Review of verification activities and developments

Clive Wilson –Expert Team on diagnostics, validation & verification

33rd EWGLAM/18th SRNWP meetings –Tallinn 10-13 October 2011

Contents

- SRNWP-V programme
  - Reading workshop ET – November 2010
  - Final reports stage 1
  - Stage 2 progress
- Consortia activities
  - COSMO
  - Hirlam – See Bent & Carl's talks Wed session
  - Met Office – Wed Session
- ECMWF TAC Subgroup on verification
  - SEEPS precipitation score

SRNWP-V Programme

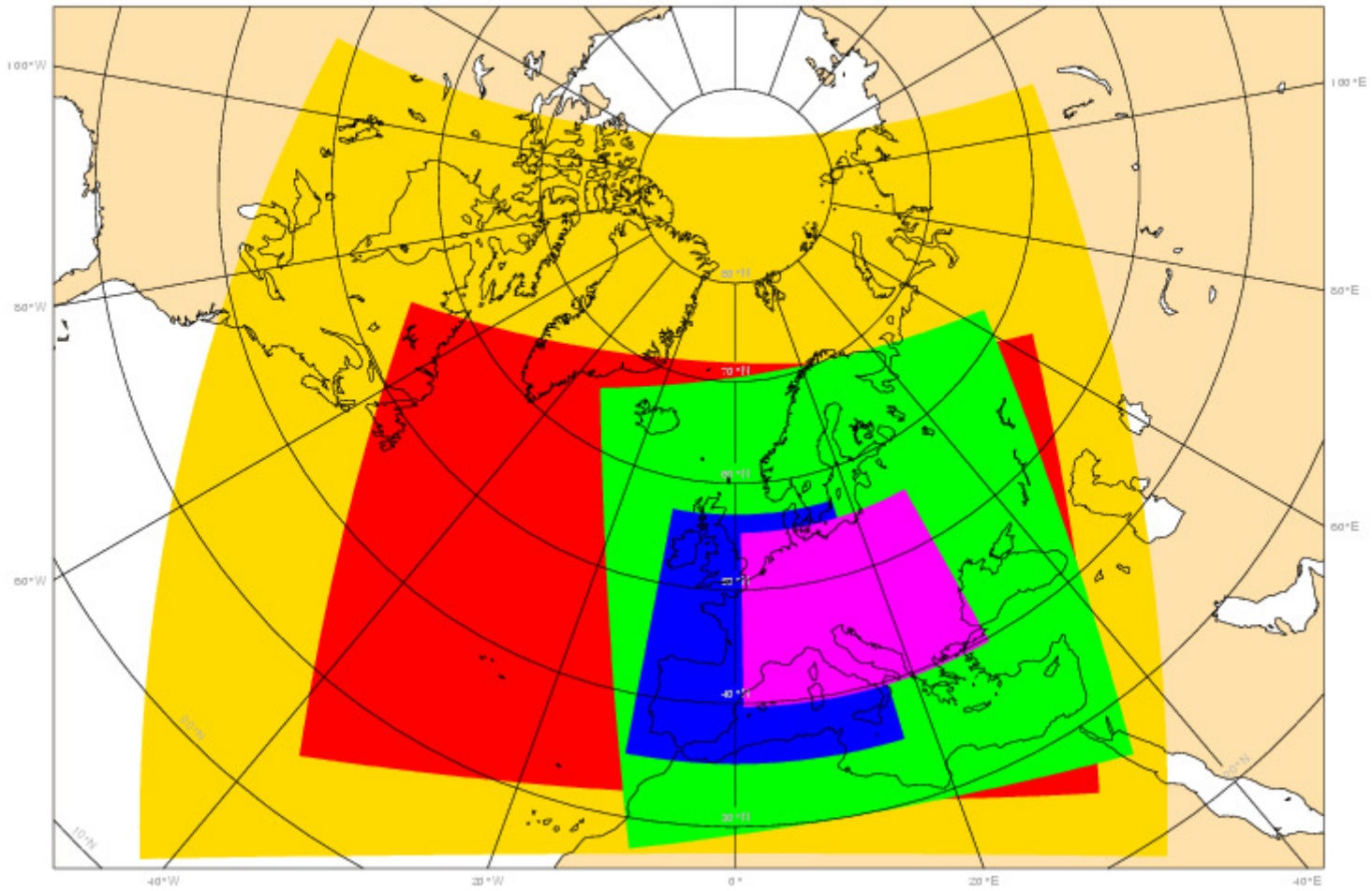
Desire to benchmark operational models cf CBS global exchange

- Phase 1 2009-2010
  - Establish framework, recommend methods
- Phase 2 2011-2012
  - Continue comparison, more parameters
  - Use radar/extra observations
  - Extremal measures

EUMETNET/SRNWP programme - Deliverables

- **D1: Operational verification comparison of deterministic forecasts from one version of each of the 4 regional models of Europe (available for all the participating members)**
- **D2: Additional intercomparison of other versions of the consortia models including high resolution models**
- **D3: Inventory and recommendations of "new" scale-selective verification methods.**
- **D4: Catalogue of sources of non-GTS data**
- **D5 Exchange methods and code for verification of severe weather forecasts**

Domains of 5 consortia reference models



15km

12km

7km

10km

5 km

(10km)

Hirlam UM COSMO ALADIN Aladin-Lace

Parameter	Scores
Mean sea level pressure	mean bias and root mean square errors
2m temperature	Bias, rmse
2m relative humidity	Bias, rmse
10m winds	mean bias speed error and root mean square vector wind error
6 hourly total precipitation	equitable threat score and frequency bias for 0.5, 1.0 and 4.0 mm $6h^{-1}$

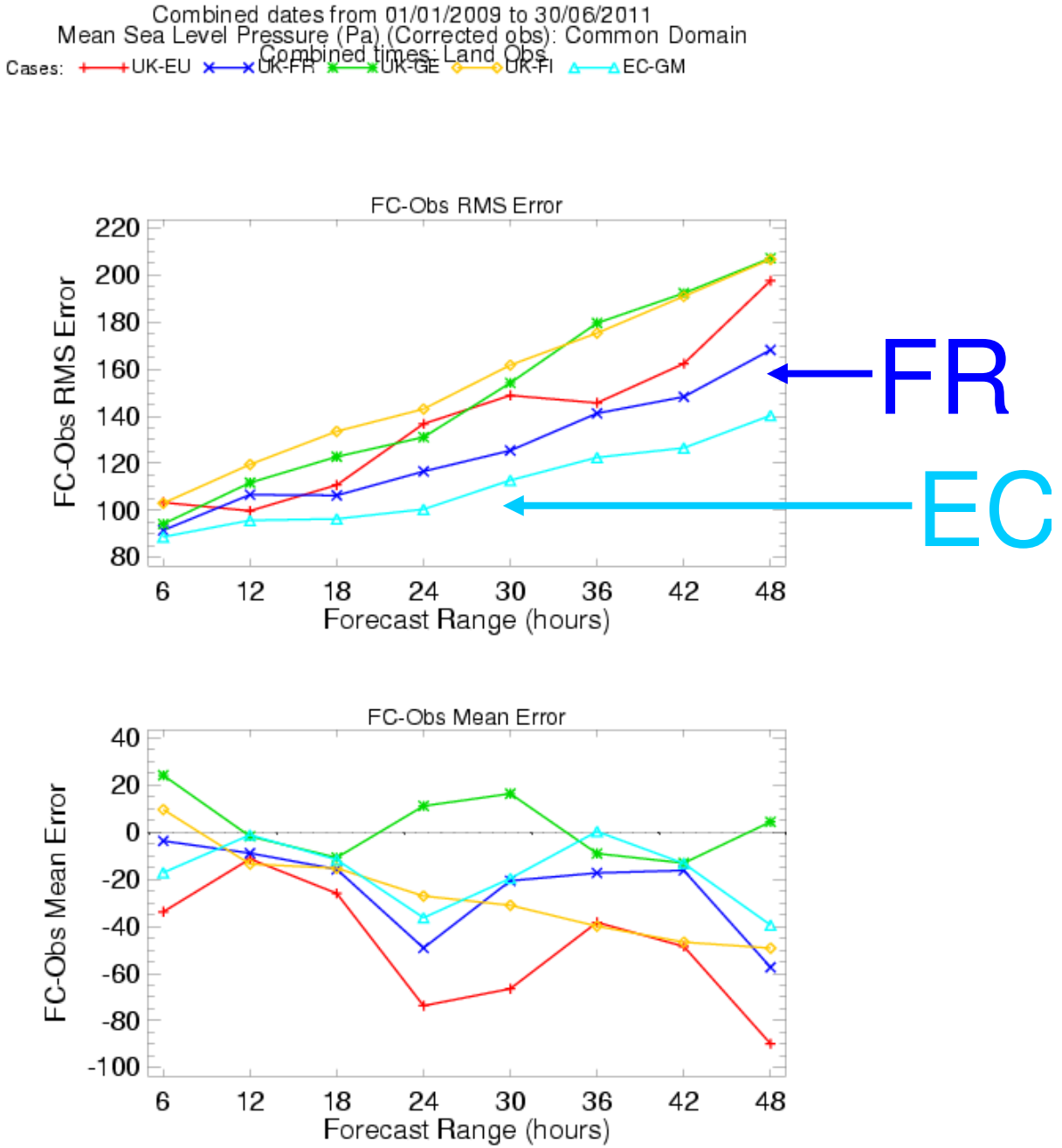
Key

Model	Label
Hirlam reference run by FMI	UK-FI
Aladin-France run by Meteo-France	UK-FR
COSMO Europe run by DWD	UK-GE
The North Atlantic European configuration of the Unified Model run by the Met Office	UK-EU
Aladin-Czech (LACE) run by CHMI	*UK-LC*
ECMWF high resolution global model	EC-GM

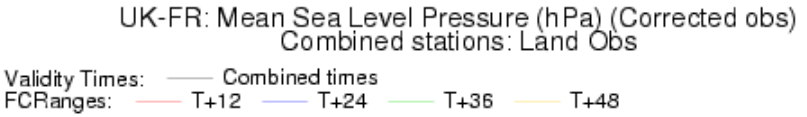
Comparison over ALADIN-France domain unless otherwise stated

Sea level Pressure

Mean 01/2009-06/2011

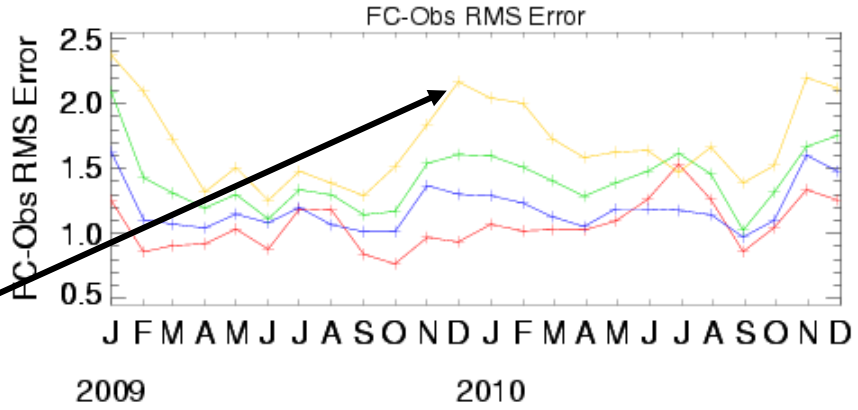


Sea level Pressure

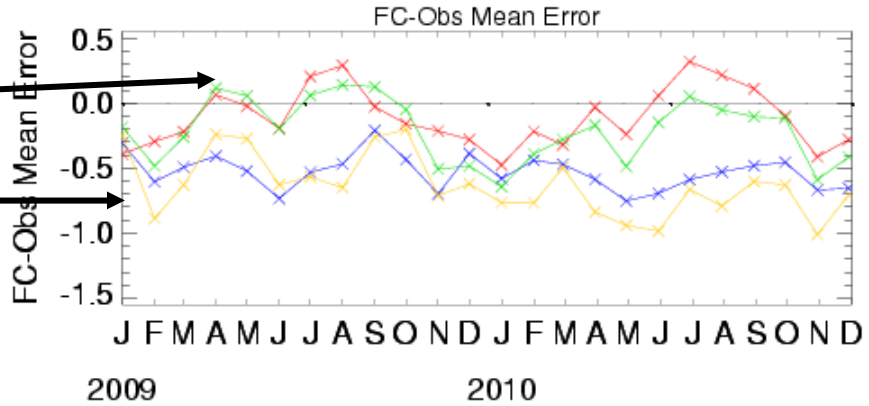


ALADIN-FR

- ONLY 00Z forecasts exposes diurnal variation
- Winter errors larger



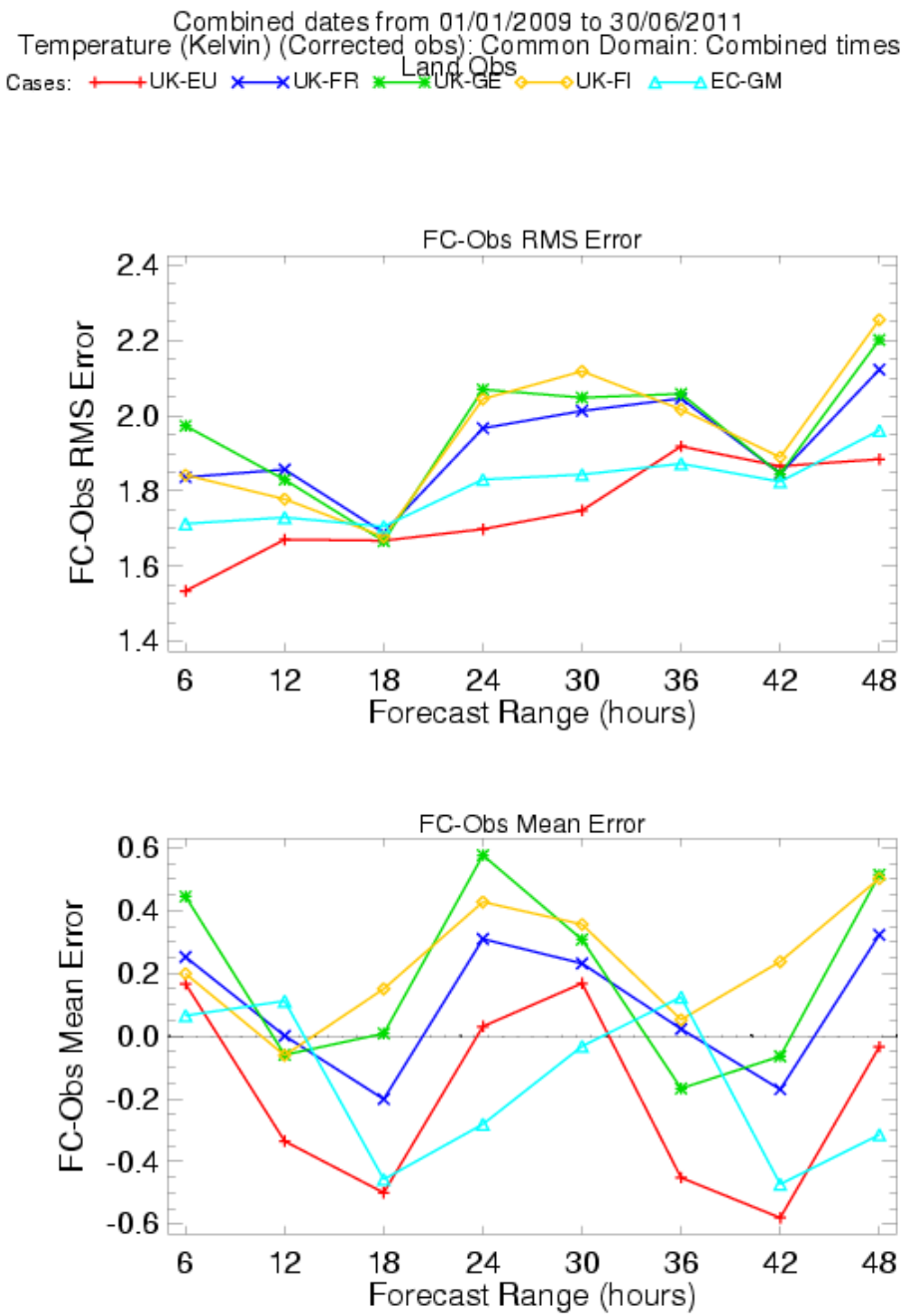
- Smaller daytime bias summer
- Larger bias at night



