



Verification of forecasts from the SWFDP – Southern Africa *and E. Africa*

Laurence Wilson

lawrence.wilson@ec.gc.ca

Co-chair, WMO Joint Working Group on Forecast
Verification Research (JWGFVR)

6th International Verification Methods Workshop

6th International Verification Methods Workshop New Delhi, India



Tutorial Session: 13-15 March 2014
National Centre for Medium Range Weather Forecasting
Noida

Scientific Workshop: 17-19 March 2014
India Meteorological Department
New Delhi

The tutorial will include lectures and exercises on the following topics:

- Introduction to R
- Basic verification concepts
- Verification of continuous, categorical, probability and ensemble forecasts
- Statistical inference
- Verification of very high resolution NWP
- Warning verification
- Tropical cyclone verification
- Heavy rain verification
- Climate projection evaluation
- Operational verification systems

In addition, the participants will be offered the opportunity to submit data sets for use in projects to be carried out during the workshop. Space for the tutorial workshop is limited.

Deadline for tutorial applications :
30 November 2013

Additional information :
www.ncmrwf.gov.in/verif2014

The scientific workshop will include both invited and contributed oral and poster presentations. Papers are solicited on all aspects of forecast verification methodology, including:

- Verification of high impact weather forecasts and warnings
- Verification of ensembles and probability forecasts
- Evaluation of seasonal and long-term climate forecasts and projections
- Propagation of uncertainty
- Regional verification
- Aviation forecast verification
- Hydrological verification
- User focused verification for impacts and special events
- Verification tools
- New data sources for verification

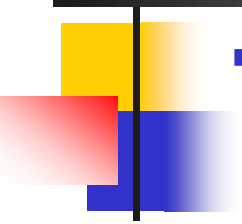
Deadline for submission of abstracts :
15 January 2014

Scientific program committee chair :
Dr. Beth Ebert, BOM, Australia, e.berth@bom.gov.au

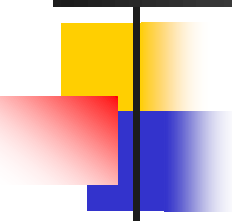
Local organizing committee chair:
Gopal Iyengar, NCMRWF, India, gopal.iyengar@nic.in



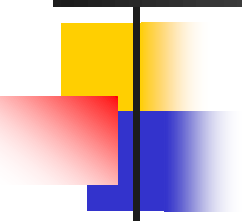
Resources

- 
- Resources:
 - The EUMETCAL training site on verification – computer aided learning:
 - <http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/courses/msgcrs/index.htm>
 - The website of the Joint Working Group on Forecast Verification Research:
 - <http://www.cawcr.gov.au/projects/verification/>
 - This contains definitions of all the basic scores and links to other sites for further information
 - For the SWFDP
 - Presentation on RSMC website
 - Document “Verification of forecasts from the African SWFDPs” to be published soon by WMO.

Outline

- 
- Introduction: What is verification?
 - Why verify? Purposes and Principles of verification
 - Gathering the data – the event form
 - Hits, misses, false alarms and correct negatives – the Contingency table
 - EXERCISE – Building the table
 - Some relevant verification measures: Scores from the table and what they mean
 - EXERCISE – Interpreting the table and scores
 - Verification of the Regional severe weather charts (S. Landman)
 - Verification of other products from the SWFDP
 - Verification of probability forecasts (if time)

What do we mean by forecast *verification*?

- 
- To measure the quality of a forecast by comparison with observations

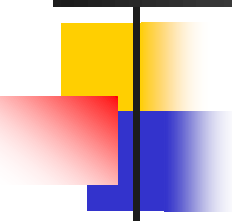
A forecast is like an experiment...

You make a hypothesis about what will happen.

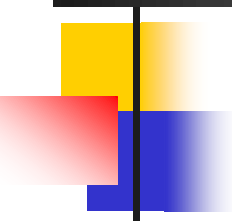
You would not consider an experiment to be complete until you found out what happened.

→ VERIFICATION

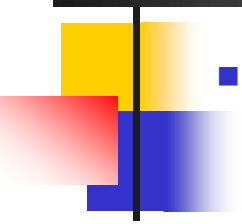
Introduction

- 
- “Verification activity has value only if the information generated leads to a decision about the forecast or system being verified” – A. Murphy
 - Corollary: Verification systems should be set up so they are useful to someone.
 - THAT IS, Verification must have a user
 - Influences the design of the verification
 - “Users” are those who are interested in verification results, and who will take action based on verification results
 - Forecasters, modelers are users too.
 - Importance of verification
 - Increasing tendency to put out graphical forecasts directly from models (quality unknown)
 - Increasing tendency to put out forecasts for populated areas around the world via the web (quality unknown)
 - Models tend not to be verified for countries or regions outside the country(ies) for which they are developed
 - THEREFORE, verification has become more essential.
 - Assume that noone else is going to verify with respect to your stations – Push for it, make it as easy as possible for others.

Why Verify?

- 
- Do you verify your own NMS forecasts?
 - If SO:
 - Whom do you verify for?
 - Why are you doing it?
 - If NOT:
 - Why not?
 - Are you interested in knowing the quality of the guidance products that you use?
 - What do you already know about their quality by looking at them over the months or years?
 - What would you like to know?

Why verify? - Goals of Verification

- 
- Administrative
 - Justify cost of provision of weather services
 - Justify additional or new equipment
 - Monitor the quality of forecasts and track changes
 - Usually means summarizing the verification into few numbers (scoring)
 - Scientific
 - To identify the strengths and weaknesses of a forecast product in sufficient detail that actions can be specified that will lead to improvements in the product, ie to provide information to direct R&D.
 - Demands more detail in verification methodology “diagnostic verification”
 - SWFDP: Both administrative goals and scientific goals

Summary – Products to be verified



- IDEALLY, ALL the products in the SWFDP that are used would be verified objectively

- Requires data – observations

- GTS data, non GTS data

- Derived products such as the Hydroestimator or TRMM

- QC important – generally should not involve models

- WHO? Generally easier to do it at the forecast issuing location to avoid transfer of large data volumes

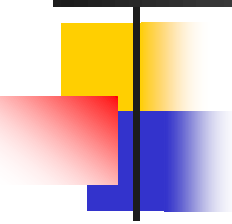
- BUT, hasn't really worked out that way.

- Compromise: Global centers prepare datasets and GTS observations for RSMCs and NMSs to verify – how?

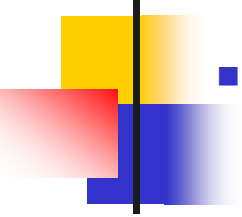
Rest of the presentation – exercise sessions is about HOW to verify

Goal: To encourage verification activity and to make it is easy, painless, and interesting as possible

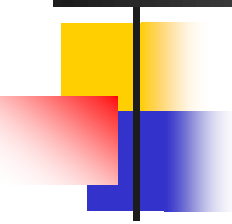
What is truth? Some comments on observations

- 
- Station observations
 - Valid at points – a sample of local weather
 - Generally accurate for the points they represent
 - BUT must be quality controlled
 - For verification, QC should be independent of models
 - Satellite-derived precipitation estimates such as HE
 - Space and time coverage good if from geostationary
 - NOT representative of points – some averaging e.g. HE is about 12km. Limited by satellite footprint
 - For verification – use of model in estimation is a problem – incestuous if model is used in forecasting process
 - Most users of forecasts live at points
 - Station-based verification fundamental, and best
 - Averaging/incestuous effects important – will lead to “optimistic” verification, but not necessarily realistic

How to verify: Verification Procedure

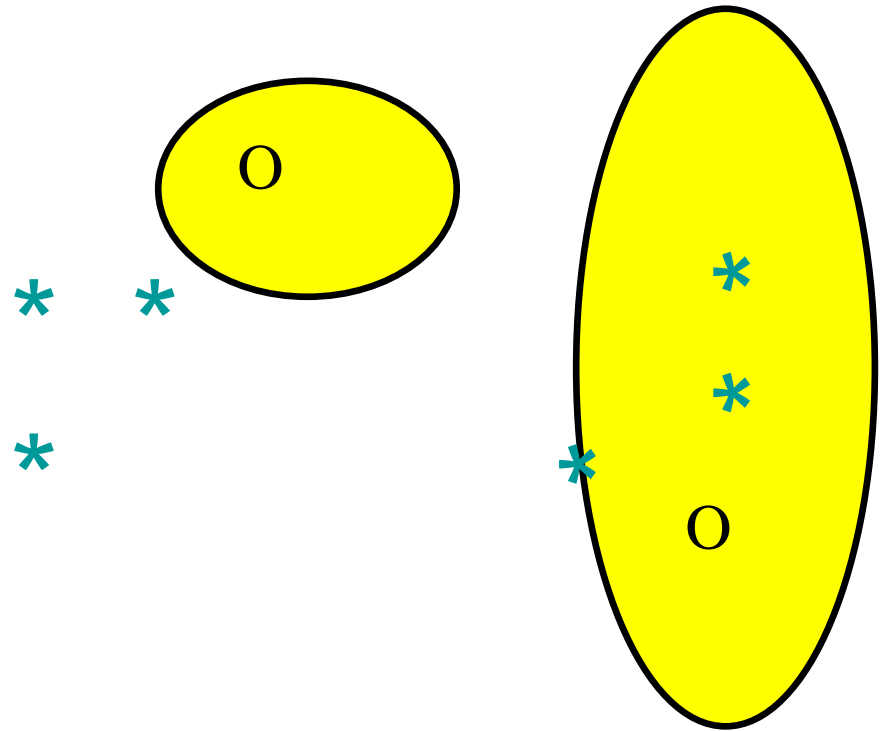
- 
- Build a matched dataset of forecasts and observations
 - From events table – your forecasts
 - From datasets supplied from global and regional centers
 - SWFDP: Predicted variables are **categorical**: Extreme events, where extreme is defined by thresholds of precipitation and wind. Some **probabilistic** forecasts are available too
 - Build contingency tables from matched data
 - Scores
 - Interpretation and decisions about model being verified.

What is the Event?

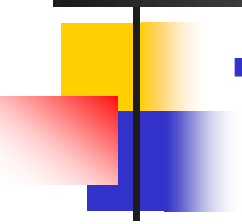
- 
- For categorical and probabilistic forecasts, one must be clear about the “event” being forecast
 - Location or area for which forecast is valid
 - Time range over which it is valid
 - Definition of category
 - Example?
 - And now, what is defined as a correct forecast?
 - The event is forecast, and is observed – anywhere in the area? Over some percentage of the area?
 - Scaling considerations
 - Discussion:

Verification of NMS warnings: What is the Event?

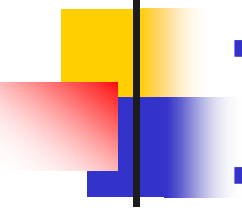
- For categorical and probabilistic forecasts, one must be clear about the "event" being forecast
 - Location or area for which forecast is valid
 - Time range over which it is valid
 - Definition of category
- And now, what is defined as a correct forecast?
 - The event is forecast, and is observed – anywhere in the area? Over some percentage of the area?
 - Scaling considerations



Events for the SWFDP

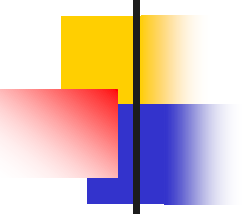
- 
- Best if “events” are defined for similar time period and similar-sized areas
 - One day 24h
 - Fixed areas; should correspond to forecast areas and have at least one reporting stn.
 - The smaller the areas, the more useful the forecast, potentially, BUT...
 - Predictability lower for smaller areas
 - More likely to get missed event/false alarm pairs
 - Data density a problem
 - Best to avoid verification where there is no data.
 - Non-occurrence – no observation problem

Preparation of the table

- 
- Start with matched forecasts and observations
 - Forecast event is precipitation >50 mm / 24 h Next day
 - Count up the number of each of hits, false alarms, misses and correct negatives over the whole sample
 - Enter them into the corresponding 4 boxes of the table.

Day	Fcst to occur?	Observed ?
1	Yes	Yes
2	No	Yes
3	No	No
4	Yes	No
5	No	No
6	Yes	Yes
7	No	No
8	No	Yes
9	No	No

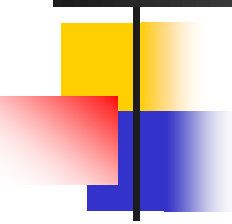
The contingency Table



Observations

		Observations		
		Yes	No	
Forecasts	Yes	HITS	FALSE ALARMS	Total Events Forecast
	No	MISSED EVENTS	CORRECT NEGATIVES	Total non-events Forecast
	Total Events Observed	Total Non-Events Observed	Sample size	

Mozambique exercise

- 
- You have a spreadsheet called Mozambique exercise – for manual CT generation – open this
 - The data for this comes from the events table. Events forecast or observed or both are shown, then for all days with no forecast or observed events, one “event” has been added for each day, total cases = 116
 - There are two forecasts represented: The Mozambique forecast and the RSMC forecast for all events.
 - Your job is to determine the number of hits, misses and false alarms and complete the table

