Tendency diagram, linear probability score and hit score

Caio Coelho

caio.coelho@cptec.inpe.br

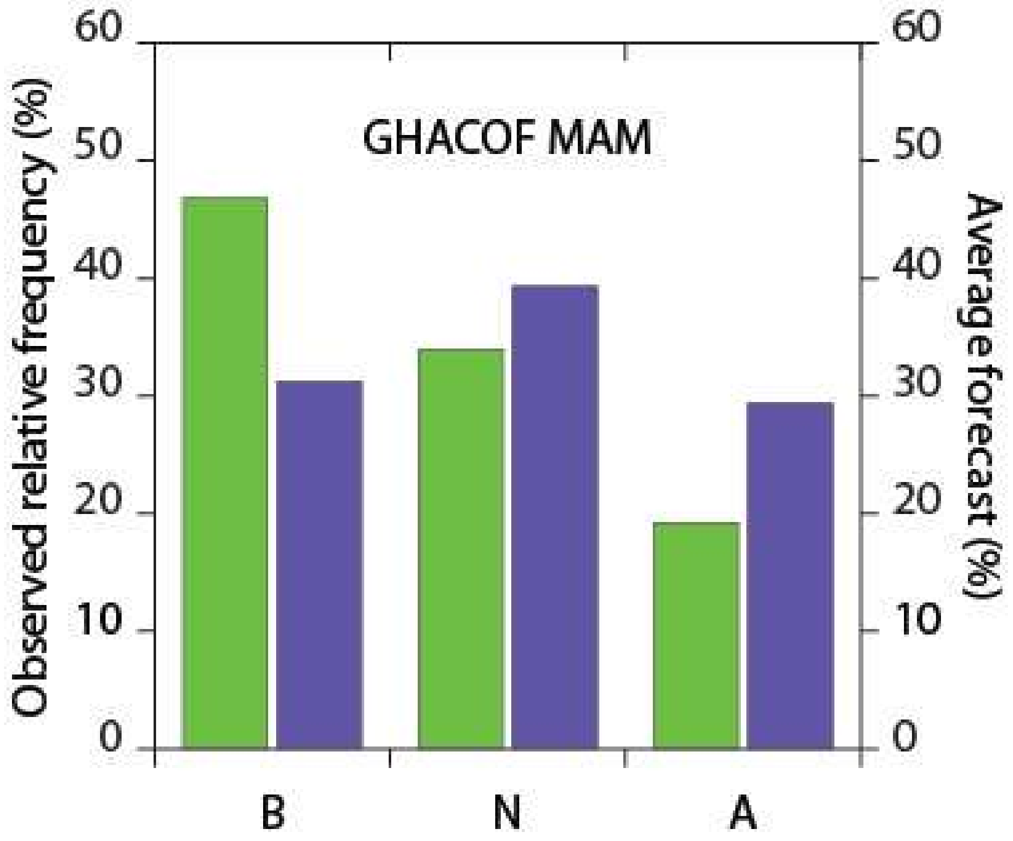
CPTEC/INPE

Mostly based on: Guidance on Verification of Operational Seasonal Climate Forecasts By Simon J. Mason (IRI), prepared under the auspices of the World Meteorological Organization (WMO), Commission for Climatology XIV, Expert Team on CLIPS Operations, Verification and Application Service

MedCOF Training Workshop on Verification of Operational
Seasonal Forecasts in the Mediterranean region
Rome, Italy, 15-18 November 2016

Tendency diagram for diagnosing unconditional bias/consistency

- Calculating the obs frequency of each category over the verification period gives a simple indication as to whether the period has been unusual. Any **shift** may or may not be permanent.
- Are probabilities **consistently** too high or too low?



Green bar: Observed
Blue bar: Forecast

Courtesy: Simon Mason (IRI)

Tendency diagram

- Useful to illustrate unconditional biases in the forecasts, and particularly to demonstrate hedging of normal category
- Can be plotted for each station or for aggregated station averages
- Hedging on normal category is an indication of lack of consistency, i.e. forecasters do not necessarily forecast what they really think, but instead play safe and issue conservative forecasts putting highest probability in the normal category because they know that by doing this if the forecast does not verify, it will at least be just one category away from the observed category

Linear probability score (LPS)

$$\text{LPS} = 100\% \times \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{j,i}\,p_{j,i}$$

(Wilson et al, 1999)

Option for communicating forecast quality of individual maps for non-specialists

n: number of points (locations i) on the map at which the forecasts are to be verified

m: number of categories j

$y_{i,j}$: is 1 if the observation at location i was in category j, and is 0 otherwise

