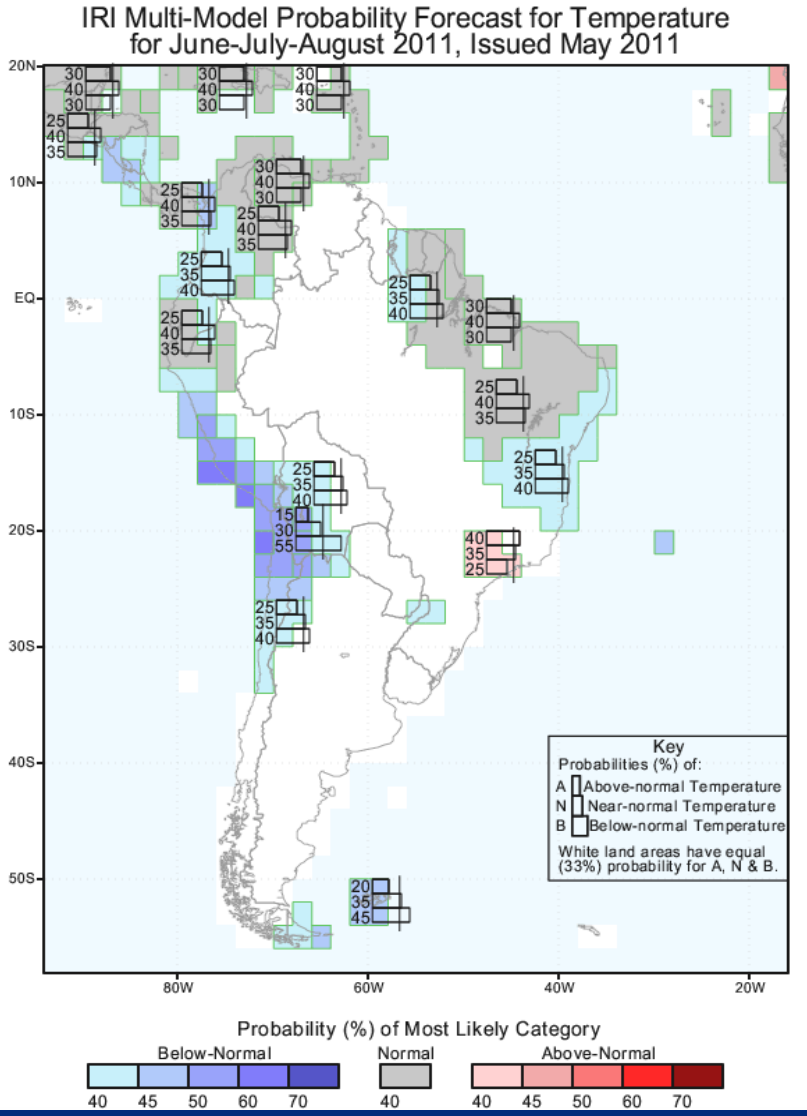
Commission for Climatology Guidance on Verification of Seasonal Forecasts

Simon Mason
simon@iri.columbia.edu
International Research Institute
for Climate and Society
EARTH INSTITUTE | COLUMBIA UNIVERSITY

MedCOF 2015 Training Workshop
Madrid, Spain, 26 – 30 October 2015

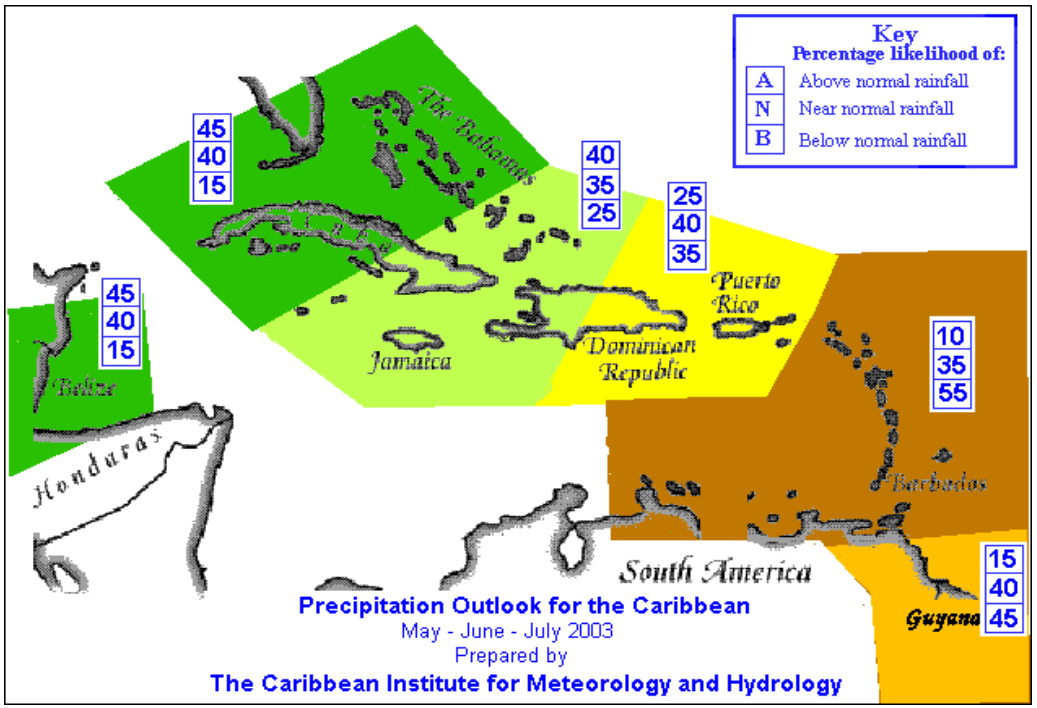
Seasonal forecast formats

2. **(a) Maps showing probabilities of the verification falling within one of two or more categories (by grid)**



Seasonal forecast formats

2. (b) Maps showing probabilities of the verification falling within one of two or more categories (by region)

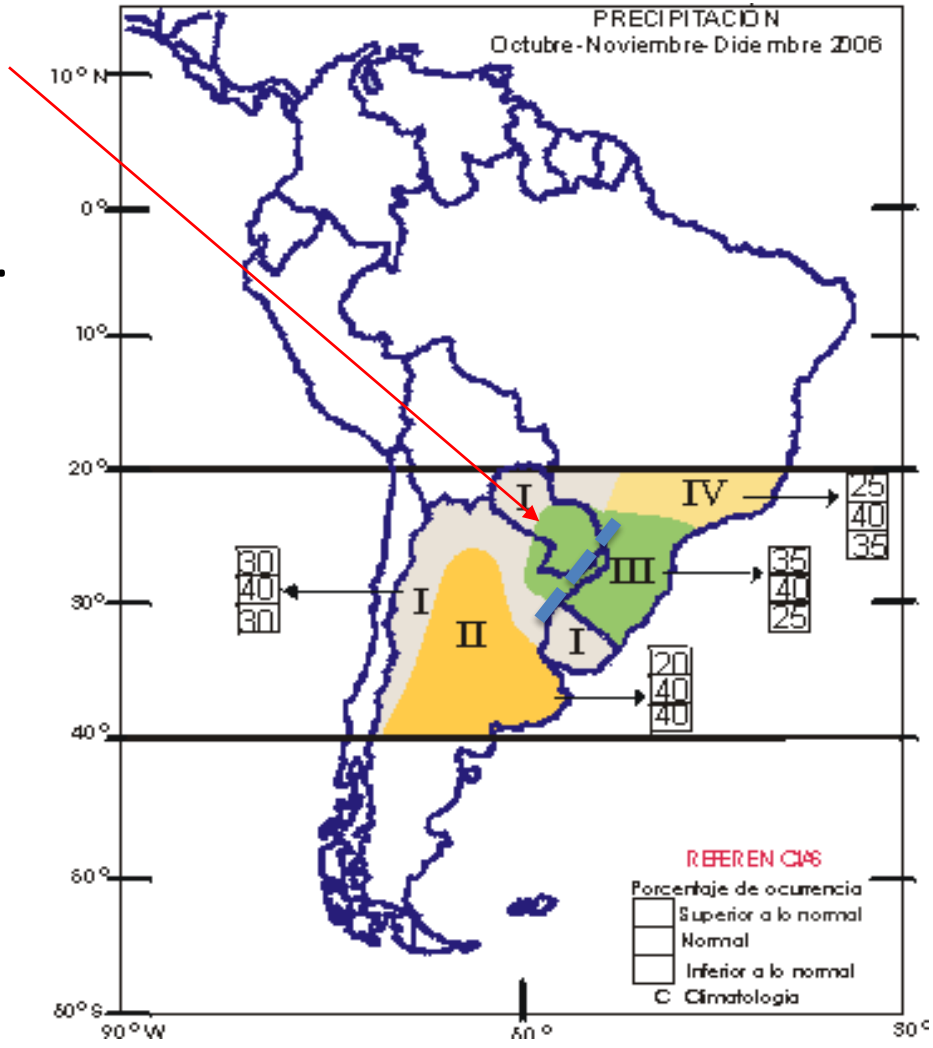


What is the predictand?

Imagine that the green region was originally two regions with the same forecast (A=35; N=40; B=25): S. Paraguay and S. Brazil.

Assume that S. Brazil is Below, and S. Paraguay is Above, and the combined area average is Normal. The forecast verifies well, but the original two forecasts verify badly!

We *must* work towards eliminating ambiguity in seasonal forecasts.



What makes a “good” forecast?

Forecast: This afternoon’s lecture will be so boring it will not be worth attending.

