Forecast Verification

Simon Mason
simon@iri.columbia.edu

International Research Institute
for Climate and Society
EARTH INSTITUTE | COLUMBIA UNIVERSITY

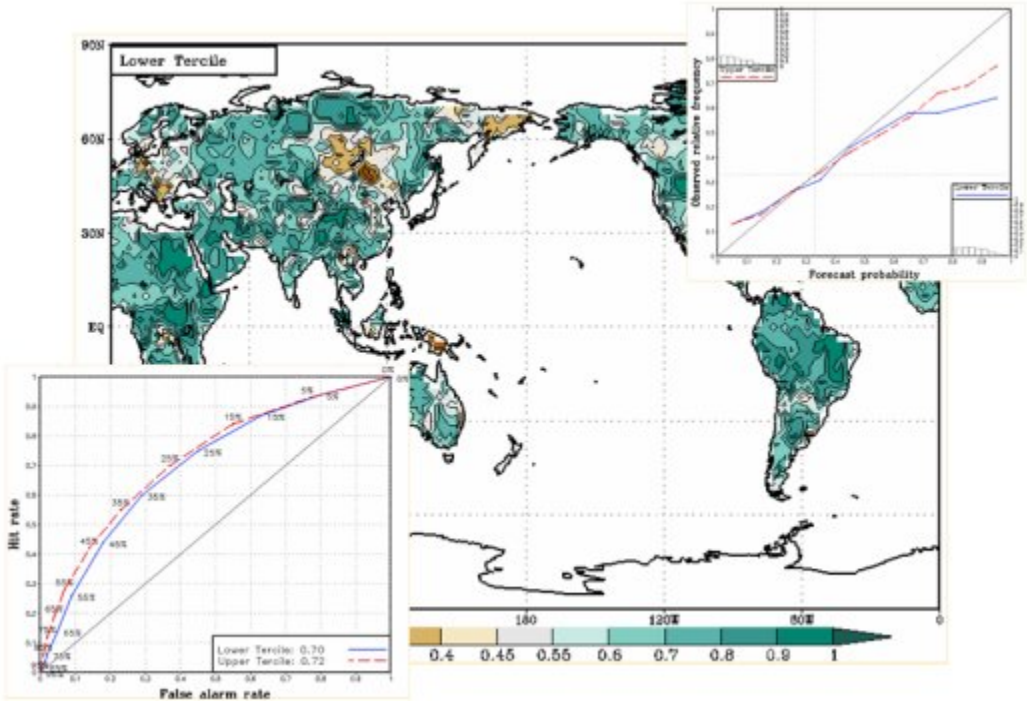
WMO Workshop on Global Review of Regional Climate Outlook Forums
Guayaquil, Ecuador, 5 – 7 September 2017

WMO Standards: the SVSLRF

Procedures for the verification of WMO Global Producing Centres (GPCs) model outputs are defined in the WMO Standardized Verification System.

These procedures are of only limited applicability to RCOF forecasts.

Some use for model selection, but not for forecast interpretation.



