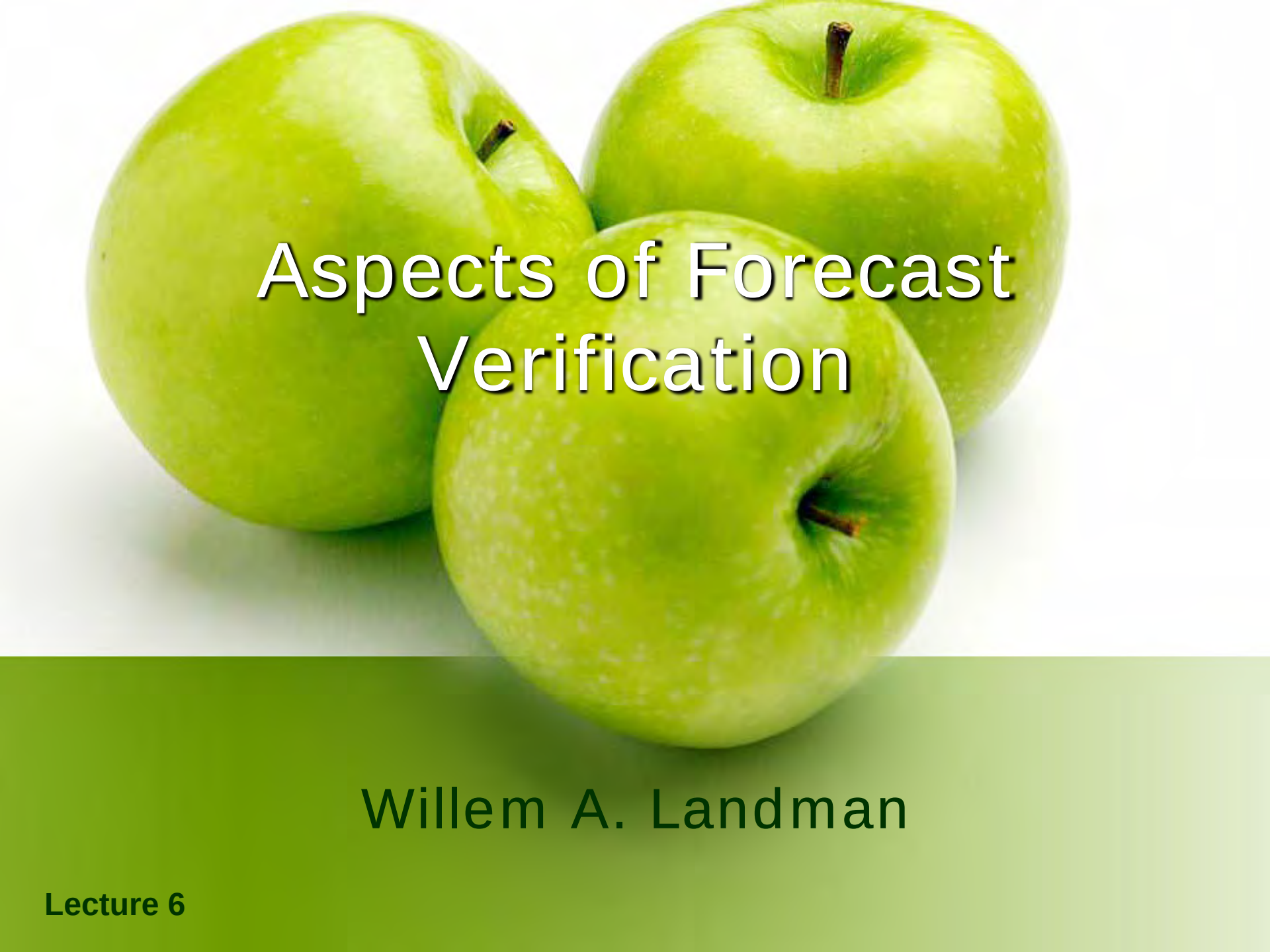


Aspects of Forecast Verification

Willem A. Landman

Lecture 6

Definition…

It is the process of determining the ***QUALITY*** of forecasts

Any forecast verification method involves **comparison** between **matched pairs** of forecasts and the observations to which they pertain

Forecast Quality and Forecast Value

- Value: the economic (or societal) worth of forecasts
  - Forecasts only have value if people use them
- Quality: the correspondence between forecasts and observations
  - accuracy, skill, reliability, …
  - “use” - make a decision or take an action which would not otherwise have been made
- Quality and value are not the same. There is not even a simple relationship between them
  - a forecast may have high skill, but no value
  - a low skill forecast could give high value to some users

The Contingency Table

	O_{yes}	O_{no}
F_{yes}	a	b
F_{no}	c	d

Accuracy Measures of Binary Forecasts: Perfect Forecasts

	O_{yes}	O_{no}
F_{yes}	a	b
F_{no}	c	d

Perfectly accurate forecasts will have:

$b = c = 0$

Accuracy Measures of Binary Forecasts: Hit Rate

	O_{yes}	O_{no}
F_{yes}	a	b
F_{no}	c	d

$$H = (a + d)/n$$

$$n = a + b + c + d$$

Accuracy Measures of Binary Forecasts: False-Alarm Ratio

	O_{yes}	O_{no}
F_{yes}	a	b
F_{no}	c	d

That proportion of forecast events that fail to materialize

FAR = b/(a + b)

Which skill score?

- Different skill scores perform differently
- Sometimes inconsistently
- Scalar skill scores are used, but are necessarily incomplete representations of forecast performance

Accuracy Measures of Multicategory (three) Forecasts

	O_A	O_N	O_B
F_A	R	S	T
F_N	U	V	W
F_B	X	Y	Z

Above		
	a = R	b = S + T
	c = U + X	d = V + W + Y + Z

Below		
	a = Z	b = X + Y
	c = T + W	d = R + S + U + V

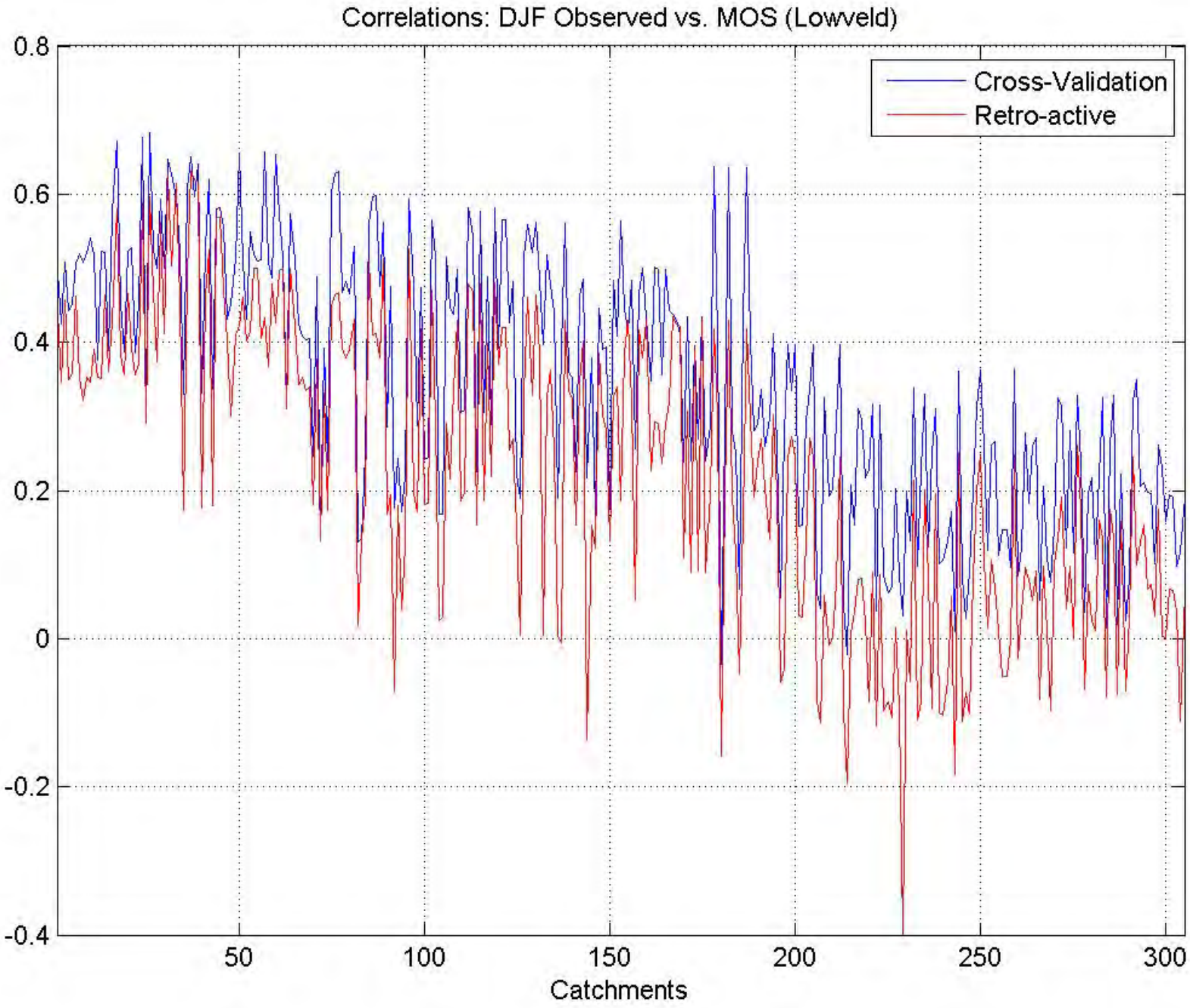
Cross-Validation

	Year1	Year 2	Year 3	Year 4	Year 5	Year 6
Model 1	omitted					
Model 2		omitted				
Model 3			omitted			
Model 4				omitted		
Model 5					omitted	
Model 6						omitted
Model 7						

