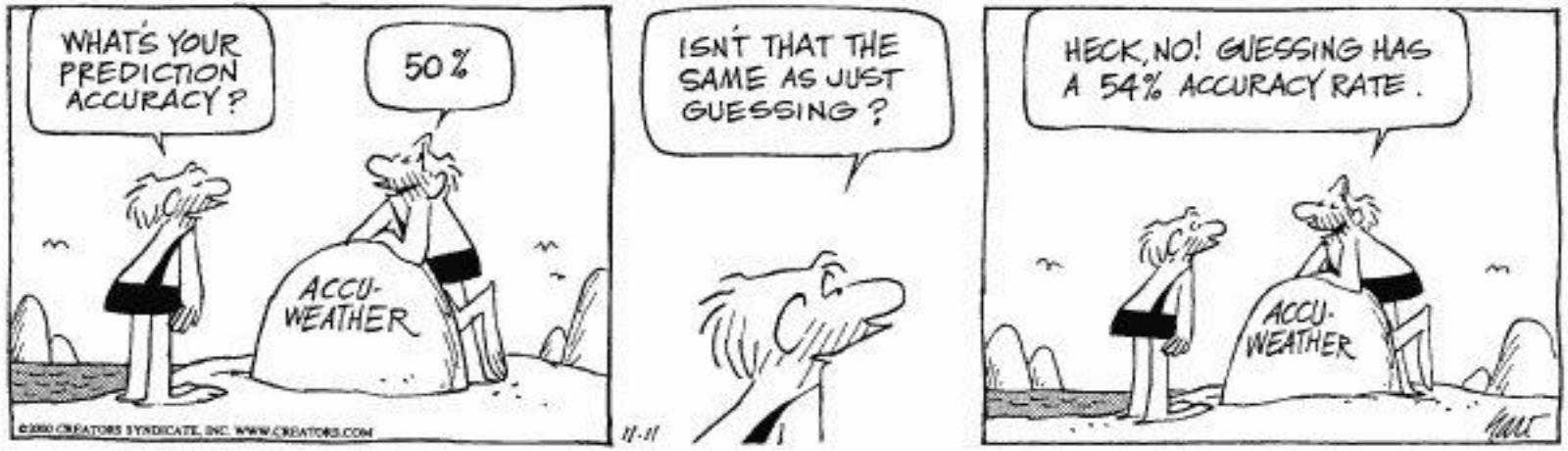
Introduction: Verification of Seasonal Forecasts

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

Verifying Forecasts



If 60% of the forecasts are correct, is that good?

1. Is 60% low?
2. What does "correct" mean if the forecasts are probabilistic?
3. Does 60% mean that I can use the forecasts?

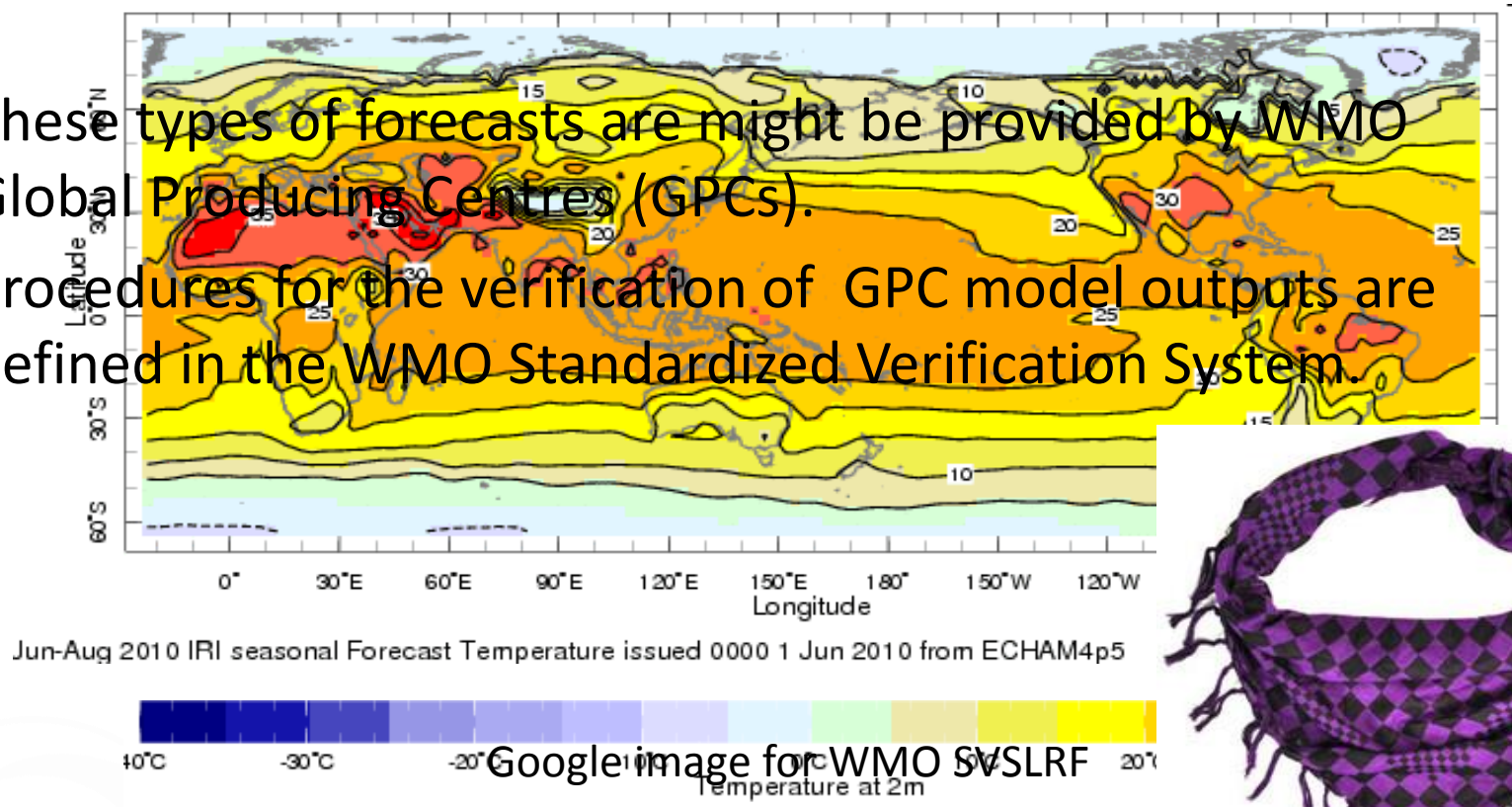
More generally:

