New Avenues in Forecast Verification

Laurie Wilson

With thanks to the members of the Joint Working Group on Forecast Verification Research:
Beth Ebert, Barb Brown, Pertti Nurmi, Anna Ghelli, Barbara Casati

Outline

- Introduction: The motivation for verification research
- New methods in “pointwise” verification
- Spatial and scale-sensitive methods
  - Types
  - Examples
- Promoting “best practice” in verification

- This is a survey of methods: Is there something in here that can be used to advantage at CMC/RPN?

Status and motivation for verification research

- "Verification activity has value only if the information generated leads to a decision about the forecast or system being verified" (Murphy)
- New emphasis on "User-oriented" verification
  - Modelers
  - Forecasters
  - Hydrological community
  - Specific users such as VANOC
- Extremes (Rare events) (High Impact Weather)
- For Ensemble Forecasts

“Traditional” methods

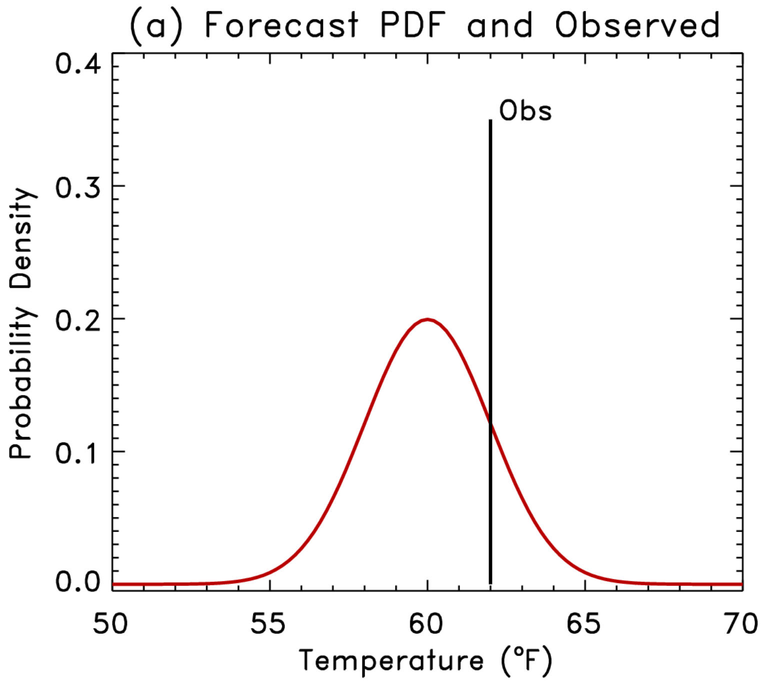
- Point-by-point matching of forecast and observation
- Summary scores:
  - Continuous variable: (R)MSE, MAE, scatter plot, linear bias
  - Categorical variable: Contingency tables and a whole lot of related scores: ETS, POD, FAR, TS(CSI), HSS, PSS(H-K)...
  - Probability forecast of a categorical variable
    - BS, BSS and reliability, resolution components.
    - Reliability diagram and the ROC
  - (Discrete) Probability distribution
    - RPS, RPSS

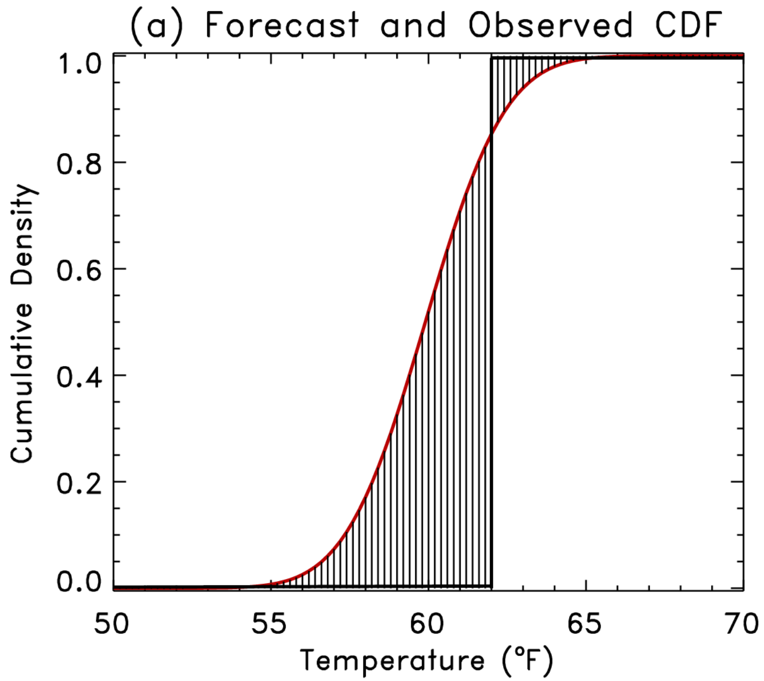
Extensions to “traditional” verification

- For ensembles: The CRPS (Herzbach, 2000)
  - Continuous form of the RPS
  - In practice is also discrete, with categories defined by the ensemble member forecasts
  - Measures the difference between the forecast cdf and the observation, represented as a cdf –example
- For extremes:
  - The extreme dependency score (EDS) and symmetric EDS (SEDS)
  - New score “SEEPS”

CRPS and CRPSS

$$CRPS = \int_{-\infty}^{\infty} \left(P_{fcst}(x) - P_{obs}(x) \right)^2 dx$$





$$CRPSS = \frac{\left(CRPS_{STD} - CRPS_{FCST} \right)}{CRPS_{STD}}$$

Extensions to “traditional” verification

- For ensembles: The CRPS (Herzbach, 2000)
  - Continuous form of the RPS
  - In practice is also discrete, with categories defined by the ensemble member forecasts
  - Measures the difference between the forecast cdf and the observation, represented as a cdf –example
- For extremes:
  - The extreme dependency score (EDS) and symmetric EDS (SEDS)
  - New score “SEEPS”

High impact (severe) weather

EDS, EDI, SEDS, SEDI ⇔ *Novelty measures!*

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

H = a / (a+c), hit rate

F = b / (b+d), false alarm rate

p = (a+c) / n, base rate

q = (a+b) / n, relative frequency of forecasted events

$$\boxed{\mathrm{EDS}} = \frac{\log p - \log H}{\log p + \log H} \qquad \boxed{\mathrm{SEDS}} = \frac{\log q - \log H}{\log p + \log H}$$

Ferro & Stephenson, 2010: Improved verification measures for deterministic forecasts of rare, binary events. *Wea. and Forecasting (submitted)*

Base rate independence ⇔ Functions of H and F

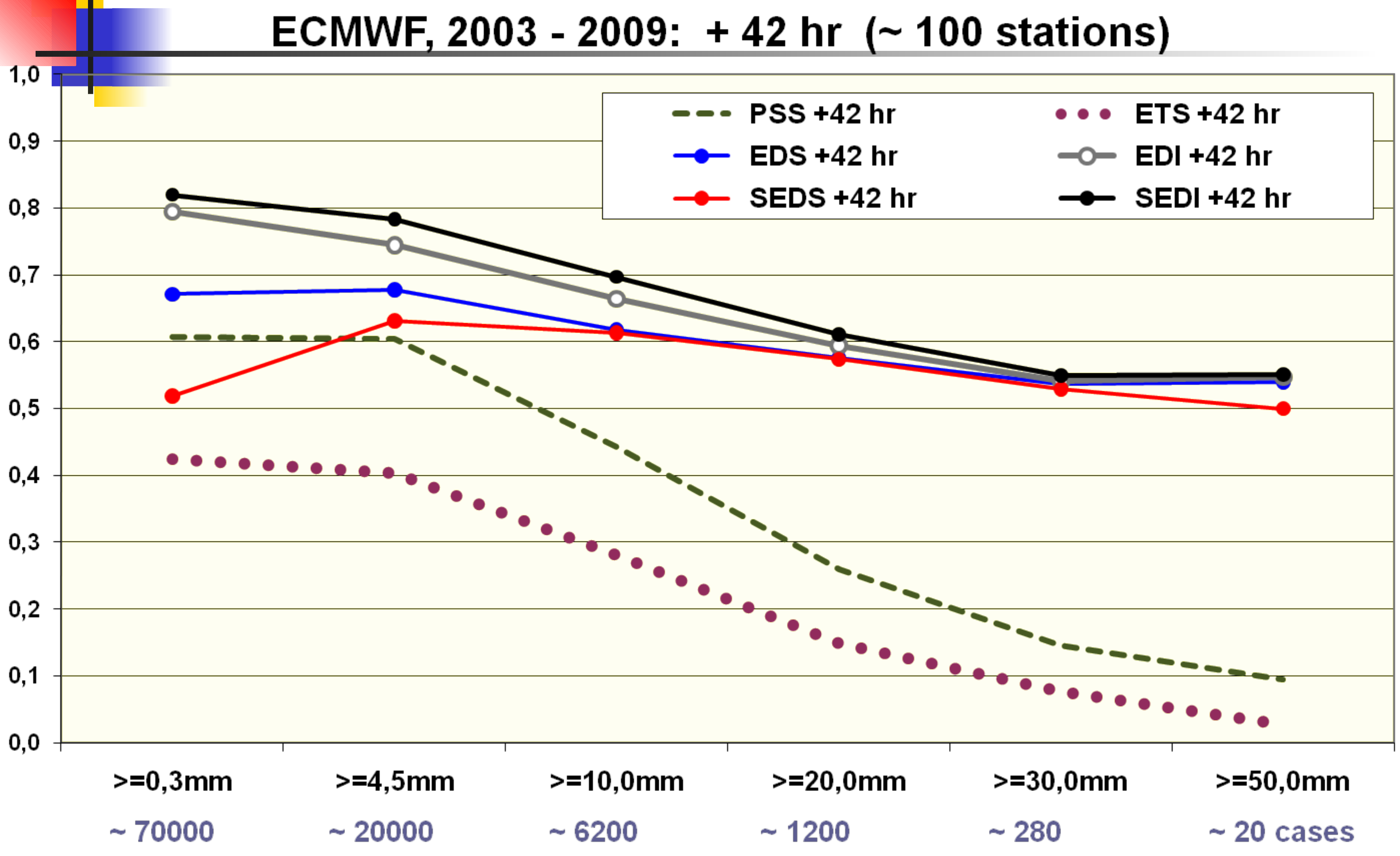
$$\boxed{\mathrm{EDI}} = \frac{\log F - \log H}{\log F + \log H}$$

Extremal Dependency Index - EDI

Symmetric Extremal Dependency Index - SEDI

$$\boxed{\mathrm{SEDI}} = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}$$

High impact (severe) weather

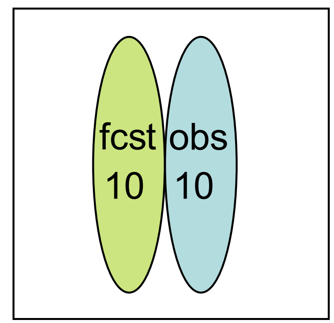


New score: SEEPS ⇔ Stable Equitable Error in Probability Space

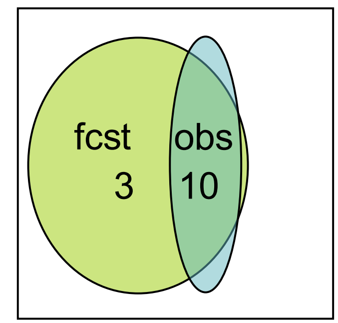
- M.J. Rodwell et al., 2010: QJRMS, 136, 1344-1363.
- Derived from LEPS score ⇔ Linear Error in Probability Space
  - Forecast error is measured in probability space using the climatological cumulative distribution function
- At each observation location, the weather is partitioned into 3 categories: (i) "dry" (ii) "light precipitation" (iii) "heavy precipitation"
  - Long-term climatological precipitation categories at given SYNOP stations are derived ⇔ Accounts for climate differences between stations
- Evaluates forecast performance across all 3 categories
- Stable to sample variations and observation error ⇔ Good for detecting trends
- Gives daily scores ⇔ Identifies a range of forecast errors, e.g.
  - Failure to predict heavy large-scale precipitation; Incorrect location of convective cells; Over-prediction of drizzle...
- Negatively oriented error measure ⇔ Perfect score =0 => **1 - SEEPS**

Why spatial verification methods?

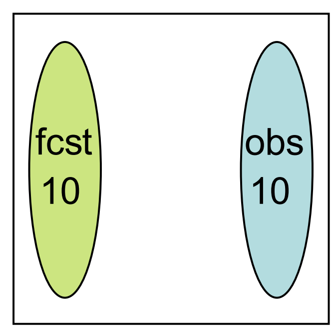
- Pointwise method specifies an exact match between forecasts and observations at every point
    - Problem of "double penalty" - event predicted where it did not occur, no event predicted where it did occur
    - But, more people receive a wrong forecast – is it really double
- Idea is to diagnose patterns predicted by models, especially high res models, which may be hindered by small scale noise



Hi res forecast
RMS ~ 4.7
POD=0, FAR=1
TS=0



Low res forecast
RMS ~ 2.7
POD~1, FAR~0.7
TS~0.3



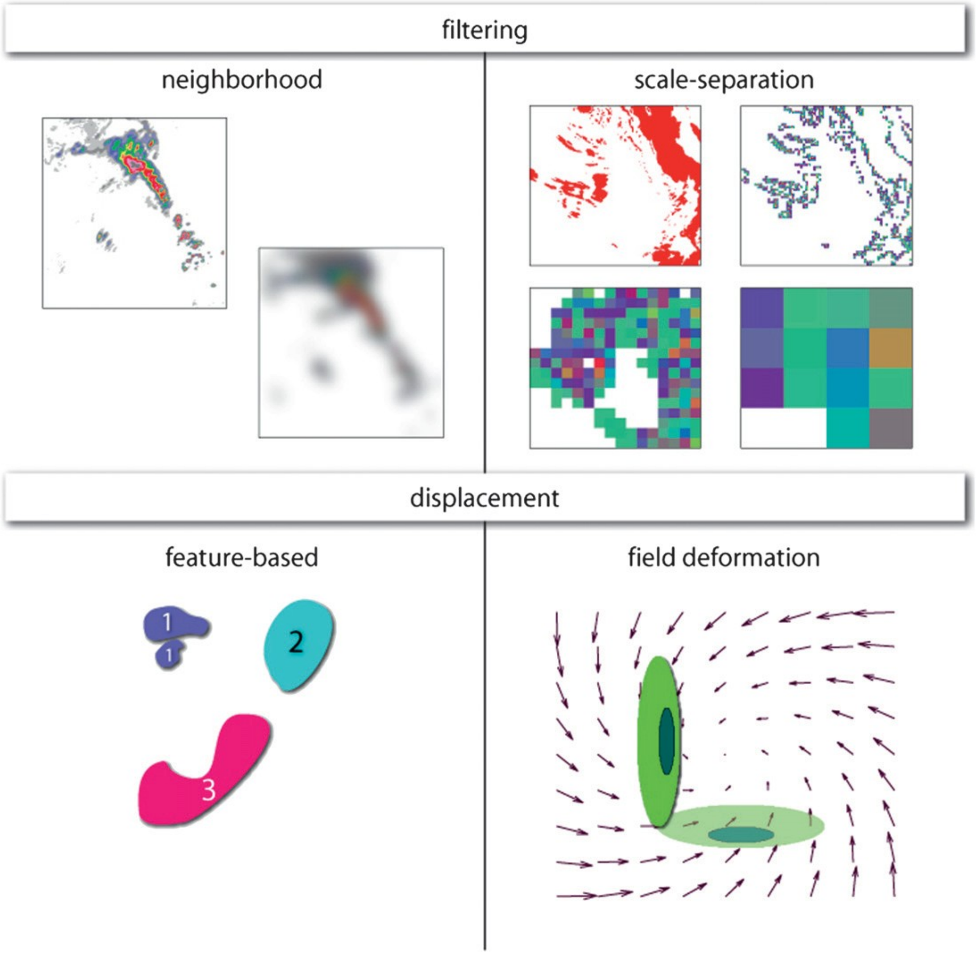
Spatial Method Intercomparison Project (ICP)