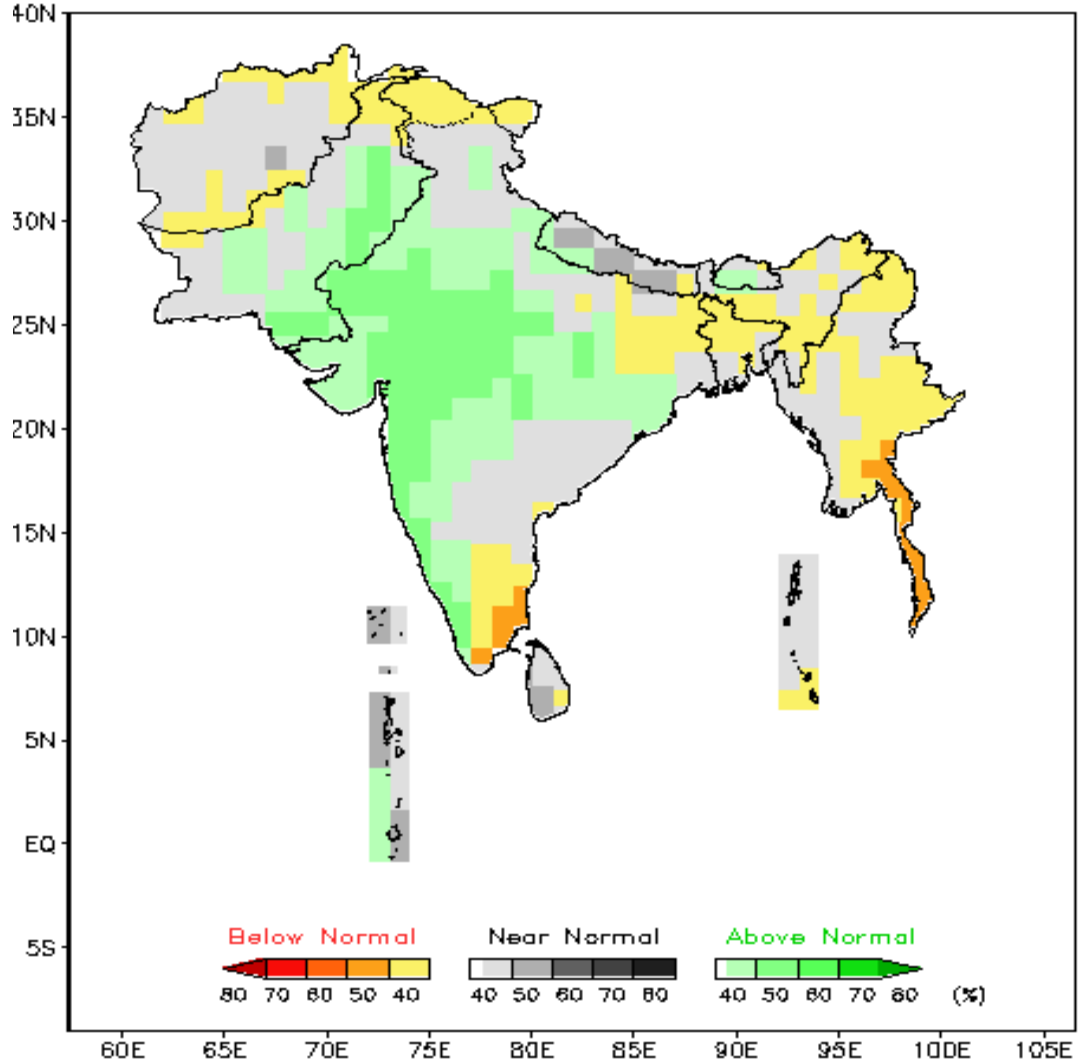
World Meteorological Organization
Lead Centre for the
Long Range Forecast Verification System

RCOF forecast formats – I

Probabilities ***by grid or fixed area***

Forecasts are spatially-averaged rainfall accumulations.

Verify by comparing with observations averaged over the same area.



