Spatial forecast verification

Manfred Dorninger
University of Vienna
Vienna, Austria
manfred.dorninger@univie.ac.at

Thanks to: B. Ebert, B. Casati, C. Keil

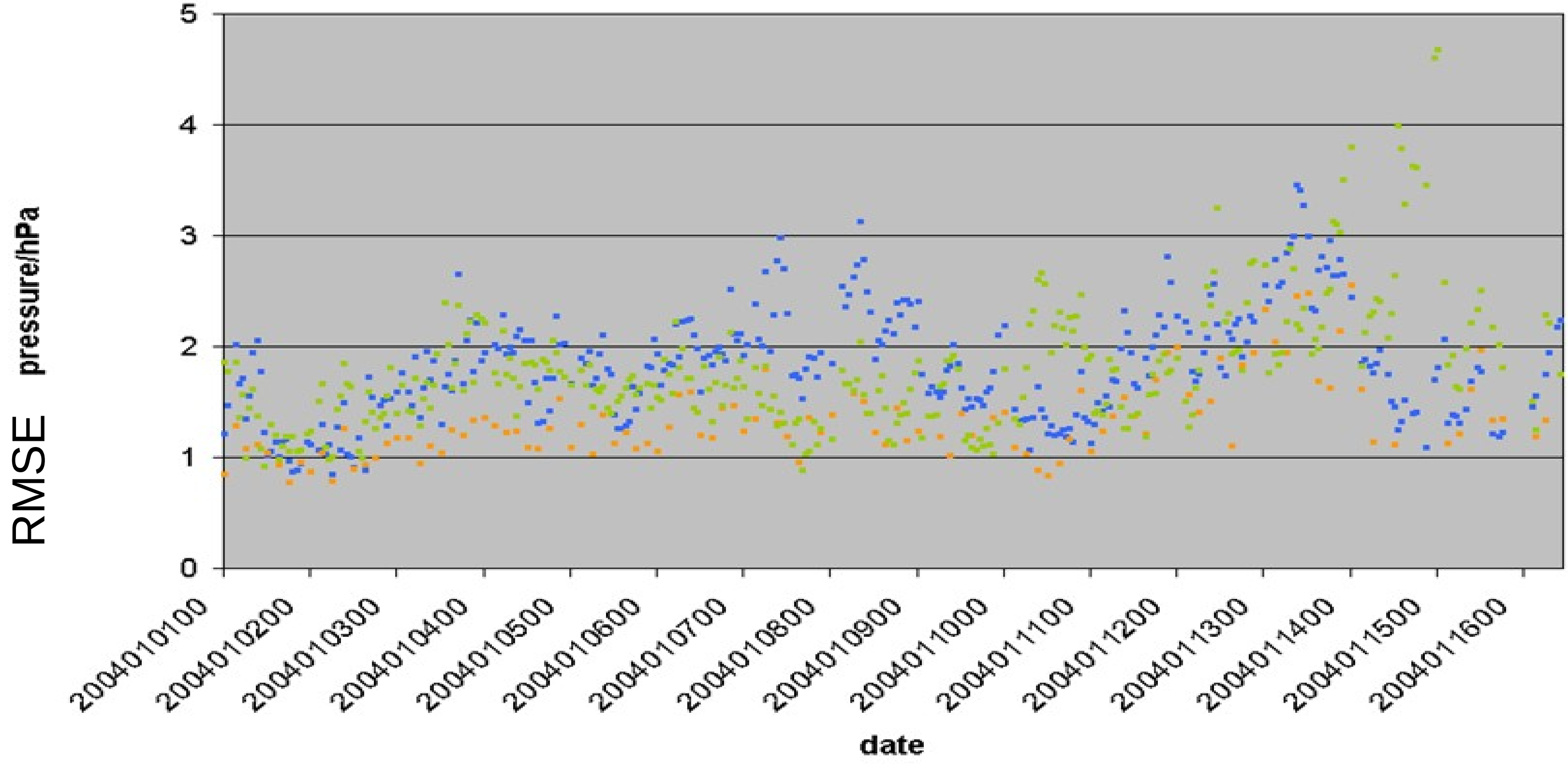
7th Verification Tutorial Course, Berlin, 3-6 May, 2017

Motivation:

Model – VERA: RMSE pmsl, 13-24 h

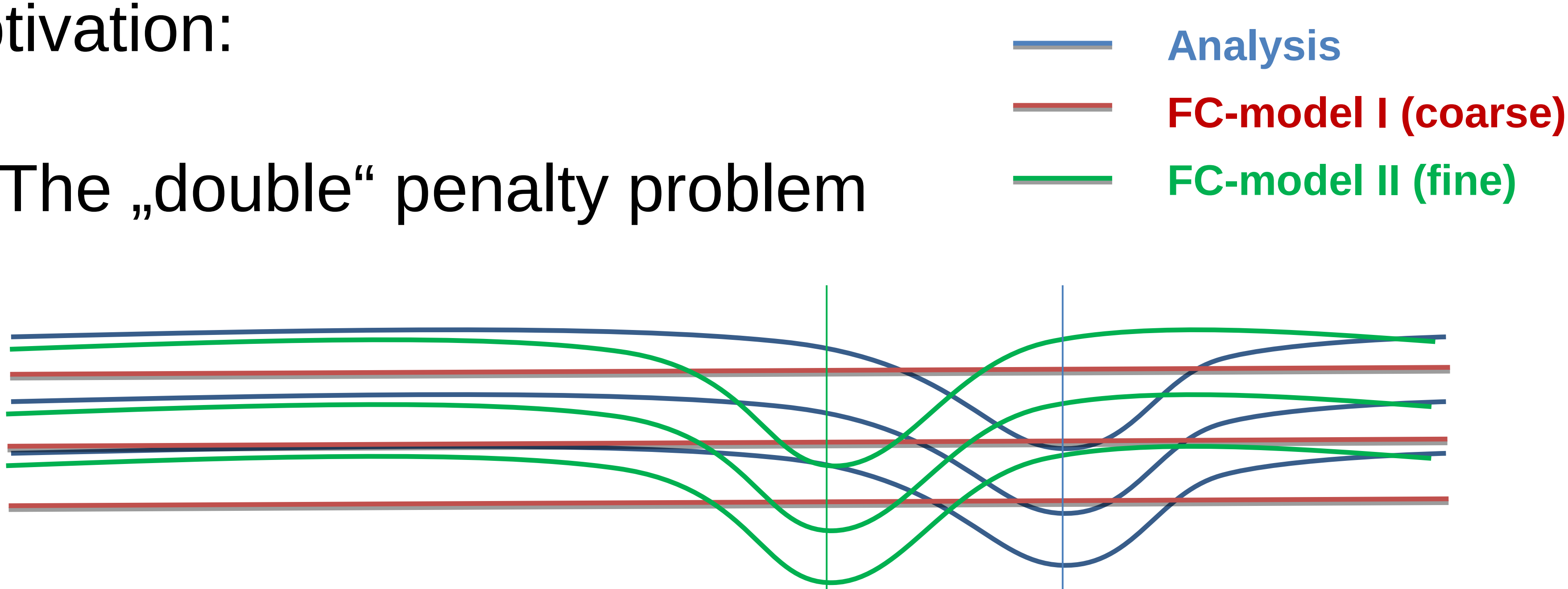
Ala 0)

~9 km ~25 km ~2 km

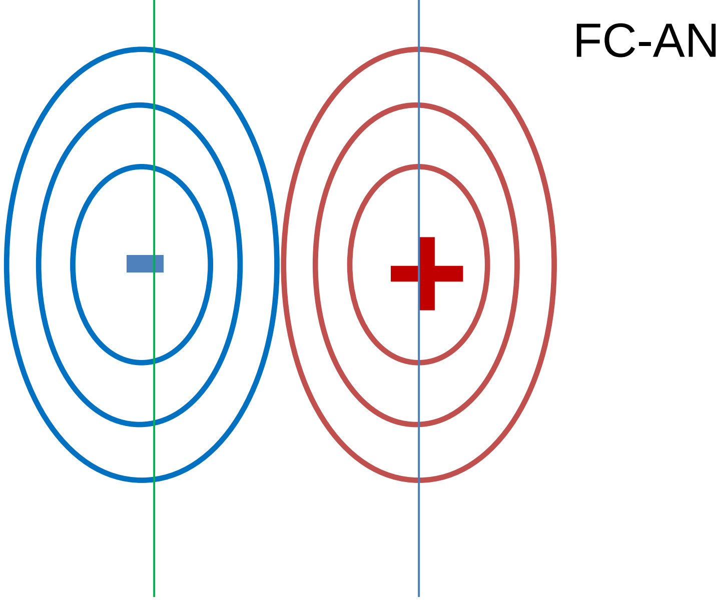


Motivation:

→ The „double“ penalty problem



- fine-scale model catches the small-scale trough but at the wrong place (or time)
- gets penalized twice
- increases quadratic measures compared to coarse model
- true for other continuous variables as well (e.g., precipitation, wind speed, etc.)



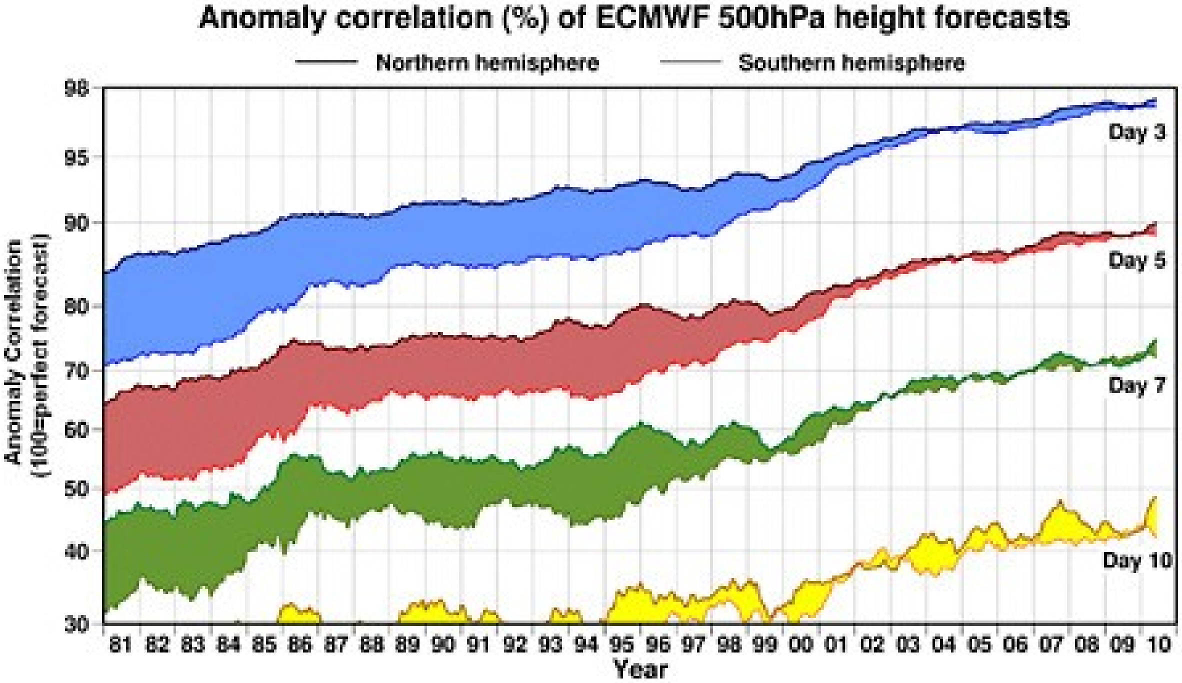
Traditional spatial verification (grid point wise approach)

Compute statistics on forecast-observation pairs

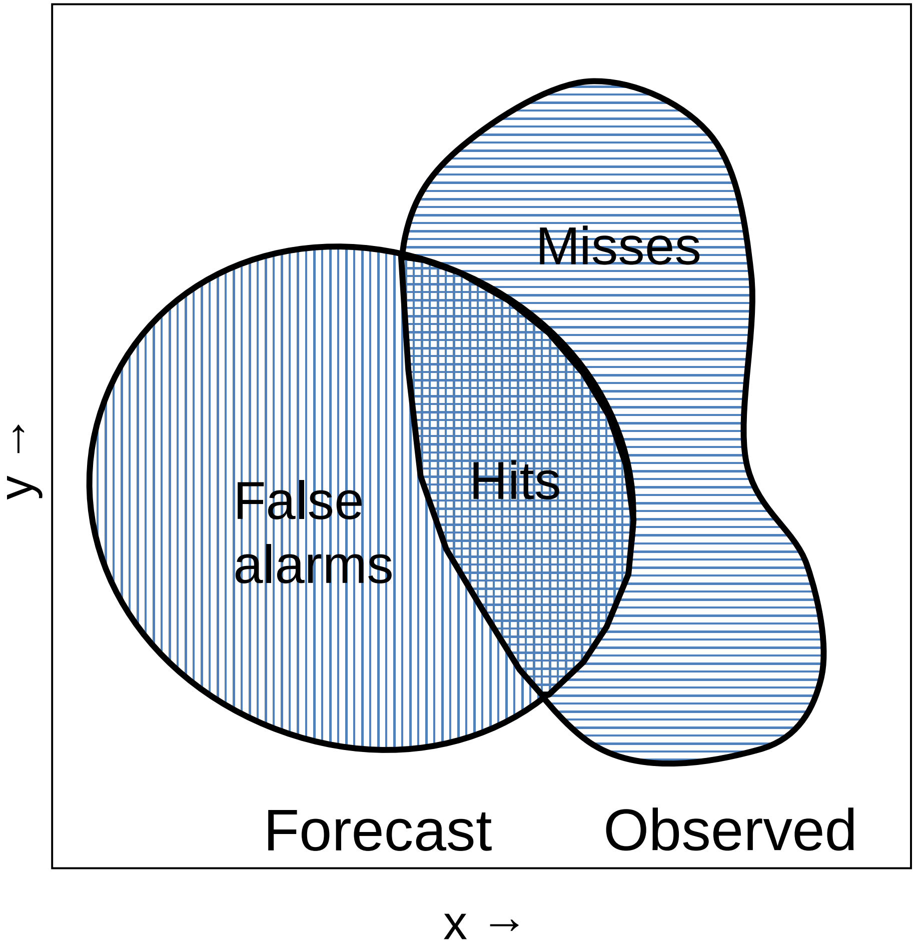
- Continuous values (e.g., precipitation amount, temperature, NWP variables):
  - mean error, MSE, RMSE, correlation
  - anomaly correlation, S1 score
- Categorical values (e.g., precipitation occurrence):
  - Contingency table statistics (POD, FAR, etc…)

Anomaly correlation

$$AC = \frac{\sum (F-C)(O-C)}{\sqrt{\sum (F-C)^2}\sqrt{\sum (O-C)^2}} \text{ (uncentered)} \quad \text{or} \quad \frac{\sum \left((F-C) - \overline{(F-C)}\right)\left((O-C) - \overline{(O-C)}\right)}{\sqrt{\sum (F-C)^2}\sqrt{\sum (O-C)^2}} \text{ (centered)}$$



Traditional spatial verification using categorical scores



Contingency Table

		Observed	
		yes	no
Predicted	yes	*hits*	*false alarms*
	no	*misses*	*correct negatives*

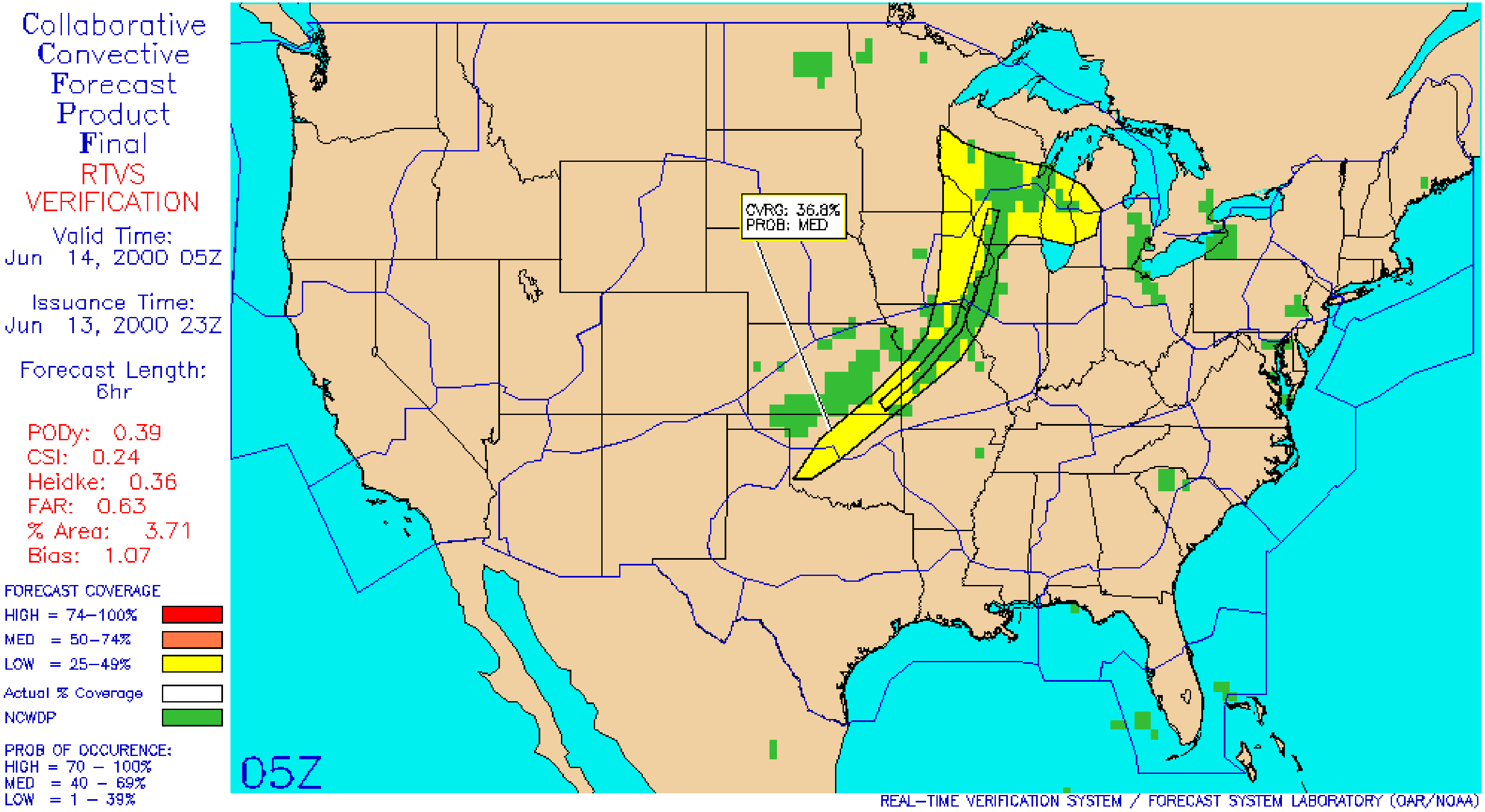
$$FBI = \frac{hits + false\ \ alarms}{hits + misses}$$

$$POD = \frac{hits}{hits + misses} \qquad FAR = \frac{false\ alarms}{hits + false\ alarms}$$

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

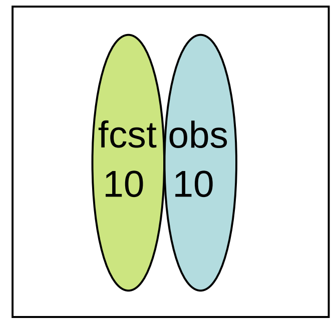
$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

POD=0.39, FAR=0.63, CSI=0.24

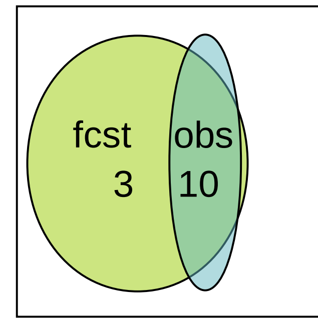


Traditional spatial verification

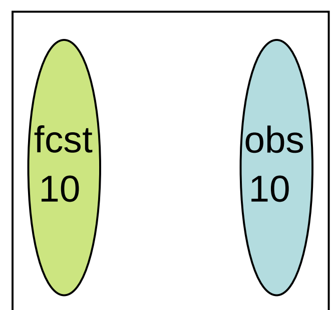
- Requires an exact match between forecasts and observations at every grid point
  - Problem of "double penalty" - event predicted where it did not occur, no event predicted where it did occur
- Traditional scores do not say very much about the source or nature of the errors



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

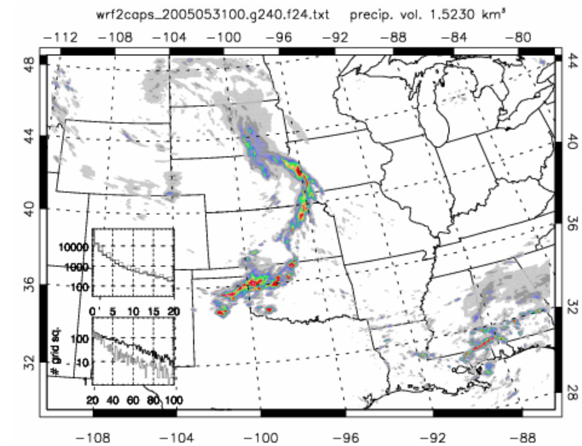


What do traditional scores not tell us ?

