Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events

NIGEL M. ROBERTS AND HUMPHREY W. LEAN

Joint Centre for Mesoscale Meteorology, Met Office, Reading, United Kingdom

(Manuscript received 19 December 2006, in final form 23 March 2007)

ABSTRACT

The development of NWP models with grid spacing down to ~1 km should produce more realistic forecasts of convective storms. However, greater realism does not necessarily mean more accurate precipitation forecasts. The rapid growth of errors on small scales in conjunction with preexisting errors on larger scales may limit the usefulness of such models. The purpose of this paper is to examine whether improved model resolution alone is able to produce more skillful precipitation forecasts on useful scales, and how the skill varies with spatial scale. A verification method will be described in which skill is determined from a comparison of rainfall forecasts with radar using fractional coverage over different sized areas. The Met Office Unified Model was run with grid spacings of 12, 4, and 1 km for 10 days in which convection occurred during the summers of 2003 and 2004. All forecasts were run from 12-km initial states for a clean comparison. The results show that the 1-km model was the most skillful over all but the smallest scales (approximately <10–15 km). A measure of acceptable skill was defined; this was attained by the 1-km model at scales around 40–70 km, some 10–20 km less than that of the 12-km model. The biggest improvement occurred for heavier, more localized rain, despite it being more difficult to predict. The 4-km model did not improve much on the 12-km model because of the difficulties of representing convection at that resolution, which was accentuated by the spinup from 12-km fields.

1. Introduction

The resolution of operational numerical weather prediction (NWP) models is continually being increased with the expectation that this will lead to improved predictions of local weather, especially precipitation. Limited area models (LAMs) with a grid spacing of less than 5 km are now common and ~1 km will be considered typical within a decade. Much of the convection that was once parameterized in coarser-resolution models will become explicitly resolved. It is hoped that this transition toward explicit representation of convection, along with the other higher-resolution benefits (e.g., more detailed orography), will result in more accurate precipitation forecasts. Several studies have already shown that models with a grid spacing of 1–4 km in which convection is explicitly resolved are capable of producing more realistic simulations of larger convective entities, such as severe thunderstorms, mesoscale convective systems, and squall lines (Weismann et al. 1997; Romero et al. 2001; Speer and Leslie 2002; Done et al. 2004).

However, greater realism does not necessarily mean more accurate forecasts. Done et al. (2004) found that, although the systems were better represented, point-specific forecasts were not necessarily improved. Mass et al. (2002) state that "decreasing grid spacing in mesoscale models to less than 10–15 km generally improves the realism of the results but does not necessarily significantly improve the objectively scored accuracy of the forecasts." The problem we may have to face is an inherent reduction in predictability at the new resolved scales as the grid spacing is reduced and convection is resolved. Lorenz (1969) argued that the ability to resolve smaller scales would result in forecast errors growing more rapidly. Zhang et al. (2003) performed an experiment in which grid-scale noise was added to a 3.3-km model simulation and found that errors initially grew rapidly at small scales in regions where convection was present, which eventually infected larger scales throughout the domain. Walser et al. (2004) also found

Corresponding author address: Nigel M. Roberts, Joint Centre for Mesoscale Meteorology, Met Office, Meteorology Building, University of Reading, P.O. Box 243, Reading, Berkshire RG6 6BB, United Kingdom.
E-mail: nigel.roberts@metoffice.gov.uk

DOI: 10.1175/2007MWR2123.1

that in convective situations forecast uncertainty increased with decreasing scale and was significant at scales up to 100 km. Findings such as these suggest that we should be very careful about the interpretation of precipitation forecasts from "storm resolving" NWP models. It is important to avoid taking literally deterministic information on scales that are expected to be unpredictable for the forecast lead time, and for that reason a probabilistic approach is more desirable for both the presentation and verification of output on those scales.

Probabilities are usually obtained from an ensemble of forecasts (Richardson 2000; Mylne 2002), but for the next few years within an operational context, this would be prohibitively expensive and scientifically difficult if convection is to be represented explicitly. An alternative is to apply suitable postprocessing to a deterministic forecast. Theis et al. (2005) describe a "nearest neighbors" method in which the probability of rain at each grid square was obtained by examining the values of the nearby surrounding grid squares (in space and time). This idea has also been used to produce rainfall products from the Nimrod nowcasting system (Golding 1998) at the Met Office and in postprocessing the Rapid Update Cycle (RUC) model (S. S. Weygandt and N. L. Benjamin 2005, personal communication). Theis et al. showed that their derived probabilities were more skillful than the raw deterministic output. However, they also noted that the verification results were sensitive to the size of the neighborhood, and that the optimal size is unknown. The highest scores were obtained from the largest neighborhood they tried. The difficulty is that the use of increasingly larger neighborhoods will eventually result in so much smoothing that the purpose of having a high-resolution forecast is lost. Techniques are required that can evaluate scales at which forecasts become sufficiently skillful and identify the scales over which increased resolution is beneficial (if any). Traditional grid-point-by-grid-point verification methods are inappropriate when the small scales are unpredictable because the structure on those scales can be regarded as noise, and this increases the measured error; yet, a forecast with little skill on small scales may still be useful over a larger area (e.g., a river catchment).

New methods for verifying quantitative precipitation forecasts (QPFs) have been developed in recent years. One approach is to classify features as objects and investigate how predicted objects differ from those observed. Techniques have been described by Ebert and McBride (2000), Done et al. (2004), and Davis et al. (2006). The advantage of the object-based approach is that, in addition to giving a measure of forecast skill, it can provide insight into the ability of a model to represent particular features; this is information that is invaluable for identifying shortcomings in NWP models. The drawback is that forecasts have to be sufficiently skillful in the first place to allow for a clear association of objects. An alternative approach is to evaluate forecast skill over different spatial scales. Briggs and Levine (1997) used a wavelet decomposition of the forecast and observed fields to obtain a multiscale verification of 500-mb geopotential height fields. Precipitation forecasts were evaluated on different spatial scales by Zepeda-Arce et al. (2000), who used the threat score and depth–area–duration curves after averaging over different sized areas to obtain the spatial scales. Casati et al. (2004) presented a technique in which they applied Haar wavelet decomposition to separate forecast errors into different spatial scales, and used the mean square error (MSE) to obtain a display of forecast error as a function of precipitation intensity and spatial scale. Bousquet et al. (2006) have also used Haar wavelet decomposition to identify the scales at which a 10-km model fails to represent the spatial variability of rainfall. Another approach has been demonstrated by Marzban and Sangathe (2006, 2008) who used a cluster analysis method to verify precipitation fields. The novel aspect of this approach is that an object-based methodology has been used to provide the means for examining forecast error on different scales.

A new scale-selective method for evaluating precipitation forecasts will be introduced here that allows us to determine the scales at which forecasts become skillful. It uses the concept of nearest neighbors as the means of selecting the scales of interest, and, like the method developed by Casati et al. (2004), it is applied to thresholds. The result is a measure of forecast skill against spatial scale for each selected threshold. A valuable spin-off from the process is that there is a direct relationship between the results obtained from the verification and the nearest-neighbors approach for probabilistic postprocessing of rainfall output.

The aims of the paper are twofold: the first is to present the verification method, and the second is to report on the impact of resolution on short-range forecast skill over scales of interest when simulations of convective events were run at 12, 4, and 1 km. In section 2 we explain the verification method, in section 3 we describe the model setup, in section 4 we present the results from the model verification, and in section 5 we discuss the implications of the results. Finally, in section 6 we draw conclusions.

2. The verification method

The purpose of this verification method is to obtain a measure of how forecast skill varies with spatial scale in

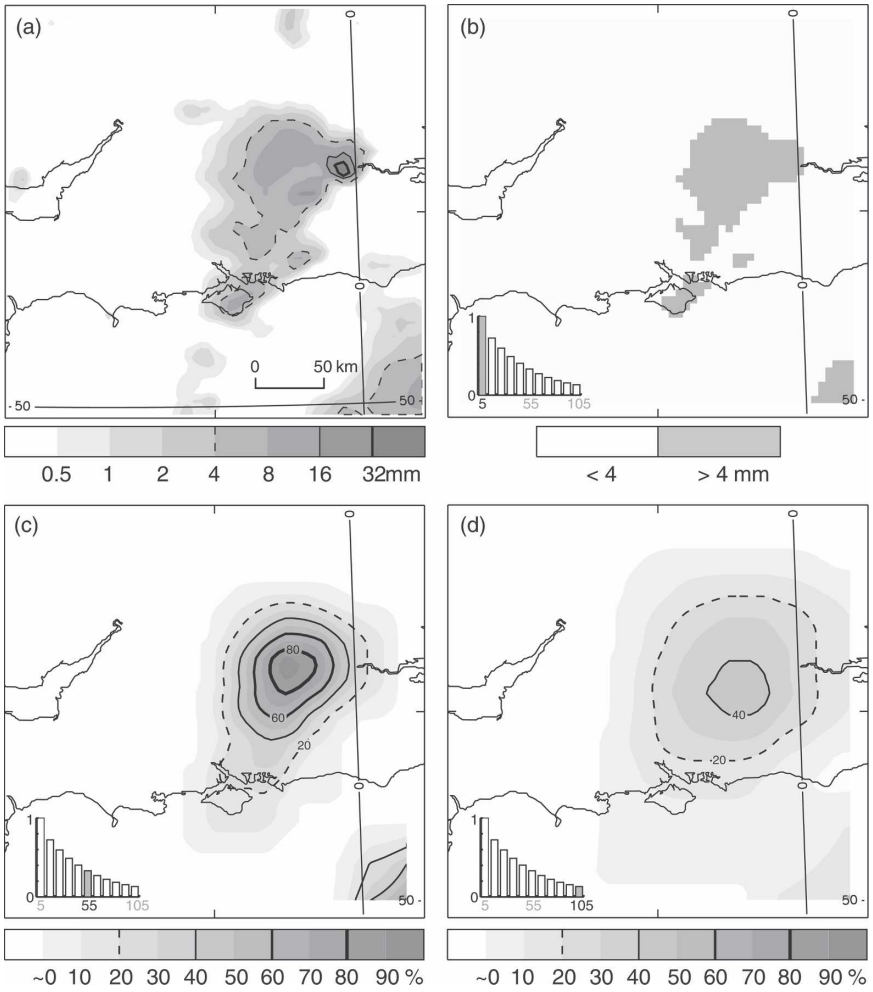

FIG. 1. (a) Nimrod composite radar rainfall accumulation from 1600 to 2200 UTC 27 Apr 2004, (b) binary image of accumulations exceeding 4 mm, and fractions computed from the 4-mm binary image using neighborhood lengths of (c) 55 and (d) 105 km. The bar charts in (b)–(d) plot the variance of the fraction fields against neighborhood size, normalized against the binary field (b), to show how sharpness is reduced with neighborhood length (km). The shaded bars match the pictures.

a way that can be intuitively understood by users and is also directly applicable for postprocessing. Radar data processed from the Met Office Nimrod system (Golding 1998; Harrison et al. 2000) are used for comparison with precipitation forecasts because of their spatial coverage. Radar error will be discussed later. The verification method will now be described.