Contingency tables

		Observations	
Forecasts	HITS a	FALSE ALARMS b	Total Events Forecast a+b
	MISSED EVENTS c	CORRECT NEGATIVES d	Total non-events Forecast c+d
	Total Events Observed a+c	Total Non-Events Observed b+d	Sample size T=a+b+c+d

$$PoD = \frac{a}{a + c}$$

range: 0 to 1
best score = 1

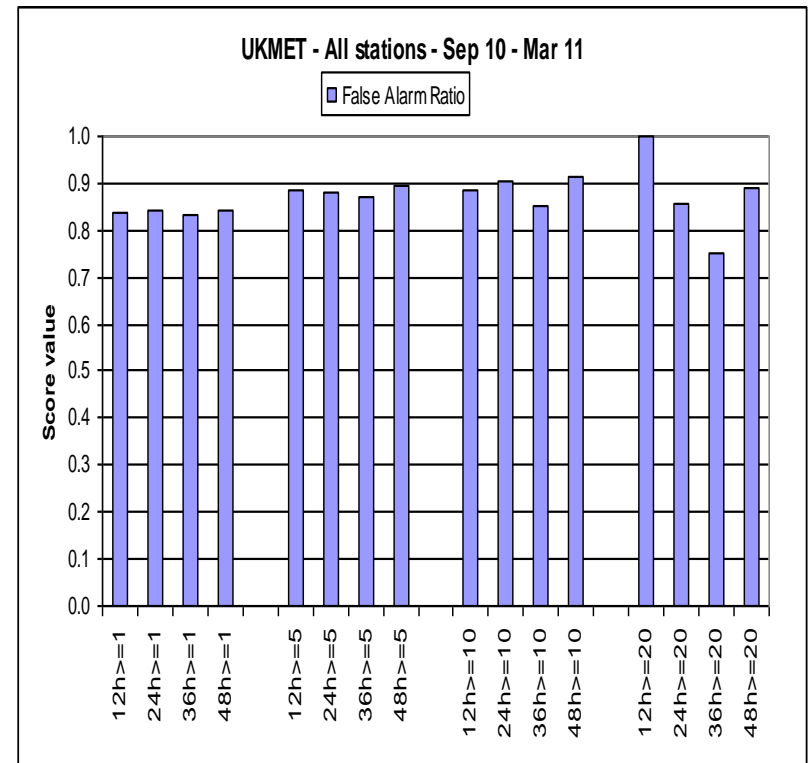
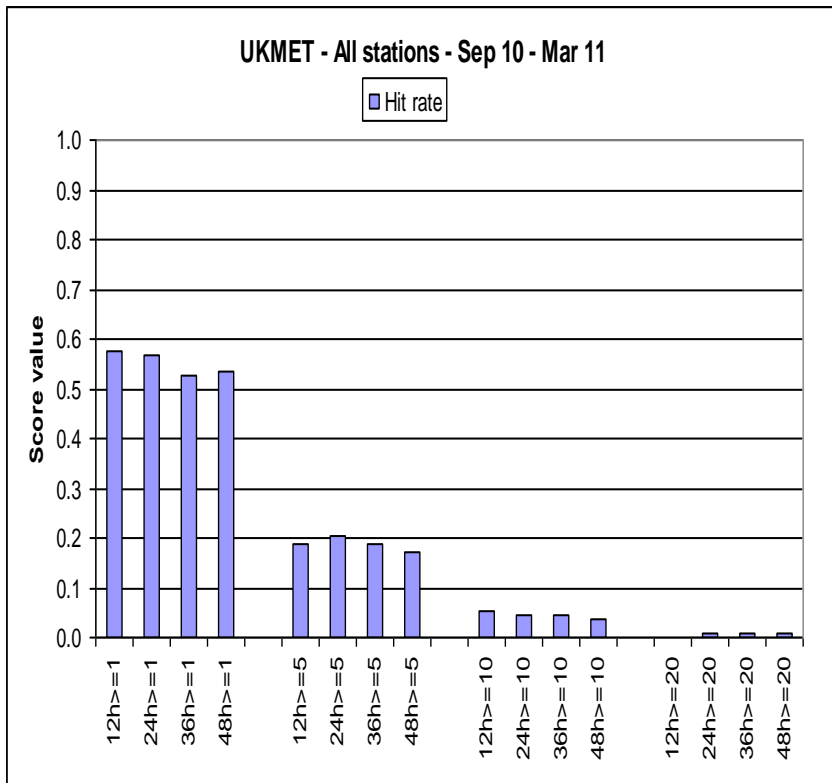
$$FAR = \frac{b}{(a + b)}$$

range: 0 to 1
best score = 0

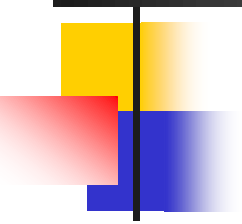
Characteristics:

- PoD= “Prefigurance” or “probability of detection”, “hit rate”
 - Sensitive only to missed events, not false alarms
 - Can always be increased by overforecasting rare events
- FAR= “False alarm ratio”
 - Sensitive only to false alarms, not missed events
 - Can always be improved by underforecasting rare events

Hit rate, false alarm ratio



What's wrong with PC - % correct? The Finley Affair (1884)



Observed

		tornado	no tornado	Total
Forecast	tornado	28	72	100
	no tornado	23	2680	2703
Total		51	2752	2803

% correct = $(28+2680)/2803 = 96.6\%$; No tornado forecast: $(2752)/2803 = 98.2\%$!

Contingency tables

		Observations	
Forecasts	HITS a	FALSE ALARMS b	Total Events Forecast a+b
	MISSED EVENTS c	CORRECT NEGATIVES d	Total non-events Forecast c+d
	Total Events Observed a+c	Total Non-Events Observed b+d	Sample size T=a+b+c+d

$$PAG = \frac{a}{a+b}$$

range: 0 to 1
best score = 1

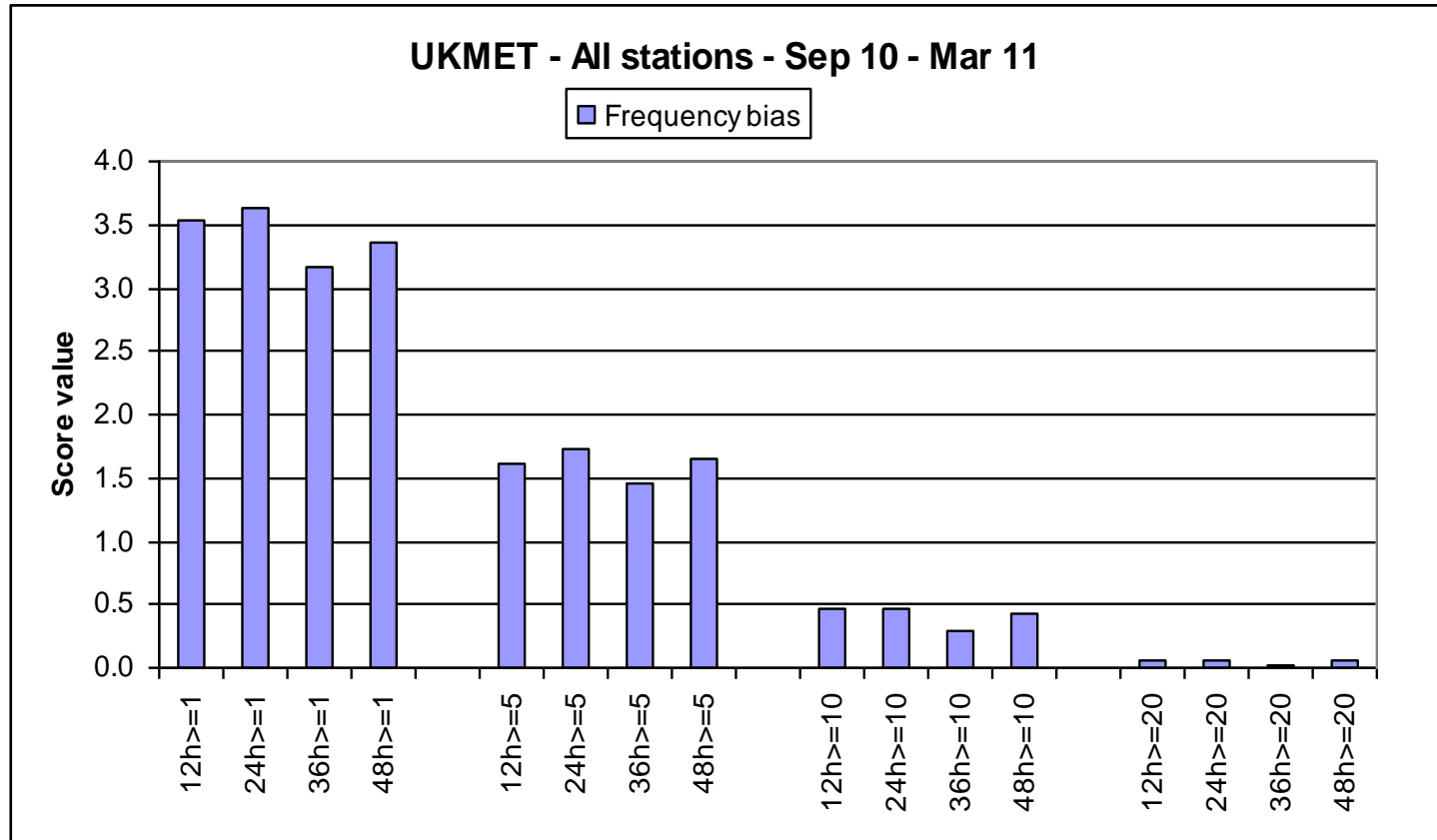
$$Bias_{frequency} = \frac{a+b}{a+c}$$

best score = 1

Characteristics:

- PAG= “Post agreement”
 - PAG= (1-FAR), and has the same characteristics
- Bias: This is frequency bias, indicates whether the forecast distribution is similar to the observed distribution of the categories (Reliability)

Frequency Bias



Contingency tables

		Observations	
Forecasts	HITS	FALSE ALARMS	Total Events Forecast
	<i>a</i>	<i>b</i>	<i>a+b</i>
	MISSED EVENTS	CORRECT NEGATIVES	Total non-events Forecast
<i>c</i>	<i>d</i>	<i>c+d</i>	
Total Events Observed	Total Non-Events Observed	Sample size	
<i>a+c</i>	<i>b+d</i>	<i>T=a+b+c+d</i>	

$$CSI = \frac{a}{a+b+c}; \frac{d}{b+c+d}$$

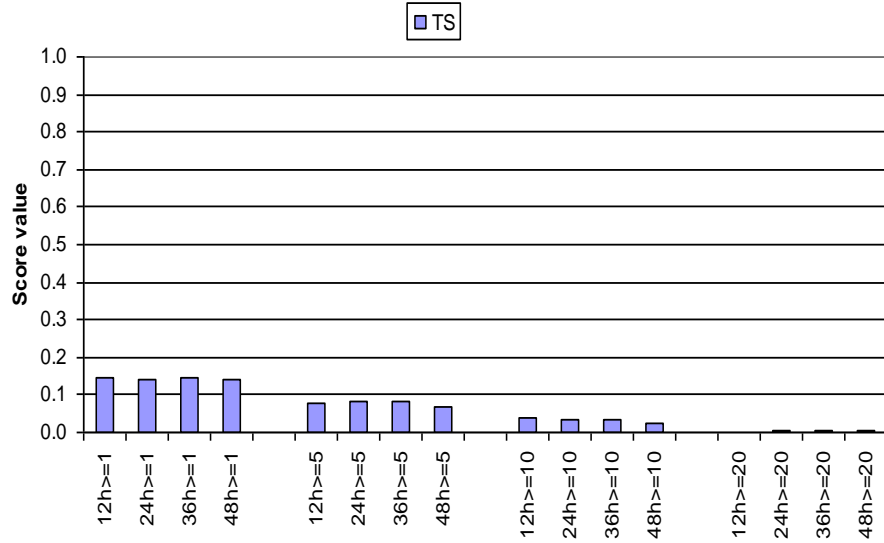
range: 0 to 1
best score = 1

Characteristics:

- Better known as the Threat Score
- Sensitive to both false alarms and missed events; a more balanced measure than either PoD or FAR

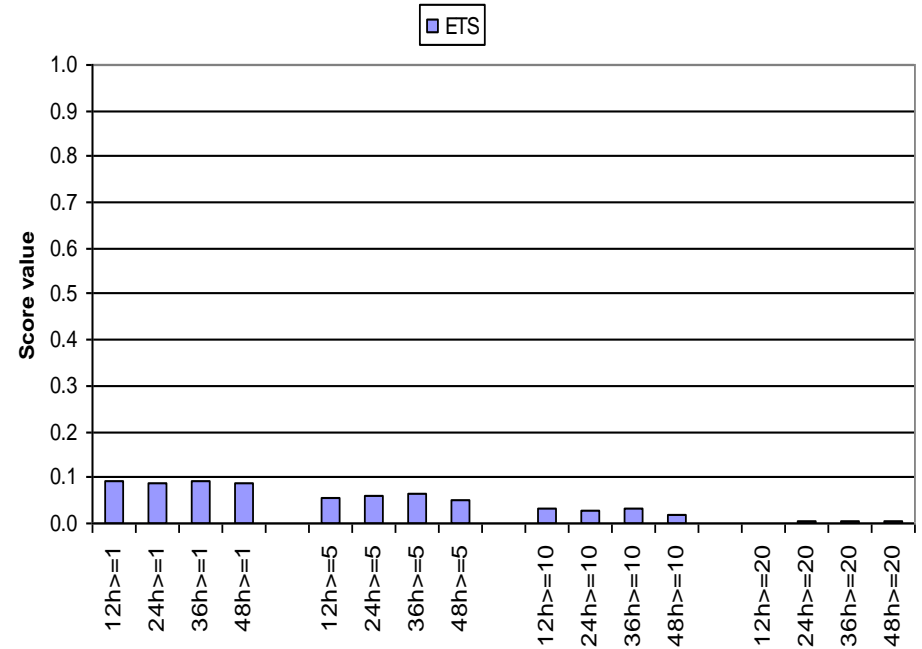
Threat Score (CSI) and Equitable Threat Score

UKMET - All stations - Sep 10 - Mar 11



ETS is smaller than TS

UKMET - All stations - Sep 10 - Mar 11



Contingency tables

		Observations	
Forecasts	HITS a	FALSE ALARMS b	Total Events Forecast a+b
	MISSED EVENTS c	CORRECT NEGATIVES d	Total non-events Forecast c+d
	Total Events Observed a+c	Total Non-Events Observed b+d	Sample size T=a+b+c+d

