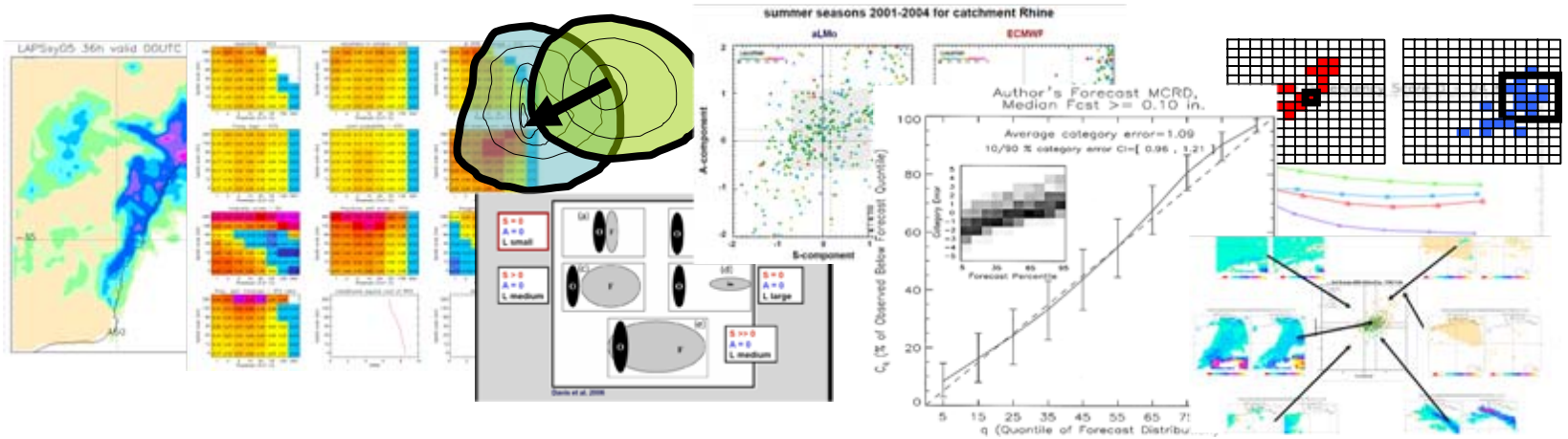
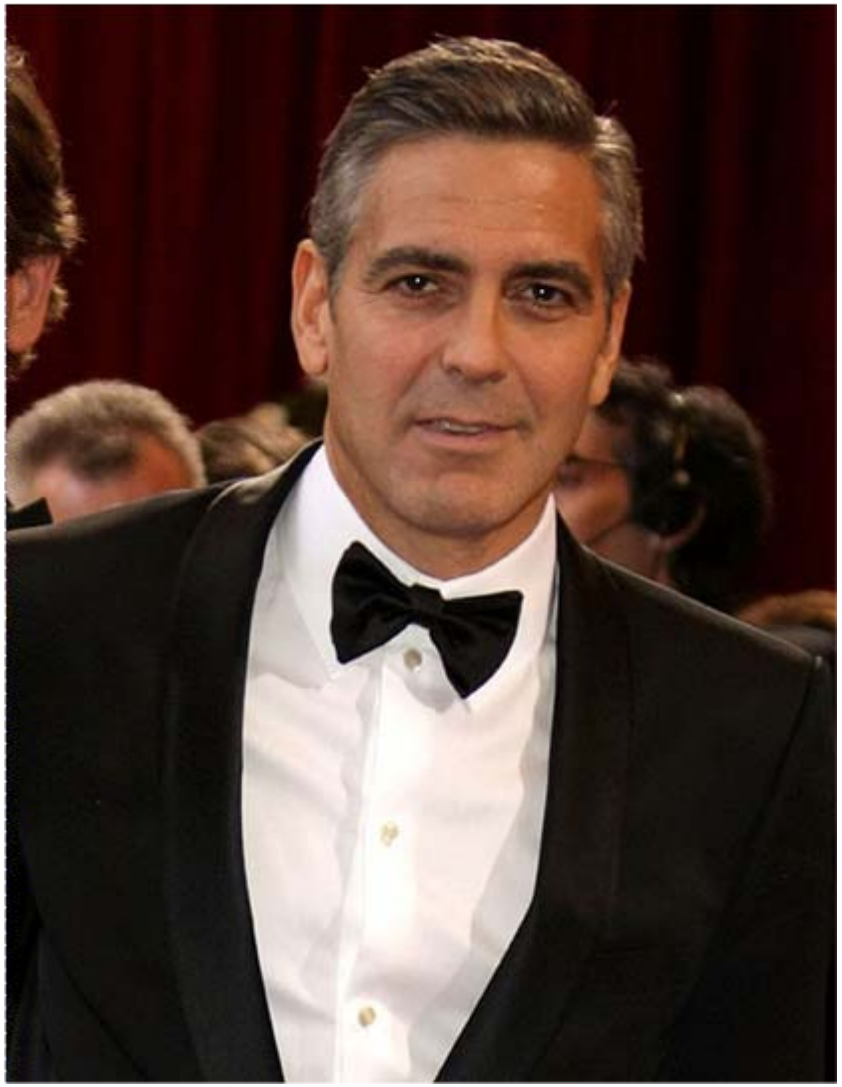
QPF verification with SAL



Carlos Santos – AEMET (Spain), Predictability Group, NWP Apps
Contributions: Arancha Amo, Anna Ghelli, Laura Ferranti, Imanol Guerrero
Contributions: J.A. García-Moya, I. Martínez, A. Amo, A. Callado, P. Escribà, J. Montero, D. Santos, J. Simarro (AEMET); P. Doblas (IC3); A. Ghelli, R. Buizza, R. Hagedorn, M. Leutbecher (ECMWF)
Acknowledgements: J. A. López, A. Chazarra, O. García, J. Calvo, B. Navascués, Climatic Database Staff, Computer Systems Staff, member and cooperating states ECMWF
This work is partially funded by project PREDIMED CGL2011-24458 from the Spanish Ministerio de Ciencia en Innovación.

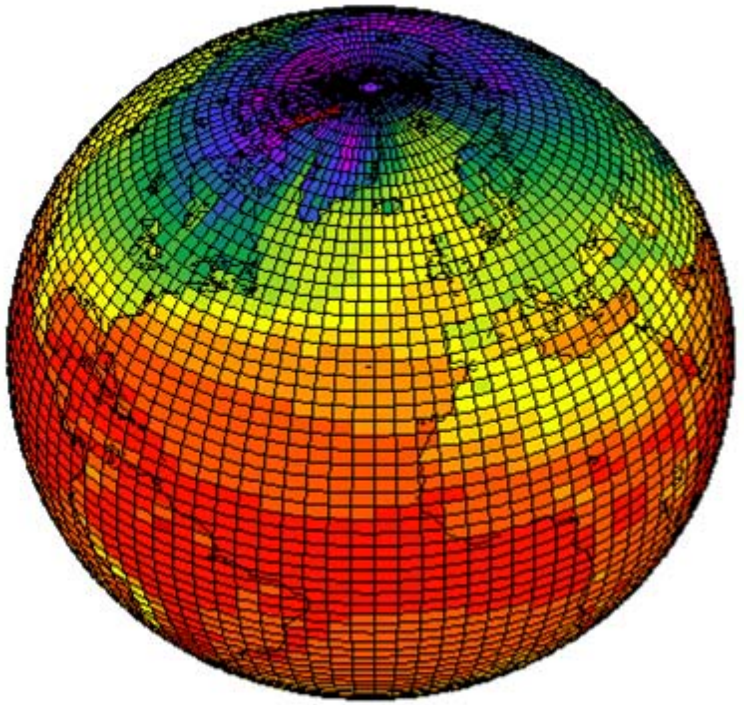
Are the models perfect?

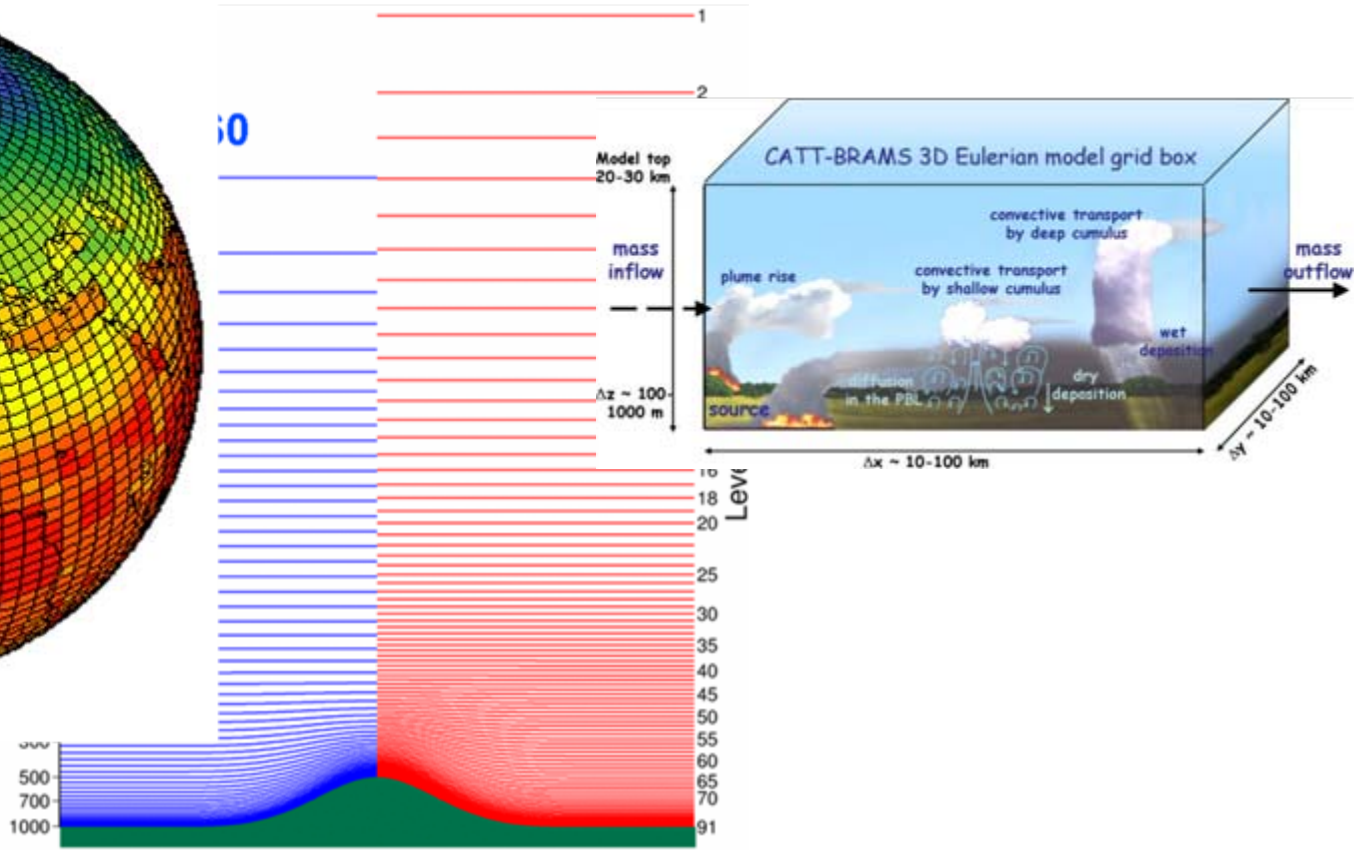


Are the models perfect?

Are models good representations of reality?

Are the models perfect?





Models are only simulations of reality

Outline

- Are the models perfect?
- Forecast verification: an introduction
- Classical methods
- A critical vision
- New spatial verification methods
- SAL: an example of feature-oriented method
- MODE: a method applicable to EPSs
- Radar & satellite data
- Software, conclusions, references

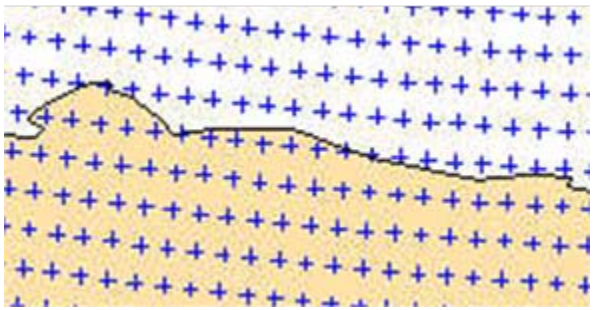
1

- Are the models perfect?
- **Forecast verification: an introduction**
- Classical methods
- A critical vision
- New spatial verification methods
- SAL: an example of feature-oriented method
- MODE: a method applicable to EPSs
- Radar & satellite data
- Software, conclusions, references

What is verification?

Forecast: of a future atmospheric state

Corresponding Observation: (or a good estimate of reality e.g. analysis)



Verification:

forecast vs observation

¿Why verify?

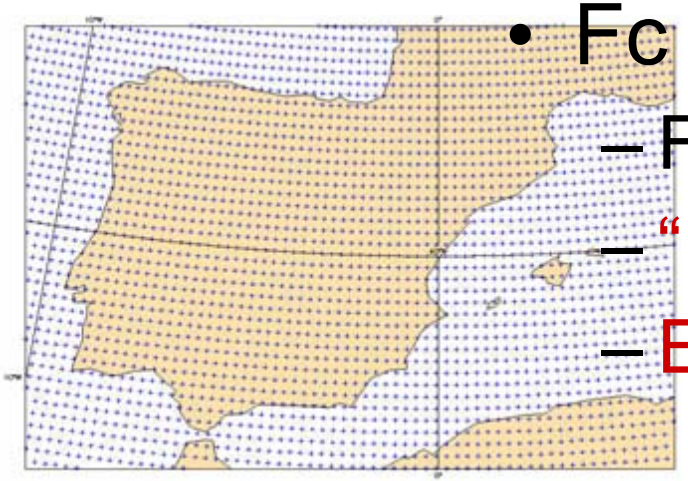
- Improve quality
- Forecaster guidance

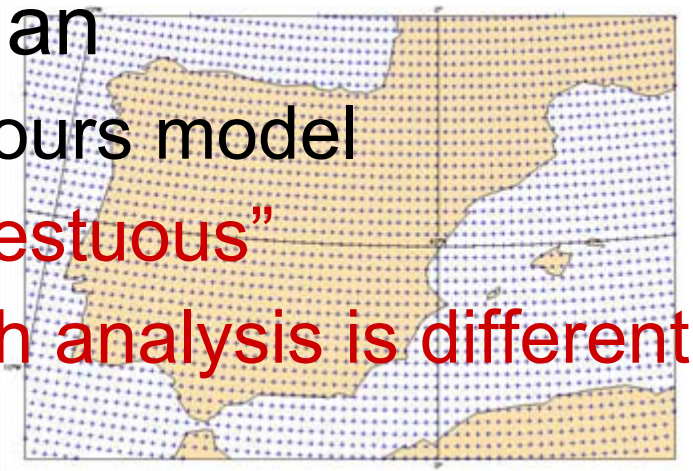
- Verification procedures should be included in a prediction system to:
  - Assess quality, accuracy, trends
  - Ubderstand failures
  - Compare quality of different prediction systems; e.g.: compare the current system with experimental improvements or other systems
- Administrative, scientific & economic (decision making)

Quality and value

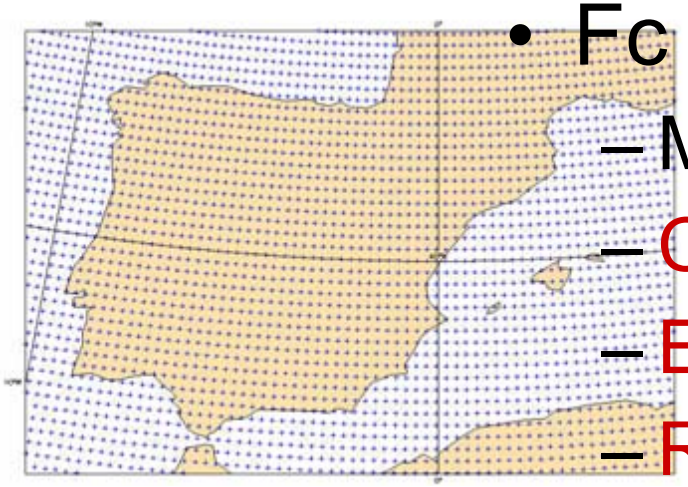
- Quality: Agreement forecast - observation
- Value: helps the user take better decisions
- Examples:
  - Clear skies forecast over Sahara Desert during dry season: high quality, little value
  - Isolated thuderstorms development forecast in a region without pointing the local areas, quality is poor bit it provides a high value to issue early warnings
- Murphy:
  - Murphy adds a third property: consistency with forecaster's judgement

References (“truths”)





- Fc vs an
  - Favours model
  - “Incestuous”
  - Each analysis is different

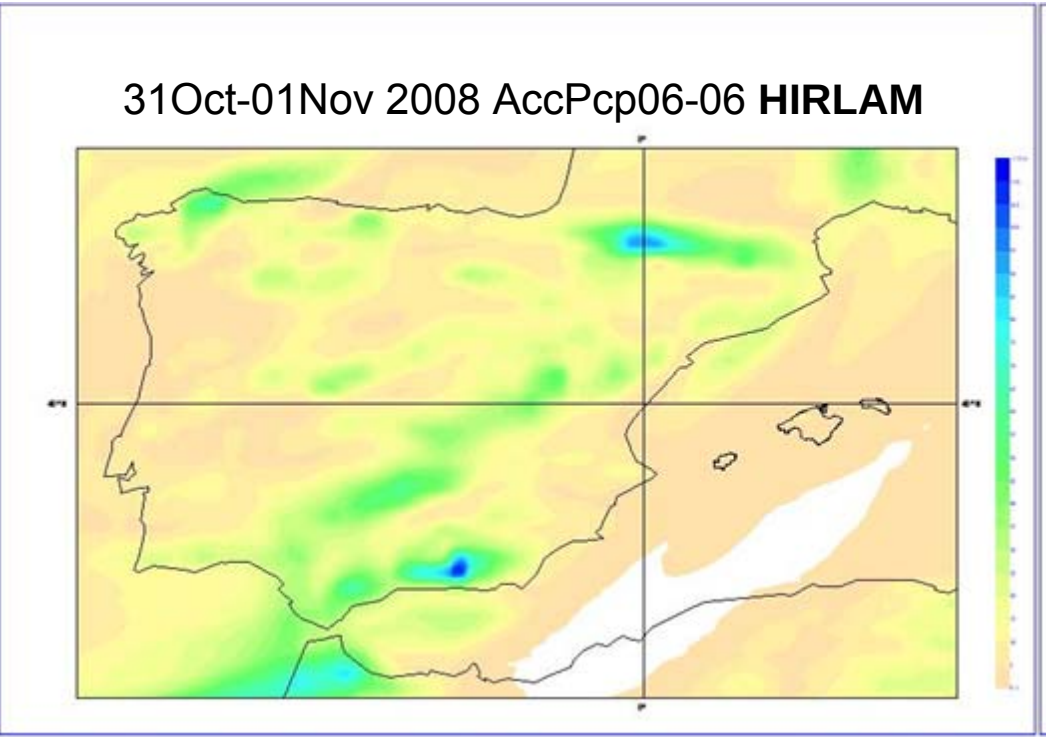


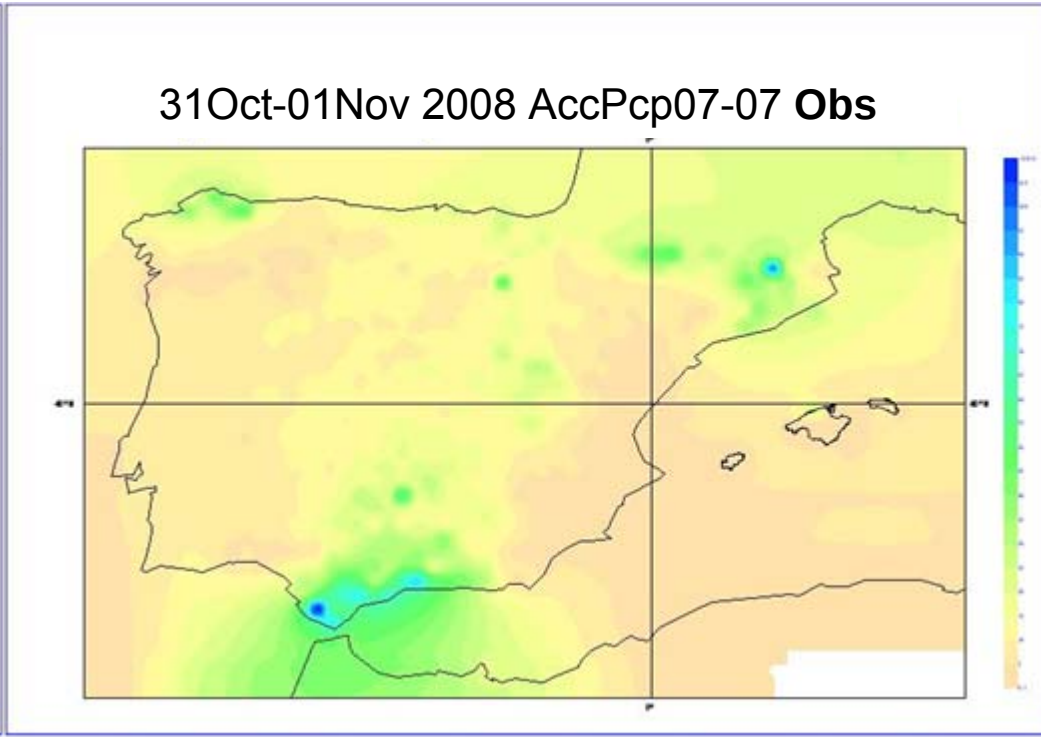
- Fc vs ob
  - More demmanding
  - Covering
  - Error
  - Representativity

2

- Are the models perfect?
- Forecast verification: an introduction
- **Classical methods**
- A critical vision
- New spatial verification methods
- SAL: an example of feature-oriented method
- MODE: a method applicable to EPSs
- Radar & satellite data
- Software, conclusions, references

Subjective Verification

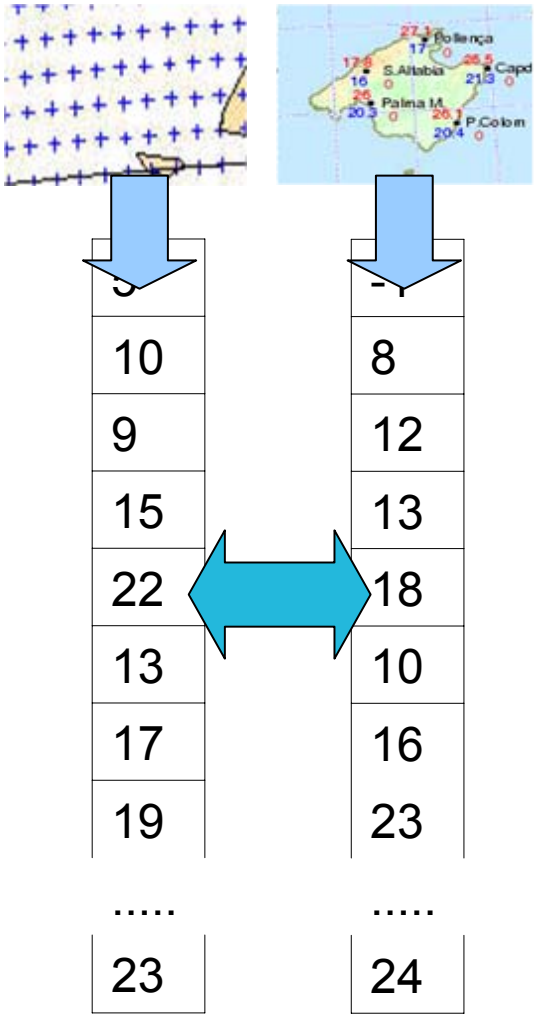




- How did the model captured patterns / structures?
- What range of scales was the forecast better / worse?
- Intensity and / or location of events?
- “Eye ball”

Objective Verification

- **Comparison by pairs (fc,ob),** and study statistical properties of the set $\{(fc,ob)_i\}$
- **Score**: metric related to some properties
- E.g.: bias = fc – ob; 2m T forecast in Sevilla was 4ºC and observed was 3ºC, then bias = 1ºC ¿What can we infer from that?

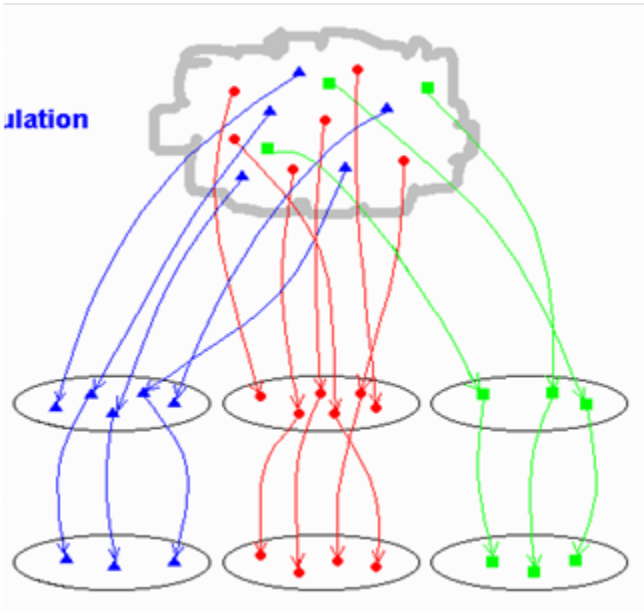


Acumulation vs Estratification

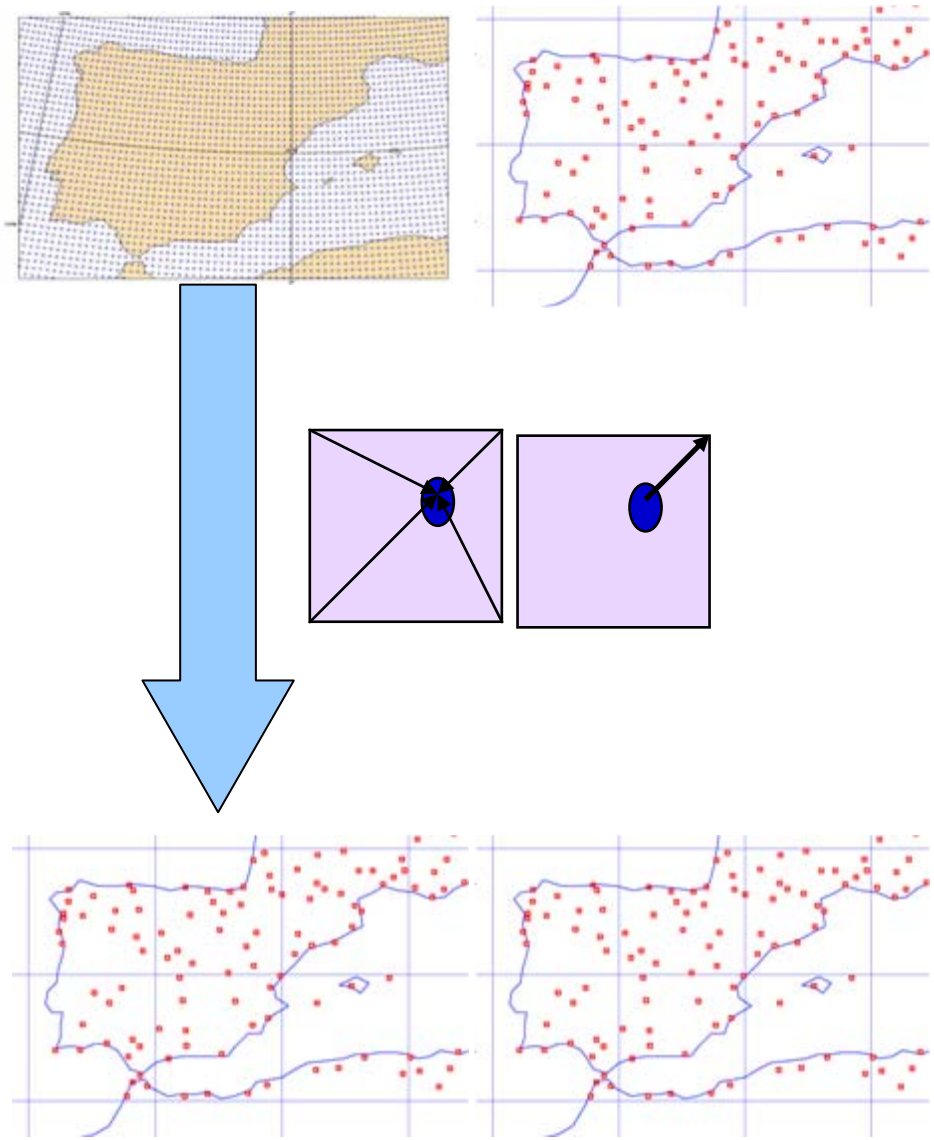
- Pairs forecast / observation

- Pooling in time and / or space
  - Differences between non-homogeneus data can be hidden
  - Biased results to the most common regime sampled (e.g. days without severe weather)

- Stratification in quasi-homogeneus subsets (seasonal, regional …)
  - High-light forecast behaviour according regime (e.g. monsoon)
  - Subsets must contain enough data for significant results
  - Biased samples can be used if different verification methods are used to compare results

