Forecast attributes and metrics

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

Acknowledgments: Arun Kumar, Alberto Arribas, Barbara Brown, Beth Ebert, David Stephenson, Debbie Hudson, Francisco Doblas-Reyes, Laura Ferranti, Nachiketa Acharya, Simon Mason and Yuhei Takaya

Lecture plan

1) Brief review of forecast goodness
2) Attributes based forecast quality assessment: examples of sub-seasonal to seasonal verification practice
3) Final remarks

Seventh WMO International Workshop on Monsoons (IWM-7)

ONLINE TRAINING WORKSHOP ON

SUBSEASONAL TO SEASONAL (S2S) PREDICTION OF MONSOONS

1-12 NOVEMBER 2021

What is a good forecast?

Good forecasts have:

- **QUALITY: Measure of correspondence btw forecasts and observations using mathematical relationship (deterministic and probabilistic measures)**
- **VALUE/UTILITY: Measure of benefit achieved (or loss incurred) through the use of forecasts**
- **CONSISTENCY: Correspondence between a forecast and the forecasters belief with appropriate representation of forecast uncertainty**

Attributes of quality:

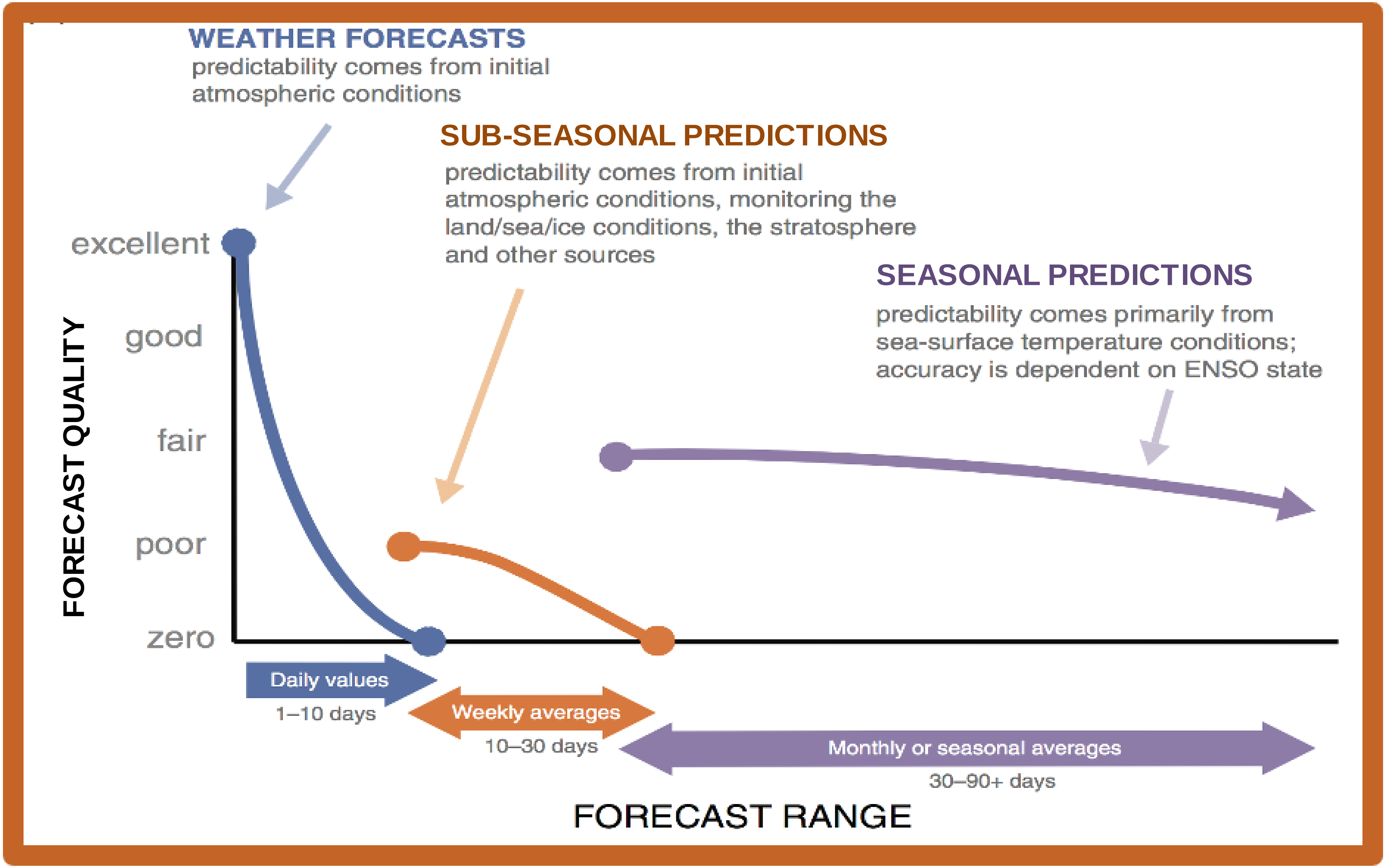
- Association
- Accuracy
- Discrimination
- Reliability
- Resolution

…

→ 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.

Forecast quality on different time ranges



Source: Adapted from the IRI

Sub-seasonal to seasonal forecast quality assessment

1. Attributes of deterministic forecasts (ensemble mean)

Association

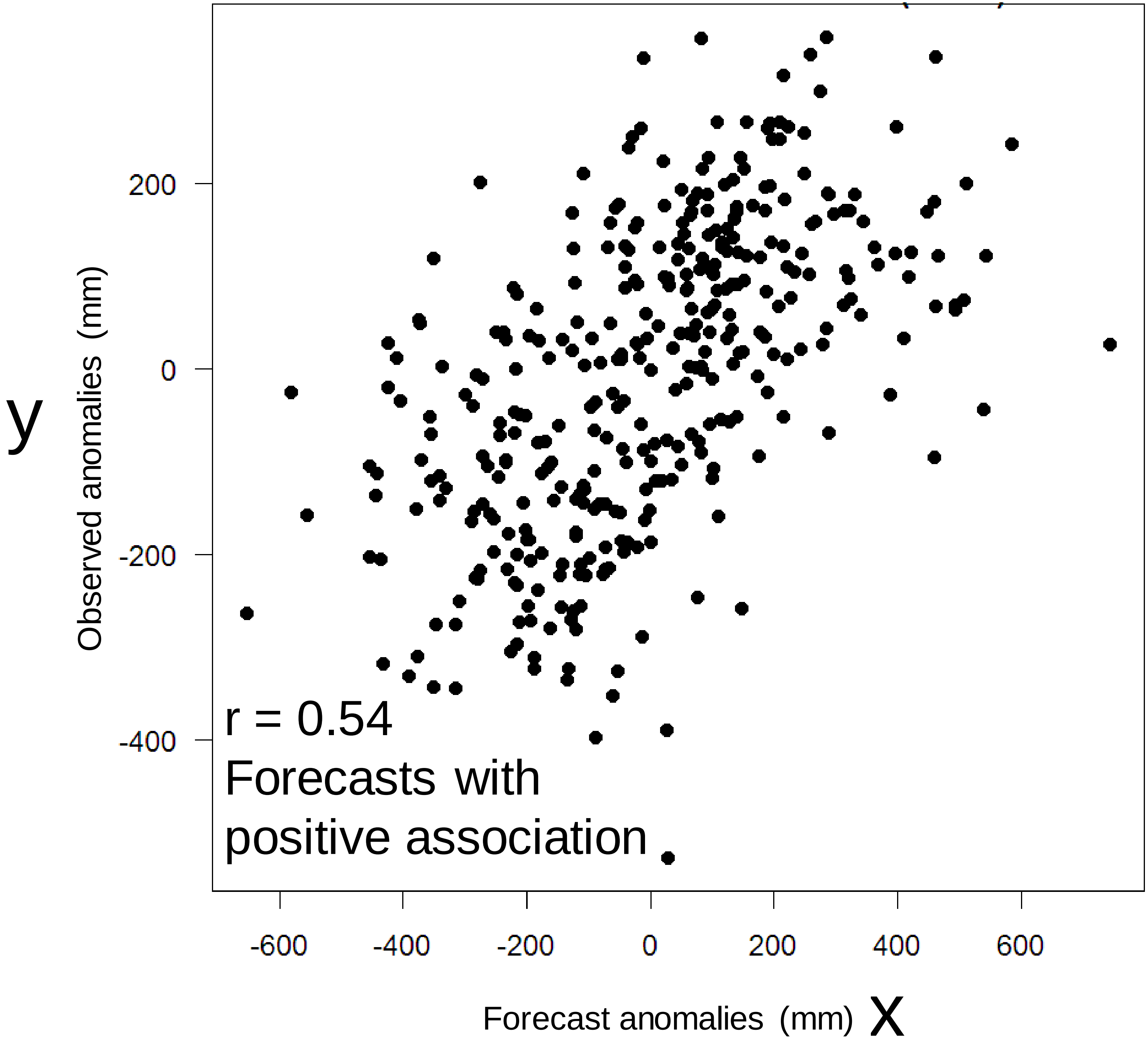
- Overall strength of the relationship between the forecasts and observations
- Linear association is often measured using the product moment **correlation coefficient**