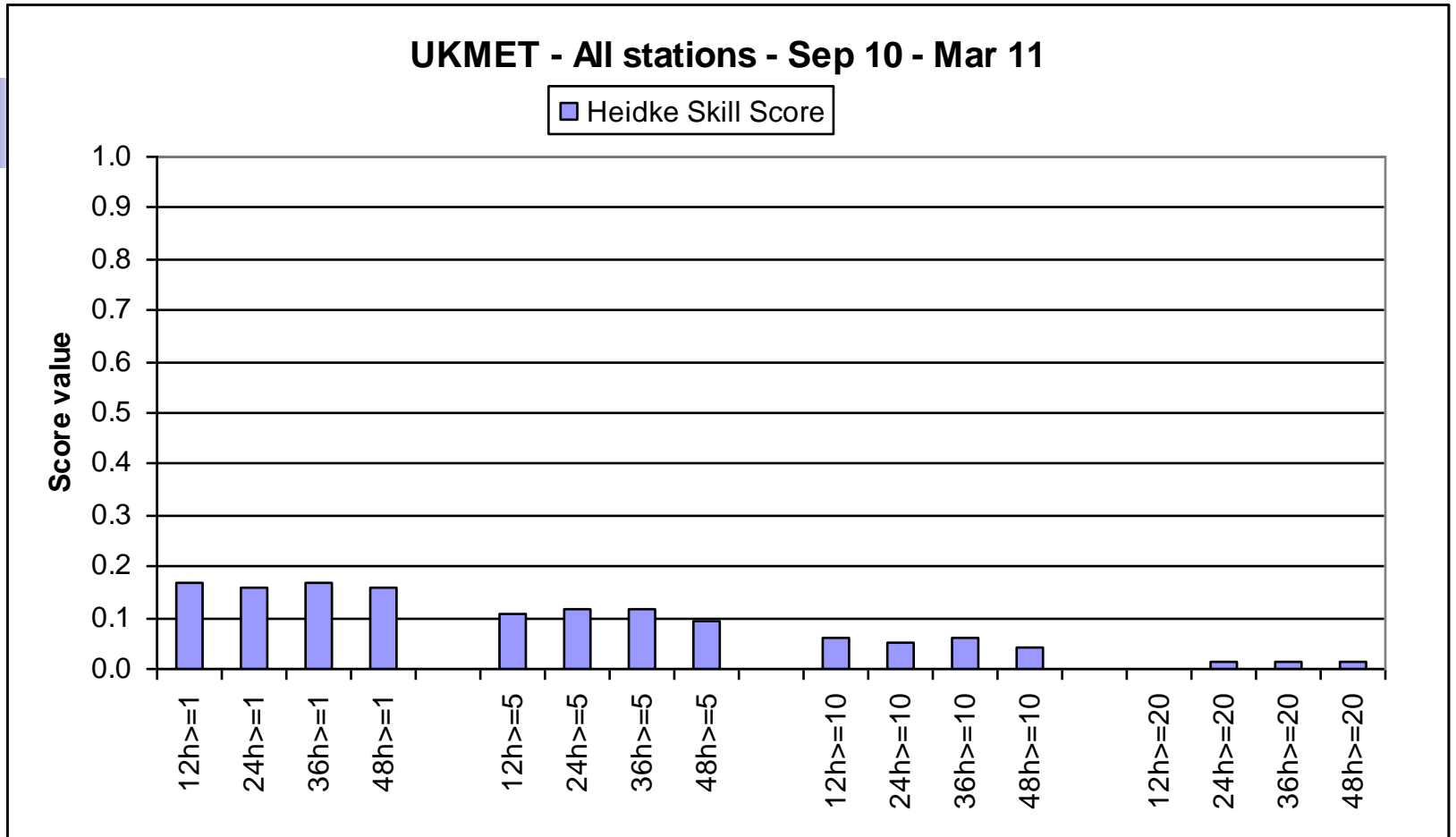
$$HSS = \frac{(a+d) - \frac{(a+b)(a+c) + (c+d)(b+d)}{T}}{T - \frac{(a+b)(a+c) + (c+d)(b+d)}{T}}$$

**range: negative value to 1
best score = 1**

Characteristics:

- A skill score against chance (as shown)
- Easy to show positive values
- Better to use climatology or persistence
 - needs another table

Heidke Skill Score



Contingency tables

	Observations		
Forecasts	HITS <i>a</i>	FALSE ALARMS <i>b</i>	Total Events Forecast <i>a+b</i>
	MISSED EVENTS <i>c</i>	CORRECT NEGATIVES <i>d</i>	Total non-event Forecast <i>c+d</i>
	Total Events Observed <i>a+c</i>	Total Non-Events Observed <i>b+d</i>	Sample size <i>T=a+b+c+d</i>

$$HR = \frac{a}{a + c}$$

range: 0 to 1
best score = 1

$$FA = \frac{b}{(b + d)}$$

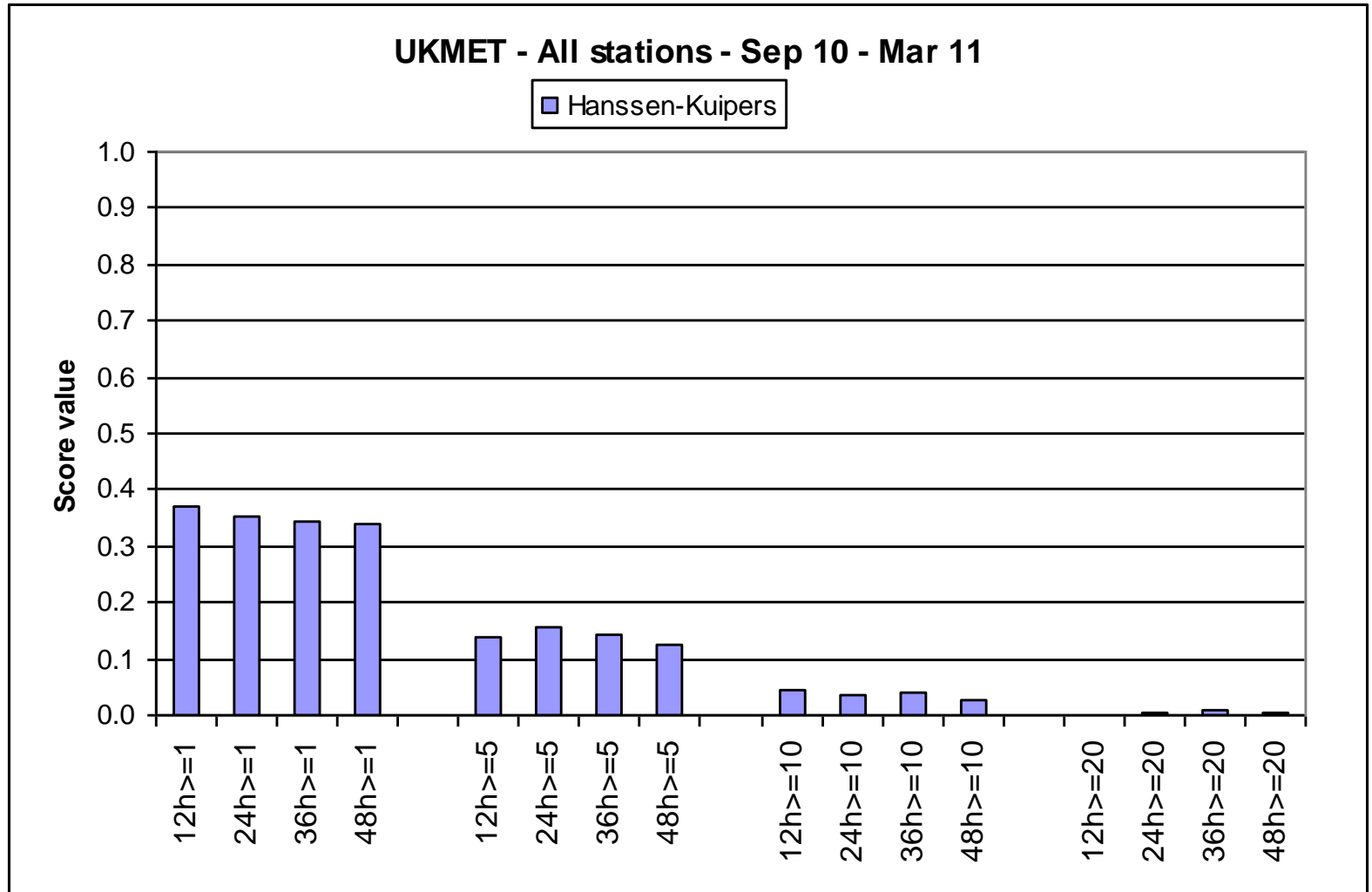
best score = 0

$$KSS = HR - FA$$

Characteristics:

- Hit Rate (HR) is the same as the PoD and has the same characteristics
- False alarm RATE. This is different from the false alarm ratio.
- These two are used together in the Hanssen-Kuipers score, and in the ROC, and are best used in comparison.

Hanssen-Kuipers (Pierce) Skill score



Verification of extreme, high-impact weather

- **EDS – EDI – SEDS - SEDI** ⇔ **Novelty categorical measures!**

Standard scores tend to zero for rare events

Event forecast	Event observed		Marginal total
	Yes	No	
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d = n

$H = a / (a+c)$, hit rate

$F = b / (b+d)$, false alarm rate

$p = (a+c) / n$, base rate

$q = (a+b) / n$, relative frequency of forecasted events

$$\boxed{\text{EDS}} = \frac{\log p - \log H}{\log p + \log H}$$

$$\boxed{\text{SEDS}} = \frac{\log q - \log H}{\log p + \log H}$$

Ferro & Stephenson, 2011: Improved verification measures for deterministic forecasts of rare, binary events. *Wea. and Forecasting*

Base rate independence ⇔ Functions of H and F

$$\boxed{\text{EDI}} = \frac{\log F - \log H}{\log F + \log H}$$

Extremal Dependency Index - EDI

Symmetric Extremal Dependency Index - SEDI

$$\boxed{\text{SEDI}} = \frac{\log F - \log H - \log(1 - F) + \log(1 - H)}{\log F + \log H + \log(1 - F) + \log(1 - H)}$$

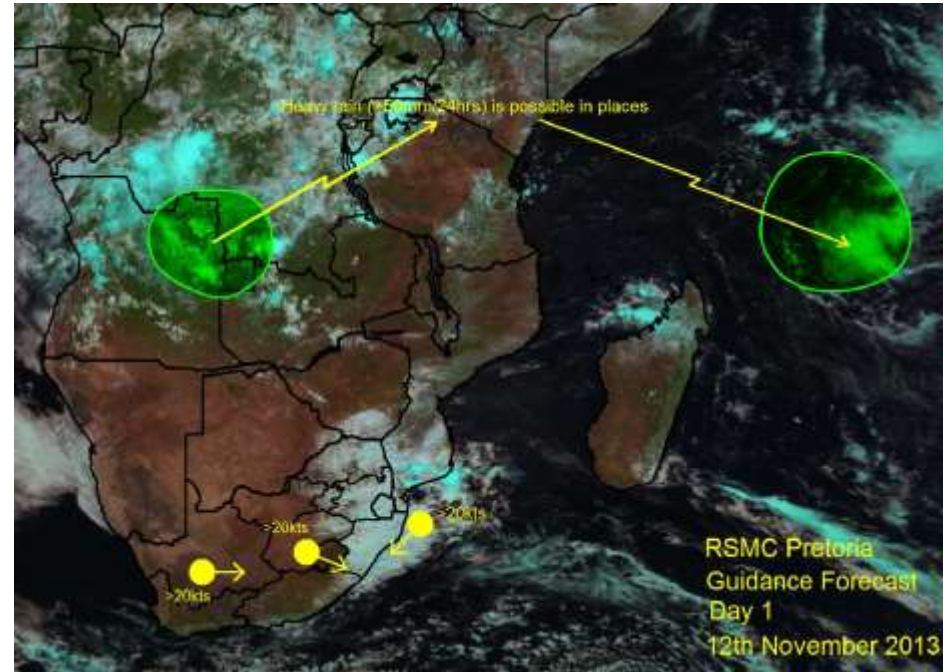
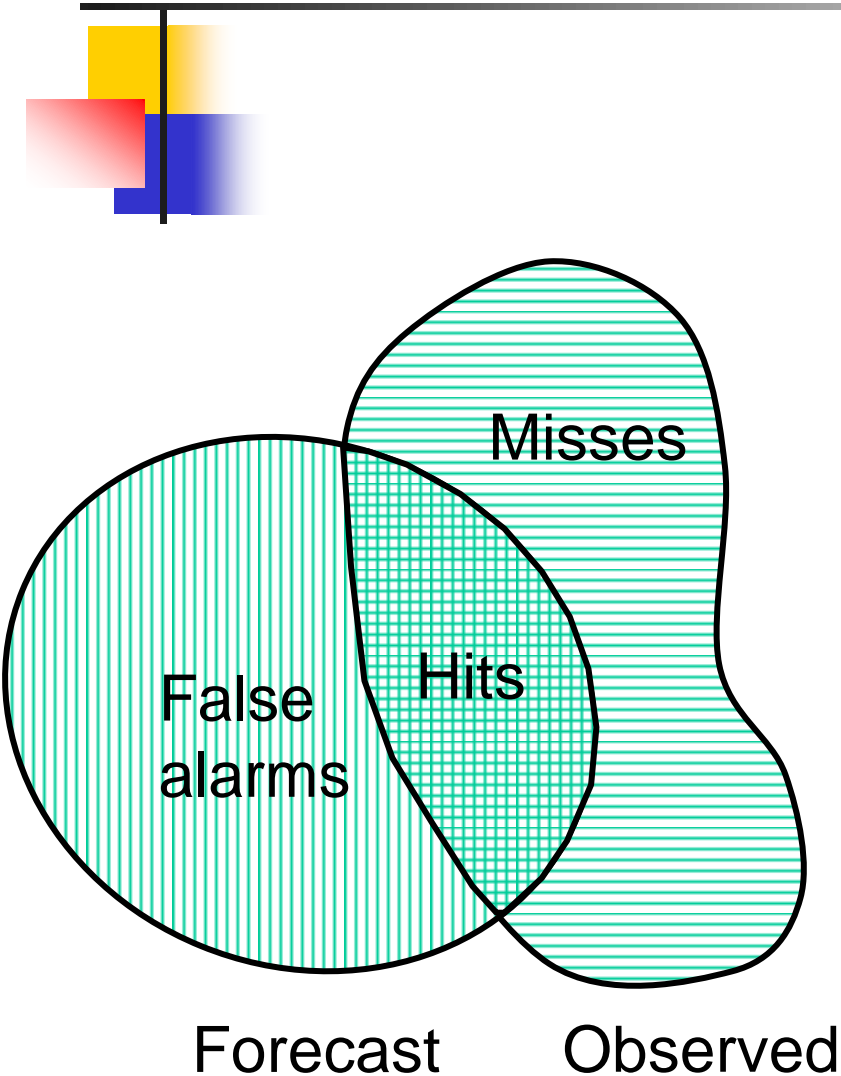


Mozambique Interpretation Exercise

- Load “Mozambique exercise – with tables and scores”
- Two forecasts, from NMS Mozambique and from RSMC Pretoria
- Goal: to decide which is better and why
- We’ll do this together