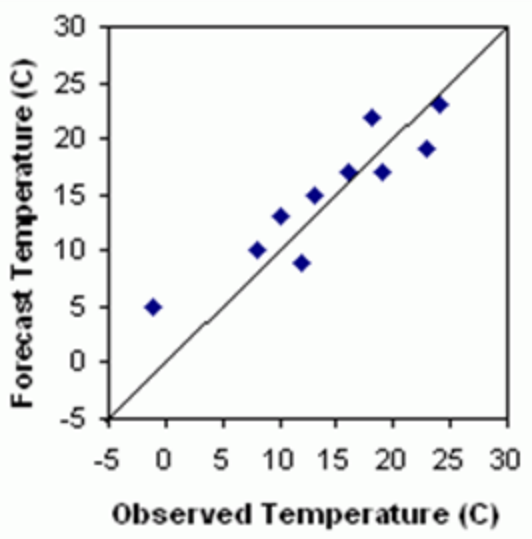


Point vs Up-scaling

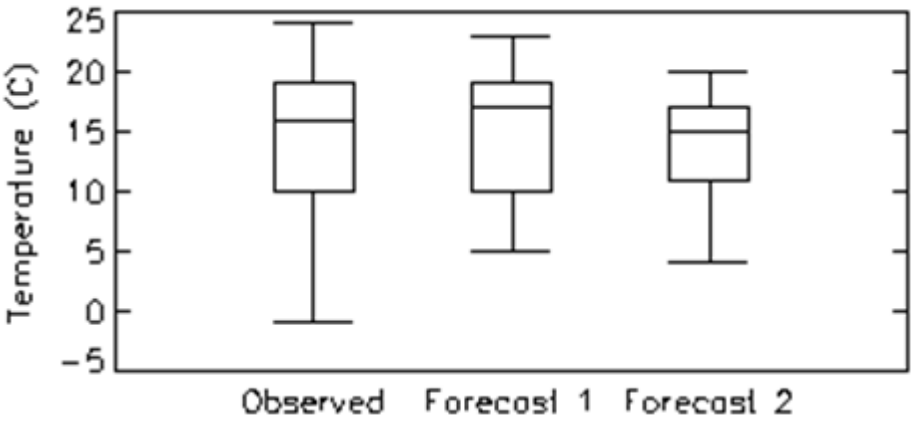


Continuous: descriptive methods

Scatter plot



Box plot

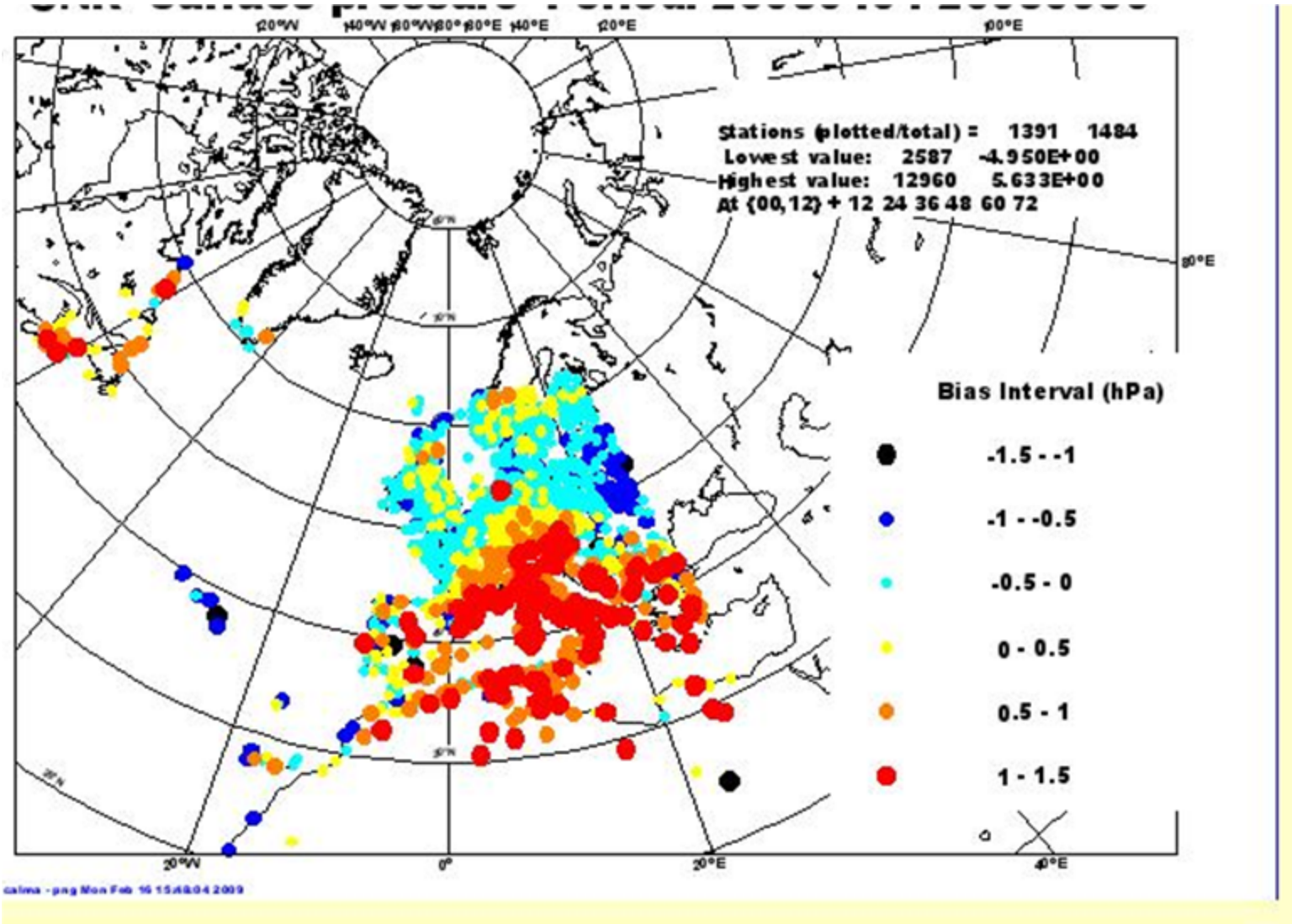


Continuous variables: scores

Score	Cómputo	Significado	Rango	Perf
Bias	$F = \frac{b}{b+d}$	Diferencia entre predicción y observación	(-∞,∞)	0
Error absoluto medio	$H = \frac{a}{a+c}$	Evita compensaciones	[0,∞)	0
Error cuadrático medio	$PC = \frac{a+d}{a+b+c+d}$	Evita compensaciones +Castiga errores mayores	[0,∞)	0
Coeficiente correlación anomalías	$ACC = \frac{\overline{(fc-c)(ob-c)}}{\sqrt{\overline{(fc-c)^2}\,\overline{(ob-c)^2}}}$	Correspondencia o diferencia de fase con respecto a la climatología	[-1,1]	1

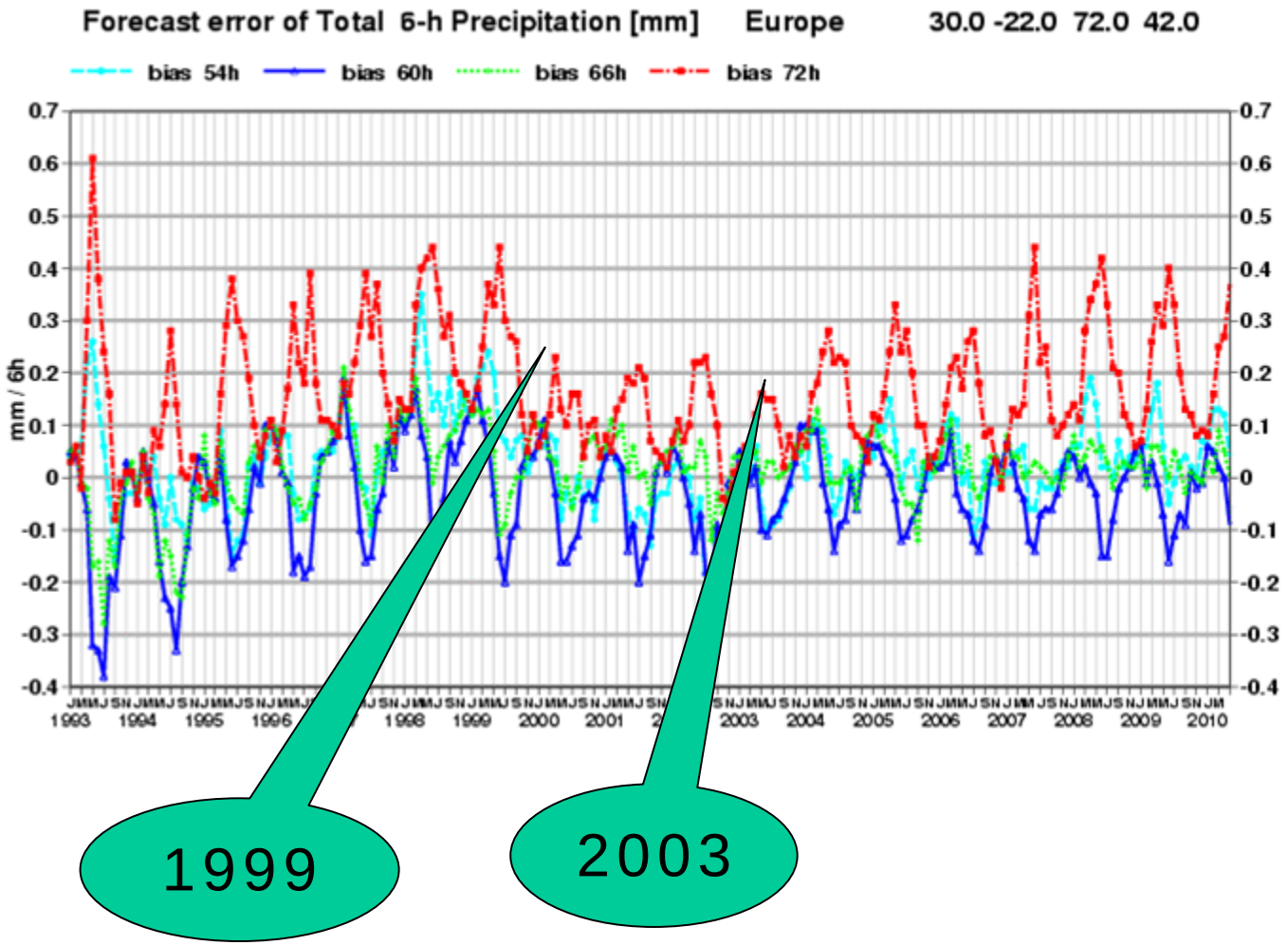
Scores: spatial distribution

E.g.: bias



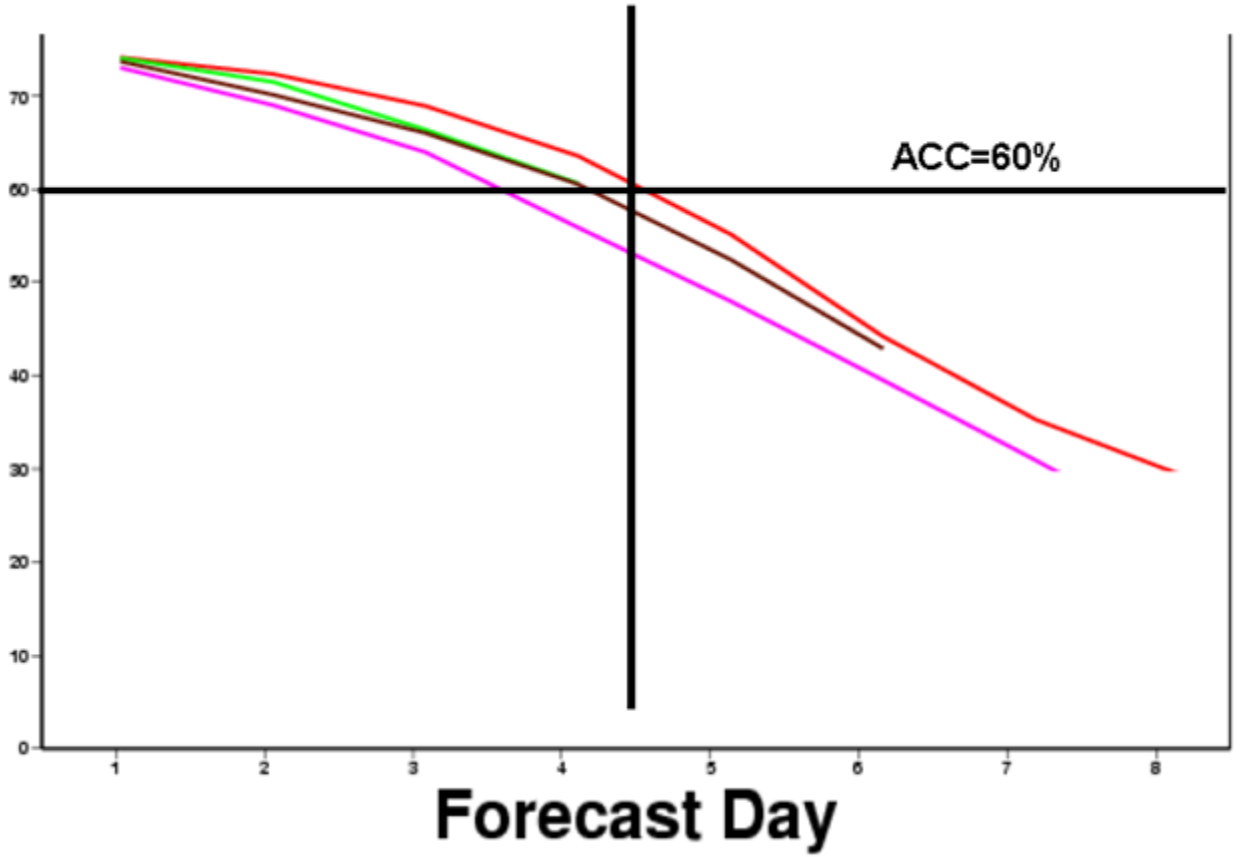
Scores: time series

- Long term trends
- 6h acc pcp
- Positive bias in summer (over fc pcp) for 12UTC run
- Impacts to notice:
  - 1999: cloud scheme
  - 2000: T511
  - 2003: convection
  - 2006: T799
  - 2010: T1279



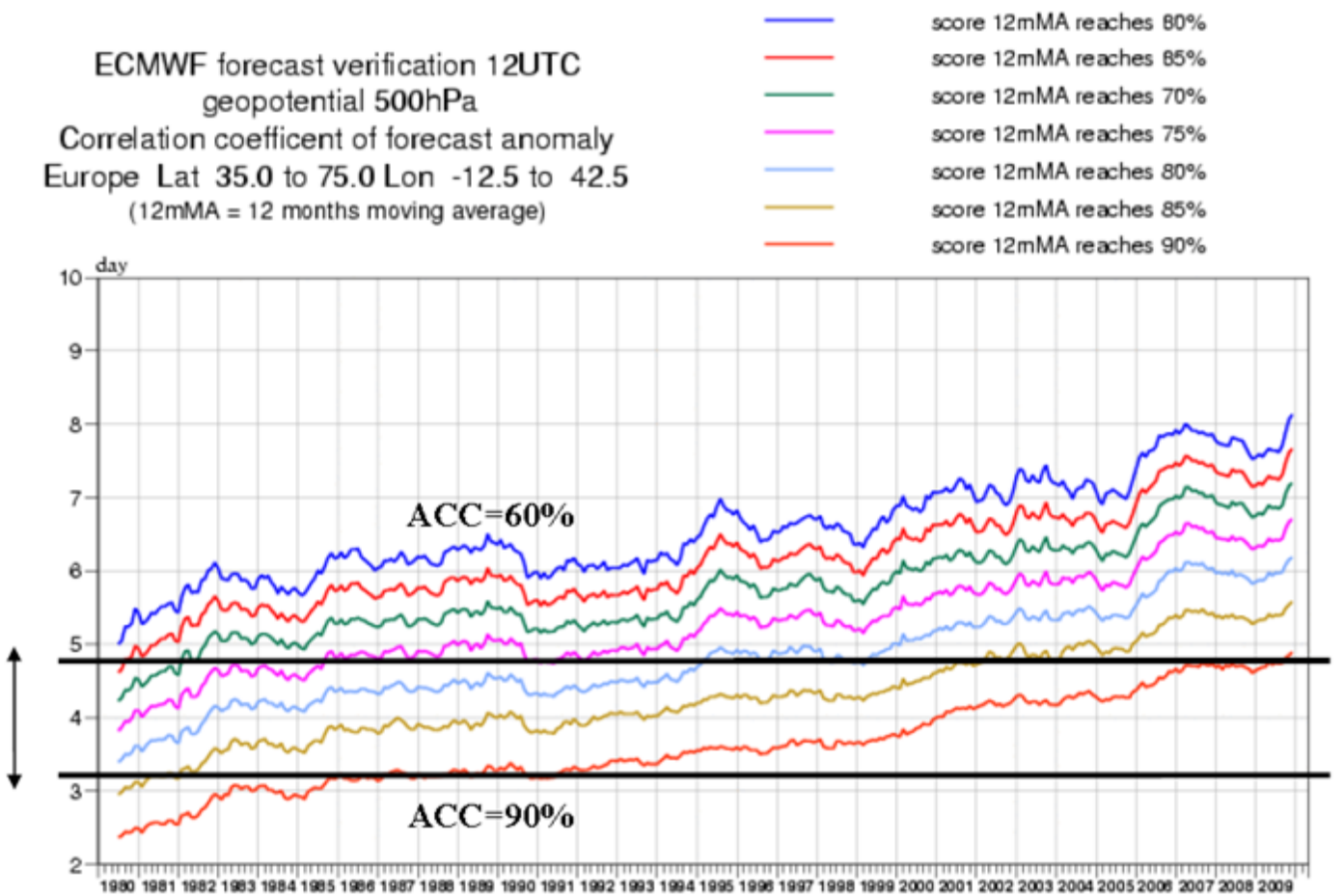
Scores: predictability limit

- "Predictability limit" ~ "validity" of forecasts
- Arbitrarily selected ACC=60%
- ACC decrease with forecast step and crosses 60% line on a certain "validity day" (e.g. D+4.5)



Scores: predictability limit

ACC will cross 60% line in different validity days (D+5, D+7…) as forecasts improve with years



Variables dicotómicas: llueve o no

		ob		
		1	0	
fc	1	82 Aciertos	38 falsas alarmas	120
	0	23 Fallos	222 negativos correctos	245
		105	260	365

Los tornados de Finley

		ob	
		1	0
fc	1	a	b
	0	c	d

$$PC = \frac{a + d}{N}$$

		ob	
		1	0
fc	1	1	6
	0	0	358

$$PC_{SPE} \approx 0.984$$

- Pensemos en la predicción de tornados (“evento raro”). El famoso Sistema de Predicción Excelente (SPE) puede obtener PC=0.984.

El camelo de Finley

		ob	
		1	0
fc	1	a	b
	0	c	d

$$PC = \frac{a \square d}{N}$$

		ob	
		1	0
fc	1	1	6
	0	0	358

$$PC_{SPE} \approx 0.984$$

		ob	
		1	0
fc	1	0	0
	0	1	364

$$PC_{SPS} \approx 0.997$$

- Pensemos en la predicción de tornados (“evento raro”). El famoso Sistema de Predicción Excelente (SPE) puede obtener PC=0.984. A su vez, el también famoso Sistema de Predicción del Sinvergüenza (SPS) obtiene un PC aún mejor con esfuerzo mínimo (no prediciendo nunca tornado)
- El PC puede estar engordado de forma espuria, no es un score realmente representativo de “precisión” ☹, tiene demasiado peso la “d”
- Hacen falta otros scores

Variables dicotómicas: scores

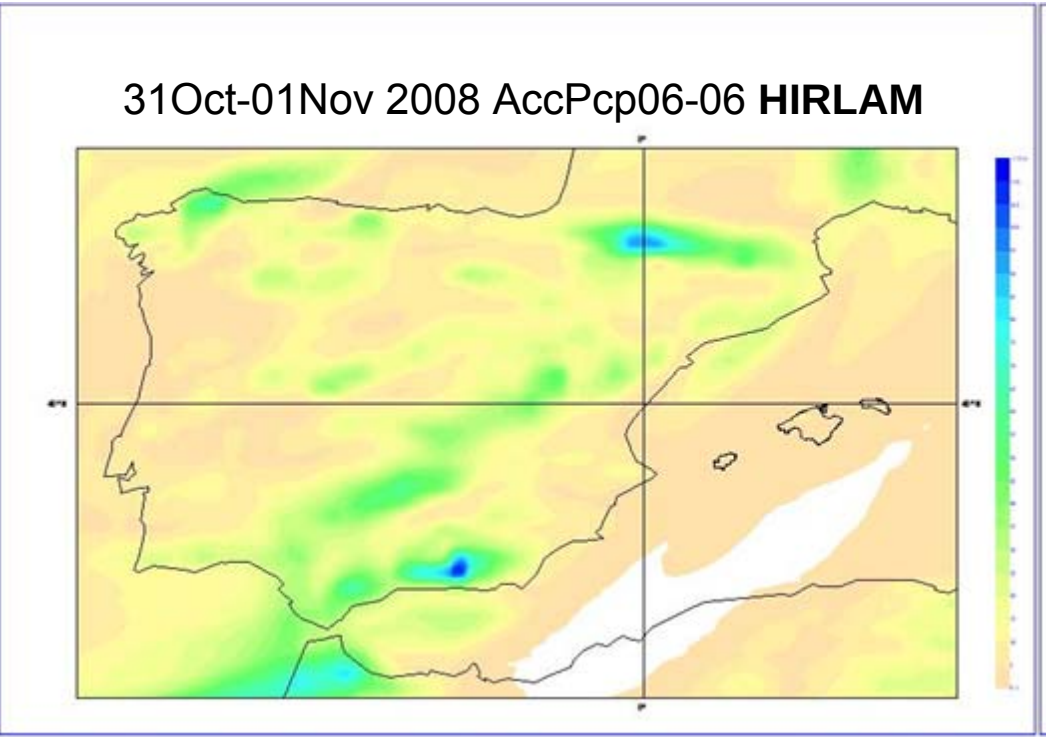
		ob	
		1	0
fc	1	**a**	**b**
	0	**c**	**d**

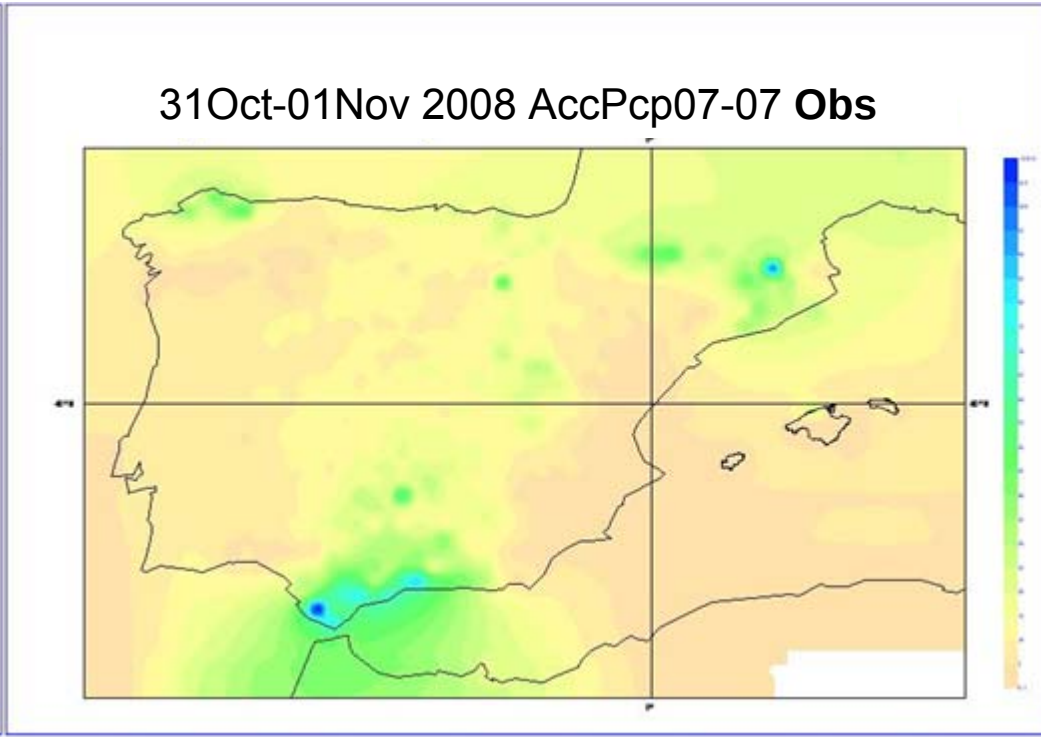
Score	Cómputo	Significado	Rango	Perf
Percentage correct	$PC = \frac{a+d}{a+b+c+d}$	No es realmente indicativo	[0,1]	1
Hit rate	$H = \frac{a}{a+c}$	SIs: se predice el evento y se da	[0,1]	1
False alarm rate	$F = \frac{b}{b+d}$	NOs: se predice el evento pero no se da	[0,1]	0
True skill score	$TSS = \frac{a}{a+c} - \frac{b}{b+d}$	Mide la habilidad para separar Sis y Nos, compara H y F	[-1,1]	1
Frequency bias index	$FBI = \frac{a+b}{a+c}$	Mide la proporción entre evento predicho y evento observado	$[0,\infty)$	1

3

- Are the models perfect?
- Forecast verification: an introduction
- Classical methods
- **A critical vision**
- New spatial verification methods
- SAL: an example of feature-oriented method
- MODE: a method applicable to EPSs
- Radar & satellite data
- Software, conclusions, references

Subjective and Diagnostic verification

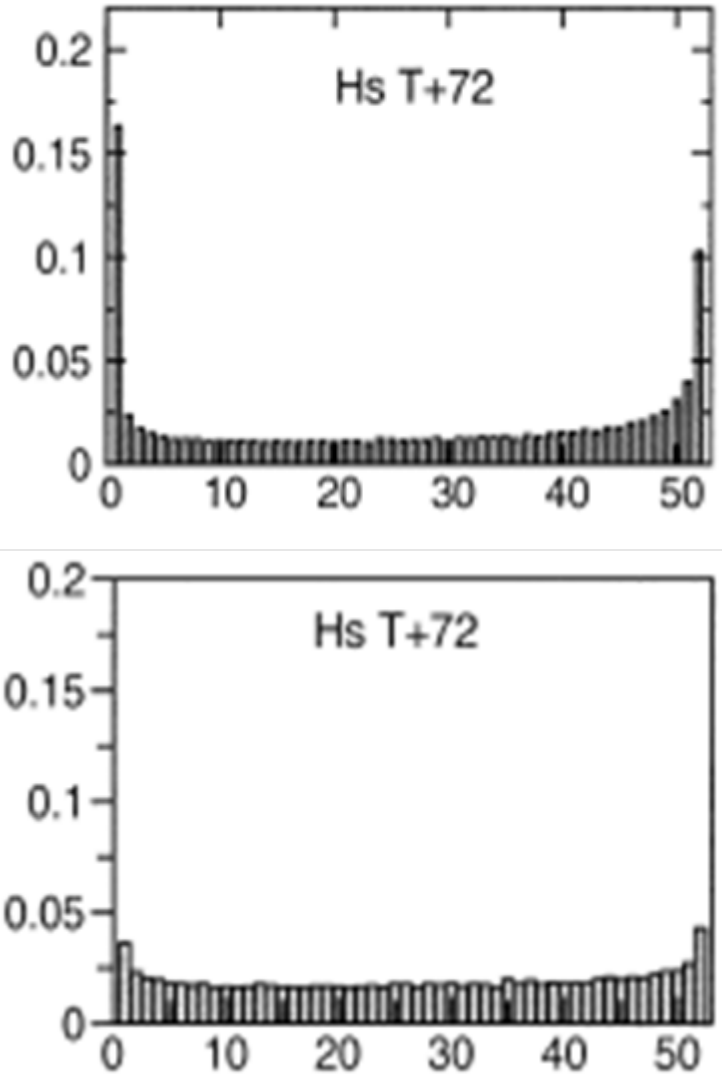




- How did the model captured patterns / structures?
- What range of scales was the forecast better / worse?
- Intensity and / or location of events?
- “Eye ball”

Error in observations

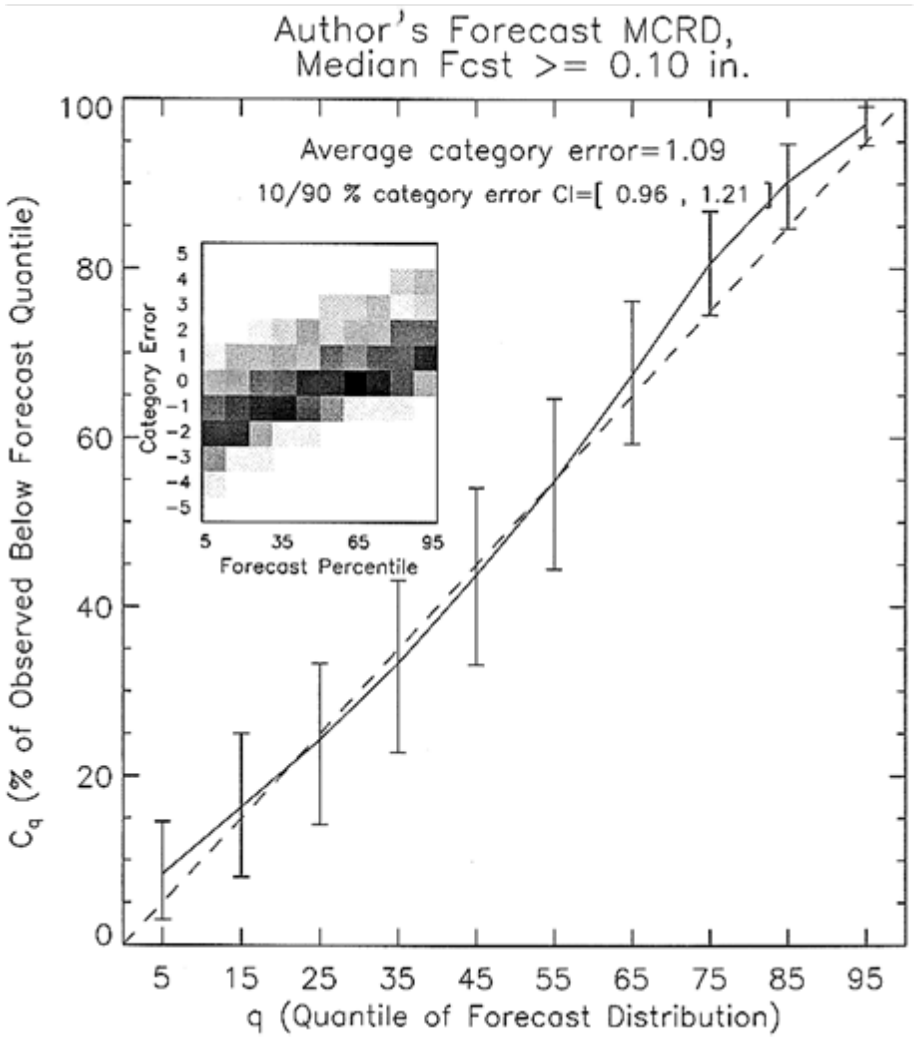
- Classical assumption: ?
  - observation error << model error
- What's happening?
  - Verification measures can give wrong (or at least poor) ideas
- Trends:
  - Estimate of observational error
  - Include obs err as gaussian noise or by sampling
  - Surprising results