$p_{i,j}$: is the forecast probability for category j at location i

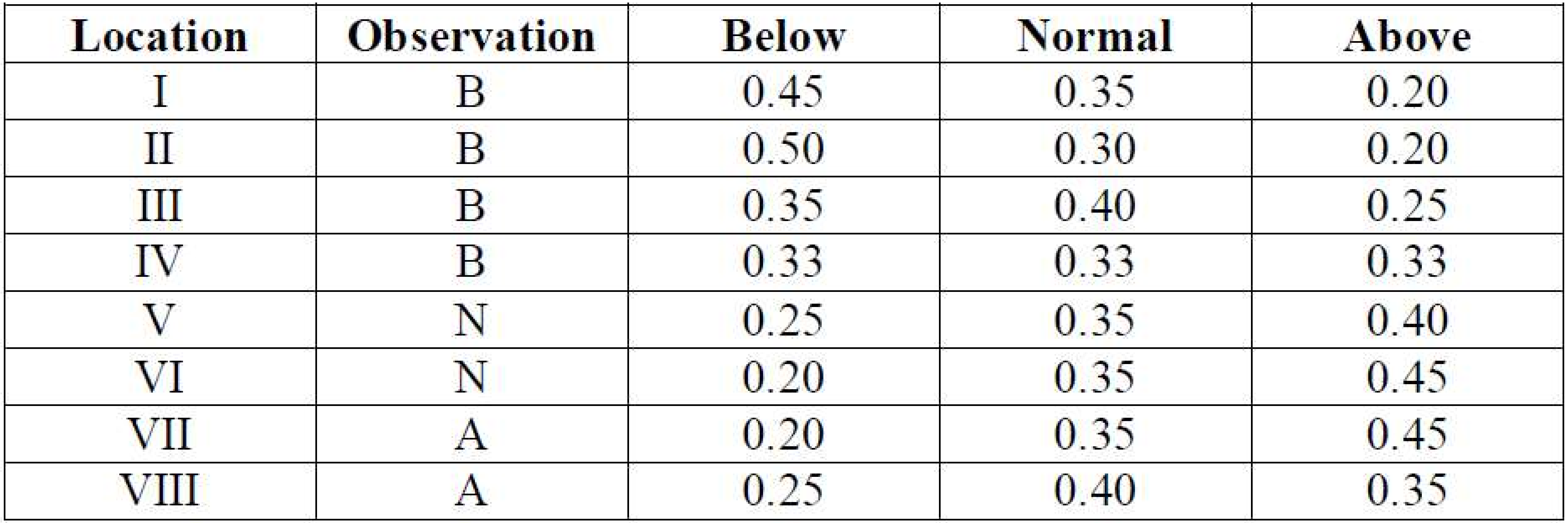
Interpretation: Average forecast probability assigned to the verifying categories

Range: from 100% for perfect forecasts (100% probability assigned to the observed category at each of the locations) to 0% for perfectly bad forecasts (0% probability assigned to the observed category at each of the locations)

A “good” forecast will score more than a strategy of using the climatological forecasts, and will beat the expected score from guessing. **Question: What is this expected score from guessing for 3 category forecasts?**

Example forecasts and observations for 3 equi-probable categories [below-normal (B), normal (N), and above-normal (A)]

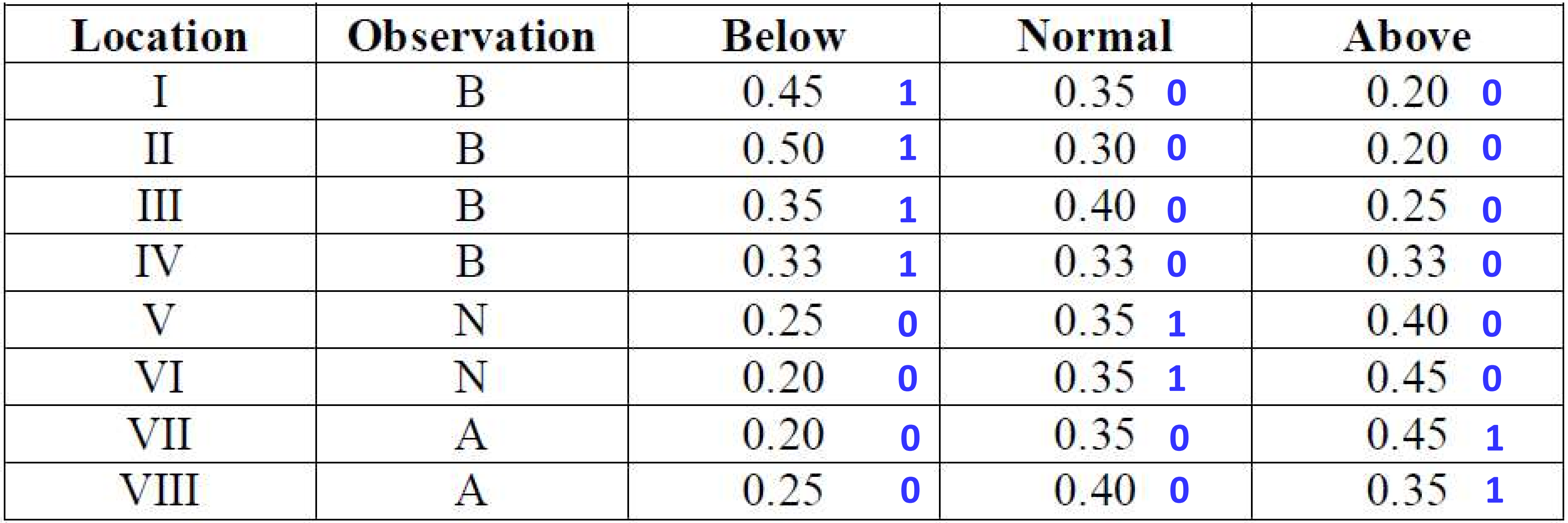
Forecast probabilities for the 3 categories



