Verification of climate forecasts

How is forecast skill or accuracy measured?

What aspects of forecast quality is measured by various scores?

Some verification measures

D: deterministic

P: probabilistic

D Heidke skill score (a hit/miss score)

D Correlation with respect to means of sample being evaluated

D Correlation using climate means, not sample means

D Rank correlation

D Root mean squared error skill score (RMSSS)

P Ranked probability skill score (RPSS); a form of Brier Score

P Likelihood skill score (LSS)

P Rate of return

P ROC area

P ROC area generalized to all categories

P Reliability analysis

Heidke Skill Score (for deterministic categorical forecasts)

Probability forecasts for 3 tercile-based categories can be scored as if they are forecasts simply for the category having the highest probability. Then, depending on the obs category, it was either a *hit* or not.

Heidke skill score = $$100\left(\frac{\#Hits - \#Expected}{\#Total - \#Expected}\right)$$

Example: Suppose for OND 1997, rainfall forecasts are made for 15 stations in southern Brazil. Suppose the forecast is defined by tercile-based category having highest probability. Suppose for all 15 stations, “above” is forecast with highest probability, and that observations were above normal for 12 stations, and near normal for 3 stations. Then Heidke score is:

100 X (12 – 15/3) / (15 – 15/3)

100 X 7 / 10

= 70

Note that the forecast probabilities did not matter, only which category had highest probability. This conversion of probability forecasts to deterministic forecasts is a major **weakness.**

Correlation: Measuring the strength of linear relationship between two variables—here, between forecasts and their corresponding observations.

Let forecasts be x, and corresponding observations be y.

$$Correlation = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathrm{x}_i - \bar{\mathrm{x}})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{z}(x_i)\mathrm{z}(y_i)$$

Mean forecast biases and linear conditional forecast biases are ignored by the correlation. (When such biases exist, the correlation describes potential, but not actual, accuracy.) The correlation therefore measures *discrimination*.

$$Correlation = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathrm{x}_i - \bar{\mathrm{x}})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{z}(x_i)\mathrm{z}(y_i)$$

Two versions of correlation:

(1) When $\bar{\mathrm{x}}$ and $\bar{y}$ are means for *just the sample being correlated*

(2) When $\bar{\mathrm{x}}$ and $\bar{y}$ are means for *some other defined base period*

Version 2: Suppose there is a warming trend, and the 1971-2000 base period is used to define the climatology (means, anomalies, and tercile boundaries). Then by forecasting enhanced probabilities for “above normal”, a somewhat positive correlation is virtually guaranteed even if the year-to-year variations are not well discriminated. So, version 2 measures discrimination, and also calibration (freedom from mean bias and conditional biases).

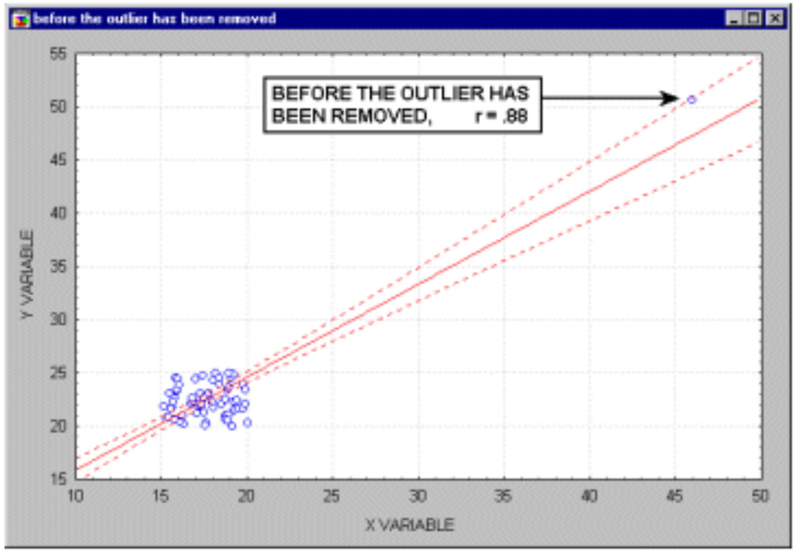
Version 1: If the means of the sample at hand are subtracted, then interannual variability becomes critical, and a positive correlation is not guaranteed by just forecasting “above normal” most of the time. Here, discrimination is measured.

Spearman rank correlation

Spearman rank correlation is the Pearson correlation between the *ranks* of X vs. the ranks of Y, treating ranks as numbers. Rank correlation measures the strength of ***monotonic*** relationship between two variables.

Rank correlation de-emphasizes outliers by not honoring original intervals between adjacent ranks. Adjacent ranks simply differ by 1. One or two influential cases become much less influential.

Original data:	2	9	189	3	21	7
Ranks:	6	3	1	5	2	4



Root-mean-Square Skill Score: RMSSS for **continuous deterministic** forecasts

Murphy (1988), Mon. Wea. Rev.

RMSSS is defined as:

$$RMSSS = 100 \times \left(1 - \left(\frac{RMS_f}{RMS_s} \right) \right)$$

where: $RMSE_f$ = root mean square error of forecasts, and $RMSE_s$ = root mean square error of a *standard* used as no-skill baseline.