Courtesy ECMWF

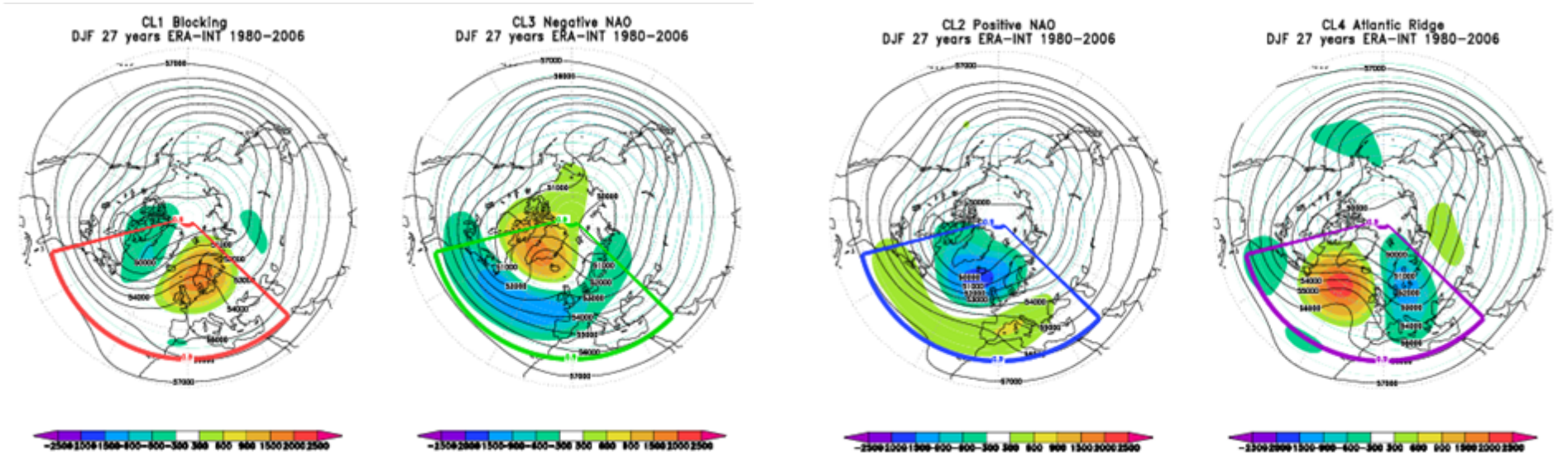
Sampling Error

- Any measure must be given with its associated uncertainty
- Verification measures are statistical estimates of the truth value (population), computed with a finite number of pairs (fc,ob)
- Without sampling error, conclusions should not be made, e.g. model “A” is better than model “B”
- Trends: error bars, confidence intervals, bootstrap



Courtesy T.M. Hamill

Flow dependent Verification



Courtesy L.Ferranti

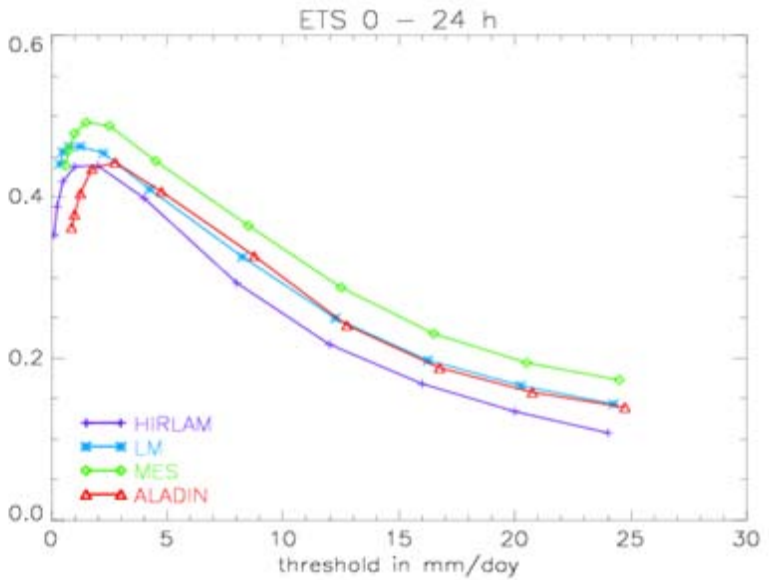
- Z h#p xvw#p suryh#wkh#wudglwrqdo#vwudwlilfdwrq#e|#vhdvrqv
- Wuhqgv=#Fxvwhulqj#ri#iorz #uhjlp hv/#dqg#fruuhvsrqglqj# yhulilfdwrq

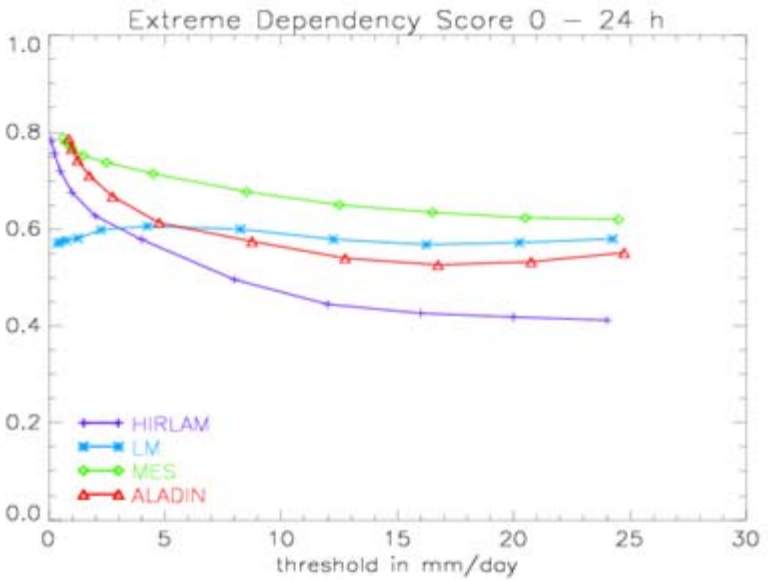
Extreme and severe weather verif

- Extreme ≠ Severe
- Low frequency → under-sampling makes traditional scores unuseful
- Equitable Threat Score (ETS) → Extreme Dependency Score (EDS)
- **No mature methods**

$$TSS = \frac{a}{a+c} - \frac{b}{b+d}$$

$$EDS = \frac{2\log(a+c/n)}{\log(a/n)} - 1 \quad p \to 0 \quad \frac{1-}{1}$$





Courtesy D.B. Stephenson

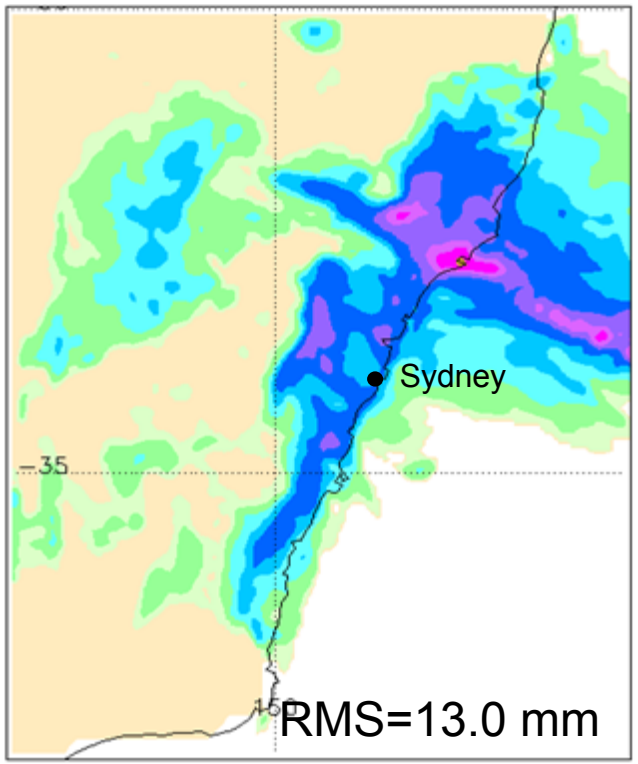
Limitations of classical verification methods

- Inconsistent with visual “eye ball”
  - It is not diagnostic (no information about reasons for errors “what was wrong”)
  - Does fc look realistic?
  - How can we improve the fc?
  - How can we take decisions with the fc?
- Observational Uncertainty/Error:
  - Specific “new” methods: Saetra&Hersbach, Candille&Talagrand “Observational Probability”
- Sampling Uncertainty/Error:
  - Can lead to false conclusions
  - Specific methods: Confidence Intervals, Bootstrap
- Severe and extreme weather
  - Severe $\neq$ Adverse
  - A forecasting system can be useful on detecting signals even without good scores
  - Distributions-oriented verification, extreme events scores e.g. EDS
- Spatial scale models and observations: insensitive to location-shape errors
  - Interpolation methods
  - Correlation
  - E.g. patterns, structures
  - E.g. double penalty can give better scores to a coarser grid model
  - “New” methods: Up-scaling, Fuzzy, Feature/Object-oriented (e.g. SAL)

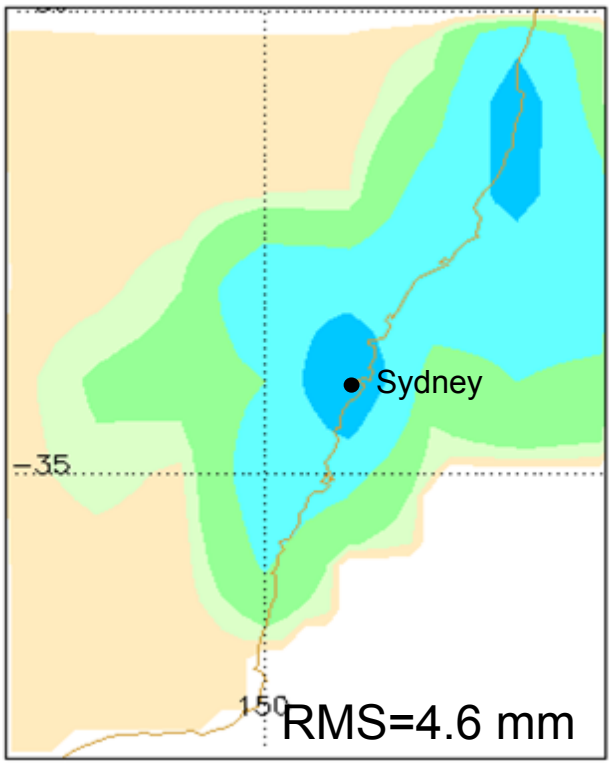
Double Penalty

Which rain forecast would you rather use?

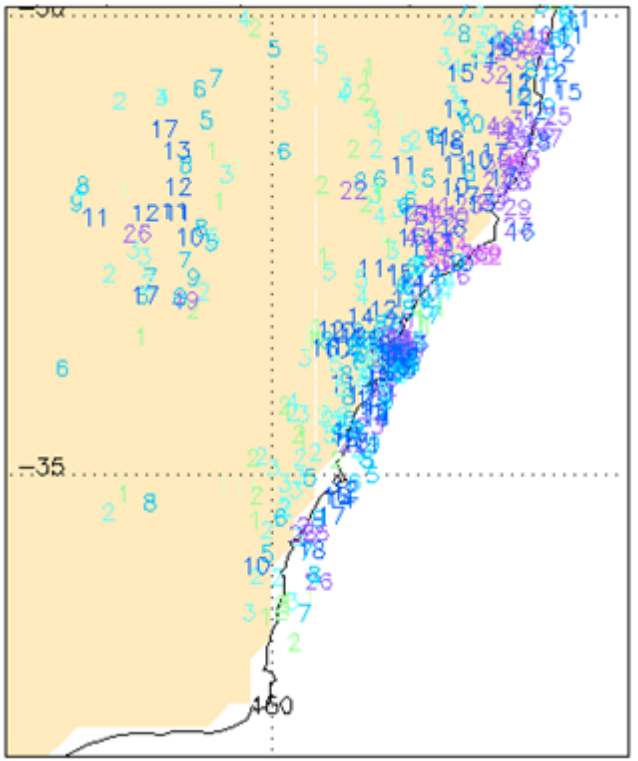
Mesoscale model (**5 km**) 21 Mar 2004



Global model (**100 km**) 21 Mar 2004



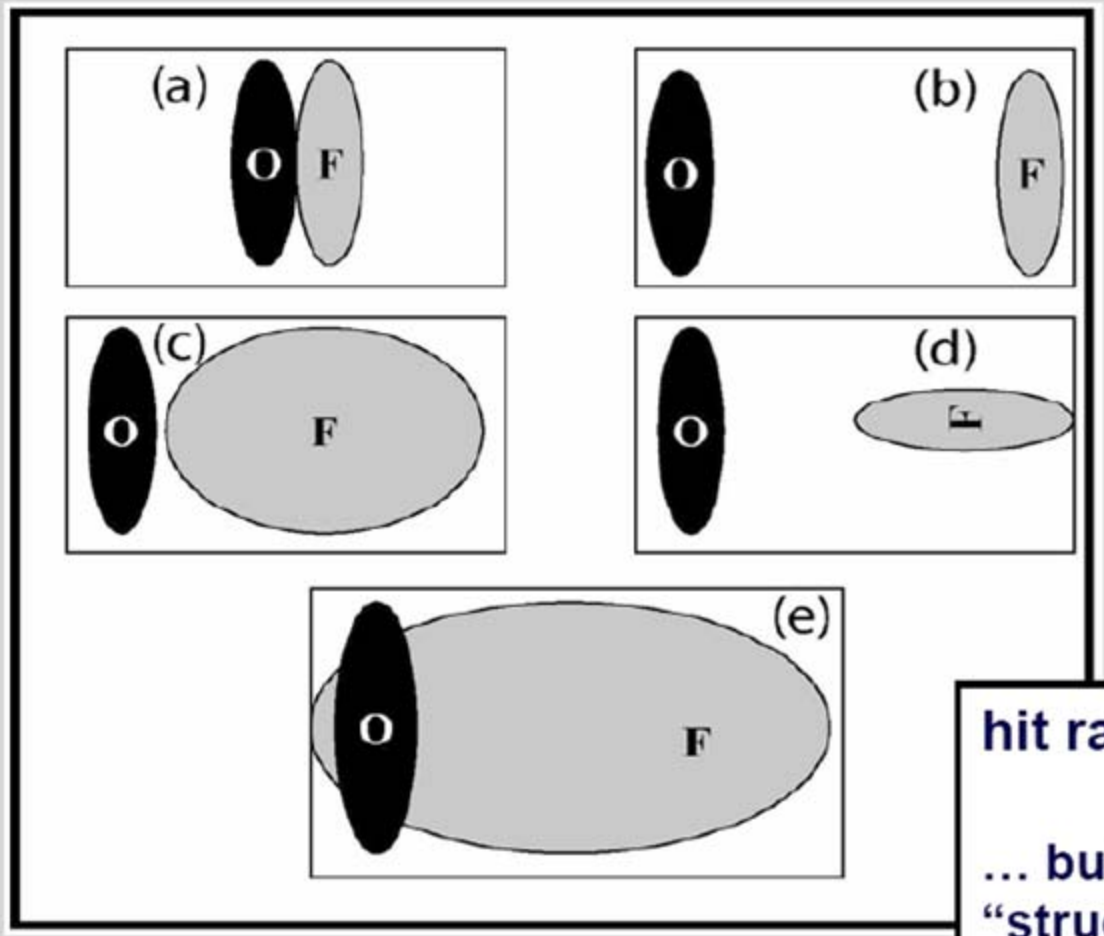
Observed 24h rain



Courtesy Beth Ebert

Double penalty

Problematic aspects of grid point based error scores



Davis et al. 2006 (MWR)

hit rate = 0

… but (a) is "better" than (b)-(d)

hit rate > 0

… but forecast has not the right "structure"

Escalas espaciales fc vs ob

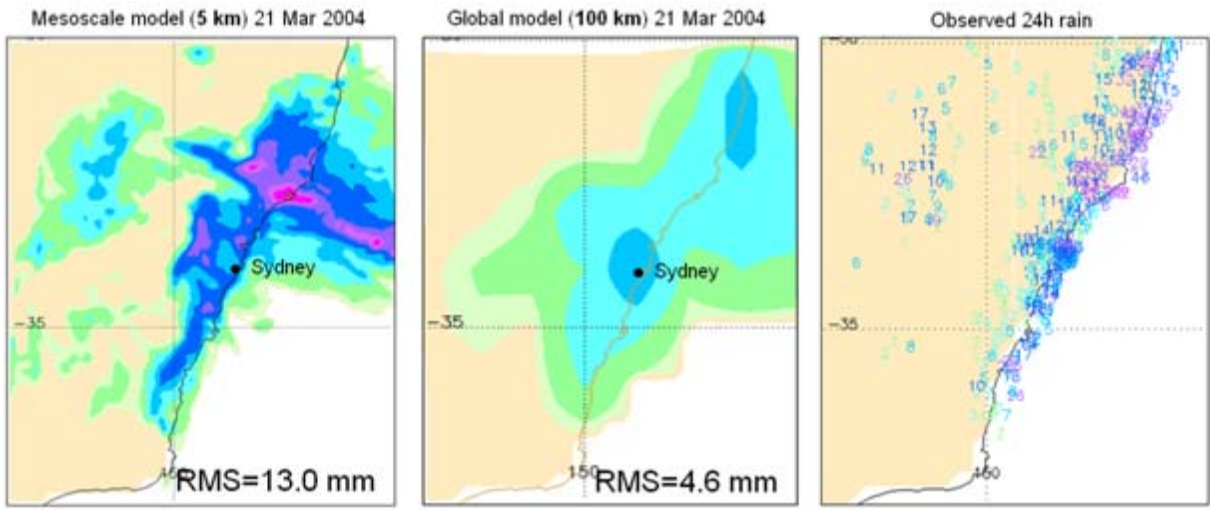
- Limitaciones de los métodos clásicos:
  - Son “reduccionistas” en cierto sentido: trabajan con conjuntos pares (fc,ob), suponiendo que sus medidas estadísticas pueden representar ciertas propiedades.
  - Métodos de interpolación, correlación, etc
  - E.g.: patrones, estructuras, problema del **“double penalty”**: los modelos son más realistas pero son penalizados
  - No es verificación diagnóstica, necesitamos valorar la utilidad de los modelos

- Se requiere nuevos métodos y una buena cobertura de observaciones
  - Redes climatológicas de alta densidad se quedan cortas <= 5km
  - **Radar y satélite**

- Software especializado:
  - Scripting
  - Plug-ins
  - OOP
  - Database, geografía
  - Formatos, volumen

NCEP op:

- 381km(1950) to 12km(2002)



Courtesy Beth Ebert

What are we looking for?

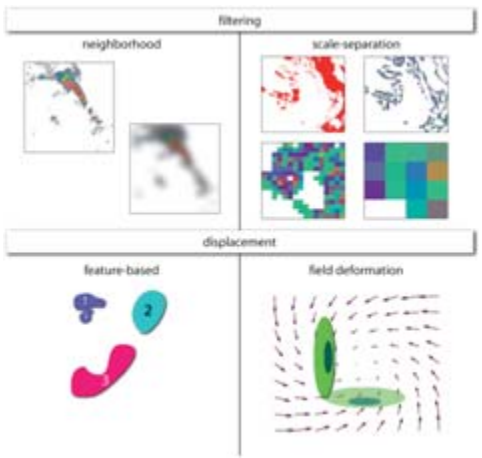
Diagnostic verification

- Explicit, detailed and quantitative information about failures, useful to improve model, able to measure quality fairly
- E.g. SAL: identify precipitation areas in the forecast pcp field, the observed pcp field, and compare giving: location and phase, amplitude, structure, etc.

Evaluate the **spatial scales** in which the model is “skillful”

- **Avoid double-penalty**
- Nowadys, this skill is not at grid-scale, but a bit coarser

Objective and subjective evaluations will become closer

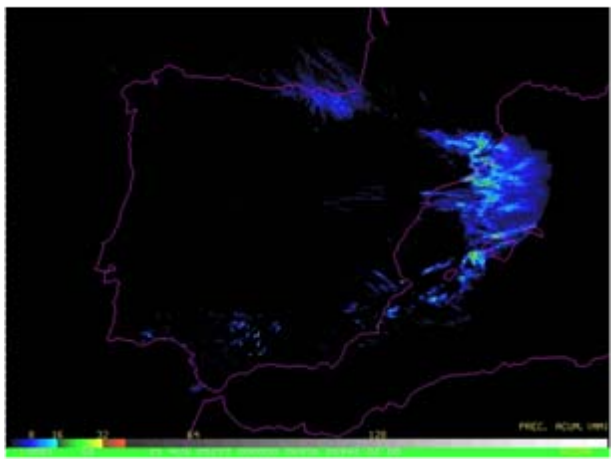


What do we need?

New methods

HR Observations

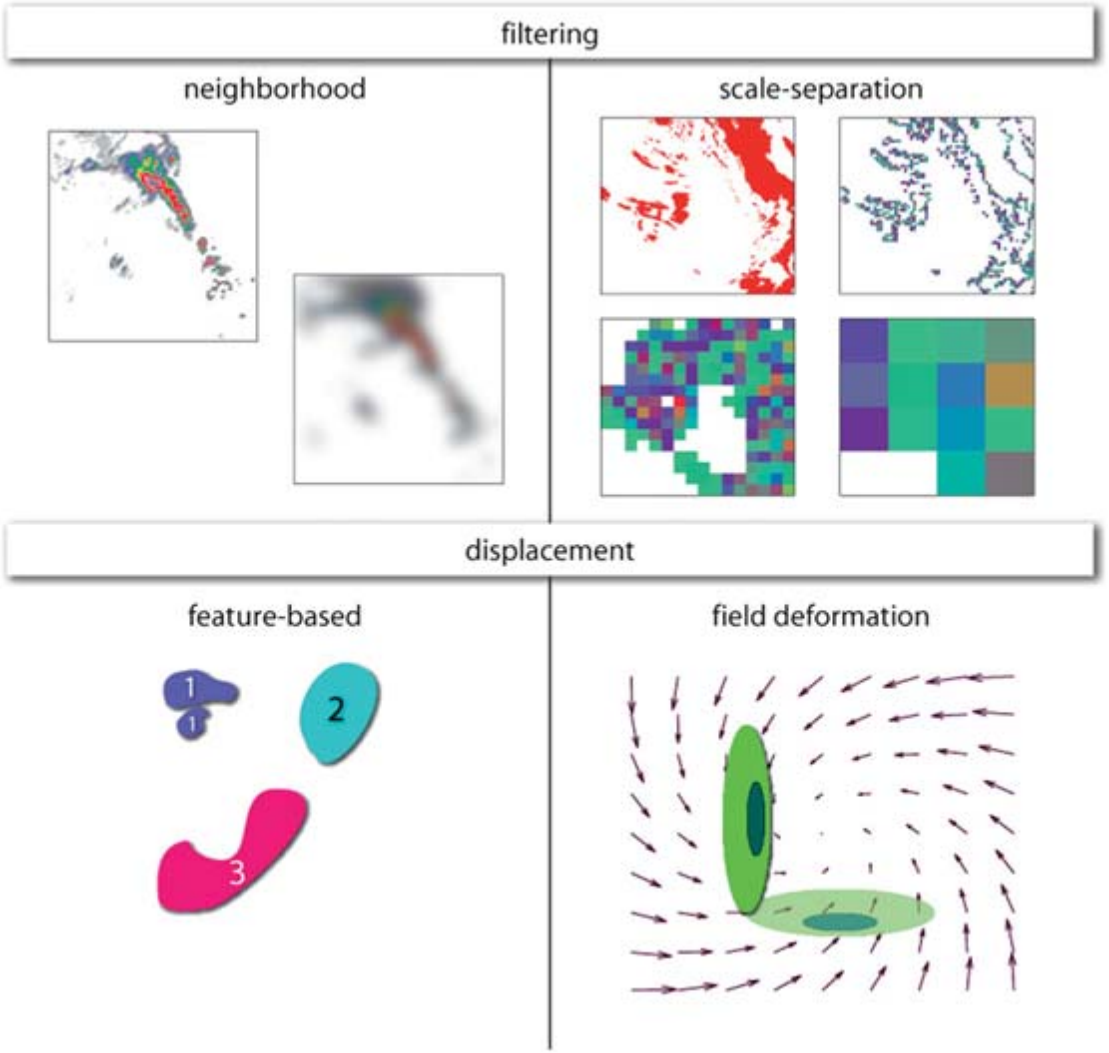
- Radar
- Satellite



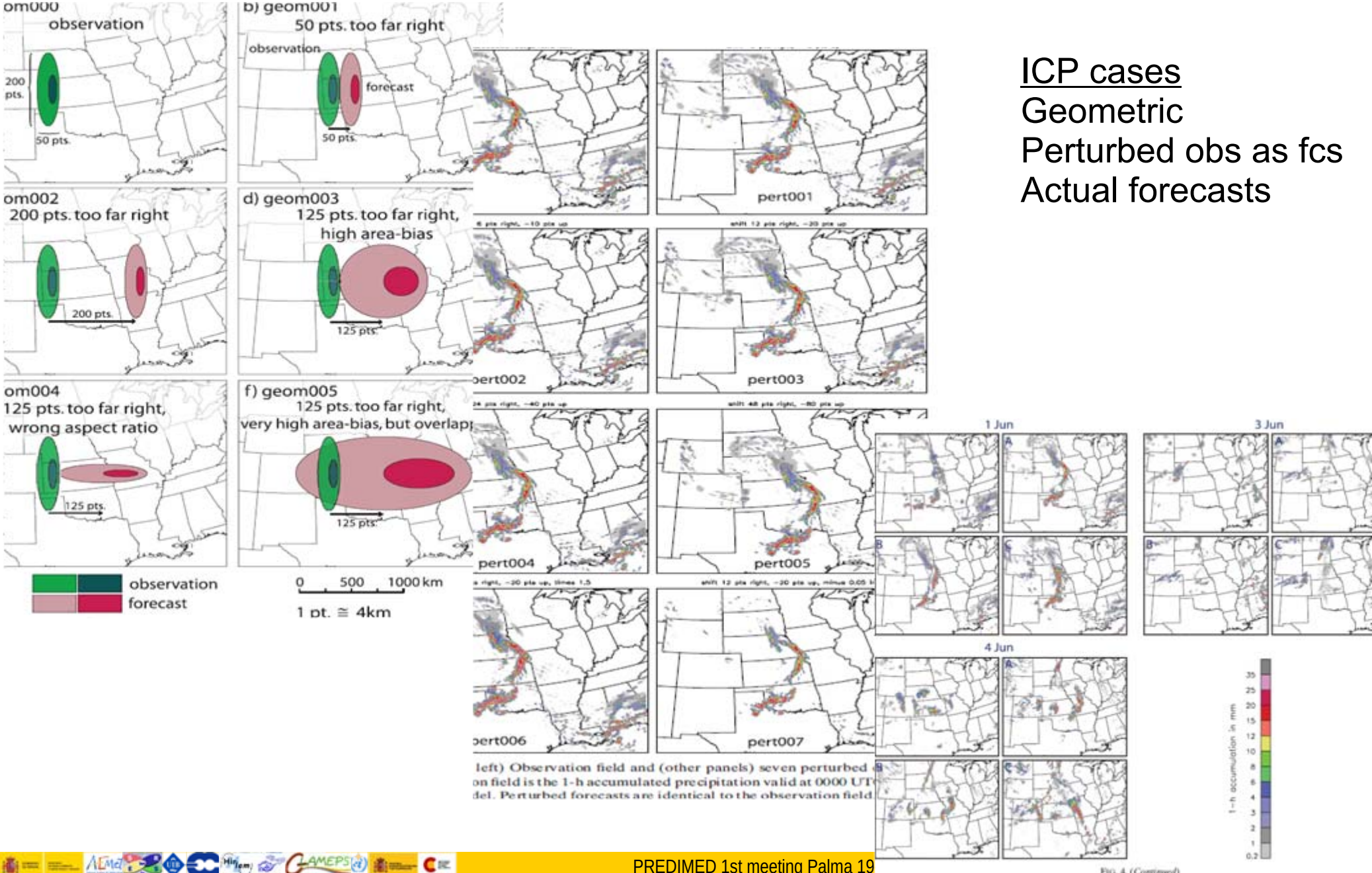
4

- Are the models perfect?
- Forecast verification: an introduction
- Classical methods
- A critical vision
- **New spatial verification methods**
- SAL: an example of feature-oriented method
- MODE: a method applicable to EPSs
- Radar & satellite data
- Software, conclusions, references

Intercomparison Of Spatial Fc Verif Methods Project



- How do the methods inform about performance at different scales?
- How do the methods provide information on location errors?
- Do the methods provide information on intensity errors and distributions?
- Do the methods provide information on structure errors?
- Do the approaches have the ability to provide information about hits, misses, false alarms, and correct negatives?
- Do the methods do anything that is counterintuitive?
- Do the methods have selectable parameters and how sensitive are the results to parameter choice?
- Can the results be easily aggregated across multiple cases?
- Can the methods be used to identify timing errors?
- Can confidence intervals or hypothesis tests be readily computed for the method?



