Using time-lag ensemble techniques to assess behaviour of high-resolution precipitation forecasts

Marion Mittermaier

3rd Int’l Verification Methods Workshop, ECMWF, 31/01/2007
Outline

Introduction

“Inconsistency” or spread ↔ Is this a consequence of letting the dynamics loose?

Constructing a lagged ensemble

How can I use this as a diagnostic tool?

Summary
UK4 time-lag ensemble: TotalPrecipitation6hr mm

VT 12 – 18Z on 10/05/2006

The fickleness of triggering convection
- **UK 4 km Unified Model** run 4 times a day with **3DVAR** at 03, 09, 15 and 21Z producing a 36h forecast.

- Inclusion of **latent heat nudging (LHN)** from radar analysis.

- **Convection is resolved dynamically.** Parameterisation scheme for shallow (non-precipitating) convection only.

- Produce **6h precipitation accumulations** for 00-06Z, 06-12Z, 12-18Z and 18-00Z from t+3h to t+33h forecasts.
## Constructing a time-lagged ensemble

### Forecast range

<table>
<thead>
<tr>
<th></th>
<th>D+1 06–12Z</th>
<th>D+1 12–18Z</th>
<th>D+1 18–00Z</th>
<th>D+2 00–06Z</th>
<th>D+2 06–12Z</th>
<th>D+2 12–18Z</th>
<th>D+2 18–00Z</th>
<th>D+3 00–06Z</th>
<th>D+3 06–12Z</th>
<th>D+3 12–18Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>03Z</td>
<td>3–9h</td>
<td>9–15h</td>
<td>15–21h</td>
<td>21–27h</td>
<td>27–33h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09Z</td>
<td>3–9h</td>
<td>9–15h</td>
<td>15–21h</td>
<td>21–27h</td>
<td>27–33h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15Z</td>
<td>3–9h</td>
<td>9–15h</td>
<td>15–21h</td>
<td>21–27h</td>
<td>27–33h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21Z</td>
<td>3–9h</td>
<td>9–15h</td>
<td>15–21h</td>
<td>21–27h</td>
<td>27–33h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>03Z</td>
<td>3–9h</td>
<td>9–15h</td>
<td>15–21h</td>
<td>21–27h</td>
<td>27–33h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09Z</td>
<td>3–9h</td>
<td>9–15h</td>
<td>15–21h</td>
<td>21–27h</td>
<td>27–33h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15Z</td>
<td>3–9h</td>
<td>9–15h</td>
<td>15–21h</td>
<td>21–27h</td>
<td>27–33h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 5-member ensemble of 6-hour accumulation forecasts for the most recent interval
Only small overlap between forecasts.

Ensemble mean gives indication of area and totals in excess of 8 mm/6h.

Ensemble max is better at capturing potentially large totals.
Truth

PoP $\geq 4$ mm/6h from ensemble at grid scale (4 km)

Latest forecast (lag 0) PoP $\geq 4$ mm/6h at grid scale (4 km)

PoP $\geq 4$ mm/6h from ensemble averaged to 12 km (3x grid scale)
Roberts and Lean (2006)

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

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

is a version of the Brier score in which fractions are compared with fractions

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

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

0 < p_j < 1 forecast fraction
0 < o_j < 1 radar fraction
FSS results for May 2006

- Skill as function of scale at a given threshold.
- Investigate individual contributions as a function of forecast range/age.
- Is the sum of the parts greater than the parts?

![Graphs showing FSS results for different thresholds and scales.](image-url)
Good skill for ensemble forecasts

Better discretisation of forecast skill
Mean ROC and ROC area

Very similar
How to discriminate?

Compare mean monthly performance
ROC$a$ differ only by $\sim 0.01$

Discrimination distance $d'$ better benchmark?

Monthly ROC plot at 4 mm/6h

Feb 2005: ROC$a = 0.93; d' = 2.12$
May 2006: ROC$a = 0.94; d' = 1.91$
Jul 2006: ROC$a = 0.92; d' = 1.78$
Oct 2006: ROC$a = 0.91; d' = 1.77$
• Some improvement in bias with spatial averaging at lower thresholds. Less clear that spatial averaging has any benefits at higher thresholds.

• Ensemble forecasts are biased BUT we know they have skill (ROC) so they can be recalibrated.
Recalibrated ...

Bias

False alarms

Log odds ratio

Misses

... interpret only at recalibration scale
Correlation

Successive forecasts are consistent overall:
There is sufficient spread to combine them.

- Not sensitive to bias.
- Captures pattern.
- How similar are forecasts verifying at the same time?
- How swiftly do they decorrelate?
- How valid is a lagged ensemble technique?

Forecasts are quite consistent
Skill vs Obs decreases slightly
In summary

- High-resolution forecasts are more variable as the detail that can be resolved is less predictable. This is why it has been difficult to demonstrate increased skill (especially for precipitation).

- An ensemble approach may maximise forecast skill (and value). It provides the forecaster with information on variability, i.e. how confident he/she should be about model guidance.

- The discrimination distance $d'$ is a better parameter for assessing how good an ensemble forecast really is, due to a clearer discretisation.

- On average forecasts six hours apart do decrease in skill but over 36 hours and for higher thresholds a monotonic trend is not always present.

- Combining forecasts of different ages is another way of accounting for forecast uncertainties, with the possibility of retaining many smaller-scale features that would be lost if a spatial averaging technique were used.

- For higher thresholds it is clear that spatial averaging may be detrimental to forecast skill. Therefore an optimal (but varying) averaging length may exist.
Questions & Answers