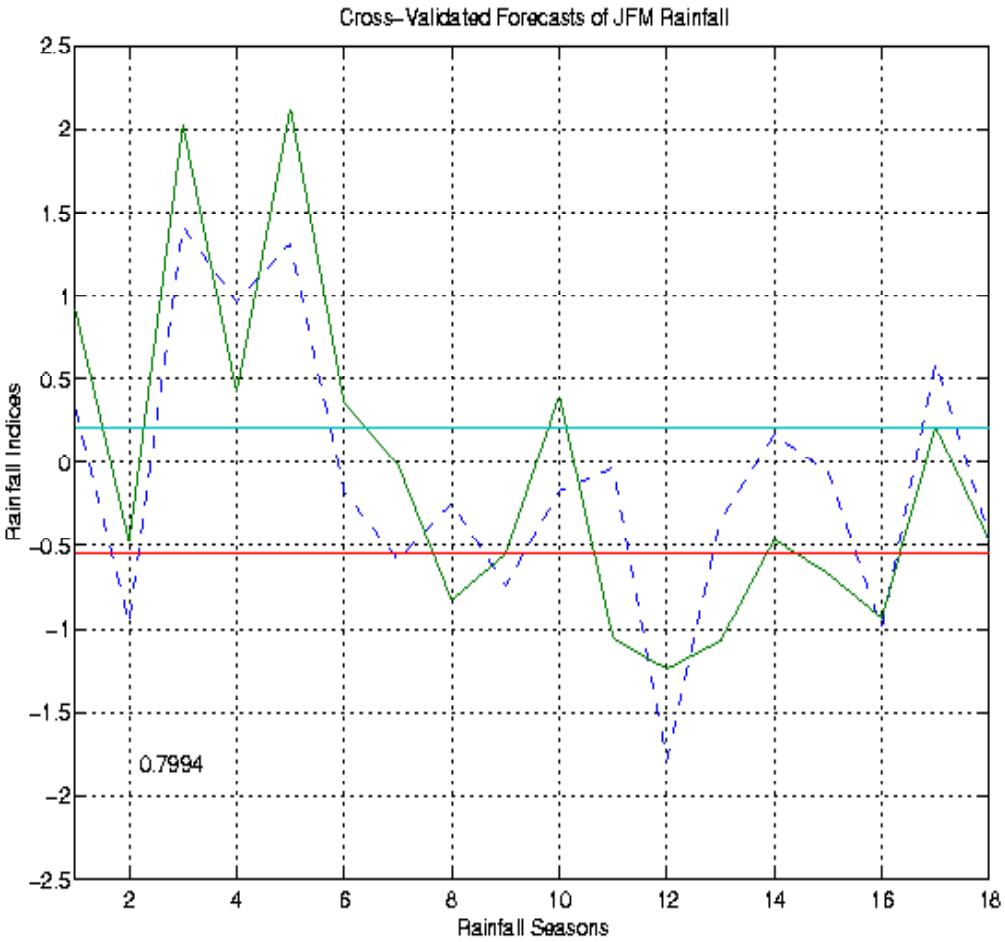
Retro-Active Forecasting

Model 1	**Climate Period**	**Predict Year 1**	**Predict Year 2**	**Predict Year 3**			
Model 2	**Climate Period**				**Predict Year 4**	**Predict Year 5**	**Predict Year 6**

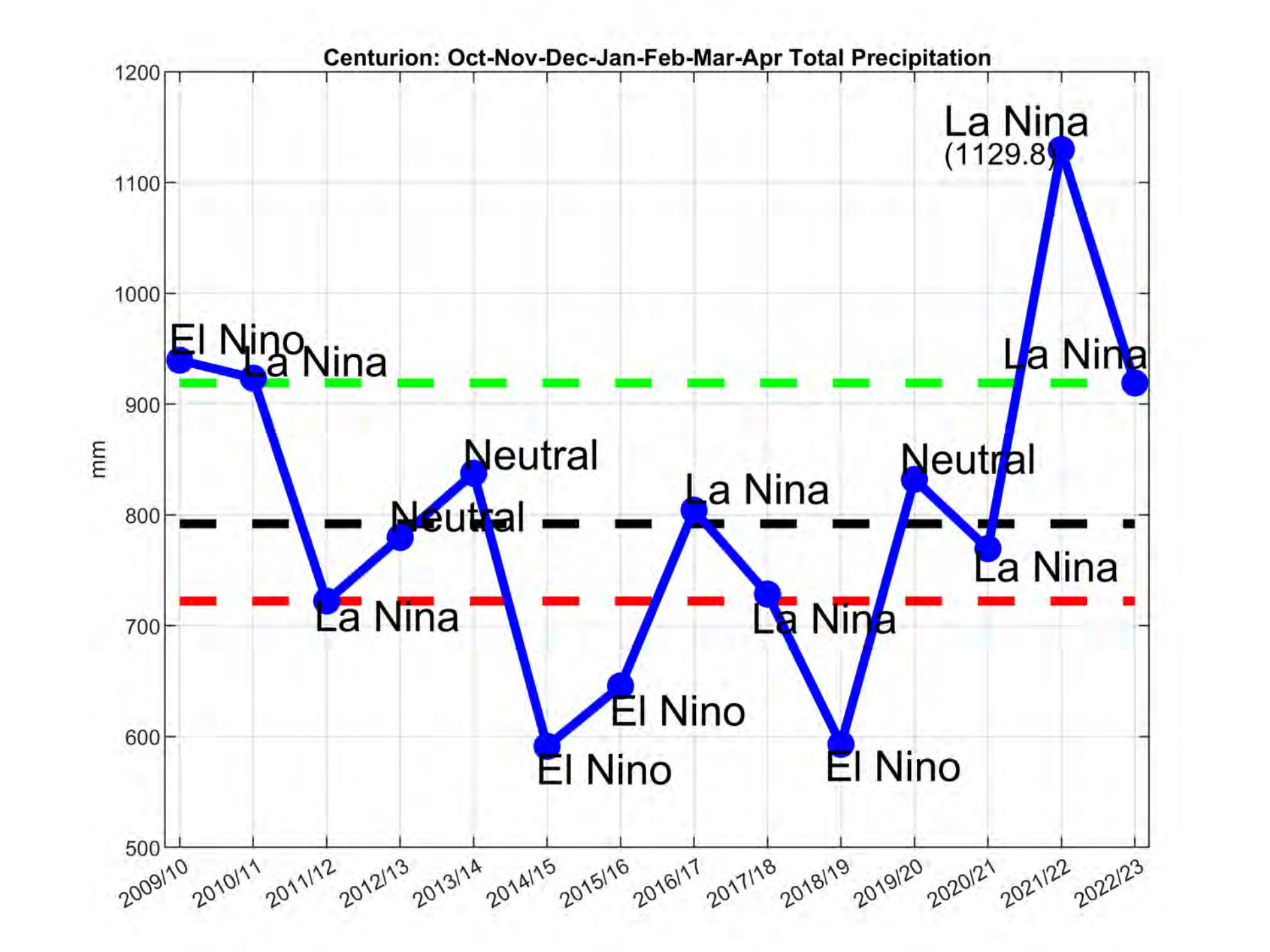
I.e., Model 1 used a 30 year climate period to predict Year 1, 2 and 3. Model 2 used the 30 years of Model 1 AND Year 1, 2 and 3 to predict Year 4, 5 and 6


Correlations: DJF Observed vs. MOS (Lowveld)
Cross-Validation
Retro-active
0.8
0.6
0.4
0.2
0
-0.2
-0.4
50
100
150
200
250
300
Catchments

Example of 3-Category Forecasts



	O_A	O_N	O_B
F_A	4	1	0
F_N	2	2	4
F_B	0	3	2


Centurion: Oct-Nov-Dec-Jan-Feb-Mar-Apr Total Precipitation
mm
1200
1100
1000
900
800
700
600
500
2009/10
2010/11
2011/12
2012/13
2013/14
2014/15
2015/16
2016/17
2017/18
2018/19
2019/20
2020/21
2021/22
2022/23
El Nino
La Nina
La Nina
Neutral
Neutral
La Nina
El Nino
El Nino
La Nina
La Nina
Neutral
La Nina
La Nina
(1129.8)

Heidke Skill Score (HSS)

- Most commonly used skill score for summarizing square contingency tables
- Based on the hit rate as the basic accuracy measure
- For perfect forecasts: HSS = 1; forecasts equivalent to reference forecasts: HSS=0; forecasts worse than ref forecasts: HSS<0
- $HSS = 2(ad-bc) \div [(a+c)(c+d)+(a+b)(b+d)]$ (2X2 situation)

Continuous (can take on any real value) Predictand

Anomaly correlation (AC) or *pattern correlation* resembles the Pearson corr. coef., but grid-point-by-grid-point climatological average values of AC play the roles of the sample mean in Pearson

Mean-Squared Error (MSE): the average squared difference between the forecast and observation k pairs:

$$\mathbf{MSE = n^{-1}\Sigma_k(y_k - o_k)^2}$$

Linear Error in Probability Space (LEPS)

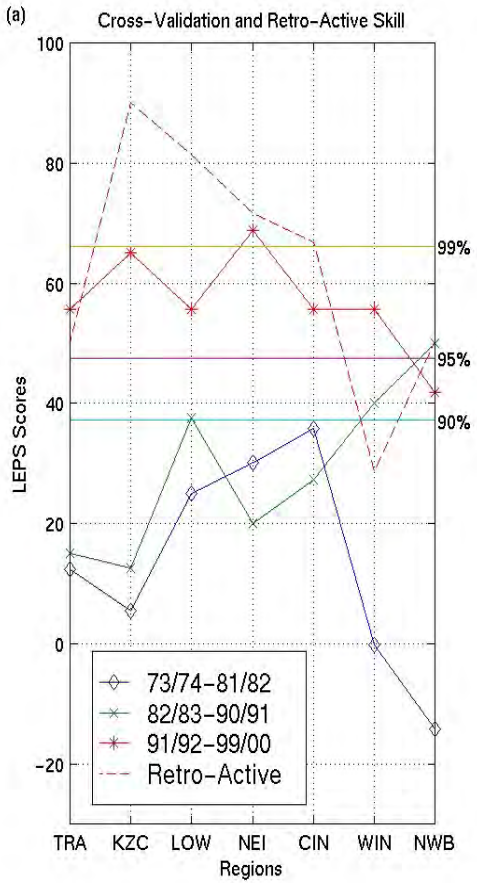
LEPS = $1 - |p_f - p_v|$;

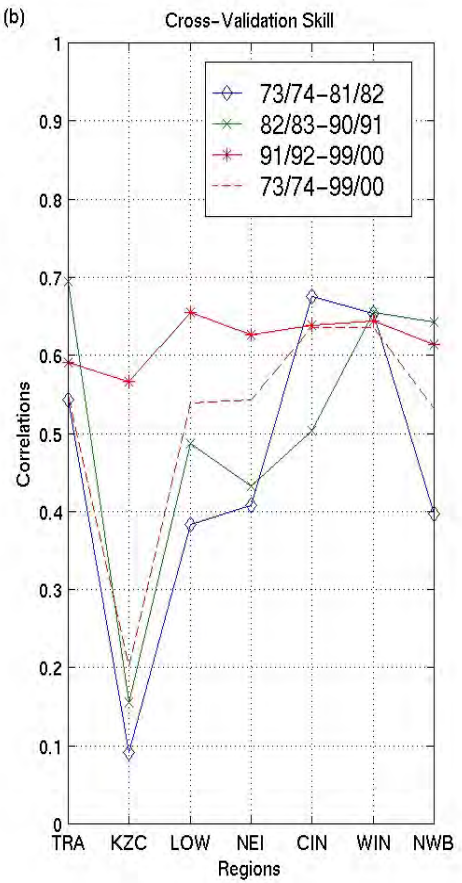
p_f and p_v the cumulative probabilities of the forecast and observed values respectively

For categorical forecasts, the expected score is decreased by:

- a forecast bias for Near-normal (issuing forecasts with a lower variance than the observations)
- issuing forecasts that are two categories out

