Nat. Hazards Earth Syst. Sci., 8, 445–460, 2008
www.nat-hazards-earth-syst-sci.net/8/445/2008/

Natural Hazards
and Earth
System Sciences

Introducing uncertainty of radar-rainfall estimates to the verification of mesoscale model precipitation forecasts

M. P. Mittermaier

Mesoscale Model Development and Diagnostics Group, Met Office, UK

Received: 15 October 2007 – Revised: 18 January 2008 – Accepted: 6 March 2008 – Published: 8 May 2008

Abstract. A simple measure of the uncertainty associated with using radar-derived rainfall estimates as "truth" has been introduced to the Numerical Weather Prediction (NWP) verification process to assess the effect on forecast skill and errors. Deterministic precipitation forecasts from the mesoscale version of the UK Met Office Unified Model for a two-day high-impact event and for a month were verified at the daily and six-hourly time scale using a spatially-based intensity-scale method and various traditional skill scores such as the Equitable Threat Score (ETS) and log-odds ratio. Radar-rainfall accumulations from the UK Nimrod radar-composite were used.

The results show that the inclusion of uncertainty has some effect, shifting the forecast errors and skill. The study also allowed for the comparison of results from the intensity-scale method and traditional skill scores. It showed that the two methods complement each other, one detailing the scale and rainfall accumulation thresholds where the errors occur, the other showing how skillful the forecast is. It was also found that for the six-hourly forecasts the error distributions remain similar with forecast lead time but skill decreases. This highlights the difference between forecast error and forecast skill, and that they are not necessarily the same.

1 Introduction

Precipitation estimates from weather radar provide the most detailed information regarding the complex spatial and temporal distribution of precipitation. Despite continued development and improvement in estimation methods, it is widely acknowledged that errors in the rainfall estimates can still be a factor of two (−50%/+100%) (e.g. Joss and Waldvogel, 1990), and this is *after* bias correction using rain-gauge data.

Among the radar community the difficulties of quantitative precipitation estimation are well documented. Besides instrument limitations and calibration issues, estimates are affected by the vertical profile of reflectivity, low-level orographic growth, anomalous propagation, ground clutter, attenuation or the presence of other spurious echos (e.g. Joss and Waldvogel, 1990; Harrison et al., 2000). Corrections for these must be applied to improve data quality. Yet other problems remain. For example, radar may under-estimate precipitation because the droplet size of drizzle or light rain drops is too small and therefore below the detection threshold (Golding, 2000). Alternatively, the radar may simply not be able to see all the precipitation, due to partial blocking or the beam over-shooting shallow precipitation at long ranges. Rain measured aloft may evaporate before reaching the ground, leading to an over-estimation. Wind drift of precipitation particles may also lead to displacement errors (Mittermaier et al., 2004). This is particularly relevant if radar measurements above the freezing level (bright band) are used for estimating precipitation at the ground. Operationally data quality may also suffer due to missing data.

Rain gauges remain the main source of precipitation observations, along with radar and satellites. Rain gauges have good time resolution and provide an accurate estimate of ground truth at specific locations. However, rain gauge networks are often very sparse and unevenly distributed in space. Radar and satellite observations have good spatial coverage, yet the precipitation estimate at the ground obtained from these measurements is indirectly inferred from radar reflectivity or brightness temperature respectively. The temporal sampling of satellites in particular can make the use of satellite precipitation unsuitable for the verification of individual forecasts from high-resolution models. Despite the uncertainties and errors associated with radar-rainfall estimates, they are being used as "truth" for the verification of mesoscale models, *without* taking the errors of the estimate into account. This is especially necessary as the horizontal

Correspondence to: M. P. Mittermaier
(marion.mittermaier@metoffice.gov.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

resolution of mesoscale models today require a verifying data set at resolutions finer than what most rain-gauge networks can provide. When it comes to the verification of extreme events, radar alone can provide the kind of detail that is required.

Traditional verification scores do not fully account for the unique characteristics of precipitation. For example, the root-mean-squared error (RMSE) is sensitive to discontinuities, noise and outliers, thus skewing or distorting the results. Moreover, verification scores for continuous forecasts do not account for the complex spatial interdependency of precipitation values. Categorical verification scores, which utilize thresholds are generally better suited to cope with precipitation fields (circumventing the problem of outlier contamination), and are more widely used for quantitative precipitation forecast (QPF) verification. Unfortunately many of these scores are overly sensitive to the base rate (frequency of occurrence) of an event and to the frequency bias (ratio of the observed to the predicted events). Göber et al. (2004) show that forecasts of extreme or rare events have more skill than those of relatively “normal” events, by using scores not sensitive to the base rate. They also state that recent trends due to model improvements show a small increase in skill but a large reduction in model bias for forecasts of light precipitation.

Recently developed verification techniques (e.g. Briggs and Levine, 1997; Zepeda-Arce and Foufoula-Georgiou, 2000) aim to provide more informative feedback on some of the physical aspects of the forecast error, by verifying at different spatial scales. In doing so, there is the recognition that precipitation phenomena occur on different scales (e.g. showers or fronts) and are driven by different physical processes (e.g. convection or large scale synoptic processes). This approach can help determine which model processes need further development. Casati et al. (2004) describe an intensity-scale verification method which uses wavelet decomposition. The technique allows the differences between the forecast and the observation to be diagnosed as a function of the spatial scale of the error and intensity of the precipitation events. This enables a separate evaluation of mesoscale and convective features (such as fronts or convective cells), drizzle and intense events. Mittermaier (2006) devised a non-parametric method for aggregating individual intensity-scale decompositions for compiling longer-term statistics. The intensity-scale method has been further expanded by Casati and Wilson (2007) to decompose the Brier score. Another approach is described by Roberts and Lean (2008), particularly addressing the needs for presenting results from very high-resolution model runs at 1, 2 and 4 km. Roberts and Lean introduce a fractional exceedance where a probability is calculated based on the number of model grid points with precipitation exceeding a given threshold, in a predefined area. By performing this analysis for predefined areas of different sizes, the optimum product presentation grid length can be determined.

Turner et al. (2004) discuss methods used for improving nowcasts of precipitation through the filtering of non-predictable scales of precipitation and how this improves skill. Seed (2003) also uses the fact that the lifetime of a precipitation pattern is dependent on the spatial scale of the pattern, linking the lifetime of the event to its predictability, or rather the lack of predictability of small-scale. Fast evolving features substantiate the need for averaging model output to a scale coarser than the model resolution to obtain the best forecast. This ties in with the idea of Roberts and Lean (2008) in attempting to find the optimum scale for presenting model output.

Crucially, the nowcasting and NWP community have converged on the use of radar data and the understanding that small-scale features are less predictable. Both have recognized that a temporally varying optimal averaging length exists where the forecast accuracy and skill are maximized, and the error minimized. Yet, the uncertainty in the radar products used has not been incorporated thus far. Fortunately this is changing. Several different approaches for incorporating uncertainty are emerging. For downstream hydrological applications, radar-generated ensembles are being explored by Germann et al. (2006) and Bowler et al. (2006). An attempt is made at quantifying the uncertainties in motion and evolution of radar-derived rainfall products so that the actual radar fields can be stochastically perturbed, and an ensemble of realizations produced. These are examples of how uncertainties in the radar measurements can be incorporated in “real-time”. Verification by contrast is very much an “after the event” exercise.

Quantifying the uncertainty in the observations used to verify NWP model forecasts has gained greater prominence in recent years. For example Bowler (2006) has used a deconvolution method to consider the impact on deterministic categorical scores whilst Saetra et al. (2004) considered the impact on ensemble probabilistic forecast verification measures. When using radar-rainfall estimates for verifying precipitation forecasts, two additional options come to mind: applying an error bound (or function) to the accumulation field (the simplest, computationally least expensive option), or generating multiple realizations of the accumulation field by adding error-bounded “noise” to the accumulation field.

In this paper, as a first attempt, the simplest but well-established constant factor-of-two error associated with radar-rainfall estimates is added to the verification process. The impact is assessed using the intensity-scale method introduced by Casati et al. (2004) and various more traditional measures of skill. A brief summary of the intensity-scale method is given in Sect. 2 together with the method for including the radar-rainfall estimate error. This is followed by a description of the model fields and the radar data using in subsequent sections. In Sect. 3 a two-day heavy rainfall event during June 2004 is used to assess the impact on individual forecasts and more extreme events. Both the daily and six-hourly accumulations are assessed. The month of June 2004 is assessed as a whole in Sect. 4. A summary and concluding remarks are presented in Sect. 5. A basic introduction to categorical statistics is provided in Appendix A.

2 Method

2.1 Intensity-scale method

For a complete description of the method the reader is referred to Casati et al. (2004). Some notable differences between the method proposed by Casati et al. and the method used here, include the following: no dithering step is performed because all values included in the error analysis are floating-point numbers; and secondly no forecast recalibration was performed. The analysis is performed on a 2^7 by 2^7 spatial domain, with L=7.