1. Are the forecasts good?
2. Is the choice of score appropriate? What does the score mean?

Seasonal forecast formats

Most seasonal forecasts are in one of two classes:

1. **A (set of) deterministic forecast(s) – outputs from a dynamical or statistical model.**



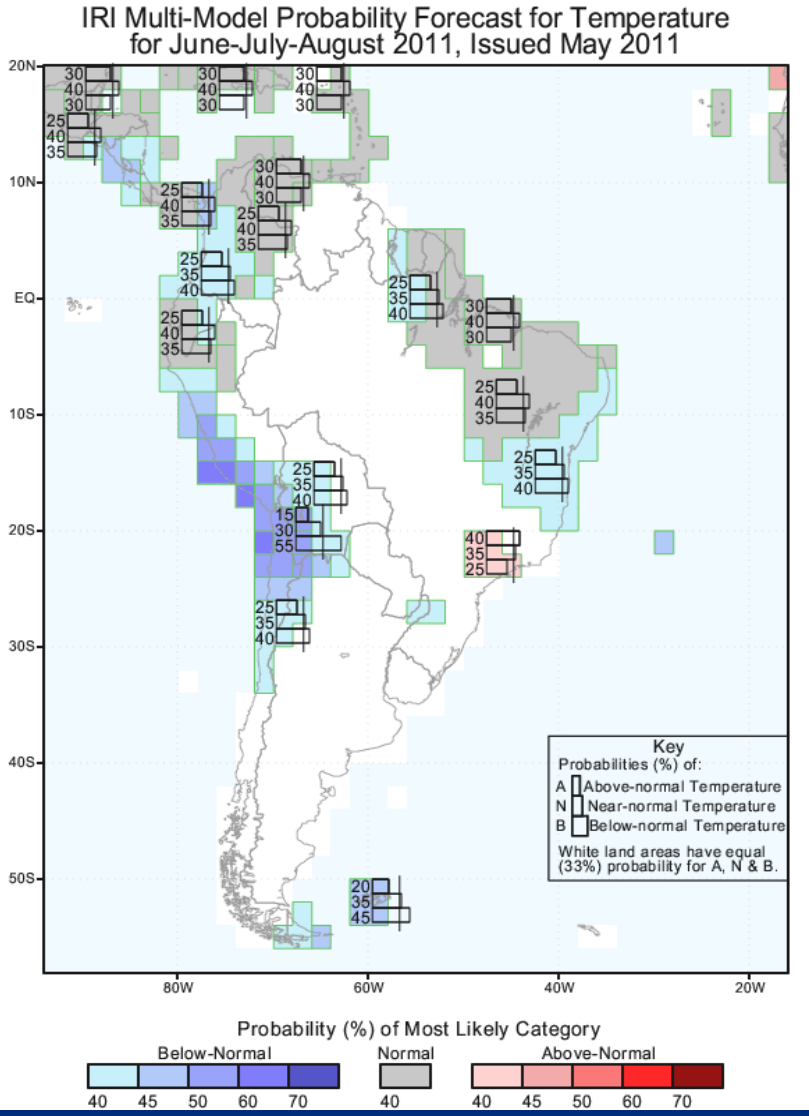
These types of forecasts are might be provided by WMO Global Producing Centres (GPCs).

Procedures for the verification of GPC model outputs are defined in the WMO Standardized Verification System.

