The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

Spring Colloquium on
'Regional Weather Predictability and Modeling'
April 11 - 22, 2005

1) Workshop **on Design and Use of Regional Weather Prediction Models,** *April 11 - 19*

2) Conference on **Current Efforts Toward Advancing the Skill of Regional Weather Prediction. Challenges and Outlook,** *April 20 - 22*

301/1652-21

Limited area ensemble prediction system of ARPA-SMR: COSMO-LEPS

S. Tibaldi
ARPA-IDRO/METEO
Bologna, Italy

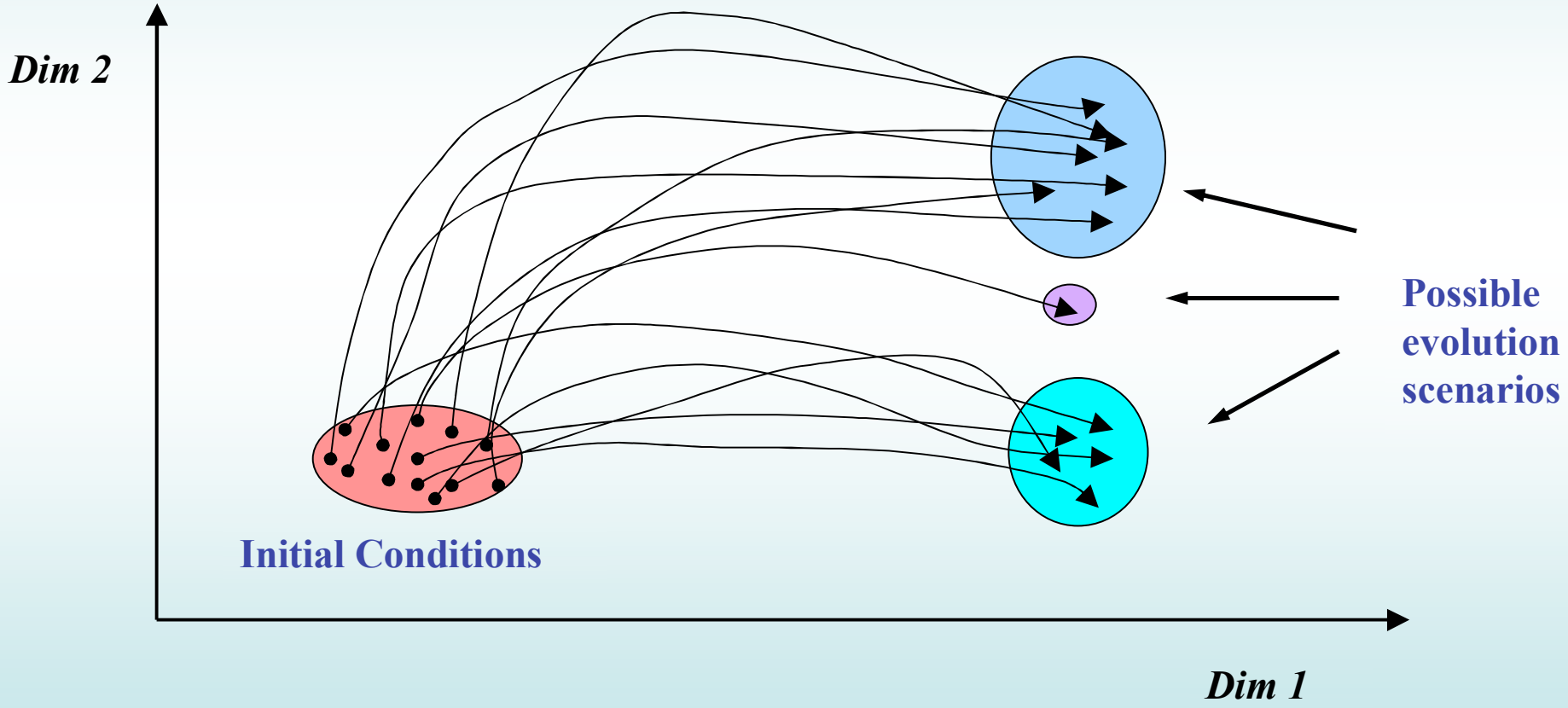
The limited area ensemble prediction system of ARPA-SMR: COSMO-LEPS

Andrea Montani, Chiara Marsigli, Tiziana Paccagnella and Stefano Tibaldi

ARPA-SIM, Bologna, Italy

ECMWF EPS

Simplified atmospheric phase space



THE NEED: REGIONALISATION OF SCENARIOS

- "Brute force" approach:
  - **T1000 ensemble with 100 members (half done?)**
  - **one LAM integration for each ensemble member**
- ARPA-SMR approach:
  - **ensemble size reduction**
  - **concept of "most significant member"**
  - **only a few LAM high- resolution runs needed**

THE LEPS APPROACH

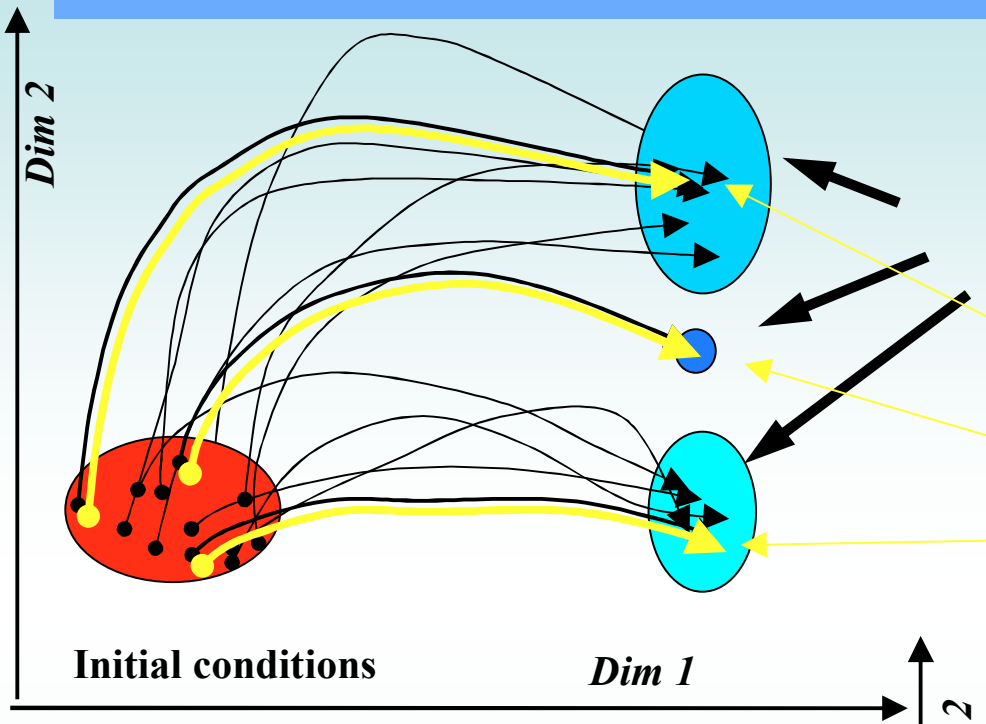
The LAM is nested in only a limited number of members selected from the global EPS, the <u>Representative Members</u>

Some of the information from global EPS is lost BUT the operation becomes feasible on an operational basis

Most Representative Member

- one per cluster
- choice is based on selected 3D fields: has to be: the closest to the mean of its own cluster AND the most distant to the other clusters' means
- 5 (or 10) runs instead of 51, 102 or 153!!

LEPS – Limited area Ensemble Prediction System

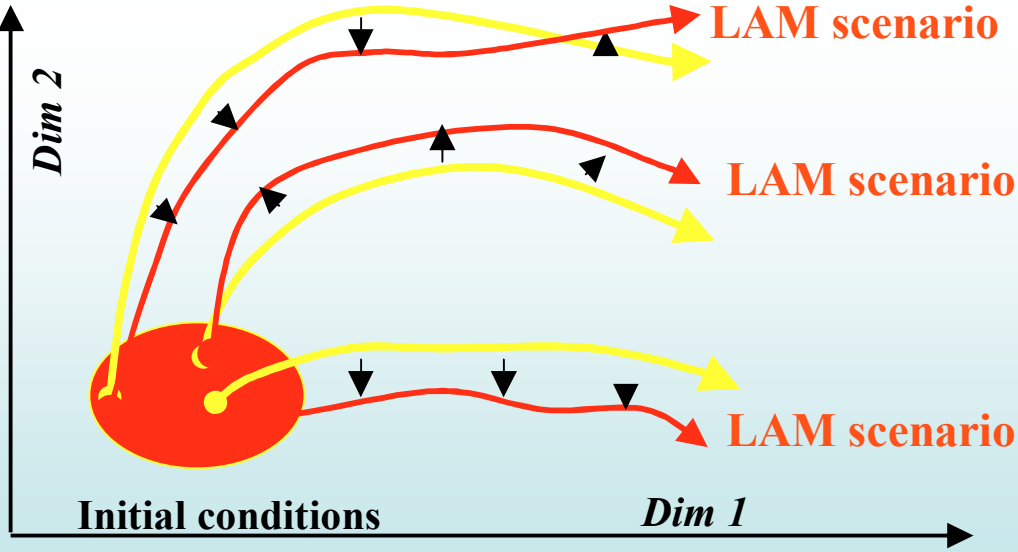


Possible evolution scenarios

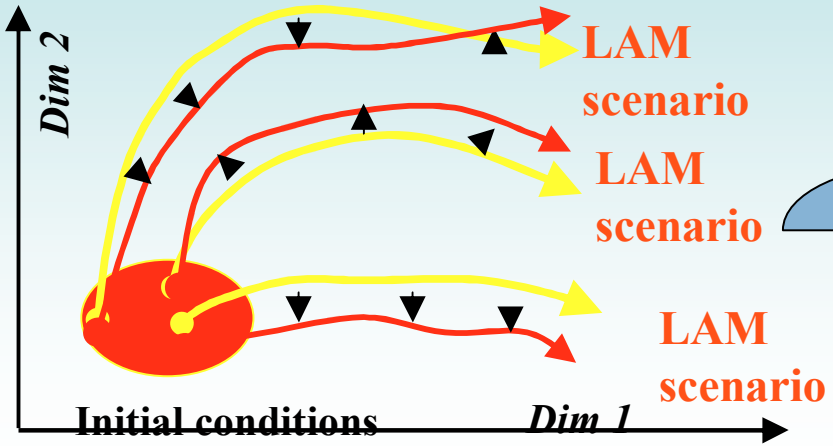
EPS and ensemble size reduction

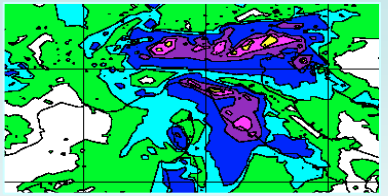
Cluster members chosen as representative members (RMs)

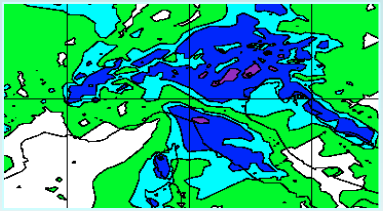
LAM integrations driven by RMs



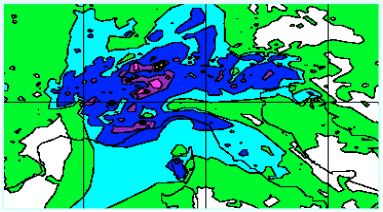
LEPS – Limited area Ensemble Prediction System



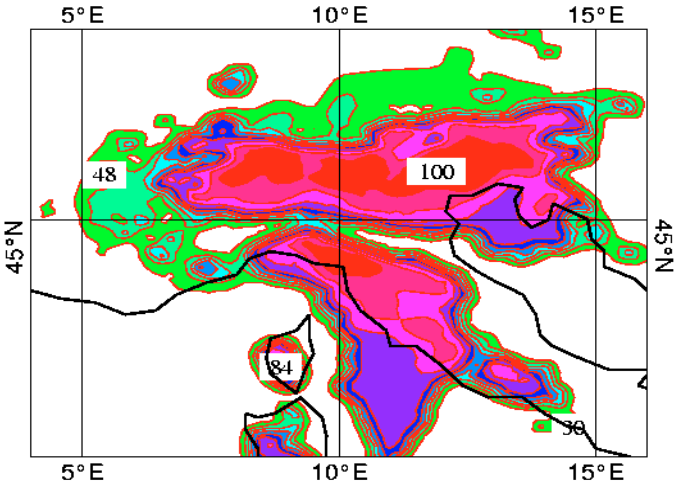


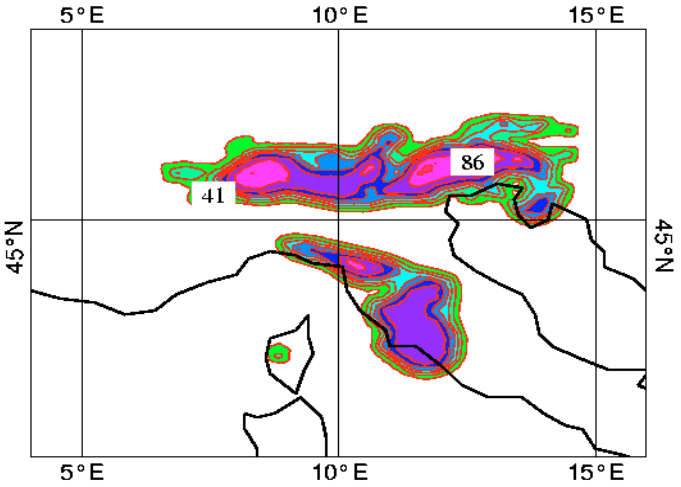


precipitation scenarios



PROBABILITY MAPS: WEIGHTING?





www.cosmo-model.org

A COSMO aderiscono

Germania, Svizzera, Italia, Grecia e Polonia

COSMO è finalizzato allo sviluppo e alla gestione operativa del modello non idrostatico

LAMI (Limited Area Model of Italy)

Deutscher Wetterdienst

MeteoSwiss

Ufficio Generale per la Meteorologia

Hellenic National Meteorological Service

Amt für Wehrgeophysik

Il Servicio Meteorologico Regionale di ARPA

LAMI
LAM Italia

COSMO

UGM

ARPA-SMR

LIGURIA

PIEMONTE

LOMBARDIA

TRENTO

VENETO

MARCHE

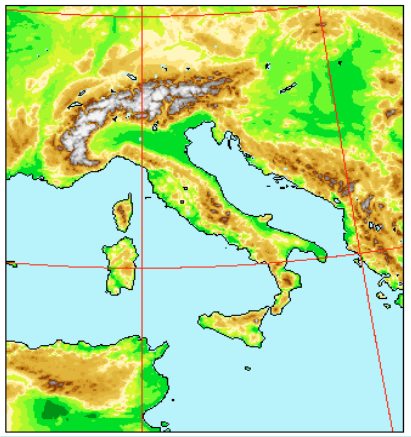
AMPANIA

CALABRIA

DPCN

SUITE LAMI 2001/2002/2003

CINECA
IBM - SP3 - 64 PE



COSMO-LEPS (developed at ARPA-SIM)

- What is it?

It is a **L**imited-area **E**nsemble **P**rediction **S**ystem (**LEPS**), based on Lokal Modell and developed within ***COSMO*** (***CO***nsortium for **S**mall-scale ***MO***delling, which includes Germany, Greece, Italy, Poland and Switzerland).

- Why?

Because the horizontal resolution of global-model ensemble systems is limited by computer time constraints and does not allow a detailed description of mesoscale and orographic-related processes.

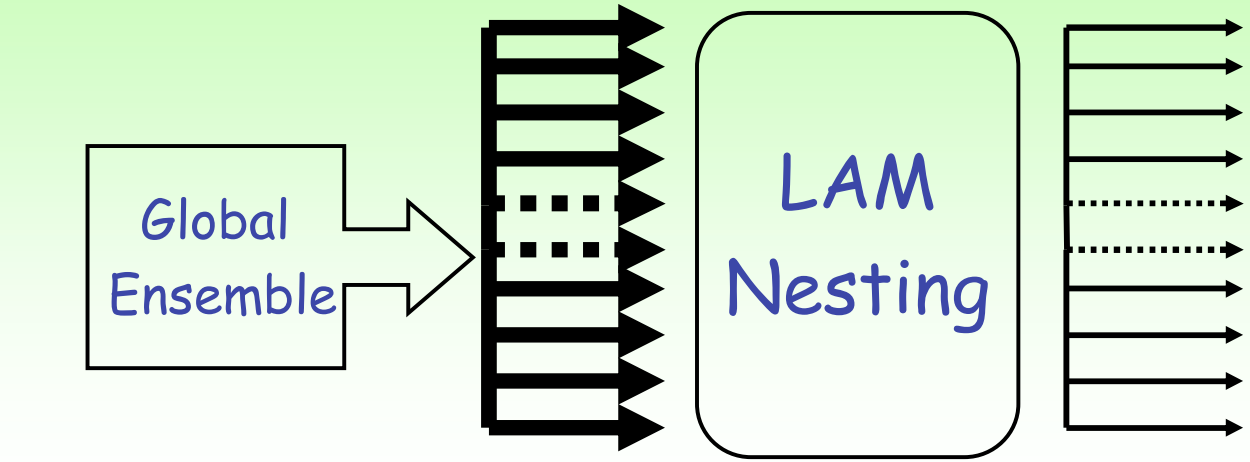
The forecast of heavy precipitation events is still inaccurate (in terms of both locations and intensity) after the short range.

COSMO-LEPS project

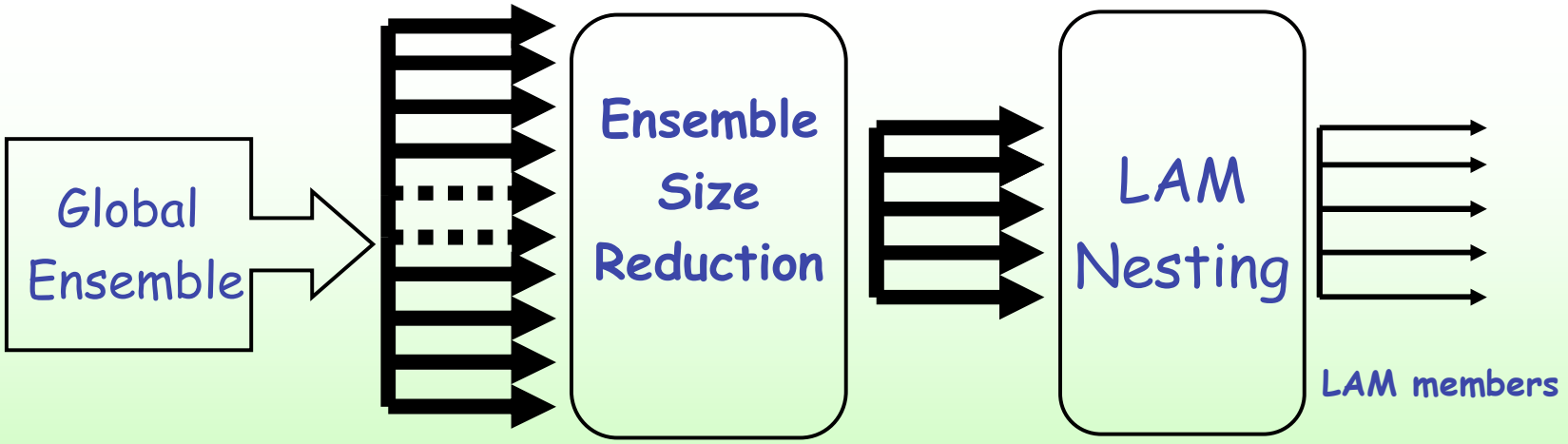
➔ combine the advantages of global-model ensembles with the high-resolution details gained by the LAMs, so as to identify the possible occurrence of **intense** and **localised** weather events (heavy rainfall, strong winds, temperature anomalies, snowfall, …);

generation of COSMO-LEPS in order to improve the Late-Short (48hr) to Early-Medium (120hr) range forecast of the so-called "severe weather events".

Downscaling

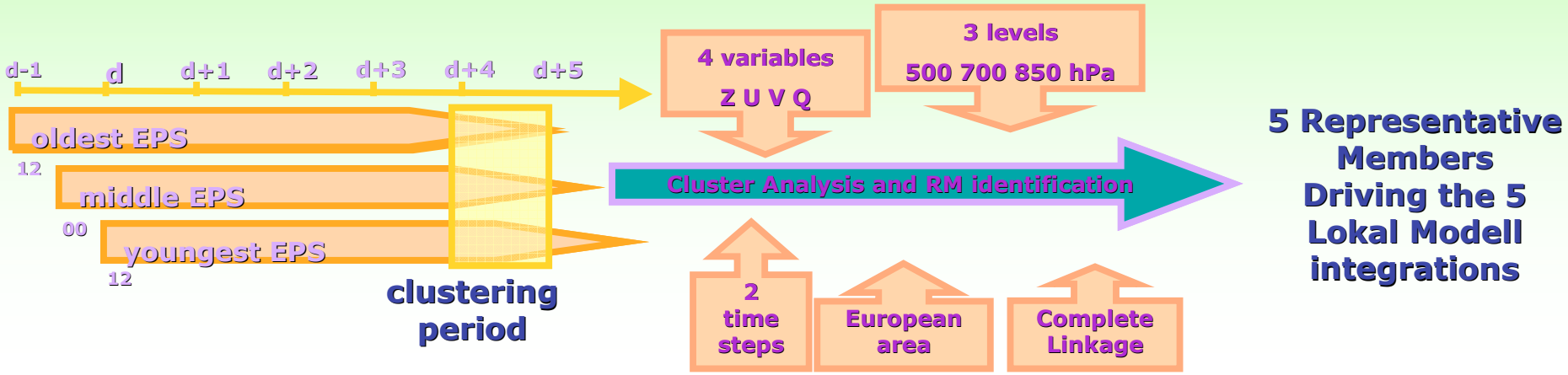


Total Downscaling (BFA)

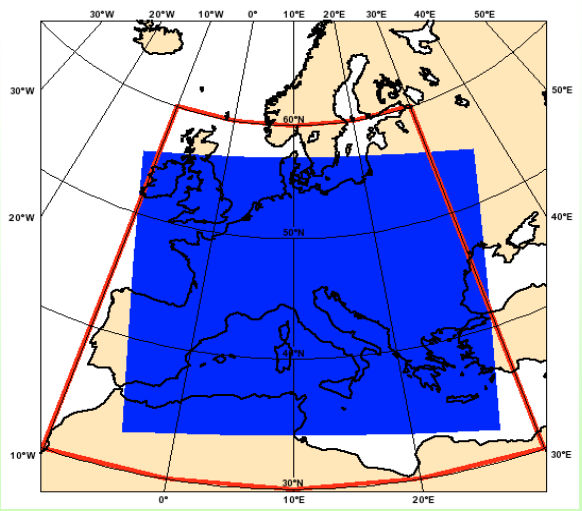


COSMO LEPS approach

The COSMO-LEPS suite @ ECMWF
November 2002 - May 2004

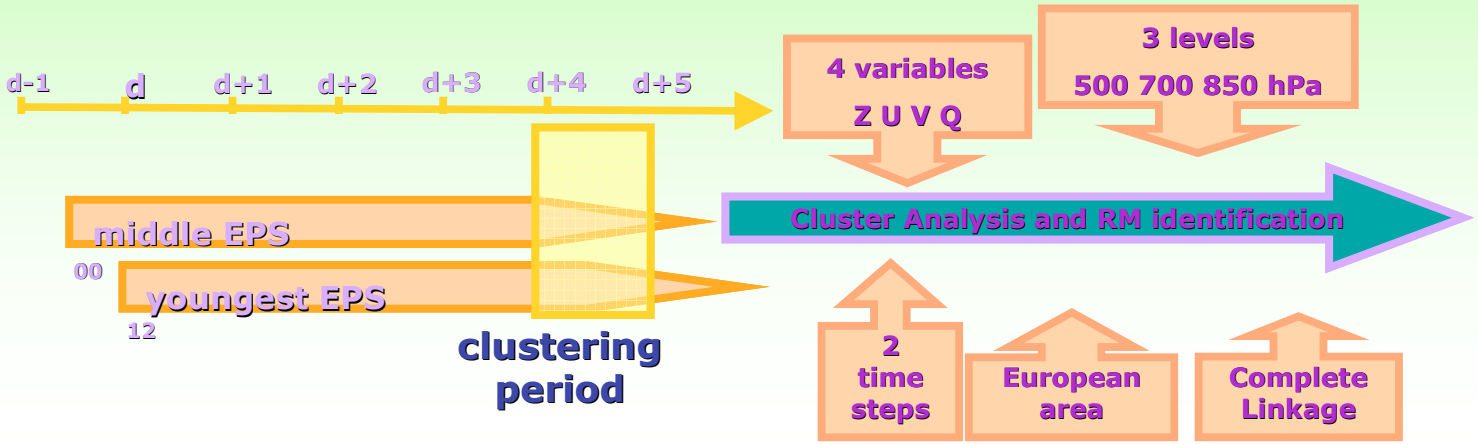


COSMO-LEPS clustering area



COSMO-LEPS Integration Domain

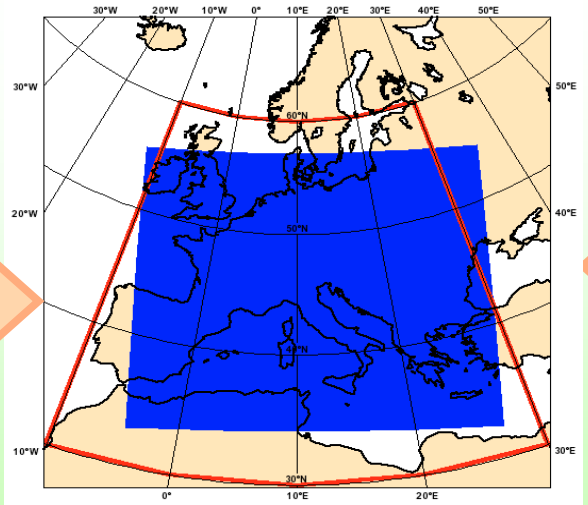
The COSMO-LEPS suite @ ECMWF since June 2004



10 Representative Members driving the 10 Lokal Modell integrations (weighted according to the cluster populations)

employing either Tiedtke or Kain-Fristch convection scheme (randomly choosen)

COSMO-LEPS clustering area



COSMO-LEPS Integration Domain

- suite running every day at ECMWF managed by ARPA-SIM;
- Δx ~ 10 km; 32 ML;
- fc length: 120h;
- Computer time provided by the COSMO partners which are ECMWF member states.

