Author response to reviewer comments on manuscript ID JRSSA-Mar-2024-0080

We would like to thank the referees and the associate editor for their valuable feedback and comments, and we appreciate the opportunity to revise and resubmit the manuscript. Below, we try to provide responses to each point of concern or inquiry from the reviewers. Text from the reviewers is marked by a left-hand bar and our responses follow inline. We refer to changes in the manuscript using magenta hyperlinks, which jump to relevant points of a revision-marked manuscript which is included below the response. Additions appear there in blue with wavy underlining (or inside of a blue box in the case of the added figure) and removed text appears in red and is struck through.

Comments from the Associate Editor

> Two referees and I have read this manuscript carefully; their reviews can be viewed in ScholarOne. I agree with them that the manuscript is well written, addresses an important problem in public health with sound statistical approaches, and is consistent with the mission of the journal.
>
> The approach is well motivated by recent epidemics (Ebola, H1N1, Covid-19). Using the theory of scoring rules, the authors propose a theoretically sound AS (allocation score) as a way to evaluate forecasts of resource allocations in terms of how much the benefit received actually met the resource needs. This formalization of the problem into one that can be approached through the use of scoring rules is creative. As Tukey once said, "Sometimes finding the question is harder than finding the problem" and the authors have formulated the problem in a clever way to enable a sensible and statistically sound approach to it. For policy makers, the AS can be related to policy impact.

We thank the AE for the encouraging words and feel that Tukey's saying frames perfectly the project from which this paper emerged.

> Both referees offer suggestions that, if adopted, will serve to benefit the authors, in that their suggestions will strengthen the paper and likely inspire greater use of their proposed AS. Both of the suggestions from Referee 2 are echoed within two of the comments from Referee 1.
>
> Referee 1, #5, and Referee 2, #1, both ask about the potentially joint impacts of multiple locations, both in terms of geography and demographics. Can forecasts from different, likely correlated, areas be assessed jointly?

This is a very subtle and important question, but it is not addressed directly by the present formulation of the AS. To clarify that the AS has no immediate dependence on a forecaster's underlying beliefs regarding joint probabilities we added a paragraph after paragraph 4 on page 8 at the end of Section 2.3. We also renamed Section 2.3 to better describe these considerations.

> The comments from Referee 1, #4, and Referee 2, #2, both relate to interpretability of AS and how other public health measures, that may be under consideration simultaneously, may affect it.
>
> The suggestions from Referee 1, #1, are good ideas that, if implemented, will surely strengthen the paper. A full-blown sensitivity analysis may be beyond the scope of the paper, but some limited illustrations will be important to reassure users of its relative robustness to assumptions.

[author response point 1] We agree that demonstrating relative robustness to assumptions is crucial for promoting practical use of the AS, and had this in mind when writing Section 3.5.4, and especially when composing Figure 6. The section title, however, only mentioned the integrated AS. To better draw attention to the issue of robustness we have split the section into two sections, 3.5.4 and 3.5.5. and begin the title of Section 3.5.4 with "Sensitivity".

We see Figure 6A as demonstrating not only how AS ranks will tend to be stable across a wide range of constraints but also illustrating a key insight about the AS — that it is, in fact, quite sensitive to the resource level outside this central range but that this is a desirable and natural feature reflecting the limited impact forecast selection will have when resources are too scarce or too abundant for allocation decisions to matter.

> Referee 1's comment 2 refers to clarification of computations and can be easily addressed. I believe the authors address Comment 3 ("practical computation of this score") in Section 3.4, which Referee 1 may have overlooked. Comment 6 ("connections with other fields") is noted briefly on MS p2, paragraph 3; perhaps a sentence or two along these lines can be added in your final Discussion section.

We are grateful to the AE for organizing the reviewers' comments and for the helpful suggestions.

> Minor: 1. I advise against starting a sentence with "And"; e.g., p2, l.8: "And" is not needed at all, although you could change "And the WIS has beeen used during the pandemic..." to "Likewise, the WIS was used during the pandemic..." Also p2, para 4, l.3: "And within this body of work we have found the discussion of such a connection to still be at a ..." -> "Even within this body of work we have found the discussion of such a connection to be at a ..." [i.e., change "And" to "Even" and delete "still"]

These changes have been adopted here and here.

> 2. In the Landau (2021) reference (p26, 4th reference), per https://joss.theoj.org/papers/10.21105/joss.02959 both "r" and "m" (in "r package" and "make-like") are capitalized. I did not verify the accuracy of all your references (but you should).

We have double checked the accuracy of all of our references, including capitalization.

> 3. Please see JRSSA submissions information for style in "References" (pp 24-27). They should be ordered alphabetically by Last Name of lead author.

We have reordered the references alphabetically by lead author last names using a version of the `rss.bst` style file with a slight modification needed to truncate the long author lists of several of our references.

Comments from Referee 1

> The manuscript presents a novel scoring rule for evaluating infectious disease forecasts based on their utility in resource allocation decisions, specifically in the allocation of medical resources. The work is timely and relevant, aiming to fill a notable gap in the existing literature by directly connecting forecast accuracy with public health decision-making, an aspect that has gained increased relevance due to recent global health crises. While the proposed allocation score (AS) is innovative and potentially impactful, especially for public health agencies, there are several areas where the manuscript could be improved to enhance its clarity, completeness, and impact.

We thank the referee for the optimistic assessment of the manuscript's practical potential.

> 1. The authors did not sufficiently discuss how the proposed AS adapts to varying public health scenarios with different resource constraints. Resource availability can vary significantly by region, disease, or over time. It would therefore be advantageous for the authors to include discussions on the following aspects:
>
>   - Calibration techniques to ensure the score remains meaningful across scenarios with different levels of resource availability.
>   - A sensitivity analysis demonstrating how the AS reacts to changes in resource constraints, its robustness, and reliability.
>   - Strategies for updating the AS in real-time, based on fluctuations in resource availability or shifts in disease spread.

We certainly agree that it is important to convey that resource availability is a parameter of the AS for which potential users are responsible, and that any forecast evaluation results using the AS should be accompanied by explanations of how the corresponding resource constraint levels were chosen (or weighted in the case of the IAS). On this issue please see author response point 1 above and the changes involving sections 3.5.4 and 3.5.5 it discusses. We see Figure 6 as an efficient means of highlighting how the ranking and relative differences in scores of the forecasts in this example were in fact stable across a contextually relevant range of constraint levels.

We are less sure how to address the concern regarding calibration, which we understand to be a property of forecasts rather than scores. Referring again to author response point 1 we would argue that it is more a feature than a bug of the AS that is becomes less meaningful (and less coherent as seen in the tails in figure Figure 6A) under extreme scarcity or abundance.

As for updating strategies, it seems worth noting that in our implementation the AS in computed in parallel for a range of constraint levels (as illustrated in sections 3.5.4/5). Thus fluctuating realized resource constraints would not, in themselves, present a computational barrier. And as disease spread shifts, we would expect the AS to change accordingly and indicate which forecasts were able to better anticipate such shifts for the purposes of resource allocation.

That said, sequential decision-making would introduce other complexities that are not captured by the formulation of the AS introduced in this paper, and we have added a new paragraph (beginning with "Secoondly") to the discussion outlining these challenges.

> 2. Appendix Section C outlines the numerical methods used for computing allocation Bayes acts, but it lacks an evaluation of the effectiveness of these methods. Assessing the accuracy and reliability of these methods is essential for verifying their applicability in public health decision- making. It would be more convincing if the authors could conduct some empirical tests, includ- ing simulated scenarios, to demonstrate the effectiveness of the proposed rule.

We have rewritten Appendix Section C and created a new figure in order to better convey the simplicity of our numerical methods (which boil down essentially to finding the root of a monotonic function) and reassure the reader that they would be unlikely to be the source of any numerical instability jeopardizing the applicability of the AS in public health decision-making. We also note that in order to refer to the closed form of the allocation levels for the examples of section 2.1 we numbered the relevant formula in Appendix section B.

> 3. The manuscript introduces the integrated allocation score (IAS) in Section 2.2.3. While the concept is well-founded, it does not describe the practical computation of this score. Have the authors developed any algorithms for these calculations? If so, it would be better to include a brief discussion.

We have not developed any special algorithms for the purpose of calculating the IAS, and see investigations as likely requiring the kind of *ad hoc* approximation we use in (what is now) section 3.5.5. To clarify this we have added text after our definition of the IAS. We also added some clarifying text to section 3.5.5 to indicate how an approximation to the IAS was computed in our example. We would expect such an approximation to be available and appropriate in general.

> 4. The proposed AS rule focuses on addressing immediate resource needs like hospital admissions and static resources. While this focus is practically significant, it may be too simplistic to capture the complex, dynamic nature of public health management. The framework notably excludes preventative measures such as vaccines, which are crucial for reducing future disease incidence. It would be beneficial for the authors to discuss the challenges and potential method- ologies for incorporating these preventative aspects and dynamic resources into the scoring rule.

In order to be more transparent regarding the scope of this paper we have rewritten part of the discussion section to collect limitations mentioned earlier (such as those relating to vaccines in paragraph 2 of section 2.2.2) and give some brief methodological suggestions on how they could be confronted.

> 5. The authors propose using unmet need as a loss function, which simplifies the model. However, this approach may not fully account for the diverse impacts of unmet needs across different locations and demographics. Could the authors explore the possibility of incorporating different weights or more complex loss functions to provide deeper insights, such as considering factors like population vulnerability.

We agree that the loss function we have formulated makes many simplifying assumptions, but we feel this simplification is necessary to keep the discussion in the paper of a manageable length. Some of the complexities the referee mentions were noted in section 2.2.2, where we first introduce the loss function: "A variety of extensions are possible. . . ." However, we agree that these are important limitations of our initial formulation of the AS, and we have therefore added an additional mention of

these possible extensions in paragraph 7 of the revised discussion section (which begins with "Secondly, our formulation of the AS...").

> 6. To make the work more accessible to a general audience, it could be beneficial to establish connections with other fields such as economics, sociology, or urban planning. This could enrich the resource allocation framework presented in this paper, while also allow the findings to potentially benefit decision-making processes in those fields.

We have added several references to the end of the first paragraph in the discussion section, indicating other fields where decision making based on forecasts might benefit from the perspective we develop. We also sought a more general tone in our concluding paragraph by addressing it to decision-makers rather than just public health officials.

Comments from Referee 2

> Comments to the Author Summary: Gerding et al. advocate for new scoring rules for infectious disease forecasts that are explicitly based on resource allocation and other policy decision considerations. The authors consider a situation, where a forecaster produces multiple forecasts for different geographical locations that are under the jurisdiction of the same policy making agency (e.g., counties in a state or states/provinces/regions in a country). This is a reasonable setting that mimics policy decisions that had to be made during the COVID-10 pandemic and that are routinely being made during flu/COVID/RSV seasons. In particular, Gerding et al. develop their scoring rule using a decision theoretic framework with a loss function equal to the number of unmet need in bed allocations across all geographic areas under consideration. I think the manuscript is definitely a step in the right direction. It is really well written, with methods rigorously evaluated. Congrats on a great paper!
>
> Paper strengths:
>
> 1. Evaluation of probabilistic infectious disease forecasts is an important area of research, so the authors' contribution to this area of research is timely.
>
> 2. The authors operate within a rigorous decision theoretic framework when deriving their score of the forecast skill and think deeply about policy decisions.

We thank the referee for the positive assessment of our work and highlighting feautures that we also feel make the larger research project worthwhile.

> 1. It is my understanding that currently most forecasts evaluated by the authors operate by producing forecasts independently across spatial locations. In other words, most forecasters do not produce joint probabilistic forecasts of all locations in the region of interest, for example using some complicated method that models all hospital beds in all graphic locations jointly. However, the score suggested by the authors explicitly evaluates these forecasts jointly. So my question is the following: is it possible that if the forecasts were produced jointly, the difference between WIS and AS would be less pronounced? Is it possible to illustrate this point on some toy forecasting problem? I think this could be important for stimulating development of multivariate forecasts across spatial locations – a difficult problem computationally, but important to solve eventually for us as a community.