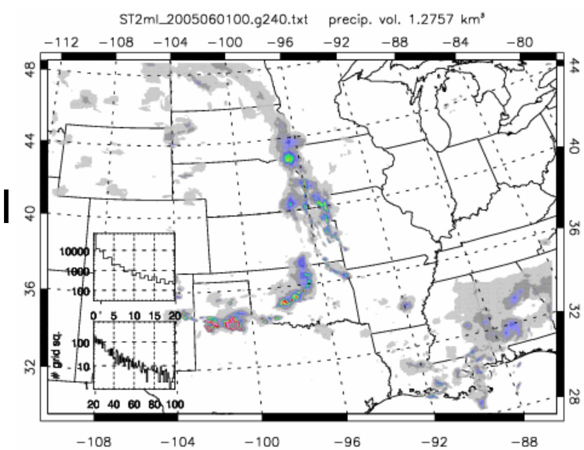
- Traditional approaches provide overall measures of skill but…
- They don't provide much ***diagnostic*** information about the forecast:
  - What went wrong? What went right?
  - How close is the forecast to observation (in terms of spatial thinking)?
  - Does the forecast look realistic?
  - How can I improve this forecast?
  - How can I use it to make a decision?
- Best performance for ***smooth*** forecasts !!!
- Some scores are insensitive to the ***size*** of the errors…

Spatial forecasts

WRF model



Stage II radar



Weather variables defined over spatial domains have **coherent spatial structure and features**

New spatial verification techniques aim to:

- account for field spatial structure
- provide information on error in physical terms
- account for uncertainties in location (and timing)

Spatial verification types

Neighborhood (fuzzy) verification methods

give credit to "close" forecasts

Scale separation methods

measure scale-dependent error

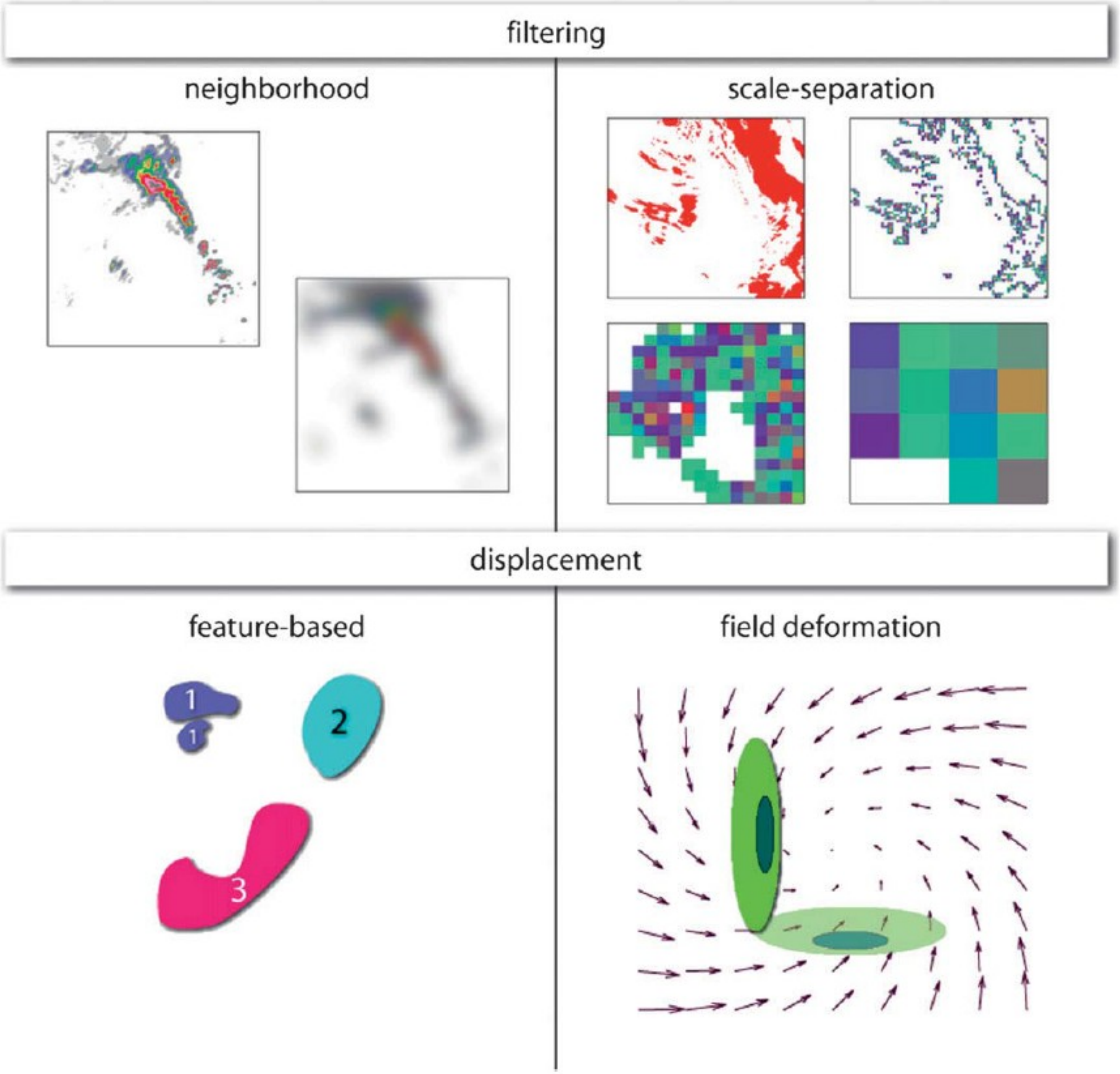
Features-based methods

evaluate attributes of identifiable features

Field deformation

evaluate phase errors

Spatial verification types



Gilleland, et al. 2009

FIG. 1. Schematic representations of the four categories of verification methods reviewed in this paper. (top) The neighborhood and scale-separation methods can both be considered "filtering" approaches while (bottom) the feature-based and field deformation methods fall under the "displacement" category.

TABLE 1. List of individual methods considered in this paper, and the ICP, along with their abbreviations used here. References listed are not comprehensive; see the text and the references for further representative works.

Abbreviation	Description	Method type	Reference(s)
BCETS	Bias-corrected ETS	Traditional	Mesinger (2008)
CA	Cluster analysis	Features based*	Marzban and Sandgathe (2006, 2008)
Composite	Composite method	Features based*	Nachamkin (2005, 2009)
CRA	Contiguous rain area	Features based	Ebert and McBride (2000); Ebert and Gallus (2009)
DIST	Distributional method	Neighborhood	Marsigli et al. (2006)
FQI	Forecast quality index	Field deformation*	Venugopal et al. (2005)
FQM–DAS	Forecast quality measure–displacement amplitude score	Field deformation	Keil and Craig (2007, 2009)
FSS	Fractions skill score	Neighborhood	Roberts (2005); Roberts and Lean (2008); Mittermaier and Roberts (2009)
IS	Intensity scale	Scale separation	Casati et al. (2004); Casati (2009)
IW	Image warping	Field deformation	E. Gilleland, J. Lindström, and F. Lindgren (2009, unpublished manuscript); Lindström et al. (2009)
MODE	Method for Object-based Diagnostic Evaluation	Features based	Davis et al. (2006, 2009)
MSV	Multiscale variability	Scale separation	Zapeda-Arce et al. (2000); Harris et al. (2001); Mittermaier (2006)
Neighborhood	Neighborhood based methods	Neighborhood	Ebert (2008, 2009)
Procrustes	Cell identification and Procrustes shape analysis	Features based	Micheas et al. (2007)
Procrustes2	Multiscale cell identification and Procrustes shape analysis	Scale separation–Features based	Lack et al. (2009)
SAL	Structure, amplitude, and location	Features based	Wernli et al. (2008, 2009)
Traditional	Point-based comparison	Point	Jolliffe and Stephenson (2003)
VGM	Variogram	Scale separation*	Marzban and Sandgathe (2009)

* A method that only loosely belongs to the given method type.

Gilleland, et al. 2009

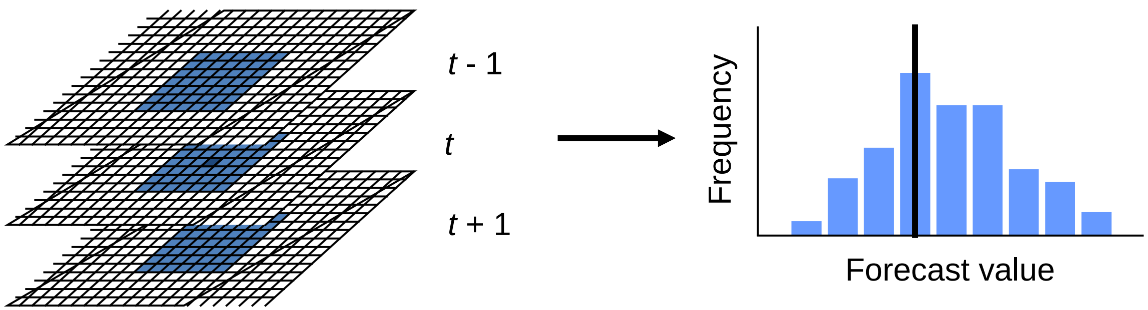
Neighborhood (fuzzy) verification methods
→ give credit to "close" forecasts

“close“ **“not close“**

Neighborhood verification methods

- Don't require an exact match between forecasts and observations
  - Unpredictable scales
  - Uncertainty in observations

- Look in a space / time neighborhood around the point of interest



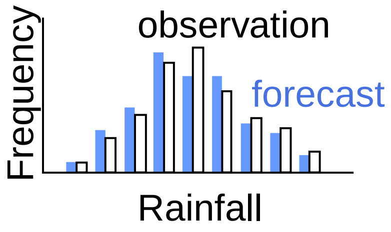
  - Evaluate using categorical, continuous, probabilistic scores / methods

Neighborhood verification methods

Treatment of forecast data within a window:

- Mean value (upscaling)
- Occurrence of event* somewhere in window
- Frequency of events in window → probability
- Distribution of values within window

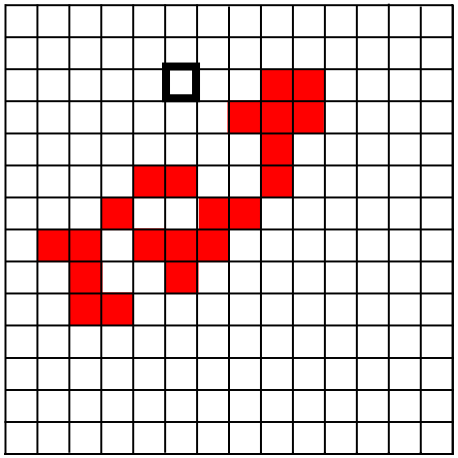
May also look in a neighborhood of observations



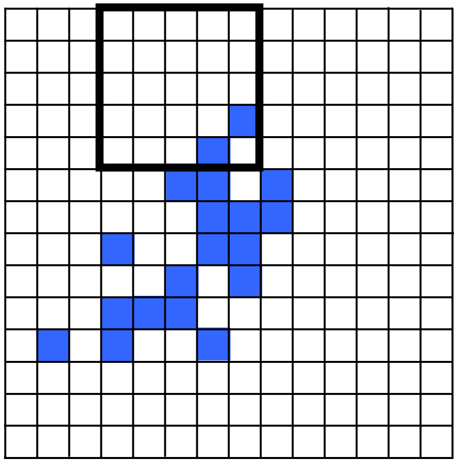
* *Event* defined as a value exceeding a given threshold, for example, rain exceeding 1 mm/hr

Moving windows

For each combination of neighborhood size and intensity threshold, accumulate scores as windows are moved through the domain



observation



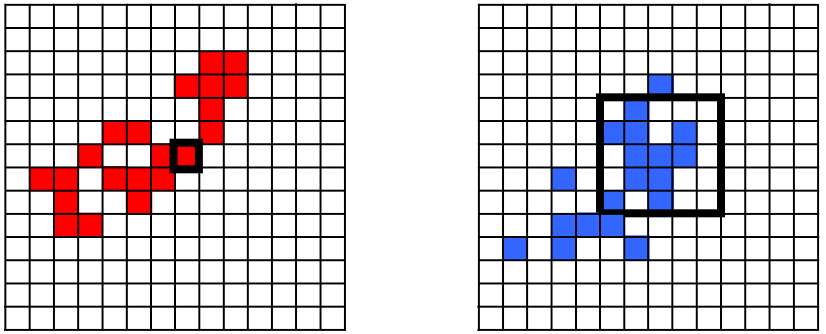
forecast

Neighborhood verification framework

Neighborhood methods use one of two approaches to compare forecasts and observations:

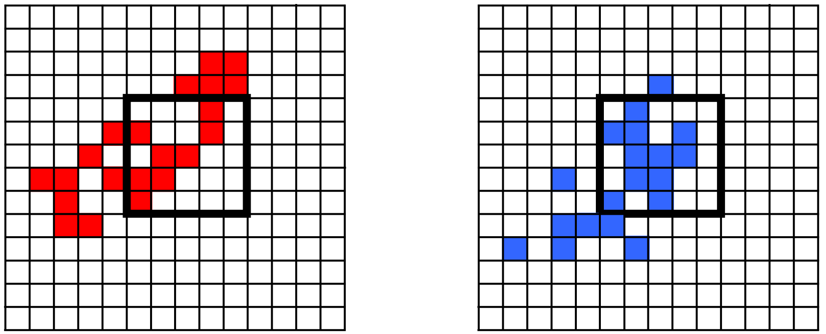
single observation – neighborhood forecast

(SO-NF, user-oriented)



neighborhood observation – neighborhood forecast

(NO-NF, model-oriented)



Different neighborhood verification methods have different decision models for what makes a *useful forecast*

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

*NO-NF = neighborhood observation-neighborhood forecast,
SO-NF = single observation-neighborhood forecast

from Ebert, *Meteorol. Appl.*, 2008

Detailed description of Fraction Skill Score (FSS) (Roberts and Lean, 2008)

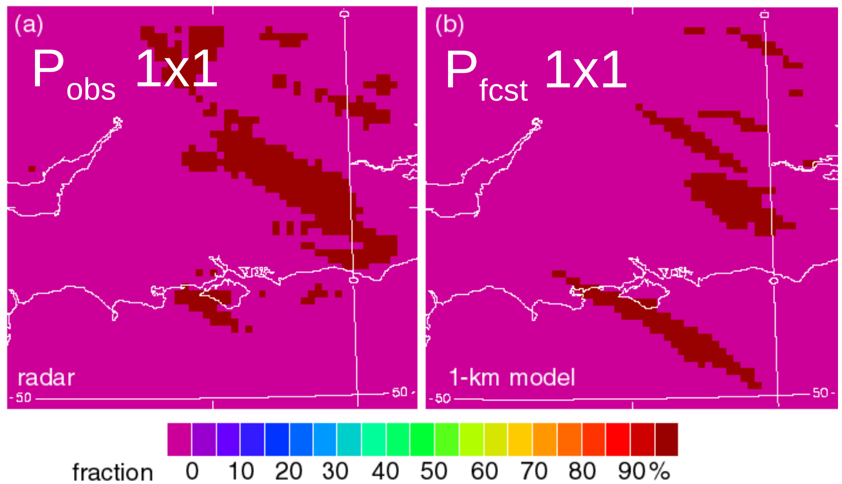
We want to know

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

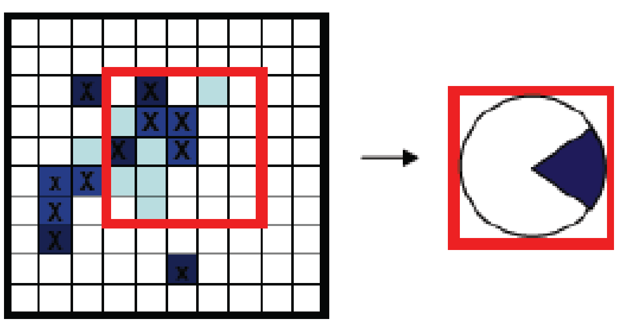
Step 1: FC and Observation/Analysis have to be on the same grid.
Step 2: Choose suitable thresholds q (e.g.: 0.5, 1, 2, 4 mm)
Step 3: Convert FC/AN fields to binary fields I_O and I_M according to threshold

$$I_o = \begin{cases} 1 & O_r \geq q \\ 0 & O_r < q \end{cases} \quad \text{and} \quad I_M = \begin{cases} 1 & M_r \geq q \\ 0 & M_r < q \end{cases}$$

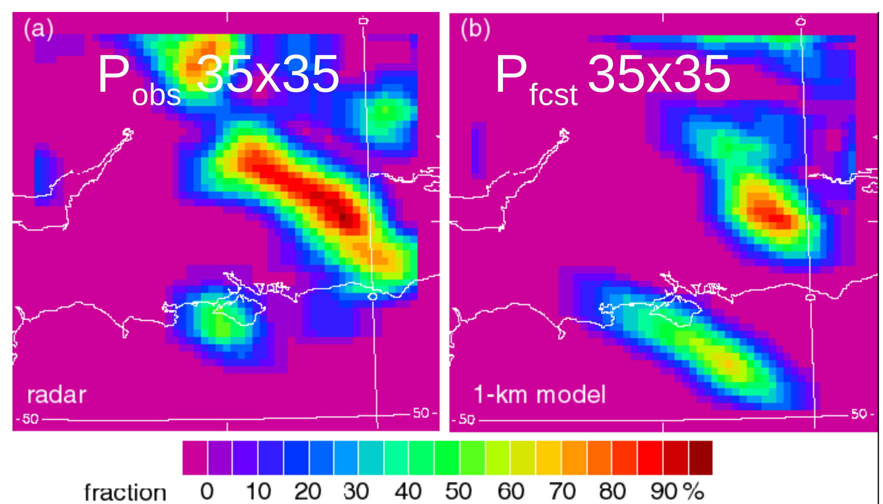
Detailed description of Fraction Skill Score (FSS) (Roberts and Lean, 2008)



Step 4: Generate fractions for all thresholds:



P_{obs} = fraction of obs grid points > threshold
P_{fcst} = fraction of fcst grid points > threshold



Detailed description of Fraction Skill Score (FSS) (Roberts and Lean, 2008)

Step 5: Compute fraction skill score for all thresholds:

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

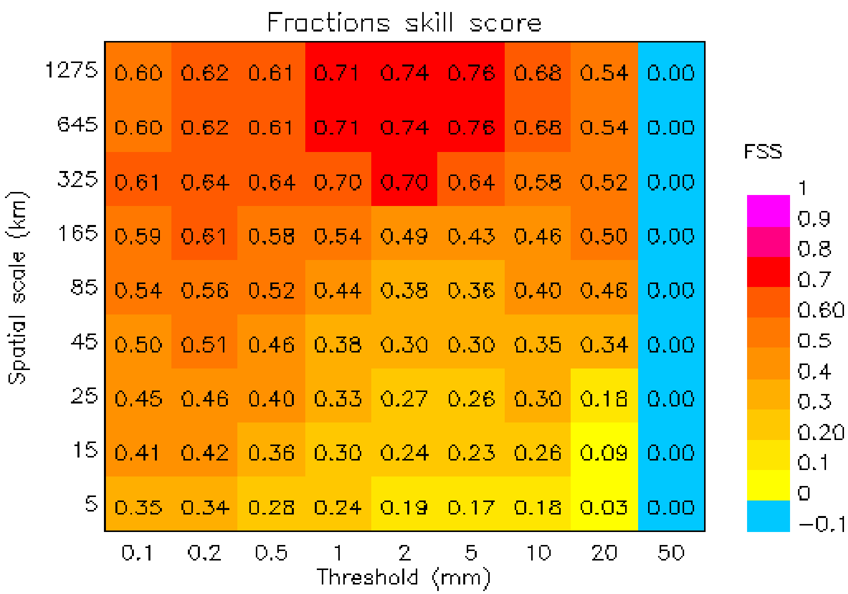
$$\mathrm{FSS}_{(n)} = \frac{\mathrm{MSE}_{(n)} - \mathrm{MSE}_{(n)\mathrm{ref}}}{\mathrm{MSE}_{(n)\mathrm{perfect}} - \mathrm{MSE}_{(n)\mathrm{ref}}} = 1 - \frac{\mathrm{MSE}_{(n)}}{\mathrm{MSE}_{(n)\mathrm{ref}}}$$

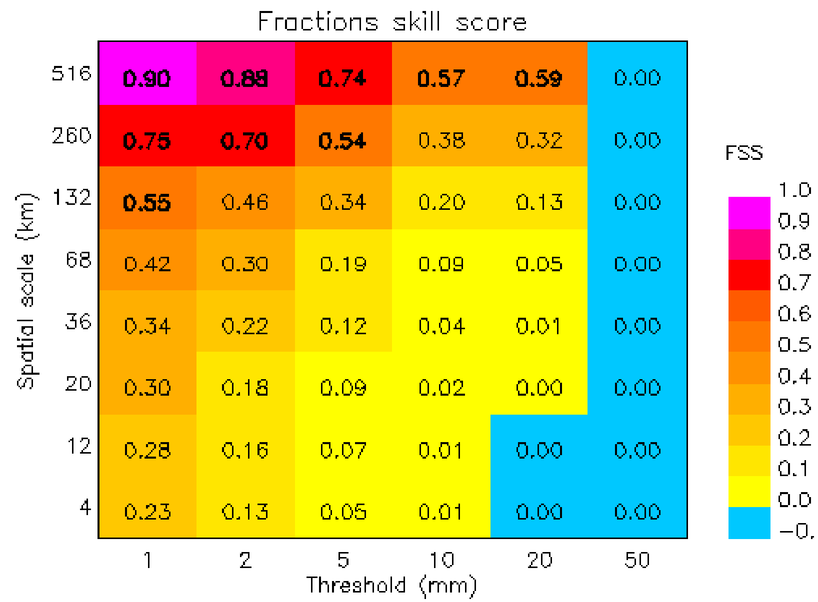
Maximum estimation (low-skill reference) of MSE:

$$(P_{fcst}-P_{obs})^2 = P_{fcst}^{\ 2} - 2P_{fcst}P_{obs} + P_{obs}^{\ 2} \sim P_{fcst}^{\ 2} + P_{obs}^{\ 2} = MSE_{ref}$$

Detailed description of Fraction Skill Score (FSS) (Roberts and Lean, 2008)

Step 6: Graphical presentation for each threshold and spatial scale:



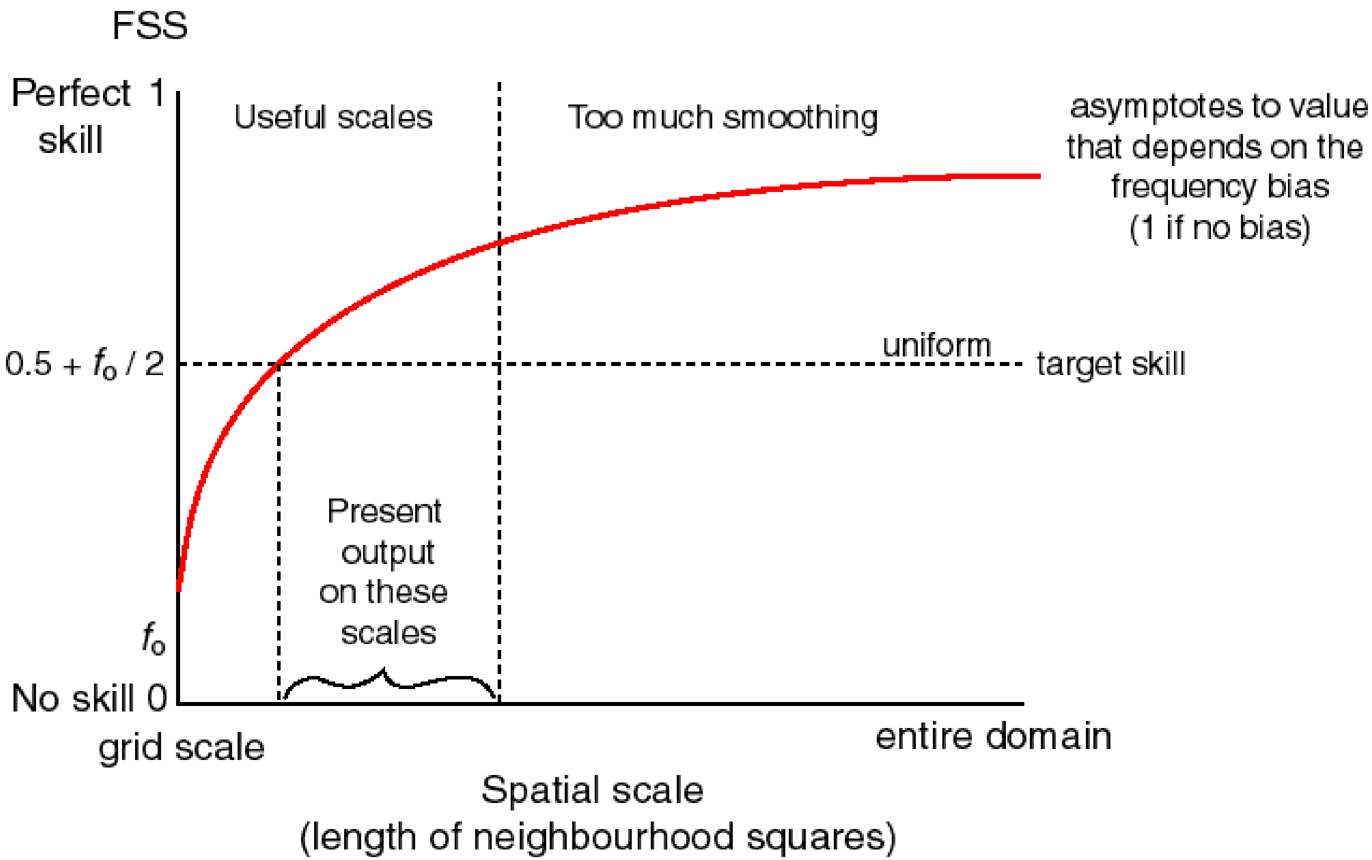


Interpretation:

- Skill increases with spatial scale
- The smaller the displacement error the faster the skill increases with increasing spatial scale
- When the length of the moving window is smaller or equal the displacment error there is no skill and FSS=0

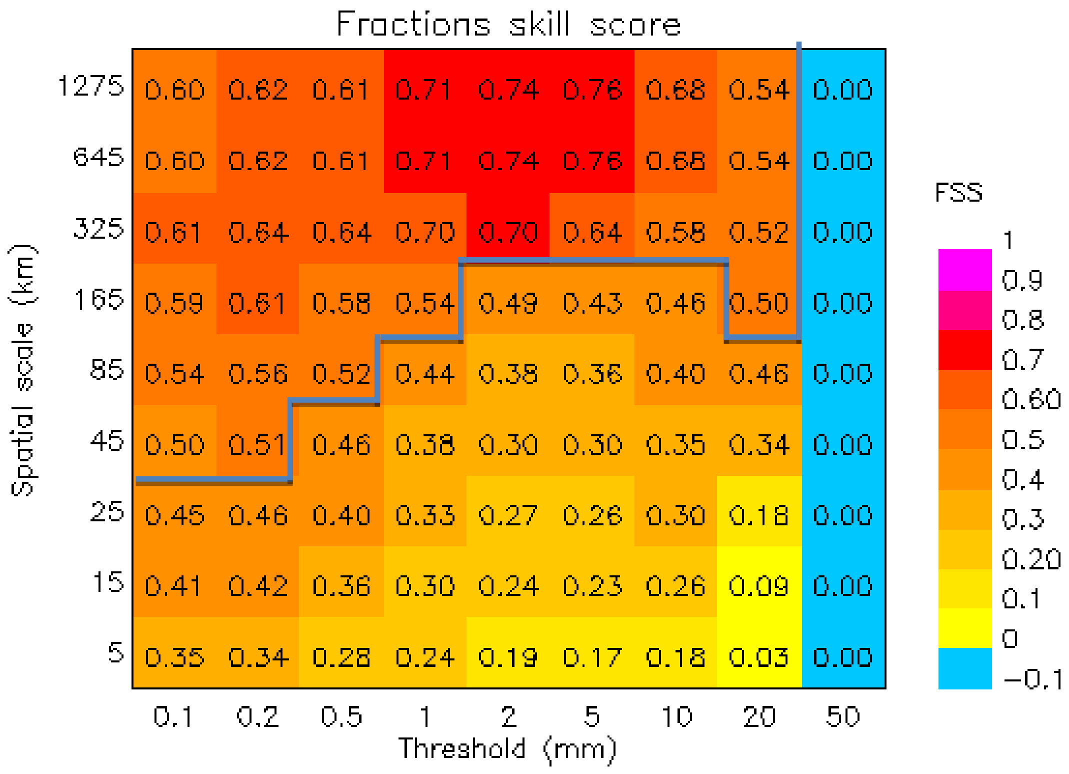
Detailed description of Fraction Skill Score (FSS) (Roberts and Lean, 2008)

Q: What happens if size of moving window is equal to domain size?
Q: What are useful (skillfull) numbers of FSS?



f_o=domain obs fraction on the grid scale (for f_0=0.2(20%) → target skill: FSS=0.5+0,2/2=0.6

Detailed description of Fraction Skill Score (FSS) (Roberts and Lean, 2008)



Scale separation methods
→scale-dependent error

1. Which spatial scales are well represented and which scales have error?
2. How does the skill depend on the precipitation intensity?

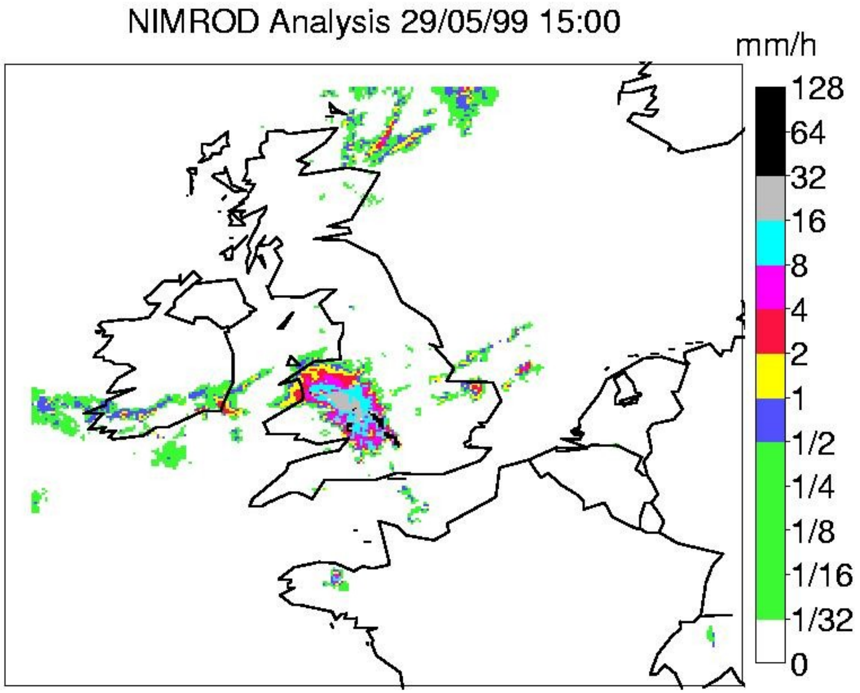
NOTE: scale = single band spatial filter → features of different scales → feedback on different physical processes and model parameterizations

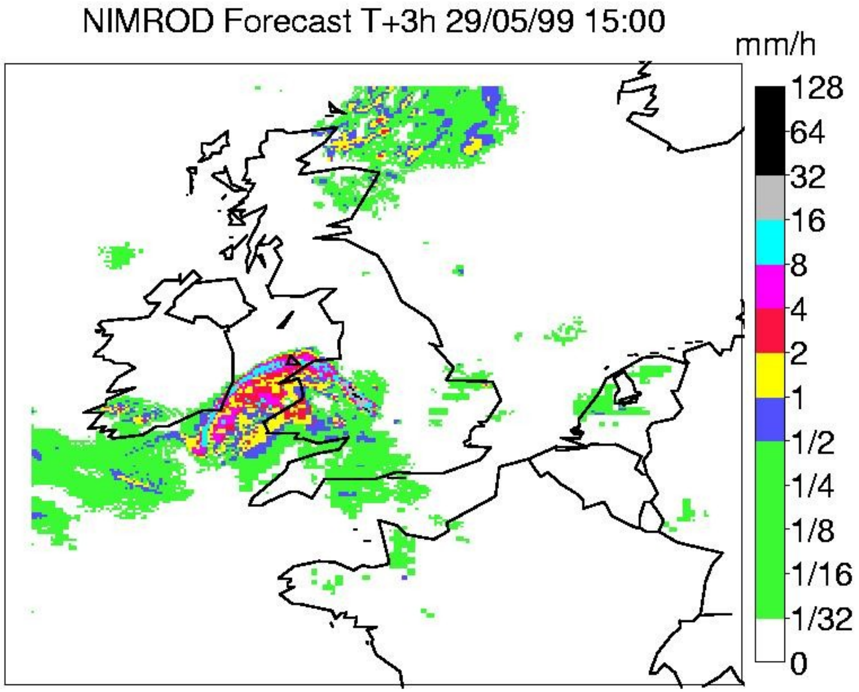
In the neighborhood based (fuzzy) verification, the **scale** is the neighborhood size (low band pass filter): as the scale increases the exact positioning requirements are more and more relaxed

What is the difference between neighborhood and scale separation approaches?

- Neighborhood verification methods
  - Get scale information by ***filtering out higher resolution scales***
- Scale separation methods
  - Get scale information by ***isolating scales of interest***

Nimrod case study: intense storm displaced



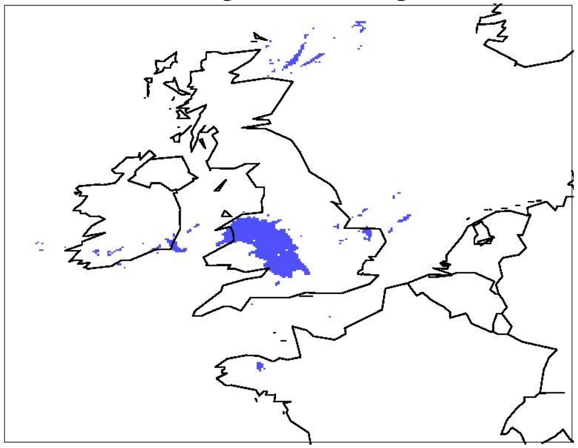


Step 1: Gridded data, square domain with dimension 2^n

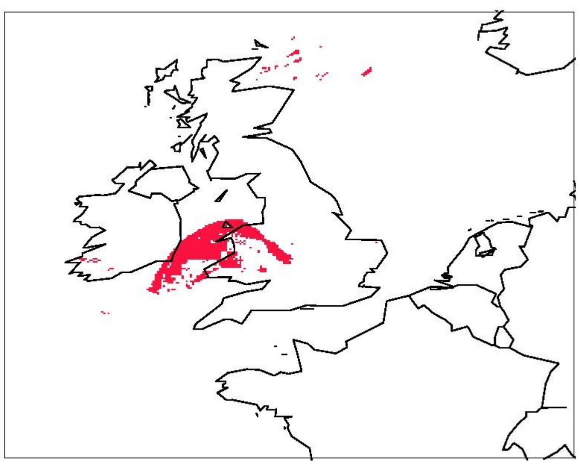
It can be applied to any meteorological field … however, it was specifically designed for spatial precipitation forecasts …

Step 2: Intensity: threshold to obtain binary images (categorical approach)

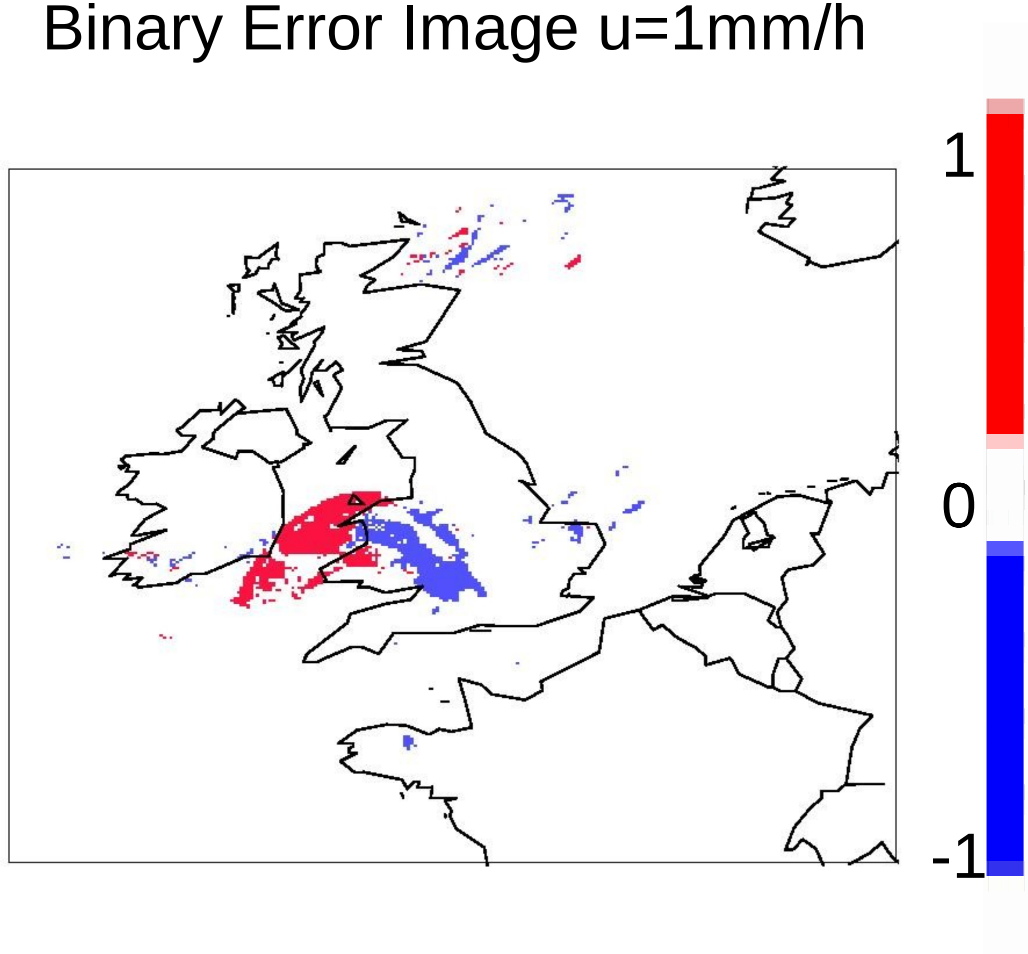
Binary Analysis



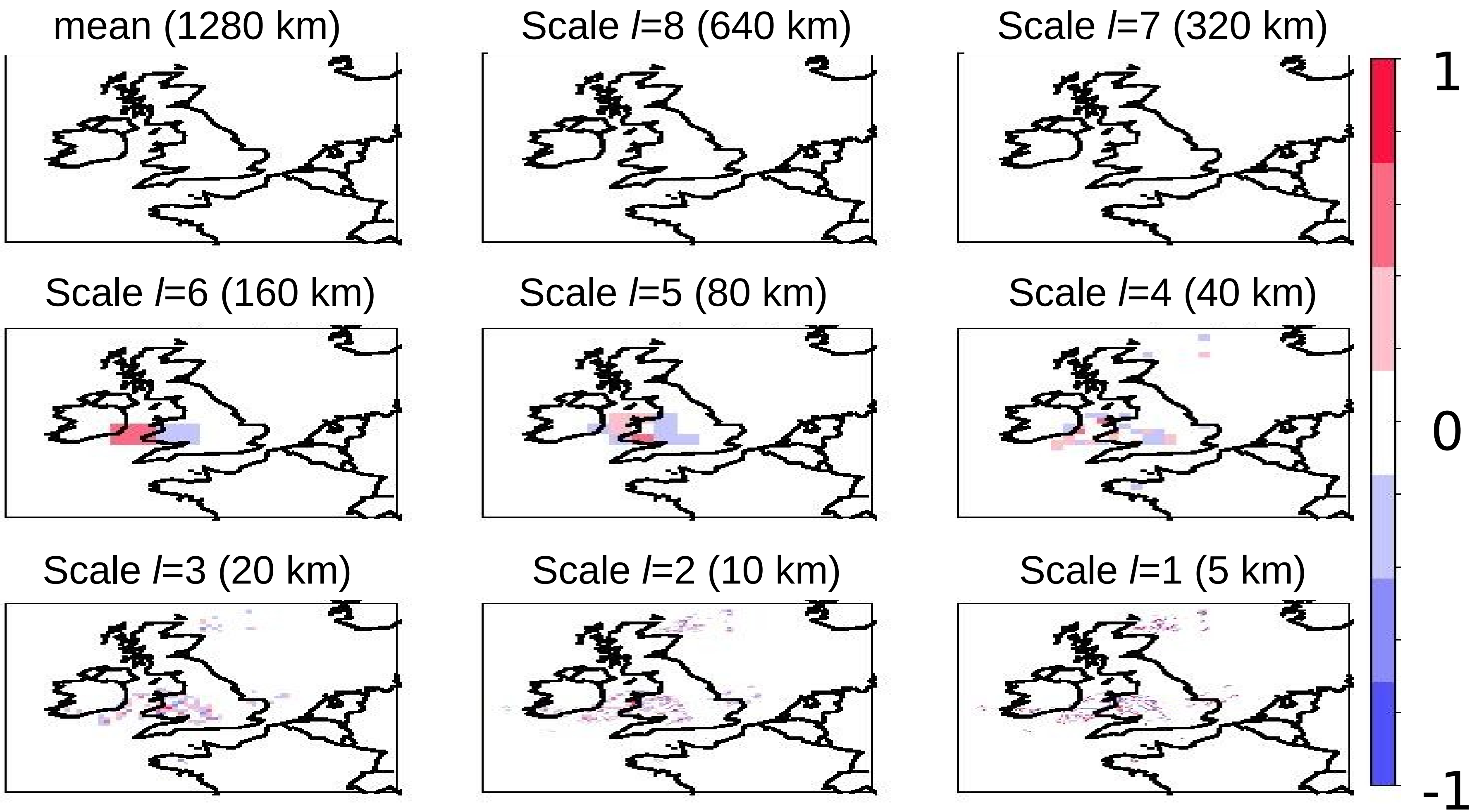
Binary Forecast



Binary Error Image u=1mm/h



Step 3: Scale → wavelet decomposition of binary error

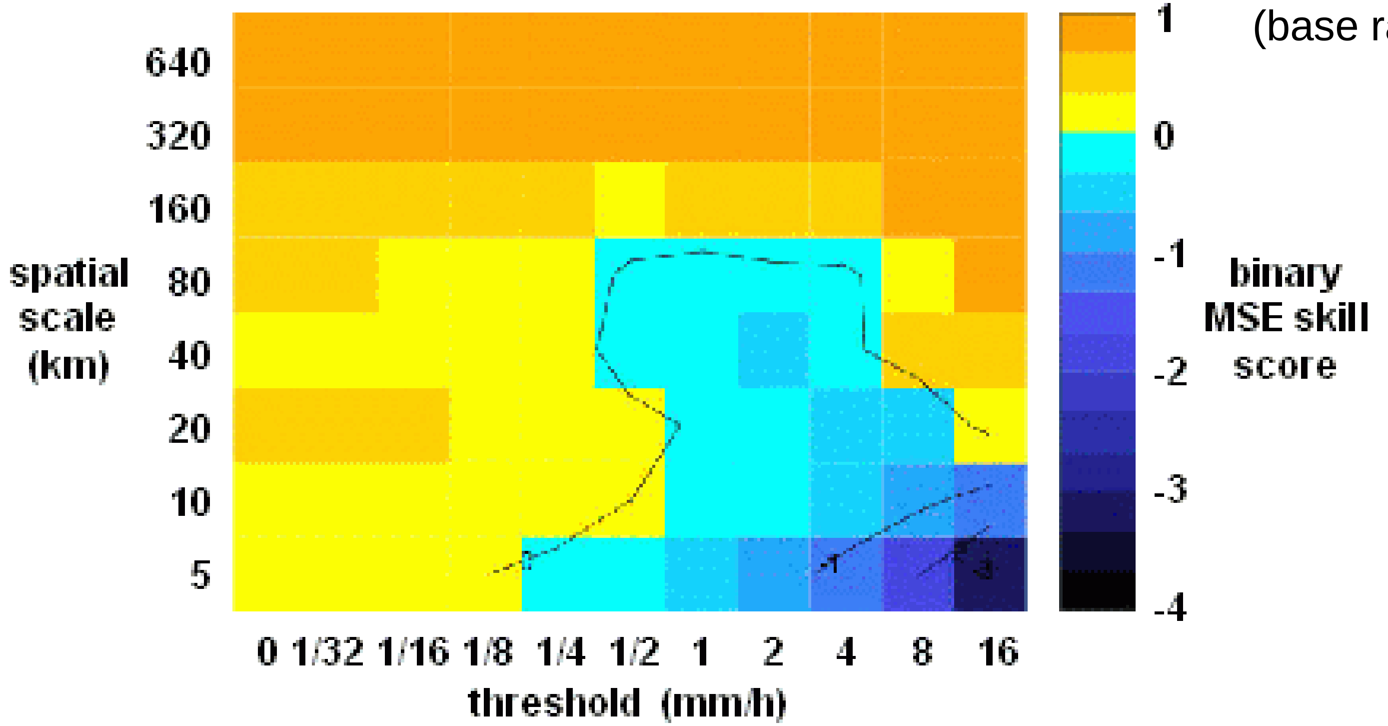


$$E_u = \sum_{l=1}^{L} E_{u,l} \qquad MSE_u = \sum_{l=1}^{L} MSE_{u,l}$$

Step 4: MSE skill score for each threshold and scale component

$$SS_{u,l} = \frac{MSE_{u,l} - MSE_{u,l,random}}{MSE_{u,l,best} - MSE_{u,l,random}} = 1 - \frac{MSE_{u,l}}{2\varepsilon(1-\varepsilon)/L}$$

Sample climatology (base rate)



Strenghts

Categorical approach → robust and resistant

Wavelets → cope with spatially discontinuous fields characterized by the presence of few sparse non-zero features

