Guidance of the WMO Commission for Climatology on verification of operational seasonal forecasts

Ernesto Rodríguez Camino

AEMET

(Thanks to S. Mason, C. Coelho, C. Santos, E. Sanchez)

Forecasts possess no intrinsic value. They acquire value through their ability to influence the decisions made by users of the forecasts.

(Murphy 1993)

Guidance on Verification of Operational Seasonal Climate Forecasts

Simon J. Mason
International Research Institute for Climate and Society

Draft: 19 November 2008
Revision: 10 February 2009
Final revision: 13 August 2013

Prepared under the auspices of

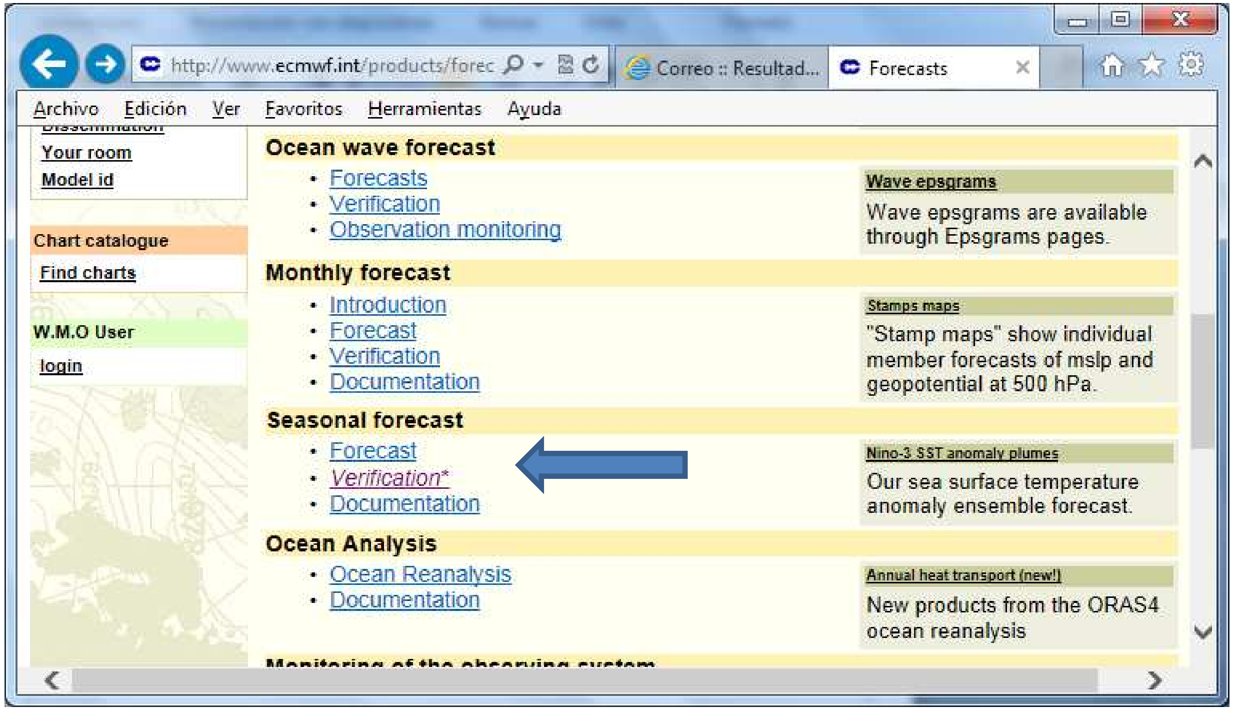
World Meteorological Organization, Commission for Climatology XIV
Expert Team on CLIPS Operations, Verification, and Application Service

Why verify operational seasonal forecasts?

- Does a new system improve the current one?
- Is the cost of the forecast justified?
- Is it a good idea to use (or pay for) the forecast?
- If so, how can they best used?

All operational forecast should be accompanied by readily available information on the quality of forecast (minimum set of diagnostics)

Example: ECMWF



Your room

Model id

Chart catalogue

Find charts

W.M.O User

login

Ocean wave forecast

- Forecasts
- Verification
- Observation monitoring

Wave epsgrams

Wave epsgrams are available through Epsgrams pages.

Monthly forecast

- Introduction
- Forecast
- Verification
- Documentation

Stamps maps

"Stamp maps" show individual member forecasts of mslp and geopotential at 500 hPa.

Seasonal forecast

- Forecast
- *Verification**
- Documentation

Nino-3 SST anomaly plumes

Our sea surface temperature anomaly ensemble forecast.

Ocean Analysis

- Ocean Reanalysis
- Documentation

Annual heat transport (new!)

New products from the ORAS4 ocean reanalysis

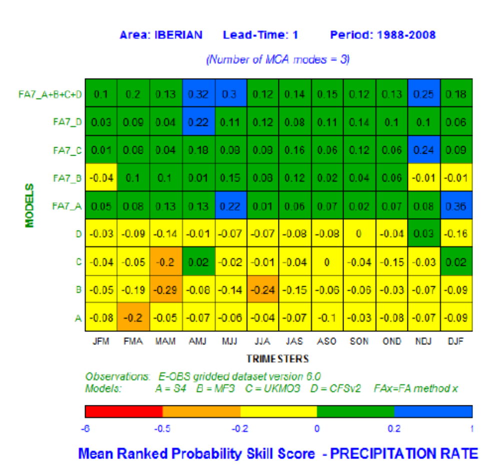
Monitoring of the observing system

Unfortunately, this is not the case for most RCOFs!

Motivation

- There is quite a lot of experience veryfing probabilistic outputs of seasonal models.
- Complement the Standardized Verification System for LRFs (SVSLRF) for GPC products.

