From metric to action:
The decision value of infectious disease forecasts

Cathal Mills

Florence Nightingale Research Fellow, University of Oxford

Collaborators:

Nicholas J. Irons, Joseph. L.-H. Tsui, Sarah Sparrow, Luiz M. Carvalho, Adam J. Kucharski, Oliver Ratmann, Ben Lambert, Christl A. Donnelly, and Moritz U. G. Kraemer

UNIVERSITY OF OXFORD

*"**Forecasts possess no intrinsic value.***

*They acquire value through their **ability to influence decisions made by users of the forecasts.**"*

Allan H. Murphy

cathal.mills@stats.ox.ac.uk

UNIVERSITY OF OXFORD

1) Why?

cathal.mills@stats.ox.ac.uk

Why ID forecasting?

THE HYPE – BENEFITS + USES

Future = **uncertain =>** Need **reliable + timely information**

Simple: Forecast = **best estimate** of what ***will*** **happen**

Ideally … **support decisions** under
- **Pressures**
- **Uncertainty**

Users + uses? Varied **policy qs**
- **Short-term**: Outbreak response
- **Seasonal** + long-term: PH planning

THE LIMITATIONS

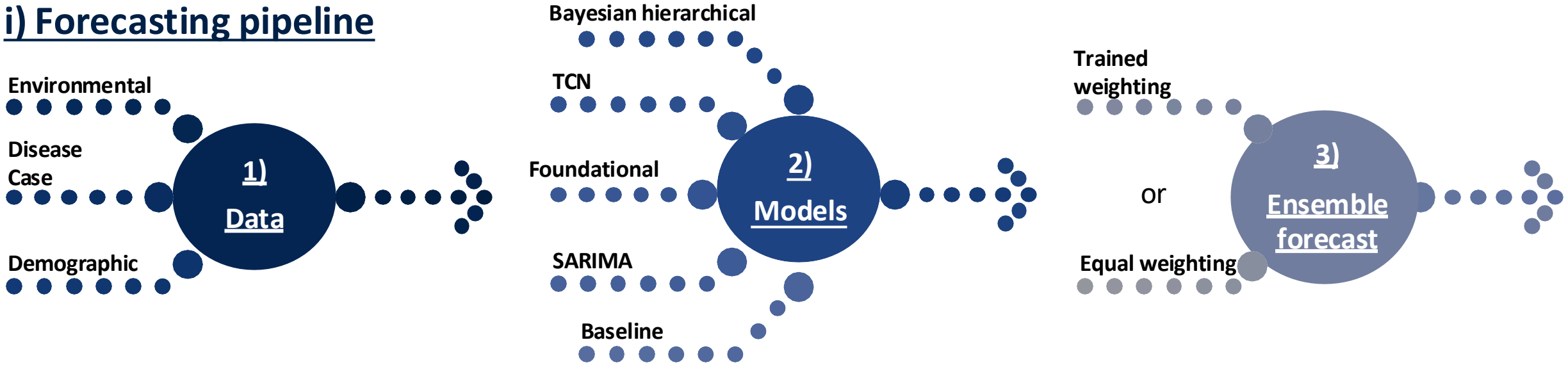
- Epidemics = highly non-linear, stochastic, + **feedback loops**
- Predict "**modifiable future**"?
- **1 part** of modellers' + decision-makers' **toolkits**
- **Understanding** = crucial

Cathal Mills

cathal.mills@stats.ox.ac.uk

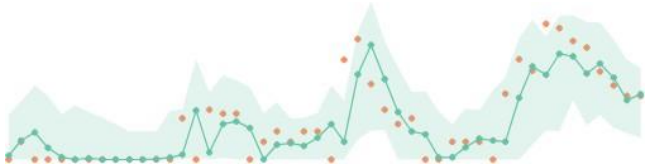
Example: Dengue forecasting

i) Forecasting pipeline



ii) Forecast evaluation

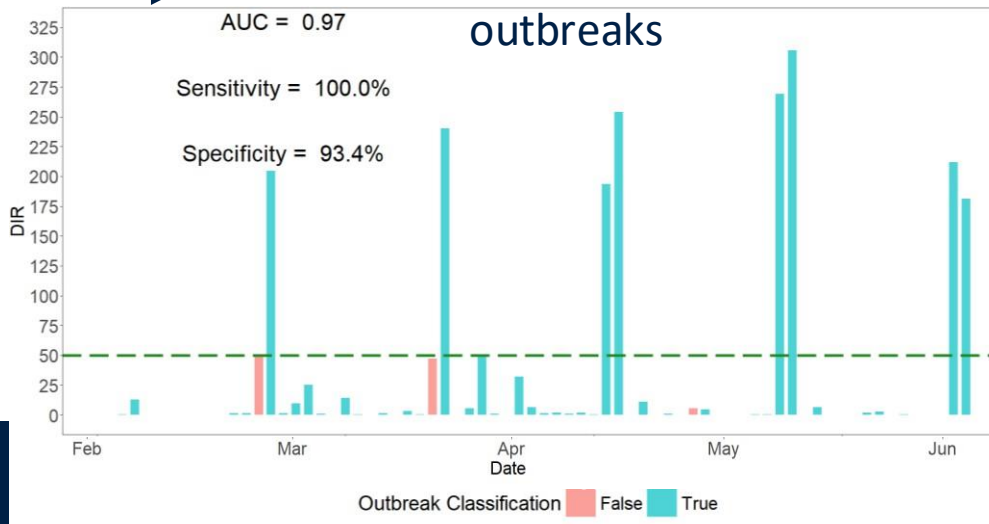
Probabilistic forecast
Predict cases with uncertainty



"Proper" scoring rules
Measure whole predictive distribution

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - 1(x \geq y))^2 dx$$

Public-health oriented metrics
Historically optimised classification of outbreaks

