Towards Standardized Verification of Seasonal Forecasts

Simon J. Mason[1], and Andreas P. Weigel[2]

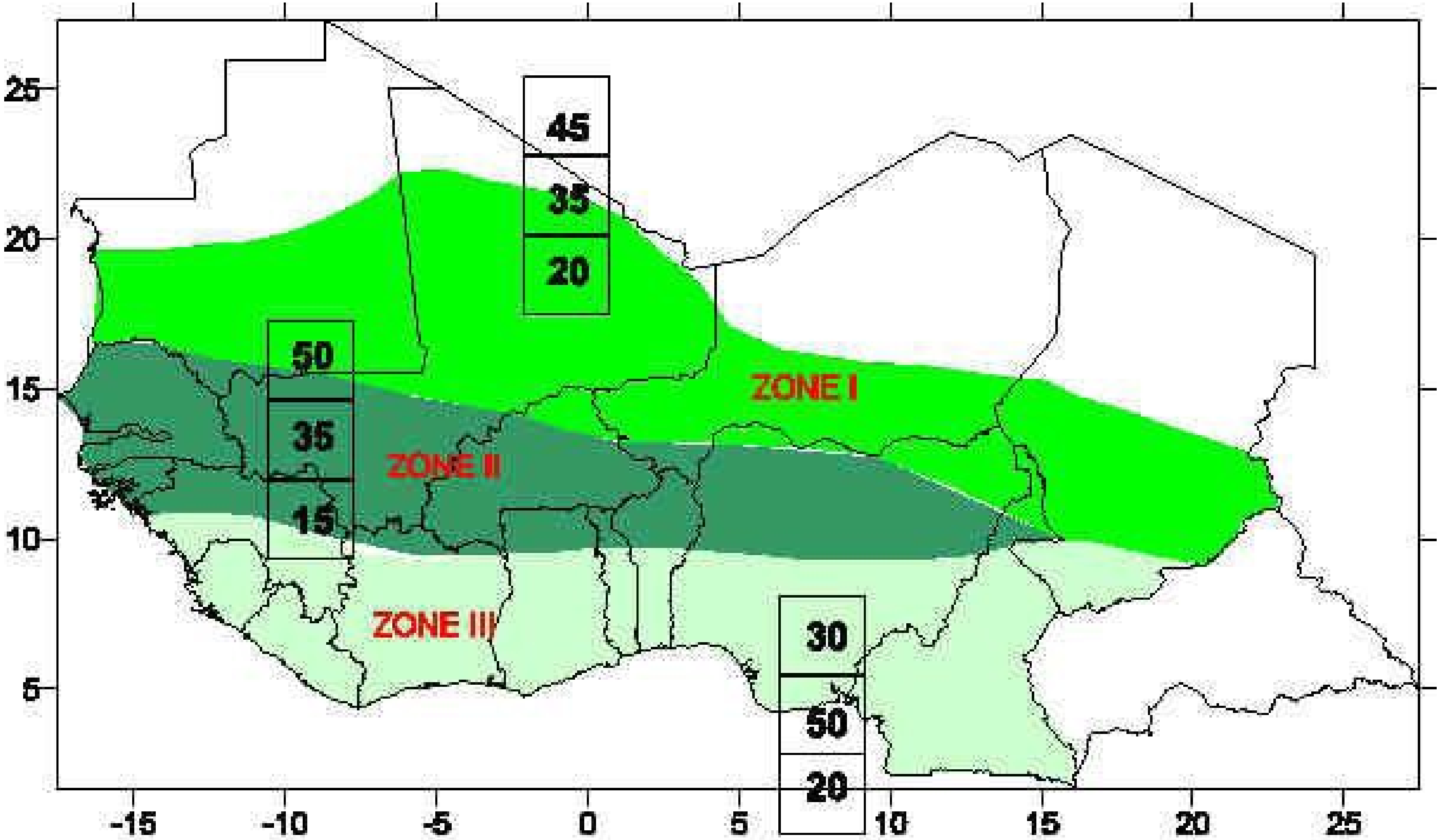
simon@iri.columbia.edu

1. International Research Institute for Climate and Society
The Earth Institute of Columbia University

2. Federal Office of Meteorology and Climatology, MeteoSwiss

Fourth International Verification Methods Workshop

Helsinki, Finland, 8 – 10 June, 2009



Key

Percentage likelihood of:

A	Above-normal rainfall
N	Near-normal rainfall
B	Below-normal rainfall

Over a decade of operational forecasts.

