Review of verification activities and developments

Clive Wilson – for Expert Team on diagnostics, validation & verification

31st EWGLAM/16th SRNWP meetings – Athens 29 Sep 2009

Contents

- Expert team members
- EUMETNET SRNWP-V – separate presentation tomorrow
- 4th workshop of WWRP/WGNE working group on verification ,Helsinki, 8-11 June 2009
- ECMWF TAC subgroup
- Consortia activities
- Plans

Expert Team on diagnostics, validation & verification

- Members
  - Clive Wilson (chair), Joël Stein, Carl Fortelius, Francis Schubiger , Dijana Klaric
  - Dave Richardson (ECMWF contact)
- Additional members
  - Marek Jerczynski, Alexander Kann , Andrea Raspanti, Ulf Andre, Xiaohua Yang, Lovro Kalin, Nigel Roberts, Marion Mittermaier

EUMETNET/SRNWP Verification programme

- **D1: Operational verification comparison of one version of each of the 4 regional models of Europe (available for all the participating members).**
- Responsible Member – Met Office
- Programme manager-Clive Wilson
- Deputy PM : Marion Mittermaier
- Commenced 1 Jan 2009
- Ends 31 December 2010
- Costs – € 32000 /year

4th International verification Methods workshop -JWGFVR

- Helsinki, 8-11 June 2009
- Pertti Nurmi (FMI) local organiser
- Aims
  - Focus on extremes & severe weather
  - Ensembles/probabilistic verification
  - Uncertainty & Value
  - High resolution forecast verification
  - Promote more focused user-oriented verification

4th International verification Methods workshop –SRNWP participation

- 9 members of ET attended + several others
- 12 oral, 4 poster
- Clive Wilson: Do key performance targets work?.
- Lovro Kalin: Is ETSS Really Equitable?
- Adriano Raspanti: VERSUS: Unified Verification Package in COSMO
- Marek Jerczynski: Some robust scale separation methods at work
- Marion Mittermaier: Time-series analysis of scale-selective verification: Can we use it for operational forecast monitoring?
- Joel Stein and Marielle Amodei: Another look at the contingency tables: Scores based on Manhattan distances in the error space

4th International verification Methods workshop –SRNWP participation -2

- Clive Wilson: A critical look at the verification of Met Office "Flash" Warnings
- Marion Mittermaier: Verifying extreme rainfall alerts for surface water flooding
- Marion Mittermaier: Identifying skillful spatial scales using the Fraction skill Score
- Ulrich Damrath: Some experiences during verification of precipitation forecasts using fuzzy techniques
- Kees Kok: Valuing information from high resolution forecasts
- Chiara Marsigli: QPF Verification of Limited Area Ensemble Systems during the MAP D-PHASE OP
- Marielle Amodei: Deterministic and fuzzy verification of the cloudiness of High Resolution operational models
- Francis Schubiger: Verification of precipitation forecasts of the MAP D-PHASE data set with fuzzy methods
- Sami Niemelä: Verification of High resolution Precipitation forecasts by Using the SAL Method

ECMWF TAC Verification Subgroup

- recommend headline measures that are suitable to complement those in the current ECMWF Strategy (namely anomaly correlation of Z500 for the deterministic forecast and probabilistic scores of T850 for the EPS);
- recommend verification procedures to aid forecasters' decision making;
- recommend measures suitable for validating forecasts of weather associated with high impact events;
- identify requirements for observational data necessary for this verification;
- review Member State and Co-operating State requirements for the development of future forecast products.

ECMWF TAC Verification Subgroup - meetings March & Sept 2009

- Chair Pertti Nurmi FMI
- Martin Goeber- DWD
- Carlos Santos AEMET
- Marielle Amodei Meteo-France
- Jim Hamilton – Met Eirean
- Kees Kok- KNMI
- Marion Mittermaier Met Office
- Clive Wilson EUMETNET
- David Stephenson, Chris Ferro, Exeter University
- ECMWF David Richardson, Mark Rodwell,Walter Zweiflhofer, Erik andersson, Laura Ferranti, Anna Ghelli, Time Hewson,Cristina Primo

ECMWF TAC Verification Subgroup- new headline score

- 24h precipitation new headline currently deterrministic fc
  - Nearest grid-point instead of interpolation
  - ACC 7 RACC sensitivity to outliers
  - 1995-2005 obs Europe 2800 → 3600 per month
    - ACC sensitive to sample size
    - Zero rainfall inlfuence needs investigation
    - Operational verification different QC
- Mark Rodwell – new score SLEEPS
- Semi-linear Equitable Error in Probability Space
  - Potential candidate for monitoring NWP trends
  - Shows reduced errors over time

SLEEPS Mark Rodwell NB still being developed

- 3 category
  - Dry, light and heavy ppn
  - Based on local climatology at stations ~5000
    - 1980-2007
  - Need at least 150 (~5 years for each month)
  - Scoring matrix for 3 cumulative prob categories
    - Equitablity imposed
    - Not symetric -> semi-linear

ECMWF TAC Verification Subgroup –extreme events

- Extreme dependency score
  - Modified –independent base rate
  - Symmetric SEDS (Hogan et al)
  - All converge for rarer events
- Theroretical/practcal properties need further investigation before most suitable score can be proposed
- Influence of bias on EDS-type scores to be investigated

ECMWF TAC Verification Subgroup – aid to forecasters

- New clustering scheme for flow dependent verification
  - Proposed compute prob. Scores verifying analyses & individual clusters – see accuracy at different forecast ranges
  - Investigate transitions between regimes
  - Stratification of standard error measures by regimes for assessing year on year changes
- Feature tracking
  - Extra tropical cyclones (Hewson)
  - Need to look at specific cyclones cf TCs
  - Emphasis on those associated with extreme weather
  - Strike probs in various classes severity cf analyses

Aladin verification activities

1) Common Aladin verification package

- operational (Slovenia)
- No significant change during last year.
- Against surface stations and radio-sounding European data.
- Allows comparison of the different versions of Aladin
- Results can be compared with the inter comparison results from the Met Office.

