Seasonal Forecast Verification
Introduction & concepts

WEATHER CLIMATE WATER
TEMPS CLIMAT EAU

JP. Céron
and contribution from S. Mason (IRI)
jpceron.wmo@gmail.com

WMO OMM

World Meteorological Organization
Organisation météorologique mondiale

lectures

1. **Introduction**
2. *Forecast Attributes*
3. *ROC & Reliability - Exercise*
4. *Significance & Robustness*
5.

Introduction

Why Verification ?

For Modelers

- Detection of problems and discrepancies
- Validation and evaluation of models,
- Improvement of models
- Comparison of models

For Users :

- Better knowledge of model performance over the region of interest
- Better use of the information
- Assessment of contribution of the forecast as additional information to the user's activity
- Assessment of the « value » of the forecasting information

Introduction

Comparing the forecast to what ?

- **Verification dataset**
  - Observations (not regularly located, needs of rules to compare « local » informations and model forecasts)
  - Users's dataset
  - Model analysis (problem of the size of the series with respect to the homogeneity of the analysis)
  - Reanalysis (most of the time for Climate Models)
  - Grided dataset, (e.g. E-Obs, GPCP, …)

Quality of the reference data crucial !

- **Forecasting strategy (including users)** :
  - Climatology
  - Persistence
  - Other forecasting system

Introduction

To answer to which question ?

- **Different aspects for Modelers**
  - Is the model Good ? Skilful ?
  - Is the uncertainty estimate correct ?
  - Is the model perform better than another existing model ?

- **Different aspects for Users**
  - Is the information useful (including for Decision) ?
  - Is the information bring added value ?
  - Has the information some value ?
  - *Has the use of the information some impact on the user's activity?*

Introduction

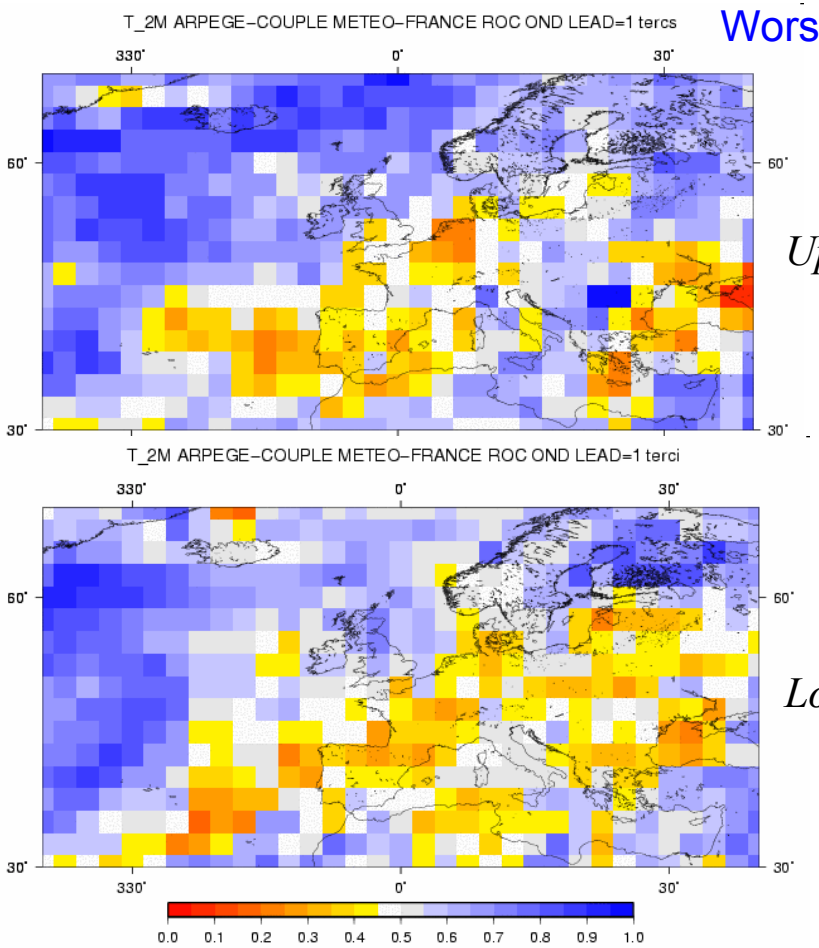
Additional consideration

- **On Verification on the hindcast**
  - Need of large sample size to get significant results
  - Numerous criteria (scores and skill-scores – see lecture on scores) which help to answer to specific questions
  - Choice of the target (question to address) crucial
  - Specific aspects related to the probabilistic nature of the forecast
  - Interpretation about little skill (predictability vs model)
  - Significance and robustness
  - **Caution in interpretation and use**

