METEOROLOGICAL APPLICATIONS
*Meteorol. Appl.* **20**: 176–186 (2013)
Published online 4 October 2011 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.296

# A long-term assessment of precipitation forecast skill using the Fractions Skill Score

Marion Mittermaier,[a]* Nigel Roberts[b] and Simon A. Thompson[a]
[a] *Weather Science, Met Office, Exeter, UK*
[b] *Met Office, Reading, UK*

**ABSTRACT:** The Fractions Skill Score (FSS) is a spatial verification metric routinely computed in the operational verification suite. It enables the comparison of forecasts of different resolutions against a common spatial truth (radar rainfall analyses) in such a way that high-resolution forecasts are not penalized for representativeness errors that arise from the 'double penalty' problem. Officially Met Office model precipitation forecast accuracy is monitored using the Equitable Threat Score (ETS) at gauge locations. These precipitation scores form part of a basket of measures assessing six surface parameters known as the UK index, which forms the basis for making decisions regarding model upgrades (especially over the UK). It is used to monitor the impact of continuous model improvements. This framework and the methodology underlying it, is less appropriate for high-resolution forecasts for reasons as described above. For precipitation forecasts in particular, a new framework for long-term monitoring is necessary and the FSS provides such a potential framework.

This paper provides an objective critique of FSS results to date. It has been shown that the 'convection-permitting' (4 km) Unified Model (MetUM) forecasts are better than the 12 km MetUM (significant at the 5% level). The scale at which the models have sufficient practical skill is typically 10 km better for the high-resolution forecasts, and are better at forecasting afternoon convection exceeding 4 mm (6 h)$^{-1}$. The use of frequency (percentile) thresholds is recommended because of the implicit bias removal this approach provides, as any rain in a forecast period is treated as 'the event of interest'. Copyright © 2011 British Crown copyright, the Met Office. Published by John Wiley & Sons Ltd.

KEY WORDS high-resolution NWP; spatial verification methods, precipitation; long-term monitoring against radar

*Received 7 February 2011; Revised 1 August 2011; Accepted 8 August 2011*

## 1. Introduction

Improving the accuracy and usefulness of precipitation forecasts is a significant challenge and is of particular importance for high-impact events. The ability to predict the positioning and quantity of flood-producing rainfall for disaster mitigation purposes is vital, so the demands for accuracy, in terms of location and rainfall quantity are very high and increasing. The difficulty is compounded by the fact that many high-impact events are convective and highly localized. Until just a few years ago convective storms could not be resolved in operational Numerical Weather Prediction (NWP) models, but the advent of kilometre-scale 'storm-permitting' NWP models has completely changed forecast capability worldwide and brought higher expectations for local forecast accuracy, especially for high-impact events, and as a result new problems for forecast interpretation and verification.

Serious flooding events in the UK over recent years have further highlighted the need for more accurate precipitation forecasts and revealed the potential benefit of high-resolution NWP models. The Boscastle flood (May *et al*., 2004), the Ottery St Mary hail storm (Grahame *et al*., 2009; Clark, 2011), the extensive floods in July 2007, and several serious flooding episodes in parts of northern England between 2005 and 2009, have all provided evidence of the improved realism and guidance provided by storm-permitting NWP models for both convective and stratiform rainfall events. The challenge now is to detect this perceived improvement in forecast capability using an objective measure. Murphy (1993) suggests that good forecasts are consistent, accurate and valuable to a user and it is the change in goodness with increased resolution that needs to be assessed for all types of precipitation events.

Traditional verification measures such as those computed form a contingency table, or the root-mean-square error (rmse) do not fully account for the unique properties of precipitation. Furthermore, any method that requires an exact match between forecasts and observations in space and time is likely to penalize even small differences in intensity and location. The potential for such differences increases as the detail in the forecast increases. The rmse is sensitive to discontinuities, noise and outliers.

In short, traditional categorical measures are not appropriate for this task. Most common descriptive measures such as the frequency bias and skill indicators such as Equitable Threat Score (ETS) are based on the population of a 2 × 2 contingency table which is obtained by counting the number of forecast and observed events exceeding a given threshold (see e.g. Wilks, 2006). From this table a multitude of different measures can be calculated. Yet many of the categorical statistics such as the ETS are sensitive to the base rate (the frequency of observed events) so that as the intensity thresholds increase the skill is perceived to go down, purely because there are fewer events (not good for high-impact events) and the true signal of the skill of the forecast is not easy to discern. The ETS is also sensitive to the bias and, subject to hedging, the ability to improve a

* Correspondence to: M. Mittermaier, Weather Science, Met Office, Exeter EX1 3PB, UK. E-mail: marion.mittermaier@metoffice.gov.uk

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

score by over- or under-forecasting the frequency of an event (e.g. Baldwin and Kain, 2006).

At the Met Office the ETS is currently used for measuring the skill of all categorical variables (as a component of the UK index). The UK index considers all 6 h forecast lead times between 0 and 48 h. Parameters included are 10 m vector wind (as rms skill score), temperature (as rms skill score), precipitation ($> = 0.5$, 1 and 4 mm (6 h)$^{-1}$ ETS), cloud cover ($>2.5$, 4.5 and 6.5 okta ETS), cloud base height ($< = 100$, 300 and 1000 m given 2.5 okta ETS), and visibility ($< = 200$, 1000 and 4000 m ETS). For the UK index North Atlantic European (NAE) model precipitation forecasts are compared to rain gauge accumulations. The nearest grid point value to an observing site is selected for populating the verification database. Is this appropriate for kilometre-scale models? Precise matching in space and time is unlikely to show true forecast goodness. This is illustrated in Figure 1(a), using 6 months of concurrent forecasts from four different Unified Model (MetUM) configurations between April and October 2010 for the 4 mm (6 h)$^{-1}$ threshold. Here the 6-h forecast accumulations valid at 0000, 0600, 1200 and 1800 Z from the Global Model (GM, 25 km), the NAE (12 km), the UK4 (4 km) and UKV (1.5 km, pre-operational acceptance) are compared. UK4 and UKV run times are offset by 3 h so, for example, forecast accumulations are for 6 h ending $t + 21$ h for the UK4 and UKV, and $t + 24$ h for the NAE and GM. The global model lacks the horizontal resolution, whilst the sub-10 km scale models suffer from the additional detail that is too imprecise and is penalized when using gauges. The GM scores (25 km) are however very comparable to the NAE (12 km) with the UKV (1.5 km) and UK4 (4 km) worse than the coarser models. This also gets more pronounced with lead time. The GM is particularly competitive at the early lead times. Does this present the right picture? The answer is no. Figure 1(b) shows the frequency bias for the models for the same period. It shows that the high-resolution models tend to over-forecasting, with progressive under-forecasting for the coarser resolution models.

Traditional scores tend to favour smoother forecasts and doubly penalize the detail at higher resolution unless it is exactly correct because a mismatch with truth gives both a miss and a false alarm (the 'double penalty'). Since it is not expected that storm-permitting NWP models be correct at the grid-scale anyway because of rapid error growth at small scales (Lorenz, 1969) the interest ought to be in the benefit they provide at the larger scales that retain predictability and for that, point by point verification is not helpful. Forecasters would typically judge a forecast by the accuracy of the positioning and realism of a front rather than the precise details of the rain within the front. For this reason, much time and energy has been devoted to developing new spatial verification methods that aim to determine the likeness of a forecast to truth (in terms of perhaps distance apart or precipitation structure) rather than just whether each pixel or observing site is correct or not. A spatial methods inter-comparison project used a set of standard test cases to assess this plethora of new methods on a common data set. This was done in an attempt to document the strengths and weaknesses of these new methods. Gilleland *et al.* (2009) point out in the summary paper for this inter-comparison project that these new spatial verification methods also raise many questions regarding how output from km-scale models should be interpreted and presented to the user. Increased spatial resolution introduces more detail with faster error growth which means that high-resolution forecasts must be upscaled to improve forecast skill.