Decrease in LEPS: Western Interior

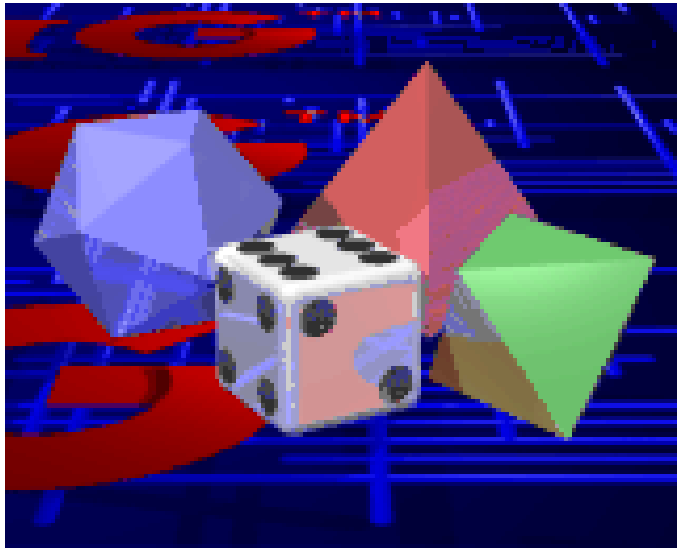




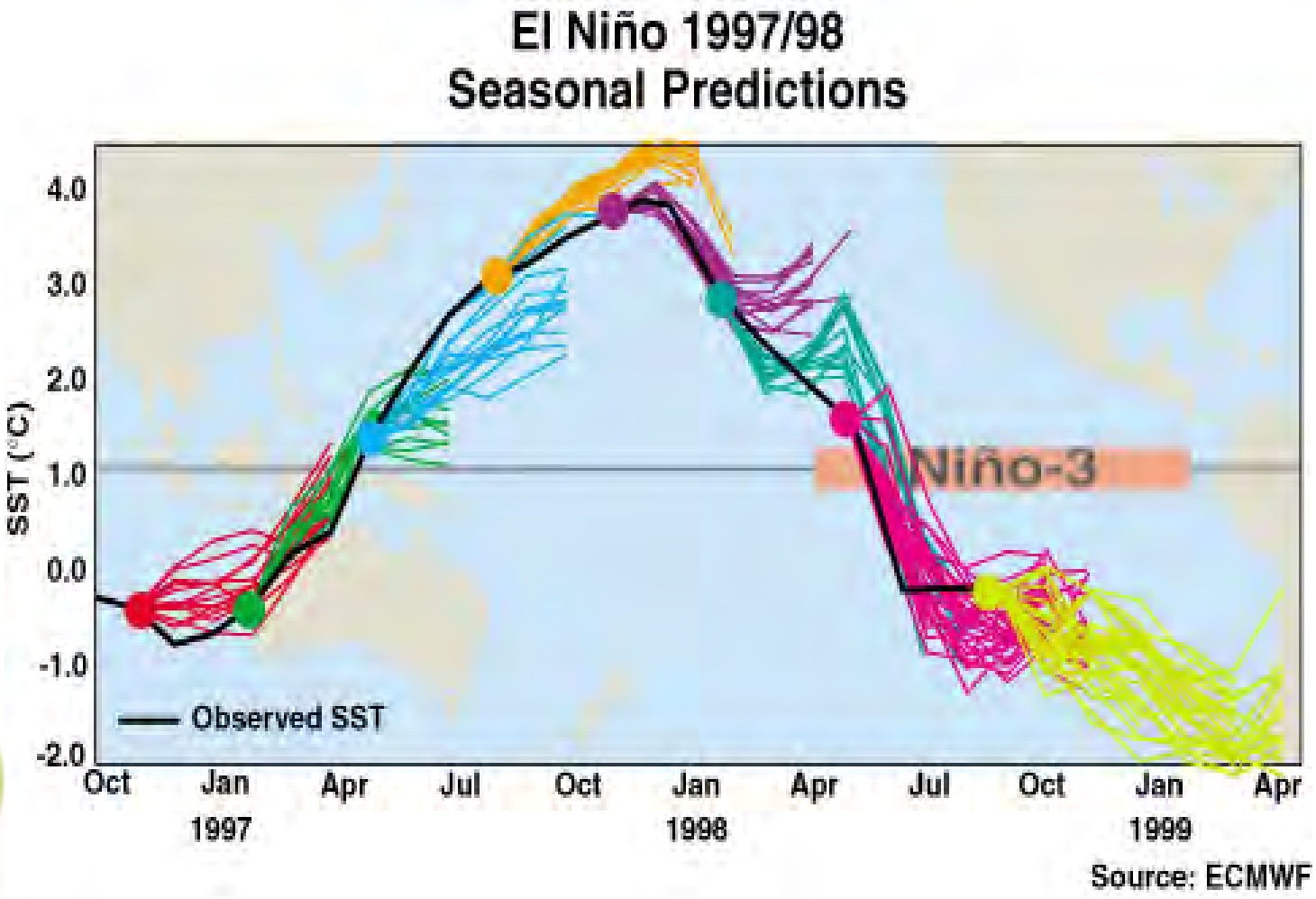
	Hits	1 Cat Out	2 Cat Out
73/74 to 81/82	5	1	3
82/83 to 90/91	5	3	1

Forecast Format

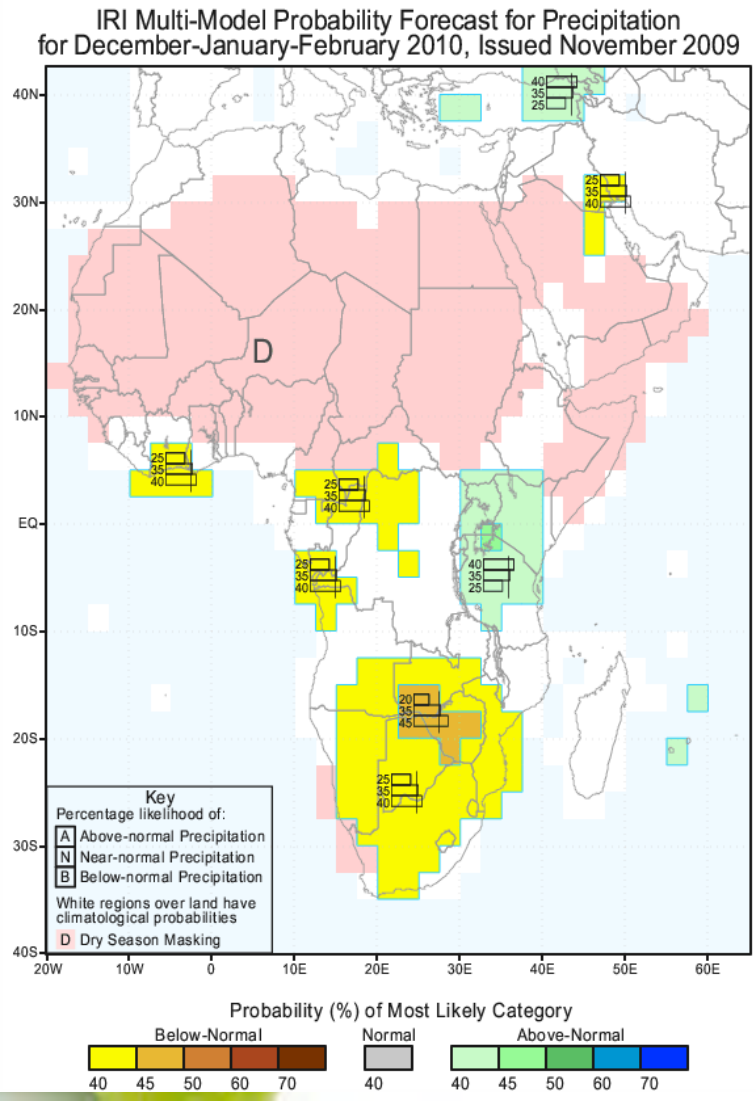
Although boundary conditions provide predictability of the atmosphere at seasonal time scales, the inherent variability of the atmosphere requires seasonal climate forecasts to be expressed ***probabilistically***

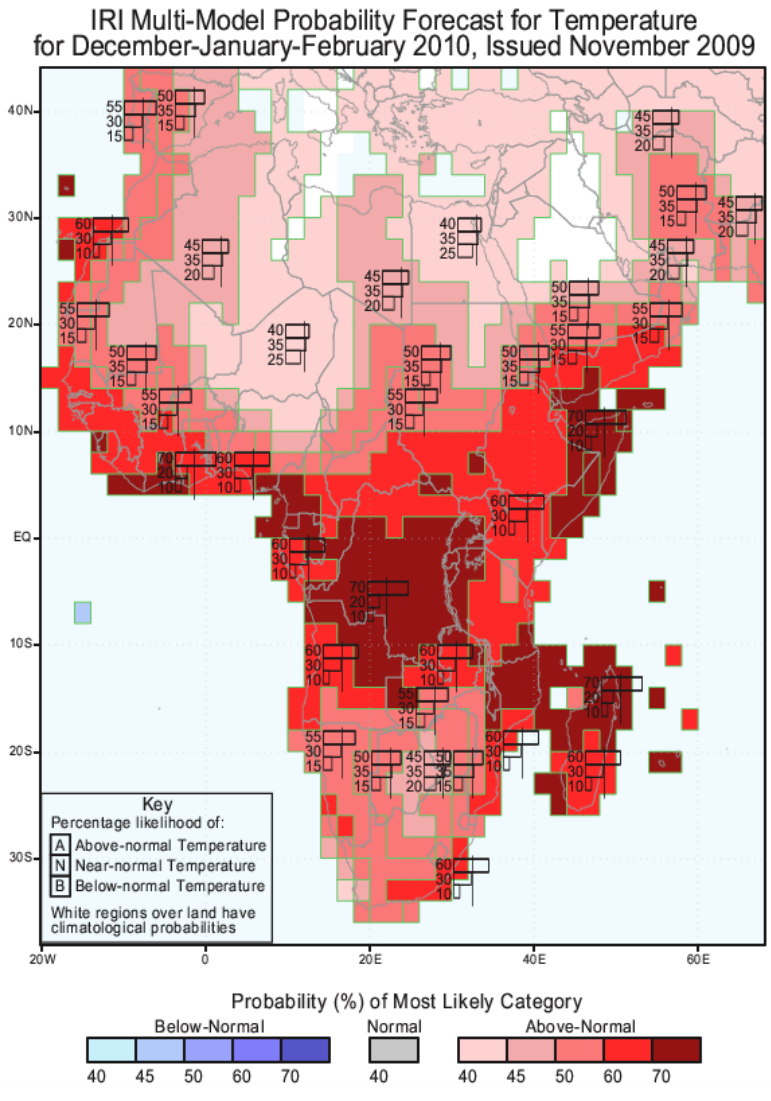


ENSO Probability Forecast



Rainfall and Temperature Probability Forecast



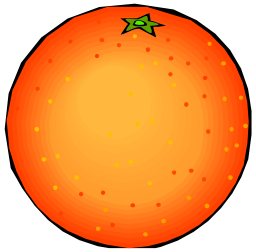
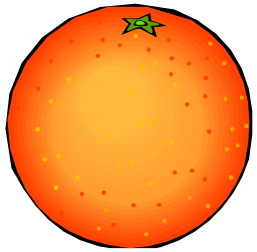


Verification of Probability Forecasts

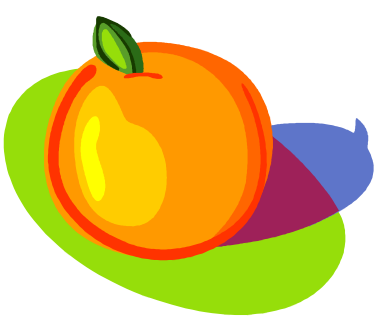
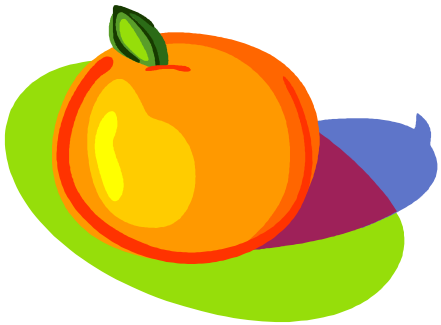
- Unless a probability forecast is either 1.0 or 0.0, it is not clear whether an individual forecast is correct
- However, for probability values between these two extremes, a single forecast is neither “right” nor “wrong”

