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?

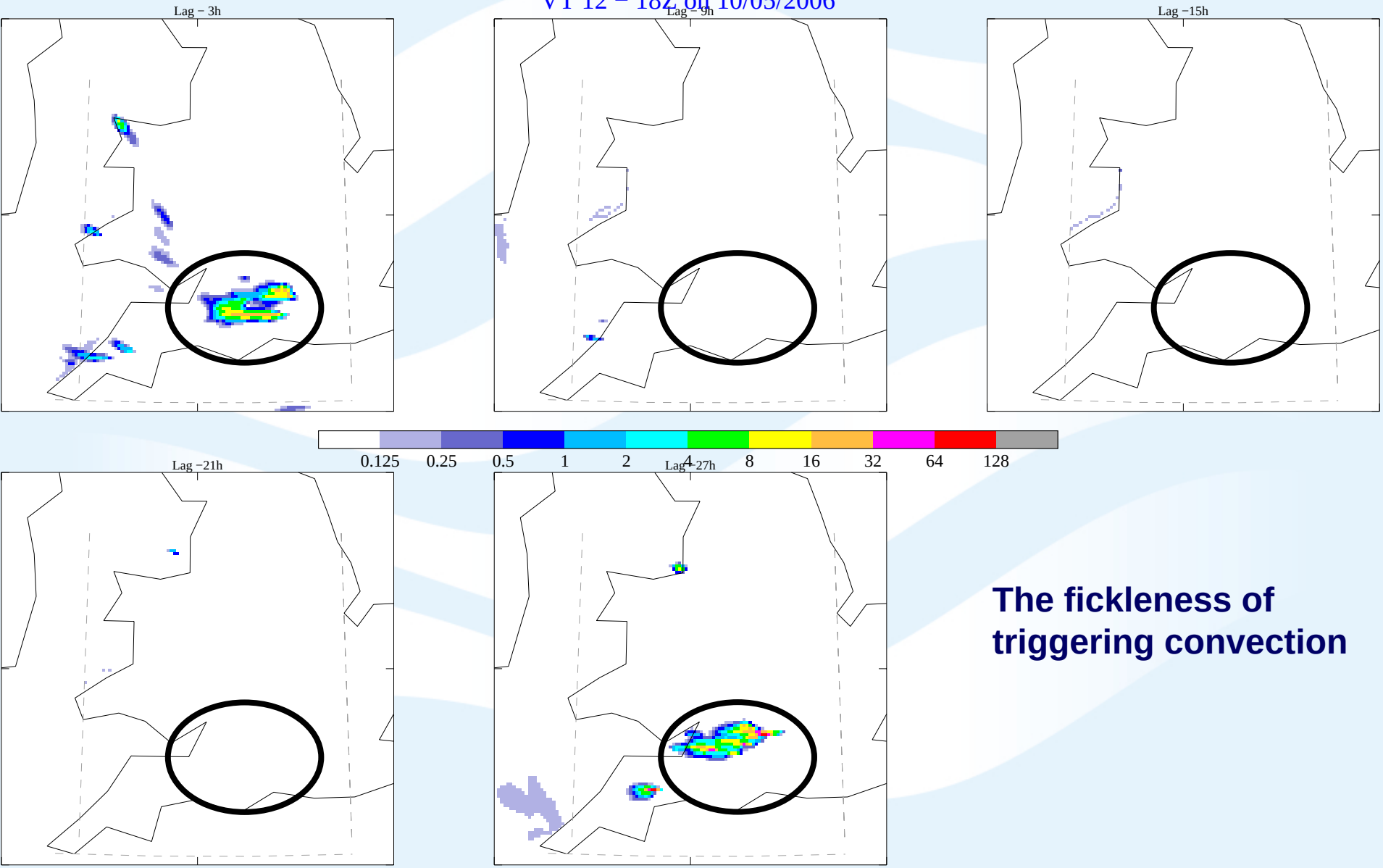
Constructing a lagged ensemble

How can I use this as a diagnostic tool?