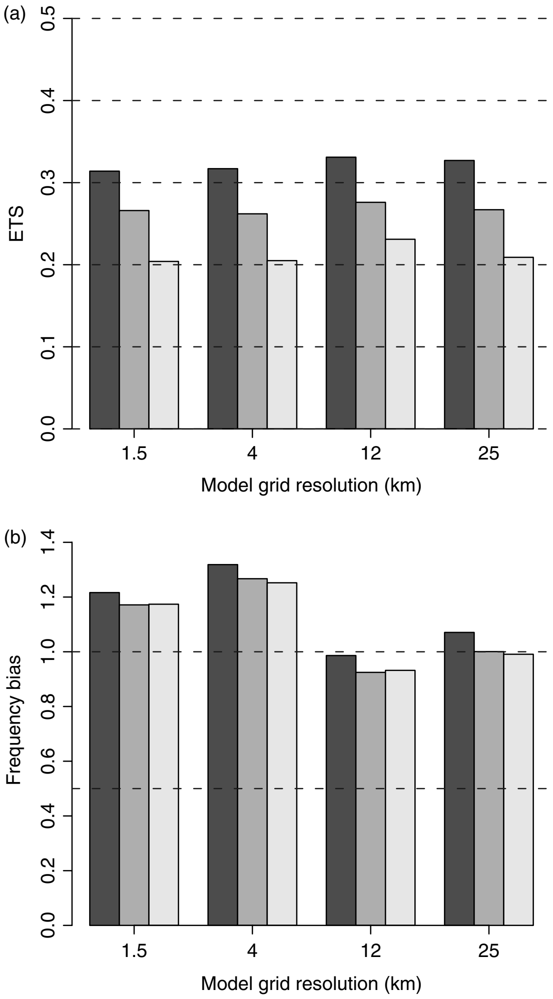

Figure 1. Six-month comparison of 6 h precipitation $> = 4$ mm (a) ETS and (b) frequency bias (both against rain gauges) for four MetUM model configurations: GM (25 km), NAE (12 km), UK4 (4 km) and the UKV (1.5 km) for a 4 mm (6 h)$^{-1}$ threshold. ■: t + 12/9; ▨: t + 24/21; □: t + 36/33.

The aim is to find the smallest scale at which a useful level of skill is reached for each resolution and if the higher resolution forecasts are better that scale will be smaller.

Gilleland *et al.* (2009, 2010) provide a classification of spatial verification methods into four groups: (1) neighbourhood, the changing relationship between surface temperatures and Indian monsoon rainfall with the phase of ESI tendency (2) scale separation (or scale decomposition), (3) features-based (or object-based), and, (4) field deformation. Neighbourhood spatial verification methods are designed to give credit for forecasts that are closer spatially. Ebert (2009) provides a comparison of a variety of neighbourhood-based methods, one of which is the Fractions Skill Score (FSS).

Roberts and Lean (2008) introduced the FSS for assessing high-resolution NWP precipitation forecasts in a way that takes account of the closeness of the forecast to truth and provides an estimate of the scale at which a forecast has acceptable skill. Roberts (2008) expands further on the uses of the FSS, and Mittermaier and Roberts (2010) discuss the characteristics of the FSS in the context of idealized and real test cases from the spatial methods inter-comparison project. The method was introduced to the Met Office operational verification suite in April 2007 and has been running routinely on a daily basis ever

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

since. The present paper provides an objective critique of the results that have been collected to date, and attempts to answer the question whether the first generation of high-resolution precipitation forecasts have realized their potential, i.e. that they are indeed better than their coarser resolution counterparts. Section 2 provides a brief overview of the Fractions Skill Score, how it is calculated, its characteristics and interpretation. A description of the two model configurations and the operational verification strategy is provided in Section 3. The time series analysis follows in Section 4, addressing aspects of skill as a function of threshold, lead time and spatial scale. The use of frequency thresholds (variable physical thresholds) is also assessed. The method also enables the calculation of a skilful spatial scale which provides valuable information on the spatial scale at which forecasts should be interpreted. An analysis of the skill as a function of the diurnal cycle is also investigated. Section 5 contains a summary and conclusions.

## 2. Method overview

To verify a rainfall forecast using the FSS, firstly a threshold (e.g. 1 mm accumulation) is applied to the forecast and verifying observed field (radar). A square 'neighbourhood' of a particular size is defined (e.g. $1 \times 1$, $3 \times 3$, $5 \times 5 \ldots$ pixels centred around each pixel). The fraction of pixels exceeding the threshold within the neighbourhood of each forecast grid point is then compared to the fraction of pixels exceeding the threshold within the neighbourhood of each equivalent observed grid point, using the FSS.

The FSS is a variation on the Brier Skill Score (Roberts, 2008) and given by:

$$FSS = 1 - \frac{FBS}{FBS_{\text{worst}}} \tag{1}$$

where the FBS is a mean square difference between observed and forecast rainfall fractions given by:

$$FBS = \frac{1}{N}\sum_{i=1}^{N}(O_i - F_i)^2 \tag{2}$$

$O_i$ and $F_i$ are the observed (radar) and forecast fractions respectively at each point and $N$ is the number of pixels in the domain.

$FBS_{\text{worst}}$ is a reference that gives the largest possible $FBS$ that could be obtained from the observed and forecast fractions when there is no collocation of non-zero fractions, i.e. the positioning of all the rain is in error by a distance greater than the size of the neighbourhood used and therefore gives the worst possible $FBS$:

$$FBS_{\text{worst}} = \frac{1}{N}\left[\sum_{i=1}^{N} O_i^2 + \sum_{i=1}^{N} F_i^2\right] \tag{3}$$

The FSS can have values between 0 and 1: 0 for a complete mismatch and 1 for a perfect forecast.

The computation of the FSS is repeated for all required neighbourhood sizes from the model pixel up to the whole domain. In this way, how the forecast skill varies with neighbourhood size can be determined first. Secondly, the smallest neighbourhood size which provides a sufficiently skilful forecast which is given by L (FSS $> 0.5 + f/2$) can be determined (Roberts and Lean, 2008), where $f$ is the observed fractional rainfall coverage over the domain (wet-area ratio). This represents a lower limit of useful scales. If $f$ is not very large (and it typically is not for a large domain) a value of 0.5 can be used. The upper limit of such useful scales is less easily defined and may well depend on the forecast application.

The FSS is sensitive to bias. Only if the FSS for the neighbourhood encompassing the whole domain is 1 can the forecast be considered unbiased. Otherwise the asymptotic value is an indication of the magnitude of the frequency bias $B$ given by $\text{FSS}(n = n_{\max}) = 2B/(1 + B^2)$, where $n_{\max}$ is the largest possible neighbourhood. So, an $\text{FSS}(n = n_{\max})$ value of 0.8 indicates a bias factor of 2. However, it does not indicate the sign of the bias, i.e. whether the forecast is over- or under-estimating precipitation totals. The only way an unbiased assessment is obtained is through the use of frequency thresholds. This is explored in Section 4.