Example verification seasonal forecasts from GCMs: RPSS



Area: IBERIAN   Lead-Time: 1   Period: 1988-2008

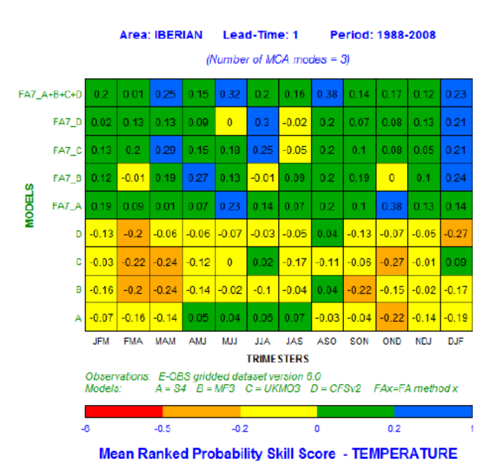
(Number of MCA modes = 3)

MODELS	JFM	FMA	MAM	AMJ	MJJ	JJA	JAS	ASO	SON	OND	NDJ	DJF
FA7_A+B+C+D	0.1	0.2	0.13	0.32	0.3	0.12	0.14	0.15	0.12	0.13	0.25	0.18
FA7_D	0.03	0.09	0.04	0.22	0.11	0.12	0.08	0.11	0.14	0.1	0.1	0.06
FA7_C	0.01	0.08	0.04	0.18	0.08	0.08	0.16	0.06	0.12	0.06	0.24	0.09
FA7_B	-0.04	0.1	0.1	0.01	0.15	0.08	0.12	0.02	0.04	0.06	-0.01	-0.01
FA7_A	0.05	0.08	0.13	0.13	0.22	0.01	0.06	0.07	0.02	0.07	0.08	0.36
D	-0.03	-0.09	-0.14	-0.01	-0.07	-0.07	-0.08	-0.08	0	-0.04	0.03	-0.16
C	-0.04	-0.05	-0.2	0.02	-0.02	-0.01	-0.04	0	-0.04	-0.15	-0.03	0.02
B	-0.05	-0.19	-0.29	-0.08	-0.14	-0.24	-0.15	-0.06	-0.06	-0.03	-0.07	-0.09
A	-0.08	-0.2	-0.05	-0.07	-0.06	-0.04	-0.07	-0.1	-0.03	-0.08	-0.07	-0.09

TRIMESTERS

Observations: E-OBS gridded dataset version 6.0
Models: A = S4 B = MF3 C = UKMO3 D = CFSv2 FAx=FA method x

Mean Ranked Probability Skill Score - PRECIPITATION RATE



Area: IBERIAN   Lead-Time: 1   Period: 1988-2008

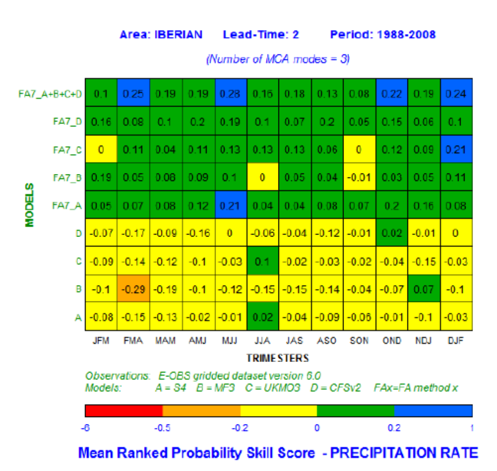
(Number of MCA modes = 3)

MODELS	JFM	FMA	MAM	AMJ	MJJ	JJA	JAS	ASO	SON	OND	NDJ	DJF
FA7_A+B+C+D	0.2	0.01	0.25	0.15	0.32	0.2	0.16	0.38	0.14	0.17	0.12	0.23
FA7_D	0.02	0.13	0.13	0.09	0	0.3	-0.02	0.2	0.07	0.08	0.13	0.21
FA7_C	0.13	0.2	0.29	0.15	0.18	0.25	-0.05	0.2	0.1	0.08	0.05	0.21
FA7_B	0.12	-0.01	0.19	0.27	0.13	-0.01	0.09	0.2	0.19	0	0.1	0.24
FA7_A	0.19	0.09	0.01	0.07	0.23	0.14	0.07	0.2	0.1	0.38	0.13	0.14
D	-0.13	-0.2	-0.06	-0.06	-0.07	-0.03	-0.05	0.04	-0.13	-0.07	-0.05	-0.27
C	-0.03	-0.22	-0.24	-0.12	0	0.02	-0.17	-0.11	-0.06	-0.27	-0.01	0.09
B	-0.16	-0.2	-0.24	-0.14	-0.02	-0.1	-0.04	0.04	-0.22	-0.15	-0.02	-0.17
A	-0.07	-0.16	-0.14	0.05	0.04	0.06	0.07	-0.03	-0.04	-0.22	-0.14	-0.19

TRIMESTERS

Observations: E-OBS gridded dataset version 6.0
Models: A = S4 B = MF3 C = UKMO3 D = CFSv2 FAx=FA method x

Mean Ranked Probability Skill Score - TEMPERATURE



Area: IBERIAN   Lead-Time: 2   Period: 1988-2008

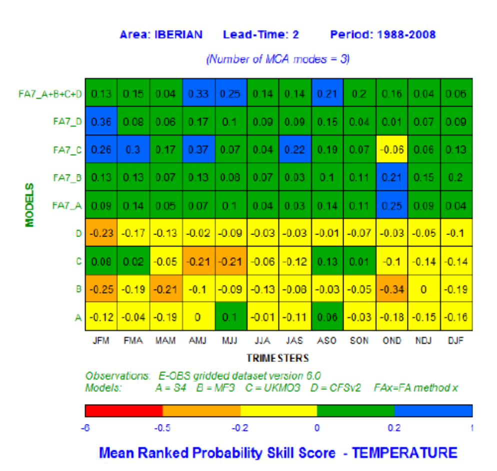
(Number of MCA modes = 3)

