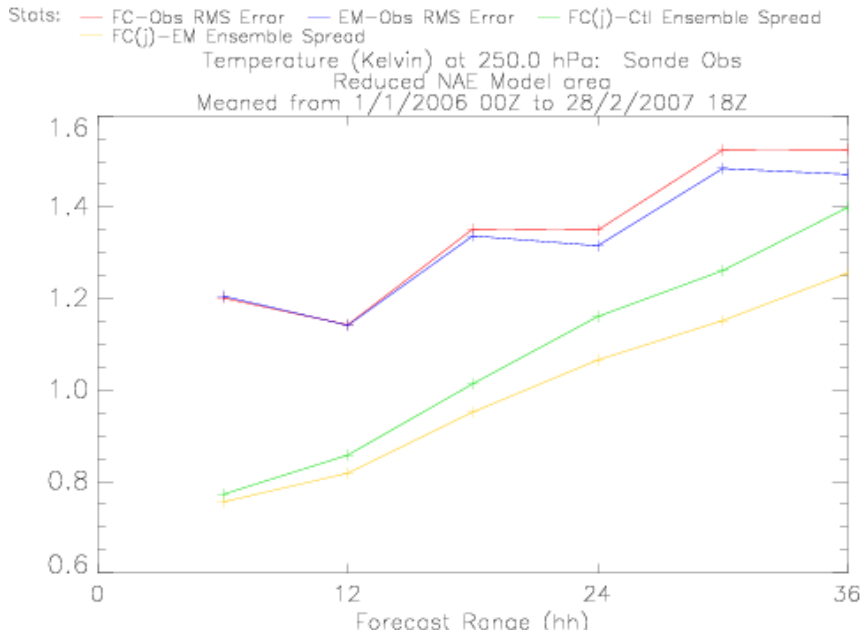


Figure 6.26, Root mean square spread and error against forecast lead time for forecasts of temperature at 250 hPa. The root mean square error about the control and the ensemble mean are shown by the red and blue lines respectively, the spread of the ensemble about the control and ensemble mean forecasts are shown by the green and yellow lines respectively.

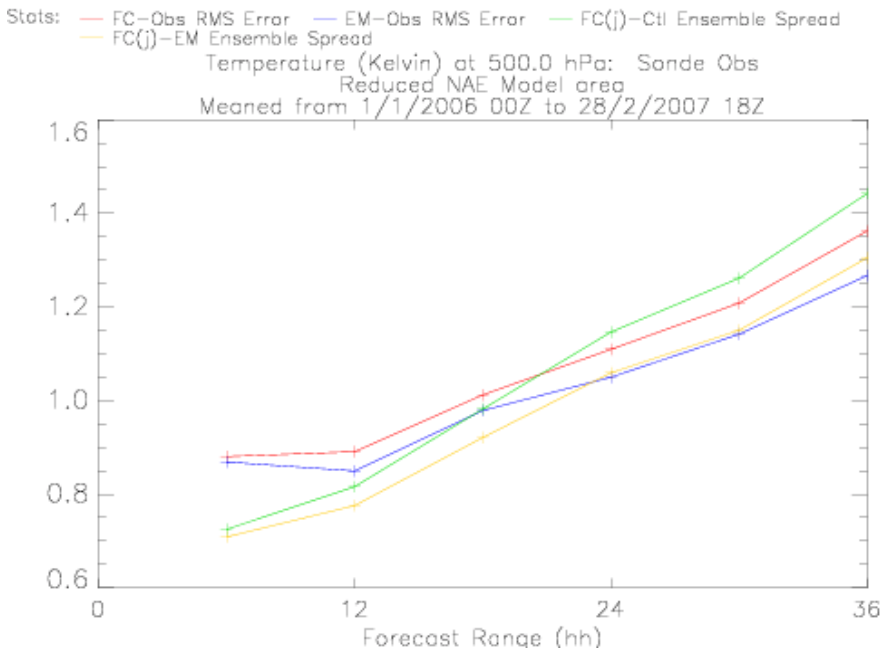


Figure 6.27, Root mean square spread and error against forecast lead time for forecasts of temperature at 500 hPa.