The FSS neighbourhood concept also has applications in post-processing of high-resolution forecasts (e.g. Theis *et al.*, 2005). Mittermaier (2007) also showed that the skill of high-resolution precipitation forecasts can be further improved through creating lagged ensembles based on the neighbourhood concept used for calculating the FSS. The FSS was used to show that the skill of such Probability of Precipitation (PoP) forecasts was greater than that of the individual forecasts, effectively improving predictability, and increasing the accuracy of (probabilistic) high-resolution forecasts as a function of spatial scale.

## 3. Model configurations and verification process

In this study two configurations of the Met Office Unified Model (MetUM) are compared: the 12 km North Atlantic European (NAE), and the 4 km UK4. Both models run four times a day. Whilst the NAE runs at 0000, 0600, 1200 and 1800 UTC, the UK4 is embedded within the NAE, and runs with a 3 h offset, i.e. 0300, 0900, 1500 and 2100 UTC. Precipitation at the short-range (0–48 h) is assessed as six hourly accumulations between 0000 and 0600 UTC, 0600 and 1200 UTC etc. To compare the same accumulation periods the forecast lead times are also offset, so models cannot be compared at the same forecast lead time. This is indicated in the text as follows: $t + 24(21)$ h indicates a 24 h lead time for the NAE and a 21 h lead time for the UK4. When the older UK4 forecast is being compared to a more recent NAE forecast it could be argued that the skill advantage should be with the coarser NAE model (purely due to the impact of loss of skill as a function of lead time). However, if the older high-resolution UK4 forecast performs better despite this, it adds further weight to the assertion that the high-resolution forecast is better. Alternatively if the most recent UK4 forecast is compared the skill advantage should be with the UK4.

The UK index precipitation components are computed for all possible 6 h accumulation periods in the 0–36 h forecast range over a quality-controlled site list of ~110 sites. The UK4 model only provides a 36 h forecast, whilst the NAE provides a 48 h forecast every 6 h, so comparisons are only possible up to $t + 36$ h. Figure 2 shows the UK index combined weighted score and the weighted precipitation score differences (since April 2008) between the UK4 and NAE. The NAE is shown as the horizontal line with values above the line indicating that the UK4 is better, and below the line that the UK4 is worse. So, whilst it can be shown that the UK4 generally provides a better forecast overall in Figure 2(a) and (b) shows that

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

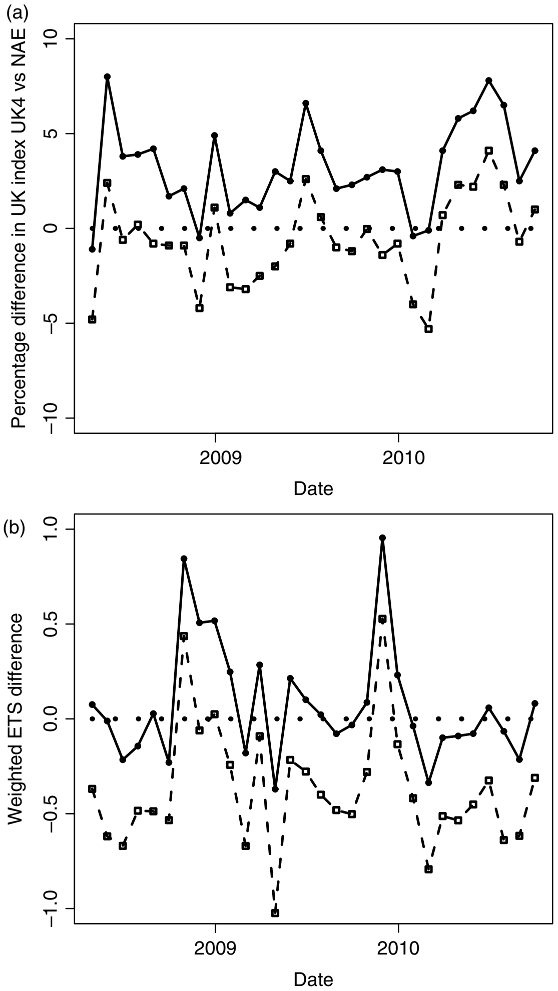

Figure 2. A comparison of the NAE (12 km) forecast performance with the UK4 forecasts 'either side', i.e. the UK4 initialized 3 h earlier, and 3 h later. (a) Percentage difference in monthly UK index scores. (b) Difference in the combined weighted ETS values for the three precipitation thresholds. —•—: most recent; -□-: previous.

the weighted ETS precipitation component differences between the NAE and UK4 (most recent and previous runs) is less positive. It illustrates the difficulties of being able to say that the UK4 precipitation forecasts are consistently better than the NAE using precise matching to rain gauge locations, and these traditional metrics.

The FSS components in the operational verification suite run four times a day. Both model forecasts are interpolated onto a regular 5 km Cartesian grid for comparison to the gridded UK radar composite accumulation. This may seem to benefit the 4 km model as the grid size is closer, but as will be seen later the main differences appear at scales significantly larger than 12 km, for which this choice should not matter. An hourly accumulation is made up of 12 5 min rainfall rate fields. Quality control is applied to ensure that the sub-hourly accumulations are representative of the hour (currently at least 10 of the 12 5 min fields must be available, i.e. >80%).

The FSS is computed over three areas, the United Kingdom (UK) as a whole, a southern England domain and a northern Scotland domain to assess regional variations. With many competing requirements, computational affordability in an operational context, involving High Performance Computing (HPC) resource, is of great importance. (The method lends itself well to code optimization and parallelization). Table 1 illustrates the lead times, thresholds and neighbourhood sizes that could be implemented in the operational verification suite. The results presented in Section 4 are for the largest area covering the whole UK, and a selection of thresholds and scale sizes in Table 1.

Table 1. Summary of the thresholds and spatial scales that are implemented in the operational verification suite to fit with available computing resource.

| Forecast range UK4 | Forecast range NAE | Physical thresholds (over 6 h) (mm) | Frequency thresholds (over 6 h) (%) | Neighbourhood sizes |
|---|---|---|---|---|
| $t + 9$ h | $t + 6$ h | 0.2 | 1.0 | 1 (5 km) |
| $t + 15$ h | $t + 12$ h | 0.5 | 2.5 | 3 (15 km) |
| $t + 21$ h | $t + 18$ h | 1.0 | 5.0 | 5 (25 km) |
| $t + 27$ h | $t + 24$ h | 2.0 | 10.0 | 7 (35 km) |
| $t + 33$ h | $t + 30$ h | 4.0 | 25.0 | 9 (45 km) |
| | $t + 36$ h | 8.0 | 50.0 | 11 (55 km) |
| | $t + 42$ h | 16.0 | | 13 (65 km) |
| | $t + 48$ h | 32.0 | | 15 (75 km) |
| | | | | 17 (85 km) |
| | | | | 19 (95 km) |
| | | | | 21 (105 km) |
| | | | | 25 (125 km) |
| | | | | 29 (145 km) |
| | | | | 31 (155 km) |
| | | | | 45 (225 km) |
| | | | | 61 (305 km) |
| | | | | 89 (445 km) |

## 4. Time series assessment

The data set comprises 4996 forecasts spanning a 41 month period. For the most part 12 month running means are used to show time series results. At other times the sample as a whole is considered. The asymptotic value of the FSS was used to filter out forecasts with potential radar data and/or sampling problems as a low value is usually associated with very low coverage for a particular threshold. This was set at a value of 0.2 (representing a factor of 10 in the frequency bias), and deemed necessary to not compromise the statistical integrity of the analysis. Excessive biases are typically related to days with little or no rain, where residual non-meteorological radar errors such as ground clutter or anomalous propagation would dominate the bias.