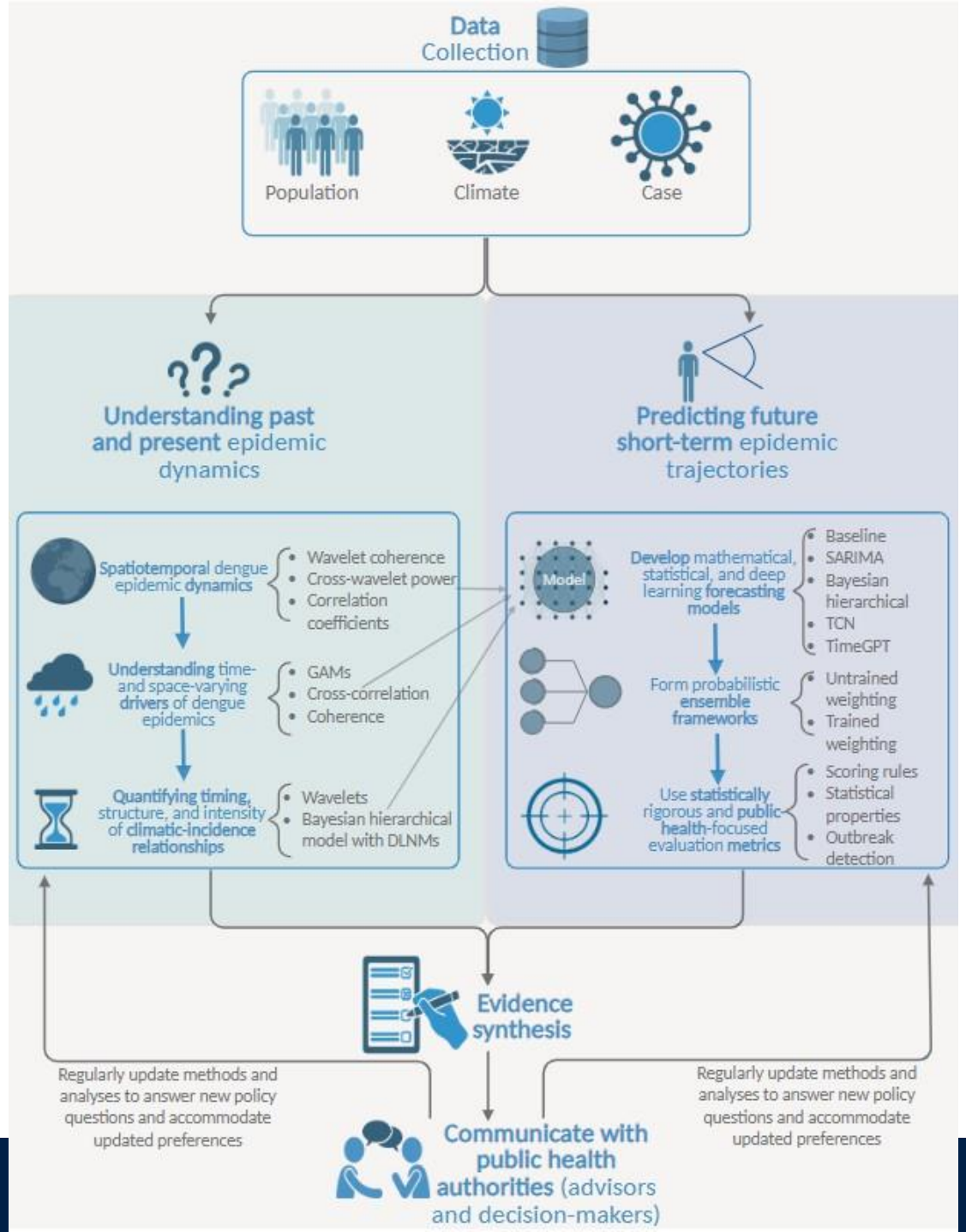


Cathal Mills, Francesca Falconi-Agapito, Jean-Paul Carrera, César V. Munayco, Moritz U. G.Kraemer, and Christl A. Donnelly.
Multi-model approach to understand and predict past and future dengue epidemic dynamics.
Royal Society Open Science, August 2025

Accurate forecast ≠ effective policy

… part of a wider multi-model approach to

i) Learn from the past: Epidemic drivers and spatiotemporal dynamics

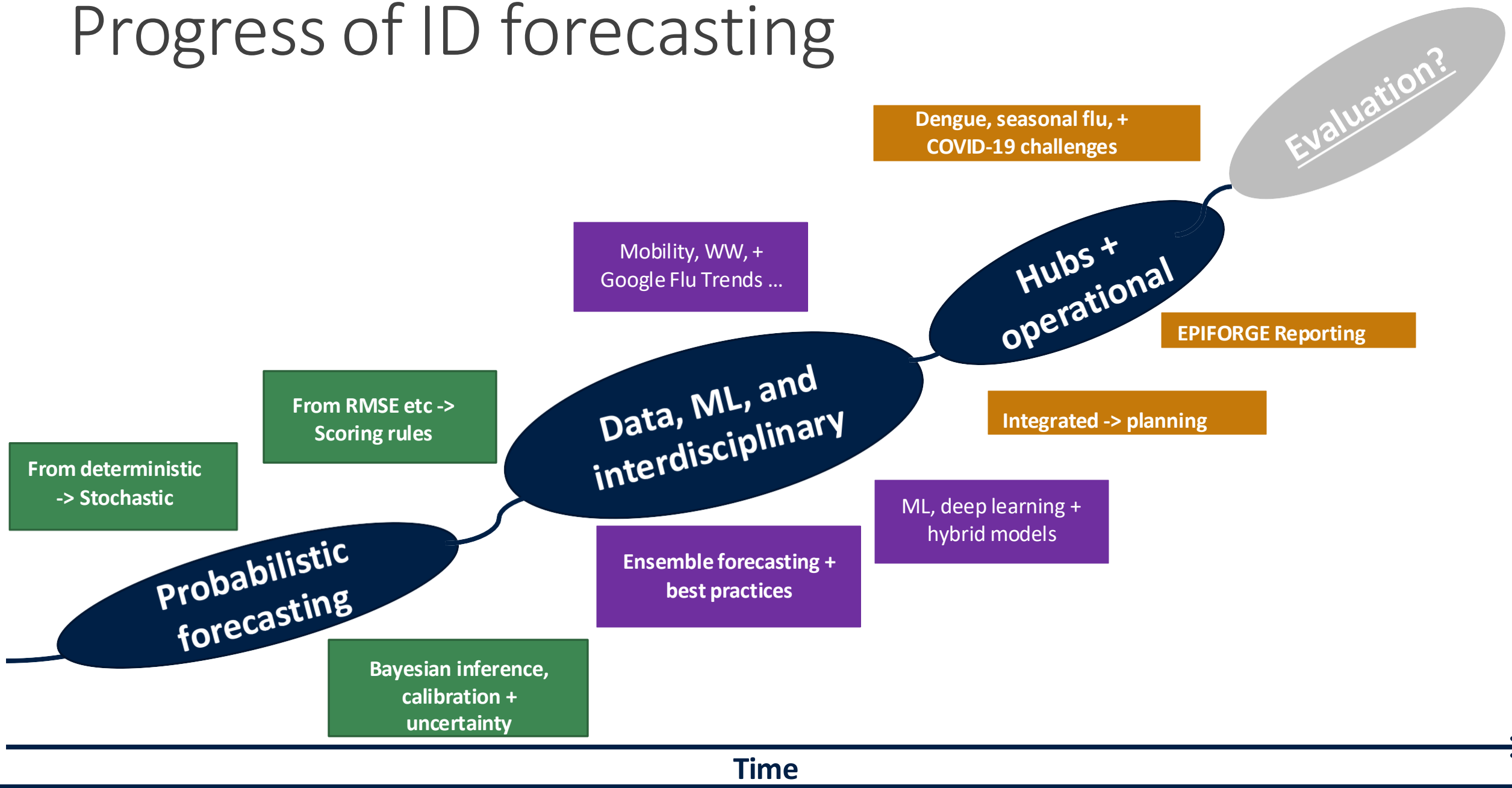


Cathal Mills, Francesca Falconi-Agapito, Jean-Paul Carrera, César V. Munayco, Moritz U. G.Kraemer, and Christl A. Donnelly. Multi-model approach to understand and predict past and future dengue epidemic dynamics.
Royal Society Open Science, August 2025

ii) Plan for the future: Forecast future short-term trajectories

cathal.mills@stats.ox.ac.uk

Progress of ID forecasting



Cathal Mills

cathal.mills@stats.ox.ac.uk

Yet ….

Forecasts + decisions <u>not</u> **epi-specific** problems	E.g. forecasts support decisions under pressure for **energy, economics, weather, finance …**
=> Our framework integrates **branches of science**:	Information theory
	Decision theory
	Weather forecasting
	Operations research

cathal.mills@stats.ox.ac.uk

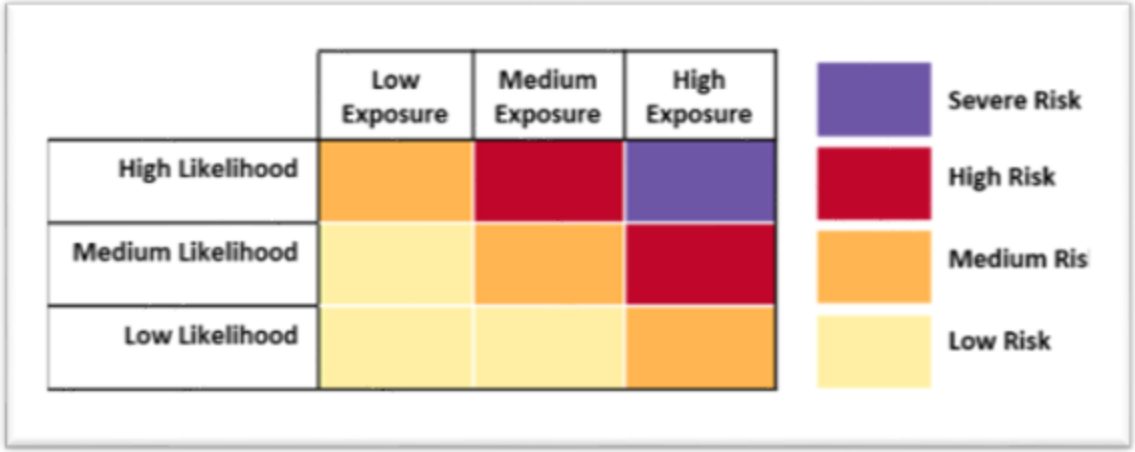
Learn from weather forecasting?