ICP cases

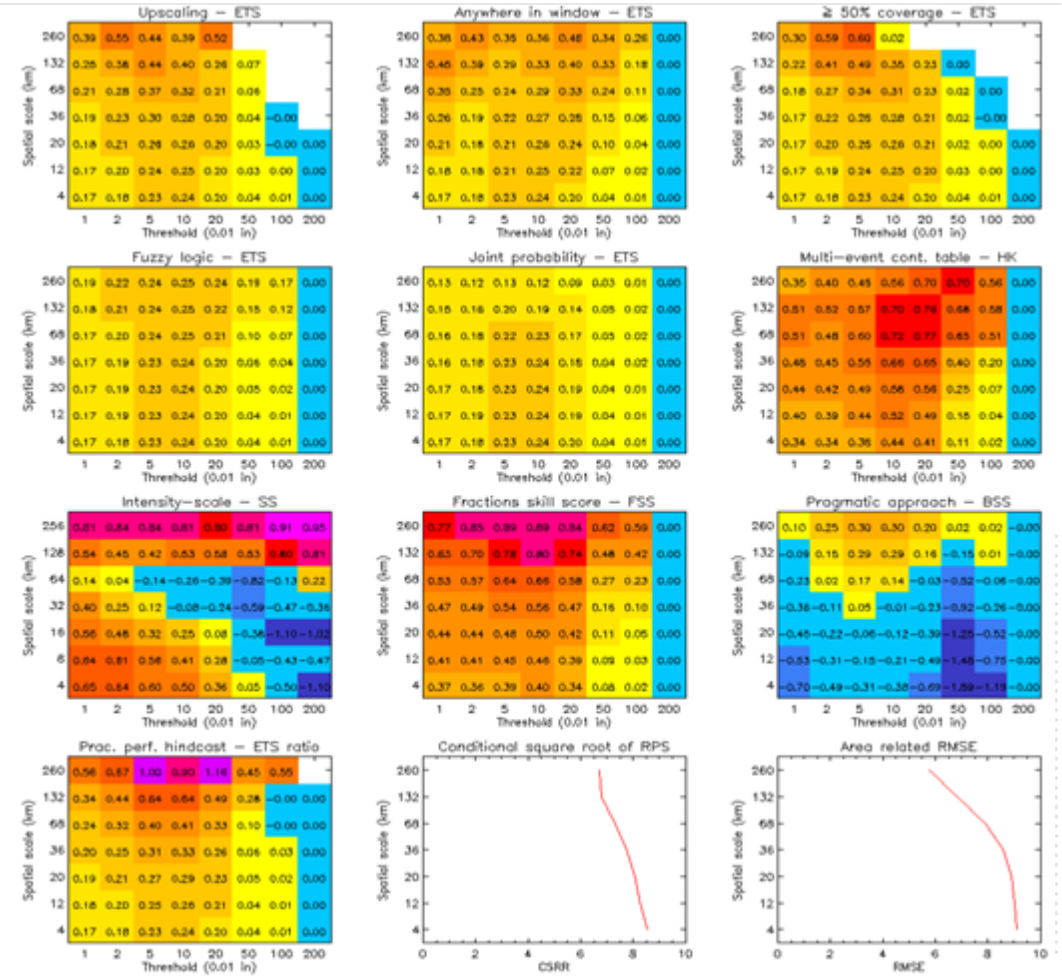
Geometric

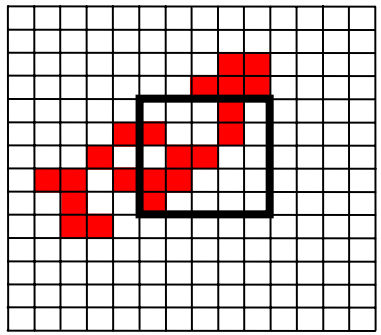
Perturbed obs as fcs

Actual forecasts

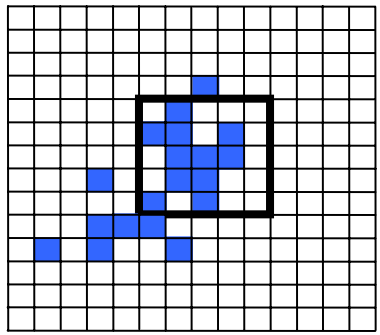
Method type						
Traditional	CETS	No*	Indirectly	Yes	No	Yes
Features-based	CA	Yes (see text)	Yes (see text)	No*	No	Yes
Features-based	Procrustes	No*	Yes	Yes	Yes	Yes (see text)
Features-based	Procrustes2	Yes	Yes	Yes	Yes	Yes
Neighborhood	Composite	No*	Yes (see text)	Average intensities	Yes	Yes
Field-deformation	CRA	No*	Yes	Yes	Yes	Yes
Field-deformation	DIST	Yes (see text)	Indirectly	Yes	No	Yes
Neighborhood	GQI	No*	Yes	Yes	No	No
Scale-separation	GQM–DAS	No*	Yes	Yes	Yes	Yes (see text)
Field-deformation	GSS	Yes (see text)	Indirectly	Yes	No	Indirectly
Features-based	IS	Yes (see text)	Indirectly	Yes	No	Indirectly
Scale / Features	MW	Yes (see text)	Yes	Yes	No*	Yes (see text)
Features-based	MSV	Yes (see text)	Indirectly	Yes	No	No
Point	MODE	Yes (see text)	Yes	Yes	Yes	Yes (see text)
Scale-separation	Neighborhood	Yes (see text)	Indirectly	Yes	No	Yes

Neighborhood





observation



forecast

good performance

poor performance

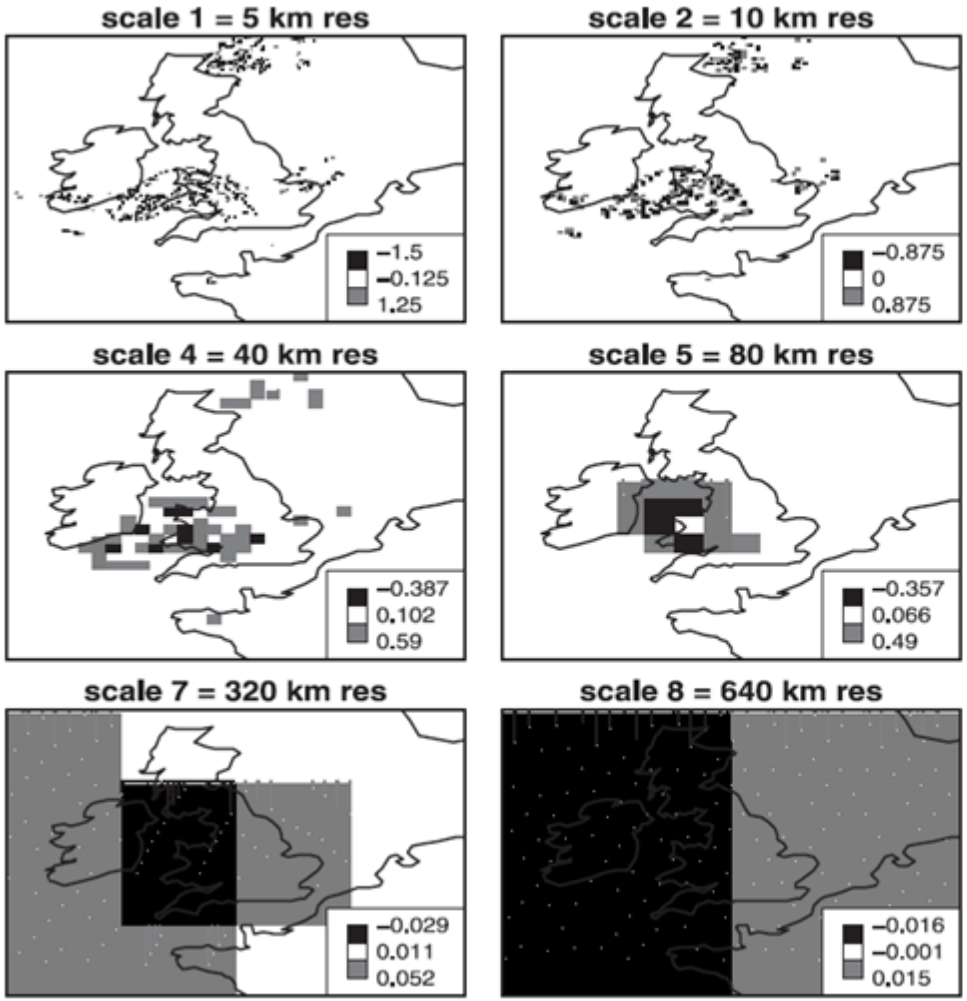
e.g

Ebert 2009: Up-scaling, FSS, etc.

Mittermaier 2008: FSS

Scale-separation

e.g
Casati 2009:
Intensity-scale



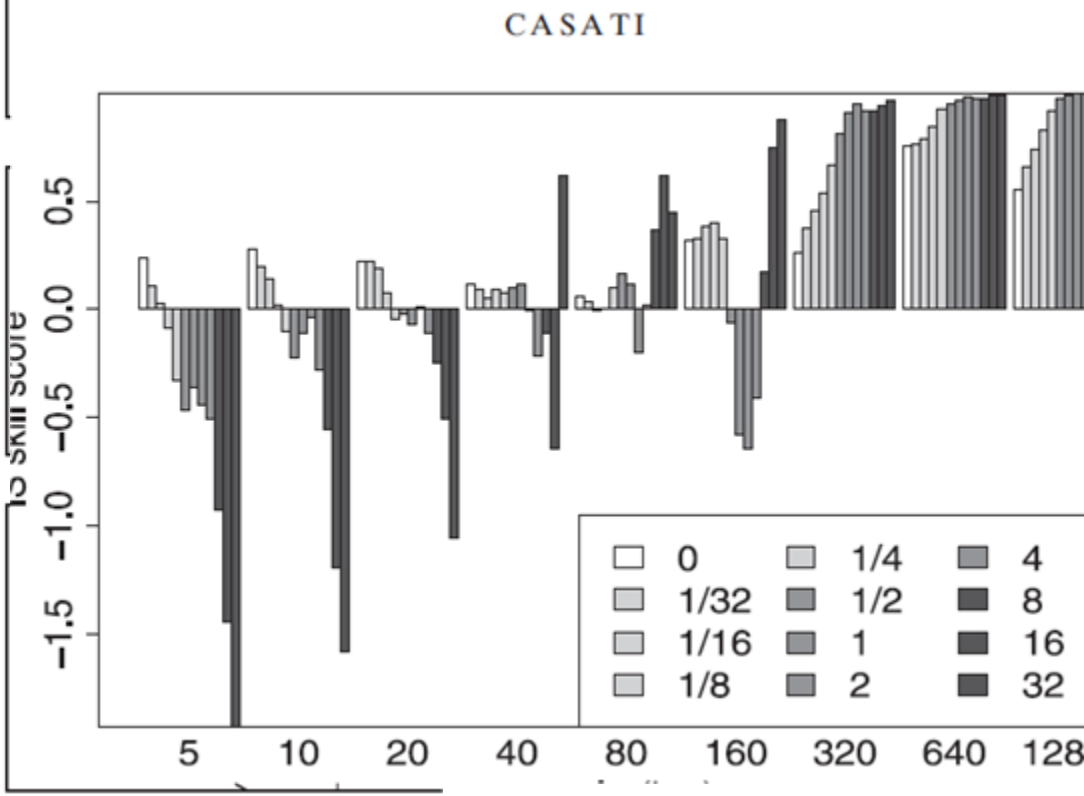
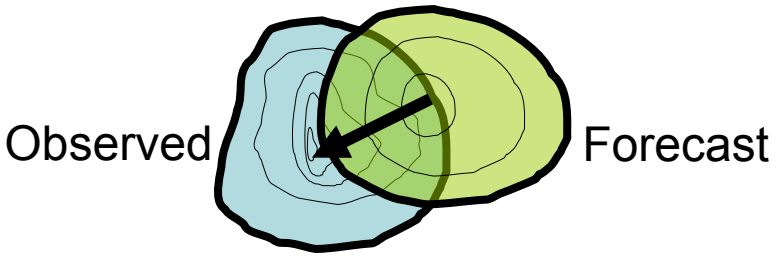


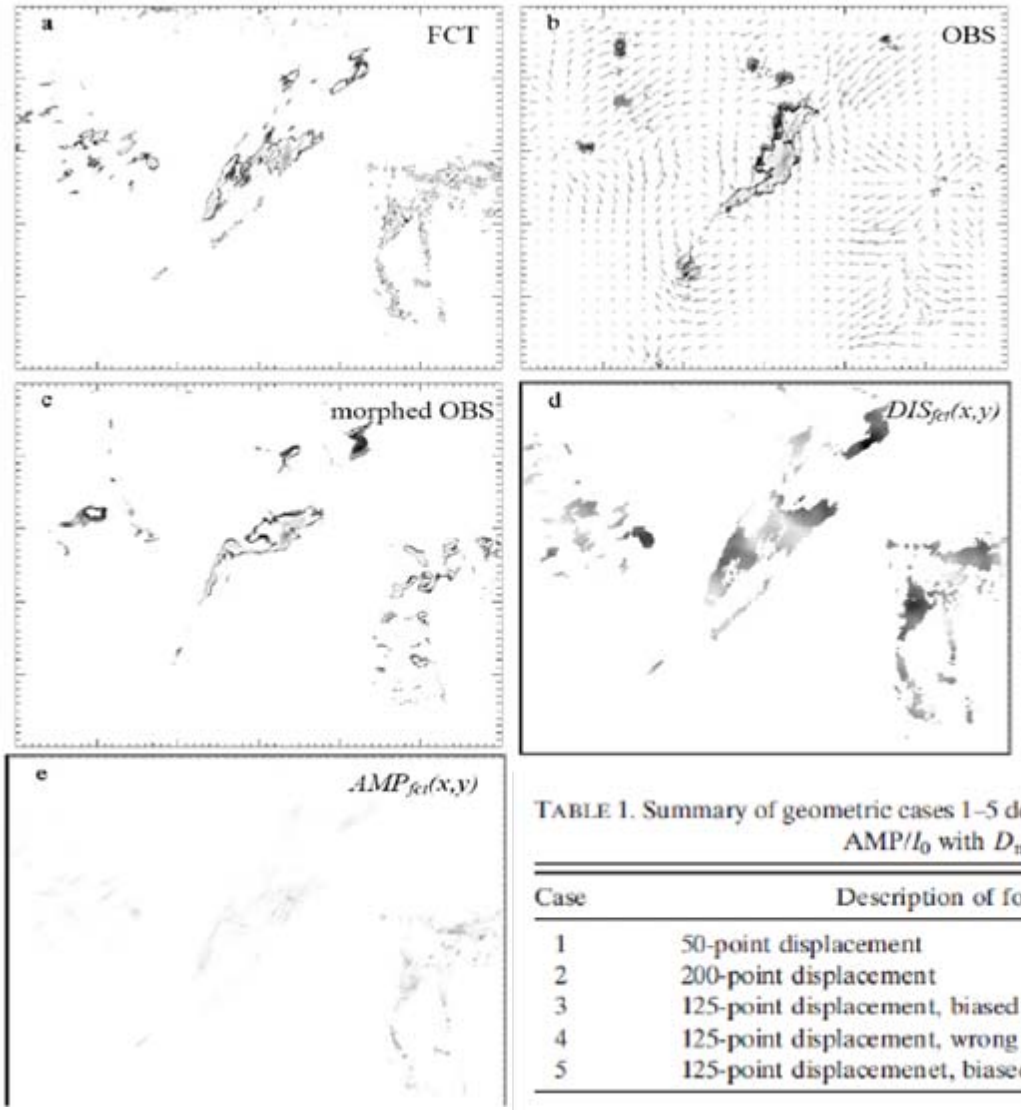
FIG. 3. Wavelet scale components of the binary field difference for the NIMROD case study, for a threshold of 1 mm h^{-1}

Feature-based



- Define *entities* using a threshold
- Horizontally translate the forecast until a *pattern matching* criterion is met
- Displacement is the vector difference between the original and final locations of the forecast
- Compare properties of whole objects
- Extensible to ensemble
- E.g.: Ebert 2009: CRA, Davis 2009: MODE (both extensible to ensemble), Wernli 2008: SAL (not matching but field as a whole)

Field-deformation



e.g
Keil and Craig 2009:
Optical Flow

TABLE 1. Summary of geometric cases 1–5 depicting a brief description, the DAS, normalized DIS and AMP values (i.e., DIS/D_{max} and AMP/I_0 with D_{max} = 360 km and I_0 = 15.4 mm), and the corresponding rank.

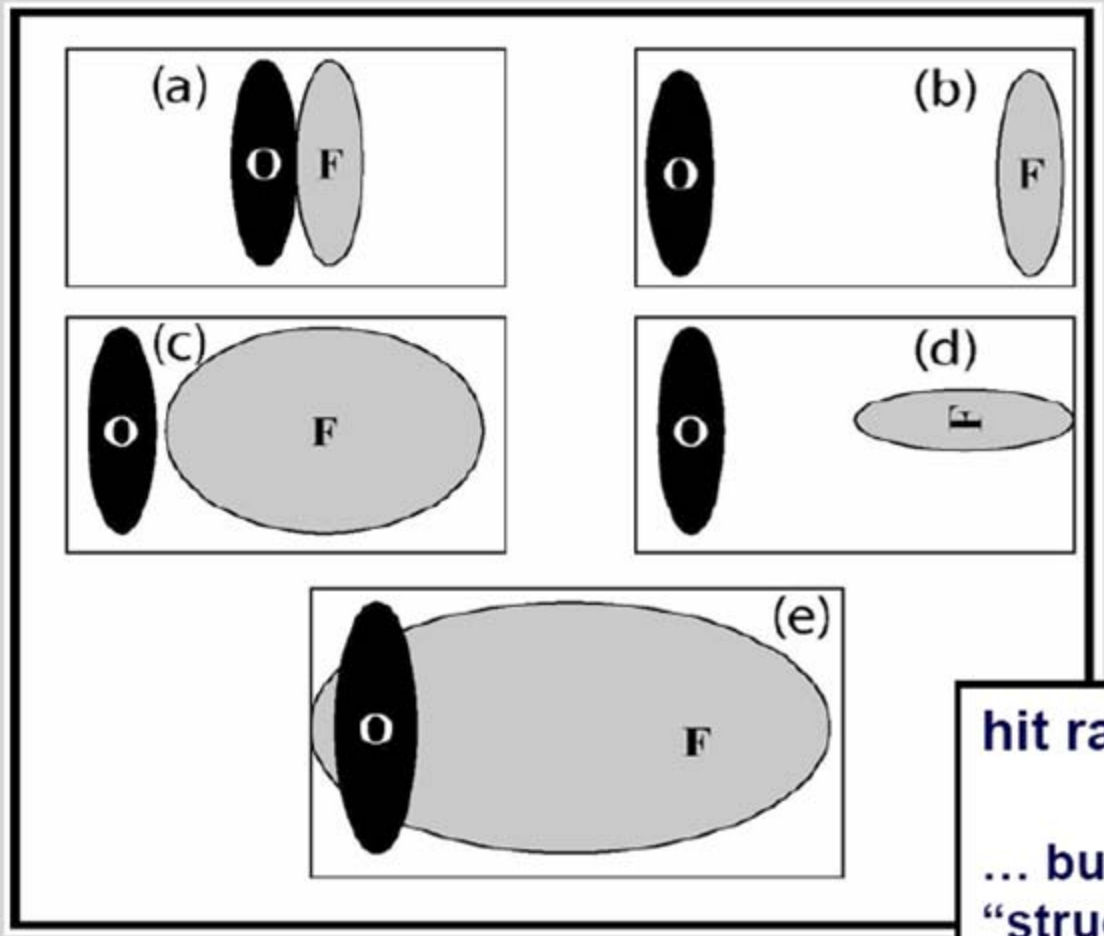
Case	Description of forecast feature	DAS	DIS/D_{max}	AMP/I_0	Rank
1	50-point displacement	0.62	0.55	0.07	1
2	200-point displacement	1.00	0.00	1.00	2
3	125-point displacement, biased high	1.11	0.21	0.91	5
4	125-point displacement, wrong aspect ratio	1.09	0.22	0.87	4
5	125-point displacemenet, biased very high, but overlapping	1.02	0.19	0.83	3

5

- Are the models perfect?
- Forecast verification: an introduction
- Classical methods
- A critical vision
- New spatial verification methods
- **SAL: an example of feature-oriented method**
- MODE: a method applicable to EPSs
- Radar & satellite data
- Software, conclusions, references

SAL

Problematic aspects of grid point based error scores



Davis et al. 2006 (MWR)

hit rate = 0

… but (a) is "better" than (b)-(d)

hit rate > 0

… but forecast has not the right "structure"

SAL

- Classical problem of double penalty
- Feature-oriented → ~ subjective verification
- E.g: SAL measure
  - S (Structure)
  - A (Amplitude)
  - L (Location)
- Perfect forecast: S = A = L = 0
- S requires patterns/objects definition, currently simple algorithms, need improvement

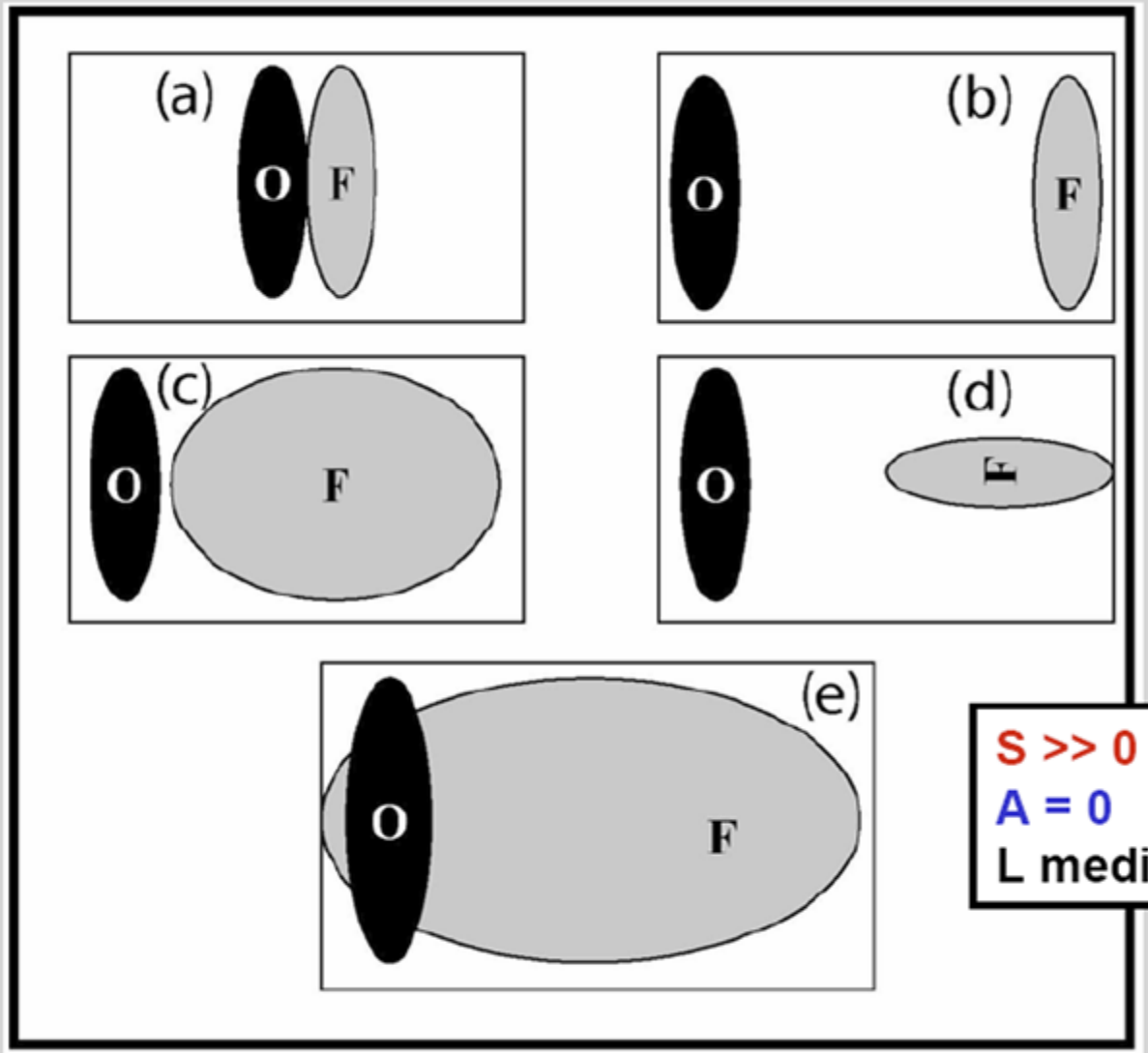
SAL

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

SAL

S = 0
A = 0
L small

S > 0
A = 0
L medium



S = 0
A = 0
L large

S = 0
A = 0
L large

S >> 0
A = 0
L medium

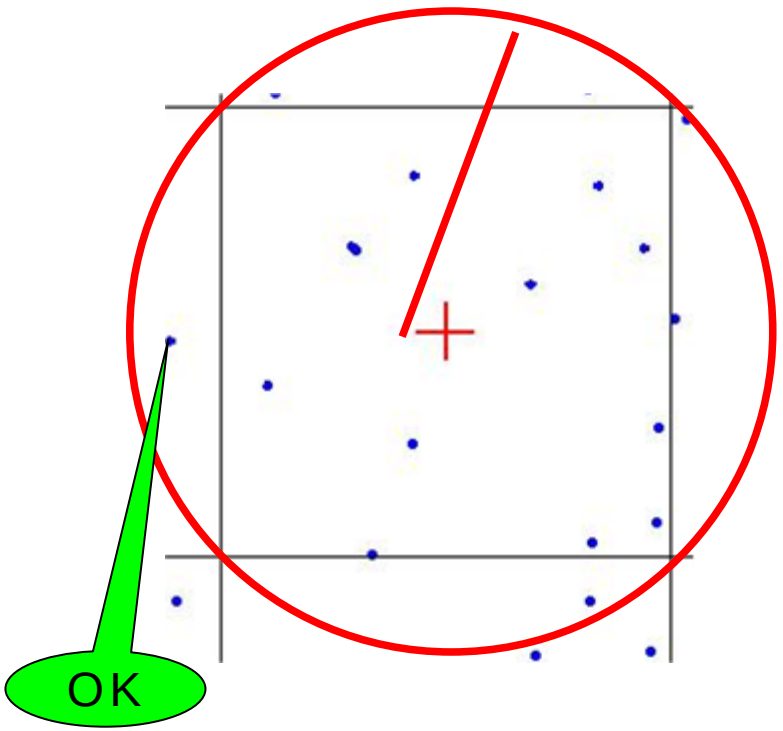
Davis et al. 2006

Up-scaling

What is the **truth**?

