Demonstrating a strategy for verifying km-scale NWP forecasts at observing sites

Marion Mittermaier
Background

• Traditional metrics can be misleading → trust in objective results undermined, especially for testing model changes.

• Representativeness of observations and model grid values → implications for highly localised events.

• Lack of predictability and rapid error growth at km-scale → impact on perceived skill.

• Difference between grid scale ($\Delta x$) and model resolution ($y \cdot \Delta x$, typically $y \geq 4$) → now even more reason that km-scale model forecasts must be treated differently (probabilistically) for product generation and verification.
The double penalty

CSI = 0 for first 4;
CSI > 0 for the 5th

Closeness not rewarded
Detail is penalised unless exactly correct
- higher resolution is more detailed!
Grid scale – Low skill/saturated

Domain scale - maintained skill

Intermediate scales – decreasing skill

Large synoptic

Predictable

Uncertain

Noise

Individual showers

From Roberts 2008
Small uncertainty at large scales = large uncertainty at small scales

Justifies the use of a downscaling ensemble (MOGREPS-UK)

5% error at 1000 km = 100% error at 50 km

Link to larger scale:
Russell et al. 2008
Hanley et al. 2011, 2012
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.

 лишь ~130 1.5 km grid points in >500 000 domain used to assess entire forecast!

Note the variability in the neighbourhoods.
Relative framework

- **Model A**
  - det or ens
  - Improvement in skill over a skillful alternative?

- **Model B**
  - det or ens
  - Improvement in skill over a skillful alternative?

- **Persistence**
  - 24h ob
  - Benefit of change?
Framework outline

- 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) [“s” denotes significant at 5%]
  - 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 → MAE</td>
<td></td>
</tr>
<tr>
<td>Vector wind</td>
<td>RMSVESS → MAE</td>
<td></td>
</tr>
<tr>
<td>(wind speed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloud cover</td>
<td>ETS → PC</td>
<td></td>
</tr>
<tr>
<td>CBH</td>
<td>ETS → PC</td>
<td></td>
</tr>
<tr>
<td>Visibility</td>
<td>ETS → PC</td>
<td></td>
</tr>
<tr>
<td>1h precip</td>
<td>ETS → PC</td>
<td></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
Comparisons:
1 GP with 12 single ensemble GPs
or
9 GP with 12 * 9 ensemble GPs → enhanced sampling

Approach

- Deterministic forecast with/without neighbourhood
- Ensemble members with/without neighbourhoods

2.2 km MOGREPS-UK ensemble
Three scenarios …
Deterministic vs deterministic (different resolution)

Cumulative 36h benefit, December 2011 03Z

- **+ve** = UKV test better than UK4
- **"none"** = 1 nearest GP UKV vs 1 nearest GP UK4

Quasi-ensemble approach beneficial even for T and wind

Counteracting GS biases

**Temperature**

**10 wind speed**

**Total cloud amount**

**Cloud base height**

**Visibility**

**Hourly Precipitation**

© Crown copyright 2013   Met Office
Deterministic test vs control (model trialling)

Detect small changes

Reduces to proportion correct at grid-scale

+ve = UKV test better than control

“none” = 1 nearest GP UKV vs 1 nearest GP UKV

Temperature

10 m wind speed

Total cloud amount

Visibility

Hourly precipitation

UKV @ 1.5 km

PS31 September 2012 03Z
Deterministic vs EPS

+ve = MOGREPS-UK ensemble better
“none” = 12 nearest GP values MOGREPS-UK vs 1 nearest GP UKV

Benefit of ensemble

Cloud base height

Temperature

10 m wind speed

Total cloud amount

Visibility

Neighbourhood greater benefit for UKV

Benefit of ensemble

1st 5 weeks of 03Z MOGREPS-UK
Skill against persistence

Temperature @~4 km

Wind speed @~4 km

Cloud cover @~4 km

CBH @~4 km

Visibility @~4 km

1h precipitation @~4 km

MOGREPS-UK @ 2.2 km
UKV @ 1.5 km
A few words on reliability ....
Visibility and cloud base height (CBH) for 3 months JFM 2013

Sampling not enough events

Reliability is not that different between UKV and MOGREPS-UK even though the number of grid-points used is 9x greater

Probabilities can be made reliable

Over-confident

Evidence of some skill for selected probability thresholds
Conclusions

- Method aims to provide **objective reflection of inherent skill within a forecast neighbourhood** in the vicinity of an observing site in a quasi-probabilistic way.

- Method **cannot fabricate “skill”** where there is none. Model deficiencies are clearly highlighted.

- Method **appears robust** for all three scenarios tested → key requirement for Met Office Unified Model R & D.

- Results point the way for **post-processing km-scale NWP output** to maximise skill of forecast products.
Conclusions (cont.)

• New verification framework illustrates benefit of km-scale ensemble over deterministic.

• Bigger neighbourhoods will improve forecast skill (for the most part) but the UKV needs (and benefits more from) neighbourhood processing, i.e. better “harvesting” of information content.
Questions?