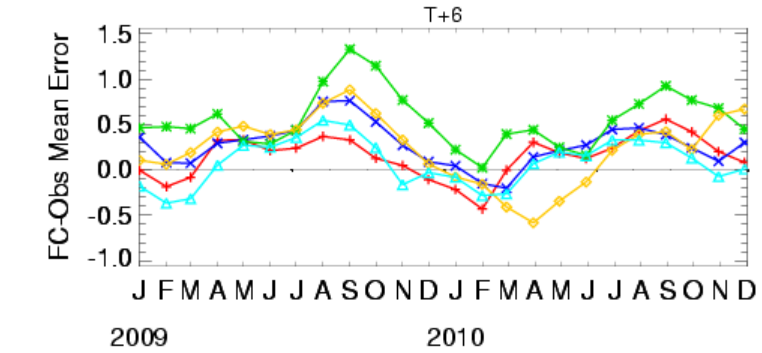
Screen temperature

Mean 01/2009-06/2011

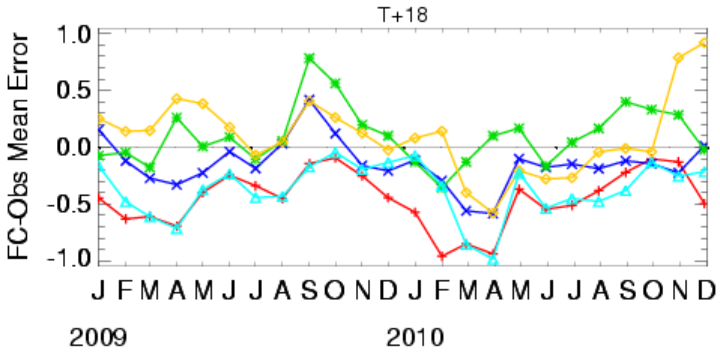
Warm night, cold day bias tendency



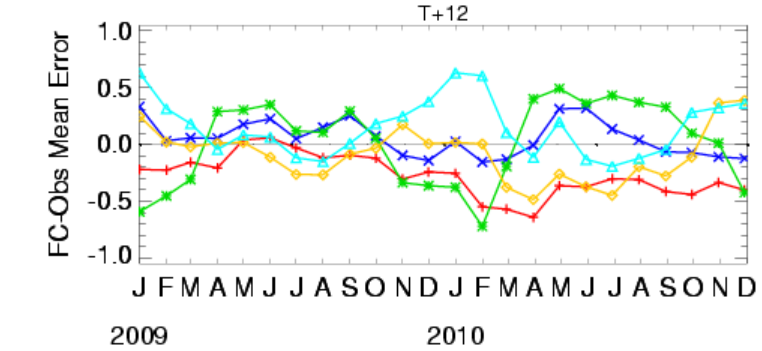
Night



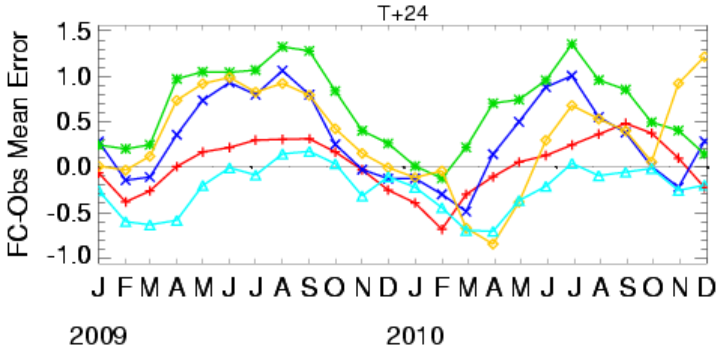
day



day



Night

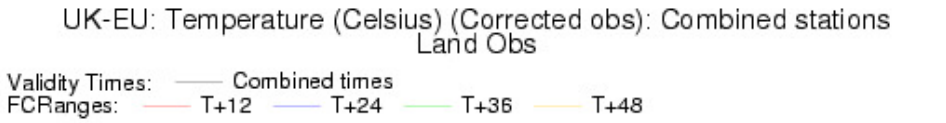


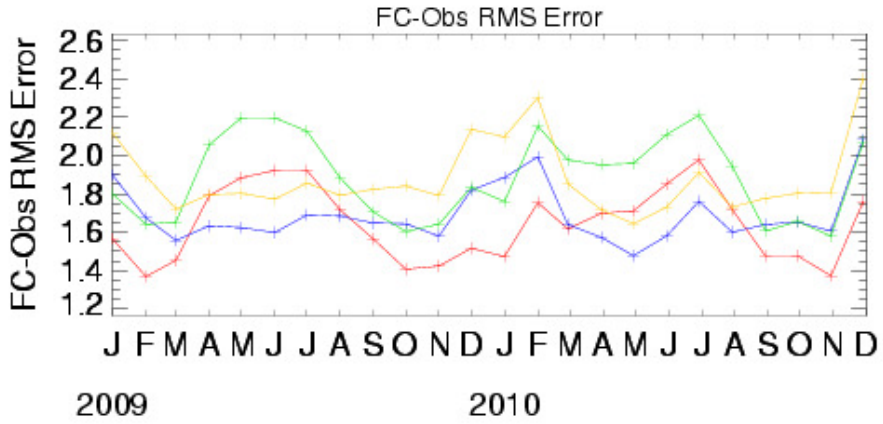
Screen temperature

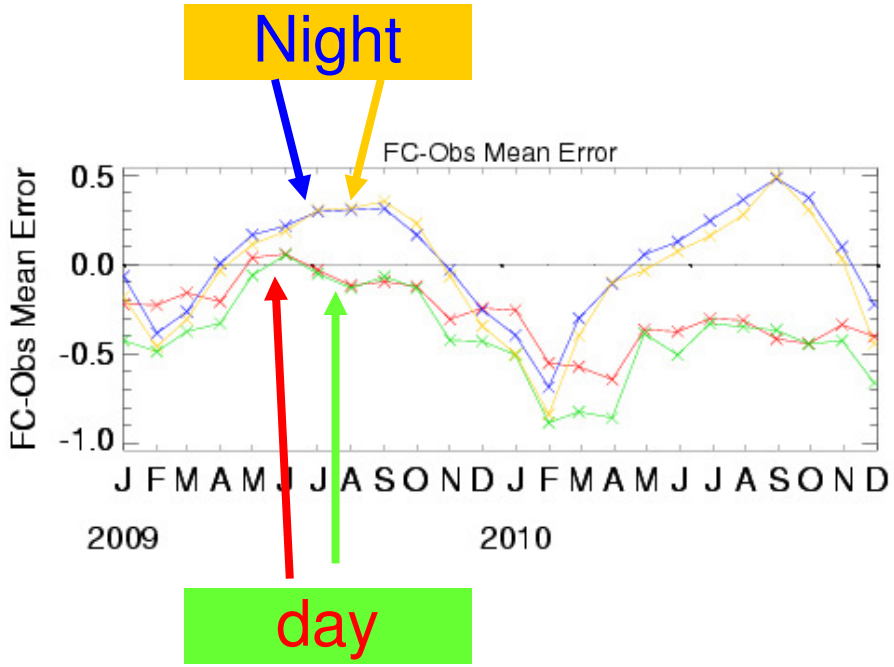
Screen temperature

NAE

Warm night, cold day bias tendency more apparent in summer



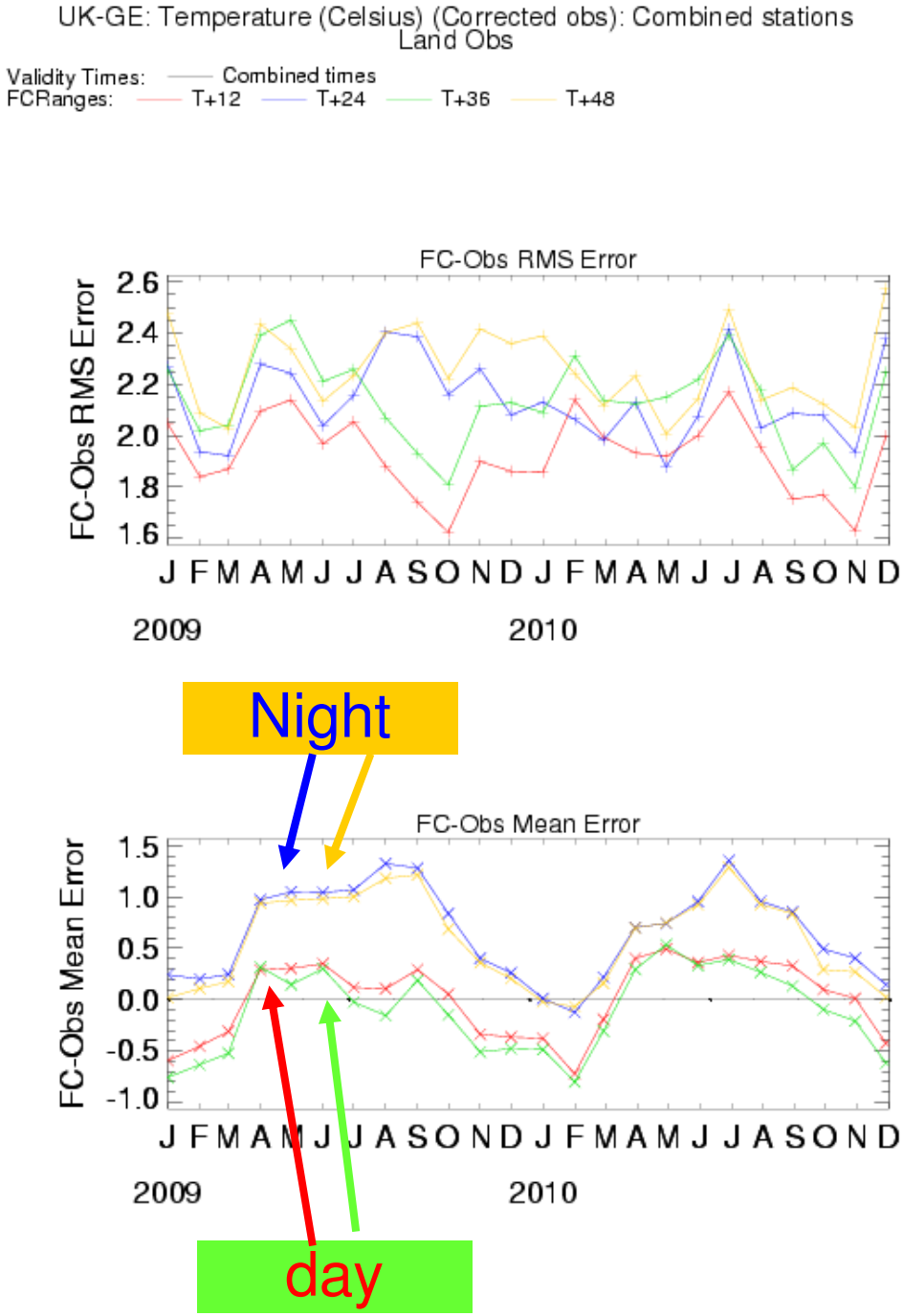




Screen temperature

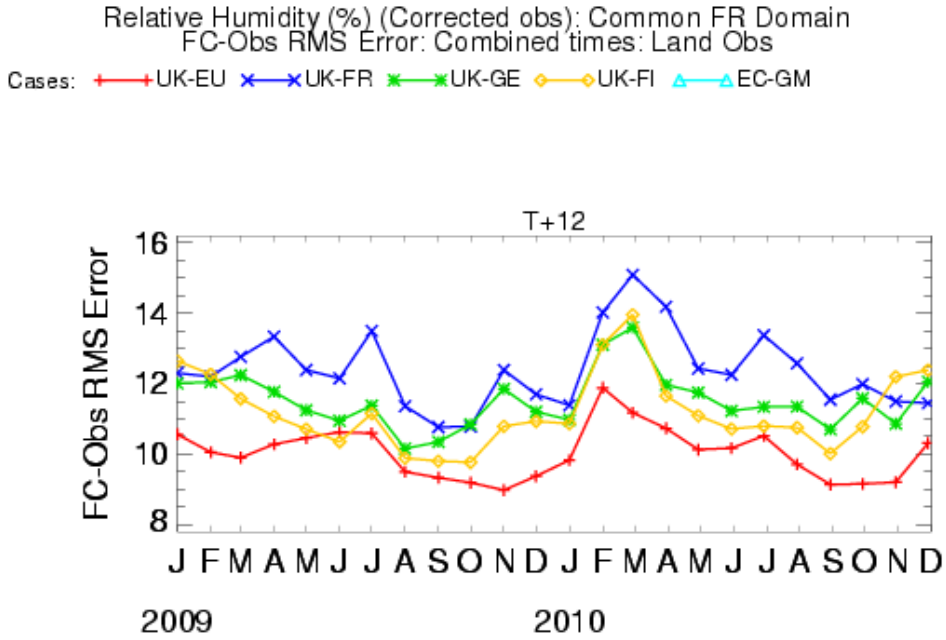
COSMO-EU

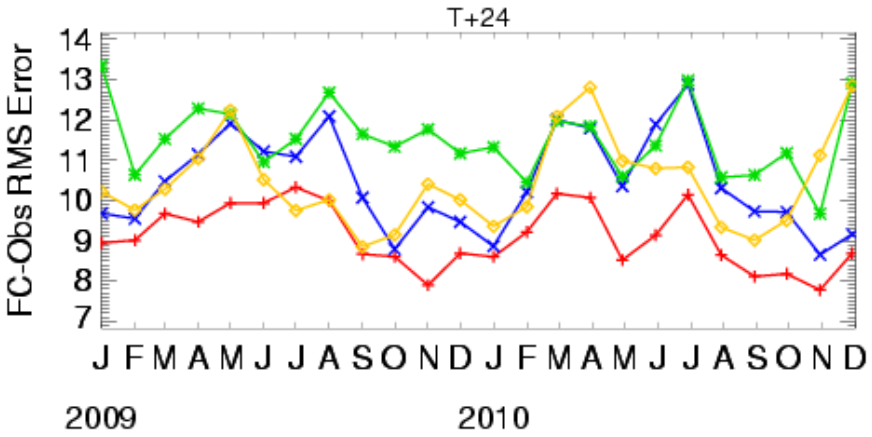
Warm night, cold day bias tendency more apparent in summer