Interpreting forecast probabilities

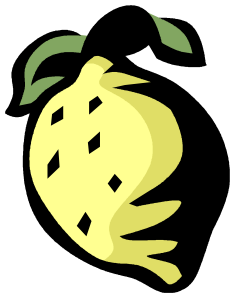
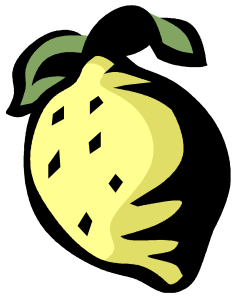
A 33.3%

N 33.3%

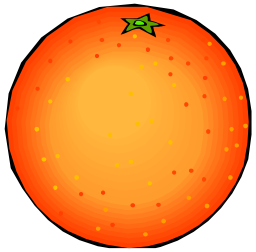



B 33.3%

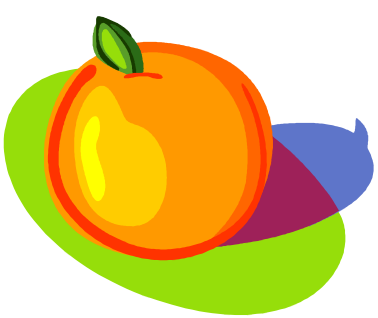



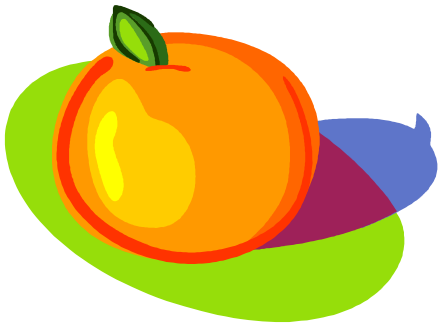
Interpreting forecast probabilities

A 16.7%

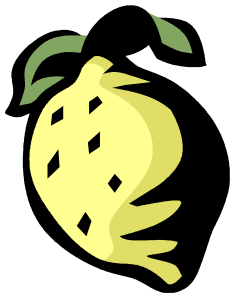


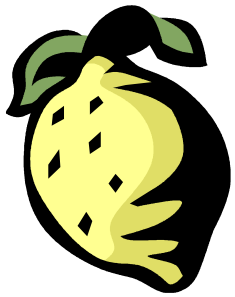
N 33.3%

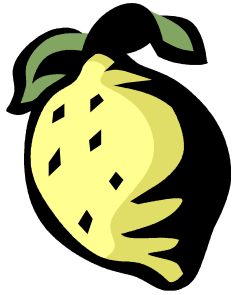




B 50.0%







Reliability

The correspondence between a given probability, and the observed frequency of an event in the case this event is forecast with this probability

For the previous "forecast" to be considered reliable, picking and replacing fruit a 100 times should produce approximately 50 lemon draws ("below-normal") and 17 grapefruit draws ("above-normal")

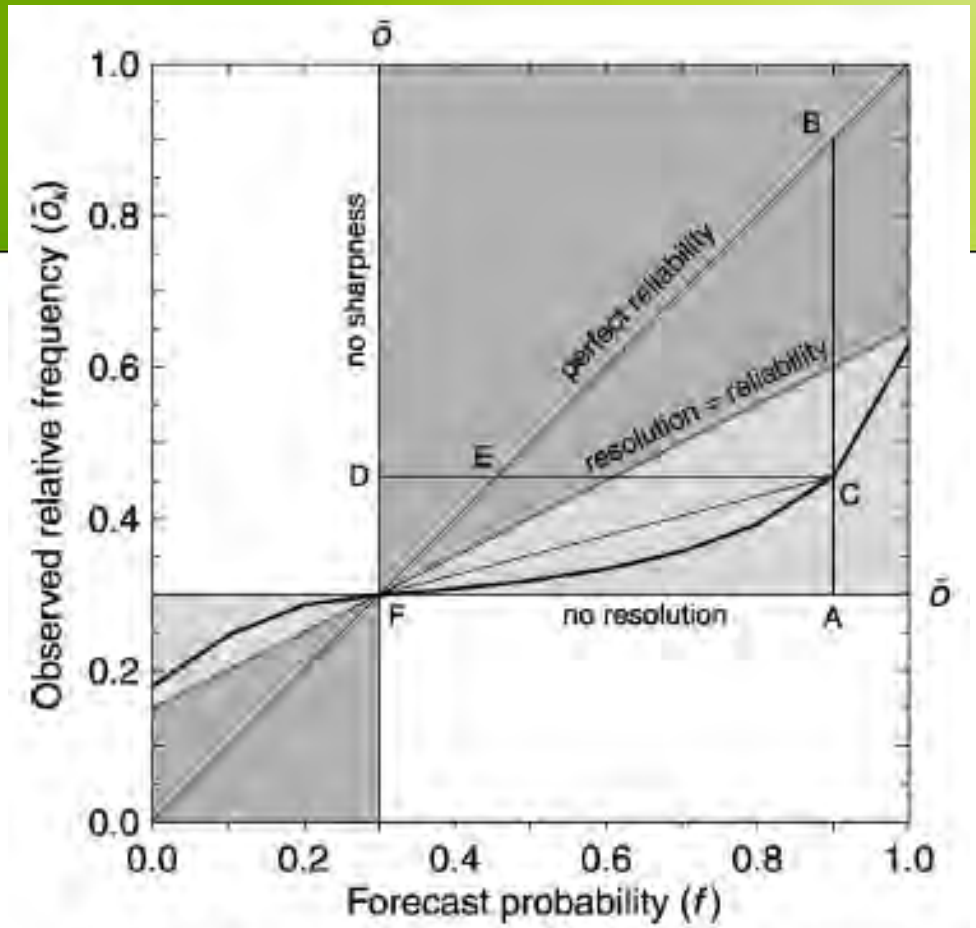


FIG. 1. Attributes diagram showing the areas of skill compared to forecasts of climatology (dark shading) and additional areas of skill compared to random guessing (light shading). The prior probability of the event, $\overline{o}$, is arbitrarily set at 0.3.

On Using "Climatology" as a Reference Strategy in the Brier and Ranked Probability Skill Scores

SIMON J. MASON

International Research Institute for Climate Prediction, Columbia University, Palisades, New York

MULTIMODEL ENSEMBLING IN SEASONAL CLIMATE FORECASTING AT IRI

BY ANTHONY G. BARNSTON, SIMON J. MASON, LISA GODDARD, DAVID G. DEWITT, AND STEPHEN E. ZEBIAK

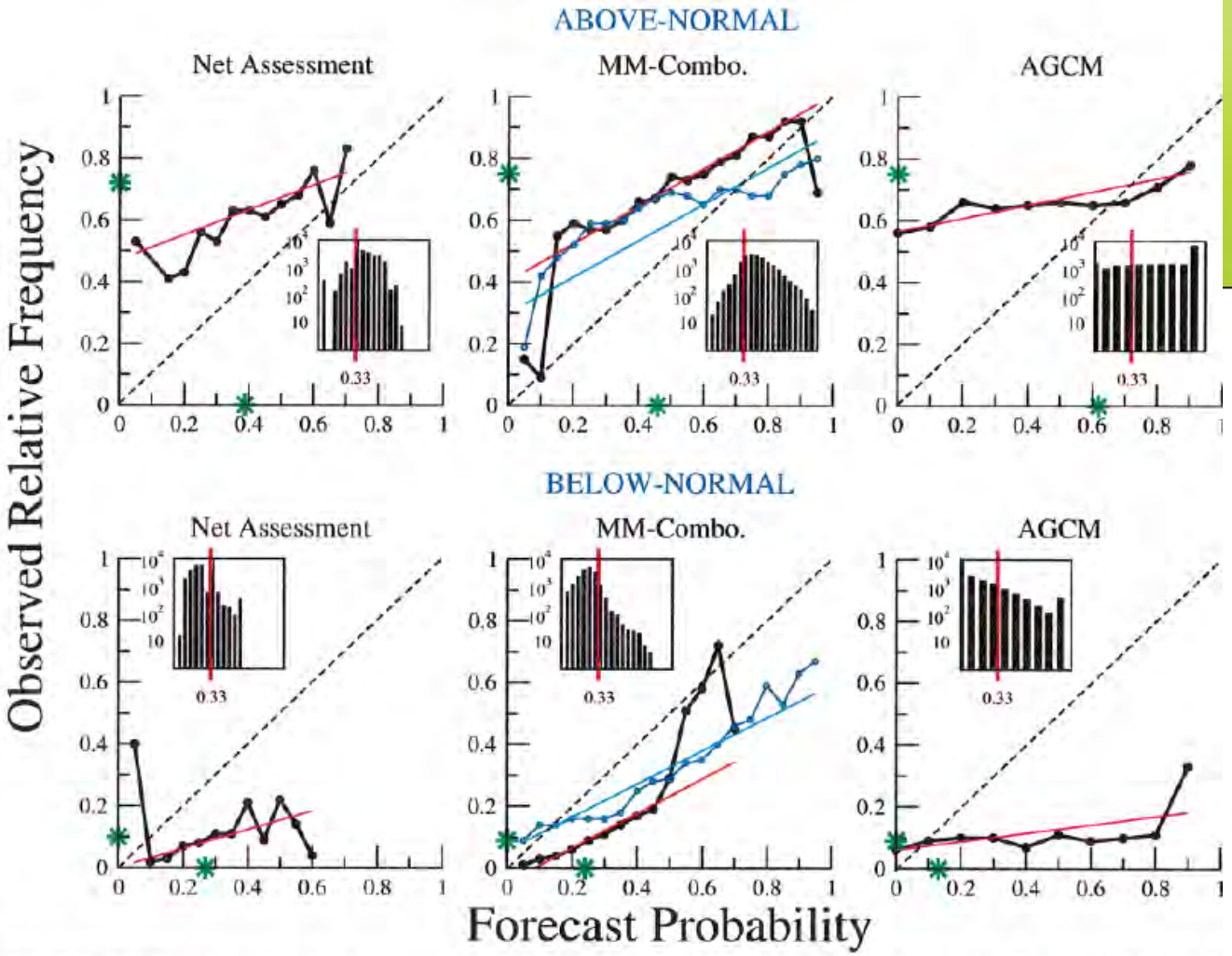


FIG. 5. Reliability diagrams for temperature forecasts at low latitude (30°N–30°S) over the 4-yr period for (left column) the IRI's issued forecasts (called Net Assessment), (middle column) the objective multimodel ensemble prediction, and (right column) for one of the three individual AGCMs. Just one of the AGCMs is included because all three of them have very similar reliability curves. (top row) Forecasts for above-normal temperature, and (bottom row) below-normal temperature. The *x* axis indicates forecast probability, and *y* axis relative observed frequency. The red line is a least-squares regression that takes into account the sample size represented by each point. The green asterisks on the axes show the overall mean of the forecast probabilities and observed relative frequencies. The inset histograms show the frequency with which each category of probability was forecast (with a logarithmic frequency scale), with the climatological probability (0.33) shown by the red vertical line. In the case of the multimodel combination (middle column), the blue dots and regression lines show results using a simple pooling (equal weighting) scheme as opposed to the two multimodel schemes. For IRI forecasts and multimodel predictions, the probability categories are centered on integer multiples of 0.05, while for the AGCM they are centered on integer multiples of 0.10.