Location	Observation	Below	Normal	Above
I	B	0.45	0.35	0.20
II	B	0.50	0.30	0.20
III	B	0.35	0.40	0.25
IV	B	0.33	0.33	0.33
V	N	0.25	0.35	0.40
VI	N	0.20	0.35	0.45
VII	A	0.20	0.35	0.45
VIII	A	0.25	0.40	0.35

Example forecasts and observations for 3 equi-probable categories [below-normal (B), normal (N), and above-normal (A)]

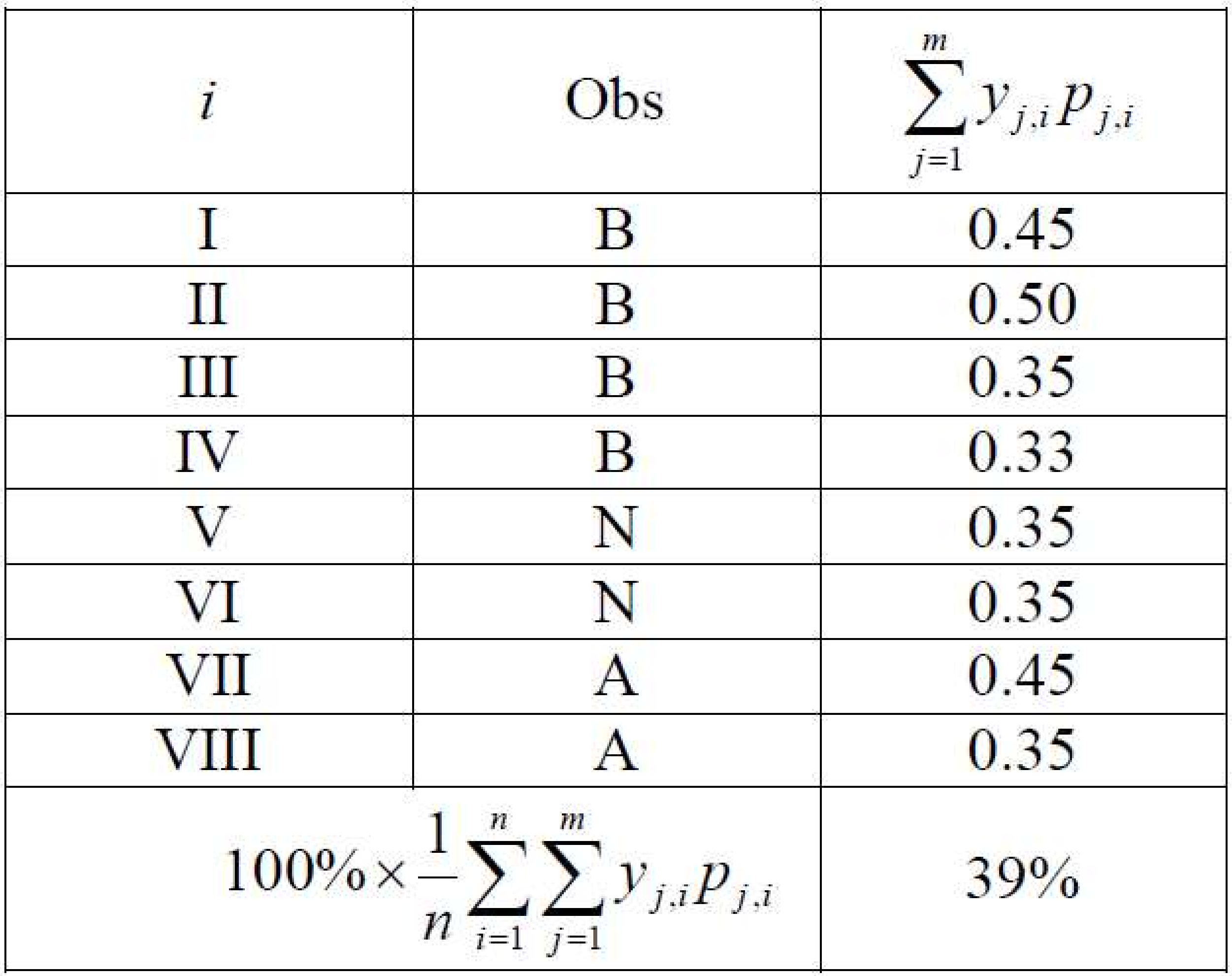
Forecast probabilities for the 3 categories