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

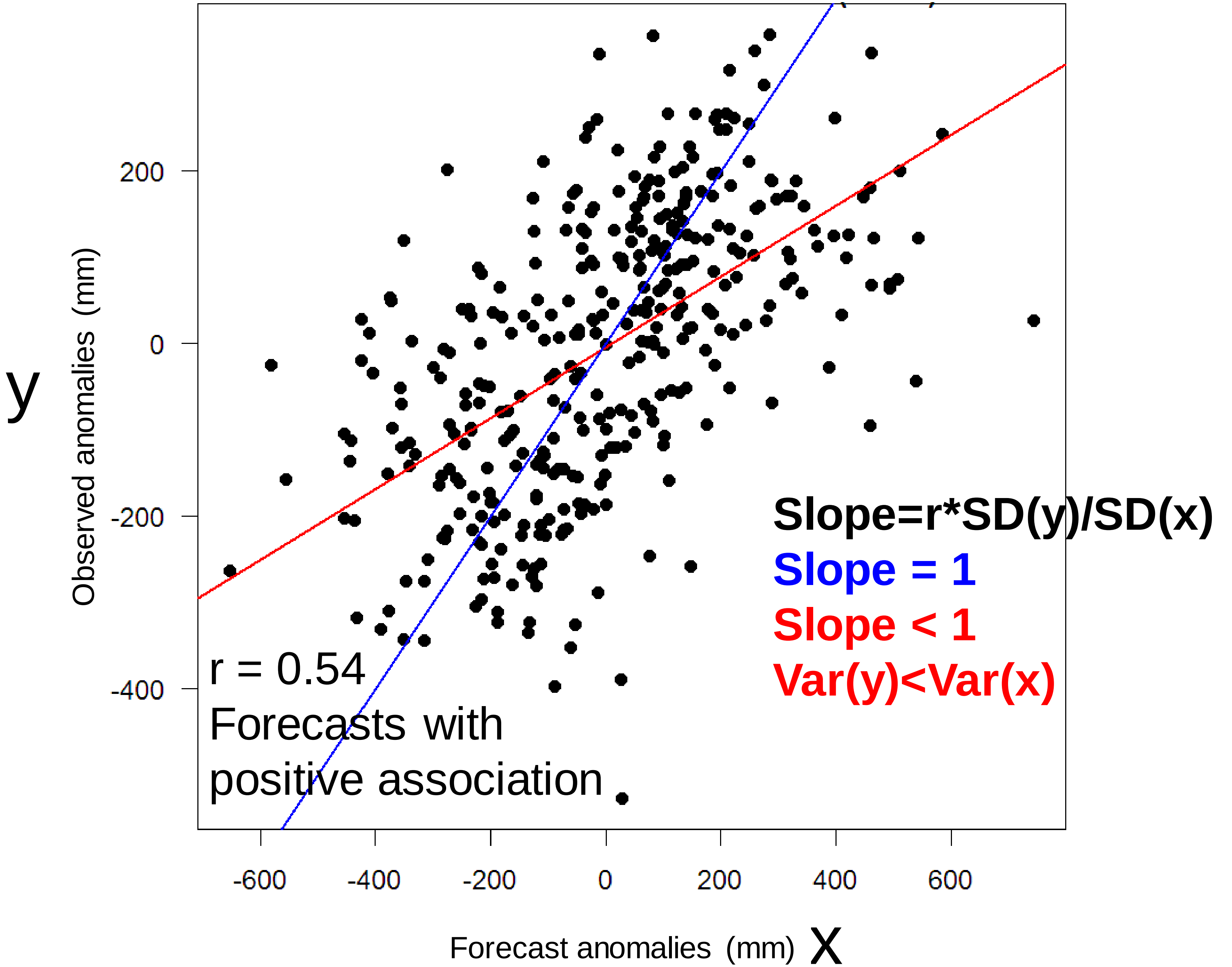
x: forecast *y: observation*

n: number of (x,y) pairs

Relationship between past forecast and past obs. anomalies



Relationship between past forecast and past obs. anomalies



Accuracy

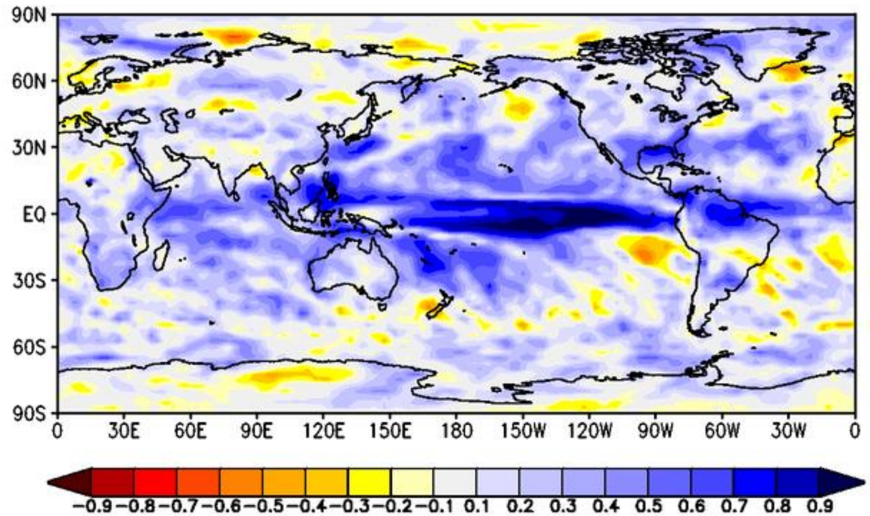
- Average difference between forecasts and observations
- Simplest measure is the **Mean Error (Bias)**

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

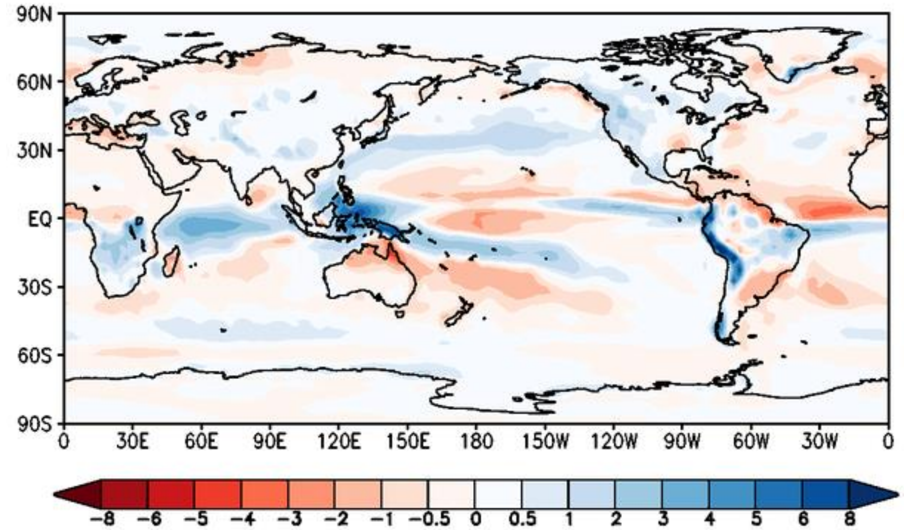
x: forecast *y: observation* *n: number of (x,y) pairs*

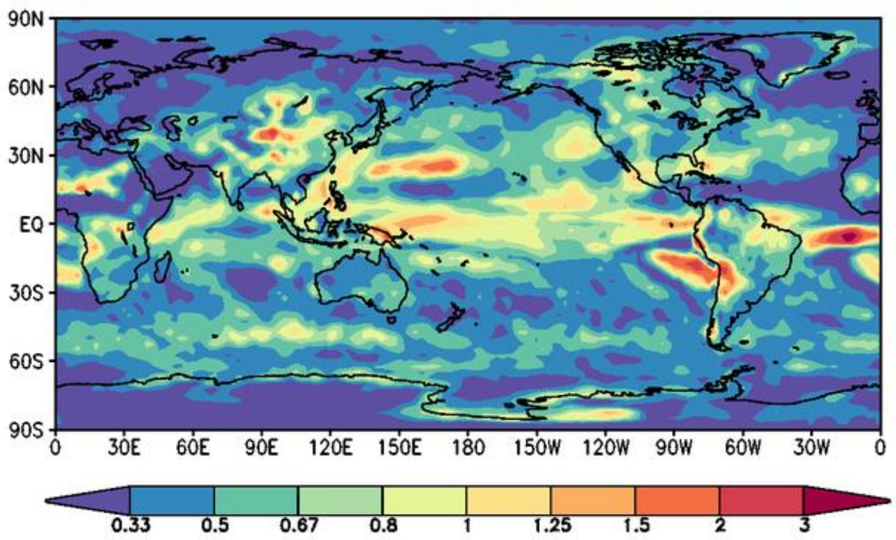
Seasonal forecast example:
JMA 1-month lead precip. fcsts for DJF

Corr. btw (F, O) anoms (against GPCP v2.2)
I.C: Nov. Valid: DJF (1981-2010)



Bias (against GPCP v2.2)
I.C: Nov Valid: DJF (1981-2010)



