From metric to action: An evaluation framework to translate infectious disease forecasts into policy decisions

Cathal Mills*[1,2], Nicholas J. Irons[1,3], Joseph L.-H. Tsui[2, 4], Sarah Sparrow[5], Luiz M. Carvalho[6], Adam J. Kucharski[7], Oliver Ratmann[8], Ben Lambert[1,2], Christl A. Donnelly[1,2], and Moritz U. G. Kraemer*[2,4]

[1]Department of Statistics, University of Oxford, Oxford, United Kingdom
[2]Pandemic Sciences Institute, University of Oxford, Oxford, United Kingdom
[3]Leverhulme Centre for Demographic Science, University of Oxford, Oxford, United Kingdom
[4]Department of Biology, University of Oxford, Oxford, United Kingdom
[5]Oxford e-Research Centre, Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[6]School of Applied Mathematics, Getulio Vargas Foundation, Rio de Janeiro, Brazil
[7]Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom
[8]Department of Statistics, Imperial College London, London, United Kingdom

July 20, 2025

* Corresponding Author, cathal.mills@linacre.ox.ac.uk, moritz.kraemer@biology.ox.ac.uk

Abstract

Like weather, energy and other areas of society, forecasts and decisions for infectious disease outbreaks are often critical. The difficult and consequential decisions are made under considerable time and societal pressure and in the face of great uncertainty. Ideally, such decisions can be supported by a probabilistic forecast – the prediction of the future value of an epidemiological quantity, together with uncertainty. Forecasting is challenged by noisy, incomplete, and delayed data alongside non-linear and evolving dynamics. Currently, performance metrics for infectious disease forecasts generally focus on the forecaster's perspective; improving calibration and sharpness of forecasts. There is a lack of principled protocols to measure a forecast's "value" – its ability to provide actionable insights for decision-makers. We develop a framework for epidemic decision-making, focusing on three aspects: i) translating forecasts and popular forecast evaluation metrics into interpretable, public-health-relevant quantities; ii) refocusing evaluations on decision-makers and epidemic events for more actionable outputs; iii) linking predictability of an epidemic to the value of forecasts for decision-making. By melding concepts from weather forecasting, information theory, and decision theory, our framework bridges conceptual gaps between forecasters and decision-makers. We illustrate the framework with an application to forecasts of weekly incident COVID-19 cases, where ensemble models often, but not necessarily, provided the most value for different decision-makers, epidemic events, and across space and time. The evaluation framework connects statistical concepts with decision-theoretic principles to promote more actionable and reliable infectious disease forecasts for enhancing public health responses.

1 Introduction

> "Forecasts possess no intrinsic value. They acquire value through their ability to influence decisions made by users of the forecasts [1]."
>
> *Allan H. Murphy*

Infectious disease forecasting involves predicting future values of epidemic indicators over short- and medium-term horizons. These indicators may be disease-induced cases, hospitalisations, healthcare capacity, or deaths, and are distinguished from *scenario projections* by

being unconditional predictions about the future. Subseasonal or short-term epidemic forecasts can be used to inform outbreak responses, including when and where to deploy resources, implement interventions, and launch public information campaigns. Seasonal (i.e., medium-term) forecasts can be used to quantitatively inform public health planning, as predicting the season's peak (timing and value) and cumulative cases/hospitalisations/deaths can provide timely information about future resource requirements. Depending on the targeted epidemic indicator and actions of decision-makers, the predicted future can be i) modifiable (e.g., COVID-19 cases following vaccination strategies with direct impacts), or ii) largely invariant to interventions (e.g., seasonal dengue cases following expansion of healthcare capacity). A unifying goal in all forecasting tasks is to support the forecast *user* (i.e., a decision-maker) by providing the scientific understanding and uncertainty necessary for making informed decisions.

Although there were pre-COVID-19 pandemic guidelines [2, 3] and forecast competitions for several infectious diseases (e.g., dengue, Ebola and seasonal influenza [4–6]), there has been rapid progress in infectious disease forecasting since the COVID-19 pandemic. Many of the key advances (e.g., the COVID-19 Forecast Hub in the United States [7–9]) were in the evaluation of model performance and the construction of ensembles for improved forecast performance. At the heart of such breakthroughs has often been the use of scoring rules (which measure the whole forecast distribution) and more detailed forecast performance evaluations. Such methods are the gold standard for evaluating the reliability of infectious disease forecasts and nowcasts (e.g., measuring the calibration and sharpness of their uncertainty intervals) [8, 10–17]. Separately, statistical decision theory has been proposed for more principled epidemic decision-making processes [18, 19].

Many advances in infectious disease forecast evaluation have focused on forecast *quality* (alignment of probabilistic forecasts with observations), yet fewer examples link infectious disease forecasts to *value* for decision-makers (i.e., benefit to the forecast user to make decisions). In weather forecasting [20] and in medical decision-making [21], there are clearer guidelines for linking model outputs to the value for the decision-maker. Meanwhile, it remains unclear how to i) systematically evaluate infectious disease forecasts for individual decision-makers under considerable uncertainty and time pressure, and ii) convert the statistical evaluation metrics to timely actions. Although there are differences with infectious disease forecasting (e.g., the better understanding of the physics of weather or the inability to influence dynamics) [22], in weather forecasting, evaluation procedures measure how well a model's forecast informs the user's ability to take action for extreme weather events (e.g., weather warnings for impact mitigation by emergency managers or for travel avoidance by the public) [20, 23].

Relevant questions within the context of infectious disease forecasting remain understudied. Is a model "good" at producing forecasts that inform important epidemic decisions? Can a decision-maker take appropriate action from statistical metrics and the outputs of a set of forecast models? How do forecast model rankings and decision-makers' decisions vary with different risk preferences? However, there are no procedures to determine *a priori* which future forecast model outputs are likely to be reliable under different epidemic regimes. Is it possible to anticipate the future reliability of model forecasts to determine which models to use, additional data to collect, or outputs to report?

We address each of these questions by building on recent advances in infectious disease settings and other research areas such as forecasting theory, information theory, and decision theory. We develop a framework for infectious disease forecasters and forecast evaluators to i) interpret the contributors to the forecast performance of models (Section 2), ii) quantify the forecast value for specific epidemic events and decision-makers (Section 3), and iii) explore how predictability can shape the forecast performance and value of a model (Section 5). The framework simultaneously emphasises statistical rigour and public health utility throughout, while explicitly recognising practical time pressures and uncertainties of epidemic decision-making.

2 Recasting scoring metrics for interpretable epidemic forecasting across space and time

The targeted quantity, model, and evaluation procedure for an infectious disease forecast should depend on the decision-maker and their policy questions and risk preferences. Before the decision-maker can decide whether to take action, both the forecaster and forecast evaluator must understand the overall quality and value of the model's forecasts and their determinants. We will explain ways to determine whether models have stable performance across different statistical properties, decision-makers, and policy questions, and how performance differs across space and time. Rather than interpreting popular scoring metrics as the end goal of evaluation, we focus on the implications and implicit assumptions of metrics for infectious disease forecasts and then discuss how we can extend the application of these concepts and integrate them into an evaluation framework (Figure 1 and Section 6).

2.1 Scoring functions and rules

Probabilistic forecasts of epidemic quantities can be comprehensively evaluated by proper scoring rules. These are loss functions that measure reliability by how well the model's forecasts capture uncertainty (using the whole forecast distribution) and quantify distributional accuracy beyond the alignment of pointwise estimates with observations. A scoring rule is a specific type of scoring function with lower values implying better performance (see Table 1). A scoring rule is (strictly) proper if the score is minimised, in expectation, by (only) the unknown, true probability distribution of the target quantity. Examples include the continuous rank probability score (CRPS) [24] and the Brier score (BS) [25]. For predictive cumulative distribution F and continuous observation y (e.g., incident hospitalisations), the CRPS is defined as:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - 1(x \geq y))^2 dx. \tag{2.1}$$

Probabilistic forecasts are often represented using quantiles of the predictive distribution F [9, 12, 26, 27], which can be scored on test data using the Weighted Interval Score (WIS) [28], a proper scoring rule that approximates the CRPS. Relative scores (e.g., relative WIS, rWIS

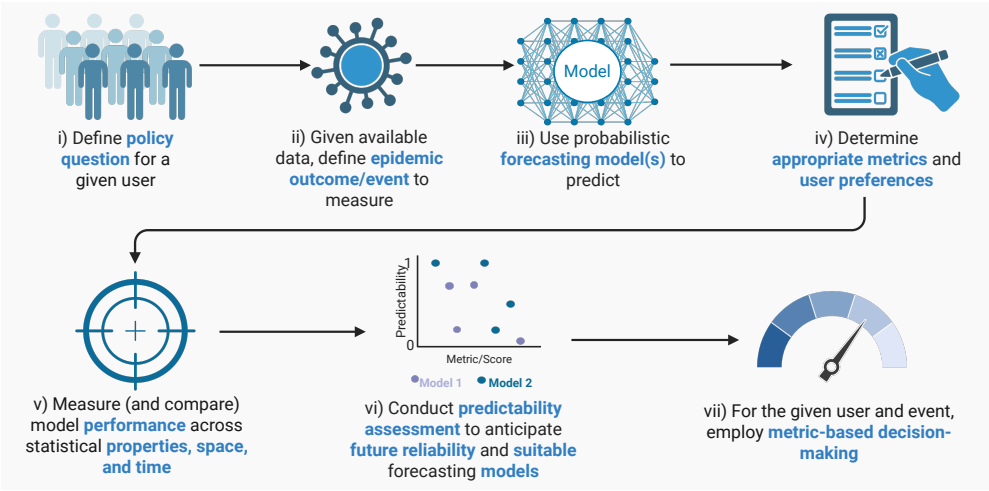

Figure 1: **Evaluation framework and workflow for forecasters and forecast evaluators:** A proposed workflow for our evaluation framework which enables forecasters and evaluators to justify why a decision-maker would use the forecasts of a specific model (or set of models) to inform a public health policy decision. The workflow starts with the forecaster or evaluator focusing on a specific policy question co-defined with a given decision-maker. The workflow is to be applied iteratively at each time point (e.g., for multiple events and metrics), and the metric-based optimal model(s) for each decision will be based on one or more chosen metrics. For determining policy questions and an appropriate metric for a given user and policy question, see Figure 2 and Table 2. We provide details of metric-based and predictability assessments throughout Sections 2–5. See Section 6 for applications of the framework to COVID-19 case forecast evaluation.

[26, 29]), while no longer proper [30], are often used to compare forecast performance against a reference model. WIS and CRPS are our framework's proper scoring rules for measuring overall predictive performance (Figure 2 B) of a model for forecasting the continuous value of a target epidemiological quantity y.

The BS is an *event-specific* (Figure 2 B) classification measure that assumes a uniform distribution of decision-maker risk preferences [25]. BS is defined for a binary outcome $y \in \{0, 1\}$ and corresponding probabilistic forecast $p \in [0, 1]$ as:

$$\mathrm{BS}(p, y) = (p - y)^2 \tag{2.2}$$

where $y = 1$ indicates an event occurring (e.g., incident cases above a threshold θ). The CRPS is the BS integrated over all possible thresholds (θ) for the binary event.

Pinball loss (or quantile score) assesses forecasts for a fixed, *user-specific* (Figure 2 B) quantile level τ:

$$\mathrm{PL}_\tau(x, y) = (\mathbb{1}\{y \leq x\} - \tau)(x - y) = \begin{cases} (1-\tau)(x-y), & y \leq x, \\ \tau(y-x), & y \geq x, \end{cases} \tag{2.3}$$

where x is the $\tau-$quantile of a predictive distribution F. Pinball loss is a consistent scoring function (see Table 1), as opposed to a scoring rule, since it assigns a score to a given quantile (τ) and not the entire forecast distribution. The CRPS is the pinball loss averaged over all quantile levels.

For any scoring function or rule, the mean score across N test set observations is often used as a summary score to rank the out-of-sample abilities of different forecasting models:

$$\overline{S} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{2.4}$$

where S_i is the score obtained for observation i, calculated using one of eqs (2.1)–(2.3).

2.2 Stratifications and interpretable scoring metrics

Alongside difficulties in interpretability and actionability for decision-makers, which we discuss below, summary scores can obscure i) variations in forecast performance under different epidemic conditions (e.g., temporal [15], feature [9], or spatial [12] dependencies) and ii) variations in the different statistical components of forecast performance (e.g., calibration versus discrimination abilities). Epidemics unfold across space and time, thereby exhibiting distinct spatial and temporal signatures. In our framework, as in previous studies [9, 12, 26, 33–35], all evaluation metrics should be applied to all available test data, alongside stratifications – trailing windows/subsets of data, different geographies, and spatial scales (see applications in Section 6).

Table 1: **Key concepts for our framework of probabilistic infectious disease forecasting:** We define important concepts discussed throughout the text and used in our framework for forecasting infectious diseases. Given the wealth of metrics and tools above, the choice of metric/tool may not be immediately obvious, and may be both user- and event-specific, which motivates our workflow (Figure 1 A) starting with the well-defined policy question(s) for a user/decision-maker and corresponding epidemic outcome to measure. See Table SI 1 for definitions of additional concepts.

Concept	Meaning
Brier Score (BS)	A strictly proper scoring rule defined as the mean squared error for a probabilistic forecast of a binary outcome [25].
Classifiers	A probabilistic classifier produces a probability forecast for a binary outcome.
Consistent	A scoring function is consistent for a given functional $T(P)$ (e.g., quantile level) if the score is minimised when the forecast aligns with $T(P)$ consistent.
Continuous Ranked Probability Score (CRPS)	A proper scoring rule that is strict when the first moment of the forecast distribution is finite, and generalises mean absolute error for probabilistic forecasts of continuous outcomes [24].
Cost-loss (C/L) ratio	In this decision model for binary outcome events, C/L ratio α is the cost (C) of preparation for an adverse event, divided by the avoidable loss (L) if the event occurs without any preparation [1, 31]. The optimal decision is to act/prepare if the forecasted probability exceeds this *decision threshold* C/L. Knowing plausible values of the ratio for a given binary event is sufficient for studying forecast model results, and thus for metric-based decision-making.
Discrimination (DSC)	A non-negative component of the decomposition for the mean value of a score, and measures the ability of a model's forecasts to distinguish events versus non-events [32]. Higher values indicate superior discrimination ability.
Event threshold	A cut-off value θ that is used for defining whether a binary outcome occurs or not.
Forecast/predictive distribution	A probabilistic forecast's distribution for the future value of a given quantity, often represented by a parametric distribution, samples from the distribution, and/or quantiles of the distribution.
Forecast horizon	The lead time at which forecasts are made.
Forecast quality	Measure of the alignment of a model's forecasted values with observed values.
Forecast value	Measure of the ability of a model's forecasted values to inform decision-making.
Horizon	The lead time at which a forecast for a future value is produced.
Miscalibration (MSC)	A non-negative component of the decomposition for the mean value of a score, and measures the calibration of a model's forecasts [32]. Lower values indicate superior calibration.
Murphy curve/diagram	For a given model, a curve of mean elementary scores (eq. (2.6)) across different C/L ratios (α) or event thresholds (θ). Lower curves indicate better performance, and a Murphy diagram compares Murphy curves for different models.
Pinball loss	Measure of forecast performance for a specific user-defined probability/quantile level. Pinball loss is also known as Quantile Score (QS) or the asymmetric piecewise linear scoring function.
Permutation entropy (PE)	Measure of predictability which works by studying temporal dependencies in both linear and nonlinear time series. PE is a model-free measure that uses Shannon's Entropy to study the frequency of distinct patterns in time series data.
Predictability	The inherent randomness of a system.
Proper	A scoring rule is proper if the expected score is minimised by using the true probability distribution.
Scoring function	A function that assigns a score to a given functional (e.g., quantile level). This is not necessarily proper, and can asymmetrically penalise probabilistic forecasts (e.g., pinball loss). Lower values for a scoring function imply better predictive performance.
Scoring rule	A scoring rule (e.g., CRPS) is a specific type of scoring function (see above) that assigns a score to a probabilistic forecast, and lower values for a scoring rule imply better predictive performance.

This will help i) identify the time- and space-varying differences in model performance (e.g., in the presence of data distribution shifts), and ii) ensure that model development, model comparison, and decision-making procedures reflect the relevant local context: the decision-maker(s), stage of epidemic, and region(s) in question.

Identifying a "best" forecasting model will always be a subjective decision, determined by the policy question (Figure 2 A) of a local decision-maker and the available data, which then guides the choice of evaluation metric and model (Figure 1). Even when appropriate metrics have been chosen (see Table 2) and applied carefully to past forecasts across time and space, it is uncovering the determinants of a model's statistical strengths and weaknesses is important for model improvements, enhanced trust, and tackling uncertain future conditions. The statistical properties are the drivers of forecast reliability for public health policy. For example, a naïve, constant model for short-term forecasts could be perfectly calibrated, yet have little discrimination ability for major surges in cases. Another model could produce precise forecasts, yet be over-confident and have poor calibration.

While rarely used in infectious disease forecasting, scoring function *decompositions* can shed light on the components of forecast reliability [32, 36–41]. There are various ways to decompose $\overline{S}$, and we use CORP (Consistent, Optimally binned, Reproducible, and Pool-adjacent-violators)-algorithm-based decompositions [32, 41, 42]. These methods use nonparametric isotonic regression to decompose $\overline{S}$ for any consistent scoring function (eq. (2.4) for e.g., CRPS) into three interpretable components:

$$\overline{S} = \text{MCB} - \text{DSC} + \text{UNC}, \tag{2.5}$$

where non-negative quantities MCB, DSC, and UNC denote miscalibration, discrimination, and uncertainty. Lower values of MCB indicate superior calibration (i.e., forecasted probabilities align with observed frequencies), while higher values of DSC indicate superior discrimination ability (mean score of the model relative to a recalibrated but constant forecast). Calibration is a measure of forecast reliability, while discrimination ability is a measure of how well a model can distinguish different epidemic events ahead of time. The uncertainty component measures the irreducible variability of the data.

The CORP decomposition (see [32, 41, 42] for further details) relies on the availability of sufficient test data to perform a forecast recalibration. This decomposition may have limited use in early outbreak stages and in short windows of epidemic data (when tracking temporal variations in forecast performance) due to the scarcity of *relevant* data. Future applications of our framework could consider using other existing decomposition methods (e.g., [43] or [39]) or developing new decomposition

techniques (e.g., using conformal inference [44–47] in the presence of data distribution shifts).

The CORP decomposition for a scoring function estimates the drivers of statistical performance (measured by the score) for a forecasting model, which are also public-health-relevant properties. Strong discrimination ability is crucial for detecting surges and declines in epidemic curves, while probabilistic forecasts must be well-calibrated to maintain reliable estimates of the uncertain future in epidemics.

In our framework, for scoring functions, we visualise CORP-based discrimination abilities alongside calibration properties (Figure 2 B), where DSC (y-axis) is plotted against MCB (x-axis), with points indicating models and lines indicating level curves of constant forecast performance (as measured by the scoring function). We encourage also plotting DSC and MCB for operational baseline/reference models (e.g., a seasonal average model for an endemic disease). Diagnosing any discrimination or calibration issues across space and time, or stratified by features, can provide insights for forecasters about where additional data or bias correction may be required and for evaluators about ensemble model weights and likely suitable models for different policy questions.

2.3 Elementary scores and utility functions

Despite the central role of cost-benefit analysis in public health planning, it is not often recognised in infectious disease forecasting that commonly used summary scores are intricately linked to cost-loss (C/L) ratios (Figure 2 B). These scores are the output of *weighting* an *elementary score* by a utility function for the user's preferences (e.g., risk aversion via their C/L ratio) for different event thresholds (e.g., epidemic events) [48, 49].

The error scoring matrix is defined as:

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

C will often represent economic costs for a decision-maker taking a specific preventive action (e.g., increasing intensive care unit (ICU) capacity), and the preventable loss L will be the *additional* burden and public health impact (e.g., overwhelmed hospital losses, or the more costly loss of lives) if the pre-specified event (e.g., severe outbreak) occurs without taking action. We can also view C as the economic *harm* of taking action and L as the maximum *benefit* of taking action.

The *elementary score* [48] is defined for an event threshold $\theta \in \mathbb{R}$ (binary event occurs if $y > \theta$) and C/L ratio α (the ratio of the cost of taking action versus the preventable loss if no action taken):

$$s_{\alpha,\theta}(x, y) = \begin{cases} 1-\alpha, & \text{if} \quad x \leq \theta < y \\ \alpha, & \text{if} \quad x > \theta \geq y \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

where $x \in \mathbb{R}$ is the $(1-\alpha)-$quantile forecast, $x := F^{-1}(1-\alpha)$.

Under the C/L decision model, $\alpha = C/L$ is the optimal probability threshold above which the decision-maker takes action. Small C/L values for an individual decision-maker indicate inexpensive risk mitigation and large potential losses for the event, and the optimal action for this risk-intolerant decision-maker will be to act even when the forecasted event probability is low.

A C/L ratio can have very different interpretations, depending on how costs and losses are defined (see Section 2.4). In general, if $\alpha < 0.5$, the economic costs incurred by an underprediction and lack of action (e.g., too few ICU beds) may be low, yet loss of life could be high as this loss is valued as much more costly than the cost of an ICU bed. Eq. (2.6) would penalise this underprediction with a higher elementary score. If $\alpha > 0.5$, the economic costs incurred by an overprediction and prompted action (e.g., too many ICU beds) may be high, relative to the defined losses incurred.

A *summary score* is generally defined by weighting the elementary score (eq. (2.6)) with a *utility* function $u(\alpha, \theta)$ [48, 49].

$$S(F, y, u) = \int_{-\infty}^{+\infty} \int_{0}^{1} s_{\alpha,\theta}(x, y) u(\alpha, \theta) \mathrm{d}\alpha \, \mathrm{d}\theta, \tag{2.7}$$

where $x \in \mathbb{R}$ is the forecast at the $(1-\alpha)-$quantile level $x := F^{-1}(1-\alpha)$, $u(\alpha, \theta)$ determines how the forecast user weights different combinations of C/L ratios (i.e., probability decision thresholds, α) and epidemic event thresholds (θ).

Eq. (2.7) establishes a key result for both forecasters and evaluators: any summary score must make an assumption about the decision-maker's utility function $u(\alpha, \theta)$. These assumed utility functions each differentially weight, simultaneously, the decision-maker's risk aversion α and epidemic event severity θ. For example, the CRPS (eq. (2.1)) is defined by uniformly weighting all probability decision thresholds α and considering all epidemic event thresholds θ. This evaluation stretches across the whole probability space of decision-makers and events (Figure 2 B), and CRPS produces the simplest single measure of forecast value for the entire uniform family of decision-maker risk preferences for all epidemic events [50]. While the CRPS

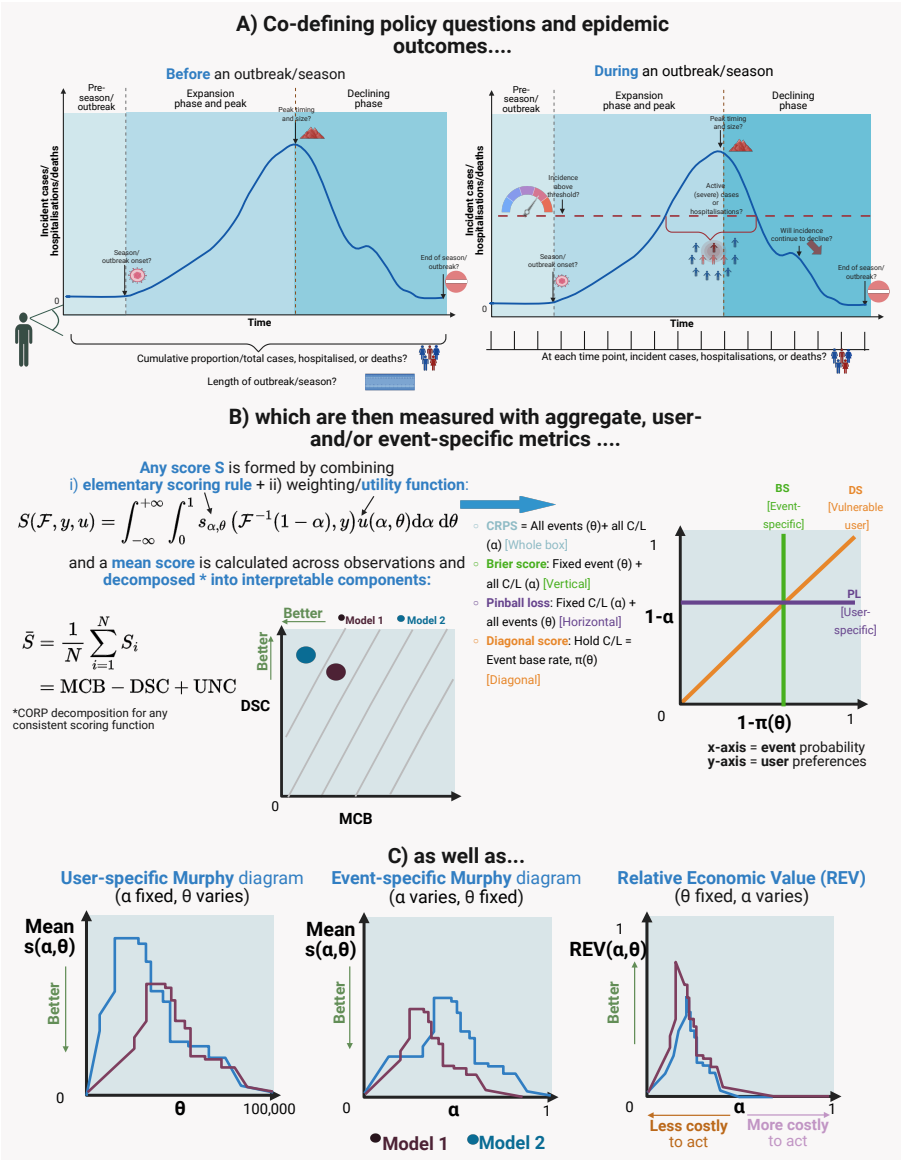

Figure 2: **Policy questions and metrics for different users and events: A)** Illustration of generic epidemic curves to illustrate policy questions, either before or during an outbreak/season. All forecast evaluation in our framework starts from these defined policy questions. **B)** Popular summary scores (e.g., CRPS) are formed by combining an elementary score and utility function, and each score makes implicit assumptions about i) the epidemic event which is captured by event threshold θ, and ii) the decision-maker whose preferences for the event are expressed by the C/L ratio α. These scores can be further decomposed into the statistical drivers of forecast performance for enhanced interpretation (see main text). **C)** User- and event-specific evaluation tools – Murphy diagrams and Relative Economic Value curves where lower and higher curves indicate better forecast performance respectively.

provides this rigorous assessment across decision-makers and epidemic events, such implicit assumptions may be less suitable for specific decision-makers and in a fast-evolving epidemic.

To illustrate Eq. (2.7), suppose an individual is considering a probabilistic weather forecast (the model) of daily rainfall exceeding a defined level (the event threshold, θ). Their decision problem is whether to bring an umbrella (the action) based on combining the forecast with their risk preferences (α) about the rainfall event threshold (θ). The value of a forecast will clearly depend on the model *and* how the individual defines both the event severity and their corresponding risk preferences. To measure this forecast value, different functions $u(\alpha, \theta)$ will apply different penalties to the elementary score for individual combinations of (α, θ).

2.4 Eliciting risk preferences for decision-makers

The effects of infectious disease outbreaks are wide-ranging. Decisions supported by forecasts–such as whether to enact additional vector control strategies or increase ICU bed capacity–can have profound health, social, and economic impacts. Forecasters and evaluators can account for these various factors via $u(\alpha, \theta)$, yet eliciting these utility functions for individual decision-makers is non-trivial. In the current work, we focus on how C/L ratios (α) could be defined for different event thresholds (θ), as a necessary first step for future input to custom functional forms for $u(\alpha, \theta)$. See [51] for a detailed exploration of how different proper scoring metrics apply different asymmetric penalties.

α and θ must be defined simultaneously and both require real-time updating. For example, suppose that a local health authority (the decision-maker) has determined that to protect against the loss of life from severe disease, it is necessary to increase/reallocate another 10 ICU beds to a district if more than 1,000 (θ) infections occur (the event) in a given week. ICU beds are made available at an economic cost if a chosen model (see below) forecasts infections exceeding the threshold in the next week with probability $\geq$ C/L (α). The economic costs of action (C) could depend not only on the additional beds themselves but also the probable length of patient stays, additional medical and staffing costs, and possibly indirect costs if beds are relocated from elsewhere. Meanwhile, if the event occurs without preventive action, preventable losses (L) could entail life costs from direct impacts of severe disease, comorbidity, and macroeconomic impacts (e.g., greater workforce absenteeism by patients and relatives). Regarding life costs, the standard approach in the United States to quantify mortality risk reductions in monetary terms is the value of a statistical life, commonly estimated at about \$10 million in USD2017 [52], while in the United Kingdom the National Institute for Health and Care Excellence cost-effectiveness threshold for medical interventions is set to £30,000 per quality adjusted life year [53]. Quantitative estimates of costs for different actions and losses from epidemic impacts may also be available (e.g., for COVID-19 [54–56] or dengue [57, 58]) which can be used to calibrate C and L.

There will inevitably be residual uncertainty, ethical judgements, and legal implications. Many costs are unforeseen or difficult to quantify, yet forecast-informed public health decisions are high-stakes. In the absence of reliable estimates for all relevant quantities, forecasters and evaluators can perform sensitivity analyses and surveys (see Discussion and Section 4) to consider reasonable ranges of their values, while recognising that decision-makers may prioritise different costs depending on their preferences (e.g., prioritising impacts from loss of life by down-weighting possible contributions to C related to operational costs).

In general, C/L ratios for different epidemic event thresholds will vary i) by pathogen and its transmission history (e.g., immune landscape), and ii) by decision-maker and available resources (e.g., a higher C/L ratio in a resource-limited setting requiring greater certainty to minimise false alarms). These will also vary across space and time as epidemics and their impacts evolve alongside economic and population health. By evaluating the value of forecasts over a range of C/L ratios and event thresholds, determined by variation in costs arising both from uncertainty and differing decision-maker priorities and risk preferences, forecasters and evaluators can explore the Pareto frontier of forecasts that are performant simultaneously relative to multiple objectives. Finally, we emphasise that by evaluation only depending on the C/L ratio, it is only the relative cost of mitigation that is required for forecast evaluation.

3 Forecast value for informing epidemic decision-making

3.1 User-and-event-focused scoring metrics

When confronted with the introduction of a new pathogen, the coming season of an endemic pathogen, or an apparent rise or decline in an epidemic trajectory, decision-makers are faced with difficult and specific decisions (Figure 2 A), for which a forecasting model's *value* may vary. These decisions require forecasters and forecast evaluators to focus model development, model selection, and forecast evaluation on specific epidemic events, specific parts of the forecast distribution, and a specific decision-maker's risk preferences. Decisions are generally about actions drawn from a wider *action set*, and may be short-term interventions (e.g., public information campaigns) or longer-term plans (e.g., vaccination strategies).

Table 2: **Metrics of our framework:** Details of different metrics for the quality and value of probabilistic forecasts, including the information needed to compute metrics and relative advantages and disadvantages of each metric.

Metrics	Data/User Requirements	Advantages	Disadvantages
Relative Economic Value (REV)	Rough sense of C/L ratio and defined event threshold	Interpretability, quantifies the relative/operational value of model, and event-specific measure	Focus on fixed event, need for sense of C/L ratios, and does not use elementary scores
Brier Score (BS)	Defined event threshold (θ)	Event-specific, proper scoring rule, and can be decomposed [32, 59] into meaningful statistical properties	Focus on fixed event and uniformly across all user risk preferences (i.e., all C/L ratios)
Pinball loss/Quantile Score	Fixed risk preferences (C/L ratio)	User-specific evaluation measure, consistent scoring function, and can also be decomposed into meaningful statistical properties	Does not measure whole predictive distribution and assumes fixed risk preferences across event thresholds
Murphy Diagram	Defined event threshold or risk preferences	User- *or* event-specific, interpretable visualisations, and can trace out forecast value under uncertain event thresholds (θ) and/or different risk preferences (α)	More formal hypothesis testing may be required and crowded visualisations
Continuous Ranked Probability Score (CRPS)	Sample or full probability distribution representation of forecasts	Proper scoring rule measuring whole predictive distribution, simplest measure of forecast value for uniform family of user risk preferences, and can be decomposed into interpretable components [41]	Neither event- nor user-specific and computationally expensive (for sample representation)
Weighted Interval Score (WIS)	Forecasts represented as quantiles	Proper scoring rule measuring the whole predictive distribution, converges to CRPS, and computationally cheaper to handle quantiles (instead of samples) of a predictive distribution	Neither event- nor user-specific and limitations (e.g., distribution tails) of quantile representation

Viewing each decision problem as a *binary* one (action or no action), the *value* of a model's probabilistic forecasts (e.g., of incident cases) is determined by its ability to inform the binary decision problem(s) for different decision-makers, given their corresponding utility functions $u(\alpha, \theta)$. Therefore, the *value* of a forecast depends on both the forecasting model and the decision-maker's utility function ($u(\alpha, \theta)$) for their policy question. This framing does not mean that a forecast is the sole basis for complex epidemic decision-making, but rather that the value of a forecast must be determined by its ability to inform a decision-maker.

3.1.1 Relative Economic Value

An approach in our framework to measure forecast value is Relative Economic Value (REV) [1, 31]. For an event threshold (θ) and a decision-maker's corresponding C/L ratio (α), REV measures the value of a model's forecasts *relative* to a baseline/reference model's forecasts and a perfect model's forecasts. REV is a decision-theoretic framework often used in weather forecasting (e.g., [31, 50, 60]). It relies on knowing a plausible range of a decision-maker's C/L ratios (α) for the defined epidemic event threshold (θ).

For an event threshold θ, REV starts by converting the probabilistic forecast x_i into a classification of the binary event $y_i \in \{0, 1\}$ based on a given rule. Under the C/L framework, a forecasting model predicts occurrence of the binary event if the probability x_i exceeds the C/L ratio α. REV then uses the expected expense (E) that would be incurred by using the forecasting model's classifications for the binary event. This expected expense is calculated by combining the correct and incorrect classifications with corresponding costs and loss incurred, yet the calculation of REV relies only on knowing the ratio of costs and losses (see Supplementary Material B):

$$\mathrm{REV}(\alpha = C/L, \theta) := \frac{E_{\text{baseline}} - E_{\text{model}}}{E_{\text{baseline}} - E_{\text{perfect}}}. \tag{3.1}$$

Bounded between 0 (for the baseline model) and 1 (for the perfect model), REV is computed for different C/L ratios (α) while holding the event threshold (θ) constant. REV is then shown as a curve for different decision-maker risk preferences (α), with higher values indicating superior relative performance for a given risk preference and event threshold θ (Figure 2 C). We use REV in our framework as an interpretable, event-focused measure of a model's operational/relative value over a baseline model for different risk preferences (Table 2).

3.1.2 User- or event-specific Murphy diagrams

A Murphy diagram [48] provides user- or event-specific measures of forecast value (Figure 2 C, Table 2), and extends beyond the REV by capturing the forecast value *and/or* overall predictive performance, using a proper scoring rule (BS) or consistent scoring function (pinball loss) without the need for a baseline model.

First, as a *user-specific* evaluation tool, the Murphy curve (Figure 2 C) visualises, for a fixed C/L ratio α, the mean elementary score (y-axis) for a given forecasting model against different epidemic event thresholds θ (x-axis) across *all* N test set observations:

$$\overline{\mathrm{S}_\alpha}(\theta) := \frac{1}{N} \sum_{i=1}^{N} S_\alpha(x_i, y_i) \tag{3.2}$$

where for observation i, $x_i \in \mathbb{R}$ is the model's forecast at quantile level $1 - \alpha$, $y_i \in \mathbb{R}$ is the observed outcome, and S_α is the corresponding elementary score (eq. (2.6)) for the fixed C/L ratio α and a specific event threshold θ. In our framework, we include Murphy diagrams, a collection of Murphy curves that trace out, for a fixed risk aversion preference (α) of a decision-maker,

how the value of different models' forecasts varies across different magnitudes of epidemic events (θ). Lower curves for a given model are preferable, and if curves for different models cross, there may be no model dominance [42, 48]. The area under each Murphy curve is the mean pinball loss for that model. This visualisation improves on BS as it measures value for a specific decision-maker's risk preferences and different epidemic event thresholds, which may be advantageous when epidemic events are difficult to *a priori* define.

Alternatively, a Murphy curve can be an *event-specific* evaluation tool (Figure 2 C), holding the epidemic event threshold θ fixed and visualising a model's mean elementary score $\overline{\mathrm{S}_\theta}(\alpha)$ (y-axis) against different C/L ratios α (x-axis):

$$\overline{\mathrm{S}_\theta}(\alpha) := \frac{1}{N}\sum_{i=1}^{N} S_\theta(x_i, y_i), \tag{3.3}$$

where S_θ is the elementary score (eq. (2.6)) for the fixed event threshold θ and a specific C/L ratio α. The area under such a Murphy curve is the mean BS, and a Murphy diagram again visualises Murphy curves for different models. This diagram traces out, for a fixed epidemic event severity (θ), how the forecast value of different models varies with different risk preferences (α) for the decision-maker.

Caution is necessary, as: a user-specific Murphy diagram does not measure the whole predictive distribution and scores could be dominated by tail behaviour, while the event-specific Murphy diagram focuses on a single event threshold; neither diagram segments forecast performance into discrimination or calibration; varying the fixed event thresholds and/or decision-maker preferences is important; and formal inference and hypothesis tests may be required.

3.1.3 Other measures of forecast value

We briefly highlight other metrics that focus forecast evaluation on specific decision-makers and their priorities. Originating from weather forecasting [49], the diagonal score (Figure 2 B) assumes a fixed relationship between C/L α and the empirical probability of epidemic events $\pi(\theta)$ (i.e., how often different events occur). This makes it suitable for risk-intolerant decision-makers, as if the cost of action (C) remains fixed across event thresholds, rare extreme large event thresholds are associated with large losses (L).

Traditional classification tools include Receiver Operating Characteristic (ROC) curves and the Area Under the ROC Curve (AUC), as implemented in [33, 61, 62]. As ROC curves can fail to satisfy concavity requirements and AUC ignores miscalibration [59, 63], we recommend the use of DSC-MCB properties in the CORP-based decompositions (e.g., of BS) and/or concave ROC curves that use the recalibrated forecasts of the CORP decomposition [59].

4 Risk-averse epidemic decision-making and distributionally robust optimisation

Epidemic decision-making is high-stakes; decision-makers may prioritise risk-averse strategies that avoid the worst outcomes and remain robust to the considerable uncertainty in the evolving dynamics of outbreaks. We described above uncertainty in a decision-maker's C/L ratios which could affect forecast value, model selection, and the metric-implied optimal decision. Meanwhile, observed data during an epidemic are imperfect, impacted by time- and space-varying issues such as reporting errors, delayed reporting, and underrepresentation. These issues, along with fast-evolving epidemic dynamics, can induce data distribution shifts and lead to metric-implied optimal decisions – from results for past test data, even if stratified by space and time (Section 2.2) – no longer being optimal when making the decision in real time on newly available data.

While the scoring metrics introduced above (Table 2) provide a general and flexible framework for evaluating forecasts, they are clearly limited and do not cover all relevant risk measures. We draw attention to distributionally robust optimisation (DRO) as an alternative or complementary setup for decision-making under uncertainty [64, 65]. *Expected tail loss* (ETL) is a quantity central to the DRO literature (see [64, 65] and Supplementary Material Section G for formulae and descriptions) that measures the risk associated with extreme events occurring at the upper tail of the loss distribution, defined as the average loss incurred among the worst α% of outcomes Y (e.g., hospitalisations). At many points during the COVID-19 pandemic, discussion and messaging surrounding public health measures implicitly invoked ETL as a focal risk measure. Emphasis on early lockdowns in China without complete knowledge of disease severity, for example, prioritised controlling tail risk. ETL and some other risk measures T of interest (e.g., the variance and mode of Y) are not *elicitable* from any proper scoring rule S, in the sense that there is no S for which predictions issued on the basis of T that uniquely minimize the score [66–68]. Therefore, evaluating forecasts based on S would yield conclusions inconsistent with the choice of T as a risk measure.

The distinction between scoring rules and risk measures discussed above points to the gap between the stochastic optimisation framework—within which scoring rules are situated—and DRO, respectively [65]. Here we draw attention to DRO as a body of work that may prove salient to the mitigation of future outbreaks. Although it has yet to take hold in the infectious disease

literature, we view DRO as an important future tool to help address the limitations of our framework and the impacts of uncertain decision-making.

5 Predictability assessments for epidemic forecasting

Predictability analyses are another way to protect against the impacts of uncertainties, tail risks, and temporal dependencies (in epidemic data and forecast value). Unlike some natural systems, predictability for an observed epidemic is not just a property of nature, as i) there can be a modifiable dynamical system with no governing laws, and ii) observed data are highly imperfect measures of the true nonlinear epidemic dynamics (e.g., incident infections). This raises the question of predictability. There is no universal definition of predictability, and we use a broad definition of predictability as the inherent randomness of the epidemic time series data (see below) that induces fundamental limits in the ability to predict the future values of the time series and therefore limits to forecast performance.

5.1 Introducing predictability for infectious disease forecasting

In our framework, we have described various metrics (see above and Table 2) to measure forecast quality and value. Forecast *quality* (or distributional accuracy) should be bounded above by the predictability of the epidemic data, yet how does the *value* of a forecasting model for informing decision-making relate to epidemic predictability? This is an open question that we begin exploring here.

Arguably, a "good" forecasting model could be highly valuable for decision-making during epidemic stages of rapid change when there may be high unpredictability and a model could have *value* over a reference/naive model (e.g., seasonal average). The association between predictability and forecast value will depend on how a forecaster or evaluator defines both predictability and forecast value. Predictability is measured as a magnitude-free property of the underlying dynamical system. In contrast, forecast value is measured for an individual decision-maker and/or epidemic event, thereby depending on both the magnitude of the observed data and the set of models. Thus, as any measure of predictability is independent of the decision-maker, the models, and the magnitude of the data, a forecaster or evaluator should not expect predictability to have a one-to-one relationship with any measure of forecast value.

However, the forecaster, evaluator, and decision-maker need to know which outputs can be reliably reported and which forecasting models to use for a decision. Procedures are needed to evaluate how predictability has shaped forecast quality and value in the past and how to anticipate future epidemic predictability and its likely impacts on the utility of a forecasting model for informing decision-making. In our framework, we hypothesise that i) assessing past predictability of the epidemic time series will help to determine whether past model forecast performance was largely caused by the model and/or the (un)predictability of the epidemic, and ii) analysing past predictability can inform the likely future predictability and forecast performance of different models.

5.2 Permutation entropy

For estimating predictability, we follow [69] and use permutation entropy (PE), a model-free measure of temporal dependence in potentially nonlinear time series data. Entropy summarises uncertainty in a system. Shannon's entropy [70], for a random variable $X \in \mathcal{X}$, where $\mathcal{X}$ is the sample sample, is defined as $H(X) = -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x)$.

For time series $\boldsymbol{X} = \{x_t : t = 1 \dots N\}$, PE works by first defining: i) the embedding dimension d which determines the segment length for measuring randomness/patterns in the data and ii) the time delay τ which determines the number of time steps between the compared measurements (see below). These are used to construct vectors: $\boldsymbol{x}^d_{t,\tau} = \left(x_t, x_{t+\tau}, \dots, x_{t+(d-1)\tau}\right)$. We fix $\tau = 1$ to focus on successive values of the time series.

Within each vector $\boldsymbol{x}^d_{t,\tau}$, entries of this segment of length d are then (totally) ordered by comparing their relative magnitudes. This produces an ordinal pattern π. Normalised PE is then defined by using the Shannon entropy on these patterns and normalising using the maximum entropy $\log(d!)$:

$$H_{d,\tau}(\boldsymbol{X}) = -\sum_{\pi} \frac{p_\pi \log(p_\pi)}{\log(d!)} \tag{5.1}$$

where p_π is the empirical probability of a specific ordinal pattern π, based on the frequency at which each pattern occurred.

To promote conservative estimates of PE, we again follow [69], and define predictability as:

$$\chi = 1 - min_{d,\tau=1} \left(H_{d,\tau}(\boldsymbol{X})\right), \tag{5.2}$$

Table 3: **Submitted model names and descriptions:** The following table includes our abbreviation, the original model name from each modelling team's submission, and the corresponding model description provided in [9].

Model name	Original name	Description
Ensemble	COVIDhub-4_week_ensemble	Unweighted average or median of submitted forecasts to the COVID-19 Forecast Hub
Baseline	COVIDhub-baseline	Median prediction at all future horizons is equal to the most recent observed incidence
CU-Select	CU-Select	A metapopulation county-level SEIR model for projecting future COVID-19 incidence cases and deaths
Karlen-pypm	Karlen-pypm	Finite time difference equations implemented as a general-purpose population modelling framework
CA-DELPHI	COVIDAnalytics-DELPHI	SEIR model augmented with underdetection and interventions
RW-ESG	RobertWalraven-ESG	Multiple skewed Gaussian mathematical fit

where the minimisation searches over values from $d = 2$ to a user-defined maximum dimension (see [69] for more details). Using eq. (5.2), we can compare epidemic predictability (χ) across space and time using trailing windows of length N and region-specific time series data, noting that i) the window length N, over which PE is calculated, can be varied, but must be greater than $d!$, and ii) PE (and therefore χ) is global-magnitude-invariant as it compares relative magnitudes of observations. Values of χ vary between 0 (low predictability) and 1 (high predictability).

5.3 Embedding predictability in our framework

In predictability assessments of our framework (Figure 1), a forecaster or evaluator can compare the values of χ to results in trailing windows for the different metrics of Sections 2–3. Although low predictability *may* suggest low forecast quality, i) high predictability does not guarantee high forecast quality or value and ii) low predictability does not guarantee low forecast value. Understanding which models provide the most relative improvements in forecast quality or value over baseline/reference models during different regimes of predictability is one way to link predictability to forecast quality and value. Then, an estimated recent or future low predictability will have setting-specific implications – this may provide caution *or* optimism about using, for decision-making, a given model's forecast of the future.

Given that i) an operational forecast is often required to support decision-makers irrespective of how predictable an outbreak is, and ii) data distribution shifts are pervasive in epidemics, these predictability analyses after applications of metrics could be a useful sense-check and additional layer to model evaluation. These analyses could help quantify the practical or operational value at different stages of the epidemic and protect against future erosion of public or decision-maker trust. For example, PE values could indicate regime changes, which could inform model selection or data collection. PE could also be used by forecasters as an input to a forecasting model or as an indicator of a future change in other epidemic indicators (e.g., in the time-varying reproduction number $R(t)$).

5.4 Extending predictability analyses

A key limitation of using eq. (5.2) is its univariate nature. An epidemic may be more predictable conditional on knowing other information, often provided by heterogeneous data streams. To quantify historical drivers of (un)predictability, we propose analyses across space and time. For example, χ could be estimated alongside $R(t)$ to understand the association between predictability and stages of epidemic growth. By further associating results with those for the forecast metrics of the previous sections, these retrospective analyses can enhance our understanding of the reasons behind historical forecast performance, and this understanding can then be used for anticipating future predictability and future performance.

6 Applying metrics to epidemic forecasts

6.1 Application of our framework

We have introduced a conceptual framework for which the workflow (Figure 1) could be as follows. Given the available data, the first step for a forecaster or evaluator is to: i) co-define a relevant policy question with a local decision-maker, ii) identify a corresponding epidemic outcome to measure (e.g., incident cases or healthcare capacity), iii) produce or collect model-based probabilistic forecasts, iv) identify appropriate metrics and a decision-maker's preferences, vi) measure forecast performance for each model across statistical properties, over time, and across space, and vii) assess historical predictability of epidemic dynamics alongside past forecast performance to anticipate future reliability and model suitability, and viii) use metric-based results to choose the "optimal" model's forecasts for future decisions where optimality is only with respect to a given user and event. As

decisions will not be made on the basis of a single question or epidemic outcome, this evaluation workflow should be repeated multiple times at each time point, using different policy questions and different outcomes and metrics targeting the same policy question.

We illustrate the framework with a brief application (Table SI 2 for EPIFORGE 2020 [2] checklist items), using state-level forecasts of weekly incident cases from the COVID-19 Forecast Hub (see [7, 9, 26] for details).

Our objective was not to repeat previous analyses, but rather to highlight i) how different metrics capture forecast value for informing simple policy questions, and ii) the implications of the statistical theory for different decision-makers and epidemic stages. There will be no one-size-fits-all approach for forecasting, evaluation, or decision-making, and any application should make clear the subjective choices made. Setting-specific considerations for the decision-maker include relevant policy questions, intervention action sets, costs and resources, and likely responses from the public. For the forecaster and evaluator, considerations for model development and selection include evaluation window sizes and spatial scales, both of which depend on practical data availability and requirements of public health authorities.

We studied the time period of 1 August 2020 – 15 January 2022 inclusive and restricted our focus to six models (Table 3) that submitted forecasts for all 73 weeks. We used the COVID-19 Forecast Hub baseline model as a reference model, which forecasts the most recently observed incidence as its median estimate with uncertainty based on historical differences between forecasted and observed values [26].

We were limited by the probabilistic forecasts being raw case counts and summarised using seven quantiles, preventing the use of recommended transformed scales for evaluation [13] and evaluation of the full forecast distribution (and all C/L ratios). We adopted a 12-week trailing window to examine predictability of the epidemic and time-varying performance of models. This enabled a realistic application, as a forecaster or evaluator would likely not want ensemble weights or models selected to change too rapidly, as there is already enough weight and model uncertainty in fast-evolving epidemics. To measure overall forecast quality, we used the WIS [29] alongside relative WIS (rWIS) – a scaled version of WIS that measures relative performance compared to the baseline model and accounts for potentially different numbers of forecasts for different time points and locations. We used the same weekly estimates as [9] for the instantaneous reproduction number, $R(t)$. All code and data used in our analysis are available at `https://github.com/cathalmills/forecast_evaluation`.

6.2 From metric to action I: Evaluation workflow results

We first evaluated the forecasts across the studied time period. Following our workflow (Figure 1), suppose a first policy question is the common task of predicting the future continuous value of weekly incident COVID-19 cases for each of the 51 studied states/territories. Using WIS and rWIS, for both forecast horizons of 1–4 weeks, the ensemble model generally outperformed other models, including across all time points (Figure SI 1 A), for individual locations (Figure SI 1 B), and in trailing windows of length 12 weeks (Figure SI 1 C, SI 3 – SI 6). Outperformance over time and across space was less clear for longer forecast horizons (Figure SI 1 C, Figures SI 5 – SI 6). As WIS approximates CRPS for quantile-represented forecasts, the ensemble model generally produced forecasts with the most value for a decision-maker with uniform risk preferences, when value was measured across all epidemic event thresholds (see Section 2.3).

These forecasts have already been extensively analysed through the lens of WIS [8, 9]. In epidemic settings, we are often interested in the many decisions for epidemic control faced by a specific decision-maker. A forecaster or evaluator can consider a risk-intolerant decision-maker with low C/L ratio (α) – action is inexpensive, relative to epidemic losses. In weather forecasting, a typical range for extreme events considered is α between 0.01 and 0.2 [60, 71]. For both forecast horizons and a range of C/L ratios, according to the pinball loss and its CORP decomposition, the ensemble model produced simultaneously more discriminative and better calibrated forecasts (Figure 3 A). This indicates higher forecast value for the ensemble model for risk-intolerant decision-makers across the entire set of observed epidemic event thresholds. Such user-specific insights may be useful if the decision-maker could not, or did not wish to, define a specific epidemic event *a priori* (e.g., evolving dynamics), but had a fixed risk preference across event thresholds and over time. We further examined pinball loss for different decision-maker preferences over time (Figures SI 15–SI 16) and across space (Figures SI 18–SI 24), where the ensemble did not always outperform the baseline model for risk-intolerant decision-makers (e.g., Figure SI 15 and SI 18). However, for binary event thresholds, the ensemble was the only model to provide a generally greater forecast value over the baseline model across different forecast horizons and most C/L ratios, aligning with the findings for WIS across the predictive distribution. Compared to risk-intolerant decision-makers and longer horizons, we found greater relative value for more risk-tolerant decision-makers (higher C/L ratios) and for shorter forecast horizons.

Focusing on the axis of epidemic events (see Figure 2 B, right panel), we found that for classifying the upper 10% of each location's observed incident cases (location-specific event thresholds θ), the CORP-based decomposition of the BS indicated similar results – the ensemble providing more discriminative and better calibrated forecasts for decision-makers (Figure 3B). This event-specific analysis focuses on state-specific extreme events for a uniform family of decision-makers, yet ignores any likely

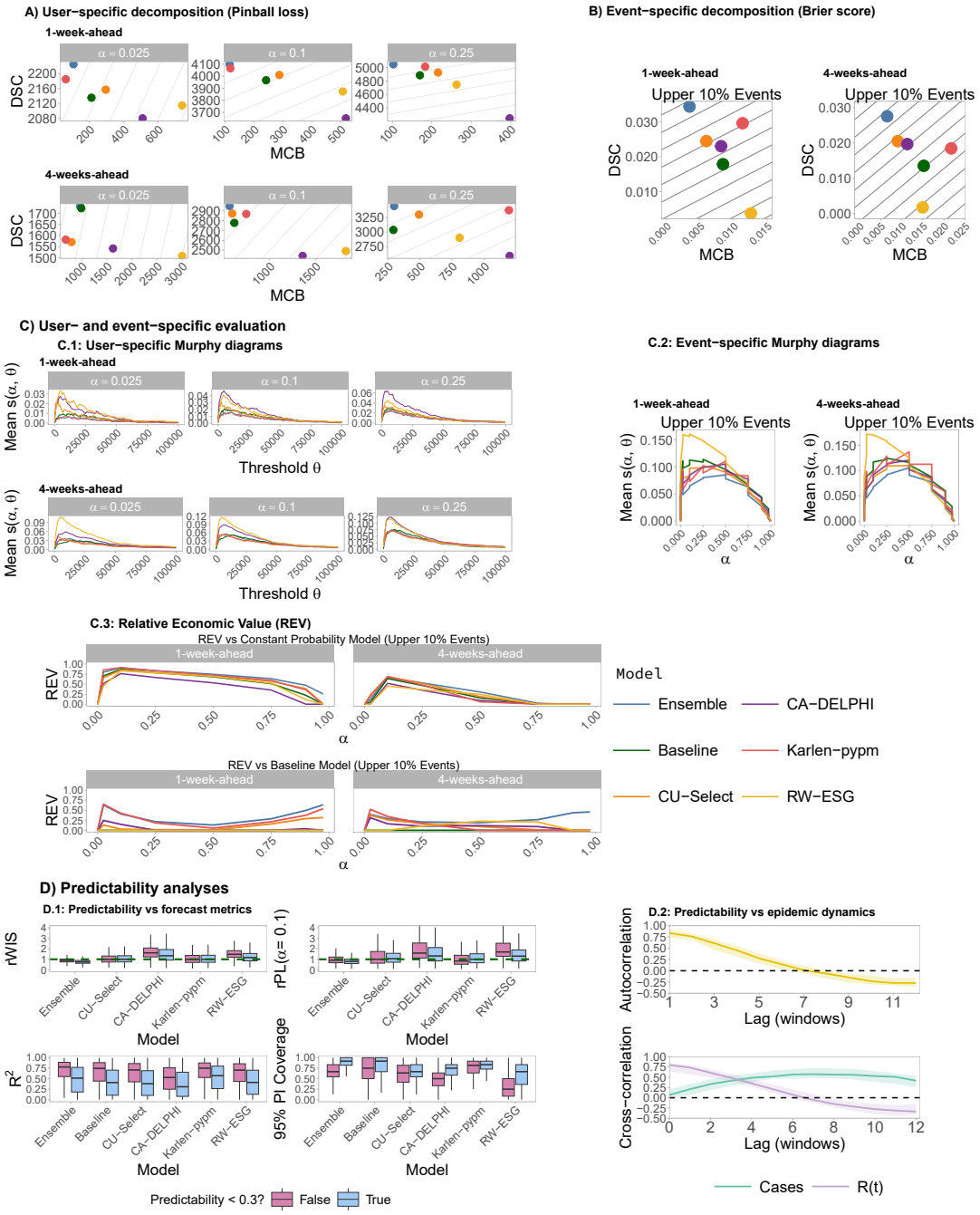

Figure 3: **COVID-19 forecast evaluation results** for application of our framework to state-level forecasts from 1 August 2020 – 15 January 2021, focusing on one-week-ahead and four-weeks-ahead forecasts. **A)** Pinball loss, a user-specific measure of forecast value, using discrimination (DSC) ability versus miscalibration (MCB). Higher DSC and lower MCB values indicate superior forecast value for a specific decision-maker's C/L ratio α and grey lines indicate constant forecast value. **B)** Analogous decomposition for the Brier Score, an event-specific measure of forecast value, focusing on the upper 10% of each location's observed cases. **C)** User- and event-specific evaluation measures of forecast value including event-specific Murphy diagrams (C.1) which capture the mean elementary score $s(\alpha, \theta)$ for a fixed C/L ratio (α) and varying event thresholds (θ), user-specific Murphy diagrams (C.2) which capture mean $s(\alpha, \theta)$ for a fixed event threshold (θ is the state-specific upper 10% threshold of cases) and varying C/L ratio (α), and Relative Economic (REV) diagrams (C.3) for the upper 10% of each state's cases. Lower values on Murphy curves and higher values on REV curves indicate superior performance of a model's forecasts for a given user-event combination. **D)** Predictability analyses include boxplots (D.1) of the relationship between predictability (binary thresholds here, further analyses in Tables SI 3 – SI 5, Figures SI 27 – SI 29) and forecast metrics, alongside median autocorrelation (D.2) of predictability and cross-correlation of predictability with cases and $R(t)$, time-varying reproduction number estimates. Shaded intervals capture the interquartile range across locations. Metric details are summarised in Table 1 and Figure 2.

risk preferences of the decision-maker that could expressed with a range of plausible C/L ratios, especially for extreme events.

User-specific (Figure 3 C.1) and event-specific (Figure 3 C.2) Murphy diagrams allow simultaneous assessment of forecast value for different epidemic events (or decision-makers) and a specific decision-maker (or a specific epidemic event). The Murphy diagrams for specific decision-makers (Figure 3 C.1, holding α fixed) trace a trend of the Karlen-pypm and ensemble model's forecasts often providing the most value for risk-intolerant decision-makers with low C/L ratios (for whom action is inexpensive, relative to the loss). We found less clear trends for rarer, larger event thresholds (θ) across forecast horizons. For more risk-tolerant decision-makers (i.e., higher C/L ratios), the ensemble model provided clearer value for longer forecast horizons (Figure SI 26). These user-specific Murphy diagrams, while crowded and likely benefitting from further focus on event thresholds θ for specific policy questions, capture the value of the models' forecasts for informing a specific decision-maker's binary decision of whether to act based on the forecasted probability of the future (one or four weeks ahead) weekly incidence exceeding different event thresholds (θ). Focusing on location-specific extrema (location-specific event thresholds θ), the event-specific Murphy diagrams for BS (Figure 3 C.2) display that the ensemble model provided the most value for classifying the upper 10% of each location's incident cases for most decision-makers across both forecast horizons (up to C/L ratios of 0.75). There is benefit in using both types of Murphy diagram, as the user- and event- specific visualisations can capture forecast value when there is uncertainty in the defined event (e.g., evolving dynamics) and/or decision-maker's C/L ratio (e.g., likely uncertainty about cost of intervention or epidemic event loss). This can inform decision-making under different uncertainties.

The REV does not capture forecast performance in terms of elementary scores, but provides an alternative, interpretable measure of relative value for different forecasting models over other simple, operational models. The choice of reference model is crucial, and we used i) a probabilistic classifier that always predicts the binary event to occur with a fixed probability (similar to the climatology commonly used in weather forecasting [20, 60]), and ii) the COVID-19 Forecast Hub baseline model [7].

Using the REV for the upper 10% of each location's incident cases (location-specific event thresholds θ), relative to a constant probabilistic classifier (forecasting occurrence of a binary event with probability 0.1, Figure 3 C.3 top row), we found that i) almost all models provided relative value for a range of decision-maker risk preferences, ii) REV was generally greater for shorter horizons, and iii) the ensemble model (alongside the Karlen-pypm model for one-week-ahead forecasts) provided the most clear relative value across the range of risk preferences. Although the relative value of models' forecasts was often lower when using the baseline model as reference, the ensemble model again provided the most clear relative value, including for risk-tolerant decision-makers (high C/L ratios). We repeated these REV analyses for different types of events, such as the upper 10% (15%) of all observed incident cases (Figures SI 11–SI 12) and for the upper 15% of each location's observed incident cases (Figures SI 9–SI 10 and Figures SI 13–SI 14), and found largely similar results (e.g., highest REV for these same two models and similar trends across C/L ratios and forecast horizons). In practice, the results from these REV diagrams could be filtered by a forecaster or evaluator to interpret them for a specific decision-maker's defined event, forecast horizon, and available action.

Relating forecast performance to predictability, higher predictability was moderately yet significantly (Table SI 3) correlated with higher R^2 for each model's forecasts across different horizons (Figures 3 D.1 and SI 27). Using relative measures such as rWIS (Figures 3 D.1, SI 29) or relative pinball loss across different C/L ratios (Figures 3 D.1 and SI 25), we estimated that several forecasting models generally had *greater* relative performance when epidemic dynamics were *less* predictable. However, greater relative performance for the given model during times of lower predictability, compared to times of high predictability, does not mean that the models necessarily provided improvements over a naive/baseline model at such times (Figure 3 D.1). It is necessary to understand their overall predictive performance during all regimes of predictability. We interpret the relative improvements alongside comparisons to the baseline (rWIS and relative pinball loss alongside predictability, Figures 3 D.1). We estimated generally consistent benefits, across space, time, and metrics, of using the ensemble model over the naive baseline model, for decision-making during less *and* more predictable times. The estimated relative improvements of the ensemble model, during high unpredictability, were often greater at longer forecast horizons (Figures SI 15 – SI 29).

These findings shed light on how predictability could have impacted the value of different models, yet we also considered whether such predictability itself can be anticipated and/or what drives different predictability regimes. Higher predictability within a window was strongly correlated with higher $R(t)$ estimates within the window (Pearson's [72] correlation $r = 0.70$, 95% CI = (0.68, 0.72)) and in past windows, while higher predictability was correlated with higher levels of incident cases in the recent past (Figures 3 D.2, SI 30). This indicates that predictability was greatest during the epidemic growth and peak phases (as classified by [9]), and lower during the declining and nadir phases. Overall, we found generally low predictability, and this varied across space and time (Figures SI 30 – SI 31). These insights, while needing further exploration beyond our scope of evaluation, may be useful for informing model development and selection and for anticipating the likely future reliability of outputs. For example, lower $R(t)$ preceding lower predictability, or past performance during growth or declining phases, may be used to guide changes to a model or to an ensemble's weights. In future works, the recent past and likely future predictability could be incorporated into the forecasting models or into the model selection procedures. For instance, we found that for the ensemble and several models, higher predictability was correlated with lower coverage of central 95% prediction intervals (PIs) (Figure 3 D.1, Table SI 4), particularly for longer forecast horizons (Figure SI 4). This reflects findings from [9] where coverage dropped during

growth phases which we estimated to be generally more predictable. During such phases, central 95% PI coverage, even from the ensemble's forecasts, was often much lower than 95%, suggesting a need for updating of ensemble weights and/or models and their inputs (e.g., data collection).

6.3 From metric to action II: Real-time decision-making

Above, we evaluated forecast quality and value *across* the 77 weeks. In practice, forecasters and evaluators will apply the framework for local decision-makers in specific locations at specific times. It is so far less clear how the forecasts could have informed decisions in real time for location-specific, important stages of the epidemic. Suppose now that a decision-maker in a given location is unknowingly at the onset of a large wave, which we define as the first date in 2021 when the median estimate of $R(t) > 1$. We focus on Arizona (a randomly chosen state) and describe pre-wave results in Section F.1.

Studying the COVID-19 wave, rWIS results for one-week-ahead forecasts remain comparable to pre-wave results (Section F.1). The same two models (Ensemble and Karlen-pypm) provided the best overall predictive performance, yet no model consistently outperformed the baseline for four-weeks-ahead forecasts (Figure 4 B). Focusing on specific events, co-defining *a priori* the policy question and epidemic event to measure before the wave is non-trivial and infinite choices exist. First, a decision-maker may be interested in whether future incident cases at a specified horizon breach a fixed extreme threshold, say $\theta = 20,000$ (an extreme value *before* the wave, Figure 4 A). The decision problem for this event could correspond to taking action based on the costs and preventable losses of maintaining a public health measure (e.g., additional ICU capacity, see Section 2.4) if incident cases $y \geq 20,000$. Using the CORP-based decomposition of the mean BS for this event (Figure 4 C, left), both the ensemble and CU-select model provided the most value across discrimination ability and calibration properties. However, if we add a decision-maker's risk preferences (Figure 4 C, middle), the event-specific Murphy diagrams capture how the optimal model varied by risk preferences. The analogous REV plots (Figure 4 C, right) are also crowded, yet illustrate differences in selected models across risk preferences and the low value for longer forecast horizons. In practice, we could provide the operational forecast value of each model for a specific decision-maker using a plausible range of their C/L ratios for this event threshold ($y \geq 20,000$).

On the other hand, an event threshold may become less meaningful if dynamics are evolving quickly and/or incident cases are predictably above/below this threshold. Varying this epidemic event threshold (θ), the user-specific Murphy diagram for pinball loss (Figure 4 D, right panel) captures how well the forecasts informed decisions using different event thresholds. These diagrams can also easily become crowded (e.g., due to a wide range of events on the x-axis, see Figure 4 D). Alternatives may be to focus on the CORP-based decompositions for BS or pinball loss (Figure 4 C–D) or to restrict x-axes of the Murphy diagram to event thresholds (θ) of interest.

Suppose a decision-maker cannot pre-define a fixed event threshold and/or will not base decisions on a single event threshold, yet has a fixed risk profile (C/L ratio α). Focusing on *all* observations of the wave for these fixed risk preferences, the pinball loss decompositions (Figures 4 D and SI 35) capture how the ensemble provided the clearest value for risk-intolerant decision-makers. However, the chosen model for decision-makers varied according to their relative cost of mitigation (C/L ratio) and the forecast horizon. The models with the most value for a given decision-maker during the wave (Figure SI 35 B) often differed to the models with the most value before (Figure SI 35 A) the wave. We found similar differences in models selected before versus during the wave when using the Brier score, REV, and relative pinball loss (Figure SI 36–SI 37). Collectively, these results suggest that for policy-focused model development and representative evaluation, forecasters and evaluators need to i) zoom in on user- and event-specific measures and ii) capture the spatiotemporal structures of epidemic data in evaluation and model selection. Otherwise, averaging over these heterogeneities may result in inaccurate measures of local, real-time forecast value and a failure to provide different decision-makers with the optimal model's forecasts.

Retrospectively, we can analyse how well the peak values were predicted, for example, using BS or REV for the upper 10% of incident cases during the wave. This assessment of wave-specific extreme events can be used by forecasters and/or evaluators to build better models going forward and/or to assign ensemble weights. Here, the best models across the C/L ratios for predicting these events again varied by different forecast horizons (Figure 4 E, right) and by different decision-makers' preferences (Figure 4 E, left), while the best models differed to those chosen for the upper 10% of events before the wave (Figure SI 32).

Compared to the pre-wave results (Section F.1), similar predictability trends persisted during the wave (Figures 4 A and SI 38–SI 39), as predictability rose and fell alongside the epidemic trajectory – $R(t)$ was strongly correlated with estimated predictability ($r = 0.80$, 95% CI: (0.62, 0.91)) during the wave. Low predictability was again correlated with lower pointwise accuracy for models, and during such low predictability regimes, some models had greater value relative to the baseline model for the uniform family of decision-maker preferences, but not necessarily individual decision-maker preferences (Figures SI 32 –SI 37 and Figure SI 39). Therefore, given that sharp predictability changes were also estimated towards the start and at the peak of the wave, it is perhaps unsurprising that the best models varied (Figures SI 32 – SI 37) i) for different decision-makers, ii) for different events, iii) over time within the wave, and iv) compared to the pre-wave period. Looking forward, these insights (e.g., predictability having consistent impacts on forecast performance), which benefit from our hindsight, suggest the need for model

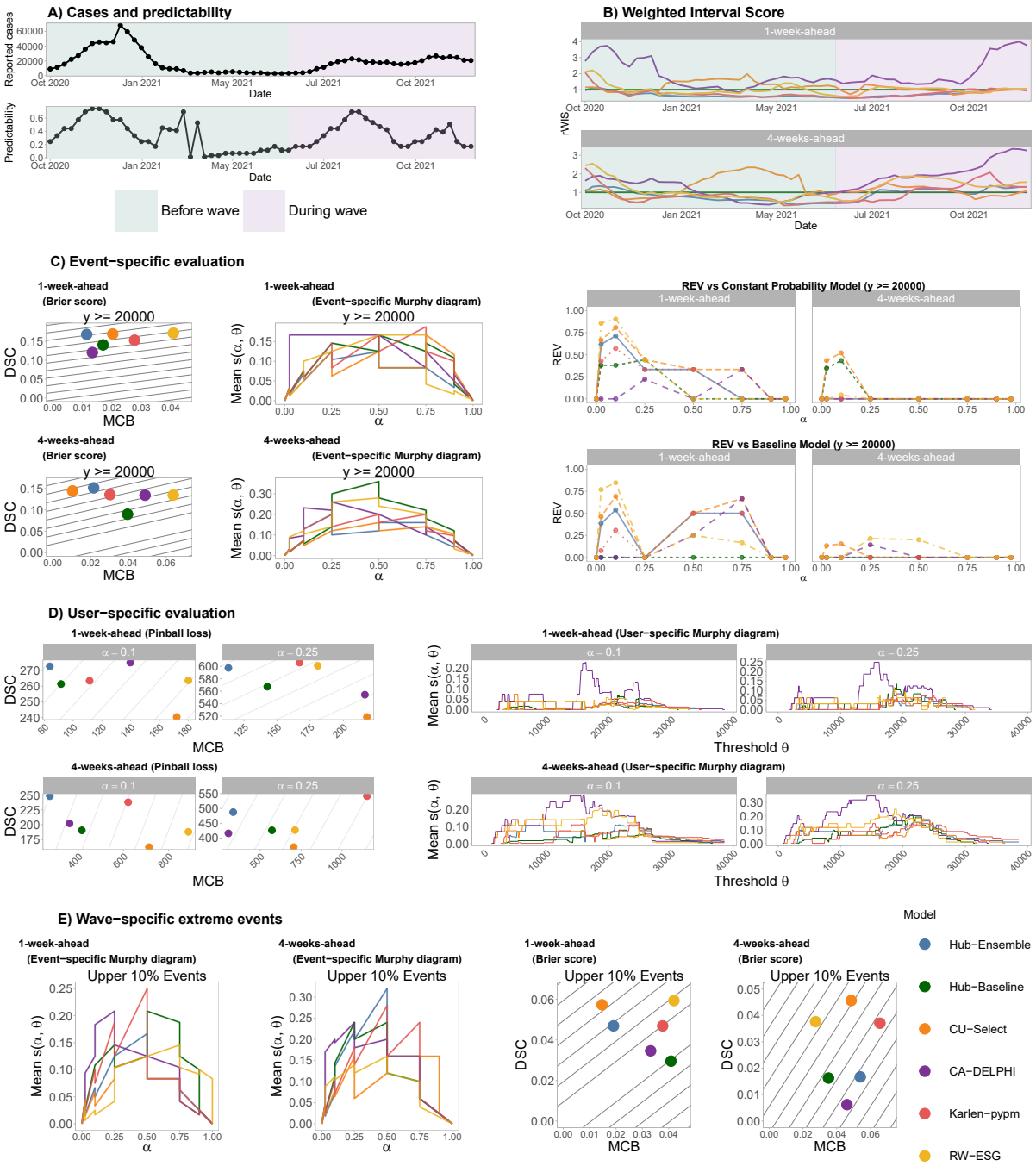

Figure 4: **Evaluation for real-time decision-making during a COVID-19 wave in Arizona: A)** Reported incident cases and estimated predictability before and during a COVID-19 wave that started in July 2021 (based on median $R(t) > 1$). **B)** 12-week trailing window relative Weighted Interval Score (rWIS) for all models across two forecast horizons before and during the wave. **C)** For two forecast horizons, Brier Score decomposition and event-specific Murphy diagrams, alongside a Relative Economic Value diagram, where the event is pre-defined as incident cases $y \geq 20{,}000$ (based on values before the wave). **D)** For two forecast horizons, Brier Score decomposition and event-specific Murphy diagrams for risk-intolerant decision-makers. We provide further visualisations for other risk-intolerant decision-maker preferences and risk-tolerant decision-makers in Figures SI 34 – SI 37. **E)** Event-specific Murphy diagrams and Brier Score decomposition and event-specific Murphy diagrams for forecasting the upper 10% of incident cases during this COVID-19 wave. The studied period for the wave ends on January 15 2022.

development and selection, ensemble weighting, and data collection to condition on predictability, in addition to time, space, and the decision-maker's preferences and priorities.

7 Discussion

Forecasting epidemics is challenging. The future that is predicted is both nonlinear and highly stochastic [73], and inherently uncertain [24, 74]. Working towards more timely and reliable predictions, our understanding of infectious disease forecasting has developed recently. Methods for evaluating forecast performance have evolved substantially, yet it was unclear how this statistical rigour translates into measuring the forecast's value for informing a decision-maker's dilemma: to act or not. A disconnect between forecasters and decision-makers and a lack of interpretable and operationally relevant forecast outputs have long been challenges in the operational use of forecasts to support decision-making [75]. Here, for forecasters aiming to build better models for decision-makers' needs and for forecast evaluators aiming to evaluate how well the models meet these needs, we provide a framework that contributes procedures to: i) recast forecasting models' statistical strengths and weaknesses into policy-relevant statistical properties across space and time, ii) select the "optimal" model for a decision-maker's individual policy question and risk preferences, and iii) link the limits of predictability for the epidemic to model selection and forecast value. The aim is that this user-focused framework provides a foundation for improving forecasting models and measuring the value of their forecasts by their ability to inform the uncertain decision-making of infectious disease outbreaks.

As shown in our applications, our new framework does not replace best practices in forecast reporting [2], evaluation techniques [8–10, 76], forecast model development (e.g., time series cross-validation [77, 78]) or standard diagnostics (e.g., calibration and reliability assessments [42, 79]).

Our framework establishes connections between many best practices and extends to other branches of science (e.g., weather forecasting and decision theory). Importantly, the user- and decision-focused conceptual framework presents evidence and does not provide advice (which would involve subjective policy guidance) [80, 81]. The framework allows the forecaster and evaluator to provide a decision-maker with the forecasts of the "optimal" model for their decision problem, based on past performance for the set of models and the defined policy question, action set, and risk preferences. The optimality conditions are highly specific and must be pre-defined, yet our applications suggest how to incorporate uncertainties and misspecification in decision-makers' risk preferences and priorities. From a decision and social theory perspective, our user-focused prioritisations reflect theory that a group-level utility function which fairly weights individual preferences is mathematically problematic and arguably unattainable [82, 83]. Given that individuals have different risk preferences, there is no single scoring function for the group that allows all individual users to select the same forecasting model for the given task. Instead, we incorporate a decision-maker's individual policy question and risk preferences to determine the model(s) that have provided the most value for informing decisions. Similarly to [19], we argue that this decision-theoretic framework enhances the credibility and defensibility of uncertain policymaking. Iterative application of our evaluation framework for different policy questions connects with other research focused on decision problems [19, 75] as part of *integrated* assessments for scientific advisors and policymakers. Decision-making before/during epidemics will never be based on a single policy question or forecast, epidemic event/outcome, or branch of science, yet if forecasting is to take place, evaluation must measure how well the forecasts inform decisions.

For operationalisation, collaboration with decision-makers will be essential to improve our proposed workflow (Figure 1). Models, evaluation metrics, and epidemic trajectories can be complex, and stakeholders have different priorities and backgrounds. Communicating uncertainty to both technical experts and decision-makers is essential yet non-trivial [81, 84], challenged by i) a lack of agreement about uncertainty and the language of probability, and ii) the many nuances of an epidemic forecast (see below) [85, 86]. Clear, timely, and succinct communication will involve decreasing levels of technicality as information is passed up the decision-making chain/process, while preserving key modelling outputs and uncertainty. Similarly, how a scientist represents data (e.g., binary versus continuous [87]) and communicates uncertainty [88] affects how a user interprets the strength of evidence. Future behavioural studies could quantify the optimal ways to communicate evaluation and forecast results to different decision-makers. This could involve traffic light schemes and warning impact matrices [89], or interactive dashboards [76, 90]. All of this will benefit from continuous dialogue between modellers, social scientists, and decision-makers [81, 84].

Recent guidelines for prediction tools advocate for a 3-U (useful, usable, and used) framework [91], which we embrace in the context of evaluation with public-health-relevant quantities, user-focused evaluations and visualisations, and metric-based decision-making. For example, the C/L model appeals to decision-maker preferences around the world for possibly crude but interpretable thresholds [92]. Future post-event surveys and analyses should investigate how well the framework addressed the specific needs of decision-makers for informing decisions, and which metrics to focus on to provide reliable *and* timely information.

A key next step will be to elicit estimates of decision-makers' C/L ratios (α) in real time for different event thresholds (θ), requiring careful setting-specific considerations (Section 2.4). Without plausible ranges for these C/L ratios, the framework could remain abstract. Eliciting preferences will benefit from i) regular consultation and surveys with decision-makers and experts across multiple disciplines, and ii) economic evaluation literature for different user perspectives [93]. Real-time surveys could define the distributions of users' C/L ratios and event thresholds for different epidemic questions. For example, [94] developed a survey-based user BS to understand how forecast users combine objective probabilities with their subjective feelings when making

decisions based on weather forecasts. Similarly, additional post-event analyses could help to better determine decision-makers' C/L ratios and policy priorities by studying their responses and actions, and the impacts on epidemic dynamics. There will always be both heterogeneity and uncertainty in such C/L ratios as decisions before and during an epidemic are multidisciplinary tasks with ethical and human components [19]. Given the feedback loops, transience and uncertainty in C/L ratios, reinforcement learning and/or model-predictive control may also be useful for iterative adjustments, alongside recent frameworks for evaluating cost-effectiveness of different interventions [55, 95] and for integrating public health ethics in models [96].

The C/L model is grounded in decision theory, yet remains a simple abstraction of reality, and future work may consider asymmetric utility functions, three-way decision problems, continuous action sets and stochastic policies, and/or multi-objective algorithms for decision-making using different policy questions and metrics. Our framework components could also be embedded in recent multi-model approaches for policy decisions under uncertainty [18, 97].

Depending on the target quantity, the predicted future can be modifiable, modifications that a forecasting model generally does not anticipate. The "best" model may be the one that promotes the decision-maker to implement interventions that have direct impacts on epidemic dynamics, which can differ from the model that produces the most accurate forecast of the realised future. Thus, the actions of a decision-maker can impact the very probability of the epidemic events that a model aims to forecast. A probabilistic forecast integrates over uncertainty in many different counterfactual scenarios and is a special, *unconditional* case of a scenario projection [85] – a model's best estimate of what *will* happen.

Forecasts may be operationally advantageous for empowering decision-makers to make decisions based on predicted probabilities for the most likely future states of the world. This avoids the difficulties of probability statements and evaluation in inherently counterfactual scenario projections. However, forecasts can suffer from irreducible error due to unknown and possibly model-based modifications in the intervening period after the forecast date. This can yield limits to predictability, and such feedback from the future means that forecast evaluation, without adjustment, can measure both the adequacy of the forecasts and the effectiveness of policy responses [98]. These issues are exacerbated (alleviated) for longer- (shorter-) term forecasts where model-based future policy changes could be more (less) likely. Two potential ways to address these feedback loops include: i) directly incorporating all different policies with corresponding probabilities in the integration over all possible future states of the world, highlighting the need for modeller and decision-maker dialogue, or ii) using scenario projections for these quantities, which can consider the "what-if" situations of different future policy decisions [85]. We suspect that scenario projections, possibly paired with our evaluation framework, may be more suitable when there is greater uncertainty about future policy changes (e.g., longer-term horizon seasonal predictions), while unconditional forecasts may be more suited to short-term horizons. We highlight here that this feedback loop is not always an issue for forecasting as targeted epidemiological quantities can be largely invariant to interventions (e.g., incident cases after adding to ICU capacity). All of this highlights the need for close collaboration with decision-makers to maintain clarity about the future policies, the future that a forecast/scenario projection is targeting, and the associated utility, limitations, and implications of the prediction model.

Forecasting and scenario projections remain two strands of the scientific toolkit to inform decision-makers. Epidemics are not simply a prediction problem. A reliable prediction does not necessarily convert to effective action. Without understanding *why* or *how* transmission has occurred, predictions cannot inform targeted strategies to control an epidemic. Our framework should be used alongside inference of other key epidemiological quantities, variant/serotype predictions, nowcasts of the present dynamics, analyses of intervention effectiveness, and many other tasks [86]. Our evaluation framework could be adapted for such tasks, for example for outbreak risk (is $R(t) > 1$?). Beyond modelling, experts from other fields (e.g., health economics and social theory) should be embedded in multidisciplinary decision-making processes for an infectious disease outbreak [19].

The applications of our framework have several limitations. We evaluated forecasts on the imperfect reported COVID-19 case data without any data on policy decisions and their impacts or time-varying and/or space-varying reporting rates. These influences may vary within a studied population and for different infectious diseases. Our applications therefore reflect the realities of data collection, modelling, and human responses in real-time outbreaks. However, our predictability metric (PE) is robust to both observational and dynamical noise, while we focus on stable performance over 12-week windows and across space to limit sensitivity to surveillance artefacts. Epidemiological delays are also commonplace, and delayed case reporting impacts the evaluation process and the communication of timely information for decision-making [6, 99]. This suggests using new methods for estimating delay distributions [100–102] and for nowcasting [90, 103, 104]. In the future, forecasters and evaluators will benefit from i) extensions to DRO (Section 4), ii) further diagnostic checks (e.g., reliability diagrams [59]) iii) formal statistical testing [105–107], and iv) forecast post-processing techniques [45]. Finally, it will be instructive to analyse the relationship between evaluation metrics and different model types, to inform both model and ensemble development.

Our predictability analyses are a stepping stone towards more concrete links between predictability and forecasting. Using historical predictability to anticipate future predictability is imperfect. Historically predictable epidemic dynamics may no longer be predictable during fast-evolving epidemics (e.g., new SARS-CoV-2 variants). We could address this with i) predictability assessments that are conditional on covariates, ii) regular collection of heterogeneous data to anticipate such changes, and iii) extensions to multivariate PE [108]. PE is also only one way to estimate predictability, and alternatives include permutation

dependence [109] or learning and irreducible errors [110].

In short, our new framework shifts the perspective of forecast evaluation from the forecaster to the decision-maker. Recognising that there is no one-size-fits-all approach to forecasting, evaluation, or decision-making, we provide principled procedures for forecasters and evaluators to measure and improve the value of forecasting models for informing and empowering decision-makers.

References

[1] Allan H. Murphy. "What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting". en. In: *Weather and Forecasting* 8.2 (June 1993), pp. 281–293. ISSN: 0882-8156, 1520-0434. DOI: 10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2. URL: http://journals.ametsoc.org/doi/10.1175/1520-0434(1993)008%3C0281:WIAGFA%3E2.0.CO;2 (visited on 01/13/2025).

[2] Simon Pollett et al. "Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines". en. In: *PLOS Medicine* 18.10 (Oct. 2021), e1003793. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1003793. URL: https://dx.plos.org/10.1371/journal.pmed.1003793 (visited on 06/13/2024).

[3] Farzaneh Sadat Tabataba et al. "A framework for evaluating epidemic forecasts". en. In: *BMC Infectious Diseases* 17.1 (Dec. 2017), p. 345. ISSN: 1471-2334. DOI: 10.1186/s12879-017-2365-1. URL: http://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-017-2365-1 (visited on 01/28/2025).

[4] Michael A. Johansson et al. "An open challenge to advance probabilistic forecasting for dengue epidemics". en. In: *Proceedings of the National Academy of Sciences* 116.48 (Nov. 2019), pp. 24268–24274. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1909865116. URL: https://pnas.org/doi/full/10.1073/pnas.1909865116 (visited on 09/21/2023).

[5] Cécile Viboud et al. "The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt". en. In: *Epidemics* 22 (Mar. 2018), pp. 13–21. ISSN: 17554365. DOI: 10.1016/j.epidem.2017.08.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S1755436517301275 (visited on 09/03/2024).

[6] Nicholas G. Reich et al. "A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States". en. In: *Proceedings of the National Academy of Sciences* 116.8 (Feb. 2019), pp. 3146–3154. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1812594116. URL: https://pnas.org/doi/full/10.1073/pnas.1812594116 (visited on 02/08/2025).

[7] Estee Y. Cramer et al. "The United States COVID-19 Forecast Hub dataset". en. In: *Scientific Data* 9.1 (Aug. 2022), p. 462. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01517-w. URL: https://www.nature.com/articles/s41597-022-01517-w (visited on 03/27/2025).

[8] Evan L. Ray et al. "Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States". en. In: *International Journal of Forecasting* 39.3 (July 2023), pp. 1366–1383. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.06.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022000966 (visited on 01/02/2024).

[9] Velma K. Lopez et al. "Challenges of COVID-19 Case Forecasting in the US, 2020–2021". en. In: *PLOS Computational Biology* 20.5 (May 2024). Ed. by Daniel B. Larremore, e1011200. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011200. URL: https://dx.plos.org/10.1371/journal.pcbi.1011200 (visited on 10/13/2024).

[10] Evan L. Ray et al. "Flusion: Integrating multiple data sources for accurate influenza predictions". en. In: *Epidemics* (Dec. 2024), p. 100810. ISSN: 17554365. DOI: 10.1016/j.epidem.2024.100810. URL: https://linkinghub.elsevier.com/retrieve/pii/S1755436524000719 (visited on 01/15/2025).

[11] Nikos I. Bosse et al. "Comparing human and model-based forecasts of COVID-19 in Germany and Poland". en. In: *PLOS Computational Biology* 18.9 (Sept. 2022). Ed. by James M McCaw, e1010405. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010405. URL: https://dx.plos.org/10.1371/journal.pcbi.1010405 (visited on 10/14/2024).

[12] Sarabeth M. Mathis et al. "Evaluation of FluSight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations". en. In: *Nature Communications* 15.1 (July 2024), p. 6289. ISSN: 2041-1723. DOI: 10.1038/s41467-024-50601-9. URL: https://www.nature.com/articles/s41467-024-50601-9 (visited on 10/20/2024).

[13] Nikos I. Bosse et al. "Scoring epidemiological forecasts on transformed scales". en. In: *PLOS Computational Biology* 19.8 (Aug. 2023). Ed. by James M McCaw, e1011393. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011393. URL: https://dx.plos.org/10.1371/journal.pcbi.1011393 (visited on 05/25/2024).

[14] Tim K. Tsang et al. "An adaptive weight ensemble approach to forecast influenza activity in an irregular seasonality context". en. In: *Nature Communications* 15.1 (Oct. 2024), p. 8625. ISSN: 2041-1723. DOI: 10.1038/s41467-024-52504-1. URL: https://www.nature.com/articles/s41467-024-52504-1 (visited on 10/10/2024).

[15] Jonathon Mellor et al. "Forecasting COVID-19, influenza, and RSV hospitalizations over winter 2023–4 in England". en. In: *International Journal of Epidemiology* 54.3 (Apr. 2025), dyaf066. ISSN: 0300-5771, 1464-3685. DOI: 10.1093/ije/dyaf066. URL: https://academic.oup.com/ije/article/doi/10.1093/ije/dyaf066/8156944 (visited on 06/08/2025).

[16] Daniel Wolffram et al. "Collaborative nowcasting of COVID-19 hospitalization incidences in Germany". en. In: *PLOS Computational Biology* 19.8 (Aug. 2023). Publisher: Public Library of Science, e1011394. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011394. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011394 (visited on 04/12/2025).

[17] Adrian Lison et al. "Generative Bayesian modeling to nowcast the effective reproduction number from line list data with missing symptom onset dates". en. In: *PLOS Computational Biology* 20.4 (Apr. 2024). Publisher: Public Library of Science, e1012021. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1012021. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012021 (visited on 04/12/2025).

[18] Beate Jahn et al. "On the role of data, statistics and decisions in a pandemic". en. In: *AStA Advances in Statistical Analysis* 106.3 (Sept. 2022), pp. 349–382. ISSN: 1863-818X. DOI: 10.1007/s10182-022-00439-7. URL: https://doi.org/10.1007/s10182-022-00439-7 (visited on 04/03/2025).

[19] Loïc Berger et al. "Rational policymaking during a pandemic". en. In: *Proceedings of the National Academy of Sciences* 118.4 (Jan. 2021), e2012704118. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2012704118. URL: https://pnas.org/doi/full/10.1073/pnas.2012704118 (visited on 04/03/2025).

[20] Martin Janousek Thomas Haiden. *Evaluation of ECMWF forecasts, including the 2023 upgrade*. en. text. 2023. URL: https://www.ecmwf.int/en/elibrary/81389-evaluation-ecmwf-forecasts-including-2023-upgrade (visited on 03/01/2025).

[21] Uwe Siebert. "When should decision-analytic modeling be used in the economic evaluation of health care?" en. In: *The European Journal of Health Economics, formerly: HEPAC* 4.3 (Sept. 2003), pp. 143–150. ISSN: 1618-7601. DOI: 10.1007/s10198-003-0205-2. URL: https://doi.org/10.1007/s10198-003-0205-2 (visited on 04/03/2025).

[22] Kelly R. Moran et al. "Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast". en. In: *Journal of Infectious Diseases* 214.suppl 4 (Dec. 2016), S404–S408. ISSN: 0022-1899, 1537-6613. DOI: 10.1093/infdis/jiw375. URL: https://academic.oup.com/jid/article-lookup/doi/10.1093/infdis/jiw375 (visited on 04/22/2025).

[23] Vanessa J. Fundel et al. "Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users". en. In: *Quarterly Journal of the Royal Meteorological Society* 145.S1 (Sept. 2019), pp. 210–231. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.3482. URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3482 (visited on 06/09/2025).

[24] Tilmann Gneiting and Adrian E Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". en. In: *Journal of the American Statistical Association* 102.477 (Mar. 2007), pp. 359–378. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214506000001437. URL: http://www.tandfonline.com/doi/abs/10.1198/016214506000001437 (visited on 04/21/2025).

[25] Glenn W. Brier. "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78 (Jan. 1950). Publisher: AMS ADS Bibcode: 1950MWRv...78....1B, p. 1. ISSN: 0027-0644. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL: https://ui.adsabs.harvard.edu/abs/1950MWRv...78....1B (visited on 02/15/2025).

[26] Estee Y. Cramer et al. "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 119.15 (Apr. 2022), e2113561119. ISSN: 1091-6490. DOI: 10.1073/pnas.2113561119.

[27] Katharine Sherratt et al. "Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations". en. In: *eLife* 12 (Apr. 2023), e81916. ISSN: 2050-084X. DOI: 10.7554/eLife.81916. URL: https://elifesciences.org/articles/81916 (visited on 03/02/2025).

[28] Johannes Bracher et al. "Evaluating epidemic forecasts in an interval format". en. In: *PLOS Computational Biology* 17.2 (Feb. 2021). Ed. by Virginia E. Pitzer, e1008618. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008618. URL: https://dx.plos.org/10.1371/journal.pcbi.1008618 (visited on 01/17/2024).

[29] Estee Y Cramer et al. *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US*. en. Feb. 2021. DOI: 10.1101/2021.02.03.21250974. URL: http://medrxiv.org/lookup/doi/10.1101/2021.02.03.21250974 (visited on 05/12/2025).

[30] Edward Wheatcroft. "Interpreting the skill score form of forecast performance metrics". en. In: *International Journal of Forecasting* 35.2 (Apr. 2019), pp. 573–579. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2018.11.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207019300093 (visited on 02/11/2025).

[31] David Richardson. "Skill and relative economic value of the ECMWF Ensemble Prediction System". In: (1998). Publisher: ECMWF. DOI: 10.21957/2XJS554NR. URL: https://www.ecmwf.int/node/11903 (visited on 02/15/2025).

[32] Timo Dimitriadis, Tilmann Gneiting, and Alexander I. Jordan. "Stable reliability diagrams for probabilistic classifiers". en. In: *Proceedings of the National Academy of Sciences* 118.8 (Feb. 2021), e2016191118. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2016191118. URL: https://pnas.org/doi/full/10.1073/pnas.2016191118 (visited on 02/11/2025).

[33] Cathal Mills et al. *Multi-model approach to understand and predict past and future dengue epidemic dynamics*. en. Pages: 2024.10.18.24315690. June 2025. DOI: 10.1101/2024.10.18.24315690. URL: https://www.medrxiv.org/content/10.1101/2024.10.18.24315690v2 (visited on 06/10/2025).

[34] Eduardo Correa Araujo et al. *Leveraging probabilistic forecasts for dengue preparedness and control: the 2024 Dengue Forecasting Sprint in Brazil*. en. May 2025. DOI: 10.1101/2025.05.12.25327419. URL: http://medrxiv.org/lookup/doi/10.1101/2025.05.12.25327419 (visited on 05/16/2025).

[35] Felipe J. Colón-González et al. "Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles". en. In: *PLOS Medicine* 18.3 (Mar. 2021). Ed. by Alex Richard Cook, e1003542. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1003542. URL: https://dx.plos.org/10.1371/journal.pmed.1003542 (visited on 01/22/2024).

[36] Allan H. Murphy. "A New Vector Partition of the Probability Score". en. In: *Journal of Applied Meteorology* 12.4 (June 1973), pp. 595–600. ISSN: 0021-8952. DOI: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. URL: http://journals.ametsoc.org/doi/10.1175/1520-0450(1973)012%3C0595:ANVPOT%3E2.0.CO;2 (visited on 04/03/2025).

[37] Hans Hersbach. "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems". en. In: *Weather and Forecasting* 15.5 (Oct. 2000), pp. 559–570. ISSN: 0882-8156, 1520-0434. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2. URL: http://journals.ametsoc.org/doi/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2 (visited on 04/03/2025).

[38] G. Candille and O. Talagrand. "Evaluation of probabilistic prediction systems for a scalar variable". en. In: *Quarterly Journal of the Royal Meteorological Society* 131.609 (July 2005), pp. 2131–2150. ISSN: 0035-9009, 1477-870X. DOI: 10.1256/qj.04.71. URL: https://rmets.onlinelibrary.wiley.com/doi/10.1256/qj.04.71 (visited on 04/03/2025).

[39] Meelis Kull and Peter Flach. "Novel Decompositions of Proper Scoring Rules for Classification: Score Adjustment as Precursor to Calibration". en. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Annalisa Appice et al. Vol. 9284. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 68–85. ISBN: 978-3-319-23527-1 978-3-319-23528-8. DOI: 10.1007/978-3-319-23528-8_5. URL: http://link.springer.com/10.1007/978-3-319-23528-8_5 (visited on 04/03/2025).

[40] Stefan Siegert. "Simplifying and generalising Murphy's Brier score decomposition". en. In: *Quarterly Journal of the Royal Meteorological Society* 143.703 (Jan. 2017), pp. 1178–1183. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.2985. URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.2985 (visited on 04/03/2025).

[41] Sebastian Arnold et al. "Decompositions of the mean continuous ranked probability score". In: *Electronic Journal of Statistics* 18.2 (Jan. 2024). ISSN: 1935-7524. DOI: 10.1214/24-EJS2316. URL: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-18/issue-2/Decompositions-of-the-mean-continuous-ranked-probability-score/10.1214/24-EJS2316.full (visited on 05/16/2025).

[42] Tilmann Gneiting et al. "Model Diagnostics and Forecast Evaluation for Quantiles". en. In: *Annual Review of Statistics and Its Application* 10.1 (Mar. 2023), pp. 597–621. ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-032921-020240. URL: https://www.annualreviews.org/doi/10.1146/annurev-statistics-032921-020240 (visited on 02/02/2025).

[43] Sam Allen, Christopher A. T. Ferro, and Frank Kwasniok. "A conditional decomposition of proper scores: quantifying the sources of information in a forecast". en. In: *Quarterly Journal of the Royal Meteorological Society* 149.754 (July 2023), pp. 1704–1725. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.4478. URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4478 (visited on 04/03/2025).

[44] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, eds. *Algorithmic Learning in a Random World*. eng. SpringerLink Bücher. Boston, MA: Springer Science+Business Media, Inc, 2005. ISBN: 978-0-387-00152-4 978-0-387-25061-8. DOI: 10.1007/b106715.

[45] Anastasios N. Angelopoulos, Emmanuel J. Candes, and Ryan J. Tibshirani. *Conformal PID Control for Time Series Prediction*. arXiv:2307.16895 [cs, eess, stat]. July 2023. URL: http://arxiv.org/abs/2307.16895 (visited on 08/23/2024).

[46] Anastasios N. Angelopoulos and Stephen Bates. *A gentle introduction to conformal prediction and distribution-free uncertainty quantification*. Version Number: 6. 2021. DOI: 10.48550/ARXIV.2107.07511. URL: https://arxiv.org/abs/2107.07511 (visited on 08/23/2024).

[47] Isaac Gibbs and Emmanuel Candès. *Conformal Inference for Online Prediction with Arbitrary Distribution Shifts*. arXiv:2208.08401 [stat]. Oct. 2023. DOI: 10.48550/arXiv.2208.08401. URL: http://arxiv.org/abs/2208.08401 (visited on 04/03/2025).

[48] Werner Ehm et al. "Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations and Forecast Rankings". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78.3 (June 2016), pp. 505–562. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/rssb.12154. URL: https://academic.oup.com/jrsssb/article/78/3/505/7040984 (visited on 02/25/2025).

[49] Zied Ben Bouallègue, Thomas Haiden, and David S. Richardson. "The diagonal score: Definition, properties, and interpretations". en. In: *Quarterly Journal of the Royal Meteorological Society* 144.714 (July 2018), pp. 1463–1473. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.3293. URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3293 (visited on 01/28/2025).

[50] Tim Palmer and David Richardson. "Decisions, decisions. . . !" In: (2014). Publisher: ECMWF. DOI: 10.21957/BYCHJ3CF. URL: https://www.ecmwf.int/node/17333 (visited on 01/25/2025).

[51] Erez Buchweitz, João Vitor Romano, and Ryan J. Tibshirani. *Asymmetric Penalties Underlie Proper Loss Functions in Probabilistic Forecasting*. arXiv:2505.00937 [math]. May 2025. DOI: 10.48550/arXiv.2505.00937. URL: http://arxiv.org/abs/2505.00937 (visited on 06/01/2025).

[52] Thomas J. Kniesner and W. Kip Viscusi. "The Value of a Statistical Life". en. In: *Oxford Research Encyclopedia of Economics and Finance*. July 2019. ISBN: 978-0-19-062597-9. DOI: 10.1093/acrefore/9780190625979.013.138. URL: https://oxfordre.com/economics/display/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-138 (visited on 05/22/2025).

[53] Christopher McCabe, Karl Claxton, and Anthony J. Culyer. "The NICE cost-effectiveness threshold: what it is and what that means". eng. In: *PharmacoEconomics* 26.9 (2008), pp. 733–744. ISSN: 1170-7690. DOI: 10.2165/00019053-200826090-00004.

[54] Nicholas J. Irons and Adrian E. Raftery. *US COVID-19 school closure was not cost-effective, but other measures were*. arXiv:2411.12016 [stat]. Nov. 2024. DOI: 10.48550/arXiv.2411.12016. URL: http://arxiv.org/abs/2411.12016 (visited on 05/22/2025).

[55] Matt J. Keeling et al. *Cost-effectiveness of routine COVID-19 adult vaccination programmes in England*. en. Nov. 2024. DOI: 10.1101/2024.11.08.24316972. URL: http://medrxiv.org/lookup/doi/10.1101/2024.11.08.24316972 (visited on 03/03/2025).

[56] Siyu Chen et al. "Mass testing for discovery and control of COVID-19 outbreaks in adult social care: an observational study and cost-effectiveness analysis of 14 805 care homes in England". en. In: *BMJ Public Health* 3.1 (Mar. 2025). Publisher: BMJ Publishing Group Ltd. ISSN: 2753-4294. DOI: 10.1136/bmjph-2024-001376. URL: https://bmjpublichealth.bmj.com/content/3/1/e001376 (visited on 04/17/2025).

[57] Hugo C. Turner et al. "An economic evaluation of Wolbachia deployments for dengue control in Vietnam". en. In: *PLOS Neglected Tropical Diseases* 17.5 (May 2023). Publisher: Public Library of Science, e0011356. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0011356. URL: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0011356 (visited on 06/02/2025).

[58] Wu Zeng et al. "Cost-effectiveness of dengue vaccination in ten endemic countries". en. In: *Vaccine* 36.3 (Jan. 2018), pp. 413–420. ISSN: 0264410X. DOI: 10.1016/j.vaccine.2017.11.064. URL: https://linkinghub.elsevier.com/retrieve/pii/S0264410X17316638 (visited on 06/02/2025).

[59] Timo Dimitriadis et al. "Evaluating probabilistic classifiers: The triptych". en. In: *International Journal of Forecasting* 40.3 (July 2024), pp. 1101–1122. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2023.09.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207023000997 (visited on 01/26/2025).

[60] Ilan Price et al. "Probabilistic weather forecasting with machine learning". en. In: *Nature* (Dec. 2024). ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-024-08252-9. URL: https://www.nature.com/articles/s41586-024-08252-9 (visited on 12/05/2024).

[61] Nicolas Banholzer et al. *A comparison of short-term probabilistic forecasts for the incidence of COVID-19 using mechanistic and statistical time series models*. arXiv:2305.00933 [cs, q-bio, stat]. May 2023. URL: http://arxiv.org/abs/2305.00933 (visited on 10/08/2024).

[62] Emilie Finch et al. *Climate variation and serotype competition drive dengue outbreak dynamics in Singapore*. en. Sept. 2024. DOI: 10.1101/2024.09.17.24313793. URL: http://medrxiv.org/lookup/doi/10.1101/2024.09.17.24313793 (visited on 11/28/2024).

[63] Tilmann Gneiting and Peter Vogel. "Receiver operating characteristic (ROC) curves: equivalences, beta model, and minimum distance estimation". en. In: *Machine Learning* 111.6 (June 2022), pp. 2147–2159. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-021-06115-2. URL: https://link.springer.com/10.1007/s10994-021-06115-2 (visited on 03/28/2025).

[64] R Tyrrell Rockafellar, Stanislav Uryasev, et al. "Optimization of conditional value-at-risk". In: *Journal of risk* 2 (2000). Publisher: Citeseer, pp. 21–42.

[65] Hamed Rahimian and Sanjay Mehrotra. "Frameworks and results in distributionally robust optimization". In: *Open Journal of Mathematical Optimization* 3 (2022), pp. 1–85.

[66] Stefan Weber. "Distribution-invariant risk measures, information, and dynamic consistency". In: *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 16.2 (2006). Publisher: Wiley Online Library, pp. 419–441.

[67] Tilmann Gneiting et al. "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation". en. In: *Monthly Weather Review* 133.5 (May 2005), pp. 1098–1118. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/MWR2904.1. URL: http://journals.ametsoc.org/doi/10.1175/MWR2904.1 (visited on 09/03/2024).

[68] Tobias Fissler and Johanna F. Ziegel. "Higher order elicitability and Osband's principle". In: *The Annals of Statistics* 44.4 (2016). Publisher: Institute of Mathematical Statistics, pp. 1680–1707. DOI: 10.1214/16-AOS1439. URL: https://doi.org/10.1214/16-AOS1439.

