Met Office

Are higher spatial resolution precipitation forecasts better ? - can we show it ?

Marion Mittermaier, Nigel Roberts and Simon A Thompson

Outline

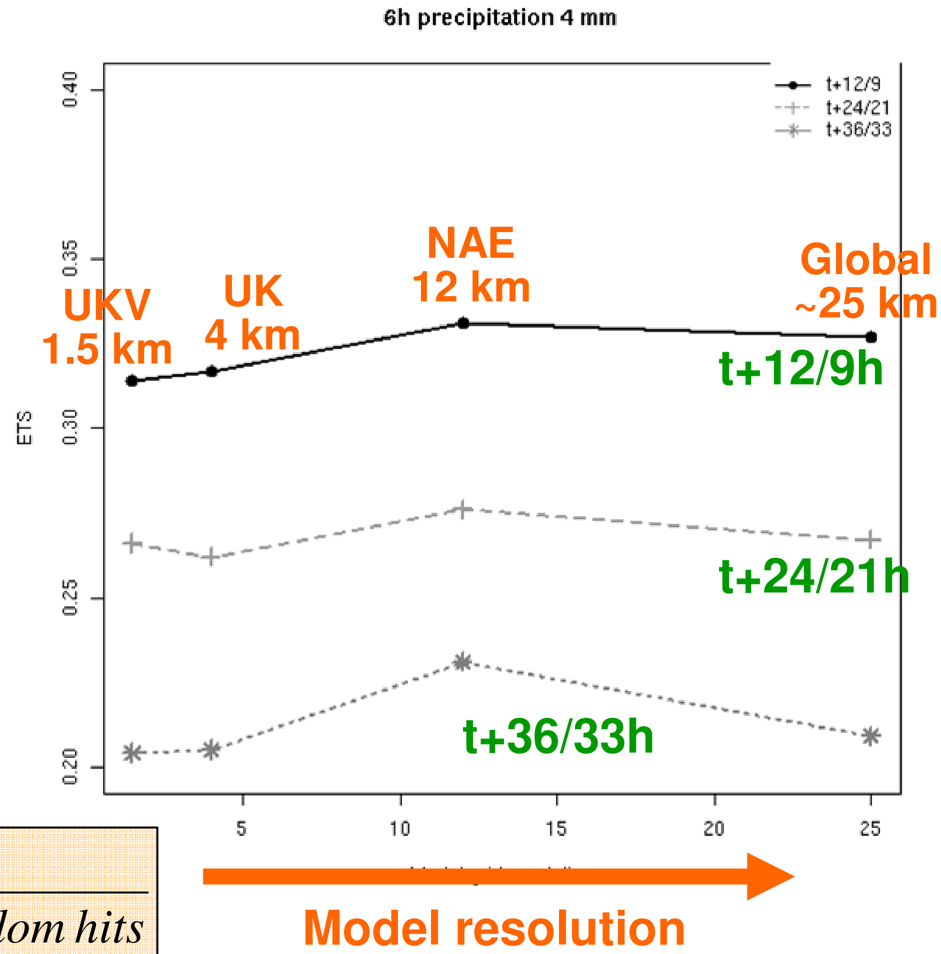
1. Introduction
2. Spatial verification methodology and Fractions Skill Score
3. Key findings from the NAE-UK4 long-term precipitation forecast assessment
4. The thorny issue of "what is truth"
5. Conclusions

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



$$ETS = \frac{hits - random\ hits}{hits + false\ alarms + misses - random\ hits}$$

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

Has this been measured the right way?

There are two main problems.

1. Double penalty effect

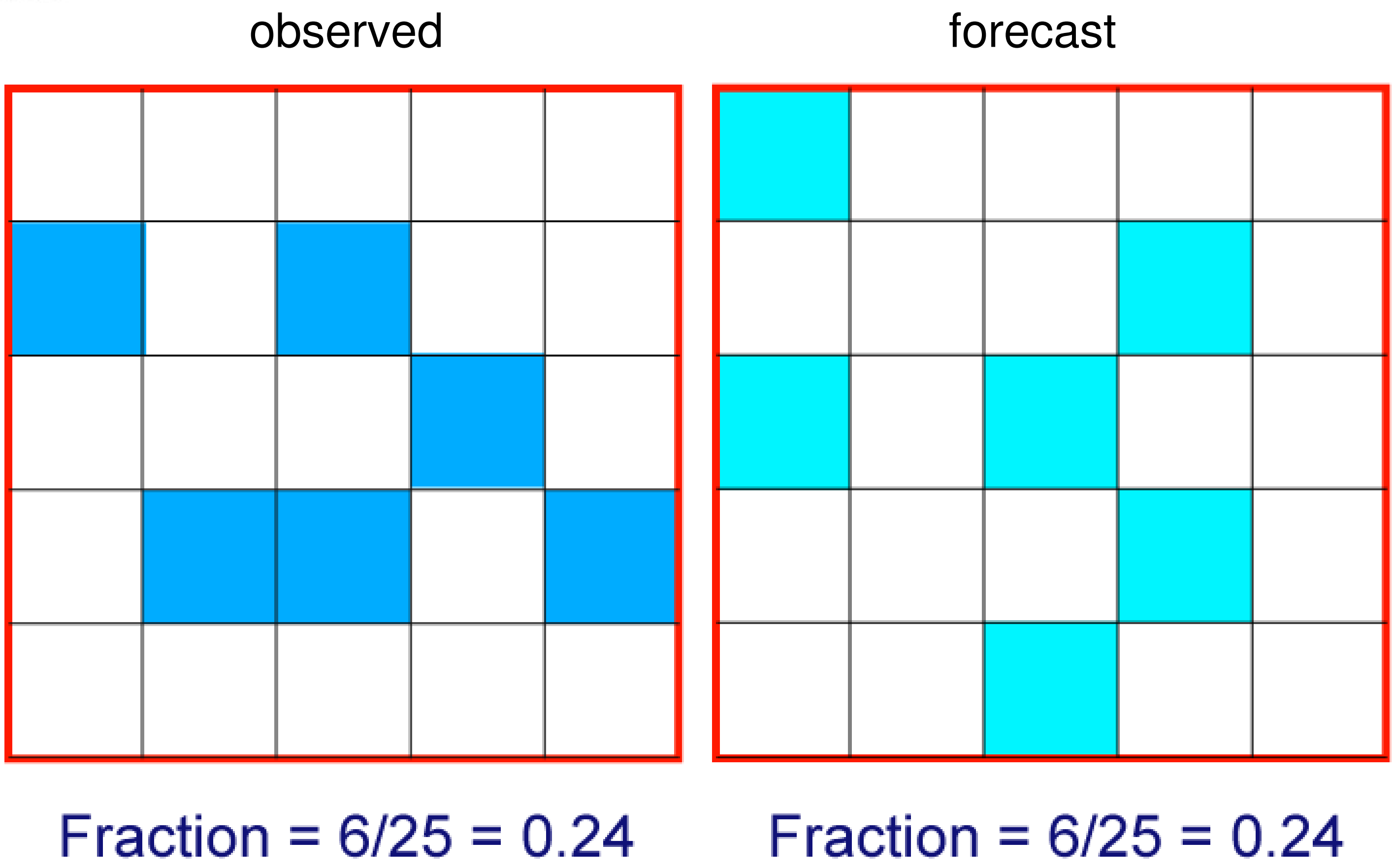
- Errors are counted as false alarms and misses.
- Detail penalised, closeness not rewarded