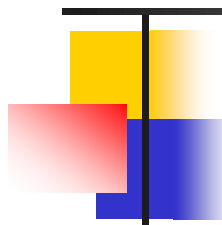
Spatial verification of RMSC products



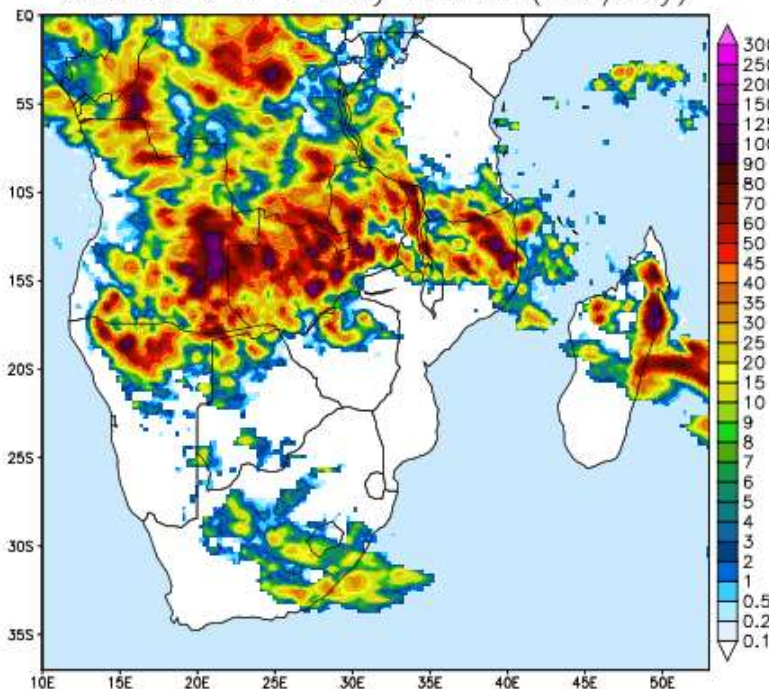
Spatial contingency table:

- Can accomplish IF one has quasi-continuous spatial observation data
- Stephanie's method

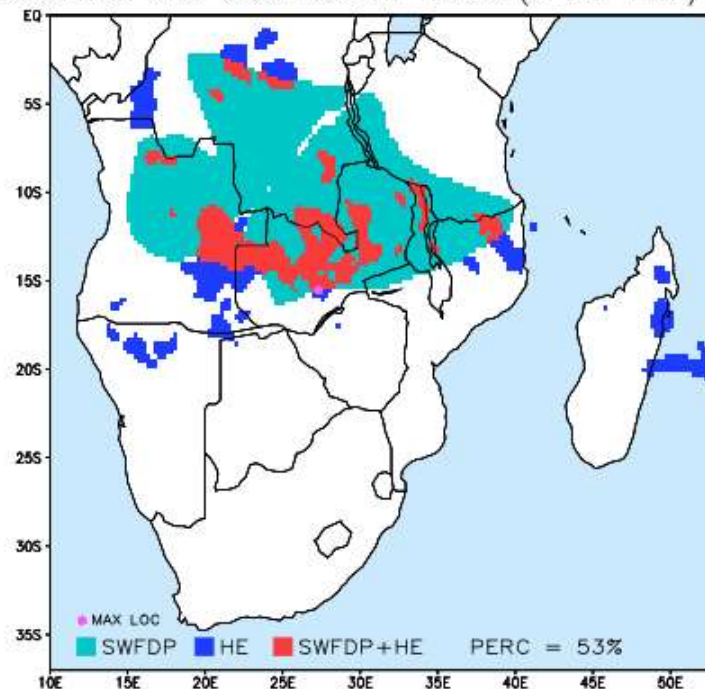
Verification of regional forecast map using HE



20121219 H-E daily rainfall (mm/day)



Guidance and Observation fields (> 50 mm/day)



Verification statistics for 20121219 : Grid Size = 0.25° : Units = mm/day : n = 25777

	Guidance	H-E
Number of gridpoints ≥ 50 mm	3294	1243
Average Rain over domain	~	19.7012
≥ 50 mm Rain Area (km ² *10 ⁴)	2.05875	0.776875
Maximum Rainfall Observed (mm)	~	151.124

	Categorical Forecasts
Frequency Bias	2.65004
Probability of Detection	0.526146
False Alarm Ratio	0.801457
Hansen & Kuipers Score	0.418541
Equitable threat score	0.132959
Spatial Correlation	0.264835

		OBSERVATION	
		≥ 50	< 50
GUIDANCE	≥ 50	654	2640
	< 50	589	21894

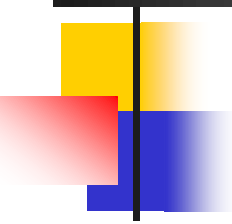
Extreme Events Verification	
Extreme Dependency Score	0.650434
Symmetric Extreme Dependency Score	0.385181
Extremal Dependency Index	0.552717
Symmetric Extremal Dependency Index	0.59486

(**Ferro and Stephenson, 2011**)

<http://rsmc.weathersa.co.za/RSMC/index.php>
Format based on IPWG verification output



Capital Cities Verification

- 
- Forecast is nearest gridpoint to Capital cities of all countries
 - Observation is HE estimated precipitation at that point (top row) and Max HE estimated precipitation within 50 km (bottom).
 - About 4 years of data – allows for enough severe precipitation cases at a single location (usually)
 - Data prepared by Stephanie, loaded into Excel via “CT calculator program” which is set up to calculate all the contingency table and all the scores from one fcst-obs matched binary dataset.
 - Results are loaded into the “summary” page for easy comparison
 - Your task:
 - Load the Excel file for your group
 - Evaluate the scores for each of your capital cities, decide which is best and why. Comment on over- under-forecast tendency at each location.
 - Consider: Hit rate, false alarm ratio, bias, ETS, SEDS, EDI
 - Nominate a presenter from each group to discuss

Example - Madagascar

211 Cases

Separate tables assuming low, medium, high risk as thresholds

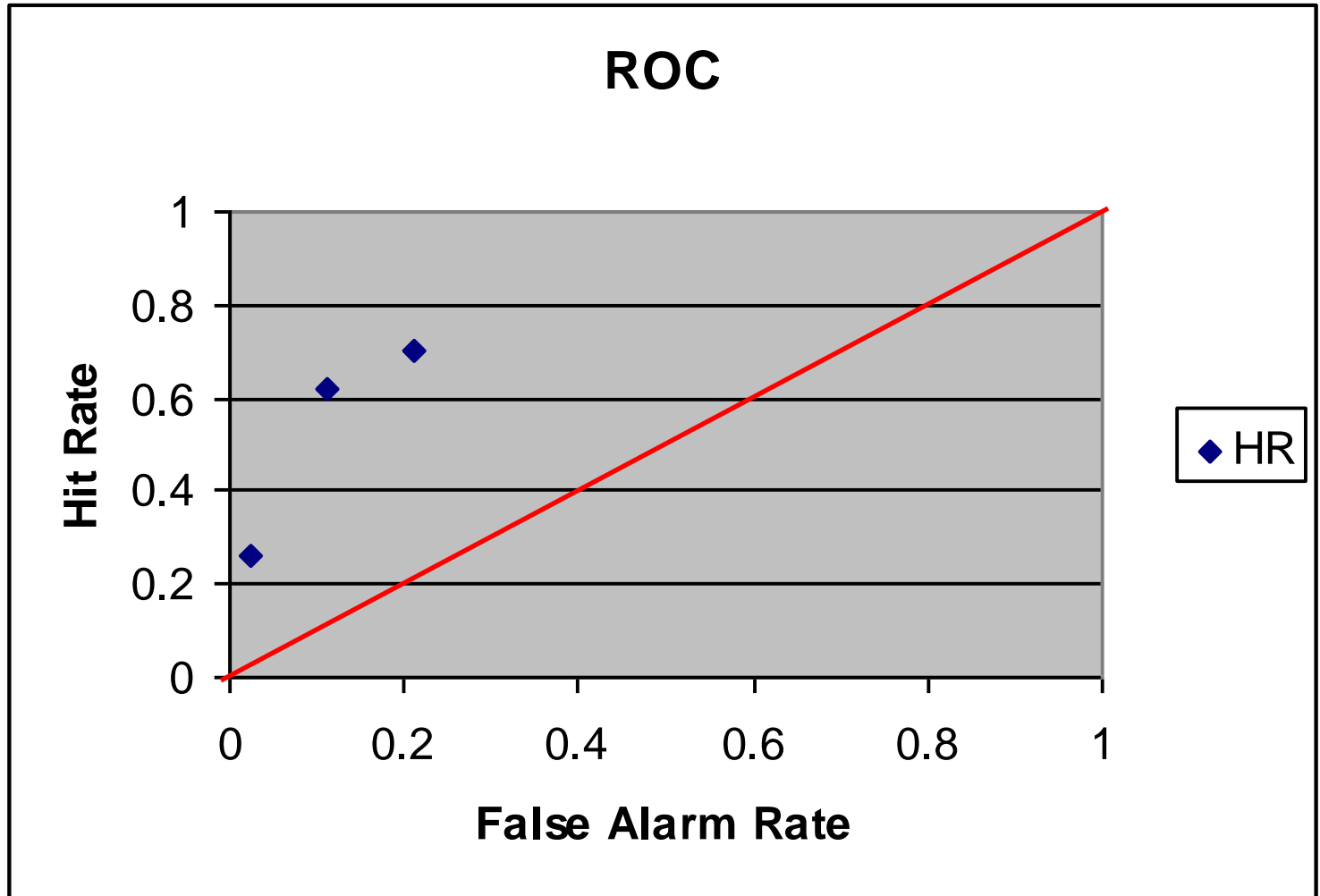
Can plot the hit rate vs the false alarm RATE = FA/total obs no

Low	Obs yes	Obs no	Totals
Fcst yes	35	34	69
Fcst no	15	127	142
Totals	50	161	211

Med	Obs yes	Obs no	Totals
Fcst yes	31	18	49
Fcst no	19	143	162
Totals	50	161	211

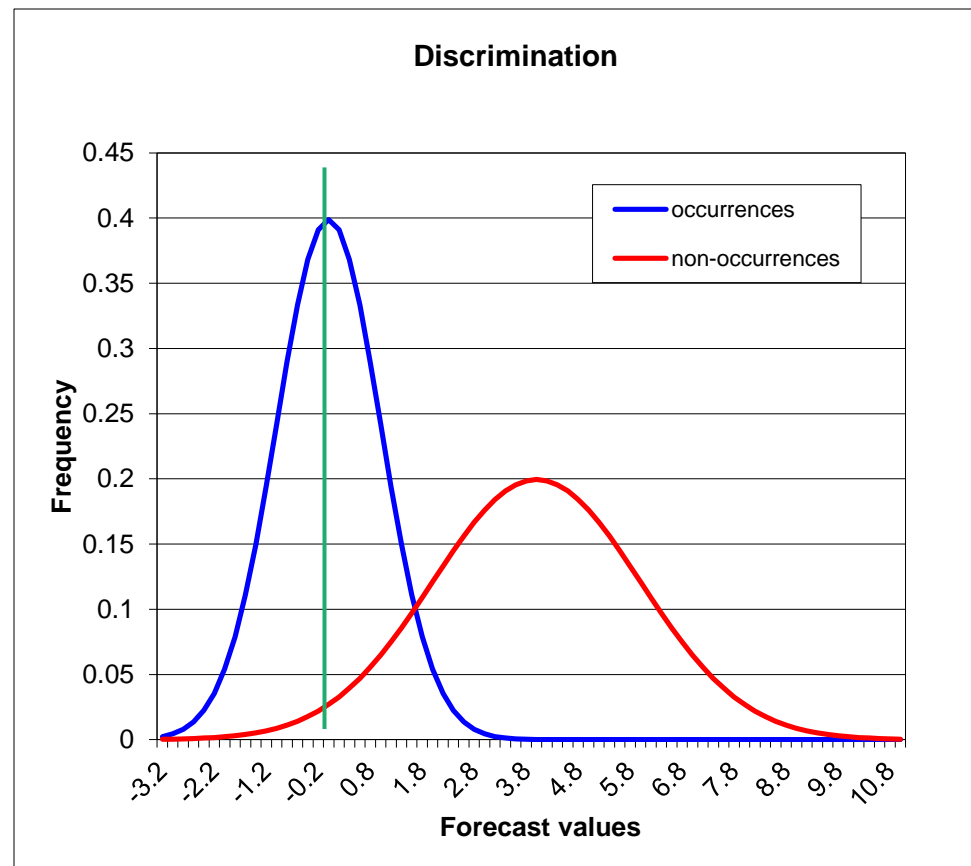
High	Obs yes	Obs no	Totals
Fcst yes	13	4	17
Fcst no	37	157	194
Totals	50	161	211

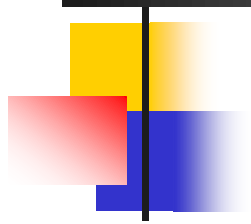
Example (contd)