Operational COSMO-LEPS set-up

Core products

➔ 10 *perturbed* LM runs (ICs and 6-hourly BCs from 10 EPS members) to generate probabilistic output (start at 12UTC; Δt = 120h);

Additional products

➔ 1 *reference* run (ICs and 6-hourly BCs from the high-resolution deterministic ECMWF forecast) to assess the relative merits between deterministic and probabilistic approach (start at 12UTC; Δt = 120h);

➔ 1 *proxy* run (ICs and 3-hourly BCs from ECMWF analyses) to "downscale" ECMWF information (start at 00UTC; Δt = 36h).

Dissemination to the COSMO community

Products disseminated to the COSMO–countries

probabilistic products:

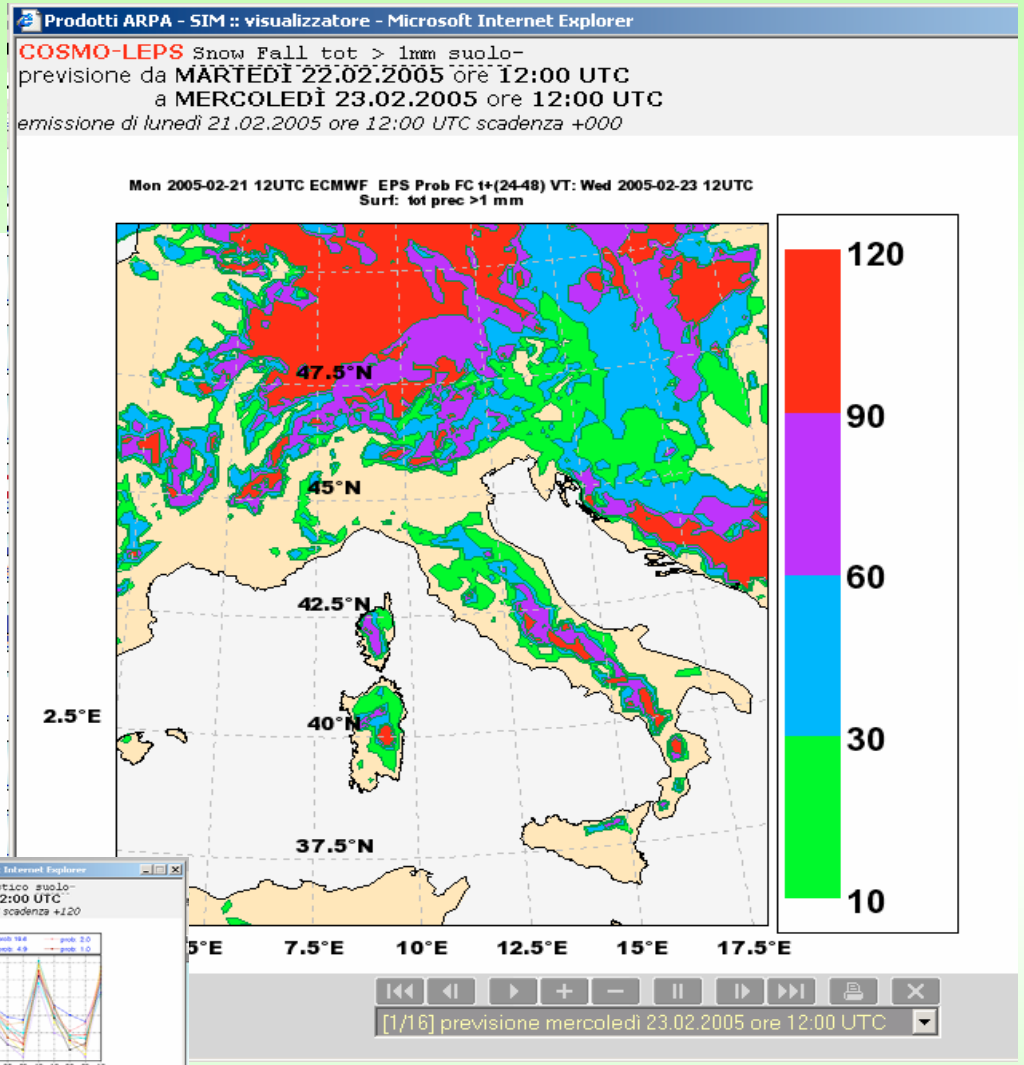
- 24h rainfall exceeding 20, 50, 100, 150 mm;
- 72h rainfall exceeding 50, 100, 150, 250 mm;
- 24h snowfall exceeding 1, 5, 10, 20 "**cm**";
- $UVmax_{10m}$ in 24h above 10, 15, 20, 25 m/s;
- $Tmax_{2m}$ in 24h above 20, 30, 35, 40 ^{0}C;
- $Tmin_{2m}$ in 24h below -10, -5, 0, +5 ^{0}C;
- min height of 0 ^{0}C isotherm in 24h below 1500, 1000, 700, 300 m;
- max–CAPE in 24h above 2000, 2500, 3000, 3500 J/kg;
- min Showalter Index in 24h below 0, -2, -4, -6;

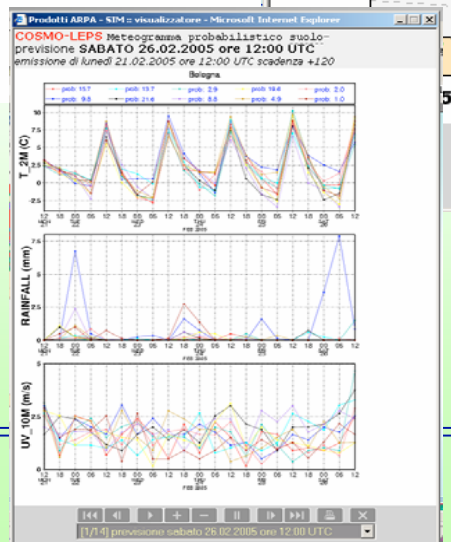
deterministic products (for each LM run):

- 24–hour cumulated rainfall; mean–sea–level pressure, Z700, T850;

meteograms (over a number of station points):

- T_{2m}, rainfall, 10m wind speed.





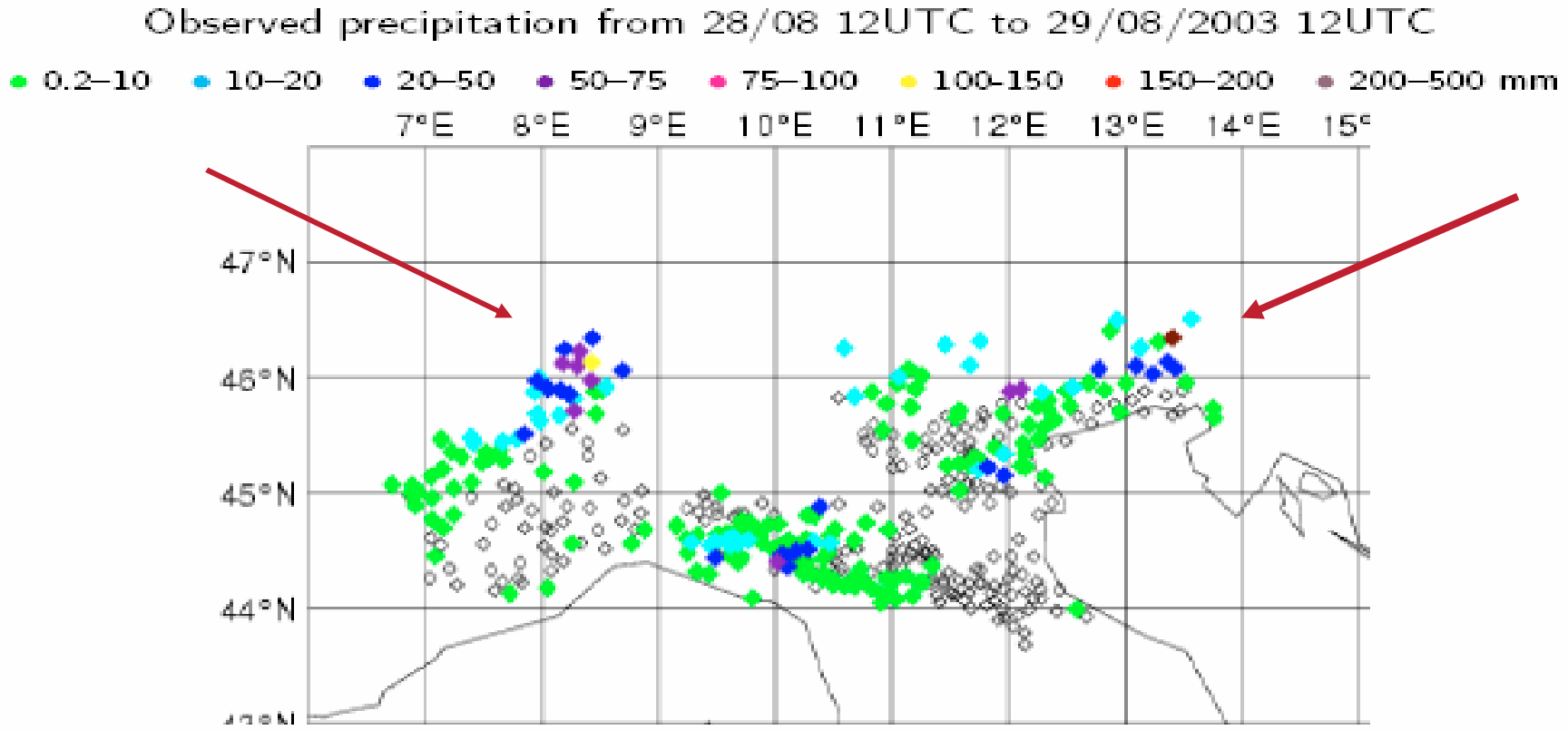
Operational COSMO-LEPS ~ Operational EPS

S.E. 153
5 RMs

"Friuli case"

The youngest EPS

Case study: Friuli–Ticino flood

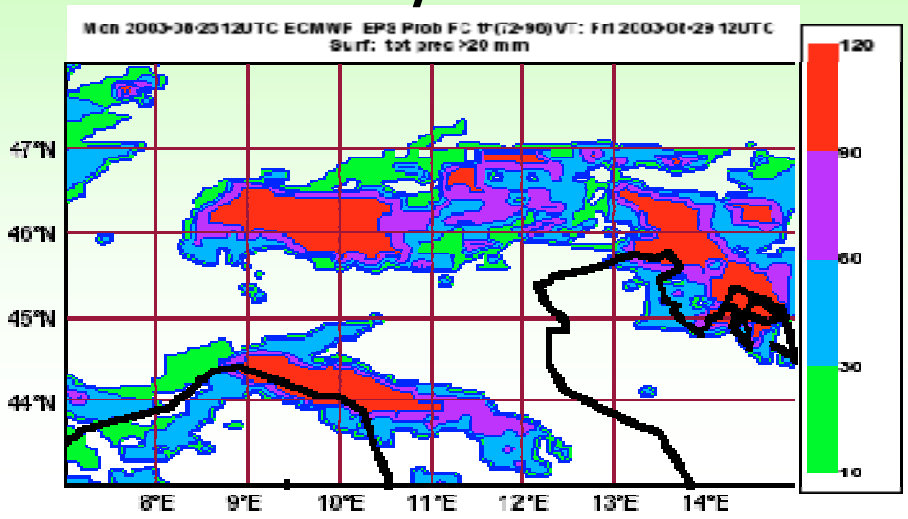


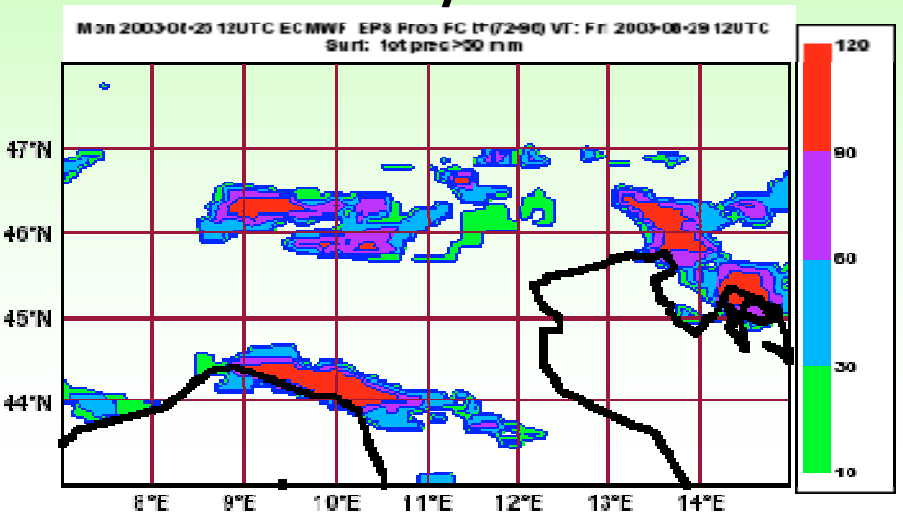
Operational COSMO-LEPS ~ Operational EPS : Friuli case probability maps – fc. Range: +96h

>20mm/24h

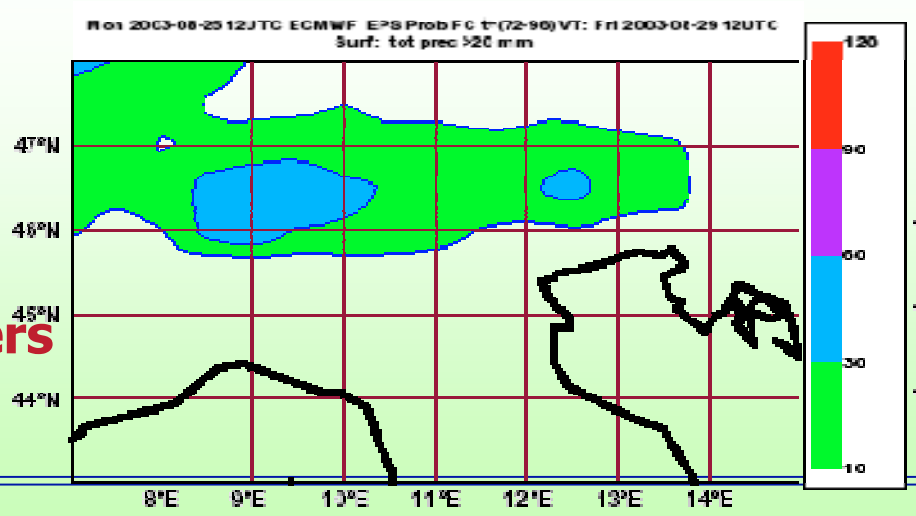
>50mm/24h

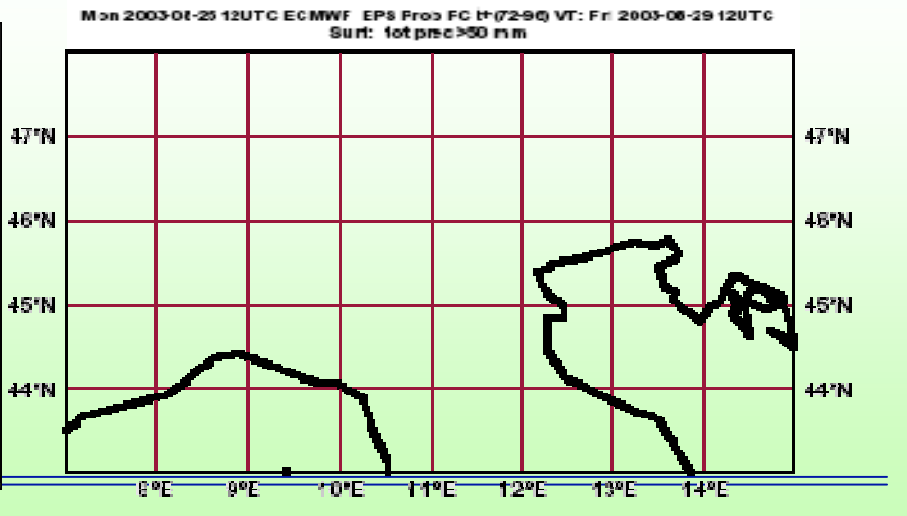
COSMO LEPS





EPS 51 members





COSMO-LEPS ongoing activities

EVALUATION OF THE METHODOLOGY with respect to:

- **ENSEMBLE SIZE REDUCTION**
- **SUPER-ENSEMBLE SIZE**
- **CLUSTERING SETTING (parameters, time range, areas)**
- impact of weighting
- **ADDED VALUE WITH RESPECT TO EPS**

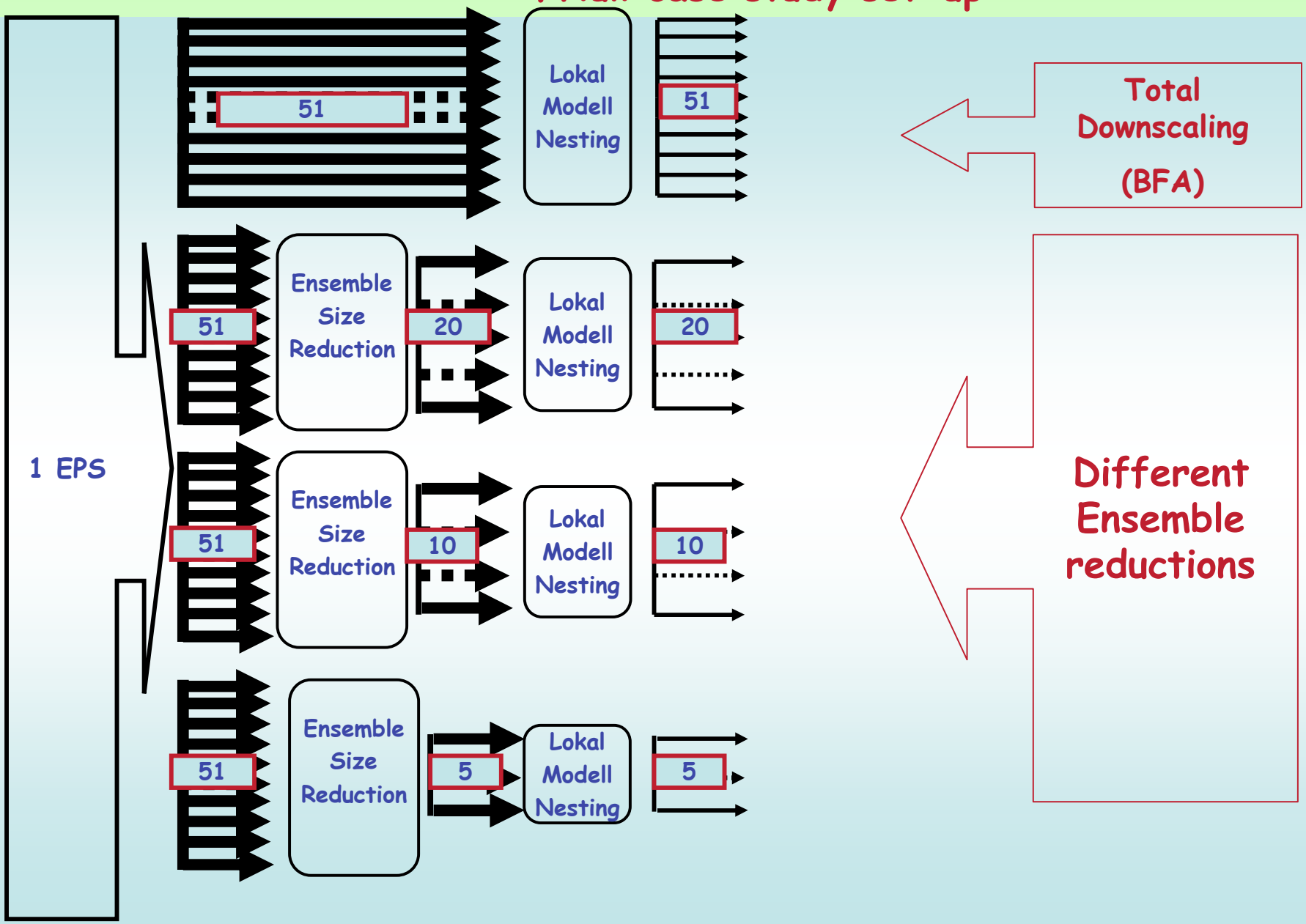
OBJECTIVE VERIFICATION OF COSMO-LEPS

- **ADDED VALUE WITH RESPECT TO EPS**
- evaluated at different spatial scales
- evaluated over different geographical regions
- evaluated for two different convection schemes

2 related ECMWF Special Projects ongoing:

- SPITLAEF in cooperation with UGM
- SPCOLEPS in cooperation with Meteo-Swiss

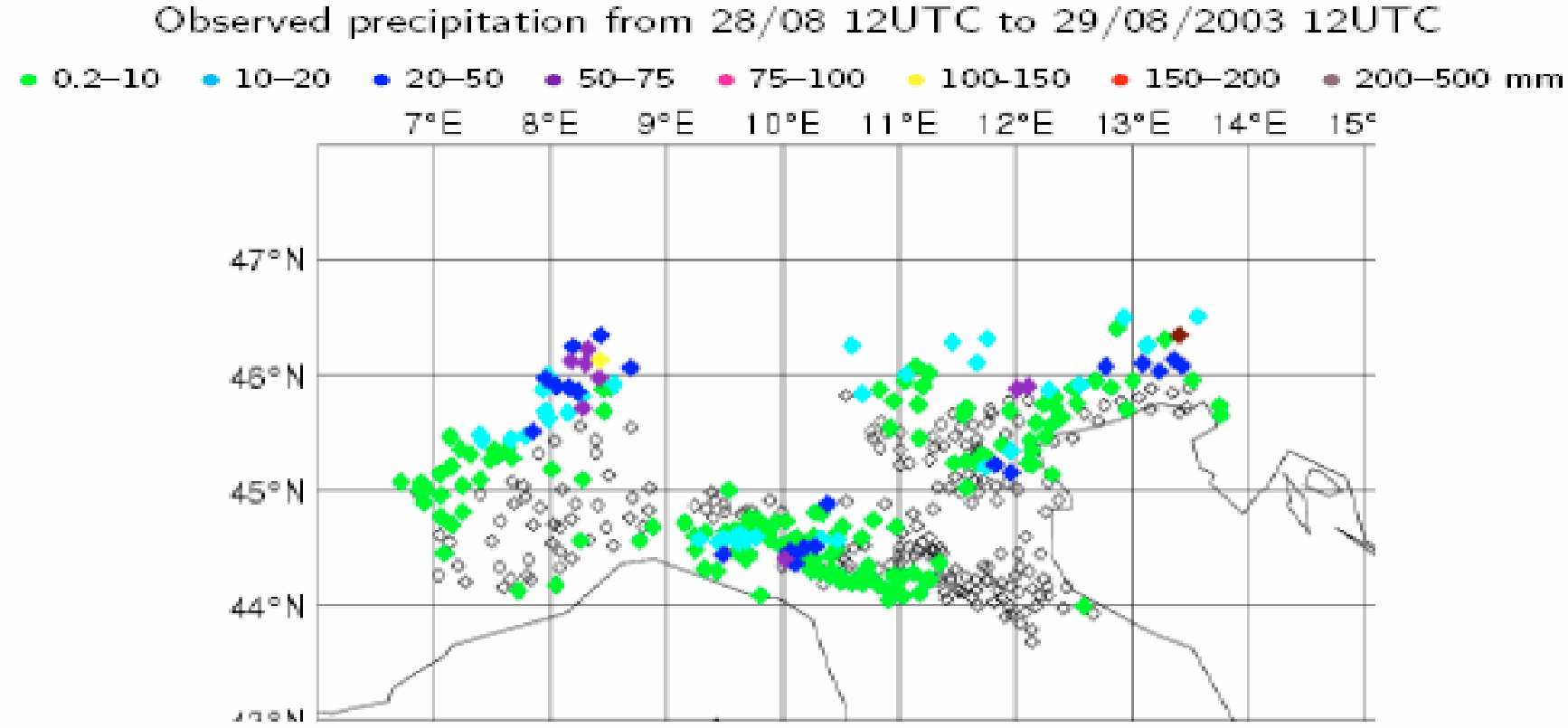
ENSEMBLE SIZE REDUCTION:
Friuli case study set-up



ENSEMBLE SIZE REDUCTION

IMPACT EVALUATED ON CASE STUDIES (1)

Case study: Friuli–Ticino flood



2003082512 Friuli

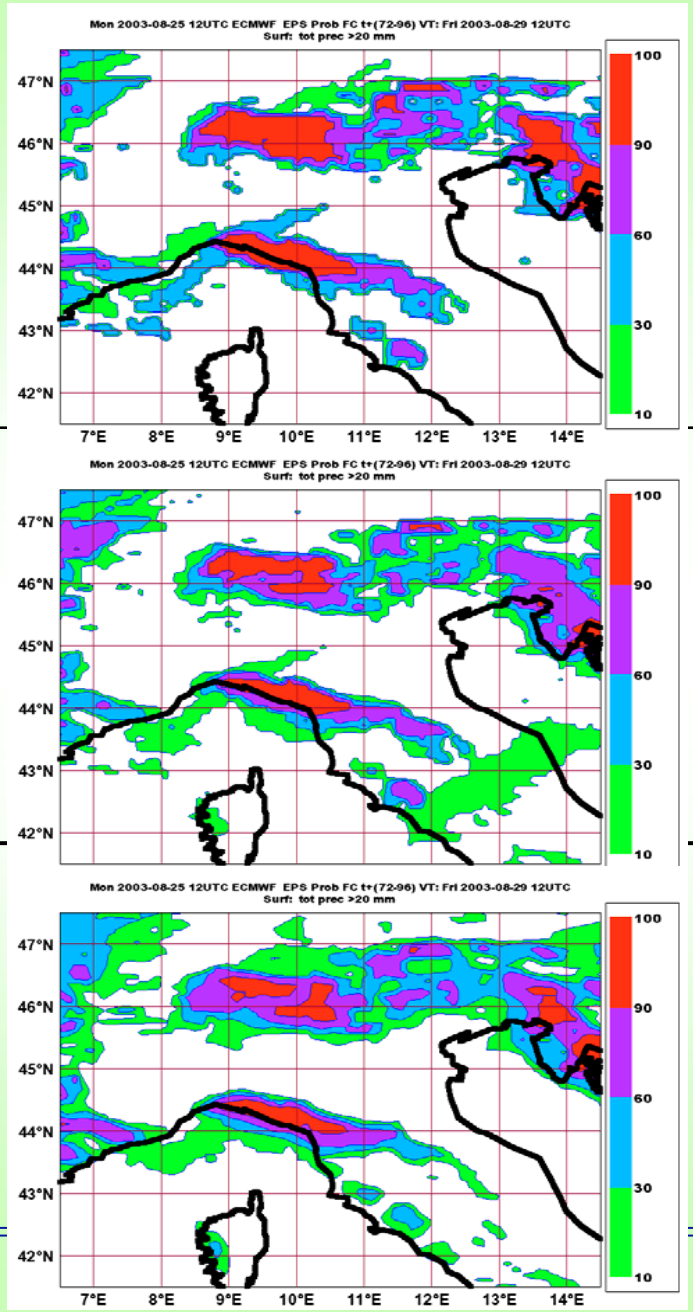
(fc+72-96h)