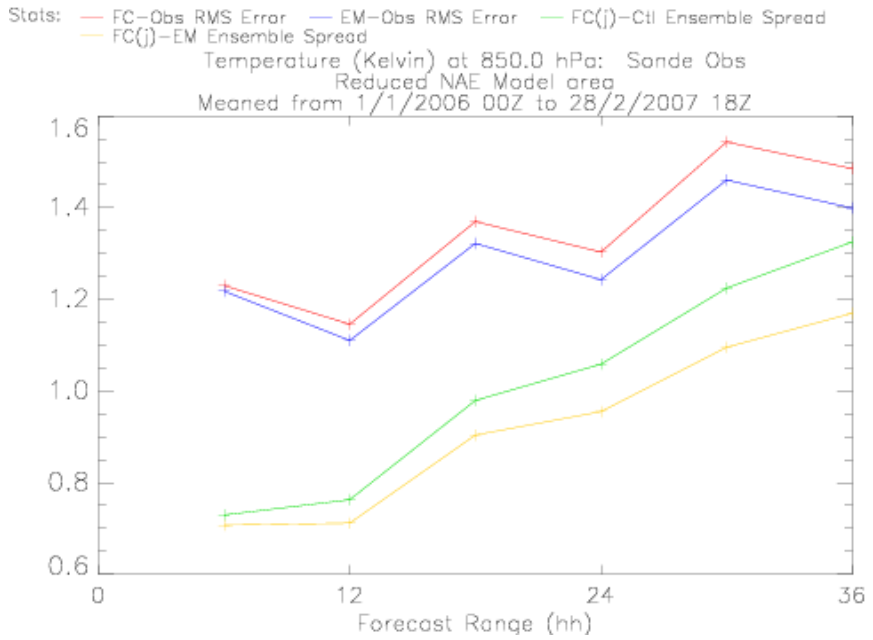


Figure 6.28, Root mean square spread and error against forecast lead time for forecasts of temperature at 850 hPa.

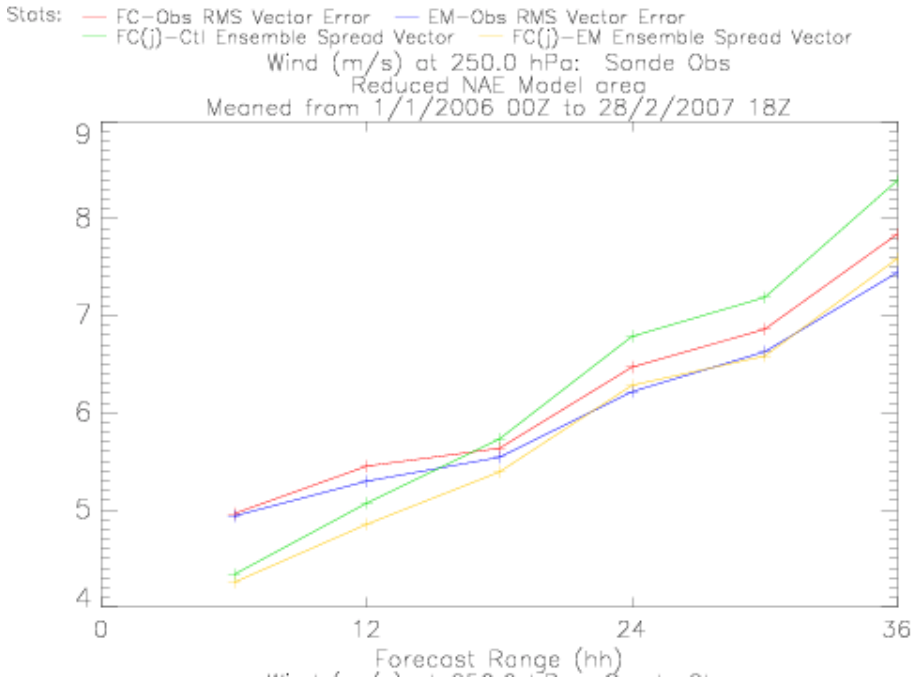


Figure 6.29, Root mean square spread and error against forecast lead time for forecasts of wind speed at 250 hPa.