Skill of Seasonal forecasts

- Quality of Seasonal T2m forecasts

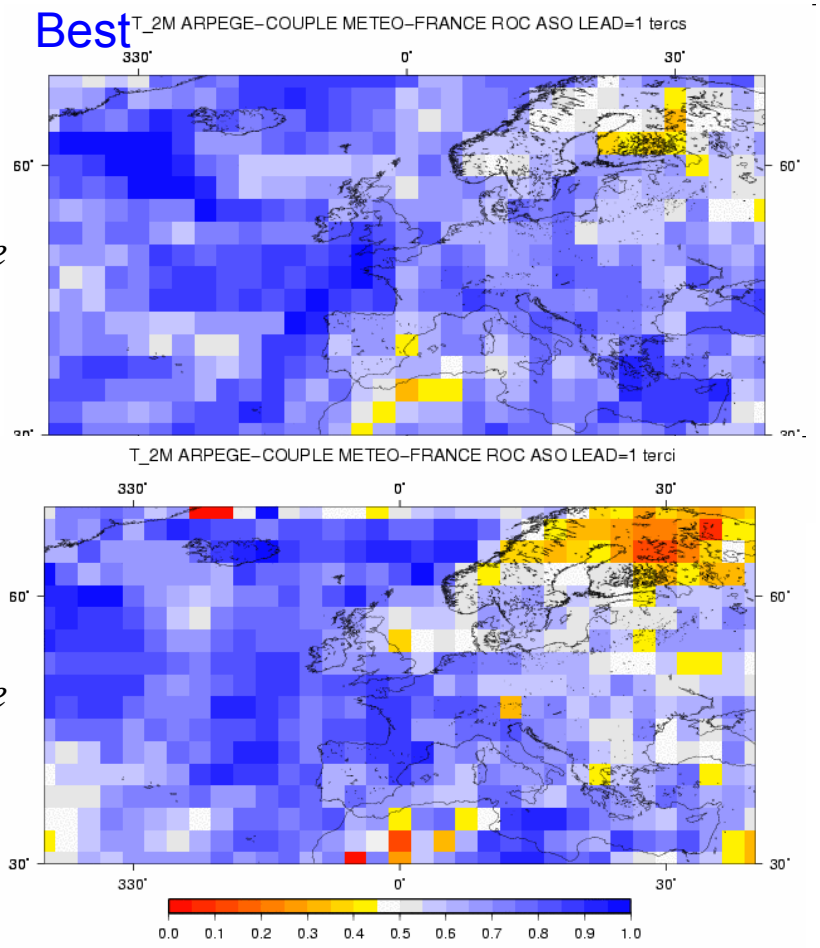
Worst



Upper tercile

Lower tercile

Best

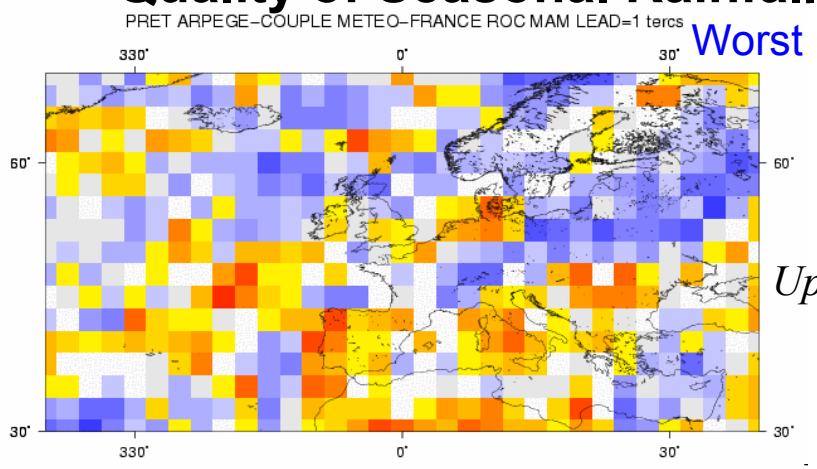


October-November-December

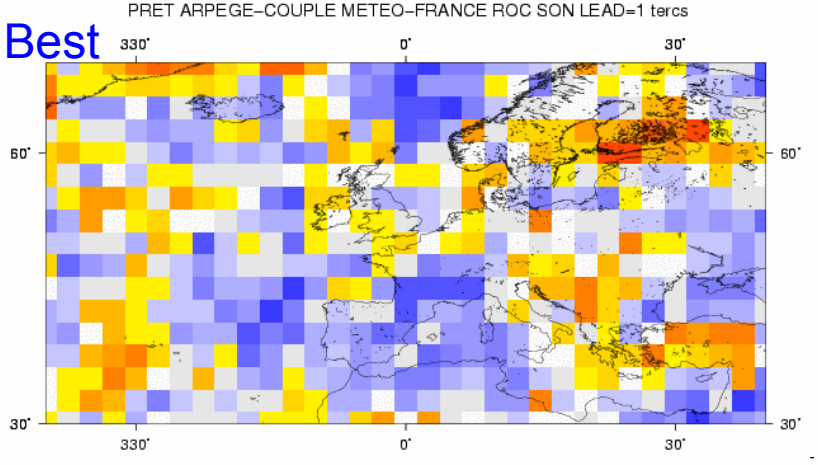
August-September-October

Skill of Seasonal forecasts

Quality of Seasonal Rainfall forecasts

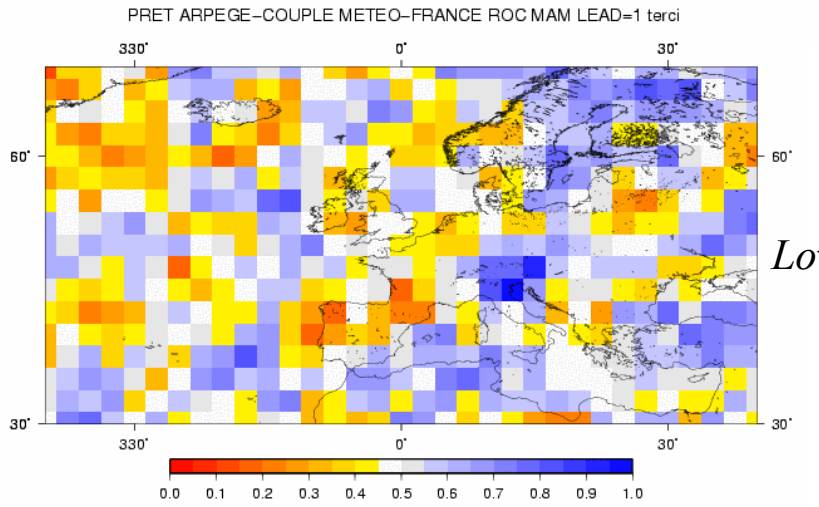


Worst

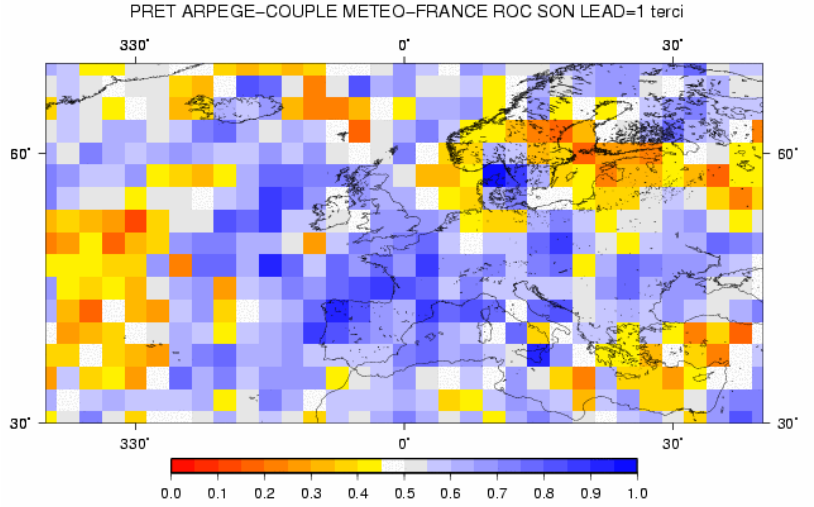


Best

Upper tercile



March-April-May



September-October-November

Lower tercile

Introduction

Additional consideration

- **Verification of the current forecast**
  - Specific criteria to be used
  - Impact of the predictability (and associated diagnosis)
  - Impact of weather vs climate
  - **Caution in interpretation and use**

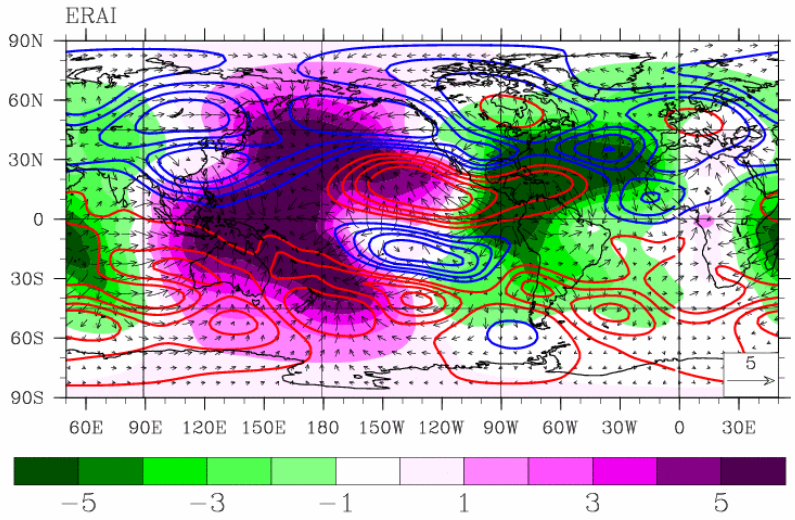
Reliability and Skill

- How can we detect the predictability ?

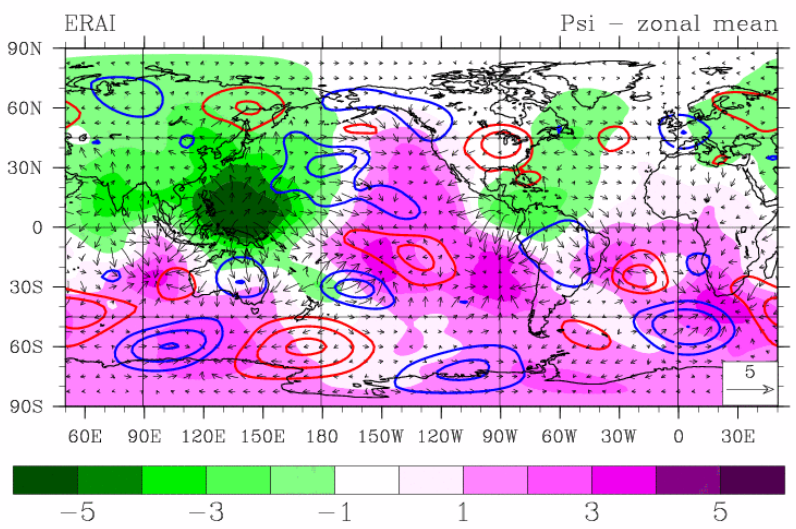
Analyse of the reaction of the atmosphere in the Tropics (direct and indirect action of SST) and beyond (especially via teleconnections to mid-latitudes)

Some periods where the predictability is :

« Good »



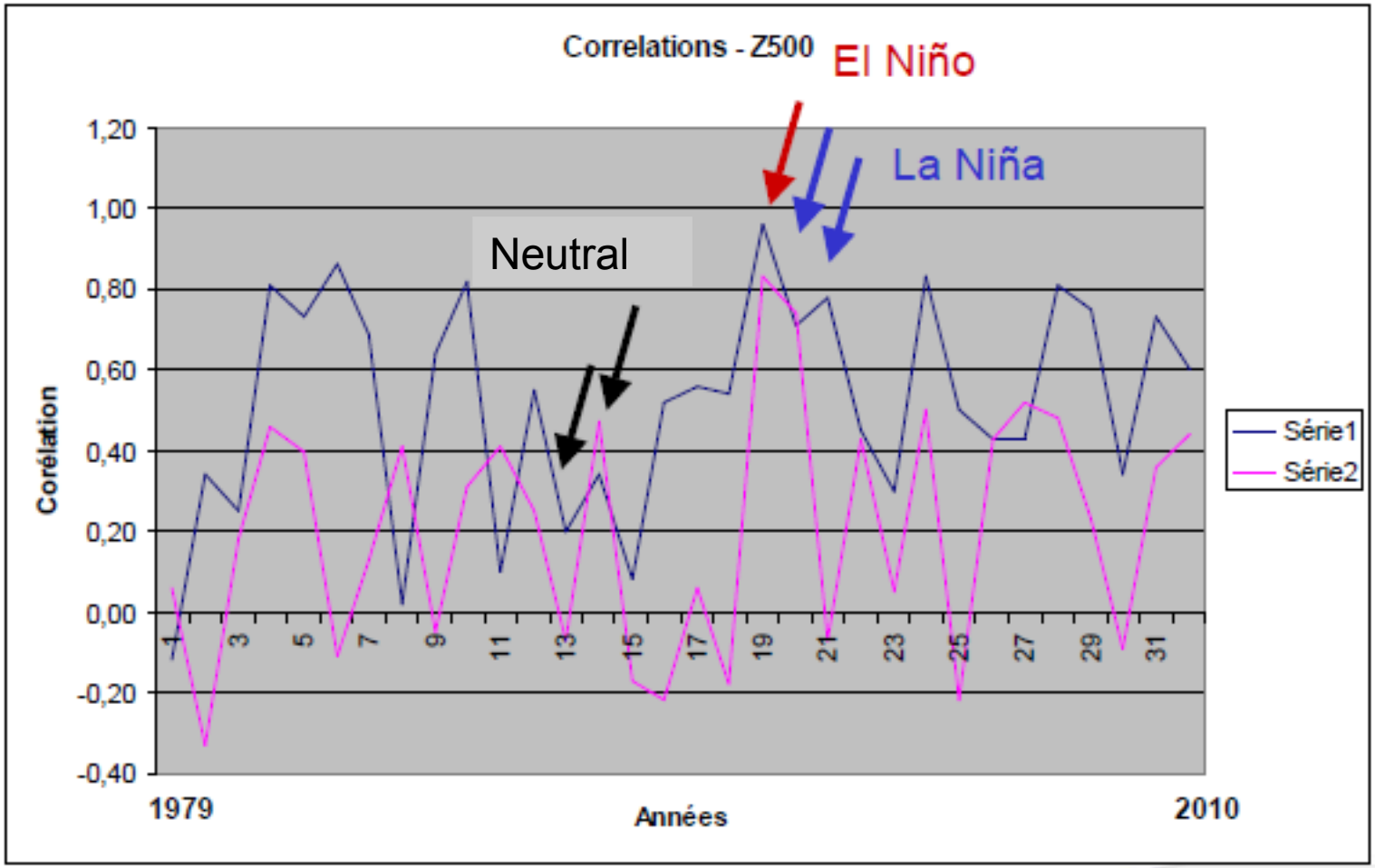
« Weak »



Observations (analyses) in the high troposphere of the components of the atmospheric circulation

Reliability and Skill

- Quality of the forecasts vs years (Geopotential Heigh)
  Winter season (DJF)