- *Weather and Forecasting* special collection WAF, 2009 and 2010
  - 13 papers on specific methods
  - 2 overview papers
- Methods applied by researchers to same datasets (precipitation; perturbed cases; idealized cases)
- Subjective forecast evaluations
- Future variables and datasets
  - Wind
  - Cloud
  - Timing errors

http://www.rap.ucar.edu/projects/icp/index.html

Spatial methods

- Types:
  - Neighbourhood: Look for feature in vicinity rather than at specific points **(High resolution models and ensembles)**
  - Scale separation: Keep track of scales represented by obs and fcsts; partition scores according to scale **(“Seamless” verification?)**
  - Feature-based methods: Characterize features and verify the characteristics **(Forecaster-oriented verification)**
  - Deformation methods: systematically deform and translate features to get best match; track statistics of differences. **(Model diagnostics?)**

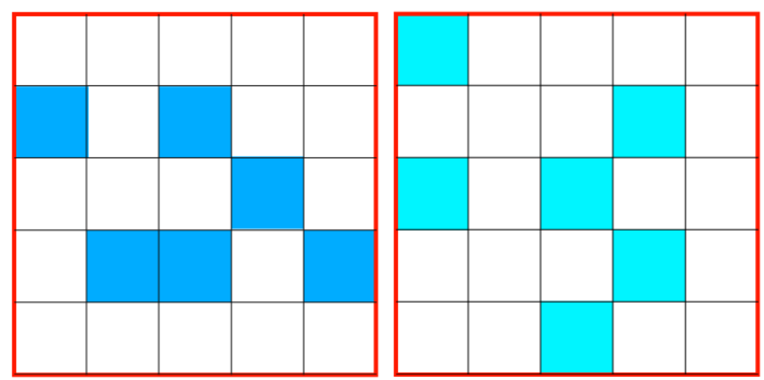


Neighbourhood methods: Fractions skill score (Roberts and Lean, 2008, MWR)

- We want to know
  - How forecast skill varies with neighborhood size
  - The smallest neighborhood size that can be can be used to give sufficiently accurate forecasts
  - Does higher resolution NWP provide more accurate forecasts on scales of interest (e.g., river catchments)

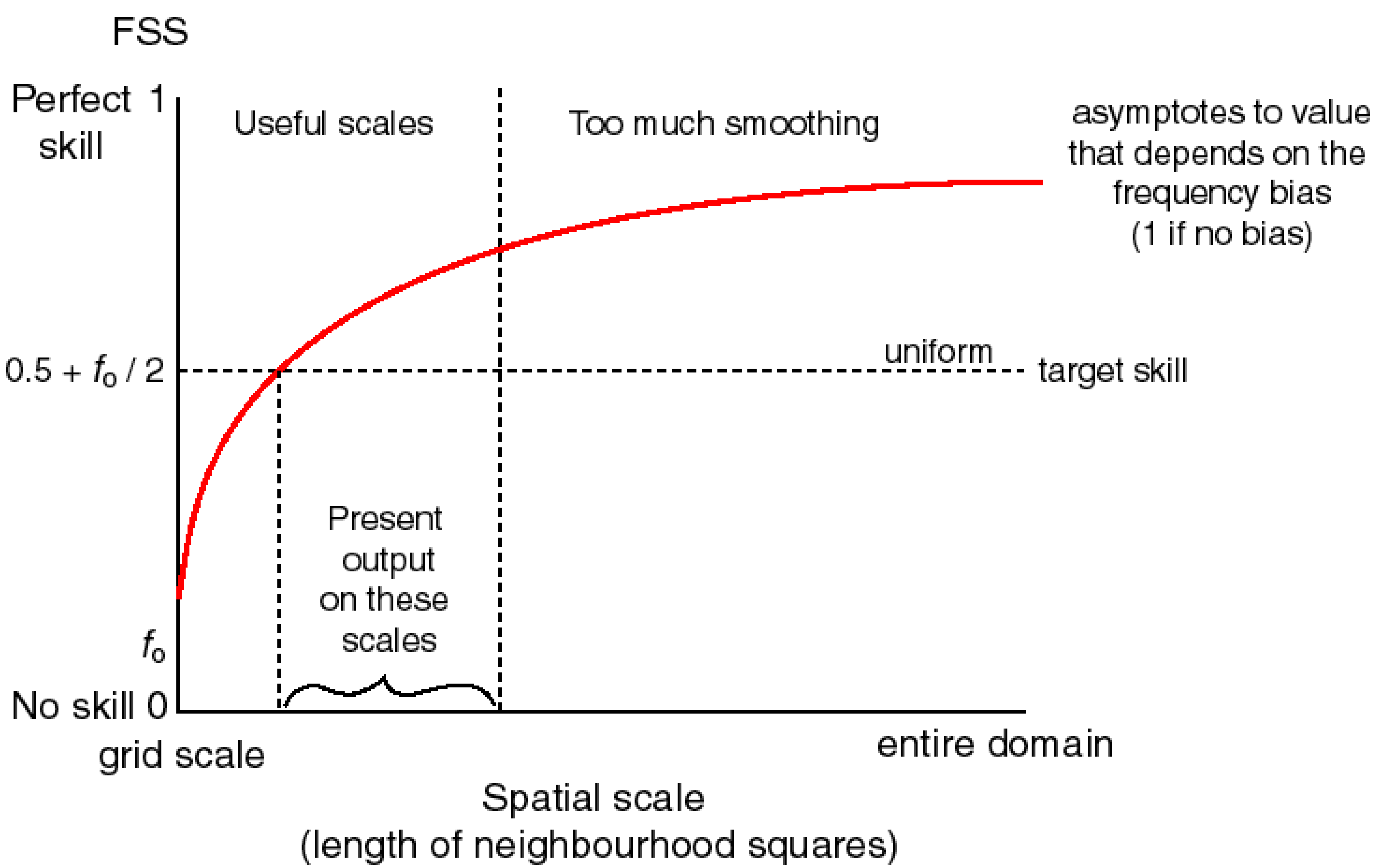
Compare forecast fractions with observed fractions (radar) in a *probabilistic* way over different sized neighbourhoods

$$\mathrm{FSS} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(P_{fcst} - P_{obs}\right)^2}{\frac{1}{N}\sum_{i=1}^{N} P_{fcst}{}^2 + \frac{1}{N}\sum_{i=1}^{N} P_{obs}{}^2}$$



Fraction = 6/25 = 0.24 | Fraction = 6/25 = 0.24
observed | forecast

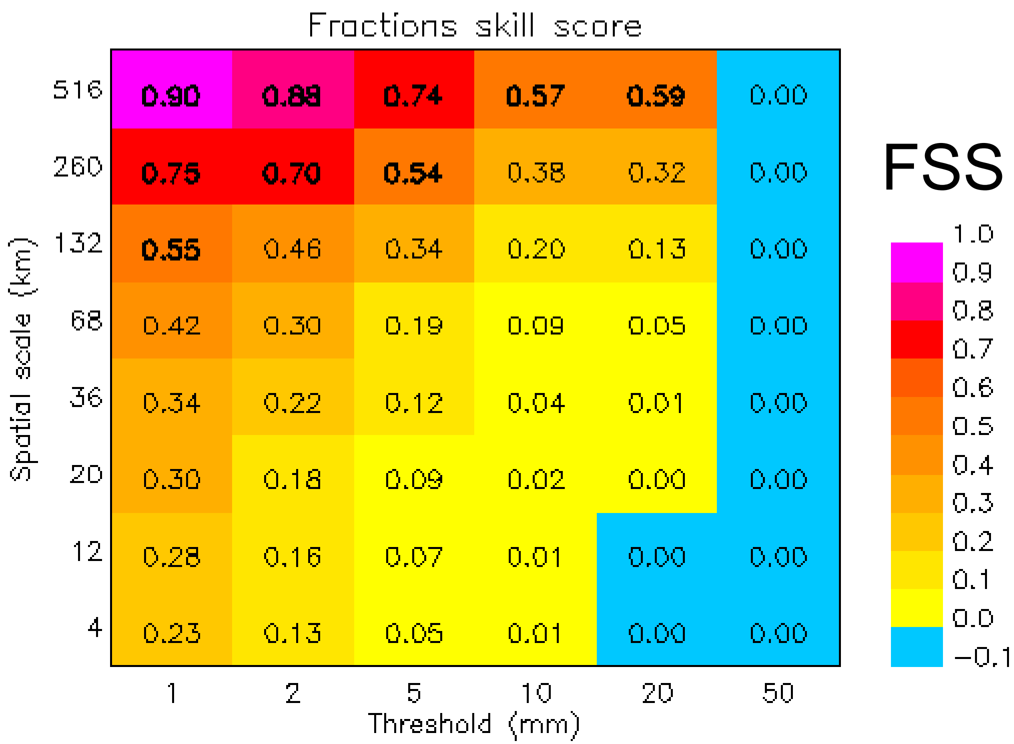
Fractions skill score (Roberts and Lean, *MWR*, 2008)



