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)

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" also to be put on the SWFDP website.

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
- Verification of the Regional severe weather charts (S. Landman)
- EXERCISE – Interpreting the table and scores
- Diversion into statistical interpretation
- 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
- Corollory: 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

- SWFDP: Both administrative goals and scientific goals
  - Administrative: WMO wants to know the impact of the program on the quality of severe weather forecasts
  - Scientific:
    - To decide which of the global center products are best to use for different forecasting problems.

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

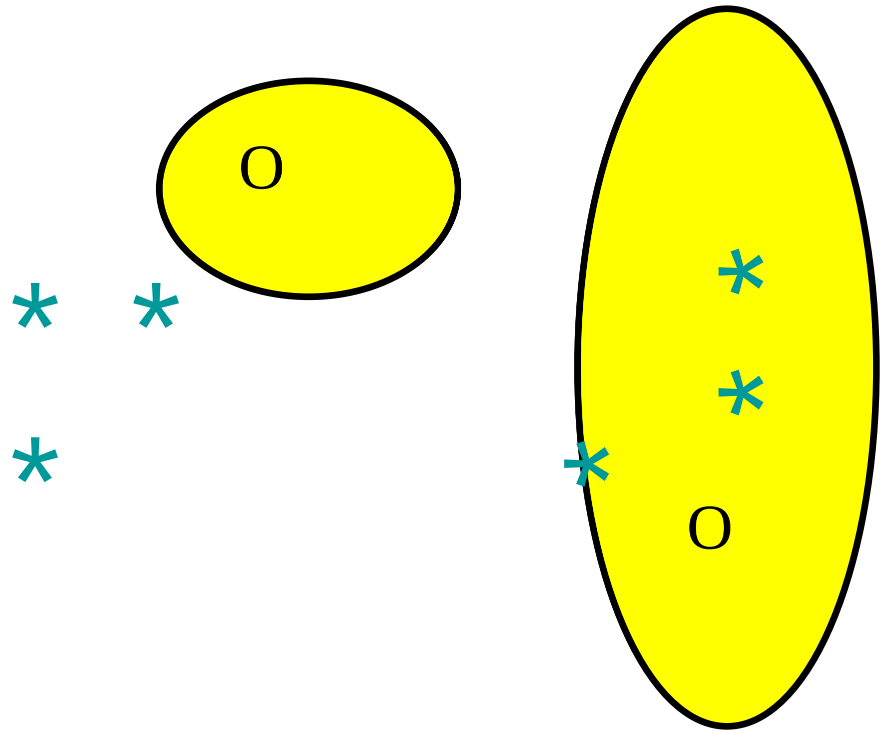
- Build a matched dataset of forecasts and observations
  - From events table – your forecasts
  - From datasets supplied from global and regional centers
- SWFDP: Predicted variables are **categorical**: Extreme events, where extreme is defined by thresholds of precipitation and wind. Some **probabilistic** 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

Progress Evaluation Table = Events Table

Event No.	Event type	Region	OBS start time (to nearest h in UTC)	OBS end time (to nearest h)	observations (list all reports in region)	Severe weather observed? (Yes=1, No=0)	Warning Issued? (Yes=1, No=0)	FCST start time (to nearest h)	FCST end time (to nearest h)	Lead time of warning (0=time of observed start)	Impact of event	Impact of the warning
	strong wind >20KT	ENTIRE COAST	06UTC	13UTC	Pemba 20KT Zanzibar 25KT D'salaam 20KT Mtwara 20KT	1	1	18/07/12 06UTC	18/07/12 18UTC	1	A BOAT SANK IN OCEAN	WARNING DELIVERED
EVENT DATE 18/07/2012						Guidance:	RSMC: (check each one used)	Evaluation: 1 to 4 (1=useless, 4=best) 4	Other Products (check each one used)	Evaluation: 1 to 4 (1=useless; 4=best)		
							Severe weather chart: ✓ Prob Table ✓	4 4	ECMWF: ✓ NCEP: UKMO global: UKMO regional:	4		
1	strong wind >20KT	ENTIRE COAST				0	1	18/07/12 18UTC	19/07/12 18UTC			WARNING DELIVERED
EVENT DATE 19/07/2012						Guidance:	RSMC: (check each one used)	Evaluation: 1 to 4 (1=useless, 4=best)	Other Products (check each one used)	Evaluation: 1 to 4 (1=useless; 4=best)		
							Severe weather chart: ✓ Prob Table ✓	3 3	ECMWF: ✓ NCEP: UKMO global: UKMO regional: ✓	ECMWF: 3 NCEP: UKMO global: UKMO regional: 3		
2	strong wind >20KT	ENTIRE COAST	10 UTC	17 UTC	Zanzibar 20-25KT Pemba 20KT D'salaam 20KT Mtwara 20KT	1	1	25/07/12 18UTC	26/07/12 18UTC	4	NO IMPACT REPORTED	WARNING DELIVERED
EVENT DATE 26/07/2012						Guidance:	RSMC: (check each one used)	Evaluation: 1 to 4 (1=useless, 4=best)	Other Products (check each one used)	Evaluation: 1 to 4 (1=useless; 4=best)		
							Severe weather chart: Prob Table		ECMWF: ✓ NCEP: UKMO global: UKMO regional: ✓	ECMWF: 4 NCEP: UKMO global: UKMO regional: 4		
3	strong wind >20KT	ENTIRE COAST	09 UTC	13 UTC	Zanzibar 20KT Tanga 20KT	1	1	26/07/12 18UTC	27/07/12 18UTC	4	NO IMPACT REPORTED	WARNING DELIVERED
EVENT DATE 27/07/2012						Guidance:	RSMC: (check each one used)	Evaluation: 1 to 4 (1=useless, 4=best)	Other Products (check each one used)	Evaluation: 1 to 4 (1=useless; 4=best)		
							Severe weather chart: Prob Table		ECMWF: ✓ NCEP: UKMO global: ✓ UKMO regional: ✓	ECMWF: 4 NCEP: UKMO global: 3 UKMO regional: 4		
4	strong wind >20KT	ENTIRE COAST	08 UTC	13 UTC	Mtwara 20KT D'salaam 20KT Zanzibar 20KT	1	1	28/07/12 18UTC	29/07/12 18UTC	4	NO IMPACT REPORTED	WARNING DELIVERED
EVENT DATE 29/07/2012						Guidance:	RSMC: (check each one used)	Evaluation: 1 to 4 (1=useless, 4=best)	Other Products (check each one used)	Evaluation: 1 to 4 (1=useless; 4=best)		
							Severe weather chart: ✓ Prob Table ✓	4 4	ECMWF: ✓ NCEP: UKMO global: UKMO regional: ✓	ECMWF: 4 NCEP: UKMO global: UKMO regional: 3		
5	strong wind >20KT	ENTIRE COAST				0	1	29/07/12 18UTC	30/07/12 13UTC			WARNING DELIVERED
EVENT DATE 30/07/2012						Guidance:	RSMC: (check each one used)	Evaluation: 1 to 4 (1=useless, 4=best)	Other Products (check each one used)	Evaluation: 1 to 4 (1=useless; 4=best)		
							Severe weather chart: ✓ Prob Table ✓	3 2	ECMWF: ✓ NCEP: UKMO global: UKMO regional:	ECMWF: 3 NCEP: UKMO global: UKMO regional:		

Preparation of the contingency 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

Forecasts		Yes	No	
	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

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)

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

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

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
- ETS = Equitable threat score is the TS adjusted for number correct by chance

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

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

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

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.

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		
	Yes	No	Marginal total
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

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

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

Base rate independence ⇔ Functions of H and F

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

Extremal Dependency Index - EDI

Symmetric Extremal Dependency Index - SEDI

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

JWGFVR

Comments on the extreme dependency family

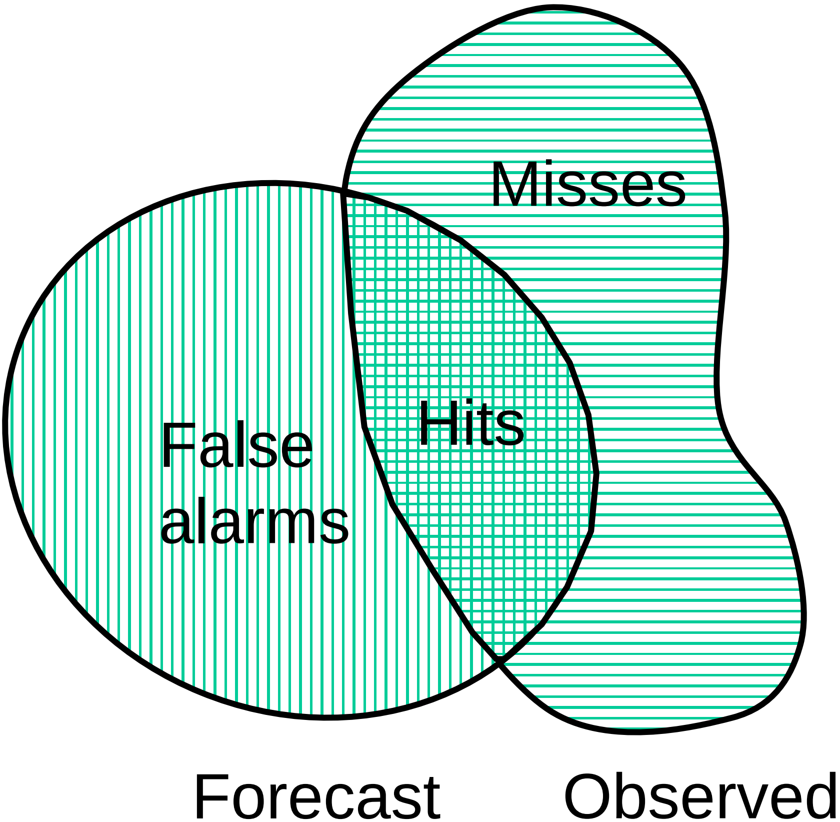
- EDS now discredited
  - Sensitive to base rate
  - NOT sensitive to false alarms
- SEDS
  - Weakly sensitive to base rate, but useful
  - Useful to forecasters because uses the forecast frequency
- EDI
  - User-oriented, function of HR and FA like HK and ROC
  - Absolutely independent of base rate
- SEDI
  - Like EDI, but has additional property of symmetry; not necessarily important for our purposes