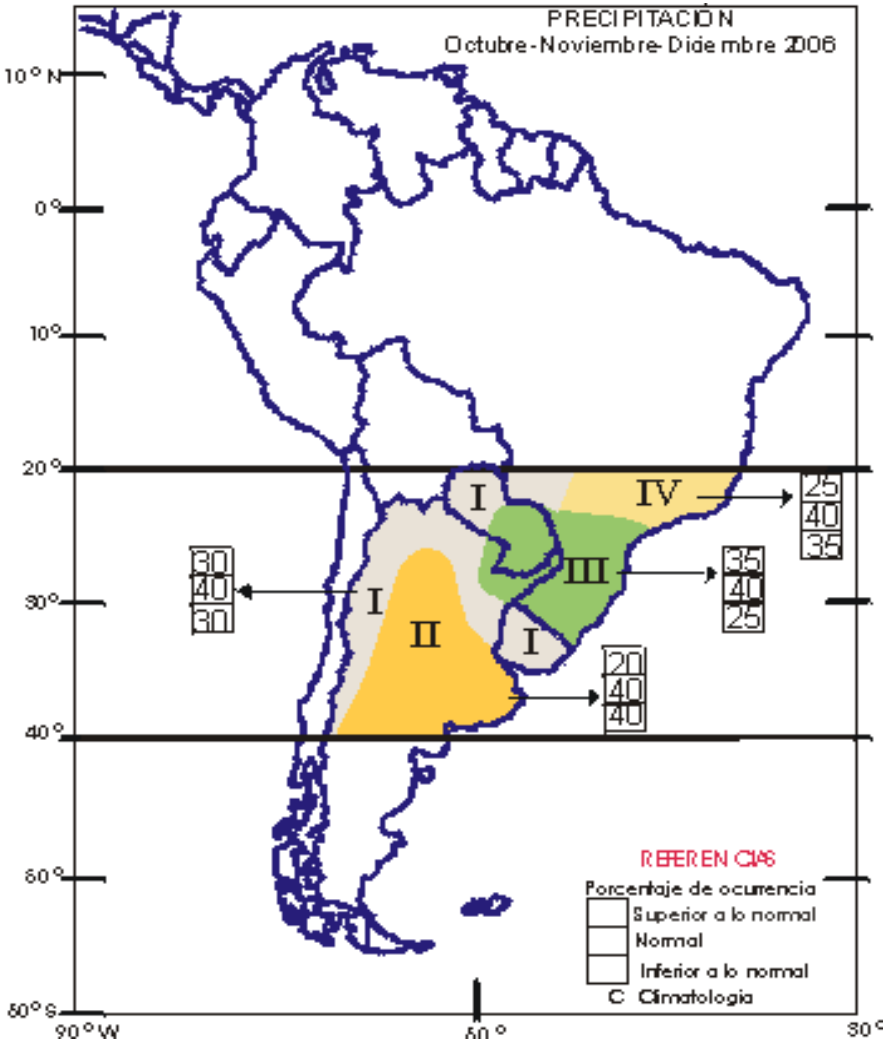
Verification: I lied, so I do not have to work hard today!

Consistency

Ambiguous forecasts

What is the region?

- Encourage consistency of predictand (whether a ll stations, representative stations or area averages)
- Indicate the subregions, if the forecasts are the same. I.e., do **not** combine regions.



Different verification questions

- How good were **these** forecasts?
- How good was **this** forecast?

Different verification questions

- How good were **these** forecasts?
- How good was **this** forecast?

Forecast "goodness"

What makes a "good" forecast?

1. Consistency
2. Quality
3. Value

Murphy AH 1993; *Wea. Forecasting* 8, 281

Forecast "goodness"

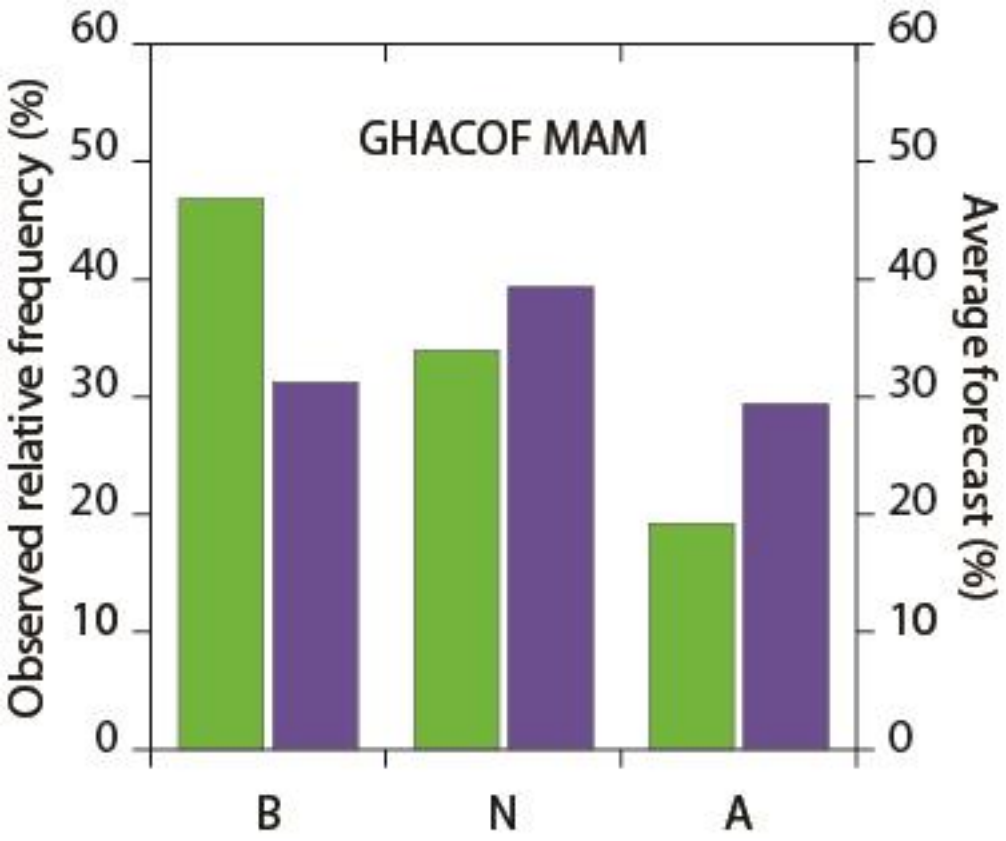
What makes a "good" forecast?

1. Consistency
2. Quality
3. Value

Murphy AH 1993; *Wea. Forecasting* 8, 281

Unconditional bias

- Calculating the frequency of each category over the verification period gives a simple indication of possible hedging. It also indicates whether the verification period has been unusual. Any shift may or may not be permanent.
- Are probabilities consistently too high or too low?



Forecast "goodness"

What makes a "good" forecast?

1. Consistency
2. Quality
3. Value

Murphy AH 1993; *Wea. Forecasting* 8, 281

Hit scores

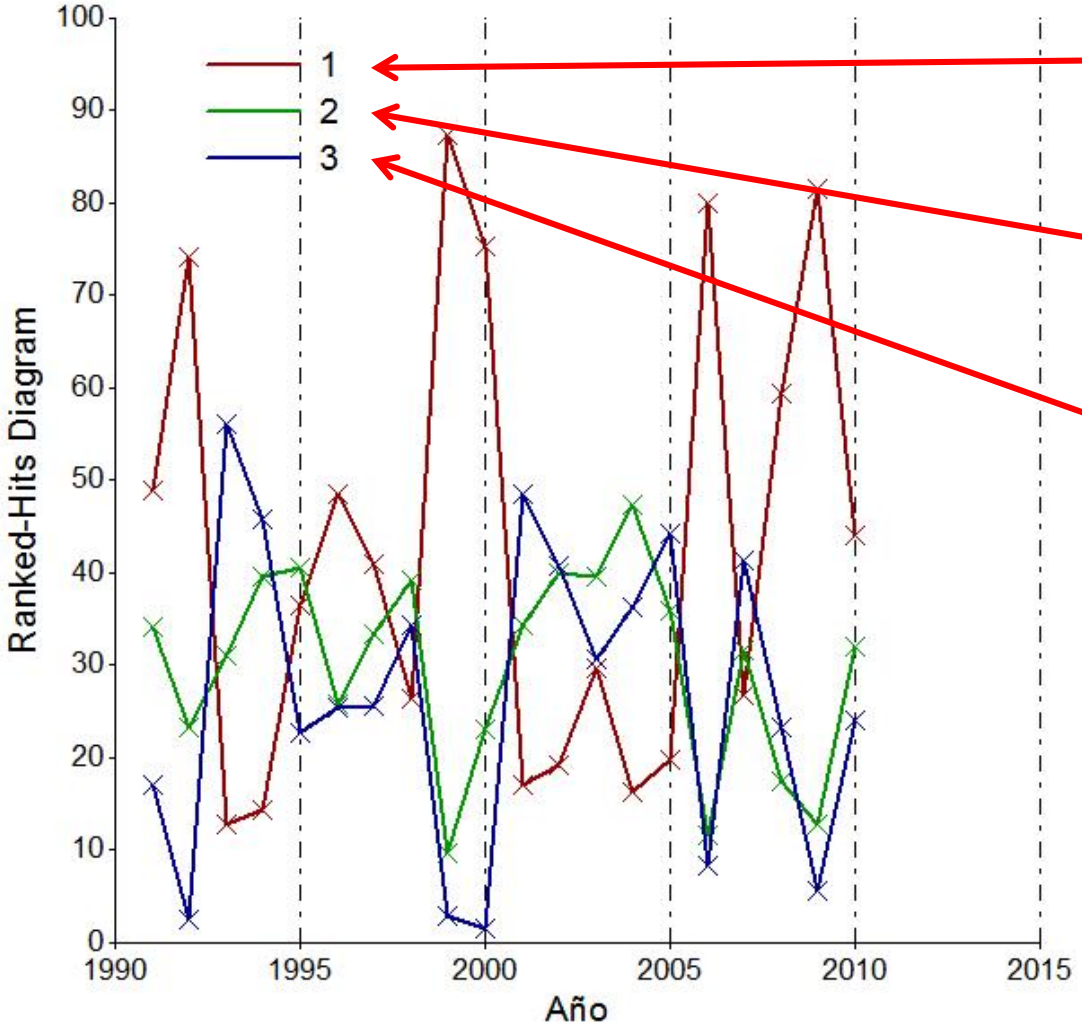
Above	30
Normal	45
Below	25

Above	45
Normal	30
Below	25

If Below occurs, in both cases the least likely category occurs and the two forecasts should be rated equally badly.