Thresholding is used to convert the forecast (Y) and analysis (X) into binary images. In line with the original method, thresholds are factors of two beginning with one-eighth of a millimetre up to 128 mm (this is done to achieve log-normality). The difference between the binary forecast and analysis defines the *binary error* Z=I_Y−I_X. The binary error image is then decomposed and expressed as the sum of components on different spatial scales by performing a two-dimensional discrete Haar wavelet decomposition. Alternatively the two-dimensional discrete Haar wavelet decomposition can also be obtained more simply by averaging over square regions on different scales. The mean-squared error (MSE) of the binary error image is given by the average of all the differences over all the pixels in the domain. Therefore the MSE of the binary error image is equal to

$$\mathrm{MSE} = \sum_{\mathrm{l}=1}^{\mathrm{L}} \mathrm{MSE_l}, \tag{1}$$

where $\mathrm{MSE_l}=\overline{\mathrm{Z}_\mathrm{l}^2}$ is the MSE of the l-th spatial scale component of the binary error image. For each precipitation rate threshold, the binary MSE skill score SS can be calculated, relative to the MSE of a random forecast:

$$SS = 1 - \frac{\mathrm{MSE}}{B\,\epsilon(1-\epsilon) + \epsilon(1-B\epsilon)} \tag{2}$$

where the denominator represents the random MSE calculated from the base rate ϵ (observed frequency of occurrence) and the bias B of the sample at a given threshold. As with most skill scores a perfect forecast has a value of 1. When the score is zero it means that the forecast is no better than a random forecast. Negative values imply that the model is worse than the random forecast, in terms of the MSE, although this does not necessarily mean that the model forecast has no skill. As shall be seen in following sections, the horizontal scale where the error is "eliminated" is often of such a magnitude that it could be argued whether a forecast averaged to such a length scale would be *useful*. It is worth noting that the intensity-scale method is not a tool to show strength of skill. It merely shows that when the SS is positive, it *is* skillful, but not *how* skillful. Therefore the outcome of this method is strongly biased towards understanding the error and identifying the source through the scale and intensity of where, and when it occurs. It is a diagnostic tool.

2.2 Introducing uncertainty

The thresholding process produces binary fields I_X and I_Y for each of the N accumulation thresholds u=u_1, ..., u_N for n=1, ..., N as shown in Eq. (3).

$$I_Y = \begin{cases} 1 & \text{if } Y > u_n \\ 0 & \text{if } Y \le u_n \end{cases} \quad I_X = \begin{cases} 1 & \text{if } X > u_n \\ 0 & \text{if } X \le u_n \end{cases} \tag{3}$$

The +100% error (representing the potential over-estimation) is implicitly included as the thresholding is open-ended. However, no effort is made to account for the possible under-estimation of precipitation by radar. This could be important as, given the spatial coverage of radars, they provide the best means of testing the rain-no rain boundary. As mentioned in the Sect. 1, there is a lingering question of whether it is the model that produces too large areas of drizzle and light rain, or whether the radar, at longer ranges (due to beam over-shooting), or due to droplet size, fails to detect this very light rain and drizzle. To address this potential underestimation, we can implement a −50% uncertainty, using "lagged" thresholding. So now Eq. (3) applies only to the first threshold u_1. For all subsequent thresholds:

$$I_Y = \begin{cases} 1 & \text{if } Y > u_n \\ 0 & \text{if } Y \le u_n \end{cases} \quad I_X = \begin{cases} 1 & \text{if } X > u_{n-1} \\ 0 & \text{if } X \le u_{n-1} \end{cases} \tag{4}$$

This is illustrated in Fig. 1. This simple implementation is illustrated using the power-of-two thresholding sequence required to achieve log-normality. Thus the previous threshold is half of the current one. Using this "lagged" threshold should enlarge the sample size at the higher rainfall thresholds. As the model tends to produce larger precipitation areas (especially at lower thresholds) than observed by radar, this increase in the radar area may possibly favour the model's skill. This aspect is considered in the analysis that follows. Given the non-linearity of all forecasting systems, the introduction of uncertainty should not affect the *skill* or the *errors* in a uniform manner. Some descriptive measures such as the bias may show some systematic trends.

2.3 Description of model output

The model forecasts evaluated in this study are 24-h or 6-hourly precipitation accumulations of the mesoscale (MES) version of the Met Office Unified Model (UM) spanning the first 24 h of the forecast (0–24 h). The UM provides a seamless nested forecasting system that can be run at multiple resolutions. All resolutions share the same dynamic core and (relevant) parameterizations. The model is non-hydrostatic

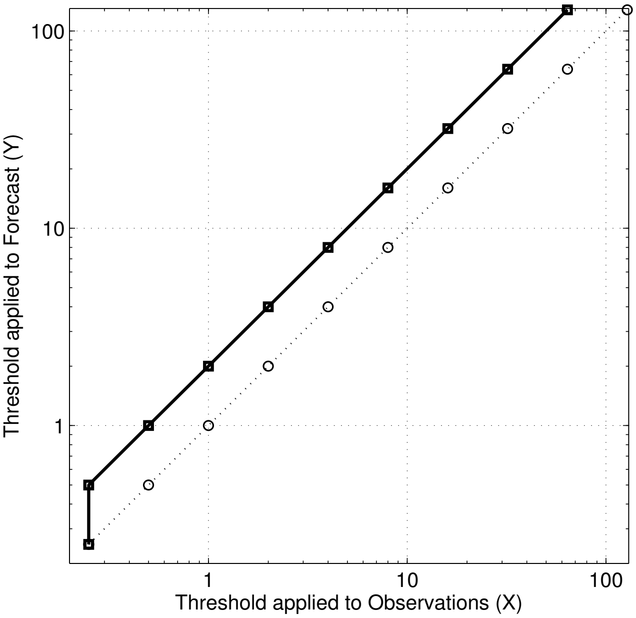

Fig. 1. Schematic showing how the lagged thresholding described in Eqs. (3) and (4) is used to introduce uncertainty. The solid line refers to thresholds for this study, whereas the dashed line shows when equal thresholds are applied.

and is based on semi-implicit, semi-Lagrangian numerics. At this time it was run with 38 levels in the vertical, of which 13 are in the lowest 5 km. It uses terrain-following co-ordinates. Further details of the model can be found in Davies et al. (2005).

It is recognized that using 24-h accumulations may damp timing errors at shorter time scales, such as the time of a frontal passage at a given location. Therefore the four six-hourly accumulations spanning the same 24-h period are also studied. The model in its current operational configuration has a 0.11° resolution which translates to around 12 km over the UK. The precipitation accumulation is the sum of the convective and large scale, liquid and snow amounts. The model is run four times a day but only the 12:00 UTC run was evaluated in this study.

2.4 Radar data

NIMROD is an automated short-range mesoscale nowcasting system used operationally at the UK Met Office (Golding, 1998). NIMROD produces hourly precipitation rate forecasts and analyses with a resolution of 5 km, every 15 min, up to 6 h ahead. The precipitation rate nowcast are produced using radar and satellite data, along with surface observations. In this study stand-alone deterministic forecasts are being verified against the NIMROD baseline product which is the quality-controlled UK radar composite. Various near-continent radars are also part of the composite but these were excluded from this study. The 5 km radar data are averaged onto the MES grid, and given the radar coverage, comparisons can only be made where radar data are available.

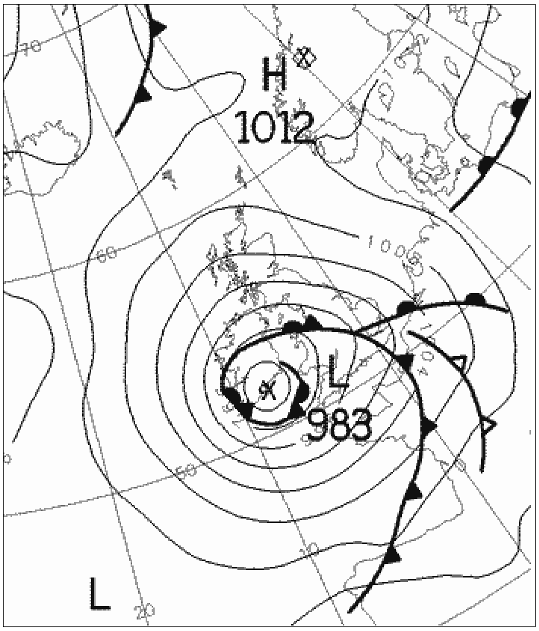

Fig. 2. The analysis field from the global UM at 00:00 UTC on 23 June 2004.

3 A heavy precipitation event case study

The main precipitation event for the month of June 2004 occurred on the 22nd and 23rd when a deep depression tracked across Wales and northern England. The central pressure at 06:00 UTC on the 23rd was 982 hPa, making it one of the deepest depressions recorded in June over England and Wales. Many parts of south-western, southern and central England and Wales received over 25 mm of rain in 24 h, based on the NIMROD radar-rainfall product. On the 23rd gale force winds along the English Channel coast gave gusts in excess of 50 knots. The low passed over the country from south-west to north-east eventually tracking over the North Sea on the 24th but not before giving another 50 mm (radar) of rain over north-eastern England and gusts of 40–50 knots over eastern England. Figure 2 shows the midnight analysis on the 23rd from the Unified global model, the position of the low centre just south of Ireland.

Rainfall was concentrated more in the south and west on the 22nd, shifting to the central and northern parts on the 23rd. The twenty-four hour rainfall accumulations shown in Fig. 3 are for the 23rd. Accumulations for the two days ranged between 25–50 mm for most parts. As described in Sect. 2.4 the MES fields are masked using the area covered by radar, with non-UK radars excluded.