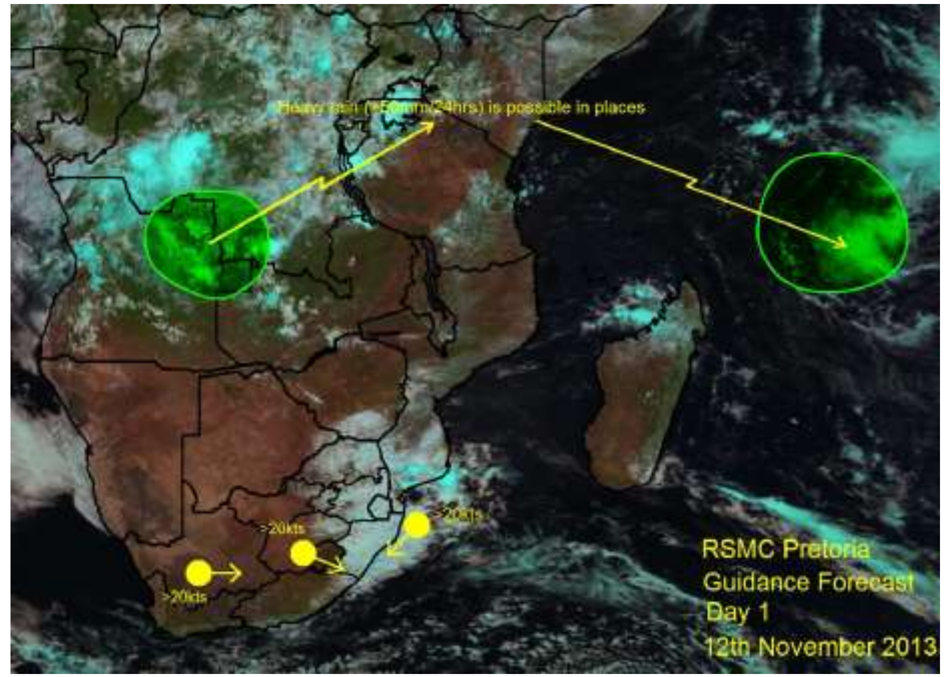
Mozambique Interpretation Exercise

- 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

JWGFVR

Spatial verification of RMSC products



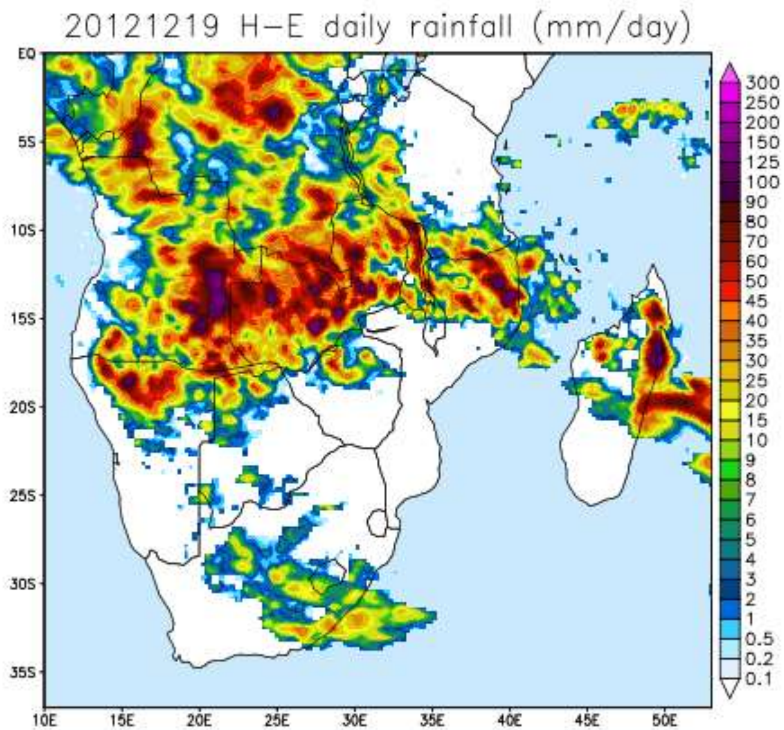


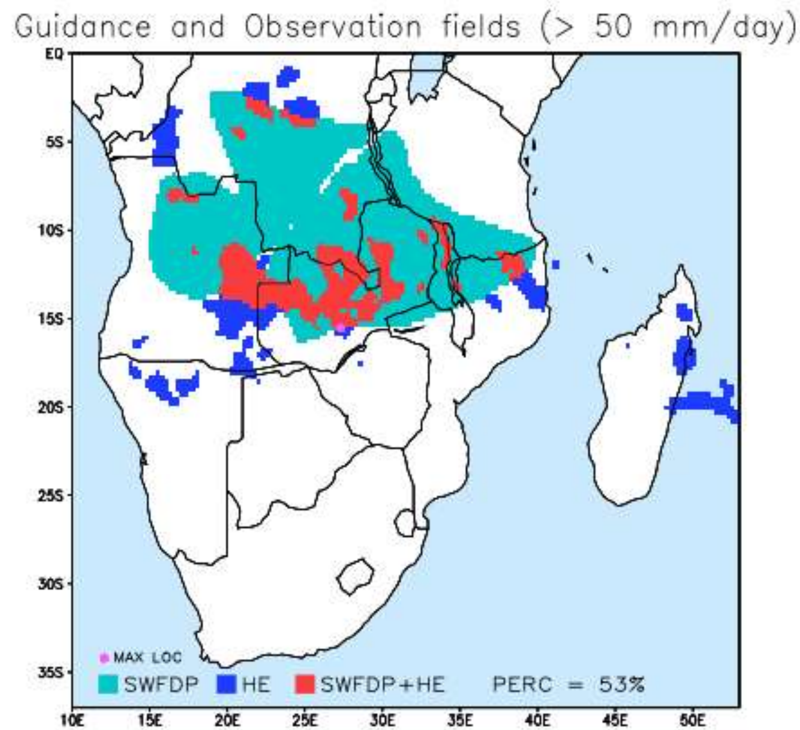
Spatial contingency table:

-Can accomplish IF one has quasi-continuous spatial observation data

-Stephanie's method

Verification of regional forecast map using HE





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 (km²*10⁴)	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

GUIDANCE \ OBSERVATION	>=50	<50
>=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

WMO
South African Weather Service

Capital Cities Verification

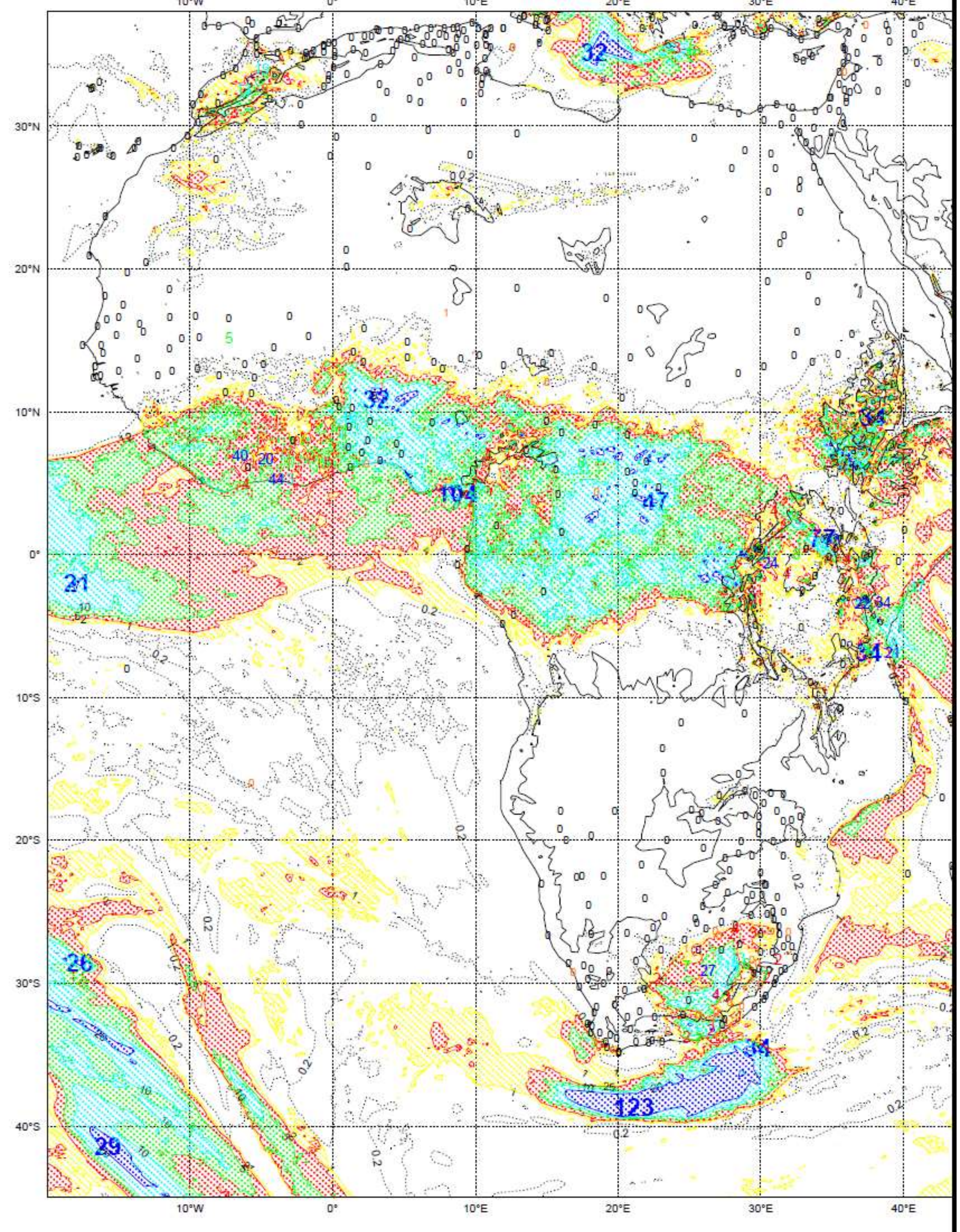
- Forecast is nearest gridpoint to Capital cities of all countries
- Observation is HE estimated precipitation at that point (top row) and Max HE estimated precipitation within 50 km (bottom).
- About 5 years of data – allows for enough severe precipitation cases at a single location (usually) about 1825 cases.
- 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.

Capital Cities Verification

- Results are loaded into the "summary" page for easy comparison
- Summary page setup:
  - Top 2 rows of results: "nearest point" and "50km radius" verification - 2014 dataset – 5 years of data
  - Bottom 2 rows: data from 2013 lab, 3.5 years of data
    - Can check to see if forecasts have improved on average in last 1.5 yr.
- 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.
  - REMEMBER: The observations are an interpretation of satellite data with influence from a model.
  - Consider: Hit rate, false alarm ratio, bias, ETS, SEDS, EDI
  - Nominate a presenter from each group to discuss

ECMWF Diagnostic chart:

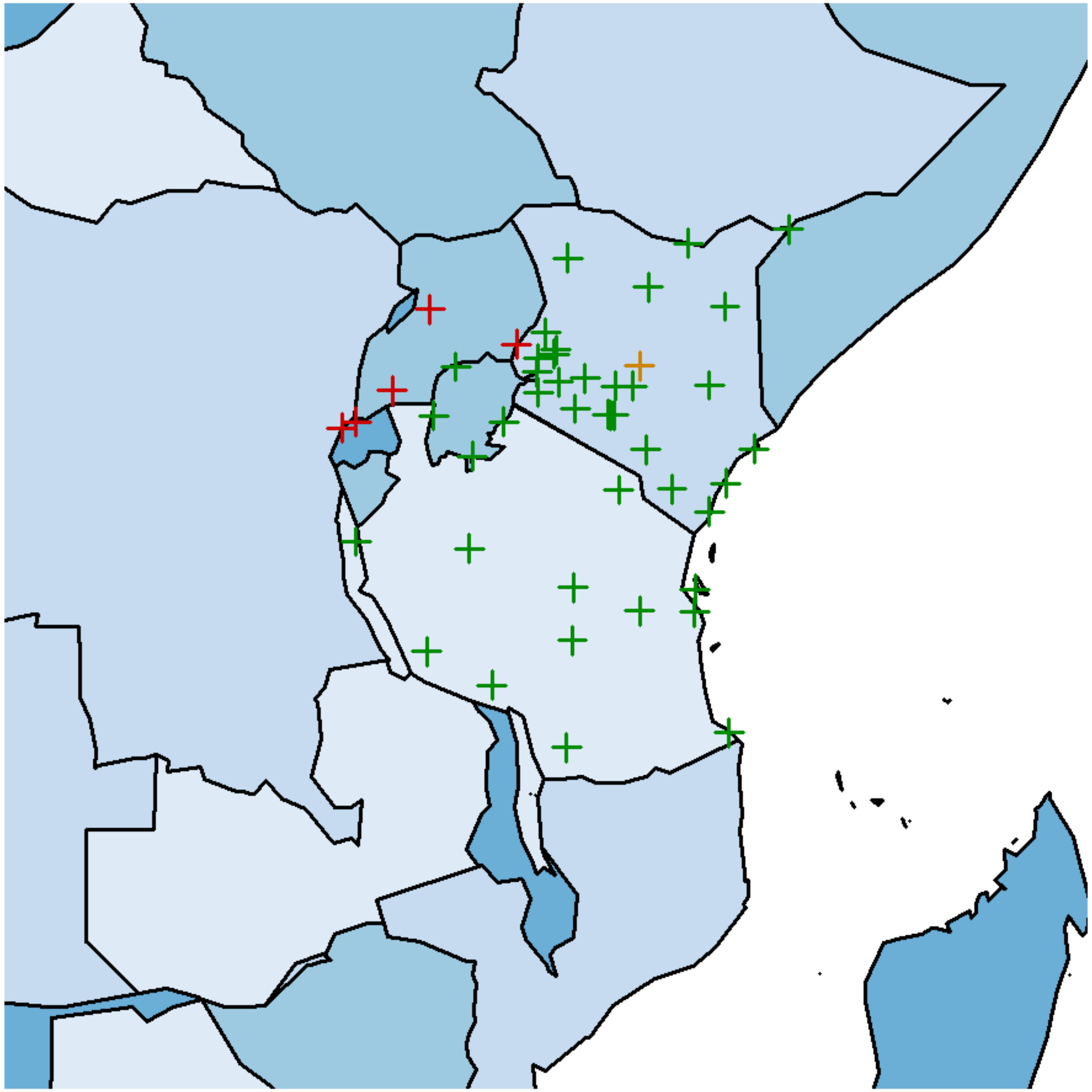
-Daily precipitation values plotted vs forecast amts.