Instead of counting “near-misses”, count how often the category with the second highest probability occurs.

Ranked Hits diagrams



highest probability

second highest probability

lowest probability

Retroactive forecasts of JFM 1991 – 2010 Colombia rainfall using December Pacific SSTs.

Category with highest probability is occurring most frequently.

What makes a "good" probabilistic forecast?

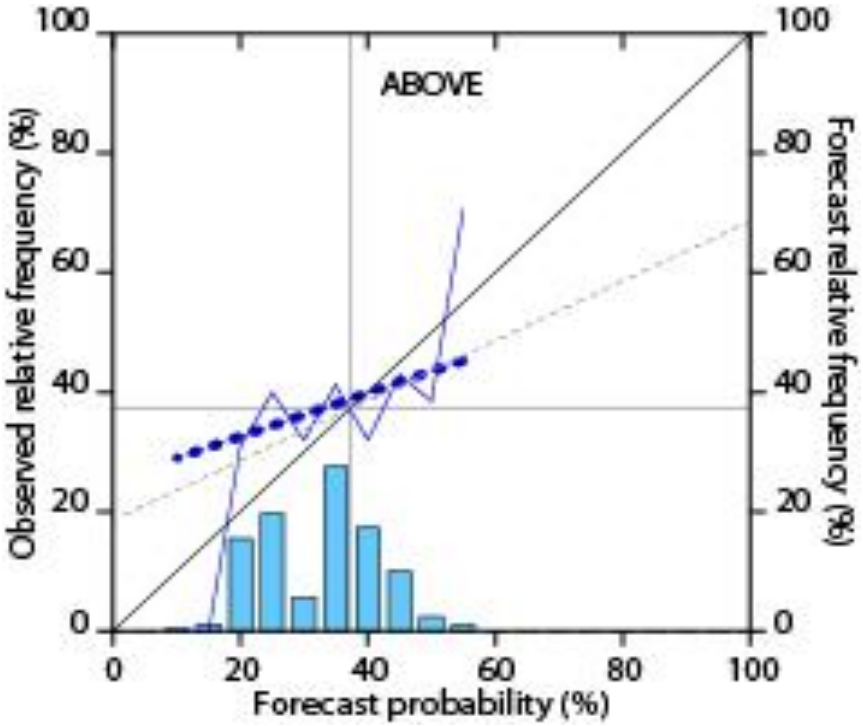
Reliability	the event occurs as frequently as implied by the forecast
Sharpness	the forecasts frequently have probabilities that differ from climatology considerably
Resolution	the outcome differs when the forecast differs
Discrimination	the forecasts differ when the outcome differs

Which attribute(s) is the hit score measuring?

Measure the attributes separately.

Discrimination is easier to measure than resolution.

Attributes diagrams



The histograms show the sharpness.

The vertical and horizontal lines show the observed climatology and indicate the forecast bias.

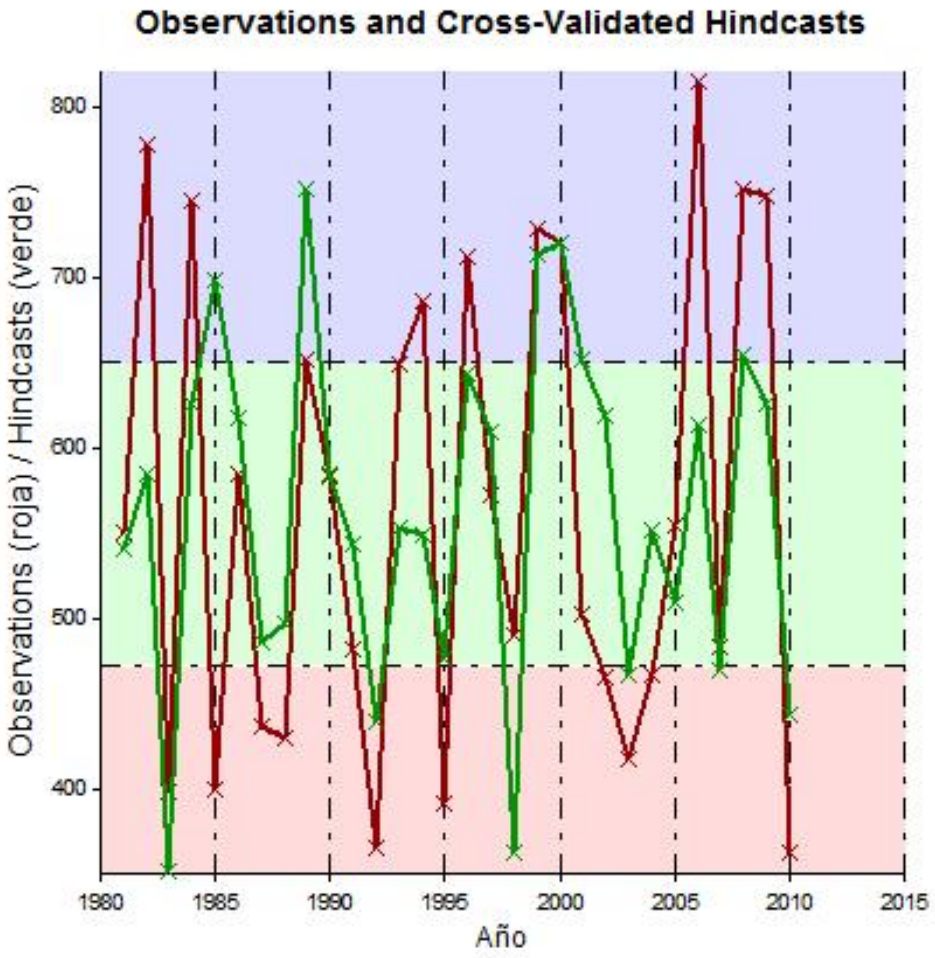
The diagonal lines show reliability and "skill".

The coloured line shows the reliability and resolution of the forecasts.

The dashed line shows a smoothed fit.

What if there are only a few forecasts or if we want results for individual locations …

ROC



Forecasts of JFM rainfall for Colombia.

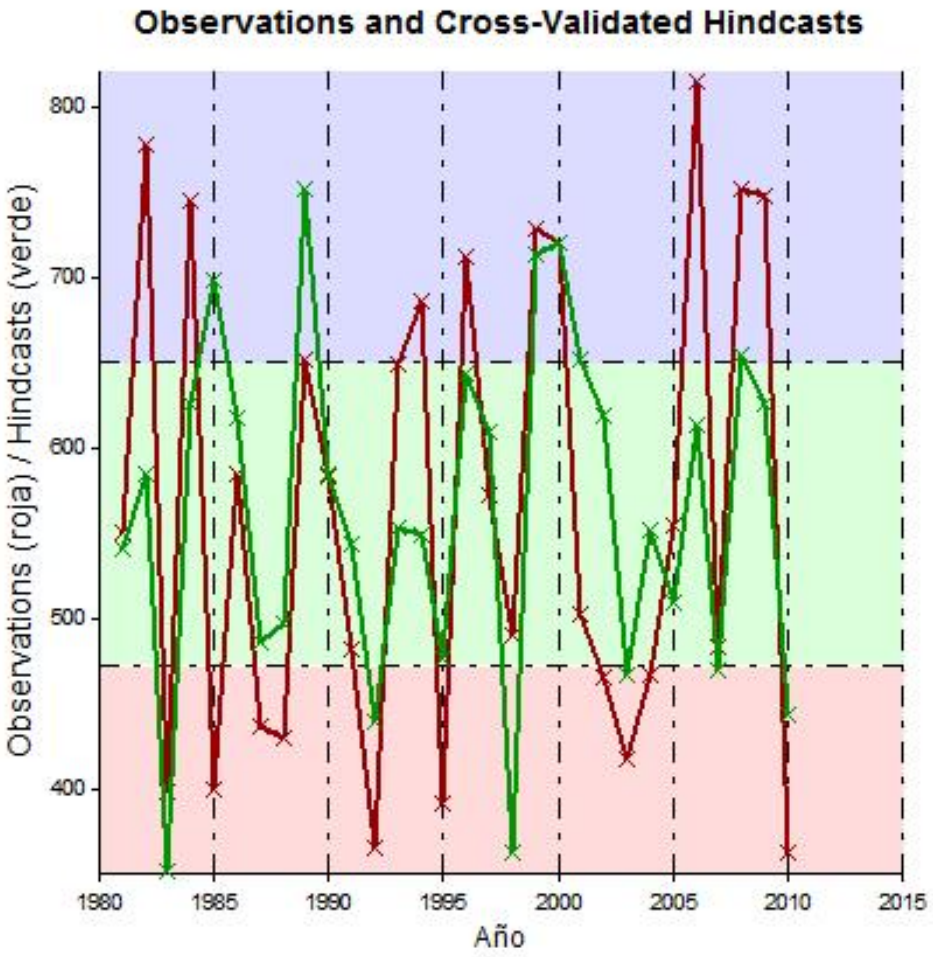
Looking only at the forecasts, which year are you most confident is a dry year?

Was it dry?