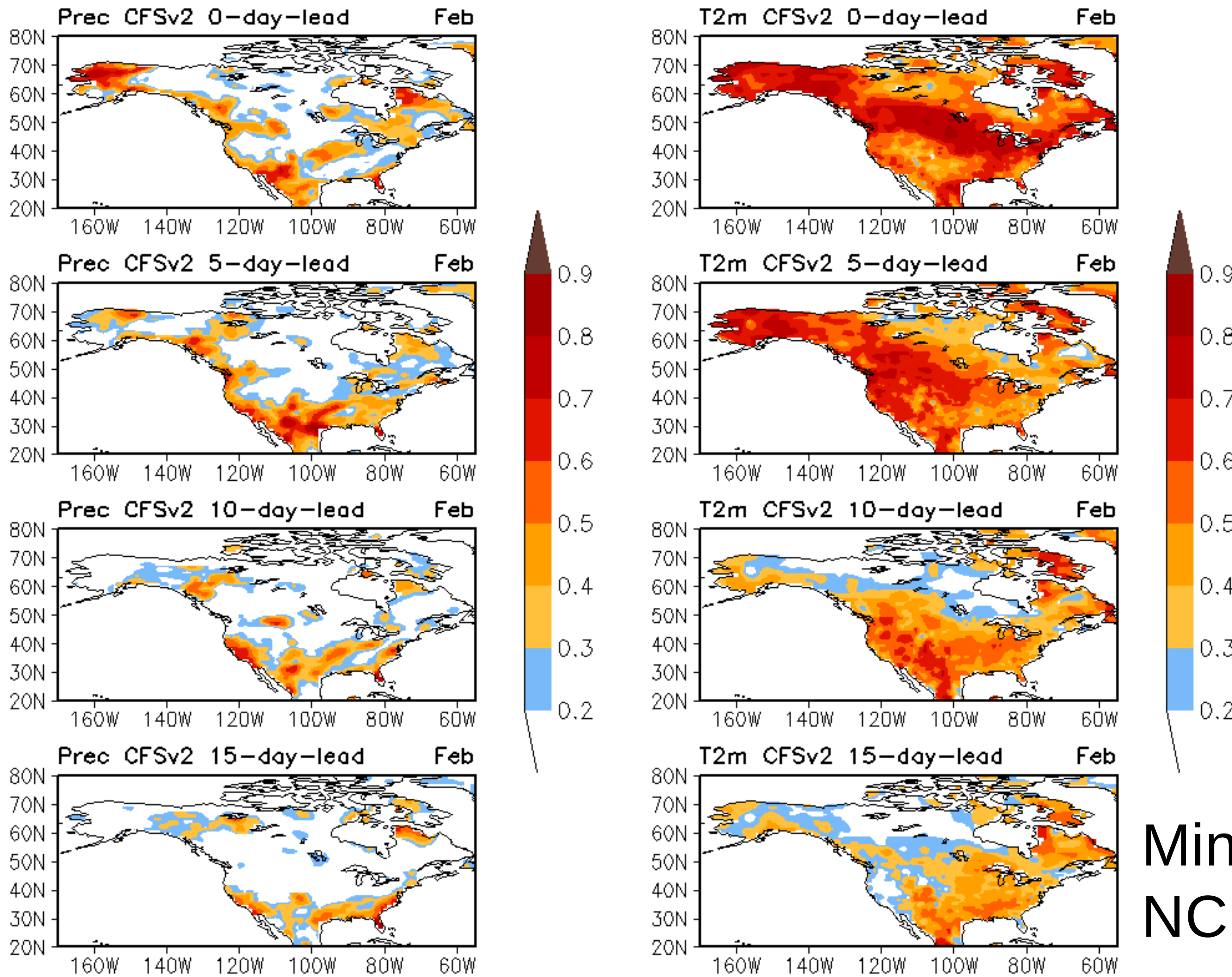


St. dev ratio (F/O) (against GPCP v2.2)
I.C: Nov Valid: DJF (1981-2010)

Source: JMA/MRI

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

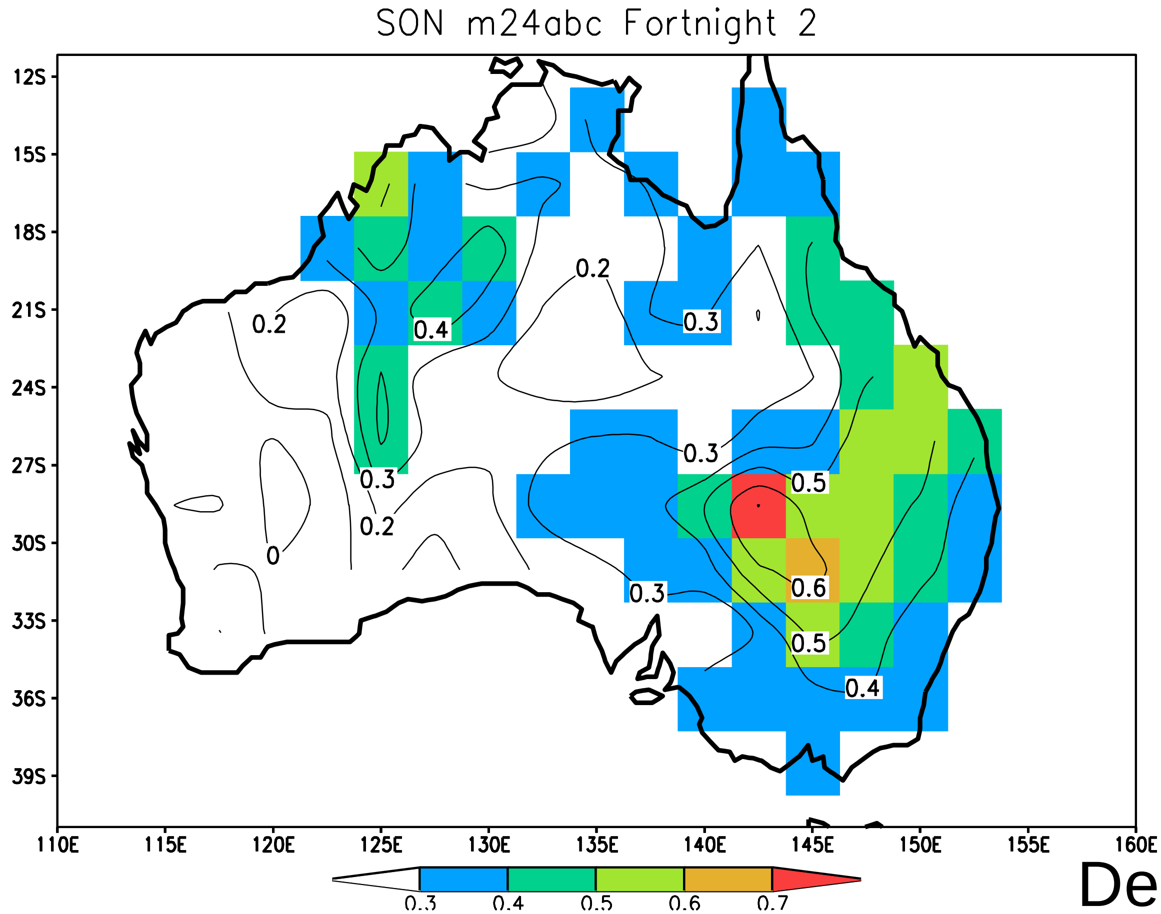
Precipitation CFSv2 Correlation Feb (1982–2009) **2m Temperature**



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

Sub-seasonal to seasonal forecast quality assessment

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

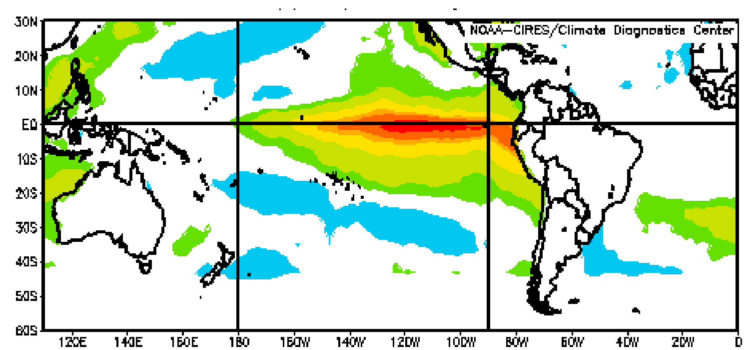
Discrimination

- 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

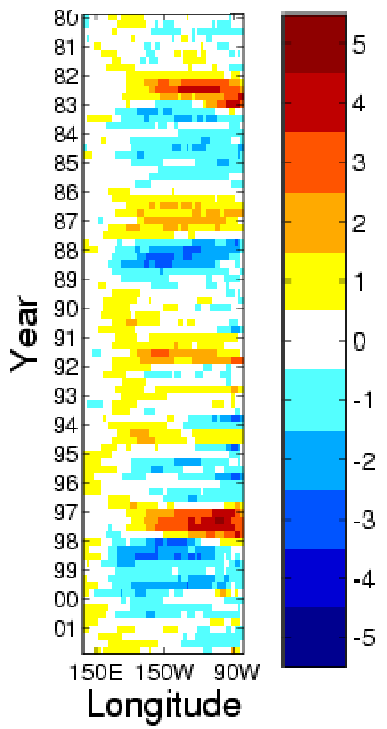
Example:Equatorial Pacific SST anomaly forecasts

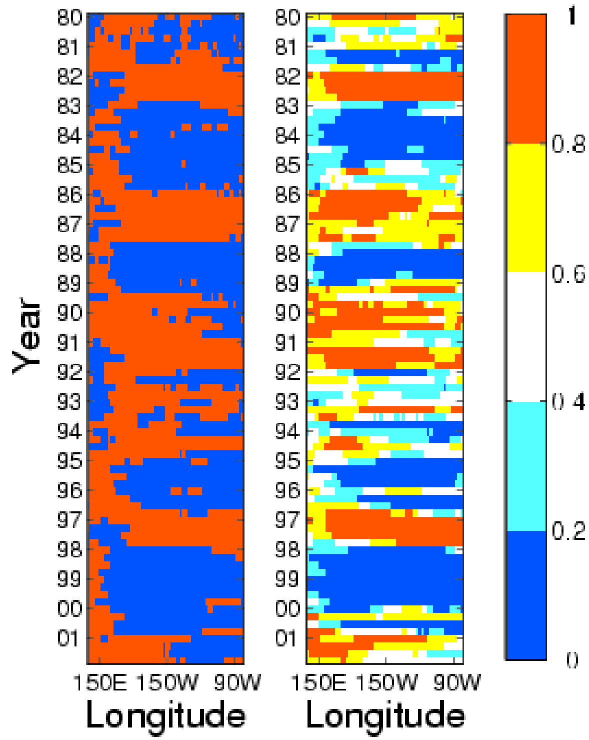
88 seasonal probability forecasts of binary SST anomalies at 56 grid points along the equatorial Pacific. Total of 4928 forecasts.

6-month lead forecasts for 4 start dates (F,M,A,N) valid for (Jul,Oct,Jan,Apr)