Summary

UK4 time−lag ensemble: TotalPrecipitation6hr mm

VT 12 − 18Z on 10/05/2006

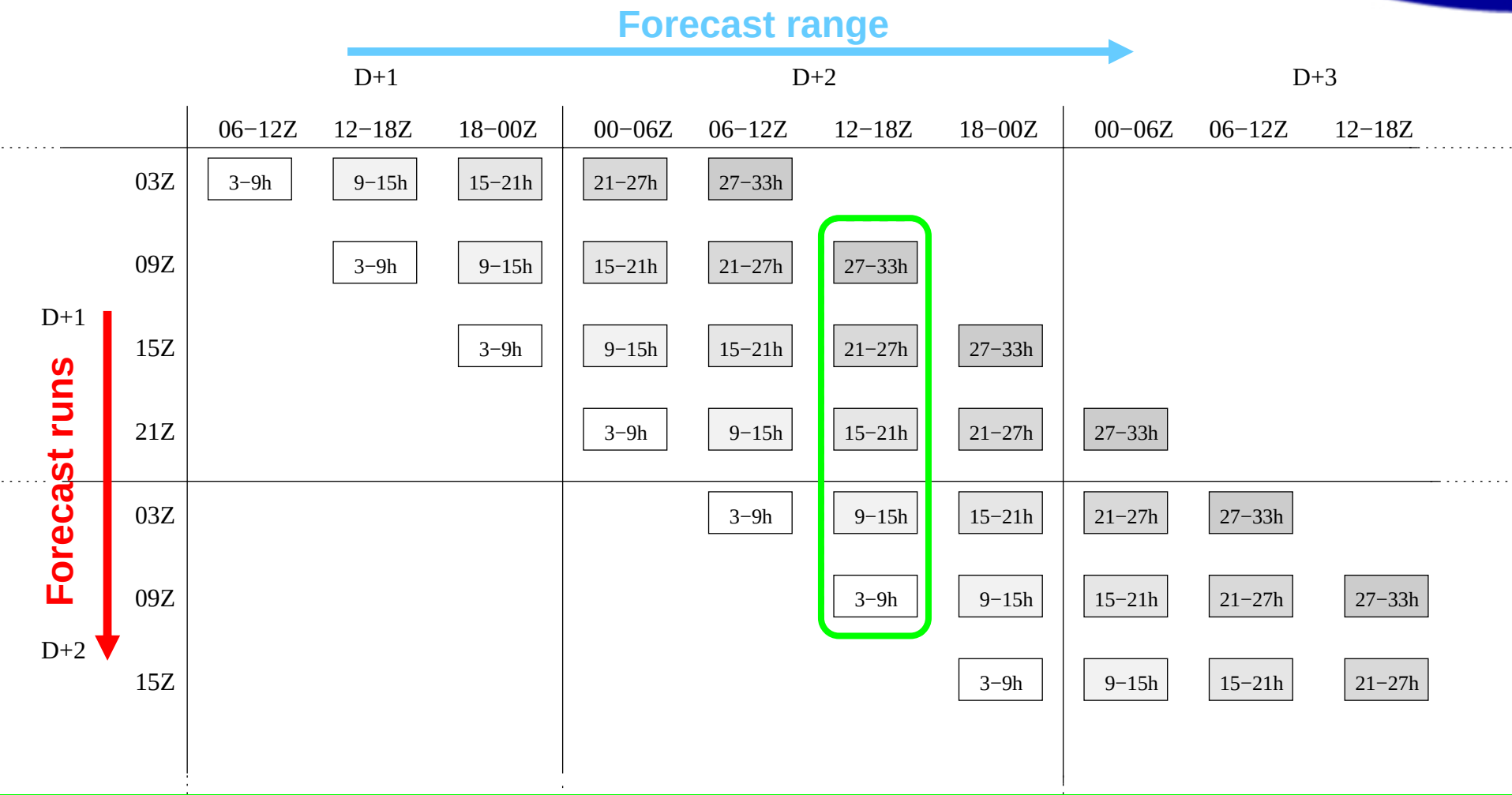


The fickleness of triggering convection

Model description

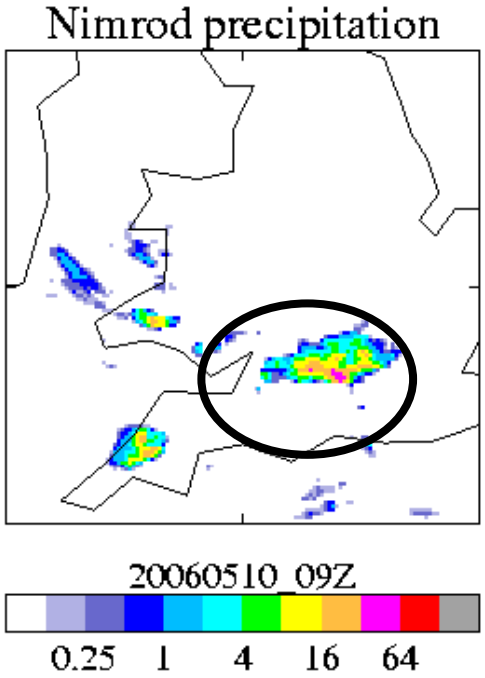
- **UK 4 km Unified Model** run 4 times a day with **3DVAR** at 03, 09, 15 and 21Z producing a 36h forecast.
- Inclusion of **latent heat nudging (LHN)** from radar analysis.
- **Convection is resolved dynamically**. Parameterisation scheme for shallow (non-precipitating) convection only.
- Produce **6h precipitation accumulations** for 00-06Z, 06-12Z, 12-18Z and 18-00Z from t+3h to t+33h forecasts.

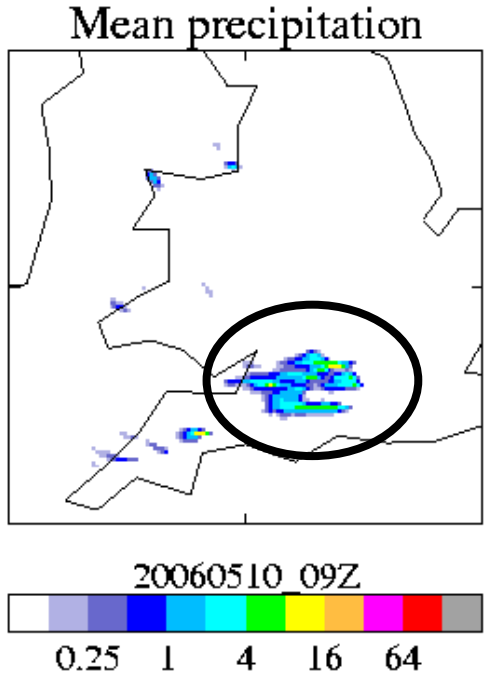
Constructing a time-lagged ensemble



5-member ensemble of 6-hour accumulation forecasts for the most recent interval

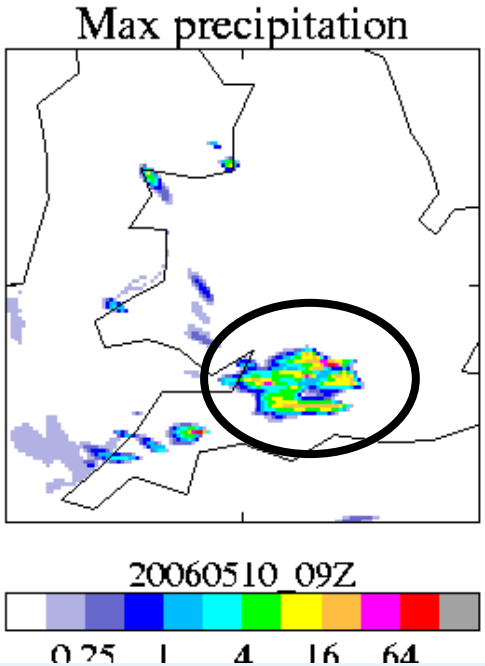
Truth

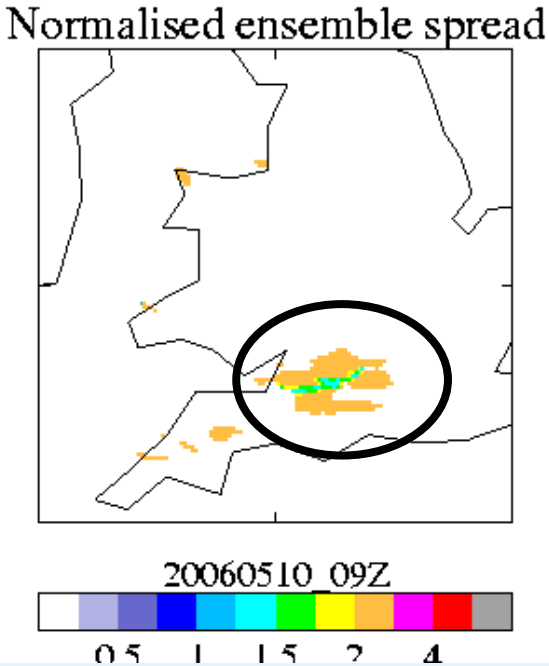




Ensemble mean gives indication of area and totals in excess of 8 mm/6h.