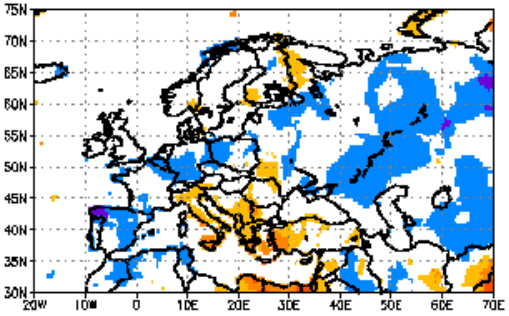
Tropics

(20°N,20°S)

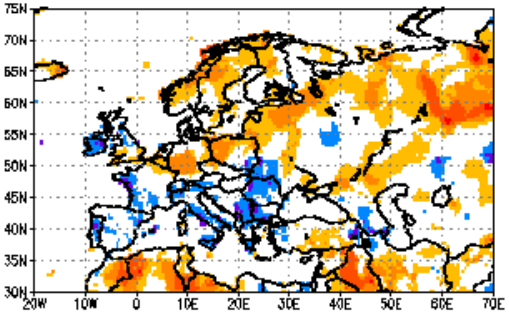
North Hemisphere

Temperature Probabilities for DJF
associated with La Nina (Min. 10 NINO3.4 SSTa D

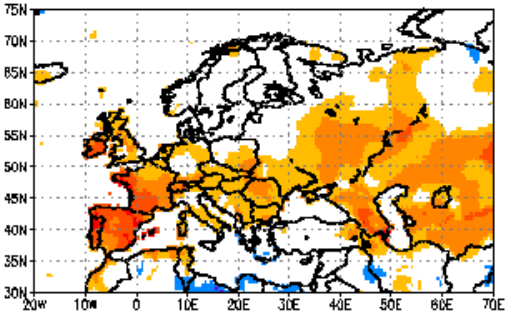
"ABOVE-NORMAL"

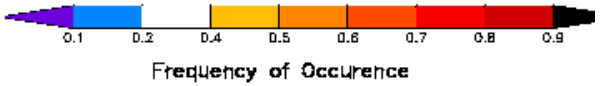


"NEAR-NORMAL"



"BELOW-NORMAL"

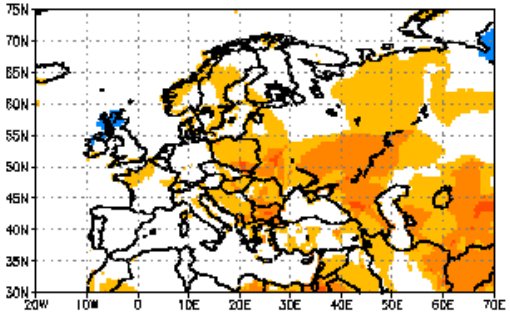




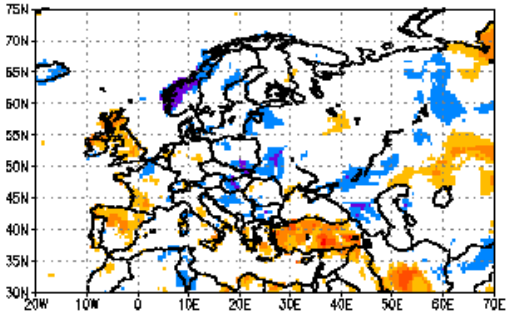
Frequency of Occurence

Temperature Probabilities for DJF
associated with El Nino (Max. 10 NINO3.4 SSTa DJF 1950-1995)

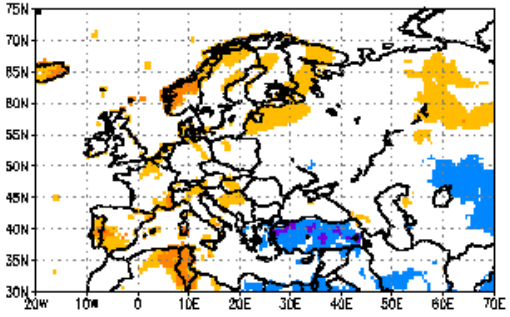
"ABOVE-NORMAL"

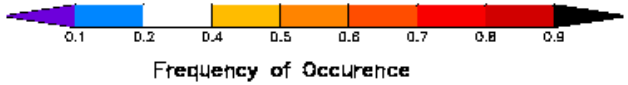


"NEAR-NORMAL"



"BELOW-NORMAL"





Frequency of Occurence

Cold NINO3.4 Yrs (incr. magnitude): 1963, 1985, 1965, 1951, 1955, 1956, Warm NINO3.4 Yrs (incr. magnitude): 1970 1991 1988 1969 1987 1966 1992 1973 1958 1983

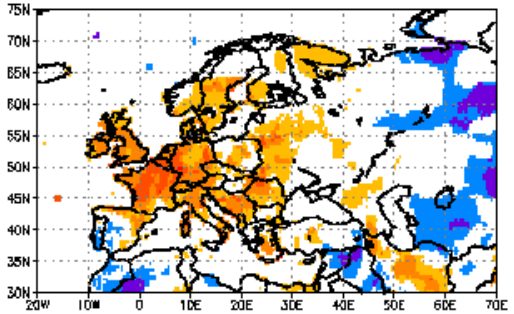
IRI International Research Institute

IRI International Research Institute

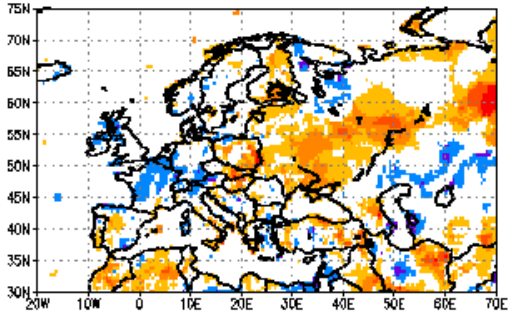
Temperature Probabilities for JJA

associated with La Nina (Min. 10 NINO3.4 SSTa JJ

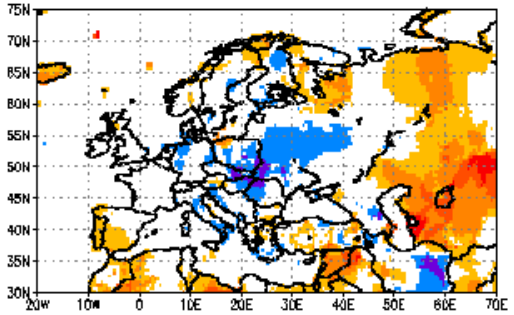
"ABOVE-NORMAL"



"NEAR-NORMAL"



"BELOW-NORMAL"

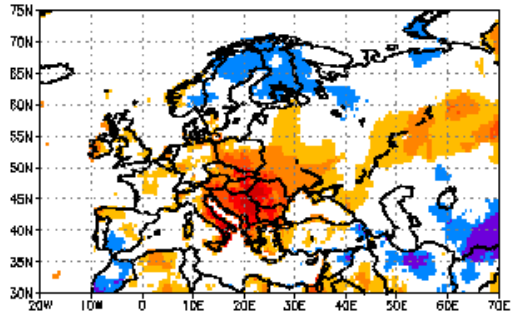


Cold NINO3.4 Yrs (incr. magnitude): 1971, 1970, 1950, 1954, 1964, 1956, 1

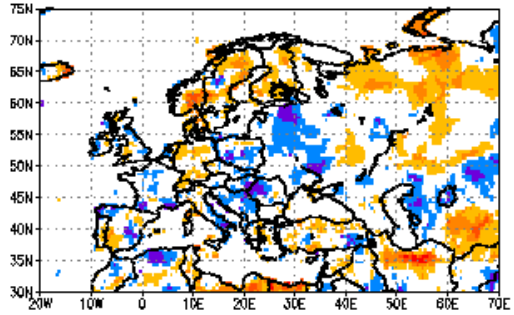
Temperature Probabilities for JJA

associated with El Nino (Max. 10 NINO3.4 SSTa JJA 1950-1995)

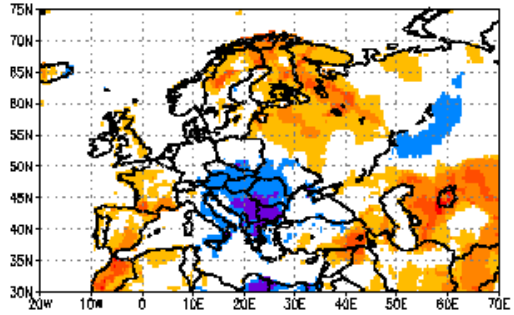
"ABOVE-NORMAL"



"NEAR-NORMAL"



"BELOW-NORMAL"



Warm NINO3.4 Yrs (incr. magnitude): 1963 1958 1992 1994 1957 1982 1965 1972 1991 1987

IRI International Research Institute

IRI International Research Institute

Preliminary question

How do we know that a forecast is « good » ?

- **The method of verification depends upon the nature of information provided by the forecast**
  - An event

    Forecast : It will rain tomorrow **Verification** : Yes / no.
  - A quantitative information (deterministic)

Forecast : There will be 5mm of rain tomorrow.

Verification : Calculate error in amount of rain..

  - A probabilistic information

Forecast : The probability of No significant rain will be 75% tomorrow.

Verification : Calculate error ?

Introduction

- How do we know if a probabilistic forecast was “correct”?

“A probabilistic forecast can never be wrong!”

As soon as a forecast is expressed probabilistically, all possible outcomes are forecasted. However, the forecaster’s level of confidence can be “correct” or “incorrect” = **reliable**.

Is the forecaster **over-** / **under-confident**?

Whenever a forecaster says there is a high probability of rain tomorrow, it should rain more frequently than when the forecaster says there is a low probability of rain (see reliability diagrams).