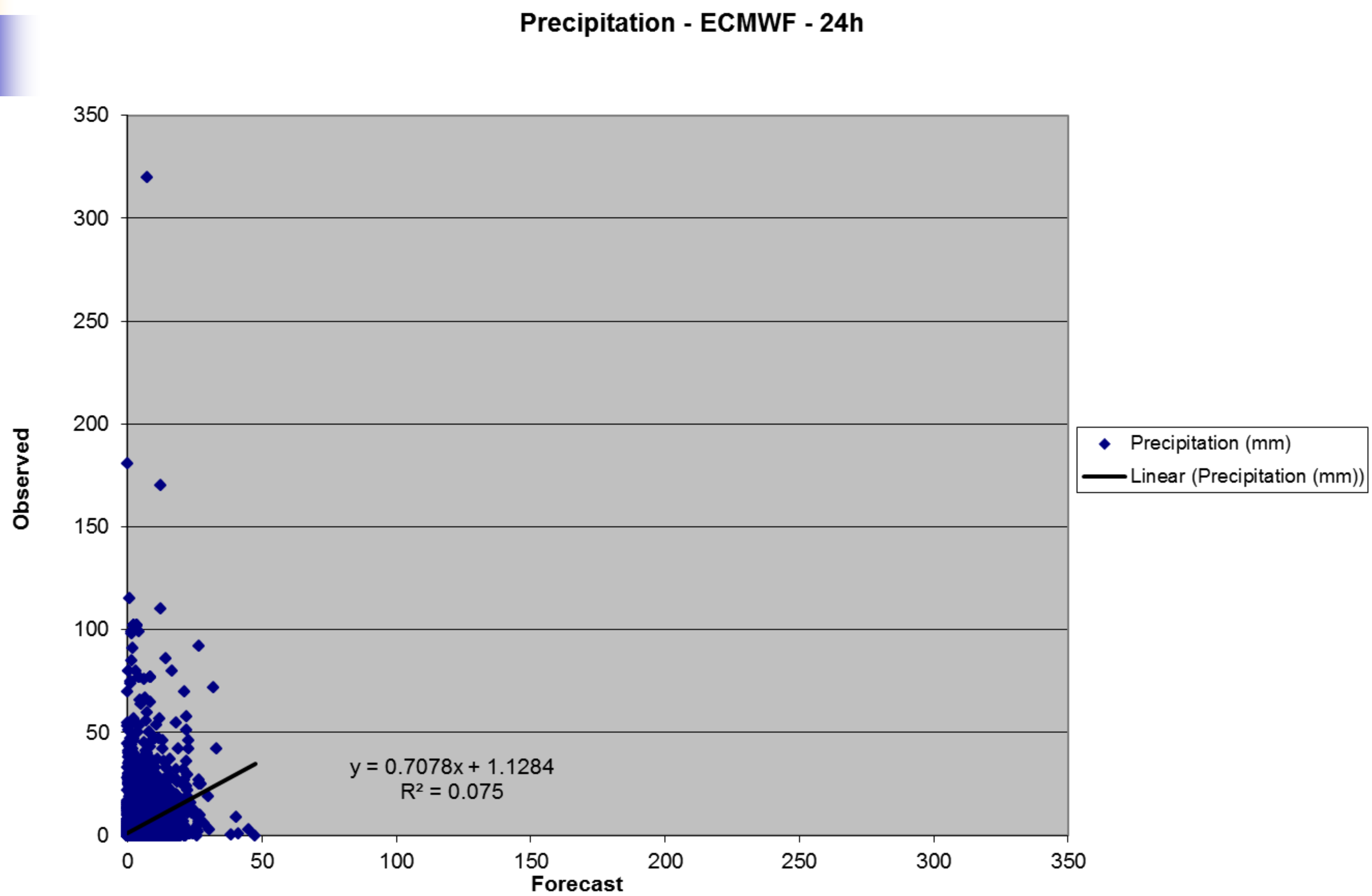
Verification in E. Africa project

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

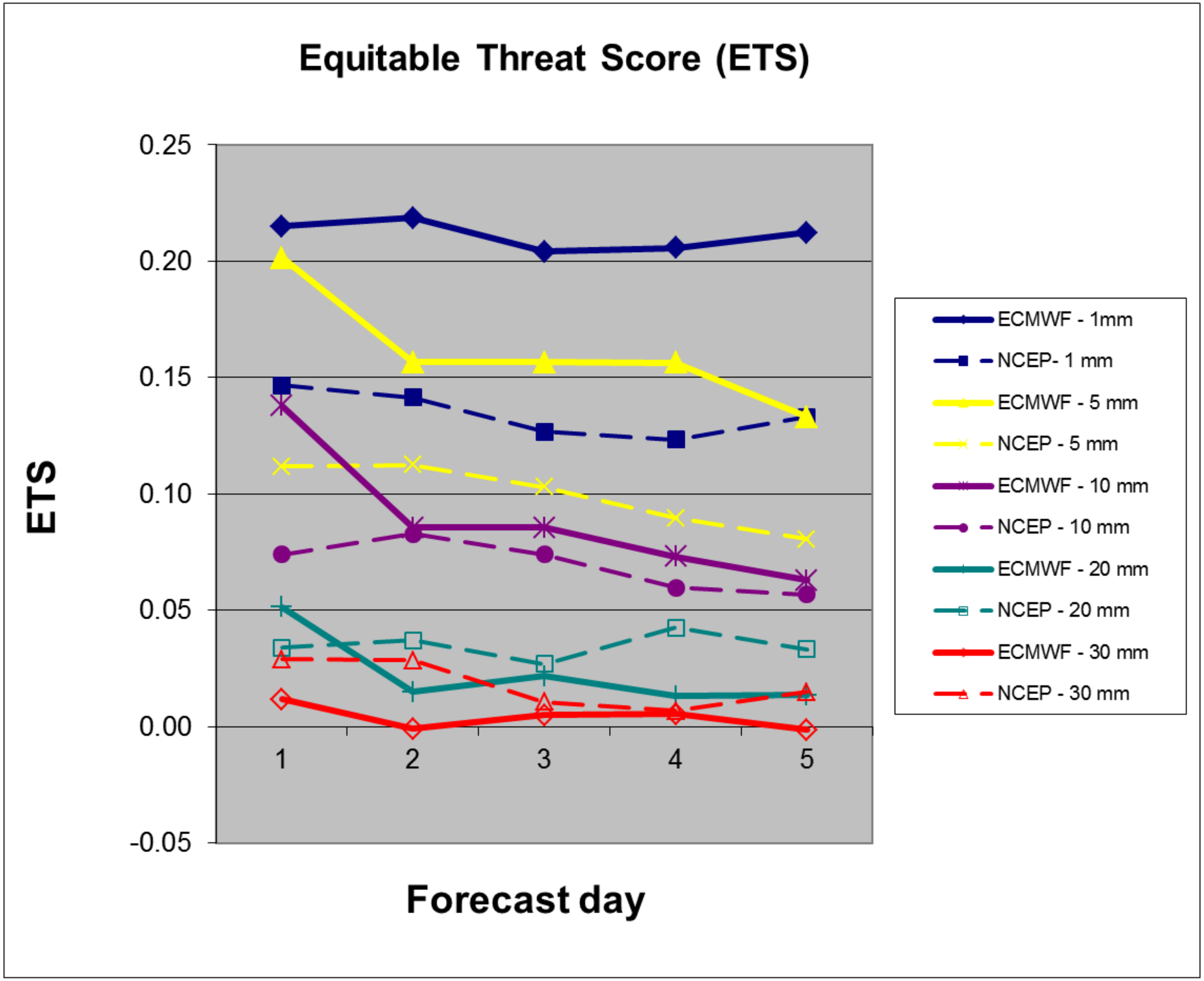
Global model verification Sept 2010 to May, 2011
Stations available