Either persistence or climatology can be used as baseline. Persistence, for a given parameter, is the persisted anomaly from the forecast period immediately prior to the period being verified. For example, for example, for seasonal forecasts, persistence is the seasonal anomaly from the season prior to the season being verified. (For SST, it can be from the prior month.) Climatology is equivalent to a forecast having an anomaly of zero.

$$RMS_f = \sqrt{\frac{\sum_{i=1}^{N} \left[\left(f_i - O_i \right)^2 W_i \right]}{\sum_{i=1}^{N} W_i}}$$

$$\mathrm{RMS}_f = \sqrt{\frac{\sum_{i=1}^{N}\left[\left(f_i - O_i\right)^2 W_i\right]}{\sum_{i=1}^{N} W_i}}$$

where: i stands for a particular location (grid point or station).

f_i = forecasted anomaly at location i
O_i = observed or analyzed anomaly at location i.

W_i = weight at grid point i. When verification is done on a grid, to equalize grid areas, set weight using W_i = cos(latitude)

N = total number of grid points or stations where verification is carried.

RMSSS is given as a percentage, while RMS scores for f and for s are given in the same units as the verified parameter.

The RMS and the RMSSS are made larger by three main factors:

(1) The mean bias
(2) The conditional bias (including an amplitude bias)
(3) The lack of correlation between forecast and obs

(1) and (2) are calibration factors; (3) involves discrimination.

It is easy to correct for (1) using a hindcast history. This will improve the score. In some cases (2) can also be removed, or at least decreased, and this will improve the RMS and the RMSSS farther. Improving (1) and (2) does not improve (3). It is most difficult to increase (3). If the tool is a dynamical model, a spatial MOS correction can increase (3), and help improve RMS and RMSSS.

Verification of Probabilistic Categorical Forecasts: The **Ranked Probability Skill Score (RPSS)**

Epstein (1969), J. Appl. Meteor.

RPSS measures cumulative squared error between categorical forecast probabilities and the observed categorical probabilities relative to a reference (or standard baseline) forecast.

The observed categorical probabilities are 100% in the observed category, and 0% in all other categories.

$$RPS = \sum_{cat=1}^{Ncat} (Pcum_{F(cat)} - Pcum_{O(cat)})^2$$

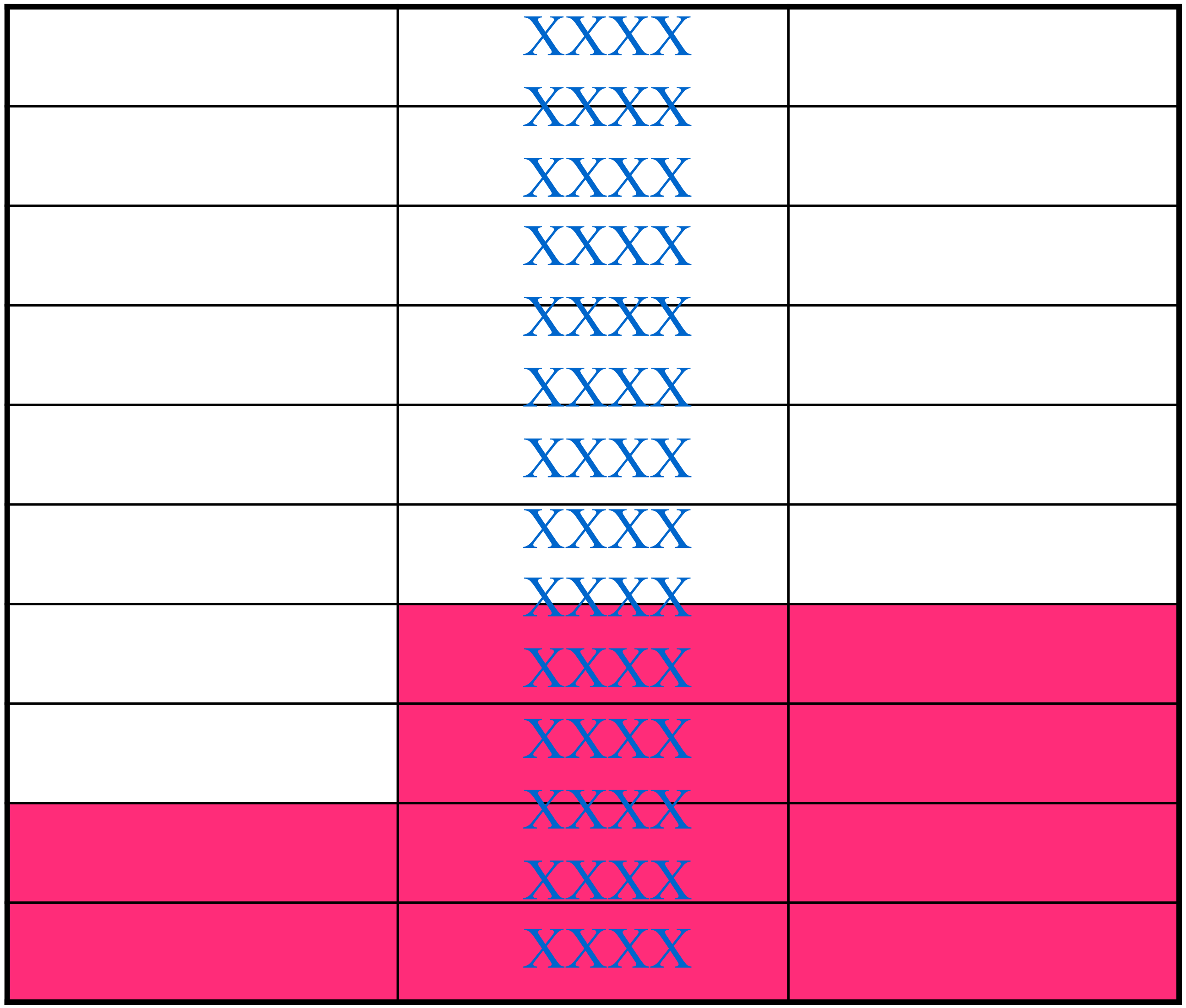
Where Ncat = 3 for tercile forecasts. The “cum” implies that the summation is done first for cat 1, then cat 1 and 2, then cat 1 and 2 and 3.