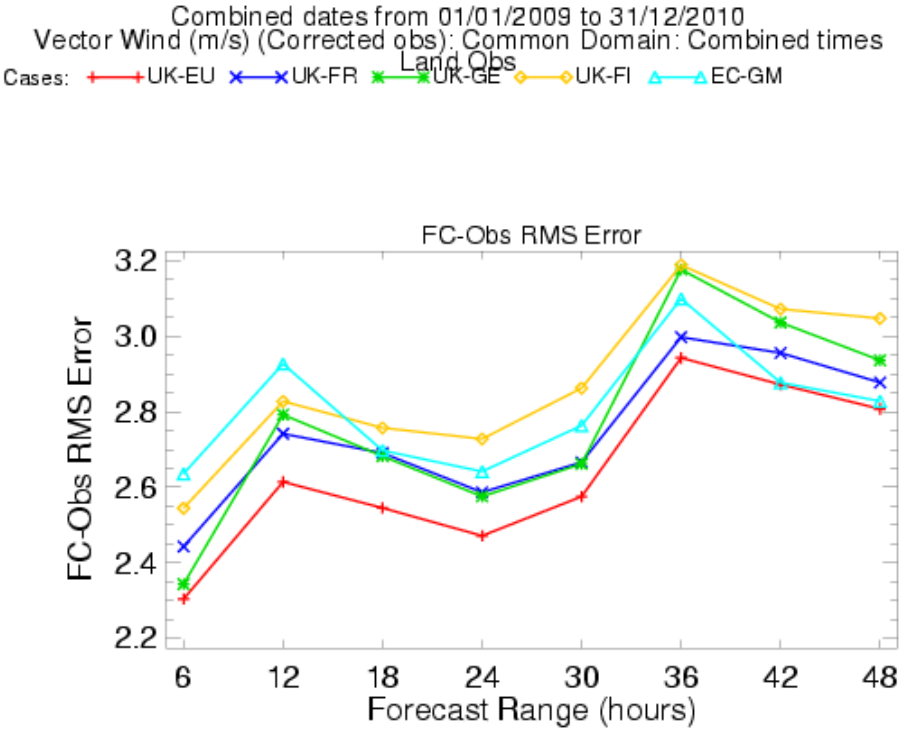
Relative humidity

NB screen temperature and humidity assimilated in NAE

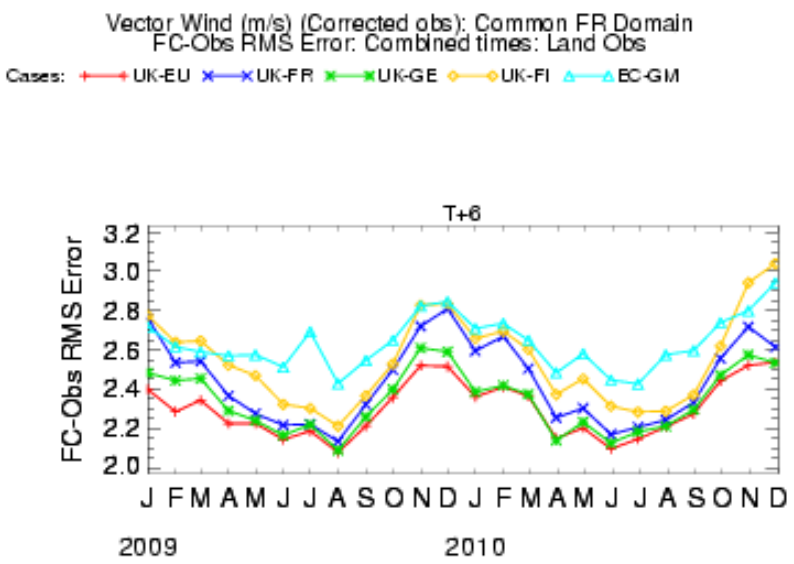


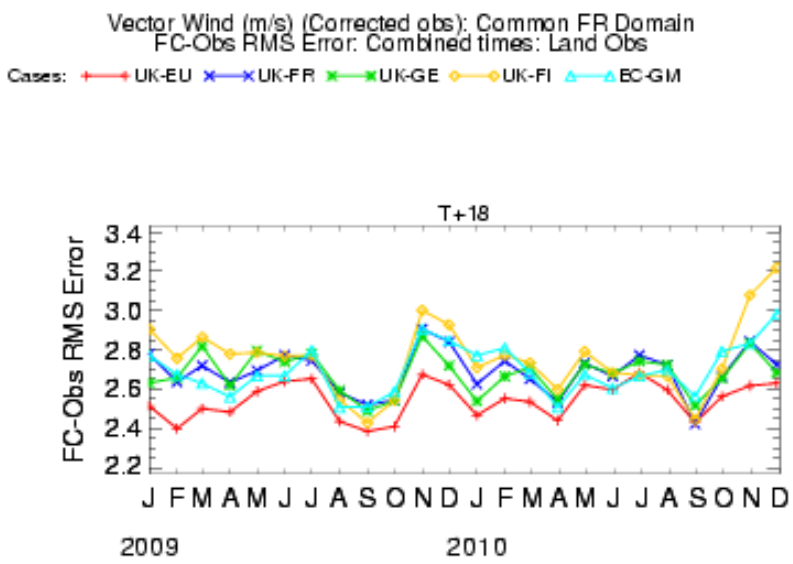


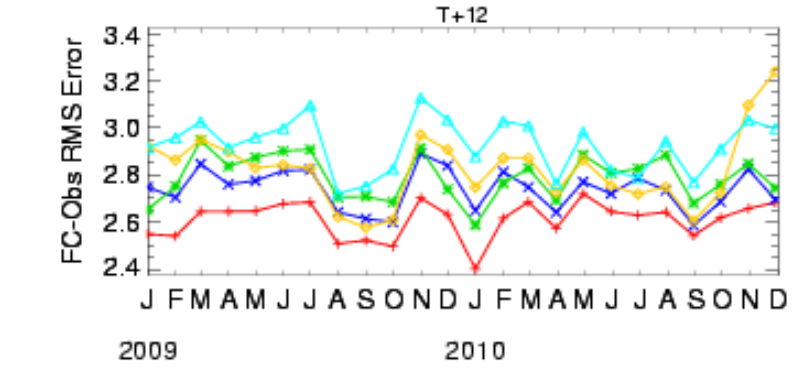
10m vector wind rmse

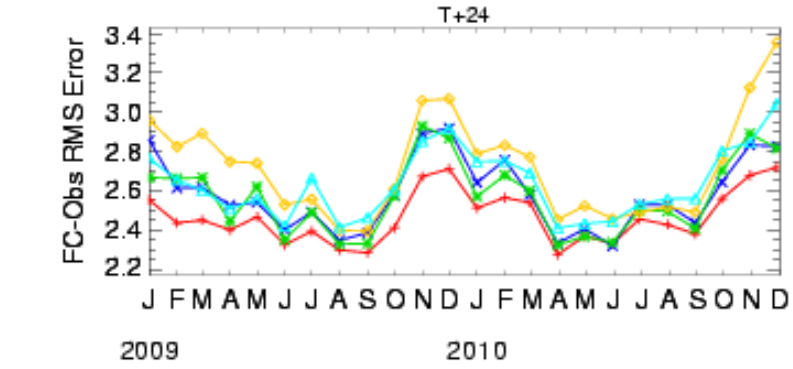


Regional models better than global especially shorter range

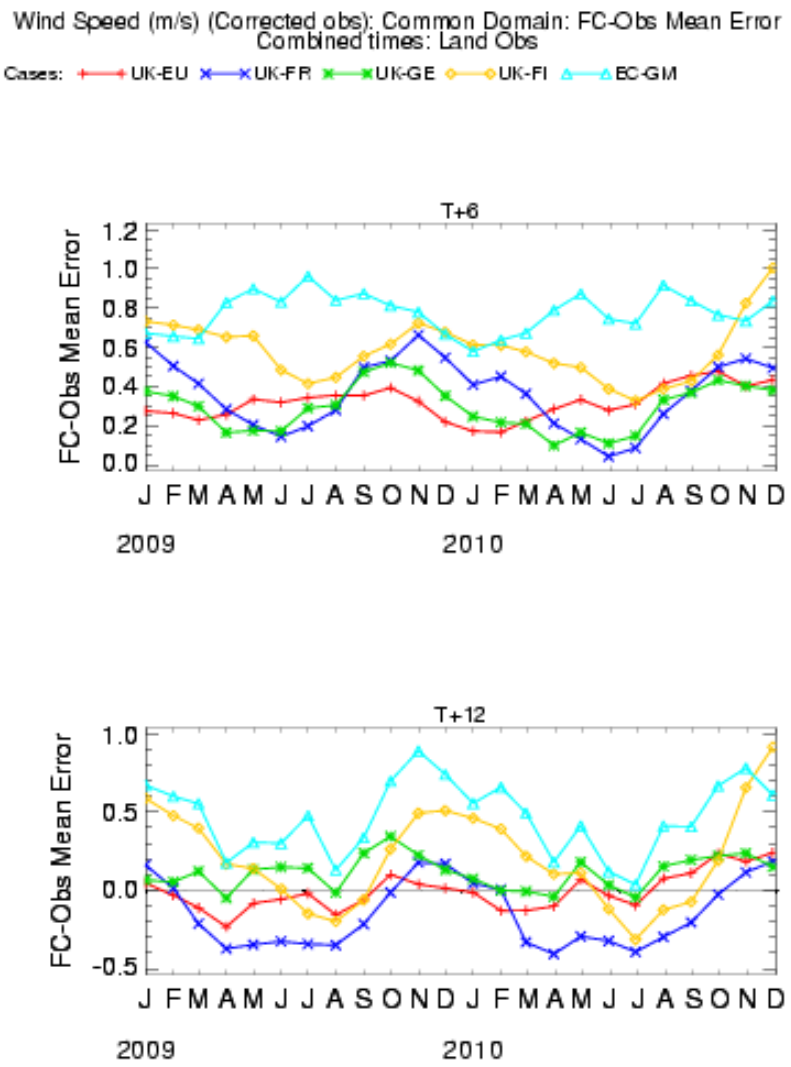


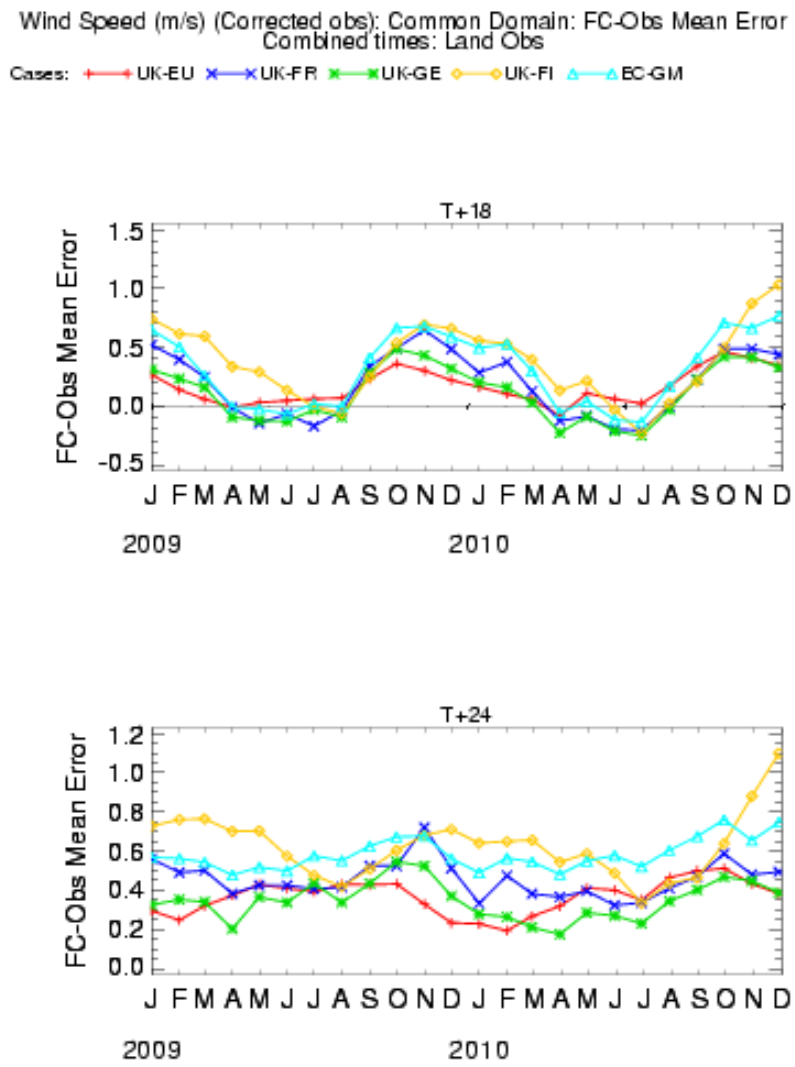






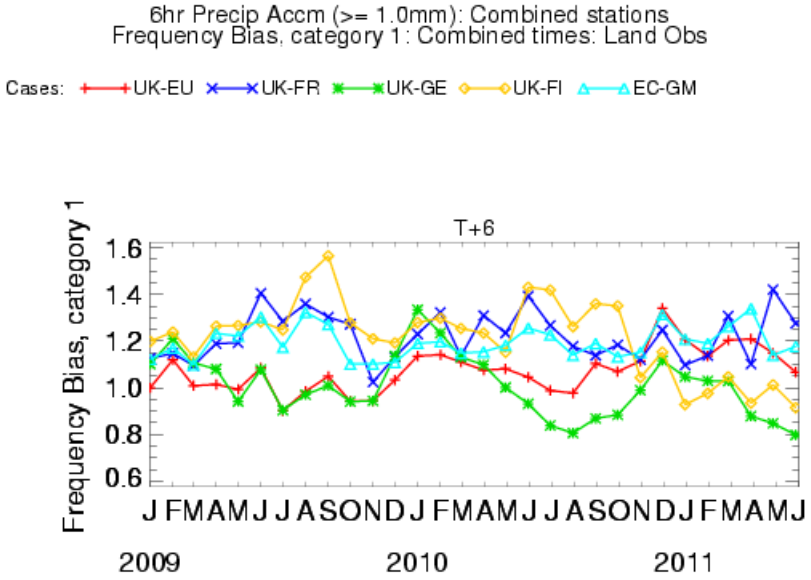
Seasonal variation more evident at night