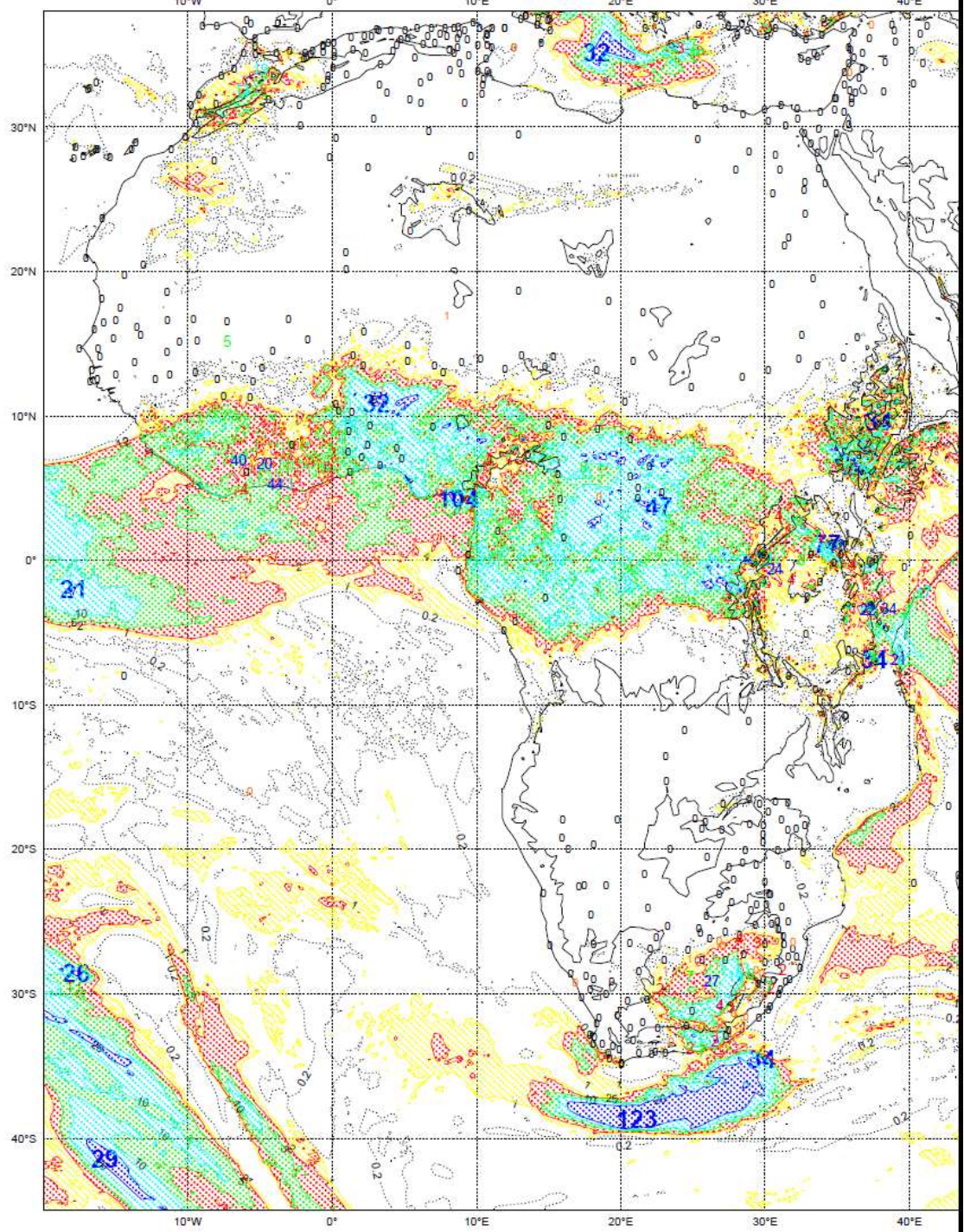
Discrimination

- User-perspective:
 - Event: temperature remains above freezing
 - Does the forecast tend to give higher temperature values when the event occurs than when it doesn't?

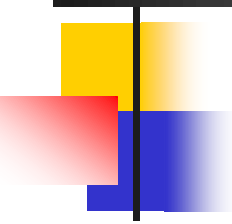




ECMWF Diagnostic chart:
-Daily precipitation values
plotted vs forecast amts.

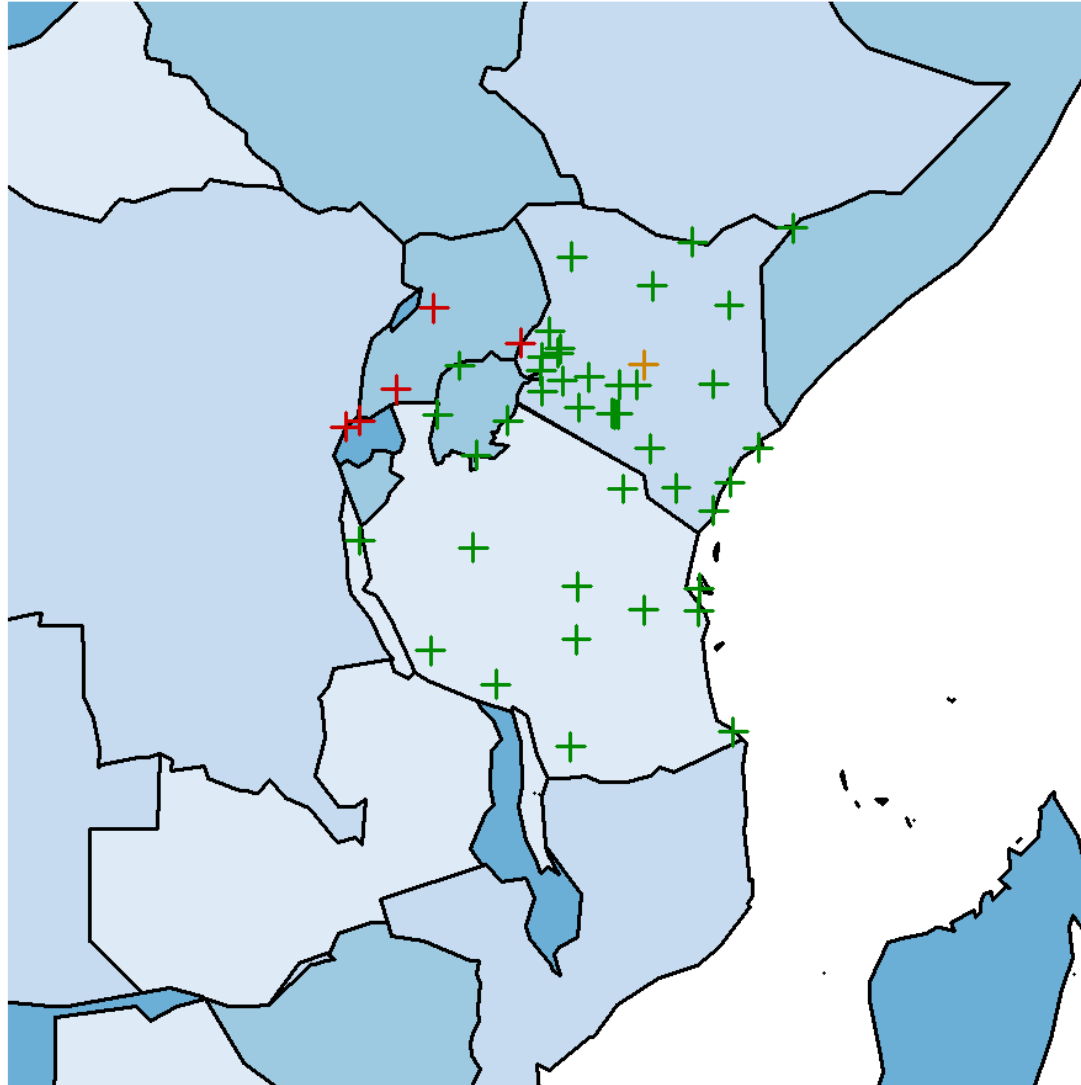
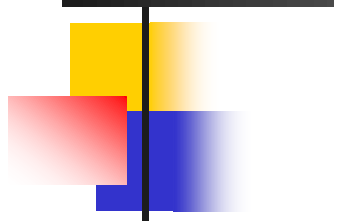


Verification in E. Africa project

- 
- NCEP and ECMWF comparison
 - “The Africas Cup” ECMWF vs NCEP eps verification
 - Verification study of 4km UM over L. Victoria

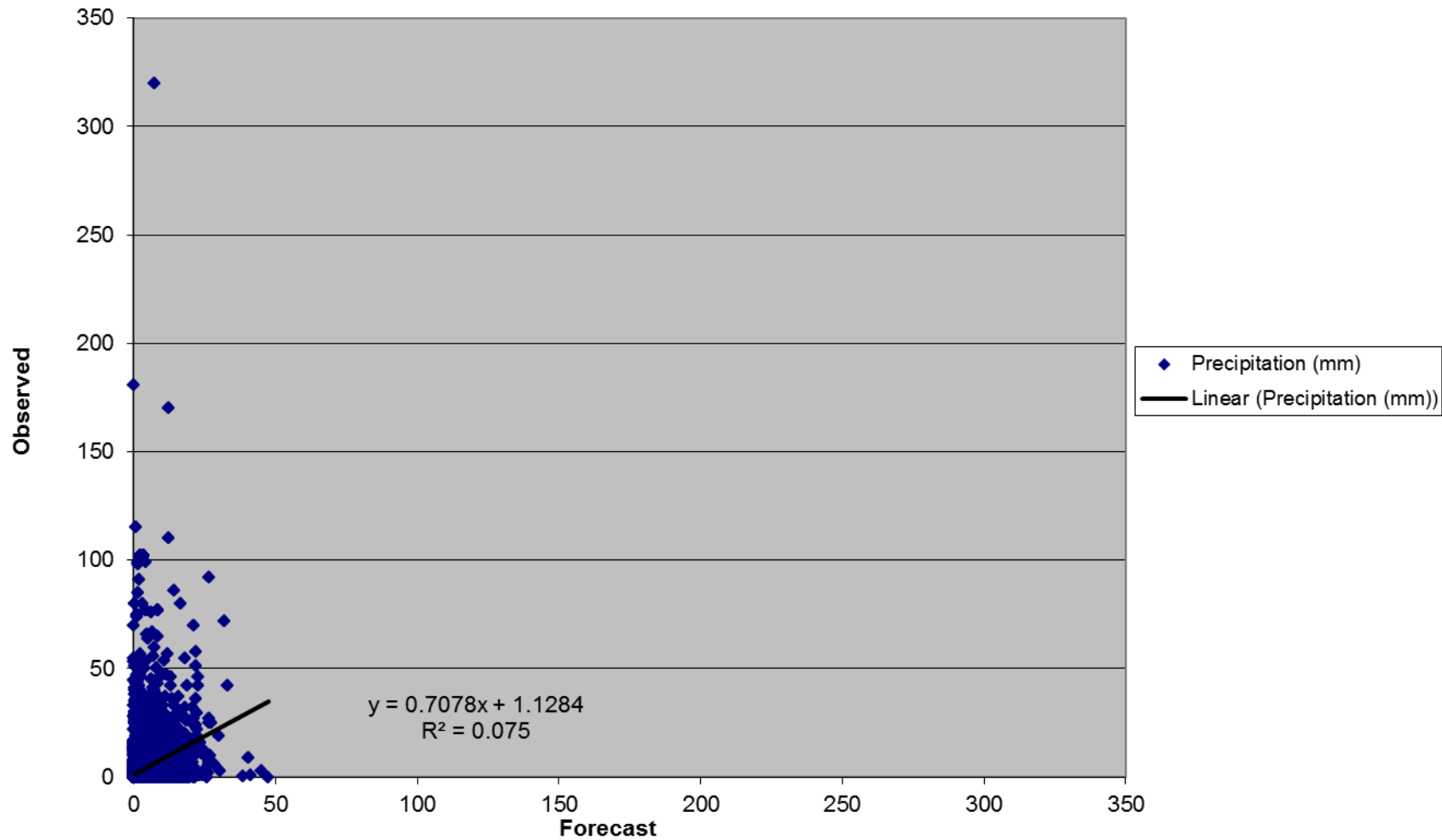
Global model verification Sept 2010 to May, 2011