### 4.1. Scores as a function of intensity thresholds

Figure 3 shows the 12 month running mean time series of the FSS differences for a range of lead times, expressed as UK4 score minus NAE score. The three thresholds used for the UK index are used. Most score differences are positive, showing that the UK4 has more skill. The exception was during the first 18 months of the time series for the 0.5 and 1 mm (6 h)$^{-1}$ thresholds at longer lead times where the differences were marginal. At 4 mm (6 h)$^{-1}$ the differences have always been positive, suggesting that the UK4 is superior in capturing heavy rain including embedded or isolated convection. What is striking is the apparent growth in the difference in scores. It is generally accepted that the biggest upward impetus to

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

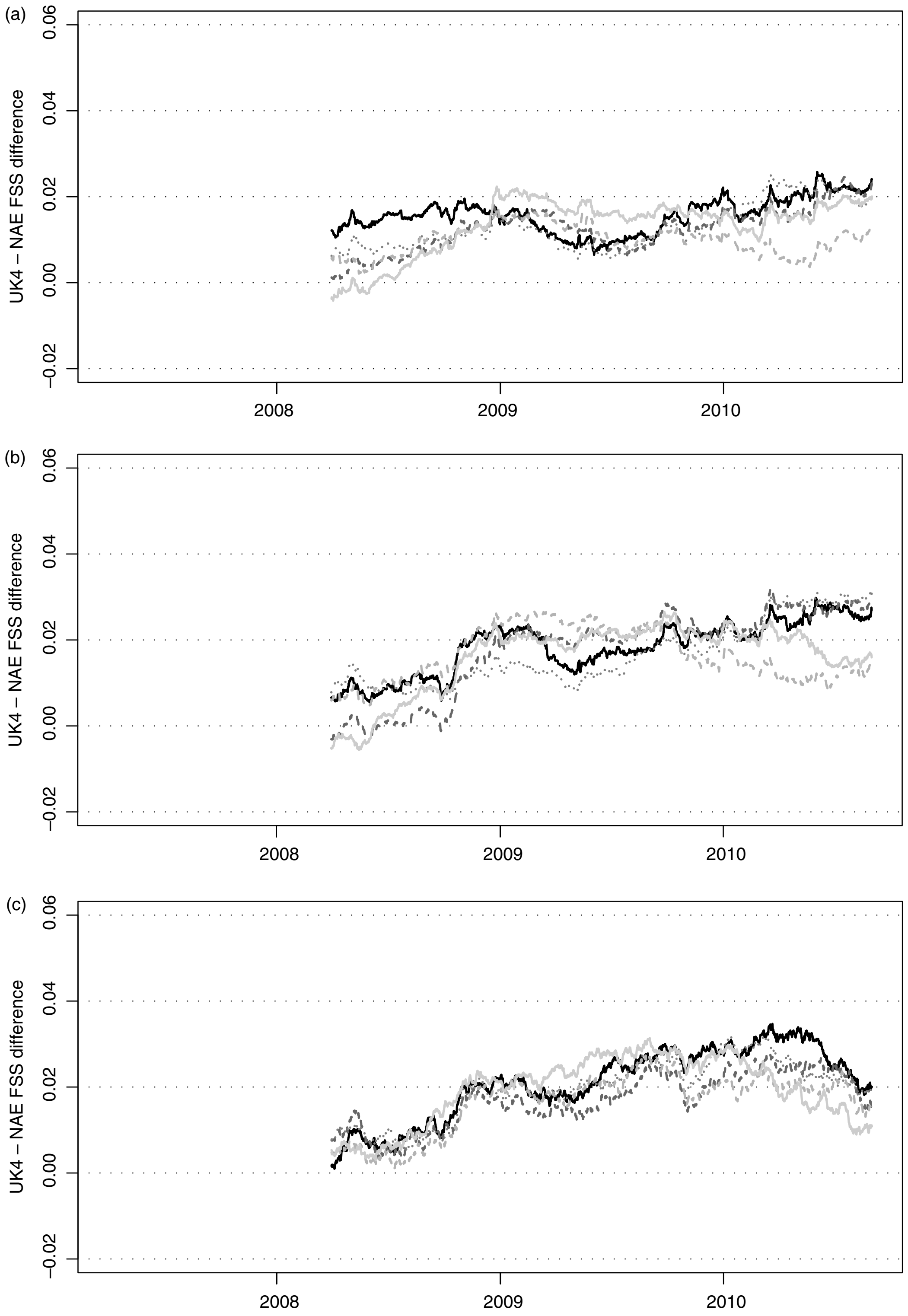

Figure 3. Time series of 12-month running mean FSS (UK4 - NAE) differences at a spatial scale of 25 km for the three UK index thresholds: (a) 0.5 mm (6 h)$^{-1}$, (b) 1 mm (6 h)$^{-1}$ and (c) 4 mm (6 h)$^{-1}$. ——: t + 12/9; – – : t + 18/15; · · : t + 24/21; – – : t + 30/27; ——: t + 36/33.

precipitation scores is given through increases in horizontal resolution, e.g. from long time series of verification statistics such as those computed at the European Centre for Medium Range Weather Forecasts (ECMWF, 2009) and Rodwell *et al.* (2010). The observed divergence in the scores found in this study must be due to some other effect. What it does suggest is that one model is doing much better (or worse) relative to the other, or possibly a combination of both these effects.

A list of model changes (introduced as parallel suites, PS) and impacts on precipitation scores in particular (as identified during trialling) are given in Table 2. These impacts are objectively assessed using the UK index. It must be stressed that it is quite unusual for a package of model changes to have a positive impact on all forecast parameters. Therefore, the UK index impact is augmented with pragmatic decision-making, deciding what sort of detrimental impacts

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

Table 2. List of relevant model change cycles during the last 4 years, with an indication of the impact on precipitation forecast performance only.

| Parallel suite | Date | NAE ppn | UK 4 ppn |
|---|---|---|---|
| 15 | Q1 2007 | Negative | Neutral |
| 16 | Q2 2007 | Neutral | Neutral |
| 17 | Q4 2007 | Neutral | Neutral |
| 18 | Q1 2008 | Neutral | Neutral |
| 19 | Q3 2008 | Neutral | Neutral |
| 20 | Q4 2008 | Positive | Neutral |
| 22 | Q4 2009 | Neutral | Neutral |
| 23 | Q1 2010 | Negative | Neutral |
| 24 | Q3 2010 | Positive | Neutral |
| 25 | Q4 2010 | Positive | Neutral |

The NAE impact is measured using the UK index during trialling. Whilst the UK index is also used to assess the impact of UK4 changes, the impact on UK4 precipitation forecasts is assessed qualitatively. (To date trialling methods only allow comparison against gauges which this is not considered to give a reliable signal of improvements).

are deemed acceptable for a package of model changes to be made operational. If the impact on other parameters is very beneficial for example, a small hit on precipitation scores will be acceptable, provided the overall UK index impact is positive, or at least neutral. Therefore, for example PS15 changes implemented in the first quarter (Q1, winter) of 2007 introduced a detrimental impact on NAE precipitation which largely persisted until the positive impact introduced in PS20 in Q4 2008. The impact of upgrades on UK4 precipitation forecasts was considered to be largely neutral, albeit that it was only qualitatively assessed at the time. This provides some indication that recent NAE upgrades are at least partially responsible for a degradation of precipitation forecast performance. There also appears to be a slight threshold and lead time dependence with later lead times at 0.5 and 1 mm (6 h)$^{-1}$ performing differently since early 2010, the difference being reduced. At 4 mm (6 h)$^{-1}$ the reduction in the differences is evident at all lead times. The reasons for this change in behaviour are as yet unclear.