a. Convert to binary fields

All of the model and radar data (either precipitation accumulation or rates) are projected on to the same verification grid. Suitable thresholds (q) are chosen (e.g., $q = 0.5$, 1, 2, and 4 mm) and used to convert the radar-observed (O_r) and forecast-model (M_r) rainfall fields into binary fields I_o and I_M. All grid squares exceeding the threshold have a value of 1 and all others a value of 0,

$$I_o = \begin{cases} 1 & O_r \geq q \\ 0 & O_r < q \end{cases} \quad \text{and} \quad I_M = \begin{cases} 1 & M_r \geq q \\ 0 & M_r < q \end{cases}. \qquad (1)$$

An example of this conversion into a binary field for an accumulation threshold of 4 mm is shown in Figs. 1a,b. Percentile rather than accumulation thresholds are also used for conversion into a binary field. For example, the 95th percentile threshold selects the highest 5% of observed and forecast accumulations (over all grid squares) for comparison. The purpose of doing this is to remove the impact of any bias in rainfall amounts when we wish to focus on the spatial accuracy of the forecasts.

b. Generate fractions

The process for generating the fractions is essentially the same as the nearest-neighbors method used by Theis et al. (2005) to obtain probabilities. For *every* grid point in the binary fields obtained from Eq. (1) we compute the fraction of surrounding points within a given square of length n that have a value of 1 (i.e., have exceeded the threshold). This is described by Eqs. (2) and (3) below, in which $O_{(n)}(i, j)$ is the resultant field of observed fractions for a square of length n obtained from the binary field I_o and $M_{(n)}(i, j)$ is the resultant field of model forecast fractions obtained from the binary field I_M. These quantities assess the spatial density in the binary fields,

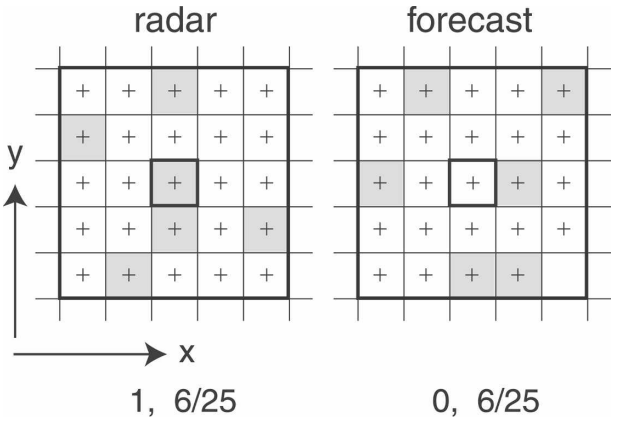

FIG. 2. Schematic example of radar and forecast fractions (see text).

$$O(n)(i,j) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} I_O\left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2}\right], \tag{2}$$

$$M(n)(i,j) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} I_M\left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2}\right]. \tag{3}$$

Here i goes from 1 to N_x, where N_x is the number of columns in the domain and j goes from 1 to N_y, where N_y is the number of rows. Fractions are generated for different spatial scales by changing the value of n, which can be any odd value up to $2N - 1$, where N is the number of points along the longest side of the domain. A square of length $2N - 1$ is the smallest that can encompass all points in the domain for squares centered at any point in the domain. Figure 2 provides a visual interpretation. The pictures are schematic representations of radar and forecast binary fields on the same grid. At the central grid square the binary forecast is wrong; the radar field has a value of 1, and the forecast field has a value of 0. However, when fractions are computed over the 5×5 ($n = 5$) neighborhood, the radar and forecast fractions are both 6/25 (six shaded grid squares in each). In this example the forecast is deemed correct over the area of that $n = 5$ neighborhood.

Figures 1c,d show fractions generated for a real case. As the size of the neighborhood is increased, the sharpness (see Potts 2003) is reduced (bar charts in Fig. 1), and the fractions obtained from the larger neighborhood in Fig. 1d give a smoother picture than in that in Fig. 1c. Over the largest neighborhood required, length $= 2N - 1$, the same fraction would be obtained at every grid square; then, there is no sharpness. Points outside the domain are assigned a value of zero.

By using squares, we have applied the convolution kernel for a mean filter to the binary field, which is something used often in image processing. Equation (2) can be rewritten as

$$O(n)(i,j) = \sum_{k=1}^{n}\sum_{l=1}^{n} I_O\left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2}\right]K(n)(k,l), \tag{4}$$

where $K(n)(k, l)$ is the $n \times n$ kernel for a (square) mean filter. It might be preferable to use a different kernel, such as a circular mean filter or a Gaussian kernel. However, we do not believe it would alter the key results enough to warrant the additional complexity, because an important priority is to keep the method as simple as possible.

c. Compute fractions skill scores

The MSE for the observed and forecast fractions from a neighborhood of length n is given by

$$\mathrm{MSE}_{(n)} = \frac{1}{N_x N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}[O_{(n)i,j} - M_{(n)i,j}]^2. \tag{5}$$

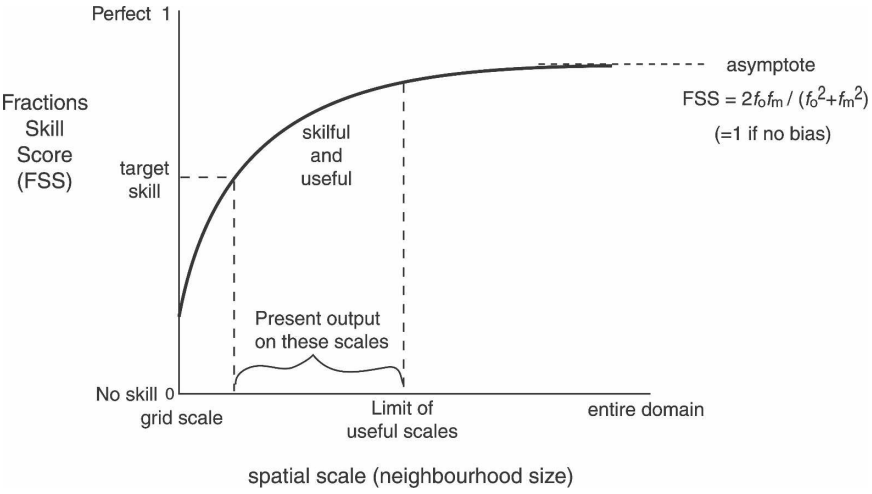

FIG. 3. Schematic graph of skill against spatial scale (see text).

The MSE is not in itself very useful because it is highly dependent on the frequency of the event itself. A MSE skill score has been computed relative to a low-skill reference forecast (Murphy and Epstein 1989). This is defined as the fractions skill score (FSS),

$$\mathrm{FSS}_{(n)} = \frac{\mathrm{MSE}_{(n)} - \mathrm{MSE}_{(n)\mathrm{ref}}}{\mathrm{MSE}_{(n)\mathrm{perfect}} - \mathrm{MSE}_{(n)\mathrm{ref}}} = 1 - \frac{\mathrm{MSE}_{(n)}}{\mathrm{MSE}_{(n)\mathrm{ref}}}, \tag{6}$$

where $\mathrm{MSE}_{(n)\mathrm{perfect}} = 0$ is the MSE of a perfect forecast for neighborhood length n. The reference used ($\mathrm{MSE}_{\mathrm{ref}}$) for each neighborhood length (n) is given by,

$$\mathrm{MSE}_{(n)\mathrm{ref}} = \frac{1}{N_x N_y}\left[\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} O^2_{(n)i,j} + \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} M^2_{(n)i,j}\right]. \tag{7}$$

It can be thought of as the largest possible MSE that can be obtained from the forecast and observed fractions. The relationship between the FSS and the more conventional reference forecasts (e.g., random forecast) is introduced in section 2d.

Figure 3 shows the way the FSS typically varies with neighborhood length n, given a sufficiently large sample. It has a range from 0 to 1. A forecast with perfect skill has a score of 1; a score of 0 means zero skill. Skill is lowest at the grid scale, that is, when the neighborhood is only one grid point and the fractions are binary ones or zeros. As the size of the neighborhood is increased, skill increases until it reaches an asymptote at $n = 2N - 1$. If there is no bias (an equal number of observed and forecast pixels exceeding the threshold) the asymptotic fractions skill score (AFSS) (FSS at $n = 2N - 1$) has a value of 1, indicating perfect skill over the whole domain. If there is a bias, then the observed frequency f_o (fraction of observed points exceeding the threshold over the domain) is not equal to the model-forecast frequency f_M, and from Eqs. (5), (6), and (7) it can be shown that

$$\mathrm{AFSS} = 1 - \frac{(f_o - f_M)^2}{f_o^2 + f_M^2} = \frac{2 f_o f_M}{f_o^2 + f_M^2}. \tag{8}$$

This descriptor of the bias is useful because it relates the bias to the spatial accuracy of a forecast and is linked to the conventional frequency bias (f_o/f_M), with the advantage of being less sensitive to biases from small frequencies (AFSS = 0.8 is a factor of 2, AFSS = 0.5 is a factor of 4, and AFSS = 0.2 is a factor of 10 frequency bias).

The practical benefit of plotting the FSS against spatial scale is also demonstrated by Fig. 3. Skill increases with spatial scale until there comes a point at which some desired level of skill has been reached (see, e.g., $\mathrm{scale}_{\mathrm{min}}$ below). This is the smallest scale at which output from the forecast system should be presented, although this scale should always exceed five grid lengths (Lean and Clark 2003; Skamarock 2004; Bousquet et al. 2006). At larger scales, skill increases further, but the information content of the forecasts is limited by the additional smoothing. The largest scale over which output should be presented becomes a compromise between user requirements, cost effectiveness, and forecast skill; it may be considered a waste of resources to run a 1-km model for forecasting on scales as large as

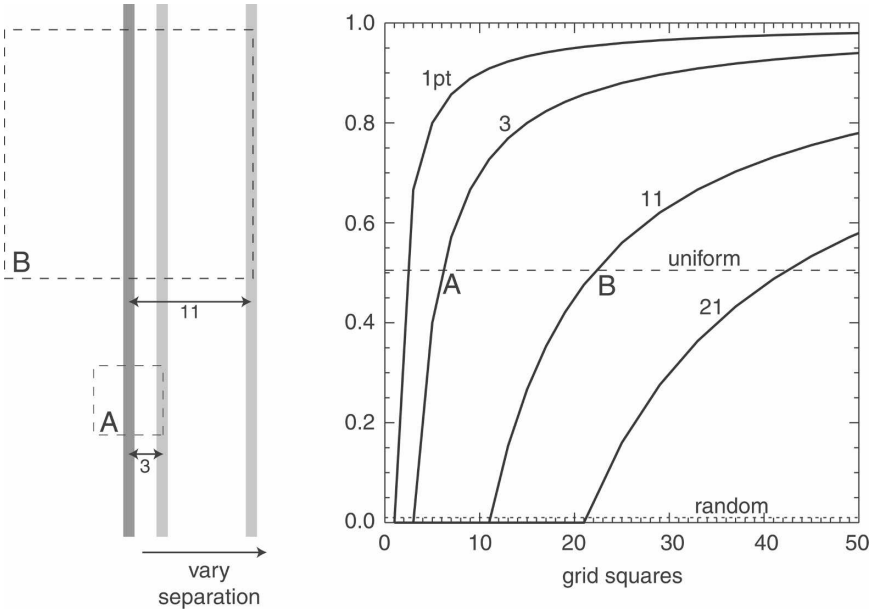

FIG. 4. Idealized situation in which forecasts of a band of rain 1 pixel wide are misplaced by varying distances. The dark gray band represents the observed rainfall. The light gray bands represent forecast bands shifted by 3 and 11 pixels. The curves on the right show the variation of FSS with neighborhood length for forecast bands misplaced by 1, 3, 11, and 21 pixels. The horizontal dashed lines are $FSS_{uniform}$ and FSS_{random}; A and B show where the 3 and 11 pixel separation curves cross $FSS_{uniform}$. The dashed squares show the neighborhood sizes for A and B.

100 km, but it may be perhaps useful for flood prediction on a scale of 50 km.

d. An idealized example

Figure 4a shows an idealized situation in which a band of rain, one grid square wide, is predicted with different displacement errors. The structure, alignment, and coverage of the "observed" and "forecast" rainbands are identical; only the distance between them varies. FSS is plotted against neighborhood size in Fig. 4b for forecast displacements of 1, 3, 11, and 21 grid squares. When the length of the sampling square is less than or equal to the displacement error, there is no skill and the FSS = 0. For spatial scales (sampling squares) longer than the displacement error the FSS increases with spatial scale, eventually reaching a value of 1 (because there is no bias in this experiment). The smaller the forecast error, the more rapidly the skill increases with scale.

The curves intercept two horizontal dashed lines. The first (labeled "random") denotes the FSS that would be obtained from a random forecast with the same fractional coverage over the domain as that of the rainband (i.e., equal to the base rate, f_o). It is given by $FSS_{random} = f_o$. In this example the rainband covered 1% of the domain, so a random forecast has $FSS_{random} = 0.01$. The other dashed line (labeled "uniform") represents the FSS that would be obtained *at the grid scale* ($n = 1$) from a forecast with a fraction/probability equal to f_o at every point. It is given by $FSS_{uniform} = 0.5 + f_o/2$ (i.e., halfway between random forecast skill and perfect skill); so, in this example $FSS_{uniform} = 0.505$. Whereas the random forecast has low skill unless the base rate is large, the uniform forecast is always reasonably skillful, but has zero sharpness. The FSS curve reaches $FSS_{uniform}$ at a scale termed $scale_{min}$. If the domain is large, $f_o \to 0$ and $FSS_{uniform} \to 0.5$; then, for a displacement distance D, $scale_{min} = 2D$. This is shown visually by the sampling squares A and B in Fig. 4. $scale_{min}$ represents the smallest scale over which the forecast output contains useful information (unless specific user requirements dictate otherwise). It is also possible to compute categorical scores for this idealized situation. When FSS = $FSS_{uniform}$ (at $scale_{min}$) the hit rate becomes 0.5 and the critical success index (CSI) is 0.33. Here $FSS_{uniform}$ is considered to be a suitable value for the "target skill" in Fig. 3.

3. The NWP model and experimental setup

The Met Office's Unified Model (UM) solves nonhydrostatic, deep-atmosphere dynamics using a semi-

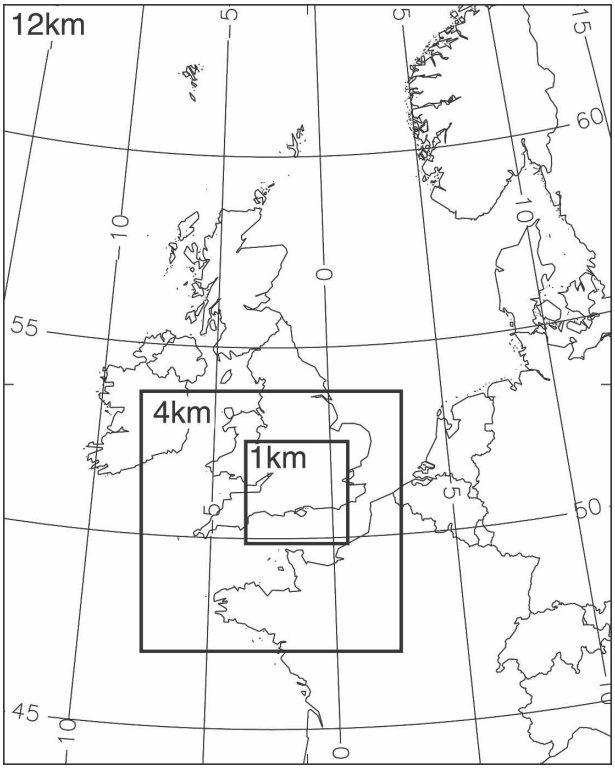

FIG. 5. The domains used for the 12-, 4-, and 1-km models.

implicit, semi-Lagrangian numerical scheme (Cullen et al. 1997; Davies et al. 2005). The model includes a comprehensive set of parameterizations, including surface exchange (Essery et al. 2001), boundary layer (Lock et al. 2000), mixed phase cloud microphysics (Wilson and Ballard 1999), and convection (Gregory and Rowntree 1990). The model runs on a rotated latitude–longitude horizontal grid with Arakawa C staggering and a terrain-following hybrid-height vertical coordinate with Charney–Philips staggering. Soil moisture fields are generated offline from observations using the UM surface exchange scheme (Smith et al. 2006).

For the trials reported here, the 4-km model was one-way nested inside the 12-km model and the 1-km model was one-way-nested inside the 4-km model (Fig. 5). The differences in configuration between the models are summarized in Table 1. It was considered reasonable to run the model without a convection parameterization scheme at 1 km. It is less clear whether this is sensible at 4 km. Convection schemes are not designed for this resolution, but without one it is not possible to represent smaller showers. The approach adopted here was to use the Gregory and Rowntree scheme, but limit its activity by restricting the cloud-base mass flux (Roberts 2003). Sensitivity experiments have shown that this approach usually gives more realistic simulations than either including the unrestricted scheme or switching it off altogether, but the underlying problem remains.

The operational 12-km model is run without any horizontal diffusion but does apply diffusion locally where there are high vertical velocities in order to suppress gridpoint storms. At 4 and 1 km, diffusion is ∇^4 in the horizontal with a fixed diffusivity designed to damp two grid-length waves with an *e*-folding time of eight time steps; this is derived on the basis of the maximum shear that should be permitted to occur. Results have been compared with different options (∇^2 and different diffusivities), and the current choice produces the most acceptable power spectra in winds. The 12-km model assumes that rain falls straight out of the model without being advected by the winds. It is likely to be a poor approximation on horizontal scales of less than 10 km, and for this reason prognostic rain was included in the 4- and 1-km model runs.

To obtain aggregated statistics, 10 days with convective activity were chosen from the summers of 2003 and 2004 (Table 2). For each convective day, four forecasts at each resolution were run at 3-h intervals. The 12-km forecasts were run for 7 h and used three-dimensional variational data assimilation (3DVAR) (Lorenc et al. 2000) for most observation types, in addition to the Moisture Observation Preprocessing System (MOPS) (Macpherson et al. 1996) and latent heat nudging (LHN) (Jones and Macpherson 1997). The 4- and 1-km

TABLE 1. Model configurations.

	12 km	4 km	1 km
Time step	5 min	100 s	30 s
Vertical levels	38	38	76
Domain	146 × 182 points	190 × 190 points	300 × 300 points
Horizontal diffusion	None	Eight time steps ∇^4	Eight time steps ∇^4
Prognostic rain	No	Yes	Yes
Convection scheme	Standard mass flux	Restricted mass flux	None
Assimilation	3DVAR, MOPS, and LHN	Initialize from 12 km $T + 1$ h	Initialize from 12 km $T + 1$ h

TABLE 2. Summary of cases.

Case no.	Date	Model runs	Description
1	13 May 2003	6, 9, 12, 15	Organized thunderstorms
2	25 May 2003	6, 9, 12, 15	Scattered showers
3	1 Jul 2003	6, 9, 12, 15	Organized showers
4	28 Aug 2003	6, 9, 12, 15	Bands of convective rain
5	27 Apr 2004	9, 12, 15, 18	Localized thunderstorms
6	8 Jul 2004	3, 6, 9, 12	Bands of convective rain
7	10 Jul 2004	3, 6, 9, 12	Scattered showers
8	20 Jul 2004	6, 9, 12, 15	Scattered showers
9	3 Aug 2004	6, 9, 12, 15	Organized thunderstorms
10	20 Aug 2004	3, 6, 9, 12	Organized showers

models were initialized from the 12-km model at $T + 1$ and run for 6 h. The 4- and 1-km models were initialized with interpolated 12-km data in order to keep the comparison clean. It is understood that this reduces the expected benefits of running the model at finer resolution.

4. Verification results

First, results from the forecasts of two individual cases will be shown to allow a comparison of FSS curves with a visual interpretation of forecast skill. Then, aggregated results will be shown to provide some insights into model skill and behavior. The forecast and radar data have been projected onto a 5-km verification grid over an area covering most of the 1-km domain (excluding the 10 km around the edge).

a. Visual comparison for two cases

Often, it is difficult to make a reliable subjective evaluation of forecast skill because rainfall patterns are complex. However, on the occasion when a visual assessment was more clear cut, the FSS curves were found to be in good agreement with the perceived skill. Two examples are shown to demonstrate this. The first is

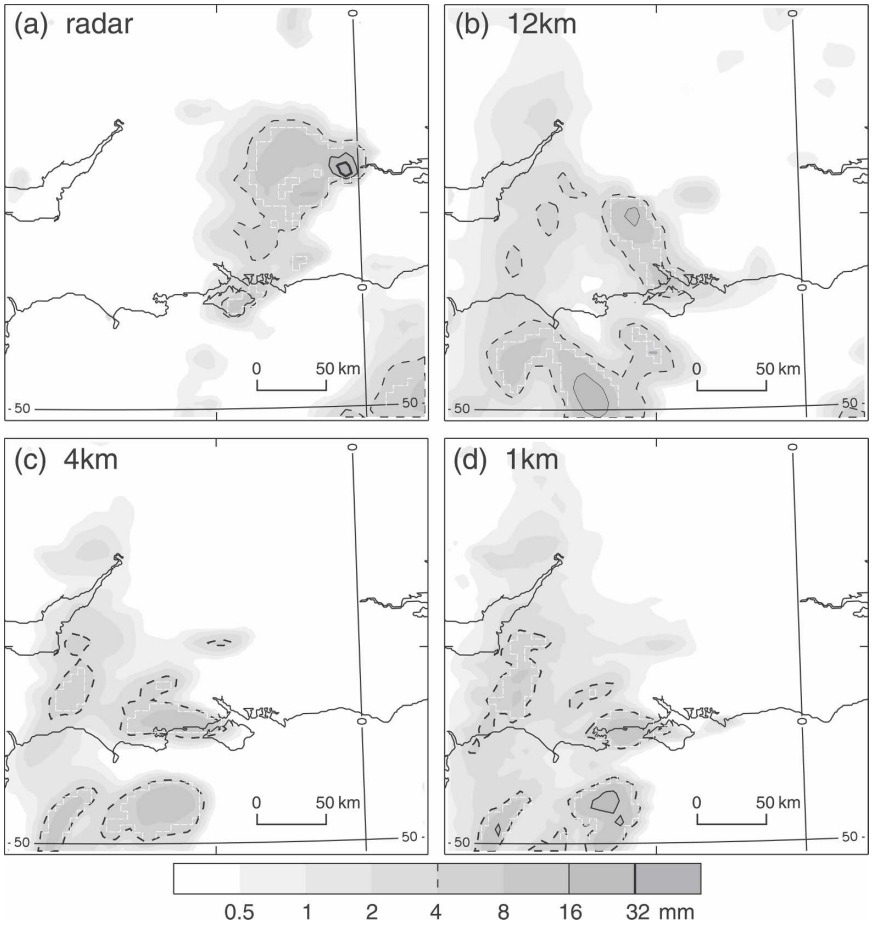

FIG. 6. Rainfall accumulations over the period 1600–2200 UTC 27 Apr 2004 (case 5) from (a) radar, (b) the 12-km model forecast from 1500 UTC, (c) the 4-km model forecast from 1600 UTC, and (d) the 1-km model forecast from 1600 UTC. White dashed lines enclose the top 5% of accumulations (>95th percentile).

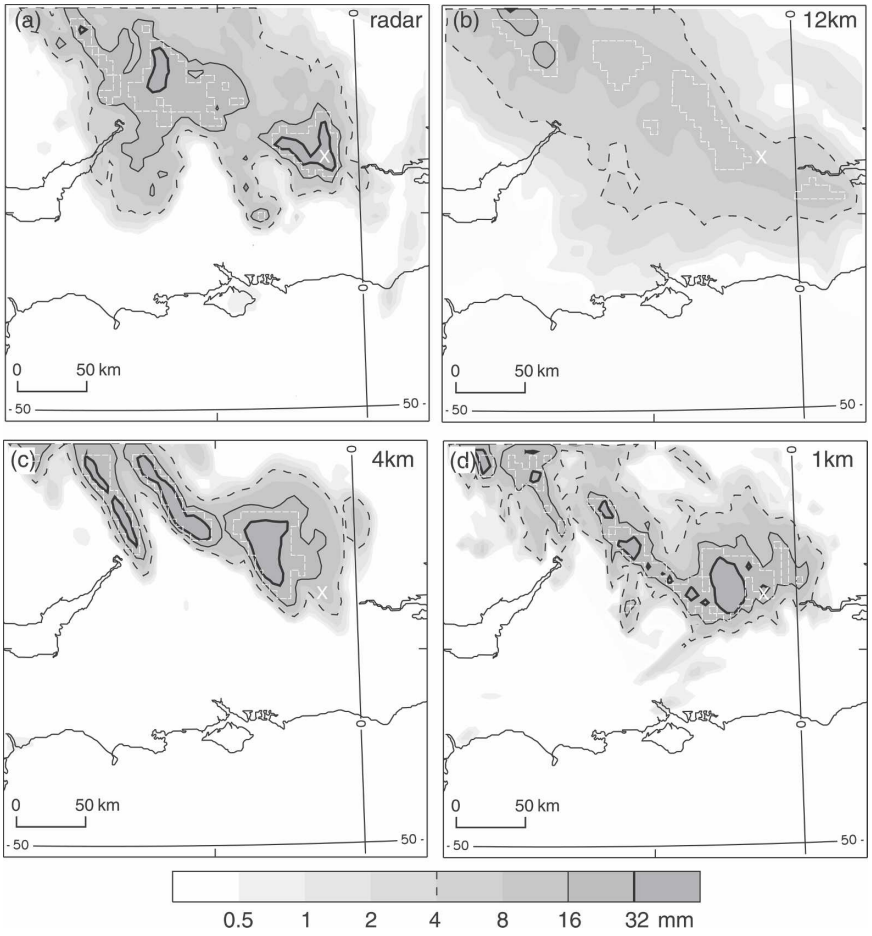

FIG. 7. Rainfall accumulations over the period 1300–1900 UTC 3 Aug 2004 (case 9) from (a) radar, (b) the 12-km model forecast from 1200 UTC, (c) the 4-km model forecast from 1300 UTC, and (d) the 1-km model forecast from 1300 UTC. White dashed lines enclose the top 5% of accumulations (>95th percentile). The cross marks the location of the highest observed accumulations.

from case 5, in which all of the forecasts predicted rain in the wrong place (Fig. 6), and includes the worst individual forecast in the trial. The second is from case 9, in which all of the forecasts compared well with the radar (Fig. 7). Skill curves for each of these forecasts are shown in Fig. 8. Here we have used a 95th percentile threshold (see section 2a) rather than an accumulation threshold to focus on the spatial accuracy of the forecasts.

Starting with case 5 (Fig. 6), it is clear that none of the forecasts were able to predict the area of rain in the correct place. Figure 8a supports this view by showing that at scales up to 40 km, the 12-km forecast was less skillful than a random forecast, and the 4- and 1-km forecasts were worse than random at scales up to ~85 km. The 4- and 1-km forecasts look very similar, and this is reflected in their almost identical FSS curves. Both forecasts achieved $FSS_{uniform}$ at a scale of around 180 km. This is consistent with an observed misplacement of the rain area. The visual impression given by the 12-km forecast is that it was somewhat more skillful than the others because it predicted some rain farther east. The FSS curve is in agreement with this perception of improved skill. It is more skillful at all scales, with $FSS_{uniform}$ achieved at a scale of 160 km.

Turning to case 9 (Fig. 7), all of the models predicted the broad distribution of the rainfall very well, although the 12-km forecast was unable to reproduce the highest totals. Figure 8b shows that the 12- and 4-km models had similar skill over all scales, but at the smaller scales (<60 km) the 1-km forecast was the most skillful. This reflects the visual impression that the 1-km forecast was locally more accurate in the London, United Kingdom, area, where much of the heaviest rain fell. The 1-km FSS exceeded $FSS_{uniform}$ at a scale of ~15 km, compared with ~40 km for the 12- and 4-km forecasts. Note

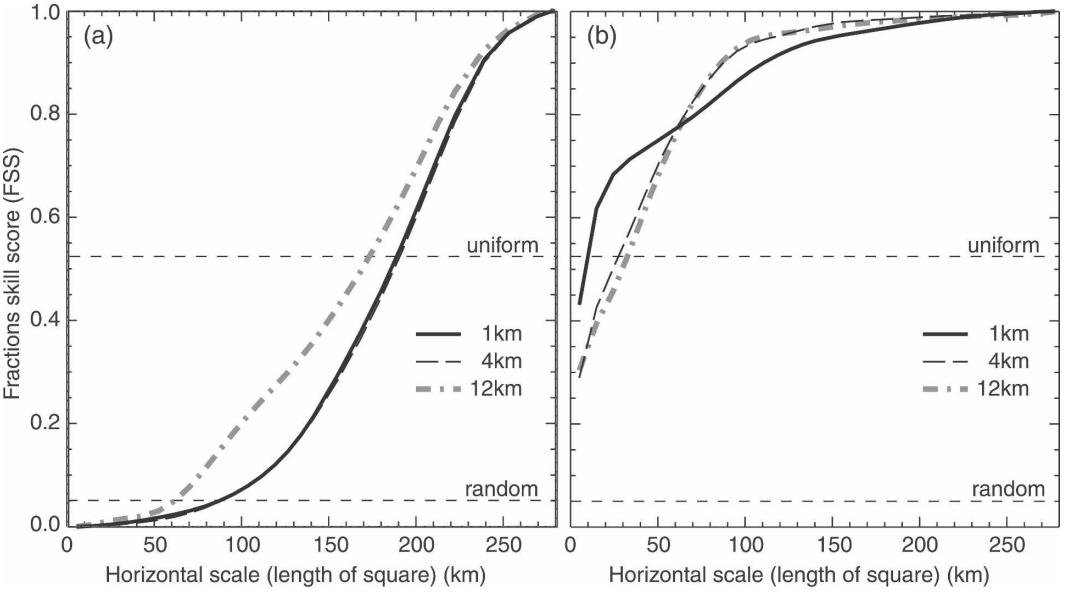

FIG. 8. Graphs of FSS against neighborhood length for 12-, 4-, and 1-km model forecasts using a 95th percentile threshold for (a) 27 Apr 2004 (case 5) and (b) 3 Aug 2004 (case 9).

that the use of a percentile threshold has removed the impact of the underprediction in the 12-km forecast. The use of a high accumulation threshold (e.g., 16 mm) would lead to a low FSS for the 12-km model at all scales. Both examples have shown that the verification results are in agreement with the visual interpretation for both the gross spatial errors and those that are more subtle on smaller scales.

b. Aggregated results from 40 forecasts

Results will be presented both for aggregations of 4-h accumulations from the last 4 h of each forecast (hours 3–7; to avoid the initial spinup period of the 4- and 1-km forecasts), and for aggregations of hourly accumulations from the entire 6 h of each forecast.

1) ERROR BARS

The error bars in Figs. 9 and 10 (and Fig. 14, below) represent the effect of uncertainty in radar from (among several contributors) (i) spurious rainfall resulting from anomalous propagation (anaprop) or ground clutter, (ii) inadequacies in the conversion from reflectivity to rain rate, and (iii) rain drift below the radar beam. Most of the anaprop and ground clutter is automatically removed during the Nimrod quality control process (Harrison et al. 2000), but some is missed. Other quality control procedures within the Nimrod system are also described in Harrison et al. (2000). To understand how these errors may affect the results, three types of modification were made to the radar fields: 1) A random increment of up to a factor of 2 was added to every nonzero grid-square accumulation to represent errors in rain rate, 2) a small area of high accumulations (5×5 grid squares) was added to represent anaprop, and 3) every grid square was shifted two grid squares to account for spatial misplacement. The error bars were obtained by measuring the largest FSS deviations at each scale that resulted from the modifications to the radar fields.

2) AGGREGATED ACCUMULATIONS OVER THE FINAL 4 H (HOURS 3–7)

FSS curves for accumulation thresholds of 0.2, 1.0, 4.0, and 16.0 mm are displayed in Fig. 9. For the low thresholds of 0.2 and 1.0 mm, both the 1- and 12-km models were significantly more skillful over all scales than the 4-km model. For a 1.0-mm threshold, $FSS_{uniform}$ was reached at scales of 17 (12-km forecasts), 36 (4-km forecasts), and 14 (1-km forecasts) km. The 4-km model was poorer because it tended to underpredict light rain (discussed later). For the higher thresholds of 4.0 and 16.0 mm, the 4- and 1-km models have comparable skill, but the 12-km model is considerably worse, especially for the 16.0-mm threshold, for which $FSS_{uniform}$ is not reached at any scale. The FSS of only 0.05 at scales > 100 km is the result of the bias resulting from an underprediction of locally heavy rain. In circumstances such as this where the model grossly underpredicts, it is very difficult to extract meaningful information on the spatial accuracy of the model. To focus

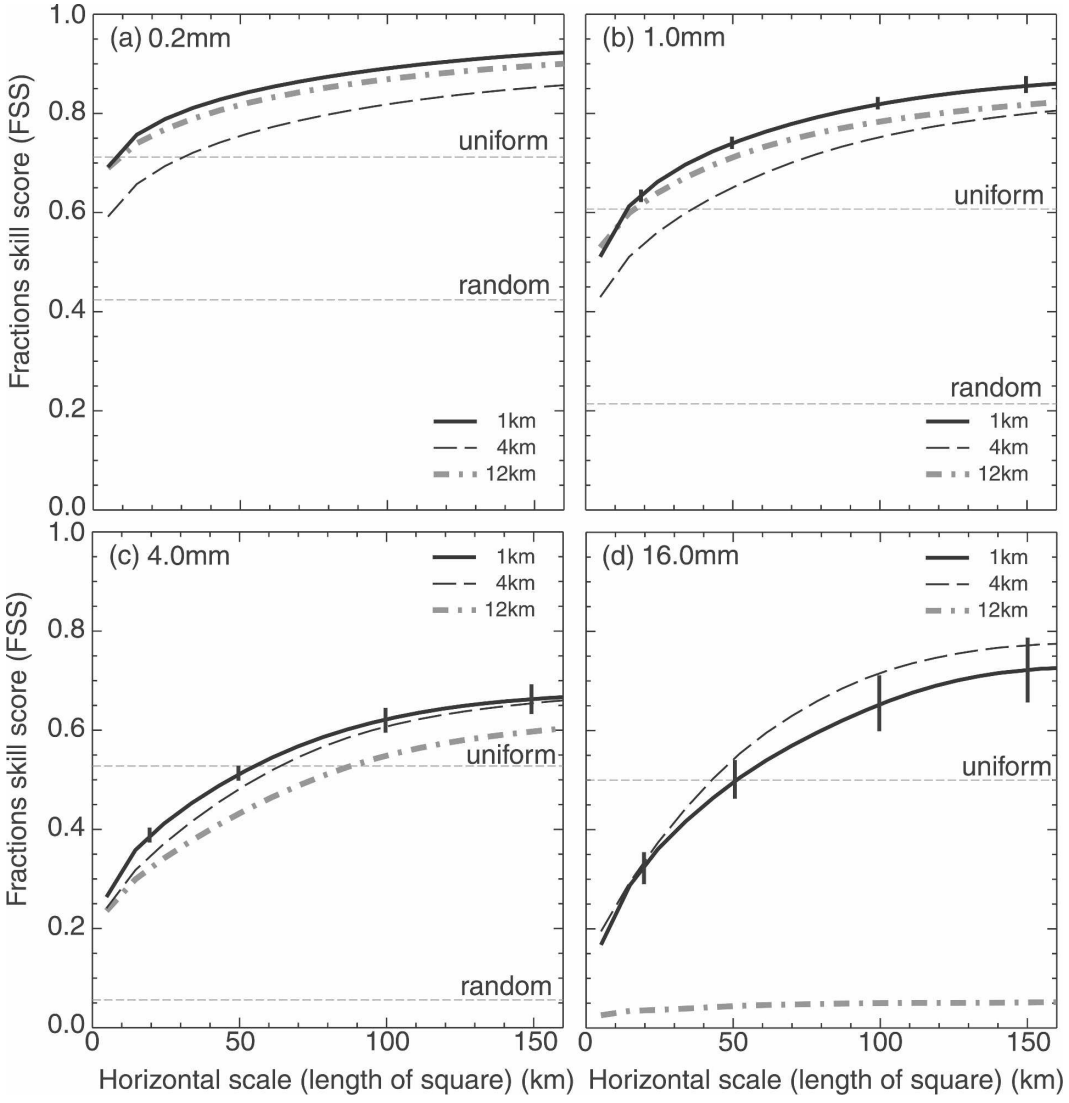

FIG. 9. Graphs of aggregated FSS against neighborhood length for rainfall accumulations over the last 4 h of the 12-, 4-, and 1-km model forecasts using accumulation thresholds of (a) 0.2, (b) 1.0, (c) 4.0, and (d) 16.0 mm. The error bars in each panel apply to all curves.

on the spatial accuracy, we have used percentile thresholds.

FSS curves using percentile thresholds are shown in Fig. 10. The 75th percentile locates widespread rainfall accumulations that occupy a quarter of the domain. Increasingly higher percentiles sample less extensive rain areas. The 99th percentile threshold picks out localized features in the rainfall pattern that occupy only 1% of the domain. Two patterns are evident. First, the more localized rainfall features are more difficult to predict accurately for any given spatial scale or model resolution. Second, the 1-km model is the most skillful over all of the thresholds, and the gain is greatest for more localized rainfall. For the 95% (75%) threshold, $FSS_{uniform}$ is reached at a scale of 65 km (27 km) with the 1-km model compared with 84 km (33 km) with the 12-km model. The greater improvement in skill at 1 km for the more localized rainfall is not surprising, because it is more likely to respond to improvements in the representation of orography and local dynamics, as well as the transition to explicit convection; whereas the distribution of more widespread rainfall (an envelope of convective activity) is dependant on the larger-scale mesoscale forcing, which should vary less between resolutions, especially when the initial conditions are identical. In contrast to the trend in FSS seen in Fig. 9 (using accumulation thresholds), the 4- and 12-km models have similar skill over all scales for all of the percentile

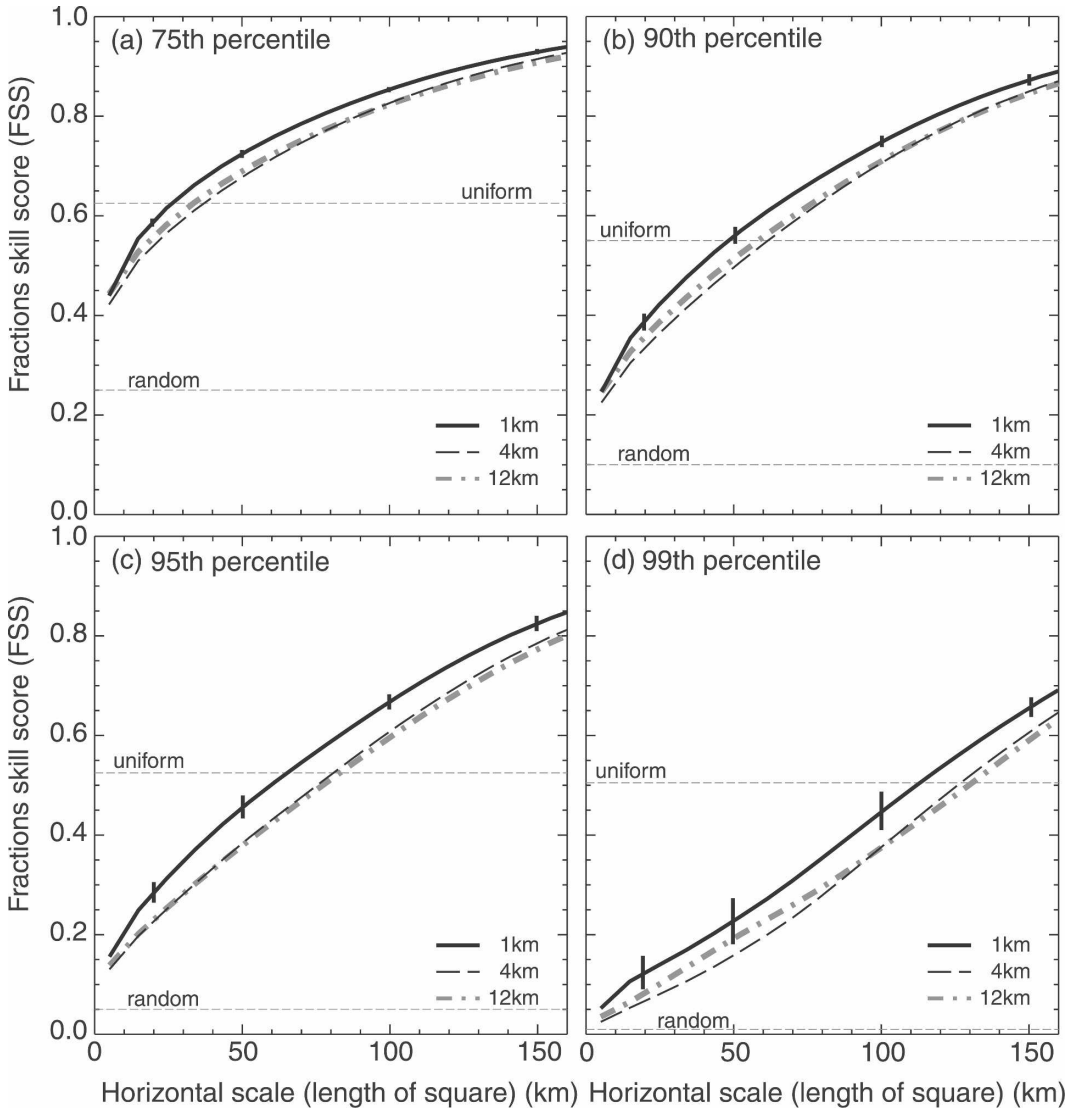

FIG. 10. Graphs of aggregated FSS against neighborhood length for rainfall accumulations over the last 4 h of the 12-, 4-, and 1-km model forecasts using percentile thresholds of (a) 75%, (b) 90%, (c) 95%, and (d) 99%. The error bars in each panel apply to all curves.

thresholds, indicating that the differences found in Fig. 9 were due to differences in the biases.

The scale at which $FSS_{uniform}$ is reached ($scale_{min}$) is plotted as a function of percentile thresholds to examine the variation of skill with both resolution and threshold (Fig. 11). All three model resolutions become less skillful as the areas of rain being sampled become more localized. The graph also confirms that, however widespread or localized the rain, the 1-km model reaches $FSS_{uniform}$ at smaller scales than the 12- and 4-km models, and that the improvement is greater when more localized rainfall is sampled (e.g., an improvement of 7 km for the 70th percentile compared to 18 km for the 95th percentile).

3) BIAS

It is instructive to investigate how the frequency bias of the forecast binary fields f_M/f_o varies with accumulation threshold. This is shown in Fig. 12a and can be compared with another measure of the bias, the AFSS (see section 2), which is given by $2f_o f_m/(f_o^2 + f_M^2)$ for each threshold (Fig. 12b). Starting with the 12-km model, Fig. 12a shows an overprediction of the number of low-accumulation pixels (<5 mm) and an underprediction of the number of high-accumulation pixels. Figure 12b shows that the 12-km model becomes progressively more biased (lower FSS) as the accumulation threshold gets larger. The two graphs may appear to be

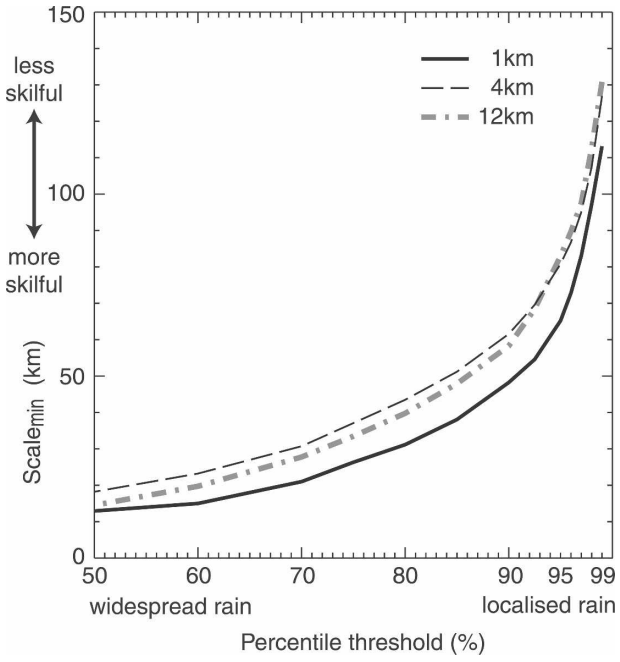

FIG. 11. Graph of aggregated scale$_{min}$ (see text) against percentile threshold for the 12-, 4-, and 1-km model forecasts.

somewhat contradictory, but they provide a different perspective. The frequency bias (Fig. 12) is computed from all of the binary pixels added together over the whole trial for each given threshold. It says how much a model under- or overpredicts, on average, but includes a cancellation between over- and underpredicting individual forecasts. The AFSS shown is an average from each of the forecasts. It does not have the problem of cancellation because it is always positive and therefore does not record the sense of the bias.

Returning to the 12-km forecasts, the overprediction (underprediction) of low (high) thresholds is characteristic of a model that does not have sufficient resolution to represent most of the convection explicitly, and instead has to rely on a convection parameterization scheme. The decrease of AFSS with threshold (Fig. 12b) indicates that the individual forecast biases increase with threshold. The low scores for high thresholds are the combined result of two modes of behavior in the 12-km model. Most of the convection is parameterized and large rainfall totals are underpredicted. However, on a few occasions the 12-km model attempts to resolve the more intense storms, but inadequate convective parameterization leads to intense dynamical ascent, resulting in excessive resolved rain at small scales.

In contrast to the 12-km model, the 4-km model underpredicts (overpredicts) the low (high) thresholds. This behavior is characteristic of a model that tries to represent the convection explicitly, but still lacks sufficient resolution. In general, it does not generate enough small showers, but delays initiation and then generates larger, more intense, and well-separated storms. Petch (2006) describes similar behavior when clouds are underresolved in a cloud-resolving model (CRM). The AFSS curve decreases with increasing threshold, and then reaches a minimum for accumulations of 6–10 mm. At these thresholds, the worst individual forecast biases

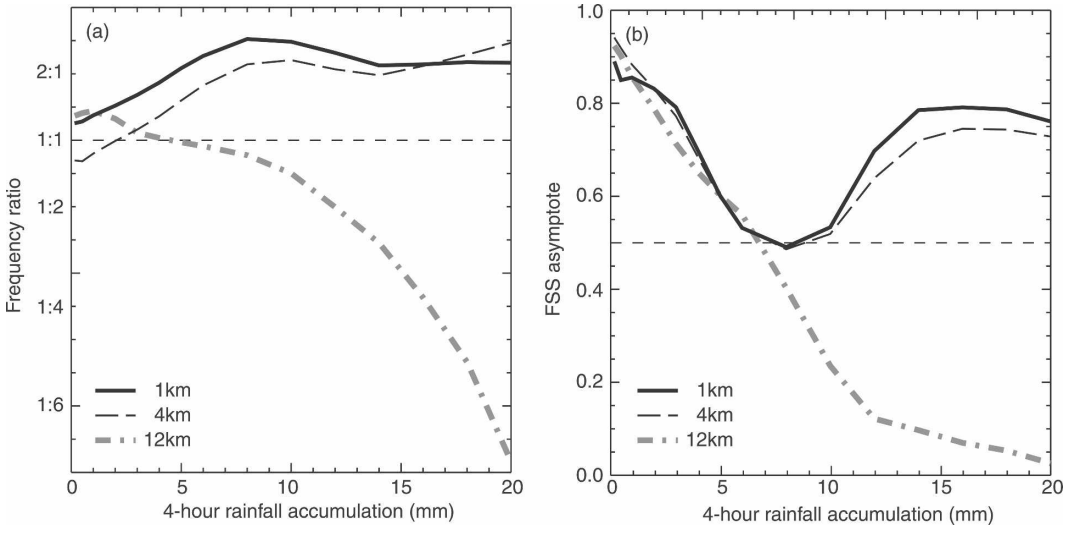

FIG. 12. (a) Graph of the ratio of model frequency f_M to observed frequency f_o (the base rate) against rainfall accumulation threshold for each model resolution for hours 4–7. Above (below) the 1:1 ratio line $f_M > f_o$ ($f_o > f_M$). (b) A graph of FSS$_{asymptote}$ (see text) against rainfall accumulation threshold for each model resolution.

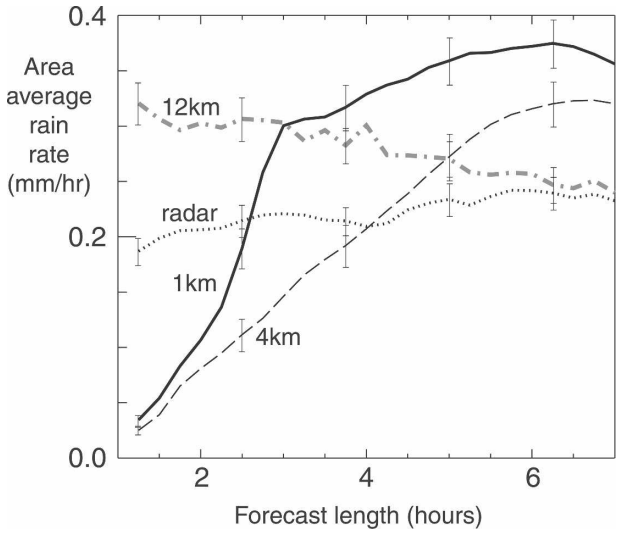

FIG. 13. Graph of aggregated domain-averaged rainfall rate against time from radar and the 12-, 4-, and 1-km models, including standard errors taken over the 40 forecast periods.

occurred, but not all forecasts overpredicted; at higher thresholds the 4-km model overpredicted most of the time.

The 1-km model overpredicts the amount of rain for all thresholds; otherwise, the curves in both graphs exhibit similar behavior to that of the 4-km model, including the minimum in asymptotic FSS at 6–10 mm. The main difference between the two is that the bias indicated by both measures is larger in the 1-km model for most of the thresholds (>2 mm). A large part of this is because of a faster spinup at 1 km.

The spinup period for the 4- and 1-km models can be seen in the domain average rain rates (Fig. 13). They both start with very little rain, and then as convection develops rain rates increase; this happens more quickly in the 1-km model. Later, when the early buildup of instability is released, both models are characterized by an "overshoot," because too much convective rain is produced.

4) HOURLY ACCUMULATIONS

The variation in FSS with time is shown in Fig. 14 for a neighborhood length (spatial scale) of 55 km. This scale was chosen because it is far enough away from the grid scale of the models, yet small enough to be of interest for forecast applications. Figures 14a,b,c show results using accumulation thresholds of 0.2, 1.0, and 4.0 mm, respectively. For very small accumulations (0.2 mm), the 12-km model maintains skill at a scale of 55 km throughout the 6-h period, whereas skill decreases for moderate accumulations (1.0 mm) and decreases rapidly for high accumulations (4.0 mm). Much of the reduction in skill for high accumulations in the 12-km model comes from a worsening of the underprediction of heavy rain with time. This is thought to be due to 1) a premature decay of parameterized rain because of the inability of a convection scheme to organize convection, and 2) the reduction of any "resolved" convection that was introduced by the data assimilation. The 4- and 1-km models are initially less skillful than the 12-km model, because they have not had time to spin up from the 12-km initial state. However, as the forecasts progress, the 1-km model initiates convection and becomes more skillful than the 12-km model after 1–2 h. The 4-km model takes longer to spin up; for the 0.2-mm threshold it takes the entire 6 h to achieve comparable skill with the 12-km model, for 1.0 mm it takes 2–3 h, and for 4.0 mm it behaves much like the 1-km model (taking 1–2 h). Examination of the fields has revealed that this variation of spinup time with threshold is the result of the tendency of the 4-km model to initially produce a small number of intense cells, and therefore to generate a signal at higher thresholds more rapidly.

Figures 14d,e,f show results using percentile thresholds. All of the models become more skillful as the percentile threshold becomes lower (i.e., more rain is sampled). The skill of the 12-km model decreases with time for all three percentile thresholds, as it did for the 1.0- and 4.0-mm accumulation thresholds. However, the decline is not as dramatic as that for the 4.0-mm threshold. The removal of the impact of the bias accounts for some of the slowing of the loss of skill, but it is still evident that the spatial accuracy of the forecasts has also diminished with forecast length. At 4 km, the spinup period that is so evident in the accumulation thresholds is hardly noticeable in the percentile thresholds. Over the first approximately 2–3 h, the skill does not change, suggesting that the 4-km model spinup is characterized by an improvement in the bias as new cells develop, but there is little change in spatial accuracy. Later on there is an improvement in FSS and the spatial accuracy exceeds or matches that of the 12-km model over the last 2 or 3 h for all percentile thresholds.

In contrast, the spatial distribution of the rain in the 1-km model improves rapidly over the first hour for the 90th and 75th percentile thresholds. This occurs as the development of new convective cells shifts some of the emphasis away from rain that was initially resolved in the coarser 12-km fields toward new regions of convective activity. It is not seen for the 98th percentile threshold (more localized rainfall) because, despite new triggering, the very highest accumulations remain largely within the areas where rain was resolved at 12 km. Over the final 4 h, the 1-km model is more accurate than the

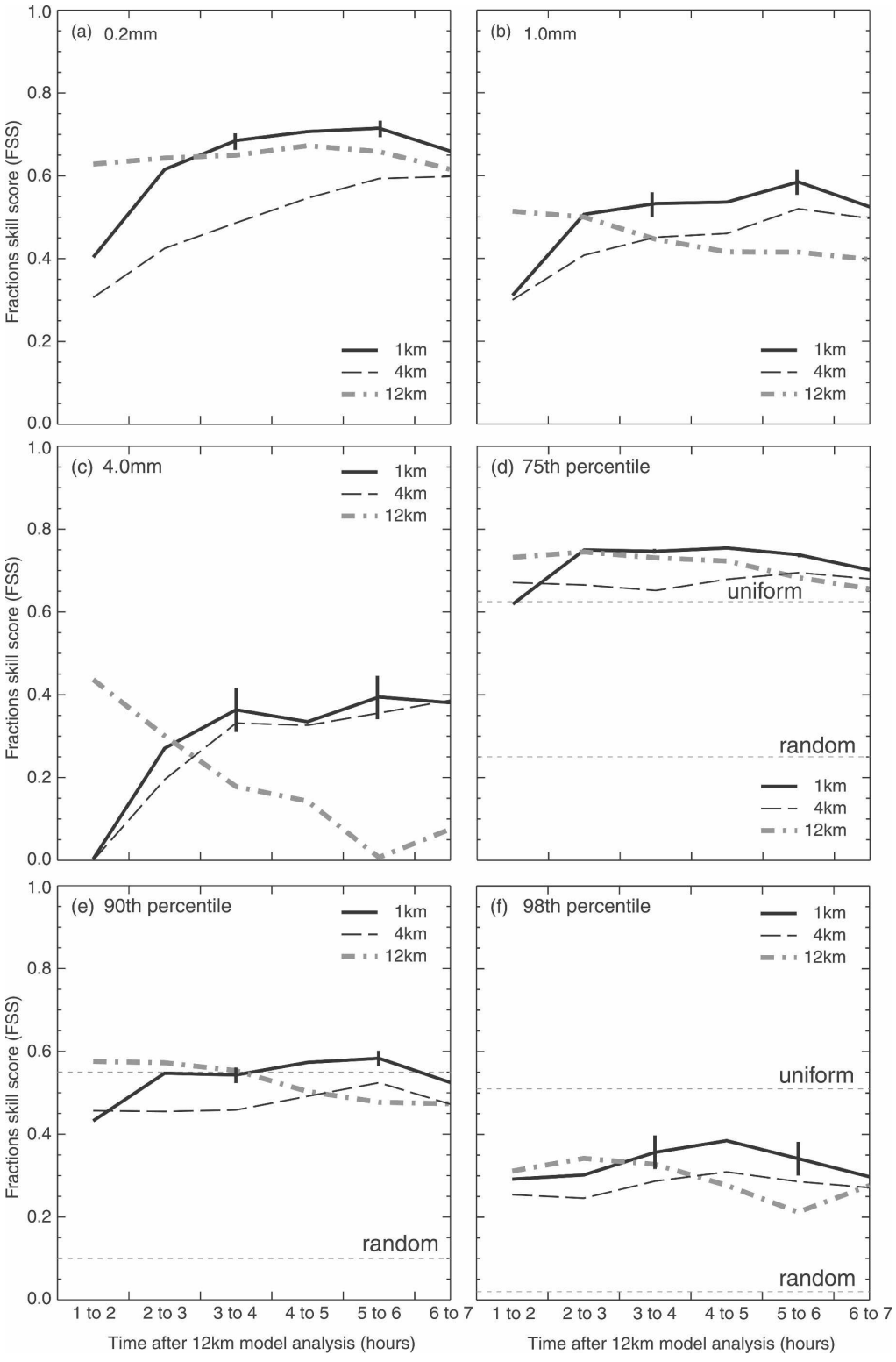

FIG. 14. Graphs of aggregated FSS against time for hourly accumulations over the period from $T + 1$ to $T + 7$ (relative to the 12-km model analysis time) for the 12-, 4-, and 1-km models, using a neighborhood square of length 55 km, with (a)–(c) accumulation thresholds of 0.2, 1.0, and 4.0 mm and (d)–(f) percentile thresholds of 75%, 90%, and 98%, respectively.

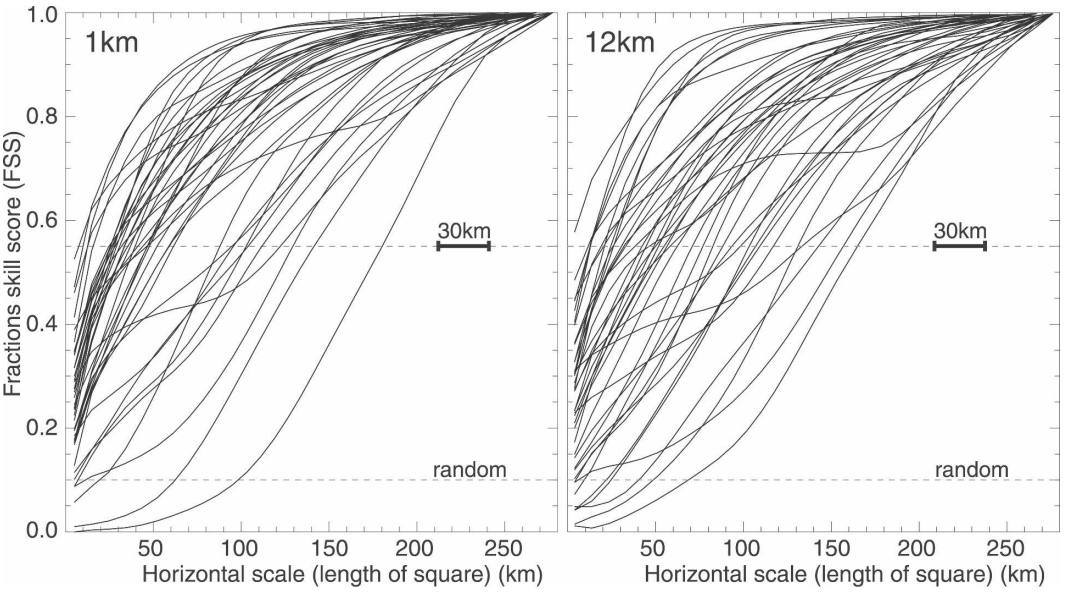

FIG. 15. Graph of FSS against neighborhood length for the 90th percentile threshold and 4-h accumulation period for each of the 12- and 1-km model forecasts.

other two resolutions for all of the percentile thresholds.

Neighborhoods bigger and smaller than 55 km have also been examined, but are not shown. At larger scales the same differences between the models are evident for a given threshold, but the overall skill is greater. At smaller scales (~20 km), the 1-km model is still the most skillful, but the improvement is less and all of the models are less skillful. At scales approaching the verification grid length (5 km), forecast skill is considerably less in all models and the differences between the models becomes small.

5) FORECAST SPREAD FOR 4-H ACCUMULATIONS (HOURS 3–7)

Figures 15 and 16 show the large variability in skill from forecast to forecast in both the 1- and 12-km models, which is something that is often overlooked when evaluating model performance. An examination of the individual forecast skill is used here to reveal how the aggregated improvement in skill at 1 km is achieved. We see that most of the 1-km forecasts exceed the target skill ($FSS_{uniform}$) at scales of <50 km, whereas the 12-km model has a larger proportion of less accurate forecasts. Figure 16 shows that the aggregated improvement in FSS from the 1-km model (shown in Fig. 10) is the result of a shift to a greater concentration of forecasts with higher skill ($scale_{min} < 50$ km). The 1-km model is more accurate on average, but individual 1-km forecasts can still be worse than their 12-km counterparts.

The scale at which $FSS_{uniform}$ is achieved ($scale_{min}$) has been extracted from hourly accumulation FSS curves for each of the forecasts. If each good forecast tends to remain good and each poor forecast tends to remain poor, then the value of $scale_{min}$ at one hour should be strongly correlated with $scale_{min}$ over subsequent hours. Figure 17 shows how the correlation co-

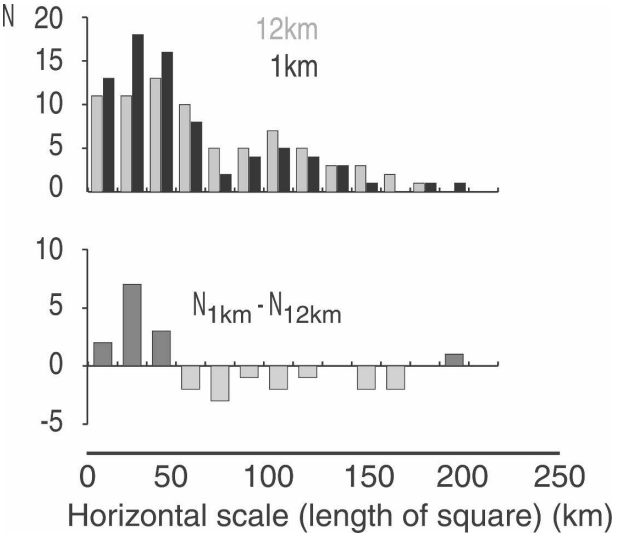

FIG. 16. (top) The number of 1- and 12-km forecasts that intercept the $FSS_{uniform}$ line in Fig. 15 as a function of horizontal scale (neighborhood length). Each bar represents the number of intercepts within a 30-km section displaced by 15 km from the next. (bottom) The differences between the lengths of the corresponding 12- and 1-km bars in the top graph.

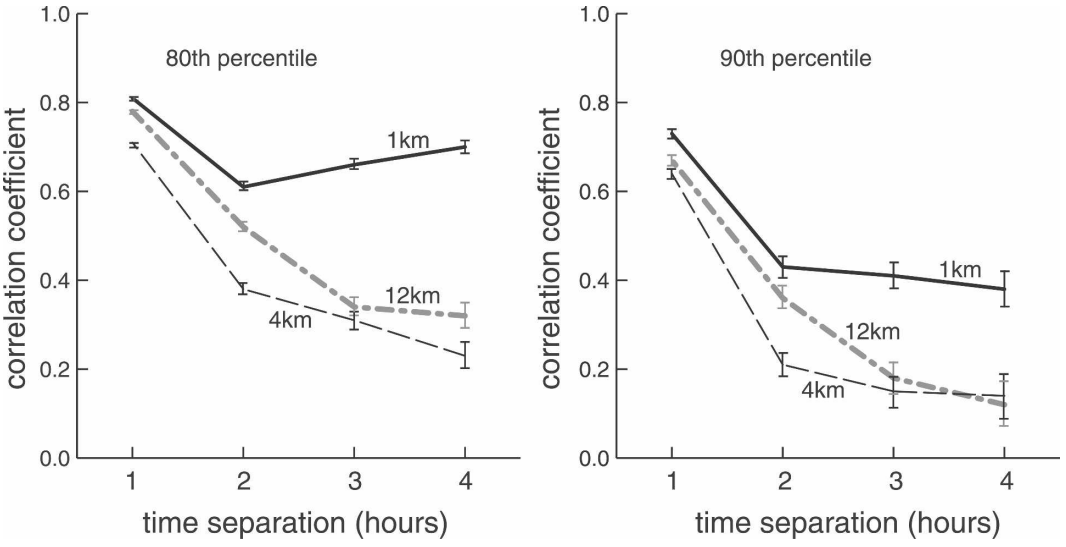

FIG. 17. Graphs of how the correlation coefficient for $\text{scale}_{\text{min}}$ (see text) varies with the time interval between hourly accumulations for each model resolution, using percentile thresholds of (a) 80% and (b) 90%. The error bars were obtained by resampling using random variations within the range of the radar error.

efficient between the $\text{scale}_{\text{mins}}$ varies with time separation. Focusing on the 80th percentile threshold, the correlation drops off quickly with time separation for the 12- and 4-km models, but remains high at 1 km (~0.7). The difference comes from the ability of the 1-km model to generate realistic showers that retain continuity, whereas the 12-km rainfall is largely generated by a convection scheme with no memory and the 4-km model is suffering from delayed initiation through the spinup period. The 90th percentile threshold shows similar behavior, but has lower values, which is consistent with more localized rainfall retaining less continuity.

The correlation between the three resolutions for a range of percentile thresholds is shown in Fig. 18. Not surprisingly, there is a very high correlation between all resolutions when widespread rain is being sampled because all of the models start from the same initial conditions and the preferred regions of convection are modulated by the same mesoscale dynamics. When more localized rainfall is sampled the correlation is smaller because there is more freedom for differences between the forecasts to emerge. The 4- and 1-km models are the most similar at these small scales because the differences between two models that represent convection explicitly are less than that between the parameterized and explicit realizations. The correlation between resolutions implies that an improvement in the skill of the 12-km model should also have a substantial impact at 4 and 1 km and highlights the importance of getting the larger scales correct. This is an important factor to take into account when introducing independent data assimilation into a 4- or 1-km model.

5. Discussion

a. 1-km performance

The 1-km model was more skillful than the 12-km model over all but the smallest scales for both 4-h and

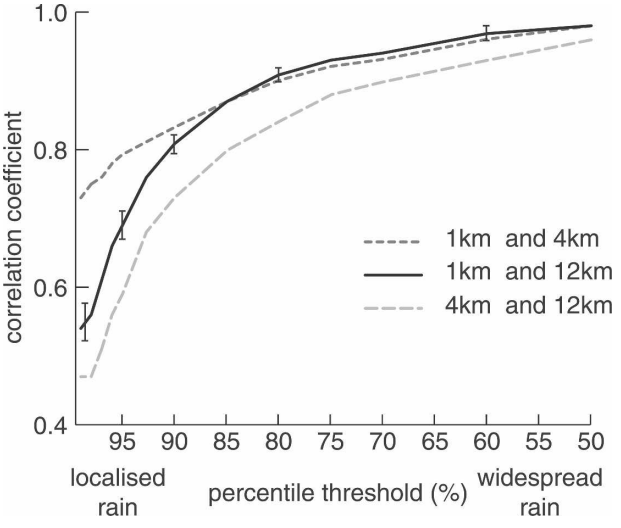

FIG. 18. Graph of how the correlation coefficient for $\text{scale}_{\text{min}}$ (see text) between the different resolutions varies with percentile threshold, for rainfall accumulations over the final 4 h of the forecasts. The error bars were obtained by resampling using random variations within the range of the radar error and are applicable for all three lines.

hourly accumulations. If verification had only been performed at the grid scale, the improvement from the 1-km model would not have been detected. Indeed, it may even have appeared worse, supporting the view that much of the extra small-scale detail can be regarded as noise. The 1 (12)-km models reached the target level of skill ($FSS_{uniform}$) at scales of 45–60 km (50–80 km) for the 90th–95th percentile thresholds (Fig. 10). In terms of an application, such as flood warning, useful skill was achieved on the scales of significantly smaller river catchments. Furthermore, the greatest improvement in skill occurred for the higher-accumulation thresholds, which have the biggest societal impact. Much of the improvement was due to convection being explicitly represented rather than parameterized and due to a more accurate representation of predictable local effects (e.g., orographic uplift, sea breezes). The 1-km model is still under development, and further improvements will be made. Unlike the 12-km model, it has not been tuned for operational performance. A moist turbulence parameterization for cloud mixing outside the boundary layer was not included at this stage, which may account for much of the overprediction of rainfall amounts. In addition, the forecasts were spun up from 12-km fields, which had a significant impact on the first few hours of the forecasts. An operational 1-km model is not likely be used in a nowcasting context unless it is part of a continuous cycle with data assimilation.

b. 4-km performance

The 4-km model performs poorly compared with the 1-km model, and shows little or no improvement on the 12-km model at any scale apart from the high-accumulation thresholds. The initiation of explicit convection is delayed, which results in a longer spinup period. Once showers are formed they are too large, intense, and well spaced; they then persist for too long. These characteristics lead to errors in the location and amount of rainfall, which are signaled in the verification scores. It is an inherently difficult resolution to use for predicting convective rainfall, as noted by Deng and Stauffer (2006), because convection is neither adequately resolved nor satisfactorily parameterized. However, a grid spacing of ~3–5 km is not a resolution we can ignore because it is the finest that is currently affordable for operational forecasting in the United Kingdom, and in future will be required to provide boundary information for higher-resolution models.

c. Forecast skill and presentation

The skill of all the models increases with spatial scale, but the practical benefit reduces as sharpness is lost (see Fig. 1), which means that there is an optimal range of scales over which model output should be used. In addition, there is a variation with threshold—the higher the accumulation or the smaller the percentile threshold the less accurate the models become, but the greater the benefit from higher resolution. A measure of acceptable skill, $FSS_{uniform}$, has been introduced to define the smallest scale over which a model might be considered useful ($scale_{min}$). That scale can be used to define an appropriate smoothing kernel for generating probabilistic output from deterministic forecasts using a nearest-neighborhood method. Smaller scales are then regarded as being unpredictable and are treated as stochastic noise. The drawback with this approach is that it does not take account of day-to-day variations in skill between forecasts (Fig. 15), which applies to all resolutions, whatever the average value of $scale_{min}$. A single filtering scale cannot be appropriate for every occasion, but adjustments to $scale_{min}$ on a forecast-by-forecast basis would require a priori information about the expected accuracy of each forecast. Yet, it appears that this is possible for short forecasts using the 1-km model. We have seen from Fig. 17 that $scale_{min}$ for one hour is correlated with $scale_{min}$ in successive hours, provided that large enough rain areas are sampled; that is, a good (poor) forecast at one time leads to a good (poor) forecast at later times. Such a relationship opens up the possibility of real-time predictions of suitable filtering scales for presenting short-range kilometer-scale NWP model output.

6. Conclusions

A verification method is presented that is designed to measure how the skill of precipitation forecasts varies with spatial scale and determine what scales should be believed. It has been used to assess the performance of 12-, 4-, and 1-km versions of the UM from a sample of 40 forecast periods. The purpose was to examine the improvement to forecast skill from increased resolution alone. Data assimilation is a separate issue for subsequent papers.

The results from this trial have shown that the 1-km model is indeed more skillful than the 12-km model (after the spinup period) at all scales for which a comparison is meaningful (>15 km). The improvement comes from a more accurate distribution of the rain and a better prediction of high accumulations, although there is an overprediction of rainfall amounts. A satisfactory level of skill ($scale_{min}$, defined in section 2d) is reached at scales around 20%–30% shorter than that achieved by the 12-km model. The indications are that a 1-km model is capable of a significant improvement in

rainfall predictions over scales that are useful for flood prediction, even if skill close to the grid scale is low. The 4-km model does not achieve the same level of performance because of inherent difficulties in representing convection at that resolution. It is anticipated that the introduction of data assimilation at a high resolution will improve skill further, particularly over the first few hours when the spinup from a coarse-resolution initial state is a problem. The verification approach described here will provide a tool for assessing how new developments impact scale-dependent forecast skill and for defining the scales over which output should be presented. It may be possible to develop an adaptive presentation of forecasts in which the scales over which output is filtered vary over the domain according to the variation in spatial accuracy of an earlier time. Such a system may be able to "lock on" to more predictable features, such as showers tied strongly to orography, and assign more uncertainty elsewhere. Whether the improvement in skill brought about by increased resolution is sufficient to warrant the extra cost will ultimately depend on the requirements of a forecast system.

Acknowledgments. The authors thank Andrew Macallan for his technical assistance, and Brian Golding, Peter Clark, Mark Dixon, Chris Ferro, Ian Jolliffe, and Barbara Casati for useful comments about this work.

REFERENCES

Bousquet, O., C. A. Lin, and I. Zawadzki, 2006: Analysis of scale dependence of quantitative precipitation forecast verification: A case-study over the Mackenzie River basin. *Quart. J. Roy. Meteor. Soc.,* **132,** 2107–2125.

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

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

Cullen, M. J. P., T. Davies, M. H. Mawson, J. A. James, S. C. Coulter, and A. Malcolm, 1997: An overview of numerical methods for the next generation UK NWP and climate model. *Numerical Methods in Atmospheric and Ocean Modelling: The Andre J. Robert Memorial Volume,* C. A. Lin, R. Laprise, and H. Ritchie, Eds., Canadian Meteorological and Oceanographic Society, 425–444.

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

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.,* **134,** 1772–1784.

Deng, A., and D. R. Stauffer, 2006: On improving 4-km mesoscale model simulations. *J. Appl. Meteor. Climatol.,* **45,** 361–381.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmos. Sci. Lett.,* **5,** 110–117.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

Essery, R., M. Best, and P. Cox, 2001: MOSES 2.2 technical documentation. Met Office, Hadley Centre Tech. Note 30, 30 pp.

Golding, B. W., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteor. Appl.,* **5,** 1–16.

Gregory, D., and P. R. Rowntree, 1990: A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure. *Mon. Wea. Rev.,* **118,** 1483–1506.

Harrison, D. L., S. J. Driscoll, and M. Kitchen, 2000: Improving precipitation estimates from weather radar using quality control and correction techniques. *Meteor. Appl.,* **7,** 135–144.

Jones, C. D., and B. Macpherson, 1997: A latent heat nudging scheme for the assimilation of precipitation data into an operational mesoscale model. *Meteor. Appl.,* **4,** 269–277.

Lean, H. W., and P. A. Clark, 2003: The effects of changing resolution on mesoscale modelling of line convection and slantwise circulations in FASTEX IOP 16. *Quart. J. Roy. Meteor. Soc.,* **129,** 2255–2278.

Lock, A. P., A. R. Brown, M. R. Bush, G. M. Martin, and R. N. B. Smith, 2000: A new boundary layer mixing scheme. Part I: Scheme description and single-column model tests. *Mon. Wea. Rev.,* **128,** 3187–3199.

Lorenc, A. C., and Coauthors, 2000: The Met Office global three-dimensional variational data assimilation scheme. *Quart. J. Roy. Meteor. Soc.,* **126,** 2991–3012.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.,* **26,** 636–646.

Macpherson, B., B. J. Wright, W. H. Hand, and A. J. Maycock, 1996: The impact of MOPS moisture data in the U.K. Meteorological Office mesoscale data assimilation scheme. *Mon. Wea. Rev.,* **124,** 1746–1766.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.,* **83,** 407–430.

Marzban, C., and S. Sandgathe, 2006: Cluster analysis for verification of precipitation fields. *Wea. Forecasting,* **21,** 824–838.

——, and ——, 2008: Cluster analysis for object-oriented verification of fields: A variation. *Mon. Wea. Rev.,* in press.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.

Mylne, K. R., 2002: Decision-making from probability forecasts based on forecast value. *Meteor. Appl.,* **9,** 307–315.

Petch, J. C., 2006: Sensitivity studies of developing convection in a cloud-resolving model. *Quart. J. Roy. Meteor. Soc.,* **132,** 345–358.

Potts, J. M., 2003: Basic concepts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley & Sons, 13–36.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667.

Roberts, N. M., 2003: The impact of a change to the use of the convection scheme in high-resolution simulations of convective events. Met Office Forecasting Research Tech. Rep. 407, 30 pp.

Romero, R., C. A. Doswell III, and R. Riosalido, 2001: Observations and fine-grid simulations of a convective outbreak in northeastern Spain: Importance of diurnal forcing and convective cold pools. *Mon. Wea. Rev.,* **129,** 2157–2182.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.,* **132,** 3019–3032.

Smith, R. N. B., E. M. Blyth, J. W. Finch, S. Goodchild, R. L. Hall, and S. Madry, 2006: Soil state and surface hydrology diagnosis based on MOSES in the Met Office Nimrod nowcasting system. *Meteor. Appl.,* **13,** 89–109.

Speer, M. S., and L. M. Leslie, 2002: The prediction of two cases of severe convection: Implications for forecast guidance. *Meteor. Atmos. Phys.,* **80,** 165–175.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.,* **12,** 257–268.

Walser, A., D. Lüthi, and C. Schär, 2004: Predictability of precipitation in a cloud-resolving model. *Mon. Wea. Rev.,* **132,** 560–577.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.,* **125,** 527–548.

Wilson, D. R., and S. P. Ballard, 1999: A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Quart. J. Roy. Meteor. Soc.,* **125,** 1607–1636.

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

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.,* **60,** 1173–1185.