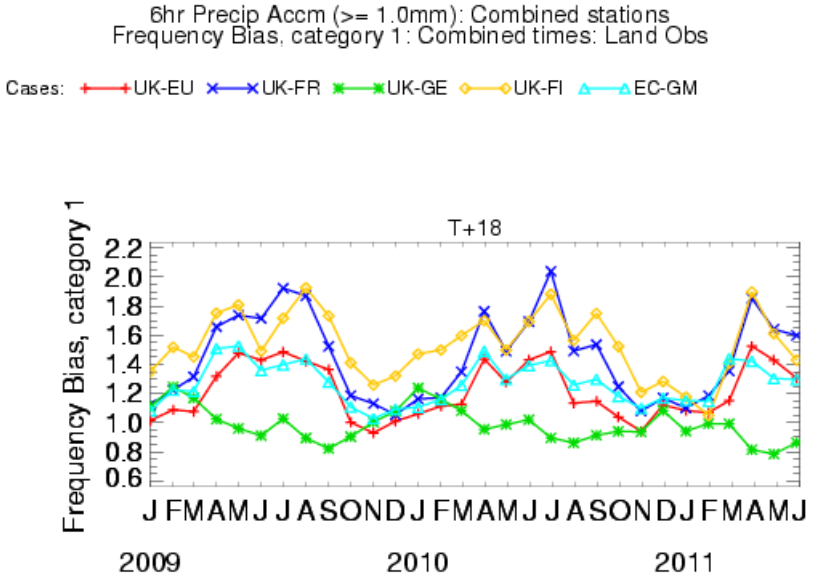


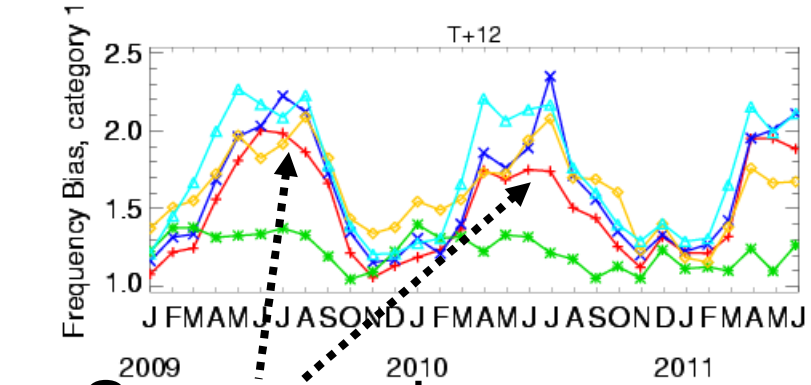


Mostly positive speed bias

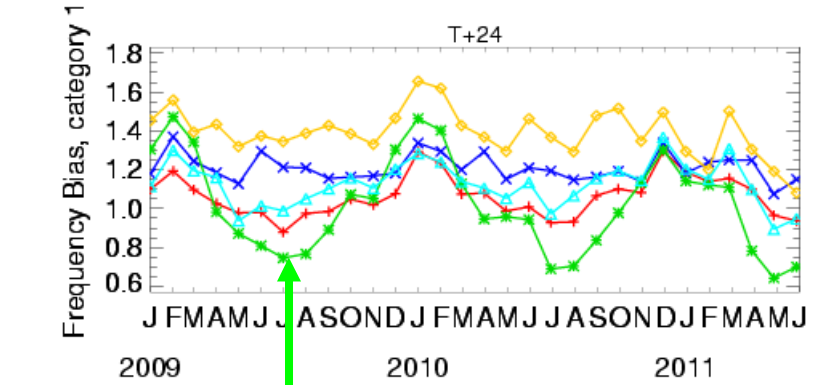
ppn frequency bias >1mm/6h





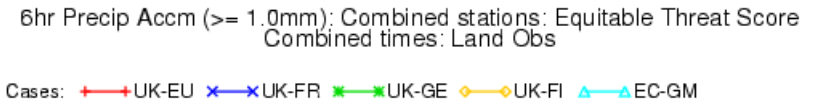


Summer day bias,convection

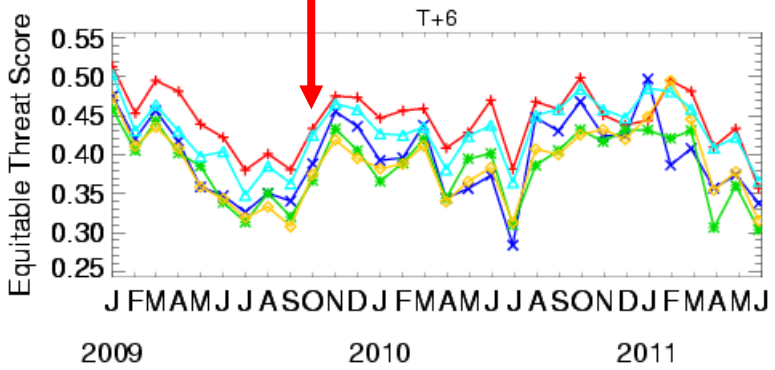


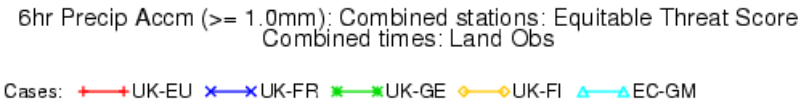
Night negative bias

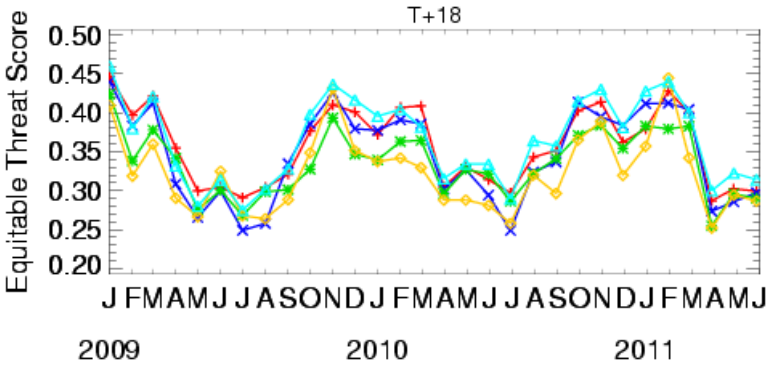
ppn ETS >1mm/6h

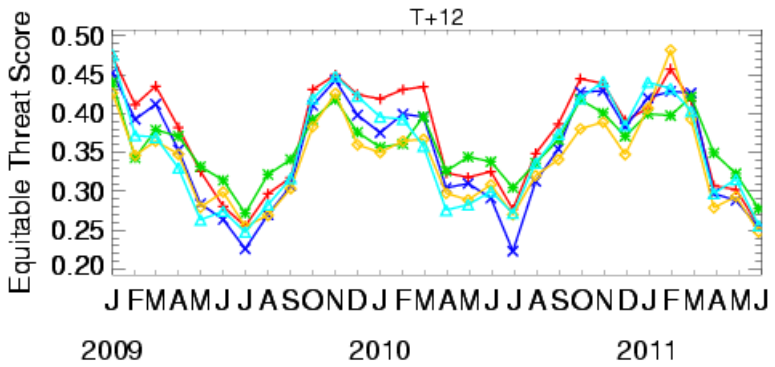


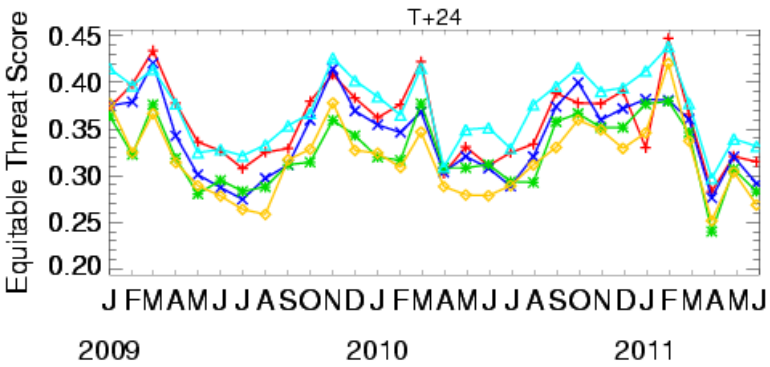
DA – radar/cloud







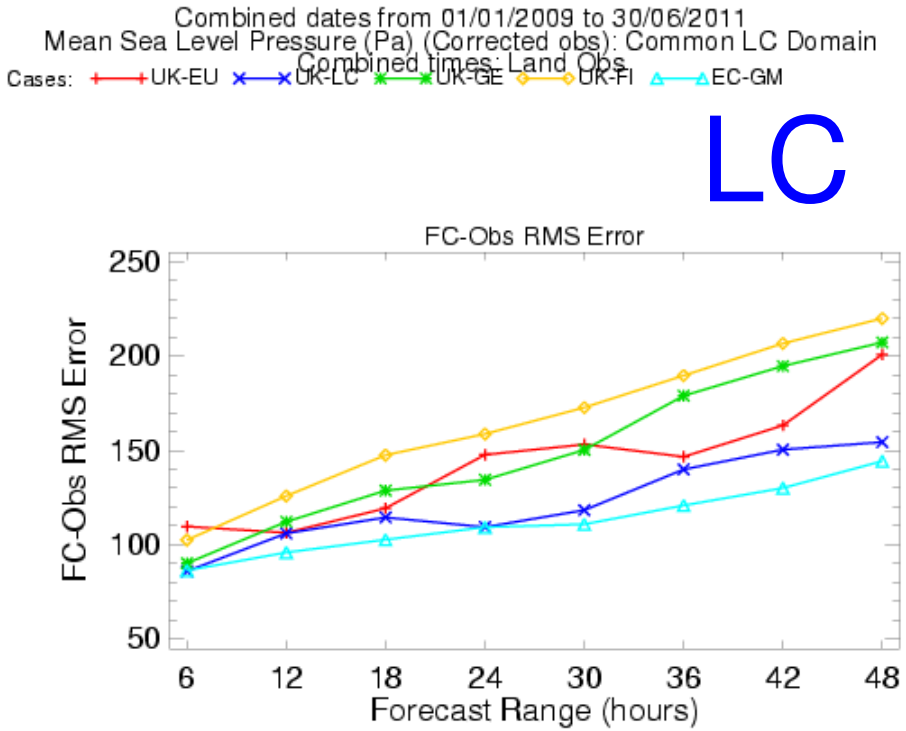




Sea level Pressure

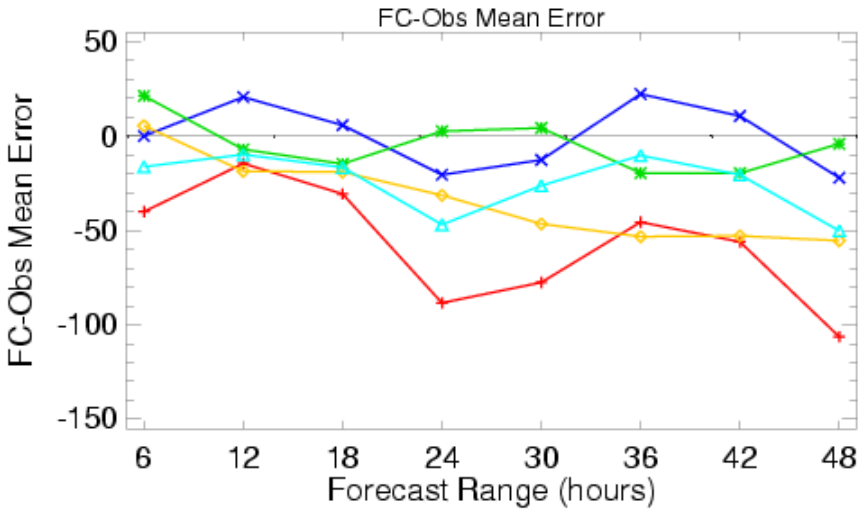
ALADIN-LACE domain

Mean 01/2009-06/2011



LC

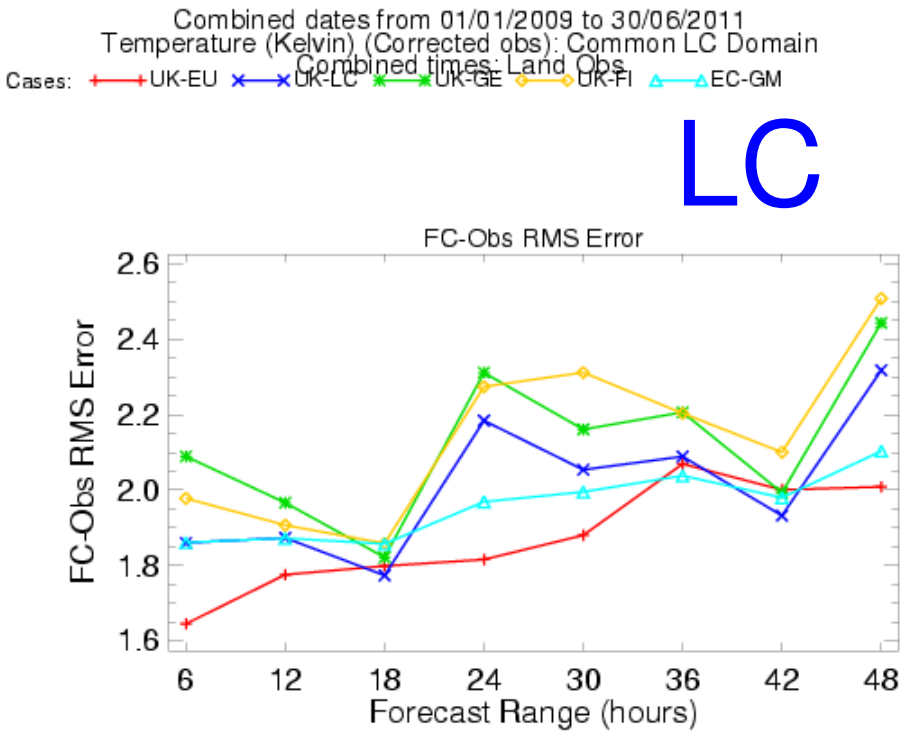
-EC

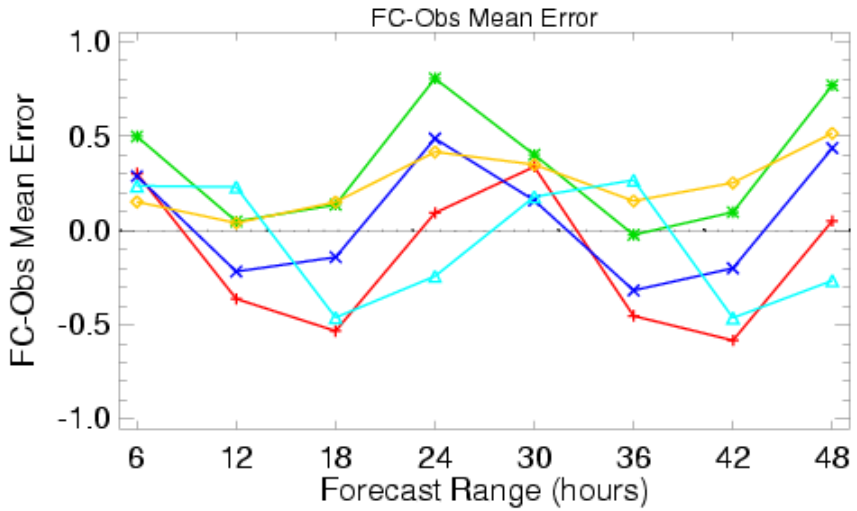


Screen temperature

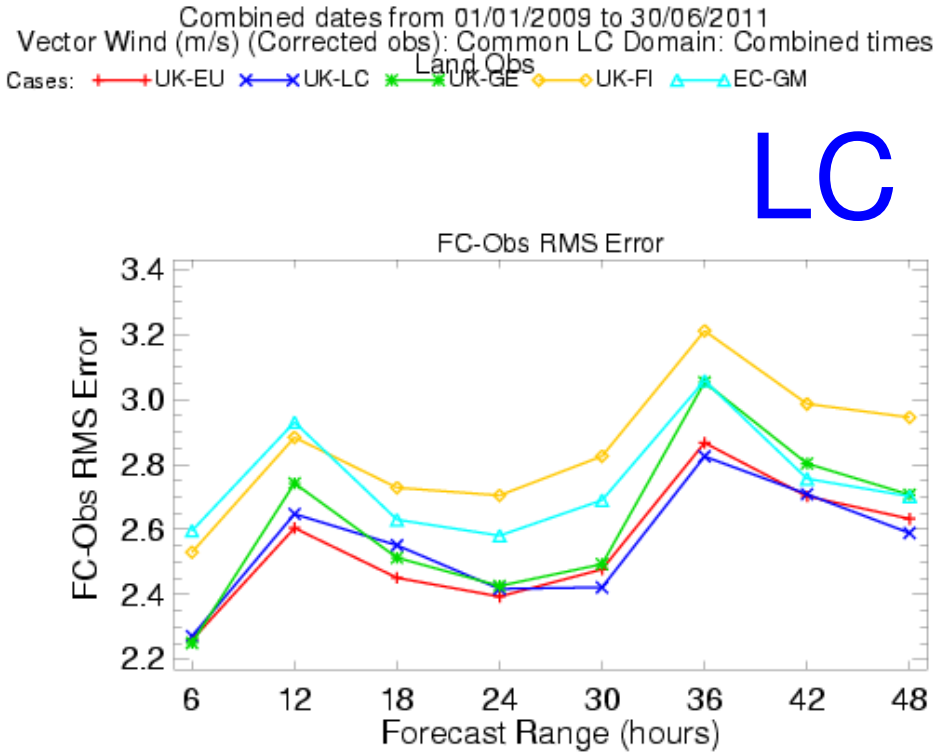
ALADIN-LACE domain

Mean 01/2009-06/2011





10m vector wind rmse ALADIN-LACE domain



Recommendations of “new” scale-selective verification methods.

- Studies/experience at Met Office, MeteoSwiss, DWD,Meteo-France
  - Fraction skill +variants (Roberts, Amodei & Stein)
  - Upscaling(+ETS etc)
  - Easiest to use & interpret

  - Structure/amplitude/location (Wernli et al)
  - Scale intensity (Casati)

  - Key aim of phase 2 to apply these to forecasts from reference/new higher resolution operational models

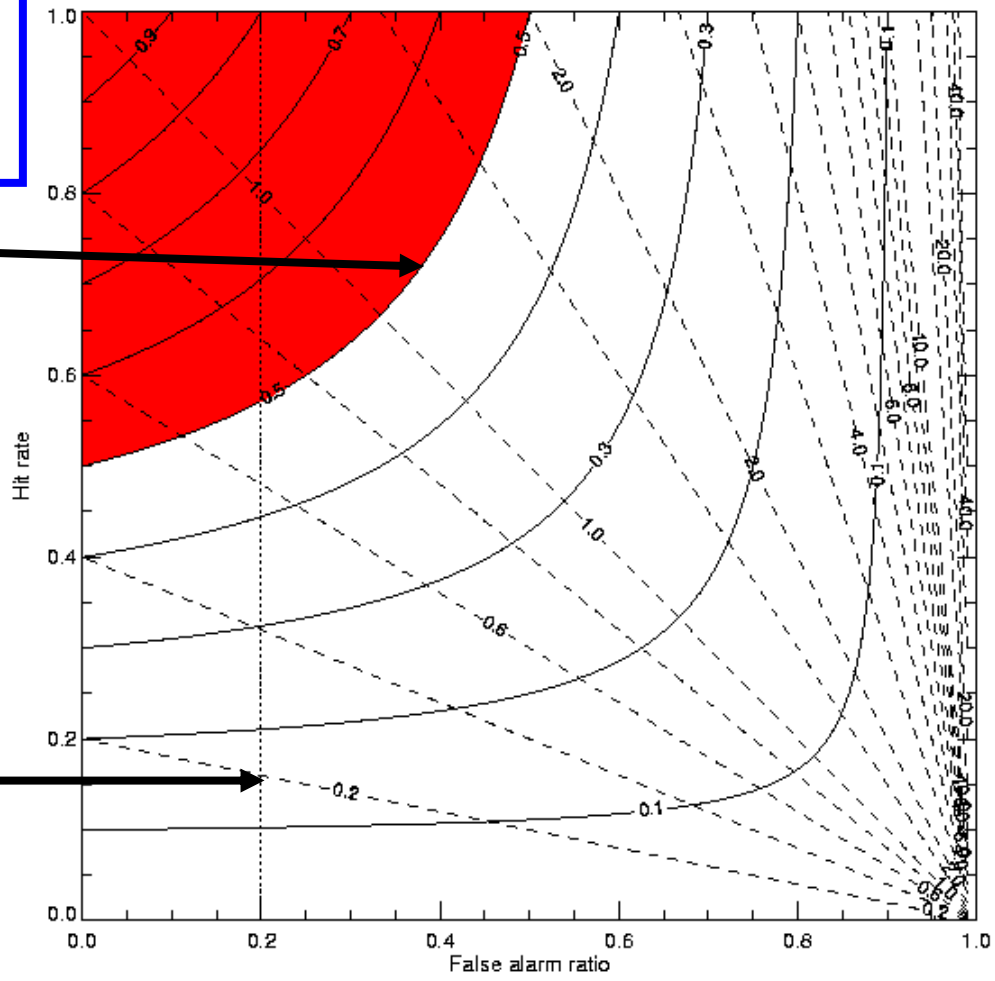
Hit rate v False alarm ratio-cf Roebber – WAF, 2009

