Meteorol. Appl. 11, 141–154 (2004) DOI:10.1017/S1350482704001239

A new intensity-scale approach for the verification of spatial precipitation forecasts

B. Casati[1], G. Ross[2] & D. B. Stephenson[1]
[1]*Department of Meteorology, University of Reading, Earley Gate PO Box 243, RG6 6BB, Reading, UK.*
Email: b.casati@reading.ac.uk; D.B.Stephenson@reading.ac.uk
[2]*Met Office, Fitzroy Road, Exeter, Devon, EX1 3PB, UK.*
Email: gross@meto.gov.uk

A new intensity-scale method for verifying spatial precipitation forecasts is introduced. The technique provides a way of evaluating the forecast skill as a function of precipitation rate intensity and spatial scale of the error. Six selected case studies of the UK Met Office now-casting system NIMROD are used to illustrate the method.

The forecasts are assessed using the Mean Squared Error (MSE) skill score of binary images, obtained from the forecasts and analyses by thresholding at different precipitation rate intensities. The skill score is decomposed on different spatial scales using a two-dimensional discrete Haar wavelet decomposition of binary error images. The forecast skill can then be evaluated in terms of precipitation rate intensity and spatial scale.

The technique reveals that loss of forecast skill in NIMROD is predominantly due to small spatial scale (< 40 km) errors of more intense events. The technique is capable of isolating specific intensity-scale errors for individual cases. As an example, in one of the case studies the displacement error of an incorrectly advected storm is well detected by a minimum negative skill score occurring at the 160 km spatial scale for thresholds between 1/2 and 4 mm/h.

1. Introduction

Verification of Quantitative Precipitation Forecasts (QPFs) is one of the most challenging task of forecast verification (for a recent review of QPF see Collier & Krzysztofowicz 2001). Precipitation is highly discontinuous in space and time; its distribution is positively skewed and characterised by the presence of many zero values; spatial maps are very noisy and often contain large outliers. These characteristics make verification of QPFs a difficult yet exciting area of research.

QPFs are traditionally assessed using a variety of both continuous and categorical verification approaches or by exploratory methods (Bougeault 2003; Ebert et al. 2003; FOAG 1993; Johnson & Olsen 1998; Airey & Hulme 1995; Osborn & Hulme 1998). The most commonly used continuous scores are the Root Mean Squared Error (RMSE) and the product moment correlation coefficient. The most commonly used categorical scores are the equitable threat score, frequency bias, hit rate, false alarm ratio and ROC curve. Exploratory methods are typically based on the comparison of forecast and observation means, standard deviations, maxima, distributions and cumulative frequencies. A recent and comprehensive review of these scores and verification methods and their interpretation can be found in Jolliffe & Stephenson (2003).

Traditional verification scores do not fully account for the unique characteristics of precipitation. For example, the widely used RMSE and the product moment correlation coefficient are sensitive to discontinuities, noise and outliers. Moreover, verification scores for continuous univariate forecasts do not account for the complex spatial interdependency of precipitation values. Categorical verification scores often deal somewhat better with some of the features of precipitation fields and are generally more widely used for QPF verification. However, many of these scores are overly sensitive to the base rate of the event and to the bias (Doswell et al. 1990; Woodcock 1976; Schaefer 1990; Marzban 1998; Göeber et al. 2003).

Previous QPF studies have shown the importance of discriminating between the different sources of forecast error (e.g. Browning et al. 1992). Recently, attention in the QPF community has focused on developing verification techniques which distinguish different types of error (e.g. position or amount of the precipitation features) and which assess separately the different attributes of individual precipitation features. Hoffman et al. (1995) introduced a verification approach based on the decomposition of the forecast error into displacement, amplitude and residual error. The horizontal displacement was obtained by translating the forecast features over the observation features until a 'best fit criterion' was satisfied (e.g. minimisation of the RMSE). The method was originally applied to precipitable water, wind and 500 hPa geopotential height fields. Later Du et al. (2000) applied the method to precipitation fields. Hoffman & Grassotti (1996) and more recently Nehrkorn et al. (2003) have extended the work of Hoffman et al. (1995) by using a spherical harmonic representation of the fields. The method has been developed into a variational analysis technique for use in data assimilation (Brewster 2003). Based on Hoffman et al. (1995), Ebert & McBride (2000) developed a QPF feature-based verification method based on the decomposition of the forecast-observation disagreement into displacement, volume and pattern error. Forecast and observed precipitation features were isolated into individual precipitation events within contiguous rain areas. For each contiguous rain area, the horizontal displacement was obtained by translating the forecast precipitation feature over the observed feature until the MSE was minimised. A novel contingency table, based on displacement and amount error categories, was used to assess Australian precipitation forecasts (Ebert 2001). Baldwin et al. (2001) developed an event-oriented verification approach in which each precipitation event (e.g. convective cell) was isolated and then described by a set of attributes. The forecast-observation disagreement was evaluated by comparing the attributes of paired forecast and observed events. Verification scores obtained from the covariance matrix and from a generalised Euclidean distance matrix associated to the attributes of forecast and observed events was used to assess the forecast performance. Brown et al. (2002) developed an alternative object-based approach in which forecast and observed precipitation events were modelled as basic geometrical shapes, such as ellipses. Comparison of the attributes of each object-shape (such as the centroid location, axis orientation, eccentricity, axis magnitude) were then used to diagnose different types of forecast-observation disagreement (such as location, orientation, shape, size).

Another issue that has recently received attention is the evaluation of forecast skill on different spatial scales. Precipitation events on different spatial scales (e.g. showers or frontal systems) are caused by different physical processes (e.g. convection or large-scale ascent). Verification of different spatial scales can provide deeper insights into model performance at simulating these different processes. Zepeda-Arce & Foufoula-Georgiou (2000) evaluated precipitation forecasts on different spatial scales by using threat score and depth-area-duration curves. The different scales were obtained by averaging the precipitation values of the original fields over the grid size (scale) of interest. The forecast ability to reproduce the multi-scale spatial structure and space-time dynamics of the precipitation field was assessed by evaluating scale-invariant parameters related to the scale-to-scale spatial variability of precipitation field and its time-scale evolution. Briggs & Levine (1997) developed a multi-scale verification technique based on wavelets. The technique was applied to geopotential height fields, but could be extended to precipitation fields. Forecast and observed 500 mb geopotential height fields were decomposed into the sum of components on different spatial scales by using a two-dimensional discrete wavelet transform. Forecast-observation disagreement was then assessed on different spatial scales using RMSE, correlation coefficient and energy ratio.

An alternative intensity-scale approach for the verification of spatial precipitation forecasts is introduced here. The technique allows one to assess the forecast skill as a function of precipitation rate intensity and spatial scale of the error. The technique is demonstrated on six carefully selected representative case studies of the Met Office now-cast forecasting system NIMROD. The case studies and selection criteria are described in Section 2. The verification technique is described in Section 3 and results are presented in Section 4. Conclusions are given in Section 5.

2. NIMROD precipitation rate forecasts

NIMROD is the very short-range mesoscale Numerical Weather Prediction (NWP) system used operationally at the Met Office, UK (Golding 1998). NIMROD produces hourly precipitation rate forecast and analysis images over the UK National Grid (435×345 grid point spatial domain, regularly gridded with a resolution of 5 km, covering the UK and surrounding areas). Both NIMROD analysis and forecast fields are produced every 15 minutes. The NIMROD precipitation rate analysis is estimated from the UK radar network images, merged and corrected by quality-control statistical algorithms and parameterisations which make use of satellite and surface observations and the Met Office mesoscale model outputs (Harrison et al. 2000). NIMROD precipitation rate forecasts are produced by combining now-casting advection techniques with the mesoscale model forecasts, and then correcting the product with a parameterisation based on the local climatology characteristics (Golding 2000). NIMROD forecasts have lead times in the nowcasting range that extends up to six hours ahead. As the lead time increases, the forecast process gives less weight to the advection techniques and more weight to the mesoscale model forecasts.

Table 1. *NIMROD case studies and their main characteristics.*

CASE	Date	Time (UTC)	Synoptic Situation
1	26/01/99	16:00	Front detected, showers missed and mis-handled.
2	13/04/99	12:00	System of showers mis-handled.
3	29/05/99	15:00	Advection of intense storm incorrect.
4	29/05/99	18:00	Better advection of the intense storm.
5	05/07/99	18:00	Heavy showers partially displaced.
6	05/11/99	14:00	Frontal system timing error.

Six NIMROD case studies were evaluated in this study. NIMROD precipitation rate forecasts at lead times of three hours were verified against their corresponding analyses. Forecasts at this lead time were chosen because they make almost equal use of both the now-casting advection technique and the mesoscale model NWP outputs. The evaluation was performed over an area of 1280×1280 km ($2^8 \times 2^8 = 256 \times 256$ pixels spatial domain), which constitutes a suitable (dyadic) number of grid points for the two-dimensional wavelet transform algorithm (Appendix A).

