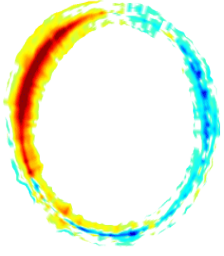
Do statistical models trade resolution for reliability?

Simon J. Mason

simon@iri.columbia.edu



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

ECMWF Seminar on Seasonal Prediction

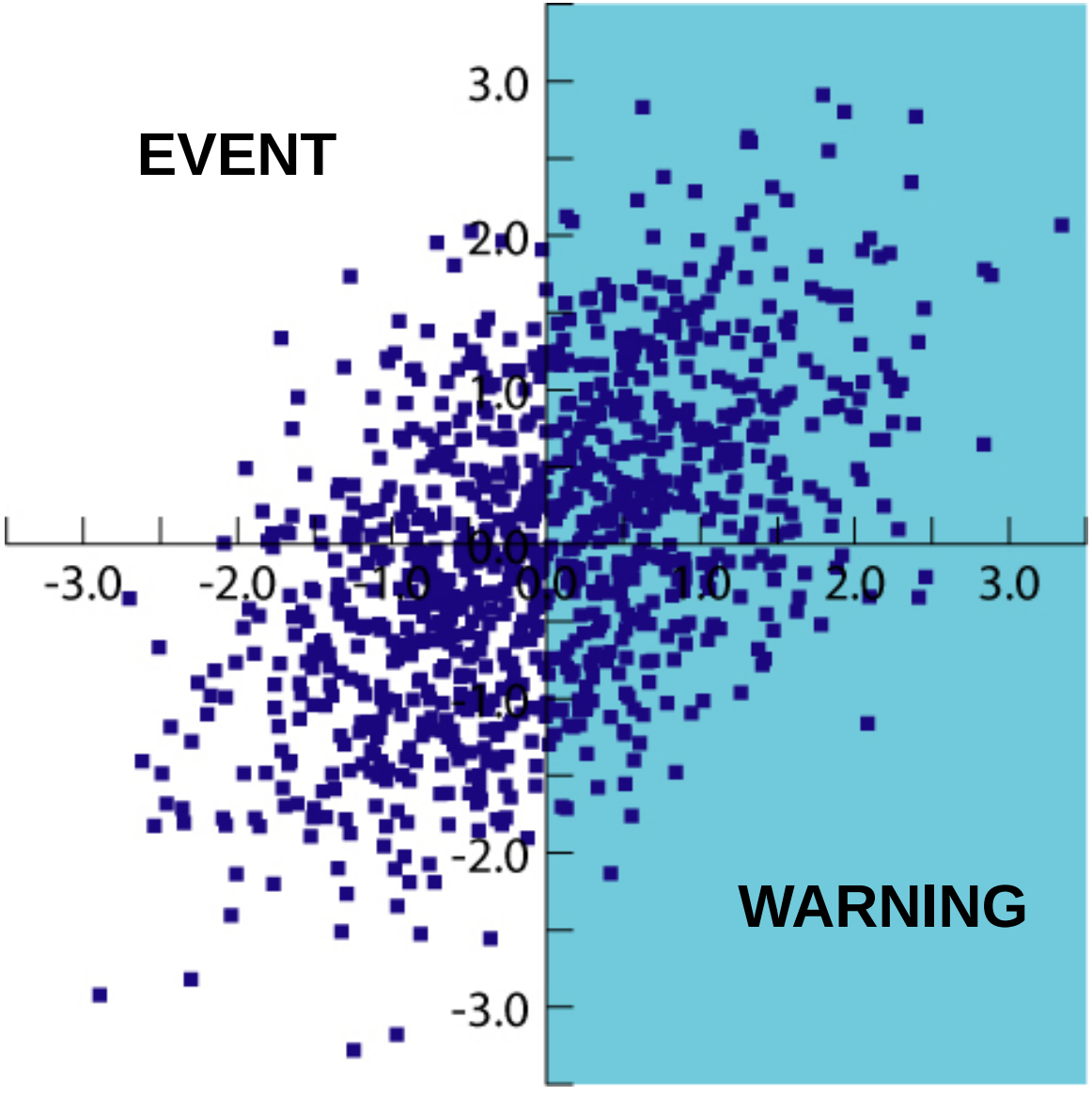
Shinfield Park, England, 3 – 7 September 2012

Definitions

"I don't like definitions"
Mark Knopfler

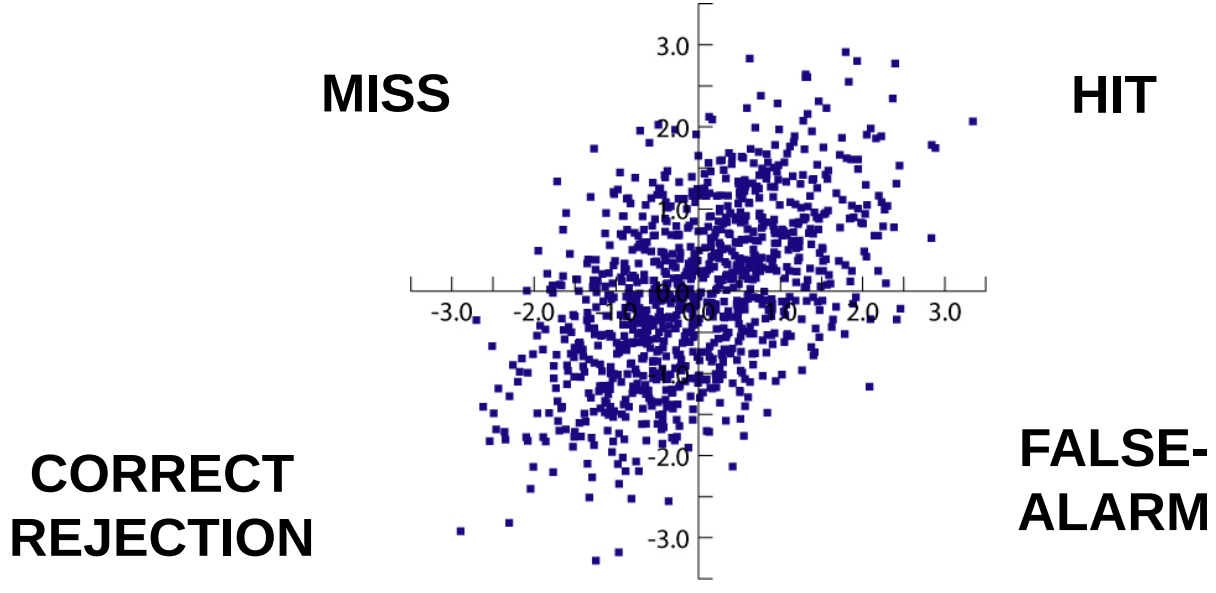
- *Reliability*
  - the outcome occurs as frequently as indicated
- *Resolution*
  - the outcome differs for different forecasts
- *Discrimination*
  - the forecast differs for different outcomes
- *Sharpness*
  - the forecasts differ sometimes

Measuring attributes: Deterministic forecasts

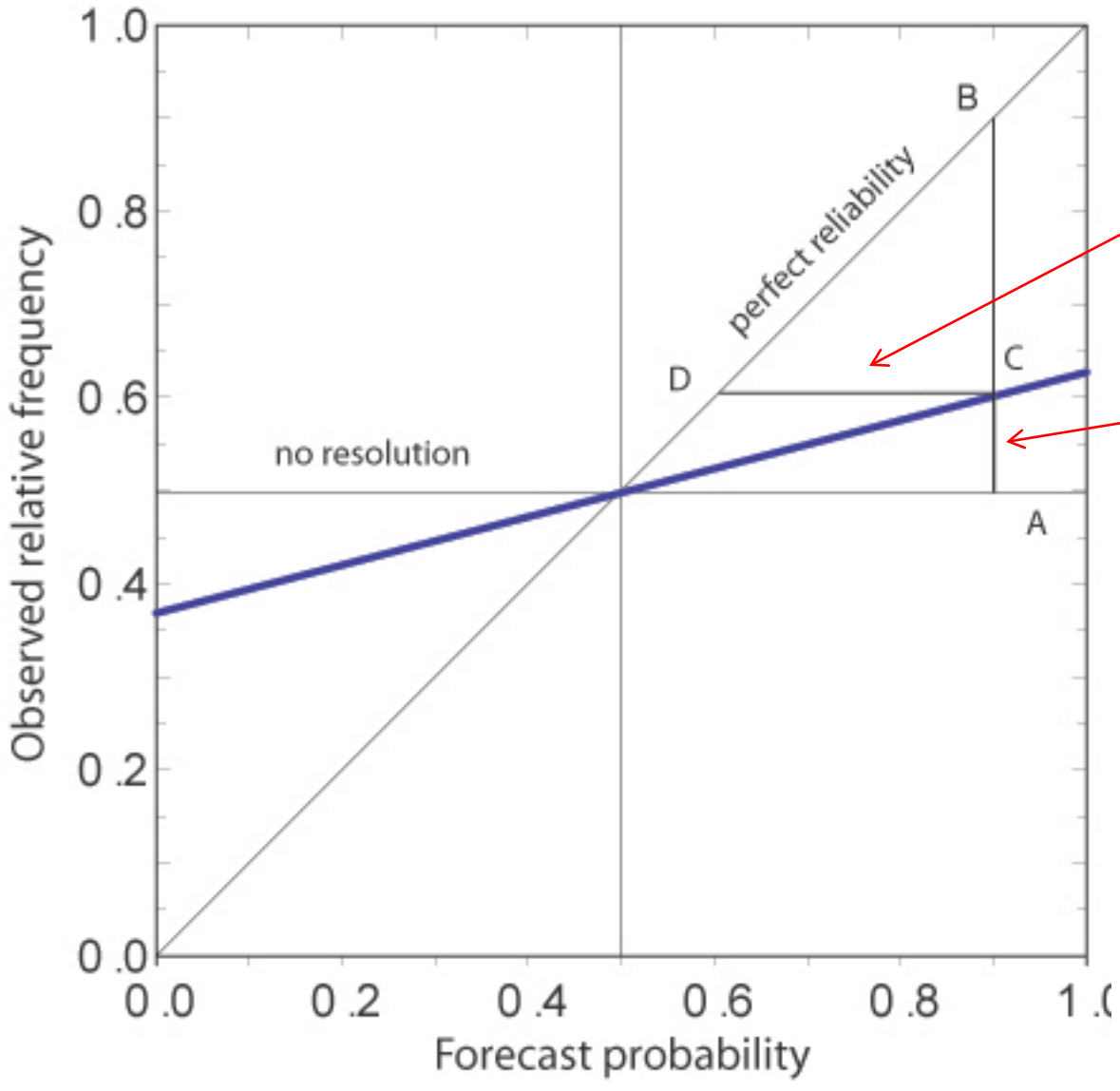


Measuring attributes: Deterministic forecasts

- *Reliability*: number of warnings = number of events
- *Discrimination*: hit rate > false-alarm rate
- *Resolution*: correct-alarm rate > miss rate (for equi-probable 2-category systems, resolution = discrimination).



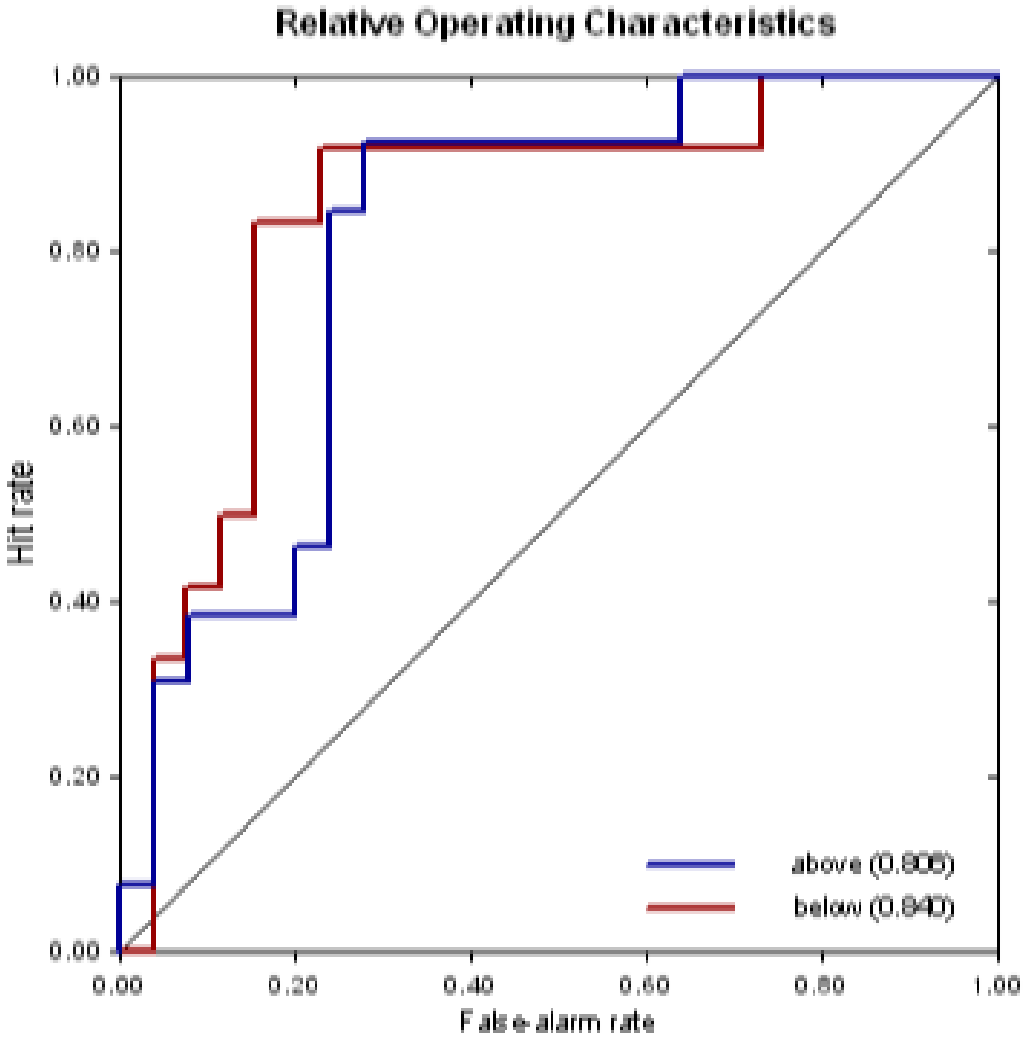
Measuring attributes: Probabilistic forecasts