RPSS

Probability forecast (B,N,A):

.20 .40 .40

Non-cumulative
Shown in Table



.00 1.00 .00

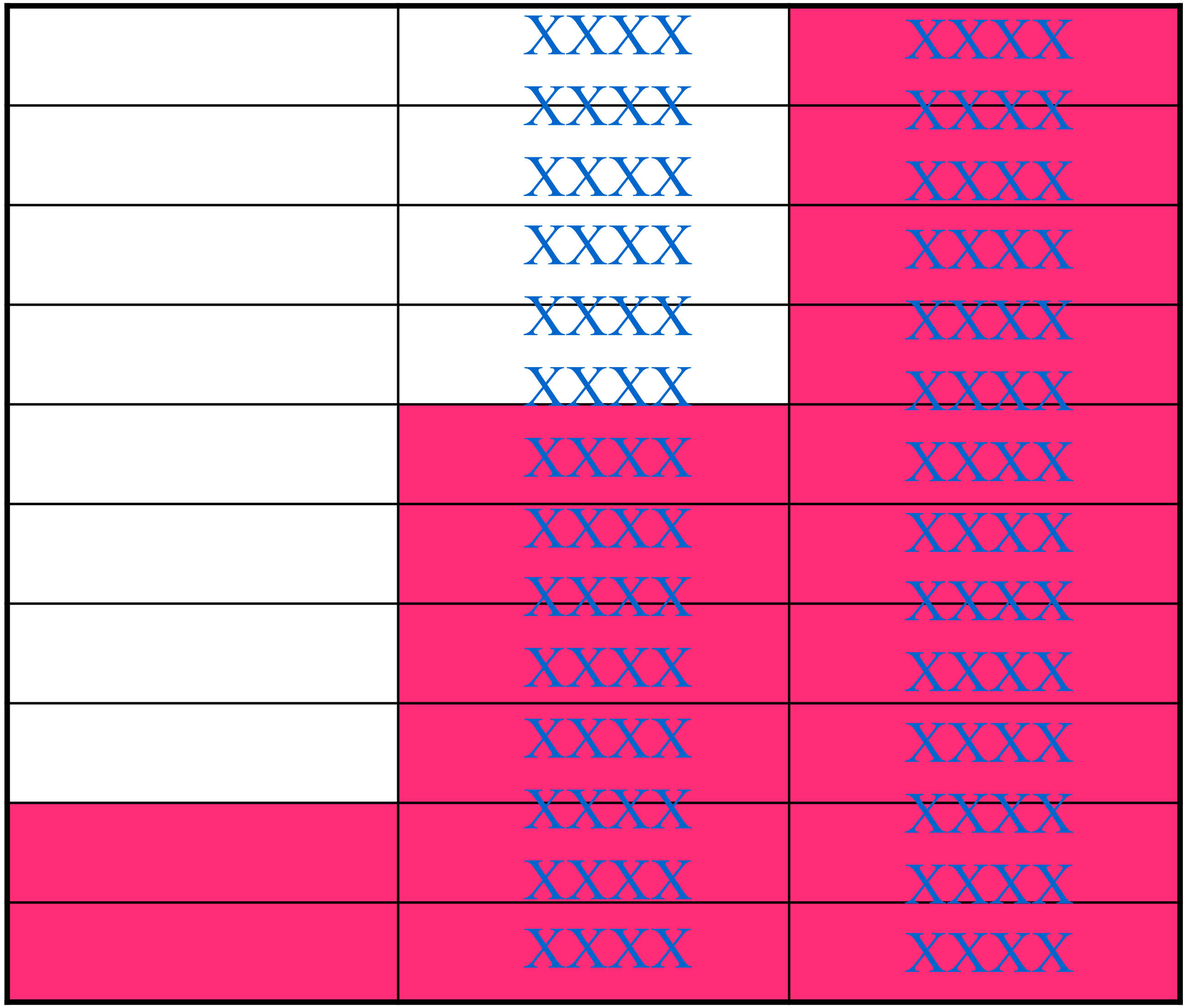
Observation “Probability” (N occurred)

