Evaluating infectious disease forecasts with allocation scoring rules

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

December 20, 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 that is often used to score quantile forecasts, 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 guide 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; Bilinski 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 2.2.1. 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 Murphy (1993) 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 (e.g., Catt et al., 2007; Petropoulos et al., 2019; Palmer, 2002; Pappenberger et al., 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; Murphy, 1993; 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. Bilinski et al. (2023) explore ways in which policymaker preferences could inform risk thresholds for predictive models using a framework for measuring the costs and losses associated with taking an action that is similar to methods that have been used in meteorology and elsewhere. Marshall et al. (2023) develop a forecast scoring rule that is informally motivated by utility considerations, but the score is not derived from a decision-theoretic set up. Separately, there is a thread of literature that quantifies the link between infectious disease modeling and policy making 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 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:

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

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 the forecast distribution of future levels 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 σ_i (see section 4 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 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 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. 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

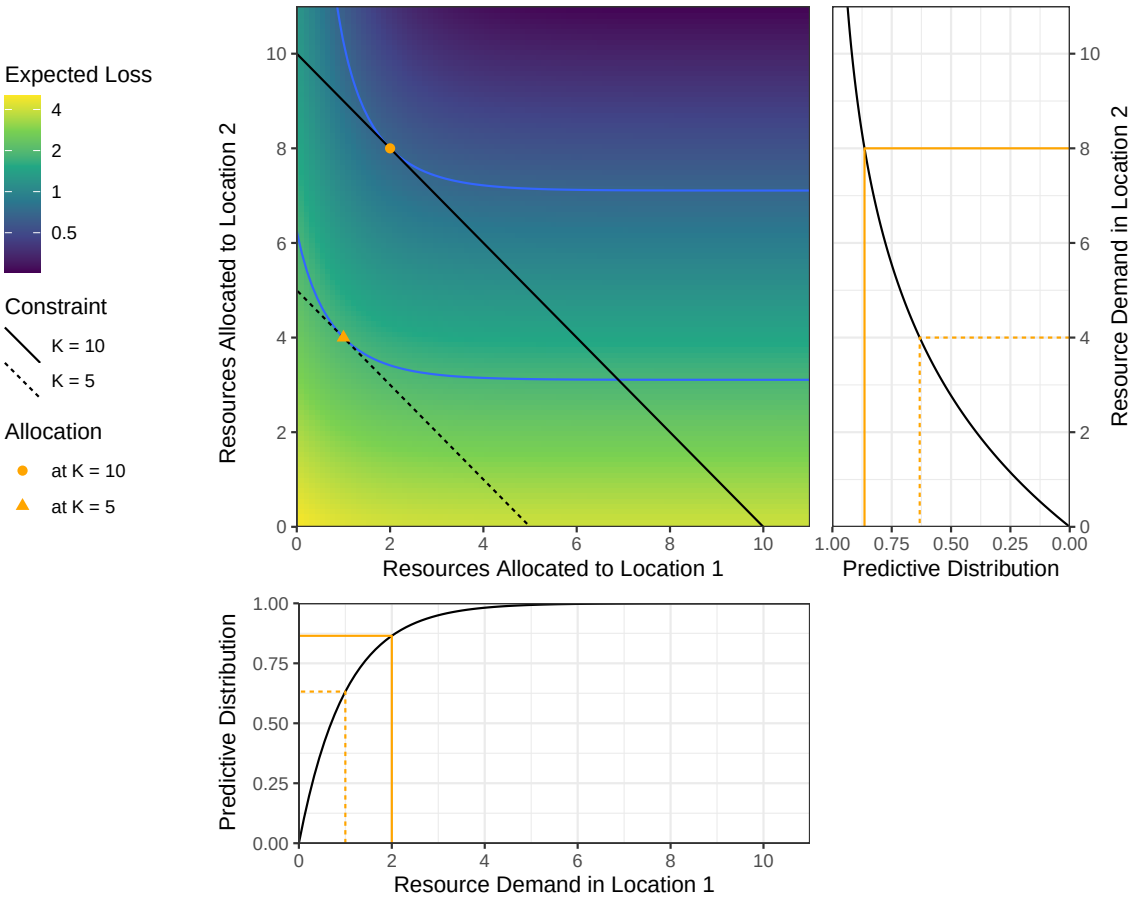

Figure 1: An illustration of the resource allocation problem in Example 1. There are $N = 2$ locations, with predictive distributions $F_1 = \text{Exp}(1)$ and $F_2 = \text{Exp}(1/4)$. The cumulative distribution functions of these distributions are illustrated in the panels at bottom and right. In the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

