1 INTRODUCTION
Precipitation is one of the atmospheric variables with the most direct impact on human activities. Precipitation forecasts play a key role in decision making: the issuing of flood warnings, agricultural policies, aviation and other transport sectors. Verification of precipitation forecasts is essential for monitoring model performance and providing useful feedback for improvement.

Validating precipitation forecasts is also one of the most challenging, and frustrating of verification tasks. Unlike most other model variables, it remains difficult to distinguish a true upward trend in the skill of precipitation forecasts from time series of skill scores. Often, the signal is trapped within the noise, indistinguishable from it.

Precipitation fields themselves are noisy and discontinuous with sharp gradients in both space and time. The noisiness is due to the natural spatio-temporal variability of precipitation. This also means that data contain statistically speaking, outliers and/or missing values. Precipitation fields are also characterised by the presence of coherent spatial structures such as fronts, introducing a complex spatial dependency. Precipitation forecast verification therefore demands robust techniques designed specifically to deal with the precipitation field characteristics. To address this, recently developed verification techniques have aimed 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 features have different scales (e.g. from individual showers or organised rainbands in fronts) and are driven by different physical processes (e.g. thermal/orographic triggering, synoptic-scale forcing). This approach can potentially help determine which aspects of the model need further development. Briggs and Levine (1997); Zepeda-Arce and Foufoula-Georgiou (2000); Alvera-Azcárate (2004) all introduce such methods.

2 THE INTENSITY-SCALE TECHNIQUE
Another such method is the intensity-scale technique introduced by Casati et al. (2004) and the interested reader is referred to the paper. In what follows only a short summary will be provided for clarity. Here, the method’s potential usefulness in assessing operational deterministic model performance is tested.

Thresholding is used to convert the forecast ($y'$) and analysis ($x$) into binary images. The difference between the binary forecast and analysis defines the binary error $Z = I_{y'} - I_X$. The binary error image is then expressed as the sum of components on different spatial scales by performing a two-dimensional discrete Haar wavelet decomposition. The same result can also be obtained more simply by averaging over square regions on different scales. The Mean-Squared-Error (MSE) of the binary 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

$$MSE = \sum_{i=1}^{N} MSE_i, \quad (1)$$

where $MSE_i = Z_i^2$ is the MSE of the $i^{th}$ spatial scale component of the binary error image. For each precipitation rate threshold, the binary MSE skill score can be calculated, relative to the MSE of a random forecast:

$$SS = \frac{MSE - MSE_{random}}{MSE_{best} - MSE_{random}} = 1 - \frac{MSE}{MSE_{random}}, \quad (2)$$

where $MSE_{best} = 0$ is the MSE associated with a perfect forecast, and $MSE_{random}$ is the MSE associated with a random forecast calculated from the bias and the base rate at each threshold. This ensures the sensitivity of the method at high rainfall thresholds where the number of occurrences compared to the sample size are small. In this manner the method is particularly sensitive to the intensity and removes the dependence on spatial scale.

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 doesn’t necessarily mean that the model forecast has no skill.

3 MODEL OUTPUT AND RADAR DATA
Both 24-hour and six-hourly precipitation accumulations of the mesoscale version of the Met Office Unified Model (MES), and the new UK 4 km version were evaluated. It is recognised that using 24-hr accumulations may damp timing errors at shorter time scales (e.g. the time of a front at a given location). Therefore the four six-hourly accumulations spanning the same 24-hr period are also studied.

Daily accumulations are worth studying as they provide insights into the climatology of the model. The MES
in its current operational configuration has a resolution of 0.11° which translates to a 12 km grid over the UK. The precipitation accumulation is the sum of the convective and large scale, liquid and snow amounts. The MES is run one times a day, whereas currently the UK 4 km model is run only once, at 00 UTC. The MES is run with 3DVAR plus the assimilation of moisture variables via a latent heat nudging scheme (Jones and Macpherson, 1997; Macpherson, 2001). Lateral boundary conditions are provided from the global run. The UK 4 km model is currently run with no assimilation and lateral boundaries are provided by the 12 km North Atlantic European (NAE) model. For a more detailed description see the paper by Bornemann et al. (2005).

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 minutes, out to 6 hours. The precipitation rate forecasts are produced using radar and satellite data, along with surface observations. Advection techniques are used to produce nowcasts from observations and these are combined with MES forecasts. As the lead time increases, the forecast process gives less weight to the nowcasts from the observations and more weight to the MES forecasts (Harrison et al., 2000).