Google image for WMO SVSLRF

Seasonal forecast formats

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

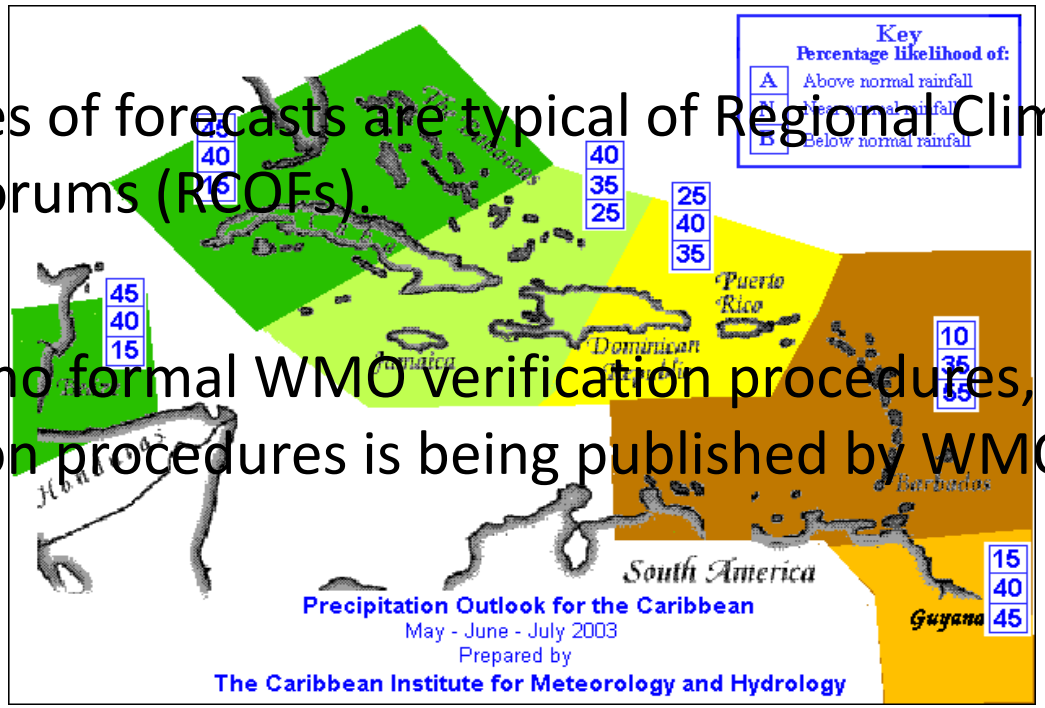


Seasonal forecast formats

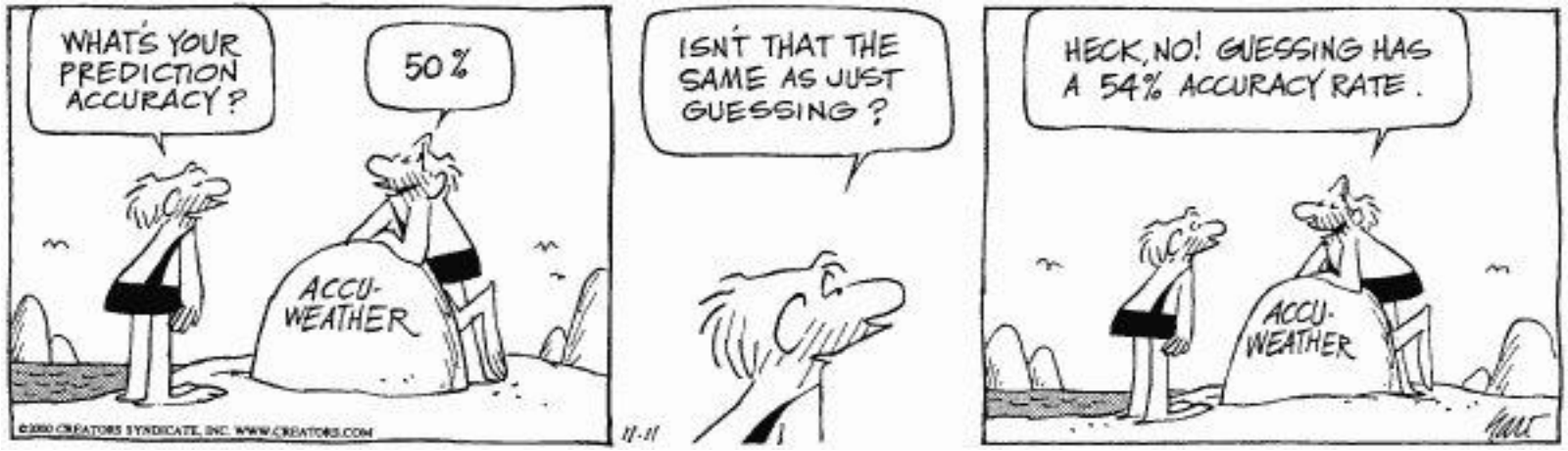
2. (b) Maps showing probabilities of the verification falling within one of two or more categories (by region)

These types of forecasts are typical of Regional Climate Outlook Forums (RCOFs).

There are no formal WMO verification procedures, but some guidance on procedures is being published by WMO CCl.



Verifying Forecasts



If 60% of the forecasts are correct, is that good?

1. Is 60% low?
2. What does "correct" mean if the forecasts are probabilistic?
3. Does 60% mean that I can use the forecasts?

More generally:

1. **Are the forecasts good?**
2. Is the choice of score appropriate? What does the score mean?

What makes a “good” forecast?

Forecast: Japan will do well in the 2015 Rugby World Cup.

Verification: ? They were knocked out by New Zealand (the possible champions) in the Quarter Finals.

