



Cite this article: Howerton E *et al.* 2023 Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology. *J. R. Soc. Interface* **20**: 20220659. <https://doi.org/10.1098/rsif.2022.0659>

Received: 7 September 2022

Accepted: 3 January 2023

Subject Category:

Life Sciences–Mathematics interface

Subject Areas:

computational biology

Keywords:

combination, uncertainty, ensemble, multi-model scenario projections, linear opinion pool, Vincent average

Author for correspondence:

Emily Howerton

e-mail: ehowerton@psu.edu

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.6384904>.

Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology

Emily Howerton¹, Michael C. Runge², Tiffany L. Bogich¹,
Rebecca K. Borchering¹, Hidetoshi Inamine¹, Justin Lessler^{3,4},
Luke C. Mullany⁵, William J. M. Probert⁷, Claire P. Smith⁴, Shaun Truelove^{4,6},
Cécile Viboud⁸ and Katriona Shea¹

¹Department of Biology and Center for Infectious Disease Dynamics, The Pennsylvania State University, University Park, PA, USA

²Eastern Ecological Science Center at the Patuxent Research Refuge, U.S. Geological Survey, Laurel, MD, USA

³Department of Epidemiology and Carolina Population Center, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

⁴Department of Epidemiology, Bloomberg School of Public Health, ⁵Applied Physics Laboratory, and

⁶Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

⁷Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, UK

⁸Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

EH, 0000-0002-0639-3728; MCR, 0000-0002-8081-536X; TLB, 0000-0002-8143-5289;

RKB, 0000-0003-4309-2913; HI, 0000-0002-6367-4630; JL, 0000-0002-9741-8109;

LCM, 0000-0003-4668-9803; WJMP, 0000-0002-3437-759X; CPS, 0000-0003-1069-9460;

ST, 0000-0003-0538-0607; CV, 0000-0003-3243-4711; KS, 0000-0002-7607-8248

Probabilistic predictions support public health planning and decision making, especially in infectious disease emergencies. Aggregating outputs from multiple models yields more robust predictions of outcomes and associated uncertainty. While the selection of an aggregation method can be guided by retrospective performance evaluations, this is not always possible. For example, if predictions are conditional on assumptions about how the future will unfold (e.g. possible interventions), these assumptions may never materialize, precluding any direct comparison between predictions and observations. Here, we summarize literature on aggregating probabilistic predictions, illustrate various methods for infectious disease predictions via simulation, and present a strategy for choosing an aggregation method when empirical validation cannot be used. We focus on the linear opinion pool (LOP) and Vincent average, common methods that make different assumptions about between-prediction uncertainty. We contend that assumptions of the aggregation method should align with a hypothesis about how uncertainty is expressed within and between predictions from different sources. The LOP assumes that between-prediction uncertainty is meaningful and should be retained, while the Vincent average assumes that between-prediction uncertainty is akin to sampling error and should not be preserved. We provide an R package for implementation. Given the rising importance of multi-model infectious disease hubs, our work provides useful guidance on aggregation and a deeper understanding of the benefits and risks of different approaches.

1. Introduction

Predictions about the future are central to describing and managing ecological systems. In these complex and uncertain settings, the decision-making process relies on identifying what could happen in the future and the likelihood of those potential outcomes [1,2]. The optimal policy decisions that use these predictions often depend strongly on the associated uncertainty. For example, a decision maker may be concerned with the risk of exceeding some threshold (e.g. estimating the chance a hospital will exceed bed capacity), or uncertainty may yield a qualitative change in the decision recommendation (e.g. implementing a mask mandate in response to a rise in projected COVID-19 transmission). As such, deliberate processes for appropriately expressing and managing uncertainty are warranted [3].

One such method involves eliciting, and then aggregating, predictions from multiple independent experts or models. In their seminal paper, Bates and Granger demonstrate the benefits of additional information provided by independent predictions, showing that an average of two predictions is more accurate than either alone [4]. Since then, aggregated predictions have been shown to more accurately and more reliably capture future outcomes than a single expert or model. Representing these predictions probabilistically (i.e. defining the probability of possible future events) provides the most complete expression of uncertainty and risk [5]. In the field of infectious diseases, efforts to elicit predictions across multiple mathematical and statistical models are becoming common to support planning and outbreak response [6,7]. Multi-model efforts have been used to predict a range of future public health outcomes (e.g. incident deaths, peak magnitude or epidemic size) [8–12] and estimate intervention effectiveness [13–15].

Importantly, the method used to aggregate multiple predictions has a meaningful, and often substantial, effect on the resulting ensemble. Despite the significance of aggregation methodology and an extensive available literature, identifying the most appropriate method for a given problem can be difficult. Furthermore, which method is most appropriate can depend on one's primary objective (e.g. some objectives proposed by Winkler [16] are overall performance, reliability and robustness to poor performance, and ease of communication and interpretation). Within multi-model infectious disease studies specifically, a range of aggregation methods have been adopted (e.g. averaging probability bins [9], averaging quantiles [12], performance-weighted average [17]). In this work, we review two classes of methods for aggregating multiple probabilistic predictions and illustrate their application in the context of infectious disease dynamics.

Validation of the performance of different aggregation methods against subsequent empirical data can inform methodological choices [18]. However, validation requires feedback between predictions and data that are impractical in many decision contexts. One particularly difficult case is that of scenario projections, namely when decision makers are interested in comparing predictions across multiple possible future situations [19]. In contrast to forecasts, which are predictions about the future (*what will happen*), scenario projections are conditional on a given set of assumptions about how the future will unfold (*what would happen if*); we use 'prediction' as a general term that encompasses both forecasts and projections. Scenarios may specify actions a decision maker could take in the future (e.g. implementation of, and assumed

compliance with, a public health intervention) or uncertainties that are out of the decision maker's control (e.g. emergence of a new virus variant). Because scenarios are not expected to occur exactly as specified, it is unclear how best to compare projections with empirical data [20], complicating potential performance assessment of different aggregation methods.

Similarly, when the timeframe over which a decision must be made is short relative to the availability of data for validation, an aggregation method must be chosen before it is possible to evaluate performance. The relative scales of these timeframes may depend on the natural history of a particular pathogen (e.g. time to symptom presentation in influenza versus tuberculosis) or the decision context (e.g. hospital capacity planning versus setting eradication and elimination targets). Some resource allocation decisions, for example, require quantitative predictions that need to be made well in advance of anticipated need, yet by the time validation of those predictions is possible, the decision-making window has passed.

While eliciting predictions from multiple models provides useful information for infectious disease management, guidance is needed on how to aggregate predictions when there are little to no empirical data available. We address this question from two perspectives: (i) a literature review on the aggregation of probabilistic predictions, and (ii) a simulation analysis aggregating predictions from multiple infectious disease models in which we control the uncertainty within and between models. We end with recommendations for the choice of an aggregation method when empirical validation is lacking and illustrate how the chosen approach should be based on a hypothesis about the expression of uncertainty within and between individual model predictions.

2. Aggregation theory and methodologies

In this section, we summarize existing theory on two classes of methods to aggregate probabilistic predictions. We searched the literature from fields including expert judgement, statistics, operations research, climate modelling and forecasting of weather and economics. Our search terms included variations of 'aggregation', 'combination' or 'ensemble' across multiple databases such as Web of Science and ProQuest. We followed up on relevant articles to the extent possible through references therewithin. As the literature on these topics is vast, it is unlikely our search was exhaustive, but we pursued leads until converging on a set of similar concepts and references. As such, this section is not intended to be a formal, systematic literature review, but instead serves as an overview of concepts necessary to understand our case study.

In addition to summarizing relevant theory, we developed a package, *CombineDistributions*, in the R statistical software for implementing the described methods (<https://github.com/eahowerton/CombineDistributions>). This package supplements other existing packages, including for aggregating point estimates [21], aggregating and evaluating interval forecasts [22] and evaluating probabilistic predictions [23].