➔ suitable for spatial precipitation forecasts

Weaknesses

need gridded data on a square domain with dimension 2^n

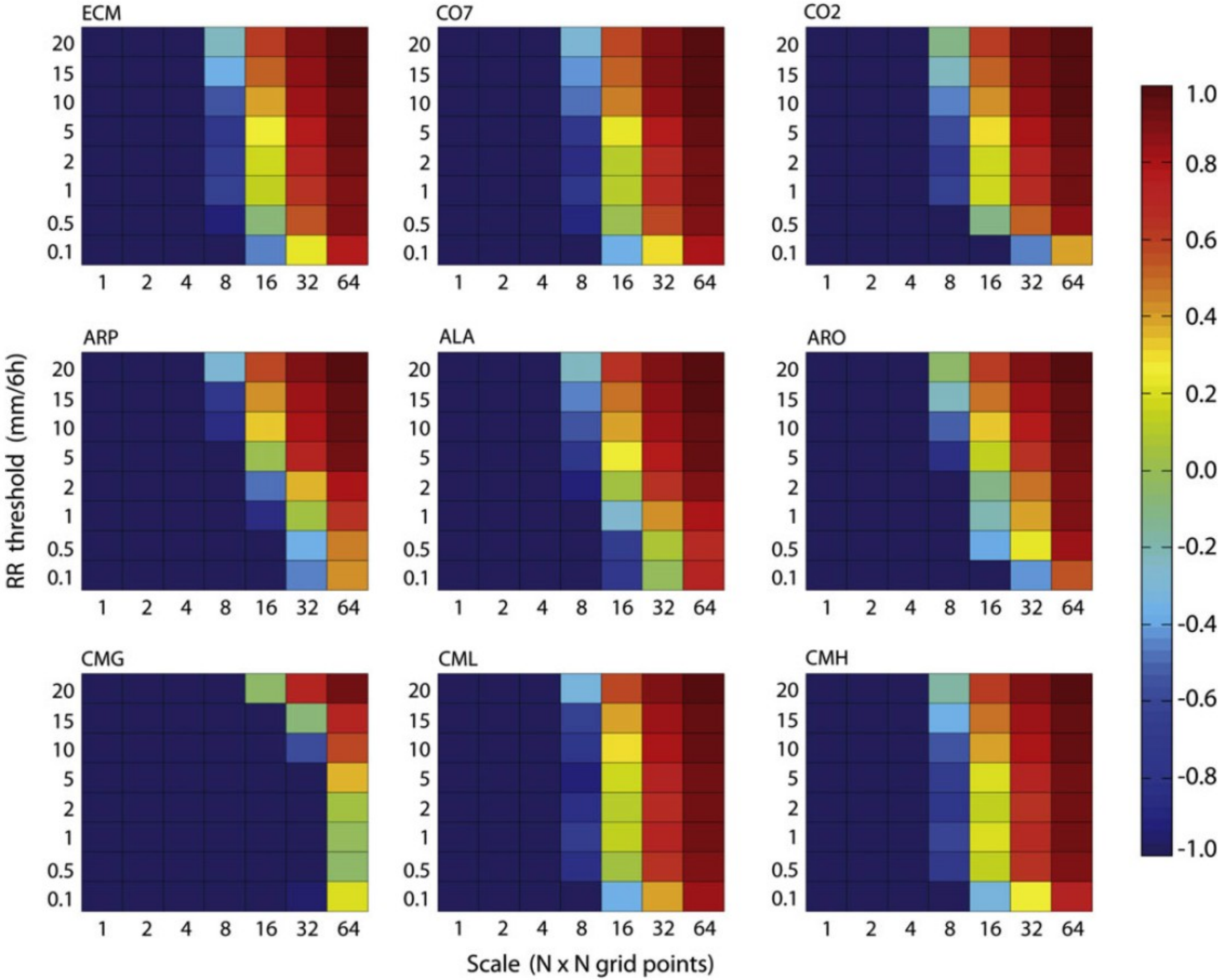


Figure 3: Intensity-scale skill score (ISS) for 6 h accumulated precipitation forecasts and up to 18 h lead time. Forecast ranges: 0-6 h, 6-12 h and 12-18 h; period: Jun-Nov 2007; x-axes: scale (1≙8 km, 2≙16 km, 4≙32 km, 8≙64 km, 16≙128 km, 32≙256 km, 64≙512 km); y-axes: precipitation thresholds in mm/6 h.

Dorninger and Gorgas, 2012

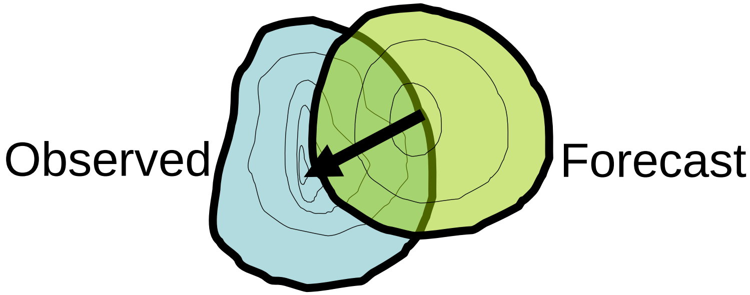
Features-based methods

→evaluate attributes of features

Feature-based approach (CRA)

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

- Define entities using (user defined) 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) before shifting

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

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

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

The *volume error* is the bias in mean intensity

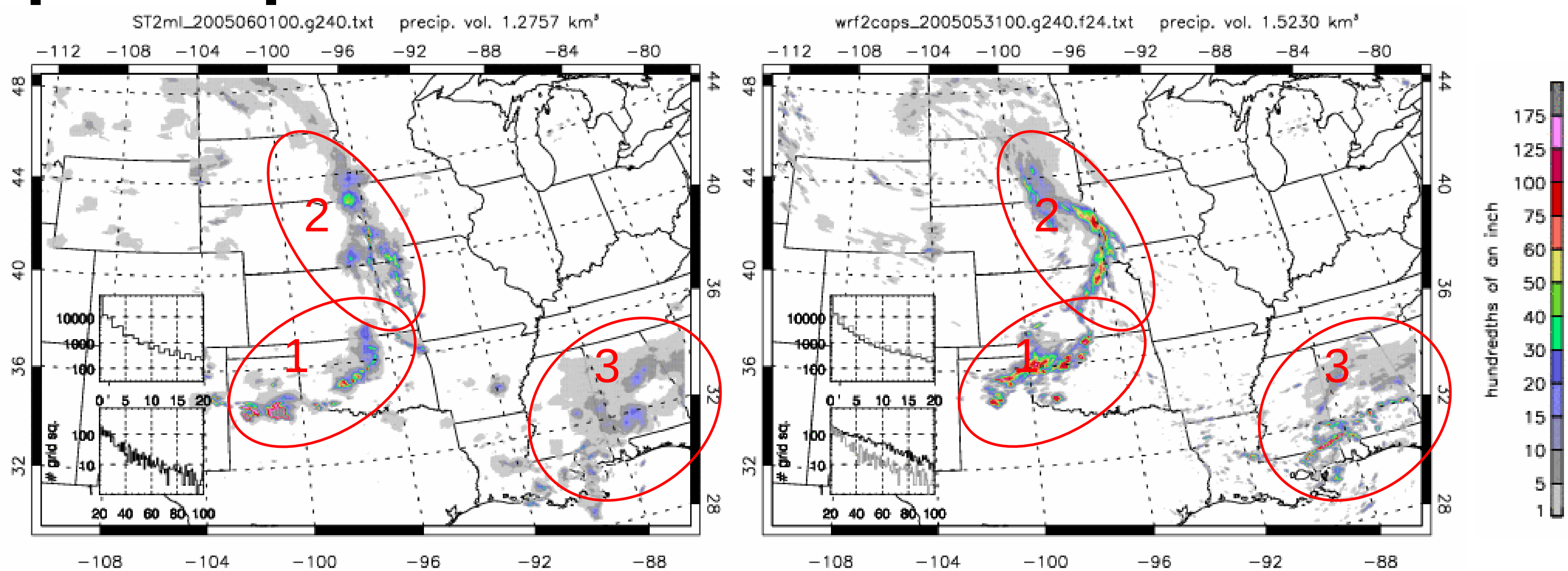
$$MSE_{volume} = (\overline{F} - \overline{X})^2$$

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,

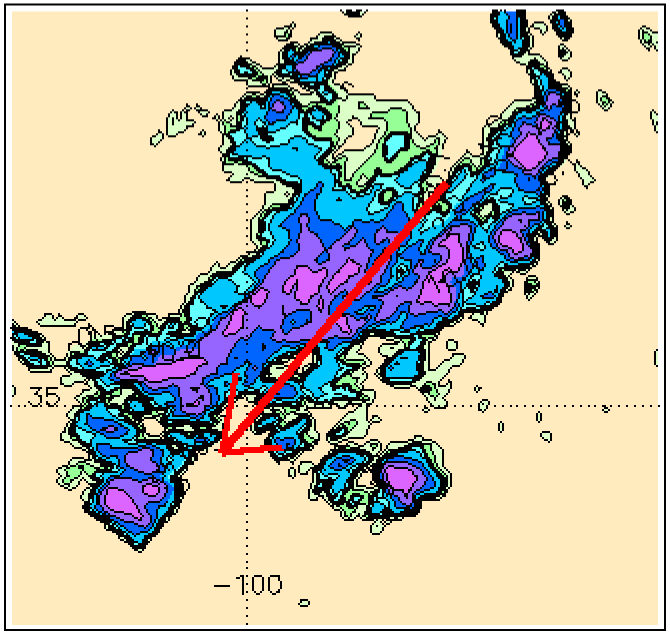
$$MSE_{pattern} = MSE_{shifted} - MSE_{volume}$$

Example: CRA verification of precipitation forecast over USA

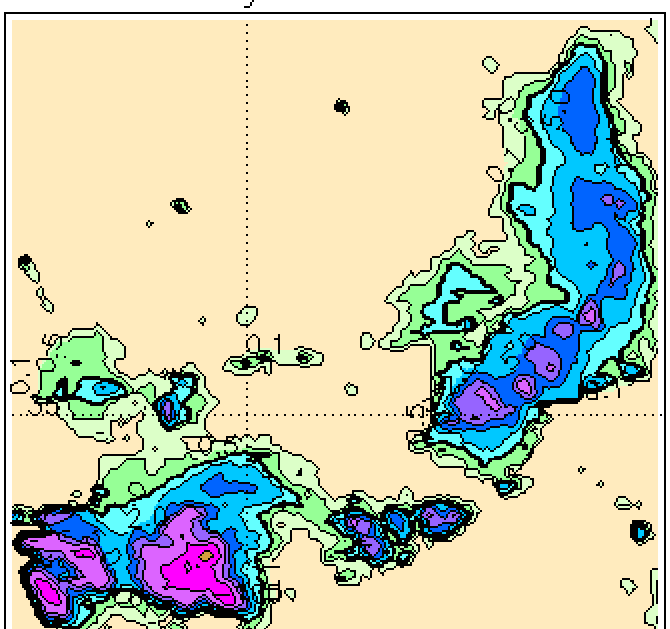


1. What is the location error of the forecast?
2. How do the forecast and observed rain areas compare? Average values? Maximum values?
3. How do the displacement, volume, and pattern errors contribute to the total error?

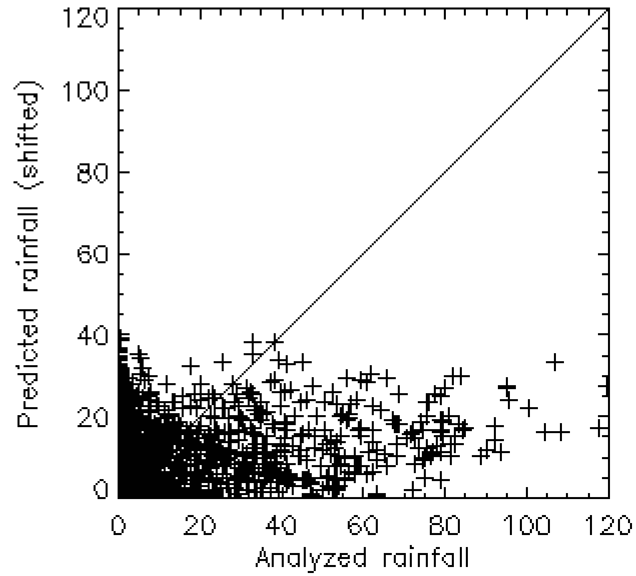
wrf2 fcst 20050601 hour 00–24



Analysis 20050601



CRA 20050601



wrf2 24h fcst 20050601 n=8423
(33.49°,–102.28°) to (37.77°,–96.00°)
Verif. grid=0.042° CRA threshold=1.0 mm/h

	Analysed	Forecast
# gridpoints ≧1 mm/h	3304	3597
Average rainrate (mm/h)	3.58	3.61
Maximum rain (mm/h)	119.63	39.12
Rain volume (km^3)	0.51	0.52

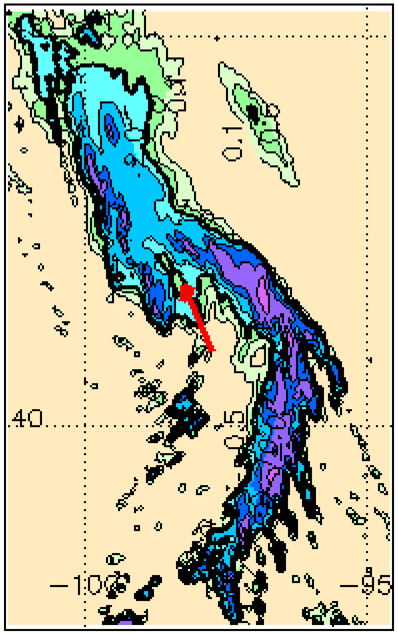
Displacement (E,N) = [2.20°,1.92°] max.corr matching

	Original	Shifted
RMS error (mm/d)	12.81	10.24
Correlation coefficient	–0.167	0.305

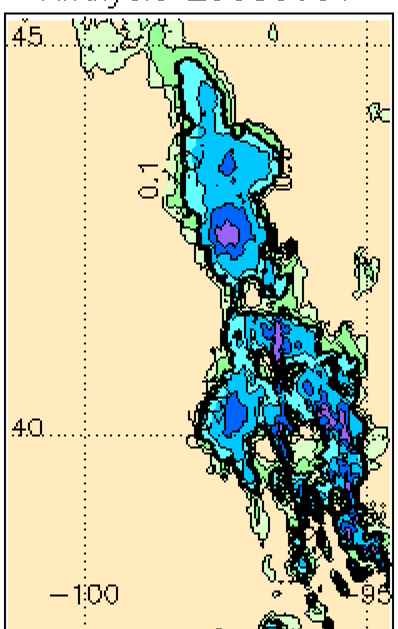
Error Decomposition:

Displacement error	36.1%
Volume error	0.0%
Pattern error	63.9%

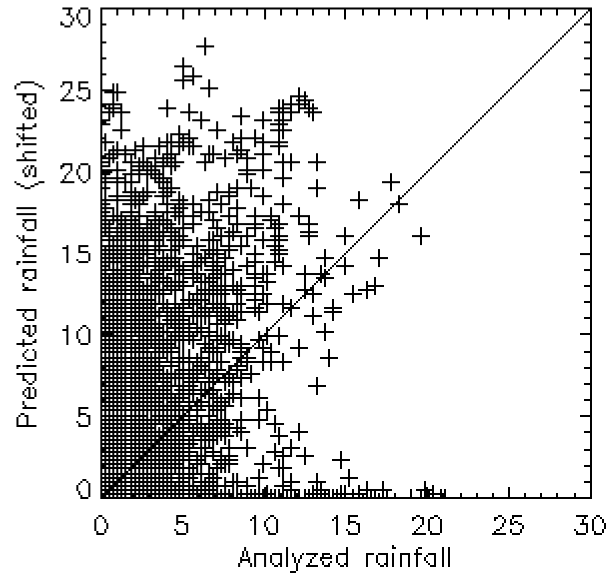
wrf2 fcst 20050601 hour 00–24



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

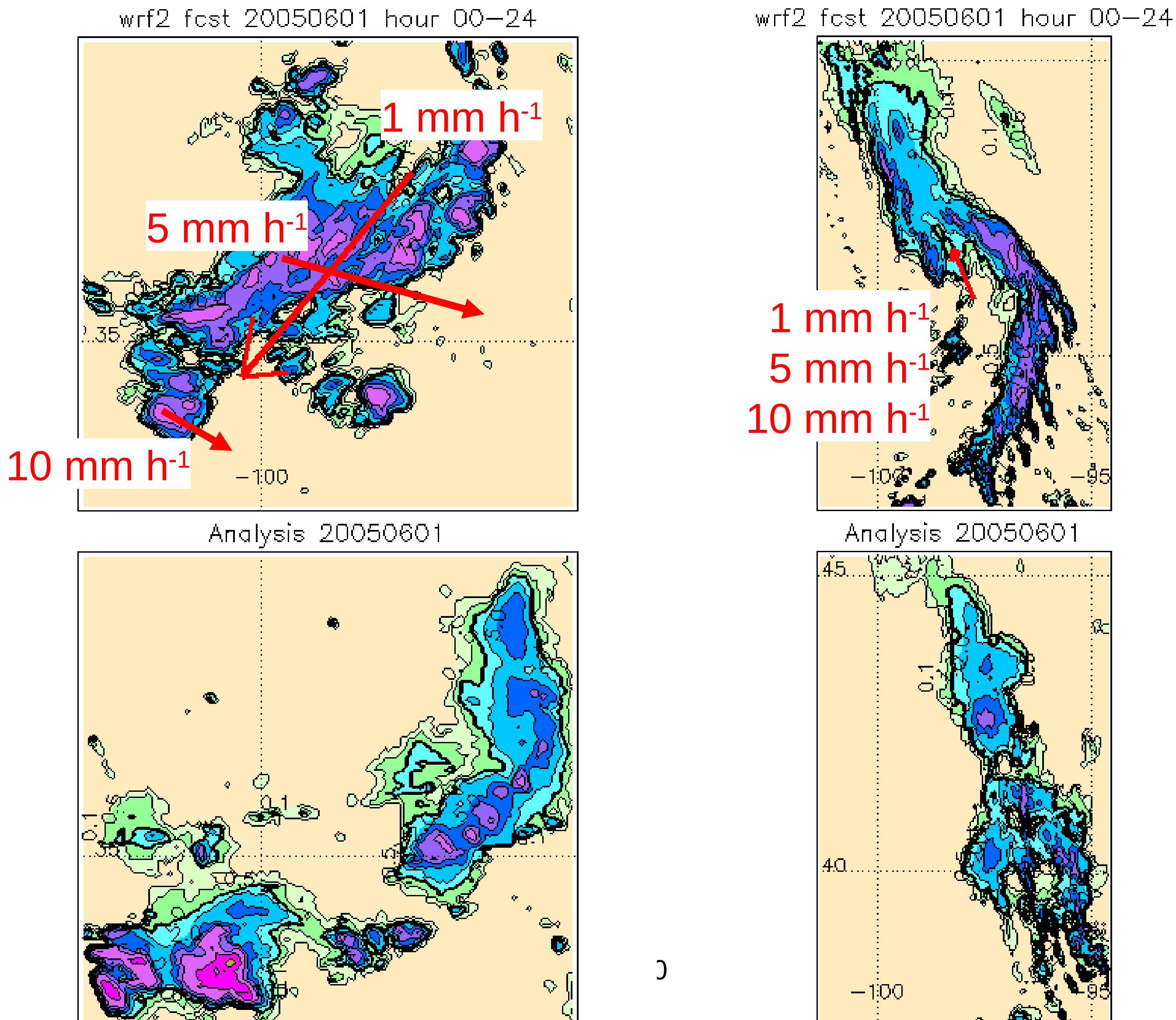
Displacement (E,N) = [0.52°,–0.84°] max.corr matching

	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



Strengths of CRA

The entity-based CRA verification method has a number of attractive features:

(a) It is intuitive, quantifying what we can see by eye
(b) it estimates the location error in the forecast,
(c) the total error can be decomposed into contributions from location, intensity, and pattern,
(d) forecast events can be categorized as hits, misses, etc.
These descriptions could prove a useful tool for monitoring forecast performance over time.

Weaknesses of CRA

There are also some drawbacks to this approach:

(a)Pattern matching: it must be possible to associate entities in the forecast with entities in the observations. This means that the forecast must be halfway decent. The verification results for a large number of CRAs will be biased toward the "decent" forecasts, i.e., those for which location and intensity errors could reliably be determined.
(b)The user must choose the pattern matching method as well as the isoline used to define the entities. The verification results will be somewhat dependent on these choices (subjective).
(c)When a forecast and/or observed entity extends across the boundary of the domain it is not possible to be sure whether the pattern match is optimal. If the CRA has a reasonably large area still within the domain then the probability of a good match is high. Ebert and McBride (2000) suggest applying a minimum area criterion to address this issue.