Ambiguity

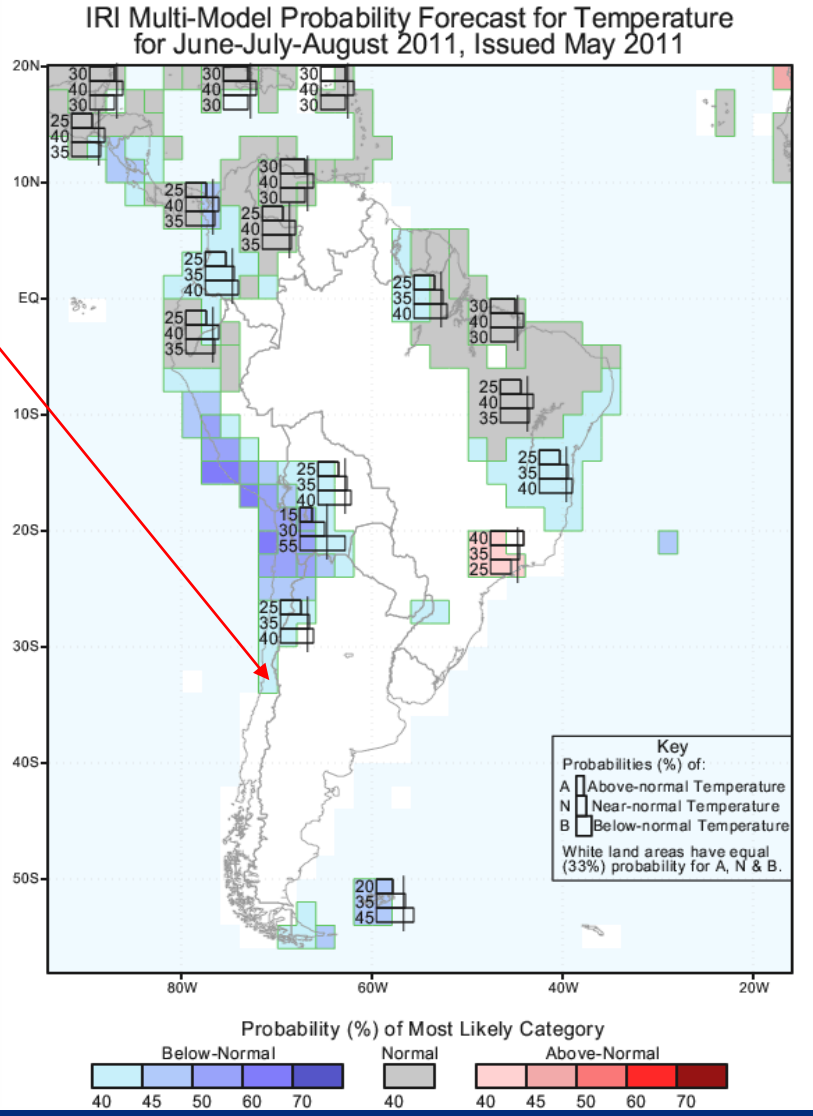
What is the predictand?

What is the seasonal forecast for Santiago?

Does the forecast apply to:

- Individual stations?
- Grid averages?

These forecasts apply to grid averages, not to individual locations. But how well the forecasts verify at individual locations may be an interesting question anyway.

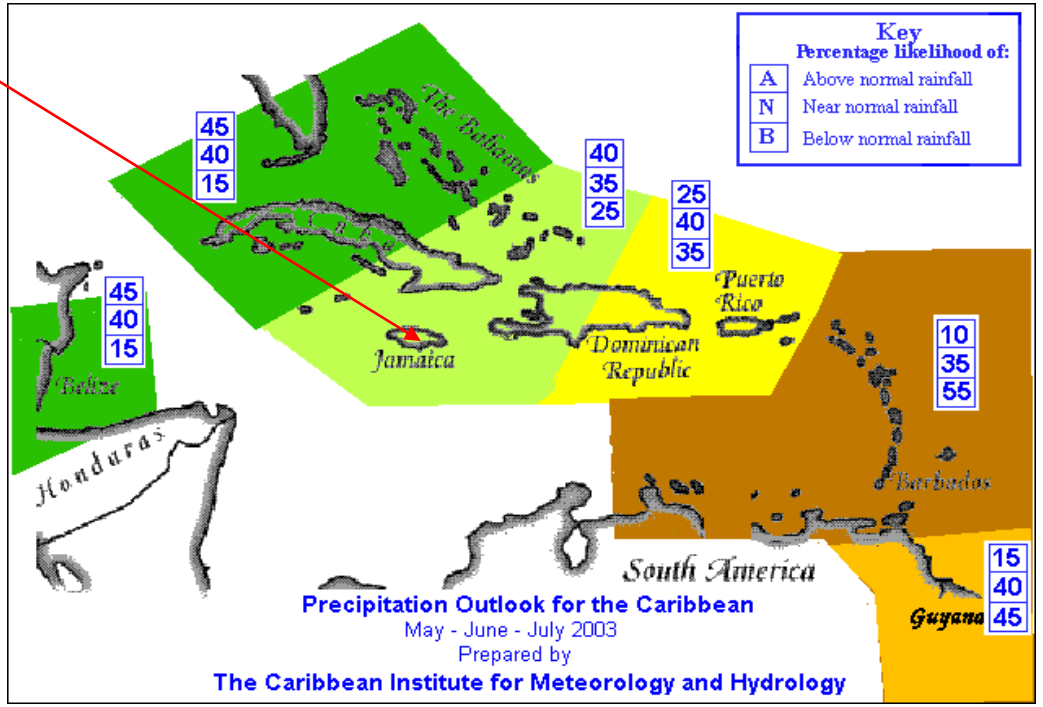


What is the predictand?

What is the seasonal forecast for Kingston? Does the forecast apply to:

- Individual stations?
- Grid averages?

These forecasts apply to area averages (I think), but what are the areas?



What is the predictand?