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



Threshold exceeded where squares are blue

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

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

(Fractions Brier Score)

$0 \leqslant p_j \leqslant 1$ forecast fractions

$0 \leqslant o_j \leqslant 1$ radar fractions

N number of points

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

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

Courtesy of Nigel Roberts

Characteristics of the FSS

Range from 0 to 1 → 0 for zero skill, 1 for perfect skill

Typically increases with spatial scale (always for large sample)

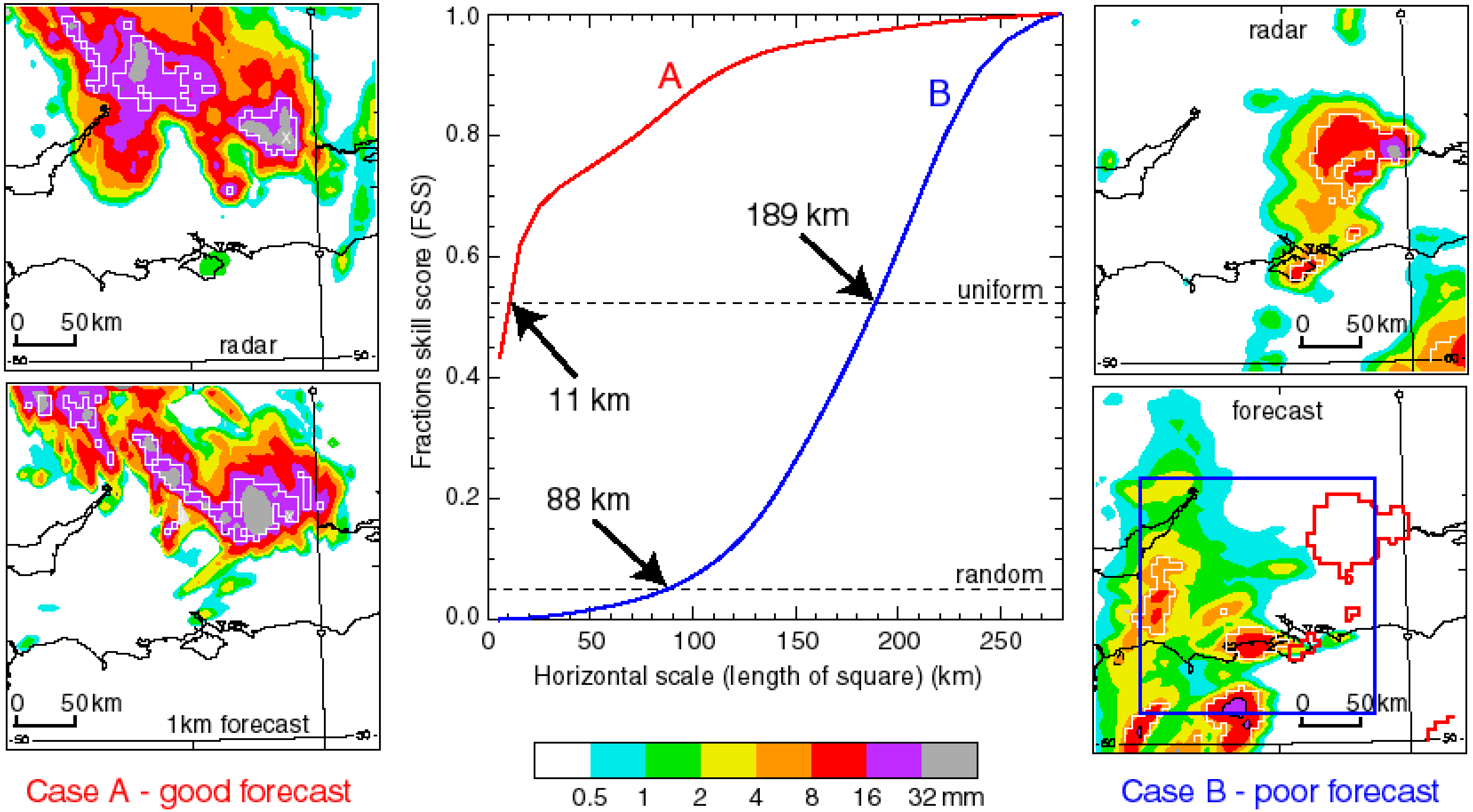
Only asymptotes to 1 in the domain average limit if the forecast is **unbiased or for frequency thresholds**. Typically < 1 for physical thresholds.

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

Courtesy of Nigel Roberts

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.

Met Office

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: $$T = \frac{\overline{D}}{s_D / \sqrt{n}}$$

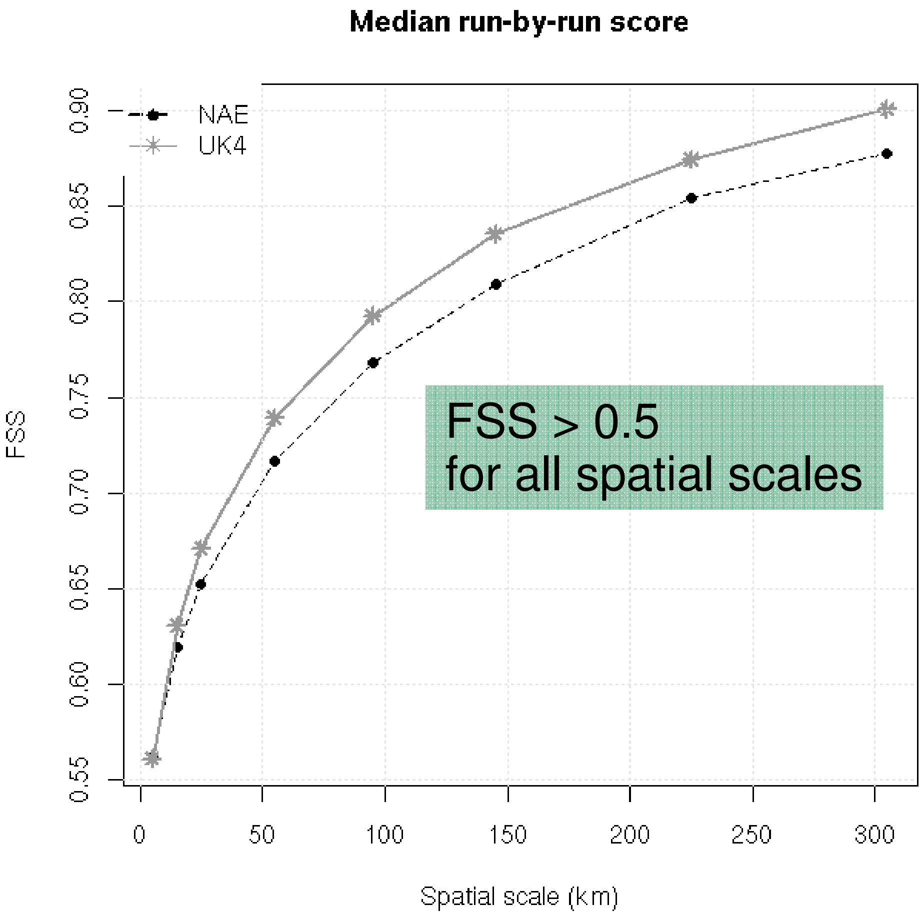
where $\overline{D}$ is the mean of the differences in scores

and s_D is the standard deviation.

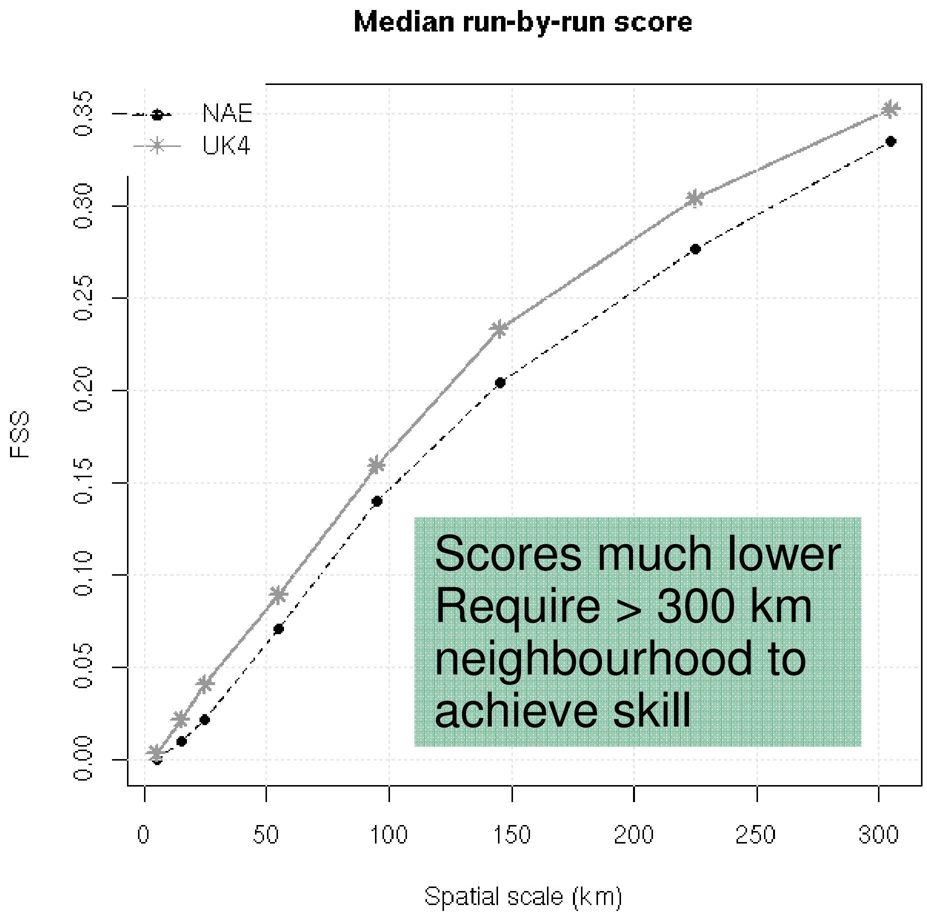
- Test the null hypothesis that H_0: $\mu 1 = \mu 2$ where H_0 is rejected if $t <= t_{n-1,\alpha/2}$ or $t >= t_{n-1,\alpha/2}$.

FSS (neighbourhood size)

0.5 mm/6h



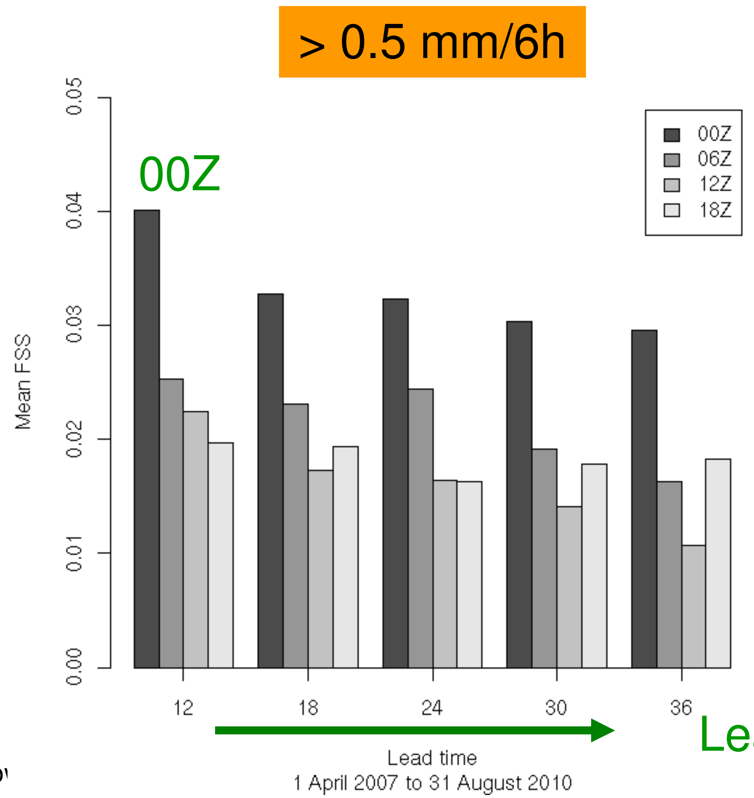
16 mm/6h

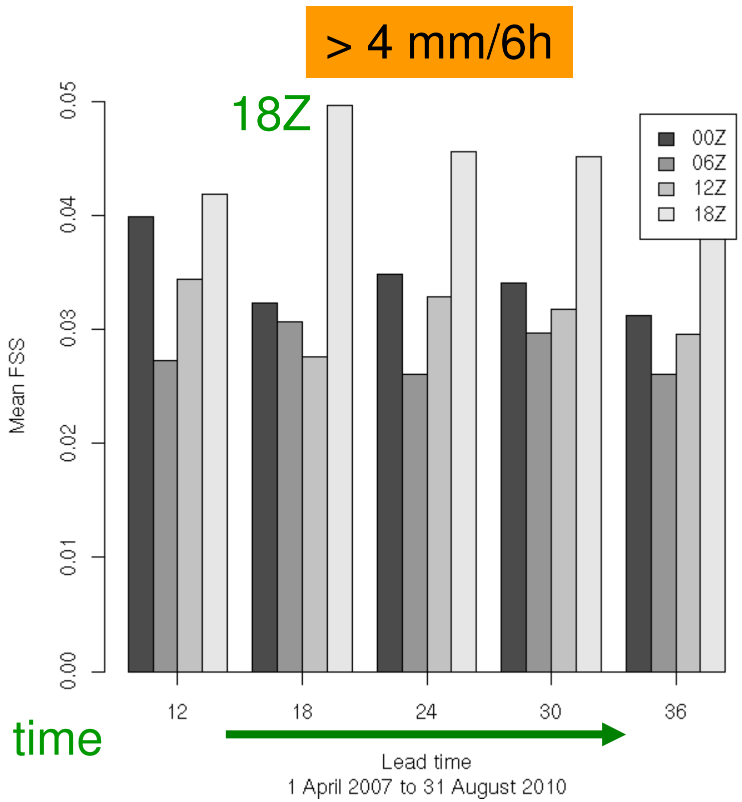


M Mittermaier, N Roberts & S Thompson
submitted to Met Apps

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.





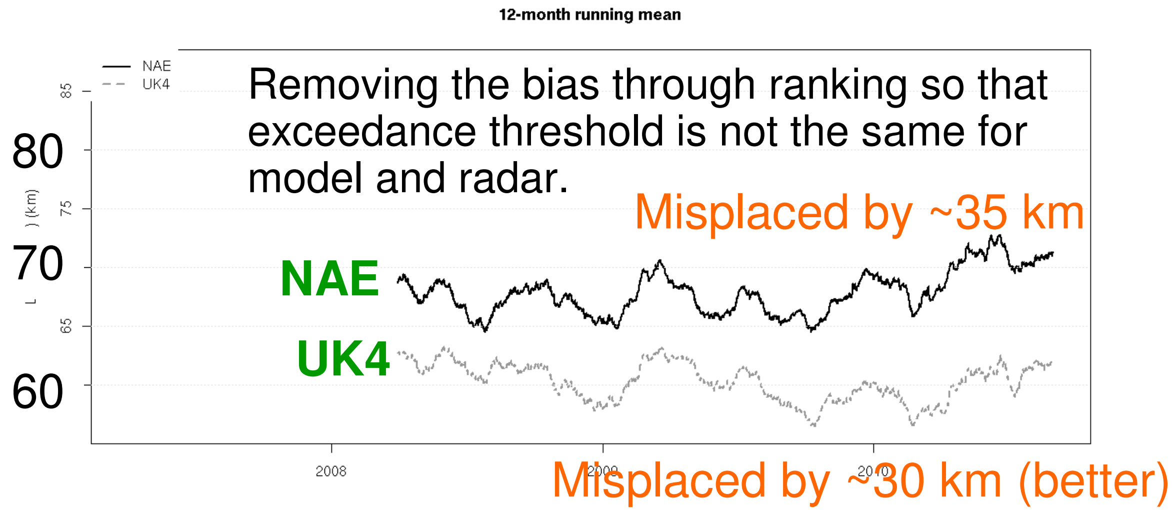
Lead time

L(FSS>0.5) for 10% threshold and 0.5 mm/6h

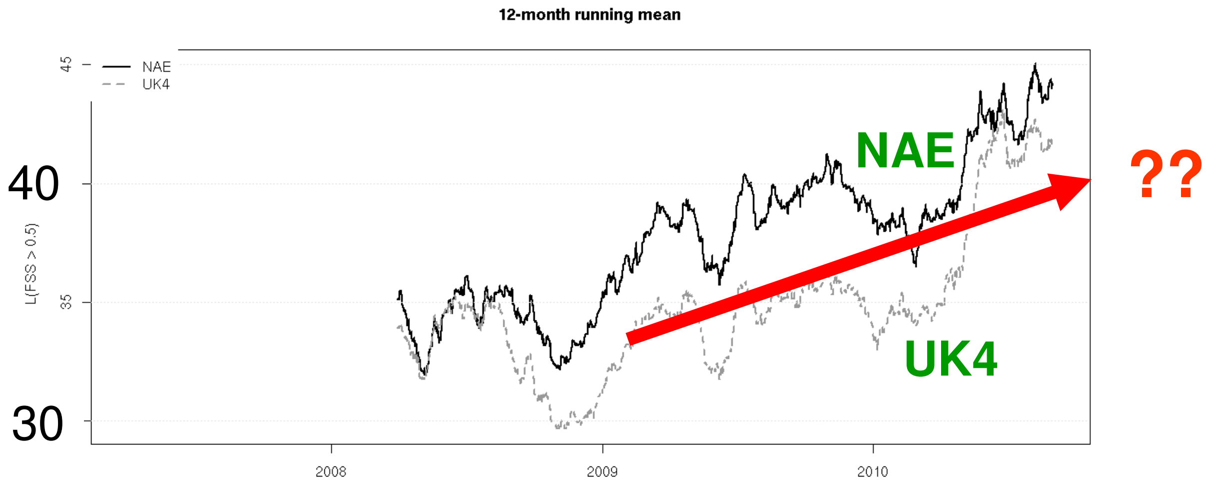
The expectation is that through model improvements L(FSS>0.5) DECREASES over time….. or at least stays constant

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 <u>both</u> 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.

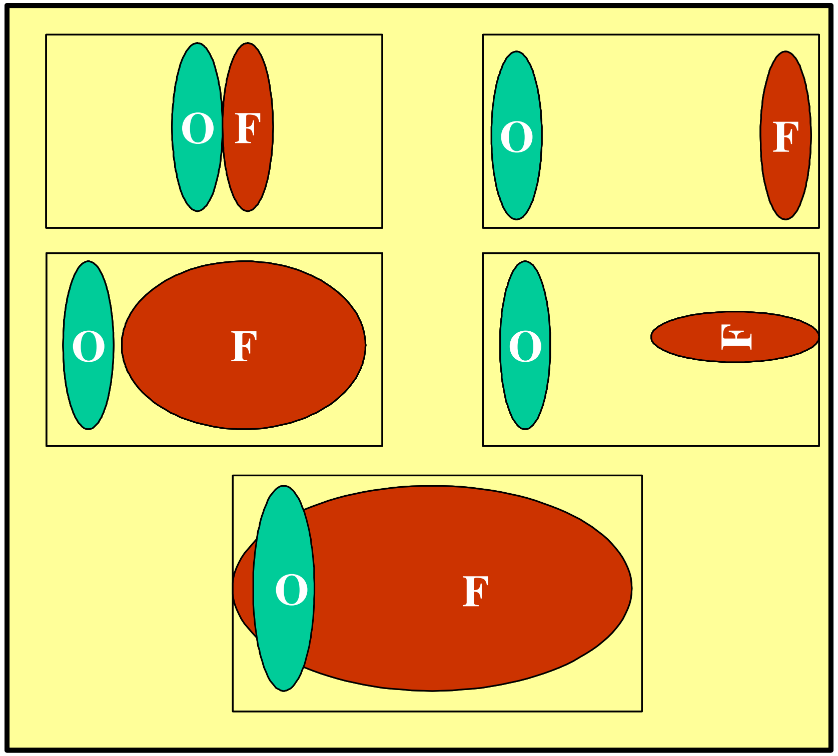
Key findings

- 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!

A long-term assessment of precipitation forecast skill using the Fractions Skill Score.
Mittermaier M., N. Roberts and S. A. Thompson.
Accepted *Meteorol. Apps*. August 2011.

The double penalty



Closeness not rewarded

Detail is penalised unless exactly correct

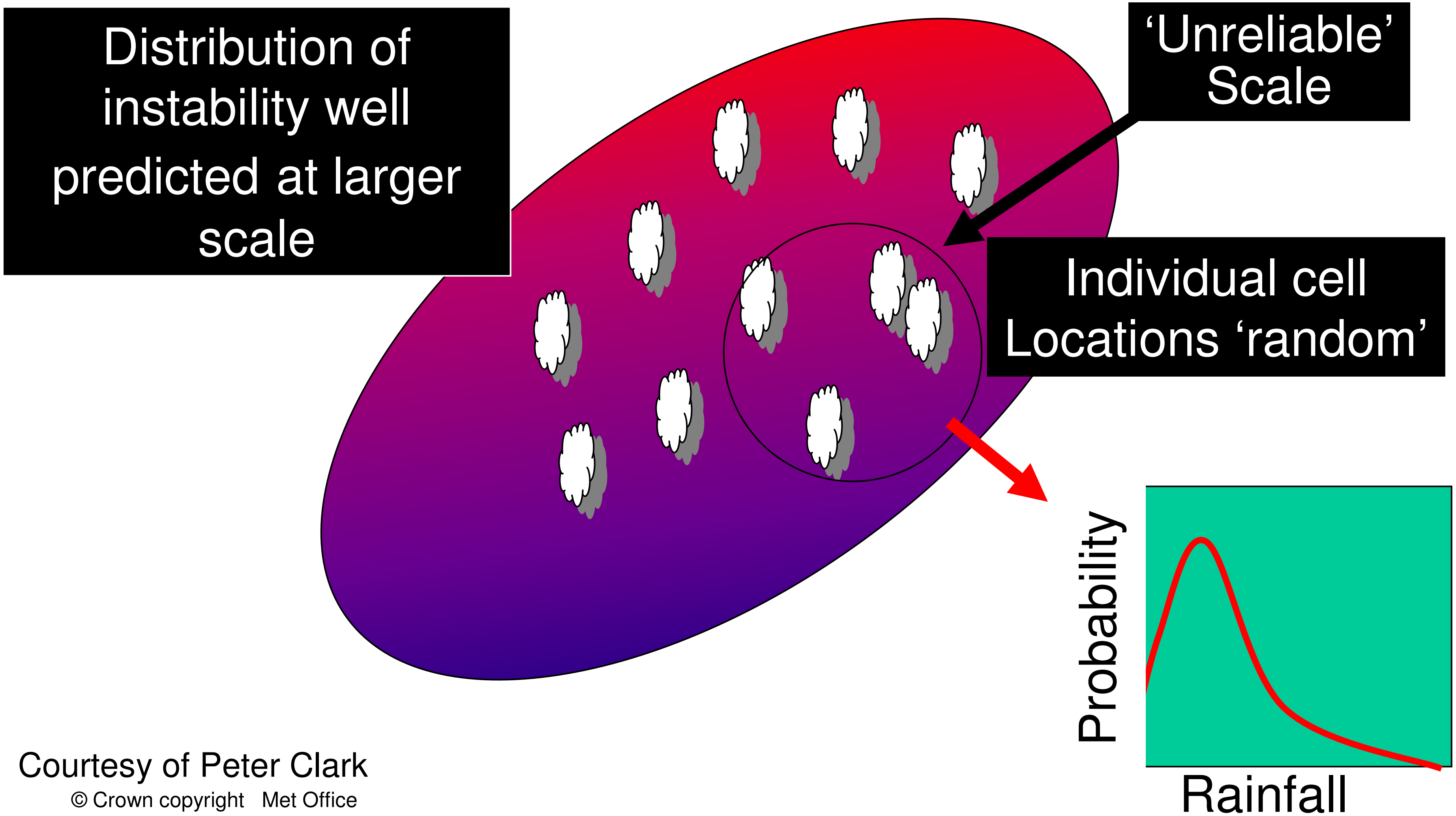
- higher resolution is more detailed!

CSI = 0 for first 4;

CSI > 0 for the 5th

$$CSI = \frac{hits}{hits + false\ alarms + misses}$$

We shouldn't believe high-resolution (at or near the grid scale)



Courtesy of Peter Clark

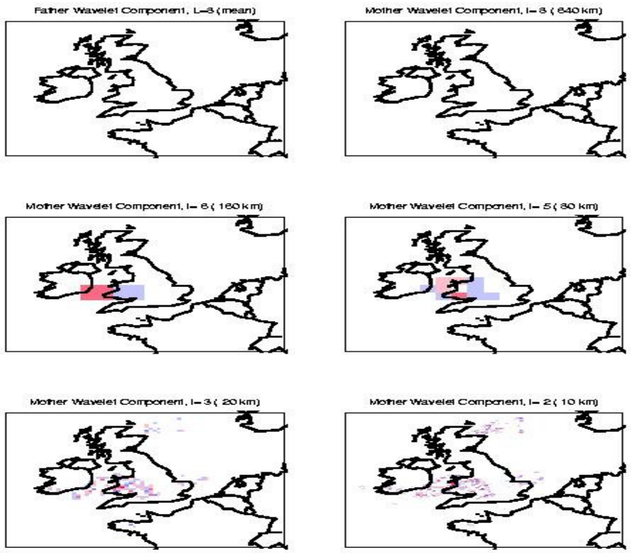
Spatial verification methods

Inter-comparison special issue Wea. Forecasting

Neighbourhood

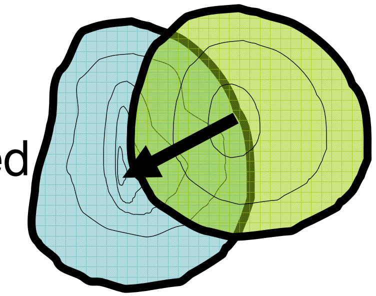
Neighborhood method	Matching strategy*	Decision model for useful forecast
Upscaling (Zepeda-Arce et al. 2000; Weygandt et al. 2004)	NO-NF	Resembles obs when averaged to coarser scales
Minimum coverage (Damrath 2004)	NO-NF	Predicts event over minimum fraction of region
Fuzzy logic (Damrath 2004), joint probability (Ebert 2002)	NO-NF	More correct than incorrect
Fractions skill score (Roberts and Lean 2008)	NO-NF	Similar frequency of forecast and observed events
Area-related RMSE (Rezacova et al. 2006)	NO-NF	Similar intensity distribution as observed
Pragmatic (Theis et al. 2005)	SO-NF	Can distinguish events and non-events
CSRR (Germann and Zawadzki 2004)	SO-NF	High probability of matching observed value
Multi-event contingency table (Atger 2001)	SO-NF	Predicts at least one event close to observed event
Practically perfect hindcast (Brooks et al. 1998)	SO-NF	Resembles forecast based on perfect knowledge of observations

Scale-separation



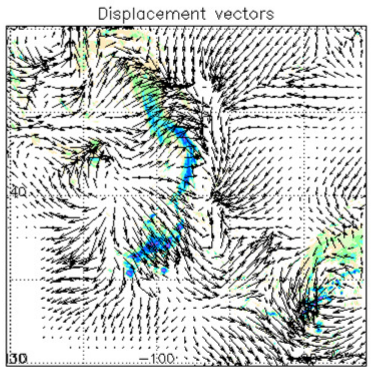
SAL

MODE



CRA

Object-oriented



DAS

Field deformation

Impact of PS changes on precip

Parallel Suite	*Date*	*NAE ppn*	*UK4 ppn*
15	Q1 2007	Negative	Neutral
16	Q2 2007	Neutral	Neutral
17	Q4 2007	Neutral	Neutral
18	Q1 2008	Neutral	Neutral
19	Q3 2008	Neutral	Neutral
20	Q4 2008	Positive	Neutral
22	Q4 2009	Neutral	Neutral
23	Q1 2010	Negative	Neutral
24	Q3 2010	Positive	Neutral
25	Q4 2010	Positive	Neutral

Thanks to Jorge Bornemann and Mike Bush

Why does “truth” have to be so complicated?

What is truth anyway?

Rain gauges

- **Relatively precise and stable**
- Sparse network – not sufficient spatial information
- Point measurement - not a grid box average
- Occasional QC issues: e.g. snow melt
- Accumulation periods too long from many gauges

Radar

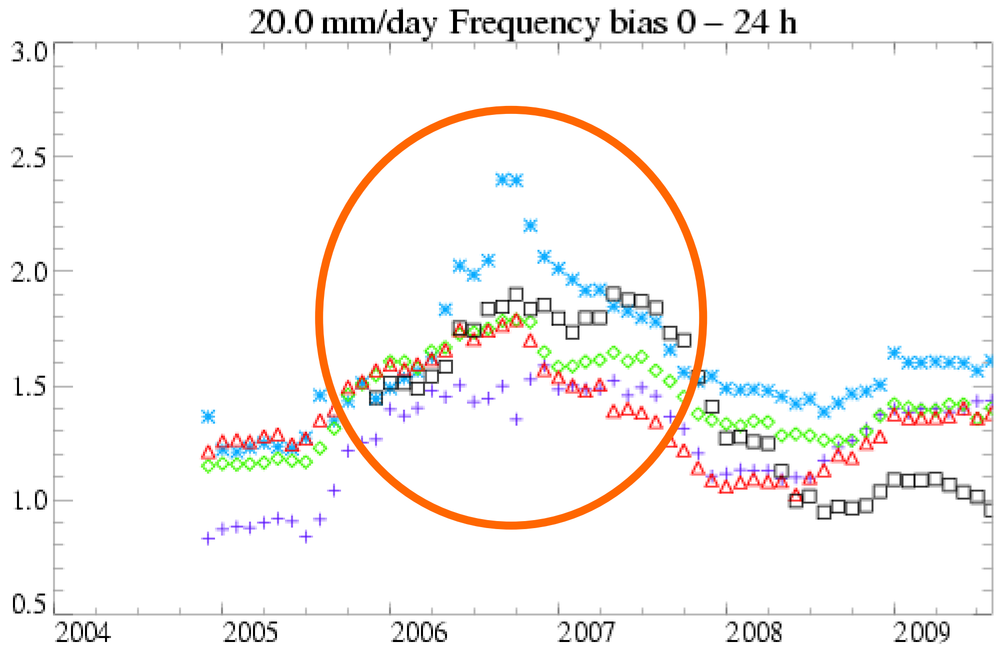
- **Good spatial coverage**
- **Grid square average**
- **Good temporal resolution**
- Assumptions in converting reflectivity to rain
- Clutter, anaprop – can be serious
- **Hardware and software upgraded; enhancements**
- Old network to be upgraded – not stable
- Attenuation in heavier rain
- Orographic enhancement

Nevertheless – if the forecasts looked like radar we'd be delighted

Courtesy of Nigel Roberts

The European Model Intercomparison of Precipitation (EMIP) …

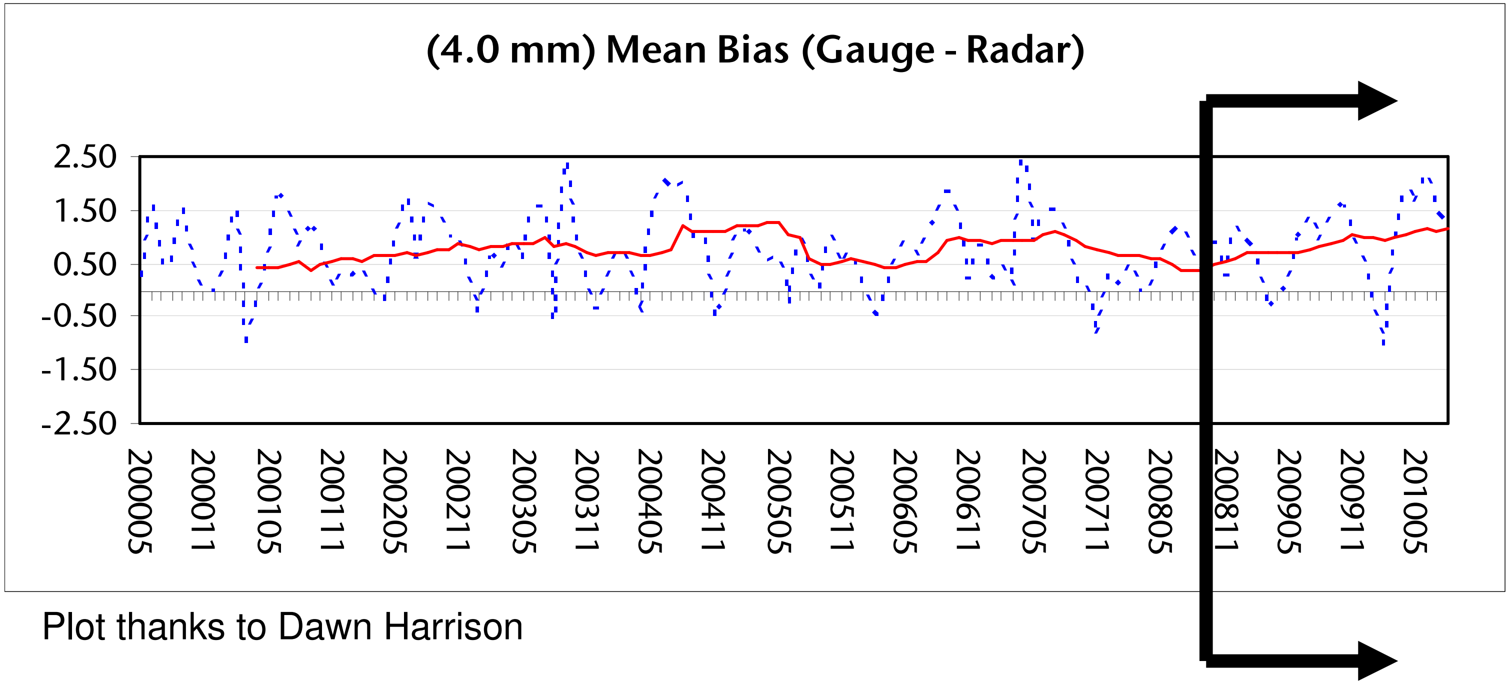
- … showed the power of using several models for monitoring the radar baseline.



Traced to an issue of 5-min data used for hourly accumulations being deleted before the hour ended, so hourly accumulations only consisted of 45 min or 9 5-min slices.

From Mittermaier *et al* in prep to ASL with SRNWP collaborators

Gauge-radar bias against calibrating gauges



Plot thanks to Dawn Harrison

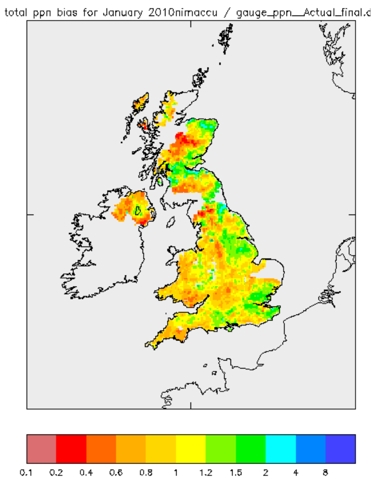
Caveats:
- Calibrating gauges not representative.
- Some radars have none in domain!

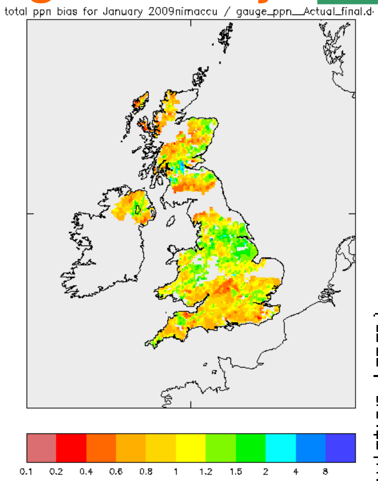
- A gradual increase in the bias towards **greater under-estimation by radar** means that fewer events breach a physical exceedance threshold, introducing a bias through the observations into the model frequency bias and scores.

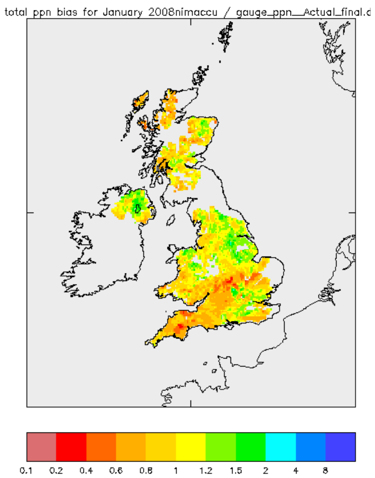
Monthly maps and time series

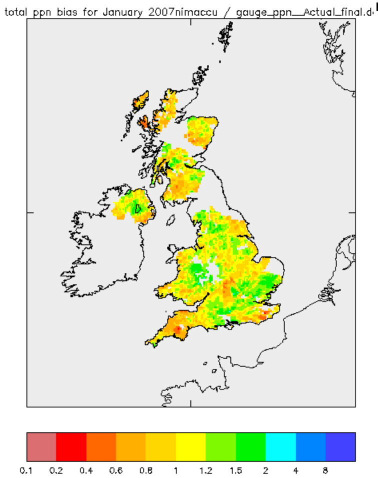
CAVEAT: not equally matched. Bias highly variable in space.

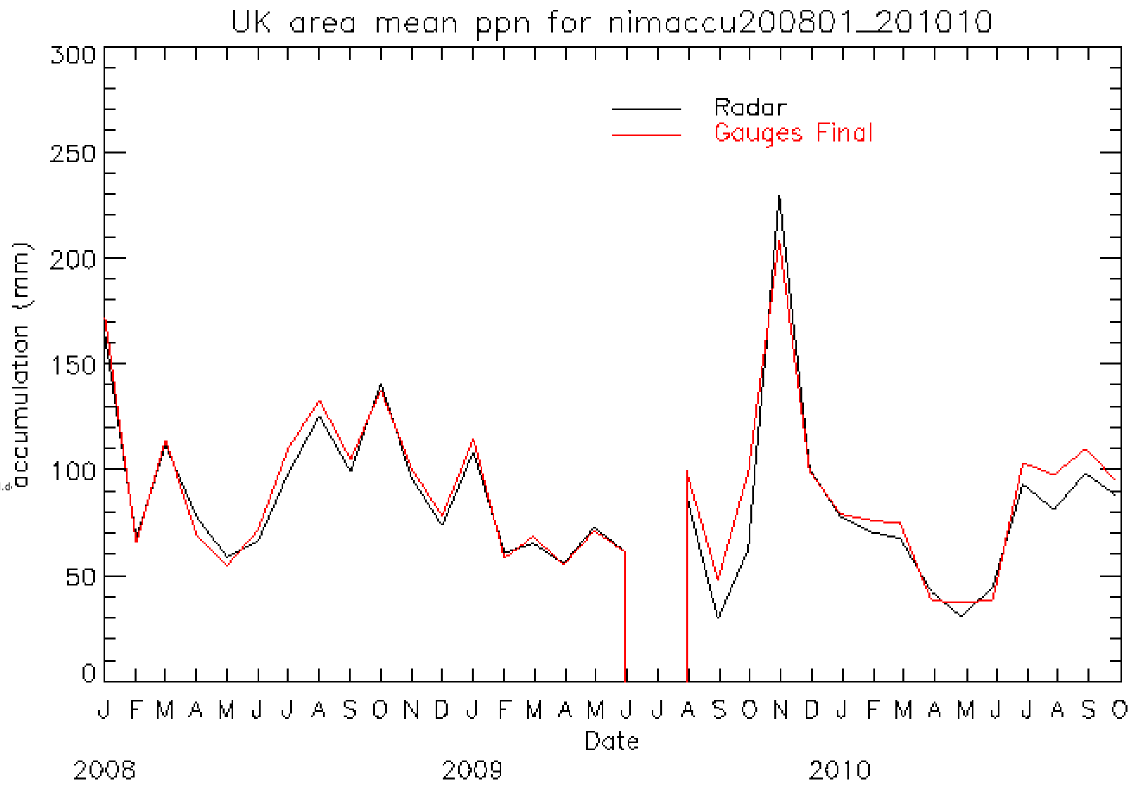
Bias Radar/Gauge January









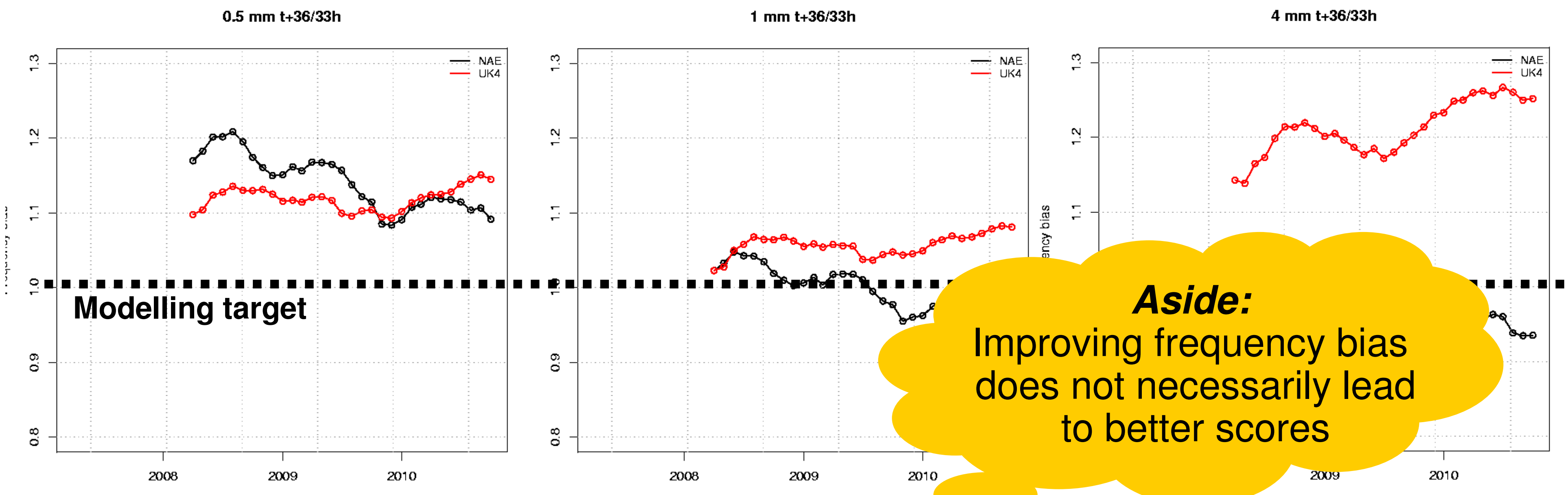


Radar more likely to be “under”.

All plots Clive Wilson

Model bias against gauges

12-month means



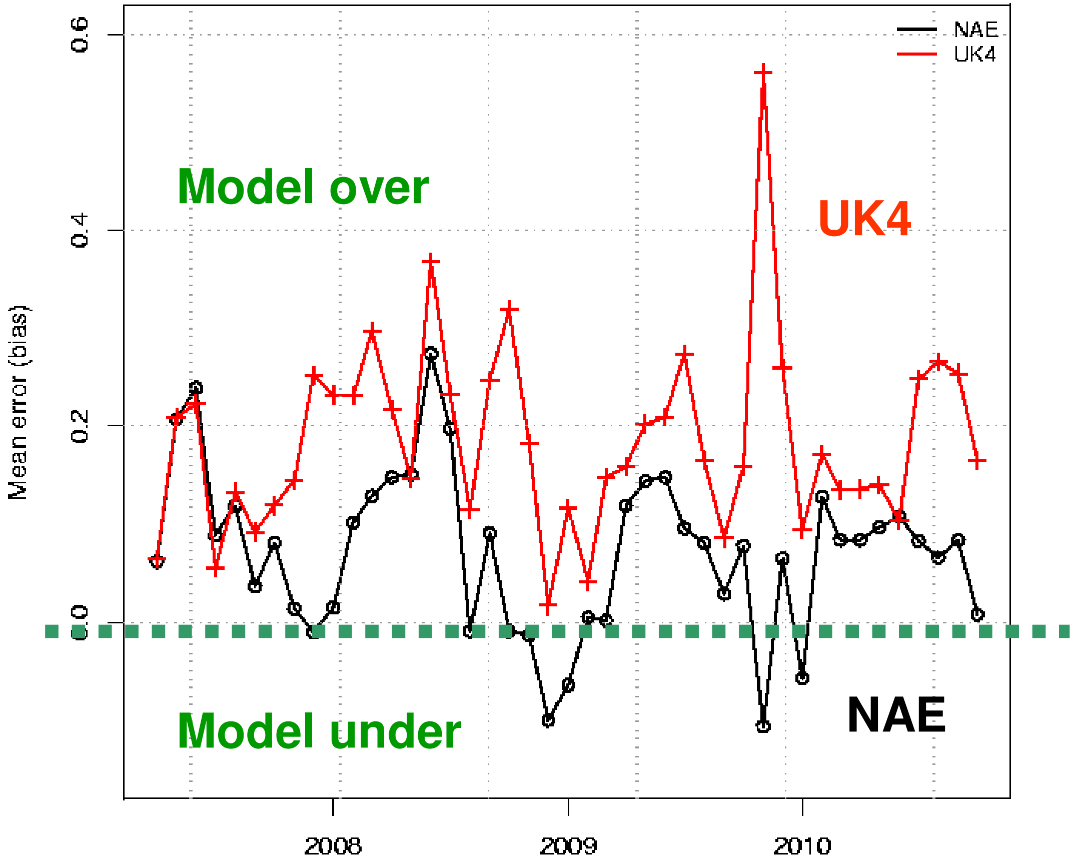
Aside:
Improving frequency bias does not necessarily lead to better scores

- Gradual improvement in NAE bias.
- Under-estimation of NAE for larger thresholds (expected)
- Over-estimation of UK4 at larger thresholds (expected). Worsening trend possibly not expected?

Model bias against gauges 2

(calculated more like the gauge-radar bias)

- Monthly ME values
- Not conditional (so slightly different to radar-gauge metric)
- In millimetres



Met Office

What would help?

- A **better operational change process** (like OPCHANGE) and understanding of what impact radar changes may have on downstream quantitative users (whether it's Cyclops changes, compositing changes, calibration changes etc etc etc).

Imminent

- Invest in the development of a **high-resolution gridded gauge analysis** which enables a wider comparison of processing changes, and the development of an optimally merged gauge-radar product.

- **Better automated QC control for the radar network as a whole** (in relation to how IT(Ops) control the radar network), e.g. understanding the implications of taking radars out of the network → it may make the product worse.

In more detail

- **Both model trends are behaving similarly** which points to a characteristic of the baseline. One does not expect them to behave in exactly the same way as they are not at the same resolution.

- Even if the baseline is changing **a comparison is valid because both models are compared against the same baseline**. Using absolute (physical) values is potentially dangerous.

- What happens if we don't use it comparatively (as for long-term monitoring)? **Baseline changes invalidate the results in physical terms because changes can not be attributed with certainty to model changes alone.**

- **Frequency thresholds are preferred.** They encompass the full range of precipitation and all rain is counted.