The differences plotted in Figure 3 are assessed for statistical significance to quantify the benefit of high-resolution convection-permitting model forecasts over and above a coarser resolution model forecast. The methodology for assessing the statistical significance is provided in Appendix A. As noted in Appendix A the need to account for the dependence of successive forecasts requires the sample size to be adjusted. Table 3 lists the effective sample size, which is typically less than half of the total number of forecasts, with samples getting smaller with increasing threshold. Table 3 also lists the test statistic. It shows that all score differences are statistically significant except for 32 mm (6 h)$^{-1}$ where the sample sizes are insufficient to achieve statistical significance at the 5% level. This provides a reminder of the difficulties associated in robustly verifying more extreme thresholds (e.g. Ferro, 2007).

### 4.2. Scores as a function of spatial scale

When implementing the FSS the decision was taken to use the observation resolution as the basis for comparison. In this case the radar analyses are at 5 km. This implies that to perform a comparison of the NAE at 5 km, the 12 km forecast is interpolated down to a 5 km grid. Hence, NAE results for the first two neighbourhood sizes (1 and 3) equivalent to 5 and 15 km scales should be treated with care.

Figure 4 shows the FSS over all forecasts as a function of spatial scale at $t + 24$ (21) h for four thresholds. These are typical of those at other lead times. There is little difference between the models at the observation scale of 5 km, even though it is below the 12 km grid length, which indicates that near the grid scale the UK4 adds detail that is no improvement on using smoother interpolated NAE values. As Mittermaier (2006) found, only at two to three times the coarsest model resolution are forecast errors sufficiently eliminated to show underlying forecast skill. In this instance this equates to ~25 km. Results for 0.5 and 1 mm (6 h)$^{-1}$ suggest that the skilful spatial scale L (FSS $> 0.5 + f/2$) may be less than 50 km (depending on $f$) whilst for 4 mm (6 h)$^{-1}$ this is only reached when averaging to 100 km or more. For the 16 mm (6 h)$^{-1}$ threshold, it would seem unlikely that forecasts achieve a skilful scale (that would be practically useful) because the scores are considerably less than 0.5. Despite this the UK4 scores are still better at each spatial scale considered. It can be seen then that scores decrease with increasing threshold through a combination of increased biases (accentuated because the sign is not taken into account) and mismatch in location. Higher rainfall amounts are more likely to be associated with embedded or isolated convection which is often prone to lower predictability and larger errors.

### 4.3. Use of frequency thresholds

Using frequency thresholds is an effective method for removing the bias to focus on the spatial error. In doing so a spectrum of thresholds is assessed. The other benefit of using frequency thresholds is that it focuses on verifying the ‘rain of interest at the time’ in a specific forecast period. It is therefore a more inclusive event-based assessment. In this manner a 6 h period dominated by drizzle, and one with heavy convective downpours can be assessed together. The disadvantage is that the association with a physical threshold, that remains constant,

Table 3. Effective sample size $n$ ($N = 4996$) for significance testing as a function of lead time and 6 h accumulation threshold for a spatial scale of 25 km.

| | 0.2 mm | 0.5 mm | 1 mm | 2 mm | 4 mm | 8 mm | 16 mm | 32 mm |
|---|---|---|---|---|---|---|---|---|
| $t + 12$ (9) | 1551 (**9.664**) | 1971 (**9.234**) | 1435 (**8.649**) | 1351 (**8.388**) | 2052 (**10.383**) | 1504 (**11.534**) | 504 (**4.663**) | 47 (1.009) |
| $t + 18$ (15) | 1712 (**7.541**) | 1159 (**7.128**) | 1281 (**7.386**) | 1407 (**8.342**) | 1878 (**9.157**) | 1262 (**9.836**) | 544 (**4.463**) | 85 (0.991) |
| $t + 24$ (21) | 1742 (**7.087**) | 1833 (**6.959**) | 1514 (**7.975**) | 1392 (**8.871**) | 1873 (**9.592**) | 1380 (**12.814**) | 479 (**4.994**) | 37 (0.549) |
| $t + 30$ (27) | 1593 (**5.419**) | 648 (**5.259**) | 1264 (**7.163**) | 1451 (**8.601**) | 1350 (**10.966**) | 1399 (**11.008**) | 421 (**4.465**) | 41 (0.423) |
| $t + 36$ (33) | 2334 (**4.402**) | 1567 (**5.381**) | 1035 (**6.498**) | 1467 (**8.573**) | 1526 (**10.435**) | 1292 (**10.609**) | 464 (**3.371**) | 49 (1.289) |

The test statistic value $T$ is shown in brackets. For the null hypothesis to be rejected (i.e. differences in FSS are significant), $T$ must exceed the critical value 1.96 (indicated in bold).

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

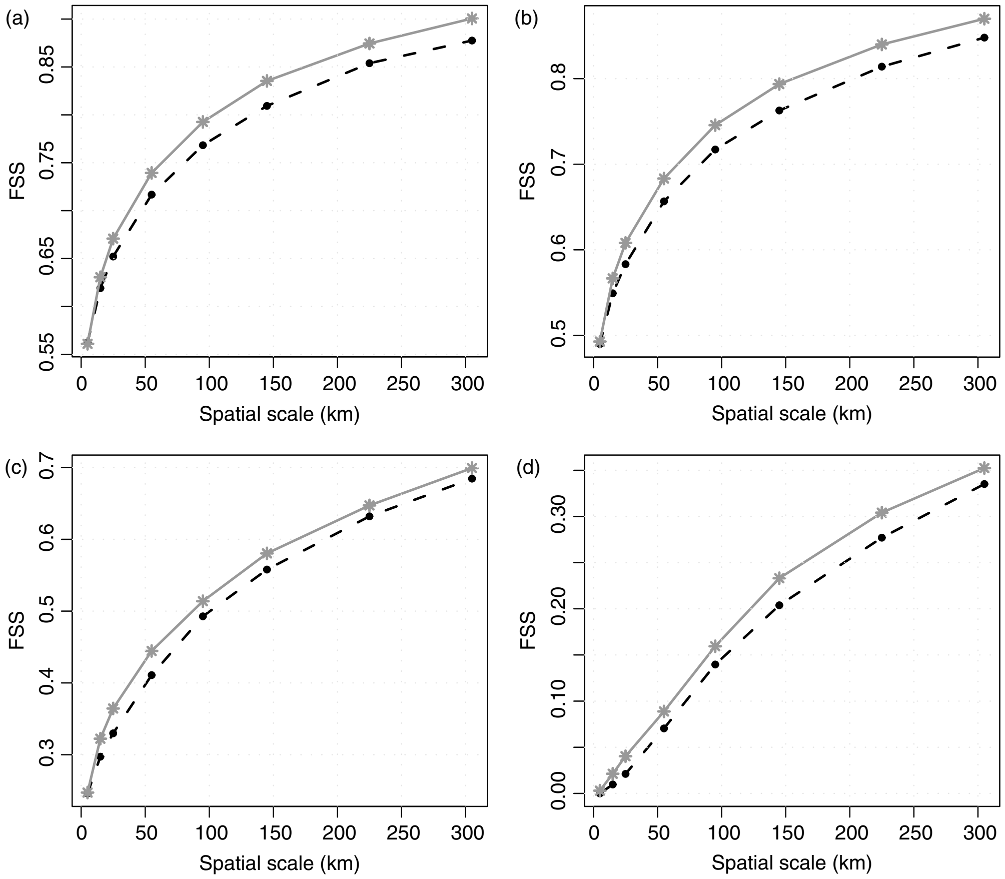

Figure 4. Median FSS as a function of spatial scale and threshold for $t + 24$ (21) h. —● ·; NAE; —✱—: UK4.