5 RMs

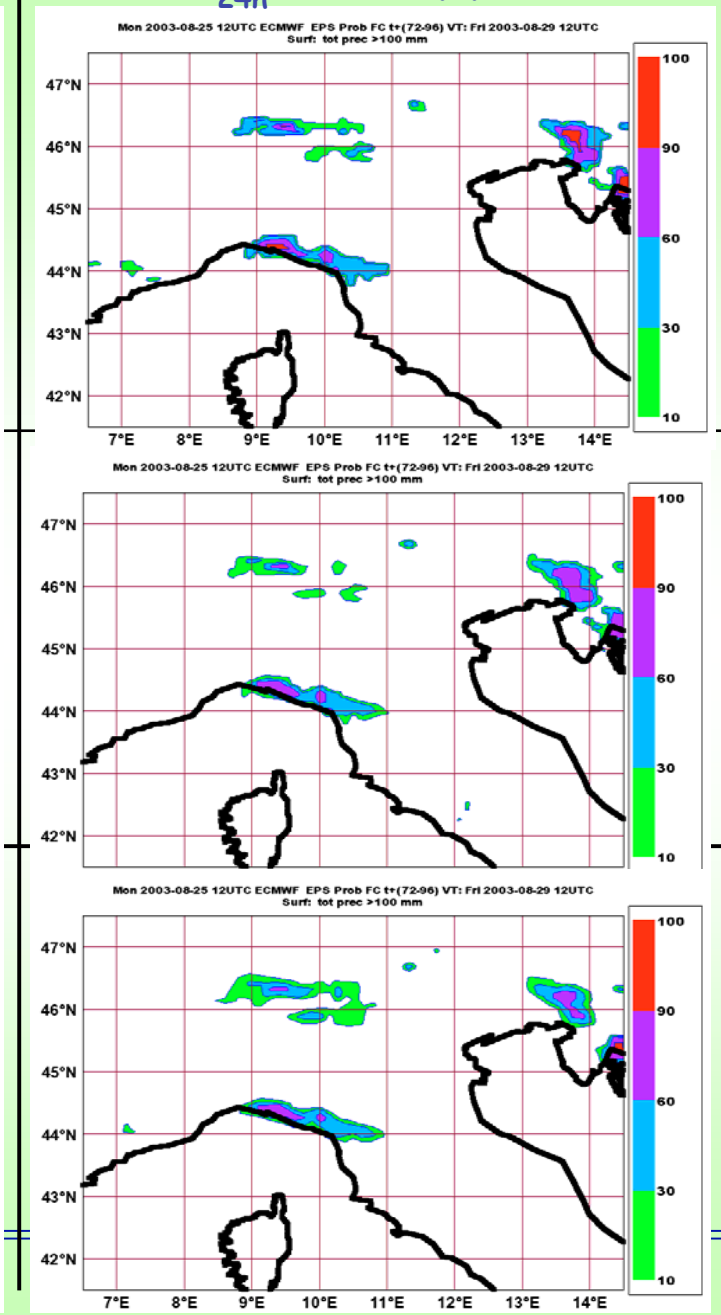
10 RMs

All 51

TP_{24h} > 20 mm



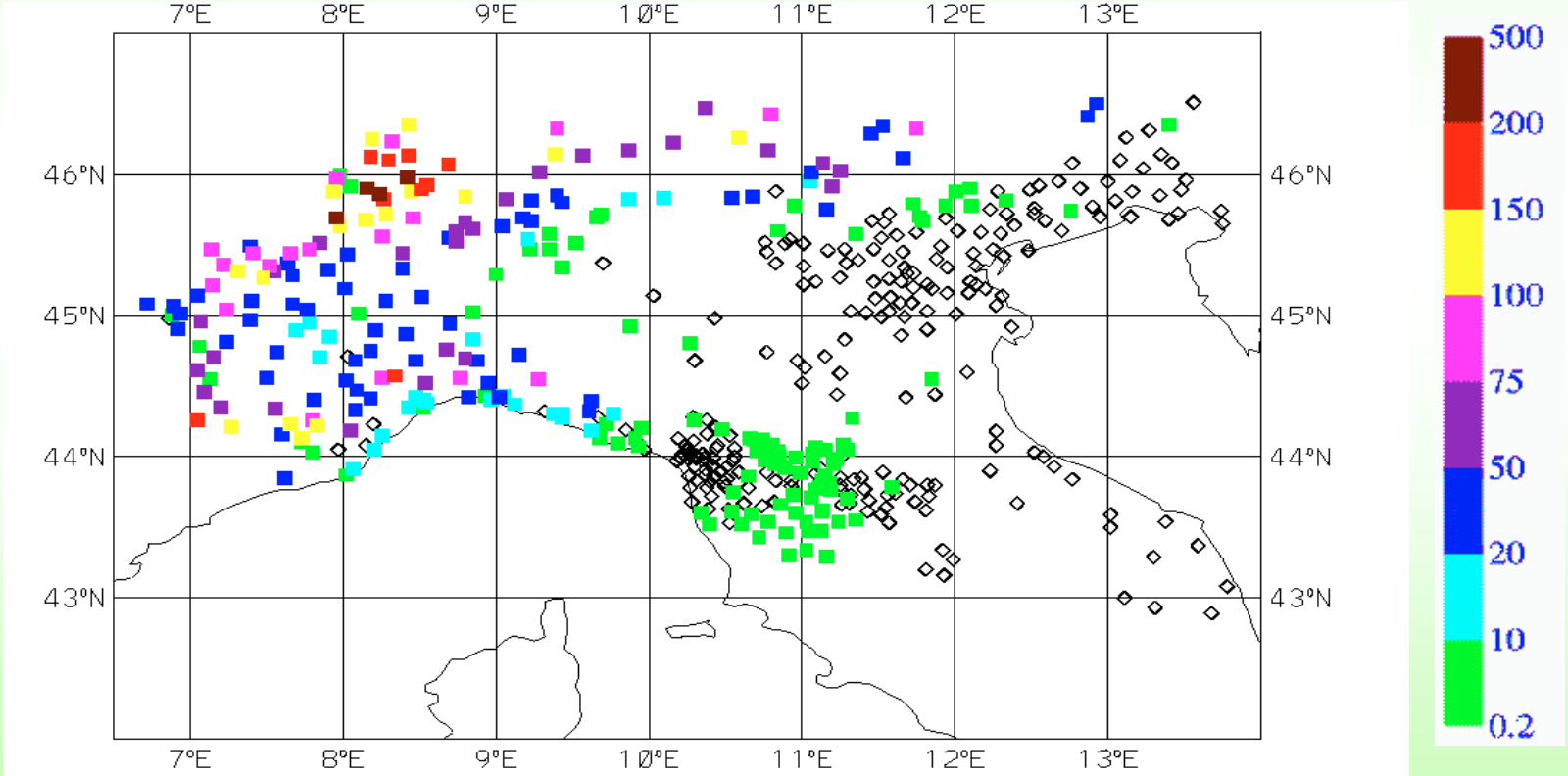
TP_{24h} > 100 mm



ENSEMBLE SIZE REDUCTION
IMPACT EVALUATED ON CASE STUDIES (2)

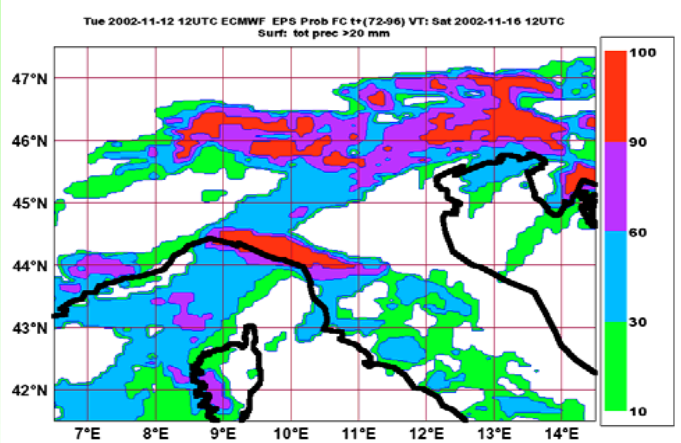
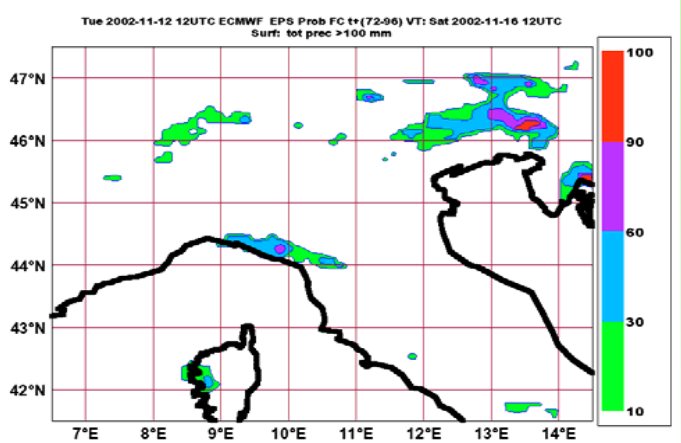
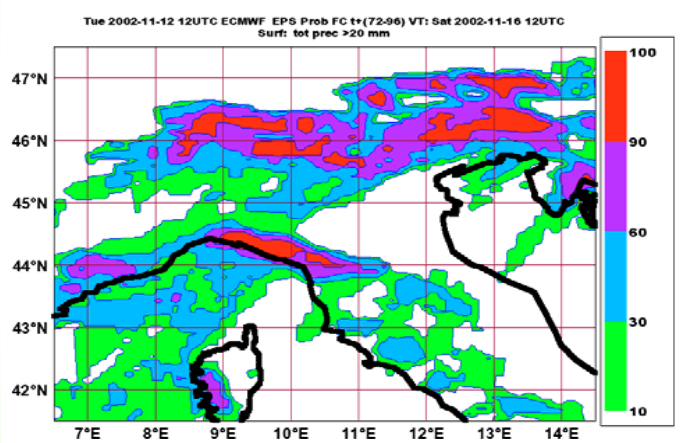
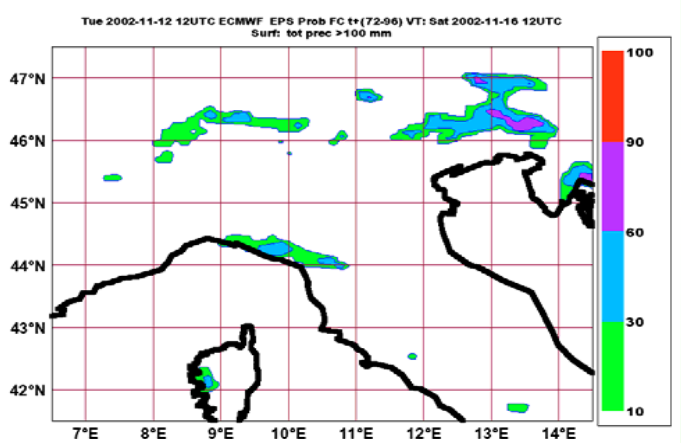
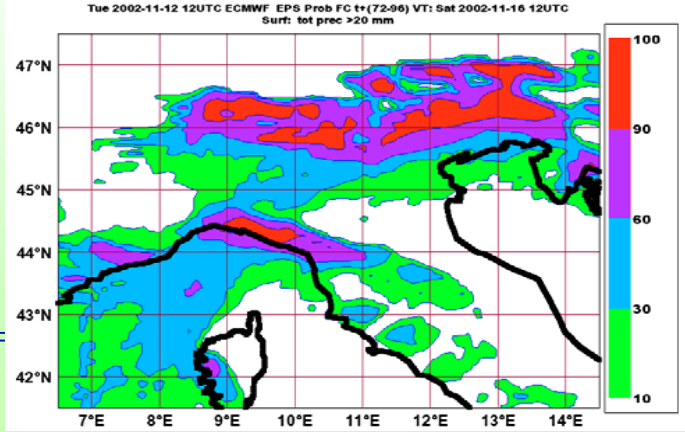
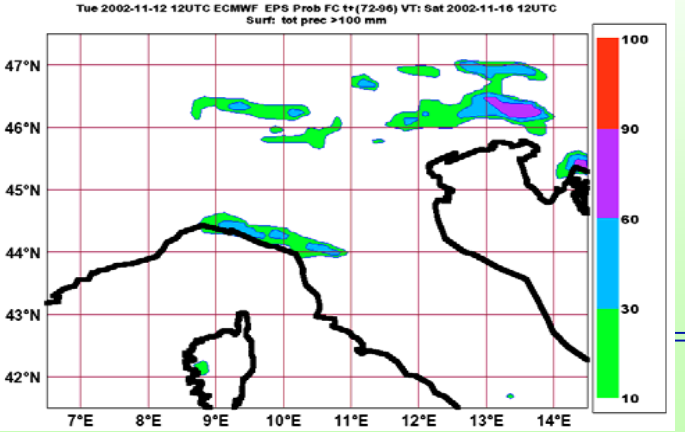
Observed precipitation between 15-11-2002 12UTC and 16-11-2002 12 UTC

Piedmont case

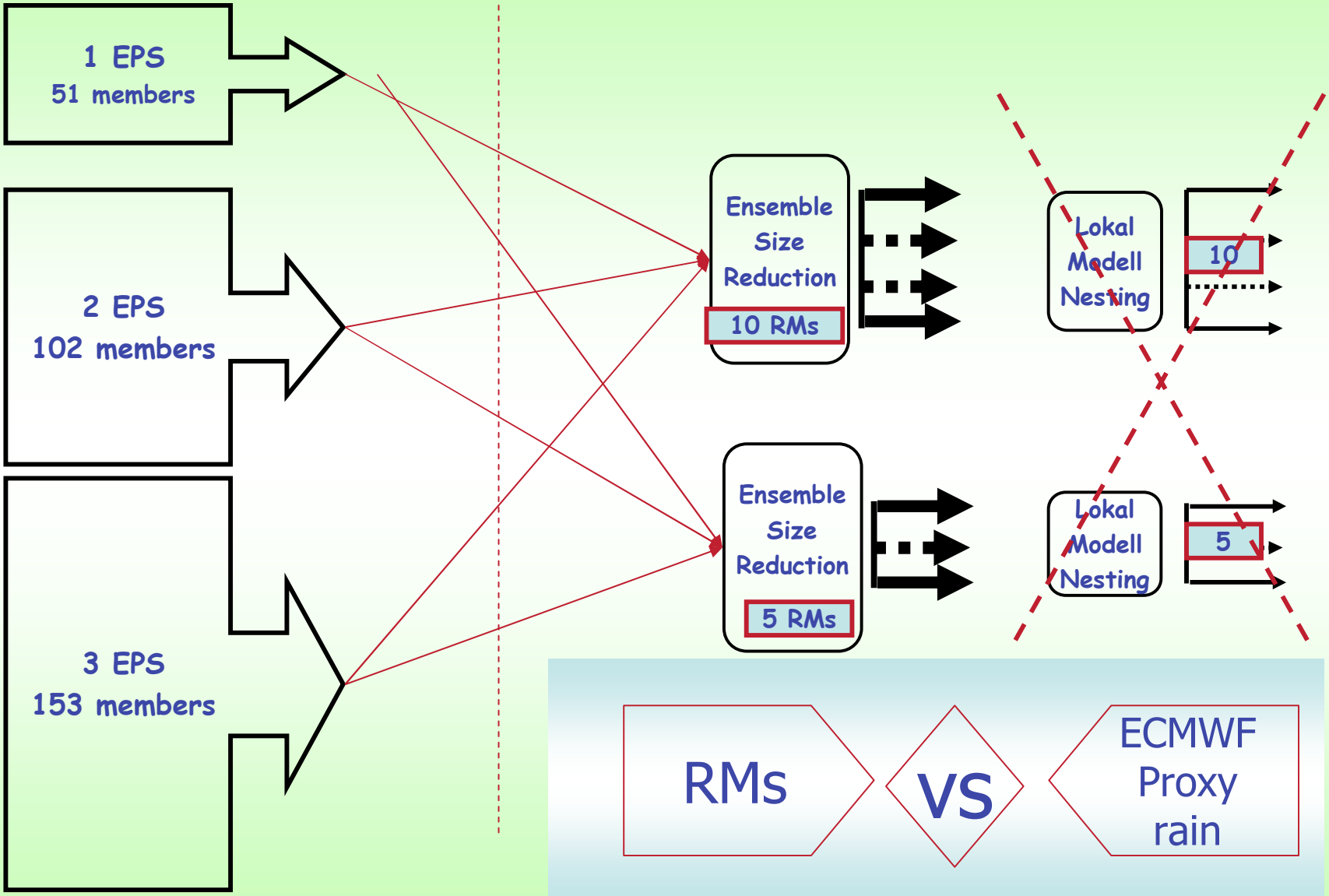


2002111212 Piedmont

(fc+72-96)

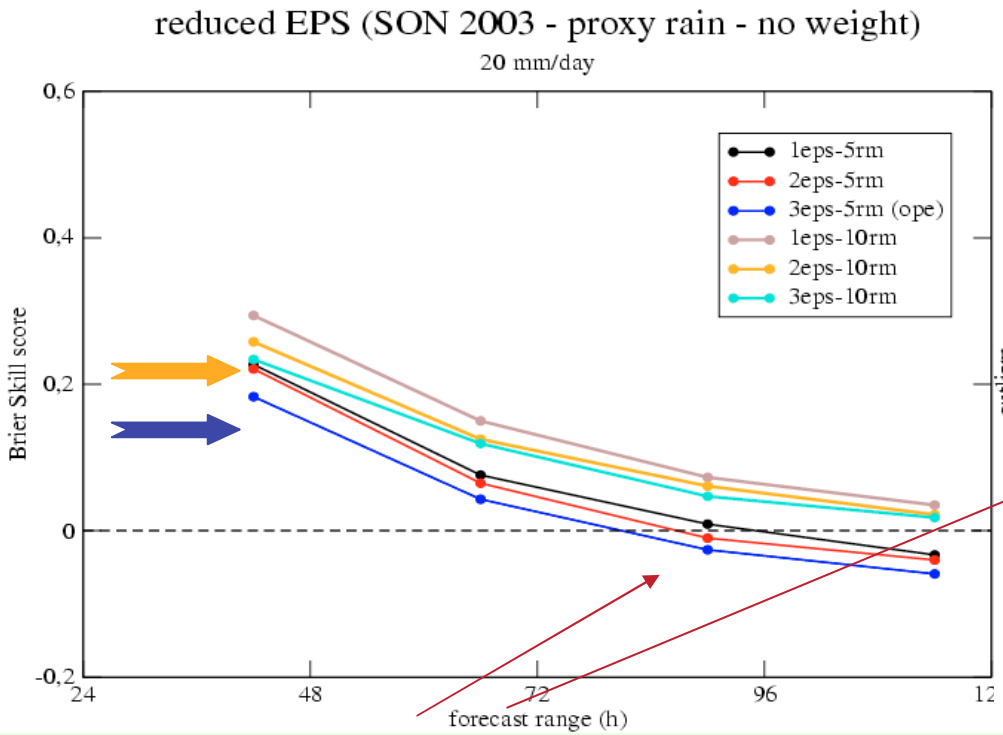
	20 mm	150 mm
5 RMs		
10 RMs		
All 51		

EVALUATION OF SUPER-ENSEMBLE (S.E.) SIZE & ENSEMBLE SIZE REDUCTION

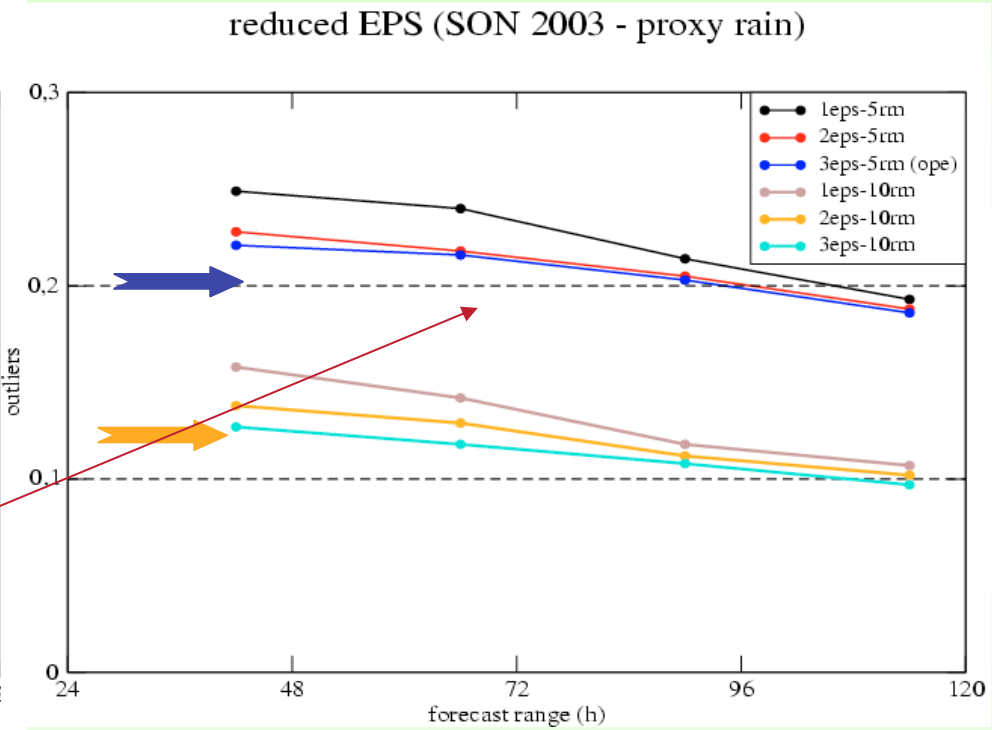


EVALUATION OF S.E. SIZE (either 51, or 102, or 153) & ENSEMBLE SIZE REDUCTION (either 5 or 10 RMs)

BSS



outliers



- Regarding the 5-member ensembles, results seem to suggest that the use of just two EPSs in the super-ensemble can be a reasonable compromise, permitting to decrease the percentage of outliers significantly (with respect to the use of 1 EPS), "paying" only a small decrease of the skill.

- Regarding the impact of the ensemble size, the difference between each 5-member ensemble and the correspondent 10-member ensemble is remarkable. The impact of doubling the ensemble size is almost the same for every configuration and is larger than the impact of changing the number of EPSs on which the Cluster Analysis is performed (either 2 or 3).

TEST OF DIFFERENT CLUSTERING INTERVALS

- Consider a fixed configuration in terms of ensemble size (10 RMs selected out of 2 EPS sets, **2eps-10rm**) and the properties of the "reduced" (10-member) global ensemble in 4 different cases:

OPE: the 10 members are selected like in the operational set-up (clustering variables: z,u,v,q; clustering levels: 500, 700, 850 hPa; clustering times: fc+96h, fc+120h);

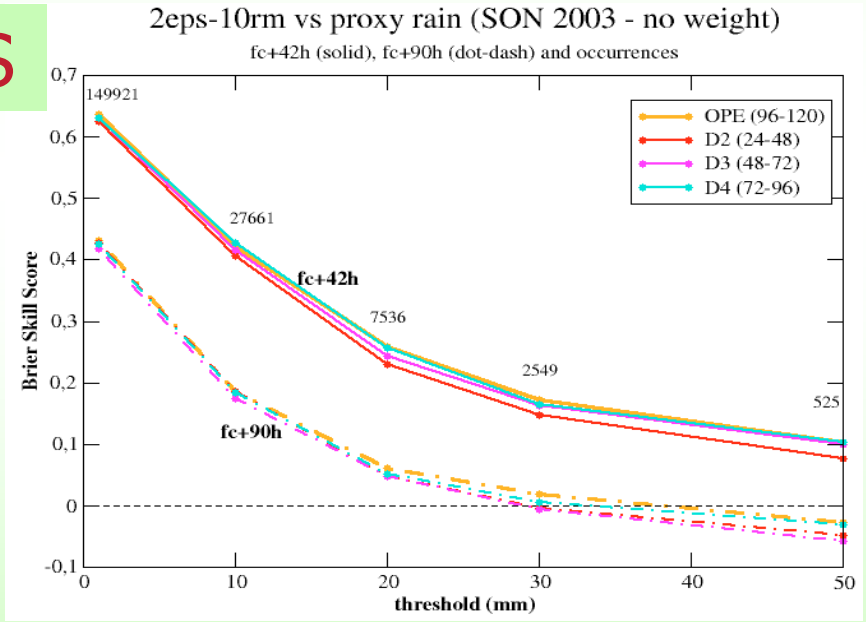
D2: like OPE, but clustering times: fc+24h, fc+48h;

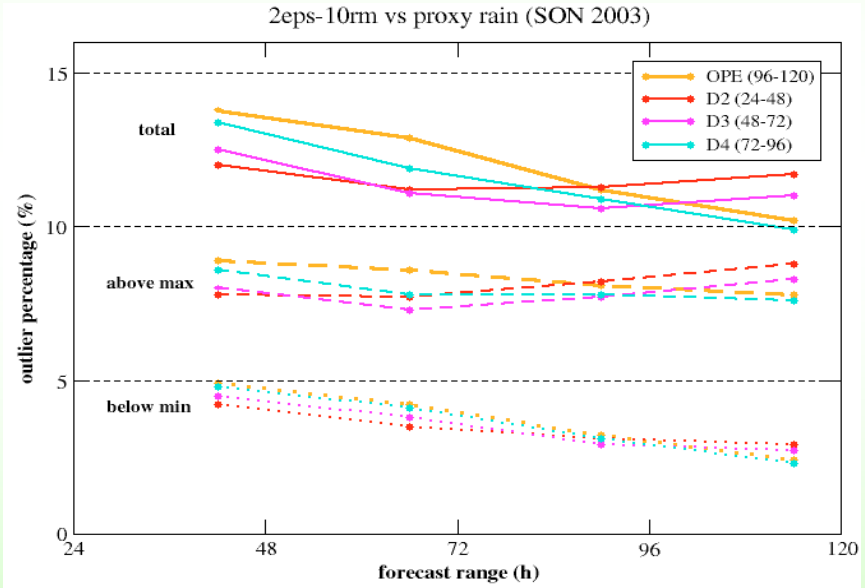
D3: like OPE, but clustering times: fc+48h, fc+72h;

D4: like OPE, but clustering times: fc+72h, fc+96h.

outliers

BSS





Brier Skill Score: OPE has slightly better scores at all verification ranges *(less evident for ROC area .. not shown);*

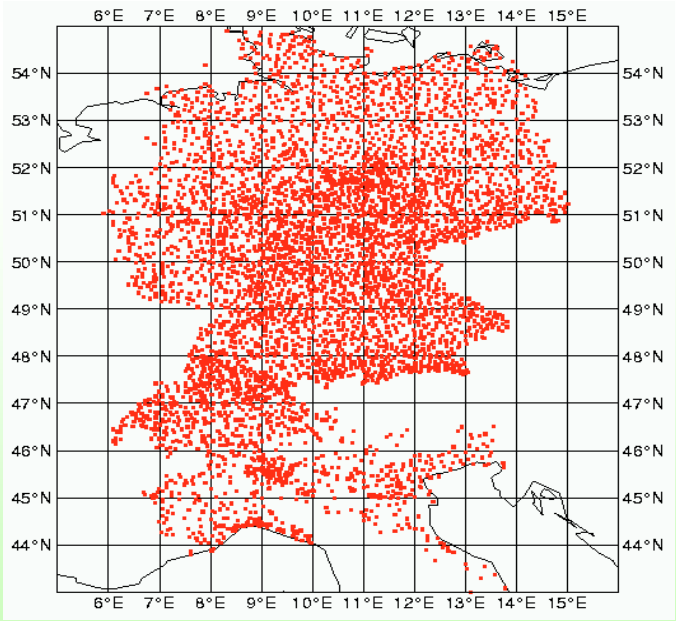
Outliers percentage: results heavily depend on the verification range.