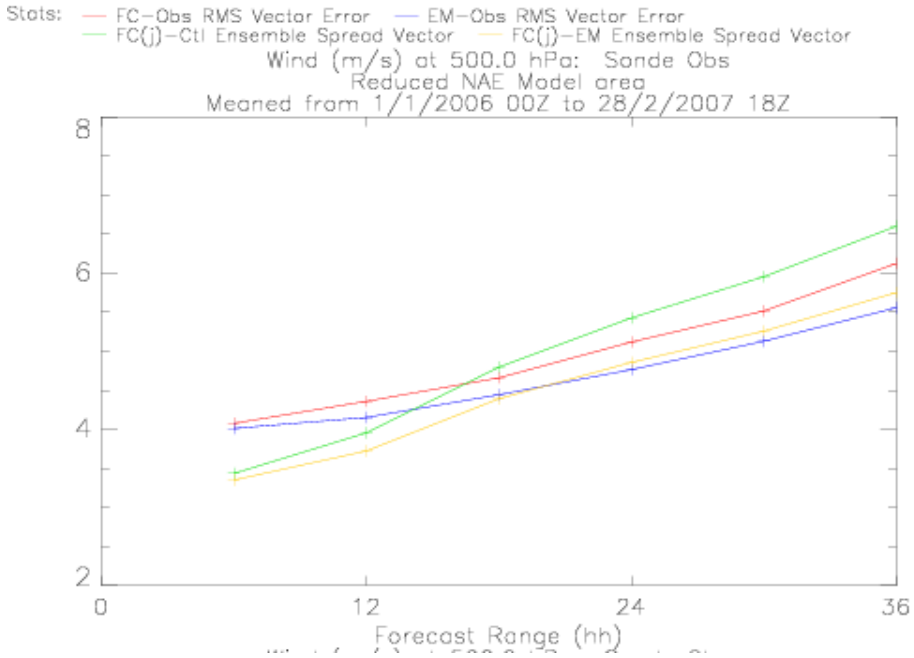


Figure 6.30, Root mean square spread and error against forecast lead time for forecasts of wind speed at 500 hPa.