SST $\qquad o = (SST > 0) \qquad f = \Pr(\hat{o})$





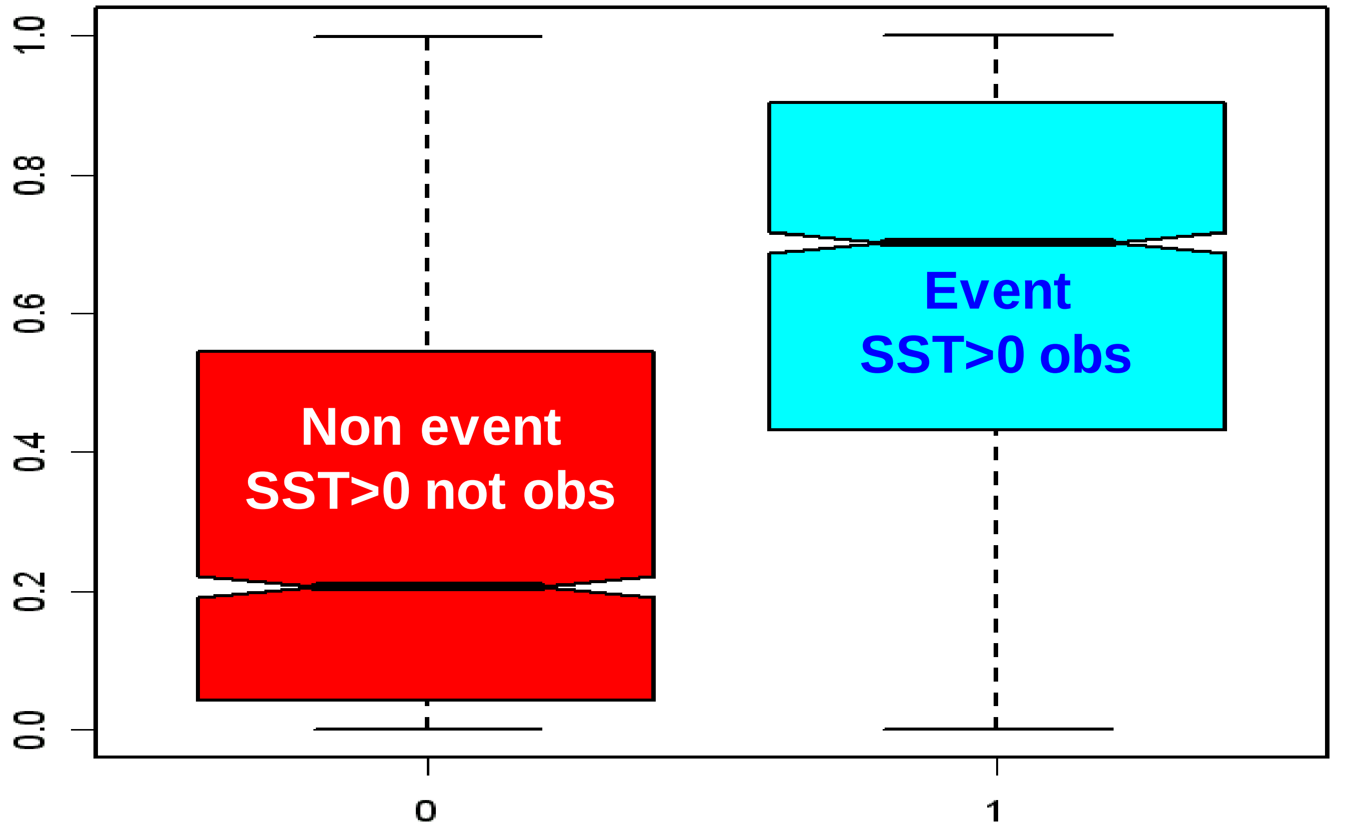
The probability forecasts were constructed by fitting Normal distributions to the ensemble mean forecasts from the 7 DEMETER coupled models, and then calculating the area under the normal density for SST anomalies greater than zero.

SST anomalies (°C)

Forecast probabilities: f

Prob. forecasts conditioned/stratified on observations

Forecast probability Pr(SST>0)



Observed binary event X

→ Forecasts do differ given different outcomes

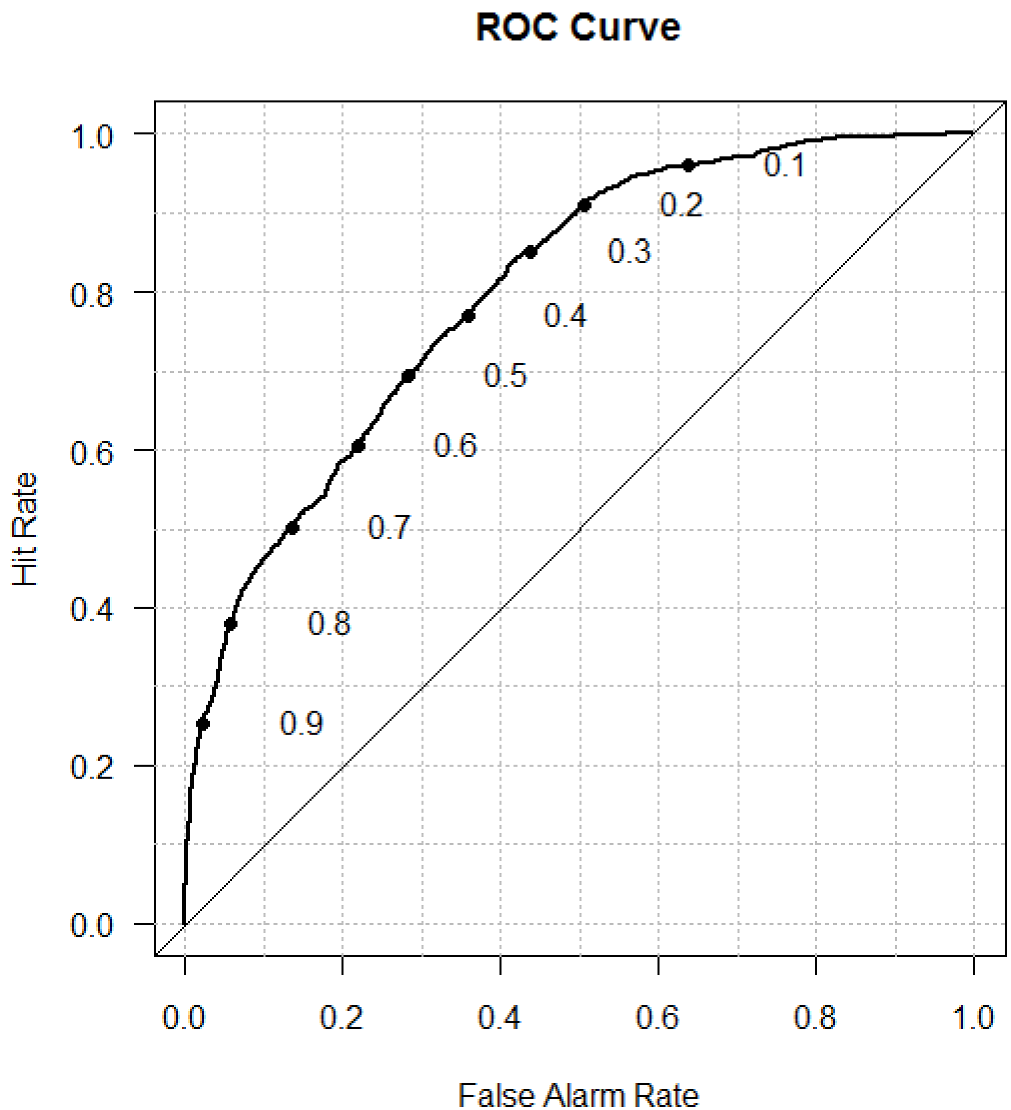
→ Forecast system has discrimination (distinguish event from non-event)

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
No	c (Miss)	d (Correct rejection)	c+d
Total	a+c	b+d	a+b+c+d=n

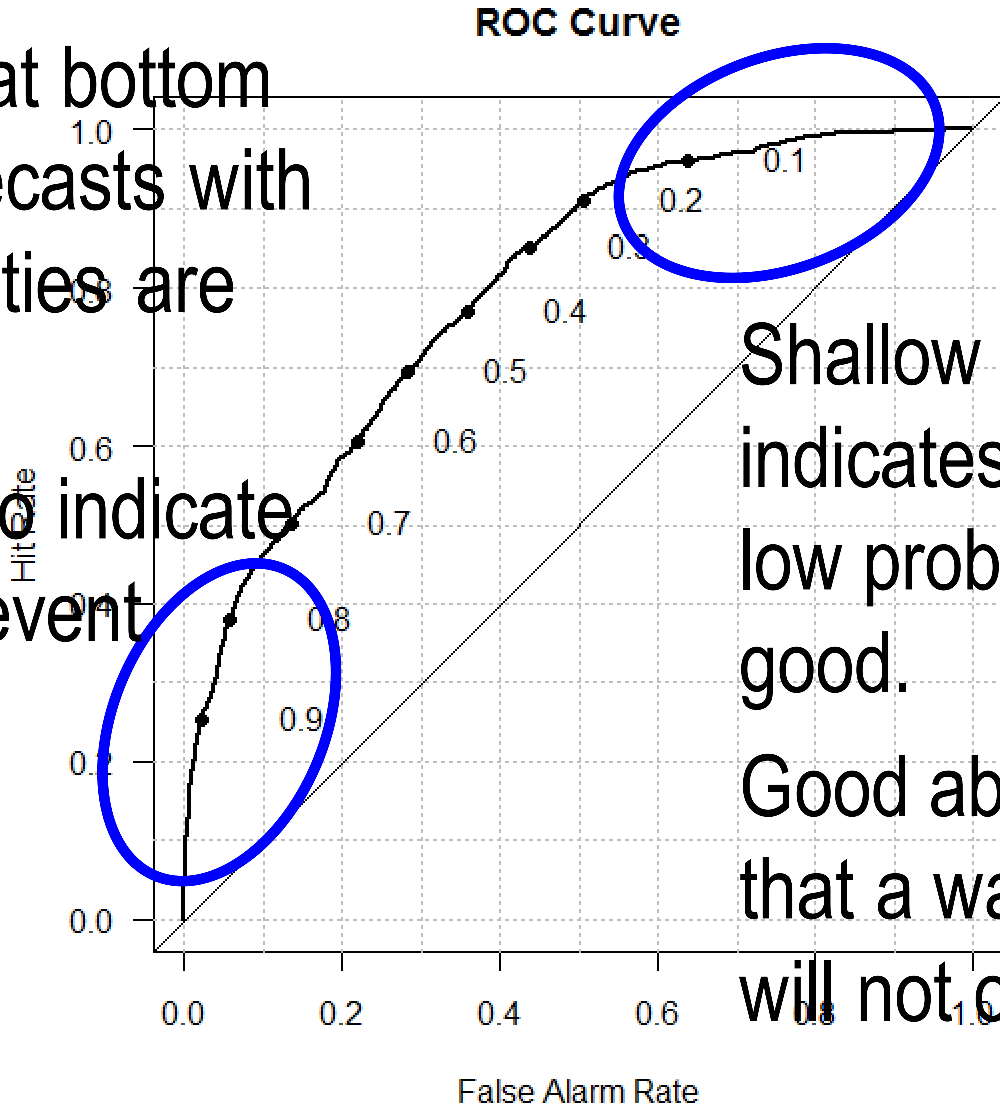
Hit rate=a/(a+c)
False alarm rate=b/(b+d)
ROC curve: plot of hit versus false-alarm rates for various prob. thresholds



- The ROC curve is constructed by calculating the hit and false-alarm rates for various 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.