The Reliability Diagram

Event A was forecast to occur with the probabilities on the x-axis

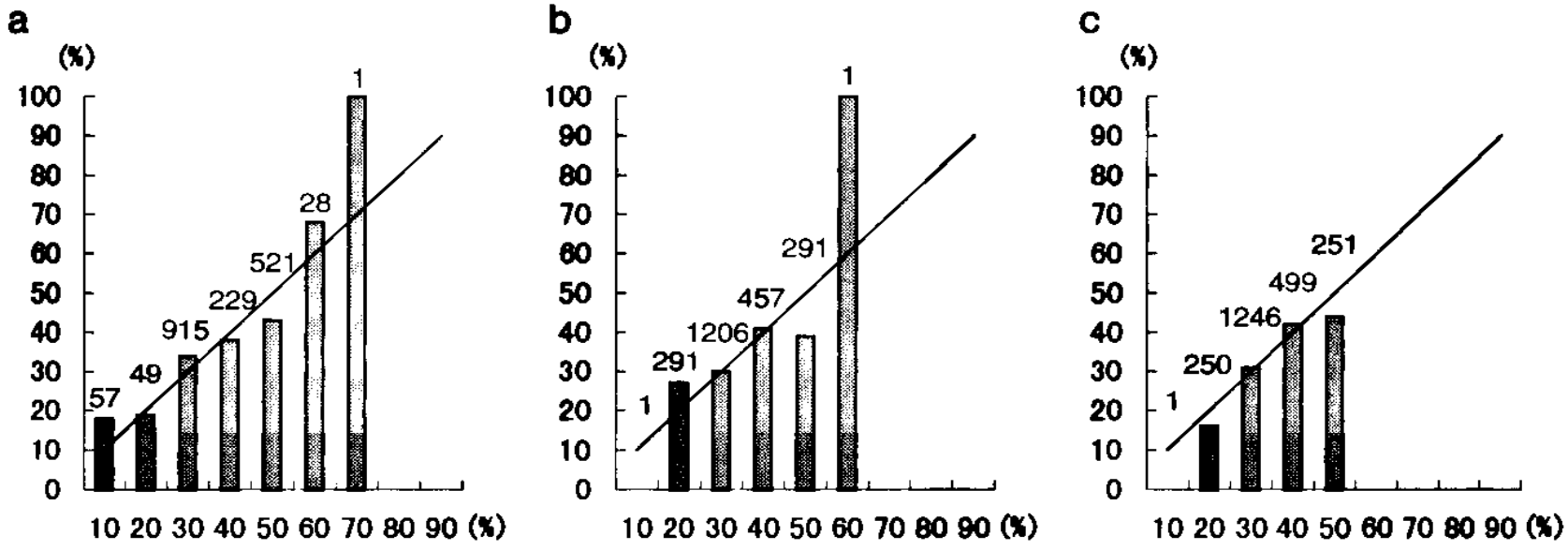
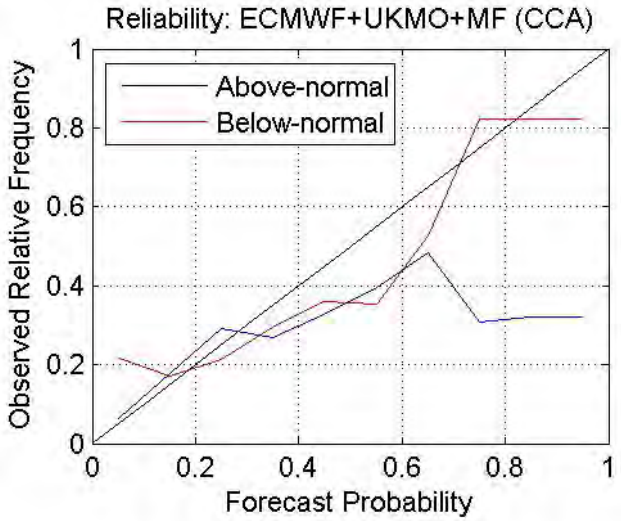


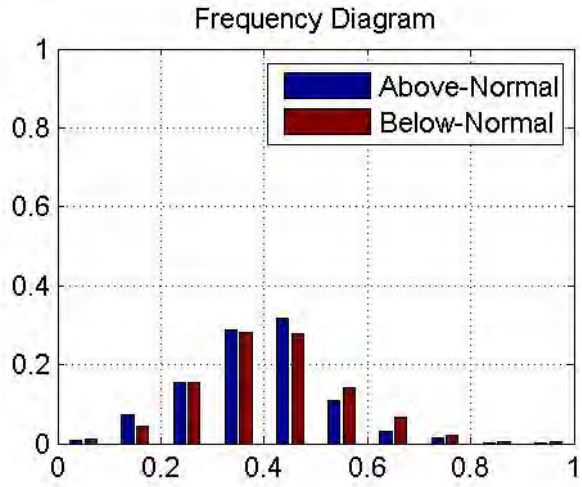
Figure 8 Reliability diagrams based on operational forecasts since March 1996. Abscissa is predicted probability, number at the top of column is total number of prediction of the probability, and ordinate is the ratio of observed occurrence. (a) Surface temperature, (b) precipitation, and (c) sunshine hours. Auxiliary line is the line when predicted probability is equal to the observed occurrence rate.

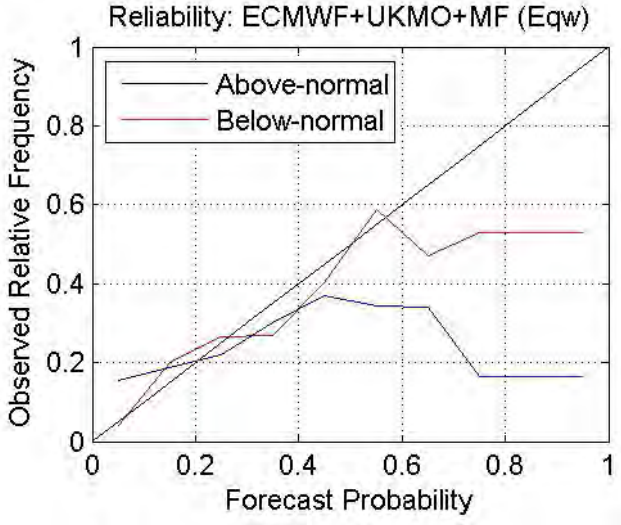
Ranked Probability Score

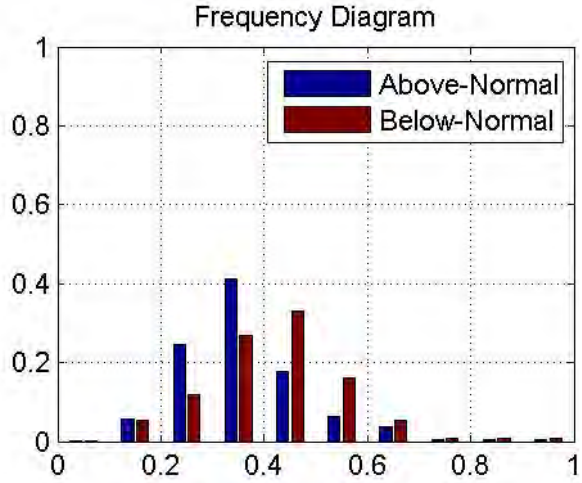
- Desired for verification of probability forecasts and is a measure that is sensitive to distance: forecasts are penalized increasingly as more probability is assigned to event categories further removed from the actual outcome
- Many scalar scores sensitive to distance exist, but of these the Ranked Probability Score (RPS) is preferred

Reliability: Southern Africa









Calculation of RPS

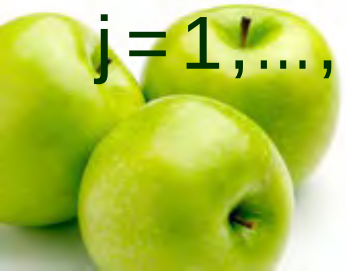
$$RPS = \Sigma_m [(\Sigma_j y_j) - (\Sigma_j o_j)]^2$$

Σy_i = cumulative forecasts

Σo_i = cumulative observations

$m = 1, \ldots, j$ the number of forecast categories

$j = 1, \ldots, m$ the number of components



Example of RPS Calculation

Forecast Category	Probability Forecast for three seasons		Season 1		Season 2		Season 3	
			Obs Category	Cum Obs Category	Obs Category	Cum Obs Category	Obs Category	Cum Obs Category
A	20%	0.2	1	1	0	0	0	0
N	50%	0.7	0	1	1	1	0	0
B	30%	1.0	0	1	0	1	1	1

- Season 1: RPS=$(0.2-1)^2+(0.7-1)^2=0.73$
- Season 2: RPS=$(0.2-0)^2+(0.7-1)^2=0.13$
- Season 3: RPS=$(0.2-0)^2+(0.7-0)^2=0.53$

Forecasts that are less than perfect receive scores that are high, so the RPS has a *negative orientation*

Example of RPS Calculation

Forecast Category	***Deterministic*** Forecast for three seasons		Season 1		Season 2		Season 3	
			Obs Category	Cum Obs Category	Obs Category	Cum Obs Category	Obs Category	Cum Obs Category
A	0%	0.0	1	1	0	0	0	0
N	0%	0.0	0	1	1	1	0	0
B	100%	1.0	0	1	0	1	1	1

- Season1: $RPS=(0.0-1)^2+(0.0-1)^2=2.0$
- Season 2: $RPS=(0.0-0)^2+(0.0-1)^2=1.0$
- Season 3: $RPS=(0.0-0)^2+(0.0-0)^2=0.0$

  For a perfect forecast, RPS=0

Definition of Forecast ***Skill***

- The relative accuracy (the average correspondence between individual forecasts and the events they predict) of a set of forecasts, with respect to some set of standard control, or reference, forecasts

- Choices for reference forecasts:
  - climatological average values of predictand
  - persistence forecasts
  - baseline

The Skill Score (SS)

For a particular measure of accuracy A:

$$SS_{ref} = (A - A_{ref})/(A_{perf} - A_{ref})$$

If $A = A_{perf}$ then $SS_{ref} = 1$

If $A = A_{ref}$ then $SS_{ref} = 0$

Ranked Probability Skill Score

- For a collection of n forecasts:

$$mRPS = n^{-1}\Sigma_k RPS_k$$

- $RPSS = (mRPS - mRPS_{clim})/(0 - mRPS_{clim})$
- $RPSS = 1 - mRPS/mRPS_{clim}$

Example of RPS_{clim} Calculation

			Season 1		Season 2		Season 3	
Forecast Category	Probability Forecast		Obs Category	Cum Obs Category	Obs Category	Cum Obs Category	Obs Category	Cum Obs Category
A	33.3	0.33	1	1	0	0	0	0
N	33.3	0.67	0	1	1	1	0	0
B	33.3	1.00	0	1	0	1	1	1

- Season 1: $RPS_{clim}=(0.33-1)^2+(0.67-1)^2=0.56$
- Season 2: $RPS_{clim}=(0.33-0)^2+(0.67-1)^2=0.22$
- Season 3: $RPS_{clim}=(0.33-0)^2+(0.67-0)^2=0.56$

RPSS for the seasons presented above:

- RPSS = 1.0 – [1/3(0.73+0.13+0.53)]/[1/3(0.56+0.22+0.56)]
- RPSS = 1.0 – 0.463/0.444
- RPSS = 1.0 – 1.0425
- RPSS = -0.0425
- A negative value of RPSS implies that the skill of the estimated probabilities as the forecast is worse than the use of climatological probabilities as the forecast

Relative Operating Characteristics

- The ROC is a representation of the skill of a forecast system in which the hit rate and the false-alarm rate are compared
- For skilful forecast systems ROC curves bend towards the top left where hit rates are higher than false-alarm rates. Curves close to the diagonal imply little or no useful information, while curves below the diagonal imply negative skill.