Structure-Amplitude-Location (SAL)

Wernli et al., *Mon. Wea. Rev.*, 2008

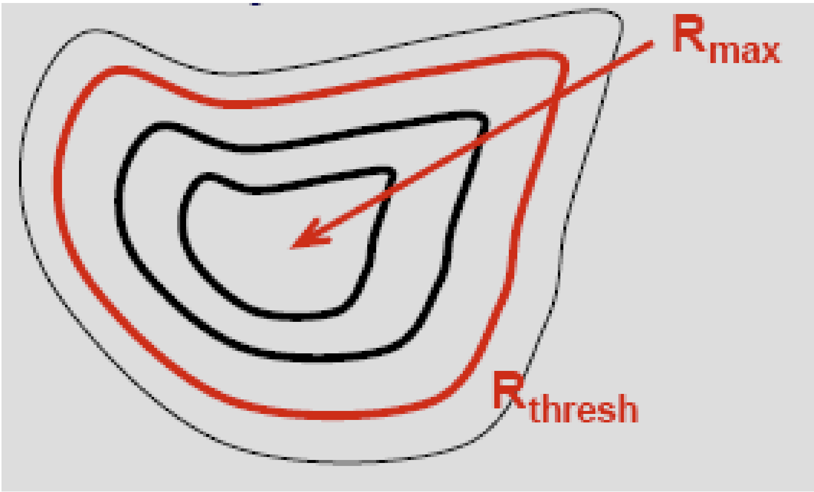
- Verification of rain forecasts in a defined domain
- No match of objects

SAL consists of three components:

- S structure
- A amplitude
- L location

Perfect forecast: S=A=L=0

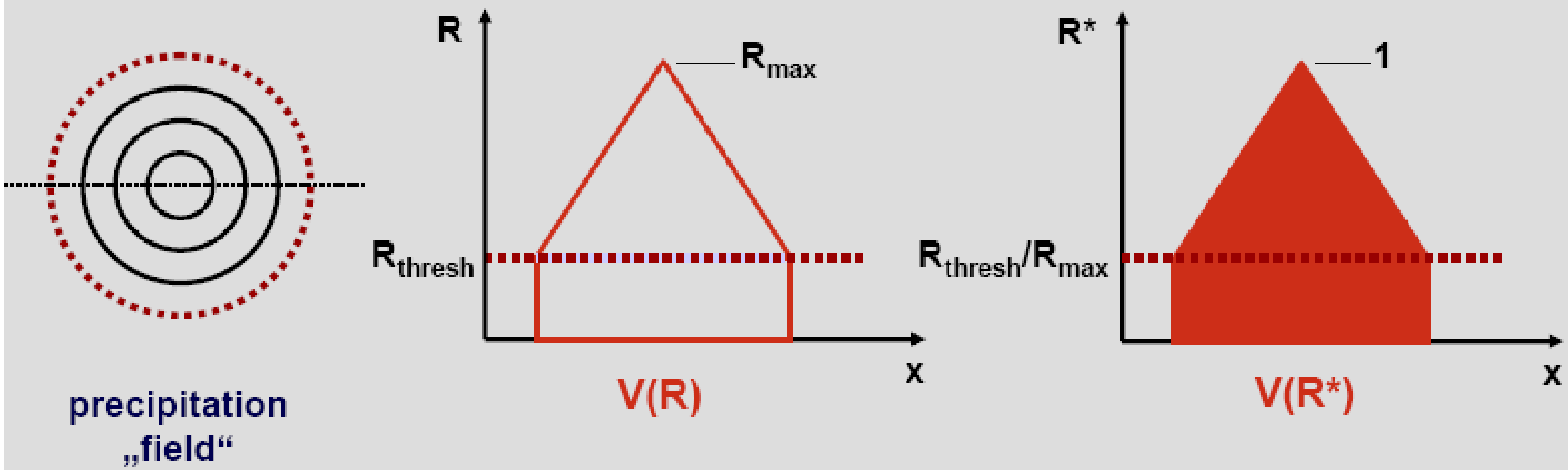
Step 1: Definition of precipitation objects



S A L – defintion of the S-component

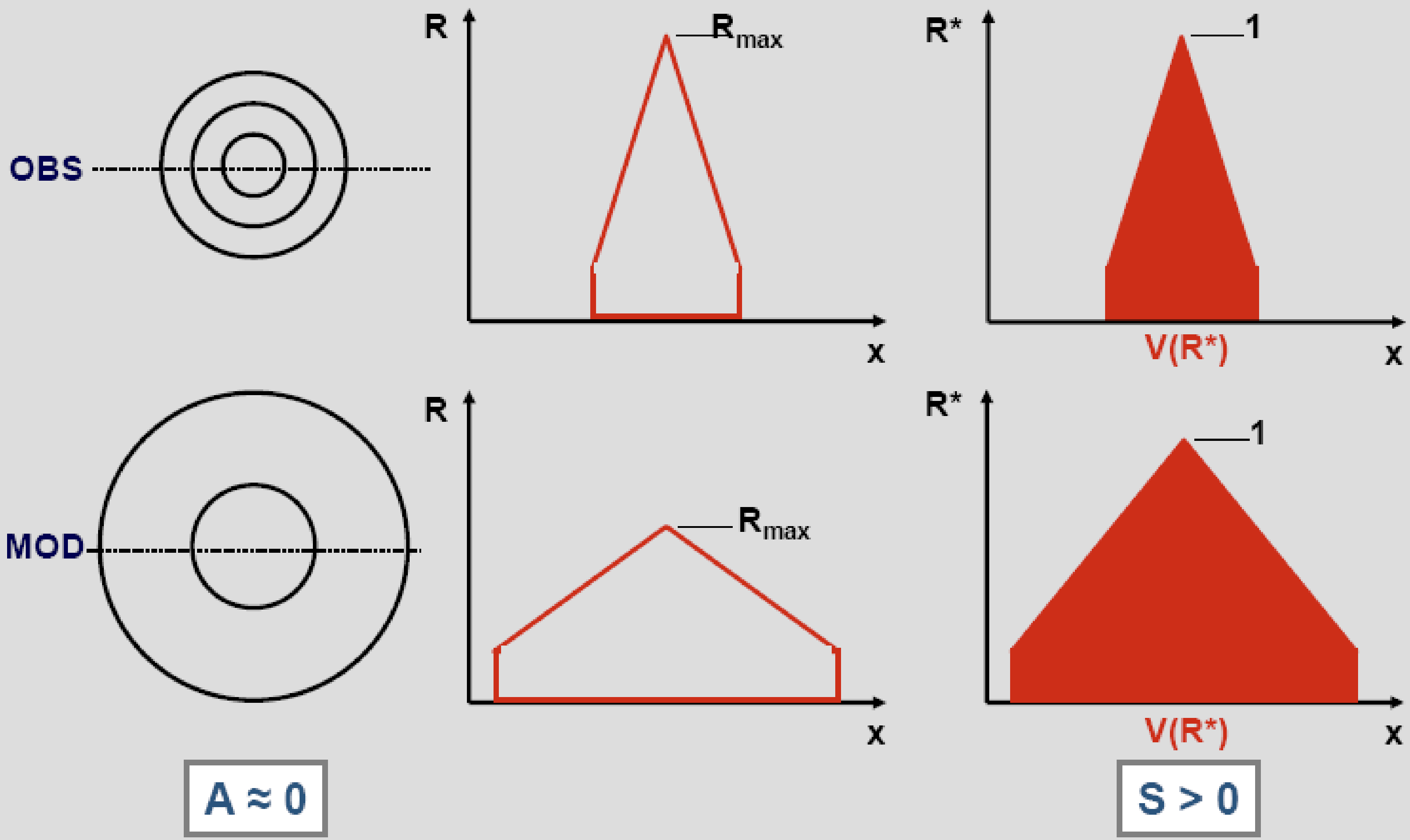
$S = V(R_{mod}^*) - V(R_{obs}^*)$

V(…) denotes the mean (weighted) volume of all scaled precipitation objects
structure error in the chosen area

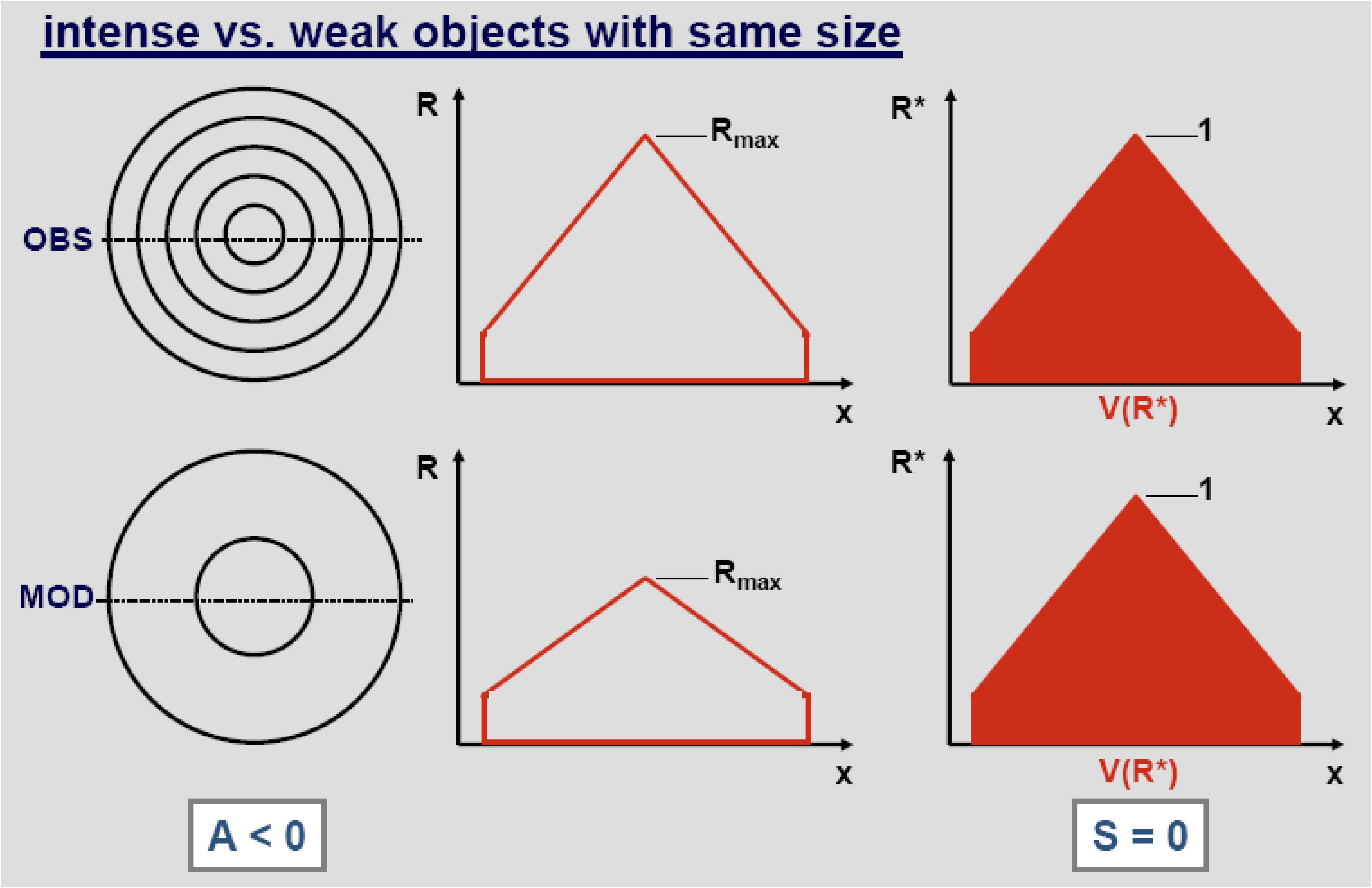


scaling for each object: $R^* = R / R_{max}$; $R^* \in [R_{thresh}/R_{max}, \ldots, 1]$

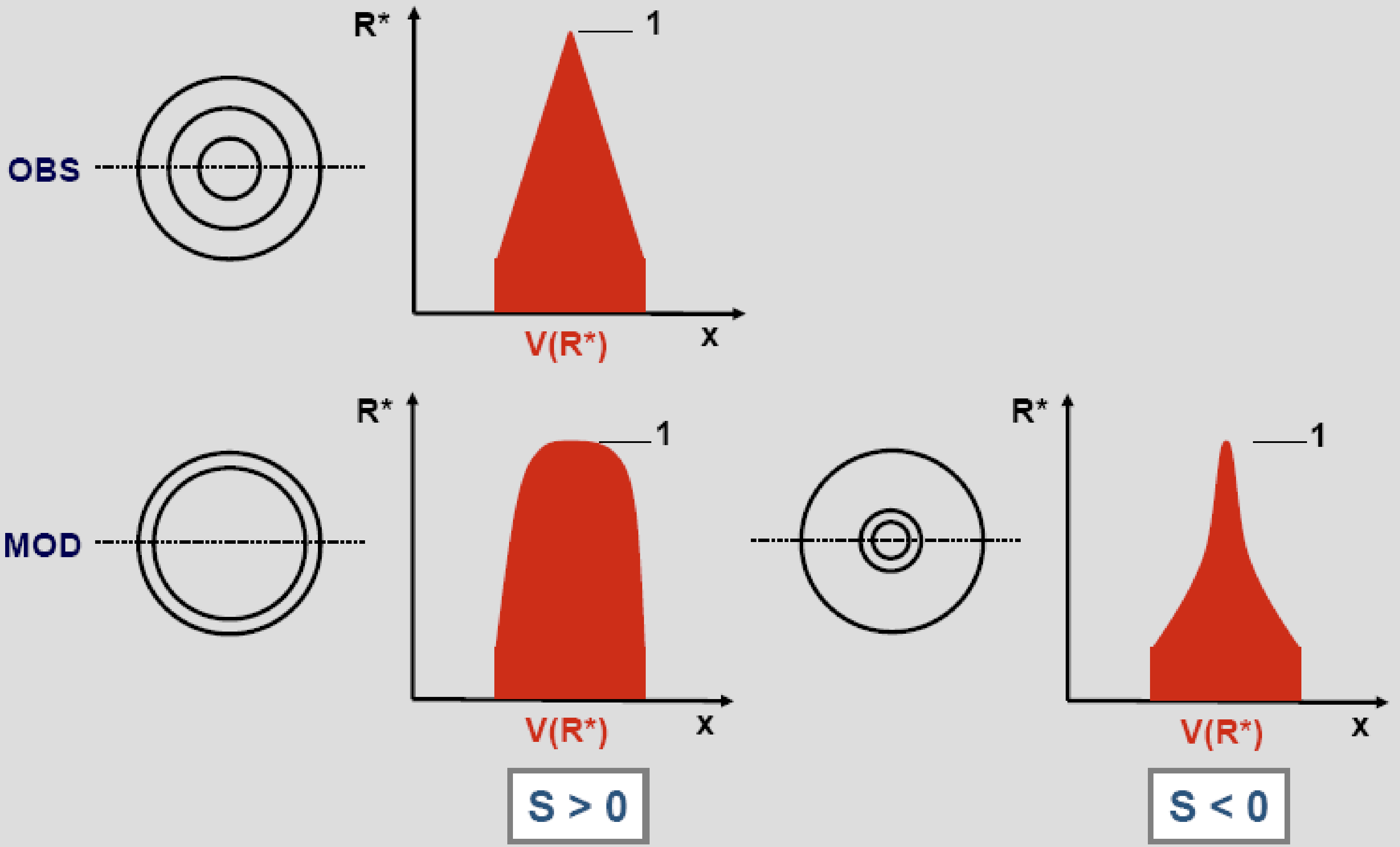
small intense vs. large weak objects



4.2.1 SAL


intense vs. weak objects with same size
OBS
MOD
R
R_{max}
x
R*
1
V(R*)
A < 0
S = 0

sensitivity to the object structure

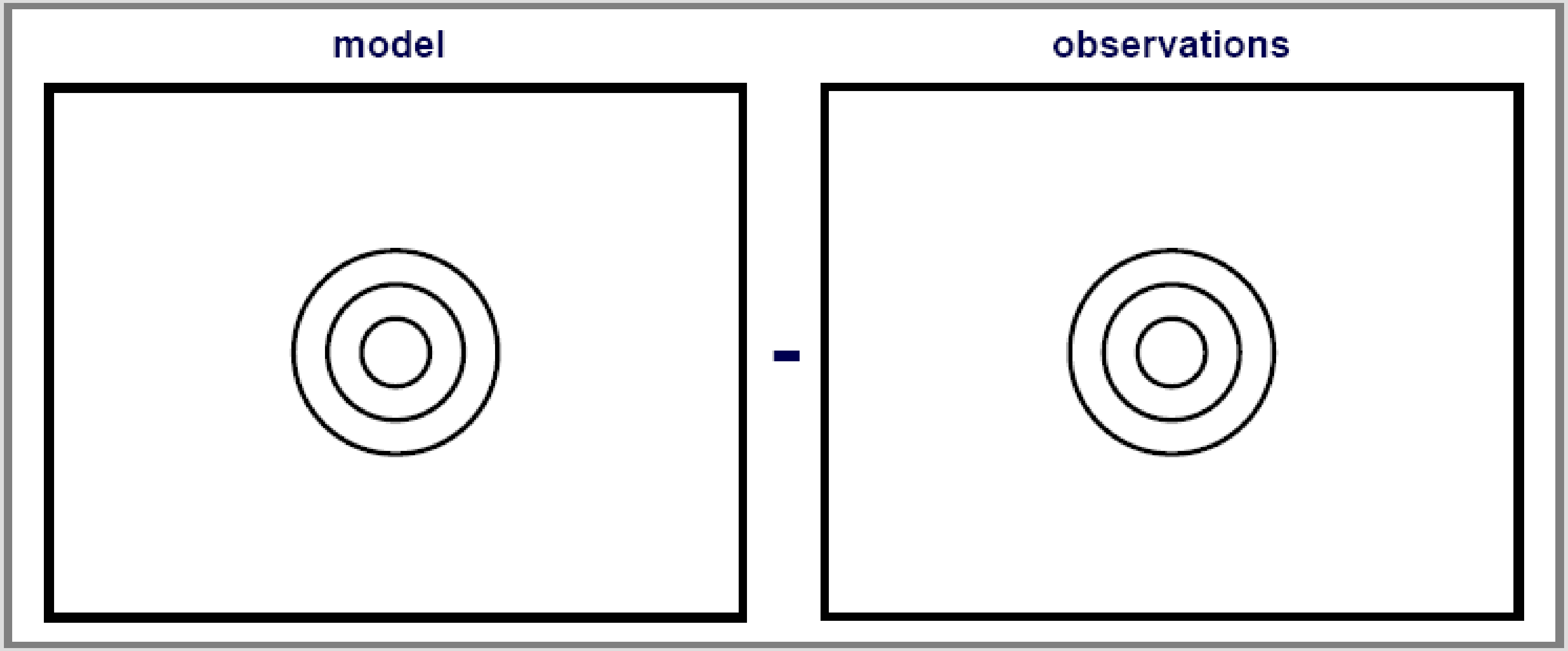


S A L – defintion of the A-component

$A = D(R_{mod}) - D(R_{obs}) = 0$

$$D(R) = \frac{1}{N} \sum_{(i,j)\in \mathcal{D}} R_{ij},$$

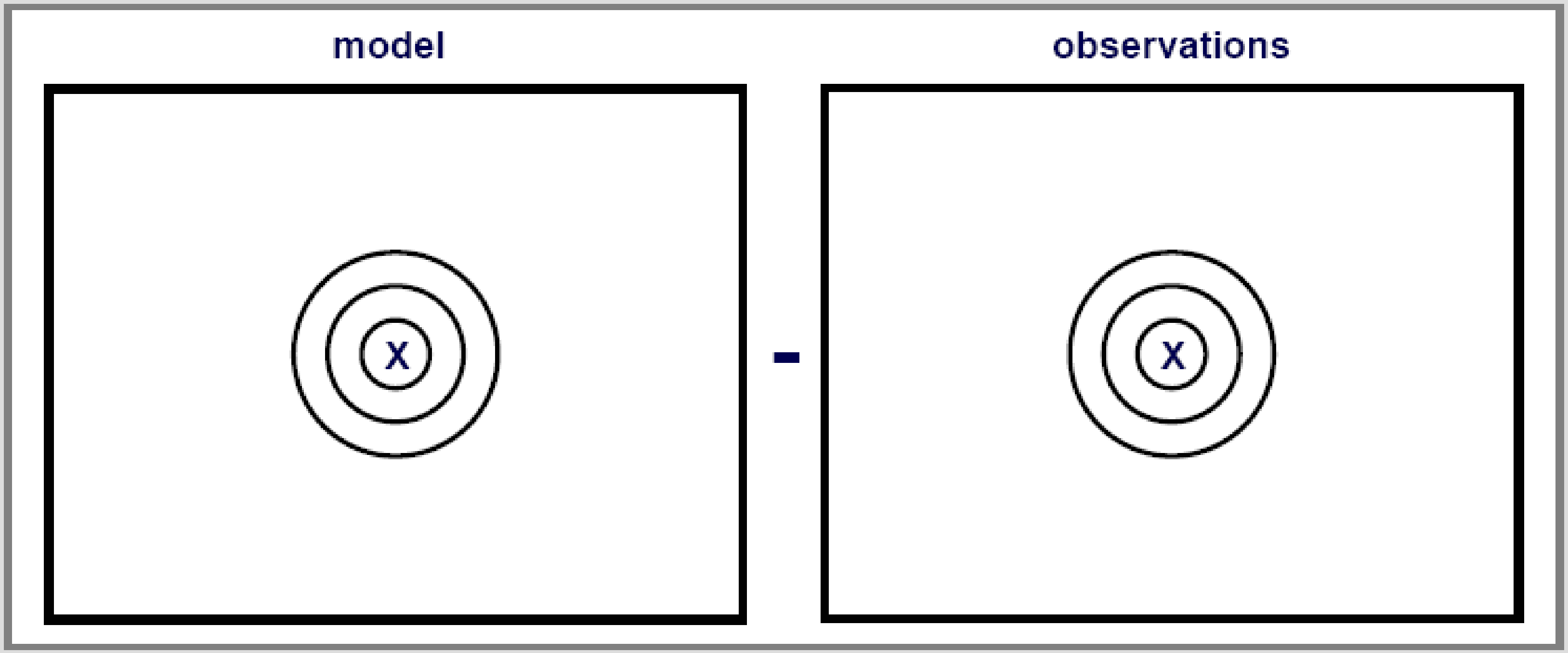
R denotes precipitation
D(…) denotes the area mean in the chosen area
amplitude error



S A L – defintion of the L-component

$L = |r(R_{mod}) - r(R_{obs})| = 0$

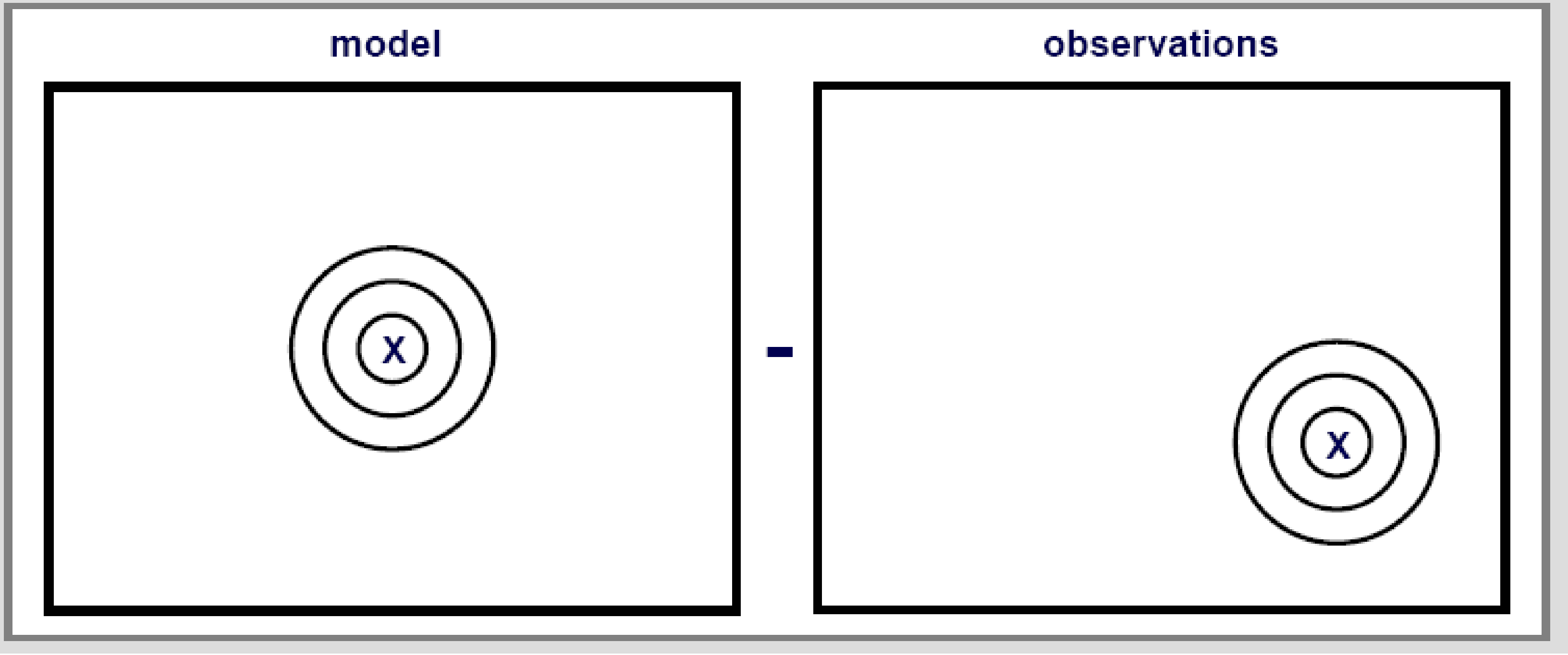
r(…) denotes the precipitation center of mass in the chosen area
displacement error of the center of mass