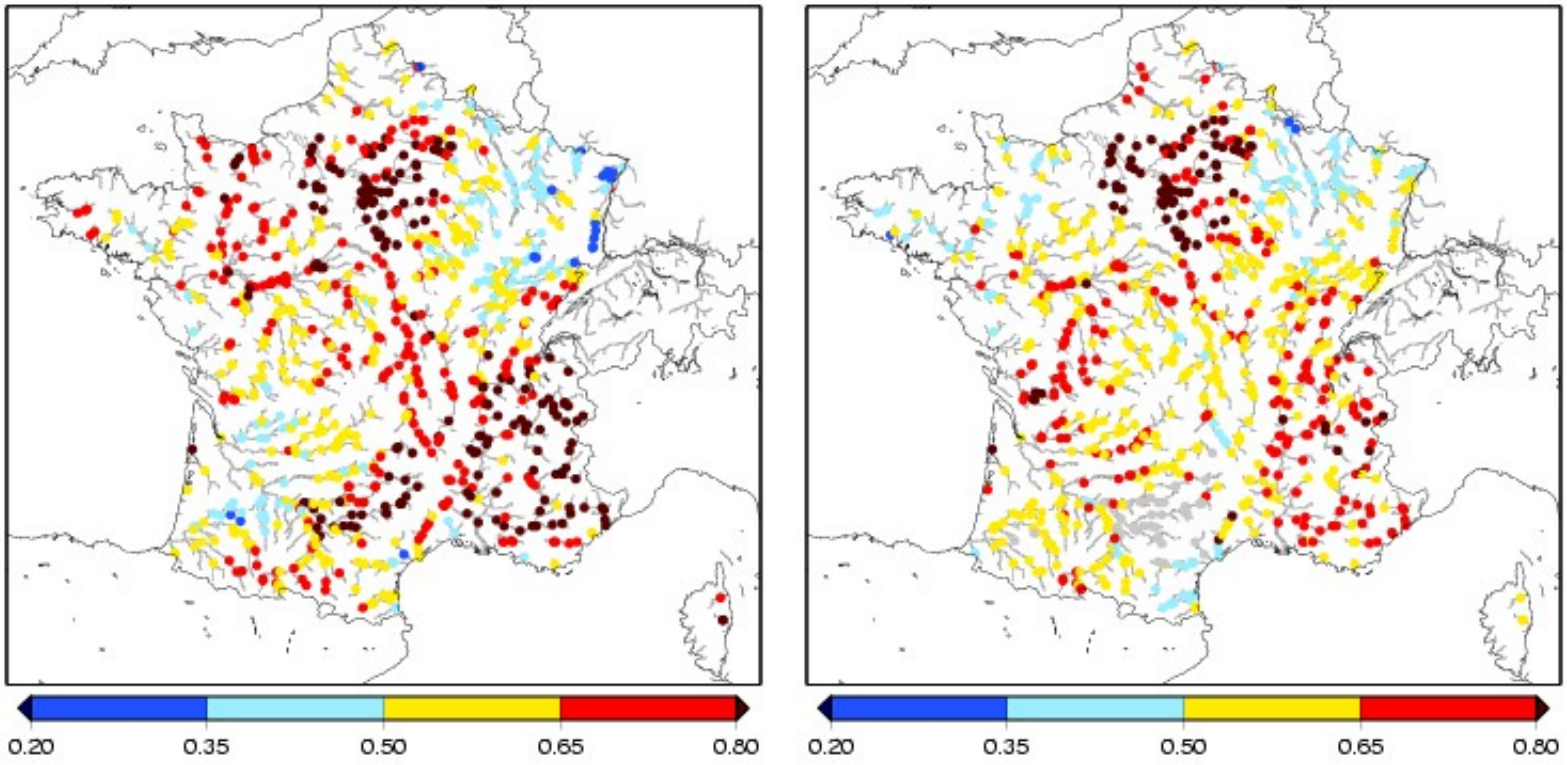
Preliminary question

How do we know that a forecast is « good » ?

- **In case of Impact Forecast** (tailored e.g. for DMP)
  - **Verification ?**
  - Depends on the usefulness for the user
  - Needs of reference dataset from the user side (Impacts, Decisions, …)
  - Verification of the use and better decision still to be developped (e.g. Placebo protocol). **The problem is more complex !**

Other component of the climate system

ROC scores for Hydro-SF (1979-2007 – IC from 1st of April)



Upper Tercile

Lower Tercile

Skills can significantly better for River Flow and SWI than for Temperature and Rainfall

(Ref : Singla et al. 2012)

Preliminary question

How do we know that a forecast is « good » ?

- **What makes a « good » forecast ?**
  - Quality : the outcome should correspond with the forecast
  - Timeliness : the forecast must be issued early enough for response
  - Uncertainty : the forecast must be about something that was not inevitable
  - Salience : the forecast must be about something of interest (including timeliness)
  - No ambiguity : the precise meaning of the forecast should not be subject to interpretation
  - Consistency : the forecast should indicate what the forecaster believes will happen

Preliminary question

How do we know that a forecast is « good » ?

- **What make a « good » forecast ?**
  - Quality

    Good forecast should corresponds to the outcome
    - Examples :

    Obama will win the US election President for the second time

    ONDJ1997 rainfall over Nairobi will be above-normal
    - Note that correctness is only one aspect of the quality.
  - Timeliness

    Good Forecast should be issued at a relevant time for use
    - Examples :

    F. Hollande will win the 2012 French President election (the day of election at 20:01 when announcemed on the TVs)

    October SWIOCOF forecasts are too late for Tourism sector

Preliminary question

How do we know that a forecast is « good » ?

- **What make a « good » forecast ?**
  - Uncertainty

    Good forecast should address something uncertain
    - Examples : JPC will win the French President Election next year

      It will rain less than 2000 mm in Niamey next rainy season
  - Salience

    Good Forecast should target something of interest
    - Examples : There will be some Orchids in my garden next September (Who cares?)

      The T500 hPa anomaly of November 2016 over France will be 1°C more than 20 years ago
    - Often the relevancy of a forecast is not obvious because of the way the forecast is presented

Preliminary question

How to score a « good » forecast ?

- **Properties of scoring rules**
  - Equitability : **Equitable scores** must score all « unskilful » forecasts equally badly
  - Propriety : **Proper scores** are those that are optimized when the forecaster forecasts what (s)he thinks is the correct forecast.
    - If the score is not proper, the forecaster can cheat or hedge (issue a different forecast to get a better score).
  - Effectiveness : **An effective score** must give a better score to a “better” forecast.

Equitability

Near-misses

	FORECASTS		
OBSERVATIONS	**B**	**N**	**A**
A	**-1.0**	**0.0**	**1.0**
N	**0.0**	**1.0**	**0.0**
B	**1.0**	**0.0**	**-1.0**

Exercise: Is the scoring table above a good idea?

Gerrity Score

	FORECASTS		
OBSERVATIONS	B	N	A
A	-1.00	-0.25	1.25
N	-0.25	0.50	-0.25
B	1.25	-0.25	-1.00

This solution has some simpler properties.

Propriety

How many of the events were forecast?

$$\text{Hit rate} = \frac{\text{number of hits}}{\text{number of events}} \times 100\%$$

A score of 100% can be guaranteed by always forecasting an event!

False alarms are incorrect forecasts.

Equitability v propriety.

Preliminary question

How to score a « good » forecast ?

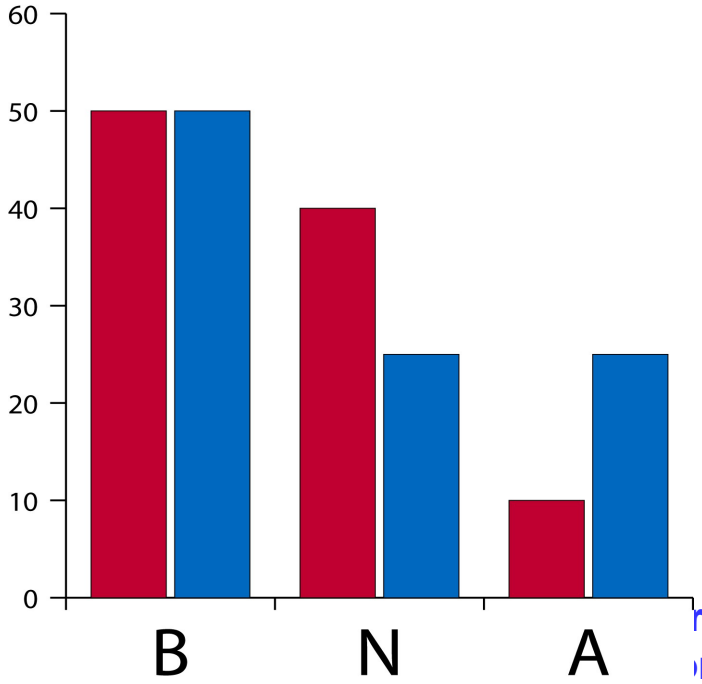
- **Properties of scoring rules**
  - Consideration of distances : **A score which consider distances** should credit forecasts that issue high probabilities for values close to the verification.
  - Understandability : It is essential to **define exactly what is the purpose of the verification** analysis so as to choose an appropriate score.
  - Locality : **A score which consider locality** must only score the forecast on the basis of the probability assigned to the verification.
    - The property of locality is inconsistent with the consideration of distance.

Effectiveness

Consider the probability score (the average squared probability error over all categories):

$$S = \frac{1}{m}\sum_{j=1}^{m}\left(v_j - p_j\right)^2$$

Red: {50%, 40%, 10%}
Blue: {50%, 25%, 25%}



If B occurs, by most considerations the red forecast is as good as or better than the blue, but, the scores are:

Red: 0.140
Blue: 0.125

Linear Error in Probability Space (LEPS)

	FORECASTS		
OBSERVATIONS	B	N	A
A	-0.78	-0.11	0.89
N	-0.11	0.22	0.11
B	0.89	-0.11	-0.78

These weights are defined to ensure that forecasts of climatology AND perpetual forecasts of one category AND random guessing have an expected score of zero.

Preliminary question

How do score a « good » forecast ?

- **What make a « good » forecast ?**
  - Ambiguity

    Good forecast should not be ambiguous (not subject to interpretation)
    - Examples : France will do well for the next Rugby World Cup

      SEE region will have a good winter season
  - Consistency

    Good Forecast should be consistent with the believes of the forecaster
    - Examples : P Ryan will win the 2012 US Vice-President.