MODELS	JFM	FMA	MAM	AMJ	MJJ	JJA	JAS	ASO	SON	OND	NDJ	DJF
FA7_A+B+C+D	0.1	0.25	0.19	0.19	0.28	0.16	0.18	0.13	0.08	0.22	0.19	0.24
FA7_D	0.16	0.08	0.1	0.2	0.19	0.1	0.07	0.2	0.05	0.15	0.06	0.1
FA7_C	0	0.11	0.04	0.11	0.13	0.13	0.13	0.06	0	0.12	0.09	0.21
FA7_B	0.19	0.05	0.08	0.09	0.1	0	0.05	0.04	-0.01	0.03	0.05	0.11
FA7_A	0.05	0.07	0.08	0.12	0.21	0.04	0.04	0.08	0.07	0.2	0.16	0.08
D	-0.07	-0.17	-0.09	-0.16	0	-0.06	-0.04	-0.12	-0.01	0.02	-0.01	0
C	-0.09	-0.14	-0.12	-0.1	-0.03	0.1	-0.02	-0.03	-0.02	-0.04	-0.15	-0.03
B	-0.1	-0.29	-0.19	-0.1	-0.12	-0.15	-0.15	-0.14	-0.04	-0.07	0.07	-0.1
A	-0.08	-0.15	-0.13	-0.02	-0.01	0.02	-0.04	-0.09	-0.06	-0.01	-0.1	-0.03

TRIMESTERS

Observations: E-OBS gridded dataset version 6.0
Models: A = S4 B = MF3 C = UKMO3 D = CFSv2 FAx=FA method x

Mean Ranked Probability Skill Score - PRECIPITATION RATE



Area: IBERIAN   Lead-Time: 2   Period: 1988-2008

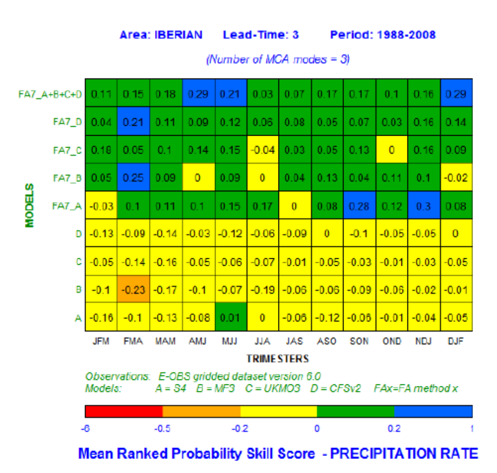
(Number of MCA modes = 3)

MODELS	JFM	FMA	MAM	AMJ	MJJ	JJA	JAS	ASO	SON	OND	NDJ	DJF
FA7_A+B+C+D	0.13	0.15	0.04	0.33	0.25	0.14	0.14	0.21	0.2	0.16	0.04	0.06
FA7_D	0.36	0.08	0.06	0.17	0.1	0.09	0.09	0.15	0.04	0.01	0.07	0.09
FA7_C	0.26	0.3	0.17	0.37	0.07	0.04	0.22	0.19	0.07	-0.06	0.06	0.13
FA7_B	0.13	0.13	0.07	0.13	0.08	0.07	0.03	0.1	0.11	0.21	0.15	0.2
FA7_A	0.09	0.14	0.05	0.07	0.1	0.04	0.03	0.14	0.11	0.25	0.09	0.04
D	-0.23	-0.17	-0.13	-0.02	-0.09	-0.03	-0.03	-0.01	-0.07	-0.03	-0.05	-0.1
C	0.08	0.02	-0.05	-0.21	-0.21	-0.06	-0.12	0.13	0.01	-0.1	-0.14	-0.14
B	-0.25	-0.19	-0.21	-0.1	-0.09	-0.13	-0.08	-0.03	-0.05	-0.34	0	-0.19
A	-0.12	-0.04	-0.19	0	0.1	-0.01	-0.11	0.06	-0.03	-0.18	-0.15	-0.16

TRIMESTERS

Observations: E-OBS gridded dataset version 6.0
Models: A = S4 B = MF3 C = UKMO3 D = CFSv2 FAx=FA method x

Mean Ranked Probability Skill Score - TEMPERATURE



Area: IBERIAN   Lead-Time: 3   Period: 1988-2008

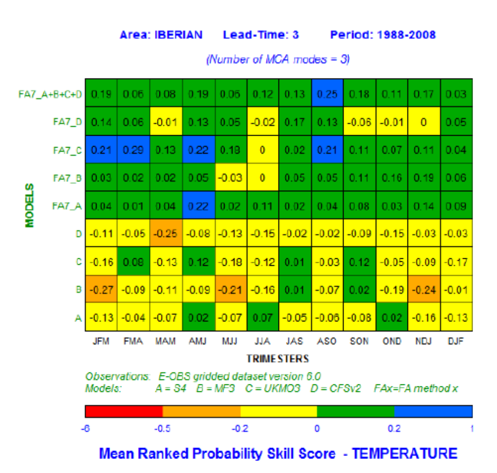
(Number of MCA modes = 3)

MODELS	JFM	FMA	MAM	AMJ	MJJ	JJA	JAS	ASO	SON	OND	NDJ	DJF
FA7_A+B+C+D	0.11	0.15	0.18	0.29	0.21	0.03	0.07	0.17	0.17	0.1	0.16	0.29
FA7_D	0.04	0.21	0.11	0.09	0.12	0.06	0.08	0.05	0.07	0.03	0.16	0.14
FA7_C	0.18	0.05	0.1	0.14	0.15	-0.04	0.03	0.05	0.13	0	0.16	0.09
FA7_B	0.05	0.25	0.09	0	0.09	0	0.04	0.13	0.04	0.11	0.1	-0.02
FA7_A	-0.03	0.1	0.11	0.1	0.15	0.17	0	0.08	0.28	0.12	0.3	0.08
D	-0.13	-0.09	-0.14	-0.03	-0.12	-0.06	-0.09	0	-0.1	-0.05	-0.05	0
C	-0.05	-0.14	-0.16	-0.05	-0.06	-0.07	-0.01	-0.05	-0.03	-0.01	-0.03	-0.05
B	-0.1	-0.23	-0.17	-0.1	-0.07	-0.19	-0.06	-0.06	-0.09	-0.06	-0.02	-0.01
A	-0.16	-0.1	-0.13	-0.08	0.01	0	-0.06	-0.12	-0.06	-0.01	-0.04	-0.05

TRIMESTERS

Observations: E-OBS gridded dataset version 6.0
Models: A = S4 B = MF3 C = UKMO3 D = CFSv2 FAx=FA method x

Mean Ranked Probability Skill Score - PRECIPITATION RATE



Area: IBERIAN   Lead-Time: 3   Period: 1988-2008

(Number of MCA modes = 3)