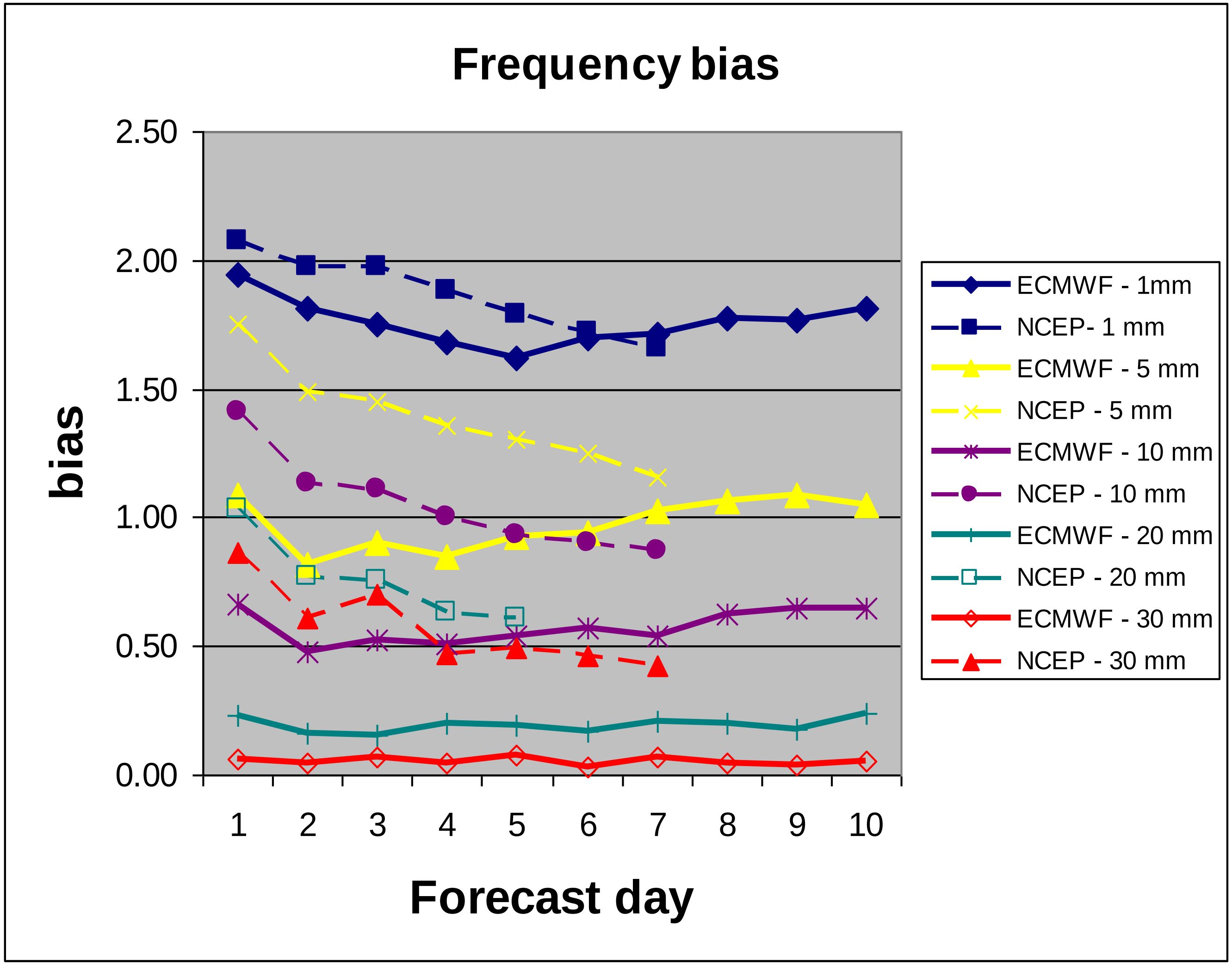
Scatterplot - ECMWF



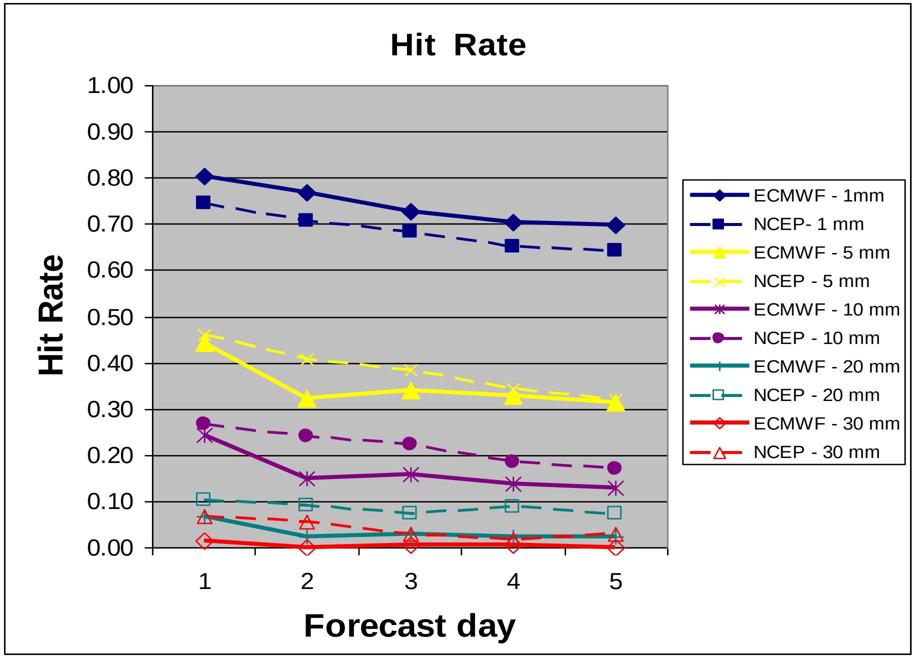
ECMWF vs NCEP 24h precipitation

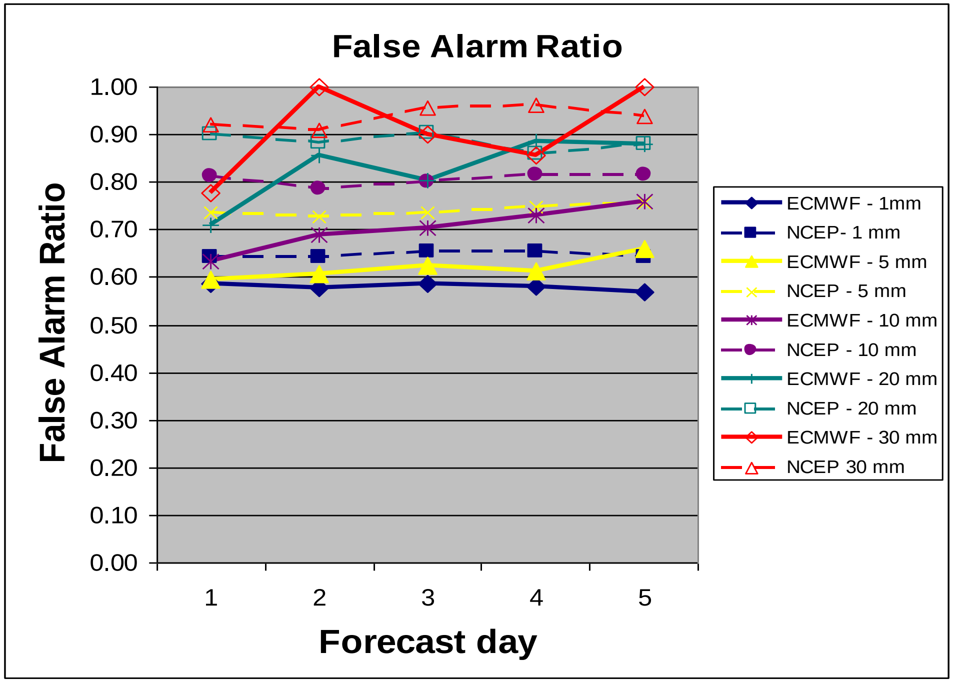


ECMWF vs. NCEP 24 h precipitation (2)

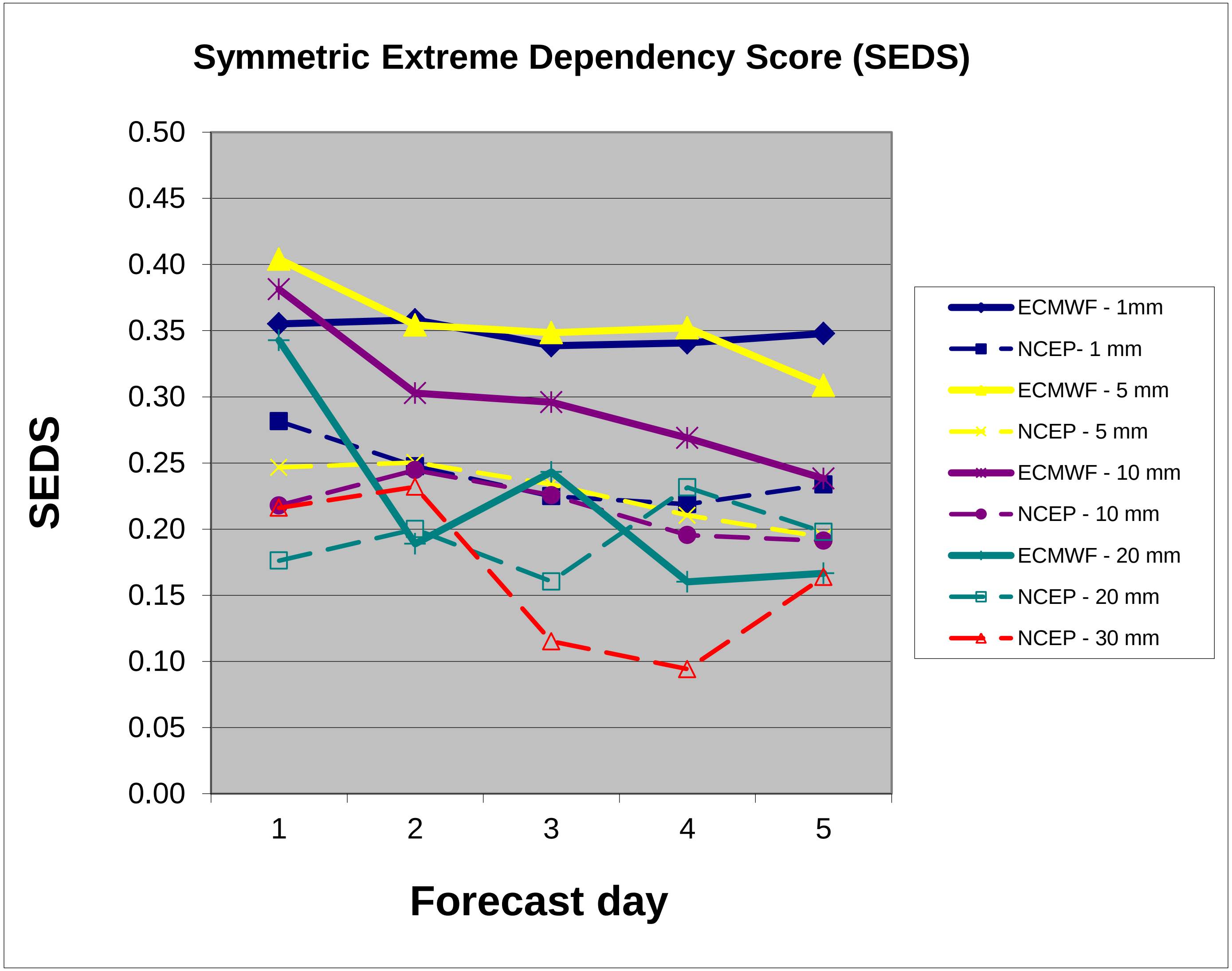


ECMWF vs NCEP 24h precipitation (3)





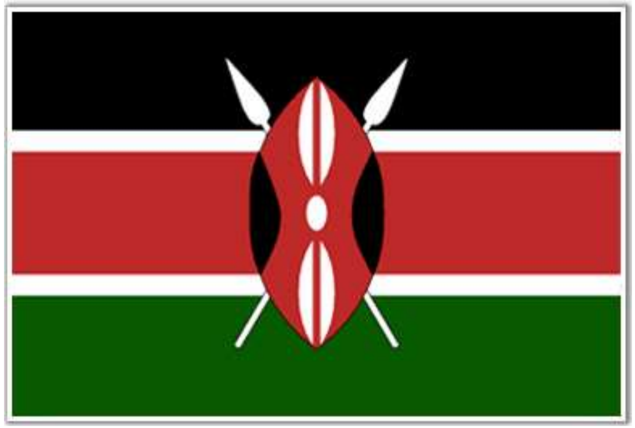
ECMWF vs NCEP 24 h Precipitation (4)