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 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 (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. 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 y_i 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 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. We also note that our problem formulation 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 derive a general form of the Bayes act by beginning with the fact that a condition for minimizing expected loss subject to the resource constraint is that there is no way to decrease expected loss further by shifting a small amount of the resource from one location to another. From this starting point, we can 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 $l_{\text{oracle}}(y; K)$ 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 $L \cdot \max(0, \sum_{i=1}^{N} y_i - K) = 0$ and there is no adjustment. Thus

the allocation score coincides with the raw score,

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

On the other hand, when the oracle cannot cover all need and incurs a loss of $L \cdot \sum_{i=1}^{N} y_i - L \cdot K > 0$, we adjust the raw score by this loss to get

$$\begin{aligned} S_A(F, y; K) &= S_A^{\text{raw}}(F, y; K) - L \cdot \sum_{i=1}^{N} y_i + L \cdot K \\ &= L \cdot \sum_{i=1}^{N} \Big\{ \max(0, y_i - x_i^{F,K}) - y_i \Big\} + L \cdot K \\ &= L \cdot K - L \cdot \sum_{i=1}^{N} \min(x_i^{F,K}, y_i) \quad \text{if } \sum_{i=1}^{N} y_i > K. \end{aligned} \tag{7}$$

This 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 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 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 be derived using the decision theoretic 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) 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. Many early COVID-19 modeling efforts, including the US COVID-19 Forecast Hub and other collaborative forecasting projects, adopted a quantile forecast format as a matter of convenience (e.g., this format does not require the forecaster to pre-specify an upper bound for disease counts as would be required to collect forecasts in the format of bin probabilities) (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 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, 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. **APG:[Maybe a rewording and/or a supplement reference here to mention that any WIS approximates the a corresponding weighted CRPS, and that WIS actually is the wCRPS for a discrete measure/weighting.]** *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. **APG:[Refs to supp for quantile score derivation and detail about averaging.]**

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 score can be defined in terms of the shortage or excess of resource procurement levels against realized need where these levels are expressable as quantiles of the respective forecast distributions by virtue of minimizing expected shortages or excesses.

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

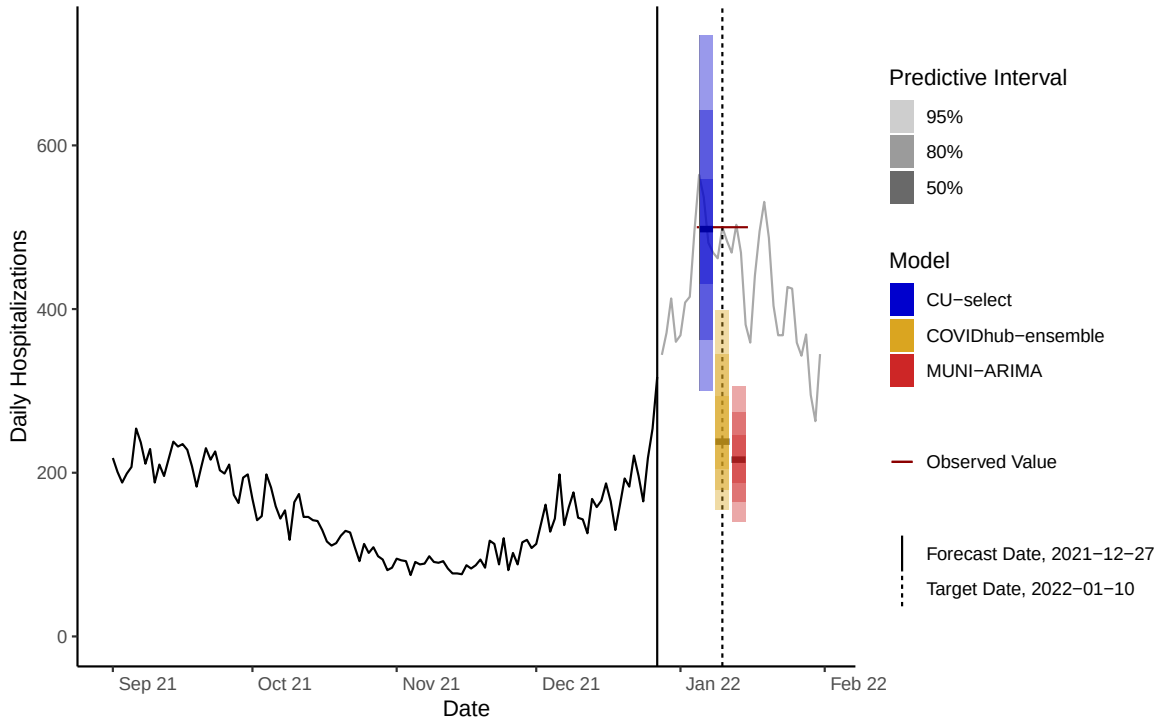

Figure 2: Hospitalization data and example forecasts for Virginia, US. New daily hospital admissions are shown by the line of data. The dark line indicates data up to the time the forecast was made on December 27, 2021; the grey line indicates data available in retrospect. The set of colored bars represent forecasts made for 14 days into the future, with a target date of January 10, 2022 (dashed vertical line). The horizontal dashed line indicates the level of the observed value on this day. Note that all three forecasts correspond to forecasts for the same day, but they are spread out for better visibility.

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. When constraints are likely to be large, and therefore of less importance, a composite score for this general scenario would converge to the WIS or CRPS. And if the constraints are likely to be tight, and therefore to be a determining factor, this composite score would converge to the IAS. We pursue this direction further in other work that is in progress.

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 states.