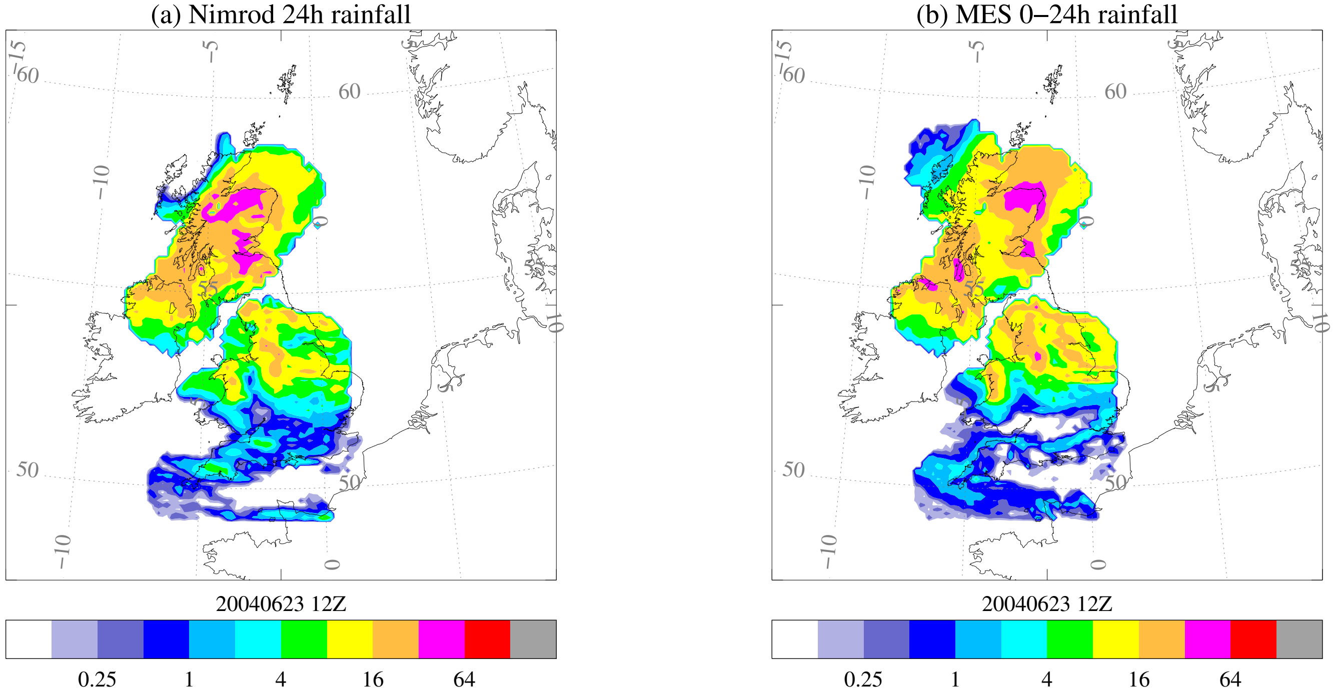

Fig. 3. Twenty-four hour rainfall accumulations from 12:00 UTC to 12:00 UTC from the radar composites **(a)** and the 12:00 UTC MES run **(b)**.

3.1 Event analysis

By inspection of Fig. 3 it can be seen that the UM predicted the general pattern of precipitation correctly but appears to over-estimate rainfall exceeding 16 mm.d^{-1}. The areas with most intense rainfall (>32 mm.d^{-1}) over Scotland are placed further to the east by the model.

The forecasts were analyzed with a selection of traditional measures of skill (see Appendix A). A slightly different set of rainfall thresholds with more thresholds between 4–48 mm were used. For robustness, a minimum sample size of 5 was set for each of the contingency table entries. The Equitable Threat Score (ETS) was used, along with ROC-like (relative operating characteristic) curves, the frequency bias and the logarithm of the odds ratio. Note that ROC is usually used for probabilistic forecast verification, plotting the skill of different probabilities for a single threshold. Here we are plotting the hit rate versus the false alarm rate as a function of thresholds (in the strictest sense there is only one point on a ROC plot for a deterministic forecast as there is only one outcome). It must be kept in mind that a "0" (such as may be the case for individual forecasts) in any entry of the contingency table means that the odds ratio is undefined and inappropriate for use (e.g. Jolliffe and Stephenson, 2003; Wilks, 2006). At the monthly time scale this may not be an issue because there should be sufficient spread in the sample. A perfect forecast would score an ETS of 1. A no-skill forecast has an ETS of zero. The ETS can be skillful when negative; it implies that the values in the contingency table are reversed. For the ROC curve the balance is between maximizing the hit rate and minimizing the false alarm rate per threshold. Values of the log-odds ratio over 1 imply skill. The log-odds ratio has been shown to be more useful for verifying extreme events (Göber et al., 2004) and is also preferable because error intervals can easily be calculated and plotted. The frequency bias on the other hand is not a measure of skill but a descriptive measure that will be equal to 1 if the forecasts of an event occur as frequently as they are observed (base rate).

The various measures were calculated for model and radar fields averaged to twice the grid length, i.e. ~25 km to eliminate most grid-scale errors as suggested by several recent studies (see e.g. Mittermaier, 2006; Vasić et al., 2007). Figure 4 shows various skill scores for the event. All the different scores appear to agree that the 4 mm.d^{-1} threshold is the most skillful. This is where the bias in panel (a) appears to be closest to 1 and appears to be fairly consistent for all rainfall accumulations. The bias may increase with threshold if it is no longer dominated by the hits but by the misses and false alarms. The bias becomes much larger than 1 when the false alarms increase disproportionately relative to the misses. The introduction of uncertainty has a large and reverse impact on the bias, as seen in panel (a). Whilst the sum of the hits and the false alarms remains constant the base rate increases, with the net result a decrease in the bias as the accumulations increase. Introducing a lagged threshold (which increases the radar area, base rate) causes the misses and false alarms to trade places (but not necessarily in proportion).

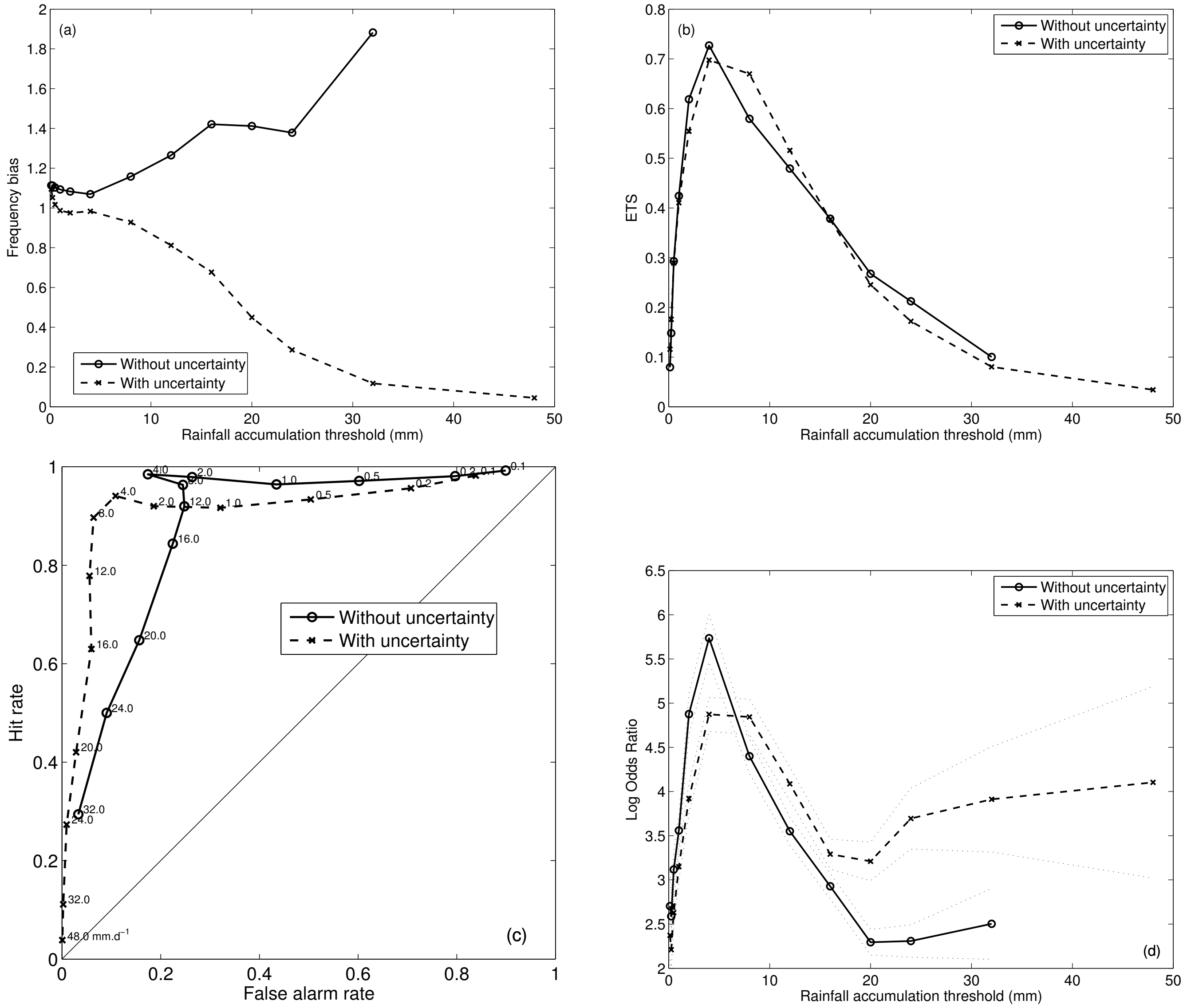

Fig. 4. Traditional measures of skill for the event 22–23 June 2004, showing results for uncertainty included and without: **(a)** Bias, **(b)** ETS, **(c)** ROC and **(d)** log odds ratio ± log-odds error.