$$TS = \frac{(1 - FAR)H}{1 - FAR(1 - H)}$$

$$B = \frac{H}{1 - FAR} \quad ---$$

Deterministic limit

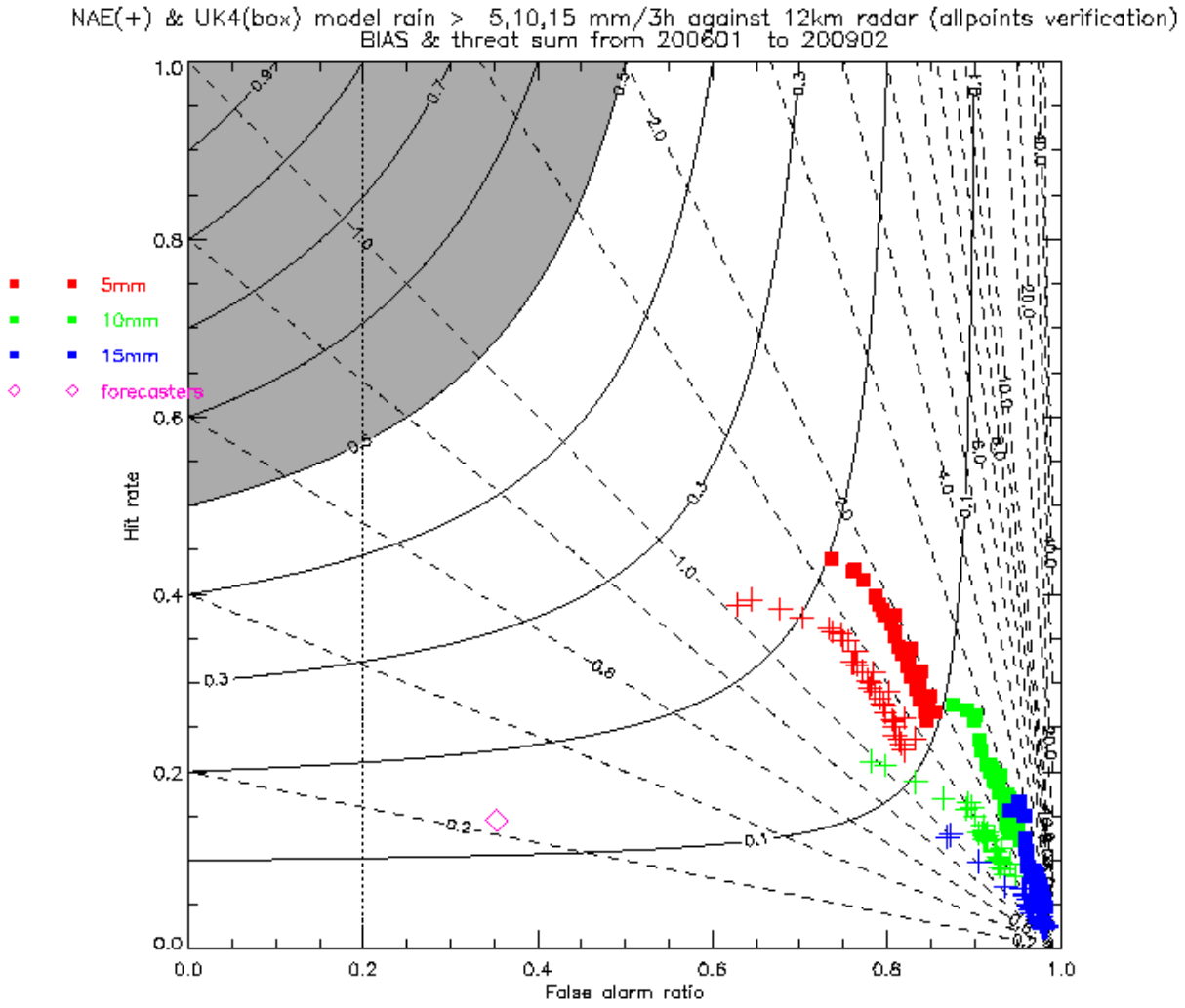
Confidence limit



NAE(12km) & UK 4km models (12km radar verification) 200601-200902

12km +

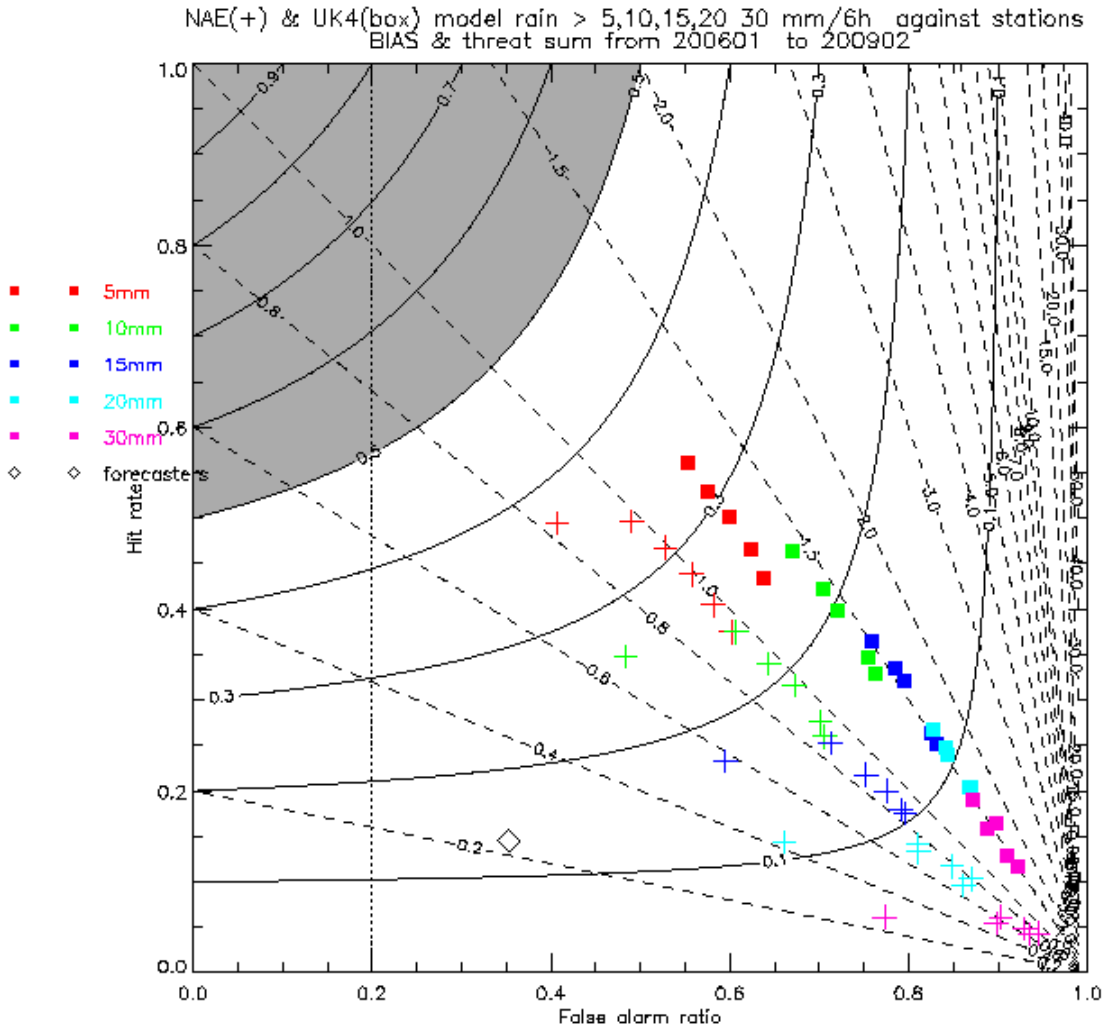
4km ■



NAE(12km) & UK 4km models (Nearest point to station verification) 200601-200902

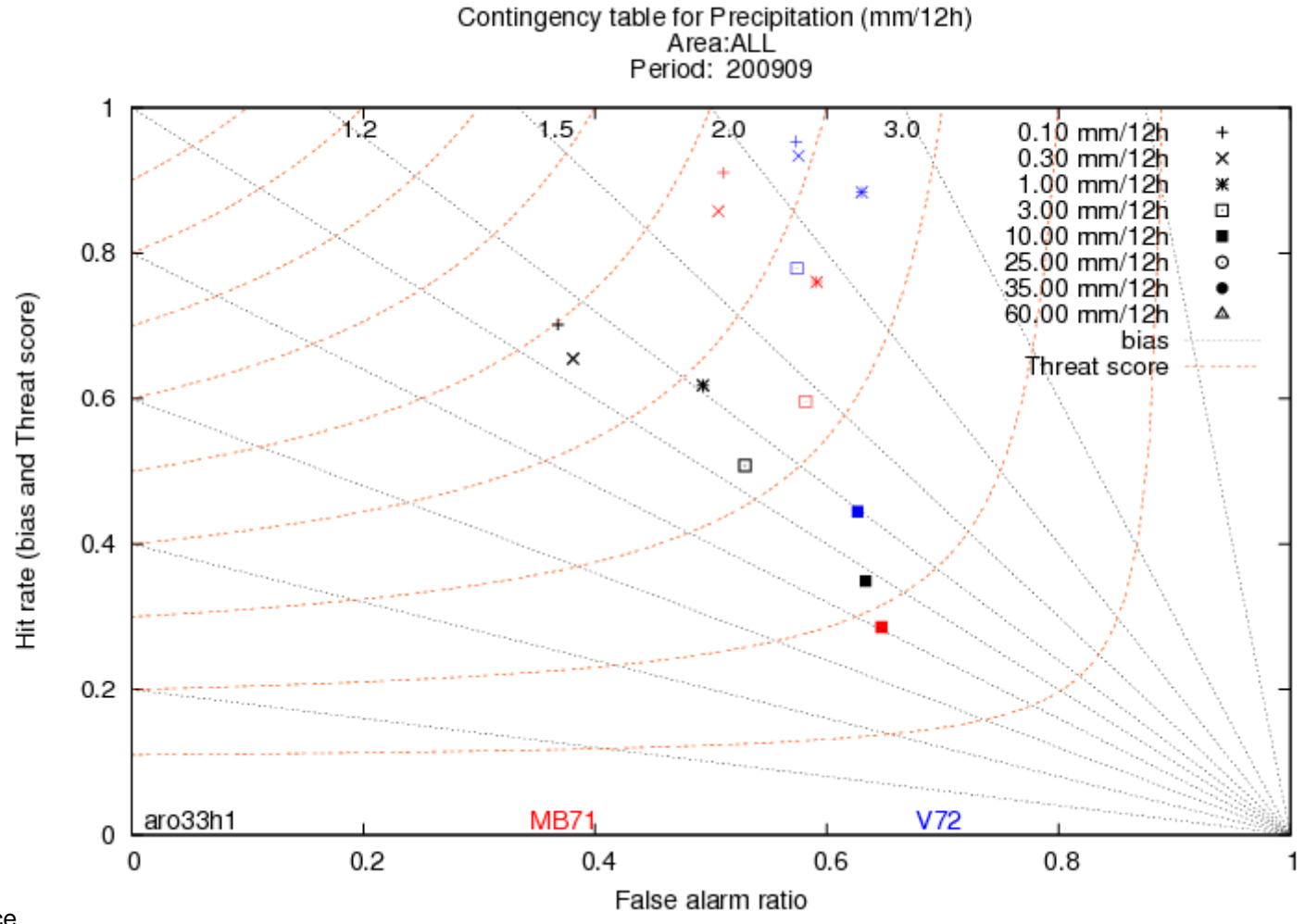
12km +

4km ■



FMI comparison

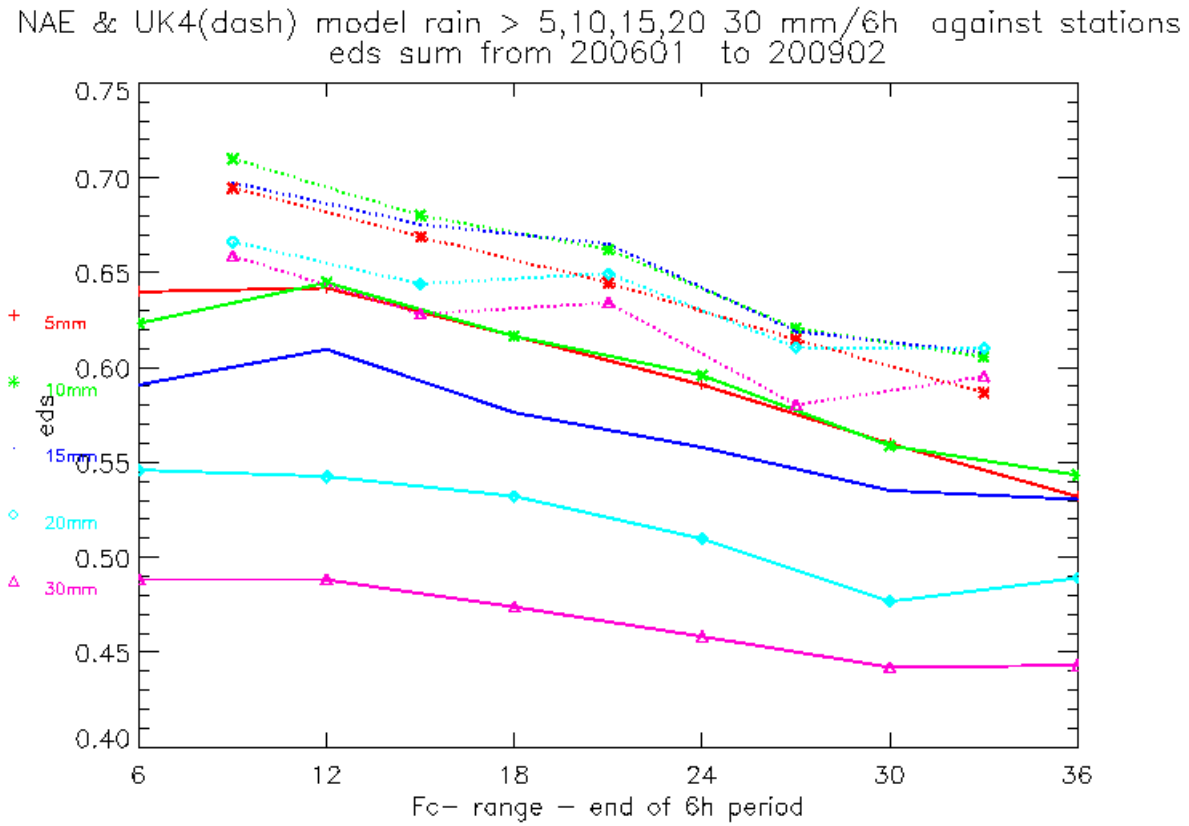
aro33h1: AROME 33h1 (2.5km L40)
MB71: HIRLAM 7.1.4 (7.5km L60)
V72 (RCR): HIRLAM 7.2 (16.5km L60)



Extreme dependency score

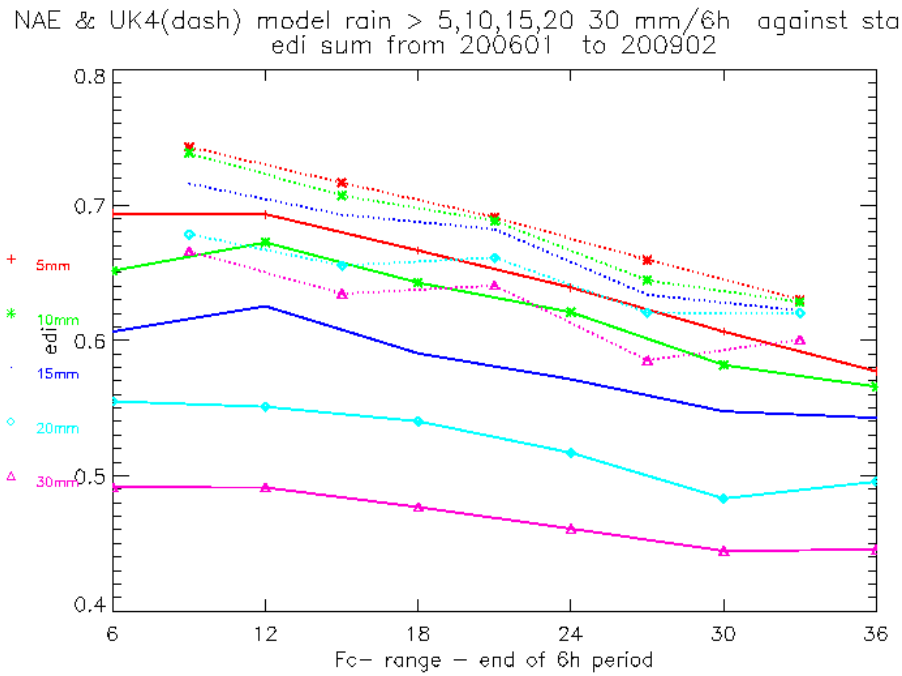
$$EDS = \frac{\log(p) - \log(H)}{\log(p) + \log(H)} \rightarrow EDI = \frac{\log(F) - \log(H)}{\log(F) + \log(H)}$$

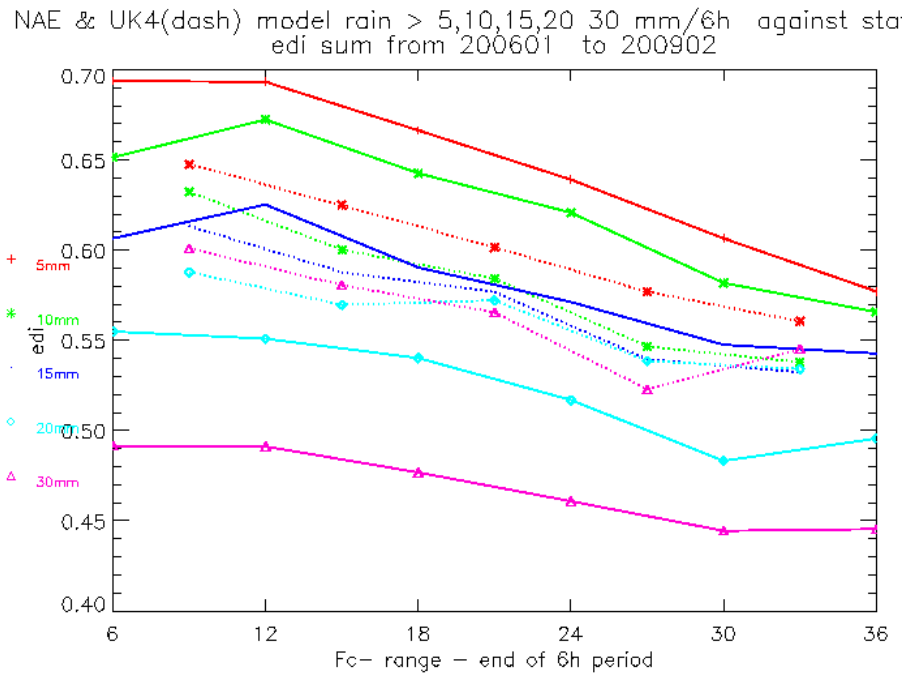
$$\text{since } F = \frac{p(B-H)}{1-p} \rightarrow p, \text{ as } p \rightarrow 0, \text{ if } B = 1, \text{ie recalibrated}$$



Extremal dependency index

—— 12km - - - 4km





Bias correction 1 Bias correction 2

SRNWP-V 1 Conclusions

- Established framework comparison
  - No single model with clear large advantage
  - Continue with new higher resolution operational models
    - Identify benefits
  - Identified most promising methods for operational monitoring with spatial methods
  - Catalogue of useful data sets non-GTS
  - Warnings / extreme verification - started

SRNWP-V Phase 2 Programme progress in red

- Continue & expand comparison
