The Abdus Salam International Centre for Theoretical Physics

United Nations Educational, Scientific and Cultural Organization

IAEA International Atomic Energy Agency

1945-1

Conference on African Drought: Observations, Modeling, Predictability, Impacts

2 - 6 June 2008

Predictability of African Drought
Part 1: Verification Methods

Andre' KAMGA Foamouhoue

ACMD
Niamey
Niger

akamgaf@yahoo.com

PREDICTABILITY OF THE AFRICAN DROUGTH

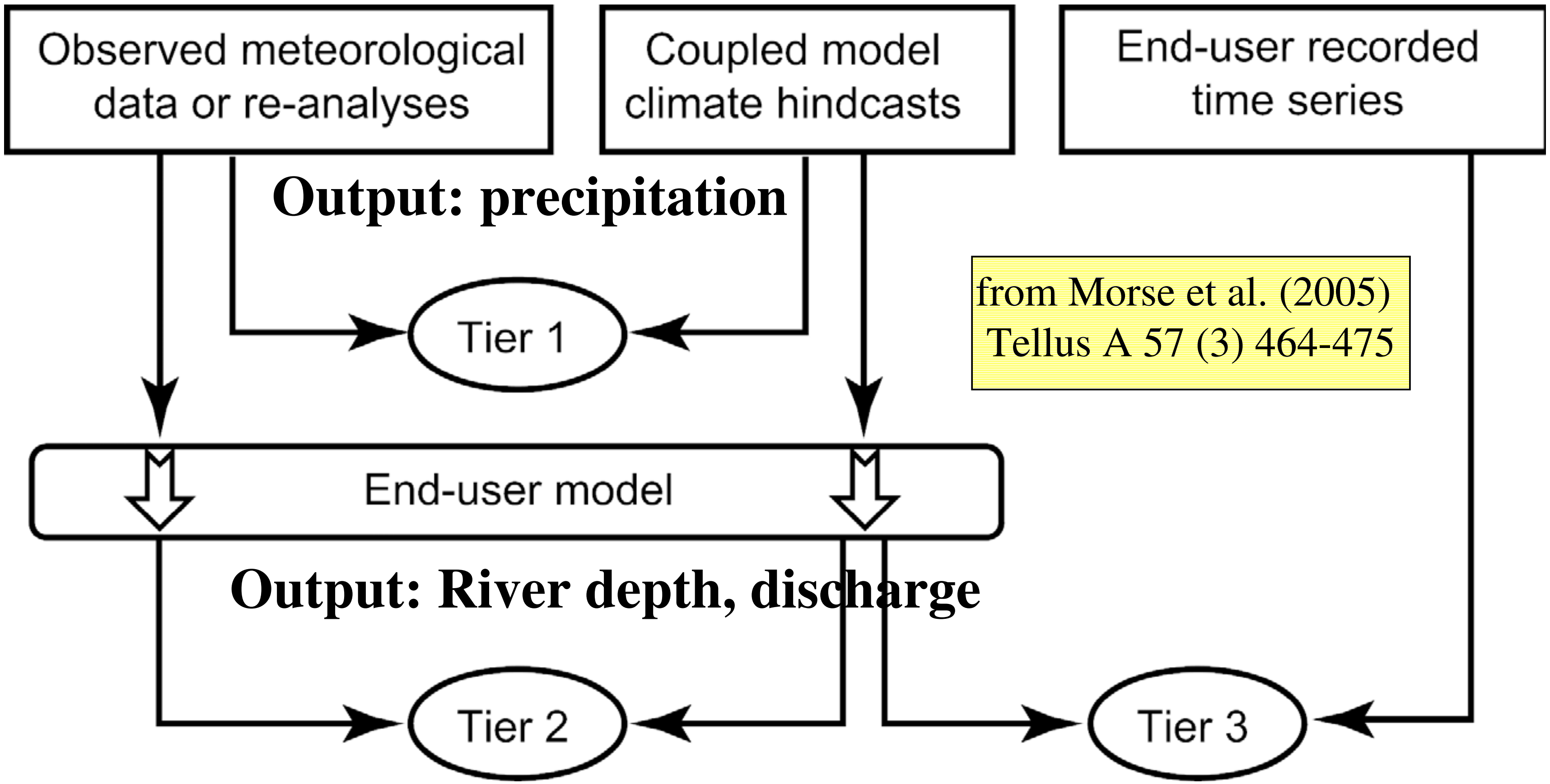
PART I:VERIFICATION METHODS

Andre KAMGA FOAMOUHOUE
ACMAD
Africa Drought Conference
ICTP-June 2008.