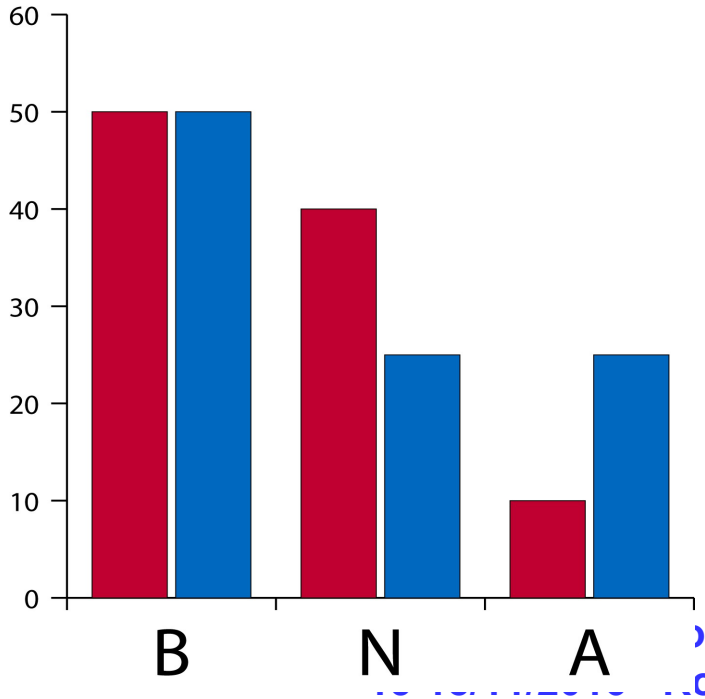
      The next OND 1997 rainy season in Ruiru (Kenya) will be close to Normal (Reluctance to forecast high probabilities of Above-normal rainfall?)

Distance

The ranked probability score resolves the lack of effectiveness of the probability score:

$$S = \frac{1}{m}\sum_{j=1}^{m}\left(V_j - P_j\right)^2$$

Red: {50%, 40%, 10%}
Blue: {50%, 25%, 25%}



If B occurs, the scores are:

Red: 0.087
Blue: 0.104

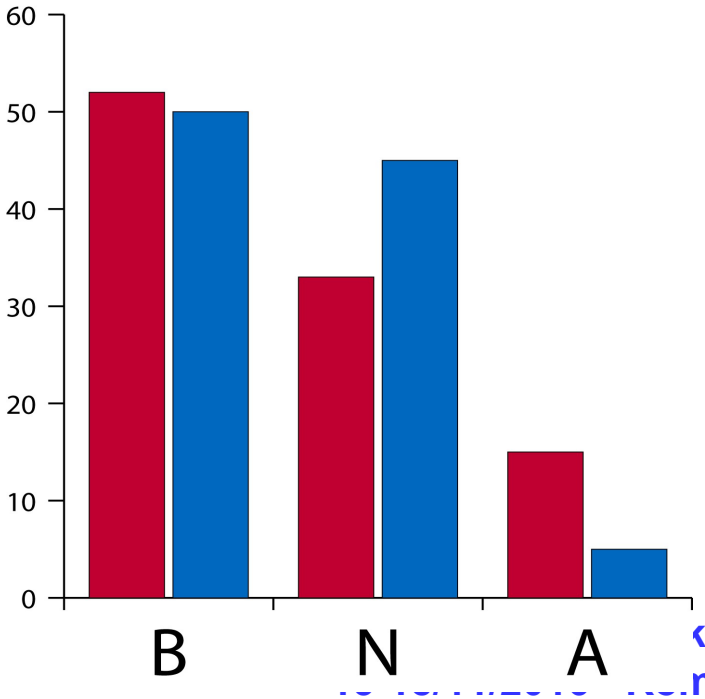
Locality

Considering distance does not necessarily give the best score to the forecast with the highest probability on the verifying category:

Red: {52%, 33%, 15%}
Blue: {50%, 45%, 5%}

If B occurs, the scores are:

Red: 0.0843
Blue: 0.0842



Interpretation of scores

$$MSSS = 1 - \frac{MSE_F}{MSE_c}$$

where : MSE=Mean Squared Error

F for Forecasts

c for Climatology

Murphy decomposition :

$$MSSS = \left\{ 2\frac{\text{var}_F}{\text{var}_o} cor_{Fo} - \left(\frac{\text{var}_F}{\text{var}_o}\right)^2 - \left(\frac{\left[mean_F - mean_o\right]^2}{\text{var}_o} + \frac{2n-1}{(n-1)^2} \right) \right\} \left\{ 1 + \frac{2n-1}{(n-1)^2} \right\}$$

Phase error

Amplitude error

Systematic error

avec : **var** = variance

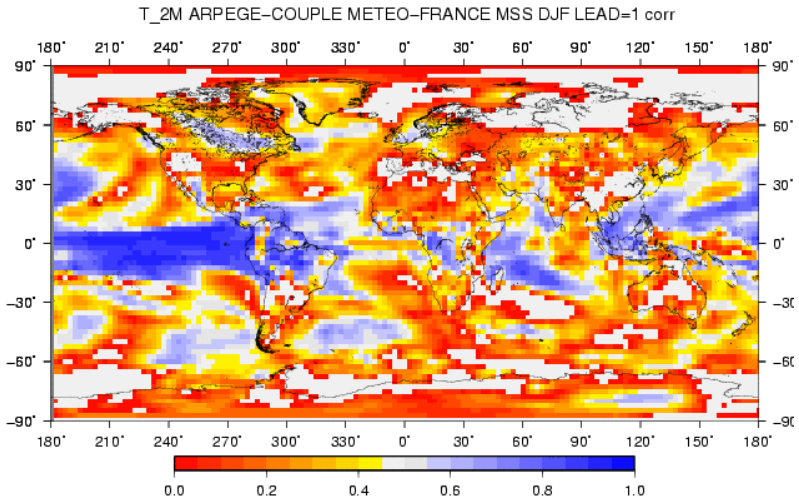
mean = mean

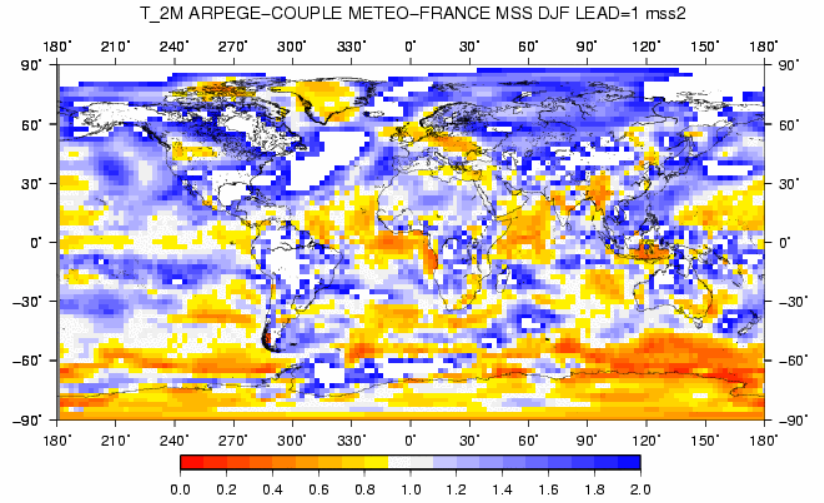
cor = correlation

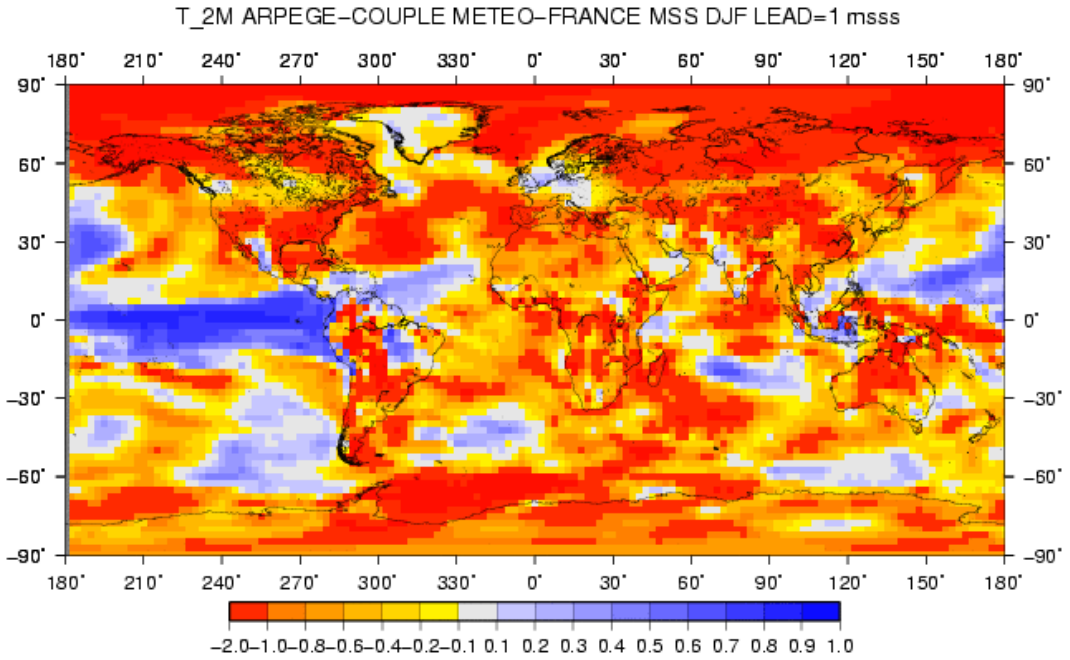
n = size of sampling

o for observation

MSSS


T_2M ARPEGE–COUPLE METEO–FRANCE MSS DJF LEAD=1 corr
0.0 0.2 0.4 0.6 0.8 1.0


T_2M ARPEGE–COUPLE METEO–FRANCE MSS DJF LEAD=1 mss2
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0