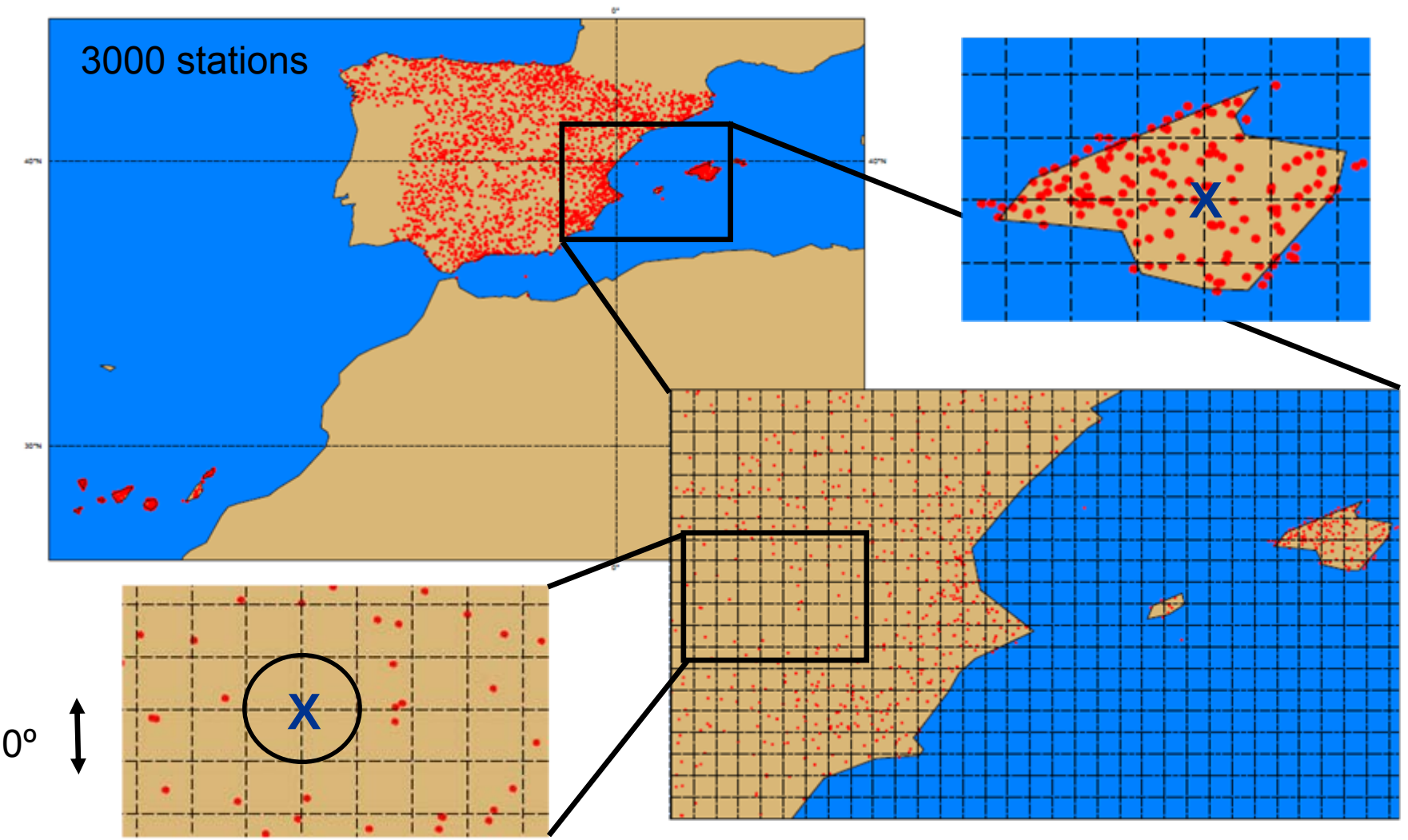
Up-scaling Europe HR obs available at ECMWF: a first simple approach:

- For each grid point consider d
- obs r < d → ob considered
- $R = \sum r^{-\alpha} R_i / \sum r^{-\alpha}$ with e.g. $\alpha = 2$
- Overcome missing data at most resolutions

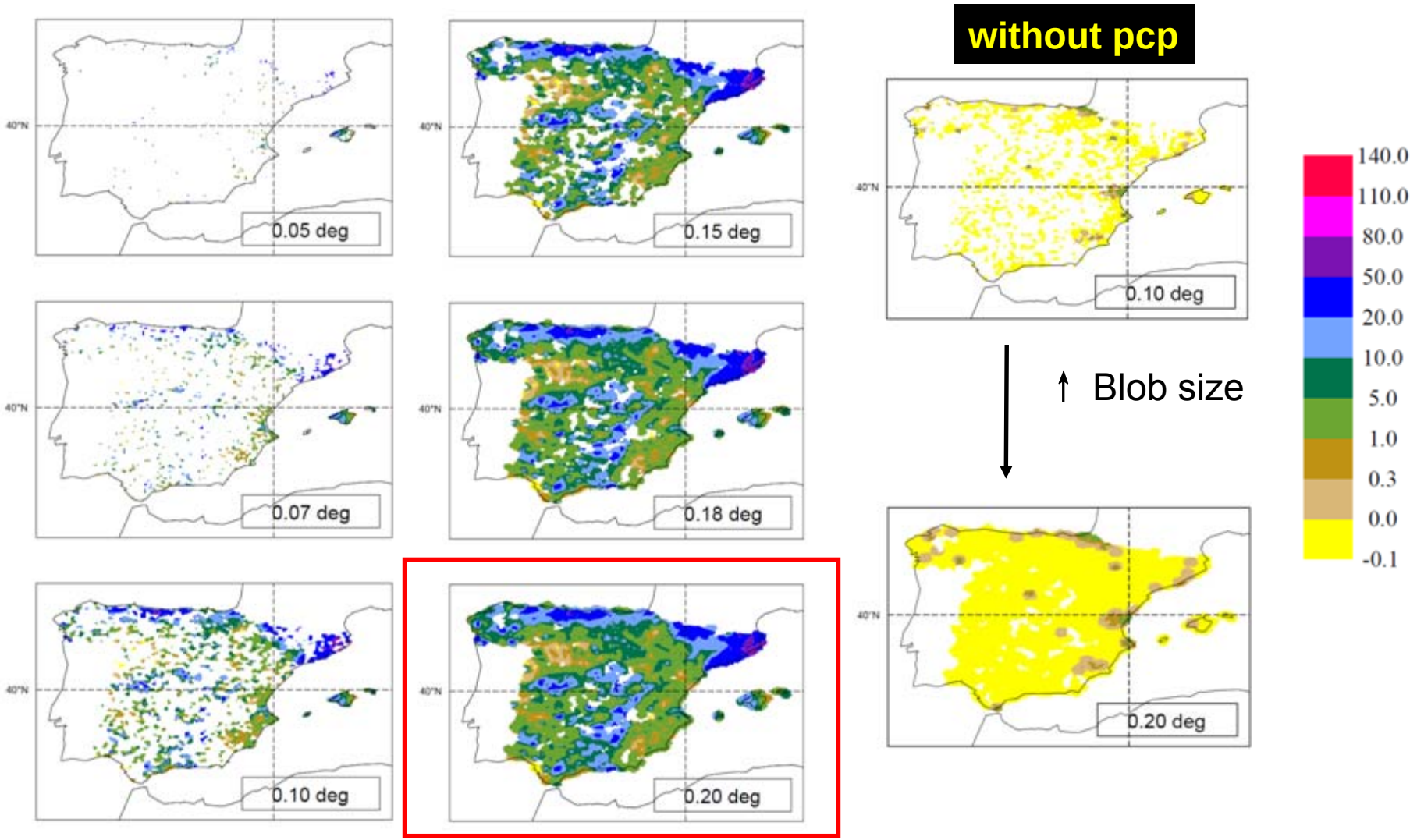


- In this work
  - Each model is compared with its own "natural up-scaling"
  - e.g. T799 with up-scaling 0.25
  - e.g. T399 with up-scaling 0.50

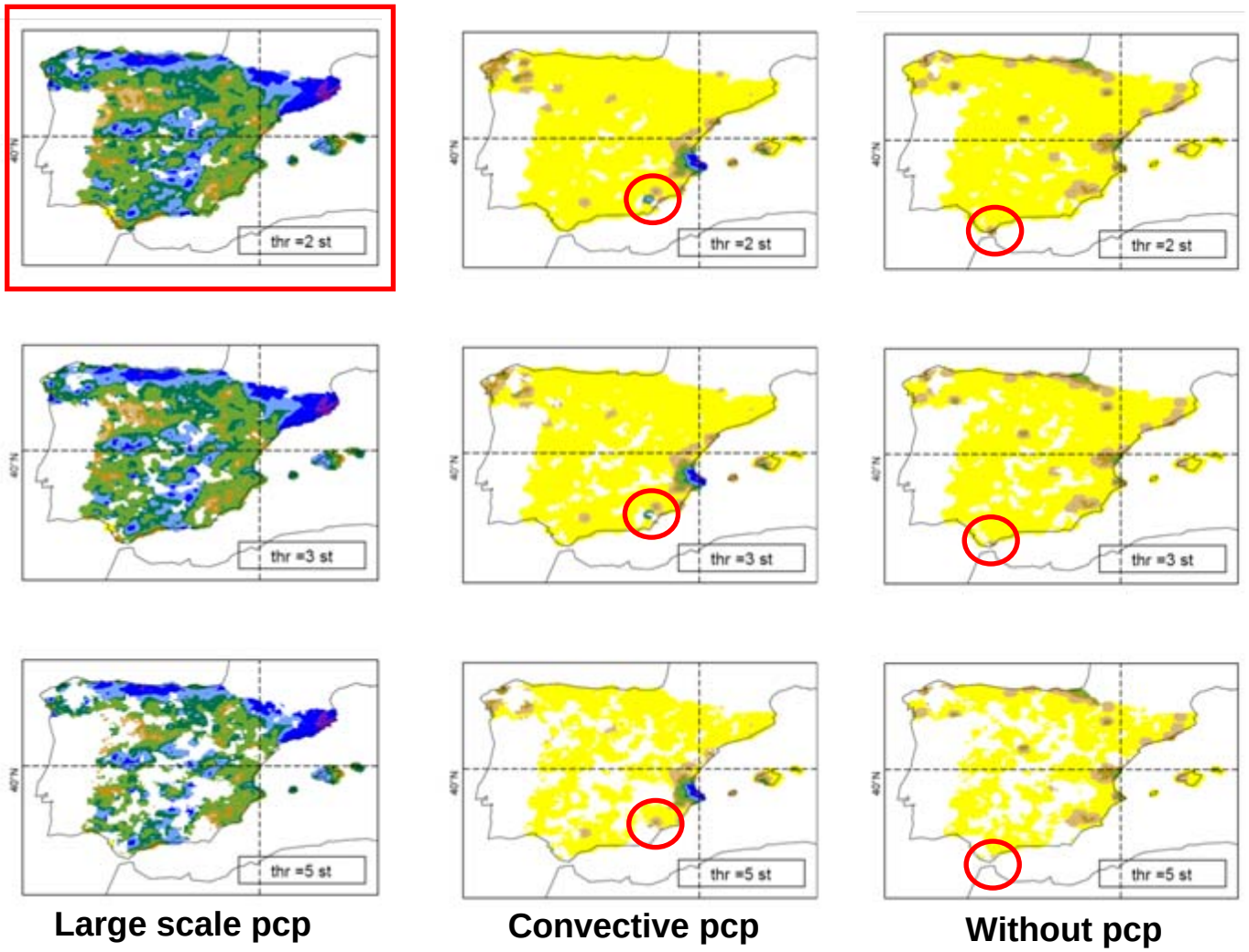
OBSERVATIONS



UPSCALING parameters for HIRLAM HNR(0.05º): distance



UPSCALING PARAMETERS: threshold in num. of stations

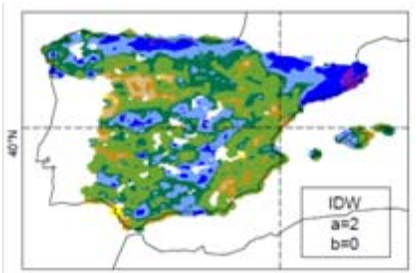


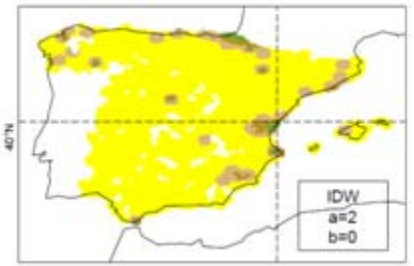
UPSCALING PARAMETERS: distribution function

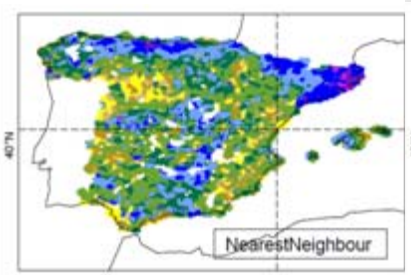
Inverse Distance Weighting

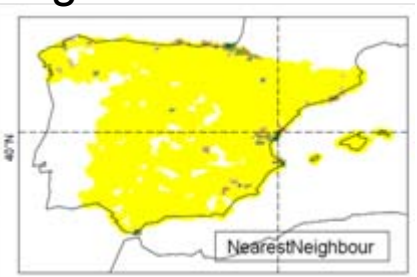
Nearest Neighbor

$$\frac{\sum R_i / r_i^2}{\sum 1/r_i^2}$$

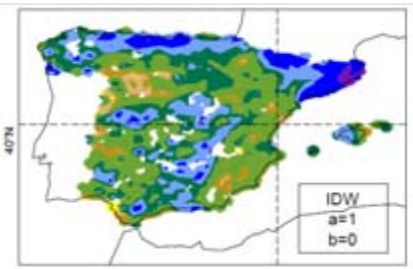


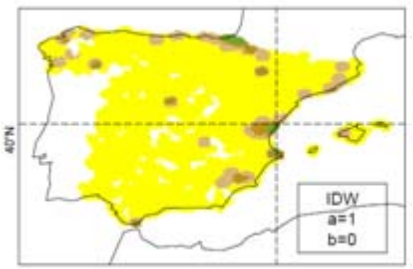




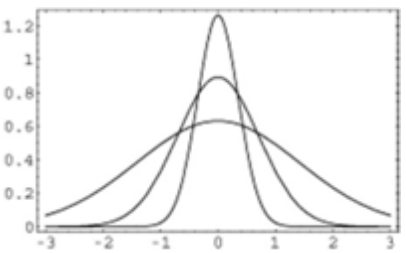


$$\frac{\sum R_i / r_i}{\sum 1/r_i}$$

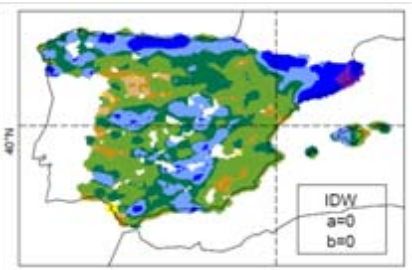


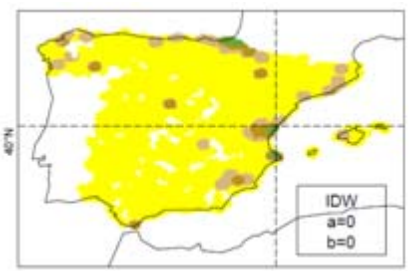


$$\frac{\sum R_i e^{-b r_i^a}}{\sum e^{-b r_i^a}}$$



$$\frac{\sum R_i}{N}$$





Exponential IDW
a=2 → gaussian dist.
↑ sigma → ↑ smoothing

$$\sigma^2 = \frac{1}{2b}$$

Up-scale obs to fit model

Pcp fields Rij for QPF(f) and QPE(o)

Find objects

Thresholding Rij $\geq$ R* = f R95 , f=1/15 empirical

Clustering → N objects

Object properties

Rn = $\sum$R(n)ij

Vn = Rn / R(n)max

xn = $\sum$R(n)ij**x**(n)nij / $\sum$R(n)ij = center of mass

Pcp field properties

D = (1/IJ)$\sum$Rij ~ E[R]

V = $\sum$RnVn / $\sum$Rn ~ E[R2]

x = $\sum$Rij**x**ij / $\sum$Rij = center of mass

r = $\sum$Rn**|x-x**n**|** / $\sum$Rn

d = max distance inside domain

SAL measures

S = 2 (Vf-Vo) / (Vf+Vo)

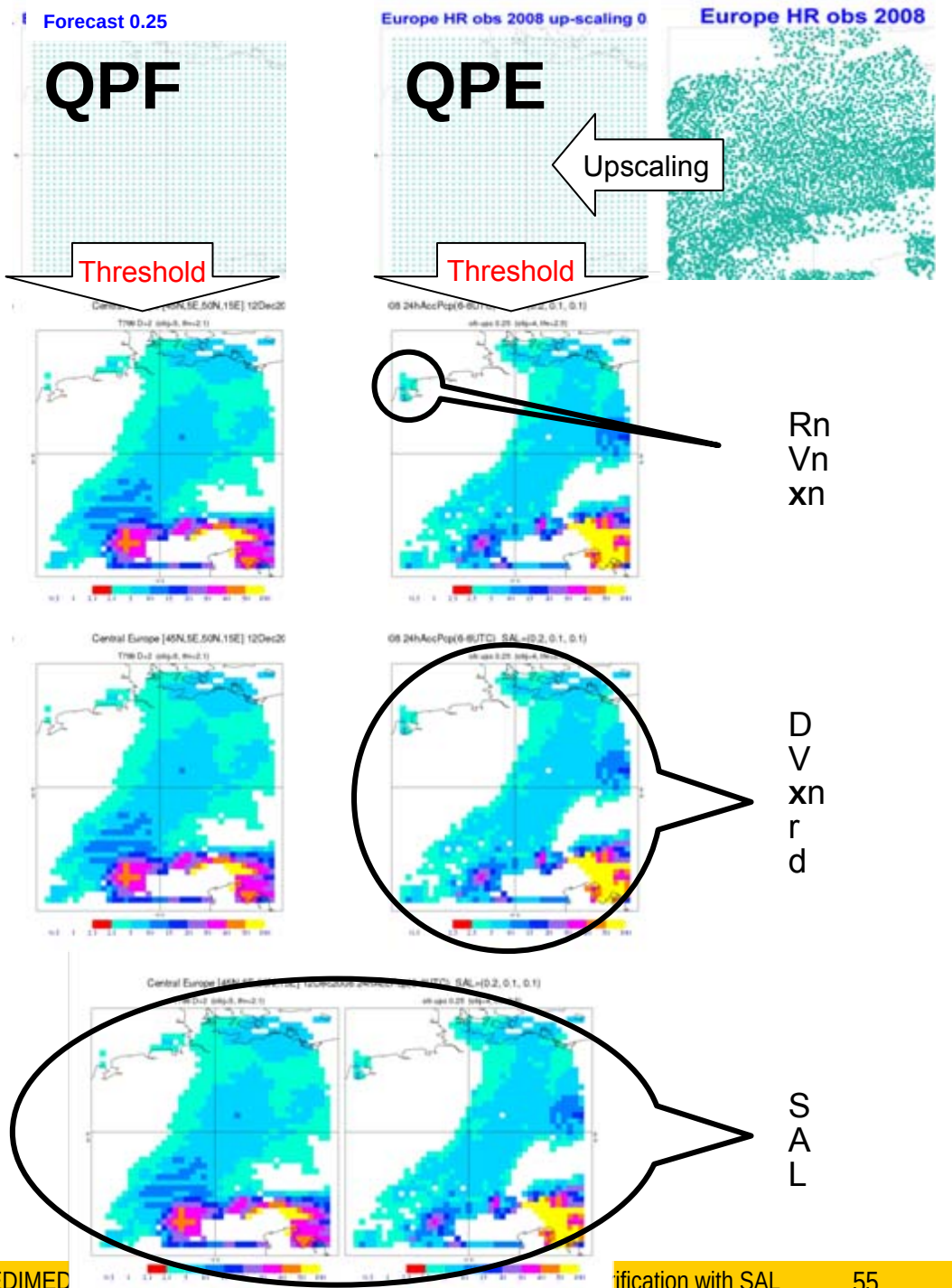
A = 2 (Df-Do) / (Df+Do)

L1 = |**x**f-**x**o| / d

L2 = 2 (rf-ro) / d

L = L1 + L2

SAL plot



SAL: measured aspects of forecast quality (I)

- **AMPLITUDE A**

Normalized difference of the domain-average pcp values between obs and fc fields.

Measure of **quantitative accuracy** of the **total amount of pcp** in the region.

A[-2,2]

A = 0 total agreement

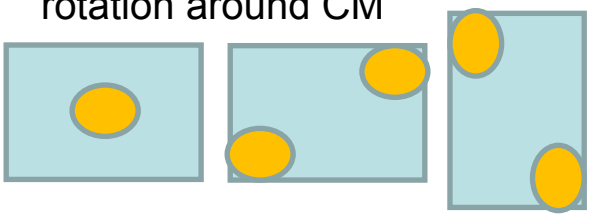
A > 0 model overestimates

A < 0 model underestimates

$$A = \frac{D(R_{\text{mod}}) - D(R_{\text{obs}})}{0.5[D(R_{\text{mod}}) + D(R_{\text{obs}})]}.$$

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

Not sensitive to rotation around CM



$$L_1 = \frac{|\mathbf{x}(R_{\text{mod}}) - \mathbf{x}(R_{\text{obs}})|}{d},$$

$$r = \frac{\sum_{n=1}^{M} R_n |\mathbf{x} - \mathbf{x}_n|}{\sum_{n=1}^{M} R_n}.$$

- **LOCATION L=L1+L2**

L1→ Normalized distance between the CM of the obs/fc pcp fields

First order indication of the **accuracy** of **the pcp distribution**

$$L_2 = 2\left[\frac{|r(R_{\text{mod}}) - r(R_{\text{obs}})|}{d}\right].$$

L2 → Takes into account the average distance between the CM of the total pcp field and infividual pcp objects → **relative positions** of objects in the field.

L[0,2] with **L = 0** → CM and average distance objects-CM are equal in obs and fc fields

SAL: measured aspects of forecast quality (II)

- **STRUCTURE S**

Compare the volume of the normalized pcp objects. $V_n = \sum_{(i,j)\in \mathcal{R}_n} R_{ij}/R_n^{\max} = R_n/R_n^{\max},$

Information about **size and shape** of objects.

Individual object volume → total pcp of the object normalized by its max value.

A weighted mean of all objects pcp volume is calculated for obs and fc fields.

$$V(R) = \frac{\sum_{n=1}^{M} R_n V_n}{\sum_{n=1}^{M} R_n}.$$

$$S = \frac{V(R_{\text{mod}}) - V(R_{\text{obs}})}{0.5[V(R_{\text{mod}}) + V(R_{\text{obs}})]}.$$

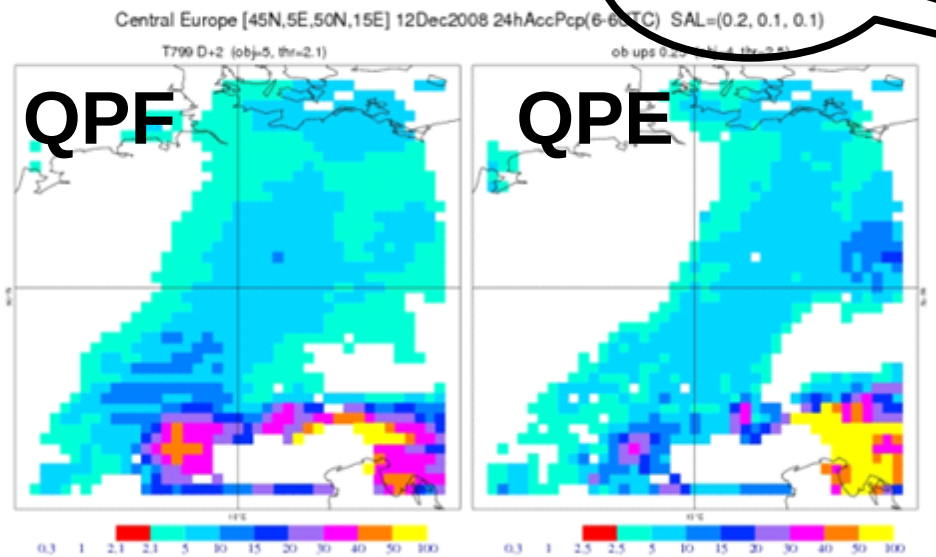
S → normalized difference between obs and fc weighted mean volumes.

S[-2,2]

S >> 0 → **model** predicts **widespread pcp** but observations show small convective events

S << 0 → **model** predicts **small and/or picked pcp** objects compare to observations

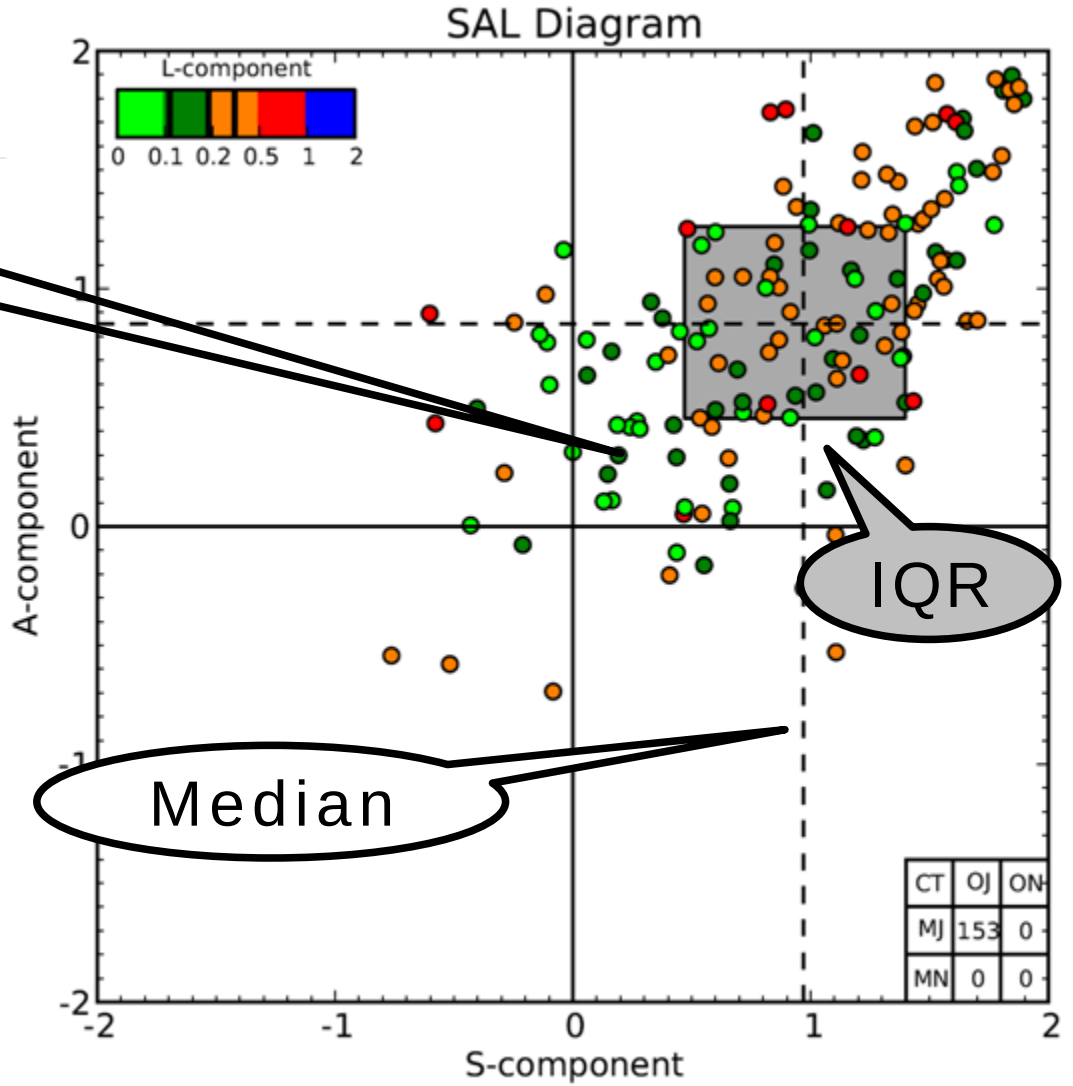
SAL plot



S: Structure — -2 … objects too small or too peaked; 0 … Perfect; +2 … objects too large or too flat

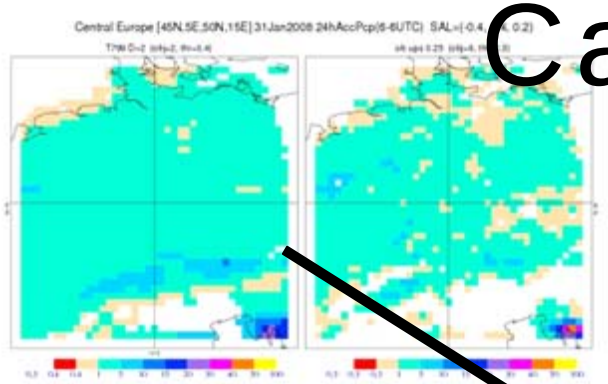
A: Amplitude — -2 … averaged QPF under-estimated; 0 … Perfect; +2 … averaged QPF over-estimated

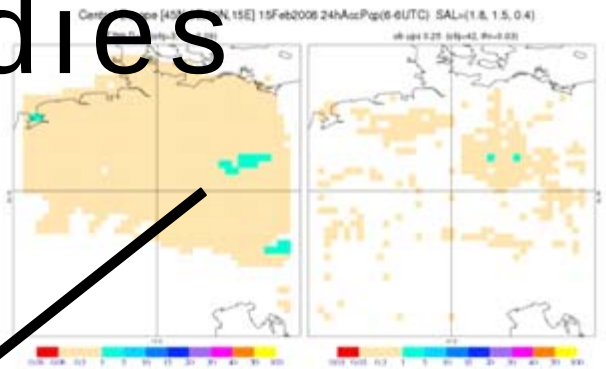
L: Location — 0 … Perfect; +2 … wrong location of Total Center of Mass (TCM) and / or of objects relative to TCM

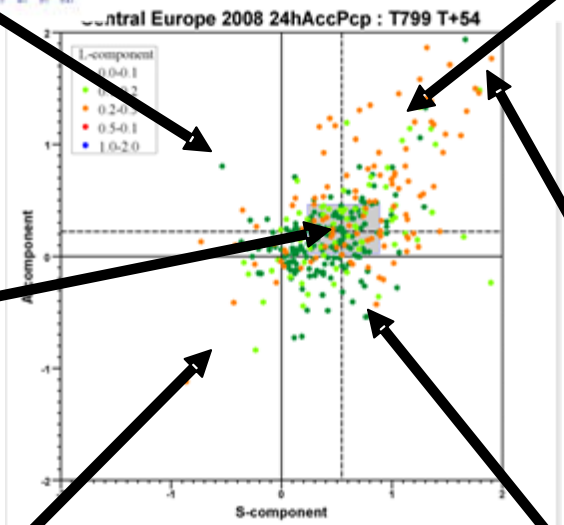


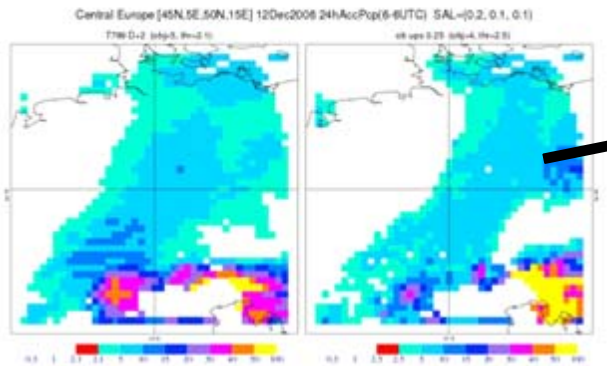
Quantitative, explicit and accumulative information about different aspects of precipitation forecast performance