3.1 Data

3.1.1 Hospitalization data

Starting in the summer of 2020, the US Health and Human Services (HHS) began reporting counts of daily new admissions to hospitals for individuals with COVID-19 (HealthData.gov). 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 Wang et al. (2023). The hospitalization data were downloaded for analysis on December 03, 2023.

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 the University of Massachusetts-Amherst, collected short-term forecasts of new hospital admissions at the daily scale starting in December 2020, using the HHS data as a source of "ground truth" (Cramer et al., 2022b). 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 (Ray et al., 2023). 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 represented using a set of 23 quantiles for each individual prediction. The submitted quantiles included 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.

We established a set of 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), both defined above.

3.2.1 Allocation Score

For this analysis, we fixed the resource constraint K to be 15,000, based roughly on a reported number of ventilators available for reallocation in the US (Ajao et al., 2015). For each week, we computed the allocation score for the 14 day-ahead forecast.

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.

$$sr^{AS}_{m,w} = 1 - \frac{r^{AS}_{m,w} - 1}{n_w - 1} \tag{8}$$

3.2.2 Weighted Interval Score (WIS)

The Weighted Interval Score (WIS), as described in earlier sections, measures the alignment of a single probabilistic forecast (F) with an observation (y). We computed the mean WIS across all L locations for each model and each forecasted week as

$$MWIS_{m,w} = \frac{1}{L}\sum_{l=1}^{L} WIS(F_{l,m,w}, y) \tag{9}$$

where $F_{l,m,w}$ is the probabilistic forecast from model m for location l and week w. Using the same procedure as for allocation scores described above, we computed standardized ranks for MWIS ($sr^{MWIS}_{m,w}$).

3.3 Data and code availability

All forecast data used in this evaluation are available through the COVID-19 Forecast Hub (Cramer et al.). An R package implementing the allocation score is available at `https://github.com/aaronger/alloscore` and all code for the analyses presented in this manuscript is available at `https://github.com/aaronger/utility-eval-papers`.

3.4 Application results

3.4.1 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 Omicon wave peak nationally, with individual states typically observing a peak at or after January 3, 2022.

For many locations, forecasts predicted lower values than were eventually observed, as this was during the period of steep increase of viral transmission across many states. Of the 10 models evaluated,

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
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: 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.

Forecasts and allocations for a selection of states with high numbers of hospitalizations on January 3, 2022 reveal mechanics about how allocations are made (Figure 3). As described in the methods above, allocations for a given location from a specific model are generated by finding the single quantile of the probabilistic forecast across all locations that reaches the resource allocation limit (in this case, assumed to be 15,000). A direct comparison between the forecasts from the JHUAPL-Bucky and CU-select models show 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 3).

For JHUAPL-Bucky, the allocations were assigned at the 28th percentile of the predictive distribution and for CU-select they were at the 89th percentile. This reflects that in general the JHUAPL-Bucky forecasts assigned more probability to higher values and thus a low quantile value across all locations reached the allocation limit of 15,000. In California, New York and Texas, the eventually observed number of hospital admissions was closer to the predicted median from JHUAPL-Bucky than the median from CU-select, and was contained in JHUAPL-Bucky's 80% PI but not in the 80% PI from CU-select. However for each of these three states, the resource allocation was lower for JHUAPL-Bucky than for CU-select and therefore resulted in more hypothetical unmet need. In Florida, which was the state that saw the largest number of reported hospitalizations on this target date, the forecast from JHUAPL-Bucky missed by a wide margin (as did the forecast from CU-select) and the allocation was 716 units lower than the allocation derived from the CU-select model.

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 4A). These allocation errors resulted because that model's forecasts did not consistently capture the relative resource needs across different states. (Figure 3 & 4A). 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.