RPSS

Probability forecast (B,N,A):

.20 .40 .40

Cumulative
Shown in Table



.00 1.00 .00

Observation “Probability” (N occurred)

RPS(fcst) =

$(.20 - .00)^2$
$+ (.60 - 1.00)^2$
$+ (1.00-1.00)^2$

= .04 + .16
= .20

:

RPS(clim) =

$(.333 - .00)^2$
$+ (.667 - 1.00)^2$
$+ (1.00-1.00)^2$

= .111 + .111
= .2222

RPSS =
1 – (.20 / .222)
= .10

$$RPS = \sum_{cat=1}^{Ncat} (Pcum_{F(cat)} - Pcum_{O(cat)})^2$$

The higher the RPS, the poorer the forecast. RPS=0 means that the probability given to the category that was observed was 100%.

The RPSS is based on the RPS for the forecast compared to the RPS For a **reference forecast** such as one that simply gives climatological probabilities.

$$RPSS = 1 - \frac{RPS_{f\ orecast}}{RPS_{ref\ erence}}$$

RPSS > 0 when RPS for actual forecast is smaller than RPS for the reference forecast.

Suppose that the forecast probabilities for terciles for 15 stations in OND 1997 in Kenya, and the observations were:

	forecast(%)			obs(%)			RPS calculation
1	20	30	50	0	0	100	$RPS=(0-.20)^2+(0-.50)^2+(1.-1.)^2=.04+.25+.0=.29$
2	25	35	40	0	0	100	$RPS=(0-.25)^2+(0-.60)^2+(1.-1.)^2=.06+.36+.0=.42$
3	25	35	40	0	0	100	
4	20	35	45	0	0	100	$RPS=(0-.20)^2+(0-.55)^2+(1.-1.)^2=.04+.30+.0=.34$
5	15	30	55	0	0	100	
6	25	35	40	0	0	100	
7	25	35	40	0	100	0	$RPS=(0-.25)^2+(1-.60)^2+(1.-1.)^2=.06+.16+.0=.22$
8	25	35	40	0	0	100	
9	20	35	45	0	0	100	
10	25	35	40	0	0	100	
11	25	35	40	0	100	0	
12	20	35	40	0	100	0	
13	15	30	55	0	0	100	$RPS=(0-.15)^2+(0-.45)^2+(1.-1.)^2=.02+.20+.0=.22$
14	25	35	40	0	0	100	
15	25	35	40	0	0	100	

Finding RPS for reference (climatology baseline) forecasts:

for 1st forecast, $RPS(clim)=(0-.33)^2+(0-.67)^2+(1.-1.)^2=.111+.444+0=.556$

for 7th forecast, $RPS(clim)=(0-.33)^2+(1.-.67)^2+(1.-1.)^2=.111+.111+0=.222$

for a forecast whose observation is “below” or “above”, PRS(clim)=.556

forecast(%)		obs(%)	RPS and RPSS(clim)		RPSS
1	20 30 50	0 0 100	RPS= .29	RPS(clim)= .556	1-(.29/.556) = .48
2	25 35 40	0 0 100	RPS= .42	RPS(clim)= .556	1-(.42/.556) = .24
3	25 35 40	0 0 100	RPS= .42	RPS(clim)= .556	1-(.42/.556) = .24
4	20 35 45	0 0 100	RPS= .34	RPS(clim)= .556	1-(.34/.556) = .39
5	15 30 55	0 0 100	RPS= .22	RPS(clim)= .556	1-(.22/.556) = .60
6	25 35 40	0 0 100	RPS= .42	RPS(clim)= .556	1-(.42/.556) = .24
7	25 35 40	0 100 0	RPS= .22	RPS(clim)= .222	1-(.22/.222) = .01
8	25 35 40	0 0 100	RPS= .42	RPS(clim)= .556	1-(.42/.556) = .24
9	20 35 45	0 0 100	RPS= .34	RPS(clim)= .556	1-(.34/.556) = .39
10	25 35 40	0 0 100	RPS= .42	RPS(clim)= .556	1-(.42/.556) = .24
11	25 35 40	0 100 0	RPS= .22	RPS(clim)= .222	1-(.22/.222) = .01
12	20 35 40	0 100 0	RPS= .22	RPS(clim)= .222	1-(.22/.222) = .01
13	15 30 55	0 0 100	RPS= .22	RPS(clim)= .556	1-(.22/.556) = .60
14	25 35 40	0 0 100	RPS= .42	RPS(clim)= .556	1-(.42/.556) = .24
15	25 35 40	0 0 100	RPS= .42	RPS(clim)= .556	1-(.42/.556) = .24

Finding RPS for reference (climatol baseline) forecasts:

When obs=“below”, RPS(clim) = $(0-.33)^2+(0-.67)^2+(1.-1.)^2$ =.111+.444+0=.556

When obs=“normal”, RPS(clim)=$(0-.33)^2+(1.-.67)^2+(1.-1.)^2$ =.111+.111+0=.222

When obs=“above”, RPS(clim)= $(0-.33)^2+(0-.67)^2+(1.-1.)^2$ =.111+.444+0=.556

RPSS for various tercile probability forecasts,
when **observation is “above”.**

Forecasted Tercile Probabilities			
-	0	+	RPSS
100	0	0	-2.60
90	10	0	-2.26
80	15	5	-1.78
70	25	5	-1.51
60	30	10	-1.11
50	30	20	-0.60
40	35	25	-0.30
33	33	33	0.00
25	35	40	0.24
20	30	50	0.48
10	30	60	0.69
5	25	70	0.83
5	15	80	0.92
0	10	90	0.98
0	0	100	1.00

Note: issuing overly confident forecasts causes high penalty when incorrect. Skills are highest for “true” probabilities, which would be revealed in a reliability plot.

The RPSS is made worse by three main factors:

(1) Mean probability biases
(2) Conditional probability biases (including amplitude biases)
(3) The lack of correlation between forecast probabilities and obs

(1) and (2) are calibration factors; (3) involves discrimination.

The RPSS (Epstein 1969, J. Appl. Meteor.) is an extension of the **Brier Score** (Brier, 1950, Mon. Wea. Rev.), which is the same calculation except for only two categories. The tercile category system can be seen as a two category system if the two tercile boundaries are considered one at a time:

below normal vs. not below normal
above normal vs. not below normal.

The likelihood score

The likelihood score is the nth root of the product of the probabilities given for the event that was later observed. for example, using terciles, suppose 5 forecasts were given as follows, and the category in red was observed:

45	35	20
33	33	33
40	33	27
15	30	55
20	40	40

The likelihood score disregards what probabilities were forecast for categories that did not occur.

The likelihood score for this example (n=5) would be

$$\sqrt[5]{(0.35)(0.33)(0.40)(0.55)(0.40)} = \sqrt[5]{0.0102} = \mathbf{0.40}$$

This score could then be scaled such that 0.333 would be 0%, and 1 would be 100%. A score of 0.40 would translate to (0.40 - 0.333) / (1.00 - 0.333) = 10.0%.

The likelihood skill score

If the likelihood score comes out to be $\sqrt[5]{0.0102}$ **= 0.40,**

then a likelihood skill score can be computed using a no-skill reference forecast:

$$LSS = \frac{LS_{forecast} - LS_{reference}}{1 - LS_{reference}}(100)$$

Then, LS=0.40 is positioned on a scale where 0.333 becomes 0%, and 1 becomes 100%:

$$LSS = \frac{0.40 - 0.3333}{1 - 0.3333}(100)$$ **= 10%**

The LSS is used in verifying the IRI's probability forecasts (Barnston et al. 2010). It is closely related to the *ignorance score* (Roulston and Smith, Mon. Wea. Rev., 2001) and derived scores such as *rate of return (*Hagedorn and Smith, Meteor. Applic., 2008; Tippett and Barnston, Mon. Wea. Rev., 2008).

The rate of return

The rate of return also begins with the nth root of the product of the probabilities given for the event that was later observed. Using terciles, suppose 5 forecasts were again as follows, and the category in red was observed:

45 35 20
33 33 33
40 33 27
15 30 55
20 40 40

The rate of return disregards what probabilities were forecast for categories that did not occur.

The rate of return for this example (n=5) begins as:

$$\sqrt[5]{(0.35)(0.33)(0.40)(0.55)(0.40)} = \sqrt[5]{0.0102} = \mathbf{0.40}$$

The rate of return will then divide this geometric mean by that which would result if 0.3333 were forecast perpetually: that is, 0.40 / 0.3333.

The rate of return

If the geometric mean comes out to be $\sqrt[5]{0.0102}$ **= 0.40,**

then the rate of return (ROR) is computed using a no-skill reference forecast:

$$ROR = \left[\frac{GeomMean_{forecast}}{GeomMean_{reference}} - 1\right](100)$$

So, 0.40 is divided by the same computation for climatology forecasts (0.3333), and then 1 is subtracted, and the result is multiplied by 100 to give a percentage:

$$ROR = \left[\frac{0.40}{0.3333} - 1\right](100) = \mathbf{20\%}$$

For tercile-based forecasts, then, the rate of return is exactly double LSS, so that "perfect" forecasts would result in a rate of return of 200%.

Use of logarithm in likelihood score and rate of return

Both the likelihood skill score and the rate of return involve computing a geometric mean. When there are many forecasts, multiplying many numbers less than 1 can result in underflow. Therefore, the logs are computed and summed, followed by division by n, and then taking the antilog (exponentiation). Hence the term “log likelihood”.

Then

$$\sqrt[5]{(0.35)(0.33)(0.40)(0.55)(0.40)}$$

would become

$$e^{\left[\frac{\ln(0.35)+\ln(0.33)+\ln(0.40)+\ln(0.55)+\ln(0.40)}{5}\right]}$$

The rate of return seen in the literature uses base 2 instead of the base e (2.718), and defines the ignorance score as $-\log_2$(prob)

The likelihood skill score and the rate of return are both made worse by three main factors:

(1) Mean probability biases
(2) Conditional probability biases (including amplitude biases)
(3) The lack of correlation between forecast probabilities and obs

(1) and (2) are calibration factors; (3) involves discrimination.