Yes: score a hit

No: score a false-alarm

ROC

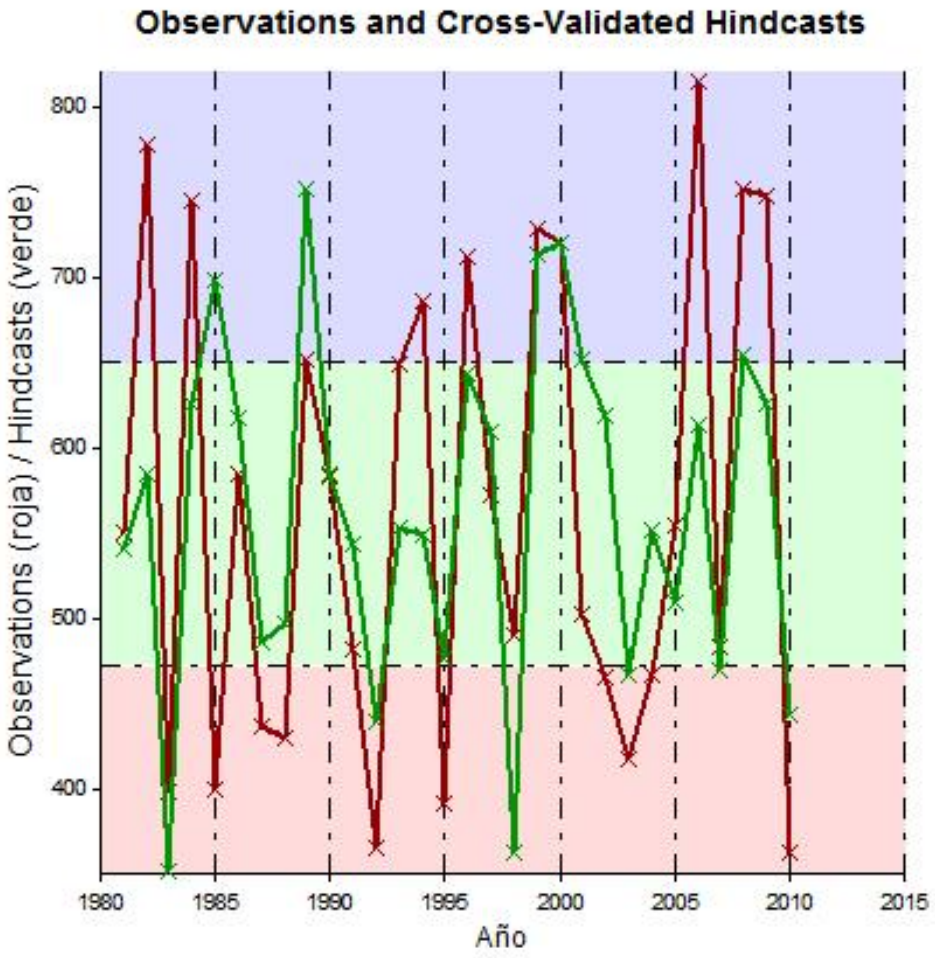


For the first guess:

$$\text{Hit rate} = \frac{\text{number of hits}}{\text{number of events}} = \frac{1}{10}$$

$$\text{FAR} = \frac{\text{number of false alarms}}{\text{number of non-events}} = \frac{0}{20}$$

ROC



Forecasts of JFM rainfall for Colombia.

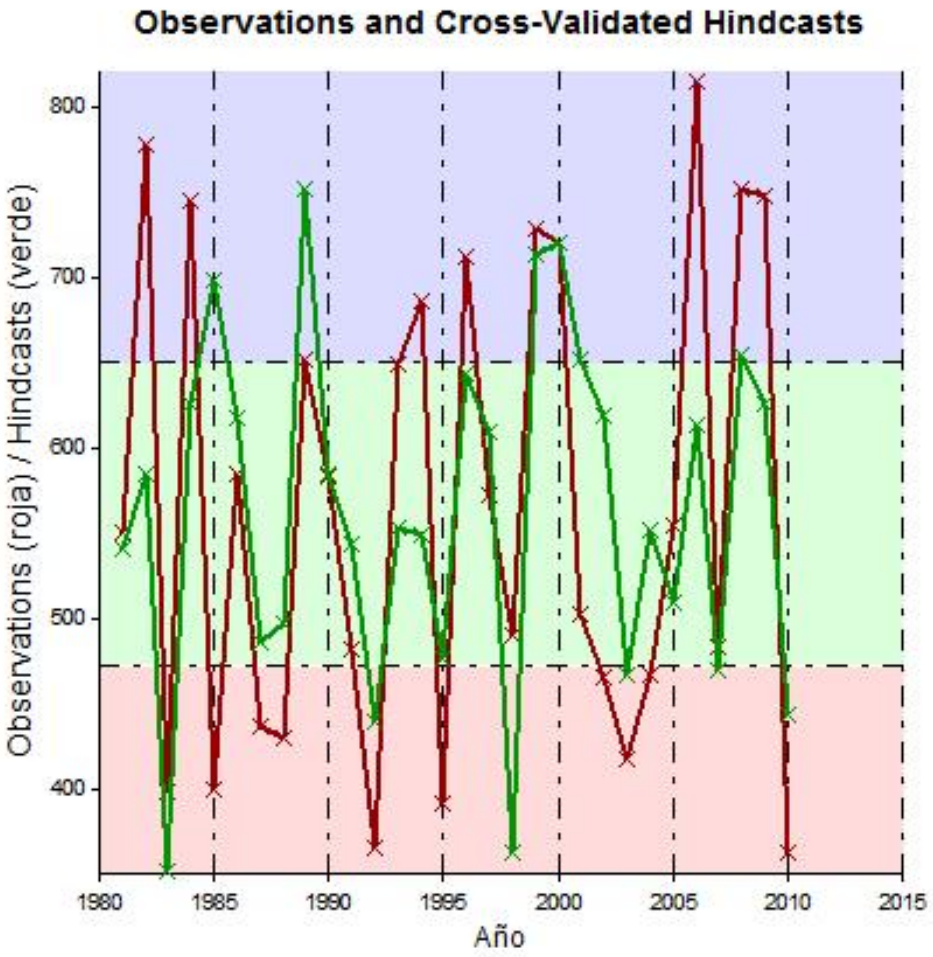
Looking only at the forecasts, which year are you next most confident is a dry year?

Was it dry?

Yes: score a hit

No: score a false-alarm

ROC



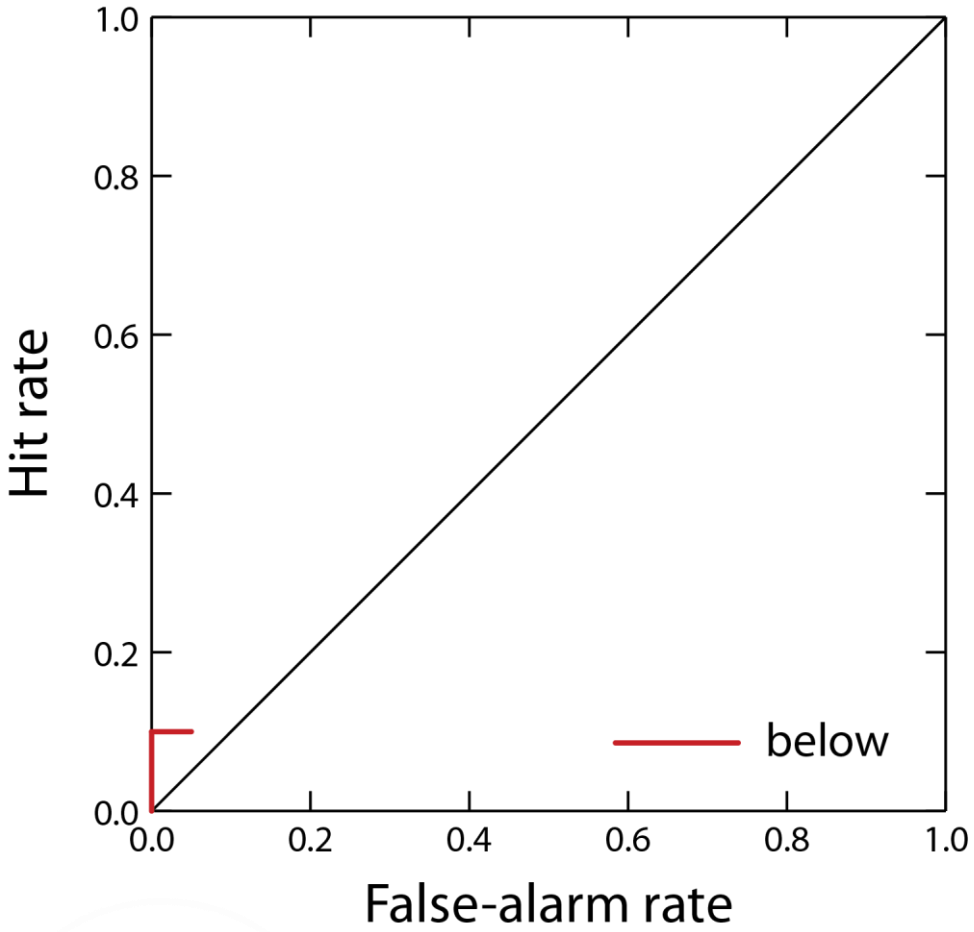
For the second guess:

$$\text{Hit rate} = \frac{\text{number of hits}}{\text{number of events}} = \frac{1}{10}$$

$$\text{FAR} = \frac{\text{number of false alarms}}{\text{number of non-events}} = \frac{1}{20}$$

Repeat for all forecasts.