ROC curves

Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels

SIMON J. MASON AND NICHOLAS E. GRAHAM

International Research Institute for Climate Prediction, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

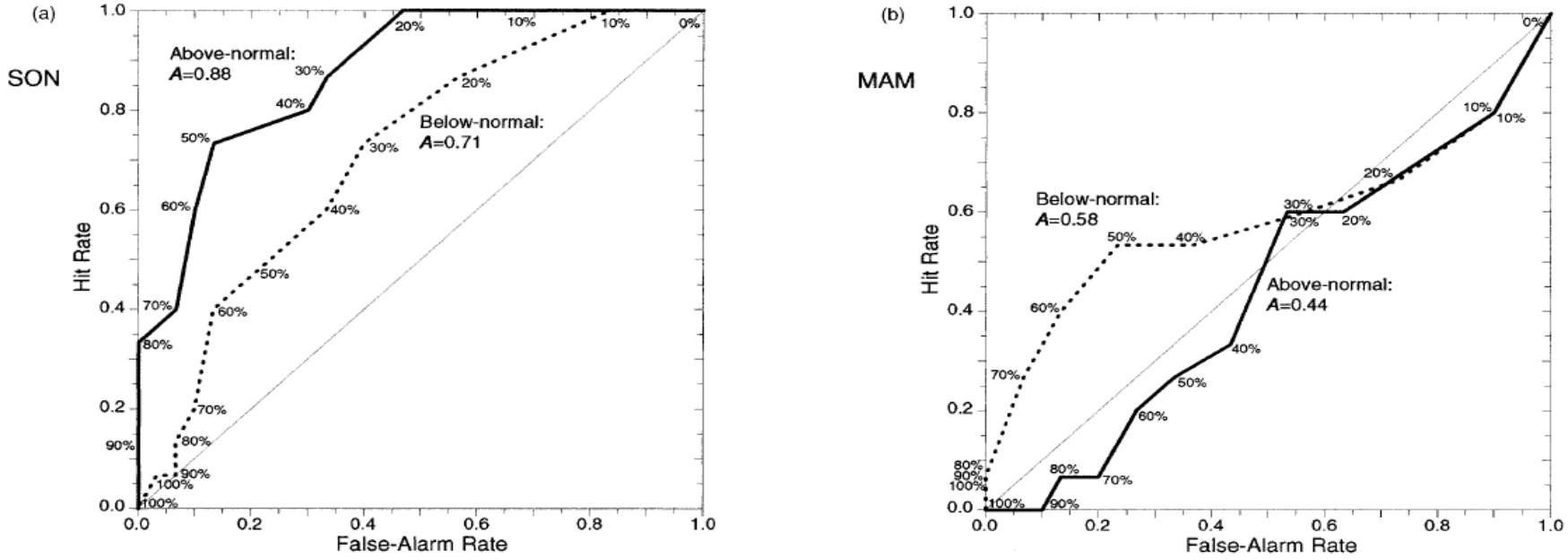
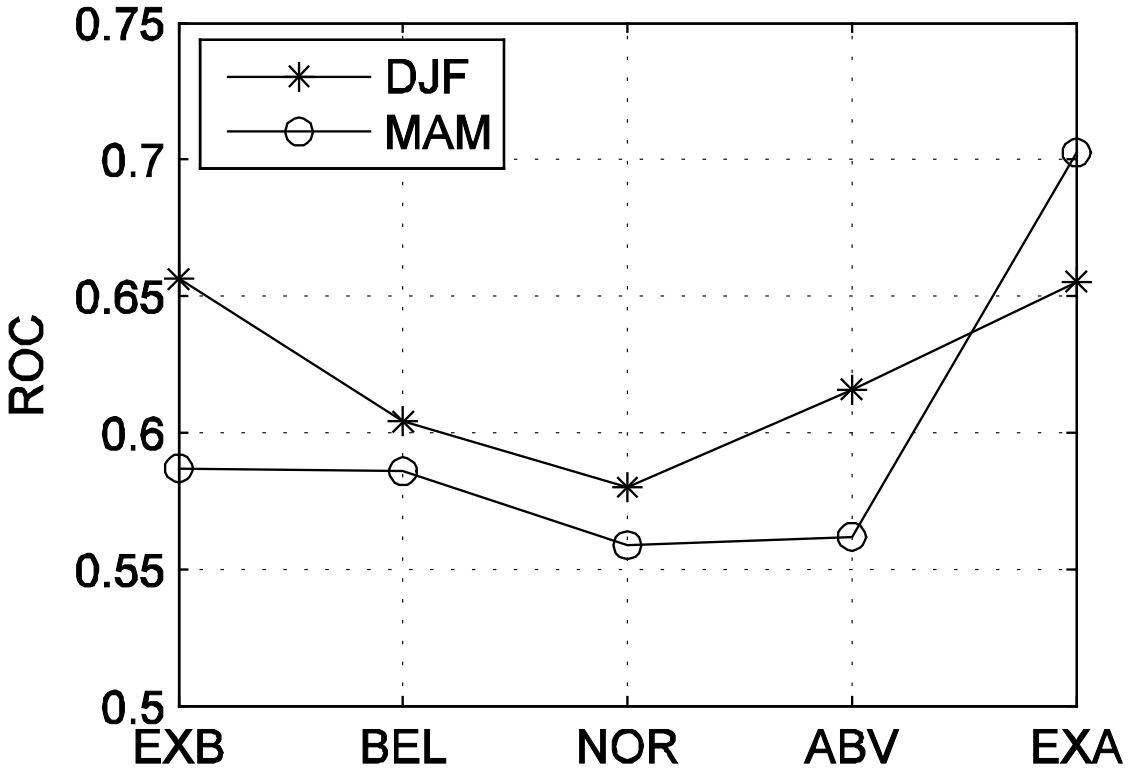


FIG. 1. Hit rates vs false-alarm rates for (a) Sep–Nov and (b) Mar–May area-averaged rainfall for eastern Africa (10°N–10°S, 30°–50°E) from 1950 to 1994. The hit and false-alarm rates were calculated using rainfall simulated by the ECHAM3-T42 general circulation model forced with observed sea surface temperatures and using 10 ensemble members. Results are shown for the simulation of rainfall in the upper (solid line) and lower (dotted line) terciles. Rates are indicated

V-shape of ROC curve



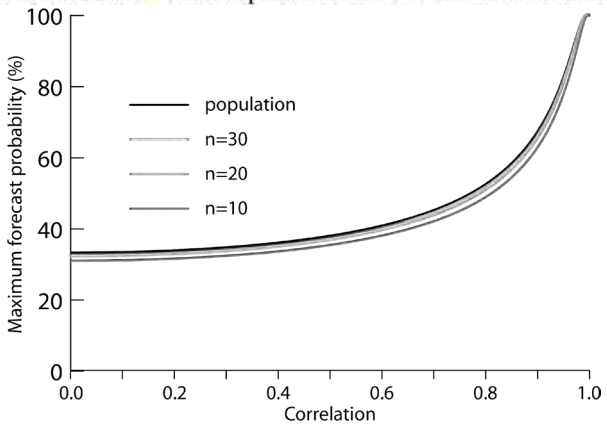
5 categories

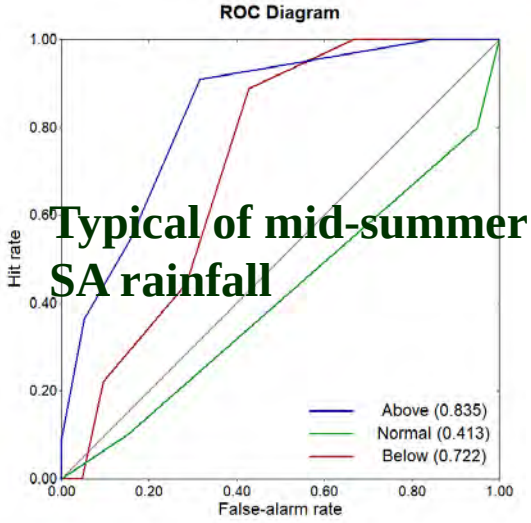
Lessons regarding "normal" forecasts

Received: 25 June 2020 | Revised: 11 November 2020 | Accepted: 19 December 2020 | Published on: 10 February 2021
DOI: 10.1002/qj.3968
Quarterly Journal of the Royal Meteorological Society RMetS
RESEARCH ARTICLE

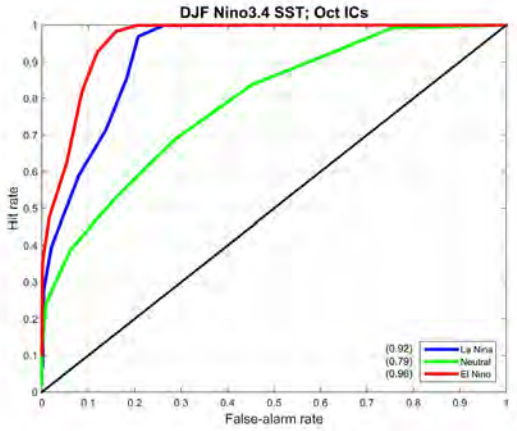
Forecasts of "normal"

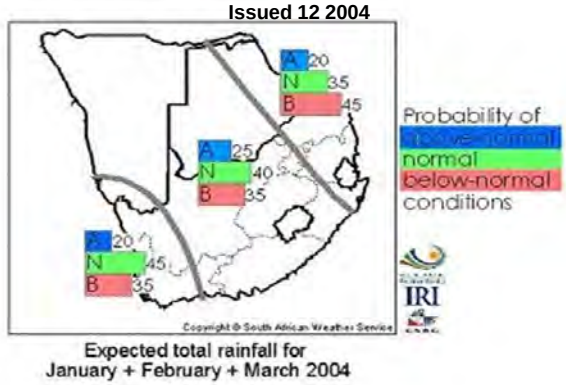
Simon J. Mason[1] | Christopher A. T. Ferro[2] | Willem A. Landman[3]

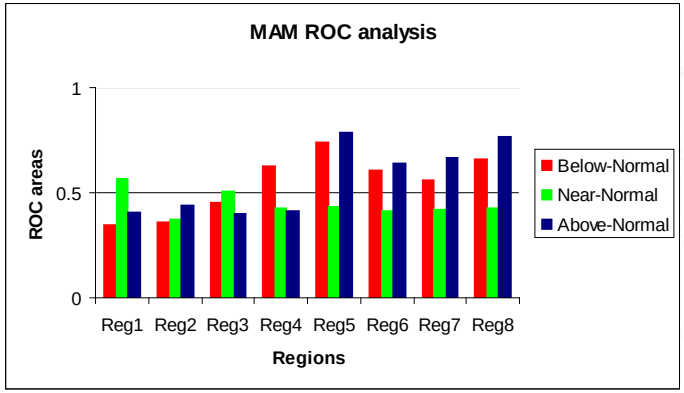


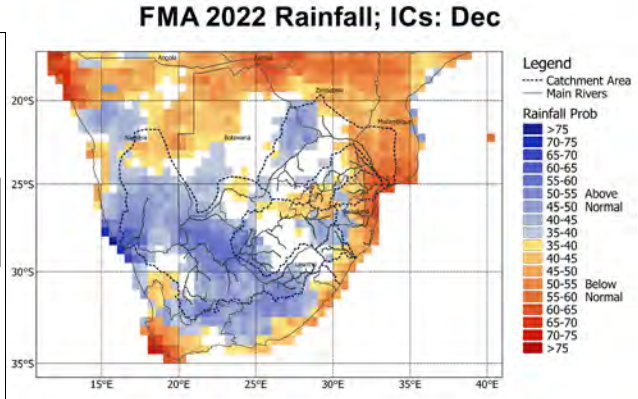


Typical of mid-summer SA rainfall



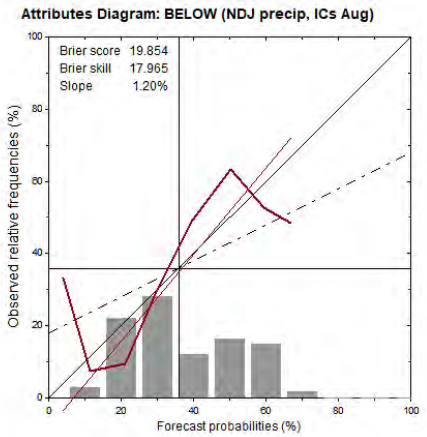


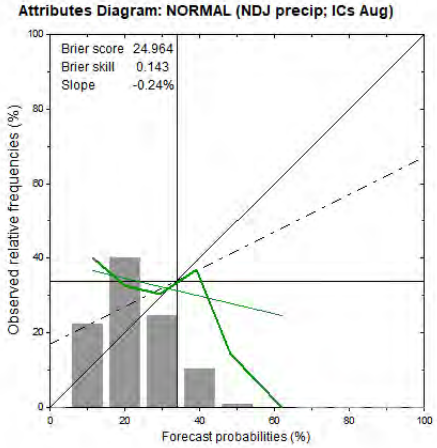


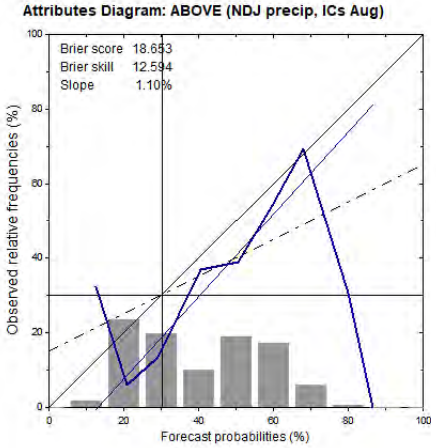


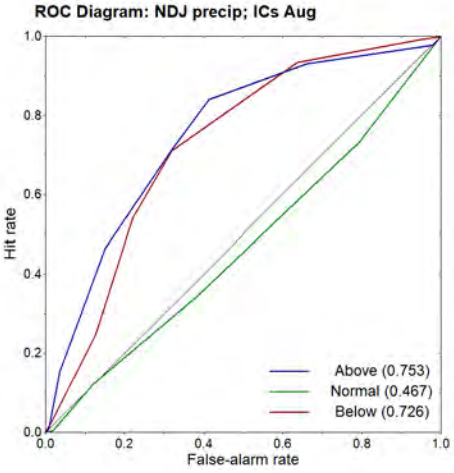
Seasonal Forecast Worx

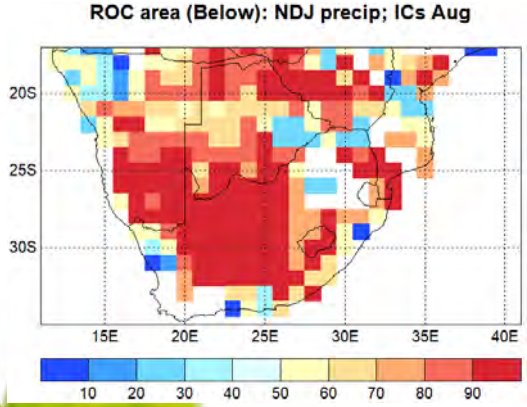
Verification of *real-time* seasonal forecasts: 2018/19 – 2022/23

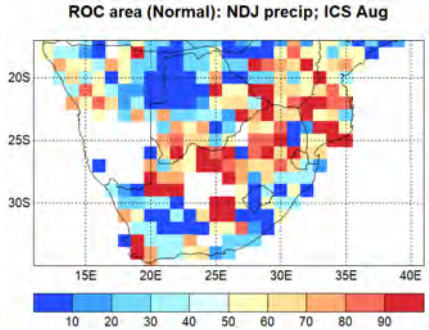


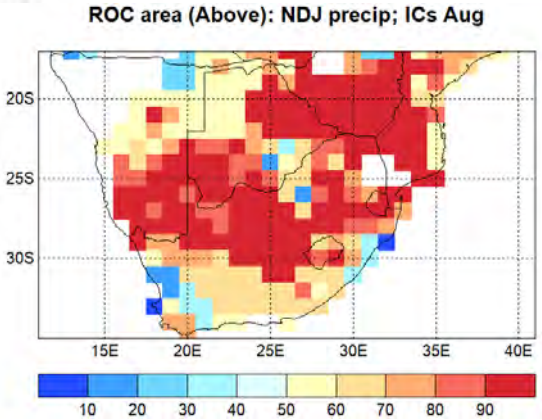












Can users understand verification statistics?

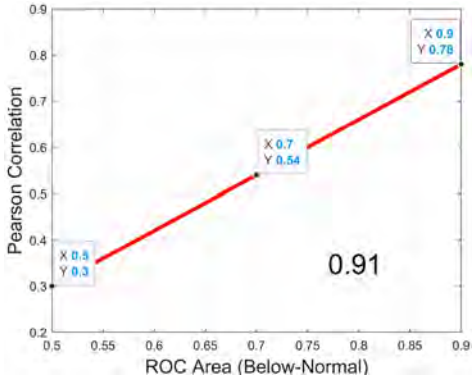
Water SA 49(3) 192–198 / Jul 2023
https://doi.org/10.17159/wsa/2023.v49.i3.4058

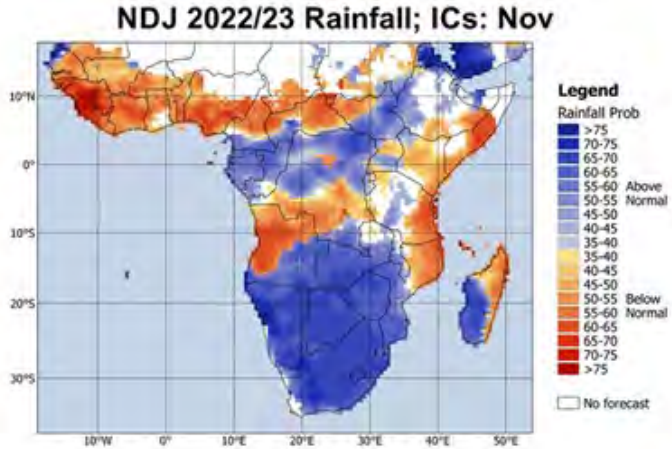
Research paper

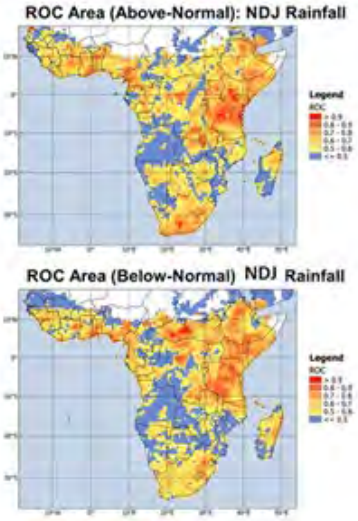
Probabilistic vs deterministic forecasts – interpreting skill statistics for the benefit of users

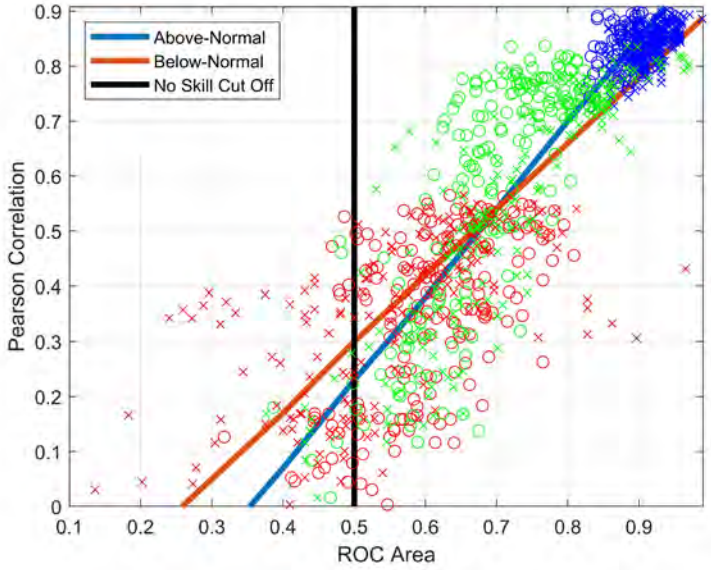
Willem A Landman[1], Mark Tadross[2], Emma Archer[1] and Peter Johnston[2]

[1]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa
[2]Climate System Analysis Group, University of Cape Town, Cape Town, South Africa