The six case studies were carefully selected to represent the typical NIMROD forecast errors (personal communication, Will Hand). Moreover, the case studies were chosen to include a variety of precipitation features on different spatial scales representative of synoptic situations of interest. Table 1 lists the six case studies and their main characteristics. Figures 1 and 2 show the analyses and corresponding forecasts, respectively, for the six case studies. Case 1 is an example of a well-detected frontal system. However, intense rainfall rates within the front were forecast with reduced intensity and too much drizzle (1/32–1/2 mm/h) was forecast. Some showers in Wales were missed and other showers to the north of Ireland were misplaced. Case 2 is an example of a mis-handled system of showers. The amplitude of intense rainfall rates was under-estimated and too much drizzle was forecast. Case 3 shows an intense storm of about 100–200 km spatial scale displaced nearly its entire length and slightly rotated. Intense precipitation within the storm was forecast with reduced intensity and too much drizzle was forecast. Case 4 shows the same storm, three hours later, better advected, but the forecast is still late. Intense precipitation within the storm was forecast with larger intensities and extent than observed. Once again, too much drizzle was still forecast. Case 5 shows five shower systems that were detected, but heavy showers within these were partially displaced. Too much drizzle was forecast. The precipitation features in the east of France and in the south of Norway were forecast, respectively, with reduced and larger intensities. Case 6 shows a front timing error. Drizzle and low rainfall rates up to 2 mm/h were forecast with larger intensities and extent than observed.

3. The verification method

3.1. Data pre-processing and forecast recalibration

NIMROD analyses and forecasts were first pre-processed, to obtain more reliable data before verification. All the non-zero precipitation values were *dithered* by adding a very small amount of uniformly distributed noise in the range $(-1/64, 1/64)$ mm/h, equal to the discretisation round-off error of the data (further information on the dithering process can be found on the web at *http://www.cadenzarecording.com/dither.html*). Dithering helps compensate for the discretisation effects caused by the finite precision storage of the precipitation rate values (i.e. binary integer multiples of 1/32 mm/h). The precipitation rate values were then *normalised* by performing a (base 2) logarithmic transformation. The no-rain pixels (0 mm/h) were assigned the value of -6, since the smallest non-zero precipitation value (after the dithering) is always larger than $1/64 = 2^{-6}$ mm/h. The logarithmic transformation reduces skewness and produces more normally distributed values.

Forecasts were then recalibrated by substituting each value of the forecast image with the value of the analysis image having the same empirical cumulative probability. This non-linear transformation is described by the recalibration function:

$$Y' = F_X^{-1}(F_Y(Y)), \tag{1}$$

where X is the analysis, Y the forecast, Y' the recalibrated forecast and F_X and F_Y are the empirical cumulative distribution functions of analysis and forecast, respectively. The recalibration eliminates bias in the marginal distribution of the forecast precipitation. Figure 3 shows the effect of the recalibration on case study 6. The excessive precipitation forecast over the North Sea is substantially improved by the recalibration procedure. The effect of the recalibration is revealed by the empirical recalibration function. Figure 4 shows the empirical recalibration functions $F_X^{-1} \circ F_Y$ for the six case studies. For all the case studies the curves exhibit a deviation below the main diagonal at low precipitation rates (drizzle) and a slight deviation above the main diagonal (except case 4) at

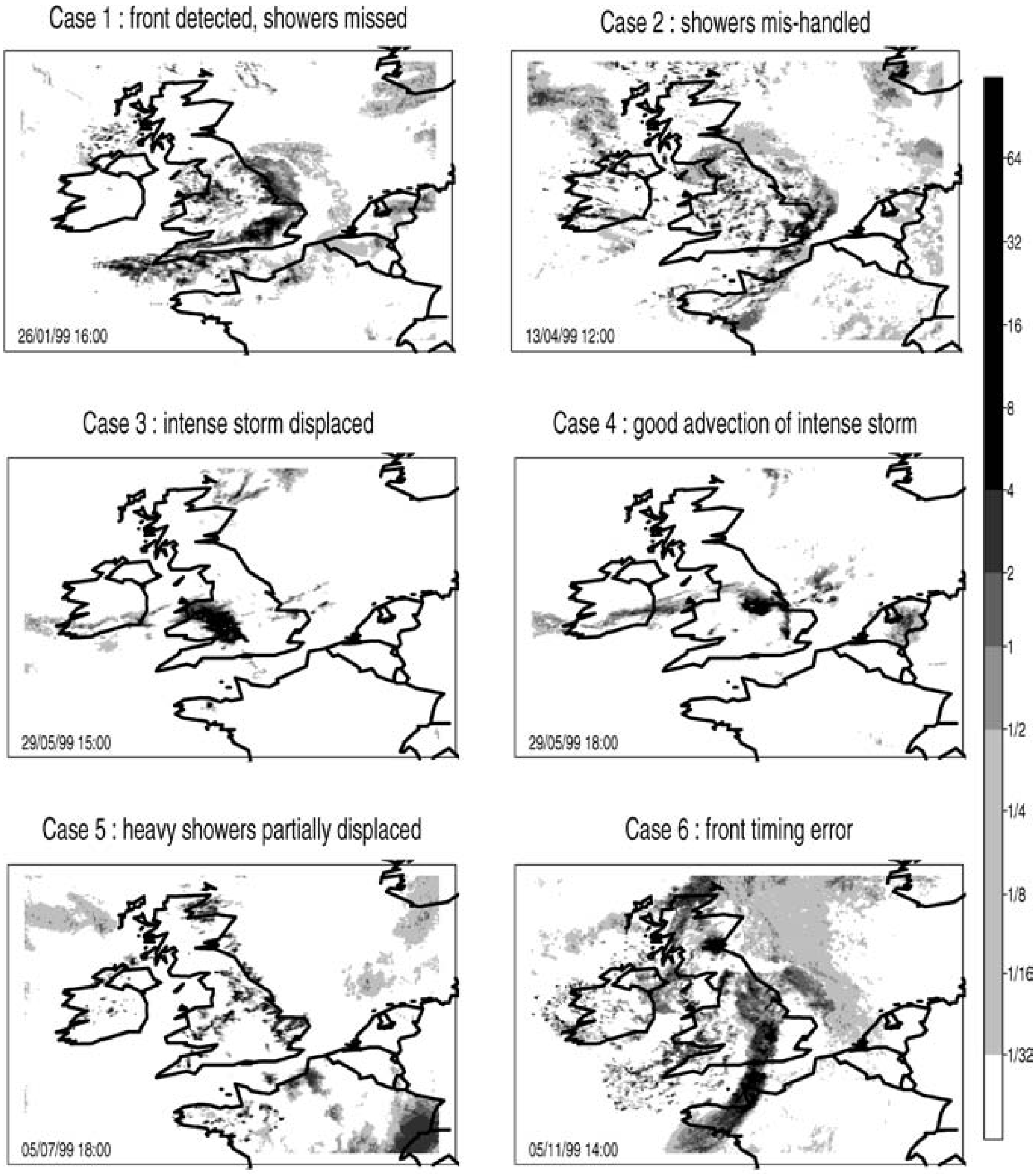

Figure 1. *NIMROD precipitation rate (mm/h) analyses for the six case studies.*

high precipitation rates. This behaviour shows that the forecasts systematically forecast too many small precipitation rate events (drizzle). Parameterisation of these recalibration functions could be used to help calibrate precipitation forecasts of future events. Note that the deviation of the recalibration functions from the main diagonal provides a measure of the bias.

3.2. Intensity-scale verification

3.2.1. Binary error decomposition

Thresholding is used to convert the recalibrated forecast (Y') and analysis (X) into binary images

$$I_{Y'} = \begin{cases} 1 & Y' > u \\ 0 & Y' \leq u \end{cases} \quad I_X = \begin{cases} 1 & X > u \\ 0 & X \leq u \end{cases} \tag{2}$$

for each of the rainfall rate thresholds $u = 0, 1/32, 1/16, \ldots, 128$ mm/h. The difference between binary recalibrated forecast and analysis defines the binary error

$$Z = I_{Y'} - I_X. \tag{3}$$

Figure 5 shows an example of binary analysis (I_X), binary recalibrated forecast ($I_{Y'}$) and binary error ($Z = I_{Y'} - I_X$) for case study 6 with rainfall rate threshold $u = 1$ mm/h.