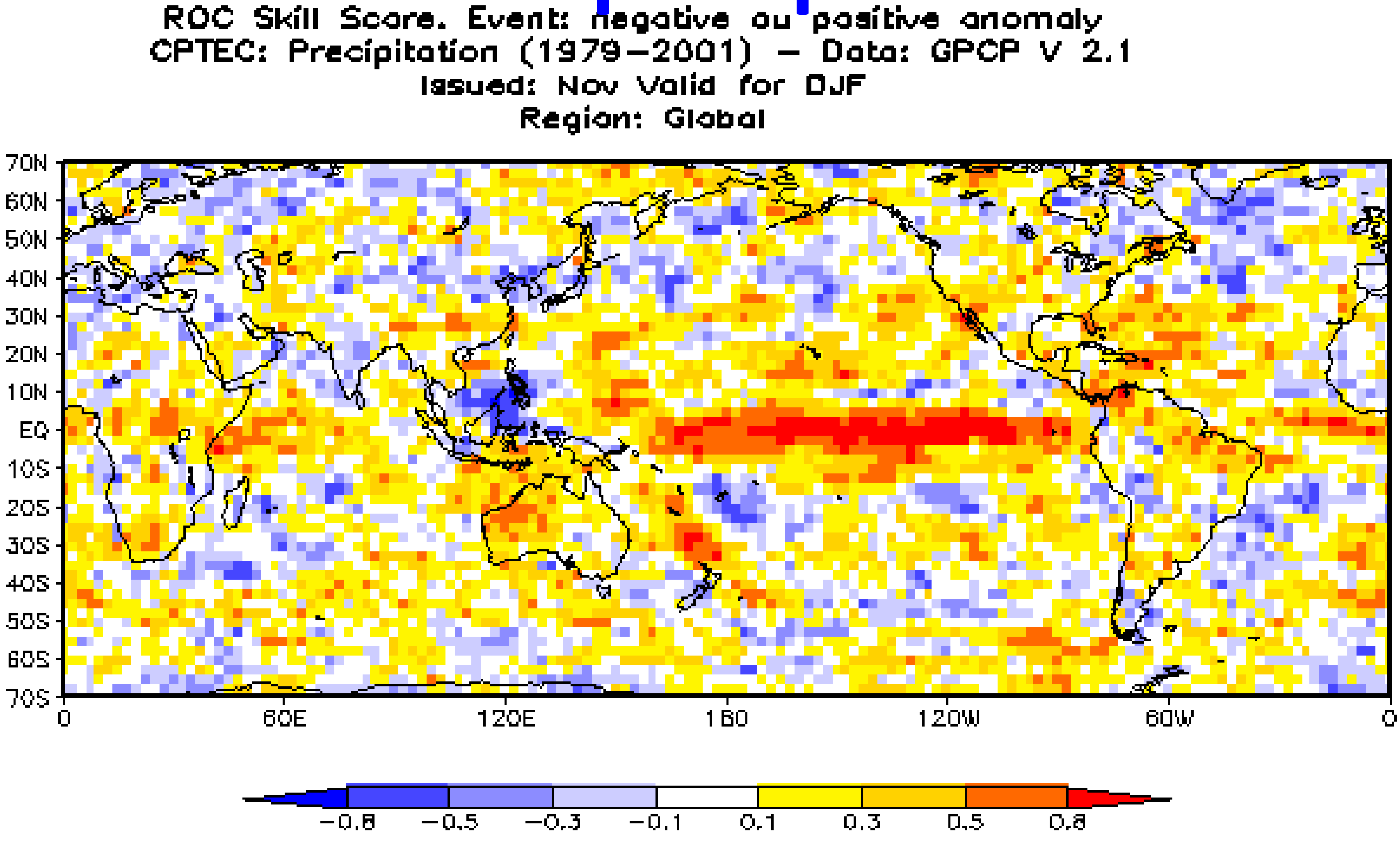
Good ability to indicate that a warm event will not occur.

- The ROC curve is constructed by calculating the hit and false-alarm rates for various 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)

Important points to remember

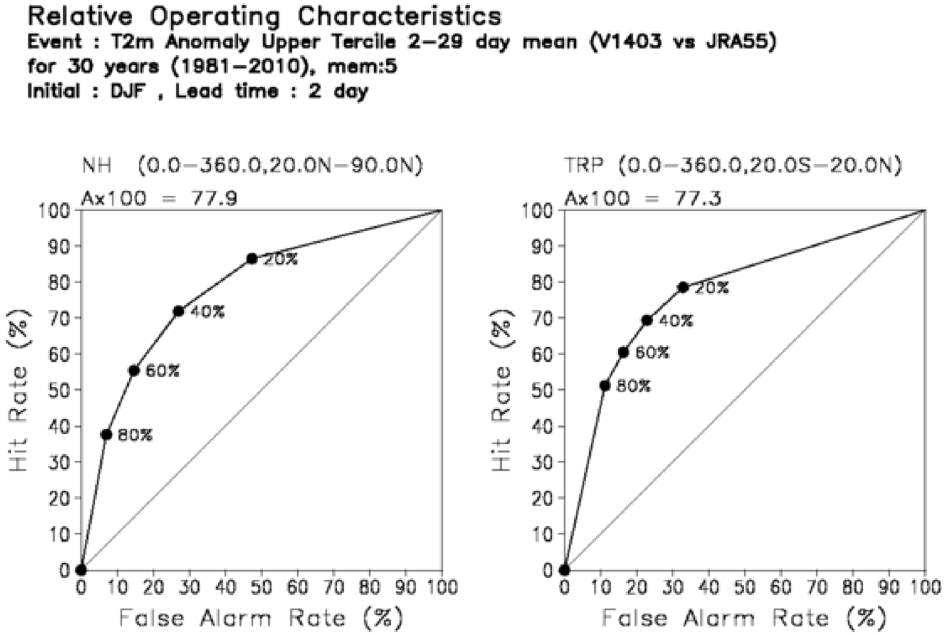
- The area under the ROC curve will tell us the probability of successfully discriminating an event from a non event. In other words, how different the forecast probabilities are for events and non events
- As events and non-events are binary (i.e have 2 possible outcomes) the probability of correctly discriminating (distinguishing) and event from a non-event by chance (guessing) is 50% and is represented by the area below the 45 degrees diagonal line in the ROC plot
- ROC is not sensitive to biases in the forecasts
- Forecast biases are diagnosed with the reliability diagram

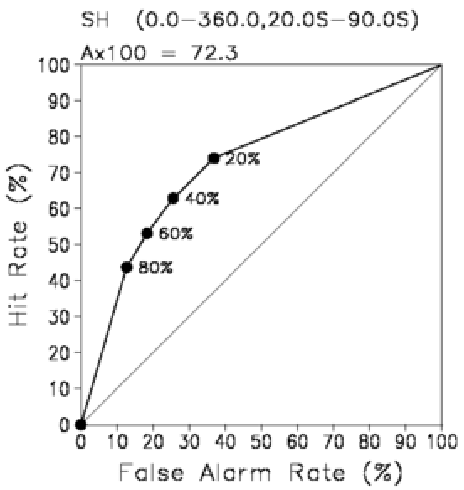
Seasonal forecast example: 1-month lead precip. fcsts for DJF



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.

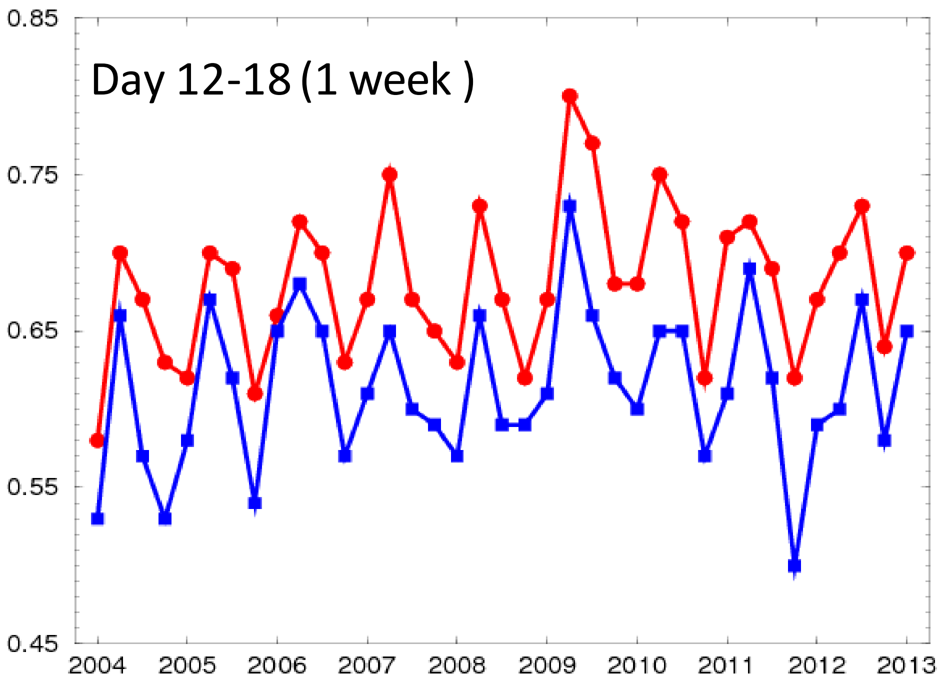
Yuhei Takaya, JMA

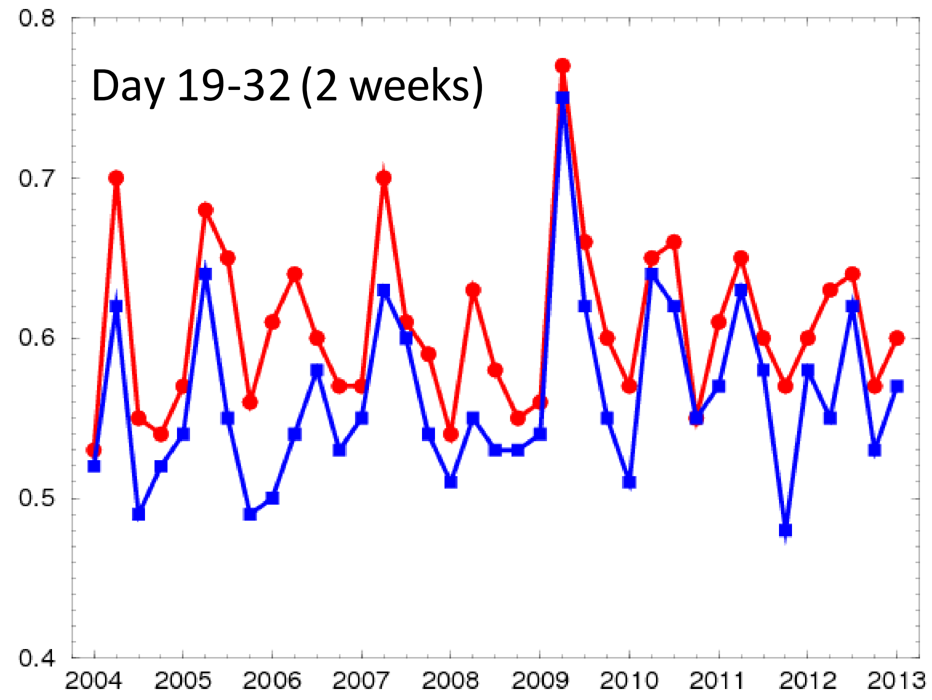
One to two weeks forecast example: Northern extratropics