is lost. Figure 5 shows the FSS for the top 10% of thresholds (or 90th percentile) at $t + 24(21)$ h. These are representative of results at other times. Part (a) shows the distribution of values as a function of spatial scale and suggests that, in a median sense, differences in forecast skill between the NAE and UK4 are small at low spatial scales, with a gradual divergence in the median scores as the neighbourhood size increases, with the UK4 having higher median scores. The distribution of physical accumulation thresholds is shown in Figure 5(b). This highlights the fact that ~30% of forecasts assessed will have focused on the rain/no-rain threshold, but some events with thresholds greater than 12.5 mm (6 h)$^{-1}$ (half inch) are also considered. The skilful spatial scale in (c) reveals that the UK4 forecasts have skill at spatial scales ~10 km less (better) than the NAE. There is a hint of an upward trend in the NAE skilful scale (suggesting a degradation in skill), causing the curves to diverge. This echoes the results shown in Figure 3. The effective sample size and test statistic values for the 75th and 90th percentile thresholds are listed in Table 4. FSS differences for both thresholds are statistically significant at the 5% level.

### 4.4. What is the skilful spatial scale as a function of threshold?

The skilful scales increase with lead time, as expected, where lower (or smaller) is better. They also vary with season, being lower in the colder months which are more likely to be dominated by large frontal systems, and larger more homogeneous rain areas.

Table 4. Effective sample size $n$ ($N = 4996$) for significance testing as a function of lead time and 6 h percentile thresholds for a spatial scale of 25 km.

| | 90th | 75th |
|---|---|---|
| $t + 12$ (9) | 1340 (6.400) | 1996 (6.863) |
| $t + 18$ (15) | 2433 (5.847) | 1588 (6.340) |
| $t + 24$ (21) | 1788 (5.359) | 1693 (5.803) |
| $t + 30$ (27) | 1146 (5.703) | 1989 (6.570) |
| $t + 36$ (33) | 1791 (5.844) | 2037 (6.819) |

See Table 3 for more details.

Whilst there may well be an issue with degradation in the NAE performance, both models are compared against the same 'truth' and the results are therefore subject to the characteristics of the radar-rainfall analyses. Figure 6 seems to suggest that something else may also be going on. Here the 12-month running means of the L(FSS $>$ 0.5) are shown for the three UK index thresholds at $t + 36$ (33) h. The results for other lead times (not shown) are similar. All the thresholds essentially show the same trend. Whilst the difference is the smallest at 1 mm (6 h)$^{-1}$ and the UK4 forecasts have a smaller skilful spatial scale in general, the plots show that instead of the skilful spatial scale decreasing over time, as one would hope, both model forecasts appear to be losing ground. This sort of signal is often symptomatic of changes, or a drift in, the verifying baseline. Any impacts related to changes in radar processing are being investigated as a matter of urgency. Changes in the observations baseline introduce external trends in the bias, and represent the biggest risk in any model assessment. If this is

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

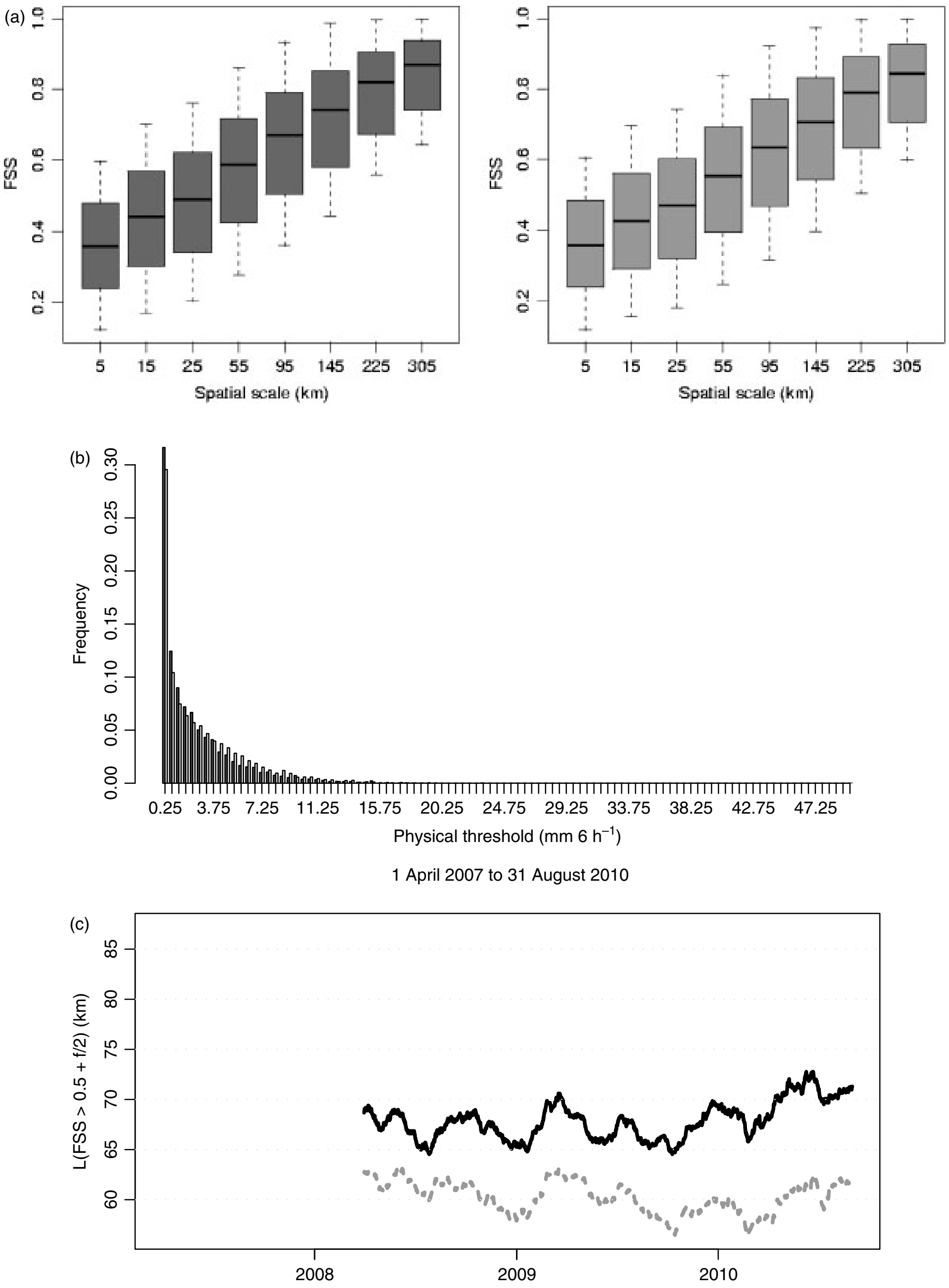

Figure 5. FSS for the top 10% of thresholds at $t + 24$ (21) h. (a) The distribution of scores as a function of spatial scale; (b) the distribution of physical thresholds; and (c) the skilful spatial scale expressed as a 12-month running mean. ■: NAE; □: UK4; ━: NAE; ╍ ·: UK4.

considered to be a considerable risk, using frequency thresholds for the FSS is strongly advocated.