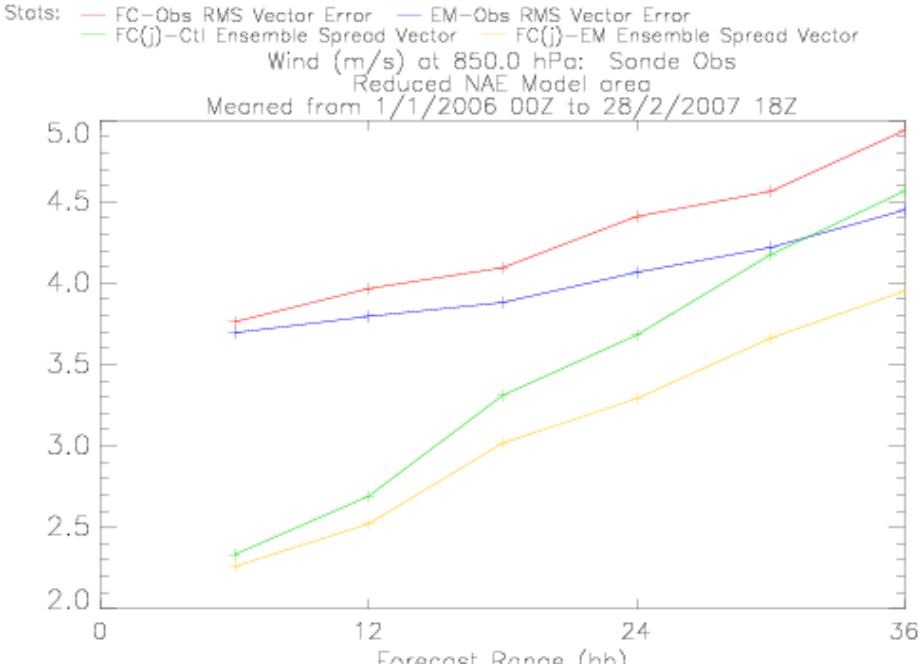


Figure 6.31, Root mean square spread and error against forecast lead time for forecasts of wind speed at 850 hPa.

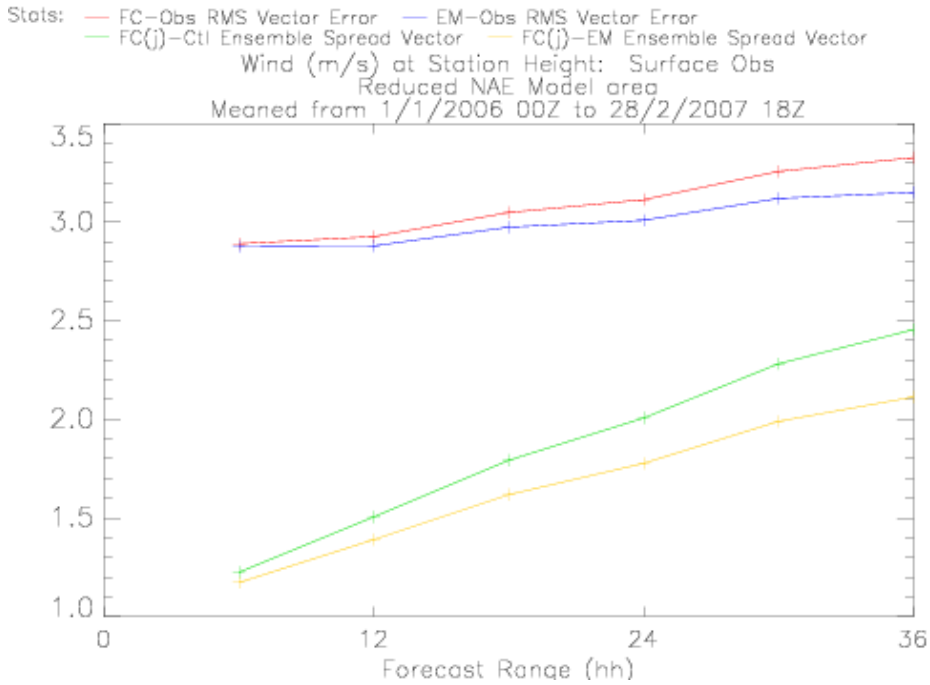


Figure 6.32, Root mean square spread and error against forecast lead time for forecasts of wind speed at Station height.

6.2.4 Summary

The results presented above indicate that the ensemble is over spread at 250 and 500hPa for temperature, wind speed and geopotential height and has approximately the correct level of spread at 850hPa in temperature and wind speed. This result is perhaps not surprising when we consider the perturbation strategy adopted in the regional ensemble. The perturbations that drive the regional ensemble are taken directly from the global ensemble and are reconfigured to the regional model resolution. However, because the analysis times of the two ensembles are offset by 6 hours and the perturbations are added over the first three hours of the forecast period, the perturbations are derived from T+7 forecasts from the global ensemble. Such perturbations to the control analysis are likely to be too large and cause the ensemble to be over spread. A new perturbation strategy, an Ensemble Transform Kalman Filter for the regional ensemble, will be adopted in the regional ensemble during May 2007. Results from trials of this approach (not shown) indicate that the spread of the regional ensemble is dramatically reduced compared to the current approach.