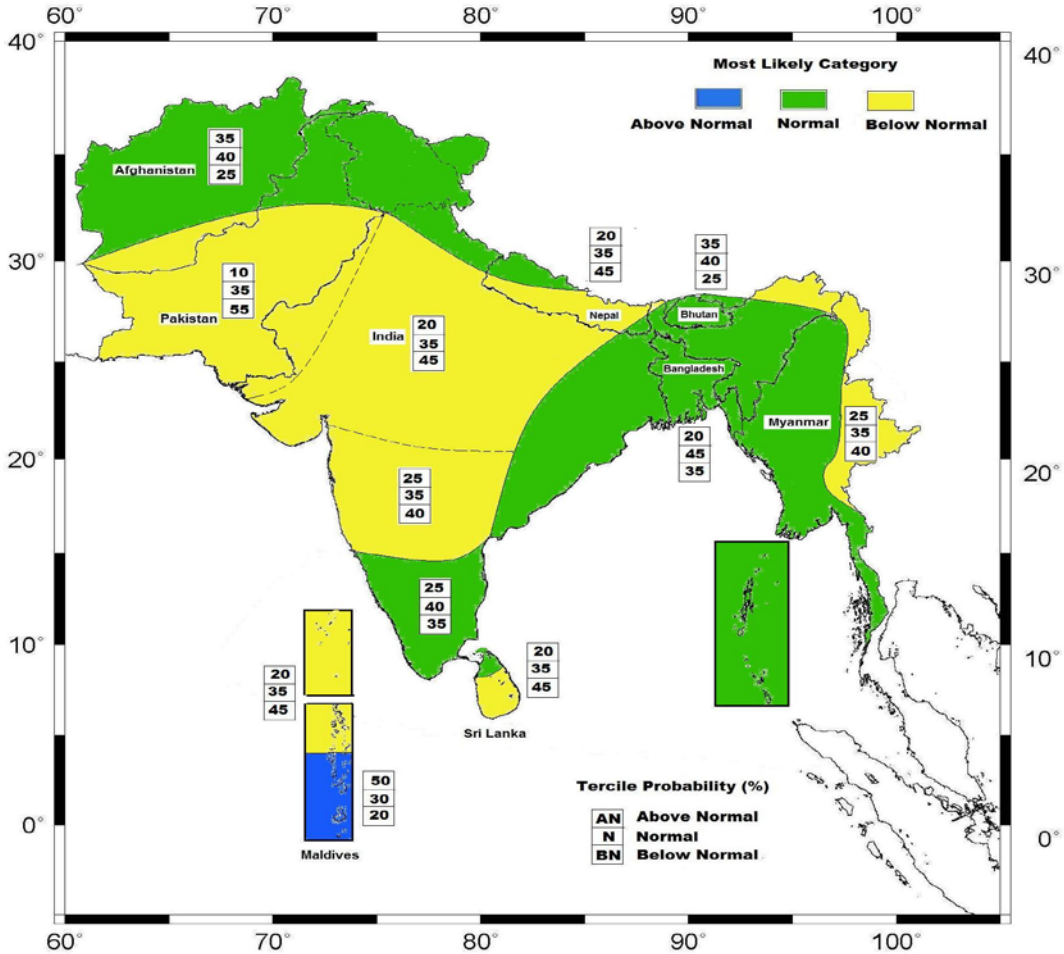
RCOF forecast formats – II

Probabilities ***by region***

Forecasts are NOT regionally-averaged rainfall accumulations.

Verify by station if made that way.

Verifiable (and verified) only by reinterpreting the forecast.