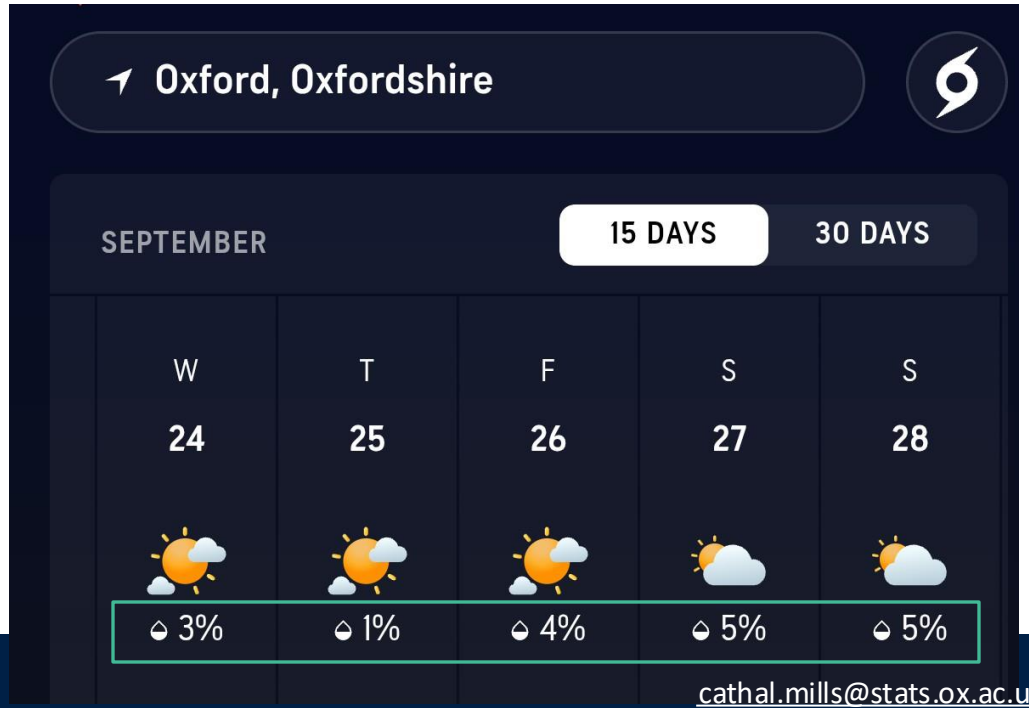
Thus far,

- Ensembles, hubs, and scoring

Epi forecasting tomorrow?

- **Systematic evaluation procedures** to measure decision value
- **Customised products** for stakeholders
- **Impact-based** forecasting
- **Communication + uncertainty** -> your phone…
- **Weather risk management**; identify, educate, + mitigate
- **Hurricane vs Epidemics….**





Cathal Mills
cathal.mills@stats.ox.ac.uk

The forecast-decision gap

- Is model **"good" for informing decisions**?
- How **interpret model rankings** e.g. by CRPS?
- Can **decision-maker take action**? ***Why* use model?**
- Do **rankings +/ actions change by decision-maker**?
- How incorporate **risk preferences, local context, resources, + priorities**?
- Public-health users care about **actionability**: should I act now, wait, or allocate resources differently?
- ….

	Model	WIS	Bias	PI Coverage	R^2	RMSE	Sensitivity	Specificity	AUC
1	Median *	0.34	0.03	95.2%	0.74	0.81	81.8%	94.3%	0.88
2	Median-NoBase *	0.35	0.02	94.8%	0.73	0.82	81.8%	95.7%	0.89
3	Median-NoBayes *	0.35	0.06	94.8%	0.74	0.82	81.8%	87.2%	0.85
4	EW-Mean *	0.35	0.17	87.8%	0.75	0.79	72.7%	95.9%	0.84
5	Median-NoCov *	0.36	0.04	95.2%	0.73	0.83	81.8%	91.0%	0.86
6	Ew-Mean-NoBayes *	0.36	0.21	87.5%	0.74	0.81	63.6%	97.3%	0.80
7	EW-Mean-NoCov *	0.36	0.14	91.2%	0.74	0.81	81.8%	86.5%	0.84

What we know?
Forecast *quality* (necessary)

?

What is less clear?
***Value:* Link Metric -> Action**

Cathal Mills
cathal.mills@stats.ox.ac.uk

Objectives today

Q = What is a "good" forecast? best model on average not best for individual time points nor individual decision-makers -> **event, user, + action can take**

<u>Answer</u> = *Value* of a forecast measured by

ability to inform decisions under uncertainty

<u>Not</u> forecast => decision, nor one-size-fits-all approach ….

Instead, **if** we are to **forecast,**

- **Evaluation** must **focus on decision-maker**, priorities and risk preferences
- Decisions always **subjective yet systematic** evaluation

Bridging gaps from statistical metrics to public health decision value:

cathal.mills@stats.ox.ac.uk

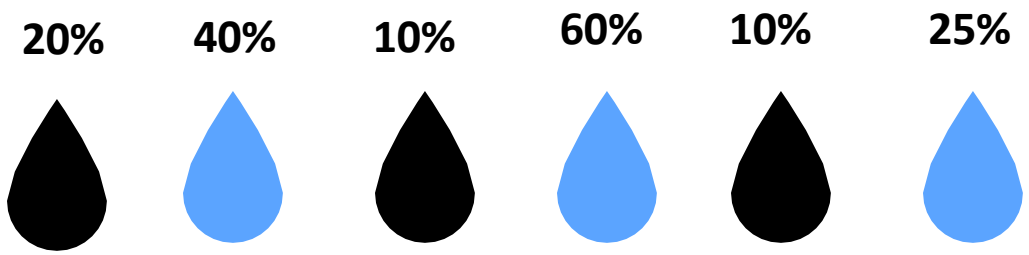
2) Decisions and forecasts: From umbrellas to scoring rules + functions….

Cathal Mills

cathal.mills@stats.ox.ac.uk

From umbrellas to scoring….

Model =
Weather forecast
P(rain)

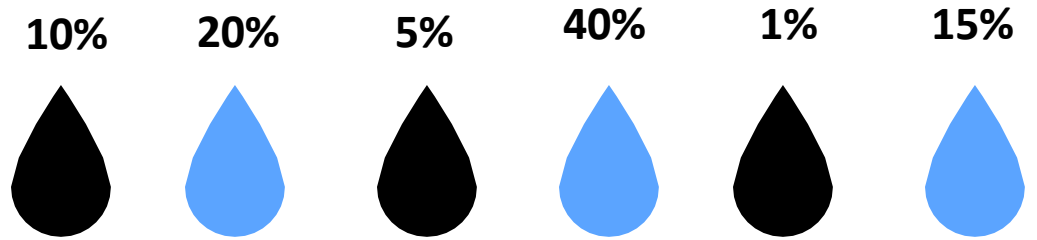


Decision:
Umbrella?

User decision rule = umbrella if P(rain) >= 0.3

but…. **how are we defining ``rain''?**
E.g. precipitation above user-defined level
> User's decision will also depend on **event** definition <

Model =
Weather forecast
P(heavy rain)



Decision:
Umbrella?

