Probabilistic verification

Chiara Marsigli

with the help of the WG and Laurie Wilson in particular

Goals of this session

- Increase understanding of scores used for probability forecast verification
  - Characteristics, strengths and weaknesses
- Know which scores to choose for different verification questions

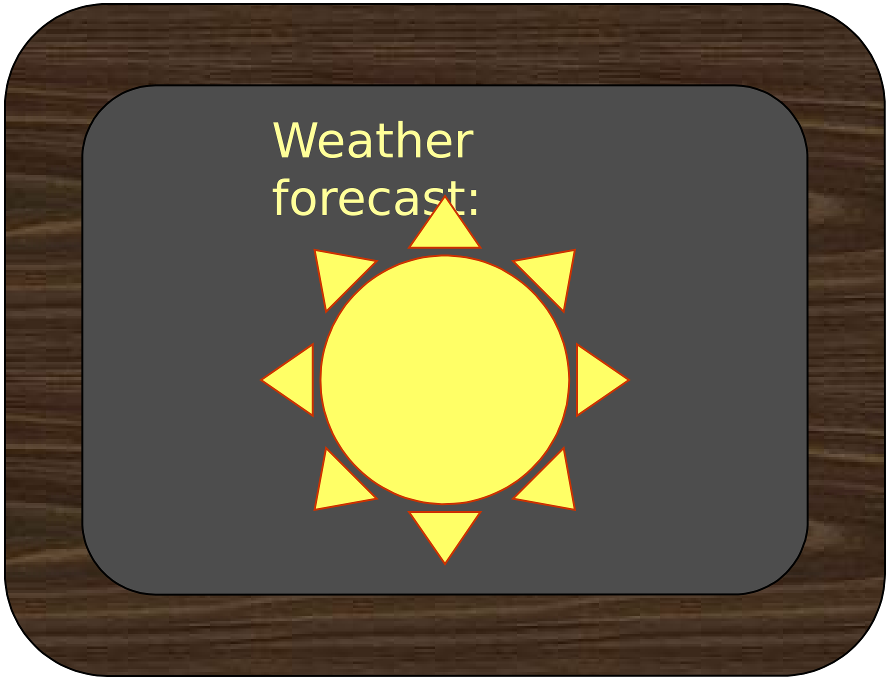
Topics

- Introduction: review of essentials of probability forecasts for verification
- Brier score: *Accuracy*
- Brier skill score: *Skill*
- Reliability Diagrams: *Reliability, resolution* and *sharpness*
  - Exercise
- *Discrimination*
  - Exercise
- Relative operating characteristic
  - Exercise
- Ensembles: The CRPS and Rank Histogram

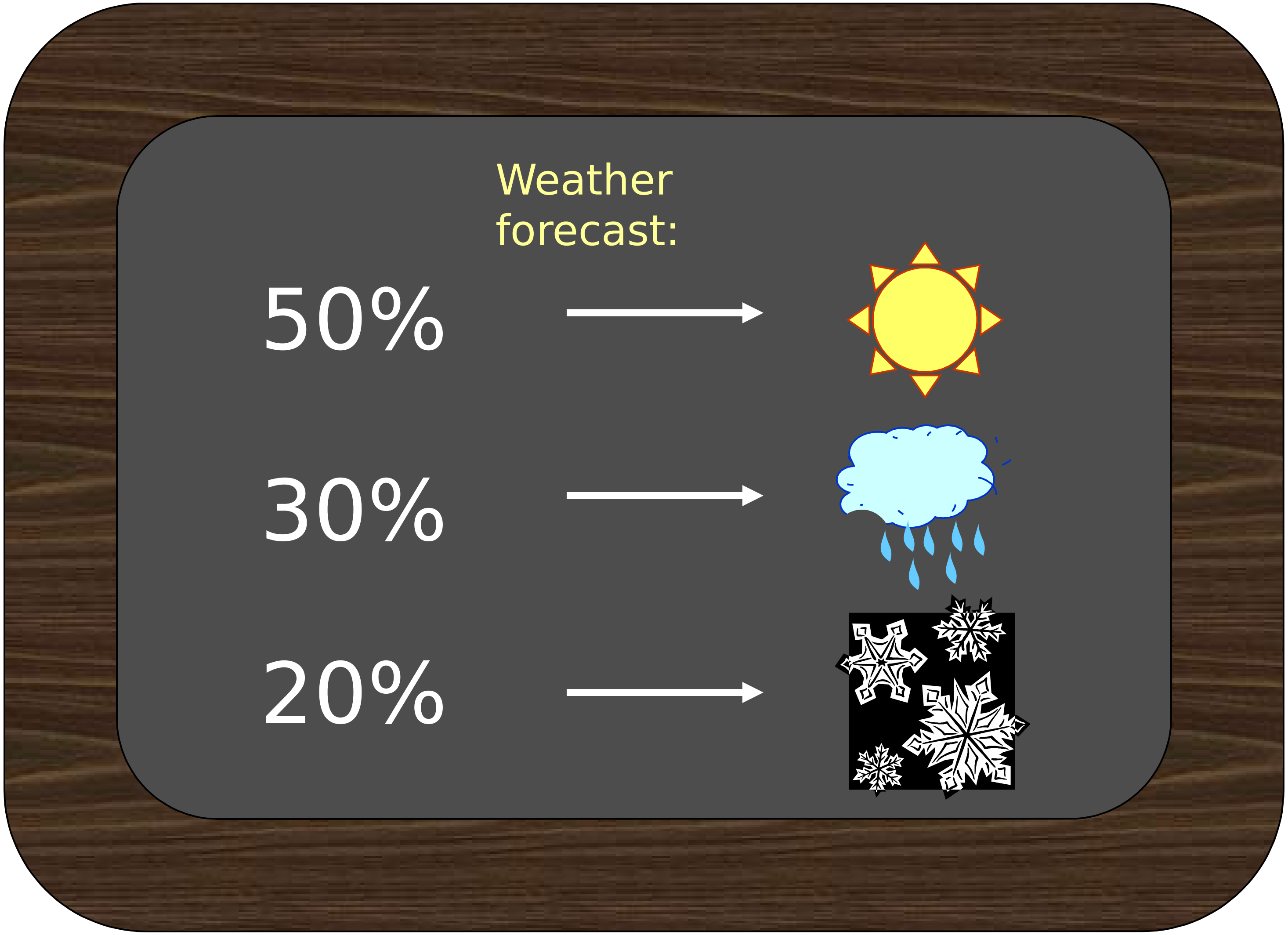
Probability forecast

- Applies to a specific, completely defined event
  - Examples: Probability of precipitation over 6h
  - …
- Question: What does a probability forecast “POP for Melbourne for today (6am to 6pm) is 0.40” mean?

Deterministic approach



Probabilistic approach



Weather forecast:

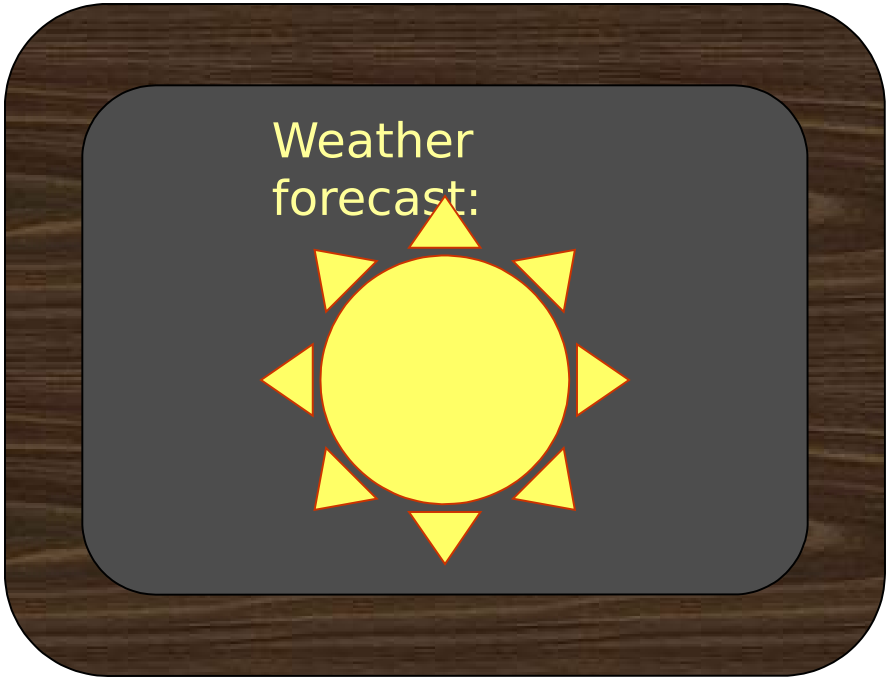
50% →

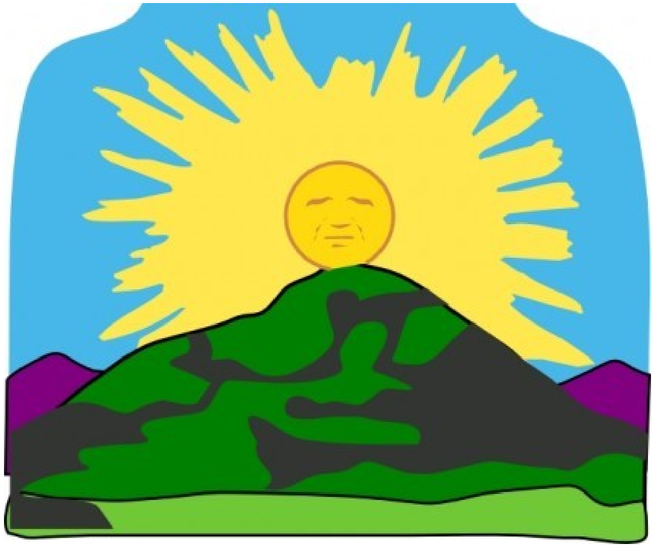
30% →

20% →

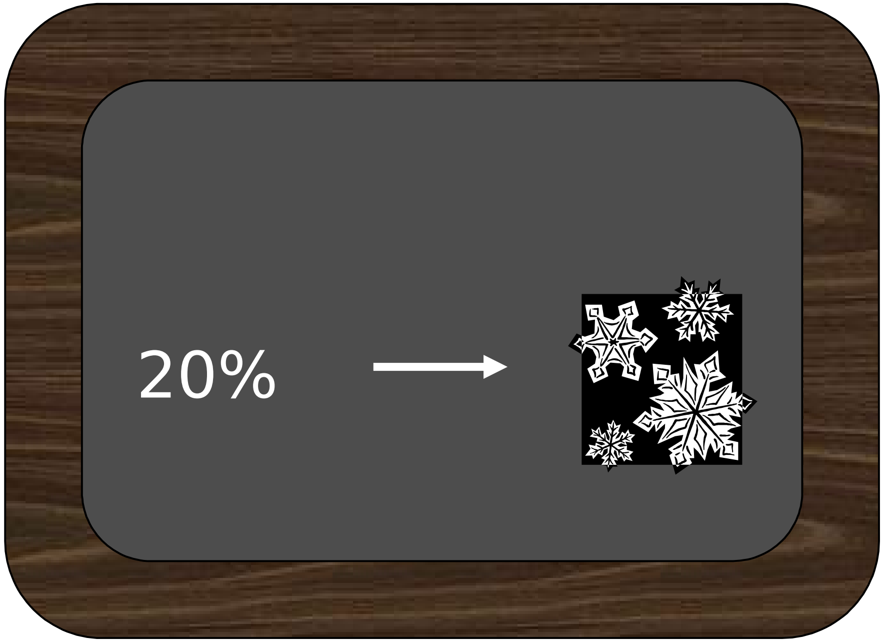
?

Deterministic approach



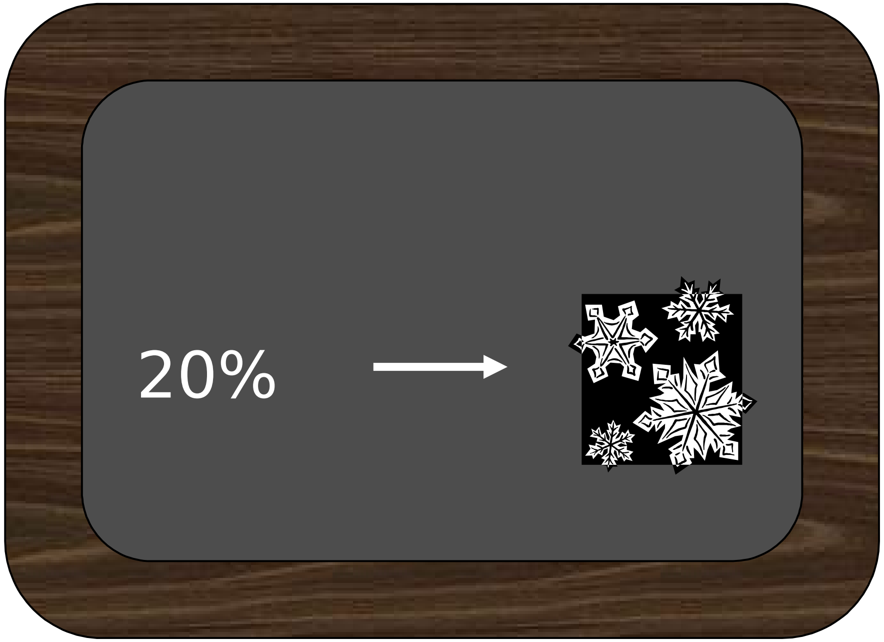


Probabilistic approach



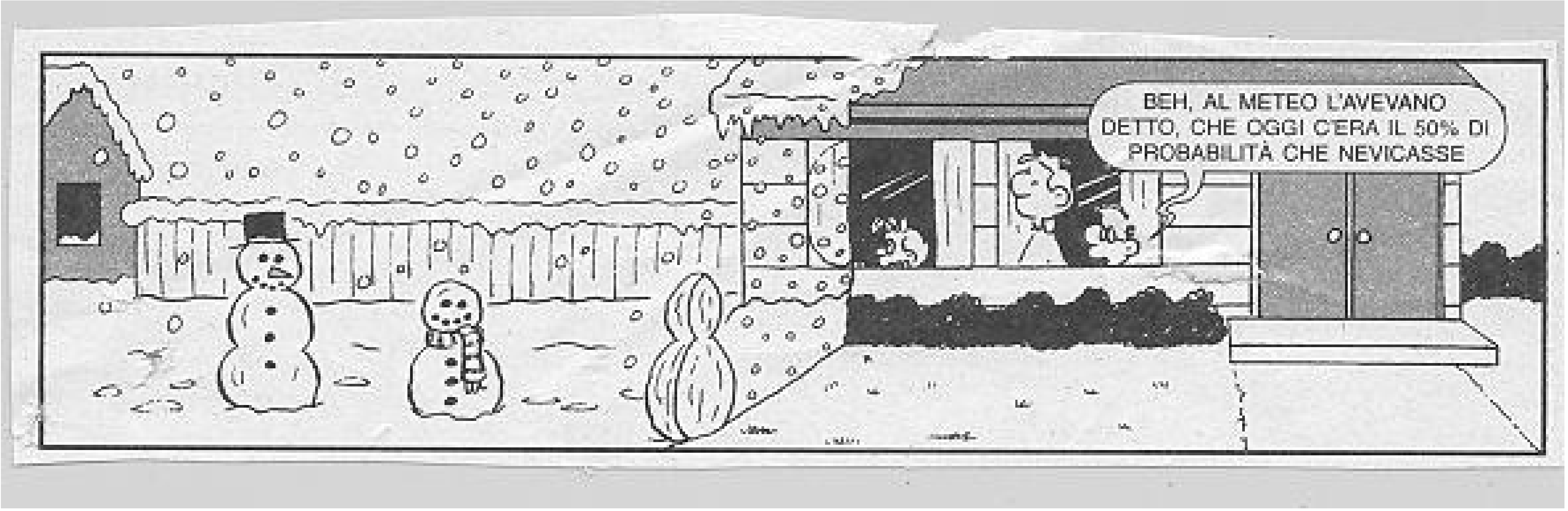
20%

Probabilistic approach



20%

Probabilistic approach


BEH, AL METEO L'AVEVANO DETTO, CHE OGGI C'ERA IL 50% DI PROBABILITÀ CHE NEVICASSE

Deterministic forecast

event E

e. g.: 24 h accumulated precipitation on one point (raingauge, radar pixel, catchment, area) exceeds 20 mm

no
o(E) = 0

event is observed with frequency o(E)

yes
o(E) = 1

no
p(E) = 0

event is forecasted with probability p(E)

yes
p(E) = 1

Probabilistic forecast

event E

e. g.: 24 h accumulated precipitation on one point (raingauge, radar pixel, catchment, area) exceeds 20 mm

no
$o(E) = 0$

event is observed with frequency o(E)

yes
$o(E) = 1$

event is forecasted with probability p(E)

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

Ensemble forecast

event E

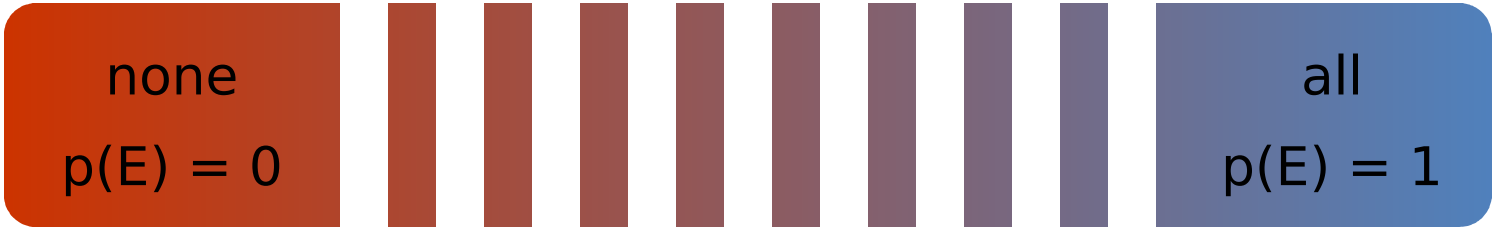
e. g.: 24 h accumulated precipitation on one point (raingauge, radar pixel, catchment, area) exceeds 20 mm

no
o(E) = 0

event is observed with frequency o(E)

sì
o(E) = 1

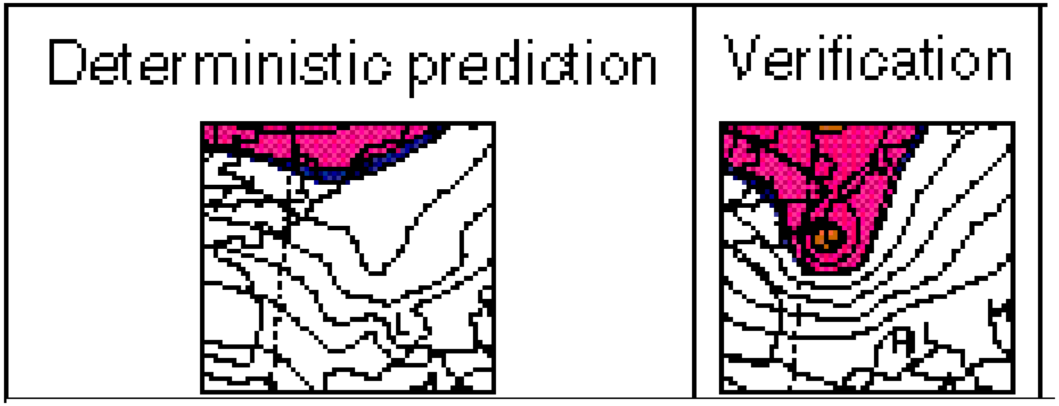
ensemble of M elements
event is forecasted with probability p(E) = k/M