In their paper, Casati et al. (2004) verified 1-3 h 5-km resolution NIMROD nowcasts against their corresponding analyses. At this lead time the radar data and mesoscale model NWP output have almost equal contributions. By contrast, in this study stand-alone deterministic forecasts are being verified against the NIMROD baseline accumulation product, produced from the quality-controlled UK radar composites. The 5 km fields are either interpolated onto the UK 4 km grid or averaged onto the 12 km MES grid. When interpolating to a higher-resolution grid the intensity levels will remain the same. Interpolation in itself can't manufacture higher peaks. On the other hand, averaging to a coarser resolution yields a smoothed result. Given the radar coverage, comparisons can only be made where radar data are available, the rest of the NIMROD analysis being masked out.

4 ASSESSING FORECAST PERFORMANCE

In the run-up to operational implementation of the 4 km model, 20 cases were studied over the reduced 4 km domain (compared to the MES) and compared to the 12 km MES output. These span the period March 2004 to January 2005.

For the intensity-scale method the domain size changes when assessing the UK 4 km model output. With the 12 km MES a 128 by 128 domain is used whereas at 4 km grid, a 256 by 256 grid is possible. The scale progression is also different. For the 12 km the scales range from 12–1536 km whereas for the 4 km the progression is from 4–1024 km.

Although it is worthwhile considering the error analysis at 4 km resolution, for comparison to the MES, and for assessing the added benefit or value of running a higher resolution model, the scale decomposition was repeated using the 4 km model output averaged to the 12 km MES grid. In doing so the same 12 km averaged radar fields used for assessing the MES are also used to assess the 4-km model averaged output. This ensures that the “basis of comparison” is the same.

4.1 A severe local flooding event

The daily and six-hourly time scales will be discussed using a case study of a severe local flooding event. On 16 August 2004 devastating floods swept through the village of Boscastle on the north Cornwall coast in southwestern England. Over 180 mm were recorded by one gauge in a five-hour period (Trevalec, 11:30–16:30 LT). This case is also discussed by Roberts and Lean (2005) using the 1 km storm scale model.

The 24-hour radar and model rainfall accumulations (00Z to 00Z) along with the intensity-scale diagrams are shown in Fig. 1. At 4-km resolution the localised storm track appears on the radar rainfall composite (a) as an accumulated value of between 32-64 mm.d⁻¹ (the 2-km radar product showed a single pixel with an accumulation in excess of 128 mm). When the radar accumulations are averaged to 12 km, a smaller maximum of the same range remains to be seen in (d).

The MES (at 12 km) shown in (e) represents this event as a smooth 8-16 mm.d⁻¹ accumulation over a larger area. The first thing to notice is that the 4 km model, shown in (b) and (g) over-estimates rainfall intensity in general. Yet it is able to resolve the observed banded structure of the showery elements, especially over England and Wales. Most showers were aligned along a trajectory, with subsequent showers following the same path as those before.

The radar accumulations clearly highlight the local nature of the event, the most intense rainfall swath only 2–4 pixels across, i.e. 8-16 km. For the MES with a 12 km resolution such features are considered grid-scale, which it can not resolve. It would therefore be fair to say that the MES performance is within its expected capabilities.

The full 4-km resolution accumulations show rainfall maxima between 64–128 mm.d⁻¹ along the Cornish coast. Averaging the 4-km model field to the MES resolution, an intense swath of rain with accumulations between 32–64 mm.d⁻¹ remains, as seen in (g). In hindsight it is clear that the 4 km forecast provided better guidance on the possibility of a localised heavy rainfall event than the MES, purely because of the added detail and structure.