reliability

resolution

Measuring attributes: Probabilistic forecasts



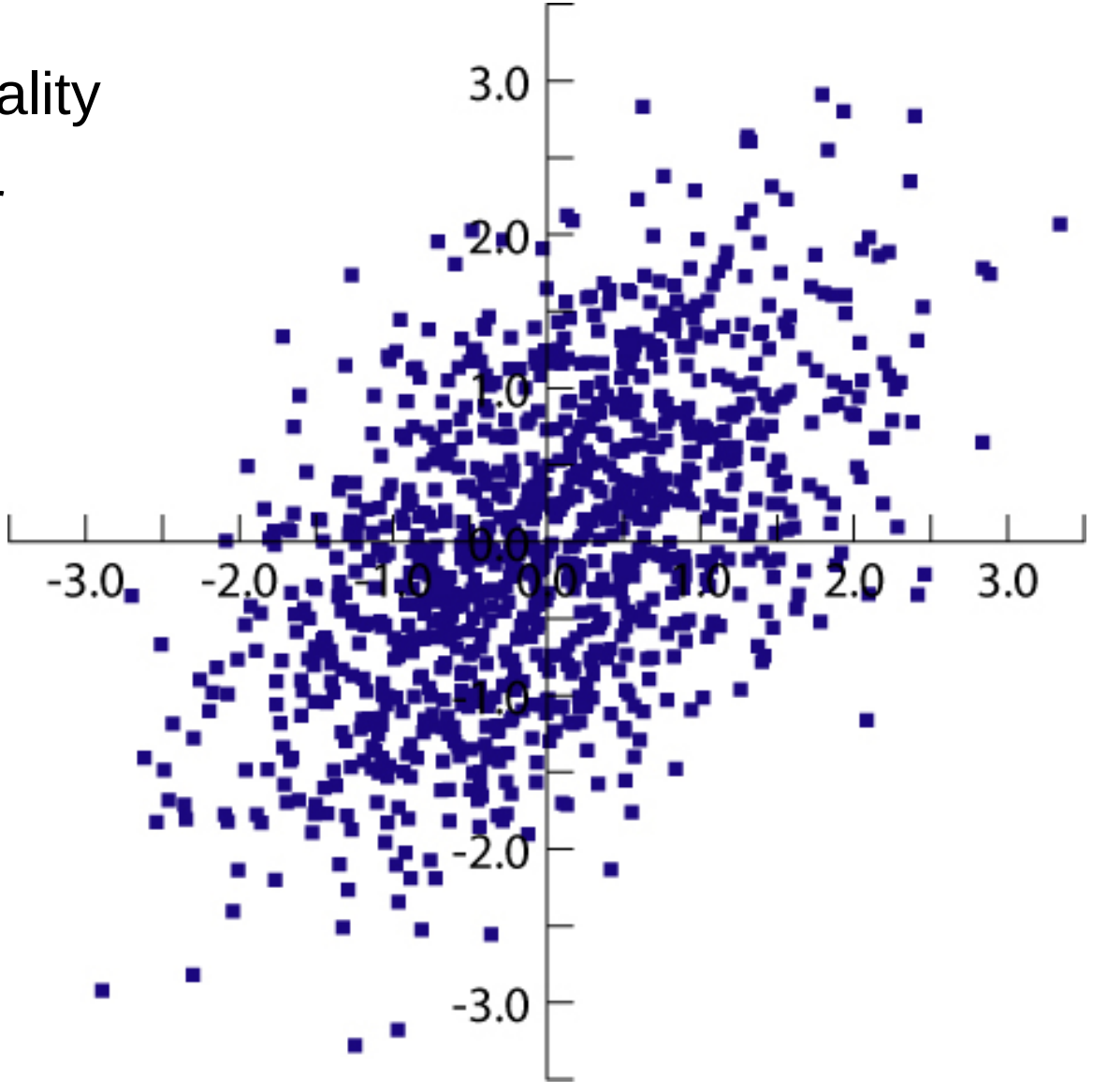
Discrimination:

Area beneath the curve indicates the probability of successfully discriminating an event from a non-event.

Idealised forecasts

Bivariate-normality

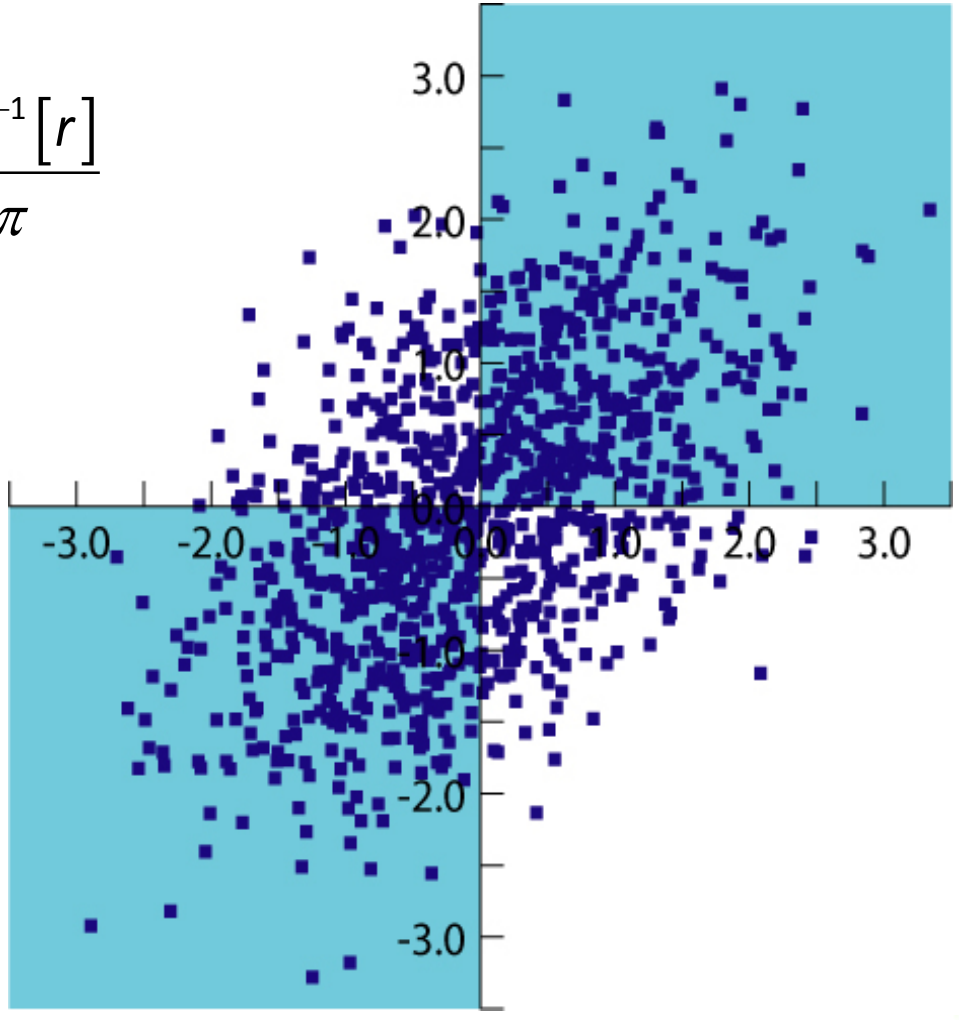
Correlation of *r*



2-category deterministic scores

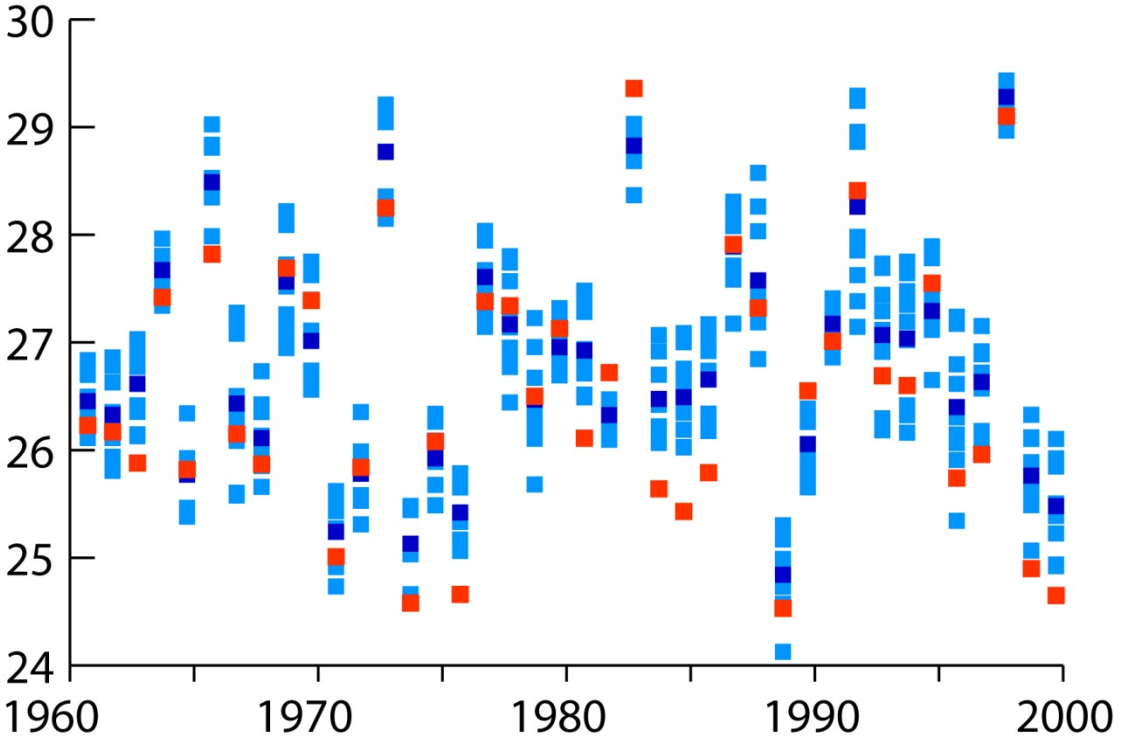
$$\text{Proportion correct} = 0.5 + \frac{\sin^{-1}[r]}{\pi}$$

$$\text{Peirce skill score} = \frac{2}{\pi}\sin^{-1}[r]$$



Data

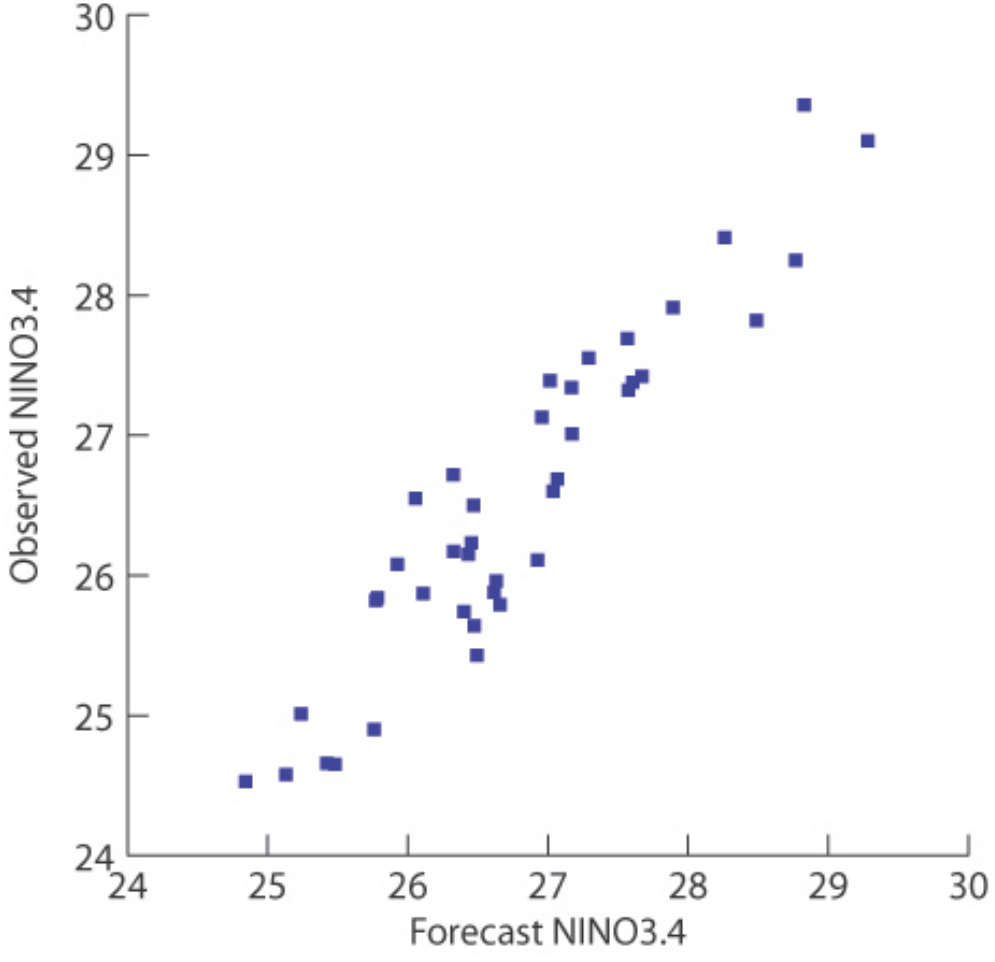
CNRM predictions of the January 1961–2000 Niño3.4 index from the previous August; 9 ensemble members.