2.1. What constitutes a 'good' prediction?

The quality of a probabilistic prediction is often assessed based on calibration [24] (figure 1). *Calibration* describes the consistency between the prediction and the observed data, where better-calibrated predictions more closely reflect future observations. Gneiting & Katzfuss [25] define calibration as 'statistical

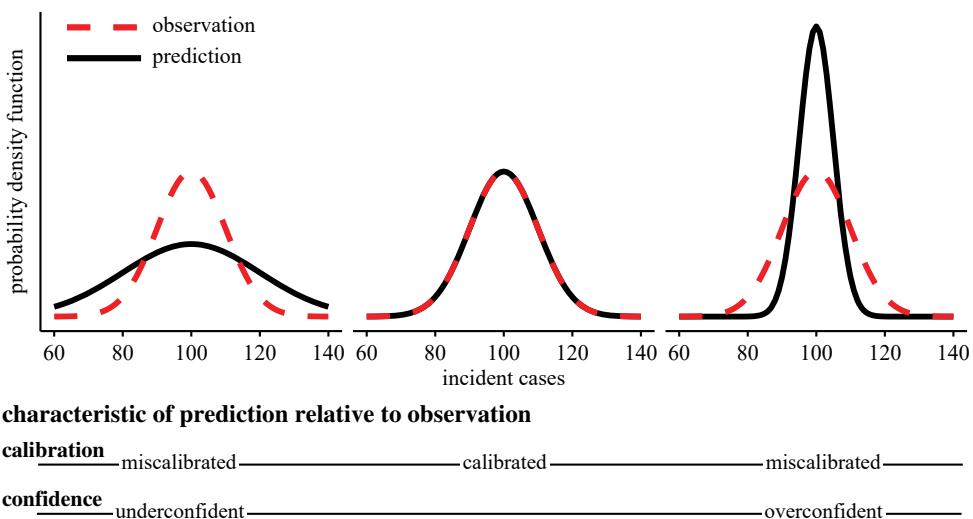


Figure 1. Black distributions represent three probabilistic predictions of incident cases and red dashed lines represent a distribution of ‘true’ observed incident cases (though in practice, we often expect to observe only one realization from this distribution). The calibration and confidence of each prediction is defined. While this figure focuses on prediction variance of unimodal distributions, miscalibration can arise in many ways (e.g. prediction mean).

compatibility of probabilistic forecasts and observations; essentially, realizations should be indistinguishable from random draws from predictive distributions’. Predictions can be miscalibrated because of misaligned mean or variance (or, indeed, higher moments). *Confidence* describes the calibration of the variance specifically. Overconfident distributions are overly concentrated (i.e. predictions are too certain about a future outcome), while underconfident distributions are overly dispersed.

In many settings, however, only a single observation is realized, making it difficult to discern whether a distribution is well calibrated. As such, calibration is typically computed over a series of predictions (e.g. weekly forecasts of incident deaths), and in this case, the sharpness of a distribution is used to distinguish between predictions from different sources that all capture an observation. *Sharpness* describes the width of the predicted distribution without reference to future observations, where narrower distributions are sharper. Together these concepts constitute a standard for determining the quality of a prediction: ‘maximize sharpness, subject to calibration’ [24]. In other words, distributions should be as narrow as possible (maximize sharpness) without sacrificing the ability to capture future observations (subject to calibration). As our discussion of this topic is brief and the literature is vast, we direct readers to a selection of other papers and reviews for more information (see [24–36] for theoretical properties and corresponding metrics for evaluation and [20,37,38] for approaches to evaluation specifically for situations when observations are not readily available).

2.2. How to combine individual predictions?

There are many ways to use multiple sources of quantitative information, including selecting a single source or aggregating across sources. While there are many methods proposed in the literature to perform such aggregation, here, we describe two classes of methods for aggregating probabilistic predictions: methods that operate on probabilities, and methods that operate on quantiles. When operating with an arithmetic mean, these methods are called the linear opinion pool, or probability averaging, and the Vincent method, or quantile averaging (table 1). Both methods yield a valid probability distribution.

The linear opinion pool (LOP) [39] is calculated by taking the arithmetic mean of cumulative probabilities (or, equivalently, probability densities) across alternative predictions for a fixed x value (averaging in the ‘vertical direction’, figure 2a). Alternatively, the Vincent average [42,43] is calculated by taking the arithmetic mean of values across alternative predictions for a fixed quantile (averaging in the ‘horizontal direction’, figure 2b). The median of the Vincent average will be the arithmetic mean of individual model medians. If the predictions to be aggregated are not defined analytically, implementing these methods numerically requires a means of interpolating the cumulative distribution function (CDF) between defined value-quantile pairs.

An array of existing theory describes these two methods, including cases in which each method is preferred. Although both methods produce an aggregate distribution with the same mean [40], there are many ways in which the resulting aggregates differ. The Vincent average will be sharper than the LOP applied to the same set of predictions [40], due to the increase in LOP variance (σ_{LOP}^2 , table 1) as individual predictions become less sharp or more dispersed in central tendency [44]. As individual prediction means are more dispersed, the two methods become increasingly dissimilar [41].

Importantly, theoretical results have shown that using LOP to aggregate a set of well-calibrated distributions will only be well calibrated if the individual CDFs are identical [45–47]. Methods that use metrics other than a weighted average have been proposed to compensate for the overdispersion of the LOP (e.g. logarithmic [41,48], beta-transformed [47], spread-adjusted [45] and generalized [29,49] linear pools). However, in cases where individual distributions are overconfident and somewhat dispersed in central tendency, the simple arithmetic mean LOP can yield a better calibrated distribution because of the increased variance [46].

Unlike the LOP, the Vincent average can preserve distributional shape across predictions for some distributions. Specifically, when all individual predictions are from the same location-scale family (i.e. distributions defined by only location and scale parameters, such as normal, logistic, Cauchy), the Vincent average will also be from that family

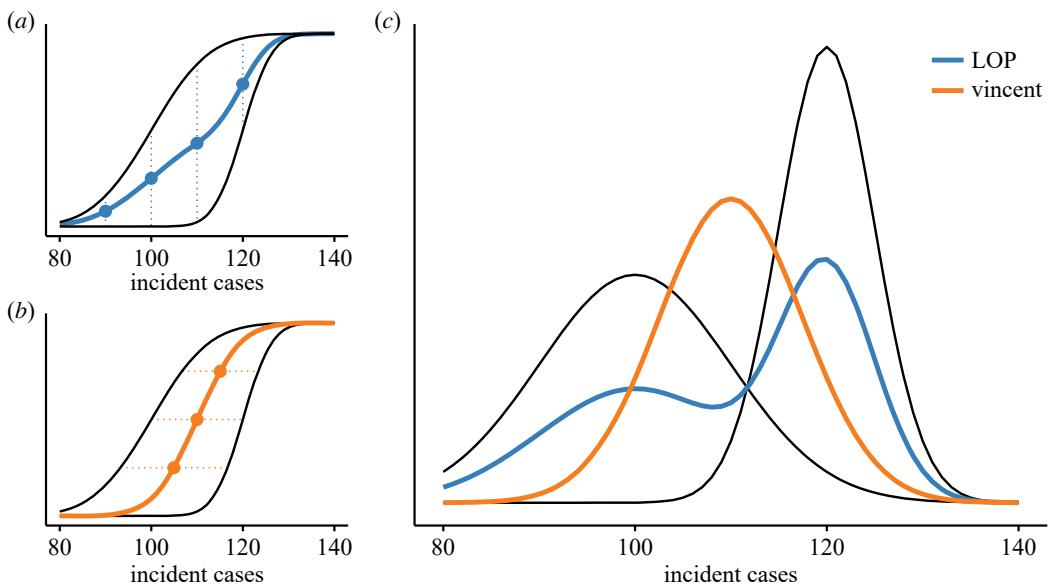


Figure 2. (a) The linear opinion pool (LOP; blue) averages *probabilities* across individual predictions (black) along a single vertical line. (b) The Vincent average (orange) averages *values* across individual predictions (black) along a single horizontal line. (c) Probability density functions for individual predictions (black) and each aggregation method (LOP, blue; Vincent average, orange). In all panels, black lines show individual predictions to be aggregated, $N(\mu = 100, \sigma = 10)$ and $N(\mu = 120, \sigma = 5)$. The Vincent average is equivalent to averaging individual distribution parameters, or $N(\mu = 110, \sigma = 7.5)$. Panels (a) and (b) are modified from [40] and panel (c) from [41].