We have tried to offer some initial clarification on the very important issue above in author response point 1, but the referee's question is a broad one and we feel that this paper is probably too confined a venue for us to do it justice. It does seem likely to us that if a forecaster tries to improve their forecast by introducing valid and identifiable inter-location dependency terms into their model, that this will tend to improve AS more than it improves WIS. It also seems likely that the AS will be more effective than WIS at discriminating between forecasters using better or worse inter-location dependency specifications. But testing such conjectures is difficult since we cannot "control" for the marginal forecasts, and these are the only direct inputs that the AS sees.

> 2. One of the strengths of the new AS is its interpretability. I would like to encourage the authors to highlight this strengths more. One common complaint about WIS that I hear from applied epidemiologists and policy makers is that it is hard to understand what it means. The AS, if I understand it correctly, doesn't have this problem. This opens an opportunity for not only ranking forecasts but also identifying if they are meaningfully different from each other. For example, two forecasts with AS scores of 400 and 380 may not be considered meaningfully different from each other if the total number of available hospital beds was 15,000. A policy maker may have prior knowledge that says that in practice implementing bed allocations across multiple locations according to a forecast is possible only up to some percentage of the total number of beds available. So +/- 20 beds could be smaller than the number of beds that would be misallocated due to logistical problems, human errors, medical personnel illnesses, etc.

We appreciate this encouragement and look forward to emphasizing the interpretability advantages of the AS in future applied work. In order to draw a bit more attention this strength we have rephrased the third paragraph of the introduction. The ability that the AS gives decision-makers to contextualize differences in forecaster scores is also now mentioned in the first paragraph of the discussion section.

Evaluating infectious disease forecasts with allocation scoring rules

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

Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts at Amherst

September 9, 2024

Abstract

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

Keywords: public health, forecast evaluation, proper scoring rules, resource allocation, epidemiology

1 Introduction

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

In the wake of the COVID-19 pandemic, this trend has been followed by calls for infectious disease forecasts to be not only designed, but also evaluated in ways that align specifically with how forecasts can be used to inform such outbreak control decisions (Marshall et al., 2023; Bilinski et al., 2023). This contrasts, however, with the historically standard practice of measuring the quality of disease forecasts using general purpose accuracy and skill scores, especially those that have implementations available in existing software when the relevant outbreak occurs. For point forecasts, the root mean square error (RMSE) (e.g., Papastefanopoulos et al. (2020)) and the mean absolute error (MAE) (e.g., Johansson et al. (2016)) are common choices. For probabilistic forecasts, which are the focus of this

*Corresponding Author: agerding@umass.edu

paper, researchers have often relied on the logarithmic score (LS). For example, the LS has been used to evaluate the skill of US seasonal influenza forecasts (McGowan et al., 2019; Reich et al., 2019) as well as forecasts targeting surveillance measures of dengue incidence in Peru and Puerto Rico (Johansson et al., 2019). More recently, the continuous ranked probability score (CRPS) and a discretized version adapted to multi-quantile forecasts, the weighted interval score (WIS) (Bracher et al., 2021), have gained prominence. For example, the CRPS was used to assess probabilistic forecasts (based on random effect models) of dengue incidence at the district level in Vietnam (Colón-González et al., 2021). ~~And~~ Likewise, the WIS has been used during the COVID-19 pandemic to evaluate forecasts of observed cases, hospitalizations and deaths in the US and Europe, as reported by municipal, state, and federal surveillance systems (Cramer et al., 2022b; Fox et al., 2022; Sherratt et al., 2023).

While ~~it should be noted that there are ways to interpret~~ any of these scores ~~abstractly~~ can be interpreted through the lens of decision theory, ~~and that all of the application-specific papers cited above benefited from direct collaboration with public health agencies, a key impetus for the present work has been that we were not able to find in any of them, nor in the literature they represent,~~ they lack explicit connections between how a forecast ~~was~~ is evaluated and how that forecast was used in practice. A key impetus for the present work was to develop a score that is interpretable to decision makers by measuring the extent to which forecasts lead to improved decisions. In addition, the allocation score we propose is expressed in units that are intelligible to decision makers.

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

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

In this work, we begin to fill this gap between the ways that infectious disease forecasts have traditionally been evaluated and the ways that they have been used to support public health policy. To do so, we consider a setting in which forecasts are used to help determine the allocation of a limited quantity of medical supplies across multiple regions. In section 2 of the paper, we define a new forecast scoring rule — the *allocation score* — that evaluates forecasts based on how beneficial resource allocations derived from them would turn out to be. In section 3, we present an illustrative analysis using the allocation score to evaluate forecasts of hospital admissions in the US that were made leading up to and during the Omicron wave that peaked in January of 2022. This analysis is "synthetic" insofar as it is not intended to correspond to any specific historical record of allocation decisions that could

have been supported by hospitalization forecasts during this period. However, we view the general allocation problem on which our framework is based as a versatile template for formalizing real-world decisions that must constantly be made in real-time by public health administrators around the globe, especially those related to hospital capacity, ventilator usage, doses needed for specific medications and other situations where an outbreak creates sudden and highly variable demand for potentially scarce resources.

2 The Allocation Score

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

2.1 Overview of Allocation Scoring

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

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

where the unmet need in a particular location is the difference between resource need in that location and the number of resource units 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 value 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 regard one allocation as better than another (with respect to the realized need) if it results in lower total unmet need, and accordingly regard one forecast as better than another (again with respect to the realized need) if the allocation derived from the former results in lower total unmet need than does the allocation derived from the latter.

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. The best and lowest possible allocation score is 0. A positive allocation score indicates that an oracle with perfect foreknowledge could have selected an allocation that met more need than the allocation suggested by F.

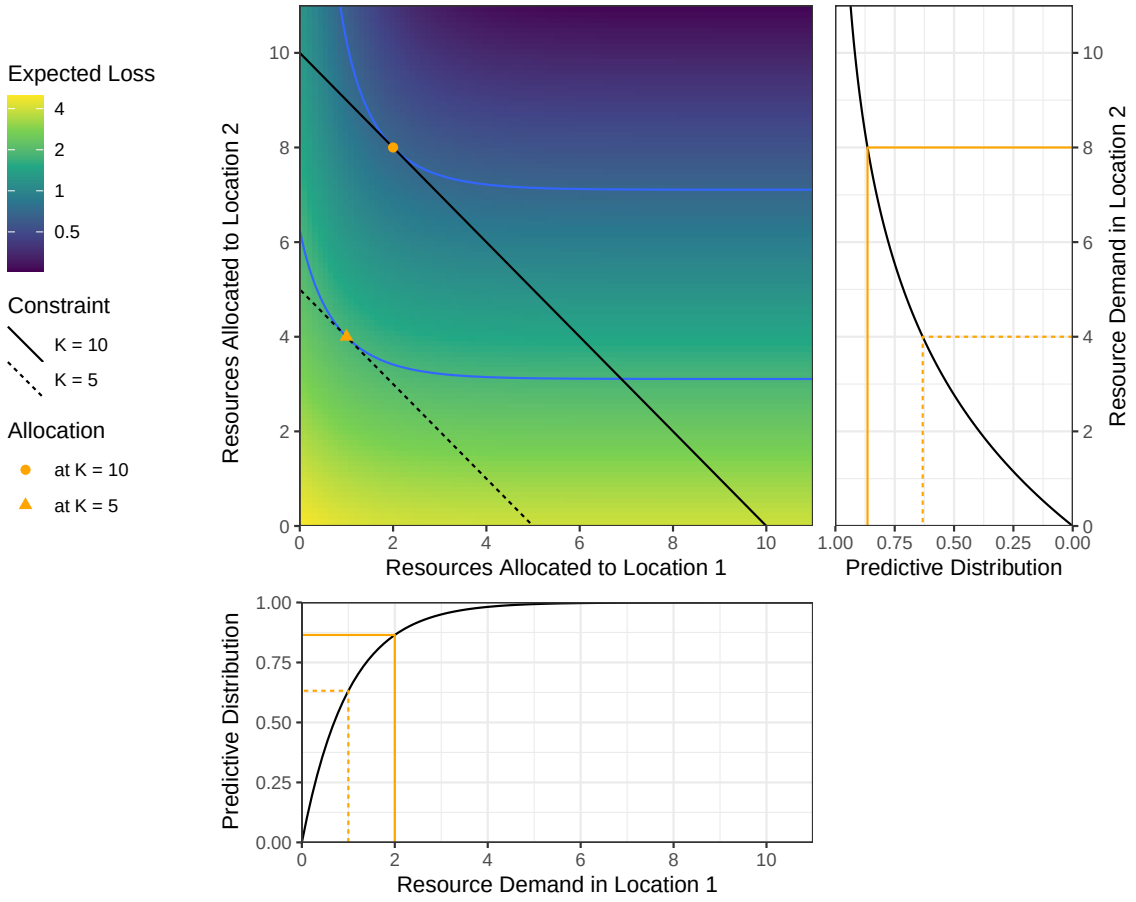

Figure 1: An illustration of the resource allocation problem in Example 1. There are $N = 2$ locations, with predictive distributions $F_1 = \text{Exp}(1)$ and $F_2 = \text{Exp}(1/4)$. The cumulative distribution functions of these distributions are illustrated in the panels at bottom and right. In the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

Example 1 Suppose we have a forecast F for need in two locations such that F_1 and F_2 have exponential marginal distribution with scale parameters $\sigma_1 = 1$ and $\sigma_2 = 4$. The means of these distributions are given by the scale parameters σ_i. When the marginal forecasts are exponential distributions, it can be shown that the optimal allocation divides the available resources among the locations proportionally to the scale parameters σ_i (see appendix B). If $K = 5$ units of the resource are available, the optimal allocation according to F would be 1 and 4 resource units in locations 1 and 2, respectively. Increasing K to 10 changes this optimal allocation to 2 and 8. 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 evaluate 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 $K = 5$ available resource units, 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. But these 6 units of unmet need are unavoidable under the constraint $K = 5$: allocating 0 units of resources to location 1 and 5 to location 2, for example, still results in a total unmet need of 6 across both locations. This gives an allocation score of 0 for the forecast when $K = 5$. On the other hand, when $K = 10$, the

forecast's allocation results in $10 - 8 = 2$ units of unmet need in location 2 despite leaving no need unmet in location 1. In this case, the oracle would be able to prevent all but 1 of the total 11 units of need from going unmet, for example by allocating 1 unit of resources to location 1 and the remaining 9 units of resources to location 2. So for $K = 10$, the forecast gets an allocaton score of 1 ($= 2$ realized $- 1$ unavoidable) in units of avoidable unmet need.

These scores illustrate a general result: allocation scores for a forecast will tend to be larger when the resource constraint is close to the observed need, because this is when it matters most which locations are allocated more or less resources. If the amount of available resources is small relative to the observed need, any allocation of those limited resources will result in a large amount of unmet need. If the amount of available resources is comparatively large, it becomes less important which locations receive relatively more or fewer resources because all locations are likely to receive enough resources to meet their need. In either of these extremes of resource availability, the avoidable unmet need that arises from the allocation suggested by a forecast (i.e., the forecast's allocation score) will tend to be small.

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

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

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

2.2 A decision theoretic development of the allocation score

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

2.2.1 The decision theoretic setup for forecast evaluation

Decision theory investigates how decisions are made by formalizing a decision as the process of selecting an *action* x from a specified set $\mathcal{X}$ of potential actions while taking into consideration the possible consequences of x under future states of the world. Let Y be a quantifiable aspect of the future which will help to determine the consequences of x. Y is a random variable inasmuch as it is indeterminate when the decision is made. We refer to a realization y of Y as an *outcome*. A fundamental strategy in decision theory is to assume that the consequences of an action x under outcome y can be assigned a numeric value, or *loss*, measuring the (lack of) success of x.