MODELS	JFM	FMA	MAM	AMJ	MJJ	JJA	JAS	ASO	SON	OND	NDJ	DJF
FA7_A+B+C+D	0.19	0.06	0.08	0.19	0.06	0.12	0.13	0.25	0.18	0.11	0.17	0.03
FA7_D	0.14	0.06	-0.01	0.13	0.05	-0.02	0.17	0.13	-0.06	-0.01	0	0.05
FA7_C	0.21	0.29	0.13	0.22	0.18	0	0.02	0.21	0.11	0.07	0.11	0.04
FA7_B	0.03	0.02	0.02	0.05	-0.03	0	0.05	0.05	0.11	0.16	0.19	0.06
FA7_A	0.04	0.01	0.04	0.22	0.02	0.11	0.02	0.04	0.08	0.03	0.14	0.09
D	-0.11	-0.05	-0.25	-0.08	-0.13	-0.15	-0.02	-0.02	-0.09	-0.15	-0.03	-0.03
C	-0.16	0.08	-0.13	0.12	-0.18	-0.12	0.01	-0.03	0.12	-0.05	-0.09	-0.17
B	-0.27	-0.09	-0.11	-0.09	-0.21	-0.16	0.01	-0.07	0.02	-0.19	-0.24	-0.01
A	-0.13	-0.04	-0.07	0.02	-0.07	0.07	-0.05	-0.06	-0.08	0.02	-0.16	-0.13

TRIMESTERS

Observations: E-OBS gridded dataset version 6.0
Models: A = S4 B = MF3 C = UKMO3 D = CFSv2 FAx=FA method x

Mean Ranked Probability Skill Score - TEMPERATURE

Motivation

- There is quite a lot of experience veryfing probabilistic outputs of seasonal models.
- Complement the Standardized Verification System for LRFs (SVSLRF) for GPC products.
- So far most RCOFs are limited their verification to qualitative procedures → need move towards use of objective scores!!
- There are no formal WMO verification procedures, but some guidance on procedures is being published by WMO CCl
- Focus on how well forecasts correspond with observations (quality), and also on attributes making forecasts potentially useful (value).
- Small sample sizes (few years, few stations) typical of seasonal forecasts → large sampling errors

What is a good forecast? (Murphy 1993)

3 types of goodness:

- CONSISTENCY → true indication of what the forecaster thinks is going to happen
- QUALITY → how well what was forecast corresponds with what happened
- VALUE/UTILITY → “value” economic, or social, or otherwise.

Probabilistic forecasts and forecast quality

- A forecaster says there is a 100% chance of rain tomorrow → It rains → Very good forecast!
- A forecaster says there is a 80% chance of rain tomorrow→ It rains → ?
- A forecaster says there is a 50% chance of rain tomorrow→ It rains → ?
- A forecaster says there is a 10% chance of rain tomorrow→ It rains → ?

How good are the different forecast?

How good are the different forecast?

- One reasonably common practice is to define probabilistic forecasts as “correct” if the category with the highest probability verified.
- Most RCOFs verify qualitatively in this way
- Forecasters typically become tempted to hedge towards issuing higher probabilities on the normal category to avoid a two category “error” → Scoring strategy is an issue!!

Verification procedures suitable for the forecasts in the format in which they are presented.

- If forecasts are delivered in form of tercile-based categories → Verification should fit to it!

Attributes of “good” probabilistic forecasts

(Murphy 1993)

- Resolution

  Does the outcome change when the forecast changes? OUTCOME CONDITIONED BY FORECAST

  Example: does above-normal rainfall become more frequent when its probability increases?

- Discrimination

  Does the forecast differ when the outcome differs? FORECAST CONDITIONED BY OUTCOME

  Example: is the probability on above-normal rainfall higher when above-normal rainfall occurs?