SVSLRF targets GPC model outputs.

Recommended Procedures

- Results should be communicable to the general public.

- Scores should measure specific attributes of good probabilistic forecasts.

Forecast Attributes

- What, precisely, do we mean by a “good” (probabilistic) forecast?

- Most of the time we measure forecast quality it is not clear which attribute we are measuring …

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%.

Skill

The problem with the Brier and the ranked probability scores is that they try to measure a number of attributes of good probabilistic forecasts at the same time.

In the previous example, the problem is that the squared reliability errors (2.5%) more than offset the gain in resolution (1%). Why should this difference be interesting?

Most generic scores are difficult to interpret because they measure multiple attributes.

“Skill” is a vague attribute – in what respect(s) is (are) one set of forecasts better than another?

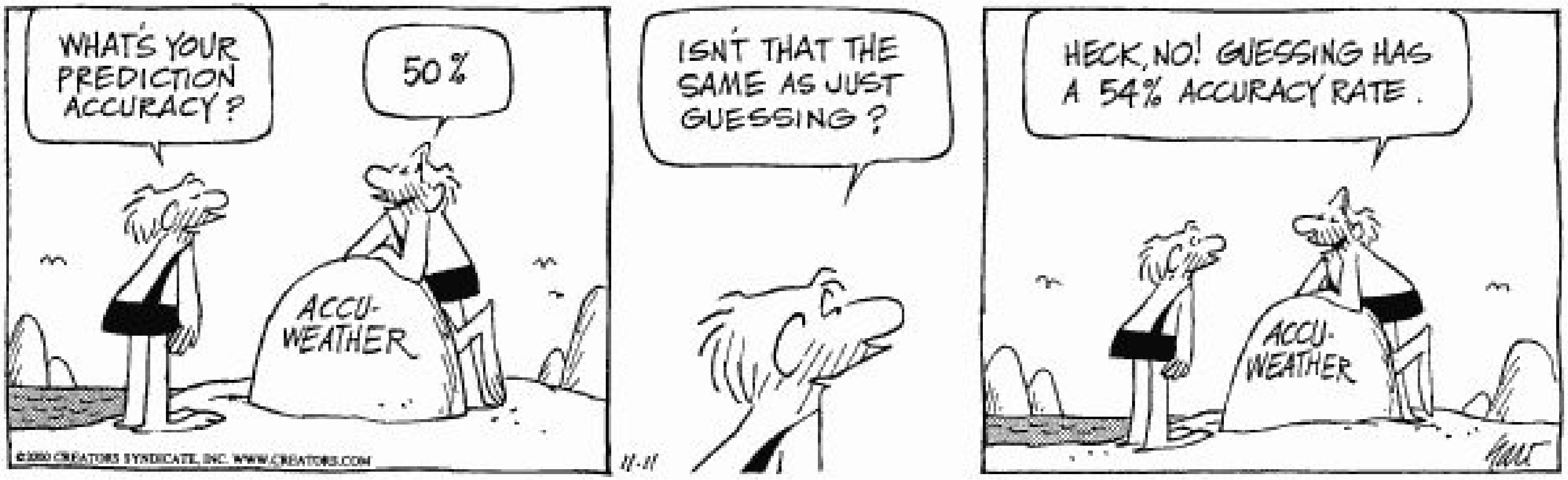
Score or procedure	**Attributes**	**References**
Generalized discrimination	Discrimination	Mason and Weigel (2009)
ROC graph	Discrimination	Mason (1982); Harvey et al. (1992)
ROC area	Discrimination	Mason (2003)
Reliability score	Reliability	Murphy (1973)
Resolution score	Resolution	Murphy (1973)
Hit scores of ranked categories	Resolution	
Effective interest rate	Value	Hagedorn and Smith (2008)
Accumulated profit graphs	Value	Hagedorn and Smith (2008)
Reliability diagrams	Reliability, resolution, sharpness	Hsu and Murphy (1986)
Tendency diagrams	Unconditional bias	
Slope of reliability curve	Resolution, conditional bias	Wilks and Murphy (1998)

Score	**References**
Verification maps as percentiles	
Linear probability score	Wilson et al. (1999)
Average interest rate	Hagedorn and Smith (2008)
Ignorance score	Roulston and Smith (2002)

It does not make much sense to verify a single probabilistic forecast using standard procedures because the attributes of interest cannot be meaningfully measured.

