Comparison of Fractional skill score and new HiRA verification

Historical trends and Experience in parallel trials

Clive Wilson, Marion Mittermaier, Lesley Allison

Wednesday, 5th October 2016. 38th EWGLAM and 23th SRNWP Meeting
Contents

• Historical record of Fractional skill score
• Basics of High resolution Assessment (HiRA)
  • Single Ob – neighbourhood forecast
• Effect of neighbourhood size
• Comparisons – regular operational monitoring
  • UKV & MOGREPS-UK
  • UKV & Global cf conventional verification
• Use in parallel trial PS38
  • Summer precipitation cf Fractional skill score
• Conclusions
Historical data

- NAE (12 km), UK4 and UKV (1.5 km) models spanning the period from April 2008.

- During this period either 5 km or 1 km (or both) radar rainfall fields were used for verification.

- Series have a 365-day running mean applied.

- Times of day are kept separate to consider the diurnal cycle.

- Lead times up to t+36h are considered.

- 6h precipitation are considered.

- Radar data has been of variable quality across the UK during the last decade through the radar renewal project. Hence the use of percentile thresholds (here the 90th or 95th) to avoid as much of any biases that may be introduced by the radar as possible. (can’t be fully eliminated)

From Mittermaier, in prep.
Precipitation: Fractional skill score @25km - historical

From Mittermaier, in prep.
Weather regime stratification

From Mittermaier, in prep.
The “regional” effect

- Scores higher than @5km (not shown)
  - “useful” skill if FSS>0.5
- Improving trend with model upgrades
  - Though detrimental PS35
- Differences due time of day

- Change in score with lead time not necessarily linear
- Strong variation in precipitation forecast skill with flow type
Spatial sampling

Make use of spatial verification methods which compare single observations to a forecast neighbourhood around the observation location. → SO-NF

Represents a fundamental departure from our current verification system strategy where the emphasis is on extracting the nearest GP or bilinear interpolation to get matched forecast-OB pair.

NOT upscaling/smoothing!

Only ~130 1.5 km grid points in >500 000 domain used to assess entire forecast! Note the variability in the neighbourhoods.
High Resolution Assessment framework*

• How to **consistently demonstrate skill** in increasingly higher-resolution models?

• Subjective assessment indicates skill, but the numbers don’t always say this.

• Rapid error growth, timing errors, misplaced detail, double penalty effect.

• **single-observation-neighbourhood-forecast** approach (SO-NF).

• Verifying at observing sites is **relevant to the user**.

SO-NF

- Verifying against observations

Traditional Interpolate/Average

HiRA e.g. for 3x3 grid

HiRA for 3x3 ensemble

Repeat for multiple neighbourhood sizes.
Create 3 scores from this data – Brier Score, RPS and CRPS (and associated skill scores)
High Resolution Assessment (HiRA) framework

- Use standard synoptic observations and a range of neighbourhood sizes
- Use 24h persisted observations as reference
- The method needs to be able to compare:
  - Deterministic vs deterministic (different resolutions, and test vs control of the same resolution)
  - Deterministic vs EPS
  - EPS vs EPS
- Test whether differences are statistically significant (Wilcoxon signed rank test)
- Grid scale calculated for reference → NOT main focus.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>RMSESS</td>
<td>MAE</td>
</tr>
<tr>
<td>Vector wind</td>
<td>RMSVESS</td>
<td>MAE</td>
</tr>
<tr>
<td>Cloud cover</td>
<td>ETS</td>
<td>PC</td>
</tr>
<tr>
<td>CBH</td>
<td>ETS</td>
<td>PC</td>
</tr>
<tr>
<td>Visibility</td>
<td>ETS</td>
<td>PC</td>
</tr>
<tr>
<td>1h precip</td>
<td>ETS</td>
<td>PC</td>
</tr>
</tbody>
</table>

RMS(V)ESS = Root Mean Square (Vector) Error Skill Score
ETS = Equitable Threat Score
BSS = Brier Skill Score
RPSS = Ranked Probability Skill Score
CRPSS = Continuous Ranked Probability Skill Score
MAE = Mean Absolute Error
PC = Proportion Correct

Mittermaier 2014, WAF.
Increasing neighbourhood size

More effective for UKV than MOGREPS-UK
Monthly median index
MOGREPS-UK, UKV
“equal pseudo-ensemble 11x11, 12x9”

Relative benefit MOGREPS-UK, UKV – “equal pseudo-ensemble”


Relative benefit of UK-EK over UK-UKV

Benefit of model2 over model1 = \frac{(C)RPSS(model2) - (C)RPSS(model1)}{1 - (C)RPSS(model1)}