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:

$$Z = \sum_{l=1}^{L} Z_l. \tag{4}$$

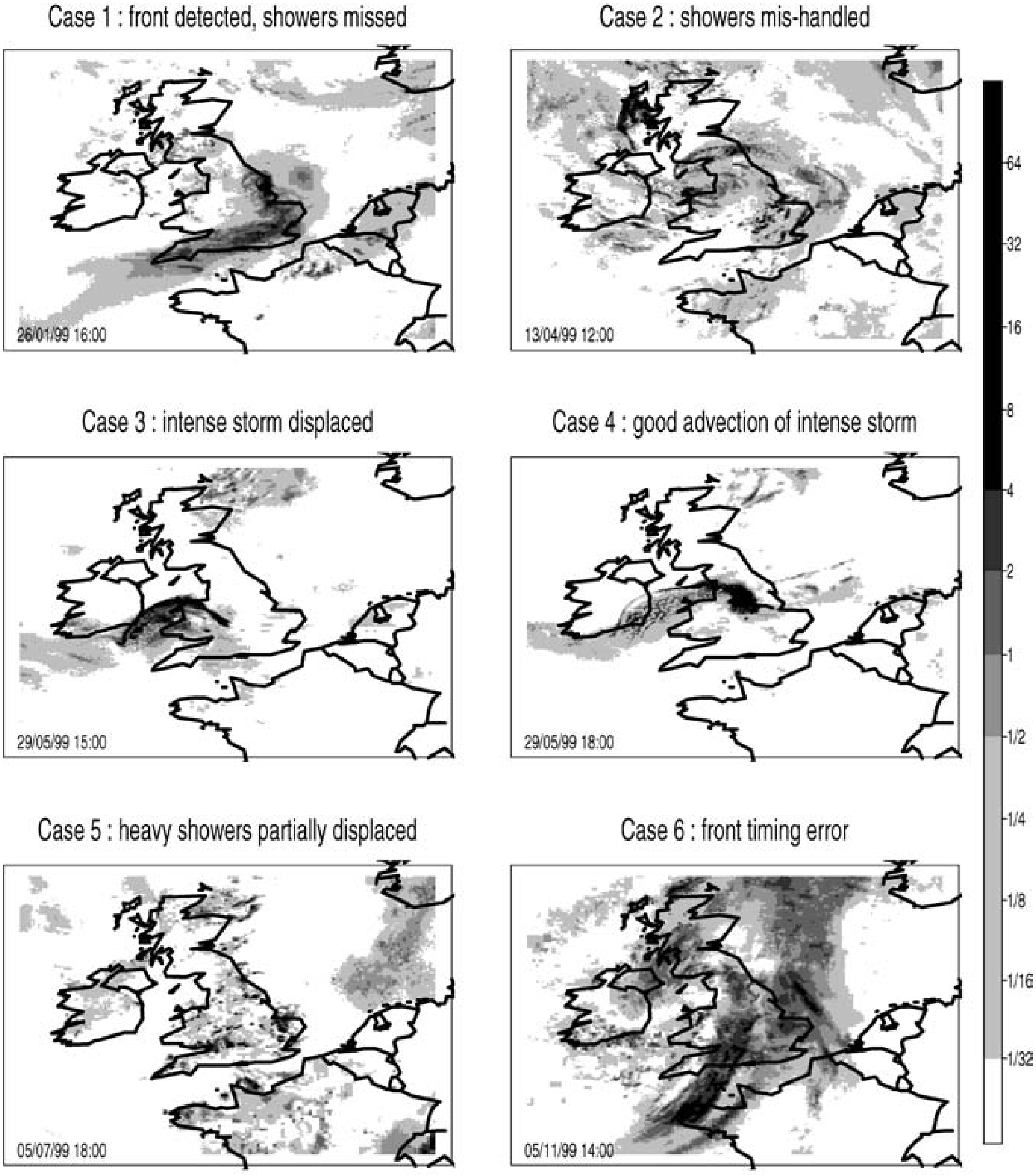

Figure 2. *NIMROD precipitation rate (mm/h) forecasts at three hours lead time for the six case studies.*

Appendix A describes in detail the two-dimensional discrete Haar wavelet decomposition algorithm. Note that although wavelets provide a rigorous and elegant mathematical framework, the two-dimensional discrete Haar wavelet decomposition can be obtained more simply by averaging over square regions on different scales. Figure 6 shows the most substantial mother wavelet components, Z_l, obtained from the two-dimensional discrete Haar wavelet decomposition of the binary error image shown in Figure 5. The spatial scales $l = 1, \ldots, L = 8$ correspond to 5, 10, 20, 40, 80, 160, 320, 640 km resolution of the binary error image mother wavelet components. It is important to note that l refers to the spatial scale of the *error*, and not to the spatial scale of the precipitation features or their displacement.

3.2.2. Binary Mean Squared Error

The MSE of the binary error image is given by

$$MSE = \overline{Z^2}, \tag{5}$$

where $\overline{\cdot}$ denotes the average over all the pixels in the domain. Since the wavelet components obtained from a discrete wavelet decomposition are orthogonal, then

$$\overline{Z_l Z_{l'}} = 0 \quad l \neq l'. \tag{6}$$

From this result and Eqn. (4), the MSE of the binary error image can be written as

$$MSE = \sum_{l=1}^{L} \sum_{l'=1}^{L} \overline{Z_l Z_{l'}} = \sum_{l=1}^{L} \overline{Z_l^2}. \tag{7}$$

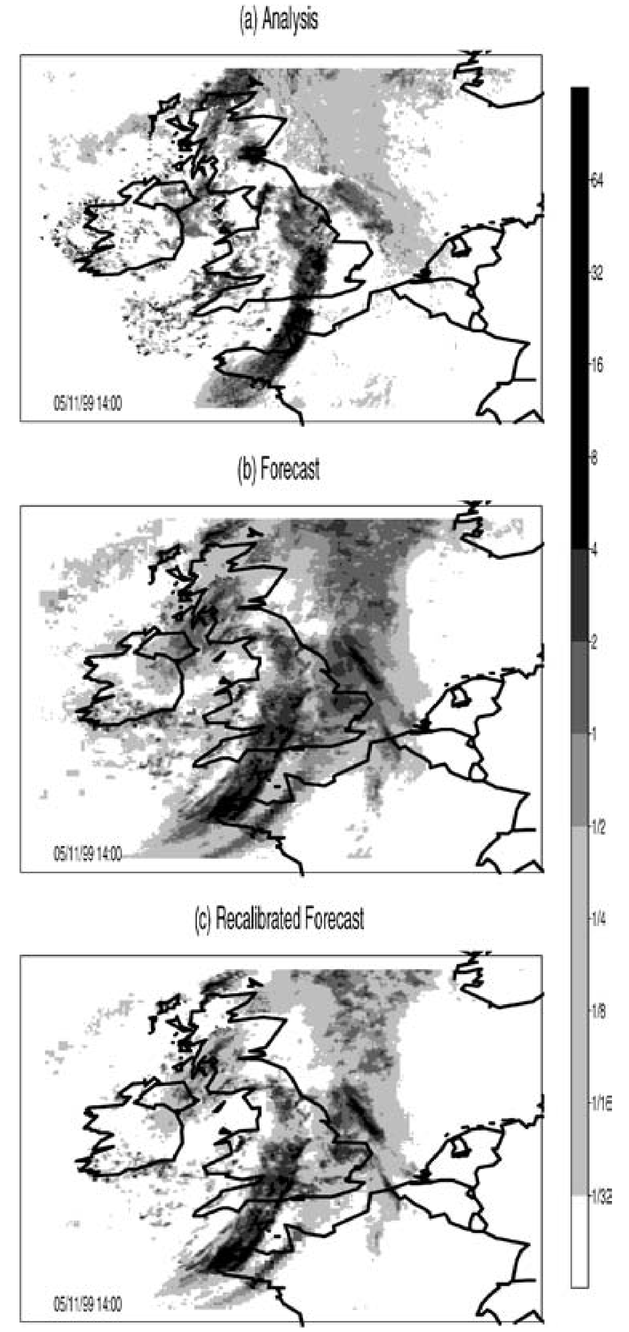

Figure 3. *(a) Analysis, (b) forecast and (c) recalibrated forecast for case study 6. The frontal system is forecast with a timing error. Drizzle and low rainfall rates up to 2 mm/h were forecast with larger intensities and extent than observed. The excessive precipitation forecast over the North Sea is substantially improved by the recalibration procedure.*

Therefore the MSE of the binary error image is equal to

$$MSE = \sum_{l=1}^{L} MSE_l, \tag{8}$$

where $MSE_l = \overline{Z_l^2}$ is the MSE of the l^{th} spatial scale component of the binary error image.

Note that the binary MSE_l depends both on the threshold u and on the spatial scale l. It therefore allows verification to be interpretated for different precipitation rate intensities and for different spatial scales. Moreover, any score which can be expressed as a linear function of the binary MSE can be written as the sum of components on different spatial scales. This also enables such scores to be evaluated as functions of precipitation rate intensity and spatial scale.