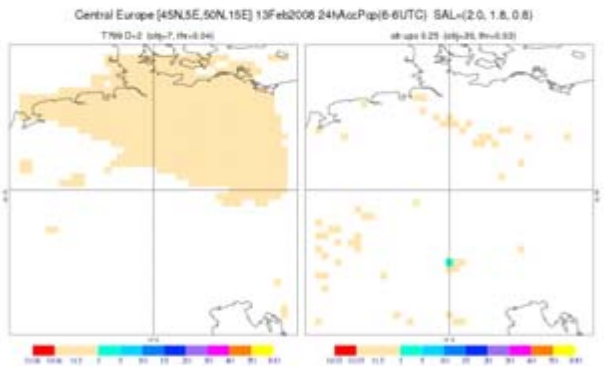
Case studies

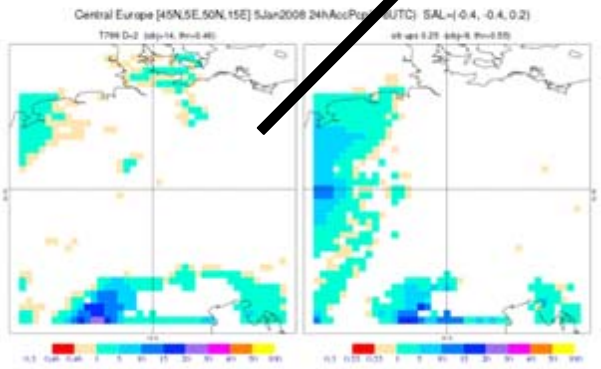


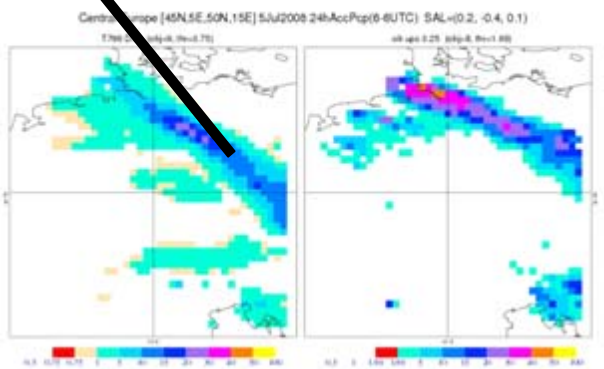


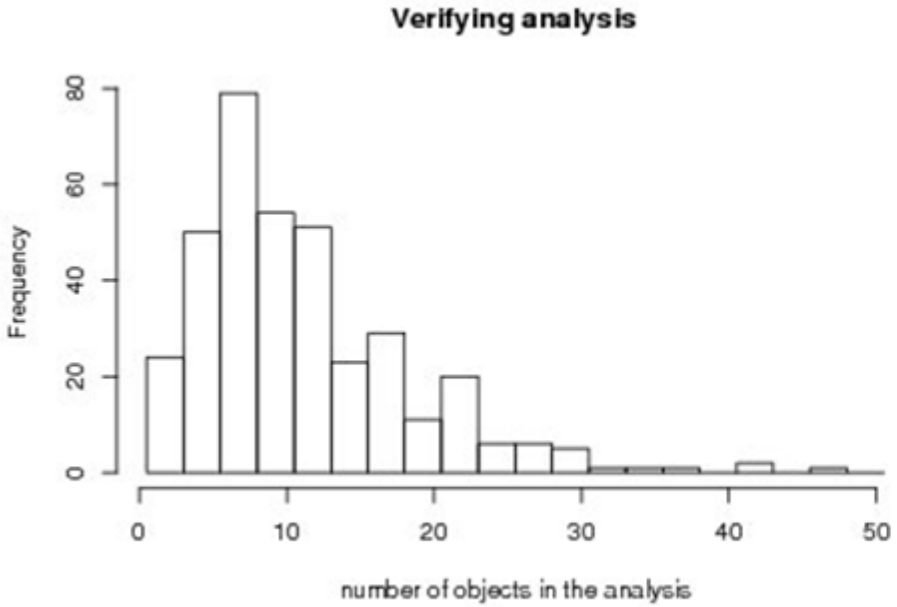


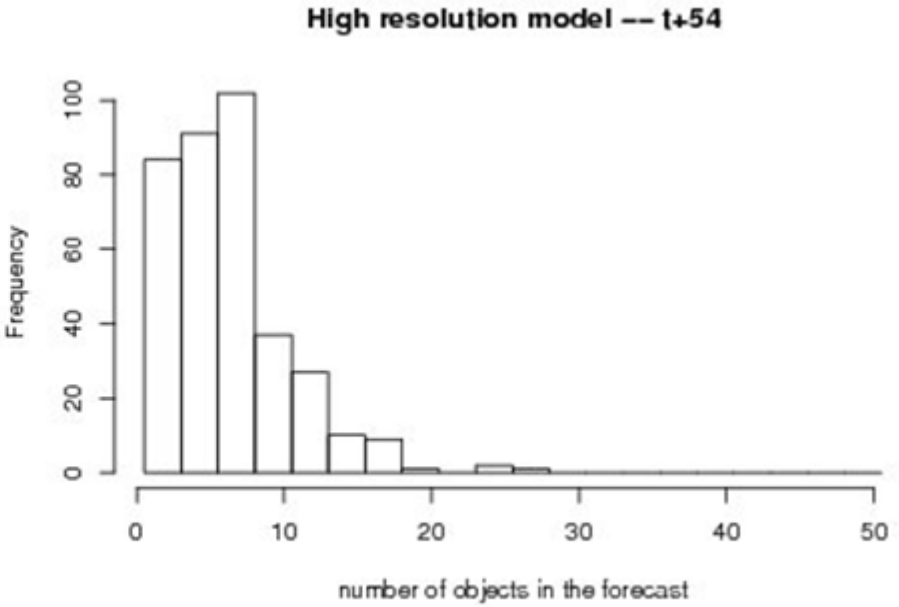


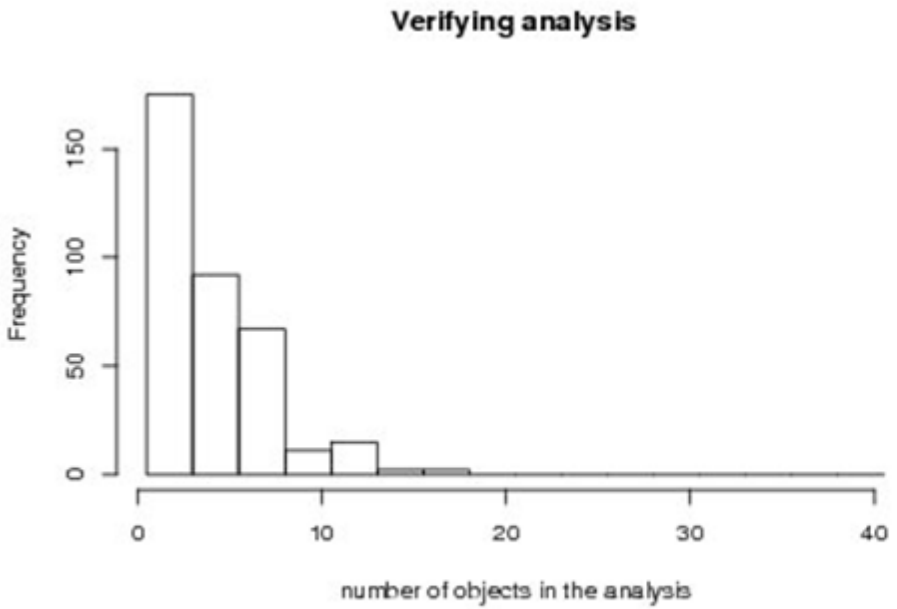


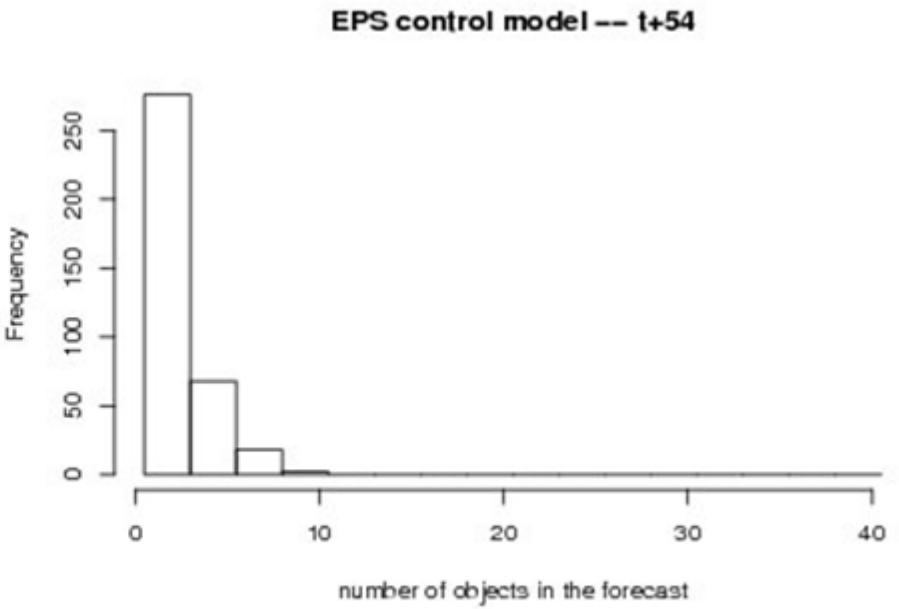




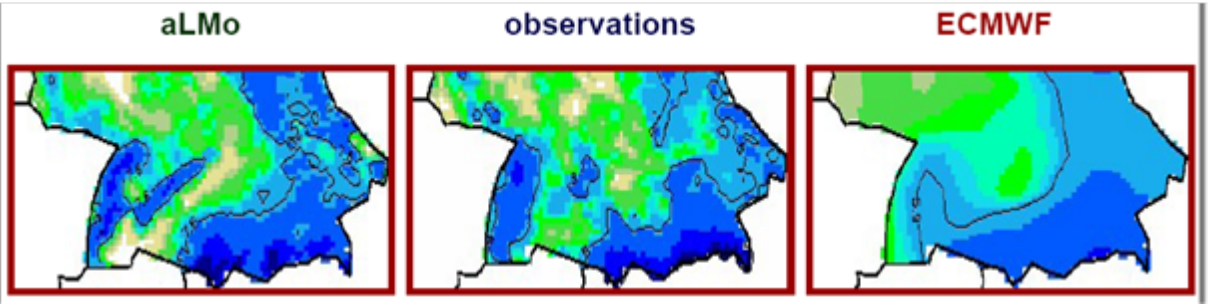

Verifying analysis
Frequency
0
20
40
60
80
0
10
20
30
40
50
number of objects in the analysis


High resolution model -- t+54
Frequency
0
20
40
60
80
100
0
10
20
30
40
50
number of objects in the forecast

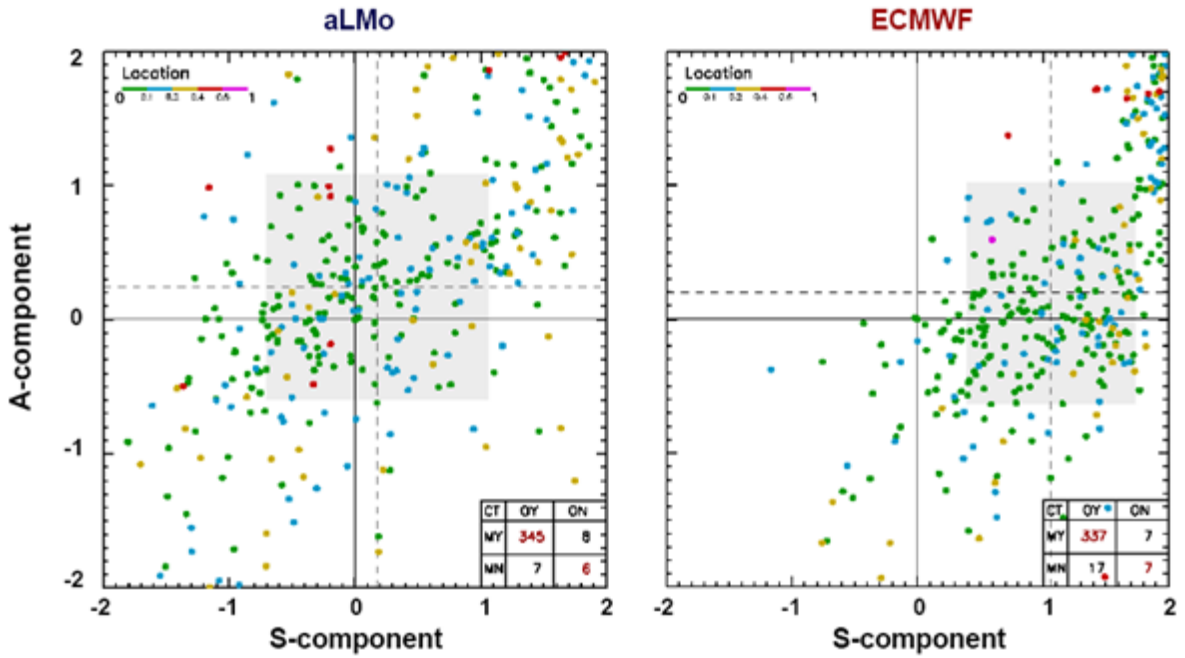

Verifying analysis
Frequency
0
50
100
150
0
10
20
30
40
number of objects in the analysis


EPS control model -- t+54
Frequency
0
50
100
150
200
250
0
10
20
30
40
number of objects in the forecast

Wernli et al: aLMo & ECMWF



summer seasons 2001-2004 for catchment Rhine



Courtesy Marcus Paulat

- QPF SAL performance
  - summer 2001-2004
  - German Elbe catchment
  - Up-scaling (3500 stations DWD, 10km)
  - aLMo 7km
  - ECMWF 0.25 interpolation
- Caveats
  - R* = f Rmax
  - where f = 1/15
- Results
  - ECMWF overestimates S

Work at AEMET

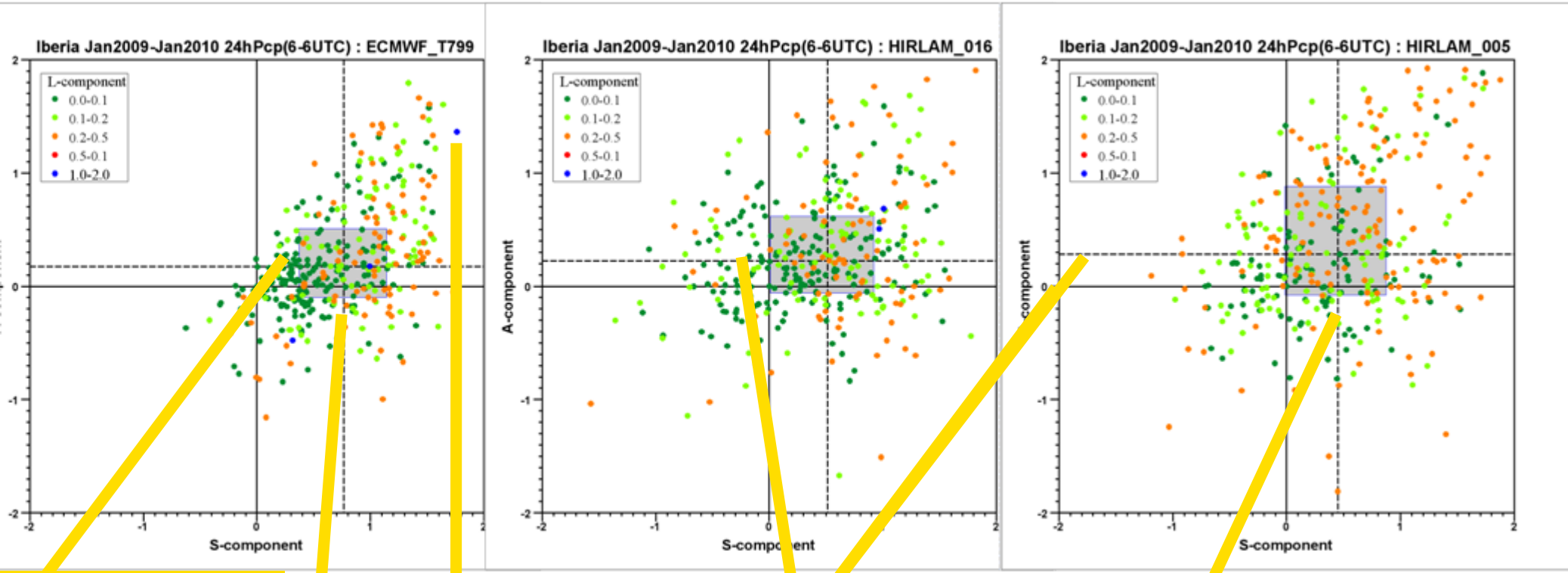
- SAL code adaptation (provided by Marcus Paulat, DWD)
- Up-scaling code implementation of two algorithms
  - Cell (with problems of missing data)
  - Structure functions $r^{-\alpha}$
- Research about models QPF SAL performance on one season
  - SON 2008
  - Iberian Peninsular
  - Up-scaling 3000 stations
- Research impact of:
  - Pcp threshold R* = f Rmax, over Spain f = 1/5..1/20
  - Model resolution: Hirlam_0.05, Hirlam_0.16, ECMWF T799
  - Model interpolation (original rotated to regular)

Work ECMWF-AEMET

- Collaboration framework ECMWF-AEMET
  - Anna Ghelli, Carlos Santos
  - Up-scaling & SAL code installed on linux cluster
- Research about models QPF SAL performance on one year
  - 2008
  - Central Europe (55N/5E/45N/15E)
  - Up-scaling 3000 stations
- Research impact of:
  - Pcp threshold R* = f Rmax, f = 1/15, stratification on 1.0mm pivot
  - Model resolution: T799, T399 (cf)
  - Forecast step: D+2, D+5

SAL

Objective and fair comparison of models with different resolution without penalty for the finer ones, e.g.: ECMWF 25km (T799), HIRLAM 16km, HIRLAM 5km



A > 0 ~ positive bias

some L>> ~ mis-location

A > ~ higher bias

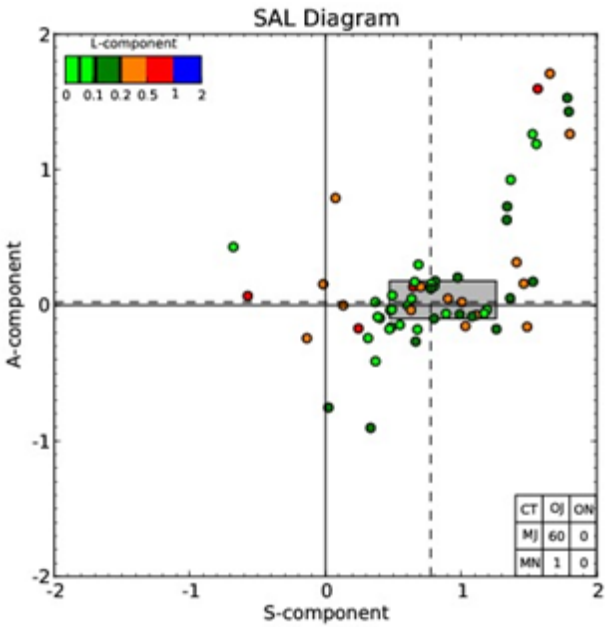
L< ~ better location

S big ~ over-population I and IV ~ structures large and/or flat

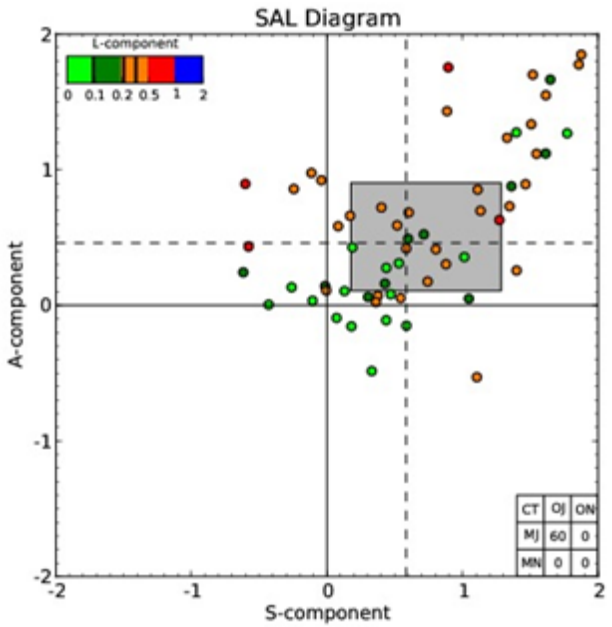
S < population more balanced over I, II, III and IV ~ better structures

SAL

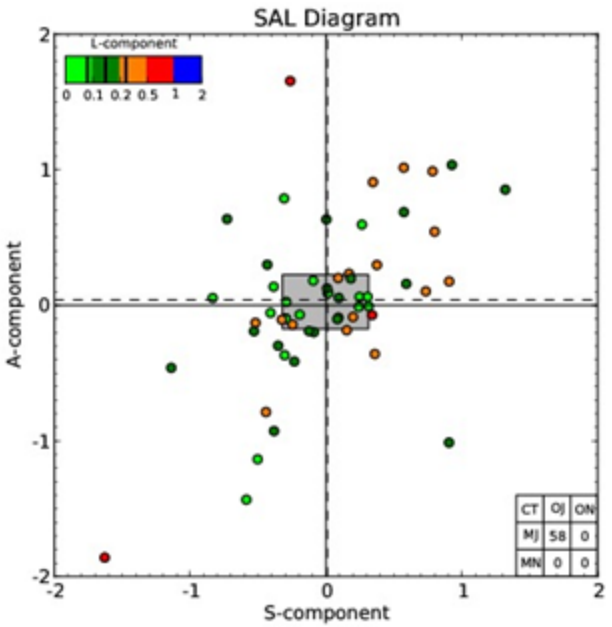
Objective and fair comparison for models of different resolution, without penalty for mesoscale ones; Structure (X) Amplitude (Y) Location (color)



ECMWF T1279 (17 km)



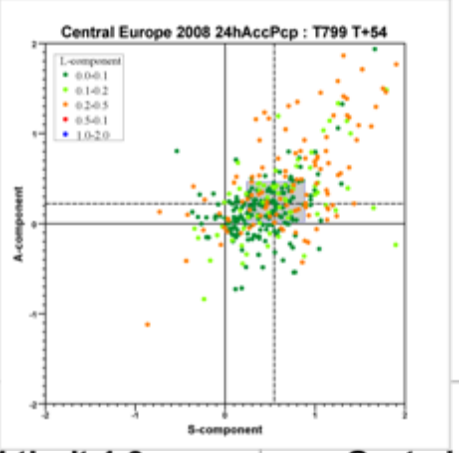
HIRLAM 0.05 (5 km)