Stations available

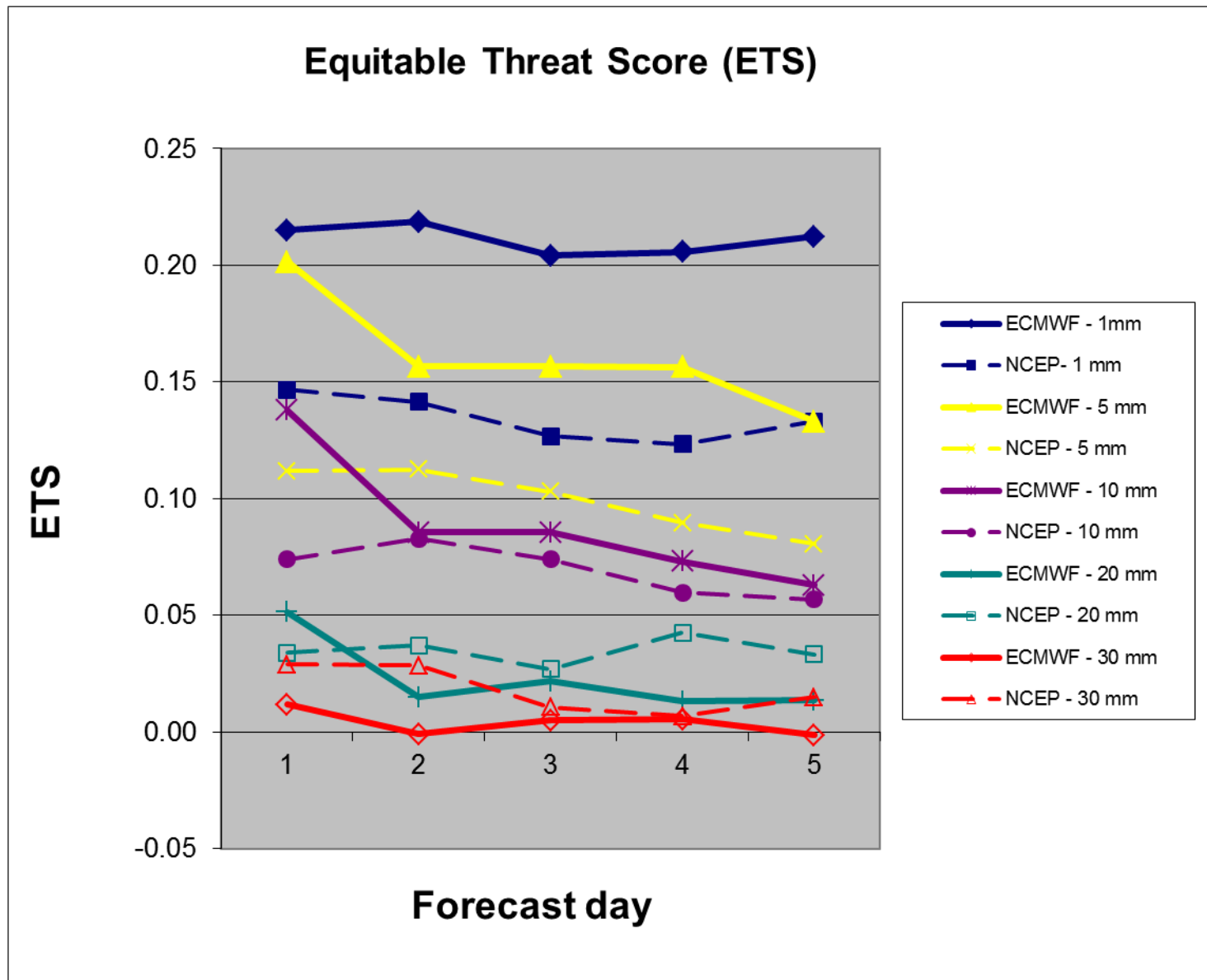


Scatterplot - ECMWF

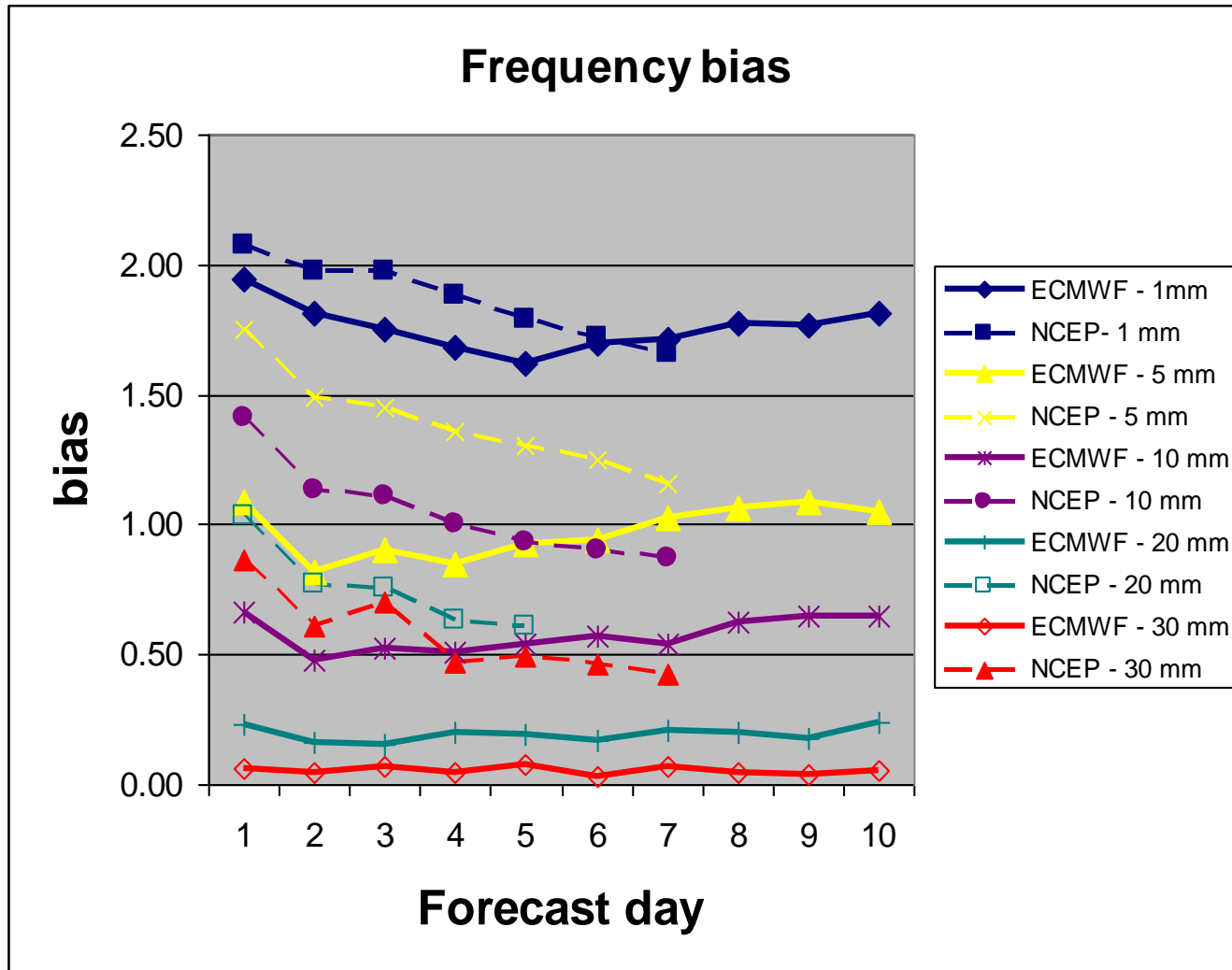
Precipitation - ECMWF - 24h



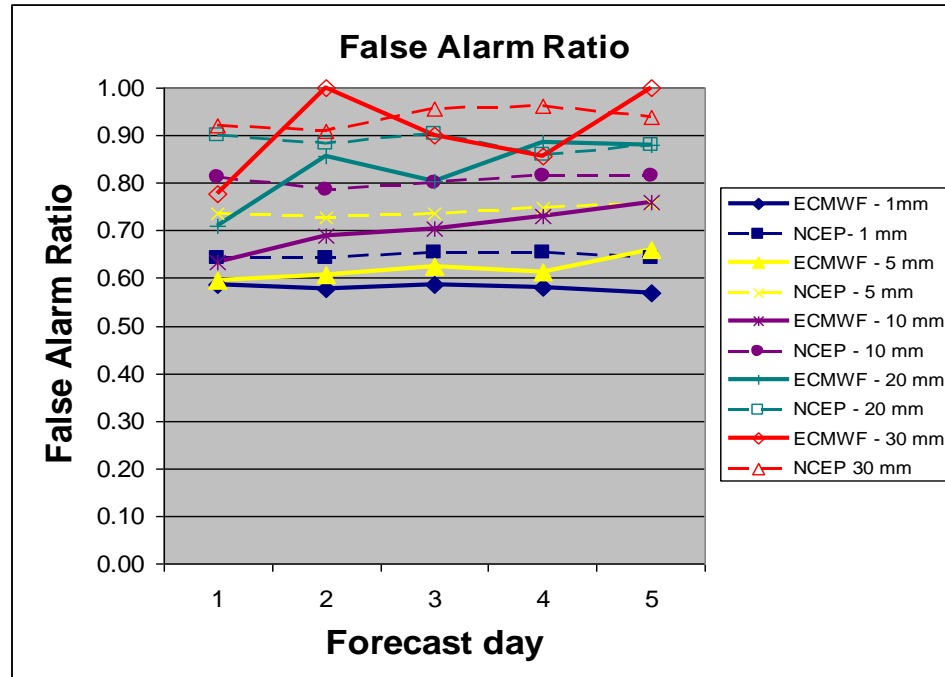
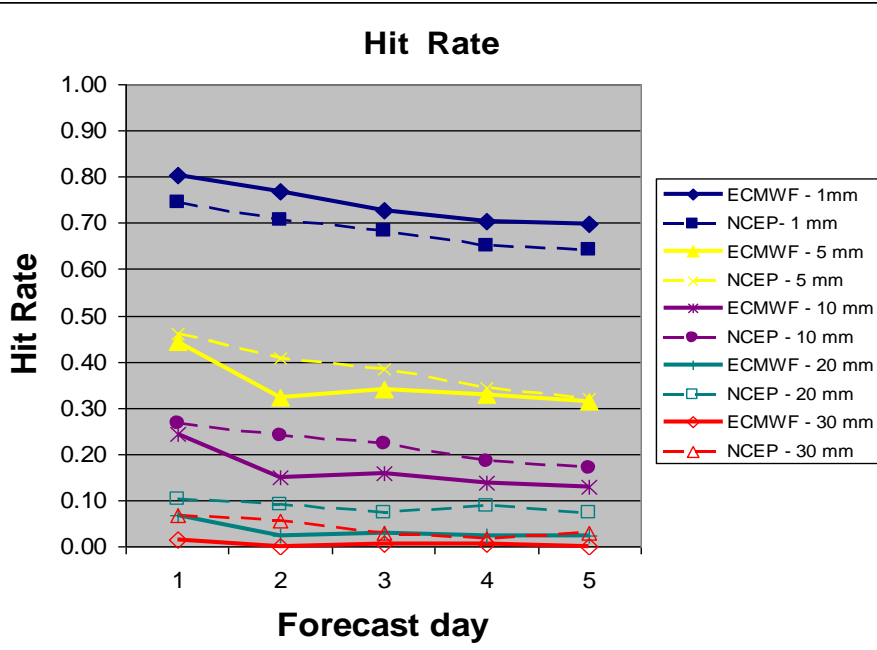
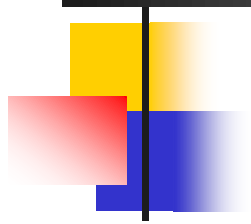
ECMWF vs NCEP 24h precipitation



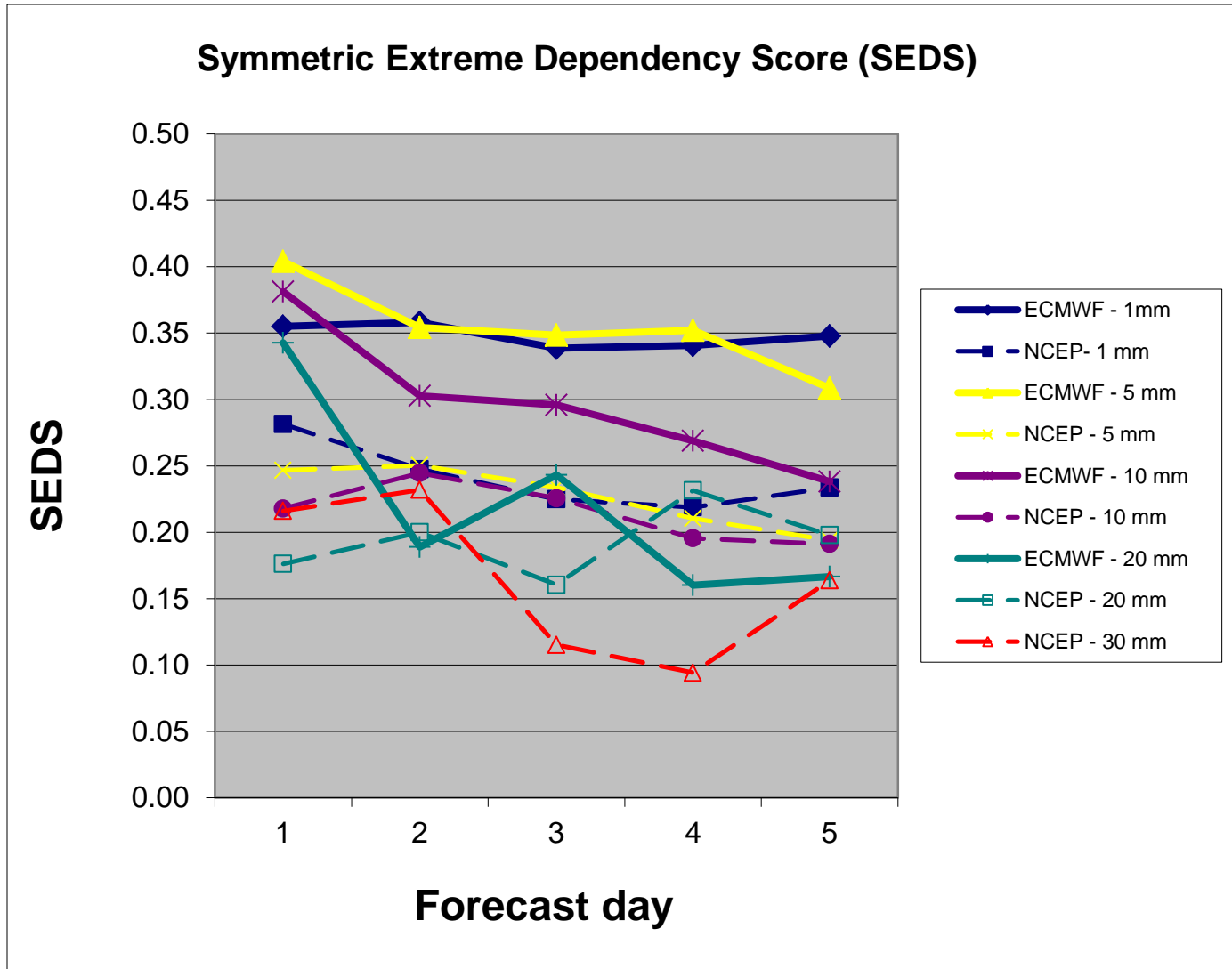
ECMWF vs. NCEP 24 h precipitation (2)



ECMWF vs NCEP 24h precipitation (3)



ECMWF vs NCEP 24 h Precipitation (4)

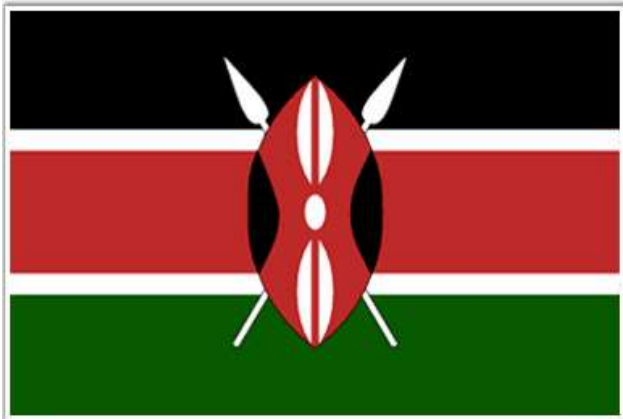




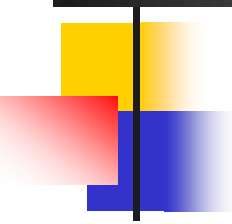
ECMWF Vs. MOGREPS

Africa Cup

Trevor Carey-Smith, Yinglin Li, Evgeny Atlaskin, Matthew Trueman,
Anatoly Muravyev

