Australia's National Science Agency

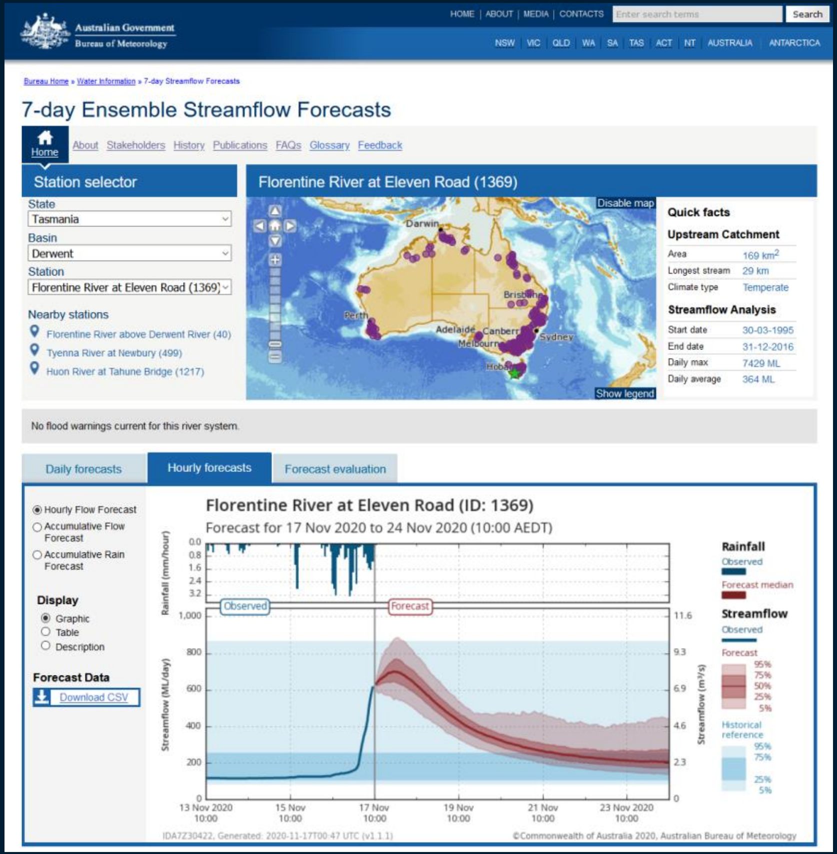
Probability integral transforms for verifying probabilistic predictions in hydrology

James Bennett, David Robertson, Andrew Schepen

9IVMW | 22 May 2024

Reliability in hydrological predictions

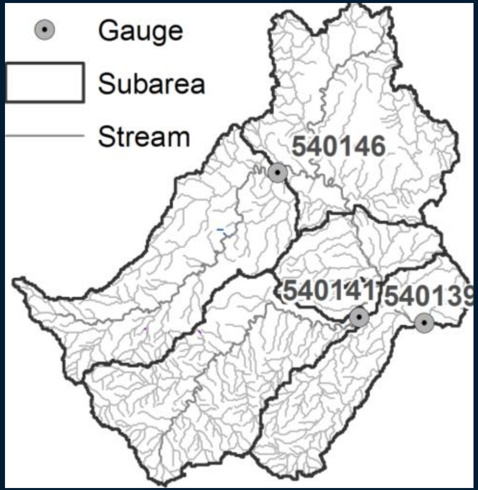
- Probabilistic/ensemble predictions increasingly common in hydrology
- Gneiting et al. 2007:
    - ‘Sharpness, subject to reliability’
    - Reliability can be checked with Probability Integral Transforms



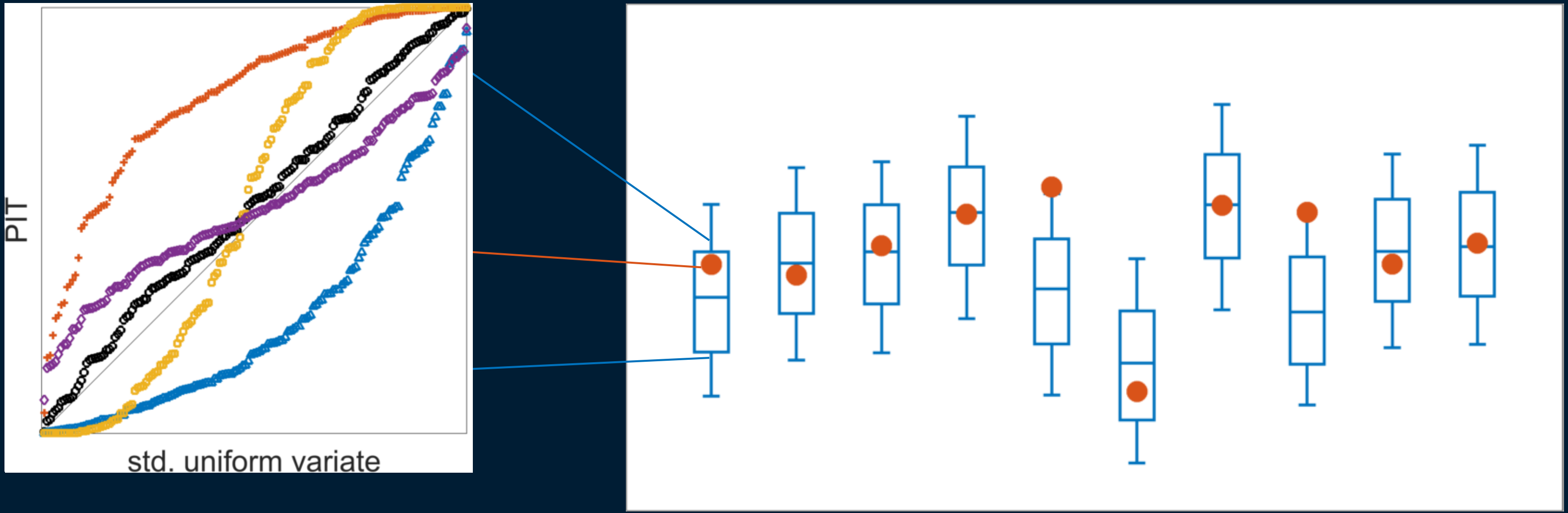
http://www.bom.gov.au/water/

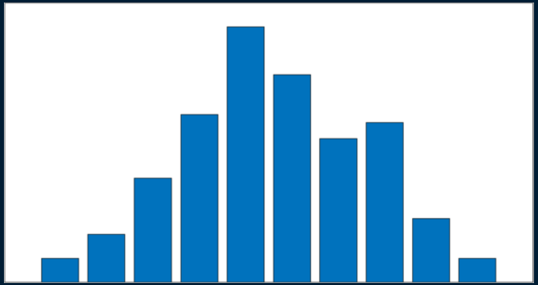
Reliability and the information value chain

- Reliable forecast probabilities translate directly to decisions
  - No hedging needed
- Uncertainty can be propagated downstream
- Outputs can be used directly in decision models



Reliability – probability integral transform

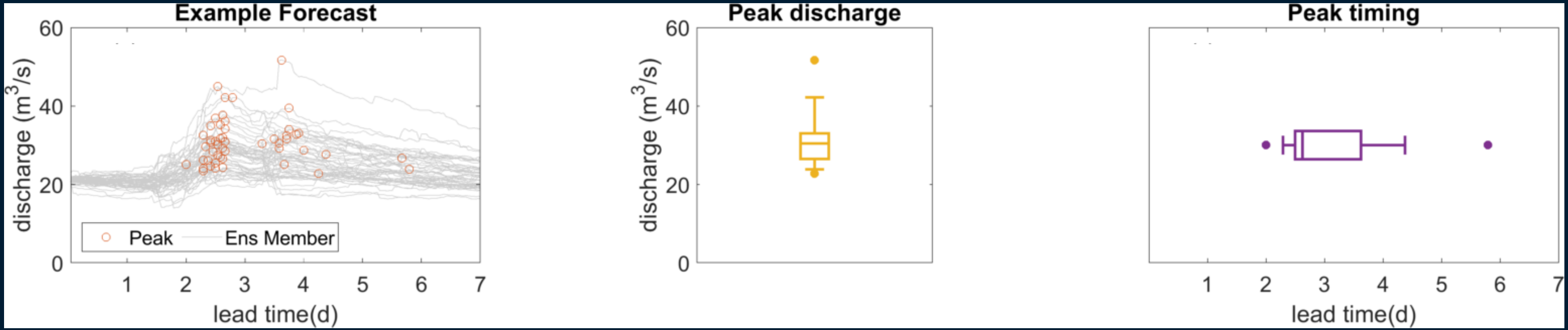




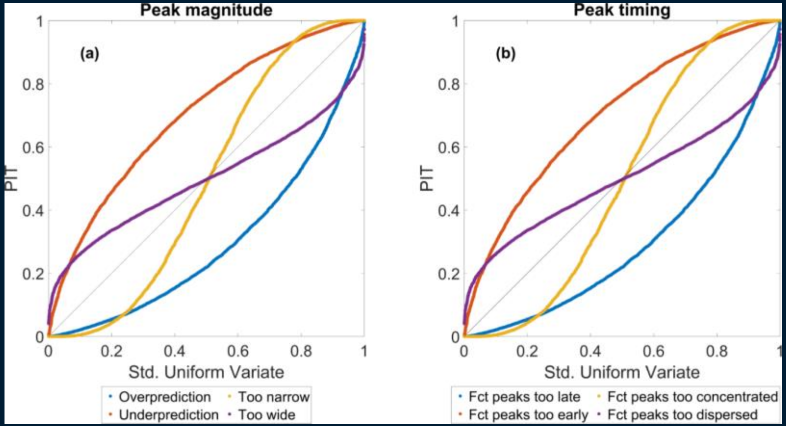
rank histogram

- PIT can use fewer forecast-obs pairs than rank histograms

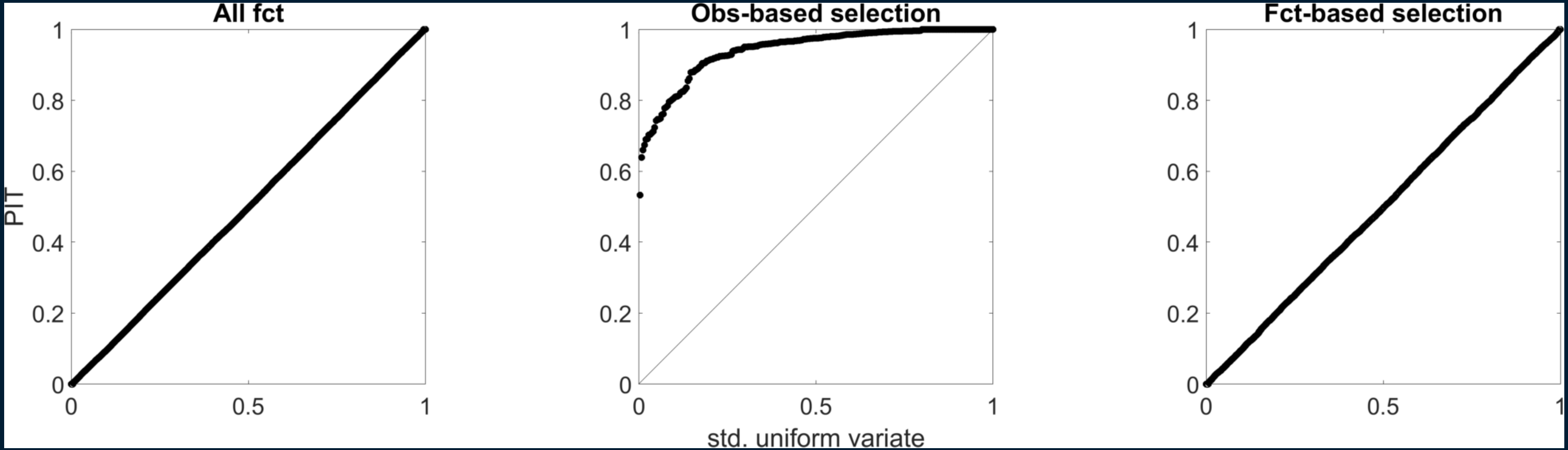
Reliability of flood peak magnitude and timing



- Must condition any stratification on forecasts (Bellier et al. 2017)



Reliability and forecast stratification



Synthetic example replicating Bellier et al 2017

- Obs and forecasts drawn from the same normal distributions
- ‘Flood threshold’ based on 99% quantile of ‘observations’

PIT summary statistics

- Renard et al. 2010
  - alpha-index
  - xi-index (coverage)
- Allov… …arison of sit… …tc.
- Alph… …sed in hydr…

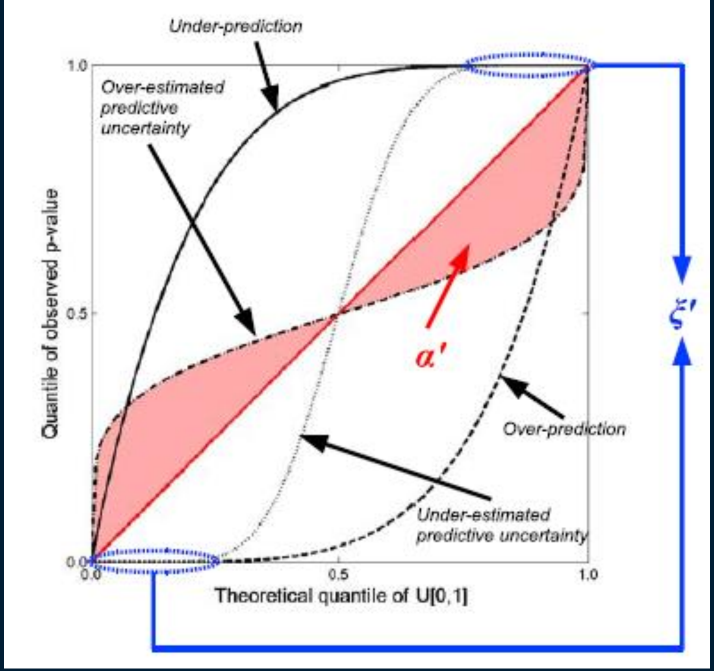
$$\alpha_x = 1 - 2\alpha_x^{'}$$

$$\alpha_x^{'} = \sum_{i=1}^{N_x} \left| p_{x(i)} - p_{x(i)}^{(th)} \right| / N_x$$

$$\xi_x = 1 - \xi_x^{'}$$

$$\xi_x^{'} = \sum_{i=1}^{N_x} \left(1_{\{0,1\}}(p_{x(i)}) \right) / N_x$$

$$1_{\{0,1\}}(z) = \begin{cases} 1 \text{ if } z = 0 \text{ or } z = 1 \\ 0 \text{ otherwise} \end{cases}$$



PIT summary statistics

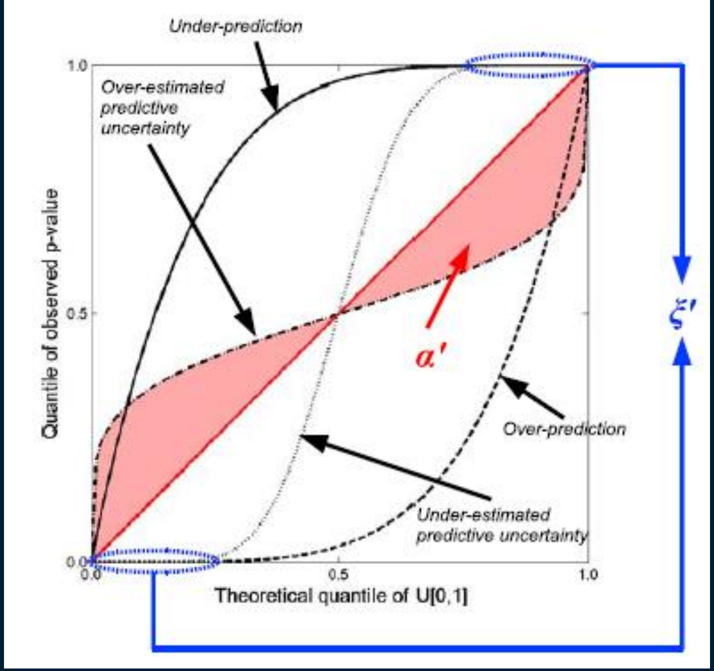
- Comparison to CRPS decomp (Hersbach 2000)

$$\overline{\text{CRPS}} = \overline{\text{Reli}} - \overline{\text{Resol}} + \overline{U}$$

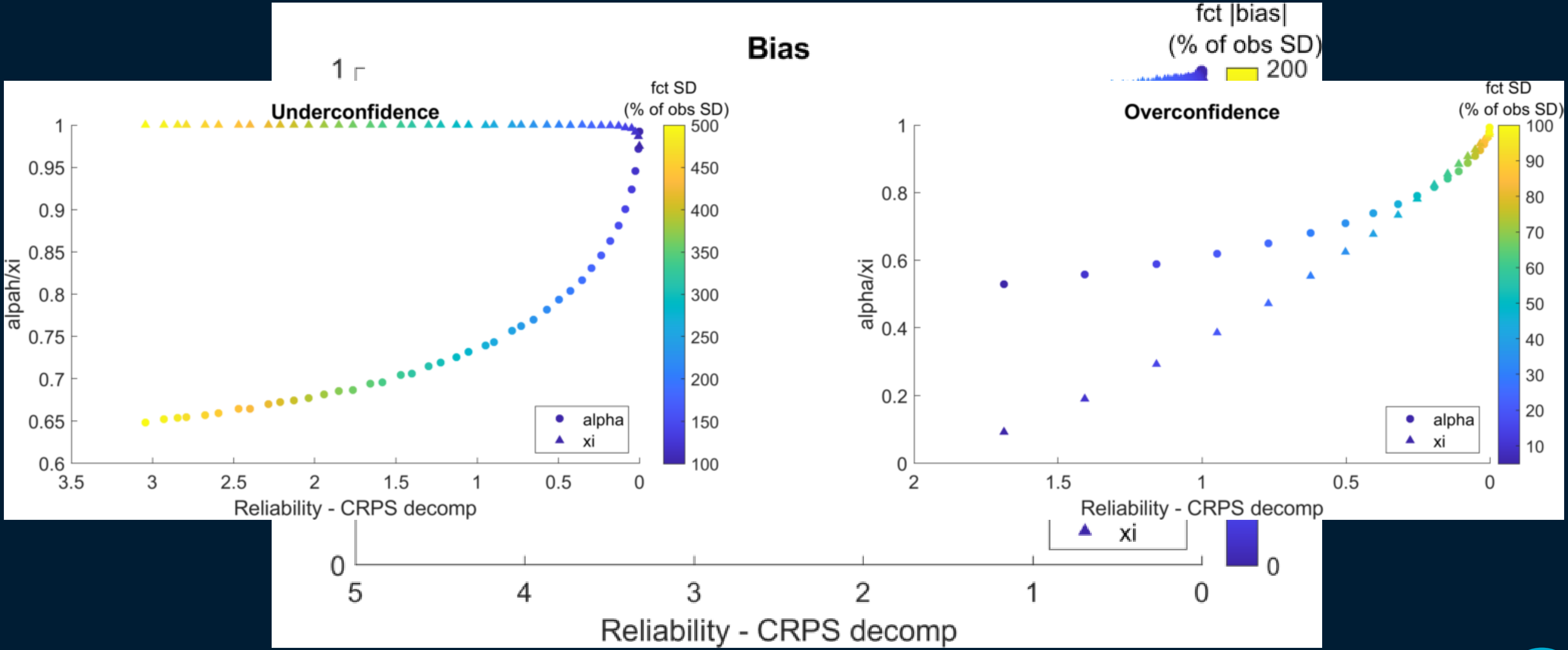
$$\overline{\text{Reli}} = \sum_{i=0}^{N} \overline{g}_i(\overline{o}_i - p_i)^2, \qquad p_i = \frac{i}{N}$$

$$\overline{U} = \int_{-\infty}^{\infty} P_{\text{sam}}(x)[1 - P_{\text{sam}}(x)]\, dx.$$

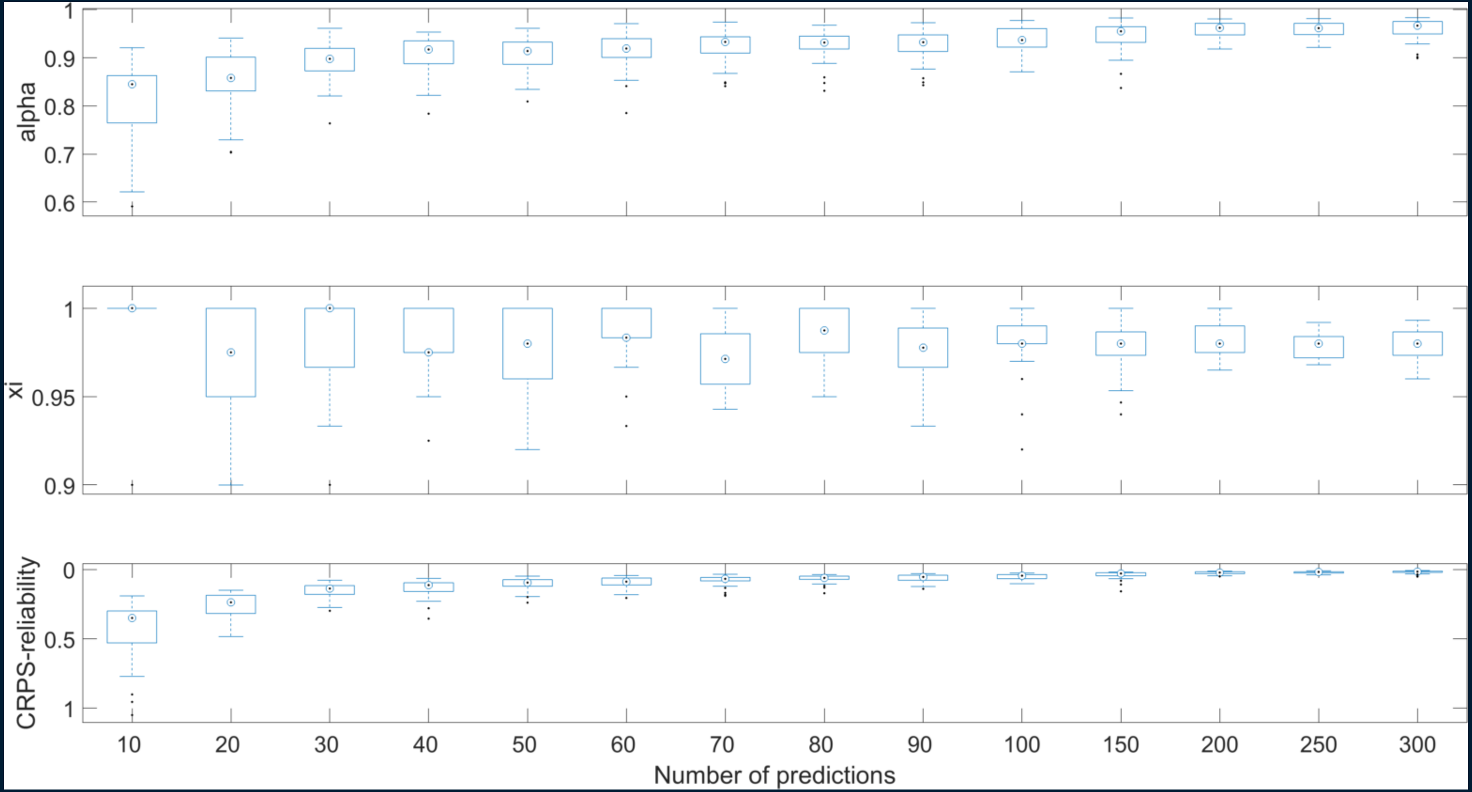
$$\overline{\text{Resol}} = \overline{U} - \sum_{i=0}^{N} \overline{g}_i \overline{o}_i (1 - \overline{o}_i).$$



PIT summary statistics behaviour



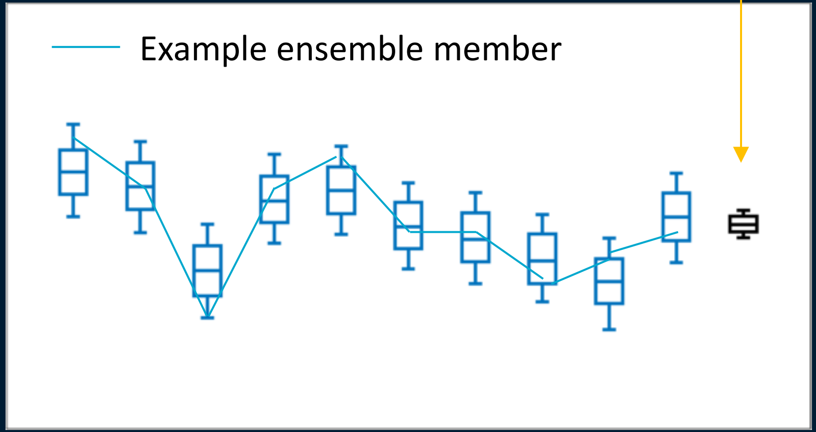
PIT summary statistics behaviour



PIT used to diagnose temporal/spatial structure

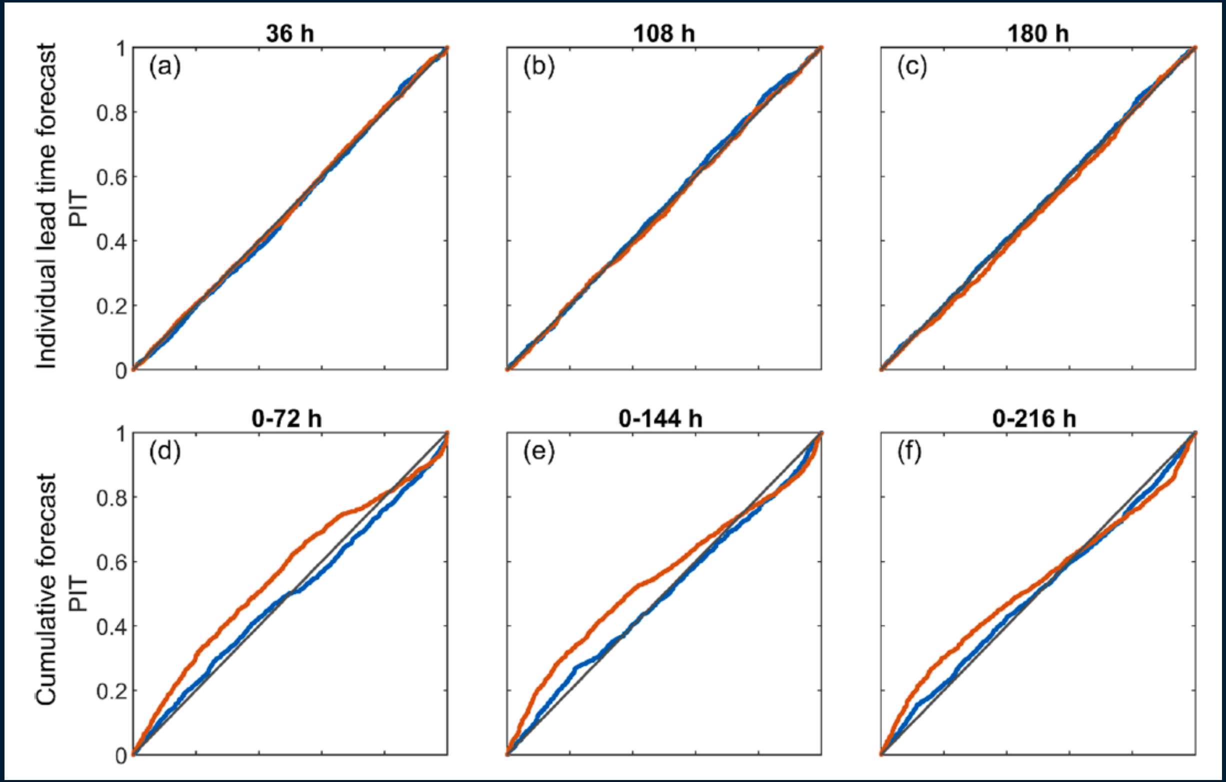
Mean of blue predictions

No autocorrelation



Strong autocorrelation

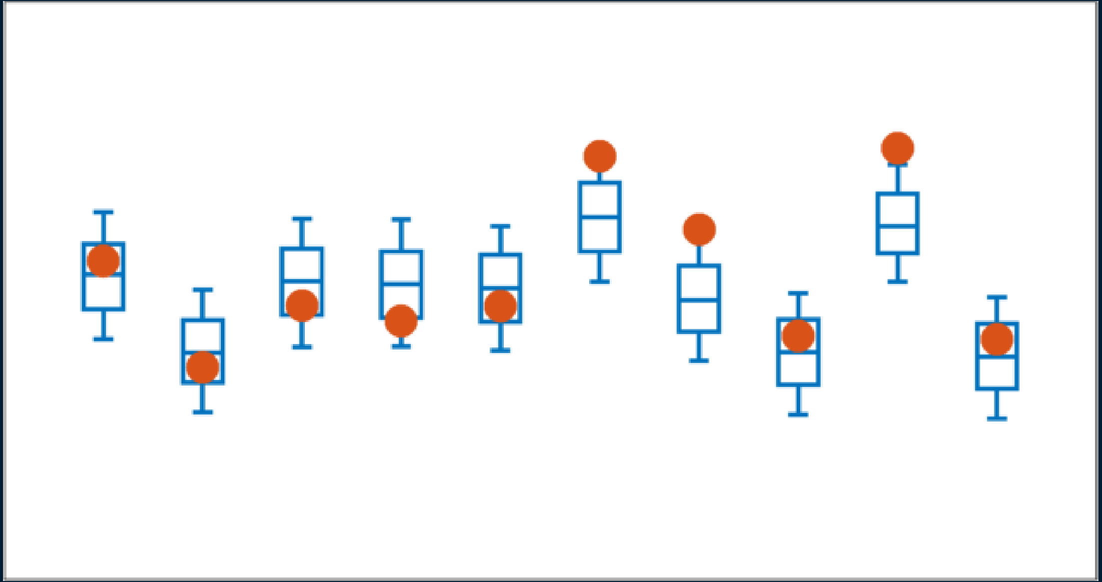
PIT used to diagnose temporal/spatial structure



Shrestha et al. 2020 *J. Hydrol*

PIT used to diagnose non-stationarity

