The Convective-scale Unified Model: Evaluating NWP precipitation forecasts

Nigel Roberts, Richard Forbes, Humphrey Lean, Peter Clark

Met Office, UK

JCMM, Met Office, Meteorology Building, University of Reading,
Reading, Berkshire, RG6 6BB, U.K.

November 2005

1 Convective-Scale NWP

Over the past few years, the Met Office has been active in research and development of the Unified Model for convective-scale Numerical Weather Prediction. A trial system was set up (called the High Resolution Trial Model, HRTM) with a 1km grid-resolution version of the UM (centred on the Chilbolton research radar in southern England), nested inside a 4km grid-resolution model for the southern UK, nested in the operational 12km model (See Fig. 1). The focus of the convective-scale UM is forecasting severe convective precipitation, particularly for advanced warnings of potential flood events. Results from a number of mainly summer convective cases during 2003 and 2004 have lead to gradual improvement of the model and in the spring of 2005 the 4km UM went operational for a domain covering the whole of the UK. In the Autumn of 2005, the operational system will be enhanced with a data assimilation cycle with 36 hour forecasts 4 times a day. Research is ongoing to improve the model and data assimilation system, understand the convective-scale processes through observational campaigns such as the Convective Storms Initiation Project (CSIP, summer 2004/2005), and investigate convective-scale predictability issues for forecasting, verification and presentation of the NWP forecasts. This note describes one method of precipitation forecast verification for high resolution models.

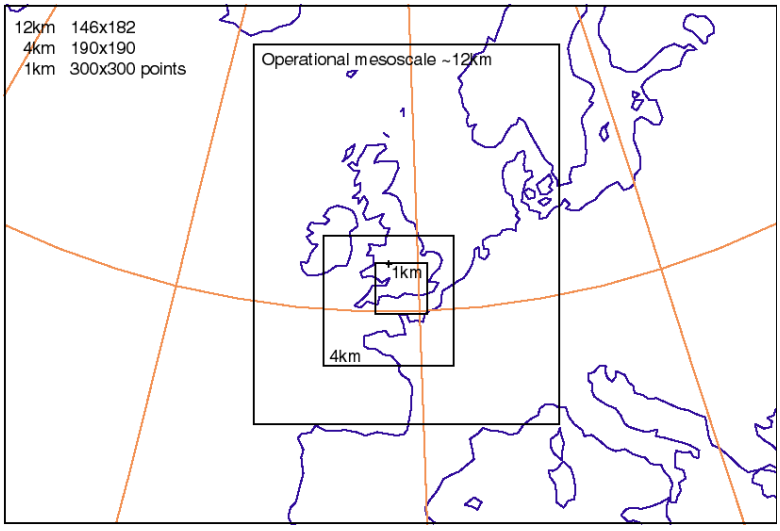


Figure 1. Test 1km and 4km grid-resolution domains for the convective-scale.

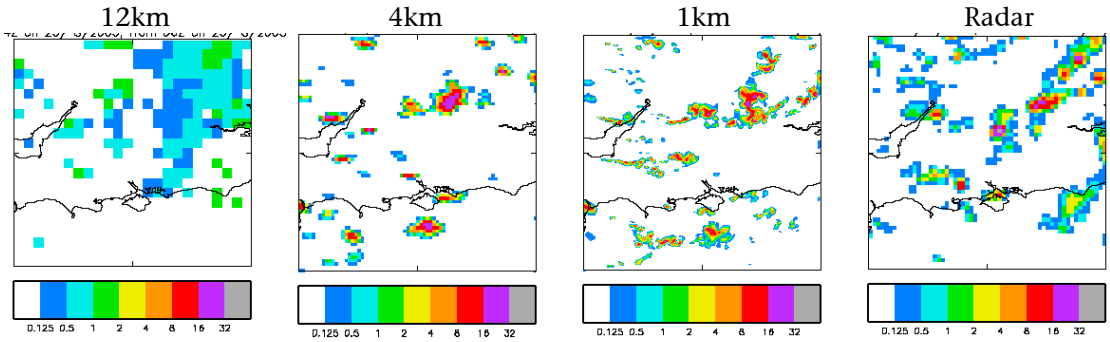


Figure 2. An example forecast of surface rain rate from different resolution models (from left to right, 12km, 4km, 1km) compared with radar-derived rain rate (far right) for a CSIP case study with severe convection (14Z on 25th August 2005). All fields are shown for the 1km domain over the central southern UK. The 12km model has parametrized convection, the 4km and 1km are "resolving" the deep convection and provide a better indication of the shower structures and intensities, but not always the exact locations.