Outline

- Tiers of verification
- Accuracy and skill
- Probabilistic measures
- Data preparation for verification
- Reliability
- RPSS
- ROC

Tiers of Verification



from Morse et al. (2005)
Tellus A 57 (3) 464-475

Morse and Kamga 'Seamless ensemble forecast users in Africa'
THORPEX STISS, December 2006

Some Measures of accuracy and skill

- Pattern correlation anomaly
- Root mean square error
- Reliability diagrams
- Ranked probability skill score
- Relative operating characteristic Score

Result of deterministic verification experiment

(deterministic and ensemble mean outputs)

From Root mean square error and pattern anomaly correlation (deterministic measures) over NH show some advantage of using ensemble mean as the estimator of the future flow particularly at medium range.

Probabilistic measures

- The most important application of ensemble forecasts is their use for the generation of probabilistic forecasts.
- We can evaluate such forecasts by determining the percentage of the ensemble members at each grid point that fall into any of the climatologically equally likely categories (tercile, quintile, …) and then using that value as the forecast probability of the event. All verification scores can be averaged over the climate events.

RPSS (compare performance of two forecasting systems)

Reliability Diagrams (forecast calibration or adjustment of probability particularly when over or under forecasting is noted)

ROCSS (suitable to assess forecast skill for a specif event. The skill map may be used to mask the forecast in areas where there is no skill)

For details for review on forecast verification:

1-Wilks, 1995, Statistical methods in Atmospheric sciences, Acamedic press Inc. 467 pp)

2- Richardson, D. S., 2000: Skill and economic value of ECMWF EPS Q.J.R. Meteorl. Soc., 126, 649-668.

Verification of probabilistic/deterministic forecasts: Data Preparation

Categorical forecasts can be converted to probabilistic forecasts for the sake of verification with EPS.A probability of 1 is assigned to the bin in which the deterministic forecasts falls at each grid point and 0 to all the other bins.

These probabilities are then calibrated resulting in a binary probabilistic forecast