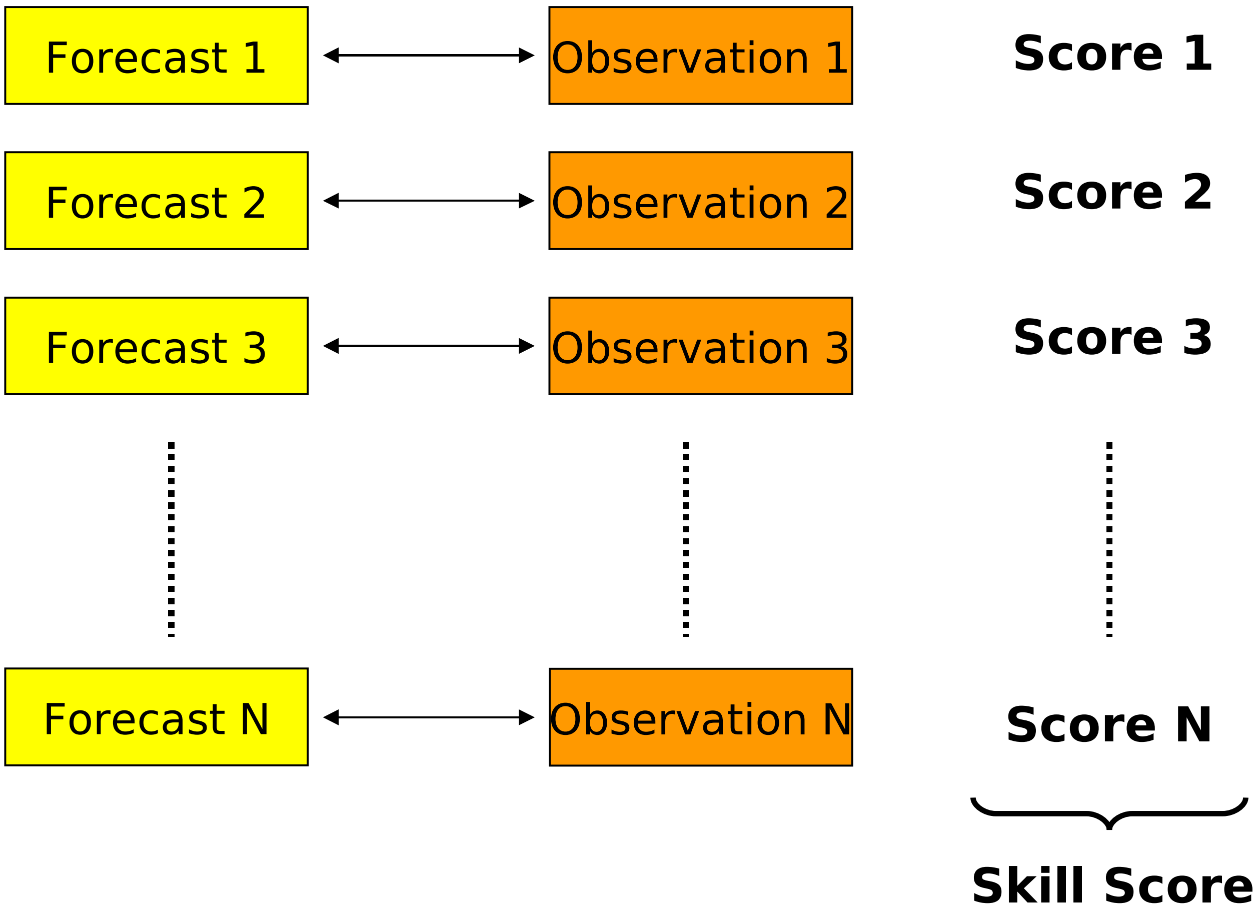
But if is perfectly reasonable to ask “Was last season’s forecast unusually good or bad?” If decisions made in response to last season’s forecasts were particularly beneficial (or detrimental), it would be helpful to know whether to expect similar levels of benefit (or loss) given subsequent forecasts.

Communication


WHAT'S YOUR PREDICTION ACCURACY?
50%
ACCU-WEATHER
©2000 CREATORS SYNDICATE, INC. WWW.CREATORS.COM
11-11
ISN'T THAT THE SAME AS JUST GUESSING?
HECK, NO! GUESSING HAS A 54% ACCURACY RATE.
ACCU-WEATHER

"How often are the forecasts correct?"