Centred about UKV analysis
Relative benefit UKV over Global

Benefit of model2 over model1 = \frac{(C)RPSS(model2) - (C)RPSS(model1)}{1 - (C)RPSS(model1)}

Relative benefit UKV over Global

Relative benefit of UK-UKV over UK-GM

31  32  33  34  35  36  37
Old index- standard GP→Ob, RMSE, ETS skill scores

Relative Impact (%), Surface Obs, UKV - GM

PS34 – ENDGame, GA6 physics, 17km
**PS38 : extended domain**

**Improved Spin-up of Convection – 10/12/14**

---

**Operational UKV1**

**Extended Domain UKV2**

---

**Figure 1** Comparison showing improved representation of showery regime to the west of Ireland. Both models have cloudy boundary conditions in the west from which convection develops. UKV2 provides a better product for the UK given that the convection has more time to develop within it.
Parallel trials
1.5m temperature CRPSS
1-2% worse early forecast range

Surface (1.5m) Temperature (deg K)
Meaned between 20150616 03:00 and 20150715 21:00

Significance threshold: 0.05 (against relevant scale)

CRPSS vs persistence, 1 = Perfect, 0 = No Skill

Forecast range (hours)

up is good
Standard bias and RMSE

PS38 package slight advantage ~ -0.4%

Summer
10m wind RPSS
2.5% worse early forecast range

Surface (10m) Wind Speed (m/s)
Meaned between 20150616 03:00 and 20150715 21:00

HiRA Summer

up is good
Standard speed bias and RMSVE

PS38 package slight detriment ~ +0.4%

Summer
Cloud amount $\geq 0.8125$

Ceilometer Cloud Amount ($\geq 0.8125$)
Meaned between 20150616 03:00 and 20150715 21:00

HiRA Summer

up is good
Cloud amount $\geq 0.8$, Standard ETS and frequency bias

PS38 package slight detriment to T+18, better T+24 onwards

Ceilometer cloud amount(0.8): LNDSYN Obs (unknown sub-category)
WMO Block 03 station list
Equalized and Meaned from 16/6/2015 00Z to 15/7/2015 23Z

Summer
1hr Precipitation Accumulation (mm) (>= 1mm) Measured between 20150618 03:00 and 20150715 21:00

Significance threshold: 0.05 (against relevant scale)

BSS vs persistence, 1=Perfect, 0=No Skill

Difference (+ve = Testbetter)
PPN ≥4mm/6h Standard ETS and frequency bias

PS38 package ETS neutral to T+18; slight detriment to bias to T+24, different impact to HiRA

Summer
1 hr Ppn Accumulation FSS: T+6, T+12, T+24

Disagrees with HiRA

Summer
MOGREPS-UK
CRPS 1.5m temperature

Continuous ranked probability score, T 1.5 m

Summer

Winter

down is good
MOGREPS-UK
RPS precipitation (hourly)

Ranked probability score, hourly precip.

Summer

Winter

down is good

good
Conclusions

• FSS monitor long term trends
• HiRA
  • Unified approach deterministic/ensemble
  • Show ensemble benefit
  • Show high resolution benefit (convective scale)
• Parallel trials use:
  • Some discrepancies with conventional GP-Ob
  • Summer (convective) disagreement with FSS
    • HiRA – smaller scale verification
Grazie Questions?
Extra slides
Compare fractional coverage over different sized areas

Threshold exceeded where squares are blue

Fraction = 6/25 = 0.24
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

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

\[
FBS \quad (\text{Fractions Brier Score})
\]

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

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

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

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

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

© Crown copyright   Met Office

Courtesy of Nigel Roberts
Winter HiRA

1hr Precipitation Accumulation (mm) (>= 1mm)
Meaned between 20150112 03:00 and 20150305 21:00

Significance threshold: 0.05 (against relevant scale)

BSS vs persistence, 1=Perfect, 0=No Skill

Forecast range (hours)

Control 1.5km
Control 4.5km
Control 10.5km
Control 16.5km
Test 1.5km
Test 4.5km
Test 10.5km
Test 16.5km

(Test)-(Control) 1.5km
(Test)-(Control) 4.5km
(Test)-(Control) 10.5km
(Test)-(Control) 16.5km

Difference (+ve = Test better)
PPN $\geq 4\text{mm/6h}$ Standard ETS and frequency bias

PS38 package ETS better to $T+18$; slight detriment to bias to $T+24$, similar impact to HiRA to $T+18$
1 hr Ppn Accumulation FSS: T+6, T+12, T+24

Winter
MOGREPS-UK
RPS 10m wind speed

Summer                                      Winter

Ranked probability score, wind 10 m

- contr
- cons
- stretch

down is good