In many RCOFs it is unclear what the predictand is because:

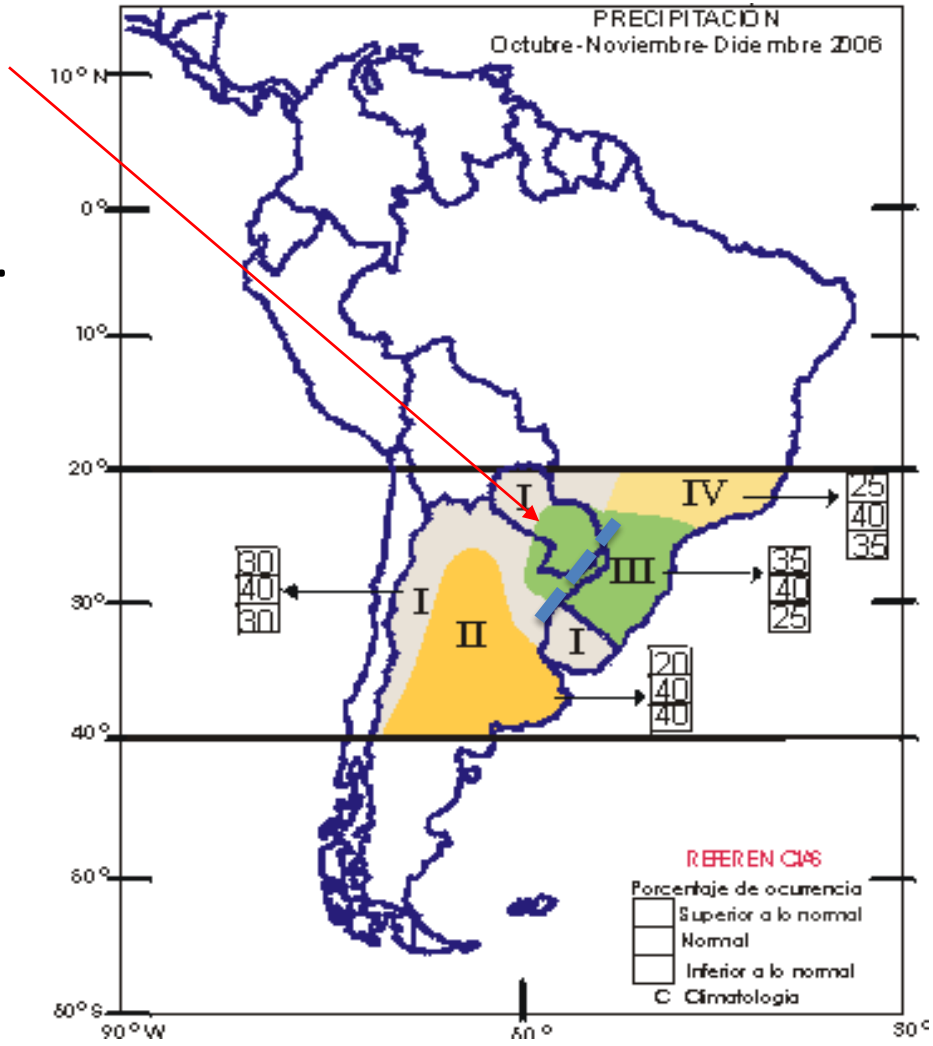
- The forecast inputs are made in different ways:
  - representative stations;
  - regional indices;
  - multiple stations.
- The original regions are often combined, and verifying using original area averages does not necessarily give the same result as verifying using the combined area-average.

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

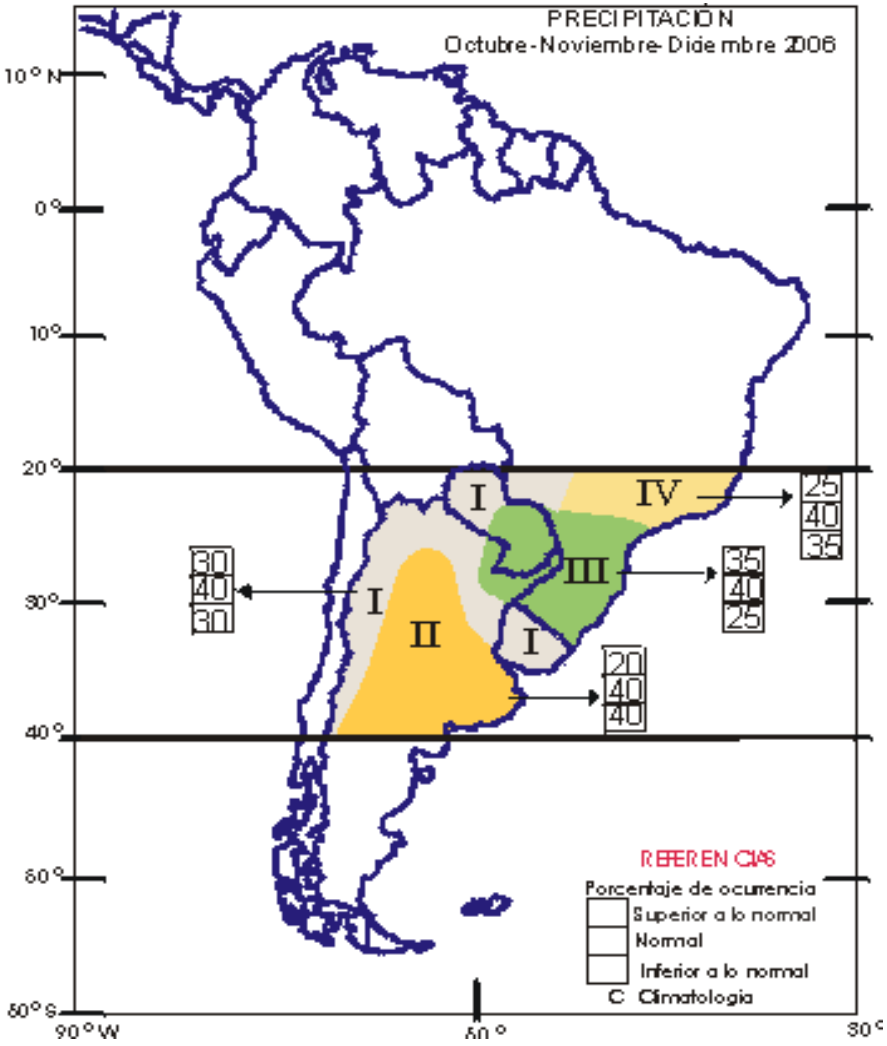
Consistent forecasts?