- Reliability

  if observation falls in the category as FREQUENTLY as the forecast implies

- Sharpness

  Probabilities differing MARKEDLY from the climatology

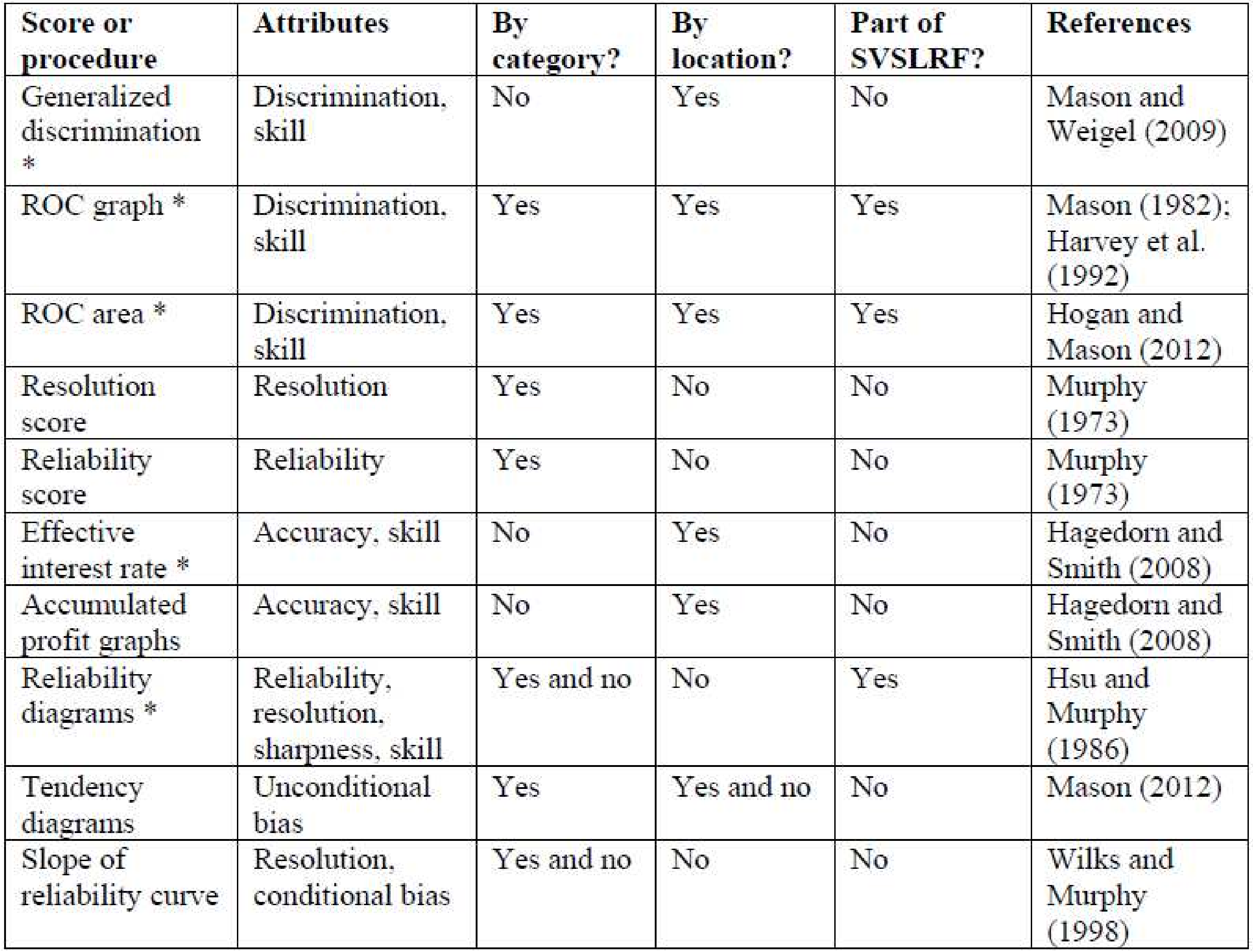
- Skill

  It COMPARES two forecasts with some metric

From EUMETCAL(http://www.eumetcal.org)

	High reliability	High resolution	High Sharpness	Discriminatory	High Skill
The forecaster predicts the long term climatological frequency on each occasion	✓	✗	✗	✗	✗
The forecaster predicts categorically, that is, he assigns a forecast of 100% to the category he thinks is most likely, and 0 to the other.	✗	✗	✓	✗	✗
The forecaster manages to forecast 45% probability when the event does not occur and 55% when it does.	✗	✓	✗	✓	✗
A forecaster who is sure, but never absolutely certain, forecasting 80% when he thinks rain will occur and 20% when he thinks it won’t.	✗	✗	✓	✗	✗
The forecaster sits back with a smile on his face: He went out on a limb and predicted 90% probability of rain in his dry climate where it normally rains on only 10% of the days. And sure enough, it rained.	✗	✗	✗	✗	✓

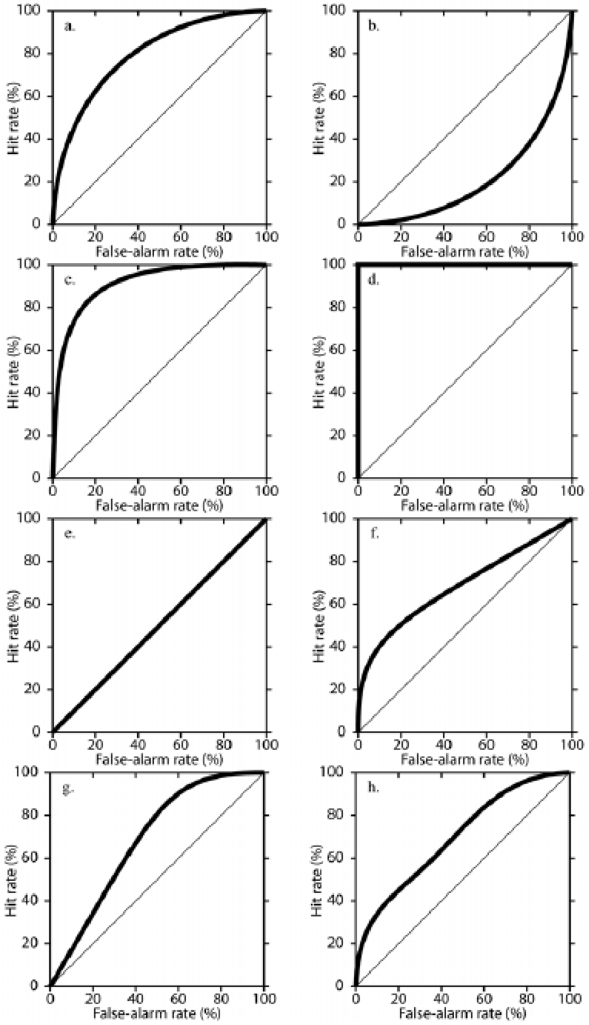
Recommended scores/procedures for series of forecasts



Score or procedure	Attributes	By category?	By location?	Part of SVSLRF?	References
Generalized discrimination *	Discrimination, skill	No	Yes	No	Mason and Weigel (2009)
ROC graph *	Discrimination, skill	Yes	Yes	Yes	Mason (1982); Harvey et al. (1992)
ROC area *	Discrimination, skill	Yes	Yes	Yes	Hogan and Mason (2012)
Resolution score	Resolution	Yes	No	No	Murphy (1973)
Reliability score	Reliability	Yes	No	No	Murphy (1973)
Effective interest rate *	Accuracy, skill	No	Yes	No	Hagedorn and Smith (2008)
Accumulated profit graphs	Accuracy, skill	No	Yes	No	Hagedorn and Smith (2008)
Reliability diagrams *	Reliability, resolution, sharpness, skill	Yes and no	No	Yes	Hsu and Murphy (1986)
Tendency diagrams	Unconditional bias	Yes	Yes and no	No	Mason (2012)
Slope of reliability curve	Resolution, conditional bias	Yes and no	No	No	Wilks and Murphy (1998)

(*) Minimum set for an operational centre

ROC curves: idealized examples



(a) good discrimination and good skill

(b) good discrimination but bad skill

(*c*) excellent discrimination

(*d*) good discrimination

(*e*) no discrimination

(*f*) good discrimination for high probability forecasts

(*g*) good discrimination for low probability forecasts

(*h*) good discrimination for confident (high and low probability) forecasts.