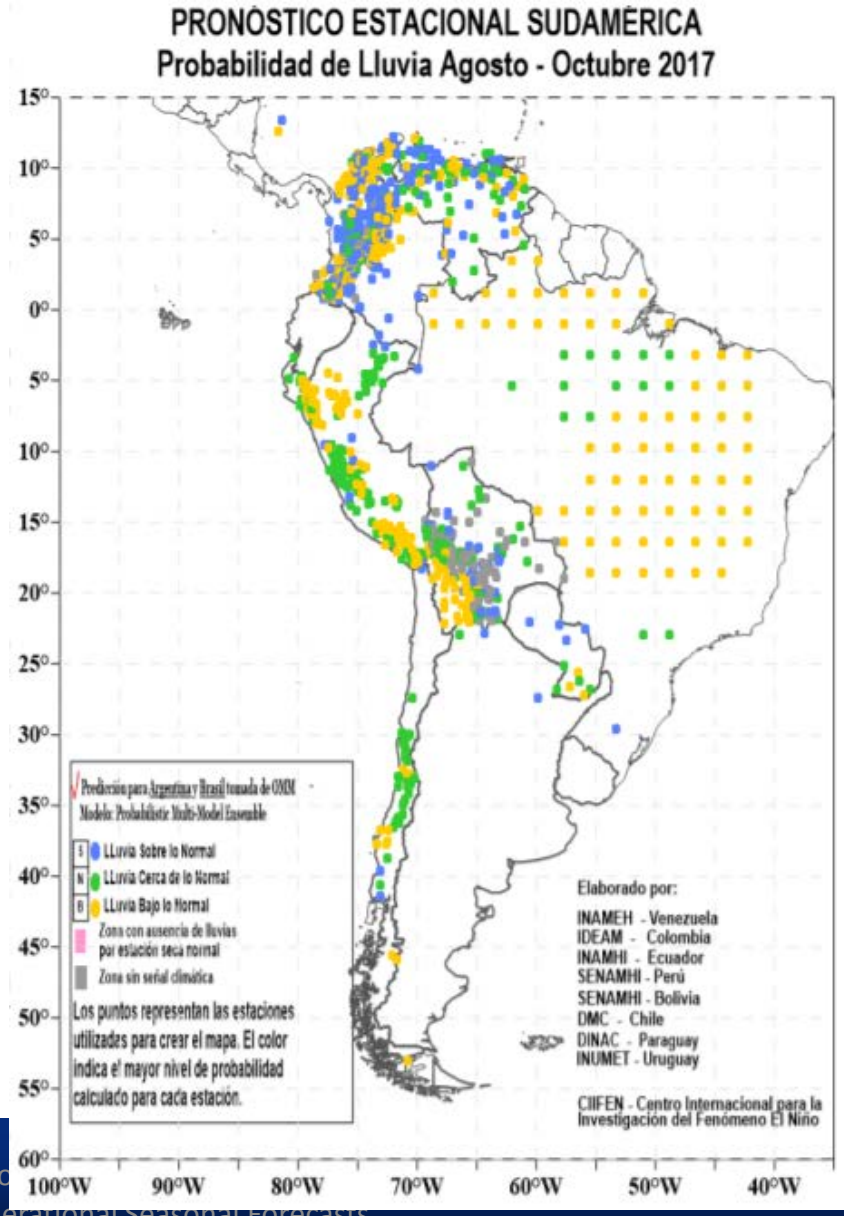
RCOF forecast formats – III

*Probabilities **by station***

Forecasts are location-specific.

Verify by comparing with station observations.

Weight by clustering?



Different verification questions

- How good are the RCOF forecasts? (Not addressed by any RCOF)
- How good was **this** RCOF forecast?

Skill

Is one set of forecasts (A) better than another (B)?

1. If forecasts A are skillful are they good? Forecasts B may be really bad (e.g., guessing for temperature)
2. If forecasts B are unskillful are they bad? Forecasts B may be really good, or made to look artificially good (e.g., "climatology" in RPSS and Brier skill score)

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

Attributes of probabilistic forecasts to verifiy

Reliability the category occur as frequently as implied by the forecasts (40% of below-normal should be below-normal)

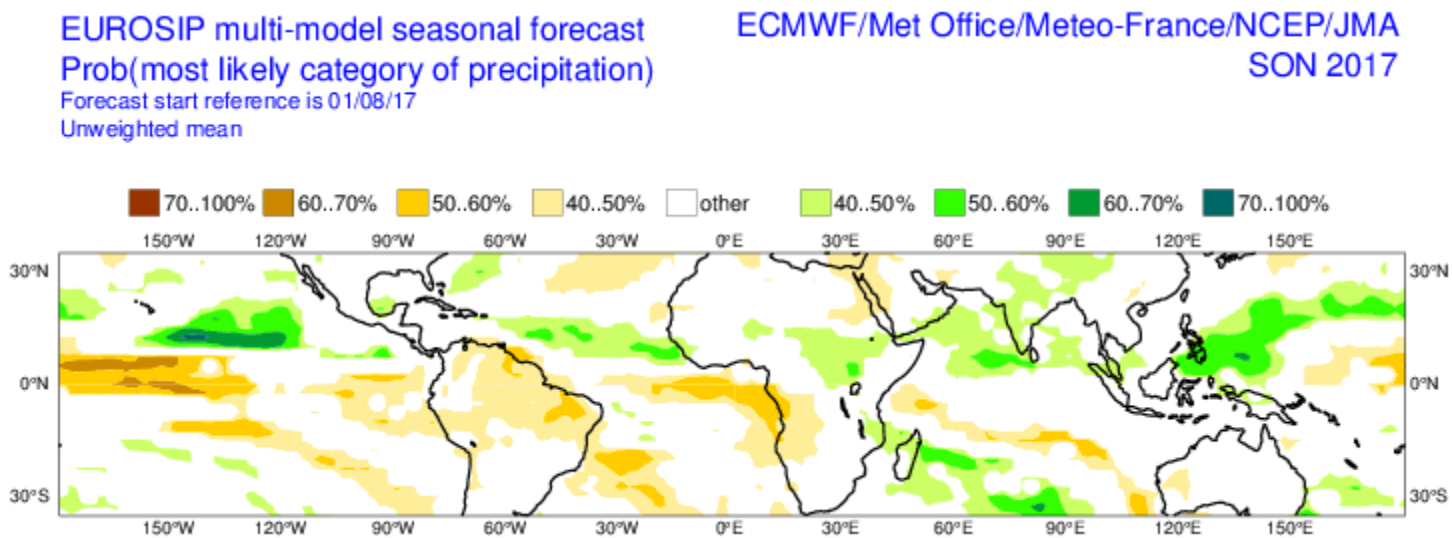
Resolution the outcome differs when the forecast differs; (below-normal should occur more frequently at 40% than at 20%)

Discrimination the forecasts differ when the outcome differs; (below-normal probabilities should be higher when below-normal occurs than when normal and above-normal occur)

Reliability diagrams; ROC (as per SVSLRF); others (CCl Guidance)

Using the SVSLRF

Forecast Interpretation



How do you convert a model probability + a score to a forecast? Is the score a reflection of:

- Poor calibration, or
- Low predictability?