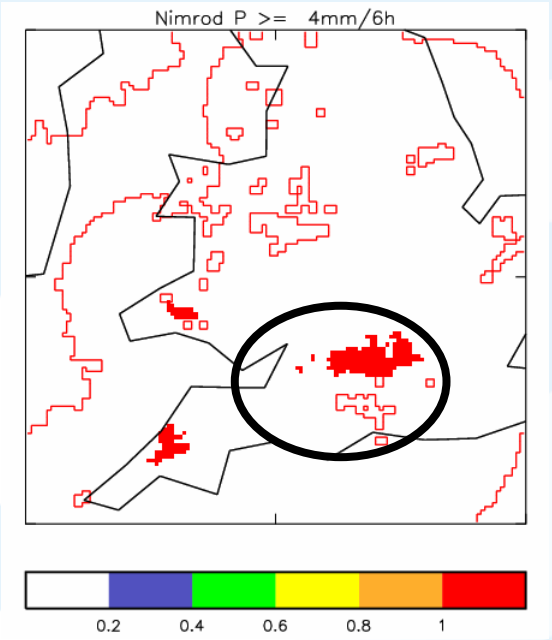
Ensemble max is better at capturing potentially large totals.

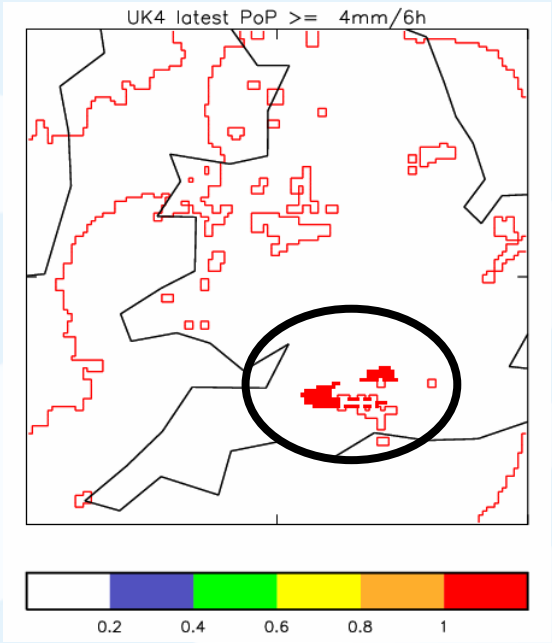




Only small overlap between forecasts

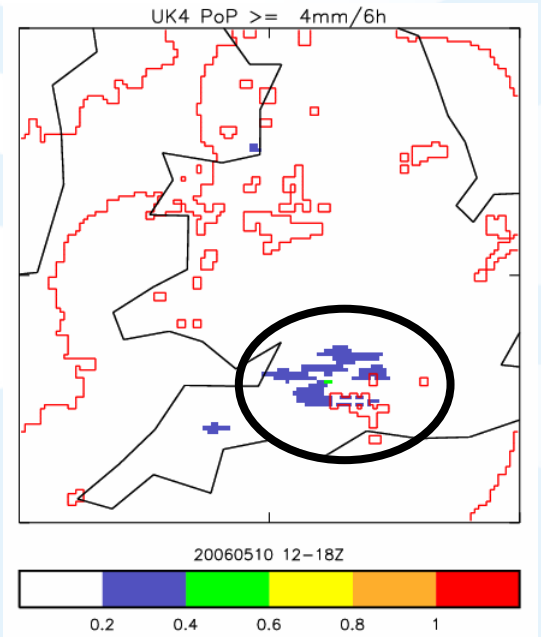
Truth

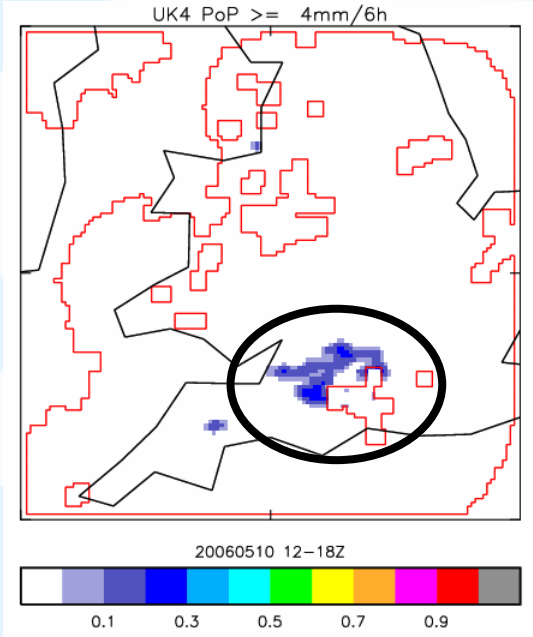




Latest forecast (lag 0) PoP >= 4 mm/6h at grid scale (4 km)

PoP >= 4 mm/6h from ensemble at grid scale (4 km)





PoP >= 4 mm/6h from ensemble averaged to 12 km (3x grid scale)

Fractions Skill Score

Roberts and Lean (2006)

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

$0 \leqslant p_j \leqslant 1$ forecast fraction

$0 \leqslant o_j \leqslant 1$ radar fraction

$$\underset{\text{(Fractions Brier Score)}}{FBS} = \frac{1}{N}\sum_{j=1}^{N}(p_j - o_j)^2$$

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

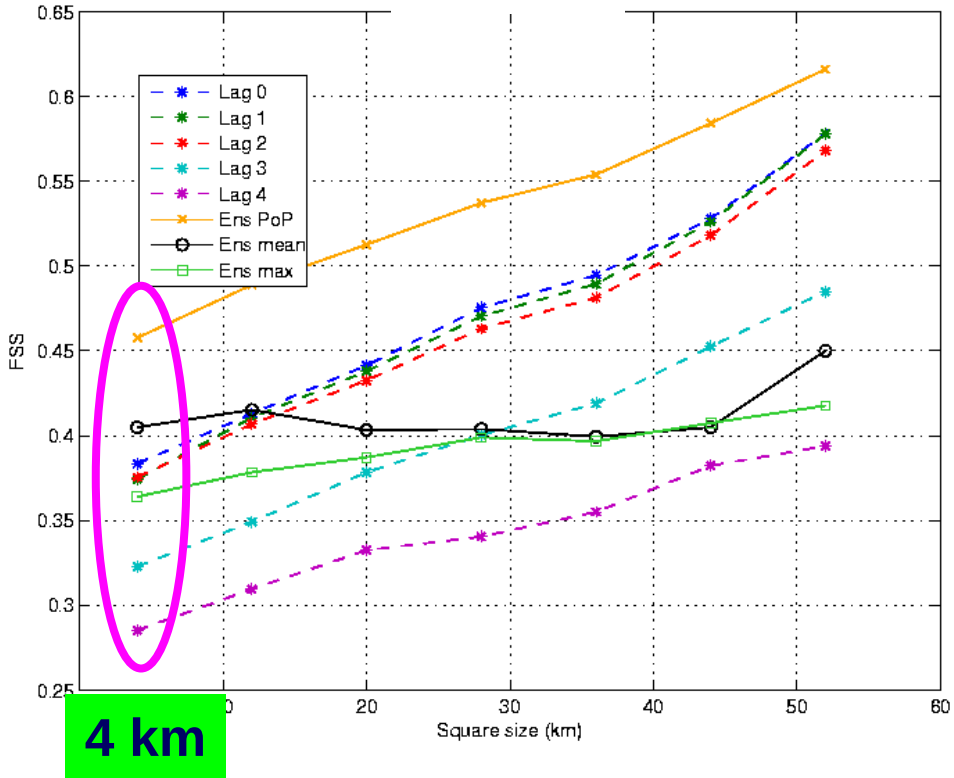
$$\frac{1}{N}\left[\sum_{j=1}^{N}(p_j)^2 + \sum_{j=1}^{N}(o_j)^2\right]$$

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

FSS results for May 2006

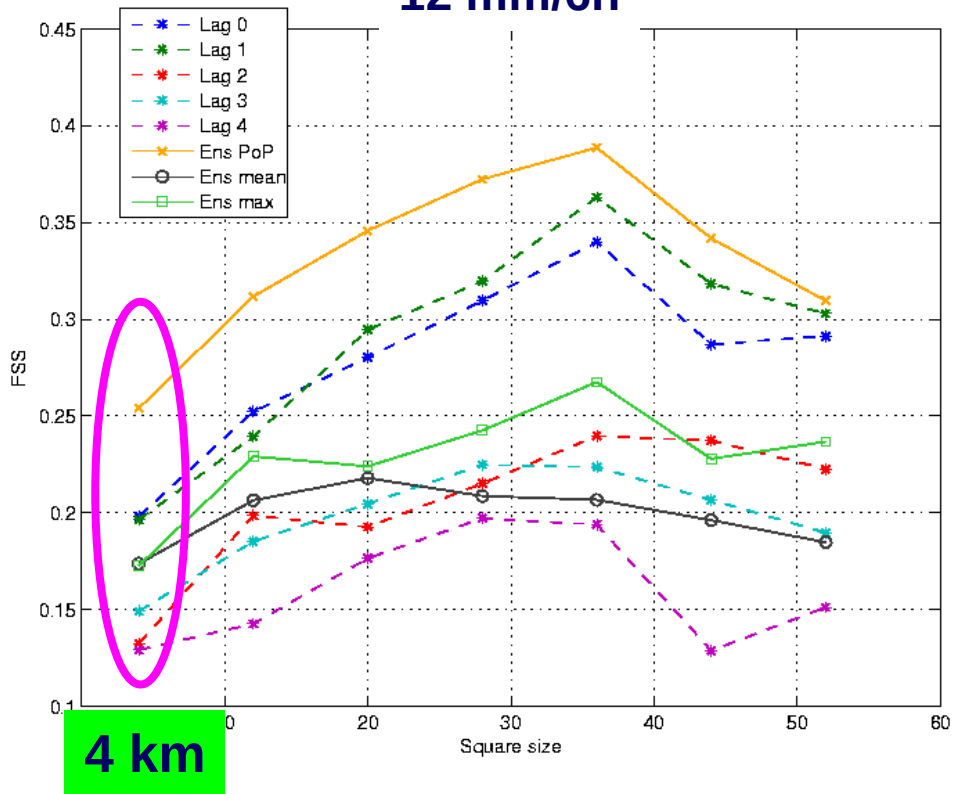
- Skill as function of scale at a given threshold.
- Investigate individual contributions as a function of forecast range/age.
- **Is the sum of the parts greater than the parts?**

4 mm/6h



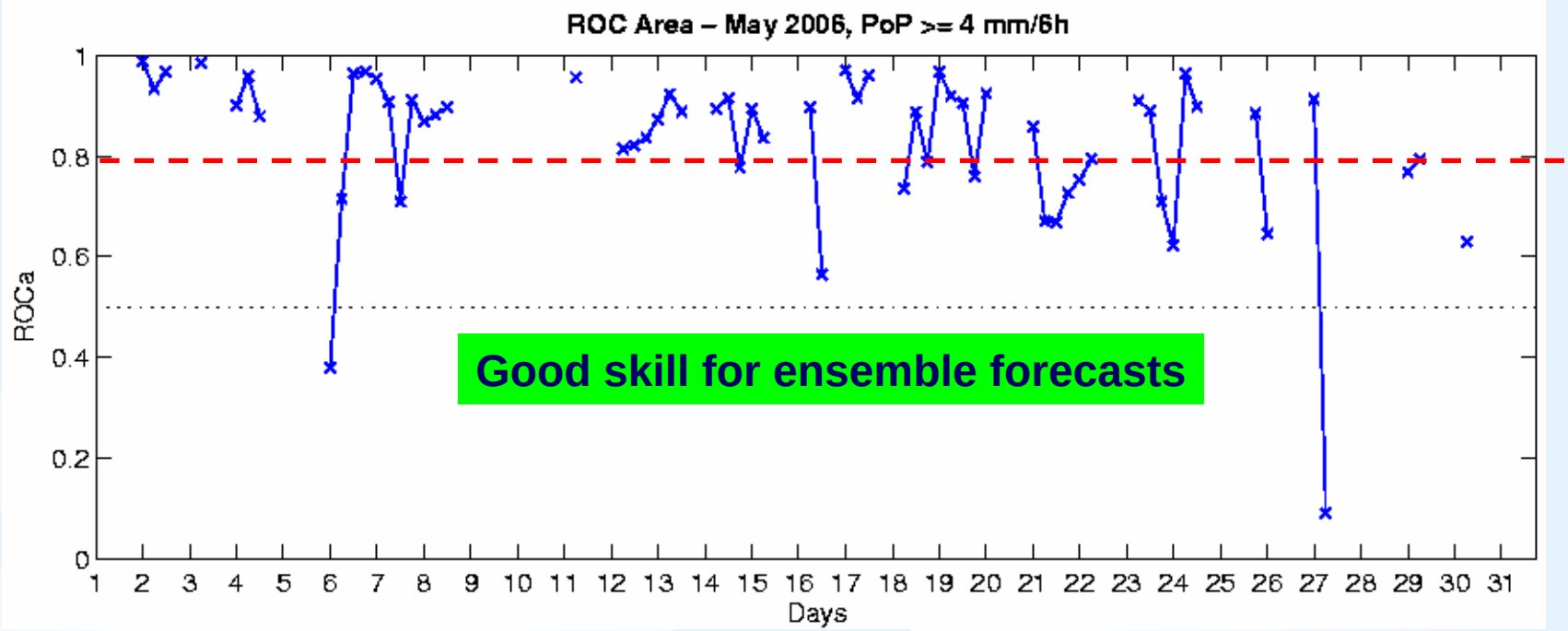
4 km

12 mm/6h

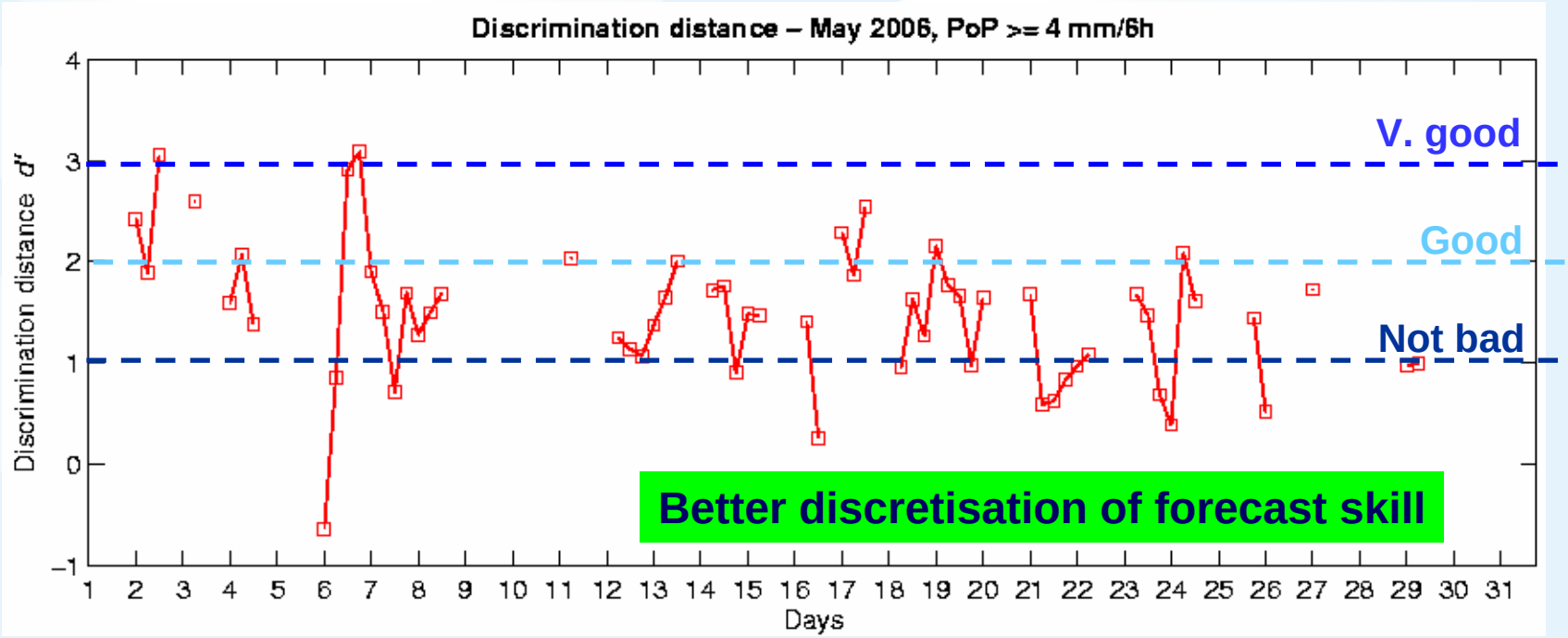


4 km

ROCa

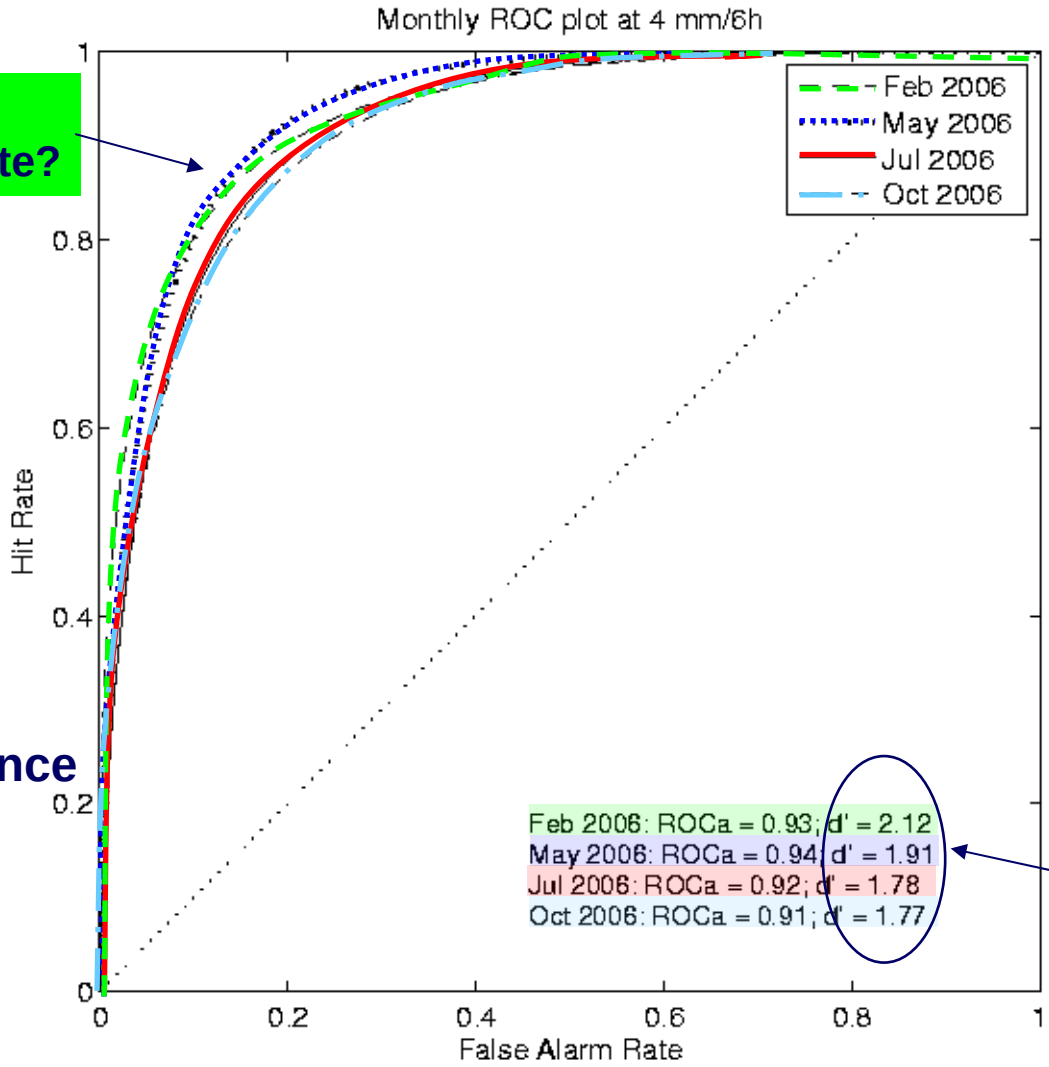

ROC Area – May 2006, PoP >= 4 mm/6h
ROCa
Days
Good skill for ensemble forecasts

d'


Discrimination distance – May 2006, PoP >= 4 mm/6h
Discrimination distance d'
Days
V. good
Good
Not bad
Better discretisation of forecast skill

Mean ROC and ROC area

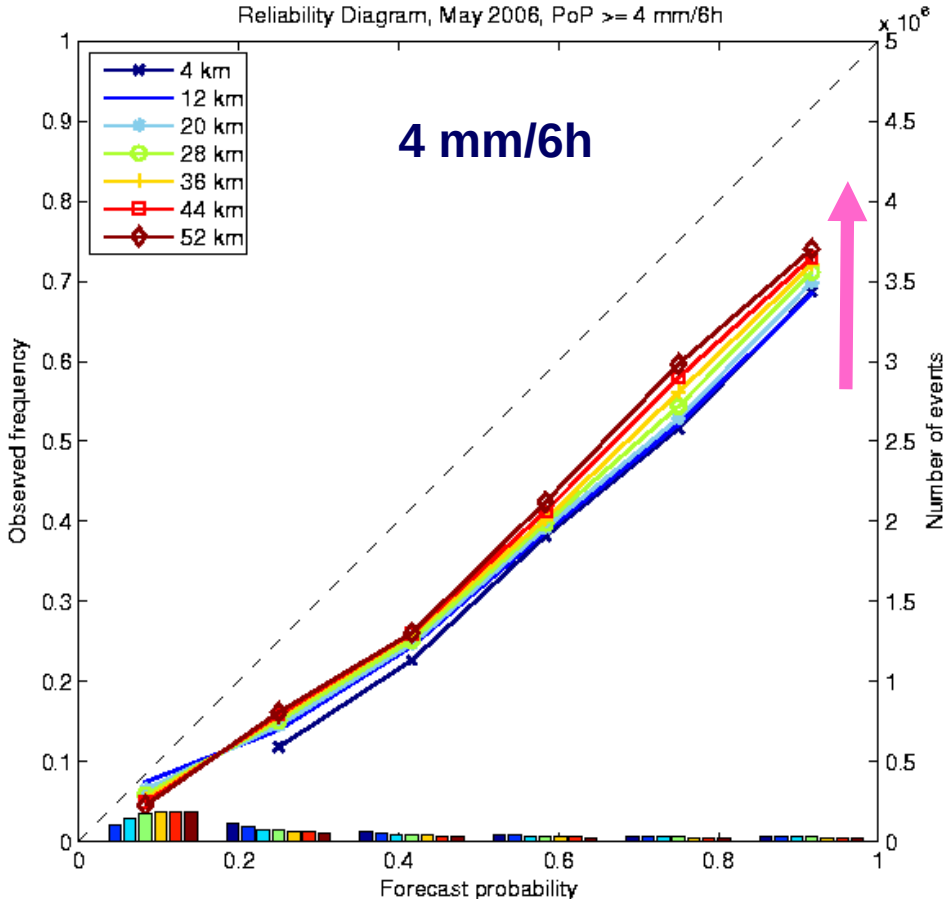
Very similar
How to discriminate?

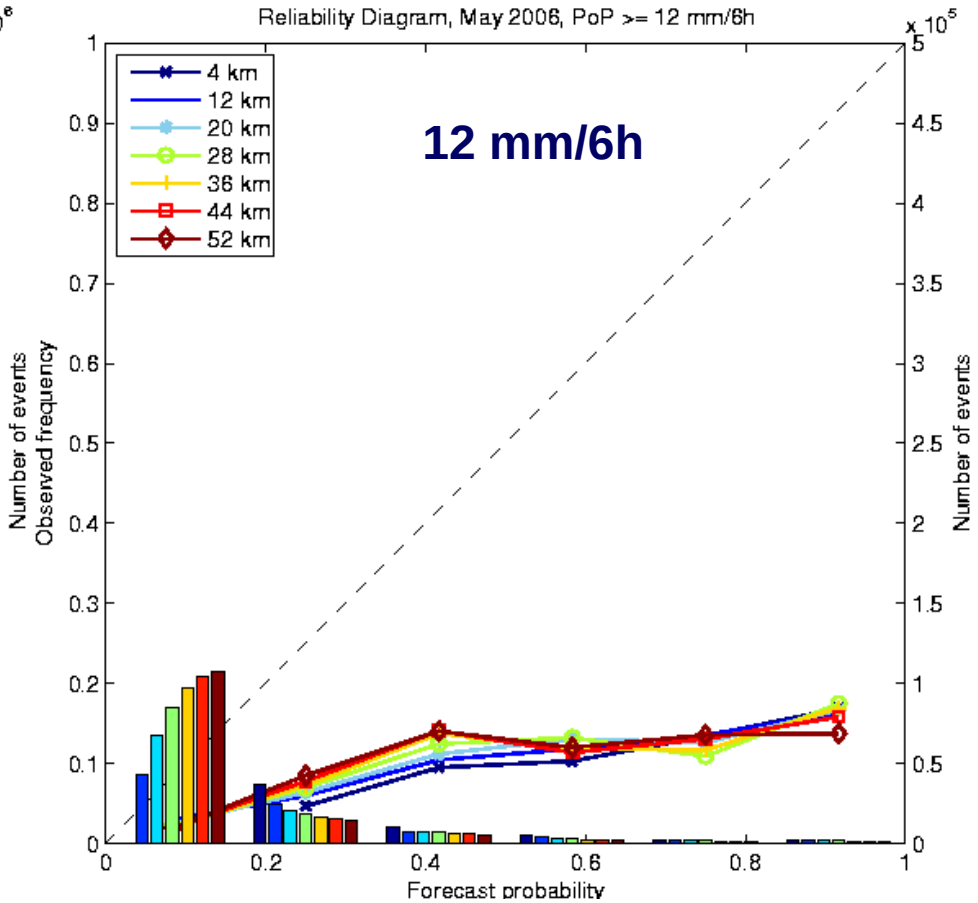


Compare mean monthly performance ROCa differ only by ~0.01

Discrimination distance *d'* better benchmark?

Reliability diagrams



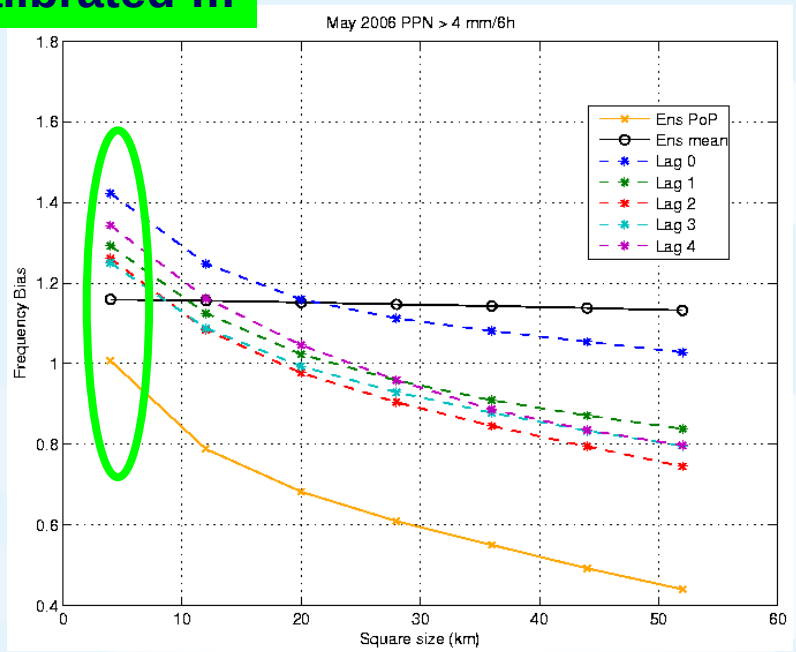


- Some improvement in bias with spatial averaging at lower thresholds. Less clear that spatial averaging has any benefits at higher thresholds.

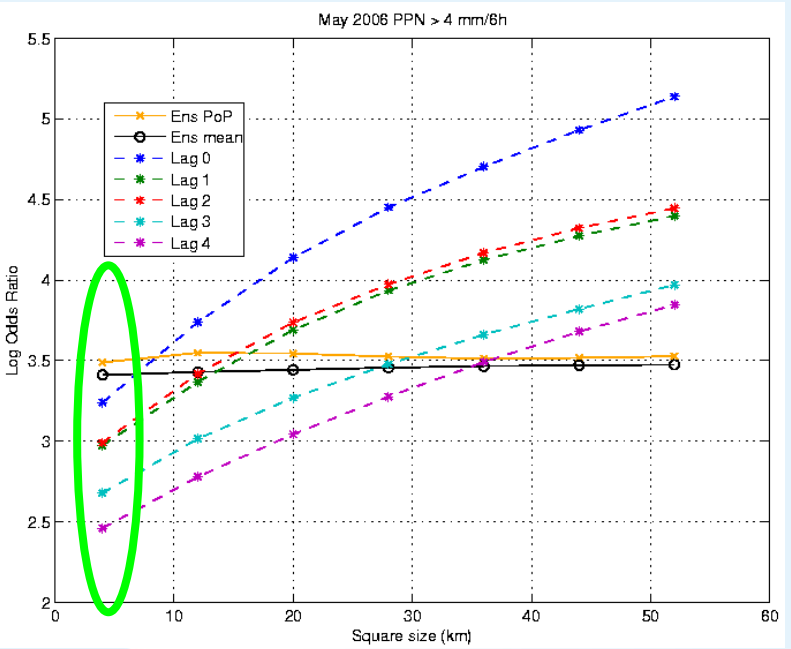
- Ensemble forecasts are biased BUT we know they have skill (ROC) so they can be recalibrated.