  - Longer more robust results – up-to-date, publication of results every Quarter on EUMETNET Portal
  - higher resolution of future operational models
    - AROME and ARPEGE results processed since start of 2011
  - Overlap models in pairs
  - SEEPS scores being calculated over common domain
- Additional products verified
  - Cloud amount/base
    - truth ? Auto/manual different biases
    - Satellite mask – Ric Croker UKMO, ETS,Intensity scale, FSS,SAL
  - Visibility – which models
  - Wind gust – Met Office validation underway
  - Others as suggested by Consortia

ND3: Spatial & scale selective verification of precipitation

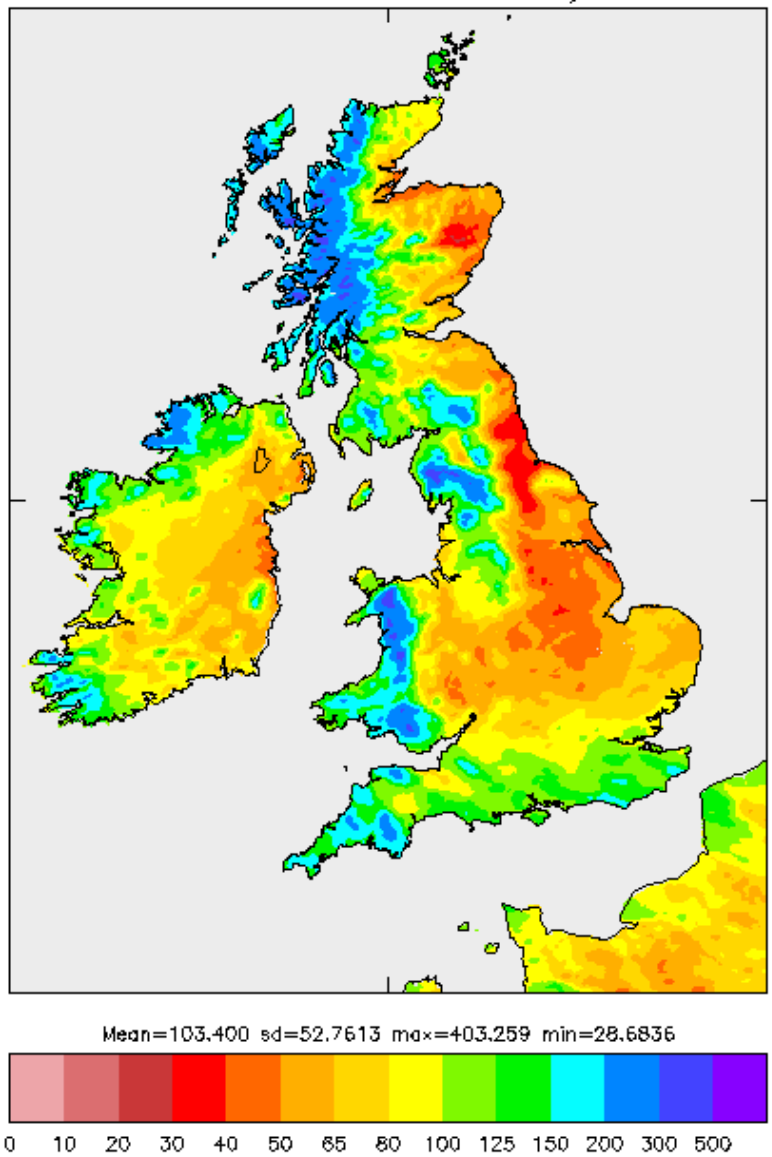
- Verify against
  - Gridded analyses- ECMWF, Meteo-France, Met Office (UK only)
    - Other national gridded sets ?
  - High resolution radar (5 min,1-2 km) -UK
  - OPERA radar composite QC / gauge bias correction

- Methods
  - Fractional skill (Roberts & Lean)
  - Upscaling
  - Intensity scale (Casati)
  - Structure,amplitude,location (SAL) (Wernli et al)
  - Contiguous rain areas (Ebert & McBride)

ND4 Inclusion of severe/high impact weather verification

- ND4 Methods as identified in SRNWP-V 1
  - Extreme dependency scores being applied to precipitation and wind forecasts from models
  - Warnings verification
- Deliverable ND5:
  - Full documentation of the methods used in the intercomparison.
  - Newer spatial methods code to be portable.