Some examples of actions are levels of investment in measures designed to mitigate severe disease outcomes such as hospital beds, ventilators, medication, or medical staff, with $\mathcal{X}$ being the set of all feasible levels of investment. An outcome y determining the consequences of such an action might be

the number of individuals who eventually become sick and would benefit from the mitigation measure. In our example, x will succeed to the extent that it meets the realized need y, so that greater unmet need will incur greater loss.

When a forecast F of Y is available, a decision maker might choose that action x which yields the lowest expected loss according to F. The loss of the action x chosen according to F can then, in turn, be interpreted as the value of F as an input to the decision making process under the realized outcome y of Y.

The strategy of minimizing the expected loss according to F might be used if the decision maker trusted that F was accurate and did not have any additional information that was not captured in F. However, here we use it only as a device for evaluating forecasts by examining the consequences of hypothetical decisions that would be made when using only F as an input. We return to this point in the discussion.

We arrive at a three-step procedure for developing scoring rules for probabilistic forecasts:

1. Specify a *loss function* $s(x, y)$ that measures the loss associated with taking action x when outcome y eventually occurs.
2. Given a probabilistic forecast F, determine the *Bayes act* x^F that minimizes the expected loss under the distribution F.
3. The *scoring rule* for F calculates the score as the loss incurred when the Bayes act was used: $S(F, y) = s(x^F, y)$.

This is a general procedure that may be applied in settings where it is possible to specify a quantitative loss function. We call a scoring rule obtained from this procedure a *Bayes scoring rule* (noting that to our knowledge, this terminology is not standard). In appendix A we review the result from the scoring rule literature (e.g., Dawid (2007); Gneiting and Raftery (2007)) that Bayes scoring rules are proper by construction.

2.2.2 The allocation score for a fixed resource constraint

In our setting an action is a vector $x = (x_1, \ldots, x_N)$ specifying the amount of resources allocated to each of N locations. We require 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 $y = (y_1, \ldots, y_N)$; this is a realization (unknown at the time of decision making) of the random vector $Y = (Y_1, \ldots, Y_N)$. A forecast of the random need Y for each location is written as $F = (F_1, \ldots, F_N)$. We assume need is non-negative and that the forecasts reflect that, i.e. the support of each F_i is a subset of $\mathbb{R}^+$. Finally, we assume that each unmet unit of need in any location incurs a fixed loss of $L > 0$, 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 decision is made about a public health resource allocation and the time t_r when resource needs that might be addressed by that allocation occur. Our setup assumes that $t_d < t_r$ and that the resource in question can only meet the need realized at t_r, not prevent it. That is, we stipulate that the choice of allocation x has no effect on the distribution of the random vector Y. In the context of infectious disease, this means that we do not consider resources, such as vaccines, that are intended to reduce the number of people who will become sick at some point in the future. Instead, our setup addresses resources like hospital beds, oxygen supply, or ventilators which are intended to meet the medical needs of patients who are already sick. Additionally, our setup addresses decision-making that is related to resource needs only at the time t_r; we do not 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 procedure in section 2.2.1.

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

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

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

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

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

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

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

The oracle adjustment aligns with a common theme in economic decision theory that opportunity loss (often known as regret or (negative) relative utility) is often a more important quantity than absolute

loss (see e.g., Diecidue and Somasundaram (2017)).

2.2.3 Integrating the allocation score across resource constraint levels

The allocation score S_A developed in the previous section evaluates 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:

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

$$S_{IAS}(F, y) = \int S_A(F, y; K)p(K)\, dK \tag{8}$$

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).

In practice, it may be challenging to compute the integral in (8). In the applied investigations below we resort to a discretization, evaluating at a grid of values of the resource constraint K. A discrete treatment along these lines will often be natural, in settings where the resource in question is non-divisible (e.g., ventilators, masks).

2.3 ~~Connections to~~ Comparisons with Other Scores

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

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

We also note that the AS, just as the (aggregate or mean) quantile score and WIS, depends only on the marginal distributions for each location of a multi-location forecast (see equations (5) and (18) in Section B of the appendix). It is however the case that when a forecaster modifies a single marginal forecast (holding those for other locations fixed), the aggregate WIS of their forecast will change only according to the WIS of that marginal forecast, whereas every term in the sum defining the AS (in

(5)) will change as the Bayes probability level defined by (18) changes. In this sense, the AS depends "jointly" on the marginal forecasts while the WIS does not.

3 Evaluating forecasts of COVID hospitalizations using the allocation score

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

3.1 Data

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

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

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

The hospitalization data used for scoring forecasts were downloaded on July 26, 2024.

3.2 Evaluation metrics

We evaluated forecasts using the allocation score (AS) and the weighted interval score (WIS), computed at a horizon of 14 days. We chose to focus much of our analysis on the AS computed for a resource constraint of $K = 15,000$. This level provides an anchor for interesting comparisons between phases of the Omicron wave by representing a national shortage at the peak and a national excess before and after the wave (see Figure 4A), assuming that the resource need corresponds directly to new hospital admissions. We additionally computed the mean WIS (MWIS) across all of the forecasted state-level locations.

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

As described above, predictions were submitted to the Forecast Hub in the form of a set of 23 quantiles of the predictive distribution. The WIS can be directly calculated from these quantiles. However, our numerical method for calculating the AS (outlined in appendix C) requires a full cumulative distribution function (CDF) for the forecast in each location. For the purpose of this analysis, we imputed CDFs based on the provided quantiles following a procedure detailed in appendix D. Briefly, this involves interpolating the provided quantiles between the lowest (.01) and highest (.99) probability levels using monotonic cubic splines, and then extending outside these levels with normal distribution tails parametrized to match the two lowest and two highest quantiles. We show in appendix E that

if this evaluation procedure had been specified prospectively, the resulting score would be proper, but that a post hoc application of this procedure is improper. We use the procedure here to illustrate the properties of the AS, and note that a collaborative forecasting hub interested in using the AS for evaluation could circumvent propriety issues by communicating the definition of the AS to forecasters and then collecting allocations at specified resource levels as part of forecast submissions.

3.3 Allocation benchmarks

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

3.4 Data and code availability

All forecast data used in this evaluation are available through the COVID-19 Forecast Hub (Cramer et al., 2022a). An R package implementing the allocation score is available at https://github.com/aaronger/alloscore. All code and data for the analyses presented in this manuscript are available at https://github.com/aaronger/utility-eval-papers. The analyses were generated using a reproducible workflow using R version 4.4.1 (2024-06-14) and the `targets` package (R Core Team, 2023; Landau, 2021).

3.5 Application results

3.5.1 Anatomy of forecast scores for one week

To illustrate the mechanics of allocation scoring, we start by focusing on how forecasts generated on or before December 20, 2021, with predictions for January 03, 2022, were scored by different metrics. This week was around the peak of the Omicron wave nationally, with individual states typically observing a peak at or after January 3, 2022.

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

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

Table 1: For one illustrative week, a comparison of allocation scores (AS), mean weighted interval scores (MWIS), and two varieties of Integrated Allocation Scores (IAS), (see Section 3.5.5). 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.

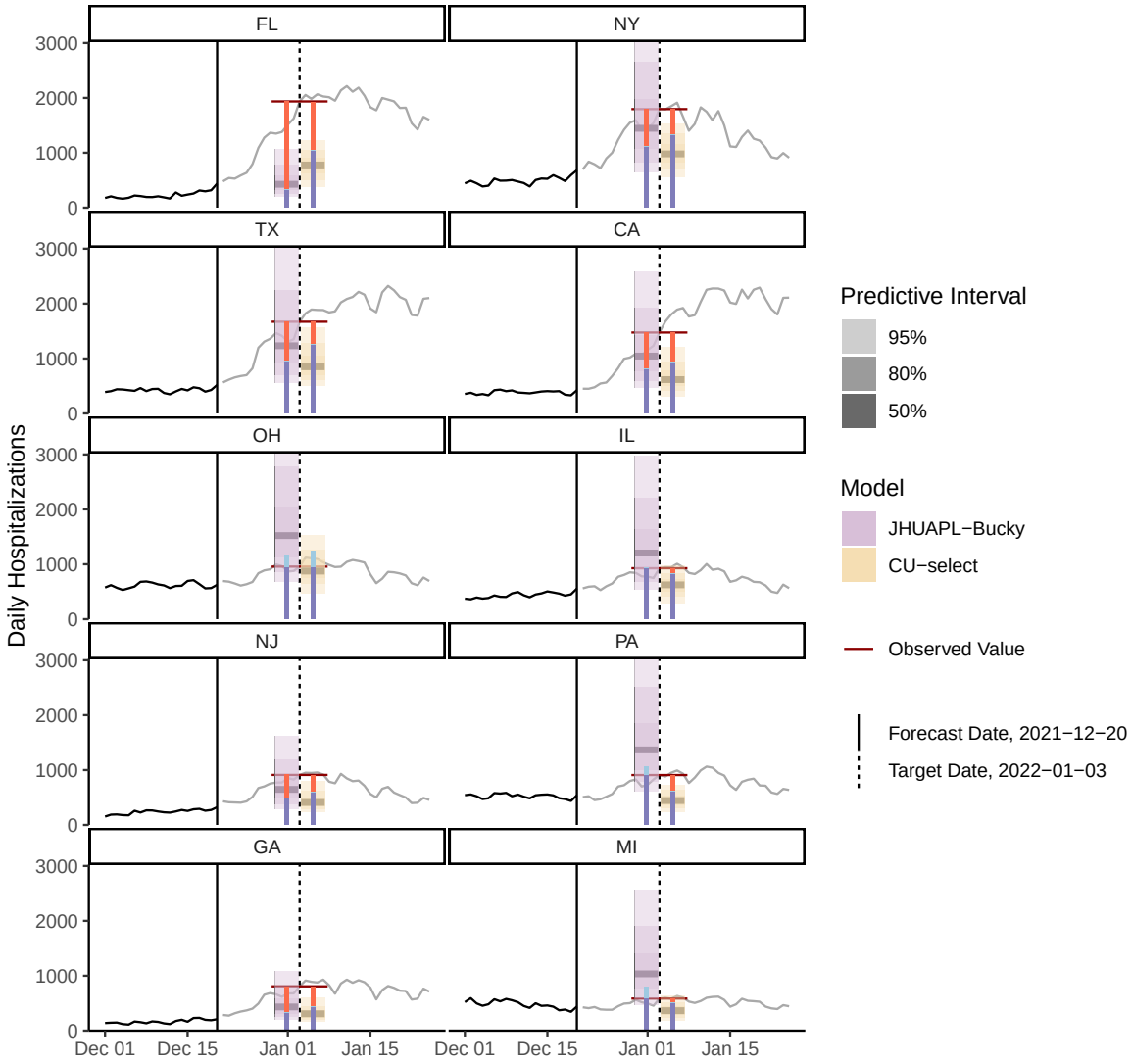

Figure 2: Probabilistic forecasts for new COVID-19 hospital admissions on January 3, 2022 and their suggested resource allocations for the states with the ten highest (eventually observed) hospitalization counts. For each state, the dark black line shows the data observed when the forecast was made,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

A state-wise comparison of the `JHUAPL-Bucky` and `CU-select` models shows that while the `JHUAPL-Bucky` forecast distributions were more centered on the eventual observations in many states, the suggested allocations of the `JHUAPL-Bucky` model were less efficient than those of the `CU-select` model (Figures 2 and 3A). Thus `JHUAPL-Bucky` achieved a worse allocation score than `CU-select` (on this particular week) by allocating excess resources to several states, such as Pennsylvania and Michigan, rather than to states such as Florida and California where these resources would have reduced unmet need. These allocation errors resulted from forecasts failing to consistently capture the relative resource needs across

different states. The CU-select model made some similar errors — most prominently, over-allocating resources to Ohio — but overall, it more successfully forecasted the resource demands across different locations in relative terms.