### 4.5. Diurnal variations

The representation of the diurnal cycle is fundamental to overall model performance. To understand any diurnal variations in precipitation forecast skill forecasts were stratified into those valid at 0000, 0600, 1200 and 1800 UTC and analysed separately. Figure 7 shows the score difference (UK4-NAE) a function of physical thresholds and lead times. The score differences are positive throughout, demonstrating that the UK4 is better, and this is independent of the time of day. However, it is striking that the biggest differences are for forecasts valid at 0000 UTC for all lead times and thresholds less than 4 mm (6 h)$^{-1}$. From 4 mm (6 h)$^{-1}$ upwards the forecasts valid at 1800 UTC show the greatest differences. This possibly points towards the better representation of afternoon convection persisting into the evening and night more realistically in the UK4. The signal is more muted at 16 mm (6 h)$^{-1}$, probably due to fewer events.

## 5. Conclusions

In this study 4996 6 h precipitation accumulation forecasts from the NAE and UK4 configurations of the MetUM were compared to answer the fundamental question: are high-resolution km-scale precipitation forecasts better than their coarser resolution counterparts?

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

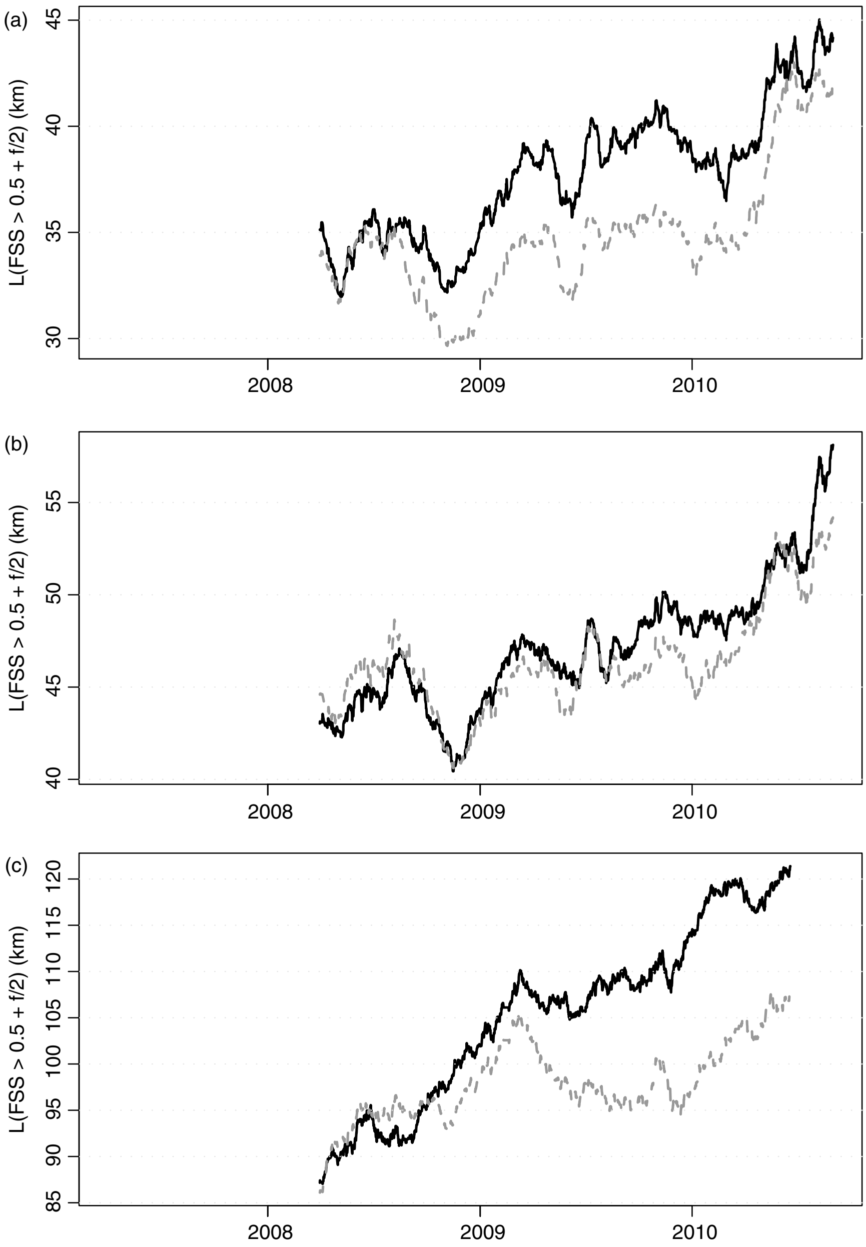

Figure 6. Twelve-month running means of the skilful spatial scale L(FSS > 0.5) for three UK index thresholds at $t + 36(33)$ h. —: NAE; – – : UK4.

Using the FSS approach the UK4 forecasts are statistically significantly better than the NAE forecasts at the 5% level (although the UK4 forecasts also benefit from a 3 h lead time advantage). Considering the time series as a whole the scores continue to diverge, suggesting that successive NAE model changes have a discernible detrimental impact on the NAE's precipitation forecast skill. Skilful spatial scales for the UK4 are at least 10 km less (therefore better) than the NAE, based on the top 10% frequency threshold.

Removal of the bias (by using a frequency threshold) is important for managing and minimizing potential impacts on the verification metrics from changes to the observations. It is important to stress that unlike gauges, radar-rainfall accumulations do not represent the same stable baseline to compare against. Just as with NWP models, radar hardware and software is subject to continual upgrade and change. Each change has an impact on the properties of the radar rainfall product, e.g. detection, bias. Therefore, for long-term monitoring of forecast performance the use of absolute intensity thresholds is potentially dangerous because the verification scores are affected by multiple sources of change, producing a mixed signal that cannot be disentangled. In the context of this study this is less important, as two models are compared against each other, both against the same (varying) truth.

Even when using frequency thresholds, understanding the bias is still important and this must be computed. In this case low thresholds appear to be almost unbiased, but the bias increases with increasing accumulation thresholds.

There are some differences in scores as a function of time of day, with UK4 forecasts being clearly superior for predicting afternoon/evening convection.

The use of the FSS has allowed determination in a meaningful way that there is an improvement in skill from increased resolution and quantify this in terms of the minimum scale over

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.