	Observed	CNRM
Average (°C)	26.5	26.8
Std. dev. (°C)	1.21	1.04 (mean) 1.10 (ensemble)

Data

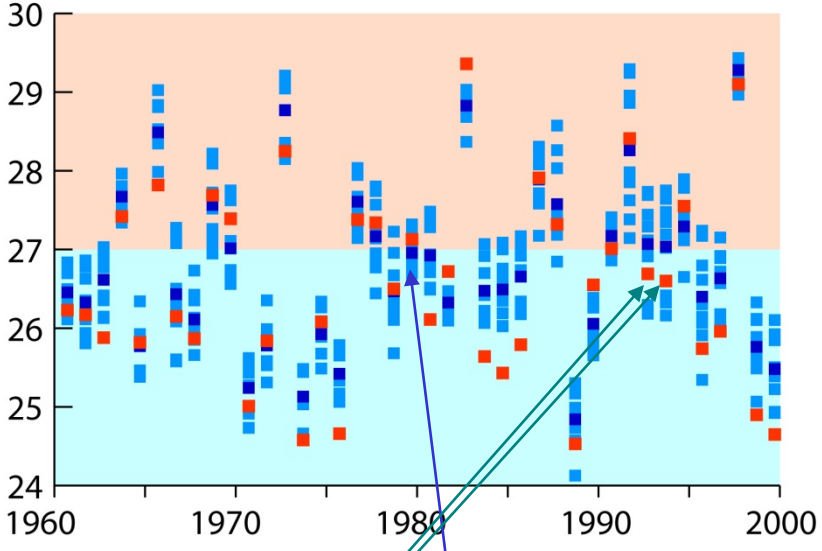
Pearson's correlation for the ensemble-mean is about 0.94.



2-category deterministic predictions

For the CNRM forecasts, consider forecasts of a "warm" event (Niño3.4 index > 27°C):

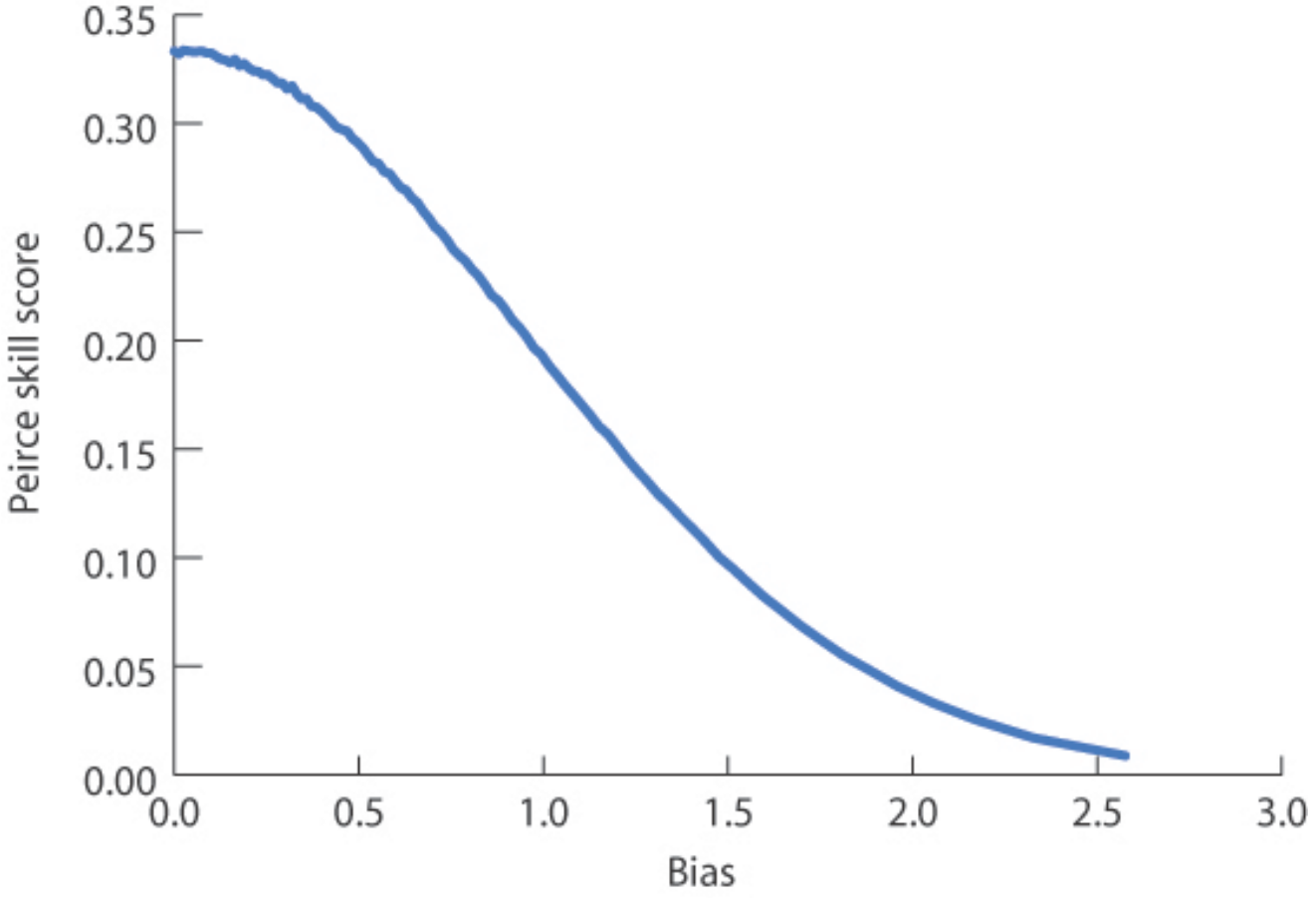
Probability of correctly discriminating "warm" from "cool" is about 93% (PSS=86%).



OBSERVATIONS	FORECASTS	
	>27°C	<27°C
>27°C	14	1
<27°C	2	23

Effects of bias

Effects of bias on discrimination, given *r*=0.5.



Correcting for a bias in the mean *should* improve discrimination.

Effects of bias

After correcting for the bias in the mean, the probability of correctly discriminating “warm” (>27°C) from “cool” (<27°C) drops to about 89% (PSS=78%).

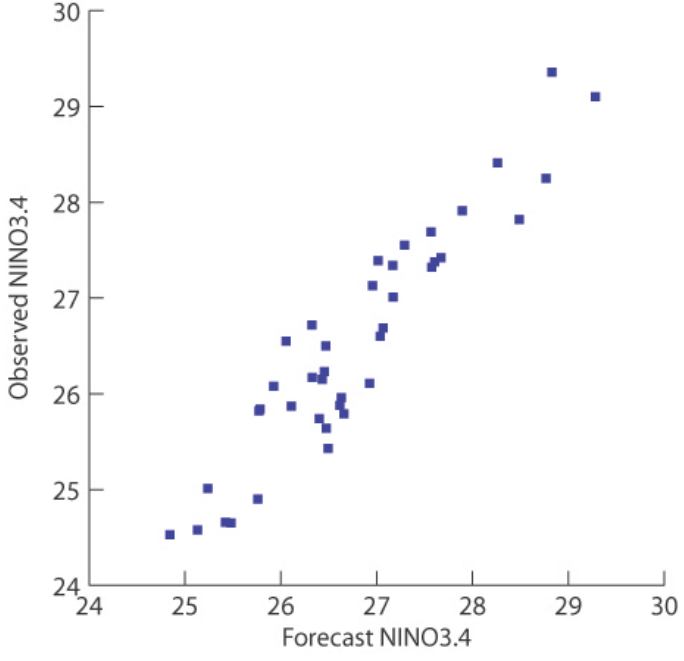
OBSERVATIONS	FORECASTS	
	>27°C	<27°C
>27°C	13	2
<27°C	2	23

Even the simplest recalibration schemes can result in a loss of discrimination.

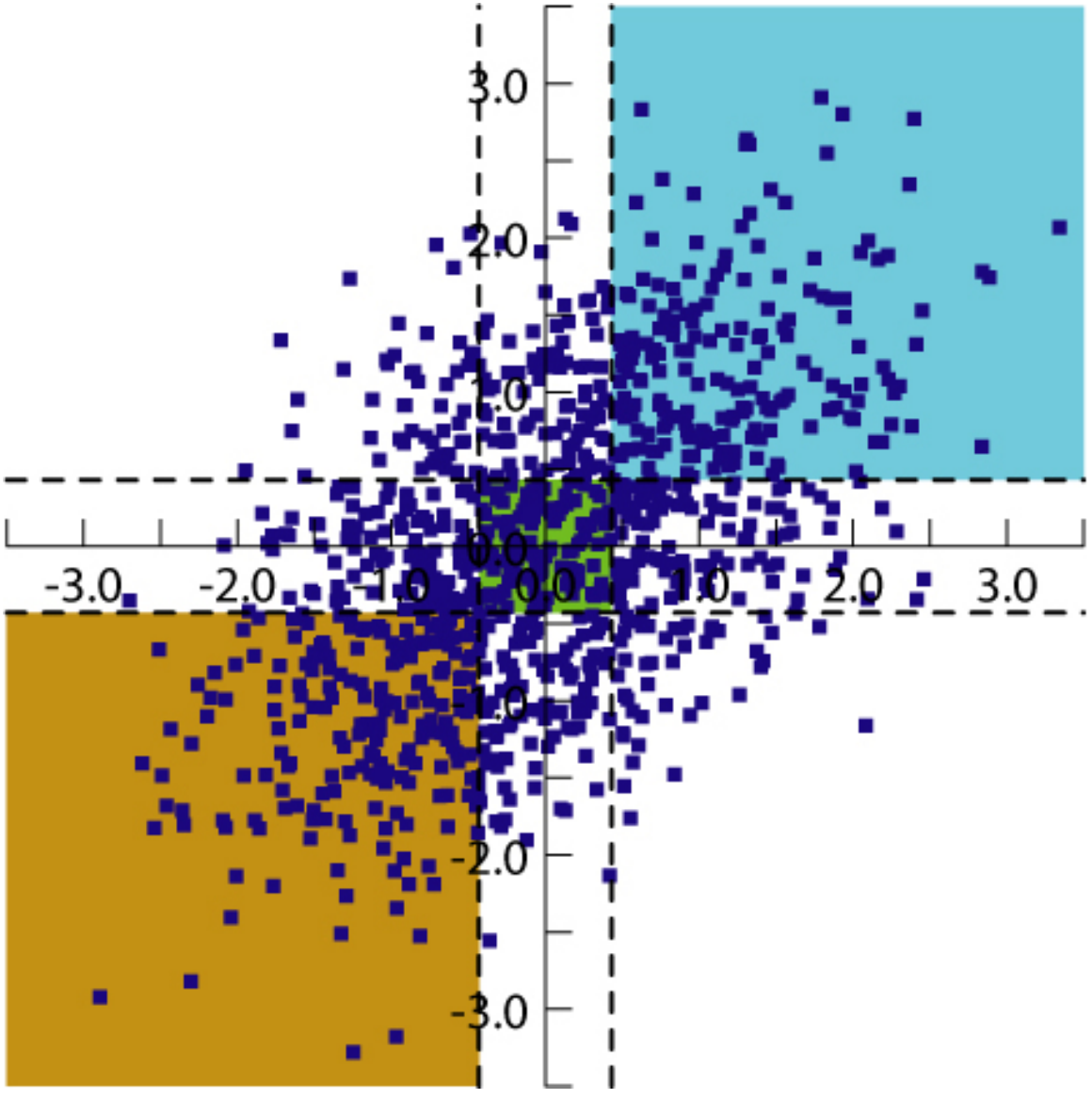
2-category deterministic scores

BUT:

If any of the assumptions of bivariate normality are violated, the relationships between the scores are not preserved, and calibration *may* come at the cost of resolution / discrimination.