A common feature of all the results in figures 6.23 to 6.32 is that the rate of growth of ensemble spread appears to be larger than the rate of growth of forecast errors. This is unusual in that for most ensembles the rate of growth is too small, and this may suggest that the stochastic physics schemes employed in MOGREPS are particularly effective. This will need to be reviewed after the regional ETKF has been established, if possible taking account of observational errors, in order to review the strategy with regard to physics perturbations.

The under dispersion observed in the surface wind speed in the ensemble is also observed in the screen level temperature (not shown). This is an indication that the perturbation strategies used in the ensemble (initial condition perturbations and random parameters scheme) are failing to represent uncertainties in the surface processes. Further research into perturbations near the model surface, for example, soil moisture perturbations should be performed to tackle this issue.

7. Tropical cyclone track verification

7.1 Introduction

All tropical cyclone forecast tracks from the global deterministic model are verified against the observed tracks. The MOGREPS 15-day ensemble produces 24 tropical cyclone forecast tracks. The ensemble mean track is calculated for every forecast and these have been verified in the same way as the deterministic forecast tracks.

7.2 Verification

Ensemble mean tropical cyclone forecast tracks have been produced for most forecasts for the period 7th February to 6th April 2007. During this period there were 15 tropical cyclones; seven in the South-West Indian Ocean, seven in the Australian region (South-East Indian Ocean and South Pacific) and one in the North-West Pacific. These forecast tracks were verified and a homogeneous comparison made with the deterministic forecast tracks. The results are shown below:

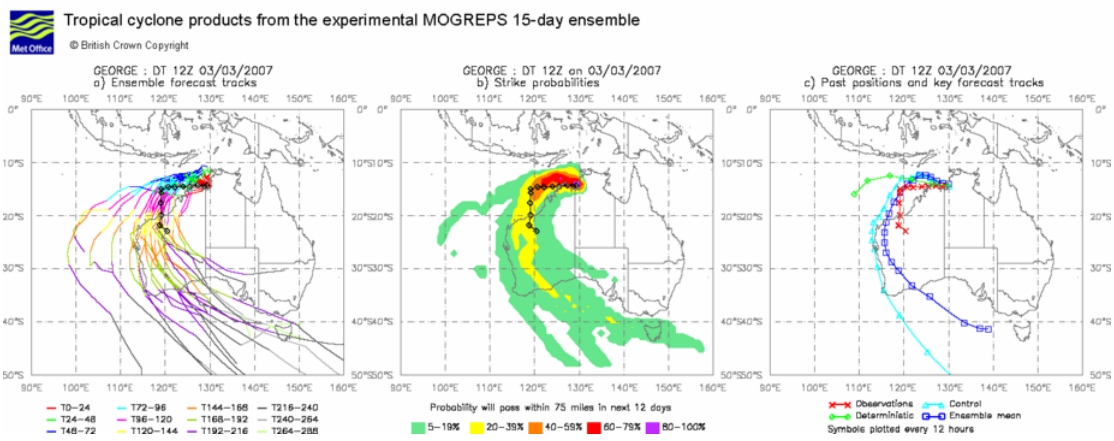
	T+0	T+24	T+48	T+72	T+96	T+120
No. of cases	100	79	55	34	20	11
Deterministic track error (km)	51	141	292	460	769	979
Ensemble mean track error (km)	36	130	286	437	609	611
Percentage reduction in error	29.4	7.8	2.1	4.9	20.8	37.6
Deterministic skill score (%)	-	29	30	19	-	-
Ensemble mean skill score (%)	-	36	32	20	-	-