none
p(E) = 0

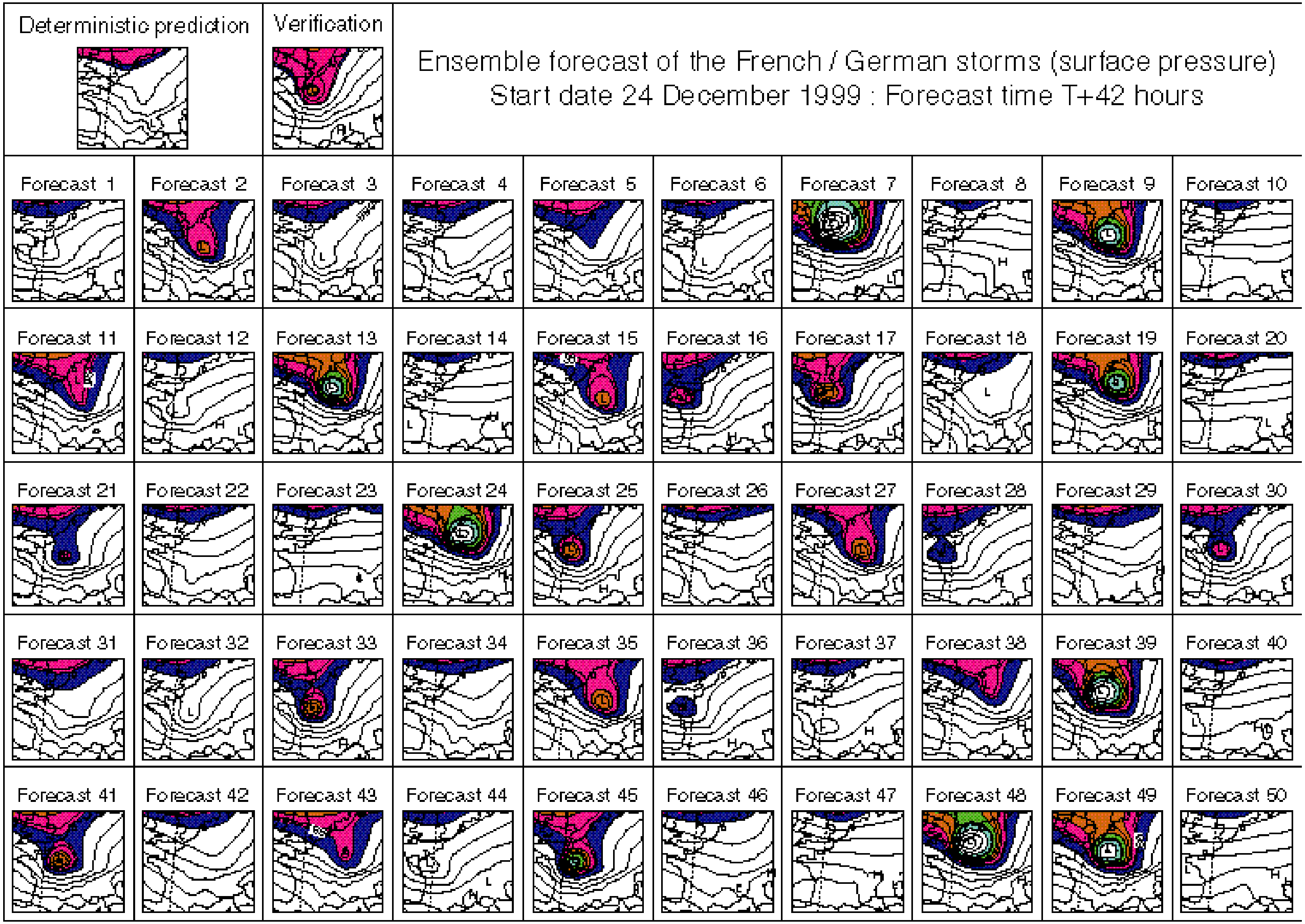
all
p(E) = 1

Deterministic approach

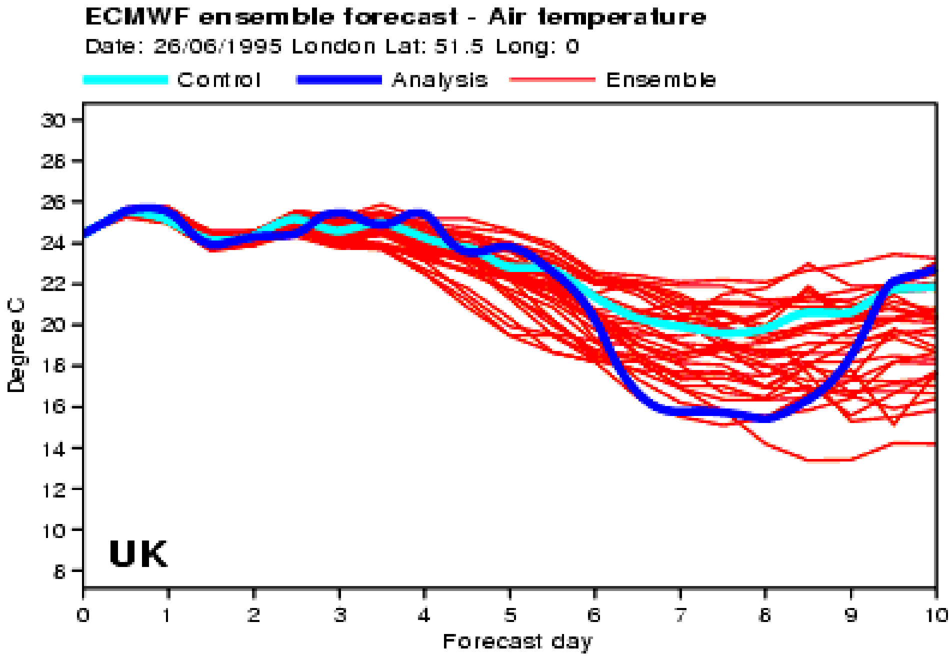


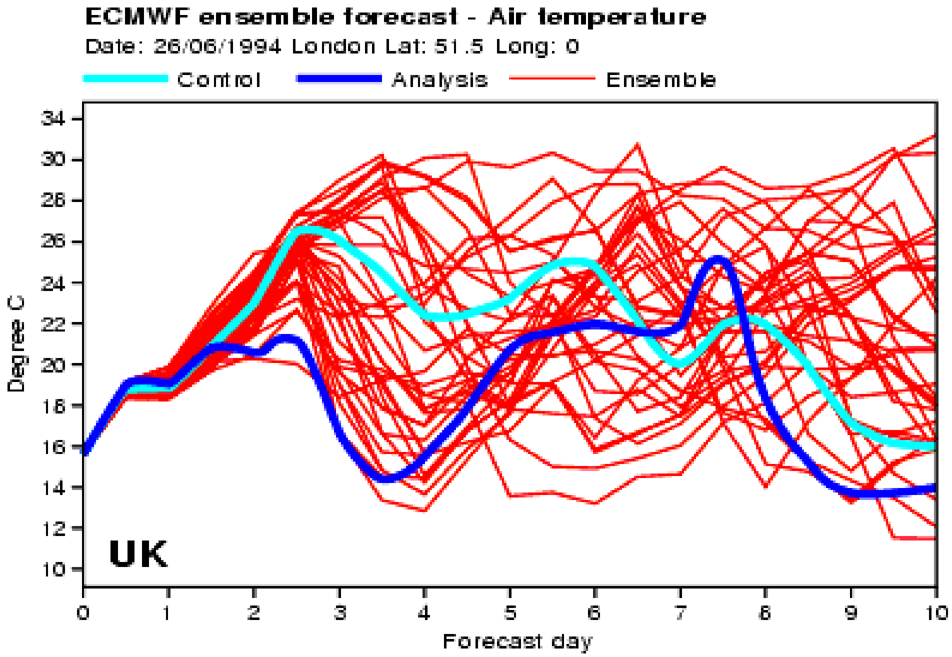
forecast of the French / German storms (surface pressure)
Start date 24 December 1999 : Forecast time T+42 hours

Probabilistic approach



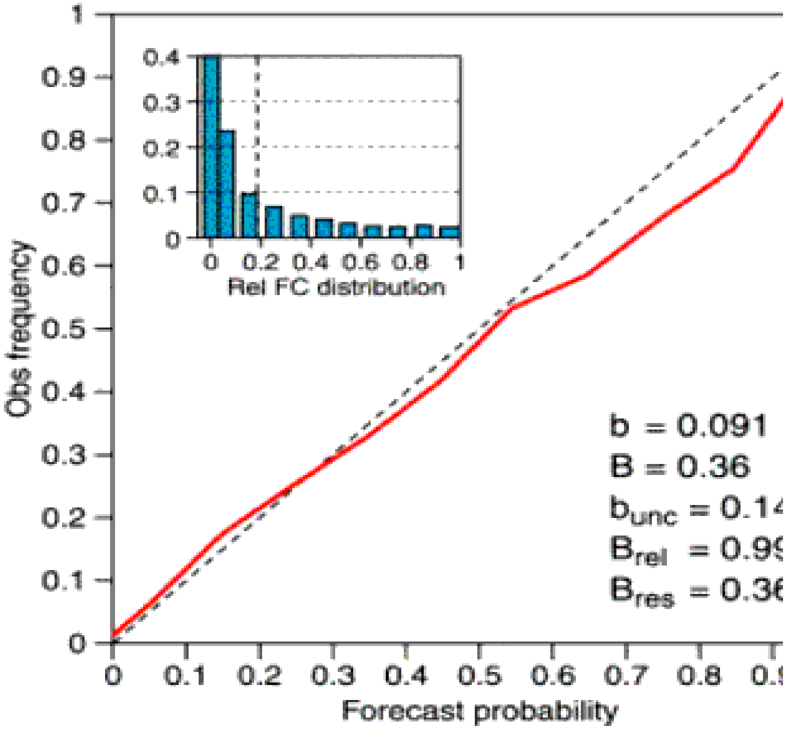
Ensemble forecast





Forecast evaluation

- Verification is possible only in statistical sense, not for one single issue
- E.g.: correspondence between forecast probabilities and observed frequencies
- Dependence on the ensemble size



Brier Score

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

Scalar summary measure for the assessment of the forecast performance, mean square 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 can take on values in the range [0,1], a perfect forecast having BS = 0

Brier Score

- Gives result on a single forecast, but cannot get a perfect score unless forecast categorically.
- A “summary” score – measures accuracy, summarized into one value over a dataset.
- Weights larger errors more than smaller ones.
- 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 – components of the error

Components of probability error

The Brier score can be decomposed into 3 terms (for *K* probability classes and a sample of size *N*):

$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k (p_k - \overline{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k (\overline{o}_k - \overline{o})^2 + \overline{o}(1 - \overline{o})$$

reliability

If for all occasions when forecast probability p_k is predicted, the observed frequency of the event is $\overline{o}_k = p_k$ then the forecast is said to be reliable. Similar to bias for a continuous variable

resolution

The ability of the forecast to distinguish situations with distinctly different frequencies of occurrence.

uncertainty

The variability of the observations. Maximized when the climatological frequency (*base rate*) =0.5

Has nothing to do with forecast quality! Use the Brier skill score to overcome this problem.

The presence of the uncertainty term means that Brier Scores should not be compared on different samples.

Probabilistic forecasts

An accurate probability forecast system has:

- reliability - agreement between forecast probability and mean observed frequency
- sharpness - tendency to forecast probabilities near 0 or 1, as opposed to values clustered around the mean
- resolution - ability of the forecast to resolve the set of sample events into subsets with characteristically different outcomes

Brier Score decomposition

Murphy (1973)

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

reliability

resolution

uncertainty

M = ensemble size

K = 0, ..., M number of ensemble members forecasting the event (probability classes)

N = total number of point in the verification domain

N_k = number of points where the event is forecast by k members

$\overline{o}_k = \sum_{i=1}^{N_k} o_i$ = frequency of the event in the sub-sample N_k

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

Brier Score decomposition

Murphy (1973)

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

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.

Brier Score decomposition

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

reliability

resolution

uncertainty

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 within the sample of years studied, then b_{unc}=0; in this case, always forecasting the climatological probability generally gives good results. When the climatological probability is near 0.5, there is substantially more uncertainty inherent in the forecasting situation: if E occurred 50% of the time within the sample, then b_{unc}=0.25. Uncertainty is a function of the climatological frequency of E, and is not dependent on the forecasting system itself.

Brier Score decomposition II

Talagrand et al. (1997)

$$BS = \bar{o} \sum_{k=0}^{M} H_k \left[1 - \frac{k}{M} \right]^2 + (1 - \bar{o}) \sum_{k=0}^{M} F_k \left[\frac{k}{M} \right]^2$$

Hit Rate term

False Alarm Rate term

M = ensemble size

K = 0, …, M number of ensemble members forecasting the event (probability classes)

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

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

Brier Skill Score

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

$$BSS = \frac{BS - BS_{ref}}{BS_{ref}}$$

The forecast system has predictive skill if BSS is positive, a perfect system having BSS = 1.

IF the sample climatology $\bar{o}$ is used, can be expressed as:

$$BSS = -\frac{\mathrm{Res} - \mathrm{Rel}}{\mathrm{Unc}} \qquad BS_{cli} = \bar{o}\left(1 - \bar{o}\right)$$

Brier Score and Skill Score - Summary

- Measures accuracy and skill respectively
- “Summary” scores
- Cautions:
  - Cannot compare BS on different samples
  - BSS – take care about underlying climatology
  - BSS – Take care about small samples

Ranked Probability Score

$$RPS = \frac{1}{M-1}\sum_{m=1}^{M}\left[\left(\sum_{k=1}^{m} f_k\right) - \left(\sum_{k=1}^{m} o_k\right)\right]^2$$

Extension of the Brier Score to multi-event situation.
The squared errors are computed with respect to the cumulative probabilities in the forecast and observation vectors.

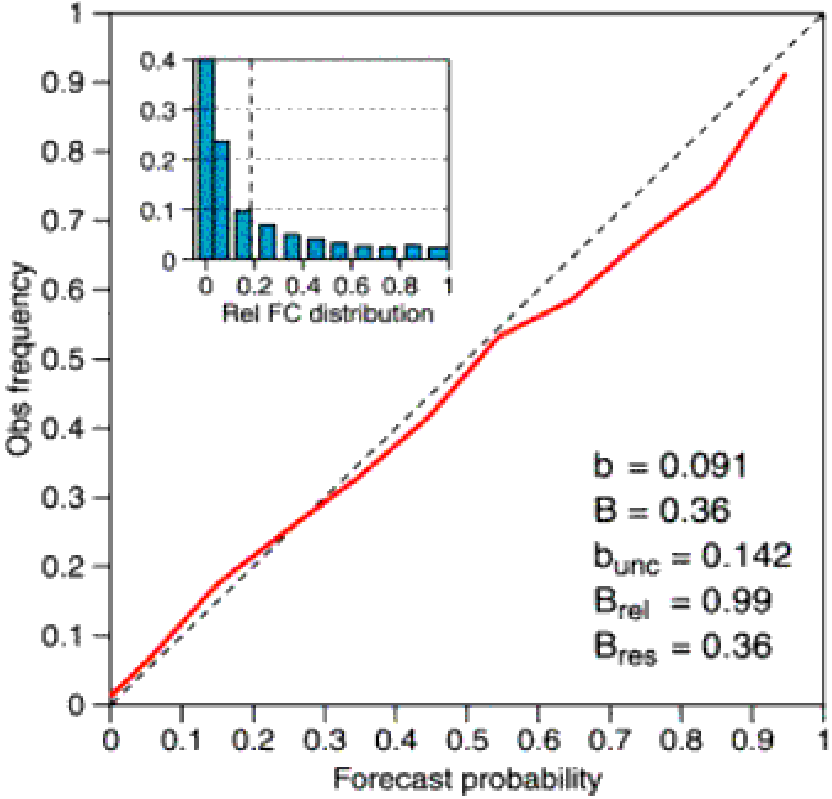
- M = number of forecast categories
- o_{ik} = 1 if the event occurs in category k

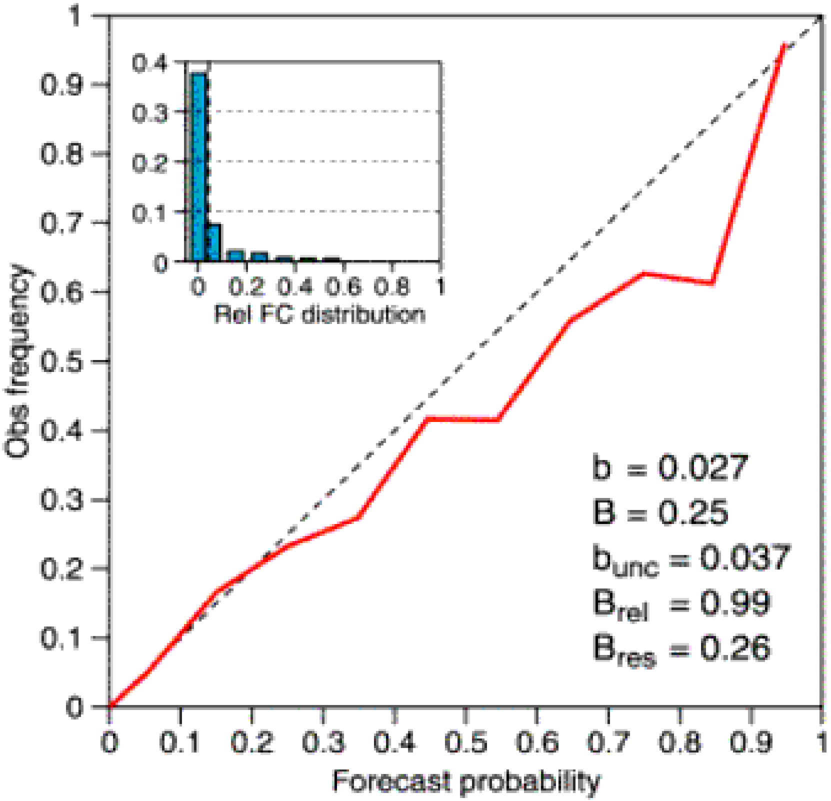
  = 0 if the event does not occur in category k
- f_k is the probability of occurrence in category k according to the forecast system (e.g. the fraction of ensemble members forecasting the event)
- RPS take on values in the range [0,1], a perfect forecast having RPS = 0