Aladin verification activities
2) “Fuzzy” methods

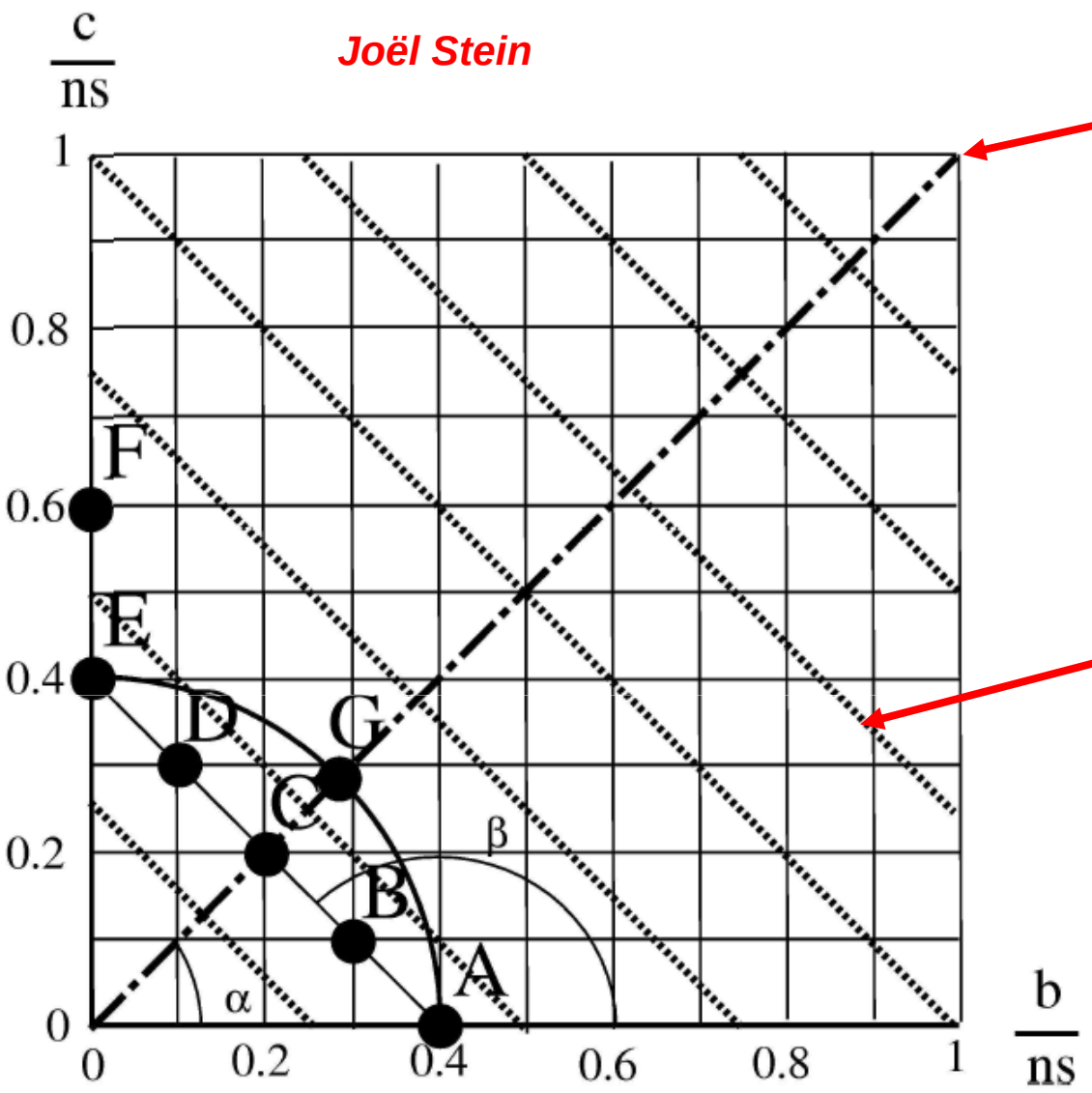
- Fuzzy, pattern recognition tests (Poland)
- Operational in Météo-France
  - compare high and low resolution models (AROME and ALADIN-FRANCE)
  - Deterministic forecasts --> probabilities
    - forecast frequency in a neighbourhood = P_N
- 2 Brier skill scores against *persistence either* :
  - compare P_N to 0 /1 value observed at the centre of the neighborhood (BSS_SO for single observation)
  - or to the observed frequency in the same neighborhood (BSS_NO for neighborhood observation).
  - Use climatological French raingauges network (Amodei and Stein 2009) and the tables of contingency corresponds to a temporal windows of 3 months.

Aladin verification activities

- Comparison of post-processed forecasted brightness temperatures with observed temperature by Meteosat 9 for ALADIN and AROME models.
  - Classical and probabilistic scores to quantify the double penalty influence on the comparison between 2 models of different resolution
  - Presentation by I. Sanchez in Athens
- “Theoretical" formulation of the scores relates them to Manhattan distances in the phase space of the possible forecasts when the observations are fixed.
  - poster presented in Helsinki (Stein and Amodei 2009) based on a preliminary version of a paper submitted to Met App.
  - Revised version restricted to 2x2 tables is now under consideration (Stein 2009).
  - The graphical representation deduced from this study of the table of contingency is now implemented on our web site to compare Arome and Aladin.

Scaled misses

Joël Stein



Heidke skill score

Bias=1

Equal error distance

Scaled false alarms

Figure 2: Graphical representation of the 2D space of errors normalized by the number of observed events ns. The abscissa correspond to b/ns and the ordinates to c/ns. Representative points of seven forecasts are plotted with full disks. Isolines of constant distance D1 (see text for details) from the origin are represented by slanted dotted lines. The line B=1 (dash-dotted line) separates the region of over-forecasts (B>1) located under this line from the region of under-forecasts (B<1) above it. The base rate is s=0.6.