The classical approach



Forecast 1 ↔ Observation 1 Score 1

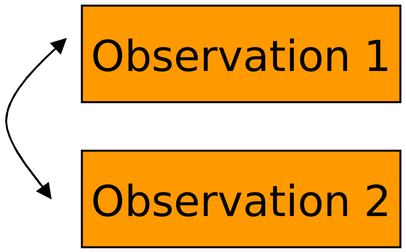
Forecast 2 ↔ Observation 2 Score 2

Forecast 3 ↔ Observation 3 Score 3

Forecast N ↔ Observation N Score N

Skill Score

An alternative approach



- Compare two observations
- One observation has a specified characteristic
- The other observation lacks this characteristic

Which of these two observations has the characteristic
=> Two-alternative forced choice (2AFC)

Two-Alternative Forced Choice Test

Who is an Oscar winning actor?

Peter O’Toole

Russell Crowe

Two-Alternative Forced Choice Test

In which of these two Januaries did El Niño occur (Niño3.4 index >27°C)?

Year
1965
1966

What is the probability of getting the answer correct?

50% (assuming that you do not have inside information about ENSO).

Two-Alternative Forced Choice Test

In which of these two Januaries did El Niño occur (Niño3.4 index >27°C)?

Year	Forecast
1965	No El Niño
1966	El Niño

What is the probability of getting the answer correct now?

That depends on whether we can believe the forecasts.

Two-Alternative Forced Choice Test

But what if we have more information in our forecasts than just “yes” or “no” warnings of El Niño? In which of these two Januaries did El Niño occur (Niño3.4 index >27°C)?

Year	Forecast
1965	25.8°C
1966	28.5°C

What is the probability of getting the answer correct now?

That again depends on whether we can believe the forecasts. Select the forecast with the higher value.

Two-Alternative Forced Choice Test

The same test can be applied if probabilistic forecasts are issued.

Year	Forecast
1965	11%
1966	89%

What is the probability of getting the answer correct now?

That again depends on whether we can believe the forecasts. Select the forecast with the higher probability.

Two-Alternative Forced Choice Test

But what if the outcome is not a simple “yes” or “no”?

The 2AFC test can also be applied to comparisons:

- does *A* score higher (or lower) on some characteristic than *B*?

Two-Alternative Forced Choice Test

Who is the richest soccer team owner?

$6.5 bn

Silvio Berlusconi
(AC Milan)

$8.5 bn

Roman Abramovich
(Chelsea)

Two-Alternative Forced Choice Test

In which of these two Januaries did the *stronger* El Niño occur?

Year	Forecast
1964	27.7°C
1966	28.5°C

What is the probability of getting the answer correct now?

That again depends on whether we can believe the forecasts. Select the forecast with the higher value.

Two-Alternative Forced Choice Test

$$p_{٢AFC} = \frac{١}{n_{٠}n_{١}} \underset{i=١}{\overset{n_{٠}}{\text{�}}} \underset{j=١}{\overset{n_{١}}{\text{�}}} I\left(p_{٠,i}, p_{١,j}\right)$$

Given an ensemble forecast:

$$I\left(p_{0,i}, p_{1,j}\right) = \begin{cases} 0.0 & \text{if } F\left(p_{0,i}, p_{1,j}\right) < 0.5 \\ 0.5 & \text{if } F\left(p_{0,i}, p_{1,j}\right) = 0.5 \\ 1.0 & \text{if } F\left(p_{0,i}, p_{1,j}\right) > 0.5 \end{cases}$$