Simple realistic example

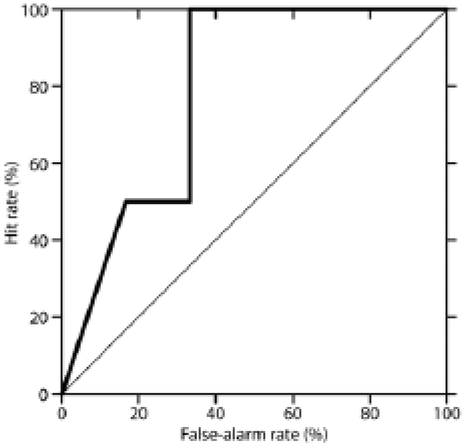
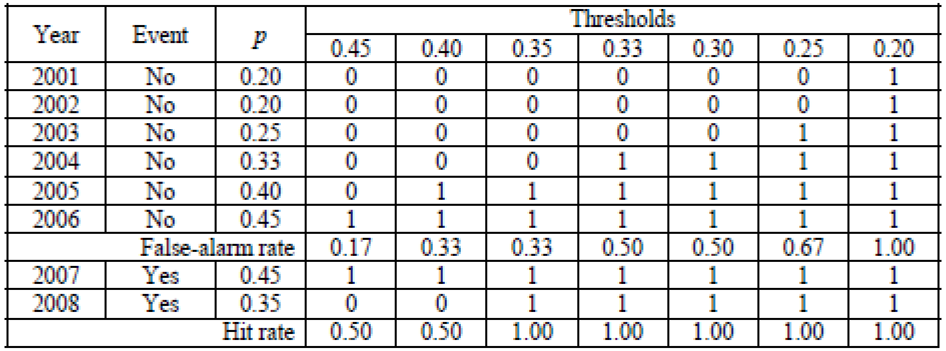


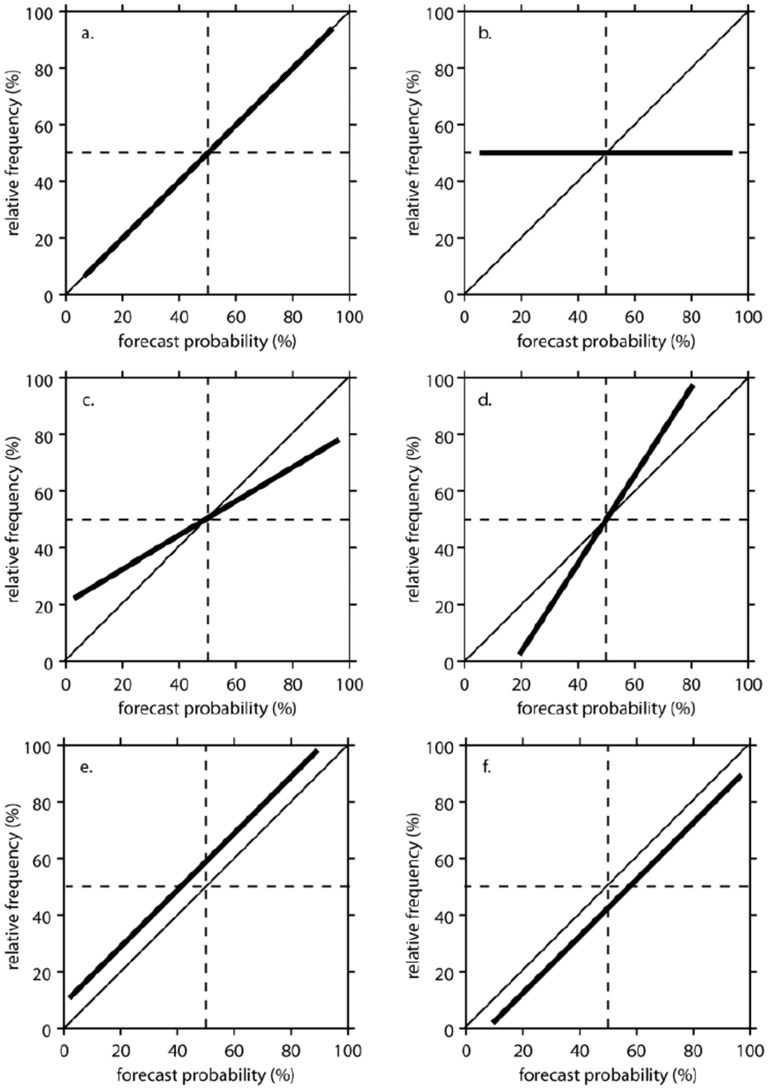
Table B.5a. Example calculation of the hit and false-alarm rates for the ROC graph.



Year	Event	*p*	Thresholds						
			0.45	0.40	0.35	0.33	0.30	0.25	0.20
2001	No	0.20	0	0	0	0	0	0	1
2002	No	0.20	0	0	0	0	0	0	1
2003	No	0.25	0	0	0	0	0	1	1
2004	No	0.33	0	0	0	1	1	1	1
2005	No	0.40	0	1	1	1	1	1	1
2006	No	0.45	1	1	1	1	1	1	1
False-alarm rate			0.17	0.33	0.33	0.50	0.50	0.67	1.00
2007	Yes	0.45	1	1	1	1	1	1	1
2008	Yes	0.35	0	0	1	1	1	1	1
Hit rate			0.50	0.50	1.00	1.00	1.00	1.00	1.00

Reliability diagrams:

observed relative freq. vs forecasted relative freq.