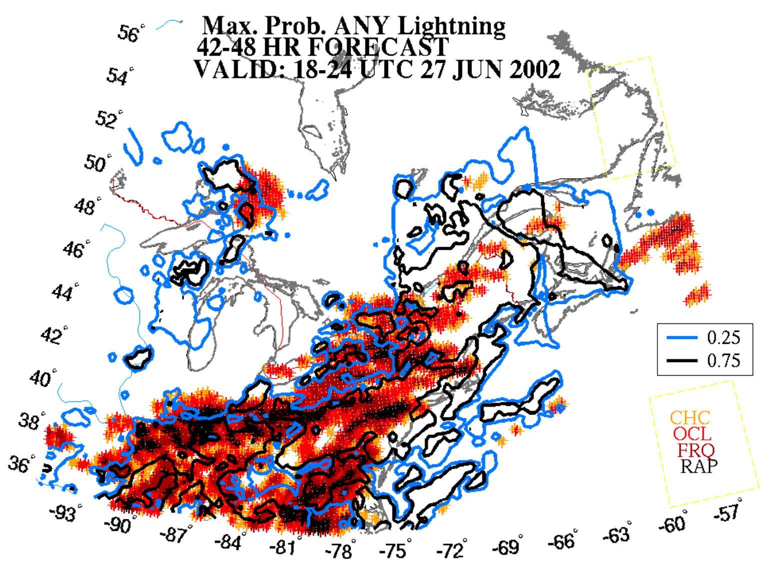
f_o=domain obs fraction

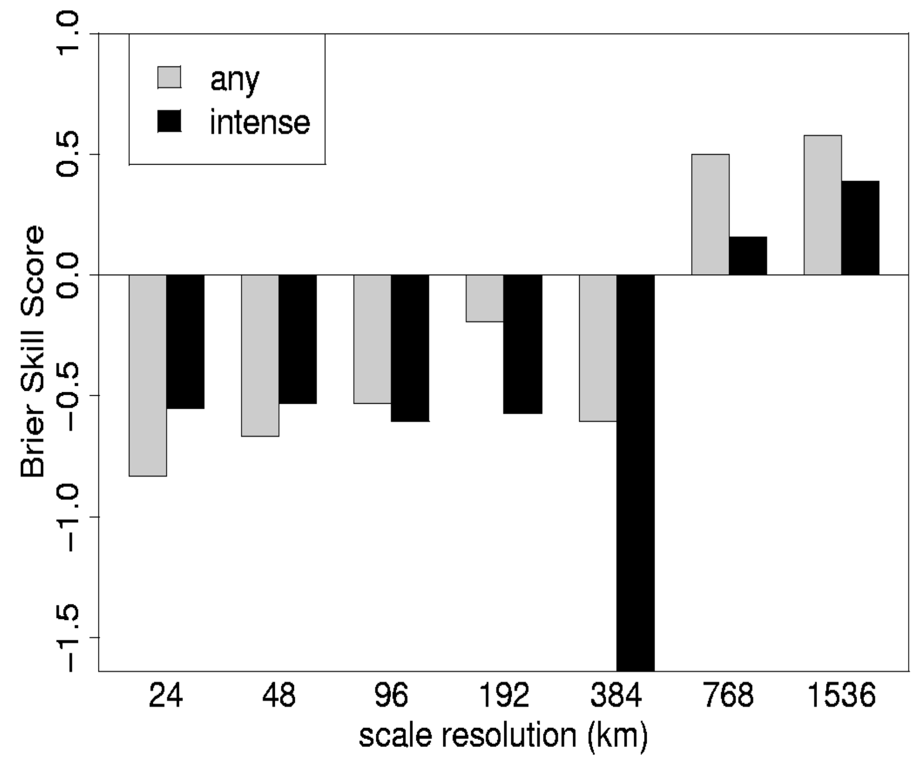
Presenting the results from the FSS

- Fractions skill score



Scale-separation methods



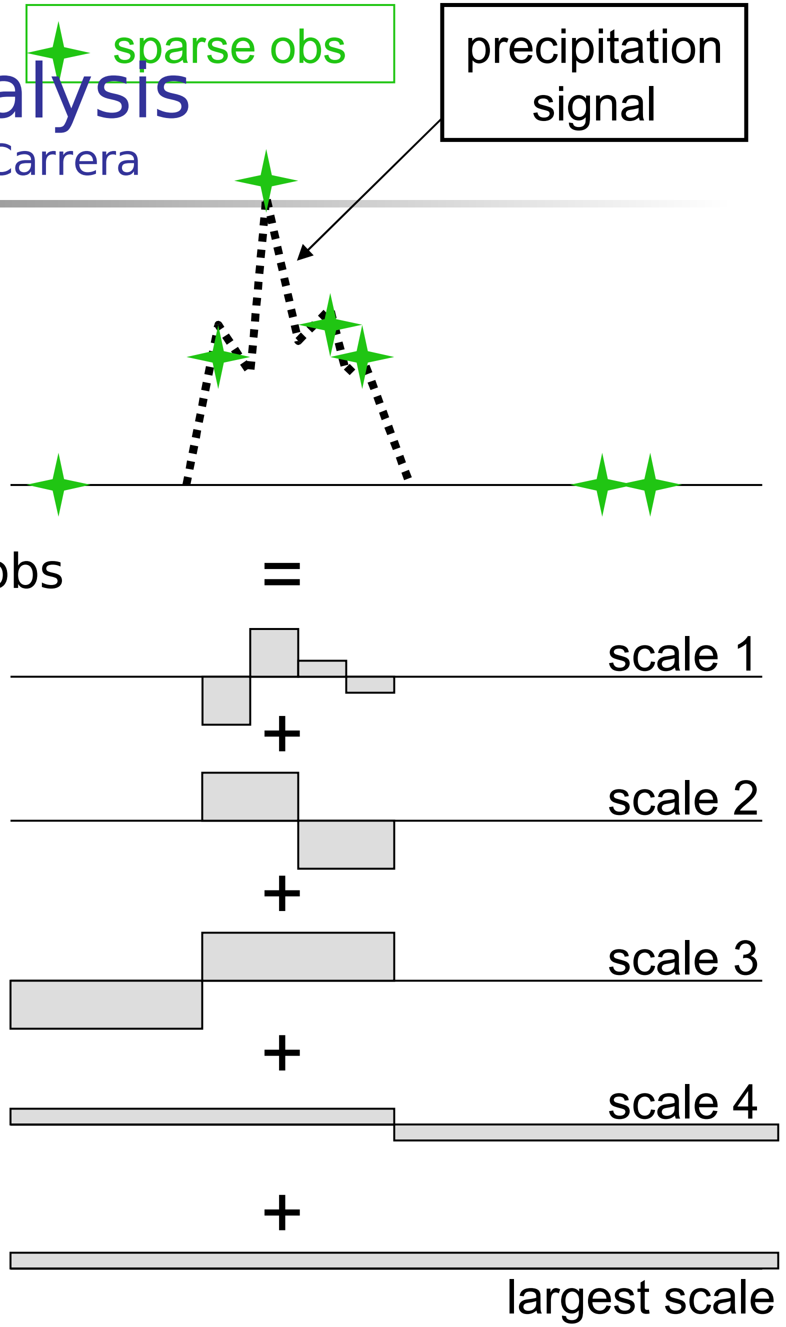


Wavelet decomposition of the Brier Skill Score

Thanks to Barbara Casati

B. Casati's Wavelet Analysis

Thanks also to Vincent fortin and Marco Carrera



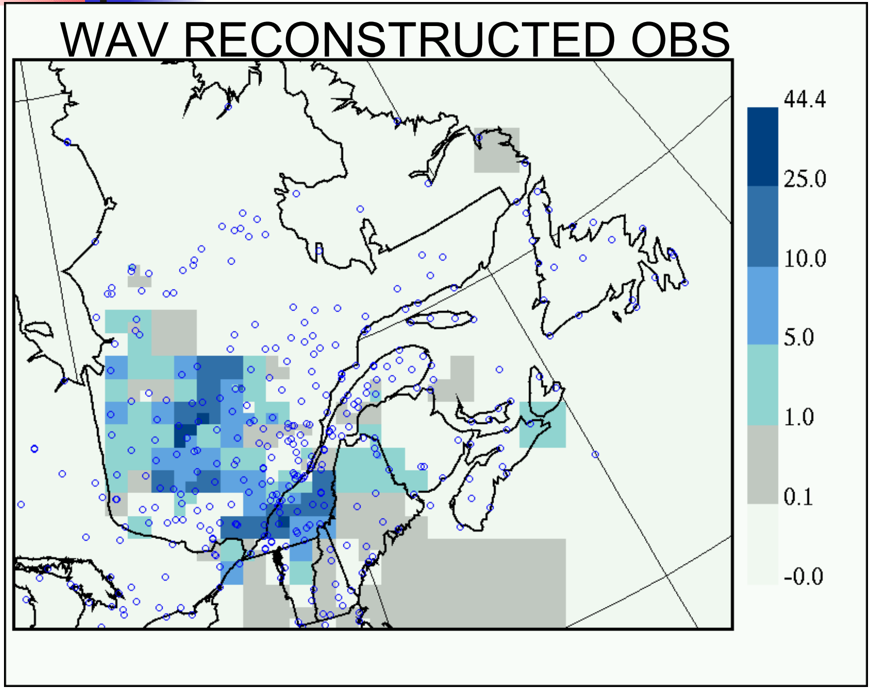
Use 2D Harr Wavelets to represent e.g. Precipitation field from network of gauges

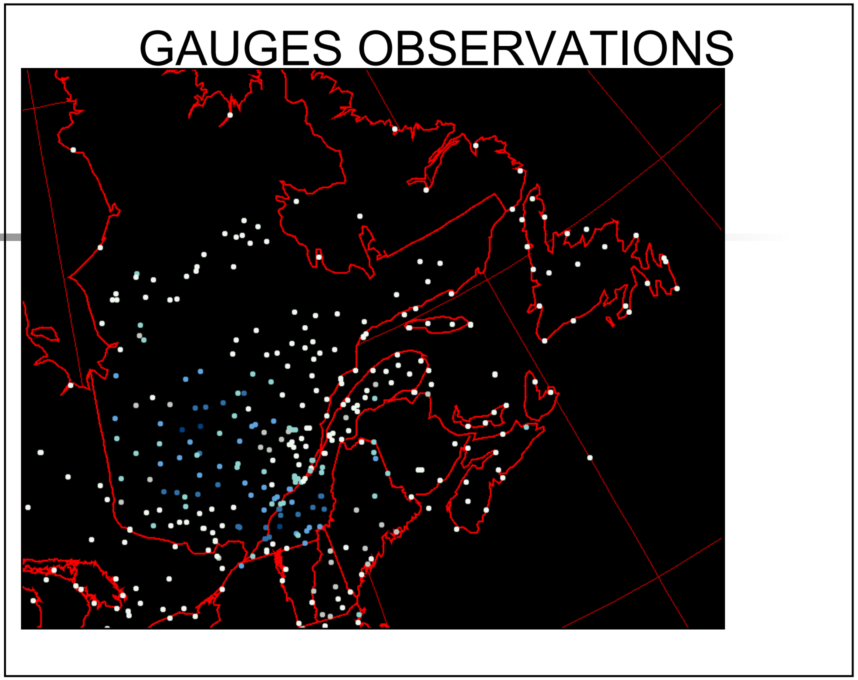
Main advantage: Keeps track of resolved Scales; for better matching of forecast and obs

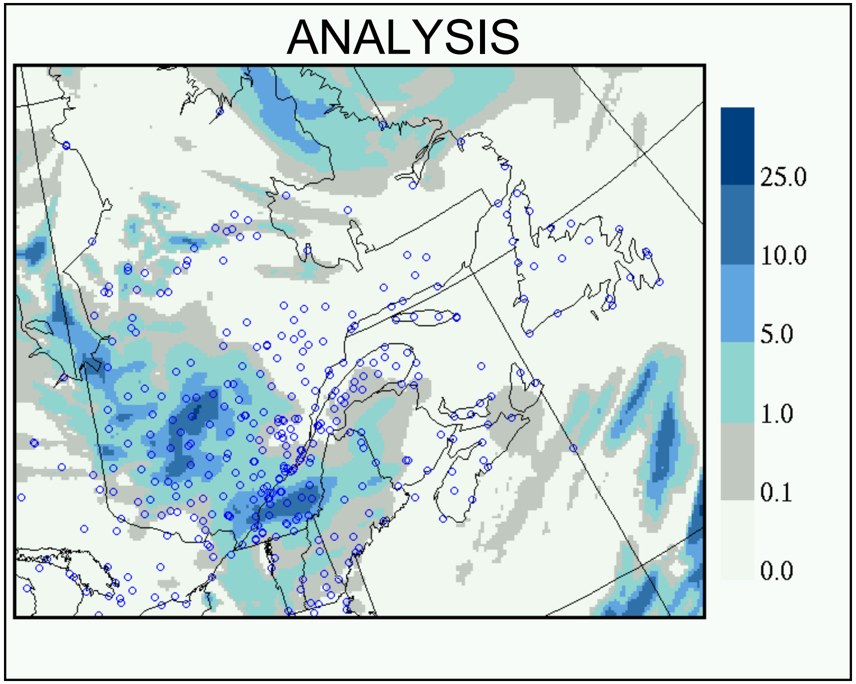
1. Compute wavelet coefficients from sparse gauge obs
2. Reconstruct field as sum of components on different scales

NOTE: no gauges = missing obs, no dense gauge network = no information on small scales, large scales only !

Example: 6h acc (mm)
27th Aug 2003, 6:00 UTC

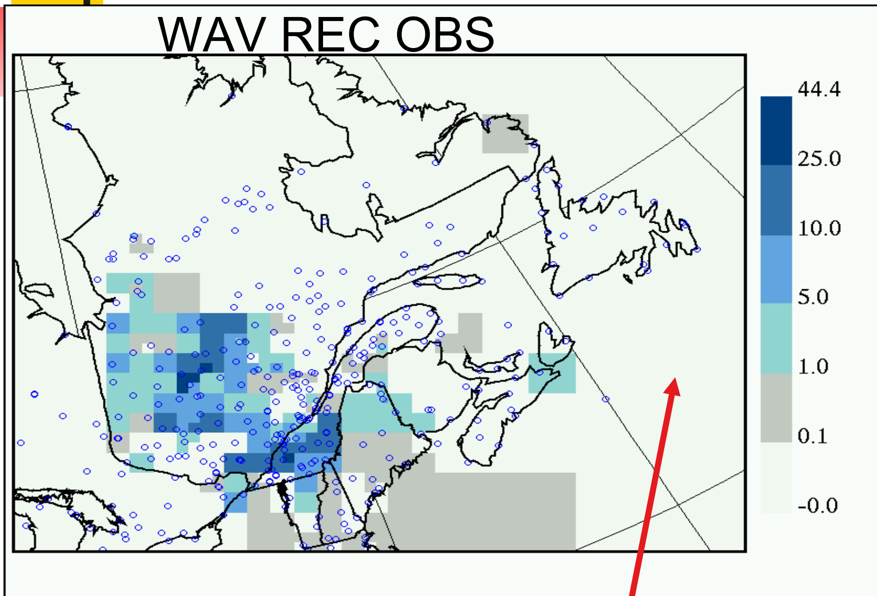


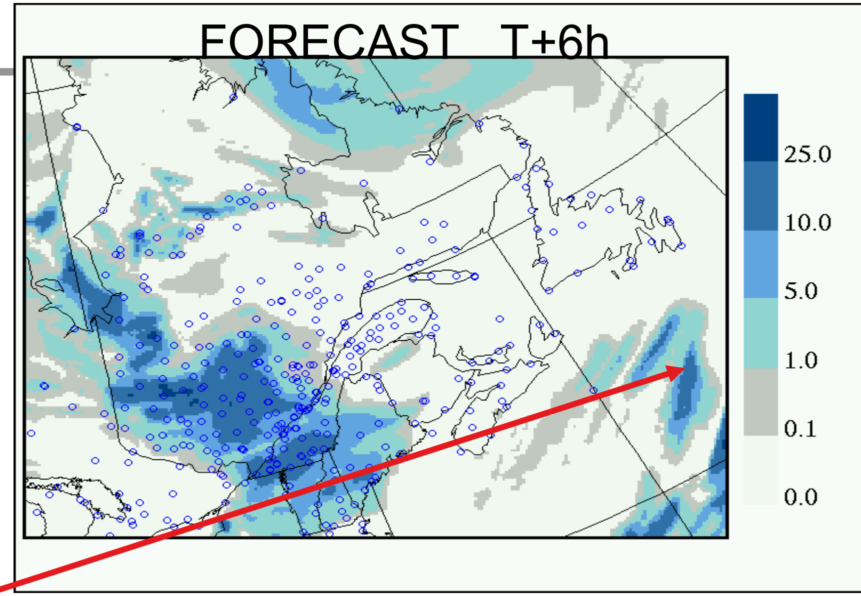




- Account for existence spatial structures on different scales
- Account for gauge network density
- Value at station location is equal to gauge value

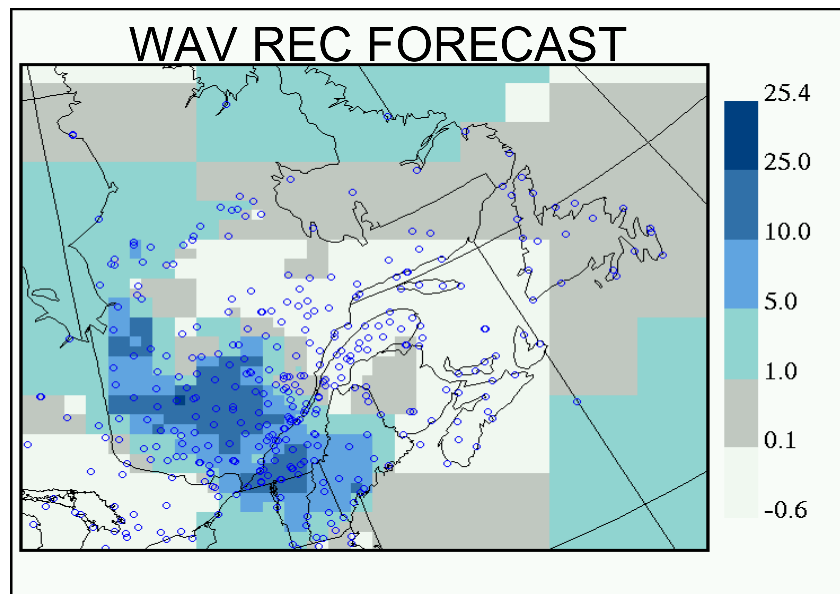
3. Representativeness and forecast filtering



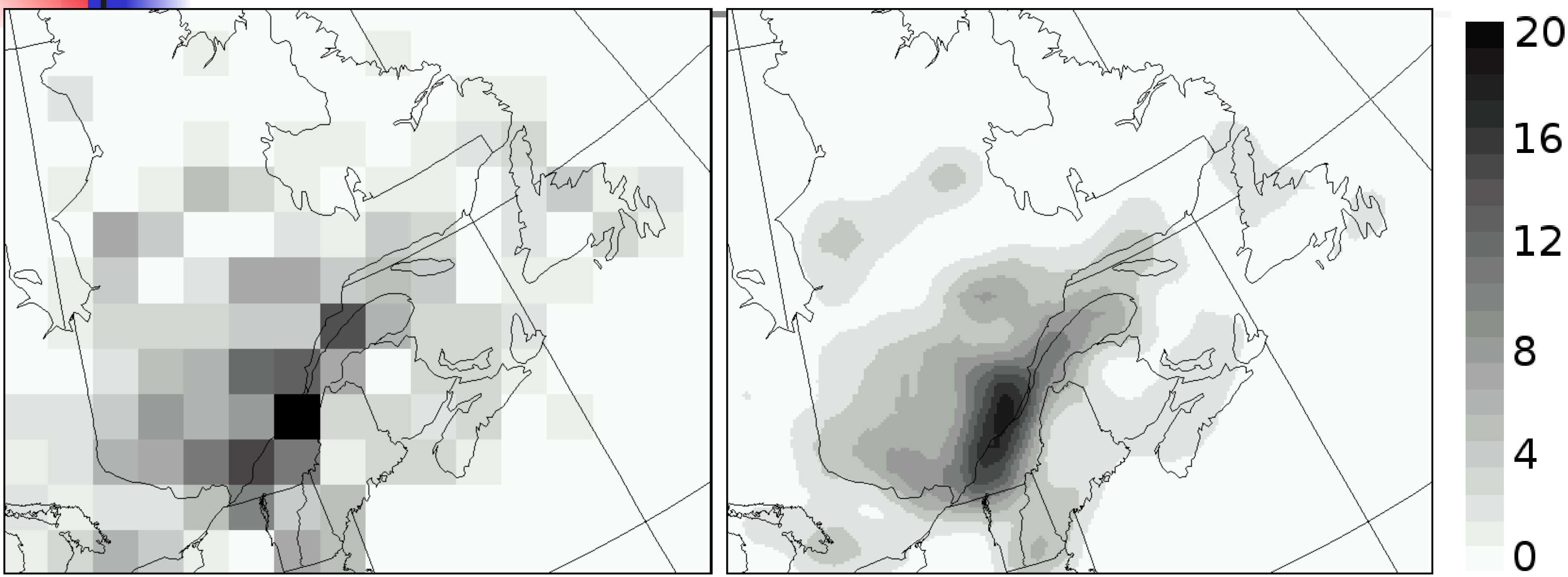


No gauges = missing obs, but forecast has features!