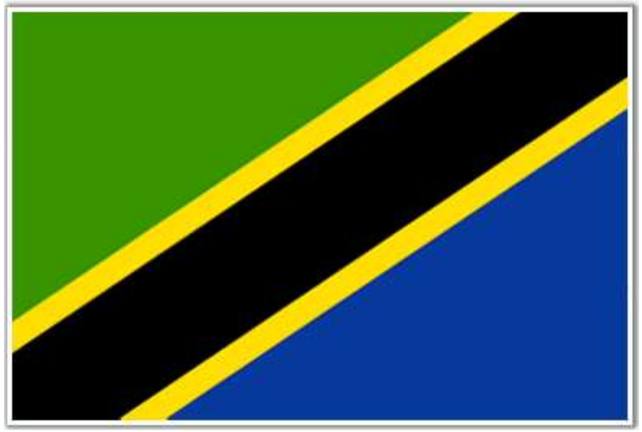


ECMWF Vs. MOGREPS

Africa Cup

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

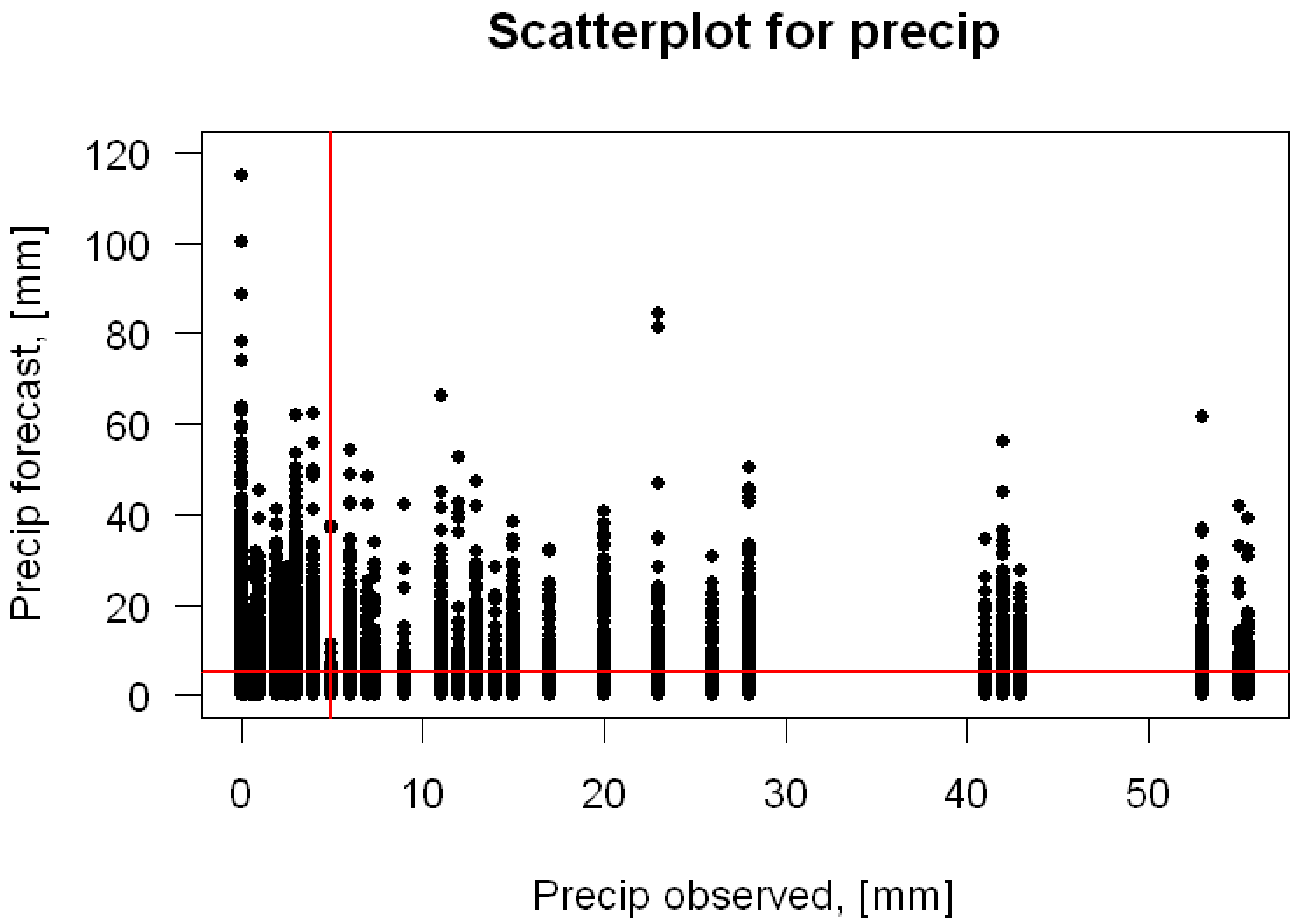




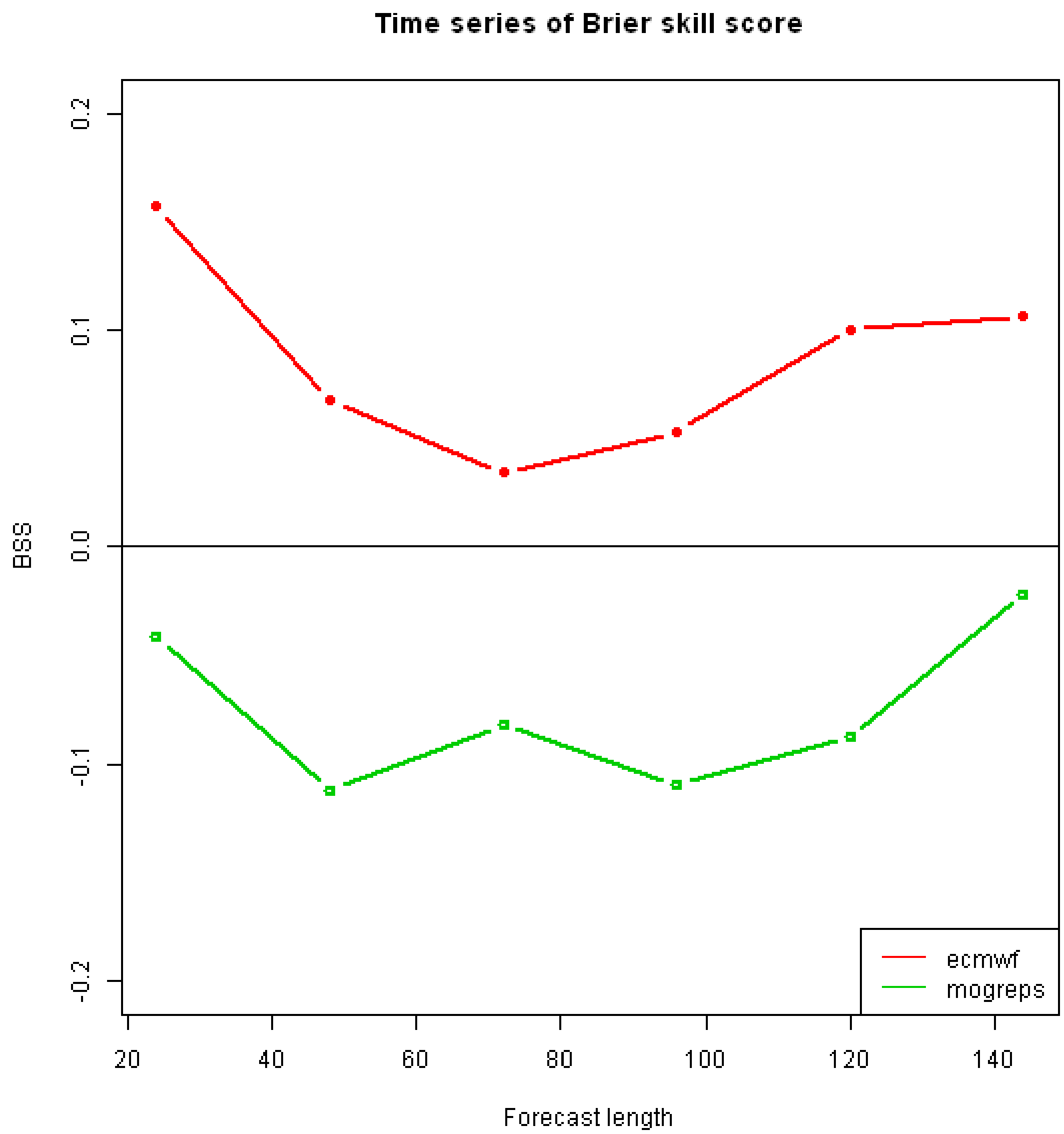
Rules of the Match

- 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

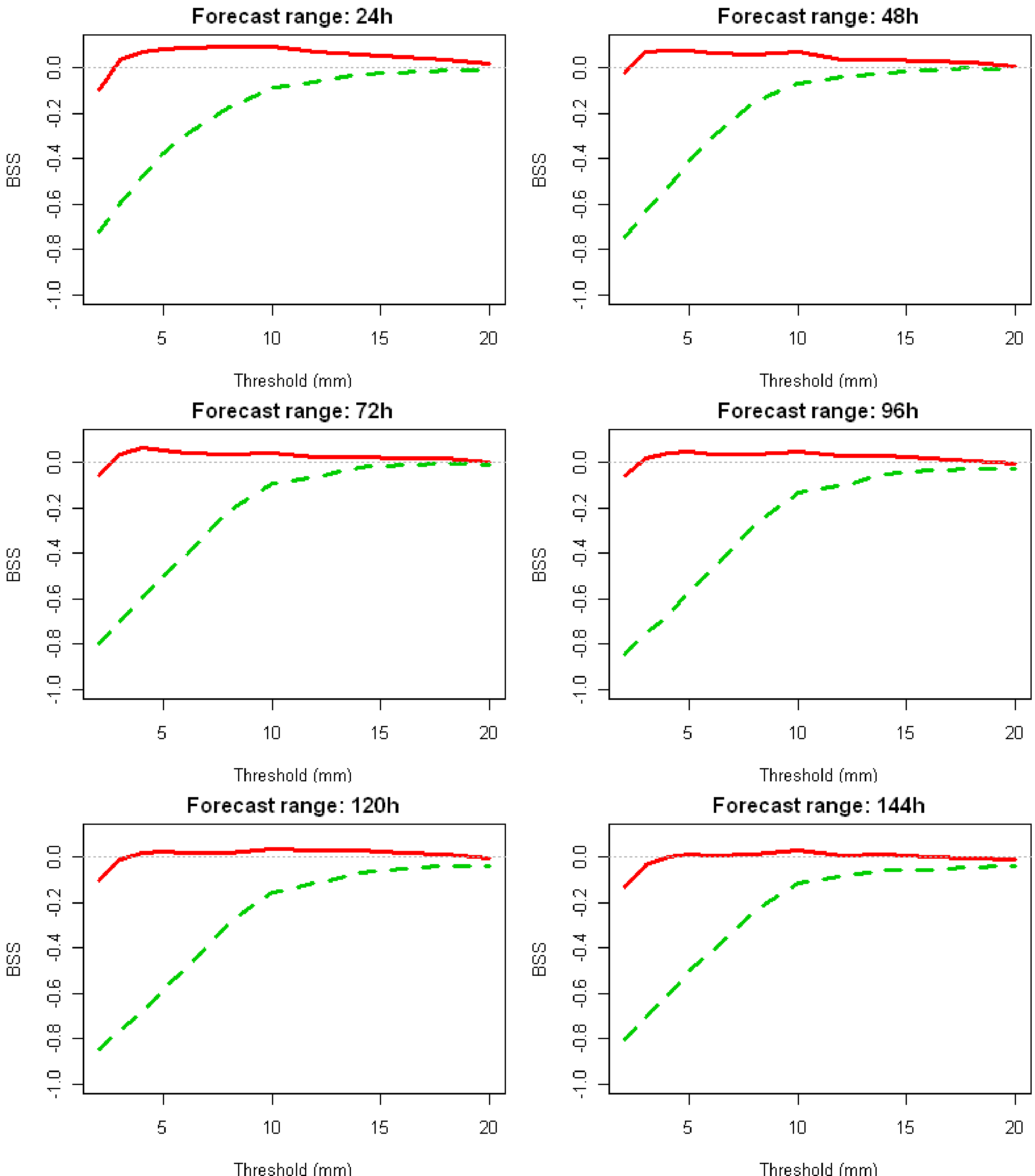
The Data



Scatterplot for precip
Precip forecast, [mm]
120
100
80
60
40
20
0
0
10
20
30
40
50
Precip observed, [mm]