(*a*) perfect reliability,

(*b*) no resolution

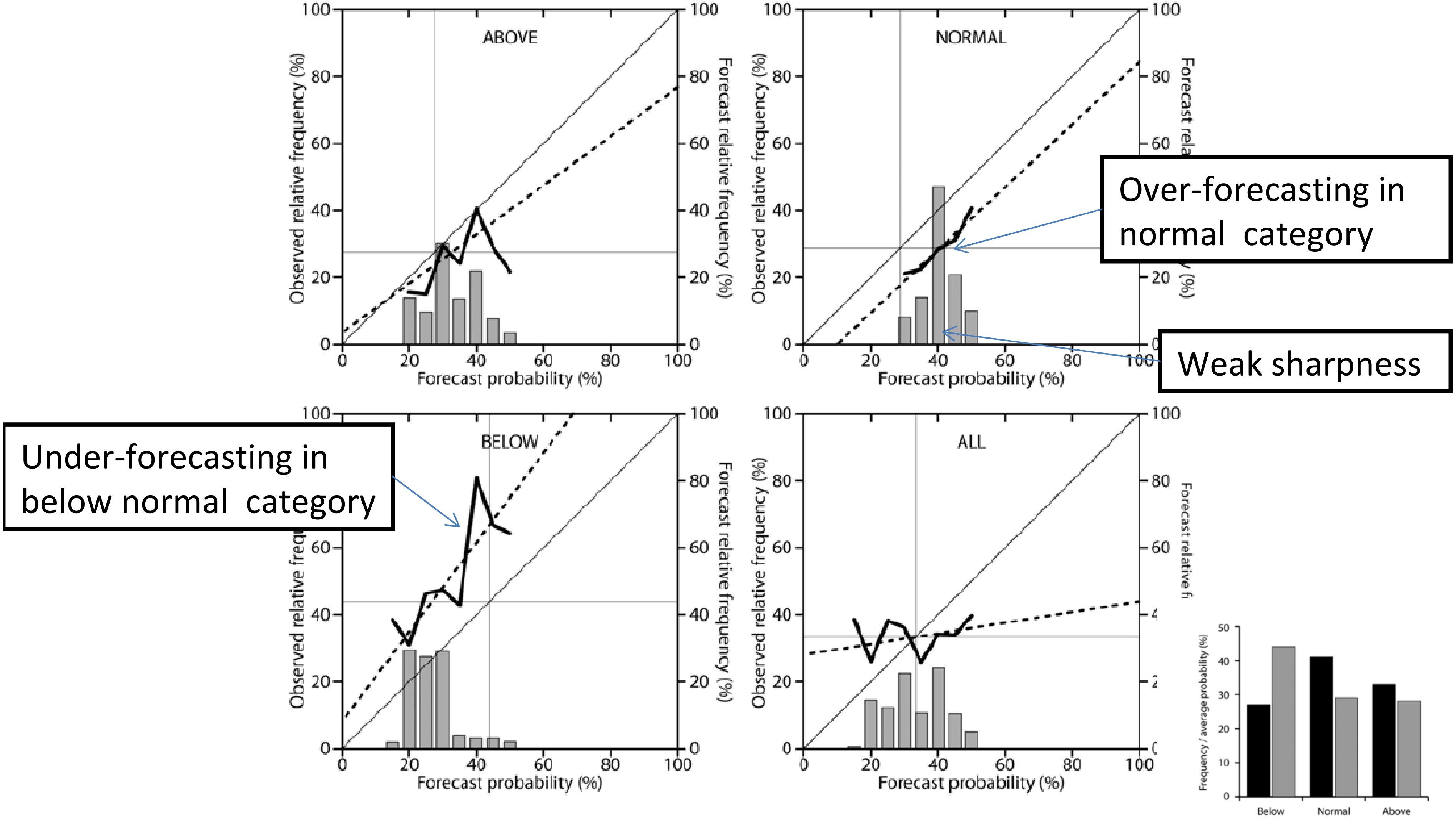
(*c*) over-confidence

(*d*) under-confidence

(*e*) under-forecasting

(*f*) overforecasting

Reliability diagrams for the first 10 years of PRESAO (seasonal rainfall forecasts Jul-Sept)