All regions have highest probability on the normal category.

Did we genuinely think that normal was the most likely category everywhere, or did we think it was the safest forecast everywhere?

70 – 80% of all the African RCOF forecasts have highest probability on normal.

Are we really forecasting what we think, or are we playing safe?



Forecast "goodness"

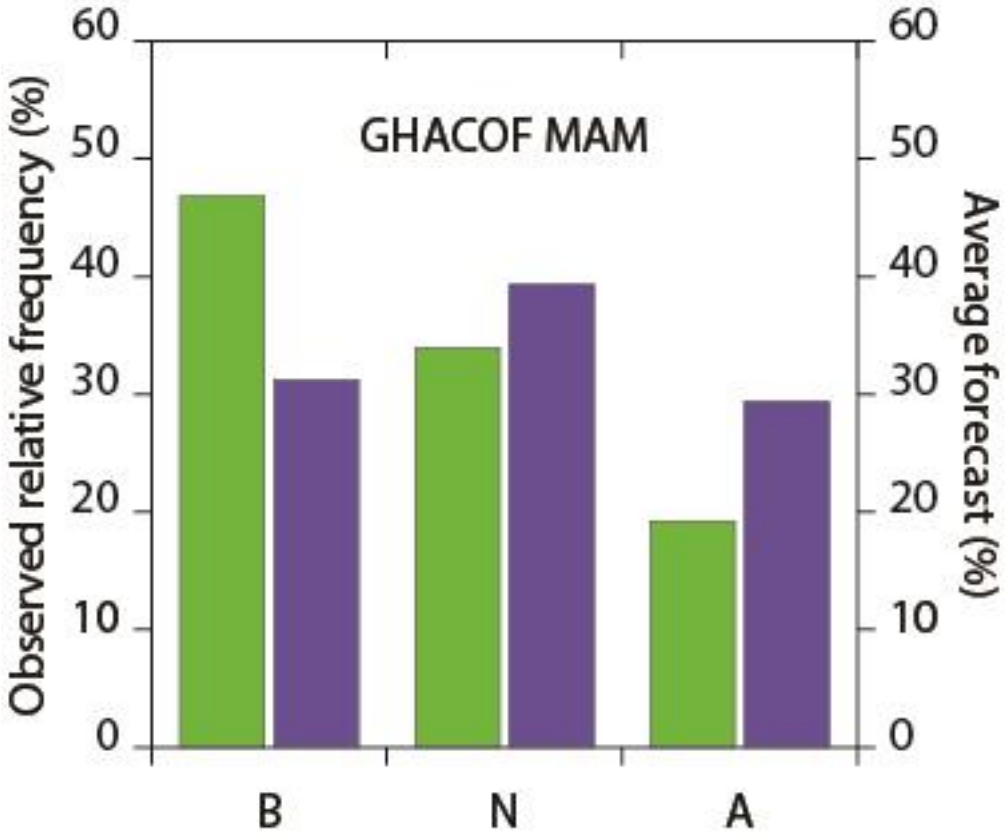
What makes a "good" forecast?

1. Consistency
2. Quality
3. Value

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

Consistency: Unconditional bias

- Are probabilities consistently too high or too low?



Reliability

Give 90% confidence limits for the mean annual rainfall in New Delhi.

706 mm (27.8 inches).

How many of these confidence intervals contain the observed annual rainfall?