The SVSLRF is being misused to reinforce the subjectivity of the forecast production process.

Using the SVSLRK

Model selection and combination

Sampling errors on all scores are *enormous* (e.g., standard error of a 0.3 correlation with n = 30 is 0.18):

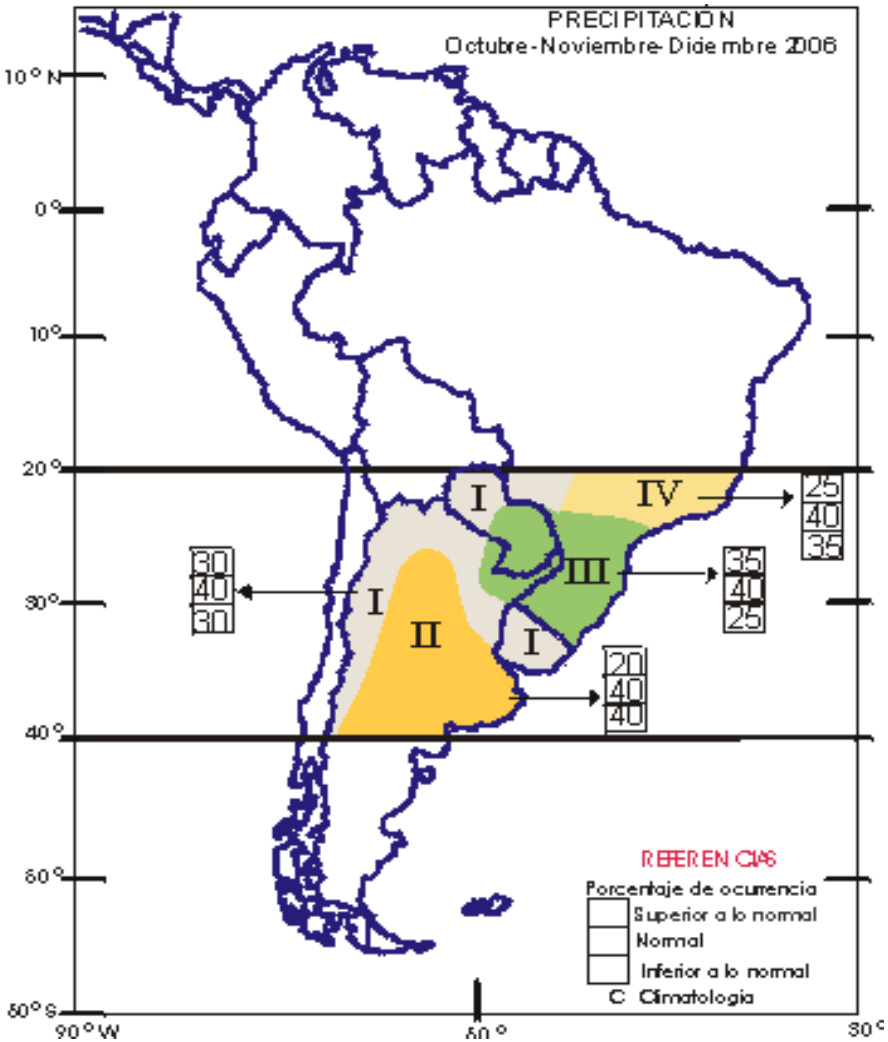
- Identifying the "best" model, or weighting models by skill is just silly!
- The "best" models are the ones that are most accessible and best-supported in your region.