HARMONIE 2.5 km

Pcp thr

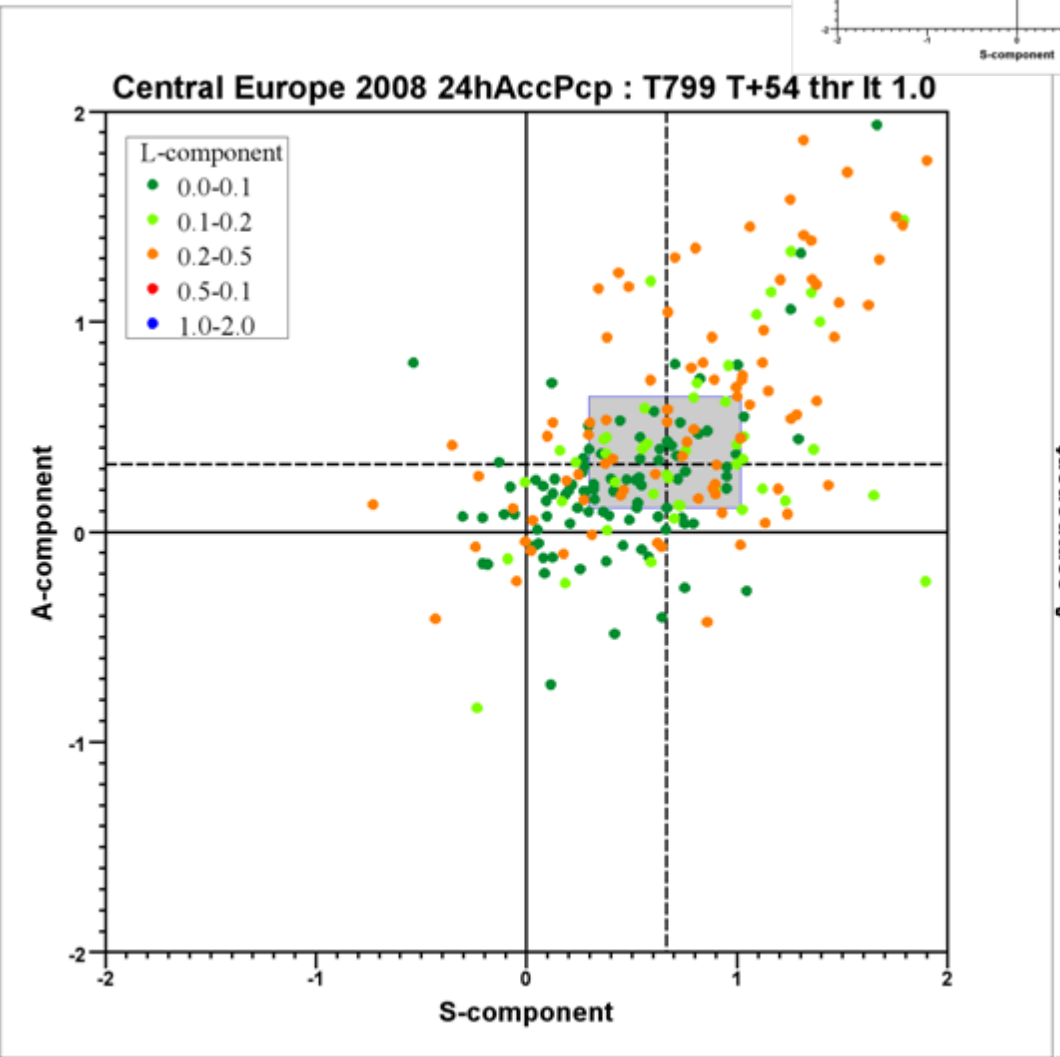
<

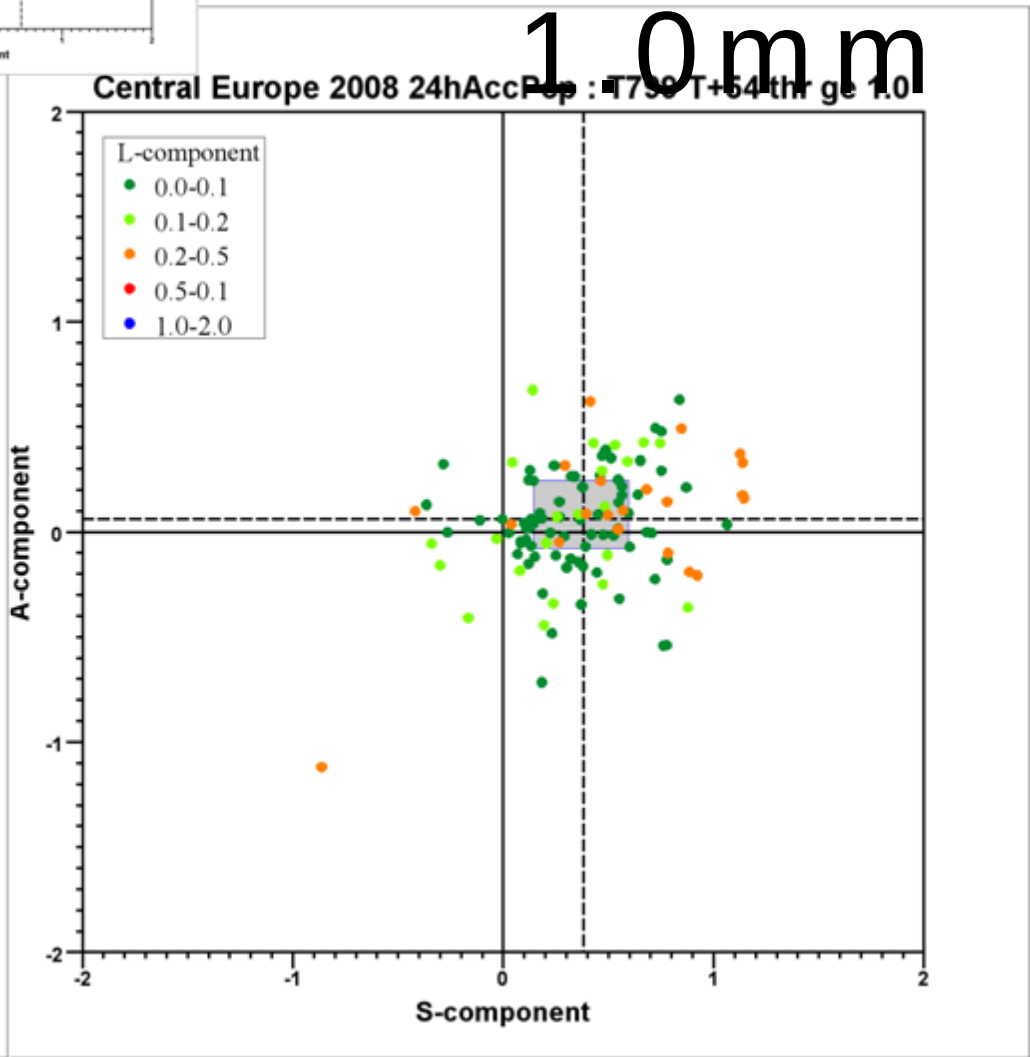


Pcp thr

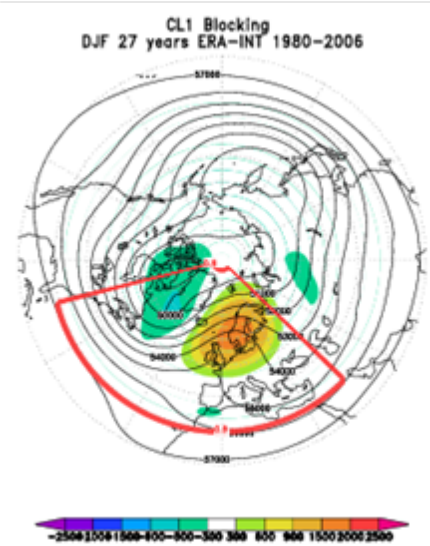
>=

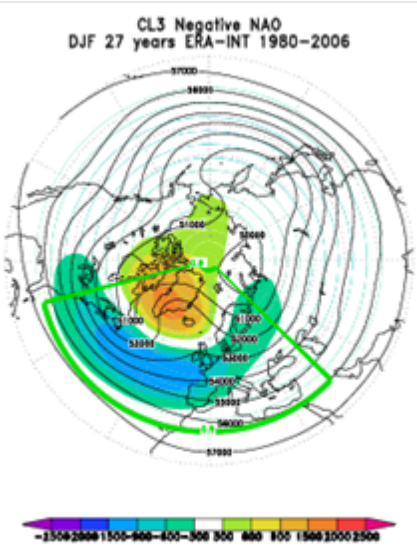
1.0mm

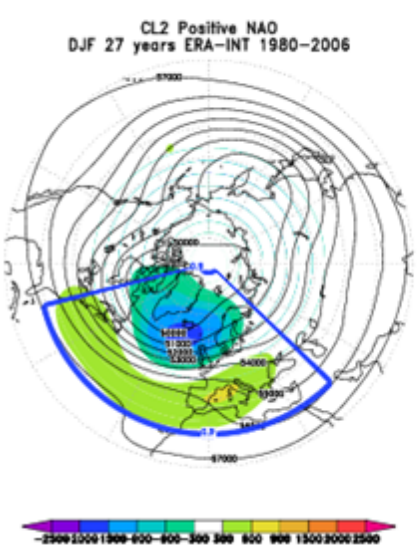


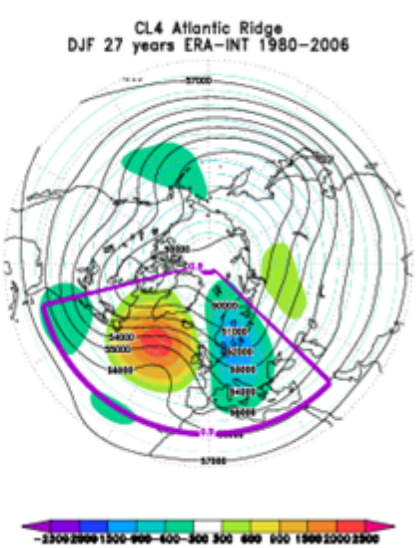


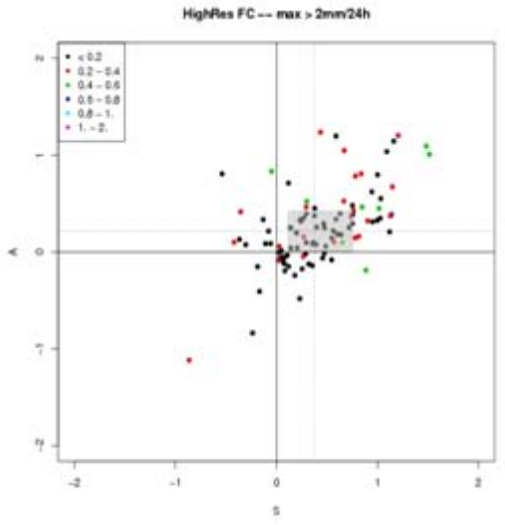
Flow-dependent SAL

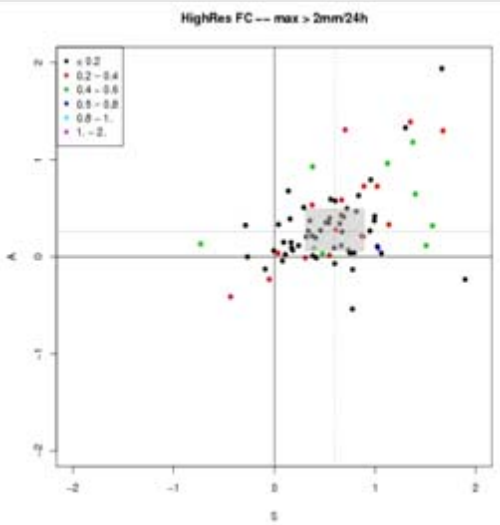


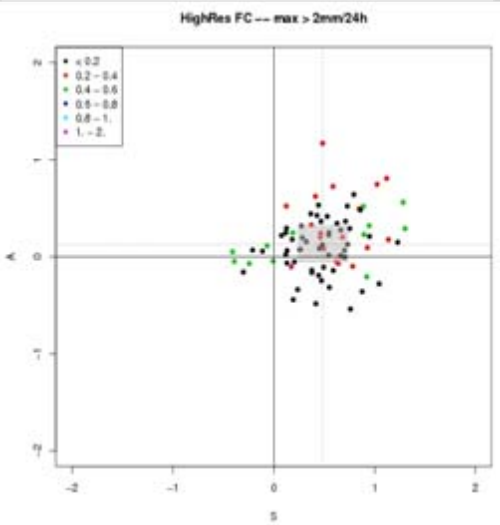


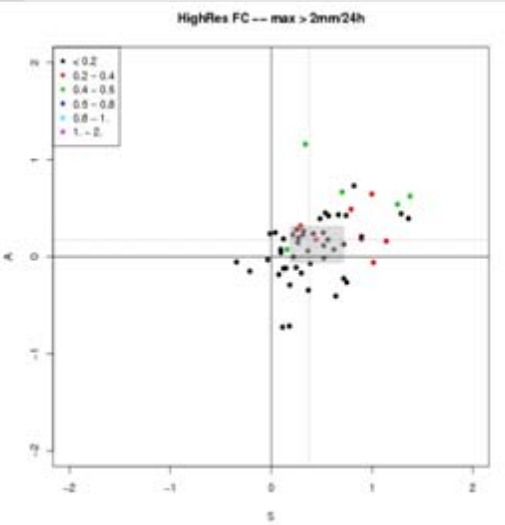




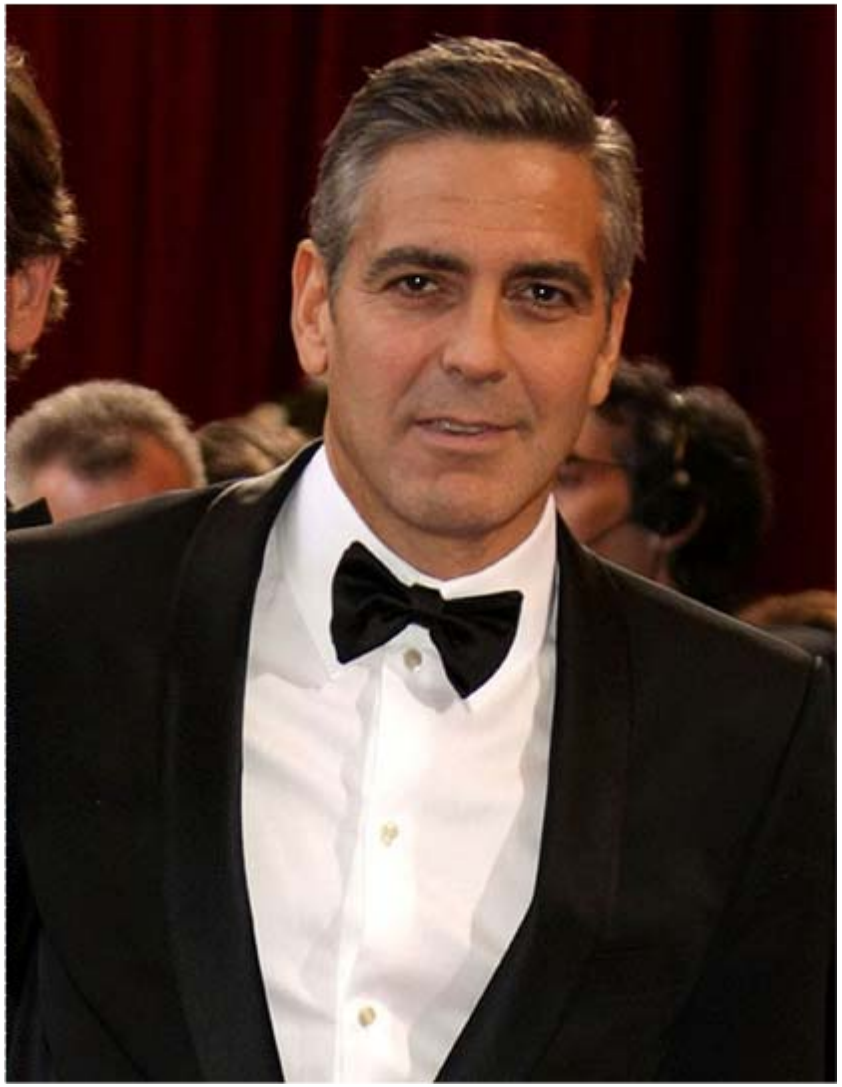








Are the models perfect?



What about the ensembles?

6

- Are the models perfect?
- Forecast verification: an introduction
- Classical methods
- A critical vision
- New spatial verification methods
- SAL: an example of feature-oriented method
- **MODE: a method applicable to EPSs**
- Radar & satellite data
- Software, conclusions, references

Object-Base Diagnostic Evaluation, MODE
(Davis et al. 2006, parts I & II)

Concepto: identificar objetos (áreas) en el campo observado y campo modelo relevantes para el observador humano, describir dichos objetos por una serie de atributos cuantificables, identificar sus contrapartidas en ambos campos, y comparar sus atributos.

Objetivo: evaluar el skill de predicciones de fenómenos localizados y episódicos (pcp's).

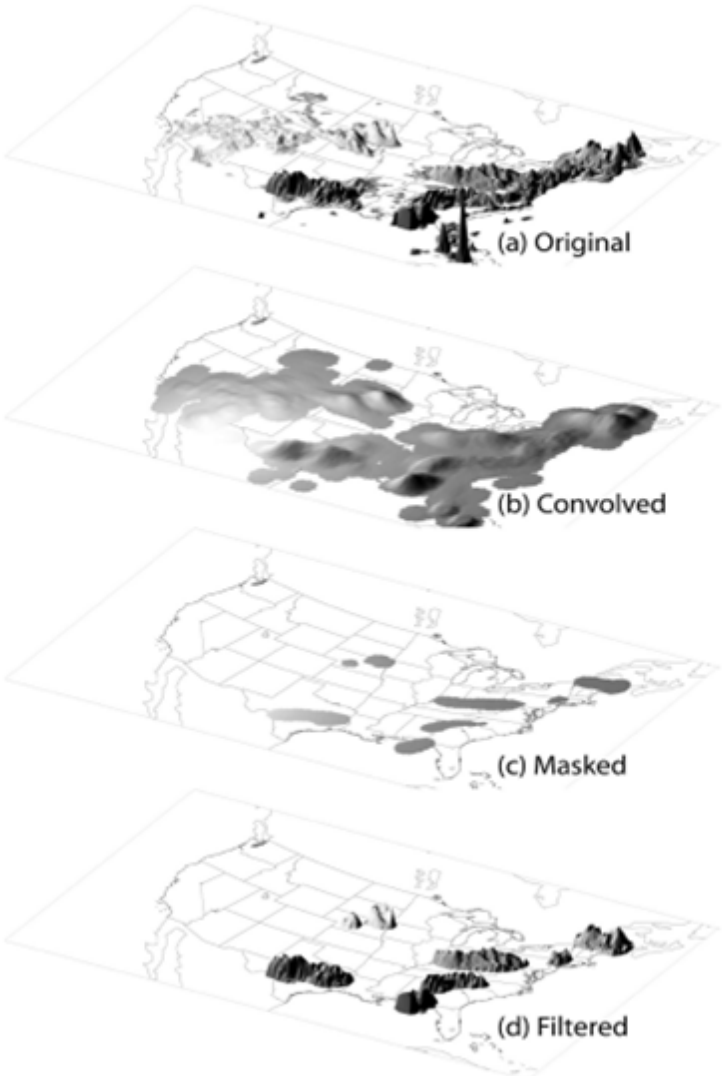
Método:

1-Identificación de objetos mediante filtrado espacial de los datos: eliminar áreas de pcp demaisado débiles en intensidad o demasiado pequeños en área.

2-Búsqueda de contrapartidas en ambos campos: criterios de coincidencia

3-Estadísticas de los objetos por separado de cada campo y en conjunto.

IDENTIFICACIÓN DE OBJETOS



Campo original del modelo

Filtrado de altas frecuencia espaciales → convolución a un disco de r = 4 x resolución

Máscara binaria → aplicar un umbral de pcp al campo filtrado (>x mm)

Máscara aplicada sobre el campo original → retiene las pcp originales. Sólo se consideran áreas > 25 pix.

PROPIEDADES DE LOS OBJETOS

Intensidad: pcp considerada como una distribución acumulativa.

Tamaño o área. Variable independiente → escalas horizontales asociadas a sistemas con diferente dinámica y predecibilidad

Centroide o localización del objeto.

Ángulo de orientación: eje mayor del objeto respecto a la dirección EW

Razón de proporcionalidad= eje menor/eje mayor

Curvatura= 1/radio

Fracción de área cubierta=area con pcp/area del objeto

Indicativo de la complejidad de la forma del objeto

Teniendo en cuenta la dimensión temporal:

Duración de sistemas de pcp.

Velocidad media de traslación del sistema

DIMENSIÓN TEMPORAL: SISTEMAS DE PCP

Una secuencia de áreas de pcp en tiempos consecutivos pueden formar parte de un único y coherente sistema de pcp que se desplaza en el tiempo y el espacio.

Cómo se definen los sistemas en el modelo?

Se buscan coincidencias entre áreas de pcp separadas temporalmente en 1h y espacialmente por < distancia umbral dada.

En sucesivas comparaciones del campo a 0h con los campos a 1h,2h,…, se asignan valores crecientes al índice k (duración) en caso de existir coincidencia entre áreas, y los valores de los atributos de dichas áreas se promedian, pasando a formar un sistema.

Las áreas con ninguna coincidencia presentan k=1.

Resultado del proceso: se favorecen los atributos de las áreas situadas en medio del ciclo del sistema, debido a los pesos resultantes de los promedios → los valores de los atributos del sistema son representativos de su etapa madura.

Distancia umbral: obtenida a partir de la velocidad máxima de traslación que consideremos para los sistemas (144 km/h).

Velocidad media de traslación del sistema:

(centroide final-centroide inicial)/k

CRITERIOS DE COINDICENCIA MODELO-OBSERVACIÓN

Dimensión espacial (áreas):

Favorecen las áreas grandes

$$D < \frac{\sqrt{Ao} + \sqrt{Af}}{2}$$

Mínimo de solapamiento

$$D < \sqrt{Ao} + \sqrt{Ao}$$

- Dimensión espacial y temporal (sistemas):
  Error de posición = (xf-xo, yf-yo, tf-to) → errores separados

$$[(xf - xo)^2 + (yf - yo)^2]^{1/2} < 4\bar{L}$$

$$\bar{L} = (Lf + Lo)/2$$

ESPACIAL

$$|tf - to| < 3h$$

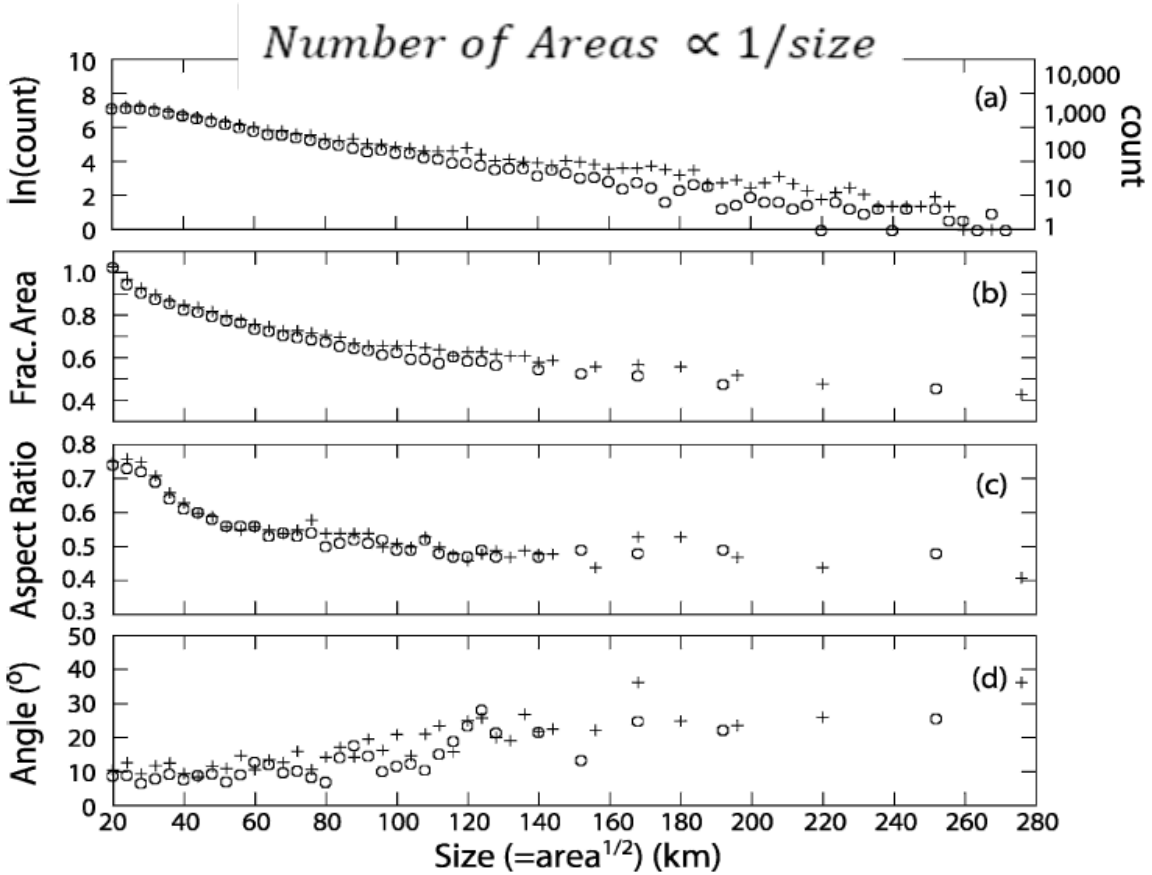
TEMPORAL

$$0.5 \leq \frac{ko}{kf} \leq 2$$

DURACIÓN

ESTADÍSTICAS INDIVIDUALES DE ÁREAS

Identificación de **sesgos** en los atributos → **errores sistemáticos** del modelo



Sesgo + para tamaños > 20-30 celdas (modelos de 22 y 4km de resolución).

Sesgo +

Ligero sesgo + a pequeñas áreas

No estadísticamente significativo

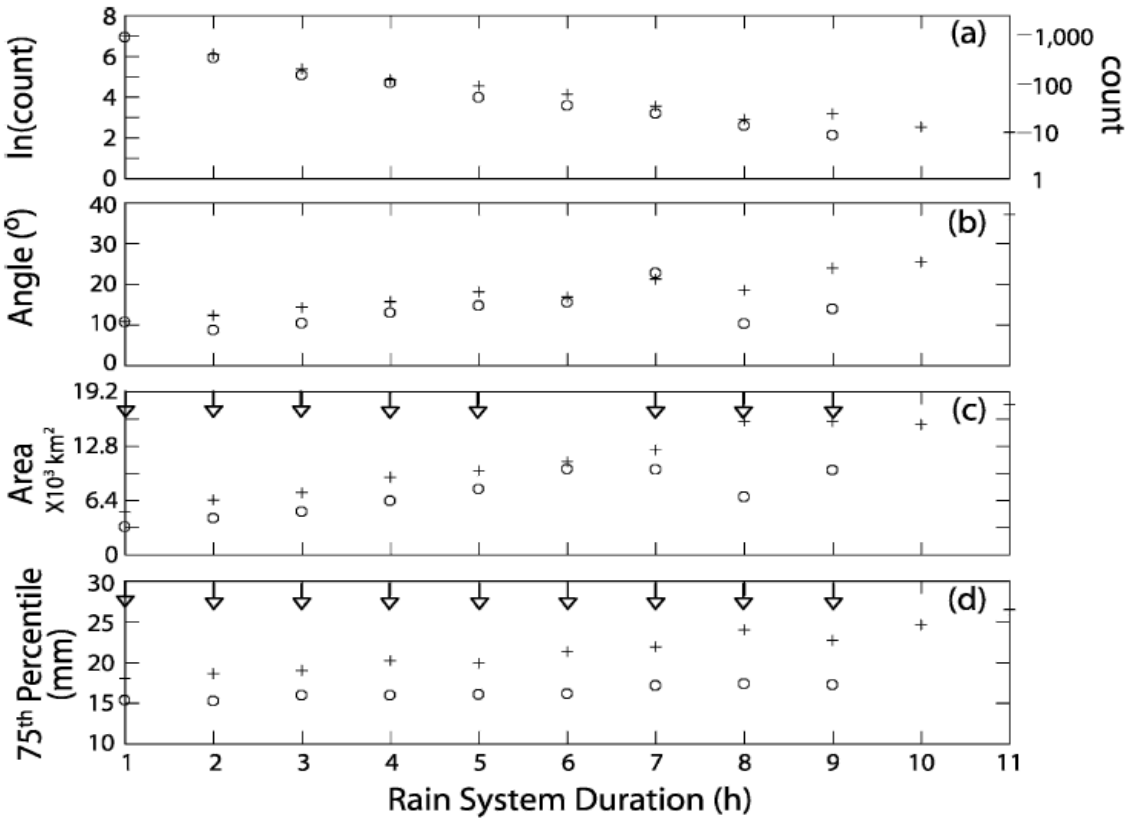
+ modelo

O observado

Ejemplo info extraíble: áreas elongadas o con similar orientación pueden indicar pcp forzada por sistemas frontales

ESTADÍSTICAS DE SISTEMAS DE PCP

En función de la duración del sistema, k



$$Number\ of\ Systems \propto 1/k$$

Estadísticamente no significativo

Sesgo + del modelo

$$Size \propto k$$

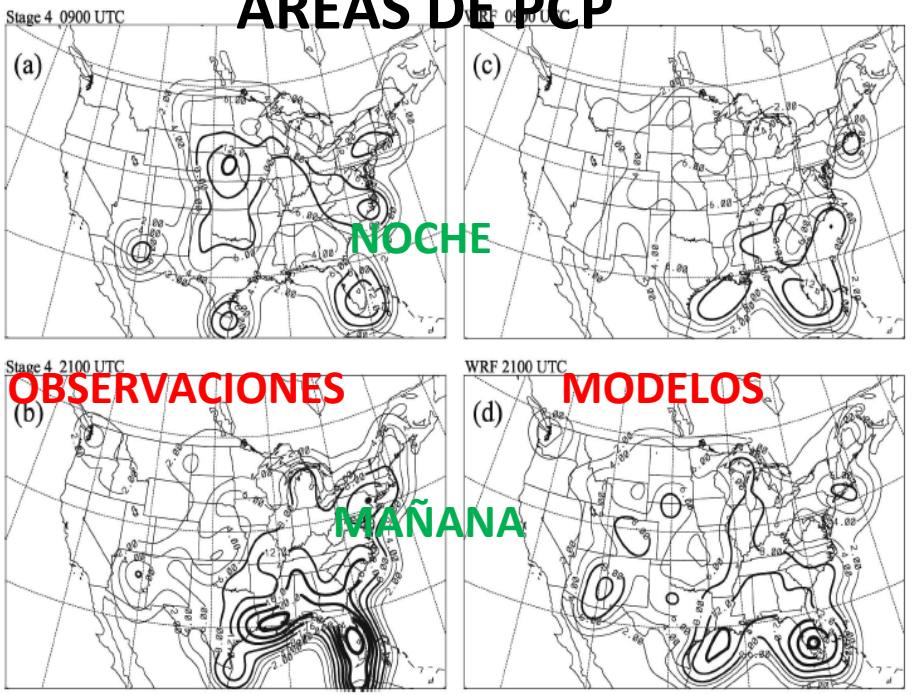
Modelo predice más pcp intensa

+ modelo

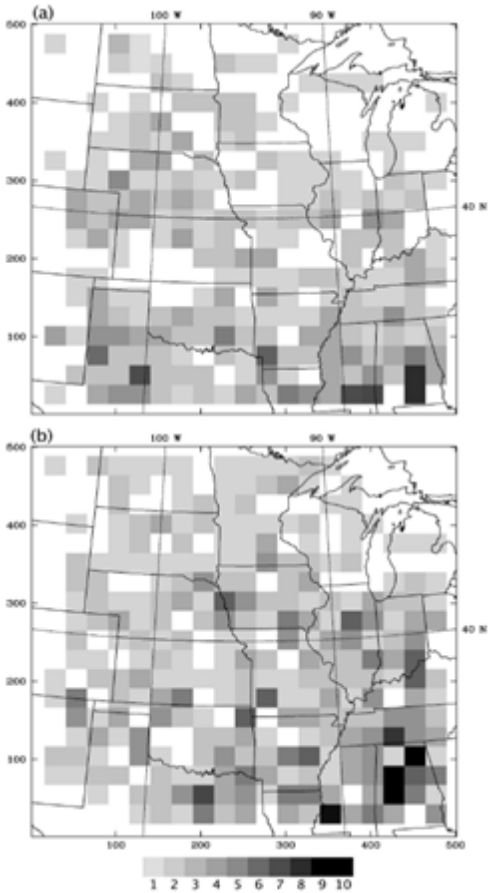
O observado

DISTRIBUCIÓN ESPACIAL

AREAS DE PCP



SISTEMAS DE PCP



En cada pto de grid se evalúa el número de areas de pcp próximas (a una distancia s) mediante:

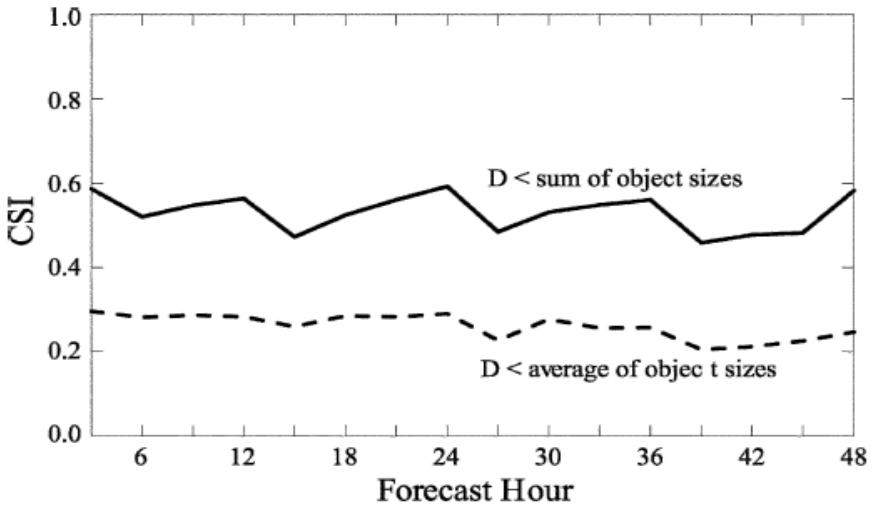
$$Numb.Rain\ Areas = \sum_{m} exp(-d_m^2/s^2)$$

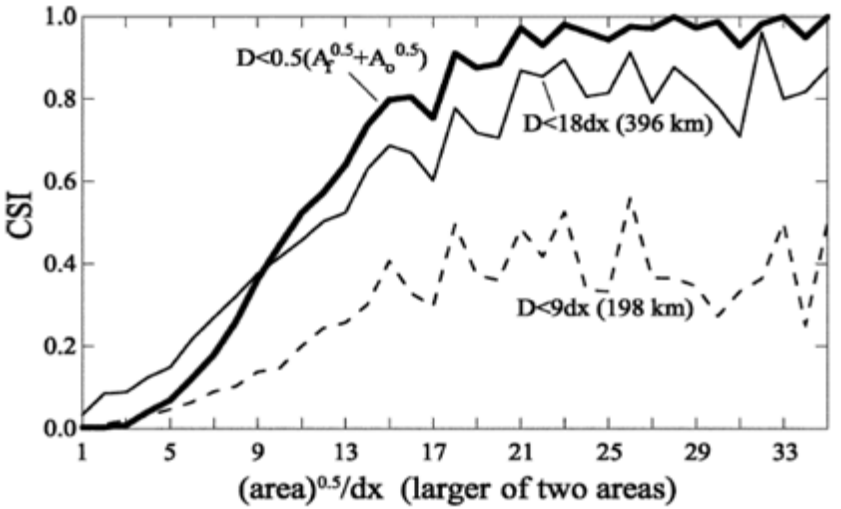
Número de sistemas con k>=3 en cada píxel de 25x25 (100 km2)

ESTADÍSTICA DE LAS COINCIDENCIAS:
Critical Success Index (CSI)

CSI: Coincicencias/(Coincicencias+FalsasAlarmas+Pérdidas)

Decrecimiento lento con el tiempo



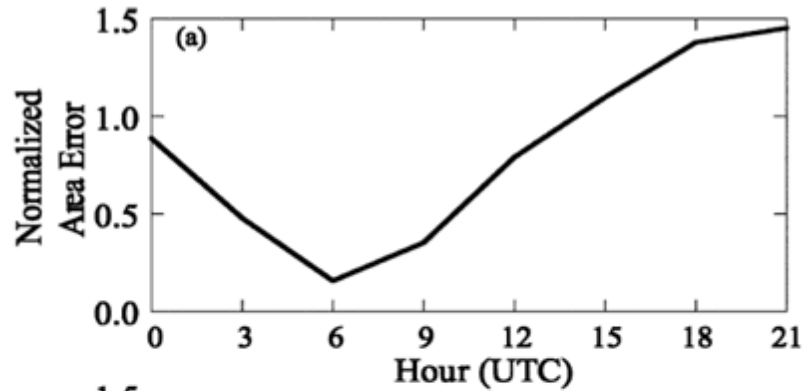


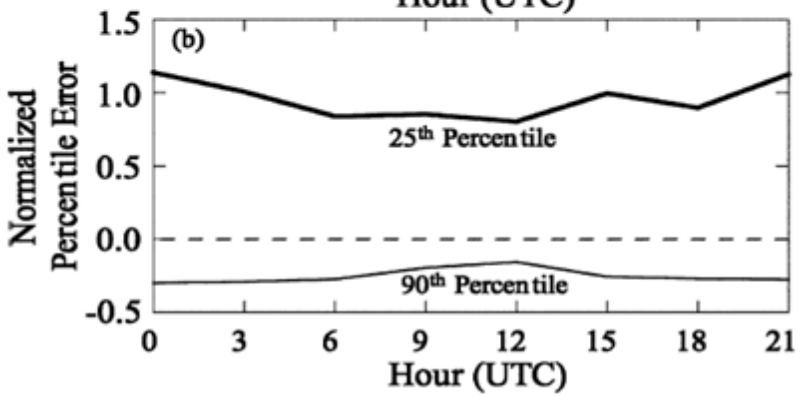
Nº coincidencias depende del tamaño de las áreas

Plateau a mayores tamaños → áreas de pcp de mayor tamaño asociadas a sistemas de gran escala → mejor predecibilidad

ESTADÍSTICA DE COINCIDENCIAS:
sesgos normalizados

$$B_{\text{area}}^{(h)} = \frac{\sum_i [A_{f,i}^{(h)} - A_{o,i}^{(h)}]}{\sum A_{o,i}^{(h)}}.$$





$$B_{\text{intensity}}^{(N)} = \frac{\sum_i [R_{f,i}^{(N)} - R_{o,i}^{(N)}]}{\sum_i R_{o,i}^{(N)}}$$

Parámetros en función del momento del día:

TAMAÑO DEL ÁREA DE PCP:

Sesgo de área + y fuerte durante el día

El modelo sobrestima las tamaños, exagerando la amplitud de la variación diaria observada.