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



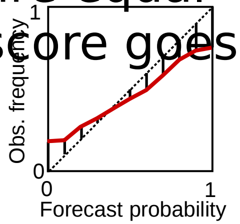


Reliability Diagram

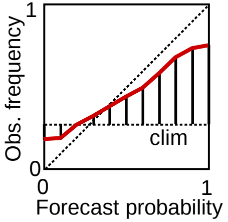
Reliability: Proximity to diagonal

Resolution: Variation about horizontal (climatology) line

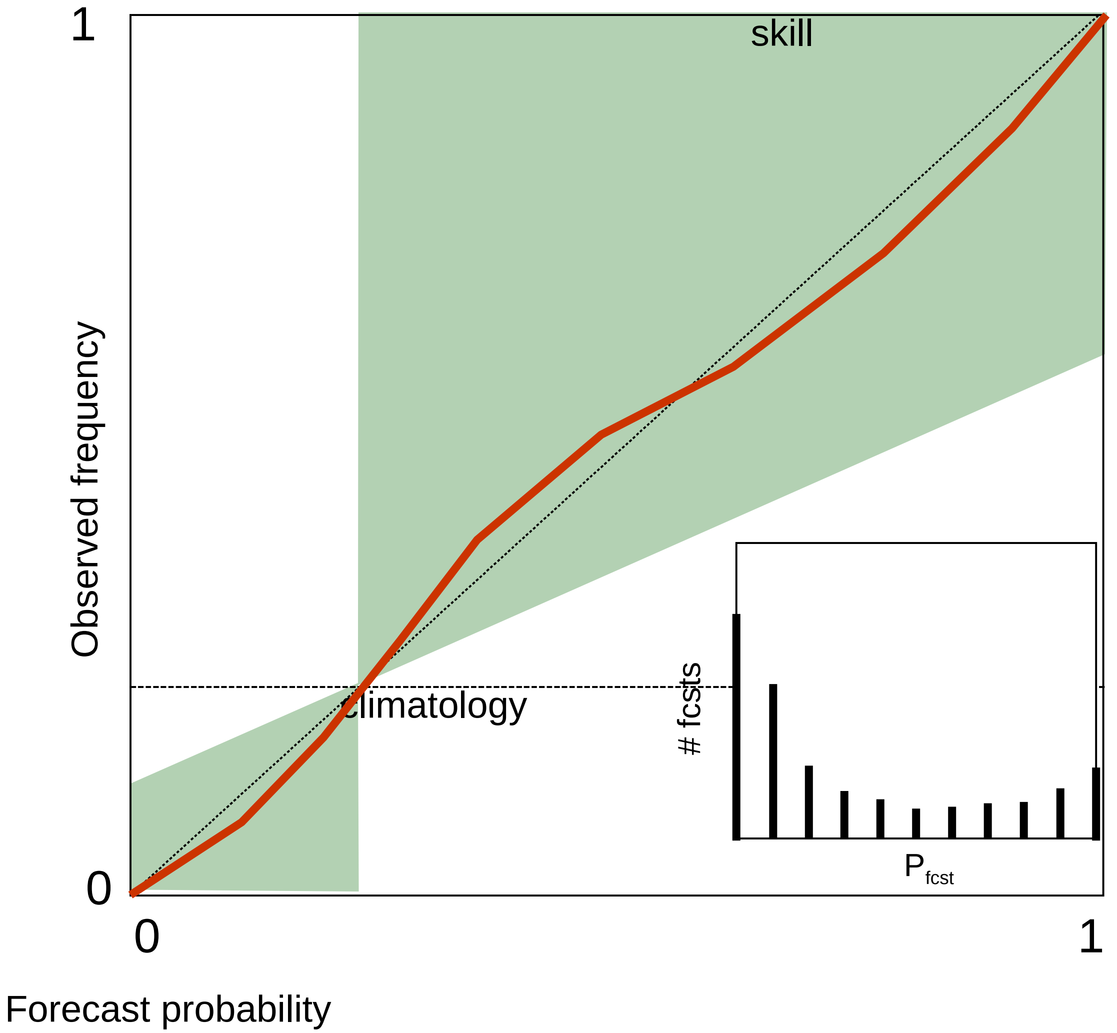
No skill line: Where reliability and resolution are equal – Brier skill score goes to 0



Reliability

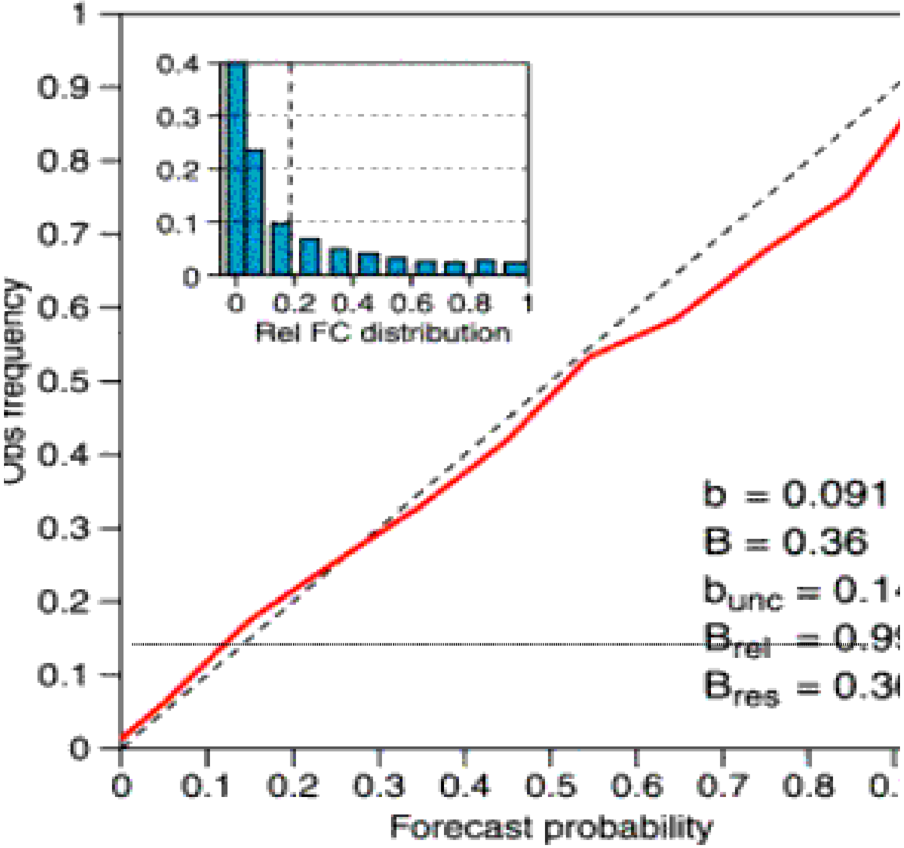


Resolution



Reliability Diagram and Brier Score

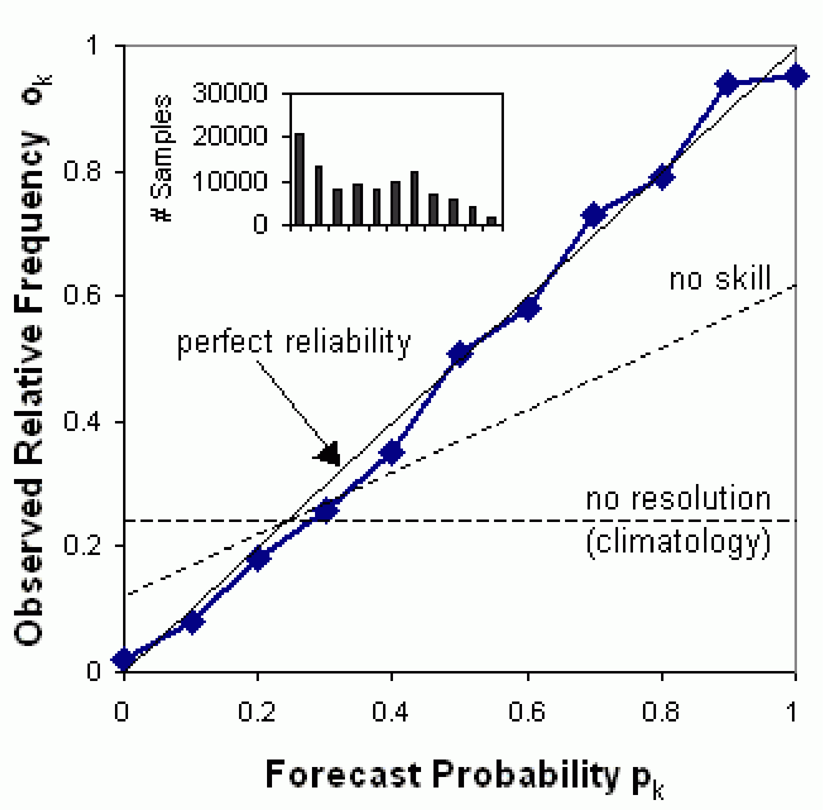
The reliability term measures the mean square distance of the graph of o(p) to the diagonal line.

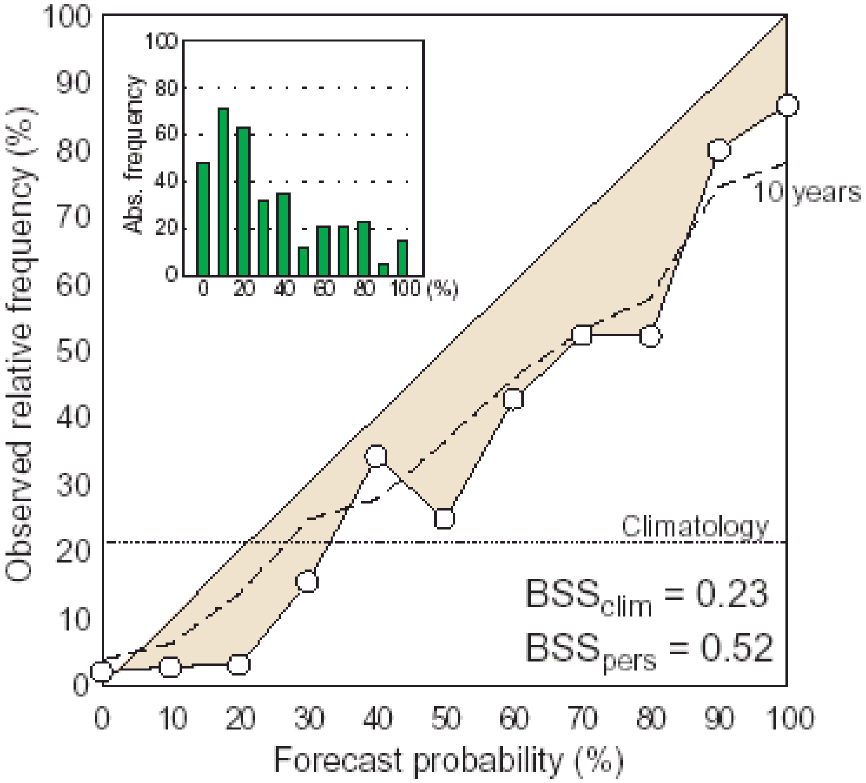


Points between the "no skill" line and the diagonal contribute positively to the Brier skill score.

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

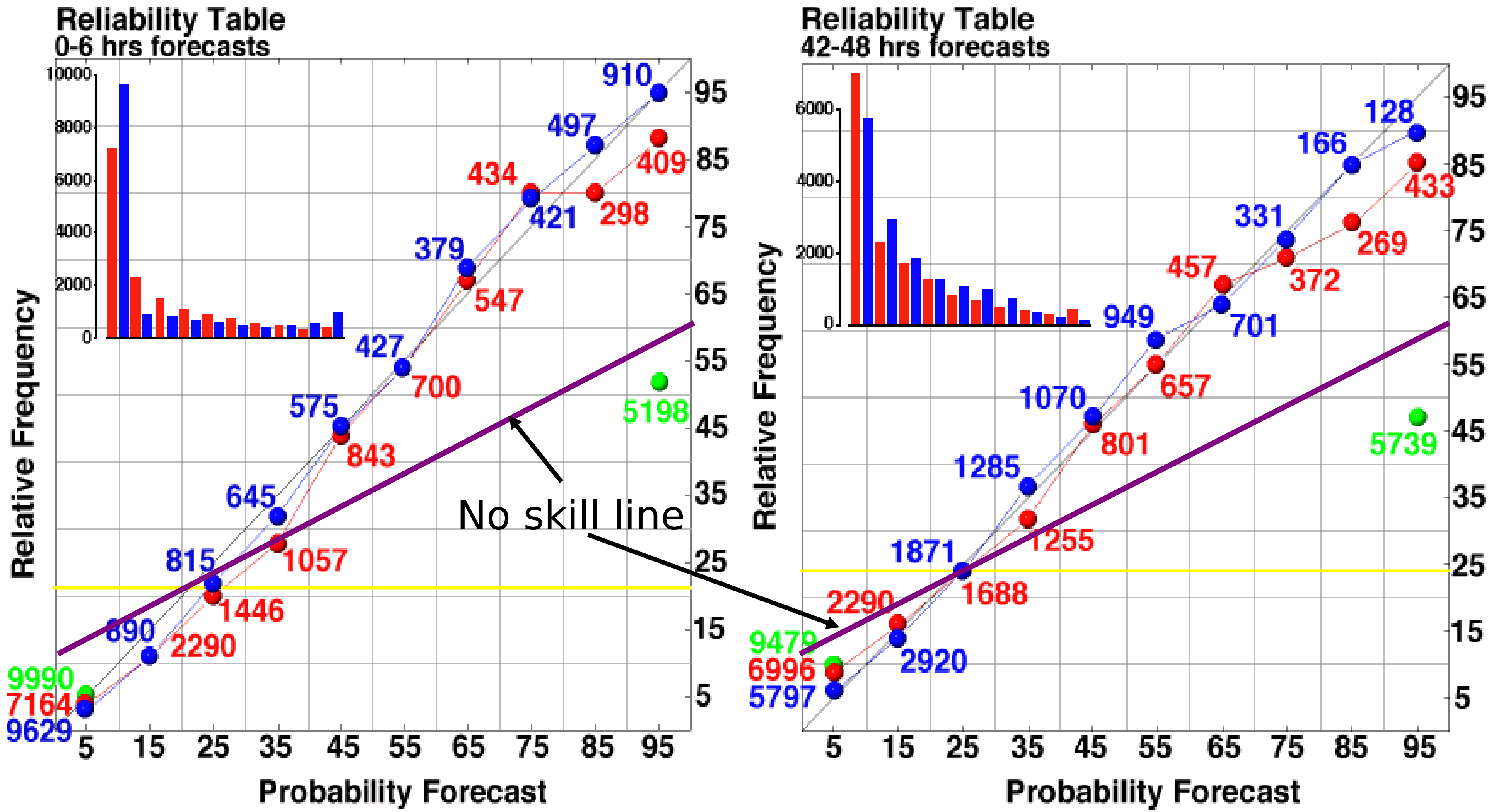
Reliability Diagram





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

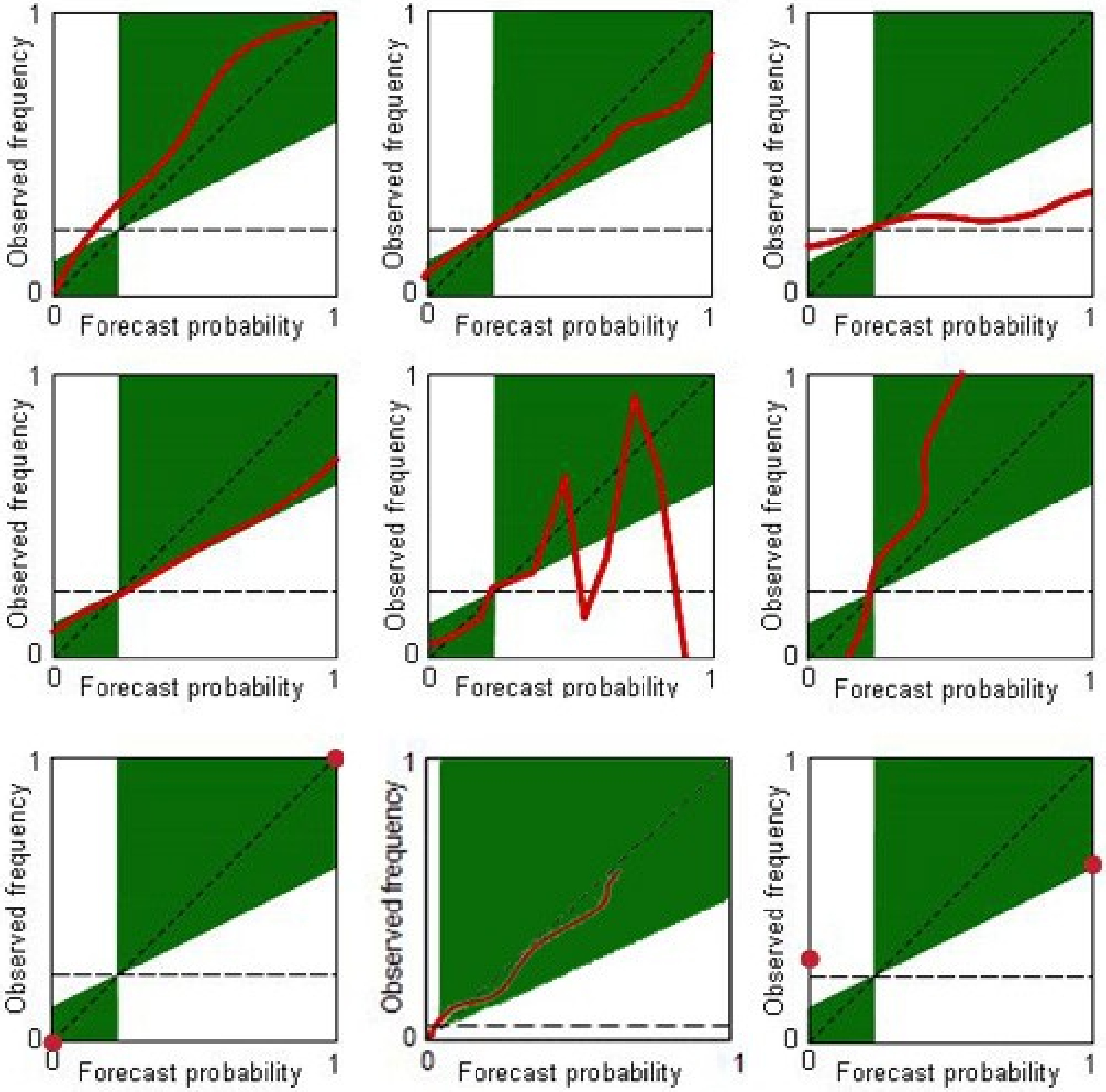


Reliability Table
0-6 hrs forecasts

Reliability Table
42-48 hrs forecasts

No skill line

Reliability Diagram Exercise



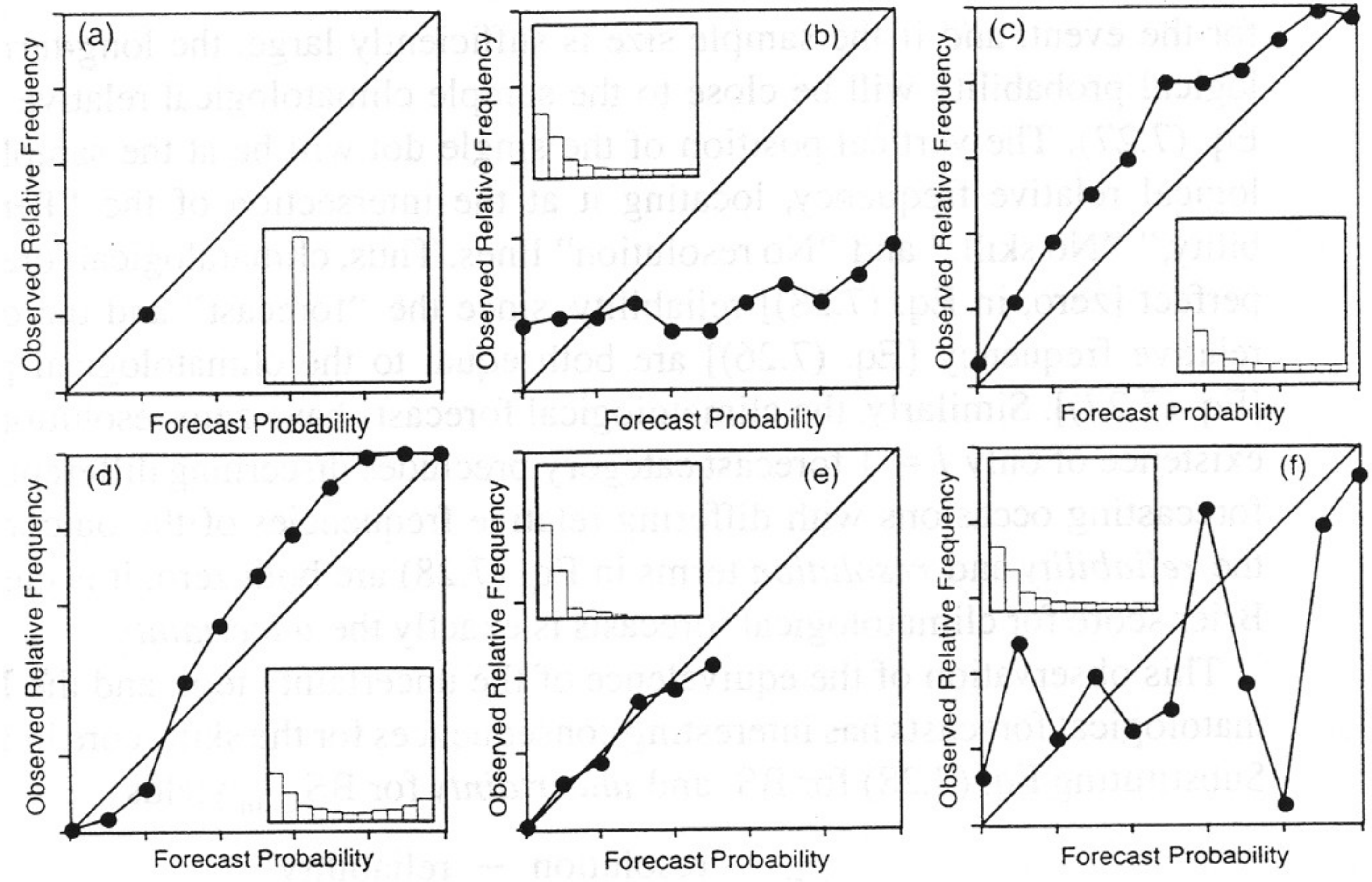
Reliability Diagram

Wilks (1995)