Table 1. Comparing across averaging techniques, ‘method’ gives the mathematical definition, ‘properties’ gives formulas for mean (μ) and variance (σ^2), and ‘underlying premise’ provides a description of the theoretical basis. Throughout, $F(x)$ is a cumulative distribution function (CDF) defined for values, x , $F^{-1}(\theta)$ is a quantile function (inverse of the CDF) defined for quantiles θ , N is the number of predictions to be aggregated, and w is a weight for averaging. Subscripts $_{\text{LOP}}$, $_{\text{V}}$ or $_{i}$ indicate the linear opinion pool, Vincent average, or individual prediction, respectively.

	linear opinion pool (LOP; probability averaging)	Vincent average (quantile averaging)
method	calculate average cumulative probability F , at each value, x , or $F_{\text{LOP}}(x) = \sum_{i=1}^N w_i F_i(x)$	calculate average value, x , at each quantile, F^{-1} , or $F_{\text{V}}^{-1}(\theta) = \sum_{i=1}^N w_i F_i^{-1}(\theta)$
properties	both methods yield distributions with the same mean, which is the average of individual distribution means $\mu_{\text{LOP}} = \sum_{i=1}^N w_i \mu_i$	$\mu_{\text{V}} = \sum_{i=1}^N w_i \mu_i$
	the variance of the Vincent average will always be less than or equal to the variance of the LOP $\sigma_{\text{LOP}}^2 = \sum_{i=1}^N w_i \sigma_i^2 + \sum_{i=1}^N w_i (\mu_i - \mu_{\text{LOP}})^2$ $= \left(\begin{array}{c} \text{mean of} \\ \text{individual} \\ \text{variances} \end{array} \right) + \left(\begin{array}{c} \text{variance of} \\ \text{individual} \\ \text{means} \end{array} \right)$	although σ_{V}^2 follows no general form, $\sigma_{\text{V}}^2 \leq \sigma_{\text{LOP}}^2$ and $\sigma_{\text{V}}^2 = \sum_{i=1}^N w_i \sigma_i^2$ if all predictions are from the same location-scale family
underlying premise	each prediction captures a possible outcome, and therefore <i>between-prediction uncertainty is retained</i>	each prediction represents a noisy sample from a higher-order distribution, and therefore <i>between-prediction uncertainty is not retained</i>

and the parameters of the aggregate will be the average of individual distribution parameters [50]. When the true distribution is from the same family as the individual distributions, this shape-preserving property can offer calibration benefits [40]. Other reviews provide further discussion of the theoretical properties of these methods [40,41,51].

Importantly, these properties emerge from assumptions underlying each method about how uncertainty is expressed between predictions. Here, and throughout, we use ‘between-

prediction’ uncertainty to refer to the uncertainty captured across the set of independent predictions; this does not imply predictions are related in any way. The LOP treats individual predictions as alternative possible futures across which uncertainty should be retained [52], yielding an aggregate distribution that superimposes the shapes of each individual prediction. The Vincent average assumes individual predictions are each an imperfect representation of a single distribution of interest, appropriately capturing

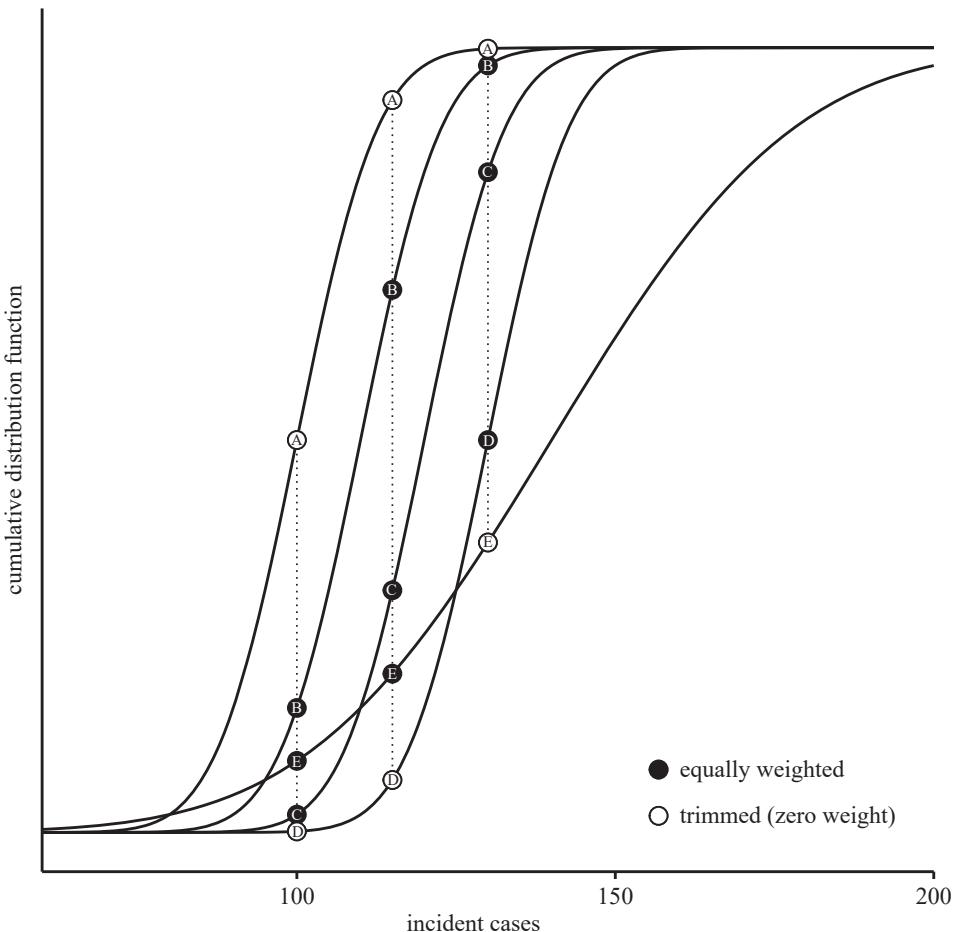


Figure 3. Solid black lines show individual predictions to be aggregated (A: $N(\mu = 100, \sigma = 10)$, B: $N(\mu = 110, \sigma = 10)$, C: $N(\mu = 120, \sigma = 10)$, D: $N(\mu = 130, \sigma = 10)$, E: $N(\mu = 140, \sigma = 30)$). At three values of incident cases (100, 115, 130), circles show which individual predictions were trimmed (open circles) and which were included (closed circles) when aggregating. Figure modified from [58].

uncertainty despite random noise across predictions. As such, the Vincent average cancels high and low predictions, yielding an intermediate aggregate distribution. We demonstrate this philosophical difference by aggregating two normal distributions (figure 2c). The LOP is bi-modal with modes centred at each of the two individual distributions, whereas the Vincent average is centred between the two individual distributions.

2.3. How should predictions be weighted?

In addition to the averaging direction, the weights assigned to each prediction (i.e. w_i in table 1) affect the resulting aggregate distribution. Weighting schemes can recalibrate the aggregate by giving more weight to individual predictions that perform well (e.g. [4,53,54]). Despite the prevalence of sophisticated optimization schemes (e.g. machine learning algorithms [18]), giving all predictions equal weight is surprisingly robust in many applications [55–57] including some infectious disease forecasting settings [54]. However, averaging can be sensitive to outlying predictions, and in the LOP framework, this sensitivity can lead to a highly dispersed aggregate distribution. We describe one method, called trimming, which can address this concern without the need for feedback with observations.

Trimming methods adjust the sharpness of the aggregate distribution by excluding some values (i.e. assigning a weight of zero) and equally weighting all remaining values

[58]. Exterior trimming, which gives zero weight to outermost values, increases the sharpness of the aggregate distribution. This approach is expected to improve performance in cases where the unweighted aggregate is underconfident, including for example, aggregating predictions that vary greatly in central tendency [58]. Aggregating with a median instead of a mean is the most extreme form of exterior trimming, where all values except the centremost are given zero weight. The median LOP and median Vincent average are equivalent in most cases, and the reduced variance of the median LOP can lead to performance improvements in cases where the unweighted (mean) LOP is underconfident [59].