The ETS shown in panel (b) on the other hand shows a decrease of up to 0.1 for thresholds between 1–4 mm; forecasts with accumulations greater than 8 mm.d^{-1} are the most skillful. The interesting result comes from the ROC curve in panel (c). The false alarm rate is reduced for all thresholds, and the hit rate is lower for all thresholds. The log-odds ratio in panel (d) shows that the forecast is skillful at all thresholds, being above 1. The odds-ratio shows a similar reduction in skill to the ETS. The log-odds ratio shows an increase in skill whereas the ETS produces a rather low score of 0.4 or less. Introducing uncertainty appears to improve the skill according to the log-odds ratio for thresholds greater than 8 mm.d^{-1}. This must be linked to the reduction of the false alarm rate in panel (c).

An intensity-scale representation of the MSE reveals that error distribution patterns can vary greatly. Errors can be spread over a range of rainfall thresholds but contained to, say, the grid scale or to a multiple of the grid scale, yielding a relatively “flat” error distribution. This tends to suggest that there are few timing or displacement problems and that the overall distribution of the precipitation pattern is well captured. The differences are small-scale and localized. Alternatively similar magnitude errors at the grid scale may be found at larger scales but only selected thresholds. This is a recognizable signature for precipitation features that are displaced either linearly or rotated (in case of fronts that have a different orientation in the forecast than observed). The error “propagation” is a function of the underlying method of

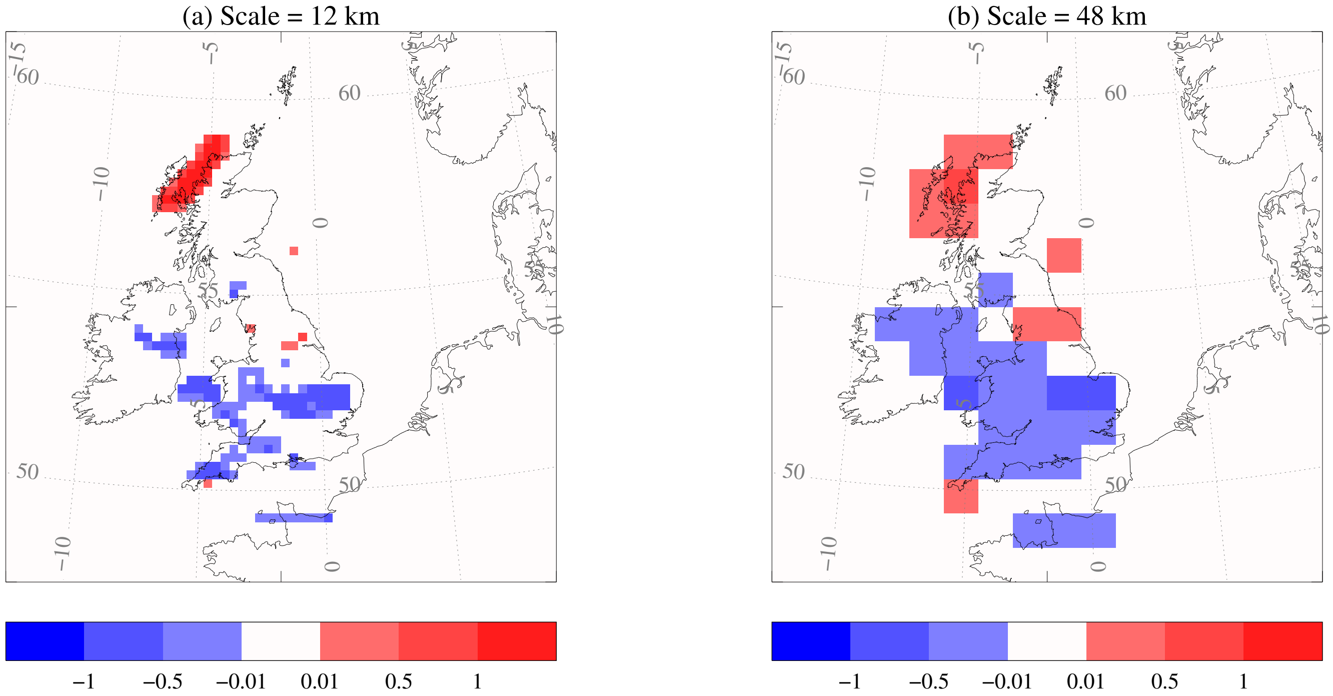

Fig. 5. Mother-wavelet components for the 4 mm/day binary error image ($Z{=}I_Y{-}I_X$) obtained from the accumulations in Fig. 3 at **(a)** 12 km and **(b)** 48 km. The binary error image is equal to the sum of the mother wavelets over all spatial scales.

averaging to a coarser resolution. The error will be found at this coarser resolution so long as the differences between the observed and model forecasts remain large enough.

Before analysing the intensity-scale diagrams for this event, consider first the decomposition of the binary error at the 4 mm threshold at two different scales, as shown in Fig. 5. These mother-wavelet plots correspond to the accumulations shown in Fig. 3. The MSE components capture the main differences between the forecast and the observed: there is too little rain over the south in the forecast (negative differences) and the larger accumulations in the north-west are not observed (positive differences).

The intensity-scale diagrams in Fig. 6 suggest that the forecast for the event ought to have been quite skillful (as the scores in Fig. 4 show). The intensity-scale diagrams indicate that the largest errors occur at the grid scale, for the largest rainfall accumulations. Most of the contours are clustered in the bottom-right hand corner suggesting that even averaging to twice the grid length would reduce the error considerably. It should be noted that averaging alone can not eliminate the errors entirely, but acts to minimize it. Some of the errors are still evident at length scales of 96–192 km for rainfall accumulations of 32 mm.d^{-1} or more, which relates to the fact that the most intense rainfall was displaced. The introduction of uncertainty as shown in Fig. 6b has marginally worsened the error at the grid scale and increased the length scale of the worst errors, bearing in mind that the threshold range of the analysis has been extended through the lagged thresholding.

3.2 Six-hourly analysis

The traditional scores for the six-hourly accumulations are shown in Fig. 7. The decrease in the bias with threshold when including uncertainty is still evident when comparing panels (a) and (e). The 12–18 h bias is the most consistently good over all thresholds. The decrease in forecast skill with lead time is evident in all the panels, and the skill is generally lower than for the daily totals. The best ETS was greater than 0.7 for the daily accumulations, but less than 0.6 for the six-hourly totals, indicating the non-additive properties of the system. The inclusion of uncertainty produces slight decreases and increases in skill at different rainfall accumulations which are not statistically significant. The differences in the ETS may also be related to the non-additive properties of the ETS. The false alarm rate is reduced, improving the skill as represented by the ROC in panels (c) and (g). The hit rate is also reduced. From the log-odds ratio graphs in panels (d) and (h), there are some gains in skill at the higher accumulation thresholds ($>$12 mm) where the introduction of uncertainty appears to improve the model's performance through the increase in the radar area. Another interesting aspect of the plots without uncertainty is the clear separation between forecasts 0–12 h and 12–24 h. This is evident for the ETS, ROC and log-odds ratio.

The intensity-scale diagrams for the six-hourly accumulations are given in Fig. 8. There is no evidence of a systematic pattern in the error distribution when uncertainty is included.

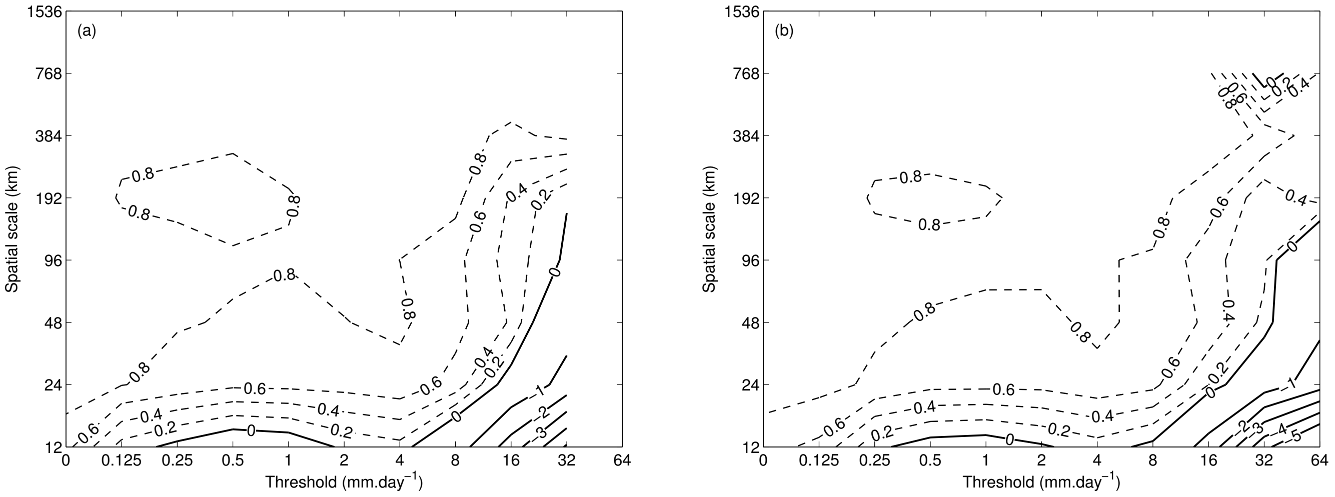

Fig. 6. Two-dimensional plots of the MSE per threshold and scale, known as "intensity-scale diagrams", for 23 June 2004 showing the results without uncertainty in **(a)** and with uncertainty in **(b)**. Note that the point where a contour terminates implies that no data are available for analysis beyond that point.

