Australian Government
Bureau of Meteorology

BOM Climate Forecast Verification Strategy and Metrics

William Wang
Long Range Forecast, Bureau of Meteorology

NESP Project 2.3 / CSC Decadal Prediction Workshop
15/05/2018

Outline

- A recall of basic concept of verification
- BOM climate forecasts
- Verification strategy, metrics and comparison
- Some discussion of challenges

Why verify?

- A forecast is not complete until you find out how successful the forecast is (BOM service policy)
- The three most important reasons to verify forecasts:
  1. to ***monitor*** forecast quality - how accurate are the forecasts?
  2. to ***improve*** forecast quality - the first step toward getting better is discovering what you're doing wrong.
  3. to ***compare*** the quality of different forecast systems - to what extent does one forecast system give better forecasts than another, and in what ways is that system better? – hence to justify model/system upgrade.
- BOM Climate forecast and verification started from 1989

What makes a forecast "good"?

- **Consistency**
  - – Forecast system should be properly developed and stable; avoiding jumping from one system to another
- **Quality**
  - – the degree to which the forecast corresponds to what actually happened
- **Value**
  - – the degree to which the forecast helps a decision maker to realize some incremental economic and/or other benefit.

QUALITY aspects of forecasts?

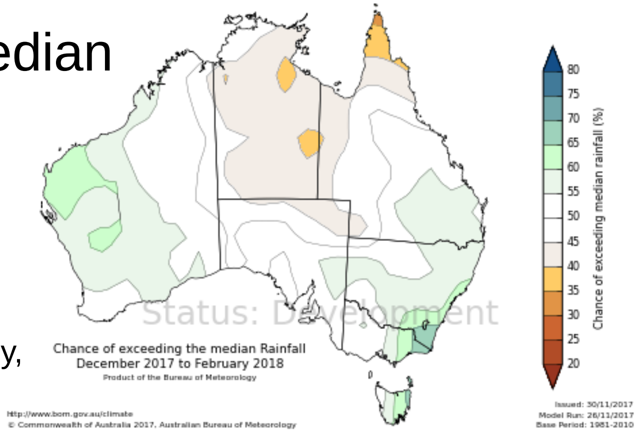
- **Bias** - a systematic error between the mean forecast and mean observation.
- **Association** - the strength of the linear relationship between the forecasts and observations
- **Accuracy** - the level of agreement between the forecast and the truth
- **Skill** - refers to the increase in accuracy due purely to the "smarts" of the forecast system.
- **Reliability** - measures how close the forecast probabilities are to the true frequency
- **Resolution** - the distribution of outcomes when "A" was forecast is different from the distribution of outcomes when "B" is forecast
- **Sharpness** - the tendency of the forecast to predict extreme values. To use a counter-example, a forecast of "climatology" has no sharpness.
- **Discrimination** - ability of the forecast to discriminate among observations, that is, to have a higher prediction frequency for an outcome whenever that outcome occurs.
- **Uncertainty** - the variability of the observations. The greater the uncertainty, the more difficult the forecast will tend to be.

Verification strategy of forecasts QUALITY

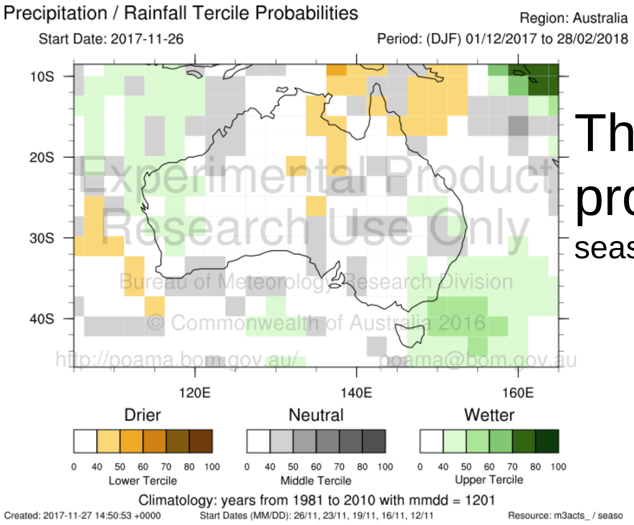
- **Forecast**
  - Various types of forecasts, such as above median probability forecast
- **Truth**
  - data that we use to verify a forecasts generally comes from observational data, for example AWAP.
- **What aspects of quality we want to know**
  - any of the qualities listed, such as percent consistency as skill and reliability
- **Metric**
  - the mathematics and diagram designed to measure the specific forecast quality
- **Validity of verification results**
  - trustworthy when the quantity and quality of the verification data are high, use of some error bounds on the verification results themselves
  - Use of hindcast skill assessment, skill may vary from one period to another, real time skill may be different from hindcast skill

Types of forecasts to be verified

Probability above median forecast (Monthly, seasonal)



Status: Development
Chance of exceeding the median Rainfall
December 2017 to February 2018
Product of the Bureau of Meteorology
http://www.bom.gov.au/climate
© Commonwealth of Australia 2017, Australian Bureau of Meteorology
80
75
70
65
60
55
50
45
40
35
30
25
20
Chance of exceeding median rainfall (%)
Issued: 30/11/2017
Model Run: 26/11/2017
Base Period: 1981-2010

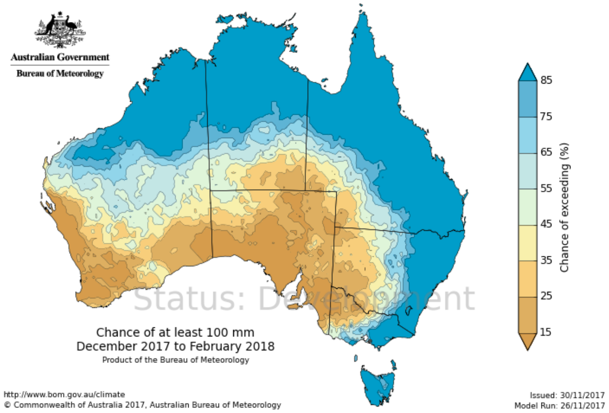


Precipitation / Rainfall Tercile Probabilities
Region: Australia
Start Date: 2017-11-26
Period: (DJF) 01/12/2017 to 28/02/2018
10S
20S
30S
40S
Experimental Product
Research Use Only
Bureau of Meteorology Research Division
© Commonwealth of Australia 2016
http://poama.bom.gov.au/
poama@bom.gov.au
120E
140E
160E
Drier
Neutral
Wetter
0 40 50 60 70 80 100
Lower Tercile
0 40 50 60 70 80 100
Middle Tercile
0 40 50 60 70 80 100
Upper Tercile
Climatology: years from 1981 to 2010 with mmdd = 1201
Created: 2017-11-27 14:50:53 +0000
Start Dates (MM/DD): 26/11, 23/11, 19/11, 16/11, 12/11
Resource: m3acts_ / seaso

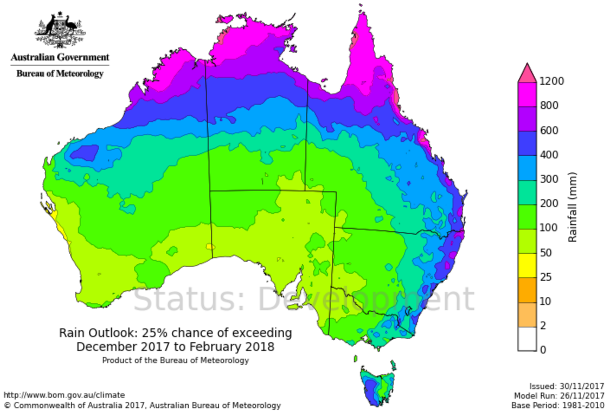
Three category/tercile probability forecast (Monthly, seasonal)

Probability of exceedance

25% of Outlook scenarios



Australian Government
Bureau of Meteorology
Status: Development
Chance of at least 100 mm
December 2017 to February 2018
Product of the Bureau of Meteorology
http://www.bom.gov.au/climate
© Commonwealth of Australia 2017, Australian Bureau of Meteorology
85
75
65
55
45
35
25
15
Chance of exceeding (%)
Issued: 30/11/2017
Model Run: 26/11/2017