T_2M ARPEGE–COUPLE METEO–FRANCE MSS DJF LEAD=1 msss
–2.0 –1.0 –0.8 –0.6 –0.4 –0.2 –0.1 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

Interpretation of scores

Brier score (*Murphy's decomposition*)

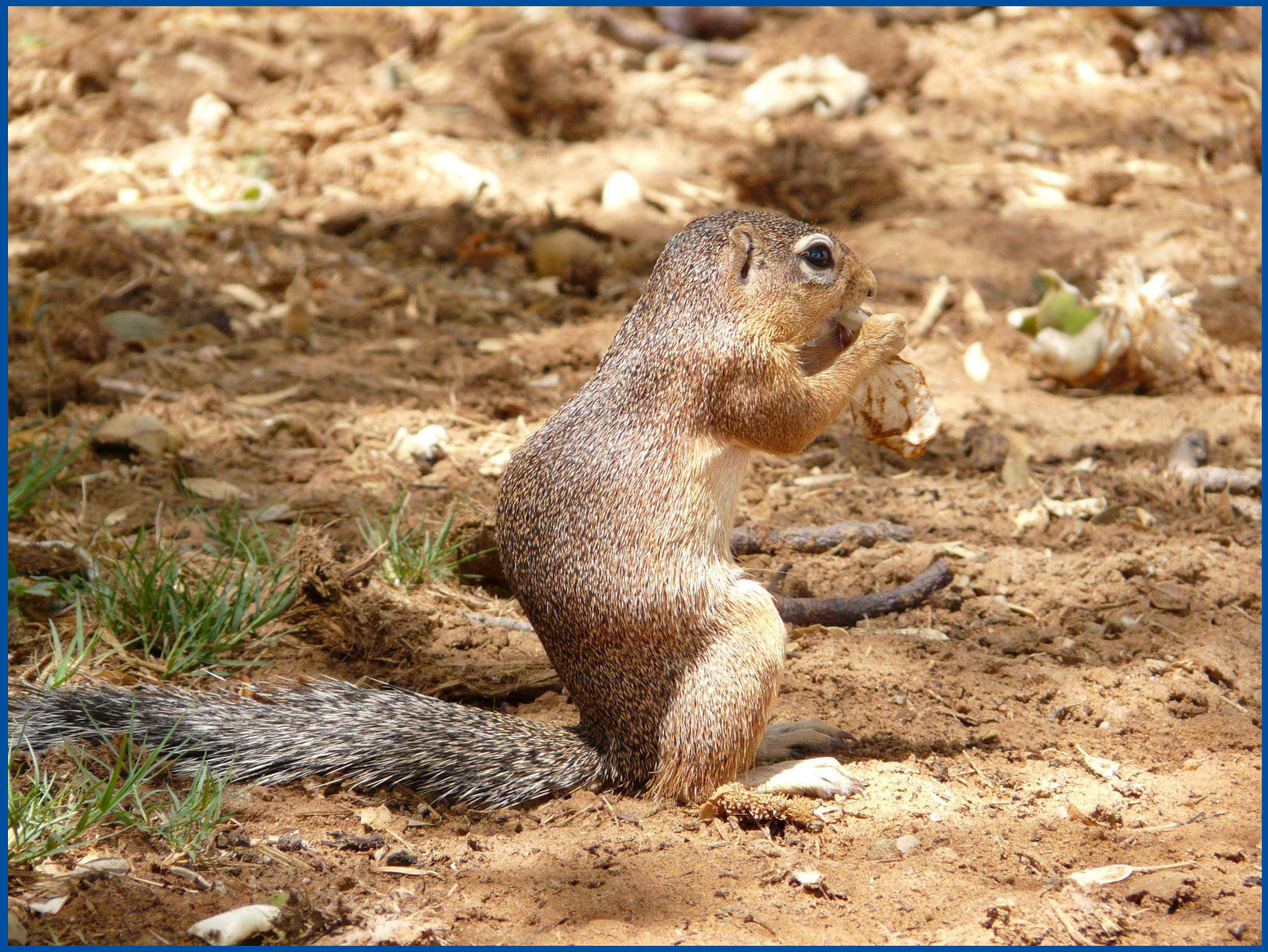
Brier score = reliability – resolution + uncertainty

Resolution : when the forecast is 60% for dry, is the outcome the same as when the forecast is 10% for dry?

Reliability : when the forecast is 60% for dry, do dry conditions occur 60% of the time?

Uncertainty : what is the climatological probability of dry conditions occurring?

WEATHER CLIMATE WATER
TEMPS CLIMAT EAU

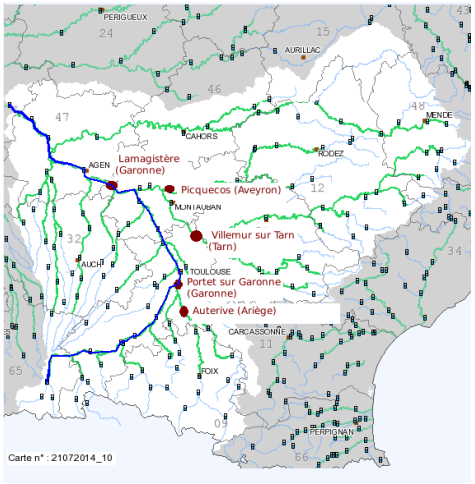


WMO OMM

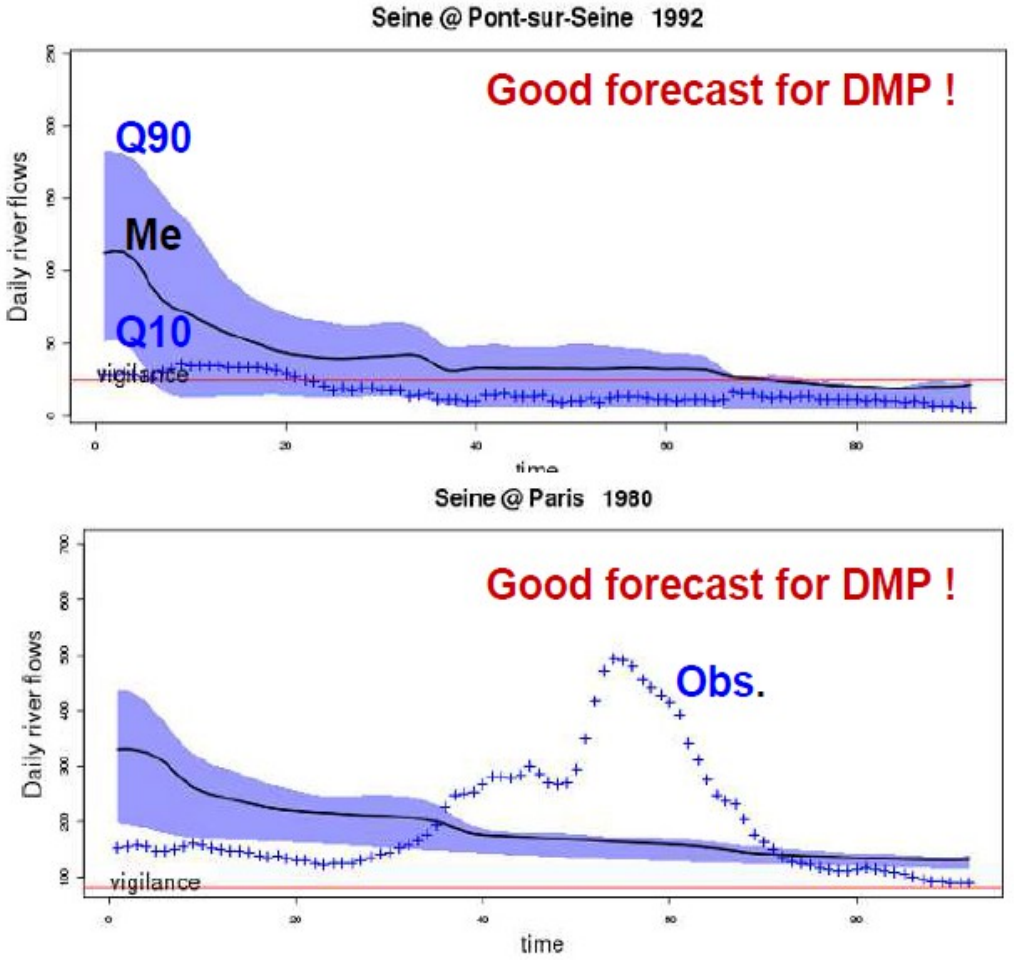
World Meteorological Organization
Organisation météorologique mondiale

Verification for tailored information

- Some examples



Key Stations used by the SMEAG



Forecast - Daily Time Series of ensemble Median, Q10 and Q90

Interpretation of scores

Brier score

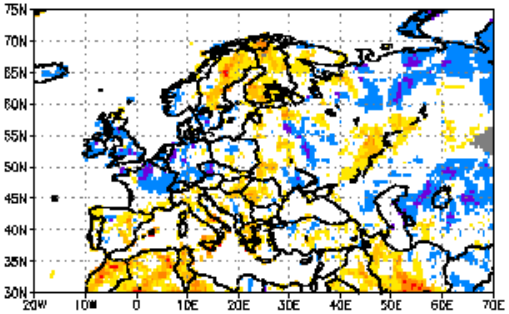
Measures the mean-squared error of probability forecasts *(equivalent of MSE for deterministic forecast)*.

$$\text{Brier score} = \frac{\text{total of squared probability errors}}{\text{number of forecasts}}$$

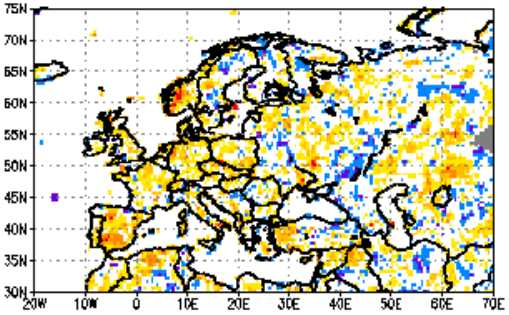
If an event was forecast with a probability of 60%, and the event occurred, the probability error is:
60% - 100% = -40% and BS contribution is 0.16