Interior trimming gives zero weight to central values, decreasing the sharpness of the resulting distribution by increasing weight on outer values. This trimming method is appropriate when individual predictions are overconfident and concentrated in central tendency [58]. The effect of trimming a single value depends on the number of predictions being aggregated (e.g. trimming one value in a set of five predictions excludes 20% of available information compared with 2% in a set of 50 predictions).

These trimming strategies depend on a method for determining which values are ‘interior’ or ‘exterior’ (i.e. ranking). Here, we focus on CDF trimming [58] which ranks cumulative probabilities at each value. Figure 3 illustrates exterior CDF trimming, where open circles indicate cumulative probabilities that will be given zero weight in the LOP average. The individual prediction being trimmed can vary across

Table 2. Individual model parameters across three cases of uncertainty expression within and between models. Assumptions are shown for each model (A, B, C, D, E). For simplicity, we assume that all models define $\sigma_\beta = 0.2$ and $1/\gamma \sim N(\mu = 1, \sigma = 0.1)$. For the susceptible-infected-recovered (SIR) model, $\rho = 0$, for the susceptible-infected-recovered-susceptible (SIRS) model, we assume a 26-week mean time to waning (i.e. $\rho = 1/26$), and for models that represent both waning possibilities, SIR and SIRS, we assume $\rho \in \{0, 1/26\}$.

	model	parametric uncertainty	structural uncertainty	
case 1	A	within & between models	$\mu_\beta = 1.2$	none
	B		$\mu_\beta = 1.4$	SIR
	C		$\mu_\beta = 1.6$	SIR
	D		$\mu_\beta = 1.8$	SIR
	E		$\mu_\beta = 2.4$	SIR
case 2	A	within & between models	$\mu_\beta = 1.2$	between models
	B		$\mu_\beta = 1.4$	SIRS
	C		$\mu_\beta = 1.6$	SIR
	D		$\mu_\beta = 1.8$	SIRS
	E		$\mu_\beta = 2.4$	SIR
case 3	A	within & between models	$\mu_\beta = 1.2$	SIR & SIRS
	B		$\mu_\beta = 1.4$	SIR & SIRS
	C		$\mu_\beta = 1.6$	SIR & SIRS
	D		$\mu_\beta = 1.8$	SIR & SIRS
	E		$\mu_\beta = 2.4$	SIR & SIRS

values; for example, the predictions being trimmed are different at 100 and 130 incident cases (A and D versus A and E; figure 3). As a result, CDF trimming often includes information across more individual predictions than alternative trimming methods (e.g. mean trimming, which excludes entire distributions from aggregation based on the mean, details in electronic supplementary material, S1.1). More discussion and examples of trimming are provided in [58,59].

3. Illustrating the choice of aggregation method in a simulated outbreak setting

While it is preferable to evaluate the performance of different aggregation methods based on past empirical data, it is not always possible, and results can be difficult to interpret. In these cases, theory (as outlined in the previous section) provides a useful ground on which to make methodological decisions.

However, the needs of the infectious disease setting may not be fully met by the guidance available in existing theory. Much theory on aggregating probability distributions depends on assumptions about the form of the underlying predictive distribution (e.g. normal distribution), but infectious disease predictions may not be of the studied forms or any particular form at all. Here we illustrate how the choice of aggregation approach can affect the resulting predictions, specifically for an infectious disease setting. We simulate the aggregation of predictions from multiple infectious disease models to predict two disease burden outcomes and investigate the properties of LOP- and Vincent-aggregated distributions. We focus on the equally weighted aggregation methods to illustrate the assumptions of each approach and then explore the properties of exterior trimming when an individual model with outlying assumptions is included in the set.

3.1. Methods

We use different versions of a simple SIRS (susceptible-infected-recovered-susceptible) epidemiological model [60]. In this model, susceptible individuals become infected based on the transmission rate, β . Infected individuals clear infection and move to the recovered class based on the recovery rate, γ . Once recovered, individuals are immune and cannot be reinfected for some time; the immunity of such recovered individuals, however, can wane at rate ρ , moving that individual back to the susceptible class. When $\rho = 0$, the SIRS model is equivalent to an SIR (susceptible-infected-recovered) model for fully immunizing infections (i.e. infections for which immunity does not wane).

Within our simulation, we define four individual models that make predictions about an outbreaking infectious disease in a population of 1000 individuals who are all susceptible at the start of the outbreak. Individual models make a single prediction of cumulative and peak cases that will occur in a 52-week period (i.e. the total number of new infections in that timeframe and the maximum number of new infections in a single week of that timeframe, respectively). Our simulations capture uncertainty about both model parameters (i.e. parametric uncertainty) and model structure (i.e. structural uncertainty) (table 2). We implement parametric uncertainty within and between models. For within-model parametric uncertainty, we let individual models draw the transmission rate from a normal distribution with some mean, μ_β , and variance, σ_β^2 . To implement between-model parametric uncertainty, we assume μ_β varies across models.

We assume the primary source of structural uncertainty is about waning of immunity, specifically whether the infection process follows an SIR ($\rho = 0$) or SIRS (ρ greater than 0) model. To understand the impact of structural uncertainty on aggregation, we consider three cases about structural uncertainty:

- (1) Structural uncertainty is *not represented* across individual models: this case may arise in situations where there is consensus on a particular biological or epidemiological feature. Alternatively, all models may improperly make the same assumption due to a shared ignorance of the system (i.e. unknown unknowns). Here, we represent this case with all models assuming no waning immunity.
- (2) Structural uncertainty is represented *between models*: as there are multiple ways to represent complex biological processes with mathematical prediction models, this case is probably common in multi-model prediction settings. Here, we represent this case with two models assuming waning and two assuming no waning.
- (3) Structural uncertainty is represented *within models*: this case may arise in situations where an uncertainty that is known to be important (but has not yet been resolved) is included by independent models. Here, we represent this case with all individual models assuming with equal probability there will be no waning or waning.

For real multi-model prediction efforts, where there are many uncertainties affecting dynamics, it is most likely that multiple cases will be represented for a single set of predictions.

Within each case, we perform 100 000 stochastic replicates for each model using the chain binomial model [61]. This stochastic simulation framework assumes that the number of individuals transitioning between compartments (e.g. S to I) is randomly drawn from a binomial distribution. The probability of transitioning is based on the rate of that transition (defined by β , γ and ρ). Our simulation uses weekly timesteps, which may obscure early stochastic fade-out dynamics (such dynamics are not the focus of this simulation study).

Then, we summarize the stochastic replicates into a CDF approximated by a set of 999 equally spaced quantiles. We aggregate the CDFs using both LOP and Vincent average (with an equal weighting scheme) and test the performance of each aggregate under different assumptions about the future. To generate 'future observations' against which to test our aggregates, we make assumptions about the 'true' values of each parameter.

We consider realized futures where each individual model has captured μ_β (four futures), and where the true μ_β is the mean of individual model μ_β s (one future). We cross these realized futures with two future waning scenarios, where immunity either wanes or does not wane (corresponding precisely to model assumptions, i.e. no waning implies $\rho = 0$ and waning implies $\rho = 1/26$). For each of these 10 realized futures (5 transmission futures \times 2 waning futures), we generate 1000 synthetic observations against which we measure the performance of each aggregate distribution using the continuous rank probability score (CRPS) [62].

Last, we examine an alternative set of scenarios where an outlier model is included in the set of individual models to assess the utility of trimming. Outliers bias aggregate predictions, especially when there are only a handful of solicited models. But exterior trimming should, in theory, alleviate such biases. To simulate the effects of the outliers on aggregate distributions in an outbreak setting, we included an outlier model with $\mu_\beta = 2.4$, deliberately outside of the range of individual model and truth scenario values. We address the same three structural uncertainty scenarios: the outlier model assumes no waning immunity when structural

uncertainty is not represented or represented between models; and assumes waning and no waning are equally likely when structural uncertainty is represented within models. We again aggregate these (now five) distributions using both LOP and Vincent average. We contrast equal weighting of both the LOP and Vincent average with CDF exterior trimming (excluding the highest and lowest values). We compare the performance of these four aggregate distributions in the 10 truth scenarios (defined in the previous paragraph) using CRPS.

All technical details for this simulation experiment, including model structure, output aggregation, evaluation and outlier case are provided in electronic supplementary material, S2. Code to implement this case study can be found in the SIRS vignette of the CombineDistributions R package.

3.2. Results

Our multi-model simulation illustrates the effect of uncertainty on individual model predictions and the corresponding aggregate under various methods. Predictions of cumulative cases depend on both parametric and structural assumptions of an individual model (figure 4a–c). For all models, cumulative case predictions increase with transmission rate (and corresponding R_0). When a model assumes there is no waning, the population size (1000 individuals) serves as an upper bound on cumulative cases and predictions are left-skewed. Alternatively, models that assume immunity wanes yield multi-modal predictions with alternative states corresponding to stochastic fade-out after the first wave (500–1000 cumulative cases) and endemicity (1500–2000 cumulative cases). In very few cases, simulated epidemics fade out before an outbreak can take off (because our simulation is implemented in weekly timesteps). Although also yielding a multi-modal prediction, models that consider both waning possibilities assign different probabilities to each mode compared with those assuming immunity strictly wanes.

Differences in individual predictions translate into different aggregate distributions across the three structural uncertainty scenarios. In cases where assumptions about model structure are consistent across individual models, the LOP and Vincent average generate similar distributions (figure 4d,f). However, when individual models make different assumptions about model structure, the aggregate distributions are qualitatively different (figure 4e). In this case, although both methods generate multi-modal distributions with one mode at 500–1000 cumulative cases, the second mode in the LOP distribution is between 1500 and 2000 cumulative cases compared with 1000 and 1500 cumulative cases in the Vincent distribution. This second mode of the Vincent average leads to poor performance against possible future observations. Specifically, when compared against 10 possible truth scenarios varying the true mean transmission rate and waning immunity, the LOP (generated when individual models make different assumptions about structural uncertainty) has better CRPS values than the Vincent average for 84% of simulated observations (see §3.1 and electronic supplementary material, S2.2 for additional details).

Predictions of peak cases are less sensitive to structural uncertainty (as subsequent infection waves due to waning are smaller than the initial outbreak). As a result, parametric uncertainty is the major driver of differences between

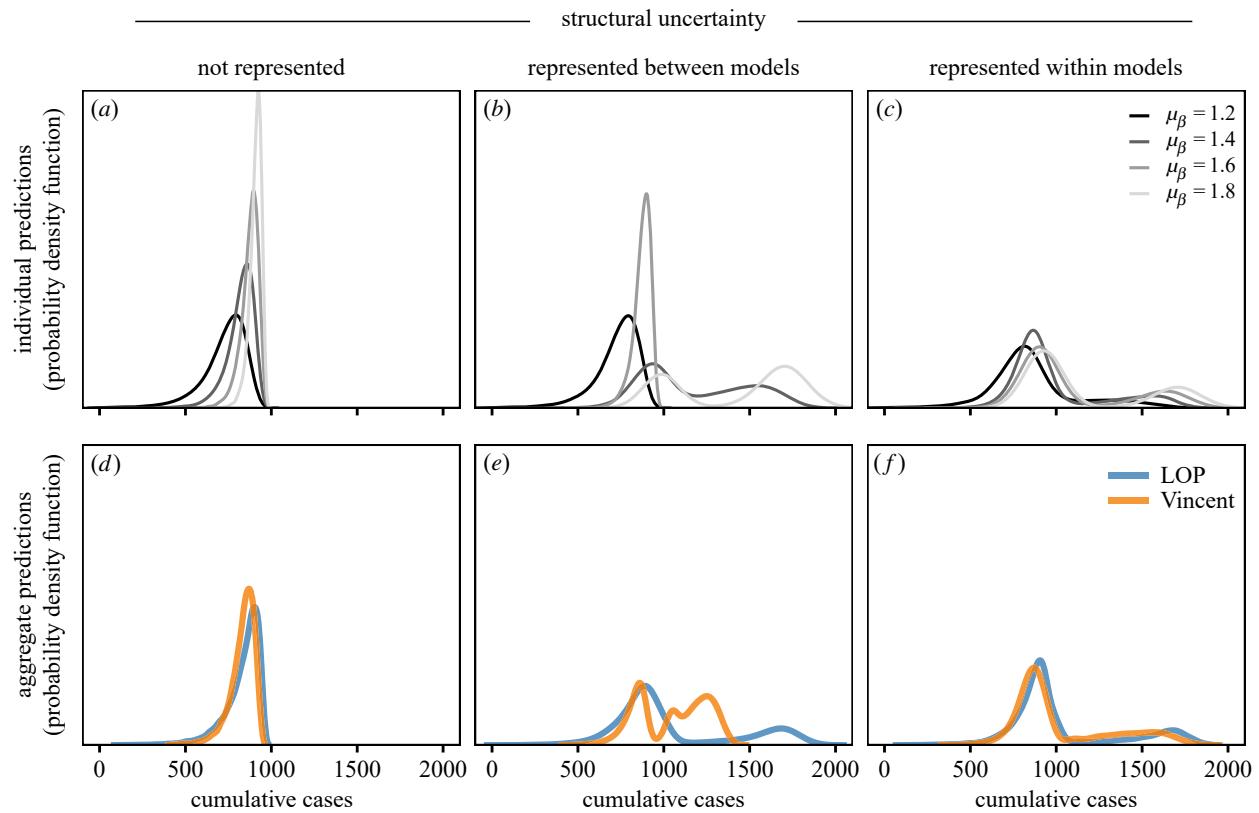


Figure 4. (a–c) Individual model predictions of cumulative cases over a 52-week period, for three assumptions about how structural uncertainty is represented across models. Each curve shows the distribution of cumulative cases generated from 100 000 stochastic replicates for four models that make different assumptions about transmission rate (grey colours, models A–D in table 2). (d–f) Aggregate distributions are shown when each set of four predictions are aggregated using equally weighted linear opinion pool, LOP (blue) and Vincent average (orange).

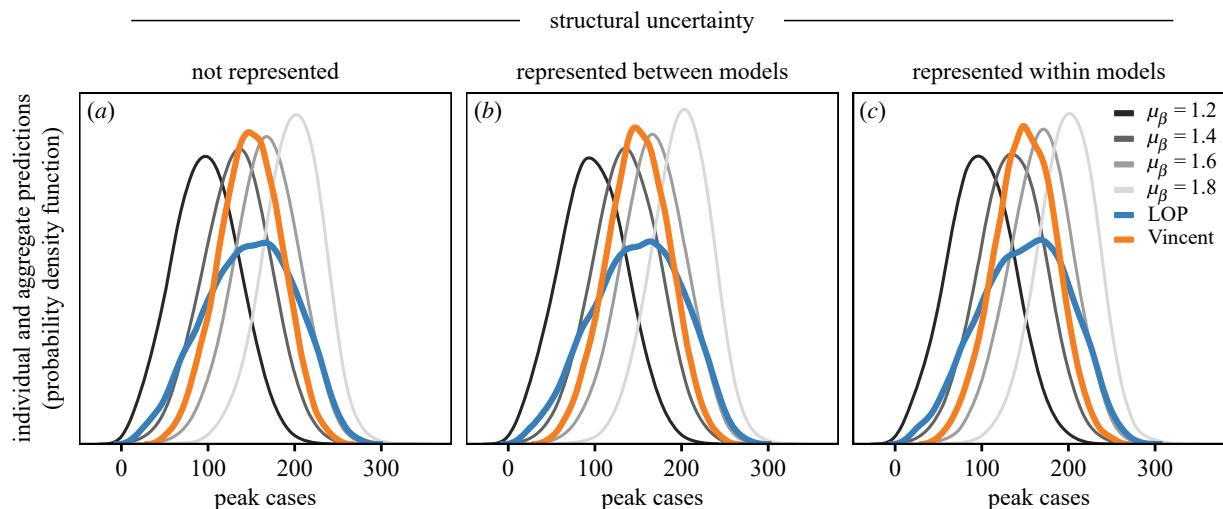


Figure 5. (a–c) Individual model predictions of peak cases over a 52-week period, for three assumptions about how structural uncertainty is represented across models. Each curve shows the distribution of peak cases generated from 100 000 stochastic replicates for four models that make different assumptions about transmission rate (grey colours, models A–D in table 2). Aggregate distributions when each set of four predictions are aggregated using equally weighted linear opinion pool, LOP (blue) and Vincent average (orange).

individual model predictions of peak cases (figure 5a–c). Both LOP and Vincent averages yield distributions with similar central tendency; however, the Vincent average retains the shape of the individual distributions and is sharper than the LOP, resulting in better CRPS values than LOP in 54% of the future observations considered. When the true transmission rate is the mean of individual model transmission rates (i.e. the case where individual model variation is akin to sampling error), the Vincent average performs better for

71% of simulated observations (electronic supplementary material, S2.2).

The presence of an outlier in the set of individual model predictions can affect the aggregate distribution and its performance when all models are weighted equally (our base scenario). Using peak cases as an example, the central tendency of the LOP and Vincent average aggregate distributions are shifted toward the outlier. While the sharpness of the Vincent average is relatively unaffected, the LOP becomes

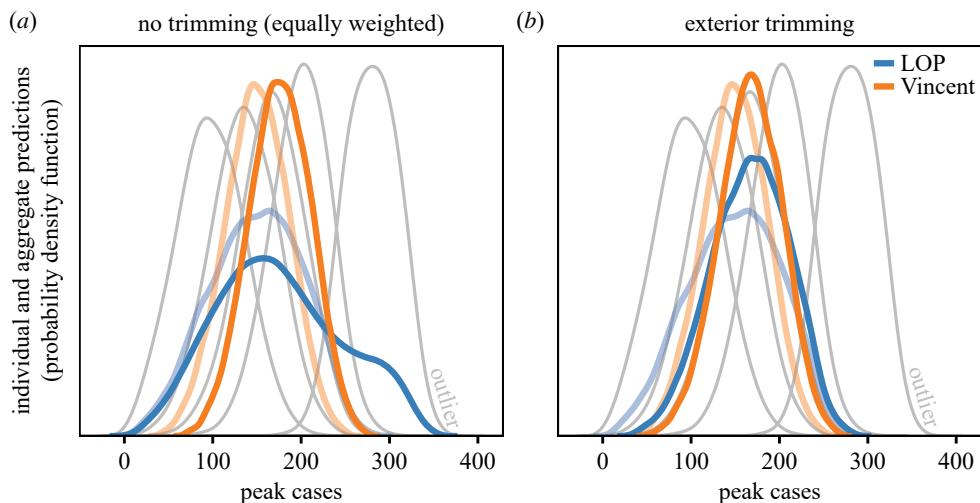


Figure 6. Two versions of LOP (blue) and Vincent average (orange) aggregate distributions for peak cases when an outlier is included in the set of individual predictions; (a) equally weighted (i.e. no trimming) and (b) exterior trimmed. For reference, the five models being aggregated are shown in grey, including the outlier (with $\mu_\beta = 2.4$, model E in table 2) which is labelled as 'outlier'. The other four models match those in figure 5 ($\mu_\beta = 1.2, 1.4, 1.6, 1.8$, models A–D in table 2). The LOP and Vincent average aggregate distributions without an outlier present are shown in lighter blue and orange, respectively. Results are shown for only the case where structural uncertainty is represented between models, as the three structural uncertainty scenarios have very similar predictions of peak cases (see electronic supplementary material, S2.3).

more dispersed with an outlier in the set (figure 6a). Exterior trimming reduces the effect of the outlier, slightly compensating for central tendency and sharpening the resulting LOP aggregate, compared with the untrimmed LOP (figure 6b). The trimmed LOP performs better than the untrimmed LOP for 69% of future observations considered; however, the Vincent average has the best performance across the majority of future observations (30% trimmed and 29% untrimmed). Similar results hold for predictions of cumulative cases, but here, a version of the LOP aggregate performs best for 70% of observations (44% trimmed and 26% untrimmed). See electronic supplementary material, S2.3 for detailed performance results, including the scenarios in which each method performs best, the magnitude of this performance advantage, and the variability of performance over multiple replicates.

4. Discussion

To be most useful for decision making and planning, predictions about the future (or possible futures) need to appropriately capture our current understanding of uncertainty, so the decision makers can assess risk. Representing predictions probabilistically is important for quantifying uncertainty, and aggregating predictions across multiple independent sources can help us more accurately express the uncertainty and incumbent risk.

When aggregating, methodological choices can significantly affect the way uncertainty is expressed. Identifying which aggregation method provides the most appropriate expression of uncertainty is a key challenge, especially in the absence of feedback with observations. When empirical validation is not possible (either not on relevant time scales or not at all, e.g. for scenarios that never materialize), we argue the chosen aggregation approach should align methodological theory with a hypothesis about the uncertainty represented within and between individual model predictions.

In particular, based on our review, theory suggests that when the uncertainty expressed between individual predictions should be retained, the properties and assumptions of

LOP are more appropriate. However, when the between-prediction uncertainty is akin to sampling error and therefore should be averaged away, the Vincent average is better suited and the LOP will be underconfident. Beyond the direction of averaging, weighting schemes provide an additional means to calibrate the aggregate and modulate its variance. In cases where individual predictions vary in levels of confidence (with some probably underconfident) or are overly dispersed in central tendency, exterior trimming (including aggregating with a median) will reduce the variance of the aggregate. Identifying an effective weighting scheme is best achieved when predictions can be directly confronted with observations, yet simple rules may still be preferred [16].

There are challenges to identifying which theoretical case is more appropriate. Details about the assumptions and implementation of individual models may not be available, inhibiting identification of individual and collective expressions of uncertainty. Multiple uncertainties represented within the same set of predictions may need to be treated differently, or predictions may group into subsets based on similar assumptions, requiring one approach within a subset and a different approach across subsets.

Crucially, the choice of aggregation method can have a meaningful impact on decision making. Inappropriately aggregating with LOP will overestimate the probability of extreme events and may suggest more drastic actions than necessary. Conversely, inappropriately using the Vincent average may underestimate these probabilities, leaving decision makers unprepared. Identifying qualitative and quantitative decision outcomes across a range of aggregation methods will reduce the importance of selecting a single best method [52].

To further explore these conclusions in the context of infectious diseases, we used an outbreak simulation case study, where we defined exactly how uncertainty was expressed within each individual model and across the set. When individual models made different assumptions about waning immunity, the LOP retained important between-prediction uncertainty (namely about model structure) for

predictions of cumulative cases. The Vincent average generated an aggregate distribution with a mode between the two expected alternative states (stochastic fade-out or endemicity).

A common criticism of the LOP approach is that it can be underconfident and sensitive to outliers. Here we have shown that a simple exterior trimming procedure can appreciably sharpen the aggregate LOP distribution and restore central tendency when outliers are present, with the trimmed LOP outperforming the untrimmed LOP for the majority of future observations considered. Here, outlier predictions were due to outlying model assumptions about the mean transmission rate, but there are a range of reasons outlier predictions may be generated (e.g. anomalies in data used to generate predictions could cause erroneous predictions). The likelihood of outlying predictions, and the reason these predictions are outliers (i.e. genuine scientific uncertainty versus technical error), may influence the chosen weighting scheme for aggregation.

The effect of each source of uncertainty was not consistent across prediction targets. For predictions about peak cases, variation across individual models was driven primarily by parametric uncertainty rather than structural uncertainty. Predictions of peak cases were not affected by waning immunity (i.e. structural uncertainty) because the largest peak in the SIRS model occurs during the first epidemic wave, before waning takes place. The first peak in the SIRS model is similar to the sole wave of the SIR model. This conclusion, however, may be context dependent as the peak in both models is sensitive to assumptions about initial conditions and parameters and becomes increasingly complicated as transmission rates and waning rates vary in time (e.g. as seen in the COVID-19 pandemic). If we consider individual model assumptions about transmission rate to be variation around a true mean (as we did in two of the ten truth scenarios considered), the sharper Vincent average provides a better representation of uncertainty for predictions of peak cases, including preserving the consistent shape of the individual predictions.

Here, we have presented a simulation study where the models generating individual predictions and future observations are known by design. Yet many additional challenges exist when implementing multi-model aggregation methods in practice. First, our simulated results necessarily are consistent in how outcomes are defined. However, achieving consistency across multiple independent models is not straightforward and failing to do so can lead to discrepancies in results [13]. Expert judgement methods can help to minimize the linguistic uncertainty associated with interpretation differences [3], and when consistency across models is not possible, vote-processing methods can be used to combine decision recommendations from differing sources [63]. In addition, we have assumed that observations were generated by parameters and processes within the range defined by the individual models, which is unlikely to be the case (e.g. truth values were simulated from the same general model structure and stochastic simulation framework as the individual models that generate the aggregate distribution). Identifying failures to

capture observations among all models, including the aggregate, is an important first step to detecting and addressing changes in system dynamics that affect predictions [64].

Accurate estimates of future outcomes and related uncertainty is important for effective infectious disease management, including the integration of formal decision theory into infectious disease applications [65–68]. Aggregating predictions from multiple experts or models has proven to yield better calibrated estimates of future outcomes both for infectious disease dynamics (e.g. [9,17,69–71]) and other fields (e.g. [55]). These methods are becoming increasingly common in infectious disease management; however, many outstanding challenges exist to maximize the utility of these approaches [6,7]. Here we address one of these challenges: in some crucially important decision settings, traditional approaches to selecting an aggregation method (namely, empirical validation) are not feasible on decision-relevant timescales. By providing a theory-based guide to aggregation methodology and extending this theory to infectious disease modelling via simulations, our work provides much needed support of the use of multi-model approaches in public health planning and response.

Data accessibility. The R package to implement methods discussed in this paper can be found in the public repository <https://github.com/eahowerton/CombineDistributions>, <https://doi.org/10.5281/zenodo.7437280>. Code for the case study is available in the SIRS vignette.

Methodological details and additional results are provided in the electronic supplementary material [72].

Authors' contributions. E.H.: conceptualization, data curation, formal analysis, methodology, software, validation, visualization, writing—original draft, writing—review and editing; M.C.R.: conceptualization, methodology, validation, writing—review and editing; T.L.B.: validation, writing—review and editing; R.K.B.: conceptualization, validation, writing—review and editing; H.I.: conceptualization, validation, writing—review and editing; J.L.: validation, writing—review and editing; L.C.M.: validation, writing—review and editing; W.J.M.P.: conceptualization, software, validation, writing—review and editing; C.P.S.: validation, writing—review and editing; S.T.: validation, writing—review and editing; C.V.: conceptualization, validation, writing—review and editing; K.S.: conceptualization, funding acquisition, methodology, validation, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. Justin Lessler has served as an expert witness on cases relating to the COVID-19 pandemic and response.

Funding. This work was supported by the National Science Foundation (COVID-19 RAPID awards DEB-2028301, DEB-2037885 and DEB-2126278), the Eberly College of Science Barbara McClintock Science Achievement Graduate Scholarship in Biology at the Pennsylvania State University, and the Huck Institutes for the Life Sciences. Additional support was received from the US Department of Health and Human Services Centers for Disease Control and Prevention.

Acknowledgements. The authors thank members of the COVID-19 Scenario Modeling Hub for discussion and feedback. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the US Government.

References

1. Clemen RT. 1991 *Making hard decisions: an introduction to decision analysis*. Boston, MA: PWS-Kent Pub. Co.
2. Covello VT. 1987 Decision analysis and risk management decision making: issues and methods. *Risk Anal.* **7**, 131–139. (doi:10.1111/j.1539-6924.1987.tb00978.x)
3. Shea K, Runge MC, Pannell D, Probert WJM, Li S-L, Tildesley M, Ferrari M. 2020 Harnessing multiple models for outbreak management.

Science **368**, 577–579. (doi:10.1126/science.abb9934)

- Bates JM, Granger CWJ. 1969 The combination of forecasts. *OR* **20**, 451–468. (doi:10.2307/3008764)
- Gneiting T. 2008 Editorial: probabilistic forecasting. *J. R. Stat. Soc. Ser. A Stat. Soc.* **171**, 319–321. (doi:10.1111/j.1467-985X.2007.00522.x)
- Reich NG *et al.* 2022 Collaborative hubs: making the most of predictive epidemic modeling. *Am. J. Public Health* **112**, 839–842. (doi:10.2105/AJPH.2022.306831)
- Biggerstaff M, Slayton RB, Johansson MA, Butler JC. 2022 Improving pandemic response: employing mathematical modeling to confront coronavirus disease 2019. *Clin. Infect. Dis.* **74**, 913–917. (doi:10.1093/cid/ciab673)
- Viboud C *et al.* 2018 The RAPIDD ebola forecasting challenge: synthesis and lessons learnt. *Epidemics* **22**, 13–21. (doi:10.1016/j.epidem.2017.08.002)
- Johansson MA *et al.* 2019 An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl Acad. Sci. USA* **116**, 24 268–24 274. (doi:10.1073/pnas.1909865116)
- Reich NG *et al.* 2019 A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl Acad. Sci. USA* **116**, 3146–3154. (doi:10.1073/pnas.2200703119)
- McGowan CJ *et al.* 2019 Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Sci. Rep.* **9**, 683. (doi:10.1038/s41598-018-36361-9)
- Ray EL *et al.* 2020 Ensemble forecasts of Coronavirus disease 2019 (COVID-19) in the U.S. *medRxiv*, 2020.08.19.20177493. (doi:10.1101/2020.08.19.20177493)
- Shea K *et al.* 2020 COVID-19 reopening strategies at the county level in the face of uncertainty: Multiple Models for Outbreak Decision Support. *medRxiv*, 2020.11.03.20225409. (doi:10.1101/2020.11.03.20225409)
- Borchering RK *et al.* 2021 Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios — United States, April–September 2021. *Morb. Mortal. Wkly. Rep.* **70**, 719–724. (doi:10.15585/mmwr.mm7019e3)
- Li X *et al.* 2021 Estimating the health impact of vaccination against ten pathogens in 98 low-income and middle-income countries from 2000 to 2030: a modelling study. *Lancet* **397**, 398–408. (doi:10.1016/S0140-6736(20)32657-X)
- Winkler RL. 2015 Equal versus differential weighting in combining forecasts. *Risk Anal.* **35**, 16–18. (doi:10.1111/risa.12302)
- Reich NG *et al.* 2019 Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLOS Comput. Biol.* **15**, e1007486. (doi:10.1371/journal.pcbi.1007486)
- McAndrew T, Wattanachit N, Gibson GC, Reich NG. 2021 Aggregating predictions from experts: a review of statistical methods, experiments, and applications. *WIREs Comput. Stat.* **13**, e1514. (doi:10.1002/wics.1514)
- Vollmar HC, Ostermann T, Redaelli M. 2015 Using the scenario method in the context of health and health care – a scoping review. *BMC Med. Res. Methodol.* **15**, 89. (doi:10.1186/s12874-015-0083-1)
- Baumberger C, Knutti R, Hadorn GH. 2017 Building confidence in climate model projections: an analysis of inferences from fit. *WIREs Clim. Change* **8**, e454. (doi:10.1002/wcc.454)
- Weiss CE, Raviv E, Roetzer G. 2019 Forecast combinations in R using the ForecastComb package. *R J.* **10**, 262. (doi:10.32614/RJ-2018-052)
- Wang SY *et al.* 2021 Code for: reichlab/covidHubUtils: repository release for Zenodo. *Zenodo*. (doi:10.5281/zenodo.5207940)
- Jordan A, Krüger F, Lerch S. 2019 Evaluating probabilistic forecasts with scoringRules. *J. Stat. Softw.* **90**, 1–37. (doi:10.18637/jss.v090.i12)
- Gneiting T, Balabdaoui F, Raftery AE. 2007 Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 243–268. (doi:10.1111/j.1467-9868.2007.00587.x)
- Gneiting T, Katzfuss M. 2014 Probabilistic forecasting. *Annu. Rev. Stat. Its Appl.* **1**, 125–151. (doi:10.1146/annurev-statistics-062713-085831)
- Simonis JL, White EP, Ernest SKM. 2021 Evaluating probabilistic ecological forecasts. *Ecology* **102**, e03431. (doi:10.1002/ecy.3431)
- Winkler RL *et al.* 1996 Scoring rules and the evaluation of probabilities. *Test* **5**, 1–60. (doi:10.1007/BF02562681)
- Bracher J, Ray EL, Gneiting T, Reich NG. 2021 Evaluating epidemic forecasts in an interval format. *PLOS Comput. Biol.* **17**, e1008618. (doi:10.1371/journal.pcbi.1008618)
- Dawid AP. 1984 Present position and potential developments: some personal views: statistical theory: the prequential approach. *J. R. Stat. Soc. Ser. Gen.* **147**, 278–292. (doi:10.2307/2981683)
- Gneiting T, Raftery AE. 2007 Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378. (doi:10.1198/016214506000001437)
- Gneiting T, Ranjan R. 2011 Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econ. Stat.* **29**, 411–422.
- Parry M, Dawid AP, Lauritzen S. 2012 Proper local scoring rules. *Ann. Stat.* **40**, 561–592. (doi:10.1214/12-AOS971)
- Murphy AH. 1993 What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecast* **8**, 281–293. (doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2)
- Mitchell J, Wallis KF. 2011 Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. *J. Appl. Econom.* **26**, 1023–1040. (doi:10.1002/jae.1192)
- Jose VRR, Winkler RL. 2009 Evaluating quantile assessments. *Oper. Res.* **57**, 1287–1297.
- Hora SC. 2010 An analytic method for evaluating the performance of aggregation rules for probability densities. *Oper. Res.* **58**, 1440–1449. (doi:10.1287/opre.1100.0789)
- Knutti R. 2008 Should we believe model predictions of future climate change? *Phil. Trans. Math. Phys. Eng. Sci.* **366**, 4647–4664. (doi:10.1098/rsta.2008.0169)
- Tebaldi C, Knutti R. 2007 The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Math. Phys. Eng. Sci.* **365**, 2053–2075. (doi:10.1098/rsta.2007.2076)
- Stone M. 1961 The opinion pool. *Ann. Math. Stat.* **32**, 1339–1342.
- Lichtendahl KC, Grushka-Cockayne Y, Winkler RL. 2013 Is it better to average probabilities or quantiles? *Manag. Sci.* **59**, 1594–1611. (doi:10.1287/mnsc.1120.1667)
- Busseti F. 2017 Quantile aggregation of density forecasts. *Oxf. Bull. Econ. Stat.* **79**, 495–512. (doi:10.1111/obes.12163)
- Vincent SB. 1912 The function of the vibrissae in the behavior of the white rat. PhD thesis, University of Chicago.
- Ratcliff R. 1979 Group reaction time distributions and an analysis of distribution statistics. *Psychol. Bull.* **86**, 446–461. (doi:10.1037/0033-2909.86.3.446)
- Wallis KF. 2005 Combining density and interval forecasts: a modest proposal. *Oxf. Bull. Econ. Stat.* **67**, 983–994. (doi:10.1111/j.1468-0084.2005.00148.x)
- Gneiting T, Ranjan R. 2013 Combining predictive distributions. *Electron. J. Stat.* **7**, 1747–1782. (doi:10.1214/13-EJS823)
- Hora SC. 2004 Probability judgments for continuous quantities: linear combinations and calibration. *Manag. Sci.* **50**, 597–604.
- Ranjan R, Gneiting T. 2010 Combining probability forecasts. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72**, 71–91. (doi:10.1111/j.1467-9868.2009.00726.x)
- Genest C, Weerahandi S, Zidek JV. 1984 Aggregating opinions through logarithmic pooling. *Theory Decis.* **17**, 61–70. (doi:10.1007/BF00140056)
- Grushka-Cockayne Y, Lichtendahl KC, Jose VRR, Winkler RL. 2017 Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Oper. Res.* **65**, 712–728. (doi:10.1287/opre.2017.1588)
- Thomas EAC, Ross BH. 1980 On appropriate procedures for combining probability distributions within the same family. *J. Math. Psychol.* **21**, 136–152. (doi:10.1016/0022-2496(80)90003-6)
- Aastveit KA, Mitchell J, Ravazzolo F, van Dijk HK. 2018 The evolution of forecast density combinations in economics. *Tinbergen Institute Discussion Paper*, no. TI 2018-069/III. Amsterdam, The Netherlands: Tinbergen Institute. (doi:10.1093/acrefore/9780190625979.013.381)
- Athanassoglou S, Bosetti V. 2015 Setting environmental policy when experts disagree. *Environ. Resour. Econ.* **61**, 497–516. (doi:10.1007/s10640-014-9804-x)
- Cooke RM. 2015 The aggregation of expert judgment: do good things come to those who weight? *Risk Anal.* **35**, 12–15. (doi:10.1111/risa.12353)

54. Ray EL *et al.* 2022 Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *Int. J. Forecast.* (doi:10.1016/j.ijforecast.2022.06.005)

55. Clemen RT. 1989 Combining forecasts: a review and annotated bibliography. *Int. J. Forecast.* **5**, 559–583. (doi:10.1016/0169-2070(89)90012-5)

56. Armstrong JS. 2001 Combining forecasts. In *Principles of forecasting: a handbook for researchers and practitioners*, pp. 417–439. Boston, MA: Springer.

57. Arieli I, Babichenko Y, Smorodinsky R. 2018 Robust forecast aggregation. *Proc. Natl Acad. Sci. USA* **115**, E12135–E12143. (doi:10.1073/pnas.1813934115)

58. Jose VRR, Grushka-Cockayne Y, Lichtendahl KC. 2014 Trimmed opinion pools and the crowd's calibration problem. *Manag. Sci.* **60**, 463–475. (doi:10.1287/mnsc.2013.1781)

59. Hora SC, Fransen BR, Hawkins N, Susel I. 2013 Median aggregation of distribution functions. *Decis. Anal.* **10**, 279–291. (doi:10.1287/deca.2013.0282)

60. Keeling MJ, Rohani P. 2008 *Modeling infectious diseases in humans and animals*. Princeton, NJ: Princeton University Press.

61. Bailey NTJ. 1957 *The mathematical theory of epidemics*. London, UK: C. Griffin.

62. Matheson JE, Winkler RL. 1976 Scoring rules for continuous probability distributions. *Manag. Sci.* **22**, 1087–1096. (doi:10.1287/mnsc.22.10.1087)

63. Probert WJM, Nicol S, Ferrari MJ, Li S-L, Shea K, Tildesley MJ, Runge MC. 2022 Vote-processing rules for combining control recommendations from multiple models. *Phil. Trans. R. Soc. A* **380**, 20210314. (doi:10.1098/rsta.2021.0314)

64. Runge MC, Stroeve JC, Barrett AP, McDonald-Madden E. 2016 Detecting failure of climate predictions. *Nat. Clim. Change* **6**, 861–864. (doi:10.1038/nclimate3041)

65. Webb CT *et al.* 2017 Ensemble modelling and structured decision-making to support Emergency Disease Management. *Prev. Vet. Med.* **138**, 124–133. (doi:10.1016/j.prevetmed.2017.01.003)

66. Shearer FM, Moss R, McVernon J, Ross JV, McCaw JM. 2020 Infectious disease pandemic planning and response: incorporating decision analysis. *PLoS Med.* **17**, e1003018. (doi:10.1371/journal.pmed.1003018)

67. Berger L, Berger N, Bosetti V, Gilboa I, Hansen LP, Jarvis C, Marinacci M, Smith RD. 2021 Rational policymaking during a pandemic. *Proc. Natl Acad. Sci. USA* **118**, e2012704118. (doi:10.1073/pnas.2012704118)

68. Baker CM *et al.* 2022 From climate change to pandemics: decision science can help scientists have impact. *Front. Ecol. Evol.* **10**, 792749. (doi:10.3389/fevo.2022.792749)

69. Cramer EY *et al.* 2022 Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc. Natl Acad. Sci. USA* **119**, e2113561119. (doi:10.1073/pnas.2113561119)

70. Oidtmann RJ *et al.* 2021 Trade-offs between individual and ensemble forecasts of an emerging infectious disease. *Nat. Commun.* **12**, 5379. (doi:10.1038/s41467-021-25695-0)

71. Yamana TK, Kandula S, Shaman J. 2016 Superensemble forecasts of dengue outbreaks. *J. R. Soc. Interface* **13**, 20160410. (doi:10.1098/rsif.2016.0410)

72. Howerton E *et al.* 2023 Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology. Figshare. (doi:10.6084/m9.figshare.c.6384904)