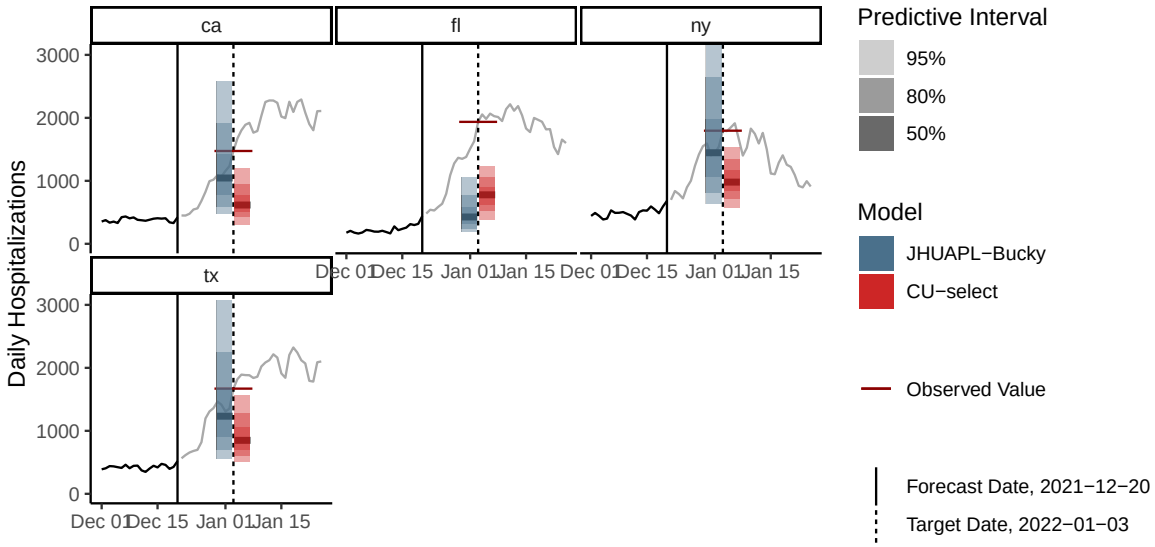

Figure 3: Probabilistic forecasts for new hospital admissions and the inferred resource allocations for COVID-19 on January 3, 2022 for the states with the nine highest hospitalization counts. For each state, the dark black line shows the data observed when the forecast was made and the grey line shows eventually observed counts. The side-by-side shaded regions show the median (solid 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 red and yellow shading show the allocations: the red bar goes from zero to the level of the observed number of hospitalizations while the yellow bar shows how many resources were suggested by the model to be allocated to that location. The allocations (yellow bar) may be more or less than the observed number, therefore exposed red bar indicates unmet need, while the yellow bar extending above the observed number of hospitalizations means that excess resources were suggested for that given location.

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 4B). Predictions from the JHUAPL-Bucky model were wider, and included the observed level of daily hospital admissions more often. Therefore, WIS did not as often assign that model severe penalties for forecasts that underpredicted or overpredicted the actual hospitalization count.

Together, these results corroborate the understanding that the allocation score and weighted interval score measure different aspects of forecast performance. The allocation score penalizes models that mischaracterize the relative magnitude of resource need across different locations, while WIS penalizes forecasts that do not capture the absolute magnitude of the target quantity in each location.

3.4.2 Forecast scores showed differences over time

Allocation scores varied substantially by date and by model (Figure 5). 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 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, 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. Predictions made during the peak week and just after showed the highest variation in allocation scores, with some models having allocation scores under 1000 and others having values over 3500.

Overall, across the first 11 weeks evaluated (we excluded the last four since nearly all the models achieved an allocation score of zero), 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 four weeks, more than any other model except the `COVIDhub-ensemble` which also had four.

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.

3.4.3 Metrics were not consistently correlated over time

Models showed differing levels of correlation between their allocation scores and MWIS values (Figure 6). 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`).
- Several models had consistently strong MWIS ranks but also had highly variable allocation score ranks with no clear association between the two (e.g., `JHUAPL-SLPHospEns` and `CU-select`)
- One model performed consistently well for both metrics with no clear association (`COVIDhub-ensemble`)
- One model performed consistently well for allocation score but had only middlings ranks for MWIS (`CMU-TimeSeries`)