Joël Stein

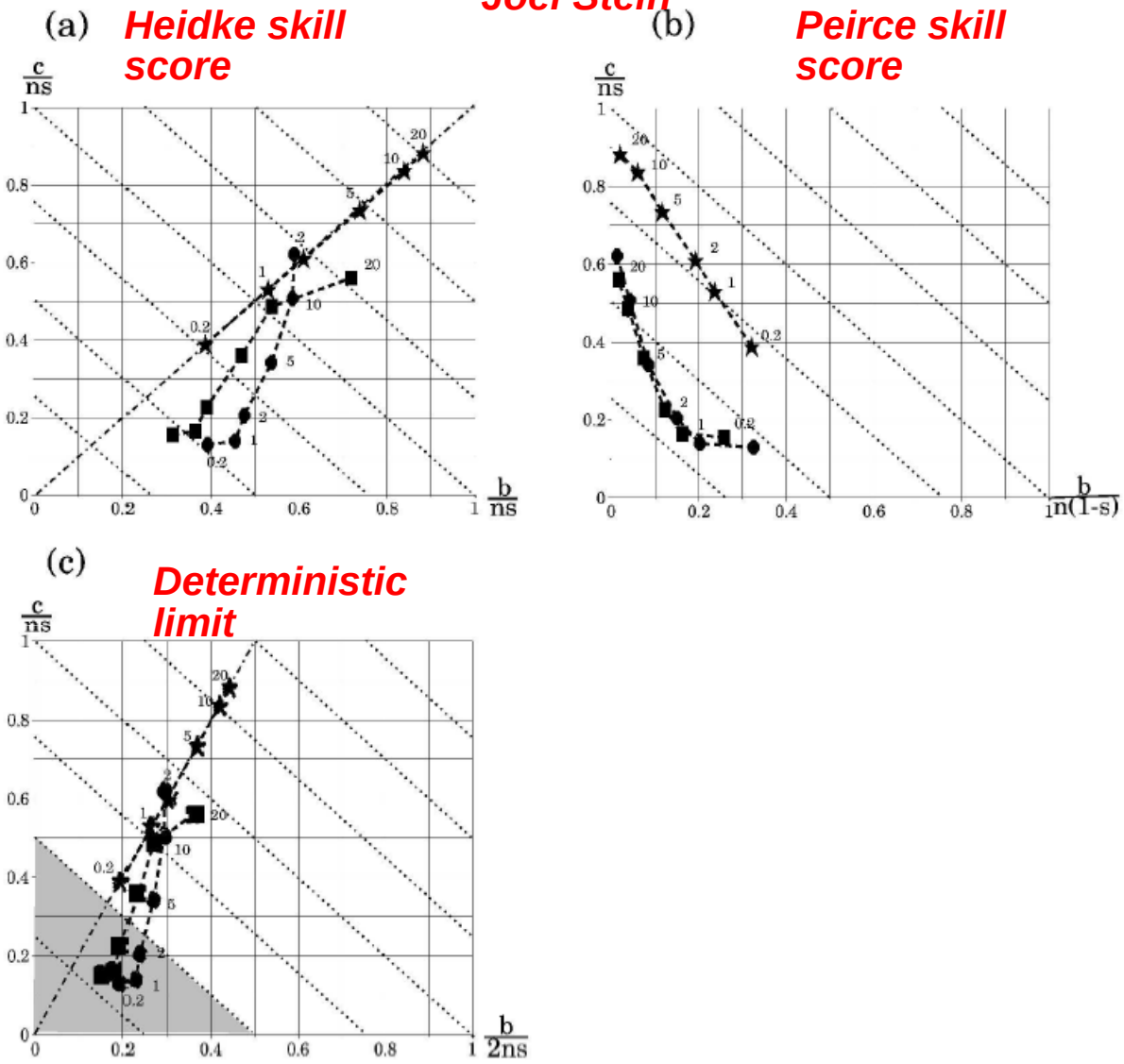


Figure 7: Comparison of the QPF performed by ARPEGE (full disk), ALADIN (full square) and the persistence forecast (star) for autumn 2005. The reference is provided by the 24 hours accumulated rain measured by the French climatological network of rain gaujes. The three different panels correspond to the same axis as in Figure 5 (a), (b) and (c). The thresholds list is the same as in Figure 6. The thresholds are recalled near every representative points. The grey area corresponds to

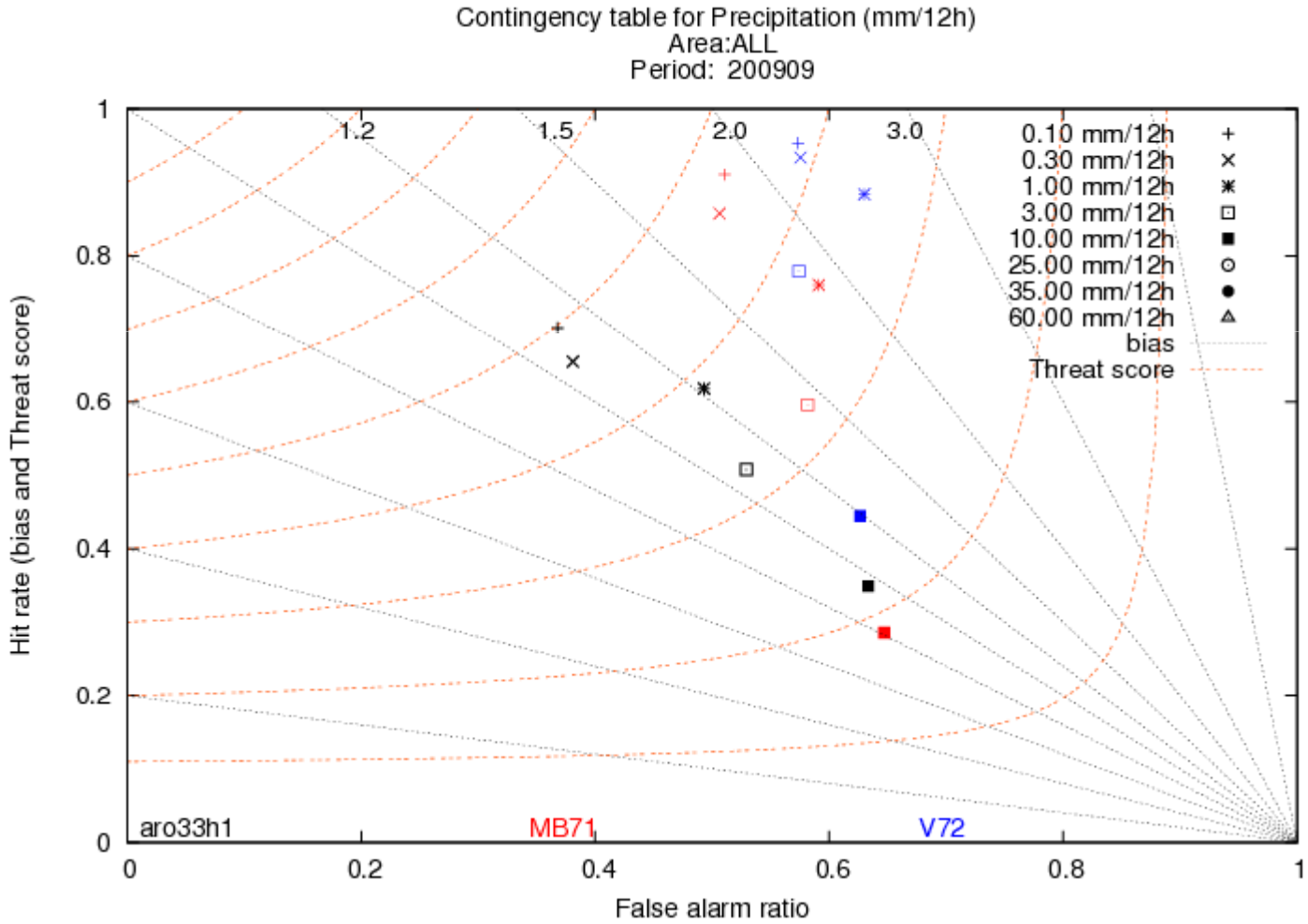
COSMO verification activities -separate slides