On the other hand, CU-select had worse performance as measured by MWIS. Its forecasts were biased downwards, and it consistently incurred a large penalty for underprediction (Figure 3B). The JHUAPL-Bucky model had wider predictive intervals which more often included the observed level of daily hospital admissions. Its MWIS was therefore less severely penalized by under- or over-predictions of the actual hospitalization counts.

3.5.2 Forecast scores showed differences in aggregate and over time

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

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

Model rankings on mean allocation score (MAS) and time-averaged mean weighted interval score (taMWIS) were broadly similar, with some places of disagreement. The four most accurate and the five least accurate models were the same according to both metrics, although not in exactly the same order (Table 2). (This comparison excludes the per-capita allocation which does not use forecasts and therefore cannot be assigned a MWIS.) Notably, the JHUAPL-SLPHospEns model ranked best on taMWIS but had only the fourth best MAS.

model	MAS	taMWIS	taMWIS rank
COVIDhub-ensemble	389	70	2
per-capita	464	-	-
COVIDhub-trained_ensemble	483	87	4
CU-select	502	81	3
JHUAPL-SLPHospEns	526	67	1
COVIDhub-baseline	594	93	5
JHUAPL-Bucky	643	112	7
MUNI-ARIMA	707	116	8
JHUAPL-Gecko	929	110	6
USC-SI_kJalpha	1473	155	9

Table 2: Time averaged mean weighted interval scores (taMWIS) and mean allocation scores (MAS) by model across 13 weeks. These results show the average performance across time for the nine models that submitted forecasts for every week from 2021-11-29 through 2022-02-21, as well as for a per-capita allocation. MWIS is not defined for the per-capita allocation which does not use forecasts.

3.5.3 Metrics were not consistently correlated over time

We can gain some insight into the discordance between MAS and taMWIS by examining the correlation between time-specific AS and MWIS values of the various models in Figure 5. We find, for example, no clear association between the consistent and high MWIS rank of the JHUAPL-SLPHospEns model and its highly variable AS rank. This contrasts with a clearly positive correlation between MWIS and AS ranks in other models such as Karlen-pypm and USC-SI_kJalpha. We also can see in more detail the consistently high performance of the COVIDhub-ensemble on both scores. And while it did not submit forecasts for enough weeks to be included in Table 2, CMU-TimeSeries is an interesting example of a model that performed consistently well on AS but had only middling MWIS ranks.

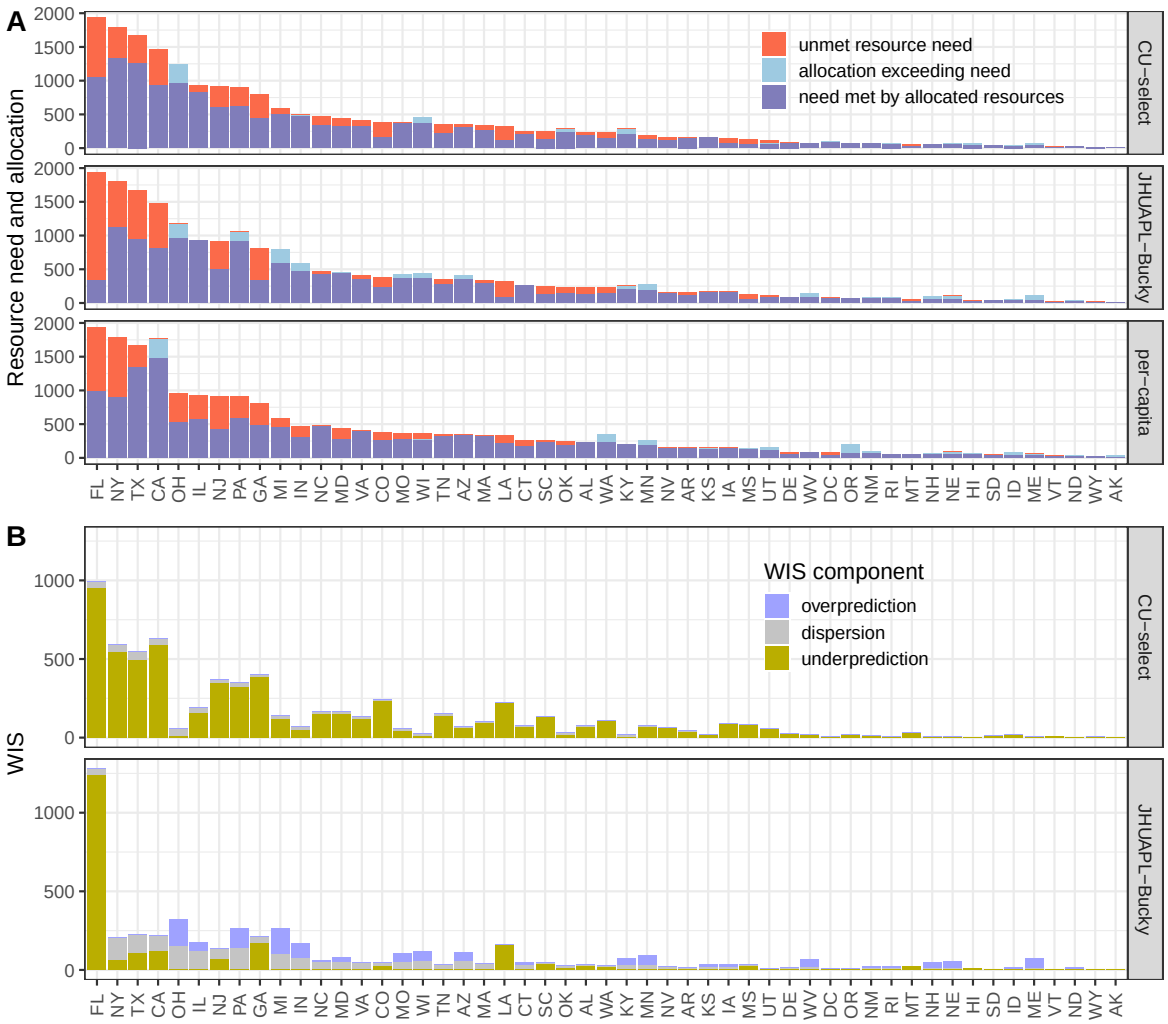

Figure 3: Component-wise breakdowns of the allocation score (Panel A) and weighted interval score (Panel B), by location for forecasts of hospitalization admissions on January 3, 2022, for two selected models (`JHUAPL-Bucky` and `CU-select`). Panel A shows the observed resource need, in this case the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

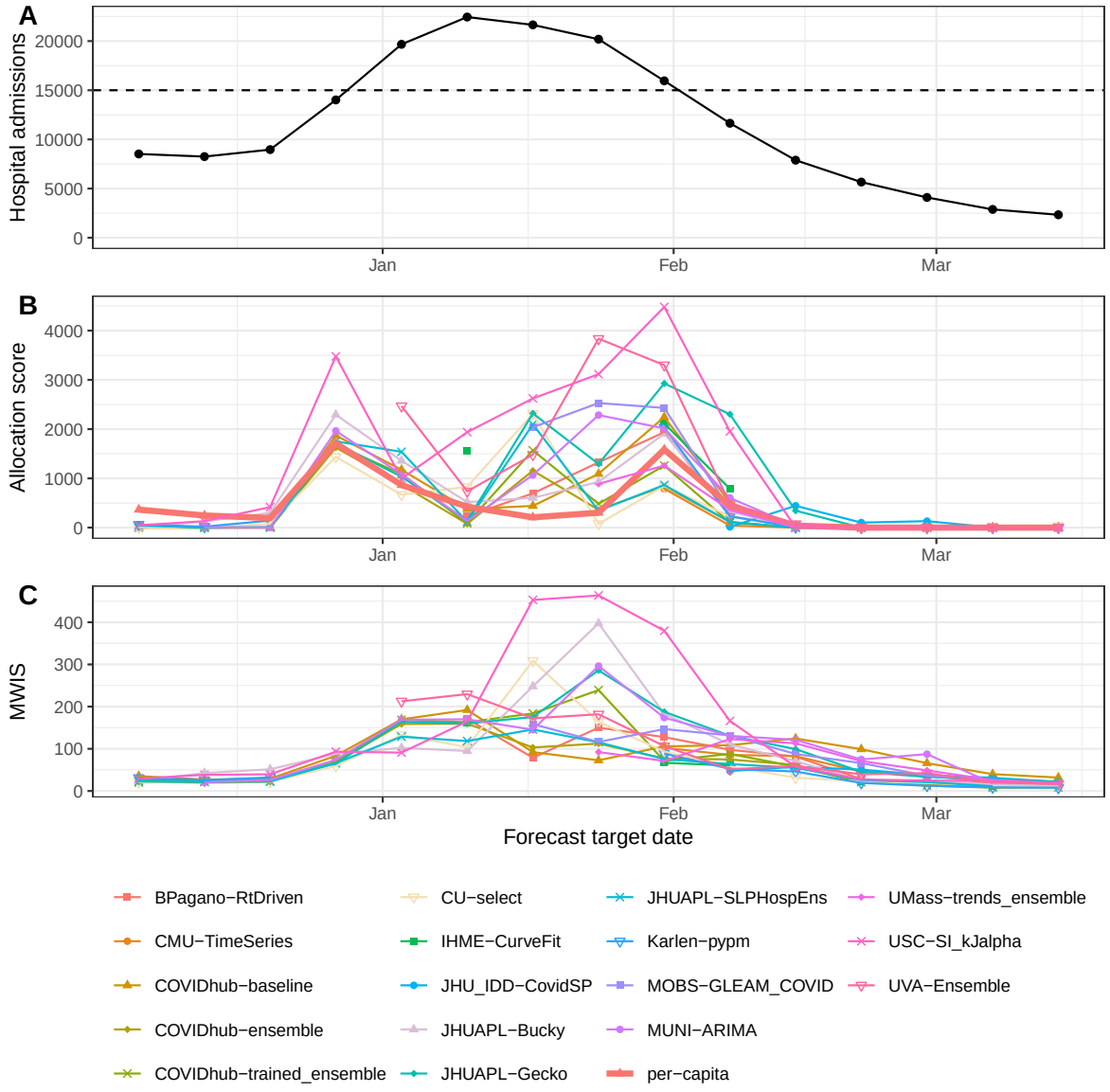

Figure 4: Hospital admissions and evaluation metrics over time. Panel A shows the number of hospital admissions in the US as a whole due to COVID-19 on a sequence of 15 Mondays from December 2021 through March 2022. These are the dates for which forecasts were made and evaluated. A horizontal dashed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