User Decision rule = umbrella if P(severe rain) >= 0.15

Cathal Mills
cathal.mills@stats.ox.ac.uk

Takeaways for epidemics

Define **simultaneously:**

i) the forecast **user (decision-maker) risk preferences**

ii) the **event threshold**

In **epidemic,** these **vary across space and time -> Local context = crucial ….**

- **Epidemics = fast-evolving**
- **Resources** and **intervention** options **differ**
- Transmission **histories differ**

Cathal Mills

cathal.mills@stats.ox.ac.uk

UNIVERSITY OF OXFORD

3) A new evaluation framework

FORECAST'S *VALUE*:=

ABILITY TO INFORM DECISION-MAKER'S DILEMMA;

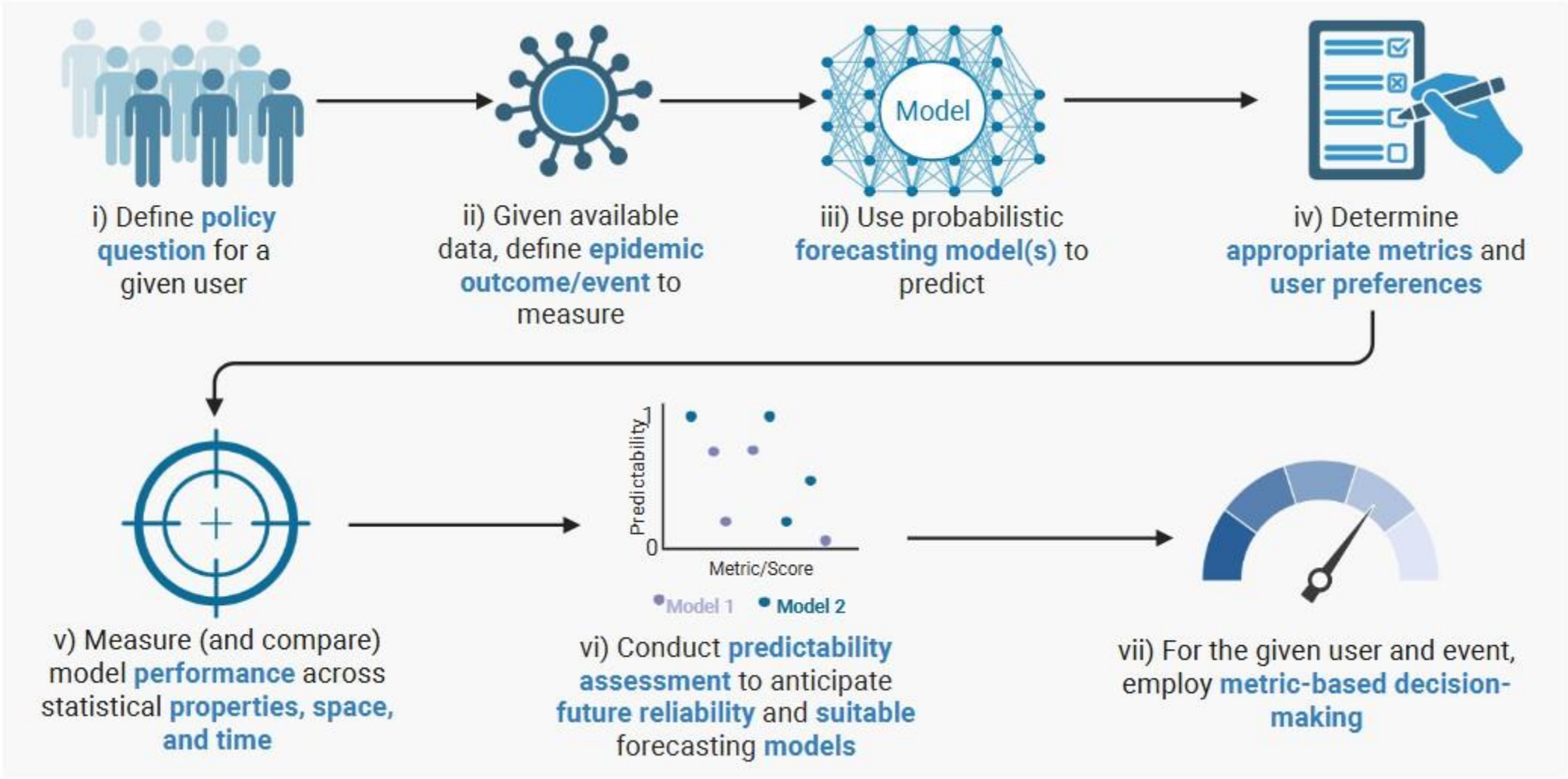
TO ACT OR NOT

Cathal Mills

cathal.mills@stats.ox.ac.uk

Forecasting workflow:

Applied iteratively per Q, space + time



i) Define **policy question** for a given user

ii) Given available data, define **epidemic outcome/event** to measure

iii) Use probabilistic **forecasting model(s)** to predict

iv) Determine **appropriate metrics** and **user preferences**

v) Measure (and compare) model **performance** across statistical **properties, space, and time**

vi) Conduct **predictability assessment** to anticipate **future reliability** and **suitable** forecasting **models**

vii) For the given user and event, employ **metric-based decision-making**

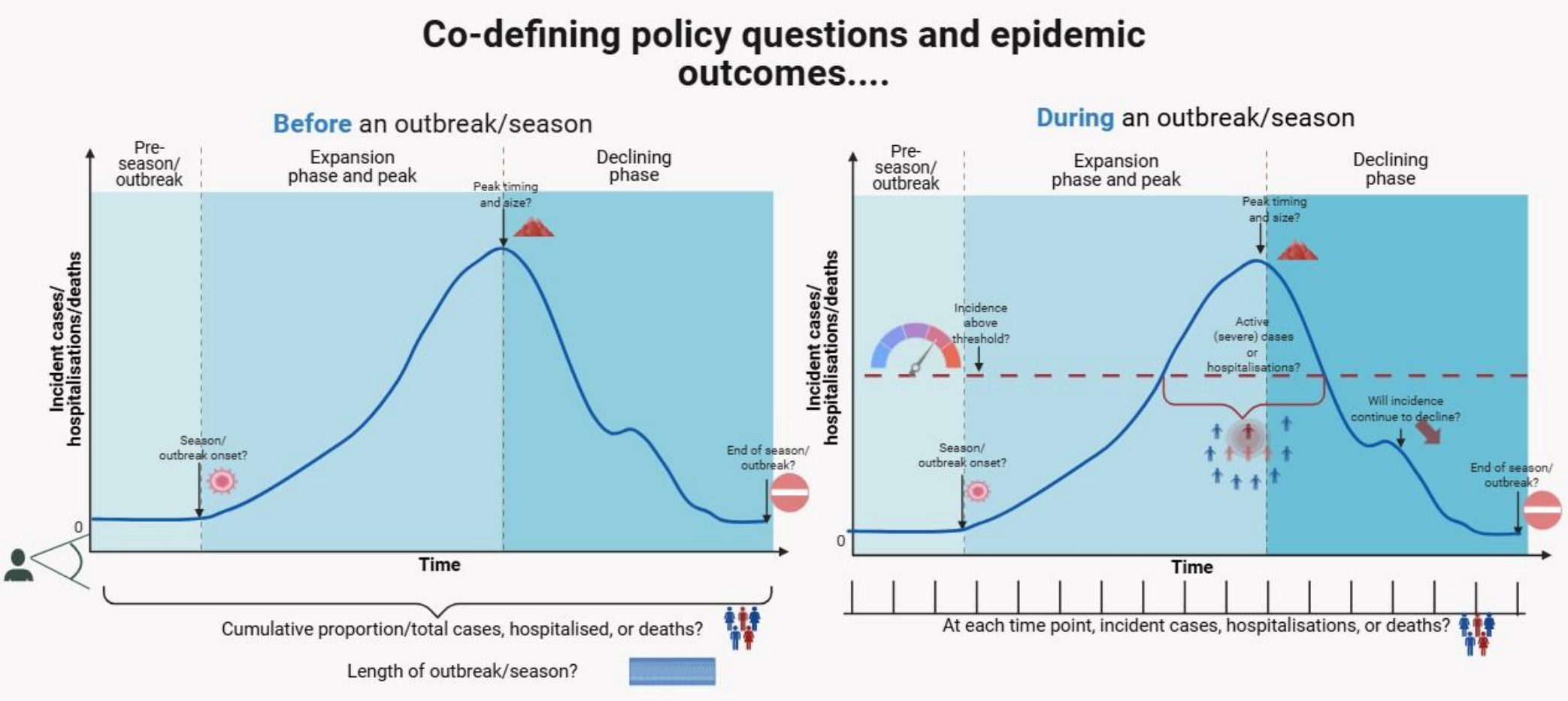
Cathal Mills

cathal.mills@stats.ox.ac.uk

Starting with decision-maker

Anchor to the policy question

"*Models need questions to answer, otherwise they are just lines on graphs.*"
Medley (2022)



cathal.mills@stats.ox.ac.uk

How to evaluate forecasts?