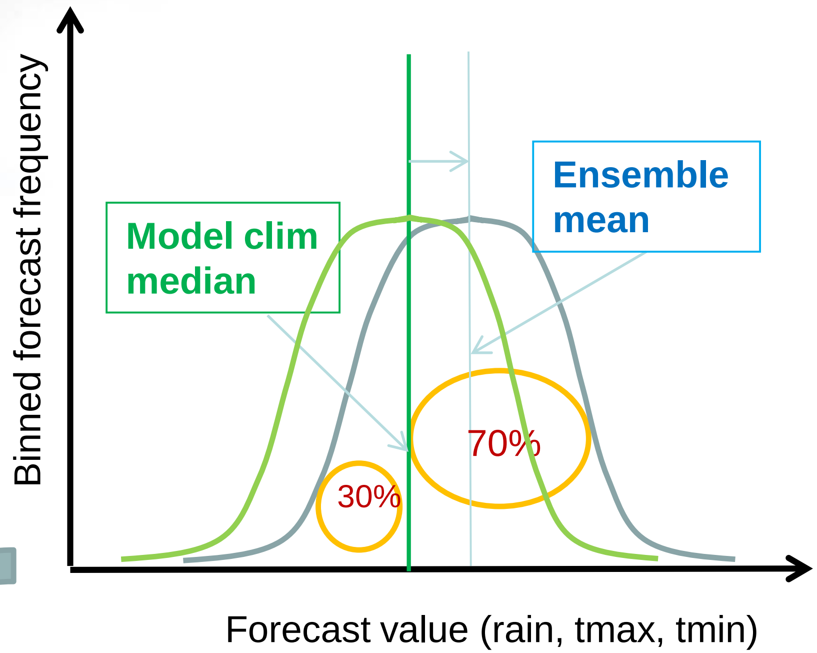
Australian Government
Bureau of Meteorology
Status: Development
Rain Outlook: 25% chance of exceeding
December 2017 to February 2018
Product of the Bureau of Meteorology
http://www.bom.gov.au/climate
© Commonwealth of Australia 2017, Australian Bureau of Meteorology
1200
800
600
400
300
200
100
50
25
10
2
0
Rainfall (mm)
Issued: 30/11/2017
Model Run: 26/11/2017
Base Period: 1981-2010

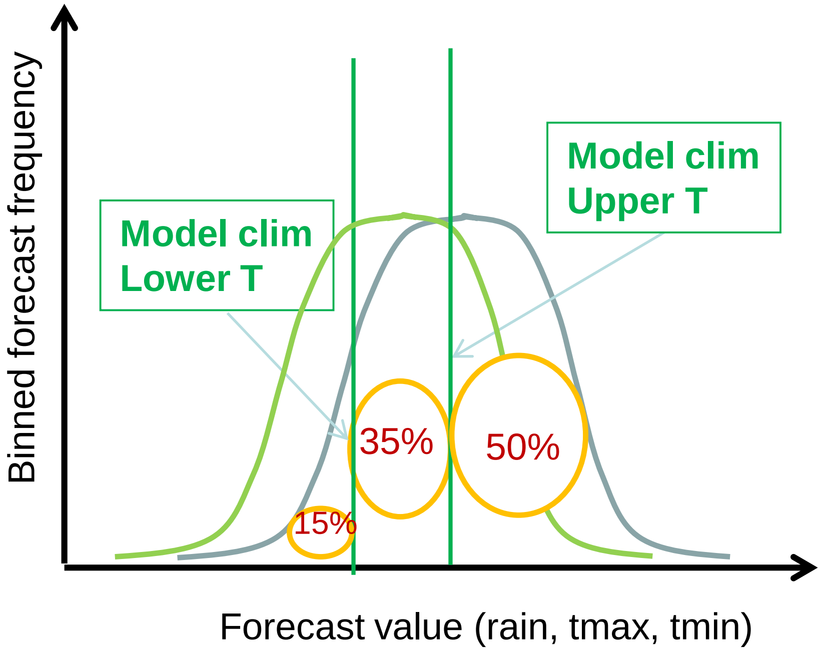
Probability of exceedance (POE)

(Monthly and seasonal)

Real time Northern Rainfall Onset, Nino Indices and IOD forecasts only get visual verification. Proper verification metrics need to be developed.

Making of probability above median and tercile forecast from model output





Forecast: 70% probability above median

Forecast: T1 15% T2 35% T3 50%

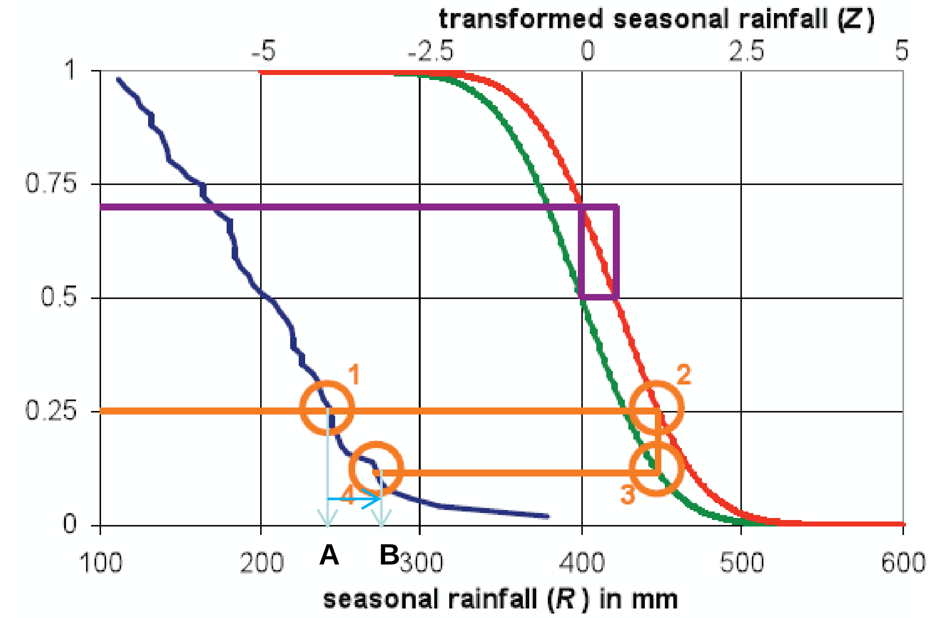
Ensemble size

Monthly: 99 Seasonal: 165 Weekly:33 Fortnightly:33

Making of POE forecast
from probability above median forecast

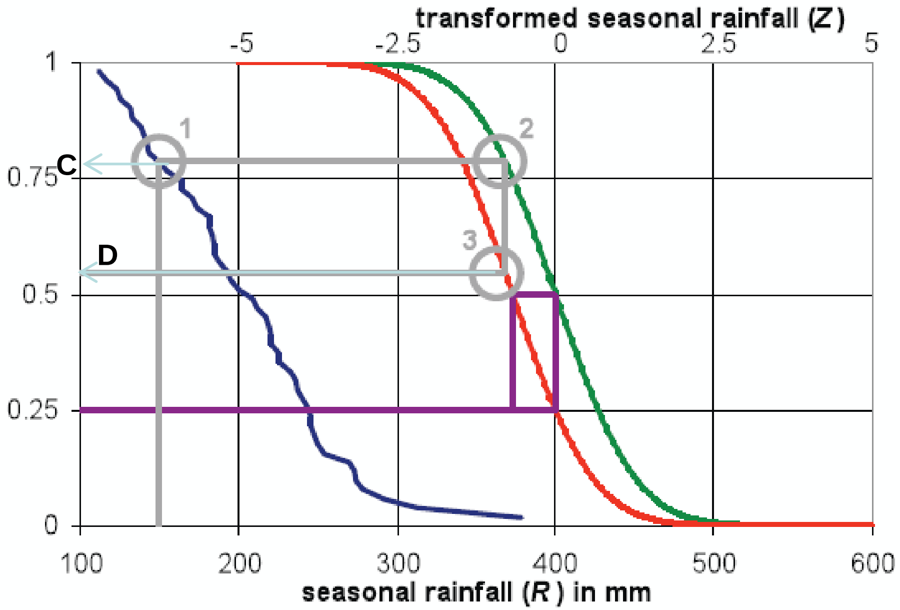
The making and verification of these forecasts were done using the scheme presented by http://cmap.bom.gov.au/papers/bamos-paper01.pdf, presuming variation of monthly/seasonal rainfall total at each grid fit normal distribution. Hence above median probability forecast can be converted to the POE forecast

Three scenarios: 25, 50 & 75%



For 70% above median forecast, rain with 25% probability increased from A to B

12 set totals: 10 25 50 100 150 200 250 300 400 500 600 700



For drier forecast, probability of 150mm rain decreased from C to D

Metrics of verification and operational verification strategy

- **Accuracy (Percent consistency)**
- **Weighted Percent Consistency (WPC)**
- Probability of detection (hit rate)
- False alarm ratio
- Probability of false detection (false alarm rate)
- Threat score (critical success index)
- Hanssen and Kuipers discriminant (true skill statistic, Peirce's skill score)
- Heidke skill score
- Mean absolute error
- Root mean square error
- **Linear error in probability space (LEPS)**
- **Anomaly correlation (AC)**
- **Brier (Skill) score**
- **Reliability diagram**
- **Relative Operating Characteristic (ROC)**
- **Ranked Probability Skill score (RPSS)**
- … …