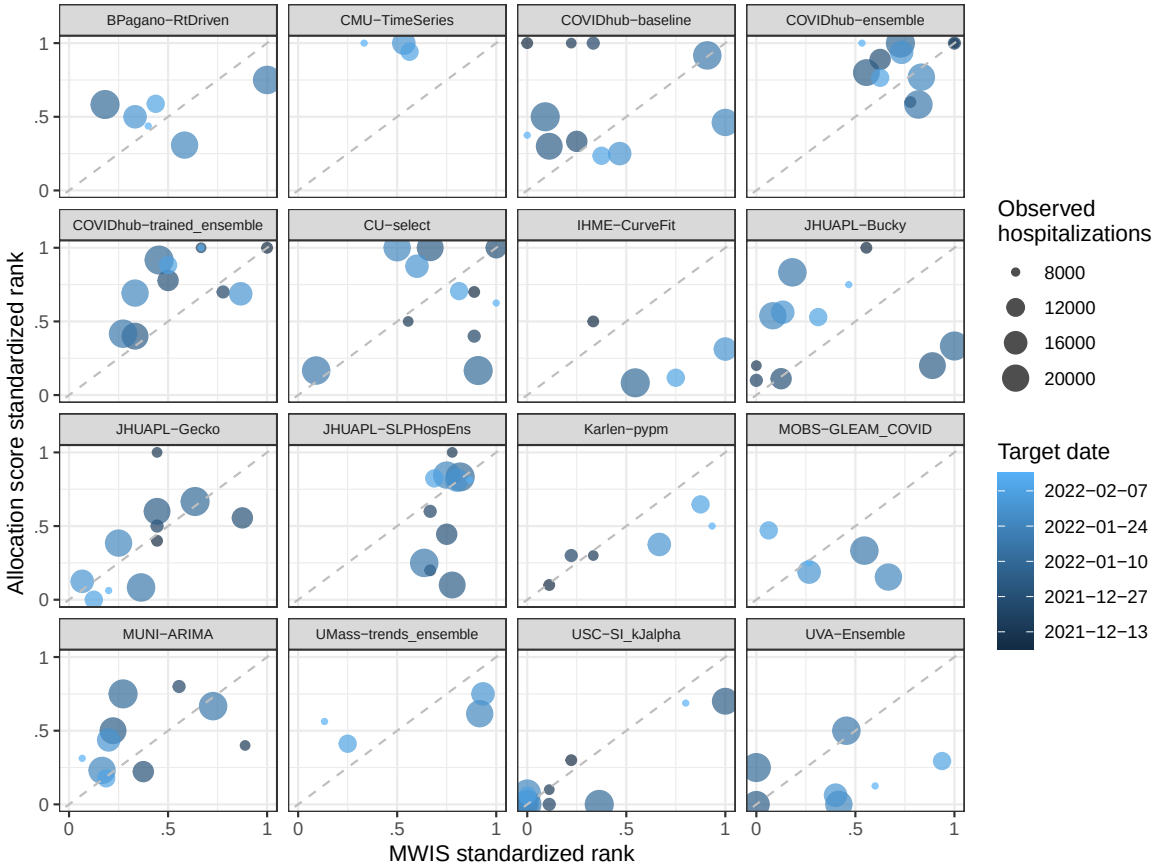

Figure 5: Association of standardized ranks for MWIS and allocation score by model and week. Each facet of the plot corresponds to one model. Within each facet, each point corresponds to a week. The x- and y-values correspond to the MWIS standardized rank and the allocation score standardized rank for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

3.5.4 ~~Integrated~~ Sensitivity of allocation score ~~across~~ to values of K

AS was computed for a range of K from 200 to 60,000 at increments of 200 for forecasts made on December 20, 2020 predicting levels of hospitalizations on January 3, 2021 as well as for the `per-capita` allocation (Figure 6A). These calculations highlight that AS was highest at at values of K near the observed nationwide total hospital admissions of 19,581 that day. Model ranks by AS were fairly stable across all K, and there appears to be a substantial interval around the observed value over which ranks are constant.

3.5.5 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), allowing one to assign higher weights to more likely values of K (Section 2.2.3). IAS was computed for two weight distributions on the same grid of values of K used in the sensitivity analysis above, one uniform across the entire range and the other with weights proportional to a truncated normal distribution centered at 15,000 (the orange and blue shaded areas in Figure 6A, respectively). With these weight distributions, the IAS defined in Equation (8) can be computed as the sum

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

Both versions of the IAS were correlated with the original AS for $K = 15,000$, with the higher correlation coming from the weighting centered at $K = 15,000$ (Figure 6B). Model rankings based on the AS and the centered IAS were roughly similar, with the top and bottom three approaches being the same for both scores (Table 1).

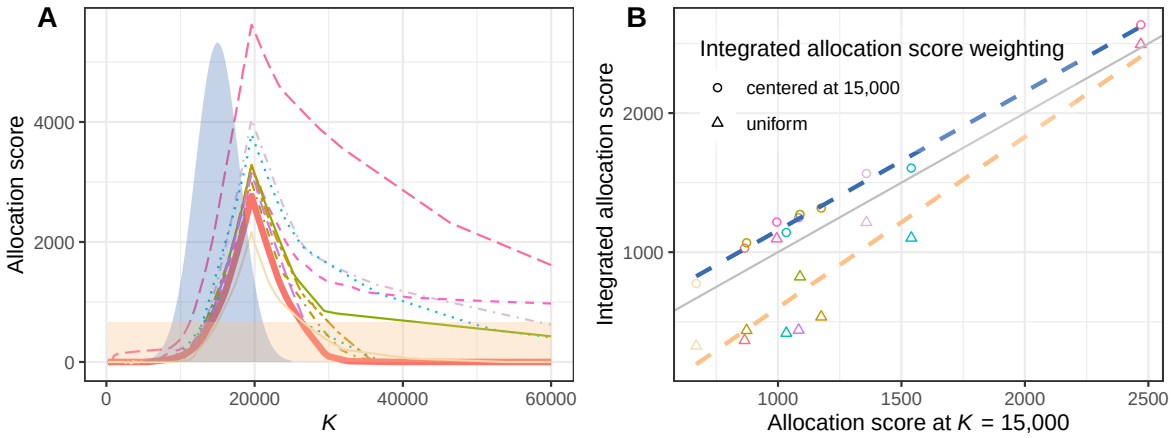

Figure 6: Allocation scores (AS) across different resource constraints (K) for 10 models that made forecasts on 2021-12-20. Panel A shows, for each model, the AS for values of K between 200 and 60,000 at increments of 200. The AS show a sharp peak just under 20,000, near the eventually observed number of hospitalizations. Two possible weighting functions for the Integrated Allocation Score (IAS) are shown. The first (blue shaded area) computes weights proportional to a normal distribution centered at 15,000 with a standard deviation of 3,000, and truncated to be between 5,000 and 25,000. The second (orange shaded area) 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 B shows how the two versions of the IAS (y-axis) correlate with the AS at $K = 15,000$ (x-axis). Every point represents the AS at 15,000 and a version of the IAS for one model. As we would expect, the IAS centered at 15,000 (circles) is more closely correlated with the AS at 15,000 than is the uniform IAS (triangles).

4 Discussion

Probabilistic forecasts of infectious disease outbreaks have typically been evaluated using well-known proper scoring rules such as the LS, CRPS, and WIS. Often models are ranked according to such a score, but without reference to any underlying decision problem from which the score might be derivable as a Bayes scoring rule or be otherwise motivated. We argue that the value of collaborative forecasting efforts, such as the COVID-19 Forecasting Hub, could be enhanced by including evaluations of forecasts with scoring rules that directly consider the success of forecasts in supporting specific ~~public health decisions. As evidence for this possibility, we~~ decisions. The allocation score was developed for this purpose. The AS is expressed in units that are directly relevant to decisions, which helps not only with ranking forecasts in the context of a specific action, but also allows decision-makers to contextualize the magnitude of differences between forecast scores. For example, it may be the case that differences in AS are small relative to the amount of available resources or the accuracy with which planned resource allocations can be executed. The benefits of the approach to forecast evaluation discussed here generalize beyond public health, to any application of probability forecasts for decision-making, such as economics, where the monetary consequences of using a forecast can be clearly juxtaposed with standard accuracy measures of the forecast (Bannigidadmath and Narayan, 2016; Zhang et al., 2019).

resource allocation in urban planning (Burnett, 2014; Liang and Wey, 2013), or sustainability initiatives such as water and energy allocation (Syme et al., 1999; Gebre et al., 2021; Colett et al., 2016).

We have demonstrated how tying forecast evaluation to a specific decision problem (e.g., via the allocation score) can yield model rankings that differ substantively from those based a current standard measure of forecast accuracy, the WIS, which does not refer to any decision problem that has been clearly connected to outbreak response. This result aligns with findings in other fields ~~, especially those where the monetary consequences of using a forecast can be clearly juxtaposed with standard accuracy measures of the forecast~~ (Leitch and Tanner, 1991; Murphy, 1993; Cenesizoglu and Timmermann, 2012). Additionally, we show that an existing ensemble forecast approach was the only method to outperform (in terms of this new evaluation method) two benchmark allocation approaches in a hypothetical application, suggesting that there is room for innovation of current epidemiological modeling techniques that might be spurred on by a reorientation of the field toward practical decision problems. For example, in order to surpass benchmark allocation scores, forecasters need to devote effort to capturing the *relative* magnitude of future resource need across different locations. Success at this task is not, however, directly rewarded by the WIS (see the discussion at the end of Section 2.3).

Our example application of the AS refers to a hypothetical and unspecified resource need that is assumed to pair with our forecasted outcome (new hospital admissions) in a manner that might appear overly simplistic and artificially direct. We are optimistic, however, that by augmenting this scheme with a probabilistic measurement model for resource need and more detailed domain knowledge, it could be meaningfully aligned with a more concrete decision problem such as the allocation of ventilators during a respiratory viral pandemic. This problem was explored from a stochastic optimization perspective (not from a forecast evaluation perspective) in Huang et al. (2017) and provided an initial motivation for this project. We see ~~it as~~ this as a promising focus for further development of the AS framework since it (a) has been identified as a setting in which allocation decisions are made during respiratory disease outbreaks, (b) involves logistical considerations depending on geographical and population scales, and (c) would likely imply a central role for hospitalization data. But again, additional modeling and data synthesis would be required since not everyone who is hospitalized needs a ventilator. Other possible real-world examples include the allocation of a limited stockpile of vaccinations (Araz et al., 2012; Persad et al., 2023) or diagnostic tests (Du et al., 2022; Pasco et al., 2023).

In practice, epidemiological forecasts are often directed at a diverse set of end-users. It may be easy for some of these users to summarize the consequences of how they use a forecast with a numerical loss function, but for others, such as officials deciding how best to update public "situational awareness" of an outbreak, this may be essentially impossible. Even those forecast uses that are easily framed as expected loss minimization may differ enough that no single score would be appropriate for all users. Ideally, targeted forecasting tools could be developed through close collaboration between modelers and public health officials. However, this may only be possible in settings with sufficient staffing on both an analytics and a public health team. Increasingly, collaborative modeling hubs are being used to generate "one-size-fits-all" forecasts for many locations at once. But it could still be beneficial in these settings to expand the set of scoring methods used by the hub to include scores, such as the AS, which are based on specific public health decision problems. This could help end users better understand the value of available forecasts as inputs to their particular decision making contexts. (We note that issues of propriety are raised when forecasts are evaluated with scoring rules for which they were not elicited or via parameters imputed by a hub after submission, both of which are done in our application example. We consider these issues briefly in appendices D and E.)

There are ~~several~~ three important limitations to the current work ~~. The allocation score we developed here~~ which we noted when introducing the AS in Section 2.2.2. We view these issues as promising avenues for future investigation.

First, the terms in our loss function (3) do not depend on geographic or demographic covariates. Policy makers may, however, wish to incorporate differences in the impact of unmet need across populations into a forecast evaluation methodology. For example, in the state-level forecasting context of Section 3, it might be appropriate to adjust the loss so that the costs of unmet need are larger in locations with greater population vulnerability. Along these lines, it should also be mentioned that the AS does not directly ~~account for important considerations such as~~ take into consideration the fairness or

equity of allocations, or more broadly, individual and group preference relations that are difficult or even impossible to encode into a loss or utility function. Ambiguity aversion (in the sense of Ellsberg (1961)), for example, seems especially relevant to the use of forecasts in outbreak response, where there can be immense social and political pressure on public health officials to maintain a transparent base of evidence for their policy choices and recommendations. ~~The proposed framework 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~~

Secondly, our formulation of the AS does not take into consideration the temporal inter-dependencies between sequential decisions that public health decision makers are likely to face as an epidemic unfolds. Such dependencies could arise from time-varying inputs such as resource availability or the current effective allocations of resources. The sequential decision problem could also be combined with others allowing constraints for different resources to be modified by time-varying funding shifts between various disease mitigation measures. ~~Finally,~~ Incorporating these possibilities into an AS framework would likely require technical developments involving either dynamic programming (see e.g., Bertsekas (2012)) or a scenario-based approach (see e.g., Pflug and Pichler (2014)).

Thirdly, we emphasize that we opted at this stage of development to explicitly rule out situations where a successful epidemiological forecast could lead to policy decisions that change the distribution of the predicted outcome Y. ~~Our framework would need considerable enhancements before being applicable~~ For example, this excludes the important problem of allocation of doses of a preventative vaccine. We have not developed a formal approach to forecast evaluation ~~beyond horizons for which causal feedback can be neglected~~using scores given by (7) when both disease burden and unmet need are considered as endogenous variables. Such an approach would require some form of accounting for the causal effect of the allocation decision on observed measures of disease burden, e.g. using the available observed data to estimate the (counterfactual) number of cases or deaths averted from one or another allocation strategy.

~~An~~ Another opportunity for further investigation is to more carefully evaluate whether forecasts add value to standard procedures already in place for making public health decisions. In the context of allocations, such a procedure might extrapolate need from current observed need and population levels (similarly to the two benchmark approaches presented above), with adjustments based on other political or real-world considerations. For example, in many settings public health stakeholders will make decisions after synthesizing information from a variety of quantitative and qualitative sources coupled with expert judgment. The allocation score presented in this work does not directly measure whether a given forecast adds useful information to such an existing decision-making process. While the scoring procedures as presented do not directly address this question, they could be modified (say, by comparison to a baseline model or expert-elicited allocations in the absence of forecast data) to quantify the benefit of using a forecast to inform a specific decision.

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

Acknowledgements

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

This work has been supported by the National Institutes of General Medical Sciences (R35GM119582)

and the U.S. CDC(1U01IP001122). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIGMS, the National Institutes of Health, or CDC.

Appendices

We address some technical and methodological points from the main text. We begin in section A by defining proper scoring rules and showing that the allocation score is proper. Section B gives a justification for the result that the Bayes act for the allocation problem is given by a vector of quantiles of the forecast distributions in each location at a shared probability level, which was stated in section 2.2.2. We describe the algorithm that we use to compute allocations given a forecast distribution in each location in section C. In section D, we describe the methods for approximating a full distribution from a set of quantiles, implemented in the R package `distfromq`, that we used to support computation of allocation scores from quantile forecasts that were submitted to the US COVID-19 Forecast Hub for the application in section 3. Finally, in section E we examine implications for propriety of an analysis that uses summaries of forecasts (such as predictive quantiles) rather than the full forecast distributions for computation of allocation scores.

A Proper scoring rules

In decision theory, a loss function l is used to formalize a decision problem by assigning numerical value $l(x, y)$ to the *result* of taking an *action* x in preparation for an *outcome* y. A *scoring rule* S is a loss function for a decision problem where the action is a probabilistic forecast F of the outcome y (or the statement of F by a forecaster). We refer to the realized loss $S(F, y)$ as the *score* of F at y.

Probabilistic forecasts can be seen as a unique kind of action in that they can be used to generate their own (simulated) outcome data, against which they can be scored using S. A probabilistic forecast F is thus committed to the "self-assessment" $\mathbb{E}_F[S(F, Y)] := \mathbb{E}[S(F, Y^F)]$, where $Y^F \sim F$ is the random variable defined by sampling from F, as well to an assessment $\mathbb{E}_F[S(G, Y)]$ of any alternative forecast G.

A natural consistency criterion for S is that, for observations assumed to be drawn from F, it will not assess any other forecast G as being better than F itself, that is, that

$$\mathbb{E}_F[S(F, Y)] \leq \mathbb{E}_F[S(G, Y)] \tag{9}$$

for any F, G. A scoring rule meeting this criterion is called *proper*. If S were improper, then from the perspective of a forecaster focussed (solely) on expected loss minimization, the decision to state a forecast G other than the forecast F which they believe describes Y could be superior to the decision to state F. S is *strictly proper* when (9) is a strict inequality, in which case the *only* optimal decision for a forecaster seeking to minimize their expected loss is to state the forecast they believe to be true.

A.1 The allocation score is proper

Our primary decision theoretical procedure, outlined in section 2.2.1, uses a decision problem with loss function $s(x, y)$ to define a scoring rule

$$S(F, y) := s(x^F, y) \tag{10}$$

where $x^F := \operatorname{argmin}_x \mathbb{E}_F[s(x,Y)]$ is the Bayes act for F with respect to s. Such scoring rules, which we call *Bayes scoring rules*, are proper by construction since

$$\begin{aligned}
\mathbb{E}_F[S(F,Y)] &= \mathbb{E}_F[s(x^F,Y)] \\
&= \min_x \mathbb{E}_F[s(x,Y)] \quad \text{(by definition of } x^F\text{)} && (11)\\
&\leq \mathbb{E}_F[s(x^G,Y)] && (12)\\
&= \mathbb{E}_F[S(G,Y)].
\end{aligned}$$

The allocation scoring rule is Bayes and therefore proper.

We note that in the probabilistic forecasting literature (see e.g., Gneiting (2011a), Theorem 3) what we have termed Bayes scoring rules typically appear via (10) where x^F is some given functional of F which can be shown to be *elicitable*, that is, to be the Bayes act for some loss function s. Such a loss function is said to be a *consistent loss (or scoring) function* for the functional $F \mapsto x^F$, and many important recent results in the literature (e.g., Fissler and Ziegel (2016)) address whether there *exists* any loss function that is consistent for x^F. Our orientation is different from this insofar as we *begin* by specifying a decision problem and a loss function of subject matter relevance and use the Bayes act only as a bridge to a proper scoring rule. Consistency is never in doubt.

B Allocation Bayes acts as vectors of marginal quantiles.

Here we study the form of the Bayes act for the allocation problem (AP) (equation (4) in section 2.2.2) of the text:

$$\underset{0\leq x}{\text{minimize}}\ \mathbb{E}_F[s_A(x,Y)] = \sum_{i=1}^N L\cdot\mathbb{E}_{F_i}[\max(0,Y_i-x_i)] \text{ subject to } \sum_{i=1}^N x_i = K, \tag{13}$$

where the marginal forecasts F_i for $i=1,\ldots,N$ represent forecasts for N distinct locations. We show that the Bayes act $x^{F,K} = (x_1^{F,K},\ldots,x_N^{F,K})$ for a forecast F and resource constraint level K is a vector of quantiles of the marginal forecast distributions F_i at a single probability level $\tau^{F,K}$, that is, $x_i^{F,K} = q_{F_i,\tau^{F,K}}$. An immediate consequence used in the examples in Section 2.1 is that if $F_i = \text{Exp}(1/\sigma_i)$ for all i, then the Bayes act is proportional to $(\sigma_1,\ldots,\sigma_N)$, since ~~$q_{\text{Exp}(1/\sigma),\tau} = -\sigma\log(1-\tau).$~~

$$q_{\text{Exp}(1/\sigma),\tau} = -\sigma\log(1-\tau). \tag{14}$$

We begin by noting that a key feature of each term of the loss function $s_A(x,Y)$ defining the AP is the presence of a *shortage*: an amount $\max\{0,y-x\}$ by which a resource demand y exceeds a supply decision variable x, which, for convenience, we write as $(y-x)_+$. This is a feature shared with decision problems used to define quantiles and related scoring rules such as the CRPS and the WIS (see e.g., Gneiting (2011b), Jose and Winkler (2009), and Royset and Wets (2022), sections 1.C and 3.C). In particular, a quantile at probability level α of a distribution F on $\mathbb{R}$ (which we assume to have a well-defined density $f(x)$) is a Bayes act for the loss function

$$l(x,y) = Cx + L(y-x)_+$$

where $\alpha = 1-C/L$ and C and L can be interpreted as the cost per unit of a resource (such as medicine) and the loss incurred when a unit of demand (such as illness) cannot be met due to the shortage $(y-x)_+$. This follows because a Bayes act, as a minimizer of $\mathbb{E}_F[l(x,Y)]$, must also be a

vanishing point of the derivative

$$\begin{aligned}\frac{d}{dx}\,\mathbb{E}_F\left[l(x,Y)\right] &= \mathbb{E}_F\left[\frac{d}{dx}l(x,Y)\right]\\ &= C + L\,\mathbb{E}_F\left[\frac{d}{dx}(Y-x)_+\right]\\ &= C - L\,\mathbb{E}_F\left[\mathbf{1}\{Y>x\}\right]\\ &= C + L(F(x)-1),\end{aligned} \tag{15}$$

so that $1 - C/L = F(x)$. The formula $\frac{d}{dx}\,\mathbb{E}_F\left[(Y-x)_+\right] = F(x) - 1$ for the derivative of the shortage, used above in (15), can be obtained from an application of the "Leibniz Rule":

$$\begin{aligned}\frac{d}{dx}\,\mathbb{E}_F[(Y-x)_+] &= \frac{d}{dx}\int_x^\infty (y-x)f_Y(y)dy\\ &= \int_x^\infty \frac{d}{dx}(y-x)f_Y(y)dy - (x-x)f_Y(x) = -\int_x^\infty f_Y(y)dy = F(x)-1.\end{aligned} \tag{16}$$

Note that more care is required when F does not have a density.

Returning to the AP (13), notice that in order for $x^\star \in \mathbb{R}_+^N$ to be a Bayes act it must be true that reallocating $\delta > 0$ units of the resource from location i to location j will lead to a net increase in expected shortage — in other words, the reallocation increases the expected shortage in location i at least as much as it decreases the expected shortage in location j:

$$\begin{aligned}&\mathbb{E}_{F_i}[(Y_i - (x_i^\star - \delta))_+] - \mathbb{E}_{F_i}[(Y_i - x_i^\star)_+] \text{ (detriment in } i)\\ &\quad\geq \mathbb{E}_{F_j}[(Y_j - x_j^\star)_+] - \mathbb{E}_{F_j}[(Y_j - (x_j^\star + \delta))_+] \text{ (benefit in } j)\ .\end{aligned}$$

Dividing by δ and letting $\delta \searrow 0$, this implies from (16) that

$$\begin{aligned}1 - F_i(x_i^\star) &= -\frac{d}{dx_i}\mathbb{E}_{F_i}[(Y_i - x_i^\star)_+] \text{ (rate of detriment in } i)\\ &= \lim_{\delta\searrow 0}\frac{1}{\delta}\left\{\mathbb{E}_{F_i}[(Y_i - (x_i^\star - \delta))_+] - \mathbb{E}_{F_i}[(Y_i - x_i^\star)_+]\right\}\\ &\geq \lim_{\delta\searrow 0}\frac{1}{\delta}\left\{\mathbb{E}_{F_j}[(Y_j - x_j^\star)_+] - \mathbb{E}_{F_j}[(Y_j - (x_j^\star + \delta))_+]\right\}\\ &= -\frac{d}{dx_j}\mathbb{E}_{F_j}[(Y_j - x_j^\star)_+] = 1 - F_j(x_j^\star) \text{ (rate of benefit in } j).\end{aligned} \tag{17}$$

(Negative derivatives appear here because our optimality condition addresses how a *decrease* in resources will *increase* the expected shortage in i and vice versa in j.)

Since (17) also holds with i and j reversed, a number λ (a *Lagrange multiplier*) exists such that $L(1 - F_k(x_k^\star)) = \lambda$ for all $k \in 1, \ldots, N$. (We scale by L to facilitate possible future interpretations of λ in terms of the partial derivatives of $\mathbb{E}_F[s_A(x,Y)]$.) That is, $x_k^\star$ is a quantile q_{τ,F_k} for $\tau = 1 - \lambda/L$. The value of τ is then determined by the constraint equation

$$\sum_{i=1}^N q_{\tau,F_i} = K. \tag{18}$$

It is important to note that τ depends on F and K and is *not* a fixed parameter of the allocation scoring rule.

