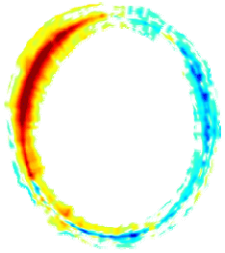
Presenting Forecast Verification Information to User Communities

Simon J. Mason

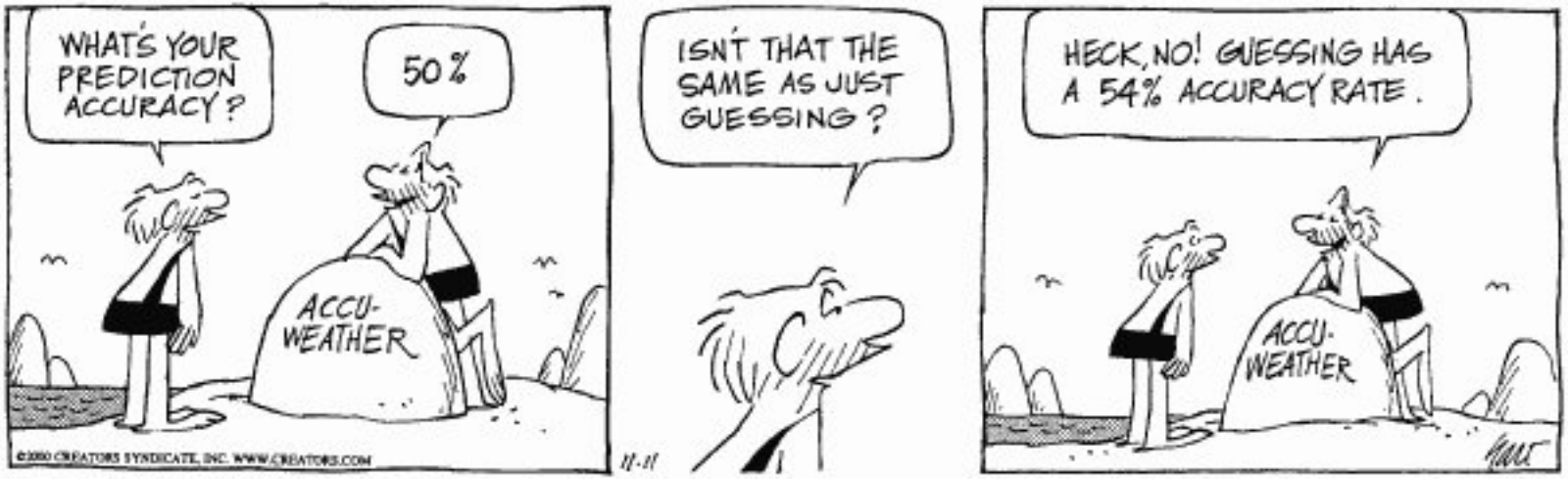
simon@iri.columbia.edu



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

APEC Climate Symposium 2008

Lima, Peru, 19 – 21 August, 2008



“How often are the forecasts correct?”

Introduction

“It is not possible to summarise adequately the quality of a set of forecasts in a single number.”

Murphy, 1991

“Any verification scheme should be … chosen to answer the questions of interest.”

Jolliffe and Stephenson, 2003

Introduction

- Users' interests in forecast verification can be classified into two levels of complexity:
  - *How should I use the forecasts?*

    The answer to this question is specific to each case, and cannot be answered generically.
  - *Should I even believe the forecasts anyway?*

    This question is preliminary to any economic evaluation of the forecasts: one would want to know whether the forecasts are potentially useful before conducting a more detailed assessment of how to optimize their value.
- So the initial question is: *Are the forecasts good*?

Two-Alternative Forced Choice Test

A 2AFC test is a test to correctly identify which of two options has a characteristic of interest. (One, and only one, of these choices should have the characteristic.)

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

If we have forecasts.

Year	Forecast
1965	25.8°C
1966	28.5°C

What is the probability of getting the answer correct now?