Comparing panels (a) and (e) at 0–6 h lead times the differences are minimal, with a slight worsening of the errors at the grid scale. At 6–12 h lead times panels (b) and (f) show that the worst spatial error at 8 mm is reduced but errors at 16 mm have increased. Perhaps this is more of a redistribution of the error. Panels (c) and (g) which represent lead times of 12–18 h show that the inclusion of uncertainty worsens the spatial error for totals greater than 8 mm. At 18–24 h lead time, panels (d) and (h) again indicate the worsening of the spatial error for totals greater than 8 mm. The errors at or near the grid scale are largely the same, especially for small accumulations, typically less than 4 mm. Another point of comparison is how the model performs at these shorter accumulation periods as opposed to the daily scale. One could expect the model to perform less well at the shorter time scale as timing errors become more pronounced. This is indeed the case, when comparing Figs. 6 and 8. The difference lies in the errors at the grid scale, which are worse for the shorter time scale. Certainly errors at the grid scale and near twice the grid scale across all thresholds appear to suggest the need for averaging to at least twice the grid length.

4 Monthly behaviour

Thus far the impact of introducing uncertainty to the verification process has been considered for a two-day heavy rainfall event. In this section the effect on longer-term statistics is assessed. The forecast errors and skill of both daily and six-hourly accumulations up to a lead time of t+24h are assessed using traditional verification measures and spatial error decomposition methods. First the method for aggregating individual intensity-scale diagrams is briefly described.

4.1 A modified sign test statistic

It is highly likely that Numerical Weather Prediction (NWP) model forecast errors behave non-linearly, which implies that the skill score SS used to construct individual intensity-scale diagrams can not simply be added together and averaged to obtain, say, a monthly "mean" intensity-scale diagram. Mittermaier (2006) proposed a method for aggregating individual intensity-scale decompositions using the non-parametric sign test. For more detail please refer to the paper. In brief, an array containing the sign test statistic $\mathcal{B}$ (which is the number of positive skill scores SS for a given intensity and scale out of a possible m forecasts) is constructed in intensity-scale phase space. The null hypothesis H_0, is rejected if $\mathsf{b} \leq \mathsf{b}_{m,\alpha}$, where $\mathcal{B}$ is binomially distributed as $\mathcal{B} \sim bi(m, 0.5)$ for small samples ($m<40$), or approximately normally distributed for large samples. The significance level α has been set to 0.025. The result can be visually expressed in intensity-scale phase space as a modified sign test statistic $(m-\mathcal{B})/m$ which is the proportion of the scores that are negative. For each scale and intensity where H_0 is rejected, the location is shaded based on the modified sign test statistic.

4.2 Monthly statistics based on daily accumulations

Figure 9 shows the monthly behaviour of the daily accumulation errors as expressed by SS during June 2004. As explained in the previous section, shaded regions in this diagram indicate the scales and intensities where the null hypothesis has been rejected, i.e. the SS is negative at a confidence level of 97.5%. The degree of shading shows the proportion of negative scores. Regions that are shaded black indicate the proportion is equal to 1, i.e. SS was negative for every single forecast included in the analysis, and for the

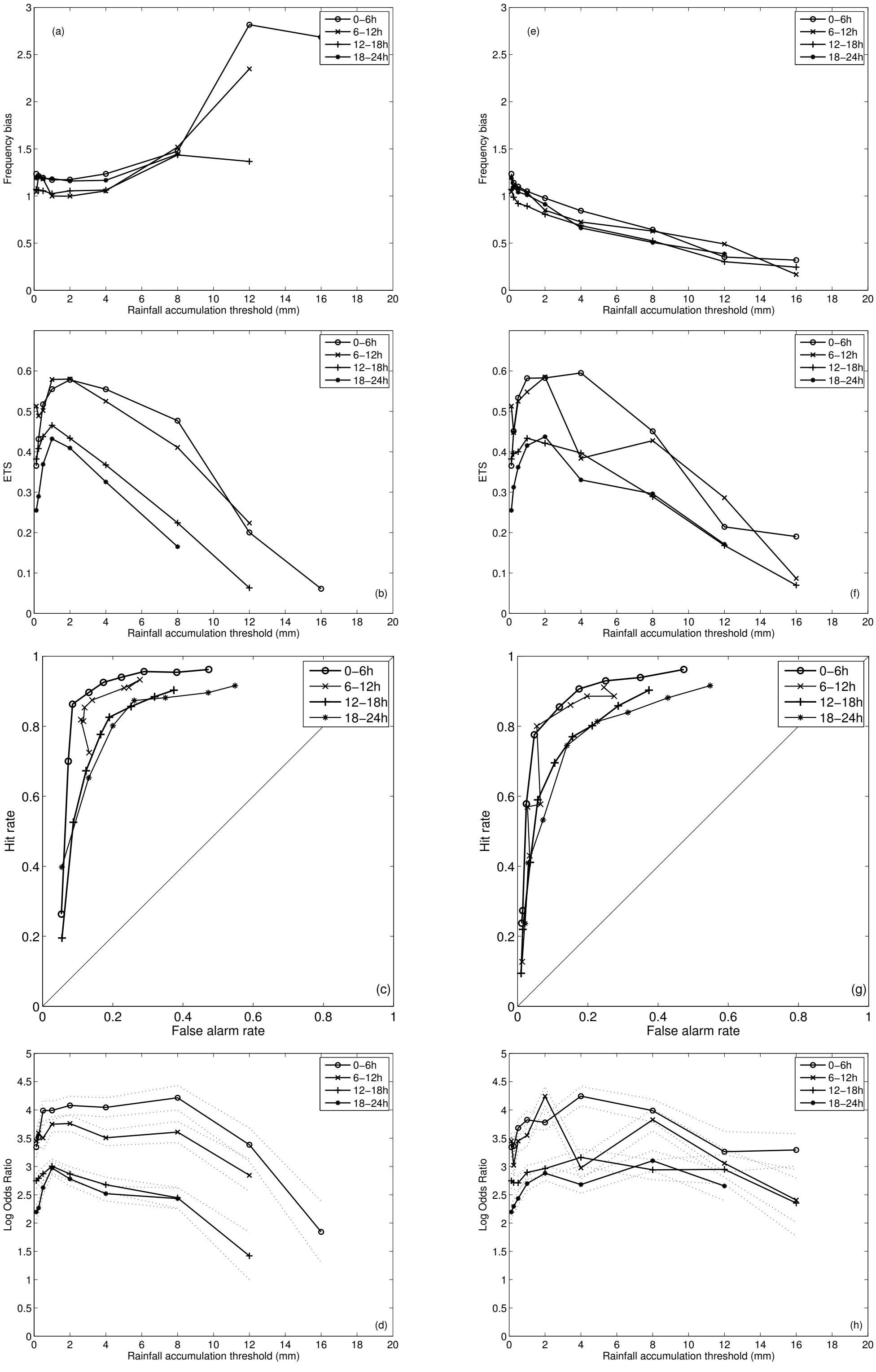

Fig. 7. Traditional measures (bias, ETS, ROC and log odds ratio) for the six-hourly accumulations from 12:00 UTC on 22 June 2004 to 12:00 UTC on 24 June 2004. Panels **(a)** to **(d)** show results without uncertainty, whereas panels **(e)** to **(h)** show the results with uncertainty.