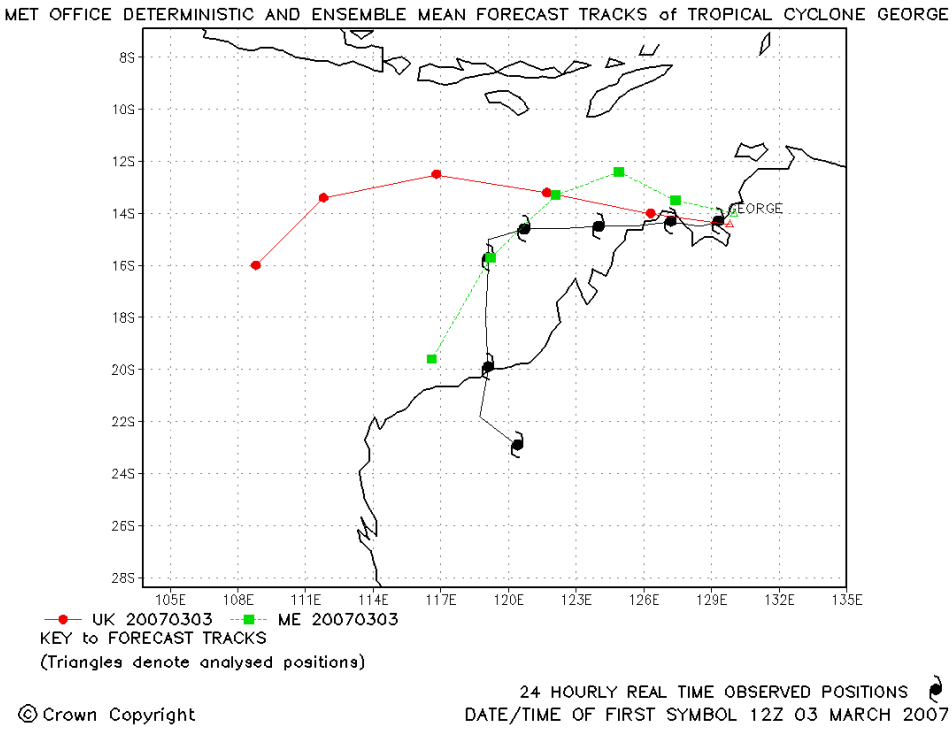
The skill score is the model's track error relative to that of a climatology/persistence model (CLIPER). A positive skill score indicates that the model performs better than CLIPER.
Skill is calculated thus:- (CLIPER error - Model error) / CLIPER error x 100%
Full details of the tropical cyclone verification method can be found here:
http://www.metoffice.gov.uk/weather/tropicalcyclone/method/index.html

These results indicate that the ensemble mean track errors were lower than the deterministic model at all lead times and by a wide margin at T+96 and T+120. However, at these longer lead times just 20 and 11 forecasts were verified. When averaged over all forecasts (199 cases, excluding analyses), the ensemble mean track error was 8.7% lower than the deterministic model error. Skill scores were higher for the ensemble mean by 7% at T+24 and 2% and 1% at T+48 and T+72.

7.3 Case Study

A closer examination of the data reveals that a major reason for the better performance of the ensemble mean over the deterministic forecast is the performance for Tropical Cyclone George. The mean improvement by the ensemble mean drops from 8.7% to 2.3% if George is excluded from the results. George formed over northern Australia and moved westwards before making a sharp turn towards the north-west Australian coast. This turn was not predicted by any deterministic models. Although the MOGREPS ensemble did not fully capture the leftwards turn. It gave sufficient indication for the ensemble mean to perform much better than the deterministic model. Charts below show the ensemble forecast tracks for 12Z 3rd March 2007 and a comparison of the deterministic track and the ensemble mean relative to the actual track.





7.4 Conclusion

Results from verification a small initial sample of tropical cyclone forecasts are promising, showing greater skill in the ensemble mean track compared to the deterministic forecasts. In the current small sample of cases most of this benefit is seen from a single storm which was poorly forecast by the deterministic model, but there is also some small benefit in other cases. The benefit in the case of the poorly forecast storm is valuable as it shows the potential to give some alert to areas at risk even in difficult forecast situations. Further cases are required in order to draw firmer conclusions on the overall benefit, and this data will be accumulated over coming tropical cyclone seasons.

