Verification of Quantile Forecasts
A Journey
(or: Mistakes I have made)

Deryn Griffiths, Robert Taggart, Michael Foley, Nicholas Loveday,
Alistair McKelvie, Ben Price

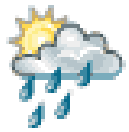
Bureau of Meteorology, Australia

(the mistakes are mine)

Deryn.Griffiths@bom.gov.au

A forecast

Saturday 17 October



Min 12 Max 19

Showers.

Possible rainfall: **5 to 10 mm**

Chance of any rain: **80%**

Melbourne area

Cloudy. High (80%) chance of showers, most likely in the morning and afternoon. The chance of a thunderstorm about the nearby hills in the afternoon and evening. Winds north to northeasterly 15 to 20 km/h shifting south to southwesterly during the morning then becoming light during the evening.

Sun protection recommended from 9:50 am to 4:20 pm, UV Index predicted to reach 7 [High]

To verify: Chance of any rain

Possible rainfall range

How should I measure my errors?

I want my measure to be *Proper (Consistent)*, unable to be *Gamed*

- I want Forecasters to minimise their expected error by forecasting what they believe.
- If forecasters predict something they don't believe, I want them to expect a worse error.

E.g. "% within 5 mm" is not Proper for expected rainfall

- Being confident of 0 mm, you are still better off to forecast 4.9 mm.

For a forecast of "Chance of <an event>"

- Brier Score is well known to be consistent

If Observation, o, is 1 or 0 (event occurred or not)

And Forecast, p, is in [0, 1]

Then the Brier Score is $(p - o)^2$

- A standard decomposition to show Reliability and Resolution

Chance of any rain Verification Results

ETA Verification Latest Results - DailyPoPX + PoP

Home | FAQs | Station Groups Info ----- **DailyPoPX + PoP** | (Daily)PrecipYPct -----

Sources
- OCF 18Z vs Official pm
- **Official pm vs Current FCF (2.0) 18Z**
- OCF 18Z vs Current FCF (2.0) 18Z

Plot

Region
- **Cross-region**
- NSW/ACT
- NT
- QLD
- SA
- TAS

Partition
- **District**
- Finer Topo
- Regions
- Special
- Topographical

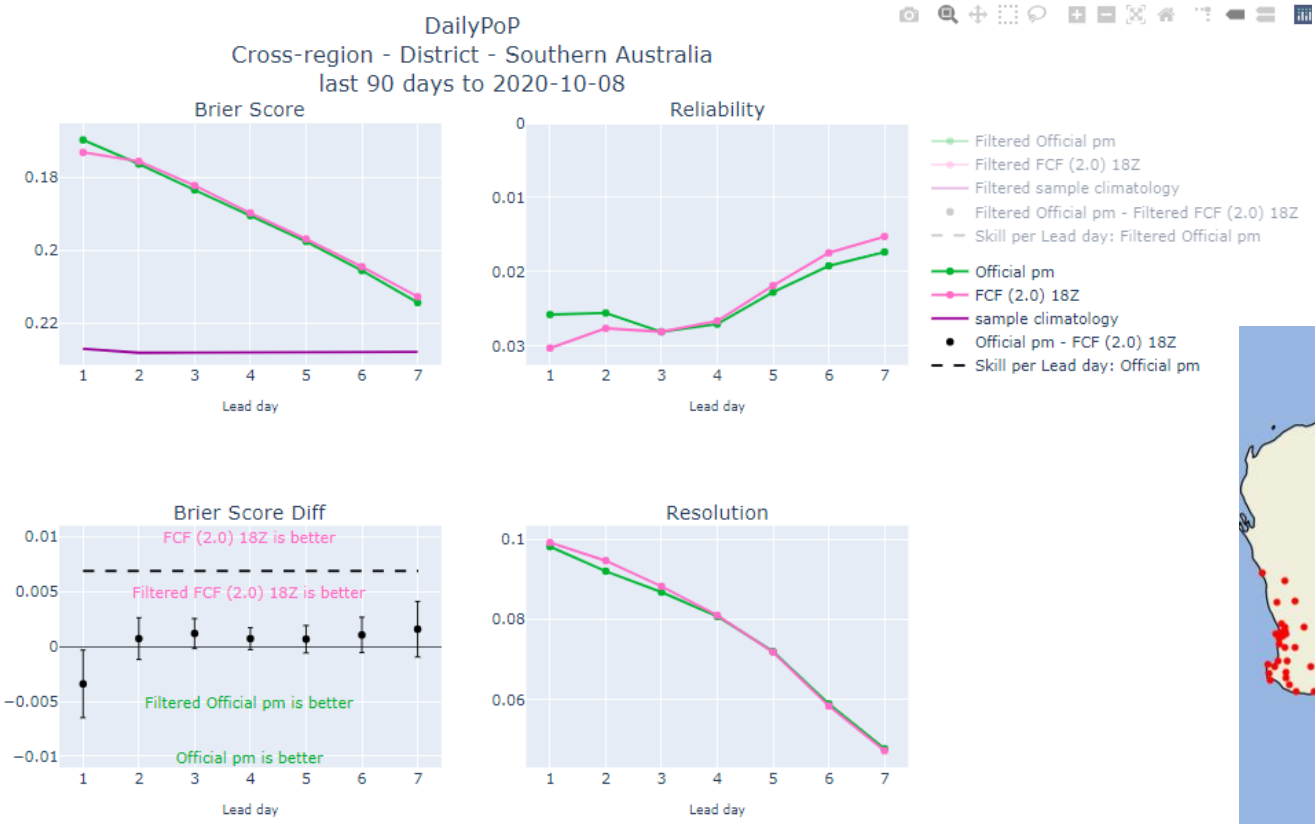
Station group
- East Coast & Ranges Australia
- Inland Australia
- **Southern Australia**
- Tropics Australia

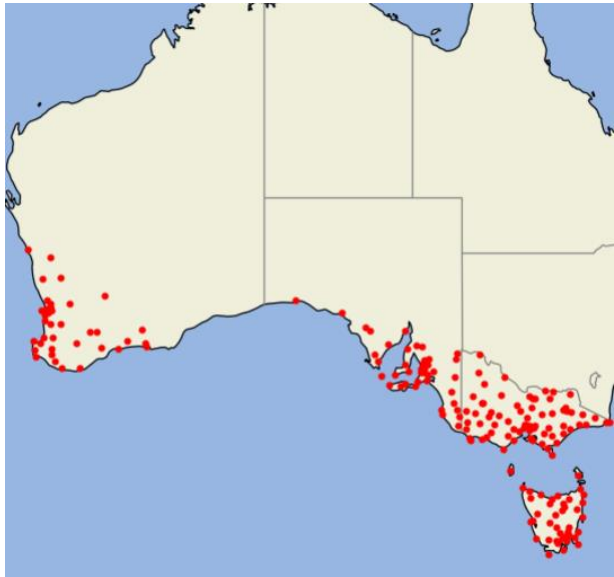
Season
- **Last 90 days**
- Autumn 2020
- Summer 2019-20
- Spring 2019
- Winter 2019
- Autumn 2019

Param
- **DailyPoP**
- DailyPoP1
- DailyPoP5
- DailyPoP10
- DailyPoP15
- DailyPoP25

Metric
- Scorecard (chance better than)
- Scorecard (mean difference)
- **Brier Score and Components**
- Reliability Diagram
- Relative Economic Value

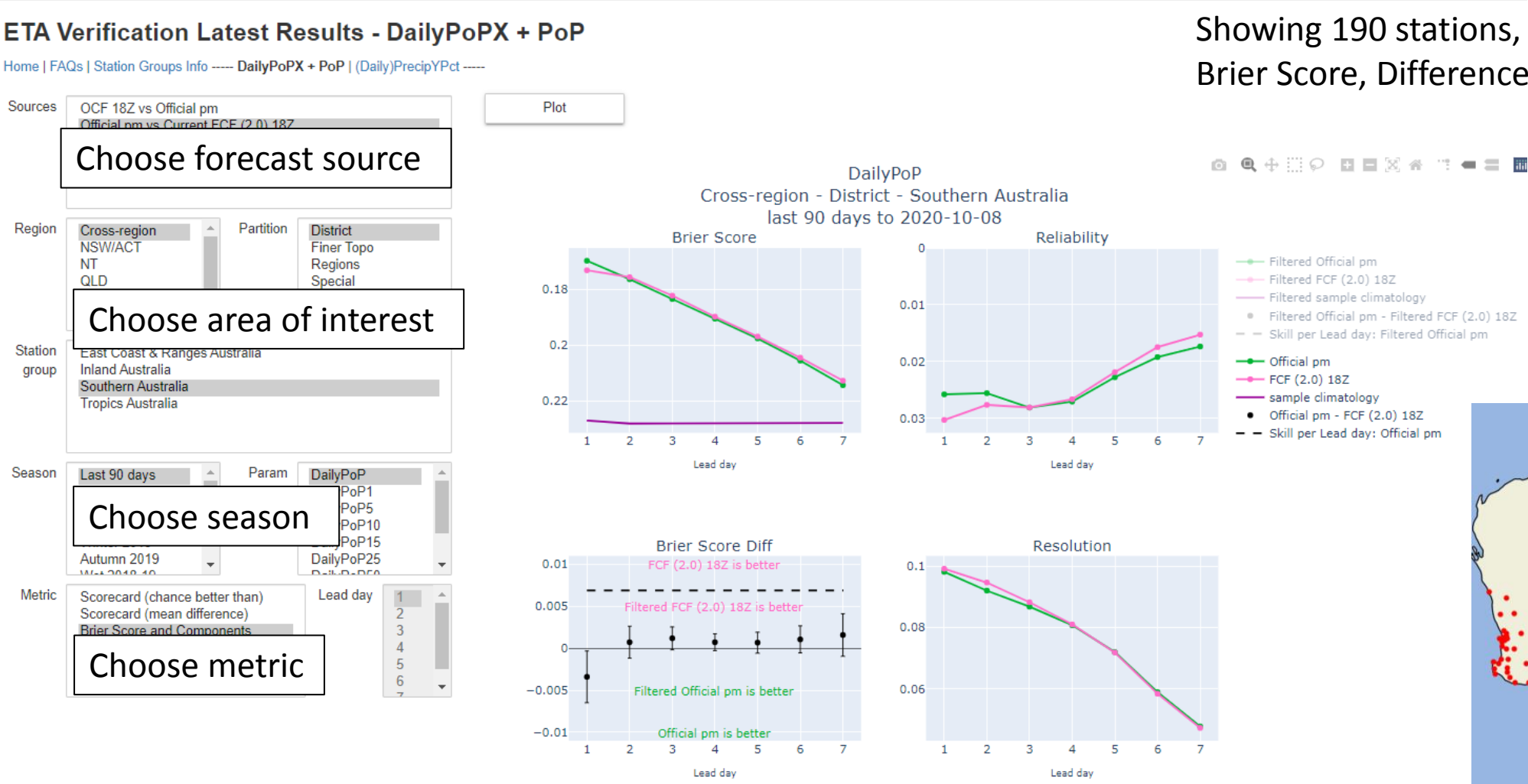
Lead day
- **1**
- 2
- 3
- 4
- 5
- 6

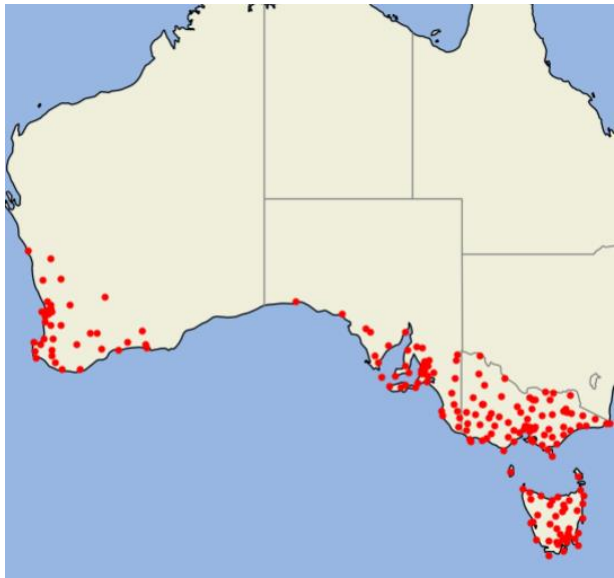




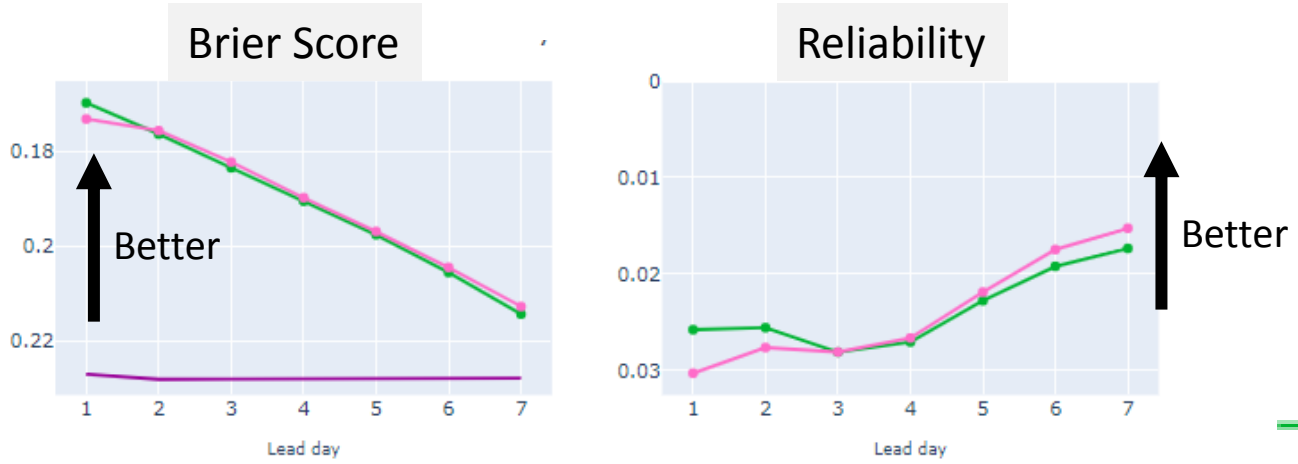
Chance of any rain Verification Results

Showing 190 stations, Winter (wet season), Brier Score, Difference and Components

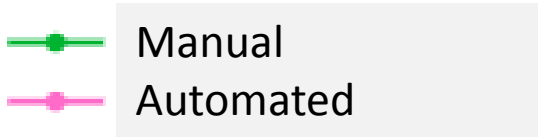


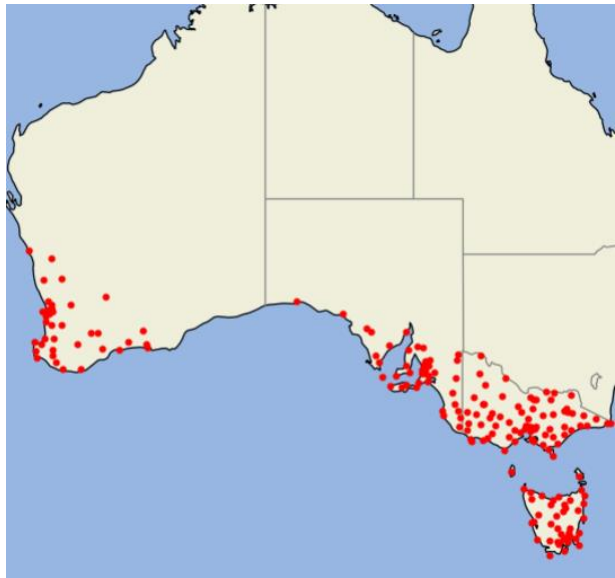


Chance of any rain Verification Results



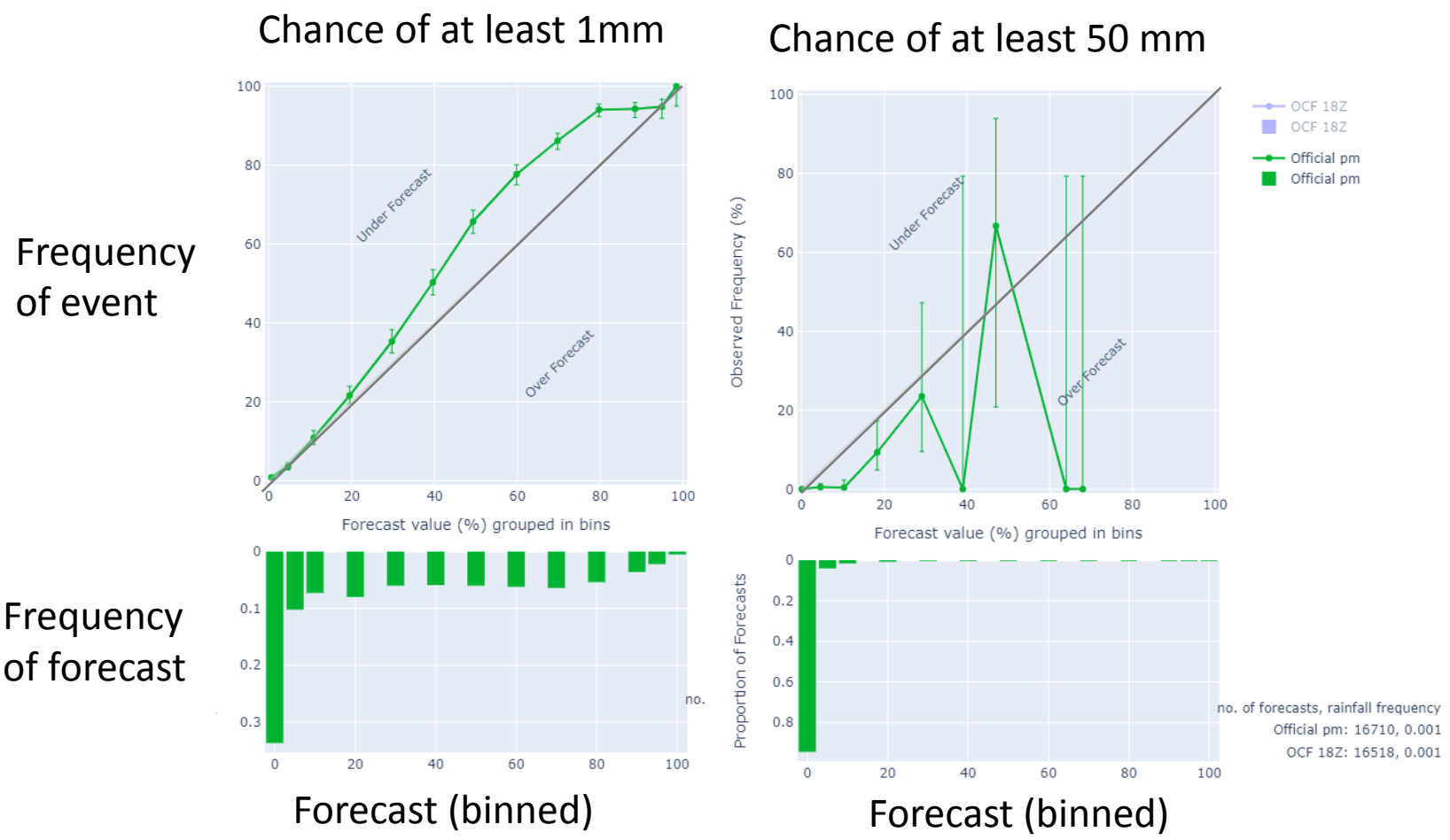
- Both better than climatology to Day 7
- Difference not substantial
- Difference not highly significant
- Signal Days 2 to 7 suggests we rely on the pink forecast
- Strong signal green better at Day 1 (tomorrow), of order less than one lead-day of skill





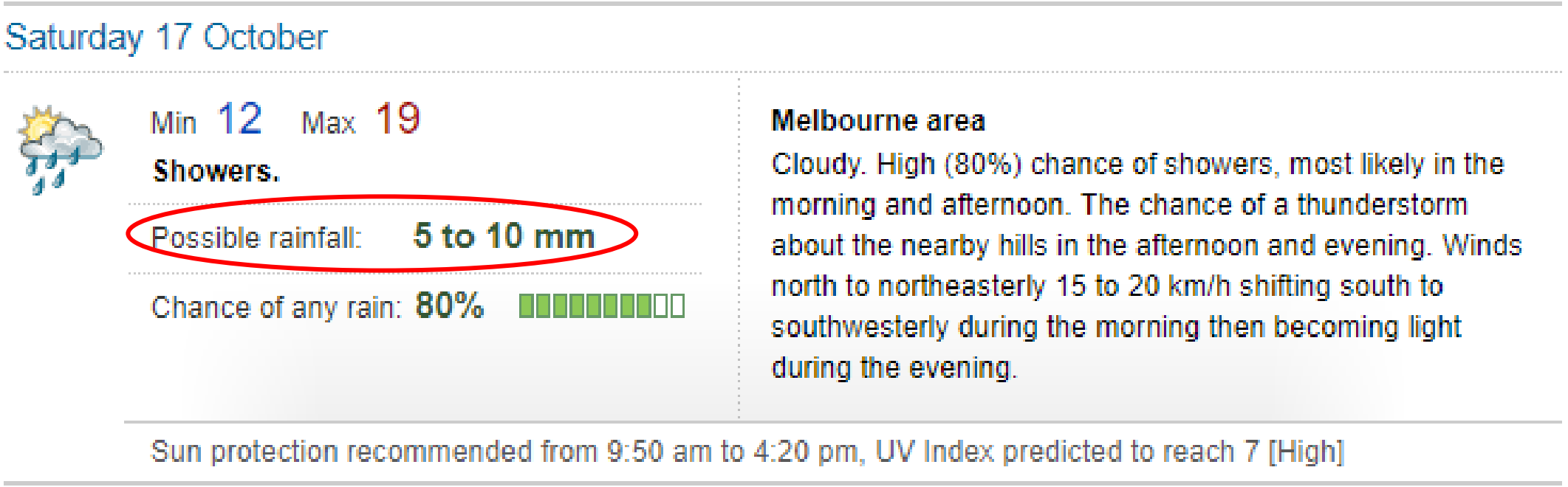
Reliability

In addition to the Reliability Component of the Brier Score, we create Reliability Diagrams



When bins have too few forecasts, Reliability Diagrams are noisy.

The Reliability and Resolution components of the Brier Score are sensitive to the bins used.



To verify: Possible rainfall range

First need definition: Lower value (5mm) is median (50th percentile)

Upper value (10 mm) is 75th percentile.

Initial efforts to assess percentiles
- assessed reliability only, no measure of resolution

ETA Verification Latest Results - (Daily)PrecipYPct

Home | FAQs | Station Groups Info ----- DailyPoPX + PoP | **(Daily)PrecipYPct** ----- (CTRL + click for multiple sources)

Sources: Official pm; OCF 18Z; Current FCF (2.0) 18Z

Filtered: False; True

Plot

Region: Cross-region; NSW/ACT; NT; QLD; SA; TAS; VIC

Partition: District; Finer Topo; Regions; Special; Topographical

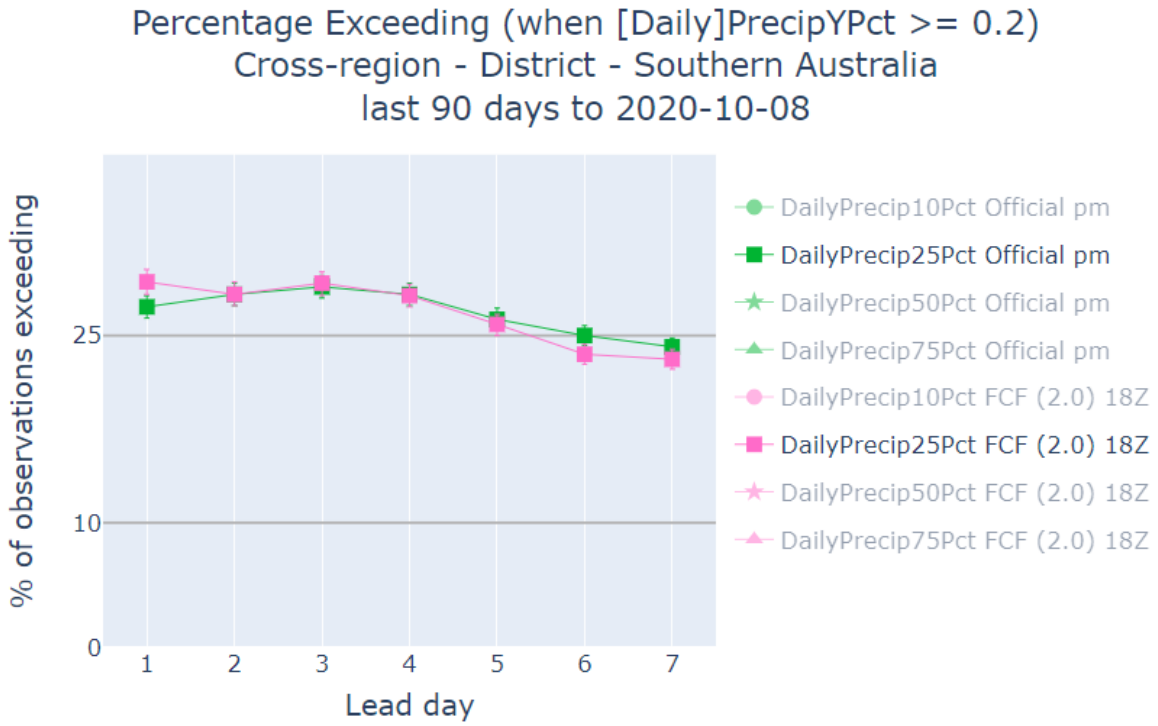
Station group: East Coast & Ranges Australia; Inland Australia; Southern Australia; Tropics Australia

Season: Last 90 days; Autumn 2020; Summer 2019-2(; Spring 2019; Winter 2019; Autumn 2019; Wet 2018-19

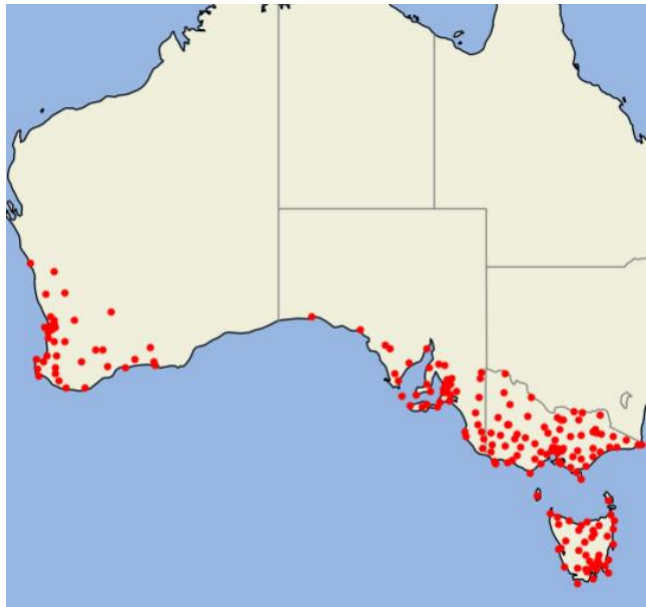
Param: DailyPrecip10Pct; DailyPrecip25Pct; DailyPrecip50Pct; DailyPrecip75Pct; Precip10Pct (by lead day); Precip25Pct (by lead day); Precip50Pct (by lead day)

Metric: Percentage Exceeding (by lead time); Percentage Exceeding (by fcst mm)

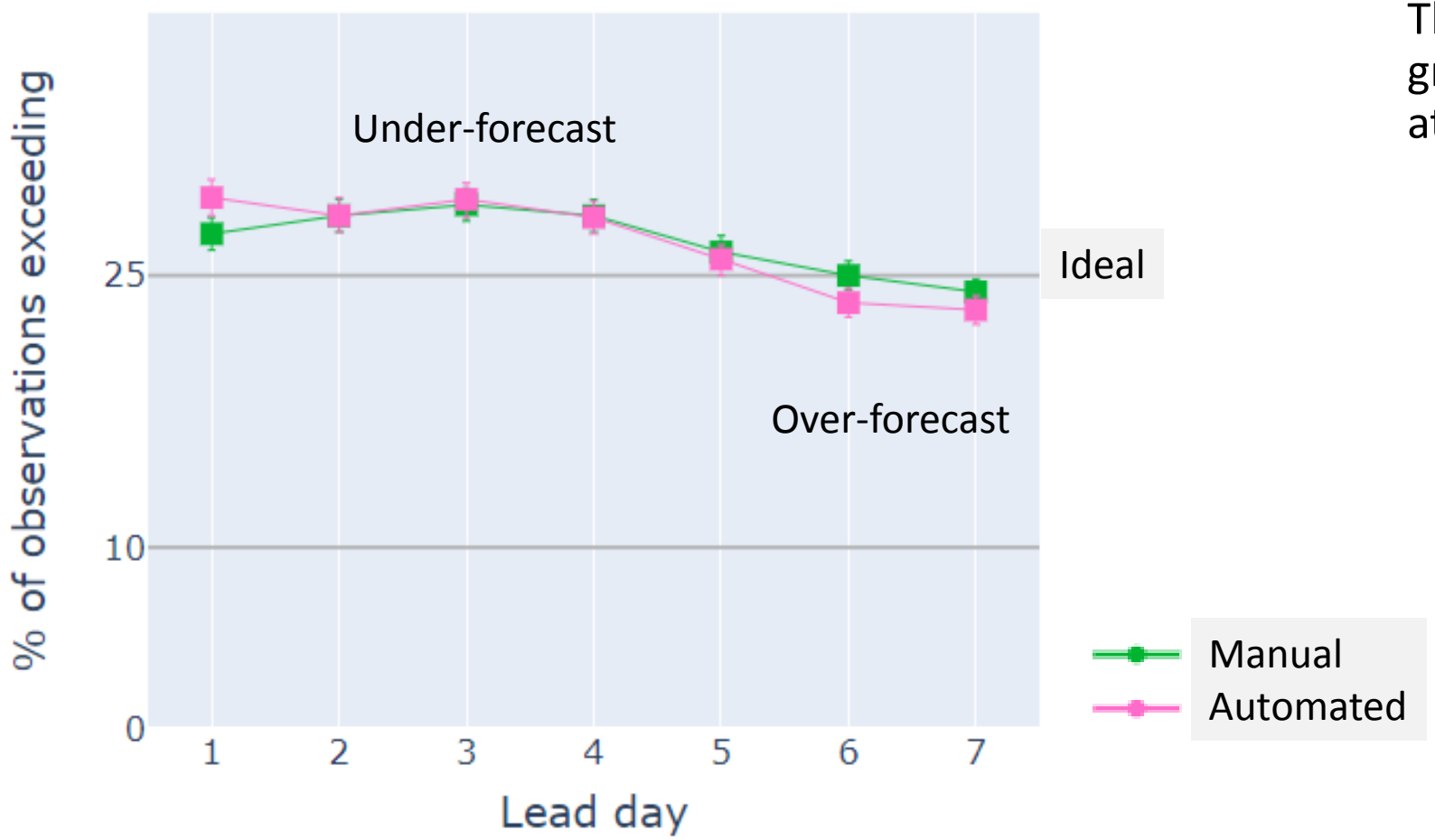
Lead day: 1; 2; 3; 4; 5; 6; 7



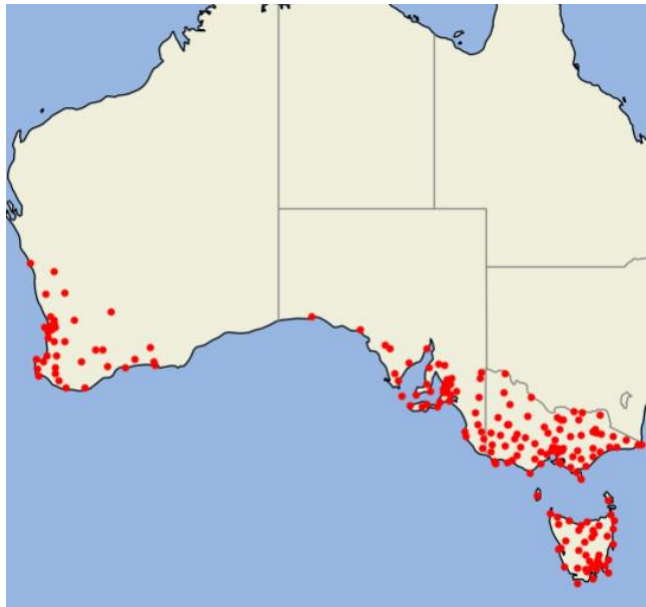
This graph shows pink and green are under-forecasts at lead days 1 to 4.



Initial efforts to assess percentiles
- assessed reliability only, no measure of resolution

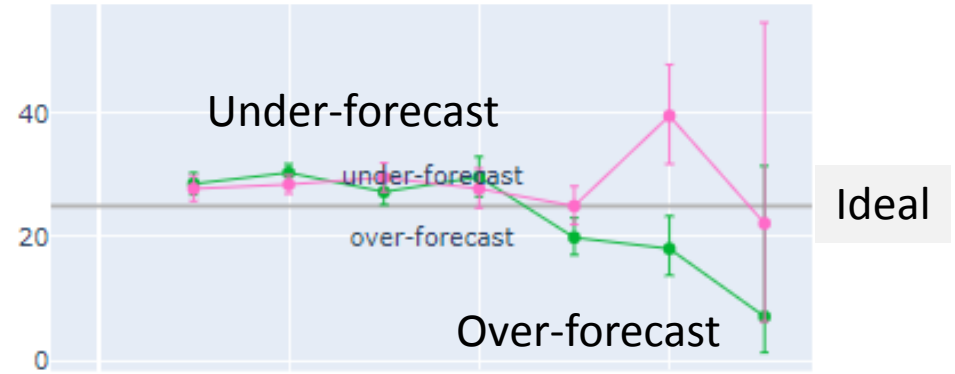


This graph shows pink and green are under-forecasts at lead days 1 to 4.



75th percentile – Reliability as a function of forecast value

Lead Day 2



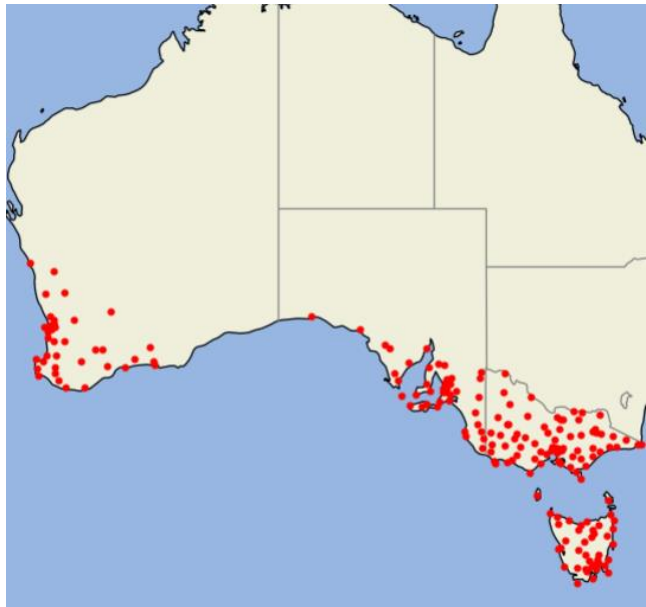
This graph shows that the green forecasts above 15 mm were likely to be too high.

It is still a very coarse view.

The forecasts have different characteristics.

Which is better?

Manual
Automated



Common error measures

Root Mean Square Error

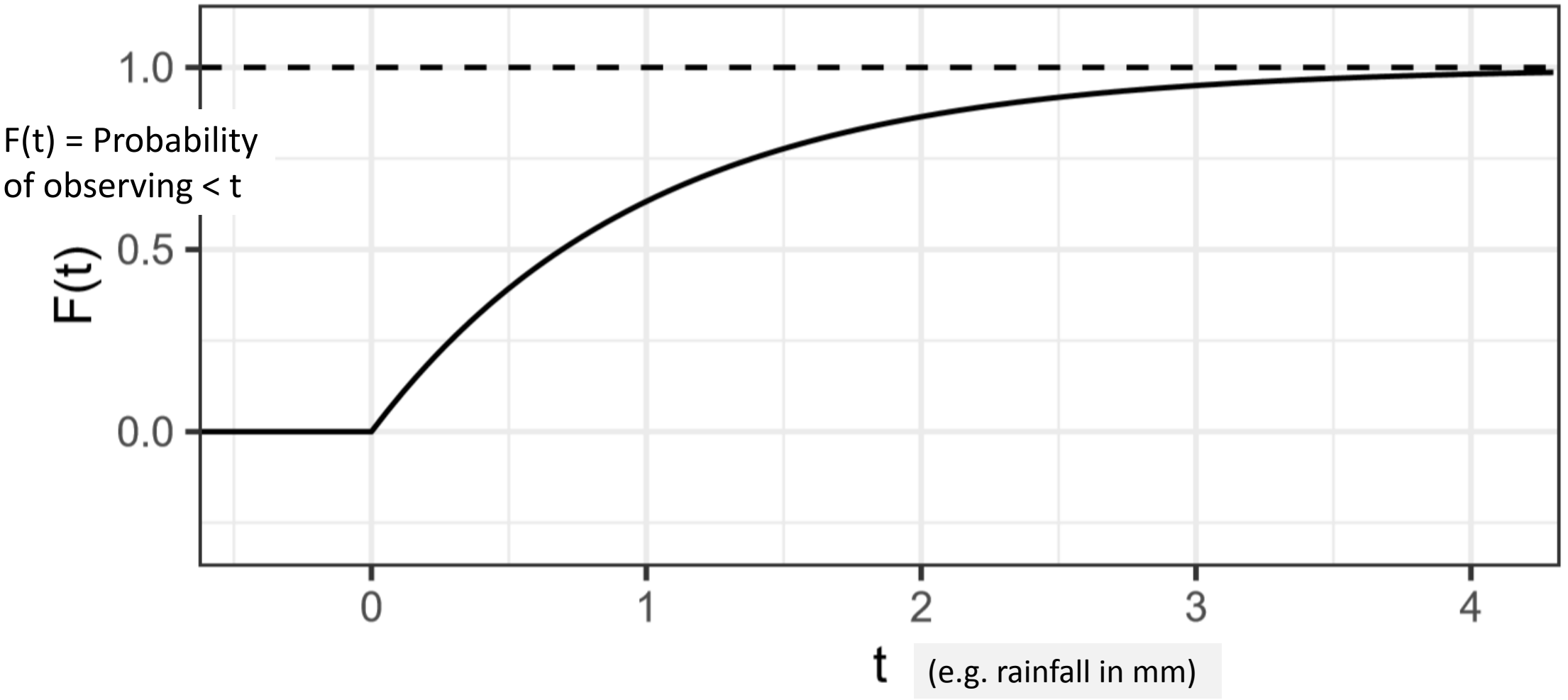
- An error of 10 is considered 4 times as bad as an error of 5.
- Very popular

Mean Absolute Error

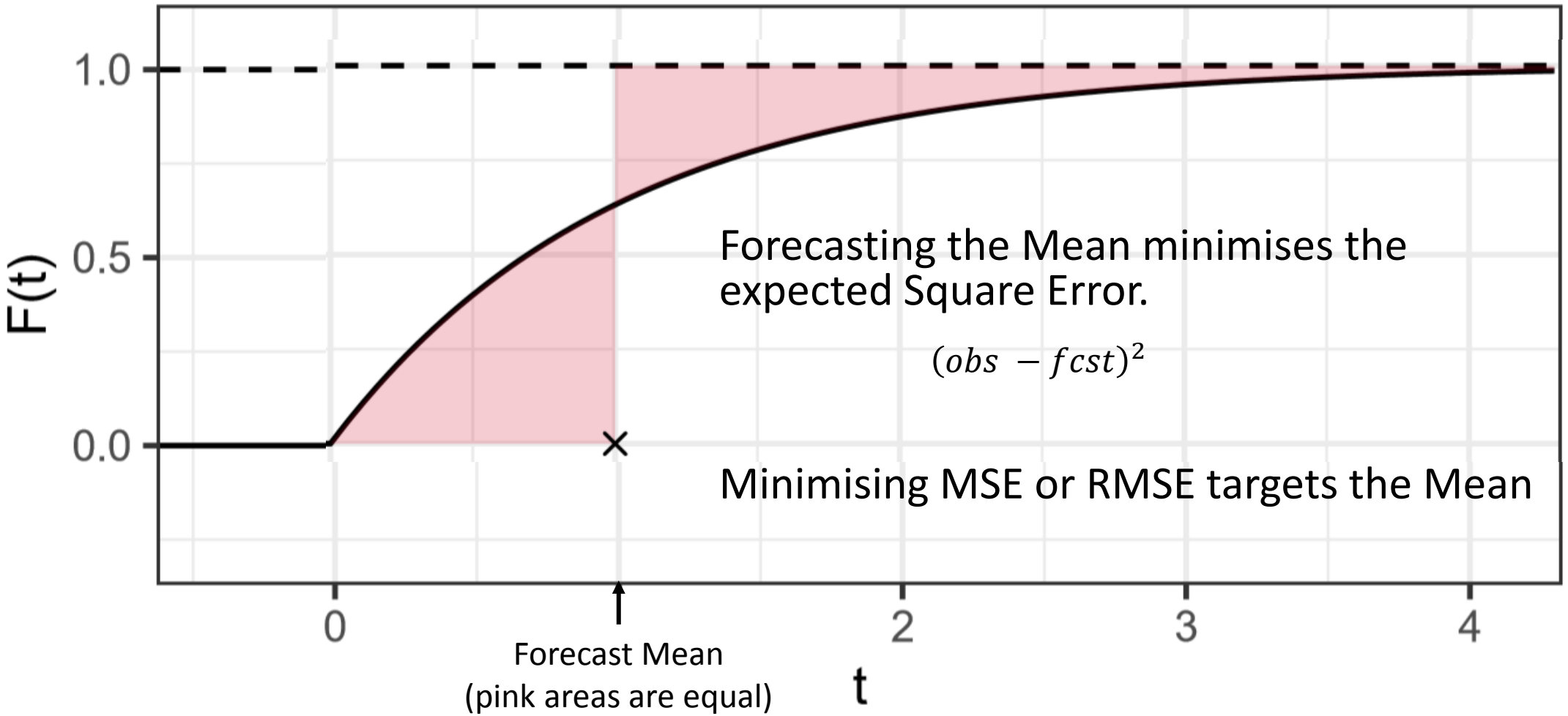
- An error of 10 is considered twice as bad as an error of 5.
- Used when you don't want sensitivity to large errors

But what are they actually targeting?
Are they relevant to rainfall forecasts (skewed distribution, bounded by zero)?

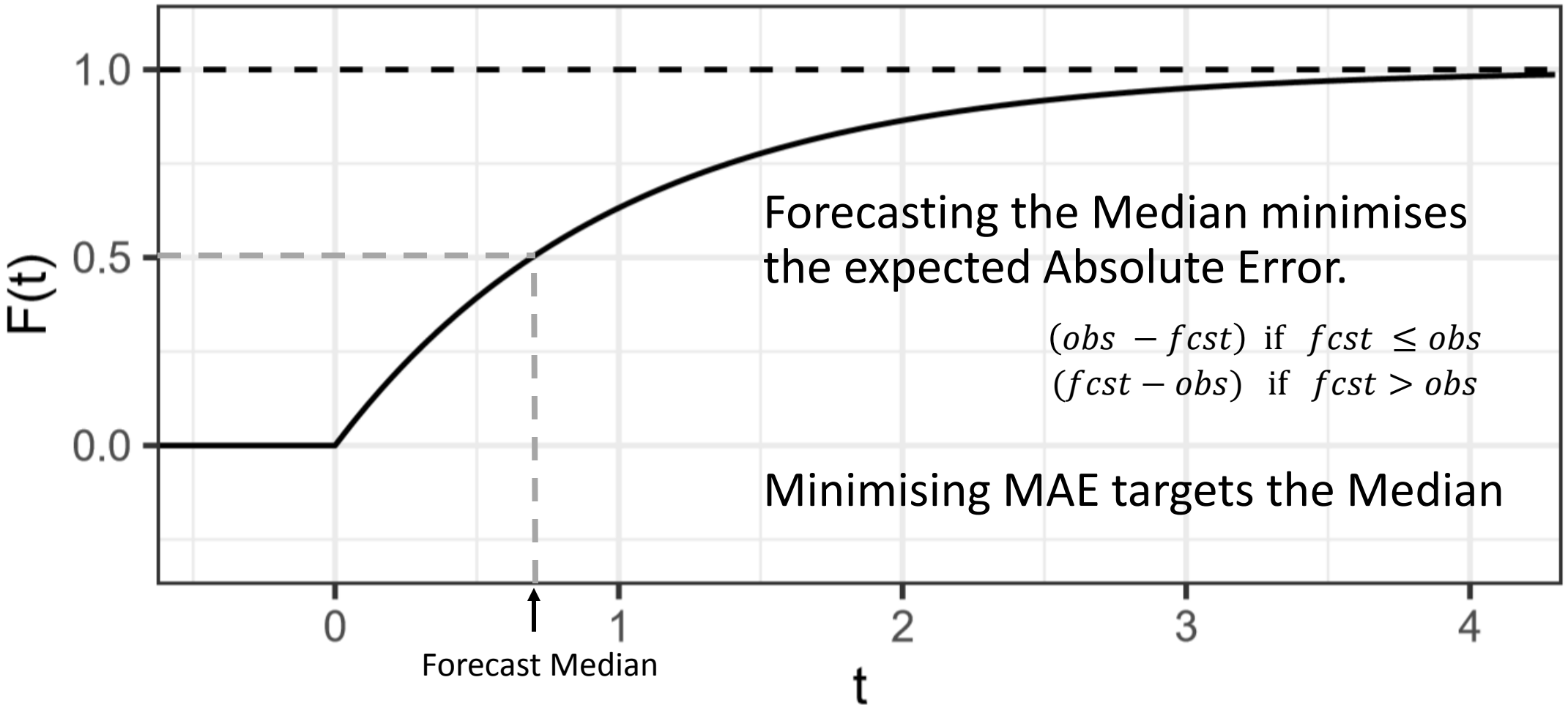
Example Cumulative Density Function (the forecaster's belief)
F(t) = Forecast probability {observing < t}



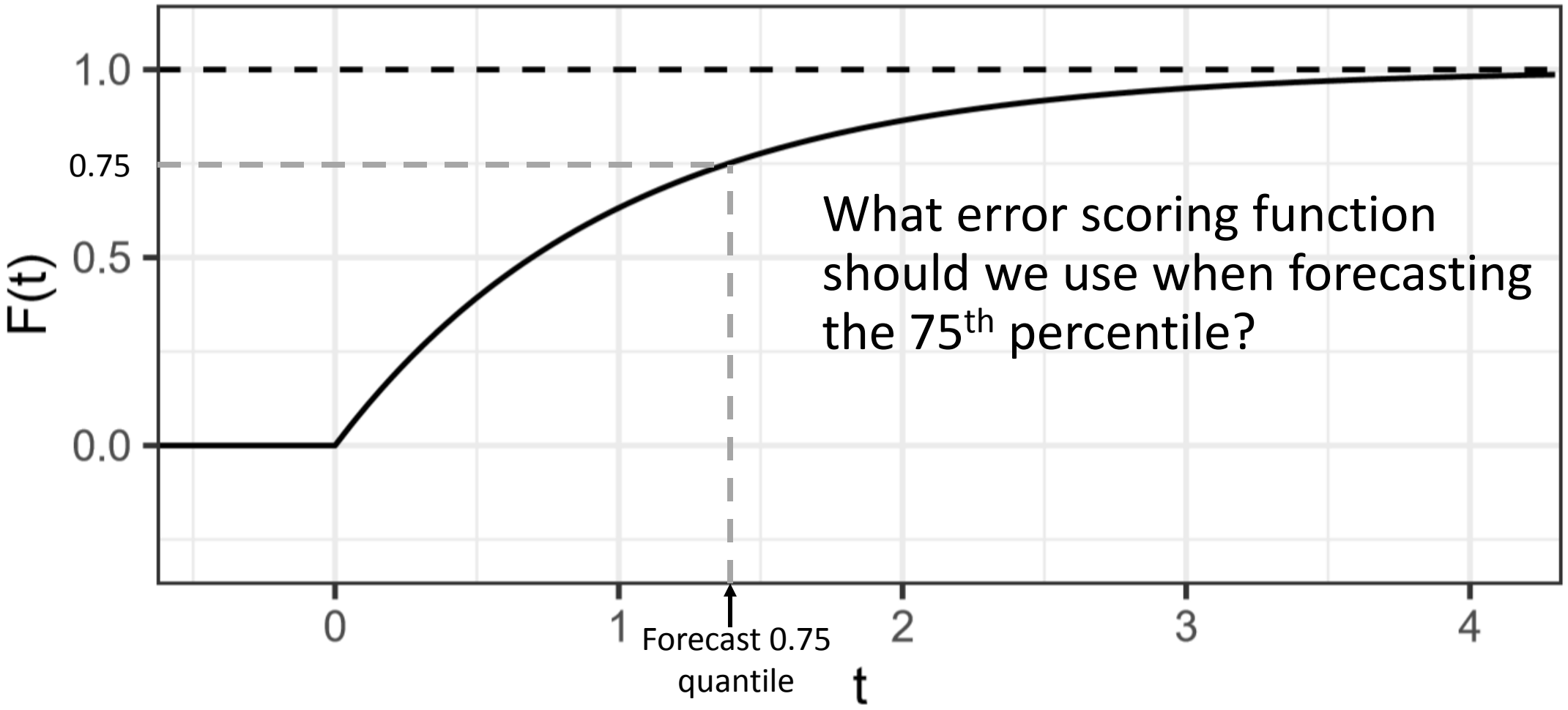
Example Cumulative Density Function
F(t) = Forecast probability {observing < t}



As it turns out…



Example Cumulative Density Function
F(t) = Forecast probability {observing < t}



Quantile Scoring Function

For a forecast of the 75th percentile of the forecast distribution, we can score the forecast with an error of

$$0.75\ (obs\ -fcst)\ \text{if}\ \ fcst\ \leq obs$$
$$0.25\ (fcst - obs)\ \ \text{if}\ \ fcst > obs$$

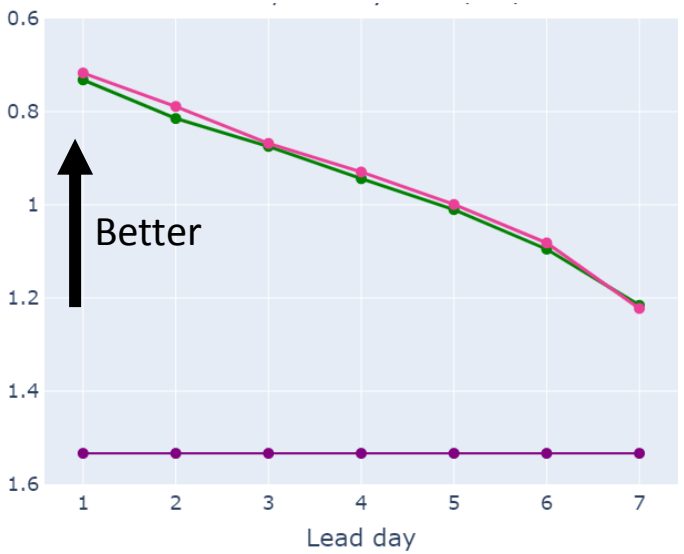
Minimising this error ensures we are targeting the 75th percentile.

See abstract for formula generalised to any quantile.

Reference: Gneiting, J. Amer Statist Assoc. 2011, *Making and evaluating point forecasts*, https://www.tandfonline.com/doi/abs/10.1198/jasa.2011.r10138

Quantile Score and Difference vs Lead Day

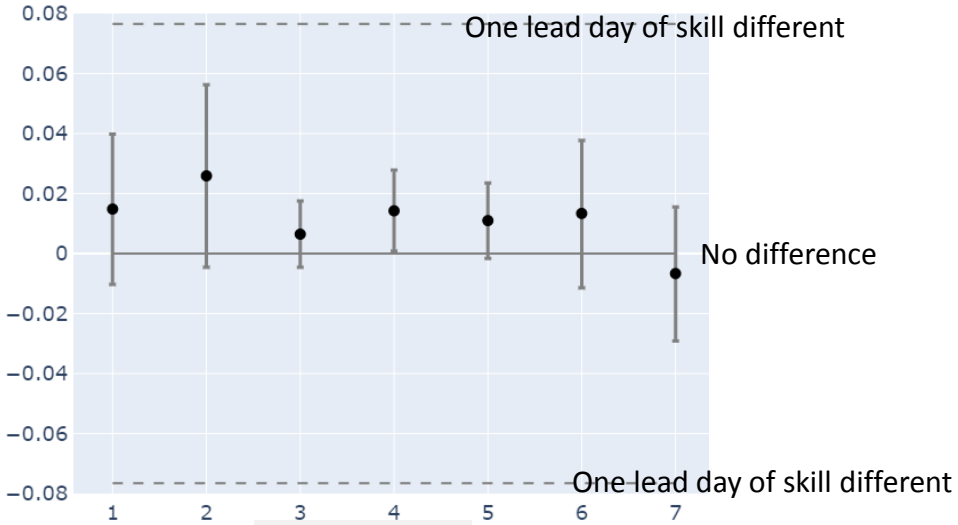
Mean Quantile Score

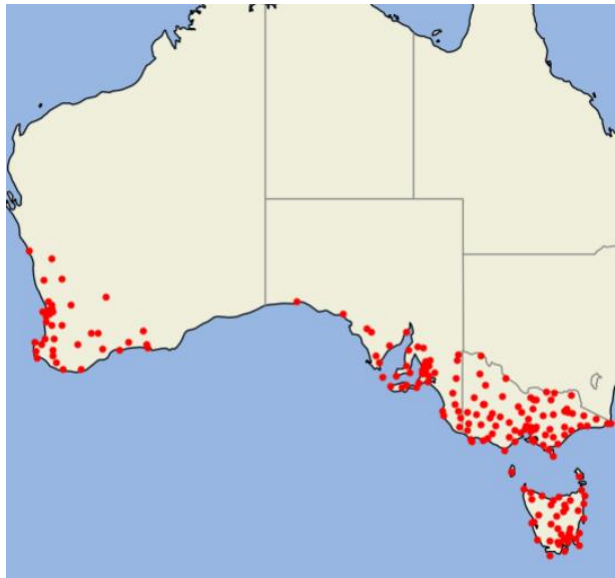


- Both better than climatology to Day 7
- Difference not substantial
- Suggests we can rely on the pink forecast

Manual
Automated
Sample climatology

Difference in Quantile Score





Lesson learnt – do a better literature review

Particularly Interesting and Useful references:

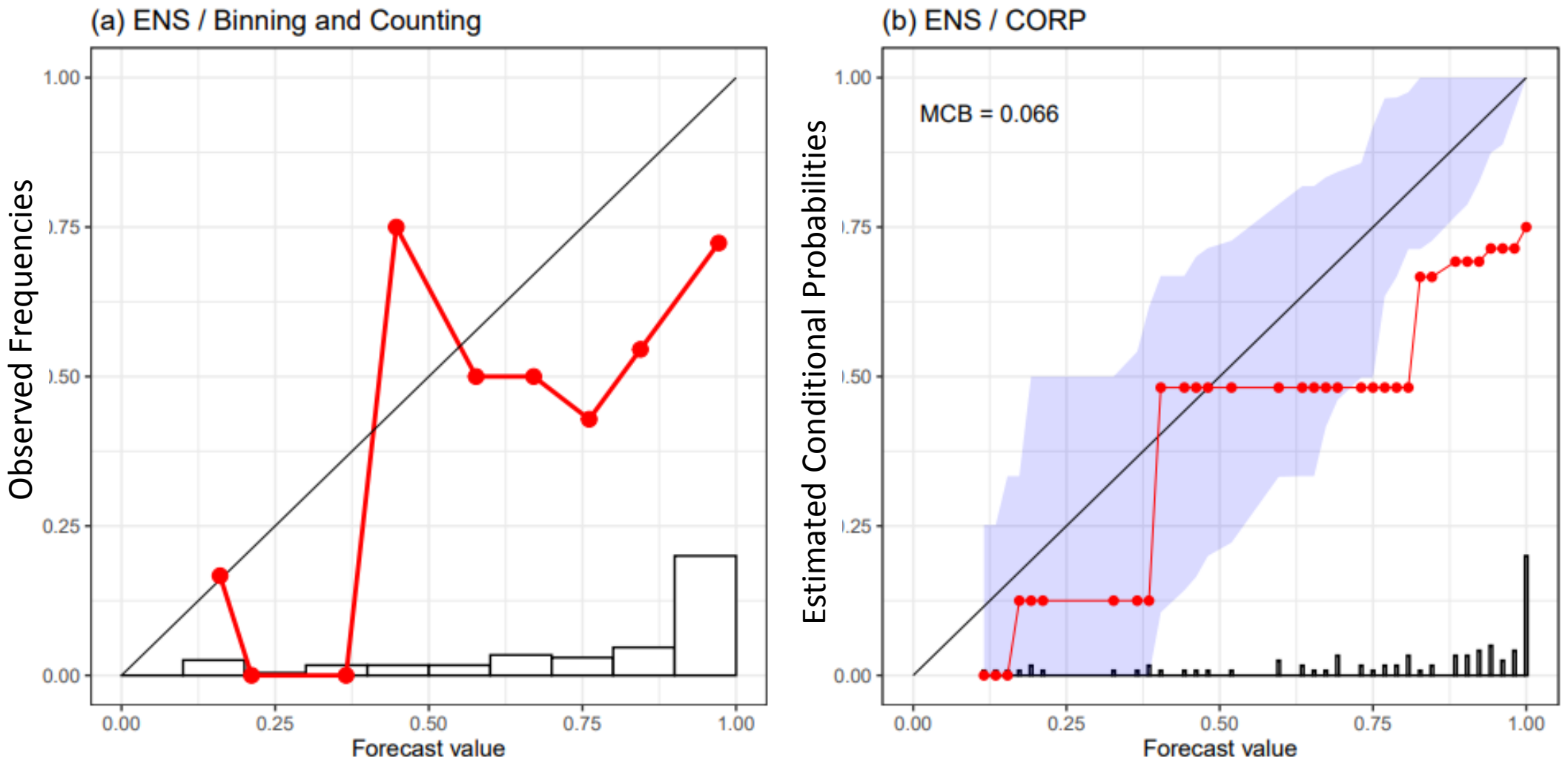
Bentzien & Friederichs, QJRMS 2014, *Decomposition and graphical portrayal of the quantile score*, https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.2284

> Technique for determining reliability and resolution components of quantile scoring function

Dimitriadis, Gneiting & Jordan, 2020, *Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited*, https://arxiv.org/pdf/2008.03033.pdf

> Technique that removes sensitivity to choice of bins.
> Reliability curve is forced to be non-decreasing via isotonic regression.
> Reliability and Resolution Components are more confidently meaningful.

From Fig 1 of Dimitriadis et al.



Planned Adjustment to our Verification

Add MSE (or RMSE) analysis of forecast Expected Precipitation

Add Quantile Scoring Function analysis of forecast quantiles for Precipitation

Explore using technique of Dimitriadis et al to explore Reliability and Resolution (for all forecasts).

References

Gneiting, International Journal of Forecasting 2011, *Quantiles as optimal point forecasts,* https://doi.org/10.1016/j.ijforecast.2009.12.015

Geniting and Katzfuss, Annual Review of Stat. Appl. 2014, *Probabilistic forecasting,* https://doi.org/10.1146/annurev-statistics-062713-085831

Already mentioned:

Bentzien & Friederichs, QJRMS 2014, *Decomposition and graphical portrayal of the quantile score*, https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.2284

Dimitriadis, Gneiting & Jordan, 2020, *Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited*, https://arxiv.org/pdf/2008.03033.pdf

Gneiting, J. Amer Statist Assoc. 2011, *Making and evaluating point forecasts*, https://www.tandfonline.com/doi/abs/10.1198/jasa.2011.r10138

Abstract

The Bureau of Meteorology issues various forecasts for daily rainfall including the mean, the median, other quantiles, and the chance of exceeding various thresholds.

We verify the chance of exceeding a given threshold using the Brier Score. However, our initial attempts to assess forecasts such as the 90th percentile of daily rainfall was very simply showing the proportion of times the observation exceeded the forecast. This showed some measure of the reliability of the forecasts but gave no overall measure of its skill. A climatological forecast would score perfectly on this measure.

Recently, we learnt about consistent scoring functions for single value quantile forecasts. For forecast x predicting the α quantile of the forecast distribution, and observation y, we can score the forecast as follows

$$\alpha|x - y| \quad \text{if} \quad x \leq y$$

$$(1-\alpha)\,|x - y| \quad \text{if} \quad x > y$$

For the median forecast, the score is essentially the mean absolute error.

By introducing this score, we will be able to track improvement in forecasts of a particular quantile and to compare two forecasts of the same event in a meaningful way. To compare the whole forecast distribution, we use the (Continuous) Ranked Probability Score. However, verifying a point of the forecast probability distribution is important if that value is a prominent aspect of one's forecast service, or known to be used by a client for a particular decision.

This talk will showcase techniques for verifying a rainfall probability distribution, including point values from the distribution, and discuss the decisions being informed by the verification.