[69] Samuel V. Scarpino and Giovanni Petri. "On the predictability of infectious disease outbreaks". en. In: *Nature Communications* 10.1 (Feb. 2019), p. 898. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08616-0. URL: https://www.nature.com/articles/s41467-019-08616-0 (visited on 02/06/2025).

[70] C. E. Shannon. "A Mathematical Theory of Communication". en. In: *Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. ISSN: 00058580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://ieeexplore.ieee.org/document/6773024 (visited on 05/15/2025).

[71] Richard W. Katz and Jeffrey K. Lazo. "Economic Value of Weather and Climate Forecasts". In: *The Oxford Handbook of Economic Forecasting*. Ed. by Michael P. Clements and David F. Hendry. Oxford University Press, July 2011, p. 0. ISBN: 978-0-19-539864-9. DOI: 10.1093/oxfordhb/9780195398649.013.0021. URL: https://doi.org/10.1093/oxfordhb/9780195398649.013.0021 (visited on 05/21/2025).

[72] Karl Pearson. "Note on Regression and Inheritance in the Case of Two Parents". In: *Proceedings of the Royal Society of London Series I* 58 (Jan. 1895). ADS Bibcode: 1895RSPS...58..240P, pp. 240–242. URL: https://ui.adsabs.harvard.edu/abs/1895RSPS...58..240P (visited on 03/23/2025).

[73] Hans Heesterbeek et al. "Modeling infectious disease dynamics in the complex landscape of global health". In: *Science* 347.6227 (Mar. 2015). Publisher: American Association for the Advancement of Science, aaa4339. DOI: 10.1126/science.aaa4339. URL: https://www.science.org/doi/full/10.1126/science.aaa4339 (visited on 04/23/2025).

[74] Inga Holmdahl and Caroline Buckee. "Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us". en. In: *New England Journal of Medicine* 383.4 (July 2020), pp. 303–305. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMp2016822. URL: http://www.nejm.org/doi/10.1056/NEJMp2016822 (visited on 04/20/2025).

[75] Oliver Morgan. "How decision makers can use quantitative approaches to guide outbreak responses". eng. In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 374.1776 (July 2019), p. 20180365. ISSN: 1471-2970. DOI: 10.1098/rstb.2018.0365.

[76] Alexis Robert et al. "Predicting subnational incidence of COVID-19 cases and deaths in EU countries". en. In: *BMC Infectious Diseases* 24.1 (Feb. 2024), p. 204. ISSN: 1471-2334. DOI: 10.1186/s12879-024-08986-x. URL: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-024-08986-x (visited on 02/12/2025).

[77] Rob J. Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. eng. Third print edition. Melbourne, Australia: Otexts, Online Open-Access Textbooks, 2021. ISBN: 978-0-9875071-3-6.

[78] Paul-Christian Bürkner, Jonah Gabry, and Aki Vehtari. "Approximate leave-future-out cross-validation for Bayesian time series models". en. In: *Journal of Statistical Computation and Simulation* 90.14 (Sept. 2020), pp. 2499–2523. ISSN: 0094-9655, 1563-5163. DOI: 10.1080/00949655.2020.1783262. URL: https://www.tandfonline.com/doi/full/10.1080/00949655.2020.1783262 (visited on 05/31/2025).

[79] Tilmann Gneiting and Johannes Resin. "Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination". In: *Electronic Journal of Statistics* 17.2 (Jan. 2023). ISSN: 1935-7524. DOI: 10.1214/23-EJS2180. URL: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-17/issue-2/Regression-diagnostics-meets-forecast-evaluation--conditional-calibration-reliability-diagrams/10.1214/23-EJS2180.full (visited on 02/15/2025).

[80] Cabinet Office. *Independent review into the response to the 2009 swine flu pandemic*. Tech. rep. Cabinet Office, July 2010.

[81] Graham F. Medley. "A consensus of evidence: The role of SPI-M-O in the UK COVID-19 response". In: *Advances in Biological Regulation*. COVID-19 models and expectations 86 (Dec. 2022), p. 100918. ISSN: 2212-4926. DOI: 10.1016/j.jbior.2022.100918. URL: https://www.sciencedirect.com/science/article/pii/S2212492622000586 (visited on 03/03/2025).

[82] Charles Blackorby, Walter Bossert, and David J. Donaldson. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. 1st ed. Cambridge University Press, Aug. 2005. ISBN: 978-0-521-53258-7 978-0-521-82551-1 978-1-139-05224-5. DOI: 10.1017/CCOL0521825512. URL: https://www.cambridge.org/core/product/identifier/9781139052245/type/book (visited on 02/24/2025).

[83] Kenneth J. Arrow. "A Difficulty in the Concept of Social Welfare". en. In: *Journal of Political Economy* 58.4 (Aug. 1950), pp. 328–346. ISSN: 0022-3808, 1537-534X. DOI: 10.1086/256963. URL: https://www.journals.uchicago.edu/doi/10.1086/256963 (visited on 02/24/2025).

[84] Ruth McCabe et al. "Communicating uncertainty in epidemic models". en. In: *Epidemics* 37 (Dec. 2021), p. 100520. ISSN: 17554365. DOI: 10.1016/j.epidem.2021.100520. URL: https://linkinghub.elsevier.com/retrieve/pii/S1755436521000669 (visited on 08/18/2024).

[85] Michael C. Runge et al. "Scenario design for infectious disease projections: Integrating concepts from decision analysis and experimental design". en. In: *Epidemics* 47 (June 2024), p. 100775. ISSN: 17554365. DOI: 10.1016/j.epidem.2024.100775. URL: https://linkinghub.elsevier.com/retrieve/pii/S1755436524000367 (visited on 02/12/2025).

[86] Anne Cori and Adam Kucharski. "Inference of epidemic dynamics in the COVID-19 era and beyond". en. In: *Epidemics* 48 (Sept. 2024), p. 100784. ISSN: 17554365. DOI: 10.1016/j.epidem.2024.100784. URL: https://linkinghub.elsevier.com/retrieve/pii/S1755436524000458 (visited on 10/09/2024).

[87] Grace Liu et al. "Binary climate data visuals amplify perceived impact of climate change". en. In: *Nature Human Behaviour* (Apr. 2025). ISSN: 2397-3374. DOI: 10.1038/s41562-025-02183-9. URL: https://www.nature.com/articles/s41562-025-02183-9 (visited on 05/12/2025).

[88] Claudia R. Schneider et al. "The effects of communicating scientific uncertainty on trust and decision making in a public health context". en. In: *Judgment and Decision Making* 17.4 (July 2022), pp. 849–882. ISSN: 1930-2975. DOI: 10.1017/S1930297500008962. URL: https://www.cambridge.org/core/product/identifier/S1930297500008962/type/journal_article (visited on 05/02/2025).

[89] Robert J. Taggart and David J. Wilke. *Warnings based on risk matrices: a coherent framework with consistent evaluation*. Mar. 2025. DOI: 10.5194/egusphere-2025-323. URL: https://egusphere.copernicus.org/preprints/2025/egusphere-2025-323/ (visited on 05/03/2025).

[90] C. Codeco et al. "Infodengue: A nowcasting system for the surveillance of arboviruses in Brazil". en. In: *Revue d'Épidémiologie et de Santé Publique* 66 (July 2018), S386. ISSN: 03987620. DOI: 10.1016/j.respe.2018.05.408. URL: https://linkinghub.elsevier.com/retrieve/pii/S0398762018311088 (visited on 06/02/2024).

[91] Dung Phung et al. "Advancing adoptability and sustainability of digital prediction tools for climate-sensitive infectious disease prevention and control". en. In: *Nature Communications* 16.1 (Feb. 2025), p. 1644. ISSN: 2041-1723. DOI: 10.1038/s41467-025-56826-6. URL: https://www.nature.com/articles/s41467-025-56826-6 (visited on 02/17/2025).

[92] Liza Hadley et al. "Visual preferences for communicating modelling: a global analysis of COVID-19 policy and decision makers". en. In: *Infectious Disease Modelling* 10.3 (Sept. 2025), pp. 924–934. ISSN: 24680427. DOI: 10.1016/j.idm.2025.04.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468042725000314 (visited on 06/01/2025).

[93] Manit Sittimart et al. "An overview of the perspectives used in health economic evaluations". en. In: *Cost Effectiveness and Resource Allocation* 22.1 (May 2024), p. 41. ISSN: 1478-7547. DOI: 10.1186/s12962-024-00552-1. URL: https://resource-allocation.biomedcentral.com/articles/10.1186/s12962-024-00552-1 (visited on 01/28/2025).

[94] M. J. Rodwell et al. "User decisions, and how these could guide developments in probabilistic forecasting". en. In: *Quarterly Journal of the Royal Meteorological Society* 146.732 (Oct. 2020), pp. 3266–3284. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.3845. URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3845 (visited on 02/10/2025).

[95] Yuyang Chen et al. "Measuring the effects of COVID-19-related disruption on dengue transmission in southeast Asia and Latin America: a statistical modelling study". en. In: *The Lancet Infectious Diseases* 22.5 (May 2022), pp. 657–667. ISSN: 14733099. DOI: 10.1016/S1473-3099(22)00025-1. URL: https://linkinghub.elsevier.com/retrieve/pii/S1473309922000251 (visited on 07/06/2023).

[96] Alexander E. Zarebski et al. *Including frameworks of public health ethics in computational modelling of infectious disease interventions*. arXiv:2502.00071 [physics]. Jan. 2025. DOI: 10.48550/arXiv.2502.00071. URL: http://arxiv.org/abs/2502.00071 (visited on 06/07/2025).

[97] Katriona Shea et al. "Multiple models for outbreak decision support in the face of uncertainty". en. In: *Proceedings of the National Academy of Sciences* 120.18 (May 2023), e2207537120. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2207537120. URL: https://pnas.org/doi/10.1073/pnas.2207537120 (visited on 03/26/2025).

[98] Rob J. Hyndman. "Forecasting, causality and feedback". en. In: *International Journal of Forecasting* 39.2 (Apr. 2023), pp. 558–560. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.09.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022001388 (visited on 01/29/2025).

[99] Christopher J. M. Whitty. "What makes an academic paper useful for health policy?" In: *BMC Medicine* 13.1 (Dec. 2015), p. 301. ISSN: 1741-7015. DOI: 10.1186/s12916-015-0544-8. URL: https://doi.org/10.1186/s12916-015-0544-8 (visited on 04/11/2025).

[100] Kelly Charniga et al. *Best practices for estimating and reporting epidemiological delay distributions of infectious diseases using public health surveillance and healthcare data*. Version Number: 1. 2024. DOI: 10.48550/ARXIV.2405.08841. URL: https://arxiv.org/abs/2405.08841 (visited on 07/01/2024).

[101] Sang Woo Park et al. *Estimating epidemiological delay distributions for infectious diseases*. en. Jan. 2024. DOI: 10.1101/2024.01.12.24301247. URL: http://medrxiv.org/lookup/doi/10.1101/2024.01.12.24301247 (visited on 05/15/2024).

[102] Leonardo S Bastos et al. "A modelling approach for correcting reporting delays in disease surveillance data". en. In: *Statistics in Medicine* 38.22 (Sept. 2019), pp. 4363–4377. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.8303. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.8303 (visited on 03/02/2025).

[103] Kelly Charniga et al. "Nowcasting and forecasting the 2022 U.S. mpox outbreak: Support for public health decision making and lessons learned". In: *Epidemics* 47 (June 2024), p. 100755. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2024.100755. URL: https://www.sciencedirect.com/science/article/pii/S1755436524000161 (visited on 03/02/2025).

[104] Velma K. Lopez et al. *A roadmap to account for reporting delays for public health situational awareness – a case study with COVID-19 and dengue in United States jurisdictions*. en. Nov. 2024. DOI: 10.1101/2024.11.09.24315999. URL: http://medrxiv.org/lookup/doi/10.1101/2024.11.09.24315999 (visited on 05/17/2025).

[105] Yo Joong Choe and Aaditya Ramdas. "Comparing Sequential Forecasters". In: *Operations Research* 72.4 (July 2024). Publisher: INFORMS, pp. 1368–1387. ISSN: 0030-364X. DOI: 10.1287/opre.2021.0792. URL: https://pubsonline.informs.org/doi/10.1287/opre.2021.0792 (visited on 06/01/2025).

[106] Axel Gandy, Kaushik Jana, and Almut E. D. Veraart. "Scoring predictions at extreme quantiles". en. In: *AStA Advances in Statistical Analysis* 106.4 (Dec. 2022), pp. 527–544. ISSN: 1863-818X. DOI: 10.1007/s10182-021-00421-9. URL: https://doi.org/10.1007/s10182-021-00421-9 (visited on 06/01/2025).

[107] Raffaella Giacomini and Ivana Komunjer. "Evaluation and Combination of Conditional Quantile Forecasts". In: *Journal of Business & Economic Statistics* 23.4 (Oct. 2005). Publisher: ASA Website _eprint: https://doi.org/10.1198/073500105000000018, pp. 416–431. ISSN: 0735-0015. DOI: 10.1198/073500105000000018. URL: https://doi.org/10.1198/073500105000000018 (visited on 06/03/2025).

[108] John Stewart Fabila-Carrasco, Chao Tan, and Javier Escudero. *Multivariate permutation entropy, a Cartesian graph product approach*. arXiv:2203.00550 [math]. Mar. 2022. DOI: 10.48550/arXiv.2203.00550. URL: http://arxiv.org/abs/2203.00550 (visited on 05/03/2025).

[109] Xin Huang, Han Lin Shang, and David Pitt. "Permutation entropy and its variants for measuring temporal dependence". en. In: *Australian & New Zealand Journal of Statistics* 64.4 (Dec. 2022), pp. 442–477. ISSN: 1369-1473, 1467-842X. DOI: 10.1111/anzs.12376. URL: https://onlinelibrary.wiley.com/doi/10.1111/anzs.12376 (visited on 02/18/2025).

[110] Ian Lundberg et al. "The origins of unpredictability in life outcome prediction tasks". en. In: *Proceedings of the National Academy of Sciences* 121.24 (June 2024), e2322973121. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2322973121. URL: https://pnas.org/doi/10.1073/pnas.2322973121 (visited on 02/18/2025).

[111] Nikos I. Bosse et al. *Evaluating Forecasts with scoringutils in R*. arXiv:2205.07090 [stat]. Nov. 2024. DOI: 10.48550/arXiv.2205.07090. URL: http://arxiv.org/abs/2205.07090 (visited on 04/14/2025).

[112] Alexander Jordan, Fabian Krüger, and Sebastian Lerch. "Evaluating Probabilistic Forecasts with **scoringRules**". en. In: *Journal of Statistical Software* 90.12 (2019). ISSN: 1548-7660. DOI: 10.18637/jss.v090.i12. URL: http://www.jstatsoft.org/v90/i12/ (visited on 01/04/2024).

[113] Timo Dimitriadis and Alexander I. Jordan. *triptych: Diagnostic Graphics to Evaluate Forecast Performance*. en. Institution: Comprehensive R Archive Network Pages: 0.1.3. Oct. 2023. DOI: 10.32614/CRAN.package.triptych. URL: https://CRAN.R-project.org/package=triptych (visited on 06/07/2025).

[114] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2025. URL: https://www.R-project.org/.