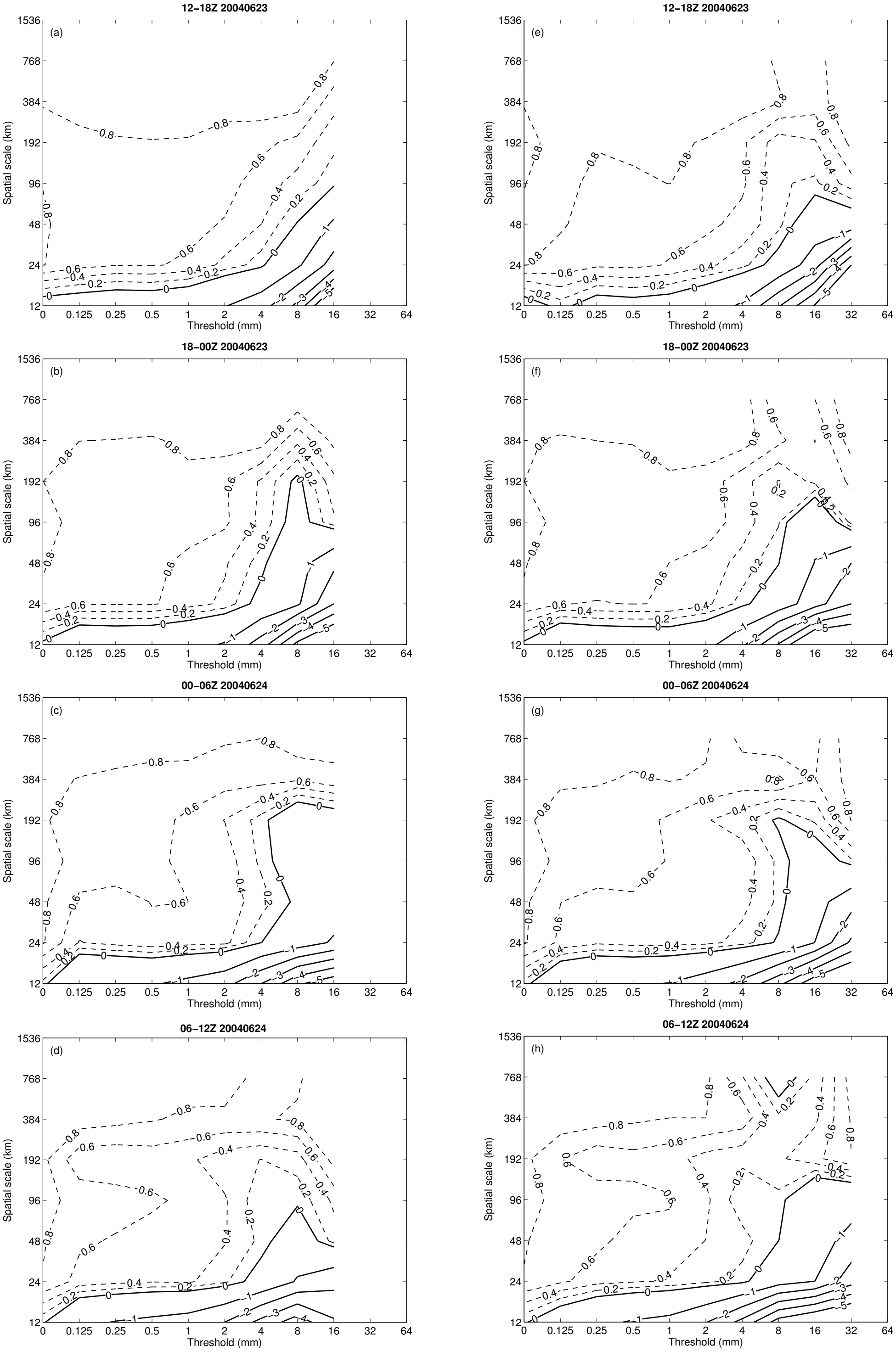

Fig. 8. Intensity-scale diagrams for six-hourly accumulations from 12:00 UTC on 23 June 2004 to 12:00 UTC on 24 June 2004. Panels **(a)** to **(d)** show results without uncertainty, whereas panels **(e)** to **(h)** show the results with uncertainty.

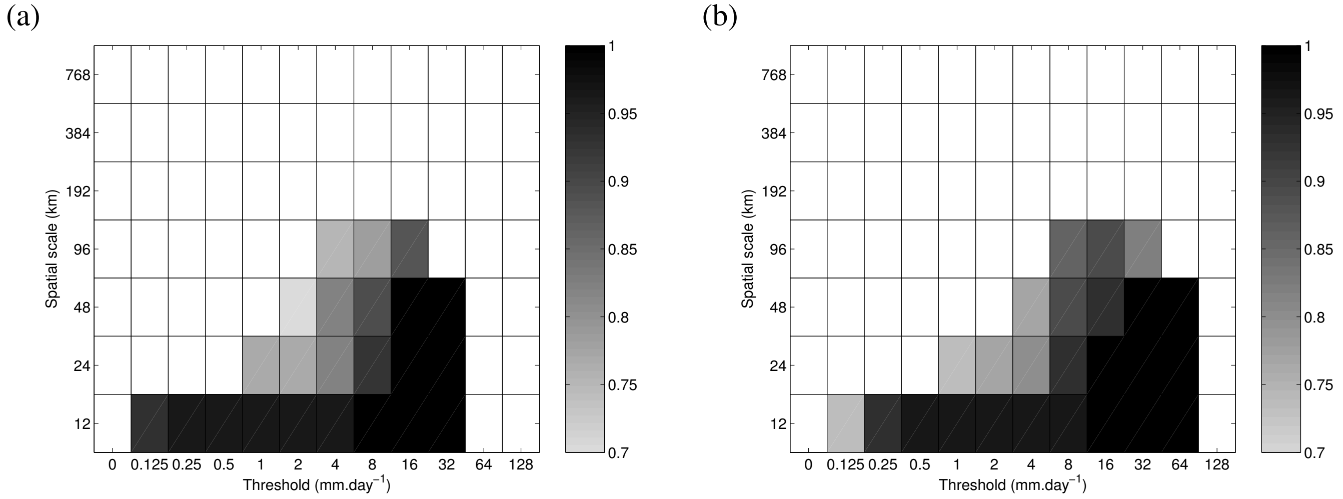

Fig. 9. Monthly behaviour of the skill score for daily accumulations during June 2004 as determined from the sign test statistic. Shaded regions show the scales and intensities where the null hypothesis has been rejected. The shading corresponds to the proportion of the scores that were negative. Panel **(a)** shows the results without the inclusion of uncertainty and **(b)** with uncertainty in radar-rainfall estimates included.

given scale and intensity. This could perhaps be explained as the "background" error distribution, where if the accumulation was forecast, the errors are prevalent or persistent over the month. These are confined to the grid scale, or up to four times the grid scale for accumulations greater than 16 mm.d^{-1}. Lighter shaded regions indicate scales and intensities where the SS was significantly negative some of the time, and are referred to as "transient errors". The graphs suggest that the inclusion of uncertainty has a small impact on the distribution of the persistent errors. There are some differences for the transient errors. The lagged thresholding also means that the model can be evaluated at a higher threshold, which is why the errors with uncertainty span all rainfall totals up to 64 mm.d^{-1}. The error distribution has been shifted or redistributed to some extent.

Linking these results to the physical, the persistent errors at the higher thresholds, and at the grid scale suggest that subgrid scale parameterizations such as the convection scheme still introduce significant errors to the output. Displacement and timing errors are more due to how well the model captures the atmosphere's dynamical evolution. These errors are clearly discernible on daily and six-hourly intensity-scale diagrams. They appear on the monthly diagrams as the grey shaded regions of transient errors around the black shaded region of persistent error.

Although the errors as captured in the intensity-scale diagrams may not be greatly affected by the inclusion of uncertainty at the monthly time scale, the skill as measured by the traditional scores, shows more variation. The scores are plotted in Fig. 10. It is also clear that not all scores show the same level of response. The trend in the bias in panel (a) is reasonably consistent and greater than 1, i.e. there are more false alarms than misses. Again the bias with uncertainty exhibits the same behaviour as seen previously for individual forecasts on the daily and six-hourly time scale, decreasing with increasing threshold, suggesting that there are progressively more misses than false alarms relative to the hits. The main effect of the inclusion of uncertainty is therefore to greatly increase the observed area of rainfall, across all thresholds, and the model does not (rightly or wrongly) capture this. It hints at the fact that the error in the radar-rainfall estimates is probably not constant for all thresholds and that, possibly, the model's perceived skill may be adversely affected if the uncertainty in the observations is treated equally for all intensities. The ETS in panel (b) shows that the inclusion of uncertainty increased skill for the low and higher accumulation totals (<4 mm and >=16 mm). This perhaps hints at closing the gap between the over-estimation of the light rainfall areas by the model and the under-estimation of the light rainfall areas by the radar. The confidence intervals obtained using a bootstrap re-sampling method show the differences are not statistically significant. Yet whilst the differences in the ETS may seem small, it is worth bearing in mind that the characteristics of the ETS (see e.g. Mittermaier, 2008) imply that small changes in the ETS are related to marked differences in the forecasts. In fact, to achieve a very large (near to 1) ETS requires a near-perfect forecast, and is therefore very rarely achieved. The ROC in panel (c) is much smoother for the monthly sample. The results are rather mixed (as a function of threshold), with no clear systematic trends in skill. The log-odds ratio in panel (d) also suggests some differences in skill but the analytical error suggests the differences are not significant.