3-category deterministic predictions



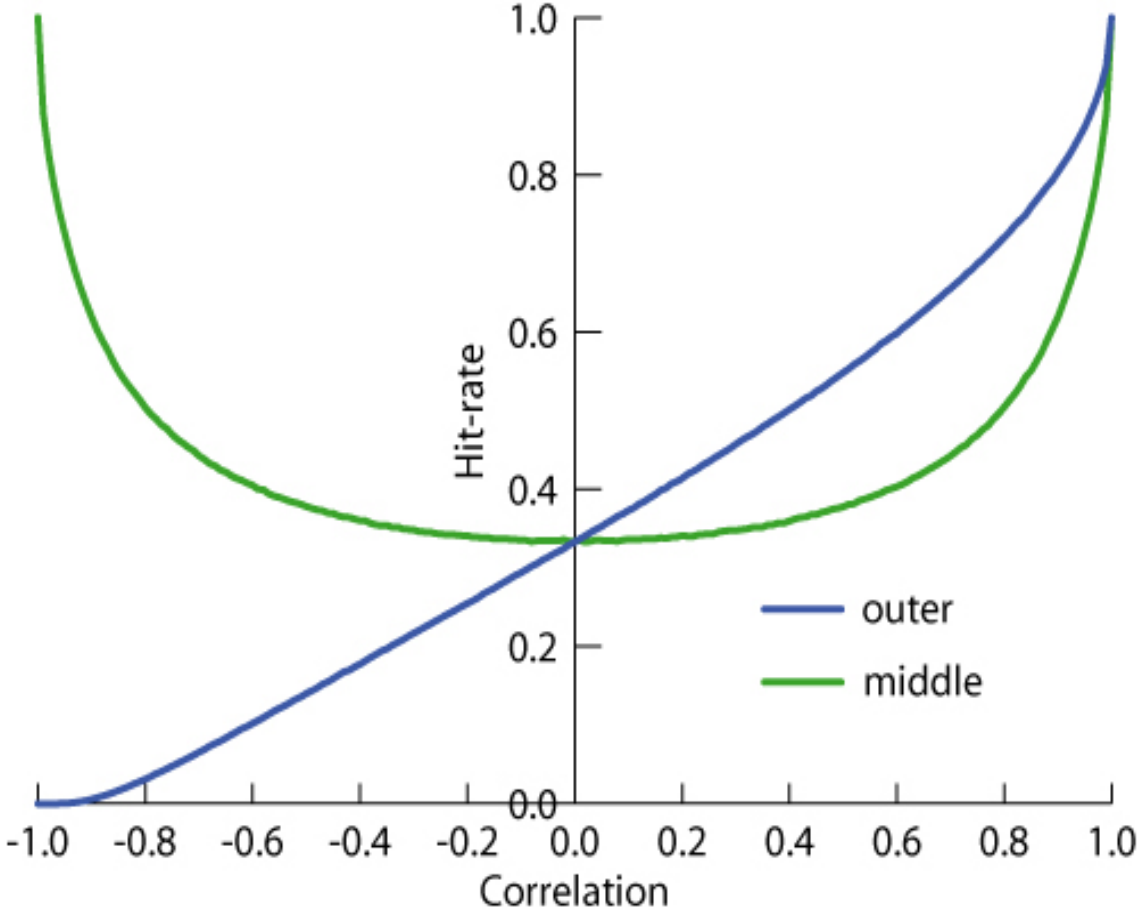
3-category deterministic predictions

In a three-category forecast system, the probabilities of correctly discriminating “above” (>27.13°C) “normal, and “below” (<25.88°C) using the uncorrected forecasts are:

OBS.	Scores (%)	
	Uncorrected	Corrected
A	94	91
N	83	81
B	79	79

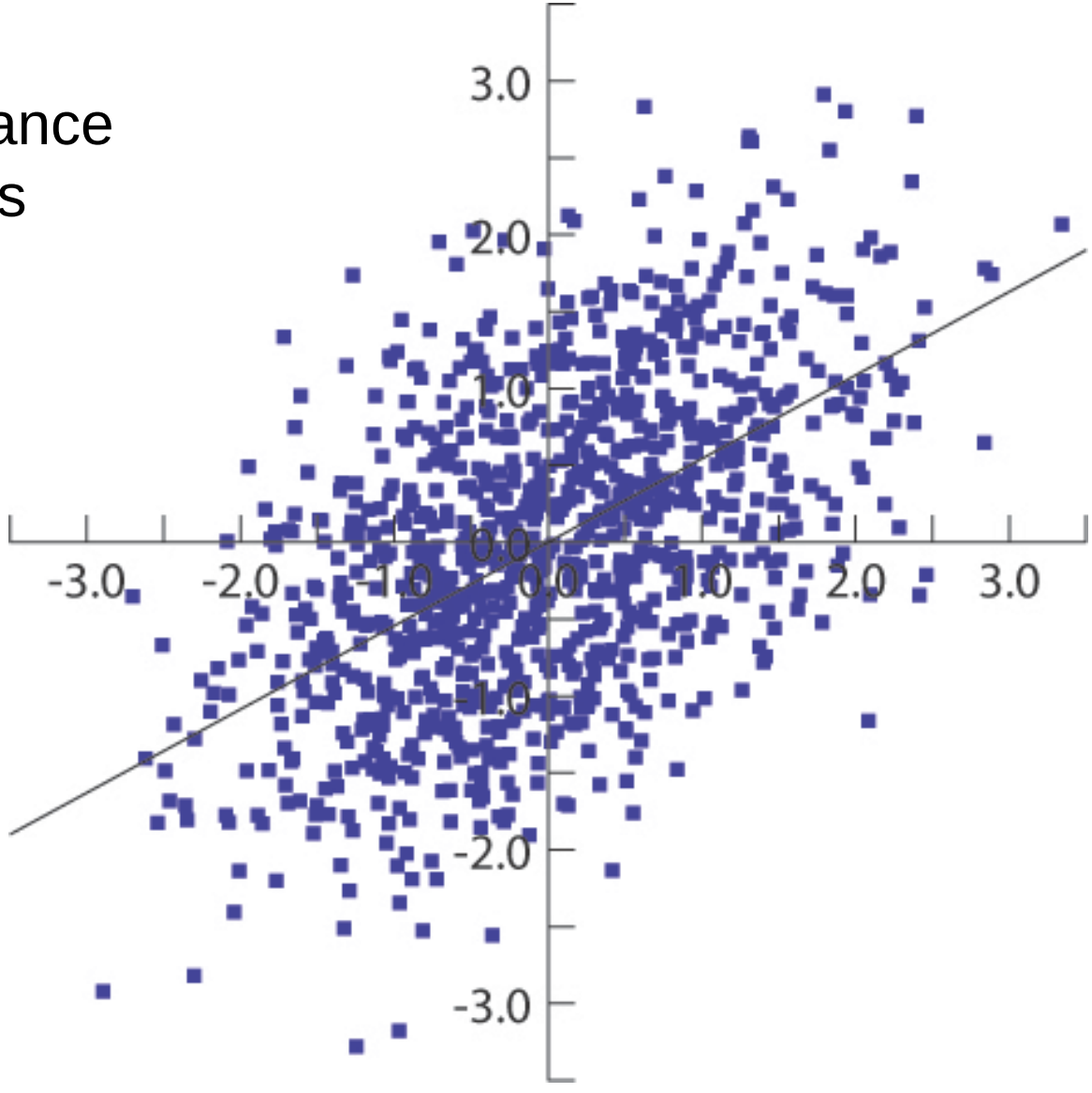
3-category deterministic predictions

The middle category (“normal”) has a lower hit-rate than the outer-categories (“above”, “below”) for $0 < r < 1$. The hit rate for “normal” is only marginally skillful unless r is very strong.



Probabilistic forecasts

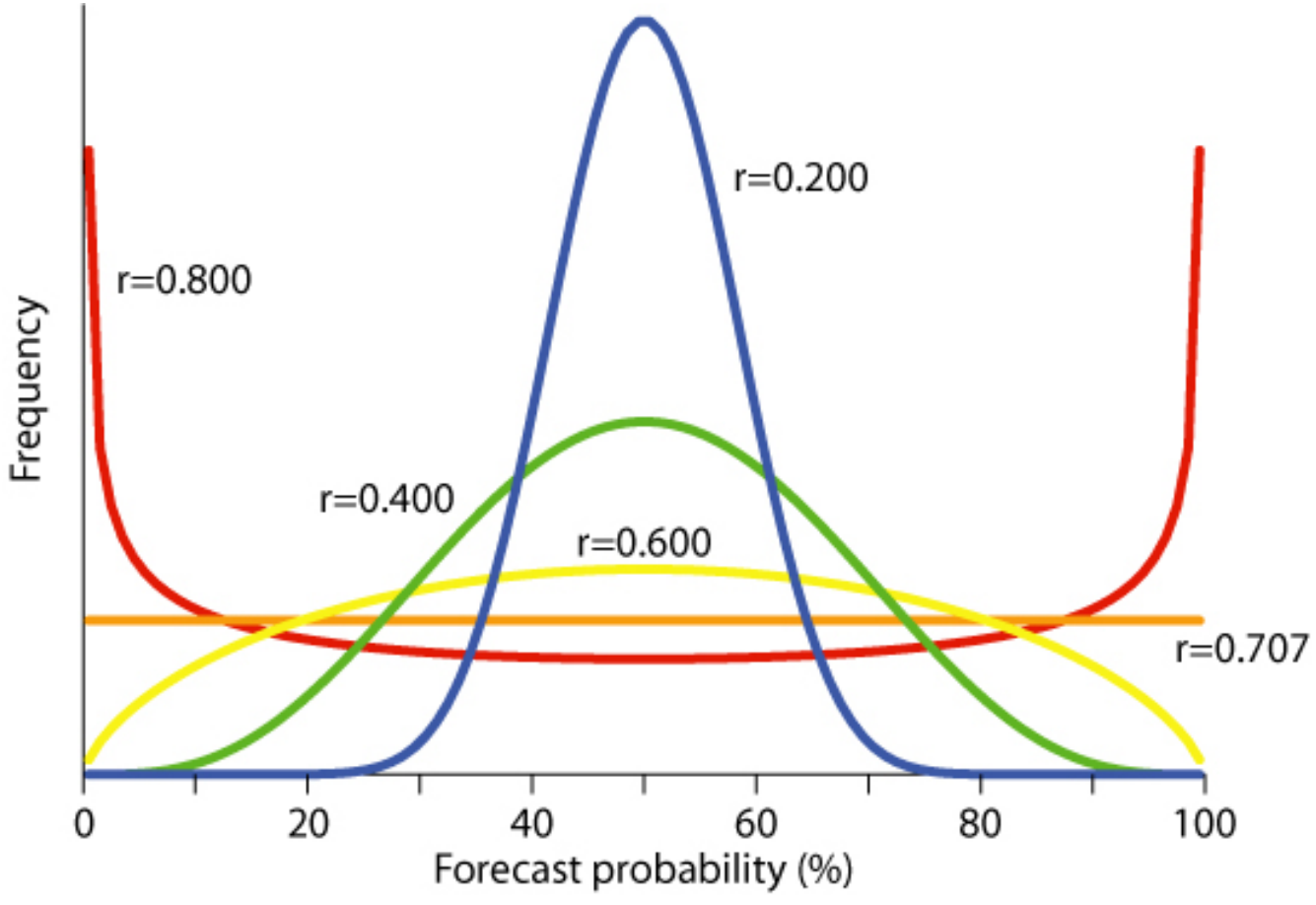
Use error variance
for probabilities



Forecast probabilities

Forecast probabilities are U-shaped when $r < 1/\sqrt{2}$

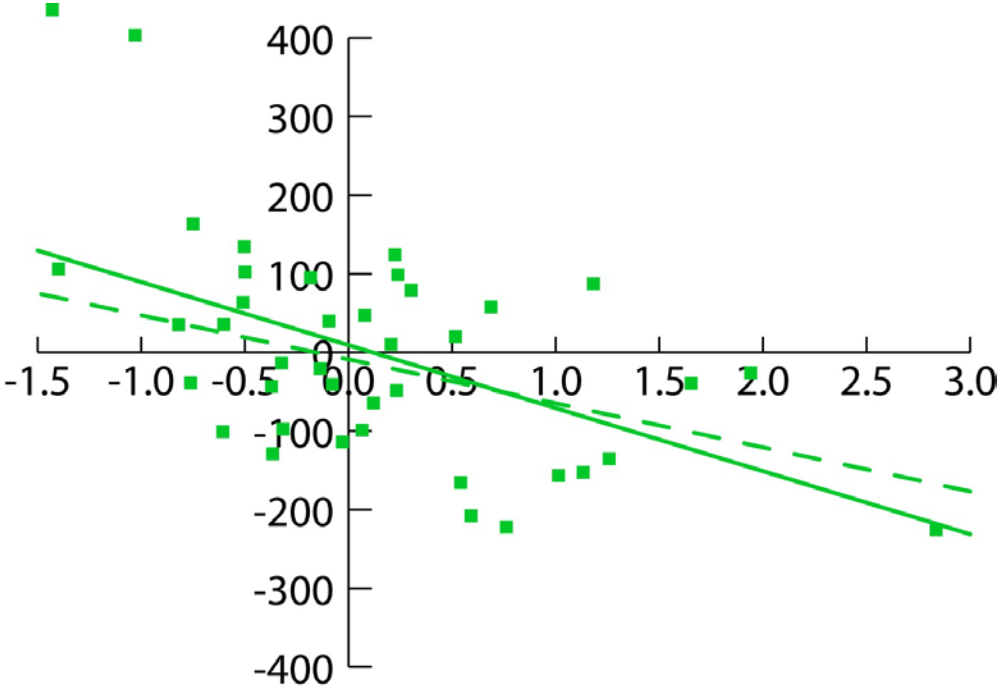
The variance of the forecast probabilities is a measure of the resolution (given assumptions …).