Relative Operating Characteristics (ROC) for Probabilistic Forecasts

Mason, I. (1982) Australian Met. Magazine

An event, such as observing “below normal” precipitation, is defined.

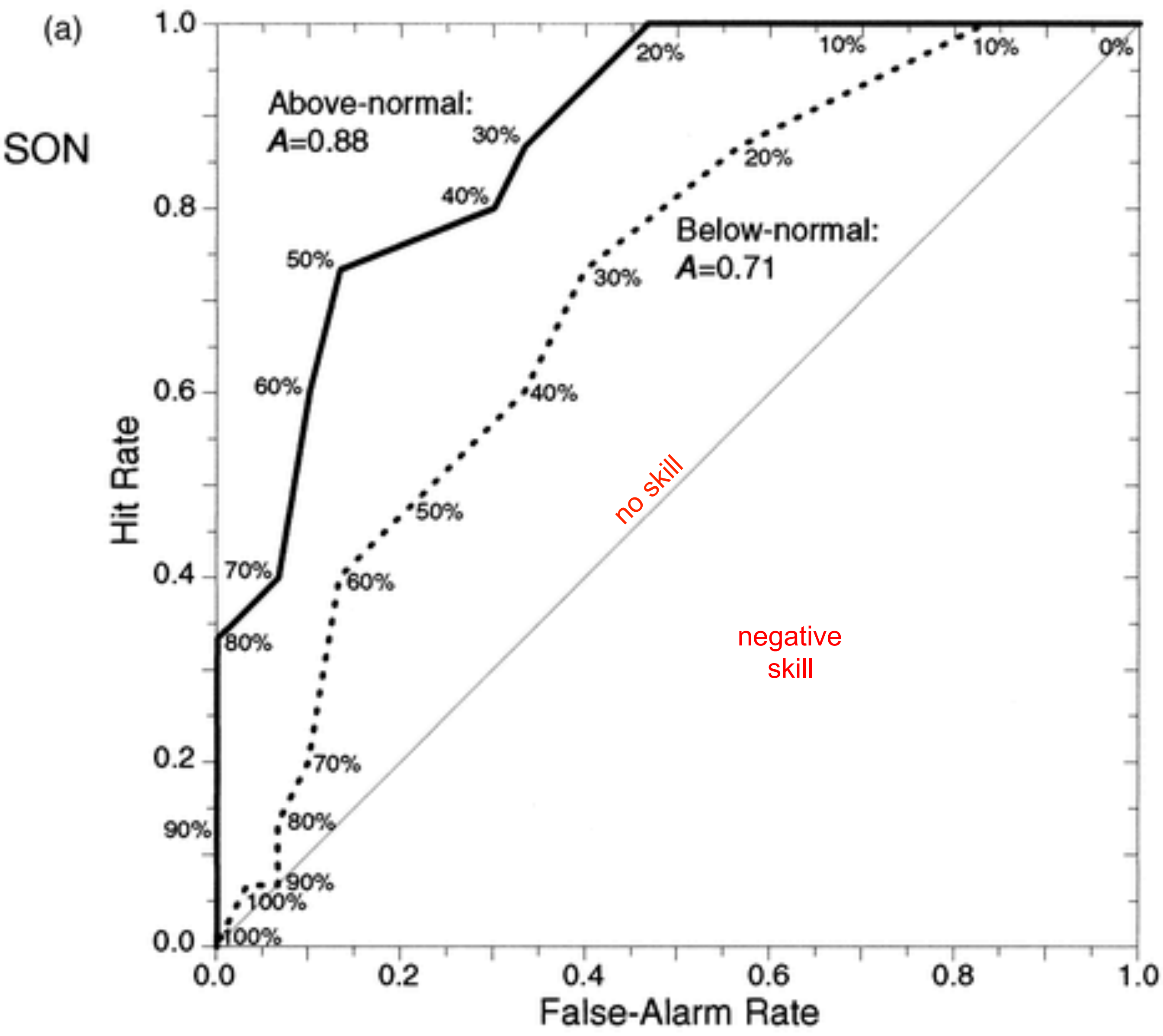
The contingency table that ROC verification is based on:

	Observation Yes	Observation No
Forecast: Yes	**O_1** (hit)	NO_1 (false alarm)
Forecast: NO	O_2 (miss)	**NO_2** (correct rejection)

Hit Rate = $0_1 / (O_1+O_2)$

False Alarm Rate = $NO_1 / (NO_1+NO_2)$

The hit rate and false alarm rate are determined for descending categories of forecast probability, cumulatively. For high forecast probabilities, we hope hit rate rate will be high and false alarm rate low; and for low forecast probabilities, we hope hit rate will be low and false alarm rate will be high. For in-between probabilities, we expect some of each.



The curves are cumulative from left to right. For example, “30%” really means “100% + 90% +80% + ….. +30%”.

Curves farther to the upper left show greater skill, since hit rate should be higher for the higher probabilities alone.