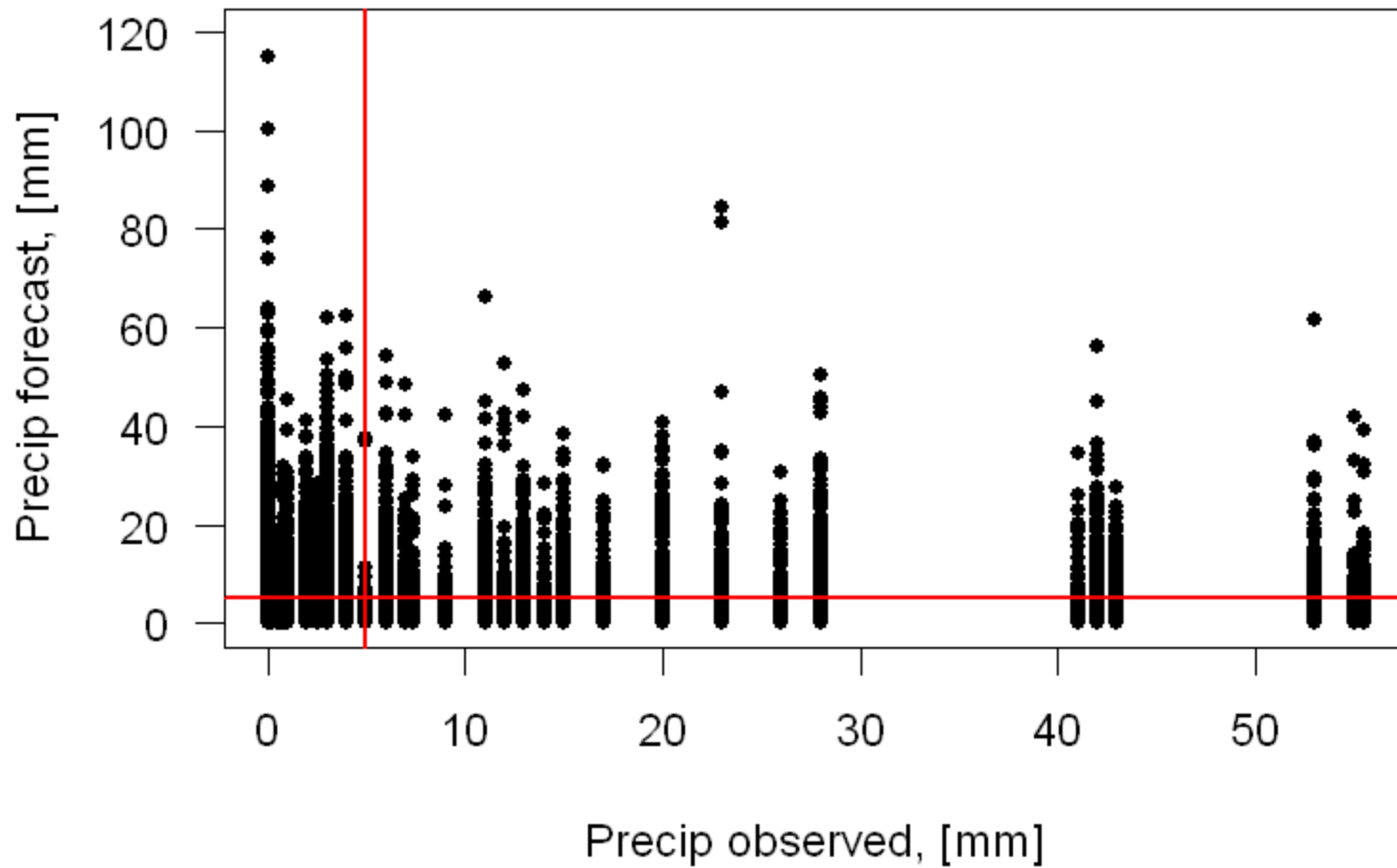


Rules of the Match

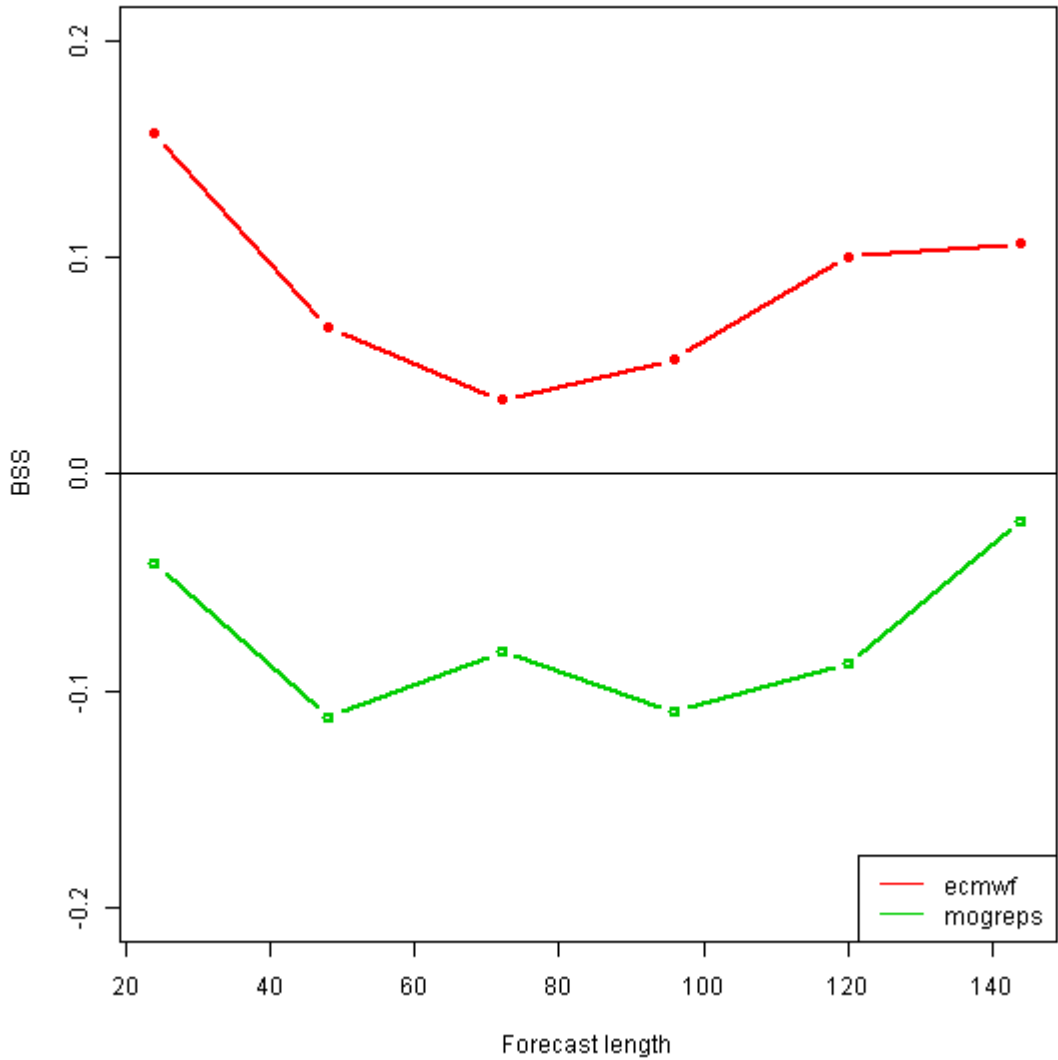
- 
- Two global ensemble prediction systems
 - ECMWF (A-squad)
 - MOGREPS (B-squad)
 - One rainy season (8.5 months)
 - 24hr precipitation accumulations
 - MOGREPS data goes to T+144

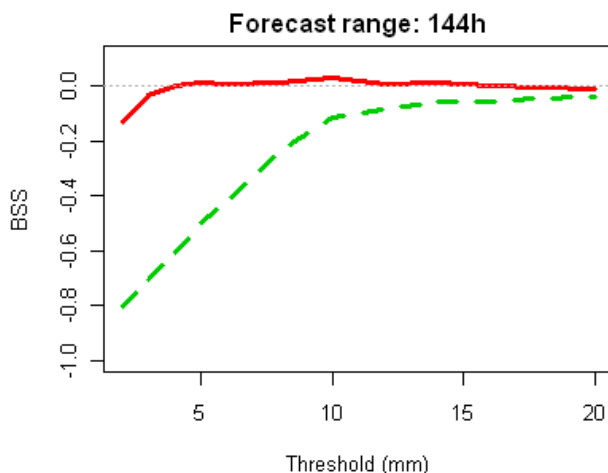
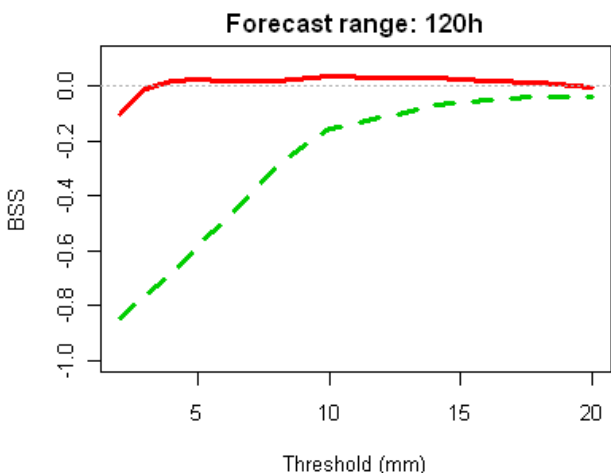
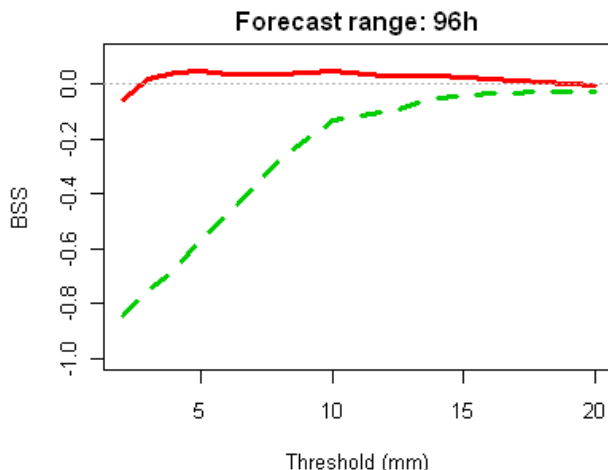
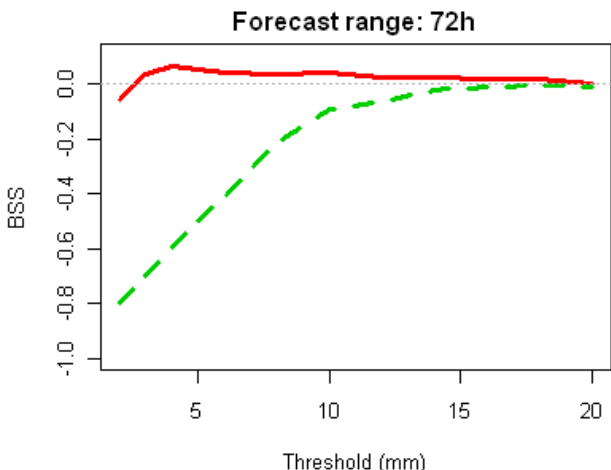
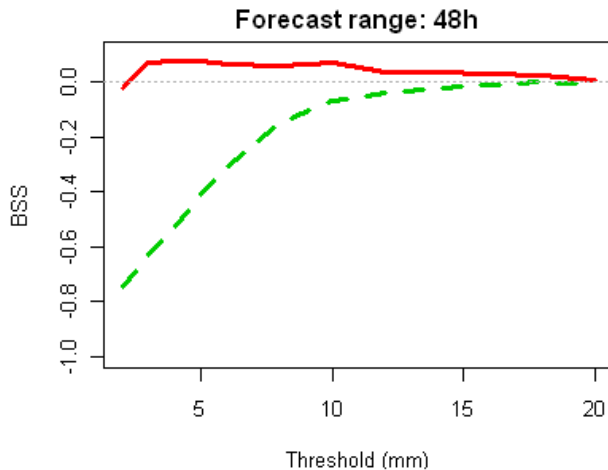
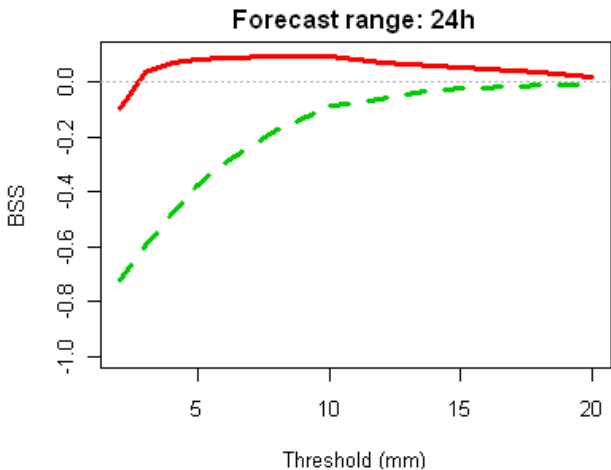
The Data