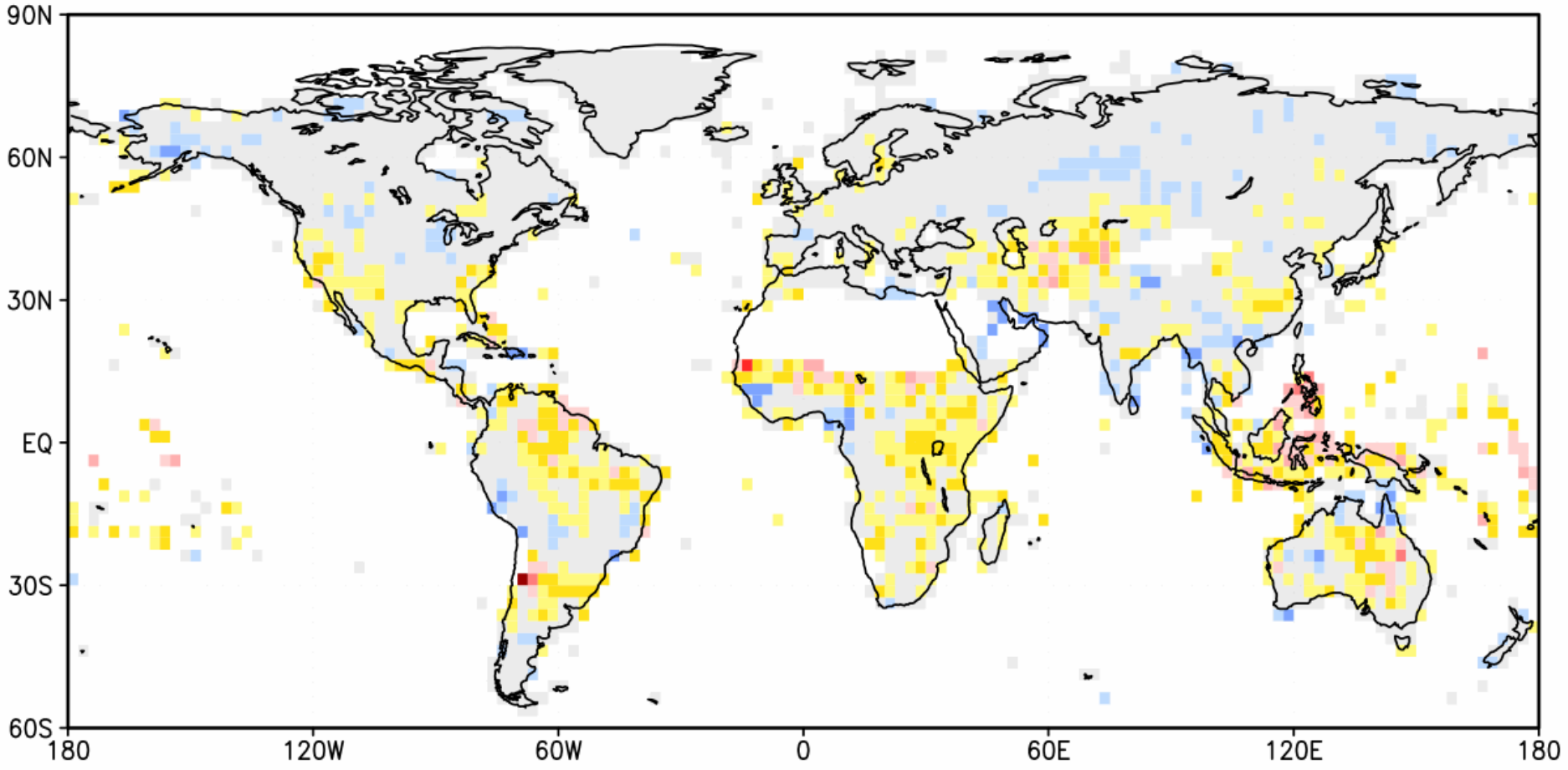
That again depends on whether we can believe the forecasts. Select the forecast with the higher probability.

Why the 2AFC score?