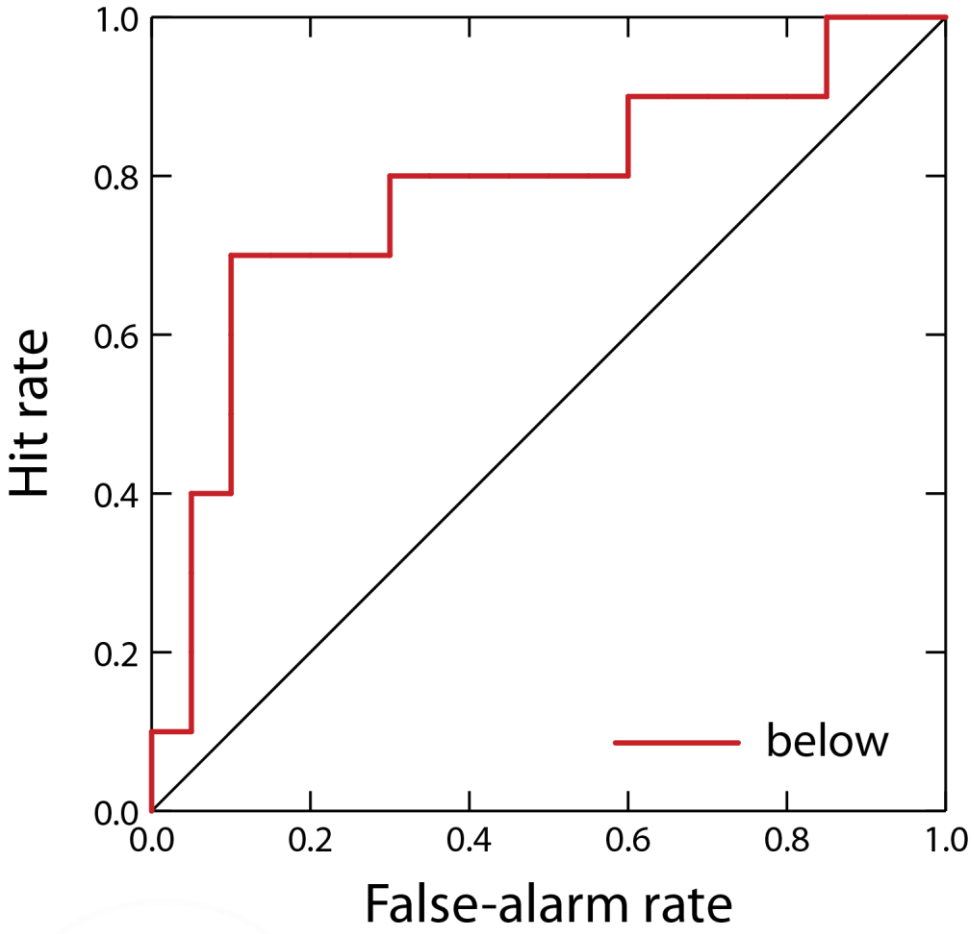
ROC



Plot the correct scores (hit-rate) against the incorrect (false-alarm rate) scores.

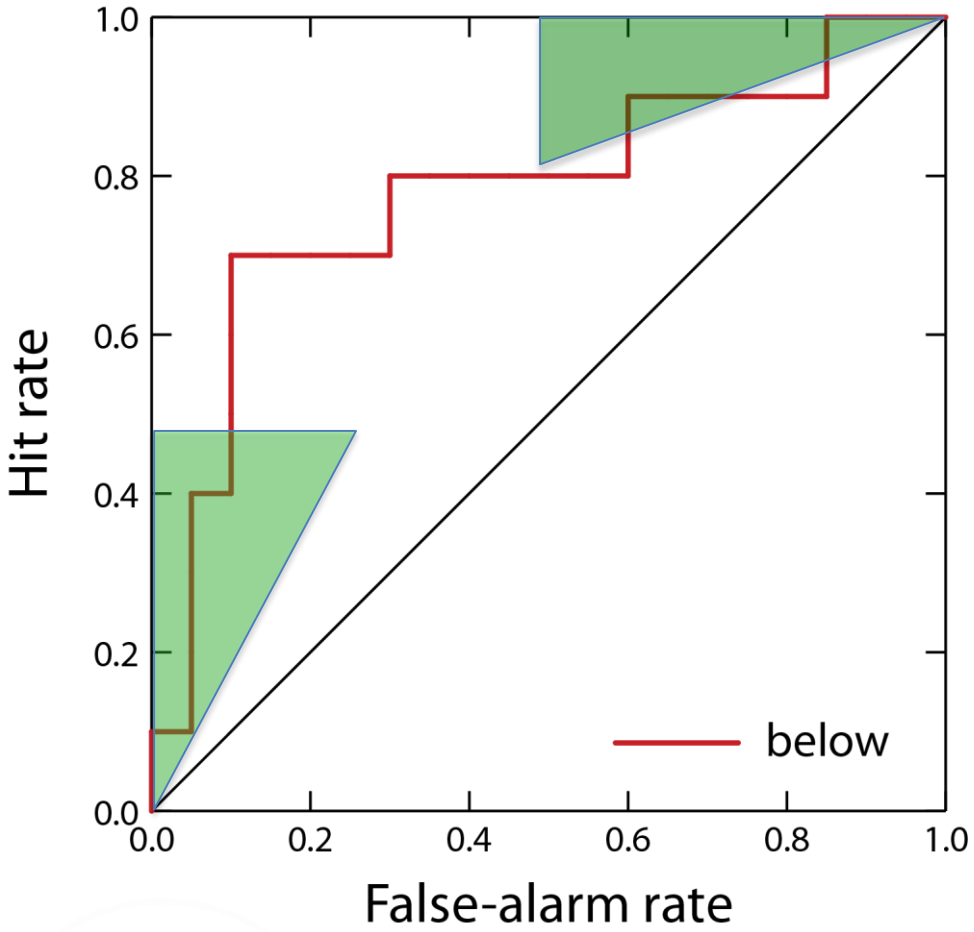
We want the correct scores to be larger than the incorrect scores, i.e., for the graph to be above the diagonal.

ROC



Continue calculating the scores until all the years have been selected when all the events and all the non-events have been selected.

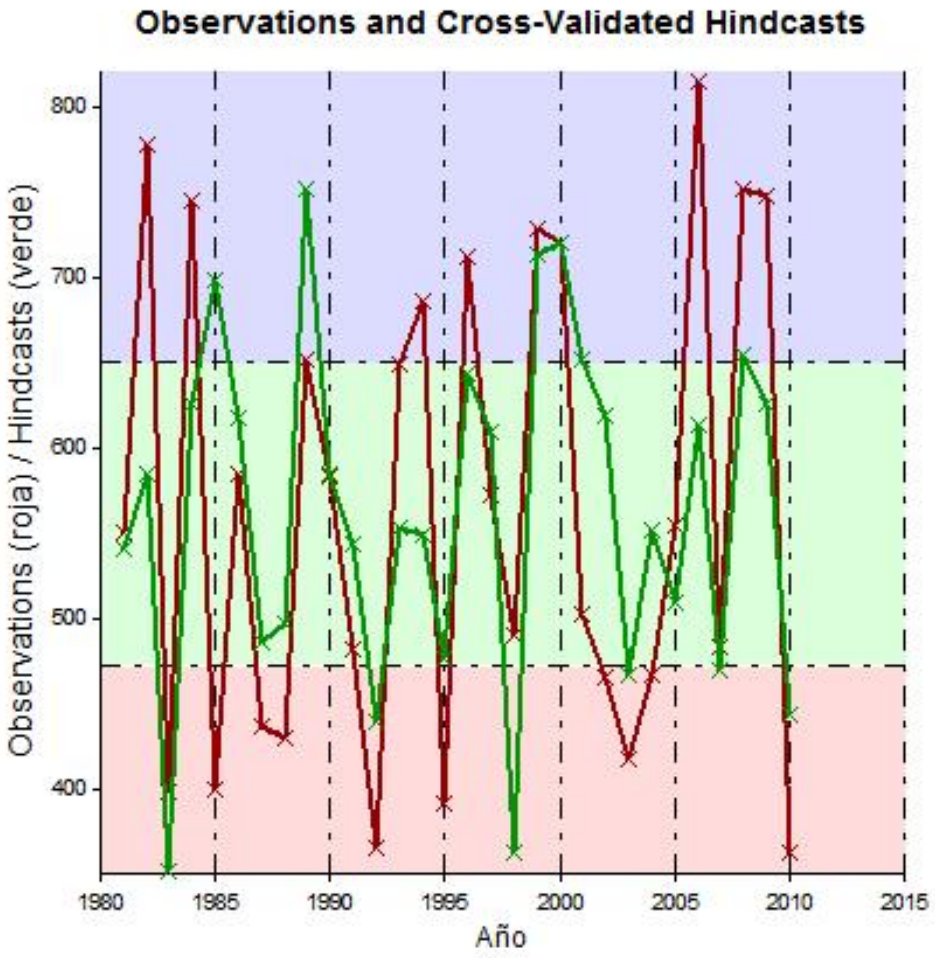
ROC



The bottom left indicates whether the forecasts with strong indications of dry (or wet) are good. Can they indicate that an event will occur?

The top right indicates whether the forecasts with strong indications of **not** dry (or **not** wet) are good. Can they indicate that an event will **not** occur?

ROC



Forecasts of JFM rainfall for Colombia.

Looking only at the forecasts, which year are you most confident is **not** a dry year?

Was it dry?