- 200910_Clive_EWGLAM-SRNWP.ppt

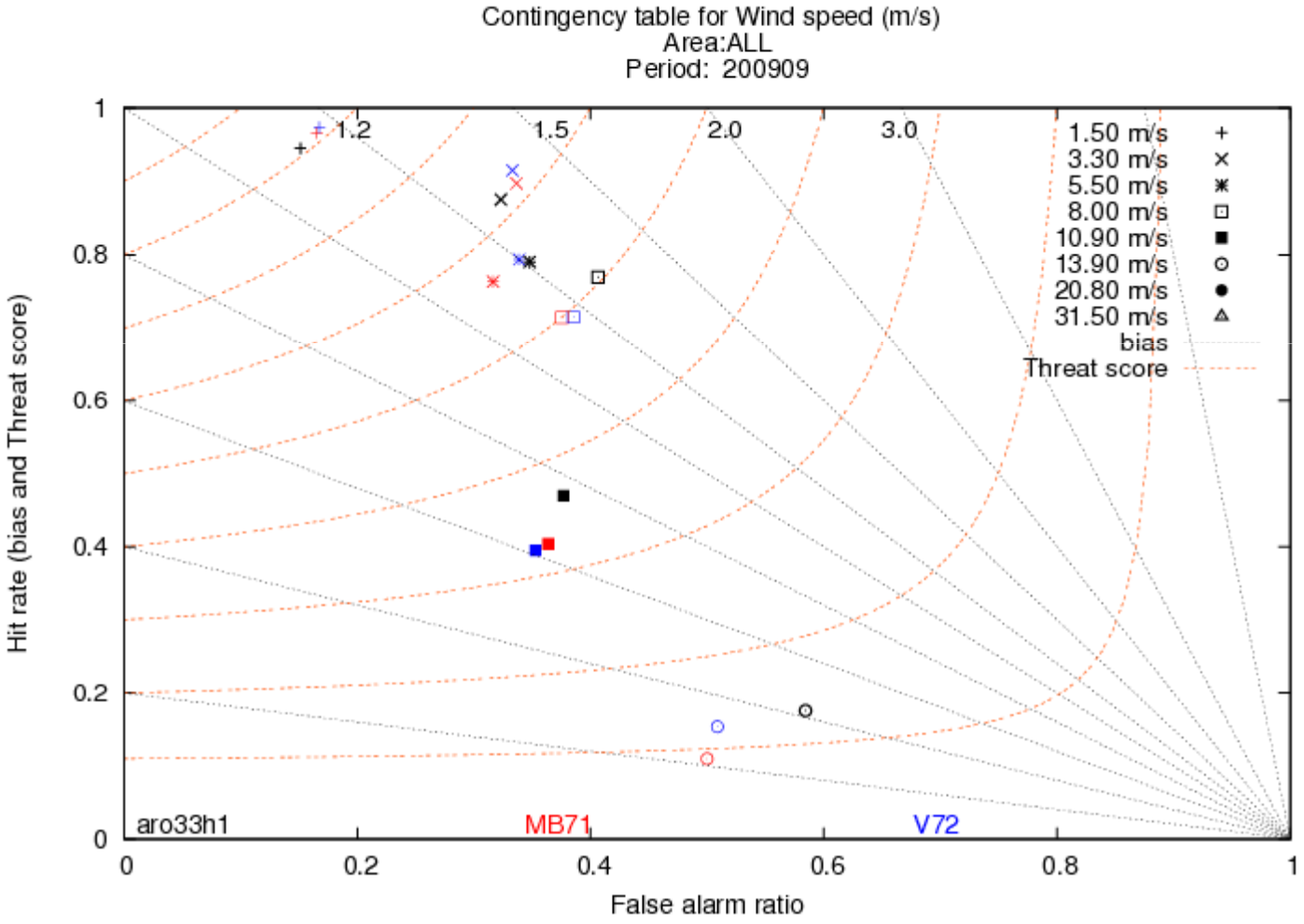
Hirlam verification activities

- HARMONIE ver package (Andrae 2007)
  - Powerful & flexible
  - compare models & to observations
  - Tables, maps,time-series,vertical profiles,histograms,scatter, du=iurnal & seasonal cycles
  - Standards scores & contingency table scores
  - Includes SAL (Wernli et al, 2008)
  - Recent addition graphs of freq. bias & threat score on hit rate v false alarm ration (Wilson, 2009)

aro33h1: AROME 33h1 (2.5km L40)
MB71: HIRLAM 7.1.4 (7.5km L60)
V72 (RCR): HIRLAM 7.2 (16.5km L60)



aro33h1: AROME 33h1 (2.5km L40)
MB71: HIRLAM 7.1.4 (7.5km L60)
V72 (RCR): HIRLAM 7.2 (16.5km L60)



