Verification of Sub-seasonal to Seasonal Predictions

Caio Coelho
INPE/CPTEC, Brazil
caio.coelho@cptec.inpe.br

Acknowledgments: Arun Kumar, Alberto Arribas, Barbara Brown, Beth Ebert, David Stephenson, Debbie Hudson, Laura Ferranti, Matthew Wheeler, Simon Mason and Yuhei Takaya

Plan of talk

1) Introduction to forecast goodness
2) Attributes based forecast quality assessment: definitions and examples of S2S practice
3) JWGFVR: Aims and opportunities for S2S forecast verification
4) Final remarks

International Conference on Sub-seasonal to Seasonal Prediction
WMO WWRP/THORPEX-WCRP joint S2S research project
NOAA Center for Weather and Climate Prediction
10-13 February 2014, College Park, MD, USA

What is a good forecast?

Good forecasts have:

- QUALITY
- VALUE/UTILITY
- CONSISTENCY

Attributes of quality:

- Association
- Accuracy
- Discrimination
- Reliability
- Resolution

...

→ No single score can be used to summarize a set of forecasts

A. H. Murphy 1993
"What is a good forecast ?
An essay on the nature of goodness in weather forecasting"
Weather and Forecasting, 8, 281-293.

Some definitions

- Quality: Measure of correspondence between forecasts and observations using mathematical relationship (deterministic and probabilistic scores)

- Value: Measure of benefit achieved (or loss incurred) through the use of forecasts

- Consistency: Correspondence between a forecast and the forecasters belief. If consistent, the forecast must communicate what the forecaster thinks will happen, and correctly indicate the associated level of uncertainty

S2S forecast quality assessment

1. Attributes of deterministic forecasts (ensemble mean)

Association

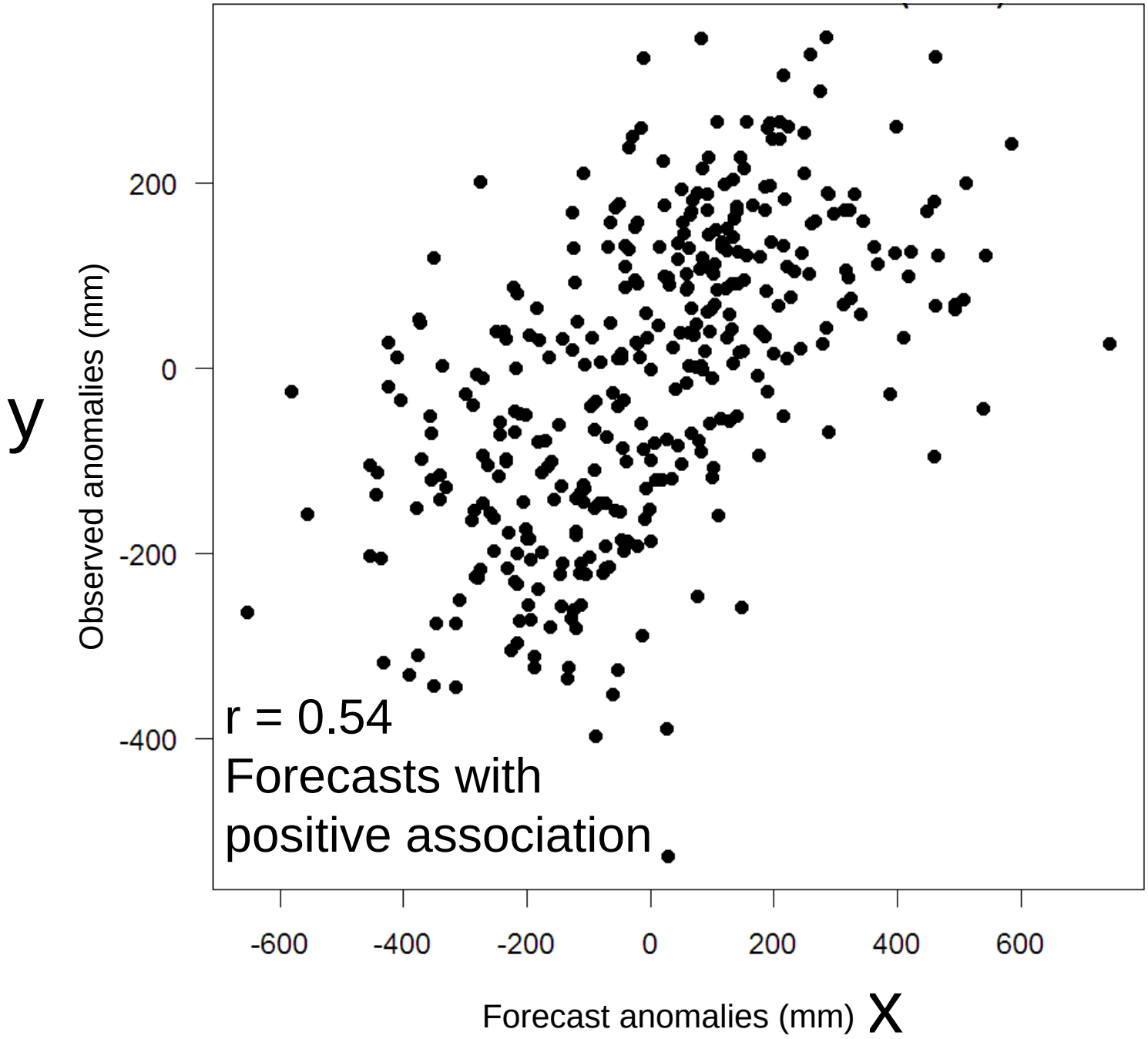
- Overall strength of the relationship between the forecasts and observations
- Linear association is often measured using the product moment **correlation coefficient**

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

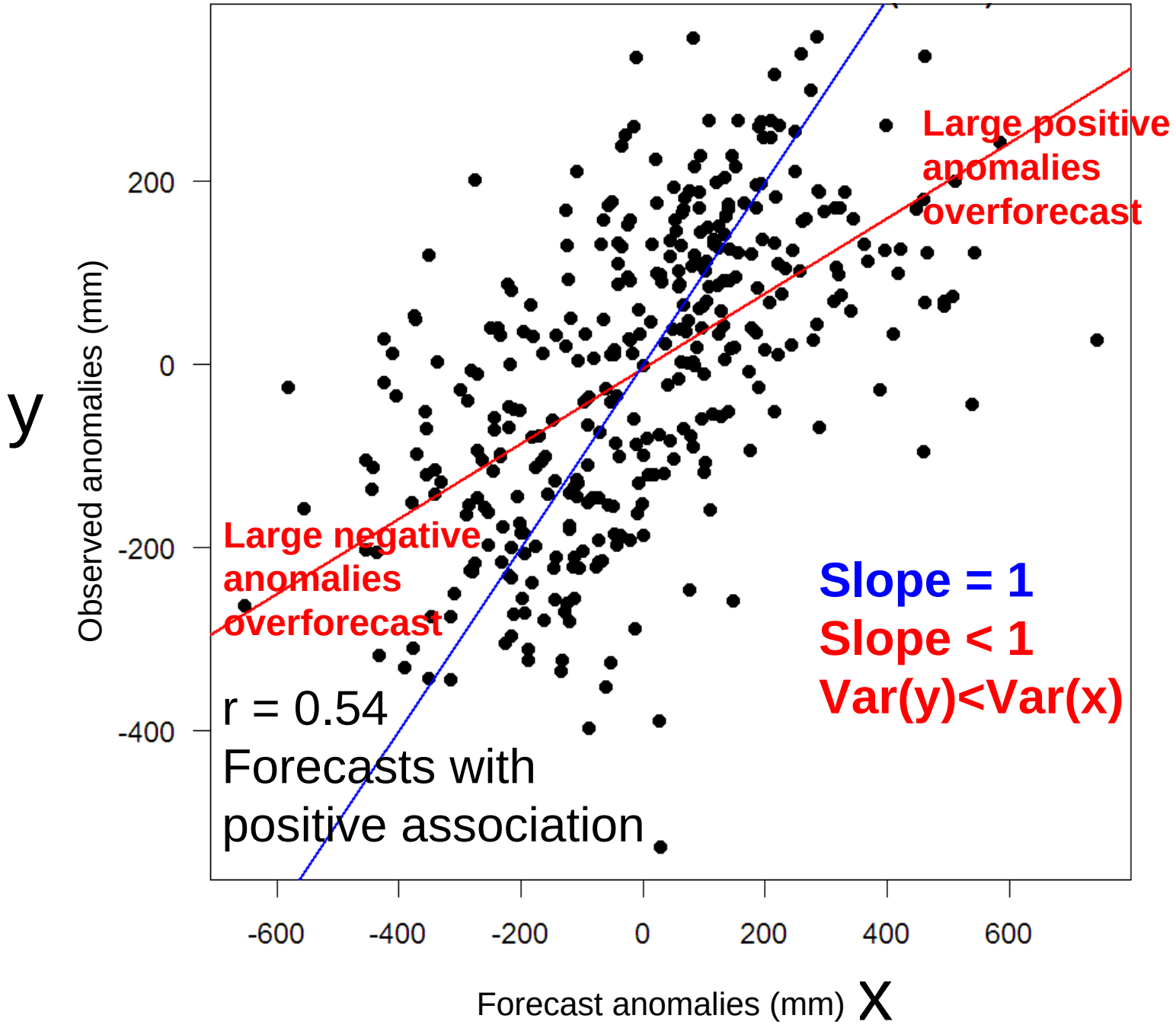
x: forecast *y: observation*

n: number of (x,y) pairs

Relationship between past forecast and past obs. anomalies



Relationship between past forecast and past obs. anomalies



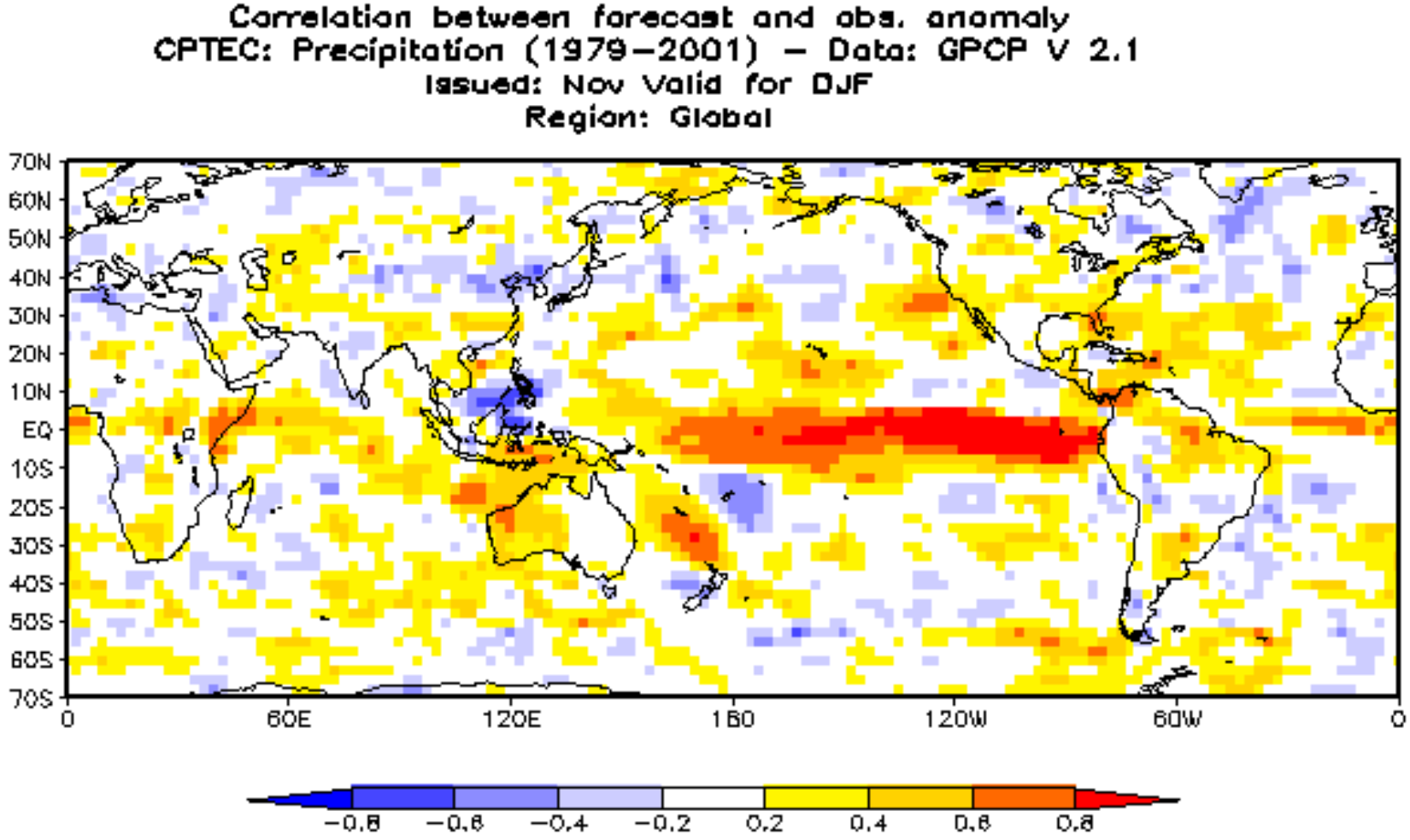
Accuracy

- Average distance between forecasts and observations
- Simplest measure is the **Mean Error (Bias)**

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)$$

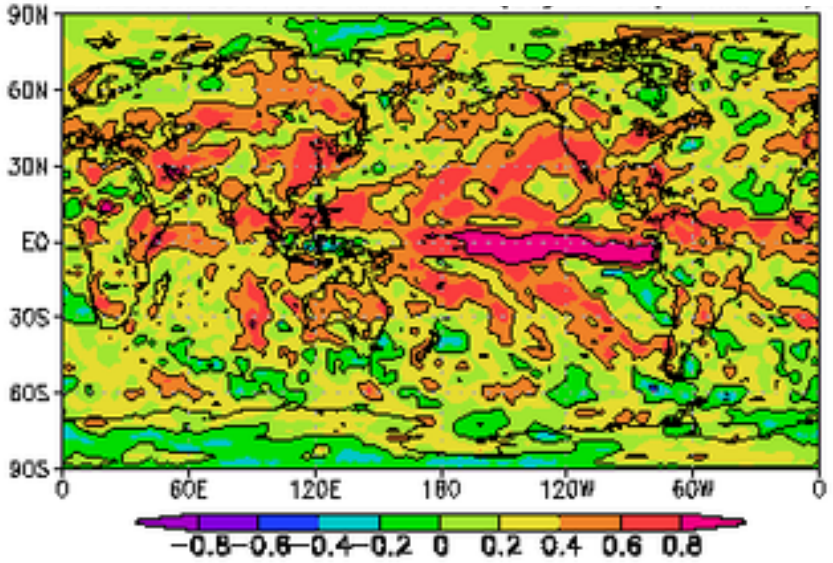
x: forecast y: observation n: number of (x,y) pairs

Seasonal forecast example: 1-month lead precip. fcsts for DJF

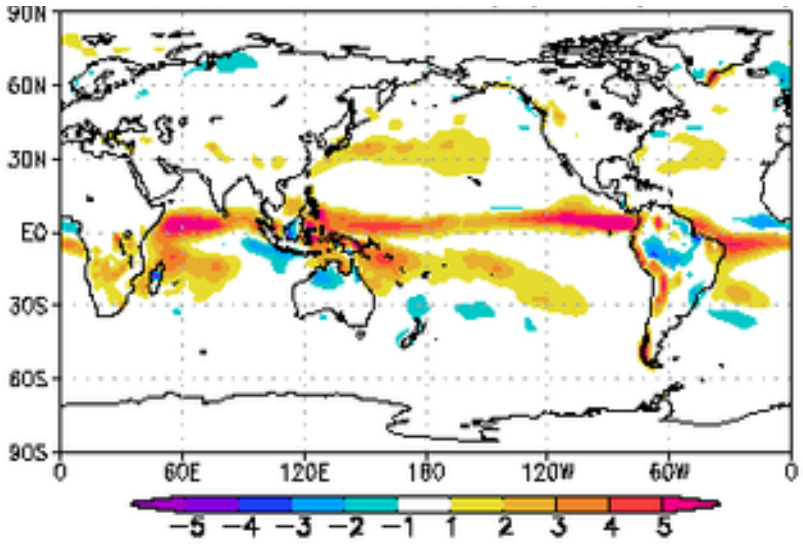


Monthly forecast example:
0-day lead precip. fcsts for next 30 days

ACC (against GPCP v2 monthly)
Day 1-30 mean
I.C. : Dec.-Feb. 1981-2010



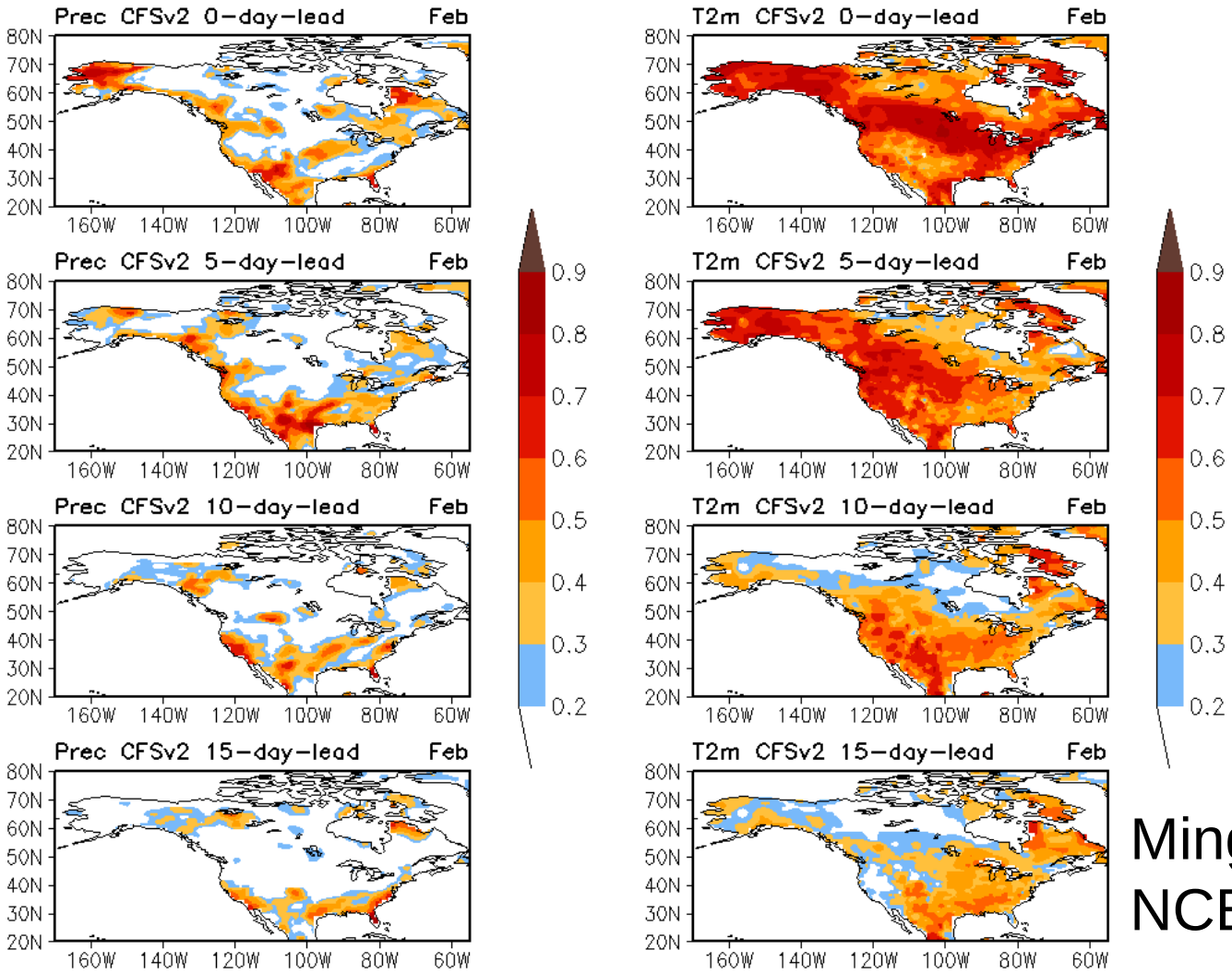
Bias (against GPCP v2 monthly)
Day 1-30 mean
I.C. : Dec.-Feb. 1981-2010



Yuhei Takaya, JMA

Monthly forecast example: 0, 5, 10 and 15-day lead fcsts for Feb

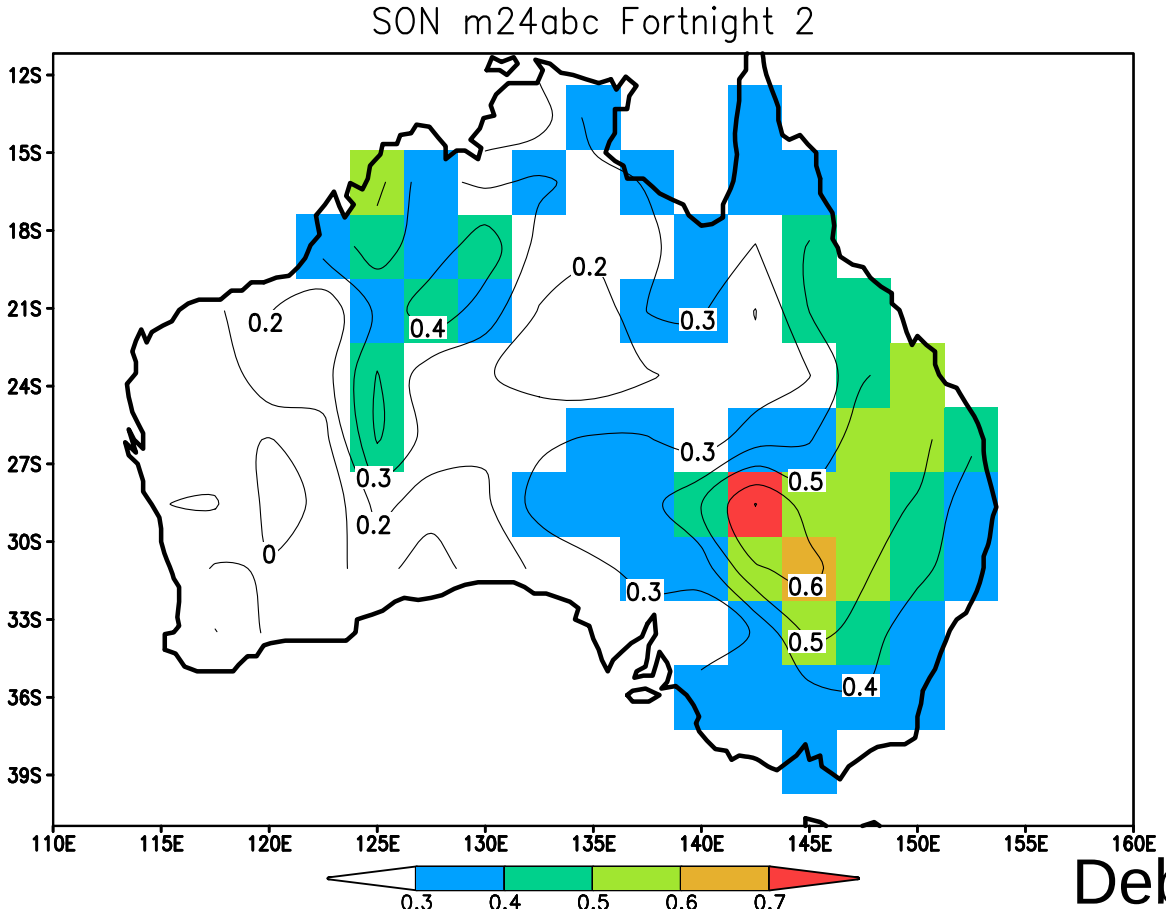
Precipitation CFSv2 Correlation Feb (1982–2009) **2m Temperature**



Mingyue Chen
NCEP/NOAA

Two weeks forecast example: ½ month lead precip. fcsts

Correlation between forecast and observed precipitation anomalies
Fortnight 2: Sep, Oct, Nov forecast start months. Hindcasts: 1980-2006



Debbie Hudson
BOM, Australia

S2S forecast quality assessment

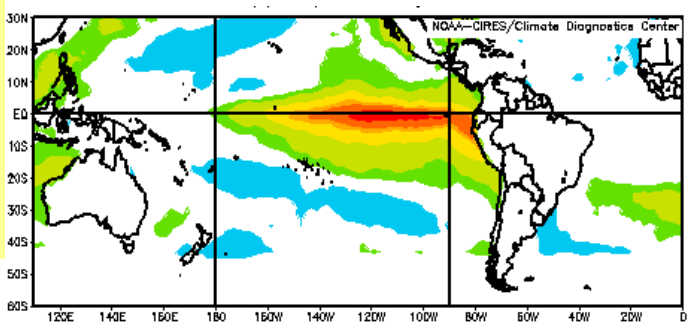
2. Attributes of probabilistic forecasts (derived from ensemble members)

Discrimination

- Conditioning of forecasts on observed outcomes
- Addresses the question: Does the forecast differ given different observed outcomes? Or, can the forecasts distinguish an event from a non-event?
- If the forecast is the same regardless of the outcome, the forecasts cannot discriminate an *event* from a *non-event*
- Forecasts with no discrimination ability are useless because the forecasts are the same regardless of what happens

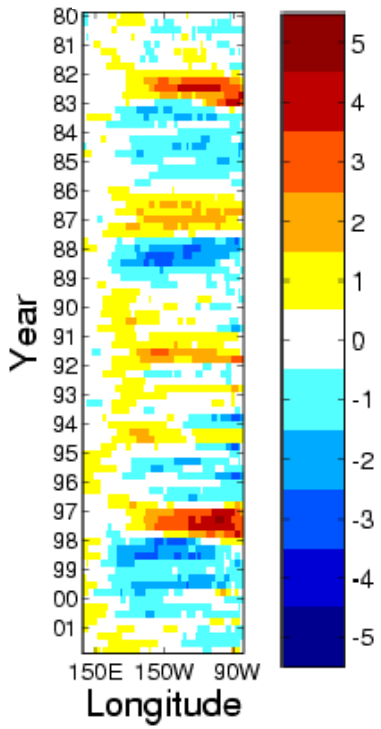
Example:Equatorial Pacific SST

88 seasonal probability forecasts of binary SST anomalies at 56 grid points along the equatorial Pacific. Total of 4928 forecasts.
6-month lead forecasts for 4 start dates (Feb,May,Aug,Nov) valid for (Jul,Oct,Jan,Apr)

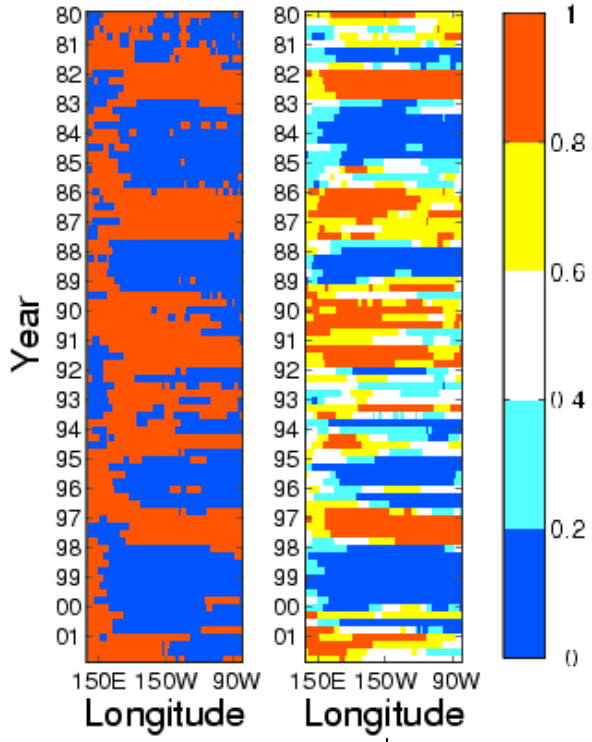


SST *Event:* $o = (SST > 0)$ $f = \Pr(\hat{o})$

OBS



OBS ENS

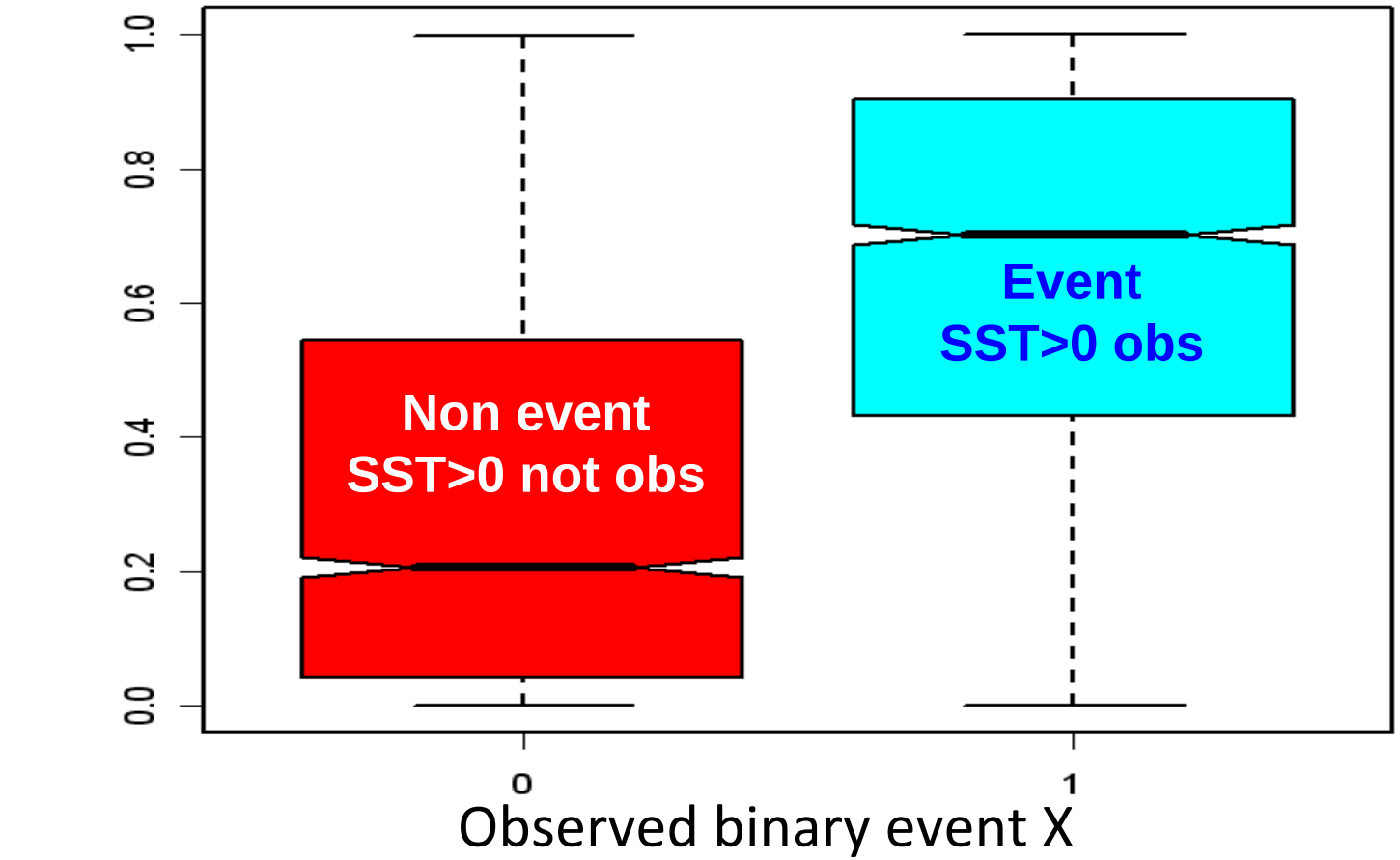


The probability forecasts were constructed by fitting Normal distributions to the ensemble mean forecasts from the 7 DEMETER coupled models, and then calculating the area under the normal density for SST anomalies greater than zero

SST anomalies (°C)

Forecast probabilities: f

Prob. forecasts conditioned/stratified on observations



→ Forecasts do differ given different outcomes

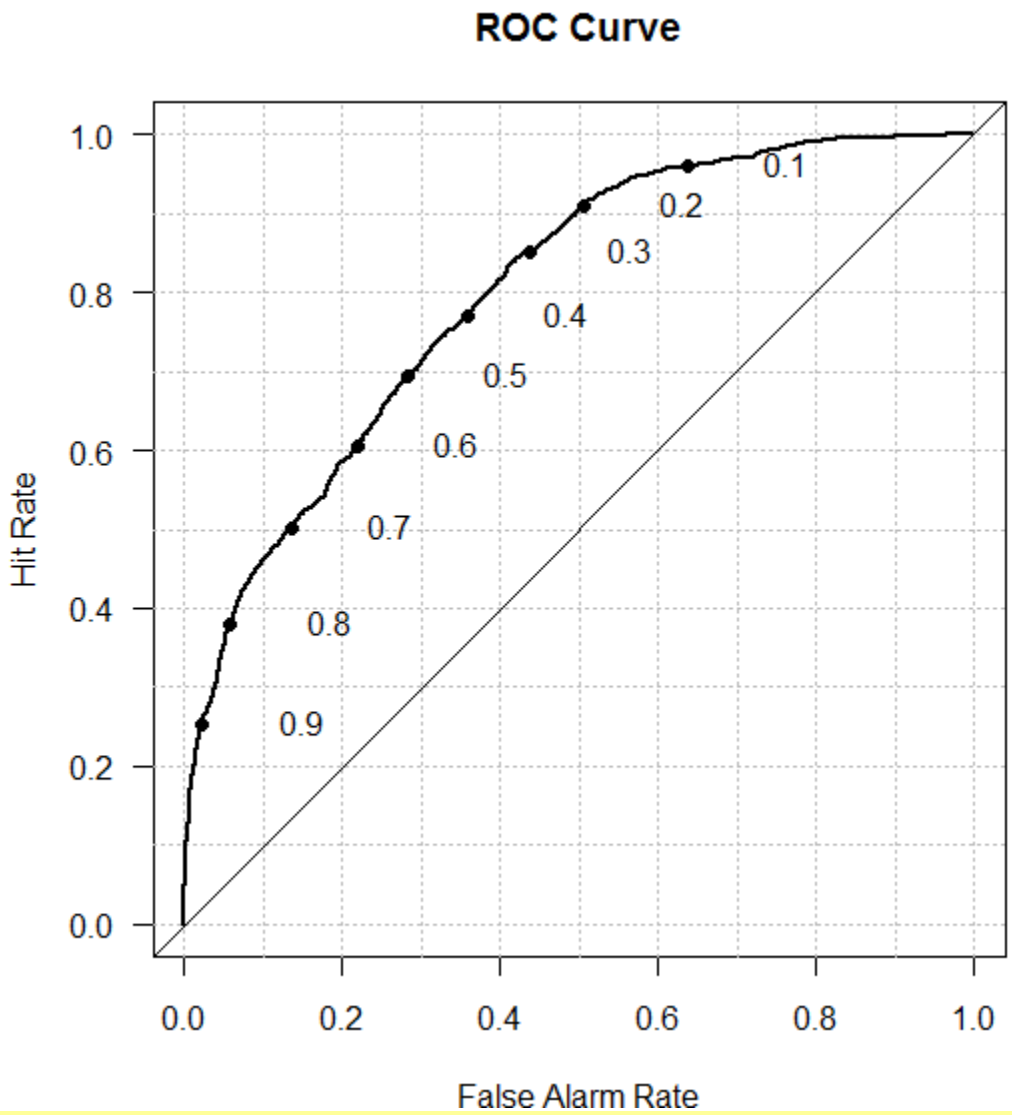
→ Forecast system has discrimination (distinguish event from non-event)

ROC: Relative operating characteristics

Measures discrimination (ability of forecasting system to detect the event of interest)

Forecast	Observed		
	Yes	No	Total
Yes	a (Hit)	b (False alarm)	a+b (Fcst)
No	c (Miss)	d (Correct rejection)	c+d (Not Fcst)
Total	a+c (Obs)	b+d (Not Obs)	a+b+c+d=n

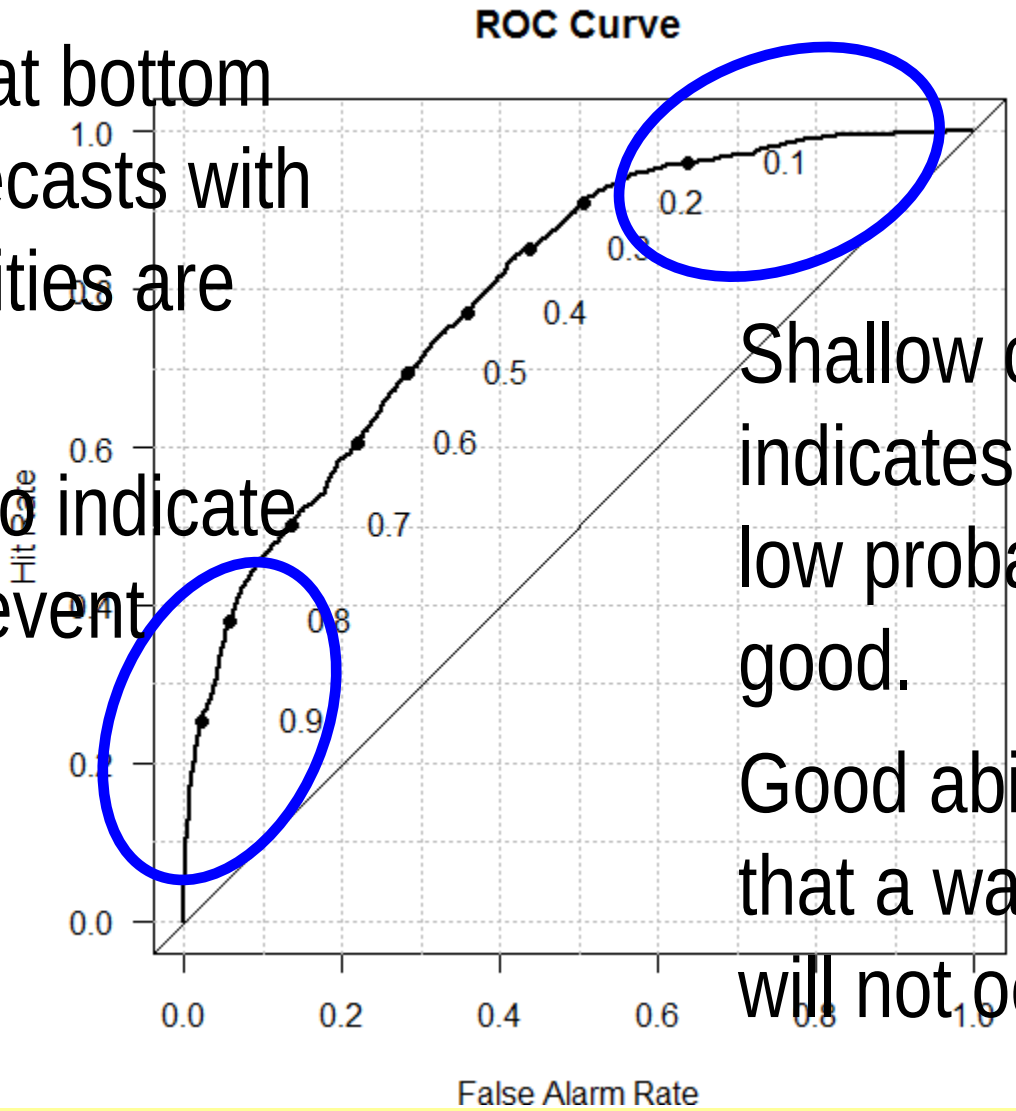
Hit rate=a/(a+c)
False alarm rate=b/(b+d)
ROC curve: plot of hit versus false-alarm rates for decreasing prob. thresholds



- The ROC curve is constructed by calculating the hit and false-alarm rates for decreasing probability thresholds
- Area under ROC curve (A) is a measure of discrimination: A = 0.79 (prob. of successfully discriminating a warm (SST>0) from a cold (SST<0) event)

Steep curve at bottom indicates forecasts with high probabilities are good.

Good ability to indicate that a warm event will occur.

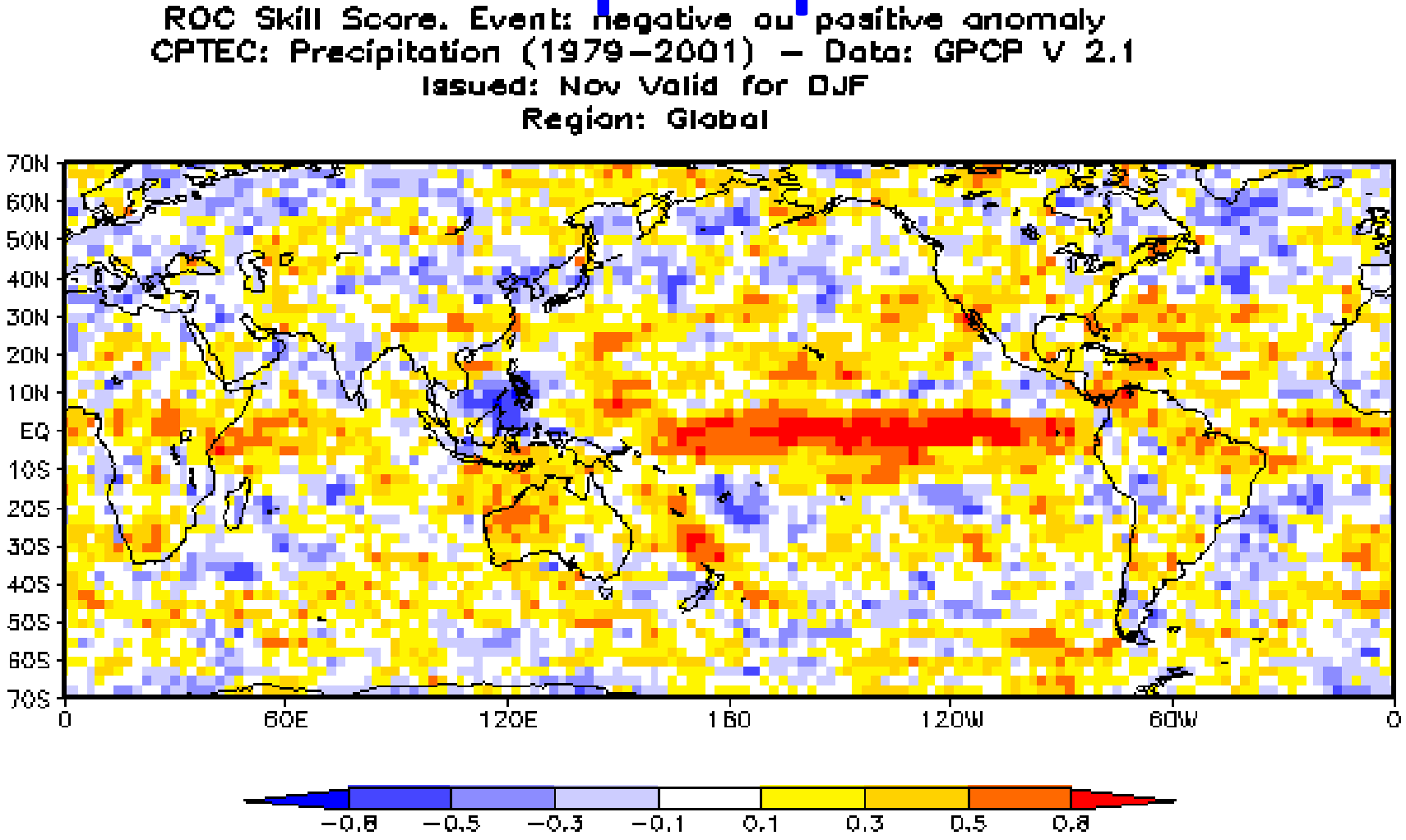


Shallow curve at top indicates forecasts with low probabilities are good.

Good ability to indicate that a warm event will not occur.

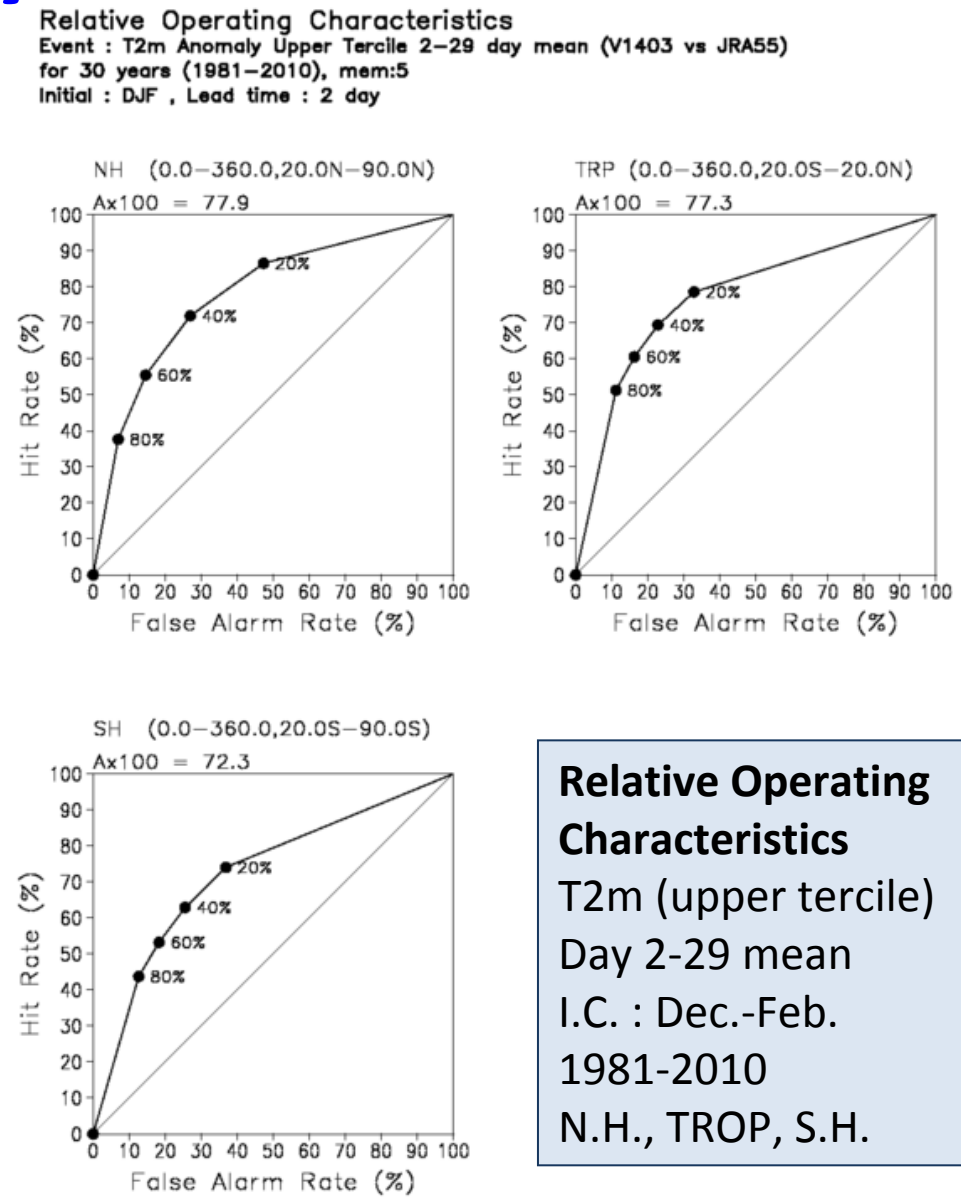
- The ROC curve is constructed by calculating the hit and false-alarm rates for decreasing probability thresholds
- Area under ROC curve (A) is a measure of discrimination: A = 0.79 (prob. of successfully discriminating a warm (SST>0) from a cold (SST<0) event)

Seasonal forecast example:
1-month lead precip. fcsts for DJF



ROC Skill Score = 2 A - 1

Monthly forecast example:
1-day lead 2mT fcsts for day 2-29 mean



Relative Operating Characteristics
T2m (upper tercile)
Day 2-29 mean
I.C. : Dec.-Feb.
1981-2010
N.H., TROP, S.H.

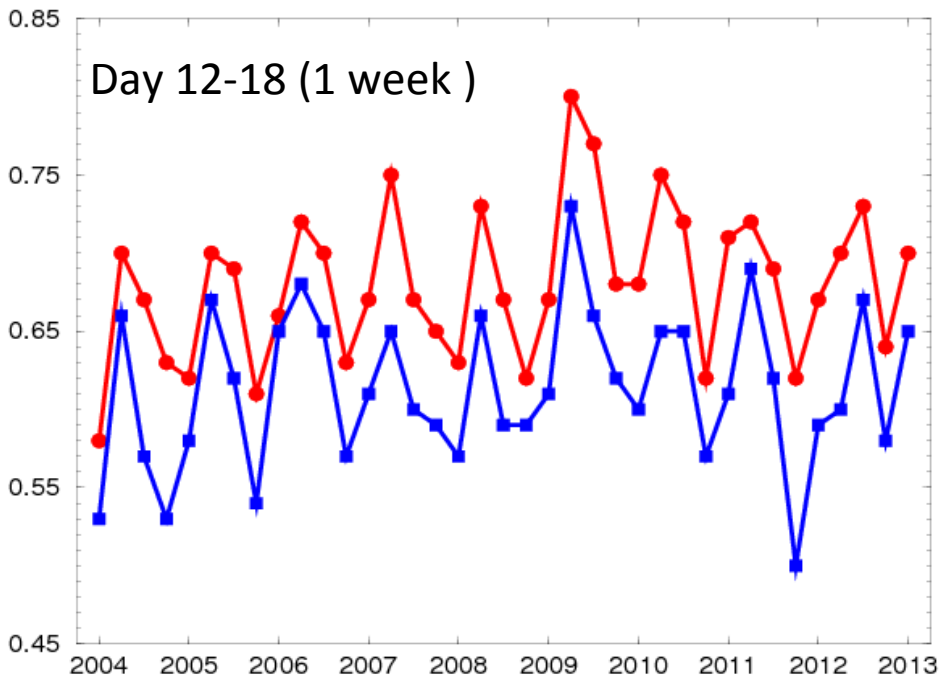
Yuhei Takaya, JMA

One to two weeks forecast example: Northern extratropics

ROC score: 2-metre temperature in the upper tercile

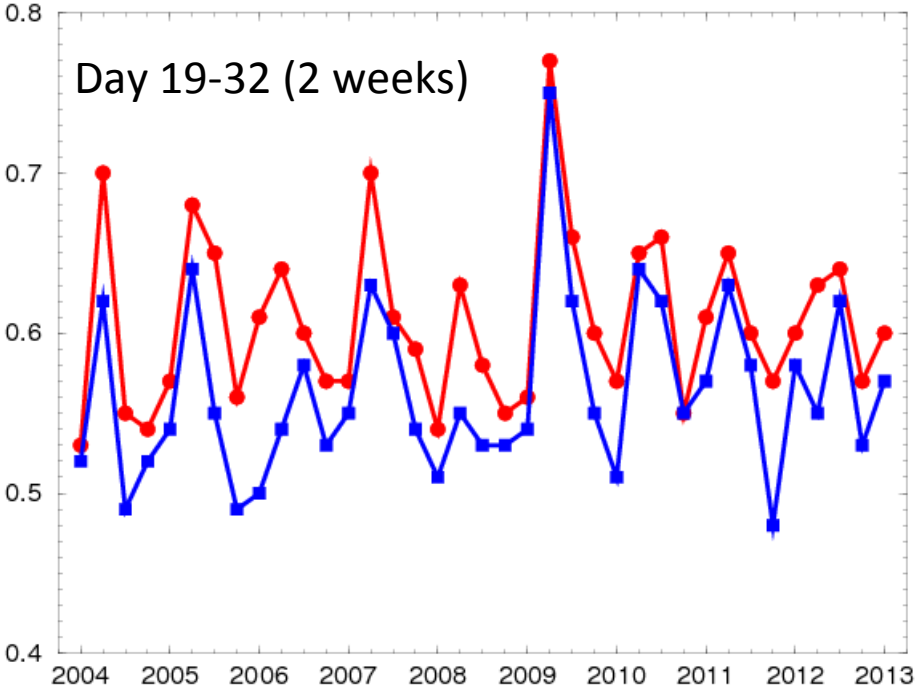
Monthly Forecast

Persistence of day 5-11



Monthly Forecast

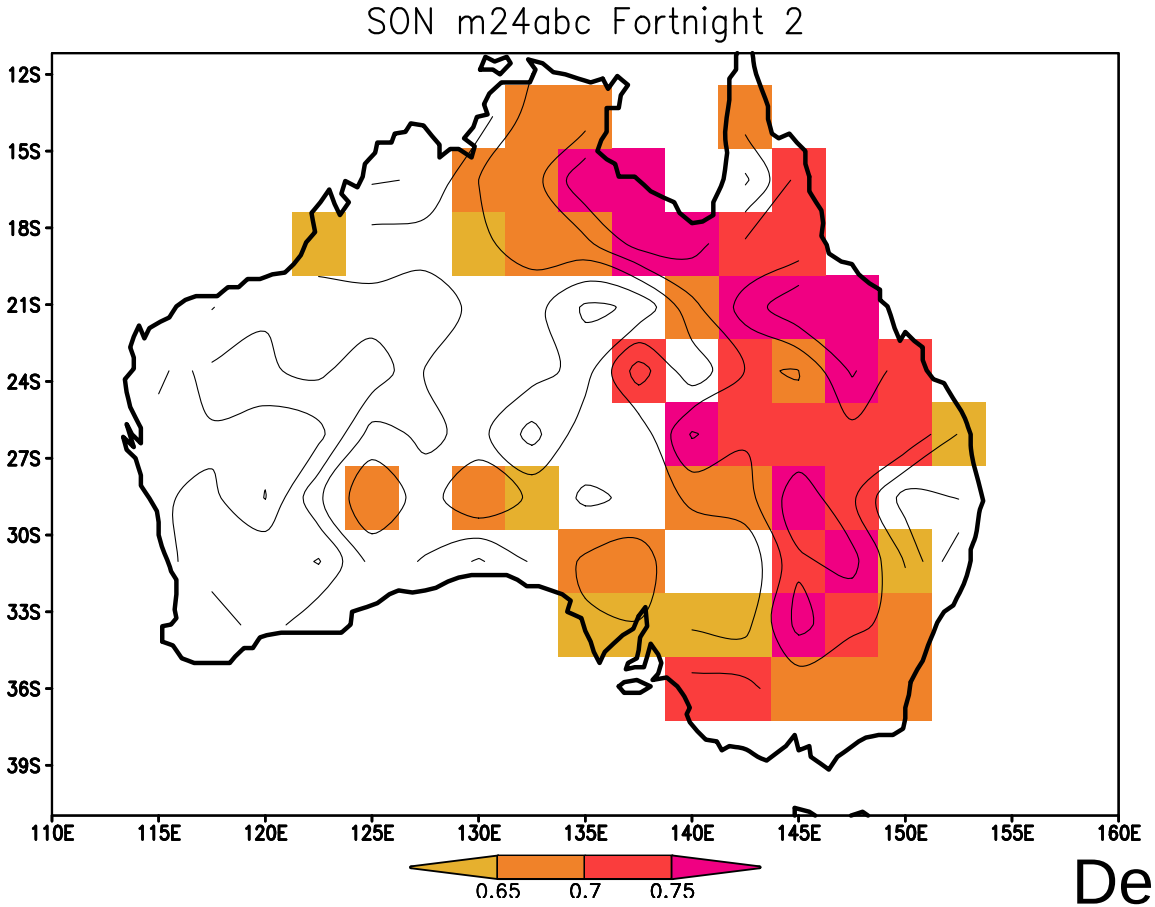
Persistence of day 5-18



Frederic Vitard and Laura Ferranti, ECMWF

Two weeks forecast example: ½ month lead precip. fcsts

ROC area: Precipitation anomalies in the upper tercile
Fortnight 2: Sep, Oct, Nov forecast start months. Hindcasts: 1980-2006



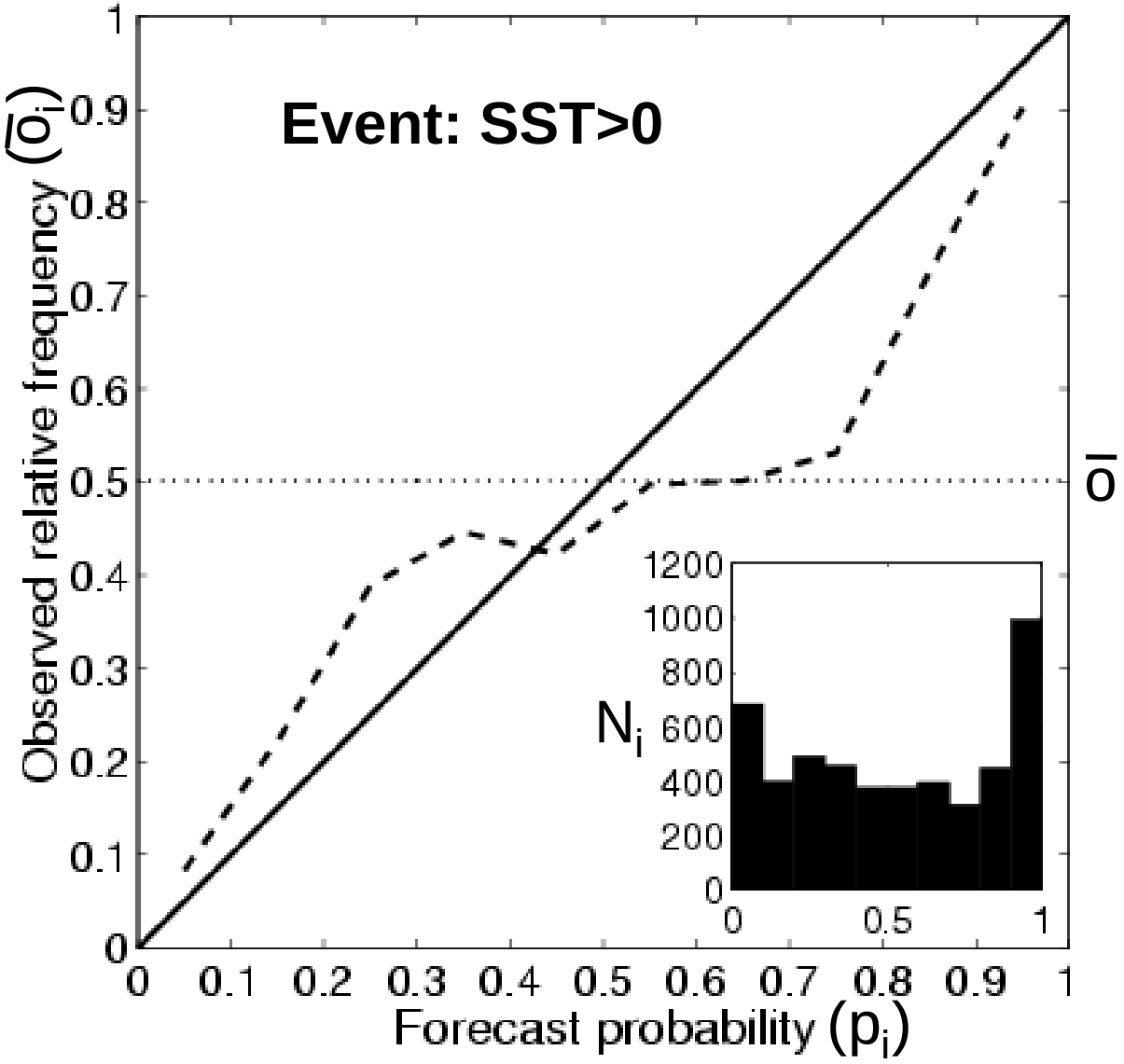
Debbie Hudson
BOM, Australia

Reliability and resolution

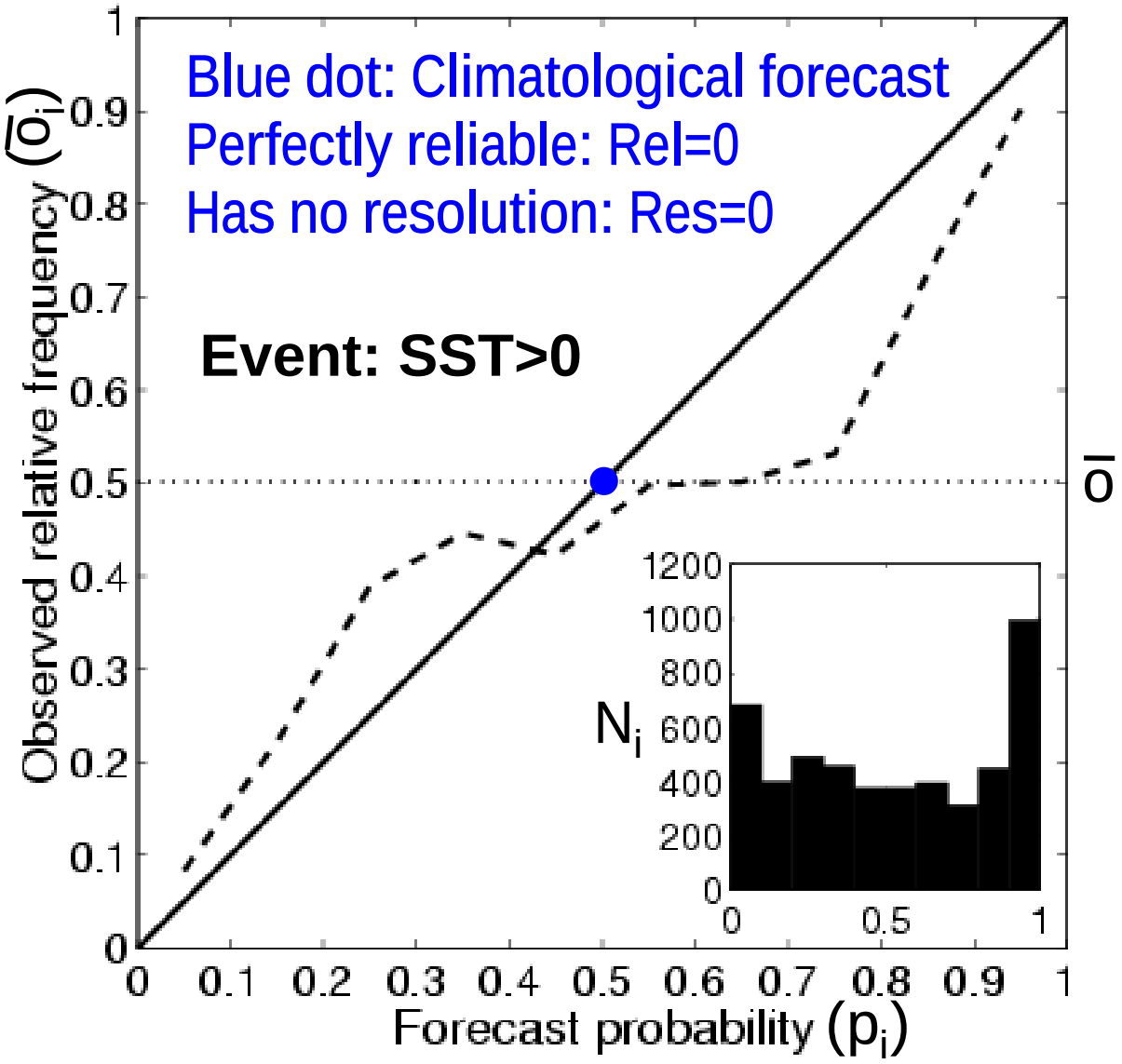
- Reliability: correspondence between forecast probabilities and observed relative frequency (e.g. an event must occur on 30% of the occasions that the 30% forecast probability was issued)

- Resolution: Conditioning of observed outcome on the forecasts
- Addresses the question: Does the frequency of occurrence of an event differs as the forecast probability changes?
- If the event occurs with the same relative frequency regardless of the forecast, the forecasts are said to have no resolution
- Forecasts with no resolution are useless because the outcome is the same regardless of what is forecast

Reliability diagram



Reliability diagram

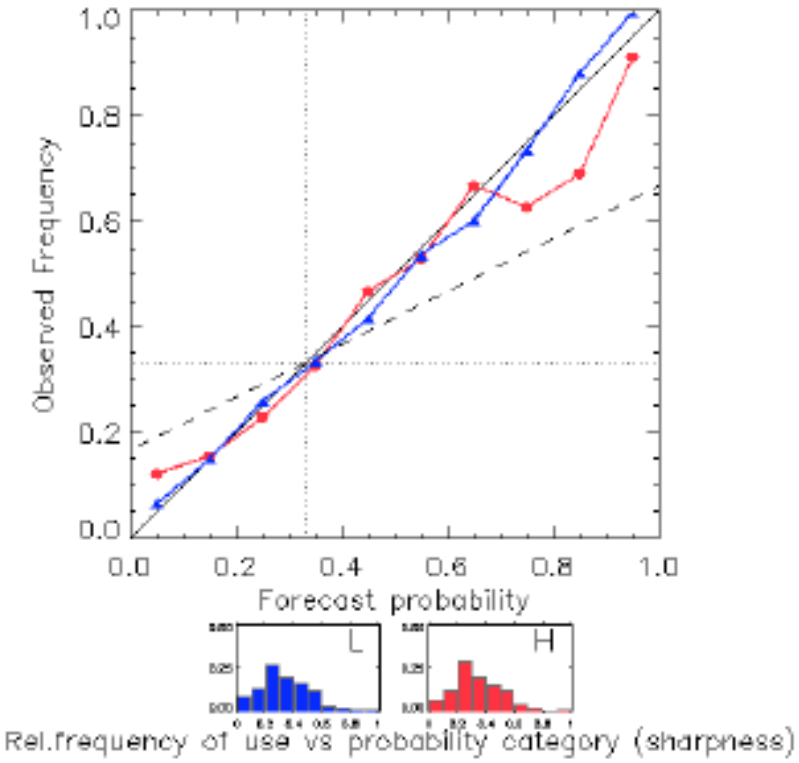


Seasonal forecast example:
1-month lead MSLP fcsts for DJF

GLOSEA5 Hindcast Probabilistic skill

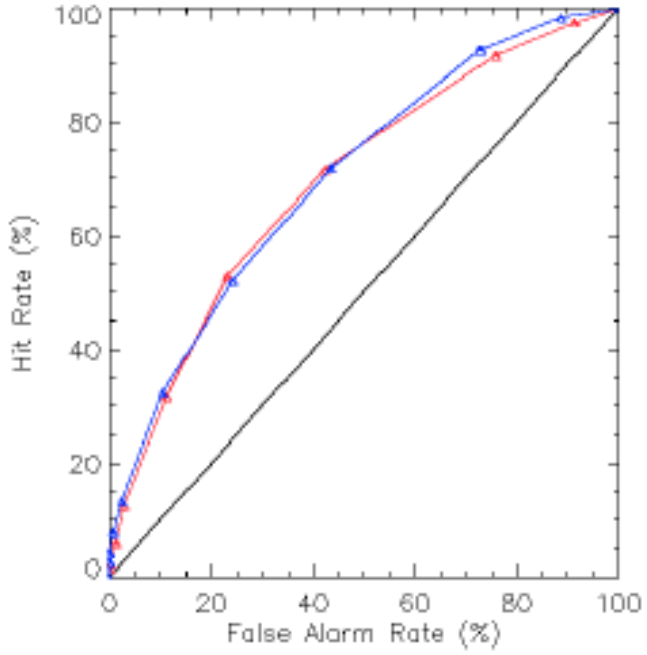
MSLP in N. Atlantic in upper and lower tercile

Reliability



(a) Reliability diagram for mean sea level pressure in GloSea5 over the North Atlantic. The red line shows the upper tercile and the blue line is the lower tercile.

ROC area



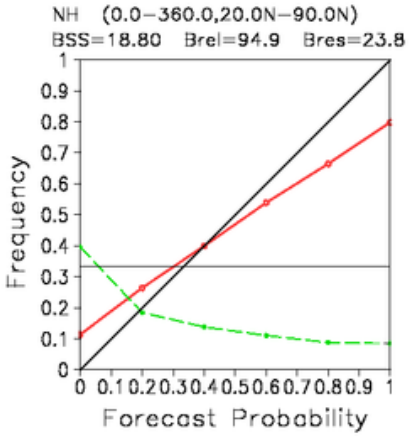
(b) Relative Operating Characteristics (ROC) diagram for the mean sea level pressure in GloSea5 over the North Atlantic. The red line shows the upper tercile and the blue line is the lower tercile.

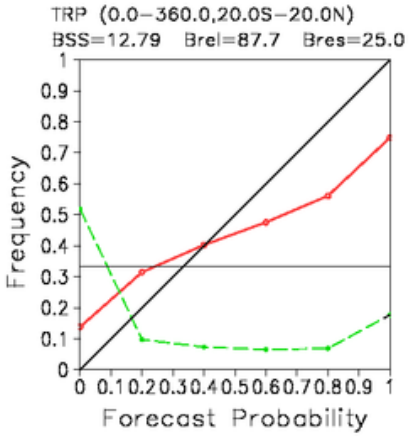
Figure 6. Statistical scores for the Northern Atlantic region.

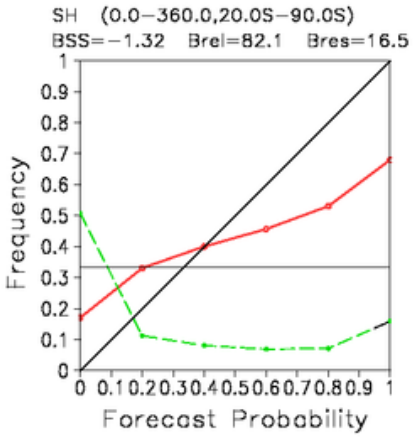
MacLachlan et al., QJRMS, 2014 (under review)

Monthly forecast example:
2-day lead 2mT fcsts for day 2-29 mean

< Reliability Diagram >
Event : T2m Anomaly Upper Tercile 2–29 day mean (V1403 vs JRA55)
BSS, Brel,Bres for 30 years (1981–2010) mem:5
Initial : DJF , Lead time : 2 day
Full(Red)=Reliability Dash(Green)=Forecast Frequency Brier Skill Scores x 100





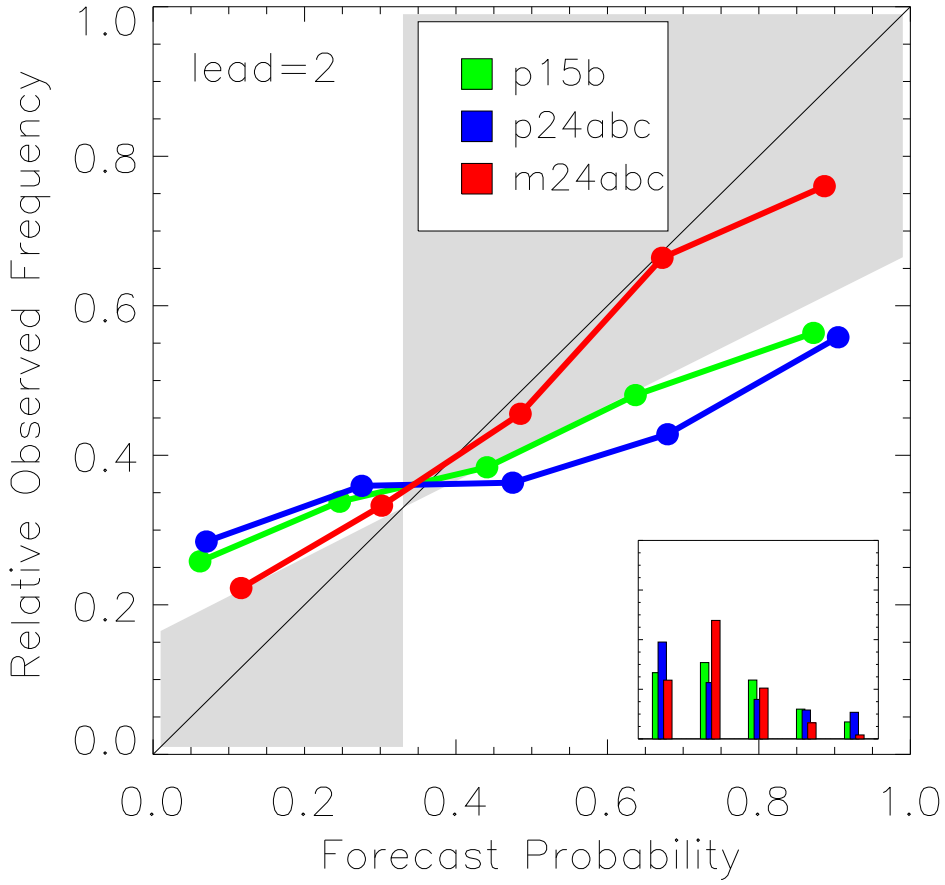


Reliability Diagrams
T2m (upper tercile)
Day 2-29 mean
I.C. : Dec.-Feb.
1981-2010
N.H., TROP, S.H.

Yuhei Takaya, JMA

Two weeks forecast example: ½ month lead precip. fcsts

Precipitation anomalies in the upper tercile
Fortnight 2: Sep, Oct, Nov forecast start months. Hindcasts: 1980-2006



Debbie Hudson
BOM, Australia

Joint Working Group on Forecast Verification Research

Verification working group members

Beth Ebert (BOM, Australia)

Laurie Wilson (CMC, Canada)

Barb Brown (NCAR, USA)

Barbara Casati (Ouranos, Canada)

Caio Coelho (CPTEC, Brazil)

Anna Ghelli (ECMWF, UK)

Martin Göber (DWD, Germany)

Simon Mason (IRI, USA)

Marion Mittermaier (Met Office, UK)

Pertti Nurmi (FMI, Finland)

Joel Stein (Météo-France)

Yuejian Zhu (NCEP, USA)

Aims

Verification component of WWRP, in collaboration with WGNE, WCRP, CBS

("Joint" between WWRP and WGNE)

- Develop and promote new verification methods
- Training on verification methodologies
- Ensure forecast verification is relevant to users
- Encourage sharing of observational data
- Promote importance of verification as a vital part of experiments
- Promote collaboration among verification scientists, model developers and forecast providers

6th WMO International Verification Methods Workshop

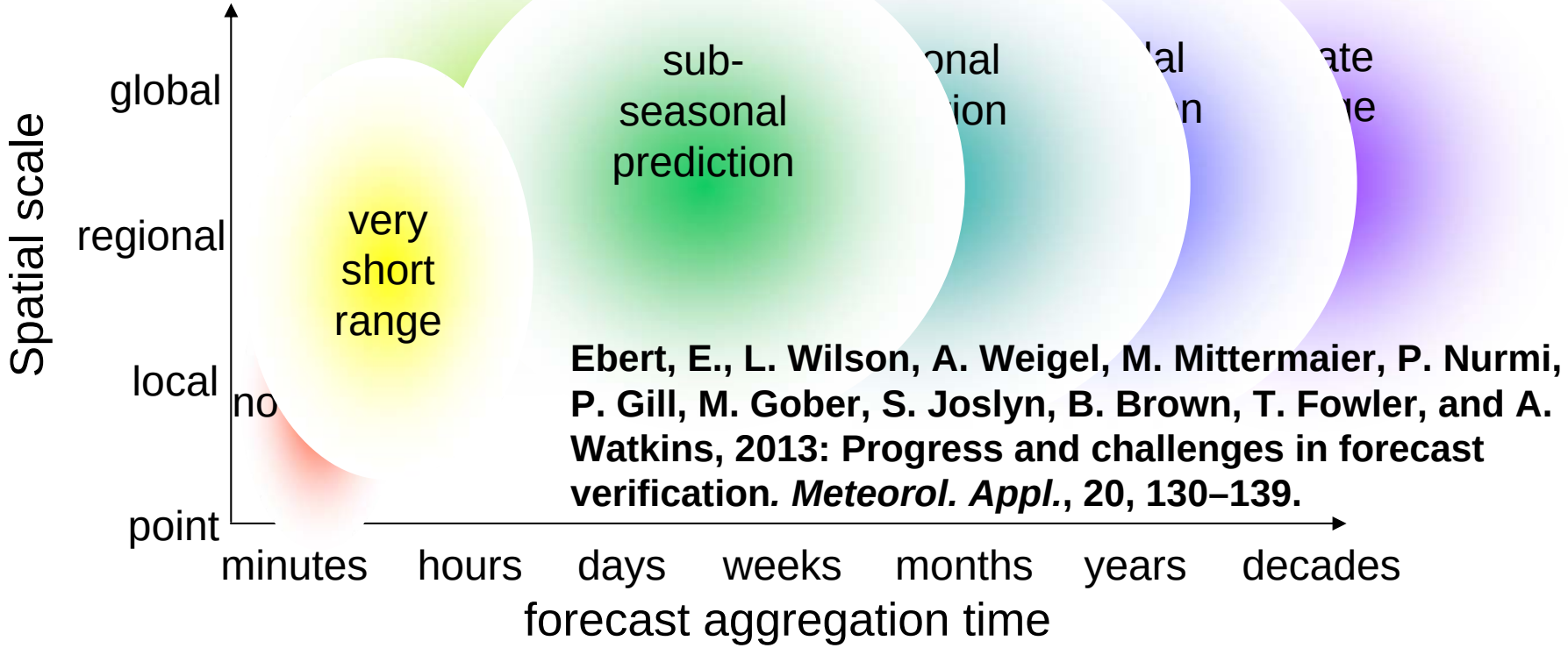
13-19 March 2014
New Delhi, India

Tutorial session: March 13-15
Scientific program: March 16-19

Seamless verification

Seamless forecasts - consistent across space/time scales

single modelling system or blended



Ebert, E., L. Wilson, A. Weigel, M. Mittermaier, P. Nurmi, P. Gill, M. Gober, S. Joslyn, B. Brown, T. Fowler, and A. Watkins, 2013: Progress and challenges in forecast verification. *Meteorol. Appl.*, 20, 130–139.

Final remarks

- S2S verification is naturally leaning towards the seamless consistency concept

- Clear need for attributes-based verification for a complete forecast quality view

- Need for use more than a single score for more detailed forecast quality diagnostics

- Standards for coordinated seasonal forecast verification already exist (WMO SVSLRF)

- S2S project offers opportunity for establishing standards for seamless S2S forecast verification using single and multi-model ensembles and also link verification activities performed by research and operational communities

Relevant recent development in S2S verification

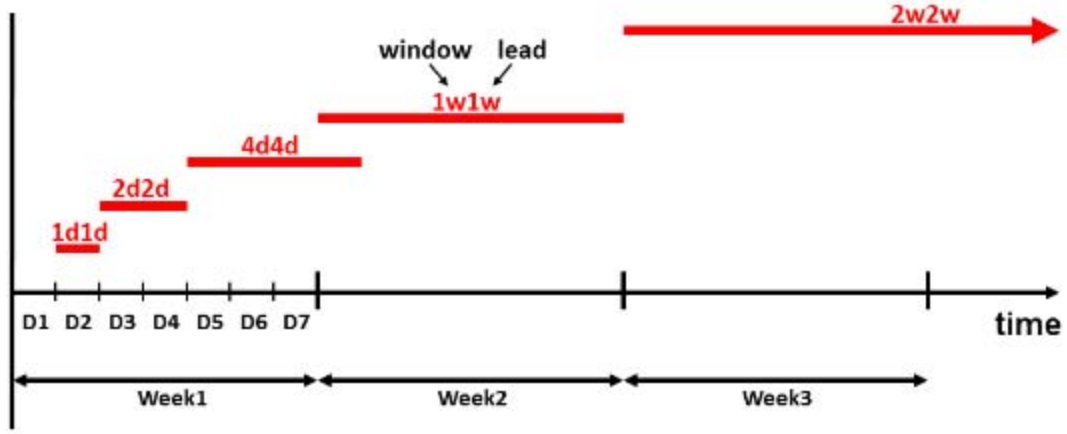


Fig. 1. Schematic of the time window and lead time definitions used in this analysis. The horizontal axis represents forecast time from the initial condition. "1d1d" refers to an averaging window of 1 day at a lead time of 1 day. Similarly, "2d2d" represents an averaging window of 2 days at a lead time of 2 days, and so on. Note that "1d1d" is what is usually called "day 2" in other papers, and "1w1w" is what is usually called "week 2".

Hongyan Zhu, Matthew C. Wheeler, Adam H. Sobel, and Debra Hudson (2014) Seamless precipitation prediction skill in the tropics and extratropics from a global model. MWR (in press)

Thank you all for your attention!

Brier Score decomposition (Murphy, 1973)

$$BS = \frac{1}{n}\sum_{k=1}^{n}(p_k - o_k)^2 \qquad 0 \le BS \le 1$$

$$BS = \underbrace{\frac{1}{n}\sum_{i=1}^{l} N_i(p_i - \bar{o}_i)^2}_{\text{Reliability}} - \underbrace{\frac{1}{n}\sum_{i=1}^{l} N_i(\bar{o}_i - \bar{o})^2}_{\text{Resolution}} + \underbrace{\bar{o}(1-\bar{o})}_{\text{Uncert.}}$$

$$\bar{o}_i = p(o_1 \mid p_i) = \frac{1}{N_i}\sum_{k \in N_i} o_k \qquad \bar{o} = \frac{1}{n}\sum_{k=1}^{n} o_k$$

$$i = 1,\ldots,l = 11 : p_1 = 0,\ p_2 = 0.1,\ p_3 = 0.2,\ldots,\ p_{10} = 0.9, p_{11} = 1$$

The Brier score can be improved (reduced):

- forecasting events of small var(o)=$\bar{o}$(1-$\bar{o}$) (reduced uncertainty)
- increasing *resolution* (e. g. combining forecasts)
- improving *reliability* (e. g. calibrating forecasts)

Note: It is common practice to decompose the Brier score in reliability and resolution for examining which component can be improved

Reliability diagram