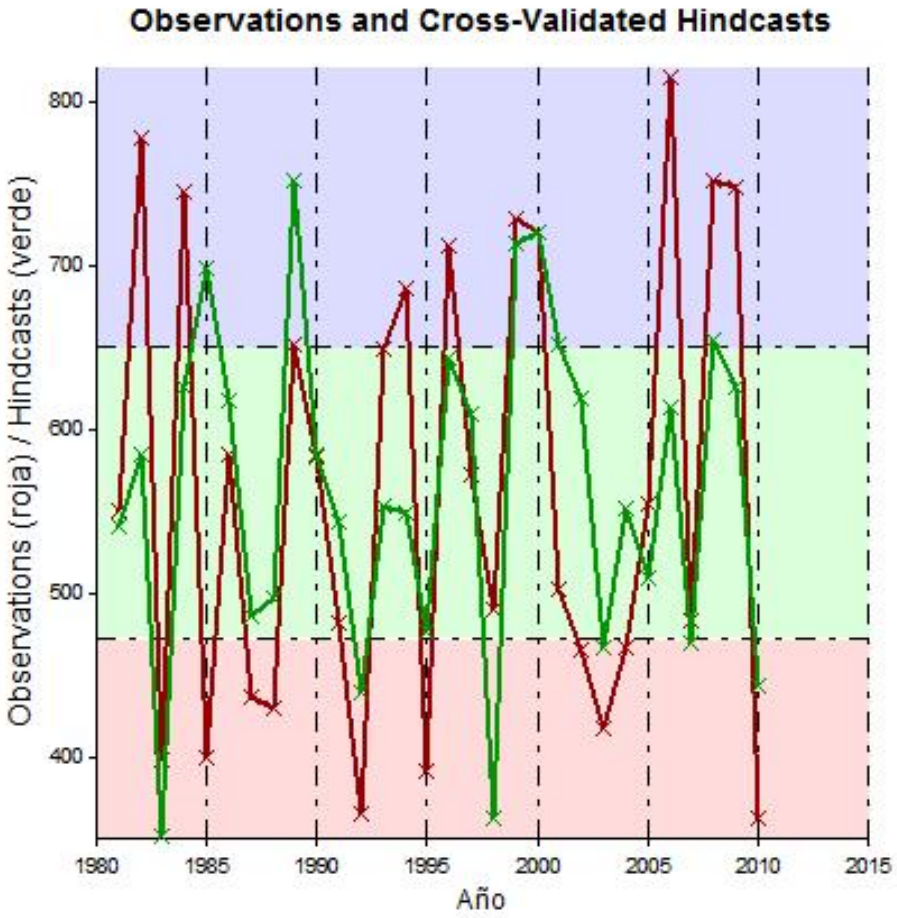
Yes: score a hit

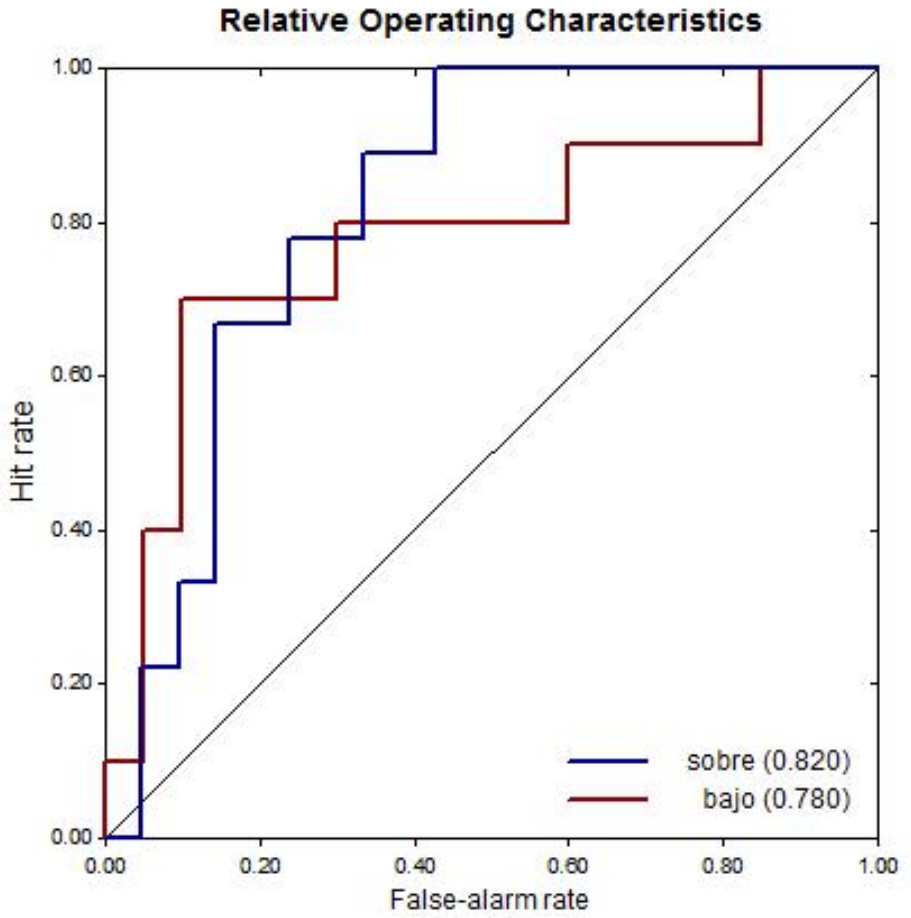
No: score a false-alarm

NB We want to score a false-alarm

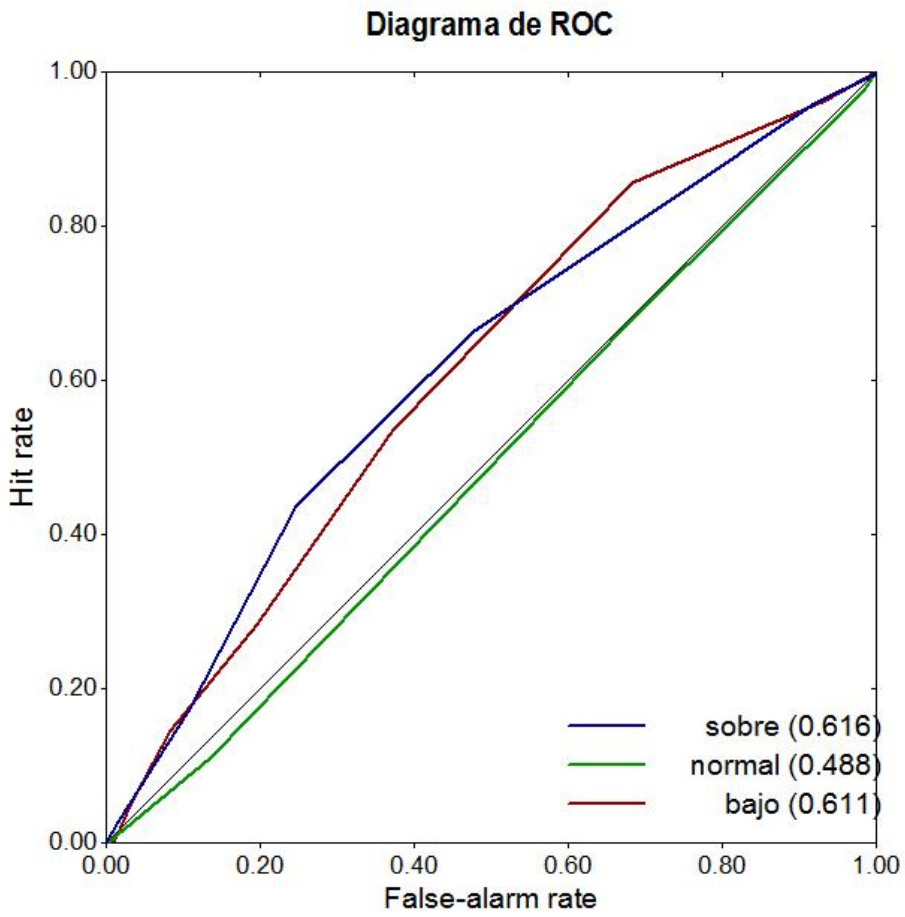
ROC

Repeat for the above-normal category.





ROC diagrams



Retroactive forecasts of JFM 1991 – 2010
Columbia rainfall using December Pacific SSTs

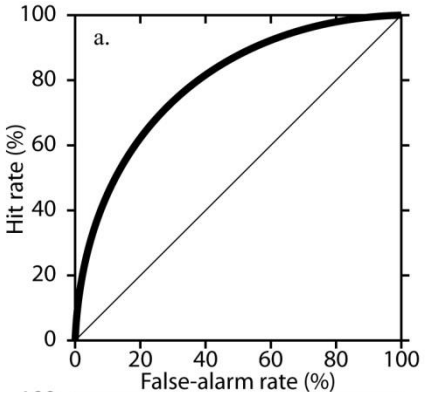
The ROC can be calculated for probabilistic forecasts

ROC areas: do we issue a higher probability when the category occurs?

Graph bottom left: when the probabilities are high, does the category occur?

Graph top right: when the probabilities are low, does the category not occur?

Relative Operating Characteristics



Forecast "goodness"

What makes a "good" forecast?

1. Consistency
2. Quality
3. Value

Murphy AH 1993; *Wea. Forecasting* 8, 281

"Weather Roulette"

Imagine that you are able to invest in climate-sensitive sectors, but you need to decide whether to invest more in sectors that will succeed if rainfall is below-normal, or normal, or above-normal.

The investments return fair odds against climatology.

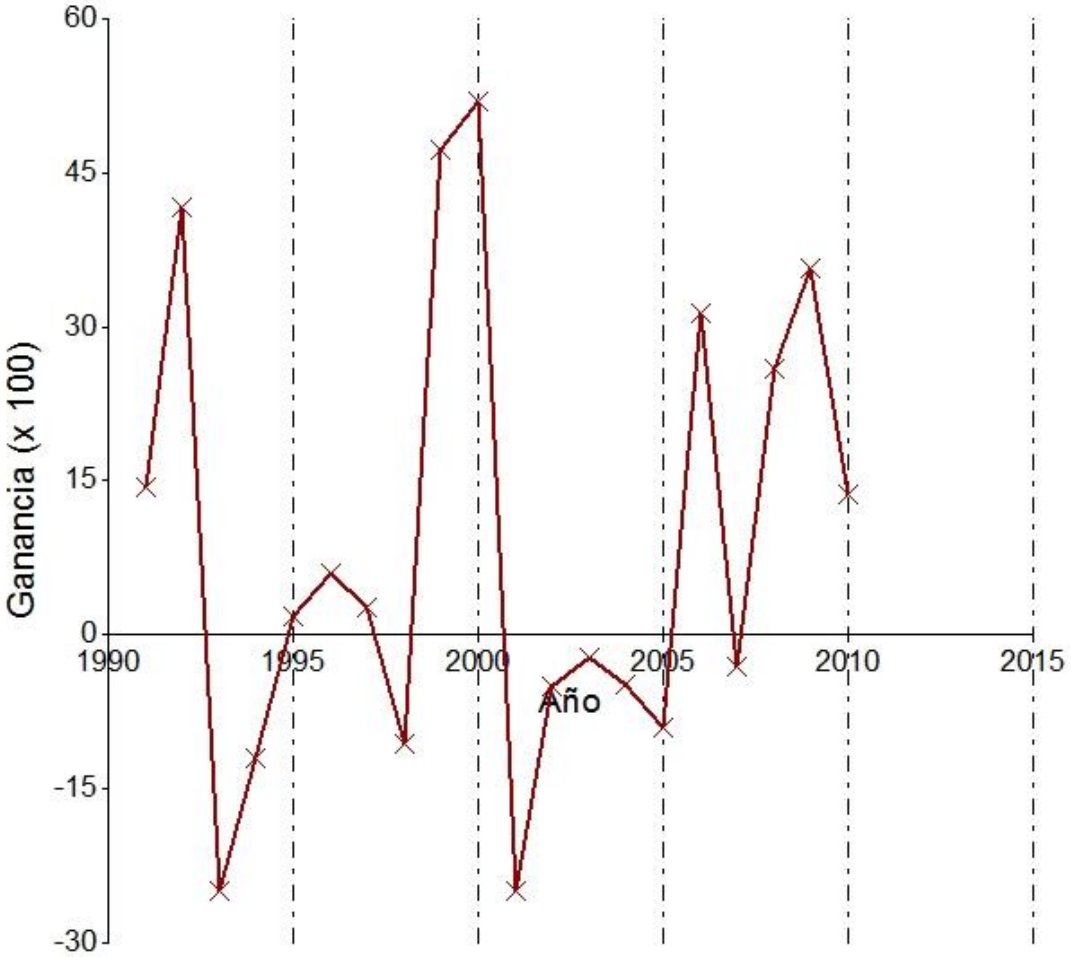
The odds of a given category occurring are:

$$\text{odds} = \frac{p}{1-p} = \frac{0.333}{1-0.333} = \frac{1}{2}$$

i.e., for every one time you win, you will lose twice.

If you invest €1m on below-normal and below-normal occurs, you would make a profit of €2m (and get the €1m back), but if below-normal does not occur you would lose the €1m.

Weather roulette – profits diagram

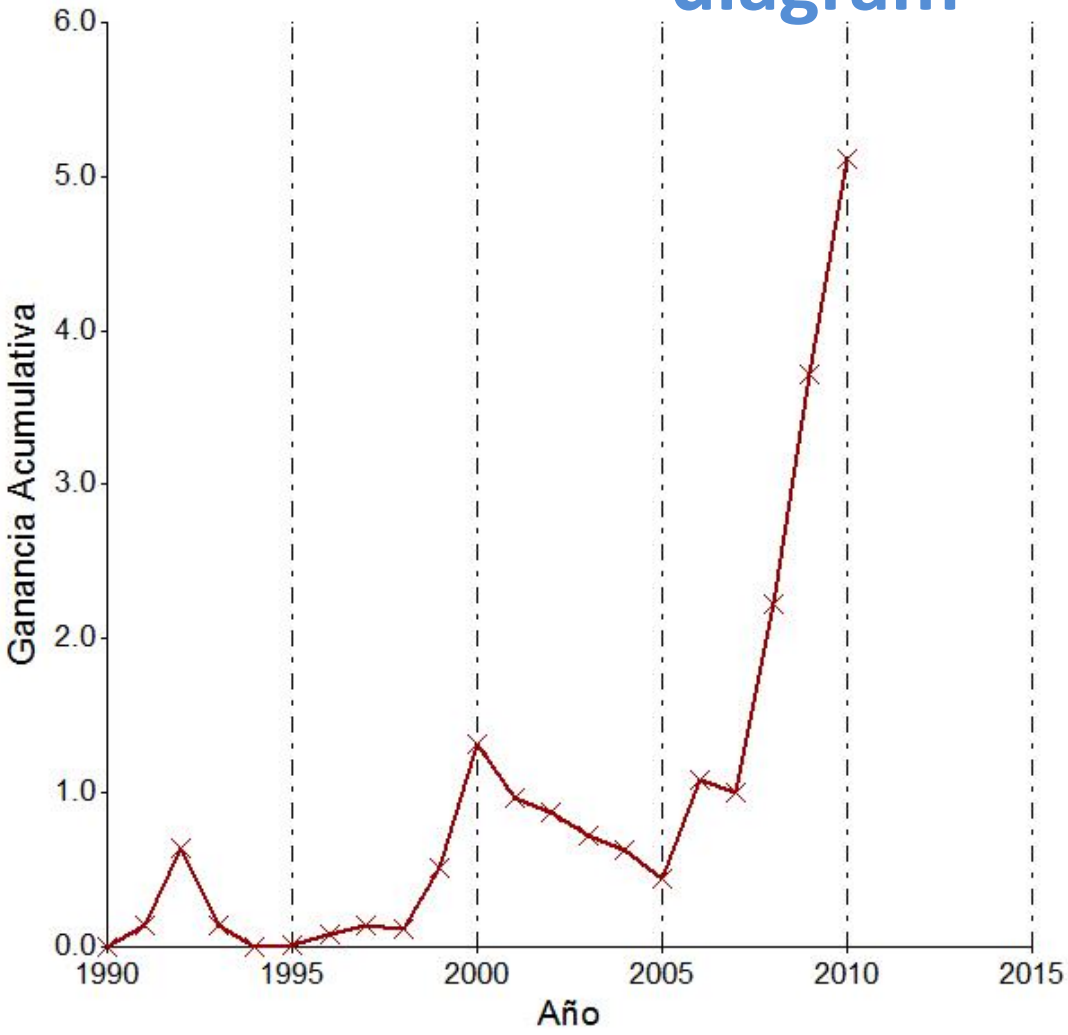


Given fair odds:

profit = 1 ÷ odds

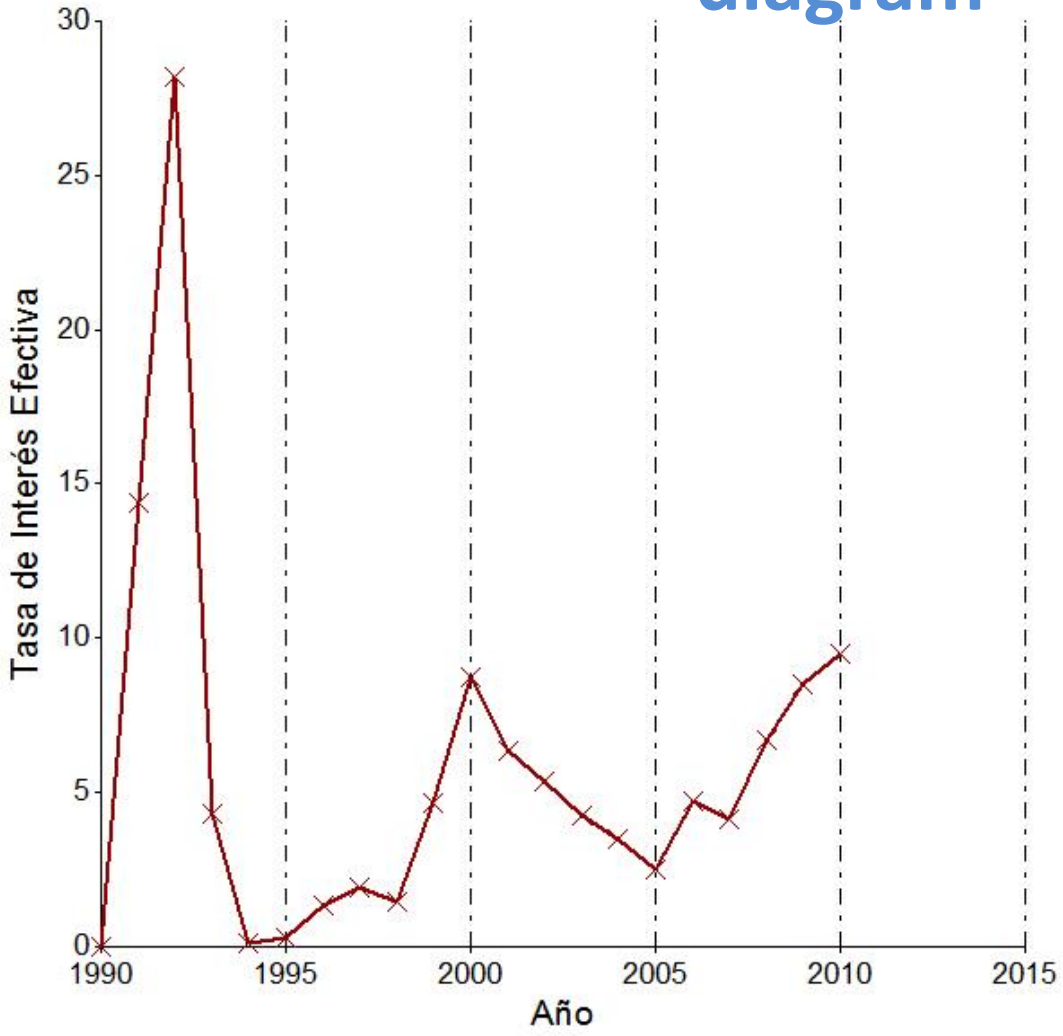
Multiply the investment by the profit (or loss) to indicate how much money would be made (or lost). Average over all locations.

Weather roulette – cumulative profits diagram



Multiply the initial investment by the profit (or loss) carried over each year to indicate how much money would be made (or lost).

