At last, proof that higher spatial resolution precipitation forecasts are better

Marion Mittermaier, Nigel Roberts and Simon A Thompson
Outline

1. Introduction
2. Spatial verification methodology and Fractions Skill Score
3. Key findings from the NAE-UK4 long-term precipitation forecast assessment
4. The thorny issue of “what is truth”
5. Conclusions
Introduction
Does higher resolution give more skilful forecasts?

Apparently not! Has it all been a waste of time?

• April to Oct 2010

• Equitable Threat Score (ETS)

• Using Block 03 gauges

\[
ETS = \frac{\text{hits} - \text{random hits}}{\text{hits} + \text{false alarms} + \text{misses} - \text{random hits}}
\]
Has this been measured the right way?

*There are two main problems.*

1. **Double penalty effect**
   - Errors are counted as false alarms and misses.
   - Detail penalised, closeness not rewarded

2. **Unskilful scales**
   - Grid-scale detail should not be believed
   - Lorenz (1969) argued that the ability to resolve smaller scales would result in forecast errors growing more rapidly -> more noise
Spatial verification methodology
Compare fractional coverage over different sized areas

observed

forecast

Threshold exceeded where squares are blue

Fraction = 6/25 = 0.24  Fraction = 6/25 = 0.24

Courtesy of Nigel Roberts
The Fractions Skill Score (FSS) for comparing fractions with fractions

Roberts and Lean (2008), Roberts (2008), Mittermaier and Roberts (2010)

Mean square error for the fractions – variation on the Brier score

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

\[
0 \leq p_j < 1 \text{ forecast fractions}
\]

\[
0 \leq o_j \leq 1 \text{ radar fractions}
\]

\[
N \text{ number of points}
\]

Skill score for fractions/probabilities - Fractions Skill Score (FSS)

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

\[
\dagger \text{ indicates the square root}
\]

© Crown copyright   Met Office

Courtesy of Nigel Roberts
Range from 0 to 1 → 0 for zero skill, 1 for perfect skill

Typically increases with spatial scale (always for large sample)
Only asymptotes to 1 in the domain average limit if the forecast is unbiased or for frequency thresholds. Typically < 1 for physical thresholds.

Can define an ‘acceptable’ value of FSS which is halfway between random skill (FSS = observed frequency) and perfect skill (FSS=1)

In idealised experiments $FSS_{target}$ is reached at a scale that is twice the length of the spatial error in the forecast
Real examples

Case A - good forecast

Case B - poor forecast

Courtesy of Nigel Roberts
Comparing the UK4 and NAE

"An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts – for support rather than for illumination. "--After Andrew Lang
NAE-UK4 long term assessment

• 41 months of forecasts (~5000) assessed using radar accumulations.
• For time series consider 25 km neighbourhood size.
• Determine whether UK4 is statistically significantly better than NAE.
• Assess the use of radar composites as truth for long-term monitoring.
• Consider the use of frequency thresholds.
• Consider skill as a function of the diurnal cycle.

Thanks to Rob Darvell for help with VER stats files.
A short note on statistical significance …

• When comparing two models against the same truth the easiest way to test whether model A is better than model B is to test whether the difference in the scores is significant.

• The test statistic: \[ T = \frac{\bar{D}}{s_D / \sqrt{n}} \]

where \( \bar{D} \) is the mean of the differences in scores and \( s_D \) is the standard deviation.

• Test the null hypothesis that \( H_0: \mu_1 = \mu_2 \) where \( H_0 \) is rejected if \( t \leq t_{n-1,\alpha/2} \) or \( t \geq t_{n-1,\alpha/2} \).
FSS (neighbourhood size)

0.5 mm/6h

Median run-by-run score

FSS > 0.5 for all spatial scales

16 mm/6h

Median run-by-run score

Scores much lower
Require > 300 km neighbourhood to achieve skill

M Mittermaier, N Roberts & S Thompson
submitted to Met Apps
Diurnal cycle

- Higher resolution beneficial for diurnal cycle, especially triggering of afternoon convection.

- UK4 –NAE FSS always positive (better) but **bigger for larger thresholds.**

- For < 2 mm/6h score differences bigger for 18-00Z accumulations; > 4 mm/6h 12-18Z score differences biggest.
L(FSS>0.5) for 10% threshold and 0.5 mm/6h

The expectation is that through model improvements L(FSS>0.5) DECREASES over time..... or at least stays constant

10% threshold

Metric is impacted through the physical exceedance threshold applied at the grid scale.

0.5 mm/6h

From Mittermaier et al 2010
Concluding remarks
Interpretation of verification statistics

• Long-term monitoring requires a **stable baseline**.

• If there are changes in bias in both the forecast and the verifying observations it becomes difficult to attribute changes in the verification results to source.

• We **expect the model bias to change (improve!)** and have some understanding of the impact of model upgrade changes on the frequency bias through the trialling and parallel suites.

• This sort of information for changes made to radar processing is not widely known/accessible.
Key findings

• Based on 41 months of forecasts (~5000) 6-h UK4 precipitation forecasts are statistically significantly better than NAE at all lead times.

• Recommend that FSS or L(FSS>0.5) (the so-called “skilful spatial scale”) be used as metric for measuring precipitation forecast skill, but using frequency thresholds.

• Despite the use of frequency thresholds the lack of stability of a radar baseline could jeopardise the use of radar for long-term monitoring for precipitation forecast skill, except in a comparative sense.

• Frequency thresholds are preferred. They encompass the full range of precipitation and all rain is counted.
Thanks for listening!