Time series of Brier skill score
BSS
0.2
0.1
0.0
-0.1
-0.2
20
40
60
80
100
120
140
Forecast length
ecmwf
mogreps

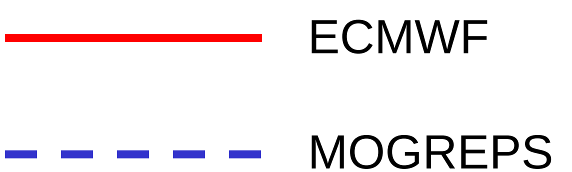


Brier Skill Scores by threshold

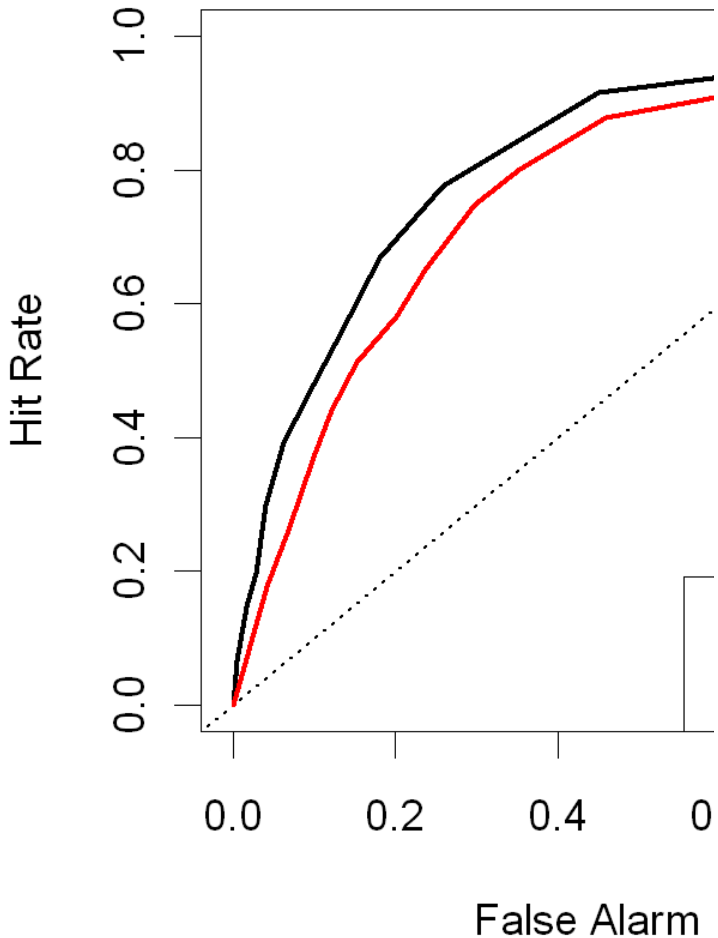
Sample climatology reference

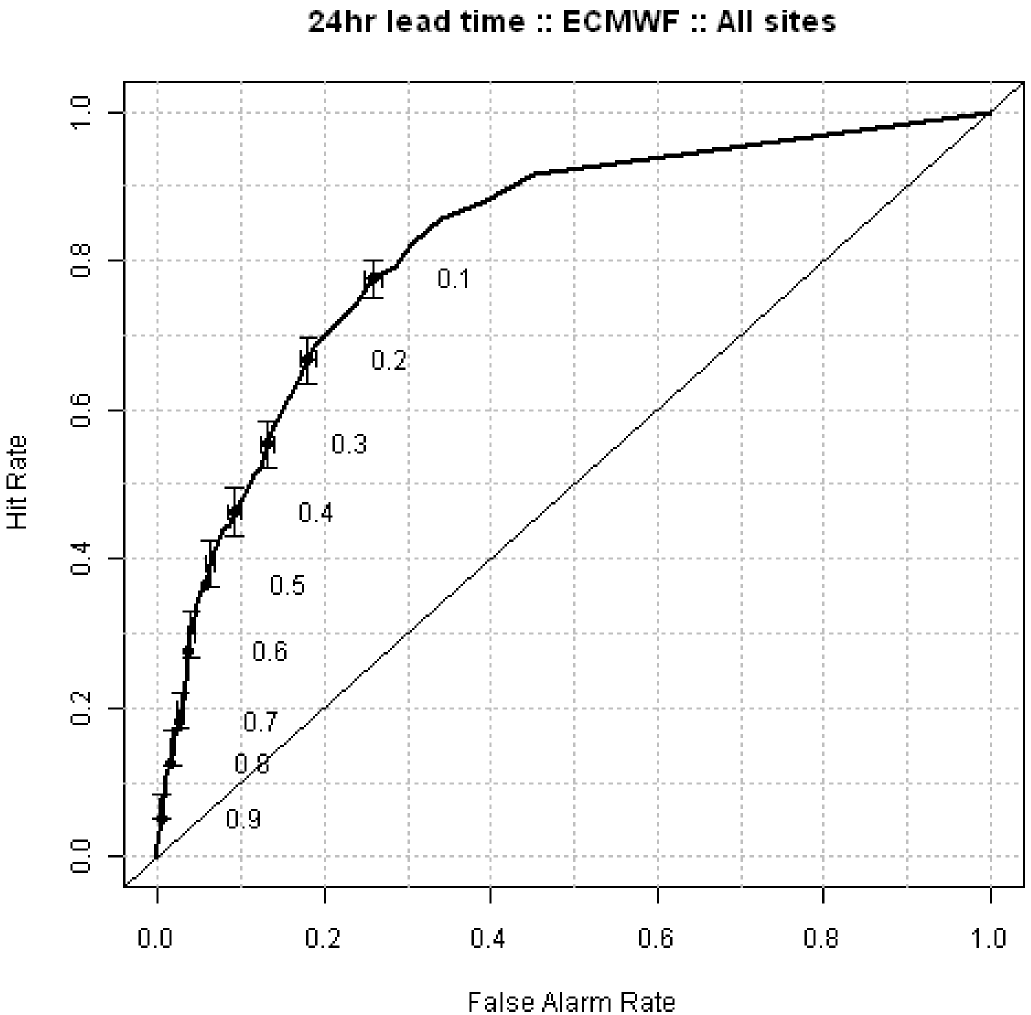
MOGREPS remains unskilful at all forecast ranges and lead times

Both models tend towards climatology



24hr lead time :: All sites



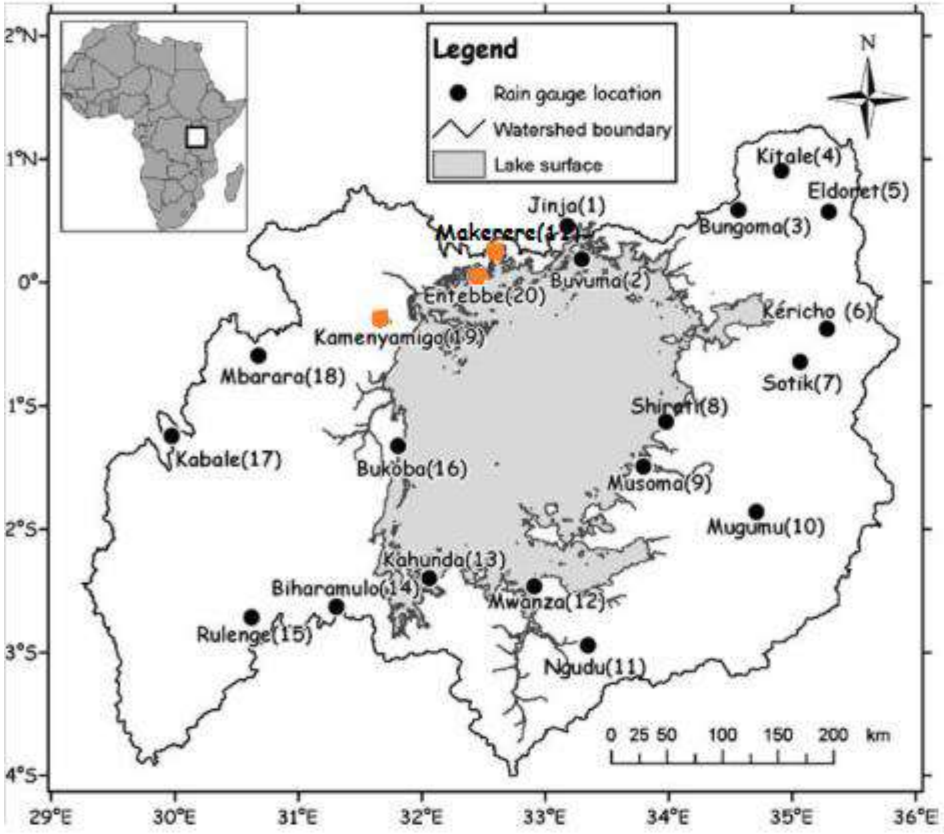


Verification of MWA forecasts over L. Victoria

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

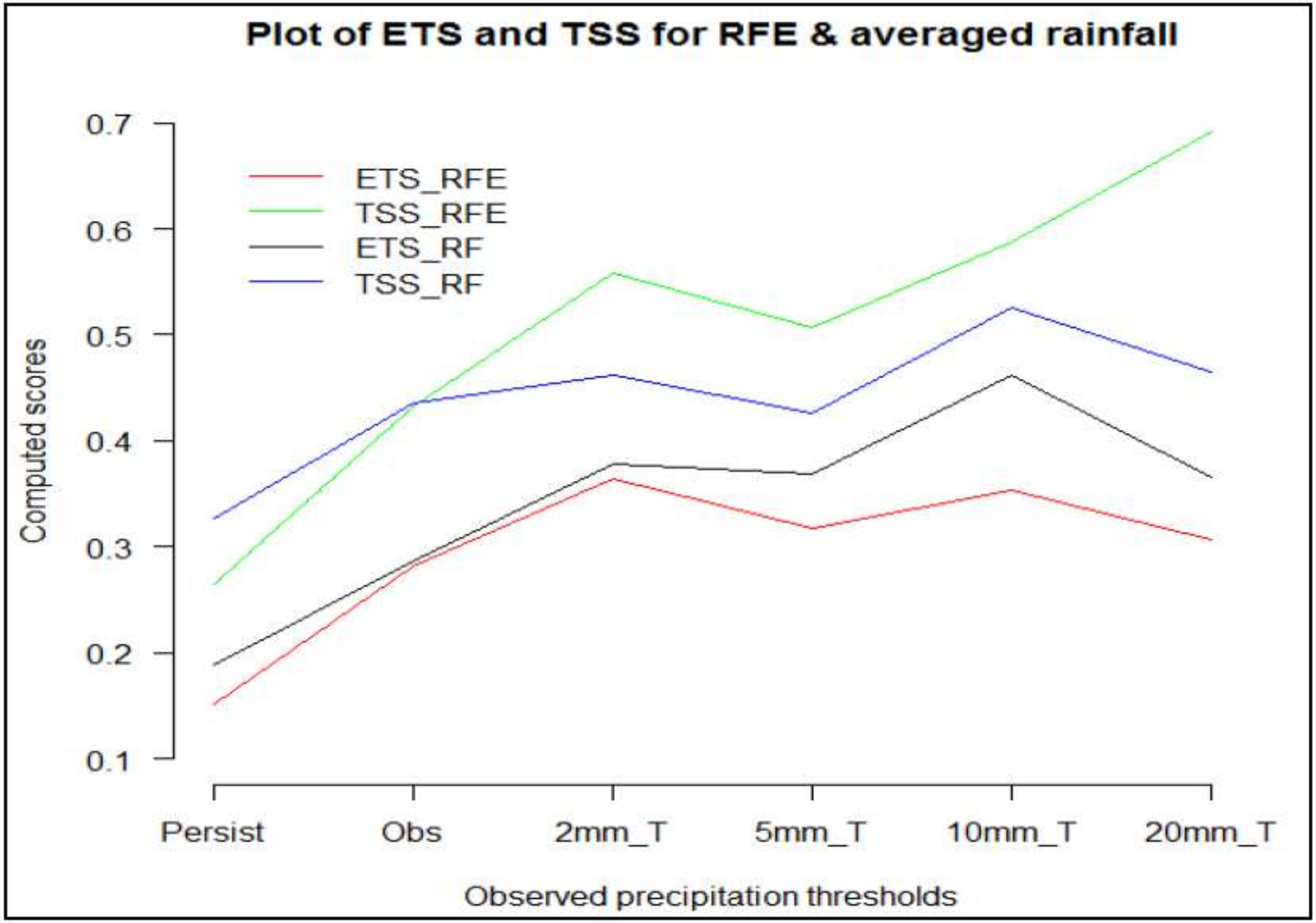


Figure 20: Computed ETS and TSS plotted against their respective observed threshold for both averaged rainfall and RFE