ROC area: 2-metre temperature in the upper tercile

Monthly Forecast
Persistence of day 5-11

Monthly Forecast
Persistence of day 5-18

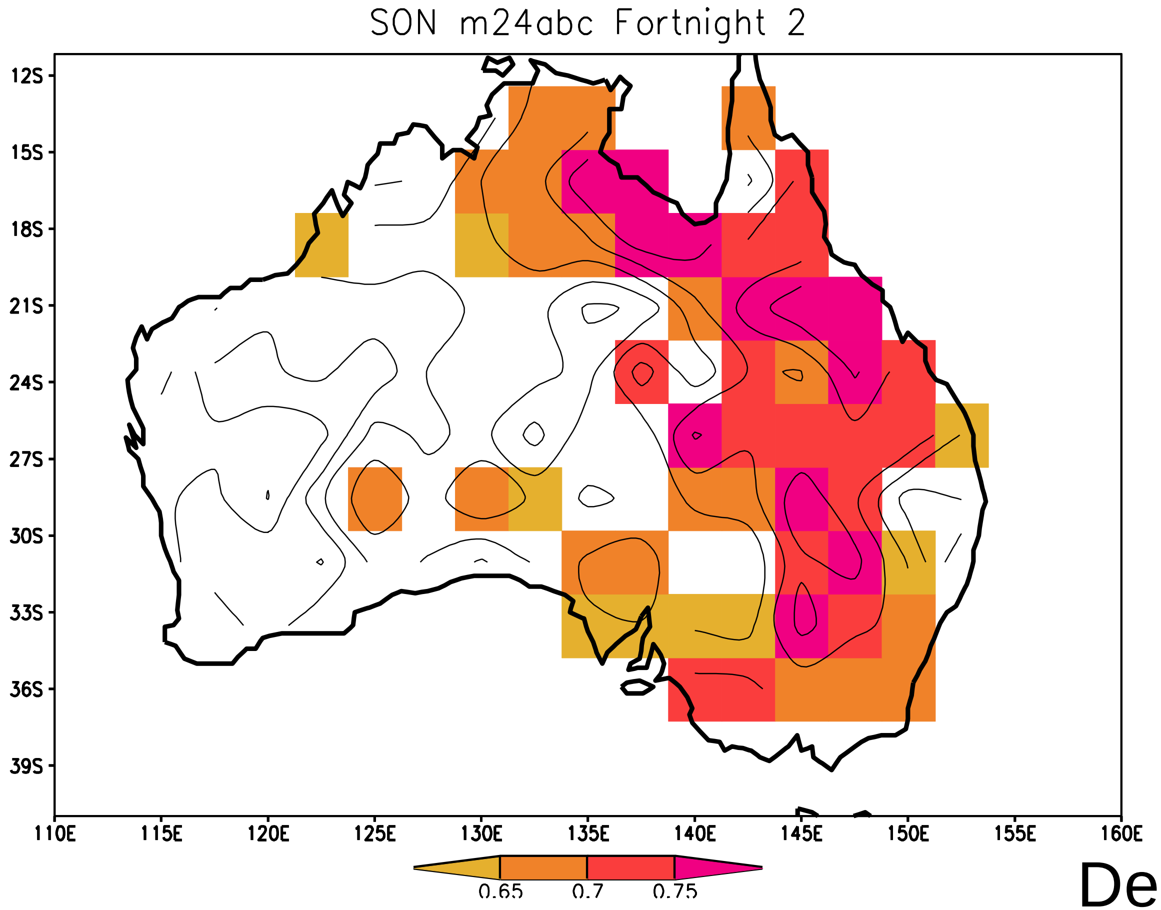




Frederic Vitard and Laura Ferranti, ECMWF

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

Brier Score decomposition (Murphy, 1973)

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

Murphy A. H., 1973: A New Vector Partition of the Probability Score. J. of App. Meteorol. and Climatol. 12(4), 595-600.

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

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

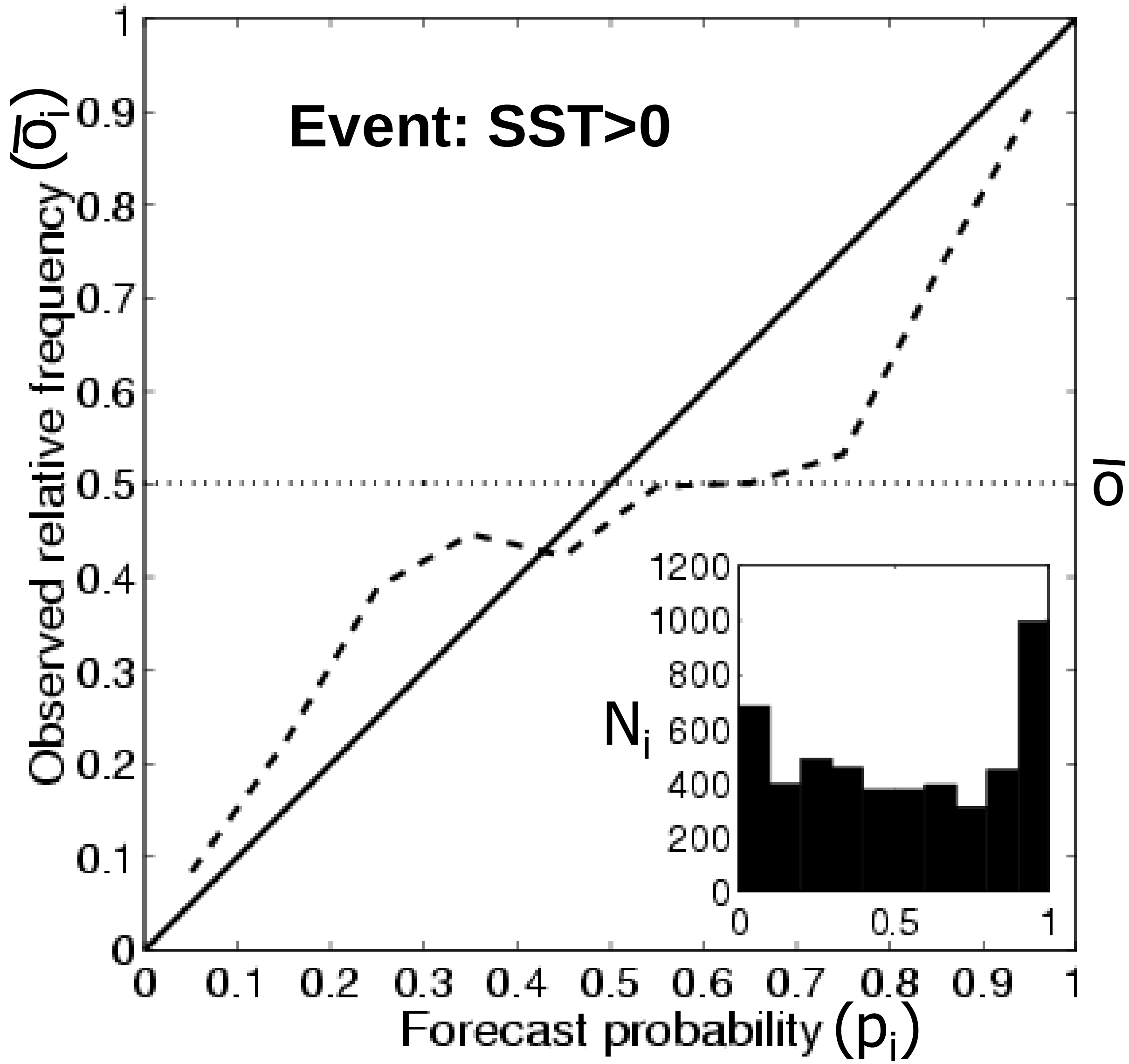
$$i = 1,\ldots, l = 11 : p_1 = 0, p_2 = 0.1, p_3 = 0.2, \ldots, p_{10} = 0.9, p_{11} = 1$$