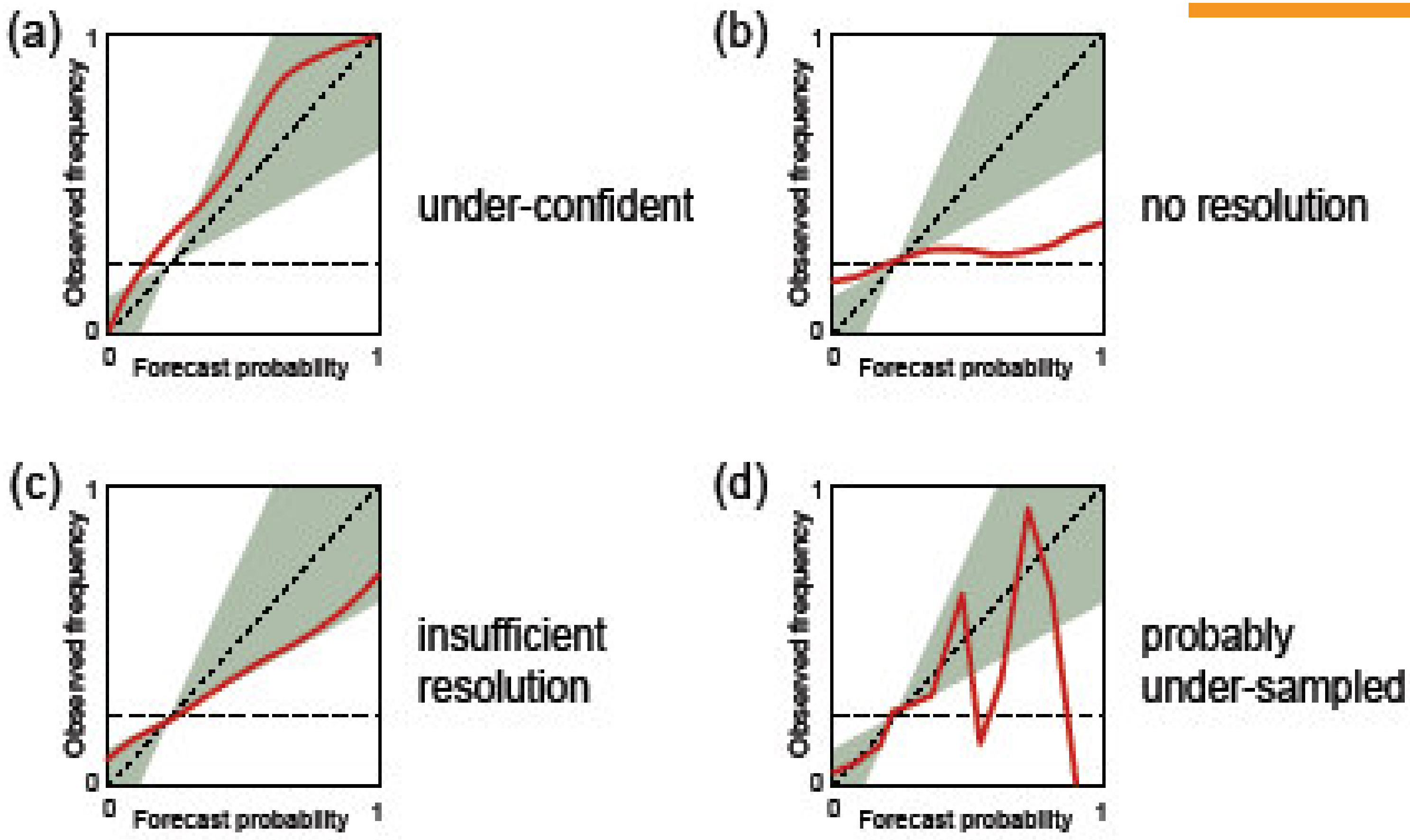
To make the verification fair between deterministic and EPS in terms of probabilistic measures, the EPS forecasts can be degraded by retaining probability value for the ensemble mode(i.e the most likely climate bin) and distributing the remaining probability 1-p equally to other bins. On can also give prob of 1 to the most likely bin and zero to other.

The difference between deterministic and degraded probabilistic forecast is that p varies depending on how predictable the weather is whereas the deterministic forecast value is fixed.

Reliability Diagram

Given a particular forecast probability of an event (that is a climate bin at a grid point), on can determine the relative frequency at which an event with that forecasts probability is observed.

Interpretation of reliability diagram



- The reliability diagram is conditioned on the forecasts (i.e., *given that X was predicted, what was the outcome?*)

Calibration

The probabilistic ensemble forecast can be well calibrated.

The calibrated forecasts probabilities are given as the observed frequencies corresponding to the forecast probabilities from a previous time period.

Example

When 8 of 15 members fall in a climate bin, the forecast probability is not only 8/15 but rather the observed frequency at which the verifying observation fell into bins with 8 ensemble members during a preceding verification time period