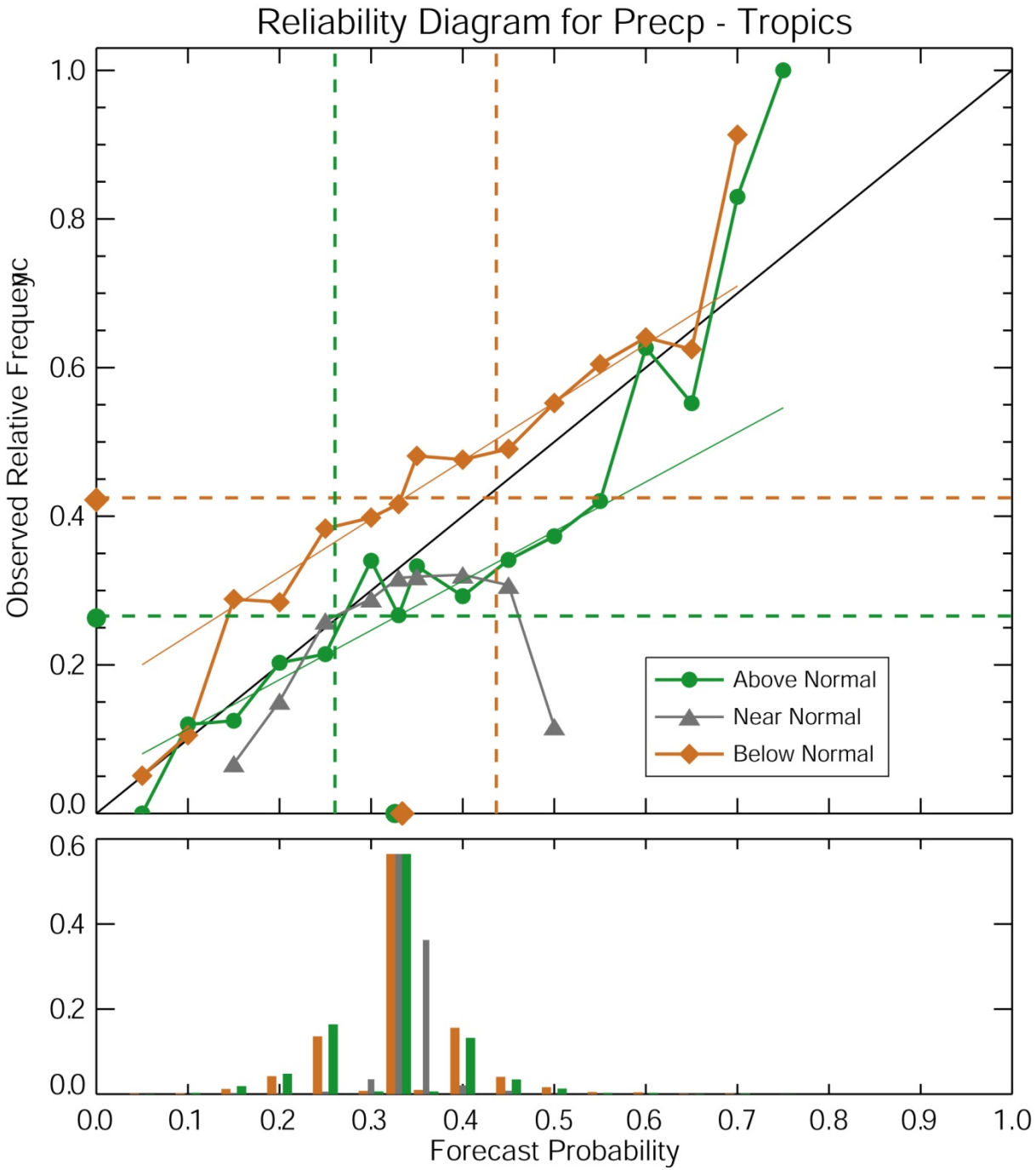
- It has an intuitive scale – 50% for no skill, 100% for perfect, 0% for perfectly bad.
- It has a simple interpretation – loosely understood as how often the forecasts are correct, but more strictly as how often the forecasts successfully distinguish different outcomes. It can be related to forecast questions such as: *will this season be wetter than last year?*
- It is equitable – the naïve forecasting strategies of guessing and perpetual forecasts (including “climatology”) both score 50%.
- It measures only one forecast attribute: “discrimination”, which is the most basic indicator of whether the forecasts contain any potentially usable information.
- It is highly flexible, and can be applied to all types of forecasts, and most types of observational data formats.