Verification of RCOF forecasts

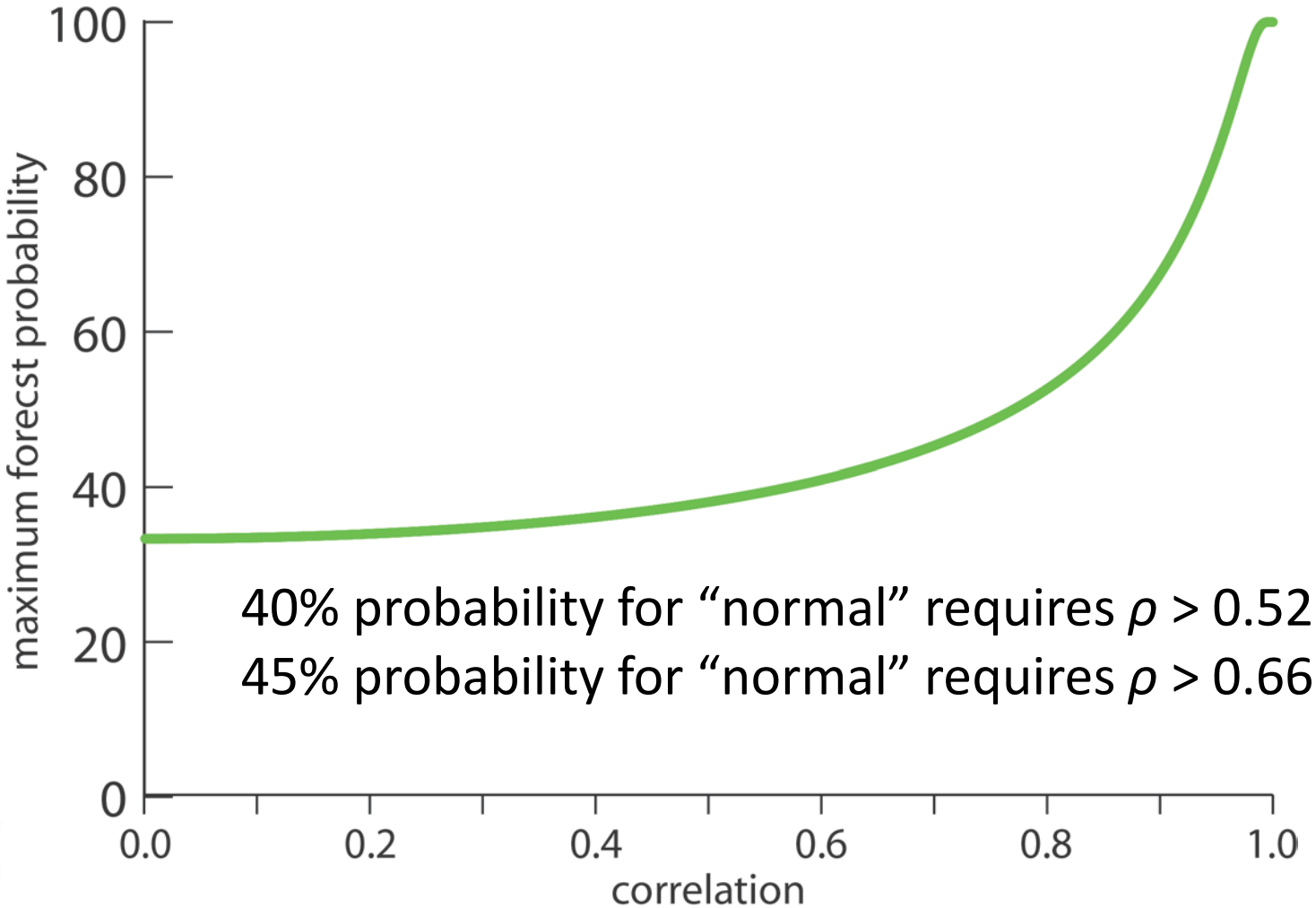
One loud and clear result: hedging on normal.

Do we genuinely think that normal is almost always the most likely category, or do we think it is the safest forecast?