OBJECTIVE VERIFICATION OF COSMO-LEPS

A verification package was developed keeping into account two measures of precipitation:

- the cumulative volume of water deployed over a specific region,
- the rainfall peaks which occur within that region.

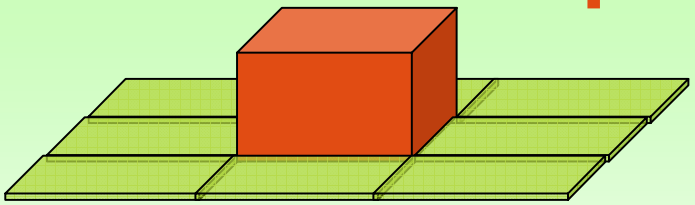
COSMO observations



The verification package includes the traditional probabilistic scores:

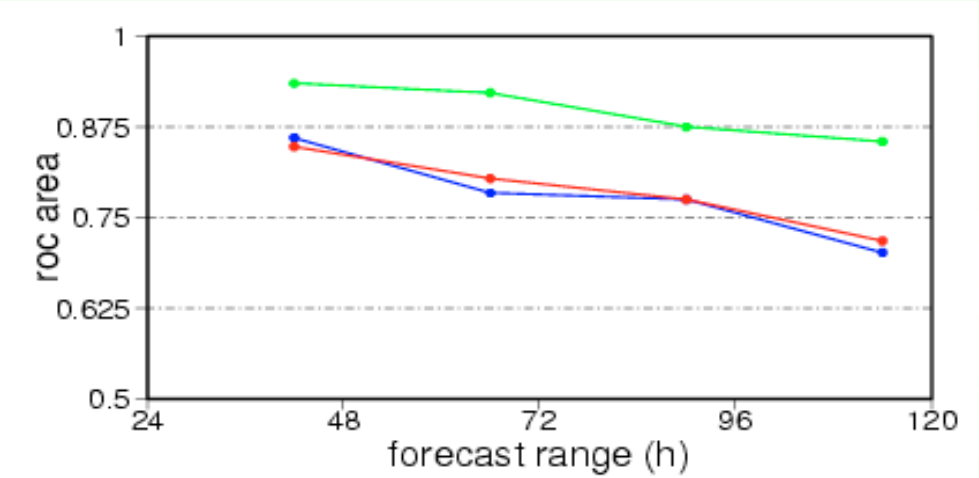
- Brier Skill Score (Wilks, 1995)
- ROC area (Mason and Graham, 1999)
- Cost-loss Curve (Richardson, 2000)
- Percentage of Outliers (Buizza, 1997)

Precipitation: average over 1.5 x 1.5 boxes

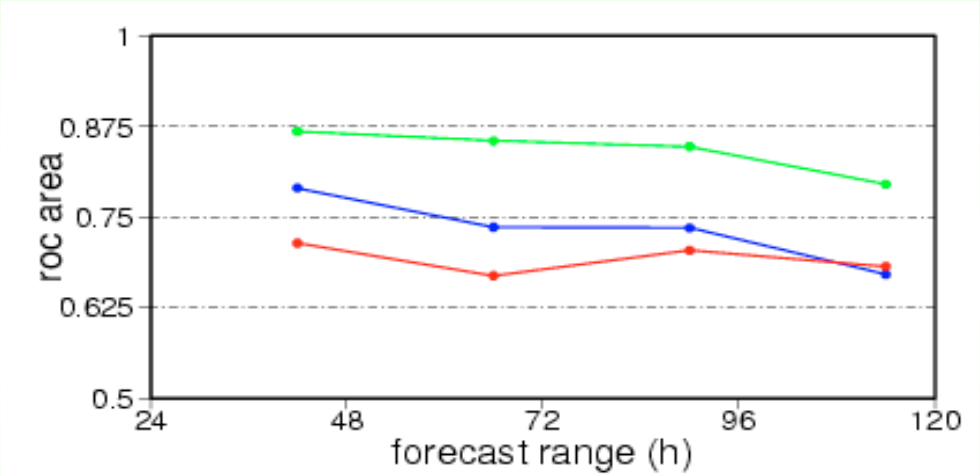


ROC area

tp > 10mm/24h



tp > 20mm/24h



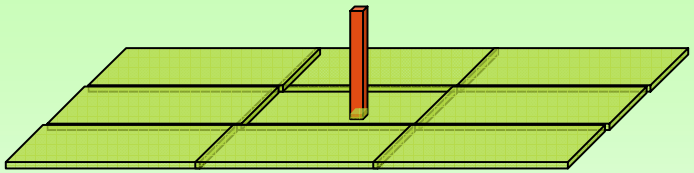
COSMO-LEPS

5-MEMBER EPS

51-MEMBER EPS

- As regards AVERAGE precipitation above these two threshols, EPS wins.
- Worsening due to the ensemble-size reduction.
- Positive impact of LM downscaling.

maxima over 1.5 x 1.5 boxes



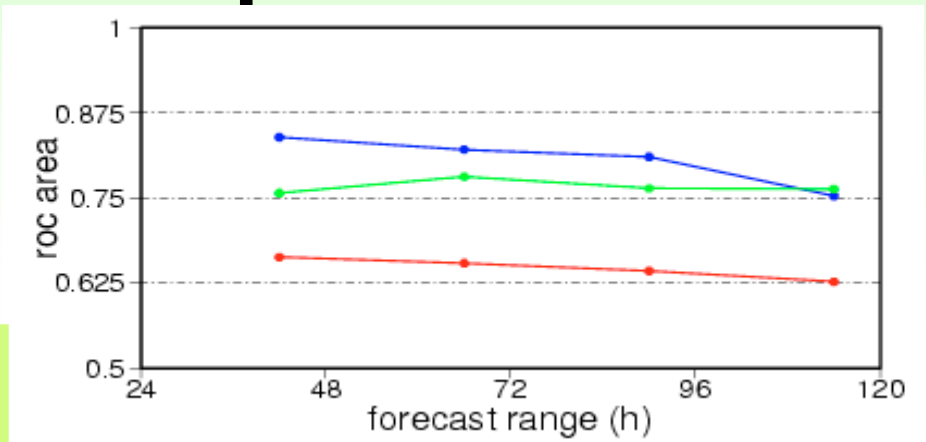
ROC area

COSMO-LEPS

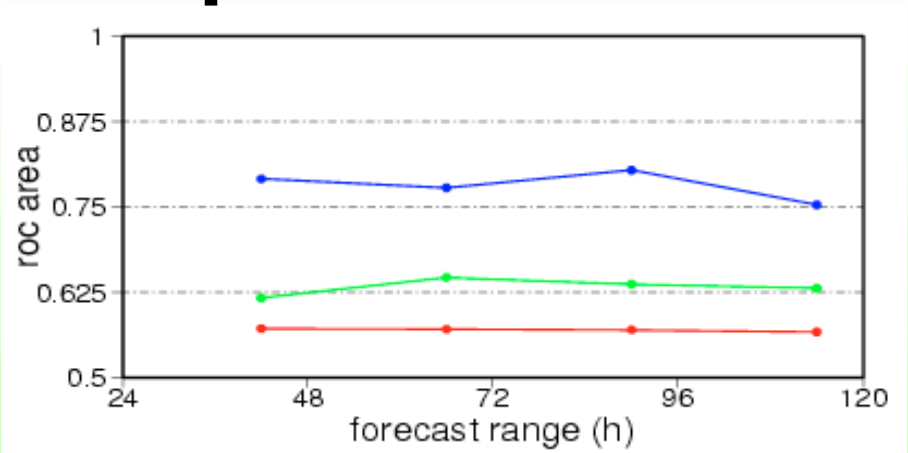
5-MEMBER EPS

51-MEMBER EPS

tp > 20mm/24h



tp > 50mm/24h



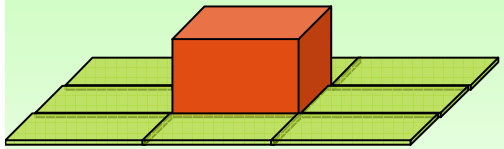
- **COSMO-LEPS is more skilful than EPS in forecasting correctly high precipitation values over a rather large area.**

Number of occurrences: 600 (20 mm threshold) and 150 (50 mm).

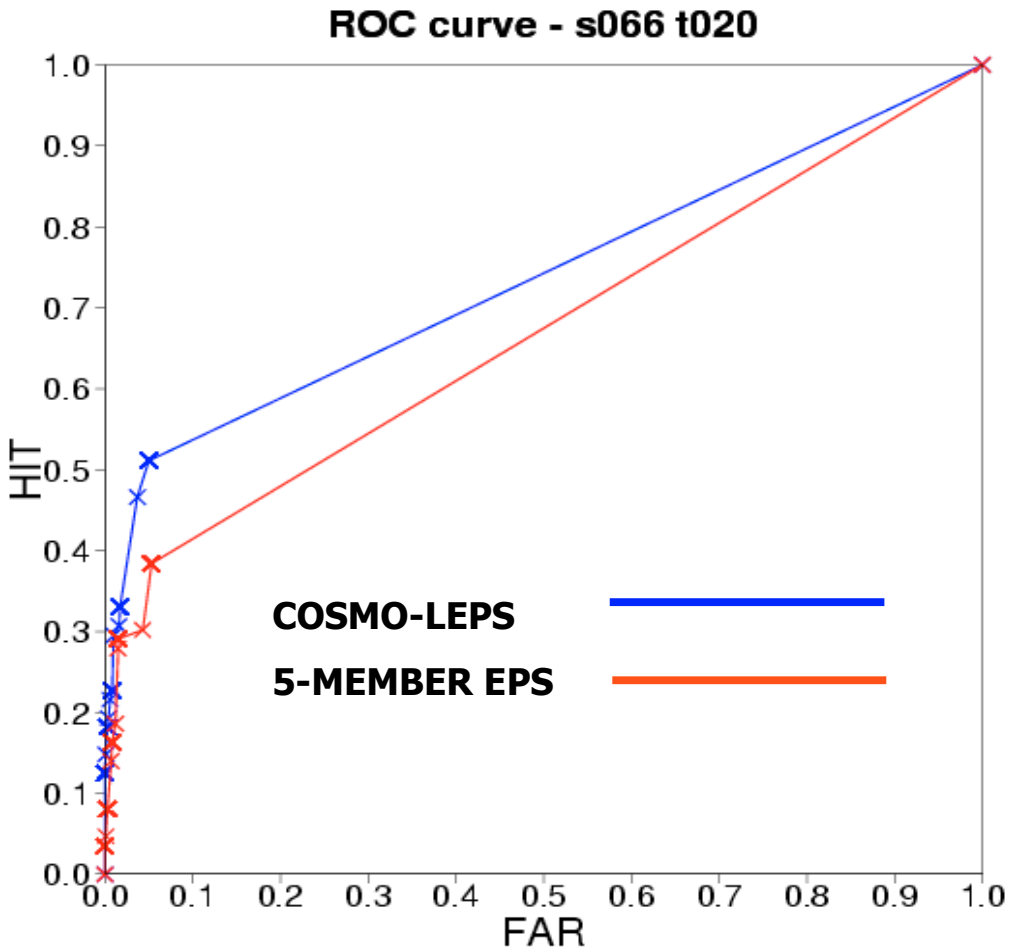
SON 2003

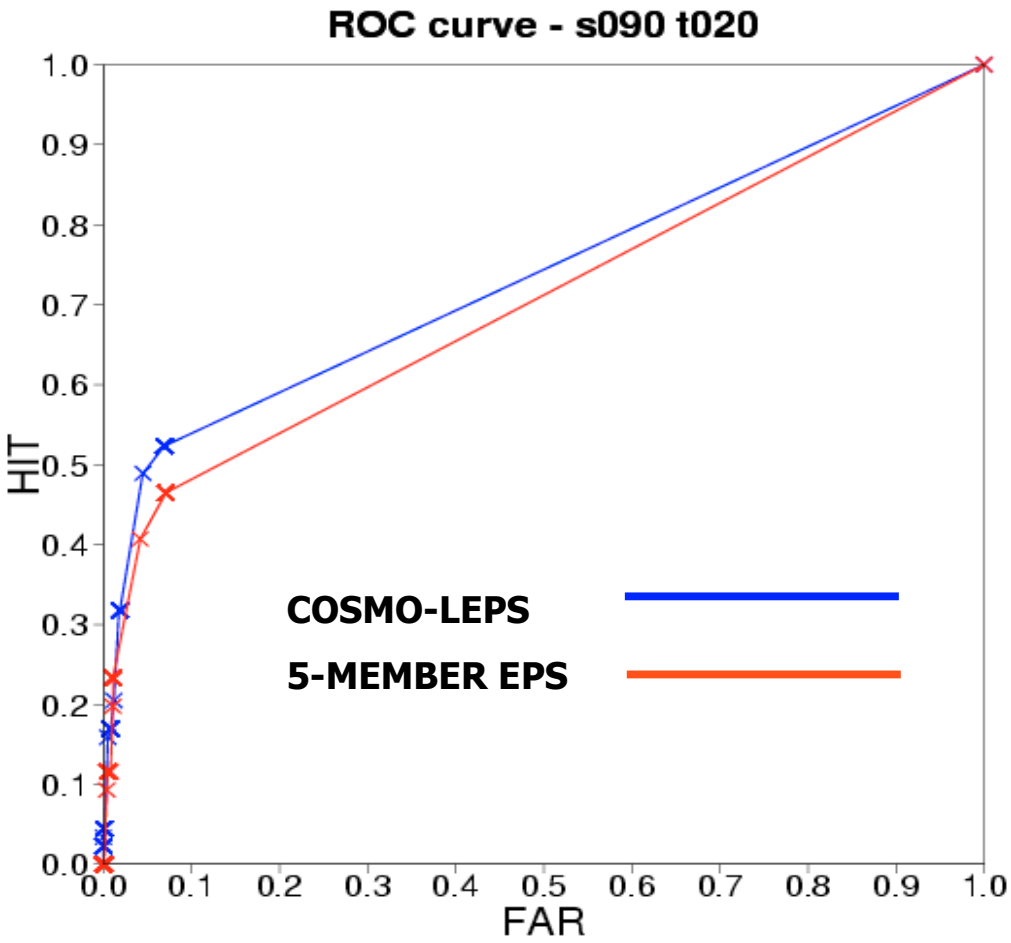
COSMO-LEPS vs ECMWF 5 RM

ROC average on 1.5 x 1.5 boxes



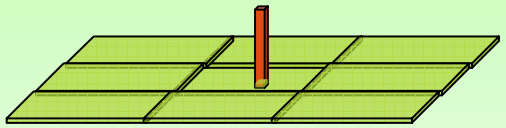
fc. range +66 **tp > 20mm/24h** **fc. range +90**



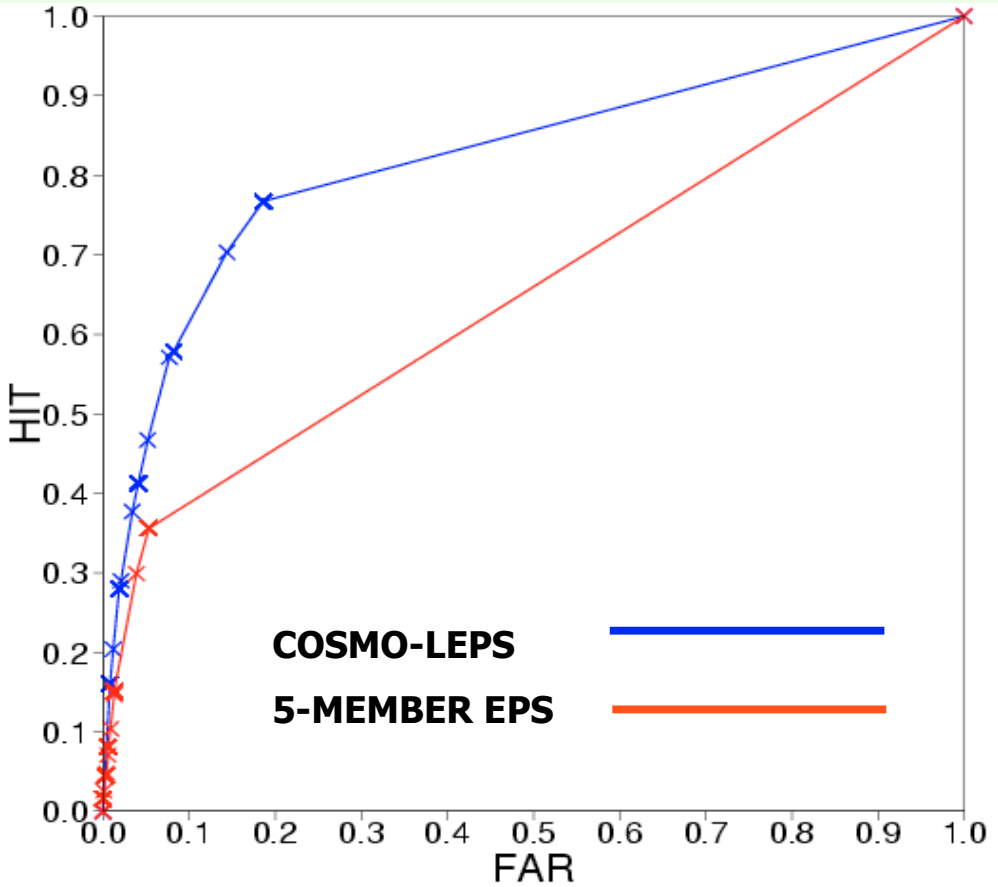


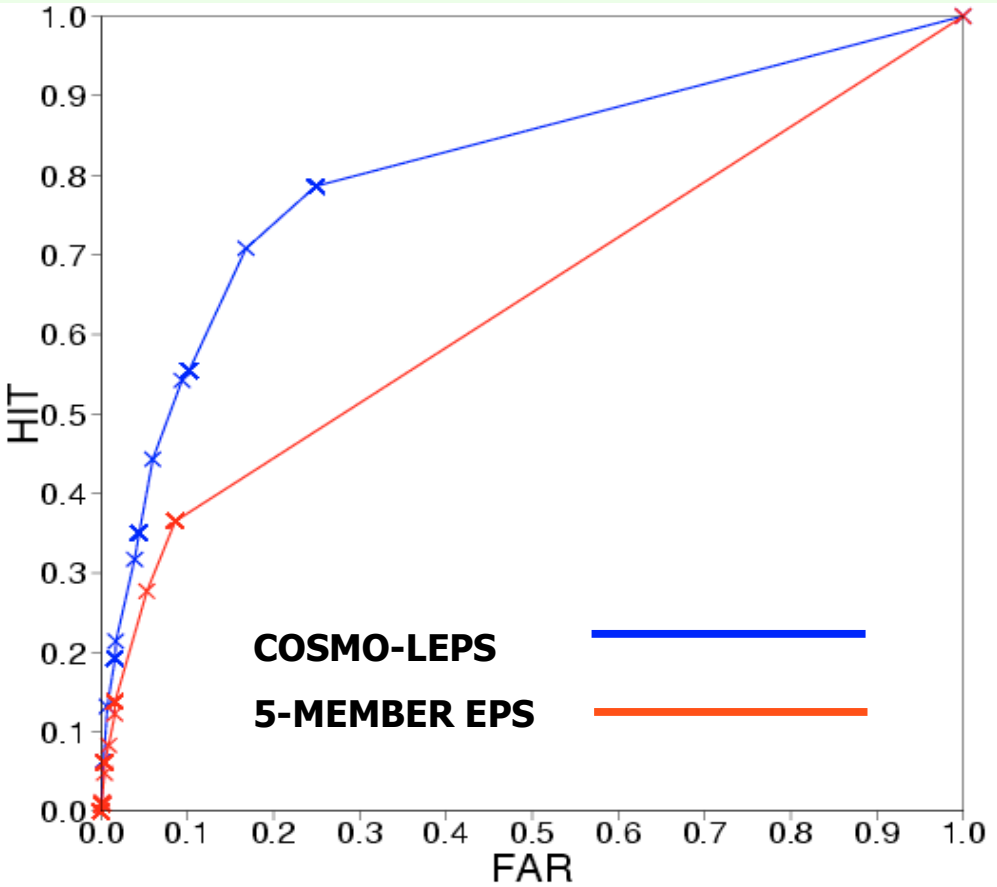
COSMO-LEPS vs ECMWF 5 RM

ROC maxima on 1.5 x 1.5 boxes



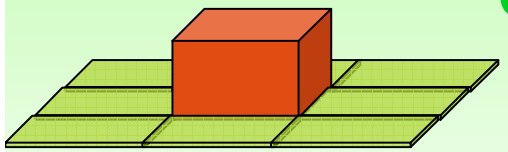
fc. range +66 **tp > 20mm/24h** **fc. range +90**



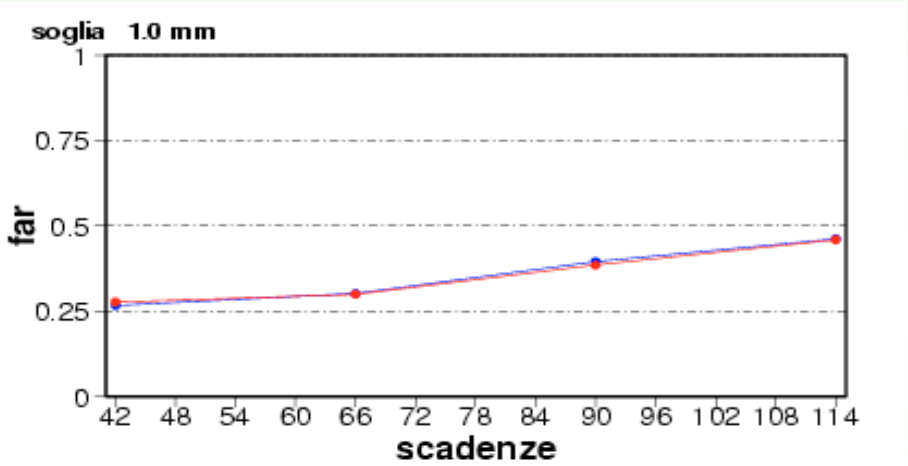


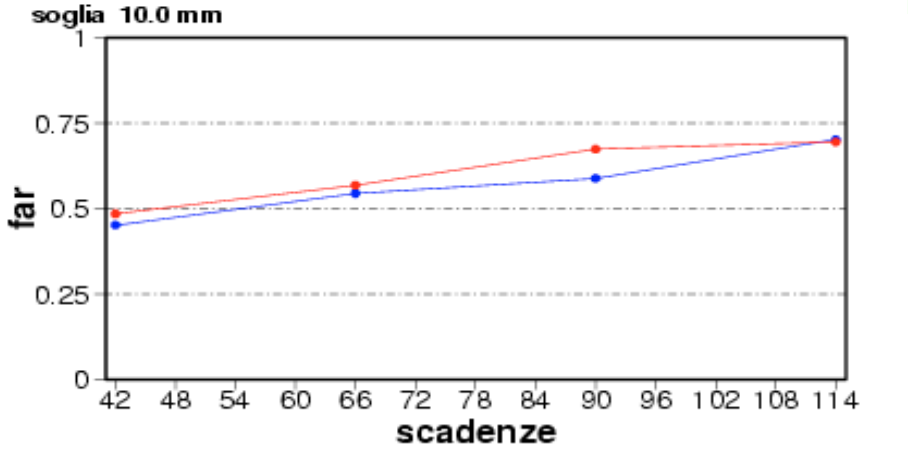
COSMO-LEPS vs ECMWF 5 RM

detscores average on 1.5 x 1.5 boxes



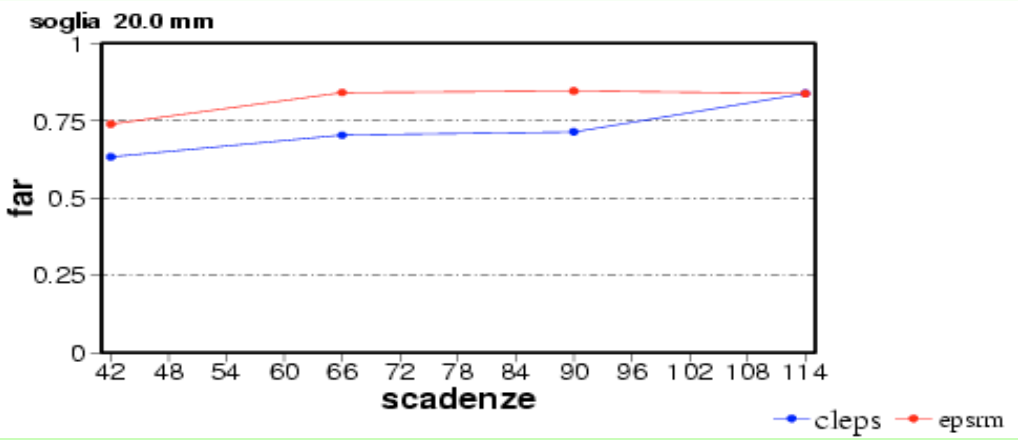
false alarm rate





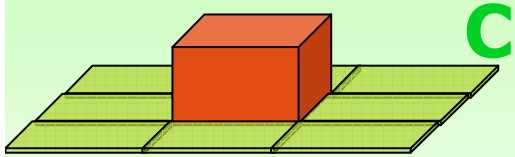
forecast \ observed	yes	no
yes	a	b
no	c	d

$$FAR = \frac{b}{a+b}$$

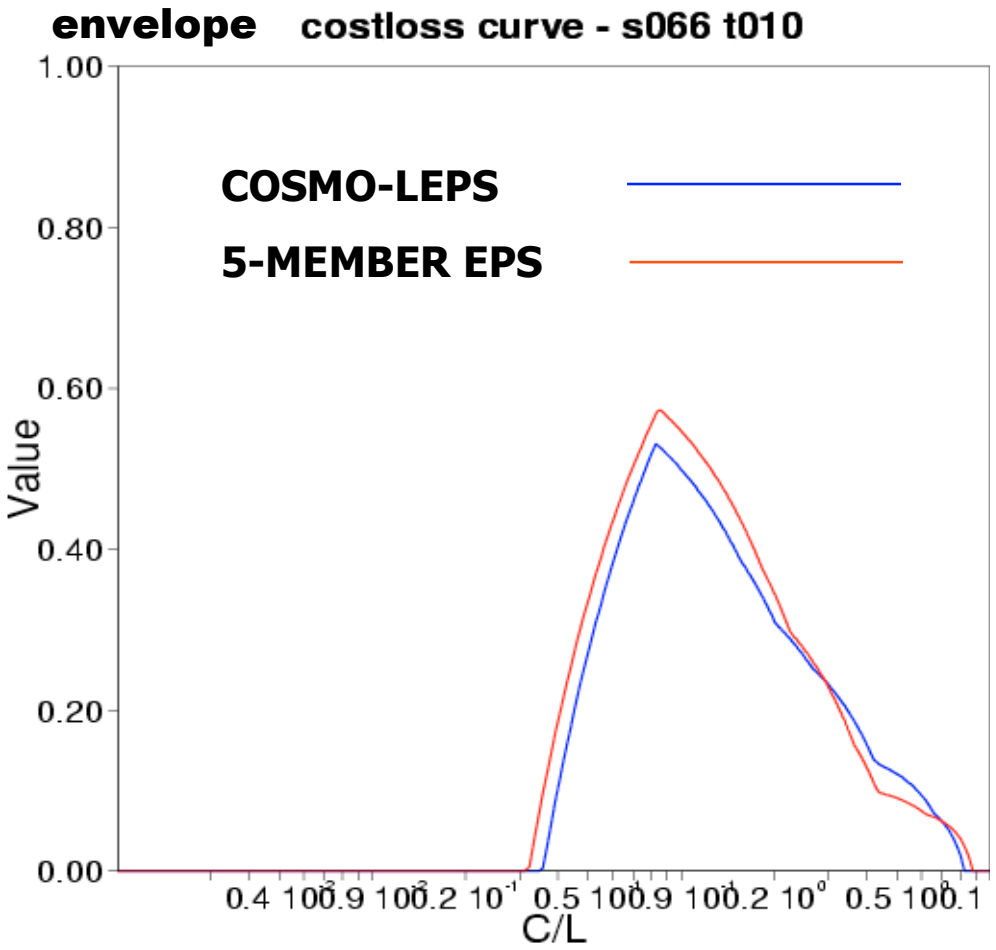


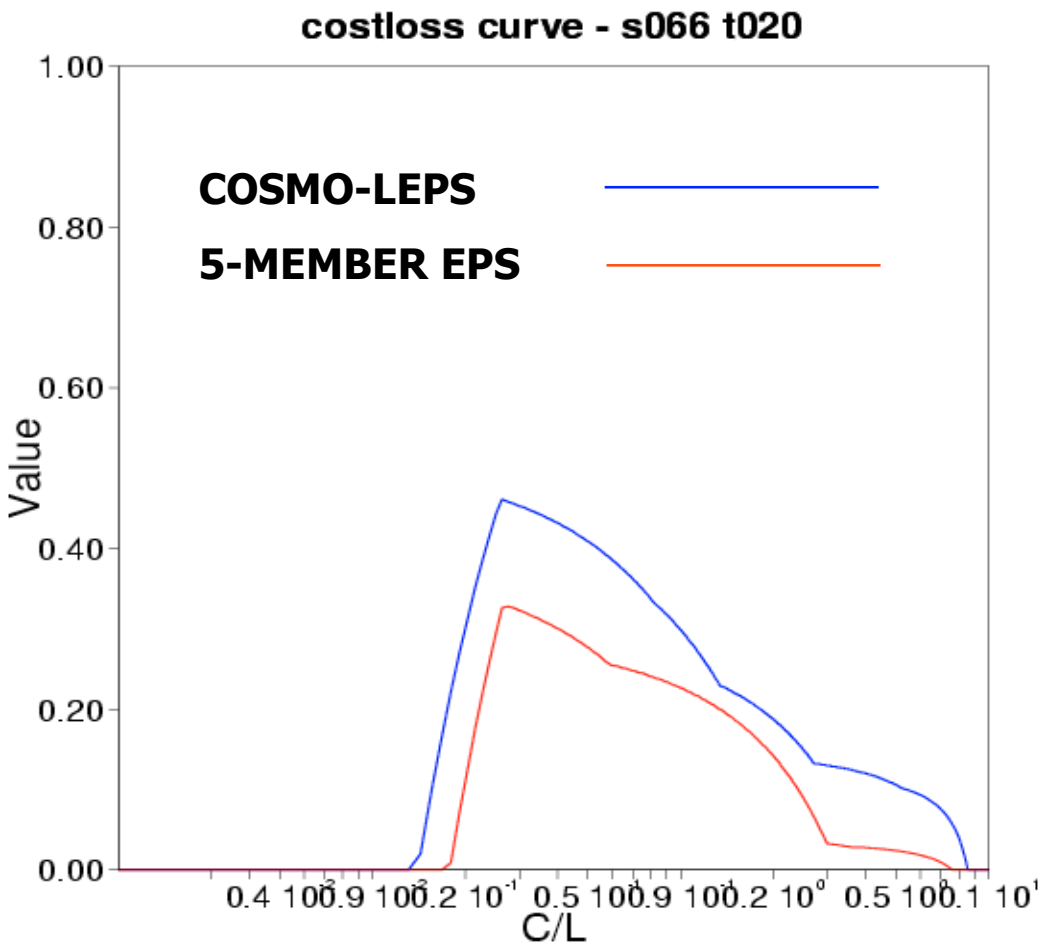
COSMO-LEPS vs ECMWF 5 RM

COST-LOSS average on 1.5 x 1.5 boxes



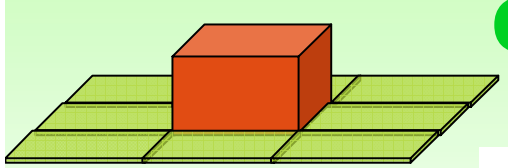
tp > 10mm/24h **fc. range +66** **tp > 20mm/24h**

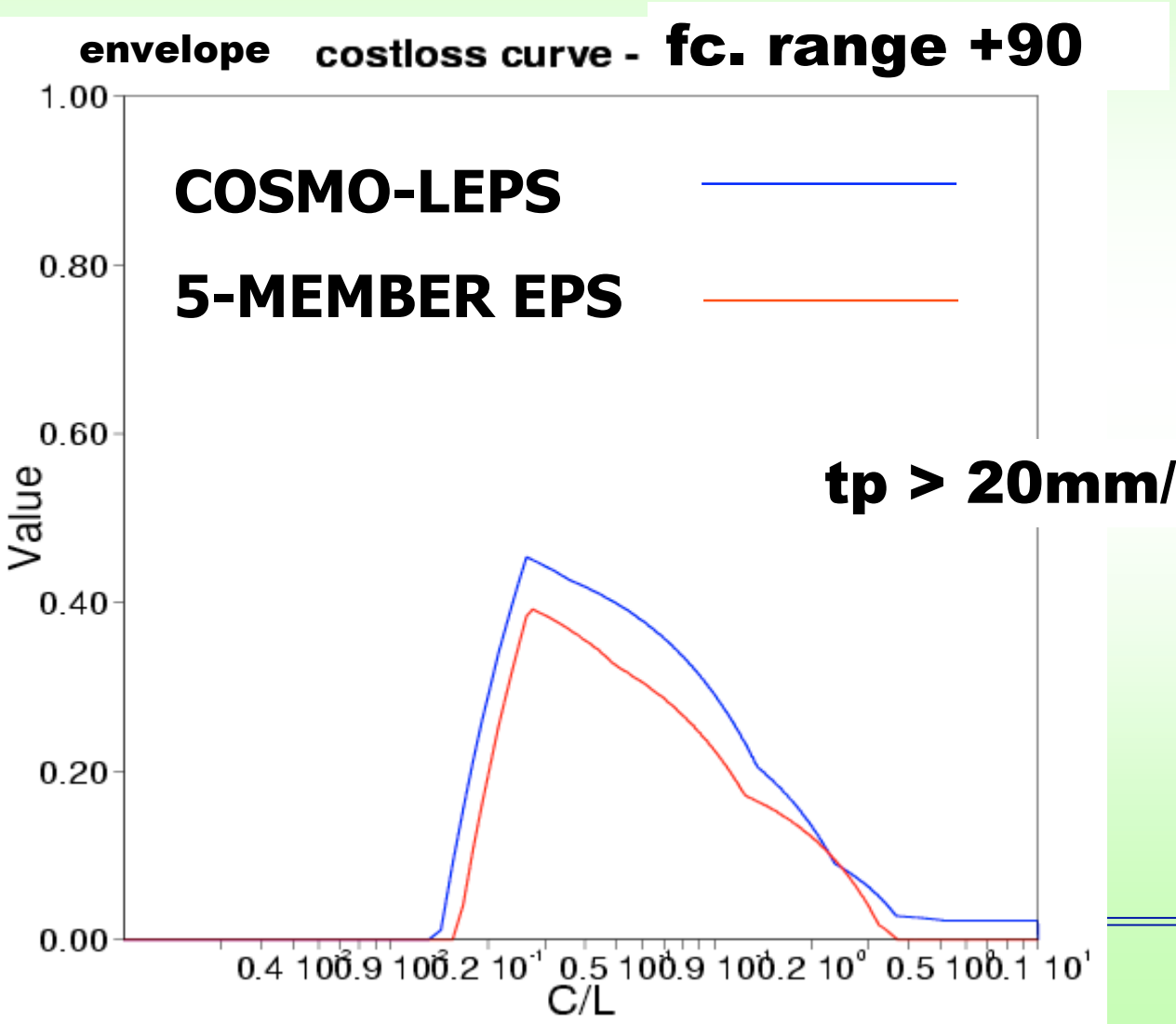




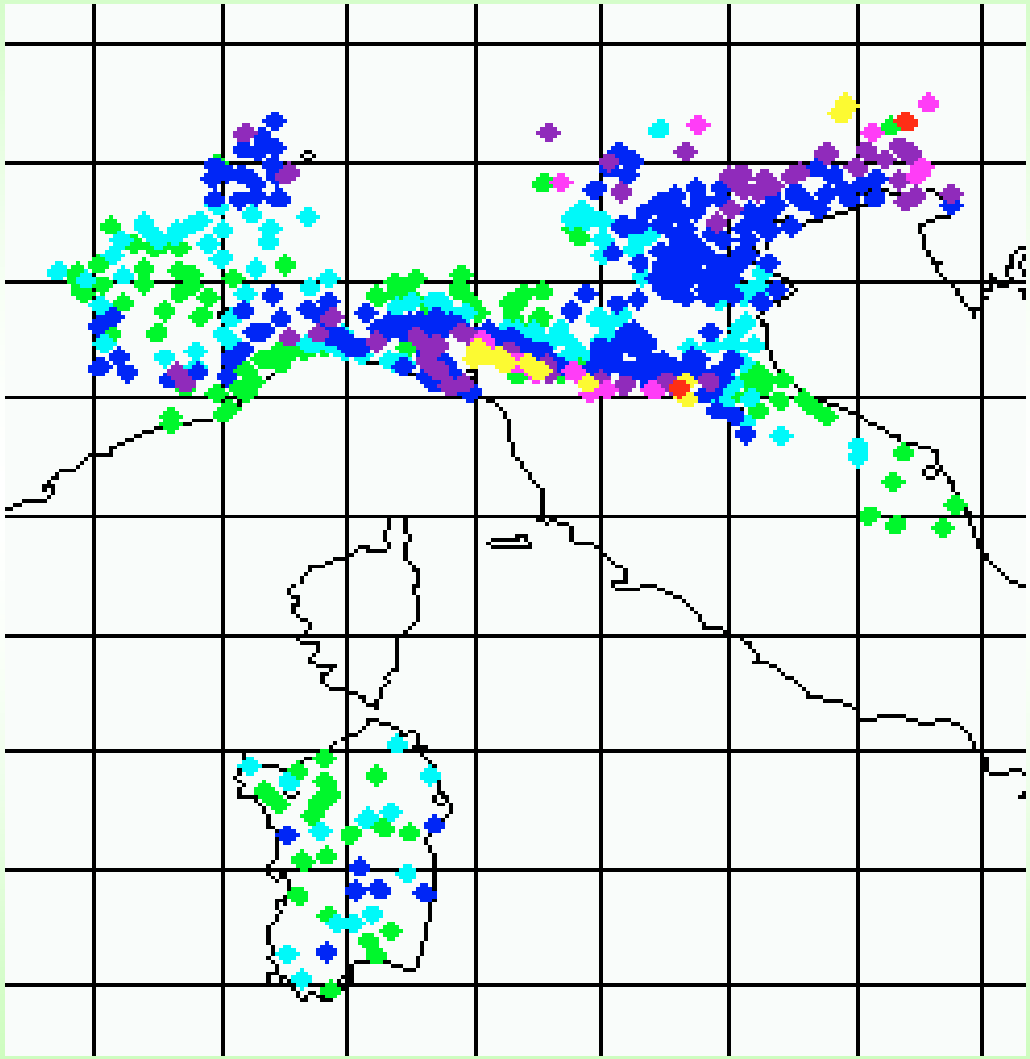
COSMO-LEPS vs ECMWF 5 RM

COST-LOSS average on 1.5 x 1.5 boxes



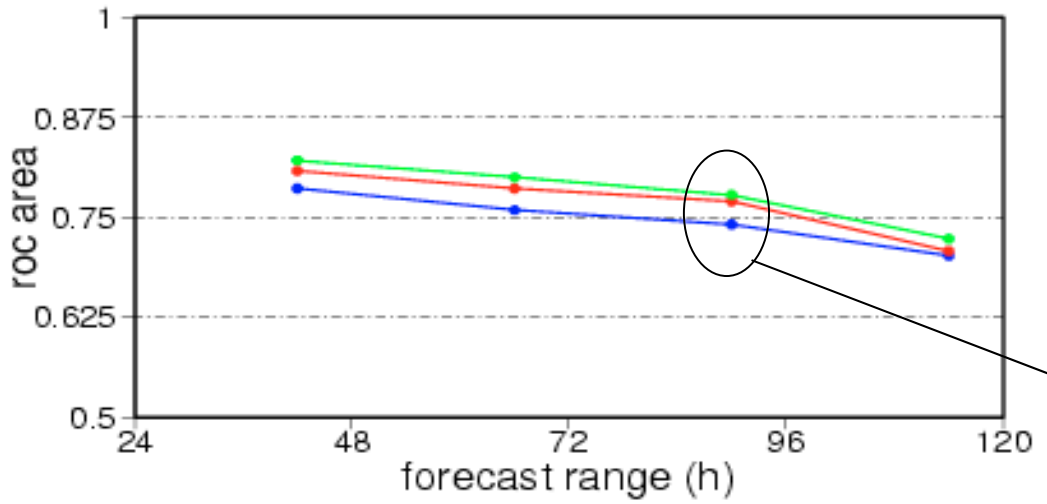


italian observations



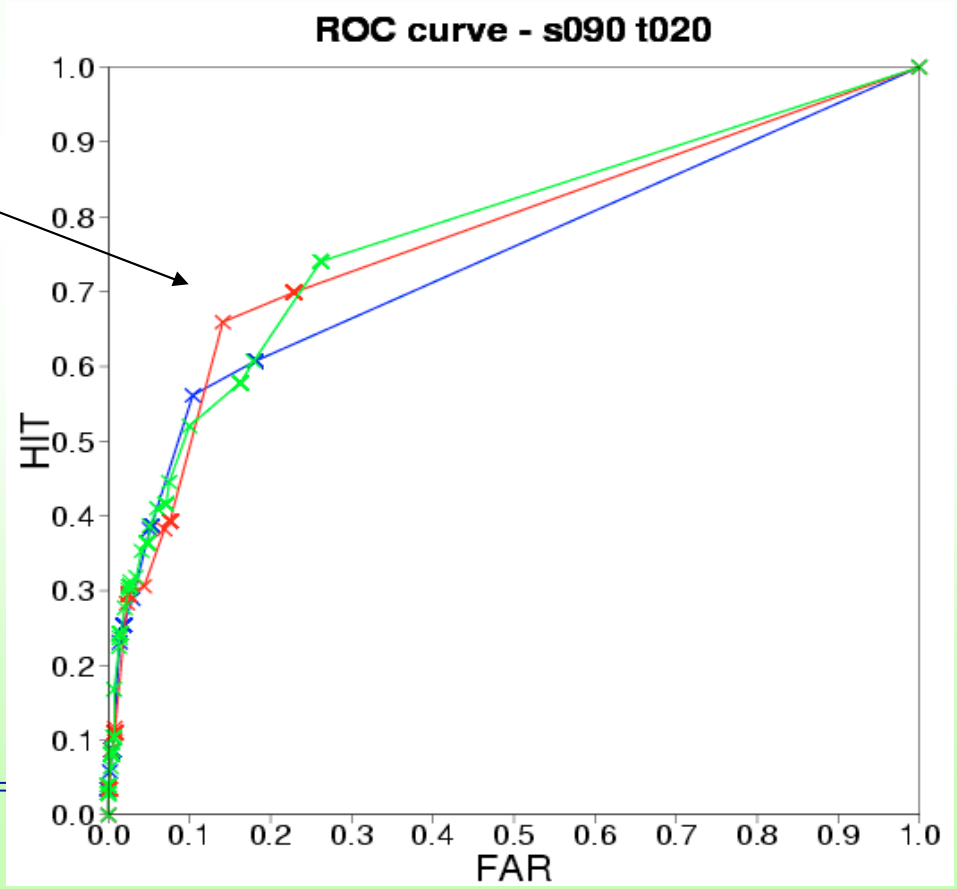
COSMO-LEPS - parallel suite

ROC average on 0.5 x 0.5 boxes



tp > 20mm/24h

fc. range +90



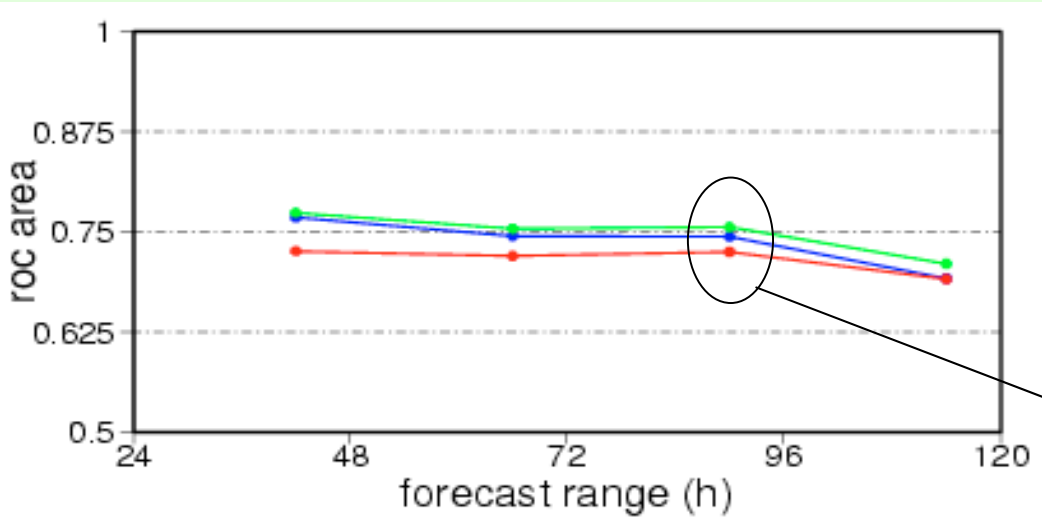
Tiedtke

Kain-Fritsch

both (10-m)

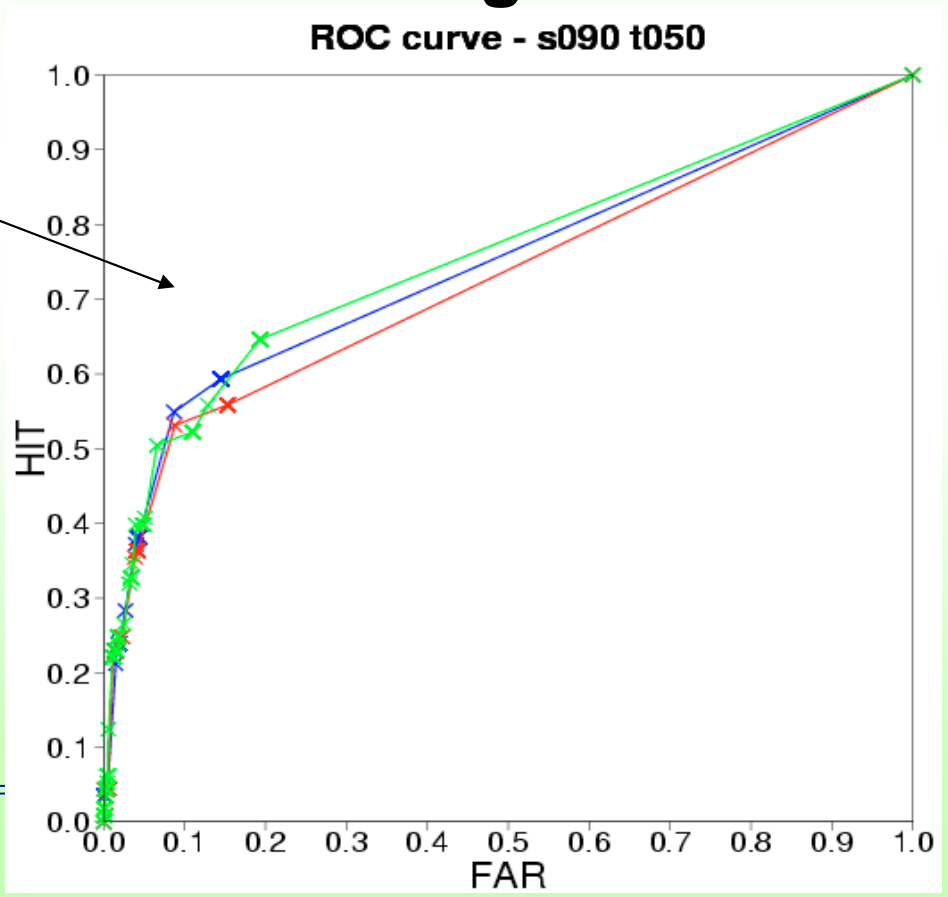
COSMO-LEPS - parallel suite

ROC maxima on 0.5 x 0.5 boxes



tp > 50mm/24h

fc. range +90



Tiedtke

Kain-Fritsch

both (10-m)

Main results

- Positive impact of COSMO-LEPS with respect to EPS in forecasting precipitation maxima;
- good performance of the ensemble size reduction technique (on case studies);
- the use of 2 EPSs and 10 RMs seems to be the "best" configuration;

and (not shown):

- no positive impact of the weighting procedure as regards high resolution precipitation;
- no relevant impact on using either Tiedke or KF convection scheme;
- differences in the scores computed in different areas (results still too preliminary but supporting the idea that Limited Area Ensemble System set-up should be designed taking into account the features of each area).

Future plans

- **COSMO-LEPS suite as "time-critical" application at ECMWF:** stronger involvement of ECMWF in the operational management of the system (+ MARS archiving of COSMO-LEPS products).
- **Participation to EURORISK-PREVIEW project:**
  - integration domain will be enlarged to include Northern Europe;
  - clustering on different areas will be tested to focus better on different scenarios (Central-North & Central-Mediterranean).
- **Partecipation to MAP D-PHASE project:**
  - further downscaling (around 2 km hor.res.) on specific areas where severe events are likely to occur (→ methodology to be evaluated also for TIGGE);
  - introduction of model perturbations to reveal uncertainty on smaller scales.
- **Carry on tests on clustering** (impact of different time ranges and different variables).
- **Verification will be further developed → new variables verified.**

Thank you for your attention

Verification of ensemble systems

Chiara Marsigli

ARPA-SIM

Deterministic forecasts

Event **E** (dichotomous event)

e.g.: the precipitation cumulated over 24 hours at a given location (raingauge, radar pixel, hydrological basin, area) exceeds 20 mm

no **o(E) = 0**	the event is observed with frequency **o(E)**	**yes** **o(E) = 1**
no **p(E) = 0**	the event is forecast with probability **p(E)**	**yes** **p(E) = 1**

Probabilistic forecasts

Event **E** (dichotomous event)

e.g.: the precipitation cumulated over 24 hours at a given location (raingauge, radar pixel, hydrological basin, area) exceeds 20 mm

no
o(E) = 0

the event is observed with frequency
o(E)

yes
o(E) = 1

the event is forecast with probability
p(E)

$$p(E) \in [0,1]$$

Ensemble forecasts

Event **E** (dichotomous event)

e.g.: the precipitation cumulated over 24 hours at a given location (raingauge, radar pixel, hydrological basin, area) exceeds 20 mm

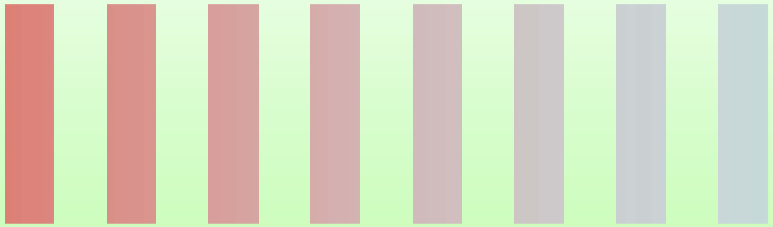
no
o(E) = 0

the event is observed with frequency
o(E)

yes
o(E) = 1

M member ensemble
the event is forecast with probability **p(E) = k/M**

no member
p(E) = 0



all members
p(E) = 1

Quality of the forecast

Murphy (1993)

Degree of correspondence between forecasts and observations

Distribution-oriented approach: the joint distribution of forecasts and observations p(f,x) contains all the non-time-dependent information relevant to evaluating forecast quality (Murphy and Winkler, 1987).
This information becomes more accessible when p(f,x) is factored into conditional and marginal distributions:

- conditional distribution of the observations given the forecasts **p(x|f)**
- conditional distribution of the forecasts given the observations **p(f|x)**
- marginal distribution of the forecasts **p(f)**
- marginal distribution of the observations **p(x)**

Quality of probabilistic forecasts

The accuracy of a probability forecast system is determined by:

- **reliability**
- **resolution**

which can be assessed by examining the conditional distribution p(x|f) and the marginal distribution p(f)

Reliability

- capability to provide unbiased estimates of the observed frequencies associated with different forecast probability values
- p(x), compiled over the cases when the forecast probability density is p(f), equals p(f)
- answers: is the relative frequency of precipitation on those occasions on which the precipitation probability forecast is 0.3 equal to this probability?

Not sufficient: a system always forecasting the climatological probability of the event is reliable but not useful

And: it can always be improved by **calibration**, re-labeling the forecast probability values

Resolution

- ability of a forecast system to *a priori* separate cases when the the event under consideration occurs more or less frequently than the climatological frequency
- measures the difference between the conditional distribution of the observations and the unconditional distribution of the observations (climatology)

Resolution cannot be improved by simply post-processing forecast probability values

Reliability and Resolution

Toth et al. (2003)

- a useful forecast system must be able to *a priori* separate cases into groups with as different possible outcome as possible, so each forecast group is associated with a distinct distribution of verifying observations (res)
- then it is necessary to label properly the different groups of cases identified by the forecast system (rel). This can be done by "renaming" the groups according to the frequency distributions associated with each forecast group, based on a long series of past forecasts (calibration)
- is the series sufficient?

Sharpness and Uncertainty

Sharpness

- expressed by the marginal distribution of the forecasts p(f)
- capability of the system to forecast extreme values (near 0 or 1); variability of the forecast probability distribution around the climatological pdf

Uncertainty

- expressed by the marginal distribution of the observations p(x)
- a situation in which the events are apporximately equally likely is indicative of high uncertainy

Brier Score

Brier (1950)

$$BS = \frac{1}{N}\sum_{i=1}^{N}\left(f_i - o_i\right)^2$$

Scalar summary measure for the assessment of the probabilistic forecast performance, mean-squared error of the probability forecast

- N = number of points in the "domain" (spatio-temporal)
- o_i = 1 if the event occurs
  = 0 if the event does not occur
- f_i is the probability of occurrence according to the forecast system (e.g. the fraction of ensemble members forecasting the event)

❖ **BS** takes on values in the range [0,1], a perfect (deterministic) forecast having BS = 0

❖ Sensitive to climatological frequency of the event: the more rare an event, the easier it is to get a good **BS** without having any real skill

Brier Score decomposition

$$BS = \frac{1}{N}\sum_{k=0}^{M} N_k (f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=0}^{M} N_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$

reliability resolution uncertainty

The first term is a **reliability** measure: for forecasts that are perfectly reliable, the sub-sample relative frequency is exactly equal to the forecast probability in each sub-sample.

The second term is a **resolution** measure: if the forecasts sort the observations into sub-samples having substantially different relative frequencies than the overall sample climatology, the resolution term will be large. This is a desirable situation, since the resolution term is subtracted. It is large if there is resolution enough to produce very high and very low probability forecasts.

The **uncertainty term** ranges from 0 to 0.25. If E was either so common, or so rare, that it either always occurred or never occurred, then b_{unc}=0. When the climatological probability is near 0.5, there is more uncertainty inherent in the forecasting situation (b_{unc}=0.25).

on the Brier Score

Talagrand et al. (1999)

Sources of uncertainty in the evaluation of the accuracy of a probabilistic prediction system:

- errors in the verifying observations
- finitness of the sample
- **finitness of the ensembles** from which predicted probabilities are estimated (N members)

$$B_N = B + \frac{1}{N}\int_0^1 p(1-p)g(p)dp$$

- increasing N will result in a decrease of the Brier Score, i.e. in a increase of the quality of the system, which results from a smoothing of the noise due to the finiteness of the ensembles
- the numerical impact of increasing N will be larger if the predicted probabilities have small dispersion (small sharpness)

on the Brier Score

Candille and Talagrand (2004)

- the system of **N members** produces probabilities p, p'(p) is the frequency of occurrence of E when p is predicted.
- **M** (**realisations** on which the statistics is computed) must be large enough so that a significant estimate of p'(p) is obtained for each p; if ε is the precision of the reliability diagnosis the condition is:

$$M \geq \frac{2}{\varepsilon^2} N \ln(N)$$

- increasing **N** without increasing M improves the resolution but degrades the reliability.

e.g. ε=10%

N =	5	10	20	50	100	1000
M >=	1963	5549	14087	44690	103447	$1.5\ 10^6$

Brier Skill Score

Measures the improvement of the probabilistic forecast relative to a reference forecast (e. g. sample climatology)

$$BSS = 1 - \frac{BS}{BS_{cli}}$$

$$BS_{cli} = \bar{o}\left(1 - \bar{o}\right)$$

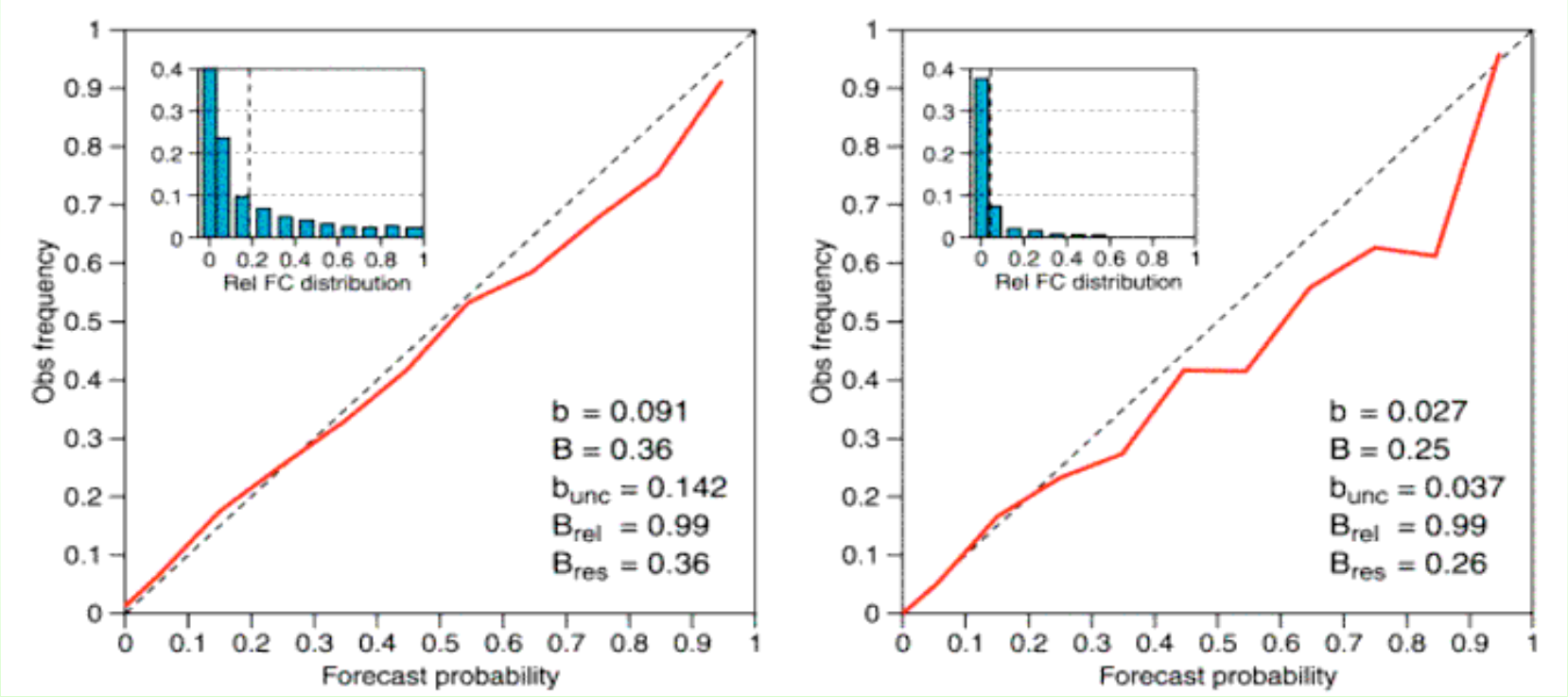
$\bar{o}$ = total frequency of the event (sample climatology)

The forecast system has predictive skill if BSS is positive (better than climatology), a perfect system having BSS = 1.

Reliability Diagram

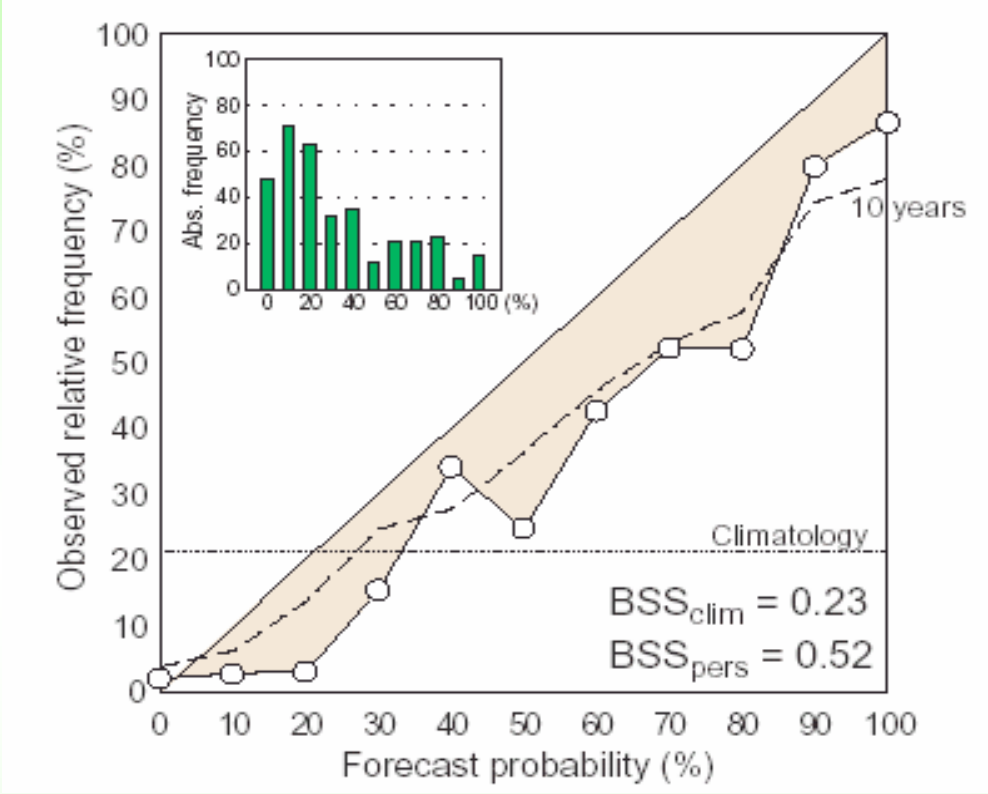
o(p) is plotted against p for some finite binning of width dp

In a perfectly reliable system o(p)=p and the graph is a straight line oriented at 45° to the axes



If the curve lies below the 45° line, the probabilities are overestimated
If the curve lies above the 45° line, the probabilities are underestimated

Reliability Diagram



Sharpness histogram:

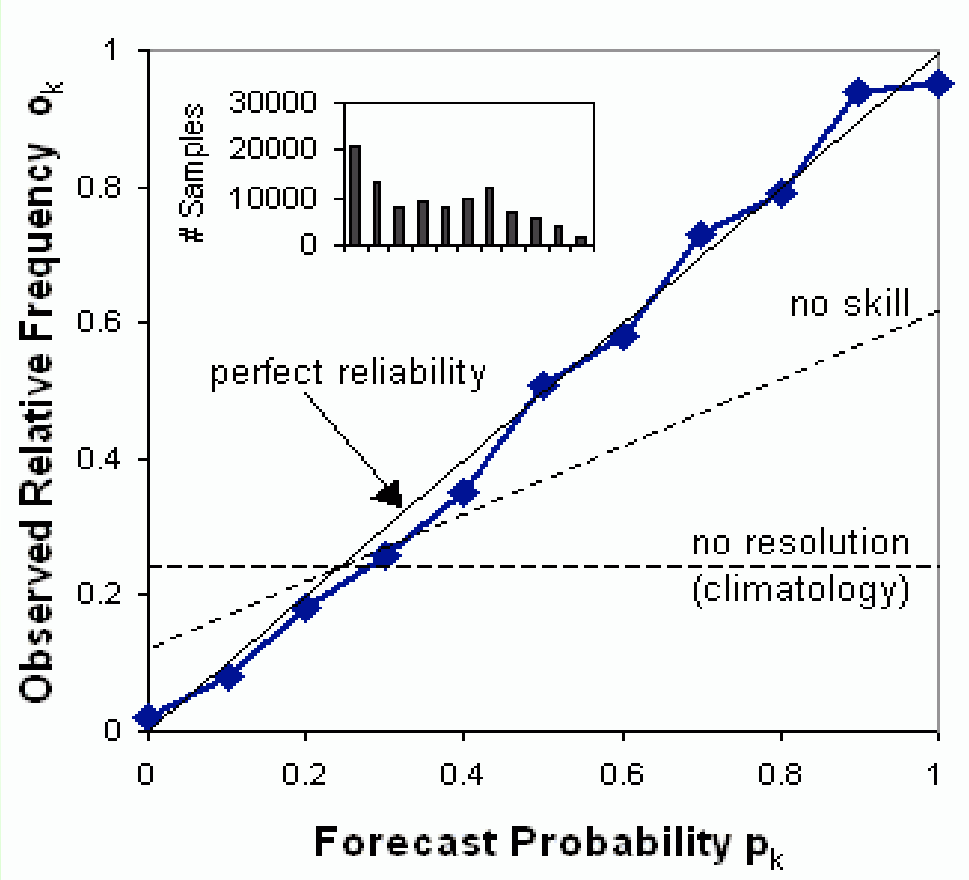
the frequency of forecasts in each probability bin (histogram) shows the sharpness of the forecast.

- the reliability diagram is conditioned on the forecasts, p(x|f), then it is a good partner to the ROC, which is conditioned on the observations, p(f|x).

- the histogram is the unconditional distribution of the forecasts p(f) => compact display of the full distribution of forecasts and observations

Reliability (attributes) Diagram

The **reliability term** measures the mean square distance of the graph of o(p) to the diagonal line.



$$\mathrm{BSS} = \frac{\text{resolution} - \text{reliability}}{\text{uncertainty}}$$

Points between the "no skill" line and the diagonal contribute positively to the Brier skill score (resolution > reliability).

The **resolution term** measures the mean square distance of the graph of o(p) to the sample climate horizontal dotted line.

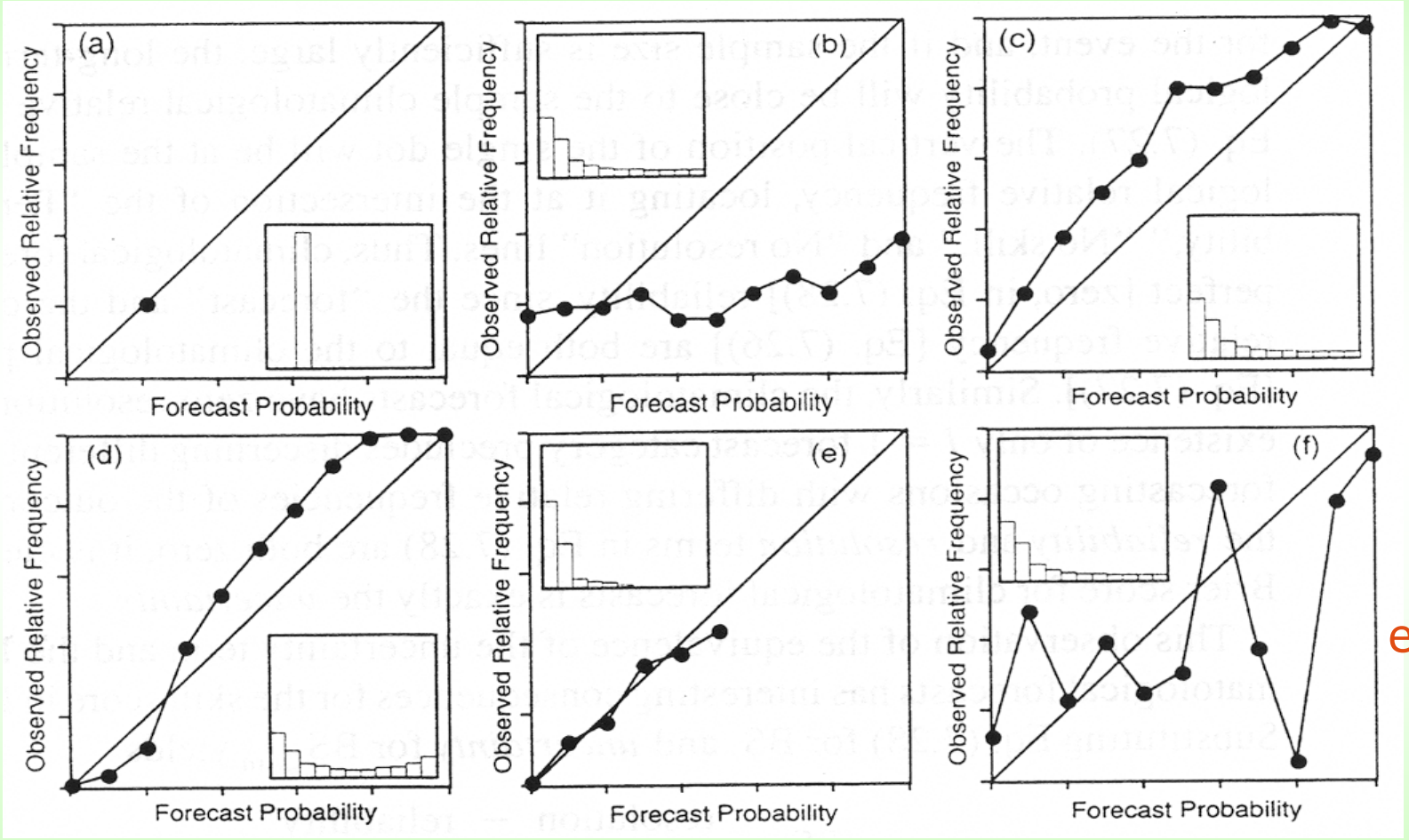
Reliability Diagram

Wilks (1995)

climatological forecast

minimal resolution

underforecasting bias



+ small ensemble

Good resolution at the expense of reliability

reliable of rare event

small sample size

Ranked Probability Score

Epstein (1969), Murphy (1971) + continuous (Hersbach, 2000)

$$RPS = \frac{1}{J-1}\sum_{m=1}^{J}\left[\left(\sum_{j=1}^{m} f_j\right)-\left(\sum_{j=1}^{m} o_j\right)\right]^2 \qquad \text{N = 1}$$

Extension of the Brier Score to the multi-event situation, taking into account the ordered nature of the variable (e.g.: TP <1mm, 1mm-20mm, >20mm)

- J = number of forecast categories
- o_j = 1 if the event occurs in category j
  = 0 if the event does not occur in category j
- f_j is the probability of occurrence in category j

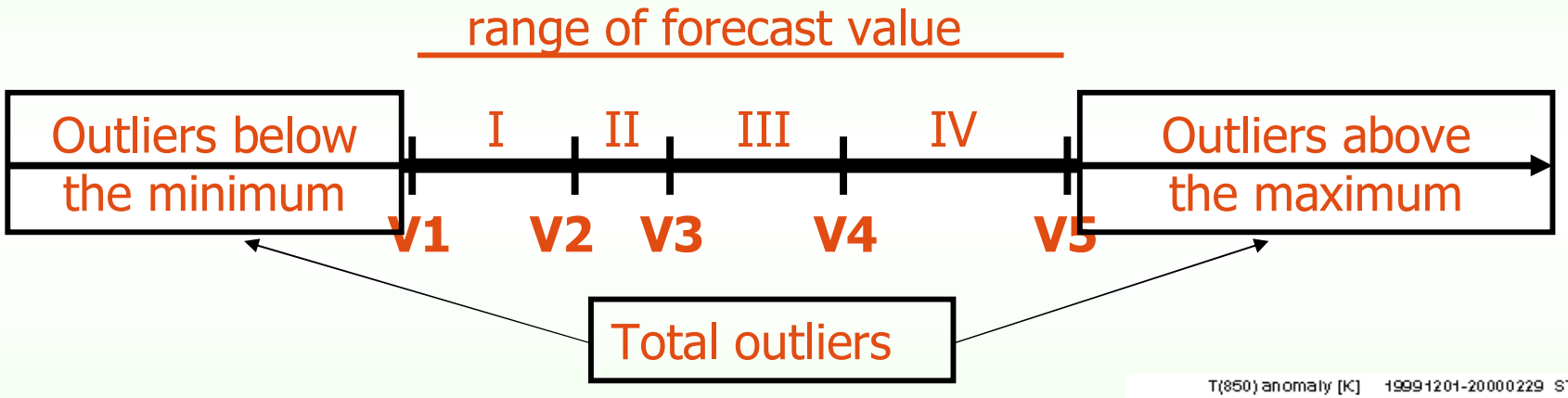
❖ sensitive to the distance: the squared errors are computed with respect to the cumulative probabilities in the forecast and observation vectors (penalise "near misses" less than larger errors, rewards small spread)

❖ **RPS** take on values in the range [0,1], a perfect forecast having RPS = 0

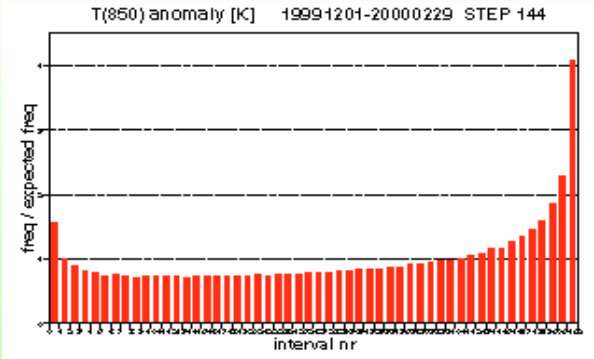
Rank histogram (Talagrand Diagram)

Talagrand et al. (1999)

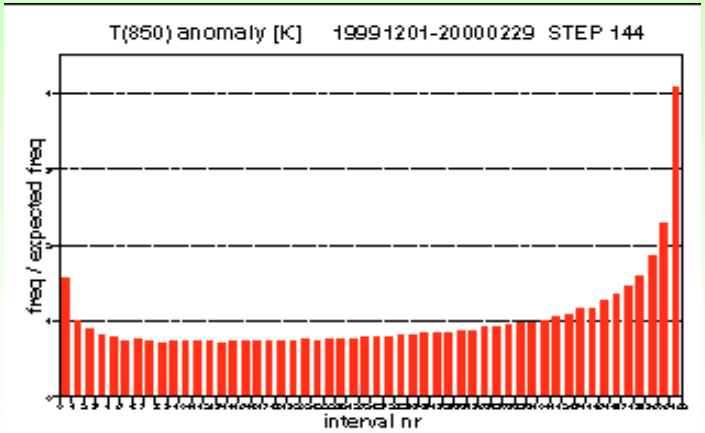
Frequency of occurrence of the observation in each bin of the rank histogram of the distribution of the values forecast by an ensemble



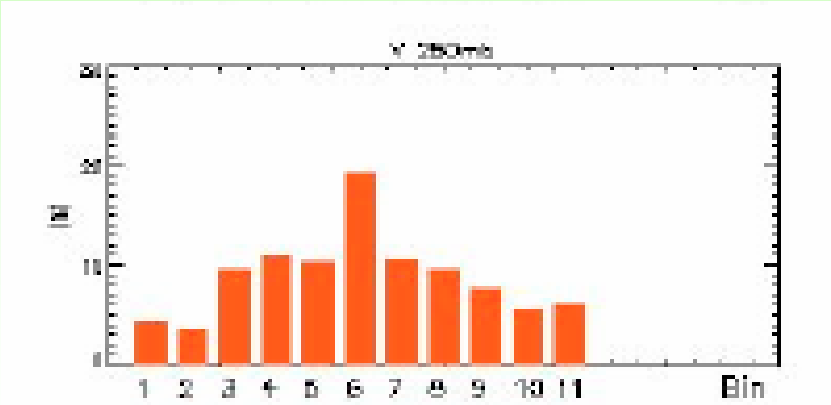
If the ensemble members and the verifying observation are independent realisations of the same probability distribution, each interval is equally likely to contain the verifying observed value (measure of reliability)