Percent consistency/accuracy

Observed

Forecast

	yes	no	total
Yes >50%	hits	False alarms	Forecast yes
No <=50%	misses	Correct negatives	Forecast no
total	Observed yes	Observed no	total

$$Accuracy = \frac{hits + correct\ \ negatives}{total}$$

Observed

Forecast

	L	M	H	total
L >33%	hits	misses	misses	Fcst L
M >33%	misses	hits	misses	Fcst M
H >33%	misses	misses	hits	Fcst H
Total	Obs L	Obs M	Obs H	total

$$Accuracy = \frac{\Sigma(hits)}{\Sigma(hits)+\Sigma(misses)}$$

Answers the question: Overall, what fraction of the forecasts were correct?
Range: 0 to 1. **Perfect score:** 1.
Reference level: 0.5 for above median and 1/3 for tercile forecast
Characteristics: Simple, intuitive. Can be misleading since it is heavily influenced by the most common category (for example desert climate), usually "no event" in the case of rare weather. Can't further quantify how good/bad a probability forecast is.
Shortfall: Probability was converted to deterministic "yes/no" or "L/M/H" for such verification metric to be applied. The true value of forecast is compromised

Weighted Percent Consistency

(two category)

$$PC = \sum(hit\ or\ correct\ negative)\ /total$$

$$WPC = \sum((hit\ or\ correct\ negative) * abs(anomaly))\ /\sum abs(anomaly))$$

PC is a categorical focused metric with abnormality of variation not considered.
The underpinning idea of WPC:

> The skilful climate forecasts are expected to be capable of capturing the significant and predictable part of the meteorological/climate variation
>
> **** Anomaly is observed above median anormaly***

1. For a single forecast at a station or grid (i,j)
   $WPC(i,j) = hitrate(i,j) = PC(i,j)$
   It becomes 0 for a miss or 1 for a hit or correct negative, just like PC

2. Its national average (overall skill) is
   $WPC = \sum_{i,j}(hitrate(i,j) * abs(anomaly(i,j))\ /\sum_{i,j} abs(anomaly(i,j))$

3. For a series of forecasts, the overall spatial WPC is
   $WPC(i,j) = \sum_{t}(hitrate(i,j,t) * abs(anomaly(i,j,t))\ /\sum_{t} abs(anomaly(i,j,t))$

4. The overall national average for a series of forecasts is then
   $WPC = \sum_{t}\sum_{i,j}(hitrate(i,j,t) * abs(anomaly(i,j,t))\ /\sum_{t}\sum_{i,j} abs(anomaly(i,j,t))$

Characteristics of WPC

1. Range of WPC is 0 → 1, same as PC

2. 0.5 is the "no skill" benchmark or reference skill obtained when climatological median being used as forecasts, which is also the situation of PC

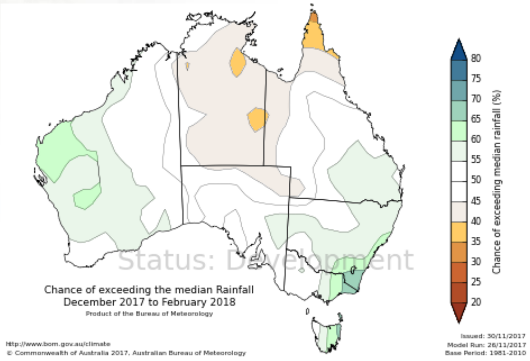
   In hindcast, using climatological median as forecasts makes half forecasts correct and half wrong—the sum of the absolute values of anomalies which are forecasted correctly should be about the same as that of those forecasted incorrectly, provided the sample size is large enough, even if the distribution of the predictand is skewed. This is because there is no reason that the correctly forecasted half cases should be wetter or warmer than the other wrongly forecasted half cases under hindcast circumstances. Hence, the corresponding benchmark WPC should be or very close to 0.5 or 50%

3. WPC = 1 or 100% the best forecasts can get.

Australian Government
Bureau of Meteorology

Application of Accuracy, Percent Consistency, and WPC

(for single forecast)

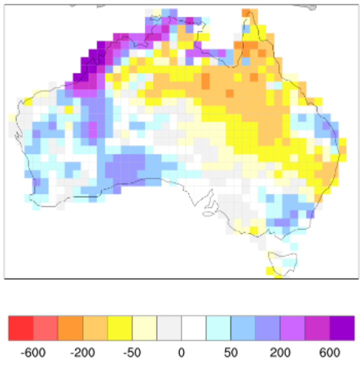


Chance of exceeding the median Rainfall
December 2017 to February 2018
Product of the Bureau of Meteorology

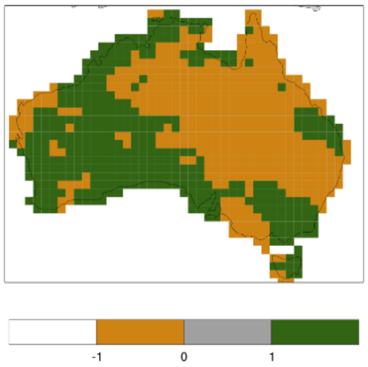
Http://www.bom.gov.au/climate
© Commonwealth of Australia 2017, Australian Bureau of Meteorology

Issued: 30/11/2017
Model Run: 26/11/2017
Base Period: 1981-2010

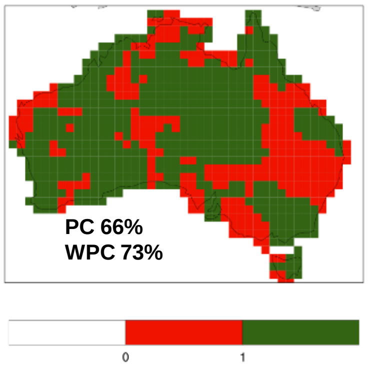
Anomaly above median for 2017.12-02 rain



above median outcome for 2017.12-02 rain

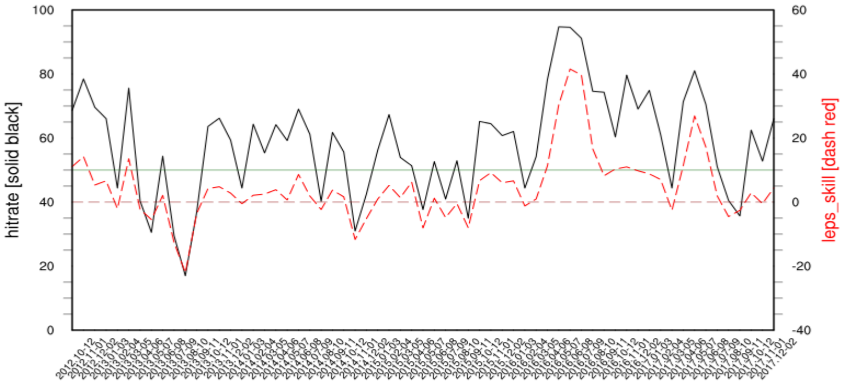


hitrate for 2017.12-02 rain



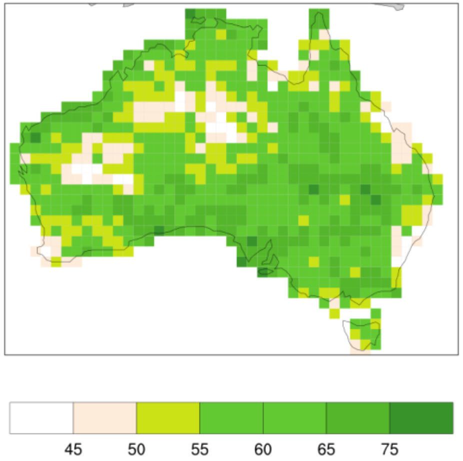
PC 66%
WPC 73%

National averaged hitrate/leps_skill for seasonal rain median forecast



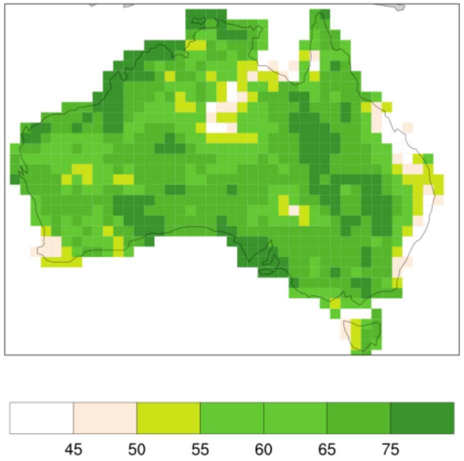
Overall percent consistency

hitrate for 63 seasonal rain median forecast



Overall weighted percent consistency

hitrate for 63 seasonal rain median forecast



LEPS Calculation formulas for probability above median and terciles forecast

- **For above median probability forecast**

LEPS SKILL = $\frac{1}{N}\sum_{i=1}^{N} sign(i)(2P(i)/100.-1)$

Here sign(i) equals "1" if the outcome is above median; "-1" otherwise. P(i) is forecast probability