C Numerical computation of allocation Bayes acts

To compute an allocation score $S_A(F, y; K) := s_A(x^{F,K}, y)$, we require numerical values for a Bayes act solving the AP (13) — that is, we must find the specific resource allocations for each location that

are determined by the forecast F under the resource constraint K. Assuming we have reliable means of calculating quantiles q_{α,F_i} of the marginal forecasts F_i, these allocations are given by $q_{\tau^\star,F_i}$ where $\tau^\star$ solves the equation (18). ~~However, this~~ This equation is not analytically tractable ~~and we must resort to a numerical method for finding an approximation $\tilde{\tau}$ of $\tau^\star$~~in general, but because the left-hand side of (18) is non-decreasing in τ, it is straightforward to find arbitrarily accurate approximations to a solution using an iterative bisection method.

We have implemented ~~an iterative bisection method that makes use of the fact that $\sum_{i=1}^N q_{\tau,F_i}$ is an increasing function of τ. The algorithm~~ such a method along with the resulting score computations in the R package `alloscore` (Gerding and Ray, 2023) which provided all allocation score values used in the analysis of section 3. The procedure begins with an initial search interval $[\tau_{L,1}, \tau_{U,1}]$ (such as $[0, \max_i F_i(K)]$) that clearly contains the solution $\tau^\star$. At each step j of the algorithm, we evaluate the total allocation $\sum_{i=1}^N q_{\tau_{M,j},F_i}$ at the midpoint of the search interval, $\tau_{M,j} = \frac{1}{2}(\tau_{L,j} + \tau_{U,j})$ and continue the search on the narrowed sub-interval

$$[\tau_{L,j+1}, \tau_{U,j+1}] = \begin{cases} [\tau_{L,j}, \tau_{M,j}] & \text{if } \sum_{i=1}^N q_{\tau_{M,j},F_i} \geq K \\ [\tau_{M,j}, \tau_{U,j}] & \text{if } \sum_{i=1}^N q_{\tau_{M,j},F_i} < K. \end{cases}$$

This search continues until $\tau_{U,j+1} < (1+\varepsilon)\tau_{L,j+1}$ for a suitably small $\varepsilon > 0$. ~~We have implemented this procedure along with the resulting score computations in the R package `alloscore` (Gerding and Ray, 2023) which provided all allocation score values used in the analysis of section 3.~~

Figure 7 illustrates this bisection method in the context of the examples of Section 2.1, where it quickly finds close approximations to the analytic solutions $\tau^\star = 1 - e^{-1}, 1 - e^{-2}$ (for $K = 5, 10$ respectively) which, in this case, are available (cf. (14)).

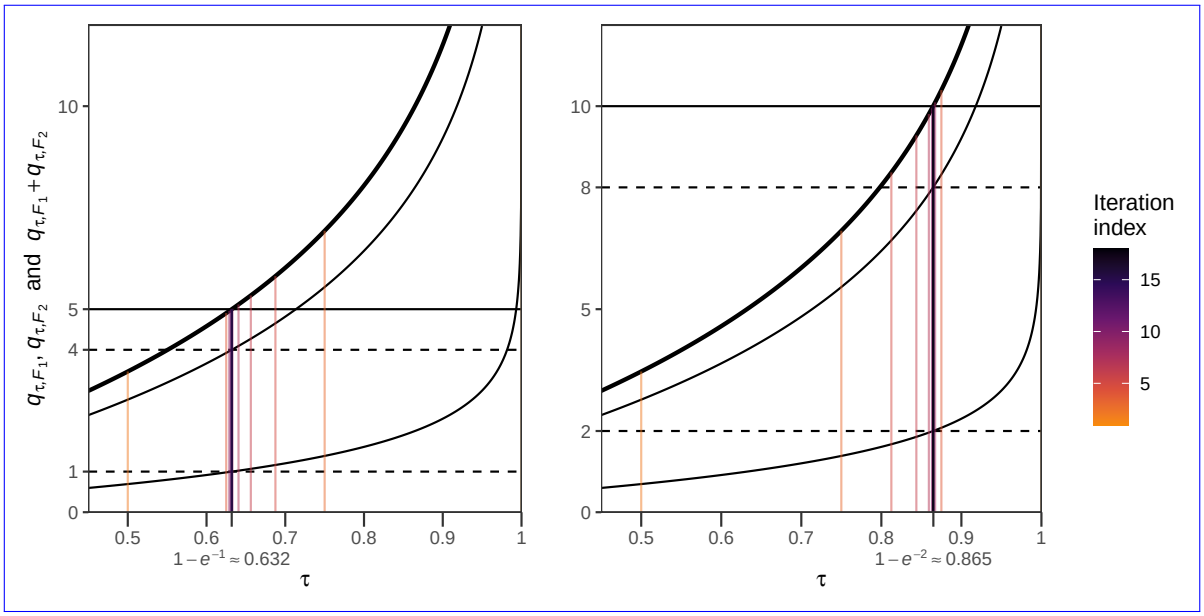

Figure 7: Illustrations of bisection method for the examples of Section 2.1. In each plot the thinner black curves are the quantile functions of the exponential forecasts for Location 1, with $\sigma_1 = 1$ (lower), and Location 2, with $\sigma_2 = 4$ (upper), while the thicker black curve gives the sum of these quantiles, which must meet the resource constraints of $K = 5$ on the left and $K = 10$ on the right. The red segments show the iterates $\tau_{M,j}$ and their associated total resource demands as given by the function `allocate` from the package `alloscore` applied to these examples and using an initial search interval $[0, 1]$. (The $\tau_{M,j}$ are in this case just sums of terms $\pm 2^{-j}$.) The horizontal dashed lines show the resulting allocations (cf. Figure 1).

Subtleties can arise when the forecast densities f_i vanish or are very small, in which case quantiles are non-unique or highly variable near a probability level, leading to ambiguity or numerical instabilities

in the evaluation of $\sum_{i=1}^{N} q_{\tau,F_i}$. Additionally, if point masses are present in any of the F_i, (18) will not have a unique solution for some discrete set of constraint levels K. We have adopted conventions for detecting such levels and enforcing consistency in score calculations near them. Through extensive experimentation, we have determined that these conditions seem to address these challenges with the forecasts we are working with, but we leave a more rigorous approximation error analysis for later work.

D Computing allocations from finite quantile forecast representations

In section 3, we used the allocation score to evaluate forecasts of COVID-19 hospitalizations that have been submitted to the US COVID-19 Forecast Hub. These forecasts are submitted to the Hub using a set of 23 quantiles of the forecast distribution at the 23 probability levels in the set $\mathcal{T} = \{0.01, 0.025, 0.05, 0.1, 0.15, \ldots, 0.9, 0.95, 0.975, 0.99\}$, which specify a predictive median and the endpoints of central $(1-\alpha)\times 100\%$ prediction intervals at levels $\alpha = 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. For a given week and target date, we use $q_{i,k}$ to denote the submitted quantiles for location i and probability level $\tau_k \in \mathcal{T}$, $k = 1, \ldots, 23$.

In the event that there is some $k \in \{1, \ldots, 23\}$ for which $\sum_i q_{i,k} = K$, i.e., the provided predictive quantiles at level τ_k sum across locations to the resource constraint K, the solution to the allocation problem is given by those quantiles. However, generally this will not be the case; the optimal allocation will typically be at some probability level $\tau^\star \notin \mathcal{T}$.

To address this situation and support the numerical allocation algorithm outlined in section C, we need a mechanism to approximate the full cumulative distribution functions F_i, $i = 1, \ldots, N$ based on the provided quantiles. We have developed functionality for this purpose in the `distfromq` package for R (Ray and Gerding, 2023). This functionality represents a distribution as a mixture of discrete and continuous parts, and it works in two steps:

1. Identify a discrete component of the distribution consisting of zero or more point masses, and create an adjusted set of predictive quantiles for the continuous part of the distribution by subtracting the point mass probabilities and rescaling.
2. For the continuous part of the distribution, different approaches are used on the interior and exterior of the provided quantiles:
   (a) On the interior, a monotonic cubic spline interpolates the adjusted quantiles representing the continuous part of the distribution.
   (b) A location-scale parametric family is used to extrapolate beyond the provided quantiles. The location and scale parameters are estimated separately for the lower and upper tails so as to obtain a tail distribution that matches the two most extreme quantiles in each tail. In this work, we use normal distributions for the tails.

The resulting distributional estimate exactly matches all of the predictive quantiles provided by the forecaster. We use the cumulative distribution function resulting from this procedure as an input to the allocation score algorithm.

We refer the reader to the `distfromq` documentation for further detail (Ray and Gerding, 2023).

E Propriety of parametric approximation

In practice, open forecasting exercises are generally not able to collect a perfect description of the forecast distribution F other than in simple settings such as for a categorical variable with a relatively small number of categories. In settings where the outcome being forecasted is a continuous quantity (such as the proportion of outpatient doctor visits where the patient has influenza-like illness) or a count (such as influenza hospitalizations), forecasting exercises have therefore resorted to collecting summaries of a forecast distribution such as bin probabilities or predictive quantiles. In this section, we address two practical concerns raised by this. First, we discuss conditions under which it is possible

to calculate the allocation score when only summaries of a forecast distribution are recorded in a submission to a forecast hub. Second, we show that a post hoc attempt to compute the allocation score based on submitted predictive quantiles may in fact compute an alternative score that is not proper.

E.1 Propriety when scoring methods are announced prospectively

We consider a setting where a forecasting exercise (such as a forecast hub) pre-specifies that forecasts will be represented using a parametric family of forecast distributions $G_\theta(y)$, and the task of the forecaster is to select a particular parameter value θ. We use $\mathcal{P}$ to denote the collection of all distributions G_θ in the given parametric family. For instance, it has recently been proposed that mixture distributions could be used to represent forecast distributions (Wadsworth et al., 2023). Additionally, we note that the functionality in `distfromq` can be viewed as specifying a parametric family $\mathcal{P}_{\text{dfq}}$ where the parameters θ of G_θ are its quantiles at pre-specified probability levels, and where the shape of any $G_\theta \in \mathcal{P}_{\text{dfq}}$ over the full range of its support is entirely controlled by these quantiles.

We find it helpful now to formally distinguish between two decision making problems. The first is the public health decision maker's allocation problem where the task is to select an allocation x, with the allocation loss $s_A(x, y) = \sum_{i=1}^{N} L \cdot \max(0, y_i - x_i)$ as described in section 2.2.2. The second is the forecaster's reporting problem where the task is to select parameter values θ to report. The forecaster's loss is given by

$$s_R(\theta, y) = s_A(x^{G_\theta}, y), \tag{19}$$

where x^{G_θ} is the Bayes act for the allocation problem under the distribution G_θ. In words, the loss associated with reporting θ is equal to the loss associated with taking the Bayes allocation corresponding to the distribution G_θ.

Following our usual construction, the Bayes act for the forecast reporting problem is the parameter set that minimizes the forecaster's expected loss. Breaking with our earlier notation for improved legibility, we use $\theta^\star(F)$ to denote this Bayes act:

$$\begin{aligned}\theta^\star(F) &= \text{argmin}_\theta \, \mathbb{E}_F[s_R(\theta, Y)] \\ &= \text{argmin}_\theta \, \mathbb{E}_F[s_A(x^{G_\theta}, Y)]\end{aligned}$$

We then arrive at the scoring rule

$$S_R(F, y) = s_R(\theta^\star(F), y) = s_A(x^{G_{\theta^\star(F)}}, y).$$

It follows from the discussion in section A.1 that this is a proper scoring rule for F. Although the full forecast distribution F is not available in the forecast submission, the score $S_R(F, y)$ can be calculated from the reported parameter values as long as the forecaster submits the optimal parameters $\theta^\star(F)$.

We emphasize that the forecaster's true predictive distribution F does not need to be a member of the specified parametric family $\mathcal{P}$ for this construction to yield a proper score. It is, however, necessary to specify the parametric family to use and the foundational scoring rule s_A (including any relevant problem parameters such as the resource constraint K) in advance, so that forecasters can identify the Bayes act parameter set $\theta^\star(F)$ to report.