S A L – defintion of the L-component

$L = |r(R_{mod}) - r(R_{obs})| > 0$

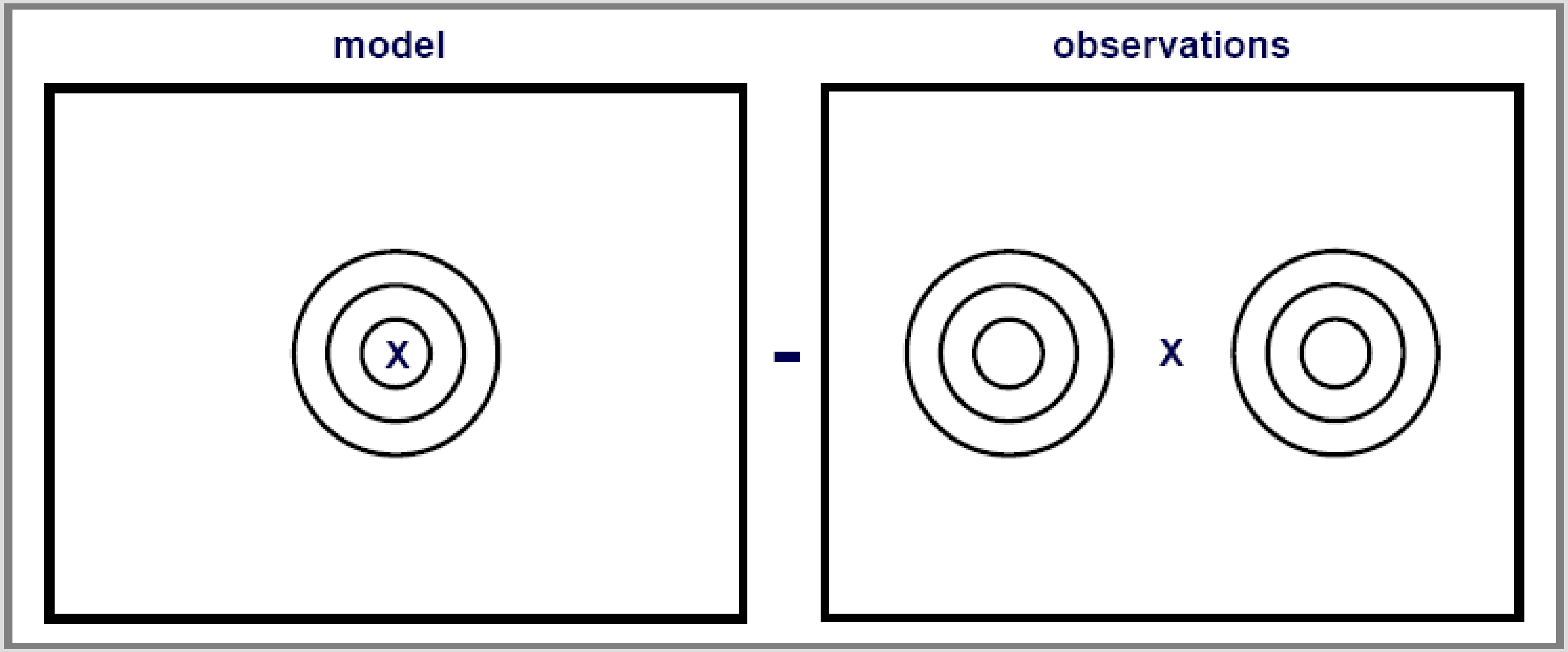
r(...) denotes the precipitation center of mass in the chosen area
displacement error of the center of mass



S A L – defintion of the L-component

$L = |r(R_{mod}) - r(R_{obs})| = 0$!?

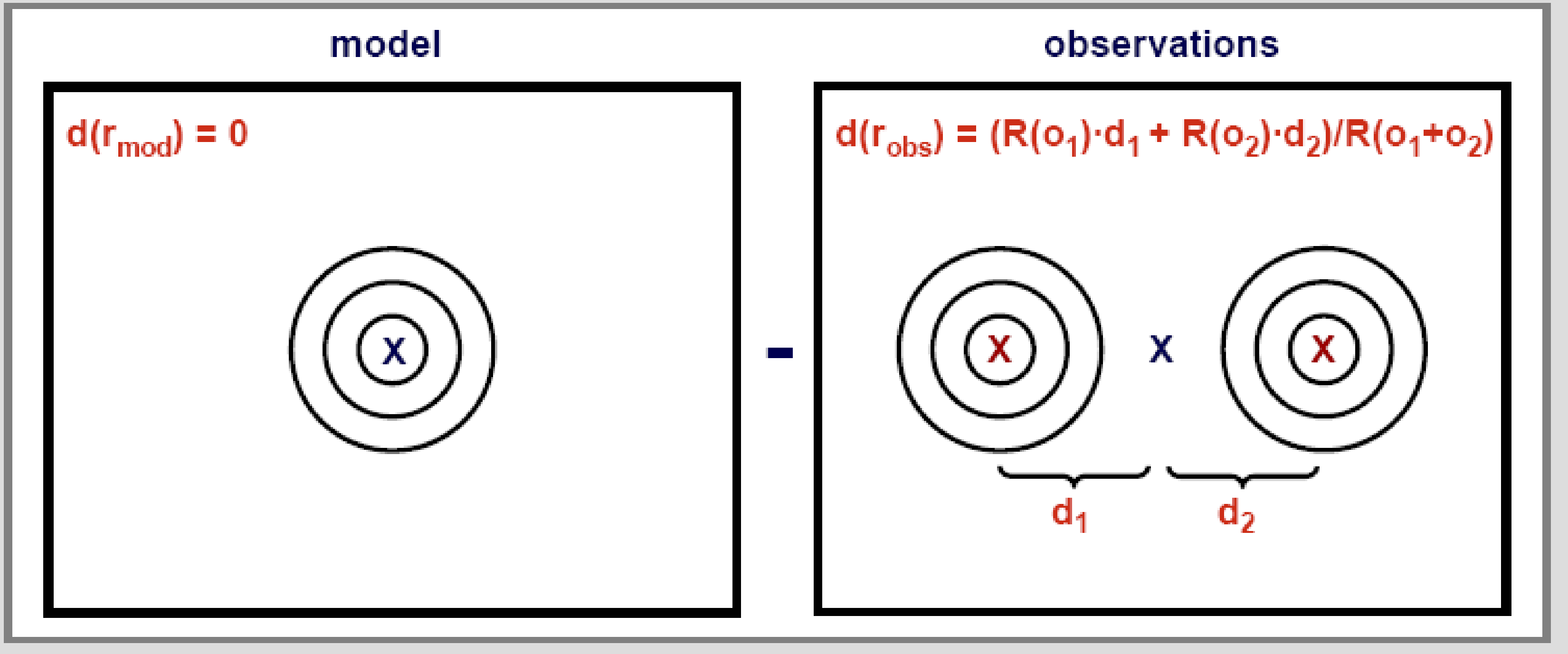
r(…) denotes the precipitation center of mass in the chosen area
displacement error of the center of mass



S A L – defintion of the L-component

$$L = |r(R_{mod}) - r(R_{obs})| + |d(r_{mod}) - d(r_{robs})| > 0$$

r(…) denotes the precipitation center of mass in the chosen area
d(…) mean (weighted) distance between objects- and area-center of mass
displacement error with impact of object distance to center of mass



$$S = (V(R_{mod}^*) - V(R_{obs}^*)) / 0.5^*(V(R_{mod}^*) + V(R_{obs}^*))$$

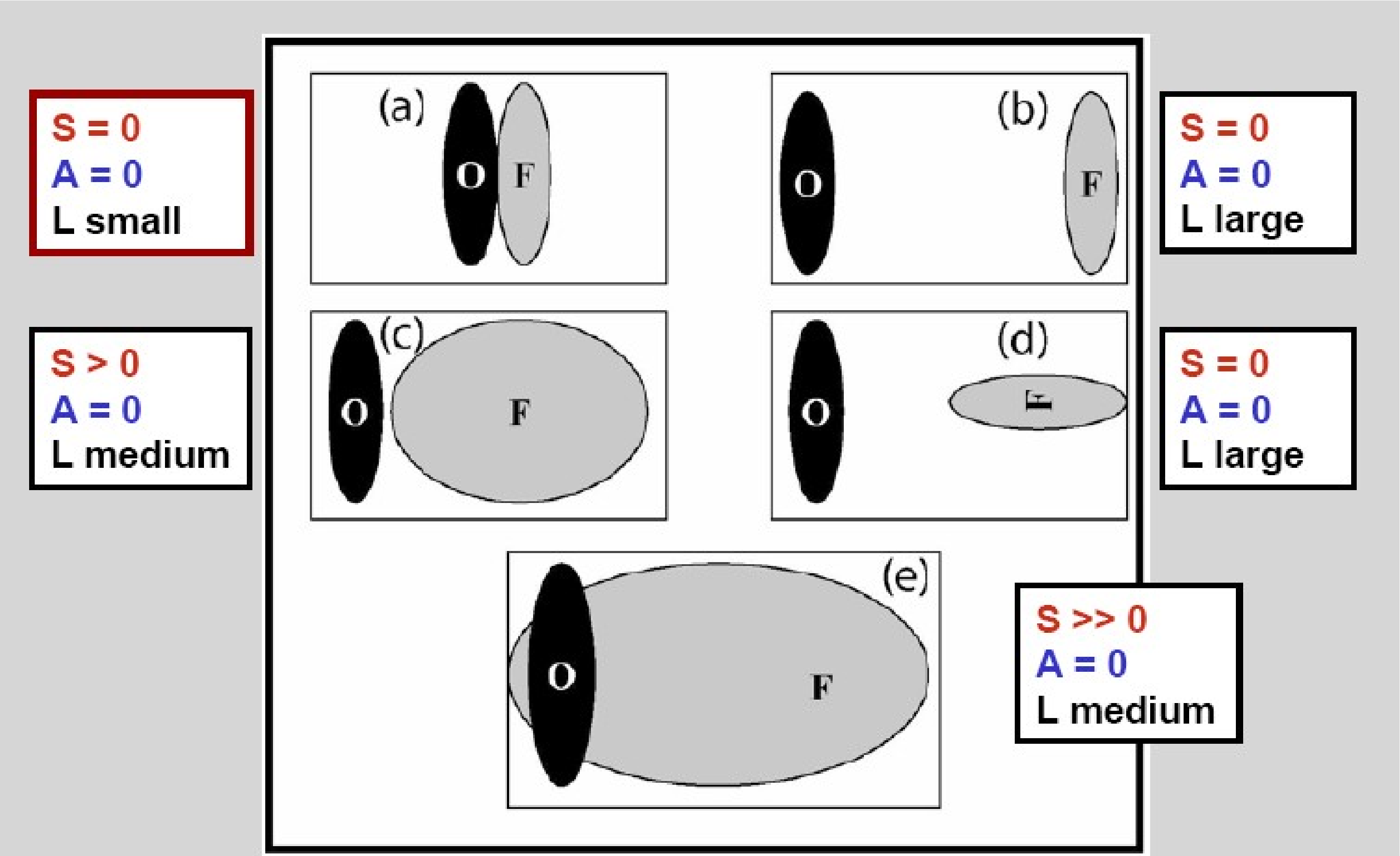
V(…) denotes the averaged volume of all scaled precipitation objects
scaled structure error in a catchment
$S \in [-2, \ldots, 0, \ldots, +2]$

$$A = (D(R_{mod}) - D(R_{obs})) / 0.5^*(D(R_{mod}) + D(R_{obs}))$$

D(…) denotes the area mean in a catchment
scaled amplitude error
$A \in [-2, \ldots, 0, \ldots, +2]$

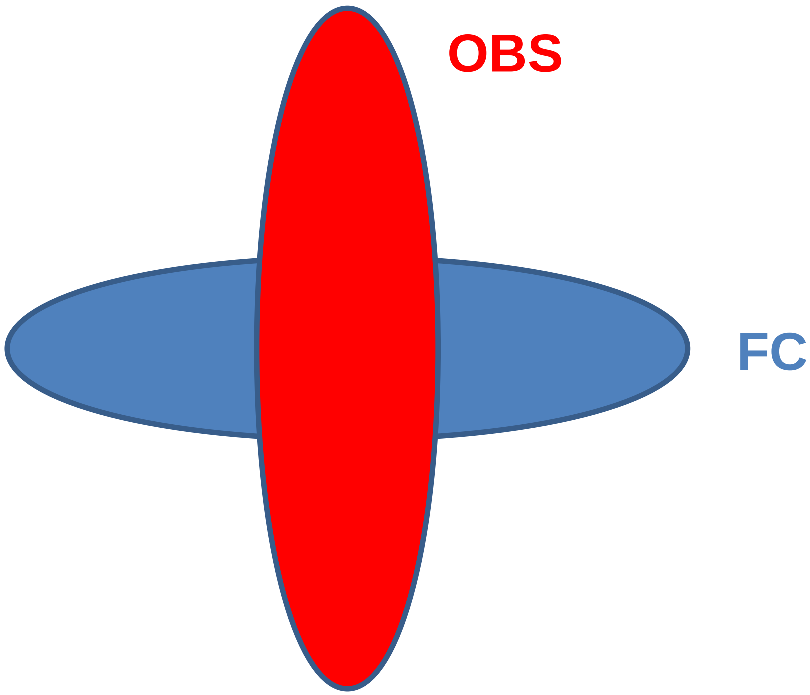
$$L = (|r(R_{mod}) - r(R_{obs})| + 2\cdot|d(r_{mod}) - d(r_{obs})|) / dist_{max}(area)$$

r(…) denotes the precipitation center of mass in a catchment
d(…) mean (weighted) distance between objects- and area-center of mass
scaled displacement error of the center of mass
$L \in [0, \ldots, 2]$

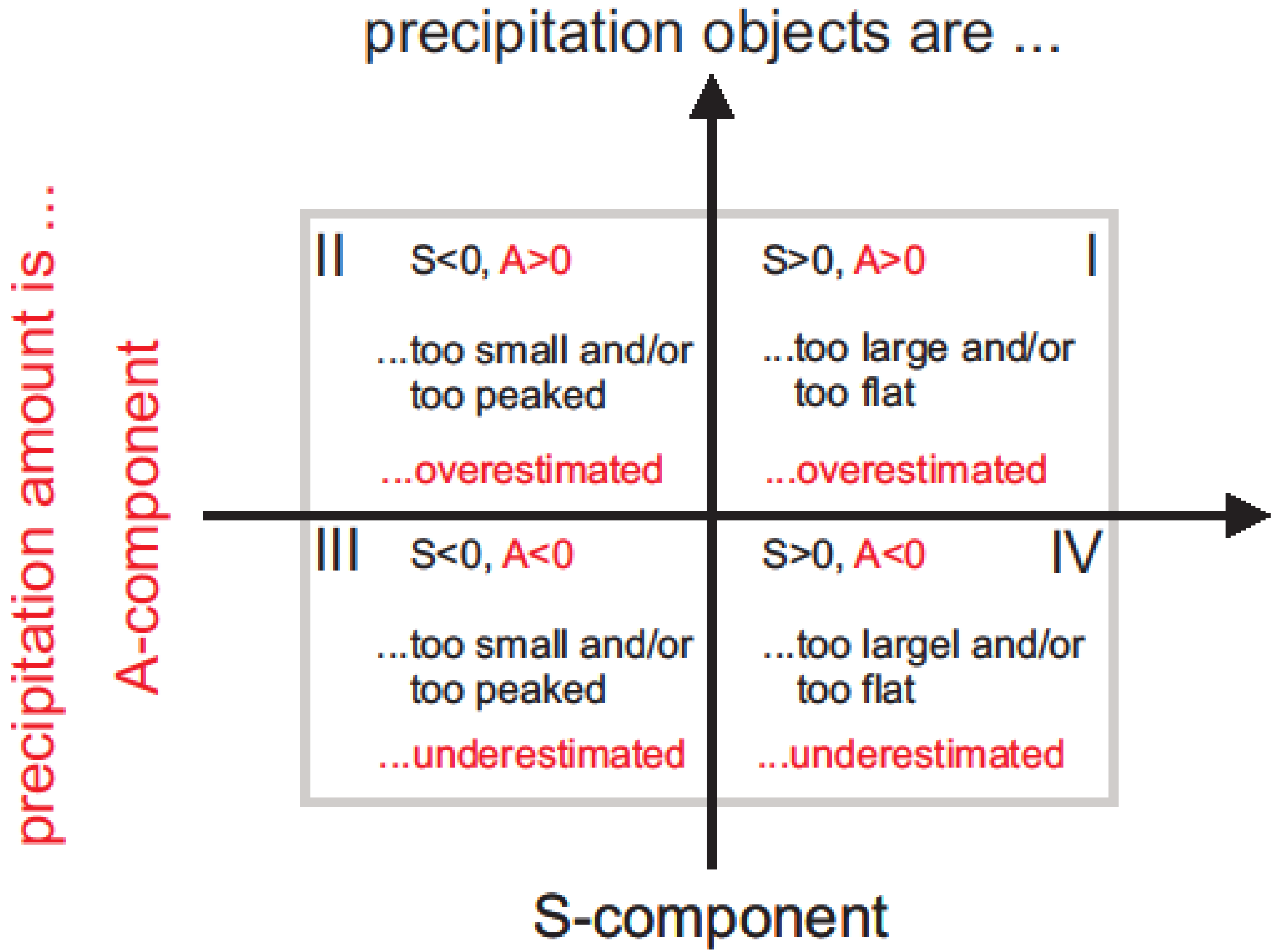


Davis et al. 2006

Q: Look at precip fields. What do you expect for S, A and L?

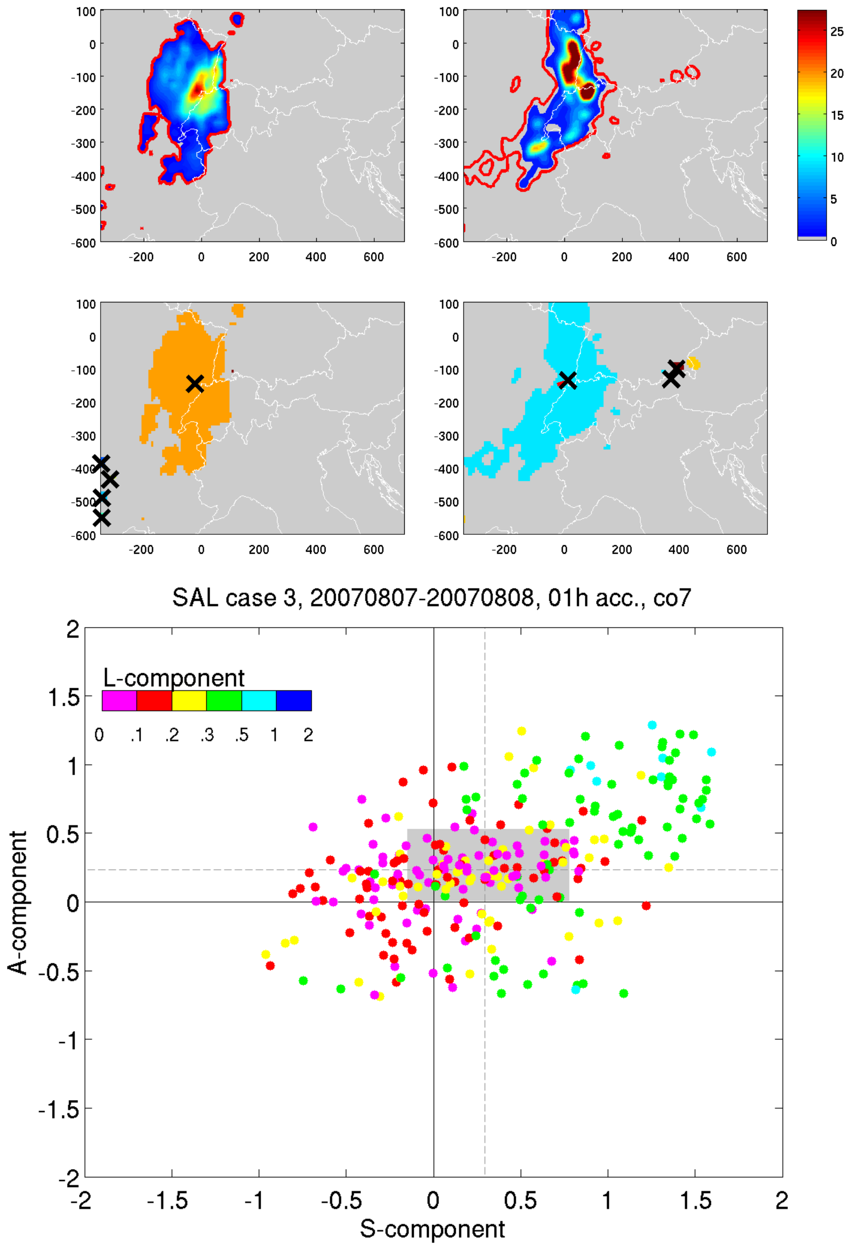


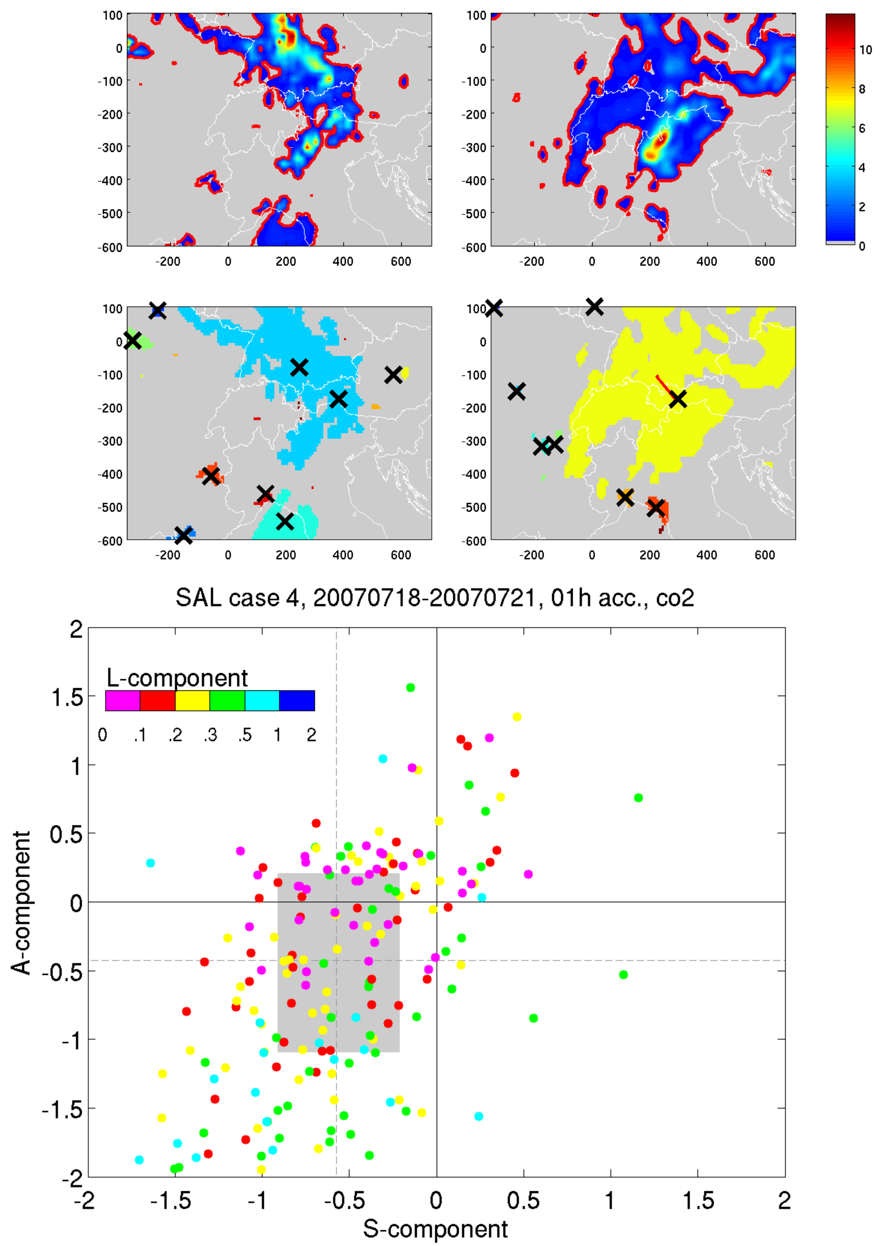
A: S=A=L=0; SAL is invariant against pure rotation.