Forecast probabilities

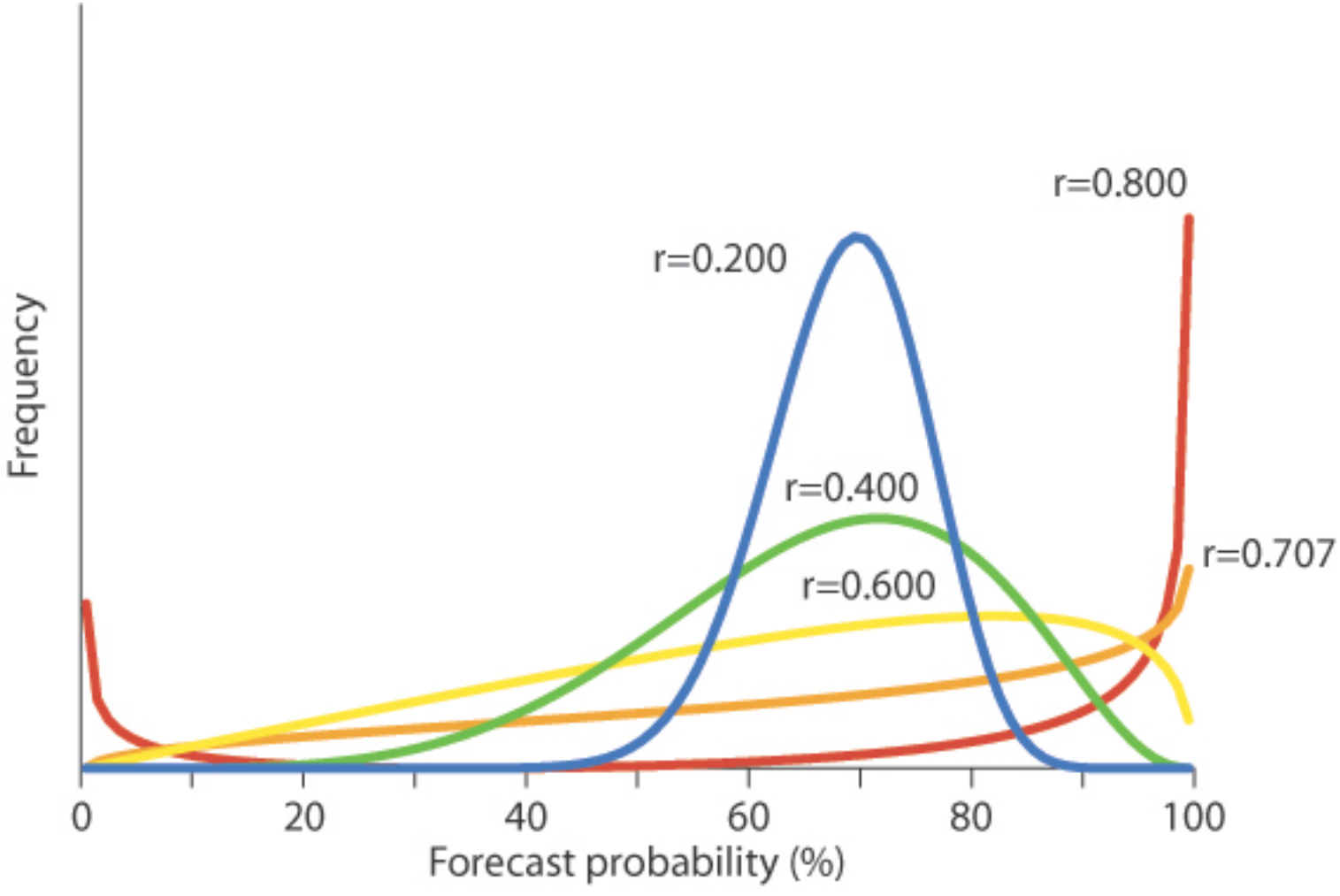
Errors are introduced if the parameters of the regression are imperfectly estimated:

- Errors in estimating the climatological probability (intercept)
- Errors in estimating the skill (slope)



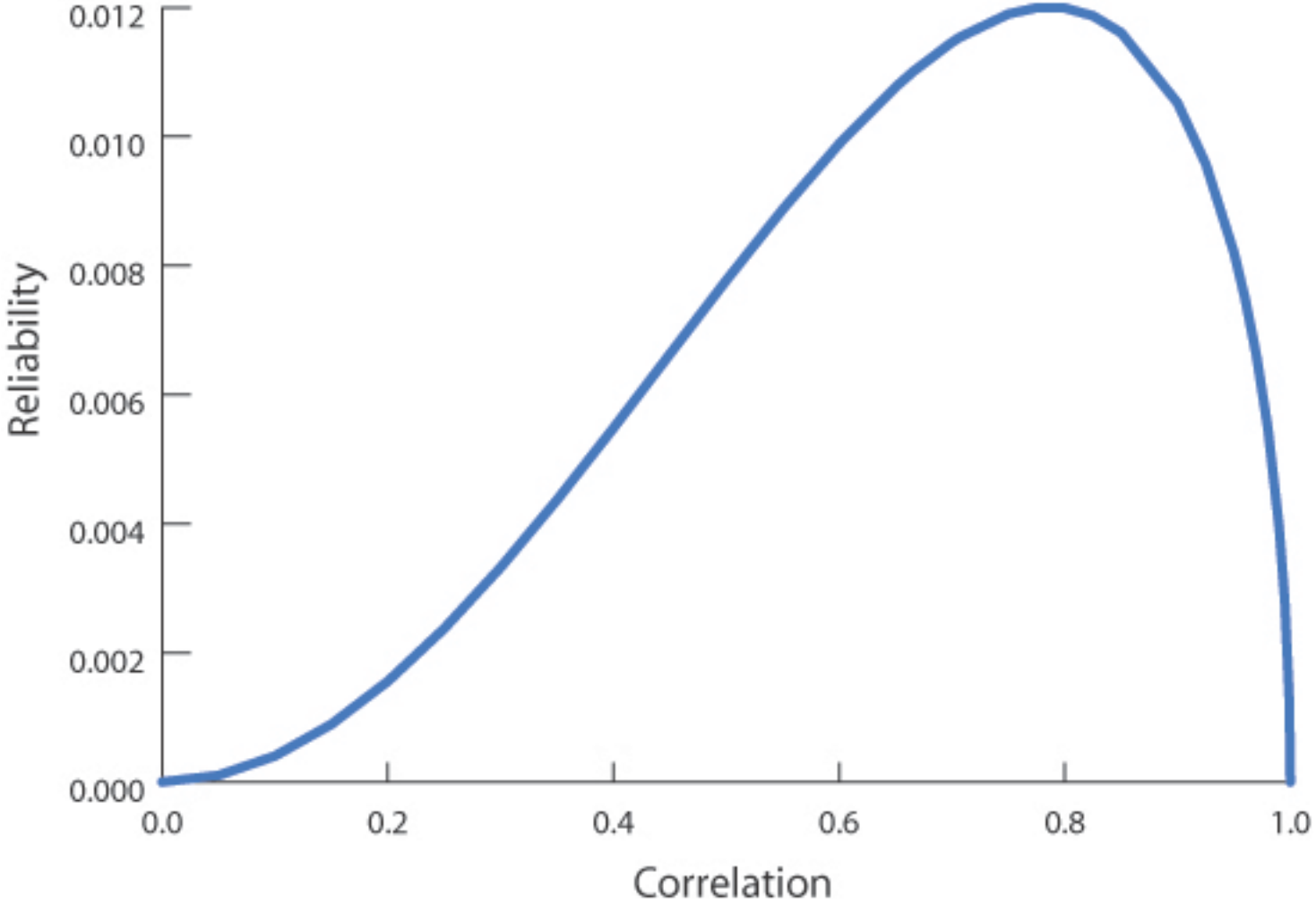
Biased forecasts

Forecast probabilities become skewed if the forecasts are biased.



Biased forecasts

Reliability errors as a function of correlation.



Incorrect skill

Discrimination is calculated using the ranks for the forecasts / forecast probabilities, so the score is insensitive to increases or decreases in variance.

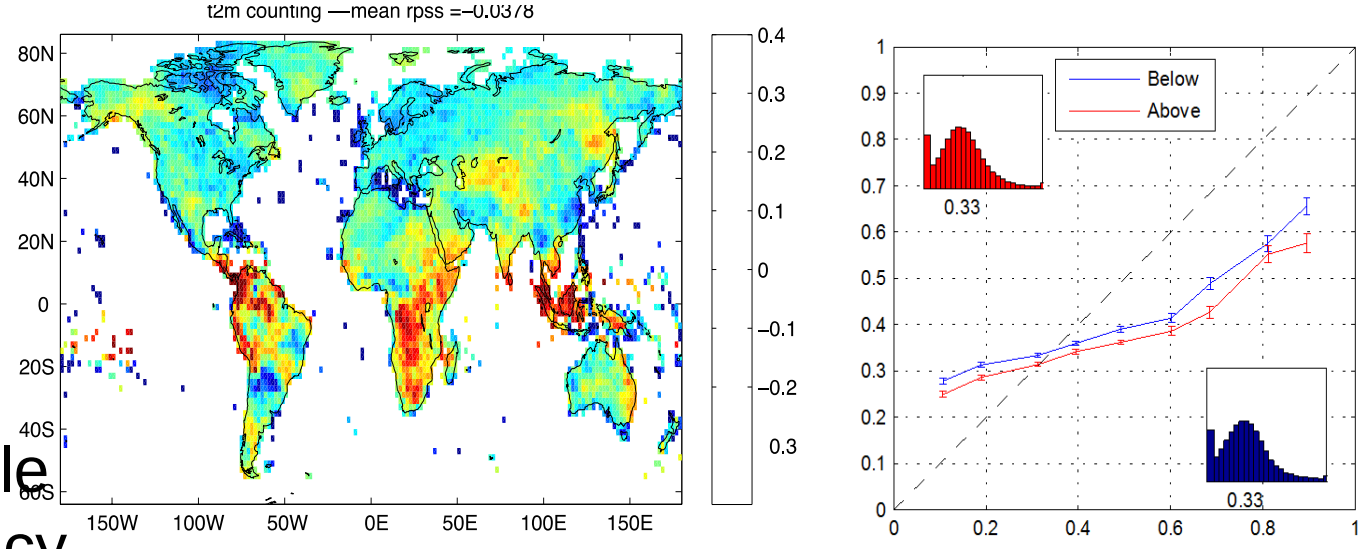
Similarly, the resolution score (Murphy, 1973) is not a function of the forecast probabilities per se.

So only procedures that affect that ranking of the forecasts (e.g., non-linear, multivariate) will leave resolution unchanged, at best.