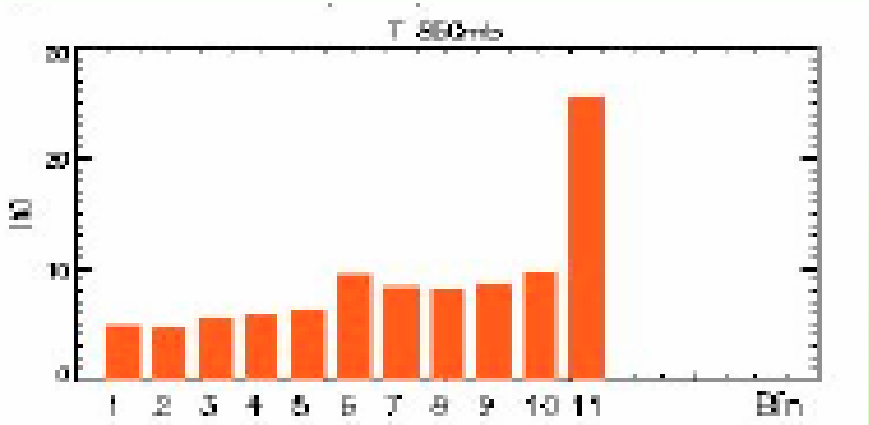
Rank histogram (Talagrand Diagram)

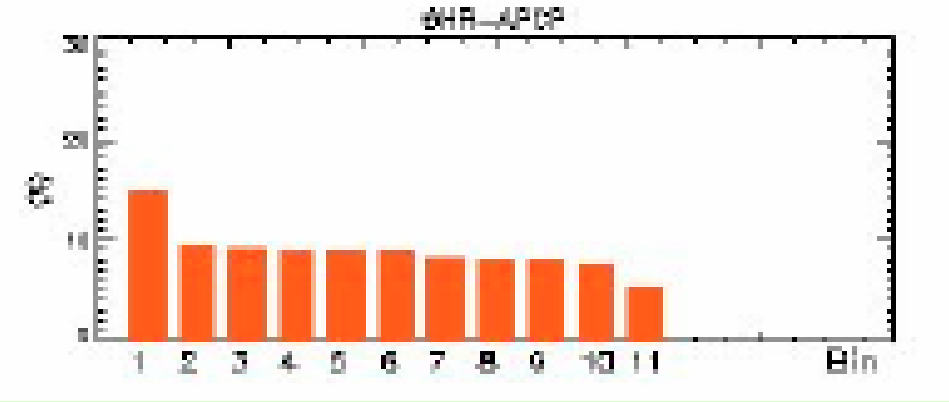


U-shape: negative bias in the variance



dome-shape: positive bias in the variance



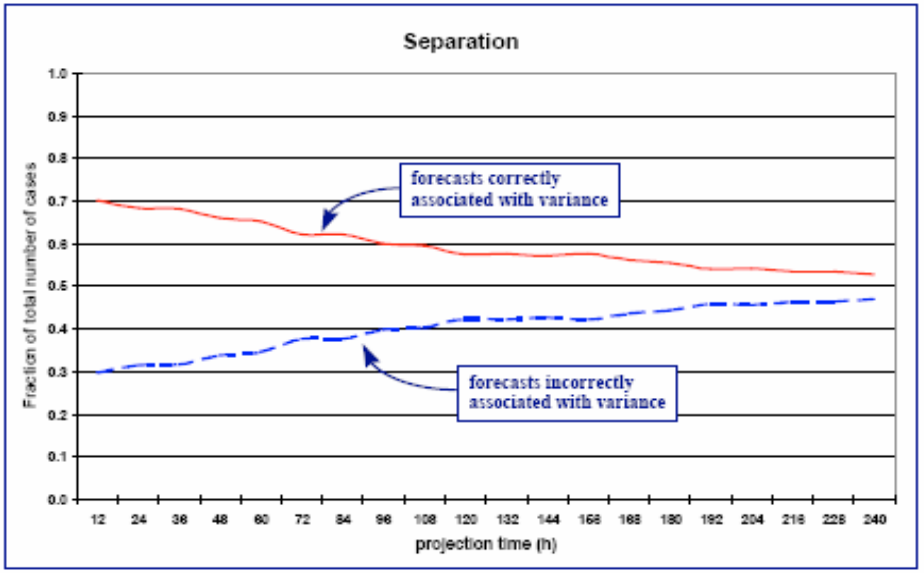


Asymmetrical: bias in the mean

Spread-skill relationship

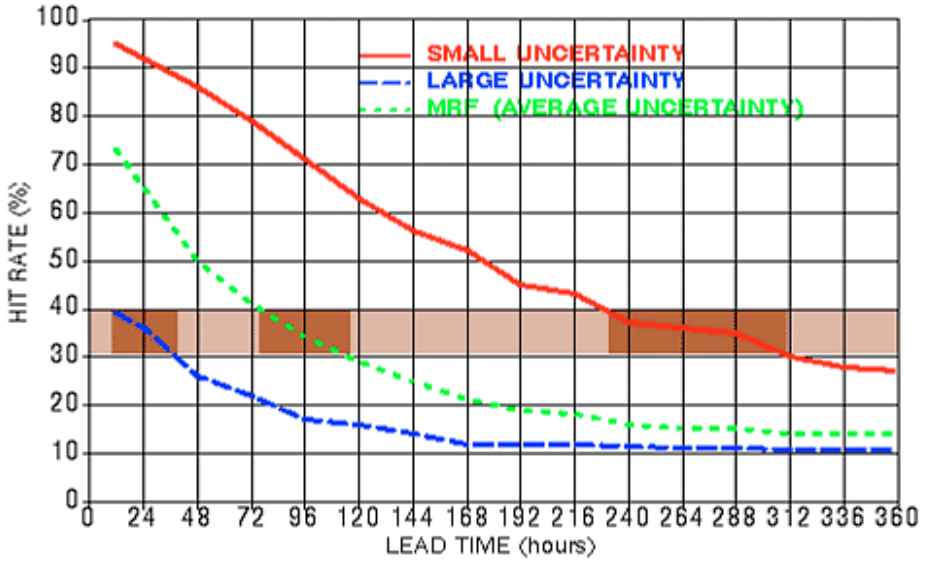
Is it possible to obtain from a probabilistic prediction system an estimate, even if qualitative, of the confidence to be given to the forecast?

If the spread of the predicted pdf is small (large), the correspondent uncertainty of the forecast si small (large)



http://ams.confex.com/ams/annual2002/techprogram/paper_26835.htm

+ Ziehmann, 2001



Toth et al., 2001

Relative Operating Characteristics (ROC)

❖ For a given probability threshold p_t, probability forecast can be converted into deterministic forecast:

if	$\hat{p} \geq p_t$	=>	$\hat{X} = 1$	the event is forecast
	otherwise		$\hat{X} = 0$	the event is not forecast

❖ It can be used the Signal Detection Theory, which permits to evaluate the ability of the forecast system to discriminate between occurrence and non-occurrence of an event (to detect the event) on the basis of information which is not enough for certainty. A powerful analysis tool is the Relative Operating Characteristic (ROC).

ROC Curves

(Mason and Graham 1999)

contingency table		Observed	
		Yes	No
Forecast	Yes	a	b
	No	c	d

Hit Rate

$$H = \frac{a}{a+c} = \frac{\text{number of correct forecasts of the event}}{\text{total number of occurrences of the event}}$$

False Alarm Rate

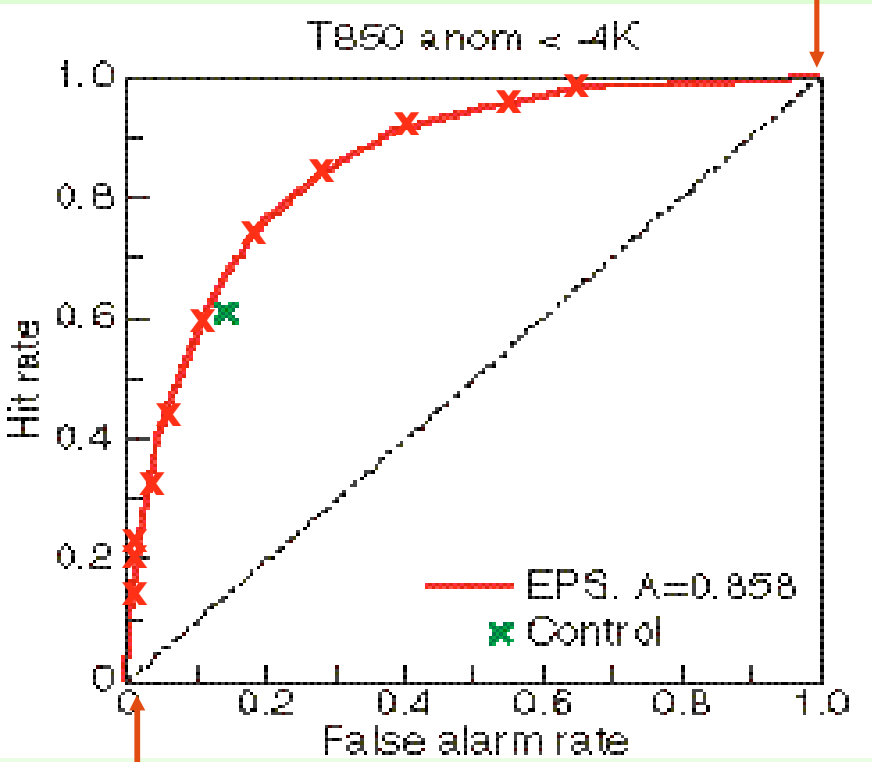
$$F = \frac{b}{b+d} = \frac{\text{number of non correct forecasts of the event}}{\text{total number of non - occurrences of the event}}$$

A contingency table can be built for each probability class (a probability class can be defined as the % of ensemble elements which actually forecast a given event)

N.B. F is defined as $F = \frac{b}{b+d}$ and not $F = \frac{b}{a+b}$

ROC Curve

"At least 0 members" (always)



"At least M+1 members" (never)

k-th probability class: E is forecast if it is forecast by at least k ensemble members

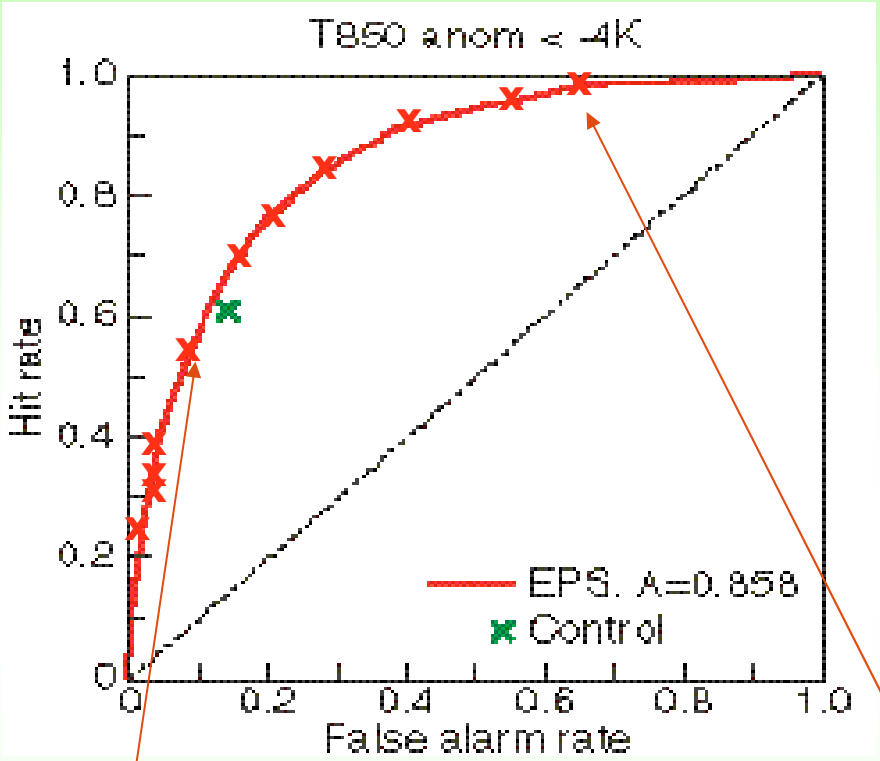
=> a warning can be issued when the forecast probability for the predefined event exceeds some threshold

For the k-th probability class:

$$H_k = \sum_{i=k}^{M} H_i \qquad F_k = \sum_{i=k}^{M} F_i$$

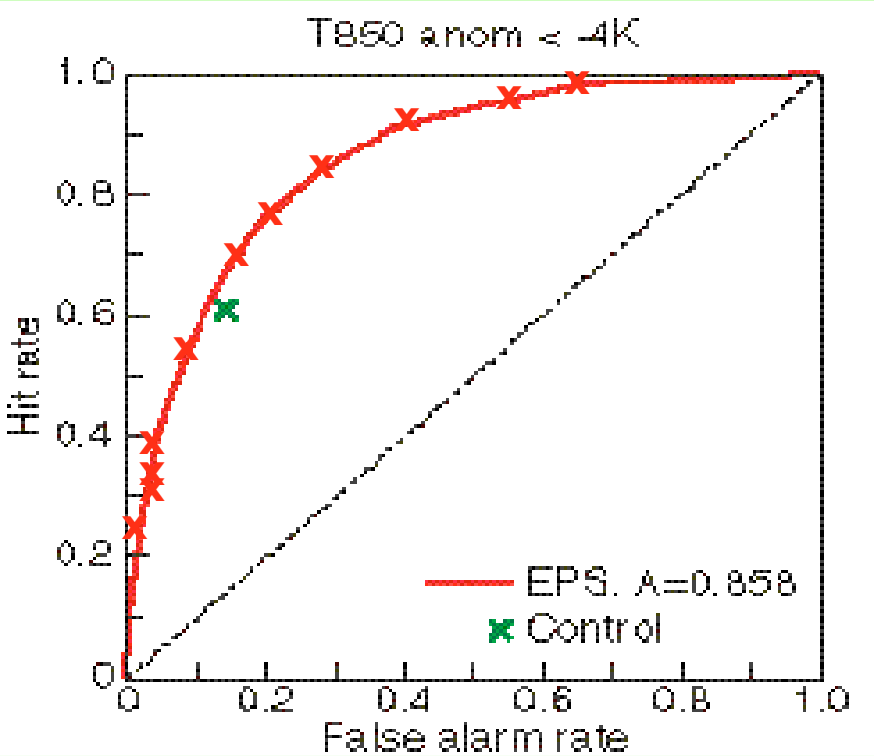
Hit rates are plotted against the corresponding false alarm rates to generate the **ROC Curve**

ROC Curve



The ability of the system to prevent dangerous situations depends on the decision criterion: if we choose to alert when **at least one member** forecasts precipitation exceeding a certain threshold, the Hit Rate will be large enough, but also the False Alarm Rate. If we choose to alert when this is done by **at least a high number of members**, our FAR will decrease, but also our HR

ROC Curve



The area under the ROC curve is used as a statistic measure of forecast usefulness. A value of 0.5 indicates that the forecast system has no skill. In fact, for a system that has no skill, the warnings (W) and the events (E) are independent occurrences:

$$H = p(W|E) = p(W) = p(W|\overline{E}) = F$$

❖ **ROC curve** measures the ability of the forecast to discriminate between two alternative outcomes, thus measuring **resolution**. It is not sensitive to bias in the forecast, so is independent of reliability.

❖ Advantage: is directly related to a decision-theoretic approach and can be easily related to the **economic value** of probability forecasts for forecast users.

Cost-loss Analysis

Is it possible to individuate a threshold for the skill, which can be considered a "usefulness threshold" for the forecast system?

Decisional model		E happens	
		yes	no
U take action	yes	C	C
	no	L	0

- The event E causes a damage which incur a loss **L**. The user U can avoid the damage by taking a preventive action which cost is **C**.
- U wants to minimize the mean total expense over a great number of cases.
- U can rely on a forecast system to know in advance if the event is going to occur or not.

Cost-loss Analysis

Richardson (2000)

contingency table		Observed	
		Yes	No
Forecast	Yes	a	b
	No	c	d

With a deterministic forecast system, the mean expense for unit loss is:

$$\mathbf{ME} = \frac{c * L + (a+b) * C}{L} = F\frac{C}{L}(1-\bar{o}) - H\bar{o}\left(1-\frac{C}{L}\right) + \bar{o}$$

$\bar{o} = a + c$ is the sample climatology (the observed frequency)

If the forecast system is probabilistic, the user has to fix a probability threshold k.

When this threshold is exceeded, it take protective action.

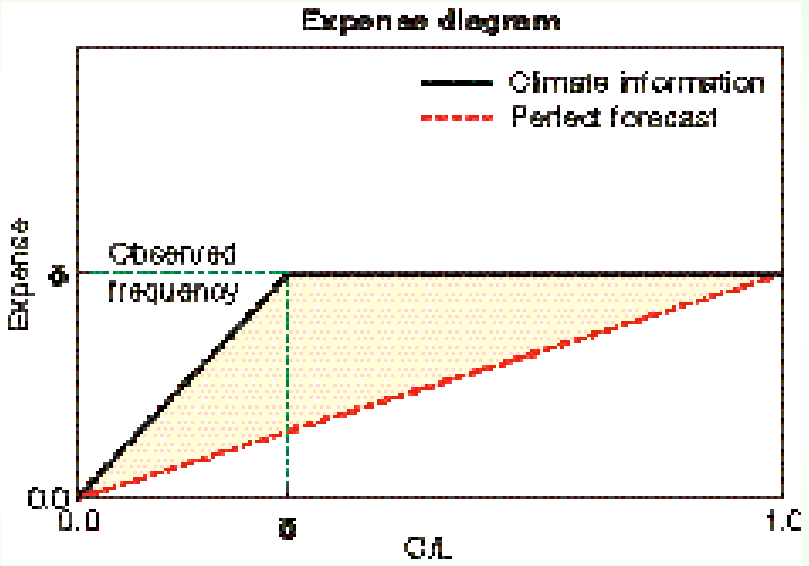
$$\mathbf{ME_k f=} \quad F_k\frac{C}{L}(1-\bar{o}) - H_k\bar{o}\left(1-\frac{C}{L}\right) + \bar{o}$$ Mean expense

Cost-loss Analysis

$$V_k = \frac{MEcli - ME_k f}{MEcli - MEp}$$

Value

Gain obtained using the system instead of the climatological information, percentage with respect to the gain obtained using a perfect system



ME with a perfect forecast system

$$MEp = \bar{o}\frac{C}{L}$$

the preventive action is taken only when the event occurs

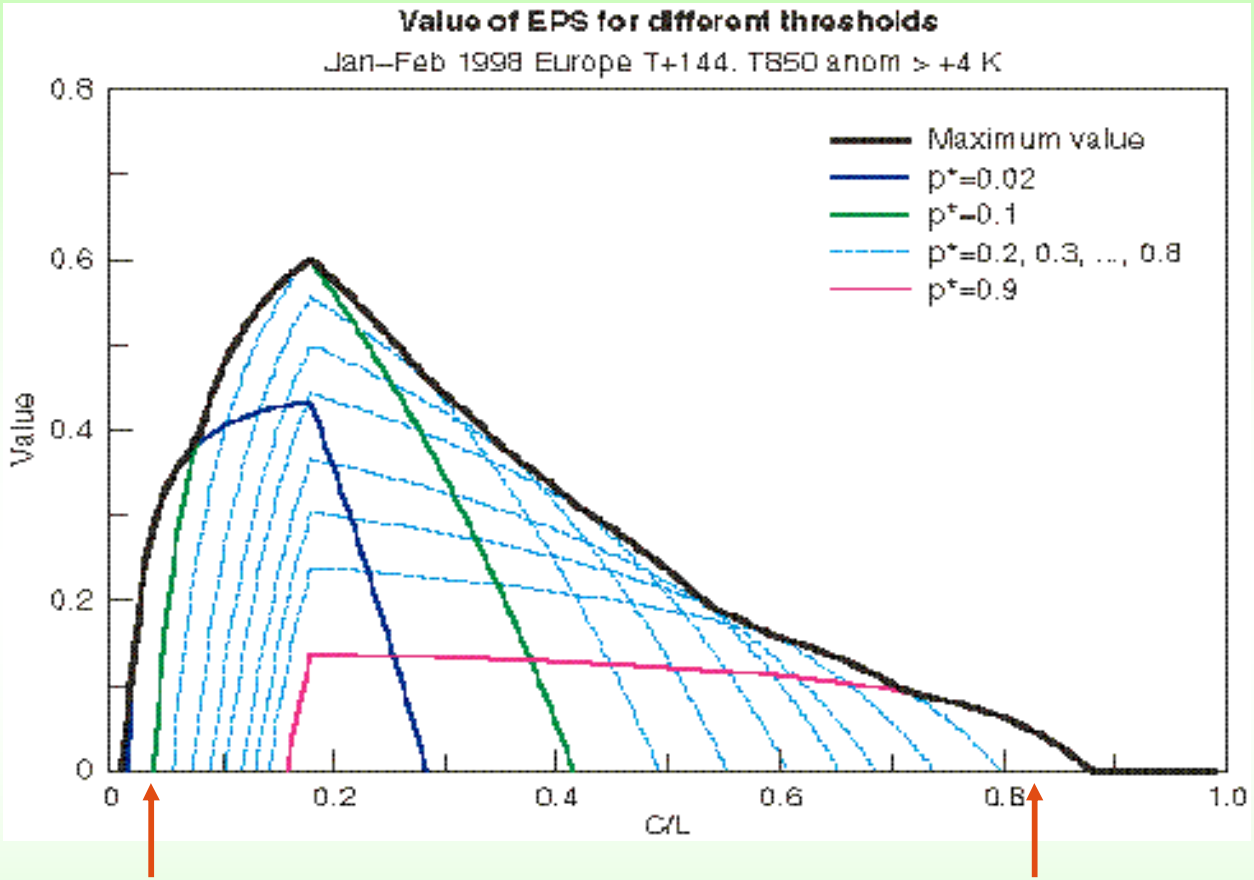
ME based on climatological information

$$MEcli = \min(\bar{o}, \frac{C}{L})$$

the action is always taken if $\frac{C}{L} < \bar{o}$

it is never taken otherwise

Cost-loss Analysis



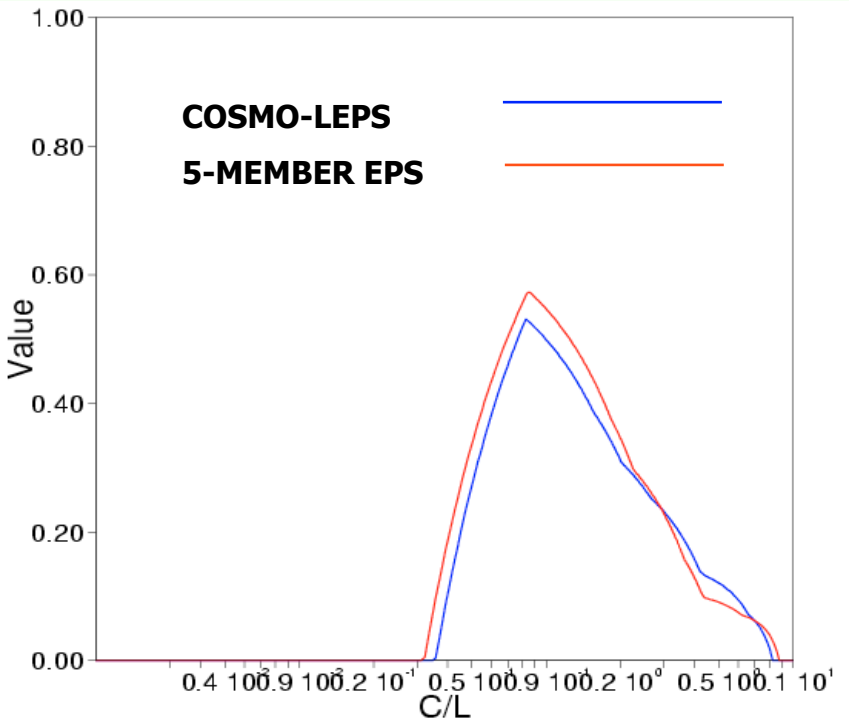
Curves of V_k as a function of C/L, a curve for each probability threshold. The area under the envelope of the curves is the **cost-loss area** (optimum maximum value).

The appropriate probability threshold p_t is equal to C/L (reliable fcs).

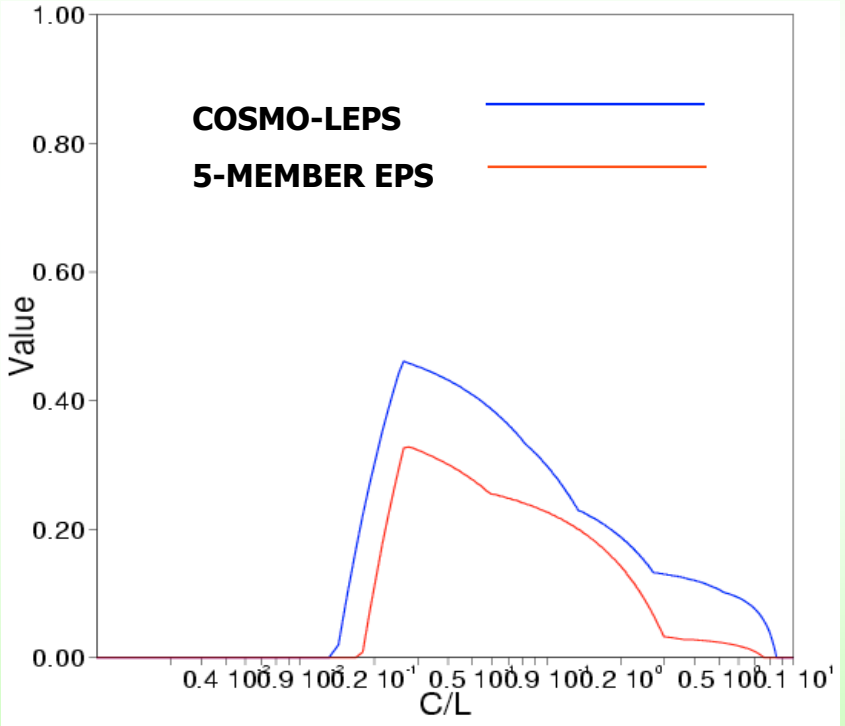
Cost-loss Analysis

The maximum value is shifted towards lower cost-loss ratios for the rarer higher precipitation events. Users with small C/L ratios benefit more from forecasts of rare events.

tp > 10mm/24h



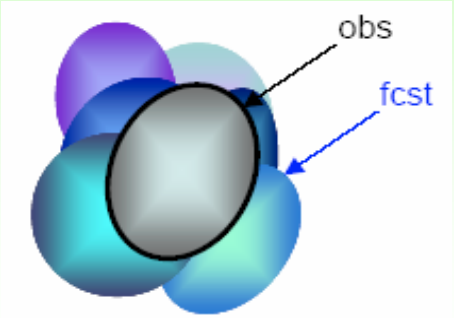
tp > 20mm/24h



Average precipitation fc. range +66h

Object oriented verification

Ebert and McBride (2000)



- verification of the properties of spatial forecast of entities (e.g. contiguous rain areas – CRAs)
- for each entity that can be identified in the forecast and in the observations, a pattern matching technique is used to determine the location error and errors in area, mean and maximum intensity, spatial pattern
- the verified entities can be classified as “hits”, “misses”, etc.