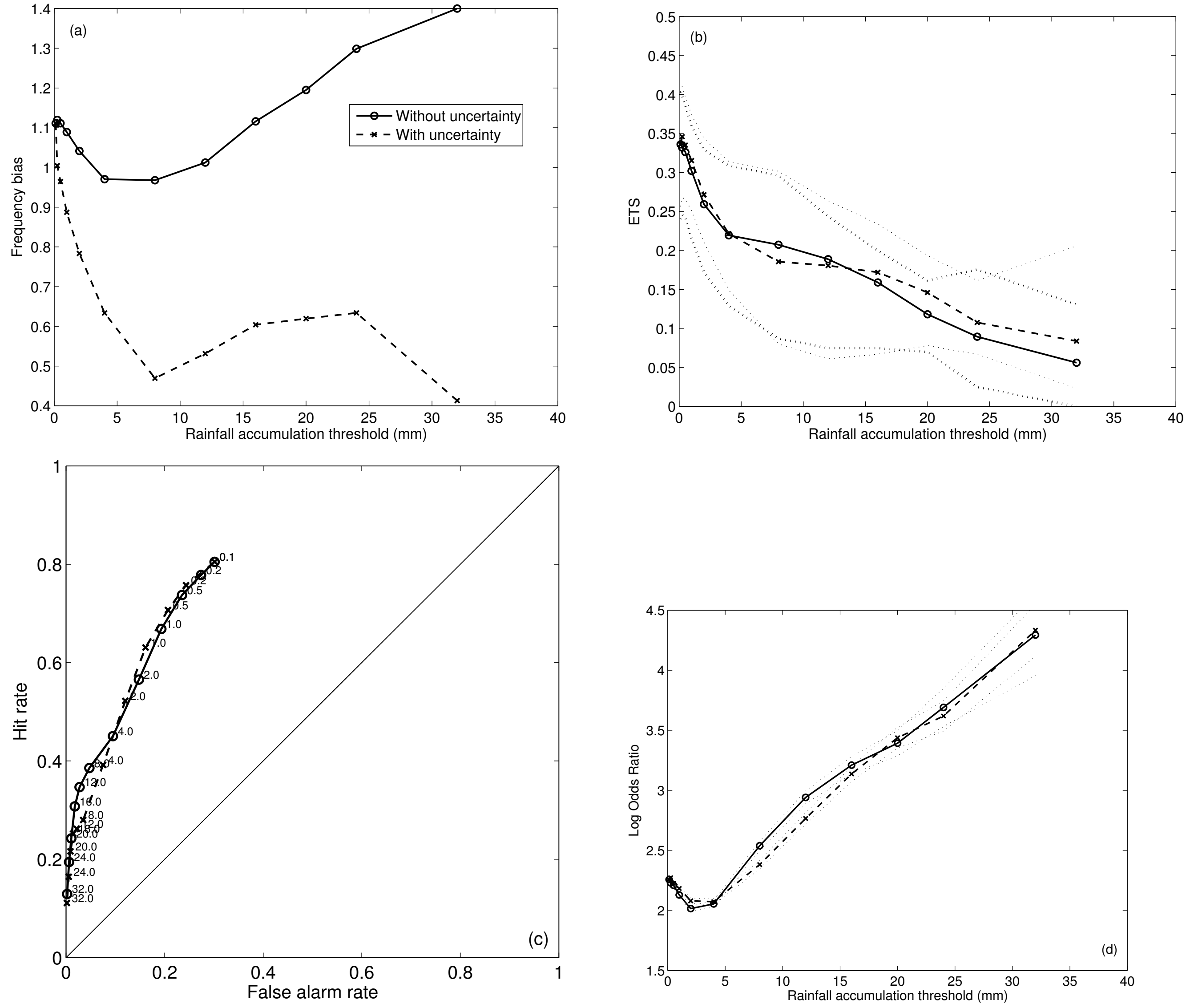

Fig. 10. Traditional scores for the 24-h accumulations of June 2004: **(a)** Bias, **(b)** ETS (with 5% and 95% bootstrap confidence intervals), **(c)** ROC and **(d)** log odds ratio ± log odds error. Results for the exclusion and inclusion of uncertainty are shown.

4.3 From six-hourly accumulations

The monthly behaviour of the intensity-scale diagrams obtained for the six-hourly accumulations spanning the first twenty-four hours of the forecast has also been considered. As the differences between results with and without uncertainty are small they are not shown.

The traditional scores for the mean six-hourly accumulations for June 2004 are given in Fig. 11. Overall the scores are very similar to the daily ones shown in Fig. 10 and differences can not be considered to be statistically significant. There are interesting trends in the evolution of the forecast skill with lead time that are hidden when analyzing the fields at the daily time scale. Considering Fig. 11a and e, the reversal in the behaviour of the bias is still present, which is essentially a function of the lagged thresholding method and the large increases in radar rainfall areas. Furthermore the 12–18 h forecast has the best bias, being closest to 1, and consistent for all thresholds. The inclusion of uncertainty does affect the evolution of model performance with lead time. Consider for instance the ETS shown in panels (b) and (f). (Note that bootstrap confidence intervals have not been included here for the sake of clarity.) The ETS oscillates with accumulation threshold and the 0–6 h and 6–12 h forecasts are out of phase. The inclusion of uncertainty has reduced skill for accumulations less than 2 mm and has increased skill for the larger thresholds, greater than 12 mm. The ROC in panels (c) and (g) show that uncertainty has

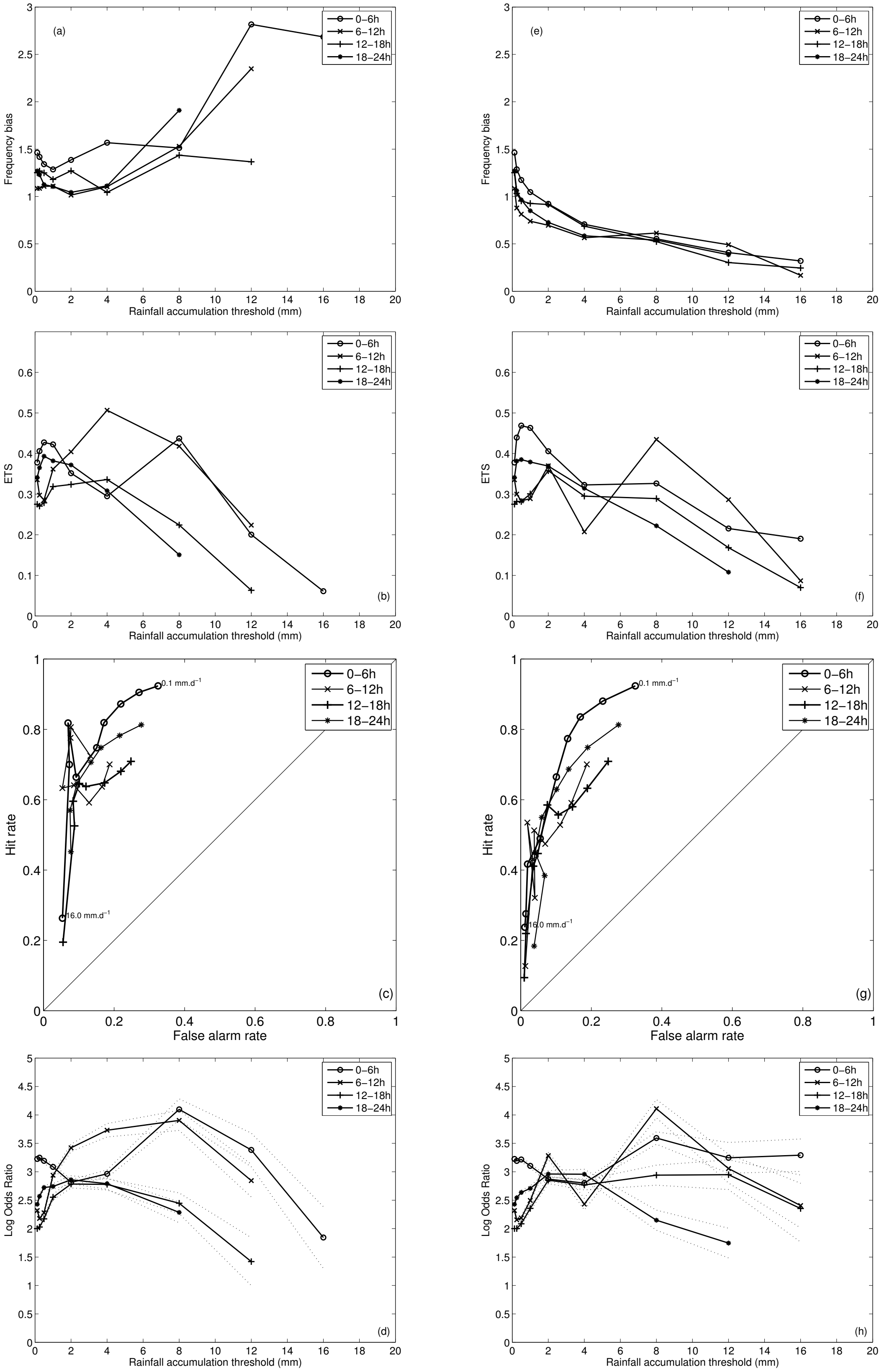

Fig. 11. Traditional measures (Bias, ETS, ROC and log odds ratio) calculated for the six-hourly accumulations for June 2004. Plots **(a)** to **(d)** show results without the inclusion of uncertainty whereas **(e)** to **(h)** include uncertainty.

reduced both the false alarm and hit rates, shifting the points down and to the left. The log-odds ratio shown in panels (d) and (h) also shows oscillatory behaviour especially at 0–6 h with an increase in skill as well as an extension in the range of thresholds that can be verified. This is due to the increase in the radar area with lagged thresholding.

5 Summary and conclusions

Uncertainty in radar-rainfall estimates has an impact on forecast skill and errors. For the UM, the effect is non-systematic, in that it may increase *or* decrease errors and skill, depending on the threshold. Some descriptive measures, such as the bias do show a trend when uncertainty is included. This effect can be considered to be a function of the methodology used, and as such, possibly somewhat artificial. Given this behaviour in the bias, any skill scores that have a dependence on the bias (e.g. log-odds ratio) may produce results that could be misleading, *if* interpreted on their own. This emphasizes the need for *always* using more than one statistic.