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

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

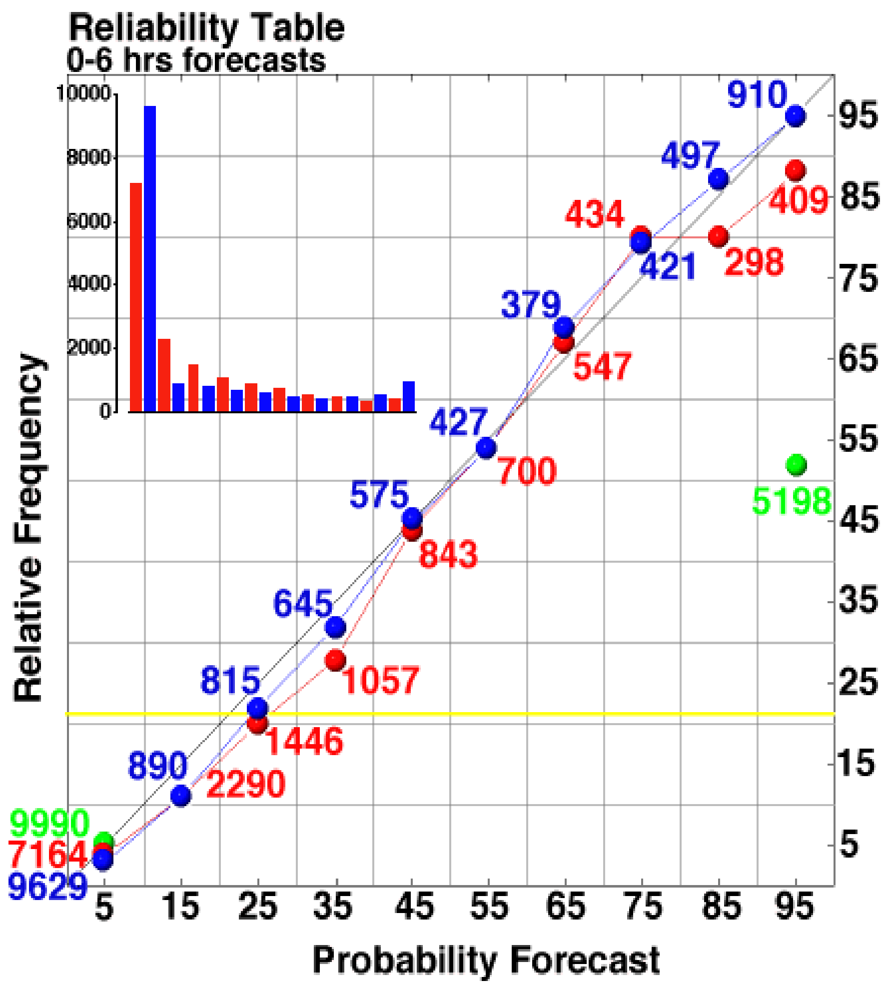
- Weights larger errors more than smaller ones

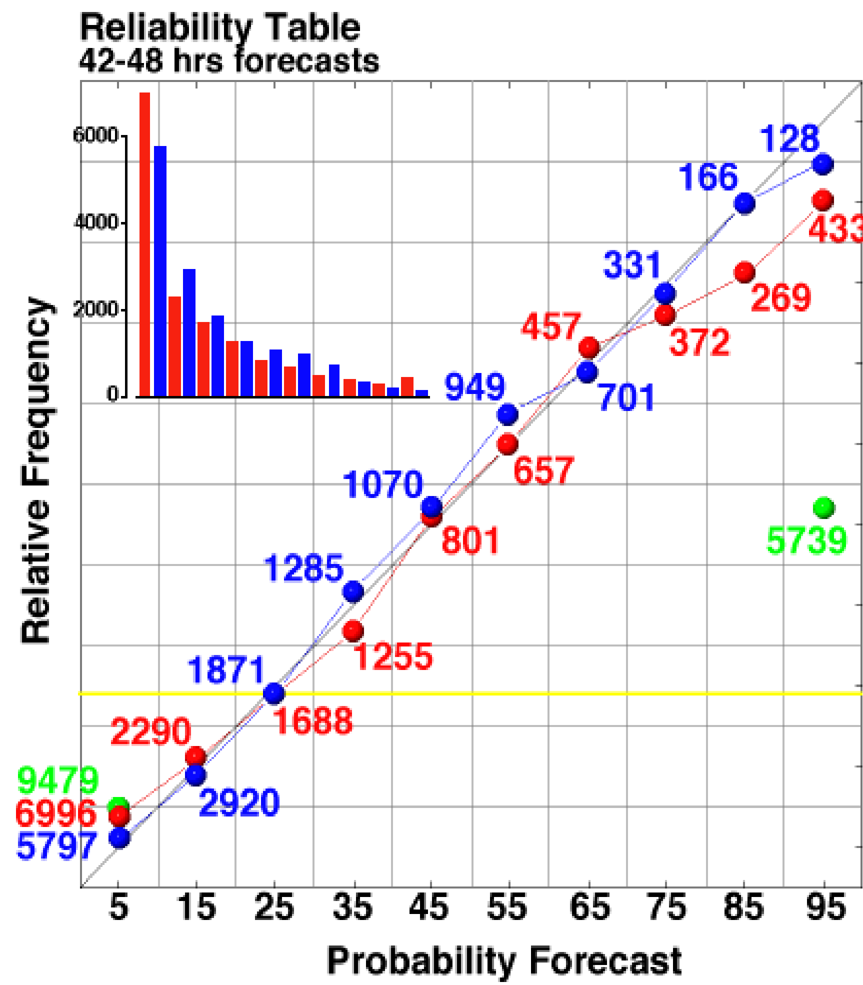
- *Sharpness*: The tendency of probability forecasts towards categorical forecasts, measured by the variance of the forecasts
  - A measure of a forecasting strategy; does not depend on obs

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**.