p_k: forecast probabilities

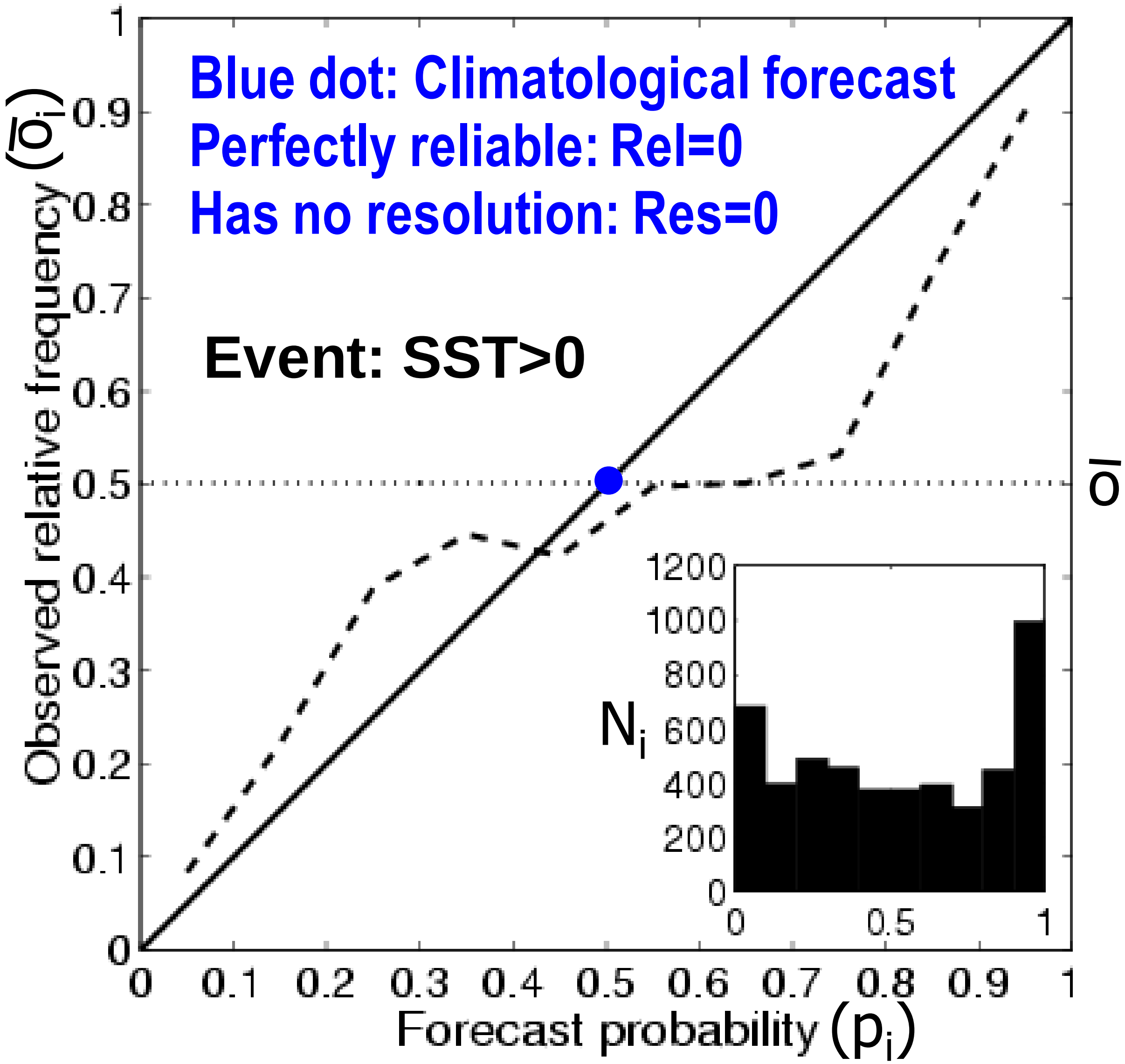
o_k: binary observations

n: number of (p_k, o_k) pairs

Reliability diagram



Reliability diagram

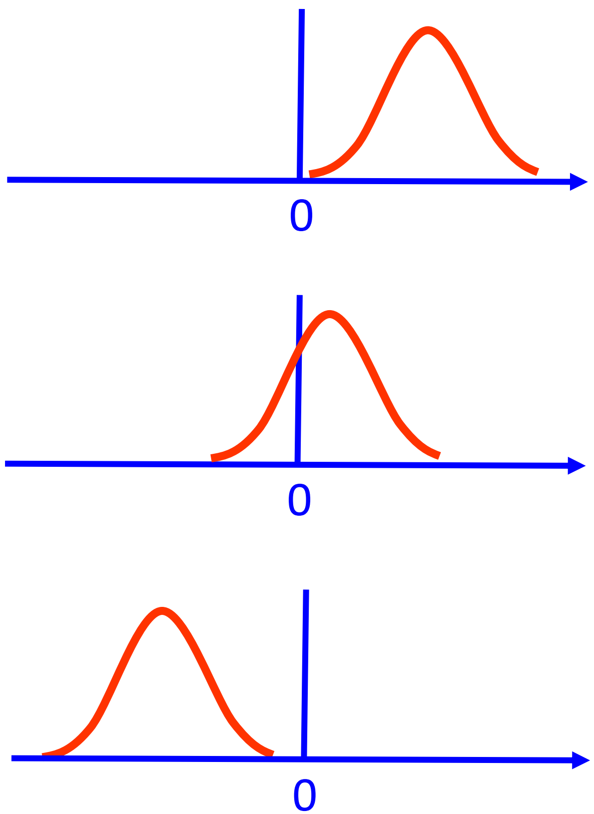


Example of how to construct a reliability diagram

Sample of probability forecasts:

22 years x 3000 grid points = 66000 forecasts

How often the event (T>0) was forecast with probability p_i?



Forecast Prob.(p_i)	# Fcsts. N_i	"Perfect fcst." OBS-Freq.($\overline{o}_i$)	"Real fcst." OBS-Freq($\overline{o}_i$)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
….	….	….	….
….	….	….	….
….	….	….	….
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

Courtesy: Francisco Doblas-Reyes

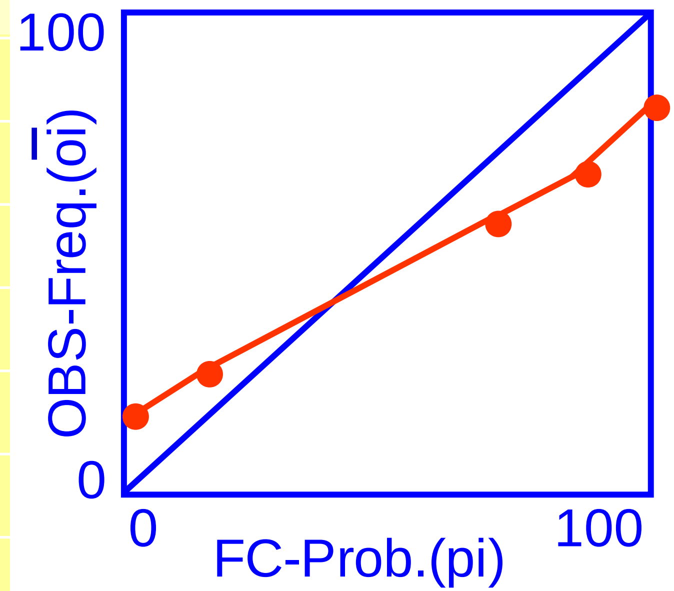
Example of how to construct a reliability diagram

Sample of probability forecasts:

22 years x 3000 grid points = 66000 forecasts

How often the event (T>0) was forecast with probability p_i?

Forecast Prob.(p_i)	# Fcsts. N_i	"Perfect fcst." OBS-Freq.($\bar{o}_i$)	"Real fcst." OBS-Freq($\bar{o}_i$)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
….	….	….	….
….	….	….	….
….	….	….	….
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)



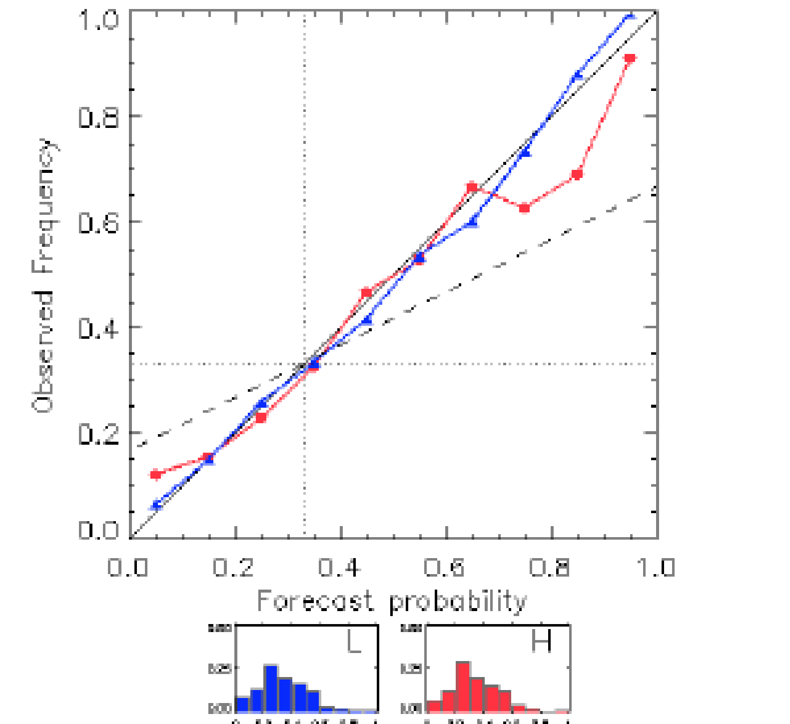
Courtesy: Francisco Doblas-Reyes

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

GLOSEA5 Hindcast Probabilistic skill

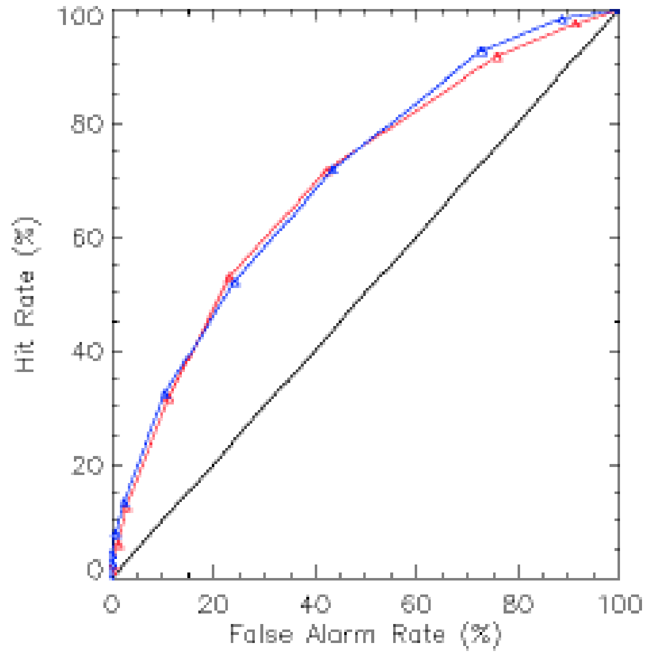
MSLP in N. Atlantic in upper and lower tercile

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

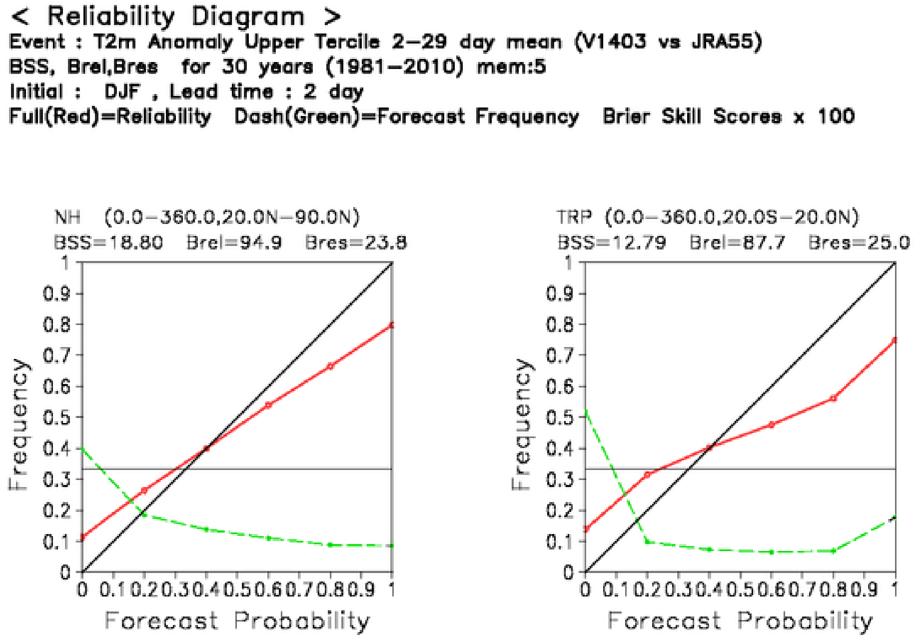


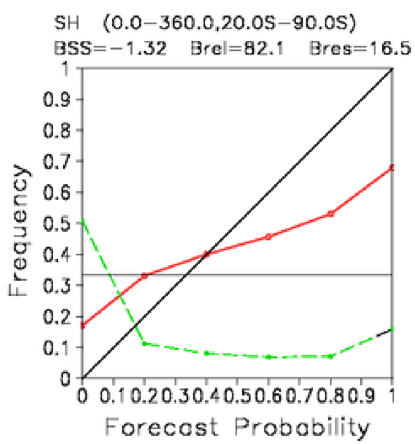
(b) Relative Operating Characteristics (ROC) diagram for the mean sea level pressure in GloSea5 over the North Atlantic. The red line shows the upper tercile and the blue line is the lower tercile.

Figure 6. Statistical scores for the Northern Atlantic region.

MacLachlan et al., QJRMS, 2015

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



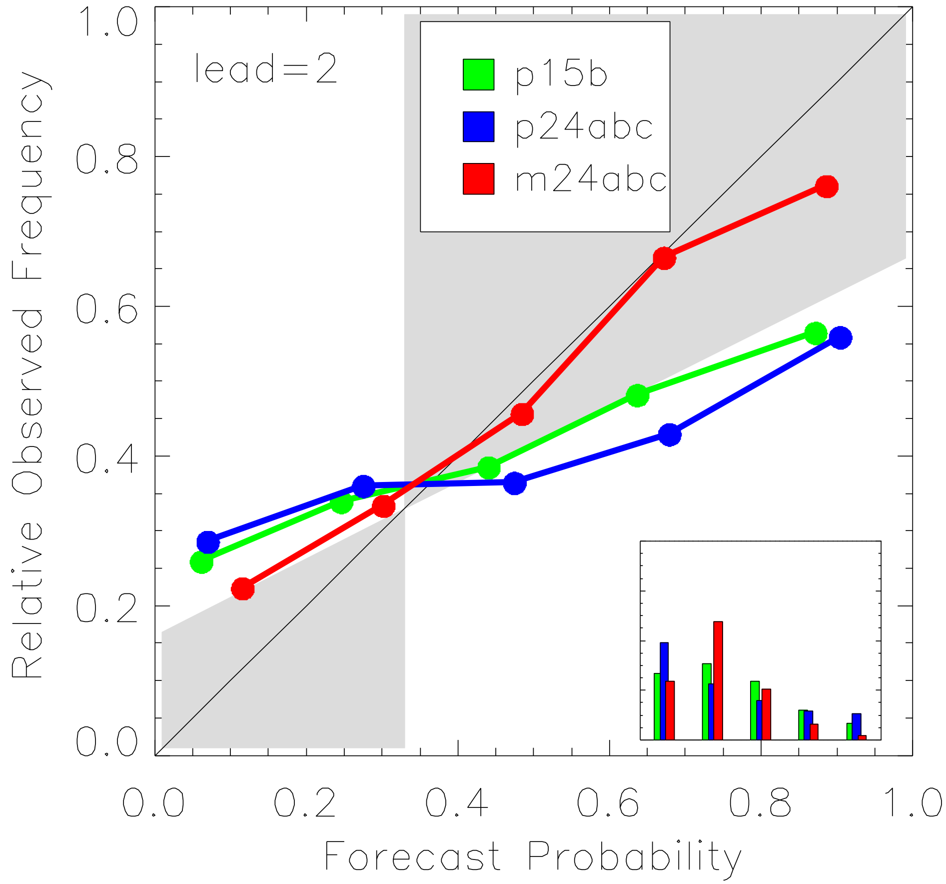


Reliability Diagrams
T2m (upper tercile)
Day 2-29 mean
I.C. : Dec.-Feb.
1981-2010
N.H., TROP, S.H.

Yuhei Takaya, JMA

Two weeks forecast example: ½ month lead precip. fcsts

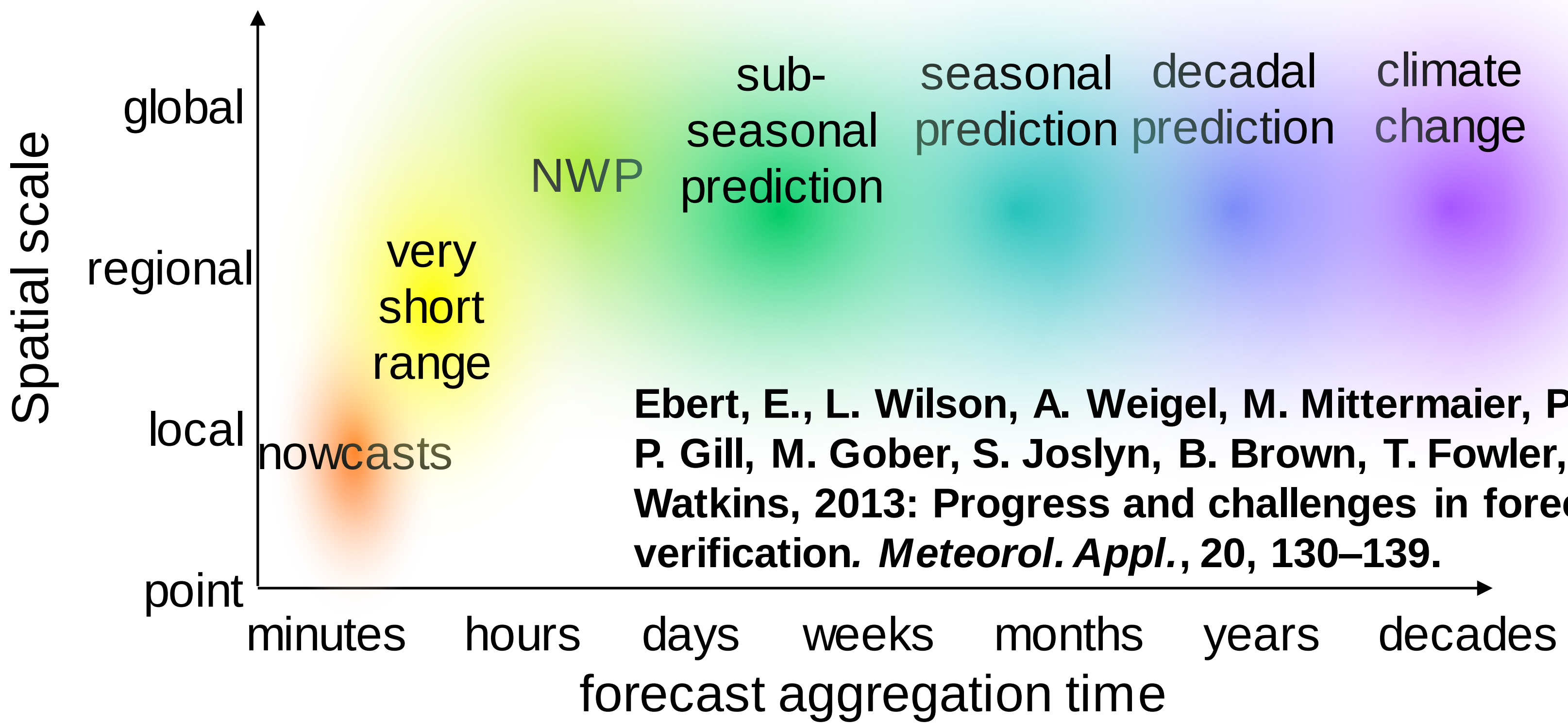
Precipitation anomalies in the upper tercile
Fortnight 2: Sep, Oct, Nov forecast start months. Hindcasts: 1980-2006



Debbie Hudson
BOM, Australia

Seamless verification

Seamless forecasts - consistent across space/time scales
single modelling system or blended
probabilistic / ensemble



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

- 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 assessment

- Sub-seasonal to seasonal verification is naturally leaning towards the seamless consistency concept addressing the question of which scales and phenomena are predictable

- As sub-seasonal to seasonal covers various forecast ranges (days, weeks and months) it naturally allows seamless verification developments

Additional references

• Mason, S, 2018: WMO **Guidance on Verification of Operational Seasonal Climate Forecasts**.

• Coelho CAS, Brown B, Wilson L, Mittermaier M, Casati B, 2019: **Forecast verification for S2S time scales**. In: Robertson AW, Vitart F (eds). Sub-seasonal to seasonal prediction: the gap between weather and climate forecasting, Book Chap. 17, 1st edn. Elsevier, Amsterdam, pp 337–361. (ISBN: 9780128117149. eBook ISBN: 9780128117156)

Thank you all for your attention!