- If trend in a model is not represented in observations, PIT values will have trend
- Can combine with standard trend assessments:
  - Sen's slope
  - Mann-Kendall test



PIT example: the TULIP model

- **T**rend and **U**ncertainty in **L**ong **I**nflow **P**redictions 🌷
- Non-stationary inflow climatology with autocorrelation

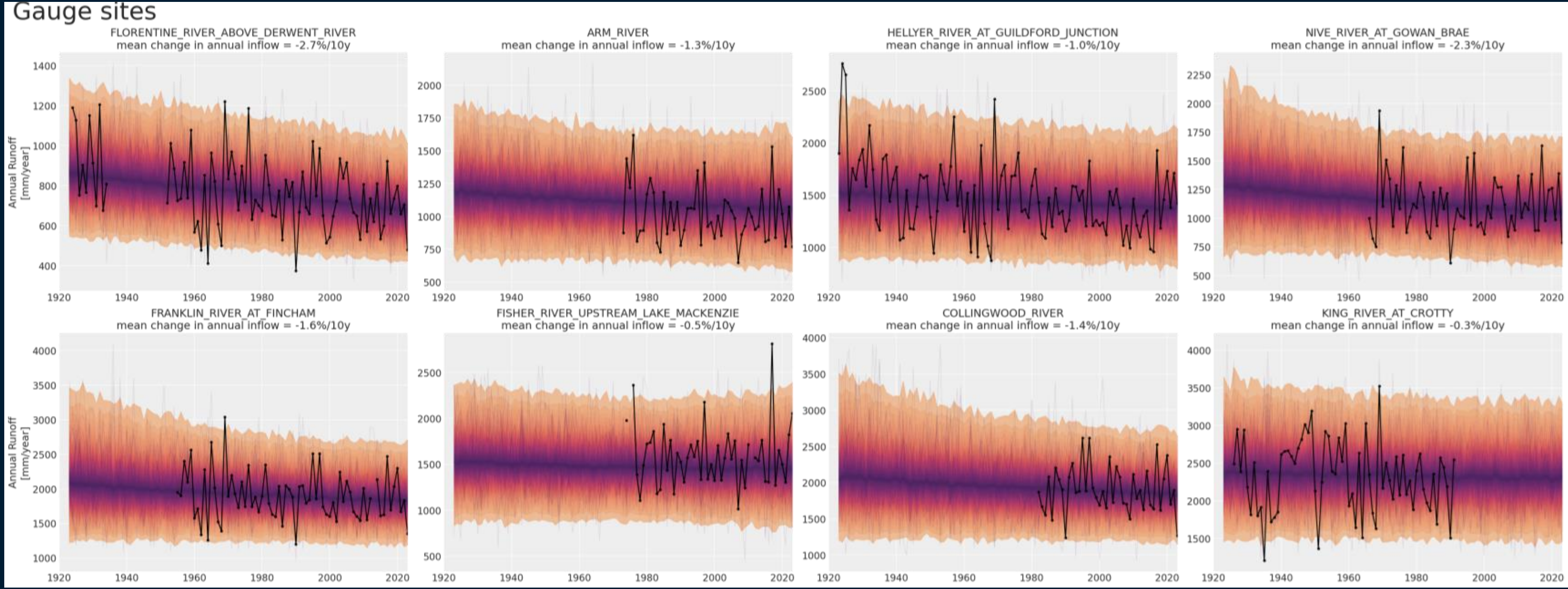
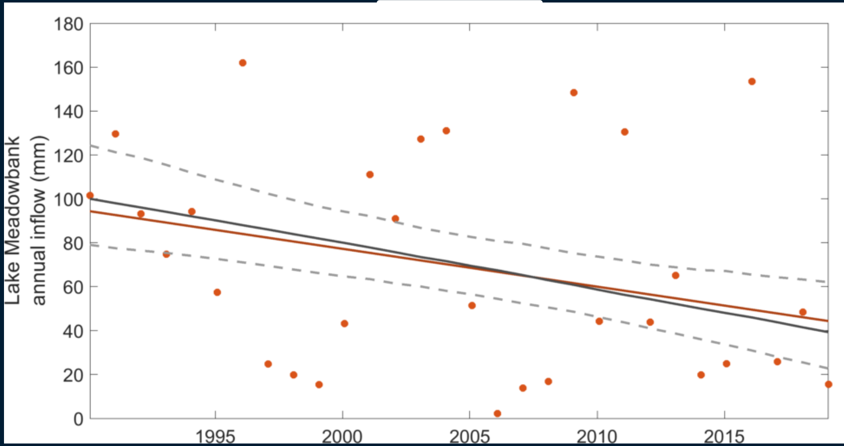


Fig courtesy David Horsley

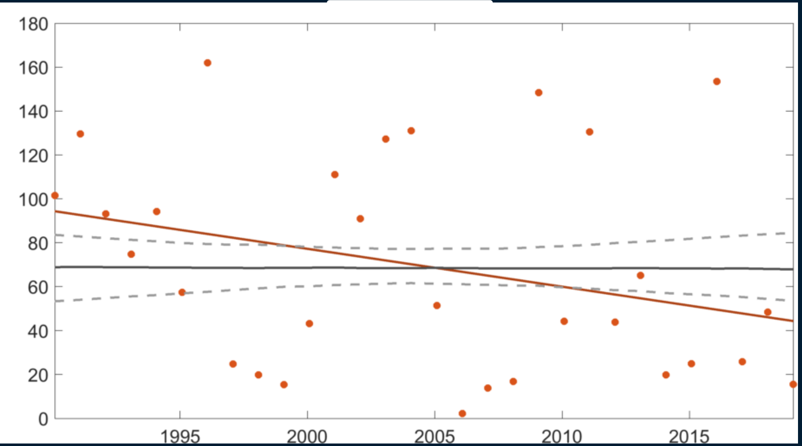
CSIRO

PIT example: the TULIP model

TULIP



Stationary climatology



- Obs
- Obs trend
- 50% CI
- TULIP trend

Meadowbank Powerstation

Summary

- PIT uniformity a formal test of reliability
- Requires fewer data points than rank histograms
- Summary statistics are available
- Diagnose issues with spatial & temporal correlations
- Diagnose problems with non-stationarity

Thank you

Land & Water
James Bennett
Principal Research Scientist

+61 2 9545 2462
james.bennett@csiro.au
https://people.csiro.au/B/J/James-Bennett

References

Forecast verification

Bellier J, Zin I, Bontron G. 2017. Sample Stratification in Verification of Ensemble Forecasts of Continuous Scalar Variables: Potential Benefits and Pitfalls. Monthly Weather Review 145: 3529-3544. DOI: 10.1175/mwr-d-16-0487.1

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69:** 243-268. DOI: 10.1111/j.1467-9868.2007.00587.x.

Hamill TM. 2001. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review* **129:** 550-560. DOI: 10.1175/1520-0493(2001)129<0550:Iorhfv>2.0.Co;2.

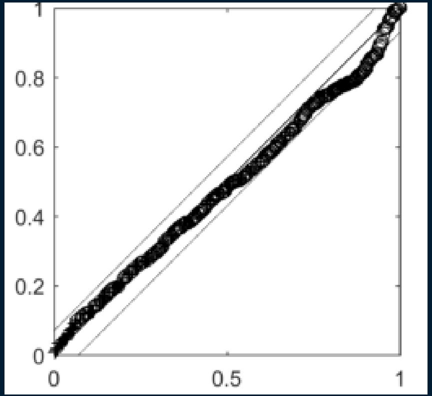
Shrestha DL, Robertson DE, Bennett JC, Wang QJ. 2020. Using the Schaake shuffle when calibrating ensemble means can be problematic. Journal of Hydrology 587: 124991. DOI: 10.1016/j.jhydrol.2020.124991.

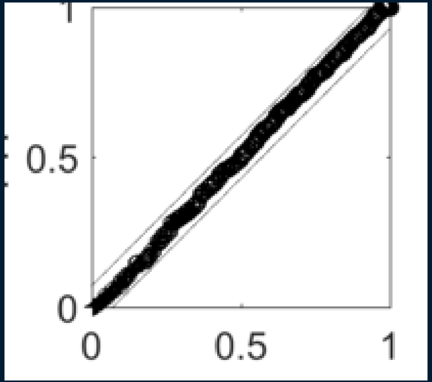
The TULIP model

- Monthly model
- Reliability of 1-year accumulated inflow
- Autocorrelation?

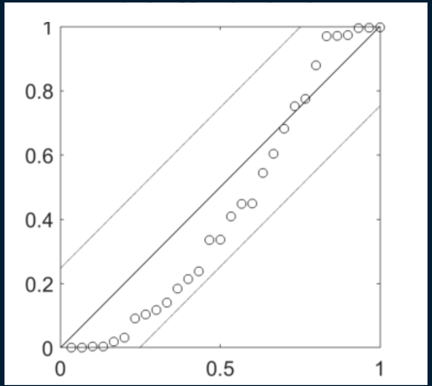
TULIP | **Old method**

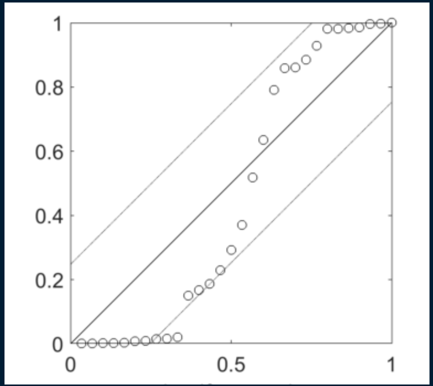
Monthly Inflow





Annual Inflow

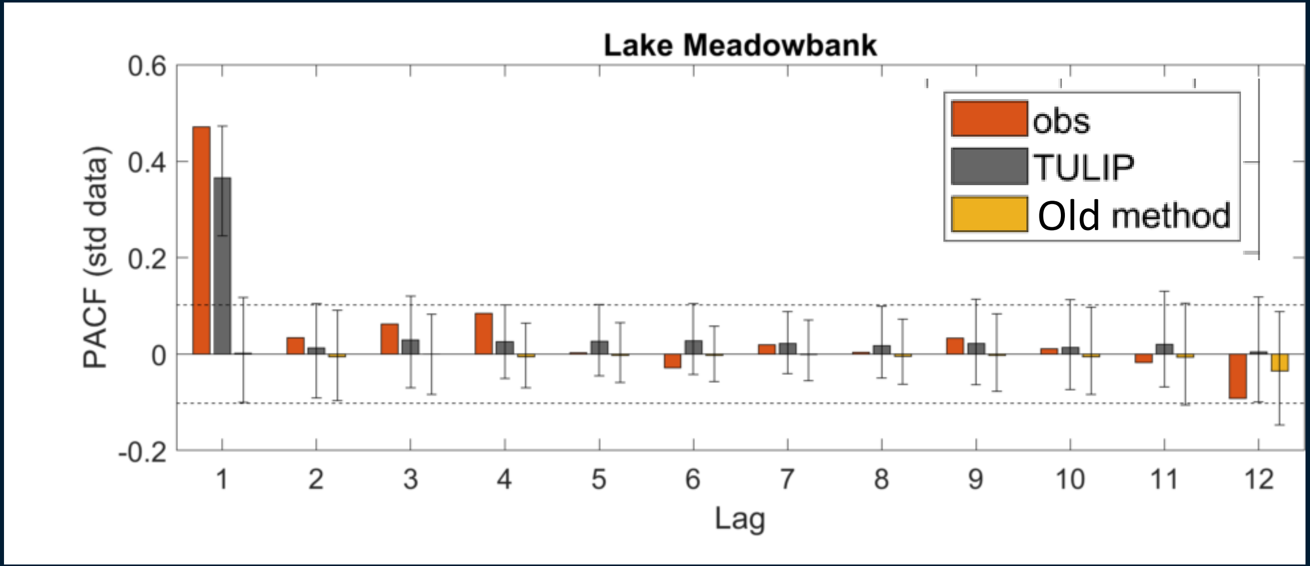




Std uniform variate

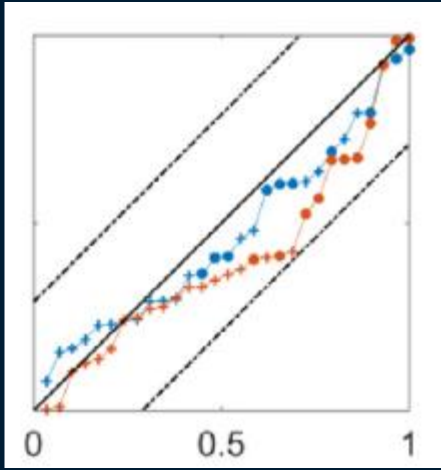
PIT example: the TULIP model

- Monthly model
- Reliability of 1-year accumulated inflow
- Autocorrelation?



Probability integral transforms with zeros

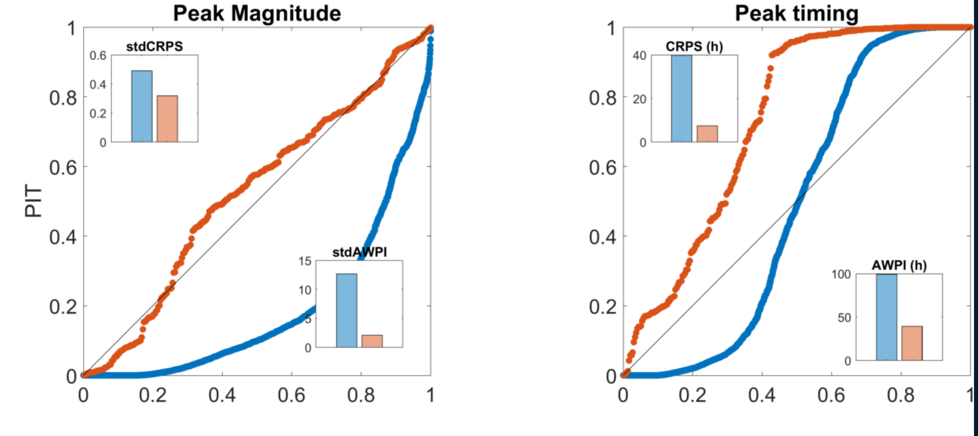
$$p(t) = \begin{cases} F(t, q_o(t)) & q_o(t) > 0 \\ U(0,1) \times F(0) & q_o(t) = 0 \end{cases}$$



PIT

Standard Uniform Variate

Reliability of flood peak magnitude and timing



- Must condition any stratification on forecasts (Bellier et al. 2017)