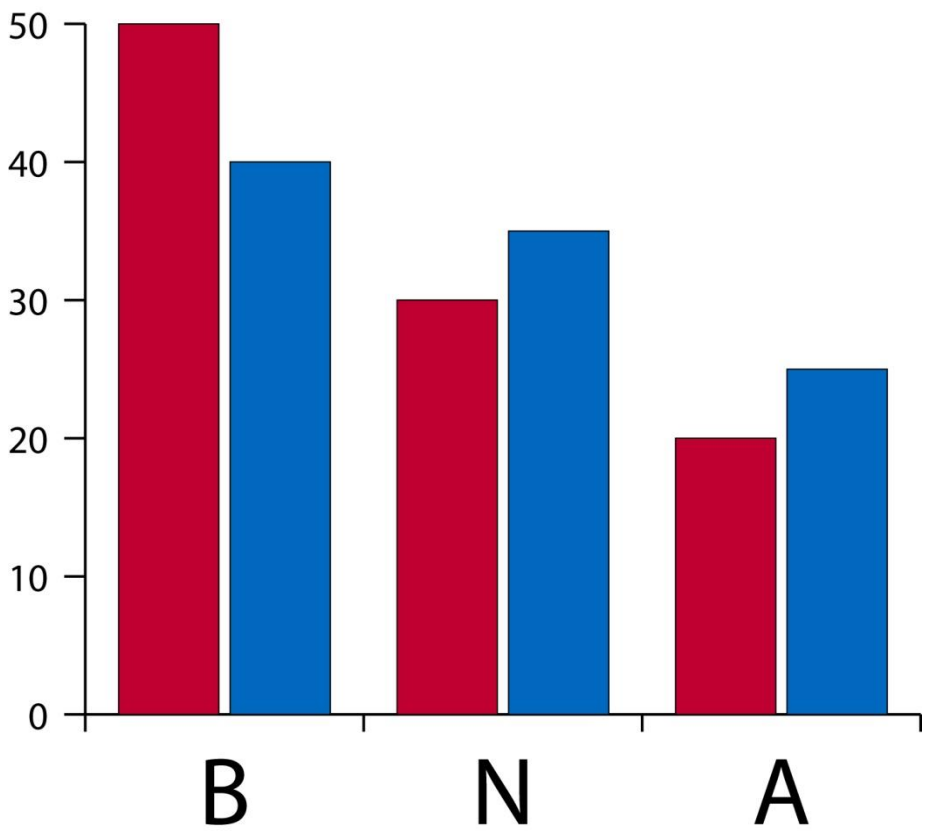
Differences between how many of these confidence intervals *do* contain the observed annual rainfall and how many *should* contain the observation are the subject of questions about reliability?

Reliability

- When we say 40% chance of below-normal, we expect a forecast of below-normal to be correct 40% of the time.

- If we take all our forecasts for a location when we said 40% chance of below-normal, 40% of them (not more or less) should be below-normal.

Verification of probabilistic forecasts



Given that below-normal occurs, which is the better forecast?

We should not get too excited when below-normal rainfall occurs, even though it is the category with the highest probability, because most of the time (60% for blue) we should want below-normal *not* to occur.

Reliability

Imagine a set of forecasts that indicates probabilities of rainfall (which has a climatological probability of 30%):

01 Feb	60%
02 Feb	60%
03 Feb	60%
04 Feb	60%
05 Feb	60%
06 Feb	10%
07 Feb	10%
08 Feb	10%
09 Feb	10%
10 Feb	10%

Suppose that rainfall occurs on 40% of the green forecasts, and 20% of the brown.

The forecasts correctly indicate times with increased and decreased chances of rainfall, but do so over-confidently.

Reliability and Sharpness

Climatological forecasts are reliable (assuming no climate change), but they lack sharpness.

Good probabilistic forecasts are reliable *and* sharp.

Resolution

Does the outcome change when the forecast changes?

01 Feb	60%
02 Feb	60%
03 Feb	60%
04 Feb	60%
05 Feb	60%
06 Feb	10%
07 Feb	10%
08 Feb	10%
09 Feb	10%
10 Feb	10%

When the forecast is 10% rain occurs 20% of the time.

When the forecast is 60% rain occurs 40% of the time.

Rain becomes more frequent when the forecast probability increases – there is resolution.

Resolution

- Does the outcome change when the forecast changes?
- Example: does above-normal rainfall become more frequent when its probability increases?
- Resolution is *the* crucial attribute of a good forecast. If the outcome differs depending on the forecast then the forecasts have useful information. If the outcome is the same regardless of the forecast the forecaster can be ignored.

Discrimination

Does the forecast change when the outcome changes?

01 Feb	60%
02 Feb	60%
03 Feb	60%
04 Feb	60%
05 Feb	60%
06 Feb	10%
07 Feb	10%
08 Feb	10%
09 Feb	10%
10 Feb	10%

When rain occurs the average forecast probability is 43%.

When it is dry the average forecast probability is 31%.

The forecast probability for rain is higher when it does rain – there is some discrimination.

Discrimination

- Does the forecast differ when the outcome differs?
- Example: is the probability on above-normal rainfall higher when above-normal rainfall occurs compared to when rainfall is normal or below-normal?
- Discrimination is an alternative perspective to resolution. If the forecast differs given different outcomes then the forecasts have useful information. If the forecast is the same regardless of the outcome the forecaster can be ignored.

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

What is “skill”?

Skill

Is one set of forecasts better than another?

- Skillful forecasts are not necessarily good; both sets of forecasts may be really bad.
- Unskillful forecasts are not necessarily bad: both sets of forecasts may be really good.

"Skill" is poorly defined. What do we mean by "better"?

Skill

Imagine a set of forecasts that indicates probabilities of rainfall (which has a climatological probability of 30%):