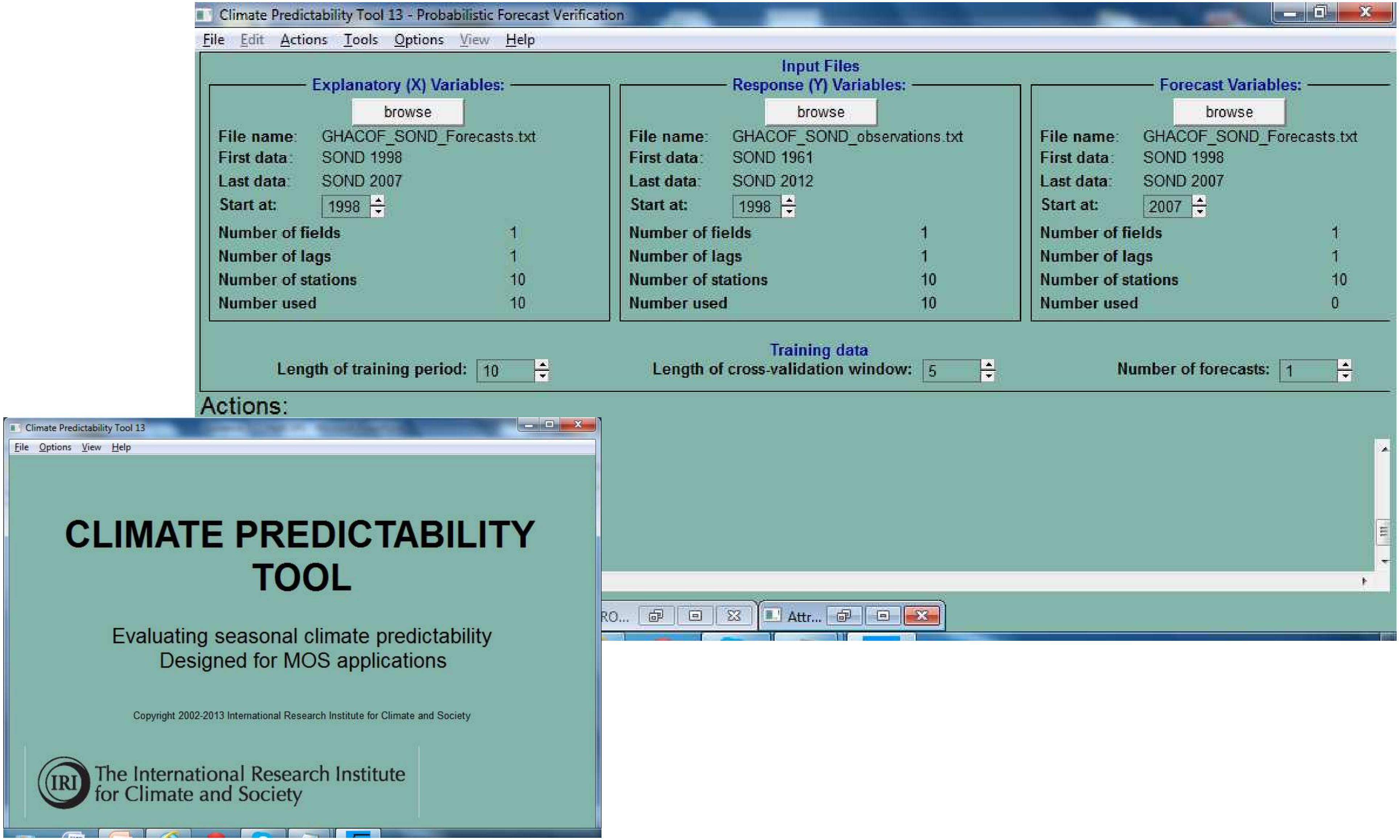


Over-forecasting in normal category

Weak sharpness

Under-forecasting in below normal category

Verification with CPT



Climate Predictability Tool 13 - Probabilistic Forecast Verification
File Edit Actions Tools Options View Help
browse
Actions:
RO...
Attr...
Climate Predictability Tool 13
File Options View Help

Input Files

Explanatory (X) Variables:		Response (Y) Variables:		Forecast Variables:	
File name:	GHACOF_SOND_Forecasts.txt	File name:	GHACOF_SOND_observations.txt	File name:	GHACOF_SOND_Forecasts.txt
First data:	SOND 1998	First data:	SOND 1961	First data:	SOND 1998
Last data:	SOND 2007	Last data:	SOND 2012	Last data:	SOND 2007
Start at:	1998	Start at:	1998	Start at:	2007
Number of fields	1	Number of fields	1	Number of fields	1
Number of lags	1	Number of lags	1	Number of lags	1
Number of stations	10	Number of stations	10	Number of stations	10
Number used	10	Number used	10	Number used	0

Training data

Length of training period: 10

Length of cross-validation window: 5

Number of forecasts: 1

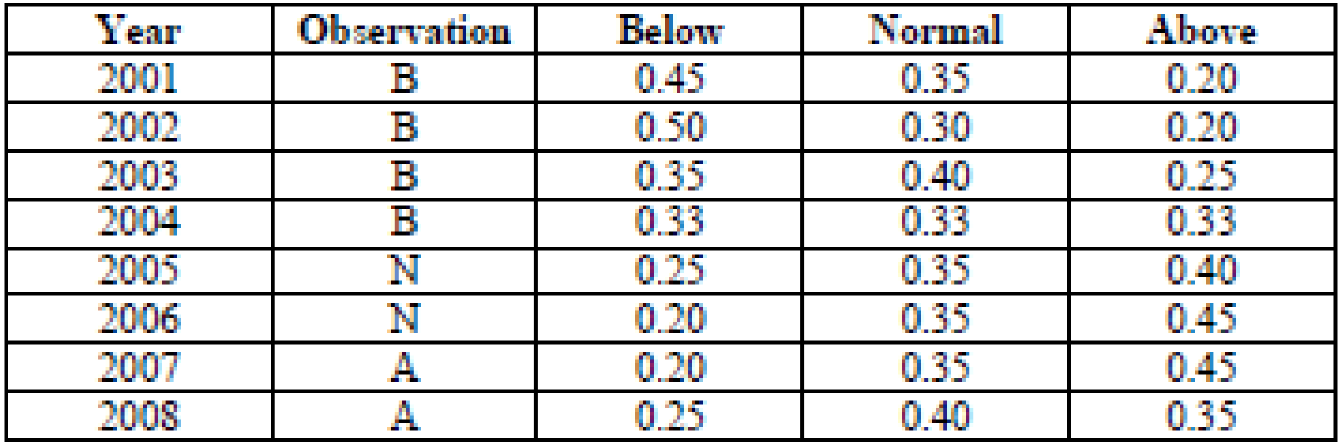
CLIMATE PREDICTABILITY TOOL

Evaluating seasonal climate predictability
Designed for MOS applications

Copyright 2002-2013 International Research Institute for Climate and Society

The International Research Institute for Climate and Society

Verification of tercile-based forecasts only requires information of the obs. category → problems related data policy circumvected



Year	Observation	Below	Normal	Above
2001	B	0.45	0.35	0.20
2002	B	0.50	0.30	0.20
2003	B	0.35	0.40	0.25
2004	B	0.33	0.33	0.33
2005	N	0.25	0.35	0.40
2006	N	0.20	0.35	0.45
2007	A	0.20	0.35	0.45
2008	A	0.25	0.40	0.35

Scores to verify the consensus forecasts

and

scores to improve the consensus process

Reference climatology is relevant!

- Paco’s tranparency!!
- Tercile-based seasonal forecasts refered to a climatology
- Climatologist → long reference periods (30 y)
- Users → short (10 y) recient periods

Recommendations

- Assess the degree to which forecasts are being hedged on normal → Eliminate, or at least reduce, the hedging:
  - Use "proper" scoring procedures
  - Review procedures for setting probabilities
- Agree upon a minimum set of verification procedures for RCOF products.
- Encourage greater standardization in forecast production

Proposal

- Start with a minimum verification package (following WMO-CCI guidelines) verifying consensus forecast (tercile-based) produced so far by SEECOF and PRESANORD
- Use initially ECA&D data from a set of selected stations and tercile-based obs. (A, N, B)
- Agree on a reference period to establish our tercile values
- Report on MedCOF-2

THANK YOU FOR YOUR ATTENTION!

and

discussion on RCOF verification
to be continued!!!

Discrimination

Perfect

2003	70%	T
2004	60%	T
2005	30%	F
2006	40%	T
2007	20%	F
2008	10%	F
2009	35%	T
2010	50%	T
2011	25%	F
2012	10%	F

If prob>35%
always T

Very bad

2003	70%	F
2004	60%	T
2005	30%	T
2006	40%	T
2007	20%	F
2008	10%	T
2009	35%	T
2010	50%	F
2011	25%	F
2012	10%	T