climatological forecast

minimal resolution

underforecasting bias



Good resolution at the expense of reliability

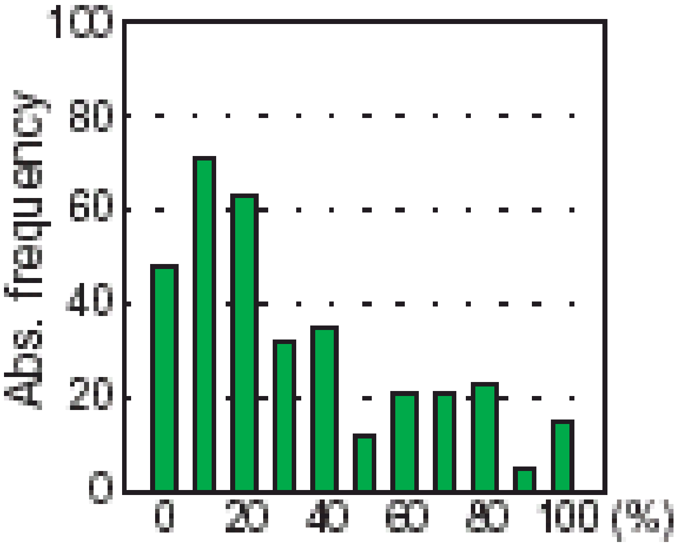
reliable rare event

small sample size

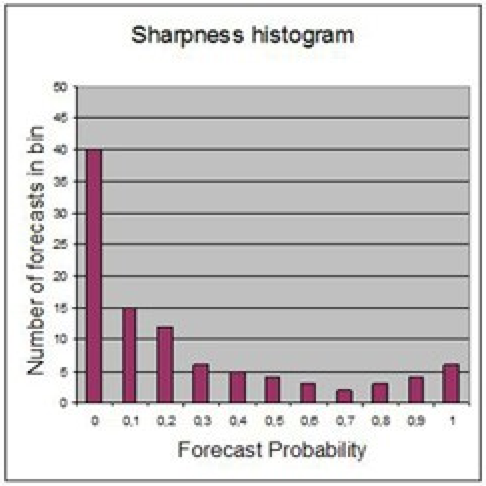
Sharpness

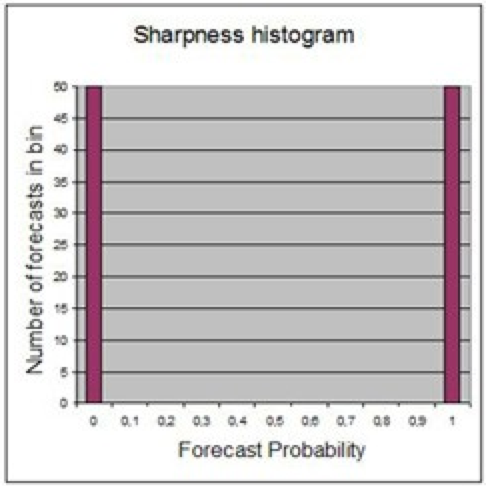
Refers to the spread of the probability distributions.

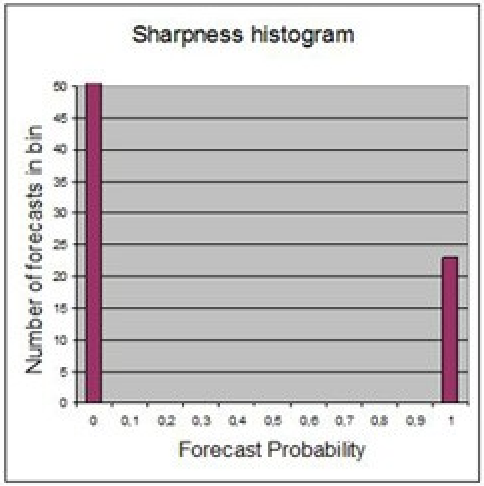
It is expressed as the capability of the system to forecast extreme values, or values close 0 or 1. The frequency of forecasts in each probability bin (shown in the histogram) shows the sharpness of the forecast.

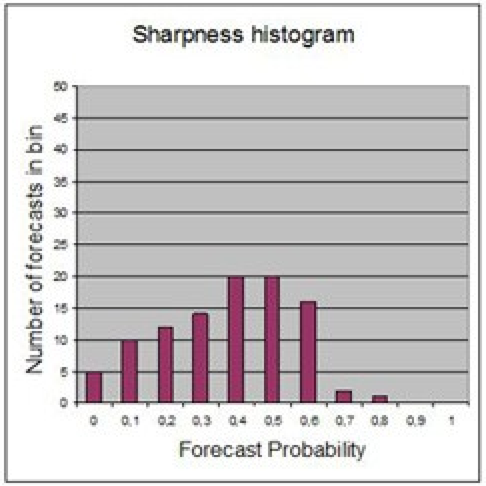


Sharpness Histogram Exercise









Reliability Diagrams - Summary

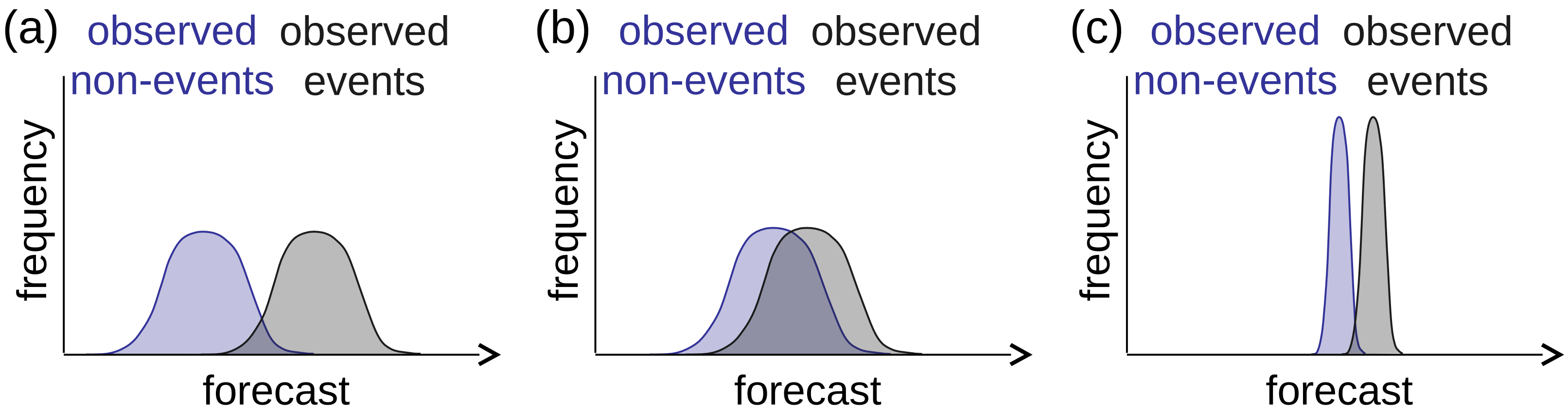
- Diagnostic tool
- Measures “reliability”, “resolution” and “sharpness”
- Requires “reasonably” large dataset to get useful results
- Try to ensure enough cases in each bin
- Graphical representation of Brier score components
- The reliability diagram is conditioned on the forecasts (i.e., given that X was predicted, what was the outcome?), and can be expected to give information on the real meaning of the forecast. It is a good partner to the ROC, which is conditioned on the observations.

Discrimination and the ROC

- Reliability diagram – partitioning the data according to the forecast probability
- Suppose we partition according to observation – 2 categories, yes or no
- Look at distribution of forecasts separately for these two categories

Discrimination

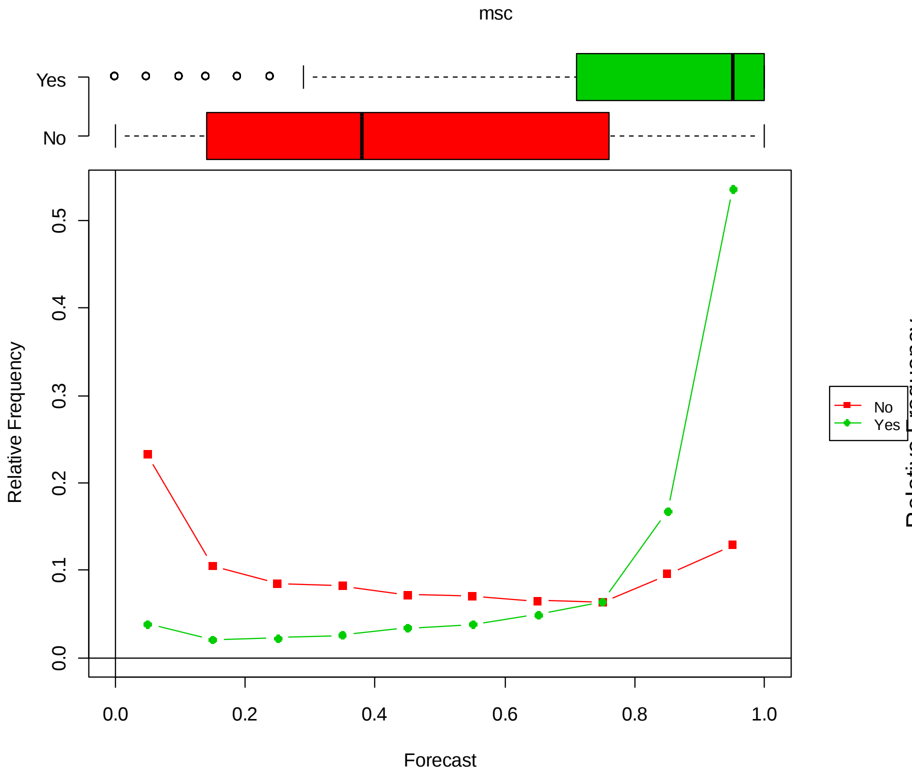
- *Discrimination*: The ability of the forecast system to clearly distinguish situations leading to the occurrence of an event of interest from those leading to the non-occurrence of the event.
- Depends on:
  - Separation of means of conditional distributions
  - Variance within conditional distributions

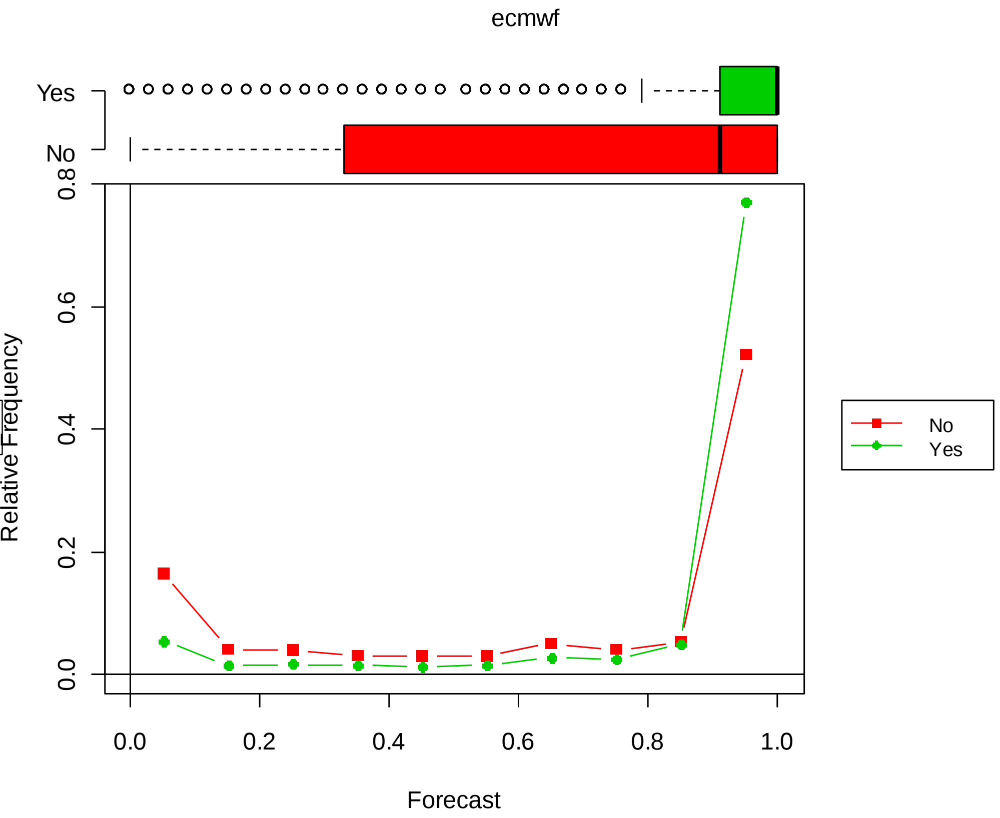


Good discrimination Poor discrimination Good discrimination

Sample Likelihood Diagrams: All precipitation, 20 Cdn stns, one year.

Discrimination: The ability of the forecast system to clearly distinguish situations leading to the occurrence of an event of interest from those leading to the non-occurrence of the event.

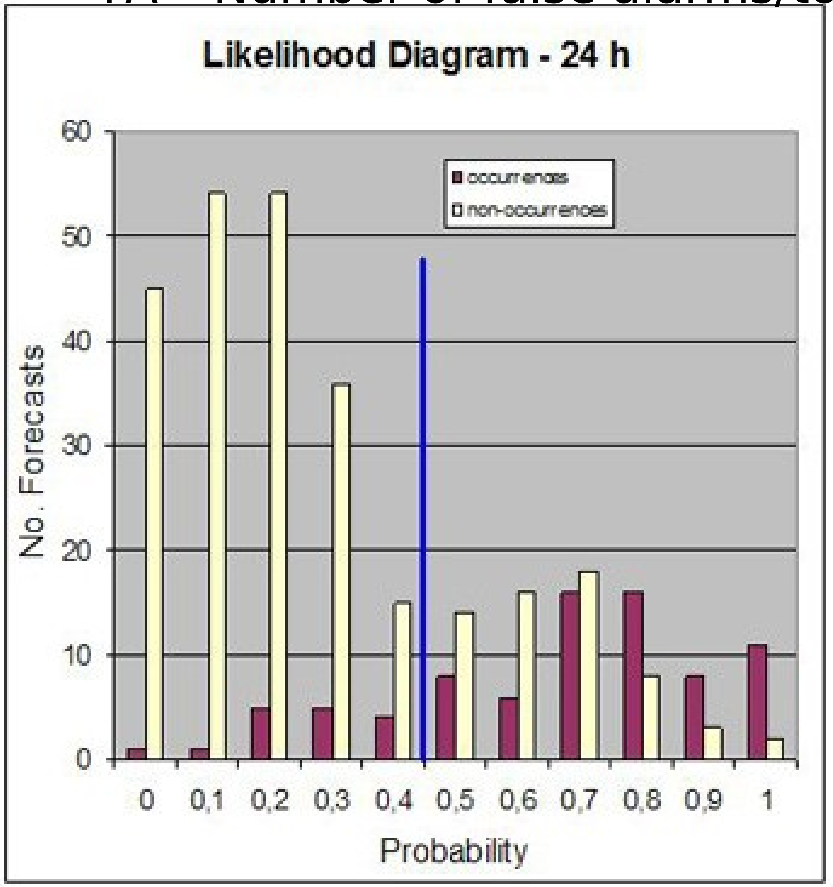


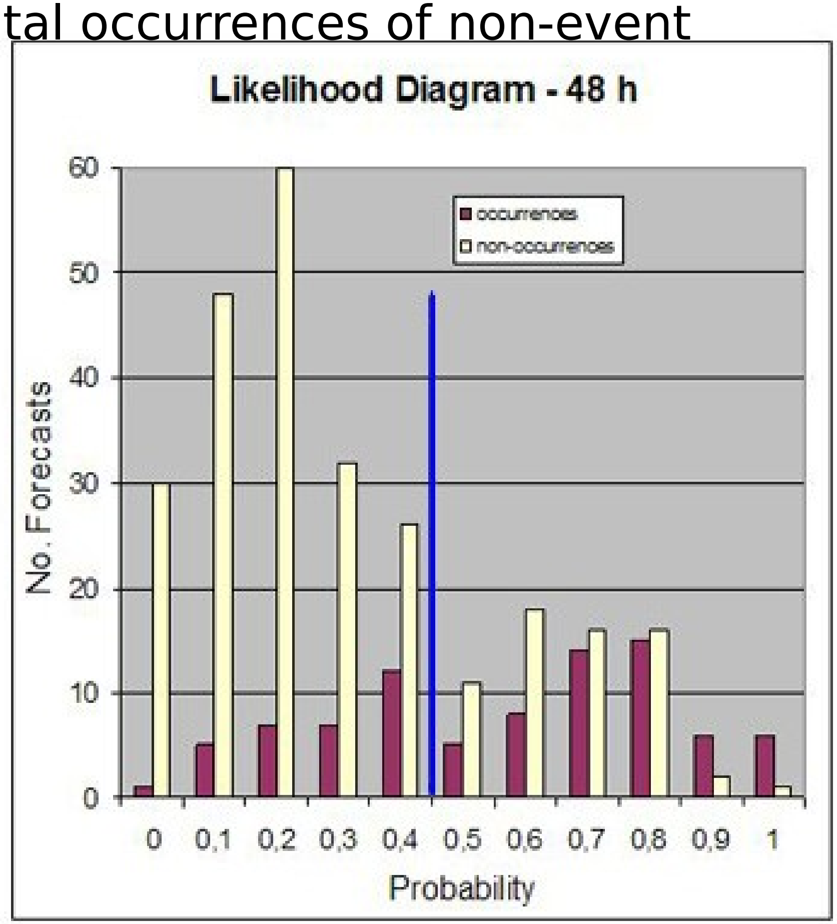


Relative Operating Characteristic curve: Construction

HR – Number of correct fcsts of event/total occurrences of event

FA – Number of false alarms/total occurrences of non-event





ROC Curves

(Relative Operating Characteristics, Mason and Graham 1999)