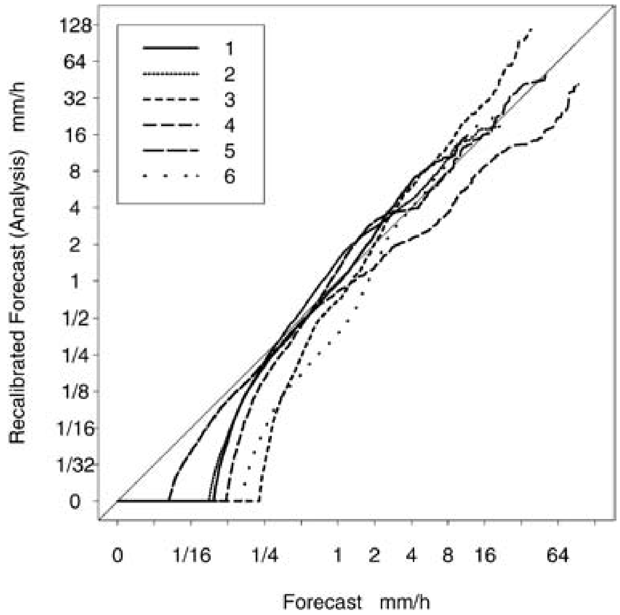

Figure 4. *Empirical recalibration functions (Eqn. 1) associated with the six case studies. For all the case studies the curves exhibit a deviation below the main diagonal at low precipitation rates (drizzle) and a slight deviation above the main diagonal (except case 4) at high precipitation rates. This behaviour shows that in general the forecasts forecast too many small precipitation rate events (drizzle) and too few intense precipitation rate events.*

3.2.3. Skill score

For each precipitation rate threshold, the binary MSE skill score can be calculated relative to the MSE of a random no-skill forecast:

$$SS = \frac{MSE - MSE_{\text{random}}}{MSE_{\text{best}} - MSE_{\text{random}}} = 1 - \frac{MSE}{2\varepsilon(1-\varepsilon)}, \tag{9}$$

where $MSE_{\text{best}} = 0$ is the MSE associated with a perfect forecast, $MSE_{\text{random}} = 2\varepsilon(1-\varepsilon)$ is the expected value of the MSE of the binary images generated randomly with no spatial dependency and ε is the base rate (fraction of rain pixels in the binary analysis image). Random binary recalibrated forecast and analysis fields are assumed to be independent Bernoulli distributed variables, $I_X \sim Be(\varepsilon)$ and $I_{Y'} \sim Be(\varepsilon)$, with (unbiased) means $E(I_X) = E(I_{Y'}) = \varepsilon$ and variances $\sigma^2_{I_X} = \sigma^2_{I_{Y'}} = \varepsilon(1-\varepsilon)$. The binary error, $Z = I_{Y'} - I_X$, can then easily be shown to have mean $E(Z) = 0$ and variance $\sigma^2_Z = \sigma^2_{I_{Y'}} + \sigma^2_{I_X} = 2\varepsilon(1-\varepsilon)$. The expected value of the MSE is then given by $MSE_{\text{random}} = E(Z^2) = E(Z - E(Z))^2 = \sigma^2_Z = 2\varepsilon(1-\varepsilon)$. Using Eqn.s (8) and (9), the skill score SS

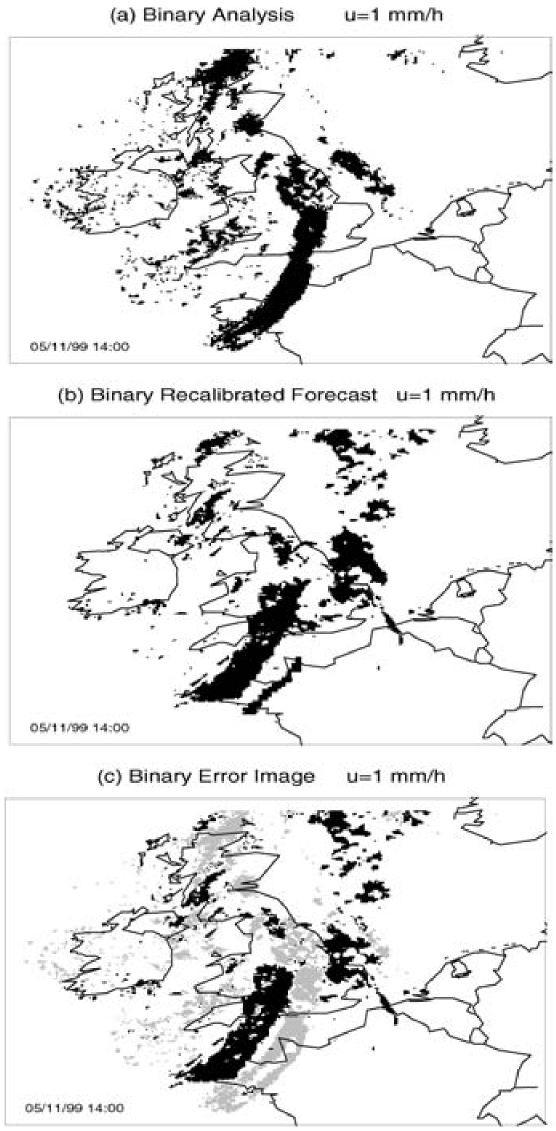

Figure 5. *(a) Binary analysis I_X, (b) binary recalibrated forecast $I_{Y'}$, and (c) binary error $Z = I_{Y'} - I_X$ for case study 6 with rainfall rate threshold $u = 1$ mm/h. Pixels with value equal to 1 are black, pixels with value equal to zero are white and pixels with value equal to -1 are shaded grey. The binary error image clearly shows the timing error that occurred in forecasting the frontal system.*

can be written as the arithmetic mean of its components on different spatial scales

$$SS_l = 1 - \frac{MSE_l}{2\varepsilon(1-\varepsilon)/L}, \tag{10}$$

where it has been assumed that the MSE of random forecast is equipartitioned over all the scales.

It is important to note that MSE alone cannot discriminate between small spatial scale errors due to small displacements of large spatial scale features and errors due to displacements of small spatial scale features. However the skill score is capable of discriminating between these cases because it also takes into account the base rate ε. Figure 7 shows one-dimensional idealised cases of a front slightly displaced and a shower displaced. The two different synoptic situations generate the same error and for both the cases $MSE = 0.04$. However the front case has a much larger base rate than the shower case ($\varepsilon_{\text{front}} = 0.5 > \varepsilon_{\text{shower}} = 0.02$). Therefore the front case has a much larger skill score than the shower case ($SS_{\text{front}} = 0.92 > SS_{\text{shower}} = -0.02$). This shows that for the front case the displacement is negligible with respect to the scale of the front and the forecast performance is good, whereas for the shower case the precipitation feature is entirely displaced and the forecast performance is worse than for the no-skill random forecast.

Table 2. *Contingency table. The counts a, b, c, d are the total number of hits, false alarms, misses and correct rejections. Note that for unbiased forecasts (such as the recalibrated forecast) the number of false alarms and misses is equal ($b = c$).*

	$X > \mathrm{u}$	$X \le \mathrm{u}$	
$Y' > \mathrm{u}$	a	b	a + b
$Y' \le \mathrm{u}$	c	d	c + d
	a + c	b + d	n

3.3. Links with categorical verification scores

The intensity-scale verification technique is a spatial generalisation of traditional binary verification (Jolliffe & Stephenson 2003, chapter 3). In this section we show that the binary MSE skill score defined in Eqn. (9) is equal to the Heidke skill score and to the Peirce skill score.

NIMROD recalibrated forecast (Y') and analysis (X) were transformed into dichotomous events on a rain/no-rain basis for the rainfall rate thresholds $u = 0, 1/32, 1/16, \ldots, 128$ mm/h. For each of these thresholds, a contingency table can be constructed by compiling over all pixels in the images (Table 2). The forecast performance can be summarised by the joint probabilities estimated by the frequencies $\{a/n, b/n, c/n, d/n\}$ (Murphy & Winkler 1987). The binary MSE is given by

$$\begin{aligned} MSE &= \overline{Z^2} = \frac{1}{n}(a \cdot 0^2 + b \cdot 1^2 + c \cdot (-1)^2 + d \cdot 0^2) \\ &= \frac{b+c}{n}. \end{aligned} \tag{11}$$

Three statistics are sufficient to fully describe the joint distribution of binary events (Stephenson 2000). However, for unbiased forecasts, such as our recalibrated forecast, only two statistics are needed to fully describe the joint distribution. For example, the base rate $\varepsilon = (a + c)/n$ and the binary MSE can be used to express the frequencies $\{a/n, b/n, c/n, d/n\}$. The base rate $\varepsilon = \varepsilon(u)$ is a monotonically decreasing function of the threshold and so can be used instead of u to denote the precipitation rate intensity.

