# Evaluating infectious disease forecasts with allocation scoring rules

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

October 17, 2023

**Abstract**

The COVID-19 pandemic has led to rapid innovation in methods for eliciting and evaluating forecasts of infectious disease burdens, with a primary goal being to help public health workers make informed decisions about how to manage these burdens. However, explicit descriptions or quantifications of the value forecasts add to society through the decisions they support are elusive. Moreover, there has only been limited discussion of how predominant forecast evaluation metrics might indicate the success of policies based in part on those forecasts.

Here we pursue one possible tether between multivariate forecasts and policy: the allocation of limited medical resources in response to COVID-19 hospitalizations in various regions so as to minimize expected unmet need. Given probabilistic forecasts of hospitalizations in each region, we formulate an allocation algorithm following techniques developed in operations research. We then score forecasts according to how much unmet need their associated allocations would have allowed. We illustrate this scheme with quantile forecasts of COVID-19 hospitalizations in the US at the state level that are recorded in the COVID-19 Forecast Hub, with the goal of determining the allocation of a hypothetical limited resource across the states. The forecast skill ranking given by this allocation scoring rule can vary substantially from the ranking given by the weighted interval score now used by the CDC, especially during surges in hospitalizations such as in late 2021 as the Omicron wave began. We see this as strong 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 research direction for epidemic forecast evaluation.

## 1 Introduction

Infectious disease forecasting models have emerged as important tools in public health. Predictions of disease dynamics have been used to inform decision making about a wide variety of measures to reduce disease spread and/or mitigate the severity of disease outcomes. For example, estimates of expected onset of flu season have been used to aid national vaccination strategies [Igboh et al., 2023], and forecasts of Ebola dynamics have been used to allocate surveillance resources [Meltzer et al., 2014, Rainisch et al., 2015]. Bertsimas et al. [2021] developed tools to inform decision making from infectious disease forecasts, which have been used to inform allocation of limited medical supplies such as ventilators, ICU capacity planning, and vaccine distribution strategy. Models developed by Fox et al. [2022] have been used to inform resource and care site planning, as well as community guidelines for masking, traveling, dining and shopping (University of Texas, 2022) . In April of 2022, the Centers for Disease Control and Prevention (CDC) announced the launch of the Center for Forecasting and

Outbreak Analytics (CFA) to translate disease forecasts into decision making (CDC, 2022), indicating that this has been identified as an important direction at the highest levels of government public health response. The value of infectious disease forecasts has typically been measured by how closely they predict disease outcomes such as cases, hospitalizations or deaths using, for example, root mean square error (RMSE) [Papastefanopoulos et al., 2020] or weighted interval score (WIS) [Bracher et al., 2021], however, recently authors have been calling for evaluating forecasts through their impact on policy [Marshall et al., 2023].

In decision making settings where it is possible to quantify the utility or loss associated with a particular action, standard tools of decision theory provide a procedure for developing forecast scoring rules that measure the value of forecasts through the quality of the decisions that they lead to. We give an overview of these procedures in Section **??**.**NGR:[This section label doesn't exist. Should it point to methods.detailed?]** There is a large history of literature applying these ideas to obtain measures of the value of forecasts that are tied to a decision making context, primarily in fields such as economics and finance, supply chain management, and meteorology. We review this work only briefly here, and we refer the reader to Yardley and Petropoulos [2021] for a general overview, and to Pesaran and Skouras [2002] and **ELR:[TODO: identify relevant review or book style summary for meteorology]** for discussions focused on applications to economics and meteorology, respectively. In finance, the value of forecasts can often be measured by the profits generated by trading decisions informed by the forecasts, perhaps adjusted for risk levels [e.g., Leitch and Tanner, 1991, Cenesizoglu and Timmermann, 2012]. In applications to supply chain management and meteorology, the value of forecasts has typically been operationalized by considering the costs associated with decisions regarding the amount of inventory to hold or the level of protection against the impacts of extreme weather events to enact Peter Catt (2007), Fotios Petropoulos and colleagues (2019), Nada Sanders and Gregory Graman (2009), T.N. Palmer (2002), Florian Pappenberger and colleagues (2015). For example, in supply chain management these decisions may incur costs related to holding inventory, labor, or providing poor service, while in meteorology we may need to balance the costs of implementing protective measures with the costs of potentially preventable weather damages. In this framework, a forecast has value if it leads to decisions with low total costs. In all of these fields, analyses have consistently found that common measures of statistical forecast accuracy do not necessarily correspond directly to measures of the value of forecasts as an input to decision making [e.g., Leitch and Tanner, 1991, Cenesizoglu and Timmermann, 2012].

However, we are aware of only a limited body of work that explicitly attempts to measure the value of infectious disease forecasts through their impact on policy, and much of this discussion has proceeded informally. For example, Ioannidis et al. [2022] discuss the possible negative consequences of inaccurate forecasts of infectious disease, but do not attempt to quantify the utility or loss incurred as a result of those forecasts. Separately, there is a thread of literature that does quantify the link between infectious disease modeling and policy making, but this work has been done outside of a forecasting context. As an example, Probert et al. [2016] develop measures of the cost of actions designed to control a hypothetical outbreak of foot-and-mouth disease and use this framework to explore policy recommendations from a variety of simulation-based projection models.

In practice, probabilistic infectious disease forecasts have most often been made for observations that emerge from public health surveillance systems and have typically been evaluated with standard, "off-the-shelf" scoring rules. For example, seasonal influenza forecasts in the US and dengue forecasts for Peru and Puerto Rico targeted public health surveillance measures of incidence over time and space,

and used log-score and mean absolute errors to evaluate forecast skill [McGowan et al., 2019, Reich et al., 2019, Johansson et al., 2019]. Pandemic COVID-19 forecasts of observed cases, hospitalizations and/or deaths in the US and Europe, as reported by municipal, state, or federal surveillance systems, were evaluated using the weighted interval score (WIS, which is an approximation of the continuous ranked probability score, or CRPS), and prediction interval coverage [Cramer et al., 2022a, Fox et al., 2022, Sherratt et al., 2023]. Similarly, CRPS was also used to assess probabilistic forecasts of dengue incidence at the district level in Vietnam [Colón-González et al., 2021]. While some of these scores can be interpreted through the lens of decision theory, and all of the application-specific papers cited above had authors from public health agencies, none of them make explicit connections between forecast evaluation and how a forecast was used in practice.

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. We consider a setting in which forecasts are used to help determine the allocation of a limited quantity of medical supplies across multiple regions. 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.

Briefly, the allocation score of a forecast is the avoidable unmet need that results from using that forecast to set resource allocations by minimizing expected unmet need. For example, suppose that a decision maker is provided with forecasts of the level of need for medical resources in each of several states or hospital systems. If there is a limited amount of the medical resource that is available to distribute, a decision maker could choose an allocation of that resource across locations that minimizes the expected unmet need according to that forecast. As measured by the allocation score, one forecast is better than another if it would lead decision makers to an allocation that results in less unmet need. If the amount of resources that is available to distribute is less than the actual need, some amount of unmet need is unavoidable. The allocation score for a forecast does not include the unmet need that was unavoidable given the resource constraint, and so it measures only the amount of unmet need that could have been prevented by using a different allocation of available resources than that suggested by the forecast. We elaborate on these ideas in Section 2.

We present an illustrative analysis using the allocation score to evaluate forecasts of hospital admissions in the US leading up to the Omicron wave in winter 2022. This analysis is "synthetic" in that it does not correspond to an actual analysis that supported decision making in real-time. However, the framework described in this paper corresponds to real-world decisions that must be made by public health administrators around the globe, and could be adapted in the future for such real-time situations. For example, forecasts for districts in Sierra Leone of bed demand to care for patients with Ebola was the subject of a real-time modeling study in late 2014 and early 2015 [Camacho et al., 2015]. And, in 2020, a model developed by an academic research group turned predictions of COVID-19 hospitalizations into estimates of ventilator usage and shortages. This framework was used by the Hartford HealthCare system in Connecticut "to align ventilator supply with projected demand at a time where the [COVID-19] pandemic was on the rise" [Bertsimas et al., 2021]. These examples illustrate the potential for forecasts to inform decisions about how to allocate limited supplies such as temporary hospital beds, ventilators, personal protective equipment, or other supplies that are known to be effective at reducing transmission or severity of disease. However, we emphasize again that these studies did not take the step of evaluating forecasts based on the quality of the allocation decisions that they supported or could have been used to support.

The remainder of this article is organized as follows. We describe the allocation score in Section 2, and in Section 3 we illustrate the use of the score in an application to evaluate short-term forecasts of COVID-19 hospital admissions in the US. Section 4 summarizes our contributions and discusses opportunities for further extensions in future work.

# 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, building on decision theoretic procedures for deriving proper scoring rules. We then comment on some connections between the allocation score that we propose and other common scores that can be derived from decision theoretic foundations, such as the quantile score, WIS, and CRPS, in section 2.3.

## 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 $\tau$ 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:

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

We emphasize that in Equation (2) the calculation of unmet need is based on the actual resource need that was realized in each location, while in Equation (1) the calculation of unmet need was based on a hypothetical potential future level of resource need. Once the actual levels of resource need have been observed, we can obtain a quantitative measure of the quality of alternative allocation decisions: one allocation is better than another if 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 $\sigma_2 = 4$. 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 $\sigma_i$ (see section *** of the supplemental materials). If we have $K = 5$ units of our resource 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, we have $K = 10$ units 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 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 $F$'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 resource constraint is very small, any allocation of those limited resources will result in a large amount of unmet need. If the resource constraint is very large, it becomes less important which locations receive relatively more or less resources because all locations will receive enough resources to meet their need.

**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 optimal allocations as the forecast in Example 1, and would therefore be assigned the same allocation score.

Note the way in which these forecasts incurred a positive (i.e., non-optimal) allocation score of 1 when $K = 10$. It was not directly due to individual misalignments of the marginal forecasts $F_i$ with the observed needs, but rather because the allocations and observed needs were not proportional as

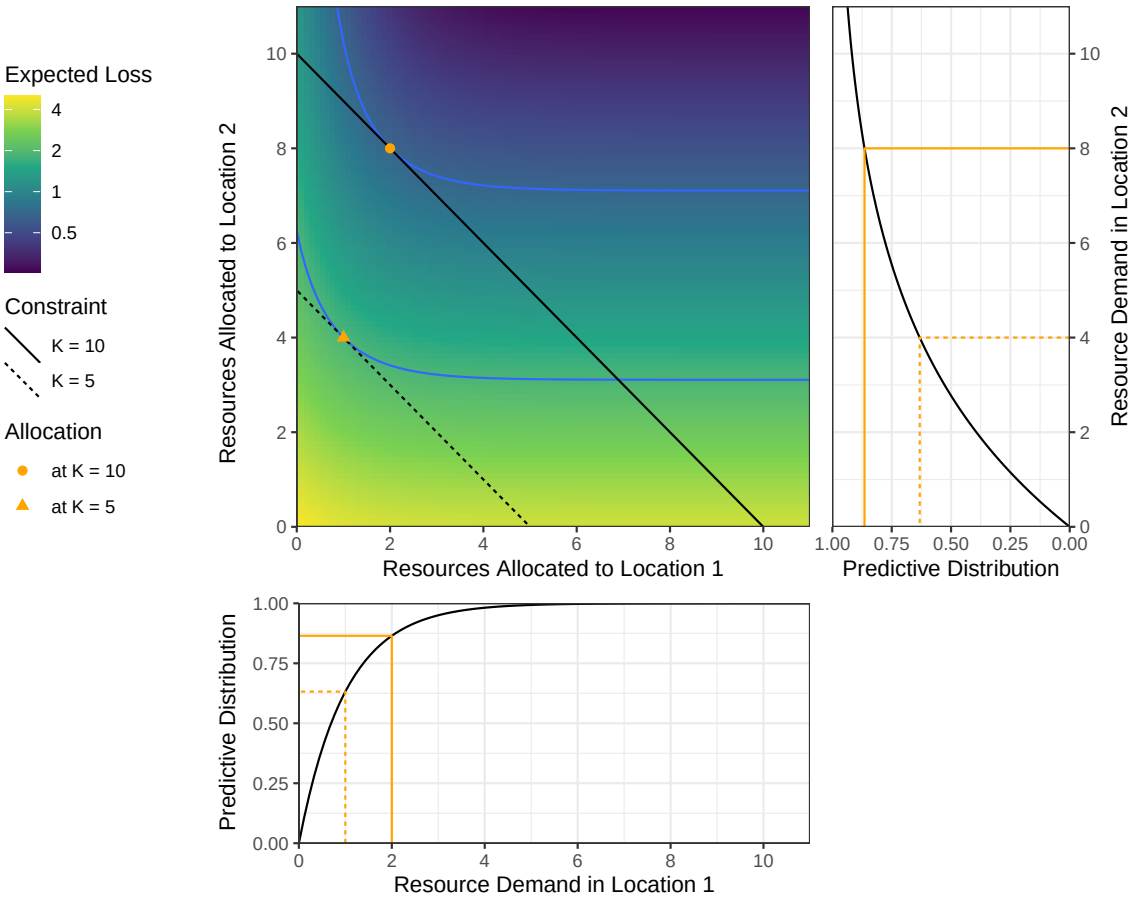


Figure 1: An illustration of the resource allocation problem in Example 1. There are $N = 2$ locations, with predictive distributions $F_1 = \text{Exp}(1)$ and $F_2 = \text{Exp}(1/4)$. The cdfs of these distributions are illustrated in the panels at bottom and right. In the center panel, the background 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.

vectors. Restating: as far as allocation decisions are concerned, with a fixed constraint $K = 10$ the fundamental problem with the forecast $F$ in Example 1 is not that it predicts a mean total resource need of 5 units; it is that the realized need was 10 times as large in location 2 as in location 1, but the forecast only indicated that the resource allocation for location 2 should be 4 times the allocation for location 1.

This illustrates a fundamental property of the allocation score: at its core, it 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. On the other hand, the allocation score is not directly sensitive to whether the forecasts in each location correctly capture the magnitude of resource need in each individual location. This stands in marked contrast to other common scoring methods for multivariate forecasts that aggregate univariate scores such as log score, CRPS, or WIS for the marginal forecasts where a poor forecast made for one unit (a location, say) is penalized regardless of alignments in other units. Note that we do not claim that the allocation score is generically preferable to these other scores —rather, it provides a view of forecast performance that is specifically tuned to the context of decision making about resource allocations.

## 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. Subject to certain technical conditions, scoring rules obtained from this procedure are proper. We refer the reader to discussion in (cite paper 2 on arxiv) for a more technically precise discussion.

### 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.

We note that the problem formulation here assumes that the resource in question does not impact the amount of demand $y_i$ that will materialize, 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. Our set up addresses resources like hospital beds, oxygen supply, ventilators, or rabies vaccines which are intended to meet the medical needs of patients who are already sick.

With this notation 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 $K$

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

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.

**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

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

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 $y_i$. But to make this a more easily interpreted measure of forecast performance, we will adjust it by the the minimum loss $l_{\text{oracle}}(y;K)$ achievable by an *oracle* allocator which has precise foreknowledge of the outcomes $y_i$ which is $l_{\text{oracle}}(y;K) = L \cdot \max(0, \sum_{i=1}^{N} y_i - K)$. Our scoring rule is then given by

$$\begin{aligned} S_A(F,y;K) &= S_A^{\text{raw}}(F,y;K) - l_{\text{oracle}}(y;K) && (6)\\ &= \sum_{i=1}^{N} L \cdot \max(0, y_i - x_i^{F,K}) - L \cdot \max(0, \sum_{i=1}^{N} y_i - K) && (7)\\ &= \begin{cases} \sum_{i=1}^{n} L \cdot \max(0, y_i - x_i^{F,K}), & \sum_{i=1}^{N} y_i \leq K \\ L \cdot K - L \cdot \sum_{i=1}^{N} \min(x_i^{F,K}, y_i), & \sum_{i=1}^{N} y_i > K. \end{cases} && (8) \end{aligned}$$

In the first case, the oracle incurs no loss so that the raw and adjusted scores coincide. The second case can be read as taking the $K$ resource units perfectly allocated by the oracle as a base penalty on the imperfect forecast $F$ and then, for each location $i$, reducing this penalty by however much of the need $y_i$ is met with the Bayes act component $x_i^{F,K}$. The oracle adjustment aligns with a central theme in modern decision theory across disciplines including economics, computer science, and psychology which is that *opportunity loss* (often known as *regret* or (negative) *relative utility*) is often a more important quantity than absolute loss from both a descriptive and normative perspective.

### 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 $\tau$. 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 $p$:

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

## 2.3 Generalizations and Connections to Other Scores

We begin this section by briefly sketching how the weighted interval score (WIS), a commonly used proper scoring rule for probabilistic forecasts during the COVID-19 pandemic, can also be derived using the approach above. Then, we discuss similarities and differences between WIS and the allocation score, and other scores in general.

### 2.3.1 The quantile loss and weighted interval score (WIS)

The weighted interval score (WIS), is a scoring rule first defined in 2020 to assist with scoring public forecasts that were being made in the early stages of the COVID-19 pandemic.[Bracher et al., 2021] Many COVID forecasts were stored in a quantile format, where probabilistic distributions are encoded by a fixed set of quantiles. This was largely a choice of convenience (e.g., all forecasts had a known, fixed data size) and was codified by the US COVID-19 Forecast Hub and other collaborative forecasting projects.[Cramer et al., 2022b] While pointing a reader interested in more mathematical detail to Bracher et al. [2021], we note simply that the WIS, is as its name suggests, 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. Interval scores themselves consist of (a) the width of the interval, with larger intervals receiving higher scores, 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. We note that the most commonly used version of WIS is one that chooses a specific set of weights so that WIS approximates the continuous rank probability score (CRPS) or pinball loss, a commonly used score for probabilistic forecasts. Other specifications of weights for the various interval scores can be used. *It is important to note that the version of WIS that has been commonly used for evaluating COVID forecasts was proposed because it approximates a commonly used forecast scoring rule, the CRPS, and not because it aligned with any particular public health decision-making rationale.*

The decision-making setup that leads to a choice of WIS as a scoring rule can be formulated using the same recipe as shown in section 2.2.1. 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.**NGR:[I might suggest adding a detailed Step 1/2/3 as an appendix and referencing it here. But I think that adding that full detail here might be too much.]** This decision-making problem leads to the pinball loss that is often used in quantile regression, and thence to the CRPS (or WIS, via approximation) as a scoring rule.**NGR:[can we make more explicit the link here to the section above? I'm not confident in my ability to describe with technical accuracy the link between quantile/pinball loss and CRPS, but I think this is a good opportunity to do so briefly.]**

### 2.3.2 Connections between scores

We have described in previous sections how both the allocation score and the weighted interval score (WIS) arise from a single standard procedure for developing proper scoring rules. The allocation score arises when the decision relates to how a fixed quantity of resources should be allocated to multiple locations. The WIS and CRPS arise when the decision relates to how much of a resource to order in the context of the cost of the resource and wanting to minimize unmet need in one or more locations. In both settings, the optimal solution sets the resource level in each location to a quantile of the respective forecast distributions.

These decision making problems differ according to the challenge faced by the decision maker: a fixed constraint on the available resources for the allocation problem, or a cost per unit of resources in the resource purchasing problem. In fact, it is possible to combine these into a more general problem where the decision maker must decide on both the total level of resources to purchase and an allocation of those resources across multiple locations, subject to a cost per unit of resources and a constraint on the total quantity of resources that can be purchased. The quantile loss and the allocation score presented in this work are both special cases of the score that emerges from that more general decision making problem that integrates across resource cost and constraint parameters together. We pursue this direction further in other work (cite arxiv version of paper 2).

# 3 Evaluating forecasts of COVID hospitalizations using the allocation score

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

Case study heading into the Omicron wave. Some more detailed discussion of implications of bad forecasts for specific decision making purposes – take a "deep dive" into one or two example states like FL.

Look at results over a broader range of time.

**NGR:[TODO: add a figure showing some forecasts and ground truth data.]**

## 3.1 Data

### 3.1.1 Hospitalization data

Starting in the summer of 2020, the US Health and Human Services began reporting counts of daily new admissions to hospitals for individuals with COVID-19.(cite HHS Protect) These data served as the source of "ground truth" data for the US COVID-19 Forecast Hub which, starting in December 2020, collected short-term forecasts of new hospital admissions at the daily scale. These daily counts were available for the US as a whole, and all states and several additional jurisdictions such as Puerto Rico and Washington DC. The data were updated daily and were available for download by the public through the HHS HealthData.gov website. For this analysis, we downloaded the hospitalizations data through the covidHubUtils R package, which connects users to the most recent version of the data.(cite package) **NGR:[can we use tar_manifest() to reproducibly display the date the data were stored?]**

### 3.1.2 Forecast data

The US COVID-19 Forecast Hub, a consortium funded by the US CDC and led by a research group at UMass-Amherst, collected short-term forecasts of hospitalizations starting in December 2020.(cite Data descriptor paper) Any team that with appropriately formatted forecasts could submit them to the Forecast Hub data repository on GitHub.(cite repo site) Forecasts were time-stamped by GitHub upon submission and passed validation checks that ensured correct formatting and that the forecasts were being submitted only for dates in the future, not for data that had already been observed.

Forecast submission followed a weekly cycle and culminated in the creation of an ensemble forecast. Forecasts could be submitted on any day during the week. However once a week on Mondays, the Forecast Hub would collect the most recent forecasts submitted by all teams that met certain inclusion criteria and create an ensemble forecast using quantile averaging.(cite Evan's paper) An ensemble that treated all models equally was created (COVIDhub-ensemble) as was a model that created weights of submitted models based on performance in the past 12 weeks (COVIDhub-trained_ensemble). One other model that combined multiple forecasts from different teams but used a different ensembling algorithm, a linear pooling method with tail extrapolation, was also included in our analyses (JHUAPL-SLPHospEns). Several other models have "ensemble" in their name, but this refers to combinations of different variations of models that the specific team created, not to a multi-model ensemble combining different submitted forecasts to the Forecast Hub.

All forecasts, including the ensemble, were submitted as probabilistic predictions about 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 specified, per the requirements established by the Forecast Hub, as a set of 23 quantiles for each individual prediction. The submitted quantiles included a median (treated as a "point" prediction) and defined 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 downloaded forecasts for the 15 weeks starting with Monday 2021-11-22 through Monday 2022-02-28.

We established a set of inclusion criteria to determine which forecasts and models to include in our analysis. Models were eligible to be included in the analysis if they were considered a "primary" model from a team. (If a team submitted multiple versions of similar models, they were required to designate one as "primary".) 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 considered eligible for inclusion.

## 3.2 Evaluation metrics

This manuscript focuses on two proper forecast scores, the allocation score and the weighted interval score (WIS).(cite Gerding and Bracher) Propriety of forecast scoring rules is desirable as it ensures that forecasters are not incentivized to modify their forecast distribution to achieve a better score.(cite Gneiting and Raftery or Raftery and Gneiting)

### 3.2.1 Allocation Score

**NGR:[TODO: general introduction to the alloscore]**

For the analysis, we fixed a resource constraint $K$ to be 15,000, based roughly on a reported number of ventilators available for reallocation in the US.(cite paper) For each week, we computed a the allocation score for each 14 day-ahead forecast based on $K = 15,000$. **NGR:[it appears that for the last two weeks selected (forecast dates: 2022-02-14 and 2022-02-28) the upper limit of K was 11400 and 9200 respectively, so we do not have alloscores computed for K=15000. Maybe they would have all been zero, since the limits were likely set based on the number of cases being low. What is the rule used to determine the max K computed? Don't think**

**we need to change it, but would be nice to state in the manuscript what it is.]**

We also computed a standardized rank for the allocation score for each model $m$ and week $w$. First, we computed the number of models that forecasted that week ($n_w$) and the rank of model $m$ among the $n_w$ models ($r^{AS}_{m,w}$). The model with the best allocation score received a rank of 1 and the worst received a rank of $n_w$. In the case of a tie between one or more models, all models received the better rank. We then rescaled these rankings to compute the allocation score standardized rank ($sr^{A}_{m,w}S$) between 0 and 1, where 0 corresponds to the worst rank and 1 to the best.

$$sr^{AS}_{m,w} = 1 - \frac{r^{AS}_{m,w} - 1}{n_w - 1} \tag{9}$$

#### 3.2.2 Weighted Interval Score (WIS)

**NGR:[TODO: general introduction to WIS and define MWIS as mean WIS.]**

We computed standardized ranks for MWIS ($sr^{MWIS}_{m,w}$) using the same procedure as for allocation scores described above.

### 3.3 Data and code availability

### 3.4 Application results

#### 3.4.1 Metrics as a function of time

Allocation scores varied substantially by date and by model (Figure XX). 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 the unnecessary unmet need was fairly low relative to the total number of hospital admissions on those days. The allocation scores are on the whole highest when the observed number of new hospital admissions is closest to the resource threshold of 15,000, as this is the time when any mistakes in allocation are costly in terms of wasting resources in one location that could have been used in another. Predictions made during the peak week and just after showed the highest variation in allocation scores, with some models having values under 1000 and others having values over 3500.

Mean weighted interval scores (MWIS) also varied by date and model, and more clearly were dependent on the scale of the observed data. MWIS values were low (all models under 100) for all Mondays in December 2021 and the final four Mondays evaluated. Across all models both the average and median MWIS value for every Monday in January was above 100, with the largest errors occurring one and two weeks after the peak was observed.

#### 3.4.2 correlation between WIS and allocation score

Models show differing levels of correlation between their allocation scores and MWIS values (Figure XX). Here are some examples of the different patterns observed:

- positive association between allocation score and MWIS ranks: `Karlen-pypm` and `USC-SI_kJalpha` models
- consistently strong MWIS ranks, highly variable allocation score ranks, no clear association: `JHUAPL-SLPHospEns` and `CU-select`

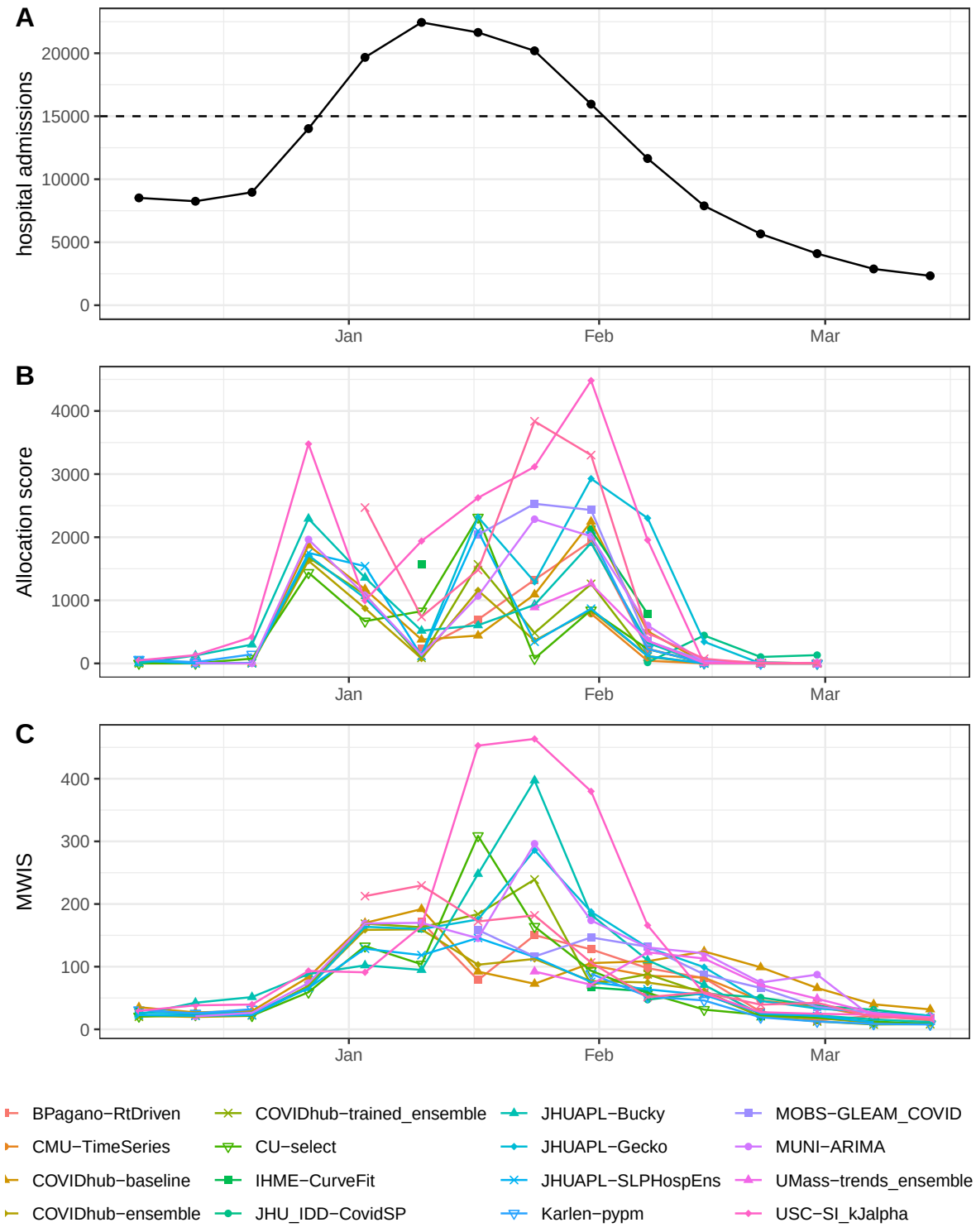


Figure 2: Hospital admissions and evaluation metrics over time. Panel A shows the number of hospital admissions in the US as a whole due to COVID-19 on a sequence of 15 Mondays from December 2021 through March 2022. These are the days for which forecasts were made and evaluated. A horizontal dashed line at 15,000 shows the assumed resource constraint $K$. Panel B shows allocation scores for each model's 14-day ahead forecast, across all US states. The x-axis corresponds to the date for which the prediction was made. Allocation scores 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). Panel C shows the MWIS metric across weeks, averaged across all states. Similarly to panel B, the x-axis corresponds to the date for which the prediction was made. MWIS values tend to scale with the observed and predicted values, and the peak MWIS values happen around and just after the pean of the Omicron wave.

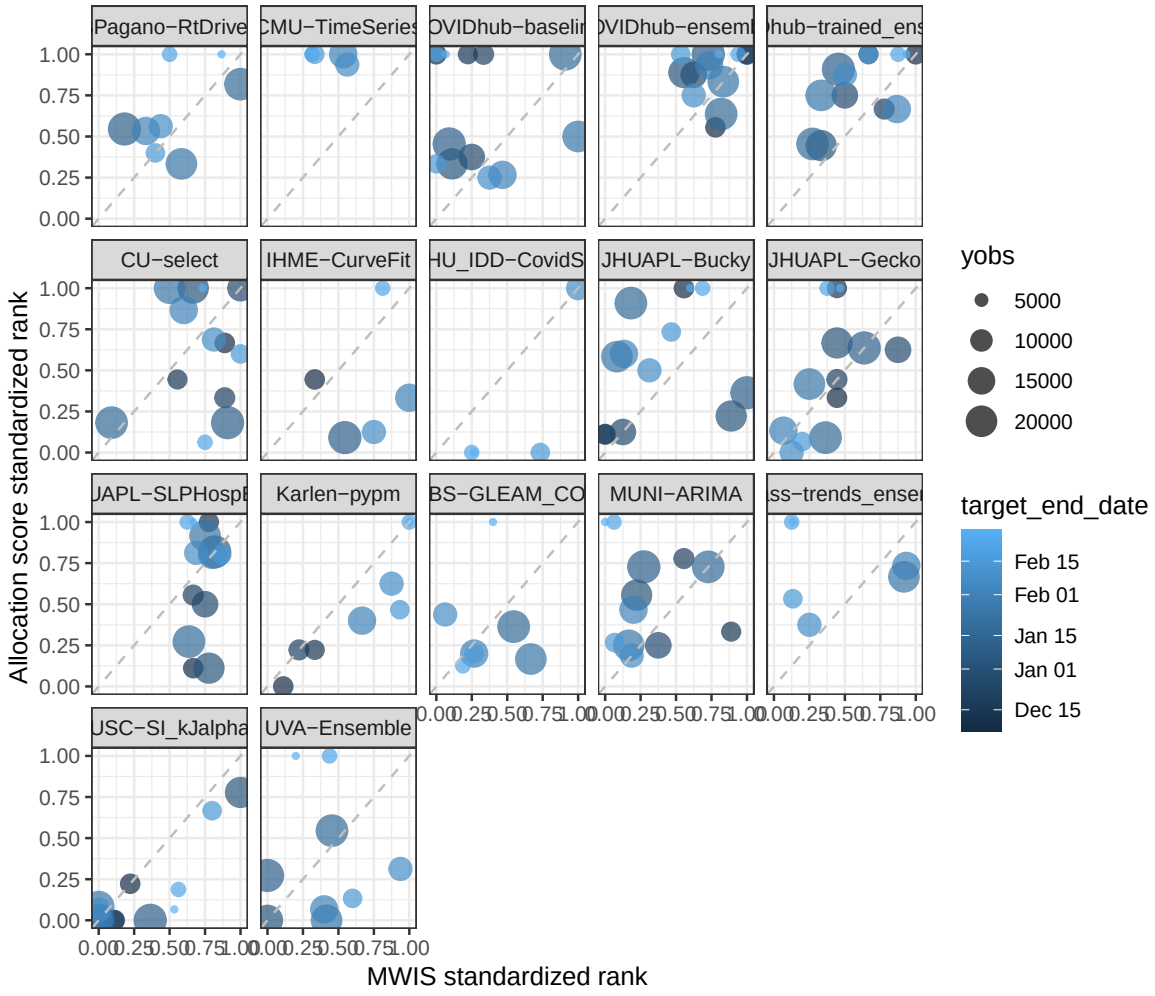


Figure 3: Association of standardized ranks for MWIS and allocation score by model and week. Each facet of the plot corresponds to one model. Within each facet, each point corresponds to a week. The x- and y-values correspond to the MWIS standardized rank and the allocation score standardized rank for 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.

- consistently strong ranks for both metrics, no clear association: `COVIDhub-ensemble`

Additionally, the `COVIDhub-baseline` model, which predicts a flat line from the most recent observation with uncertainty bounds based on a random walk, had the highest rank for allocation score in six weeks, more than any other model except the `COVIDhub-ensemble` which also had six.

- bubble figure with ranks sized by total observations

#### 3.4.3 Integrated allocation score across values of K

- something here from Ben's analysis?

## 4 Discussion

In summary, proper scoring rules for probabilistic forecasts can be traced back to a loss function motivated by a specific decision-making context. Often, forecast scores are used to rank models according to accuracy with that particular score, but without reference to the underlying decision-making process for which that score was designed. 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. In some situations, individual or ensemble models could be developed to minimize the loss for a particular setting, although this is also an area largely unexplored in the realm of public health to date.

Notes for the application

- It is clear that the metrics are capturing different aspects of performance

- WIS is scale dependent, alloscore not as much

- the fact that it is hard to beat the baseline for allocation suggests that we are not consistently adding value over just looking at the current levels. **NGR:[although is there something more to it about how the baseline accounts for uncertainty?]**

We often conceive of infectious disease forecasts as being useful for decision making purposes, but it is rare for forecast evaluation to be tied directly to the value of the forecasts for informing those decisions. This work seeks to address that gap.

We have demonstrated that evaluation methods that are tied to decision making context can yield model rankings that are substantively different from generic measures of forecast skill like WIS.

In practice, there are many users of forecasts with many different decision making problems. Not all can be easily quantified. Those that can be easily quantified may differ enough that no single score is appropriate for all users. We suggest reporting multiple scores. This may be tricky to operationalize in the setting of a general forecast hub. It matters how you elicit and represent probabilistic forecasts (quantiles? samples? cdfs?).

The allocation score we developed here does not directly account for important considerations such as fairness/equity of allocations.

The allocation score we developed also 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.

Forecaster's dilemma: a successful forecast may lead to decisions that change the distribution of the outcome $Y$. Our framework cannot be used in those settings.

There is much more to do in this general area.

# 5 References

# References


Ledor S Igboh, Katherine Roguski, Perrine Marcenac, Gideon O Emukule, Myrna D Charles, Stefano Tempia, Belinda Herring, Katelijn Vandemaele, Ann Moen, Sonja J Olsen, 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, Barbara Knust, Brett W Petersen, Elizabeth D Ervin, Stuart T Nichol, Inger K Damon, and Michael L Washington. Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014–2015. *WWMR Surveill*, 63:1–14, Sep 2014.

Gabriel Rainisch, Manjunath Shankar, Michael Wellman, Toby Merlin, and Martin I Meltzer. Regional spread of Ebola virus, West Africa, 2014. *Emerging Infectious Diseases*, 21(3):444, 2015.

Dimitris Bertsimas, Leonard Boussioux, Ryan Cory-Wright, Arthur Delarue, Vassilis Digalakis, Alexandre Jacquillat, Driss Lahlou Kitane, Galit Lukin, Michael Li, Luca Mingardi, Omid Nohadani, Agni Orfanoudaki, Theodore Papalexopoulos, Ivan Paskov, Jean Pauphilet, Omar Skali Lami, Bartolomeo Stellato, Hamza Tazi Bouardi, Kimberly Villalobos Carballo, Holly Wiberg, and Cynthia Zeng. 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, Remy Pasco, Spencer Woody, Zhanwei Du, Xutong Wang, Tanvi A. Ingle, Emily Javan, Maytal Dahan, Kelly Gaither, Mark E. Escott, Stephen I. Adler, S. Claiborne Johnston, James G. Scott, and Lauren Ancel Meyers. 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. URL `https://www.pnas.org/doi/10.1073/pnas.2111870119`. Publisher: Proceedings of the National Academy of Sciences.

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-` (2023/05/26), 2022.

Centers for Desease Control and Prevention. CDC launches new center for forecasting and outbreak analytics. Available at `https://www.cdc.gov/media/releases/2022/p0419-forecasting-center.html` (2022/05/26), 2022.

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.

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.

Maximilian Marshall, Felix Parker, and Lauren Marie Gardner. When are predictions useful? a new method for evaluating epidemic forecasts. *medRxiv*, pages 2023–06, 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.

M Hashem Pesaran and Spyros Skouras. Decision-based methods for forecast evaluation. *A companion to economic forecasting*, pages 241–267, 2002.

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.

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, Michael C. Runge, Tim E. Carpenter, Salome Dürr, M. Graeme Garner, Neil Harvey, Mark A. Stevenson, Colleen T. Webb, Marleen Werkman, Michael J. Tildesley, and Matthew J. Ferrari. Decision-making for foot-and-mouth disease control: Objectives matter. *Epidemics*, 15:10–19, 2016. ISSN 1755-4365. doi: https://doi.org/10.1016/j.epidem.2015.11.002. URL `https://www.sciencedirect.com/science/article/pii/S175543651500095X`.

Craig J. McGowan, Matthew Biggerstaff, Michael Johansson, Karyn M. Apfeldorf, Michal Ben-Nun, Logan Brooks, Matteo Convertino, Madhav Erraguntla, David C. Farrow, John Freeze, Saurav Ghosh, Sangwon Hyun, Sasikiran Kandula, Joceline Lega, Yang Liu, Nicholas Michaud, Haruka Morita, Jarad Niemi, Naren Ramakrishnan, Evan L. Ray, Nicholas G. Reich, Pete Riley, Jeffrey Shaman, Ryan Tibshirani, Alessandro Vespignani, Qian Zhang, and Carrie Reed. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, 9(1):683, January 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-36361-9. URL `https://www.nature.com/articles/s41598-018-36361-9`.

Nicholas G. Reich, Logan C. Brooks, Spencer J. Fox, Sasikiran Kandula, Craig J. McGowan, Evan Moore, Dave Osthus, Evan L. Ray, Abhinav Tushar, Teresa K. Yamana, Matthew Biggerstaff, Michael A. Johansson, Roni Rosenfeld, and Jeffrey Shaman. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 116(8):3146–3154, February 2019. ISSN 1091-6490. doi: 10.1073/pnas.1812594116.

Michael A. Johansson, Karyn M. Apfeldorf, Scott Dobson, Jason Devita, Anna L. Buczak, Benjamin Baugher, Linda J. Moniz, Thomas Bagley, Steven M. Babin, Erhan Guven, Teresa K. Yamana, Jeffrey Shaman, Terry Moschou, Nick Lothian, Aaron Lane, Grant Osborne, Gao Jiang, Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, Roni Rosenfeld, Justin Lessler, Nicholas G. Reich, Derek A. T. Cummings, Stephen A. Lauer, Sean M. Moore, Hannah E. Clapham, Rachel Lowe, Trevor C. Bailey, Markel García-Díez, Marilia Sá Carvalho, Xavier Rodó,

Tridip Sardar, Richard Paul, Evan L. Ray, Krzysztof Sakrejda, Alexandria C. Brown, Xi Meng, Osonde Osoba, Raffaele Vardavas, David Manheim, Melinda Moore, Dhananjai M. Rao, Travis C. Porco, Sarah Ackley, Fengchen Liu, Lee Worden, Matteo Convertino, Yang Liu, Abraham Reddy, Eloy Ortiz, Jorge Rivero, Humberto Brito, Alicia Juarrero, Leah R. Johnson, Robert B. Gramacy, Jeremy M. Cohen, Erin A. Mordecai, Courtney C. Murdock, Jason R. Rohr, Sadie J. Ryan, Anna M. Stewart-Ibarra, Daniel P. Weikel, Antarpreet Jutla, Rakibul Khan, Marissa Poultney, Rita R. Colwell, Brenda Rivera-García, Christopher M. Barker, Jesse E. Bell, Matthew Biggerstaff, David Swerdlow, Luis Mier-Y-Teran-Romero, Brett M. Forshey, Juli Trtanj, Jason Asher, Matt Clay, Harold S. Margolis, Andrew M. Hebbeler, Dylan George, and Jean-Paul Chretien. An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 116(48):24268–24274, November 2019. ISSN 1091-6490. doi: 10.1073/pnas.1909865116.

Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, Johannes Bracher, Andrea Brennen, Alvaro J. Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H. House, Yuxin Huang, Dasuni Jayawardena, Abdul H. Kanji, Ayush Khandelwal, Khoa Le, Anja Mühlemann, Jarad Niemi, Apurv Shah, Ariane Stark, Yijin Wang, Nutcha Wattanachit, Martha W. Zorn, Youyang Gu, Sansiddh Jain, Nayana Bannur, Ayush Deva, Mihir Kulkarni, Srujana Merugu, Alpan Raval, Siddhant Shingi, Avtansh Tiwari, Jerome White, Neil F. Abernethy, Spencer Woody, Maytal Dahan, Spencer Fox, Kelly Gaither, Michael Lachmann, Lauren Ancel Meyers, James G. Scott, Mauricio Tec, Ajitesh Srivastava, Glover E. George, Jeffrey C. Cegan, Ian D. Dettwiller, William P. England, Matthew W. Farthing, Robert H. Hunter, Brandon Lafferty, Igor Linkov, Michael L. Mayo, Matthew D. Parno, Michael A. Rowland, Benjamin D. Trump, Yanli Zhang-James, Samuel Chen, Stephen V. Faraone, Jonathan Hess, Christopher P. Morley, Asif Salekin, Dongliang Wang, Sabrina M. Corsetti, Thomas M. Baer, Marisa C. Eisenberg, Karl Falb, Yitao Huang, Emily T. Martin, Ella McCauley, Robert L. Myers, Tom Schwarz, Daniel Sheldon, Graham Casey Gibson, Rose Yu, Liyao Gao, Yian Ma, Dongxia Wu, Xifeng Yan, Xiaoyong Jin, Yu-Xiang Wang, YangQuan Chen, Lihong Guo, Yanting Zhao, Quanquan Gu, Jinghui Chen, Lingxiao Wang, Pan Xu, Weitong Zhang, Difan Zou, Hannah Biegel, Joceline Lega, Steve McConnell, V. P. Nagraj, Stephanie L. Guertin, Christopher Hulme-Lowe, Stephen D. Turner, Yunfeng Shi, Xuegang Ban, Robert Walraven, Qi-Jun Hong, Stanley Kong, Axel van de Walle, James A. Turtle, Michal Ben-Nun, Steven Riley, Pete Riley, Ugur Koyluoglu, David DesRoches, Pedro Forli, Bruce Hamory, Christina Kyriakides, Helen Leis, John Milliken, Michael Moloney, James Morgan, Ninad Nirgudkar, Gokce Ozcan, Noah Piwonka, Matt Ravi, Chris Schrader, Elizabeth Shakhnovich, Daniel Siegel, Ryan Spatz, Chris Stiefeling, Barrie Wilkinson, Alexander Wong, Sean Cavany, Guido España, Sean Moore, Rachel Oidtman, Alex Perkins, David Kraus, Andrea Kraus, Zhifeng Gao, Jiang Bian, Wei Cao, Juan Lavista Ferres, Chaozhuo Li, Tie-Yan Liu, Xing Xie, Shun Zhang, Shun Zheng, Alessandro Vespignani, Matteo Chinazzi, Jessica T. Davis, Kunpeng Mu, Ana Pastore y Piontti, Xinyue Xiong, Andrew Zheng, Jackie Baek, Vivek Farias, Andreea Georgescu, Retsef Levi, Deeksha Sinha, Joshua Wilde, Georgia Perakis, Mohammed Amine Bennouna, David Nze-Ndong, Divya Singhvi, Ioannis Spantidakis, Leann Thayaparan, Asterios Tsiourvas, Arnab Sarker, Ali Jadbabaie, Devavrat Shah, Nicolas Della Penna, Leo A. Celi, Saketh Sundar, Russ Wolfinger, Dave Osthus, Lauren Castro, Geoffrey Fairchild, Isaac Michaud, Dean Karlen, Matt Kinsey, Luke C. Mullany, Kaitlin Rainwater-Lovett, Lauren Shin, Katharine Tallaksen, Shelby Wilson, Elizabeth C. Lee, Juan Dent, Kyra H. Grantz, Alison L. Hill, Joshua Kaminsky, Kathryn Kaminsky, Lindsay T. Keegan, Stephen A. Lauer, Joseph C. Lemaitre, Justin Lessler, Hannah R. Meredith, Javier Perez-Saez, Sam Shah, Claire P. Smith, Shaun A. Truelove, Josh Wills,

Maximilian Marshall, Lauren Gardner, Kristen Nixon, John C. Burant, Lily Wang, Lei Gao, Zhiling Gu, Myungjin Kim, Xinyi Li, Guannan Wang, Yueying Wang, Shan Yu, Robert C. Reiner, Ryan Barber, Emmanuela Gakidou, Simon I. Hay, Steve Lim, Chris Murray, David Pigott, Heidi L. Gurung, Prasith Baccam, Steven A. Stage, Bradley T. Suchoski, B. Aditya Prakash, Bijaya Adhikari, Jiaming Cui, Alexander Rodríguez, Anika Tabassum, Jiajia Xie, Pinar Keskinocak, John Asplund, Arden Baxter, Buse Eylul Oruc, Nicoleta Serban, Sercan O. Arik, Mike Dusenberry, Arkady Epshteyn, Elli Kanal, Long T. Le, Chun-Liang Li, Tomas Pfister, Dario Sava, Rajarishi Sinha, Thomas Tsai, Nate Yoder, Jinsung Yoon, Leyou Zhang, Sam Abbott, Nikos I. Bosse, Sebastian Funk, Joel Hellewell, Sophie R. Meakin, Katharine Sherratt, Mingyuan Zhou, Rahi Kalantari, Teresa K. Yamana, Sen Pei, Jeffrey Shaman, Michael L. Li, Dimitris Bertsimas, Omar Skali Lami, Saksham Soni, Hamza Tazi Bouardi, Turgay Ayer, Madeline Adee, Jagpreet Chhatwal, Ozden O. Dalgic, Mary A. Ladd, Benjamin P. Linas, Peter Mueller, Jade Xiao, Yuanjia Wang, Qinxia Wang, Shanghong Xie, Donglin Zeng, Alden Green, Jacob Bien, Logan Brooks, Addison J. Hu, Maria Jahja, Daniel McDonald, Balasubramanian Narasimhan, Collin Politsch, Samyak Rajanala, Aaron Rumack, Noah Simon, Ryan J. Tibshirani, Rob Tibshirani, Valerie Ventura, Larry Wasserman, Eamon B. O'Dea, John M. Drake, Robert Pagano, Quoc T. Tran, Lam Si Tung Ho, Huong Huynh, Jo W. Walker, Rachel B. Slayton, Michael A. Johansson, Matthew Biggerstaff, and Nicholas G. Reich. 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. URL `https://www.pnas.org/doi/full/10.1073/pnas.2113561119`. Publisher: Proceedings of the National Academy of Sciences.

Katharine Sherratt, Hugo Gruson, Rok Grah, Helen Johnson, Rene Niehus, Bastian Prasse, Frank Sandmann, Jannik Deuschel, Daniel Wolffram, Sam Abbott, et al. Predictive performance of multi-model ensemble forecasts of covid-19 across european nations. *Elife*, 12:e81916, 2023.

Felipe J. Colón-González, Leonardo Soares Bastos, Barbara Hofmann, Alison Hopkin, Quillon Harpham, Tom Crocker, Rosanna Amato, Iacopo Ferrario, Francesca Moschini, Samuel James, Sajni Malde, Eleanor Ainscoe, Vu Sinh Nam, Dang Quang Tan, Nguyen Duc Khoa, Mark Harrison, Gina Tsarouchi, Darren Lumbroso, Oliver J. Brady, and Rachel Lowe. Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. *PLOS Medicine*, 18 (3):e1003542, March 2021. ISSN 1549-1676. doi: 10.1371/journal.pmed.1003542. URL `https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003542`. Publisher: Public Library of Science.

Anton Camacho, Adam Kucharski, Yvonne Aki-Sawyerr, Mark A White, Stefan Flasche, Marc Baguelin, Timothy Pollington, Julia R Carney, Rebecca Glover, Elizabeth Smout, et al. Temporal changes in ebola transmission in sierra leone and implications for control requirements: a real-time modelling study. *PLoS currents*, 7, 2015.

Estee Y. Cramer, Yuxin Huang, Yijin Wang, Evan L. Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J. Castro Rivadeneira, Aaron Gerding, Katie House, Dasuni Jayawardena, Abdul Hannan Kanji, Ayush Khandelwal, Khoa Le, Vidhi Mody, Vrushti Mody, Jarad Niemi, Ariane Stark, Apurv Shah, Nutcha Wattanchit, Martha W. Zorn, and Nicholas G. Reich. The United States COVID-19 Forecast Hub dataset. *Scientific Data*, 9(1):462, August 2022b. ISSN 2052-4463. doi: 10.1038/s41597-022-01517-w. URL `https://www.nature.com/articles/s41597-022-01517-w`. Number: 1 Publisher: Nature Publishing Group.