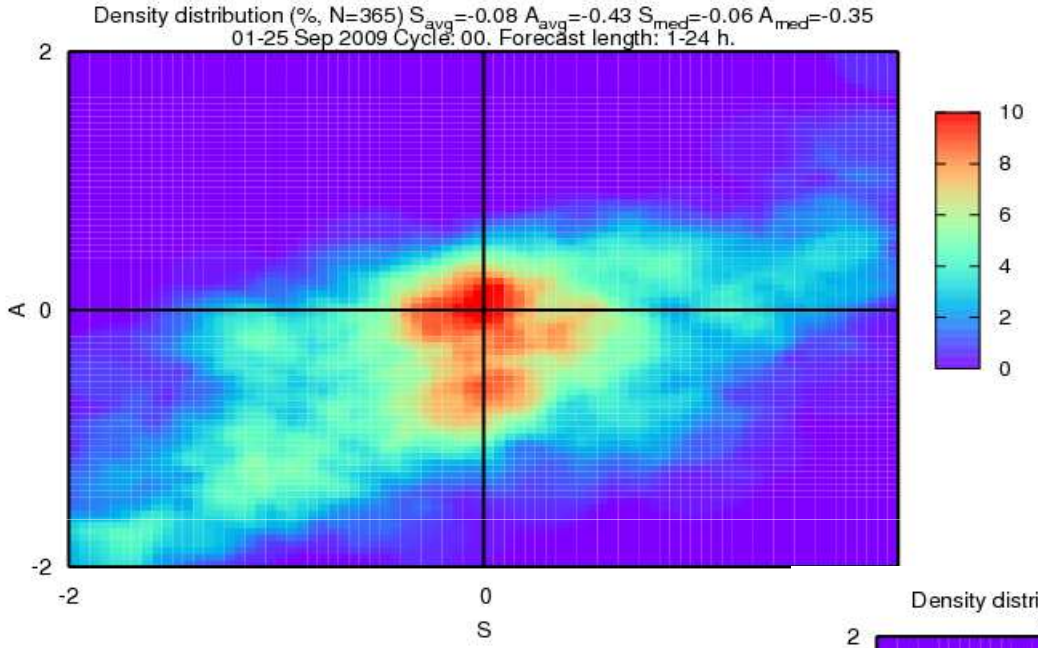
Hirlam verification (contd)

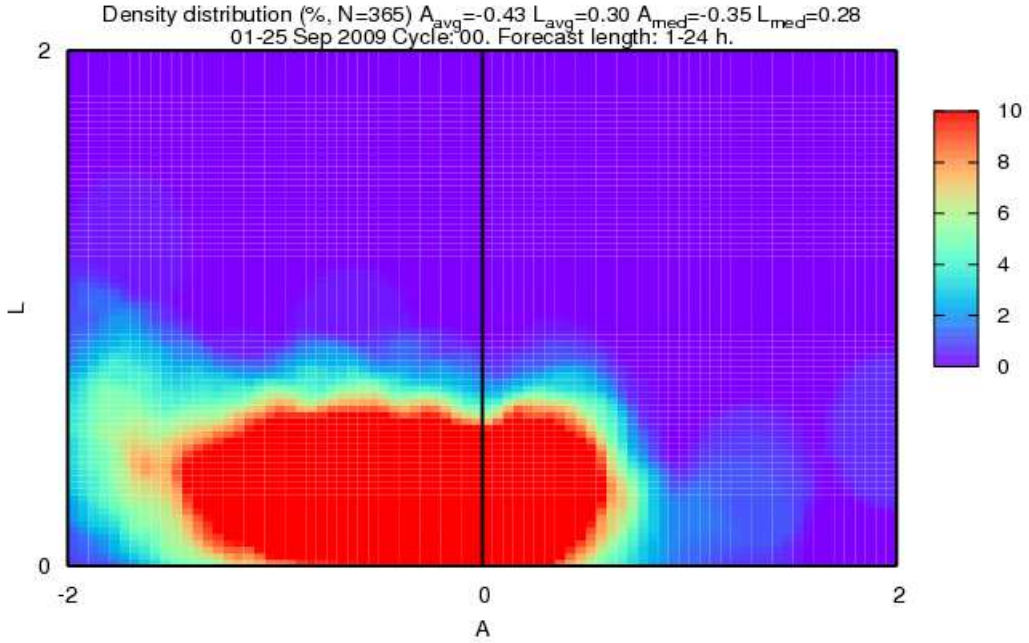
- GLAMEPS calibration & validation at AEMET
  - HPPV (Santos & Hagel, 2007)
  - Multi-model
  - Rank histograms, PIT histograms, spread-skill, Brier SS, ROC, reliability, sharpness, RV
- Feature based verifcation :
  - FMI- Finnish radar reflectivity compared to Finnish AROME using radar simulation model (Niemela, 2009 4th Intern. Workshop))
  - SAL
- Fuzzy – MOS, traditional scores (Kok et al 2008)

SAL features

S:	**Structure**	**-2**	**…**	**0**	**…**	**+2**
		objects too small or too peaked		Perfect		objects too large or too flat
A:	**Amplitude**	**-2**	**…**	**0**	**…**	**+2**
		averaged QPF under-estimated		Perfect		averaged QPF over-estimated
L:	**Location**			**0**	**…**	**+2**
				Perfect		wrong location of Total Center of Mass (TCM) and / or of objects relative to TCM

SAL


Density distribution (%, N=365) S_{avg}=-0.08 A_{avg}=-0.43 S_{med}=-0.06 A_{med}=-0.35
01-25 Sep 2009 Cycle: 00. Forecast length: 1-24 h.
2
0
-2
A
-2
0
S
10
8
6
4
2
0


Density distribution (%, N=365) A_{avg}=-0.43 L_{avg}=0.30 A_{med}=-0.35 L_{med}=0.28
01-25 Sep 2009 Cycle: 00. Forecast length: 1-24 h.
2
0
L
-2
0
2
A
10
8
6
4
2
0

Met Office verification activities

- Operational verification package now extended for ensembles- MOGREPS
  - Reliability, rank histograms, ROC, Brier, value
  - Multimodel ensembles
- Fractional skill score (FSS) now operational for NAE & UK4 (soon UK1.5km)
- Evaluation of new Flood forecasting - Extreme Rainfall Alert service
  - Probabilistic from “fuzzy” UK4 & lagged average
- Evaluation of new 1.5km
- Review of warnings (Exeter University, Stephenson & Joliffe, also 4th Intern. workshop)

Fractional skill score (FSS) Verification approach (Mittermaier & Roberts)

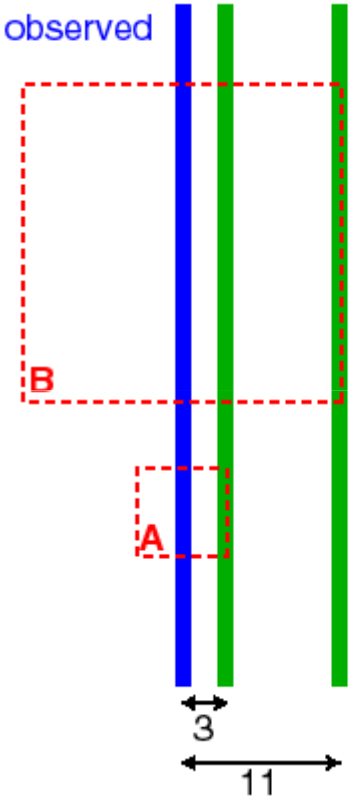
We want to know:

1. How forecast skill varies with neighbourhood size.
2. The smallest neighbourhood size that can be used to give sufficiently accurate forecasts.
3. Does higher resolution provide more accurate forecasts on scales of interest (e.g. river catchments)

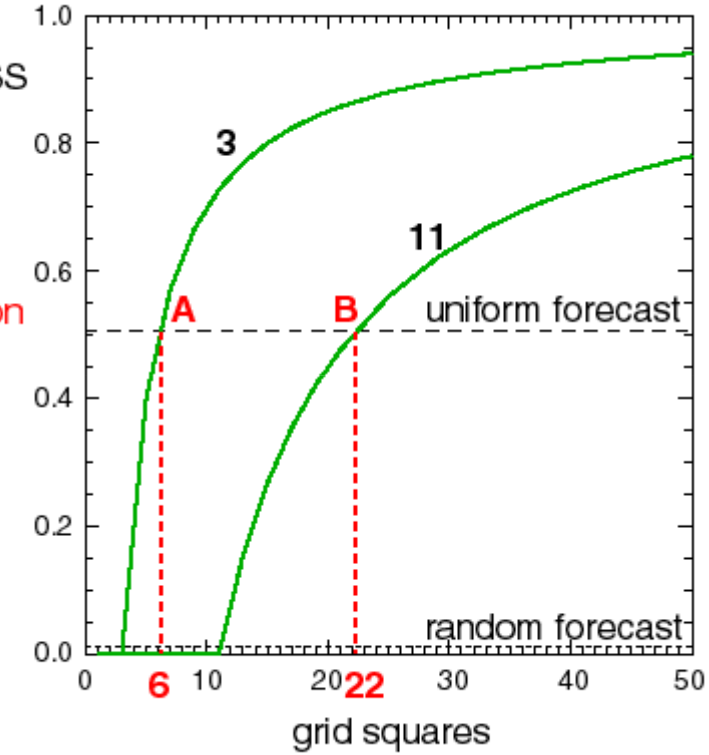
Compare forecast fractions with fractions from radar over different sized neighbourhoods (squares for convenience)

Use rainfall accumulations to apply temporal smoothing

Idealised example



twice the separation



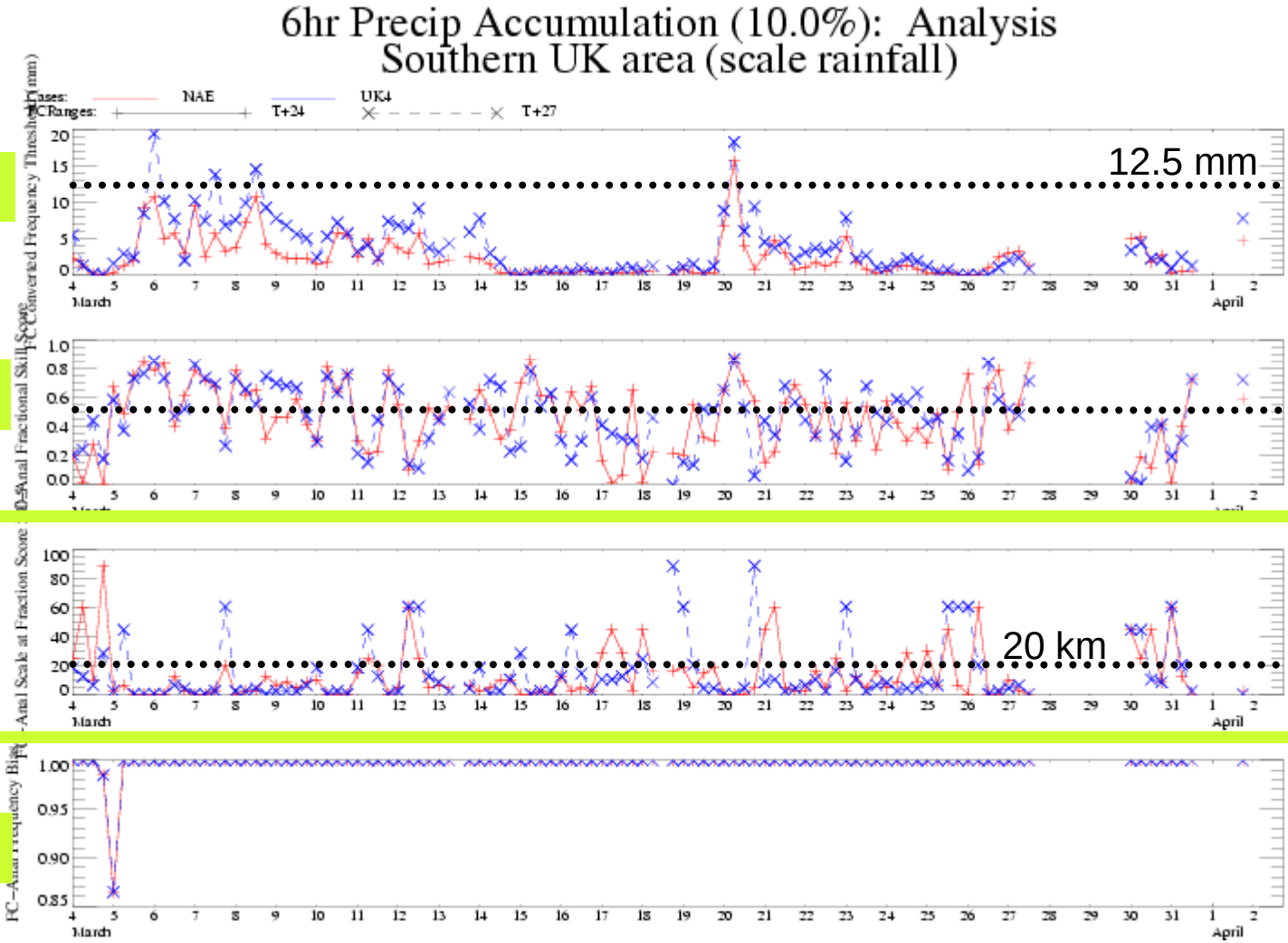
In summary

This verification method provides a way of answering **some** important questions about forecasts from ‘storm-resolving’ NWP models.

- How does forecast skill vary with spatial scale?
- At what scales are higher resolution forecasts more skilful (if any)?
- At what scales are forecasts sufficiently accurate?.....

(There are other questions that need different approaches)

How we are using it



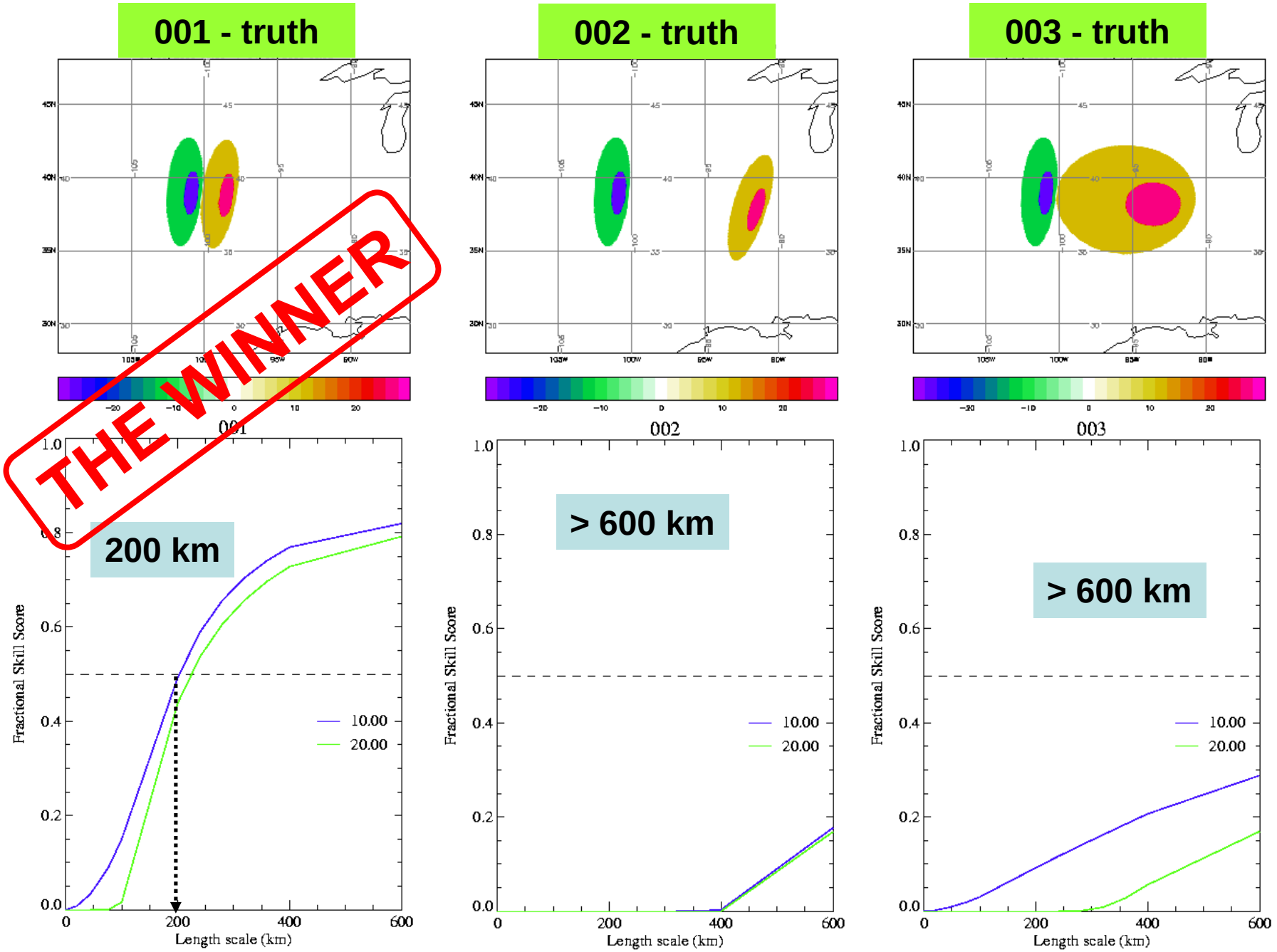
10% threshold

FSS

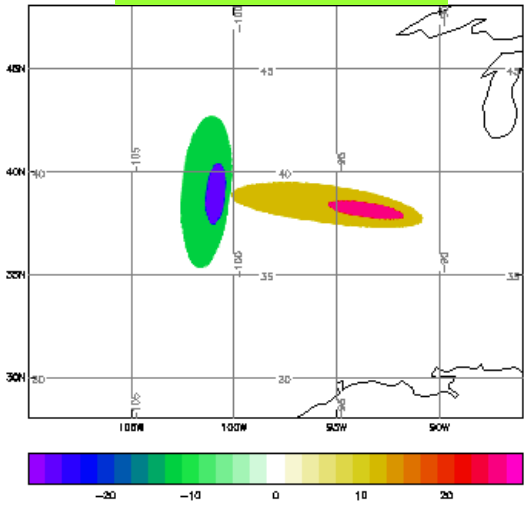
L(FSS>0.5)

Freq. Bias

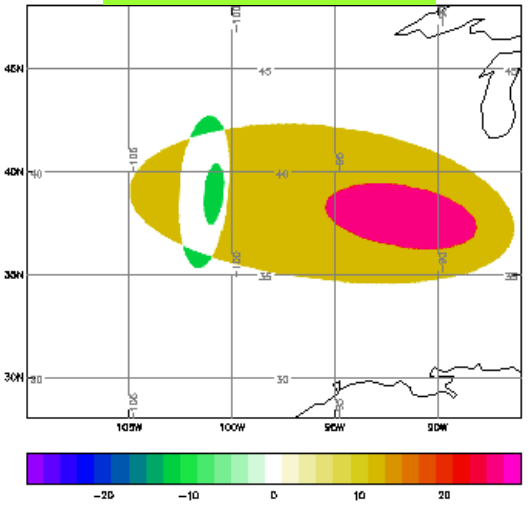
See poster by Mittermaier and Thompson

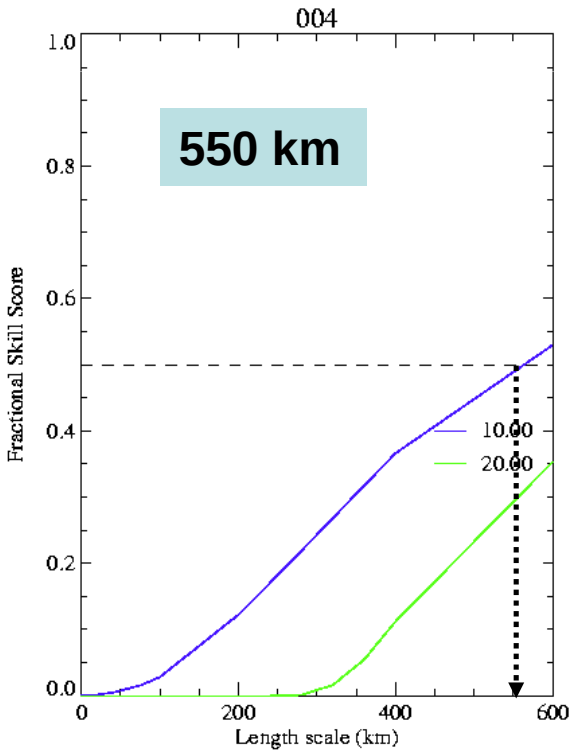

001 - truth
002 - truth
003 - truth
THE WINNER
001
002
003
200 km
> 600 km
> 600 km
Fractional Skill Score
Length scale (km)
10.00
20.00
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600

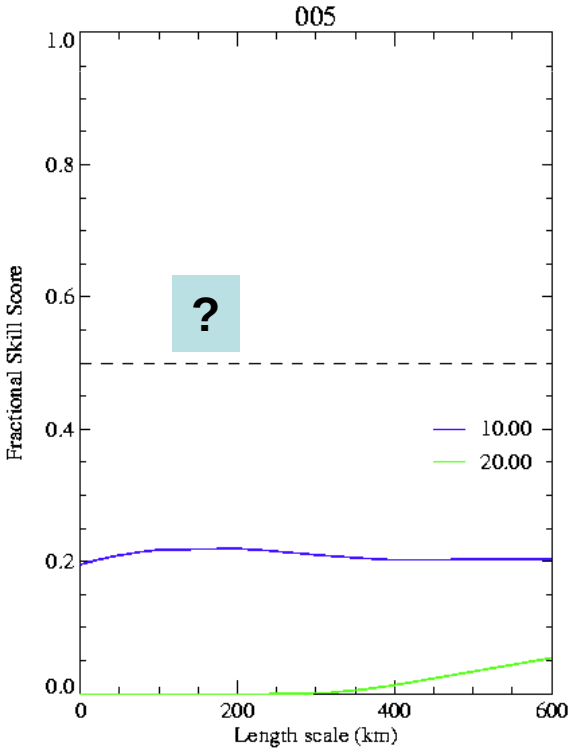
004 - truth



005 - truth





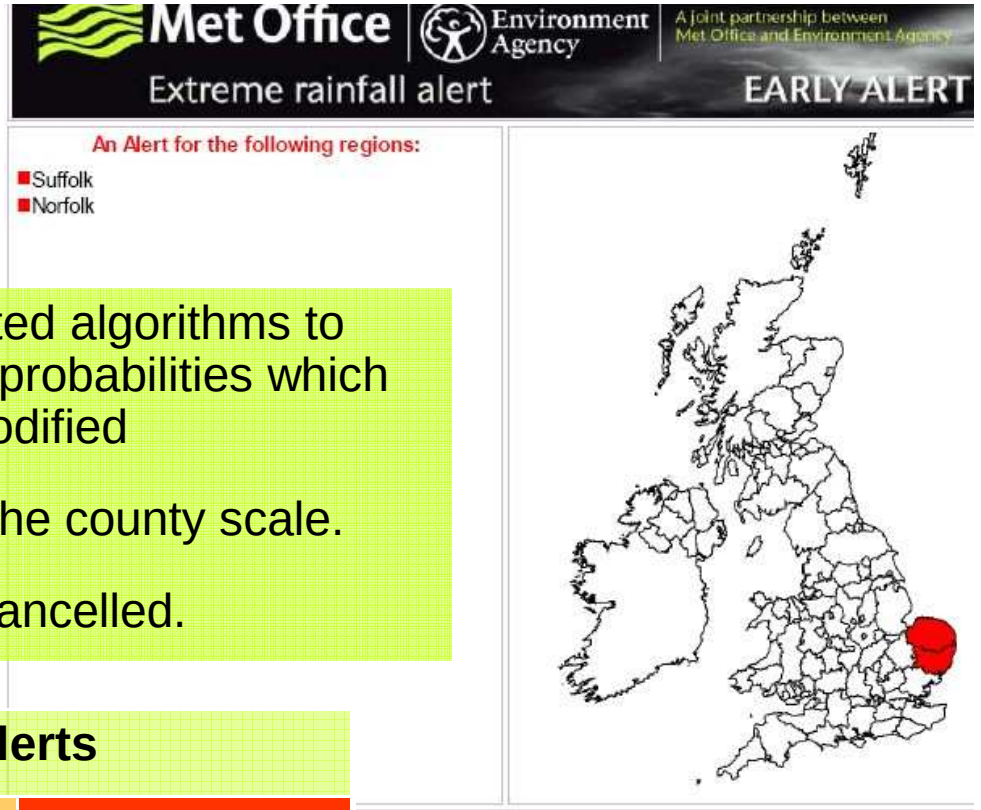


Summary : L(FSS > 0.5)

Case	> 10 mm
001	200 km
002	> 600 km
003	> 600 km
004	550 km
005	?

What are they?

- Based on sophisticated algorithms to generate first-guess probabilities which can be forecaster-modified
- Alerts are issued at the county scale.
- Can be updated or cancelled.



	Extreme Rainfall Alerts		
	Advisory	**Early**	**Imminent**
30 mm/h or 40 mm/3h or 50 mm/6h	Very low but prob >= 10%	Low with prob 20-40%	Moderate with prob >= 40%
Issued	14LT valid for the 24h starting from the next midnight	Lead time of 8 – 11h	Lead time of 1-3h

Two approaches have been considered …

- taking the **“event”** view, and

  (did an event occur anywhere in the alert area during the time that the alert was in force)

- taking the **“time series”** (continuum) view

  (comparing the county accumulation totals hour-by-hour during the time that the alert was in force to establish if the threshold was exceeded)

Caveat: both of these approaches are inherently deterministic

ET verification plans

- SRNWP EUMETNET comparison
  - QC-ed Results by end of year
  - Add others next year ?
- Agree on best (better ?) methods for high resolution forecast verification
- Link to operations – best methods of presenting forecasts, especially high resolution and EPS

Structure,S, Amplitude, A, Location , L (=L_1+L_2)

Wernli, Paulat, Hagen, Frei, 2008 (MWR)