2. Decompose forecast with wavelets
3. Set to NA wavelet coefficients where no obs
4. Reconstruct forecast field

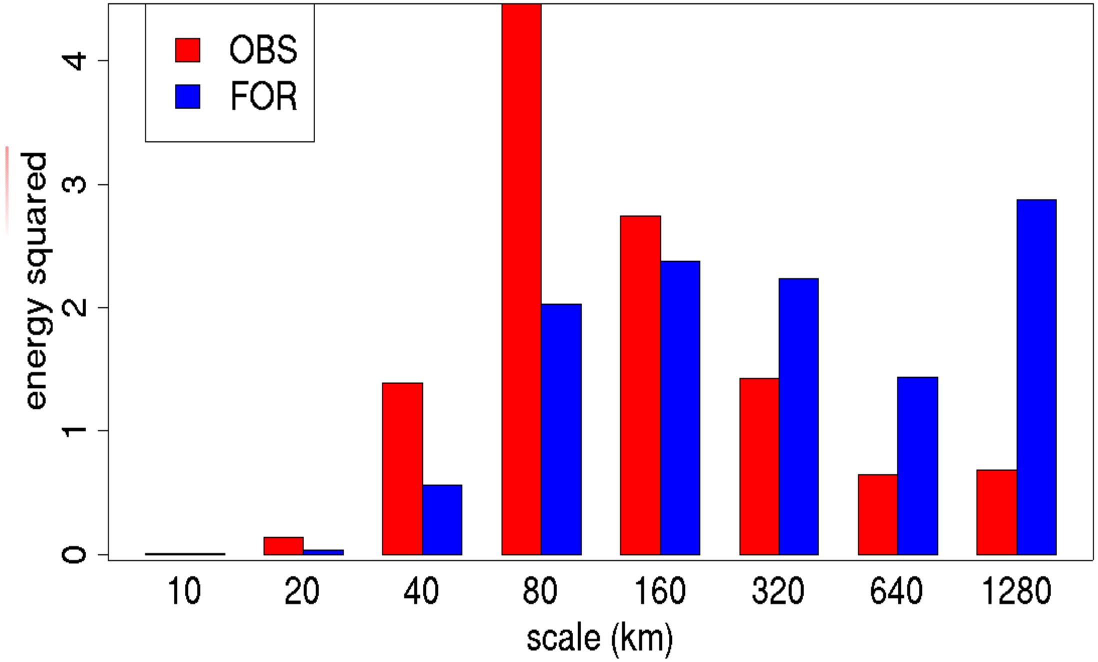


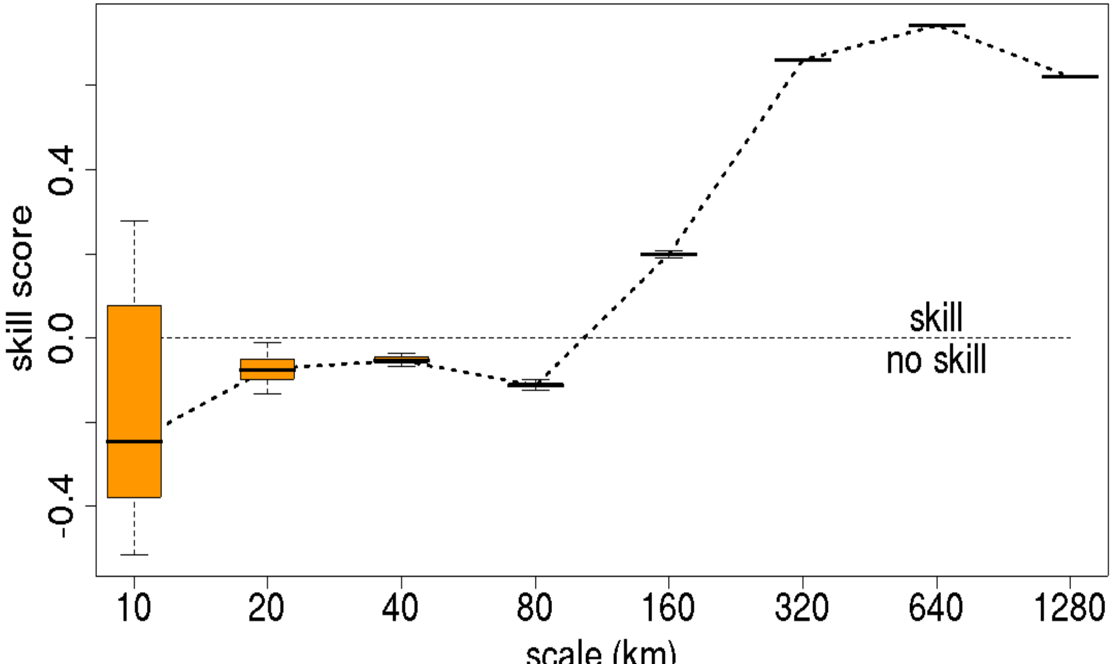
Confidence (uncertainty) mask



For each scale (e.g. 160 km resolution scale) provide <u>confidence/uncertainty associated to reconstructed fields</u>

large number of gauges ←→ confidence
small number of gauges ←→ uncertainty





5. Verification

on different scales, but only **where obs are available**

1. Energy squared:

$$En^2(X) = \langle X^2 \rangle$$

Measures the quantity of events and their intensity at each scale => BIAS, scale structure

2. MSE Skill Score:

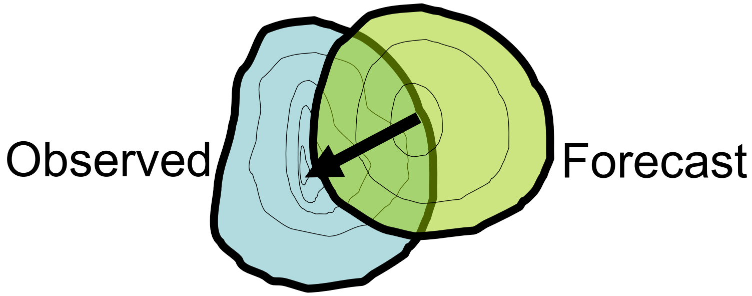
$$1 - \frac{MSE(Y,X)}{En^2(X) + En^2(Y)}$$

(related to correlation)

Feature-based approach (CRA)

Ebert and McBride, *J. Hydrol*., 2000

- Define entities using threshold (Contiguous Rain Areas)
- Horizontally translate the forecast until a *pattern matching* criterion is met:
  - minimum total squared error between forecast and observations
  - maximum correlation
  - maximum overlap
- The displacement is the vector difference between the original and final locations of the forecast.



CRA error decomposition

Total mean squared error (MSE)

$$MSE_{total} = MSE_{displacement} + MSE_{volume} + MSE_{pattern}$$

The *displacement error* is the difference between the mean square error before and after translation

$$MSE_{displacement} = MSE_{total} - MSE_{shifted}$$

The *volume error* is the bias in mean intensity

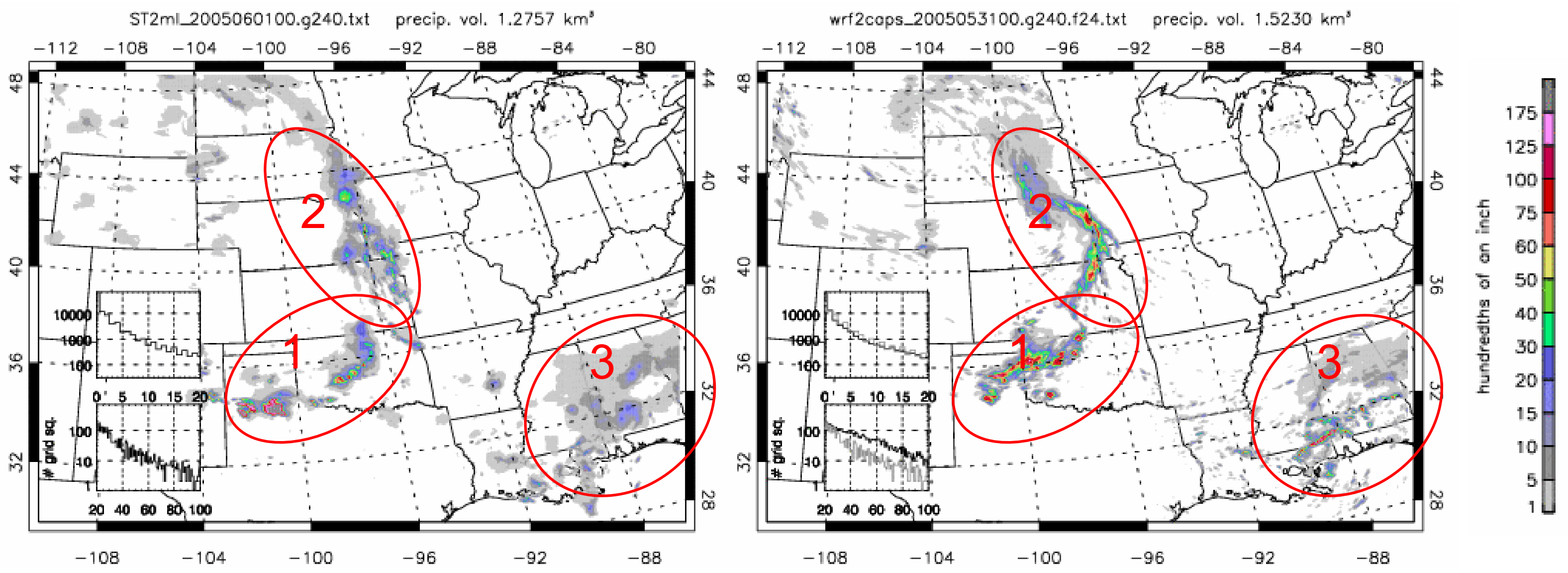
$$MSE_{volume} = (\overline{F} - \overline{X})^2$$

where $\overline{F}$ and $\overline{X}$ are the mean forecast and observed values after shifting.

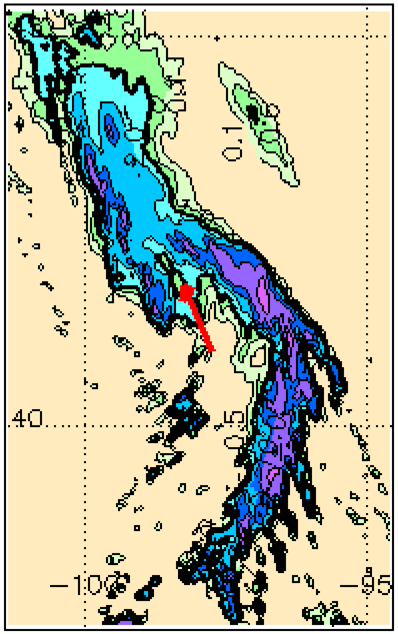
The *pattern error*, computed as a residual, accounts for differences in the fine structure,

$$MSE_{pattern} = MSE_{shifted} - MSE_{volume}$$

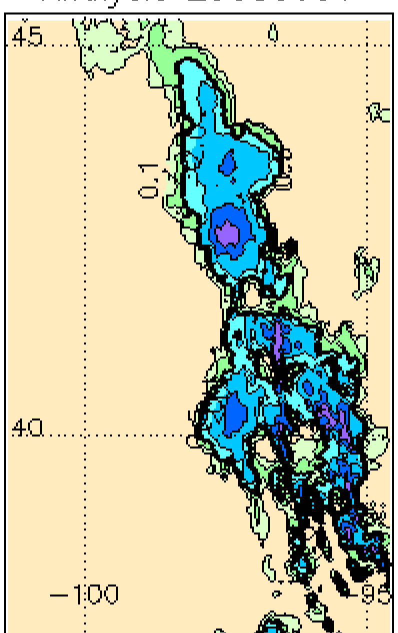
CRA verification of precipitation forecast over USA



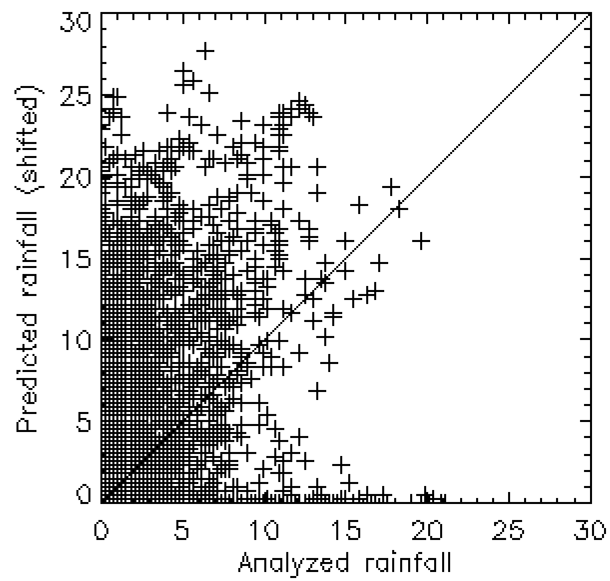
wrf2 fcst 20050601 hour 00–24



Analysis 20050601



CRA 20050601



wrf2 24h fcst 20050601 n=11007
(37.52°,−101.29°) to (45.29°,−94.65°)
Verif. grid=0.042° CRA threshold=1.0 mm/h

	Analysed	Forecast
# gridpoints ≧1 mm/h	4840	5699
Average rainrate (mm/h)	1.52	2.68
Maximum rain (mm/h)	21.08	27.69
Rain volume (km^3)	0.26	0.46

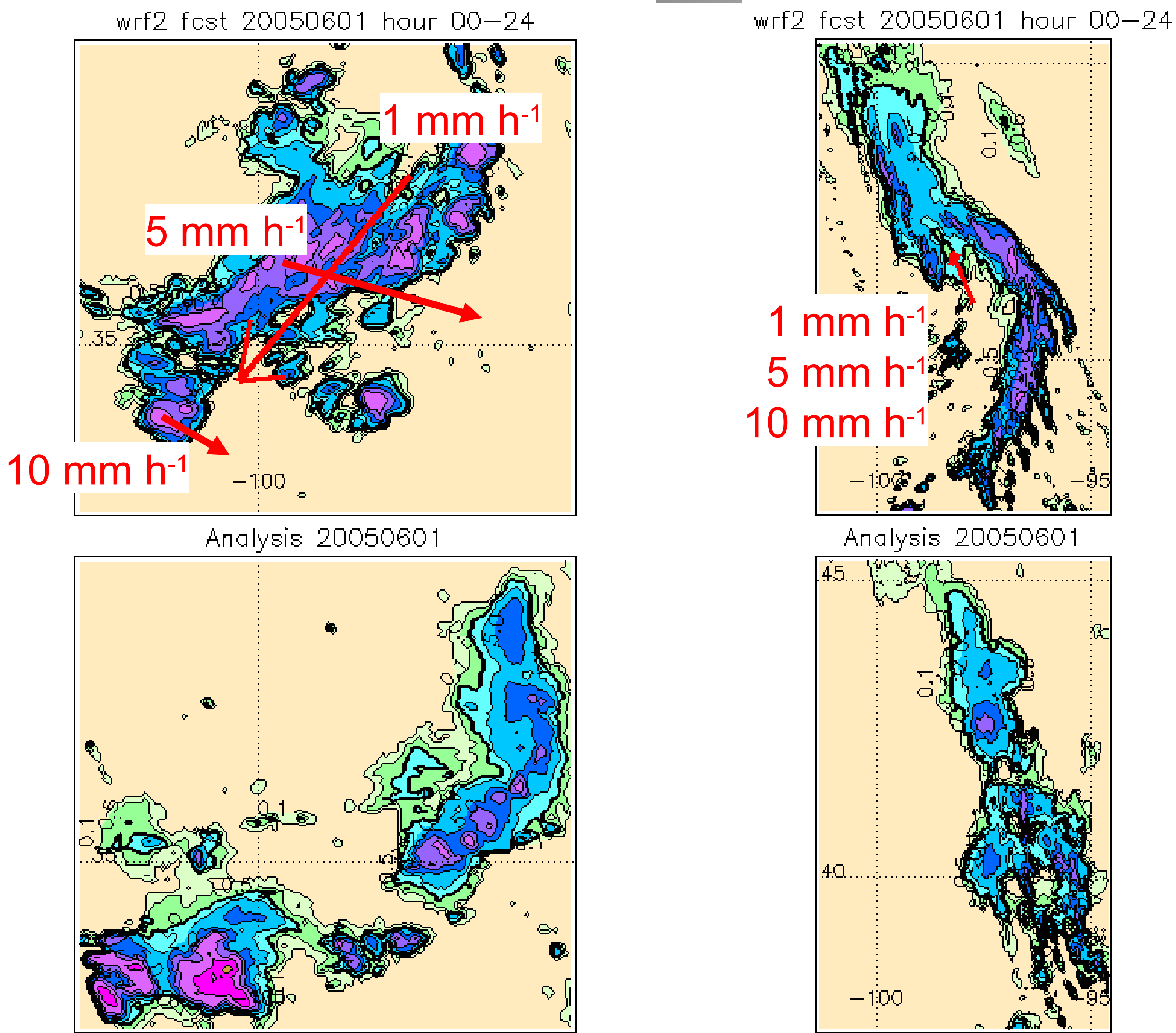
Displacement (E,N) = [0.52°,−0.84°] max.corr matching

	Original	Shifted
RMS error (mm/d)	5.11	4.65
Correlation coefficient	−0.040	0.193

Error Decomposition:

Displacement error	18.7%
Volume error	4.9%
Pattern error	76.4%

Sensitivity to rain threshold



SAL (Wernli et al, MWR, 2008)

- 3 parameter characterization of field of objects
- Structure – Amplitude – Location
- Applied to precipitation

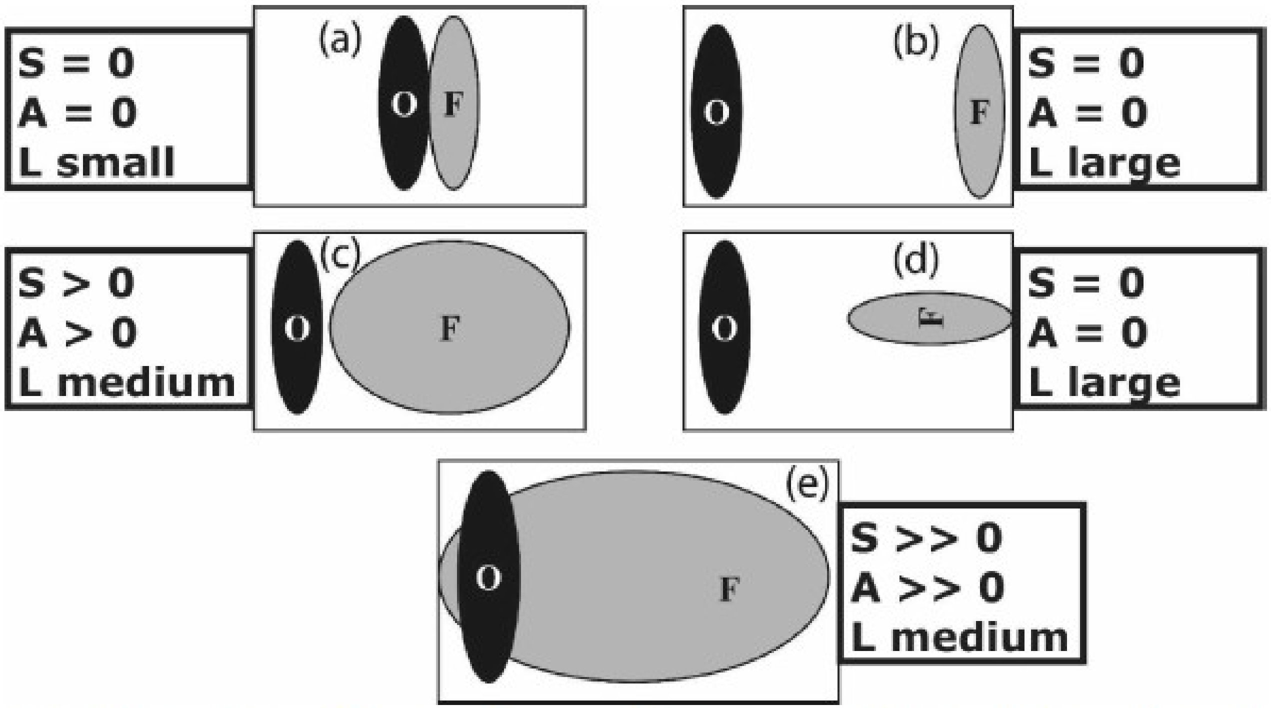
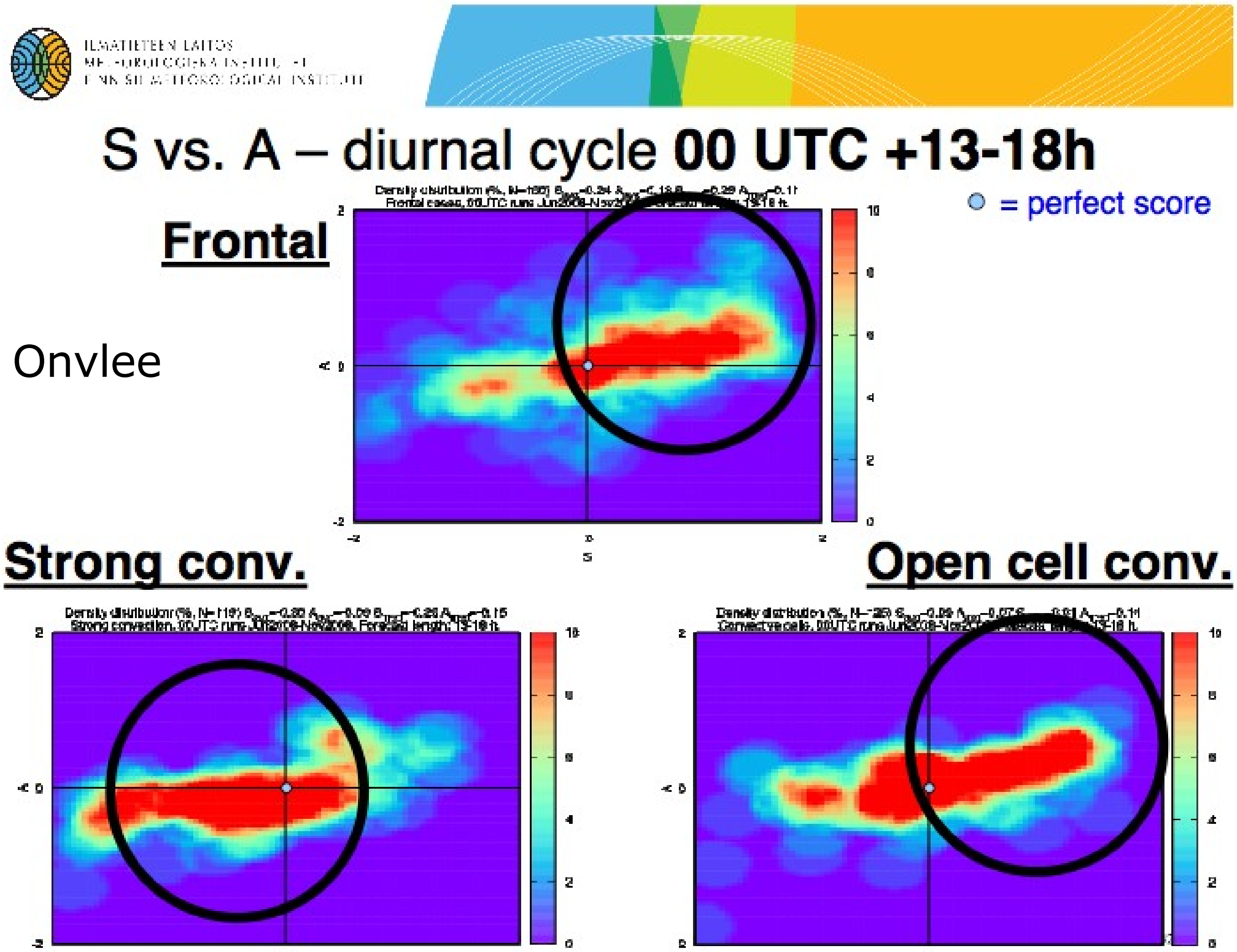


FIG. 1. A schematic example of various forecast and observation combinations, modified from Davis et al. (2006a). For the qualitative application of SAL, it was assumed that precipitation rates are uniform and the same in all objects.

Diagnostic research using SAL – The Grey Zone

Courtesy Jeanette Onvlee



SAL for Midwest US precipitation case

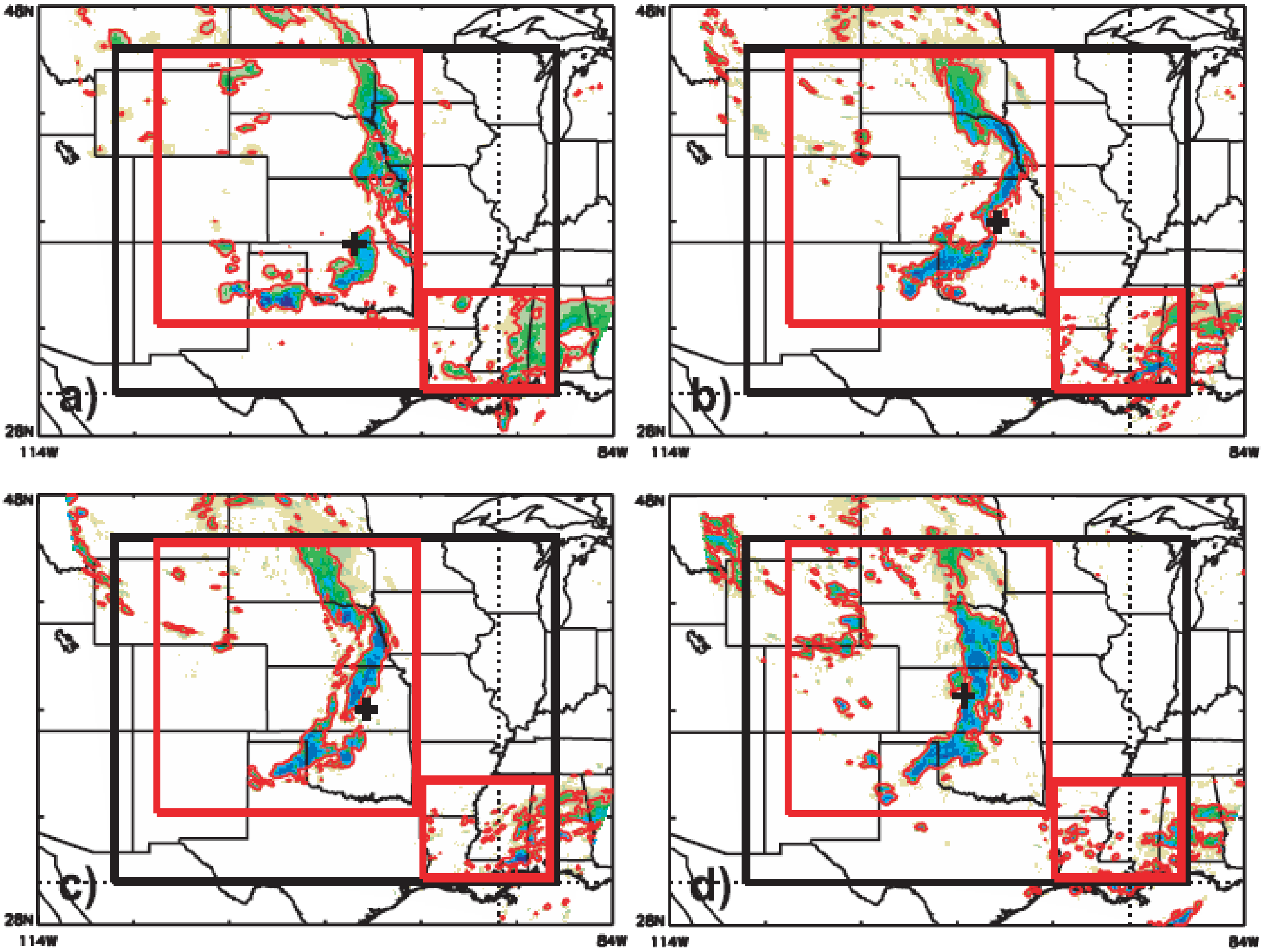
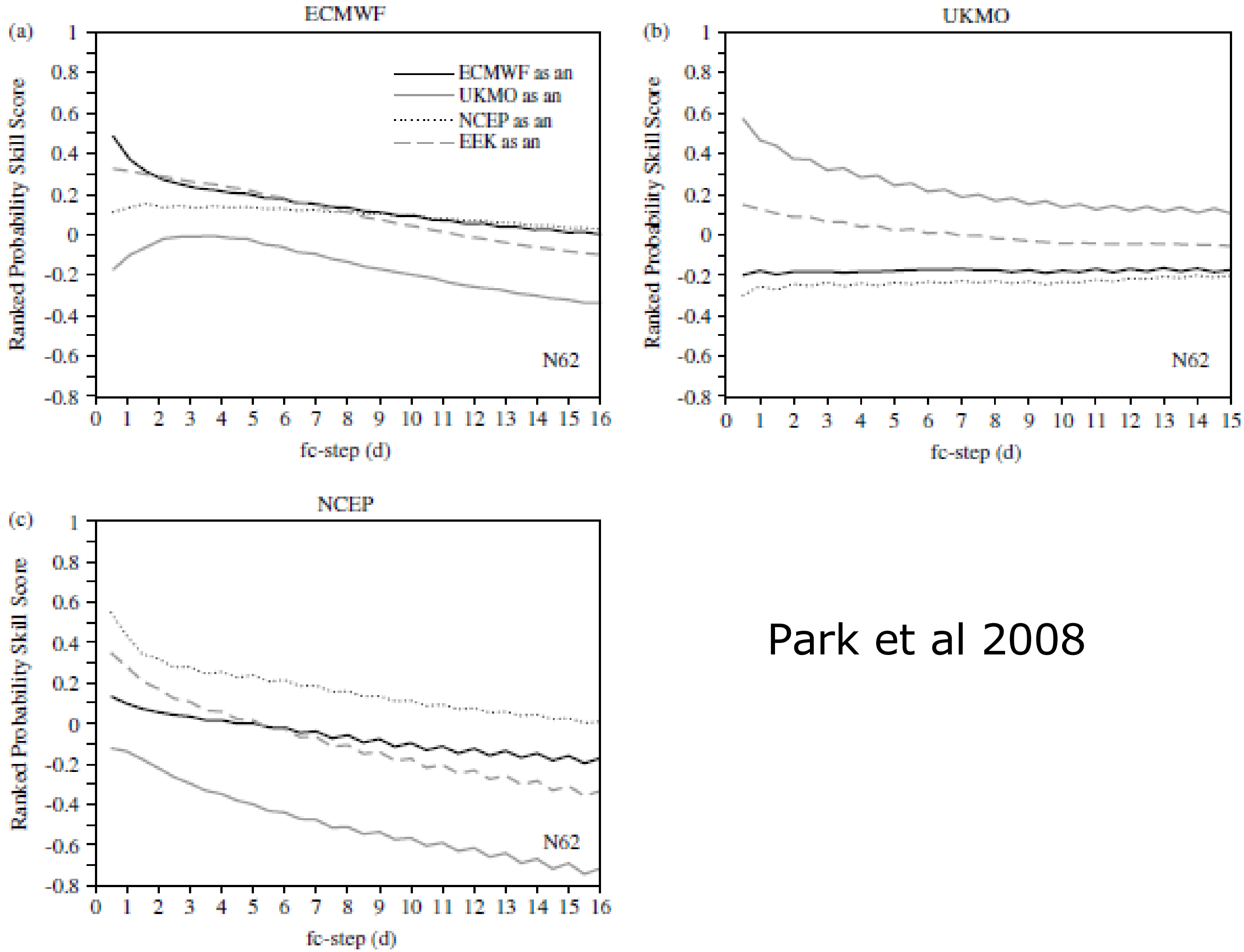


FIG. 4. Precipitation fields on 1 Jun 2005: (a) The observations and forecasts from the (b) 2CAPS, (c) 4NCAR, and (d) 4NCEP models. The rectangles show the three domains used for the SAL analysis, referred to as the large (black line), and northern and southern domains (both red lines), respectively. The black plus sign denotes the center of mass of the precipitation in the large domain.

Towards proper verification practice: When or not to use model-tainted observation data

- Data assimilation systems are designed to merge models and data
- Verification: Ideally need data that are from completely independent sources
- Verification against analysis
  - Fine when only one model is involved, depending on user of verification
- For comparison
  - Each own analysis (WMO method)
- Verification against observations
  - Model dependent too if model used in qc (WMO method)
  - Remotely sensed data
- More complicated when models or ensembles are combined
  - Use ensemble of analyses
  - Randomly select analysis from among models in multimodel ensemble
- Also for reanalysis data used as climatology

Verification results depend on analysis used



Park et al 2008

Verification and the goals of TIGGE

- Goals:
  - Enhance collaborative research
  - Enable evolution towards GIFS
  - Develop ensemble combination methods; bias removal
- Essential question: If we are going to move towards a GIFS, then we must demonstrate that the benefits of combined ensembles are worth the effort with respect to single-center ensembles. OR: Do we get a “better” pdf by merging ensembles?
- Verification – Relevant, user-oriented

European Precipitation Verification

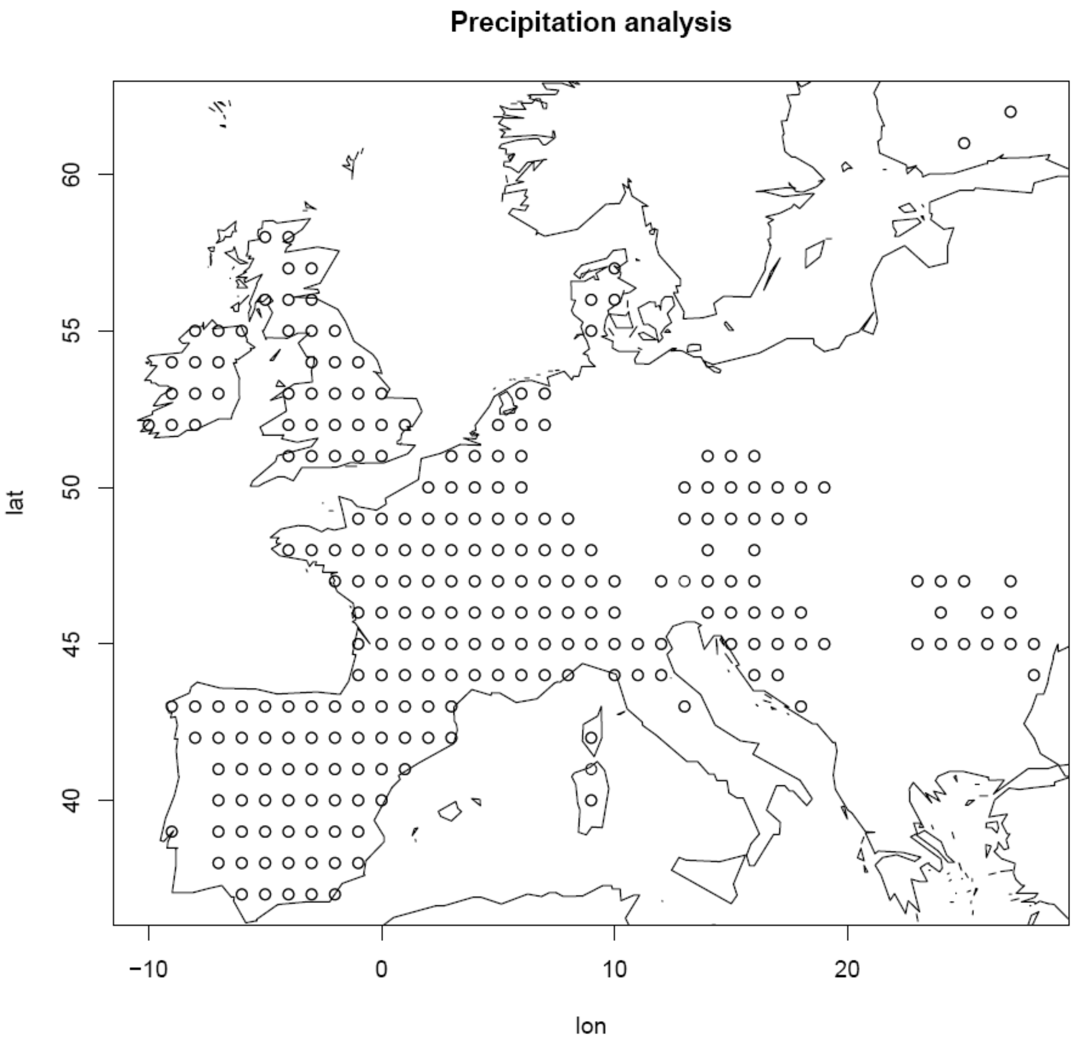
-Upscaled observations according to Cherubini et al (2002)

-OBS from gauges in Spain, Portugal, France, Italy, Switzerland, Netherlands, Romania, Czech Republic, Croatia, Austria, Denmark, UK, Ireland, Finland and Slovenia

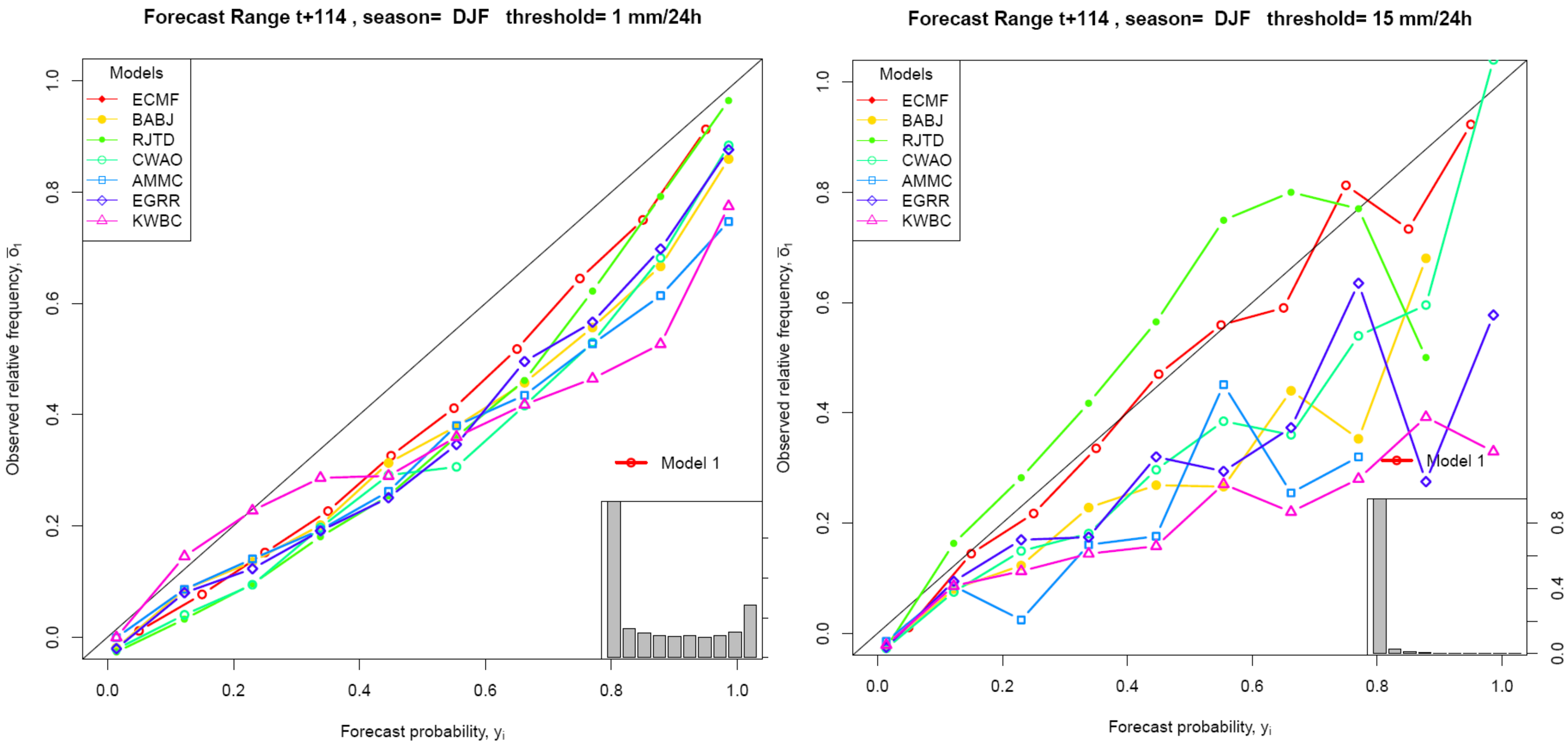
-At least 9 stns needed per grid box to estimate average

-24h precip totals, thresholds 1,3,5,10,15,20,25,30 mm

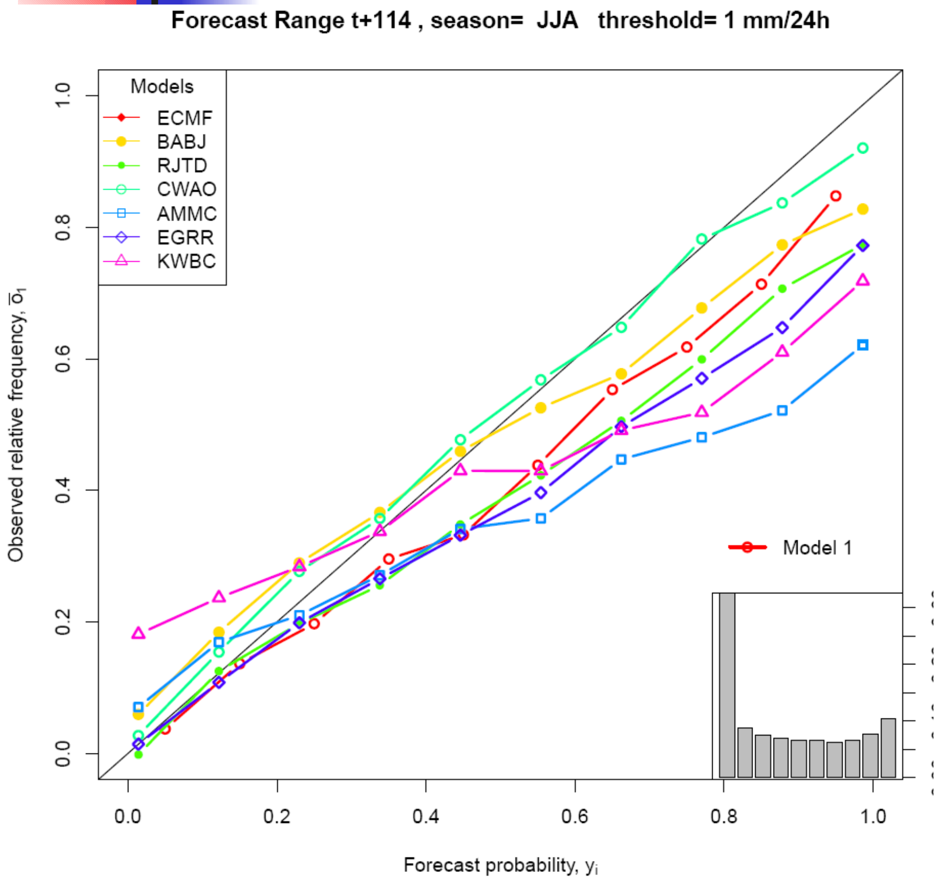
-one year (oct 07 to oct 08

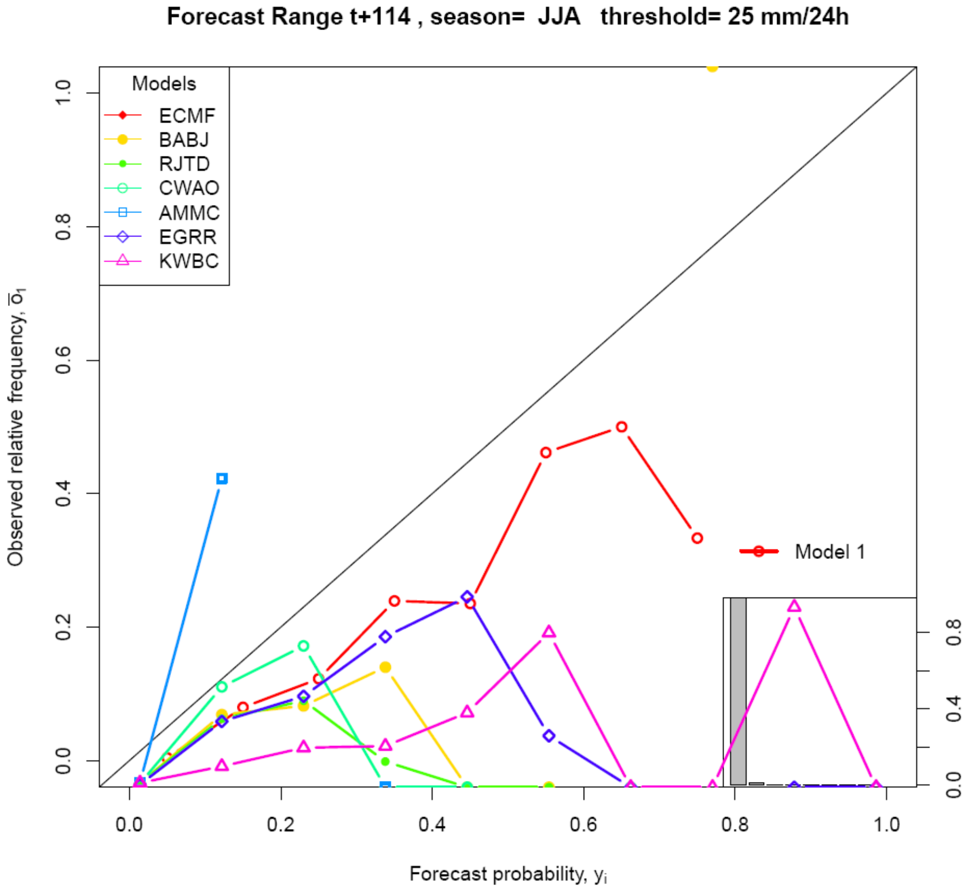


Reliability – Winter 07-08 – Europe – 114h

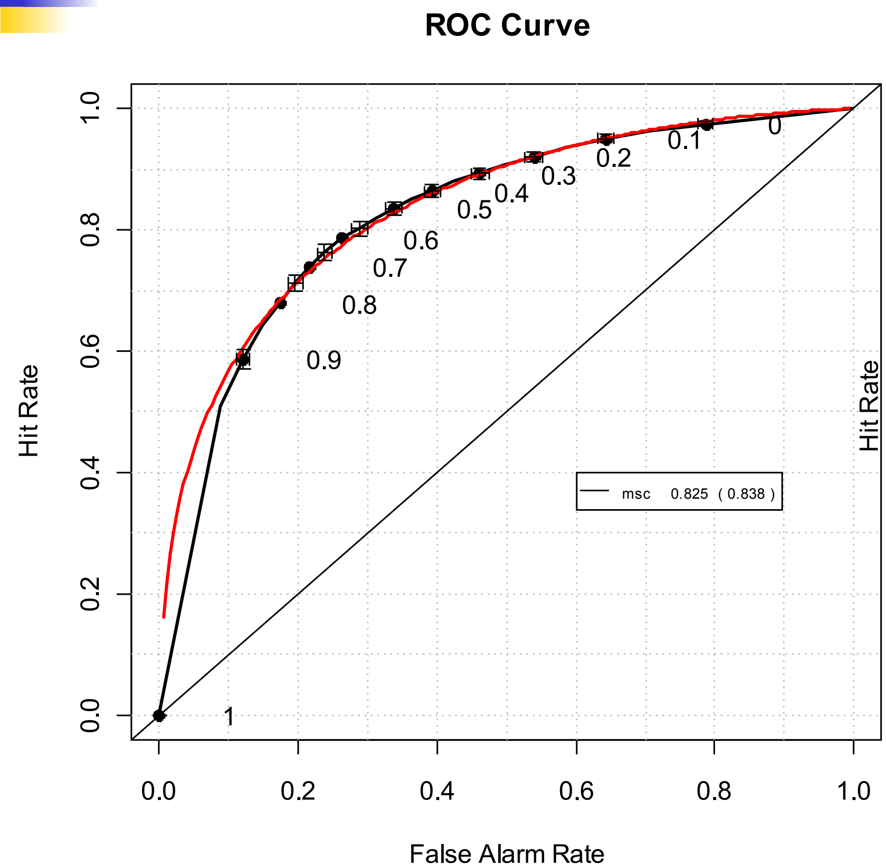


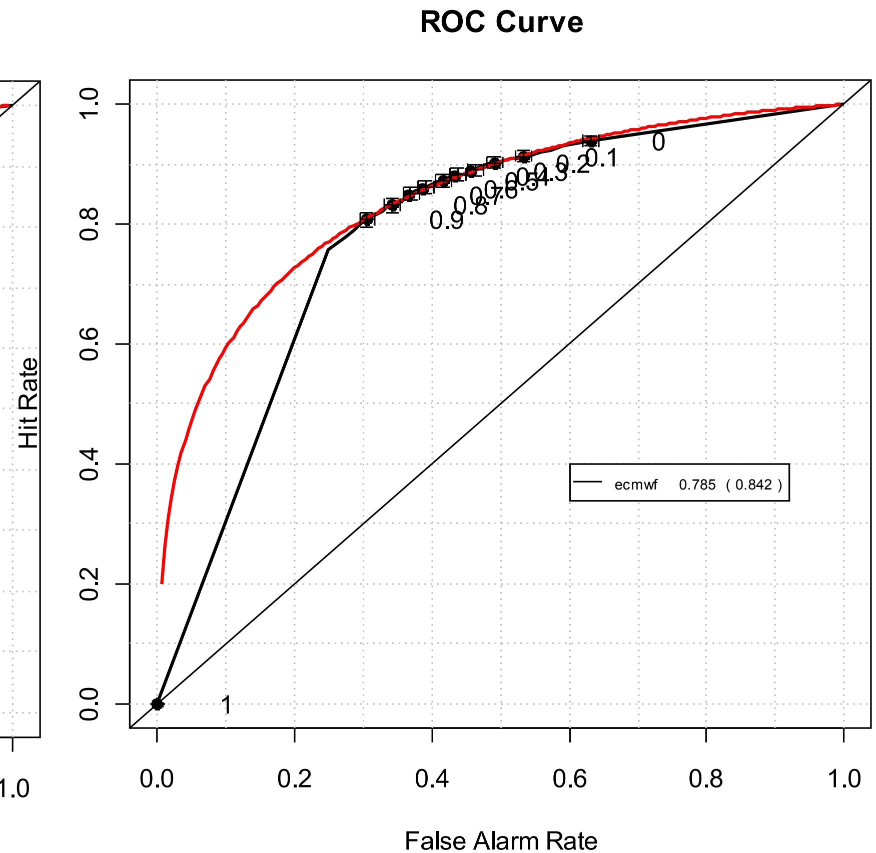
Reliability – Summer 08- Europe 114 h



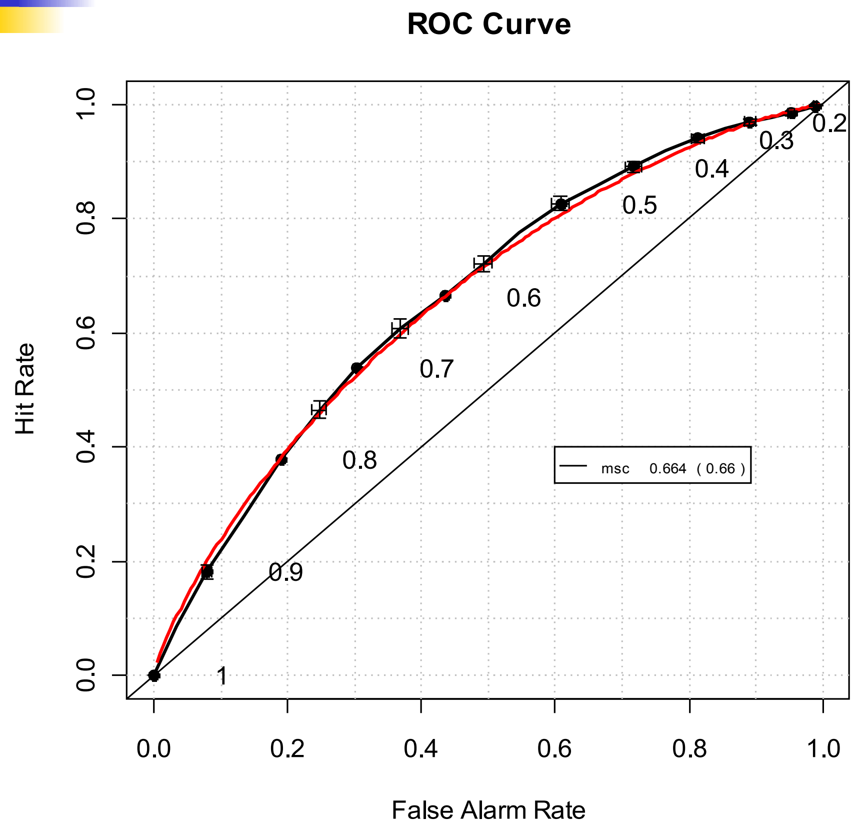


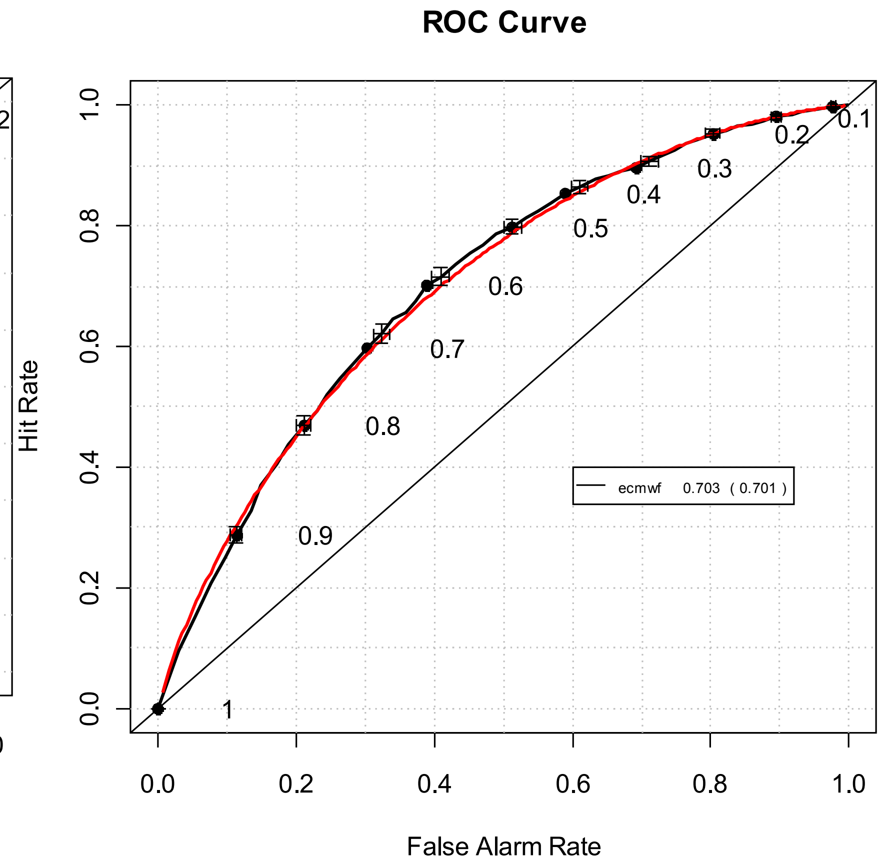
Results – Canada – ROC curves – 24h



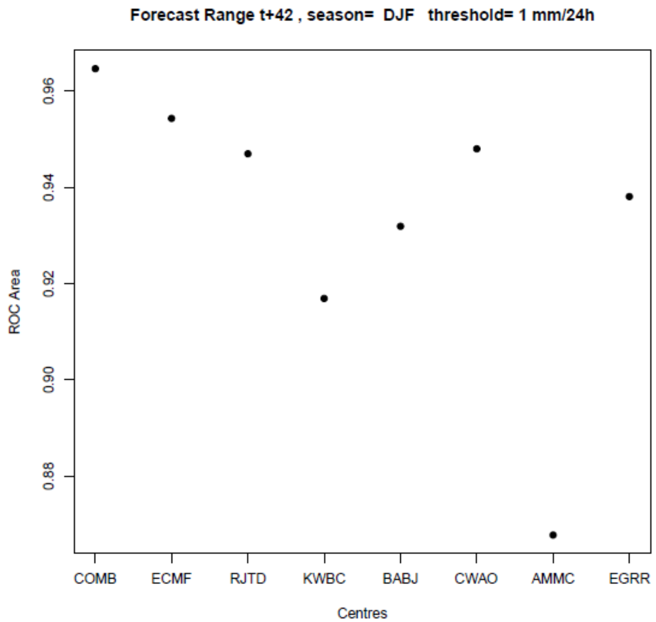


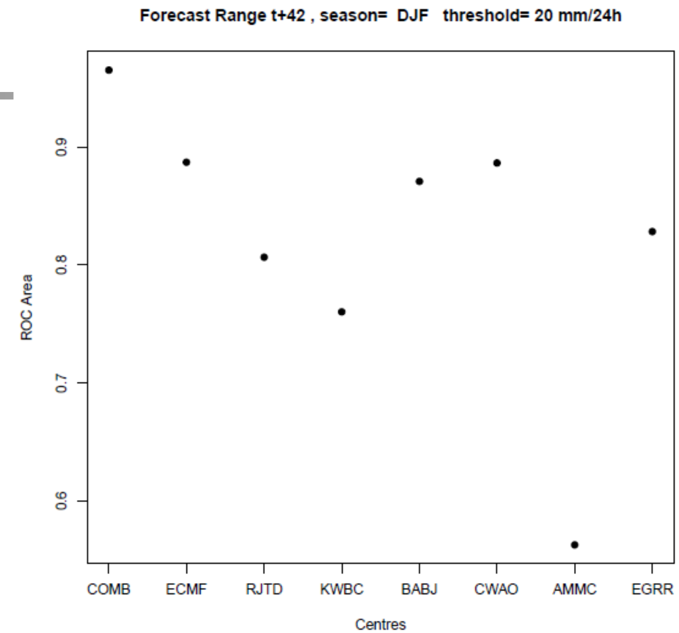
Results – Canada – ROC Curves – 144h

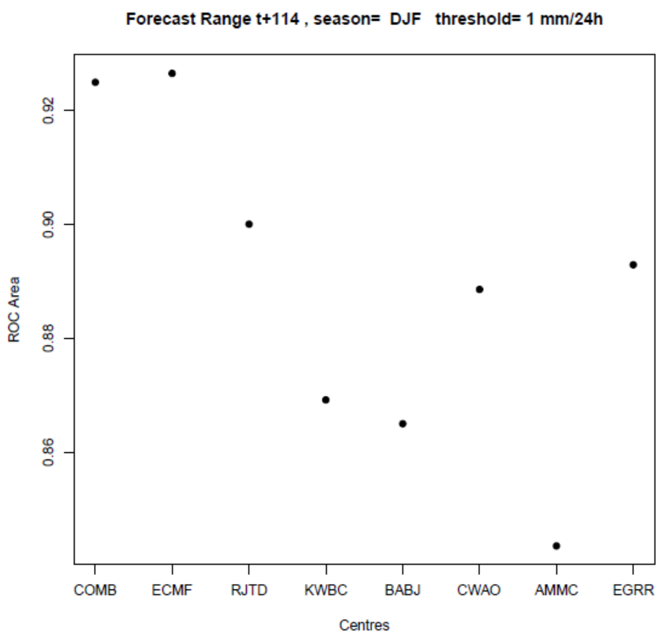


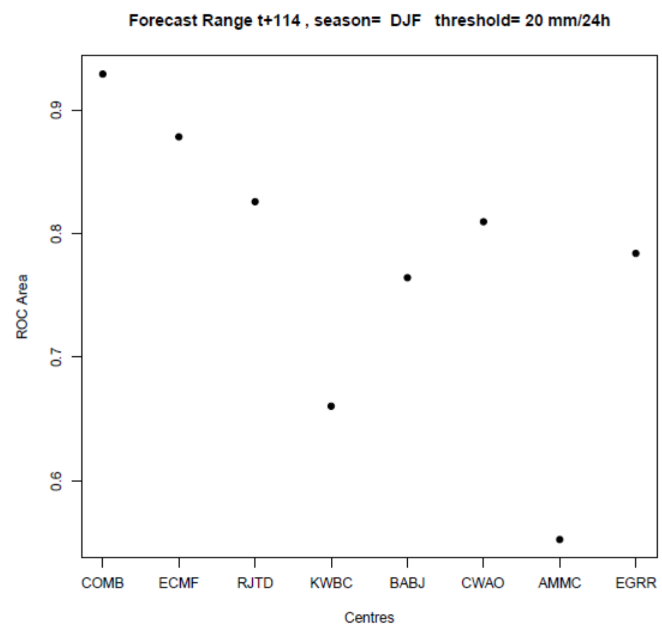


With Ensemble Combination









5th International Verification Methods Workshop Melbourne, Australia, Dec 1-7 2011

-Anticipate joint SERA participation, with overlap

-can accommodate 40 students

-similar format to previous: 3 day tutorial, one day break, then 3 day scientific conference

View from break-out area

Summary

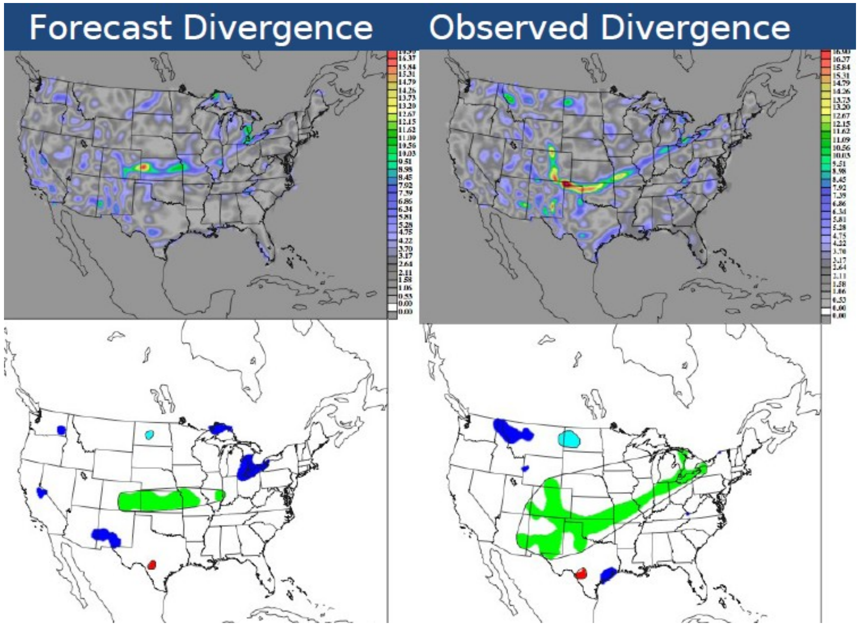
- Verification is becoming more user-oriented
- Extensions of standard verification methods to ensembles and for extreme weather
- Lots of spatial verification methods proposed, some are beginning to catch on in the broader community
- Still striving for “best verification practices” in the international community (and here too!)
  - Model-tainted data
  - Confidence intervals on verification results

Thanks!

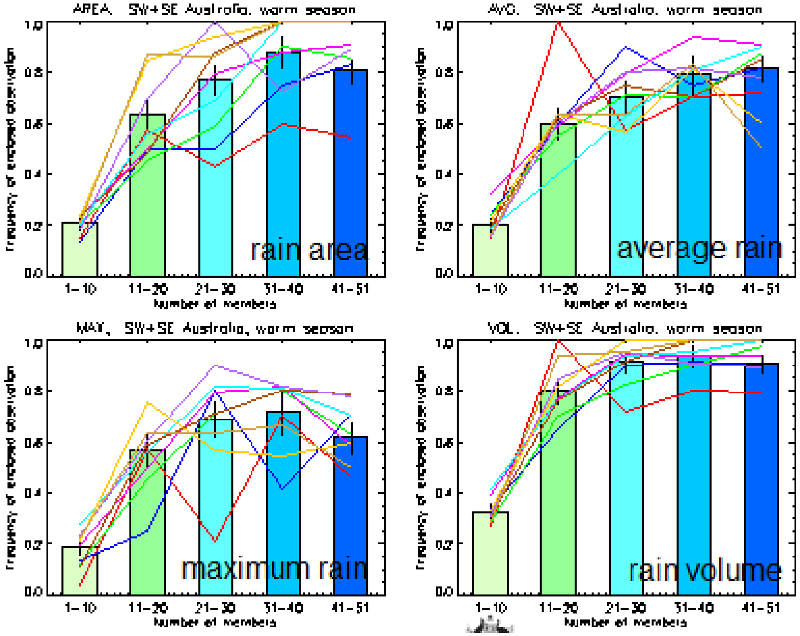
Workshop: New verification research

Spatial methods applied to:

Wind fields

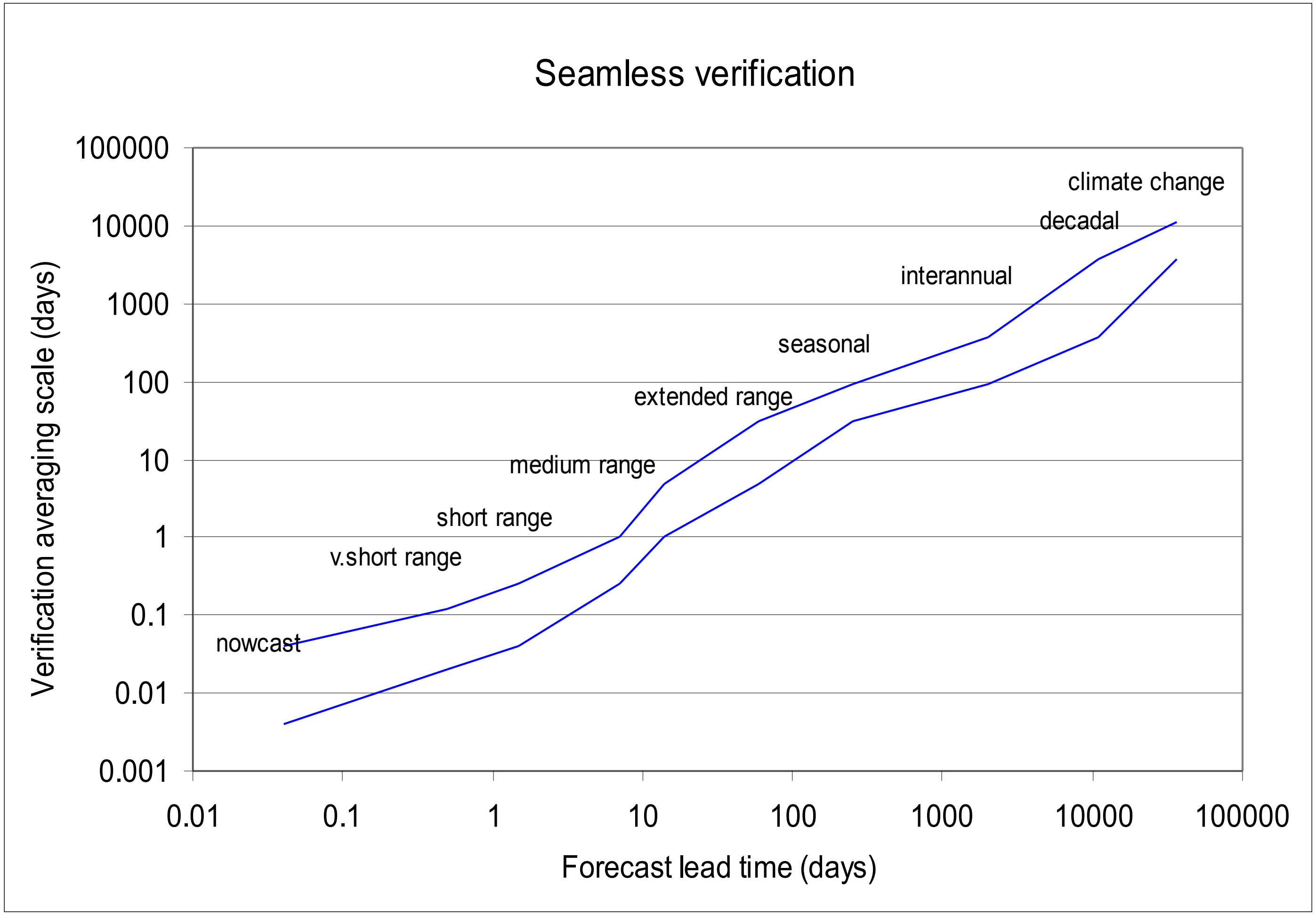


Ensemble forecasts



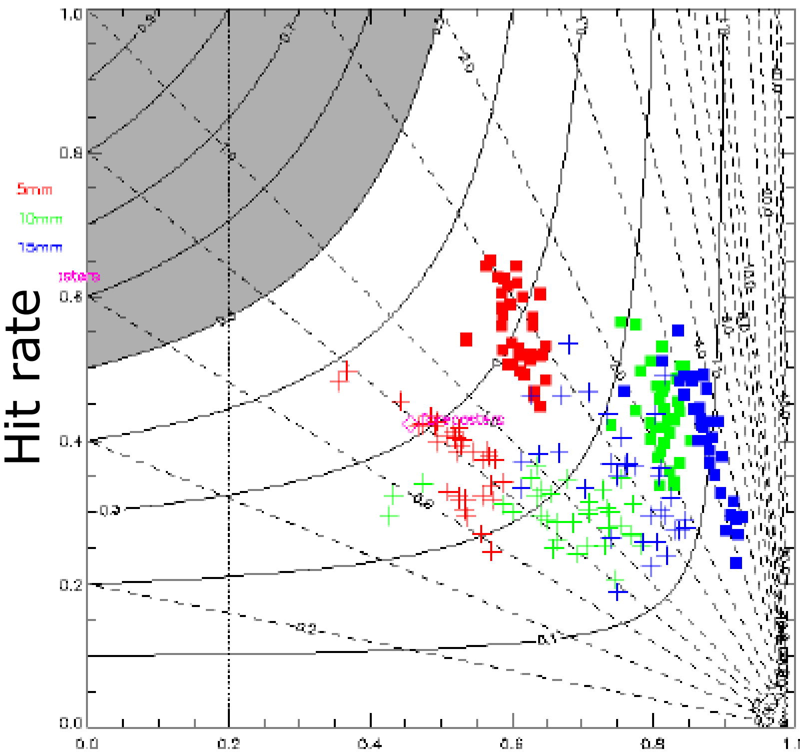
http://www.space.fmi.fi/Verification2009/

Verification across space and time scales (a.k.a. "seamless")



Workshop: New verification research

Diagnostics

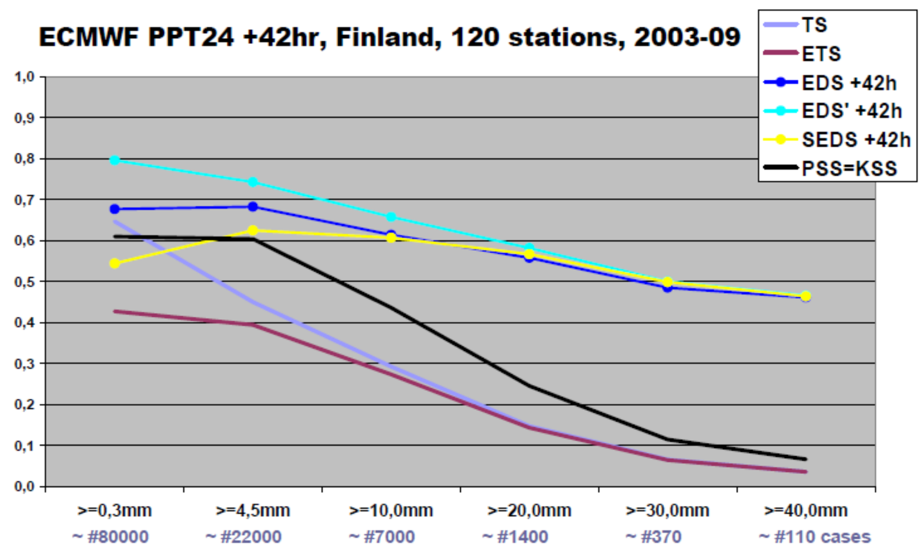


Hit rate

False alarm ratio

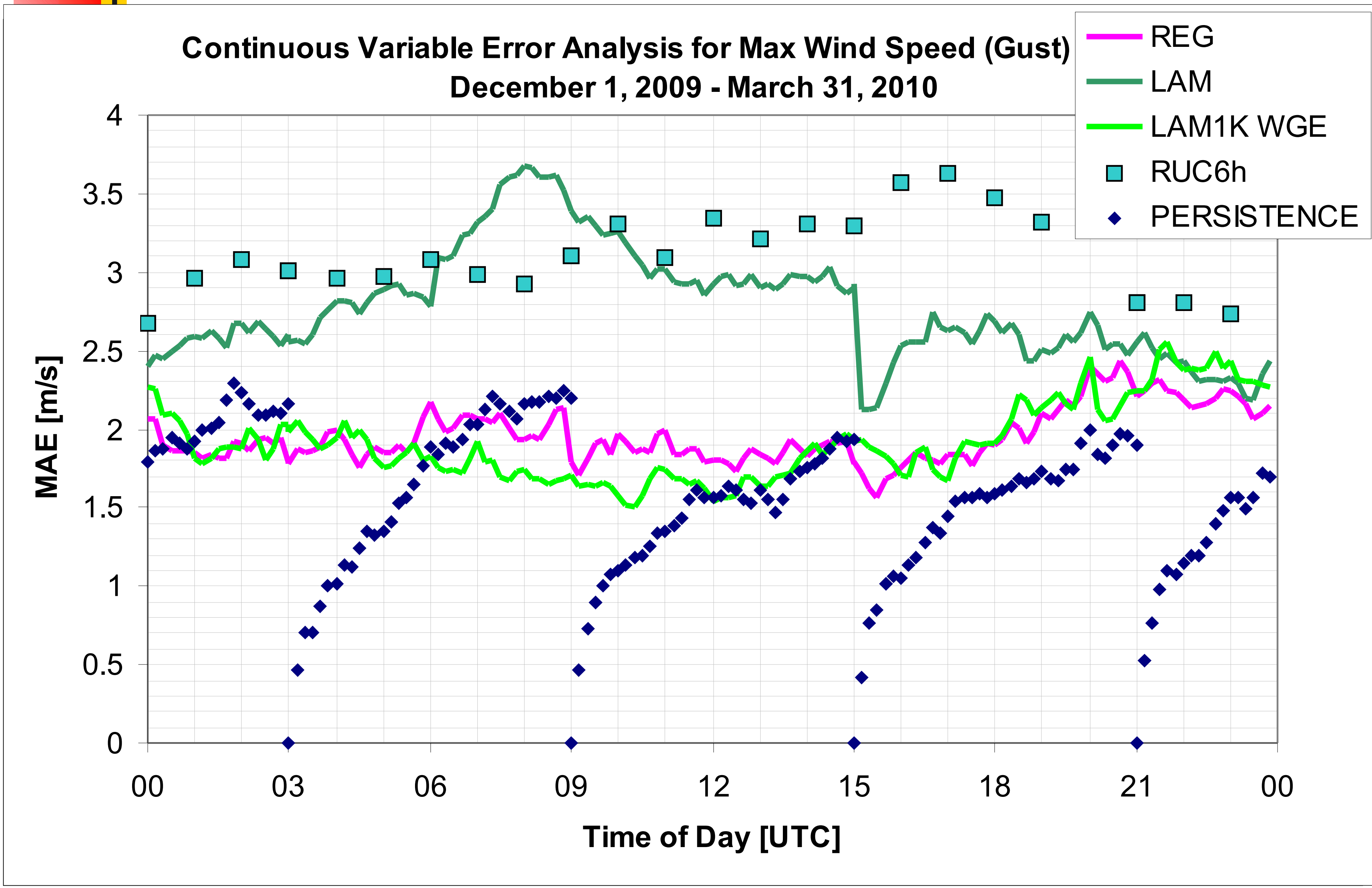
Extremes

$$SEDS = \frac{\log[(a+b)/n] + \log[(a+c)/n]}{\log(a/n)} - 1$$



http://www.space.fmi.fi/Verification2009/

Verification



Aerosol Verification