precipitation objects are ...
precipitation amount is ...
A-component
II
S<0, A>0
...too small and/or too peaked
...overestimated
S>0, A>0
I
...too large and/or too flat
...overestimated
III
S<0, A<0
...too small and/or too peaked
...underestimated
S>0, A<0
IV
...too largel and/or too flat
...underestimated
S-component

SAL Examples





Field deformation

→ evaluate phase errors

DAS in a nutshell

- <u>D</u>isplacement and <u>A</u>mplitude <u>S</u>core DAS
- constitutes a spatial measure belonging to the *field verification technique*
- is based on an areal image matcher using classical *optical flow* technique
- has *two* components: DIS and AMP (normalized with characteristical values)
- is applied in both *observation* and *forecast space* (to account for misses and false alarms)
- has been used in *deterministic* mode
- is coded in *python* and freely available

(Keil and Craig, WAF 2009)

Optical flow algorithm: Pyramidal Matching

1. **Project** observed and simulated images to same grid

2. **Coarse-grain** both images by averaging of 2^F pixels onto one pixel element

3. Compute a displacement vector field that **minimizes the RMSE** within the range of +/- 2 pixel elements

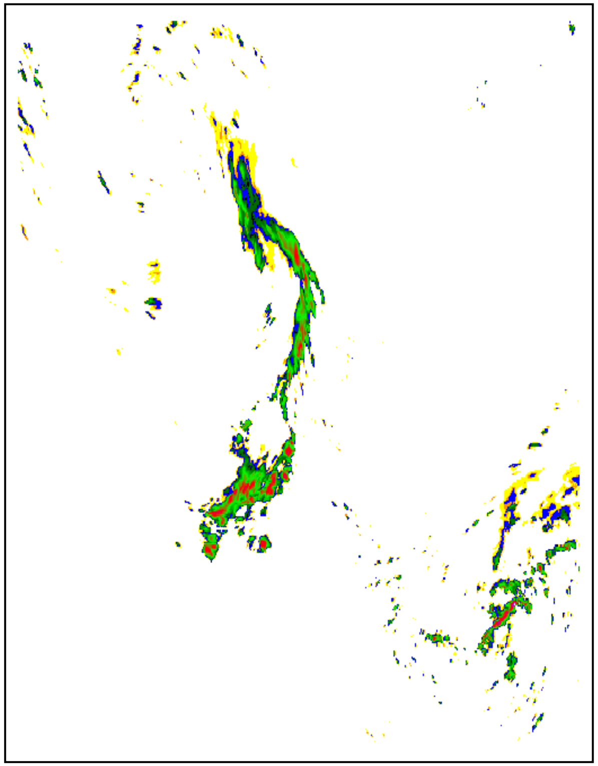
4. Repeat step 2 at successively **finer scales**

5. **Displacement vector** for every pixel results from the sum over all scales

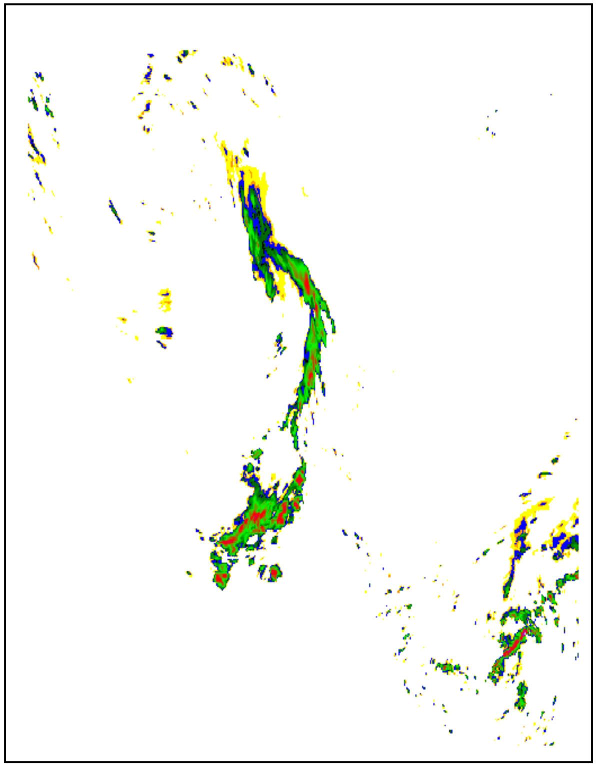
(Mannstein et al., 2002

Pyramidal Image Matching

Step 1: projection on same grid



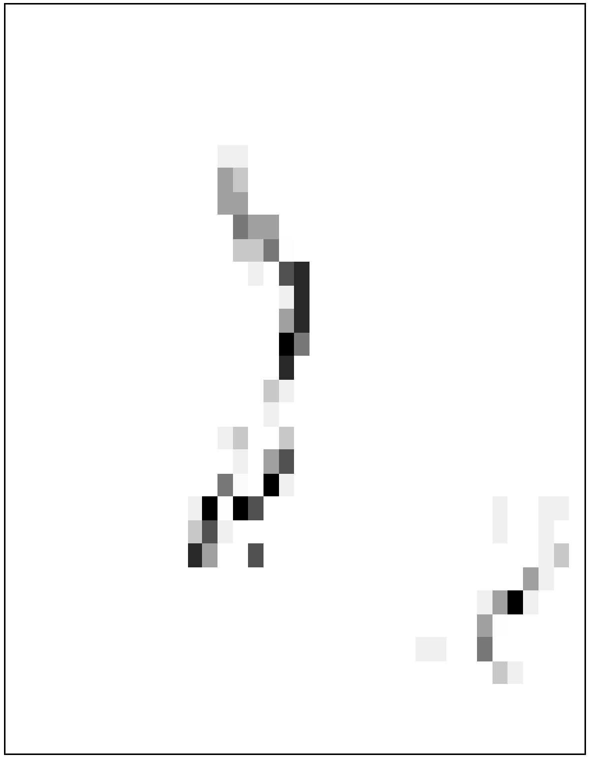
observation



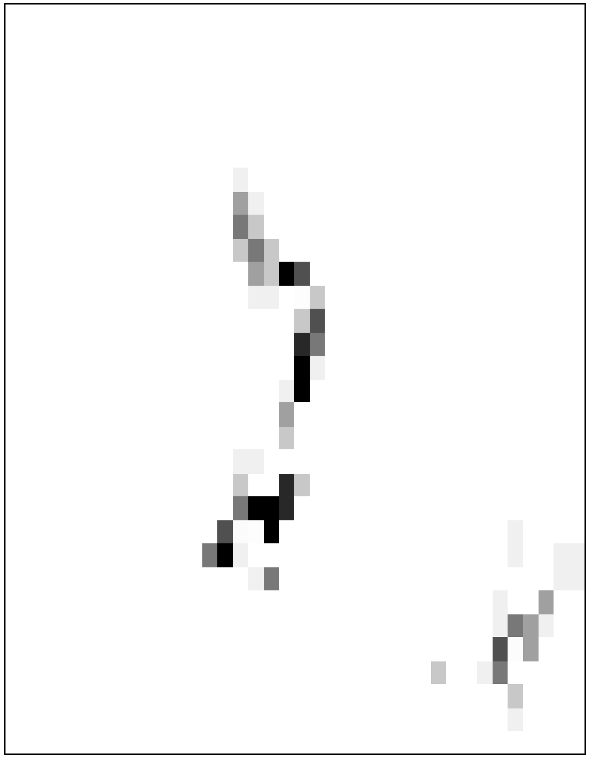
forecast

Pyramidal Image Matching

Step 2: coarse graining F=4



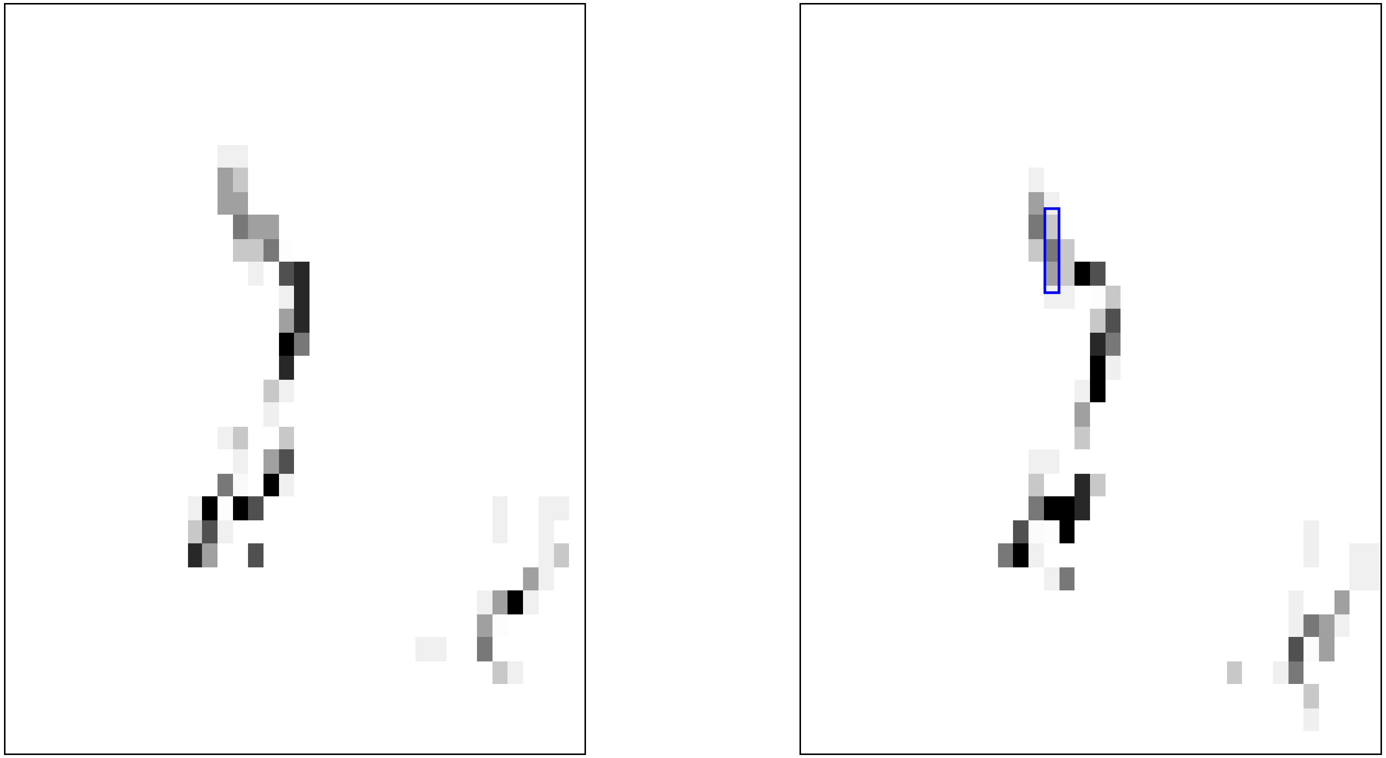
observation



forecast

Pyramidal Image Matching

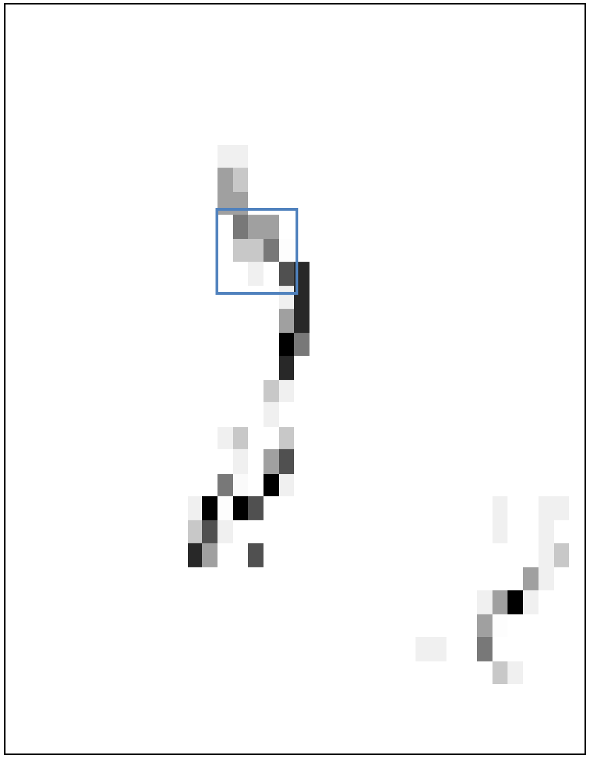
Step 3: compute displacement vector by minimizing RMSE