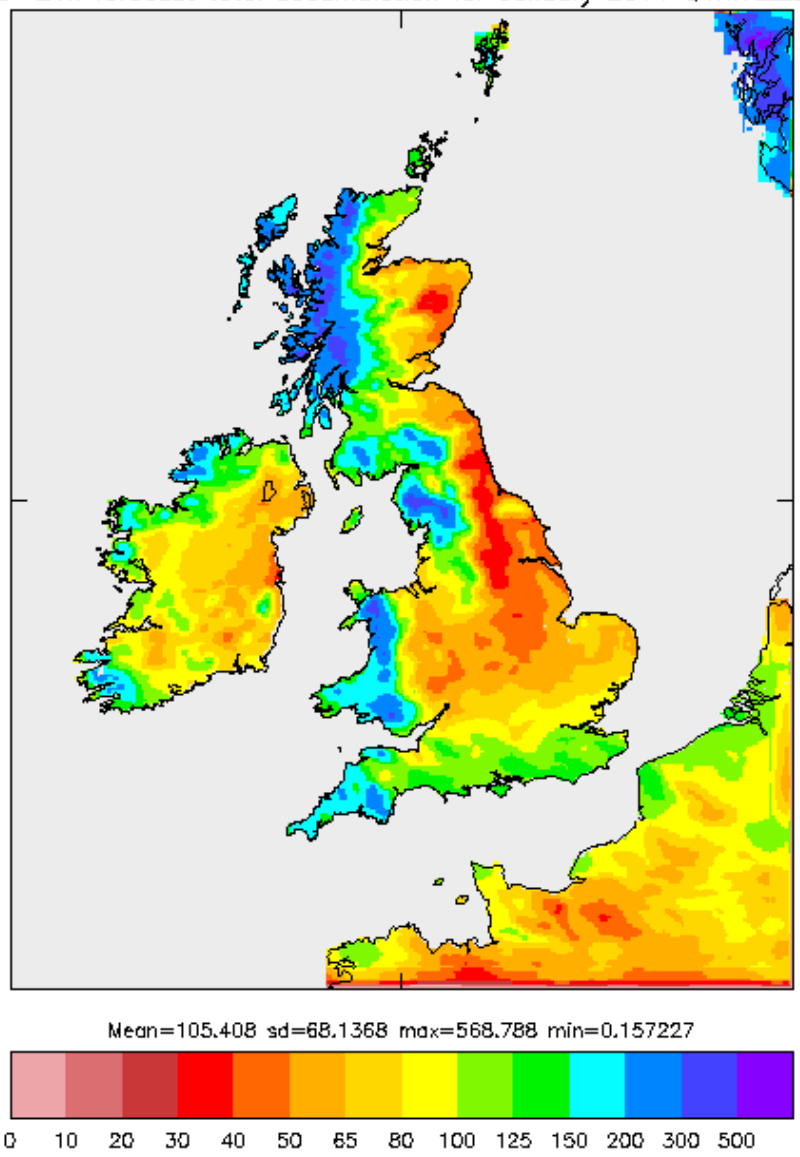
all runs 24h forecast total accumulation for January 2011 QVmvt__acc



Mean=103.400 sd=52.7613 max=403.259 min=28.6836

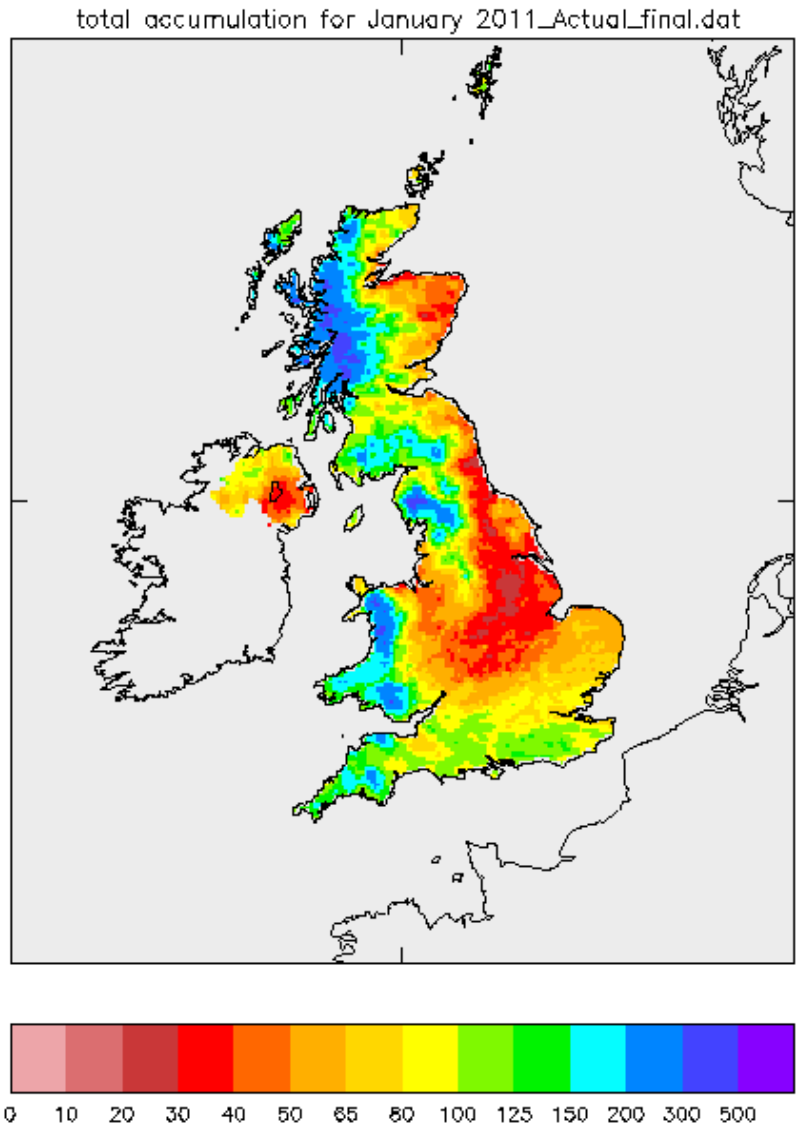
0 10 20 30 40 50 65 80 100 125 150 200 300 500

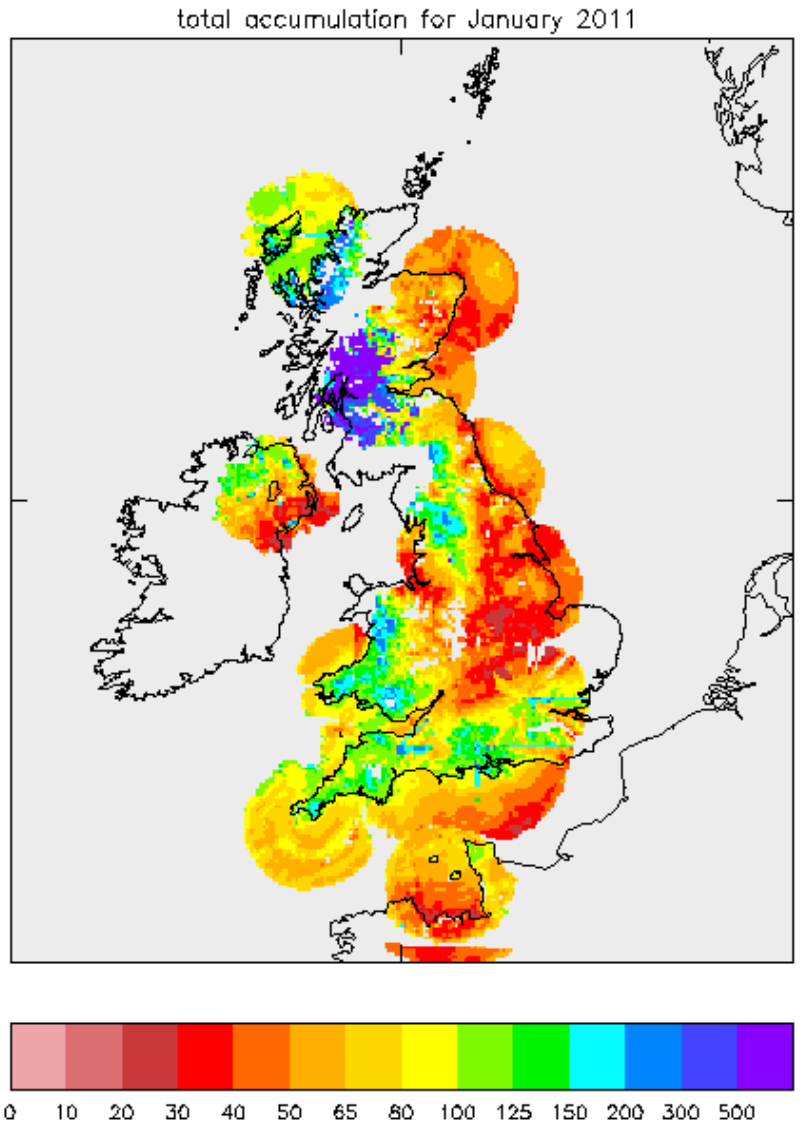
all runs 24h forecast total accumulation for January 2011 Q4mvt__acc

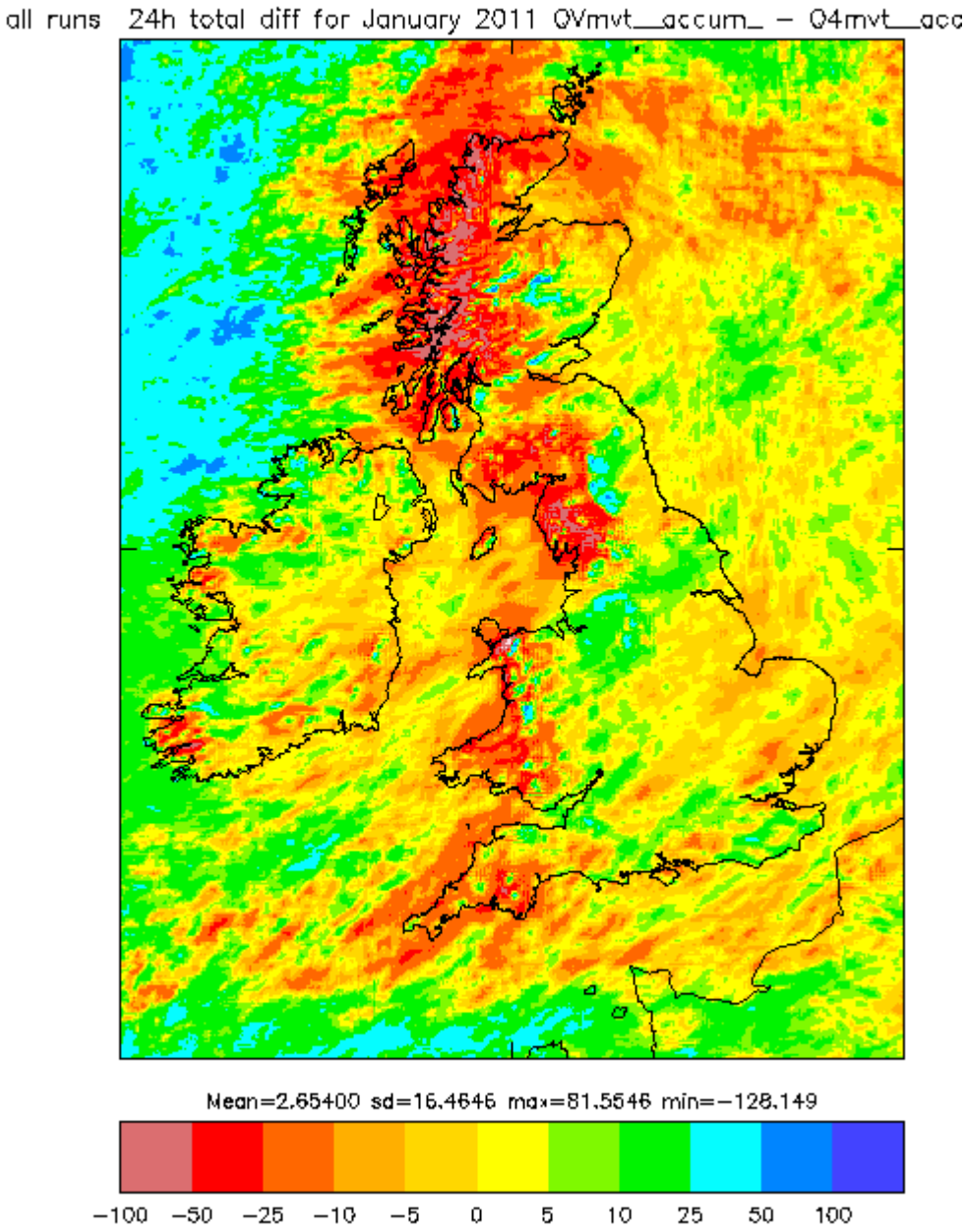


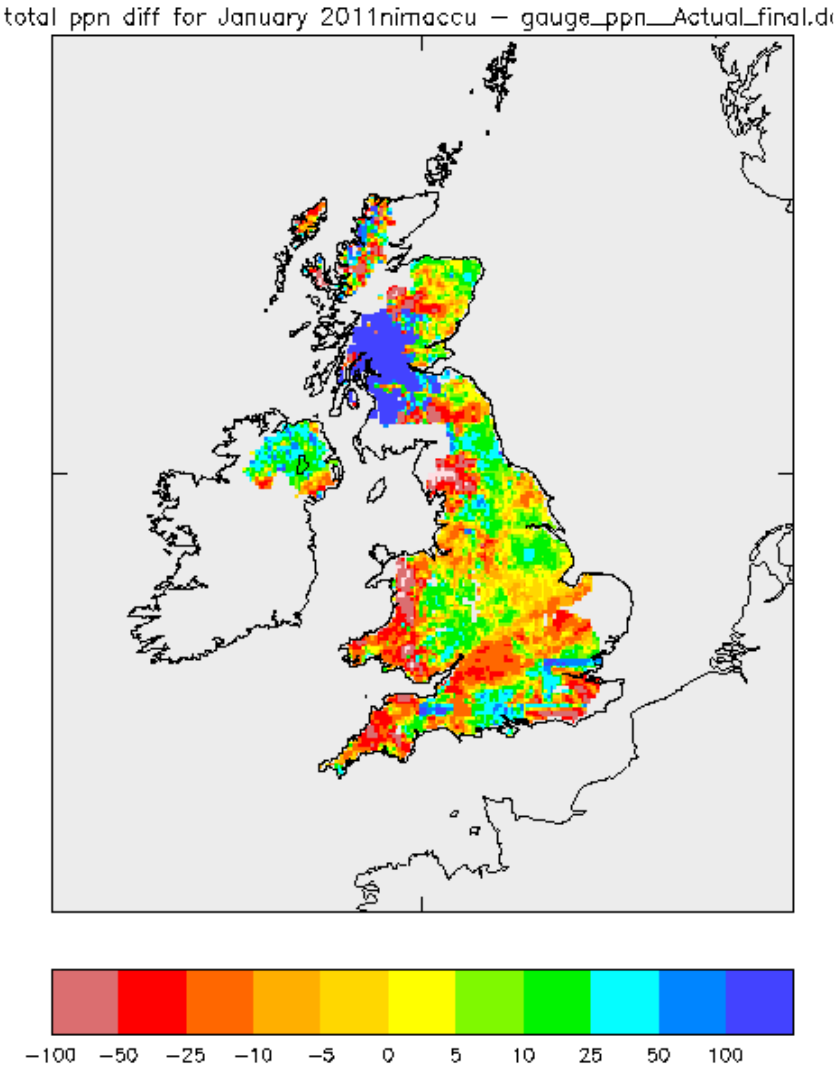
Mean=105.408 sd=68.1368 max=568.788 min=0.157227