Regardless of the effect, uncertainty ought to be included, due to the potential downstream impact. The uncertainty interval used in this study (−50%/+100% implicit) is large but corresponds to the value frequently quoted in the literature. More constrained estimates of the error are not widely available and may not be universally applicable (e.g. different radar networks and correction algorithms will yield different error bounds). The results of this study may also suggest that the "factor-of-two" error used in this study may not be appropriate for all thresholds and that the error varies as a function of threshold. The method can be applied to other error bounds, even threshold-varying ones, if such were available. The only constraint is that the threshold progression for the intensity-scale method ought to be approximately log-linear.

The inclusion of uncertainty is not enough to gloss over gross observational errors. When using radar data, one must have confidence in the data source and the data quality checks and corrections that have been applied. Otherwise the (−50%/+100%) interval will not be enough. The inclusion of spurious radar data will cause the results to be badly skewed.

Intensity-scale diagrams calculated for individual daily and six-hourly forecasts highlight the errors at the grid scale, and at longer length scales for large accumulation thresholds. One can potentially differentiate between persistent (attributable to sub-grid scale processes and parameterizations) and transient errors (attributable to timing errors and differences in larger-scale dynamic evolution).

For the six-hourly forecasts, the error distribution evolves only marginally with increasing forecast lead time, but the skill scores do show a decrease in skill. For June 2004 the 0–6 h interval performs best for most rainfall thresholds greater than 2 mm. This could be attributed to the adjustment of the model moisture fields using the latent heat nudging assimilation scheme (Macpherson, 2001).

The results show that it is necessary to average model output, even if only over two grid lengths. Most UM forecasts shown here show skill at this scale, based on the skill scores, but also from the intensity-scale diagrams which show that many errors are eliminated at this length scale. Sometimes averaging to 4–8 times the grid length is required. This begs asking the question whether such an upscaled forecast would still be useful, even if most of the errors have been eliminated. The results also show that single model forecasts should not be interpreted at the grid scale. There is some evidence to suggest that model verification scores for the low thresholds improve when uncertainty in the radar-rainfall estimates is included. More investigation is required to determine whether this conclusion is robust. Of course these results are for just one month and one model (UM). Other models may produce similar but not identical results.

The non-linearity between error and skill has been highlighted, and it has been shown that methods such as the intensity-scale method are complementary to more traditional measures of skill such as the ETS and log-odds ratio. Forecasts have errors but these do not make them fundamentally unskillful.

The results shown here are based on "worst case" error statistics. It is hoped that individual radar networks, including the UK network have errors far smaller than this. Yet, actual error statistics are not readily available. Demonstrating that NWP models are providing accurate and detailed forecasts is becoming increasingly challenging with every resolution upgrade. Differentiating between model and observation errors is emerging as an important activity in establishing whether the higher resolution forecasts are yielding the desired benefits. Radar-rainfall estimates alone can provide the detailed observations required for verifying high-resolution (<5 km) model precipitation forecasts. Detailed error statistics are desperately needed. This work has only just begun and many other avenues of including and quantifying the uncertainty in the observations we use for verification have yet to be explored.

Appendix A

Some basic categorical statistics

A 2×2 contingency table is populated by applying a threshold to both observed and forecast fields as follows:

	Observed yes	Observed no
Forecast yes	*a* *hits*	*b* *false alarms*
Forecast no	*c* *misses*	*d* *correct rejections*

From the contingency table we can calculate many different descriptive measures and skill scores. The most common descriptive measure is the frequency bias B which is defined as:

$$B = \frac{a+b}{a+c}. \tag{A1}$$

One of the most frequently used skill scores is the Equitable Threat Score (ETS), which is defined as:

$$\mathrm{ETS} = \frac{a-a_r}{a+b+c-a_r} \quad \text{where} \quad a_r = \frac{(a+b)(a+c)}{n}. \tag{A2}$$

The odds ratio is defined as:

$$\mathrm{OR} = \frac{ad}{bc}, \tag{A3}$$

which is often expressed in a log form and has the advantage of having an analytical error formula, $(1/a+1/b+1/c+1/d)^{0.5}$.

The relative operating characteristic (ROC) is a plot of the hit rate (HR) against the false alarm rate (FAR), and is most commonly used for evaluating probabilistic forecasts. It can be used for deterministic forecasts as well where each data point is the hit rate (HR) and false alarm rate (FAR) for a given threshold. The interpretation is still the same. The idea is to minimize the FAR whilst maximizing the HR, thus concentrating points in the top left corner of the plot domain (where HR=1 and FAR=0).

$$\mathrm{HR} = \frac{a}{a+c} \tag{A4}$$

$$\mathrm{FAR} = \frac{b}{b+d} \tag{A5}$$

For more information on categorical statistics see texts such as Wilks (2006) or Jolliffe and Stephenson (2003).

Acknowledgements. I would like to thank C. Wilson (Met Office) and B. Casati (CMC) for their input and advice in putting this paper together. I am also grateful to the constructive comments provided by the reviewers. This work is Crown Copyright and may not be amended in any way.

Edited by: P. P. Alberoni
Reviewed by: two anonymous referees

References

Bowler, N.: Explicitly accounting for observation error in categorical verification of forecasts, Mon. Weather Rev., 134, 1600–1606, 2006.

Bowler, N., Pierce, C., and Seed, A.: STEPS:A probabilistic precipitation forecasting scheme which merges extrapolation nowcast with downscaled nwp, Q. J. Roy. Meteorol. Soc., 132, 2127–2155, 2006.

Briggs, W. M. and Levine, R. A.: Wavelets and field forecast verification, Mon. Weather Rev., 125, 1329–1341, 1997.

Casati, B. and Wilson, L.: A new spatial-scale decomposition of the Brier score: Application to the verification of lightning probability forecasts, Mon. Weather Rev., 135, 3052–3069, 2007.

Casati, B., Ross, G., and Stephenson, D.: A new intensity-scale approach for the verification of spatial precipitation forecasts, Meteorol. Appl., 11, 141–154, 2004.

Davies, T., Cullen, M., Malcolm, A., Mawson, M., Staniforth, A., White, A., and Wood, N.: A new dynamical core for the Met Office's global and regional modelling of the atmosphere, Q. J. Roy. Meteorol. Soc., 131, 1759–1782, 2005.

Germann, U., Berenguer, M., Sempere-Torres, D., and Salvade, G.: Ensemble radar precipitation estimation – a new topic on the radar horizon, Proceedings 4th European Conference on Radar in Meteorology and Hydrology, pages 559–562, 2006.

Göber, M., Wilson, C., Milton, S., and Stephenson, D.: Fairplay in the verification of operational quantitative precipitation forecasts, J. Hydrol., 288(1–2), 225–236, 2004.

Golding, B. W.: NIMROD: a system for generating automated very short range forecast, Meteorol Appl, 5, 1–16, 1998.

Golding, B. W.: Quantitative Precipitation Forecasting in the UK, J. Hydrol., 239, 286–305, 2000.

Harrison, D. L., Driscoll, S. J., and Kitchen, M.: Improving precipitation estimates from weather radar using quality control and correction techniques, Meteorol. Appl., 6, 135–144, 2000.

Jolliffe, I. T. and Stephenson, D. B. (Eds.): Forecast Verification: A Practitioner's Guide in Atmospheric Science, John Wiley and Sons, 2003.

Joss, J. and Waldvogel, A.: Precipitation measurements in hydrology, in: Radar in Meteorology: Battan Memorial and 40th Anniversary Radar Meteorology Conference, edited by: D. Atlas, pages 577–606, AMS, 1990.

Macpherson, B.: Operational experience with assimilation of rainfall data in the Met Office mesoscale model, Meteor. Atmos. Phys., 76, 3–8, 2001.

Mittermaier, M.: Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts, Atmos. Sci. Lett., 7(2), 35–42, 2006.

Mittermaier, M.: The potential impact of using persistence as a reference forecast on perceived forecast skill, Wea. Forecasting, in press, 2008.

Mittermaier, M., Hogan, R., and Illingworth, A.: Using mesoscale model winds for correcting wind-drift errors in radar estimates of surface rainfall, Q. J. Roy. Meteorol. Soc., 130, 2105–2125, 2004.

Roberts, N. and Lean, H.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Mon. Weather Rev., 136(1), 78–97, 2008.

Saetra, O., Hersbach, H., Bidlot, J.-R., and Richardson, D.: Effects of observation errors on the statistics for ensemble spread and reliability, Mon. Weather Rev., 132, 1487–1501, 2004.

Seed, A.: A dynamic and spatial scaling approach to advection forecasting, J. Appl. Meteor., 42(3), 381–388, 2003.

Turner, B., Zawadzki, I., and Germann, U.: Predictability of precipitation from continental radar images. Part III: Operational nowcasting implementation (MAPLE), J. Appl. Meteor., 43(2), 231–248, 2004.

Vasić, S., Lin, C., Zawadzki, I., Bousquet, O., and Chaumont, D.:

Evaluation of precipitation from Numerical Weather Prediction models and satellites using values retrieved from radars, Mon. Weather Rev., 135, 3750–3766, 2007.

Wilks, D.: Statistical Methods in Atmospheric Sciences, Academic Press, 2nd edition, 2006.

Zepeda-Arce, J. and Foufoula-Georgiou, E.: Space-time rainfall organization and its role in validating quantitative precipitation forecasts, J. Geophys. Res., 105, 10 129–10 146, 2000.