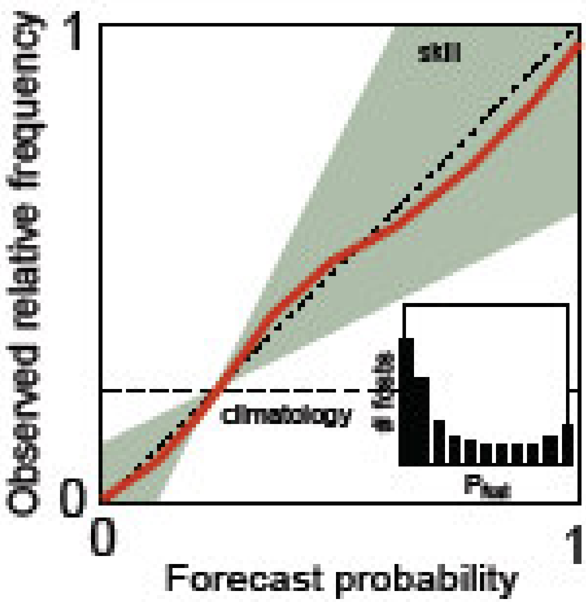
Perfect reliability = the forecasts probabilities matches the observed frequency

Reliability

Measure agreement between predicted probabilities and observed frequencies

If the forecast system is ***reliable***, then whenever the forecast probability of an event occurring is *P*, that event should occur a fraction *P* of the time.

- For each probability category plot the frequency of observed occurrence

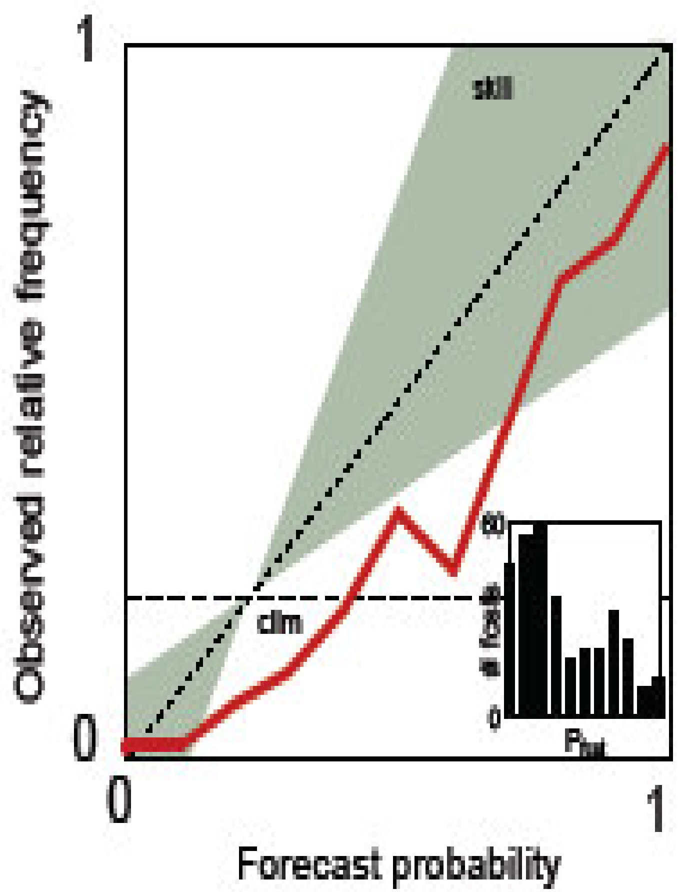


Sample dataset
Forecast Probability of rain

Date 2003	Observed rain	24h forecast POP	48h forecast POP
Jan 1	no	0.3	0.1
Jan 2	no	0.1	0.1
Jan 3	no	0.1	0.2
Jan 4	no	0.2	0.2
Jan 5	no	0.2	0.2
...	...	...	...
Dec 27	yes	0.8	0.8
Dec 28	yes	1.0	0.5
Dec 29	yes	0.9	0.9
Dec 30	no	0.1	0.3
Dec 31	no	0.1	0.1

Table and reliability diagram

Forecast probability	# fcsts	# observed occurrences	Obs. relative frequency
0.0	46	1	0.02
0.1	55	1	0.02
0.2	59	5	0.08
0.3	41	5	0.12
0.4	19	4	0.21
0.5	22	8	0.36
0.6	22	6	0.27
0.7	34	16	0.47
0.8	24	16	0.67
0.9	11	8	0.73
1.0	13	11	0.85
Total	346	81	0.23