If the parametric family used to represent forecast distributions is flexible enough, the reporting scoring rule S_R and the allocation score will coincide. Suppose that for a given resource constraint K, for any forecast distribution F it is possible to find a member $G_{\theta^\star}$ of the specified parametric family $\mathcal{P}$ with the same allocation as F (i.e., $x^F = x^{G_{\theta^\star}}$). Then $\theta^\star$ is a Bayes act for the reporting problem since for

any other parameter value θ,

$$\begin{aligned}\mathbb{E}_F[s_R(\theta^\star, Y)] &= \mathbb{E}_F[s_A(x^{G_{\theta^\star}}, Y)] \\ &= \mathbb{E}_F[s_A(x^F, Y)] \quad \text{(since } x^F = x^{G_{\theta^\star}}\text{)} \\ &\leq \mathbb{E}_F[s_A(x^{G_\theta}, Y)] \quad \text{(by definition of } x^F\text{)} \\ &= \mathbb{E}_F[s_R(\theta, Y)].\end{aligned}$$

Thus $S_R(F, y) = s_R(\theta^\star, y) = s_A(x^{G_{\theta^\star}}, y) = s_A(x^F, y) = S_A(F, y)$.

For the particular choice of the parametric family $\mathcal{P}_{\text{dfq}}$ (i.e., using the `distfromq` package), this flexibility requirement is satisfied. For instance, the forecaster could pick one required quantile level (such as 0.5, for which the corresponding predictions are predictive medians), and set the submitted quantiles of their forecast distribution in each location at that level to be the desired allocations, which sum to K across all locations. However, this representation of the forecast may be quite different from the actual forecast distribution F. For example, for the actual forecast distribution F the allocations may occur at some quantile level other than 0.5.

As another alternative for practical forecasting exercises, a forecast hub could ask forecasters to directly provide the Bayes allocations associated with their forecasts for one or more specified resource constraints K. At the cost of increasing the number of quantities solicited by the forecast hub, this would have several advantages: it would prevent any artificial distortion of the forecast distributions, allow for direct calculation of scores, and narrow the gap between model outputs and public health end users. For this to be feasible, implementations of the allocation algorithm would have to be provided to participating forecasters in the computational languages being used for modeling.

E.2 Impropriety of post hoc allocation scoring with quantile forecasts

A *post hoc* evaluation of quantile forecasts that combines the parametric family specified by `distfromq` with the allocation score does not yield the allocation score of the forecast distribution F. Instead, it computes an alternative score that is improper. This is because the forecast distribution F and the distribution $G^q \in \mathcal{P}_{\text{dfq}}$ with the same quantiles as F may determine different resource allocations. In our investigations, these discrepancies appear to be relatively minor on the interior of the provided quantiles, but could be severe if the tail extrapolations performed by `distfromq` do not match the tail behavior of F and the allocations are in the tails of the predictive distribution.

We define

$$G^\star(F) := \operatorname*{argmin}_{G \in \mathcal{P}_{\text{dfq}}} E_F[S_A(G, Y)].$$

Since S_R is defined as the Bayes scoring rule for the forecaster's loss (19), $G^\star(F)$ coincides with $G_{\theta^\star(F)}$, the distribution in $\mathcal{P}_{\text{dfq}}$ given by the optimal submission parameters $\theta^\star(F)$ for the forecaster with predictive distribution F. In general, $G^q(F)$ and $G^\star(F)$ will be different distributions: matching F at specific quantiles does not require $G^q(F)$ to match F at the quantiles for $\tau^{F,K}$ (~~e.f~~cf. (18)), which would be necessary for it to share x^F as an optimal allocation.

When an analyst attempts a post hoc computation of the allocation score using $G^q(F)$ (implicitly assuming that $G^q(F) = G^\star(F)$), they in fact compute the alternative score

$$\tilde{S}(F, y) = S_A(G^q(F), y) = s_A(x^{G^q(F)}, y).$$

This score is improper because $E_F[S_A(G, Y)]$ is minimized by $G^\star(F)$, not $G^q(F)$. In general, we have

$$\begin{aligned}E_F[\tilde{S}(G^\star(F), Y)] &\leq E_F[S_A(G^q(F), Y)] \\ &= E_F[\tilde{S}(F, Y)]\end{aligned} \tag{20}$$

However, the inequality in (20) will typically be strict. For example, if F has heavy upper tails (such as for a lognormal distribution), but normal distributions are used for tail extrapolations in `distfromq`, then the resource allocations based on the distribution $G^q(F)$ may be quite different from the optimal allocations under the distribution F, leading to a strict inequality. This demonstrates that

$\tilde{S}$ is improper.

References

Araz, O. M., Galvani, A. and Meyers, L. A. (2012) Geographic prioritization of distributing pandemic influenza vaccines. *Health Care Management Science*, **15**, 175–187.

Bannigidadmath, D. and Narayan, P. K. (2016) Stock return predictability and determinants of predictability and profits. *Emerging Markets Review*, **26**, 153–173.

Bertsekas, D. (2012) *Dynamic programming and optimal control: Volume I*, vol. 4. Athena scientific.

Bertsimas, D., Boussioux, L., Cory-Wright, R. et al. (2021) From predictions to prescriptions: A data-driven response to COVID-19. *Health Care Management Science*, **24**, 253–272.

Bilinski, A. M., Salomon, J. A. and Hatfield, L. A. (2023) Adaptive metrics for an evolving pandemic: A dynamic approach to area-level COVID-19 risk designations. *Proceedings of the National Academy of Sciences*, **120**, e2302528120.

Bracher, J., Ray, E. L., Gneiting, T. and Reich, N. G. (2021) Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, **17**, e1008618.

Burnett, A. (2014) Urban political processes and resource allocation. In *Progress in Political Geography (Routledge Revivals)*, 177–215. Routledge.

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

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

Colett, J. S., Kelly, J. C. and Keoleian, G. A. (2016) Using nested average electricity allocation protocols to characterize electrical grids in life cycle assessment. *Journal of Industrial Ecology*, **20**, 29–41.

Colón-González, F. J., Bastos, L. S., Hofmann, B. et al. (2021) Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. *PLOS Medicine*, **18**, e1003542.

Cramer, E. Y., Huang, Y., Wang, Y. et al. (2022a) The United States COVID-19 Forecast Hub dataset. *Scientific Data*, **9**, 462.

Cramer, E. Y., Ray, E. L., Lopez, V. K. et al. (2022b) Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, **119**, e2113561119.

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

Diecidue, E. and Somasundaram, J. (2017) Regret theory: A new foundation. *Journal of Economic Theory*, **172**, 88–119.

Du, J., J Beesley, L., Lee, S. et al. (2022) Optimal diagnostic test allocation strategy during the COVID-19 pandemic and beyond. *Statistics in Medicine*, **41**, 310–327.

Ellsberg, D. (1961) Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, **75**, 643–669.

Finger, F., Funk, S., White, K. et al. (2019) Real-time analysis of the diphtheria outbreak in forcibly displaced Myanmar nationals in Bangladesh. *BMC Medicine*, **17**, 58.

Fissler, T. and Ziegel, J. F. (2016) Higher order elicitability and Osband's principle. *The Annals of Statistics*, **44**, 1680 – 1707.

Fox, S. J., Lachmann, M., Tec, M. et al. (2022) Real-time pandemic surveillance using hospital admissions and mobility data. *Proceedings of the National Academy of Sciences*, **119**, e2111870119.

Gebre, S. L., Cattrysse, D. and Van Orshoven, J. (2021) Multi-criteria decision-making methods to address water allocation problems: A systematic review. *Water*, **13**, 125.

Gerding, A. and Ray, E. (2023) *alloscore: Tools for Implementing Allocation Scoring Rules*. URL: https://github.com/aaronger/alloscore. R package version 0.0.9001.

Gneiting, T. (2011a) Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**, 746–762.

— (2011b) Quantiles as optimal point forecasts. *International Journal of forecasting*, **27**, 197–207.

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

Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, **29**, 411–422.

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

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

Huang, H.-C., Araz, O. M., Morton, D. P. et al. (2017) Stockpiling Ventilators for Influenza Pandemics. *Emerging Infectious Diseases*, **23**.

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

Ioannidis, J. P., Cripps, S. and Tanner, M. A. (2022) Forecasting for COVID-19 has failed. *International Journal of Forecasting*, **38**, 423–438.

Johansson, M. A., Apfeldorf, K. M., Dobson, S. et al. (2019) An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, **116**, 24268–24274.

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

Jose, V. R. R. and Winkler, R. L. (2009) Evaluating quantile assessments. *Operations research*, **57**, 1287–1297.

Landau, W. M. (2021) The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, **6**, 2959.

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

Liang, S. and Wey, W.-M. (2013) Resource allocation and uncertainty in transportation infrastructure planning: A study of highway improvement program in taiwan. *Habitat International*, **39**, 128–136.

Marshall, M., Parker, F. and Gardner, L. M. (2023) When are predictions useful? A new method for evaluating epidemic forecasts. *medRxiv*.

McGowan, C. J., Biggerstaff, M., Johansson, M. et al. (2019) Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, **9**, 683.

Meltzer, M. I., Atkins, C. Y., Santibanez, S. et al. (2014) Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014–2015. *MMWR*, **63**, 1–14.

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

Papastefanopoulos, V., Linardatos, P. and Kotsiantis, S. (2020) COVID-19: a comparison of time series methods to forecast percentage of active cases per population. *Applied Sciences*, **10**, 3880.

Pasco, R., Johnson, K., Fox, S. J. et al. (2023) COVID-19 Test Allocation Strategy to Mitigate SARS-CoV-2 Infections across School Districts. *Emerging Infectious Diseases*, **29**.

Persad, G., Leland, R. J., Ottersen, T. et al. (2023) Fair domestic allocation of monkeypox virus countermeasures. *The Lancet Public Health*, **8**, e378–e382.

Pesaran, M. H. and Skouras, S. (2002) Decision-based methods for forecast evaluation. In *A Companion to Economic Forecasting* (eds. M. P. Clements and D. F. Hendry), chap. 11, 241–267. Oxford: Blackwell.

Pflug, G. C. and Pichler, A. (2014) *Multistage stochastic optimization*, vol. 1104. Springer.

Probert, W. J., Shea, K., Fonnesbeck, C. J. et al. (2016) Decision-making for foot-and-mouth disease control: Objectives matter. *Epidemics*, **15**, 10–19.

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

Rainisch, G., Shankar, M., Wellman, M. et al. (2015) Regional spread of Ebola virus, West Africa, 2014. *Emerging Infectious Diseases*, **21**, 444.

Ray, E. L., Brooks, L. C., Bien, J. et al. (2023) Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *International Journal of Forecasting*, **39**, 1366–1383.

Ray, E. L. and Gerding, A. (2023) *distfromq: Reconstruct a Distribution from a Collection of Quantiles*. URL: https://github.com/reichlab/distfromq. R package version 1.0.2.

Reich, N. G., Brooks, L. C., Fox, S. J. et al. (2019) A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences*, **116**, 3146–3154.

Royset, J. O. and Wets, R. J.-B. (2022) *An optimization primer*. Springer.

Sherratt, K., Gruson, H., Grah, R. et al. (2023) Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *eLife*, **12**, e81916.

Syme, G. J., Nancarrow, B. E. and McCreddin, J. A. (1999) Defining the components of fairness in the allocation of water to environmental and human uses. *Journal of environmental management*, **57**, 51–70.

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

University of Texas at Austin (2022) 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).

Wadsworth, S., Niemi, J. and Reich, N. G. (2023) Mixture distributions for probabilistic forecasts of disease outbreaks. *arXiv preprint arXiv:2310.11939*.

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

Zhang, Y., Zeng, Q., Ma, F. and Shi, B. (2019) Forecasting stock returns: Do less powerful predictors help? *Economic Modelling*, **78**, 32–39.