

Constructing a summary skill score from scale-decomposition methods for monitoring operational NWP forecasts



Marion Mittermaier

1. Introduction

The horizontal resolution of limited-area forecast models today requires a verifying data set at resolutions finer than what most rain-gauge networks (or any point-measuring instrument) can provide. Therefore for precipitation, spatial estimates from weather radar can provide information regarding its distribution in **space and time**.

Recently-developed verification techniques aim to provide more informative feedback on some of the physical aspects of the forecast error, by verifying at different spatial scales. In doing so, there is the recognition that forecast phenomena occur on many scales and are driven by different physical processes.

2. The intensity-scale technique (Casati et al. 2004)

We assess **operational stand-alone deterministic forecasts** of six-hourly and daily accumulations from the 12km UK12, 12km NAE (North-Atlantic European) and the UK 4km Unified Models against the NIMROD baseline product which is the quality-controlled UK radar composite. The 5km radar data and UK4 model forecasts are averaged onto the 12km grid and, given the radar coverage, comparisons can only be made where radar data are available.

At the Met Office the method has been implemented to help quantify the potential benefits high-resolution forecasts have to offer for QPF (Mittermaier, 2006). Particular aspects include:

- no dithering step is performed because all values are floating point numbers;
- the normalisation occurs naturally when applying a factor-of-2 threshold series (rainfall is log-normally distributed);
- no forecast recalibration was performed as this was judged unnecessary (B.Casati, pers. comm.);
- the analysis is performed on a square 2^n by 2^n spatial domain.

$$I_a = \begin{cases} 1 & a > t \\ 0 & a \leq t \end{cases} \text{ and } I_f = \begin{cases} 1 & f > t \\ 0 & f \leq t \end{cases} \Rightarrow Z = I_f - I_a \Rightarrow Z = \sum_{l=1}^L Z_l \Rightarrow MSE = Z^2$$

A series of thresholds (t) is used to convert the analysis (a) (1) and forecast (f) (2) into binary images. The difference between the binary forecast and analysis defines the binary error (3). The binary error image can then be expressed as the sum of components on different spatial scales by performing a two-dimensional discrete Haar wavelet decomposition (4). The mean-squared error (MSE) of the binary error image is given by the average of all the differences over all the pixels in the domain (5).

A **binary skill score** can be calculated for each threshold, relative to a random forecast.

$$SS_t = \frac{MSE_t - MSE_{t,random}}{MSE_{t,best} - MSE_{t,random}} = 1 - L \frac{MSE_t}{MSE_{t,random}}$$

Negative values imply that the model is worse than a random forecast, in terms of the MSE (although this doesn't necessarily mean that the model has no skill). The skill score is contoured with threshold on the x-axis and the spatial scale on the y-axis. Only the negative scores are contoured. The output and interpretation will be illustrated with an example.

3. A flash flood event

Thunderstorms developing on 19 June 2005 resulted in heavy rainfall and flash flooding over northern England. Hourly rain gauge totals showed that the village of Hawby in N. Yorkshire reported 59.8mm in just one hour. **Figure 1** shows some of the devastation.



Figure 1: Some images of the de

Figure 2 shows the error analysis and 6h accumulations. The 4km model output from the 12Z run, averaged to the 12km grid shows rainfall of the observed intensity with quite good agreement in terms of the location of the most intense rainfall over Yorkshire. However, the area of intense rain extends further southwards with a considerable number of false alarms for thresholds exceeding 8mm in six hours over southern Wales. This is reflected by the cut-off feature on the intensity-scale diagram which suggests that the displacement length scale error is of the order of 100-200km.

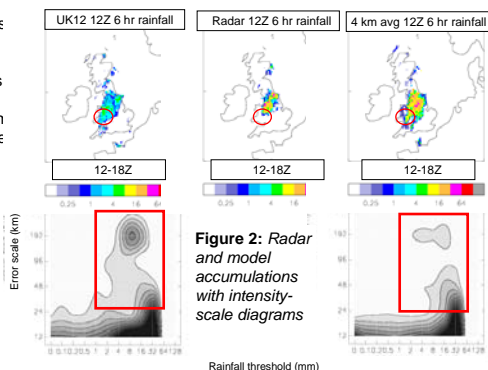


Figure 2: Radar and model accumulations with intensity-scale diagrams

4. A modified sign test statistic (Mittermaier, 2006)

We use a **distribution-free test** as normality of the errors can't be assumed.

B = number of +ve skill scores SS for a given scale and intensity during a given time interval, e.g. 1 month.

We then test the hypothesis:

- $H0$: $SS \geq 0$ (implicit positive and skillful)
- $H1$: $SS < 0$ (less skill than a random forecast)

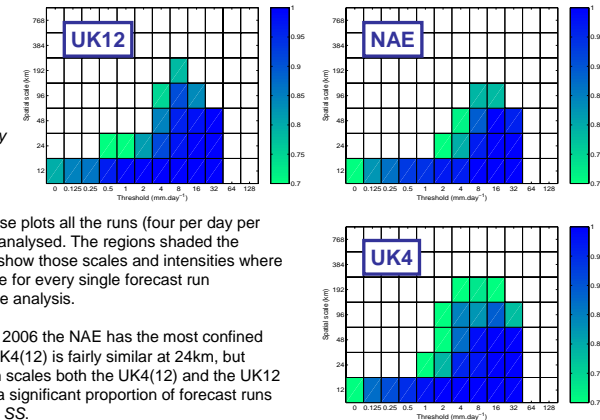
$H0$ is rejected if $b < b_{n,\alpha}$ where $B \sim b(n, 0.5)$ for small samples ($n < 40$), $\alpha = 0.025$

The value of $B' = (n - B) / n$ is shaded in the intensity-phase space for each scale and intensity where $H0$ is rejected.

5. Model inter-comparison of persistent errors

Using the **modified sign test statistic** the errors for different models can be compared at the monthly time scale. **Figure 3** shows the 'Manhattan' monthly error plots for **February 2006**. Three models are shown, the **UK12, NAE and UK4(12)**.

Figure 3: Error plots showing persistent errors at the monthly time scale



To create these plots all the runs (four per day per model) were analysed. The regions shaded the darkest blue show those scales and intensities where SS is negative for every single forecast run included in the analysis.

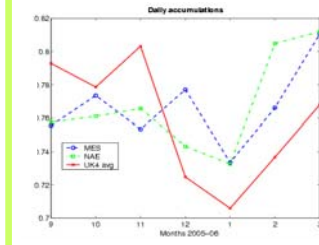
For February 2006 the NAE has the most confined errors. The UK4(12) is fairly similar at 24km, but at long length scales both the UK4(12) and the UK12 are showing a significant proportion of forecast runs with negative SS .

6. Constructing a skill score

There is always demand for a single ('simple') measure that can be tracked over time to monitor model performance. Here a sign test skill score (**STSS**), based on the **modified sign test statistic** B' has been constructed using a 'standard' method. All intensities and scales may be weighted differently using a weighting matrix w .

$$STSS = 1 - \frac{\sum_{t=1}^T \sum_{l=1}^L w_{t,l} B'_{t,l} - T \cdot L \cdot \sum_{t=1}^T \sum_{l=1}^L w_{t,l}}{0 - T \cdot L \cdot \sum_{t=1}^T \sum_{l=1}^L w_{t,l}}$$

In this case, all intensities and scales have been given equal weighting of 1. The score is bounded: the worst score is obtained if all the scales and intensities had a $B' = 1$ (errors present all the time), represented by the product TL . The best score is 0, i.e. when all B' are zeros. The final score is positively oriented.



We only have a relatively short time series for comparing all three models. **Figure 4** shows the **STSS** for daily accumulations between September 2005 and March 2006. It will take some time for the time series to reach a 'critical mass' before making a critical assessment and to compare with other verification measures and methods.

Figure 4: Sign Test Statistic Skill Score (STSS) for daily (0-24h) precipitation forecasts from three operational configurations

7. Concluding remarks

- The **4km model contains much more detail** (even when averaged to 12km)
- Detail does not necessarily equal accuracy. **Raw model output needs to be averaged**
- A **modified sign test statistic** can be used to **aggregate individual intensity-scale diagrams** to indicate persistent errors
- The monthly error ('Manhattan') plots can be converted into a **bulk skill score** for monitoring and comparing different models

References

Casati B., Ross G. and Stephenson D.B., 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Met. Apps.*, 11, 141-154.
Mittermaier M.P., 2006: Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *In print, Atmos. Sci. Let.*, April 2006.