Ranges from -1 for worst forecast to 1 for a perfect forecast

- **For Tercile probability forecast**

LEPS SKILL = $\sum_{i=1}^{N} s_i / \sum_{i=1}^{N} u_i$ if $\sum_{i=1}^{N} s_i \geq 0$

LEPS SKILL = $\sum_{i=1}^{N} s_i / \sum_{i=1}^{N} l_i$ if $\sum_{i=1}^{N} s_i \leq 0$

Ranges from -1 for worst forecast to 1 for a perfect forecast

Here s_i=8/27 p_1 − 1/27 p_2 − 7/27 p_3 ; if outcome is tercile 1

= -1/27 p_1 + 2/27 p_2 − 1/27 p_3 ; if outcome is tercile 2

= -7/27 p_1 − 1/27 p_2 $+8/27$ p_3 ; if outcome is tercile 3

$p_1 + p_2 + p_3 = 1$

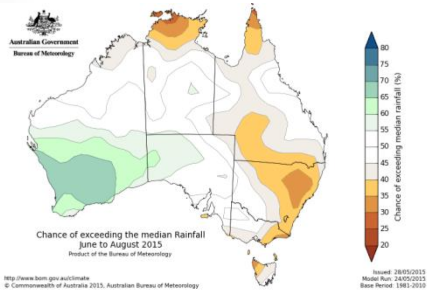
$u_i = \frac{8}{27}$ and $l_i = -\frac{7}{27}$ if outcome for the ith forecast is tercile 1 or 3;

$u_i = \frac{2}{27}$ and $l_i = -\frac{1}{27}$ if outcome for the ith forecast is tercile 2.

Fawcett, R. J. B., etc. , 2005: A verification of publicly issued seasonal forecasts issued by the Australian Bureau of Meteorology: 1998-2003. Aust. Meteor. Mag. 54, 1-13

LEPS skill score application

(for single forecast)

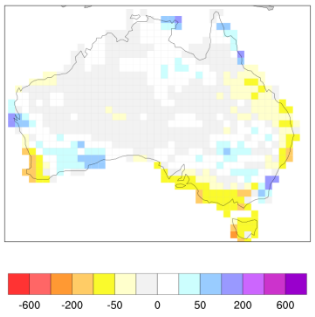


Chance of exceeding the median Rainfall
June to August 2015
Product of the Bureau of Meteorology

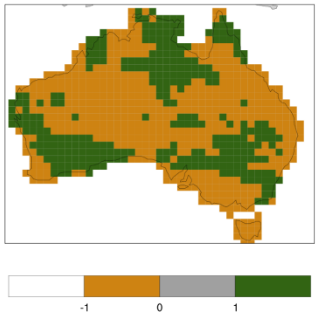
http://www.bom.gov.au/climate
© Commonwealth of Australia 2015, Australian Bureau of Meteorology

Issued: 28/05/2015
Model Run: 24/05/2015
Base Period: 1981-2010

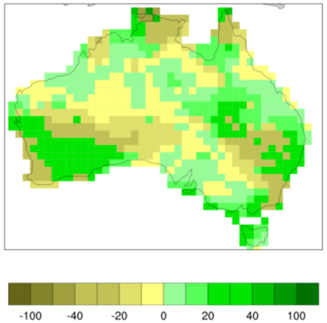
Anomaly above median for 2015.06-08 rain



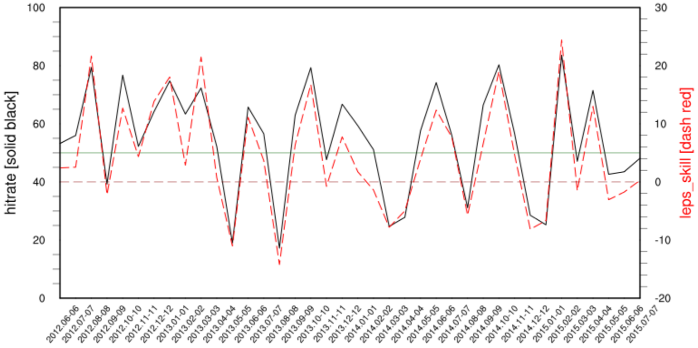
above median outcome for 2015.06-08 rain



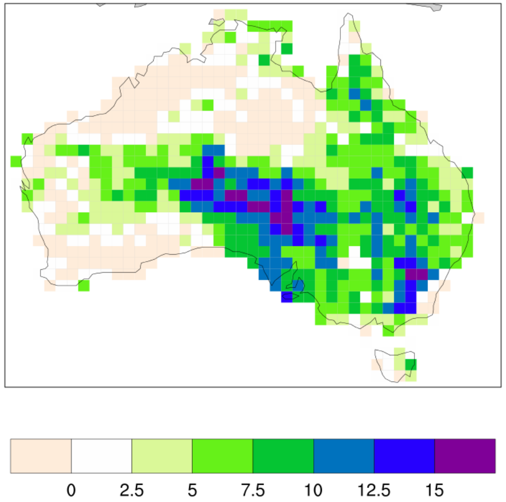
leps2 skill for 2015.06-08 rain



National averaged hitrate/leps_skill for monthly rain median forecast



leps_skill for 38 monthly rain median forecast



Discrete Ranked Probability Skill Score Calculation for POE forecast

DRPSS = (1 – (S_{prob})/(S_{ref}))*100%

For 12 category exceedance probability

$$S_{prob} = \frac{1}{30}\sum_{n=1}^{n=30}\left[\frac{1}{12}\sum_{k=1}^{k=12}(F_{prob(k)}/100.-outcome(k))^2\right]$$

$$S_{ref} = \frac{1}{30}\sum_{n=1}^{n=30}\left[\frac{1}{12}\sum_{k=1}^{k=12}(F_{ref(k)}/100.-outcome(k))^2\right] \quad ;$$

here the $F_{ref(k)}$ is corresponding observed percentage of each rainfall category. Outcome can vary from lower categories to higher ones.

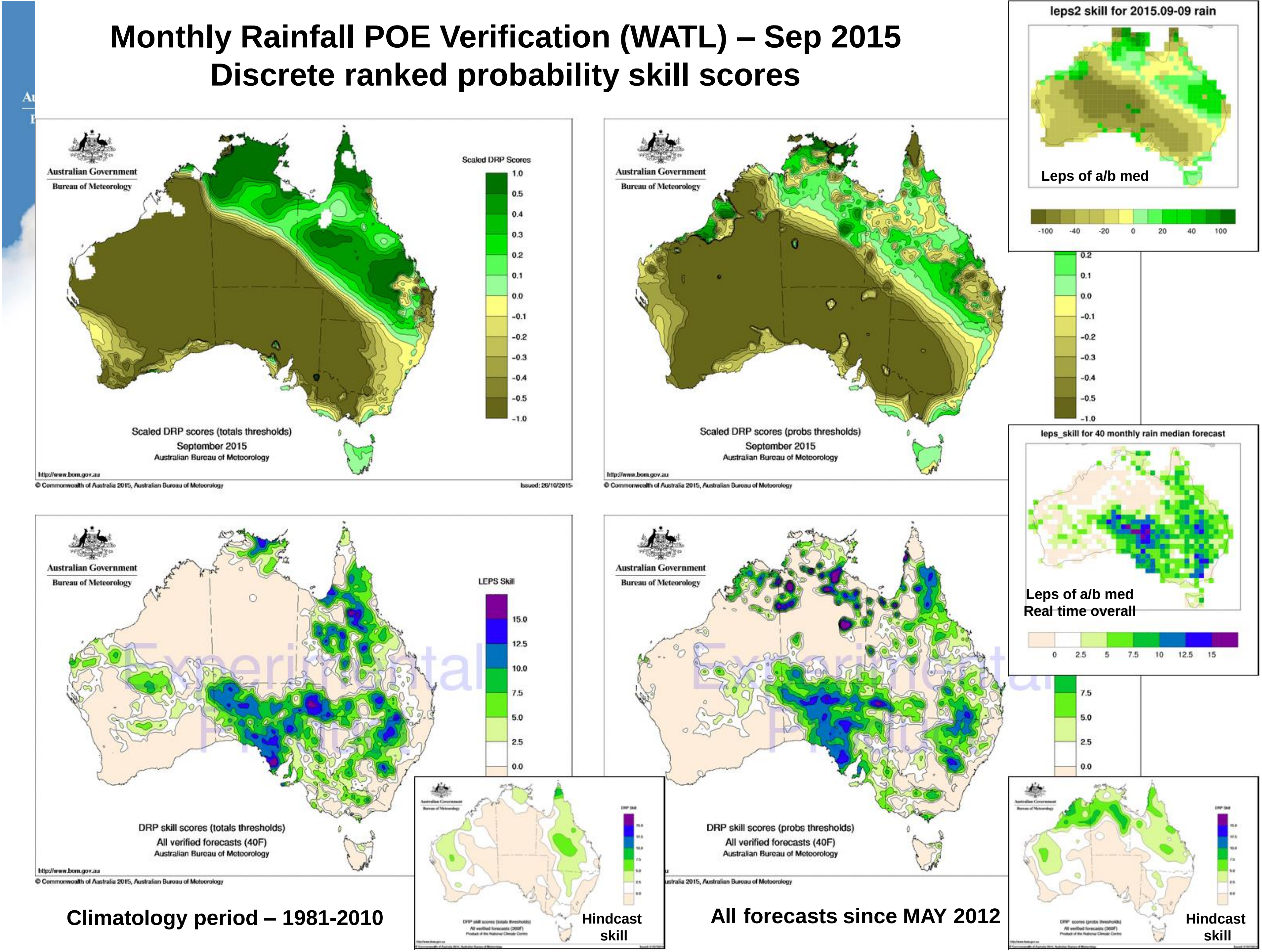
For 3 category chance of exceedance probability

$$S_{prob} = \frac{1}{30}\sum_{n=1}^{n=30}\left[\frac{1}{3}\sum_{k=1}^{k=3}(F_{prob(k)}/100.-outcome(k))^2\right]$$

$$S_{ref} = \frac{1}{30}\sum_{n=1}^{n=30}\left[\frac{1}{3}\sum_{k=1}^{k=3}(F_{ref(k)}/100.-outcome(k))^2\right]$$

here the $F_{ref(k)}$ is corresponding percentage of forecasted rainfall for each probability category.

Monthly Rainfall POE Verification (WATL) – Sep 2015
Discrete ranked probability skill scores


Australian Government
Bureau of Meteorology
Scaled DRP Scores
1.0
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-1.0
Scaled DRP scores (totals thresholds)
September 2015
Australian Bureau of Meteorology
http://www.bom.gov.au
© Commonwealth of Australia 2015, Australian Bureau of Meteorology
Issued: 26/10/2015
Scaled DRP scores (probs thresholds)
September 2015
Australian Bureau of Meteorology
leps2 skill for 2015.09-09 rain
Leps of a/b med
-100
-40
-20
0
20
40
100
leps_skill for 40 monthly rain median forecast
Leps of a/b med
Real time overall
0
2.5
5
7.5
10
12.5
15
LEPS Skill
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Experimental Product
DRP skill scores (totals thresholds)
All verified forecasts (40F)
Australian Bureau of Meteorology
DRP skill scores (probs thresholds)
All verified forecasts (40F)
Australian Bureau of Meteorology
DRP scores (totals thresholds)
All verified forecasts (300F)
DRP scores (probs thresholds)
All verified forecasts (300F)

Climatology period – 1981-2010

Hindcast skill

All forecasts since MAY 2012

Hindcast skill

Brier score and the application of reliability diagram

Definition of the Brier score

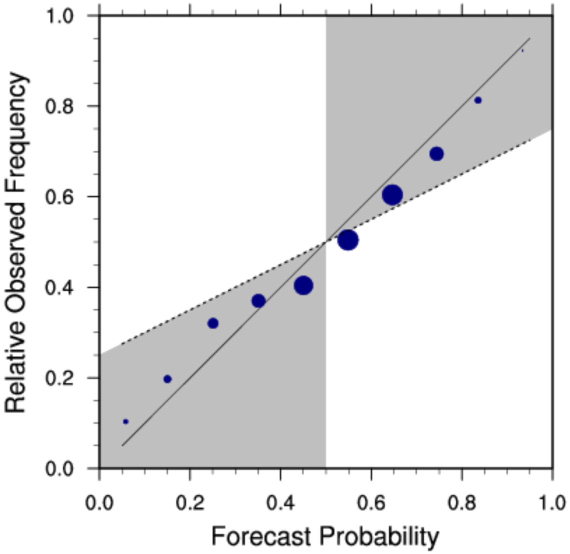
$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

3-component decomposition (grouped to K bins)

$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k(\mathbf{f_k} - \bar{\mathbf{o}}_\mathbf{k})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{\mathbf{o}}_\mathbf{k} - \bar{\mathbf{o}})^2 + \bar{\mathbf{o}}\,(1 - \bar{\mathbf{o}})$$

$$BS = REL - RES + UNC$$

Hence Brier score can be decomposed into 3 additive components: **Reliability, Resolution and Uncertainty.** (Murphy 1973)



- **Reliability:** forecast probability/number of forecasts in a bin fit the frequency of observation, small is better;
- **Resolution:** averages in forecast bins significantly different from overall average, larger is better
- **Uncertainty:** 2 category is most uncertain

$$Brier\ Skill\ Score = 1 - \frac{BS_f}{BS_r}$$

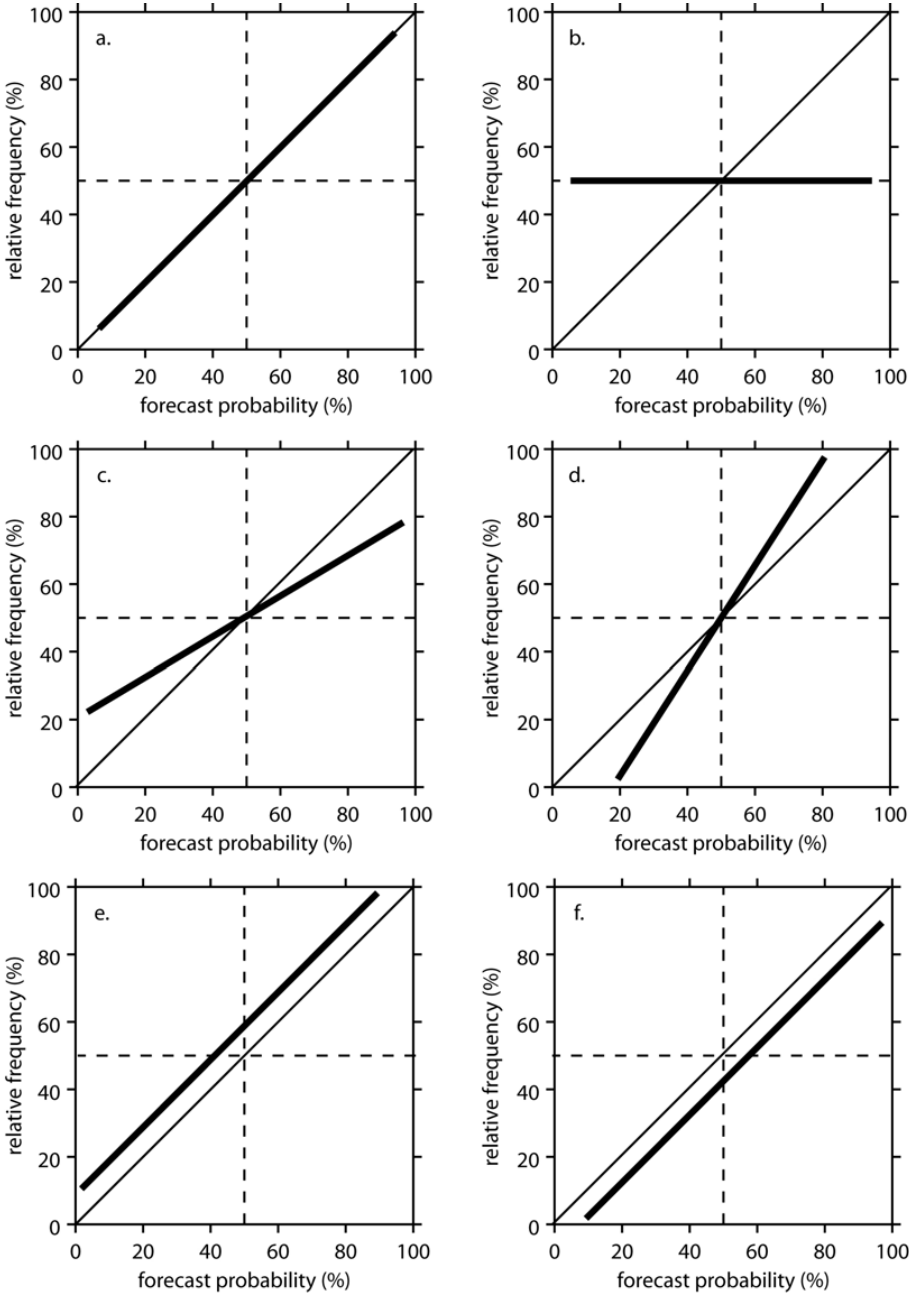
BS_f is forecast Brier Score, BS_r is Brier Score of reference forecast, from -∞ to 1

Reliability diagram Illustration

Idealized examples of reliability curves showing forecast systems with
(*a*) perfect reliability,
(*b*) no resolution,
(*c*) over-confidence,
(*d*) under-confidence,
(*e*) under-forecasting,
(*f*) over forecasting.

The horizontal dashed line indicates the observed relative frequency of the event for all forecasts, which is shown also as a dashed vertical line.

from Simon J. Mason, 2014: Guidance on Verification of Operational Seasonal Climate Forecasts, International Research Institute for Climate and Society

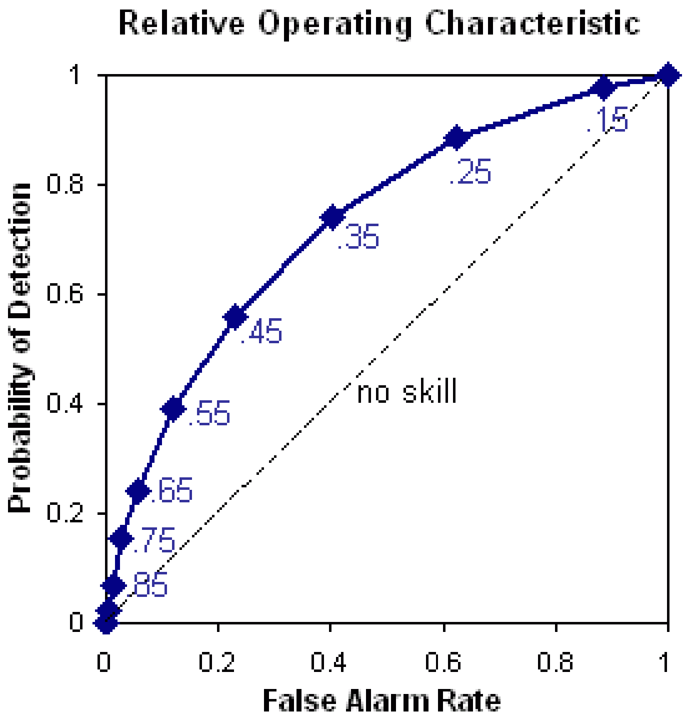


Relative operating characteristic (ROC)

- Plot *hit rate (POD)* vs *false alarm rate*

Probability of detection (hit rate) $POD = \frac{hits}{hits + misses}$ $HR_n = \sum_{i=n}^{N} O_i \Big/ \sum_{i=1}^{N} O_i$

False alarm ratio $FAR = \frac{false\ alarms}{hits + false\ alarms}$ $FAR_n = \sum_{i=n}^{N} NO_i \Big/ \sum_{i=1}^{N} NO_i$



Answers the question: The ability of the forecast to discriminate between events and non-events?

Range: 0 to 1. **Perfect score:** 1., 0.5 is no skill

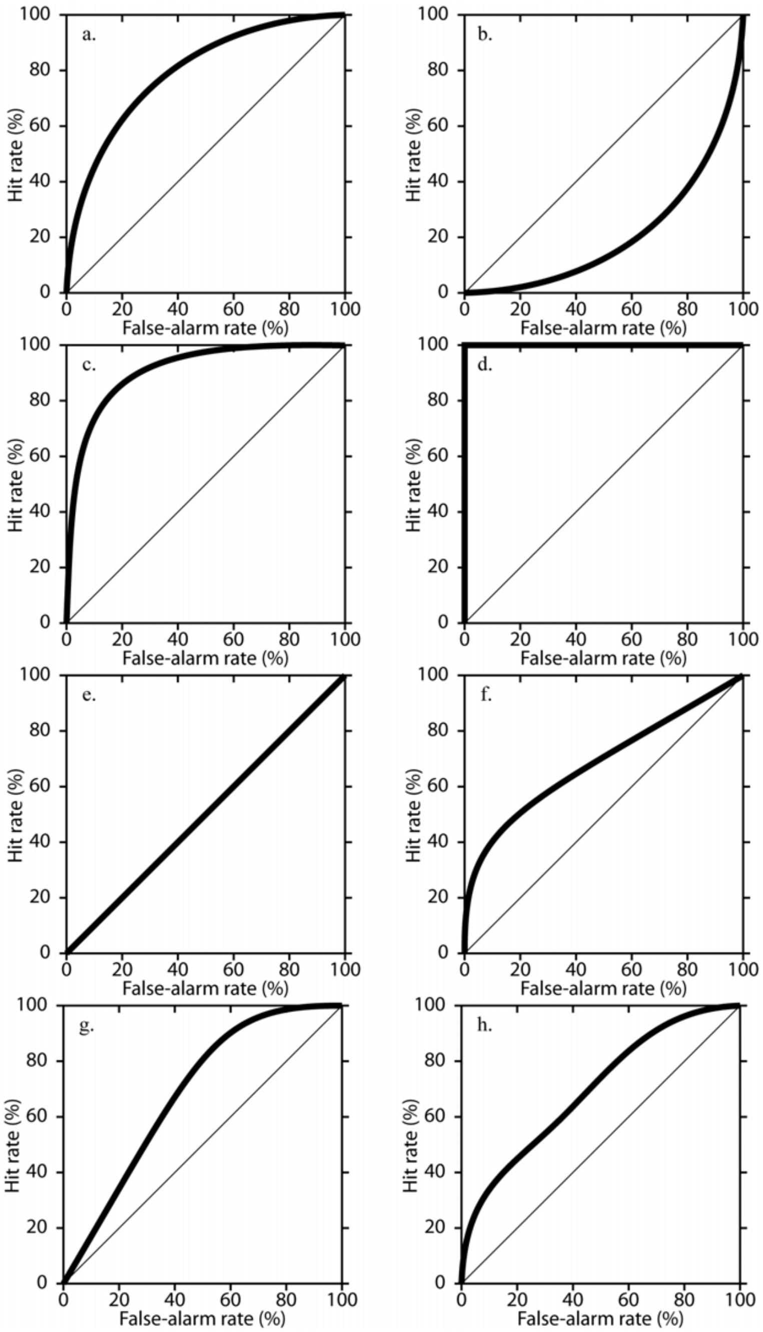
Characteristics: ROC measures the ability of the forecast to discriminate between two alternative outcomes, thus measuring resolution. It is not sensitive to bias in the forecast, so says nothing about reliability. A biased forecast may still have good resolution and produce a good ROC curve, which means that it may be possible to improve the forecast through calibration. The ROC can thus be considered as a measure of potential usefulness.

ROC diagram Illustration

Idealized examples of ROC curves showing forecast systems with

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

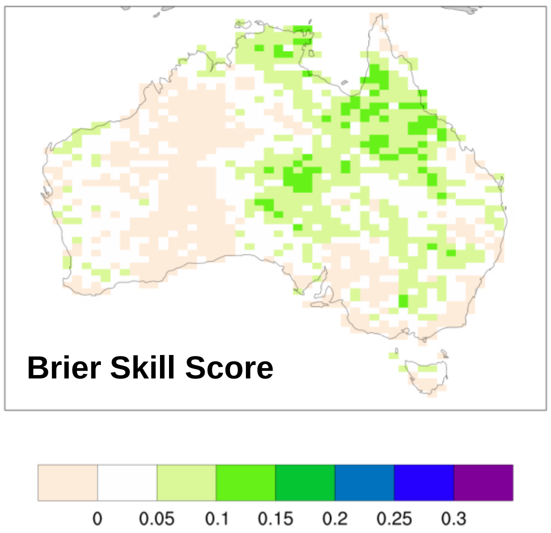
from Simon J. Mason, 2014: Guidance on Verification of Operational Seasonal Climate Forecasts, International Research Institute for Climate and Society



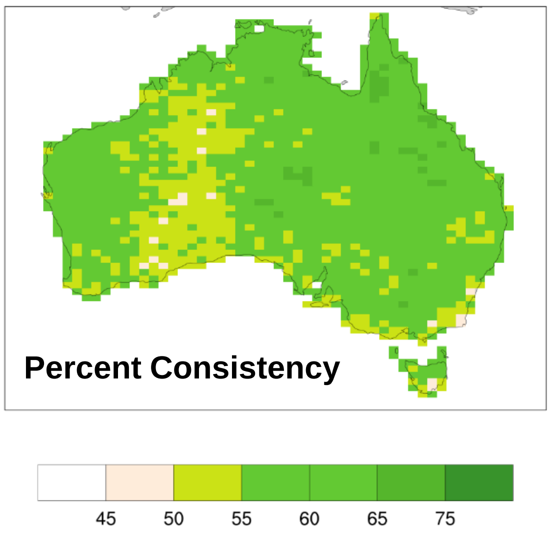
Australian Government
Bureau of Meteorology