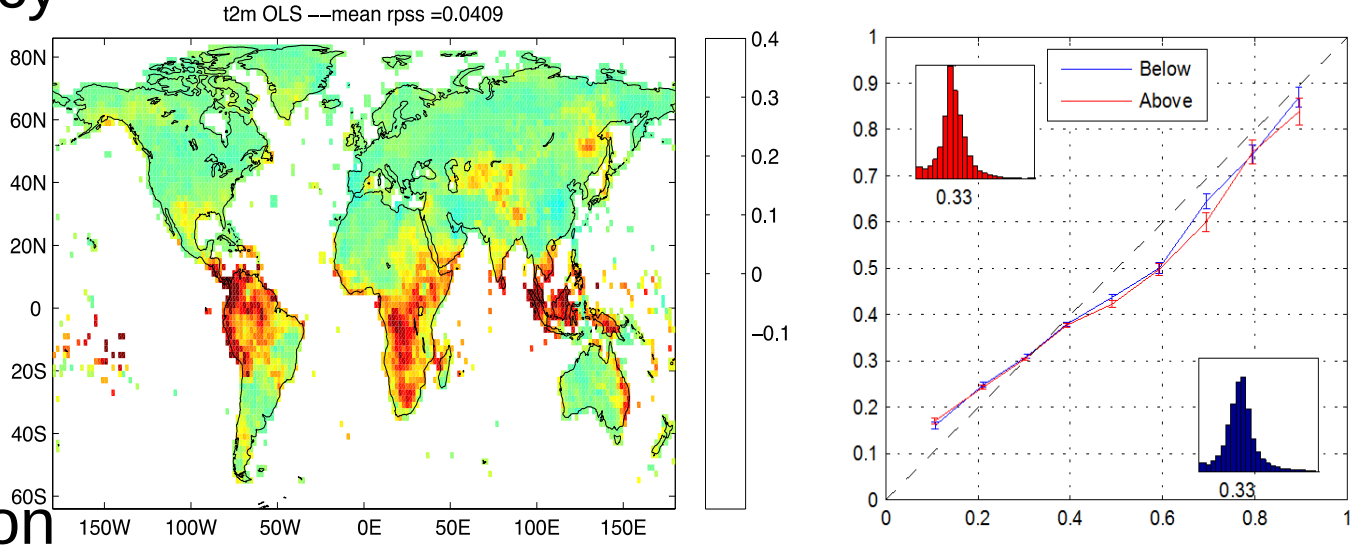
The effects of regression in practice?

DJF t2m Nov 1 start

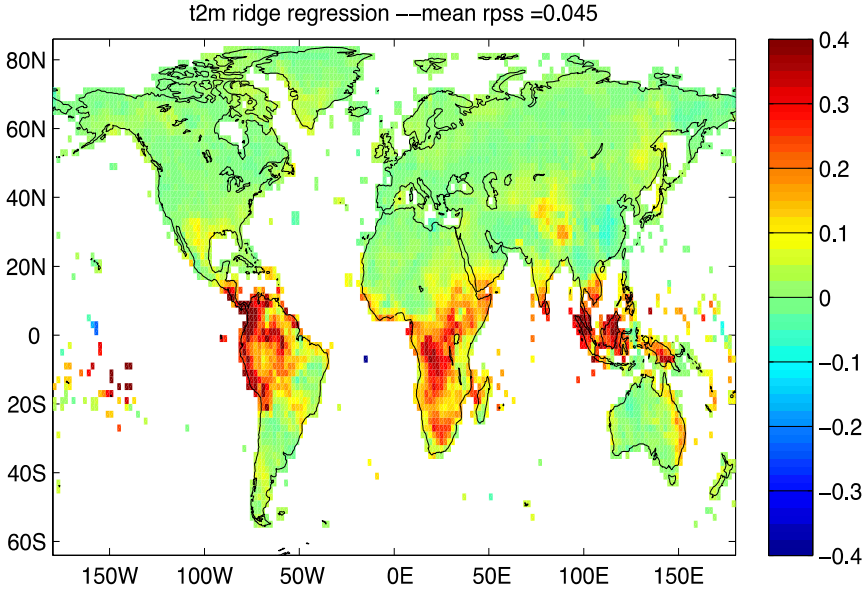
Ensemble Frequency

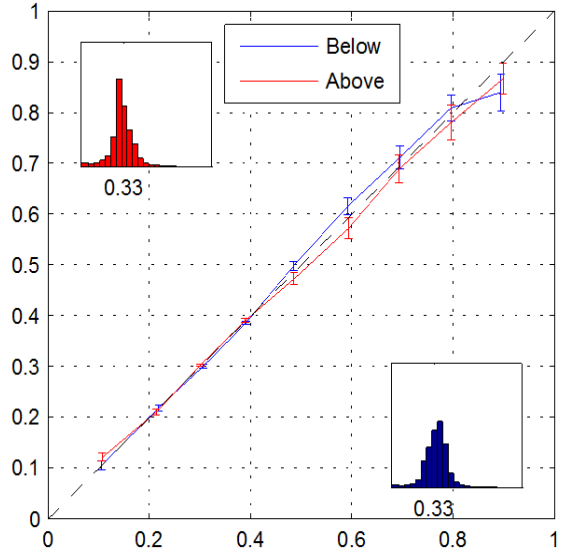


Linear regression



Ridge regression





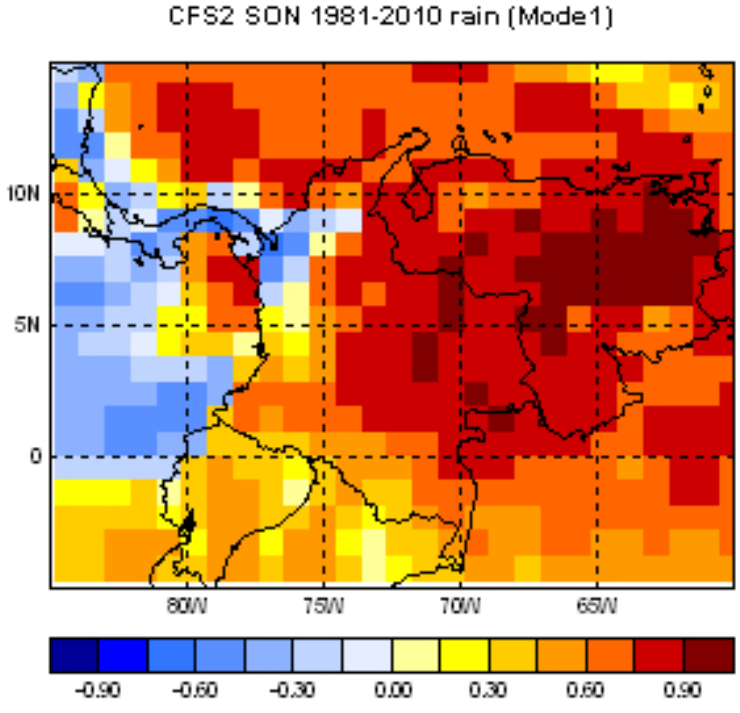
Using linear regression models for more than recalibration

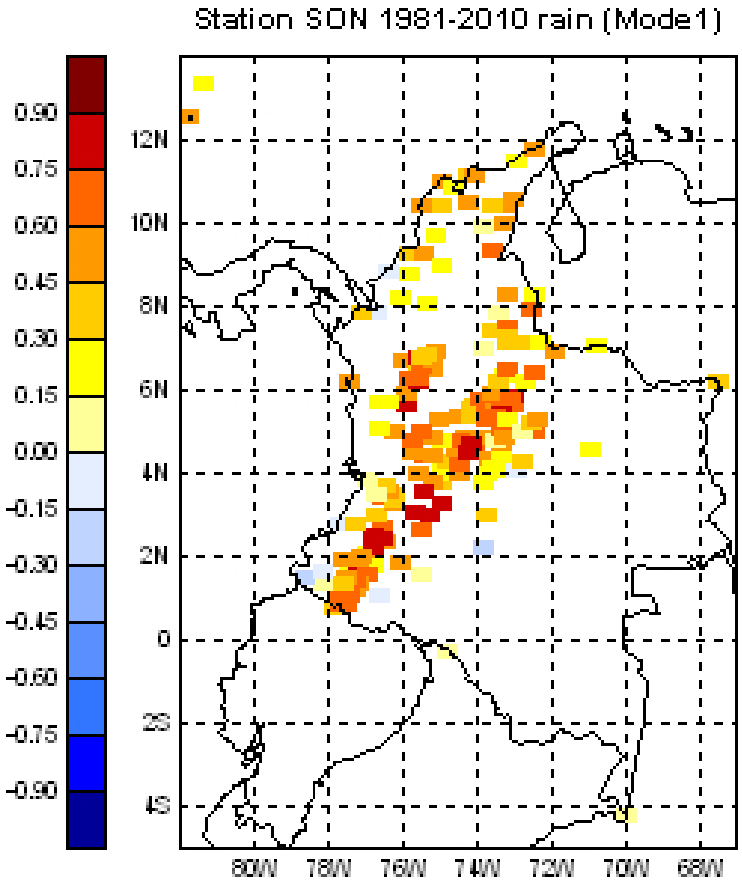
Gridbox-by-gridbox recalibration can reduce reliability errors.

Discrimination and resolution cannot be improved for continuous forecasts.

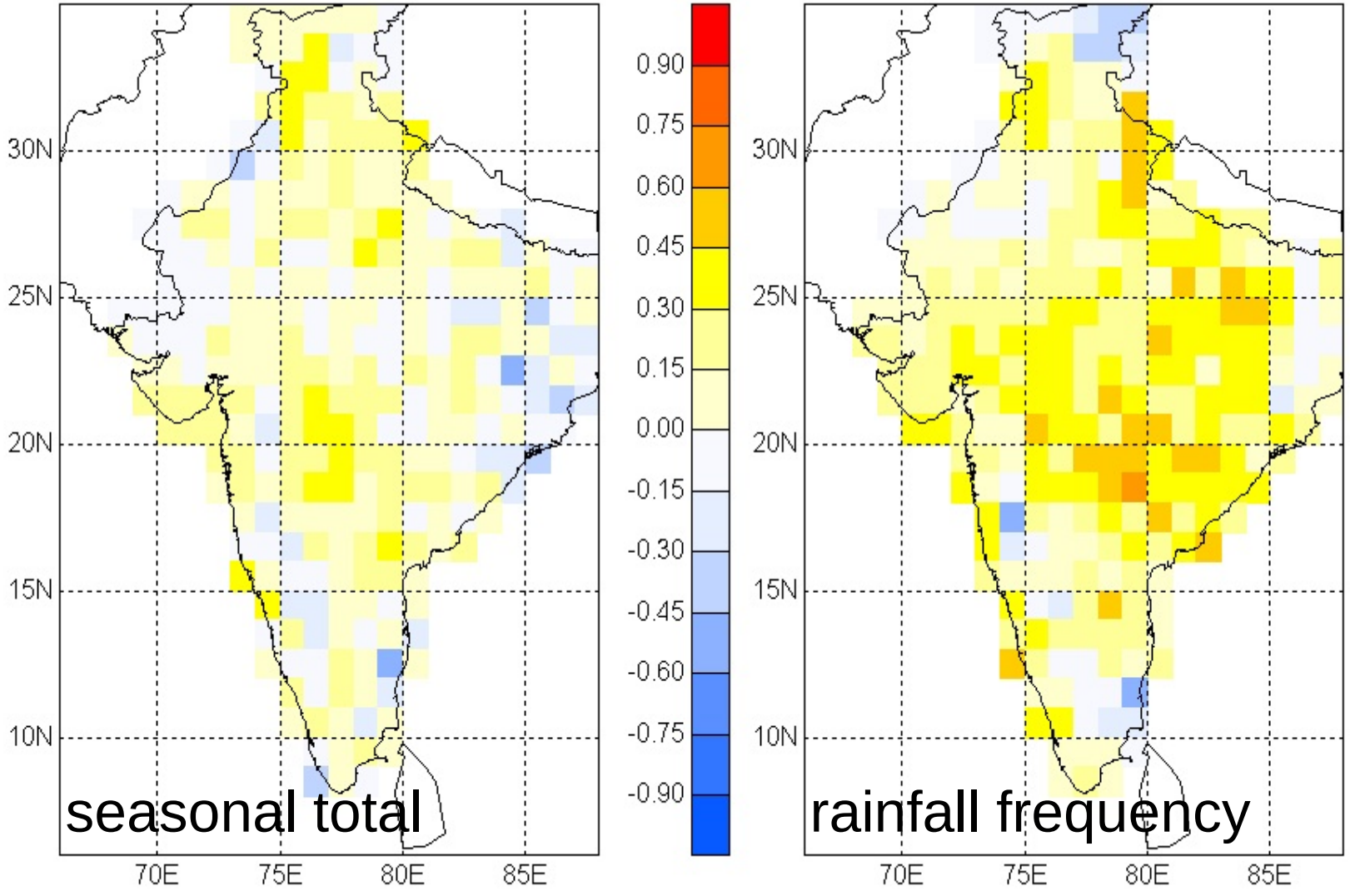
They can be improved for categorical forecasts because forecast ranks are affected (through the redefinition of ties).