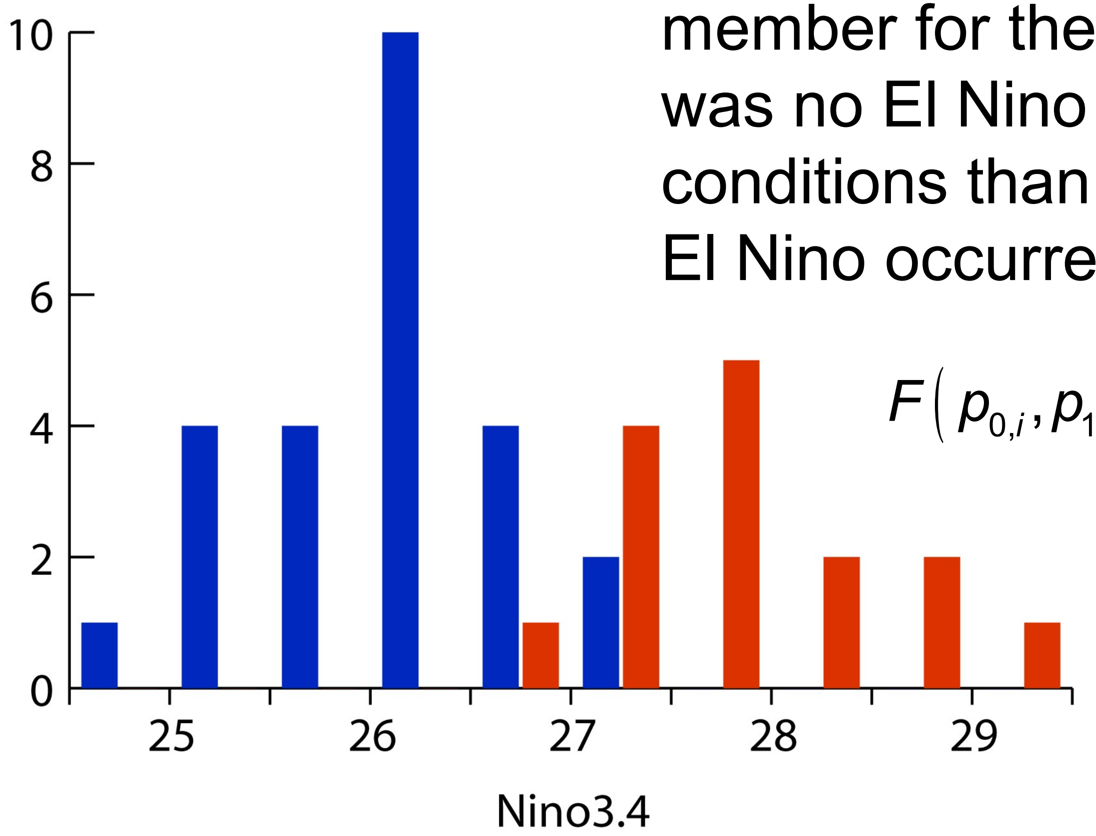
$F(p_{0,i}, p_{1,j})$ is the probability that a the forecast value for a randomly drawn ensemble member from the forecast for the non-event is larger than that for one drawn from the forecast for the event.

Two-Alternative Forced Choice Test

Two 25-member ensemble forecasts – one for when El Nino occurred, the other for when there was no El Nino.

The probability that an ensemble member for the year when there was no El Nino forecasts warmer conditions than one for when there El Nino occurred is 1%.

$$F\left(p_{0,i}, p_{1,j}\right) < 0.5 \therefore I\left(p_{0,i}, p_{1,j}\right) = 1$$



2AFC Tests

If there’s one thing Alan Murphy did abhor,
’Twas an up-start who derived a “new” score.
For, lo and behold!
It was always quite old,
For Sir Gilbert had defined it before.

In many forecasting contexts, the 2AFC-score is not new but can be reduced to an established statistical test

FORECASTS	OBSERVATIONS		
	dichotom.	polychot.	continuous
dichotomous	Pierce	*NA*	*NA*
polychotomous	ROC	Somer	*NA*
probabilities	ROC	Somer	*NA*
continuous	U		Kendall
pdf	Welch*	Welch*	Welch*

* A version of the Student’s *t*-test for cases with unequal variance. The 2AFC score is based on this test, but is not equivalent to it.

Summary

- Verification procedures that measure individual attributes are recommended.
- Summary measures tend to be too abstract to be easily interpretable.
- If the question “How good are the forecasts?” is vague then the question “How skillful are the forecasts?” (i.e., “How much better are the forecasts than the reference?”) is equally vague.
- Most naïve users want to know how often forecasts are correct? Prevarication is not an option.
- A good starting point is to address the question “Are the forecasts worth listening to?” by focusing on resolution or discrimination. Only if the answer is “yes” is it worth asking more detailed questions about whether the forecasts can be taken at face value.

The Verification Blues

Am I blue
Am I blue
Ain't these tears in my eyes telling you
Am I blue
You'd be too
If the hits (one admits) are too few

Was a time,
I thought I’d forecast it
But now I know
I mustn’t broadcast it (Lordy)

Was I gay
Till today
They’re too few
Now I’m through
Am I blue

Questions and prevarications