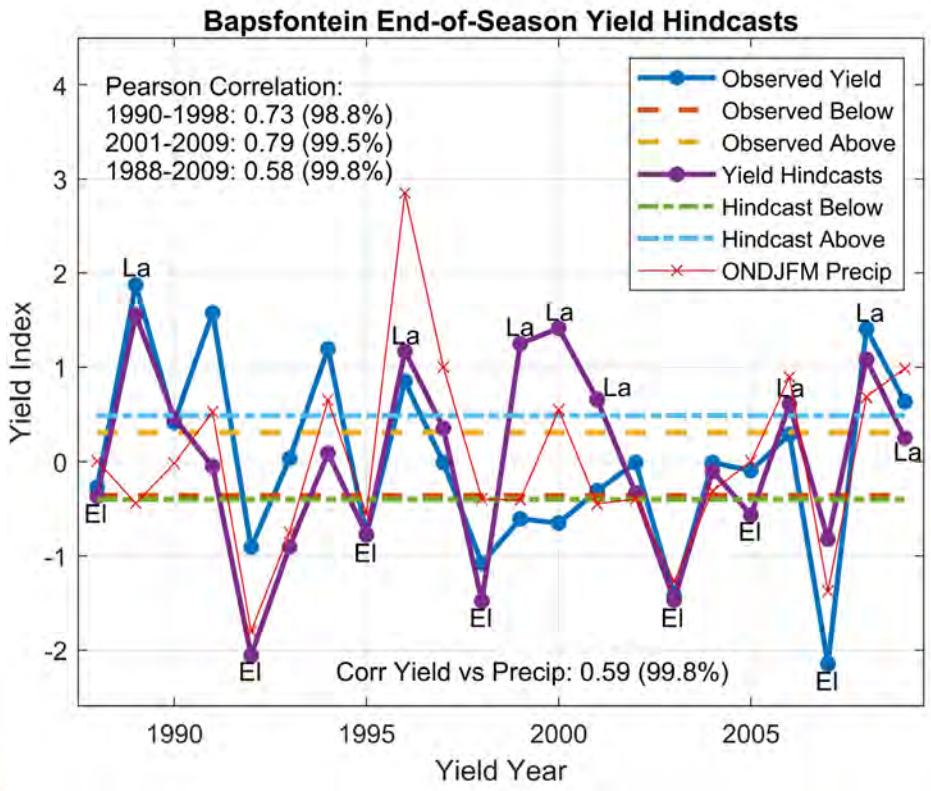


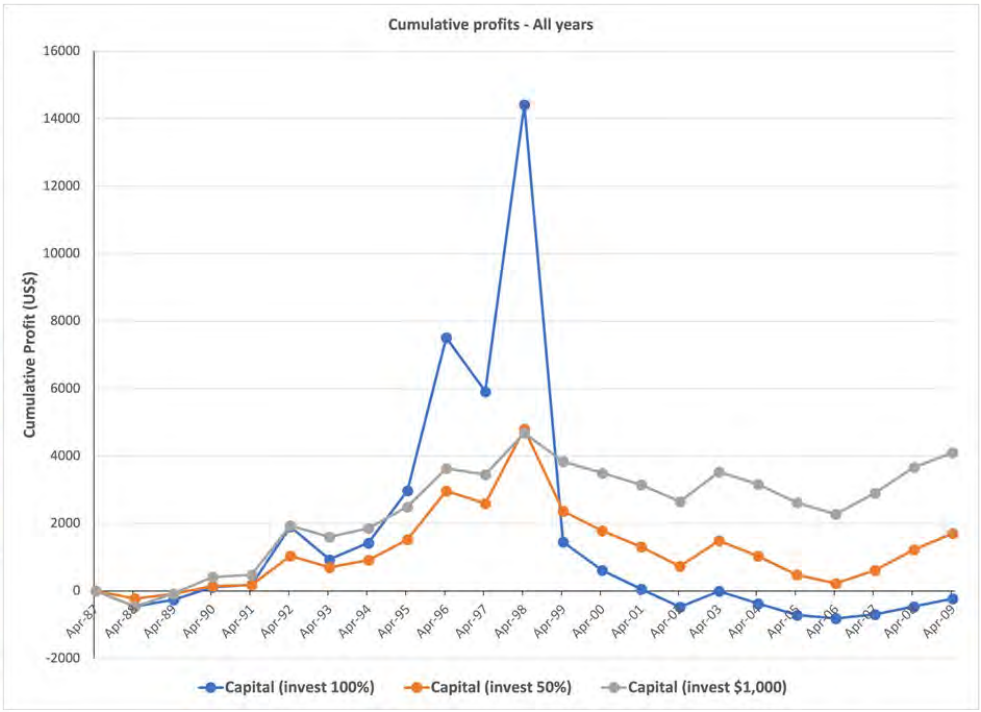




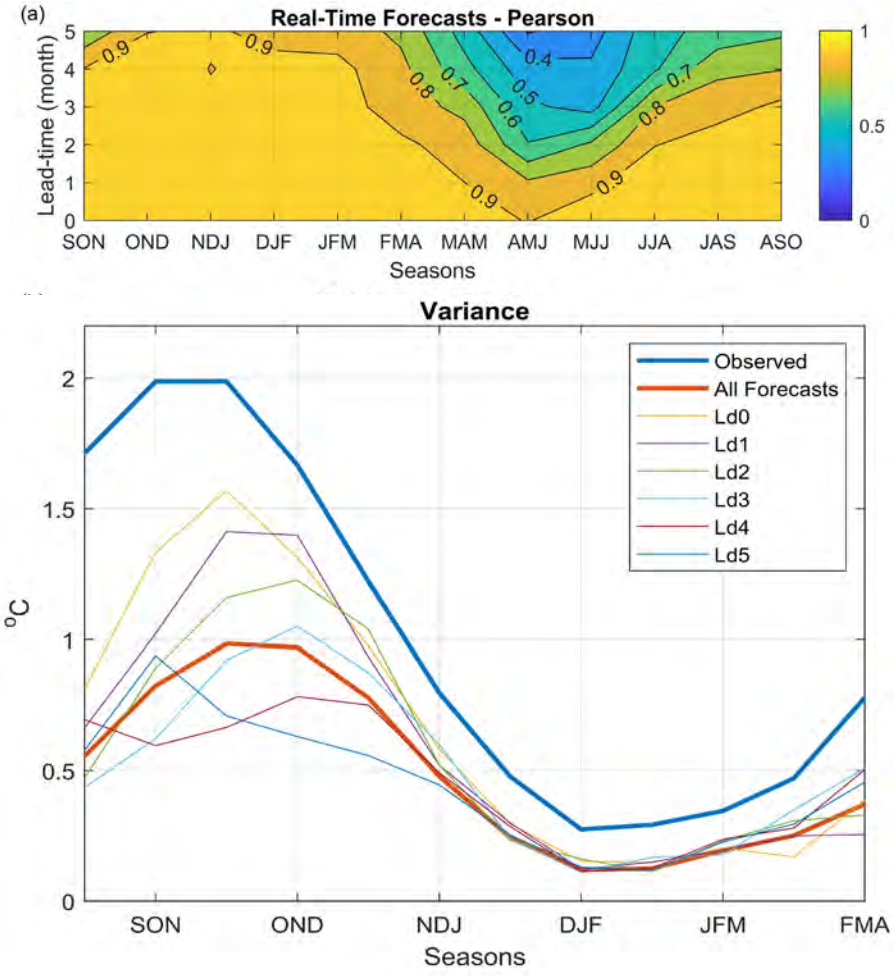
Seasonal forecast characteristics influence the financial success of farming strategies

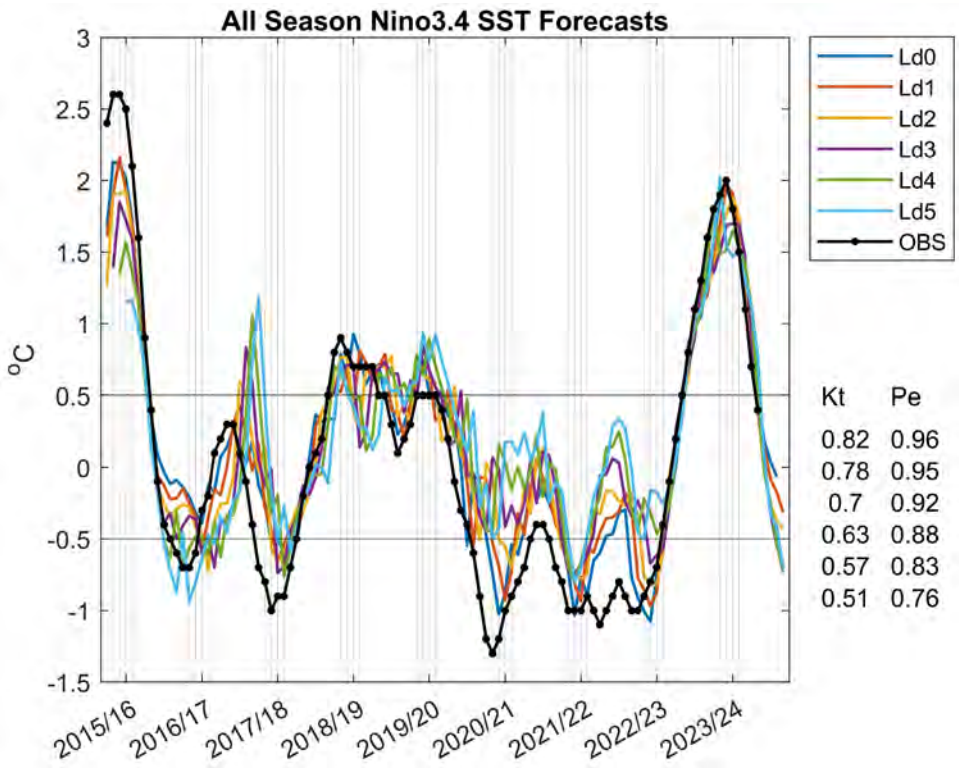
Willem A. Landman, Mark Tadross, Peter Johnston, Olivier Crespo, Emma Archer





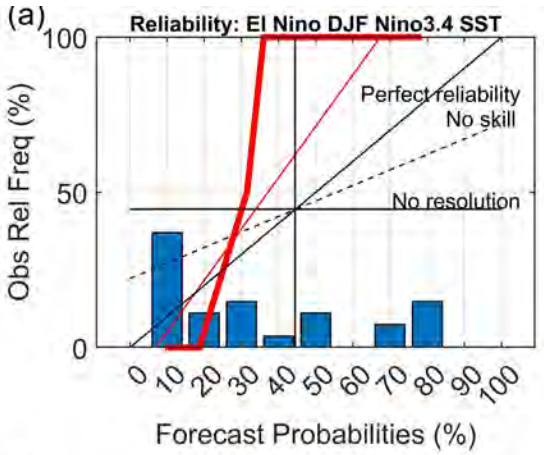
ENSO prediction capability in SA (1)

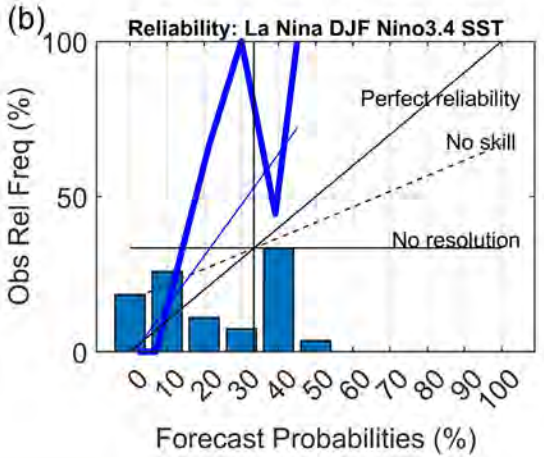


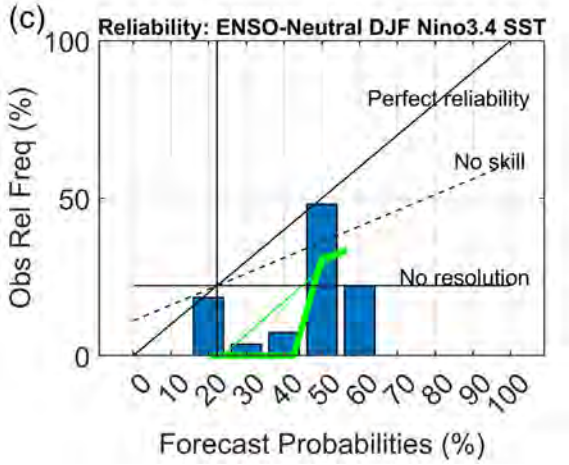


Kt	Pe
0.82	0.96
0.78	0.95
0.7	0.92
0.63	0.88
0.57	0.83
0.51	0.76

ENSO prediction capability in SA (2)







Leads 3 to 5 combined, resulting in 9 x 3 cases per season