2 Predictability and verification

Although the locations of initiation of some convective storms are more predictable than others (e.g. orographic initiation), predictability of the small spatial scales over the period of an NWP forecast is a major issue. For example, the 1km-gridlength model can not be accurate at the gridscale or even over a few kilometres because features are not properly resolved and predictability is low on those scales. Errors at smallest scales grow fastest and there will always be a scale below which even a good forecast is essentially random in nature. Put another way, individual showers can not be predicted in exactly the correct locations even if a forecast is good (e.g. Fig. 2). This can result in poor verification scores when traditional verification methods based on gridpoint comparisons are used. Although we may not be able to predict the exact location of individual convective cell, we do want the region of shower activity to be well positioned with the right type of structure and intensity. Alternative verification scores that provide information on the skill of a forecast over a range of spatial scales can therefore be more informative.

As far as the convective-scale version of the Met Office Unified Model is concerned, we would like to be able to at least attempt to answer the following questions in an objective way.

1. How accurately can we forecast precipitation over areas the size of counties, river catchments or urban areas using a particular model?
2. How does the predictable scale change with forecast time? This may be in terms of defining the smallest river-catchment area for which forecasts are useful.
3. Does the rainfall analysis agree with the radar picture at the scale of the data assimilation? Data assimilation methods are designed to add observational information to a model over particular spatial scales.
4. What are sensible products to generate for customers?
5. Does a change to either model resolution or formulation make a difference to the predictable scale?

This list of questions justifies the need to have a verification system that can be used to determine the relationship between forecast skill and spatial scale.

3 The verification approach

Figure 3 shows an example of rainfall accumulations measured by radar on a particular day and two NWP forecasts – one with a 12-km grid length and the other with a 1-km grid length (note this is a different case to that shown in Fig 2.). The 1-km forecast produced a great deal more structure than the 12-km forecast and subjectively looks closer to the radar. It gave a much better indication of the higher accumulations that we are most likely to be interested in. It also, correctly, produced rain over the sea in this case because the convection was explicitly resolved rather than parametrised by a convection scheme. However, if Figure 3 is examined more closely, it is clear that the areas of higher accumulations in the 1-km forecast are not in exactly the same place as observed by radar. It is very likely that if a standard grid-square by grid-square verification were performed, the 1-km forecast might come out worse, even though we can see that it is a 'better' (or more useful) forecast. This is the challenge – to be able to objectively verify rainfall forecasts by the same criteria we use in subjective assessment. It is one reason why verification over different spatial scales is necessary.

It was decided to verify threshold exceedance of precipitation accumulations rather than precipitation rates as it is the former that matters most for flooding and it is sensible to smooth in time if we are also smoothing in space. Verifying against radar rather than rain gauges provides a much better spatial coverage.

For every grid square, we compute the fraction of surrounding points within a given area that exceed a particular accumulation threshold over a given period. This will give a fraction for every grid square. The fractions can be considered as probabilities. They give an indication of the chance of an accumulation threshold being exceeded at each grid square, given that we think the model could be in error on a scale of the size of the area used to produce the fractions.

Fractions/probabilities can be generated over different spatial scales by changing the size of the area. For the purposes of verification, squares of different sizes are used to compute fractions for different spatial scales.

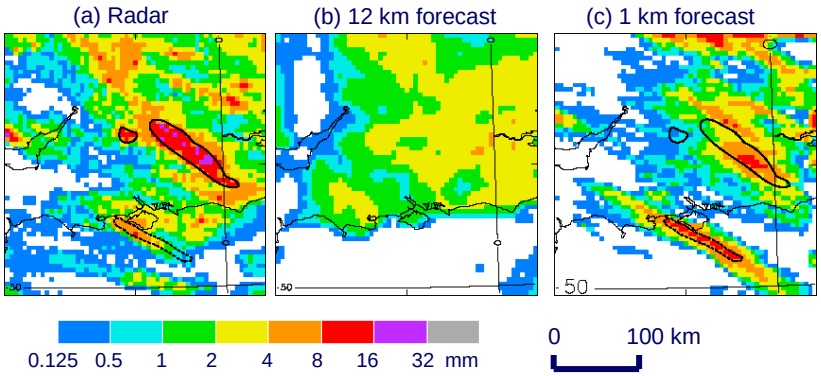


Figure 3. Rainfall accumulations over 6 hours, 13 to 19UTC 13th May 2003 interpolated/averaged to a 5 km grid spacing on a square 300x300km over southern England. (a) Observed by radar, (b) 12 km gridlength model forecast, (c) 1 km gridlength model forecast.

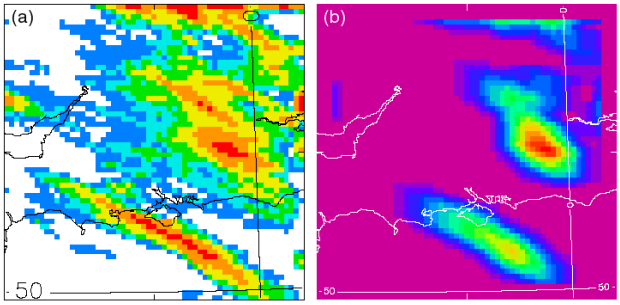


Figure 4. (a) the same as Figure (c). (b) shows the fractions for accumulations > 4 mm computed using squares of 35x35 km.

Figure 4(b) shows how fractions/probabilities have been generated using the approach described at every grid point (except around the edge) for a 1-km forecast (averaged to a 5-km grid). In this example the accumulation threshold used is 4 mm, the squares are 35x35km and the accumulation period is 6 hours.

One of the verification scores to compare forecast fractions over an area between the model and radar is the Fractions Skill Score (**FSS**); a variation on the Brier Skill Score. It is given by: -

$$FSS = 1 - \frac{FBS}{\frac{1}{N}\left[\sum_{j=1}^{N}(p_j)^2 + \sum_{j=1}^{N}(o_j)^2\right]}$$

$0 \le p_j \le 1$ forecast fraction

$0 \le o_j \le 1$ radar fraction

where,

$$FBS = \frac{1}{N}\sum_{j=1}^{N}(p_j - o_j)^2$$

(Fractions Brier Score) is a version of the Brier score in which fractions are compared with fractions

and,

$$\frac{1}{N}\left[\sum_{j=1}^{N}(p_j)^2 + \sum_{j=1}^{N}(o_j)^2\right]$$

is the worst possible FBS in which there is no colocation of non-zero fractions

The Fractions Skill Score has the following characteristics

- It has a range of 0 to 1; 0 for a complete forecast mismatch, 1 for a perfect forecast.
- If either there are no events forecast and some occur, or some occur and none are forecast the score is always 0.
- As the size of the squares used to compute the fractions gets larger, the score will asymptote to a value that depends on the ratio between the forecast and observed frequencies of the event. I.e. the closer the asymptotic value is to 1, the smaller the forecast bias.
- The score is most sensitive to rare events (or for small rain areas).

As with any verification score, this one has characteristics that are both helpful and misleading. It is not necessarily the best way of comparing fractions with fractions, but it has proved useful for providing the sort of information we are interested in.

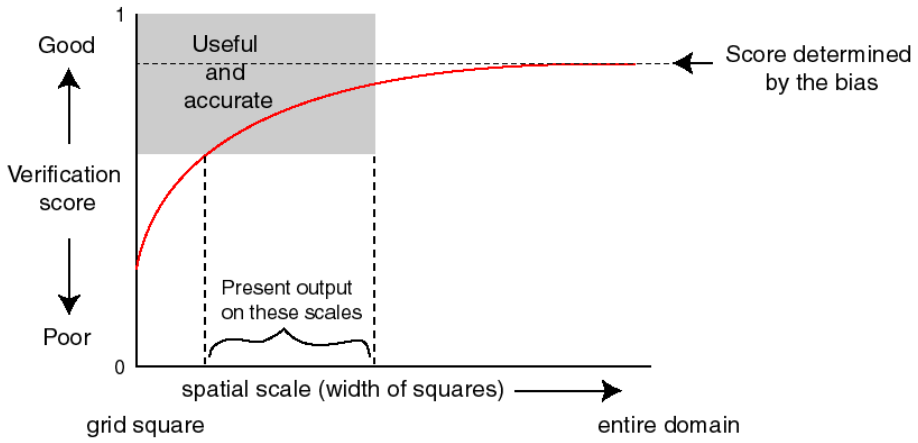


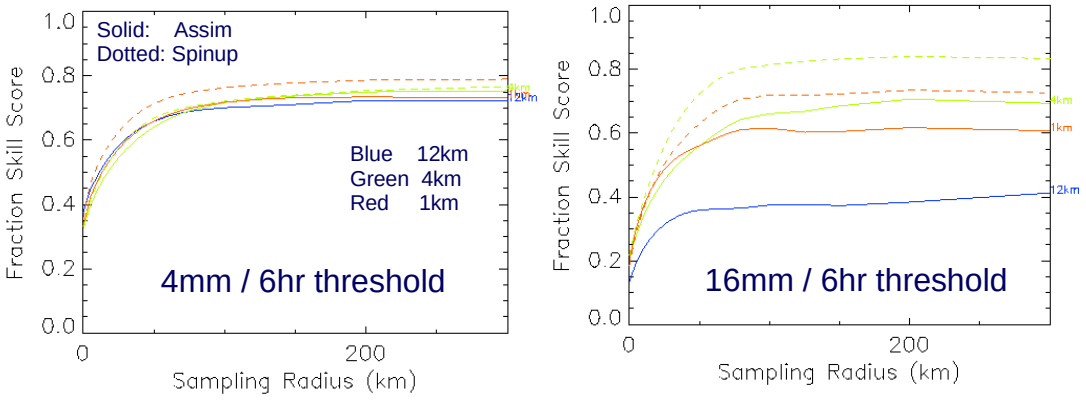
Figure 5. A schematic graph of the behaviour of the Fractions Skill Score with spatial scale.

Figure 5 shows how the Fractions Skill Score is expected to behave over a large number of forecasts for a particular accumulation period and threshold. The least skill is expected to be at the grid scale. Skill should increase with spatial scale (square size) until it reaches an asymptote that is determined by the forecast bias. The grey shading depicts the part of the graph where the score is deemed high enough for the forecast to be regarded as skilful and the spatial scale is small enough for the forecast to be useful. (There is little benefit to be gained from a forecast that is detailed but inaccurate or accurate but lacking detail). We can then, in principle, pick out the range of spatial scales over which the forecast should be presented to users/customers – i.e. the size of squares used to generate probabilities.
For a more detailed discussion, see Roberts 2004(a,b).

4 Some results for summer 2004 convective case studies

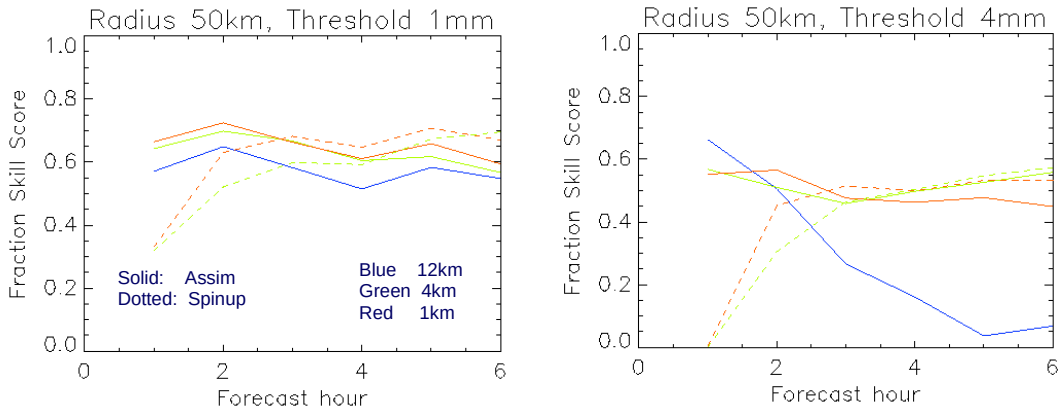
A graph of Fractions Skill Scores is shown in Figure 6 for a comparison of 6-hourly rainfall accumulations from seven summer 2004 case studies. The intention here is to show that the verification method can provide useful information about the differences in performance of the models over different spatial scales. All models have the least skill at the grid-scale and asymptote to a value less than one over large spatial scales – an indication that there is a bias (for the particular accumulation threshold). For the relatively small threshold of 4mm/6hours accumulation, the 1km and 4km models have higher skill than the 12km model, but only just. For the higher threshold of 16 mm/6 hr, the level of skill of the 4km and 1km models is much higher than for the 12km model.

Figure 7 shows the Fractions Skill Scores for 1 hour rainfall accumulations for the three models as a function of forecast hour for a spatial scale of 50km. The benefit of the high resolution models becomes clearer for the shorter timescales and higher accumulation thresholds. Note the differences between the high resolution model forecasts "spinning up" from the 12km analysis (which does not have "resolved" convection) (dotted lines in Figs. 6 and 7) compared to the forecasts starting from a 4km resolution analysis (which includes "resolved" convection) (solid lines in Figs. 6 and 7). The skill of the "spin-up" runs is generally slightly higher than the "assimilation" runs for this particular trial, due to a larger bias in the latter. This bias is being addressed. However, Fig. 7 shows the advantage of the high resolution assimilation (3DVAR with latent heat nudging) in the early part of the forecast; models spinning up from the 12km analysis take 2 to 3 hours to reach the same forecast skill as the models with high resolution data assimilation.



(a) (b)

Figure 6. A graph of the variation of the Fractions Skill Score with spatial scale for 6-hour accumulation forecasts (T+0 to T+6) from the 1km, 4km and 12-km gridlength models for seven case studies during the Summer of 2004, compared with radar accumulations. The accumulation threshold was (a) 4mm/6hr and (b) 16 mm/6hr.



(a) (b)

Figure 7. A graph of the variation of the Fractions Skill Score with forecast hour for 1 hour accumulation forecasts from the 1km, 4km and 12-km gridlength models for seven case studies during the Summer of 2004, compared with radar accumulations. The spatial scale was 50 km and the accumulation thresholds were (a) 1mm and (b) 4 mm.

5 Summary

Convective-scale NWP is now operational at the Met Office and one of the questions for high resolution forecasts of convection is how to verify them in a meaningful way. Small scale features such as individual showers are inherently unpredictable and it is not sensible to verify raw model output on a grid point by grid point basis, because we should not expect skill at that scale. We need to be able to assess the accuracy of forecasts over different spatial scales (as a human observer would).

A simple method has been described, in which rainfall accumulation forecasts are compared with radar measured accumulations by converting both fields into fractions/probabilities of a particular accumulation threshold being exceeded. The results from a number of case studies have been encouraging, and importantly, the scores have

agreed with subjective evaluation. The methodology is intuitive and should be reasonably easy to convey to users/customers. Verification from seven cases during the summer of 2004 shows there is increasing skill for higher rain accumulations and shorter timescales as the resolution is increased. The impact of an initial implementation of 3DVAR data assimilation with latent heat nudging at 4km resolution is encouraging, resulting in a significant improvement in the early part of the forecast.

There are alternative approaches to verification of precipitation forecasts that could be compared with the approach described here. For example, Casati (2004) describes a technique based on wavelet analysis. The use of "top-hat" wavelet functions in this method is likely to give very similar results. It is also important to examine the behaviour of different verification scores that could be used to compare forecast and radar fractions (e.g. odds ratio). Radar data has so far been regarded as truth, and the issue of incorporating radar error needs to be addressed. A future application might be to investigate how forecast skill varies with spatial scale for different meteorological situations (e.g. scattered convection, organised convection or frontal)

6 References

Casati, B. 2004: New approaches for the verification of spatial precipitation forecasts. PhD Thesis, Department of Meteorology, University Of Reading, UK

Roberts, N. M., 2004: Measuring the fit of rainfall analyses and forecasts to radar. *JCMM internal report.,* **146**. *(NWP Technical Report.,* **432***), Met Office, UK. http://www.metoffice.gov.uk/research/nwp/publications/papers/technical_reports/2004.html*

Roberts, N. M., 2004: Verification of the fit of rainfall analyses and forecasts to radar. *JCMM internal report.,* **148**. *(NWP Technical Report.,* **442***), Met Office, UK. http://www.metoffice.gov.uk/research/nwp/publications/papers/technical_reports/2004.html*