Example from Mason and Graham (2002), QJRMS, for eastern Africa OND simulations (observed SST forcing) using ECHAM3 AGCM

Mean forecast probability biases and linear conditional forecast probability biases are ignored by the ROC. For example, all of the probabilities for “below normal” could be double what they should be, or 35% higher than they should be, and ROC would be the same as if they did not have that problem. But changes in the probability might still correspond to the same directional changes in the observed result.

The ROC therefore measures *discrimination*.

Generalized ROC area is integrated over categories
(Mason and Weigel 2009; Mon. Wea. Rev.).

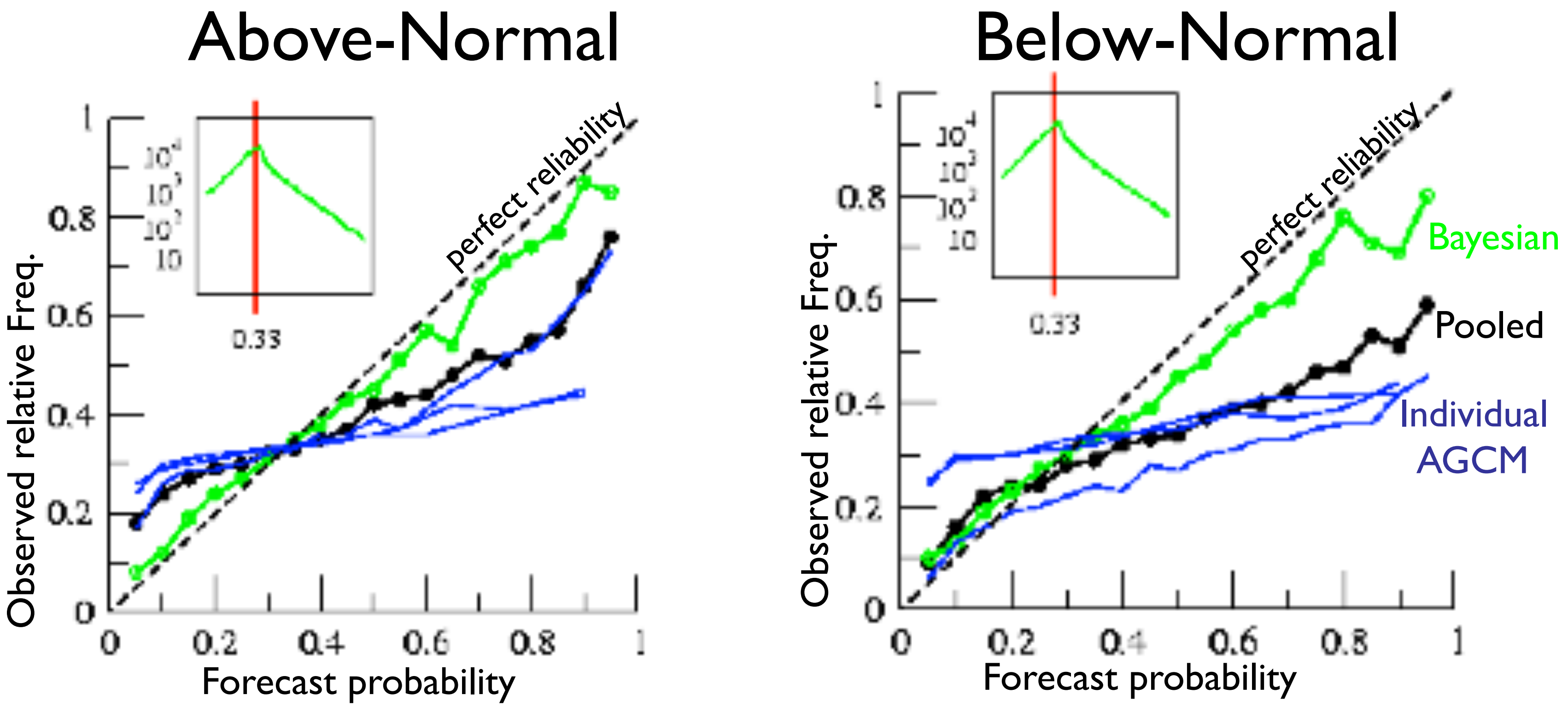
$$\text{GROC} = \frac{\sum_{k=1}^{m-1} \sum_{l=k+1}^{m} \sum_{i=1}^{n(k)} \sum_{j=1}^{n(l)} I[(P_{ki}(k), P(_{lj}(k)]}{\sum_{k=1}^{m-1} \sum_{l=k+1}^{m} n_k n_l}$$

Number of unique forecast pairs having differing obs results

Where $$I[(P_{ki}(k), P(_{lj}(k)] = \begin{cases} 0 & \text{if } P_{lj}(k) < P_{ki}(k) \\ 0.5 & \text{if } P_{lj}(k) = P_{ki}(k) \quad \text{(tie)} \\ 1 & \text{if } P_{lj}(k) > P_{ki}(k) \end{cases}$$

Each pair of forecasts having differing observation categories is examined to see if the forecast for the higher observed category case was “higher” than that for the lower observed category. The I[…] tells if so.