Scatterplot for precip



Time series of Brier skill score



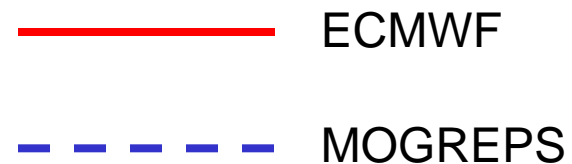


Brier Skill Scores by threshold

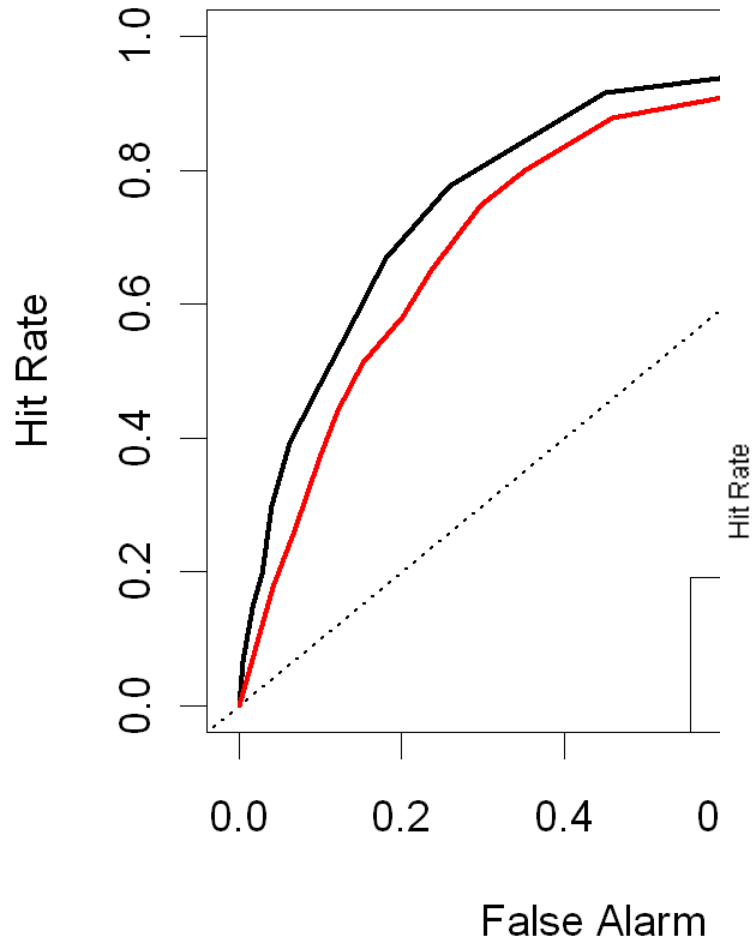
Sample climatology reference

MOGREPS remains unskilful at all forecast ranges and lead times

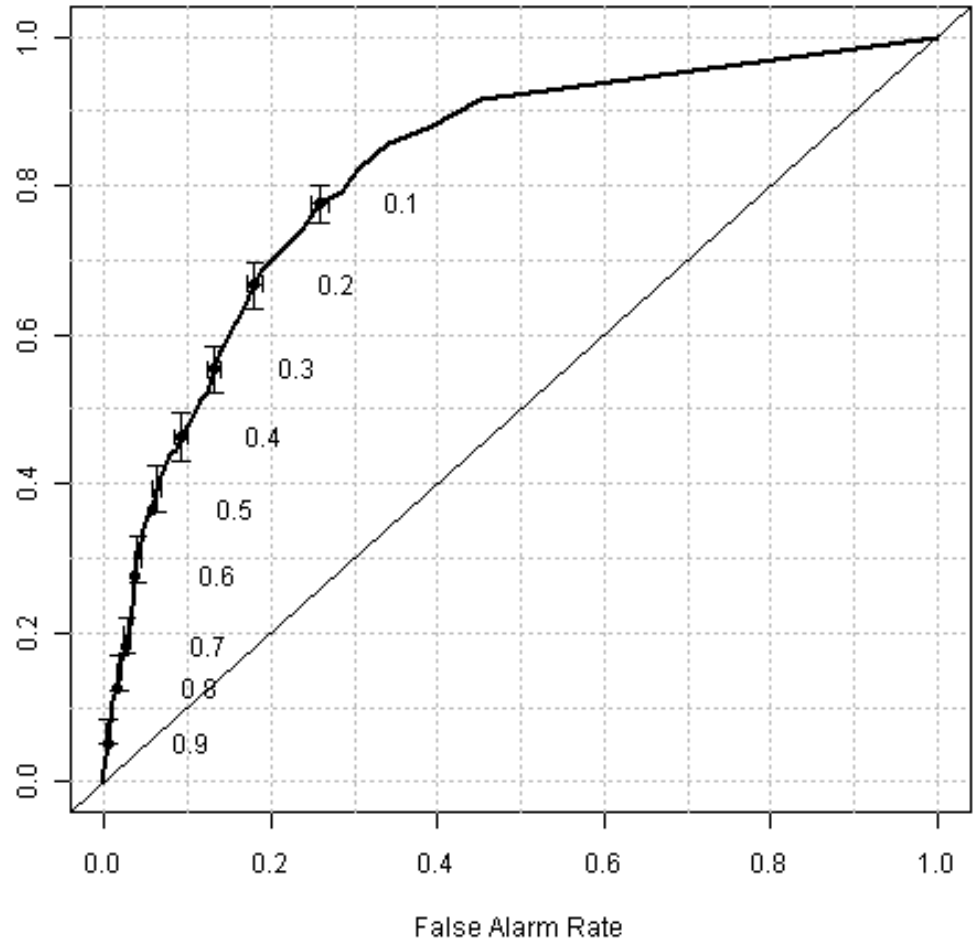
Both models tend towards climatology



24hr lead time :: All sites



24hr lead time :: ECMWF :: All sites



Verification of MWA forecasts over L. Victoria

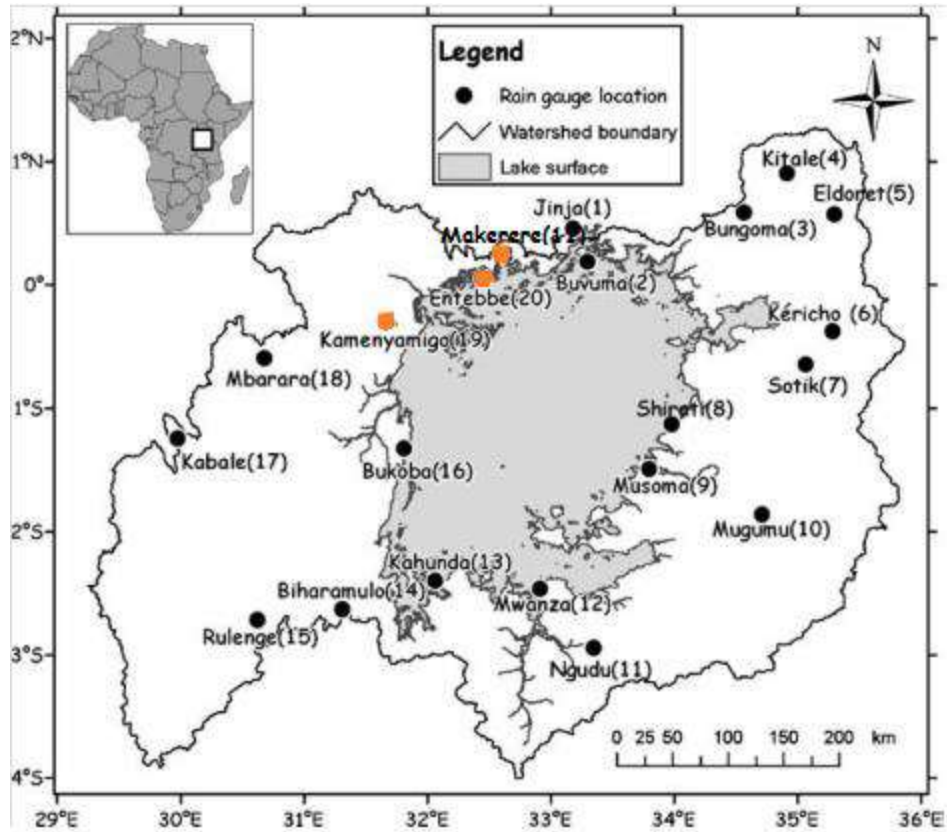
VERIFICATION OF MOBILE WEATHER ALERT FORECASTS OVER LAKE VICTORIA IN UGANDA

KHALID Y. MUWEMBE

MSc. Applied Meteorology and Climate with Management

September 2012

Stations used in L. Victoria study



Verification of UK 4 km L. Victoria model

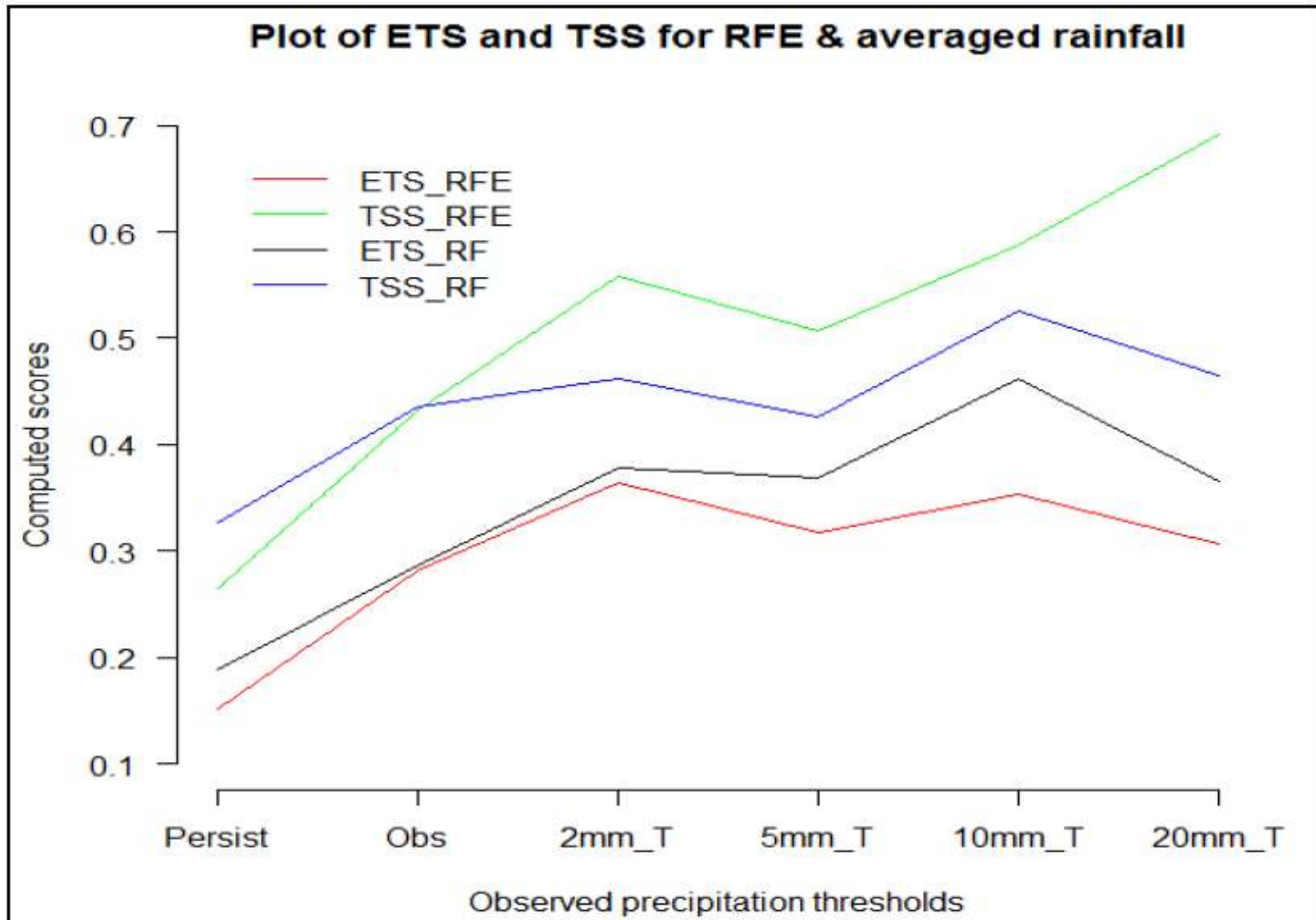


Figure 20: Computed ETS and TSS plotted against their respective observed threshold for both averaged rainfall and RFE



Summary and discussion....

- Summary
 - Keep the data!
 - Be clear about all forecasts!
 - Know why you are verifying and for whom!
 - Keep the verification simple but relevant!
 - Just do it!
- Discussion.....
- THANKS!

Verification of Probability forecasts

- 
- Brier Score (accuracy)
 - Reliability and reliability diagrams

The Brier Score

- Mean square error of a probability forecast

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2$$

- Weights larger errors more than smaller ones



- *Sharpness*: The tendency of probability forecasts towards categorical forecasts, measured by the variance of the forecasts
 - A measure of a forecasting strategy; does not depend on obs

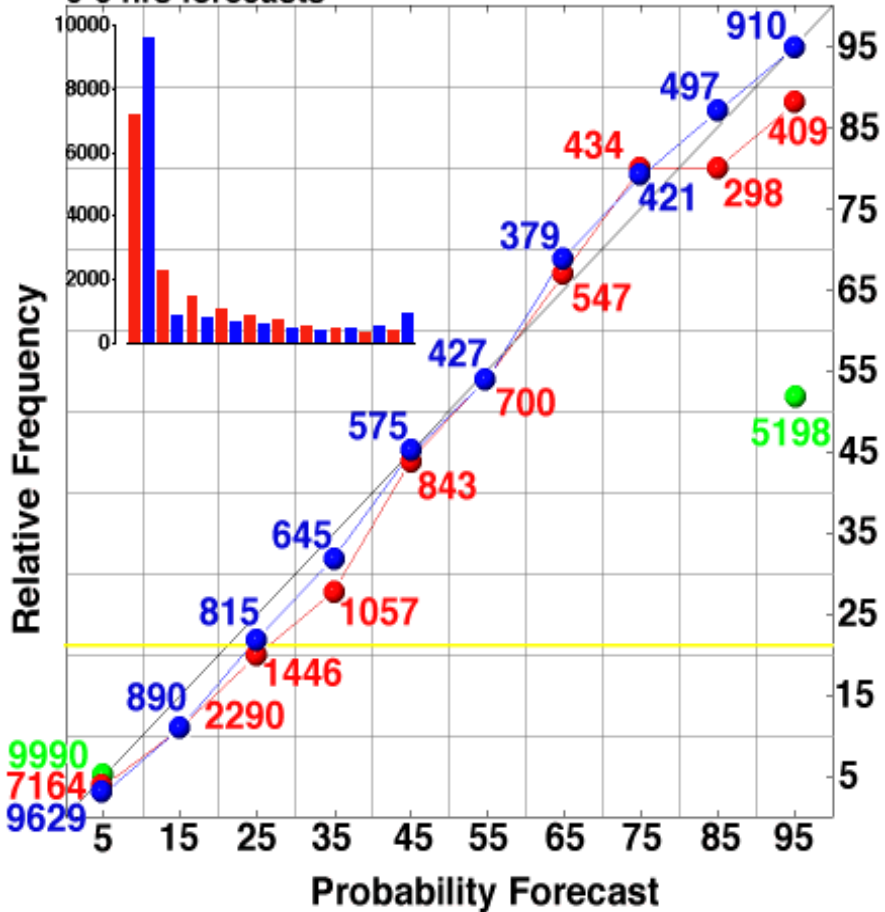


Probability forecast verification – Reliability tables

- Reliability:
 - The level of agreement between the forecast probability and the observed frequency of an event
 - Usually displayed graphically
 - Measures the bias in a probability forecast: Is there a tendency to overforecast or underforecast.
 - **Cannot be evaluated on a single forecast.**

Reliability Diagram

Reliability Table
0-6 hrs forecasts



Reliability Table
42-48 hrs forecasts