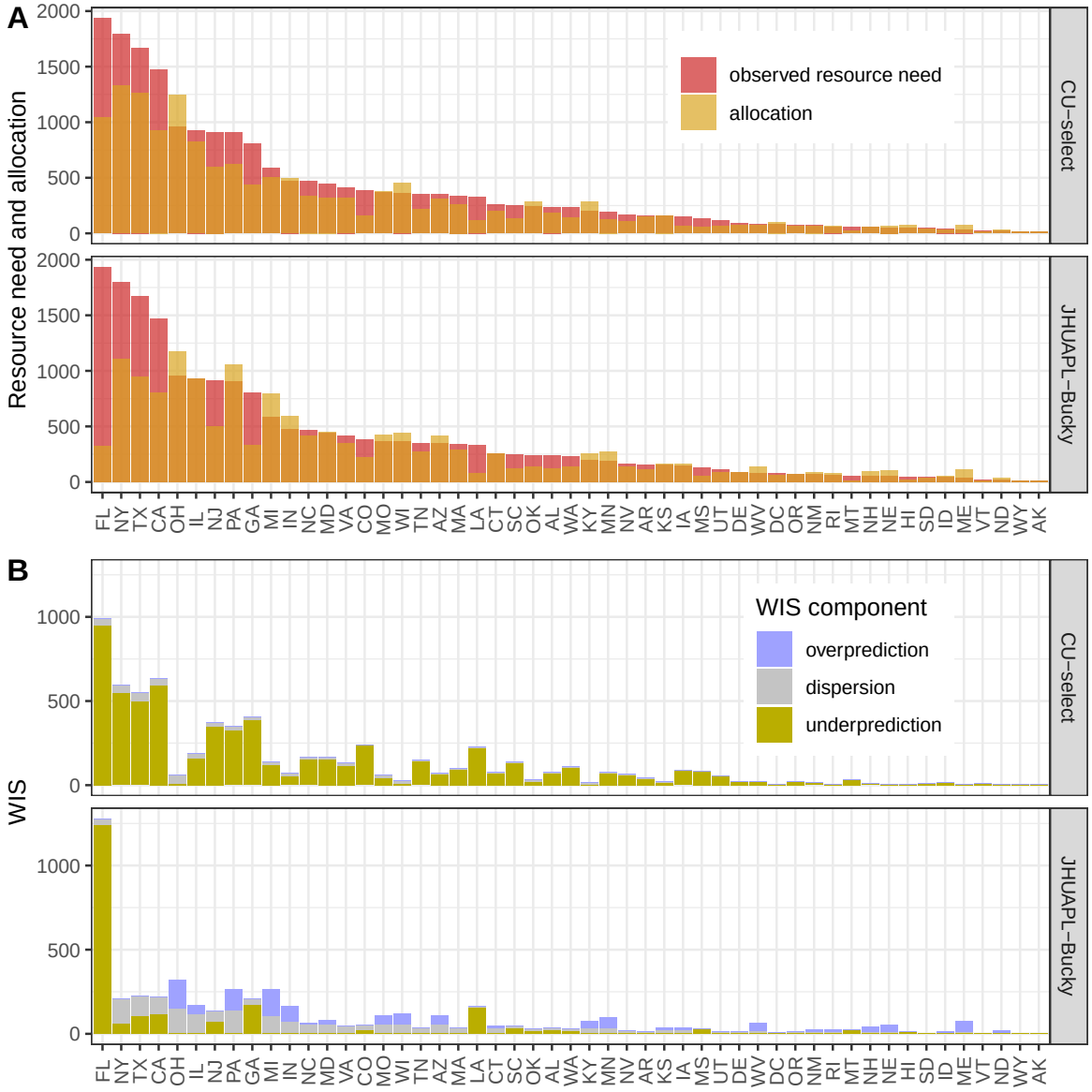

Figure 4: Component-wise breakdowns of the allocation score (Panel A) and weighted interval score (Panel B), by location for forecasts of hospitalization admissions on January 3, 2022, for two selected models (JHUAPL-Bucky and CU-select). Panel A shows the observed resource need, in this case the observed 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.