Results of Bayesian multi-model ensemble are evidenced in an analysis of **reliability** (the correspondence between forecast probability and relative observed frequency of occurrence). Simple pooling (assignment of equal weights to all AGCMs) gives more reliability than that of individual AGCMs, but the Bayesian method produces even more reliability. Note that flattish lines show model overconfidence, and 45º line shows perfect reliability.



(3-model) JAS Precipitation, 30S-30N

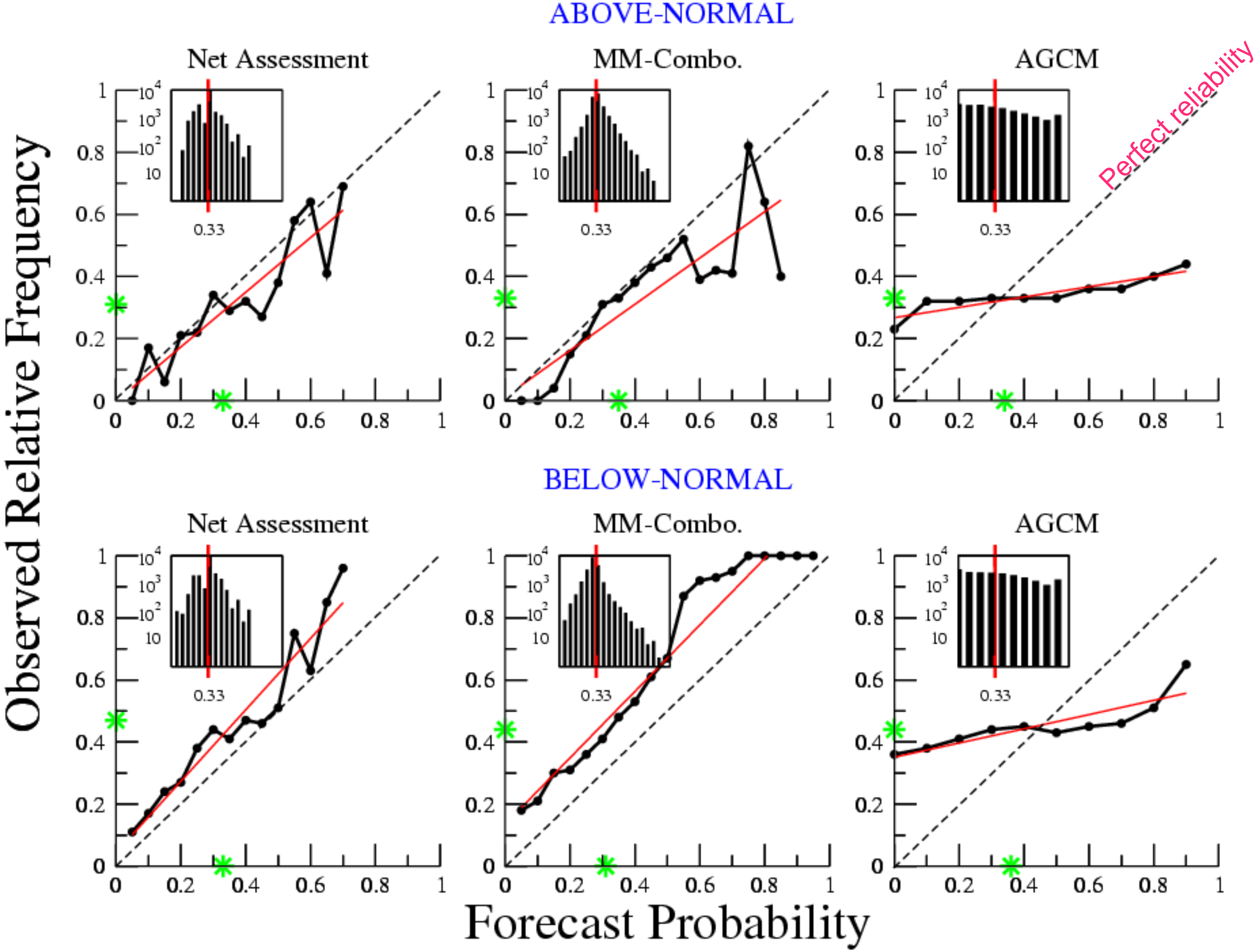
from Goddard et al. 2003

EGS-AGU-EGU Joint Assembly, Nice, France, April

RELIABILITY ANALYSIS

Precipitation Forecasts (30S-30N)

RELIABILITY DIAGRAM



ABOVE-NORMAL

Net Assessment

MM-Combo.

AGCM

BELOW-NORMAL

Net Assessment

MM-Combo.

AGCM

Perfect reliability

Use of Multiple verification scores is encouraged.

We have seen that different skill scores emphasize different aspects of skill. It is usually a good idea to use more than one score, and determine more than one aspect of skill. A reliability plot is also informative.

At least one score that measures overall quality (discrimination and calibration together), such as RPSS, or likelihood, or rate or return, and one score that measures discrimination alone (ROC), is encouraged. When discrimination is good but calibration (biases) is making overall quality less good, a reliability diagram can reveal the nature of the calibration problems.