Cost-Loss Model

Binary actions
+
binary events
(y >= θ)

Action / event matrix

	Event occurs	Event does not occur
Action taken	C	C
Action not taken	L	0

Cost-loss model:

- **C = Cost of action**
- **L = Preventable loss** (event w/o action)
- **Ratio = α**
- **Action**? E.g. Additional ICU capacity
- **Event**? E.g. Cases >= 10,000

Optimal decision:

- **"Act" if** predicted **P(event) >= C/L Ratio (α)**
- **Small α -> Vulnerable** decision-maker
- **Big α -> Risk-tolerant** decision-maker
- **Only** need to **know ratio ->** Relative costs vs losses

Cathal Mills
cathal.mills@stats.ox.ac.uk

Why α and θ? Some theory

Make assumptions about forecasts score explicit

All popular scoring rules always make **assumptions** about:

- how decision-maker **weights each combination of α and θ**
- i.e. assumptions about α (**risk preferences**) **for each** θ (**event threshold**)

Useful but obscure event-probability space -> operational needs

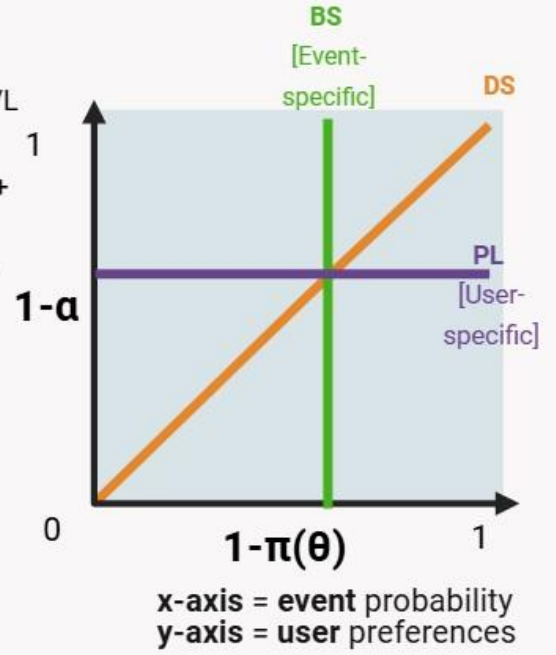
Any score S is formed by combining
i) **elementary scoring rule** + ii) weighting/**utility function**:

$$S(\mathcal{F}, y, u) = \int_{-\infty}^{+\infty} \int_{0}^{1} s_{\alpha,\theta}\left(\mathcal{F}^{-1}(1-\alpha), y\right) u(\alpha, \theta) \mathrm{d}\alpha \, \mathrm{d}\theta$$

and a **mean score** is calculated across observations

$$\bar{S} = \frac{1}{N} \sum_{i=1}^{N} S_i$$

- **CRPS** = All events (θ)+ all C/L (α) [Whole box]
- **Brier score**: Fixed event (θ) + all C/L (α) [Vertical]
- **Pinball loss**: Fixed C/L (α) + all events (θ) [Horizontal]
- **Diagonal score**: Hold C/L = Event base rate, π(θ) [Diagonal]



$$s_{\alpha,\theta}(x, y) = \begin{cases} 1-\alpha, & \text{if} \quad x \le \theta < y \\ \alpha, & \text{if} \quad x > \theta \ge y \\ 0, & \text{otherwise.} \end{cases}$$

Punish misses + false alarms
Asymmetrically,
Depending on **C/L ratio**

Cathal Mills
cathal.mills@stats.ox.ac.uk

For more on scoring...

Theory

Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings

Werner Ehm, Tilmann Gneiting ✉, Alexander Jordan, Fabian Krüger

First published: 10 May 2016 | https://doi.org/10.1111/rssb.12154 | VIEW METRICS

Model Diagnostics and Forecast Evaluation for Quantiles

Tilmann Gneiting[1,2], Daniel Wolffram[1,3], Johannes Resin[1,2], Kristof Kraus[1,2], Johannes Bracher[1,3], Timo Dimitriadis[1,4], Veit Hagenmeyer[5], Alexander I. Jordan[1], Sebastian Lerch[1,3], Kaleb Phipps[5] and Melanie Schienle[1,3]
View Affiliations and Author Notes

Vol. 10:597-621 (Volume publication date March 2023) | https://doi.org/10.1146/annurev-statistics-032921-020240

Applications

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

The diagonal score: Definition, properties, and interpretations

Zied Ben Bouallègue ✉, Thomas Haiden, David S. Richardson

First published: 30 March 2018 | https://doi.org/10.1002/qj.3293 | VIEW METRICS

Decisions, decisions...!

Tim Palmer (University of Oxford), David Richardson (ECMWF)

For IDs: Why + how define **α for each θ?**

Incorporate risk appetite for each epidemic event threshold

α = C/L

Why?

Already one **goal** = control **tail risks, e.g. COVID-19 onset**

Now,

i) no fair, group-level utility function,

ii) focus = decision-maker,

iii) "**optimality**"

How?

E.g.: C - Economic action costs, L - Life + macroeconomic costs …

Non-trivial, time- + space-varying, **always uncertain ->** Immense **ethical + human components**

Just need **plausible ranges**

Factors = i) **decision-maker (who?)** + local context, ii) actions + **events**, iii) **model** calibration

Propose: i) **decision-maker** + modeller **surveys** , ii) economic evaluation + iii) **real-time** collaboration + updating

Cathal Mills

cathal.mills@stats.ox.ac.uk

UNIVERSITY OF OXFORD

4) Metrics, visualisations, and applications

Cathal Mills

cathal.mills@stats.ox.ac.uk

Metric i) – Relative Economic Value

For **fixed event, vary risk prefs:**

- **Operational value** of model
- Compare **"expenses"** incurred by model **vs a baseline**

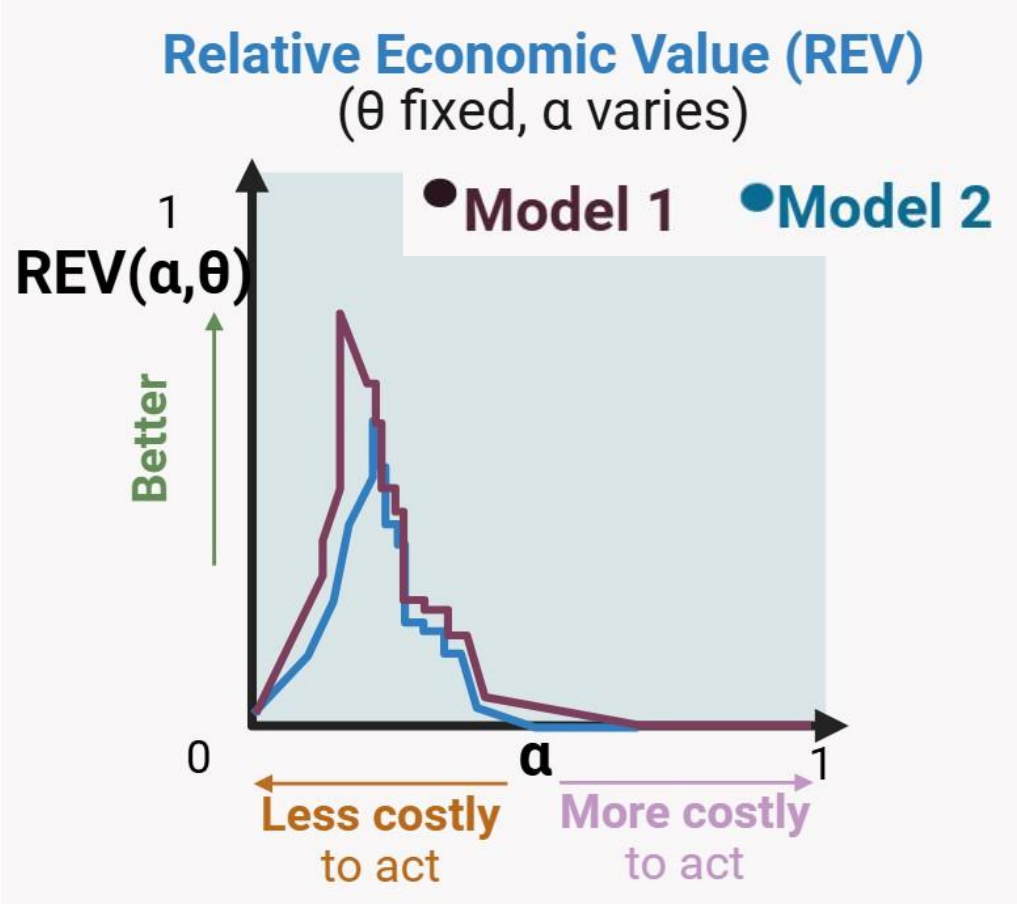
$$\mathrm{REV}(\alpha = C/L, \theta) := \frac{E_{\text{baseline}} - E_{\text{model}}}{E_{\text{baseline}} - E_{\text{perfect}}}.$$