Statistical significance - bootstrap

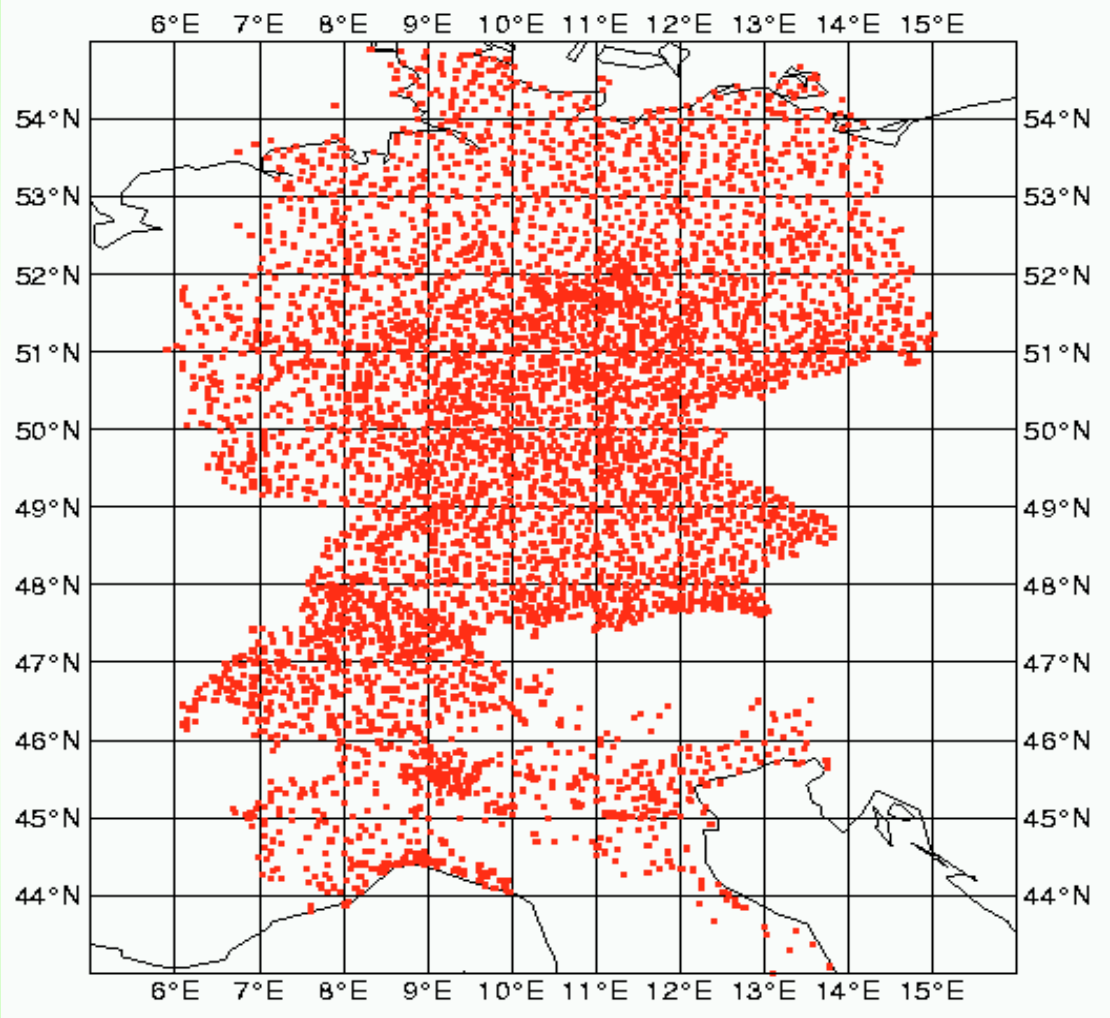
Wilks (1995), Hamill (1999)

comparison between two systems: does one ensemble perform significantly better than another? Is BSS_{M1} significantly different from BSS_{M2}?

- re-sampled test statistics consistent with the null hypothesis are generated after randomly choosing (e.g. 1000 times) either one or the other ensemble for each point and on each case day. Then, 1000 BSS* have been computed over all points and over all days and the difference between each couple of BSS* has been calculated (BSS^*_1-BSS^*_2)

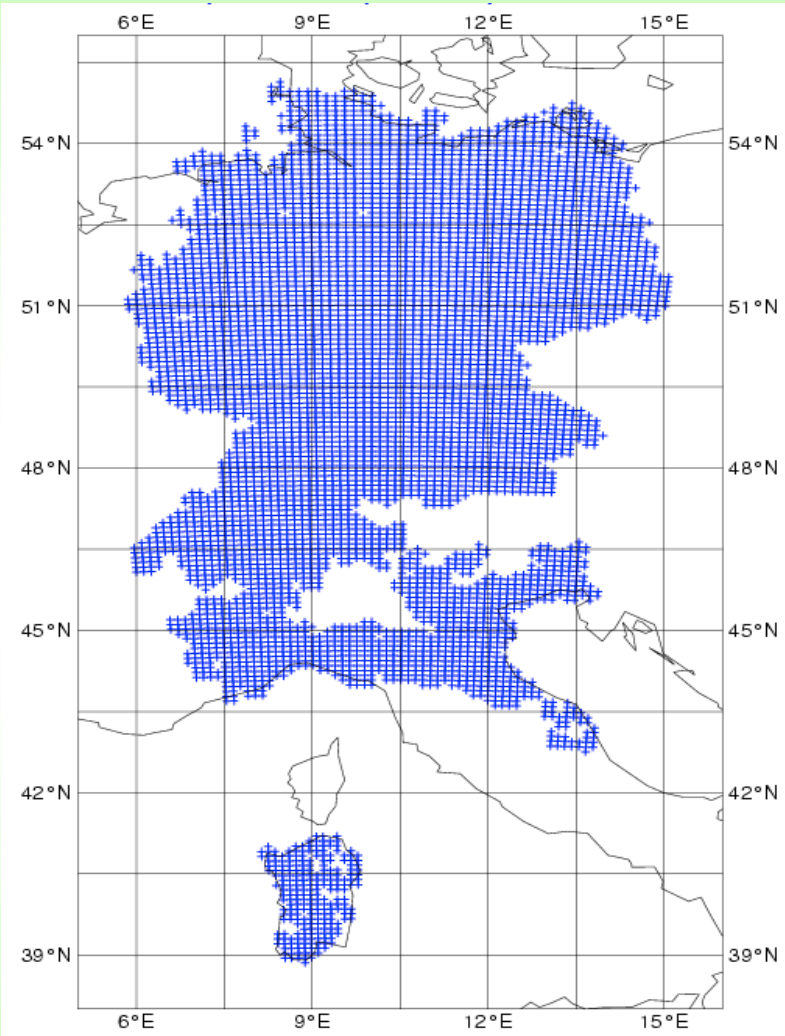
- compare the test statistic with the null distribution: determine the location of BSS_{M1}-BSS_{M2} in the re-sampled distribution

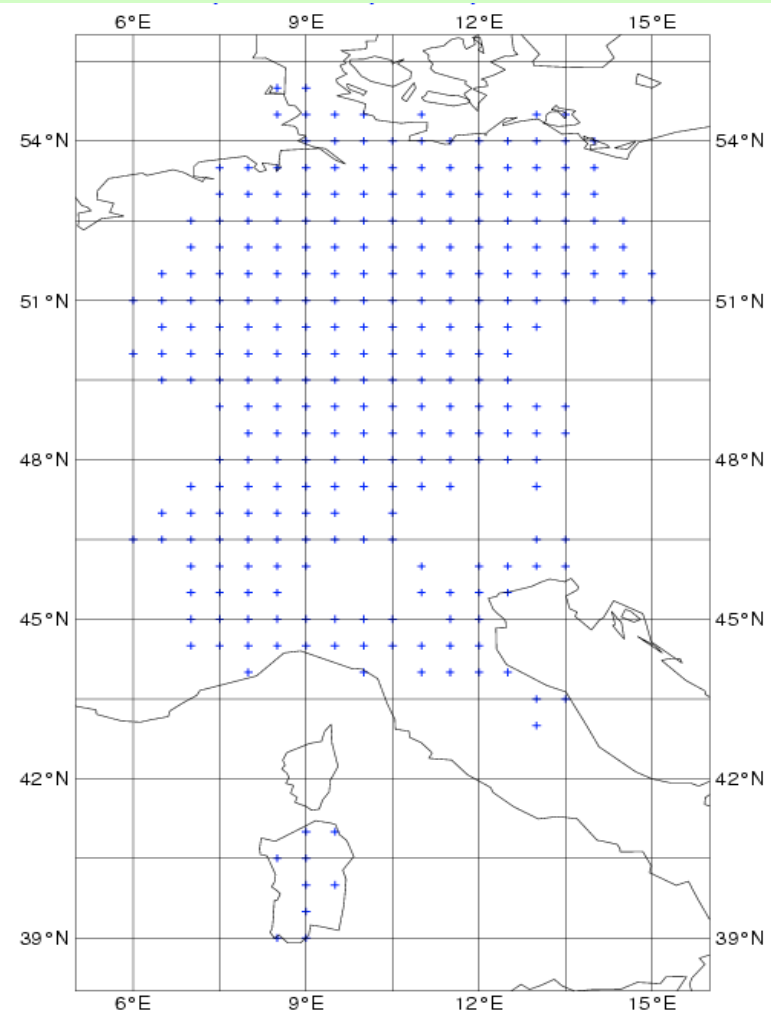
COSMO observations



+ Poland

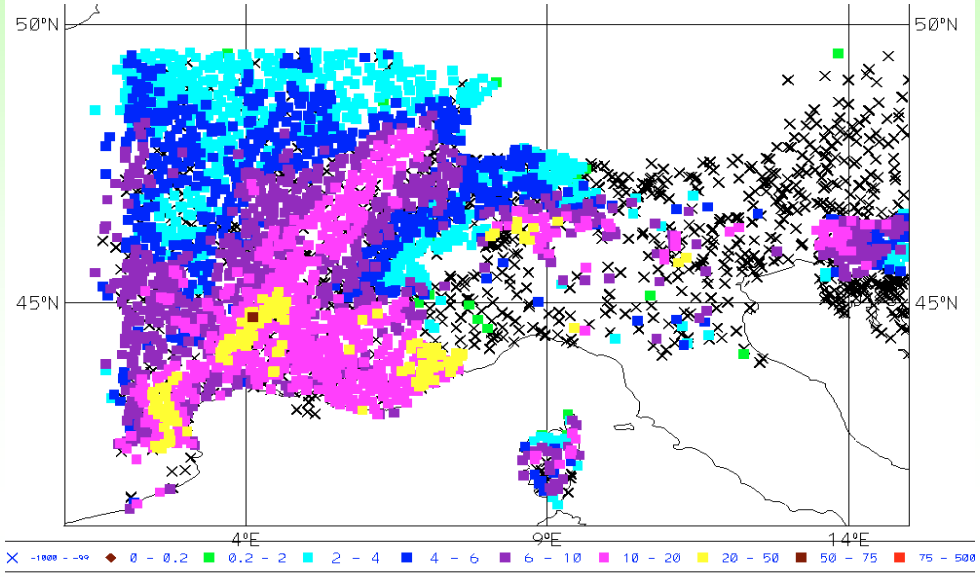
obs mask 1.5 x 1.5



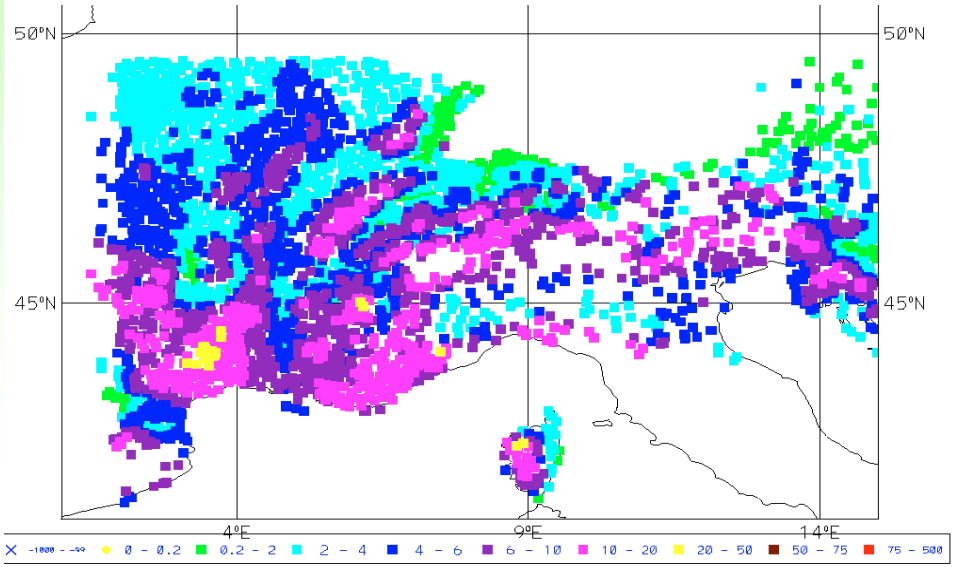


15-cases "climate"

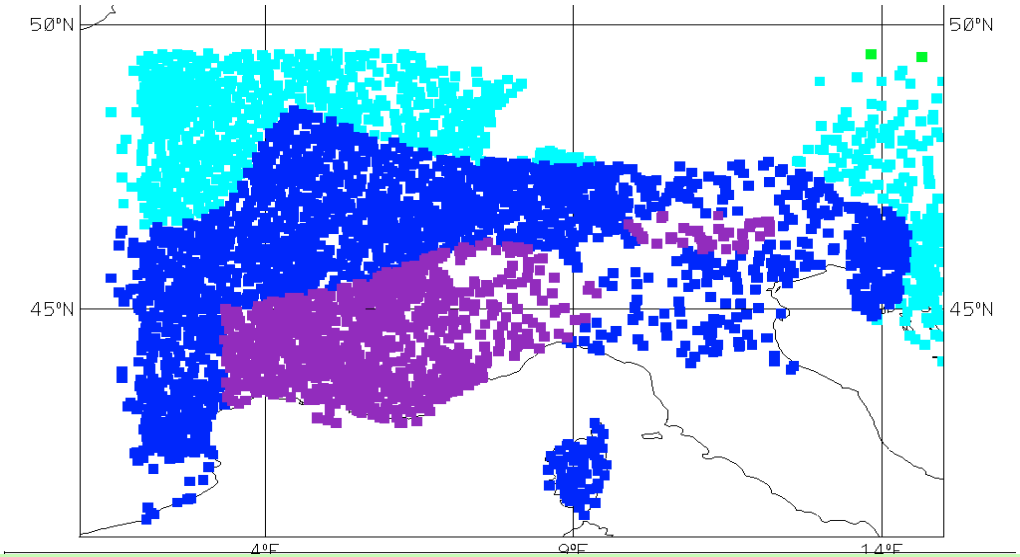
observed



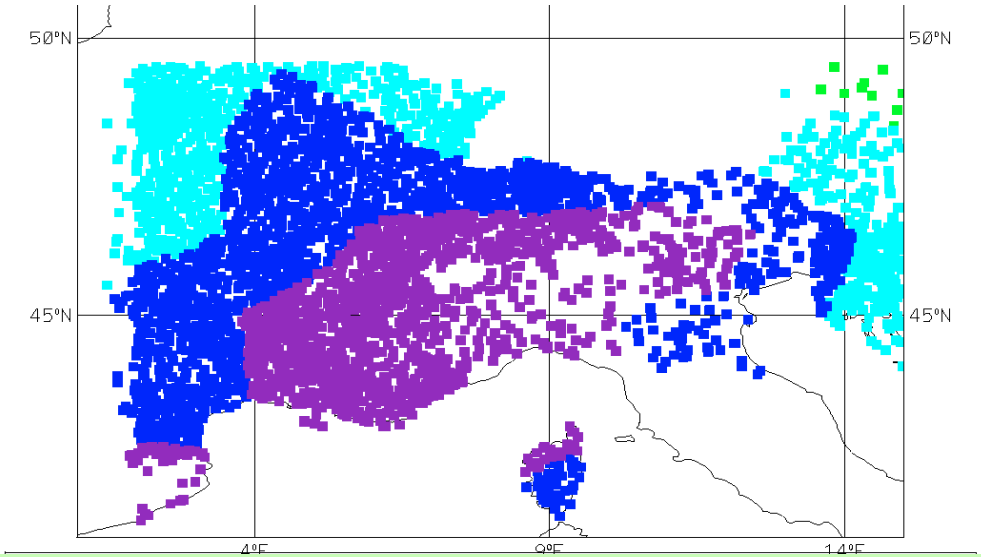
LEPS forecast



EPS forecast

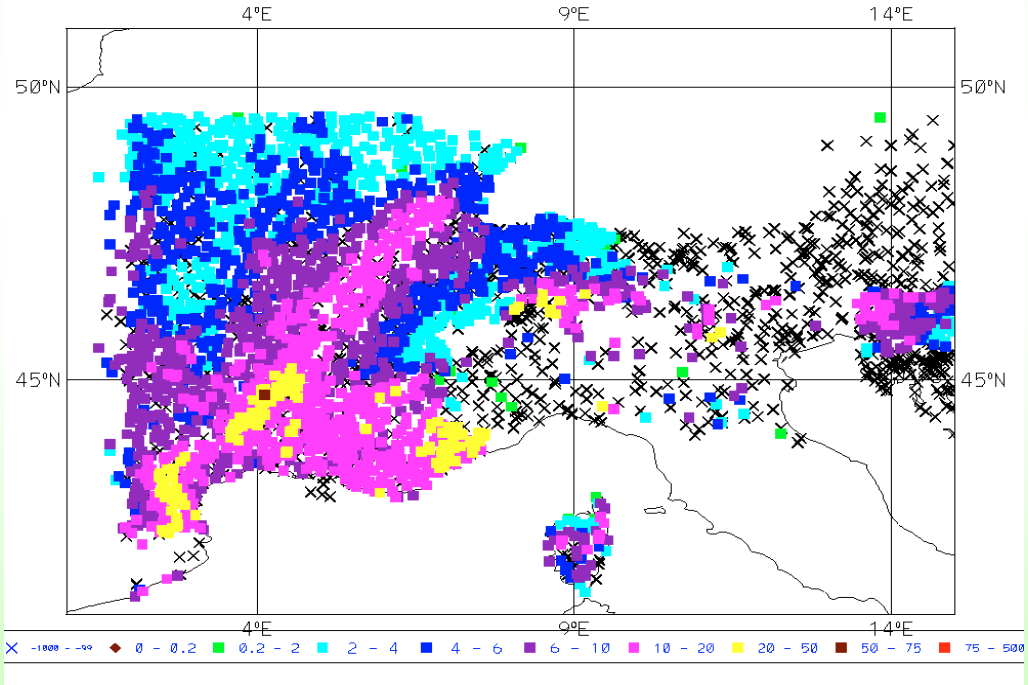


EPSRM forecast

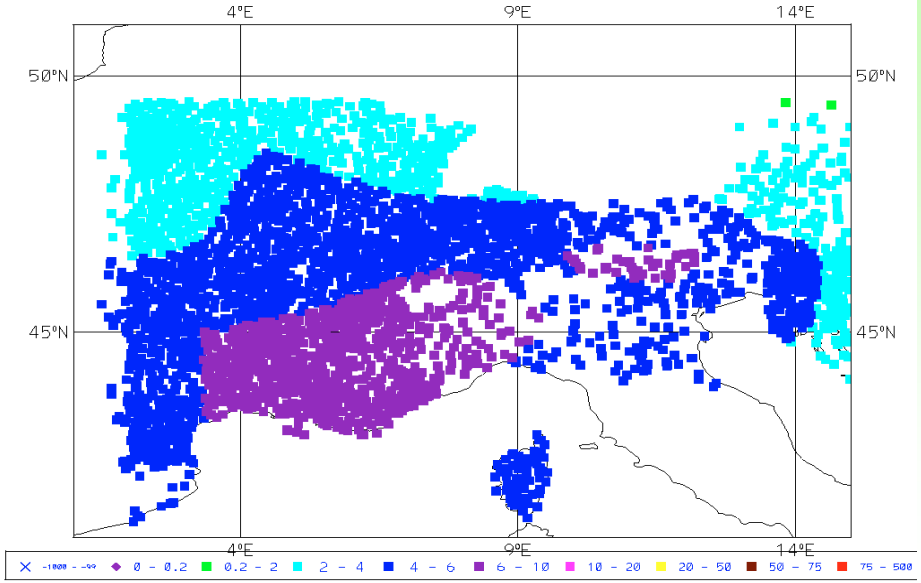


15-cases observed vs. forecast "climate" (average)

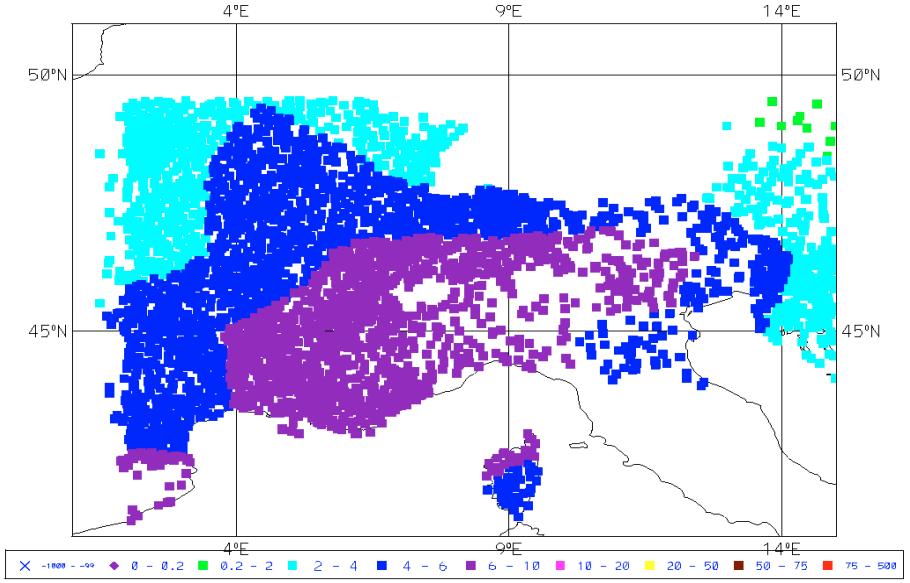
observed



EPS forecast



EPSRM forecast



Fine

bibliography - review

- www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html
- www.ecmwf.int
- http://meted.ucar.edu/nwp/pcu1/ensemble/print.htm#5.2.2
- Bougeault, P., 2003. WGNE recommendations on verification methods for numerical prediction of weather elements and severe weather events (CAS/JSC WGNE Report No. 18)
- Jolliffe, I.T. and Stephenson D.B. (Editors), 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Sciences. Wiley, 240 pp.
- Pertti Nurmi, 2003. Recommendations on the verification of local weather forecasts. ECMWF Technical Memorandum n. 430.
- Stanski, H.R., Wilson L.J. and Burrows W.R., 1989. Survey of Common Verification Methods in Meteorology (WMO Research Report No. 89-5)
- Wilks D. S., 1995. Statistical methods in atmospheric sciences. Academic Press, New York, 467 pp.

bibliography - papers

- Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.
- Candille, G. and Talagrand, O., 2004. On limitations to the objective evaluation of ensemble prediction systems. *Workshop on Ensemble Methods*, UK Met Office, Exeter, October 2004.
- Ebert, E.E. and McBride, J.L., 2000. Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179-202.
- Ebert, E.E., 2005. Verification of ensembles. *TIGGE Workshop*, ECMWF, Reading, March 2005.
- Epstein, E.S., 1969. A scoring system for probabilities of ranked categories. *J. Appl. Meteorol.*, **8**, 985-987.
- Hamill, T.M., 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

bibliography - papers

- Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559-570.
- Mason S.J. and Graham N.E., 1999. Conditional probabilities, relative operating characteristics and relative operating levels. *Wea. Forecasting*, **14**, 713-725.
- Murphy A.H., 1971. A note on the ranked probability score. *J. Appl. Meteorol.*, **10**, 155-156.
- Murphy A.H., 1973. A new vector partition of the probability score. *J. Appl. Meteorol.*, **12**, 595-600.
- Murphy A.H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.
- Richardson D.S., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteorol. Soc.*, **126**, 649-667.

bibliography - papers

❖ Talagrand, O., Vautard R. and Strauss B., 1999. Evaluation of probabilistic prediction systems. *Proceedings of ECMWF Workshop on Predictability,* 20-22 October 1997.

❖ Toth, Z., Zhu, Y. and Marchok, T., 2001. The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting*, **16**, 463-477.

❖ Toth, Z., Talagrand O., Candille, G. and Zhu, Y., 2003. Probability and Ensemble Forecasts. In: Jolliffe, I.T. and Stephenson D.B. (Editors), 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Sciences. Wiley, 240 pp.

❖ Ziehmann, C., 2001. Skill prediction of local weather forecasts based on the ECMWF ensemble. *Nonlinear Processes in Geophysics*, **8**, 419-428.