1. For each forecast probability category count the number of observed occurrences
2. Compute the observed relative frequency in each category *k*

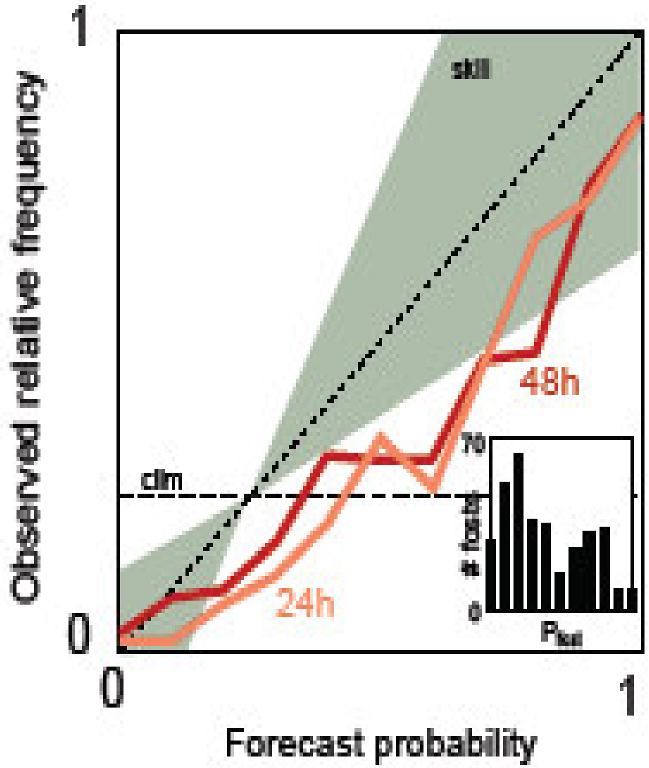
   obs. relative frequency$_k$ = *obs. occurrences*$_k$ / *num. forecasts*$_k$

3. Plot observed relative frequency vs forecast probability
4. Plot sample climatology ("no resolution" line)

   sample climatology = obs. occurrences / num. forecasts

5. Plot "no-skill" line halfway between climatology and perfect reliability (diagonal) lines
6. Plot forecast frequency separately to show forecast sharpness

Forecast probability	# fcsts	# observed occurrences	Obs. relative frequency
0.0	31	1	0.03
0.1	53	5	0.09
0.2	67	7	0.10
0.3	39	7	0.18
0.4	38	12	0.32
0.5	16	5	0.31
0.6	26	8	0.31
0.7	30	14	0.47
0.8	31	15	0.48
0.9	8	6	0.75
1.0	7	6	0.86
Total	346	86	0.25 Sample climatology



Rank Probability Skill Score (RPSS)

RPSS is a generalization of the Brier skill score for multi-categorical forecasts where categories can be ordered.

RPS

The Rank Probability Skill Score (Eipstein 1969, Wilks, 1995, Goddard et al.2003) computation begins with Rannk Probability Score (RPS) defined as:

$$RPS = \sum_{m=1}^{N_{cat}} \left(CP_{F_m} - CP_{O_m} \right)^2$$

CPfm= cumulative probabilities of the forecasts up to category m
Cpom= cumulative observed probability up to category m.
The probability distribution of the observation is 100% for the category that was observed. Ncat= 3 for terciles

RPSS= 1- RPSfcst/RPSref

RPSfcst is the RPS of the forecast;

and RPSref is the RPS of the benchmark forecast.

Applications of RPSS have shown that for some models beyond a few days (3) ensemble mode forecast, which is able to distinguish highly predictable events from poorly predictable situations, has a large advantage over deterministic .

Relative operating characteristic

- It's main advantage is that predictability of a forecasting system for a specific event (drougth, …) can be easily assessed.
- A contingency table for an event

		observation	
		yes	no
Forecast	yes	A	B
	no	C	D

A= hits, B= false alarm

Hit rate = A/(A+C)

False Alarm rate = B/(D+B)

The signal detection theory generalizes the concept of hits and false alarm to multi-category probability forecasts.

Suppose a forecast distribution is stratified into 10% wide categories occurrences and non occurrences of and event are tabulated for each category

The j-th category is related to a forecast probability between (j-1)*10% and j*10%

For a probability threshold j*10% the occurrences and non-occurrences can be summed to give the 4 entries of the contingency table.

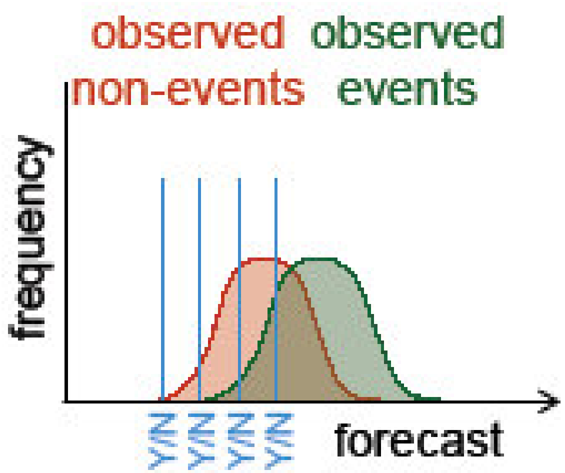
The hit and false alarm rate can be calculated and a point plotted on the ROC curve.

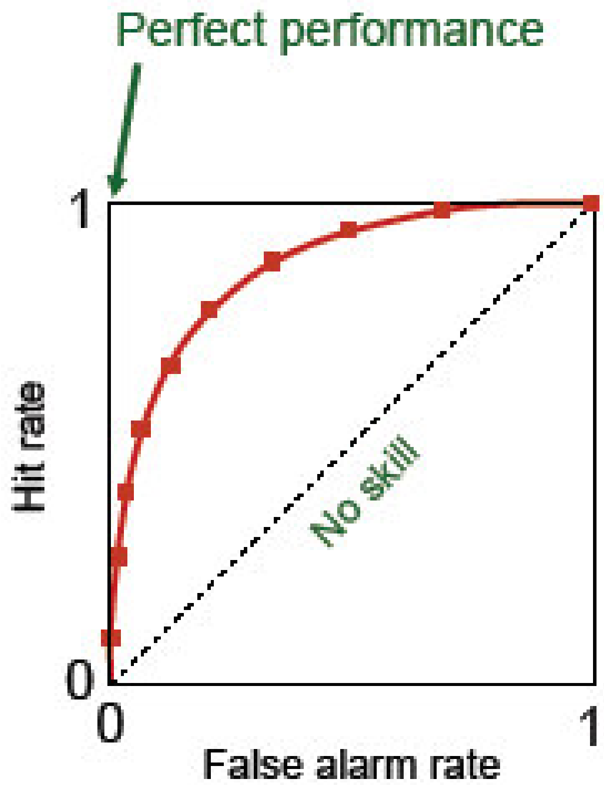
The process is repeated for all thresholds j*10% , j=1 to 10 to obtain all the points.

ROC

Measure success using Relative Operating Characteristic (ROC)

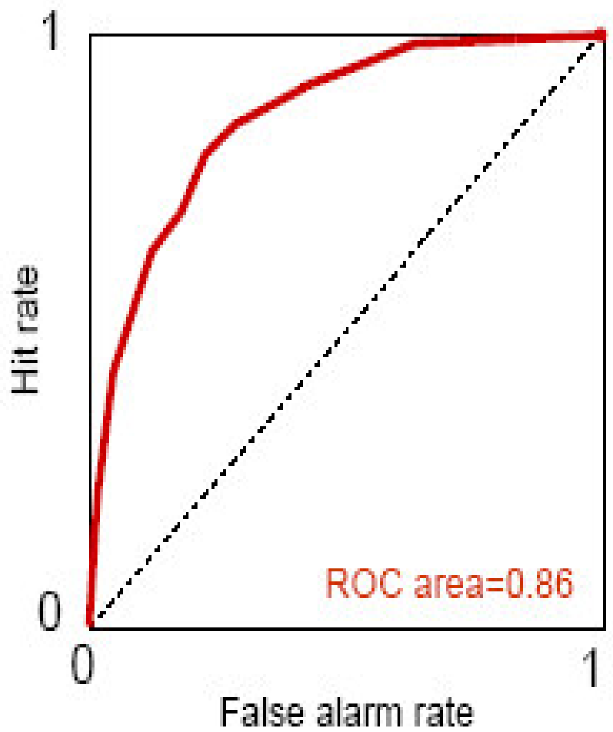
- Plot the hit rate against the false alarm rate using increasing probability thresholds to make the yes/no decision