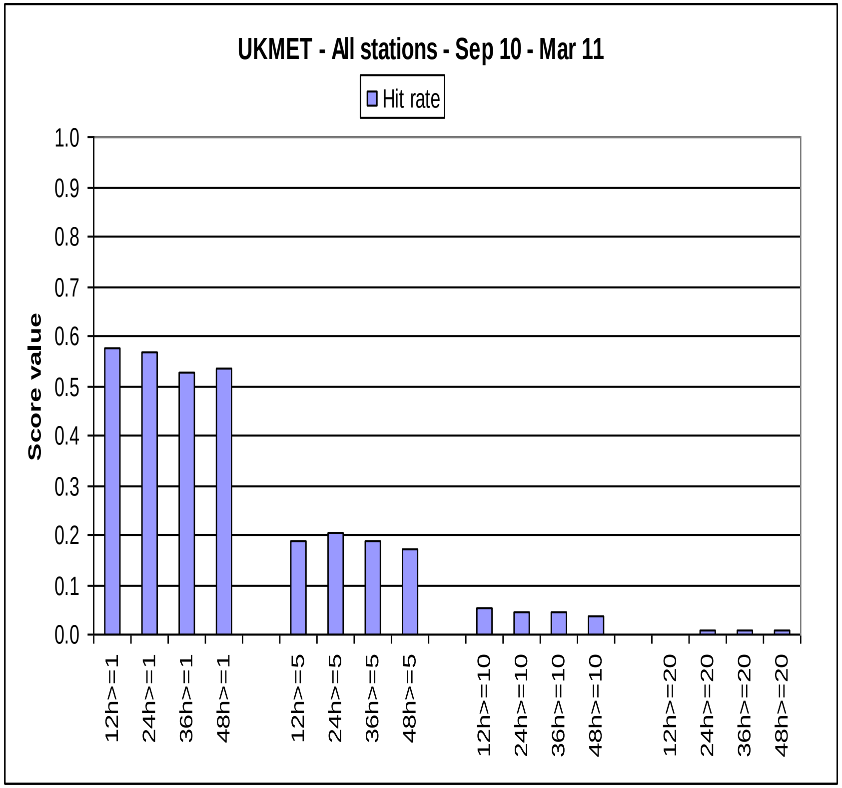
Reliability Diagram

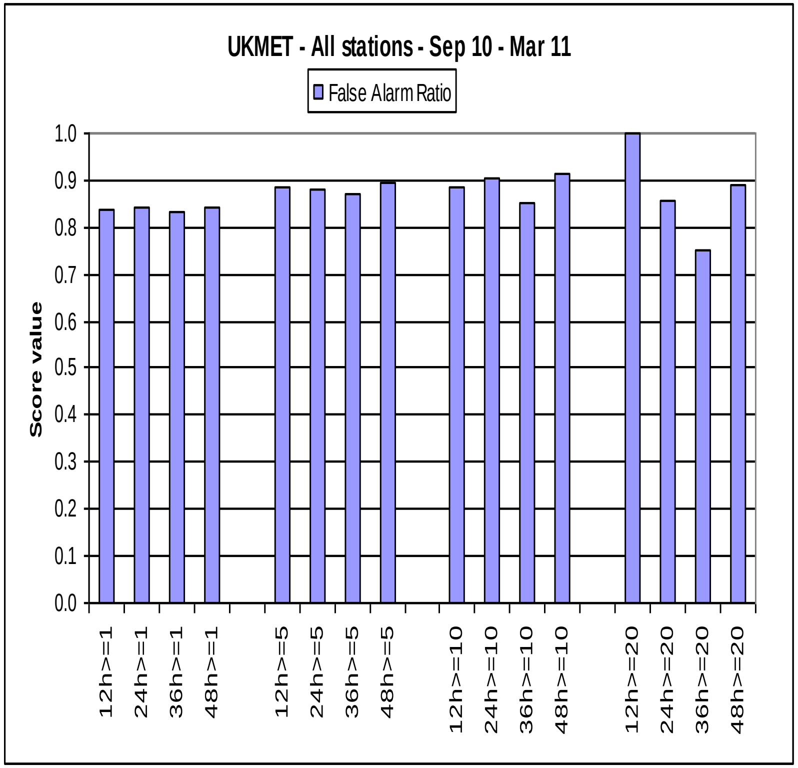




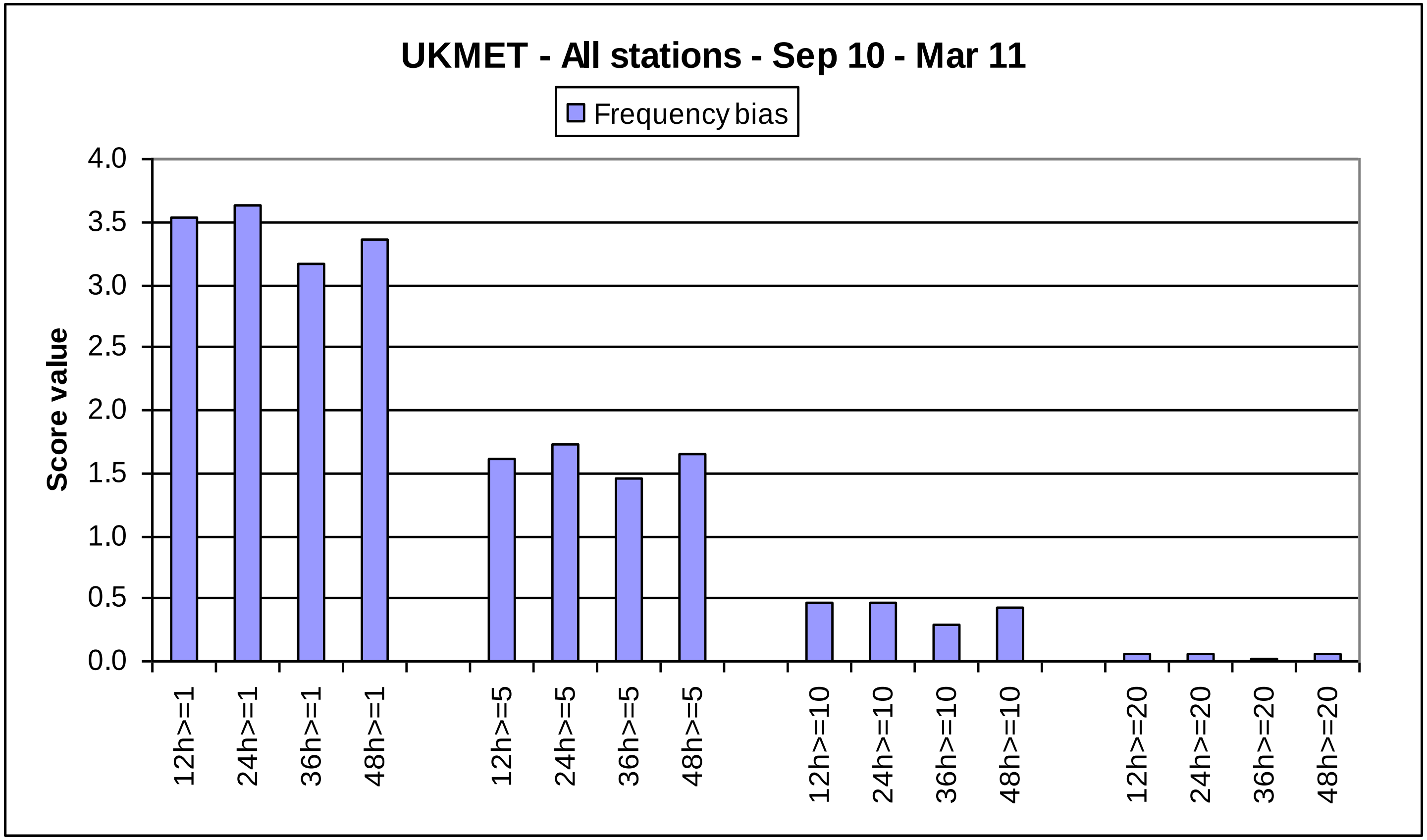
UK MET RESULTS – E AFRICA

Hit rate, false alarm ratio

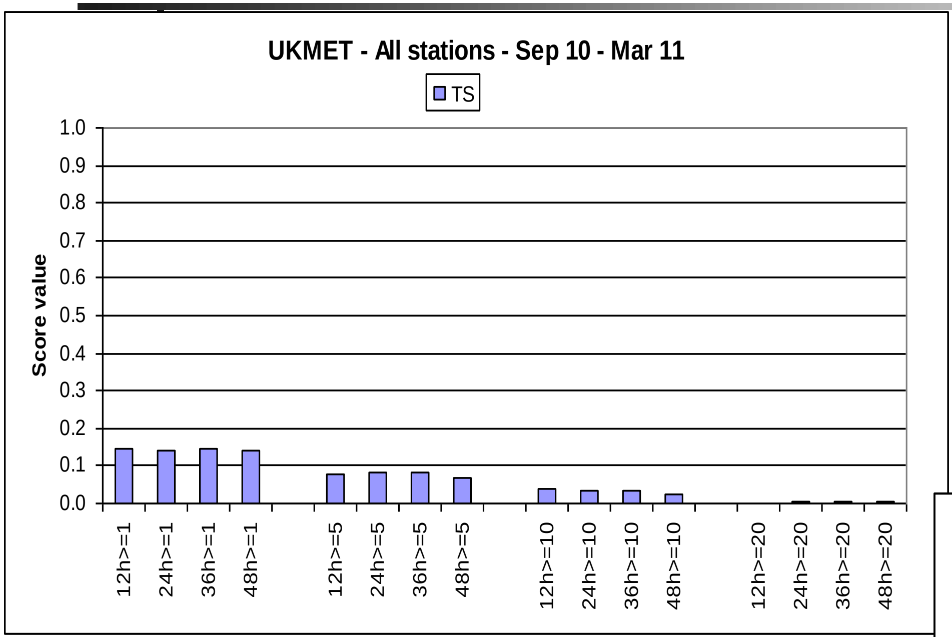


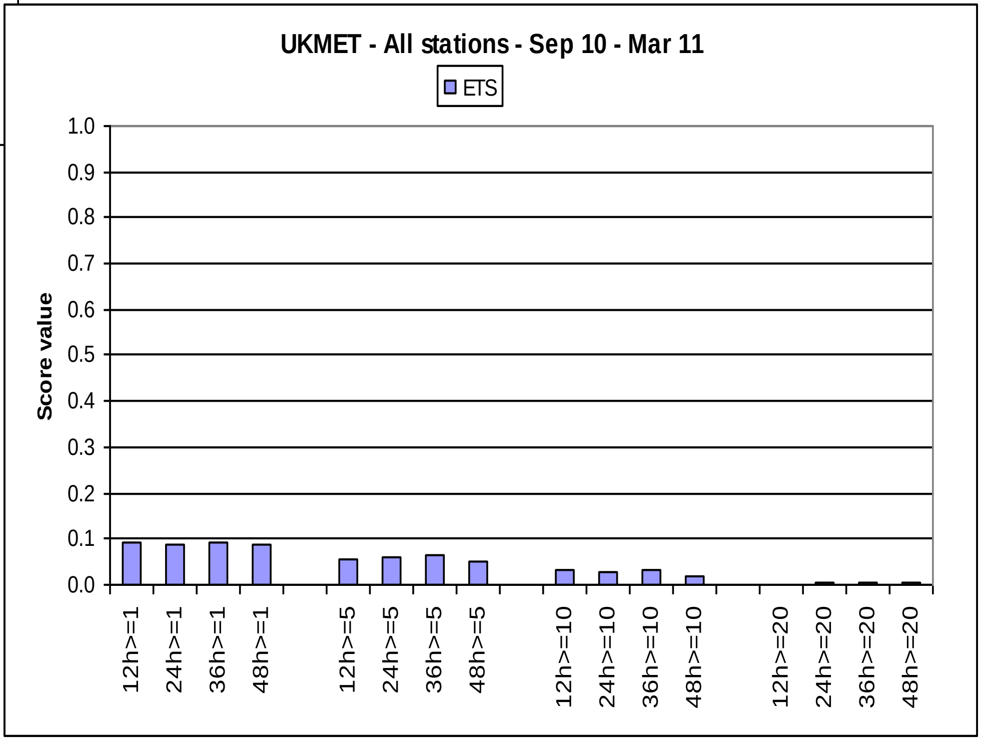


Frequency Bias



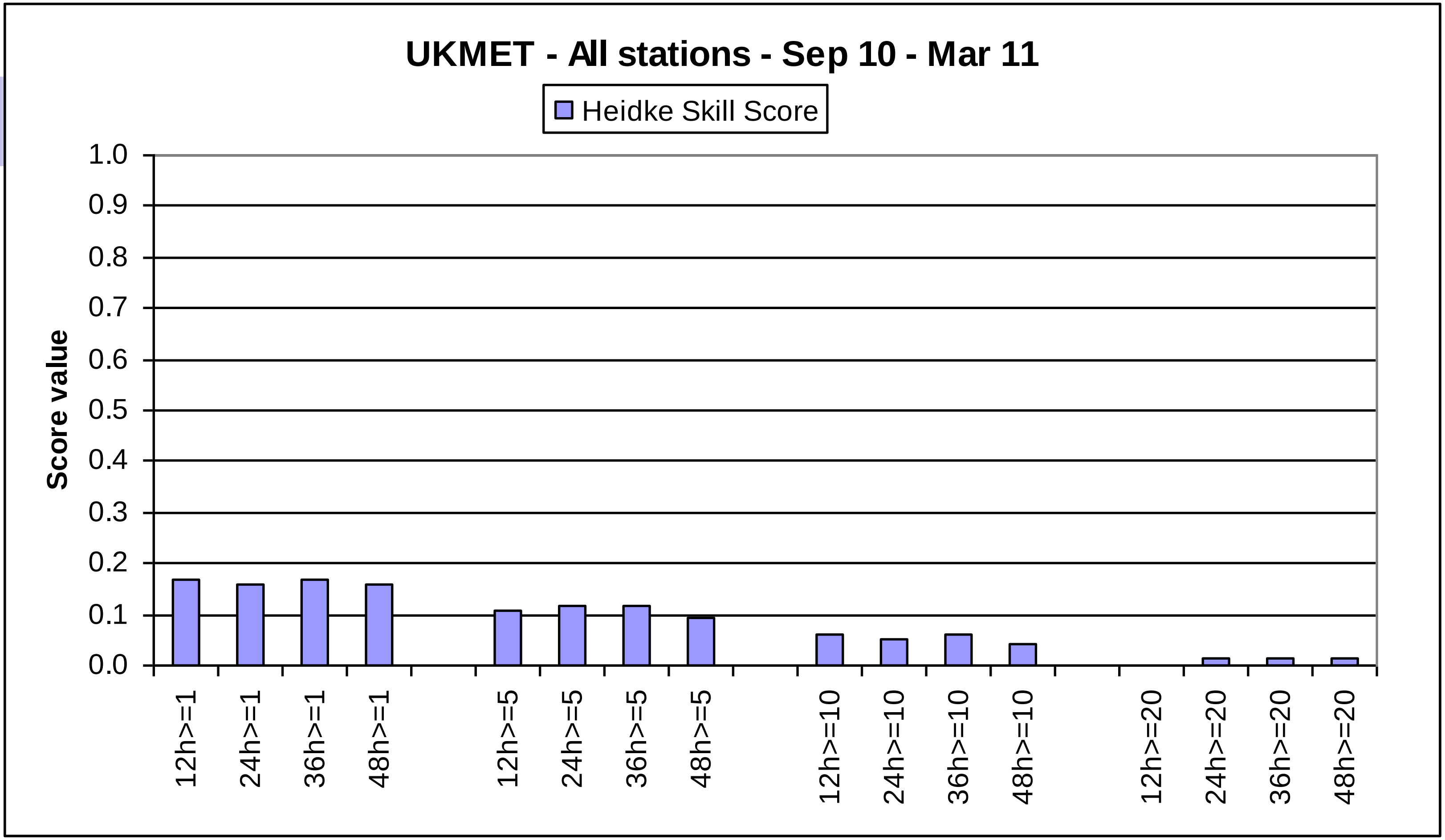
Threat Score (CSI) and Equitable Threat Score





ETS is smaller than TS

Heidke Skill Score



Hanssen-Kuipers (Pierce) Skill score