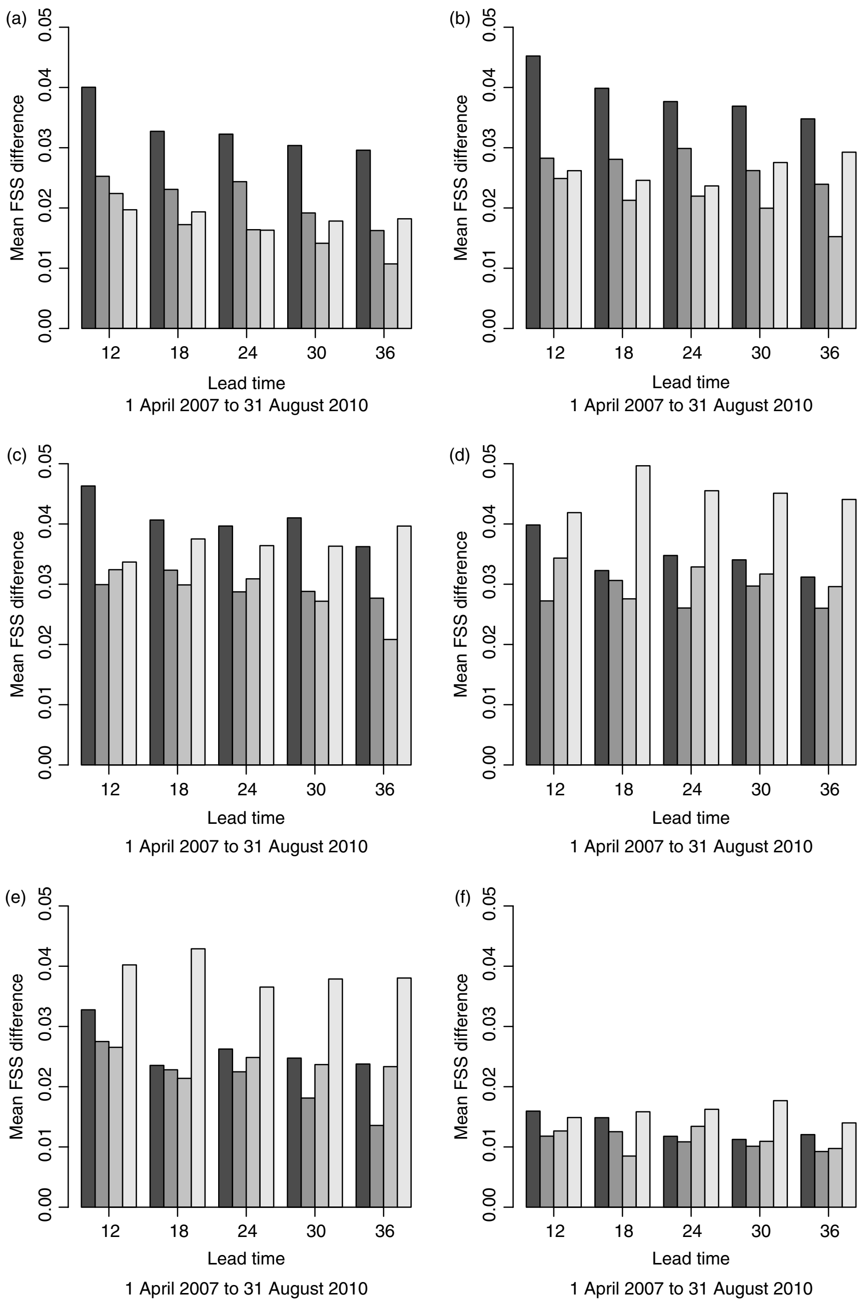

Figure 7. Diurnal mean FSS differences (UK4 - NAE) as a function of lead time and six-hourly accumulation thresholds. (a) 0.5, (b) 1.0, (c) 2.0, (d) 4.0, (e) 8.0 and (f) 16.0 mm (6 h)$^{-1}$. ■: 00Z; ▩: 06Z; ▨: 12Z; □: 18Z.

which a forecast is sufficiently skilful to be useful. In contrast, the continued use of the ETS for verifying km-scale models does not add to the understanding of true forecast skill, goodness or usefulness. In fact it can lead to very wrong decision making. It is recommended that the ETS be replaced by the FSS for the purpose of Met Office routine assessment of precipitation forecasts, with a strong endorsement for the use of a frequency thresholds for two reasons: greater immunity to bias effects (introduced by a variable observation baseline) and an enhanced event-based focus.

## Acknowledgements

The authors would like to thank Rob Darvell for his help with data extraction; Mike Bush for his insights into the NAE upgrades and Jorge Bornemann for the UK4 upgrades.

## A1. Appendix A: Testing for significant differences in verification scores

When establishing whether a package of model changes (applied to the same model, as test *versus* control) is an improvement, or assessing whether model 1 is better than model 2, it is recommended to test whether the change in scores is significant (although often operational parallel testing produces samples too small for detecting significant differences).

Verification scores computed for forecasts from different model configurations are not independent, as they are linked through the observations used, and forecasts are compared over the same time window. In this context the test for significance is based on testing whether the difference between dependent pairs is significant, as compared to simply testing that the mean $\mu_1 = \mu_2$. Testing the null hypothesis $H_0 : \ \mu_1 - \mu_2 = 0$ for two dependent samples. A stochastic variable $D$ is defined such that $D_i = i$th value in sample 1 minus the $i$th value in sample 2 for all $i = 1, 2, \ldots, N$ forecasts in the sample. The mean $\overline{D}$ and standard deviation $s_D$ of the differences (in scores) then form the basis of the test statistic $T$:

$$T = \overline{D}/SE \tag{A.1}$$

where the standard error $\mathrm{SE} = s_D/\sqrt{n}$. The effective sample size $n$ is only equal to $N$ if there is no autocorrelation in the time series. Generally $n < N$. Estimating $n$ requires estimating the spectral density at frequency zero which is achieved through fitting an autoregressive model (see e.g. Barnett, 2004). Without adjusting the sample size the statistical significance may be over-estimated, and may not even exist.

$T$ is used to evaluate the null hypothesis $H_0: \mu_1 - \mu_2 = 0$, i.e. that the scores are the same, and that the difference is not significant. The sample size determines the distribution from which critical values are drawn. For samples less than 30 the Student's $t$-distribution $t(n-1)$ should be used. For larger samples the normal distribution is appropriate. For a two-tailed test (i.e. testing inequality) $H_0$ is rejected if $t \leq -t_{n-1,\alpha/2}$ or $t \geq -t_{n-1,\alpha/2}$.

## References

Baldwin ME, Kain JS. 2006. Sensitivity of several performance measures to displacement error, bias, and event frequency. *Weather and Forecasting* **21**: 636–648.

Barnett V. 2004. *Environmental Statistics. Methods and Applications*. John Wiley and Sons: Chichester, UK; 293 pp.

Clark MR. 2011. An observational study of the exceptional "Ottery St Mary" thunderstorm of 30 October 2008. *Meteorological Applications* **18**(2): 137–154.

Ebert EE. 2009. Neighborhood verification – a strategy for rewarding close forecasts. *Weather and Forecasting* **24**(6): 1498–1510.

ECMWF. 2009. Annual Report. http://www.ecmwf.int/publications/annual_report/2009/pdf/Annual-report-2009-small.pdf

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

Gilleland E, Ahijevych D, Brown BG, Casati B, Ebert EE. 2009. Intercomparison of spatial forecast verification methods. *Weather and Forecasting* **24**(5): 1416–1430.

Gilleland E, Ahijevych D, Brown BG, Ebert EE. 2010. Verifying forecasts spatially. *Bulletin of the American Meteorological Society* **91**(10): 1365–1373.

Grahame N, Riddaway B, Eadie E, Hall B, McCallum E. 2009. Exceptional hailstorm hits Ottery St Mary on 30 October 2008. *Weather* **64**(10): 255–263.

Lorenz E. 1969. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Science* **26**: 636–646.

May B, Clark P, Cooper A, Forbes R, Golding B, Hand W, Lean H, Pierce C, Roberts N, Smith R. 2004. Flooding at Boscastle, Cornwall on 16 August 2004 – A study of Met Office Forecasting Systems. Forecasting Technical Report 429. Exeter, UK; 46 pp.

Mittermaier MP. 2006. Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmopheric Science Letters* **7**(2): 35–42.

Mittermaier MP. 2007. Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quarterly Journal of the Royal Meteorological Society* **133**: 1487–1500.

Mittermaier MP, Roberts N. 2010. Inter-comparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score. *Weather and Forecasting* **25**: 343–354.

Murphy AH. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**: 281–293.

Roberts NM. 2008. Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications* **15**: 163–169.

Roberts NM, Lean HW. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review* **136**: 78–96.

Rodwell M, Richardson D, Hewson T, Haiden T. 2010. A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* **136**: 1344–1363.

Theis SE, Hense A, Damrath U. 2005. Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications* **12**: 257–268.

Wilks DS. 2006. *Statistical Methods in the Atmospheric Sciences*, 2nd edn. Academic Press: Burlington, MA; 627 pp.

Copyright © 2011 British Crown copyright, the Met Office.
Published by John Wiley & Sons Ltd.