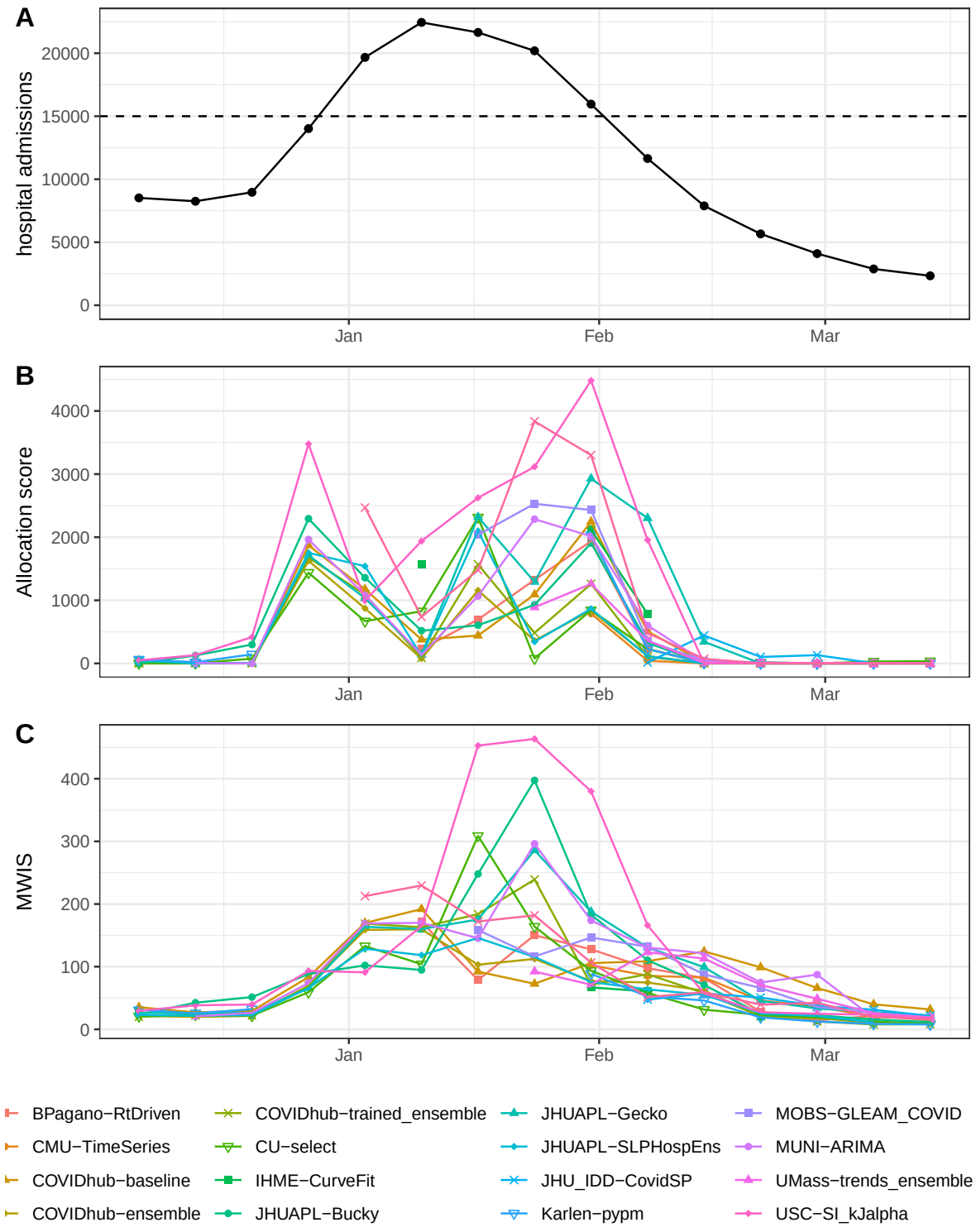

Figure 5: Hospital admissions and evaluation metrics over time. Panel A shows the number of hospital admissions in the US as a whole due to COVID-19 on a sequence of 15 Mondays from December 2021 through March 2022. These are the dates for which forecasts were made and evaluated. A horizontal dashed line at 15,000 shows the assumed 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 for which the prediction was made. 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). 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.

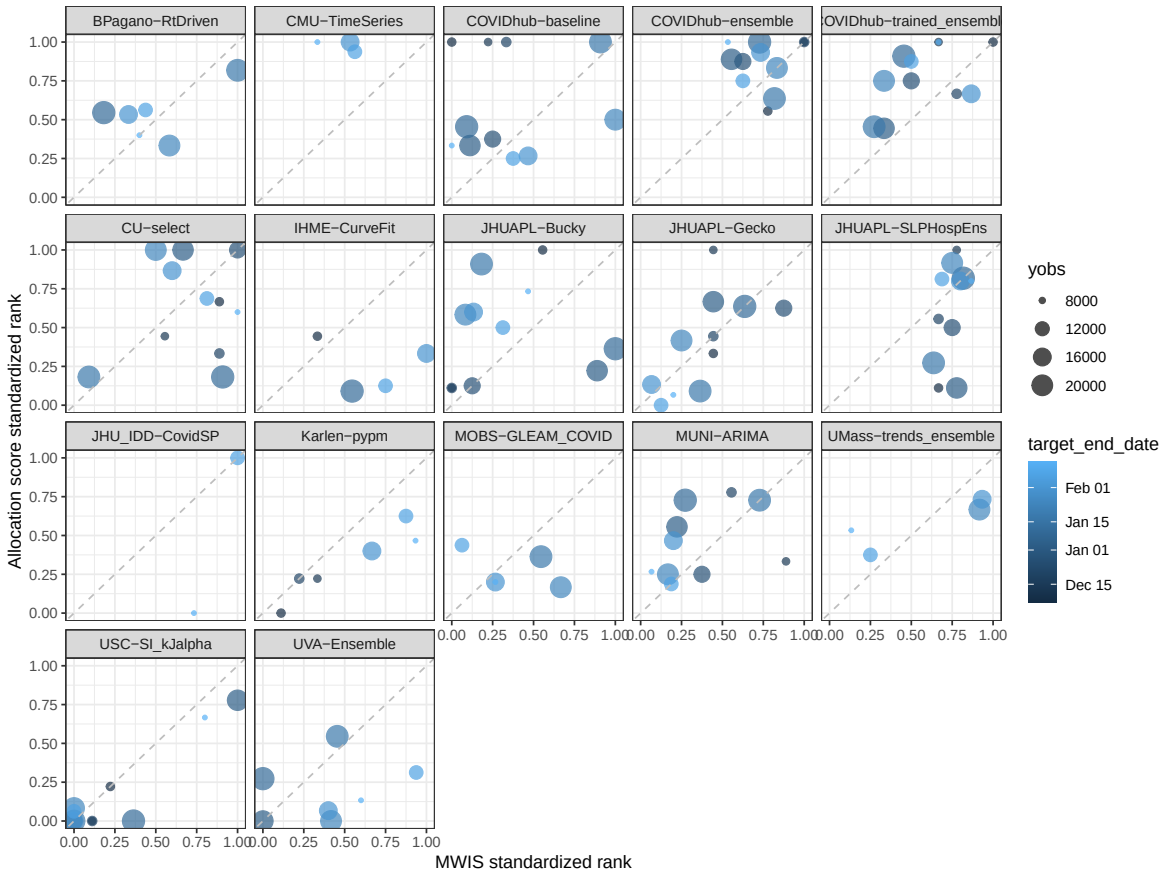

Figure 6: Association of standardized ranks for MWIS and allocation score by model and week. Each facet of the plot corresponds to one model. Within each facet, each point corresponds to a week. The x- and y-values correspond to the MWIS standardized rank and the allocation score standardized rank for 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.

3.4.4 Integrated allocation score across values of K

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). This could be useful in situations where the actual constraint may not be known precisely, or where we wish to consider results across a range of decision making settings with different constraints. Unlike in the above analyses where we conducted analyses assuming that K was known to be 15,000, the analyses in this subsection are conducted assuming different distributions on 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 (Figure 7A). These calculations highlight that AS was highest at the K nearest to the observed value of 19,581. 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.

IAS was computed for two distributions on K, one uniform across the entire range and the other a symmetric distribution around 15,000 (Figure 7B). 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 7C). Model rankings based on the AC and the centered IAS were roughly similar, with the top and bottom two models 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 rank probability score (such as the weighted interval score, or WIS) have been frequently utilized to evaluate probabilistic forecasts. Often, these 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. We have demonstrated that forecast evaluation methods that are tied to a specific decision making context can yield model rankings that are substantively different from standard measures of forecast skill.

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. However, we do note that the decision-making context presented in this work, while motivated by examples from real-world public health resource allocation problems, has not been used to inform an actual real-time decision-making process.

Resource allocation decisions are a realistic example of the kinds of decisions that could motivate more targeted forecasting exercises. One real-world example is allocation of ventilators during a respiratory viral pandemic(Huang et al., 2017). This was the motivating example for the examples presented in the work above. However, other examples where allocation is a central concern exist as well, such as 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).

In practice, there are many potential users of forecasts with many different decision making problems.

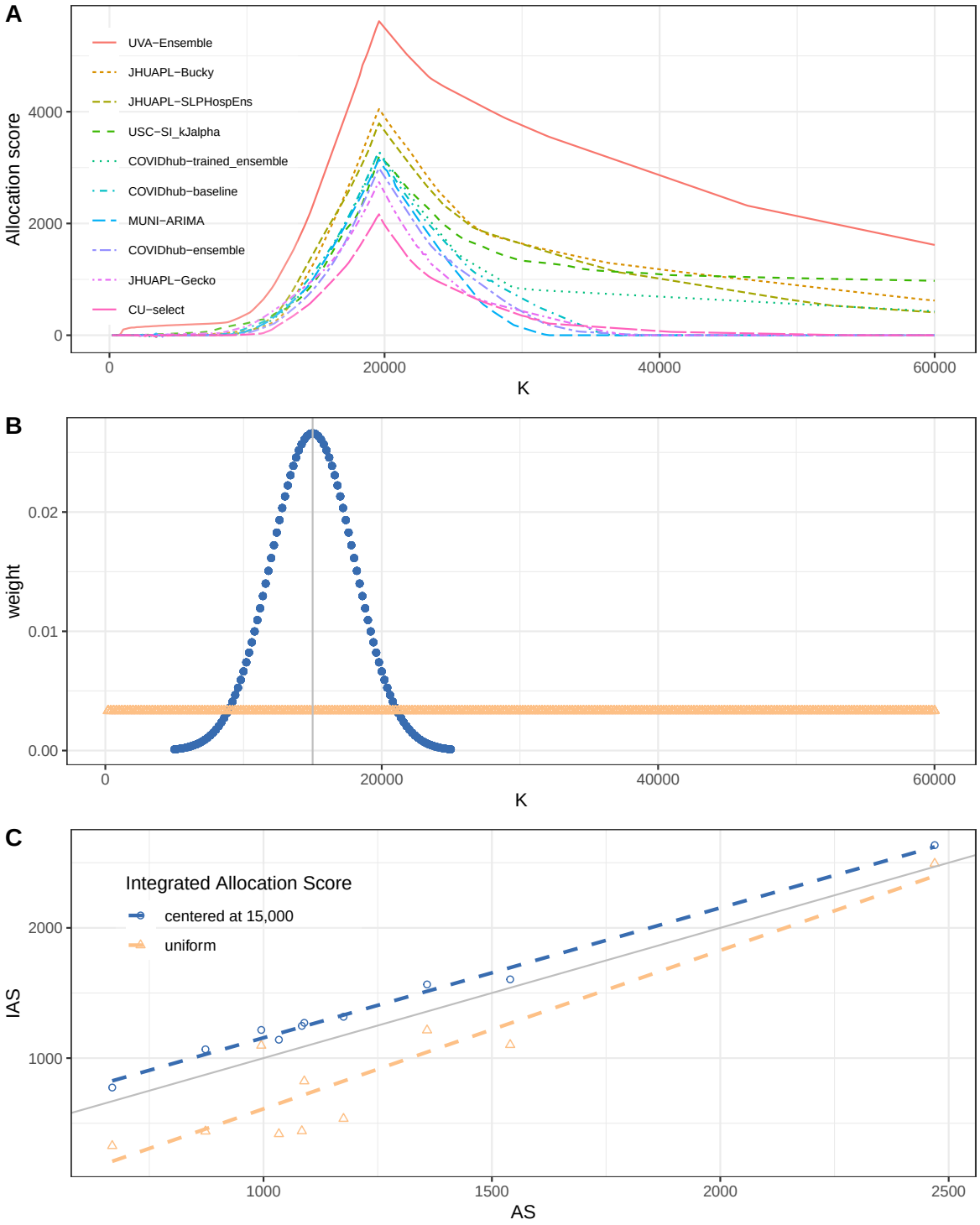

Figure 7: Allocation scores (AS) across different resource constraints (K) for 10 models that made forecasts on 2021-12-20. Panel A shows, for each model, the AS for values of K between 200 and 60,000 at increments of 200. The AS show a sharp peak just under 20,000, near the eventually observed number 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.

Not all can be easily quantified. Those that can be easily quantified may differ enough that it seems likely that no single score would be appropriate for all users. Ideally, a forecasting tool could be developed through close collaboration between modelers and public health officials. However, this may only be possible in settings with 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.

In our specific application, we made several key observations that we believe should inform future work in this area. First, it is clear that different metrics (in our case allocation score and mean weighted interval score) captured different aspects of forecast performance. The mean WIS (MWIS) was strongly dependent on the scale of the forecasted quantity (e.g., when hospitalizations were high, so were MWIS scores). The allocation score was not as scale-dependent, as the highest/worst allocation scores were observed when the number of hospitalizations was closest to the allocation constraint. We also note that contributed models found it hard to achieve a better allocation score than a naïve baseline model that just predicted a flat line from the last previous observation with wide uncertainty. This suggests suggests that contributed models (other than the ensemble, which combined forecasts from all contributed models) were not consistently adding value over just extrapolating from the current levels.

There are several important limitations to the current work. The allocation score we developed here does not directly account for important considerations such as fairness or equity of allocations. Nor does the proposed framework 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 cannot be used to evaluate a forecast at a horizon where the outcome may have been impacted by such a decision.

An additional limitation of the current work and opportunity for further investigation is to more carefully evaluate whether a forecast adds value to existing decision making processes. In the context of allocation-based decisions, standard procedures might involve extrapolating need based on a current observed data (similar to the 'baseline' model presented above), with or without adjustments based on other political or real-world considerations. For example, in many settings public health stakeholders will synthesize multiple sources of information, e.g. real-time data from a wide variety of quantitative and qualitative sources coupled with expert judgment based on a situational understanding of past and current experience. The allocation score presented in this work assesses the optimality of the allocation with reference an omniscient view of the future captured by the oracle. However, what we care about is whether a given forecast adds useful information to 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) to quantify the benefit—if any—of using a forecast to inform a specific decision.

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. We note that a few papers have begun exploring similar linkages as described in the literature review–but

we see much room for additional work in this area. In some situations, individual or ensemble models could be developed to optimize scores that are attuned to a particular decision making setting. This area is also largely unexplored in the realm of public health to date, with some initial methodological development in econometrics (Loaiza-Maya et al., 2021).

In conclusion, we argue that the way modelers and policymakers view and evaluate forecasts should change depending on the specific decision-making context. Using standard forecast evaluation metrics can mask the utility of certain forecasts, or lead to forecasts being used to inform a particular decision that are not well matched to that decision making context. 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.

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.

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. 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-through-pandemic-threats/` (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.

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. URL `https://www.pnas.org/doi/10.1073/pnas.2302528120`. Publisher: Proceedings of the National Academy of Sciences.

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.

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.

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.

Peter Maurice Catt et al. Assessing the cost of forecast error: A practical example. *Foresight: The International Journal of Applied Forecasting*, 7:5–10, 2007.

Fotios Petropoulos, Xun Wang, and Stephen M Disney. The inventory performance of forecasting methods: Evidence from the m3 competition data. *International Journal of Forecasting*, 35(1): 251–265, 2019.

Tim N Palmer. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 128(581):747–774, 2002.

Florian Pappenberger, Hannah L Cloke, Dennis J Parker, et al. The monetary benefit of early flood warnings in europe. *Environmental Science & Policy*, 51:278–291, 2015.

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. 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, et al. 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, et al. 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, et al. 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, 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. 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, 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, et al. 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, et al. Temporal changes in ebola transmission in sierra leone and implications for control requirements: a real-time modelling study. *PLoS currents*, 7, 2015.

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. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2017.08.006. URL `https://www.sciencedirect.com/science/article/pii/S0022053117300844`.

Tao Hong, Pierre Pinson, Shu Fan, et al. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016.

Estee Y. Cramer, Yuxin Huang, Yijin Wang, et al. 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.

HealthData.gov. COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries | HealthData.gov. URL `https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh`.

Yijin Serena Wang, Ariane Stark, Evan L Ray, et al. *covidHubUtils: Utility functions for the COVID-19 forecast hub*, 2023. URL `https://github.com/reichlab/covidHubUtils`. R package version 0.1.7.

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. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2022.06.005. URL https://www.sciencedirect.com/science/article/pii/S0169207022000966.

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. ISSN 1935-7893. doi: 10.1017/dmp.2015.105. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4636910/.

Estee Y. Cramer, Yuxin Huang, Yijin Wang, et al. reichlab/covid19-forecast-hub: release for zenodo, 20210816. URL https://zenodo.org/record/5208210.

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. URL https://wwwnc.cdc.gov/eid/article/23/6/16-1417_article.

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. ISSN 1572-9389. doi: 10.1007/s10729-012-9199-6. URL https://doi.org/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. ISSN 2468-2667. doi: 10.1016/S2468-2667(23)00061-0. URL https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(23)00061-0/fulltext. Publisher: Elsevier.

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. ISSN 1097-0258. doi: 10.1002/sim.9238. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9238. _eprint: https://onlinelibrary.wiley.com/doi/pdf/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. URL https://wwwnc.cdc.gov/eid/article/29/3/22-0761_article.

Ruben Loaiza-Maya, Gael M. Martin, and David T. Frazier. Focused Bayesian prediction. *Journal of Applied Econometrics*, 36(5):517–543, 2021. ISSN 1099-1255. doi: 10.1002/jae.2810. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2810. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.2810.