- w/ expenses E = C and L incurred by TP, FP, FN

Pros/cons

✓Interpretable, operational, event-specific

X Single event, need baseline, no elementary scores



cathal.mills@stats.ox.ac.uk

Metric ii) – Murphy diagrams

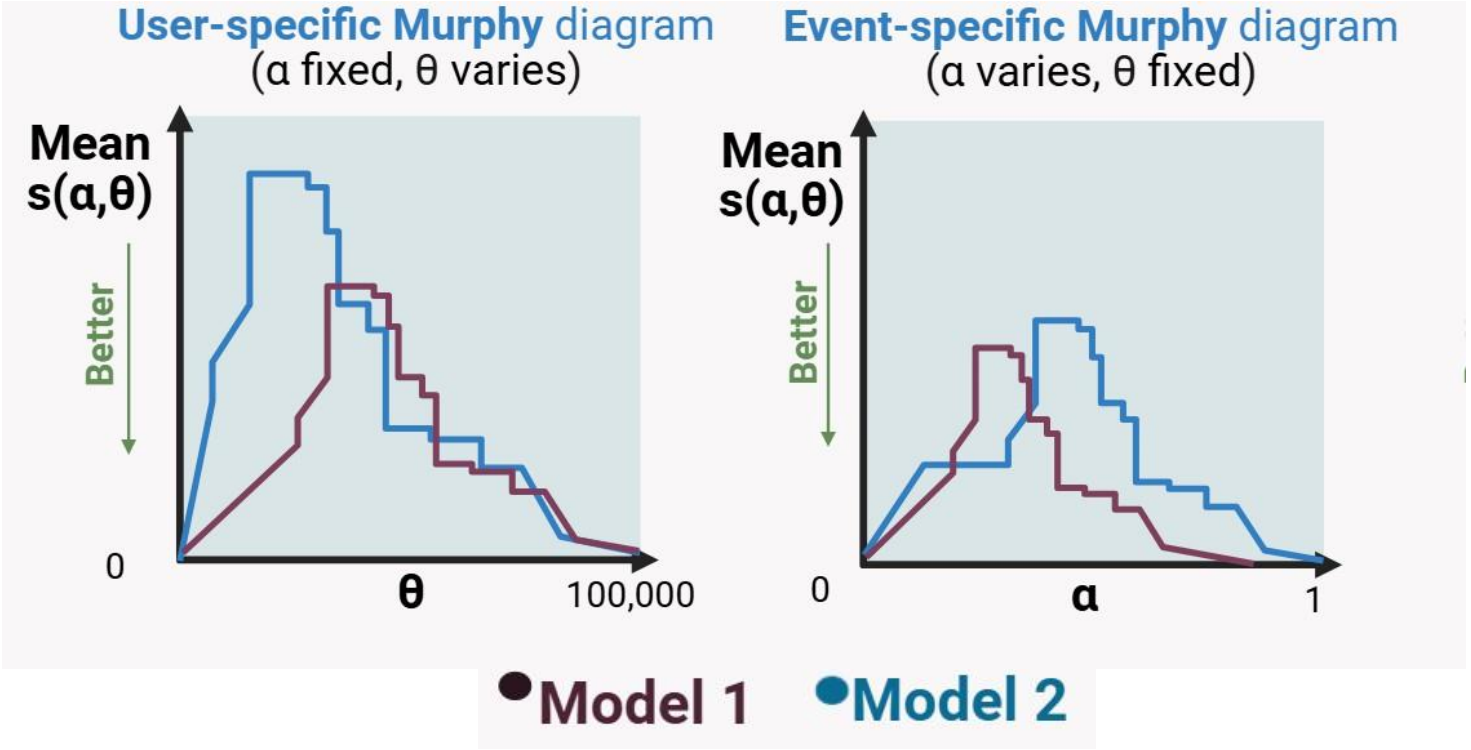
For either **fixed event/risk preferences**:

- Uses raw **elementary scores** s(α, θ)
- **Mean score** across for each α/θ
- Allows **uncertain** priorities/risk preferences

Pros/cons

✓Interpretable, either event-/user specific

X Single event/risk preferences, summary score ≠ universally optimal



cathal.mills@stats.ox.ac.uk

APPLICATION TO COVID-19 Forecast Hub:

The ensemble was usually best on average, but decision-focused rankings depended on the question asked

Case study shows why a **single leaderboard** is **not enough.**

Main results

- Using **WIS/rWIS**, the ensemble generally **outperformed** (horizons, locations, and trailing windows)
- When zoomed into **specific risk preferences and extreme events**, the ensemble, baseline, and Karlen-pypm models often led, but not for every user-event combination.
- **Similar models** for DSC-MCB decompositions (**sharpness-calibration**)