8. Conclusions

We have assessed the performance of the MOGREPS ensemble system using a number of different methods.

Section 4 focussed on the comparison of the performance of the NAE and global MOGREPS ensembles with the ECMWF ensemble. In general the NAE ensemble performed better than the other two ensembles and the global ensemble performed worst. The NAE ensemble performed better than the other models for forecasts of light precipitation amounts, particularly in the summer. The performance of the NAE ensemble was also notably better than the other ensembles for forecasts of wind speed, particularly at force 5. After KFMOS post-processing the differences between the ensemble forecasts was much less, with differences rarely being statistically significant. For the post-

processed forecasts the ECMWF and NAE ensembles performed similarly, with each performing better for 2 out of the 4 thresholds considered.

In section 5 the results of an analysis of the spread-skill relationship in the NAE ensemble were presented. Such an analysis is not common-place in the ensemble forecasting literature, and hence the results are more difficult to interpret. The results indicate that the NAE ensemble has high levels of correlation between the spread and skill for wind speed, and for temperature in the winter-time. For summer-time temperatures the spread-skill relationship is much weaker.

In section 6, we saw that the NAE ensemble has a propensity to be over-spread, and that this is most prevalent at upper levels. We also saw that (when verified against up-scaled Nimrod analyses) the forecasts of light rain are very close to perfect reliability. For forecasts of higher rain amounts, or other variables, such as cloud cover and visibility, the reliability of the forecasts were less. This may be related to the difficulty the NAE model (and NWP models in general) has representing these kind of variables.

In section 7, the performance of the MOGREPS global system for tropical cyclone track forecasting was assessed for a small number of cases. The ensemble mean forecast on average provided a better track than the deterministic forecast. Most of this benefit is seen from a single storm which was poorly forecast by the deterministic model, but there is also some small benefit in other cases.

Overall, these results show that the MOGREPS ensembles are providing a useful contribution to Met Office forecasts. The NAE ensemble is at least competitive with the ECMWF ensemble, and on many occasions its performance is superior. This is a remarkable achievement for such a new system, and justifies the decision to implement MOGREPS operationally to meet customers' needs. Furthermore, there is a strong relationship between the spread and the skill of the ensemble for some variables. There still remain a number of areas where the MOGREPS ensembles can be improved, such as the excessive spread of the ensemble at upper levels, leading us to expect that the skill of the MOGREPS ensembles will increase in future.

References

Bishop, C. H., B.J. Etherton and S.J Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part 1: theoretical aspects, *Monthly Weather Review,* **129,** 420-436.

Hamill, T.M., and Juras, J., 2007: Measuring forecast skill: Is it real skill or is it the varying climatology. Submitted to *Quarterly Journal of Royal Meteorological Society.*

http://www.cdc.noaa.gov/people/tom.hamill/skill_overforecast_QJ_v2.pdf).

Houtekamer, P.L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.,* **121**, 1834-1846.

Ingleby B., 1998 Generic Quality Control OPS Scientific Documentation Paper 2. Crown Copyright.

Mason, S.J. and Graham, N.E., 1999: Conditional probabilities, relative operating characteristics and relative operating levels, *Weather and Forecasting*, **14**, 713-725.

Simmons AJ, Hollingsworth A (2002): Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal Of The Royal Meteorological Society* **128** (580): 647-677

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

Wang, X.G. and C.H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes, *Journal of the Atmospheric Sciences,* **60,** 1140-1158.

Watkin, H., N. Savage and R. Swinbank, 2007: Comparison between Met Office and ECMWF medium-range ensemble forecast systems, *Met Office internal project report.*

Wilks, D.S., 1995: Statistical methods in the atmospheric sciences, *Academic Press*, 263-267.

Wilson, L.J., 2000: Comments on “Probabilistic predictions of precipitation using the ECMWF ensemble prediction system”, *Weather and Forecasting*, **15**, 361-364.