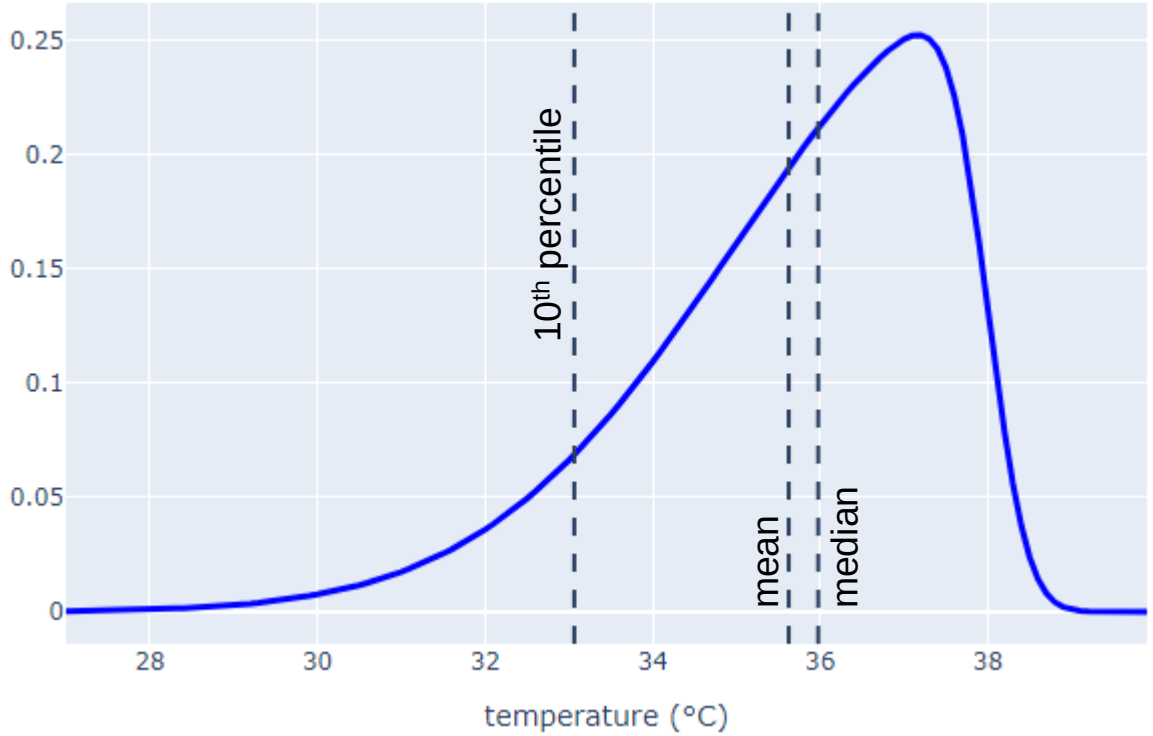
Evaluation of predictive performance with emphasis on extremes using proper scoring rules

Rob Taggart

Australian Bureau of Meteorology

What is this talk about?

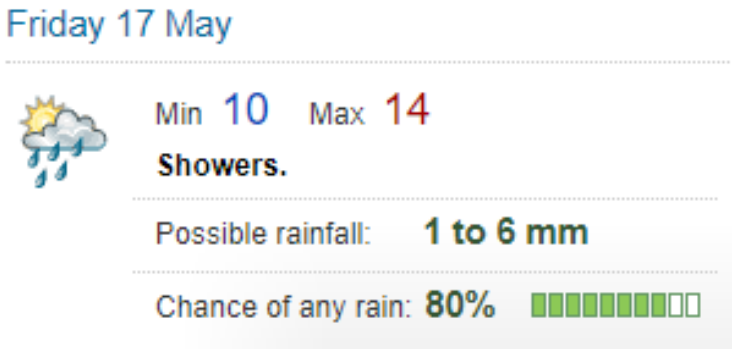
- Rigorous assessment of predictive performance with emphasis on extremes
- Focusses on single-valued forecasts that are defined statistically (e.g. mean, median, 10th percentile of a predictive distribution).



10th percentile:	33.1°C
mean:	35.6°C
median:	36.0°C

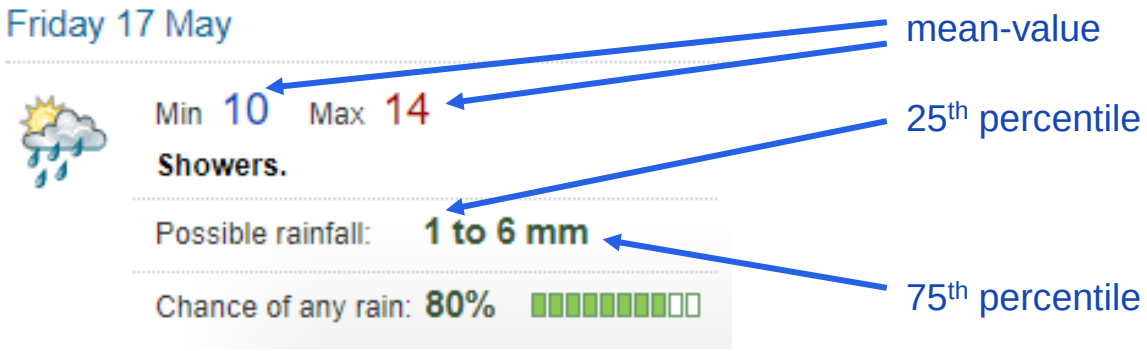
What is this talk about?

- Rigorous assessment of predictive performance with emphasis on extremes
- Focusses on single-valued forecasts that are defined statistically (e.g. mean, median, 10th percentile of a predictive distribution).



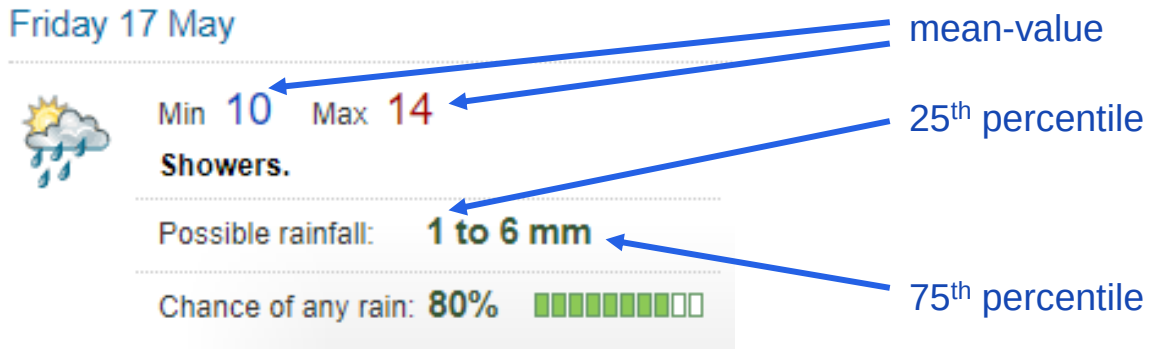
What is this talk about?

- Rigorous assessment of predictive performance with emphasis on extremes
- Focusses on single-valued forecasts that are defined statistically (e.g. mean, median, 10th percentile of a predictive distribution).



What is this talk about?

- Rigorous assessment of predictive performance with emphasis on extremes
- Focusses on single-valued forecasts that are defined statistically (e.g. mean, median, 10th percentile of a predictive distribution).



- To do assess forecasts rigorously, we want a **consistent scoring function**.

  e.g. A consistent scoring function for median forecasts provides an incentive for forecasters to issue their median-value forecast, and not some other forecast.

Consistent scoring functions are to single-valued forecasts what proper scoring rules are to fully probabilistic forecasts.

Outline

1. Motivation for this work
2. A way of thinking about the problem
3. Threshold-weighted absolute loss (for median-value forecasts)
4. Threshold-weighted squared loss (for mean-value forecasts)
5. Other threshold-weighted scores
6. Summary

This presentation is based on:

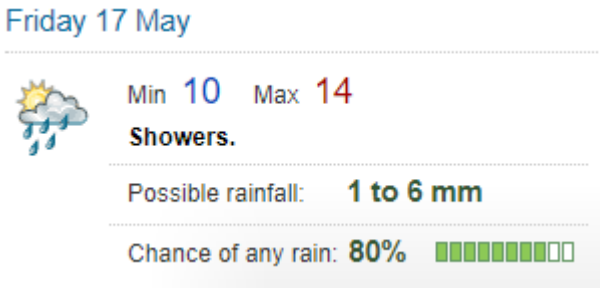
> Taggart, R., 2022. Evaluation of point forecasts for extreme events using consistent scoring functions. *Quart. J. Royal Meteorol. Soc.*, 148, 306-320.

This stands on the shoulders of giants:

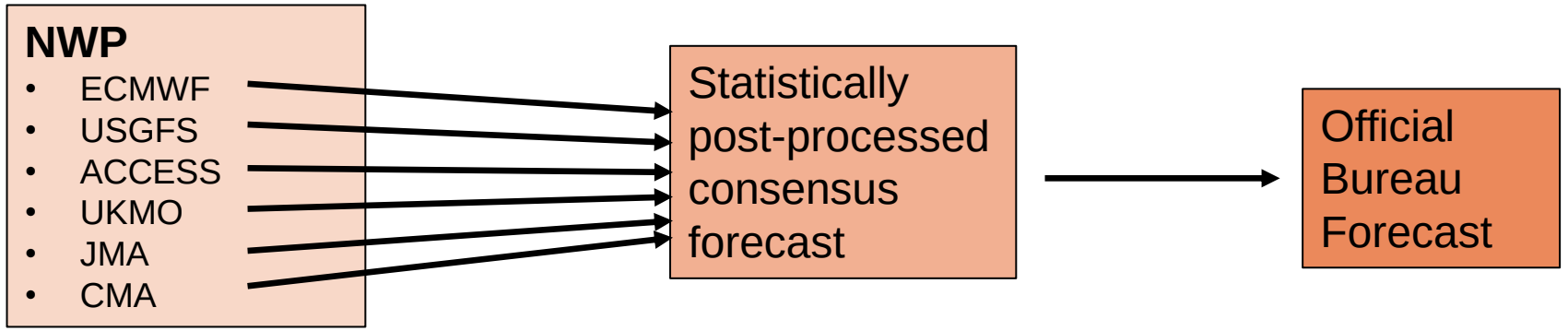
> Ehm, W., Gneiting, T., Jordan, A. and Krüger, F., 2016. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(3), pp.505-562.

Motivation

Bureau forecast process for public weather forecasts

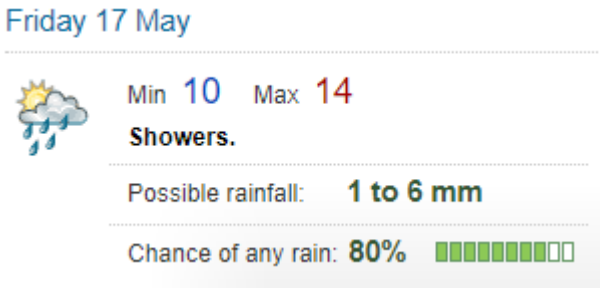


A potential automated process

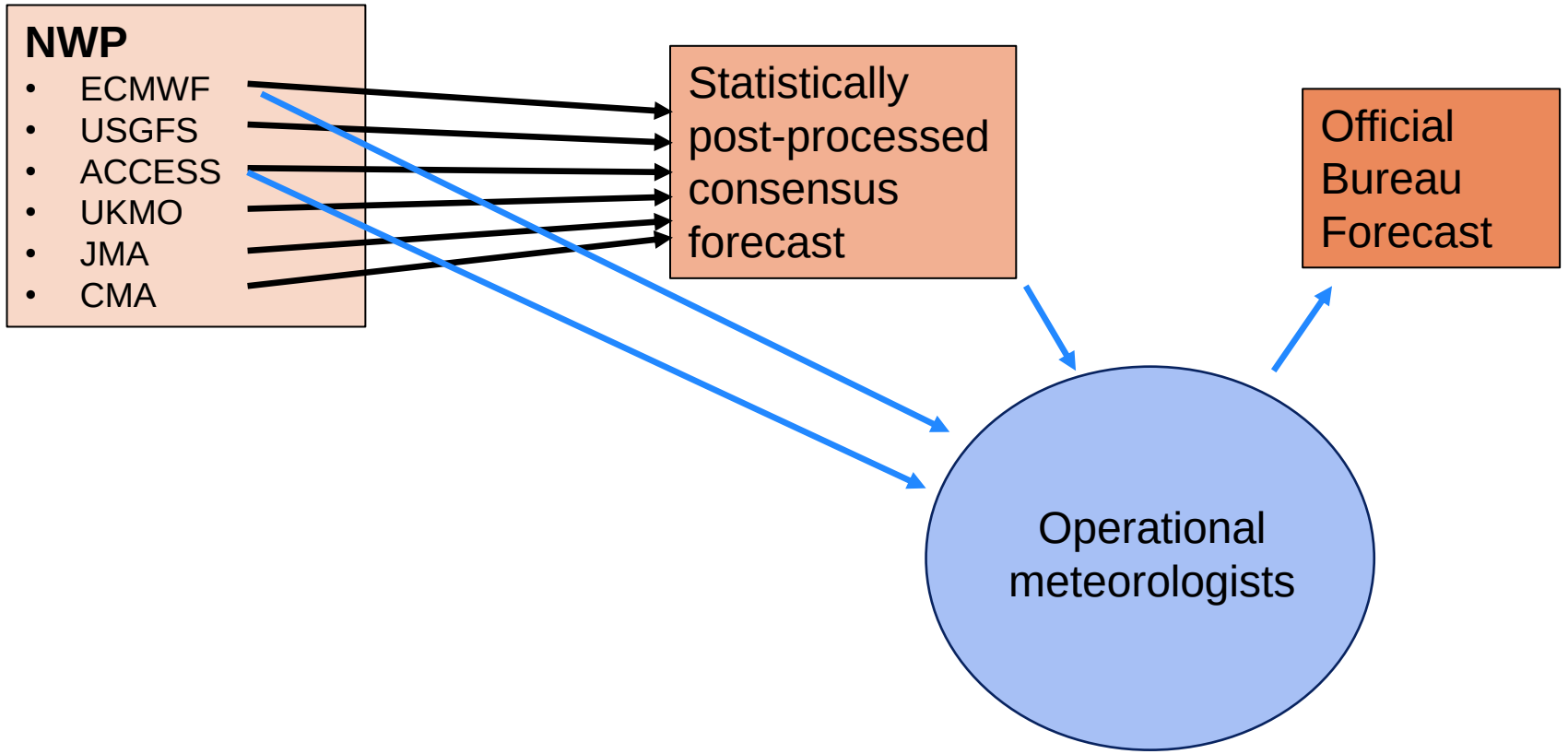


Motivation

Bureau forecast process for public weather forecasts

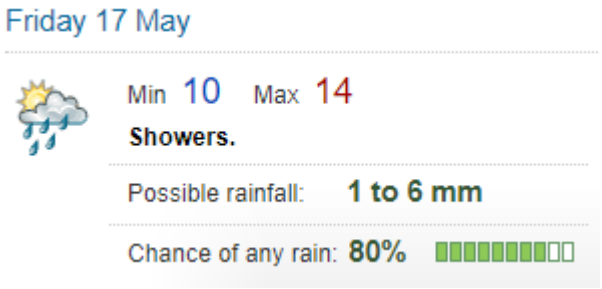


Actual process about 5 years ago

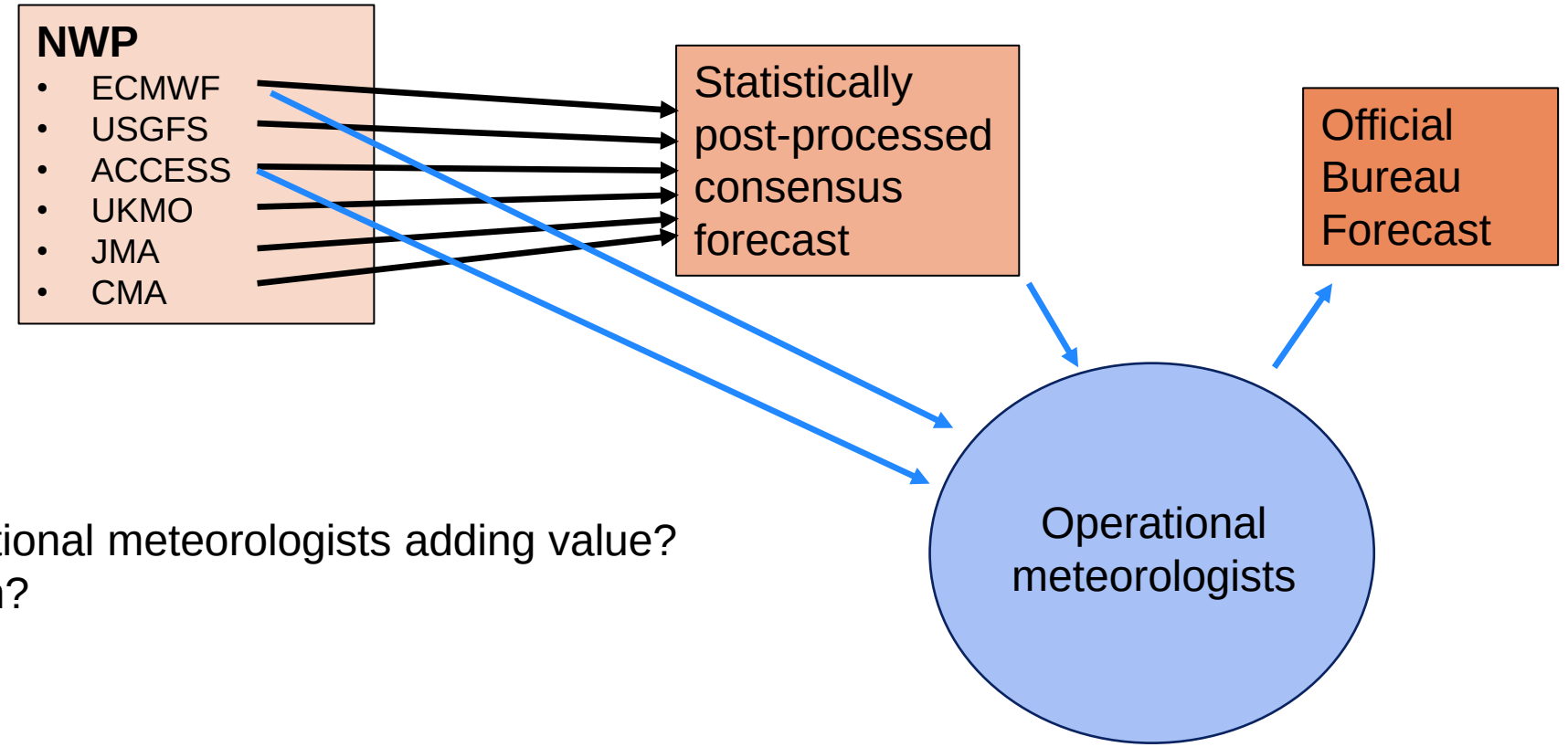


Motivation

Bureau forecast process for public weather forecasts



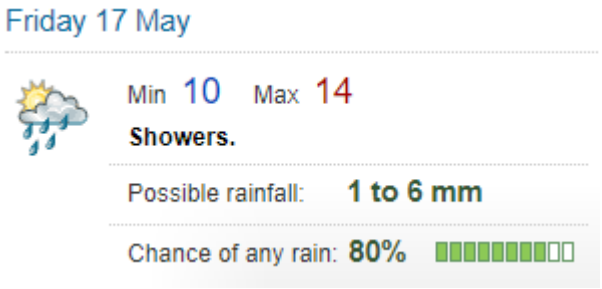
Actual process about 5 years ago



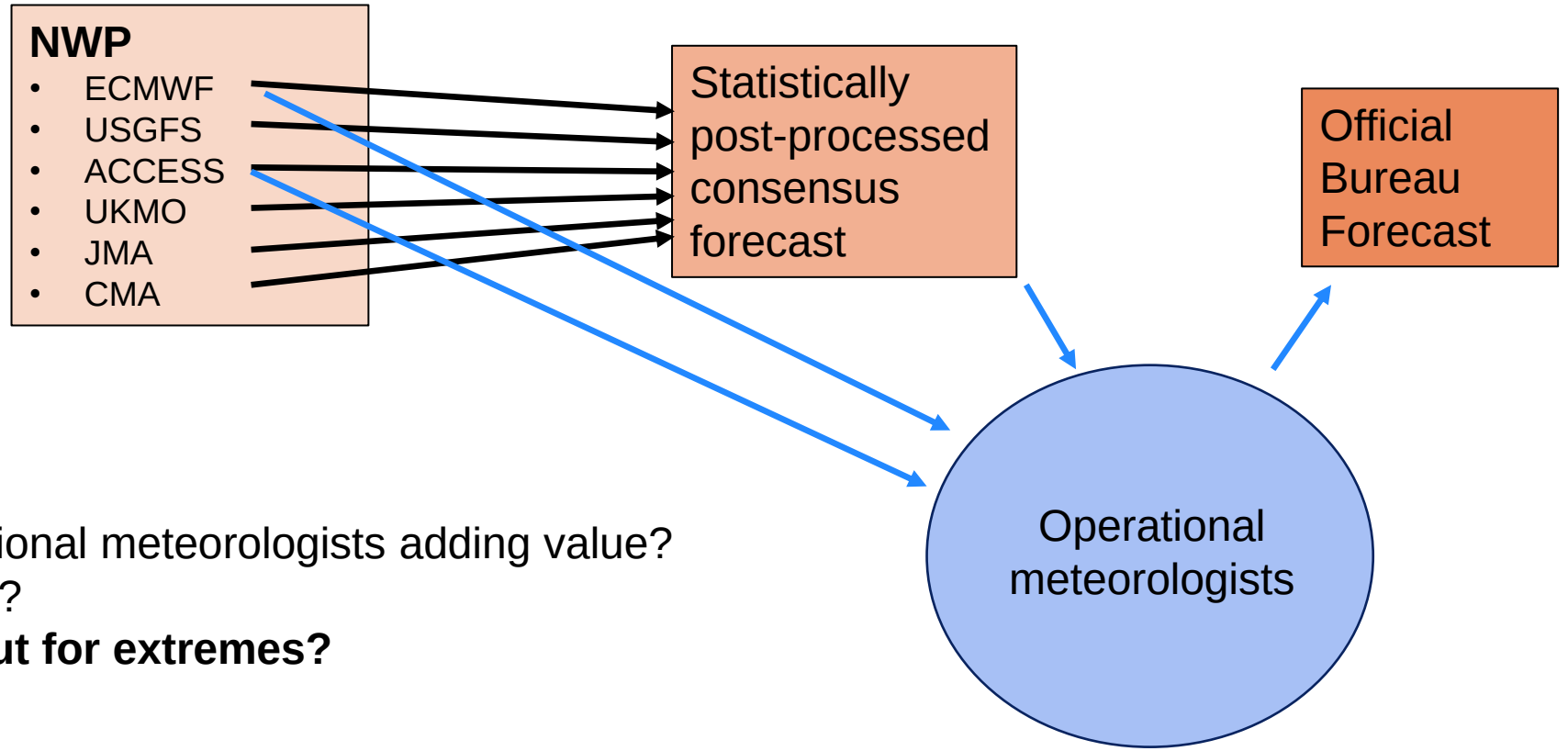
Are operational meteorologists adding value?
How much?

Motivation

Bureau forecast process for public weather forecasts



Actual process about 5 years ago

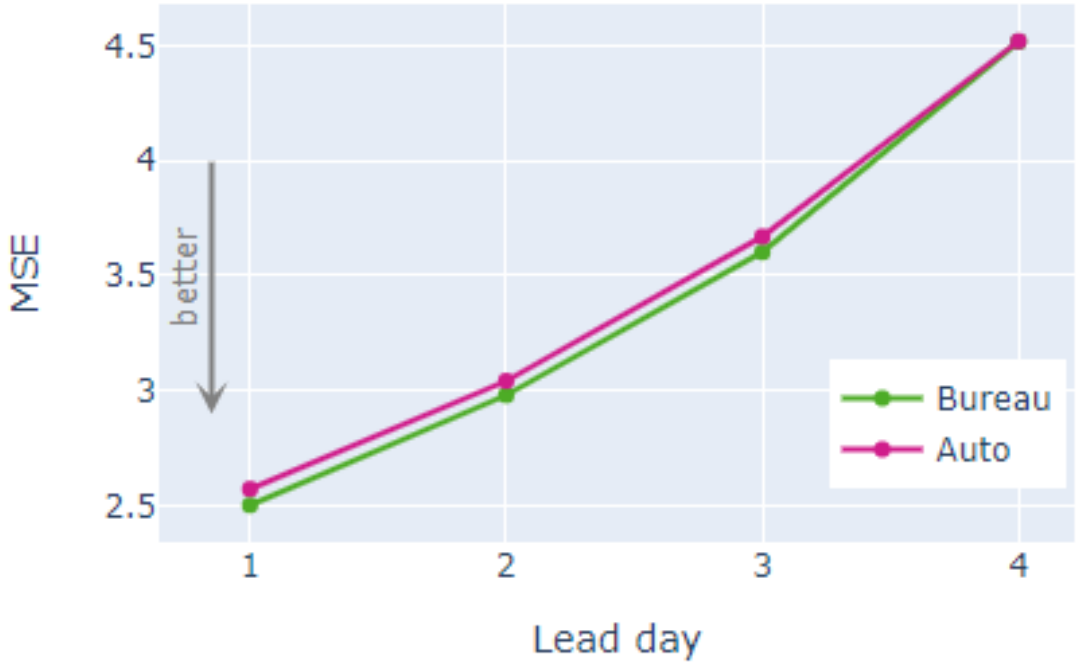


Are operational meteorologists adding value?
How much?
What about for extremes?

Black Summer (2019-20) Daily Maximum Temperature

Looking at bulk verification statistics using MSE, Bureau forecasts did slightly better than Auto.

Is this small improvement sufficient to justify hundreds of hours of manual forecasting?



Are meteorologists adding value for extremes?

How does one evaluate predictive performance for extremes in a rigorous way?

Suppose daily precipitation $\geq$ 50 mm is considered extreme.

Suppose we are assessing median-value forecasts.

A **consistent scoring function** S for median-value forecasts is absolute loss:

$$S(\mathrm{fcst}, \mathrm{obs}) = |\mathrm{fcst} - \mathrm{obs}|$$

How does one evaluate predictive performance for extremes in a rigorous way?

Suppose daily precipitation ≥ 50 mm is considered extreme.

Suppose we are assessing median-value forecasts.

A **consistent scoring function** S for median-value forecasts is absolute loss:

$$S(\mathrm{fcst}, \mathrm{obs}) = |\mathrm{fcst} - \mathrm{obs}|$$

Approach 1 for extremes: evaluate forecasts using S only when $\mathrm{obs} \geq 50$.

How does one evaluate predictive performance for extremes in a rigorous way?

Suppose daily precipitation ≥ 50 mm is considered extreme.

Suppose we are assessing median-value forecasts.

A **consistent scoring function** S for median-value forecasts is absolute loss:

$$S(\text{fcst}, \text{obs}) = |\text{fcst} - \text{obs}|$$

Approach 1 for extremes: evaluate forecasts using S only when $\text{obs} \geq 50$.

Mr Wormwood will never forecast anything less than 50mm.

How does one evaluate predictive performance for extremes in a rigorous way?

Suppose daily precipitation ≥ 50 mm is considered extreme.

Suppose we are assessing median-value forecasts.

A **consistent scoring function** S for median-value forecasts is absolute loss:

$$S(\text{fcst}, \text{obs}) = |\text{fcst} - \text{obs}|$$

Approach 1 for extremes: evaluate forecasts using S only when $\text{obs} \geq 50$.

Mr Wormwood will never forecast anything less than 50mm.

Approach 2 for extremes: evaluate forecasts using S only when $\text{fcst} \geq 50$.

How does one evaluate predictive performance for extremes in a rigorous way?

Suppose daily precipitation ≥ 50 mm is considered extreme.

Suppose we are assessing median-value forecasts.

A **consistent scoring function** S for median-value forecasts is absolute loss:

$$S(\text{fcst}, \text{obs}) = |\text{fcst} - \text{obs}|$$

Approach 1 for extremes: evaluate forecasts using S only when $\text{obs} \geq 50$.

Mr Wormwood will never forecast anything less than 50mm.

Approach 2 for extremes: evaluate forecasts using S only when $\text{fcst} \geq 50$.

Mr Wormwood will forecast less than 50mm whenever he wants an easy day.

How does one evaluate predictive performance for extremes in a rigorous way?

Conditioning on extreme observations or forecasts incentivise poor behaviour!!!

Suppose daily precipitation ≥ 50 mm is considered extreme.

Suppose we are assessing median-value forecasts.

A **consistent scoring function** S for median-value forecasts is absolute loss:

$$S(\text{fcst}, \text{obs}) = |\text{fcst} - \text{obs}|$$

Approach 1 for extremes: evaluate forecasts using S only when $\text{obs} \geq 50$.

Mr Wormwood will never forecast anything less than 50mm.

Approach 2 for extremes: evaluate forecasts using S only when $\text{fcst} \geq 50$.

Mr Wormwood will forecast less than 50mm whenever he wants an easy day.

How does one evaluate predictive performance for extremes in a rigorous way?

Suppose daily precipitation ≥ 50 mm is considered extreme.

Suppose we are assessing median-value forecasts.

A **consistent scoring function** S for median-value forecasts is absolute loss:

$$S(\text{fcst}, \text{obs}) = |\text{fcst} - \text{obs}|$$

Conditioning on extreme observations or forecasts incentivise poor behaviour!!!

Approach 1 for extremes: evaluate forecasts using S only when $\text{obs} \geq 50$.

Mr Wormwood will never forecast anything less than 50mm.

Approach 2 for extremes: evaluate forecasts using S only when $\text{fcst} \geq 50$.

Mr Wormwood will forecast less than 50mm whenever he wants an easy day.

Instead, condition on decision thresholds that lie in the extremes

Simple cost-loss decision model

Harry's decision threshold θ for daily precipitation is 50 mm.
If Harry takes protective action but the obs $\leq \theta$, it costs Harry an additional $1000.
If Harry does not take protective action but obs $> \theta$, it costs Harry an additional $1000.

Simple cost-loss decision model

Harry's decision threshold θ for daily precipitation is 50 mm.
If Harry takes protective action but the obs $\leq \theta$, it costs Harry an additional $1000.
If Harry does not take protective action but obs $> \theta$, it costs Harry an additional $1000.

The scoring function S_θ in this situation is

$$S_\theta(\text{fcst}, \text{obs}) = \begin{cases} 1 & if\ \text{obs} \leq \theta < \text{fcst} \\ 1 & if\ \text{fcst} \leq \theta < \text{obs} \\ 0 & \text{otherwise.} \end{cases}$$

An optimal forecast for this model is a **median-value forecast.**
i.e., S_θ is a consistent scoring function for median-value forecasts.

Simple cost loss decision model

But what about other decision thresholds θ?

Harry: $\theta = 50$ mm

Ioanna: $\theta = 75$ mm

Michael: $\theta = 5$ mm

Ben: $\theta = 0.2$ mm

Simple cost loss decision model

But what about other decision thresholds θ?

Harry: $\theta = 50$ mm
extreme

Ioanna: $\theta = 75$ mm
extreme

Michael: $\theta = 5$ mm
not extreme

Ben: $\theta = 0.2$ mm
not extreme

Integrating over all decision thresholds in the extreme range

Suppose that 50 mm or higher is considered extreme.

We combine all scoring functions S_θ, where $\theta \geq 50$, using integration:

$$S_{\theta \geq 50}(\text{fcst}, \text{obs}) = \int_{50}^{\infty} S_\theta(\text{fcst}, \text{obs})\, \mathrm{d}\theta$$

$$= |\mathbf{1}\{\text{fcst} \geq 50\}(\text{fcst} - 50) - \mathbf{1}\{\text{obs} \geq 50\}(\text{obs} - 50)|$$

Integrating over all decision thresholds in the extreme range

Suppose that 50 mm or higher is considered extreme.
We combine all scoring functions S_θ, where $\theta \geq 50$, using integration:

$$S_{\theta \geq 50}(\text{fcst}, \text{obs}) = \int_{50}^{\infty} S_\theta(\text{fcst}, \text{obs})\, \mathrm{d}\theta$$

$$= |\mathbf{1}\{\text{fcst} \geq 50\}(\text{fcst} - 50) - \mathbf{1}\{\text{obs} \geq 50\}(\text{obs} - 50)|$$

This is a **consistent scoring function** for median-value forecasts.

Integrating over all decision thresholds in the extreme range

Suppose that 50 mm or higher is considered extreme.
We combine all scoring functions S_θ, where $\theta \geq 50$, using integration:

$$S_{\theta \geq 50}(\text{fcst}, \text{obs}) = \int_{50}^{\infty} S_\theta(\text{fcst}, \text{obs})\, \mathrm{d}\theta$$

$$= |\mathbf{1}\{\text{fcst} \geq 50\}(\text{fcst} - 50) - \mathbf{1}\{\text{obs} \geq 50\}(\text{obs} - 50)|$$

This is a **consistent scoring function** for median-value forecasts.

Case	Score
fcst and obs both < 50	0
fcst and obs both ≥ 50	|fcst – obs|
fcst ≥ 50, obs < 50	|fcst – 50|
obs ≥ 50, fcst < 50	|obs – 50|

Threshold-weighted absolute loss

The scoring function $S_{\theta \geq 50}$ is an example of **threshold-weighted absolute loss**.

This name is appropriate because if we integrated over all θ, we obtain **absolute loss**:

$$S_{\text{all } \theta}(\text{fcst}, \text{obs}) = \int_{-\infty}^{\infty} S_{\theta}(\text{fcst}, \text{obs}) \, \mathrm{d}\theta$$

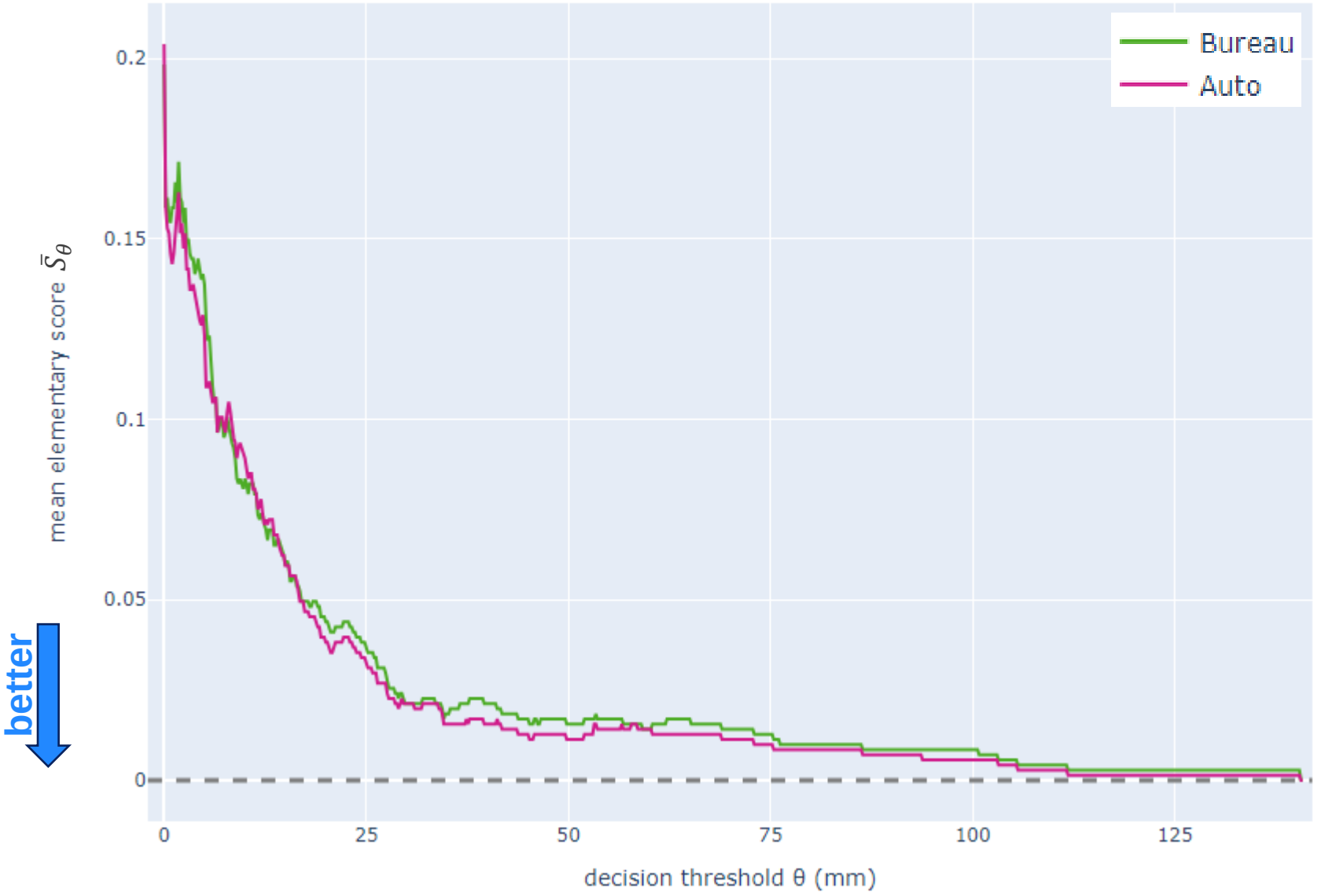
$$= |\text{fcst} - \text{obs}|$$

This gives **decompositions of MAE**, e.g.

$$\text{MAE} = \bar{S}_{\text{all } \theta} = \bar{S}_{\theta < 50} + \bar{S}_{\theta \geq 50}$$

where all terms in the decomposition are *consistent* for median-valued forecasts.

Murphy diagram



The **Murphy diagram** is a plot of mean elementary score $\bar{S}_\theta$ against decision threshold θ.

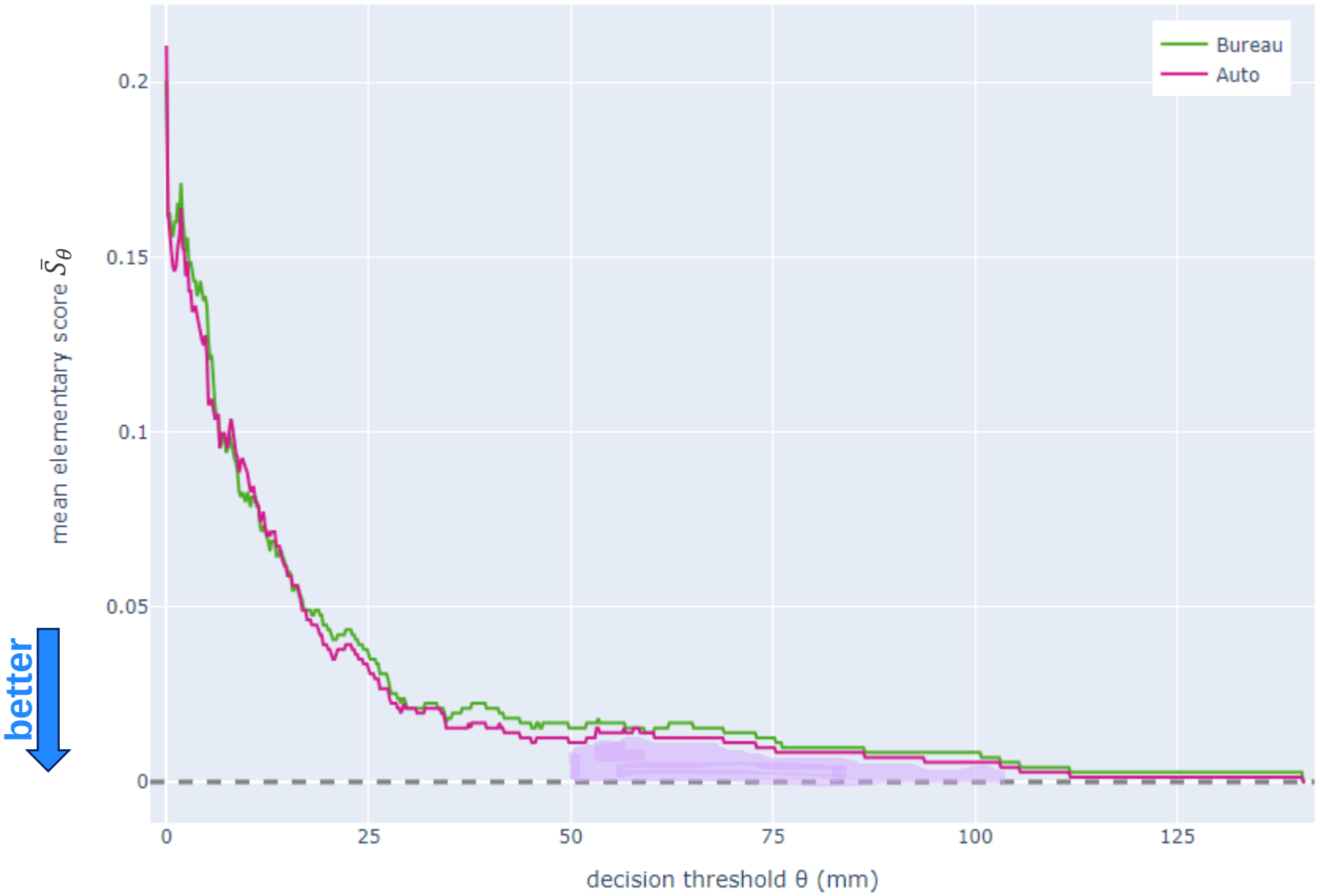
Ehm et. al. (2016), *Of Quantiles and Expectiles*

Median daily rainfall forecasts for Sydney Airport 2021-2022.

Bureau is the official forecast, curated by meteorologists.

Auto is a statistically post-processed blend of NWP.

Murphy diagram

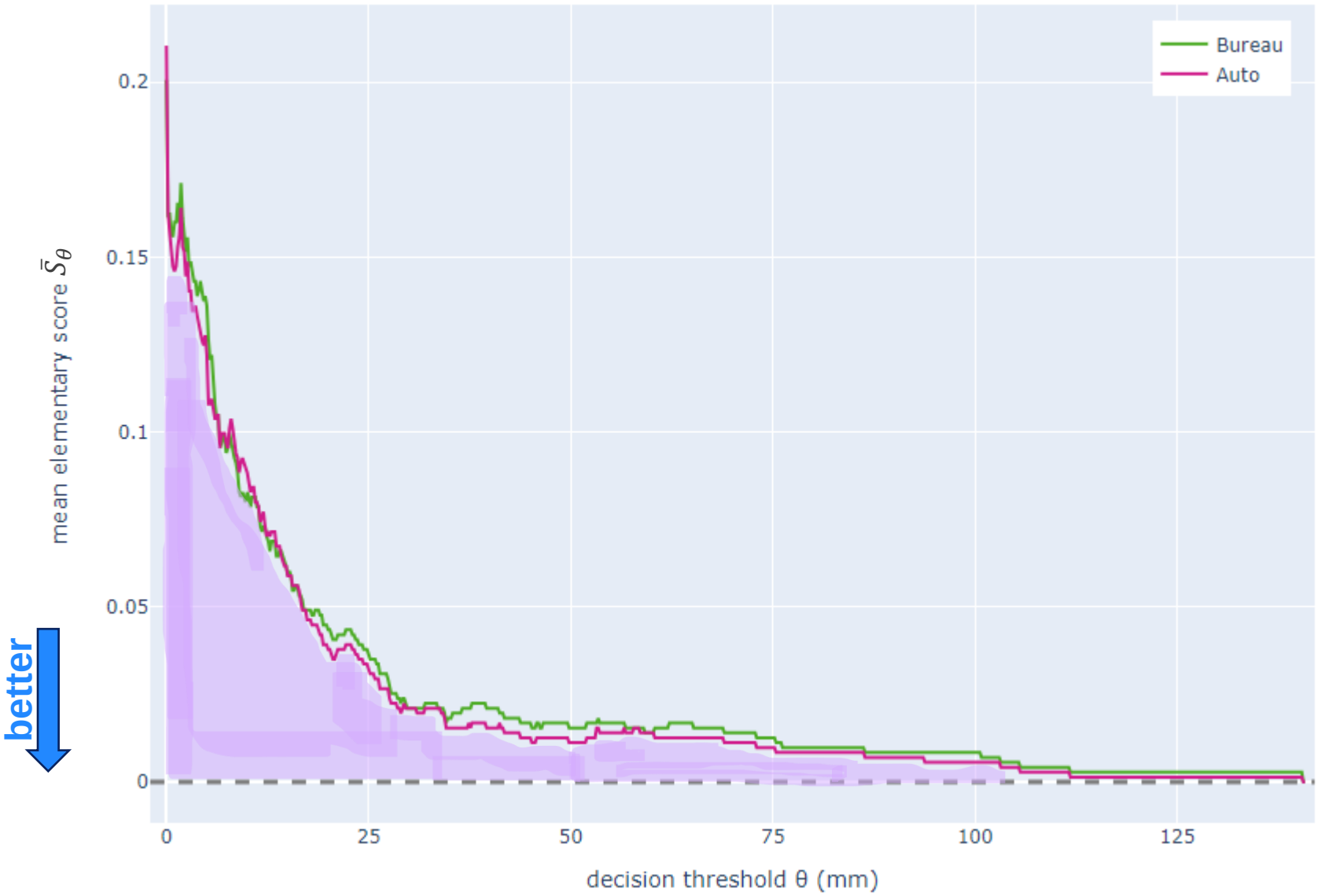


Shaded area under the Auto curve is proportional the mean threshold-weighted score $\bar{S}_{\theta \geq 50}$ for Auto.

A lower area is better.

Forecast	$\bar{S}_{\theta \geq 50}$
Bureau	0.764
Auto	0.588

Murphy diagram



Shaded area under the Auto curve is proportional the mean absolute error (MAE) $\bar{S}_{\text{all }\theta}$ for Auto.

A lower area is better.

Fcst System	MAE
Bureau	3.398
Auto	3.067

Other threshold-weighted scores

Threshold weighted squared loss for mean-valued forecasts.

e.g. For scoring with emphasis on extreme heat ($\theta \geq 40$°C):

$$S_{\theta \geq 40}(\text{fcst}, \text{obs}) = (\text{obs} - 40)^2 \, \mathbf{1}\{\text{obs} \geq 40\} - (\text{fcst} - 40)^2 \, \mathbf{1}\{\text{fcst} \geq 40\} - 2(\text{obs} - \text{fcst})(\text{fcst} - 40) \, \mathbf{1}\{\text{fcst} \geq 40\}$$

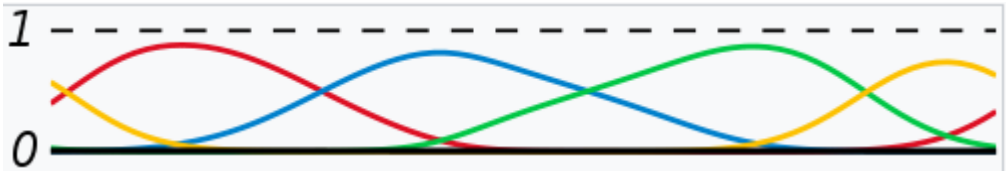
Threshold-weighted scores for percentile (quantile) forecasts

Applications to multi-categorical forecasts (FIRM score: Taggart, Loveday & Griffiths 2022)

Murphy diagrams for each of these.

Threshold-weighting functions don't have to be 0-1:

Smooth positive weights give *strictly* consistent scores and have better smoothness properties for ML applications



Example: Black Summer (2019-20) Tmax forecasts

Threshold is the 97th percentile of annual climatology over 20-year period

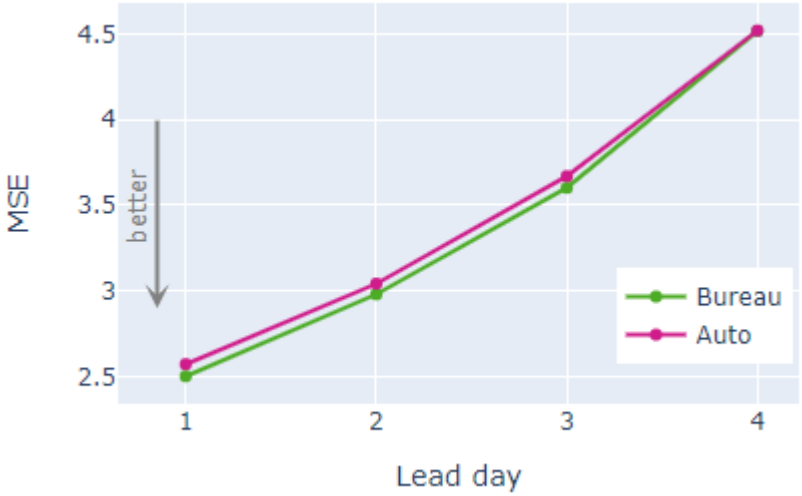
- Adelaide City: 37.4°C
- Melbourne City: 34.5°C

Results for lead days 1-3:

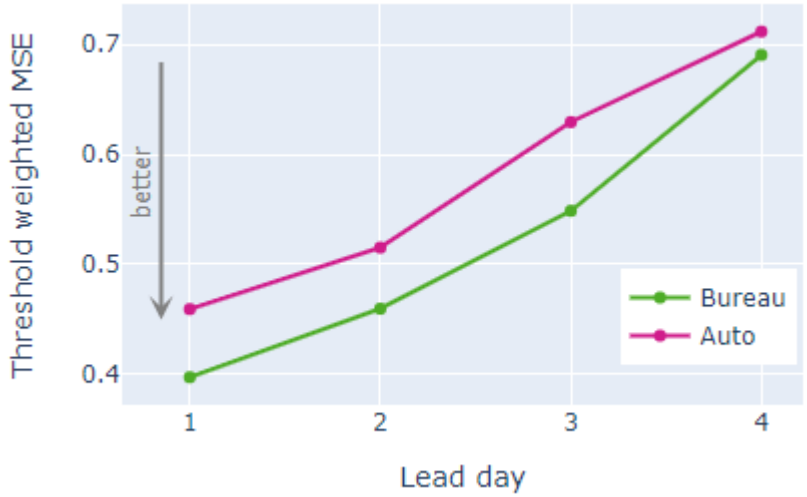
- MSE: BoM slightly better (by about 0.1 lead days of skill)
- Threshold-weighted MSE: BoM much better (by up to 1 lead day of skill)

Threshold-weighted MSE, MAE and quantile scores (check loss) to appear in the **scores python package**

Overall performance



Performance for extremes (hot days)



Summary

- Threshold-weighted scores provide a rigorous way of assessing performance with emphasis on extremes
- They are available for mean-value and percentile (including median-value) forecasts, etc
- Commonly used scores (MSE, MAE) can be decomposed into sums of threshold-weighted scores
- Murphy diagrams are an important accompanying visual tool

References

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F., 2016. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(3), pp.505-562.

Taggart, R., 2022. Evaluation of point forecasts for extreme events using consistent scoring functions. *Quart. J. Royal Meteorol. Soc.*, 148, 306-320.