WIS / rWIS

Ensemble = Best average performer across most settings

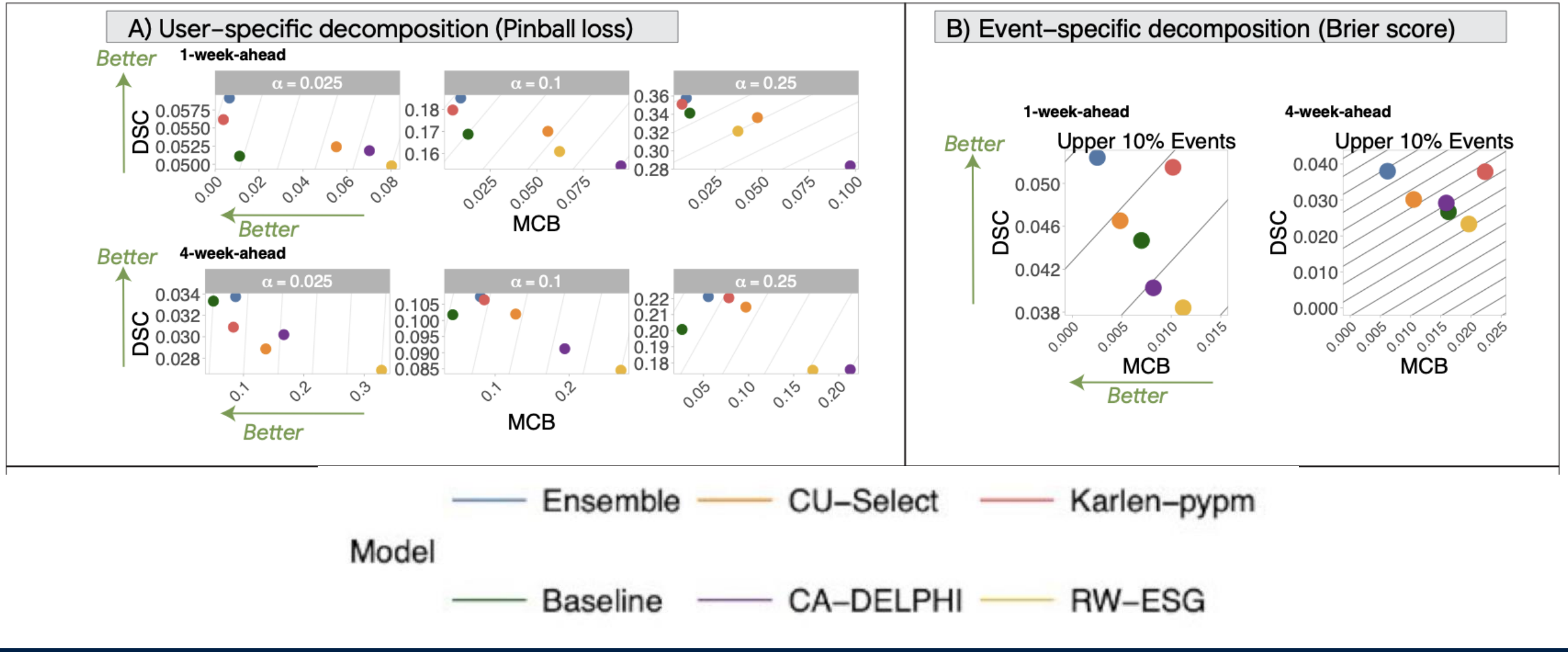
Pinball + BS

Value depends on the user preferences and the event threshold

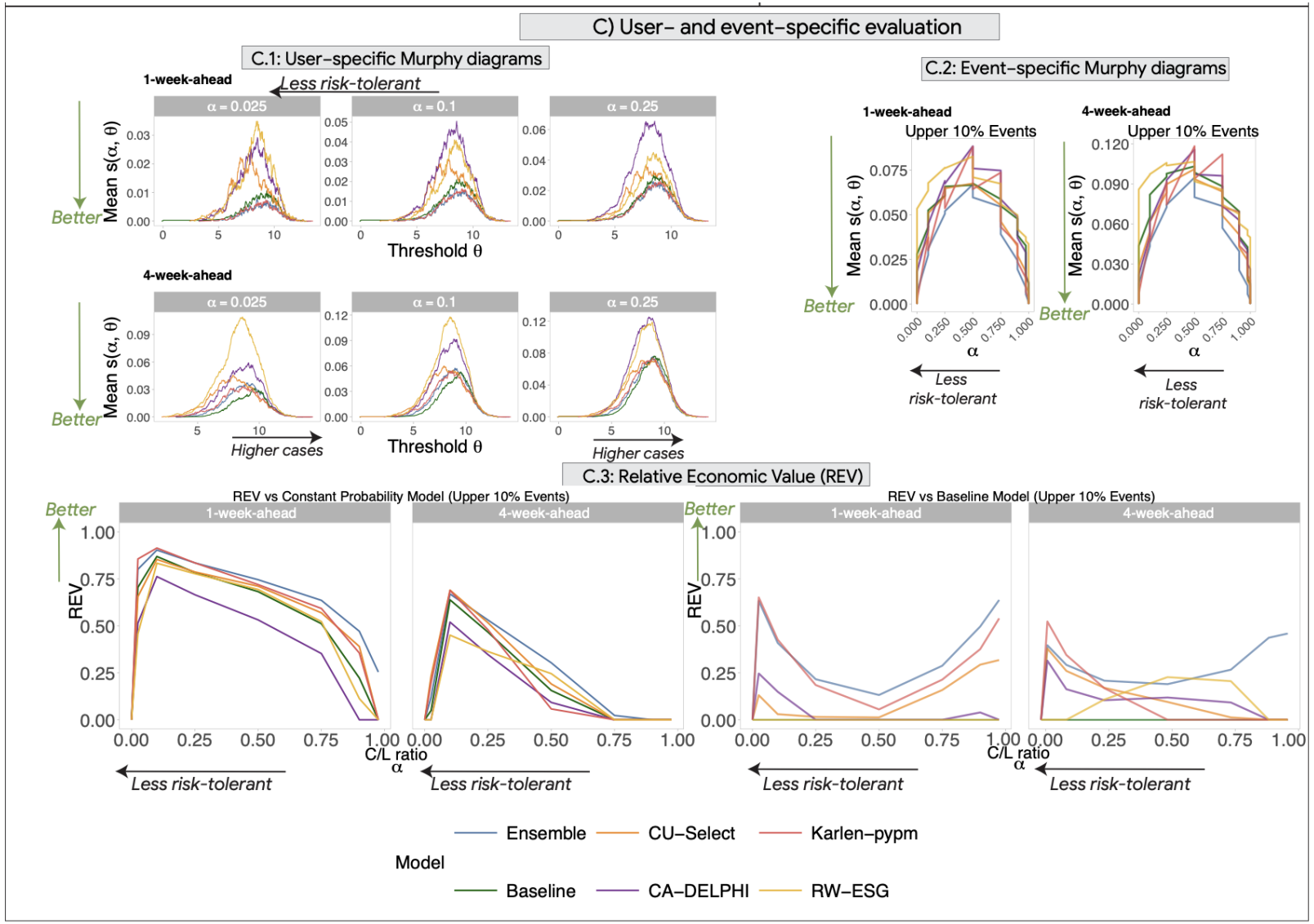
Murphy + REV

Decision value can differ sharply by model, risk profile, and horizon

Some visualisations (1/2)



Some visualisations (2/2)



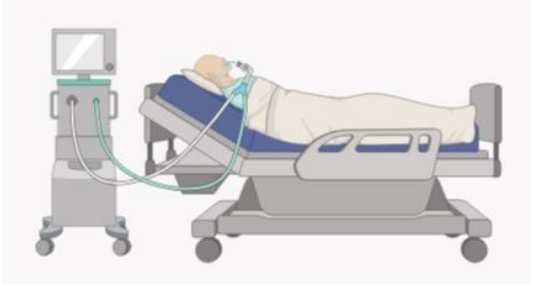
Guardrails: Predictability + DRO

UNIVERSITY OF OXFORD

- Q: **When** + **why** models be ***less*** **useful?**
- **Historically "optimal" model may not be optimal in future**
- Predictability := inherent randomness in epidemic
- **Why <u>predictability</u>**?
  1. Is poor **performance** due to **model or data**?
  2. Epidemics = evolving. Can **safeguard** vs **data distribution shifts?**

