Australian Government

Bureau of Meteorology

Sub-seasonal and seasonal forecast verification

Young Scientists School, CITES 2019

Debbie Hudson (Bureau of Meteorology, Australia)

Australian Government
Bureau of Meteorology

Overview

1. Introduction
2. Attributes of forecast quality
3. Metrics: full ensemble
4. Metrics: probabilistic forecasts
5. Metrics: ensemble mean
6. Key considerations: sampling issues; stratification; uncertainty; communicating verification

Australian Government
Bureau of Meteorology

Purposes of ensemble verification

User-oriented

- How accurate are the forecasts?
- Do they enable better decisions than could be made using alternate information (persistence, climatology)?

Intercomparison and monitoring

- How do forecast systems differ in performance?
- How does performance change over time?

Calibration

- Assist in bias removal and downscaling

Diagnosis

- Pinpoint sources of error in ensemble forecast system
- Diagnose impact of model improvements, changes to DA and/or ensemble generation etc.
- Diagnose/understand mechanisms and sources of predictability

↑ Operations ↕ Research ↓

Australian Government
Bureau of Meteorology

Evaluating Forecast Quality

Need large number of forecasts and observations to evaluate ensembles and probability forecasts

Forecast quality vs. value

Attributes of forecast quality:

- Accuracy
- Skill
- Reliability
- Discrimination and resolution
- Sharpness

Australian Government
Bureau of Meteorology

Accuracy and Skill

Accuracy

Overall correspondence/level of agreement between forecasts and observations

Skill

A set of forecasts is skilful if better than a reference set, i.e. skill is a comparative quantity

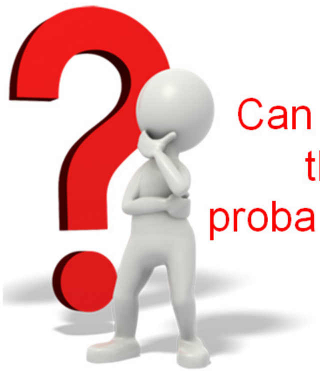
Reference set e.g., persistence, climatology, random

$$Skill\ Score = \frac{score_{\ forecast} - score_{\ reference}}{score_{\ perfect\ forecast} - score_{\ reference}}$$

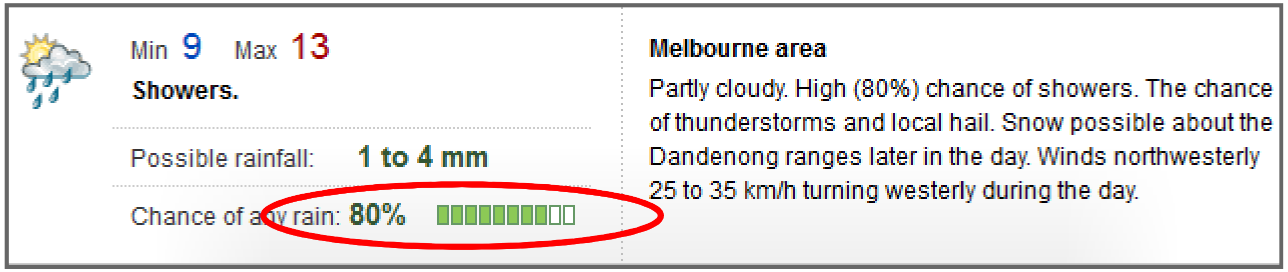
Australian Government
Bureau of Meteorology

Reliability

Ability to give unbiased probability estimates for dichotomous (yes/no) forecasts



Can I trust the probabilities?



Defines whether the certainty communicated in the forecasts is appropriate

Forecast distribution represents distribution of observations

Reliability can be improved by calibration

Australian Government
Bureau of Meteorology

Discrimination and Resolution

Resolution

- How much does the observed outcome change as the forecast changes i.e., "Do outcomes differ given different forecasts?"
- Conditioned on the forecasts

Discrimination

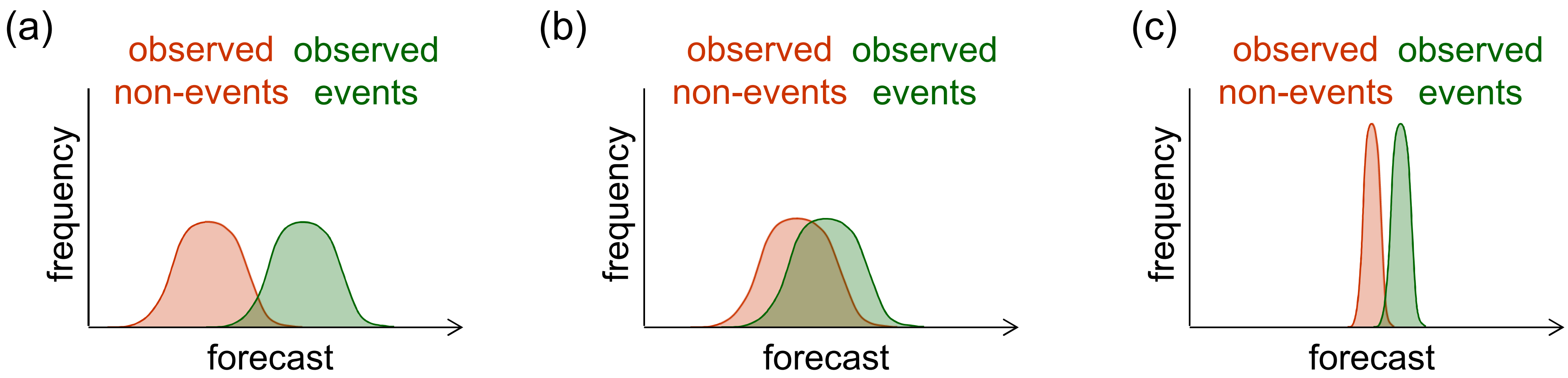
- Can different observed outcomes can be discriminated by the forecasts.
- Conditioned on the observations

Indicates potential "usefulness"

Cannot be improved by calibration

Australian Government
Bureau of Meteorology

Discrimination



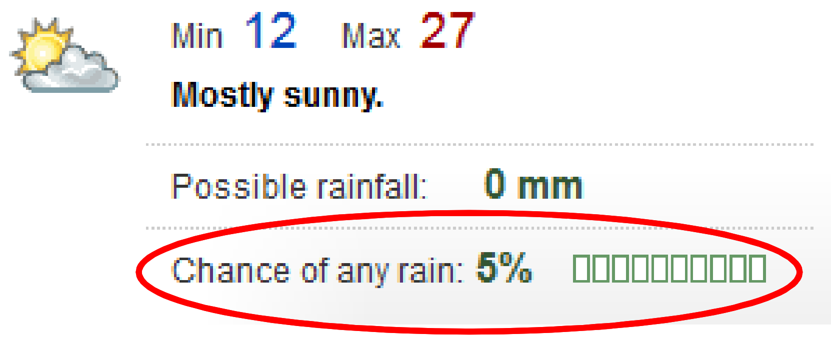
Good discrimination   Poor discrimination   Good discrimination

Australian Government
Bureau of Meteorology

Sharpness

Sharpness is tendency to forecast extreme values (probabilities near 0 or 100%) rather than values clustered around the mean (a forecast of climatology has no sharpness).

A property of the forecast only.



Sharp forecasts are "useful" BUT don't want sharp forecasts if not reliable. Implies unrealistic confidence.

What are we verifying?
How are the forecasts being used?

Ensemble distribution

Set of forecasts making up the ensemble distribution

Use individual members or fit distribution

Probabilistic forecasts generated from the ensemble

Create probabilities by applying thresholds

Ensemble mean

Australian Government
Bureau of Meteorology

Commonly used verification metrics

Characteristics of the full ensemble

- Rank histogram
- Spread vs. skill
- Continuous Ranked Probability Score (CRPS)
  (discussed under probability forecasts)

Australian Government
Bureau of Meteorology

Rank histogram

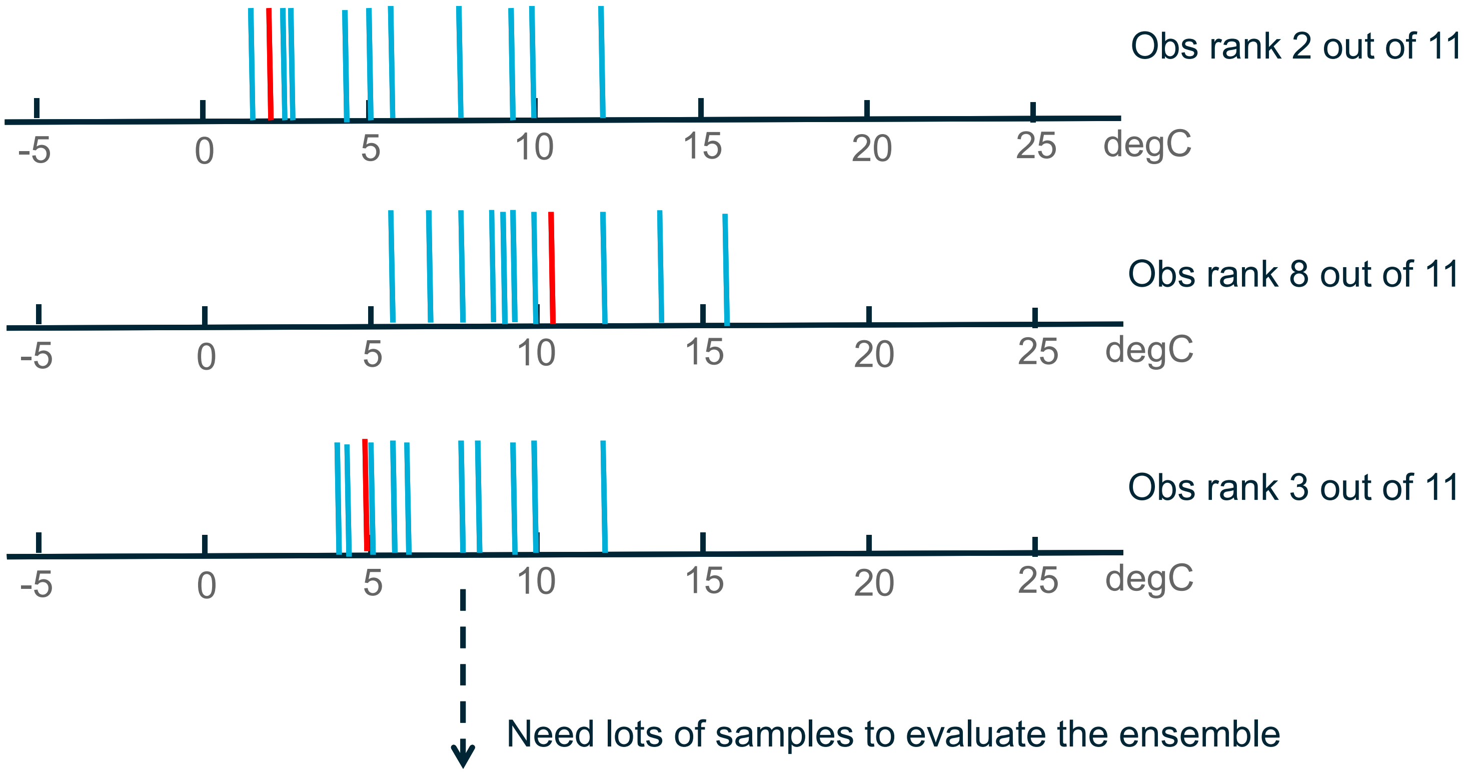
Measures consistency and reliability: the observation is statistically indistinguishable from the ensemble members

→ For each observation, rank the N ensemble members from lowest to highest and identify rank of observation with respect to the forecasts

Example for 10 ensemble members

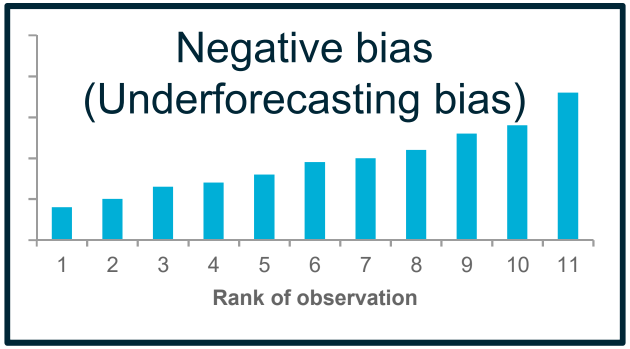
Ensemble

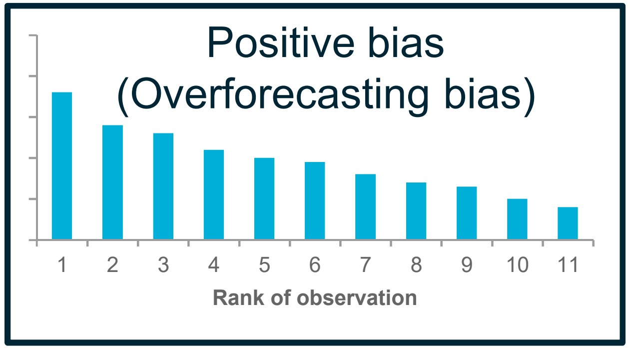
Observation

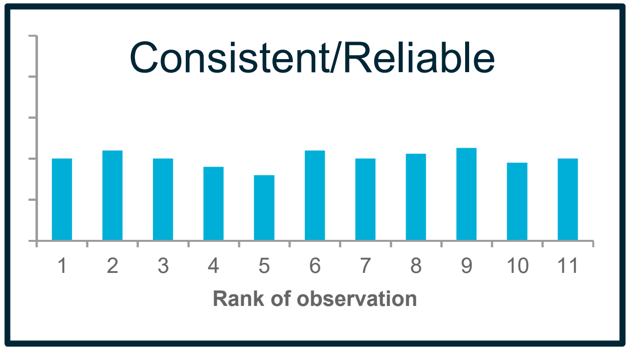


Australian Government
Bureau of Meteorology

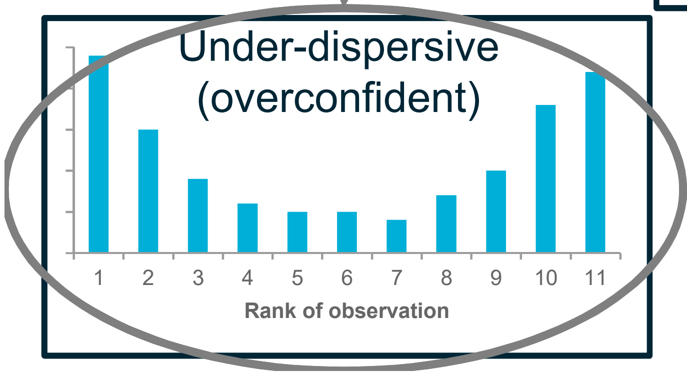
Rank histogram

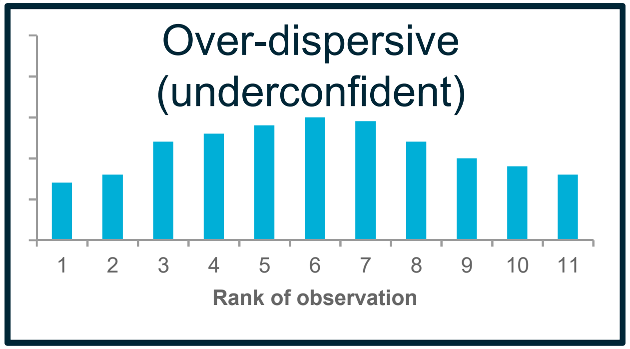






Common problem in seasonal forecasting: ensemble does not have enough spread





Australian Government
Bureau of Meteorology

Rank histogram

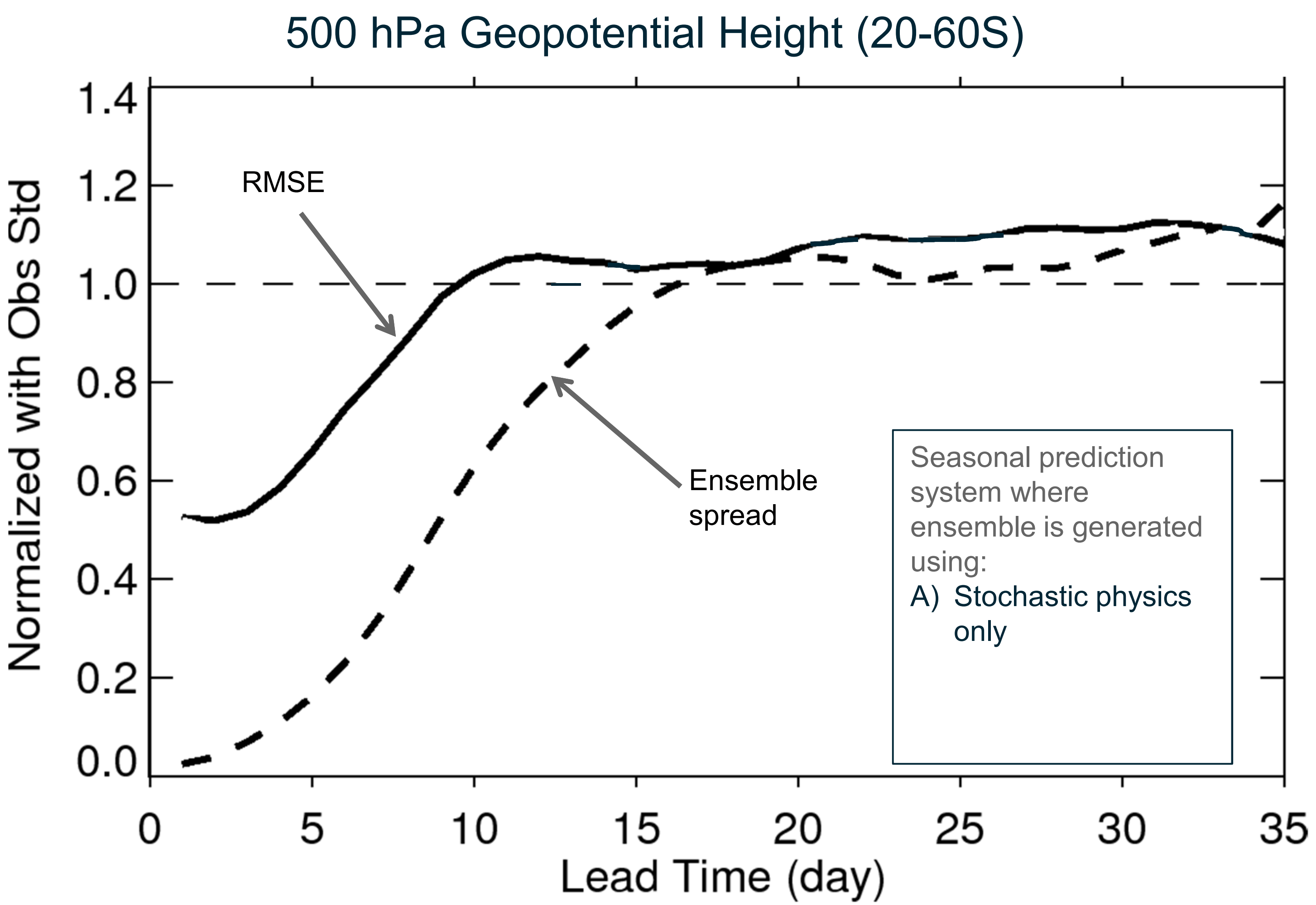
Flat rank histogram does not necessarily indicate a skillful forecast.

Rank histogram shows conditional/unconditional biases BUT not full picture

- Only measures whether the observed probability distribution is well represented by the ensemble.
- Does NOT show sharpness – climatological forecasts are perfectly consistent (flat rank histogram) but not useful

Australian Government
Bureau of Meteorology

Spread-skill evaluation

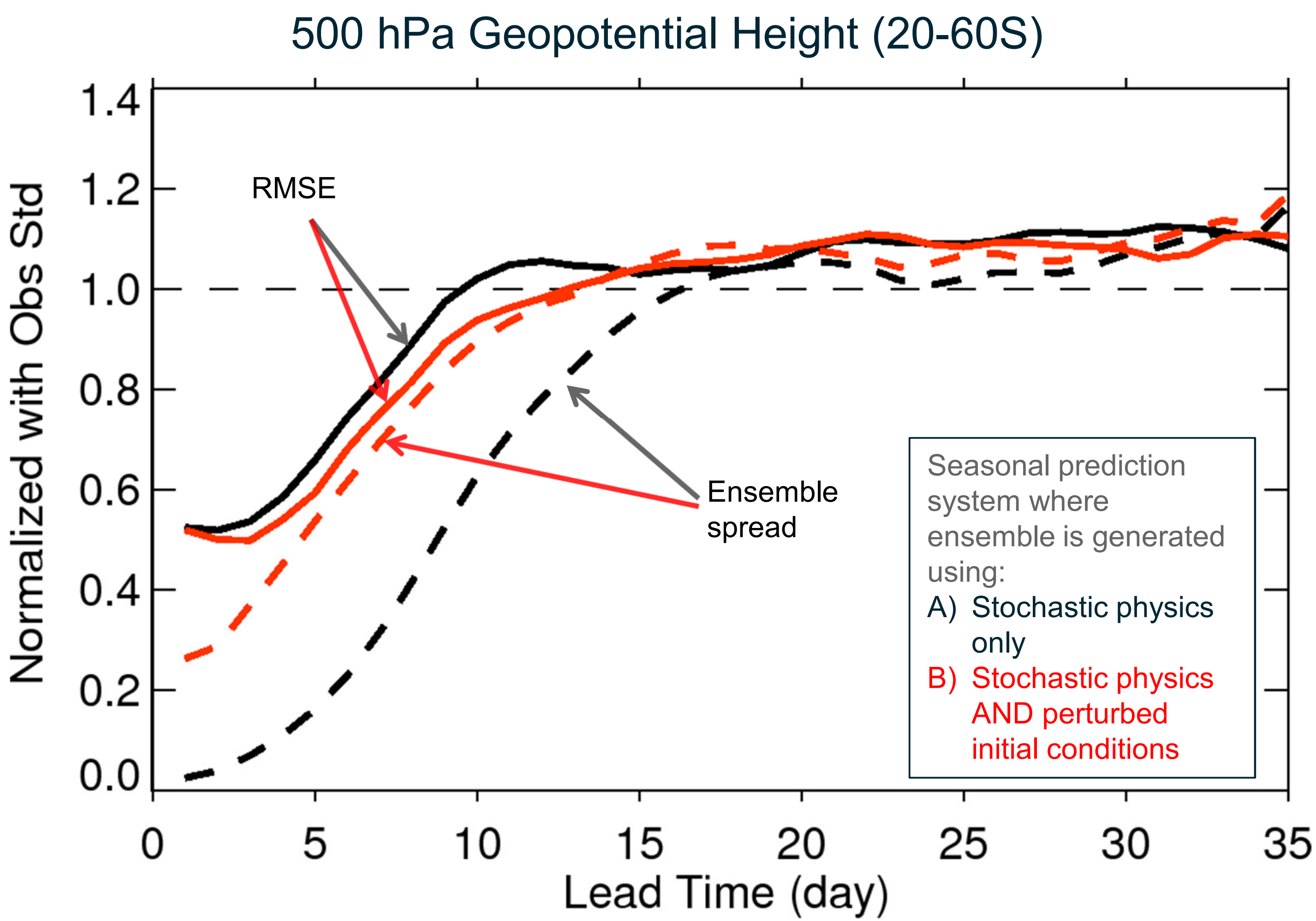


Underdispersed (overconfident)

S_{ens} < RMSE

Australian Government
Bureau of Meteorology

Spread-skill evaluation



Underdispersed (overconfident)

S_{ens} < RMSE

Consistent/reliable

$S_{ens} \approx$ RMSE

Overdispersed (underconfident)

S_{ens} > RMSE

Hudson et al (2017)

Australian Government
Bureau of Meteorology

Commonly used verification metrics

Probability forecasts

- Reliability/Attributes diagram
- Brier Score (BS and BSS)
- Ranked Probability Score (RPS and RPSS)
- Continuous Ranked Probability Score (CRPS and CRPSS)
- Relative Operating Characteristic (ROC and ROCS)
- Generalized Discrimination Score (GDS)

Australian Government
Bureau of Meteorology

Reliability (attributes) diagram

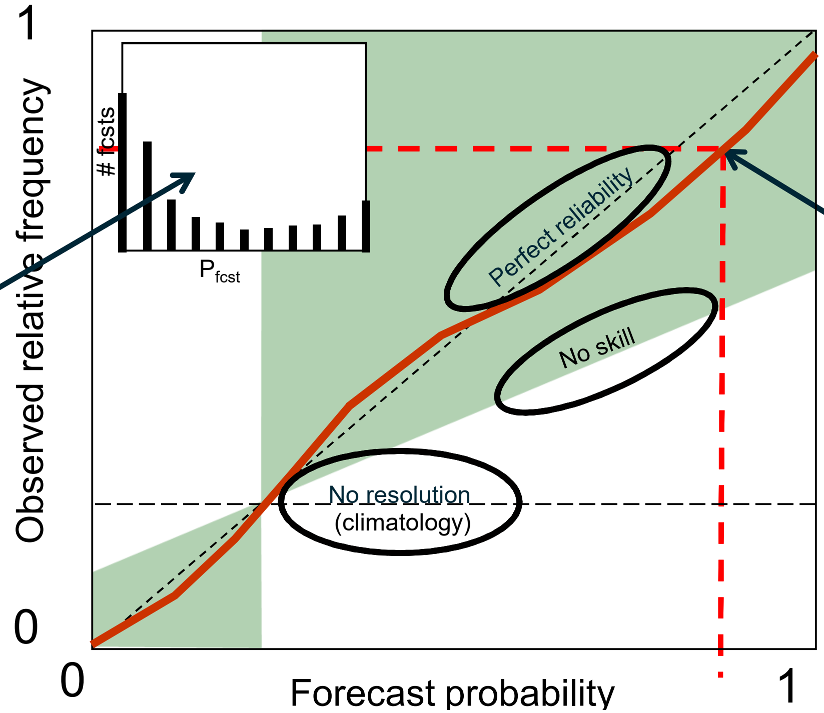
Dichotomous forecasts

Measures how well the predicted probabilities of an event correspond to their observed frequencies (reliability)

→ Plot observed frequency against forecast probability for all probability categories
→ Need a big enough sample

Histogram: how often each probability was issued.

Shows sharpness and potential sampling issues

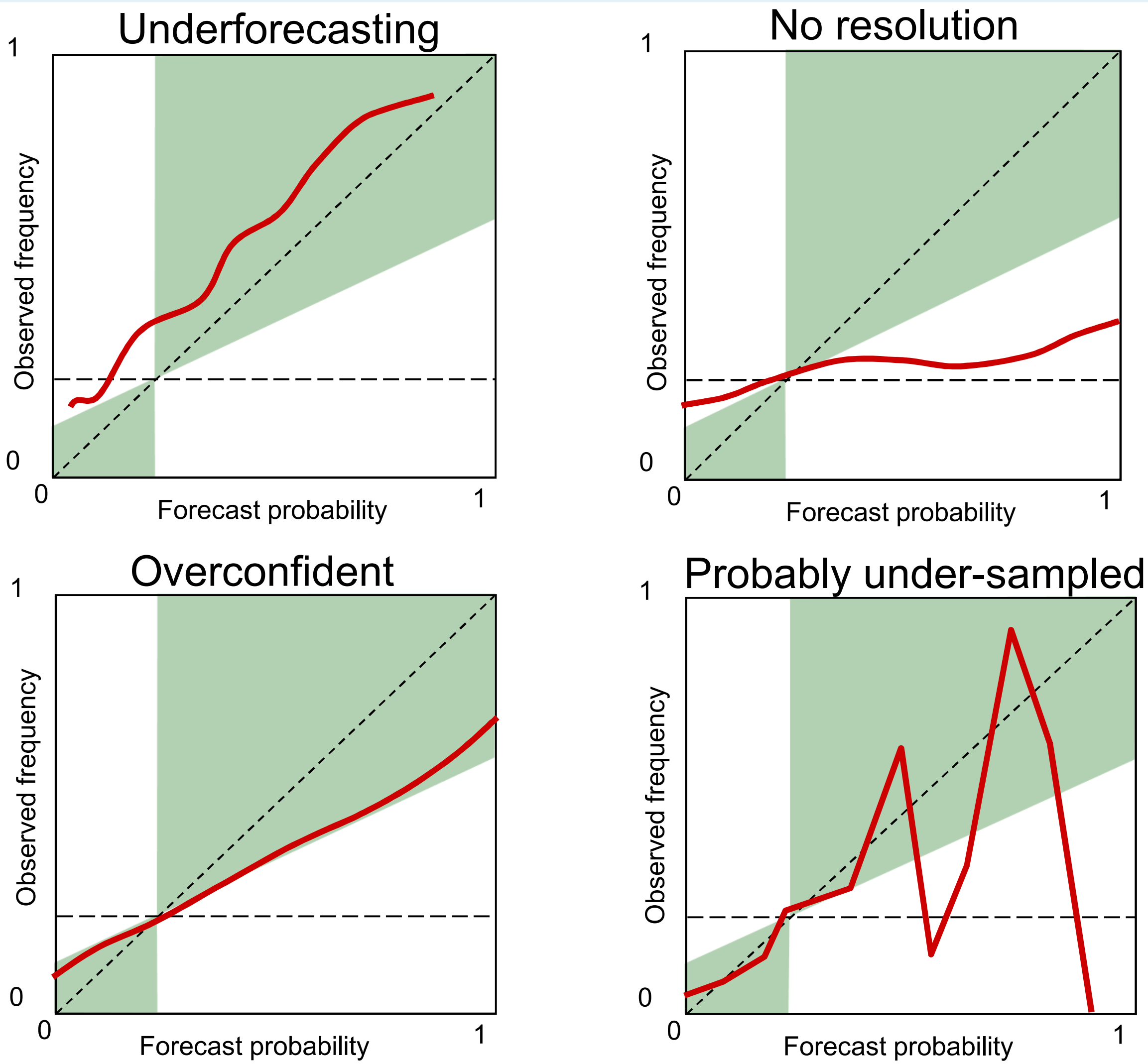


Curve tells what the observed frequency was for a given forecast probability.

Conditioned on the forecasts

Australian Government
Bureau of Meteorology

Interpretation of reliability diagrams

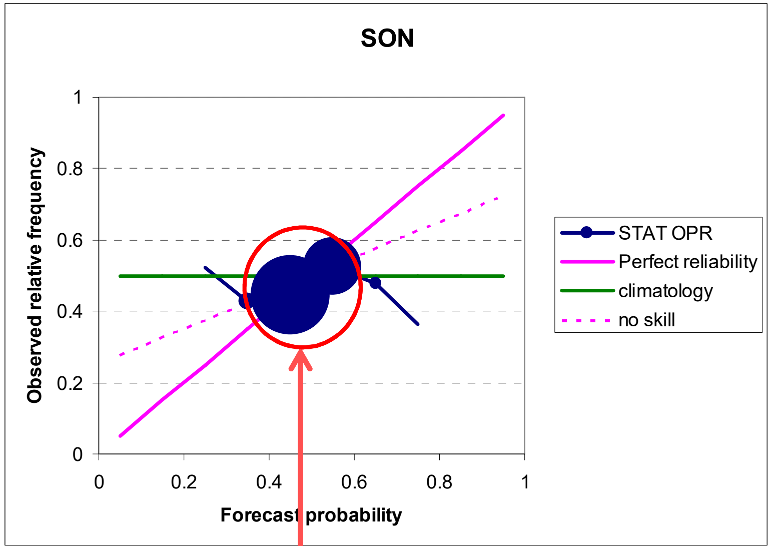


Australian Government
Bureau of Meteorology

Reliability diagram: Example

Predictions of above normal seasonal SON rainfall

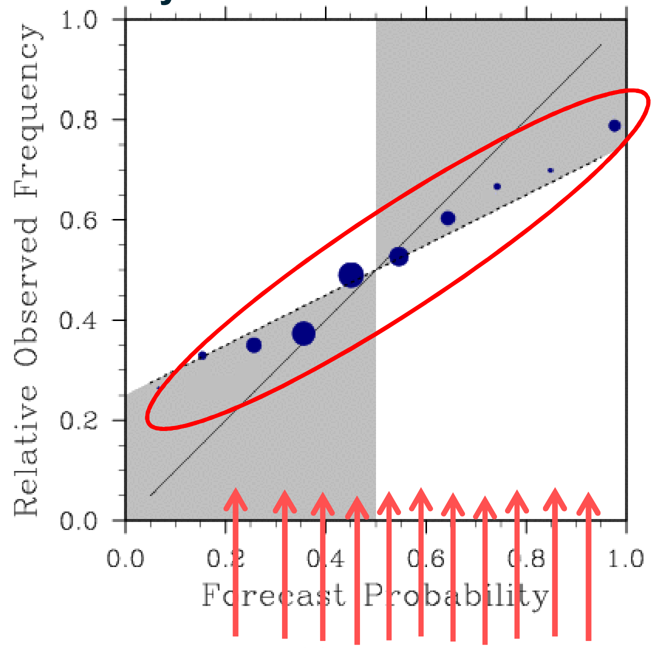
Statistical forecast scheme



Most of the forecasts issued have probabilities near 50%

Size of the circles are proportional to the number of forecasts issuing that probability

Dynamical forecast scheme



A range of forecast probabilities are issued

The statistical system often gave forecasts close to climatology – reliable BUT poor sharpness. Of limited use for decision-makers!

Australian Government
Bureau of Meteorology

Brier score (BS)

Dichotomous forecasts
Brier score measures the mean squared probability error

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

p_i: Forecast probability; o_i: Observed occurrence (0 or 1)

- Score range: 0 to 1; Perfect BS: 0

Murphy's (1973) decomposition into 3 terms (for *K* probability classes and *N* samples):

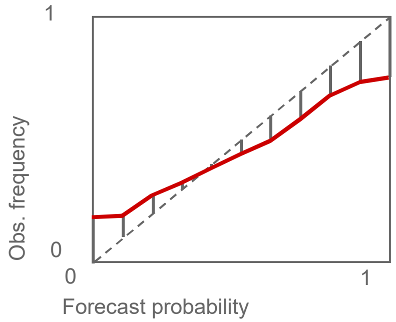
$$BS = \underbrace{\frac{1}{N}\sum_{k=1}^{K}n_k(p_k - \overline{o}_k)^2}_{\text{reliability}} - \underbrace{\frac{1}{N}\sum_{k=1}^{K}n_k(\overline{o}_k - \overline{o})^2}_{\text{resolution}} + \underbrace{\overline{o}(1-\overline{o})}_{\text{uncertainty}}$$

- Useful for exploring dependence of probability forecasts on ensemble characteristics
- Uncertainty term measures the variability of the observations. Has nothing to do with forecast quality!
- BS is sensitive to the climatological frequency of an event: the more rare an event, the easier it is to get a good BS without having any real skill

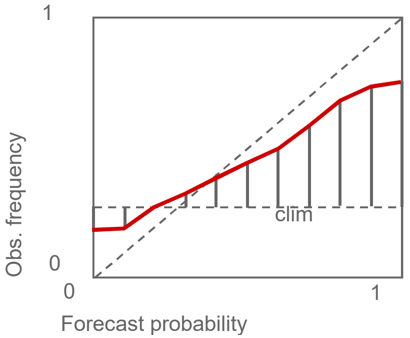
Australian Government
Bureau of Meteorology

BS, Brier Skill Score (BSS) and the Attributes diagram

Reliability term (BS_{rel}): measures deviation of the curve from the diagonal line – error in the probabilities.



Resolution term (BS_{res}): measures deviation of the curve from the sample climate horizontal line – indicates degree to which forecast can separate different situations

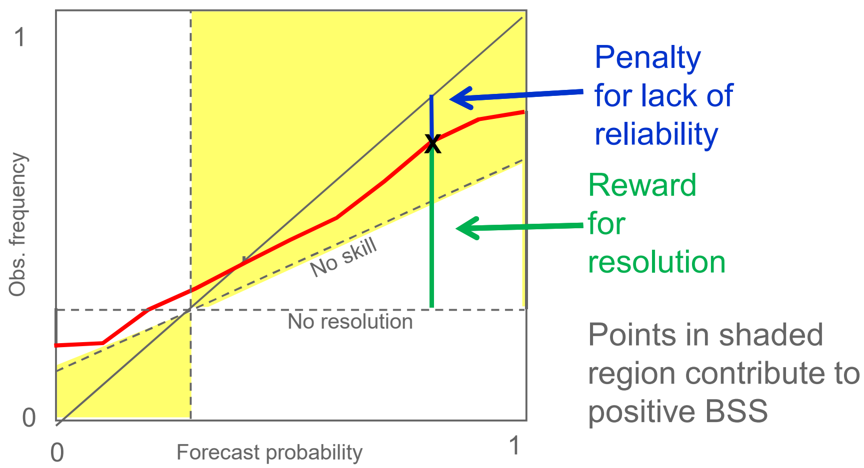


Brier *skill* score: measures the relative skill of the forecast compared to climatology

$$BSS = 1 - \frac{BS}{BS_{clim}}$$

$$BSS = \frac{resolution - reliability}{uncertainty}$$

Perfect: $BSS = 1.0$
Climatology: $BSS = 0.0$

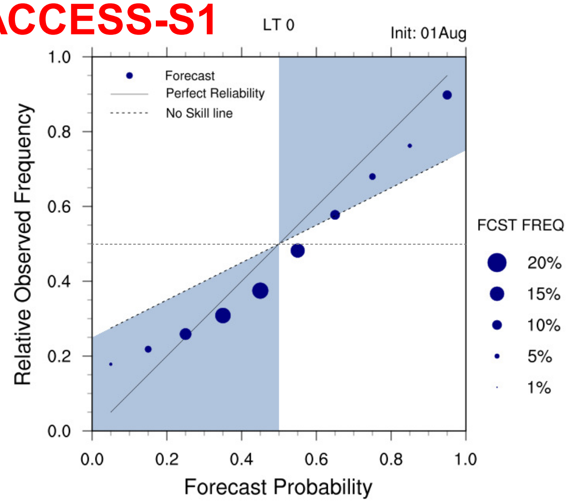


Australian Government
Bureau of Meteorology

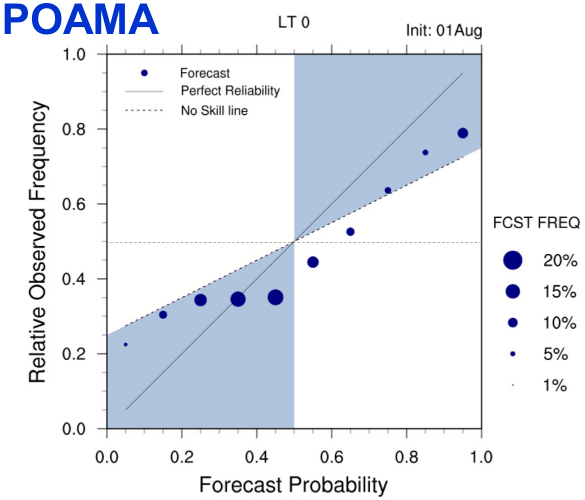
BS_{rel} and BS_{res}: Example

Aug-Sep-Oct season

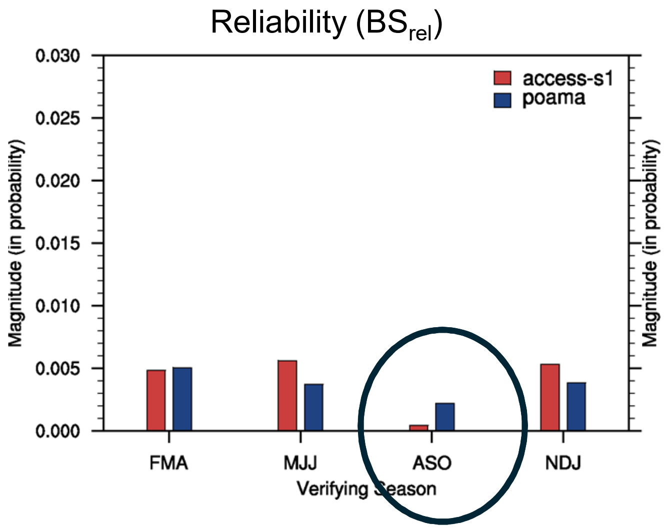
ACCESS-S1

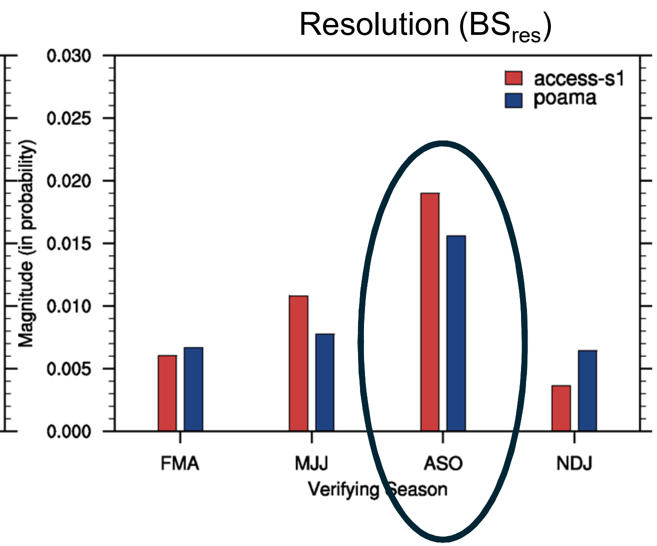


POAMA



Probability seasonal mean rainfall above-average over Australia





Smaller is better

Bigger is better

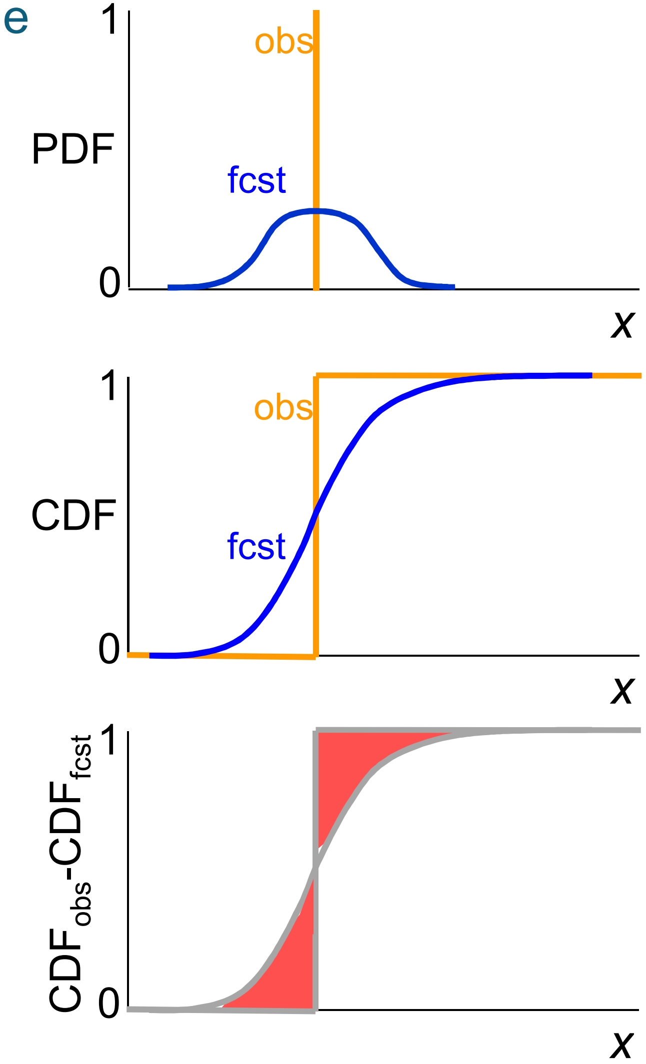
Australian Government
Bureau of Meteorology

Continuous ranked probability score (CRPS)

Continuous ranked probability score (CRPS) measures the difference between the forecast and observed CDFs

$$CRPS = \int_{-\infty}^{\infty} \left(P_{fcst}(x) - P_{obs}(x)\right)^2 dx$$

- Same as Brier score integrated over all thresholds
- On continuous scale: does not need reduction of ensemble forecasts to discrete probabilities of binary or categorical events (for multi-category use Ranked Probability Score)
- Same as Mean Absolute Error for deterministic forecasts
- Has dimensions of observed variable
- Perfect score: 0
- Rewards small spread (sharpness) if the forecast is accurate
- Skill score wrt climatology: $CRPSS = 1 - \frac{CRPS}{CRPS_{clim}}$



Australian Government
Bureau of Meteorology

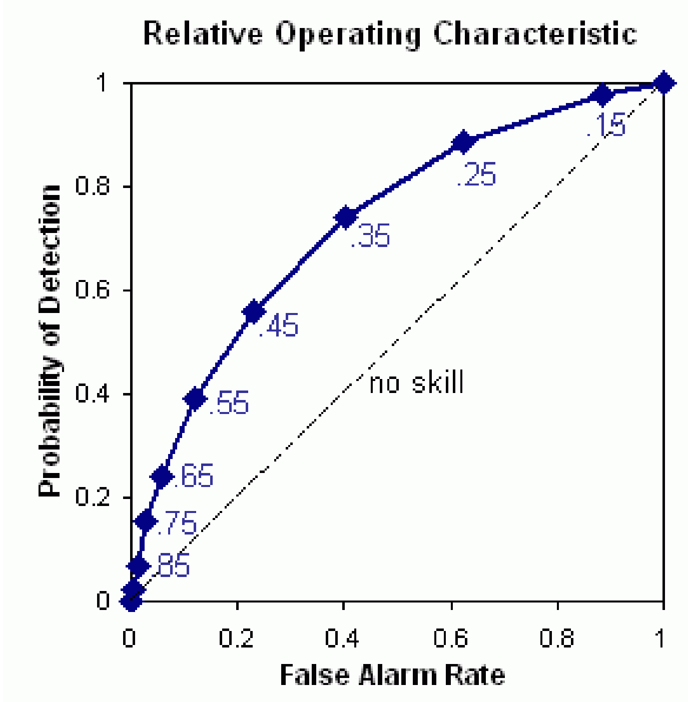
Relative Operating Characteristic (ROC)

Dichotomous forecasts
Measures the ability of the forecast to discriminate between events and non-events (discrimination)

→ Plot hit rate vs false alarm rate using a set of varying probability thresholds to make the yes/no decision.

Close to upper left corner – good discrimination

Close to or below diagonal – poor discrimination



Australian Government
Bureau of Meteorology

Relative Operating Characteristic (ROC)

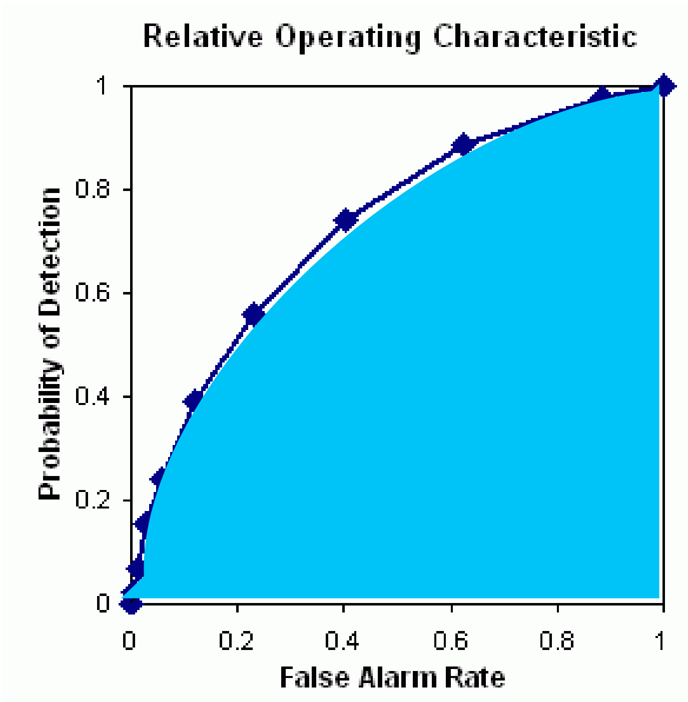
Dichotomous forecasts
Measures the ability of the forecast to discriminate between events and non-events (discrimination)

→ Plot hit rate vs false alarm rate using a set of varying probability thresholds to make the yes/no decision.

Close to upper left corner – good discrimination

Close to or below diagonal – poor discrimination

- Area under curve ("ROC area") is a useful summary measure of forecast skill



Australian Government
Bureau of Meteorology

Relative Operating Characteristic (ROC)

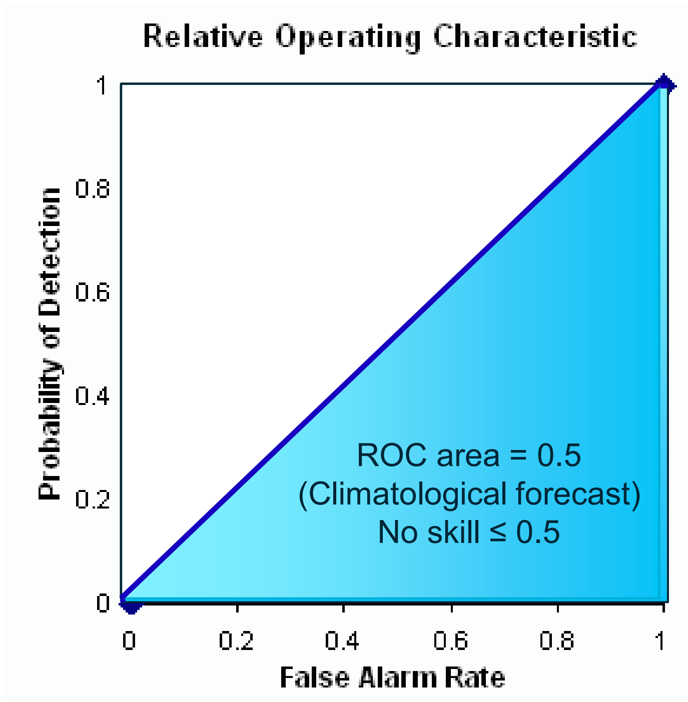
Dichotomous forecasts
Measures the ability of the forecast to discriminate between events and non-events (discrimination)

→ Plot hit rate vs false alarm rate using a set of varying probability thresholds to make the yes/no decision.

Close to upper left corner – good discrimination

Close to or below diagonal – poor discrimination

- Area under curve ("ROC area") is a useful summary measure of forecast skill



Australian Government
Bureau of Meteorology

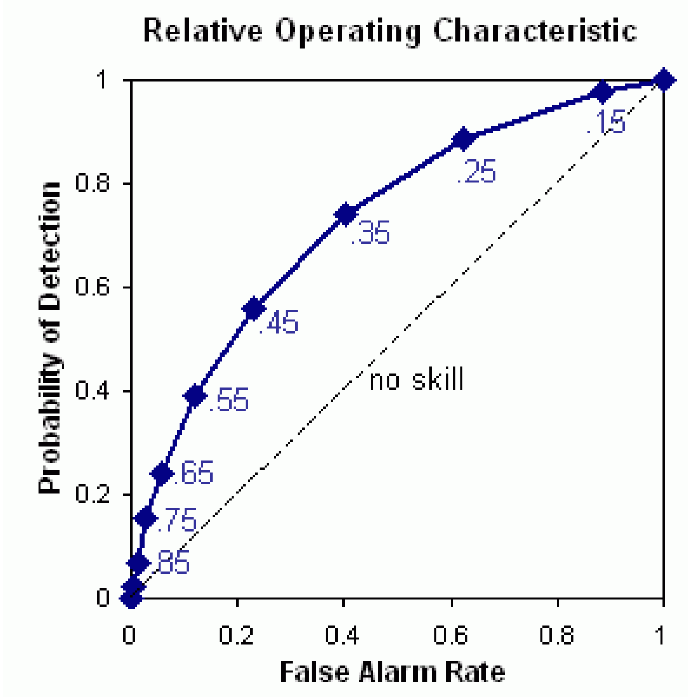
Relative Operating Characteristic (ROC)

Dichotomous forecasts
Measures the ability of the forecast to discriminate between events and non-events (discrimination)

→ Plot hit rate vs false alarm rate using a set of varying probability thresholds to make the yes/no decision.

Close to upper left corner – good discrimination

Close to or below diagonal – poor discrimination



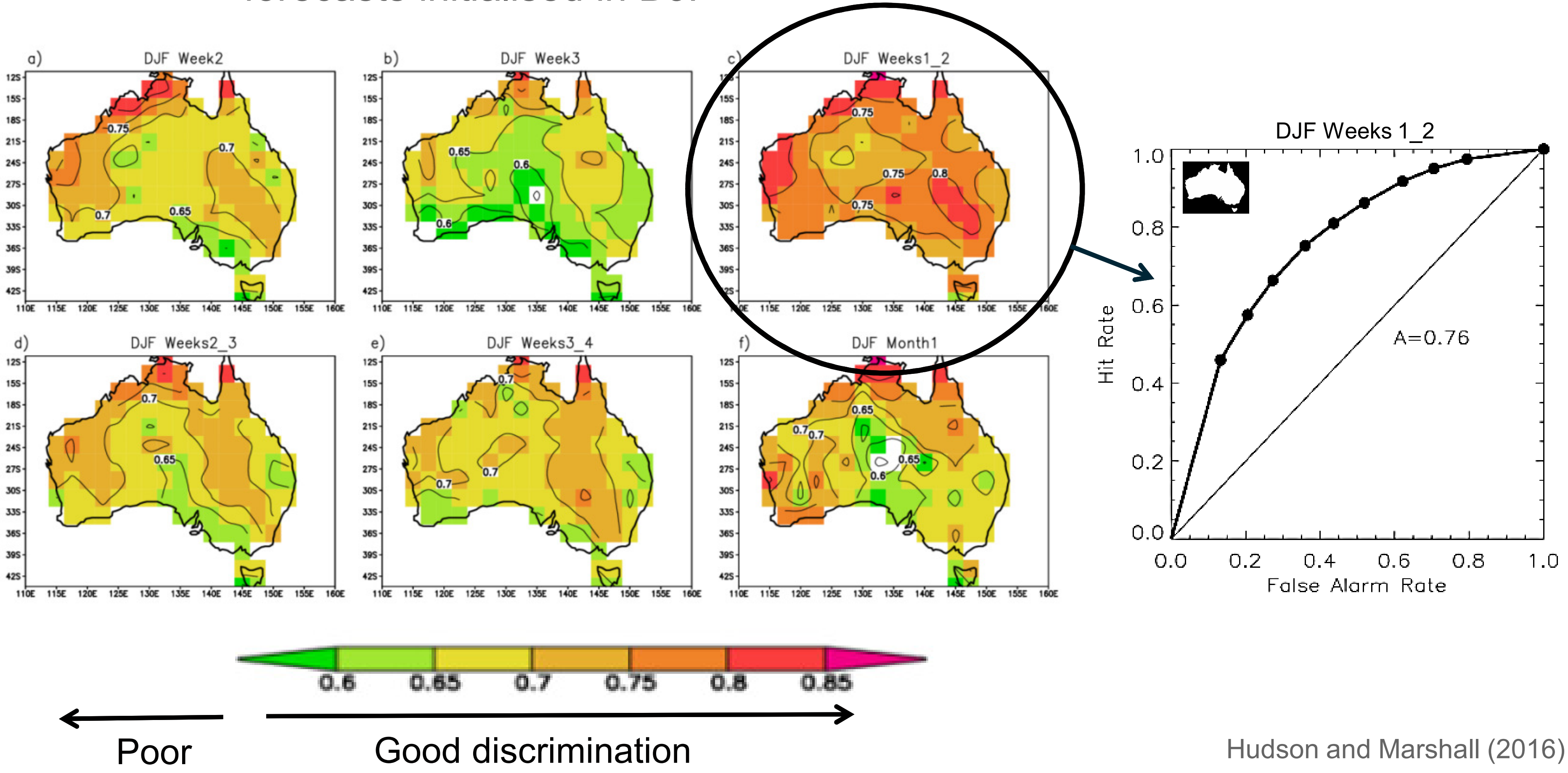
- Area under curve ("ROC area") is a useful summary measure of forecast skill

- ROC skill score: ROCS = 2(ROCarea-0.5)
- The ROC is conditioned on the observations
- Reliability and ROC diagrams are good companions

Australian Government
Bureau of Meteorology

ROC: Example

ROC area of probability of a heatwave for all forecasts initialised in DJF



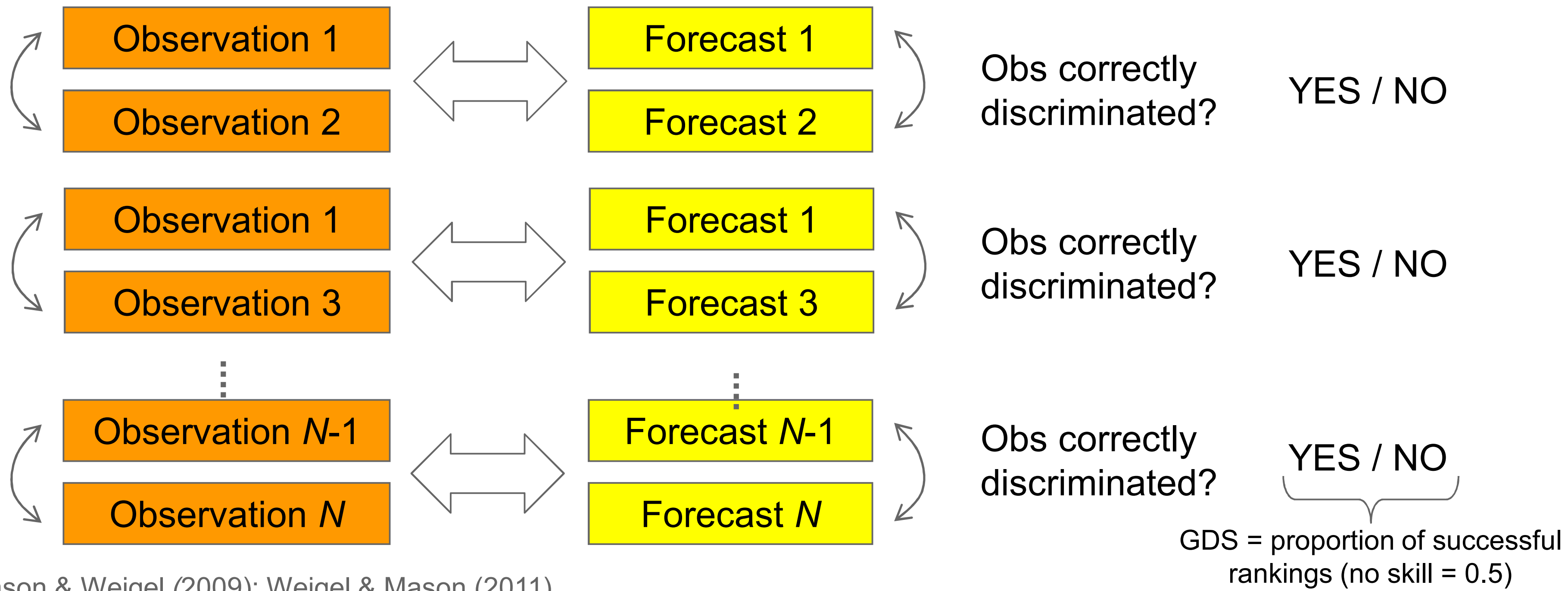
Hudson and Marshall (2016)

Australian Government
Bureau of Meteorology

Generalized Discrimination Score (GDS)

Binary, multi-category & continuous
Rank-based measure of discrimination - does the forecast successfully rank (discriminate) the two different observations?

GDS equivalent to ROC area for dichotomous forecasts & has same scaling



Mason & Weigel (2009); Weigel & Mason (2011)

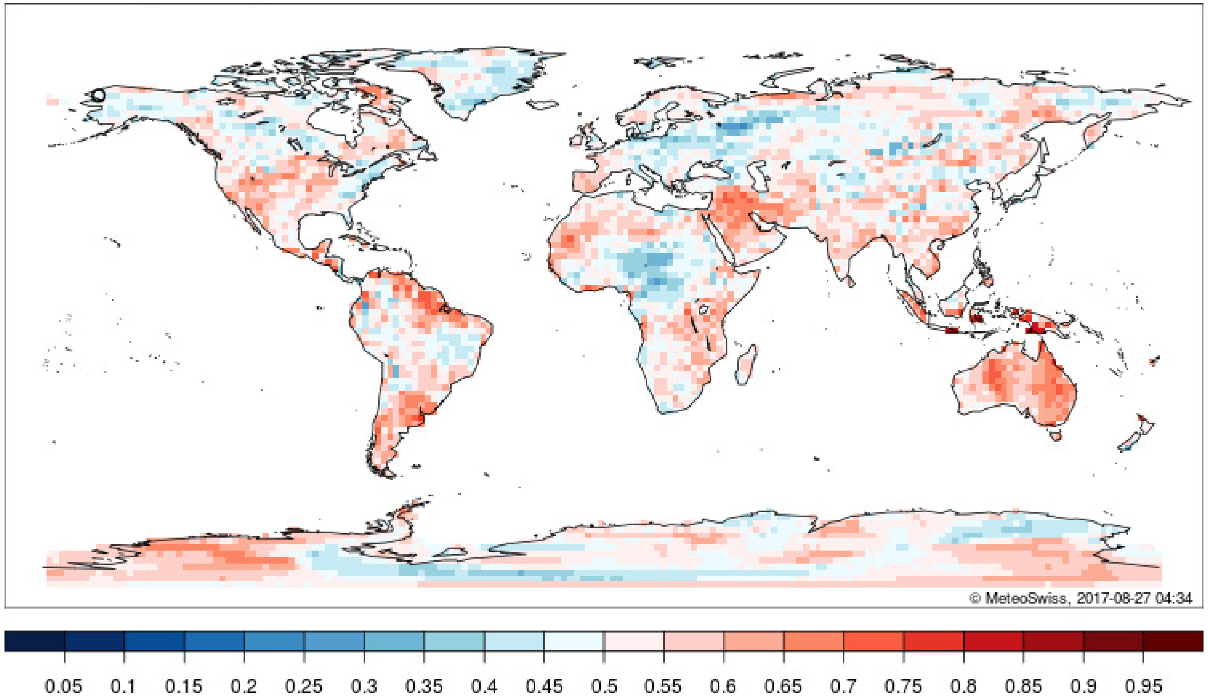
Australian Government
Bureau of Meteorology

GDS (and ROC): Example

Forecast of seasonal SON rainfall

Generalized discrimination score

Seasonal (SON) precipitation from ECMWF SYSTEM4 forecasts
initialised in August verified against ERA-INT for 1981-2014



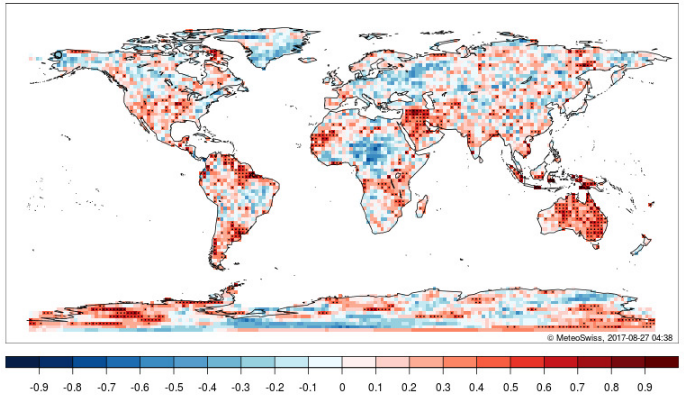
No/Poor Discrimination

Good discrimination

https://meteoswiss-climate.shinyapps.io/skill_metrics/

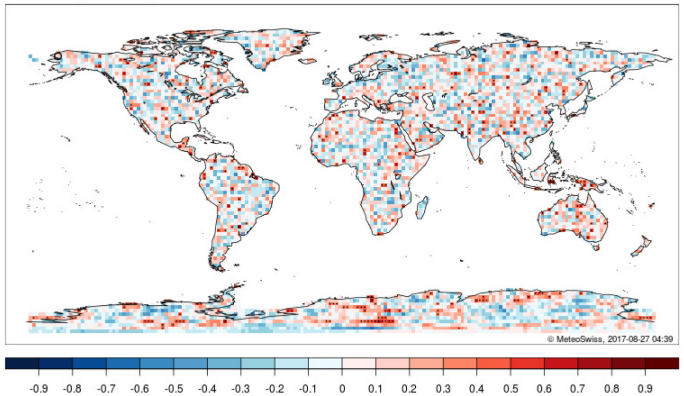
ROC area skill score (lower tercile)

Seasonal (SON) precipitation from ECMWF SYSTEM4 forecasts
initialised in August verified against ERA-INT for 1981-2014



ROC area skill score (middle tercile)

Seasonal (SON) precipitation from ECMWF SYSTEM4 forecasts
initialised in August verified against ERA-INT for 1981-2014



Australian Government
Bureau of Meteorology

Commonly used verification metrics

Ensemble mean

e.g., RMSE, correlation

Australian Government
Bureau of Meteorology

Verification of ensemble mean

Debate as to whether or not this is a good idea:

Pros:

- Ensemble mean filters out smaller unpredictable scales
- Needed for spread – skill evaluation
- Forecasters and others use ensemble mean

Cons:

- Not a realization of the ensemble
- Different statistical properties to ensemble and observations

Scores:

- RMSE
- Anomaly correlation
- Other deterministic verification scores

Australian Government
Bureau of Meteorology

Key considerations: Sampling issues

Rare and extreme events

Difficult to verify probabilities on the "tail" of the PDF

- Too few samples to get robust statistics, especially for reliability
- Finite number of ensemble members may not resolve tail of forecast PDF

Use of weighted fair scores

Gneiting, Ranjan (2011) Comparing density forecasts using threshold- and quantile weighted scoring rules. Journal of Business & Economic Statistics, 29, 411–422

Lerch, Thorarinsdottir, Ravazzolo, Gneiting (2017) Forecaster's dilemma: extreme events and forecast evaluation. Statistical Science, 32, 106–127

Ferro (2014) Fair scores for ensemble forecasts. QJRMS, 140, 1917–1923

Ferro, Richardson, Weigel (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorological Applications, 15, 19–24

Size of ensemble vs number of verification samples

Robustness of verification depends on both!!!

Australian Government
Bureau of Meteorology

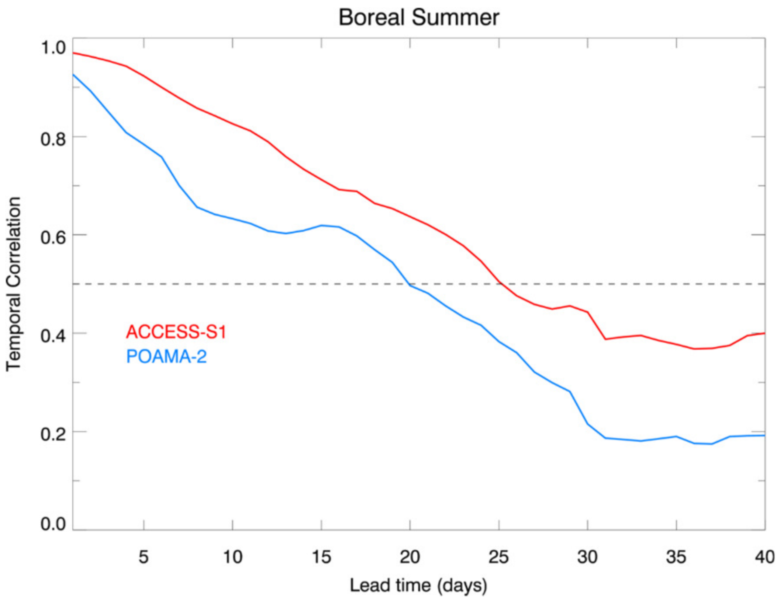
Key considerations: Stratification

Verification results vary with region, season, climate driver……

Pooling samples can mask variations in forecast performance

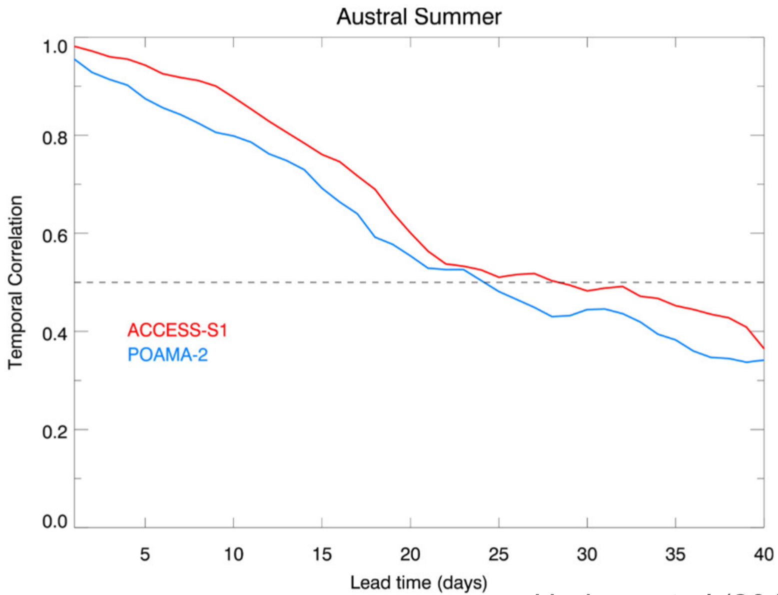
Stratify data into sub-samples

- BUT must have enough samples to give robust statistics!



MJO

Example:
MJO Bivariate correlation for RMM index



Hudson et al (2017)

Australian Government
Bureau of Meteorology

Key considerations: Uncertainty

Are the forecasts significantly better than a reference forecast?

Does ensemble A perform significantly better than ensemble B?

- Take into account sampling variability
- Significance levels and/or confidence intervals
- Non-parametric resampling methods (Monte Carlo, bootstrap)

Effects of observation errors

- Adds uncertainty to verification results
- True forecast skill unknown
- Extra dispersion of observed PDF
- Active area of research

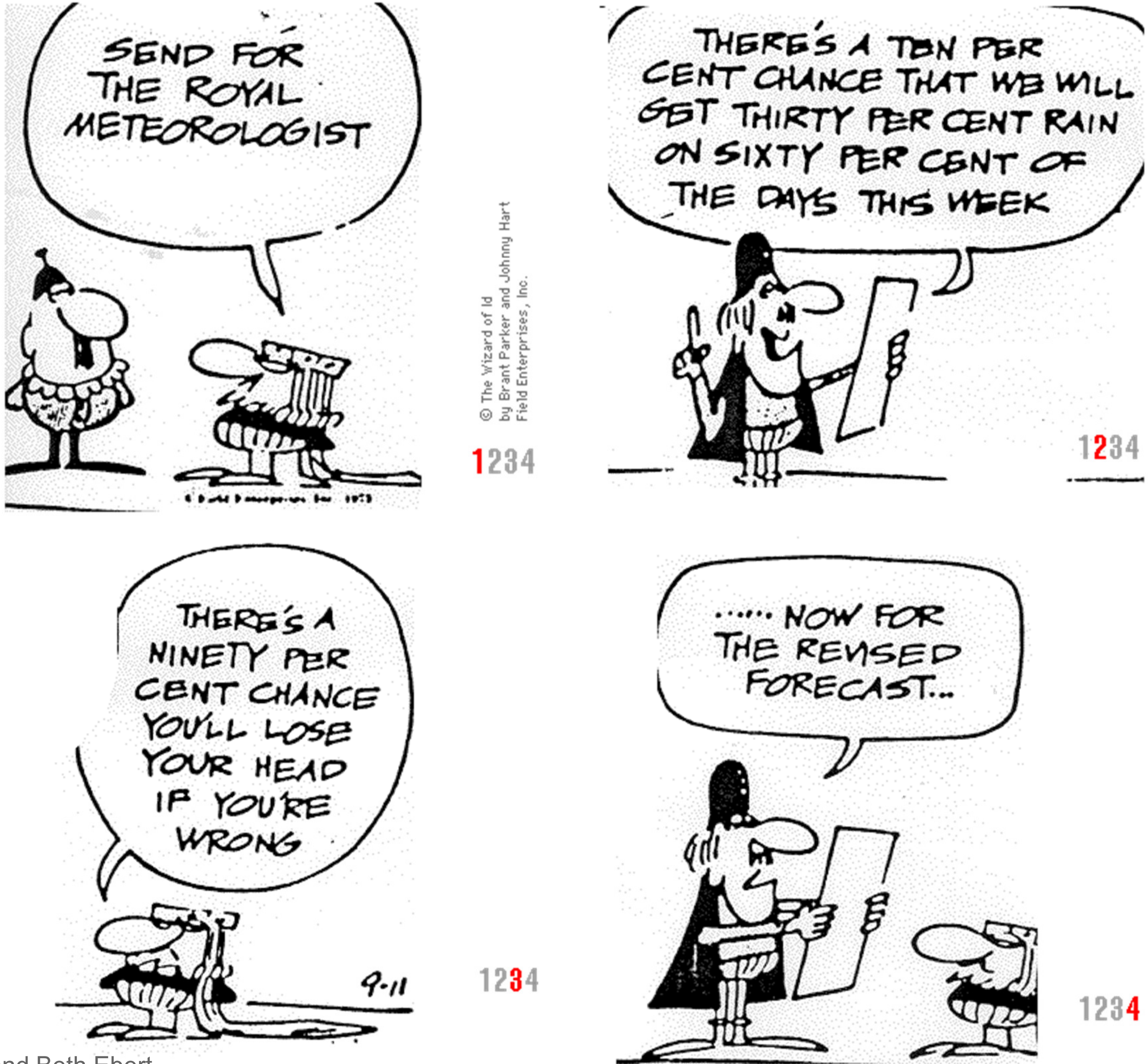
Australian Government
Bureau of Meteorology

Key considerations: Communicating verification to users

- Challenging to communicate ensemble verification
- Forecast quality does not necessarily reflect value
- Summary skill measure – average skill over hindcasts. Does not show how skill changes over time (windows of forecast opportunity)
- Large sampling uncertainty around scores for quantities that are of most interest to the user e.g. regional rainfall

Related considerations:

- Using hindcasts to estimate skill (smaller ensemble size that real-time)
- Models becoming more computationally expensive – constraints on hindcast size. What is optimal hindcast size – # years; start dates and ensemble size?



Thanks Ian Jolliffe and Beth Ebert

Australian Government
Bureau of Meteorology

Useful general references

WMO Verification working group forecast verification web page:
http://www.cawcr.gov.au/projects/verification/

Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Sciences. 3rd Edition*. Elsevier, 676 pp.

Jolliffe, I.T., and D.B. Stephenson, 2012: *Forecast Verification. A Practitioner's Guide in Atmospheric Science*., 2ndEdition, Wiley and Sons Ltd.

Special issues of *Meteorological Applications* on Forecast Verification (Vol 15 2008 & Vol 20 2013)

Thank you…

Debbie Hudson
Debbie.Hudson@bom.gov.au