Evaluating infectious disease forecasts with allocation scoring rules

Aaron Gerding, Nicholas G. Reich, Benjamin Rogers, Evan L. Ray

February 9, 2024

Abstract

Recent years have seen increasing efforts to forecast infectious disease burdens, with a primary goal being to help public health workers make informed policy decisions. However, there has only been limited discussion of how predominant forecast evaluation metrics might indicate the success of policies based in part on those forecasts. We explore one possible tether between forecasts and policy: the allocation of limited medical resources so as to minimize unmet need. We use probabilistic forecasts of disease burden in each of several regions to determine optimal resource allocations, and then we score forecasts according to how much unmet need their associated allocations would have allowed. We illustrate with forecasts of COVID-19 hospitalizations in the US, and we find that the forecast skill ranking given by this allocation scoring rule can vary substantially from the ranking given by the weighted interval score. We see this as evidence that the allocation scoring rule detects forecast value that is missed by traditional accuracy measures and that the general strategy of designing scoring rules that are directly linked to policy performance is a promising direction for epidemic forecast evaluation.

1 Introduction

Infectious disease forecasting models have emerged as important tools in public health outbreak response. The predictions they provide increasingly inform decisions regarding a wide variety of countermeasures intended to reduce transmission and mitigate the severity of disease outcomes. For example, estimates of the onset time of the flu season have been used in developing national vaccination strategies (Igboh et al., 2023), and forecasts of Ebola and diptheria dynamics have been made with the clearly stated goal of helping local public health workers choose the timing and location of interventions in settings where resources are severely constrained (Meltzer et al., 2014; Rainisch et al., 2015; Camacho et al., 2015; Finger et al., 2019). More recently, in the context of the outbreak of COVID-19 across the US, Bertsimas et al. (2021) used forecasts as inputs to decision tools for the interstate reallocation of ventilators and ICU capacity, and to recommend vaccine trial sites to a major trial sponsor. Fox et al. (2022) similarly used predictive models to inform intrastate resource and care site planning, as well as local community guidelines for masking, traveling, dining and shopping (University of Texas, 2022).

In the wake of the COVID-19 pandemic, this trend has been followed by calls for infectious disease forecasts to be not only designed, but also evaluated in ways that align specifically with how forecasts can be used to inform such outbreak control decisions (Marshall et al., 2023; Bilinski et al., 2023). This contrasts, however, with the historically standard practice of measuring the quality of disease forecasts using general purpose accuracy and skill scores, especially those that have implementations

available in existing software when the relevant outbreak occurs. For point forecasts, the root mean square error (RMSE) (e.g., Papastefanopoulos et al. (2020)) and the mean absolute error (MAE) (e.g., Johansson et al. (2016)) are common choices. For probabilistic forecasts, which are the focus of this paper, researchers have often relied on the logarithmic score (LS). For example, the LS has been used to evaluate the skill of US seasonal influenza forecasts (McGowan et al., 2019; Reich et al., 2019) as well as forecasts targeting surveillance measures of dengue incidence in Peru and Puerto Rico (Johansson et al., 2019). More recently, the continuous ranked probability score (CRPS) and a discretized version adapted to multi-quantile forecasts, the weighted interval score (WIS) (Bracher et al., 2021), have gained prominence. For example, the CRPS was used to assess probabilistic forecasts (based on random effect models) of dengue incidence at the district level in Vietnam (Colón-González et al., 2021). And the WIS has been used during the COVID-19 pandemic to evaluate forecasts of observed cases, hospitalizations and deaths in the US and Europe, as reported by municipal, state, and federal surveillance systems (Cramer et al., 2022a; Fox et al., 2022; Sherratt et al., 2023).

While it should be noted that there are ways to interpret any of these scores abstractly through the lens of decision theory, and that all of the application-specific papers cited above benefited from direct collaboration with public health agencies, a key impetus for the present work has been that we were not able to find in any of them, nor in the literature they represent, explicit connections between how a forecast was evaluated and how that forecast was used in practice.

A general phenomenon at play here — one that has been observed repeatedly over the past few decades in other fields such as finance, supply chain management, and meteorology — is that while scores such as RMSE, MAE, LS, CRPS, and WIS can describe the *quality* of a forecast in terms of how well it corresponds to the observed disease outcome, they will often fail to register the *value* of a forecast in the context of a specific decision. We refer the reader to Yardley and Petropoulos (2021) and the references collected therein (especially the foundational Murphy (1993)) for a general overview touching on a wide range of forecasting contexts and also to Pesaran and Skouras (2002) for a clear discussion from an econometric perspective.

Despite this now well-developed discussion of the quality-value distinction in the larger forecasting community, we are aware of only a limited literature attempting to connect the value of *infectious disease* forecasts to their impact on and through policy. And within this body of work we have found the discussion of such a connection to still be at a formally and quantitively imprecise stage. In Ioannidis et al. (2022), the possible negative consequences of inaccurate forecasts of infectious disease are discussed, but there is no attempt to quantify the utility or loss incurred as a result of those forecasts. Bilinski et al. (2023) explore ways in which predictive classifiers of local COVID-19 risk levels in the US could be tuned to policymaker preferences for different costs associated with over- and under-reaction to disease dynamics, but they do not clearly identify the source of these costs or how they depend on quantifiable policy choices. A similar discussion related to dengue countermeasures in Vietnam appears in Colón-González et al. (2021). A novel version of the WIS informally motivated by utility considerations is developed in Marshall et al. (2023), but the score is not derived in a decision-theoretic manner. There is also a thread of literature that frames infectious disease modeling as a component of a larger system for understanding how policy goals, means, and choices interact and constrain one another. As an example, Probert et al. (2016) explore how policy recommendations ought to flow from a possibly incongruous set of simulation-based projection models of a hypothetical foot-and-mouth disease outbreak when there are ranges of plausible responses and stakeholder interests. Decision theory plays a prominent role here, but not explicitly as a way to evaluate the choices made

in developing the models.

In this work, we begin to fill this gap between the ways that infectious disease forecasts have traditionally been evaluated and the ways that they have been used to support public health policy. To do so, we consider a setting in which forecasts are used to help determine the allocation of a limited quantity of medical supplies across multiple regions. In section 2 of the paper, we define a new forecast scoring rule — the *allocation score* — that evaluates forecasts based on how beneficial resource allocations derived from them would turn out to be. In section 3, we present an illustrative analysis using the allocation score to evaluate forecasts of hospital admissions in the US that were made leading up to and during the Omicron wave that peaked in January of 2022. This analysis is "synthetic" insofar as it is not intended to correspond to any specific historical record of allocation decisions that could have been supported by hospitalization forecasts during this period. However, we view the general allocation problem on which our framework is based as a versatile template for formalizing real-world decisions that must constantly be made in real-time by public health administrators around the globe, especially those related to hospital capacity, ventilator usage, doses needed for specific medications and other situations where an outbreak creates sudden and highly variable demand for potentially scarce resources.

2 The Allocation Score

We begin with an informal description of the allocation score and some examples illustrating its key characteristics in section 2.1. In section 2.2 we develop the allocation score more carefully, using a decision theoretic procedure for deriving proper scoring rules. (See section 2 of the supplement for a definition of a proper scoring rule and a more technical discussion of the procedure). In section 2.3, we note that another group of common scores including the quantile score, WIS, and CRPS, can also be derived from decision theoretic foundations —starting from a different decision making context.

2.1 Overview of Allocation Scoring

Suppose that a decision maker is tasked with determining how to allocate K available units of a resource across N locations. If the decision maker is provided with a multivariate forecast F where each marginal forecast distribution F_i predicts resource need in a particular location, one option is to choose the resource allocation that minimizes the expected total unmet need according to the forecast. We will give a more precise mathematical statement in section 2.2, but informally, the total expected unmet need according to the forecast is

$$\sum_{i=1}^{N} \mathbb{E}_{F_i}[\text{unmet need in location } i], \tag{1}$$

where the unmet need in a particular location is the difference between resource need in that location and the number of resources that were allocated there. This allocation problem has an intuitively appealing solution: allocate so that the probabilities of need exceeding allocation in various locations are as close to each other as possible. This will lead to the allocations provided by F being quantiles of the marginal distributions F_i for some *single* probability level τ that is shared in common for all locations.

After time passes and the actual level of resource need has been observed, the quality of a selected

allocation can be measured by comparing the actual need in each location to the amount of resources that were sent there. Specifically, we compute the total unmet need that resulted from the selected allocation:

$$\sum_{i=1}^{N} \text{unmet need in location } i. \tag{2}$$

One allocation is better than another if it results in lower total unmet need, and one forecast is better than another if the allocation derived from it results in lower total unmet need.

The **allocation score** of the forecast F is the avoidable unmet need that results from using the allocation that minimizes the expected unmet need according to that forecast. By "avoidable unmet need", we mean that the allocation score does not include the amount of unmet need that was inevitable simply because the amount of available resources K was less than the need for resources. Rather, the allocation score measures the unmet need that could have been avoided by an oracle that knows exactly how much need will occur in each location and divides the amount K so that nothing is wasted in one location while it could be put to use in another. An allocation score of 0 is optimal, and indicates that no other allocation of resources could have met need better than the allocation suggested by F. A larger allocation score indicates that it would have been possible to improve upon the allocation suggested by F.

Example 1 Suppose we have a forecast F for need in two locations with $F_1 = \text{Exp}(1/\sigma_1)$ and $F_2 = \text{Exp}(1/\sigma_2)$, where $\sigma_1 = 1$ and $\sigma_2 = 4$. The means of these distributions are given by the scale parameters σ_i. When the marginal forecasts are exponential distributions, it can be shown that the optimal allocation divides the available resources among the locations proportionally to the scale parameters σ_i (see section 4 of the supplemental materials). If $K = 5$ units of the resource are available, the optimal allocation according to F would be 1 unit of resources in location 1 and 4 units of resources in location 2. If, on the other hand, $K = 10$ units are available, we will allocate 2 units of resources to location 1 and 8 units to location 2. Figure 1 illustrates the situation.

Next suppose that we observe resource needs of 1 and 10 in locations 1 and 2, respectively. Based on these observed needs, we can measure the quality of the allocation suggested by the forecast by calculating the amount of unmet need that resulted from that allocation over and above what was unavoidable given the resource constraint. With $K = 5$ units of the resource, the allocation based on the forecast exactly meets the observed need in location 1, but it leaves 6 units of need unmet in location 2. However, working within the resource constraint, no other allocation could have done better: for example, allocating 0 units of resources to location 1 and 5 to location 2 still results in a total unmet need of 6 across both locations. Therefore, the forecast's allocation score is 0 with $K = 5$. On the other hand, when $K = 10$, the forecast's allocation results in $10 - 8 = 2$ units of unmet need in location 2 despite leaving no need unmet in location 1. In this case, the oracle would be able to prevent all but 1 of the total 11 units of need from going unmet, for example by allocating 1 unit of resources to location 1 and the remaining 9 units of resources to location 2. The allocation score for the forecast when $K = 10$ would therefore be 1 ($= 2$ realized $- 1$ unavoidable) in units of avoidable unmet need.

These scores illustrate a general result: allocation scores for a forecast will tend to be larger when the resource constraint is close to the observed need, because this is when it matters most which locations are allocated more or less resources. If the amount of available resources is small relative

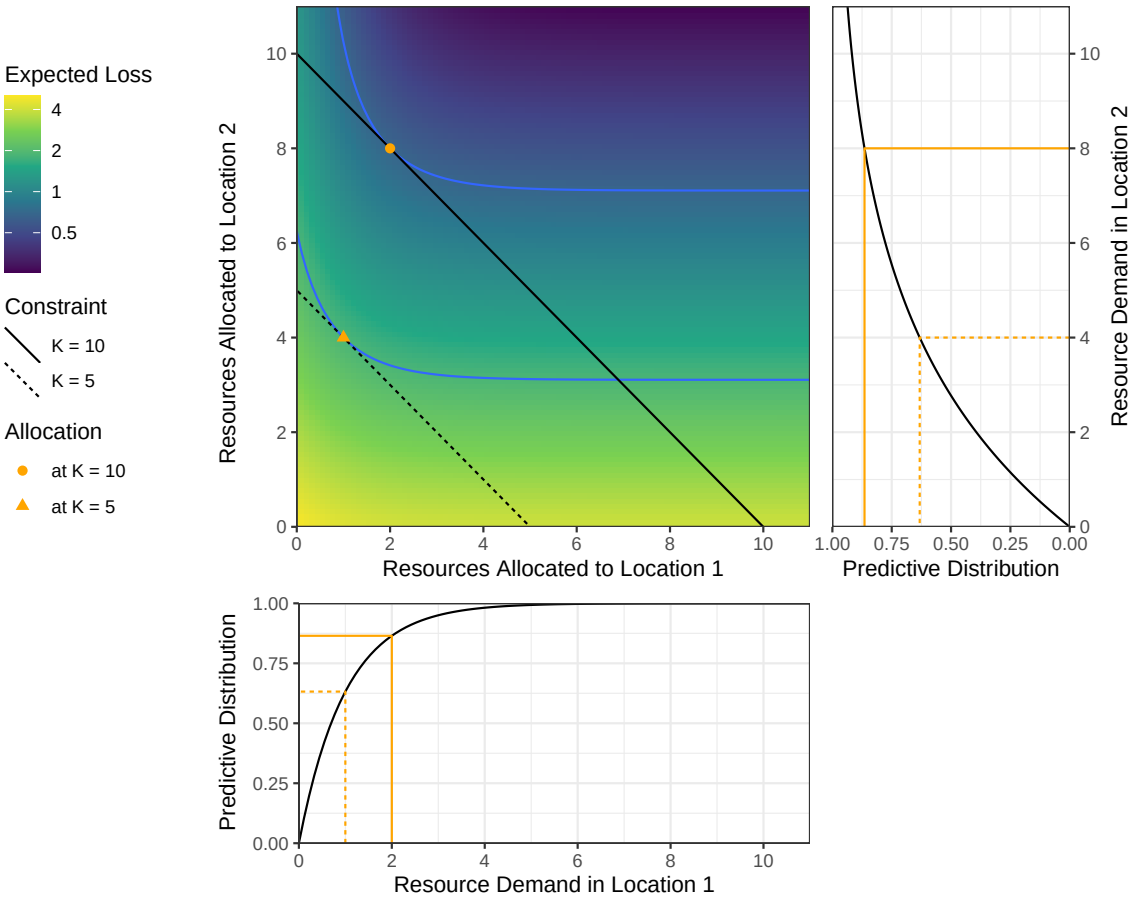

Figure 1: An illustration of the resource allocation problem in Example 1. There are $N = 2$ locations, with predictive distributions $F_1 = \text{Exp}(1)$ and $F_2 = \text{Exp}(1/4)$. The cumulative distribution functions of these distributions are illustrated in the panels at bottom and right. In the center panel, the background shading corresponds to the expected loss according to these forecasts. Diagonal black lines indicate resource constraints at $K = 5$ and $K = 10$ units; any point along those lines corresponds to an allocation that meets the resource constraint. For these forecasts, the optimal allocations are $(1, 4)$ for $K = 5$ and $(2, 8)$ for $K = 10$. These allocations are at the point on the constraint line where the expected loss is smallest, which also corresponds to the point where a level set of the expected loss surface (blue curve) is tangent to the constraint.

to the observed need, any allocation of those limited resources will result in a large amount of unmet need. If the amount of available resources is comparatively large, it becomes less important which locations receive relatively more or fewer resources because all locations will receive enough resources to meet their need. In either of these extremes of resource availability, the avoidable unmet need that arises from the allocation suggested by a forecast (i.e., the forecast's allocation score) will tend to be small.

Example 2 Now consider a different forecast that also has exponential distributions for resource need in each location, but that has the scale parameters $\sigma_1 = 2$ and $\sigma_2 = 8$, twice as large as the scale parameters of the forecast in Example 1. Because the optimal allocation is proportional to the scale parameters, this forecast would lead to the same allocations as the forecast in Example 1, and would therefore be assigned the same allocation score.

Examination of these results leads to two observations. First, the reason that these forecasts had a positive (i.e., non-optimal) allocation score at $K = 10$ is that they did not get the relative magnitude of resource need across the two locations right: the realized need was 10 times as large in location 2 as in location 1, but the forecasts only indicated that the resource allocation for location 2 should be 4 times the allocation for location 1. At its core, the allocation score measures whether the forecast accurately captures the relative magnitudes of resource need across different locations, which is precisely the information that is needed to allocate resources to those locations subject to a fixed resource constraint.

A second observation is that the forecasts in examples 1 and 2 predicted different mean levels of resource needs, but had the same allocation score. The allocation score does not directly measure whether the forecasts correctly capture the absolute magnitude of resource need in each individual location. This stands in contrast to other common scoring methods that aggregate scores such as log score, CRPS, or WIS for each location, where a poor forecast for one location is penalized regardless of alignments in other units.

2.2 A decision theoretic development of the allocation score

We give a high-level review of a general procedure for developing proper scoring rules that are tailored to specific decision making tasks in section 2.2.1, and then in section 2.2.2 we apply that procedure to develop the allocation score based on the task of deciding on how to allocate a fixed supply of resources across multiple locations. In 2.2.3 we consider a small extension where the resource constraint is not known, or it is desired to consider the value of forecasts across a range of decision making scenarios. This gives rise to the *integrated allocation score*.

2.2.1 The decision theoretic setup for forecast evaluation

In the framework of decision theory, a decision corresponds to the selection of an action x from some set of possible actions $\mathcal{X}$. For example, x may correspond to the level of investment in a measure designed to mitigate severe disease outcomes such as hospital beds, ventilators, medication, or medical staff, with $\mathcal{X}$ being the set of all possible levels of investment that we might select. The quality of a decision to take a particular action x is measured in relation to an outcome y that is unknown at the time the decision is made. For example, y may correspond to the number of individuals who eventually become sick and would benefit from the mitigation measure, and informally, an action x is successful to the extent that it meets the realized need. In the face of uncertainty, a decision maker may use a forecast F of the random variable Y to help inform the selection of the action to take. We measure

the value of a forecast as an input to this decision making process by the quality of the decisions that it leads to.

We can formalize the preceding discussion with the following three-step procedure for developing scoring rules for probabilistic forecasts:

1. Specify a *loss function* $s(x, y)$ that measures the loss associated with taking action x when outcome y eventually occurs.

2. Given a probabilistic forecast F, determine the *Bayes act* x^F that minimizes the expected loss under the distribution F.

3. The *scoring rule* for F calculates the score as the loss incurred when the Bayes act was used: $S(F, y) = s(x^F, y)$.

This is a general procedure that may be applied in settings where it is possible to specify a quantitative loss function. We call a scoring rule obtained from this procedure a *Bayes scoring rule* and in section 2 of the supplement we demonstrate that Bayes scoring rules are proper by construction (a result also shown by Dawid (2007); Gneiting and Raftery (2007)).

2.2.2 The allocation score for a fixed resource constraint

In the decision making setting that we consider, an action $x = (x_1, \ldots, x_N)$ is a vector specifying the amount that is allocated to each of N locations. We require that $0 \leq x$, i.e., that each x_i is non-negative, and that the total allocation across all locations equals the amount of available resources, K: $\sum_{i=1}^{N} x_i = K$. The set $\mathcal{X}$ consists of all possible allocations that satisfy these constraints. The eventually realized resource need in each location is denoted by $y = (y_1, \ldots, y_N)$. These levels of need are not known at the time of decision making, so we define the random vector $Y = (Y_1, \ldots, Y_N)$ where Y_i represents the as-yet-unknown level of resource need in location i. Forecasts of need in each location are collected in $F = (F_1, \ldots, F_N)$. We assume that resource need is non-negative and the forecasts reflect that, i.e. the support of each F_i is a subset of $\mathbb{R}^+$. Finally, we assume that each unit of unmet need incurs a loss denoted by L, so that if the selected resource level x_i in location i is less than the realized need y_i, a loss of $L \cdot (y_i - x_i)$ results. A variety of extensions to this setup are possible; for example, we might account for storage costs for resources that go unused, allow for a different loss per unit of unmet need in each location, or account for resource transportation costs. In this work, we choose to keep the loss function relatively simple to focus on the core ideas.

It is helpful to clearly distinguish between the time t_d when a *d*ecision is made about a public health resource allocation and the time t_r when *r*esource needs that might be addressed by that allocation occur. Our setup assumes that $t_d < t_r$. Additionally, the structure of our loss captures a setting where the resource in question does not impact the amount of demand that will materialize at time t_r, but rather it is a resource that satisfies that demand. In the context of infectious disease, this means that we do not consider resources that are intended to reduce the number of people who will become sick at some point in the future, such as a preventative influenza or COVID-19 vaccine. Instead, our setup addresses resources like hospital beds, oxygen supply, or ventilators which are intended to meet the medical needs of patients who are already sick. Additionally, our setup addresses decision-making that is related to resource needs only at the time t_r; we do not explicitly consider sequences of multiple decisions that are made over time or account for the impact of decisions on resource needs at any time other than t_r. We outline some opportunities to extend our work to more complex decision making

settings in the discussion.

With this problem formulation in place, we can develop a proper scoring rule following the outline in section 2.2.1.

Step 1: specify a loss function. The loss associated with a particular allocation is calculated by summing contributions from unmet need in each location:

$$s_A(x, y) = \sum_{i=1}^{N} L \cdot \max(0, y_i - x_i). \tag{3}$$

Here, $\max(0, y_i - x_i)$ is the unmet need in location i, which is given by $y_i - x_i$ if the realized need y_i in location i is greater than the amount x_i allocated to that location, or 0 if the amount x_i allocated to unit i is greater than or equal to the realized need. Also, L is a constant scalar value, the same across all locations, specifying the "cost" of one unit of unmet need.

Step 2: Given a probabilistic forecast F, identify the Bayes act. The Bayes act associated with the forecast, $x^{F,K}$, is the allocation that minimizes the expected loss, that is, the solution of the *allocation problem* associated with K:

$$\underset{0 \le x}{\text{minimize}}\ \mathbb{E}_F[s_A(x, Y)] \text{ subject to } \sum_{i=1}^{N} x_i = K, \tag{4}$$

where $\mathbb{E}_F[s_A(x, Y)] = \sum_{i=1}^{N} L \cdot \mathbb{E}_{F_i}[\max(0, Y_i - x_i)]$ sums the expected loss due to unmet need across all locations.

In the supplement we show that the components of the Bayes act are quantiles $x_i^{F,K} = F_i^{-1}(\tau^{F,K})$ at a probability level $\tau^{F,K}$ that depends on the forecast F and the resource constraint K, but is shared across all locations. This probability level is the level at which the resource constraint is satisfied: $\sum_{i=1}^{N} F_i^{-1}(\tau^{F,K}) = K$. This tells us that in order to allocate optimally (according to F), we must divide resources among the locations so that there is an equal forecasted probability in every location that the allocation is sufficient to meet resource need. This solution to the allocation problem is well-known in inventory management and is often attributed to Hadley and Whitin (1963).

Step 3: Define the scoring rule. We can now use the Bayes act to define a proper scoring rule for the probabilistic forecast F. Consider first the "raw" score defined as

$$S_A^{\text{raw}}(F, y; K) = s_A(x^{F,K}, y) = \sum_{i=1}^{N} L \cdot \max(0, y_i - x_i^{F,K}). \tag{5}$$

This measures the total unmet need across all locations that results from using the Bayes allocation associated with the forecast F when the actual level of need in each location is observed to be y_i.

To make this a more easily interpreted measure of forecast performance, we will adjust the raw score by subtracting the minimum loss achievable by an *oracle* allocator which has precise foreknowledge of the outcomes y_i. When the oracle has sufficient resources to meet the total need, i.e., when $\sum_{i=1}^{N} y_i \le K$, the oracle's loss is zero and allocation score coincides with the raw score. On the other hand, when the oracle cannot cover all need and incurs a loss of $L \cdot \left(\sum_{i=1}^{N} y_i - K\right) > 0$, we adjust the raw score

by this loss. The oracle-adjusted score can therefore be written as

$$S_A(F, y; K) = S_A^{\text{raw}}(F, y; K) - L \cdot \max\left(0, \sum_{i=1}^{N} y_i - K\right) \tag{6}$$

$$= L\left\{\sum_{i=1}^{N} \max(0, y_i - x_i^{F,K}) - \max\left(0, \sum_{i=1}^{N} y_i - K\right)\right\}. \tag{7}$$

The oracle adjustment aligns with a common theme in economic decision theory that *opportunity loss* (often known as *regret* or (negative) *relative utility*) is often a more important quantity than absolute loss (see e.g., Diecidue and Somasundaram (2017)).

2.2.3 Integrating the allocation score across resource constraint levels

The allocation score S_A that we developed in the previous section measures the skill of the forecast distributions F based on a single probability level $\tau^{F,K}$. This is appropriate if the resource constraint K is a known constant. However, if K is not precisely known at the time of decision making or there is interest in measuring the value of forecasts across a range of decision making scenarios with different resource constraints, we can use an *integrated allocation score* (IAS) that integrates the allocation score across values of K, weighting by a distribution p:

$$S_{IAS}(F, y) = \int S_A(F, y; K) p(K)\, dK$$

We note that the device of considering a range of hypothetical decision makers or decision making problems with different problem parameters has been employed in the past (e.g., Murphy, 1993).

2.3 Connections to Other Scores

The weighted interval score (WIS) was proposed in 2020 as a way to score forecasts that were being made in the early stages of the COVID-19 pandemic (Bracher et al., 2021); equivalent scores had also been used in previous forecast evaluation efforts (e.g., Hong et al., 2016). The WIS is a proper scoring rule for forecasts that use a set of quantiles to represent a probabilistic forecast distribution. While pointing a reader interested in more mathematical detail to Bracher et al. (2021), we note simply that the WIS is a weighted sum of interval scores at different probability levels (e.g., 50% prediction intervals, 80% PIs, 95% PIs, etc...). Larger interval scores indicate less skillful forecasts. An interval score consists of (a) the width of the interval, with larger intervals receiving higher scores (higher scores indicate less accuracy), and (b) a penalty if the interval does not cover the eventual observation, which increases the further away the interval is from the observed value. Equivalently, the WIS can also be characterized as a weighted sum of quantile scores for each individual predictive quantile. The quantile score for a particular quantile level assigns an asymmetric penalty to predictions that are too high or too low, with the relative sizes of the penalties set so that in expectation the score is minimized by the given quantile of the distribution. The most commonly used version of WIS is one that uses an equal weighting of all quantile levels, in which case WIS approximates the continuous ranked probability score (CRPS), a commonly used score for probabilistic forecasts. *It is important to note that this weighting was proposed because the resulting score approximates the CRPS, and not because it aligned with any particular public health decision-making rationale.*

That said, the quantile score and WIS can be derived using the same decision theoretic procedure that

we outlined in section 2.2. In fields such as meteorology and supply chain management, a great deal of attention has been given to the problem where a decision must be made about the quantity of a resource to purchase for a single location in the face of a fixed cost C for each unit of the resource and a loss L that will be incurred for each unit of unmet need. This leads to the quantile score for the probability level $\tau = 1 - C/L$. From this point, the WIS or CRPS can be obtained by averaging across a range of decision making settings with different cost and loss parameters, using a similar motivation that we used to obtain the IAS from the AS in section 2.2.3 (Gneiting and Ranjan, 2011).

3 Evaluating forecasts of COVID hospitalizations using the allocation score

We illustrate with an application to hospital admissions in the U.S., considering a hypothetical problem of allocating a limited supply of medical resources to states.

3.1 Data

The US COVID-19 Forecast Hub collected short-term forecasts of daily new hospital admissions for individuals with COVID-19 starting in December 2020 (Cramer et al., 2022b). The target data for these forecasts were hospital admissions as reported by the US Department of Health and Human Services through the HealthData.gov website. Forecasts were probabilistic predictions of the number of new hospital admissions on a particular day in the future, in a specific jurisdiction of the US (national level, state, or territory). Probability distributions were represented using a set of 23 quantiles for each individual prediction, including a median and the lower and upper limits of 11 central prediction intervals, from a 99% to a 10% prediction interval.

The analysis in this work focuses on forecasts made before and during the first wave of the Omicron SARS-CoV-2 variant in the US. As such, we analyzed forecasts for the 15 weeks starting with Monday November 22, 2021 through Monday February 28, 2022.

Submission to the Forecast Hub followed a weekly cycle, and each Monday the Hub collected the most recent forecasts submitted by all teams that met certain inclusion criteria and created ensemble forecasts using quantile averaging (Ray et al., 2023). Our analysis includes these ensembles (COVIDhub-ensemble and COVIDhub-trained_ensemble) as well as one other ensemble of hub models created by another team (JHUAPL-SLPHospEns) and several other individual models. Models were eligible to be included in the analysis if they were designated as a "primary" model from a team. For a model to have a complete, eligible submission in a given week, it had to have a 14 day-ahead forecast for all 50 states plus Washington DC. Models had to have a complete forecast for at least 4 of the 15 weeks in the analysis to be eligible for inclusion.

The hospitalization data used for scoring forecasts were downloaded on February 07, 2024.

3.2 Evaluation metrics

We measured forecast skill using two forecast scores, the allocation score (AS) and the weighted interval score (WIS), both defined above. We computed these scores for the 14 day ahead forecasts made each week. For this analysis, we fixed the resource constraint used for the AS to be $K = 15,000$, based

roughly on a reported number of ventilators available for reallocation in the US (Ajao et al., 2015). We computed the mean WIS (MWIS) across all of the forecasted state-level locations.

For both scores, we also computed models standardized rank among all models that submitted forecasts each week. The standardized rank is between 0 and 1, where 0 corresponds to the worst rank and 1 to the best. In the case of a tie between one or more models, all models received the better rank.

As described above, predictions were submitted to the Forecast Hub in the form of a set of 23 quantiles of the predictive distribution. The WIS can be directly calculated from these quantiles. However, calculation of the AS requires that the full cumulative distribution functions of the forecast distributions for each location are available. For the purpose of this analysis, to calculate the allocation score we approximated the full cumulative distribution functions based on the provided quantiles. On the interior of the provided quantiles we used a monotonic cubic spline to interpolate the quantiles, and in the lower and upper tails we used normal distributions with parameters selected so as to match the two lowest and two highest quantiles (see the supplement for further details). In the supplement, we show that if this evaluation procedure had been specified prospectively, the resulting score would be proper —but a post hoc application of this procedure is improper. We use the procedure here to illustrate the properties of the score, and note that a forecaster or collaborative forecasting exercise interested in using the allocation score for evaluation could circumvent issues with propriety through thoughtful elicitation of representations of forecast distributions or by collecting the allocations at specified resource levels as part of forecast submissions.

3.3 Allocation benchmarks

We used two benchmark methods as references to allow for comparison of model-derived allocations with existing standard practices. First, we evaluated forecasts from the `COVIDhub-baseline` model, which predicts a flat line from the most recent observation with uncertainty bounds based on a random walk (Cramer et al., 2022a). Second, we generated a proposed set of allocations where the quantity allocated to each state was proportional to that state's population (using US Census data of vintage 2022 (United States Census Bureau, 2022)), referred to below as `per-capita`. These two approaches generally reflect common choices for "best practices" of allocation: either allocating resources based on the most recent observed data, or in proportion to the population of each location.

3.4 Data and code availability

All forecast data used in this evaluation are available through the COVID-19 Forecast Hub (Cramer et al., 2022b). An R package implementing the allocation score is available at `https://github.com/aaronger/alloscore`. All code and data for the analyses presented in this manuscript are available at `https://github.com/aaronger/utility-eval-papers`. The analyses were generated using a reproducible workflow using R version 4.3.1 (2023-06-16) and the `targets` package (R Core Team, 2023; Landau, 2021).

3.5 Application results

3.5.1 Anatomy of forecast scores for one week

To illustrate the mechanics of allocation scoring, we start by focusing on how forecasts generated on or before December 20, 2021, with predictions for January 03, 2022, were scored by different metrics. This

week was around the peak of the Omicon wave nationally, with individual states typically observing a peak at or after January 3, 2022.

Of the 10 models evaluated for this one week, the CU-select model had the most accurate forecasts according to the allocation score while the USC-SI_kJalpha model had the most accurate forecasts based on MWIS (Table 1). The JHUAPL-Bucky model had the second best MWIS but the third worst allocation score.

Model	AS	MWIS	IAS centered at 15k	IAS uniform
CU-select	669	133	774	326
per-capita	865	-	1029	366
COVIDhub-ensemble	873	159	1067	438
USC-SI_kJalpha	995	91	1216	1097
JHUAPL-Gecko	1034	164	1141	418
MUNI-ARIMA	1084	169	1248	440
COVIDhub-trained_ensemble	1089	169	1271	823
COVIDhub-baseline	1175	170	1317	535
JHUAPL-Bucky	1358	102	1566	1214
JHUAPL-SLPHospEns	1540	129	1604	1102
UVA-Ensemble	2469	213	2635	2494

Table 1: For one illustrative week, a comparison of allocation scores (AS), mean weighted interval scores (MWIS), and two varieties of Integrated Allocation Scores (IAS). All metrics are shown for 10 models that made forecasts of hospital admissions for 2022-01-03. Results are sorted by AS. For all metrics, lower scores indicate better accuracy.

A comparison of the forecasts from the JHUAPL-Bucky and CU-select models shows that while the JHUAPL-Bucky forecast distributions were closer to the eventual observations in many states, the allocations suggested by those forecasts often were more inefficient than those from the CU-select model (Figure 2). In the example of this one week, JHUAPL-Bucky had a worse allocation score than CU-select because its forecasts led to allocations that sent excess resources to several states, such as Ohio, Pennsylvania, and Michigan, that would have been more effectively allocated to states that did receive enough resources to meet their needs, such as Florida and California (Figure 3A). These allocation errors resulted because that model's forecasts did not consistently capture the relative resource needs across different states. (Figure 2 & 3A). The CU-select model made some similar errors — most prominently, over-allocating resources to Ohio — but overall, it did a better job of forecasting the relative resource demands across different locations.

On the other hand, CU-select had worse performance as measured by WIS. Its forecasts were biased downwards, and it consistently incurred a large penalty for underprediction (Figure 3B). Predictions from the JHUAPL-Bucky model were wider, and included the observed level of daily hospital admissions more often. Therefore, it did not receive as severe penalties for forecasts that underpredicted or overpredicted the actual hospitalization count.

3.5.2 Forecast scores showed differences in aggregate and over time

Allocation scores varied substantially by date and by model (Figure 4). For predictions made for the first three Mondays in December 2021 and the last three Mondays in February 2022 all models had allocation scores under 500 (and the mean across all models was less than 100), indicating that unnecessary unmet need was fairly low on those days. The allocation scores were on the whole highest when the observed number of new hospital admissions was closest to the resource threshold of 15,000,

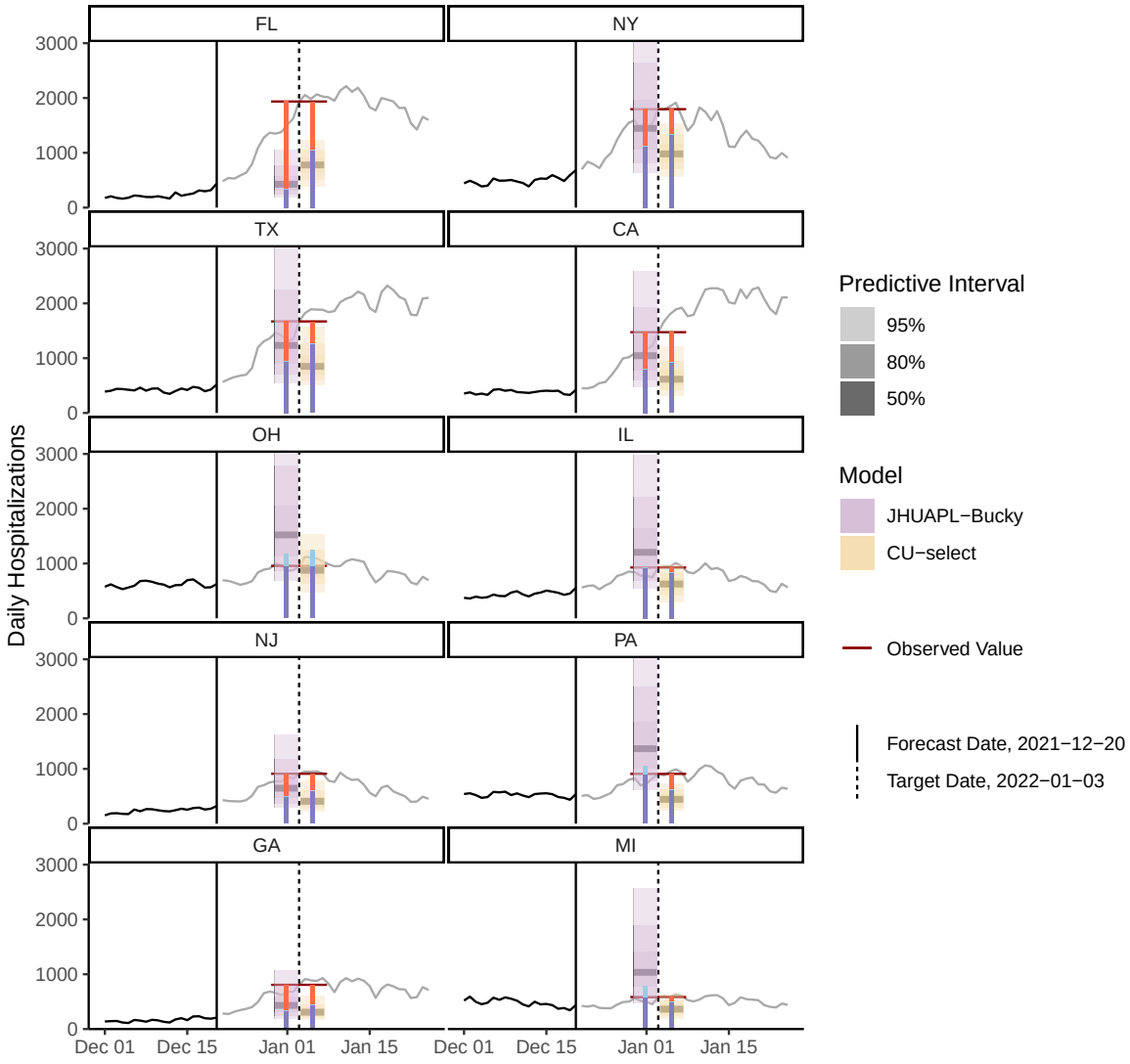

Figure 2: Probabilistic forecasts for new hospital admissions and the inferred resource allocations for COVID-19 on January 3, 2022 for the states with the ten highest hospitalization counts. For each state, the dark black line shows the data observed when the forecast was made and the grey line shows eventually observed counts. The side-by-side shaded regions show the median (solid red horizontal line) and 50%, 80% and 95% prediction intervals for the two selected models. The forecasts were made for new hospitalizations on January 3, 2022 (vertical dashed line, with number of hospitalizations indicated by red horizontal line segment at the intersection of the dashed line and the grey line of data). The vertical bars with purple, blue and red shading show the allocations. The purple bar goes from zero to the amount of need that was met by the allocation from that model to a specific location. A red bar indicates need that exceeded the resource allocation for a location, and a light blue bar shows the amount by which the resource allocation suggested by that model exceeded the need for that location.

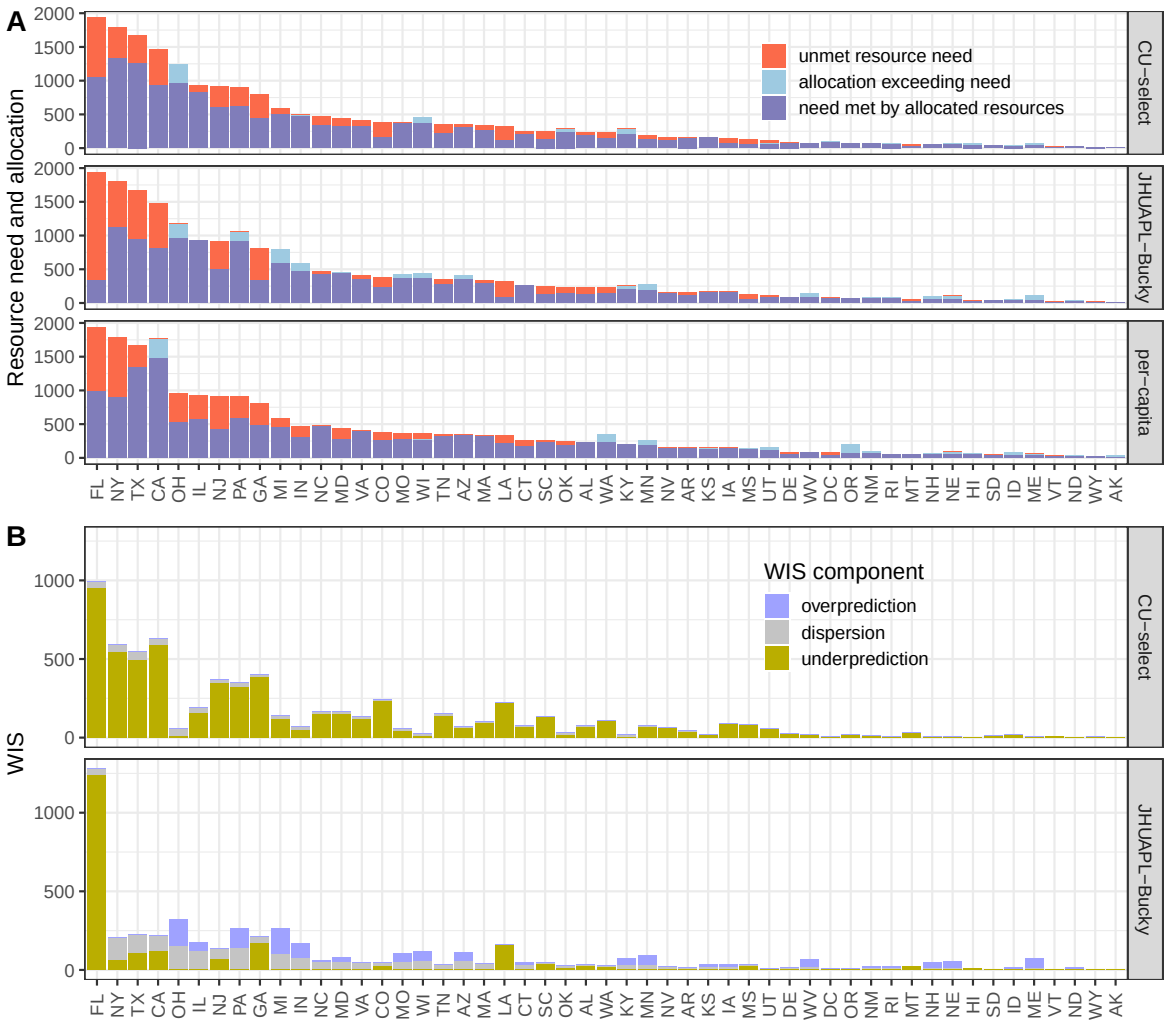

Figure 3: Component-wise breakdowns of the allocation score (Panel A) and weighted interval score (Panel B), by location for forecasts of hospitalization admissions on January 3, 2022, for two selected models (JHUAPL-Bucky and CU-select). Panel A shows the observed resource need, in this case the observed number of hospitalizations, for each state, along with the hypothetical number of resources allocated to the given location based on the forecasts from each model. The number of available resources was fixed at 15,000 and forecasts from each model were used to determine an optimal allocation strategy before the resource need was known. For most locations the resource need exceeded the resources allocated, indicated by the 'observed resource need' bar being larger than the 'allocation' bar. Panel B shows the breakdown of the weighted interval score (WIS) into components of underprediction, overprediction and dispersion. Larger values of WIS indicate more error, and the full WIS score for each location can be decomposed into the three components shown here.

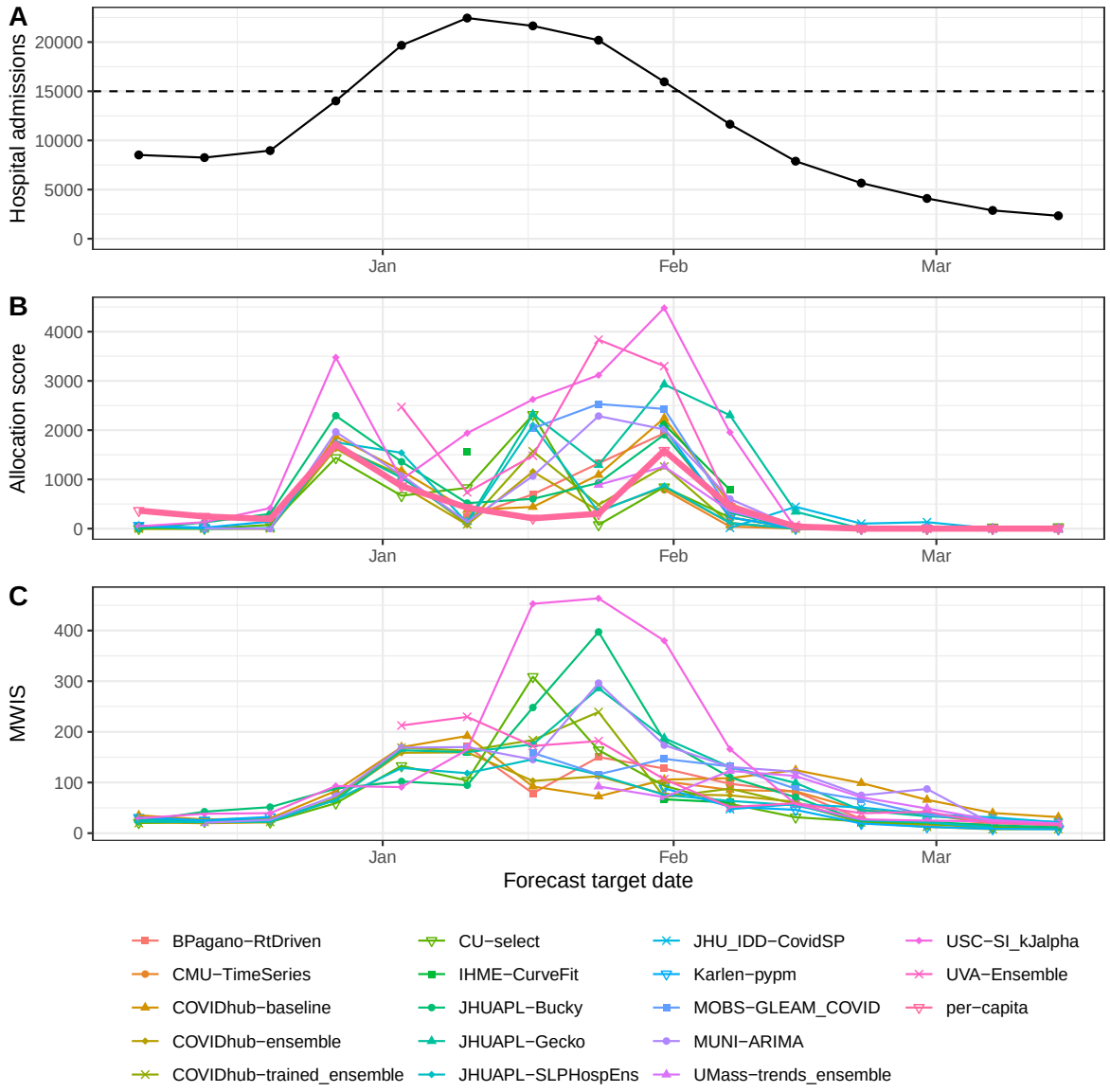

Figure 4: Hospital admissions and evaluation metrics over time. Panel A shows the number of hospital admissions in the US as a whole due to COVID-19 on a sequence of 15 Mondays from December 2021 through March 2022. These are the dates for which forecasts were made and evaluated. A horizontal dashed line at 15,000 shows the hypothetical resource constraint K. Panel B shows allocation scores (AS) for each model's 14 day-ahead forecast, across all US states. The x-axis corresponds to the date of the observation that a model's prediction was targeting (e.g., the date the forecast was made plus the forecast horizon). AS typically are high when the observed value is near to the constraint, which occurs during the last Monday in December (on the way up) and the last Monday in January (on the way down). In Panel B, the per-capital allocation is drawn in a heavier solid line. Panel C shows the MWIS metric across weeks, averaged across all states. MWIS values tend to scale with the observed and predicted values, and the peak MWIS values happen around and just after the peak of the Omicron wave.

as those are the times when any mistakes in allocation are costly in terms of wasting resources in one location that could have been used in another.

Averaging across all weeks, the ensemble forecast achieved a better mean allocation score (MAS) than two benchmark methods. The COVIDhub-ensemble had the best MAS across all weeks, the per-capita allocation had the second best MAS, and the COVIDhub-baseline model had the sixth best allocation (Table 2). No individual model had a better MAS than the per-capita allocation.

Most models were ranked similarly when evaluated using mean allocation scores (MAS) and mean weighted interval scores (MWIS), with some places of disagreement. The the four most accurate and the five least accurate models were the same according to both metrics, although not in exactly the same order (Table 2). (This comparison excludes the per-capita method which can be used to suggest allocations but does not make probabilistic forecasts and therefore cannot have a WIS.) Notably, the best model according to MWIS was the JHUAPL-SLPHospEns model, but this model only had the fourth best allocation scores.

model	MAS	MWIS
COVIDhub-ensemble	389	70
per-capita	464	-
COVIDhub-trained_ensemble	483	87
CU-select	502	81
JHUAPL-SLPHospEns	526	67
COVIDhub-baseline	594	93
JHUAPL-Bucky	643	112
MUNI-ARIMA	707	116
JHUAPL-Gecko	929	110
USC-SI_kJalpha	1473	155

Table 2: Mean weighted interval scores (MWIS) and mean allocation scores (MAS) by model across 13 weeks. For the nine models that submitted forecasts for every week from 2021-11-29 through 2022-02-21, plus a per-capita allocation, these results show the average performance of each model across all weeks. Lower scores indicate better performance, and models are sorted with the most accurate allocations at the top. WIS is not defined for the per-capita model as this approach only provides a mechanism to allocate resources, it does not provide probabilistic forecasts.

3.5.3 Metrics were not consistently correlated over time

Models showed differing levels of correlation between their allocation scores and MWIS values (Figure 5), explaining in part why aggregate MWIS and MAS values did not always agree. Here are some examples of the different model-specific patterns observed:

- Several models showed a positive association between allocation score and MWIS ranks (e.g., `Karlen-pypm` and `USC-SI_kJalpha`).
- One model had consistently strong MWIS ranks but also had highly variable allocation score ranks with no clear association between the two (e.g., `JHUAPL-SLPHospEns`)
- One model performed consistently well for both metrics (`COVIDhub-ensemble`)
- One model performed consistently well for allocation score but had only middling ranks for MWIS (`CMU-TimeSeries`)

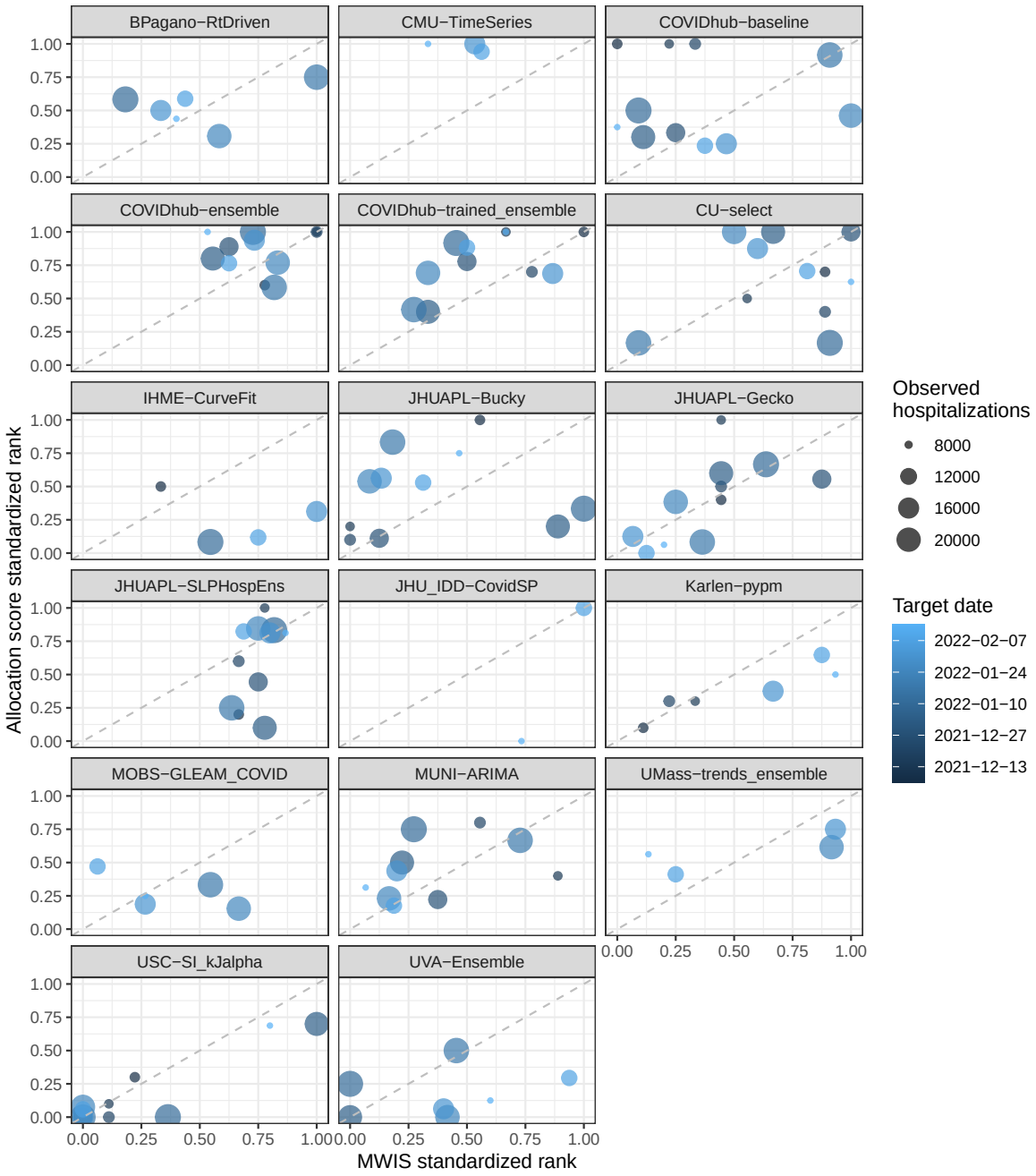

Figure 5: Association of standardized ranks for MWIS and allocation score by model and week. Each facet of the plot corresponds to one model. Within each facet, each point corresponds to a week. The x- and y-values correspond to the MWIS standardized rank and the allocation score standardized rank for that week. Points corresponding to earlier dates have darker shading. The size of the point corresponds to the observed value on the date for which the prediction was made. Models show different degrees of association between the two metrics.

3.5.4 Integrated allocation score across values of K

AS was computed for a range of K from 200 to 60,000 for forecasts made on December 20, 2020 predicting levels of hospitalizations on January 3, 2021 as well as the per-capita allocation (Figure 6A). These calculations highlight that AS was highest at at values of K near the observed nationwide total hospital admissions of 19,581 that day. Rankings of AS from all models were fairly stable across a range of values for K, with some crossings, especially at K values further away from the observed value.

The integrated allocation score (IAS) summarizes allocation scores (AS) across a range of possible values of the constraint (K), possibly taking weights into account for different values of K that might be more or less likely (Section 2.2.3). IAS was computed for two distributions on K, one uniform across the entire range and the other a symmetric distribution around 15,000 (Figure 6B). Both versions of the IAS were correlated with the original AS conducted at $K = 15,000$, with the higher correlation coming from the distribution that was centered at $K = 15,000$ (Figure 6C). Model rankings based on the AC and the centered IAS were roughly similar, with the top and bottom three approaches being the same for both scores (Table 1).

4 Discussion

In epidemiological forecasting, well-known proper scoring rules such as the log score or variations on the continuous ranked probability score (such as the weighted interval score, or WIS) have been frequently utilized to evaluate probabilistic forecasts. Often models are ranked with respect to accuracy according to a particular score, but without reference to any underlying decision-making process for which that score was designed or from which it might be derivable as a Bayes scoring rule. With careful thought and collaboration between modelers and public health officials, we argue that scores that are more aligned with public health decisions could be developed to inform specific problems. We have demonstrated that forecast evaluation methods that are tied to a specific decision making context can yield model rankings that differ substantively from those based on standard measures of forecast skill, a result that aligns with findings in other fields (Leitch and Tanner, 1991; Murphy, 1993; Cenesizoglu and Timmermann, 2012). Additionally, we show that an existing ensemble forecast approach was the only method to outperform two benchmark allocation approaches in a hypothetical application.

While our example used a hypothetical scenario to illustrate a potential application of the allocation score, resource allocation decisions are a realistic example of the kinds of decisions that could motivate more targeted forecasting exercises. We used ventilators as a notional example for a resource allocation problem because it (a) has been identified as a scarce resource during outbreaks (Huang et al., 2017), (b) could be allocated across different locations, with logistical considerations depending on the geographical scales, and (c) would likely be correlated (although not perfectly so) with hospitalization data. However, this example has limitations in that not everyone who is hospitalized needs a ventilator, and new hospital admissions may not be the best stand in for a need that corresponds to a prevalence of severe cases. Other possible real-world examples include the allocation of a limited stockpile of vaccinations(Araz et al., 2012; Persad et al., 2023) or diagnostic tests(Du et al., 2022; Pasco et al., 2023). While there may be limitations in this specific pairing of a resource and forecasted outcome, we emphasize that the important result from this work is how evaluating forecasts using a fundamentally different kind of score can illuminate previously hidden aspects of value in forecasts. In this setting, the feature of a forecast that emerges as being important is that the probabilistic forecasts predict the

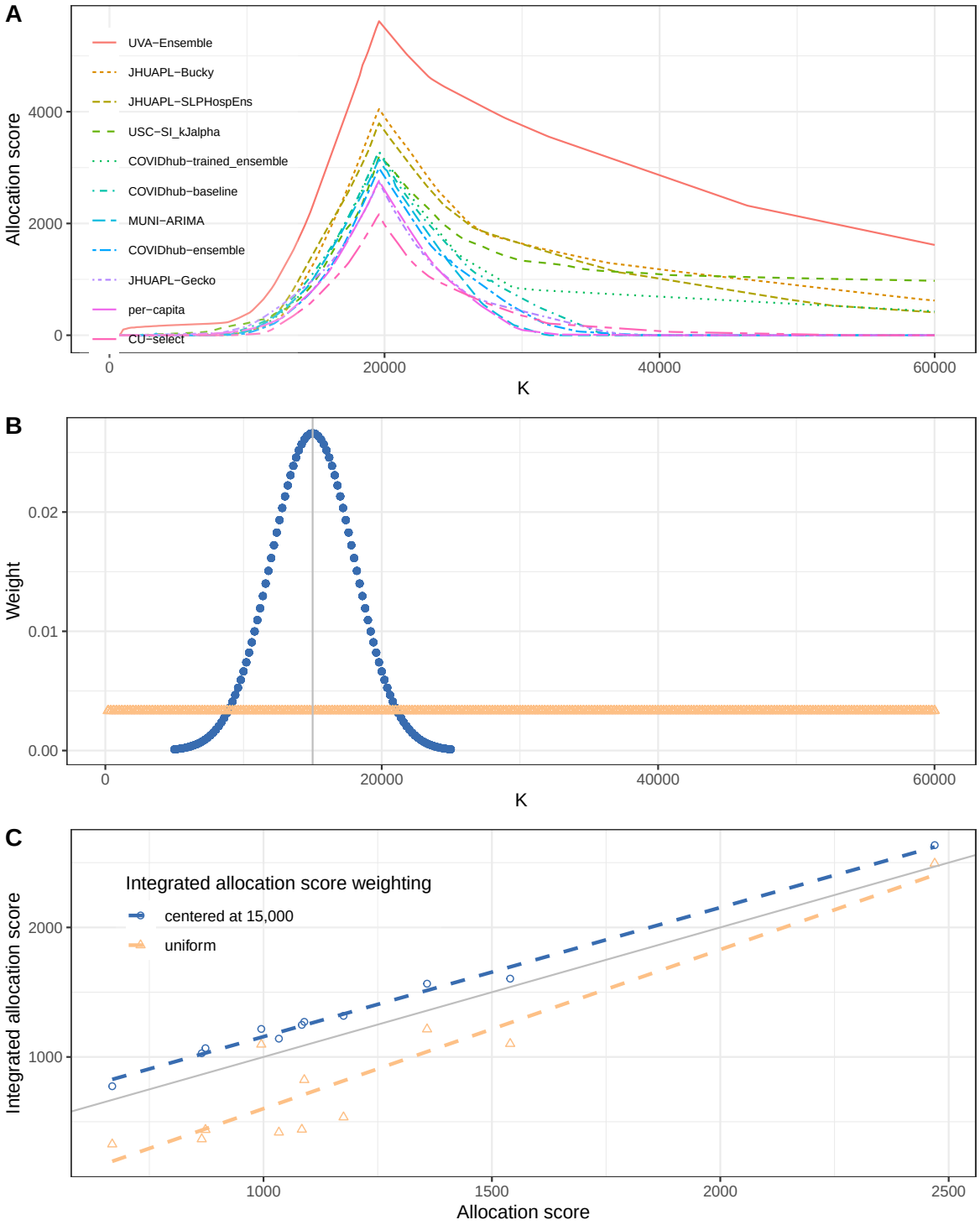

Figure 6: Allocation scores (AS) across different resource constraints (K) for 10 models that made forecasts on 2021-12-20. Panel A shows, for each model, the AS for values of K between 200 and 60,000 at increments of 200. The AS show a sharp peak just under 20,000, near the eventually observed number of hospitalizations. Panel B shows two possible weighting functions for the Integrated Allocation Score (IAS). The first (dark blue circles) computes weights proportional to a normal distribution centered at 15,000 (solid vertical gray line) with a standard deviation of 3,000, and truncated to be between 5,000 and 25,000. The second (light orange triangles) uses a uniform weight for all possible values of K. Note that the AS used in earlier sections of the application uses the single fixed value of K=15,000. Panel C shows how either method of computing the IAS (y-axis) is correlated with AS at K=15,000 (x-axis). Every point represents the AS and IAS for one model. The IAS centered at 15,000 are more closely correlated with the AS values.

relative magnitudes between locations accurately, even if the absolute magnitudes are inaccurate.

In practice, there are many potential users of forecasts with many different decision making problems. While outcomes of some decisions could potentially be specified by loss functions, other uses for forecasts may be less quantifiable, such as when forecasts are used to inform a general sense of situational awareness that may feed in unmeasurable ways into other decisions. Even those that are easily framed as expected loss minimization may differ enough that no single score would be appropriate for all users. Ideally, targeted forecasting tools could be developed through close collaboration between modelers and public health officials. However, this may only be possible in settings with sufficient staffing on both an analytics and a public health team. Increasingly, collaborative modeling hubs are being used to generate "one-size-fits-all" forecasts for many locations at once. In these settings, where tailored models are not available, it still could be possible to evaluate contributed models using a set of multiple scores to support public health end users in understanding the value of forecasts as an input to their particular decision making contexts.

There are several important limitations to the current work. The proposed framework does not attempt to capture the broader context of decision making. For example, in practice it may be possible to increase the resource constraint K by shifting funding from other disease mitigation measures. We also note that in some settings, a "successful" epidemiological forecast may lead to policy decisions that change the distribution of the predicted outcome Y. Our framework would need considerable enhancements before being applicable to forecast evaluation beyond horizons for which causal feedback can be neglected.

An opportunity for further investigation is to more carefully evaluate whether a forecast adds value to existing decision making processes. In the context of decisions about allocations, standard procedures might involve extrapolating need based on a current observed data (similar to the benchmark approaches presented above), with or without adjustments based on other political or real-world considerations. For example, in many settings public health stakeholders will make decisions after synthesizing information from a variety of quantitative and qualitative sources coupled with expert judgment. The allocation score presented in this work does not directly measure whether a given forecast adds useful information to such an existing decision-making process. While the scoring procedures as presented do not directly address this question, they could be modified (say, by comparison to a 'baseline' model or expert-elicited allocations in the absence of forecast data) to quantify the benefit of using a forecast to inform a specific decision.

In conclusion, we argue that the way modelers and policymakers view and evaluate forecasts should change depending on the specific decision-making context. Defaulting to standard forecast evaluation metrics can mask the utility (or disutility) of certain forecasts, or lead to forecasts being used in decision making contexts very different from those for which they might offer useful guidance. New collaborative work between public health officials and modeling teams is needed to assess the value and relevance of the initial findings presented here, including real-time pilot studies or simulation exercises that could be used to inform further development of new or alternative scoring metrics. We see this work as an initial overture for what we hope will grow to be a large, collaborative body of work more closely coupling applied epidemiological forecasting with public health decision making.

Acknowledgements

We wish to thank the following individuals who contributed valuable comments and feedback on early versions of this work: Matthew Biggerstaff, Rebecca Borchering, Sebastian Funk, Melissa Kerr, and Jeffrey Shaman.

This work has been supported by the National Institutes of General Medical Sciences (R35GM119582) and the U.S. CDC(1U01IP001122). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIGMS, the National Institutes of Health, or CDC.

References

Ledor S Igboh, Katherine Roguski, Perrine Marcenac, et al. Timing of seasonal influenza epidemics for 25 countries in Africa during 2010–19: a retrospective analysis. *The Lancet Global Health*, 11(5): e729–e739, 2023.

Martin I Meltzer, Charisma Y Atkins, Scott Santibanez, et al. Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014–2015. *MMWR*, 63:1–14, Sep 2014.

Gabriel Rainisch, Manjunath Shankar, Michael Wellman, et al. Regional spread of Ebola virus, West Africa, 2014. *Emerging Infectious Diseases*, 21(3):444, 2015.

Anton Camacho, Adam Kucharski, Yvonne Aki-Sawyerr, et al. Temporal changes in Ebola transmission in Sierra Leone and implications for control requirements: a real-time modelling study. *PLOS Currents*, 7, 2015.

Flavio Finger, Sebastian Funk, Kate White, et al. Real-time analysis of the diphtheria outbreak in forcibly displaced Myanmar nationals in Bangladesh. *BMC Medicine*, 17(1):58, March 2019. doi: 10.1186/s12916-019-1288-7.

Dimitris Bertsimas, Leonard Boussioux, Ryan Cory-Wright, et al. From predictions to prescriptions: A data-driven response to COVID-19. *Health Care Management Science*, 24:253–272, 2021.

Spencer J. Fox, Michael Lachmann, Mauricio Tec, et al. Real-time pandemic surveillance using hospital admissions and mobility data. *Proceedings of the National Academy of Sciences*, 119(7):e2111870119, February 2022. doi: 10.1073/pnas.2111870119.

University of Texas at Austin. COVID forecasting method using hospital and cellphone data proves it can reliably guide us cities through pandemic threats. Available at `https://news.utexas.edu/2022/02/02/covid-forecasting-method-using-hospital-and-cellphone-data-proves-it-can-reliably-guide-us-cities-through-pandemic-threats/` (2023/05/26), February 2022.

Maximilian Marshall, Felix Parker, and Lauren Marie Gardner. When are predictions useful? A new method for evaluating epidemic forecasts. *medRxiv*, 2023. doi: 10.1101/2023.06.29.23292042.

Alyssa M. Bilinski, Joshua A. Salomon, and Laura A. Hatfield. Adaptive metrics for an evolving pandemic: A dynamic approach to area-level COVID-19 risk designations. *Proceedings of the National Academy of Sciences*, 120(32):e2302528120, August 2023. doi: 10.1073/pnas.2302528120.

Vasilis Papastefanopoulos, Pantelis Linardatos, and Sotiris Kotsiantis. COVID-19: a comparison of

time series methods to forecast percentage of active cases per population. *Applied Sciences*, 10(11): 3880, 2020.

Michael A Johansson, Nicholas G Reich, Aditi Hota, et al. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for mexico. *Scientific reports*, 6(1):33707, 2016.

Craig J. McGowan, Matthew Biggerstaff, Michael Johansson, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, 9(1):683, January 2019. doi: 10.1038/s41598-018-36361-9.

Nicholas G. Reich, Logan C. Brooks, Spencer J. Fox, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, February 2019. doi: 10.1073/pnas.1812594116.

Michael A. Johansson, Karyn M. Apfeldorf, Scott Dobson, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, November 2019. doi: 10.1073/pnas.1909865116.

Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):e1008618, 2021.

Felipe J. Colón-González, Leonardo Soares Bastos, Barbara Hofmann, et al. Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. *PLOS Medicine*, 18(3): e1003542, March 2021. doi: 10.1371/journal.pmed.1003542.

Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, April 2022a. doi: 10.1073/pnas.2113561119.

Katharine Sherratt, Hugo Gruson, Rok Grah, et al. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *eLife*, 12:e81916, 2023.

Elizabeth Yardley and Fotios Petropoulos. Beyond error measures to the utility and cost of the forecasts. *Foresight: The International Journal of Applied Forecasting*, (63):36–45, 2021.

Allan H Murphy. What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8(2):281–293, 1993.

M Hashem Pesaran and Spyros Skouras. Decision-based methods for forecast evaluation. In Michael P. Clements and David F. Hendry, editors, *A Companion to Economic Forecasting*, chapter 11, pages 241–267. Blackwell, Oxford, 2002.

John PA Ioannidis, Sally Cripps, and Martin A Tanner. Forecasting for COVID-19 has failed. *International Journal of Forecasting*, 38(2):423–438, 2022.

William J.M. Probert, Katriona Shea, Christopher J. Fonnesbeck, et al. Decision-making for foot-and-mouth disease control: Objectives matter. *Epidemics*, 15:10–19, 2016. doi: 10.1016/j.epidem.2015.11.002.

A Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

G. Hadley and Thomson M. Whitin. *Analysis of Inventory Systems.* Prentice-Hall international series in management. Prentice-Hall, 1963.

Enrico Diecidue and Jeeva Somasundaram. Regret theory: A new foundation. *Journal of Economic Theory*, 172:88–119, 2017. doi: 10.1016/j.jet.2017.08.006.

Tao Hong, Pierre Pinson, Shu Fan, et al. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, 2016.

Tilmann Gneiting and Roopesh Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011. doi: 10.1198/jbes.2010.08110.

Estee Y. Cramer, Yuxin Huang, Yijin Wang, et al. The United States COVID-19 Forecast Hub dataset. *Scientific Data*, 9(1):462, August 2022b. doi: 10.1038/s41597-022-01517-w.

Evan L. Ray, Logan C. Brooks, Jacob Bien, et al. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *International Journal of Forecasting*, 39(3):1366–1383, July 2023. doi: 10.1016/j.ijforecast.2022.06.005.

Adebola Ajao, Scott V. Nystrom, Lisa M. Koonin, et al. Assessing the Capacity of the US Health Care System to Use Additional Mechanical Ventilators During a Large-Scale Public Health Emergency. *Disaster Medicine and Public Health Preparedness*, 9(6):634–641, December 2015. doi: 10.1017/dmp.2015.105.

United States Census Bureau. Annual Population Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, District of Columbia, and Puerto Rico: April 1, 2020 to July 1, 2022. US Census Bureau, 2022. URL `https://www2.census.gov/programs-surveys/popest/datasets/2020-2022/state/totals/NST-EST2022-ALLDATA.csv`. Accessed: 2024-02-01.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

William Michael Landau. The targets r package: a dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, 6(57):2959, 2021. doi: 10.21105/joss.02959.

Gordon Leitch and J Ernest Tanner. Economic forecast evaluation: profits versus the conventional error measures. *The American Economic Review*, 81(3):580–590, 1991.

Tolga Cenesizoglu and Allan Timmermann. Do return prediction models add economic value? *Journal of Banking & Finance*, 36(11):2974–2987, 2012.

Hsin-Chan Huang, Ozgur M. Araz, David P. Morton, et al. Stockpiling Ventilators for Influenza Pandemics. *Emerging Infectious Diseases*, 23(6), 2017. doi: 10.3201/eid2306.161417.

Ozgur M. Araz, Alison Galvani, and Lauren A. Meyers. Geographic prioritization of distributing

pandemic influenza vaccines. *Health Care Management Science*, 15(3):175–187, September 2012. doi: 10.1007/s10729-012-9199-6.

Govind Persad, R. J. Leland, Trygve Ottersen, et al. Fair domestic allocation of monkeypox virus countermeasures. *The Lancet Public Health*, 8(5):e378–e382, May 2023. doi: 10.1016/S2468-2667(23)00061-0.

Jiacong Du, Lauren J Beesley, Seunggeun Lee, et al. Optimal diagnostic test allocation strategy during the COVID-19 pandemic and beyond. *Statistics in Medicine*, 41(2):310–327, 2022. doi: 10.1002/sim.9238.

Remy Pasco, Kaitlyn Johnson, Spencer J. Fox, et al. COVID-19 Test Allocation Strategy to Mitigate SARS-CoV-2 Infections across School Districts. *Emerging Infectious Diseases*, 29(3), 2023. doi: 10.3201/eid2903.220761.