Location	Observation	Below	Normal	Above
I	B	0.45 1	0.35 0	0.20 0
II	B	0.50 1	0.30 0	0.20 0
III	B	0.35 1	0.40 0	0.25 0
IV	B	0.33 1	0.33 0	0.33 0
V	N	0.25 0	0.35 1	0.40 0
VI	N	0.20 0	0.35 1	0.45 0
VII	A	0.20 0	0.35 0	0.45 1
VIII	A	0.25 0	0.40 0	0.35 1

Blue numbers:
1 indicates the observed category
0 indicates non-observed category

LPS example for tercile probability forecasts



i	Obs	$\sum_{j=1}^{m} y_{j,i} p_{j,i}$
I	B	0.45
II	B	0.50
III	B	0.35
IV	B	0.33
V	N	0.35
VI	N	0.35
VII	A	0.45
VIII	A	0.35
$100\% \times \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{j,i} p_{j,i}$		39%

Average probabiliy about 6% greater than climatology

Data files for practical session

- File GHACOF_SOND_observations.txt contains precipitation observations for 10 stations (columns 2 to 11) covering the period 1961-2012 (52 years)
- Will use 1961-1990 period (30 years) to define climatology and compute tercile boundaries

- Files bforecastsGHACOF1.csv
  nforecastsGHACOF1.csv
  aforecastsGHACOF1.csv

contain forecast probabilities for below normal, normal and above normal categories for the same 10 stations (columns 2 to 11) covering the period 1998-2007 (10 years)

Hits

- How often des the category with the highest probability verify? But with more than 2 categories we like to give credit for “near-misses”.
- Implicitly or explicitly, we often use the following table to score the forecasts.

	FORECASTS		
OBSERVATIONS	**Above**	**Normal**	**Below**
Above-normal	1.0	0.0	-1.0
Normal	0.0	1.0	0.0
Below-normal	-1.0	0.0	1.0

Courtesy: Simon Mason (IRI)

Hit score

- Measures how often we did forecast the highest forecast probability on the category that was observed, how often we did forecast the second highest forecast probability on the category that was observed, and how often we did forecast the lowest forecast probability on the category that was observed

- Useful to indicate quality of a single forecast map, or the collection of forecasts over the years

Hit score

To compute hit score on needs to compute the rank of the forecast probabilities for the 3 categories and generate a scoring rule for the highest, second highest and lowest hits

Ranks can be defines as follows:

- rank is 1 for highest forecast probability
- rank is 1.5 if there are ties between highest and 2nd highest fcst prob
- rank is 2 for 2nd highest forecast probability
- rank is 2.5 if there are ties between lowest and 2nd lowest fcst prob
- rank is 3 for lowest forecast probability
- rank is -1 if climatological (1/3,1/3,1/3) forecast is issued

The scoring rule is as follows:

Highest: $H=R_1+0.5(R_{1.5})+0.33(R_{-1})/T$

Second: $S=R_2+0.5(R_{1.5})+0.5(R_{2.5})+ 0.33(R_{-1})/T$

Lowest: $L= R_3+0.5(R_{2.5})+0.33(R_{-1})/T$

Where:

- T is the total number of forecasts being verified
- R_1 is the number of times when the highest forecast probability was issued for the observed category (i.e. the counts of rank equal 1)
- R_2 is the number of times when the second highest forecast probability was issued for the observed category (i.e. the counts of rank equal 2)
- R_3 is the number of times when the lowest forecast probability was issued for the observed category (i.e. the counts of rank equal 3)
- $R_{1.5}$ is the number of times when the highest or second highest fcst prob. were issued for the observed category (i.e. the counts of rank equal 1.5)
- $R_{2.5}$ is the number of times when the lowest or second lowest fcst prob. were issued for the observed category (i.e. the counts of rank equal 2.5)
- R_{-1} is the number of times when the climatological forecast probability was issued for the observed category (i.e. the counts of rank equal -1)

Hit score

One would like to see a large number of rank 1 (highest forecast probability issued for the observed category) and a small number of rank 3 (lowest forecast probability issued for the observed category), so the percentage for H is high and the percentage for L is small