Spatial errors



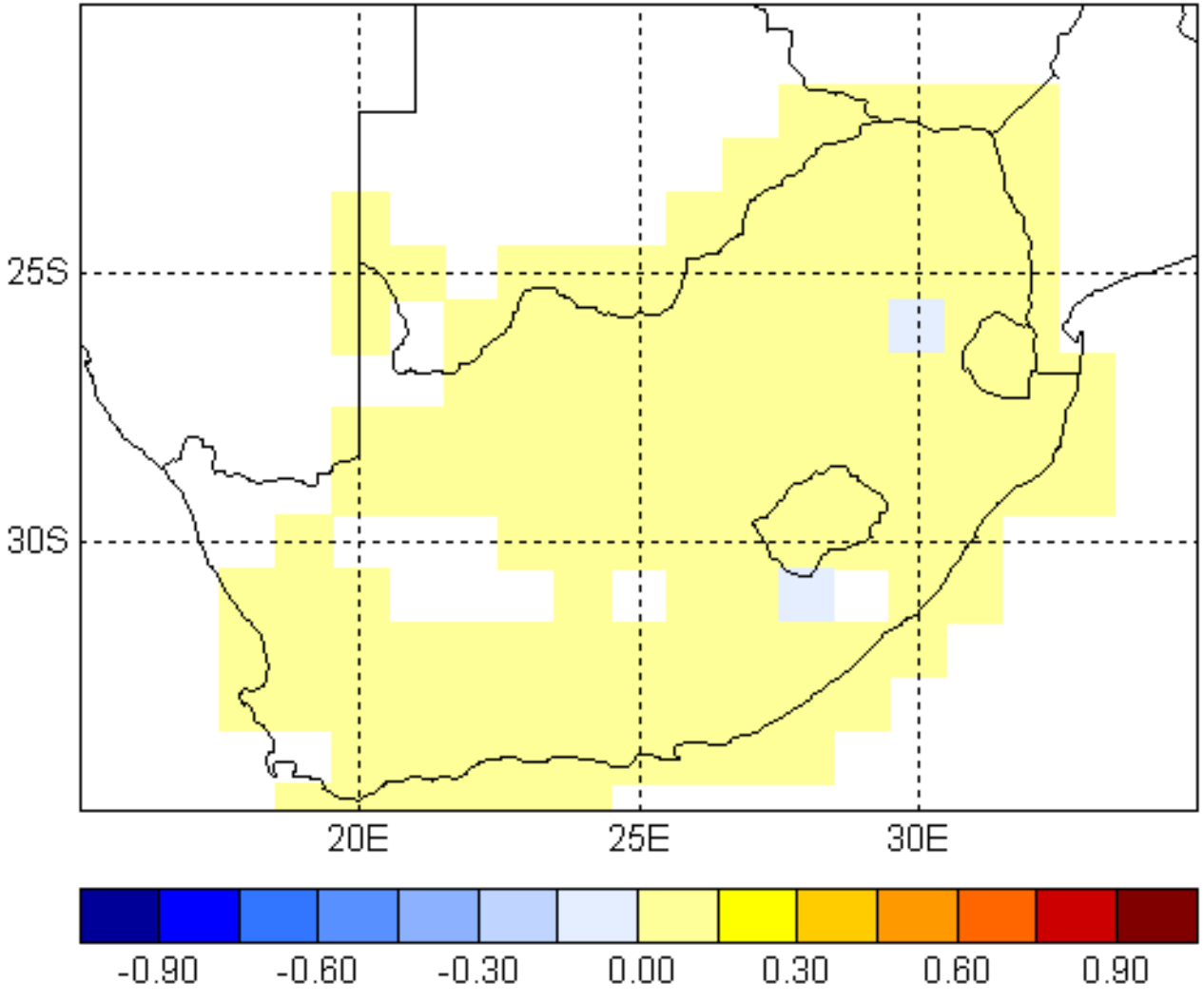


Predicting rainfall occurrence



JJAS rainfall correlation skill (ECHAM4-CA: made from June 1)

Predicting heavy rainfall occurrence



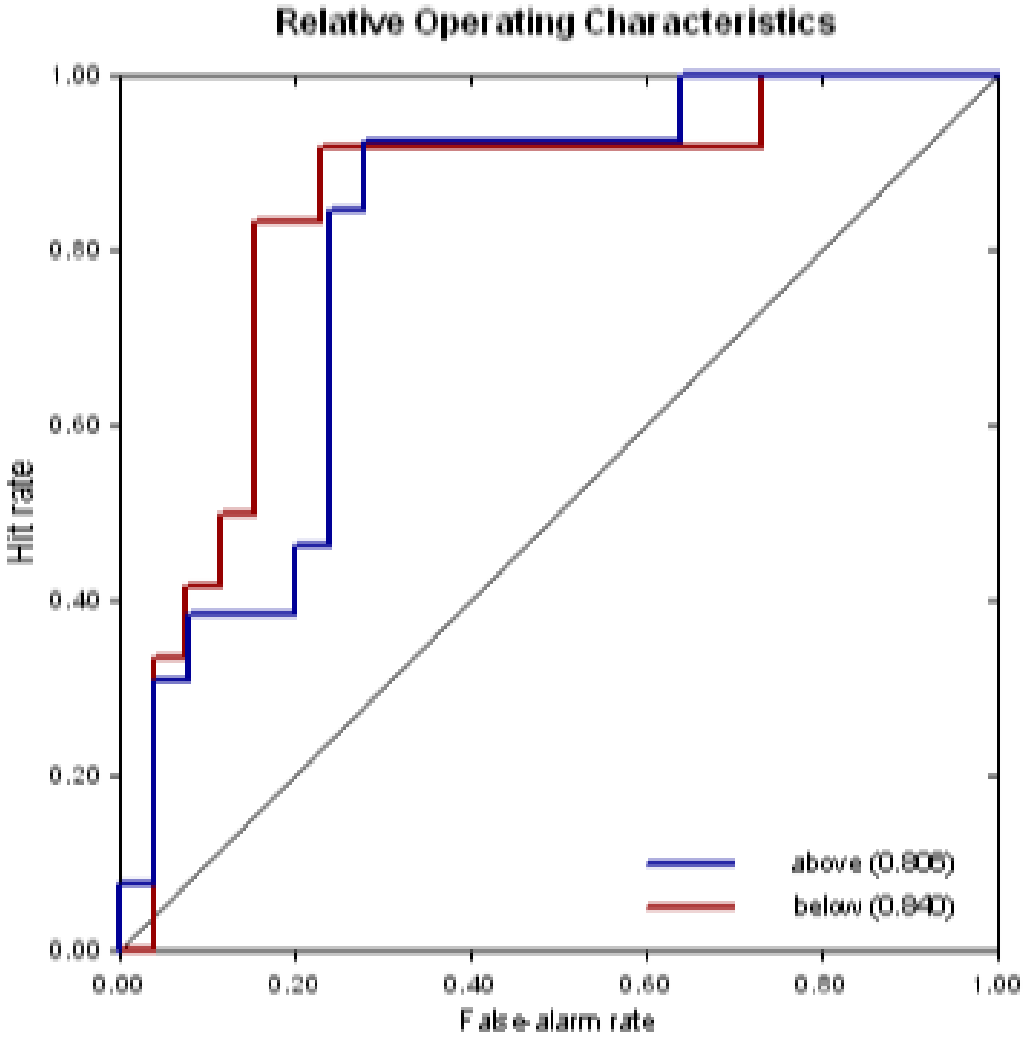
DJF heavy rainfall correlation skill

Predicting heavy rainfall occurrence

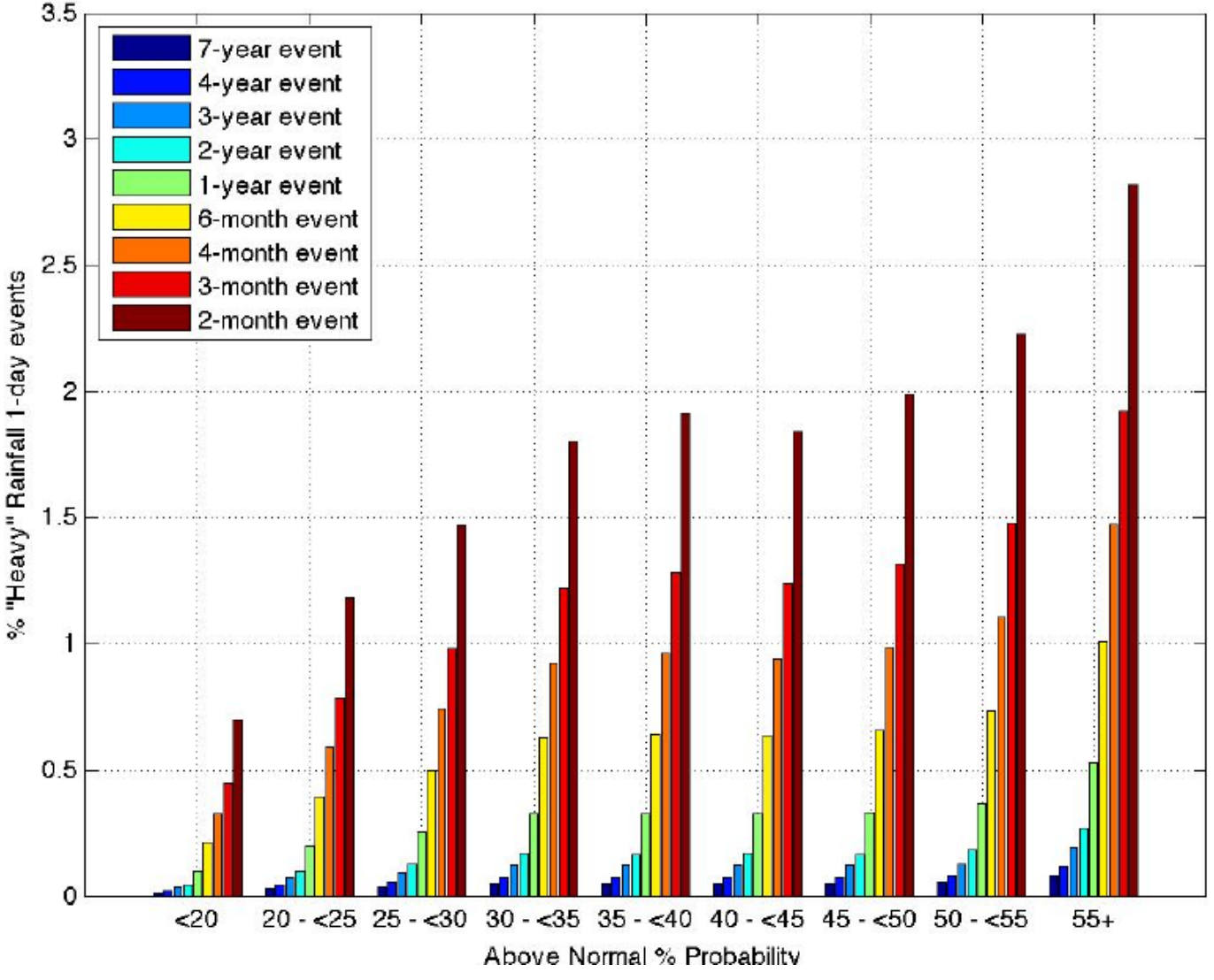
Predict occurrence of heavy-rainfall events anywhere within the country.

Predictand reflects frequency and extent.

r=0.54.

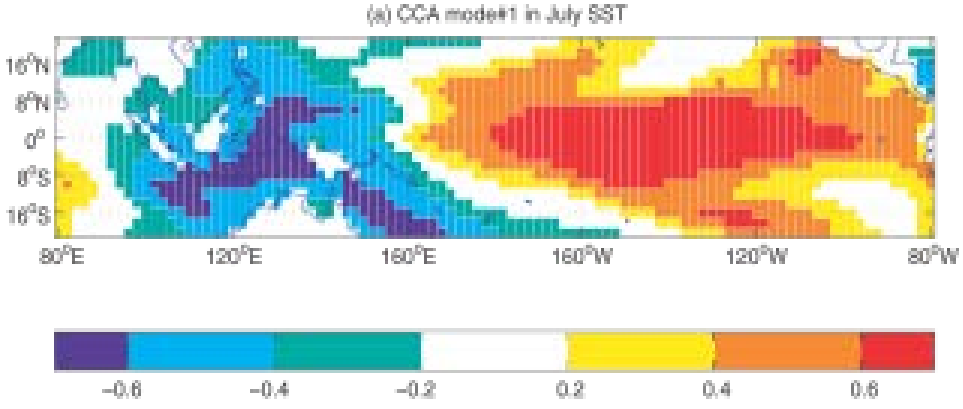


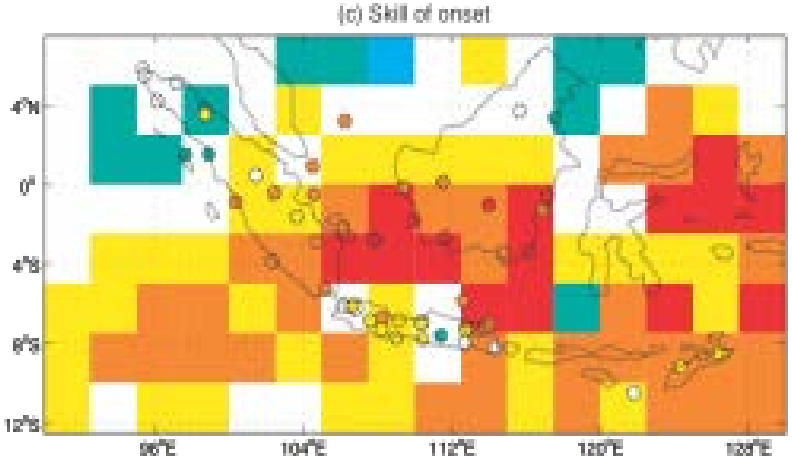
Predicting heavy rainfall occurrence

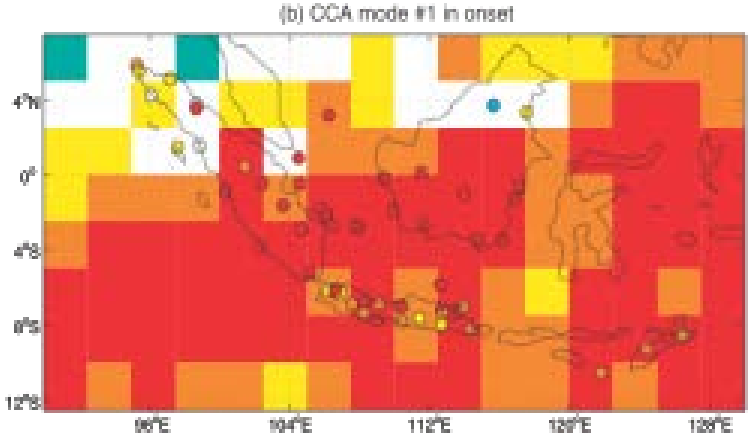


Predicting onset dates

Predictability of monsoon onset date over Indonesia from July SST





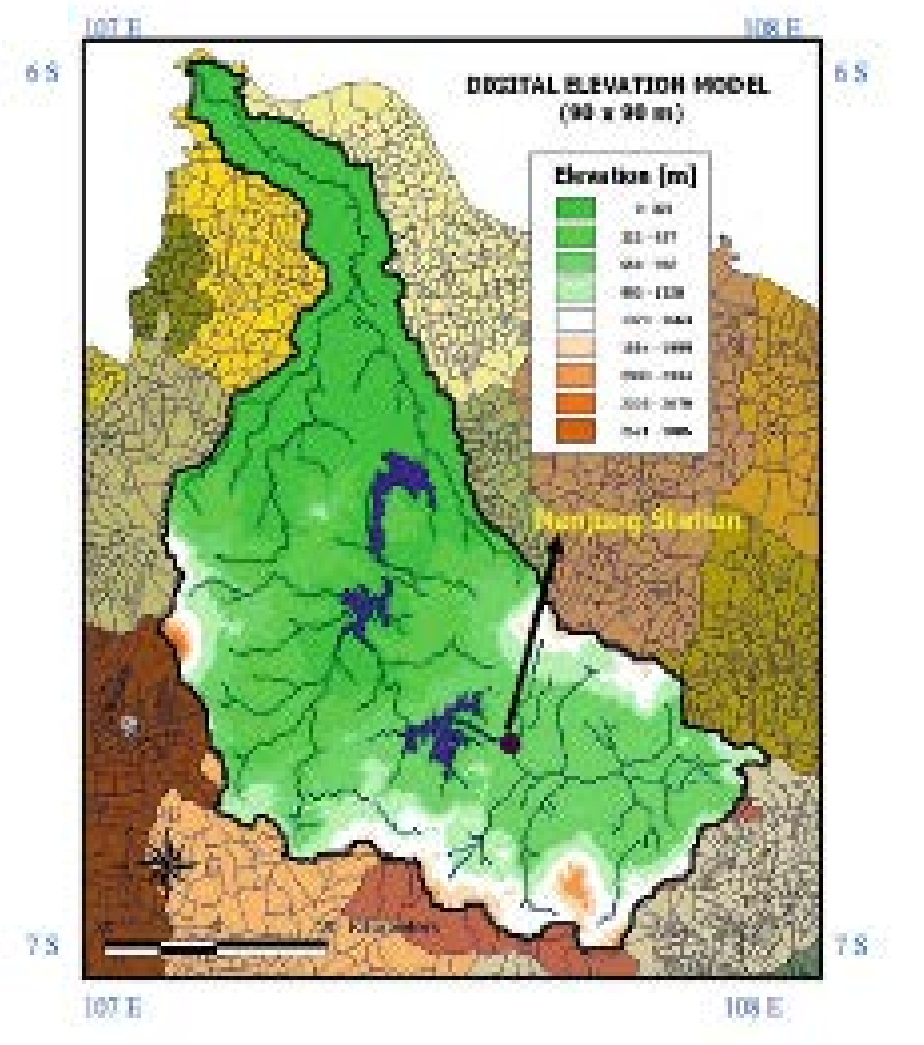


Moron, Robertson, Boer (2009)

Predicting dam inflow

The Citarum River is the largest river basin in west Java with a catchment area of 12,000 km^2.

It supplies 80% of the water demands in Jakarta alone.



Predicting dam inflow

Hindcasts of Citarum discharge (Sep–Nov) based on Aug 1 CFS hindcasts 1982–2006.

Cross-validated scores

Index: 1

Index: A

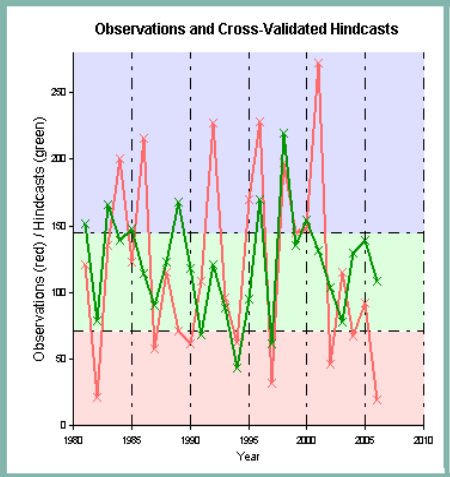
CFSv1

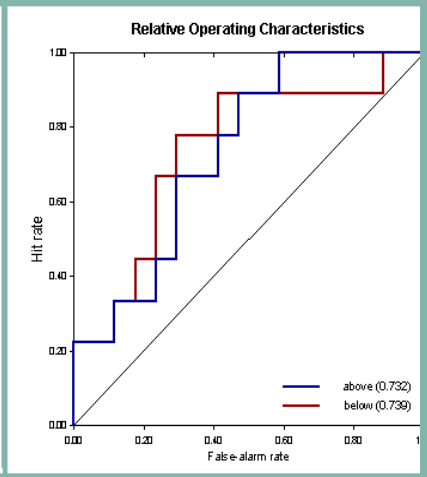
Continuous measures:

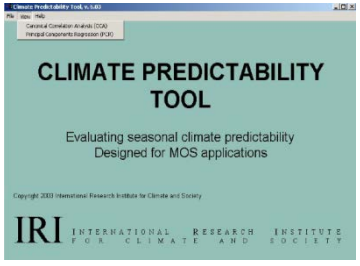
Pearson's correlation	0.4752
Spearman's correlation	0.5325
2AFC score	68.00%
% variance	22.58%
Variance ratio	0.3220
Mean bias	-0.21
Root mean squared error	60.80
Mean absolute error	50.37

Categorical measures:

Hit score	46.15%
Hit skill score	19.23%
LEPS	32.50%
Gerrity score	29.74%
2AFC (forecast categories)	70.67%
2AFC (continuous forecasts)	70.22%
ROC area (below-normal)	0.7386
ROC Area (above-normal)	0.7320







Cross-validated scores

Index: 1

Index: A

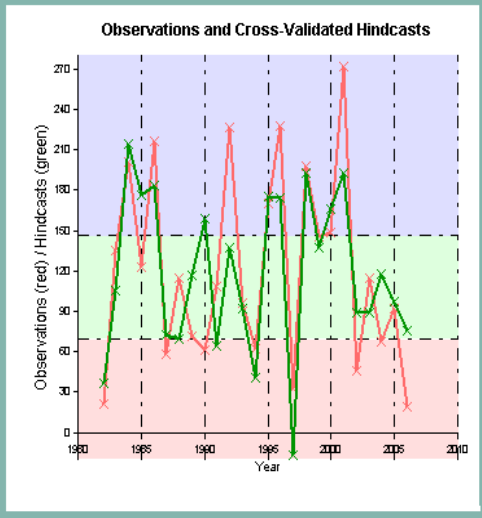
CFSv2

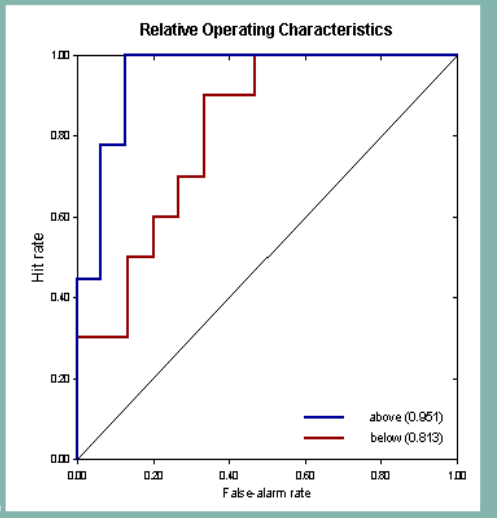
Continuous measures:

Pearson's correlation	0.7738
Spearman's correlation	0.7767
2AFC score	78.33%
% variance	59.87%
Variance ratio	0.6679
Mean bias	-2.73
Root mean squared error	44.58
Mean absolute error	36.19

Categorical measures:

Hit score	60.00%
Hit skill score	40.00%
LEPS	52.44%
Gerrity score	51.53%
2AFC (forecast categories)	79.90%
2AFC (continuous forecasts)	83.82%
ROC area (below-normal)	0.8133
ROC Area (above-normal)	0.9514

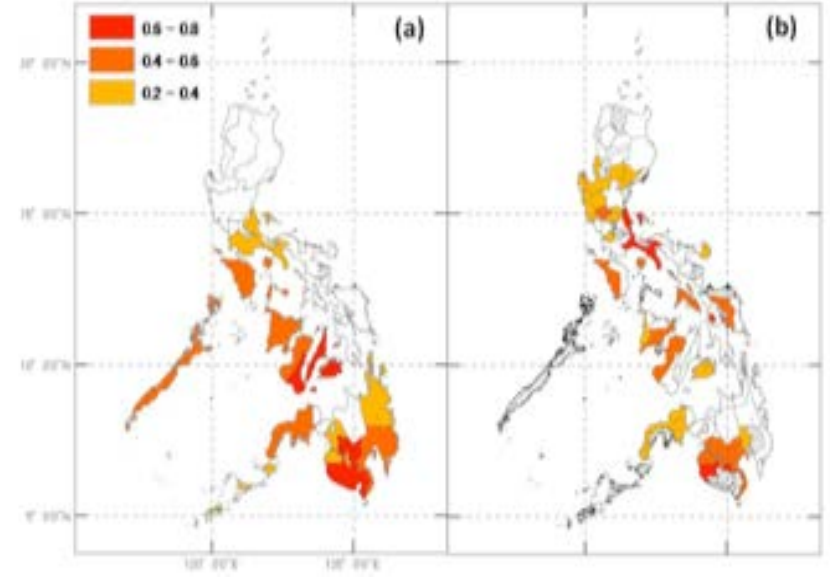




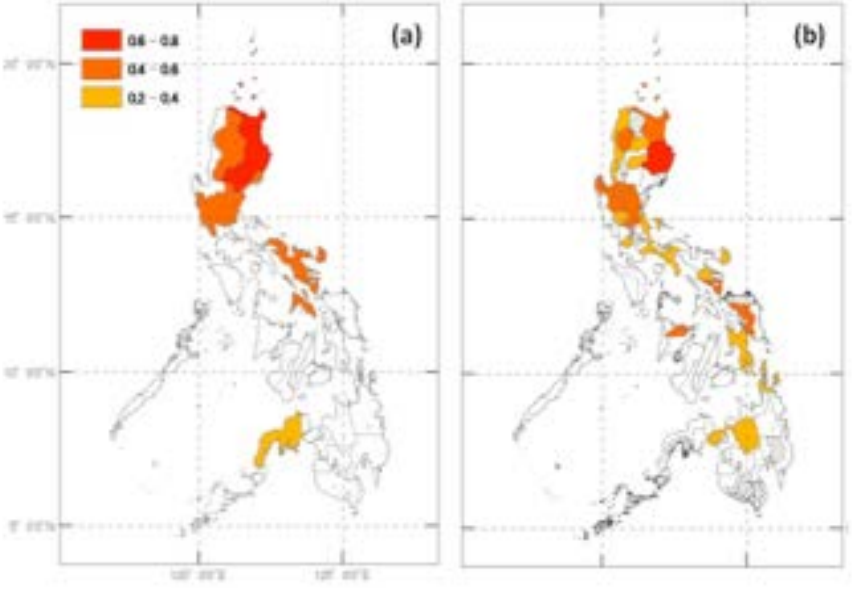
Predicting rice production

ACC Skill of (a) regional & (b) provincial production

Jan–Jun
(Dry Season)
from prev. Jun 1



Jul–Dec
(Rainy Season)
from prev. Mar 1



Conclusions

- Even the simplest of statistical models may result in loss of resolution / discrimination because of sampling errors in estimating model parameters, and invalidity of assumptions.
- More generally, recalibration schemes can often deteriorate the forecasts.
- Multivariate or non-linear statistical models can add resolution by correcting for other systematic errors, *if* the model parameters can be estimated sufficiently accurately.
- Statistical models can be useful for non-standard predictands.