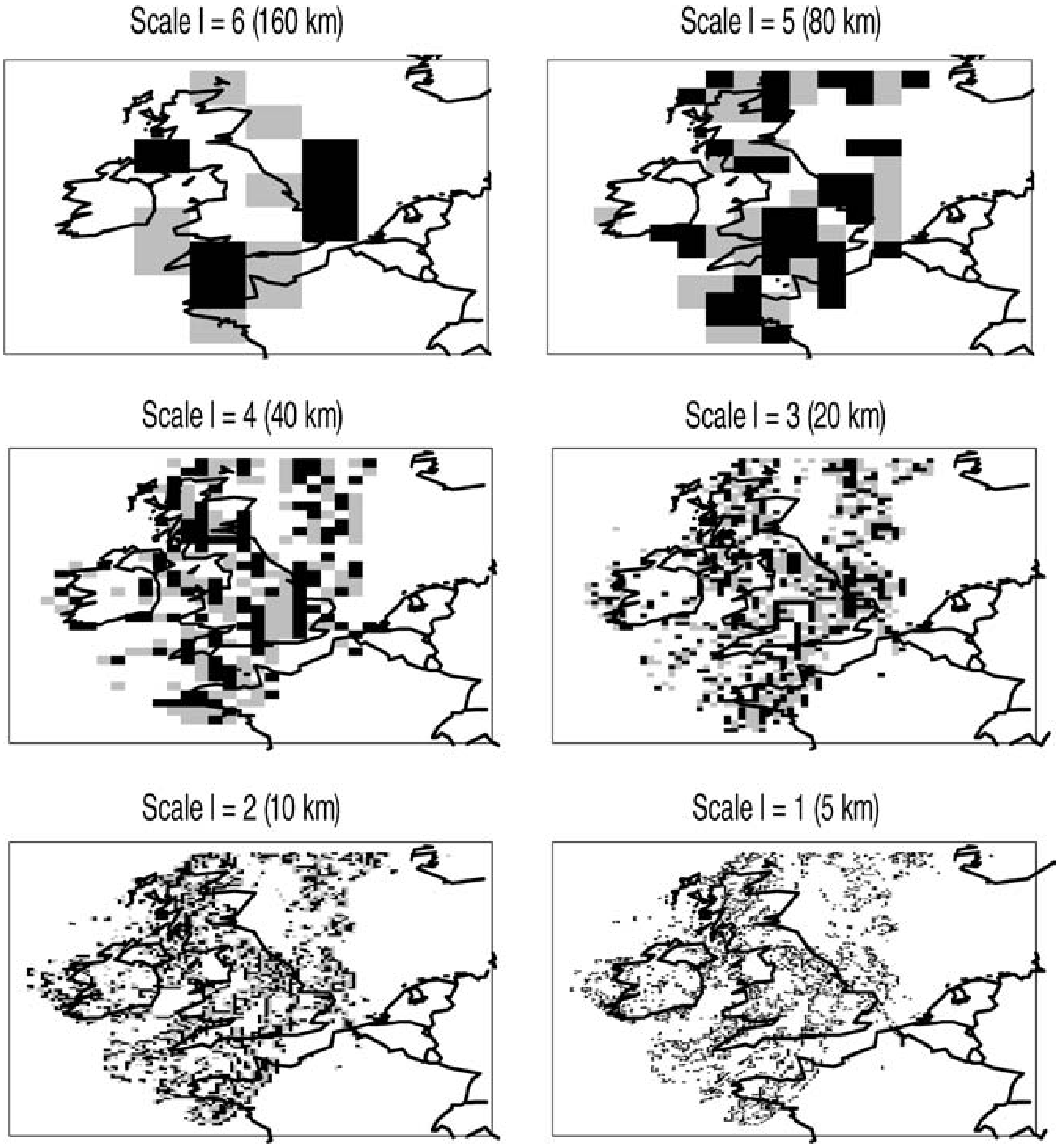

Figure 6. *Mother wavelet components obtained from the two-dimensional discrete Haar wavelet decomposition of the binary error image for case study 6 with rainfall rate threshold $u = 1$ mm/h. Pixels with values larger than 0.1 (up to 1) are black; pixels with values between -0.1 and 0.1 are white; pixels with values smaller than -0.1 (down to -1) are shaded grey. The binary error image is equal to the sum of its mother wavelet components on the spatial scales $l = 1, \ldots, L = 8$ (corresponding to 5, 10, 20, 40, 80, 160, 320, 640 km resolution). The components of the two largest spatial scales ($l = 7, 8$) have much smaller magnitudes (< 0.1) and so are not shown.*

The joint probabilities of binary recalibrated forecast and analysis are given in terms of ε and MSE by:

$$\frac{a}{n} = \varepsilon - \frac{MSE}{2}; \quad \frac{b}{n} = \frac{c}{n} = \frac{MSE}{2}; \quad \frac{d}{n} = 1 - \varepsilon - \frac{MSE}{2}. \tag{12}$$

Any categorical verification statistic, score or skill score evaluated from these joint probabilities can be expressed as a function of the base rate and the binary MSE. It can be shown that the Heidke skill score is identical to the binary MSE skill score given by Eqn. (9):

$$HSS = \frac{\frac{a}{n} + \frac{d}{n} - \frac{a+b}{n}\frac{a+c}{n} - \frac{d+b}{n}\frac{d+c}{n}}{1 - \frac{a+b}{n}\frac{a+c}{n} - \frac{d+b}{n}\frac{d+c}{n}} = 1 - \frac{MSE}{2\varepsilon(1-\varepsilon)} = SS. \tag{13}$$

Furthermore, since the recalibrated forecast is unbiased, the Peirce skill score is also equal to the binary MSE skill score. Therefore, the intensity-scale evaluation based on the binary MSE skill score is equivalent to an intensity-scale evaluation of the Heidke skill score or Peirce skill score.

4. Results

Figure 8 shows two-dimensional plots of the binary MSE components MSE_l as a function of threshold u and spatial scale l for the six case studies. For all the case studies, the MSE decreases for large spatial scales and for larger thresholds (rarer more intense events). Most MSE is due to commonly occurring errors on small

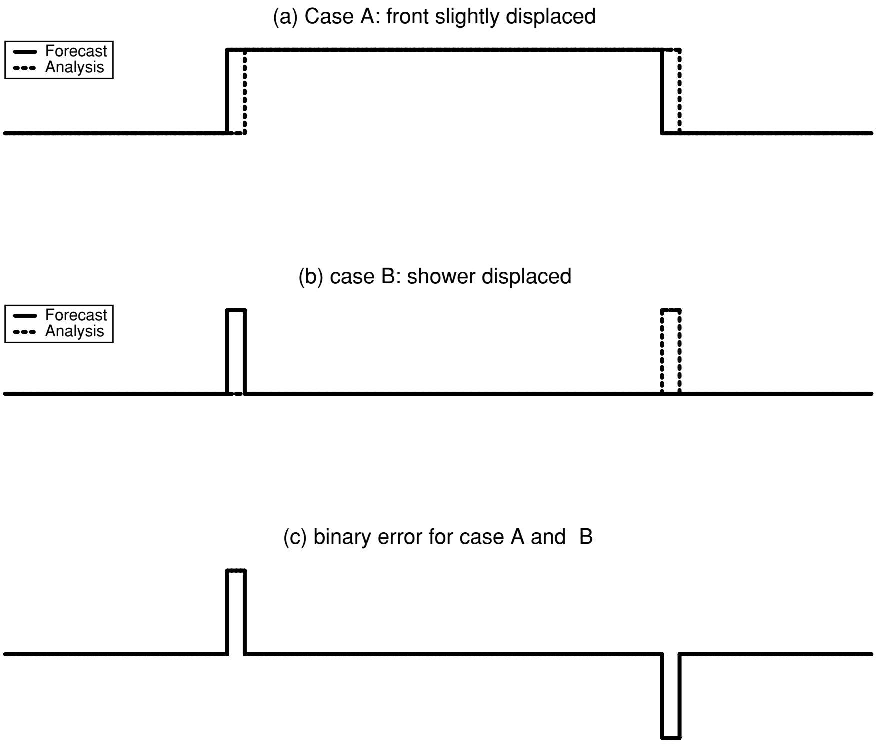

Figure 7. *One-dimensional idealised cases of (a) a front slightly displaced, (b) a shower displaced and (c) their binary error. The two different synoptic situations generate the same error and for both the cases MSE = 0.04. However the front case has a much larger base rate than the shower case ($\varepsilon_{\mathrm{front}} = 0.5 > \varepsilon_{\mathrm{shower}} = 0.02$). Therefore the front case has a much larger skill score than the shower case ($SS_{\mathrm{front}} = 0.92 > SS_{\mathrm{shower}} = -0.02$).*

spatial scales. The decrease of the *MSE* with threshold is due to the lower base rate for more intense precipitation events, and not due to improved skill.

To take account of the base rate effect, Figure 9 shows the binary MSE skill score components SS_l as functions of threshold u and spatial scale l for the six case studies. For all cases, the skill score components for errors with small spatial scale ($l < 40$ km) are negative, i.e. the forecasts are worse than random. Small spatial scale errors decrease the overall forecast skill. The large spatial scales ($l \geq 40$ km) are positive and lead to positive overall forecast skill. The separation of forecast skill at 40 km spatial scale corresponds to a separation between mesoscale and convective precipitation events. Convective precipitation features, such as rain cells, have typical spatial scales smaller than 40 km. Forecasts of such small-scale events are often dominated by displacement errors (e.g. case 2 and 5) which penalise the forecast skill. On the other hand, mesoscale features, such as frontal bands, have typical scales larger than 40 km (Bluestein 1993). Mesoscale features are generally well captured by the forecasts (e.g. case 1) and, although small-scale displacements still occur, these events contribute to the positive skill of the forecast.

The skill score in Figure 9 is less dependent on threshold for large spatial scales. However, for small scales the negative skill becomes more negative for higher thresholds. Small spatial scale errors in more intense (rarer) events give the worst skill.