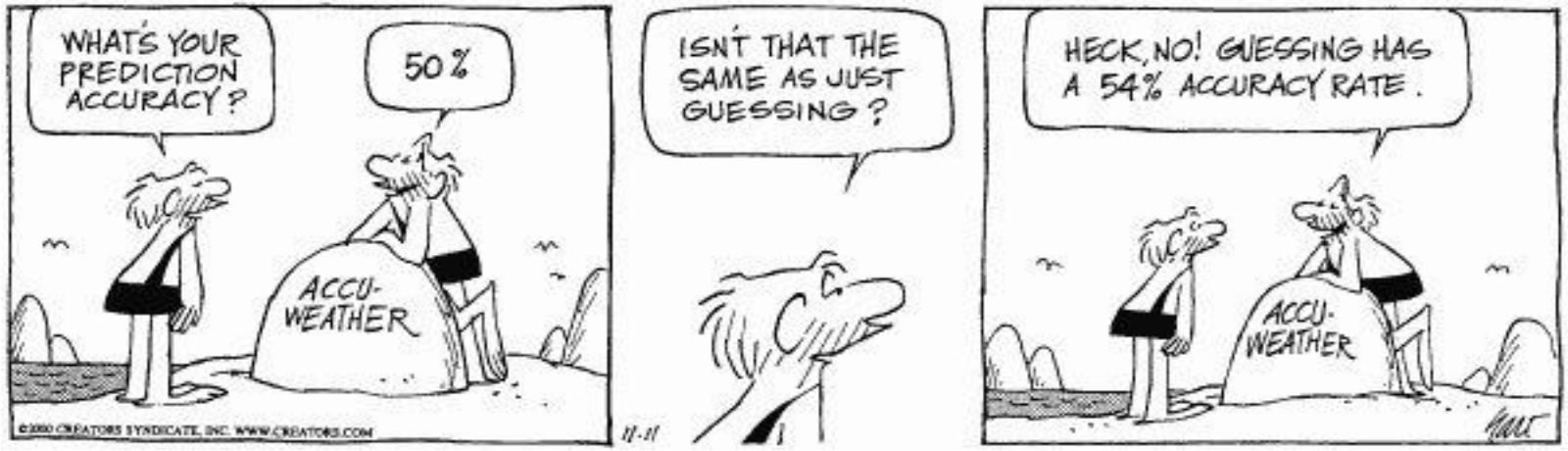
01 Feb	60%
02 Feb	60%
03 Feb	60%
04 Feb	60%
05 Feb	60%
06 Feb	10%
07 Feb	10%
08 Feb	10%
09 Feb	10%
10 Feb	10%

Suppose that rainfall occurs on 40% of the green forecasts, and 20% of the brown.

The forecasts correctly indicate times with increased and decreased chances of rainfall, but do so over-confidently.

The Brier skill score is -7%.

Verifying Forecasts



If 60% of the forecasts are correct, is that good?

1. Is 60% low?
2. What does "correct" mean if the forecasts are probabilistic?
3. Does 60% mean that I can use the forecasts?

More generally:

1. Are the forecasts good?
2. **Is the choice of score appropriate?** What does the score mean?

Propriety

A forecaster has to predict rainfall occurrence. Forecasts are issued probabilistically.

On day-10 the forecaster thinks rainfall has a probability of 80%, but wants to know what probability, *P*, to state.

Day	Rain?	For
1	Y	80%
2	Y	80%
3	Y	70%
4	Y	70%
5	N	70%
6	N	30%
7	N	30%
8	N	10%
9	N	20%
10		?

Propriety: the Linear Probability Score

$$\text{LPS} = |P - V|$$

Let $V = 1$, if the event occurs, and $V = 0$ otherwise.

If $P = 80\%$, the expected linear score is: 0.32

But if $P = 100\%$ the expected score is: 0.20

A better score is obtained by hedging (scores near 0 are good).

Day	Rain?	For
1	Y	80%
2	Y	80%
3	Y	70%
4	Y	70%
5	N	70%
6	N	30%
7	N	30%
8	N	10%
9	N	20%
10		?

Propriety: the Brier Score

$$\mathrm{BS} = (P - V)^2$$

If P = 80%, the expected Brier score is: 0.16

But if P = 100% the expected score is: 0.20

A better score is obtained by hedging (scores near 0 are good).

Day	Rain?	For
1	Y	80%
2	Y	80%
3	Y	70%
4	Y	70%
5	N	70%
6	N	30%
7	N	30%
8	N	10%
9	N	20%
10		?

Propriety: the Brier Skill Score

$$\mathrm{BSS} = 1 - \frac{(P - V)^2}{(C - V)^2}$$

If P = 80%, the expected Brier skill score is: 0.36

But if P = 100% the expected score is: 0.20

A worse score is obtained by hedging (large scores are good).

But only because C = 1- C (c.f. seasonal forecasts)

Day	Rain?	For
1	Y	80%
2	Y	80%
3	Y	70%
4	Y	70%
5	N	70%
6	N	30%
7	N	30%
8	N	10%
9	N	20%
10		?

Propriety: the Linear Probability Score

$$\mathrm{BSS} = 1 - \frac{(P - V)^2}{(C - V)^2}$$

What if $C = 0.1$?

If $P = 80\%$, the expected Brier skill score is: -0.12

If $P = 100\%$ the expected score is: -0.19

A worse score is again obtained by hedging.

But if $P = 4.7\%$ the expected score is: 0.06

A better score is obtained by careful hedging!

Hit scores

Above	30
Normal	45
Below	25

If Above occurs is this a good, neutral or bad forecast?
What if Below occurs?

Above	45
Normal	30
Below	25

If Normal occurs is this a good, neutral or bad forecast?
What if Below occurs?
If Below occurs which of the two forecasts is worse?

Hit scores and hedging

- How often does 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**

Propriety

- Any score that encourages the forecaster to forecast something other than what they really think is not a “proper” score.
- There are better ways of scoring hits that do not encourage hedging …

Summary

- Many consensus forecasts are ambiguous: this problem *must* be addressed.
- Probabilistic forecasts should be verified in terms of:
  - Reliability
  - Resolution
  - Discrimination
  - Sharpness
- Scores can be misleading: some may be too pessimistic, others may encourage the forecasters to hedge. In general we should avoid such scores.