Precipitation forecast skill : ALL

Overall Discrimination (Generalized ROC)

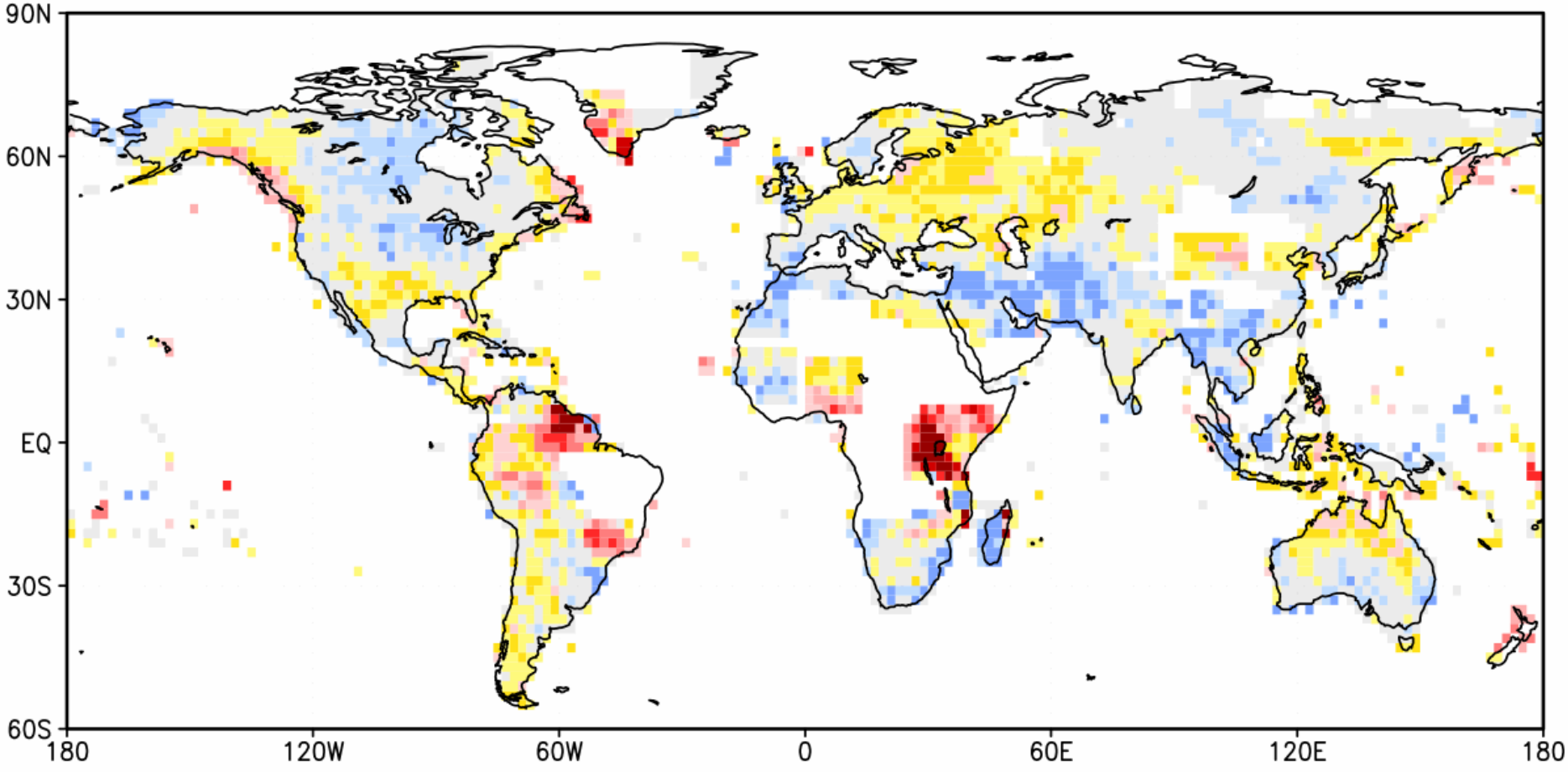


0.40 0.45 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90

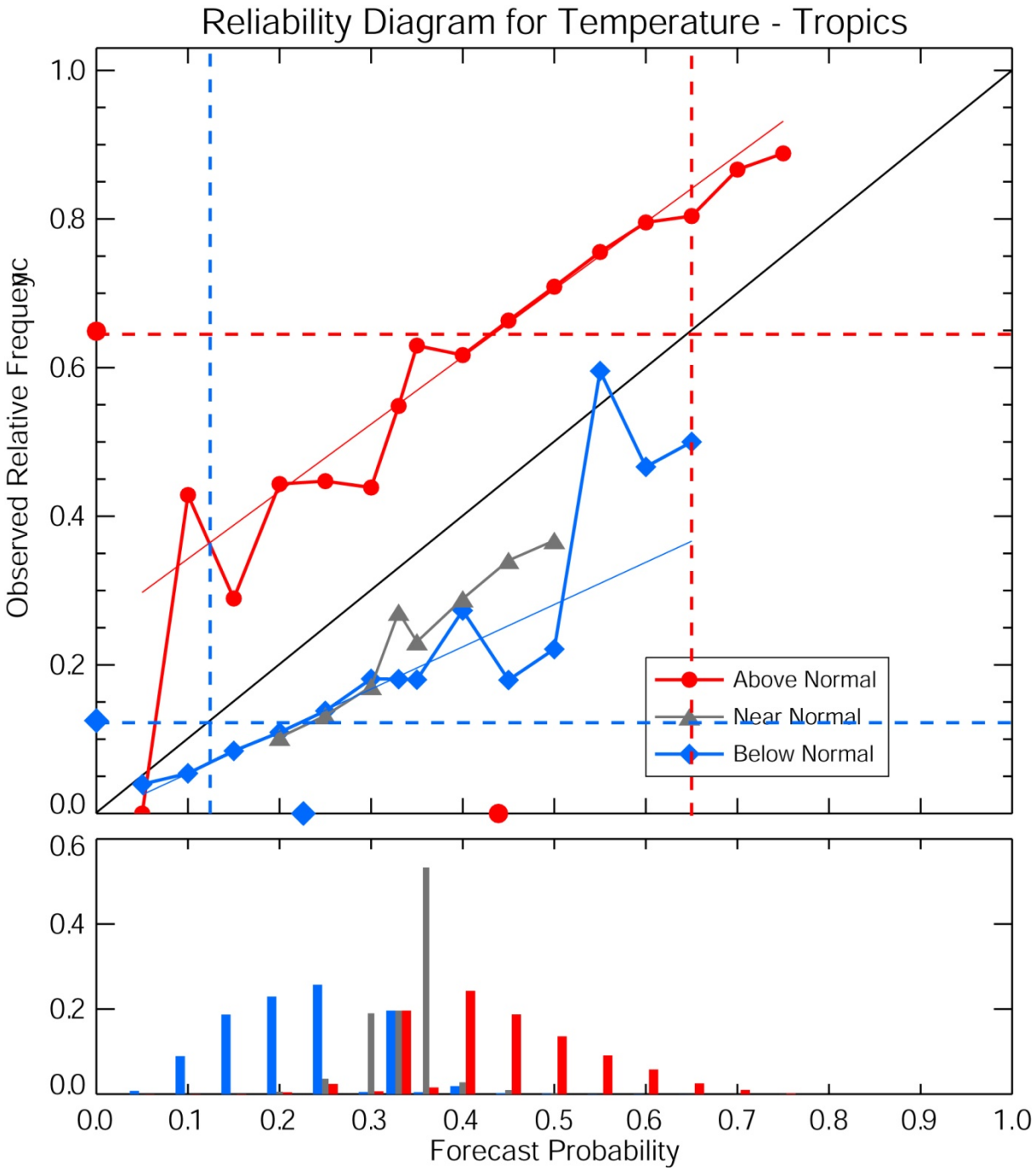

Reliability Diagram for Precp - Tropics
Observed Relative Frequency
Forecast Probability
Above Normal
Near Normal
Below Normal
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7
0.9

Temperature forecast skill : ALL

Overall Discrimination (Generalized ROC)



0.40 0.45 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90


Reliability Diagram for Temperature - Tropics
Observed Relative Frequency
Forecast Probability
Above Normal
Near Normal
Below Normal
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7
0.9