Precipitation Probabilities for DJF
associated with La Nina (Min. 10 NINO3.4 SSTa I

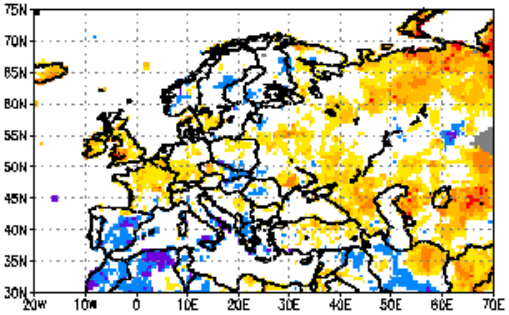
"ABOVE-NORMAL"

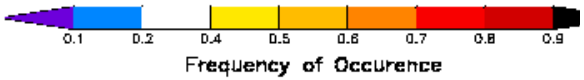


"NEAR-NORMAL"



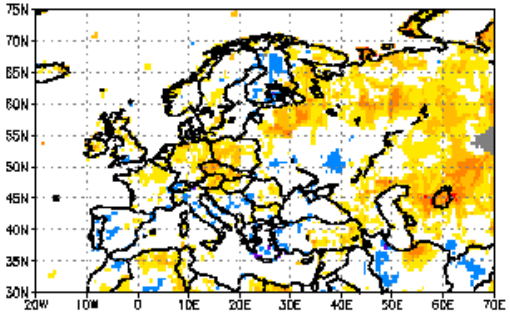
"BELOW-NORMAL"



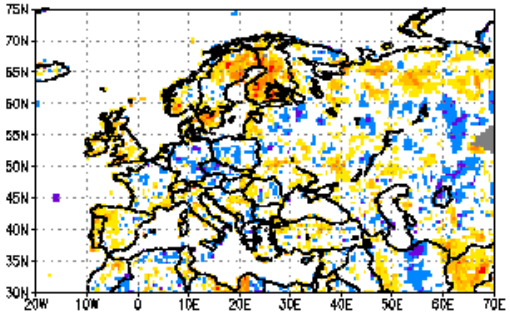


Precipitation Probabilities for DJF
associated with El Nino (Max. 10 NINO3.4 SSTa DJF 1950-1995)

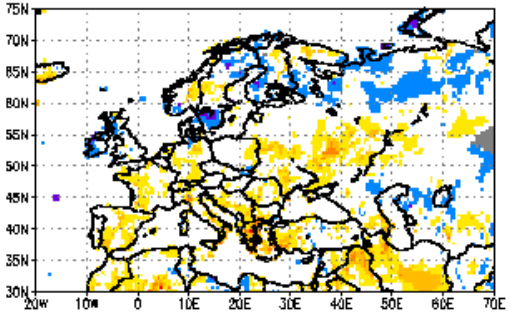
"ABOVE-NORMAL"

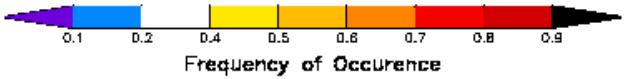


"NEAR-NORMAL"



"BELOW-NORMAL"





GREY areas indicate dry season (seasonal avg. <5cm & GREY areas indicate dry season (seasonal avg. <5cm & <15% annual avg.)

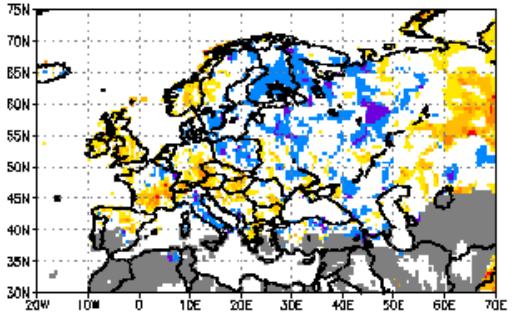
Cold NINO3.4 Yrs (incr. magnitude): 1963, 1985, 1965, 1951, 1955, 1956, Warm NINO3.4 Yrs (incr. magnitude): 1991 1988 1969 1987 1995 1966 1992 1973 1958 1983

IRI International Research Institute

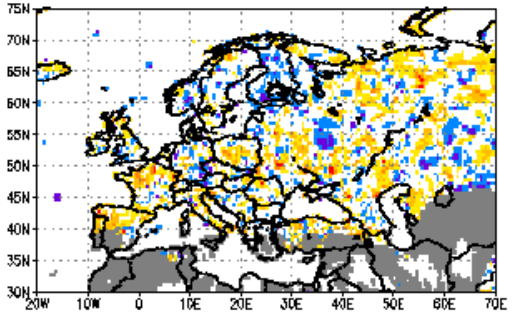
Precipitation Probabilities for JJA

associated with La Nina (Min. 10 NINO3.4 SSTa Ju

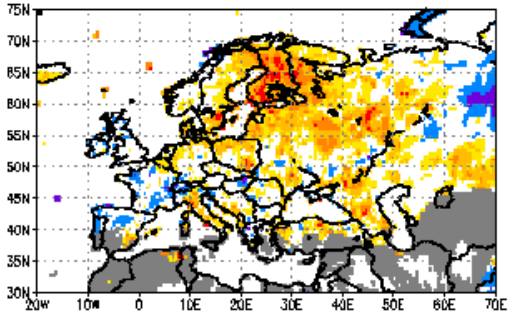
"ABOVE-NORMAL"



"NEAR-NORMAL"



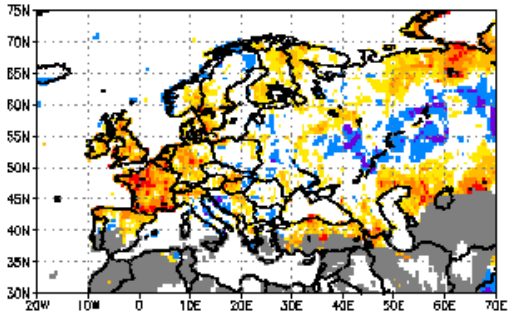
"BELOW-NORMAL"



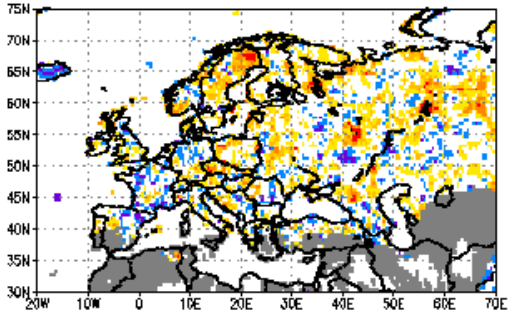
Precipitation Probabilities for JJA

associated with El Nino (Max. 10 NINO3.4 SSTa JJA 1950-1995)

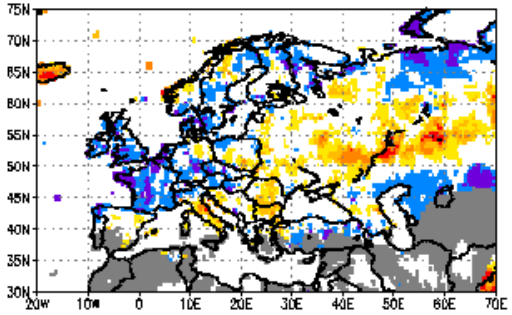
"ABOVE-NORMAL"

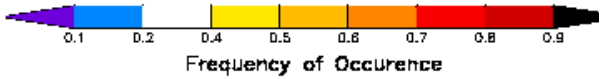


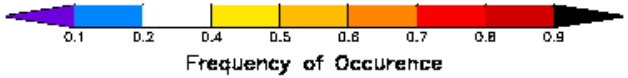
"NEAR-NORMAL"



"BELOW-NORMAL"







GREY areas indicate dry season (seasonal avg. <5cm & <15% annual avg.)

GREY areas indicate dry season (seasonal avg. <5cm & <15% annual avg.)

Cold NINO3.4 Yrs (incr. magnitude): 1971, 1970, 1950, 1954, 1964, 1956, 1 Warm NINO3.4 Yrs (incr. magnitude): 1963 1958 1992 1994 1957 1982 1965 1972 1991 1987

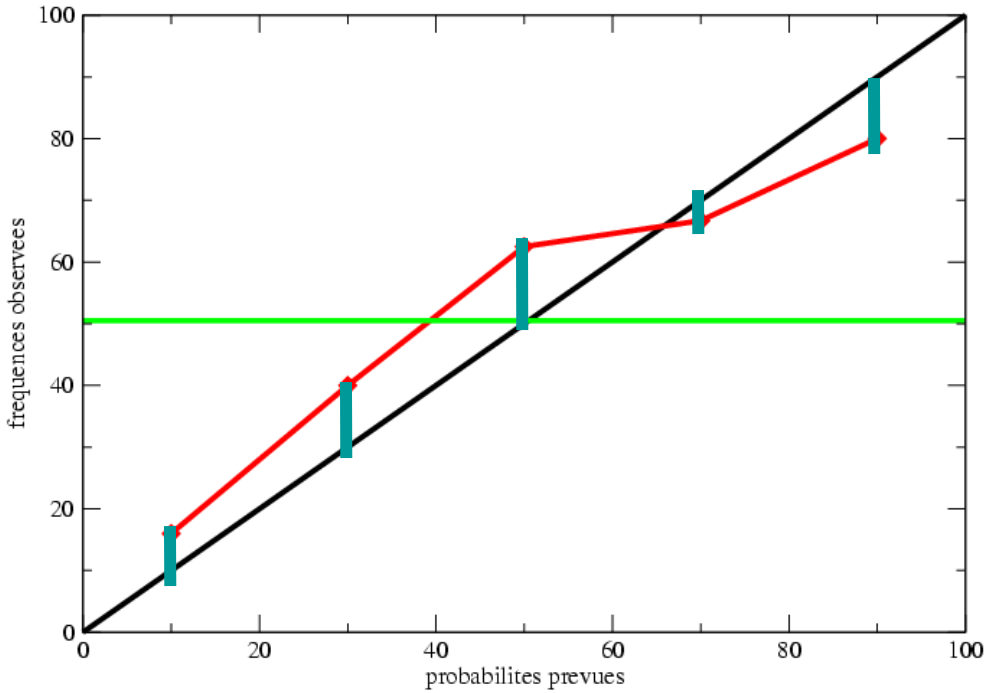
Cold NINO3.4 Yrs (incr. magnitude): 1971, 1970, 1950, 1954, 1964, 1956, 1 Warm NINO3.4 Yrs (incr. magnitude): 1963 1958 1992 1994 1957 1982 1965 1972 1991 1987