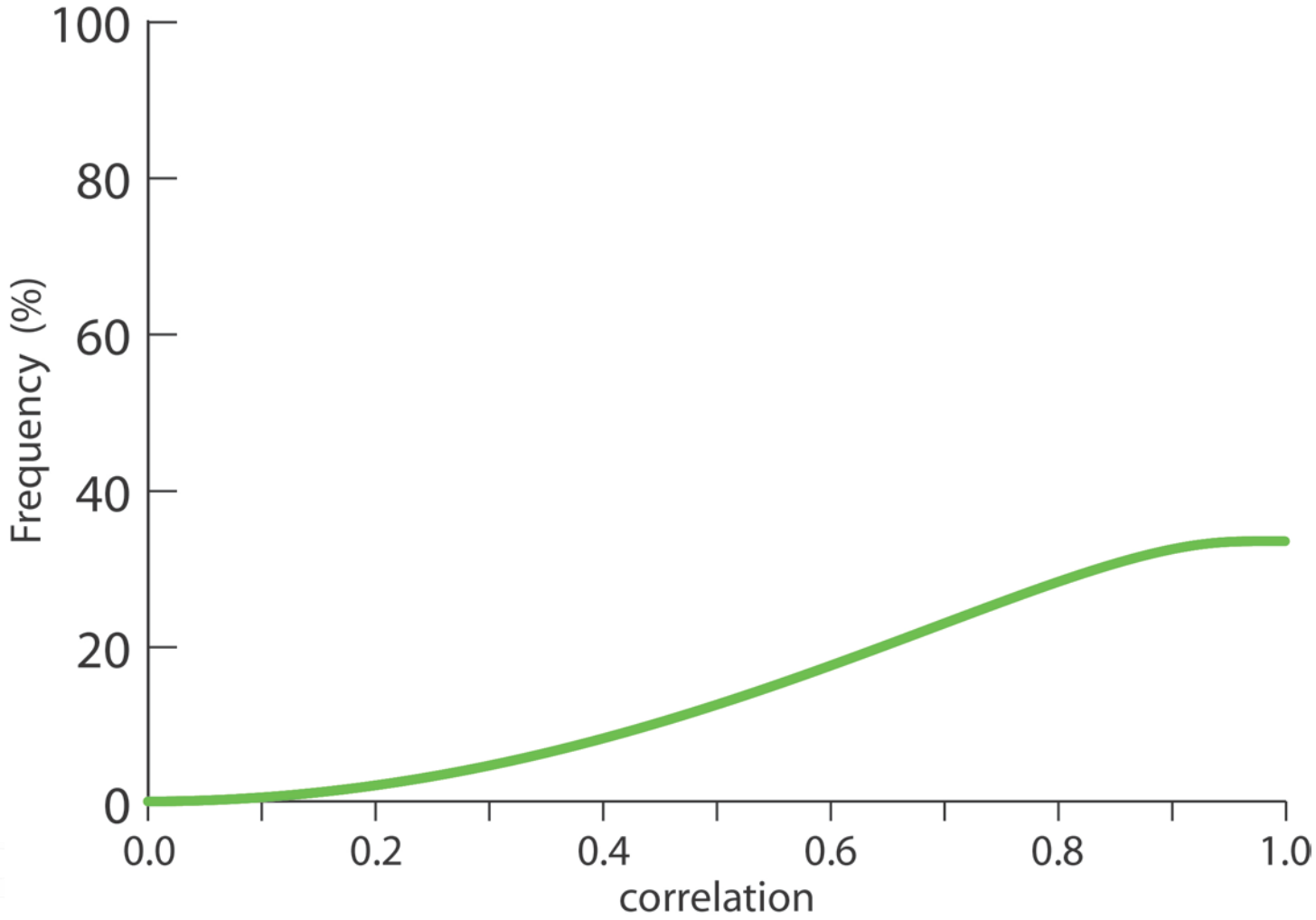
70 – 80% of all the African RCOF forecasts had highest probability on normal.



Maximum probability for “normal”



Frequency of “normal” as the most likely



Different verification questions

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

Attributes of a probabilistic forecast

But these attributes cannot be measured when we look only at one forecast:

Reliability the forecast does ***not*** mean that the category will occur over the indicated % of the area

Resolution it is impossible to estimate the observed frequencies for all the different forecasts

Discrimination it is impossible to discriminate if the whole region is dry

There is NO satisfactory way of verifying single probabilistic forecasts.

So what can we do?

Hit scores

Above	30
Normal	45
Below	25

If Above occurs is this a good forecast?

Combining categories creates all kinds of problems.

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

RCOF v1.4 or v2.0?
some radical suggestions

Hedging on normal needs to stop. Solutions?

- Training in reliable calibration

But that assumes that the problem is technical.

- Scrap the terciles (many RCOFs are effectively trying to forecast two categories)
- Scrap the probabilities in place of hit scores. The forecast probability, if reliable, is the hit score.