The intensity-scale diagram for the 4-km full resolution accumulation shown in Fig. 1(c) suggests that whilst errors are present at the grid scale for all accumulation totals, they spill over to larger length scales for all accumulations greater than 4 mm.d⁻¹ with considerable mismatches in placement of accumulations of 16 mm.d⁻¹. When comparing the 4-km output averaged to the MES...
Figure 1: Twenty-four hour rainfall accumulations for 16 August 2004 showing in (a) the 5-to-4 km interpolated NIMROD radar accumulation, and (b) 00Z UK 4km run with its intensity-scale diagram in (c). The 5-to-12 km averaged radar accumulation is shown in (d), along with the 00Z 12 km MES totals in (e) and the corresponding intensity-scale diagram in (f). The 4-to-12 km averaged 00Z UK 4km accumulation is shown in (g) along with the intensity-scale diagram in (h). Only the negative values of the intensity-scale diagrams are contoured. Positive values (>0) are white. Contours are at 0.5 intervals between -4 and 0. Values less than -4 are shaded black.

grid with the MES (f) and (h), the length scale of the errors is very similar. Two notable differences are the reduction in the error length scale for accumulations between 4–8 mm d⁻¹, and the increase in errors for trace
Figure 2: Traditional scores for the six-hourly accumulations from the MES and 4 km model 00Z runs, for 16 August 2004 including the bias, ETS, ROC and log odds ratio. The left-hand column contains results for the MES, the right-hand column for the 4-km model averaged to the MES grid.
Figure 3: Six-hourly model accumulations from the MES and the averaged 4-km model output together with the corresponding intensity-scale diagrams for 12–18h on 16 August 2004.

(< 1 mm.d$^{-1}$) rainfall amounts for the 4-km model. Quite unusually the 4-km model has less light rainfall over central England than was observed.

The traditional measures of skill for the six-hourly accumulations from 00Z are shown in Fig. 2. The left-hand column contains the figures for the MES, whereas the right-hand column contains those for the 4-km model. The bias for the MES is relatively stable across all four 6-hourly intervals in (a) whereas the bias in (b) climbs sharply for the 4-km model, i.e. the number of forecast occurrences far exceeds those observed. For both models the 0–6h and 18–24h intervals have biases (uncharacteristically) less than 1 for thresholds less than 2–4 mm, implying that the misses are greater than the false alarms. Many more occurrences were observed than predicted.

The calculation of skill scores is also sample size dependent, i.e. they can only be calculated if the sample sizes are large enough to do so. There are more occurrences of higher rainfall accumulations for the 4-km model which enables the calculation of scores for these higher thresholds.

From the Equitable Threat Score (ETS) shown in (c) and (d), the change in skill as a function of forecast lead time is apparent. The performance of both the MES and 4 km models is quite similar according to the ETS, but rather indifferent for the latter time periods, with values less than 0.2. The same can be said for the Relative Operating Curves (ROC) shown in (e) and (f), and the log-odds ratio in (g) and (h). There are signs that the False Alarm Rate (FAR) for the 4-km model is lower, but then the Hit Rate (HR) is lower too. The log-odds ratio shows a slight improvement in skill at thresholds greater than 4 mm for lead times beyond 0–6h.

The traditional scores therefore do not paint a very favourable picture in terms of value added and improved forecast skill when comparing the 12 km MES and the new 4 km UK model output for this case. Given that one of the purposes of introducing high-resolution models is to improve the forecast of small-scale, localised events, at possibly short lead times, these results are not sufficient to establish their usefulness. Next, consider the 6-hourly intensity-scale diagrams just for the 12–18h forecast lead time shown in Fig. 3.

For the full resolution 4-km model accumulation in
(a) and the error distribution in (d) there is a sharp increase in errors not confined to the grid-scale for thresholds greater than 1 mm, with length scale errors up to 128 km, 32 times the grid length. When comparing the MES in (e) and the 4-km model output averaged to the MES grid in (f), at first glance the error distribution seems very similar. However the absolute differences between the plotted intensity-scale diagrams (e) and (f) in Fig. 3 show an increase in the magnitude of the error at low thresholds (< 4 mm), at the grid scale (12 km). This is offset by a marked decrease in error at all spatial scales up to 96 km for thresholds 4–32 mm. The fact that the model under-predicts the low threshold occurrences is anomalous given that it is more usual for the models to over-predict the extent of light rainfall.

This reduction in errors at moderate to high thresholds at the grid scale and longer length scales is encouraging and provides a truer reflection of the detail, and the value that has been added by running a 4 km model. Not all grid-scale detail is accurate, which is why the grid-scale errors may increase for the low and high thresholds.

Another encouraging aspect is that errors that manifest at, say twice the raw model grid length, in fact represent a reduction in the error length scale, i.e. 8 km instead of 24 km. In this case the worst errors are almost entirely confined to the grid-scale for the 4-km model. Averaging the output to even twice the grid-length would reduce or even eliminate the grid-scale errors. This example yet again illustrates just how difficult it is to verify localised, intense events, both due to small sample sizes and the need for averaging raw model output, which the intensity-scale diagrams clearly highlight, to eliminate grid-scale noise, and reduce forecast error.