IRI International Research Institute

IRI International Research Institute

Reliability

$$reliability = \frac{1}{N}\sum_{i=1}^{I} n_i (p_i - o_i)^2$$

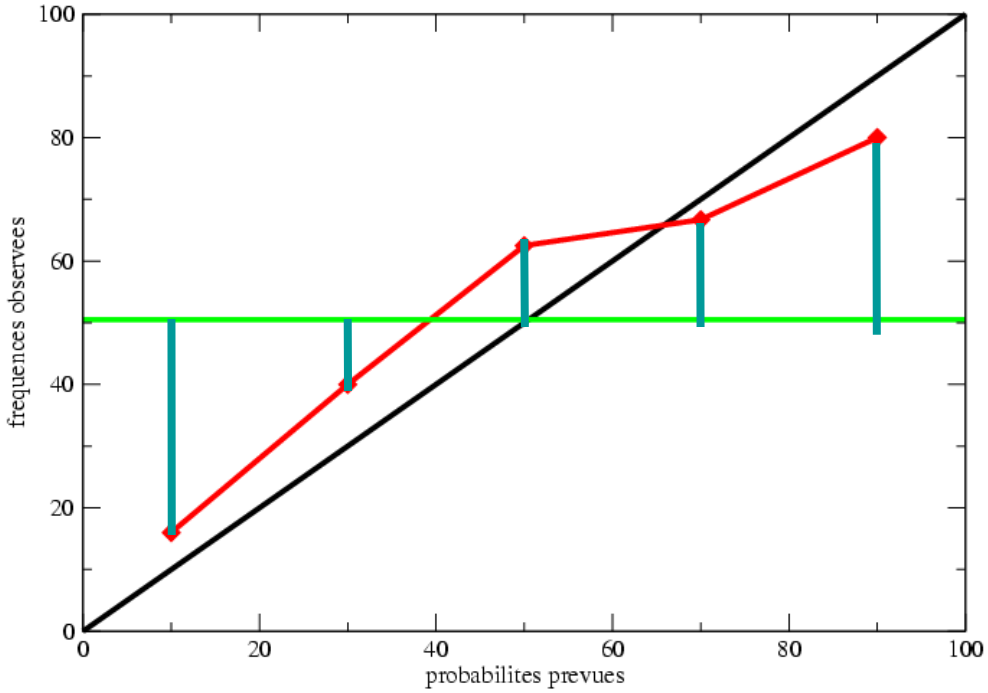


c

The lower the reliabity, the better it is.

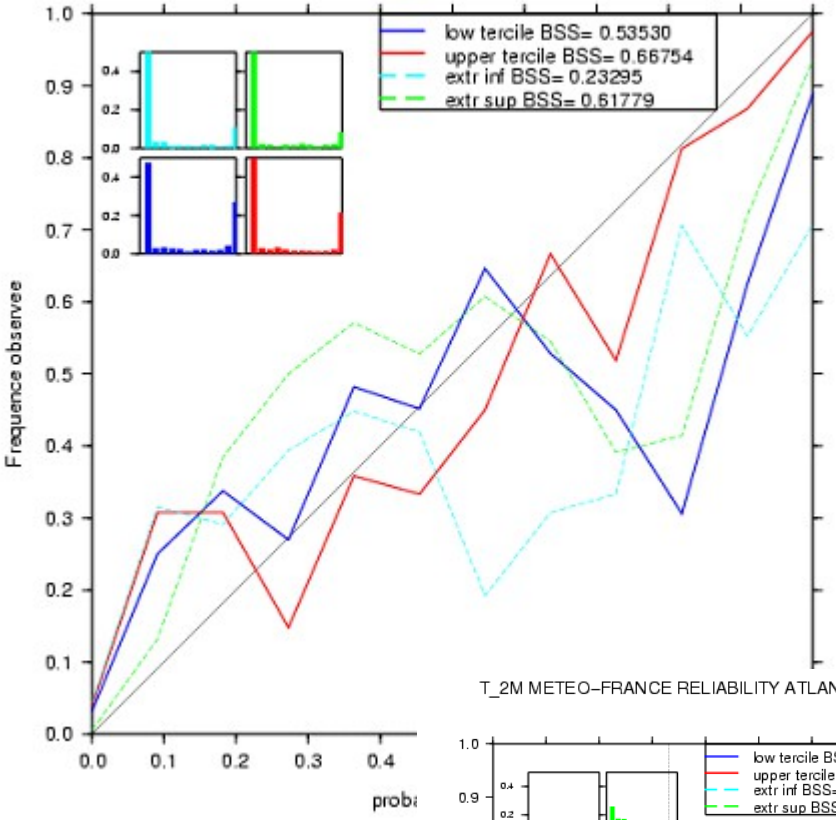
Resolution

$$resolution = \frac{1}{N}\sum_{i=1}^{I} n_i (o_i - c)^2$$

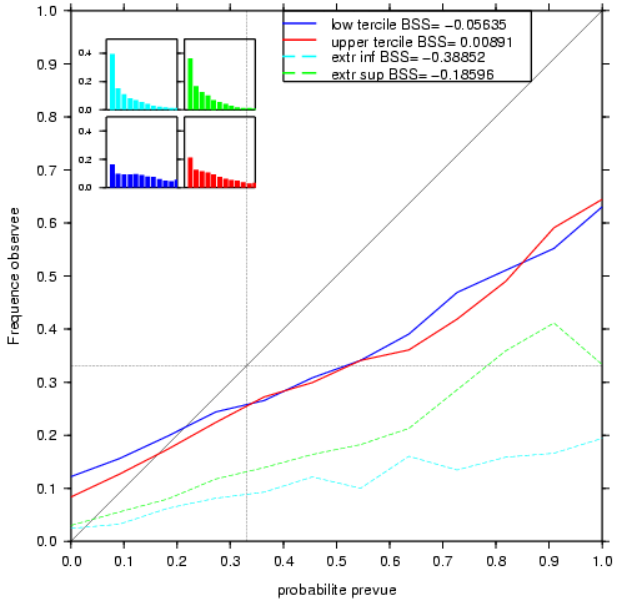


The greater the resolution, the better it is

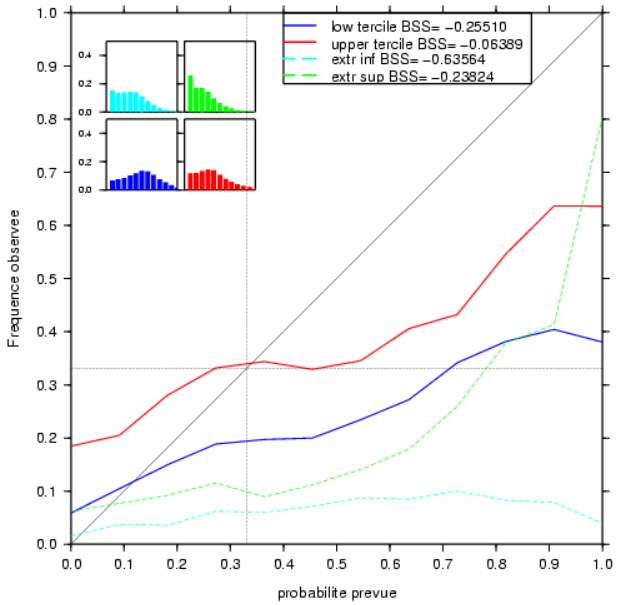
TSOL METEO-FRANCE RELIABILITY NINO3.4 DJF LEAD=1



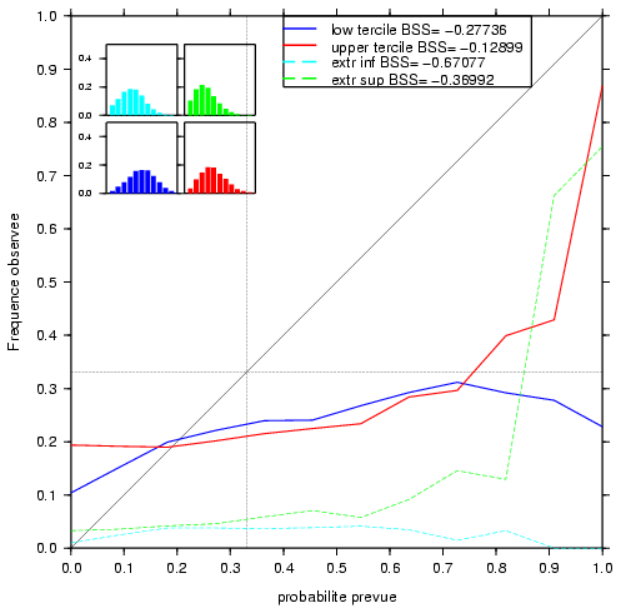
PRET METEO-FRANCE RELIABILITY PACIFIQUE TROPICAL JJA LEAD=1



T_2M METEO-FRANCE RELIABILITY ATLANTIQUE NAO DJF LEAD=1



PRET METEO-FRANCE RELIABILITY ATLANTIQUE NAO DJF LEAD=1



WMO OMM

Measures of Reliability and Sharpness

Ranked probability score

The same as the Brier score, but for multiple categories.

The Brier score and the ranked probability score can be expressed as **skill scores** in the same way as for the Heidke (hit) score.

Verification measures for continuous probabilistic forecasts are experimental – there are very few attempts to estimate the full probability distribution of possible outcomes.