Recalibrated …

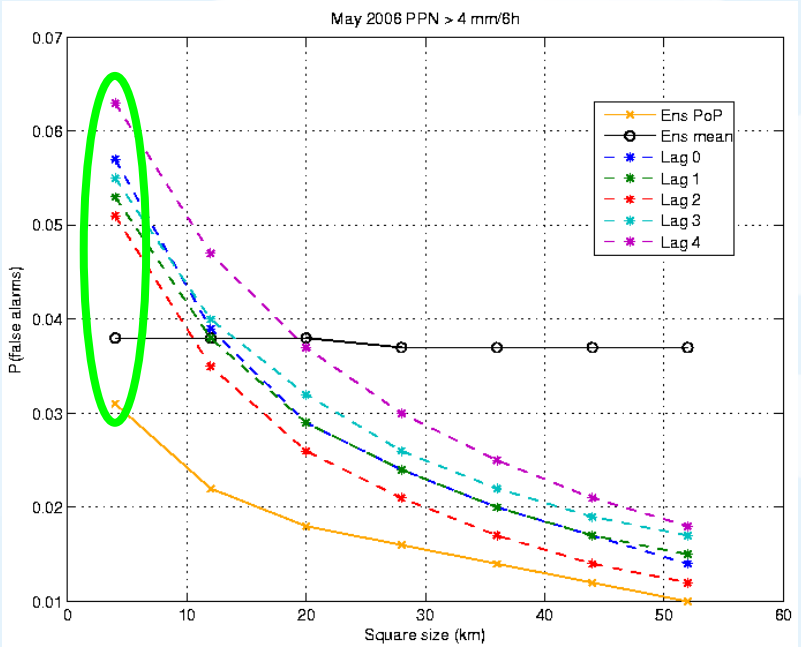
Bias



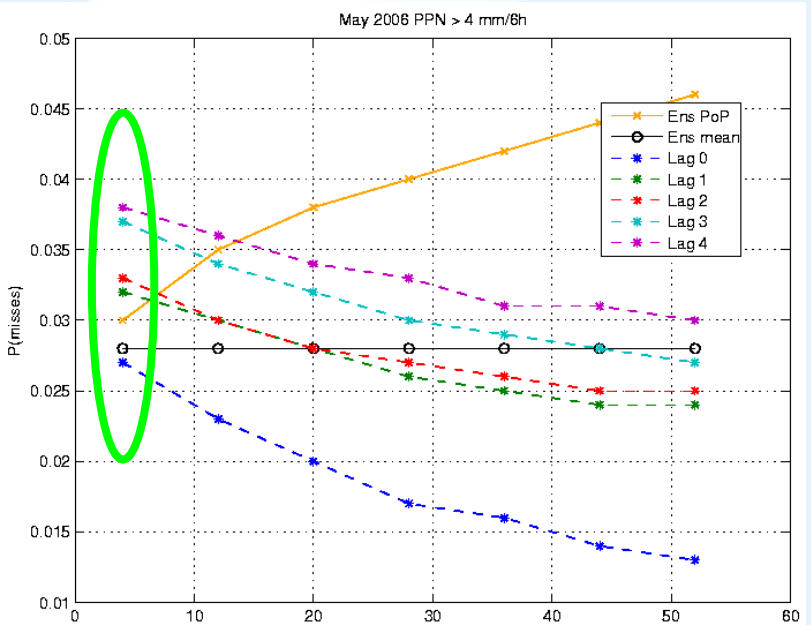
Log odds ratio



False alarms

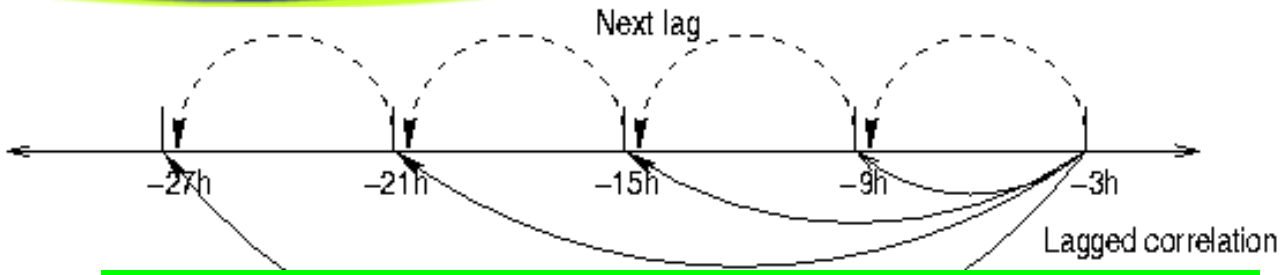


Misses



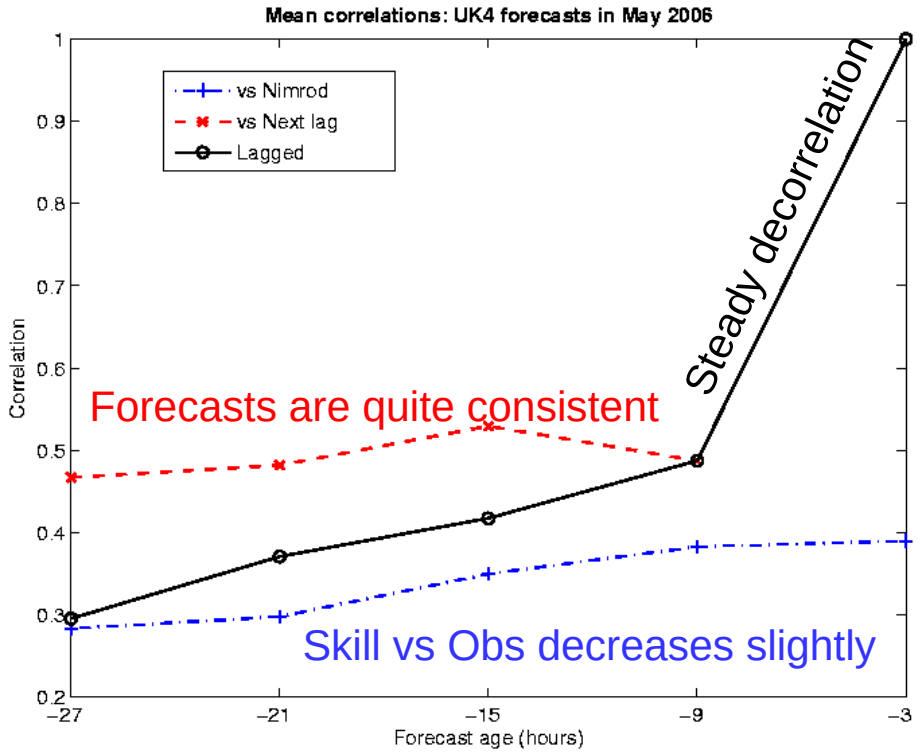
… interpret only at recalibration scale

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?



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

Mittermaier, M.P., 2006: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles.
Submitted to QJ, 15 December 2006