4.2 A monthly point of view

It is highly likely that the forecast errors described above behave non-linearly, which implies that the skill score $SS$ used to construct the intensity-scale diagram can not simply be added together and averaged to obtain, say, a monthly “mean” intensity-scale diagram. As it can also not be assumed that the underlying errors are normally distributed, the use of a simple distribution-free test (see e.g. Steyn et al., 1994), such as the sign test is more appropriate for determining the behaviour of the median of the skill score, for a month, at each intensity and scale. The hypothesis can be expressed as:

$$
H_0 : SS \geq 0 \quad \text{(implicit positive and skillful)}
$$

$$
H_1 : SS < 0 \quad \text{(less skill than a random forecast)}.
$$

(3)

An array containing the sign test statistic $B$ (which is the number of positive skill scores $SS$ for a given intensity and scale) is constructed in intensity-scale phase space. The null hypothesis $H_0$, is rejected if $b \leq b_{n,\alpha}$, where $B$ is binomially distributed as $B \sim bi(n, 0.5)$ for small samples ($n \leq 40$). The significance level $\alpha$ has been set to 0.025. The result can be expressed visually in intensity-scale phase space as a modified sign test statistic $(n - B)/n$ 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. For example, if an error is present in each intensity-scale diagram included in the sample, the shading is black. This is illustrated in Fig. 4 with the first full month of 00Z 24-hour precipitation forecasts produced in the operational suite. The diagrams show the scale and intensity of the prevalent or persistent model errors.

Figure 4: Modified sign test statistic plot showing the presence of persistent error in the intensity-scale phase space for May 2005: (a) MES and (b) averaged UK 4 km model.

The “averaged” grid-scale errors for the 4-km model in Fig. 4(b) are slightly worse than in (a), being present in every daily scale-intensity diagram generated for the 00Z forecasts. This corroborates the observation that the 4-km model seems to produce more light precipitation although the total precipitation area is smaller. For the 4-km model, errors at twice and four times the MES grid length appear to be less prevalent than for the MES for the 2–8 mm.d$^{-1}$ range. Both models appear to have persistent errors at multiple grid length scales for accumulations 16–32 mm.d$^{-1}$. More prevalent errors appear at scales of 96 km for the 4-km model in the 4–8 mm.d$^{-1}$ range. This
could be attributed to the fact that the 4-km model produces larger rainfall totals and the errors extend to higher thresholds, or are more spread out.

5 SUMMARY

In the preceding sections a model comparison, between the "predecessor" (12 km MES) and the "successor" (the UK 4 km model) has been presented. Deterministic precipitation accumulation forecasts were compared to radar-rainfall accumulations from the NIMROD nowcasting system. The UM offers a seamless nested modelling system, with the global, mesoscale and high-resolution versions all sharing the same dynamical core and parameterisations, making model comparisons easier. General preliminary conclusions would include:

- The scale-intensity technique highlights the need for averaging raw model output, to reduce grid-scale noise and improve forecast skill, regardless of grid length.

- Traditional scores and results from the scale-intensity method complement each other. Even when the skill appears very similar, the error analysis may show that there are differences in the forecast error.

The main points of the model comparison thus far can be summarised as follows:

- The 4-km model forecast contains much more detail. Unfortunately more detail does not necessarily mean it’s more accurate!

- The results show that a high-resolution forecast adds value when attempting to predict potentially severe local events. Averaging raw model output may be more important for high-resolution models given the discrepancy between detail and accuracy. Still, the advantage of averaging a high-resolution forecast to say, twice the grid length is that it will still yield forecast products with more detail, at a higher resolution than currently available from the MES.

- When comparing the 4-km model output averaged to the MES grid, the 4-km model forecast errors appear to be less for mid-intensity ranges for both daily and six-hourly forecasts. This improvement also holds for the error length scales.

- There are hints that the forecast error may be scale-invariant, i.e. that the results are more dependent on grid length than on other potential contributing factors. This implies, for example, that an error that manifests itself at twice the grid length for the 4 km model is still better than for the 12 km MES because the error length scale is 8 km rather than 24 km.

A parallel analysis using rain gauge data instead of radar-rainfall accumulations has been initiated.

REFERENCES