  -> Better: models + evaluation + decisions

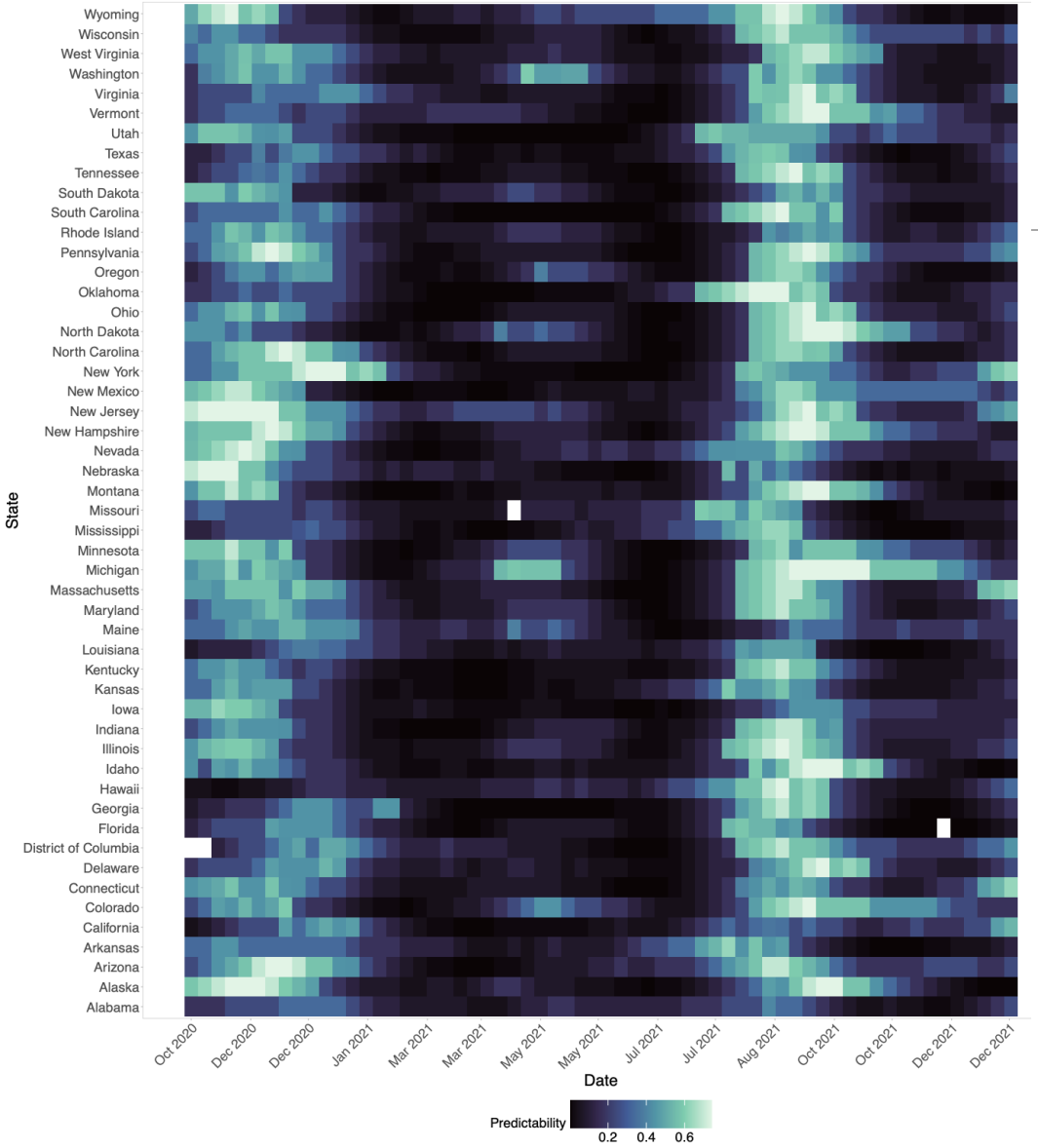
- **<u>DRO</u> = Distributionally Robust Optimisation**
  - Avoid **worst outcomes, robust -> uncertainty**
  - Tackle framework limitations -> **Future work**

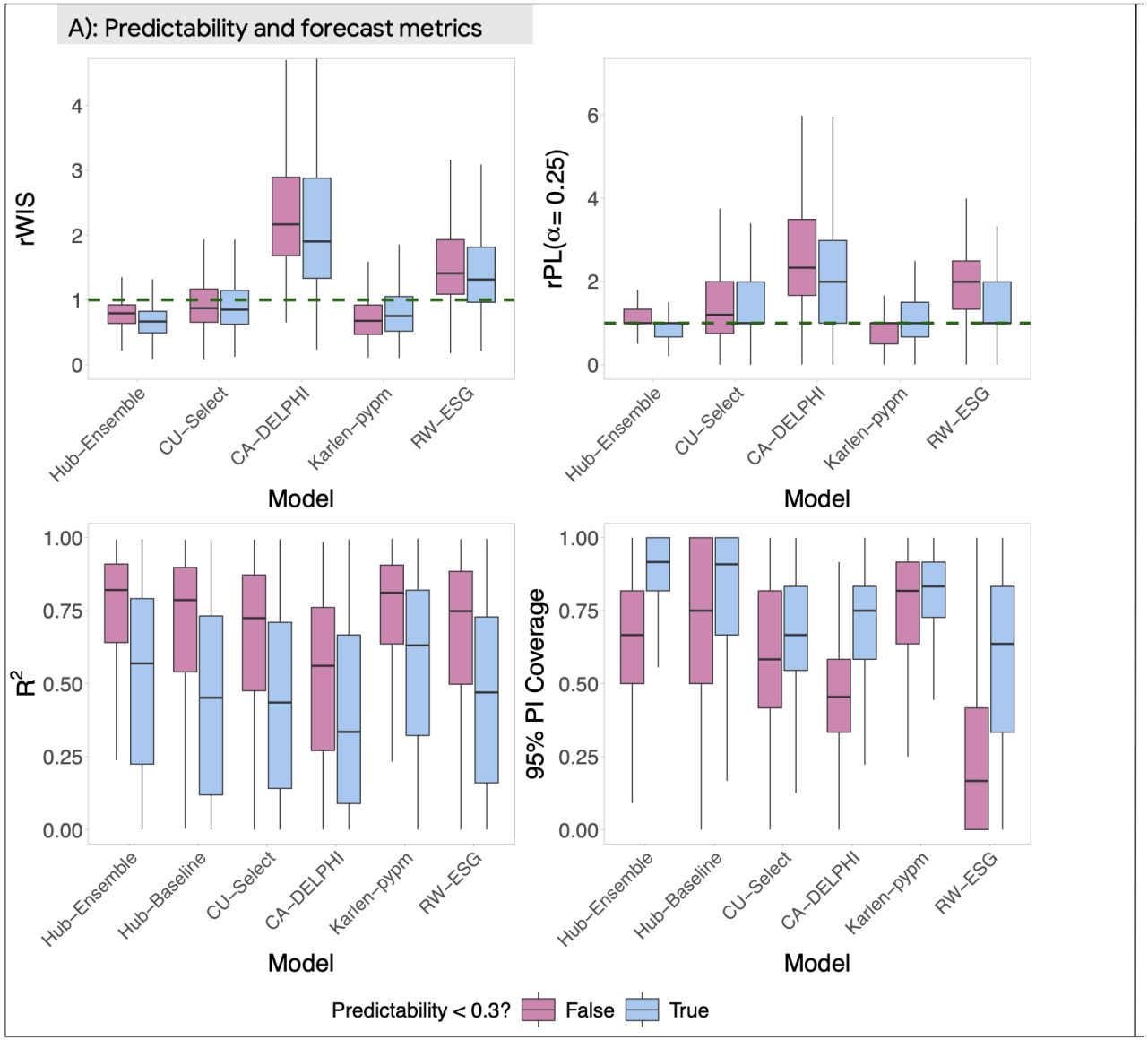


Cathal Mills

cathal.mills@stats.ox.ac.uk

Predictability in practice





Cathal Mills

cathal.mills@stats.ox.ac.uk

Limitations + next steps

LIMITATIONS

Forecasting = **limited**…

- W/o **understanding why**, what action?
- W/o dealing with **delays**, utility diminished
- W/o **understanding policy** + interventions, how?
- W/o **R(t)**, not clear sense of how to control?

….

NEXT STEPS

- **Visualisations** + extensions (testing)
- New **scoring rules**
- **Real-time**? Data? **Forecast feedback?**
- **Scenario** projections
- Decision-maker **surveys, economic evaluation** + **collaboration**

+ many more ….

Cathal Mills

cathal.mills@stats.ox.ac.uk

Conclusions – so what?

Decisions not based **solely** on forecasts, metric, model ….

Forecasts + evaluation = **part** of **integrated assessment**

Recall Q = What is a "good" forecast?

i. **Translate** popular forecast evaluation **metrics -> actionable quantities**;

ii. **Refocus** predictions + **evaluations on decision-makers -> build trust + better models**;

iii. **Procedures** to **safeguard** forecast-based decision making.

Message = Measure ***value*** of a **forecast** by
ability to **inform decisions (*i.e. by how they're used*)**

Cathal Mills

cathal.mills@stats.ox.ac.uk

Thank you for listening

- Further thank you to **my collaborators**

- **Please reach out** now or later with thoughts, questions, and collaborations**!**

cathal.mills@stats.ox.ac.uk

Some papers

Affiliations: