The Bureau
of Meteorology

The FIxed Risk Multicategorical (FIRM) Scoring Framework

Nicholas Loveday, Robert Taggart and Deryn Griffiths

Example of an ordered multicategorical warning

The Bureau has a categorical heatwave warning service

Lead day	0	1	2	3	4	5
Category	Severe	Extreme	Severe	Severe	No warning	No warning

Three categories

1. Extreme: $3 \leq \text{Heat Index} < \infty$
2. Severe: $1 \leq \text{Heat Index} < 3$
3. No warning: $-\infty < \text{Heat Index} < 1$

Example of an ordered multicategorical warning

The Bureau has a categorical heatwave warning service

Lead day	0	1	2	3	4	5
Category	Severe	Extreme	Severe	Severe	No warning	No warning

Three categories

1. Extreme: 3 ≤ Heat Index < ∞
2. Severe: 1 ≤ Heat Index < 3
3. No warning: -∞ < Heat Index < 1

Forecast directive:
"Forecast the highest category for which the probability of observing that category or higher exceeds 50%"

Existing multicategorical verification methods

Textbooks/literature recommend *equitable* scores such as the Gerrity score for evaluating multicategorical forecasts.

Equitable score: all constant forecasts and random forecasts receive the same expected score.

However, the warning strategy that optimises equitable scores:

1. Is not tied to a fixed risk.
2. It is related to the risk of observing warning conditions exceeding the sample base rate.

Warning strategy to optimise the expected dichotomous Gerrity score:

Warn if probability > sample base rate

Existing multicategorical verification methods

Textbooks/literature recommend *equitable* scores such as the Gerrity score for evaluating multicategorical forecasts.

Equitable score: all constant forecasts and random forecasts receive the same expected score.

However, the warning strategy that optimises equitable scores:

1. Is not tied to a fixed risk.
2. It is related to the risk of observing warning conditions exceeding the sample base rate.
3. For climatologically rare events, this would lead to a large amount of False Alarms.

Warning strategy to optimise the expected dichotomous Gerrity score:
If sample base rate = 0.01, warn if the probability of the event ≥ 1%

Existing multicategorical verification methods

Textbooks/literature recommend *equitable* scores such as the Gerrity score for evaluating multicategorical forecasts.

Equitable score: all constant forecasts and random forecasts receive the same expected score.

However, the warning strategy that optimises equitable scores:

1. Is not tied to a fixed risk.
2. It is related to the risk of observing warning conditions exceeding the sample base rate.
3. For climatologically rare events, this would lead to a large amount of False Alarms.
4. The more categories, the harder it is to derive the optimal probability to issue a warning on.

Existing multicategorical verification methods

Textbooks/literature recommend *equitable* scores such as the Gerrity score for evaluating multicategorical forecasts.

Equitable score: all constant forecasts and random forecasts receive the same expected score.

However, the warning strategy that optimises equitable scores:

1. Is not tied to a fixed risk.
2. It is related to the risk of observing warning conditions exceeding the sample base rate.
3. For climatologically rare events, this would lead to a large amount of False Alarms.
4. The more categories, the harder it is to derive the optimal probability to issue a warning on.

Existing multicategorical verification methods

Textbooks/literature recommend *equitable* scores such as the Gerrity score for evaluating multicategorical forecasts.

Equitable score: all constant forecasts and random forecasts receive the same expected score.

However, the warning strategy that optimises equitable scores:

1. Is not tied to a fixed risk.
2. It is related to the risk of observing warning conditions exceeding the sample base rate.
3. For climatologically rare events, this would lead to a large amount of False Alarms.
4. The more categories, the harder it is to derive the optimal probability to issue a warning on.

The Gerrity score is not a consistent score for the forecast directive:
"Forecast the highest category for which the probability of observing that category or higher exceeds 50%"

The FIxed Risk Multicategorical (FIRM) Framework

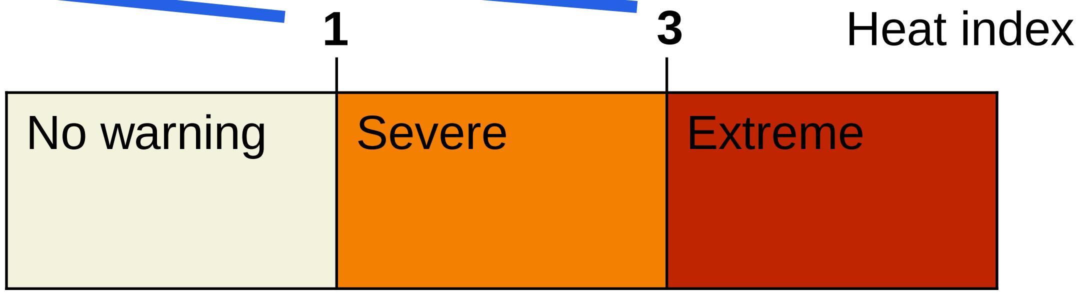
Specify the following:

1. Categorical thresholds
2. Corresponding weights for each threshold
3. Risk parameter (α)

The FIxed Risk Multicategorical (FIRM) Framework

Specify the following:

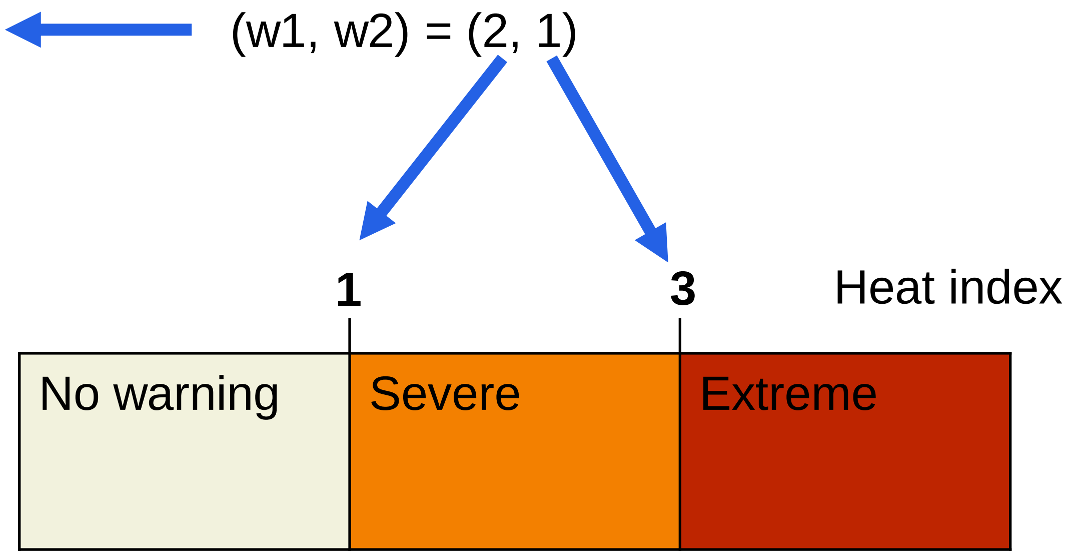
1. **Categorical thresholds**
2. Corresponding weights for each threshold
3. Risk parameter (α)



The FIxed Risk Multicategorical (FIRM) Framework

Specify the following:

1. Categorical thresholds
2. **Corresponding weights for each threshold**
3. Risk parameter (α)



(w1, w2) = (2, 1)

1

3

Heat index

No warning	Severe	Extreme

The FIxed Risk Multicategorical (FIRM) Framework

Specify the following:

1. Categorical thresholds
2. Corresponding weights for each threshold
3. **Risk parameter (α)** ←

Specify the cost of a miss relative to a false alarm. $\frac{\alpha}{1-\alpha}$

This is the equivalent to specifying a fixed threshold probability $1 - \alpha$

Directly related to the cost-loss ratio $\alpha = 1 - \frac{C}{L}$

The FIxed Risk Multicategorical (FIRM) Framework

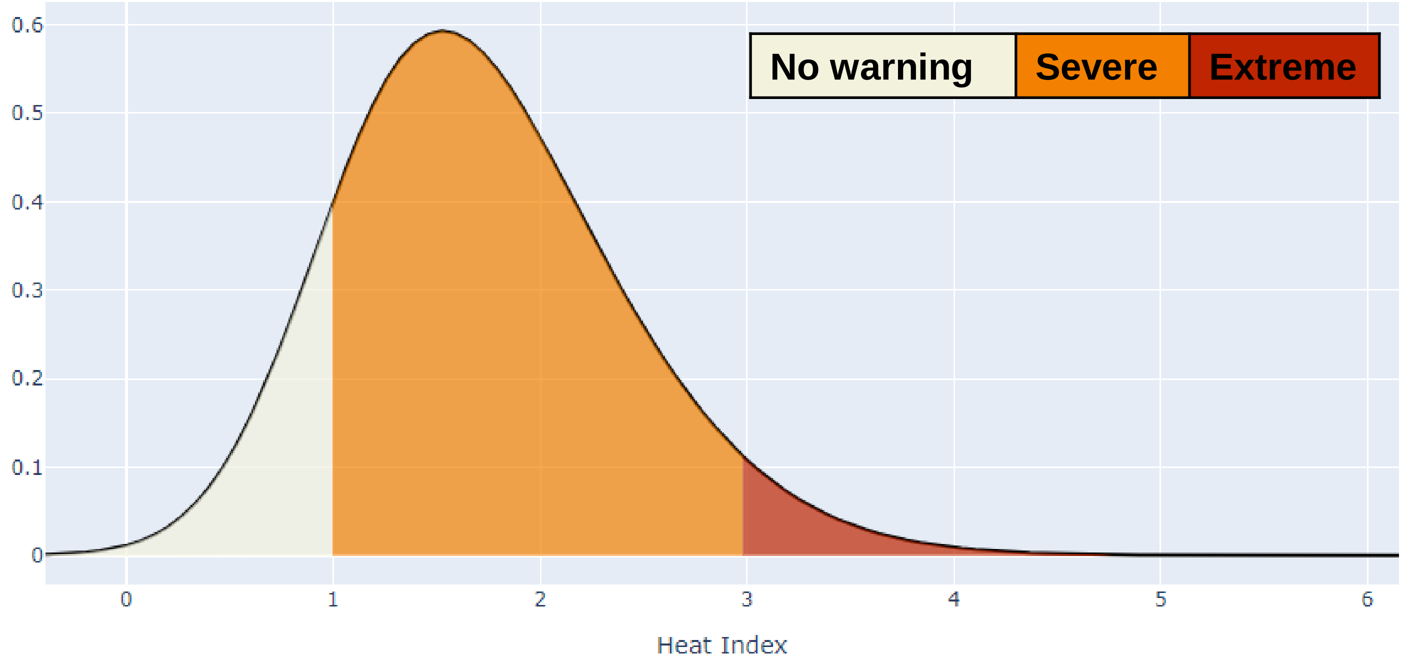
Specify the following:

1. Categorical thresholds
2. Corresponding weights for each threshold
3. **Risk parameter (α)**

Forecast directive:
"Forecast a category which contains an α-quantile of the predictive distribution"

If $\alpha = 0.5$, forecast severe.

If $\alpha = 0.95$, forecast extreme.



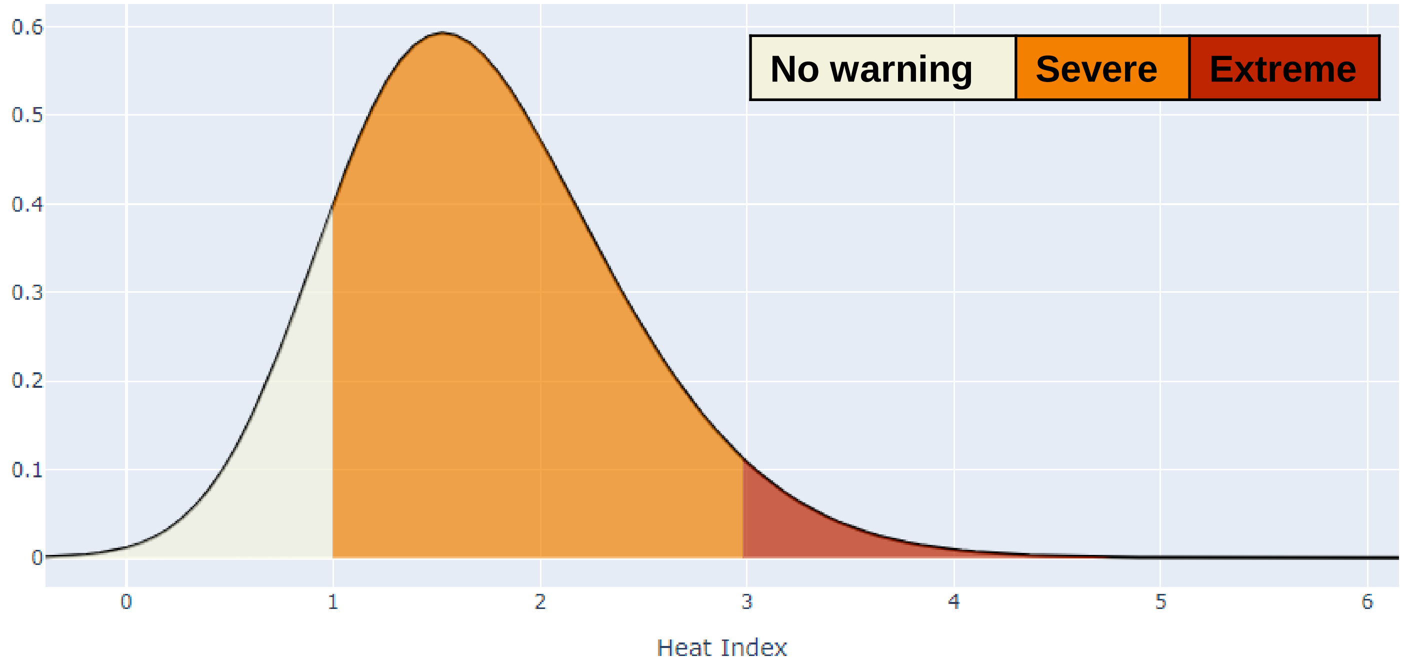
The FIxed Risk Multicategorical (FIRM) Framework

Specify the following:

1. Categorical thresholds
2. Corresponding weights for each threshold
3. **Risk parameter (α)**

Forecast directive:
"Forecast a category which contains an α-quantile of the predictive distribution"

Alternatively
"Forecast the highest category for which the probability of observing that category or higher exceeds $1 - \alpha$"



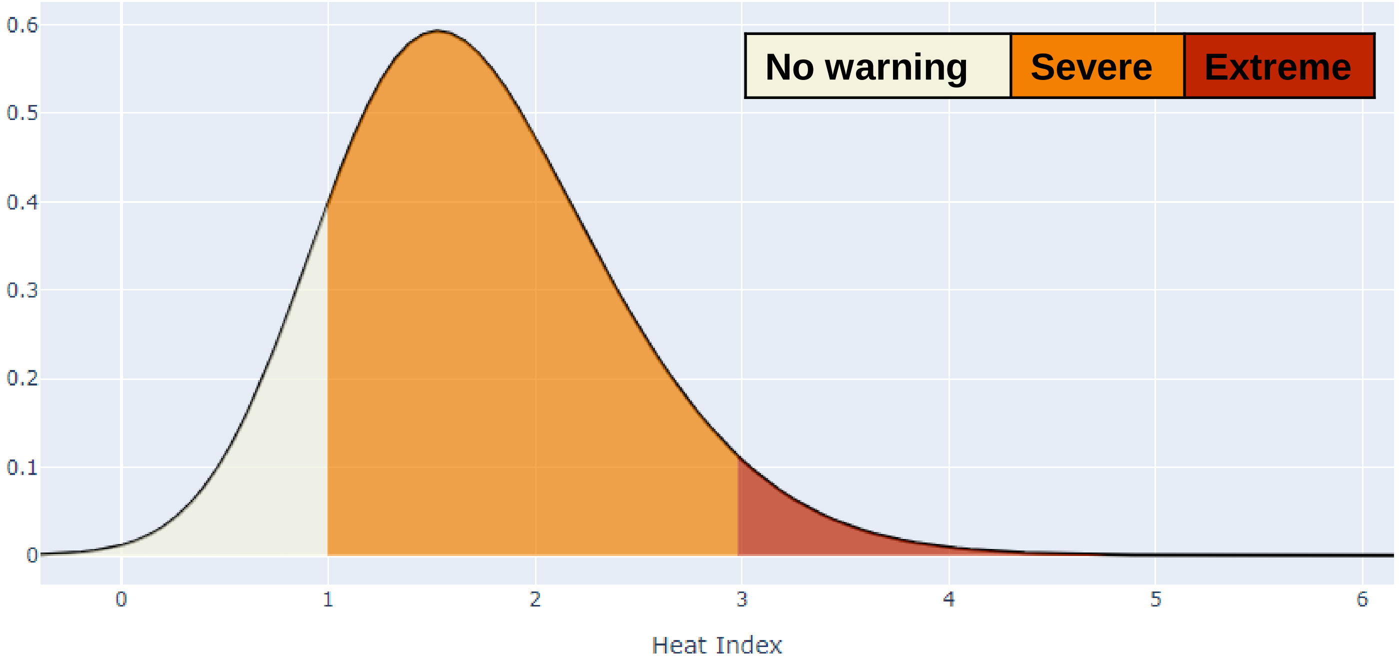
The FIxed Risk Multicategorical (FIRM) Framework

Specify the following:

1. Categorical thresholds
2. Corresponding weights for each threshold
3. **Risk parameter (α)**

Forecast directive:
*"Forecast a category which contains an **0.5** quantile of the predictive distribution"*

Alternatively
*"Forecast the highest category for which the probability of observing that category or higher exceeds **50%"***



The FIxed Risk Multicategorical (FIRM) Framework

Specify the following:

1.	Categorical thresholds	[1, 3]
2.	Corresponding weights for each threshold	[2, 1]
3.	Risk parameter (α)	0.5

The FIxed Risk Multicategorical (FIRM) Framework

Scoring functions

For the two-category case:

$$S^{\mathrm{Q}}_{\theta,\alpha}(x,y) = \begin{cases} 1-\alpha, & y \le \theta < x, \\ \alpha, & x \le \theta < y, \\ 0, & \text{otherwise.} \end{cases}$$

Penalty of False Alarm

Penalty of Miss

θ=decision threshold

For multiple categories:

$$S^{\mathrm{Q}}(x,y) = \sum_{i=1}^{N} w_i \; S^{\mathrm{Q}}_{\theta_i,\alpha}(x,y)$$

Weights

A score closer to 0 is better, similar to Mean Square Error

The FIxed Risk Multicategorical (FIRM) Framework

Scoring matrix

Forecast category

Observed category

	No warning	Severe	Extreme
No warning	0	1	1.5
Severe	1	0	0.5
Extreme	1.5	0.5	0

The FIxed Risk Multicategorical (FIRM) Framework

Scoring matrix

$$S^{\mathrm{Q}}(x,y) = \sum_{i=1}^{N} w_i \, S^{\mathrm{Q}}_{\theta_i,\alpha}(x,y)$$

Forecast category

Observed category

	No warning	Severe	Extreme
No warning	0	1	1.5
Severe	1	0	0.5
Extreme	1.5	0.5	0

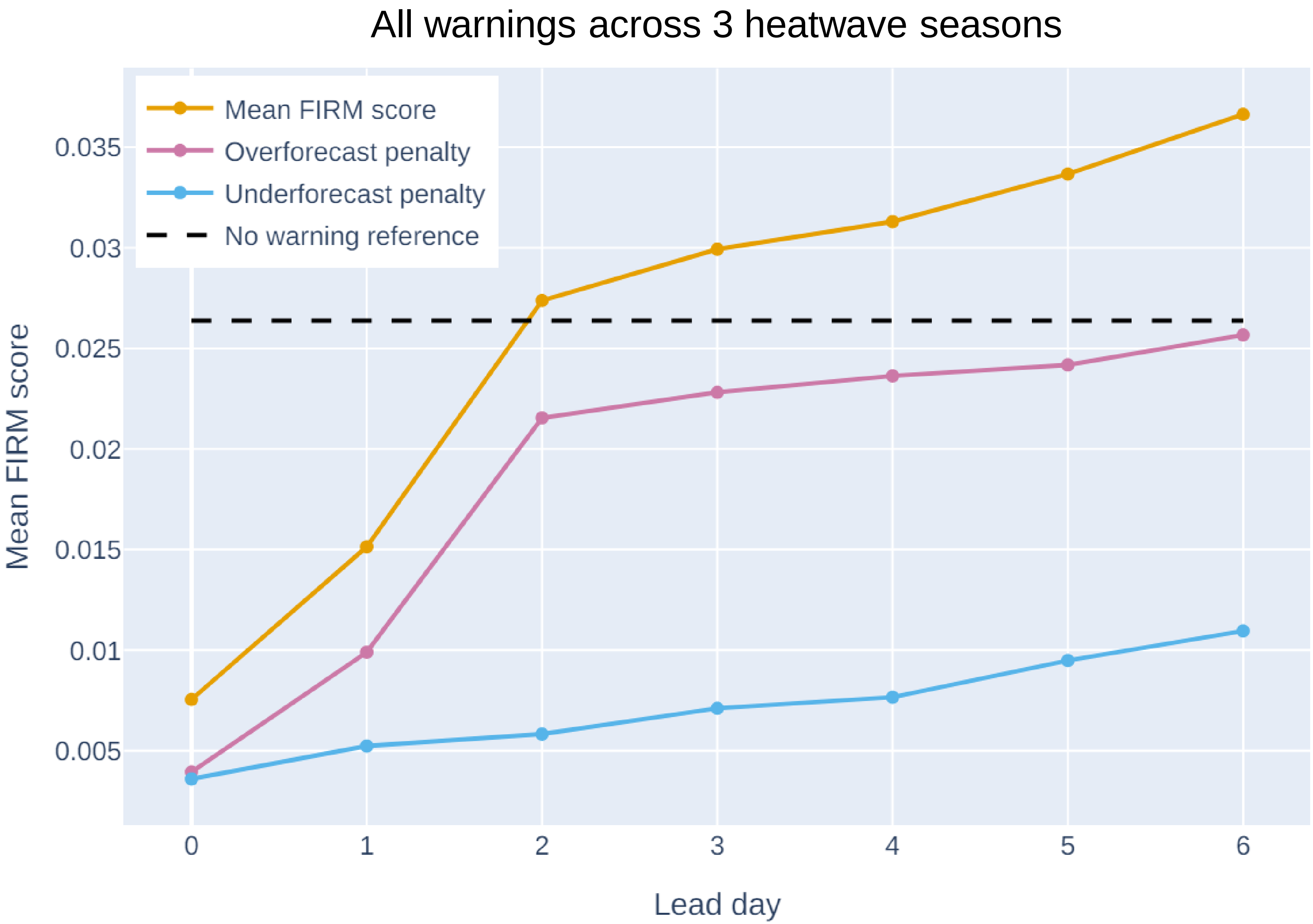
The FIxed Risk Multicategorical (FIRM) Framework

Scoring matrix

$$S^{\mathrm{Q}}(x, y) = \sum_{i=1}^{N} w_i \, S^{\mathrm{Q}}_{\theta_i,\alpha}(x, y)$$

		Forecast category		
		No warning	Severe	Extreme
Observed category	No warning	0	1	1.5
	Severe	1	0	0.5
	Extreme	1.5	0.5	0

Heatwave warning verification results



The FIxed Risk Multicategorical (FIRM) Framework

Scoring matrix

Forecast category

Observed category

	No warning	Severe	Extreme
No warning	0	1	1.5
Severe	1	0	0.5
Extreme	1.5	0.5	0

Over-forecast penalties

Under-forecast penalties

The FIRM score is consistent with the forecast directive:

"Forecast the highest category for which the probability of observing that category or higher exceeds x%"

For a proof of consistency, see

Taggart, R., Loveday, N. and Griffiths, D., 2022. A scoring framework for tiered warnings and multicategorical forecasts based on fixed risk measures. *Quarterly Journal of the Royal Meteorological Society*, *148*(744), pp.1389-1406.

Now for some extensions

Extensions

Discount penalty of near misses and close false alarms

$$S^{\mathrm{H}}_{\theta,\alpha,a}(x,y) = \begin{cases} (1-\alpha)\min(\theta - y, a), & y \le \theta < x, \\ \alpha \min(y-\theta, a), & x \le \theta < y, \\ 0, & \text{otherwise}, \end{cases}$$

Penalty of False Alarm

Penalty of Miss

θ=decision threshold

a = discounting distance parameter

$$S^{\mathrm{H}}(x,y) = \sum_{i=1}^{N} w_i \; S^{\mathrm{H}}_{\theta_i,\alpha,a}(x,y)$$

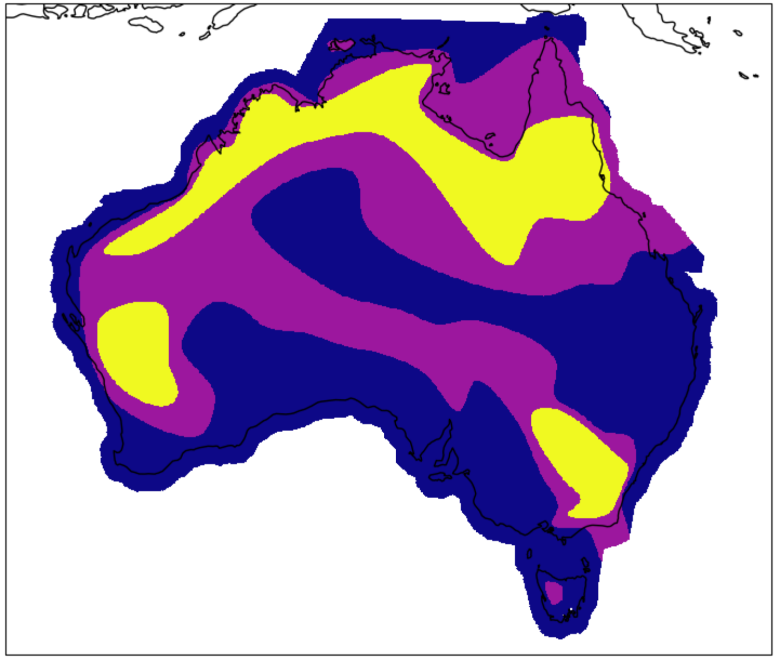
Forecast directive:
"Forecast any category that contains a Huber quantile H(F)"

Still works if forecasts are categorical, but observations are real valued.

Can't visualise a scoring matrix

Extensions

Categorical forecasts for the likelihood of an event



3 categories

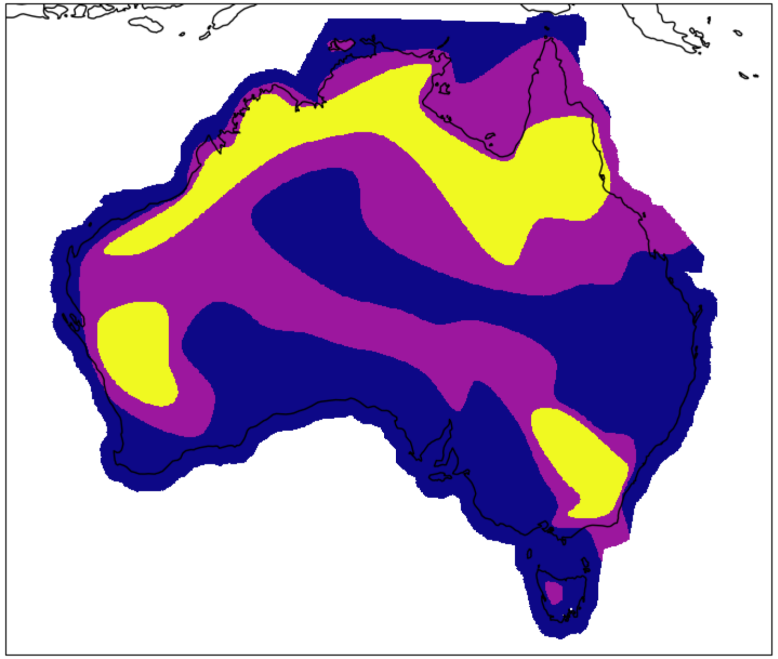
🟦 1. Nil thunderstorm. [0, 10]% chance

🟪 2. Thunderstorm possible. (10, 30]% chance

🟨 3. Thunderstorm likely. (30, 100]% chance

Extensions

Categorical forecasts for the likelihood of an event



3 categories

🟦 1. Nil thunderstorm. [0, 10]% chance

🟪 2. Thunderstorm possible. (10, 30]% chance

🟨 3. Thunderstorm likely. (30, 100]% chance

For the two-category case:

$$S^{\mathrm{B}}_{\theta_i}(p,y) = \begin{cases} \theta, & y = 0,\ p > \theta \quad \longleftarrow \text{Penalty of False Alarm} \\ 1-\theta, & y = 1,\ p \le \theta, \quad \longleftarrow \text{Penalty of Miss} \\ 0, & \text{otherwise.} \end{cases}$$

θ = probabilistic decision threshold

For multiple categories:

$$S^{\mathrm{B}}(p,y) = \sum_{i=1}^{N} w_i\, S^{\mathrm{B}}_{\theta_i}(p,y),$$

Extensions

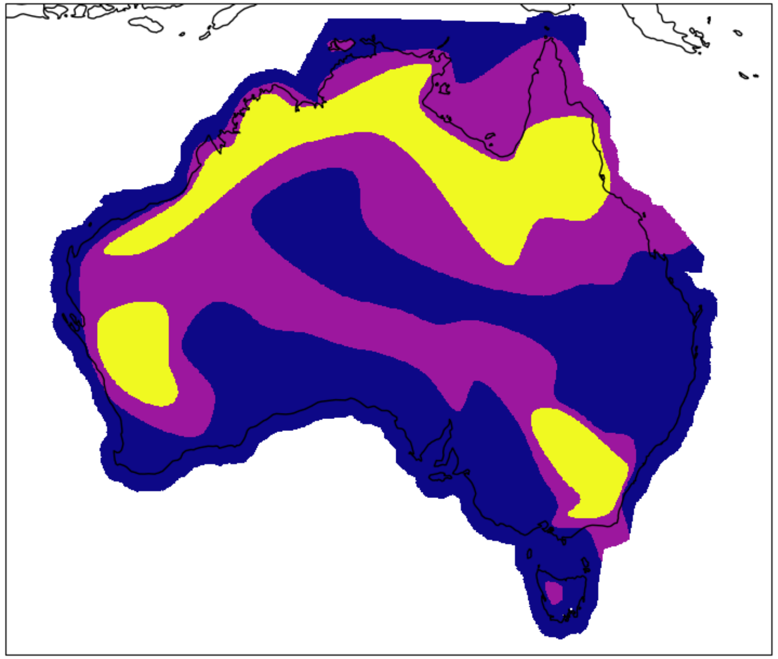
Categorical forecasts for the likelihood of an event

Forecast category	Observed non-event	Observed event
Nil thunderstorm 0-9%	0	$w_1(1-\theta_1) + w_2(1-\theta_2)$
Thunderstorm possible 10-29%	$w_1\theta_1$	$w_2(1-\theta_2)$
Thunderstorm likely 30-100%	$w_1\theta_1 + w_2\theta_2$	0

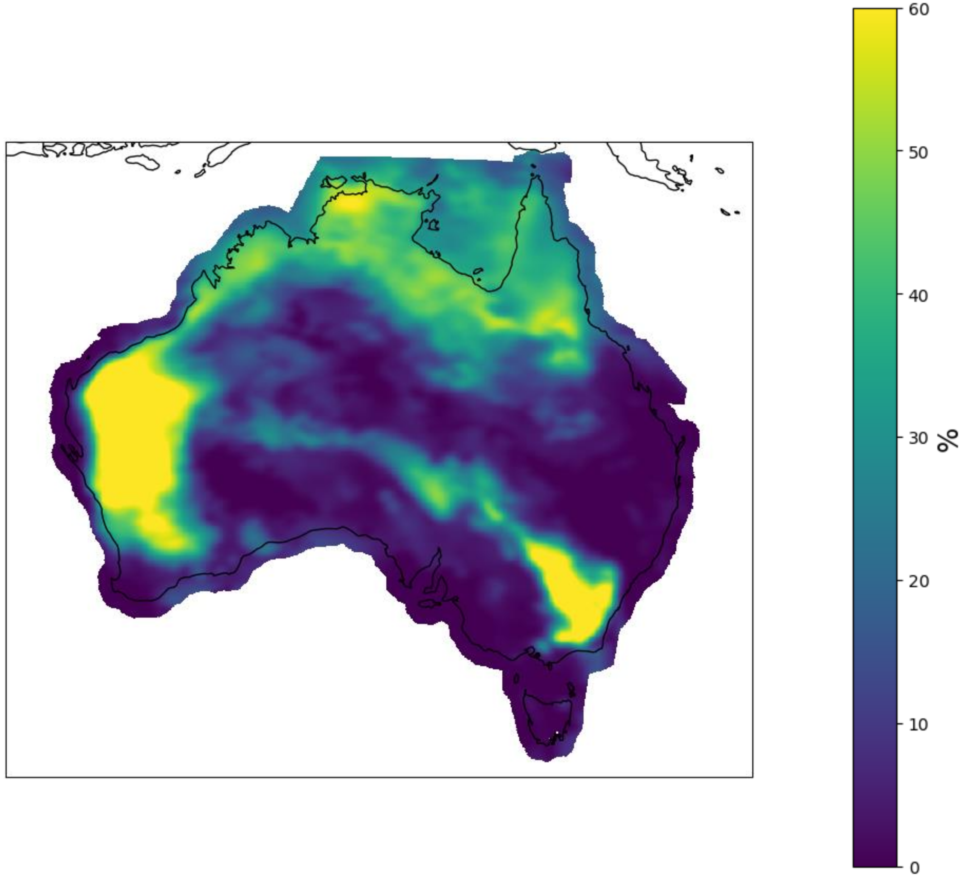
Forecast directive that optimises the expected score:
"*Forecast the category that the likelihood of the event falls within*"

Extensions

Categorical forecasts for the likelihood of an event



VS



3 categories

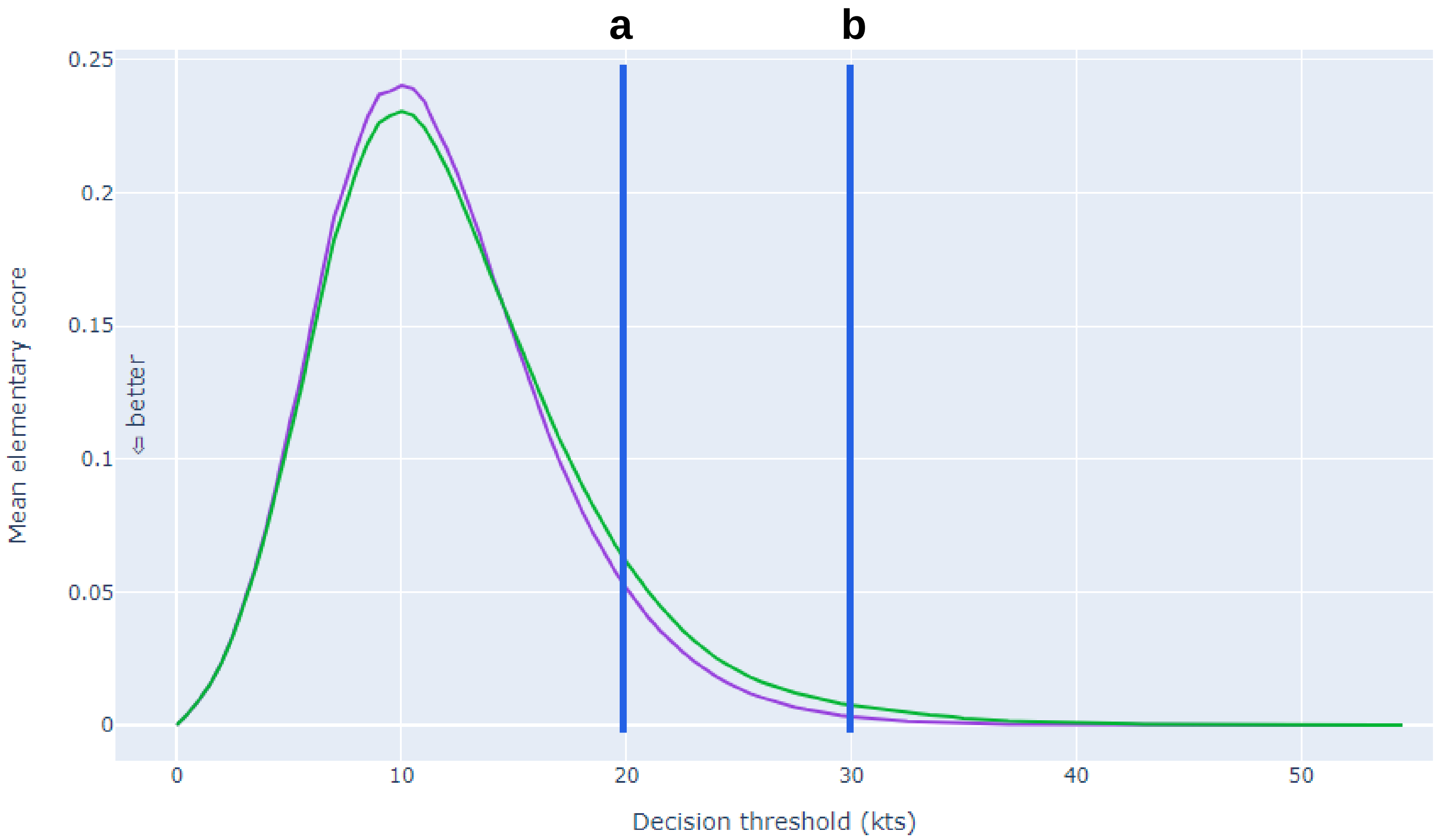
🟦 1. Nil thunderstorm. [0, 10]% chance

🟪 2. Thunderstorm possible. (10, 30]% chance

🟨 3. Thunderstorm likely. (30, 100]% chance

See **Loveday, N., Taggart, R. and Khanarmuei, M., 2024. A User-Focused Approach to Evaluating Probabilistic and Categorical Forecasts.** ***Weather and Forecasting***

Relationship to Murphy Diagrams



Summary

- If issuing warnings based on fixed risk is important, then consider using FIRM rather than an equitable score.
- The FIRM score is consistent for the forecast directive:

> *"Forecast the highest category for which the probability of observing that category or higher exceeds x%"*

- You can control the weights of the importance of each decision threshold and the ratio of the penalties for misses vs false alarms.
- There are extensions to handle near misses and close false alarms, as well as categorical probabilities of an event.

Taggart, R., Loveday, N. and Griffiths, D., 2022. A scoring framework for tiered warnings and multicategorical forecasts based on fixed risk measures. *Quarterly Journal of the Royal Meteorological Society*, *148*(744), pp.1389-1406.

Python code at https://github.com/nci/scores

Contact: nicholas.loveday@bom.gov.au