INTENSIDAD DE LA PCP:

P25 es sobrestimado por el modelo.

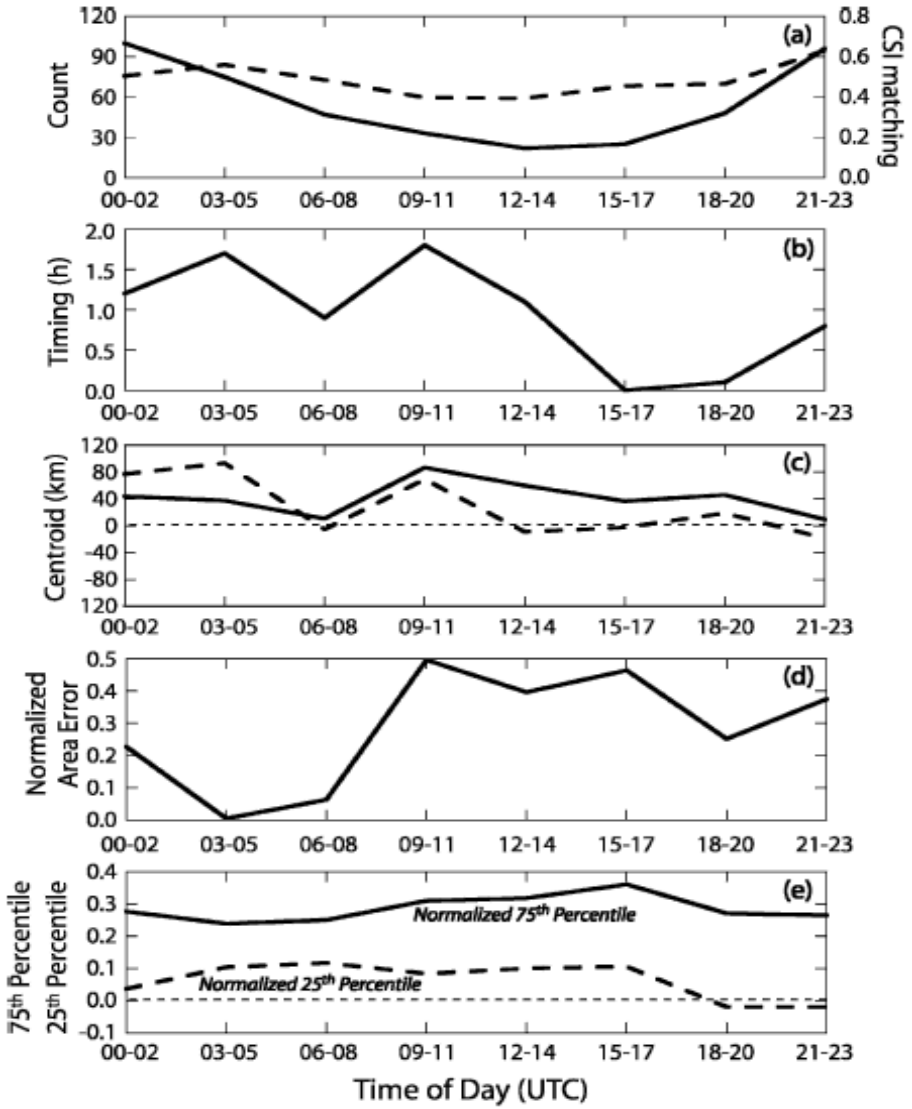
P90 es subestimado

El modelo predice una distribución de pcp demasiado estrecha.

Posible causa: parametrización de la convección (bias mayor a final de la tarde)

También se pueden representar CSI, errores de sincronización y localización en función de la hora del día

ESTADÍSTICA DE COINCIDENCIAS:



MODE proporcionado por el paquete de verificación de NCAR: Model Evaluation Tools (MET).

Modificaciones posteriores del MODE (interés total):
Davis et al. 2009

MODE aplicado a ensembles:
Gallus, 2010

7

- Are the models perfect?
- Forecast verification: an introduction
- Classical methods
- A critical vision
- New spatial verification methods
- SAL: an example of feature-oriented method
- MODE: a method applicable to EPSs
- **Radar & satellite data**
- Software, conclusions, references

HR obs data

SEVIRI

- Roebeling et al 2011: TRIPLE COLLOCATION OF SUMMER PRECIPITATION RETRIEVALS FROM SEVIRI OVER EUROPE WITH GRIDDED RAIN GAUGE AND WEATHER RADAR DATA

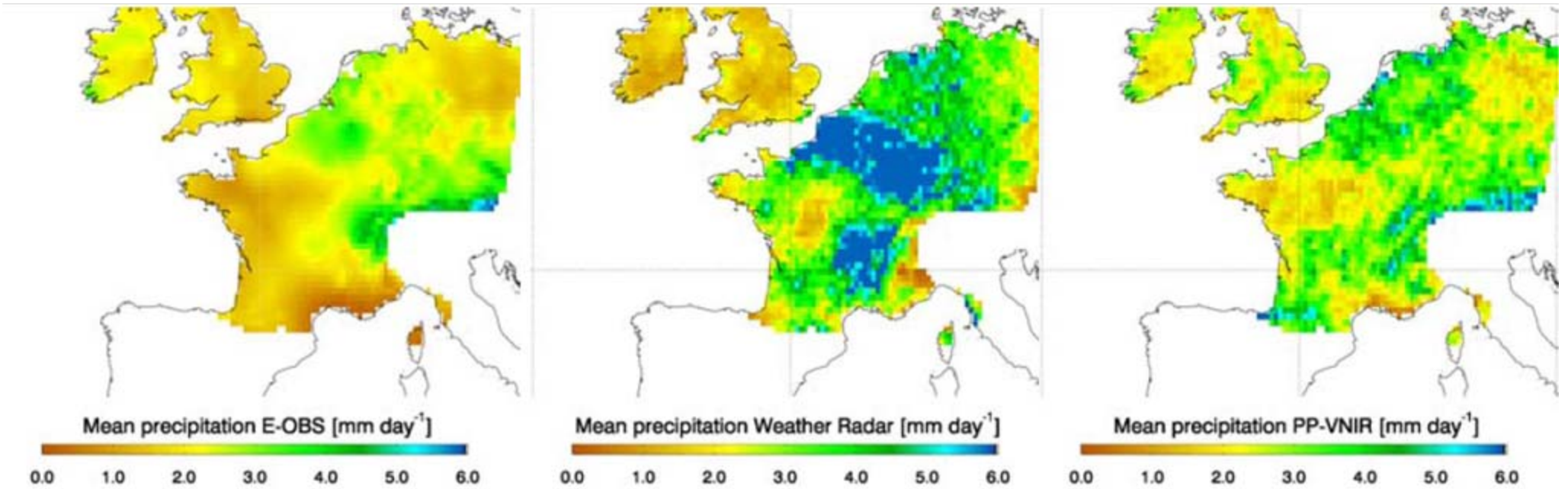
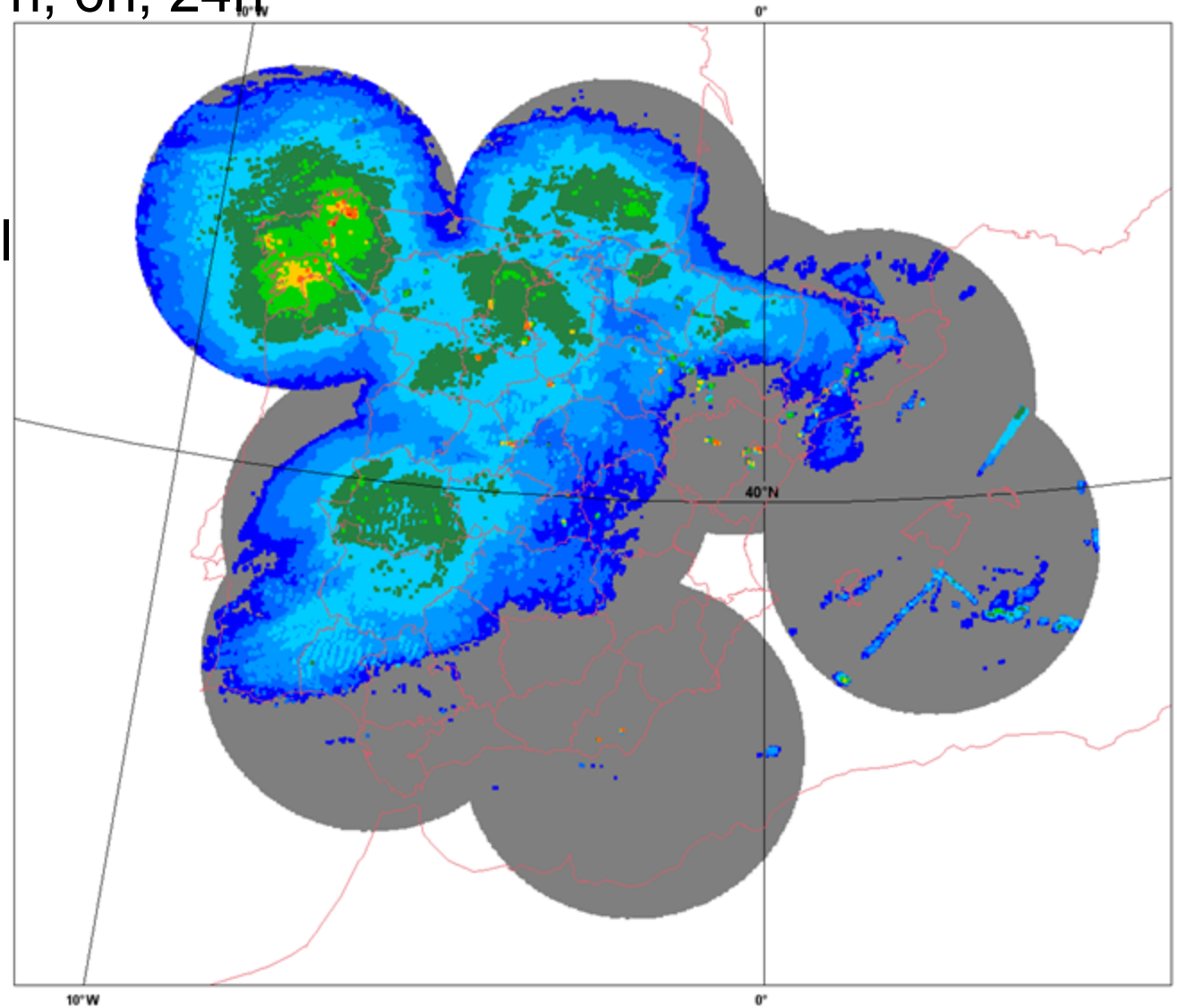


FIG. 1. Example of the mean daily precipitation amounts from E-OBS (left panel), Weather radar (middle panel), and PP-VNIR (right panel) in mm day^{-1} over the period May-August 2006.

HR obs data

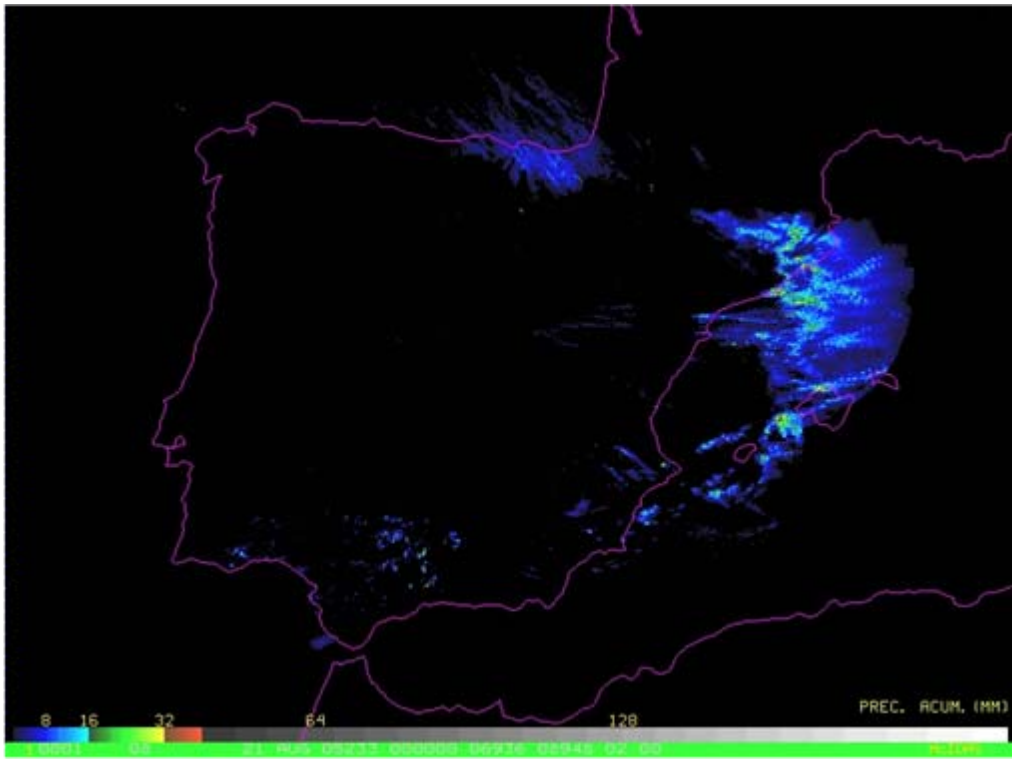
RADAR

- Acc Pcp estimates 1h, 6h, 24h
- Spain Composition
- 2 km, 1 km
- Issue: quality control



Are the observations perfect?

- Are the observations good representations of reality?
- Quality control



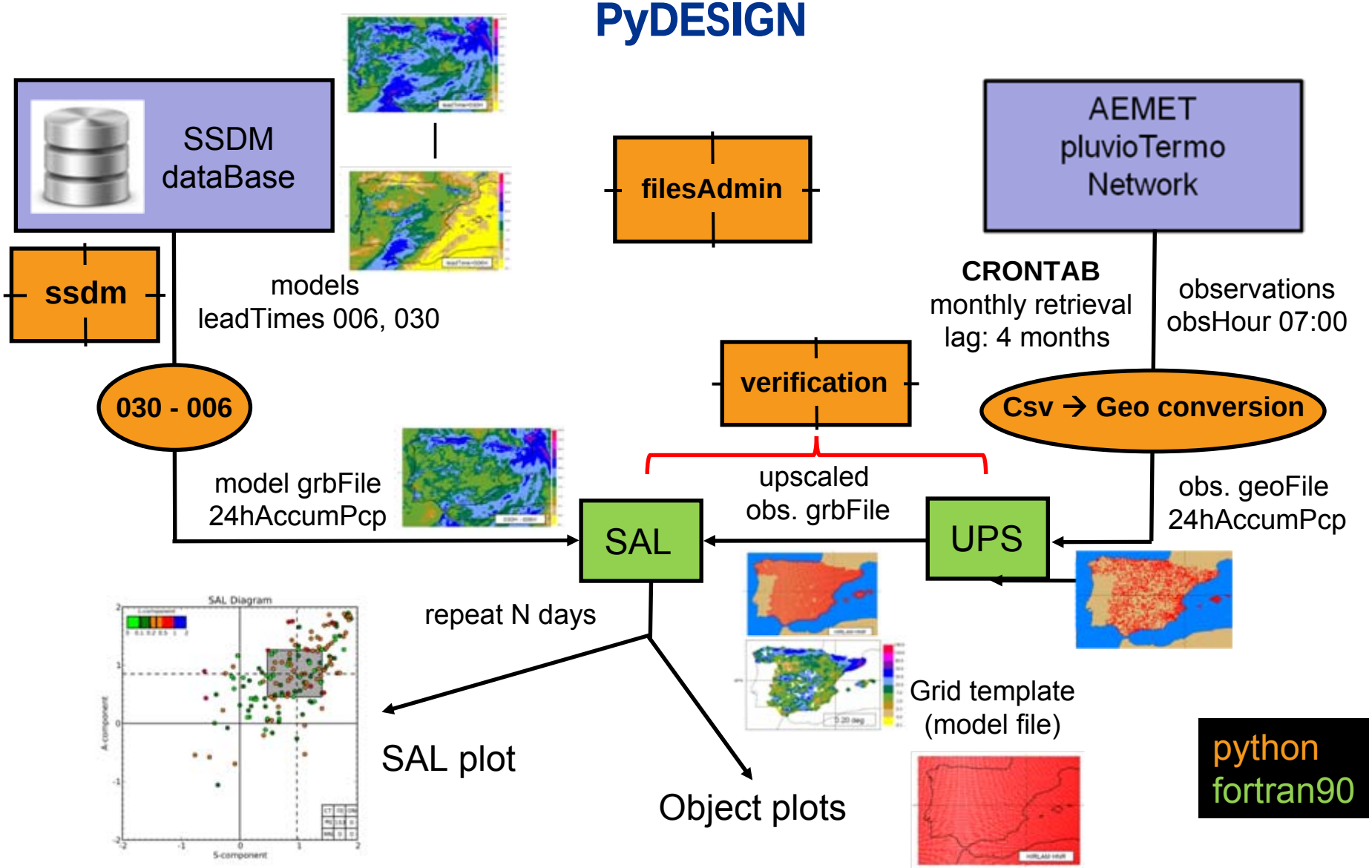
8

- Are the models perfect?
- Forecast verification: an introduction
- Classical methods
- A critical vision
- New spatial verification methods
- SAL: an example of feature-oriented method
- MODE: a method applicable to EPSs
- Radar & satellite data
- **Software, conclusions, references**

Software

- No hay paquetes de verificación completos ni estandarizados (de hecho no hay un sistema abstracto estándar de verificación)
  - Inmensos volumen y diversidad de datos: GRIB, netCDF, BUFR, etc
  - Volumen colosal de metadatos: SQL o similar
  - Entorno de desarrollo
  - Lenguaje de programación OO
  - Soporte array, estadístico, geográfico, gráfico
- Algunas opciones
  - Model Evaluation Toolkit (MET, NCAR)
  - Paquete de verificación R (CRAN)
  - MetPy+Verify (ECMWF pero no liberado)
  - HIRLAM
  - **Home-made: python [+ C++]**

PyDESIGN


SSDM
dataBase
ssdm
models
leadTimes 006, 030
030 - 006
filesAdmin
AEMET
pluvioTermo
Network
CRONTAB
monthly retrieval
lag: 4 months
observations
obsHour 07:00
verification
Csv → Geo conversion
model grbFile
24hAccumPcp
upscaled
obs. grbFile
obs. geoFile
24hAccumPcp
SAL
UPS
repeat N days
SAL Diagram
SAL plot
Object plots
Grid template
(model file)
python
fortran90

Conclusions

- Models (ensembles, observations) are not perfect
  - Verification can reveal strong and weak points, assess **quality and value**, improve and guide forecaster
  - Verification against **observations (point or up-scaling)** or against **analysis**
  - Many techniques, each forecast type may need an specific method
- Observations are not perfect, either
  - Mesoscale, new high density observations (Radar, Seviri)
- Classical methods
  - **Statistical, scores-oriented** (e.g. bias, RMSE, ACC)
  - **Acumulation vs stratification**
  - **No unique, no perfect score**
- Limitations in classical methods:
  - Observational uncertainty
  - Sampling uncertainty
  - Seasonal stratification → Flow-dependent stratification
  - Extreme and severe weather
  - **Space and time scales, e.g.: double penalty** can give better scores to a coarser grid model, so new methods must be explored for mesoscale
- Software: very young

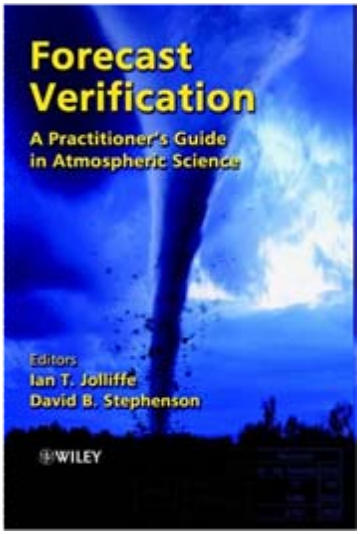
Conclusions

- New spatial verification methods for pcp: ICP
  - **Diagnostic verification** needed
  - Neighborhood, scale-separation, feature-based, field-deformation

- Example of feature-oriented novel method: **SAL**
  - Provides quantitative, detailed and accumulative information about different aspects of model QPF performance: **Structure-Amplitude-Location**
  - **Avoids double-penalty,** allows fair comparison of different resolution models without penalty for finer ones
  - Help understand (modellers and users) features improvement in the forecasts
  - AEMET - ECMWF: Research models 24h QPF SAL performance 2008 Central Europe and 2009 Spain

- Results look promising
  - GCMs and LAMs overall behaviour: overestimation of structure size (S), overestimation of pcp (A), location to improve (L), number of objects distribution shape to improve
  - Pcp threshold: Above 1mm much better performance not only on A, but also S and L
  - Model resolution: **finer models perform better**
- On-going work
  - Explore other clustering algorithms
  - Research on factor f for R* = f Rmax (regional sensitivity, introduce variability...)
  - Other patterns: seasonal, flow-dependent, etc.

References



Ian T. Jolliffe and D. B. Stephenson: **Forecast Verification**: A Practitioner's Guide in Atmospheric Science

John Wiley and Sons, Chichester (2003)

Forecast Verification – Issues, Methods and FAQ: JWGFVR/WWRP/WCRP

4470 MONTHLY WEATHER REVIEW

SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts

HEINI WERNLI AND MARCUS PAULAT
Institute for Atmospheric Physics, University of Mainz, Mainz, Germany
MARTIN HAGEN
Institut für Physik der Atmosphäre, DLR Oberpfaffenhofen, Germany
CHRISTOPH FREI
Federal Office of Meteorology and Climatology (MeteoSwiss), Zürich, Switzerland

(Manuscript received 12 October 2007, in final form 25 February 2008)

ABSTRACT

A novel object-based quality measure, which contains three distinct components that consider aspects of the structure (*S*), amplitude (*A*), and location (*L*) of the precipitation field in a prespecified domain (e.g., a river catchment) is introduced for the verification of quantitative precipitation forecasts (QPF). This quality measure is referred to as SAL. The amplitude component *A* measures the relative deviation of the domain-averaged QPF from observations. Positive values of *A* indicate an overestimation of total precipitation; negative values indicate an underestimation. For the components *S* and *L*, coherent precipitation objects are separately identified in the forecast and observations; however, no matching is performed of the objects in the two datasets. The location component *L* combines information about the displacement of the predicted (compared to the observed) precipitation field's center of mass and about the error in the weighted-average distance of the precipitation objects from the total field's center of mass. The structure component *S* is constructed in such a way that positive values occur if precipitation objects are too large and/or too flat, and negative values if the objects are too small and/or too peaked. Perfect QPFs are characterized by zero values for all components of SAL. Examples with both synthetic precipitation fields and real data are shown to illustrate the concept and characteristics of SAL. SAL is applied to 4 yr of daily accumulated QPFs from a global and finer-scale regional model for a German river catchment, and the SAL diagram is introduced as a compact means of visualizing the results. SAL reveals meaningful information about the systematic differences in the performance of the two models. While the median of the *S* component is close to zero for the regional model, it is strongly positive for the coarser-scale global model. Consideration is given to the strengths and limitations of the novel quality measure and to possible future applications, in particular, for the verification of QPFs from convection-resolving weather prediction models on short time scales.

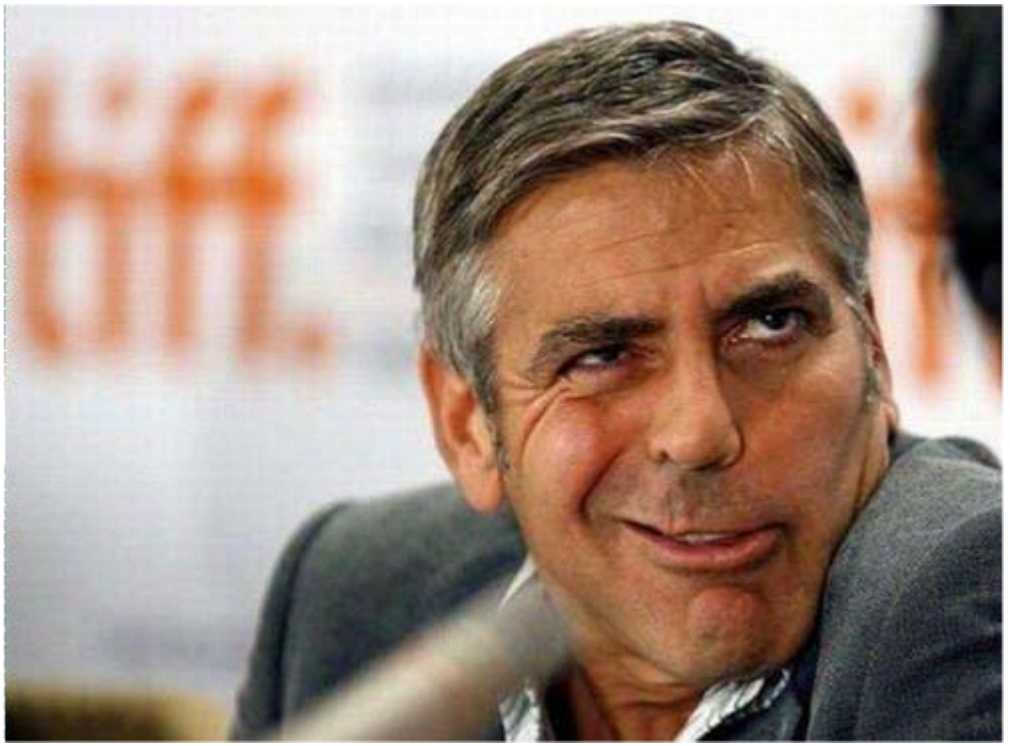
1. Introduction

Verification of numerical forecasts is an essential part of the numerical weather prediction (NWP) enterprise. On the one hand, it helps identify model shortcomings and systematic errors; on the other hand, it is key for a quantitative assessment of the im with time of current forecasting systems ar predictability limits. Quality measures like mean-square (RMS) difference or anomaly c are simple in terms of implementation and fore routinely used to monitor and compa forecast quality at operational prediction ce Simmons and Hollingsworth 2002). The quantitative precipitation forecasts (QPF) measured in terms of categorical verificat (Jolliffe and Stephenson 2003), a process th

Corresponding author address: Heini Wernli, Institute for Atmospheric Physics, University of Mainz, Becherweg 21, D-55099 Mainz, Germany.
E-mail: wernli@uni-mainz.de

DOI: 10.1175/2008MWR2415.1

Wernli et al , 2008: SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487, doi: 10.1175/2008MWR2415.1



Thanks