ACCESS-S1 seasonal precipitation hindcast skills verified by various metrics

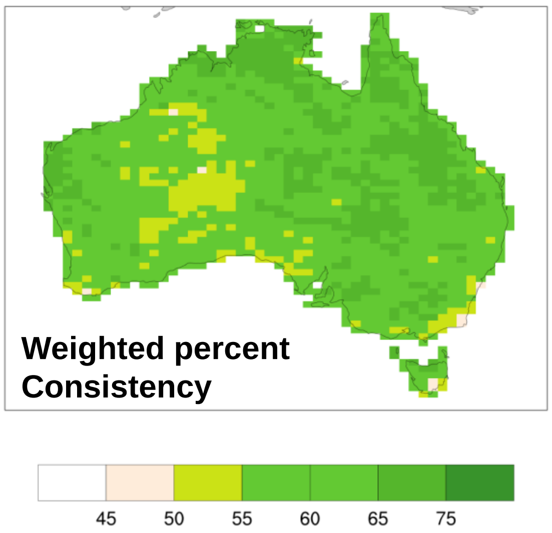
BSKILL for pr A/M forecast of ens 33 ld 1 inid 25



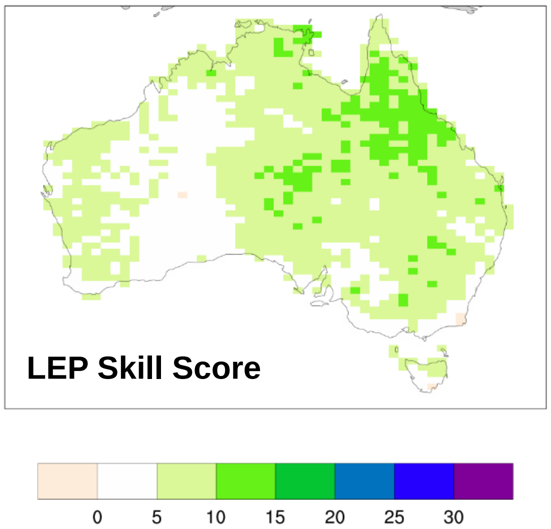
PC for pr A/M forecast of ens 33 ld 1 inid 25



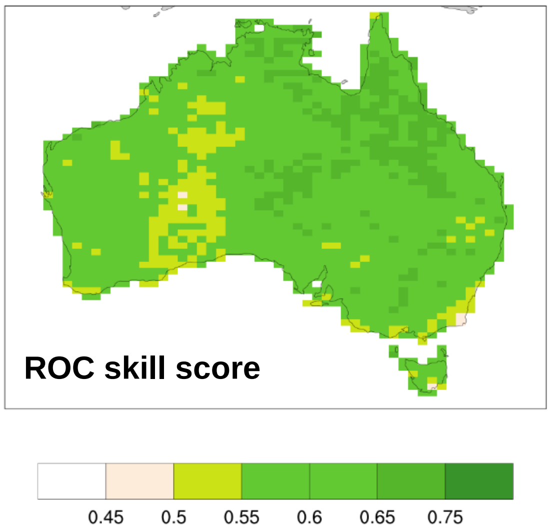
WPC for pr A/M forecast of ens 33 ld 1 inid 25



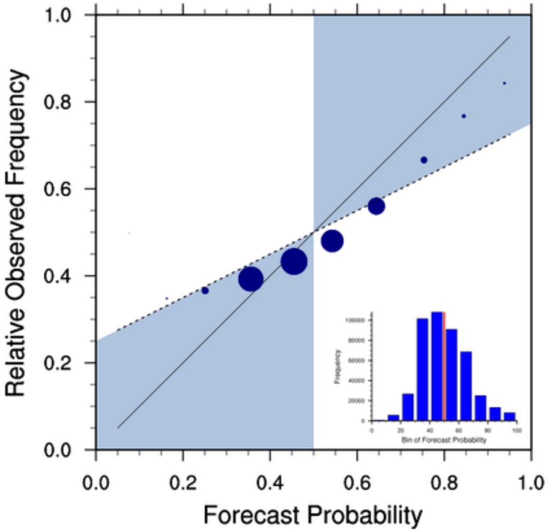
LEPS for pr A/M forecast of ens 33 ld 1 inid 25



ROCS for pr A/M forecast of ens 33 ld 1 inid 25

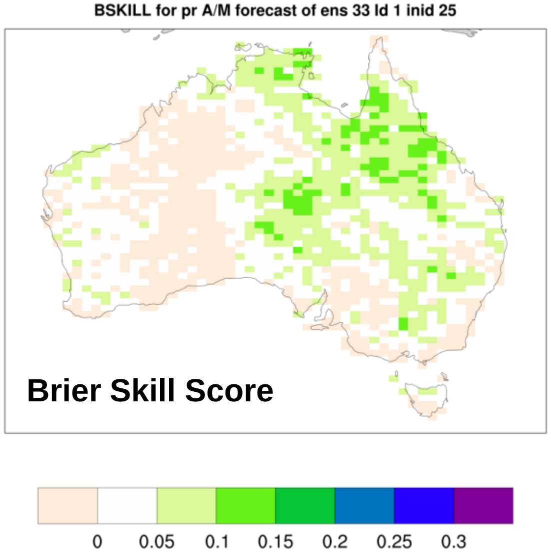


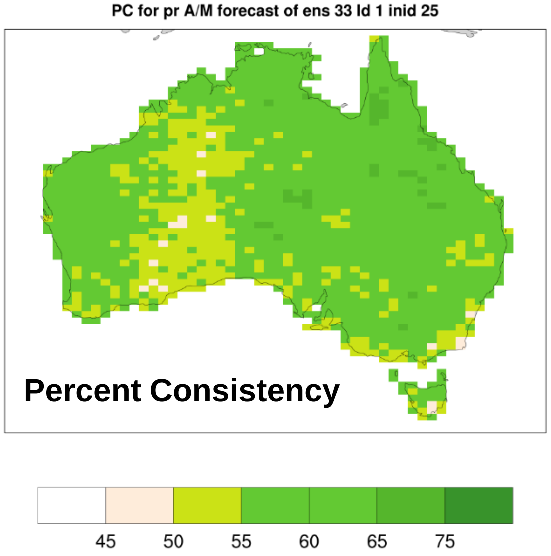
A/M pr 33 ens forecasts of inid 25 for LT1

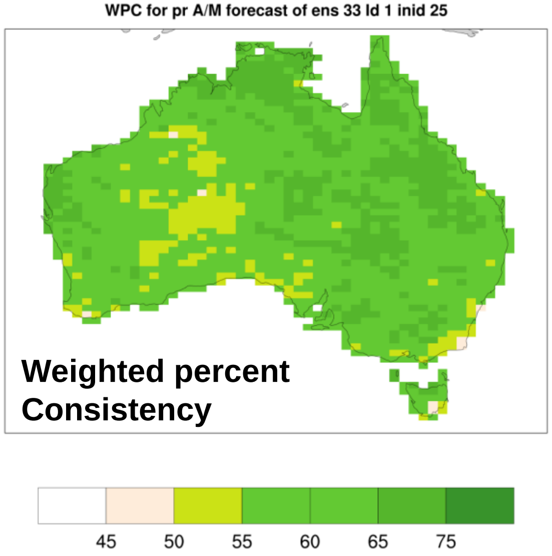


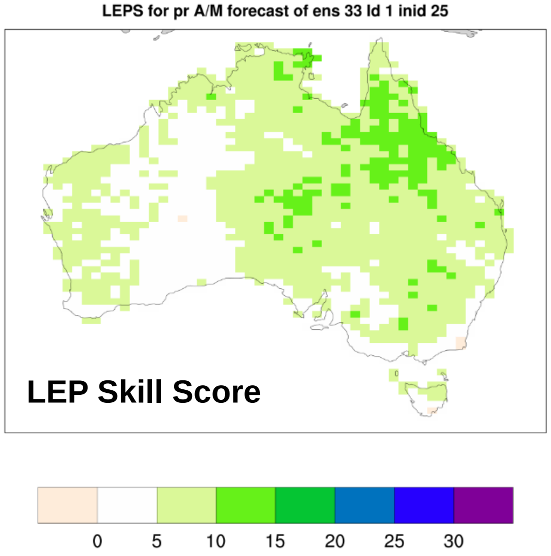
Australian Government
Bureau of Meteorology

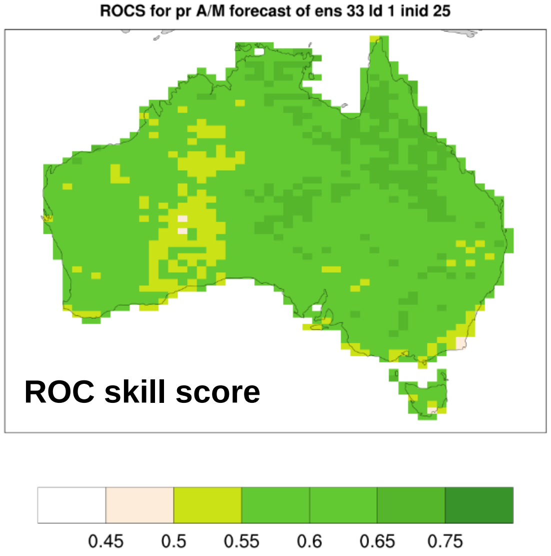
ACCESS-S1 seasonal precipitation hindcast skills verified by various metrics

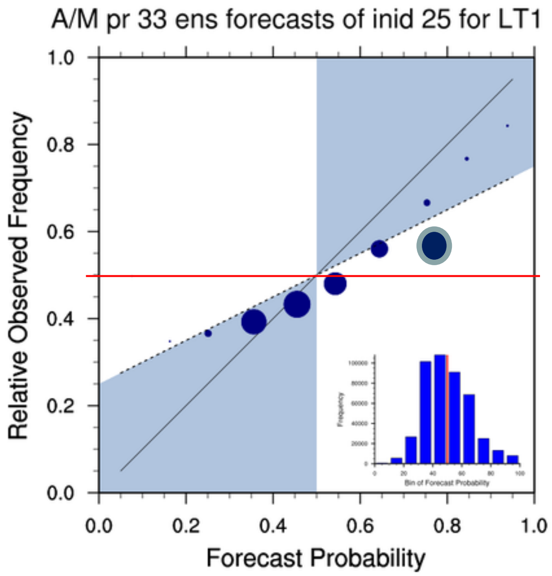




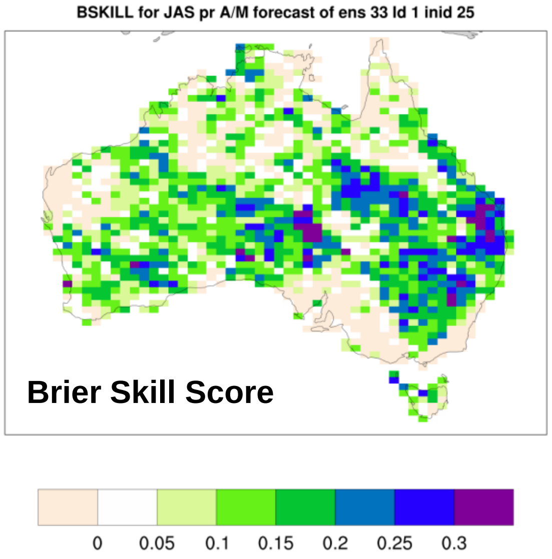

WPC for pr A/M forecast of ens 33 ld 1 inid 25
Weighted percent Consistency
45
50
55
60
65
75

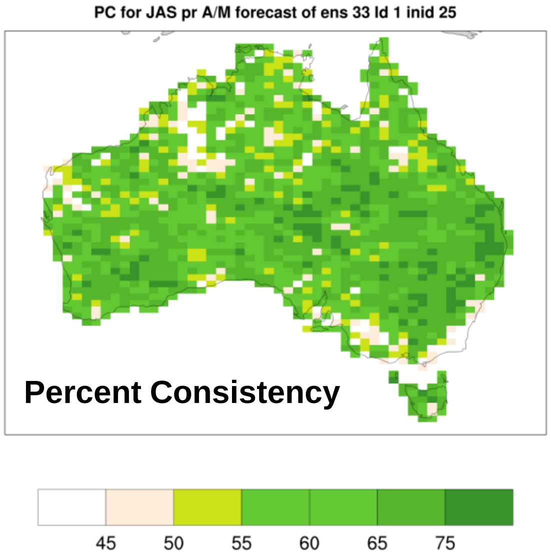


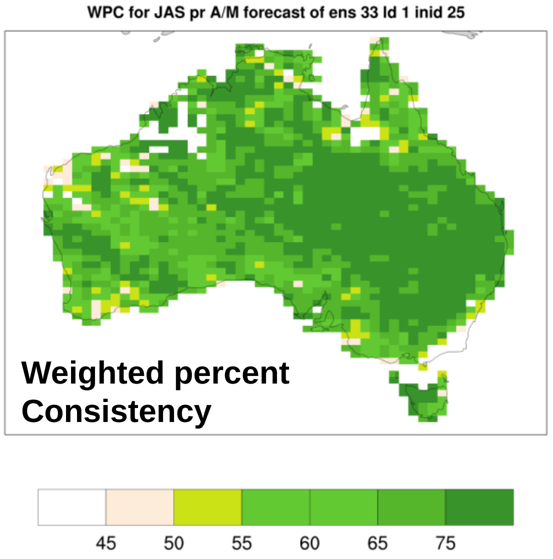


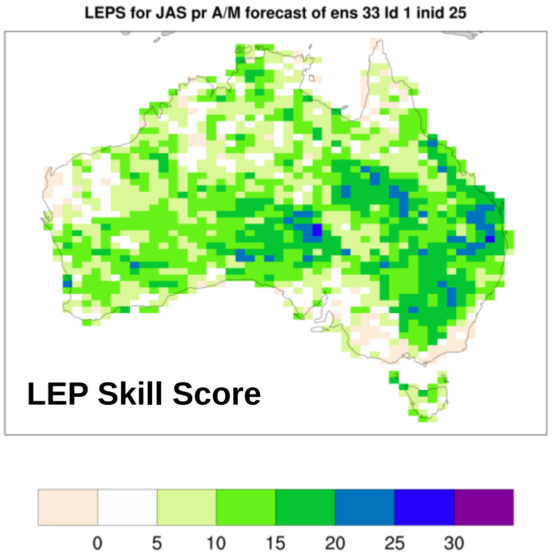


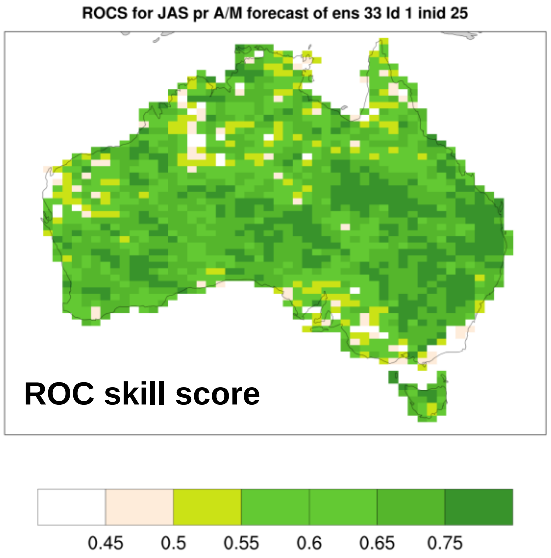
ACCESS-S1 seasonal precipitation hindcast JAS skills verified by various metrics

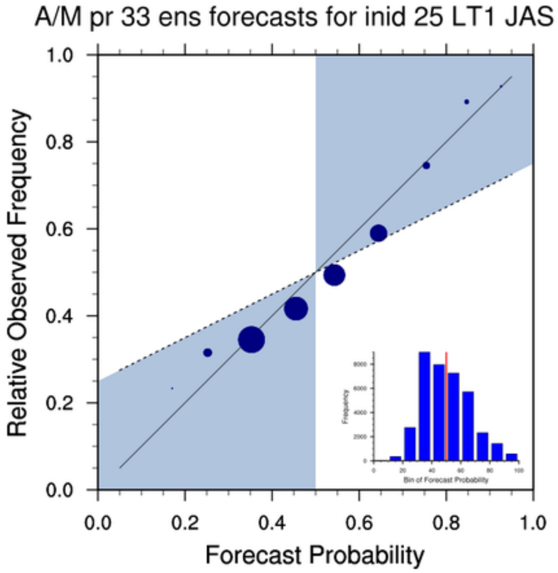





WPC for JAS pr A/M forecast of ens 33 ld 1 inid 25
Weighted percent Consistency
45
50
55
60
65
75







Summary

- It's the Bureau's policy that all forecast products have to be verified
- Verification and the understand, communication and application of skill are still quite challengeable
- Relationship between predictability and skill should be explored,
- Can signal be separated from noise for verification
- Some products are still waiting for proper verification metrics
- New products will need new verification strategy