0 10 20 30 40 50 65 80 100 125 150 200 300 500






all runs 24h total diff for January 2011 QVmvt__accum_ – Q4mvt__acc
Mean=2.65400 sd=16.4646 max=81.5546 min=−128.149
−100 −50 −25 −10 −5 0 5 10 25 50 100


total ppn diff for January 2011nimaccu – gauge_ppn__Actual_final.d
−100 −50 −25 −10 −5 0 5 10 25 50 100

ECMWF TAC subgroup on verification- final recommendations for headline measures

- 500hPA ACC (deterministic) and 850Temp CRPSS (EPS)
  - against analyses , extra-tropics, new climatology (ERA-I 1989-2008)
- Recommends 4 subsidiary scores
  - SEEPS 24h precipitation extra-tropics (deterministic)
  - Tropical cyclone position error (deterministic)
  - EFI 10m wind (EPS)
  - CRPSS for daily precipitation extra-tropics (EPS)

High-impact weather

- ✓ Comprehensive review of available verification scores for high-impact (severe) weather was performed by the Group, including presently on-going developments
- ✓ The Group identified two major issues:
  1. Lack of observations at sufficient temporal and spatial resolution
  2. Lack of fundamental research into related verification in meteorological services and universities
- ✓ The Group identified a set of properties that verification scores (especially for extreme events) should possess
  - No currently available measures satisfy these requirements
- ⇨ Substantial research is needed to develop suitable verification scores - both for deterministic and probabilistic forecasts

Long-term trends are monitored by:

The number of days for which the forecast skill remains above a pre-specified threshold (i and ii)

Proposal

b) Introduce 4 supplementary headline measures

i. New **1 - SEEPS** score for the *deterministic forecasts* of 24-hour precipitation over the extra-tropics

➔ An appropriate threshold should be 45% currently achieved for the 24-hour period ~ 3.5 Days ahead

ii. **CRPSS**, for the *EPS probabilistic forecasts* of 24-hour precipitation over the extra-tropics

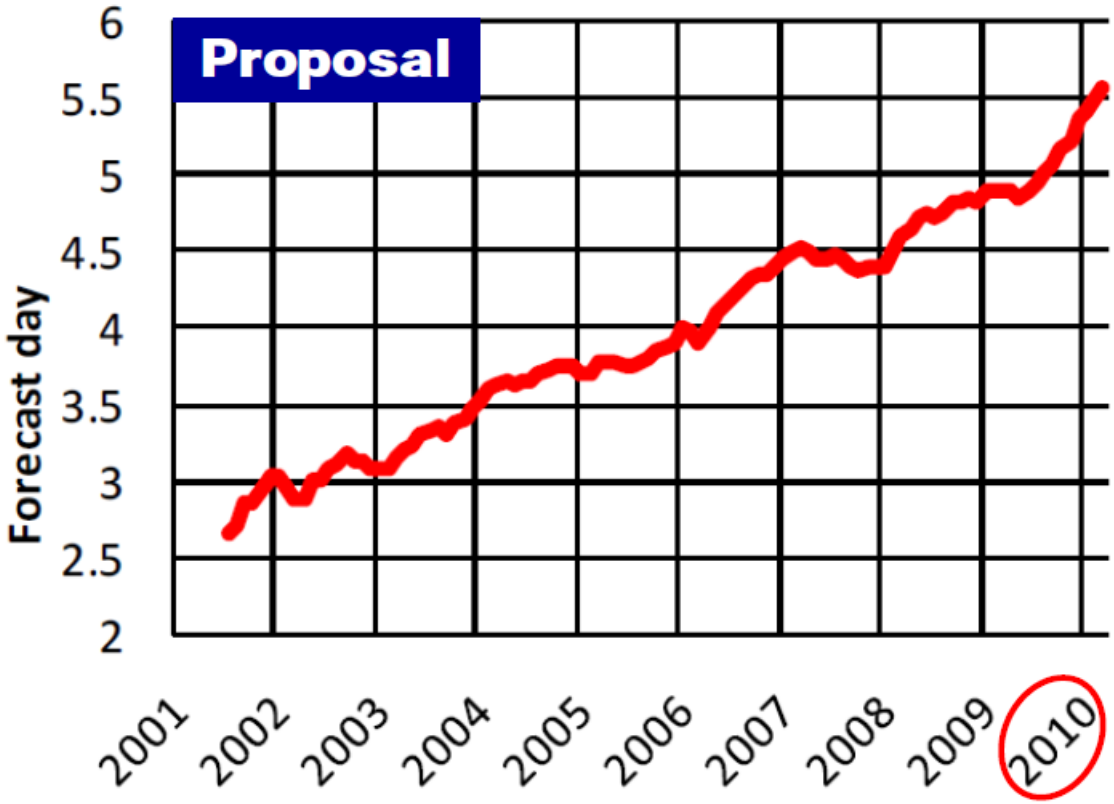
➔ An appropriate threshold should be 10% currently achieved at ~ Day 5

iii. Severe weather: **ROC Area** for the *EFI for 10 m wind*

iv. Severe weather: **Tropical cyclone position error** for the *deterministic forecast*

Proposal

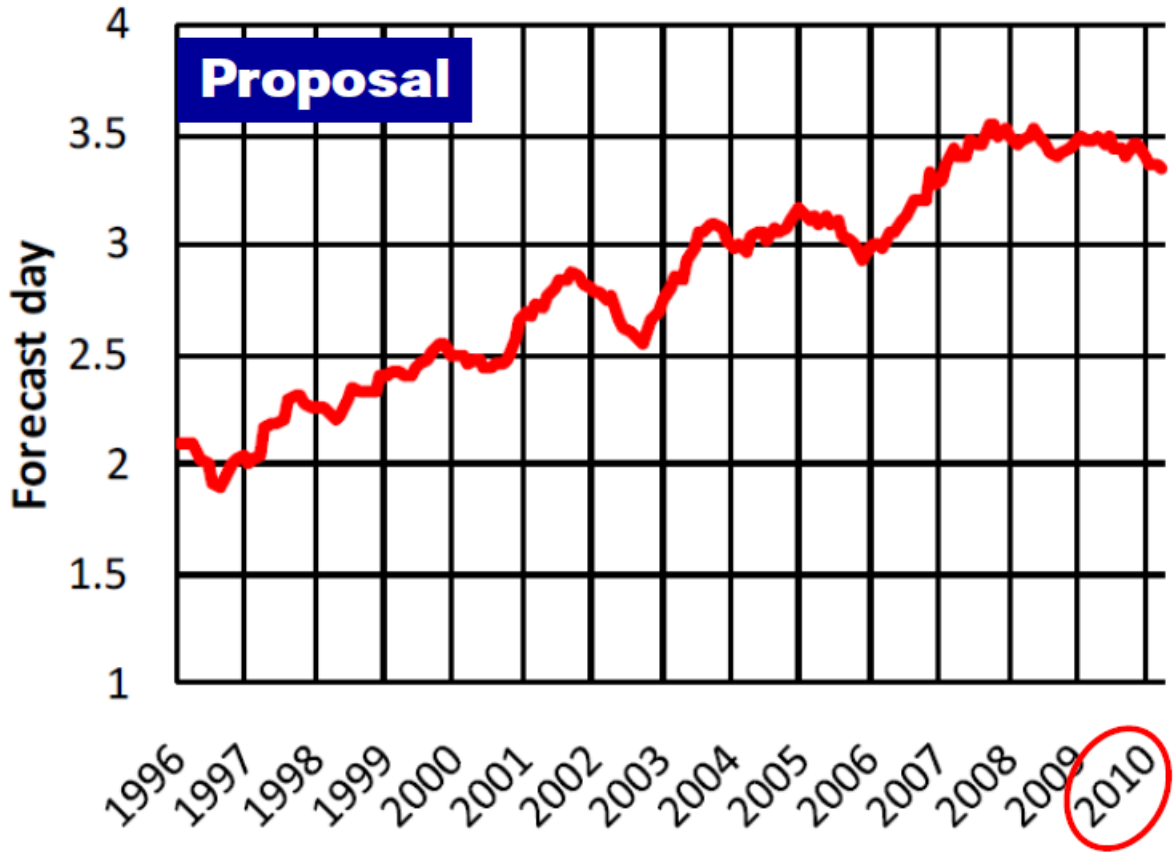
Supplementary headline measure (ii): CRPSS for 24-hr EPS Precipitation



CRPSS remains above 10 %

Supplementary headline score for ***probabilistic*** precipitation forecasts. The curve shows the number of days for which the centered 12-month mean skill remains above a specified threshold for precipitation forecasts over the extra-tropics. The verification is for 24-hour total precipitation verifying against available synoptic observations. The forecast day on the y-axis is the end of the 24-hour period over which the precipitation is accumulated. The threshold is chosen to reflect the forecast skill that is achieved at approximately ***day 5.5*** at the beginning of the strategy period.

Supplementary headline measure (i):
1 – SEEPS for 24-hr deterministic Precipitation

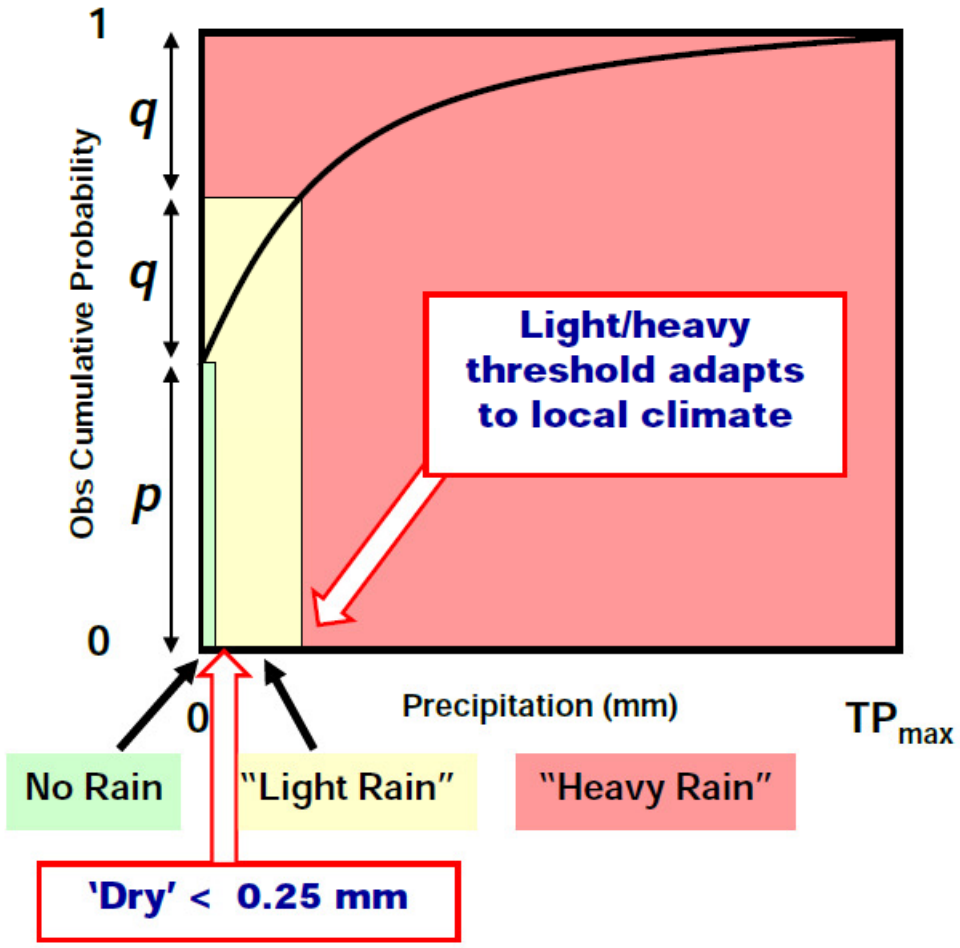


1 – SEEPS remains above 45%

Supplementary headline score for ***deterministic*** precipitation forecasts. The curve shows the number of days for which the centered 12-month mean skill remains above a specified threshold for precipitation forecasts over the extra-tropics. The verification is for 24-hour total precipitation verifying against available synoptic observations. The forecast day on the y-axis is the end of the 24-hour period over which the precipitation is accumulated. The threshold is chosen to reflect the forecast skill that is achieved at approximately ***day 3.5*** at the beginning of the strategy period.

Supplementary headline measure (i): 1 - SEEPS

Use of Cumulative Distribution



The characteristics and benefits of SEEPS

A

Stable: SEEPS is designed to be as insensitive as possible to sampling uncertainty (for sufficiently skilful forecast systems). This allows more accurate trends to be extracted from noisy data.

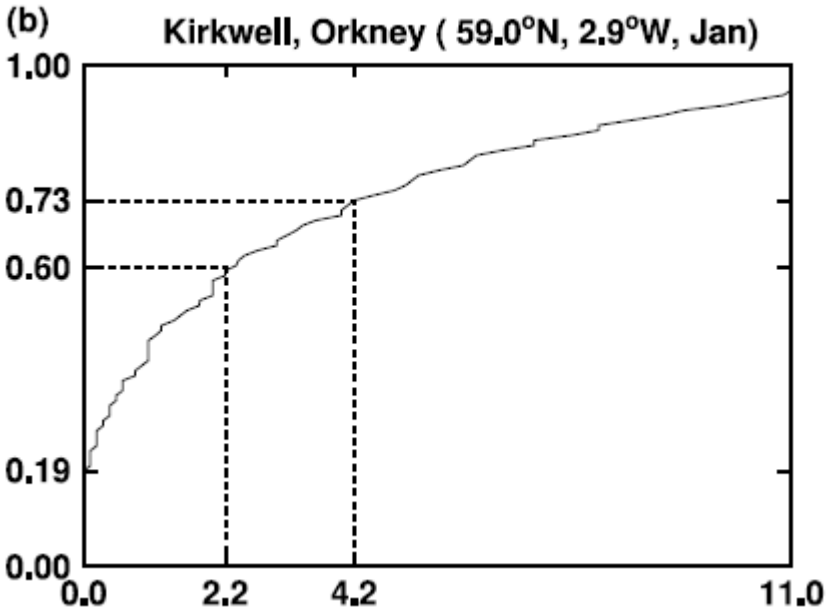
Equitable Error: A perfect forecast has a SEEPS score of 0. The expected score increases linearly with the unskilled component of the forecast towards a maximum value of 1.

Probability Space: This is used to define precipitation categories; SEEPS adapts to the underlying climate to assess the pertinent aspects of local weather. It can be aggregated over heterogeneous climate regions.

SEEPS=Stable Equitable Error in Probability Space

Rodwell et al, 2010, QJRMS 136

- Dry, light , heavy based on observed climatology (24h) at station – p_1 , p_2 , p_3
- Contingency table probabilities based on these categories
- Scoring matrix – stable, equitable
    - SEEPS=0 (perfect) , =1 (no skill - , eg constant)
- Now applying to 6h accumulations in SRNWP-V
    - 6h climatology (courtesy Mark Rodwell)



Thanks - Questions