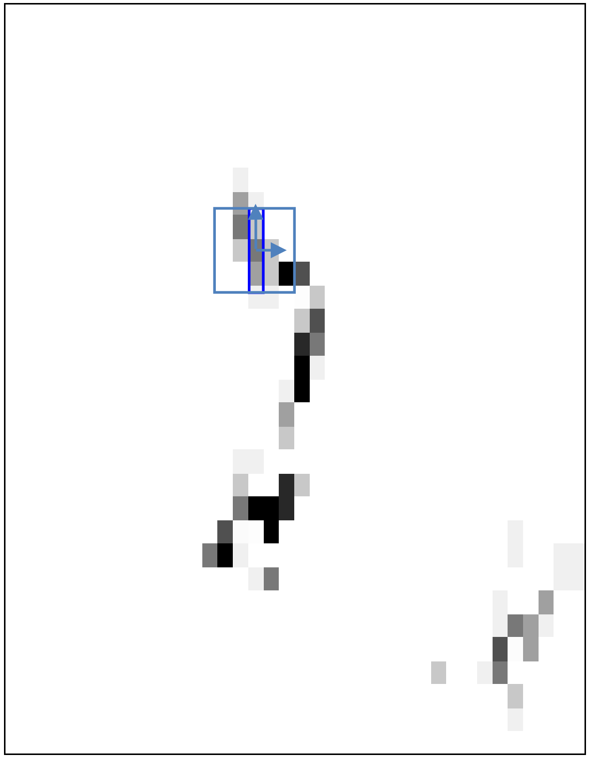
observation forecast

Pyramidal Image Matching

Step 3: compute displacement vector by minimizing RMSE



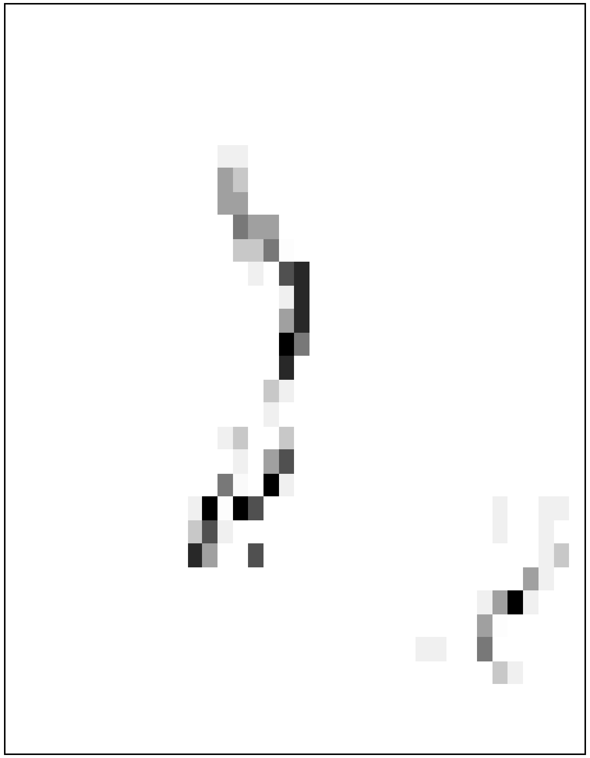
observation



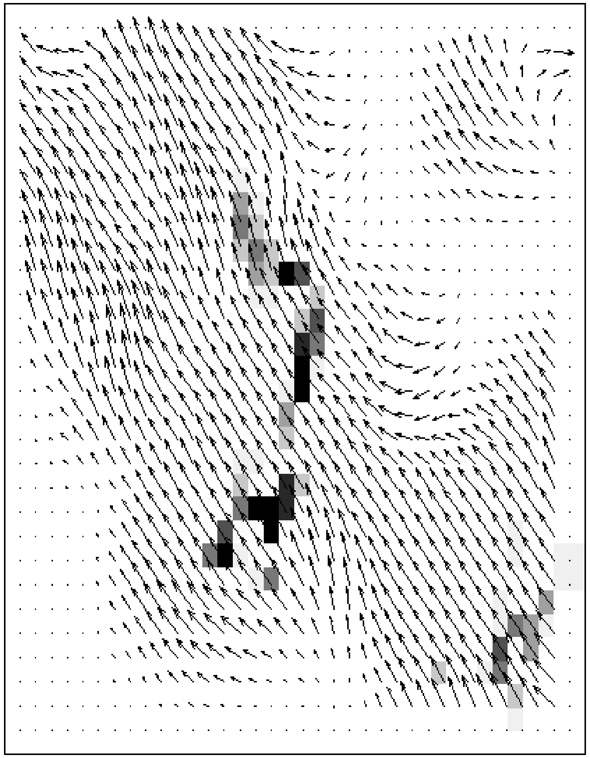
forecast

Pyramidal Image Matching

Step 3: compute displacement vector by minimizing RMSE



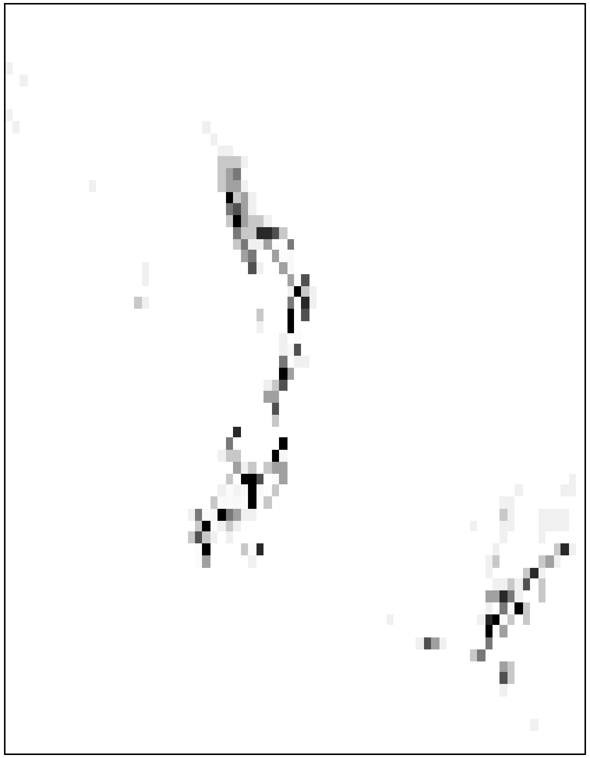
observation



forecast

Pyramidal Image Matching

Step 4: cycle on finer scales using morphed image

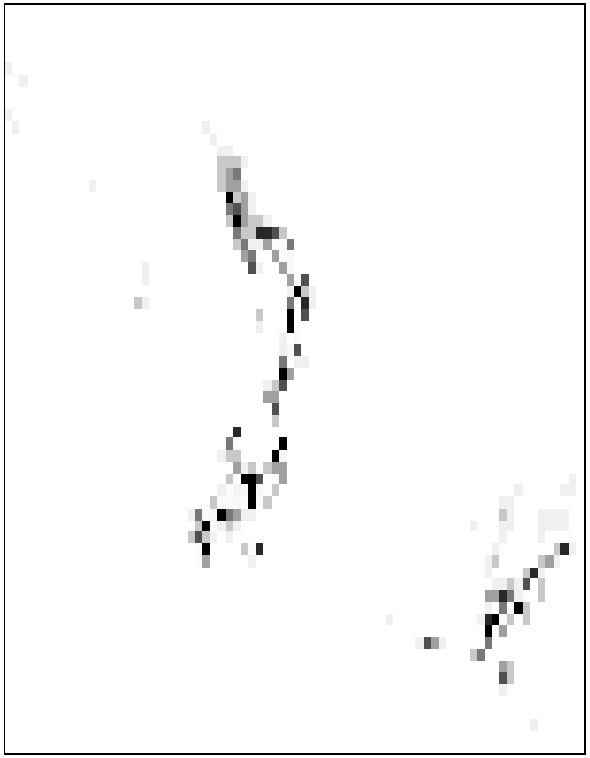


observation

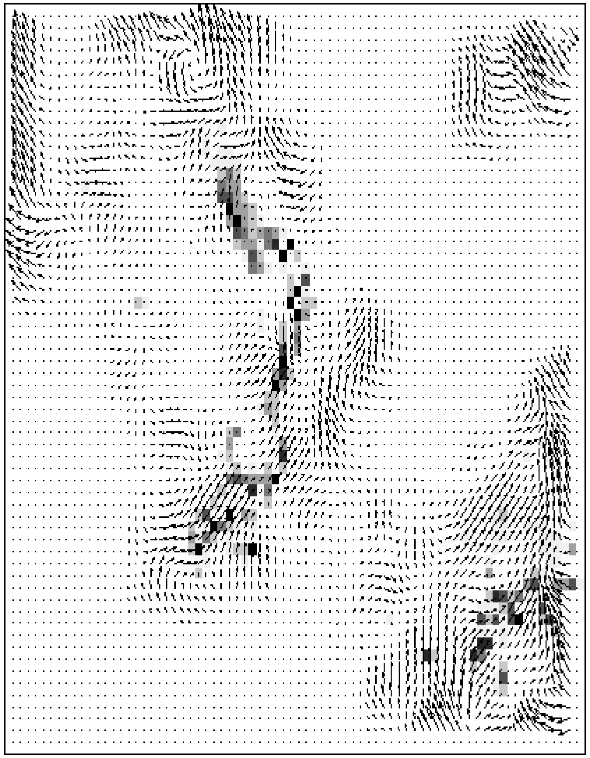
morphed forecast

Pyramidal Image Matching

Step 4: cycle on finer scales using morphed image



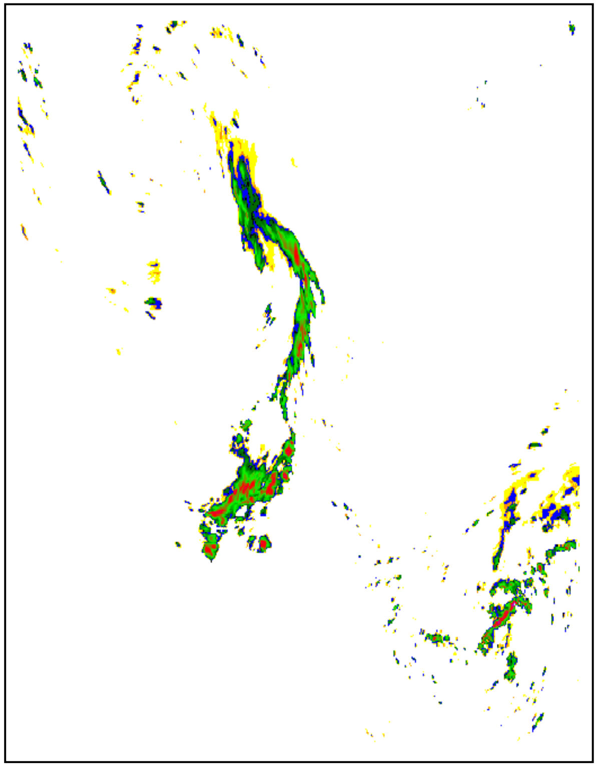
observation

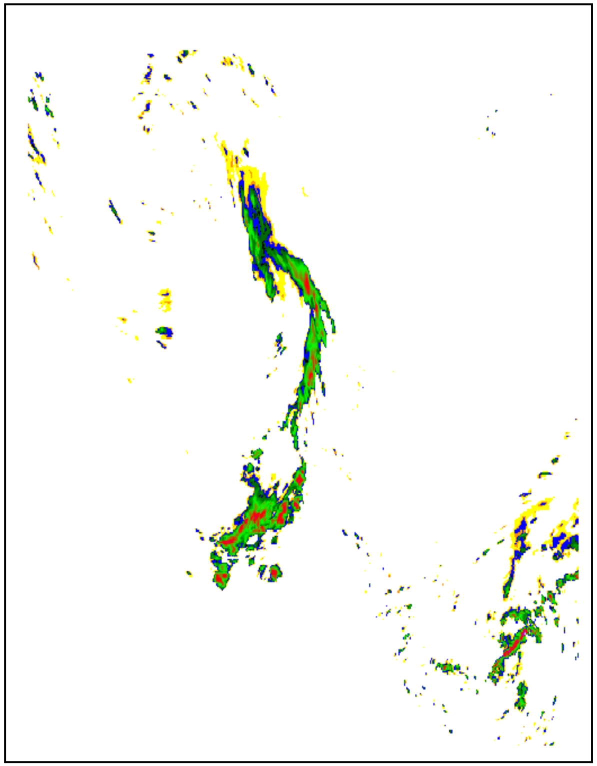


morphed forecast

Pyramidal Image Matching

Step 5: sum over all scales



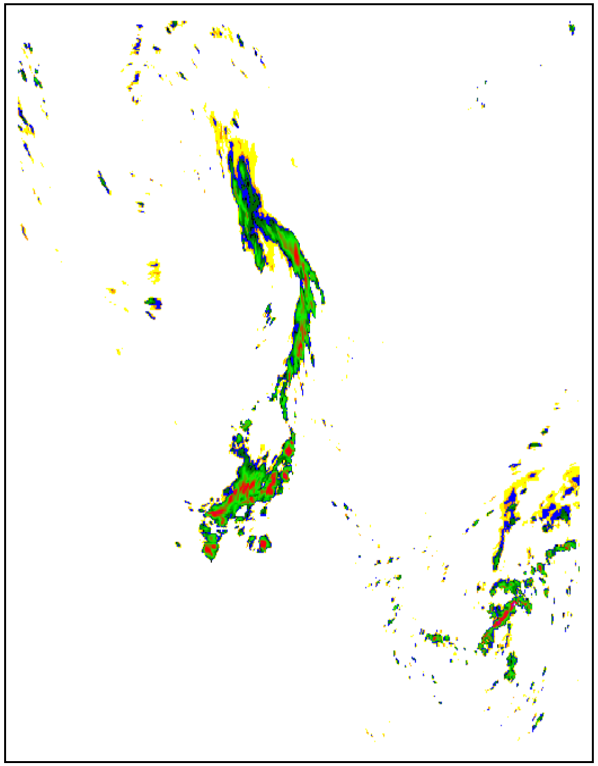


observation

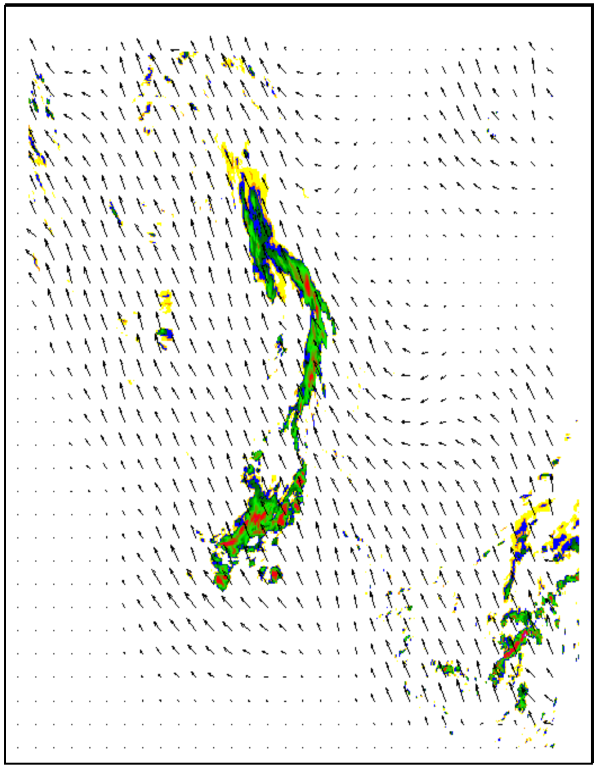
morphed forecast

Pyramidal Image Matching

Step 5: sum over all scales



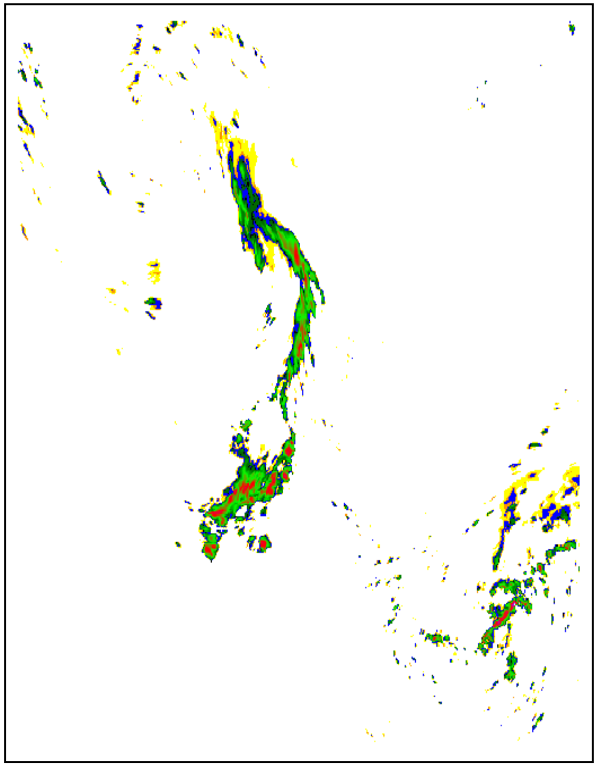
observation



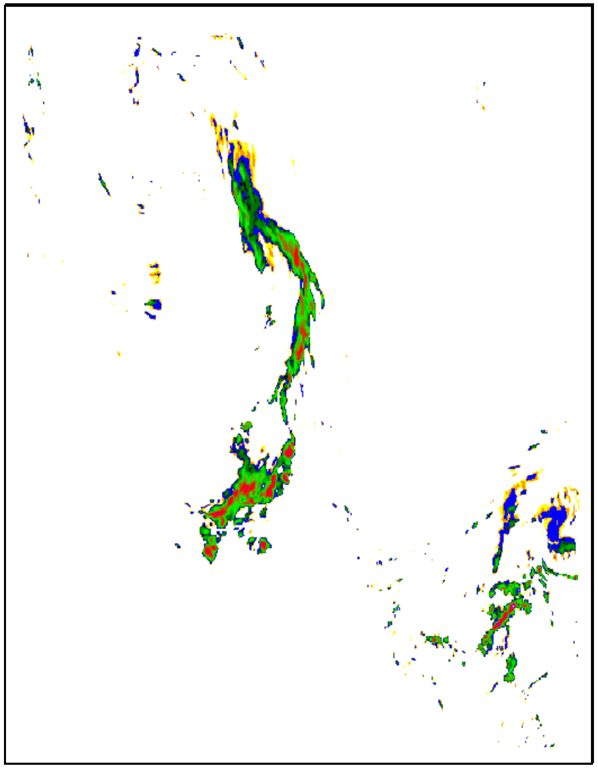
morphed forecast

Pyramidal Image Matching

Step 5: sum over all scales



observation



morphed forecast

Displacement error field DIS

DIS

Displacement error field DIS

and Amplitude error field AMP

DIS

AMP

DAS field: combined DIS and AMP fields

DAS

Displacement and Amplitude Score DAS

DAS has two components:

1. displacement error (of observed and forecast imagery)

$$\overline{DIS} = \frac{1}{n}\sum_{A} DIS(x,y)$$

2. amplitude error (RMSE of observed and morphed forecast imagery)

$$\overline{AMP} = \left(\frac{1}{n}\sum_{A} AMP^2(x,y)\right)^{1/2}$$

- **DAS is applied in observation and forecast space:**

$$DIS = \frac{1}{\left(n_{obs} + n_{fct}\right)}\left(n_{obs}\overline{DIS}_{obs} + n_{fct}\overline{DIS}_{fct}\right)$$

(Keil and Craig, WAF 2009)

Displacement and Amplitude Score DAS

- **DAS shall be a single valued measure** of forecast quality:

➢ underlying principle: complete miss = 100% AMP error

$$DAS = \frac{DIS}{D_{\max}} + \frac{AMP}{I_0}$$

- D_{max} : maximum search distance
- I_0 : characteristic intensity chosen to be typical of amplitude of the observed features

(Keil and Craig, WAF 2009)

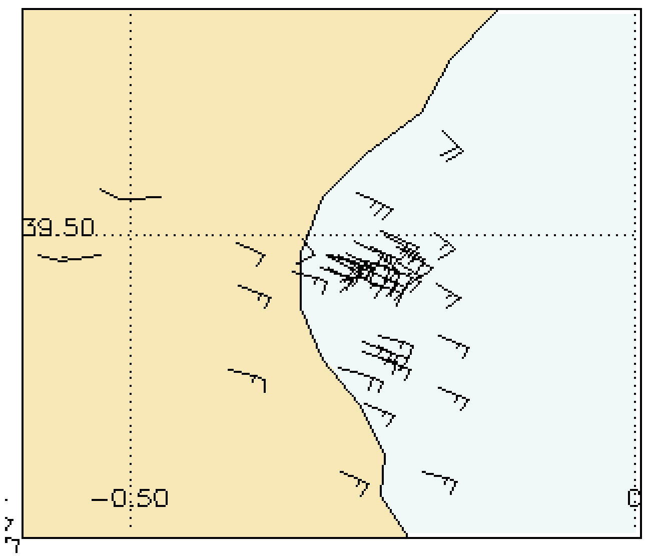
Conclusions

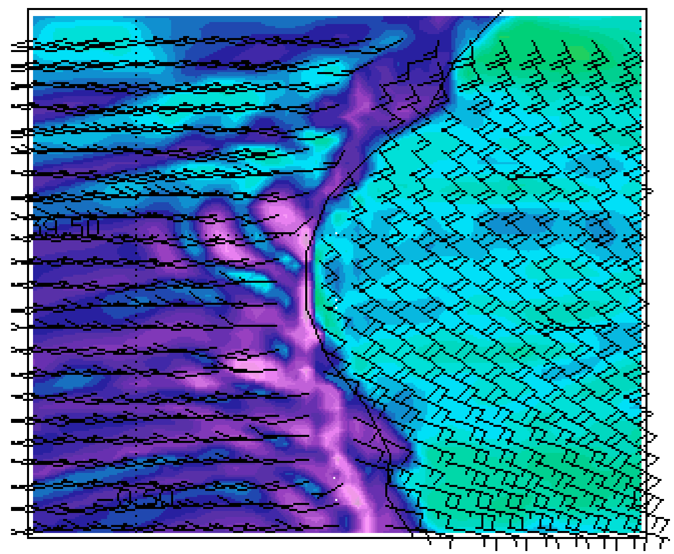
- What method should you use for spatial verification?
  - Depends what question(s) you would like to address

- Many spatial verification approaches
  - Neighborhood – credit for "close" forecasts
  - Scale separation – scale-dependent error
  - Features-based – attributes of features
  - Field deformation – phase and amplitude errors

What method(s) could you use to verify

Wind forecast (sea breeze)

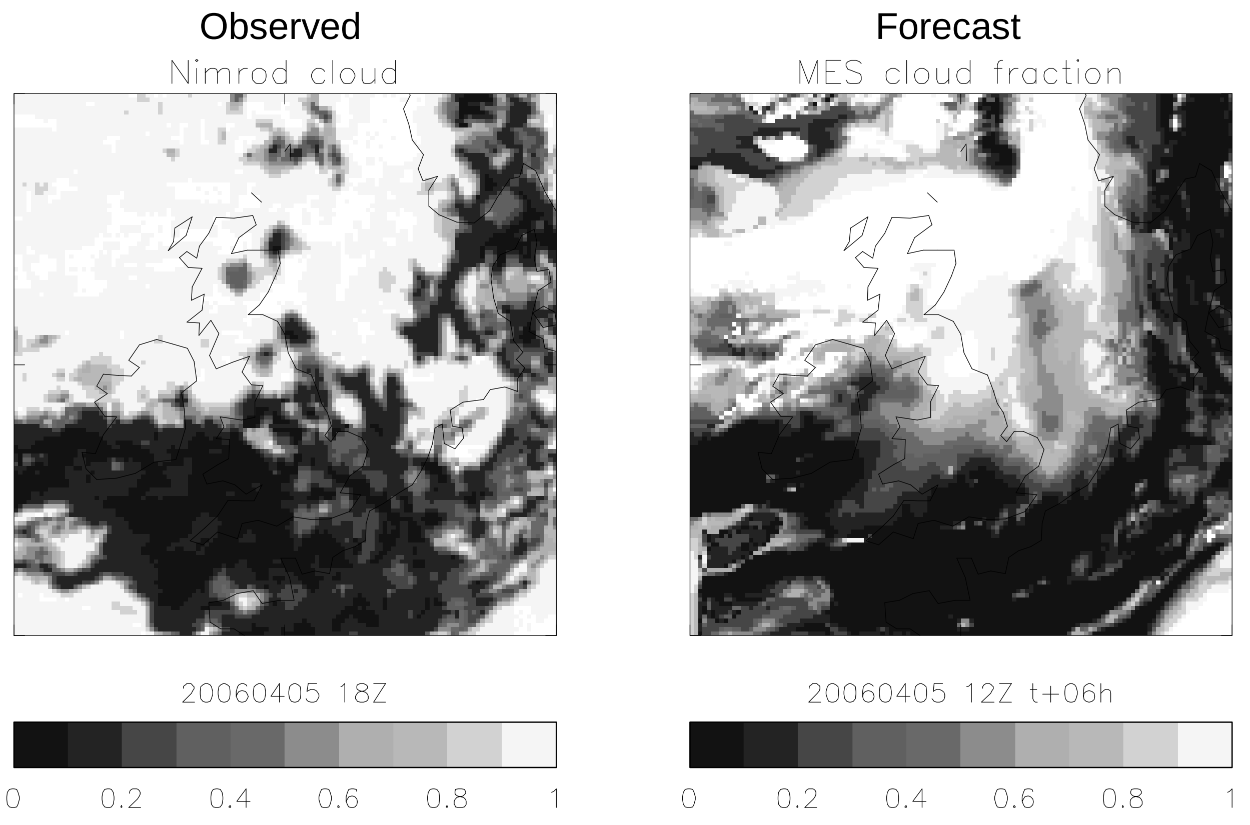




Neighborhood – credit for "close" forecasts
Scale separation – scale-dependent error
Features-based – attributes of features
Field deformation – phase and amplitude errors

What method(s) could you use to verify

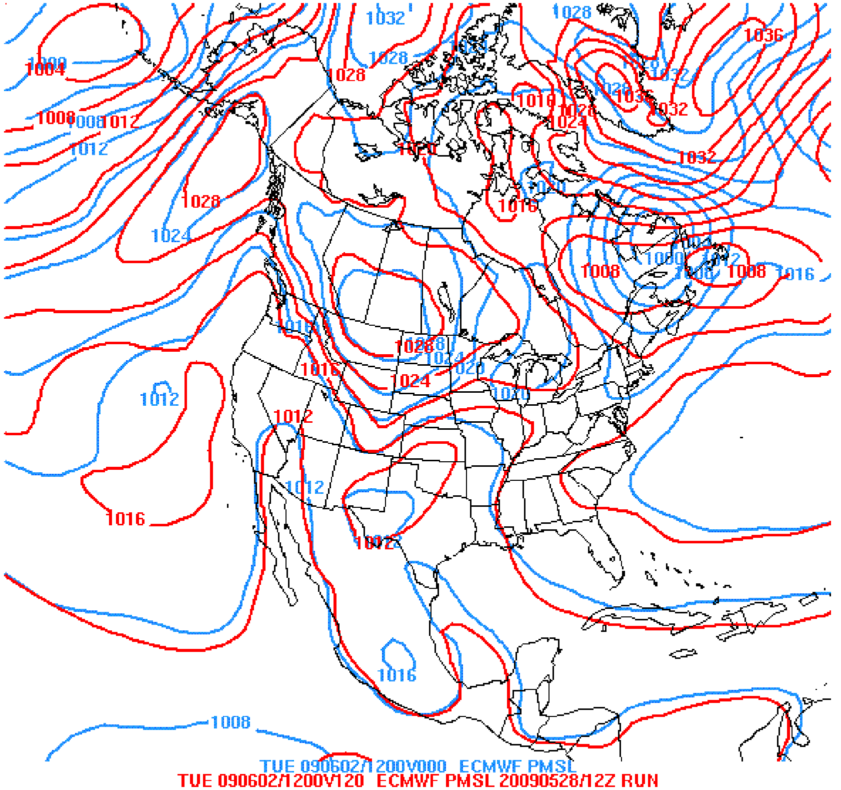
Cloud forecast



Neighborhood – credit for "close" forecasts
Scale separation – scale-dependent error
Features-based – attributes of features
Field deformation – phase and amplitude errors

What method(s) could you use to verify

Mean sea level pressure forecast



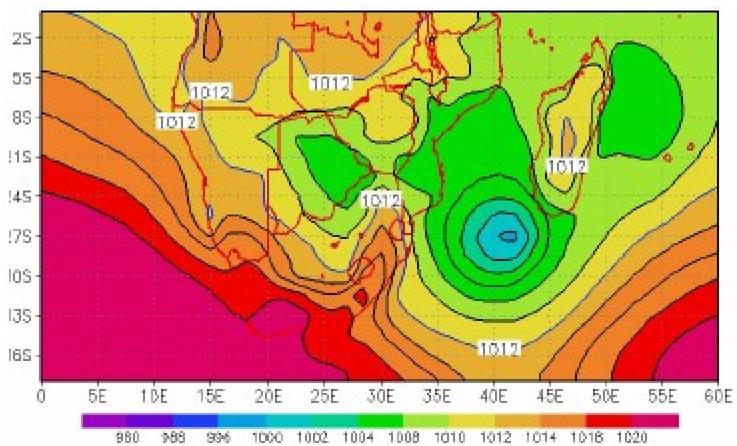
5-day forecast
Analysis

Neighborhood – credit for "close" forecasts
Scale separation – scale-dependent error
Features-based – attributes of features
Field deformation – phase and amplitude errors

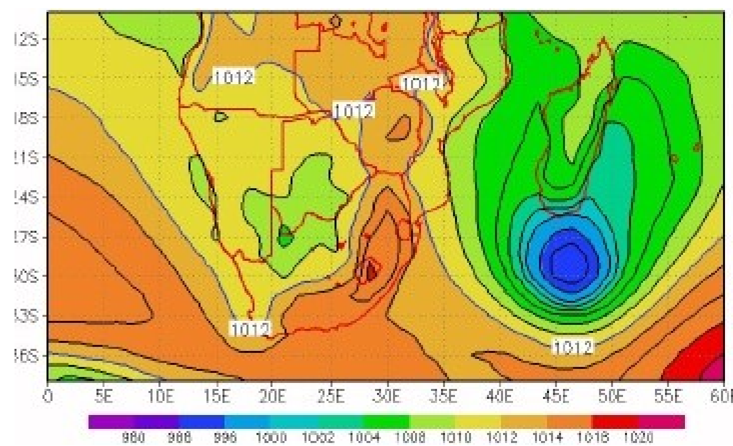
What method(s) could you use to verify

Tropical cyclone forecast

Observed



3-day forecast



Neighborhood – credit for "close" forecasts
Scale separation – scale-dependent error
Features-based – attributes of features
Field deformation – phase and amplitude errors