Sample ROC curve
Forecast probability of rain

Forecast probability	Hits	Misses	False alarms	Corr. non-events	Hit rate	False alarm rate
0.0	81	0	265	0	1.00	1.00
0.1	80	1	220	45	0.99	0.83
0.2	79	2	166	99	0.98	0.63
0.3	74	7	112	153	0.91	0.42
0.4	69	12	76	189	0.85	0.29
0.5	65	16	61	204	0.80	0.23
0.6	57	24	47	218	0.70	0.18
0.7	51	30	31	234	0.63	0.12
0.8	35	46	13	252	0.43	0.05
0.9	19	62	5	260	0.23	0.02
1.0	11	70	2	263	0.14	0.01



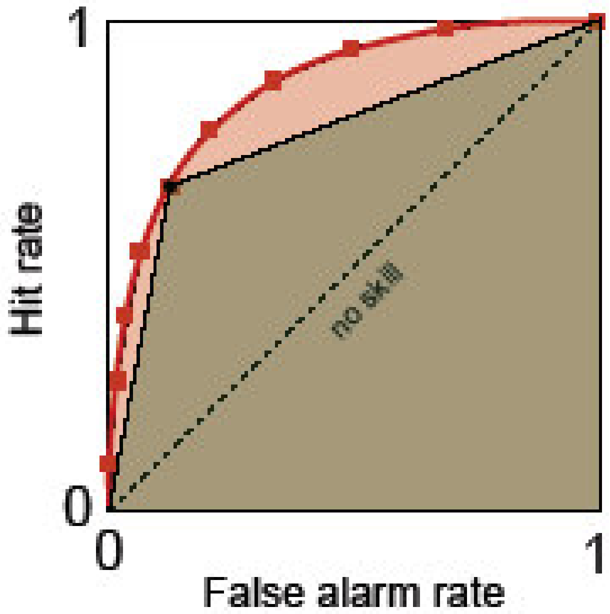
ROC curve is independent of forecast bias – is like "potential skill"

Area under curve ("ROC area") is a useful summary measure of forecast skill

- Perfect: ROC area = 1
- No skill: ROC area = 0.5
- ROC skill score

$$ROCS = 2\ (\text{ROC area} - 0.5)$$

= KSS for deterministic forecast



Assume that a decision can be taking if a summer season is likely to be abnormally wet or dry.

A forecast probability threshold above which preparedness for impacts of the abnormal climate event can be set.

- Each business may have a particular sensitivity to the anomalous event and therefore a particular probability threshold may apply to warn and trigger preparedness plan.
- If advise for action is given and the anticipated event occurs a « hit » credit provided.
- If advice for action is given and the event doesn’t occurs, it’s a « false alarm »

A ROC curve is graph with False alarm rate on the horizontal-axis and hit rate on the vertical axis.

It inform the user on the hit rates and false alarm rates to be expected from the use of different forecast probability thresholds to take decisions.

ROC curves is critical for optimal conversion of probability forecasts into statements of interest to decisions makes.

This process is forecasts interpretation and tailoring

Usefulness of a ROC curve

select the forecast probability trigger threshold for a particular user (this heavily rely on the forecaster s knowledge of the user s decision system)

Identify the optimal trade-off between hits and false alarm rates for a given decision model. This trade off should lead to a maximum benefit for the user when decision is made.

END