Case 1 is an example of a well-detected frontal system, but some showers in Wales were missed. The intensity-scale verification technique for this case exhibits positive skill at large scales and negative skill at small scales. Case 2 shows an example of a shower system that was well forecast on large scales, but the individual showers were mis-handled leading to negative skill on small spatial scales. Case 3 is characterised by an intense large spatial scale storm displaced by almost its entire length. The displacement error of the incorrectly advected storm is well detected by the intensity-scale technique as can be seen by the isolated negative skill score minimum on the 160 km spatial scale for thresholds between 1/2 and 4 mm/h. Case 4 shows the same storm of case 3 three hours later. The storm is better advected and the forecast exhibits better skill on large spatial scales, but on small spatial scales the skill is still negative. Case 5 is an example of a few shower systems. As for case 2, the systems on the large scale were detected, but the showers within the systems were partially displaced. The forecast is dominated by small-scale displacement errors and so the skill on small spatial scales is negative. Case 6 shows an example of timing error of a frontal system. The horizontal misplacement of the front is well detected as seen by the negative skill on the 40 and 80 km spatial scales for thresholds between 1/2 and 4 mm/h.

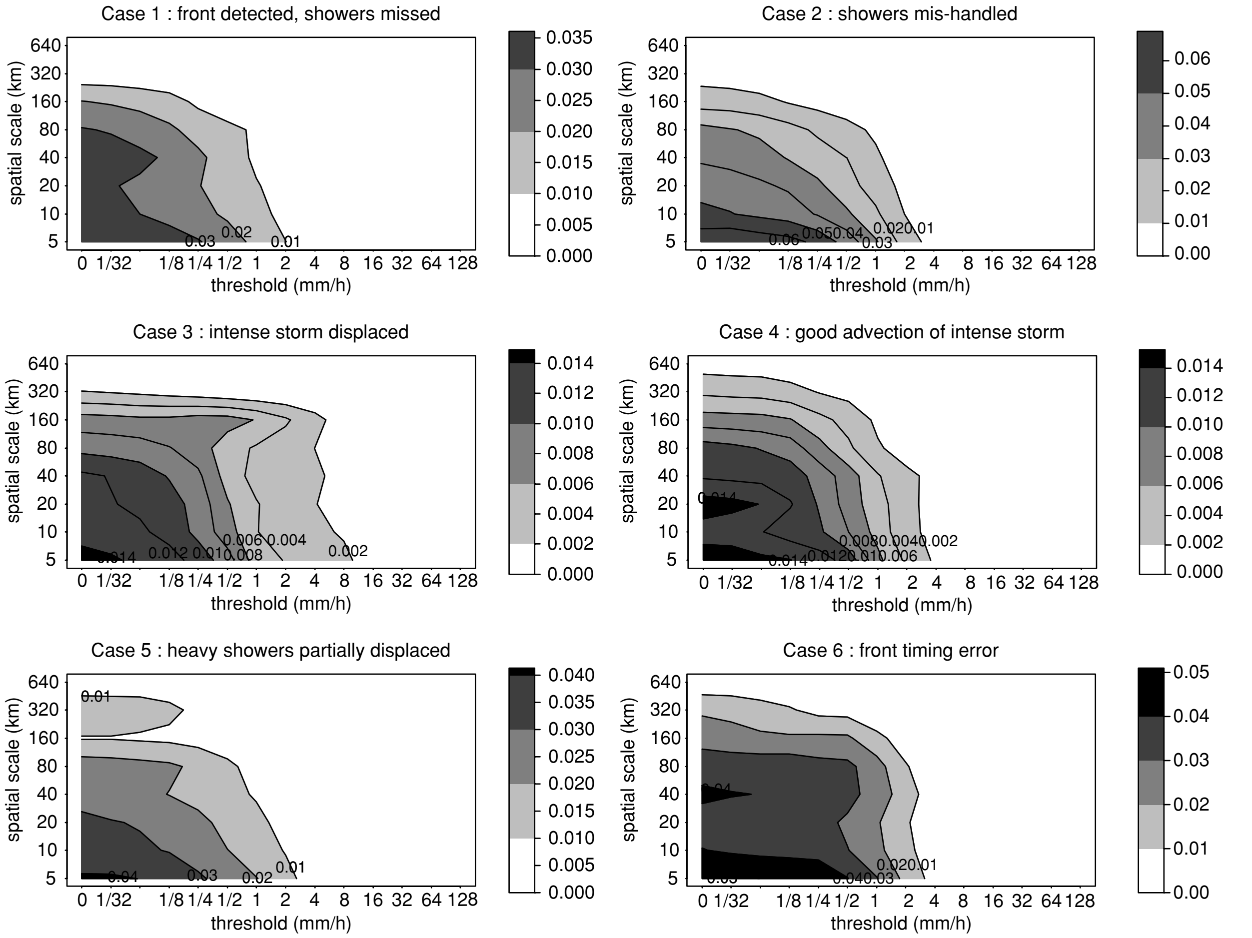

Figure 8. *Two-dimensional plots of MSE_l as function of threshold and spatial scale for the six case studies. For all the case studies, the binary MSE decreases for large spatial scales. Most of the MSE is due to errors on small scales. The binary MSE decreases as the threshold increases. This effect is due to the decrease of precipitation pixels in the binary images as the threshold is increased (base rate dependency).*

5. Conclusions

This study has developed an intensity-scale approach for the verification of spatial precipitation forecasts. The technique allows the skill to be diagnosed as a function of the scale of the forecast error and intensity of the precipitation events. Results show that reduction of skill in NIMROD is mainly due to the small-scale (< 40 km) misplacement of more intense (rarer) precipitation events. It is these features that need to be improved to improve the overall skill of the forecast. The technique provides useful insight on individual forecast cases. For example, case 3 is characterised by an intense large-scale storm displaced by almost its entire length. The displacement error of the incorrectly advected storm is clearly detected by the intensity-scale verification technique as an isolated negative skill score minimum on the 160 km spatial scale for thresholds between 1/2 and 4 mm/h.

The intensity-scale verification technique developed in this study has been applied here to now-casting precipitation forecasts, but it could easily be applied to other kinds of forecasts, such as seasonal precipitation forecast or sea surface temperature.

In this study we have applied the intensity-scale verification technique to recalibrated unbiased forecasts, but the technique can also be applied to biased forecasts by

1. Adding an extra spatial scale to the spatial scales already considered. The extra spatial scale is the largest Haar father wavelet component with spatial scale resolution equal to the whole spatial domain (e.g. 1280 km). It is the mean bias obtained by averaging over all the domain of the binary error image.
2. Substituting in Eqns (9) and (10) the MSE of the binary *unbiased* images generated by a random process $MSE_{\mathrm{random}} = 2\varepsilon(1-\varepsilon)$ with the MSE of the binary *biased* images generated by a random process $MSE_{\mathrm{random}} = B\varepsilon(1-\varepsilon) + \varepsilon(1-B\varepsilon)$, where B denotes the bias.

For biased forecasts, the binary MSE skill score is still equal to the Heidke skill score but it is no longer equal to the Peirce skill score.

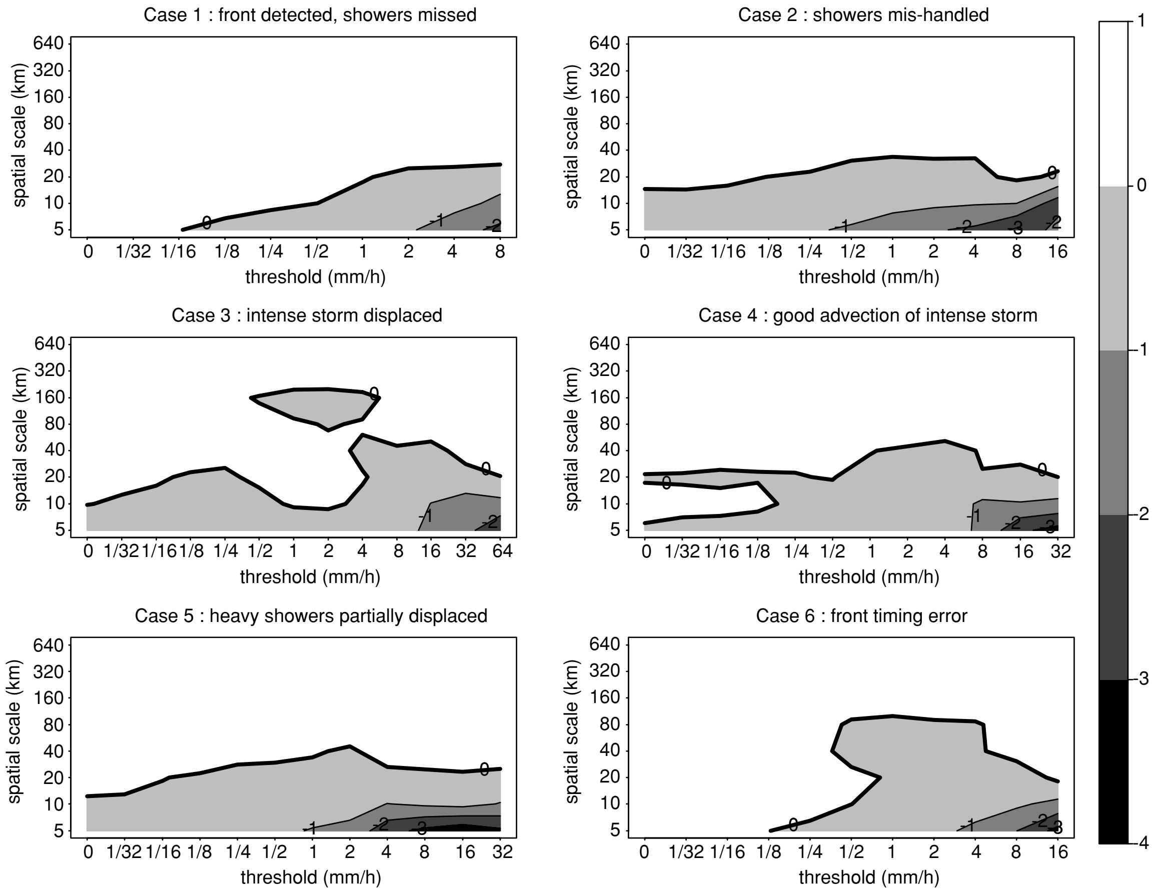

Figure 9. *Two-dimensional plots of SS_l (Eqn. 10) as functions of threshold and spatial scale for the six case studies. For all cases, the skill score components for error with small spatial scales (smaller than 40 km) are negative. Small spatial scale errors decrease the overall forecast skill. The large spatial scales are positive and increase the overall forecast skill. For small scales the negative skill becomes more negative for higher thresholds. In case study 3 the displacement error of the poorly advected storm is well detected by the isolated negative skill score minimum on the 160 km spatial scale for thresholds between 1/2 and 4 mm/h. In case study 6 the front timing error is well detected as seen by the negative skill on the 40 and 80 km spatial scale for thresholds between 1/2 and 4 mm/h.*

The intensity-scale verification technique bridges traditional categorical binary verification with the more recent techniques which evaluate the forecast skill on different spatial scales (e.g. Zepeda-Arce & Foufoula-Georgiou 2000; Briggs & Levine 1997). The intensity-scale verification technique extends the traditional categorical binary verification performed using the HSS and enables a verification on different spatial scales, as well as for different thresholds. The intensity-scale verification technique provides information about skill for different precipitation intensities, whereas Zepeda-Arce & Foufoula-Georgiou (2000) and Briggs & Levine (1997) assess only skill for different spatial scales. Moreover, the categorical approach provides a more robust probability-based verification approach than Briggs & Levine (1997), who use the MSE, which can be heavily affected by the discontinuities, noise and outliers characterising the precipitation field. Zepeda-Arce & Foufoula-Georgiou (2000) evaluate the different spatial scales by spatial averaging (Haar father wavelets), which artificially smoothes the forecast and observation fields with loss of detailed small-scale features. On the other hand, both the intensity-scale verification technique and Briggs & Levine (1997) make use of spatial variations (mother wavelets) and therefore retain small-scale features. The orthogonality of the mother wavelet components on different scales enables the verification score to be decomposed into the *sum* of scores on different spatial scales. The intensity-scale verification technique assesses the forecast on its whole domain and it is appropriate for UK weather situations characterised by the presence of many scattered precipitation events. Feature-based approaches (e.g. Ebert & McBride 2000; Brown et al. 2002; Baldwin et al. 2001) are more suitable for arid Australian or US weather scenarios, characterised by sparse isolated precipitation features.

Further work could be done to extend the intensity-scale verification technique. For example, we have evaluated analytically the expected value of the MSE of binary images generated by an homogeneous Poisson process for each spatial scale component obtained by the Haar wavelet filter. This was then used to evaluate

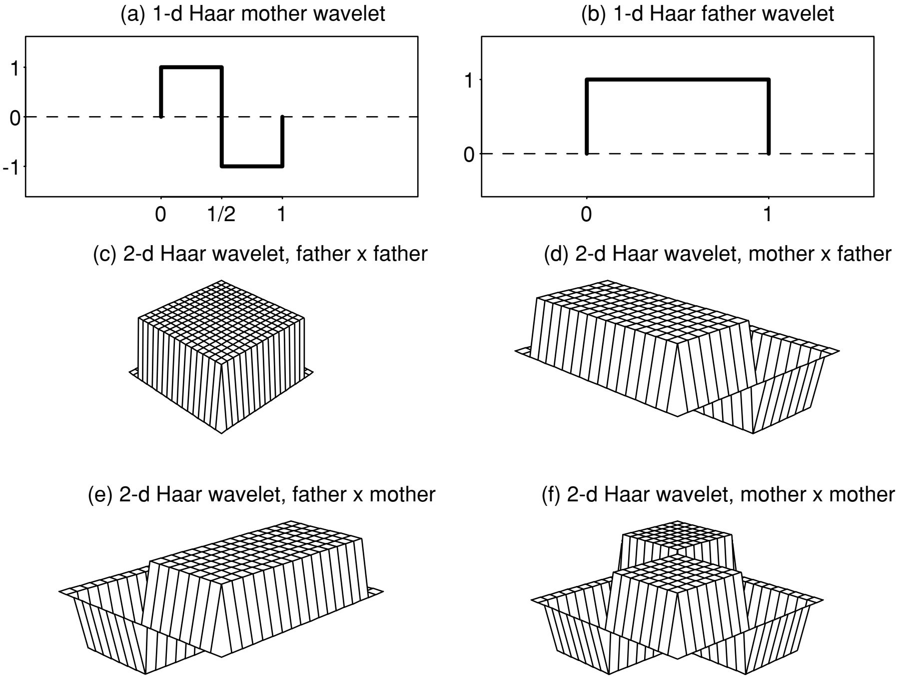

Figure 10. *One- and two-dimensional Haar wavelets. All the wavelets of the Haar family are generated from the one-dimensional mother (a) and father (b) wavelets by performing a deformation and a translation. The two-dimensional Haar wavelets are generated from the orthogonal product of one-dimensional Haar wavelets. The father wavelet components of the binary error image on the different spatial scales are obtained as a linear combination of the two-dimensional Haar wavelet shown in the panel (c). The mother wavelet components of the binary error image on the different spatial scales are obtained as a linear combination of the two-dimensional Haar wavelet shown in the panels (d), (e) and (f).*

the binary MSE skill score components for each spatial scale l. However, the resulting skill score produced misleading results, due to the neglect of spatial clustering that exists in precipitation fields. Future work could focus more attention on the spatial clustering that is responsible for some of the skill in the forecasts.

Acknowledgements

The authors wish to thank Dr Brian Golding and the UK Met Office, who helped fund this project. We also thank Will Hand for having carefully selected the NIMROD case studies, and Martin Göeber for interesting discussions about QPF verification. Barbara Casati wishes to thank Dr Barbara Brown and her colleagues in the RAP division, NCAR (Boulder, USA) for the very enjoyable visit and for many useful discussions. She also wishes to thank E. Ebert and L. Wilson for helpful and encouraging discussions on verification. In addition, we wish to thank Dr Chris Ferro and Dr Sergio Pezzulli for their feedback and help with statistics and Splus.

Appendix: Two-dimensional discrete Haar wavelet filter

Wavelets are real functions characterised by a location and a spatial scale (Daubechies 1992). Similar to Fourier analysis, wavelets can be used to express real functions as a sum of components on different spatial scales. However, because of their locality, wavelets are more suitable than Fourier transforms to deal with spatially discontinuous fields, such as precipitation.

A wavelet transform can be performed by using different types of wavelets, characterised by different shapes and mathematical properties. In this work wavelets of the Haar family (Figure 10) are used, because of their square shape which best captures the difference in binary variables. All the wavelets of the Haar family are generated from the mother and father wavelets by performing a deformation (which characterises the scale) and a translation. Two-dimensional Haar wavelets are generated from the cartesian product of one-dimensional Haar wavelets (Figure 10).

Wavelet transforms can be either continuous or discrete. Discrete wavelet transforms generate components which are orthogonal (Mallat 1989). A two-dimensional *discrete* wavelet transform can be used to decompose the binary error image (see Figure 5) into the *sum* of *orthogonal* components on different spatial scales (see Figure 6). The binary error image components on different spatial scales are obtained as a linear combination of two-dimensional Haar wavelets (Figure 10).

The two-dimensional discrete Haar wavelet filter can be explained by a simple algorithm based on spatial averaging over $2^l \times 2^l$ pixels domains. To help visualise the algorithm, Figure 11 shows an example of the

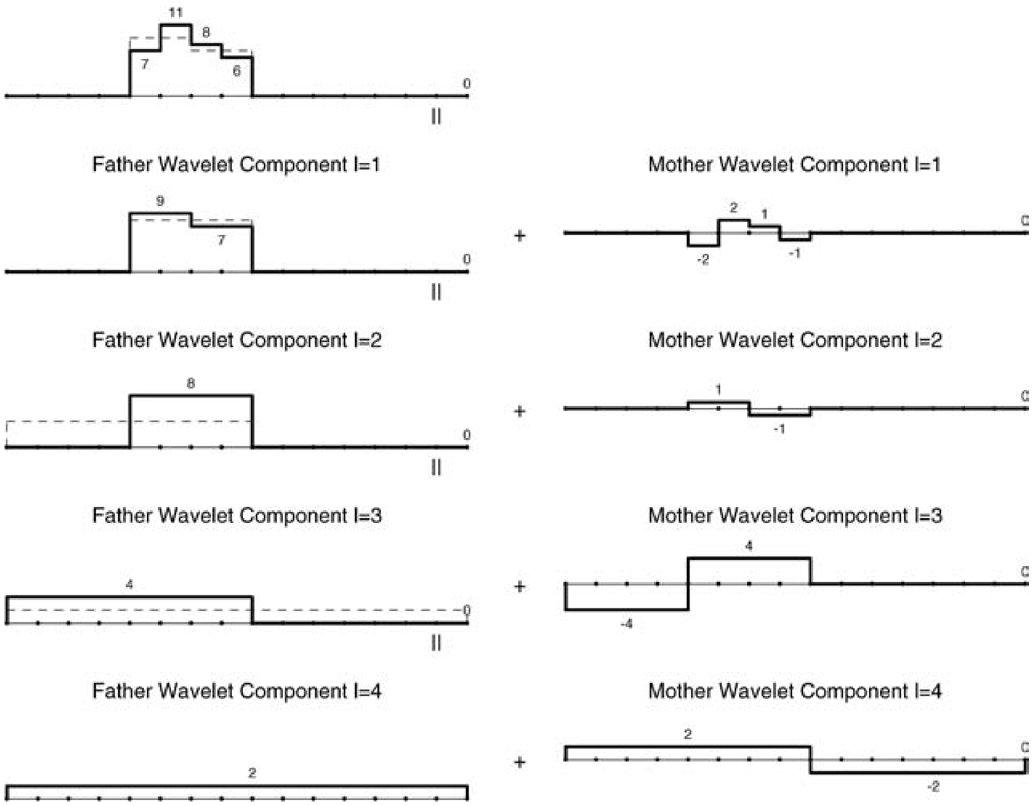

Figure 11. *Example of the one-dimensional discrete Haar wavelet filter applied to an example function (top left panel). At the first step the function is decomposed into the sum of a coarser* mean *function (the first father wavelet component) and a* variation-about-the-mean *function (the first mother wavelet component). At each step the Haar wavelet filter decomposes the father wavelet component obtained from the previous step into the sum of a coarser* mean *function (the l^{th} father wavelet component) and a* variation-about-the-mean *function (the l^{th} mother wavelet component). The l^{th} father wavelet component is obtained from the initial function by a spatial averaging over 2^l pixels. The process stops when the largest father wavelet component (mean over the whole domain) is found.*

one-dimensional discrete Haar wavelet filter applied to a function. The binary error image Z is defined on a 5 km resolution spatial domain of $2^L \times 2^L$ pixels, where $L = 8$. The Haar wavelet filter at its first step decomposes the binary error image into the sum of a coarser *mean* field (the *father wavelet component*) and a *variation-about-the-mean* image (the *mother wavelet component*). The father wavelet component is obtained from the binary error image by a spatial averaging over 2×2 pixels (10 km resolution). The mother wavelet component is obtained as the difference between the binary error image and the father wavelet component (5 km resolution). The Haar wavelet filter then decomposes the father wavelet component obtained from the first step into the sum of a coarser *mean* field (the second father wavelet component) and a *variation-about-the-mean* image (the second mother wavelet component). The second father wavelet component is obtained from the binary error image by a spatial averaging over 4×4 pixels (20 km resolution). The second mother wavelet component is obtained as the difference between the second father wavelet component and the first father wavelet component (10 km resolution).

The process is iterated and at each step the Haar wavelet filter decomposes the father wavelet component obtained from the $(l-1)^{th}$ step into the sum of a coarser *mean* field (the l^{th} father wavelet component) and a *variation-about-the-mean* image (the l^{th} mother wavelet component). The l^{th} father wavelet component is obtained from the binary error image by a spatial averaging over $2^l \times 2^l$ pixels (5×2^l km resolution). The l^{th} mother wavelet component is obtained as the difference between l^{th} and $(l-1)^{th}$ father wavelet components ($5 \times 2^{l-1}$ km resolution).

The process stops when the largest father wavelet component ($L = 8$) is found. The binary error image is decomposed into the sum of the mother wavelet components on the spatial scales $l = 1, \ldots, L$ (corresponding to 5, 10, 20, 40, 80, 160, 320, 640 km resolution) and the L^{th} father wavelet component

$$Z = \sum_{l=1}^{L} W^l_{\text{mother}}(Z) + W^L_{\text{father}}(Z). \qquad (14)$$

Note that the L^{th} father wavelet component is equal to the spatial mean of the binary error image over the whole spatial domain, i.e. the mean bias $\overline{Z}$. Since the recalibrated forecast is unbiased, the L^{th} father wavelet component is zero and so

$$Z = \sum_{l=1}^{L} W^{l}_{\text{mother}}(Z) = Z_l. \tag{15}$$

References

Airey, M. & Hulme, M. (1995) Evaluating climate model simulations of precipitation: methods, problems and performances. *Progress in Physical Geography* **19**: 427–448.

Baldwin, M. E., Lakshmivarahan, S. & Kain, J. S. (2001) Verification of mesoscale features in NWP models. In *Preprints, 9th Conference on Mesoscale Processes*, American Meteorological Society (Boston), 30 July–2 August 2001, pp. 255–258.

Bluestein, H. B. (1993) *Synoptic-Dynamic Meteorology in Midlatitudes*. Oxford: Oxford University Press.

Bougeault, P. (2003) The WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. Technical report, WMO.

Brewster, K. A. (2003) Phase-correcting data assimilation and application to storm-scale numerical weather prediction. *Mon. Wea. Rev.* **131**: 480–507.

Briggs, W. M. & Levine, R. A. (1997) Wavelets and field forecast verification. *Mon. Wea. Rev.* **125**: 1329–1341.

Brown, B. G., Davis, C. A. & Bullock, R. (2002) Verification methods for objects and fields: Standard methods, issues and enhanced approaches. Technical report, NCAR, Boulder, Colorado, USA, February.

Browning, K. A., Collier, C. G., Larke, P. R., Menmuir, P., Monk, G. A. & Owens, R. G. (1992) On the forecasting of frontal rain using a weather radar network. *Mon. Wea. Rev.* **110**: 534–552.

Collier, C. G. & Krzysztofowicz, R. (2001) Quantitative Precipitation Forecasting. *J. Hydrology* **239**: 1–2000 (Special Issue).

Daubechies, I. (1992) *Ten Lectures on Wavelets*. SIAM.

Doswell, C. A., Davies-Jones, R. & Keller, D. L. (1990) On summary measures of skill in rare event forecasting based on contingency tables. *Wea. and Forecasting* **5**: 576–585.

Du, J., Mullen, S. L. & Sanders, F. (2000) Removal of distortion error from an ensemble forecast. *J. Appl. Meteorol.* **35**: 1177–1188.

Ebert, E. E. (2001) Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.* **129**: 2461–2480.

Ebert, E. E. & McBride, J. L. (2000) Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology* **239**: 179–202.

Ebert, E. E., Damrath, U., Wergen, W. & Baldwin, M. E. (2003) The WGNE assessment of short-term Quantitative Precipitation Forecasts. *Bull. Am. Meteorol. Soc.* **84**: 481–492.

FOAG (1993) FRONTIERS evaluation of radar-based rainfall and river flow forecasts April 1992 to March 1993. Technical report, Met Office, National River Authority. FRONTIERS Operational Assessment Group (FOAG).

Göeber, M., Wilson, C. A., Milton, S. F. & Stephenson, D. B. (2003) Fairplay in the verification of operational quantitive precipitation forecasts. *J. Hydrology* (in press).

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

Golding, B. W. (2000) Quantitative Precipitation Forecasting in the UK. *J. Hydrology* **239**: 286–305.

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

Hoffman, R. N. & Grassotti, C. (1996) A technique for assimilating SSM/I observations of marine atmospheric storms: tests with ECMWF analyses. *J. Appl. Meteorol.* **35**: 1177–1188.

Hoffman, R. N., Liu, Z., Louis, J.-F. & Grassotti, C. (1995) Distortion representation of forecast errors. *Mon. Wea. Rev.* **123**: 2758–2770.

Johnson, L. E. & Olsen, B. G. (1998) Assesment of quantitative precipitation forecasts. *Wea. and Forecasting* **13**: 75–83.

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

Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **II**(7): 674–693.

Marzban, C. (1998) Scalar measures of performance in rare-event situations. *Wea. and Forecasting* **13**: 753–763.

Murphy, A. H. & Winkler, R. L. (1987) A probability model for forecast verification. *Mon. Wea. Rev.* **115**: 1330–1338.

Nehrkorn, T., Hoffman, R. N., Grassotti, C. & Louis, J. F. (2003) Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Q. J. R. Meteorol. Soc.* **129**: 195–218.

Osborn, T. J. & Hulme, M. (1998) Evaluation of the European daily precipitation characteristics from the Atmosphere Model Intercomparison Project. *Int. J. Climatol.* **18**: 505–522.

Schaefer, J. T. (1990) The Critical Success Index as an indicator of warning skill. *Wea. and Forecasting* **5**: 570–575.

Stephenson, D. B. (2000) Use of the odds ratio for diagnosing forecast skill. *Wea. and Forecasting* **15**: 221–232.

Woodcock, F. (1976) The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.* **104**: 1209–1214.

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