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

Hit Rate

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

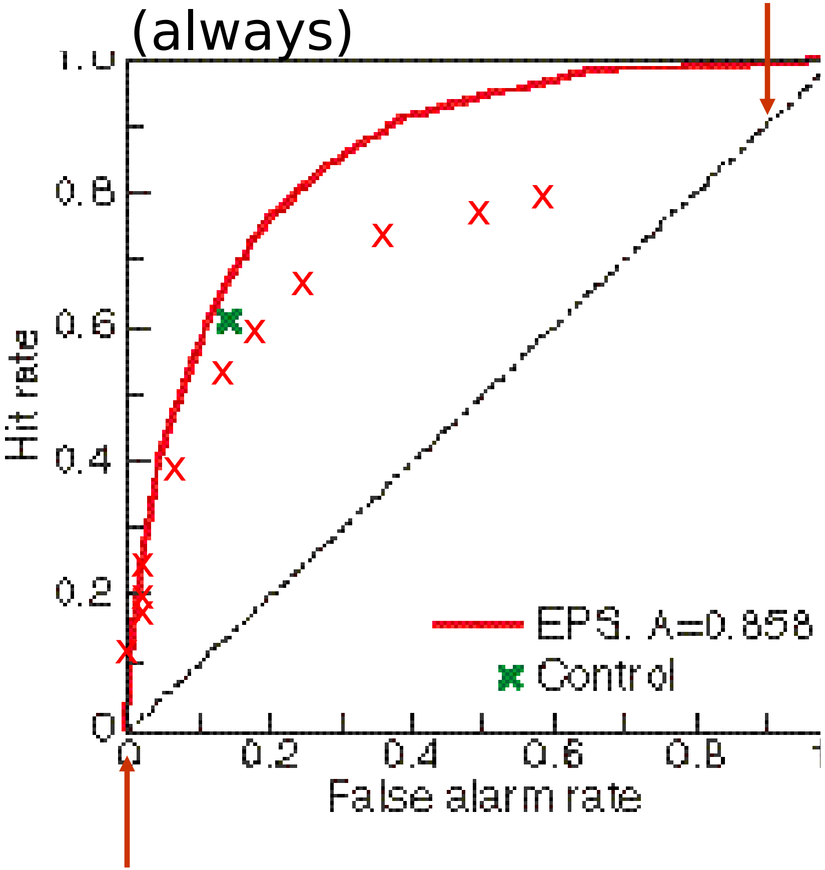
False Alarm Rate

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

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)

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

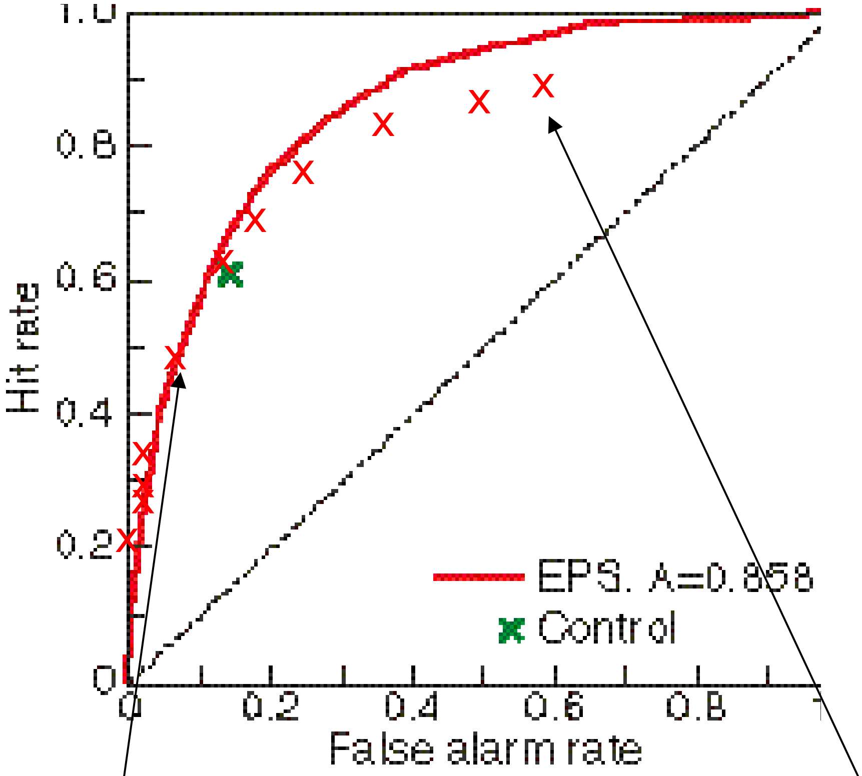
=> a warning can be issued when the forecast probability for the predefined event exceeds some threshold

For the k-th probability class:

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

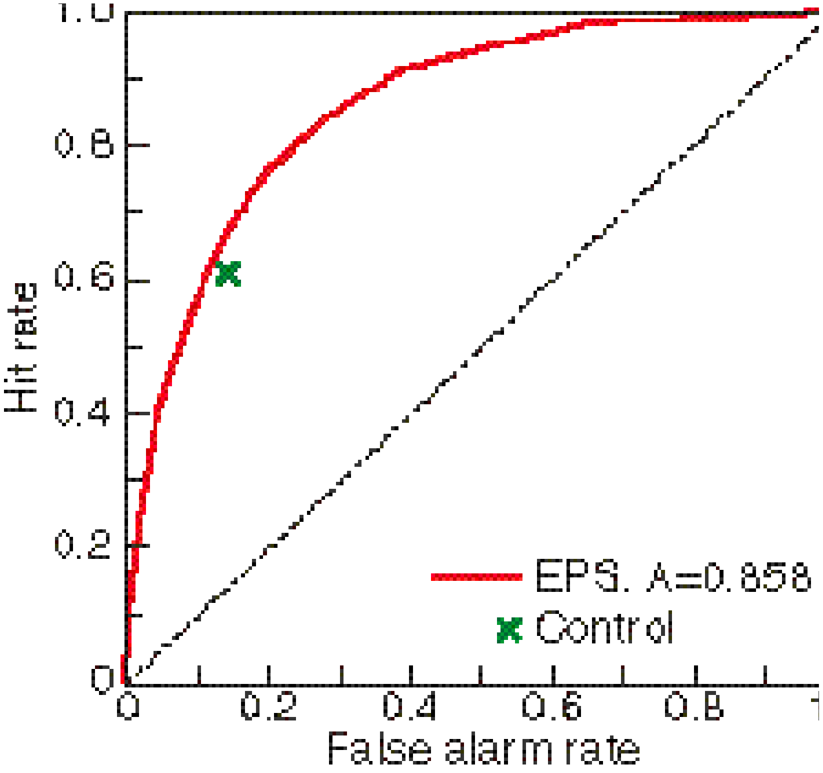
Hit rates are plotted against the corresponding false alarm rates to generate the ROC Curve

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 Area



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:

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

Construction of ROC curve

- From original dataset, determine bins
  - Can use binned data as for Reliability diagram BUT
  - There must be enough occurrences of the event to determine the conditional distribution given occurrences – may be difficult for rare events.
  - Generally need at least 5 bins.
- For each probability threshold, determine HR and FA
- Plot HR vs FA to give empirical ROC.
- Use binormal model to obtain ROC area; recommended whenever there is sufficient data >100 cases or so.
  - For small samples, recommended method is that described by Simon Mason. (See 2007 tutorial)

ROC - Interpretation

Interpretation of ROC:

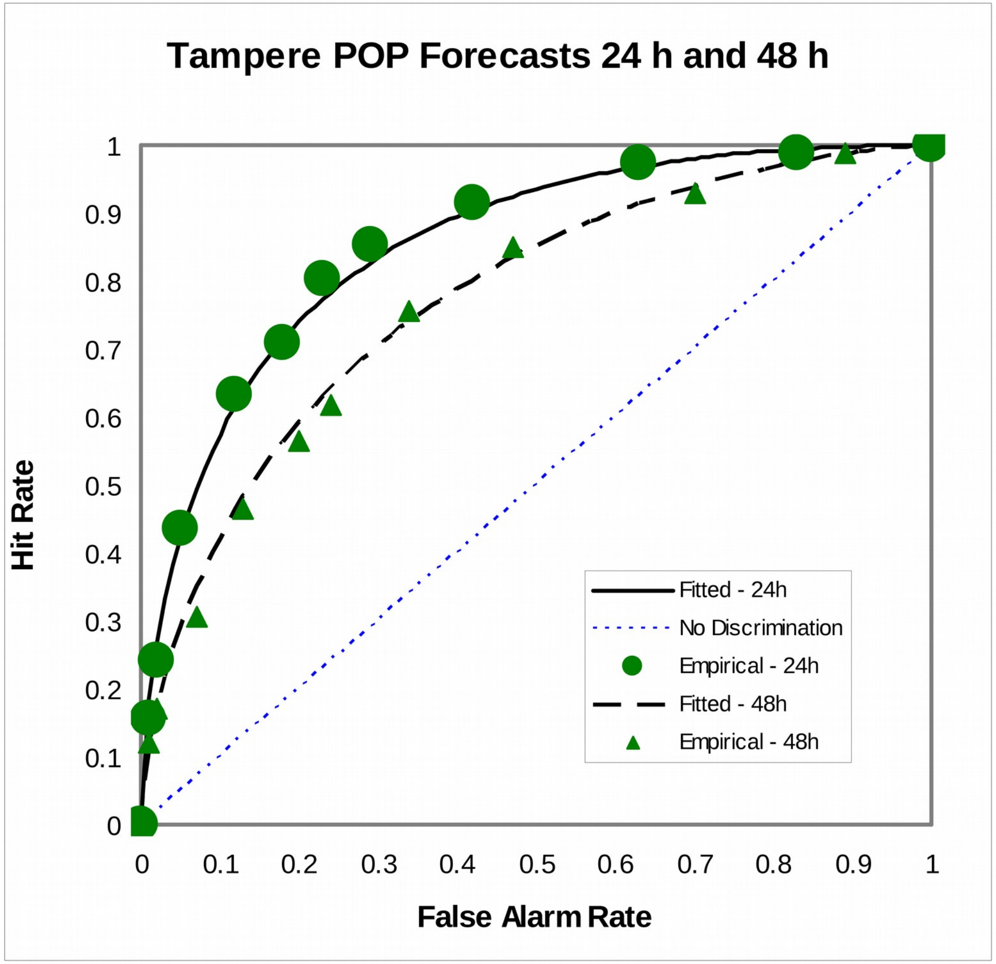
*Quantitative measure: Area under the curve – ROCA

*Positive if above 45 degree ‘No discrimination’ line where ROCA = 0.5

*Perfect is 1.0.

ROC is NOT sensitive to bias: It is necessarily only that the two conditional distributions are separate

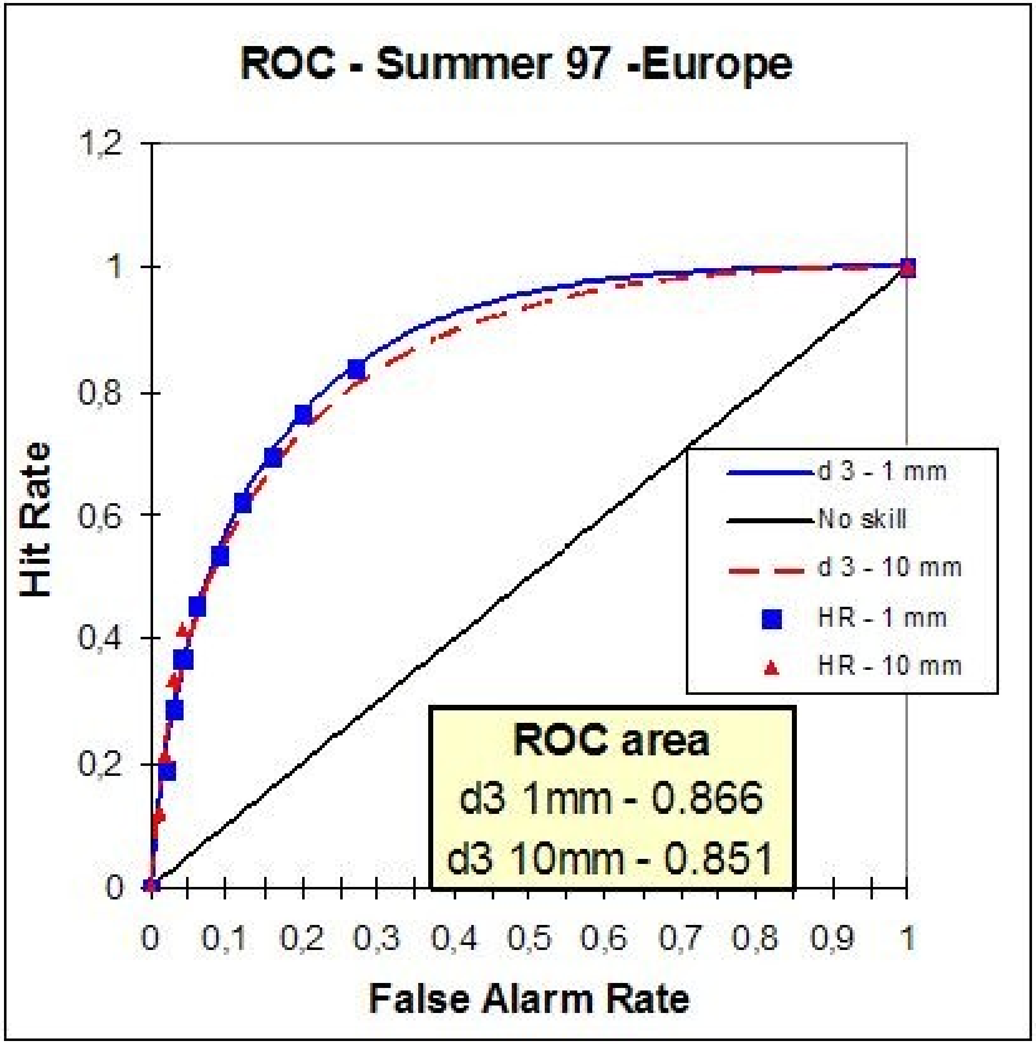
* Can compare with deterministic forecast – one point



ROC for infrequent events

For fixed binning (e.g. deciles), points cluster towards lower left corner for rare events: subdivide lowest probability bin if possible.

Remember that the ROC is insensitive to bias (calibration).



Summary - ROC

- Measures “discrimination”
- Plot of Hit rate vs false alarm rate
- Area under the curve – by fitted model
- Sensitive to sample climatology – careful about averaging over areas or time
- NOT sensitive to bias in probability forecasts – companion to reliability diagram
- Related to the assessment of “value” of forecasts
- Can compare directly the performance of probability and deterministic forecast

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 <u>over a great number of cases</u>.

❖ U can rely on a forecast system to know in advance if the event is going to occur or not.

Cost-loss Analysis

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

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

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

$\bar{o} = a + c$ 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.

$$\text{ME}_k\text{f} = \quad F_k\frac{C}{L}(1-\bar{o}) - H_k\bar{o}\left(1-\frac{C}{L}\right) + \bar{o}$$

Mean expense

Cost-loss Analysis

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

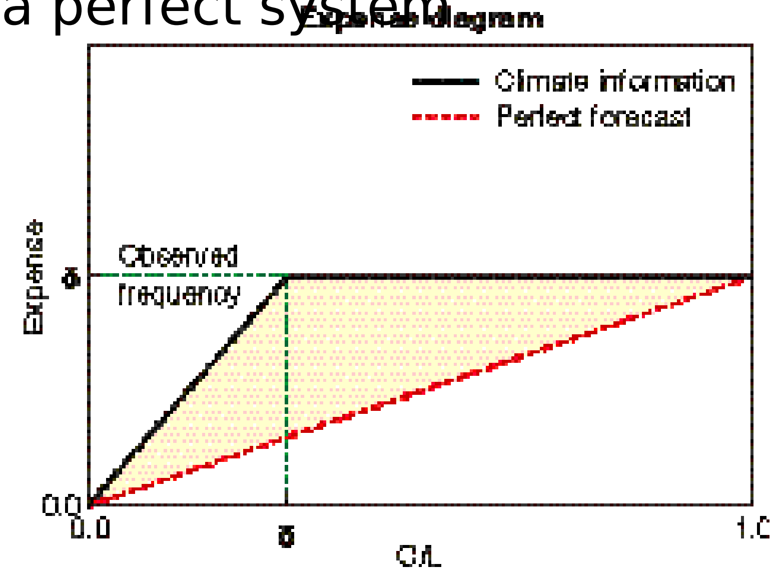
Gain obtained using the system instead of the climatological information, percentage with respect to the gain obtained using a perfect system.

Value

ME with a perfect forecast system

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

the preventive action is taken only when the event occurs



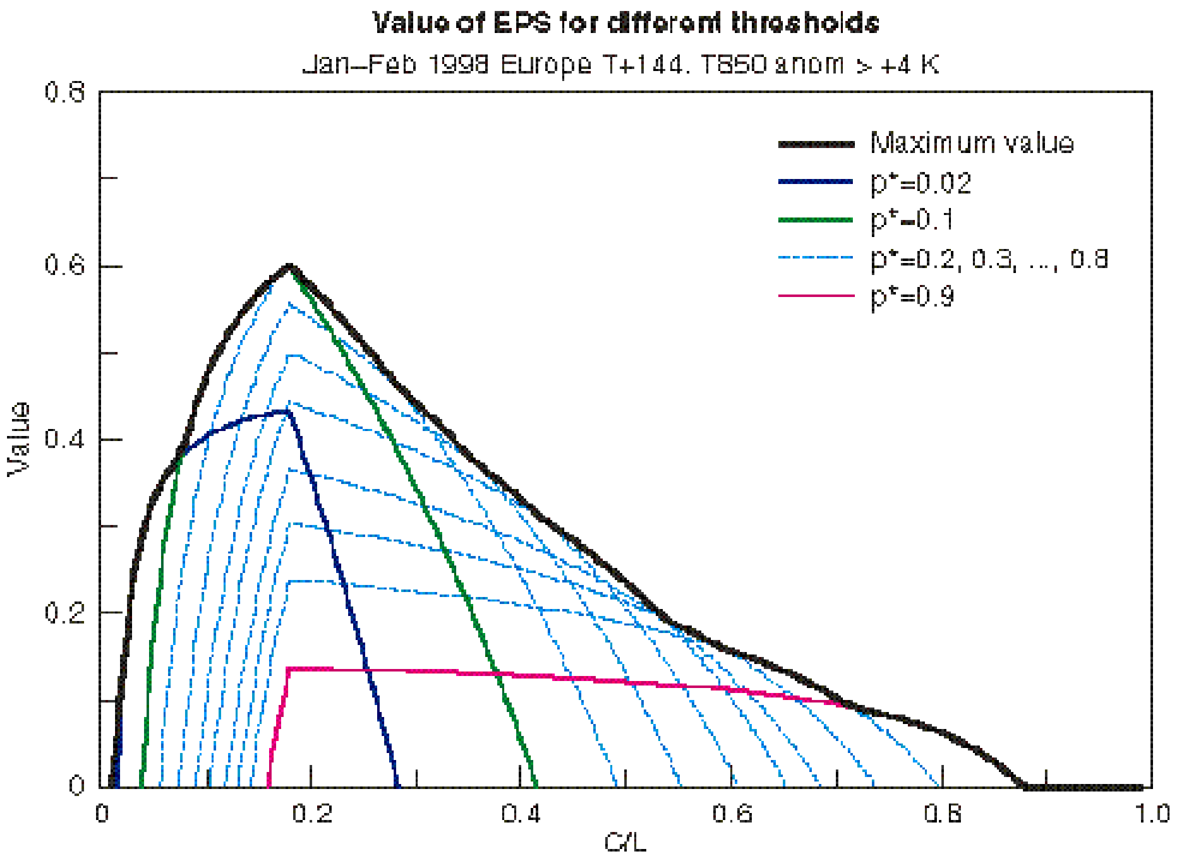
ME based on climatological information

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

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

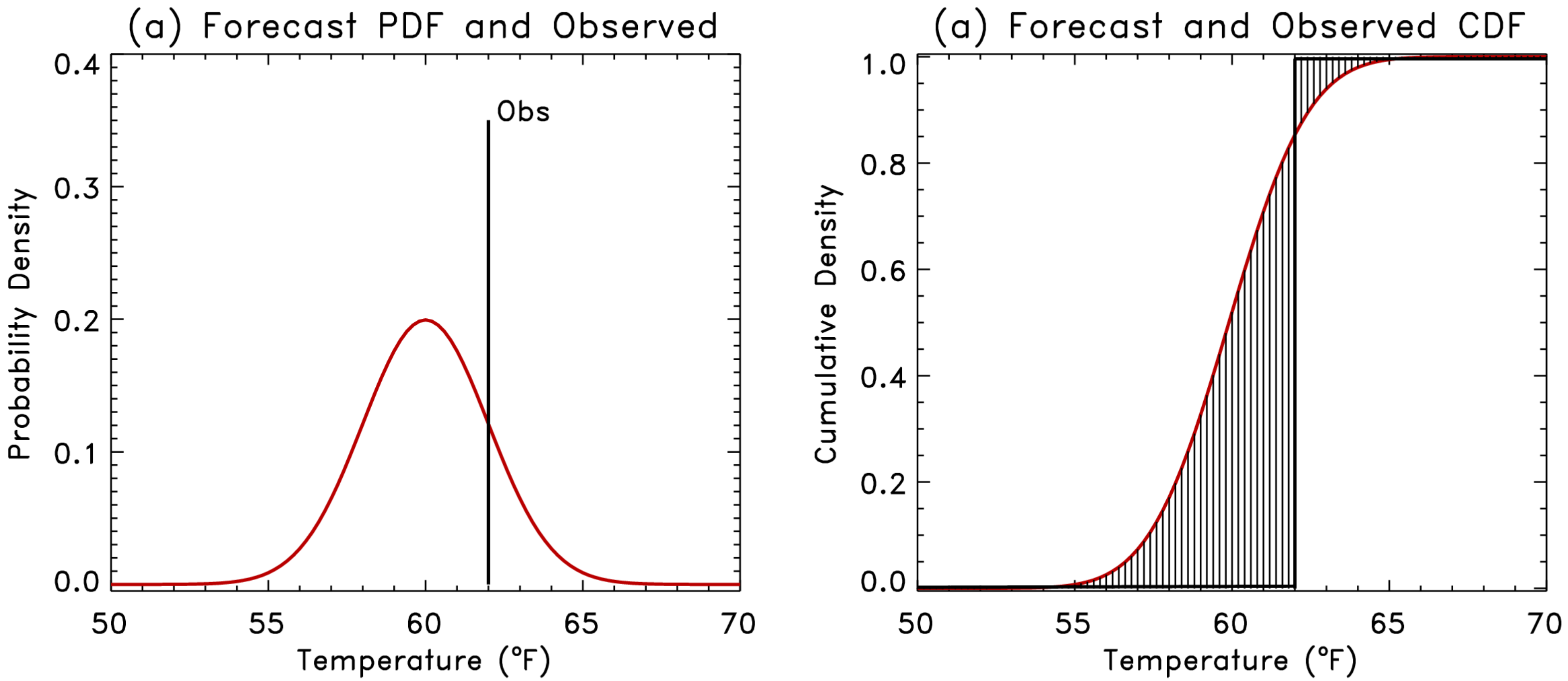
it is never taken otherwise

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.

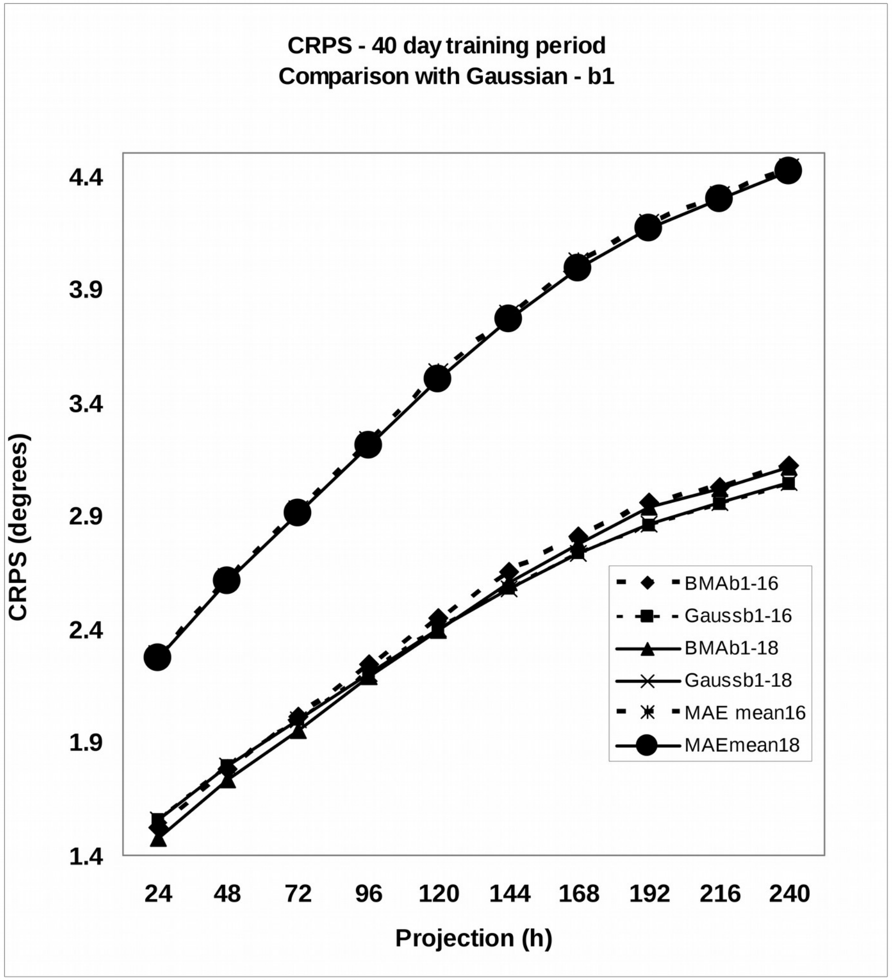
CRPS

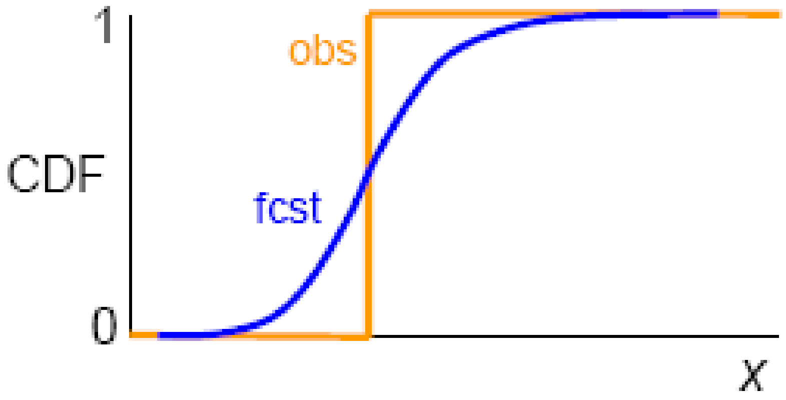


(a) Forecast PDF and Observed
Probability Density
0.4
0.3
0.2
0.1
0.0
Obs
50
55
60
65
70
Temperature (°F)
(a) Forecast and Observed CDF
Cumulative Density
1.0
0.8
0.6
0.4
0.2
0.0
50
55
60
65
70
Temperature (°F)

Continuous Rank Probability Score

$$CRPS(P, x_a) = \int_{\infty}^{\infty} \left[P(x) - P_a(x) \right]^2 dx$$





-difference between observation and forecast, expressed as cdfs

-defaults to MAE for deterministic fcst

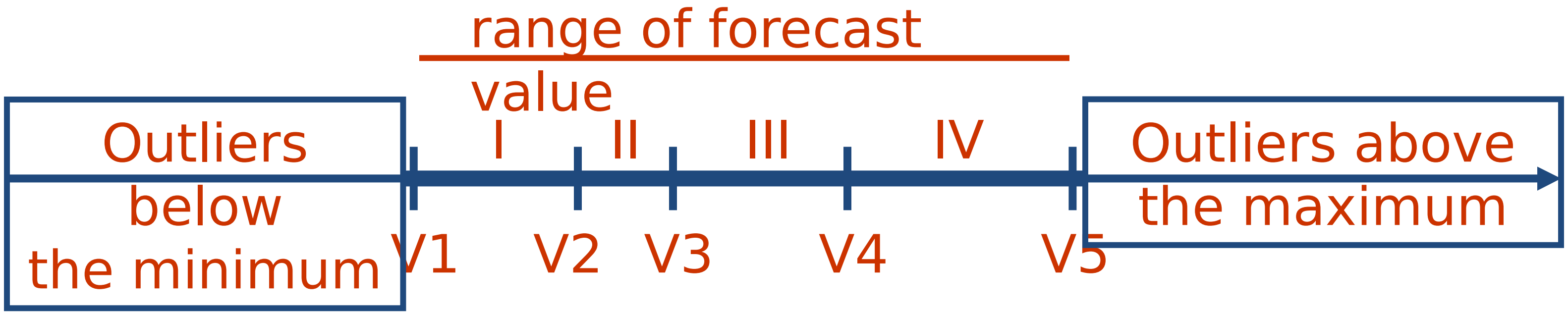
-flexible, can accommodate uncertain obs

Rank Histogram

- Commonly used to diagnose the average spread of an ensemble compared to observations
- Computation: Identify rank of the observation compared to ranked ensemble forecasts
- Assumption: observation equally likely to occur in each of n+1 bins. (questionable?)

Rank histogram (Talagrand Diagram)

Rank histogram of the distribution of the values forecast by an ensemble



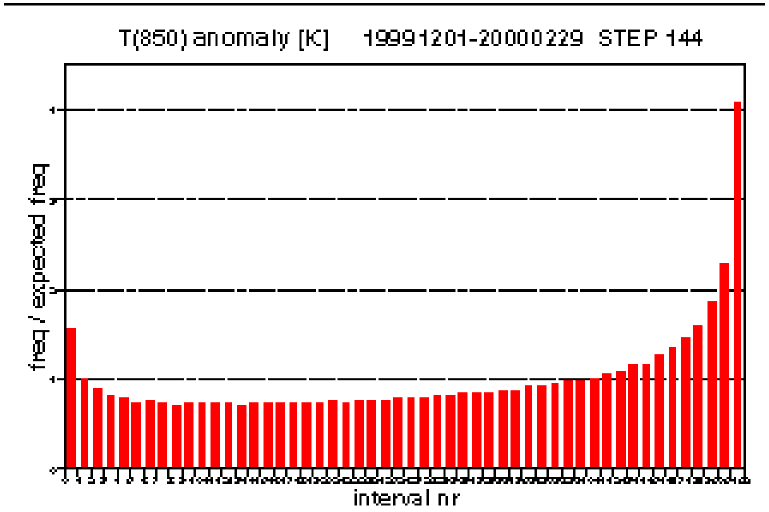
range of forecast value

Outliers below the minimum

I II III IV

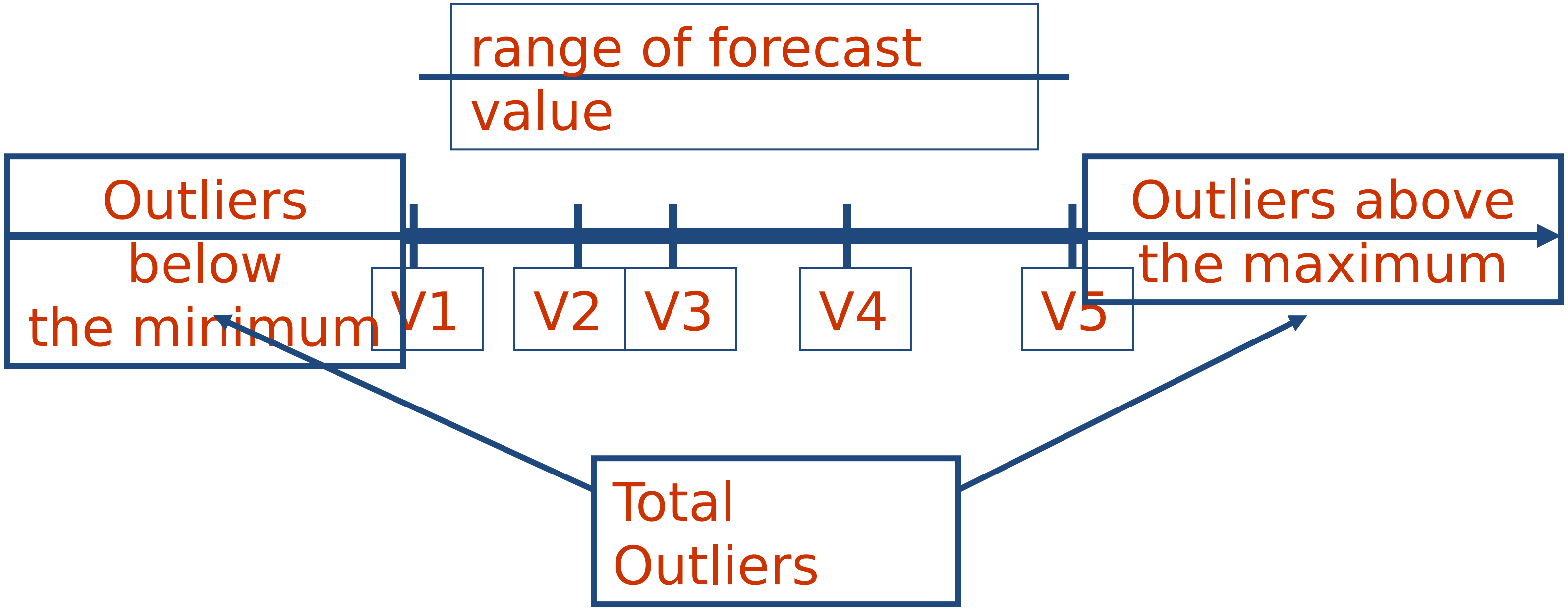
V1 V2 V3 V4 V5

Outliers above the maximum

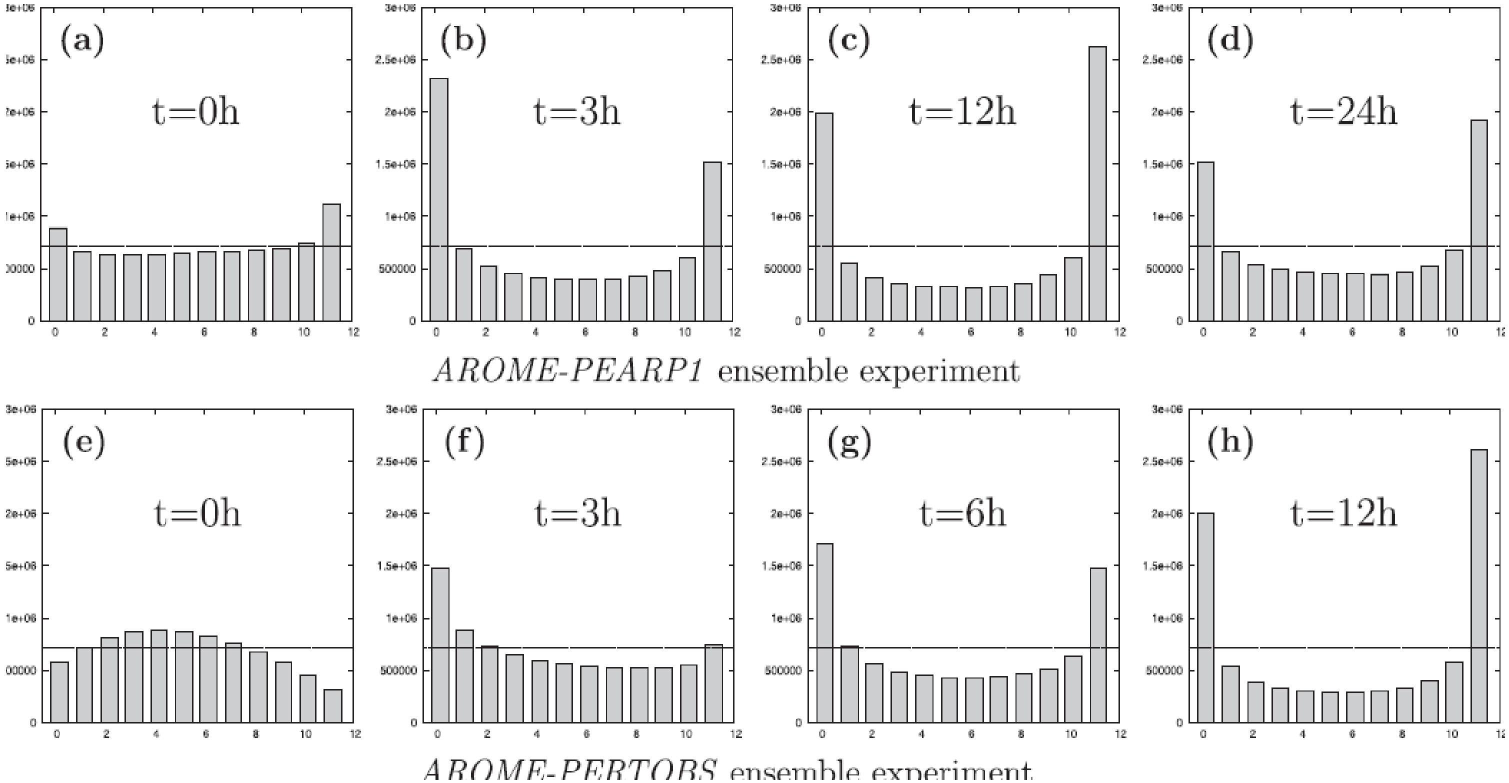


Percentage of Outliers

Percentage of points where the observed value lies out of the range of forecast values.



Rank histogram - exercise

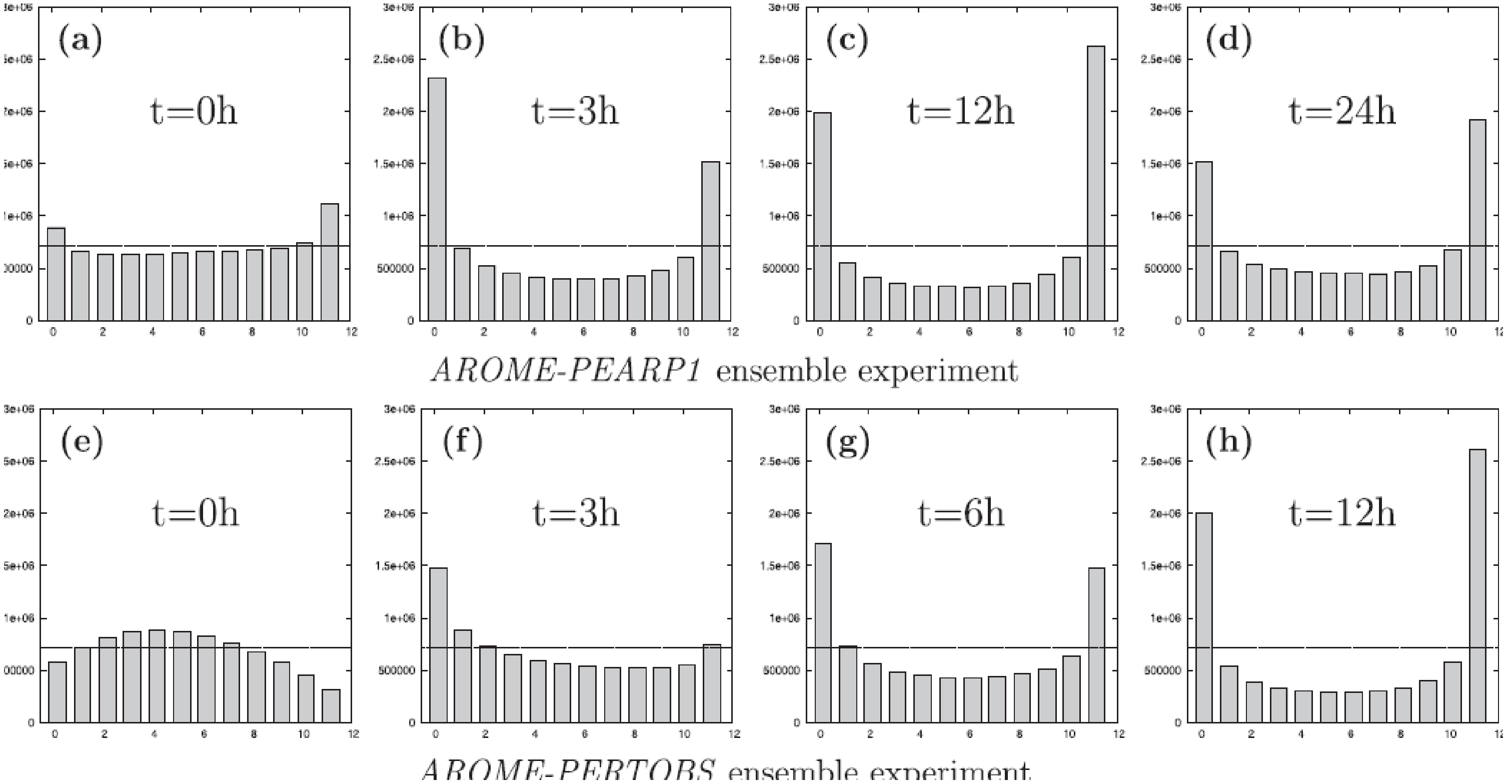


AROME-PEARP1 ensemble experiment

AROME-PERTOBS ensemble experiment

Uncertainty in LAM
Vié et al., 2011

- The uncertainty on convective scale ICs has a stronger impact over the first hours (12 h) of simulation, before the LBCs overwhelm differences in initial states. The uncertainties on LBCs have a growing impact at a longer range (beyond 12 h).



AROME-PEARP1 ensemble experiment

AROME-PERTOBS ensemble experiment

boundary condition perturbation only

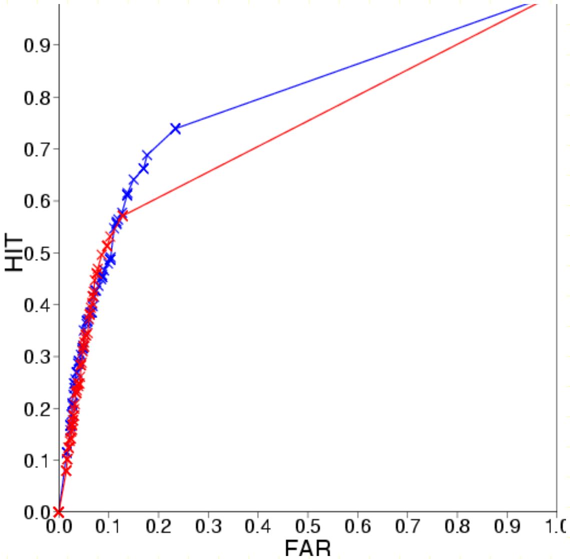
initial condition perturbation

Data considerations for ensemble verification

- An extra dimension – many forecast values, one observation value
  - Suggests data matrix format needed; columns for the ensemble members and the observation, rows for each event
- Raw ensemble forecasts are a collection of deterministic forecasts
- The use of ensembles to generate probability forecasts requires interpretation.
  - i.e. processing of the raw ensemble data matrix.

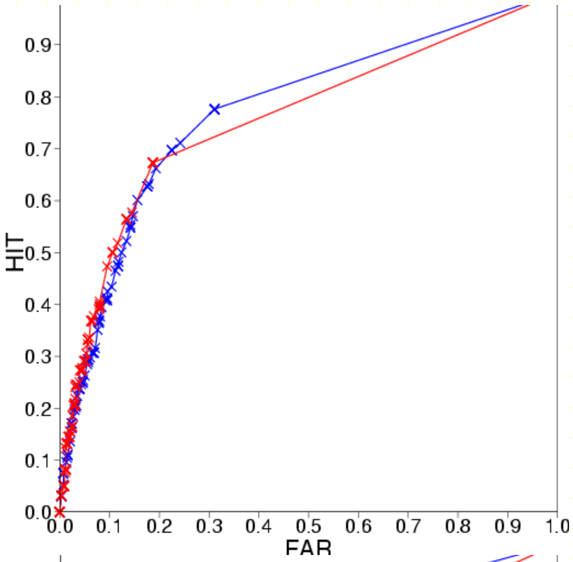
average -10mm/24h

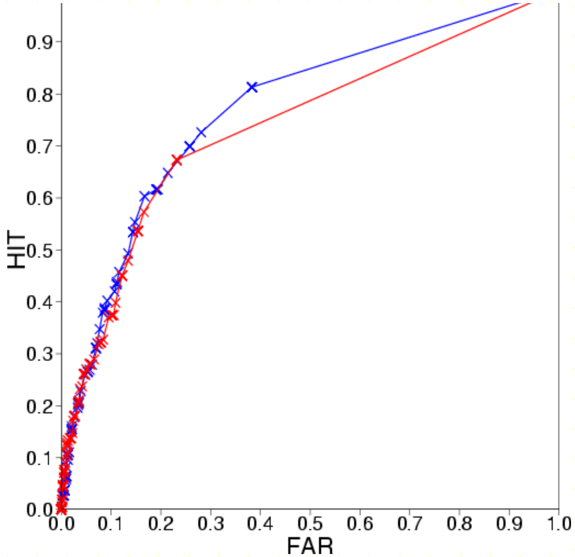
noss=234



+42

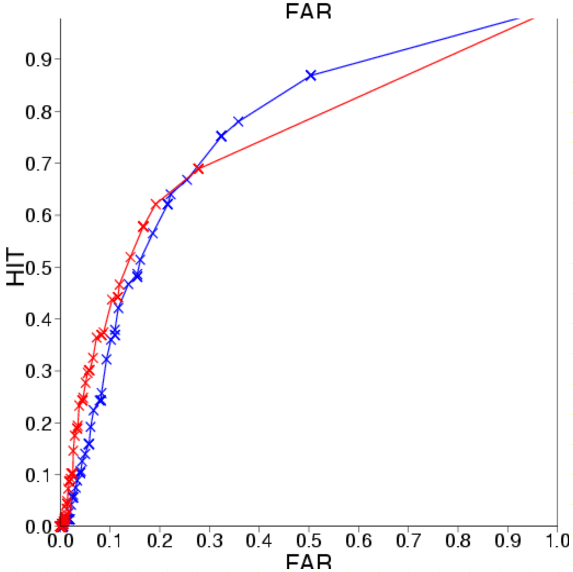
+66





+90

+114



COSMO-LEPS —

16-MEMBER EPS —

COSMO-LEPS vs ECMWF 5 RM

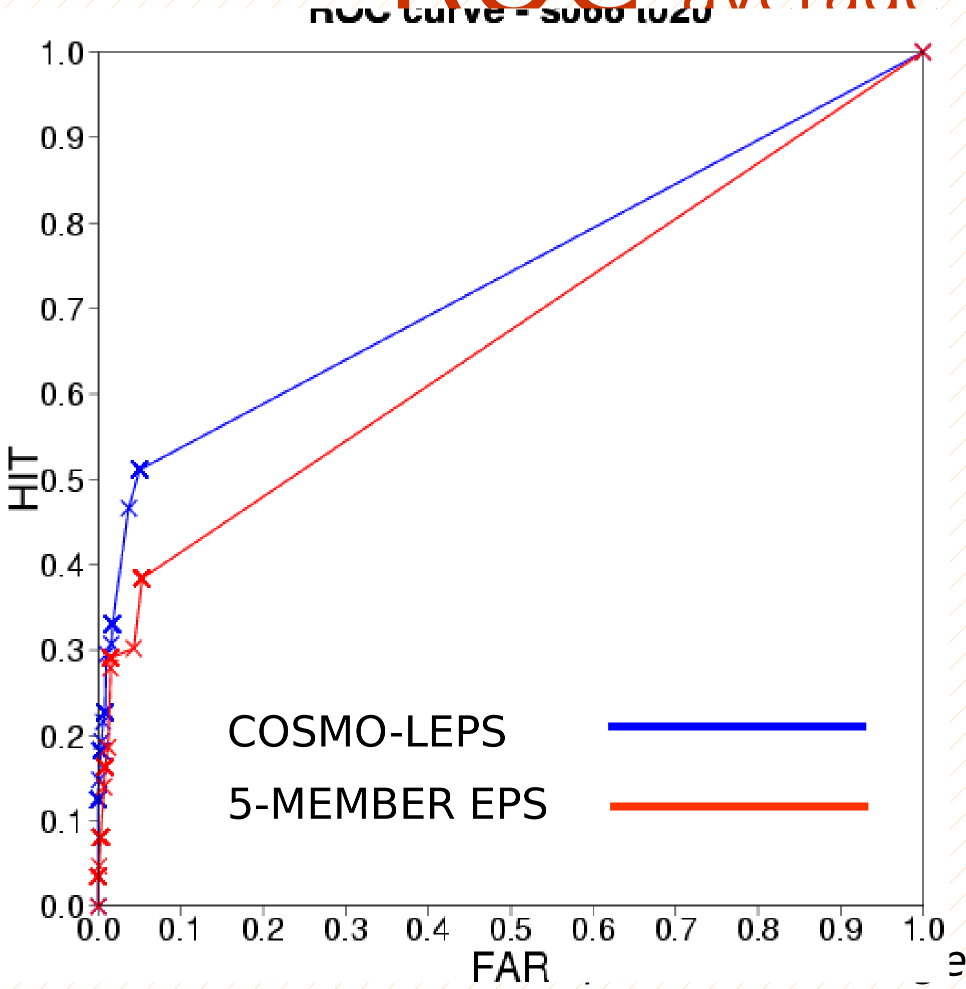
ROC

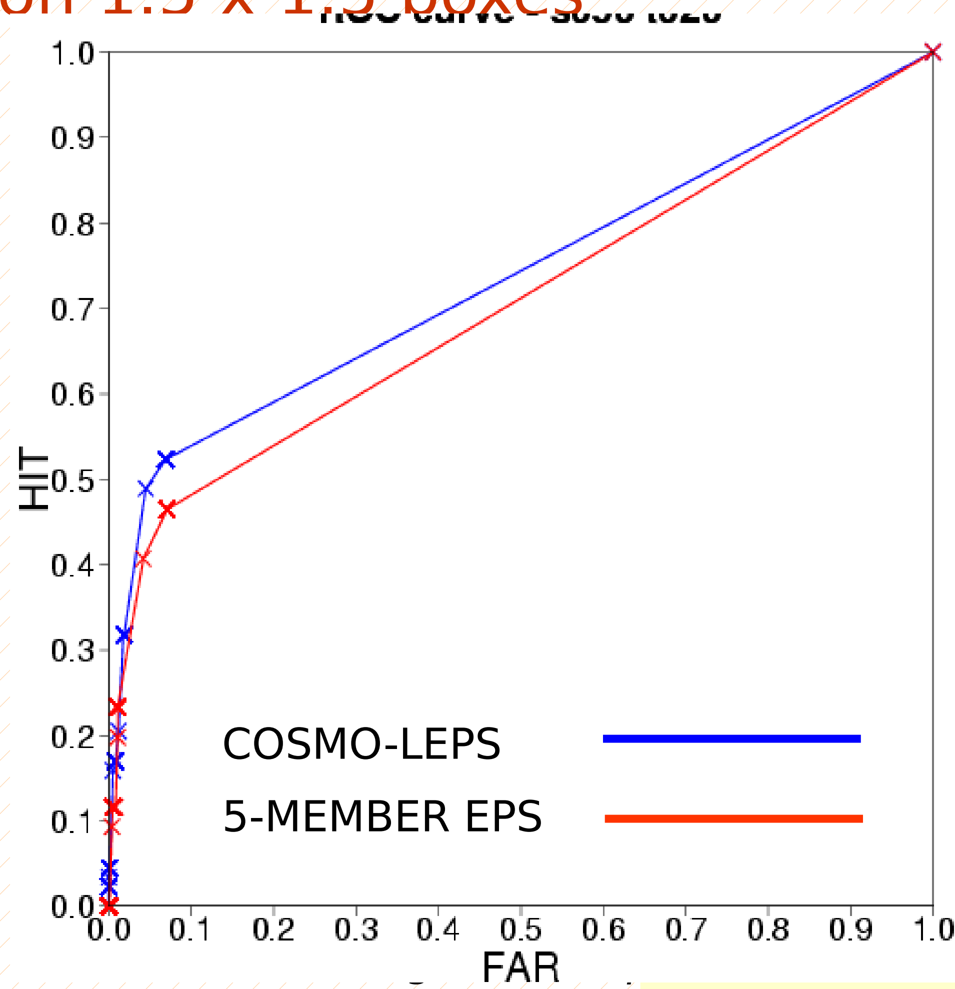
fc. range +66

tp > 20mm/24h

average on 1.5 x 1.5 boxes

fc. range +90





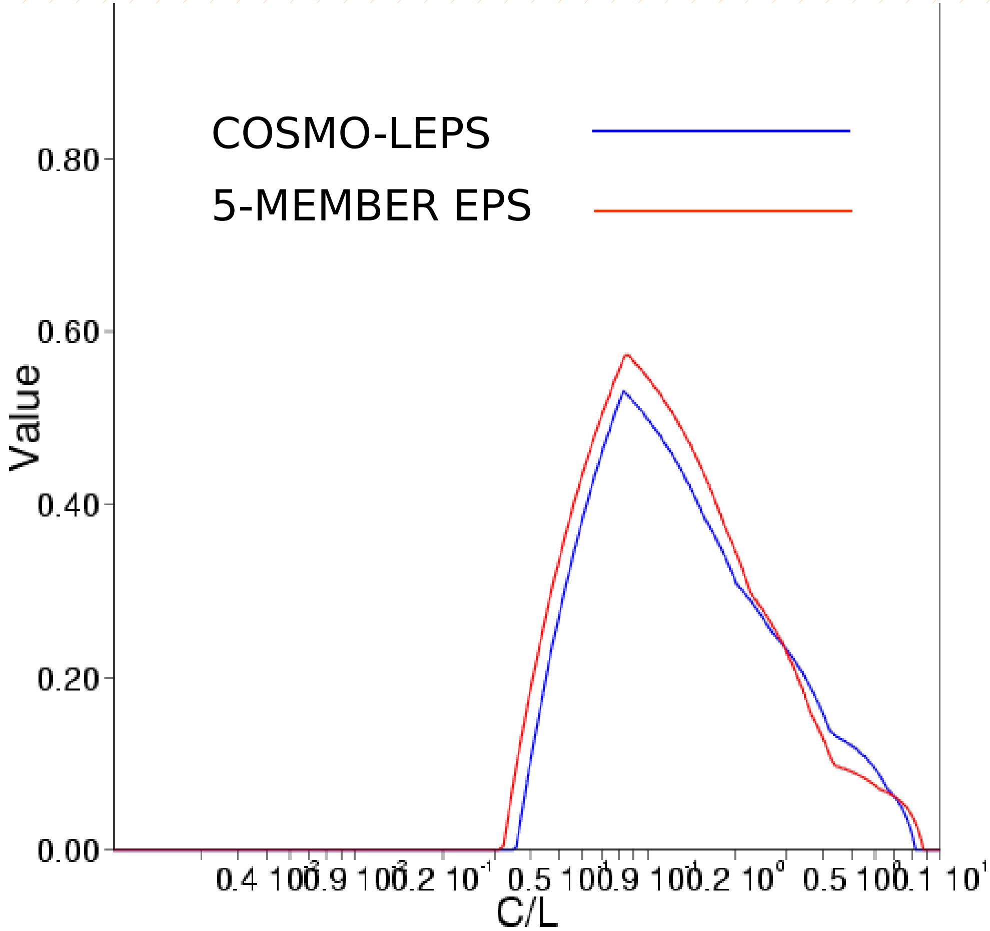
COSMO-LEPS vs ECMWF 5 RM

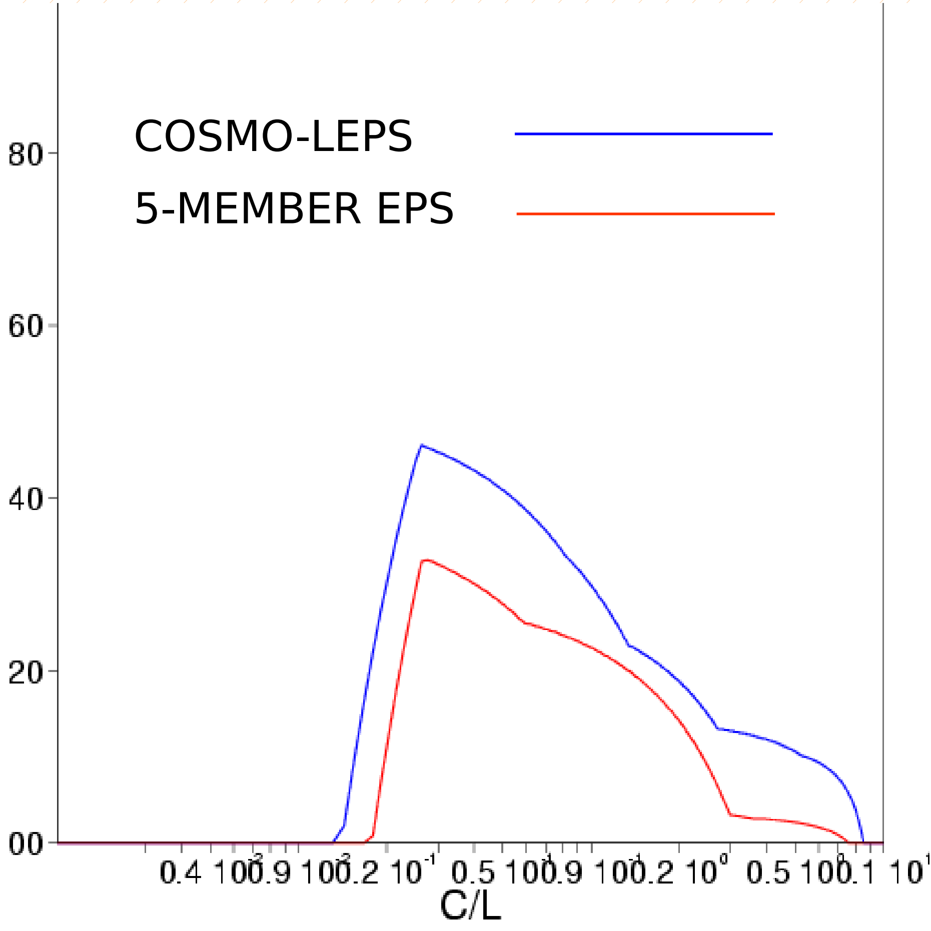
COST-LOSS (envelope) average on 1.5 x 1.5 boxes

fc. range +66

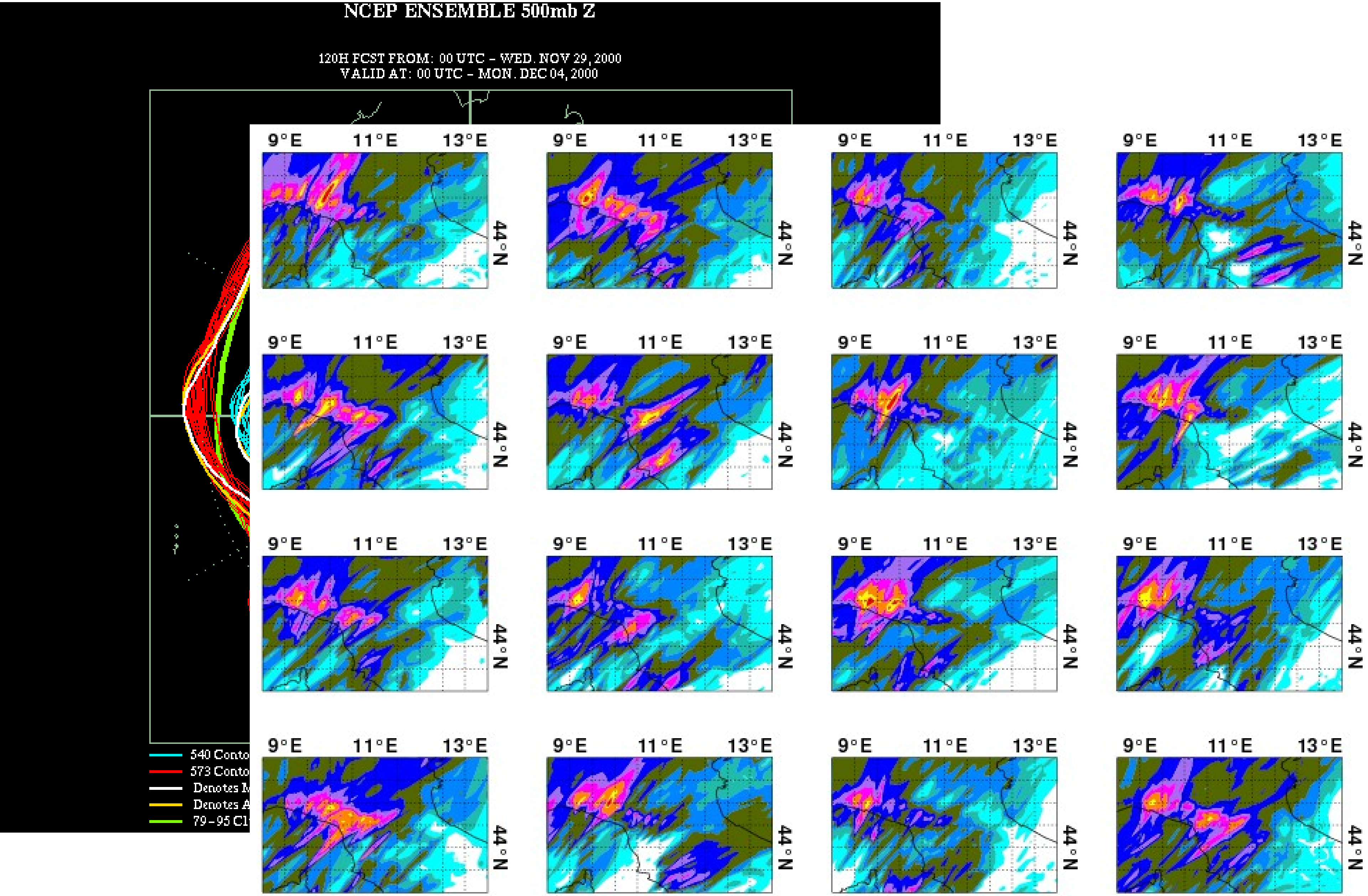
tp > 10mm/24h

tp > 20mm/24h

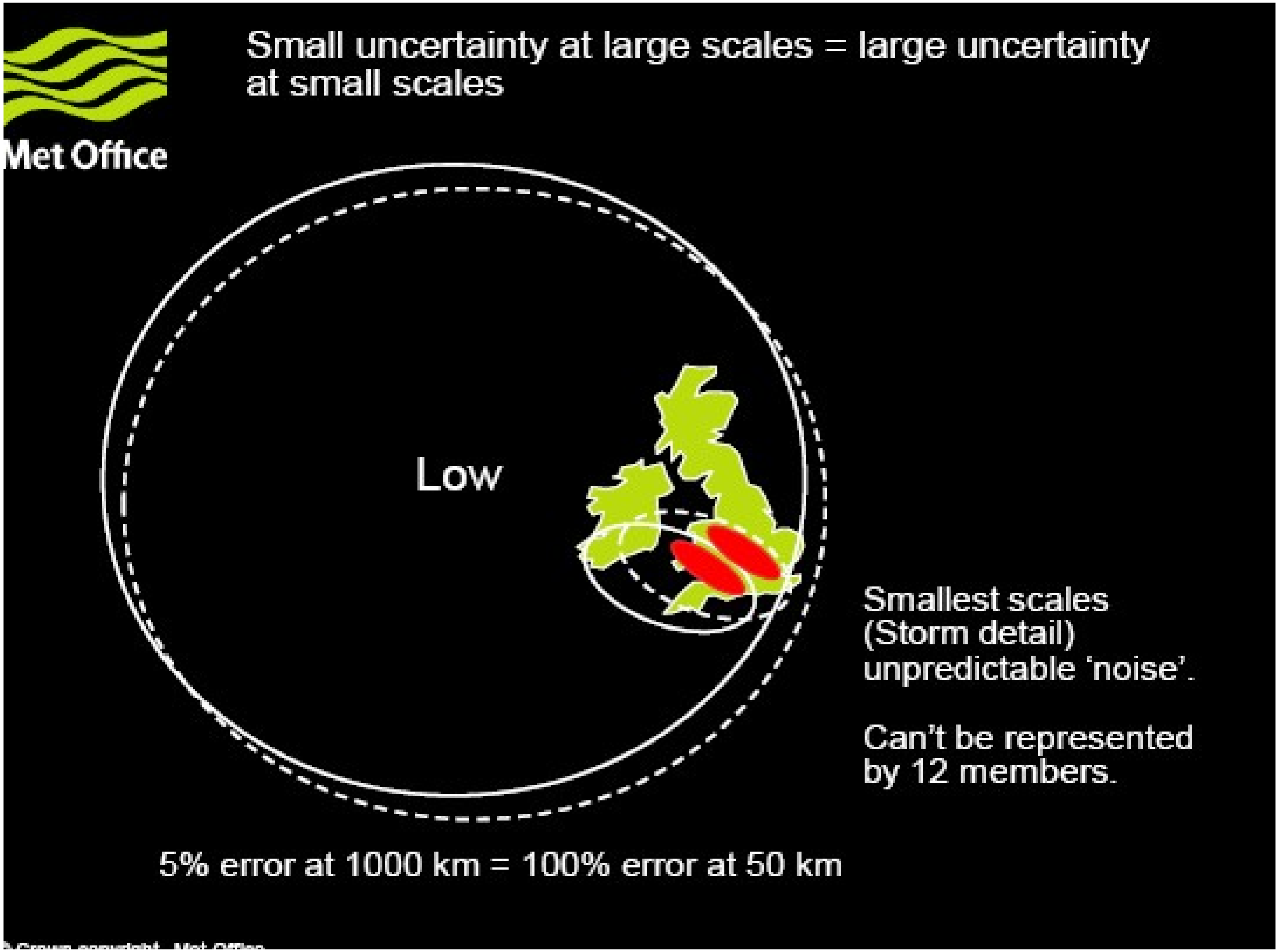




Spatial scales



Mesoscale uncertainty



Small uncertainty at large scales = large uncertainty at small scales

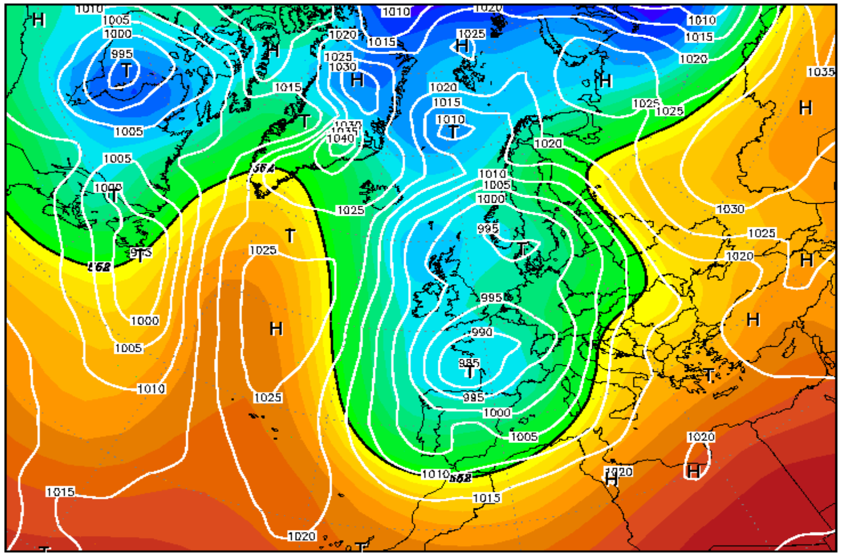
Smallest scales (Storm detail) unpredictable ‘noise’.

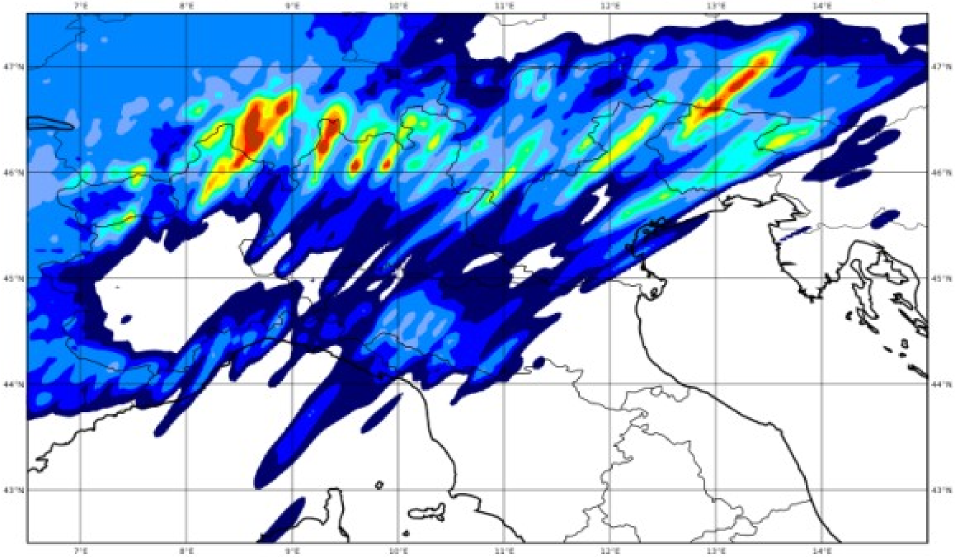
Can’t be represented by 12 members.

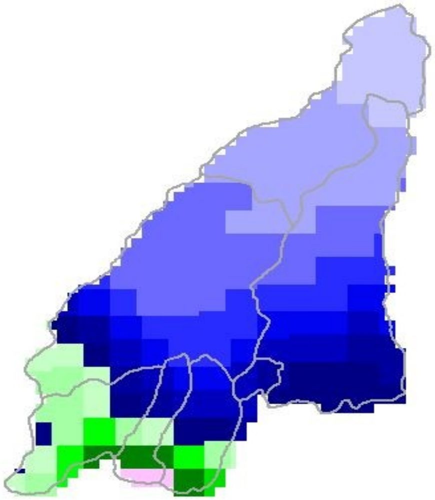
5% error at 1000 km = 100% error at 50 km

Predictability: a fractal problem

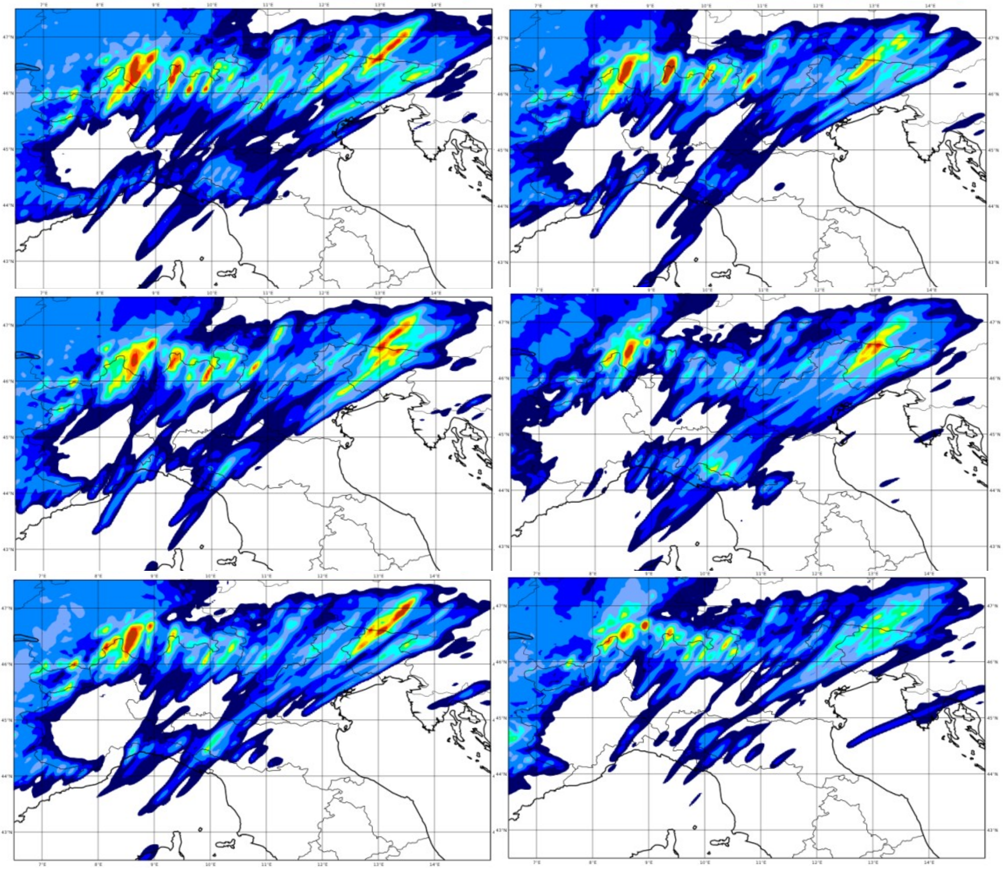
Predictability: a fractal problem

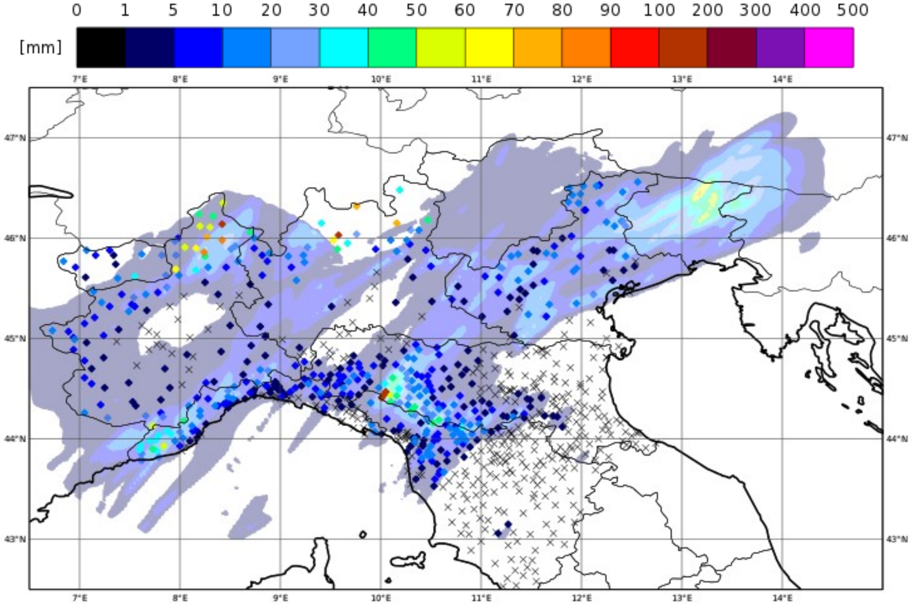




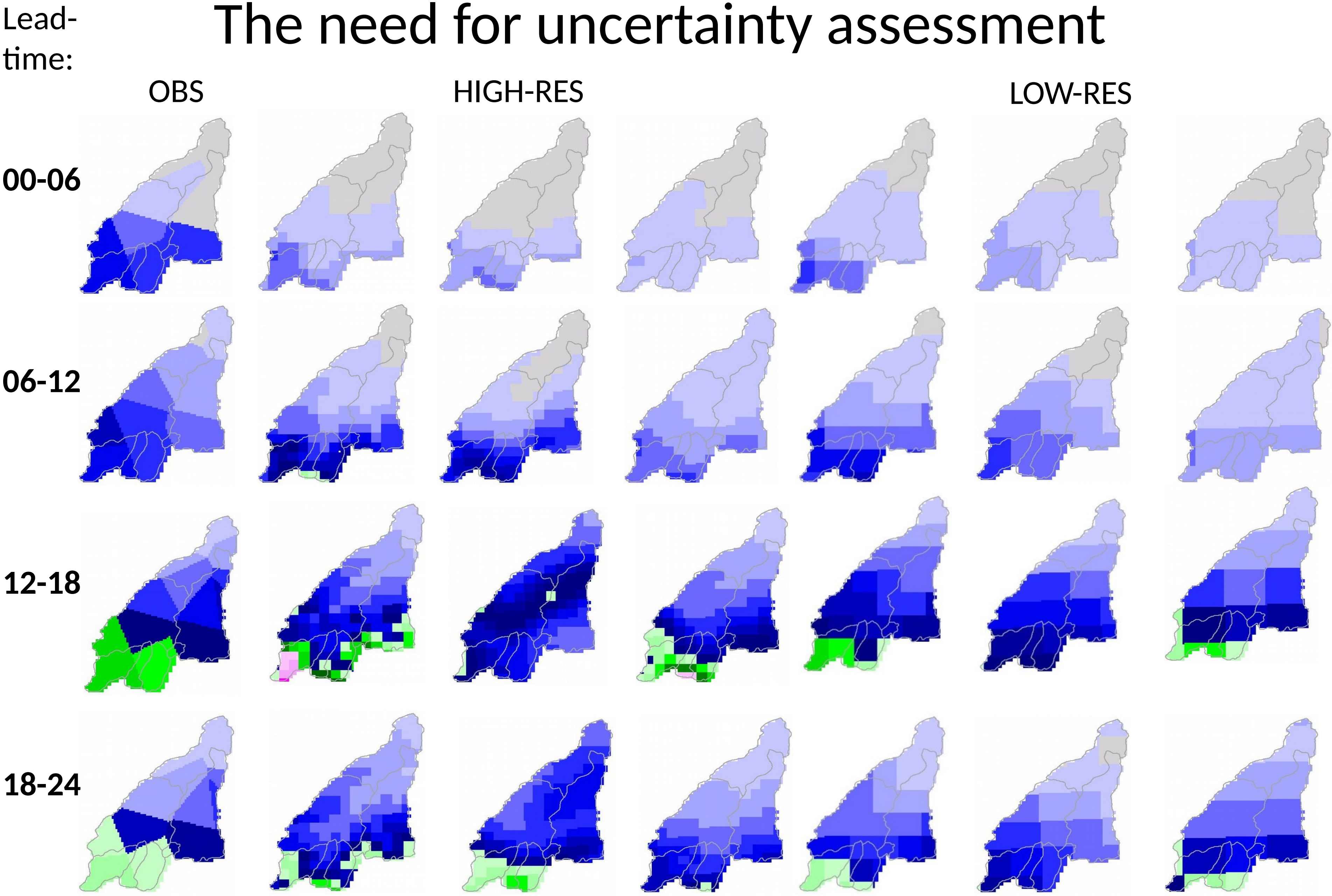


A matter of scale





The need for uncertainty assessment



Summary

- Summary score: Brier and Brier Skill
  - Partition of the Brier score
- Reliability diagrams: Reliability, resolution and sharpness
- ROC: Discrimination
- Diagnostic verification: Reliability and ROC
- Ensemble forecasts: Summary score - CRPS

Thank you!

bibliography

- www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html
- www.ecmwf.int
- 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 D.B. Stephenson, 2003. Forecast Verification: A Practitioner's Guide. In Atmospheric Sciences (Wiley).
- Pertti Nurmi, 2003. Recommendations on the verification of local weather forecasts. ECMWF Technical Memorandum n. 430.
- Stanski, H.R., L.J. Wilson and W.R. Burrows, 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

- Hamill, T.M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, 14, 155-167.
- Mason S.J. and Graham N.E., 1999. “Conditional probabilities, relative operating characteristics and relative operating levels”. *Wea. and Forecasting*, 14, 713-725.
- Murphy A.H., 1973. A new vector partition of the probability score. *J. Appl. Meteor.*, 12, 595-600.
- Richardson D.S., 2000. “Skill and relative economic value of the ECMWF ensemble prediction system”. *Quart. J. Roy. Meteor. Soc.*, 126, 649-667.
- Talagrand, O., R. Vautard and B. Strauss, 1997. Evaluation of probabilistic prediction systems. *Proceedings, ECMWF Workshop on Predictability.*