Weather roulette – effective interest rate diagram

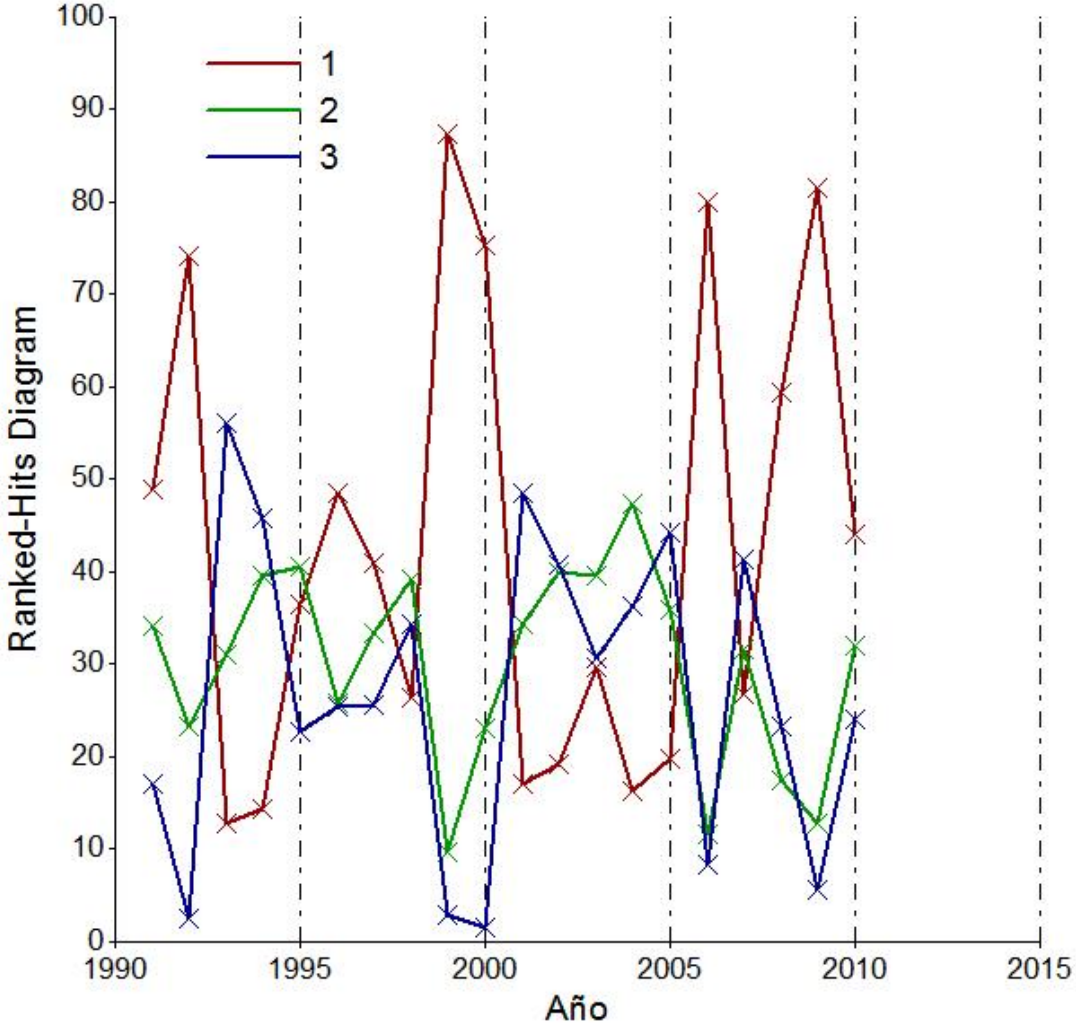


Multiply the initial investment by the profit (or loss) carried over each year, and calculate the effective interest rate.

Different verification questions

- How good were **these** forecasts?
- How good was **this** forecast?

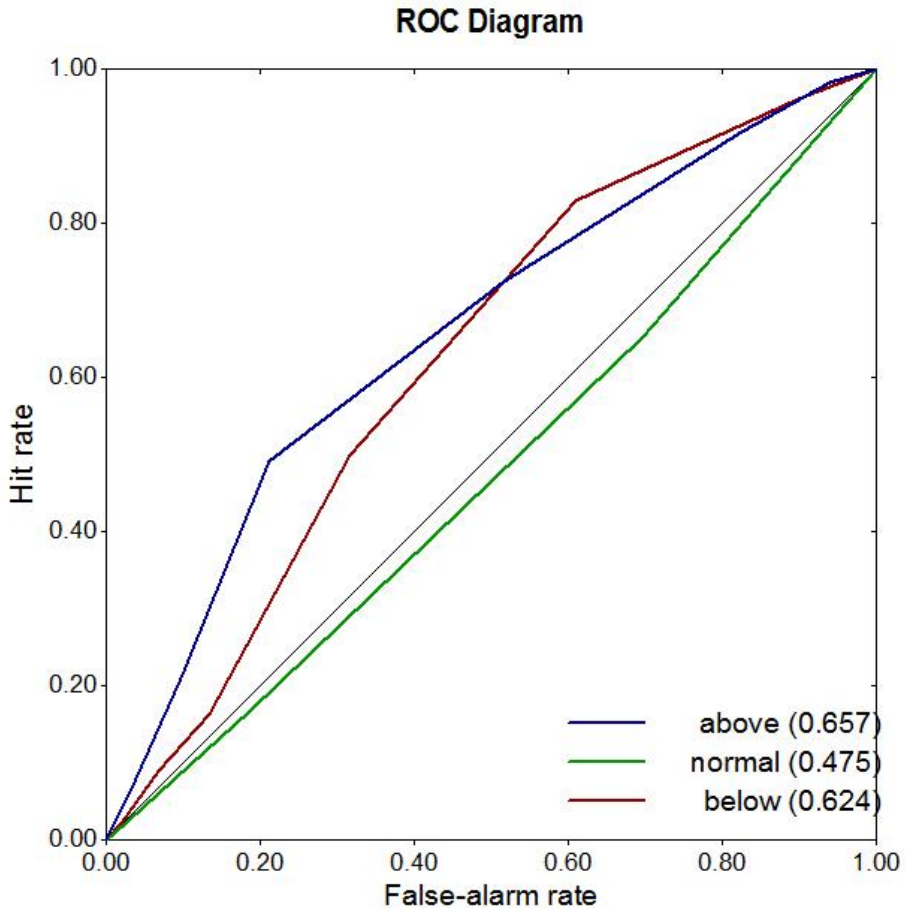
Ranked Hits diagrams



Retroactive forecasts of JFM 1991 – 2010 Colombia rainfall using December Pacific SSTs.

Category with highest probability is occurring most frequently.

ROC diagrams

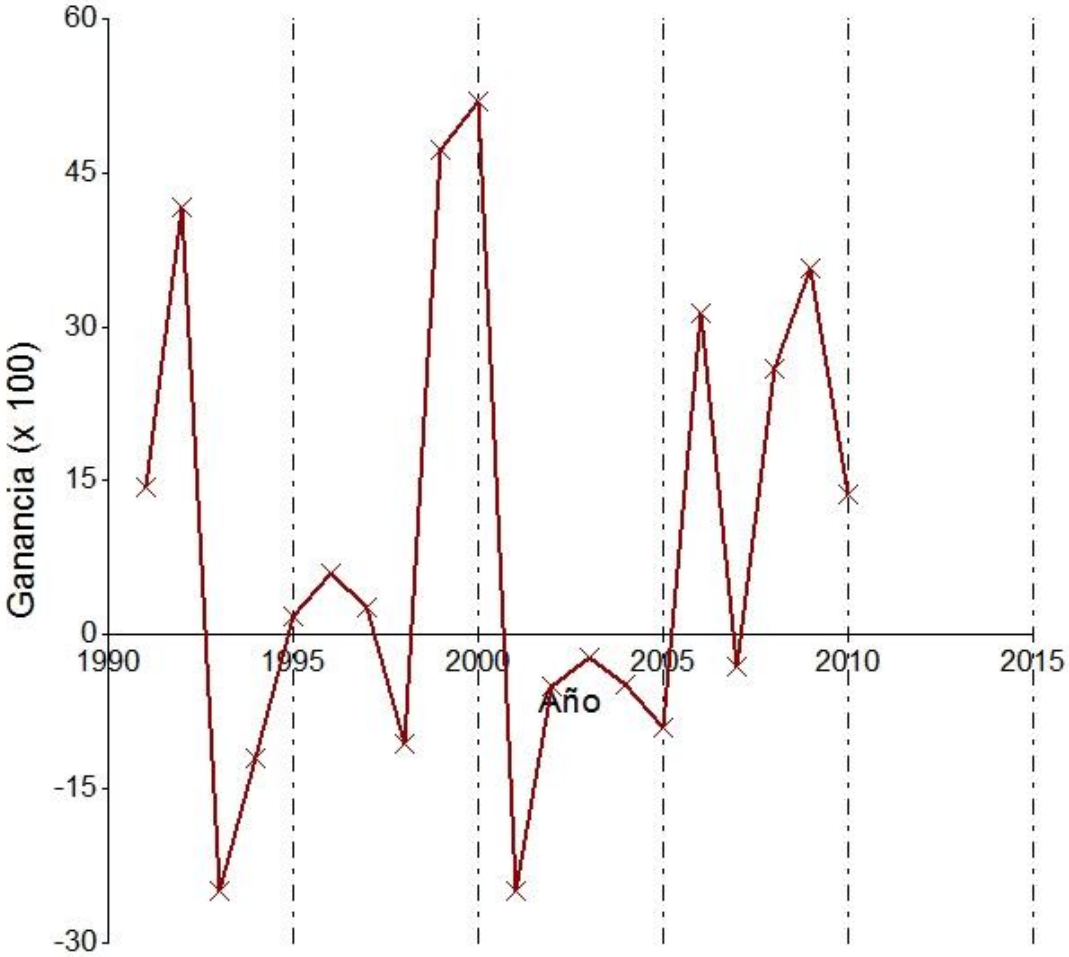


Retroactive forecasts of MAM 1986 – 2010
Thailand rainfall using February Pacific SSTs

Because they are insensitive to unconditional bias ROC diagrams do not work sensibly for forecasts individual years.

Do **not** use them in this context!

Weather roulette – profits diagram



Average across stations – but the score is then not proper.

Multiply across stations and the score is proper, but the interpretation no longer works. Instead – given the forecasts how much additional information is required to determine what the observations were?

Summary

- Many consensus forecasts are ambiguous: this problem *must* be addressed.
- Discrimination may be hard to measure if there are insufficient years of forecasts; instead try measuring discrimination.
- Hit scores (based on the ranked probabilities) give a useful, but overly simple measure of goodness.
- Some simple measures of forecast value are suggested, based on “weather roulette”.
- There are few good options for verifying individual years. Hit scores and weather roulette measures can be used.